UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS

LABORATÓRIO DE GENÉTICA CELULAR E MOLECULAR
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Tese de Doutorado

Validação de um método para predição de redes de interação proteína-proteína e sua aplicação em *Corynebacterium pseudotuberculosis* para identificar proteínas essenciais

BELO HORIZONTE
2015

Edson Luiz Folador

Validação de um método para predição de redes de interação proteína-proteína e sua aplicação em *Corynebacterium pseudotuberculosis* para identificar proteínas essenciais

Defesa de tese apresentada como requisito parcial para a obtenção do título de Doutor em Bioinformática pelo programa de pós-graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais.

Orientador: Prof. Dr. Vasco Ariston de Carvalho Azevedo

Coorientadora: Profa. Dra. Rafaela Salgado Ferreira

BELO HORIZONTE
2015

Eu dedico este trabalho principalmente a meus pais que, mal concluindo o ensino primário, com toda sabedoria sempre me motivaram a estudar e, na pessoa deles, dedico a todos os cientistas que jamais concluíram o ensino médio por não terem condições de sair dos locais de origem. Dedico também a meus filhos Jiuliane e Eduardo e, na pessoa deles, dedico a todos aqueles que permaneceram por anos distantes do conforto e abrigo de um lar familiar para conseguirem defender suas dissertações e teses. Dedico a minha esposa Adriana e ao nosso filho Arthur por serem agora motivação para eu seguir em frente.

AGRADECIMENTOS

Primeiramente e antes de tudo eu agradeço ao meu orientador professor doutor Vasco de Azevedo, não somente pela sua orientação, mas principalmente por, em um momento muito peculiar, ter acreditado em mim e em minha proposta de trabalho, ter me assistido e dado autonomia para executar o projeto proposto. Não esquecerei a oportunidade que me deste em um momento que todas as outras oportunidades me eram tiradas. Da mesma forma agradeço à professora doutora Rafaela Salgado Ferreira pelo suporte biológico e metodológico durante a orientação.

Sem citar nomes para não ser injusto, agradeço ainda a todos os membros dos grupos de pesquisa do LGCM (UFMG), do LPDNA (UFPA) e colaboradores internacionais, secos e molhados, quais direta ou indiretamente, contribuíram das mais variadas formas para a conclusão deste trabalho.

Agradeço também a toda equipe técnica e administrativa da UFMG e UFPA por todo suporte oferecido.

“A imaginação é mais importante que o conhecimento. ”

Albert Einstein

Resumo

Corynebacterium pseudotuberculosis (Cp) pertence ao grupo CMNR (*Corynebacterium, Mycobacterium, Nocardia, Rhodococcus*), é uma bactéria patogênica intracelular facultativa, gram-positiva, possui fimbrias, porém não se move, não forma capsulas e não esporula, apresenta-se nos biovares *ovis* e *equi*. O biovar *equi* infecta equinos e bovinos. O biovar *ovis* infecta principalmente rebanhos de ovinos e caprinos, sendo o agente etiológico de linfadenite caseosa (LC). Cp é prevalente em diversos países, causando significantes perdas econômicas devido à baixa qualidade de carcaças, queda na produção de carne, lã e leite. Os métodos para diagnóstico e tratamento de LC ainda não são suficientemente eficazes devido Cp apresentar baixa resposta terapêutica e habilidade em persistir no meio ambiente e no hospedeiro, sendo importante entender a biologia deste patógeno a nível sistêmico. Neste aspecto, conhecer as proteínas e suas interações é fundamental para compreender os mecanismos moleculares da célula, sendo as redes de interação proteína-proteína uma boa ferramenta para este tipo de estudo.

Visando gerar a rede de interação para Cp, nos preocupamos em validar uma metodologia para a predição de interações com dados experimentais e curados disponíveis publicamente. Como resultado, além de aumentarmos a cobertura da rede, obtivemos uma área sobre a curva (AUC) entre 0,93 e 0,96, cujo ponto de corte de 0,70 representa uma especificidade de 0,95 e a uma sensibilidade de 0,90.

Com a metodologia validada, foram geradas as redes de interação para nove linhagens do biovar *ovis* de Cp, sendo ~99% das interações mapeadas do gênero *Corynebacterium* e possuindo 15.495 interações conservadas entre as linhagens. Validação quanto ao menor caminho e distribuição do grau de interação sugerem que as redes preditas possuem características de redes biológicas. Adicionalmente, comparamos os valores do Coeficiente de Clusterização, Correlação e R^2 contra redes geradas aleatoriamente e submetemos as redes geradas ao teste de normalidade Shapiro-Wilk. Todos os resultados demonstraram que as redes de interação preditas não possuem uma distribuição aleatória, sugerindo que as redes não foram formadas por interações espúrias, existindo uma influência biológica em sua predição. Com as redes validadas, selecionamos os primeiros 15% das proteínas com maior número de interações e identificamos 181 proteínas essenciais. Apenas a proteína *DNA repair protein* (RecN) não teve homologia com a base de dados de genes essenciais (DEG) e outras três tiveram homologia em apenas um organismo em DEG: *Catalase* (KatA), *Endonuclease* III (Nth) e *Trigger factor* (Tig), sugerindo que podem ser bons alvos para diagnóstico ou desenvolvimento de drogas.

Abstract

Corynebacterium pseudotuberculosis (cp) belongs to the group CMNR (*Corynebacterium, Mycobacterium, Nocardia, Rhodococcus*), is a gram-positive facultative intracellular pathogenic bacterium, have fimbriae, is non-motile, do not form capsules and not sporulate, is presented in serovar *ovis* and *equi*. The serovar *equi* infects horses and cattle. The serovar *ovis* mainly infects herds of sheep and goats, and is the etiological agent of caseous lymphadenitis (CLA). Cp is prevalent in many countries, causing significant economic losses due to poor quality carcasses decrease in the production of meat, wool and milk. Methods for diagnosis and treatment of CLA are not yet effective enough due Cp have low therapeutic response and ability to persist in the environment, making it an important organism to be researched and understood the systemic level. In this regard, knowing the proteins and their interactions is crucial to understand the molecular mechanisms of the cell, being protein-protein interaction networks an important tool for this type of study.

Aiming to generate the Cp interaction network, we worry about validate a methodology for the prediction of interactions with experimental and cured data publicly available. As a result, in addition to increasing the coverage of the network, we obtained an area under the curve (AUC) between 0.93 and 0.96, representing the cutoff of 0.70 a specificity of 0.95 and a sensitivity 0.90.

With the validated methodology, the interaction networks were generated for nine serovar *ovis* Cp strains, being ~99% of interactions mapped from *Corynebacterium* gender, possessing 15,495 interactions conserved between strains. The shortest path and the degree interaction distribution analysis suggests the predicted networks have biological characteristics. Additionally, we compared the values of the clustering coefficient, Correlation and R^2 against randomly generated networks and submit the networks generated to the Shapiro-Wilk normality test. All results show that the predicted interaction networks do not have a random distribution, suggesting the networks were not formed by spurious interactions, existing biological bias its prediction. With validated network, we selected the first 15% of the proteins with more interactions and we identified 181 essential proteins. Only the protein DNA repair protein (RecN) had no homology against database of essential genes (DEG) and other three had homology in just one DEG organism: Catalase (KatA), Endonuclease III (Nth) and trigger factor (Tig), suggesting they may be good targets for diagnosis and drug development.

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

AUC	*Area Under Curve*
BLAST	*Basic Local Alignment Search Tool*
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CENAPAD	Centro Nacional de Processamento de Alto Desempenho
LC	Linfadenite Caseosa
CMNR	*Corynebacterium, Mycobacterium, Nocardia, Rhodococcus*
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
Cp	*Corynebacterium pseudotuberculosis*
DEG	*Database of Essential Genes*
DIP	*Database of Interacting Proteins*
DNA	Acido desorribonucleico
Fapemig	Fundação de Amparo à Pesquisa do Estado de Minas Gerais
LGCM	Laboratório de Genética Celular e Molecular
LPDNA	Laboratório do Polimorfismo do DNA
pDB	Bases de dados públicas (*public databases*)
PPI	Interação proteína-proteína (*protein-protein interaction*)
RNA	Ácido ribonucléico
ROC	*Receiver Operating Characteristic*
STRING	*Search Tool for the Retrieval of Interacting Genes/Proteins*
tRNA	RNA transportador
UFMG	Universidade Federal de Minas Gerais
UFPA	Universidade Federal do Pará

Sumário

Apresentação

Colaboradores

Este trabalho foi auxiliado pelo Centro Nacional de Processamento de Alto Desempenho (CENAPAD-MG) situado na Universidade Federal de Minas Gerais (UFMG) e foi executado no Laboratório de Genética Celular e Molecular (LGCM) da UFMG e no Laboratório de Polimorfismo e DNA (LPDNA) da Universidade Federal do Pará (UFPA) em colaboração com os seguintes pesquisadores:

- Prof. Dr. Vasco Ariston de Carvalho Azevedo, Pesquisador e Professor do LGCM/UFMG, Brasil;
- Prof. Dra. Rafaela Salgado Ferreira, Pesquisadora e Professora do Departamento de Bioquímica e Imunologia da UFMG, Brasil.
- Prof. Dr. Artur Luiz da Costa da Silva, Pesquisador e Professor do LPDNA/UFPA, Brasil.
- Prof. Dr. Debmalya Barh, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal, India.
- Prof. Dr. Richard Röttger e Dr. Jan Baumbach, Departamento de Matemática e Informática, Universidade do Sul da Dinamarca, Campusvej 55, Odense, Denmark
- Dr. Preetam Ghosh, Departamento de Ciência da Computação, Universidade Virginia Commonwealth, Richmond, VA, USA.

Este trabalho foi financiado pelas agências de fomento: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig).

Contextualização

Coordenados pelo grupo de pesquisa do Laboratório de Genética Celular e Molecular (LGCM) da Universidade Federal de Minas Gerais (UFMG) e do Laboratório de Polimorfismo e DNA (LPDNA) da Universidade Federal do Pará (UFPA), até o ano de 2014, quando esta tese começou a ser desenvolvida, haviam 21 genomas de *Corynebacterium pseudotuberculosis* sequenciados. Destes genomas, 15 estavam completos e publicamente disponíveis, sendo nove genomas do biovar *ovis* e seis genomas do biovar *equi.*

Os grupos de pesquisa, objetivando desenvolver projetos relacionados a genômica comparativa e um grande projeto de patogenômica, estavam sequenciando ainda outras novas linhagens do biovar *equi* de *C. pseudotuberculosis*, enquanto outras montagens antigas estavam sendo aperfeiçoadas e resequenciadas com as novas tecnologias.

Os vários genomas de *C. pseudotuberculosis* e outros organismos disponíveis, possibilitou ao grupo desenvolver em 2013 o primeiro trabalho de redes de interação proteína-proteína baseado no interactoma conservado entre patógeno-hospedeiro (Barh *et al.*, 2013). Com o interesse do grupo em fortalecer o desenvolvimento de projeto na área de redes de interação, foi proposto em se gerar as redes de interação proteína-proteína interna para a bactéria *C. pseudotuberculosis*. Visto que o biovar *ovis* possuía a maior quantidade de genomas disponíveis (nove) e também ser mais clonal, este biovar foi selecionado para a predição das redes de interação proteína-proteína, visando futuramente comparar estas redes com as redes de interação do biovar *equi*.

Limitações como custo e tempo foram impeditivos para realizar este trabalho experimentalmente para os nove proteomas disponíveis, optando-se assim pelo desenvolvimento *in silico* das redes de interação. A revisão bibliográfica apontou a existência de diversos métodos computacionais para a predição de rede de interação, sendo que cada método usa como entrada distintos tipos de dados biológicos. Uma característica comum entre estes métodos foi a ausência de informações na literatura sobre os detalhes de suas implementações e também sobre as formas de validação em larga escala que comprovasse a eficácia nas predições.

Assim, antes de aplicar um destes métodos para a predição das interações em *C. pseudotuberculosis* biovar *ovis*, houve a preocupação de selecionar um método que pudesse oferecer uma boa cobertura na predição das interações e, ao mesmo tempo, oferecesse uma boa razão entre sensibilidade e especificidade na predição. Adicionalmente,

houve a preocupação em validar este método com dados experimentais e curados em larga escala, visando identificar exatamente os índices de erros e acertos na predição.

Pensando em todo este contexto, ao contrário de estruturas tridimensionais de proteínas que não são abundantes para *C. pseudotuberculosis* e outros organismos não modelo, foi selecionado um método que permitisse o uso dos dados mais abundantes de *C. pseudotuberculosis*, ou seja, os seus genomas e proteomas. Assim, considerando os recursos físicos e conhecimento disponível no laboratório para a implementação do projeto, foi selecionado o método denominado mapeamento de interações ortólogas (*interolog mapping*) para ser usado nas predições das redes de interação proteína-proteína de *C. pseudotuberculosis* biovar *ovis*, cuja validação seria possível com dados experimentais e curados disponíveis publicamente.

Estrutura da Tese

Esta tese está organizada em formado de artigos e foi dividida em cinco capítulos. Mesmo estando em formato de artigo, a tese segue a linha clássica de escrita de trabalhos científico, apresentando inicialmente a introdução sobre os principais temas abordados na tese, seguido da apresentação da metodologia, dos resultados obtidos e finalizando com a discussão geral, conclusão e perspectivas.

Segue uma breve apresentação dos cinco capítulos que compõe esta tese:

a. No primeiro capítulo é apresentado a introdução da tese. Como esta tese é referente ao desenvolvimento e validação de uma metodologia para a predição de interações proteína-proteína, seguido da aplicação desta metodologia para a predição das interações de *Corynebacterium pseudotuberculosis*, a introdução foi também dividida em três seções, duas destacando as redes de interação proteína-proteína e a última destacando o organismo estudado:
   - A primeira seção, com o subtítulo “Application to a Bacterial Protein-Protein Interaction”, foi publicada em fevereiro de 2015 pela revista SM Online Publishers LLC e apresenta o capítulo de livro intitulado “Genomics”, do livro “A Textbook of Biotechnology”.
   - A segunda seção, com o título “*In silico* protein-protein interactions: avoiding data and method biases over sensitivity and specificity” foi publicado em maio de 2015 pela revista Current Protein & Peptide Science.
   - A terceira seção apresentando a introdução sobre *C. pseudotuberculosis* e as características principais deste organismo.

b. No segundo capítulo é apresentado a metodologia. O artigo referente a validação do método intitulado “An improved interolog mapping-based computational prediction of protein-protein interactions with increased network coverage”, foi publicado na revista Integrative Biology em novembro de 2014, cuja validação das métricas permitiu realizar a predição *in silico* de redes de interação proteína-proteína para *C. pseudotuberculosis.*

c. No terceiro capítulo são apresentados os resultados obtidos no desenvolvimento desta tese, relacionados à aplicação da metodologia validada para a predição das redes de interação de *C. pseudotuberculosis*. Este capítulo está dividido em dois trabalhos:

- O primeiro trabalho, com o título “*In silico* protein-protein interaction analysis reveals conserved essential proteins in nine *Corynebacterium pseudotuberculosis* serovar *ovis* strains”, submetido à revista Integrative Biology em agosto de 2015.
- O segundo trabalho, com o título “Label-free proteomic analysis to confirm the predicted proteome of *Corynebacterium pseudotuberculosis* under nitrosative stress mediated by nitric oxide”, publicado em dezembro de 2014 pela revista BMC Genomics.

d. No quarto capítulo é apresentado uma discussão geral considerando todos o conteúdo desenvolvido nesta tese.

e. No quinto capítulo são apresentadas as conclusões e as perspectivas de trabalhos futuros.

Durante o desenvolvimento desta tese, colaborando com outros integrantes dos grupos de pesquisa, outros trabalhos foram desenvolvidos. Assim, estes trabalhos publicados estão relacionados no anexo desta tese, também em formato de artigo.

Por uma questão de organização, quando constar na tese um artigo publicado, este será apresentado integralmente em seu respectivo capítulo, conforme publicado pela revista. Como as figuras, tabelas, referências bibliográficas e materiais suplementares recebem formatação e numeração própria em cada artigo, estes itens figurarão somente no respectivo artigo, no capítulo que descreve o artigo, sem serem apresentados na lista de figuras ou tabelas da tese. Da mesma forma, visando não misturar as referências bibliográficas dos artigos publicados, que são distintas na forma de apresentação e organização para cada revista, estas estarão exclusivamente ao final da apresentação de cada artigo ou do respectivo material suplementar quando este existir.

1 - Introdução

1.1- Genomics: Application to a Bacterial Protein-Protein Interaction

Flavia Figueira Aburjaile, Mariana P. Santana, Marcos Vinicius Canario Viana, Wanderson Marques Silva, Edson Luiz Folador, Artur Silva e Vasco Azevedo

Neste capítulo de livro, foi feito uma breve revisão sobre genômica estrutural, genômica funcional (transcriptomica) e proteomica, destacando os métodos de análise experimentais de cada área. Adicionalmente, foram revisados os conceitos básicos relacionados às redes de interação proteína-proteína com uma breve discussão para possíveis aplicações biotecnológicas.

Uma rede de interação é composta por nodos, no contexto deste trabalho, representando as proteínas e, por arestas, que ligam dois nodos e caracteriza uma interação. Independente do método usado, par-a-par, é possível formar uma complexa rede de interação proteína-proteína que viabiliza o estudo e compreensão de um organismo a nível de biologia de sistemas. Além de possibilitar um melhor conhecimento do organismo, uma rede de interação pode ser utilizada para direcionar o desenvolvimento de novas pesquisas em laboratório e novas aplicações biotecnológicas, bem como auxiliar na seleção de proteínas para o desenvolvimento de drogas, inclusive para inibir interações específicas.

A seção “Application to a Bacterial Protein-Protein Interaction” que integra o capítulo intitulado “Genomics” do livro “A Textbook of Biotechnology”, foi publicada em fevereiro de 2015 pela revista SM Online Publishers LLC, disponível em http://www.smgebooks.com/a-textbook-of-biotechnology/index.php.com com ISBN número 978-0-9962745-3-1.

SMGroup

Title: A Textbook of Biotechnology

Editor: Zahoorullah S MD

Published by SM Online Publishers LLC

ISBN: 978-0-9962745-3-1

First Published April, 2015

Online Edition available at www.smgebooks.com

SMGroup

Genomics

Aburjaile FF[1],Santana MP[1], Viana MVC[1], Silva WM[1], Folador EL[1], Silva A[2] and Azevedo V[1]

[1]Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Brazil

[2]Universidade Federal do Pará, Bélem, Brazil

***Corresponding author:** Vasco Azevedo, Universidade Federal de Minas Gerais/Instituto de Ciências Biológicas, Minas Gerais, Brazil, Email: vasco@icb.ufmg.br

Published Date: April 15, 2015

ABSTRACT

In the last few years, the technologies have revolutionized new areas in science especially those involving genomics, proteomics and transcriptomics. This chapter approached each of these areas showing their concepts, importance and applications.

Keywords: Genomics; Structural genomics; Functional genomics; Transcriptomics; Proteomics

STRUCTURAL GENOMICS

Introduction

Structural genomics is the analysis of sequence and structure of the genome elements as genes, regulators and mobile elements. Sequencing is necessary to obtain this information. The identification of variants among genomes has application in evolution studies, health, biotechnology and the comprehension of relation between genotype and phenotype. Knowledge of genome sequence is fundamental to obtain this kind of information. Sequencing is the process of characterization of nucleotide sequence at a region (Targeted Genome Sequencing, TGS) or in the entire genome (Whole Genome Sequencing, WGS).

The first complete genome sequence of organism was achieved in 1995, from the Gram-negative *Haemofilus influenzae*. Since then, the number of sequenced genomes is increasing, mainly because of new technologies that reduce the required cost and time, and new bioinformatics tools able to process the volume of data generated. Initiated formally in 1990, The Human Genome Project (HGP) was an international, collaborative project that sequenced the human genome in 13 years at a cost of approximately US $ 2.7 billion [1].

Genome Sequencing

The procedure for sequencing of complete genomes involves fragmentation of DNA molecule, creation of genomic libraries, reading nucleotide sequence of each fragment and reassembles the genome by alignment of the *reads*. Since the 1970s, different methods have been developed to determine the nucleotide sequence of a DNA molecule.

Sequencing generations and technologies

In 1977, Sanger and Nicklen published a DNA sequencing method by chain termination. This method required the creation of libraries by insertion of DNA fragments to be sequenced into cloning vectors (plasmids or artificial chromosomes) and amplification *in vivo*. Four sequencing reactions are performed for each fragment, each one containing a proportion of one nucleotide with no 3' OH group (dideoxy nucleotides), which causes termination of synthesis of new DNA strands in random positions. The generated fragments are separated by size on gel electrophoresis and nucleotide sequence is determined by the last nucleotide of each fragment [2]. Over the years, improvements have been incorporated into the method, such as *in vitro* amplification and automation by thermal cyclers, labeling the amplified fragments with fluorescent dideoxynucleotides (ddNTP's), as well as automatic sequencing based on reading fragments by capillary electrophoresis. The sequenced fragments generated by this method have a maximum size between 800 and 1000 bases. According to the National Human Genome Research Institute (NHGRI), in 2001, the cost to sequence 1 Mb (1,000,000 bases) and a human genome were respectively U$ 5,292 and U$ 95,293,072. In 2007, the cost was reduced to U$ 397 and U$ 7,147 [3]. The amount of data generated by chain termination technology was inadequate for sequencing projects and obtains complete genomes, even with the reduction of required costs and time.

In 2005, Next Generation Sequencing (NGS) platforms were launched, belonging to the second generation. These platforms are characterized by (i) the use of genomic libraries independent of cloning and (ii) a higher throughput compared to the first generation, by means of parallel sequencing of thousands or millions of fragments. However, the necessity of amplification for preparing libraries can bias the quantification of fragments, insert sequence errors during the replication and increase the complexity of sample preparation. In addition, the smaller size of the *reads* and the large amount of generated data make the assembly of genome more complex and creates a demand for computational structures with greater processing and storage capacity.

With the use of these platforms in 2014, the cost for sequencing 1Mb and a human genome decreased to, respectively, U$ 102 and U$ 3.063 million in 2007 to $ 0.05 to U$ 4905 [3]. The first released NGS technology was pyrosequencing technology. The sequencing reaction occurs in emulsion oil droplets distributed in wells of a fiber optic slide. Each oil droplet contains one bead linked to a single library fragment and PCR reagents. The reaction receives a flow of each nucleotide at a time followed by a wash step. The incorporation of a nucleotide is detected by the releasing of pyrophosphate. This technology is used in 454 Genome Sequencer platform (Roche), launched in 2005, being able to generate 200.00 *reads* of 110 base pairs (bp), and generating approximately 20MB of data. In 2007, Illumina/Solexa launched 1G Genetic Analyzer platform. In this technology, the fragments connected to adapters are denatured and its ends hybridized to complementary sequence adapters present on the surface of a solid blade. Complementary strand are synthesized by extension of the adapter attached to a solid plate. Also in 2007, Applied Biosystems launched SOLiD platform, based on ligation sequencing. A new complementary DNA strand is synthesized by ligation of an oligonucleotide to a primer and cleavage of oligonucleotide 3' end, releasing a fluorescent color, which represents the first two nucleotides of the incorporated oligonucleotide. The new DNA strand is removed by cleavage of the last nucleotide primer and a new cycle begins, initiated one position behind in relation to previous cycle. A color code and the sequence of each cycle colors are used to determine the nucleotide sequence. In 2010, Ion Torrent (Applied Biosystems) released the Personal Genome Machine, which sequencing chemistry is similar to pyrosequencing, however the incorporation of nucleotides is detected by pH variation. This technology uses a semiconductor in place of fluorescence and a scanning camera, resulting in higher speed, lower cost and smaller equipment. Since then, updates of each platform were launched, in order to increase the size of fragment sequences and reduce the required time, cost and labor.

The third generation is characterized by (i) non-interruption of the sequencing reaction to detect each nucleotide incorporated (flow type of each nucleotide), (ii) larger *reads* and (iii) sequencing of a single molecule. This has reduced the sequencing time and simplified genome assembly in relation to the second generation, as *reads* sizes exceeds 1000pb. This generation technologies are based on imaging of singles molecules of DNA polymerase as they synthesize a single molecule of DNA; threading DNA polymerase to or next a nanopore and detection of nucleotide passage; or imaging of individual DNA strands by advanced microscopy. One example is the Platform II PacBio RS (Pacific Biosciences), launched in 2013, that generate *reads* longer than 20kb.

Genomic libraries

The DNA library preparation is one of the fundamental steps for the success of a genome project. Despite the variation in protocols, the similarities are fragmentation of the sample to sizes between 50 and 500 nucleotides, selection of fragments by size, ligation to adapters. The adapters often contain elements for ligation on a solid surface, primer annealing sites for amplification

and sequencing, and a barcode sequence that makes possible to sequence samples of different sources.

In single-end or fragments libraries, adapters are ligated to one fragment, generating one *read*. In paired-end libraries, the ends of a DNA fragment are *read*, generating two *reads*, separated by a gap of known size. In mate-pair libraries, DNA fragments are connected to an adapter, circularized, and their ends are *read*, generating two *reads*, separated by a gap of known size. The difference between paired-end and mate-pair library is circularization and larger fragments of the second one. The latter two libraries provide an approximate distance between two *reads*, and this information facilitates the genome assembly and makes it possible to detect genome deletions, inversions, duplications and nucleotide insertions.

Applications

In studies with the genomic DNA, NGS technologies have been used in sequencing of complete genomes or fragments, analysis of genetic variation, chromatin immunoprecipitation followed by sequencing (ChIP-Seq), epigenetic markers, identification of genes associated with disease, genetic screening, monitoring of infectious agents, forensic research, metagenomics and agrogenomics. Thousands of sequencing projects and millions of genomes were released [4]. Among the RNA sequencing applications are analysis of gene expression and single cell transcriptome. In gene expression studies, NGS made the prior knowledge of genes unnecessary to the identification and quantification of transcripts, alternative splicing and sequence variations. The various applications of NGS led to the launch of benchtop sequencers, cheaper, faster, easier to use and accessible to small laboratories. These include GS Junior (Roche), Mi Seq (Illumina) and Personal Genome Machine (Life Technologies).

Genome Assembly

After sequencing the genome is reconstructed from fragments *reads*, in a process called genome assembly. The assembly is performed by means of alignment and overlap of *reads* to generate *contigs (Figure 1), and contigs* overlap. It is important that all nucleotides of a genome are represented among the sequence *reads*. The *theoretical (or expected) coverage* of a genome refers to the average expected times a nucleotide is sequenced. The calculation is performed by multiplying the size of the *reads* by the number of *reads,* and dividing the result by the size of the sequenced genome. The *depth of coverage* refers to the number of times a particular nucleotide of the genome is represented in the *reads.* This number varies because the fragment library preparation, sequencing and the process of assembling the genome generate deviations in the distribution of *reads.* The *breadth of coverage* is the percentage of a given genome sequenced by a number of *reads.*

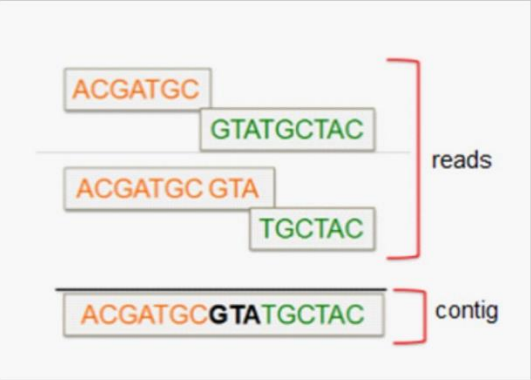


Figure 1: Assembly of *contigs* sequences by *reads* alignment.

Data analysis

The first assembly step is data analysis. Raw data from sequencer is converted to file formats used by assembly software. The main formats are FASTQ and FASTA. FASTA (Fast Alignment) format stores nucleotide sequences (Figure 2), while FASTQ stores nucleotide sequence and quality scores. Removal of poor quality sequences, and barcode *reads* is required, usually performed automatically by the assembly software. *Reads* quality is measured by the Phred index, a measure of error probability in identification of each nucleotide in a *read*.

```
>Sequencia_1
ATCGATCGATGCTAGCATCGATCGATCGATCAGCTAGCTAGCTACTACGATCGATCA
GTCAGCTAGCATGCATCGATCGATCGATCGATCGATCAGCTAGCTAGCTAGCATCGA
GCTAGCTAGCATCGATCGATCGATCAGTCAGCATGCATGCATCGATGCACACACACA
CACACCACACACACGTGTGTCAGCTAGGCTCGCGCGCGCGCCCGTACGATCGGCCAC
ATCGATCGATGCTAGCATCGATCGATCGATCAGCTAGCTAGCTACTACGATCGATCA
>Sequencia_2
ATCGATCGATGCTAGCATCGATCGATCGATCAGCTAGCTAGCTACTACGATCGATCA
GTCAGCTAGCATGCATCGATCGATCGATCGATCGATCAGCTAGCTAGCTAGCATCGA
GCTAGCTAGCATCGATCGATCGATCAGTCAGCATGCATGCATCGATGCACACACACA
ATCGATCGATGCTAGCATCGATCGATCGATCAGCTAGCTAGCTACTACGATCGATCA
GTCAGCTAGCATGCATCGATCGATCGATCGATCGATCAGCTAGCTAGCTAGCATCGA
```

Figure 2: Fasta file containing two sequences (multifasta file).

Assembly

To assembly a genome is necessary to find overlaps between *reads* and generate a consensus, or contiguous, sequence. The assembly software uses algorithms that create substrings of each *read*, called k-mers, and test alignments with k-mers of other *reads* (Figure 3) to generate contiguous sequences. The use of paired libraries facilitates the assembly process because the known distance among the pairs of readings. For example, if each paired *reading* belongs to a different *contig*, it is possible to determine the distance and orientation between these *two contigs*. A genome can be assembled using *de novo* or *ab initio* methods and by reference method.

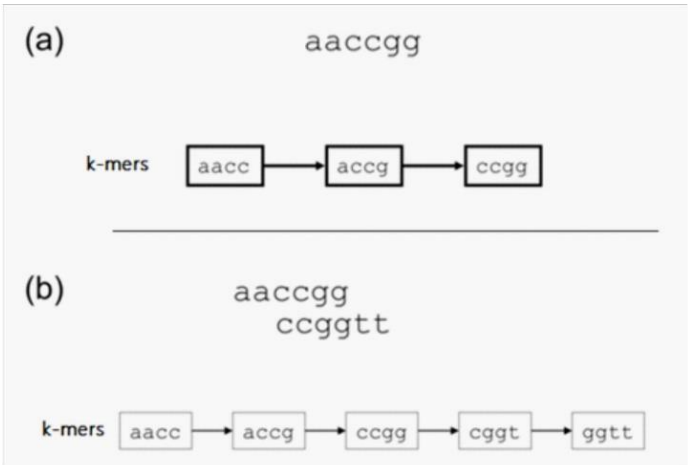


Figure 3: **(A)** k-mers of 4 nucleotides based on the *read* sequence "aaccgg".

(B) Alignment of the *reads* "aaccgg" and "ccggtt" based on k-mers.

The *de novo* assembly is performed only by *reads* overlapping, not requiring the use of a reference genome mapping (Figure 1). The most commonly used algorithms are the Greedy, OLC (Overlap-consensus) and De Bruijn graph. This method allows identification of absent regions in a reference genome. The difficulty of this method is repetitive regions where *reads* can be aligned in wrong places.

In the assembly by reference, the *reads* are mapped in a phylogenetically close genome. This method has lower computational cost and is useful in identifying SNPs, insertions and deletions between the reference and the new genomes (Figure 4). Another advantage is the resolution of repetitive regions, because the *reads* are distributed at the reference regions. However, absent sequences in the reference genome will not be mapped.

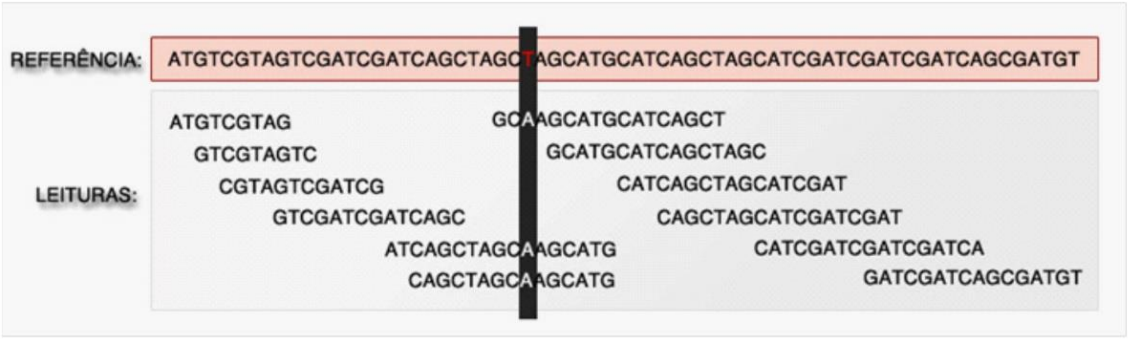


Figure 4: Mapping *reads* on a reference genome.

Scaffolding and closure of gaps

The *contigs* have to be ordered and oriented by overlaps between their ends (*de novo* assembly), mapping on reference genome or optical mapping. The latter is done by mapping restriction sites *in vitro* and comparison to the *contigs*. A contig is called *scaffold* after its position

and orientation is defined. The not identified overlaps between the *contigs* and, mainly, uncovered regions cause gaps in the assembly. Gaps can be closed *in silico* by overlapping *scaffold* ends, extension of *scaffold* by paired *reads* libraries, mapping *reads* in another reference genome in the region equivalent to the gap. The closing is accomplished by *in vitro* primer design which is long to the edge of two adjacent *scaffold*s and subsequent amplification of the region of the gap, usually by the Sanger method by generating sequences which can cover or extend the gap ends of the *scaffold*s. The third-generation sequencing platforms which generate long readings are becoming obsolete closing gaps *in vitro*, mainly due to the cost required. If a genome is still contains gaps after assembly, it is called a *draft*. The most commonly used parameters to measure the quality of an assembly is the number of *contigs*, minimum and maximum contig size and N50. In a good assembly, the number of *contigs* and its sizes is, respectively, the minimum and the maximum possible. The N50 is the size of the contig which the sum with the others, in descending order of size, is greater than or equal to 50% of the assembled genome.

Genome Annotation (Automatic and Manual Steps)

After assembly, the structure and function of genome elements as genes, regulators and other non-coding sequences have to be identified to posterior analysis. This process is called genome annotation.

Gene prediction can be made empirically, by similarity with known sequences in databases and proteins, or *ab initio*, by computational techniques, using algorithms to find stop codons, coding sequences (CDS), start codons.

Normally, the algorithms to predict coding sequences are based on Markov chain models. In Prokaryotes, gene prediction is based on Open Reading Frames (ORFs) in the six possible frames. An ORF is a sequence that initiates at a start codon and finish at the nearest 3' stop codon. The prediction of false positives, incorrect start codons and overlapping genes are the main problems of this method. This problem can be minimized by searching for ribosome binding sites (RBS) to indicate a true start site and homologies in closely-related organisms.

The occurrence of introns and alternative splicing makes Eukaryotic gene prediction more complex. The prediction is based on recognizing signal functions on DNA strand and posterior accurate prediction of gene structure and organization. Sequences are classified as coding and non-coding and functions are assigned by alignments with functionally-related documented sequences, involved in transcription, translation and splicing.

After prediction, annotation can be done automatically, by align similarity with sequences at databases and transferring the information about their function to the new sequence. Databases can hold information of nucleic acids (e.g. NCBI, EMBL), protein sequences (e.g. UniProtKB), domains (e.g. Pfam) and structures (e.g. Protein

Data Bank), metabolic pathways (e.g. Kyoto Encyclopedia of Genes and Genomes), gene

expression (e.g. dbEST), gene ontology and evolutionary relationships. Different databases are integrated by automatically annotation services, as provided by RAST (Rapid Annotation using Subsystem Technology). Automatic annotation can provide a great amount of information about a new genome from its sequence. However, not all databases are fully curated, updated or have standard terminology. Manual curation is done by checking the results for different databases to identify protein domains and functions, correct start codon positions, gene names, gene products and to identification of frameshift.

After annotation, the genome can be submitted to a database, as submission is a pre-requisite to publication in scientific journals. Once this data is available, it can be used researches about genetic diversity, evolution, biotechnology and health.

Comparative Genomics

After sequencing, assembly and annotation of a genome, different analyzes will try to find out to remove as much information as possible from a genome of an organism. In this context, there is the comparative genomics in order to seek from genomic data, which are the between intra and inter-specific relationships. This can be studied using the homology the genes, loci, functions, processes and categories.

Currently, there are a number of tools available for the comparison of genomes at the level of their genes and proteins, among them: BLAST, but their visualization is somewhat limited, hindering an overview of the genomes studied. However, emerged over the last decades other tools, which integrate different databases and integrate visualization and their processing, widely used by bioinformatics are they: CGView [5] and BLAST Ring Image Generator *(BRIG)*. With these tools you can see the structure and organization of genes in one or more genomes through a graphical interface (Figure 6). The other programs, such as the PIPS, are already able to report data pathogenicity, improving the entire information and providing the user various relevant information to do exploration of the body [6].

All of these tools require a computing power to process a large amount of data, since genomes can be processed genomes from of simple organisms (prokaryotes) to more complex organisms (eukaryotes). Another important detail to highlight is the fact that numerous incomplete genomes are being deposited in the National Center for Biotechnology Information **(**NCBI). Incomplete genomes are those assigned "*draft*", which will have a percentage of genetic information lost one time analyzed using these tools,this is because part of the "gap" cannot be correlated to anything since it is said to be unknown. Therefore, we emphasize the importance of finalizing a genome in all steps, assembly and annotation, thus avoiding the loss of biological information.

FUNCTIONAL GENOMICS

Transcriptomics

Concepts and evolution

Transcriptomics is the global study of gene expression levels of all the genes in an organism (genome-wide expression profiling), in the order words, set of gene transcripts detected in a given cell. Transcripts of the profiles are modulated by factors such as environment, time, diseases and response to adverse processes.

The first idea of transcriptome emerged in the late70s, emerged the Dot-Blot technique in order to use nucleic acids for the simultaneous analysis of some genes in arrays. In 1995, with the evolution of different Technologies, this method has been enhanced by improving their level of sensitivity in optical detection. In 2002, the Affymetrix launch the expression of identification of unknown genes and detection to splice variants. Finally, in 2008, the RNA-seq technique revolutionizes history with sequencing from new generation technologies (Figure 5 shows the timeline of transcriptome).

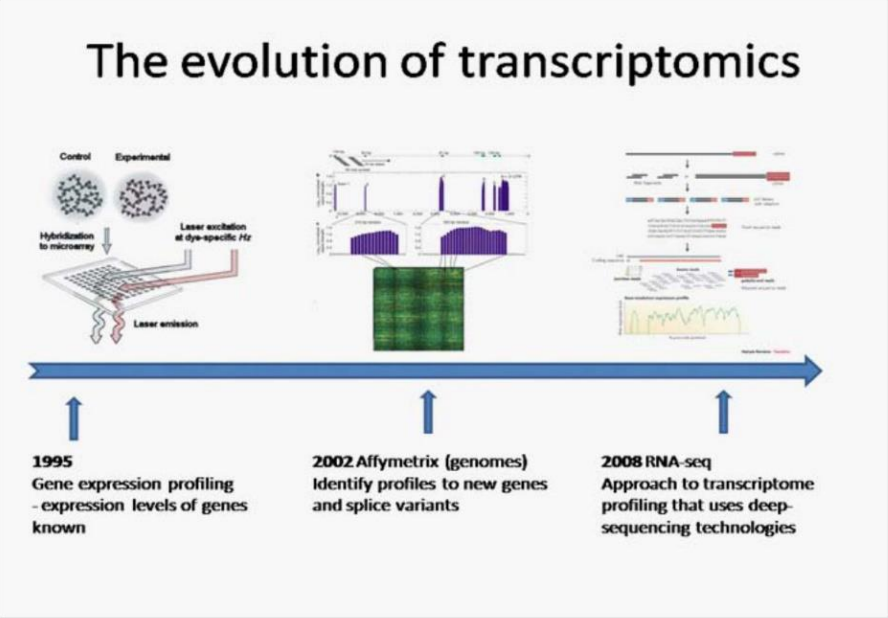


Figure 5: The evolution of transcriptomics.

Different RNAs

Total RNA is the combination of coding and noncoding; being the first comprises the messenger RNA (mRNA), transcript that carries the code information for proteins synthesis. In prokaryotes, the mRNA can be transcribed in operons or single genes and contain the exact transcript, and in eukaryotes, the mRNA has fragments that are translated in proteins (exons) and other regulatory regions (introns) that are removed by a process call splicing.

The noncoding RNAs are a combination of different classes: ribosomal RNA (rRNA), transfer RNA (tRNA), small RNA (sRNA) and long noncoding RNA (lncRNA). This topic will focuses in the non-coding ones, since the coding ones represents less than 2% of the total transcripts of human cells, and to be more precisely the emphases will be given to the regulatory RNAs, which are need to maintain genetic integrity and gene expression in a fine tune. There is still a lot to find out about those types of RNA and as a result of the use of new technologies there set are continually rising.

The tRNAs were known to be responsible for transport specific amino acids for protein synthesis and also recognize their specific codons, but recently another function were attribute to this RNA, when they are exposed to environmental stress in eukaryotes. After cleaved by ribonucleases, tRNAs derived fragments are able to inhibits translation and consequently forms stress granules where the non-translated mRNAs are storages [7]. Therewith, now the tRNAs are also recognized as a new class of regulatory RNAs in response to stress.

sRNA are responsible for mediate posttranscriptional gene silence (PTGS) and transcriptional gene silence (TGS). In PTGS, the target is cleaved or transcriptionally repress after is bound with the small RNA. TGS is a protective mechanism of genome integrity that keeps a heterochromatic state through DNA methylation or histone modifications. One of the most known type of sRNA, in eukaryotes, are the miRNA, noncoding RNAs responsible for regulates the expression of genes in a sequencing specific matter. Others types are siRNA, piRNA and sRNAs (prokaryotes). siRNAs are double-stranded RNA (dsRNA), like miRNA, which directs transcriptional silence of targets, participates in histone modification, heterochromatin formation. Piwi-interacting RNA (piRNA) is a defense system against transposons that acts through deposition of modifications in histones, inducing a heterochromatic state. The piRNAs are present in the genome in clusters and are a germ line specific class from both male and female organisms, and different from the others classes of sRNAs, piRNA requires the Piwi protein to be generated.

In prokaryotes, regulatory noncoding RNAs are transcripts of 50 to 250 nucleotides known as sRNAs that can either bound to mRNA and regulates gene expression or bind to proteins and alter their function. sRNAs can act on a target in *cis or trans*, and are involved in stress response, virulence, quorum sensing, DNA elimination, among others. A recently discover mechanism is the clustered regularly interspaced short palindromic repeat (CRISPR) pathway which are responsible for direct sequence-specific cleavage of foreign DNA.

Another type of regulatory RNA are the lncRNAs, transcripts with proximally 200 nucleotides or more that is not translate into proteins. They include different classes of RNA like enhancer RNAs (eRNA), small nucleolar RNA (snoRNA), intergenic and overlapping transcripts. Independent of the class, the lncRNA are cell specific with high selective pressure, considering the genetic sequences, the promoters and there function structure, and also can function in *cis* and in *trans*, affecting neighbor genes or more distant ones.

An important aspect of the lncRNAs is that they interact with diverse proteins, forms nucleus domains, interacting two distinct chromosomal regions and modulates the chromatin. The lncRNA can act as a regulatory element, having an influence in histone modification, activating and repressing, even silence alleles of several genes (genomic imprinting). One example is XIST, a lncRNA involved in the inactivation of the X chromosome in mammals, and aberrant XIST showed to be involved in human cancer and an alteration on a XIST regulator (TSIX – antisense noncoding regulator) result in alteration X inactivation and embryonic lethality. lncRNA are also involved in the development of central nervous system, heart among other organs and environment responses.

The field of study of the ncRNA is just getting started and there is still a lot to learn about then and what they influence in the eukaryote and prokaryotes life, and because of that the scientific world is focusing all efforts on the study of those transcripts, the real responsible for the complexly of organisms.

METHODOLOGY OF STUDY: ADVANTAGES AND DISADVANTAGES

Microarray X RNA-seq

Microarray and RNA-seq are experimental techniques to measure the expression levels of the transcripts. These technologies have driven the functional genomics research in different organisms, prokaryotes and eukaryotes, for studies of physiological or pathological conditions. Microarrays allow a quantification of nucleic acids (mRNA form of genomic DNA or cDNA) by hybridization to complementary array. However, the RNA-seq technique, by sequencing of RNAs, can apply the same protocol for various interests, such as total RNA, mRNA, small RNA, such as miRNA, tRNA and rRNA.

There are a number of advantages and disadvantages about these two techniques. Microarray techniques are limited to detect transcripts, whose genomic sequences are their known; however, the microarrays are excellent for specific identification of knownvariants, for example in diagnostic testing, such as: cystic fibrosis. Nevertheless, RNA-seq is used for detection of known and unknown transcripts and detectssplicing and variants, especially you used for further depth studies in any organism, prokaryotes and eukaryotes.

Real Time PCR

Real time PCR is a golden standard technique of gene expression that detects the product as soon it amplifies. The sample (RNA of study) is first reverse transcribed and then a mix similar with the conversional PCR plus a fluorescent agent is added in the sample for amplification. Real time PCR are applied in a variety of biological fields such as microbiology, veterinary science, pharmacology, biotechnology among others and can be useful in detecting SNP variation, allelic discrimination, forensic studies, quantification of pathogens and genetically modified organisms (GMOs), food quality, besides gene expression and ncRNA analysis.

This methodology has some advantages over the conversional PCR considering specially the measure of the DNA concentration, and another relevant fact is that the results obtained can be either qualitative or quantitative (qPCR), different from de conventional. Real time PCR is widely used technique and so became necessary an experimental protocol with detail information, being a simple and efficient instrument of research is require a maximal quality control so the results can be more reliable and possible to compare. Real-time PCR is very good for the few genes already known, however when do you want to see all a comprehensive analysis already exist other techniques are more suitable and effective.

Applied Biotechnology: Looking in to the Future

In the past, the focuses of genetics studies were or in the genome or in proteins, but the discovery of new classes of RNA and what they can do opened up a whole world of different approaches to old problems. Besides the knowledge of the biology of the specie at a given moment and situation, obtained especially by the mRNA, now the RNA is also being used in medical treatment, agriculture improvement, and in others fields involving biotechnology.

The development of global transcriptome profile using next generation sequencing as a functional genomics tool permitted a better selection of possible target for further studies. With that in mind plus the fact that constant improvement and discoveries are being made in the field of science, the techniques applied now that focused on RNA made incredible breakthrough and are a promise in a varied of areas.

Ageing

The use of non-coding RNA like miRNA was first tested in the improvement of lifespan by in *Caenorhabditis elegans* and from that different authors pursued this topic. miRNAs modulates protein homeostasis, mitochondrial metabolism, stress response, including DNA damage [7]. Using new sequencing technologies among other, researchers discovered that many *C. elegans* miRNA including *lin-4* changed their expression during life and in the absence of *lin-4* the animals decreased their lifespan [8]. This and other breakthroughs demonstrated that miRNAs as capable of predicting the longevity and remaining lifespan, and that possibly others ncRNAs are also involved in ageing [7].

Gene function and vaccine production

Next generation sequencing became an extremely important tool for the understanding of gene function, since the costs of sequencing for the whole genome and transcriptome is more cheap and fast than previously methodologies. The study of the transcriptome allows us to understand what happens with the microorganism and the host during the infection and select potential virulent factors with more precision. This selection is possible thanks to studies of different expression in environment (e.g. culture medium) that simulates the conditions faced by the pathogen or by a double transcriptome study where both the microorganism and the host are analyzed *in vitro* or *in vivo*.

Insect management

Insects pests are a major problem in agriculture and the use of pesticides are slowly decreasing since it is toxic for humans, fact that brought the necessity for nontoxic approaches. RNA interference (RNAi), a process of specific gene silence based on dsRNA, was susceptibly tested in deleting essential genes in insects and by doing so caused lethality [9]. Both miRNA and siRNA are pathways endogenous of insects and plants, and the siRNA are recognizing as a major antivirus defense mechanism of insects. Even so this strategy proved to have potential not only in killing pests but also in the control of diseases transmitted by insects [10]. The problem encounter for this method was the way of delivery, since dsRNA apply do not replicate in the host and because of that need to be constantly supplied. Two popular approaches are the use of RNAi pesticides that are safe for humans and beneficial insects, and use of transgenic plants that already produce dsRNAs [10] but in this case brings further problems that will not be discussed here.

Disease therapy

The RNAi therapy shown to be effective in treating a variety of disease from viral disease to single nucleotide polymorphisms (SNP) and oncogenes, but as mentioned in the topic above the delivery and possible toxicity of the RNAi are still a concern [11].

Human immunodeficiency virus (HIV) was one of the first target by RNAi in the attempt to silence an infectiousness targets, but since it is a highly mutate virus there is still a lot to figure it out until this approach be successful. A second method were then tested to improve the effectiveness of the treatment which instead of focusing on the virus the target were a host receptor require for the infection [12].

The use of RNAi to treat disease are in progress and apart from HIV others illness are also in experimental testing and among then are hepatitis C virus (HCV), amyotrophic lateral sclerosis (ALS), Huntington, different types of cancer and more [11,13].

Approaches that manipulates splicing also function as disease therapy by correcting aberrant splicing or inducing then. Mutation that disrupt or create new splice sites are very common, in both intronic and exonic location, and are known to cause genetic diseases (e.g. Duchenne muscular dystrophy, Miyoshi myopathy, Cancer) [13].

PROTEOMICS

The development of new technologies, involved in next-generation sequencing technology and access the complete genomes has promoted an increase of information about genomic data for several organisms. Thus, the functional genomics come with aim of complementing genomic data, where through of transcriptomic and proteomic studies, has promoted the description of the gene and protein functions as well their interactions. However, unlike of genomic and transcriptomic studies, the proteomic try to evaluate the protein expression at a given condition. Thus, a proteomic study provides valuable information about: protein synthesis, qualitative and

quantitative information of protein product, post-translational modifications, protein-protein interaction and subcellular localization.

Currently, due the great advances both in classical and high-throughput proteomic technologies, new advances has been observed in the identification of targets for infectious diseases, inflammatory diseases, chronic conditions, cancer, drug discovery. Furthermore, through of comparative proteomic screening, between two conditions, i.e normal and diseased cells, allow the identification of differentially expressed proteins, that can promoted the identification of novel targets and biomarkers.

The proteomic allows the analysis of the gene expression on a large scale; this study is basically divided into two steps: (i) separation of the protein extract through of the intrinsic fundamental properties of each protein and (ii) identification of proteins through of the sequencing of peptides predigested enzymatically. However, the success of a proteomic study is due the complimentarily by bioinformatics analysis; where several softwares are developed to auxiliary in the data analysis, i.e., correlation with genomic data, furthermore, several public data repository for *proteomics* data were created. Thus, proteomics is presented as a multidisciplinary discipline in the investigation medical, biologic and biochemistry.

Gel-Based Proteomic

Gel-based methods like gel electrophoresis are routinely used in proteomic analysis to separation of molecules or peptides. Two-dimensional gel electrophoresis (2-DE) is a classical proteome technique, applied to resolution of protein from complex protein mixture. 2-DE is used to separate proteins in two dimensions: (i) during the process of isoelectric focalization (IEF) the proteins are separated according to the isoelectric point (pI) and (ii) using a SDS-PAGE the proteins are separate by molecular weight. After the electrophoretic resolution the spot/ protein are submitted to tryptic digestion and identified by mass spectrometry (MS). Thus, the combination of the 2-DE with the technique MALDI-TOF-MS (matrix-assisted laser desorption/ ionization – time of flight), has promoted great insight in the study proteomics.

The 2-DE analysis allows the generation of proteins maps, where spots/proteins differentially expressed are visualized, as well the presence of post-translational modifications (PTM) and isoform of proteins. However, a disadvantage of 2-DE classical is gel-to-gel variation, thus to suppress this variation the two-dimensional fluorescence difference gel electrophoresis (2-D DIGE) was developed. This technique allows the simultaneous separation of two samples different pre-labeled with cyanine dye (CyDye) fluorophores in the same 2-DE. Furthermore, the great advance of this technique is the realization of a multiplex assay, where an internal standard is used for normalization of data that minimize the experimental variation between gels and favor the quantitative analysis of the proteins [14].

Gel-Free Proteomic

Due the great advances in the biotechnological, new technologies also has been developed in proteomic area, where several advances have been observed in recent years, in order to address the weaknesses observed in 2-DE. Thus, the development of gel-free methods, such as liquid chromatography (LC) MS/MS revolutionized the proteomic field, where are obtained higher sensitivity, accuracy and resolution in proteomic study of large-scale. In this approach normally utilize a microcapillary reverse phase (RP) μLC system to separate the protein extract previously cleaved into peptides and *in tandem* analyzed by MS/MS, using electrospray ionization (ESI) coupled the mass analyzer (time-of-flight (TOF), quadrupole and ion trap). Thus, LC-MS/MS analysis associated the database searching has promoted a high-throughput proteomic analysis.

The abundance changes of proteins present in complex biological mixtures is a major challenge in the proteomic study, thus with the introduction of gel-free methods, several workflows to quantification have been developed such as: (1) Chemical labeling: utilize thiol-specific tags to covalently label in cysteine residues through of the isotope-coded affinity tag (ICAT) reagent approach, in addition other approach used to quantification using the chemical labeling is the technique iTRAQ, where is based in the labeling of an amine reactive, isobaric, isotope tag. (2) Metabolic labeling: metabolic precursors (amino acids, carbohydrates, organic salts) are introduces in the culture media and changed the molecular weight of the proteins, these metabolic precursors are labeling, through of the technique of stable isotope labeling of amino acids in cell culture (SILAC). (3) Label-free quantification: no label type is used to quantify the samples. The quantification is based on (i) spectral counting; where the quantification is achieved by total number of MS/MS spectra for a protein, and (ii) intensity of the chromatographic peak area under the curve or intensity of the precursor ion MS spectra.

Proteomic in Applied Microbiology and Biotechnology

The proteomic study offer a global survey of cellular changes in different physiological states at protein level, thus several studies using distinct or combined proteomic approach has been applied both in the applied microbiology and biotechnology.

In the microbiology field several proteomic studies have been performed with lactic bacteria, due the importance of this group of bacteria in the aliment industry, act as components of probiotic products and mainly in the fermentation process of yoghurt and cheese as well in the long-term storage of these products. Thus, proteomic studies related to knowledge of the process adaptive of *Lactobacillus*, *Propionibacterium* and *Bifidobacterium*, to acid condition and bile salts, showed a set of proteins that favor the survival and adapt theses bacteria in theses environmental that are present in the gastro-intestinal tract [15-17]. In addition, the combination of bioinformatics tools and proteomic techniques has been applied to explore the host–bacteria interaction process, through of the characterization of surface-exposed proteins main protein fraction used by bacteria

to promote the interaction with receptors of the host-cell [18,19]. However, these studies favor the selection of strain bacterial that can be used in the alimentary industry.

The genomic data from cancer patients has promoted information about variations, mutations and molecular pathways this disease. However, these alterations reflect in the gene expression of the cell cancer and consequently affected the function and stability of the proteins coding by these genes. Thus, proteomic study allow unveil this alteration and identify proteins involved in this pathway. In addition, currently the proteomic has been applied in the research of potential biomarker to have utilized in diagnostic of various type of cancer. In this type of study the biological samples evaluated from blood specimens like serum or plasma, thus quantitative proteomic both label-based and label-free has promoted the identification of several regulated differentially proteins, when compared the normal cell and tumor cell [18-20].

Currently with the technology advances, great efforts have been directed to the study of sustainable energy using mainly the bioenergetics process, which aims at energy production from biological materials and photosynthetic organism. A of the applied bioenergetics process is at production of biofuels. Thus, several proteomic studies are used to characterize the main raw material used to produce biofuels. Proteomic study performed with the plants: *Sorghum* sugarcane and maize, and algae and cyanobacteria that also are used in biofuels fabrication, have promoted the characterization of proteins present in different subcellular compartments, involved in several biological process as well the identification of proteins associated the resistance and adaptation to several stress conditions. These studies combined different proteomic approach and the results obtained in these studies theses several source used in the bio-ethanol production, promote information about of the physiology of this organism as well the identification of targets that can be used in its improvement main during the cultivation in different climatic conditions, favoring the biofuels production [21].

Application to a Bacterial Protein-Protein Interaction

With the advent of second and third generation of sequencing technologies of there has a significant reduction of cost and time, in addition to increasing simultaneously the amount of readings in sequenced genomes. This made more accessible the genomes sequencing, therefore increasing the amount of biological data how DNA, RNA and proteins available in public databases about various organism [22-24].

These molecules of DNA, RNA and proteins do not act individually in a living organism they work together interacting with each other in order to carry out its regulatory or physical biological processes. By knowing how these elements are grouped and interact is a relevant area of study and research, enabling understands an organism, a disease or a biological process at systemic level. Thus, data generated from DNA, RNA, proteins and other molecules, enabled researchers to broaden the fields of research to better understand organisms and/or diseases and more, search for new pharmaceuticals, veterinary, agricultural and biotechnology compounds.

An important tool for understanding how these molecules act in complex biological systems are protein-protein interactions (PPI) that, pair-to-pair, form a complex set of biological interactions called interactome [25, 26]. A PPI is composed of nodes that represent proteins (A and B), and edges that connect these proteins (nodes). The edge that connects both proteins do not have source and destination, thus forming a non-directional interaction, being equivalent the interpretation that the protein A interacts with protein B or that protein B interacts with the protein A (Figure 6).

Figure 6: Representation of the interaction between two proteins using the graphs theory where the nodes A and B are connected by an edge.

Each interaction has two elements, being called binary. Several binary interactions form a complex network of PPI and may represent all interactions of an organism or between a pathogen and its host (Figure6).In the literature are described methods for the identification of PPI, being classified as: (i) experimental- yeast-two-hybrid (Y2H), protein chip, tandem affinity purification followed by mass spectrometry (TAP-MS); - biophysical- atomic force microscopy (AFM), analytical ultracentrifugation (UC), nuclear magnetic resonance (NMR)or (ii) computational - three-dimensional structure, text mining, phylogenetic profile, phylogenetic tree, gene co-localization, interolog mapping-basedand machine learning.

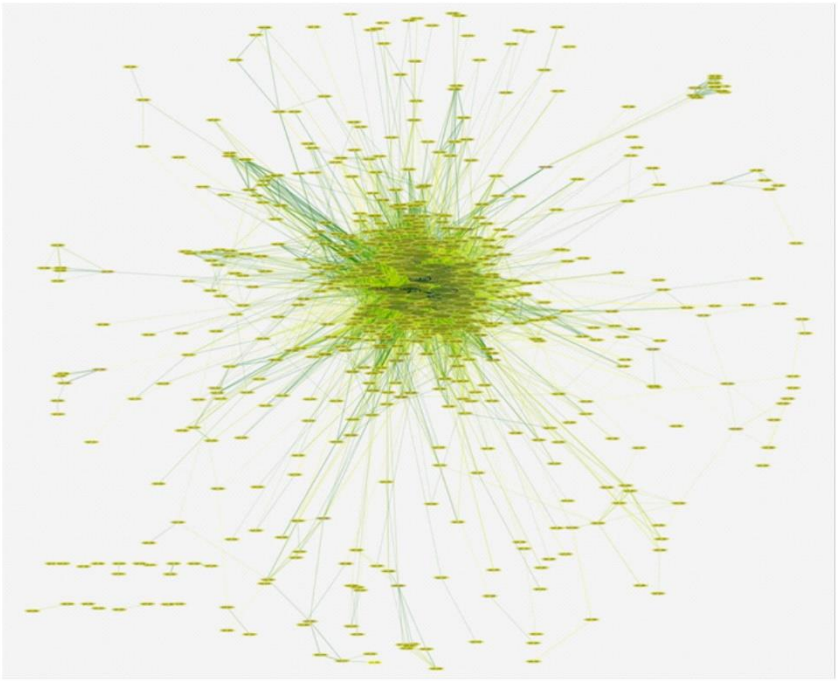


Figure 7: Example of a protein-protein interaction network, made up of several pairs of interaction.

In time, the methods can be classified also as for the volume of data generated aslow-throughput or high-throughput. High-throughput methods are able to detect a greater amount of interactions, but have low specificity. While the low-throughput methods, which test specific interactions, have high specificity. As more specific is the method, most sure has that the interaction occurs within the organism and, the more sensitive is the method, the more interactions are known. Experimental and computational methods complement each other, both being necessary for the scientific community to better explore the data [27,28].

Regardless of the method used, a network interaction can provide researchers with more knowledge about an organism, making possible to think about new biological hypotheses and direct new experiments. In this sense, though indirectly, network interaction can be used for biotechnological purposes. By better knowing the organism, this can be genetically modified and have his behavior changed, aiming at greater efficiency or quality in the production of a product of our interest, as well as, can be genetically modified to produce or metabolize a product that is harmful to human, animal health or the environment, such as bio-remediation of soils. Aided by a network of PPI this is possible.

Assuming that some gene, óperon or protein involved in a biotechnological process of interest be known, with the help of a PPI network, it is possible to identify the proteins that interact with (partner) the protein of interest. By identifying proteins an interaction, may also be known their functions allowing to better understand the biological process or even the biochemical pathway involved in biotechnological process of interest. This information, aggregated to the knowledge about the organism, can help a researcher to understand the organism at systemic level, allowing new hypotheses to be tested in the laboratory. Still, with a network of interaction topological analysis can performed to identify biologically important proteins as proteins Hubs that have a high degree of interaction with other proteins, and proteins Betweenness that make the connection between the functional modules in a network of interaction. These both protein class, Hub and Betweenness, are considered essential proteins [28].

Information of interactions, protein Hubs and between allies proteomic experiments or information to RNA-Seq, which test a condition of biotechnological interest and measure the expression of transcripts or proteins, extend the prospects of identifying targets or metabolic pathways relevant. Besides allowing know which existing proteins and the expression of these, it is possible to meet with who these proteins interact and work together for the realization of biotechnological condition tested.

References

1. National human genome research institute. 2012.
2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977; 74: 5463-5467.
3. National human genome research institute. 2014.

4. BGI.

5. Grant JR, Stothard P. The CGView Server: a comparative genomics tool for circular genomes. Nucleic Acids Res. 2008; 36: W181-184.

6. Soares SC, Abreu VA, Ramos RT, Cerdeira L, Silva A. PIPS: pathogenicity island prediction software. PLoS One. 2012; 7: e30848.

7. Kato M, Slack FJ. Ageing and the small, non-coding RNA world. Ageing Res Rev. 2013; 12: 429-435.

8. Kato M, Chen X, Inukai S, Zhao H, Slack FJ. Age-Associated Changes in Expression of Small, Noncoding RNAs, Including microRNAs, in C. Elegans. Rna. 2011; 17: 1804–1820.

9. Baum James A, Thierry Bogaert, William Clinton, Gregory R Heck, Pascale Feldmann, et al. Control of Coleopteran Insect Pests through RNA Interference. Nature Biotechnology. 2007; 25: 1322–1326.

10. Burand John P, Wayne B Hunter. RNAi: Future in Insect Management. J Invertebr Pathol. 2013; 112: S68–74.

11. Hannon GJ, Rossi JJ. Unlocking the potential of the human genome with RNA interference. Nature. 2004; 431: 371-378.

12. Boden D, Pusch O, Lee F, Tucker L, Ramratnam B. Human immunodeficiency virus type 1 escape from RNA interference. J Virol. 2003; 77: 11531-11535.

13. Havens MA, Duelli DM, Hastings ML. Targeting RNA splicing for disease therapy. Wiley Interdiscip Rev RNA. 2013; 4: 247-266.

14. Arentz G, Weiland F, Oehler MK, Hoffmann P. State of the art of 2D DIGE. Proteomics Clin Appl. 2014.

15. Hamon E, Horvatovich P, Bisch M, Bringel F, Marchioni E, et al. Investigation of Biomarkers of Bile Tolerance in Lactobacillus Casei Using Comparative Proteomics. J Proteome Res. 2012; 11: 109–118.

16. Leverrier P, Dimova D, Pichereau V, Auffray Y, Boyaval P, et al. Susceptibility and Adaptive Response to Bile Salts in Propionibacterium Freudenreichii: Physiological and Proteomic Analysis. Appl Environ Microbiol. 2003; 69: 3809–3818.

17. Borja Sánchez,Marie-Christine Champomier-Vergès,Birgitte Stuer-Lauridsen, Patricia Ruas-Madiedo,Patricia Anglade, et al. Adaptation and Response of Bifidobacterium Animalis Subsp. Lactis to Bile: A Proteomic and Physiological Approach. Applied and Environmental Microbiology. 2007; 73: 6757–6767.

18. Barinov A, Loux V, Hammani A, Nicolas P, Langella P, et al. Prediction of Surface Exposed Proteins in Streptococcus Pyogenes, with a Potential Application to Other Gram-Positive Bacteria. Proteomics. 2009; 9: 61–73.

19. Meyrand M, Guillot A, Goin M, Furlan S, Armalyte J, et al. Surface Proteome Analysis of a Natural Isolate of Lactococcus Lactis Reveals the Presence of Pili Able to Bind Human Intestinal Epithelial Cells. Mol Cell Proteomics. 2013; 12: 3935–3947.

20. KoÄevar N, Hudler P, Komel R. The progress of proteomic approaches in searching for cancer biomarkers. N Biotechnol. 2013; 30: 319-326.

21. Ndimba BK, Ndimba RJ, Johnson TS, Waditee-Sirisattha R, Baba M, et al. Biofuels as a Sustainable Energy Source: An Update of the Applications of Proteomics in Bioenergy Crops and Algae. J Proteomics. 2013; 93: 234–244.

22. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell. 2013; 155: 27-38.

23. Casals F, Idaghdour Y, Hussin J, Awadalla P. Next-Generation Sequencing Approaches for Genetic Mapping of Complex Diseases. J Neuroimmunol. 2012; 248: 10–22.

24. van Dijk EL, Auger H2, Jaszczyszyn Y3, Thermes C2. Ten years of next-generation sequencing technology. Trends Genet. 2014; 30: 418-426.

25. Ngounou Wetie AG, Sokolowska I, Woods AG, Roy U, Deinhardt K, et al. Protein–protein Interactions: Switch from Classical Methods to Proteomics and Bioinformatics-Based Approaches. Cell Mol Life Sci. 71: 205–228.

26. Rao VS, Srinivas K, Sujini GN2, Kumar GN1. Protein-protein interaction detection: methods and analysis. Int J Proteomics. 2014; 2014: 147648.

27. Harrington ED, Jensen LJ, Bork P. Predicting biological networks from genomic data. FEBS Lett. 2008; 582: 1251-1258.

28. Nicola J Mulder, Richard O Akinola , Gaston K Mazandu , Holifidy Rapanoel. Using Biological Networks to Improve Our Understanding of Infectious Diseases. Computational and Structural Biotechnology Journal. 2014; 11: 1–10.

1.2 - *In silico* protein-protein interactions: avoiding data and method biases over sensitivity and specificity

Edson Luiz Folador, Alberto Fernandes de Oliveira Junior, Sandeep Tiwari, Syed Babar Jamal, Rafaela Salgado Ferreira, Debmalya Barh, Preetam Ghosh, Artur Silva, Vasco Azevedo

O estudo de redes de interação proteína-proteína permite se ter uma visão sistêmica dos mecanismos celulares de um organismo, possibilitando conhecer o organismo a nível molecular. Considerando os diversos métodos existentes para a identificação dos pares de interação, experimentais e computacionais, aqui nos concentramos em descrever os métodos computacionais. Desconsiderando detalhes da implementação de cada método, destacamos principalmente a natureza do dado biológico usados para a predição e como estes dados causam viés sobre a sensibilidade e especificidade destes métodos, visando levar o leitor a refletir sobre os pontos positivos e negativos de cada método. Secundariamente nos preocupamos em relatar em quais organismos os métodos foram usados, citando ainda onde pode ser encontrada informações mais detalhadas sobre o funcionamento de cada método. Adicionalmente, conforme os dados usados como entrada para a predição, cada método foi classificado como primário ou não primário. Foi considerado primário o método capaz de identificar interações proteína-proteína ainda não identificadas em algum organismo e, método não primário, aquele que depende da existência de interações entre duas proteínas para que outras interações sejam preditas.

O artigo referente a esta seção foi publicado em 2015 pela revista Current Protein & Peptide Science com DOI número 10.2174/1389203716666150505235437.

Send Orders for Reprints to reprints@benthamscience.ae
Current Protein and Peptide Science, 2015, *16*, 000-000 1

In Silico Protein-Protein Interactions: Avoiding Data and Method Biases Over Sensitivity and Specificity

Edson Luiz Folador[1], Alberto Fernandes de Oliveira Junior[1], Sandeep Tiwari[1], Syed Babar Jamal[1], Rafaela Salgado Ferreira[2], Debmalya Barh[3,6], Preetam Ghosh[4], Artur Silva[5], and Vasco Azevedo[1,*]

[1]*Department of General Biology, Instituto de Ciências Biológicas (ICB), Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil;* [2]*Department of Biochemistry and Immunology, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil;* [3]*Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal, India;* [4]*Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA;* [5]*Instituto de Ciências Biológicas, Universidade Federal do Para, Belém, PA, Brazil;* [6]*InterpretOmics India Pvt. Ltd., #329 7th Main, HAL 2nd stage, Indiranagar, Bengalore, Karnataka, India*

Vasco Azevedo

Abstract: The study of protein-protein interactions (PPIs) can help researchers raise new hypotheses about an organism or disease and guide new experiments. Various methods for the identification and analysis of PPIs have been discussed in the literature. These methods are generally categorized as experimental or computational - each having its own advantages and disadvantages. Experimental methods provide insights into the real state of biological interactions but tend to be time-consuming and costly. Computational methods, on the other hand, can study thousands of PPIs at a very low cost and in much less time; however, the accuracy of such *in silico* prediction results heavily depends on the specific computational approach used. Furthermore, there is no gold standard for these computational methods; a method that works well for predicting one PPI may perform poorly (by generating false positives and false negatives) for a different PPI. Therefore, all such predictions must be carefully validated, preferably with experimental data. In this paper, we review the existing computational approaches and emphasize the use of biological data as inputs for accurate predictions of PPIs. We also discuss how such input datasets and approaches may influence the sensitivity and specificity of the predicted PPI networks.

Keywords: Computational approaches, *in silico* prediction, protein-protein interaction, protein network, system biology.

1. INTRODUCTION

Most of the data related to protein interaction networks have come from high-throughput experiments that provide information about interacting protein pairs [1]. However, such high-throughput experimental methods have their own limitations because they often produce a high number of false positives and false negatives, while also being time-consuming and labor-intensive [2]. The advent of second and third generation sequencing has allowed genomes to be sequenced with significant reductions in cost and time, which has resulted in an ever-increasing number of sequenced genomes. Consequently, the amounts of biological data related to DNA, RNA, transcripts, proteins, and other molecules have also increased, opening up new opportunities to explore these data and better understand organisms and diseases, thereby aiding the search for new pharmaceuticals or products for use in veterinary, agriculture, or biotechnology industries [3-9].

*Address correspondence to this author at the Department of General Biology, Instituto de Ciências Biológicas (ICB), Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil; Tel: +55 31 3409 2610;
Fax: +55 31 3409-2614; E-mail: vasco@icb.ufmg.br

1389-2037/15 $58.00+.00

DNA, RNA, and proteins all work together to perform the biological functions of living cells. To understand how these molecules interact and are grouped, is an important area of study that can help develop a better understanding of an organism, a disease, or a biological process at a system level. An important tool for understanding complex biological systems is the study of protein-protein interactions (PPIs), which form complex networks of biological interactions called interactomes [10-14]. In the literature, several experimental and computational methods for the identification of PPIs have been reported. Important experimental methods include yeast-two-hybrid [15-17], protein chip [18, 19], tandem affinity purification followed by mass spectrometry [20-22], and biophysics-based methods, such as, atomic force microscopy [23-27], analytical ultracentrifugation [28-30], and nuclear magnetic resonance (NMR) [31-33]. We focus on only the computational methods in this paper and provide a detailed review of such computational methods used to predict PPIs [34] specifically.

Based on the volume of data that they generate, computational methods can also be classified as low-throughput or high-throughput methods. When using computational tools to predict PPIs, an important aspect is to validate the results

with experimental data that have preferably been manually curated. In the validations, statistical measures such as sensitivity, specificity, accuracy and area under the curve (AUC), represented as Receiver Operating Characteristics (ROC) curve, have been used to associate a degree of confidence to a predicted interaction [12, 35-41]. Today, the large amount of information that is available in public PPI databases makes the validation task easier and more effective than it was some years ago. Currently, tens of thousands of positive interactions [42-46] and curated PPI data [47] are available in public databases, while datasets of non-interacting proteins have also been published [48, 49].

It is important to remember that no matter how strong the PPI validation methodology or high degree of confidence associated with the interactions, the veracity of any prediction can only be absolute when it is demonstrated experimentally in an organism. However, this fact does not reduce the importance of *in silico* predictions that can quickly identify thousands of interactions with a certain degree of confidence. Such predictions can lead to new biological hypotheses about an organism and can help to guide *in vitro* experiments to obtain the best results, thereby reducing both time and cost [50].

In the last two years, several reviews on PPI *in silico* prediction schemes were published wherein the computational methods were thoroughly discussed [10-14, 51, 52]. Our focus in this paper is to review only those computational methods that use biological theories to support the method and different types of biological data as inputs for predicting the interacting pairs. We will discuss the advantages and disadvantages of each such computational method and examine their sensitivity and specificity that are negatively correlated [53].

2. COMPUTATIONAL METHODS USED FOR PROTEIN-PROTEIN INTERACTION PREDICTION

Computational methods for predicting PPIs can be classified in various ways based on the techniques or the data sources used. However, these classifications are not mutually exclusive. In PPI predictions, the specific algorithm leverages biological data to achieve better prediction results [54-58]; for example, the biological data can be three-dimensional (3D) protein structures while the algorithm can use machine learning-based approaches. In the following, we review the popular computational methods that can be classified under five main categories: (i) docking-based methods, (ii) text-mining based methods, (iii) amino acid sequence similarity based methods, (iv) protein domain based methods, and (v) machine learning based methods.

2.1. Docking-Based Methods

Understanding the physicochemical characteristics of the proteins that interact in a network, and also the details of how an interaction is established at the molecular level, can provide details on the working of cells and organisms as a whole [59-62]. *In silico* methods can then be used to investigate such interactions and predict how these molecules might interact at the atomic level [63, 64]. In this way, important regions in a protein's structure can be identified and targeted, e.g., for drug design [65-68].

3D structures of proteins that show the arrangement of atoms in space can help researchers understand the connections between biomolecules at the atomic level. Most of the PPI prediction methods use as template a file, obtained from X-ray crystallography or NMR spectroscopy, which contains the representation of macromolecular protein structures. A pioneering work that solved the 3D structures of many proteins originated in the 20th century [69-72] and used X-ray crystallography - a method that is still widely used today. On the other hand, NMR is a newer experimental technique that is also capable of producing 3D protein structures using the different magnetic resonance frequencies of the atomic nuclei in the bio-molecule [73]. However, these experimental methods are often difficult to implement because of the long time required to determine a 3D structure at high resolution, especially for membrane proteins that present molecular weight and size limitations [68, 74, 75]. Other template-based approaches have also been used to obtain 3D protein structures. These approaches were described as "comparative modeling", "homology modeling", "template-free", or *ab initio*, and have been used to provide reasonable models for protein structures using computational tools [76, 77]. Such methods depend on the availability of experimentally determined 3D models from which an unknown structure is predicted using an algorithm that parses the 3D domains of the known structure. Today, most 3D models of structures are obtained by comparative modeling methods that follow the protocol described by Sali (1997) [76]. This protocol lists the following four fundamental steps for the creation of a model: (i) identification of homologous structures that can be used as templates for modeling; (ii) alignment of the target sequence with the template sequences; (iii) construction of the model for the target based on the alignment; and (iv) model validation.

When the 3D structures of two proteins are known, it is possible to predict the corresponding protein-protein [78] or protein-ligand interactions [79, 80]. From analyses of several protein-protein complexes at atomic resolution, it was determined that the interactions occurred at surface regions (Fig. **1**) that were highly complementary to each other [81, 82]. Molecular docking methods have been developed to help understand the degree of interatomic interaction (chemical bonds) between two or more molecules, for example, between two protein molecules. By using a scoring system, such predicted interactions can also indicate the strength of association or binding affinity between the two molecules. A wide range of tools is currently available to model molecular docking between interacting proteins [14, 83-89]. Most of these methods are based on the assumption of "hard-body" interactions, meaning that the two molecules interact as solid, rigid bodies. Geometric surface models and data structures have been developed to predict connection modes of reasonable complexity whereas heuristic functions are used to classify candidates of greater complexity [90]. However, even though such methods based on "rigid-body" structures can precisely match the surfaces of 3D structures, they have an inherent limitation as the molecular flexibility of proteins, which is important in numerous biological processes, cannot be taken into consideration [91]. Molecular dynamics based approaches, which take into account the atomic motion in biomolecules, have been proposed to

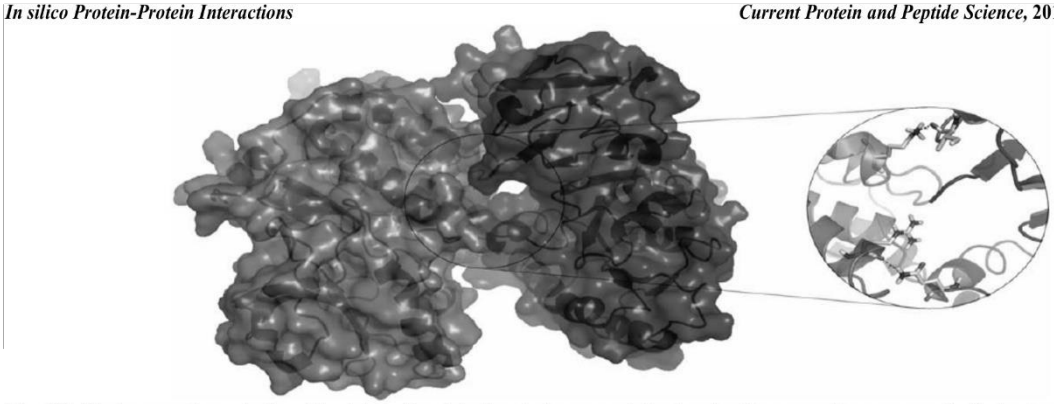


Fig. (1). Circle: complementarity at the interaction interface between proteins involved in promoting cancers. Left structure: RAF1 (PDB ID: 1C1Y_A). Right structure: MAP2K2 (PDB ID: 1S9I_A) [97].

overcome this limitation [92]. This methodology calculates the atomic positions of different constituents at consecutive time intervals in the order of femtoseconds (1fs = 10^{-5} seconds). A set of coordinates for the nucleus of each atom in the system and its temporal evolution describe the system trajectory [92-95]; hence this methodology incurs high computational costs to model an interactome having large number of interactions. Recently, the field of interactome modeling has emerged with the goal of understanding the complexity of molecular interaction networks between the proteins in a cell using 3D structures. Although interactome modeling refers to the numerous molecular contacts among intracellular proteins, in some cases, it is also possible to understand the interactions between genes. A great challenge however, is to use molecular dynamics to model a network of interactions between proteins on a large scale that would incorporate their 3D structures at the atomic level [96].

Such 3D structure based approaches can be considered primary as they can predict new interactions that have not yet been reported in the literature. However, these approaches are highly specific and depend only on the existence of 3D structures, which is a major disadvantage because of the small number of 3D structures that are currently available and also the high computational cost required to generate them *in silico* (Table **1**). Nevertheless, approaches using 3D structure data were used to predict PPIs in human cancer cells [97], human immunodeficiency virus (HIV) [98], a human-virus PPI network [99], metabolic processes and binding sites [96], and binding patterns of hub protein [100].

2.2. Text Mining-Based Methods

Text mining techniques have been used in many areas, from science to business. Text mining uses computational tools, information retrieval techniques, statistics, natural language processing, and machine learning to render, extract, organize, and view information stored in a natural (human) language. In other words, these methods use algorithms to understand and retrieve contextualized information stored in a natural format; for example, the information published in abstracts or in scientific articles [101-105]. Thus, text mining is an important tool for the extraction of knowledge in systems biology [106-108]. Grammatical rules are used to seek and recognize the co-occurrence of predefined entities and the relationships between these entities in a literary database (Fig. **2**). Here, an entity can be any object (noun) of interest, for instance, diseases, chemicals, pharmaceuticals, ligands, enzymes, inhibitors, etc. In biology, entities can be the names of genes or proteins. Relations between entities can be identified, but not limited to, action verbs and their conjugations like bind, interact, regulate, and inhibit. In the case of PPIs, an interaction can also be determined if the grammatical rules include an active or passive voice verb, such as "x regulates y" or "x is inhibited by y". The more standardized and accurately described these entities are in the literature, the more efficient the text mining technique will be. For more details please see the references [12, 45, 105, 106, 108-111].

Compared with other methods used to predict PPIs, the prediction made by the text mining approach is secondary, in the sense that a PPI can be described as soon as it is published, but interactions that have not been described in the literature cannot be retrieved. However, an advantage is that text mining does not depend on data for training or on the existence of a PPI in a public database. Thus, in the literature, it is possible to identify new interactions between entities in one organism that may also be present in a different organism of interest. Because the identification of PPIs by text mining is based on the co-occurrence of entities (genes or proteins), it is difficult to predict an interaction if it is reported as hypothetical or when a locus_tag is used as the identifier of the reported interaction (Table **1**).

Text mining has been used to identify PPIs in *Escherichia coli* and *Brucella* spp. [112], and in *Arabidopsis thaliana* [113] and was used in the STRING database to gather PPI information from the literature [44].

2.3. Similarity of Amino Acid Sequence-Based Methods

Multiple sequence alignments have been used to identify biological properties such as amino acid conservation, conservation of proteins between different organisms, homology between interacting proteins, location of nearby genes in a genome, and gene fusion [12, 13, 26]. These biological properties can be used alone or in combination to predict possible PPIs. Each property or combination of properties requires a distinct methodology with different advantages and disadvantages. Here we describe the approaches used for the prediction of PPIs based on the similarity of amino acid sequence.

Table 1. Comparison between the biological data type and methods

Method	Primary?	Sensitivity / Specificity
Docking	Yes, depends on molecular tests to determine the interaction	There are few three-dimensional structures available when compared to other types of biological data used in other methods
Text mining	No, depends on the existence of interactions described in the literature	Usually the interactions described in scientific texts are derived from experimental approaches; predictions depend on co-occurrence of genes/proteins in corresponding model organisms
Phylogenetic profile and Phylogenetic tree	Yes, depends only on sequences of amino acids	Depends on the phylogenetic proximity of organisms used in the prediction. Organisms phylogenetically very close (same species) can result in low specificity while phylogenetically very distant organisms can result in low sensitivity
Co-localization or neighborhood	Yes, depends only on the location of genes in the genome	Influenced by the number of neighbor genes considered because genes that are not considered neighbors may also interact.
Interolog mapping	No, depends on the existence of interaction to be mapped	Depends on how true are the interactions in the primary database, and also on the origin of mapping. Also depends on the cut-off value used to determine whether the mapped interaction pairs are homologous
Domain	Yes, depends on the identification of domains involved in interaction	Depends on number of considered domains and if domains are in proper functional status
Machine learning	Yes, after learning the algorithm is able to identify new interactions	Depends on whether the algorithm was trained properly to distinguish between true and false interactions; the training dataset must include both true and false interactions (negatome)

Primary: Identifies whether the method is able to predict new interactions without relying on interactions already described. Sensitivity / Specificity: Characteristics of biological data that can bias the sensitivity or specificity.

2.4. Phylogenetic Profile-Based Approach

In this approach, pairs of proteins are used to identify pairs of homologous proteins in other organisms, thus creating a phylogenetic profile that describes the presence or absence of these protein pairs in various organisms (Fig. **3**). The biological premise that underpins this approach is that if two genes are related to implementation of a function, they probably were inherited together, a situation that the loss of either of the two genes could derail a given function of the organism [26]. When using a phylogenetic profile-based approach, it is especially important to check if the conserved sequences occur as a result of selective pressure over a long period of time and not simply because the corresponding organisms were phylogenetically close. Thus, the selection of phylogenetically close organisms, without considering enough time to accommodate selective pressure, may cause the identification of false-positive interactions.

By testing the conservation of pairs of interactions in a large set of organisms, this approach tends to be more specific and less sensitive especially while selecting phylogenetically distant organisms and considering that proteins must exist in all the organisms [46, 53]. For more details, please see the references [12, 13, 46, 52].

The predictions made by this approach can be considered primary as new interactions can be identified without prior knowledge of interaction between protein participants. The predictions are dependent only on the amino acids sequences to construct a phylogenetic profile and, the presence of both proteins in organisms infers that they interact (Table **1**). This approach has been used to predict PPIs in *Mus musculus* [53], *E. coli* and yeast [114], *A. thaliana* [56], and *Plasmodium falciparum* [115].

2.4.1. Phylogenetic Tree-Based Approach

In this approach, in addition to the presence or absence of protein pairs in various organisms, a pair of proteins has to possess the same evolutionary history. The premise that underlies this approach is based on the assumption that, if only one of the proteins in a pair of interacting proteins undergoes modification as a result of evolutionary pressure, then the interaction will not exist. On the other hand, if one of the proteins in a pair undergoes modification and the other protein changes in order to adapt, the interaction can be maintained [14] and the proteins can be said to have co-evolved. When proteins co-evolve, the topology of the phylogenetic trees will be similar and they interact; thus, this approach, also called the mirror tree approach, compares the differences between these trees [10]. Each branch of a phylogenetic tree is converted to a numeric value and stored as an array or vector, enabling calculations, analyses, and large-scale computational comparisons [50, 116] (Fig. **4**). Similarities between trees can also occur when a pair of proteins evolved and maintained their similarity simply by the positive selective pressures on organisms, even though the particular protein pairs do not interact. To eliminate this bias, normalization of such phylogenetic trees using the 16S rRNA has been proposed [117]. For more details, please see the references [13, 46, 52].

This approach is also considered as primary because new interactions can be identified without relying on them having been reported in literature or existing in a public database.

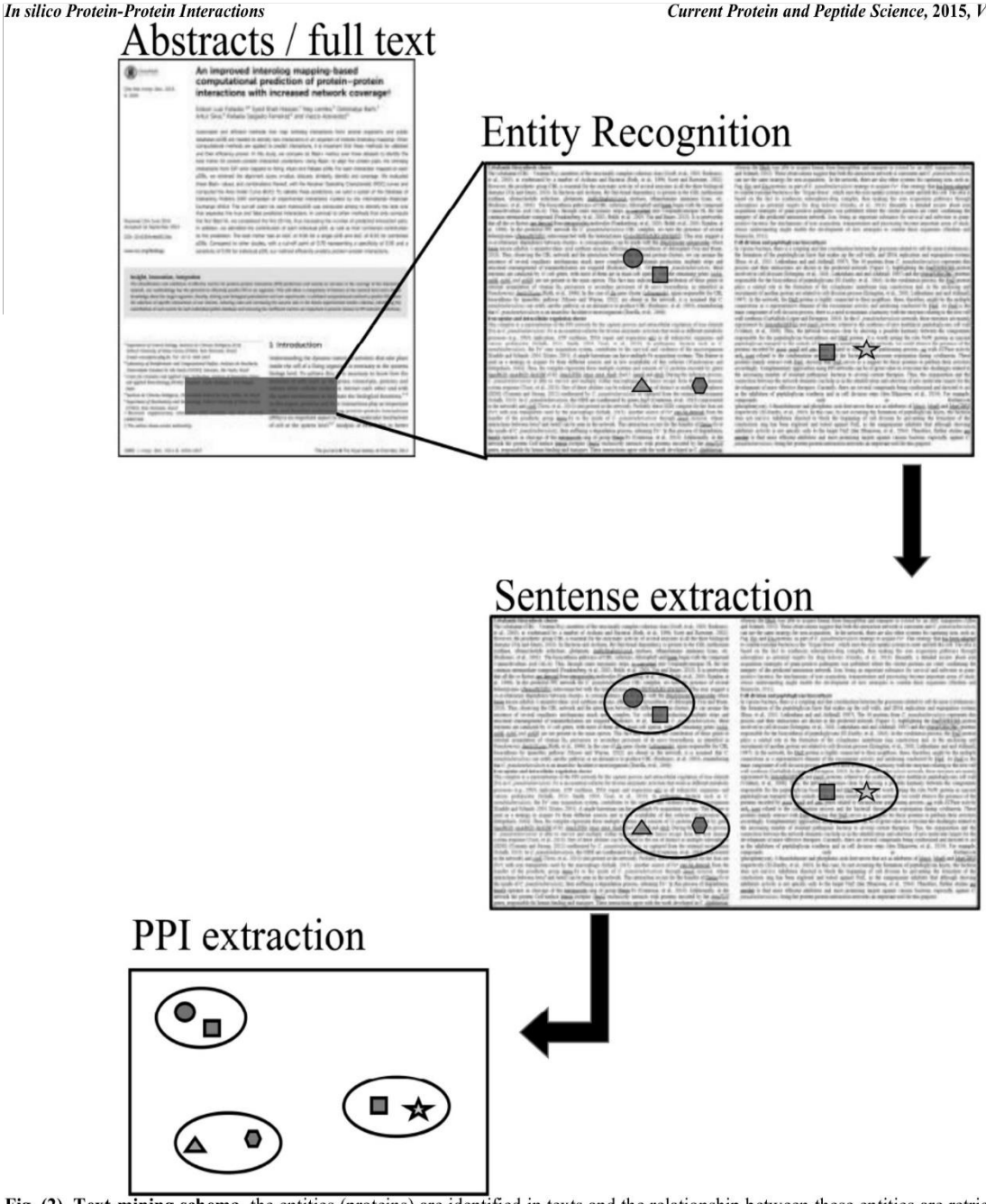


Fig. (2). Text-mining schema, the entities (proteins) are identified in texts and the relationship between these entities are retrieved from sentences allowing the extraction of the protein pair.

Depending on the number of organisms used to generate and compare the phylogenetic trees, this approach tends to be more specific and less sensitive (Table **1**). This approach has been used to predict the interactome of *E. coli* [116], human [118], and *Saccharomyces cerevisiae* [119].

2.4.2. Gene Colocalization-Based Approach

In this approach, the assumption is that when two genes are located in the same region of a genome and the region is conserved in several species, the products of these genes will probably interact [13]. This approach allows genes that may form operons to be identified, assuming that the proteins expressed by these genes interact in a cluster to perform their roles (Fig. **5**). In bacteria these neighbor genes, organized in an operon, are transcribed as a single unit while in eukaryotes, they are co-regulated usually possessing a closely related function and probably interacting [34]. The number of genome sequences used to identify the conservation of gene locations, as well as how phylogenetically close the genomes are, can generate bias; however it is more specific than the

approach in which phylogenetically distant genomes are compared [12]. For more details, please see the references [12, 13].

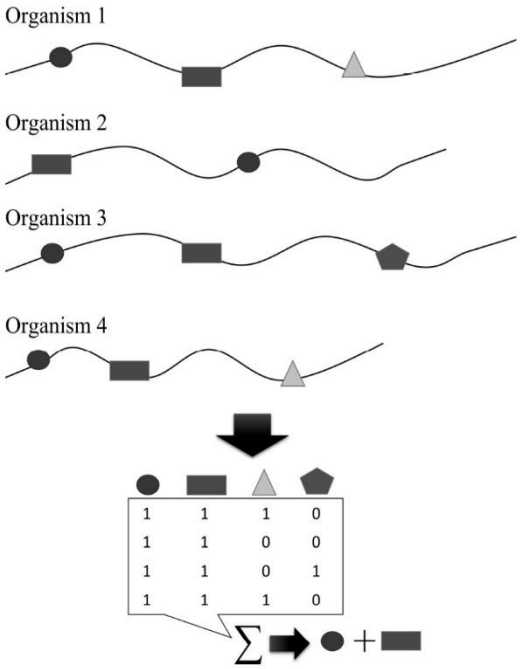


Fig. (3). Phylogenetic profile schema, representing the conservation of protein pairs in various organisms.

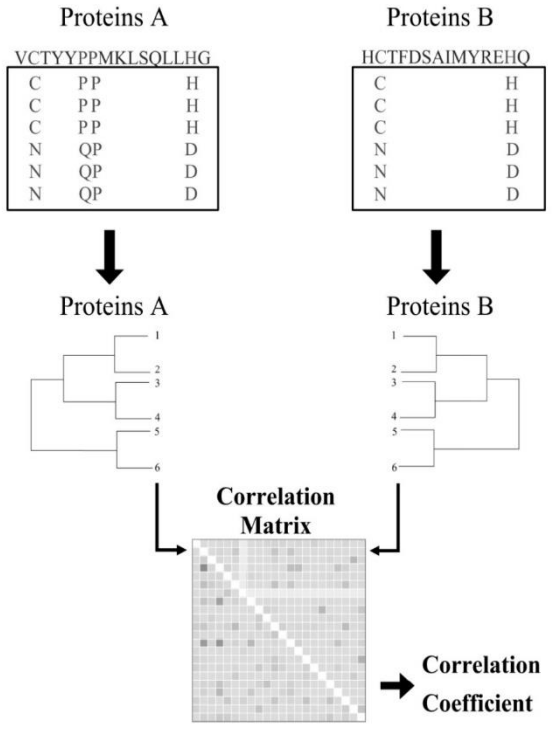


Fig. (4). Phylogenetic tree schema: represent the conservation of phylogenetic profile in various organisms besides selective pressure. When the topology of phylogenetic trees are close, their distance matrices are correlated.

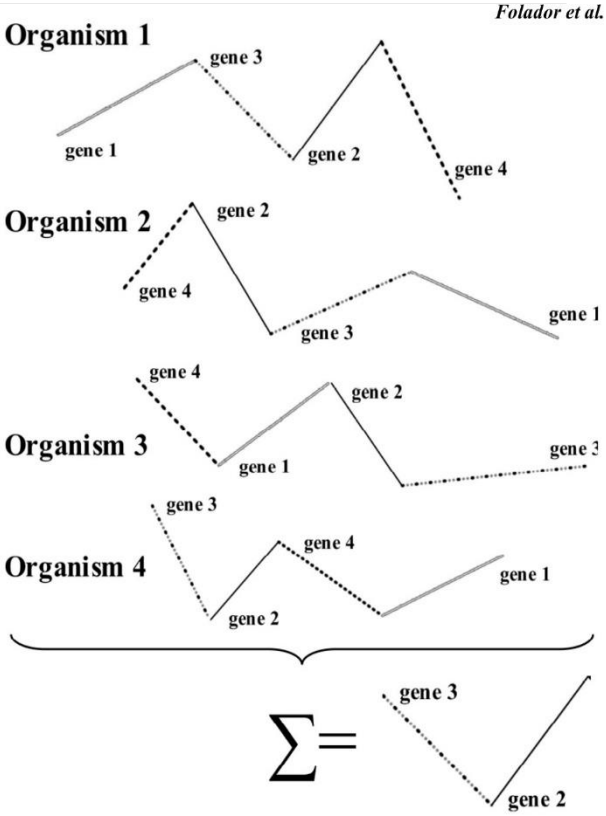


Fig. (5). Gene colocalization schema: represents the conservation of neighbor genes and their localization in various organisms.

The predictions made by this approach can also be considered as primary because new interactions can be identified without relying on them having been reported in the literature or existing in a public database. Compared with other methodologies, this approach is more specific and less sensitive, and has a tendency to generate false negatives because although neighboring genes are expected to actually interact, distant genes can also interact and these interactions would not be detected by this approach (Table **1**). This approach has been used to predict the interactome of *A. thaliana* [56].

2.4.3. Interolog Mapping-Based Approach

Interolog mapping can be used for PPI prediction by considering the mapping of orthologous interactions. This approach uses a premise similar to that used for gene annotation, where a gene function is inferred based on the function of homologous genes in other organisms. Similarly, this approach is based on the biological principle that, if two proteins (p1 and p2) from one organism are known to interact, then the homologous proteins in an organism for which we wish to predict the interactions (i1 and i2) will also interact [13]. That is, if p1 and p2 interact, and p1 is homologous to i1 and p2 is homologous to i2, then i1 and i2 will interact in the organism of interest (Fig. **6**). This premise is more realistic in prokaryotes however, in eukaryotes other factors such as the cellular location of the proteins and their possible tissue-specific expression have to be taken into account.

The basic premise used in this approach is the identification of pairs of PPI counterparts in any organism, preferably

phylogenetically close, requiring this primary source of interactions. Today, there is a large amount of public data on PPI and amino acid sequences available [42-44, 120], where the homology of these bases against the organism of interest can be identified with programs like Blast+ [121], and the homologous pairs of interactions can be mapped to the organism of interest (Fig. **6**). For more details, please see the references [14, 122].

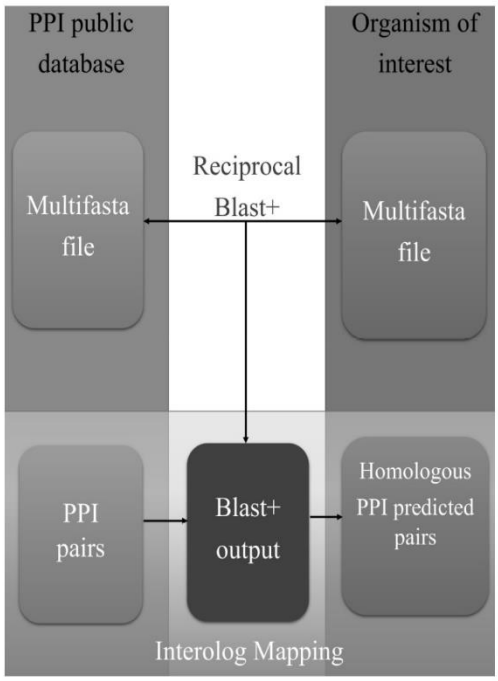


Fig. (6). Interolog mapping schema. Left box: amino acid sequence and interaction pairs from public databases. Right box: the amino acid sequence of the organism of interest. Center box: reciprocal alignment of public data sequences against sequences of the organism of interest. Down box: mapping of interactions from public database and transfer of homologous interactions to the organism of interest along with the alignment and coverage values of Blast+.

The major bottleneck of this approach is the volume of data that revolves around millions of sequences, hundreds of millions of interactions and hundreds of millions of results of alignments to be crossed, requiring expertise in Bioinformatics.

The predictions made by interolog mapping approach are secondary, because only after the PPI are known in any organism, can homologous interactions be predicted in the organism of interest. Currently, the large numbers of PPIs that are available in the various public databases [42-45, 56, 123] tend to make this approach extremely sensitive. Therefore, the specificity of the interactions should be tested carefully because, in general they have negative correlations [122]. The specificity of this approach depends on two factors: (i) whether both proteins are homologous and (ii) the reliability of the interactions in the public databases that are used to map the homologous interaction pairs (Table **1**). The use of different measures of similarity and different datasets to validate this methodology have generated conflicting results [14]. To identify the pairs of homologous proteins, combinations of metrics such as similarity, identity, and coverage seemed to better differentiate homologous protein pairs when compared with metrics such as e-value, score, or bit score; this might be because the latter metrics are sensitive to the length of the amino acid sequences and the size of the database [41, 122].

This approach is useful to transfer PPIs from model organisms to other less studied organisms. It has been used to predict PPIs in human cancer proteins [36], human [123, 124], rice [125], *A. thaliana* [56, 126], *M. musculus* [53], *Tetraselmis subcordiformis* [127], *S. cerevisiae* [128], *Caenorhabditis elegans* [129], *Danio rerio* [130], yeast [131], and *Leishmania braziliensis*, *Leishmania major*, and *Leishmania infantum* [41].

2.5. Protein Domain-Based Approach

Protein domains are functional and/or structural units in a protein and are conserved through evolution. Currently, the interest in domain-based protein interaction prediction is increasing, and several methods based on this approach have been proposed. Sprinzak and Margalit [132] were the first to use an association method for the identification of PPIs in *S. cerevisiae.* Different approaches such as the association method, domain association method, Bayesian networks method, domain pair exclusion method, and P-value method have been applied to predict PPIs [34].

The association method looks for characteristic sequences or structural motifs that are present in the proteins in an interacting pair and that are not present in non-interacting proteins [132]. Protein interaction data are used to compute log-odds scores and to find correlated domains. The log-odds score is computed as:

$$\log_2 (Pij / PiPj),$$

where, Pij is the observed frequency of domains i and j that occur in one protein pair; Pi and Pj are the background frequencies of domains i and j in the interaction data. Predicted domain interactions are defined as interactions that have positive log-odds scores and several occurrences of the given domain pair in the database.

In the association method, other domains in a given pair of interacting proteins are ignored because the method is based only on the correlated sequence signatures and each pair of interacting domains is considered separately (Fig. **7**). In Bayesian network methods, this type of data are also measured [133]. To estimate the parameters of Bayesian models, the maximum likelihood estimation method can be used. This method maximizes the probability of interactions of all putative domain pairs and incorporates the experimental errors of protein interaction data into the scoring scheme. The likelihood function is expressed by the following parameters:

$$\Theta (\Theta ij, fp, fn),$$

where Θij is the probability that domains i and j interact, fp is the false positive rate, and fn is the false negative rate derived from experimental data. Because of the large number of interacting domains, it can be difficult to maximize the likelihood function directly. This problem was resolved using the expectation maximization algorithm [34].

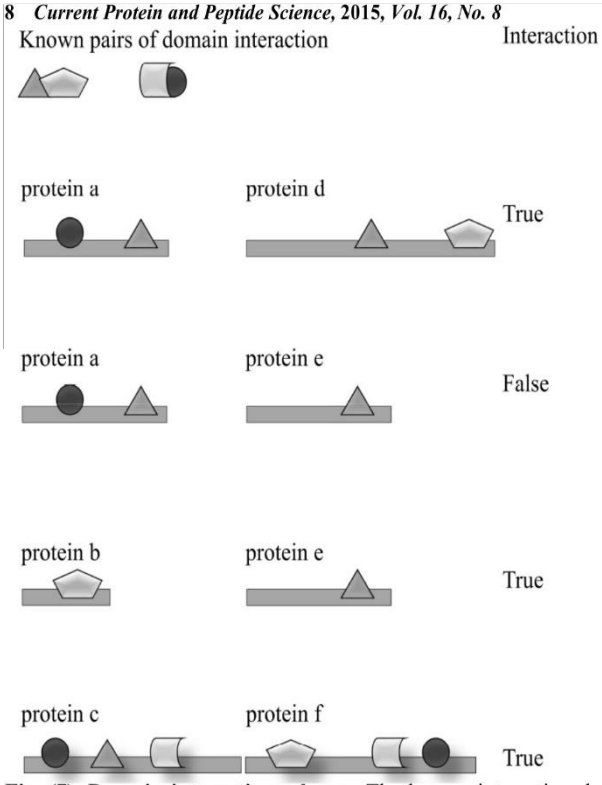


Fig. (7). Domain interaction schema. The known interacting domains are accounted for between the tested protein pairs. The prediction method uses different criteria to determine whether the pair of proteins interacts or not.

The pair exclusion method is an extension of the maximum likelihood estimation method [34], which is not capable of detecting specific domain interactions. The specific domain interactions were estimated using an Eij score function [134] that measures the evidence that domains i and j interact and is defined as the logarithm of a ratio of two probabilities. The numerator matches the probability that two proteins interact when domains i and j interact, whereas the denominator corresponds to the probability that proteins interact when domains i and j do not interact. To compute E-scores for a given domain pair, the probability in the numerator is calculated using the expectation maximization procedure. To compute the probability in the denominator, the procedure is repeated with the likelihood of a given pair of interacting domains interacting is set to zero. This allows the competing domains to maximize Θ_{ij}. A high E-score indicates a high tendency for two domains to interact, while a low E-score indicates that competing domains from the same protein pair are more likely to be responsible for the interaction. Therefore, specific domain interactions can be identified by screening for low Θ values and high E-scores [34]. Although this model does not explain the false positives and false negatives in the experimental data, the E-score has been shown to perform better than its constituent quantities by identifying 2.9 times more true positives than random assignments, whereas h values yield 1.4 times more true positives than random assignments [134].

The p-value method is based on the assessments of null hypothesis and performs well for multi-domain protein pairs where the presence of a particular domain pair in a protein pair has no effect on whether or not two proteins interact [135]. This method calculates the statistics of each domain of protein pairs and takes into account the experimental error and incompleteness of the dataset. The theoretical distribution is obtained by shuffling protein domains so that the network of protein interactions remains fixed. The obtained p-values indicate the dependability of domain interactions between two proteins.

Protein domains are highly conserved and are therefore considered for structure-based drug design. For more details please see the references [13, 136, 137]. The predictions made by domain-based methods are considered primary as new interactions can be identified without relying on interactions already reported in the literature or that exist in a public database. The use of domain interactions for the prediction of PPIs can be extremely sensitive, especially when interactions between single or multi domains are considered in protein pairs [138]. However, if only domain interactions are considered and external factors (prosthetic groups) are ignored, false positives could reduce the specificity because a particular domain may not be in a functional state because of the 3D conformation of the protein or the presence of inhibitors in the binding site [139] (Table **1**). Domain-based methods were applied to predict PPIs in *A. thaliana* [56], *S. cerevisiae* [132], and *E. coli* [137].

2.6. Machine Learning-Based Approach

Machine learning-based approaches have been used in many areas, from science to business. These approaches use a gold standard dataset as input and assume that the classifier can be trained to learn from these data. For PPI predictions based on biological data, it is assumed that the algorithm can be trained to differentiate between true and false interactions (Fig. **8**). Therefore, after training, the classifier should be able to infer if a given pair of proteins interacts or not. To train the classifier, any kind of biological information, e.g., attributes or characteristics, including those data used in the prediction approaches described here, can be used. For example, the data can be the position of the genes in the genome, the phylogenetic profile, the conservation of proteins between different species, sequence homology, domains, the positions of the amino acids in a sequence, the degree of conservation, and physicochemical or biochemical characteristics of the proteins etc. that may influence the PPIs [56, 111, 140-146]. The learning component of the algorithm uses the values of the attributes supplied. Machine learning-based methods use diverse strategies to learn from the training data set and to identify attribute values that can differentiate between true and false interactions. Such differentiation can occur through algorithms of pattern recognition, sorting, grouping, or the generation of rules for attributes and their values [140, 147-152]. The learning process aims to divide the attributes and their values into two categories: those that represent interactions and those that do not represent interactions. Sometimes, this division is not binary but rather based on a scale of statistical confidence. If the features or combinations of these attributes and values are not conclusive for determining whether an interaction is true or not (e.g., in

case of a random distribution of characteristics), the results also may not be conclusive and a high number of false positive and false negative predictions will be produced [153]. For more details, please see the references [146, 154].

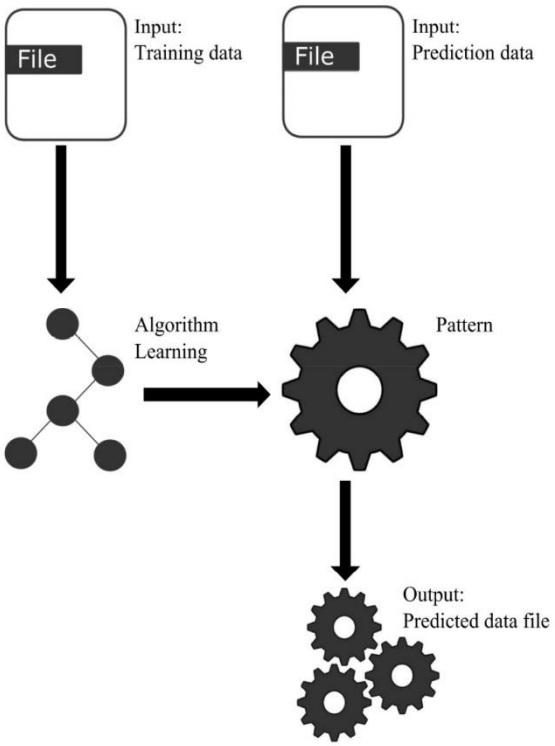


Fig. (8). Machine learning schema. The input training data set must contain true and false interactions.

To train an algorithm, it is important to have a dataset of positive interactions, preferably curated, as well as a dataset of negative interactions [35, 48] because this will help the algorithm learn the characteristics that determine an interaction and the characteristics that do not. When the positive and negative interactions are clearly differentiated in the training dataset, machine-learning is likely to be accurate, meaning that the predictions will be reasonably precise. In other words, the accuracy of a prediction is related directly to the quality of the learning of each method [12, 35]. Today, with the large amount of PPI data publicly available, including curated data [47], it is not very difficult to create a good training set.

The predictions that use machine learning can be considered as primary because such algorithms, after learning based on the information in the training dataset, are able to predict new interactions in any organism not described in the literature (Table **1**).

The most common approaches that use machine learning to predict PPIs are: (i) the random forest classifier [140], which has been used to predict PPIs in *S. cerevisiae* [2], between HIV-1 and human host cellular proteins [57], *Populus trichocarpa* [155], and human and hepatitis C virus [146]; (ii) the naïve Bayes classifier, which was used to predict PPIs in a human-microbial oral interactome [54], *P. falciparum*, *Plasmodium yoelii*, *Toxoplasma gondii*, and *Cryptosporidium parvum* [115], human cancer [145], and human and hepatitis C virus [146]; (iii) support vector machines, which have been used to predict PPIs in *A. thaliana* [56], *M. musculus* [53], human and hepatitis C virus [146], *Helicobacter pylori* [156], and human and mouse [156]; (iv) the K-nearest neighbors approach, which has been used to identify genes related to diseases in human PPIs [157], *H. pylori* and human [142], and water-mediated ligand interactions [143]; and (v) the decision tree [144, 158], which was used to predict PPIs in *Drosophila melanogaster* [159].

CONCLUSION

Experimental and computational methods that are used to test specific interactions in biological data both tend to be specific, but have the disadvantage of low sensitivity. Because the correlation between sensitivity and specificity is usually negative, a good strategy is to use combined approaches; for example, a sensitive method to identify possible interactions and a more specific method to select the most reliable interactions. Another strategy could be to use a sensitive method with a carefully selected cutoff point based on the characteristics of the data, to help achieve a good balance between sensitivity and specificity [122].

When combining different methods, it is a good idea to consider that experimental and computational approaches can complement each other [26, 160]. Experimental methods can be used to generate data that can be used as inputs to computational methods that can extract patterns and make predictions based on algorithms and machine learning techniques. The more representative the experimental dataset, the more reliable are the results obtained *in silico*. In a second step, additional experimental data can be used to validate the predictions made by the computational methods. On the other hand, computational methods can work with large volumes of data and, for example, predict millions of interactions that generate new knowledge about organisms and diseases, besides creating new biological hypotheses and guide new experiments [130, 137, 153, 157, 160].

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

We thank CAPES (Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior), CNPq (Conselho Nacional de Pesquisa), and Fapemig (Fundação de Amparo à Pesquisa do Estado de Minas Gerais) for financial support.

AUTHOR CONTRIBUTIONS

ELF conceived and designed the structure of the article. AFOJ designed the pictures. ALL, DB, PG wrote the paper and revised the draft. AS and VA contributed the materials.

REFERENCES

[1] Shoemaker, B.A.; Panchenko, A.R. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput. Biol.,* **2007,** *3*(3), e42.

[2] Chen, X.-W.; Liu, M. Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics,* **2005,** *21*(24), 4394-4400.

[3] Quail, M.A.; Smith, M.; Coupland, P.; Otto, T.D.; Harris, S.R.; Connor, T.R.; Bertoni, A.; Swerdlow, H.P.; Gu, Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genom.,* **2012,** *13*(1), 341.

[4] Koboldt, D.C.; Steinberg, K.M.; Larson, D.E.; Wilson, R.K.; Mardis, E.R. The next-generation sequencing revolution and its impact on genomics. *Cell,* **2013,** *155*(1), 27-38.

[5] Mutz, K.-O.; Heilkenbrinker, A.; Lönne, M.; Walter, J.-G.; Stahl, F. Transcriptome analysis using next-generation sequencing. *Curr. Opin. Biotechnol.,* **2013,** *24*(1), 22-30.

[6] Casals, F.; Idaghdour, Y.; Hussin, J.; Awadalla, P. Next-generation sequencing approaches for genetic mapping of complex diseases. *J. Neuroimmunol.,* **2012,** *248*(1), 10-22.

[7] van Dijk, E.L.; Auger, H.; Jaszczyszyn, Y.; Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.,* **2014,** *30*(9), 418-426.

[8] Mitra, R.; Gill, R.; Datta, S.; Datta, S. Statistical Analyses of Next Generation Sequencing Data: An Overview. In *Statistical Analysis of Next Generation Sequencing Data*, Springer: **2014**; pp. 1-24.

[9] Liu, L.; Li, Y.; Li, S.; Hu, N.; He, Y.; Pong, R.; Lin, D.; Lu, L.; Law, M. Comparison of next-generation sequencing systems. *BioMed Res. Int.,* **2012,** *2012*, 1-11.

[10] Gonzalez, M.W.; Kann, M.G. Protein interactions and disease. *PLoS Comput. Biol.,* **2012,** *8*(12), e1002819.

[11] Wetie, A.G.N.; Sokolowska, I.; Woods, A.G.; Roy, U.; Deinhardt, K.; Darie, C.C. Protein-protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. *Cell. Mol. Life Sci.,* **2014,** *71*(2), 205-228.

[12] Zahiri, J.; Hannon Bozorgmehr, J.; Masoudi-Nejad, A. Computational Prediction of Protein-Protein Interaction Networks: Algorithms and Resources. *Curr. Genom.,* **2013,** *14*(6), 397-414.

[13] Rao, V.S.; Srinivas, K.; Sujini, G.; Kumar, G. Protein-Protein Interaction Detection: Methods and Analysis. *Int. J. Proteomics,* **2014,** *2014*, 147648.

[14] Andreani, J.; Guerois, R. Evolution of Protein Interactions: From Interactomes to Interfaces. *Arch. Biochem. Biophys.,* **2014**, *554*, 65-75.

[15] Yu, H.; Braun, P.; Yıldırım, M.A.; Lemmens, I.; Venkatesan, K.; Sahalie, J.; Hirozane-Kishikawa, T.; Gebreab, F.; Li, N.; Simonis, N. High-quality binary protein interaction map of the yeast interactome network. *Science,* **2008,** *322*(5898), 104-110.

[16] Chien, C.-T.; Bartel, P.L.; Sternglanz, R.; Fields, S. The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc. Natl. Acad. Sci. U. S. A.,* **1991,** *88*(21), 9578-9582.

[17] Jankute, M.; Byng, C.V.; Alderwick, L.J.; Besra, G.S. Elucidation of a protein-protein interaction network involved in Corynebacterium glutamicum cell wall biosynthesis as determined by bacterial two-hybrid analysis. *Glycoconj. J.,* **2014**, *31*(6-7), 475-483.

[18] Zhu, H.; Snyder, M. Protein chip technology. *Curr. Opin. Chem. Biol.,* **2003,** *7*(1), 55-63.

[19] Lee, Y.; Lee, E.K.; Cho, Y.W.; Matsui, T.; Kang, I.C.; Kim, T.S.; Han, M.H. ProteoChip: A highly sensitive protein microarray prepared by a novel method of protein immobilization for application of protein-protein interaction studies. *Proteomics,* **2003,** *3*(12), 2289-2304.

[20] Sun, X.; Hong, P.; Kulkarni, M.; Kwon, Y.; Perrimon, N. In *An advanced method for identifying protein-protein interaction by analyzing TAP/MS data*, Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on, IEEE: **2012**; pp. 1-6.

[21] Bauer, A.; Kuster, B. Affinity purification-mass spectrometry. *Eur. J. Biochem.,* **2003,** *270*(4), 570-578.

[22] Shevchenko, A.; Schaft, D.; Roguev, A.; Pijnappel, W.P.; Stewart, A.F.; Shevchenko, A. Deciphering protein complexes and protein interaction networks by tandem affinity purification and mass spectrometry analytical perspective. *Mol. Cell. Proteomics,* **2002,** *1*(3), 204-212.

[23] De Las Rivas, J.; Fontanillo, C. Protein-protein interaction networks: unraveling the wiring of molecular machines within the cell. *Brief. Funct. Genomics,* **2012**, *11*(6), 489-496.

[24] Kao, F.S.; Ger, W.; Pan, Y.R.; Yu, H.C.; Hsu, R.Q.; Chen, H.M. Chip-based protein-protein interaction studied by atomic force microscopy. *Biotechnol. Bioeng.,* **2012**, *109*(10), 2460-2467.

[25] Wang, J.; Li, M.; Deng, Y.; Pan, Y. Recent advances in clustering methods for protein interaction networks. *BMC Genom.,* **2010**, *11*(Suppl 3), S10.

[26] Harrington, E.D.; Jensen, L.J.; Bork, P. Predicting biological networks from genomic data. *FEBS Lett.,* **2008,** *582*(8), 1251-1258.

[27] Barabási, A.L.; Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.,* **2004**, *5*(2), 101-113.

[28] Wetie, N.; Armand, G.; Sokolowska, I.; Woods, A.G.; Roy, U.; Loo, J.A.; Darie, C.C. Investigation of stable and transient protein-protein interactions: Past, present, and future. *Proteomics,* **2013**, *13*(3-4), 538-557.

[29] Rivas, G.; Stafford, W.; Minton, A.P. Characterization of heterologous protein-protein interactions using analytical ultracentrifugation. *Methods,* **1999,** *19*(2), 194-212.

[30] Laue, T.M. Analytical ultracentrifugation. *Curr. Protoc. Protein Sci.,* **2001**, 7.5. 1-7.5. 9.

[31] Park, K.D.; Guo, K.; Adebodun, F.; Chiu, M.L.; Sligar, S.G.; Oldfield, E. Distal and proximal ligand interactions in heme proteins: correlations between carbon-oxygen and iron-carbon vibrational frequencies, oxygen-17 and carbon-13 nuclear magnetic resonance chemical shifts, and oxygen-17 nuclear quadrupole coupling constants in [C17O-] and [13CO]-labeled species. *Biochemistry,* **1991,** *30*(9), 2333-2347.

[32] Bax, A. Multidimensional nuclear magnetic resonance methods for protein studies. *Curr. Opin. Struct. Biol.,* **1994,** *4*(5), 738-744.

[33] Xing, Q.; Huang, P.; Yang, J.; Sun, J.Q.; Gong, Z.; Dong, X.; Guo, D.C.; Chen, S.M.; Yang, Y.H.; Wang, Y. Visualizing an Ultra-Weak Protein-Protein Interaction in Phosphorylation Signaling. *Angewandte Chemie,* **2014**, *53*(43), 11501-11505.

[34] Shoemaker, B.A.; Panchenko, A.R. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.,* **2007,** *3*(4), e43.

[35] Qi, Y.; Bar-Joseph, Z.; Klein-Seetharaman, J. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins,* **2006,** *63*(3), 490-500.

[36] Jonsson, P.F.; Bates, P.A. Global topological features of cancer proteins in the human interactome. *Bioinformatics,* **2006,** *22*(18), 2291-2297.

[37] Lu, L. J.; Xia, Y.; Paccanaro, A.; Yu, H.; Gerstein, M. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.,* **2005,** *15*(7), 945-953.

[38] Burgoyne, N.J.; Jackson, R.M. Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics,* **2006,** *22*(11), 1335-1342.

[39] Li, D.; Liu, W.; Liu, Z.; Wang, J.; Liu, Q.; Zhu, Y.; He, F. PRINCESS, a protein interaction confidence evaluation system with multiple data sources. *Mol. Cell. Proteomics,* **2008,** *7*(6), 1043-1052.

[40] Wang, H.; Segal, E.; Ben-Hur, A.; Li, Q.-R.; Vidal, M.; Koller, D. InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale. *Genome Biol.,* **2007,** *8*(9), R192.

[41] Rezende, A.M.; Folador, E.L.; Resende, D.M.; Ruiz, J.C. Computational Prediction of Protein-Protein Interactions in Leishmania Predicted Proteomes. *PloS One,* **2012,** *7*(12), e51304.

[42] Xenarios, I.; Rice, D.W.; Salwinski, L.; Baron, M.K.; Marcotte, E.M.; Eisenberg, D. DIP: the database of interacting proteins. *Nucleic Acids Res.,* **2000,** *28*(1), 289-291.

[43] Hermjakob, H.; Montecchi-Palazzi, L.; Lewington, C.; Mudali, S.; Kerrien, S.; Orchard, S.; Vingron, M.; Roechert, B.; Roepstorff, P.; Valencia, A. IntAct: an open source molecular interaction database. *Nucleic Acids Res.,* **2004,** *32*(suppl 1), D452-D455.

[44] Franceschini, A.; Szklarczyk, D.; Frankild, S.; Kuhn, M.; Simonovic, M.; Roth, A.; Lin, J.; Minguez, P.; Bork, P.; von Mering, C. STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.,* **2013,** *41*(D1), D808-D815.

[45] Zanzoni, A.; Montecchi-Palazzi, L.; Quondam, M.; Ausiello, G.; Helmer-Citterich, M.; Cesareni, G. MINT: a Molecular INTeraction database. *FEBS Lett.,* **2002,** *513*(1), 135-140.

[46] de Juan, D.; Pazos, F.; Valencia, A. Emerging methods in protein co-evolution. *Nat. Rev. Genet.,* **2013,** *14*(4), 249-261.

[47] Orchard, S.; Kerrien, S.; Abbani, S.; Aranda, B.; Bhate, J.; Bidwell, S.; Bridge, A.; Briganti, L.; Brinkman, F.S.; Cesareni, G. Protein

interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods,* **2012,** *9*(4), 345-350.

[48] Blohm, P.; Frishman, G.; Smialowski, P.; Goebels, F.; Wachinger, B.; Ruepp, A.; Frishman, D. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res.,* **2013**, gkt1079.

[49] Ben-Hur, A.; Noble, W.S. Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics,* **2006,** *7*(Suppl 1), S2.

[50] Zhou, H.; Jakobsson, E. Predicting Protein-Protein Interaction by the Mirrortree Method: Possibilities and Limitations. *PloS One,* **2013,** *8*(12), e81100.

[51] Valencia, A.; Pazos, F. Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.,* **2002,** *12*(3), 368-373.

[52] Ochoa, D.; Pazos, F. Practical aspects of protein co-evolution. *Mol. Med.,* **2014,** *2*, 14.

[53] Yellaboina, S.; Dudekula, D.B.; Ko, M.S. Prediction of evolutionarily conserved interologs in Mus musculus. *BMC Genom.,* **2008,** *9*(1), 465.

[54] Coelho, E.D.; Arrais, J.P.; Matos, S.; Pereira, C.; Rosa, N.; Correia, M.J.; Barros, M.; Oliveira, J.L. Computational prediction of the human-microbial oral interactome. *BMC Syst. Biol.,* **2014,** *8*(1), 24.

[55] Cui, J.; Li, P.; Li, G.; Xu, F.; Zhao, C.; Li, Y.; Yang, Z.; Wang, G.; Yu, Q.; Li, Y. AtPID: Arabidopsis thaliana protein interactome database—an integrative platform for plant systems biology. *Nucleic Acids Res.,* **2008,** *36*(suppl 1), D999-D1008.

[56] Lin, M.; Shen, X.; Chen, X. PAIR: the predicted Arabidopsis interactome resource. *Nucleic Acids Res.,* **2011,** *39*(suppl 1), D1134-D1140.

[57] Tastan, O.; Qi, Y.; Carbonell, J.G.; Klein-Seetharaman, J. In *Prediction of interactions between HIV-1 and human proteins by information integration*, Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, NIH Public Access: **2009**; p. 516.

[58] Li, L.; Jing, L.; Huang, D. In *Protein-protein interaction extraction from biomedical literatures based on modified SVM-KNN*, Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference on, IEEE: **2009**; pp. 1-7.

[59] Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T.A.; Judson, R.S.; Knight, J.R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature,* **2000,** *403*(6770), 623-627.

[60] Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M.; Hattori, M.; Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.,* **2001,** *98*(8), 4569-4574.

[61] Giot, L.; Bader, J.S.; Brouwer, C.; Chaudhuri, A.; Kuang, B.; Li, Y.; Hao, Y.; Ooi, C.; Godwin, B.; Vitols, E. A protein interaction map of Drosophila melanogaster. *Science,* **2003,** *302*(5651), 1727-1736.

[62] Collura, V.; Boissy, G. From Protein—Protein Complexes to Interactomics. In *Subcellular Proteomics*, Springer: **2007**; pp. 135-183.

[63] Fox, J.L. Electoral campaign delays resolution of biotech issues. *Nat. Biotechnol.,* **2004,** *22*(10), 1193.

[64] Russell, R.B.; Alber, F.; Aloy, P.; Davis, F.P.; Korkin, D.; Pichaud, M.; Topf, M.; Sali, A. A structural perspective on protein-protein interactions. *Curr. Opin. Struct. Biol.,* **2004,** *14*(3), 313-324.

[65] Lahti, J.L.; Tang, G.W.; Capriotti, E.; Liu, T.; Altman, R.B. Bioinformatics and variability in drug response: a protein structural perspective. *J. R. Soc. Interface,* **2012,** *9*(72), 1409-1437.

[66] Grosdidier, S.; Totrov, M.; Fernández-Recio, J. Computer applications for prediction of protein-protein interactions and rational drug design. *Adv. Appl. Bioinform. Chem.,* **2009,** *2*, 101.

[67] Mikael, L.G.; Pawelek, P.D.; Labrie, J.; Sirois, M.; Coulton, J.W.; Jacques, M. Molecular cloning and characterization of the ferric hydroxamate uptake (fhu) operon in Actinobacillus pleuropneumoniae. *Microbiology,* **2002,** *148*(9), 2869-2882.

[68] Li, Y. Y.; An, J.; Jones, S.J. A computational approach to finding novel targets for existing drugs. *PLoS Comput. Biol.,* **2011,** *7*(9), e1002139.

[69] Piper, S.H. Some Examples of Information Obtainable From The Long Spancings of Fatty Acids. *Transactions Faraday Soc.,* **1929,** (348), 348-351.

[70] Dickinson, R.G.; Raymond, A.L. The crystal structure of hexamethylene-tetramine. *J. Am. Chem. Soc.,* **1923,** *45*(1), 22-29.

[71] Bragg, W. The Investigation of the Properties of Thin Films by Means of X-rays. *Nature,* **1925,** *115*(2886), 266-269.

[72] Muller, A. The Connection between the Zig-Zag Structure of the Hydrocarbon Chain and the Alternations in the Properties of Odd and Even Numbered Chain Compounds. *Proc. R. Soc. Lond. Series A,* **1929,** *124*(794), 317-321.

[73] Saunders, M.; Wishnia, A.; Kirkwood, J.G. The nuclear magnetic resonance spectrum of ribonuclease1. *J. Am. Chem. Soc.,* **1957,** *79*(12), 3289-3290.

[74] Li, B.; Kihara, D. Protein docking prediction using predicted protein-protein interface. *BMC Bioinformatics,* **2012,** *13*(1), 7.

[75] Wiithrich, K. Protein Structure Determination in Solution by NMR Spectroscopy. *J. Biol. Chem.,* **1990,** *265* (36), 22059-22062.

[76] Sanchez, R.; Sali, A. Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins,* **1997,** *29*(s 1), 50-58.

[77] Santos Filho, O.A.; Alencastro, R.B. d. Modelagem de proteínas por homologia. *Química Nova,* **2003,** *26*(2), 253-259.

[78] Krüger, D.M.; Garzón, J.I.; Chacón, P.; Gohlke, H. DrugScorePPI Knowledge-Based Potentials Used as Scoring and Objective Function in Protein-Protein Docking. *PloS One,* **2014,** *9*(2), e89466.

[79] Lemmon, G.; Meiler, J. Towards ligand docking including explicit interface water molecules. *PloS One,* **2013,** *8*(6), e67536.

[80] Leis, S.; Zacharias, M. ReFlexIn: a flexible receptor protein-ligand docking scheme evaluated on HIV-1 protease. *PloS One,* **2012,** *7*(10), e48008.

[81] Langridge, R.; Ferrin, T.E.; Kuntz, I.D.; Connolly, M.L. Real-time color graphics in studies of molecular interactions. *Science,* **1981,** *211*(4483), 661-666.

[82] Connolly, M.L. Shape complementarity at the hemoglobin α1β1 subunit interface. *Biopolymers,* **1986,** *25*(7), 1229-1247.

[83] Mashiach, E.; Schneidman-Duhovny, D.; Andrusier, N.; Nussinov, R.; Wolfson, H.J. FireDock: a web server for fast interaction refinement in molecular docking. *Nucleic Acids Res.,* **2008,** *36*(suppl 2), W229-W232.

[84] Levine, D.; Facello, M.; Hallstrom, P.; Reeder, G.; Walenz, B.; Stevens, F. Stalk: An interactive system for virtual molecular docking. *Comput. Sci. Eng.,* **1997,** *4*(2), 55-65.

[85] Lyskov, S.; Gray, J.J. The RosettaDock server for local protein-protein docking. *Nucleic Acids Res.,* **2008,** *36*(suppl 2), W233-W238.

[86] Hawse, W.F.; Champion, M.M.; Joyce, M.V.; Hellman, L.M.; Hossain, M.; Ryan, V.; Pierce, B.G.; Weng, Z.; Baker, B.M. Cutting edge: evidence for a dynamically driven T cell signaling mechanism. *J. Immunol.,* **2012,** *188*(12), 5819-5823.

[87] Lesk, V.I.; Sternberg, M.J. 3D-Garden: a system for modelling protein-protein complexes based on conformational refinement of ensembles generated with the marching cubes algorithm. *Bioinformatics,* **2008,** *24*(9), 1137-1144.

[88] Ghoorah, A.W.; Devignes, M.D.; Smaïl-Tabbone, M.; Ritchie, D.W. Protein docking using case-based reasoning. *Proteins,* **2013,** *81*(12), 2150-2158.

[89] Wetie, A.G.N.; Sokolowska, I.; Woods, A.G.; Roy, U.; Deinhardt, K.; Darie, C.C. Protein-protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. *Cell. Mol. Life Sci.,* **2013**, 1-24.

[90] Lengauer, T.; Rarey, M. Computational methods for biomolecular docking. *Curr. Opin. Struct. Biol.,* **1996,** *6*(3), 402-406.

[91] Norel, R.; Lin, S.L.; Wolfson, H.J.; Nussinov, R. Molecular surface complementarity at protein-protein interfaces: the critical role played by surface normals at well placed, sparse, points in docking. *J. Mol. Biol.,* **1995,** *252*(2), 263-273.

[92] McCammon, J.A.; Harvey, S.C. *Dynamics of proteins and nucleic acids*. Cambridge University Press: **1988**.

[93] Rapaport, D.C.; Blumberg, R.L.; McKay, S.R.; Christian, W. The Art of Molecular Dynamics Simulation. *Comput. Phys.,* **1996,** *10*(5), 456-456.

[94] Allen, M.P.; Tildesley, D.J. Computer simulation of liquids. **1987**.

[95] Frenkel, D.; Smit, B. Understanding molecular simulation: from algorithms to applications. *Comput. Sci. Series,* **2002,** *1*, 1-638.

[96] Monji, H.; Koizumi, S.; Ozaki, T.; Ohkawa, T. Interaction site prediction by structural similarity to neighboring clusters in protein-protein interaction networks. *BMC Bioinformatics,* **2011,** *12*(Suppl 1), S39.

[97] Kar, G.; Gursoy, A.; Keskin, O. Human cancer protein-protein interaction network: a structural perspective. *PLoS Comput. Biol.,* **2009,** *5*(12), e1000601.

[98] Doolittle, J.M.; Gomez, S.M. Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens. *Virol. J.,* **2010,** *7*(1), 82.

[99] Franzosa, E.A.; Xia, Y. Structural principles within the human-virus protein-protein interaction network. *Proc. Natl. Acad. Sci. U. S. A.,* **2011,** *108*(26), 10538-10543.

[100] Andorf, C.M.; Honavar, V.; Sen, T.Z. Predicting the binding patterns of hub proteins: a study using yeast protein interaction networks. *PloS One,* **2013,** *8*(2), e56833.

[101] Tan, A.-H. In *Text mining: The state of the art and the challenges*, Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, **1999**; pp. 65-70.

[102] Feldman, R.; Sanger, J. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press: **2007**.

[103] Elder IV, J.; Hill, T. *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press: **2012**.

[104] Billheimer, D.D.; Booker, A.J.; Condliff, M.K.; Greaves, M.T.; Holt, F.B.; Kao, A.S.-W.; Pierce, D.J.; Poteet, S.R.; Wu, Y.-J. Method and system for text mining using multidimensional subspaces. Google Patents: **2003**.

[105] Holzinger, A.; Schantl, J.; Schroettner, M.; Seifert, C.; Verspoor, K. Biomedical text mining: State-of-the-art, open problems and future challenges. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, Springer: **2014**; pp. 271-300.

[106] Fluck, J.; Hofmann-Apitius, M. Text mining for systems biology. *Drug Discov. Today,* **2014,** *19*(2), 140-144.

[107] Ananiadou, S.; Thompson, P.; Nawaz, R.; McNaught, J.; Kell, D.B. Event-based text mining for biology and functional genomics. *Brief. Funct. Genomics,* **2014**, elu015.

[108] Czarnecki, J.; Shepherd, A.J. Mining Biological Networks from Full-Text Articles. In *Biomedical Literature Mining*, Springer: **2014**; pp. 135-145.

[109] Krallinger, M.; Leitner, F.; Vazquez, M.; Salgado, D.; Marcelle, C.; Tyers, M.; Valencia, A.; Chatr-Aryamontri, A. How to link ontologies and protein-protein interactions to literature: text-mining approaches and the BioCreative experience. *Database,* **2012,** *2012*, bas017.

[110] Köster, J.; Zamir, E.; Rahmann, S. Efficiently mining protein interaction dependencies from large text corpora. *Integr. Biol.,* **2012,** *4*(7), 805-812.

[111] Thieu, T.; Joshi, S.; Warren, S.; Korkin, D. Literature mining of host-pathogen interactions: comparing feature-based supervised learning and language-based approaches. *Bioinformatics,* **2012,** *28*(6), 867-875.

[112] Xiang, Z.; Qin, T.; Qin, Z.S.; He, Y. A genome-wide MeSH-based literature mining system predicts implicit gene-to-gene relationships and networks. *BMC Syst. Biol.,* **2013,** *7*(Suppl 3), S9.

[113] Van Landeghem, S.; De Bodt, S.; Drebert, Z. J.; Inzé, D.; Van de Peer, Y. The potential of text mining in data integration and network biology for plant research: a case study on Arabidopsis. *Plant Cell Online,* **2013,** *25*(3), 794-807.

[114] Marcotte, E.M.; Pellegrini, M.; Ng, H.-L.; Rice, D.W.; Yeates, T.O.; Eisenberg, D. Detecting protein function and protein-protein interactions from genome sequences. *Science,* **1999,** *285*(5428), 751-753.

[115] Date, S.V.; Stoeckert, C.J. Computational modeling of the Plasmodium falciparum interactome reveals protein function on a genome-wide scale. *Genome Res.,* **2006,** *16*(4), 542-549.

[116] Juan, D.; Pazos, F.; Valencia, A. High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl. Acad. Sci. U. S. A.,* **2008,** *105*(3), 934-939.

[117] Pazos, F.; Ranea, J.A.; Juan, D.; Sternberg, M.J. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J. Mol. Biol.,* **2005,** *352*(4), 1002-1015.

[118] Liu, C.H.; Li, K.-C.; Yuan, S. Human protein-protein interaction prediction by a novel sequence-based co-evolution method: co-evolutionary divergence. *Bioinformatics,* **2013,** *29*(1), 92-98.

[119] Hakes, L.; Lovell, S.C.; Oliver, S.G.; Robertson, D.L. Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc. Natl. Acad. Sci. U. S. A.,* **2007,** *104*(19), 7999-8004.

[120] Gong, S.; Yoon, G.; Jang, I.; Bolser, D.; Dafas, P.; Schroeder, M.; Choi, H.; Cho, Y.; Han, K.; Lee, S. PSIbase: a database of Protein Structural Interactome map (PSIMAP). *Bioinformatics,* **2005,** *21*(10), 2541-2543.

[121] Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. BLAST+: architecture and applications. *BMC Bioinformatics,* **2009,** *10*(1), 421.

[122] Folador, E.L.; Hassan, S.S.; Lemke, N.; Barh, D.; Silva, A.; Ferreira, R.S.; Azevedo, V. An improved interolog mapping-based computational prediction of protein-protein interactions with increased network coverage. *Integr. Biol. (Camb).* **2014**, *6*(11), 1080-1087.

[123] Huang, T.-W.; Tien, A.-C.; Huang, W.-S.; Lee, Y.-C. G.; Peng, C.-L.; Tseng, H.-H.; Kao, C.-Y.; Huang, C.-Y. F. POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics,* **2004,** *20*(17), 3273-3276.

[124] Brown, K.R.; Jurisica, I. Online predicted human interaction database. *Bioinformatics,* **2005,** *21*(9), 2076-2082.

[125] Gu, H.; Zhu, P.; Jiao, Y.; Meng, Y.; Chen, M. PRIN: a predicted rice interactome network. *BMC Bioinformatics,* **2011,** *12*(1), 161.

[126] Geisler-Lee, J.; O'Toole, N.; Ammar, R.; Provart, N. J.; Millar, A. H.; Geisler, M. A predicted interactome for Arabidopsis. *Plant Physiol.,* **2007,** *145*(2), 317-329.

[127] Ji, C.; Cao, X.; Yao, C.; Xue, S.; Xiu, Z. Protein-protein interaction network of the marine microalga Tetraselmis subcordiformis: prediction and application for starch metabolism analysis. *J. Ind. Microbiol. Biotechnol.,* **2014**, *41*(8), 1287-1296.

[128] Pesch, R.; Zimmer, R. Complementing the Eukaryotic Protein Interactome. *PloS One,* **2013,** *8*(6), e66635.

[129] Li, S.; Armstrong, C.M.; Bertin, N.; Ge, H.; Milstein, S.; Boxem, M.; Vidalain, P.-O.; Han, J.-D. J.; Chesneau, A.; Hao, T. A map of the interactome network of the metazoan C. elegans. *Science,* **2004,** *303*(5657), 540-543.

[130] Alexeyenko, A.; Wassenberg, D.M.; Lobenhofer, E.K.; Yen, J.; Linney, E.; Sonnhammer, E.L.; Meyer, J.N. Dynamic zebrafish interactome reveals transcriptional mechanisms of dioxin toxicity. *PloS One,* **2010,** *5*(5), e10465.

[131] Brown, K.R.; Jurisica, I. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.,* **2007,** *8*(5), R95.

[132] Sprinzak, E.; Margalit, H. Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol.,* **2001,** *311*(4), 681-692.

[133] Deng, M.; Mehta, S.; Sun, F.; Chen, T. Inferring domain-domain interactions from protein-protein interactions. *Genome Res.,* **2002,** *12*(10), 1540-1548.

[134] Riley, R.; Lee, C.; Sabatti, C.; Eisenberg, D. Inferring protein domain interactions from databases of interacting proteins. *Genome Biol.,* **2005,** *6*(10), R89.

[135] Nye, T.M.; Berzuini, C.; Gilks, W.R.; Babu, M.M.; Teichmann, S.A. Statistical analysis of domains in interacting protein pairs. *Bioinformatics,* **2005,** *21*(7), 993-1001.

[136] Memišević, V.; Wallqvist, A.; Reifman, J. Reconstituting protein interaction networks using parameter-dependent domain-domain interactions. *BMC Bioinformatics,* **2013,** *14*(1), 154.

[137] Wojcik, J.; Schächter, V. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics,* **2001,** *17*(suppl 1), S296-S305.

[138] Jang, W.H.; Jung, S.H.; Han, D.S. A Computational Model for Predicting Protein Interactions Based on Multidomain Collaboration. *IEEE/ACM Trans. Comput. Biol. Bioinform.,* **2012,** *9*(4), 1081-1090.

[139] Watkins, A.M.; Arora, P.S. Structure-based inhibition of protein-protein interactions. *Eur. J. Med. Chem.,* **2014**, *94*, 480-488.

[140] Li, B.Q.; Feng, K.Y.; Chen, L.; Huang, T.; Cai, Y.D. Prediction of Protein-Protein Interaction Sites by Random Forest Algorithm with mRMR and IFS. *PloS One,* **2012,** *7*(8), e43927.

[141] Masso, M.; Vaisman, I.I. Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics,* **2008,** *24*(18), 2002-2009.

[142] Nanni, L.; Lumini, A. An ensemble of K-local hyperplanes for predicting protein-protein interactions. *Bioinformatics,* **2006,** *22*(10), 1207-1210.

[143] Raymer, M.L.; Sanschagrin, P.C.; Punch, W.F.; Venkataraman, S.; Goodman, E.D.; Kuhn, L.A. Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm. *J. Mol. Biol.,* **1997,** *265*(4), 445-464.

[144] Darnell, S.J.; Page, D.; Mitchell, J.C. An automated decision-tree approach to predicting protein interaction hot spots. *Proteins,* **2007,** *68*(4), 813-823.

[145] Rhodes, D.R.; Tomlins, S.A.; Varambally, S.; Mahavisno, V.; Barrette, T.; Kalyana-Sundaram, S.; Ghosh, D.; Pandey, A.; Chinnaiyan, A.M. Probabilistic model of the human protein-protein interaction network. *Nat. Biotechnol.,* **2005,** *23*(8), 951-959.

[146] Emamjomeh, A.; Goliaei, B.; Zahiri, J.; Ebrahimpour, R. Predicting of protein-protein interactions between human and hepatitis C virus via an ensemble learning method. *Mol. BioSyst.,* **2014**, *10*(12), 3147-3154.

[147] Zhang, X.; Xu, J.; Xiao, W.-x. A New Method for the Discovery of Essential Proteins. *PloS One,* **2013,** *8*(3), e58763.

[148] Craig, R.A.; Liao, L. Phylogenetic tree information aids supervised learning for predicting protein-protein interaction based on distance matrices. *BMC Bioinformatics,* **2007,** *8*(1), 6.

[149] Li, L.; Zhang, P.; Zheng, T.; Zhang, H.; Jiang, Z.; Huang, D. Integrating Semantic Information into Multiple Kernels for Protein-Protein Interaction Extraction from Biomedical Literatures. *PloS One,* **2014,** *9*(3), e91898.

[150] Zhang, S.-W.; Hao, L.-Y.; Zhang, T.-H. Prediction of Protein-Protein Interaction with Pairwise Kernel Support Vector Machine. *Int. J. Mol. Sci.,* **2014,** *15*(2), 3220-3233.

[151] Kumar, H.; Srivastava, S.; Varadwaj, P. Determination of protein-protein interaction through Artificial Neural Network and Support Vector Machine: A Comparative study. *Int. J. Comput. Biol., (IJCB),* **2014,** *3*(2), 37-43.

[152] Zhou, W.; Yan, H.; Fan, X.; Hao, Q. Prediction of Protein-Protein Interactions Based on Molecular Interface Features and the Support Vector Machine. *Curr. Bioinformatics,* **2013,** *8*(1), 3-8.

[153] Yugandhar, K.; Gromiha, M.M. Protein-protein binding affinity prediction from amino acid sequence. *Bioinformatics,* **2014**, btu580.

[154] Zahiri, J.; Bozorgmehr, J.H.; Masoudi-Nejad, A. Computational Prediction of Protein-Protein Interaction Networks: Algo-rithms and Resources. *Curr. Genom.,* **2013,** *14*(6), 397.

[155] Rodgers-Melnick, E.; Culp, M.; DiFazio, S.P. Predicting whole genome protein interaction networks from primary sequence data in model and non-model organisms using ENTS. *BMC Genom.,* **2013,** *14*(1), 608.

[156] Martin, S.; Roe, D.; Faulon, J.-L. Predicting protein-protein interactions using signature products. *Bioinformatics,* **2005,** *21*(2), 218-226.

[157] Xu, J.; Li, Y. Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics,* **2006,** *22*(22), 2800-2805.

[158] Zhang, L.V.; Wong, S.L.; King, O.D.; Roth, F.P. Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics,* **2004,** *5*(1), 38.

[159] Middendorf, M.; Ziv, E.; Wiggins, C.H. Inferring network mechanisms: the Drosophila melanogaster protein interaction network. *Proc. Natl. Acad. Sci. U. S. A.,* **2005,** *102*(9), 3192-3197.

[160] Mulder, N.J.; Akinola, R.O.; Mazandu, G.K.; Rapanoel, H. Using biological networks to improve our understanding of infectious diseases. *Comput. Struct. Biotechnol. J.*, **2014**, *11*(18), 1-10.

Received: ???????????????? **Revised:** ???????????????? **Accepted:** ????????????????

1.3 - *Corynebacterium pseudotuberculosis*

Corynebacterium pseudotuberculosis (Cp) faz parte do grupo de bactérias CMNR (*Corynebacterium, Mycobacterium, Nocardia, Rhodococcus*) (Butler, Ahearn e Kilburn, 1986). É uma bactéria patogênica intracelular facultativa, gram-positiva, possui fimbrias porém não se move, não forma capsulas e não esporula (Selim, 2001).

Cp se apresenta em dois biovares: *ovis* e *equi* (Songer *et al.*, 1988). O biovar *equi* infecta principalmente equinos e bovinos, já o biovar *ovis* é o agente etiológico de linfadenite caseosa (LC), uma doença crônica que afeta principalmente rebanhos de ovinos e caprinos, sendo a infecção em humanos associada à exposição profissional durante o manuseio dos rebanhos (Hémond *et al.*, 2009; Ivanović *et al.*, 2009).

Estudo realizado no estado de Minas Gerais, Brasil, mostrou que 78.9% dos animais testados foram soropositivos para LC (Seyffert *et al.*, 2010). Entretanto, o estudo de Cp se torna importante também pela prevalência em diversos países no globo (Windsor, 2011), como estado de Granada e ilhas Carriacou na India (Hariharan *et al.*, 2014), Coréia (Jung *et al.*, 2015), França (Trost *et al.*, 2010), Patagônia na Argentina (Cerdeira *et al.*, 2011), Brasil e Austrália (Ruiz *et al.*, 2011), Israel (Silva *et al.*, 2011), África (Hassan *et al.*, 2012), norte da Califórnia (Lopes *et al.*, 2012), Escócia (Pethick *et al.*, 2012; Voigt *et al.*, 2012), Espanha (Colom-Cadena *et al.*, 2014), Argélia (Mira *et al.*, 2014), região Selangor na Malásia (Osman *et al.*, 2015), Egito (Oreiby *et al.*, 2014), Turquia (SakmanoğLu *et al.*, 2015) e mais recentemente na Etiópia (Abebe e Sisay Tessema, 2015). A LC causa significantes percas econômicas em diversos países devido a baixa qualidade de carcaças, queda na produção de carne, lã e leite (Dorella *et al.*, 2006; Baird e Fontaine, 2007), além de mortalidade de animais causada por meningoencefalite supurativa (Santarosa *et al.*, 2015).

Até o ano de 2014, haviam sido sequenciadas e disponibilizadas publicamente pelo grupo de pesquisa do Laboratório de Genética Celular e Molecular (LGCM) da Universidade Federal de Minas Gerais (UFMG) e do Laboratório de Polimorfismo e DNA (LPDNA) da Universidade Federal do Pará (UFPA) 15 genomas de Cp, sendo nove linhagens do biovar *ovis* e seis do biovar *equi*. Mesmo com todas as informações genéticas disponíveis, os métodos desenvolvidos para diagnóstico e tratamento de LC ainda não são suficientemente eficazes devido Cp apresentar baixa resposta terapêutica aos medicamentos disponíveis e habilidade em persistir no meio ambiente (Williamson e Nairn, 1980; Dorella *et al.*, 2006; Oreiby *et al.*, 2014).

Considerando a resistência e prejuízos causados, Cp se torna um importante organismo para ser investigado, demandando ainda mais pesquisas da comunidade científica objetivando melhorar nosso conhecimento sobre os mecanismos moleculares e sua patogenicidade, viabilizando então, pensar em diferentes hipóteses e estratégias para o desenvolvimento de novos fármacos. Por estas razões, além dos genes, transcritos e proteínas, se faz necessário conhecer como estas moléculas interagem umas com as outras dentro da célula e com o meio ambiente para desempenharem suas funções biológicas (Barabási e Oltvai, 2004; Sharan *et al.*, 2005; Flórez *et al.*, 2010; Garma *et al.*, 2012; Gonzalez e Kann, 2012). Neste aspecto, conhecer as proteínas e suas interações é fundamental para entender os mecanismos moleculares da célula a nível de sistêmico (Wetie *et al.*, 2013; Peng *et al.*, 2014).

As redes de interação proteína-proteína (PPI) nos possibilitam ter uma visão sistêmica da biologia de um organismo a nível celular, viabilizando ainda fazer diversas análises. Além da identificação das interações e dos clusteres de proteínas que possibilita entender melhor o organismo, através de análise topológica da rede de interação, é possível identificar proteínas importantes, com potencial uso como alvos para drogas (Li *et al.*, 2012; Cui e He, 2014; Li *et al.*, 2014; Mulder *et al.*, 2014; Wetie *et al.*, 2014). Análises computacionais em uma rede de interação podem auxiliar no desenvolvimento de novas hipóteses sobre o organismo e no desenho de novos experimentos em laboratório conduzidos por estas hipóteses (Braun e Gingras, 2012; Zhang, Xu e Xiao, 2013).

Em caso de organismos patogênico, entender a rede de interação proteína-proteína, viabiliza a identificação de proteínas importantes, oferecendo consequentemente, oportunidades para o desenvolvimento de novas drogas, vacinas ou outros produtos biotecnológicos (Mosca *et al.*, 2013; Zoraghi e Reiner, 2013; Häuser *et al.*, 2014; Lage, 2014; Li *et al.*, 2014).

Devido à importância veterinária de *C. pseudotuberculosis* e conhecendo o potencial das redes de interação, visando fornecer recursos para que outros pesquisadores conheçam melhor este organismo a nível molecular e também identificar proteínas essenciais com potencial uso para diagnóstico ou alvos para fármacos, neste trabalho, foi validada uma metodologia para posterior aplicação na predição das redes de interação proteína-proteína de nove linhagens do biovar *ovis* de *C. pseudotuberculosis*.

2 - Metodologia

2.1 - An improved interolog mapping-based computational prediction of protein–protein interactions with increased network coverage

Edson Luiz Folador, Syed Shah Hassan, Ney Lemke, Debmalya Barh, Artur Silva, Rafaela Salgado Ferreira e Vasco Azevedo

Existem diversos métodos computacionais para a predição de interação proteína-proteína, cada um com vantagens e desvantagens, devendo cada metodologia ser cuidadosamente validada para que tenha sua viabilidade comprovada, principalmente quanto a sensibilidade e especificidade. Cada método computacional exige como entrada para a predição um determinado tipo de dado biológico, sendo as sequências de nucleotídeos e aminoácidos os tipos mais abundantes, principalmente devido ao surgimento das tecnologias de sequenciamento de nova geração.

O mapeamento de interações ortólogas (*Interolog mapping*) é um método que usa as sequências de aminoácidos como entrada para a predição de interações. Este método é baseado na premissa biológica que, se um par de proteínas interage em um organismo “a” e este par de proteínas é ortólogo no organismo “b”, a interação também ocorrerá no organismo “b”. Como existem vários bancos de dados de interação proteína-proteína disponíveis publicamente, o desafio em usar este método consiste em garantir que somente os pares de proteínas ortólogos sejam mapeados para o organismo de interesse.

Antes de usarmos este método para construirmos as redes de interação de *C. pseudotuberculosis*, tivemos a preocupação de o validar, comparando as interações preditas com interações experimentais e curadas (Xenarios *et al.*, 2000; Orchard *et al.*, 2012). Como resultado da validação, além de obtermos uma cobertura maior da rede de interação, identificamos um ponto de corte que melhor representasse a razão entre sensibilidade e especificidade.

O artigo referente a este trabalho foi publicado na revista Integrative Biology em setembro de 2014 com DOI número 10.1039/c4ib00136b, estando também disponível no endereço eletrônico http://pubs.rsc.org/en/content/articlehtml/2014/ib/c4ib00136b.

Integrative Biology

PAPER

View Article Online
View Journal | View Issue

Cite this: *Integr. Biol.*, 2014, 6, 1080

An improved interolog mapping-based computational prediction of protein–protein interactions with increased network coverage†

Edson Luiz Folador,*[a] Syed Shah Hassan,[a] Ney Lemke,[b] Debmalya Barh,[c] Artur Silva,[d] Rafaela Salgado Ferreira‡[e] and Vasco Azevedo‡[a]

Received 13th June 2014,
Accepted 1st September 2014

DOI: 10.1039/c4ib00136b

www.rsc.org/ibiology

Automated and efficient methods that map ortholog interactions from several organisms and public databases (pDB) are needed to identify new interactions in an organism of interest (interolog mapping). When computational methods are applied to predict interactions, it is important that these methods be validated and their efficiency proven. In this study, we compare six Blast+ metrics over three datasets to identify the best metric for protein–protein interaction predictions. Using Blast+ to align the protein pairs, the ortholog interactions from DIP were mapped to String, Intact and Psibase pDBs. For each interaction mapped to each pDBs, we retrieved the alignment score, *e*-value, bitscore, similarity, identity and coverage. We evaluated these Blast+ values, and combinations thereof, with the Receiver Operating Characteristic (ROC) curves and computed the Area Under Curve (AUC). To validate these predictions, we used a subset of the Database of Interacting Proteins (DIP) composed of experimental interactions curated by the International Molecular Exchange (IMEx). The cut-off point for each metric/pDB was computed aiming to identify the best one that separates the true and false predicted interactions. In contrast to other methods that only compute the first Blast hit, we considered the first 20 hits, thus increasing the number of predicted interaction pairs. In addition, we identified the contribution of each individual pDB, as well as their combined contribution to the prediction. The best metric had an AUC of 0.96 for a single pDB and AUC of 0.93 for combined pDBs. Compared to other studies, with a cut-off point of 0.70 representing a specificity of 0.95 and a sensitivity of 0.90 for individual pDB, our method efficiently predicts protein–protein interactions.

Insight, innovation, integration

The identification and validation of effective metrics for protein–protein interaction (PPI) predictions and mainly an increase in the coverage of the interaction network, our methodology has the potential to efficiently predict PPI in an organism. This will allow a comparison of features at the network level and a better knowledge about the target organism, thereby, driving new biological postulations and new experiments. A validated computational method to predict PPI allows the selection of specific interactions of our interest, reducing costs and increasing the success rate in the future experimental results. Likewise, identifying the contribution of each metric for each individual public database and removing the inefficient metrics are important to prevent misuse in PPI network predictions.

[a] *Department of General Biology, Instituto de Ciências Biológicas (ICB), Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil. E-mail: vasco@icb.ufmg.br; Tel: +55 31 3409 2610*
[b] *Laboratory of Bioinformatic and Computational Biofisic, Instituto de Biociência, Universidade Estadual de São Paulo (UNESP), Botucatu, São Paulo, Brazil*
[c] *Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal, India*
[d] *Instituto de Ciências Biológicas, Universidade Federal do Para, Belém, PA, Brazil*
[e] *Department of Biochemistry and Immunology, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c4ib00136b

‡ The author shares senior authorship.

1 Introduction

Understanding the dynamic nature of activities that take place inside the cell of a living organism is necessary at the systems biology level. To achieve this, it is necessary to know how the elements of cells such as the genes, transcripts, proteins and various other cellular molecules interact each other and with the outer environment to facilitate the biological functions.[1–5] In this aspect, proteins and their interactions play an important role, and therefore understanding protein–protein interactions (PPIs) is an important aspect to reveal the molecular mechanism of cell at the system level.[6,7] Analysis of PPIs helps in better

understanding the biology of phylogenetically close and even the distance organisms. PPI networks form complex systems and when such networks are computationally depicted in a graphical form; the nodes represent proteins and non-directional lines connecting these nodes represent the interactions between the proteins.[8,9] Computationally analyzed PPIs help in developing new hypotheses about an organism and designing the laboratory experiments driven by the hypotheses.[10,11] In the case of infectious microorganisms, studying PPI networks offers identification of pathogenic proteins and therefore offers new opportunities for developing novel drugs and vaccines.[12–14] The interactions of proteins within a cell depend on several biological or physicochemical factors[15] and the PPI can be physical interactions, regulatory associations, genetic interactions, structural interactions, functional similarity associations among others. Such associations are not mutually exclusive and may occur simultaneously.[8] Several methods have been developed for studying PPIs that can be categorized as genetic, biochemical, biophysical, high throughput, and computational approaches.[16] The important experimental methods include yeast-two-hybrid (Y2H),[17] protein chip, tandem affinity purification followed by mass spectrometry (TAP-MS),[18] atomic force microscopy (AFM)[4,8,9,19,20] and analytical ultracentrifugation (UC).[6] Each approach has its advantages and disadvantages and therefore more than one technique may require to eliminate the false positives.[16] Computational methods can handle entire proteome interactions but generate false-positive interactions similar to the high throughput techniques.[3,8,21] Computational prediction of PPIs and their analysis can be done using machine learning techniques,[11,22–26] protein sequence homology or interolog mapping,[27–29] three-dimensional protein structure analysis,[30–33] docking studies,[34] domain interactions,[35] text mining,[36–39] protein co-evolution approaches,[20,23,40] the Mirror tree method,[41] phylogenetic profile analysis[20] or a combination of these methods,[42] which have also been described and reviewed in other studies.[43–46] Computational methods, individually or in combination, have been used to develop and analyse PPI interaction networks in several organisms such as *Drosophila melanogaster*,[28] *Arabidopsis thaliana*,[29] *Leishmania brasiliensis*, *Leishmania major* and *Leishmania infantum*,[2,27] yeast,[17] *Saccharomyces cerevisiae*,[47] *Xanthomonas oryzae*,[48] *Helicobacter pylori*[49] and Human.[50] When the interaction network is predicted using sequence homology or interolog mapping, it is assumed that, if a pair of proteins interacts in a particular organism, the ortholog proteins in another organism will interact in a similar pattern[3,16] and is used to identify the conservation of protein interactions between two organisms when there is high similarity in the sequence of proteins[51] and transfer annotations between genomes.[52] But the prediction efficiency of interolog mapping is not yet satisfactory as compared to other computational methods.[33] This may be due to the use of only the first Blast hit.[53] Therefore there is scope of improving the method for its efficacy and accuracy in predicting and analyzing the PPI. Here, using publicly available PPI databases (pDB) both individually and collectively and a less stringent criterion for Blast+; we tried to increase the efficacy and sensitivity of interolog mapping based PPIs with minimal false-positive and false-negative interactions.

1.2 Materials and methods

1.2.1 Databases used. In this work, we have used four pDB: Database of Interacting Proteins (DIP),[54] String,[55] Intact,[56] and Psibase[57] (ESI† S1). Since the DIP contains experimental and curated data[58] for PPIs, it was used as the gold standard to evaluate our prediction. Aiming to increase the coverage of the interaction network prediction while also reducing the false negatives and false positives, we mapped the ortholog interactions and conducted the prediction of those interaction pairs found in the DIP database by comparing against three other pDBs instead of only one.[20]

1.2.2 Blast+. The BLASTp program from the Blast+ package[53] was used to align and map the ortholog proteins between the databases. All the six alignment values of BLASTp: the score, *e*-value, bit score, similarity, identity and coverage were considered to compose the metrics that will be evaluated. Aiming to validate a methodology that is able to classify non-orthologous and orthologous proteins, we run the Blast+ with the *e*-value parameter set to 0.1, all other parameters at their default value. To compare the metrics and how much each pDB contributes to the prediction of interaction pairs, we ran Blast+ to generate two distinct datasets: the first contains only the first Blast+ hit (num_alignments 1) and the second contains the first 20 Blast+ hits (num_alignments 20).

1.2.3 Interolog mapping. To map the ortholog proteins between pDBs using Blast+, we first used the DIP proteome as the query and the proteomes of the other pDBs (String, Intact and Psibase) as the subject. We then inverted this process, using the latter pDBs as the query and the DIP proteome as the subject. For the interaction analysis, only those proteins that had a reciprocal hit (RH), *i.e.*, when protein "a" from DIP align to protein "A" from the pDB and protein "A" from the pDB align to protein "a" from the DIP were considered. Specific datasets and metrics were generated for each pDB *versus* DIP combination. For each identified RH, we extracted six values from the Blast+ alignment results as mentioned before. For each reciprocal hit, the minimum value of its metric was calculated using the following formula:

$$\mathrm{RH(a)} = \min(\mathrm{BlastValue\ (a \rightarrow A)}, \mathrm{BlastValue\ (a \leftarrow A)})$$

Here, "BlastValue" represents each of the six values extracted from the Blast+ alignment that will be evaluated, "a" represents the protein in our gold standard (DIP), and "A" represents the pDB protein. The reciprocal hit (RH) is represented by both "a → A", indicating that the protein "a" in the DIP was used as the query and was aligned against the protein "A" in the pDB, and by "a ← A", indicating that the protein "A" in the pDB was used as the query and was aligned against the protein "a" in the DIP. The following thus represents an interaction pair:

$$\mathrm{RH(a)}, \mathrm{RH(b)}$$

Here, the proteins "a" and "b" are reciprocal hits of proteins "A" and "B", respectively. Moreover, "A" and "B" are the identifiers of the interaction pairs found in the pDBs and were used to map the interaction pairs "a" and "b" in our gold standard DIP. The metric about each predicted interaction

pairs was assessed by two distinct manners: using the average metric value and using the smallest metric value, which were, respectively, denoted by the following formulas:

$$avg(ab) = (RH(a) + RH(b))/2$$

$$min(ab) = min(RH(a), RH(b))$$

Moreover, each pDB has its own confidence score that was also evaluated both individually and in combination with the other metrics extracted from the RHs. In addition, we have evaluated the contribution of each pDB to the interaction pair, for which we combined the other metrics with the number of times that the interaction pair was predicted in the pDBs (qt_pDB), giving greater weight to interaction pairs predicted by different pDBs.

1.2.4 Validation and precision prediction. To assess the efficiency of our predictions, in addition to a positive set of interactions, a set of negative interactions is also necessary. Because the DIP database contains only positive interactions, the negative interaction pairs were randomly generated from the DIP protein identifiers through an in-house script at a ratio of five times the number of positive interactions. This negative dataset is composed of protein interaction pairs that are not found in the set of known interactions.[59–61] We created metrics with each value extracted from Blast+, with the pDB score, with the number of databases in which the interaction was predicted (qt_pDB), or by combining these values. These metrics were validated for each pDB both individually and collectively, seeking to identify which metric variation *versus* pDB best represents the set of positive and negative interactions found in our gold standard (DIP). To validate the metrics and their combinations, we used the Receiver Operating Characteristic (ROC) curve plots and calculated the Area Under Curve (AUC) for each metric using the software package ROCR.[62] For metrics with a better AUC value, when seeking to identify a cut-off point that best represented the positive and negative sets of predicted interactions, we tested values from zero to one as cut-off points and compute the sensitivity, specificity and precision by the following formulas:

$$Sensitivity = TP/(TP + FN)$$

$$Specificity = TN/(TN + FP)$$

$$Precision = TP/(TP + FP)$$

The best cut-off point was chosen using the formula

$$Sensitivity \times Specificity$$

because, aside from being easy to implement, its result is equivalent to the Matthews Correlation Coefficient (MCC).[41] The entire method is represented in the ESI,† S2.

2 Results and discussion

2.1 Comparison of predictions based on different numbers of blast alignments

One motivation for this study was the hypothesis that, when only the first hit returned by Blast+ is considered, important results might be disregarded. To test this hypothesis, we performed the analysis using two datasets: one containing only the first Blast+ hit (num_alignments 1) and another containing the first 20 Blast+ hits (num_alignments 20). We compared these two datasets and observed a general 16.95-fold increase in the number of alignments and a 5.10-fold increase in the number of distinct predicted interaction pairs. Proportionally, there was a larger increase in the number of alignments than in the number of interaction pairs. This fact is explained by comparing, especially in the case of the String pDB, the total number of interaction pairs (25 343 169) with the number of distinct interaction pairs (5 382 086), becoming evident the number of repeated interaction pairs (Table 1). When we used 20 Blast+ alignments, it is natural to expect that, if there are homolog proteins among the pDBs, these will be aligned against the same sequence in the DIP, thus mapping the same DIP identifier. Consequently, it reduces the number of distinct DIP interaction pairs identified in relation to the number of Blast+ alignments.

Consideration of first 20 Blast+ alignments generates a large number of repeated interaction pairs. But we were able to increase the number of distinct interaction pairs five times more with an aim to increase >5 times the network coverage for more informative interactions. After a significant increase in the number of distinct interaction pairs generated by Blast+ (num_alignments 20), we investigated the amount of said alignments in relation to the number of hits that Blast+ returned after each run. It was done to identify how much distinctiveness is actually contributed by increasing the parameter num_alignments to 20. From the total 812 907 alignments returned by Blast+ for

Table 1 Quantification of the alignments and interaction pairs comparing 1 and 20 blast hits dataset

	Blast+ output alignment hits			Interaction pairs mapped from the pDBs			
pDB	1 hit	20 hits	Proportion	1 hit	20 hits	20 hits(*)	Proportion(*)
String	44 660	853 234	19.10	1 651 858	25 343 169	5 382 086	3.25
Intact	41 846	450 308	10.76	101 439	5 023 022	3 518 501	34.6
Psibase	9392	322 272	34.31	112	314 280	47951	428.13
Total	95 898	1 625 814	16.95	1 753 409	30 680 471	8 948 538	5.10

1 hit: corresponds to reciprocal hits from Blast+ running with the parameter num_alignments set to 1. 20 hits: corresponds to reciprocal hits from Blast+ running with the parameter num_alignments set to 20. Proportion(*): proportion of the quantity of interaction revealed by Blast+ with num_alignments 20 had over num_alignments 1 (20 hits(*)/1 hit). Hits were counted in both the a → A and a ← A directions. (*) Represents the number of distinct interaction pairs for Blast+ 20 hits.

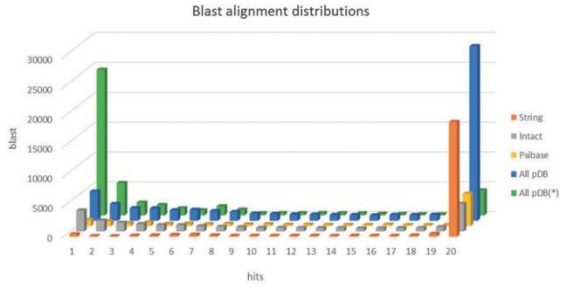


Fig. 1 Distribution of Blast+ alignments grouped by the number of hits. The alignments were generated with the Blast+ parameter num_alignments set to 20. All pDB: is the sum of String, Psibase and Intact. (*) Alignments in which the coverage to identity ratio is above 80%.

the three pDBs, 71.8% had 20 hits, indicating that an even higher cut-off value for num_alignments, may be 30 or 40, could be considered (ESI,† S3). In addition, we investigated the quality of these alignments because better alignments have a greater chance of participating in positive interactions. We then considered only those hits with >80% identity *versus* coverage ratio. Most Blast+ alignments (41.4%) had exactly 20 hits indicating that num_alignments to a value above 20 might return significant alignments too (ESI,† S3). Considering that these Blast+ alignment results are not homologous proteins, which would map identical identifiers in the DIP, they certainly should contribute to the identification of new interaction pairs. Hence, we investigated the number of distinct identifiers mapped to the DIP that would be returned when the Blast+ parameter num_alignments is set to values between 1 and 20. For this analysis, we considered that identifiers found with num_alignments 2 were unique. This was done successively until num_alignments was set to 20, and only the unique identifiers that were not found in identifier sets for num_alignments below 20 were considered (Fig. 1). As expected, most distinct DIP identifiers were found when num_alignments was set to 1 (76.65%) and only 1.4% when num_alignments 20. Of the total 23 680 distinct identifiers present in the DIP, 23 280 were found with the Blast+ parameter num_alignments set to 20, achieving a total identifier coverage of 98%. Comparing the use of num_alignment set to 1 and 20, there was an increase of approximately 23% in the number of distinct identifiers (ESI,† S3). Although it is small, this increase may contribute to increase the number of predicted interacting pairs therefore may increase the network coverage.

2.2 Analysis of interaction pairs

In our gold standard database DIP, there are positive and negative interaction pairs. The positive set consists of experimental interactions curated by the IMEX consortium,[58] whereas the negative set was randomly generated at a proportion of five times the number of positive interactions. In the DIP, all predicted interaction pairs cannot be mapped. Therefore, it is impossible to assess whether these predicted interactions are true or false. To avoid the doubtful inference, we considered only those interaction pairs predicted in the pDBs that were also mapped in the DIP to analyze our metrics. Given the difference in the number

Table 2 AUC values relating to metrics from the dataset created with Blast+ parameter num_alignments set to 1 and the average interaction pair metric value (avg(ab))

AUC Metric	pDB Intact	pDB String	pDB Psibase	All pDB
Score	0.44	0.52	?	0.51
Bitscore	0.44	0.52	?	0.51
Conserved	0.46	0.49	?	0.49
Identity	0.46	0.49	?	0.49
Expected	0.47	0.50	?	0.50
pDB_score	*0.72*	0.70	?	0.69
Combined I	0.58	*0.82*	?	0.80
Combined II	*0.72*	*0.82*	?	*0.90*

All pDB: contains the combined data of Intact, String and Psibase pDBs. The values ? of pDB Psibase column could not be computed. The ROC curves related to the AUC values are detailed in the ESI S4.

Table 3 AUC values relating to metrics from the dataset created with Blast+ parameter num_alignments set to 20 and the average interaction pair metric value (avg(ab))

AUC Metric	pDB Intact	pDB String	pDB Psibase	All pDB
Score	0.83	0.60	0.58	0.68
Bitscore	0.83	0.60	0.58	0.68
Conserved	*0.95*	0.73	0.67	0.80
Identity	*0.96*	0.74	0.68	0.81
Expected	0.88	0.61	0.60	0.71
pDB_score	0.57	0.72	0.50	0.65
Combined I	0.79	0.84	0.50	0.80
Combined II	*0.96*	*0.91*	*0.72*	*0.92*

All pDB: contains the combined data of Intact, String and Psibase pDBs. The ROC curves related to the AUC values are detailed in the ESI S5.

of Blast+ hits when comparing the two datasets generated with num_alignments set to 1 and 20, we studied the pattern of each metric in the interactions generated by each dataset. To do this, we predicted the PPI pairs, generated ROC curves and computed the respective AUC values for both the datasets: num_alignments 1 (Table 2) and num_alignments 20 (Table 3). For both the datasets, we used the metric avg(ab) to compute the six proposed blast values; score, bitscore, conserved, identity, expected and pdb_score, in addition to a combination of two other metrics. For the first dataset, the score, bitscore, conserved, identity and expected blast values displayed a random behavior with an AUC close to 0.50. Therefore, it was not possible to distinguish between positive and negative interactions. In contrast, the pDB_score metric showed considerable improvement for the String (AUC 0.70) and Intact (AUC 0.72) pDBs individually. However, when these pDBs were combined the AUC value became 0.69. We then tested the Combined I metric (pDB score*qt_pDB), which showed considerable improvement for the pDB combination (0.80) and for the String pDB (AUC 0.82), whereas the result was poorer for the Intact pDB (0.58). After observing the behavior of the metrics, we combined the best metric of each individual pDB (pDB score*qt_pDB for String and pDB score*3 for Intact) to compose the Combined II metric. This approach yielded the best result for each pDB individually (AUC of 0.82 for String and 0.72 for Intact) as well as the best result for the combined pDBs (AUC 0.90). We evaluated all metrics for the Psibase pDB in an identical manner, but only

a small number of positive interactions were mapped without a set of negative interactions as required to generate an ROC curve (Table 2). In all ROC curves, "All pDB" corresponds to the union of the data from all the other pDBs which, in theory, would be expected to contain a value close to the average AUC of the individual pDBs. However, in some cases, the AUC value was below the average. This suggested that joining the data from distinct pDBs and assessing them using the same metric will not always improve prediction and that this condition should be carefully tested. We can improve predictions by combining these metrics (Table 2 – Combined I). Still, if the best metrics of each individual pDB are normalized, they may collectively produce better results than if they are individually analyzed (Table 2 – Combined II).

Other combinations of values may generate better metrics for predicting interactions in these datasets (num_alignments 1). Our priority, however, was to perform larger analyses for the dataset generated with the Blast+ parameter num_alignments set to 20 (Table 3). This parameter value is justified by the increased number of predicted interaction pairs, the improvement in the ROC curves and the AUC values together makes this dataset more biologically relevant for analysis. Because it contains more interaction pairs, it was possible to generate the plots for the Psibase pDB, even though the AUC values for this pDB were not good. For the String and Psibase pDBs, the AUC values showed considerable improvement for all metrics. The Conserved and Identity metrics yielded the best AUC values for each individual pDB, especially for Intact, with AUC of 0.95 and 0.96, respectively. The Identity metric was used to compose the Combined II metric, which yielded the best AUC value for this dataset, both for the individual pDBs and for their combination (AUC 0.92 – Table 3). To improve the AUC values obtained with avg(ab) metrics (Table 3), we also computed the min(ab) metrics to the interaction pair (Table 4). The comparison of the plots generated for the ROC curves shows that both the metrics obtained from the average value for the interaction pair (Table 3) and those obtained from the minimum value (Table 4) yielded good results, indicating that, these two metrics are similar in predicting interaction networks. A considerable improvement is observed for the Psibase pDB when the metric is computed using the minimum value of each interaction pair. In both datasets analyzed in this study, the AUC value for the Combined II metric (0.92 – Table 4) obtained by joining all pDBs was very close to that was found in another study,[27] where an AUC equal to 0.94 was obtained (Fig. 2).

By analyzing the pDBs individually, we identified their individual contribution to the composition of the general AUC value of all pDBs. The largest contribution was from the Intact pDB (0.96), followed by the String (0.90) and Psibase pDBs (0.79) (Table 4 – Combined II). Each pDB gave a different AUC for each metric, contributing in different ways to the composition of the general AUC value. Distinct pDB combinations can also contribute differently to prediction, a fact observed when analyzing the ROC curve generated using both the String and Intact pDB. Without the Psibase pDB, the ROC curve yielded a better general AUC (0.93 – Fig. 2 – Combined II).

Table 4 AUC values relating to metrics from the dataset created with Blast+ parameter num_alignments set to 20 and the minimum interaction pair metric value (min(ab))

AUC Metric	pDB Intact	pDB String	pDB Psibase	All pDB
Score	0.88	0.61	0.73	0.71
Bitscore	0.88	0.61	0.73	0.71
Conserved	*0.95*	0.74	0.74	0.80
Identity	*0.96*	0.74	0.77	0.81
Expected	0.89	0.61	0.73	0.71
pDB_score	0.57	0.72	0.50	0.65
Combined I	0.79	0.84	0.50	0.80
Combined II	*0.96*	*0.90*	*0.79*	*0.92*

All pDB: contains the combined data of Intact, String and Psibase pDB. The ROC curves related to the AUC values are detailed in the ESI S6.

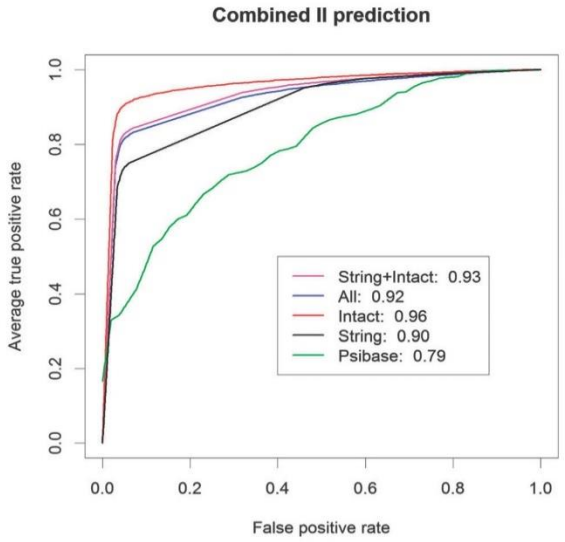


Fig. 2 Combined II ROC curve. ROC curve corresponding to the metrics generated with the Blast+ parameter num_alignments set to 20 and the minimum interaction pair metric value (min(ab)).

Independent of using the average (avg(ab)) or the minimum (min(ab)) value in the metrics, the individual values extracted from Blast+ that were most effective in predicting interaction pairs were Coverage and Identity. When an interaction pair is predicted by more than one pDB, the chances of this interaction being true are higher. We used this premise to improve the ROC curves of the String and Psibase pDBs by giving greater weight to interactions that were predicted in more than one pDB (qt_pDB in Combined II). For the Psibase pDB, this change did not improve the curve; however, it significantly improved for the combination of all pDBs (0.92) and for the String + Intact pDB combination (0.93). Individually, the Intact pDB had the best AUC value (0.96) (Fig. 2 – ESI,† S6).

For the best ROC curves, we assessed several cut-off points to choose the one having the best relationship between sensitivity and specificity. We tested cut-off points for the Combined II metric in relation to the Intact pDB on its own (Fig. 3) and for

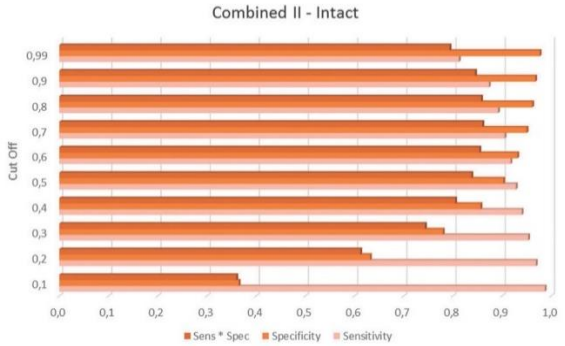


Fig. 3 Sensitivity and specificity analysis for the Combined II metric ROC curve of the Intact pDB.

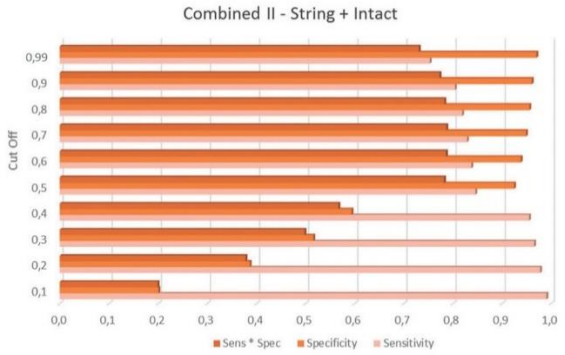


Fig. 4 Sensitivity and specificity analysis for the Combined II metric ROC curve of the String + Intact pDB.

the union of the String and Intact pDBs (Fig. 4). For both the tested sets, the sensitivity and specificity were inversely correlated, which made it difficult to choose the best suited cut-off point. We also tested the sensitivity to specificity ratio, a measure that is equivalent to the Matthews Correlation Coefficient (MCC), which has been used to predict interaction networks.[41] For both the Intact pDB dataset and the String + Intact combination, the best cut-off point of the Combined II metric was at 0.70, representing the highest sensitivity to specificity ratio (Fig. 3 and 4). The cut-off point at 0.70 corresponded to a sensitivity of 0.90 and a specificity of 0.95 for the Intact pDB and to a sensitivity of 0.83 and specificity of 0.95 for the String + Intact pDB (Table 5). This cut-off point was more specific than sensitive, which, in practice, means that less interaction pairs would be selected (0.90–0.83). However, the generated results have a higher probability of being true (0.95).

The Combined II metric consists of the identity and coverage values extracted from Blast+. The cut-off point is a ratio of these two values, *e.g.*, equivalent to a coverage of 0.837 and an identity of 0.837 or a combinations of these values for which the product is 0.70. This cut-off point was higher than those were recommended (0.30 for identity and 0.80 for coverage) to avoid the identification of false positives using the method of homolog interaction mapping.[16] The value corresponding to the score of each pDB itself (pDB score) used in the Combined I metric (Table 4) considerably improved the individual prediction for the String pDB. Thus, the pDB score could be used in combination with other values extracted from Blast+ to further improve the ROC curve of the String pDB individually or together with other pDBs. The use of the pDB score, even if justified by improvements in the ROC curve, would lead us to use different metrics for each pDB in the same ROC curve. Because this practice is not reported in the literature, we adopted a conservative posture and did not add this value for the String pDB. Each pDB sets its own criteria to classify the interactions as true, and as a consequence, the use of different metrics for each pDB may normalize these criteria and improve the prediction of interaction networks when several pDBs are used. In addition to the values extracted from the Blast+ alignments and the pDB score, the way we use the negative interaction set of the gold standard to evaluate metrics can also influence the final results (ESI,† S7 – the negative dataset).

Table 5 Summary of the Roc curve obtained by applying the Combined II metric

Data	AUC	Cut-off	Sensitivity	Specificity	Sens. × spec.	Precision
Intact	0.96	0.70	0.90	0.95	0.86	0.99
String + Intact	0.93	0.70	0.83	0.95	0.79	0.99

The following formulas were used to compute the values in this table: sensitivity = TP/(TP + FN); specificity = TN/(TN + FP); precision: TP/(TP + FP).

2.3 Comparison to similar studies

Several other methods and metrics have been developed and have shown themselves viable when applied to the prediction of interaction networks (Table 6). A comparison of the metrics found in other studies with the one presented herein, considering the different methods, techniques and datasets used by each, has shown our method to be effective: it obtained an AUC of 0.93 for the String + Intact pDB combination and an AUC of 0.96 for the Intact pDB individually. The prediction of interactions using the interolog mapping method was shown to be viable for application, due to both the results presented in this study and the comparison to other studies (Table 6).

Table 6 Comparison of the AUC value of our methodology against other methods

Method	AUC value	Ref.
Structure	Not informed	33
Support Vector Machine (SVM)	0.69	24
Support Vector Machine (SVM)	Not informed	26
Text-mining[a]	0.91	37
Interolog mapping	0.71	28
Mirrortree	0.73	41
Interolog mapping[b]	0.94	27
Interolog mapping[c]	0.96 and 0.93	This study

[a] Organism-specific method that makes predictions only for annotated genes. [b] Using only a single first hit of the Blast[63] program and only 702 interactions as the positive gold standard dataset. [c] Using the first 20 Blast+[53] hits for prediction.

Finally, we used to evaluate our work a data set consisting of 70.630 experimental and cured interactions as the gold standard.[54,58] Considering the different metrics used to measure the efficiency of the prediction methods and the cut-off point of 0.70, we obtained a precision of 0.99 for both metrics, a value higher than the precision of 0.74 obtained with a method based on text mining.[38] In addition, comparing the results from our methodology obtained here with the methodology using Support Vector Machine (SVM) and 1.500 protein interactions, though the specificity (0.98) and precision (0.8) values are approximate in both studies, the sensitivity value (0.15 and 0.28)[26] was much lower than the obtained value in this study (0.83 and 0.90, Table 5). These results, thus, reinforce the efficiency of our metrics and the good ratio between sensitivity and specificity.

3 Conclusions

This is the first study that uses the first 20 Blast+ hits to compare the combinations of values extracted from alignments for the prediction of PPIs using ortholog interaction mapping and, in addition, evaluates these values for each pDB individually and in combination. Based on our observations in this study, we concluded that each pDB contributes differently to the prediction of interactions, and when used in combinations, the results must be carefully analyzed because adding another pDB does not necessarily improve prediction. This study contributes to the scientific community the good AUC values obtained from the pDB Intact (0.96) and pDB Intact + String (0.93). Most importantly, it also contributes to the possibility of increasing the coverage of a predicted interaction network for an organism by using the first 20 Blast+ hits instead of only the single first hit, thus maintaining a decent performance. In addition, despite identifying the metrics that yield good AUC values, we also identified the metrics that are not adequate for predicting PPIs using the interolog-mapping method. The blast values such as the *e*-value, score and bit score are good metrics for indicating the best alignments for one query protein against a group, but they fail to generally differ true and false homology for all query proteins of a group. In this way, it becomes difficult to identify a cut-off point to distinguish true homologous proteins. This phenomenon is explained by the bias that these metrics are due to the size of the subject database (*e*-value) or even due to the length of the amino acid sequence (score and bit score). After all, two small proteins with good alignments receive a lower score than two larger proteins with good alignments. The combination of the coverage and identity metrics was effective to mapping orthologous interactions. It joins in a single metric, both the quality (identity) and quantity (coverage) of an alignment between two proteins. In this case, the database size does not influence these metrics and the percentage values act as normalizers for the protein size. With the results obtained in this study, we intend to use and apply our methodology to predict the *pan-interactome* of fifteen strains of the gram-positive bacterium *Corynebacterium pseudotuberculosis*, a pathogen of great veterinary and economic importance. In addition, we will use the properties of the predicted interaction network to improve the functional annotation of *C. pseudotuberculosis* genes.[7,52] Likewise, we hope that the scientific community will also make use of the *in silico* methodology that we have validated here, to predict the interaction networks of their organisms of interest. The approach we have followed can be reproduced using public-domain computer programs and databases that are freely available.

Author contributions

Conceived and designed the experiments: ELF performed the experiments: ELF analyzed the data: ELF, SSH, RSF, NL, DB wrote the paper: ELF participated in revising the draft: ALL contributed materials/analysis tools: AS, VA.

Acknowledgements

This study was conducted with support from the CENAPAD-MG. Funding: Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES), Conselho Nacional de Pesquisa (CNPq) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig).

References

1 L. Garma, S. Mukherjee, P. Mitra and Y. Zhang, *PLoS One*, 2012, **7**, e38913.
2 A. Flórez, D. Park, J. Bhak, B. C. Kim, A. Kuchinsky, J. Morris, J. Espinosa and C. Muskus, *BMC Bioinf.*, 2010, **11**, 484.
3 R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp and T. Ideker, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 1974–1979.
4 A. L. Barabási and Z. N. Oltvai, *Nat. Rev. Genet.*, 2004, **5**, 101–113.
5 M. W. Gonzalez and M. G. Kann, *PLoS Comput. Biol.*, 2012, **8**, e1002819.
6 N. Wetie, G. Armand, I. Sokolowska, A. G. Woods, U. Roy, J. A. Loo and C. C. Darie, *Proteomics*, 2014, **71**, 205–228.
7 W. Peng, J. Wang, J. Cai, L. Chen, M. Li and F.-X. Wu, *BMC Syst. Biol.*, 2014, **8**, 35.
8 J. De Las Rivas and C. Fontanillo, *Briefings Funct. Genomics*, 2012, **11**, 489–496.
9 J. Wang, M. Li, Y. Deng and Y. Pan, *BMC Genomics*, 2010, **11**, S10.
10 P. Braun and A. C. Gingras, *Proteomics*, 2012, **12**, 1478–1498.
11 X. Zhang, J. Xu and W.-X. Xiao, *PLoS One*, 2013, **8**, e58763.
12 B. Andreopoulos and D. Labudde, Protein Purification and Analysis, *Protein-protein interaction networks*, iConcept Press, DOI: 10.1586/14789450.1.2.239.
13 H. Li, V. Kasam, C. S. Tautermann, D. Seeliger and N. Vaidehi, *J. Chem. Inf. Model.*, 2014, **54**, 1391–1400, DOI: 10.1021/ci400750x.
14 K. Lage, *Biochim. Biophys. Acta, Mol. Basis Dis.*, 2014, DOI: 10.1016/j.bbadis.2014.05.028.
15 J. Luo, Y. Guo, Y. Zhong, D. Ma, W. Li and M. Li, *J. Comput.-Aided Mol. Des.*, 2014, 1–11.

16 A. G. N. Wetie, I. Sokolowska, A. G. Woods, U. Roy, K. Deinhardt and C. C. Darie, *Cell. Mol. Life Sci.*, 2013, 1–24.
17 H. Yu, P. Braun, M. A. Yıldırım, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li and N. Simonis, *Science*, 2008, **322**, 104–110.
18 X. Sun, P. Hong, M. Kulkarni, Y. Kwon and N. Perrimon, Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on, 2012.
19 F. S. Kao, W. Ger, Y. R. Pan, H. C. Yu, R. Q. Hsu and H. M. Chen, *Biotechnol. Bioeng.*, 2012, **109**, 2460–2467.
20 E. D. Harrington, L. J. Jensen and P. Bork, *FEBS Lett.*, 2008, **582**, 1251–1258.
21 R. Mrowka, A. Patzak and H. Herzel, *Genome Res.*, 2001, **11**, 1971–1973.
22 B. Q. Li, K. Y. Feng, L. Chen, T. Huang and Y. D. Cai, *PLoS One*, 2012, **7**, e43927.
23 R. A. Craig and L. Liao, *BMC Bioinf.*, 2007, **8**, 6.
24 L. Li, P. Zhang, T. Zheng, H. Zhang, Z. Jiang and D. Huang, *PLoS One*, 2014, **9**, e91898.
25 S.-W. Zhang, L.-Y. Hao and T.-H. Zhang, *Int. J. Mol. Sci.*, 2014, **15**, 3220–3233.
26 H. Kumar, S. Srivastava and P. Varadwaj, *Int. J. Comput. Biol.*, 2014, **3**, 37–43.
27 A. M. Rezende, E. L. Folador, D. M. Resende and J. C. Ruiz, *PLoS One*, 2012, **7**, e51304.
28 G. Gallone, T. I. Simpson, J. D. Armstrong and A. P. Jarman, *BMC Bioinf.*, 2011, **12**, 289.
29 J. Geisler-Lee, N. O'Toole, R. Ammar, N. J. Provart, A. H. Millar and M. Geisler, *Plant Physiol.*, 2007, **145**, 317–329.
30 W. Zhou, H. Yan, X. Fan and Q. Hao, *Curr. Bioinf.*, 2013, **8**, 3–8.
31 Q. C. Zhang, D. Petrey, J. I. Garzón, L. Deng and B. Honig, *Nucleic Acids Res.*, 2013, **41**, D828–D833.
32 R. A. Jordan, E. L. M. Yasser, D. Dobbs and V. Honavar, *BMC Bioinf.*, 2012, **13**, 41.
33 Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili and T. Hunter, *Nature*, 2012, **490**, 556–560.
34 M. Ohue, Y. Matsuzaki, T. Shimoda, T. Ishida and Y. Akiyama, *BMC Proc.*, 2013, **7**, S6, DOI: 10.1186/1753-6561-7-S7-S6.
35 W. H. Jang, S. H. Jung and D. S. Han, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2012, **9**, 1081–1090.
36 M. Krallinger, F. Leitner, M. Vazquez, D. Salgado, C. Marcelle, M. Tyers, A. Valencia and A. Chatr-Aryamontri, *Database*, 2012, **2012**, bas017, DOI: 10.1093/database/bas017.
37 Z. Xiang, T. Qin, Z. S. Qin and Y. He, *BMC Syst. Biol.*, 2013, **7**, S9.
38 J. Köster, E. Zamir and S. Rahmann, *Integr. Biol.*, 2012, **4**, 805–812.
39 J. Czarnecki and A. J. Shepherd, *Biomedical Literature Mining*, Springer, 2014, pp. 135–145.
40 T. Sato, Y. Yamanishi, M. Kanehisa and H. Toh, *Bioinformatics*, 2005, **21**, 3482–3489.
41 H. Zhou and E. Jakobsson, *PLoS One*, 2013, **8**, e81100.
42 C. Saccà, S. Teso, M. Diligenti and A. Passerini, *BMC Bioinf.*, 2014, **15**, 103.
43 A. Valencia and F. Pazos, *Curr. Opin. Struct. Biol.*, 2002, **12**, 368–373.
44 L. Skrabanek, H. K. Saini, G. D. Bader and A. J. Enright, *Mol. Biotechnol.*, 2008, **38**, 1–17.
45 V. S. Rao, K. Srinivas, G. Sujini and G. Kumar, *Int. J. Proteomics*, 2014, **2014**, 147648, DOI: 10.1155/2014/147648.
46 J. Zahiri, J. Hannon Bozorgmehr and A. Masoudi-Nejad, *Curr. Genomics*, 2013, **14**, 397–414.
47 T. Reguly, A. Breitkreutz, L. Boucher, B. J. Breitkreutz, G. C. Hon, C. L. Myers, A. Parsons, H. Friesen, R. Oughtred and A. Tong, *J. Biol.*, 2006, **5**, 11.
48 J. G. Kim, D. Park, B. C. Kim, S. W. Cho, Y. T. Kim, Y. J. Park, H. J. Cho, H. Park, K. B. Kim and K. O. Yoon, *BMC Bioinf.*, 2008, **9**, 41.
49 R. Häuser, A. Ceol, S. V. Rajagopala, R. Mosca, G. Siszler, N. Wermke, P. Sikorski, F. Schwarz, M. Schick and S. Wuchty, *Mol. Cell. Proteomics*, 2014, **13**, 1318–1329.
50 X. Yu, A. Wallqvist and J. Reifman, *BMC Bioinf.*, 2012, **13**, 79.
51 A. C. F. Lewis, N. S. Jones, M. A. Porter and C. M. Deane, *PLoS Comput. Biol.*, 2012, **8**, e1002645.
52 H. Yu, N. M. Luscombe, H. X. Lu, X. Zhu, Y. Xia, J.-D. J. Han, N. Bertin, S. Chung, M. Vidal and M. Gerstein, *Genome Res.*, 2004, **14**, 1107–1118.
53 C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer and T. Madden, *BMC Bioinf.*, 2009, **10**, 421.
54 I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte and D. Eisenberg, *Nucleic Acids Res.*, 2000, **28**, 289–291.
55 A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork and C. von Mering, *Nucleic Acids Res.*, 2013, **41**, D808–D815.
56 H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff and A. Valencia, *Nucleic Acids Res.*, 2004, **32**, D452–D455.
57 S. Gong, G. Yoon, I. Jang, D. Bolser, P. Dafas, M. Schroeder, H. Choi, Y. Cho, K. Han and S. Lee, *Bioinformatics*, 2005, **21**, 2541–2543.
58 S. Orchard, S. Kerrien, S. Abbani, B. Aranda, J. Bhate, S. Bidwell, A. Bridge, L. Briganti, F. S. Brinkman and G. Cesareni, *Nat. Methods*, 2012, **9**, 345–350.
59 V. Y. Muley and A. Ranjan, *PLoS One*, 2012, **7**, e42057.
60 E. D. Coelho, J. P. Arrais, S. Matos, C. Pereira, N. Rosa, M. J. Correia, M. Barros and J. L. Oliveira, *BMC Systems Biology*, 2014, **8**, 24, DOI: 10.1186/1752-0509-8-24.
61 A. Ben-Hur and W. S. Noble, *BMC Bioinformatics*, 2006, **7**, S2, DOI: 10.1186/1471-2105-7-S1-S2.
62 T. Sing, O. Sander, N. Beerenwinkel and T. Lengauer, *Bioinformatics*, 2005, **21**, 3940–3941.
63 S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, *J. Mol. Biol.*, 1990, **215**, 403–410.

2.1.6 - Supplementary material

Supplementary Material

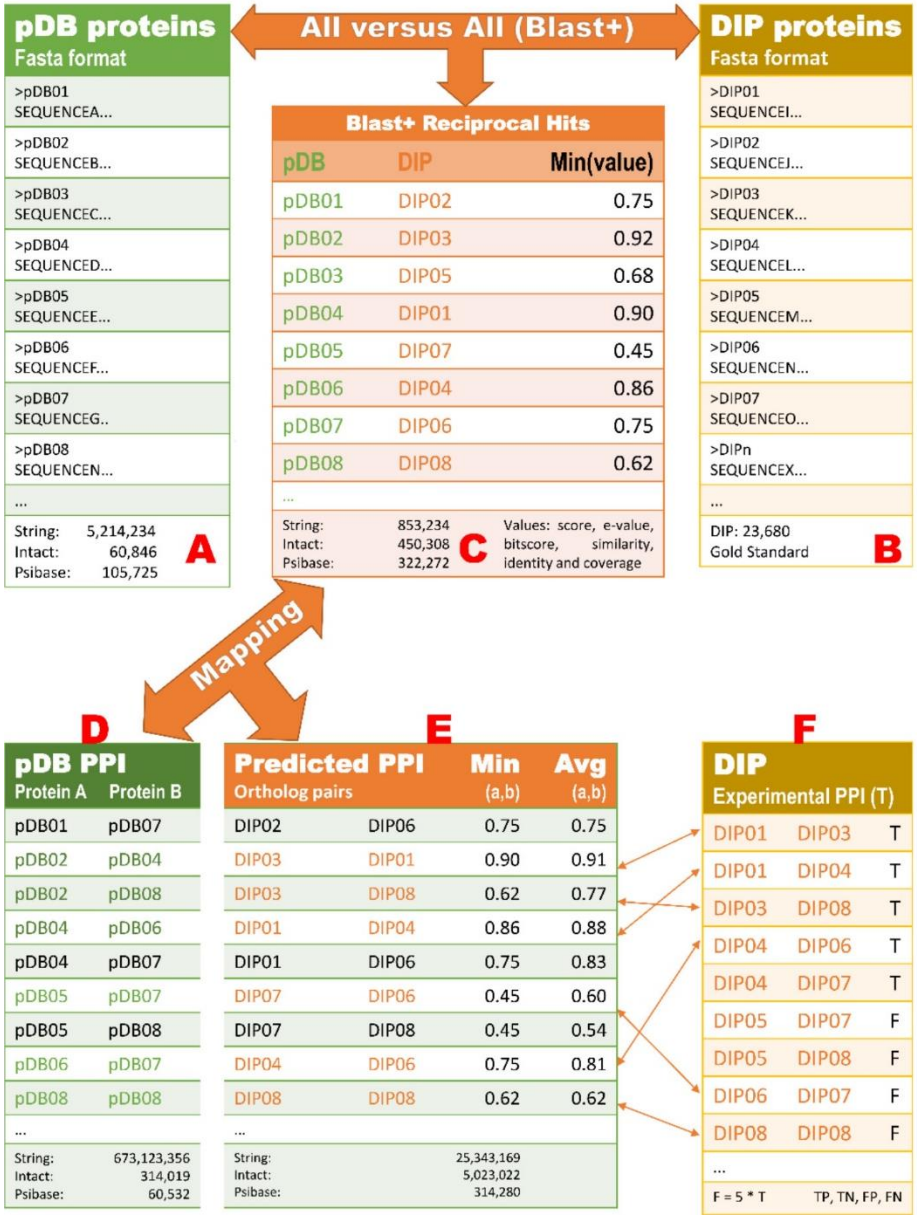
Table S1: Public databases used for predicting and validation of the interactions

Database	Proteins	Interactions
DIP(*)[54]	23,680	70,630
String[55]	5,214,234	673,123,356
Intact[56]	60,846	314,019
Psibase[57]	105,725	60,532

(*) The gold standard data used to validation. Note: Because interactions in the String database are represented both in the A -> B and B -> A directions, there are only 336,561,678 distinct interactions.

Supplementary Material

Figure S2: A) Proteins sequence in fasta format from public databases String, Intact and Psibase. B) Proteins sequence in fasta format from DIP database. C) Blast+ reciprocal hits from alignment of A versus B, and B versus A. Represents the hits generated from Blast+ running with the parameter set to 1 and 20. D) Pairs of PPI (protein-protein interactions) from public databases String, Intact and Psibase. E) Our predicted dataset, generated from the orthologous pairs of PPI mapped from D. The metrics avg(Ab) and min(ab) were used to generated the ROC curves. (C) is used to map each pDB identifier (D) to the ortholog DIP identifier. F) Our gold standard dataset, contain experimental and curated true PPI pairs (T), used to validade the predicted PPI pairs in E. The false PPI set were randomly created, having about five times more pairs than the true PPI set.



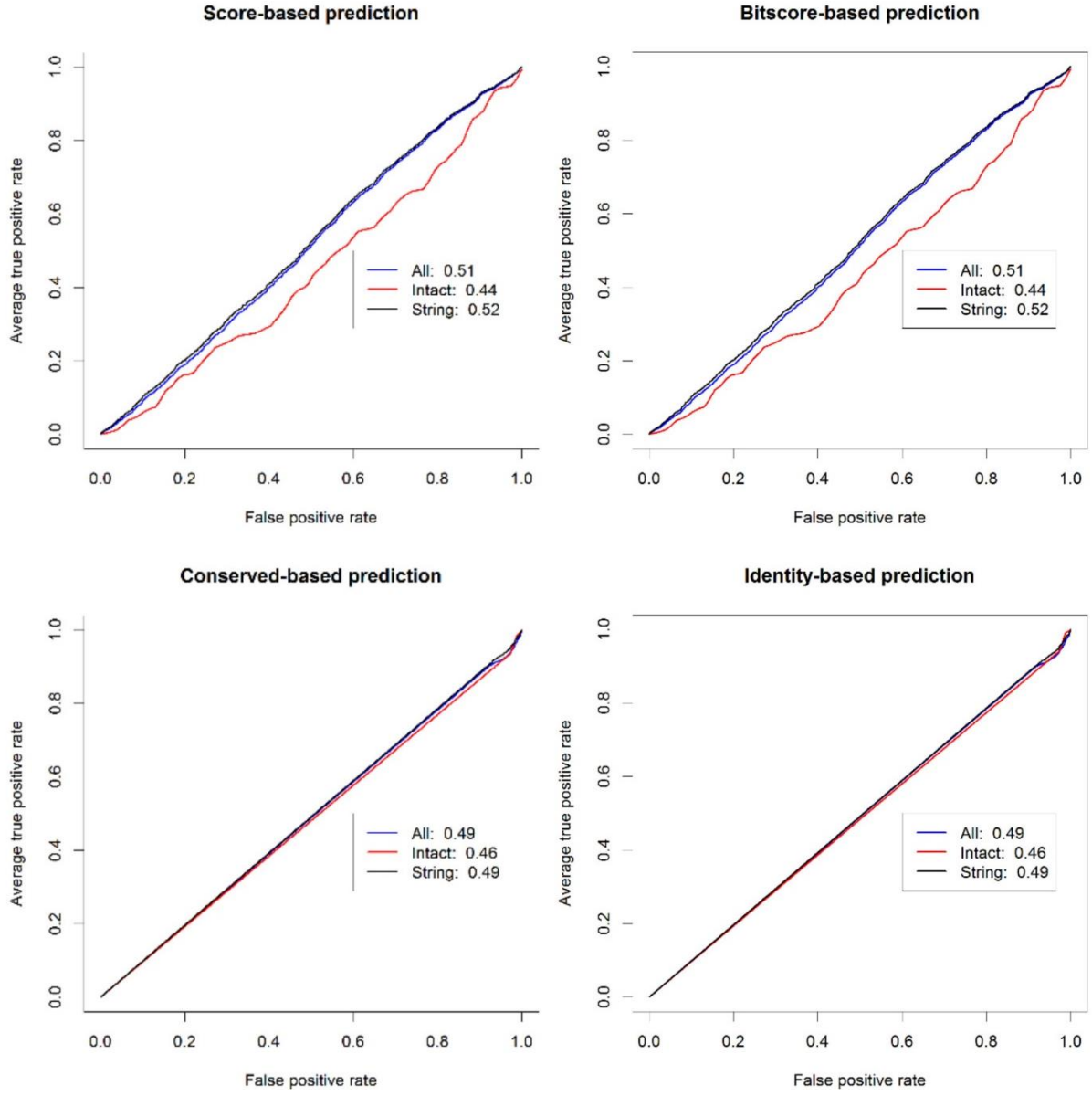
Supplementary Material

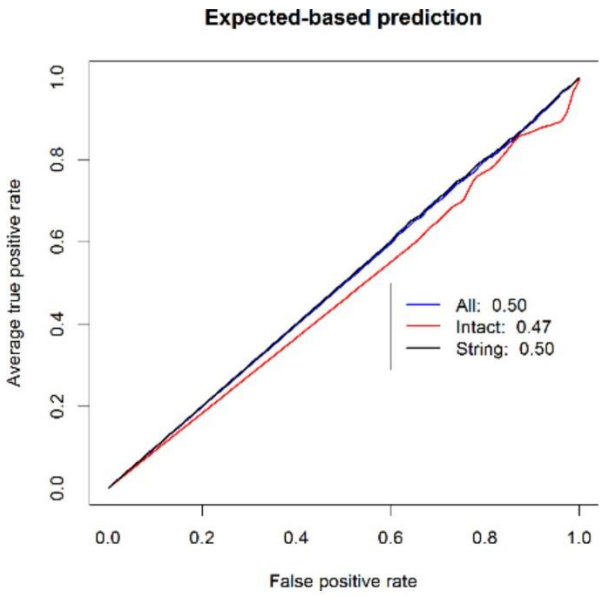
Table S3: Distribution of Blast+ alignments grouped by number of hits. Number of reciprocal hits returned by Blast+ but counted in only one direction. (*) Alignments in which the coverage to identity ratio is above 80%. All: sum of String, Psibase and Intact. Blast Total: equal to All multiplied by Hits, adding up to 812,907 Blast+ hits in one direction or 1,625,814 if reciprocal hits are counted. Contribution: number of distinct identifiers found in alignments with x number of hits, also considering that these identifiers are not found in the sets of identifiers found with x-1 or less hits.

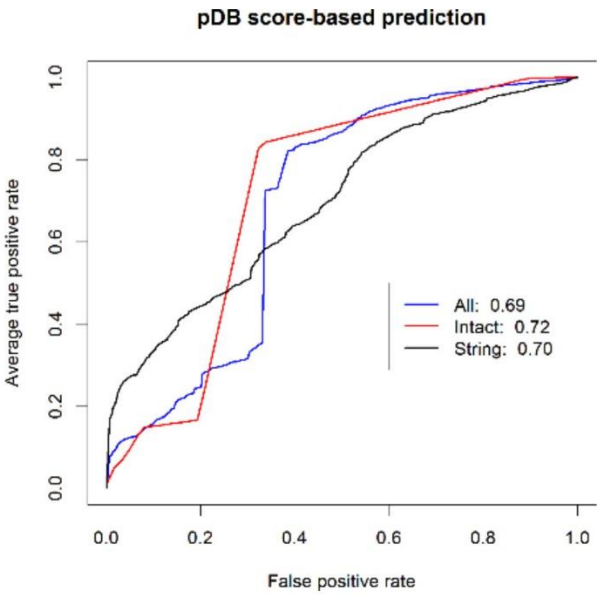
Hits	Blast1	String	Intact	Psibase	All	All(*)	Contribution(*)	Blast Total
20		19,229	4,592	5398	29,219(0.718%)	4,169(0.414%)	326(0.0140%)	584,380
19		486	732	212	1,430(0.033%)	135(0.012%)	14(0.0006%)	27,170
18		228	570	197	995(0.022%)	150(0.013%)	14(0.0006%)	17,910
17		169	585	192	946(0.019%)	166(0.014%)	16(0.0006%)	16,082
16		160	538	273	971(0.019%)	220(0.017%)	16(0.0006%)	15,536
15		175	509	197	881(0.016%)	253(0.018%)	26(0.0011%)	13,215
14		146	574	222	942(0.016%)	333(0.023%)	36(0.0015%)	13,188
13		174	565	223	962(0.015%)	357(0.023%)	42(0.0018%)	12,506
12		161	588	272	1,021(0.015%)	378(0.022%)	60(0.0025%)	12,252
11		168	629	256	1,053(0.014%)	365(0.019%)	44(0.0018%)	11,583
10		175	616	336	1,127(0.013%)	375(0.018%)	38(0.0016%)	11,270
9		163	725	294	1,182(0.013%)	347(0.015%)	73(0.0031%)	10,638
8		244	788	407	1,439(0.014%)	954(0.037%)	540(0.0231%)	11,512
7		365	891	344	1,600(0.013%)	1,501(0.052%)	163(0.0070%)	11,200
6		319	1,068	437	1,824(0.013%)	835(0.024%)	154(0.0066%)	10,944
5		244	1,079	401	1,724(0.010%)	1,167(0.029%)	238(0.0102%)	8,620
4		187	1,250	604	2,041(0.010%)	1,735(0.034%)	439(0.0188%)	8,164
3		99	1,471	522	2,092(0.007%)	2,110(0.031%)	783(0.0336%)	6,276
2		116	1,843	828	2,787(0.006%)	5,418(0.053%)	3,413(0.1036%)	5,574
1	47,949	420	3,474	993	4,887(0.006%)	24,371(0.121%)	17,845(0.7665%)	4,887
				Total hits		200,970	23,280(0.98%)	812,907

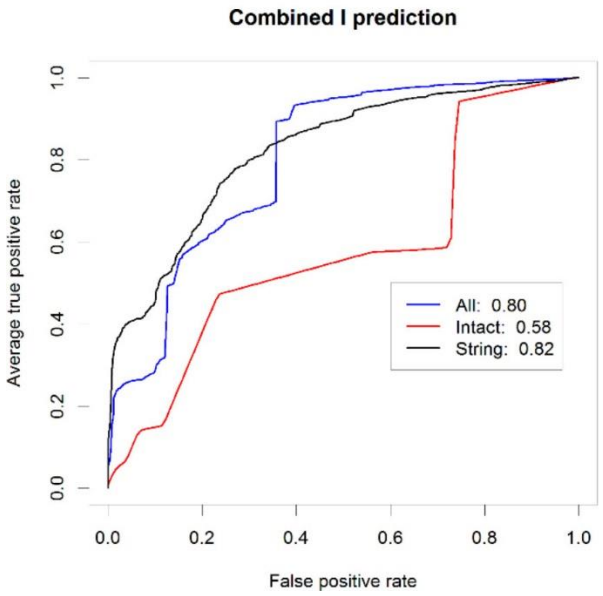
Supplementary Material

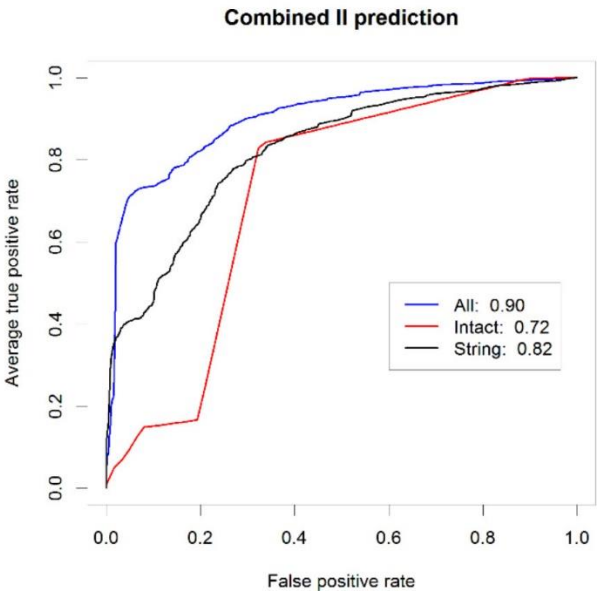
Figure S4: ROC curves corresponding to the metrics generated with the Blast+ parameter num_alignments set to 1. Conserved uses the metric Conserved/100*Coverage/100. Identity uses the metric Identity/100*Coverage/100. Combined I uses the metric pDB score * qt_pDB. Combined II uses the metrics pDB score*qt_pDB for String and pDB score*3 for Intact.




Expected-based prediction
Average true positive rate
False positive rate
All: 0.50
Intact: 0.47
String: 0.50

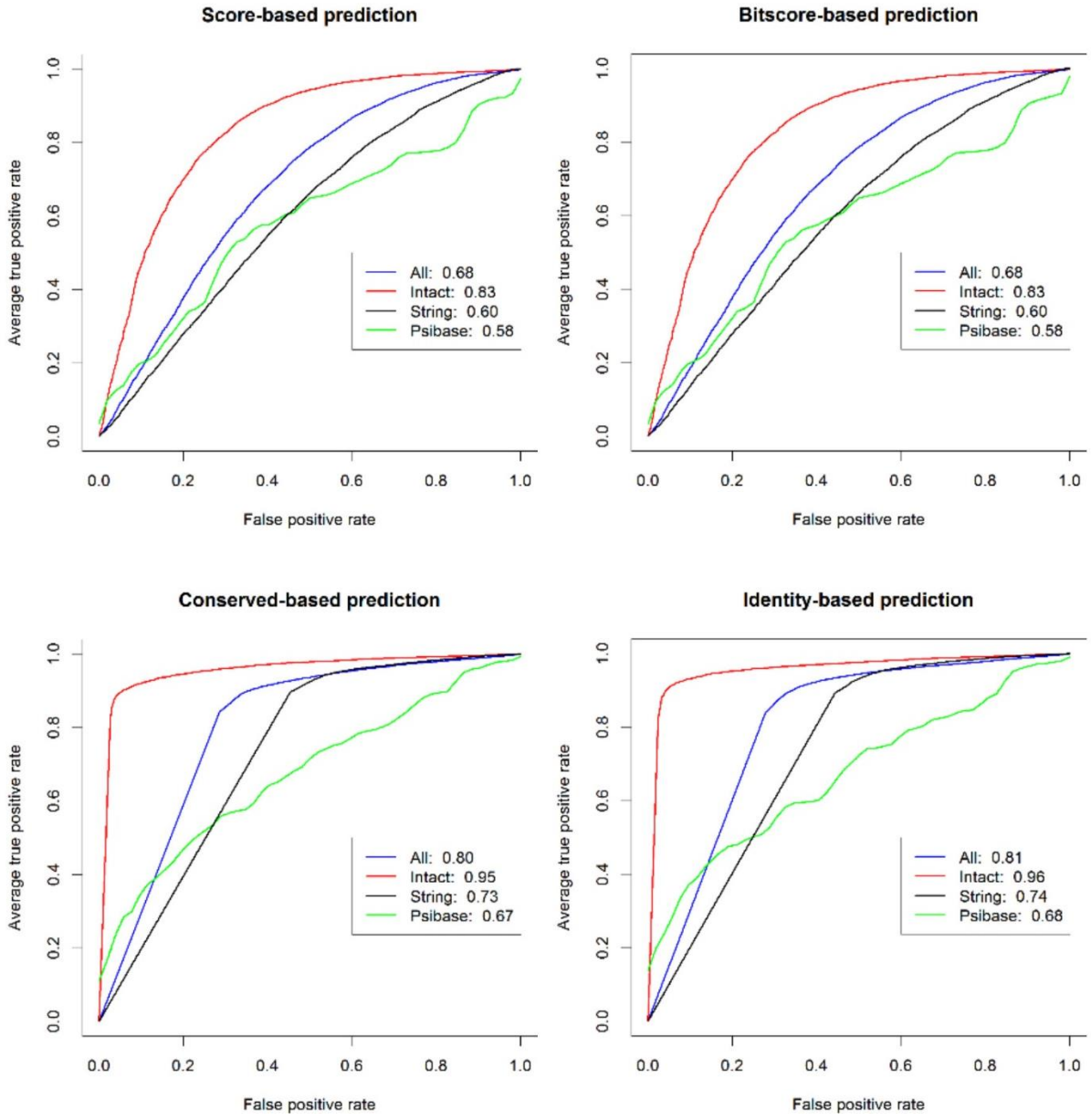

pDB score-based prediction
Average true positive rate
False positive rate
All: 0.69
Intact: 0.72
String: 0.70

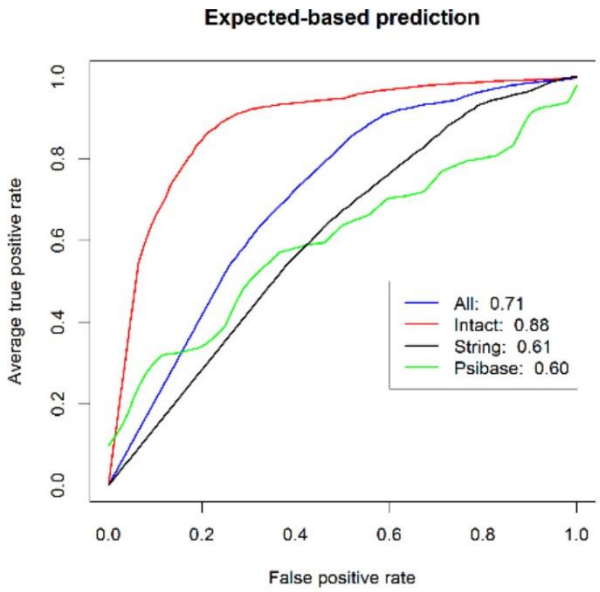

Combined I prediction
Average true positive rate
False positive rate
All: 0.80
Intact: 0.58
String: 0.82

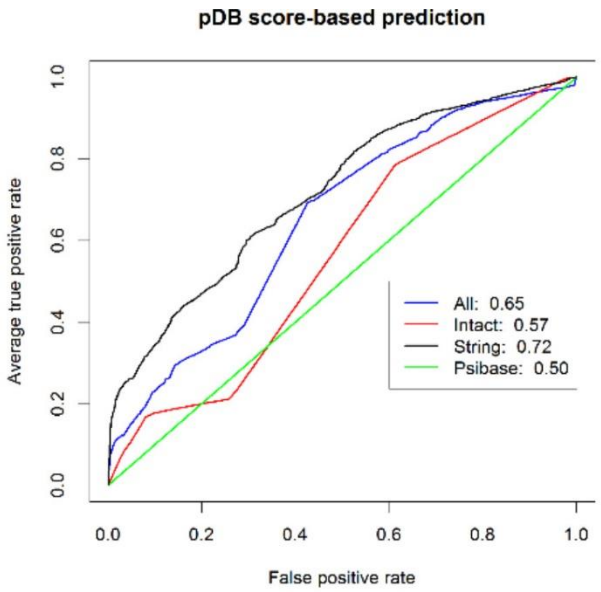

Combined II prediction
Average true positive rate
False positive rate
All: 0.90
Intact: 0.72
String: 0.82

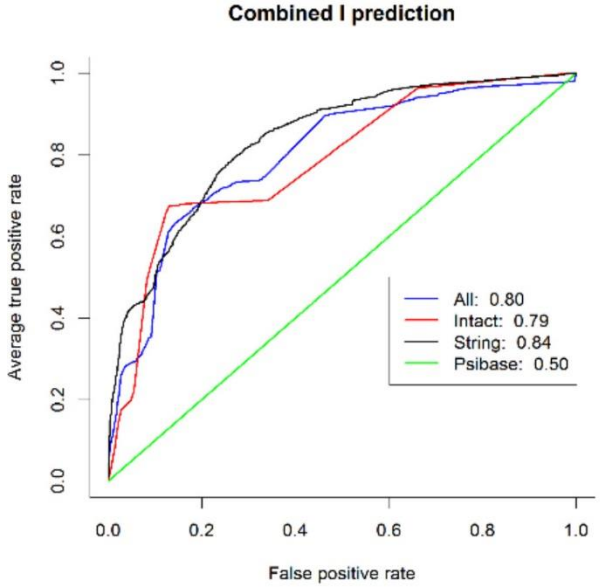
Supplementary Material

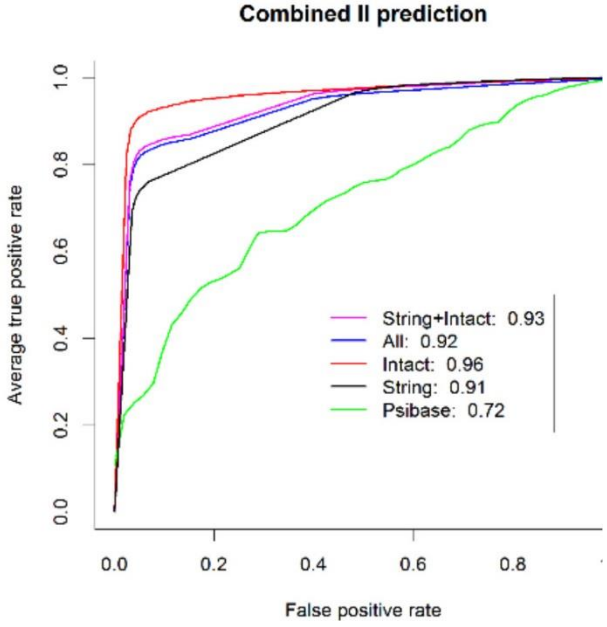
Figure S5: ROC curves corresponding to the metrics generated with the Blast+ parameter num_alignments set to 20 and average interaction pair metric value (avg(ab)). Conserved uses the metric Conserved/100*Coverage/100. Identity uses the metric Identity/100*Coverage/100. Combined I uses the metric pDB score*qt_pDB. Combined II uses the metric pc_identity/100*pc_coverage/100 for the Intact pDB and the metric pc_identity/100*pc_coverage/100*qt_pDB/2 for the String and Psibase pDBs.




Expected-based prediction
Average true positive rate
False positive rate
All: 0.71
Intact: 0.88
String: 0.61
Psibase: 0.60


pDB score-based prediction
Average true positive rate
False positive rate
All: 0.65
Intact: 0.57
String: 0.72
Psibase: 0.50


Combined I prediction
Average true positive rate
False positive rate
All: 0.80
Intact: 0.79
String: 0.84
Psibase: 0.50


Combined II prediction
Average true positive rate
False positive rate
String+Intact: 0.93
All: 0.92
Intact: 0.96
String: 0.91
Psibase: 0.72

3 - Resultados

3.1 - *In silico* protein-protein interaction analysis revels conserved essential proteins in nine *Corynebacterium pseudotuberculosis* biovar *ovis* strains

Edson Luiz Folador, Paulo Vinícius Sanches Daltro de Carvalho, Wanderson Marques Silva, Syed Shah Hassan, Rafaela Salgado Ferreira, Artur Silva, Jan Baumbach, Vasco Azevedo

Tendo uma metodologia com métricas validadas para a predição de redes de interação, a aplicamos na predição de nove redes de interação de nove linhagens do biovar *ovis* de *C. pseudotuberculosis.*

O biovar *ovis* de *C. pseudotuberculosis* é um organismo extremamente clonal (Soares *et al.*, 2013) e todas as redes preditas tiveram características semelhantes, sendo a grande maioria das interações conservadas entre as nove linhagens. As redes foram validadas considerando o menor caminho (*Shortest Path*) (Jeong *et al.*, 2001; Wang *et al.*, 2010; Taylor e Wrana, 2012) e considerando a distribuição do grau de interação (Barabási e Oltvai, 2004). As redes formadas possuem uma topologia livre de escala (*scale-free*) com a distribuição do grau de interação se aproximando a lei do poder (*power law*), demostrando possuirem características de rede biológica. Adicionalmente, comparando as redes de itneração preditas com redes de interação geradas aleatoriamente, os valores de Coeficiente de Clusterização, Correlação e R^2 foram extremamente diferentes. Em tempo, o teste de normalidade Shapiro-Wilk descartou definitivamente que as interações preditas tivessem uma distribuição normal (Shapiro e Wilk, 1965). Todas as validações sugerem que as redes não foram formadas por interações espúrias ou aleatórias, existindo um viés biológico na rede, provavelmente devido a pressão biológica exercida sobre as interações e os clusteres (Galeota *et al.*, 2015).

Este viés biológico é confirmado na análise dos clusteres, cujo apoio na literatura reforça a integridade da rede predita. Dos cinco clusteres analisados todos estavam descritos na literatura, reforçando a consistência das redes preditas e que as interações realmente podem ocorrem em *C. pseudotuberculosis*, sendo um bom exemplo o mecanismo de aquisição de ferro, recentemente revisado e que, com apoio das rede de interação, contribui para melhor entendimento da dinâmica deste mecanismo em *C. pseudotuberculosis* (Sheldon e Heinrichs, 2015).

Finalmente, pela análise do grau de interação das proteínas, foram identificadas 181 proteínas essenciais nas redes de interação de *C. pseudotuberculosis*, sendo que somente a proteína *DNA repair* (RecN) não teve sua essencialidade confirmada na base de dados de genes essenciais (DEG) (Luo *et al.*, 2014). Dentre estas proteínas, 41 não tiveram homologia contra as proteínas do hospedeiro, sendo boas candidatas para propósitos terapêuticos ou diagnóstico. Este fato faz das redes de interação uma valiosa ferramenta para pesquisadores entenderem melhor o mecanismo celular do organismo estudado e identificarem proteínas ou interações como potencial alvo para drogas (Pelay-Gimeno *et al.*, 2015).

O artigo referente a este trabalho será em breve submetido à revista Integrative Biology ou outra revista com similar importância, como para a revista BMC series, cuja avaliação prévia indicou que o artigo pode ser considerado para publicação.

In silico protein-protein interaction analysis reveals conserved essential proteins in nine *Corynebacterium pseudotuberculosis* serovar *ovis* strains

Edson Luiz Folador[1], Paulo Vinícius Sanches Daltro de Carvalho[1], Wanderson Marques Silva[1], Syed Shah Hassan[1], Rafaela Salgado Ferreira[2], Artur Silva[3], Jan Baumbach[4], Michael Gromiha[5], Preetam Ghosh[6], Debmalya Barh[7], Richard Röttger[4], Vasco Azevedo[1,*]

[1]Department of General Biology, Institute of Biological Sciences (ICB), Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil

[2]Department of Biochemistry and Immunology, Federal University of Minas Gerais (UFMG), Belo Horizonte, Brazil

[3]Institute of Biological Sciences, Federal University of Para, Belém, PA, Brazil.

[4]Department for Mathematics and Informatics, University of Southern Denmark, Campusvej 55, Odense, Denmark

[5]Department of Biotechnology, Indian Institute of Technology (IIT) Madras, Tamilnadu, India

[6]Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA

[7]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal, India

3.1.1 - Abstract

The *Corynebacterium pseudotuberculosis* is a gram-positive bacterium that belongs to the CMNR group (*Corynebacterium, Mycobacterium, Nocardia, Rhodococcus*), exhibits characteristics of both *equi* and *ovis* serovars. The serovar *ovis* is the etiological agent of caseous lymphadenitis, a chronic infection affecting sheep and goats, causing economic losses due to carcass condemnation and decrease in the production of meat, wool and milk. The protocols for diagnosis or treatment are not fully effective, requiring further research for a better understanding of *C. pseudotuberculosis* pathogenesis. In this context, the protein-protein interaction network serves as a tool for researchers to get a systemic view of an organism. We mapped the orthologous interactions from public databases to nine strains of *C. pseudotuberculosis*. The validations suggest that the interactions are not spurious and the networks possess the basic characteristics of biological networks. Based on literature support, the clustering analyses further reinforce the biological reliability of the predicted networks. For each strain we predicted on average 16,669 interactions, ~99% of which were mapped from *Corynebacterium* genus, resulting in 15,495 conserved interactions among the nine *C. pseudotuberculosis* strains. Analyzing these networks we identified 181 conserved essential

proteins, of which 41 are non-host homologous and serve as good targets for diagnosis or drug development.

Keywords: Protein-protein interaction, biologic network, system biology, essential proteins, interolog mapping, *Corynebacterium Pseudotuberculosis*, caseous lymphadenitis.

3.1.2 - Introduction

Corynebacterium pseudotuberculosis (Cp) belongs to the supra generic CMNR group (*Corynebacterium, Mycobacterium, Nocardia, Rhodococcus*) of bacteria (Butler, Ahearn e Kilburn, 1986). It is an intracellular pathogen and Gram-positive bacterium that is fimbriated, non-motile and non-capsulated (Selim, 2001) and is present in two serovars: *ovis* and *equi* (Songer *et al.*, 1988). The serovar *equi* infects mainly horses and cattle while the serovar *ovis* is the etiological agent of caseous lymphadenitis (CLA), a chronic infectious disease affecting mainly sheep and goat populations, that can lead to infection in humans associated to occupational exposure (Hémond *et al.*, 2009; Ivanović *et al.*, 2009). Furthermore, CLA disease is prevalent in several countries around the world (Jung *et al.*; Seyffert *et al.*, 2010; Trost *et al.*, 2010; Cerdeira *et al.*, 2011; Ruiz *et al.*, 2011; Silva *et al.*, 2011; Windsor, 2011; Hassan *et al.*, 2012; Lopes *et al.*, 2012; Pethick *et al.*, 2012; Voigt *et al.*, 2012; Colom-Cadena *et al.*, 2014; Hariharan *et al.*, 2014; Mira *et al.*, 2014; Oreiby *et al.*, 2014; Osman *et al.*, 2015) and causes significant economic losses due to low carcass quality, a decrease in the production of meat, wool and milk (Dorella *et al.*, 2006; Baird e Fontaine, 2007), while also causing animal mortality due to suppurative meningoencephalitis (Santarosa *et al.*, 2015). The available methods for CLA diagnosis or treatment are not effective enough, requiring further research to tackle the threats posed by *C. pseudotuberculosis*. Hence, it becomes important to know how the genes, transcripts, proteins and other molecules inside the bacterial cells interact with each other and also with the outer environment to perform their biological functions (Barabási e Oltvai, 2004; Sharan *et al.*, 2005; Flórez *et al.*, 2010; Garma *et al.*, 2012; Gonzalez e Kann, 2012). From this perspective, the study of proteins and their interactions allows for a better understanding of the molecular mechanism of cells at a system level (Wetie *et al.*, 2013; Peng *et al.*, 2014). The protein-protein interactions (PPI) form a complex network represented as a graph, where the nodes represent proteins and undirected edges connecting these nodes represent the interactions between the proteins (Wang *et al.*, 2010; De Las Rivas e Fontanillo, 2012). Computationally analyzed PPI supports developing new hypotheses and designing novel laboratory experiments driven by such hypotheses

(Braun e Gingras, 2012; Zhang, Xu e Xiao, 2013). A PPI network provides a systematic view of the biology of an organism at the cellular level, hence, essential proteins and potential drug targets can be identify by topological analysis (Li *et al.*, 2012; Cui e He, 2014; Li *et al.*, 2014; Mulder *et al.*, 2014; Wetie *et al.*, 2014), enabling the development of new drugs against pathogenic microorganisms (Mosca *et al.*, 2013; Zoraghi e Reiner, 2013; Häuser *et al.*, 2014; Lage, 2014). In this paper, we predict and validate the PPI networks of nine strains of *C. pseudotuberculosis* serovar *ovis* (Cp). Additionally, to better understand the organism and its pathogenicity we perform a cluster analysis and identify the conserved essential proteins in the PPIs, suggesting potential drug or diagnostic targets to be experimentally verified.

3.1.3 – Materials and methods

3.1.3.1 - Data sources

The prediction of the PPI networks is based on the protein sequence similarity and the information of already known PPIs. The protein sequences for the nine Cp were downloaded from NCBI, while known PPIs and their respective protein sequences were retrieved from three publicly available databases (**Table 1**).

Table 1 - Overview of the public data sources.

Data	Proteins	Interactions	Reference
DIP	23,680	70,630	(Xenarios *et al.*, 2000)
String	5,214,234	673,123,356	(Franceschini *et al.*, 2013)
Intact	60,846	314,019	(Hermjakob *et al.*, 2004)
Cp1002	2,090	n/a	(Rezende *et al.*, 2012)
Cp267	2,148	n/a	(Lopes *et al.*, 2012)
Cp3995	2,142	n/a	(Pethick *et al.*, 2012)
Cp4202	2,051	n/a	(Pethick *et al.*, 2012)
CpC231	2,091	n/a	(Ruiz *et al.*, 2011)
Cpfrc41	2,110	n/a	(Trost *et al.*, 2010)
CpI19	2,095	n/a	(Silva *et al.*, 2011)
CpP54B96	2,084	n/a	(Hassan *et al.*, 2012)
CpPAT10	2,079	n/a	(Cerdeira *et al.*, 2011)

Note: The interactions in the String database are represented both in the A -> B and B -> A directions, having 336,561,678 distinct interactions. The Cp proteomes were downloaded from ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/. The interactions for nine Cp strains (n/a) will be predicted in this work.

3.1.3.2 - The Interolog Mapping

The interolog mapping method was used to map the homologous pairs of interacting proteins from public databases to Cp biovar *ovis.* This method was already successfully applied to predict the interactions in organisms such as *Mycobacterium tuberculosis* (Liu *et al.*, 2012),

Leishmania (Rezende *et al.*, 2012) and *Mouse* (Lo *et al.*, 2015). We were already able to show that when using the method, whose previous validation with experimental interactions from DIP database (Xenarios *et al.*, 2000) cured by IMEX (Orchard *et al.*, 2012) consortium, we obtained an Area Under Curve (AUC) of 0.93, a specificity of 0.95, sensitivity exceeding 0.83 and a precision of 0.99, whose detailed flow-diagram was presented in (Folador *et al.*, 2014). The NCBI BLASTp in the latest version was used to perform the reciprocal alignment of proteins from nine Cp strains against the proteins from public databases for which there are known interactions (Camacho *et al.*, 2009). Aiming to eliminate false alignments that would only slow down the prediction process, the BLASTp *e*-value parameter was set to $1e^{-5}$ for proteins from DIP and Intact databases, and set to $1e^{-9}$ for proteins from the String database. All other BLASTp parameters were kept at their default values. To map the homologous proteins we used each of the nine Cp *ovis* proteomes as queries and the proteome of the public databases as subject. In a second step we inverted the search direction, i.e., we switched subject and query. In the remaining, we only consider those proteins alignments that yield a hit in both directions (a reciprocal hit). For each reciprocal hit, we retrieved the minimum identity and coverage values from BLASTp alignment, based on the following formula:

$$RH(a) = min(identity\ x\ coverage(a \rightarrow A),\ identity\ x\ coverage(a \leftarrow A))$$

Here ‘a’ represents a protein of Cp and ‘A’ the homologous counterpart of the known interaction. We assign for each known interaction for which we have homologous proteins in Cp an interaction conservation score. Thus, an interaction pair (IP) is represented by:

$$IP = RH(a), RH(b)$$

Here, the Cp proteins "a" and "b" are reciprocal hits of public databases proteins "A" and "B", respectively. Moreover, "A" and "B" are the public databases identifiers used to map the interaction pairs "a" and "b" to Cp *ovis*. The smallest value of each RH was assessed to compose the interaction score pair (ISP), which is denoted by the following formula:

$$ISP(ab) = min(RH(a), RH(b))$$

The *ISP(ab)* equates to the lower value of identity and coverage identified among the four alignments composing the interaction pair. Aiming to map homologous protein pairs from public databases we considered only interactions with an *ISP(ab)* greater than 0.5625 (corresponds to on average 75% identity and 75% coverage) as conserved. Furthermore, aiming to map high confidence and experimental interaction, we regarded only interactions of

the String database with a confidence score greater than 700. To ensure the accuracy of predictions, we validated the networks both statistically and with literature support.

3.1.3.3 - In silico PPI network validation

Additionally to utilizing our previously reported and validated methodology (Folador *et al.*, 2014), we verify if the nine Cp PPI networks have typical characteristics of biological networks. We submit the PPI networks to Cytoscape plugin NetworkAnalyzer (Assenov *et al.*, 2008) and analyzed the PPI distribution, the node degree distribution (Barabási e Oltvai, 2004) and the Shortest Path (Jeong *et al.*, 2001; Wang *et al.*, 2010; Taylor e Wrana, 2012). Aiming verify if the predicted interactions are spurious, we compared the clustering coefficient, correlation and R-Squared regression values from predicted networks against random networks containing 16,000 interactions for Cp267 lineage.

As an additional validation, in order to check whether the networks have random distribution, the predicted networks were subjected to distribution analysis by the Shapiro-Wilk normality test (Shapiro e Wilk, 1965), available in the statistical R package (Royston, 1982). Finally, the clusters in the predicted networks were identified by using Markov Cluster Algorithm (MCL) (Van Dongen, 2000), implemented in the ClusterMaker (Morris *et al.*, 2011) plug-in available in the Cytoscape (Shannon *et al.*, 2003) software, with MCL inflation value parameter set to 3.0. To reinforce that these interactions do occur in Cp, a literature search was performed to verify the existence of these clusters in phylogenetically close organisms.

3.1.3.4 - Essential proteins

In *Saccharomyces cerevisiae* the degree interaction of nodes was observed to be correlated with the lethality of removing such proteins from the network (Jeong *et al.*, 2001; Estrada, 2006).Large degree and centrality measures are the means for identifying the essential proteins (Betul e Eric, 2013; Tang *et al.*, 2014), explained by the disruption that knockout of one could cause in the interaction network (Han *et al.*, 2004). With the modeled interaction network, we perform topological analysis to identify the Cp essential proteins by selecting the top 15% proteins with high degree interaction, named as hub proteins. Next, to validate the essential hub proteins, we searched for homologous sequences in the bacterial protein sequences from DEG (Zhang, Ou e Zhang, 2004; Luo *et al.*, 2014) (v11.2, updated on July 3, 2015). For the alignment of Cp proteins against DEG, the BLASTp parameters were set to: *e*-value = $1e^{-5}$, low complexity filter = false and matrix = BLOSUM62. Finally, the BLASTp

program was used to align the essential proteins of Cp against the proteins from five hosts: *Ovis aries* (taxid: 9940), *Capra hircus* (taxid: 9925), *Bos Taurus* (taxid: 9913), *Equus caballus* (taixd: 9796) and *Homo sapiens* (taxid: 9606).

3.1.4 - Results and discussion

3.1.4.1 - The C. pseudotuberculosis PPI network prediction

Among the 18,890 proteins present in nine Cp strains, 10,370 participated in interactions, accounting for in total 150,019 predicted interactions (16,669 on average per Cp strain). The contribution of each public database to the formation of networks is shown in (Table 2).

Table 2 - Amount of proteins and interactions for echa serovar *ovis* strain

Linhagem	Proteins	Proteome	Interactions	DIP	Intact	String
Cp1002	1.156	2.090	16.710	103.514	121.035	39.276.922
Cp267	1.164	2.148	16.728	102.140	120.193	39.415.241
Cp3995	1.141	2.142	16.600	100.868	119.895	39.454.010
Cp4202	1.148	2.051	16.712	99.881	118.356	38.973.203
CpC231	1.151	2.091	16.647	95.314	116.142	38.866.646
cpfrc	1.165	2.110	16.897	106.993	126.679	41.393.479
CpI19	1.158	2.095	16.715	96.181	117.188	38.957.265
CpP54B96	1.149	2.084	16.537	95.231	114.476	38.776.672
CpPAT10	1.138	2.079	16.473	94.058	115.149	38.730.691

Proteins: amount of proteins participating in the interaction network for each strain. Proteome: amount of proteins for each strain. Interactions: amount of predicted interactions used for network composition. DIP: amount of interactions mapped from DIP. Intact: mount of interactions mapped from Intact. String: amount of interactions mapped from String.

Despite the large number of interaction pairs predicted from each public database individually, only a small percentage were harnessed to generate the Cp *ovis* interactome. The reduced number of harnessed interaction pairs is due the following three reasons: (i) despite the cut-off point defined for the BLASTp alignments, by having *ISP(ab)* lower than 0.5625, the majority of the interactions were not considered homologous; (ii) in addition, only the interactions with String score >= 700 (Franceschini *et al.*, 2013) were mapped and; (iii) when redundant interactions were found, the one with highest *ISP(ab)* was utilized. The latter condition occurs when the interaction is mapped to more than one public database or mapped multiple times due to the existence of homologous interactions in the same database. Hence, little more than 50% of the total proteins for each Cp strain composed the interaction networks, demonstrating the need for further research to learn about all interactions among the proteins of this organism. Only a small fraction of the interactions were mapped and, considering the predicted interactions came from organisms whose interactions are already

known (interolog mapping), we indirectly realize that we still have a lot to learn about Cp *ovis* and others phylogenetically close organisms until all interactions became known.

The phylogenetically close organisms are the most similar and hence their genotypes and phenotypes probably will also be similar. As this work uses interolog mapping to predict the interactions, we verify from which organism the Cp *ovis* interactions came. The vast majority of interactions were mapped from phylogenetically close organisms and the genus *Corynebacterium* accounted for ~99% of the mappings (Figure 2). This fact reinforce the reliability of the method and the interaction networks generated, after all, being the homologous PPI mapped from phylogenetically close organisms, greatly increases the chances they are realized in Cp.

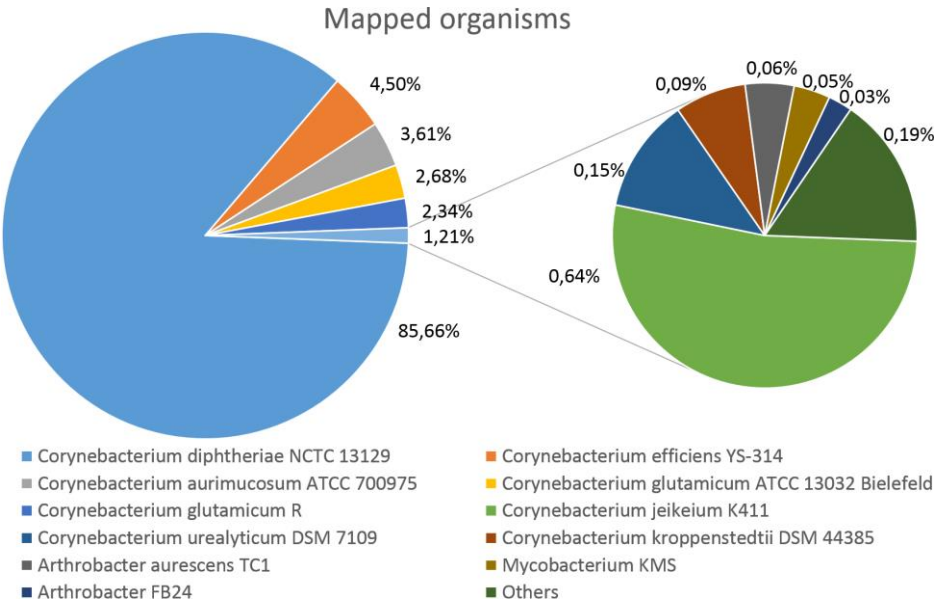


Figure 1 - Organisms from which the interactions were mapped.

Although, such evidences suggest that these interactions really occur in Cp *ovis*, we further perform both statistical and literature-based validation to check the reliability of the predicted interaction networks.

3.1.4.2 - In silico PPI network validation

We were able to show that the node degree distribution follows a power-law and together with shortest-path analysis suggest that the predicted networks have a scale-free distribution, possessing relevant characteristics pertaining to biological networks (Supplementary Material S1). Comparing the clustering coefficient, correlation and regression analysis using the R-

Squared metric from predicted Cp interaction networks, we observed that the values are higher than those obtained from random networks. With p-value < 2.2e-16 the Shapiro-Wilk normality test demonstrated that the predicted interaction networks do not show a normal distribution (Supplementary Material S2). All analyses suggest the networks were not formed by spurious interactions, and may have a biological bias, probably due to evolutionary pressure exerted over the interactions (Shapiro e Wilk, 1965). Moreover, the high Clustering Coefficient of the predicted networks suggest the existence of self-organization inside the biological cell motivated by the interactions (Galeota *et al.*, 2015). The statistical analysis values from the predicted networks are quite close to other works using the same methodology (Rezende *et al.*, 2012). Finally, based on biological literature support, we validate some conserved clusters identified in the networks, showing that the predicted interactions indeed exist in nature and therefore take place in *C. pseudotuberculosis* (Supplementary Material S3).

With the predicted and validated PPI networks, for each Cp strain we also modeled the networks (Supplementary Material S4 to S12). Almost all pairs of the predicted interactions are common to the nine Cp *ovis* strains (core-interactome), a fact which is not surprising since Cp is extremely clonal (Soares *et al.*, 2013). For each Cp *ovis* strain were predicted on average 16,669 interactions. In this work, we focused primarily on validating these interactions with computational methods or through literature support. The strain specific interactions or the accessory interactions are also important and cannot be ignored as they can explain the biology of a specific strain. However, here we focused on exploring the common PPIs for nine Cp *ovis* strains (core-interactome) aiming to better understand the serovar *ovis* instead of only a specific strain. Based on our predicted networks, we identified the conserved essential proteins in the serovar *ovis*.

3.1.4.3 - Essential proteins

The hub proteins are highly interconnected, forming a dense network of interactions, probably participating in various cellular processes and metabolic pathways. Thus, these proteins are termed essential, where the knockout of any one of them can disrupt the interaction network (Han *et al.*, 2004). From the interaction network view point, essentiality is measured by the degree of interaction of a protein (Khuri e Wuchty, 2015). So it is natural to conclude that these essential proteins interact with many other proteins, perhaps exerting various biological activities and participating in several metabolic pathways; thus the inhibition of these proteins

could interrupt their activity in various biological complexes (Han *et al.*, 2004). Laboratory studies are necessary to confirm this hypotheses in Cp because every organism may have a particular and alternative repertoire of proteins to various stress type responses (Caufield *et al.*, 2015).

In order to identify the essential proteins from Cp *ovis* PPI network, we select the top 15% proteins with more interactions, termed hubs, conserved in all nine strains. Thus, we identified 181 hub essential proteins having 68 or more interactions. In the set of essential proteins, we find proteins involved in biological processes related to carbon metabolism, cell envelope and cell wall, DNA metabolism, nucleotides biosynthesis, folding, translocation, ribosomal translation factors, tRNA synthetase, RNA metabolism and respiratory pathways, among others. Aiming to verify the essentiality of these Cp proteins, we searched for homologous proteins in the DEG database. Among the 181 essential proteins, only one had no homology against bacterial DEG proteins, showing the effectiveness of our methods for identifying the essential proteins (Supplementary Material S13). Perhaps fewer essential proteins would be identified in DEG if we used a more restrictive cut-off point, which would reveal more Cp-exclusive list of essential proteins without homologous in DEG.

The DNA repair protein (RecN), was the only Cp essential protein not found in DEG. RecN is responsible for maintaining DNA integrity when exposed to various stress conditions. Despite the conserved mechanism, both metabolic pathways and proteins can differ in each species (Eisen e Hanawalt, 1999). In *E. coli* and *Clostridium difficile*, the LexA repressor interacts with RecA regulating the DNA damage response (Walter *et al.*, 2014); LexA is also reported to regulate RecN (Rostas *et al.*, 1987), keeping the same expression pattern in *Shewanella oneidensis* when submitted to stress (Brown *et al.*, 2006). All these interactions are also found in the *C. pseudotuberculosis* PPI network, wherein the interactions between LexA and RecN in the biovar *ovis* interact with proteins encoded by the following genes: *recA, recO, recR, recF* and *recG* are too conserved (Figure 2). This suggests an important role for both RecN and LexA proteins. Using RNA-Seq data, we verified that RecN and LexA had no significant change in their expression, thereby indicating a constitutive expression in conditions of thermal shock, acid and osmotic stress (Pinto *et al.*, 2014), which is an expected characteristic for essential genes.

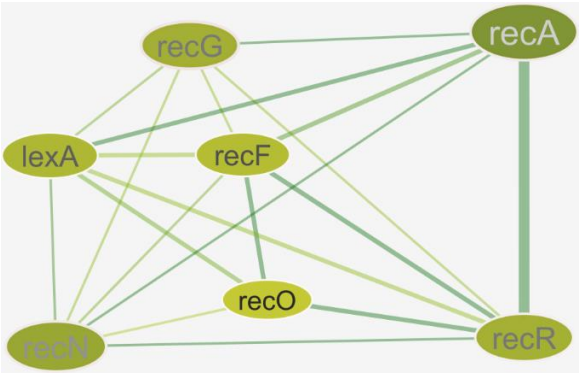


Figure 2 - Partial *C. pseudotuberculosis* DNA repair RecN interactions network.

The vast majority of proteins have homologous proteins in DEG however, this does not reduce the importance of describing their essentiality. Considering Cp is not covered by DEG till date, the description of essentiality in this organism is novel for all 181 proteins. However, while most essential proteins have homologs from over 20 organisms, three proteins have homologs in a single organism covered by DEG, showing either the lack of experiments which would support their essentiality, the lack of protein conservation across species or that the essentiality of these proteins is not conserved across species (Caufield *et al.*, 2015). These proteins are Catalase (KatA), Endonuclease III (Nth) and Trigger factor Tig (Tig). KatA has DEG homology against KatE from *Salmonella enterica*. KatA is an oxidoreductase enzyme which decomposes hydrogen peroxide (H_2O_2) at a rate of 40 million molecules per second (Nelson e Cox, 2002). In *C. glutamicum*, levels of KatA are increased quickly in response to the H_2O_2 addition (Milse *et al.*, 2014) and, was highly up-regulated for the SOS and stress response (Park *et al.*, 2014); the same occurring in *C. pseudotuberculosis* when exposed to acid medium (Pinto *et al.*, 2014). Due to the fast response to oxidative stress, KatA is an important survival mechanism in host macrophages, and therefore may have biotechnological or pharmaceutical applications (Cutler, 2005; Mitra, 2014). Endonuclease III (Nth) has DEG homology against *Haemophilus influenzae*. Nth is a base excision repair enzyme (Sahbani *et al.*, 2014) that participates in a pathway to prevent the loss of DNA functionality e.g., by spontaneous mutagenic lesion (Saito *et al.*, 1997) or near-UV radiations (Serafini e Schellhorn, 1999). This mechanism was well studied and is conserved in the *Corynebacterium* species (Resende *et al.*, 2011). Trigger factor Tig (Tig) has DEG homology against *Pseudomonas aeruginosa*. Tig participates in the protein folding process. In *Escherichia coli*, Tig cooperates with Chaperone protein DnaK to promote protein folding, however, is not essential for intermediate growth temperatures (Deuerling *et al.*, 1999). In *Exiguobacterium*

antarcticum, a gram-positive psychrotrophic bacteria, only Tig was overexpressed in response to cold; the remaining chaperone proteins were underexpressed at 0°C (Dall *et al.*, 2014). For *C. pseudotuberculosis* at 50°C, no significant change was observed in Tig expression, where the same also occurs with the Chaperonins GroEL, however DnaK was overexpressed (Pinto *et al.*, 2014). It would be necessary to submit *C. pseudotuberculosis* to lower temperature to check the behavior of Tig.

Additionally, in order to identify potential biomarkers or therapeutic targets among the essential proteins, a search for homologous proteins in the host organisms *O. aries, C. hircus, B. taurus, E. caballus and H. sapiens* was performed. Considering the Blastp alignment results (Supplementary Material S14), we identified 41 non-host homologous proteins, 24 having no alignment hit against *O. aries* and *C. hircus* proteins and 17 having both low identity (0-38%) and low coverage (0-44%) (Figure 3). Alignment details against hosts can be observed in Supplementary Material S15.

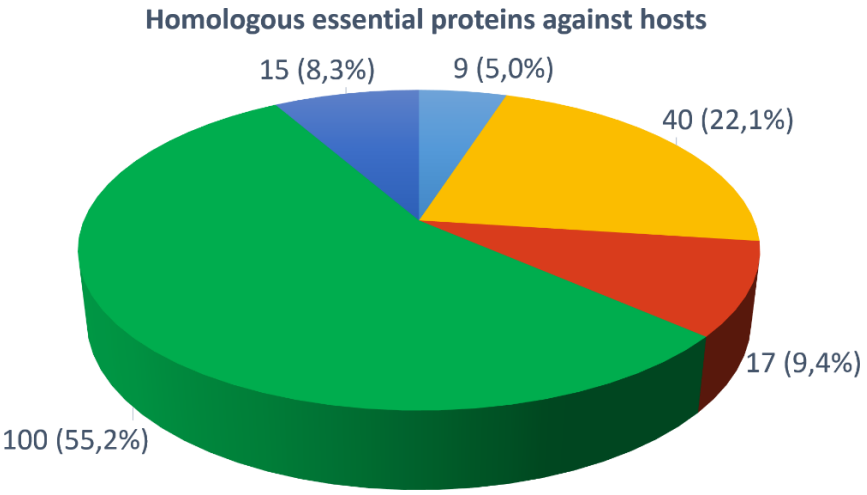


Figure 3 - Homology distribution of Cp essential proteins aligned against hosts.
Dark green: proteins homologous to host; Yellow: Proteins with low identity against hosts (identity < 30%). Dark red: non-host homologous proteins, proteins with low identity and low coverage alignment against hosts (identity x coverage <= 10%). Dark blue: non-host homologous proteins, proteins with no alignment hits against *O. aires* and *C. hircus*. Light blue: non-host homologous proteins, proteins with no alignment hits against the five hosts. The alignment details can observed in Supplementary Material S15.

The 24 non-host homologous proteins without hits against hosts are: chorismate synthase (*aroC*), dihydrodipicolinate reductase (*dapB*), DNA primase (*dnaG*), elongation factor P (*efp*), cell division protein (*ftsZ*), ATP phosphoribosyl transferase (*hisG*), dihydroxy-acid dehydratase (*ilvD*), aspartate kinase (*lysC*), UDP-N-acetylglucosamine (*murA*), transcription anti-termination protein (*nusG*), uridylate kinase (*pyrH*), DNA repair protein (*recN*),

transcription termination factor (*rho*), 50S ribosomal protein L1 (*rplA*), 50S ribosomal protein L10 (*rplJ*), 50S ribosomal protein L31 (*rpmE*), DNA-directed RNA polymerase subunit alpha (*rpoA*), 30S ribosomal protein S3 (*rpsC*), 30S ribosomal protein S6 (*rpsF*), 30S ribosomal protein S13 (*rpsM*), holliday junction DNA helicase subunit (*ruvA*), SsrA-binding protein/SmpB superfamily (*smpB*), indole-3-glycerol phosphate synthase (*trpC2*) and anthranilate synthase (*trpE*). These 41 (24+17) non-host homologous essential proteins of Cp are good choices for therapeutic and diagnostic propose, not only by the disruption which may cause in the intra-species interactions but also by having greater potential to participate in inter-species interactions with host (Zhou *et al.*, 2014). From the set of non-host homologous essential proteins, two classes draw special attention, both participating in the beginning of aromatic amino acids metabolic pathways, well characterized in *Corynebacterium glutamicum* (Ikeda, 2006), the proteins encoded by the *trp* operon, involved in tryptophan biosynthesis, and the protein prephenate dehydratase (*pheA*).

The cluster analysis draws attention to the Cp iron acquisition system, which is a well characterized system contributing to the survival and virulence of microorganisms (Köster, 2001; Kunkle e Schmitt, 2005). The Cp cluster presents the interaction among proteins of multiple iron acquisition systems, a strategy to acquire iron from different sources or in low availability (Wandersman e Delepelaire, 2004), suggesting both, alternative metabolic pathways and alternative proteins from different operons exerting the same function. In Cp networks, these multiple systems interact and consist mainly of proteins from operon *fag, ciu, fec* and *hmu* (Supplementary Material S3).

The use of interaction networks for identifying essential proteins can have a better sensitivity than other approaches. While we identified 181 essential proteins, of which 41 were non-host homologous, approaches using three-dimensional structures identify less than 10 essential protein units (Hassan *et al.*, 2014). Besides the essential proteins, the identified interactions are equally important in Cp as it allows to search for small molecules inhibitors of binding interactions (Mora e Donaldson, 2012; Zoraghi e Reiner, 2013; Villoutreix *et al.*, 2014), making feasible modern drug discovery research (Sheng *et al.*, 2015). Such interaction network can also be used with RNA-Seq or proteomics experiments to assist in data interpretation. As an example of a biological application, the PPI network from *C. pseudotuberculosis* 1002 strain was used to investigate the interactions among the proteins identified as exclusive and differentially regulated in cells exposed to nitrosative stress (Silva *et al.*, 2014). The results obtained in this work might serve as a basis for further essentiality

studies in other organisms by using the interaction network. By knowing the interaction partners of a protein, it is hence possible to provide a systemic view of the organism (Anh *et al.*, 2015).

3.1.5 - Conclusions

Here, for the first time we reported the PPI networks for nine Cp *ovis* strains and the biological relevance of the essential proteins identified in the networks. In addition to the validated networks, our contributions include the identification of 181 Cp essential proteins, 41 of them being non-host homologous, hence becoming good candidates for drug development or CLA diagnosis (Supplementary Material S13-S15). Since the essential proteins (hubs) interact with many others, it is natural to assume they associate differentially in various biological processes, in their own species well as the host, thereby participating in the formation of different clusters with other proteins to perform their functions, and hence are attractive targets for therapeutic and diagnostic propose. Similarly for the essential proteins, each specific interaction is a potential candidate to be subjected to identification of inhibitors (Villoutreix *et al.*, 2014; Gowthaman, Lyskov e Karanicolas, 2015), thus opening several drug development opportunities about *C. pseudotuberculosis*. The PPI networks reported here are valuable tools for researchers to identify proteins or interactions as potential targets that may have a better sensibility than other approaches. The experimental validation for the predicted interactome is out of the scope of this study but is, vital and will be carried out in the near future.

3.1.6 - Author Contributions

Conceived and designed the experiments: ELF. Designed and modeled the database in PostgreSQL DBMS: ELF. Developed routines in PL/PgSQL: ELF. Performed the experiments: ELF. Analyzed the data: ELF. Structured the paper: ELF, MG. Wrote the paper: ELF. Performed the clusters description: PVSDC, WMS. Performed the essential protein description ELF, Participated in revising the draft: ALL. Contributed materials/analysis tools/structure: JB, MG, RR, RSF, AS and VA.

3.1.7 - Funding

Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES), Conselho Nacional de Pesquisa (CNPq) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (Fapemig).

3.1.8 – Supplementary Material

3.1.8.1 – Shortest path and Degree distribution analysis.

Supplementary Pictures S1: Shortest path and Degree distribution analysis.

Shortest Path analysis of the nine *Corynebacterium pseudotuberculosis* serovar *ovis* strains (Figure 1-9). Degree distribution analysis of the nine *C. pseudotuberculosis* serovar *ovis* strains. The red line indicate the perfect power-law distribution (Figure 10-18).

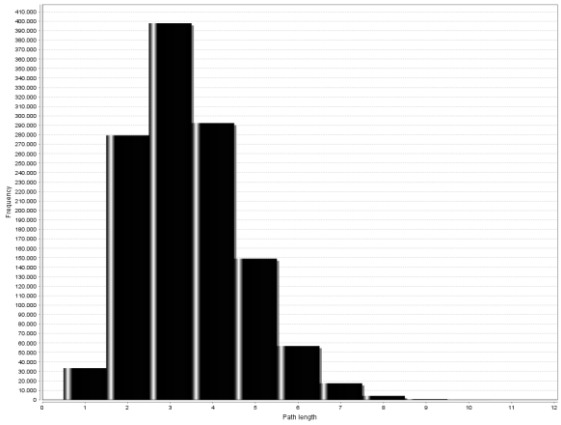


Figure 4 - Cp1002 Shortest Path analysis

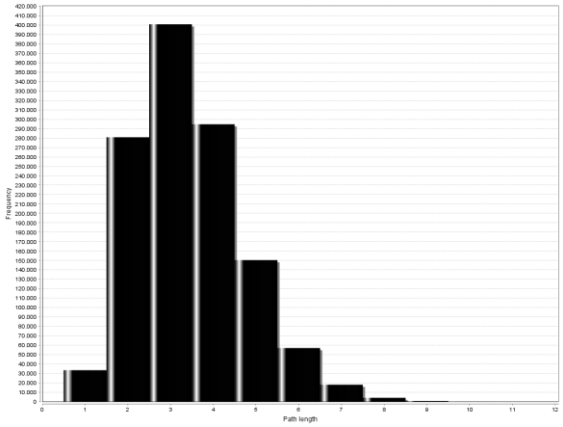


Figure 5 - Cp267 Shortest Path analysis

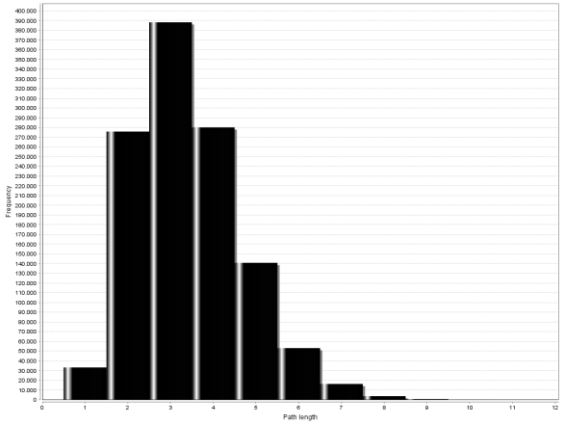


Figure 6 - Cp3995 Shortest Path analysis

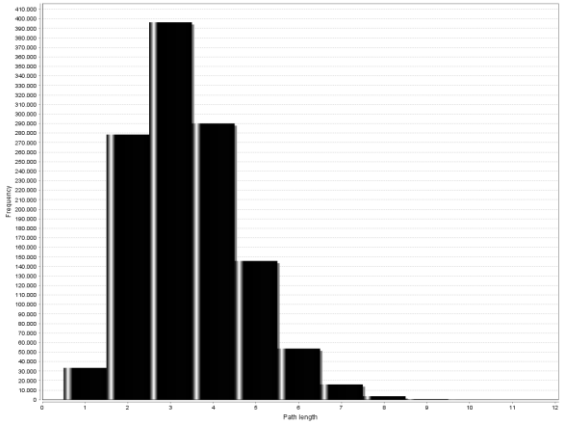


Figure 7 - Cp4202 Shortest Path analysis

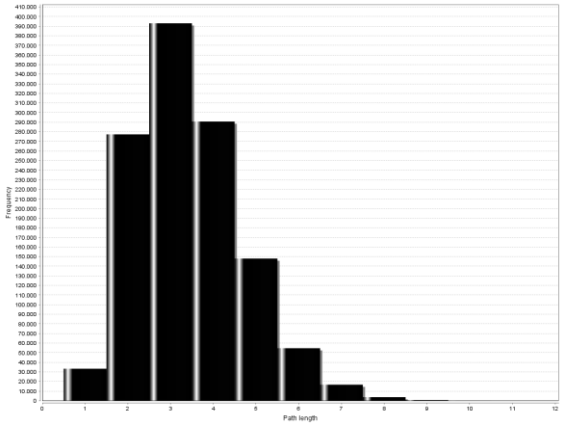


Figure 8 - CpC231 Shortest Path analysis

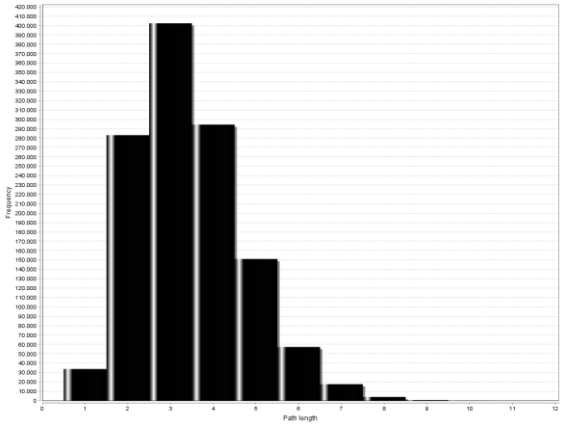


Figure 9 - Cpfrc Shortest Path analysis

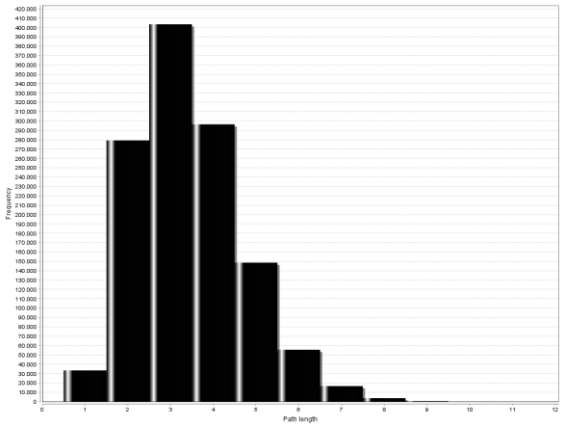


Figure 10 - CpI19 Shortest Path analysis

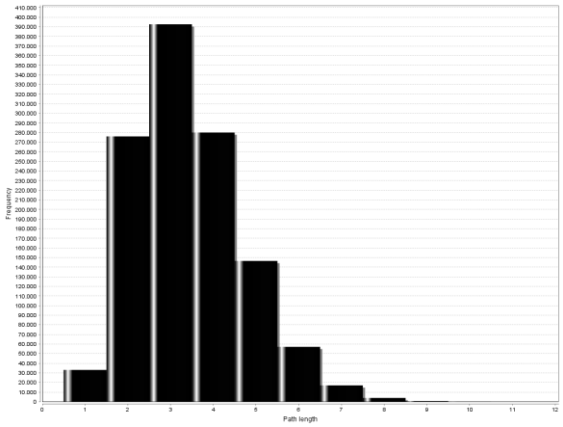


Figure 11 - CpP54B96 Shortest Path analysis

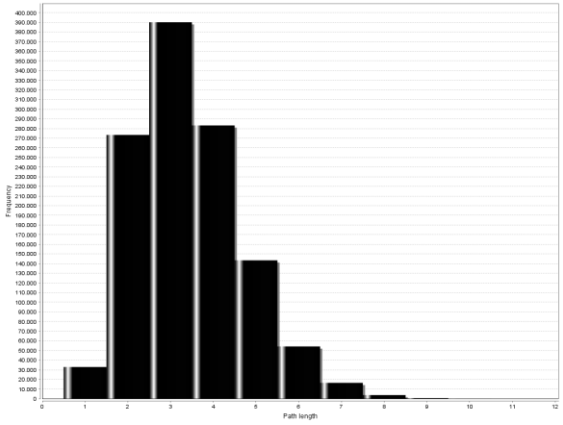


Figure 12 - CpPAT10 Shortest Path analysis

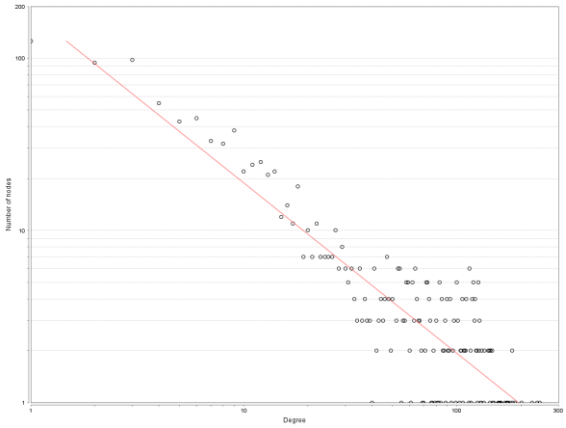


Figure 13 - CpPAT10 Degree distribution analysis. Clustering coefficient = 0.407, Correlation = 0.938, R-Squared = 0.790, Shapiro-Wilk test = p-value < 2.2e-16.

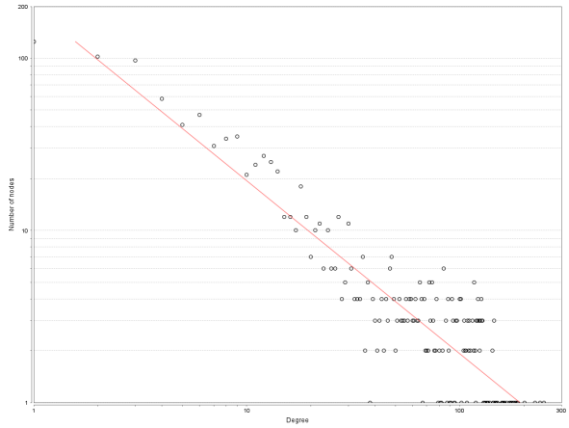


Figure 14 - Cp1002 Degree distribution analysis. Clustering coefficient = 0.408, Correlation = 0.933, R-Squared = 0.822, Shapiro-Wilk test = p-value < 2.2e-16.

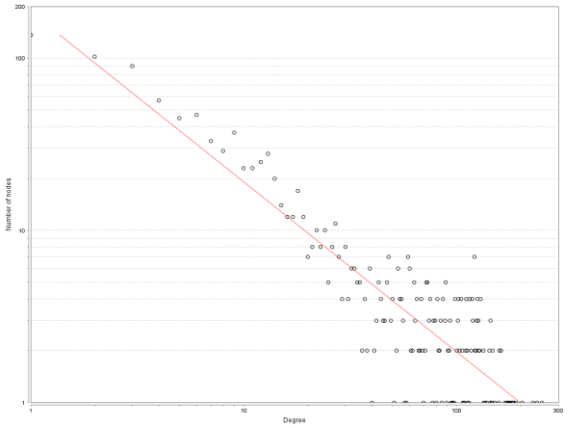


Figure 15 - Cp267 Degree distribution analysis. Clustering coefficient = 0.402, Correlation = 0.953, R-Squared = 0.785, Shapiro-Wilk test = p-value < 2.2e-16.

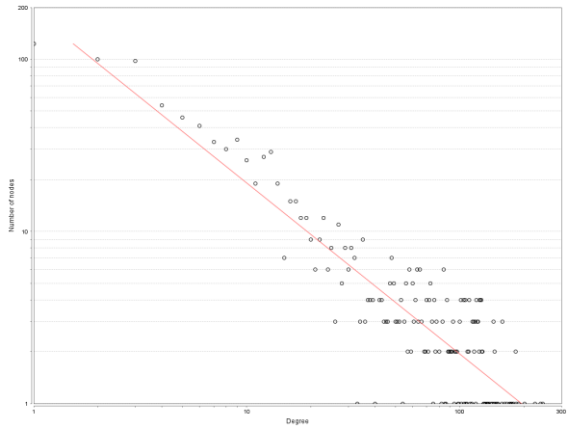


Figure 16 - Cp3995 Degree distribution analysis. Clustering coefficient = 0.410, Correlation = 0.933, R-Squared = 0.798, Shapiro-Wilk test = p-value < 2.2e-16.

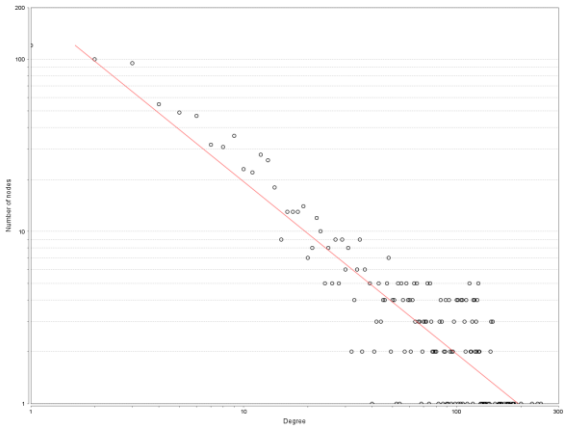


Figure 17 - Cp4202 Degree distribution analysis. Clustering coefficient = 0.410, Correlation = 0.928, R-Squared = 0.799, Shapiro-Wilk test = p-value < 2.2e-16.

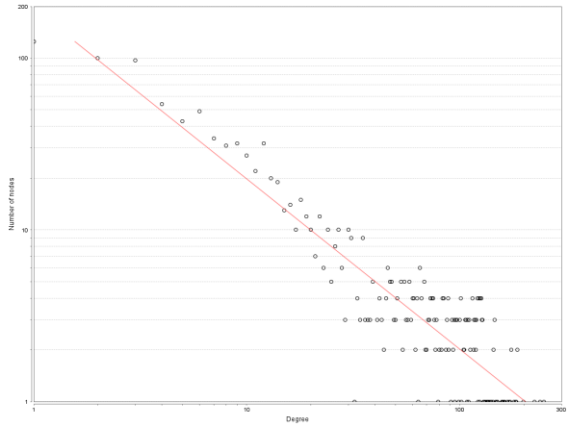


Figure 18 - CpC231 Degree distribution analysis. Clustering coefficient = 0.407, Correlation = 0.936, R-Squared = 0.825, Shapiro-Wilk test = p-value < 2.2e-16.

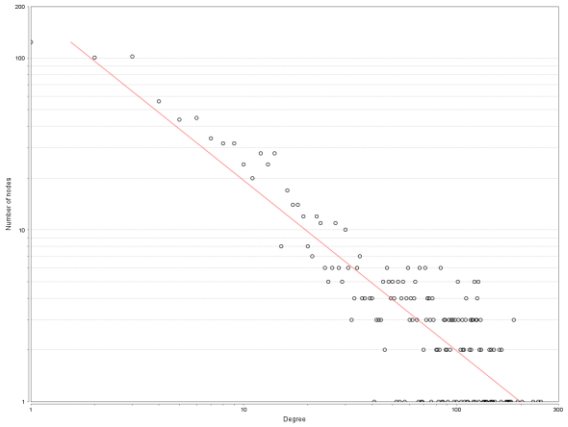


Figure 19 - Cpfrc Degree distribution analysis. Clustering coefficient = 0.408, Correlation = 0.930, R-Squared = 0.786, Shapiro-Wilk test = p-value < 2.2e-16.

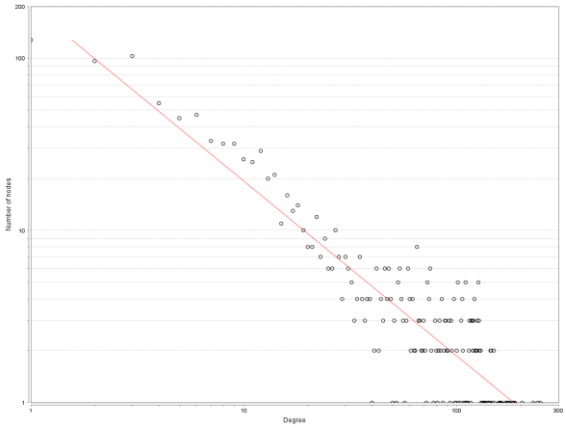


Figure 20 - CpI19 Degree distribution analysis. Clustering coefficient = 0.403, Correlation = 0.932, R-Squared = 0.813, Shapiro-Wilk test = p-value < 2.2e-16.

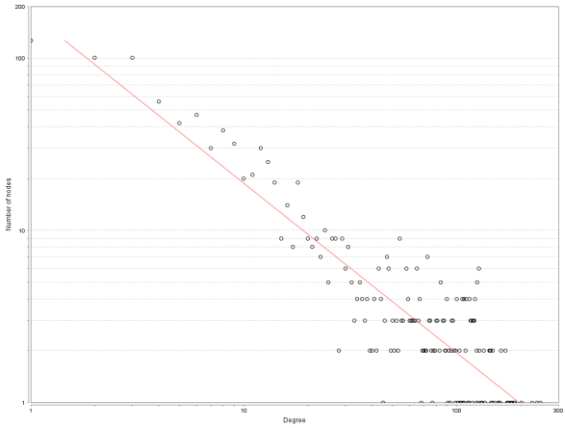


Figure 21 - CpP54B96 Degree distribution analysis. Clustering coefficient = 0.404, Correlation = 0.935, R-Squared = 0.800, Shapiro-Wilk test = p-value < 2.2e-16.

3.1.8.2 – In silico PPI network validation

Supplementary Pictures S2: In silico PPI network validation. Degree distribution analysis of nine interaction networks formed from 16,000 pairs of interactions randomly selected among all possible distinct interactions of *Corynebacterium pseudotuberculosis* Cp267 strain. The pairs distribution was analyzed by the plugin NetworkAnalyzer (Assenov *et al.*, 2008). The red line indicate the perfect power law distribution (Barabási e Oltvai, 2004). All random networks had a normal distribution and a clustering coefficient of 0.007.

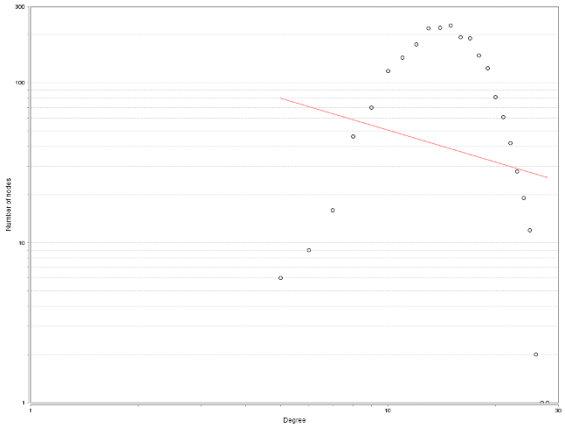


Figure 22 – Random interaction network 01. Correlation = -0.064, R-squared = 0.038

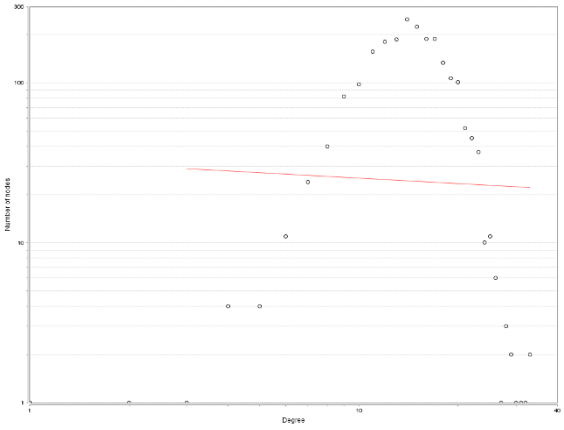


Figure 23 - Random interaction network 02. Correlation = -0.015, R-squared = 0.001

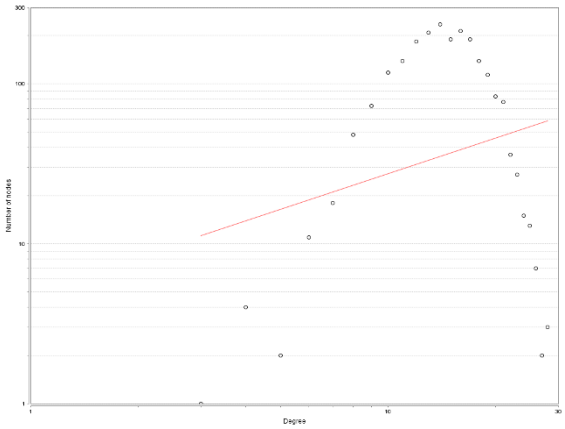


Figure 24 - Random interaction network 03. Correlation = -0.028, R-squared = 0.073

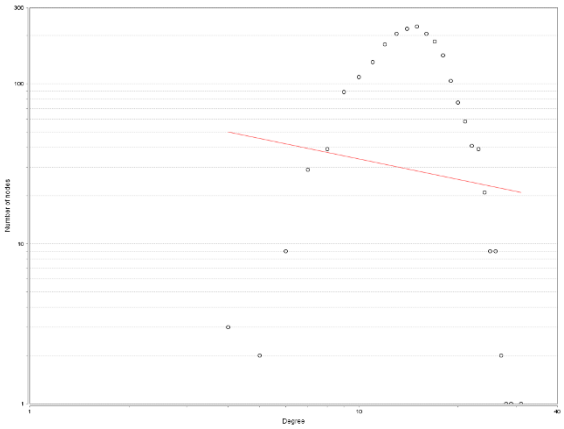


Figure 25 - Random interaction network 04. Correlation = -0.031, R-squared = 0.017

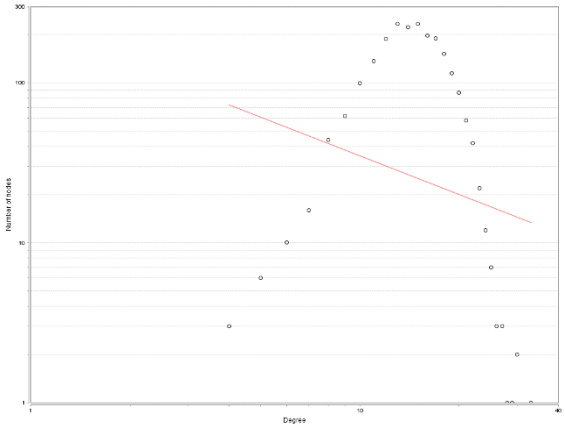


Figure 26 - Random interaction network 05.
Correlation = -0.072, R-squared = 0.059

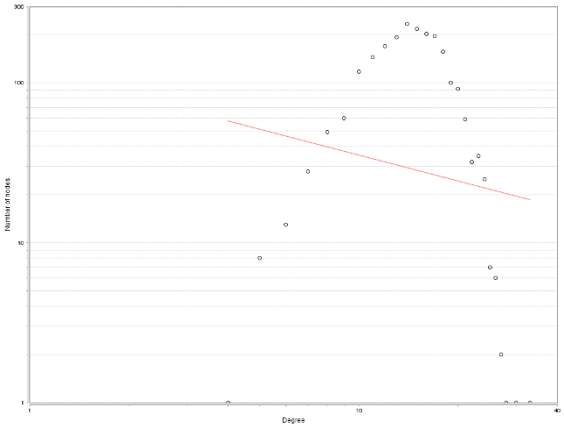


Figure 27 - Random interaction network 06.
Correlation = -0.049, R-squared = 0.027

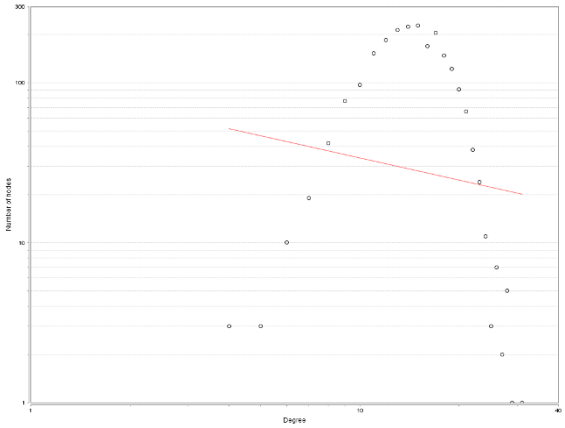


Figure 28 - Random interaction network 07.
Correlation = -0.042, R-squared = 0.021

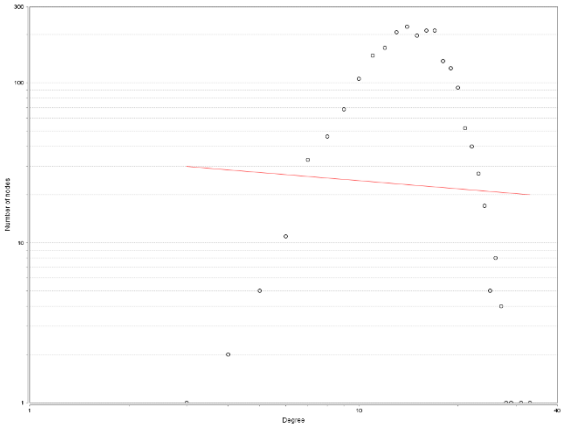


Figure 29 - Random interaction network 08.
Correlation = -0.029, R-squared = 0.003

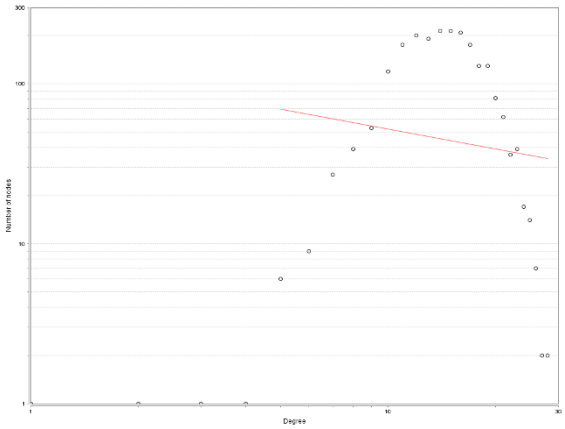


Figure 30 - Random interaction network 09.
Correlation = -0.012, R-squared = 0.020

Table 3 - Statistical comparison between the Cp *ovis* predicted networks against random networks.

Organism	Clustering Coefficient	Correlation	R-Squared	Shapiro-Wilk normality test
Cp1002	0.408	0.933	0.822	p-value < 2.2e-16
Cp267	0.402	0.953	0.785	p-value < 2.2e-16
Cp3995	0.410	0.933	0.798	p-value < 2.2e-16
Cp4202	0.410	0.928	0.799	p-value < 2.2e-16
CpC231	0.407	0.936	0.825	p-value < 2.2e-16
Cpfrc41	0.408	0.930	0.786	p-value < 2.2e-16
CpI19	0.403	0.932	0.813	p-value < 2.2e-16
CpP54B96	0.404	0.935	0.800	p-value < 2.2e-16
CpPAT10	0.407	0.938	0.790	p-value < 2.2e-16
Random Networks	0.007 (for all)	-0.012 to -0.072	0.001 to 0.073	Not performed

The Clustering Coefficient, Correlation and R-Squared were calculated by NetworkAnalyzer plugin (Assenov *et al.*, 2008). The Shapiro-Wilk test was performed in R (Royston, 1982).

3.1.8.2.1 – References

1 Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T. & Albrecht, M. Computing topological parameters of biological networks. *Bioinformatics* **24**, 282-284 (2008).

2 Barabási, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5**, 101-113 (2004).

3 Royston, J. An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, 115-124 (1982).

3.1.8.3 – Analyses of protein clusters

Supplementary Material S3: Analyses of protein clusters formed from *Corynebacterium pseudotuberculosis* biovar *ovis* protein-protein interaction network. In the figures we provide further details and information; in addition to proteins that form the cluster, same proteins were included interacting within the cluster. Their respective coding genes represent the proteins. In the network pictures, the color and size of nodes and edges were configured to show specific properties of the network, always from the lowest to the highest value of the chosen property. The node size (from smallest to largest) and color (in a range of yellow, light green to dark green) represent the property Degree. The border node size (from smallest to largest) and color (in a range of white, pink and dark red) represent the "Betweenness Centrality" property. The edge color, on a scale of red, yellow, light green to dark green, represents the score from public database where the interaction was mapped. The lowest score represents 0.70 in all networks. The edge width, from thinner to widest, represents the interaction score pair (ISP). The lowest value of ISP is 0.5625.

3.1.8.3.1 - Complex analysis

Complexes are formed by groups of identical proteins (homomers) or different proteins (heteromers), and their organization is important in performing specific biological activities in a biological process (Dai *et al.*, 2014). Such complexes are subject to evolutionary selection to form metabolic pathways (Marsh *et al.*, 2013). In an interaction network, complexes are large groups of densely connected proteins forming clusters (Morris *et al.*, 2011). To identify the clusters in the predicted networks, we used the Markov Cluster Algorithm (MCL) with inflation value set to 3.0 (Van Dongen, 2000), implemented in the Plugin ClusterMaker (Morris *et al.*, 2011) available in the Cytoscape (Shannon *et al.*, 2003) software. In addition, to validate the interaction networks, a literature search was performed to verify the existence of these clusters in other organisms, in the form of operons or metabolic pathways. For the PPI network and the complex visualization, we used the Circular or Edge-weighted Spring Embedded Cytoscape Layout (Kohl, Wiese e Warscheid, 2011).

3.1.8.3.2 - Ribosomal and RNA polymerase cluster

The complex is a network representation of protein-protein interactions (PPI) formed during the translational process of ribosomes (ribosomal RNAs + protein) in *C. pseudotuberculosis.* This complex is formed by 53 ribosomal proteins (RP) and four of the five proteins comprise the RNA polymerases (RNAP). All proteins are conserved in *C. pseudotuberculosis* biovar

ovis strains where the presence of transcriptional and translational machinery components is noted. The RPs in the network are encoded by 23 genes *rpl* (*rplBICEMKAQSDNLTFPOVJRWUXY*), 10 genes *rpm* (*rpmAEHBDCGIFJ*) and 20 genes *rps* (*rpsLBKIDEOJGCMHARSPNFQT*) (Haddadin e Harcum, 2005). The RNAP proteins are encoded by genes *rpoA, rpoB, rpoC* and *rpoZ* (Coenye e Vandamme, 2005; Teixeira *et al.*, 2008) (Figure 31). In the interaction network it can be observed that operon containing genes encode ribosomal proteins and genes encode proteins that form the subunits of RNAP, for example, the rplKAJL-rpoBC operon encoding the proteins of the large subunit of ribosome and also the β and β' subunits of RNAP (Teixeira *et al.*, 2008). As in prokaryotes, the transcriptional and translational systems are coupled and synchronized in space and time; such information may be relevant for understanding the dependence between these two processes (Mcgary e Nudler, 2013). It is because when transcripts are generated for RP, probably transcripts for RNAP proteins are also generated and therefore will join with other components to assemble the respective machinery. *Escherichia coli* was the first organism having the ribosomal component (rRNA + proteins) elucidated (Stelzl *et al.*, 2001), and hence is being widely used as a model for studies of ribosomal gene clusters in bacteria due to the similarity in the formation and organization of these clusters. In *C. Glutamicum* and *C. diphtheriae,* eleven gene clusters encoding 42 ribosomal proteins have been described. Comparing with the *E. coli* gene clusters, seven of them are organized in the same way and four have high similarity (MartíN *et al.*, 2003). Furthermore, when we look at the different bacterial genomes or even between different strains, we do not observe the conservation of all RPs (Coenye e Vandamme, 2005). This can possibly modify the pattern of interactions between the components of the translational and transcriptional machinery and somehow influence the expression of different genes in a given environmental condition.

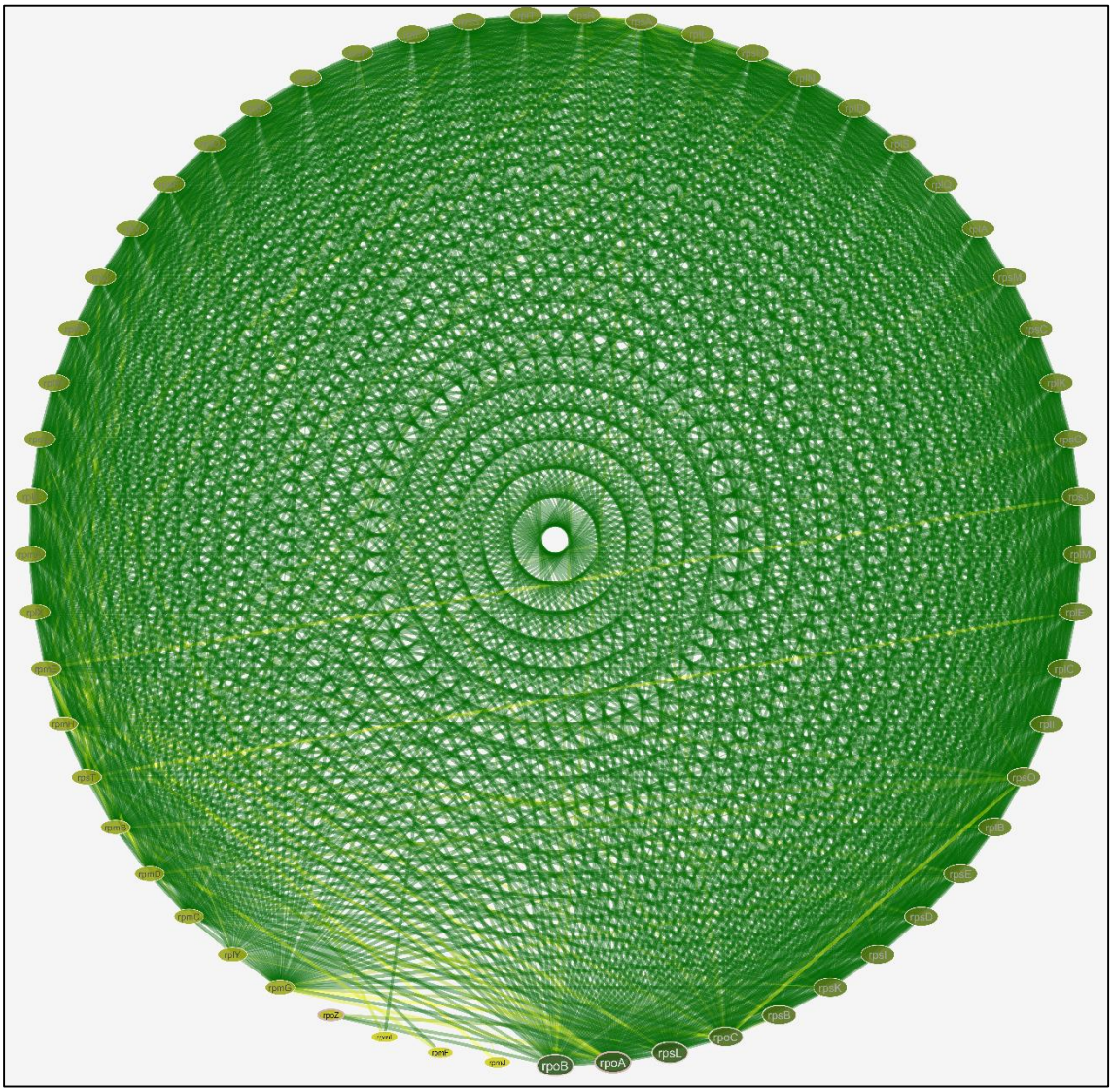


Figure 31 - Network formed by the interaction of RNA polymerase and ribosomal proteins, represented by their encoding gene.

Recent studies have attempted to identify and establish *in vitro* analyses in terms of possible physical-molecular contact between the components of the ribosomal machinery and RNAP and hence determine the influence of one machinery over the other. In one study, it was observed that the complex formed by the proteins encoded by the genes *nusG-rpsJ*, bind RNAP to the 30S subunit of the prokaryotic ribosome (Castro-Roa e Zenkin, 2012). In another study, the gene that encodes the S1 protein also binds to RNAP and stimulates transcriptional activity (Sukhodolets e Garges, 2003); these interactions are also observed in the networks of the present study. Other important observations can be found in the network, such as: large interaction of proteins encoded by genes *rpoB, rpoC* and *rpoA* with RP and no interactions of the protein encoded by the gene *rpoZ* with RP. This can be justified by the fact

that *rpoZ* is a sigma factor responsible for recognizing the binding site. After the protein beta subunits (β-encoded by *rpoB* gene), beta' (β'- encoded by gene *rpoC*) and alpha (α-encoded by *rpoA* gene) form the RNAP, *rpoZ* disconnects from the binding site. The network analysis can help us also select molecular targets for possible drug action. By observing the proteins encoded by the *rpoA* gene, *rpoB* and *rpoC*, we could note that they are highly connected proteins to RP. Thus, they can all potentially serve as candidate targets for drug development. An example in the literature is the RNAP β subunit inhibition (encoded by the *rpoB* gene) by antibiotic Rifampicin. There are also antibiotics like tetracycline, paromomycin, spectinomycin and streptomycin that exert their inhibitory activity on some proteins in the ribosomal 30S complex (Adékambi, Drancourt e Raoult, 2009).

3.1.8.3.3 - Oligopeptide transport system cluster

The Opp transporters belonging to the ABC transporters family (ATP-binding cassette) were identified and characterized in several bacterial species, both in gram-positive and gram-negative (Braibant e Gilot, 2000; Monnet, 2003). This system consists of five protein subunits: OppA, responsible for the peptides capture of extracytoplasmic means; OppB and OPPC form the transmembrane channel through which the oligonucleotides will be transported to the intracellular environment; OppD and OppF, are located in the bacterial cytoplasm and are responsible for the hydrolysis of ATP molecules generating power for the process of internalizing peptides (Braibant e Gilot, 2000). From a genetic point of view, the genes encoding these subunits are organized as an operon *oppABCDF* (Hiron *et al.*, 2007) (Figure 32). In bacteria, the main function of *Opp* is probably the peptides acquisition to be used as carbon and nitrogen source. In *E. coli*, it was demonstrated that this system is associated with the residues internalization of various amino acid types (Naider e Becker, 1975). A study of *Lactococcus lactis* has shown that the presence of a functional peptide transport system is required for the growth of bacteria in milk (Smid, Plapp e Konings, 1989). According to the generated interaction network, the Opp system is directly linked to the protein dihydrodipicolinate synthase (nanL) participating in L-lysine biosynthesis suggesting that this system may be associated with L-lysine metabolism.

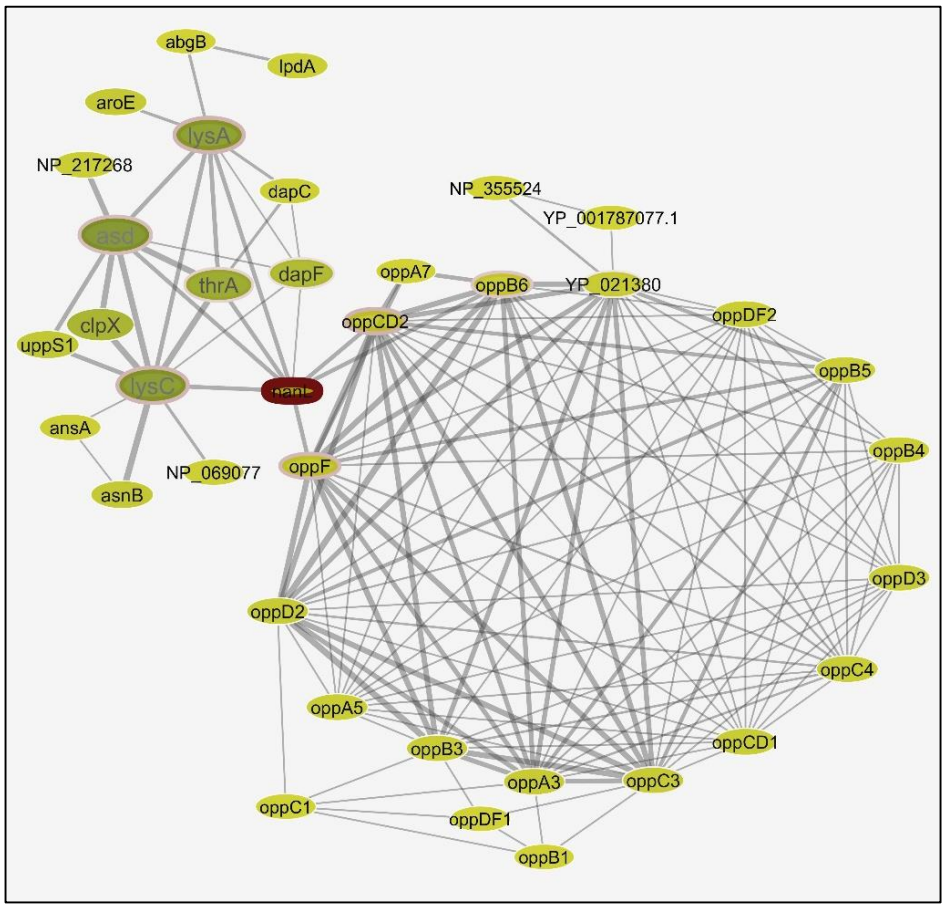


Figure 32 - Network formed by the interaction of Opp proteins, represented by their encoding genes

To date, no study was conducted to demonstrate the role of the Opp system in the transport of essential and nonessential amino acids in *C. pseudotuberculosis*. However, it was shown that the *Opp* system could contribute to the adhesion process of this pathogen. In tests conducted in experimental infection in a murine model, *oppD* mutant strains showed the same potential virulence compared to the wild type strain (Moraes *et al.*, 2014). In *Moraxella catarrhalis,* it was demonstrated that the Opp system is also involved in the acquisition of arginine and contributes to the fitness and persistence of the pathogen in the respiratory tract (Jones *et al.*, 2014). These studies demonstrate the versatility of the Opp system in pathogenic bacteria.

3.1.8.3.4 - Cobalamin biosynthesis cluster

The cobalamin (CBL - Vitamin B_{12}), members of the structurally complex cofactors class (Rodionov *et al.*, 2003; Croft *et al.*, 2005), is synthesized by a number of Archaea and Bacteria (Roth, Lawrence e Bobik, 1996; Scott e Roessner, 2002). However, the prosthetic group CBL is essential for the enzymatic activity of several enzymes in all the three biological domains (Yin e Bauer, 2013). In Bacteria and Archaea, the functional dependency is present

in the CBL methionine synthase, ribonucleotide reductase, glutamate, methylmalonyl-coA mutases, ethanolamine ammonia lyase, etc. (Rodionov *et al.*, 2003). The biosynthesis pathways of CBL cofactors, chlorophyll and haem begin with the compound 5-aminolevulinic acid (ALA). This, through some enzymatic steps, is converted into Uroporphyrinogen III, the last common intermediate compound (Frankenberg, Moser e Jahn, 2003; Heldt *et al.*, 2005; Yin e Bauer, 2013). It is noteworthy that all the co-factors are derived from tetrapyrroles molecules (Rondon, Trzebiatowski e Escalante-Semerena, 1996; Frankenberg, Moser e Jahn, 2003; Heldt *et al.*, 2005). In the predicted PPI network for *C. pseudotuberculosis* CBL complex, we note the presence of several holoenzymes (*HemABCDEL*) interconnected with the holoenzymes (*CobABDFGHJKLMNOQST*) (Figure 33).

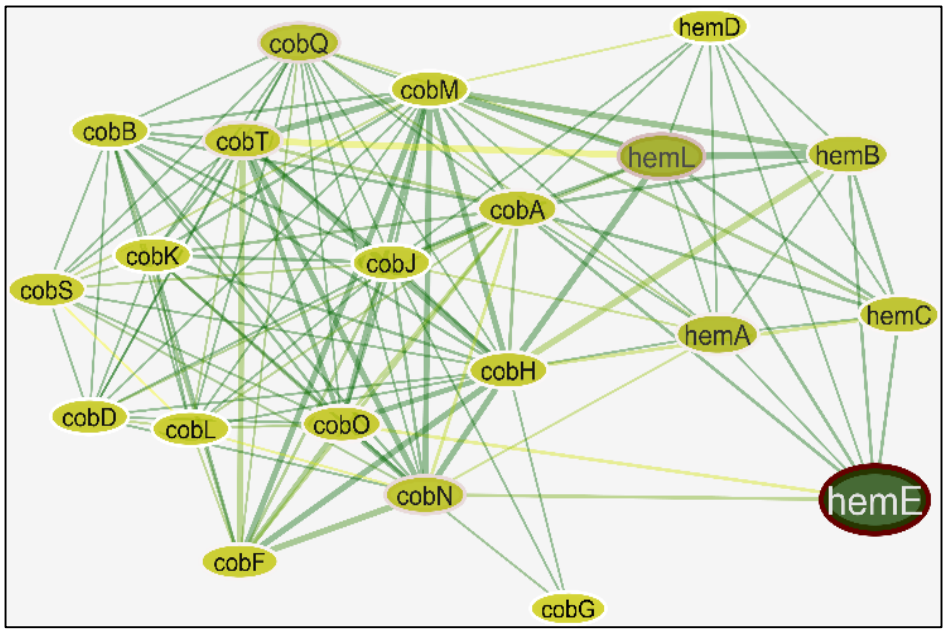


Figure 33 - Network formed by the interaction of Cob proteins, represented by their encoding genes

This may suggest a co-evolutionary dependence between clusters. A correspondence can be made with the *Rhodobacter sphaeroides* where excess haem inhibits 5-aminolevulinic acid synthase enzyme, affecting the biosynthesis of chloroplast (Yin e Bauer, 2013). Thus, observing the CBL network and the interaction between the different protein clusters, we can assume the existence of several regulatory mechanisms that are much more complicated. For cobalamin production, multiple steps and structural rearrangement of transmethylation are required (Rodionov *et al.*, 2003). In *C. pseudotuberculosis*, 15 cob genes catalyze these reactions, with most of them being in the main cob operon, while the remaining genes (*cobA, cobB, cobC and cobD*) are not present in the main operon. This fact may indicate the

contribution of these genes to external assimilation of vitamin B_{12} precursors or secondary processes of de novo biosynthesis, as identified in *Pseudomonas denitrificans* (Roth, Lawrence e Bobik, 1996). The cbi gene cluster (cobinamide), responsible for CBL biosynthesis by anaerobic pathway (Moore e Warren, 2012), is absent in the network; so we can postulate that *C. pseudotuberculosis* use solely the aerobic pathway as an alternative to produce CBL (Rodionov *et al.*, 2003), remembering that *C. pseudotuberculosis* is an anaerobic facultative microorganism (Dorella *et al.*, 2006).

3.1.8.3.5 - Iron uptake and intracellular regulation cluster

This complex is a representation of the PPI network for the capture process and intracellular regulation of iron (Fe) in *C. pseudotuberculosis*. Fe is an essential cofactor for diverse enzymatic activities that work in different metabolic processes (e.g., DNA replication, ATP synthesis, DNA repair and respiration etc.) in all eukaryotic organisms and various prokaryotes (Smith, 2004; Trost *et al.*, 2010; Schalk, 2013). In pathogenic bacteria such as *C. pseudotuberculosis*, the Fe^{+} ions acquisition system contributes to the survival and virulence of the microorganism (Köster, 2001; Kunkle e Schmitt, 2005). A single bacterium can have multiple Fe acquisition systems. This feature is used as a strategy to acquire Fe from different sources and in low availability of this cofactor (Wandersman e Delepelaire, 2004). Thus, the complex represents these multiple systems and consists of 22 proteins encoded by genes *fagABCD, ciuABCD, fecCDE (CD), hmuUVTO, htaA, pstA, fhuD, fpeC1, hemE* and *dtxR* (Figure 34).

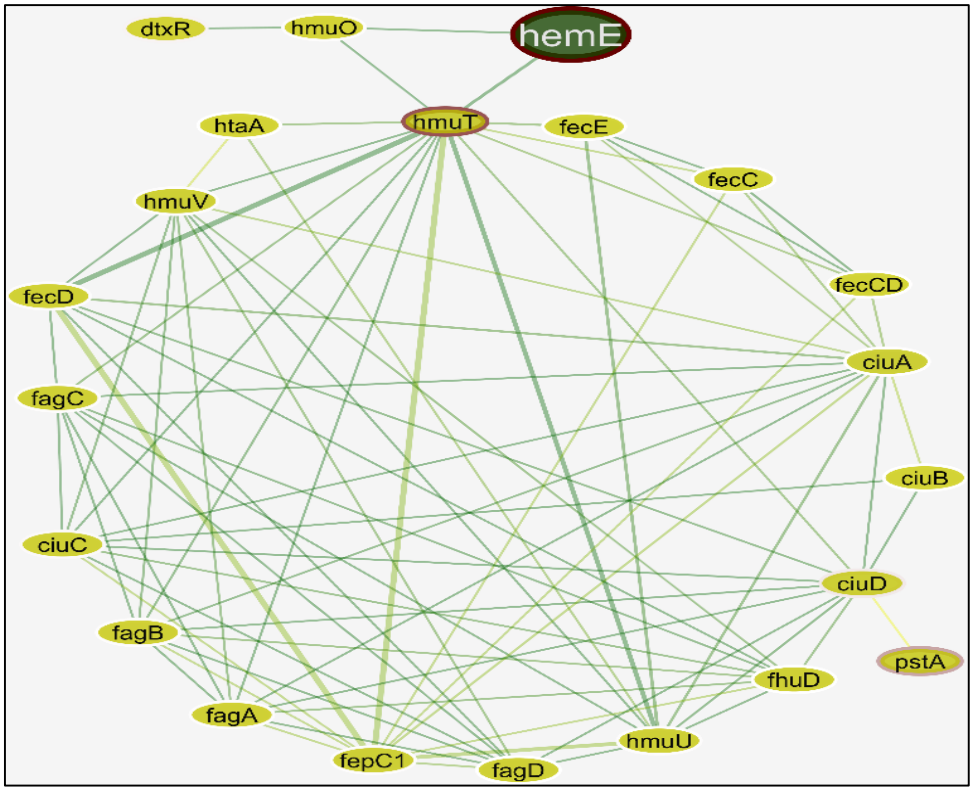


Figure 34 - Network formed by the interaction of Iron uptake proteins, represented by their encoding genes.

During the infection process, *C. pseudotuberculosis* is able to survive and multiply within macrophages and hence escape from the host immune system response (Trost *et al.*, 2010). One of these abilities can be related to the use of distinct or multiple siderophores (SIDS) (Correnti e Strong, 2012) synthesized by *C. pseudotuberculosis* or captured from the external environment (Schalk, 2013). In *C. pseudotuberculosis*, the SIDS are synthesized by genes *fagD* (Contreras *et al.*, 2014) (represented in the network) and *ciuE* (Trost *et al.*, 2010) (not present in the network). Probably these SIDS compete for the Iron ion (Fe^{+}) with iron transporters used by the macrophage (Schalk, 2013). Another source of Fe^{+} can be derived from the transfer of the prosthetic group heme-Fe to the inside of *C. pseudotuberculosis* through *hmuT* receiver, whose interactions between *hmuT* and *hemE* can be seen in the network. This interaction occurs for the transfer of Heme-Fe to the inside of *C. pseudotuberculosis*; it then suffers a degradation process, releasing Fe^{+}. In this process of degradation, hmuO operates in the cleavage of the tetrapyrrole ring of the group Heme-Fe (Contreras *et al.*, 2014). Additionally, in the network the protein Cell-surface hemin receptor (htaA) exclusively interacts with proteins encoded by the *hmuTUV* genes, responsible for hemin binding and transport. These interactions agree with the literature in *C. diphtheriae*, wherein the HtaA was able to acquire hemin from hemoglobin and transport to cytosol by an

ABC transporter (Allen e Schmitt, 2011). These observations suggest that the interaction network is consistent and also *C. pseudotuberculosis* can use the same strategy for iron acquisition. In the network, there are also other systems for capturing iron, such as: Fag, Fec and Ciu proteins, as part of *C. pseudotuberculosis* strategy to acquire Fe^{+}. One strategy that has been adopted to combat resistant bacteria is the 'Trojan Horse', which uses the iron uptake system to enter and kill the cell. The idea is based on the synthesis of the siderophore-drug complex, thus making the iron acquisition pathways through siderophore as potential targets for drug delivery (Górska, Sloderbach e Marszałł, 2014). Recently, a detailed review about iron acquisition strategies of gram-positive pathogens was published where the cluster proteins are cited, confirming the integrity of the predicted interaction network. Iron, being an important substance for survival and infection in gram-positive bacteria, the mechanisms of iron acquisition, transportation and processing become important areas of study, whose understanding might enable the development of new strategies to combat these organisms (Sheldon e Heinrichs, 2015).

3.1.8.3.6 - Cell division and peptidoglycan biosynthesis

In various bacteria, there is a coupling and fine coordination between the processes related to cell division (cytokinesis), the formation of the peptidoglycan layer that makes up the cell walls, and DNA replication and segregation systems (Lutkenhaus And e Addinall, 1997; Buss *et al.*, 2015). The 36 proteins from *C. pseudotuberculosis* represents this process and their interactions are shown in the predicted network (Figure 35), highlighting the FtsZIWHYXE protein involved in cell division (Lutkenhaus And e Addinall, 1997; Errington, Daniel e Scheffers, 2003) and the MurAFDEGIBC proteins responsible for the biosynthesis of peptidoglycans (El Zoeiby, Sanschagrin e Levesque, 2003). In the cytokinesis process, the FtsZ protein plays a central role in the formation of the cytoplasmic membrane ring constriction and, in the anchoring and recruitment of another protein set related to the cell division process (Lutkenhaus And e Addinall, 1997; Errington, Daniel e Scheffers, 2003). In the network, the FtsZ protein is highly connected to their neighbors, thereby suggesting the multiple connections as a representative element of the recruitment activity and anchoring conducted by FtsZ. As FtsZ is the main component of the cell division process, there is a need to maintain a harmony with the enzymes relating to the new cell wall synthesis (Carballido-López e Errington, 2003). In the *C. pseudotuberculosis* network, these enzymes are mainly represented by MurABCDEFGI and mraY proteins, related to the synthesis of new multilayer

peptidoglycans cell wall (Vollmer, Blanot e De Pedro, 2008). Thus, the network clearly shows a possible harmony between the components responsible for the peptidoglycan biosynthesis and FtsZ protein. It is worth noting the role of FtsW protein in nascent peptidoglycan transport to the outside of the plasma membrane. In the network, we could observe the presence of the proteins encoded by *parA, parB* and *smc* genes related to the chromosome partitioning process; *soj* with ATPase activity and *scpA* related to the condensation process and the bacterial chromosome segregation during cytokinesis. These proteins mainly interact with FtsZ, showing that FtsZ serves as a support for these proteins to perform their activities accordingly. Complementary approaches using PPI networks can be of great value to overcome the challenges related to the increasing number of resistant pathogenic bacteria to several current therapies. Thus, the organization and the connection between the network elements can help us in the identification and selection of new molecular targets for the development of more effective therapies. Currently, there are several compounds being synthesized and directed to act in the inhibition of peptidoglycan synthesis and in cell division steps (Den Blaauwen, Andreu e Monasterio, 2014). For example, compounds such as fosfomycin (phosphomycin), 4-thiazolidinone and phosphinic acid derivatives that act as inhibitors of MurA, MurB and MurCDEF respectively (El Zoeiby, Sanschagrin e Levesque, 2003). In this case, the bacteria does not survive by not forming the peptidoglycan layers. Inhibitors directed to block the beginning of cell division by preventing the formation of the constriction ring has been explored and tested against FtsZ, for example, the sanguinarine inhibitor that although showing inhibitory activity is not specific only to the target FtsZ (Den Blaauwen, Andreu e Monasterio, 2014). Therefore, further studies are needed to find more efficient inhibitors and most promising targets against various bacteria, especially against *C. pseudotuberculosis*; the protein-protein interaction networks are an important tool for this purpose.

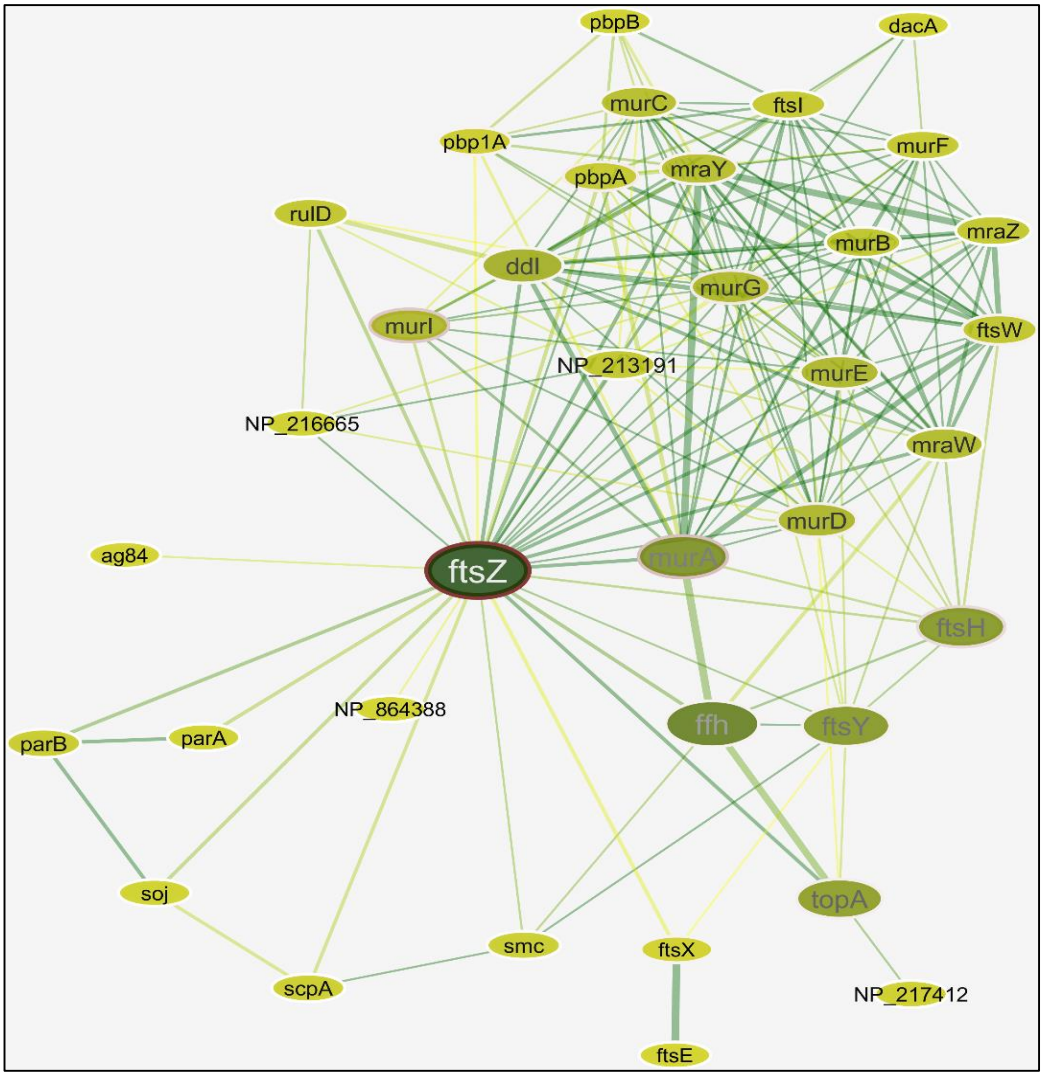


Figure 35 - Network formed by the interaction of proteins involved in cell division and peptidoglycan biosynthesis, both represented by their encoding genes.

In general, the clusters whose proteins are described in the literature (although in other organisms), prove the consistency of our predicted interaction network, reinforcing that the interactions can truly occur in Cp *ovis*. An example is the iron acquisition cluster participants whose proteins were cited in a recent review (Sheldon e Heinrichs, 2015). Despite the clusters identified and characterized individually in the interaction networks, it is common that some proteins also interact in several clusters, possibly exerting different function in each cluster. This is the case of Iron uptake, Cobalamin biosynthesis and Heme clusters, whose cooperation was characterized and described in other organisms (Köster, 2001).

Likewise, clusters or interactions not previously described, or those poorly characterized in the literature, could bring further new and relevant information about Cp *ovis*. From the

clusters analysis, we conclude the following: some proteins, operons and interaction participants in the clusters are well described in the literature for other gram-positive organisms, fortifying that the predicted interaction networks are biologically feasible for Cp *ovis* and; although some proteins and operons are well described in the literature, in some cases, the interactions between these elements are not; hence, the interaction network has the potential to contribute more information leading to a better understanding of Cp ovis, and generating new testable hypotheses. A lack of information in the literature regarding certain interactions makes the PPI networks an important tool to better understand cellular behavior and to raise new hypotheses about the biochemical processes of Cp *ovis*, making possible direct future experiments to test the essentiality or druggability of these interactions.

3.1.8.3.7 - References

1 Dai, Q.-G., Guo, M.-Z., Liu, X.-Y., Teng, Z.-X. & Wang, C.-Y. CPL: Detecting Protein Complexes by Propagating Labels on Protein-Protein Interaction Network. *Journal of Computer Science and Technology* **29**, 1083-1093 (2014).

2 Marsh, J. A. *et al.* Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell* **153**, 461-470 (2013).

3 Morris, J. H. *et al.* clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC bioinformatics* **12**, 436 (2011).

4 Van Dongen, S. A cluster algorithm for graphs. *Report-Information systems*, 1-40 (2000).

5 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498-2504 (2003).

6 Kohl, M., Wiese, S. & Warscheid, B. in *Data Mining in Proteomics* 291-303 (Springer, 2011).

7 Haddadin, F. a. T. & Harcum, S. W. Transcriptome profiles for high-cell-density recombinant and wild-type Escherichia coli. *Biotechnology and bioengineering* **90**, 127-153 (2005).

8 Teixeira, D. *et al.* The tufB–secE–nusG–rplKAJL–rpoB gene cluster of the liberibacters: sequence comparisons, phylogeny and speciation. *International Journal of Systematic and Evolutionary Microbiology* **58**, 1414-1421 (2008).

9 Coenye, T. & Vandamme, P. Organisation of the S10, spc and alpha ribosomal protein gene clusters in prokaryotic genomes. *FEMS microbiology letters* **242**, 117-126 (2005).

10 McGary, K. & Nudler, E. RNA polymerase and the ribosome: the close relationship. *Current opinion in microbiology* **16**, 112-117 (2013).

11 Stelzl, U., Connell, S., Nierhaus, K. H. & Wittmann-Liebold, B. Ribosomal proteins: role in ribosomal functions. *eLS* (2001).

12 Martín, J. F., Barreiro, C., González-Lavado, E. & Barriuso, M. Ribosomal RNA and ribosomal proteins in corynebacteria. *J. Biotechnol* **104**, 41-53 (2003).

13 Castro-Roa, D. & Zenkin, N. In vitro experimental system for analysis of transcription–translation coupling. *Nucleic acids research* **40**, e45-e45 (2012).

14 Sukhodolets, M. V. & Garges, S. Interaction of Escherichia coli RNA polymerase with the ribosomal protein S1 and the Sm-like ATPase Hfq. *Biochemistry* **42**, 8022-8034 (2003).

15 Adékambi, T., Drancourt, M. & Raoult, D. The rpoB gene as a tool for clinical microbiologists. *Trends in microbiology* **17**, 37-45 (2009).

16 Monnet, V. Bacterial oligopeptide-binding proteins. *Cellular and Molecular Life Sciences CMLS* **60**, 2100-2114 (2003).

17 Braibant, M. & Gilot, P. The ATP binding cassette (ABC) transport systems of Mycobacterium tuberculosis. *FEMS microbiology reviews* **24**, 449-467 (2000).

18 Hiron, A., Borezée-Durant, E., Piard, J.-C. & Juillard, V. Only one of four oligopeptide transport systems mediates nitrogen nutrition in Staphylococcus aureus. *Journal of bacteriology* **189**, 5119-5129 (2007).

19 Naider, F. & Becker, J. M. Multiplicity of oligopeptide transport systems in Escherichia coli. *Journal of bacteriology* **122**, 1208-1215 (1975).

20 Smid, E. J., Plapp, R. & Konings, W. Peptide uptake is essential for growth of Lactococcus lactis on the milk protein casein. *Journal of bacteriology* **171**, 6135-6140 (1989).

21 Moraes, P. M. *et al.* Characterization of the Opp Peptide Transporter of Corynebacterium pseudotuberculosis and Its Role in Virulence and Pathogenicity. *BioMed research international* **2014** (2014).

22 Jones, M. M. *et al.* Role of the Oligopeptide Permease ABC Transporter of Moraxella catarrhalis in Nutrient Acquisition and Persistence in the Respiratory Tract. *Infection and immunity* **82**, 4758-4766 (2014).

23 Croft, M. T., Lawrence, A. D., Raux-Deery, E., Warren, M. J. & Smith, A. G. Algae acquire vitamin B12 through a symbiotic relationship with bacteria. *Nature* **438**, 90-93 (2005).

24 Rodionov, D. A., Vitreschak, A. G., Mironov, A. A. & Gelfand, M. S. Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. *Journal of Biological Chemistry* **278**, 41148-41159 (2003).

25 Roth, J., Lawrence, J. & Bobik, T. Cobalamin (coenzyme B12): synthesis and biological significance. *Annual Reviews in Microbiology* **50**, 137-181 (1996).

26 Scott, A. & Roessner, C. Biosynthesis of cobalamin (vitamin B (12)). *Biochemical Society Transactions* **30**, 613-620 (2002).

27 Yin, L. & Bauer, C. E. Controlling the delicate balance of tetrapyrrole biosynthesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **368**, 20120262 (2013).

28 Frankenberg, N., Moser, J. & Jahn, D. Bacterial heme biosynthesis and its biotechnological application. *Applied microbiology and biotechnology* **63**, 115-127 (2003).

29 Heldt, D. *et al.* Aerobic synthesis of vitamin B12: ring contraction and cobalt chelation. *Biochemical Society Transactions* **33**, 815-819 (2005).

30 Rondon, M. R., Trzebiatowski, J. R. & Escalante-Semerena, J. C. Biochemistry and molecular genetics of cobalamin biosynthesis. *Progress in nucleic acid research and molecular biology* **56**, 347-384 (1996).

31 Moore, S. & Warren, M. The anaerobic biosynthesis of vitamin B12. *Biochemical Society Transactions* **40**, 581 (2012).

32 Dorella, F. A., Pacheco, L. G. C., Oliveira, S. C., Miyoshi, A. & Azevedo, V. Corynebacterium pseudotuberculosis: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. *Veterinary research* **37**, 201-218 (2006).

33 Smith, J. L. The physiological role of ferritin-like compounds in bacteria. *Critical reviews in microbiology* **30**, 173-185 (2004).

34 Schalk, I. J. Innovation and Originality in the Strategies Developed by Bacteria To Get Access to Iron. *Chembiochem* **14**, 293-294 (2013).

35 Trost, E. *et al.* The complete genome sequence of Corynebacterium pseudotuberculosis FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. *BMC genomics* **11**, 728 (2010).

36 Köster, W. ABC transporter-mediated uptake of iron, siderophores, heme and vitamin B 12. *Research in microbiology* **152**, 291-301 (2001).

37 Kunkle, C. A. & Schmitt, M. P. Analysis of a DtxR-regulated iron transport and siderophore biosynthesis gene cluster in Corynebacterium diphtheriae. *Journal of bacteriology* **187**, 422-433 (2005).

38 Wandersman, C. & Delepelaire, P. Bacterial iron sources: from siderophores to hemophores. *Annu. Rev. Microbiol.* **58**, 611-647 (2004).

39 Correnti, C. & Strong, R. K. Mammalian siderophores, siderophore-binding lipocalins, and the labile iron pool. *Journal of Biological Chemistry* **287**, 13524-13531 (2012).

40 Contreras, H., Chim, N., Credali, A. & Goulding, C. W. Heme uptake in bacterial pathogens. *Current opinion in chemical biology* **19**, 34-41 (2014).

41 Allen, C. E. & Schmitt, M. P. Novel hemin binding domains in the Corynebacterium diphtheriae HtaA protein interact with hemoglobin and are critical for heme iron utilization by HtaA. *Journal of bacteriology* **193**, 5374-5385 (2011).

42 Górska, A., Sloderbach, A. & Marszałł, M. P. Siderophore–drug complexes: potential medicinal applications of the 'Trojan horse'strategy. *Trends in pharmacological sciences* **35**, 442-449 (2014).

43 Sheldon, J. R. & Heinrichs, D. E. Recent developments in understanding the iron acquisition strategies of gram positive pathogens. *FEMS microbiology reviews*, fuv009 (2015).

44 Lutkenhaus and, J. & Addinall, S. Bacterial cell division and the Z ring. *Annual review of biochemistry* **66**, 93-116 (1997).

45 Buss, J. *et al.* A Multi-layered Protein Network Stabilizes the Escherichia coli FtsZ-ring and Modulates Constriction Dynamics. (2015).

46 Errington, J., Daniel, R. A. & Scheffers, D.-J. Cytokinesis in bacteria. *Microbiology and Molecular Biology Reviews* **67**, 52-65 (2003).

47 El Zoeiby, A., Sanschagrin, F. & Levesque, R. C. Structure and function of the Mur enzymes: development of novel inhibitors. *Molecular microbiology* **47**, 1-12 (2003).

48 Carballido-López, R. & Errington, J. A dynamic bacterial cytoskeleton. *Trends in cell biology* **13**, 577-583 (2003).

49 Vollmer, W., Blanot, D. & De Pedro, M. A. Peptidoglycan structure and architecture. *FEMS microbiology reviews* **32**, 149-167 (2008).

50 den Blaauwen, T., Andreu, J. M. & Monasterio, O. Bacterial cell division proteins as antibiotic targets. *Bioorganic chemistry* **55**, 27-38 (2014).

3.1.8.4 – Cp267 PPI network

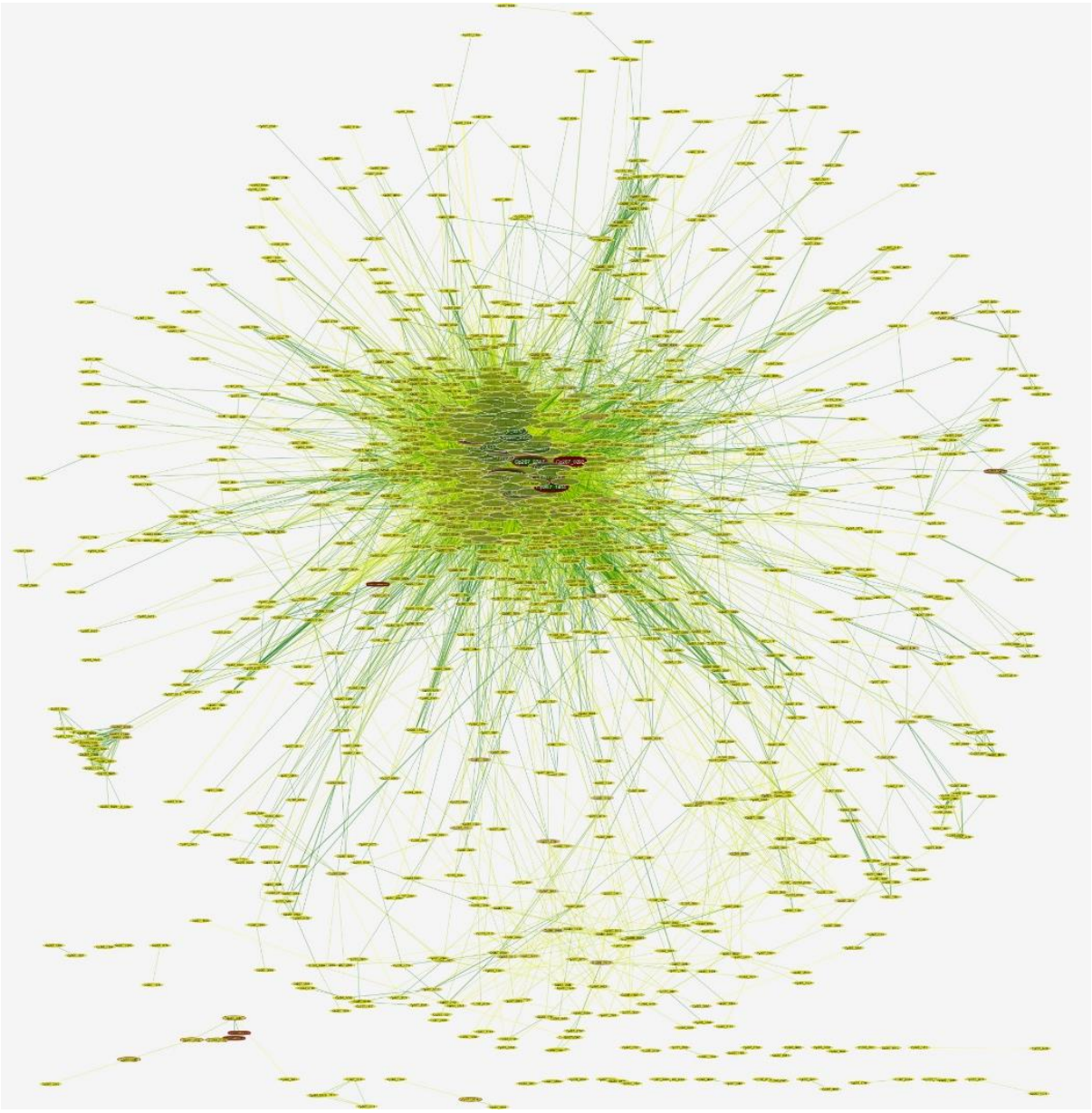


Figure 36 - Cp267 PPI network

3.1.8.5 – Cp3995 PPI network

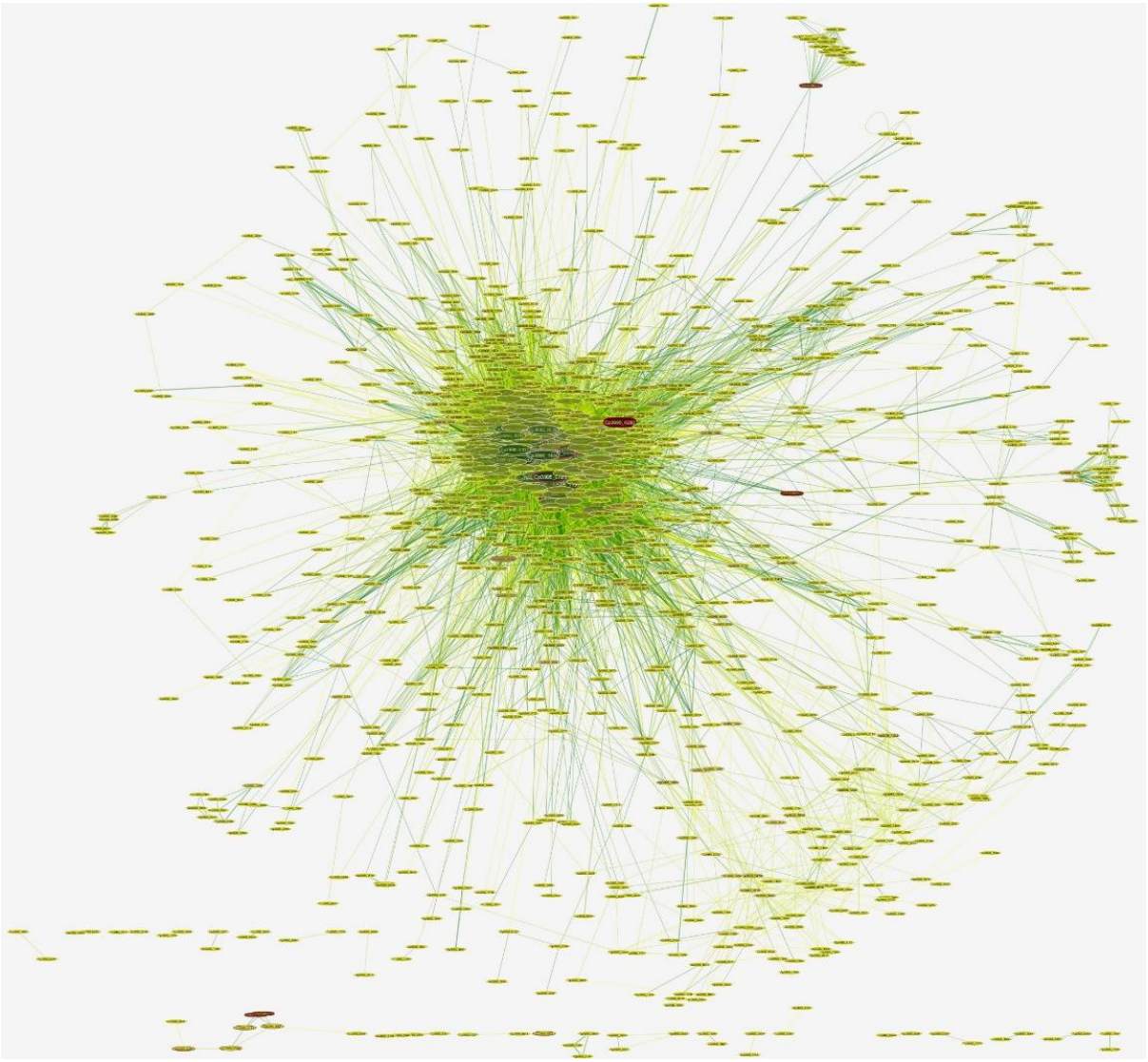


Figure 37 - Cp3995 PPI network

3.1.8.6 – Cp4202 PPI network

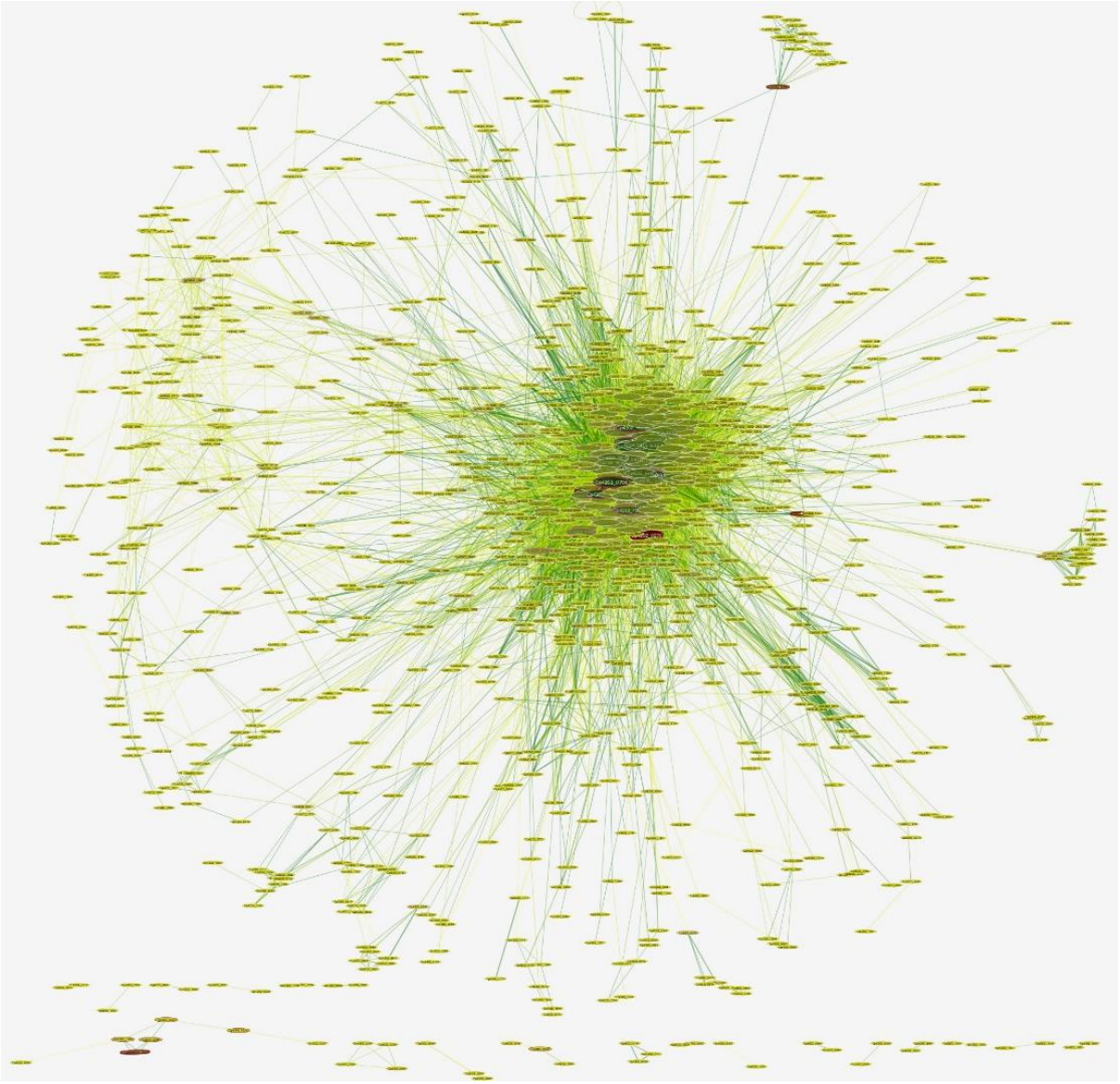


Figure 38 - Cp4202 PPI network

3.1.8.7 – CpC231 PPI network

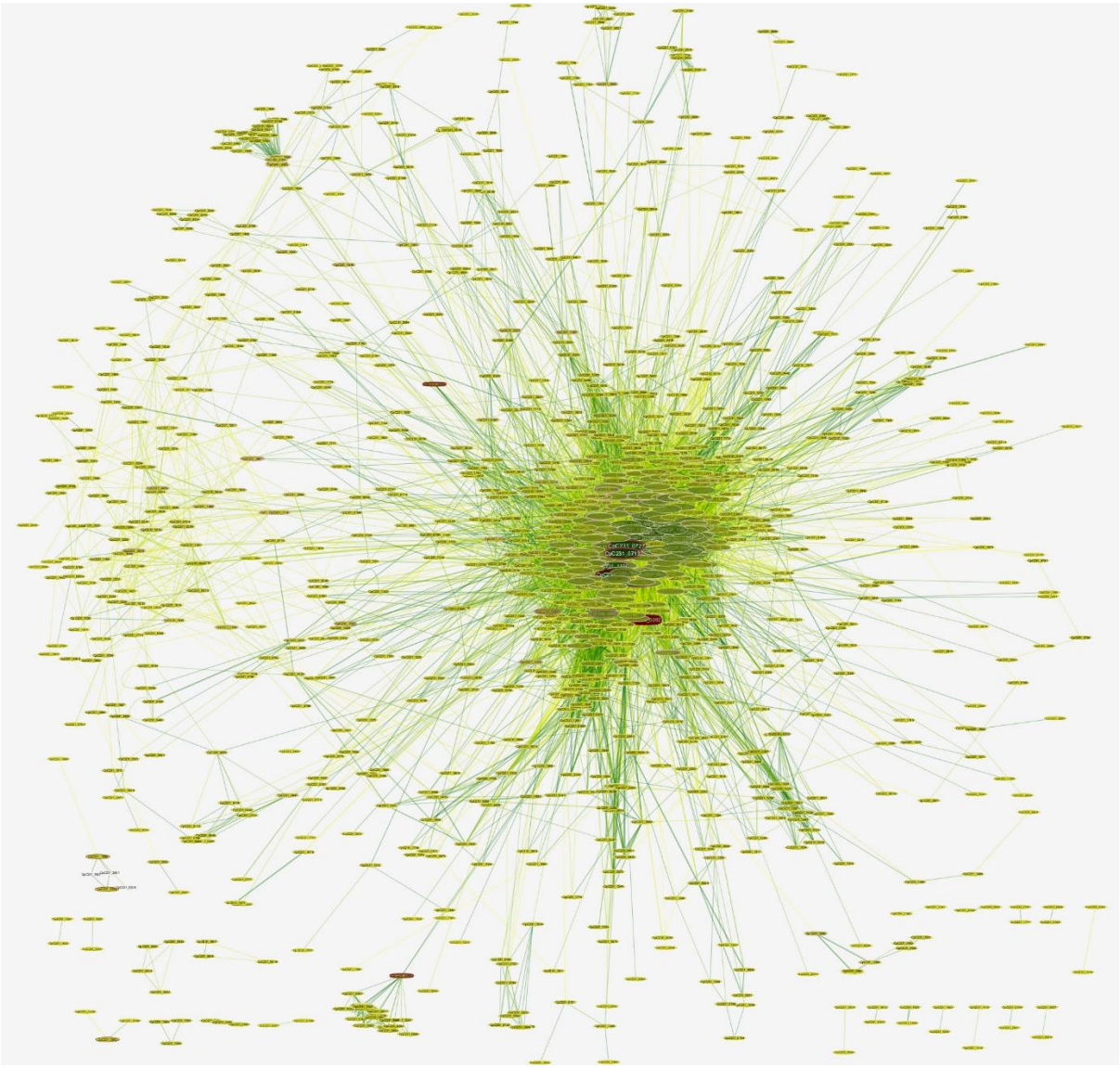


Figure 39 - CpC231 PPI network

3.1.8.8 – CpFRC PPI network

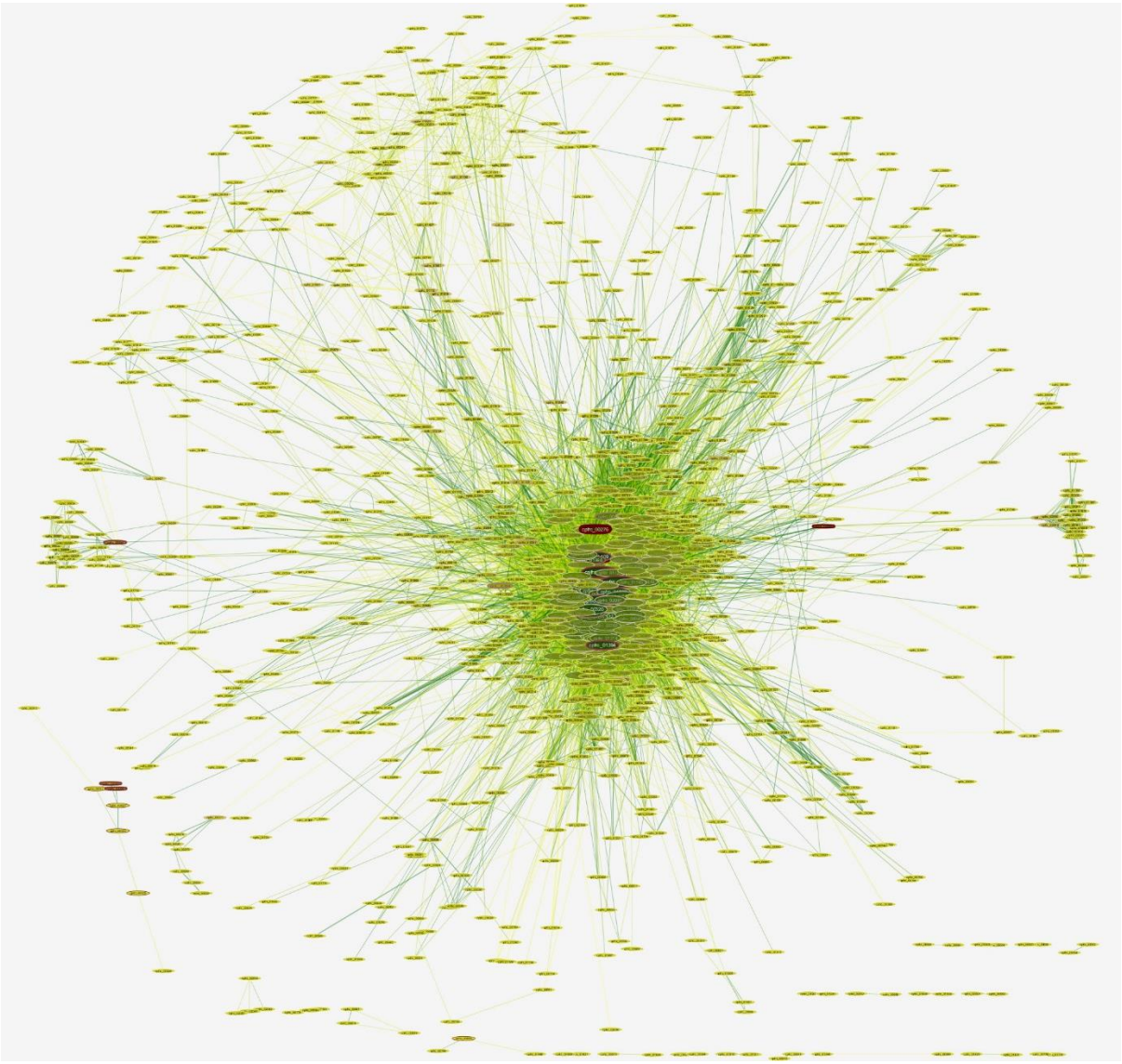


Figure 40 - CpFRC PPI network

3.1.8.9 – CpI19 PPI network

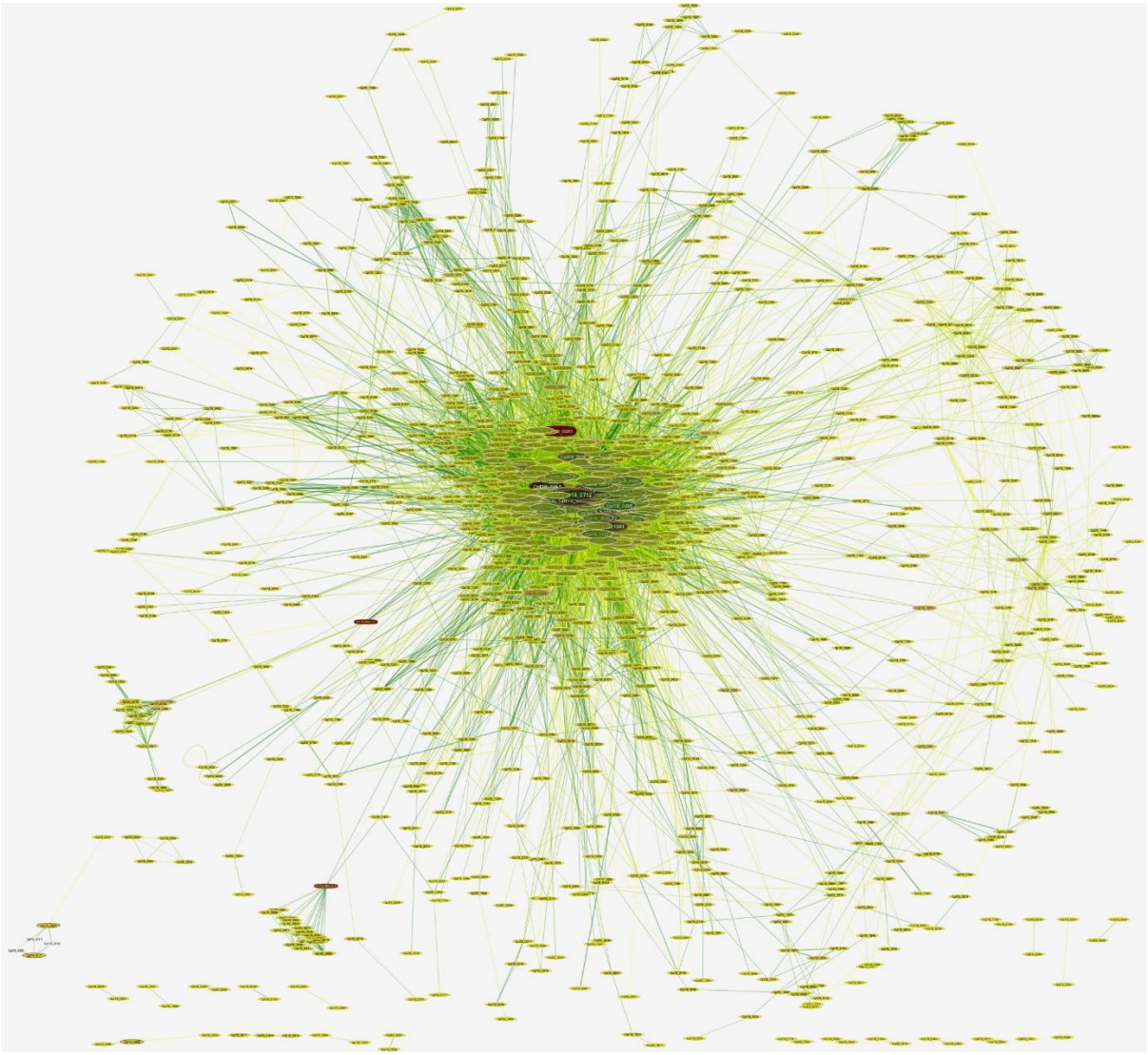


Figure 41 - CpI19 PPI network

3.1.8.10 – CpP54B96 PPI network

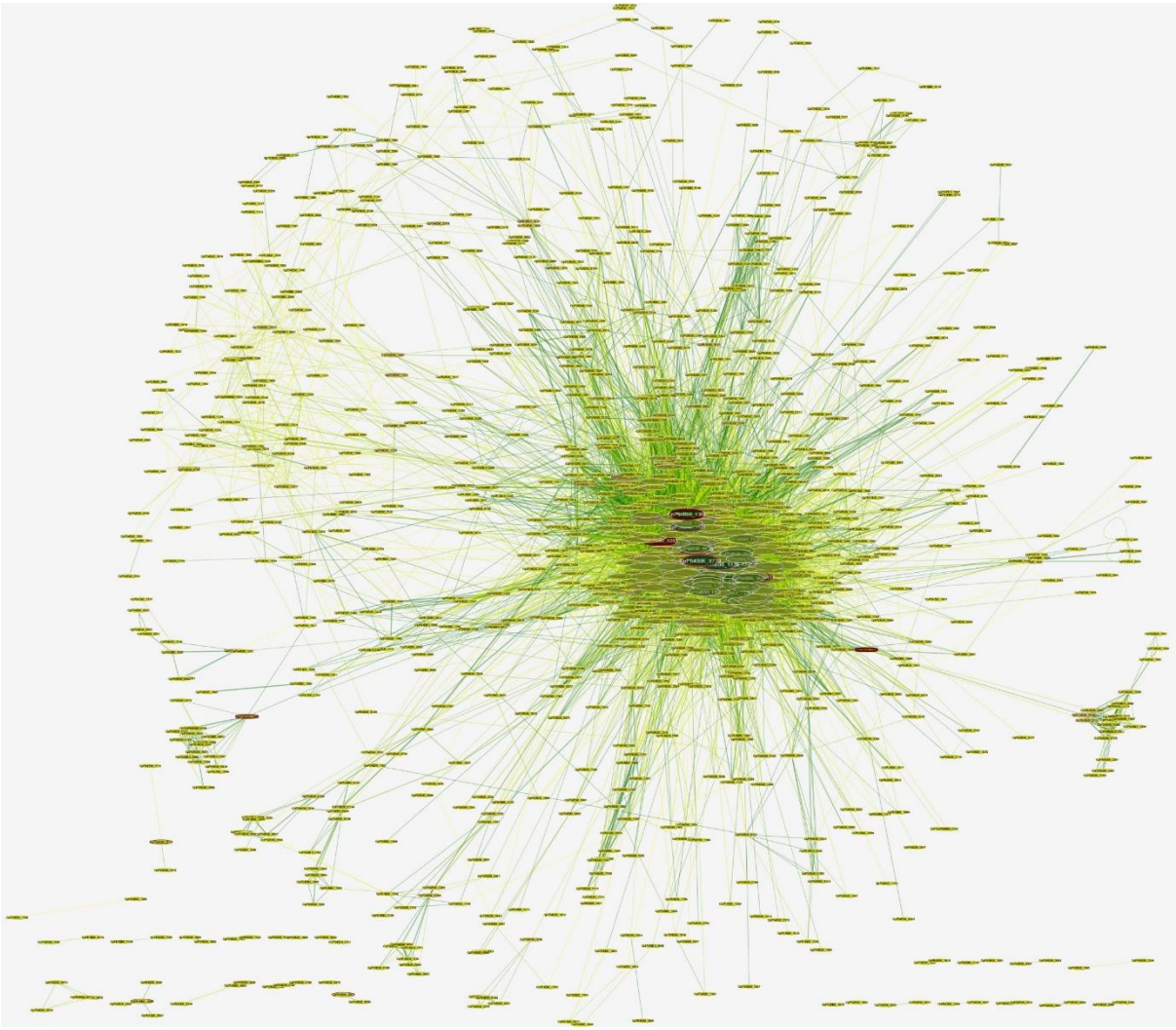


Figure 42 - CpP54B96 PPI network

3.1.8.11 – CpPAT10 PPI network

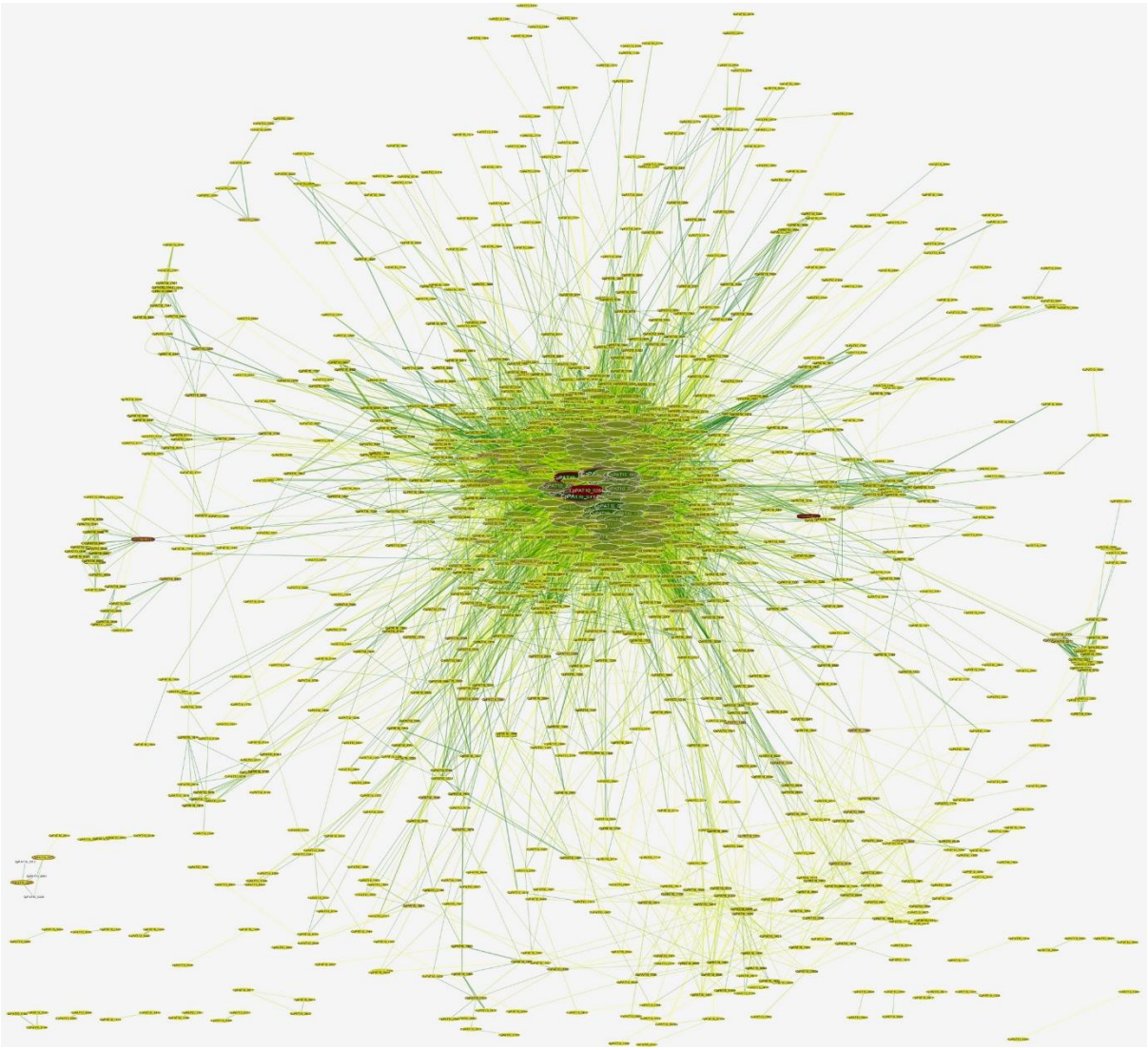


Figure 43 - CpPAT10 PPI network

3.1.8.12 – Cp1002 PPI network

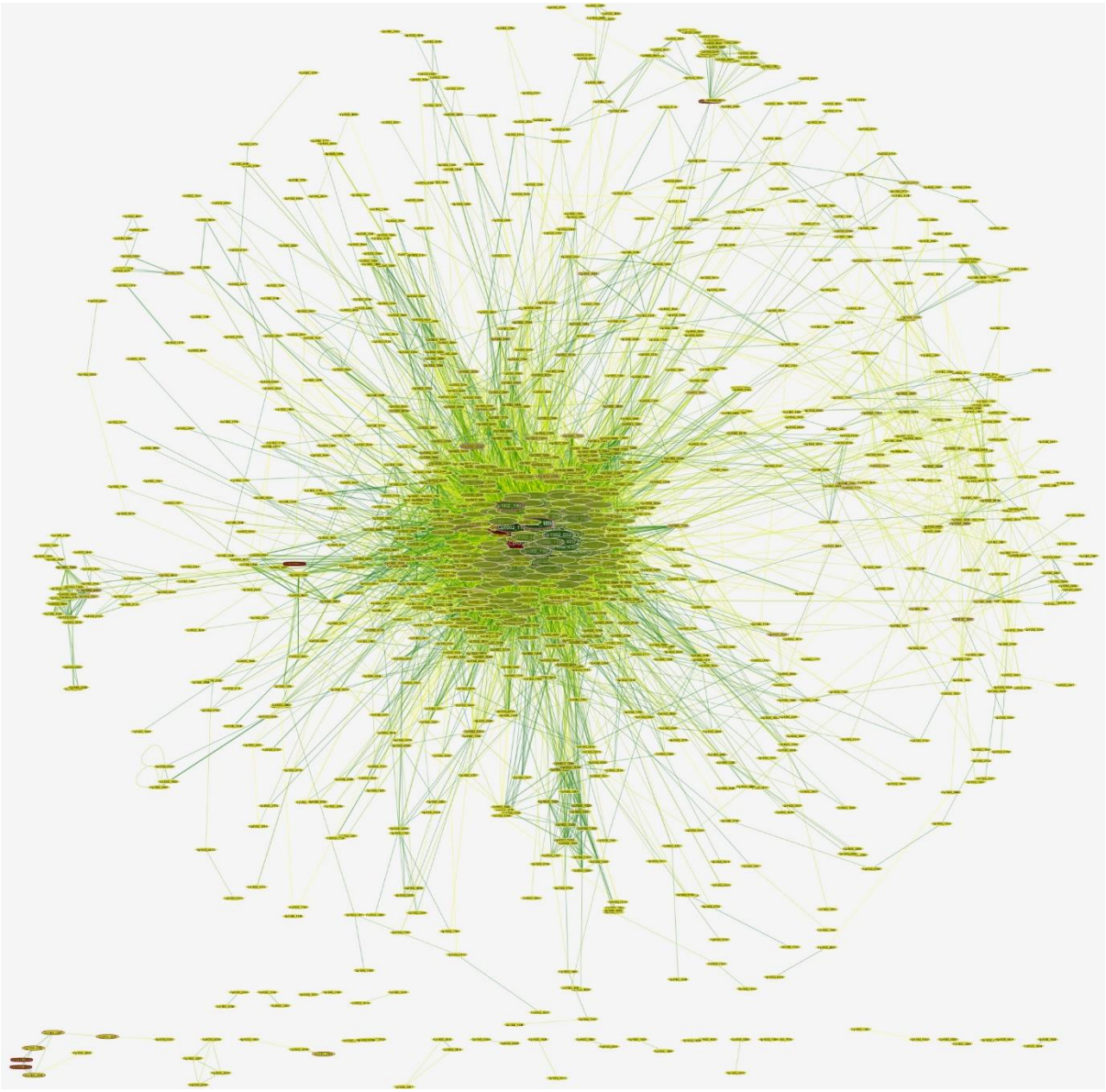


Figure 44 - Cp1002 PPI network

3.1.8.13 – List of top 15% proteins with higher degree against DEG

Supplementary Material

Supplementary Table S13: List of top 15% proteins with higher degree interaction, totaling 181 hub essential proteins. The amino acid sequence of hubs proteins was compared against bacterial proteins sequence from Database of Essential Genes (DEG) (Zhang, Ou e Zhang, 2004), v. 11.2, updated on July 3, 2015.

DEG Blast Genome Result - 8Xblk57KOz	
Your job ID: 8Xblk57KOz, which was completed in Tue Jul 28 01:32:28 2015 Beijing time.	
The result will be stored for 7 days and download Here.	
Organism:	Acinetobacter baylyi ADP1; Bacillus subtilis 168; Bacteroides fragilis 638R; Bacteroides thetaiotaomicron VPI-548 ; Burkholderia pseudomallei K96243; Burkholderia thailandensis E264; Campylobacter jejuni subsp. jejuni NCTC 11168 = ATCC 700819; Caulobacter crescentus; Escherichia coli MG1655 I; Escherichia coli MG1655 II; Francisella novicida U112; Haemophilus influenzae Rd KW20; Helicobacter pylori 26695; Mycobacterium tuberculosis H37Rv; Mycobacterium tuberculosis H37Rv II; Mycobacterium tuberculosis H37Rv III; Mycoplasma genitalium G37; Mycoplasma pulmonis UAB CTIP; Porphyromonas gingivalis ATCC 33277; Pseudomonas aeruginosa PAO1; Pseudomonas aeruginosa PAO1; Pseudomonas aeruginosa UCBPP-PA14; Salmonella enterica serovar Typhi; Pseudomonas aeruginosa PAO1; Salmonella enterica serovar Typhimurium SL1344; Salmonella enterica subsp. enterica serovar Typhimurium str. 14028S; Salmonella typhimurium LT2; Shewanella oneidensis MR-1; Sphingomonas wittichii RW1; Staphylococcus aureus N315; Staphylococcus aureus NCTC 8325; Streptococcus pneumoniae; Streptococcus pyogenes MGAS5448; Streptococcus pyogenes NZ131; Streptococcus sanguinis; Vibrio cholerae N16961;
Parameters:	deg.py -i /var/www/tubic/cgi-bin/blast/temp_seq/8Xblk57KOz/seq.txt -db /var/www/tubic/cgi-bin/blast/temp_seq/8Xblk57KOz/db -type seq -score 100 -email edson.folador@gmail.com -job 8Xblk57KOz -F F -e 0.00001 -M BLOSUM62 -g T -v 100 -b 100 -blastprogram blastp
Total protein-coding genes in your sequence:	181 genes
In your sequence, the No. of genes having homologs with DEG:	180 genes.
In DEG, the No. of genes having homologs with your sequence:	4356 genes.

Your Query Protein	No. of homologs in DEG	DEG AC Number
ackA 96 acetate kinase	6	DEG10140081; DEG10060294; DEG10180359; DEG10030589; DEG10220242; DEG10020202;
adk 172 adenylate kinase	24	DEG10170313; DEG10160162; DEG10010056; DEG10030207; DEG10060142; DEG10120257; DEG10340158; DEG10320059; DEG10130162; DEG10210026; DEG10190055; DEG10110036; DEG10310077; DEG10330165; DEG10240056; DEG10140221; DEG10290187; DEG10180089; DEG10380017; DEG10220167; DEG10020257; DEG10350054; DEG10200155; DEG10070123;

alaS 106 alanyl-tRNA synthetase	29	DEG10220278; DEG10380163; DEG10270472; DEG10050295; DEG10340311; DEG10060236; DEG10160188; DEG10360038; DEG10290291; DEG10100415; DEG10200338; DEG10030105; DEG10010189; DEG10110153; DEG10170223; DEG10340111; DEG10350227; DEG10020178; DEG10230270; DEG10130175; DEG10320224; DEG10140212; DEG10250501; DEG10330190; DEG10210070; DEG10310060; DEG10180418; DEG10370149; DEG10120181;
apt 72 Adenine phosphoribosyltransferase	8	DEG10080089; DEG10030222; DEG10290185; DEG10050439; DEG10310120; DEG10180087; DEG10140122; DEG10060229;
argC 86 N-acetyl-gamma-glutamyl-phosphate reductase	7	DEG10270317; DEG10130194; DEG10180564; DEG10280024; DEG10250346; DEG10100280; DEG10300019;
argF 115 ornithine carbamoyltransferase	9	DEG10130179; DEG10130195; DEG10280190; DEG10100218; DEG10250256; DEG10100284; DEG10280094; DEG10270250; DEG10240054;
argG 77 argininosuccinate synthase	7	DEG10280324; DEG10130167; DEG10240014; DEG10350178; DEG10270318; DEG10100285; DEG10250348;
argS 112 arginyl-tRNA synthetase	27	DEG10160081; DEG10230168; DEG10240020; DEG10180318; DEG10120144; DEG10350020; DEG10270225; DEG10380228; DEG10140084; DEG10330083; DEG10320175; DEG10340516; DEG10290353; DEG10170056; DEG10200407; DEG10060311; DEG10100191; DEG10010259; DEG10210210; DEG10360303; DEG10130019; DEG10190126; DEG10220065; DEG10250230; DEG10280448; DEG10020056; DEG10370219;
aroB 86 3-dehydroquinate synthase	7	DEG10130454; DEG10280002; DEG10050099; DEG10250498; DEG10310131; DEG10100410; DEG10200369;
aroC 125 Chorismate synthase	9	DEG10080112; DEG10050092; DEG10130248; DEG10330069; DEG10360070; DEG10250499; DEG10280523; DEG10100412; DEG10200378;
asd 98 Aspartate-semialdehyde dehydrogenase	16	DEG10010139; DEG10220113; DEG10330298; DEG10130082; DEG10360111; DEG10340074; DEG10050206; DEG10240377; DEG10230221; DEG10160294; DEG10150120; DEG10250719; DEG10100582; DEG10320294; DEG10190236; DEG10180519;
aspS 101 Aspartyl-tRNA synthetase	71	DEG10230041; DEG10290198; DEG10160082; DEG10230108; DEG10270474; DEG10070142; DEG10160126; DEG10150090; DEG10290095; DEG10060109; DEG10030264; DEG10370220; DEG10100563; DEG10010153; DEG10340426; DEG10180315; DEG10330200; DEG10120021; DEG10020181; DEG10270631; DEG10350280; DEG10140128; DEG10380068; DEG10380229; DEG10070078; DEG10280009; DEG10350040; DEG10320174; DEG10320231; DEG10330084; DEG10110114; DEG10240217; DEG10130100; DEG10220240; DEG10180155; DEG10030140; DEG10360151; DEG10370076; DEG10030238; DEG10380076; DEG10330128; DEG10170227; DEG10170035; DEG10130158; DEG10290234; DEG10340280; DEG10110060; DEG10060092; DEG10370069; DEG10080027; DEG10340389; DEG10120036; DEG10230258; DEG10190083; DEG10190125; DEG10220253; DEG10320103; DEG10220043; DEG10160197; DEG10250503; DEG10010018; DEG10010192; DEG10020159; DEG10250697; DEG10020039; DEG10360043; DEG10110164; DEG10210138; DEG10240043; DEG10200246; DEG10060025;
atpA 127 ATP synthase subunit alpha	44	DEG10030559; DEG10380085; DEG10200418; DEG10380087; DEG10060328; DEG10150334; DEG10200416; DEG10360331; DEG10120357; DEG10230082; DEG10120359; DEG10210080; DEG10130026; DEG10350417; DEG10310006; DEG10290397; DEG10350419; DEG10130028; DEG10270237; DEG10250245; DEG10140165; DEG10280104; DEG10280102; DEG10250243; DEG10100207; DEG10240356; DEG10100205; DEG10240358; DEG10060330; DEG10270239; DEG10030561; DEG10070182; DEG10020238; DEG10370084; DEG10070184; DEG10360329; DEG10080206; DEG10140095; DEG10140097; DEG10140079; DEG10370086; DEG10210078; DEG10140273; DEG10290395;
atpD 111 ATP synthase subunit beta	52	DEG10030559; DEG10380085; DEG10200418; DEG10310006; DEG10060328; DEG10150334; DEG10200416; DEG10080206; DEG10120357; DEG10360316; DEG10120359; DEG10210080; DEG10130026; DEG10240272; DEG10350417; DEG10380087; DEG10290397; DEG10290395; DEG10340350; DEG10130028; DEG10230082; DEG10250245; DEG10140165; DEG10280104; DEG10280102;

		DEG10250243; DEG10240358; DEG10240356; DEG10100205; DEG10210078; DEG10100207; DEG10060330; DEG10120308; DEG10100196; DEG10270239; DEG10030561; DEG10070182; DEG10020238; DEG10270230; DEG10270237; DEG10070184; DEG10360329; DEG10360331; DEG10140095; DEG10140097; DEG10250235; DEG10140079; DEG10220347; DEG10370086; DEG10370084; DEG10140273; DEG10350419;
atpG 85 ATP synthase subunit gamma	19	DEG10130027; DEG10350418; DEG10240357; DEG10140096; DEG10250244; DEG10290396; DEG10200417; DEG10100206; DEG10080207; DEG10060329; DEG10360330; DEG10270238; DEG10210079; DEG10030560; DEG10070183; DEG10380086; DEG10280103; DEG10370085; DEG10120358;
carA 93 carbamoyl-phosphate synthase small chain	6	DEG10250259; DEG10130343; DEG10100221; DEG10350093; DEG10240302; DEG10280051;
cysK 78 cysteine synthase	7	DEG10350174; DEG10130192; DEG10120293; DEG10270228; DEG10100194; DEG10250233; DEG10350313;
dapA 74 Dihydrodipicolinate synthase	21	DEG10330063; DEG10270496; DEG10080171; DEG10230047; DEG10350274; DEG10340286; DEG10320197; DEG10290178; DEG10220439; DEG10180377; DEG10160062; DEG10360048; DEG10130493; DEG10200132; DEG10250534; DEG10100437; DEG10280080; DEG10190142; DEG10010140; DEG10050109; DEG10150232;
dapB 72 Dihydrodipicolinate reductase	21	DEG10340026; DEG10190004; DEG10070080; DEG10240136; DEG10220433; DEG10180009; DEG10270499; DEG10200436; DEG10080085; DEG10030471; DEG10050467; DEG10280033; DEG10130496; DEG10150289; DEG10360274; DEG10330006; DEG10320008; DEG10290098; DEG10010156; DEG10250539; DEG10160006;
dnaB 90 Replicative DNA helicase	30	DEG10170008; DEG10330337; DEG10010264; DEG10050567; DEG10230028; DEG10190283; DEG10020007; DEG10370223; DEG10350103; DEG10130289; DEG10290335; DEG10120221; DEG10180580; DEG10030066; DEG10060076; DEG10270016; DEG10340304; DEG10210214; DEG10380232; DEG10140196; DEG10220277; DEG10240260; DEG10160333; DEG10280467; DEG10200206; DEG10250016; DEG10100006; DEG10320337; DEG10070109; DEG10360282;
dnaG 109 DNA primase	25	DEG10120215; DEG10130353; DEG10240367; DEG10100370; DEG10010179; DEG10210086; DEG10060211; DEG10230266; DEG10320247; DEG10170208; DEG10140167; DEG10200373; DEG10370090; DEG10250453; DEG10380091; DEG10050169; DEG10270420; DEG10220377; DEG10180456; DEG10360024; DEG10160216; DEG10020170; DEG10070161; DEG10290119; DEG10330219;
dnaK 239 Chaperone protein DnaK	46	DEG10170214; DEG10240007; DEG10290207; DEG10150262; DEG10250060; DEG10220194; DEG10290096; DEG10240157; DEG10160001; DEG10360275; DEG10210188; DEG10200006; DEG10270062; DEG10230243; DEG10080013; DEG10160241; DEG10180385; DEG10380203; DEG10340449; DEG10350009; DEG10240216; DEG10130207; DEG10100038; DEG10280144; DEG10200186; DEG10330001; DEG10370193; DEG10290351; DEG10190198; DEG10180486; DEG10060246; DEG10220183; DEG10230317; DEG10140074; DEG10360243; DEG10050129; DEG10110001; DEG10020173; DEG10320001; DEG10330244; DEG10180002; DEG10010198; DEG10030073; DEG10360162; DEG10310027; DEG10280075;
efp 133 elongation factor P	15	DEG10140194; DEG10350297; DEG10290217; DEG10180585; DEG10270470; DEG10200101; DEG10060017; DEG10100408; DEG10070028; DEG10240206; DEG10250496; DEG10120006; DEG10030245; DEG10020167; DEG10170204;
engA 105 GTP-binding protein EngA	66	DEG10130096; DEG10030494; DEG10130311; DEG10150286; DEG10200195; DEG10170166; DEG10170194; DEG10060007; DEG10010137; DEG10120258; DEG10010182; DEG10150241; DEG10280349; DEG10220018; DEG10050356; DEG10350005; DEG10230249; DEG10060268; DEG10360036; DEG10240278; DEG10160061; DEG10380040; DEG10220019; DEG10070211; DEG10180396; DEG10140009; DEG10360159; DEG10170349; DEG10140144; DEG10240204; DEG10020161; DEG10140023; DEG10320198; DEG10330062; DEG10240293; DEG10120365; DEG10290279; DEG10060274; DEG10010162; DEG10170210; DEG10340439; DEG10120110; DEG10360266; DEG10050005; DEG10260085; DEG10130059; DEG10180543; DEG10160053; DEG10100299; DEG10320205; DEG10340440; DEG10150084; DEG10230250; DEG10020137; DEG10380059;

		DEG10250362; DEG10370045; DEG10190143; DEG10200205; DEG10180380; DEG10210158; DEG10110143; DEG10190149; DEG10330054; DEG10290131; DEG10380130;
eno 232 enolase	21	DEG10010237; DEG10100156; DEG10380080; DEG10350276; DEG10130245; DEG10320227; DEG10060336; DEG10140199; DEG10290297; DEG10250192; DEG10110158; DEG10170081; DEG10370079; DEG10330196; DEG10020073; DEG10270189; DEG10160193; DEG10210097; DEG10070168; DEG10190165; DEG10120154;
ffh 123 Signal recognition particle protein	56	DEG10240125; DEG10350369; DEG10270524; DEG10100467; DEG10330294; DEG10380066; DEG10280268; DEG10060239; DEG10230083; DEG10350024; DEG10110193; DEG10320217; DEG10280321; DEG10070065; DEG10370131; DEG10080130; DEG10020123; DEG10160183; DEG10340351; DEG10020124; DEG10070070; DEG10230049; DEG10210140; DEG10180523; DEG10180407; DEG10330185; DEG10190158; DEG10120191; DEG10030109; DEG10130268; DEG10170147; DEG10170146; DEG10320298; DEG10340288; DEG10120343; DEG10010122; DEG10010123; DEG10160290; DEG10190240; DEG10240026; DEG10380142; DEG10060035; DEG10050261; DEG10200459; DEG10220251; DEG10290384; DEG10200455; DEG10220042; DEG10210115; DEG10250568; DEG10030018; DEG10370067; DEG10140132; DEG10140264; DEG10290133; DEG10250567;
fmt 116 Methionyl-tRNA formyltransferase	38	DEG10240003; DEG10330331; DEG10060300; DEG10180208; DEG10010113; DEG10340404; DEG10270171; DEG10050569; DEG10310005; DEG10320263; DEG10230268; DEG10380180; DEG10070214; DEG10160327; DEG10250267; DEG10290009; DEG10190203; DEG10030007; DEG10070114; DEG10210163; DEG10290147; DEG10370165; DEG10180489; DEG10100227; DEG10150004; DEG10020090; DEG10140017; DEG10270255; DEG10050518; DEG10220436; DEG10240263; DEG10250176; DEG10280133; DEG10280463; DEG10360007; DEG10130497; DEG10170137; DEG10120183;
folA 71 Dihydrofolate reductase	25	DEG10060193; DEG10120040; DEG10340385; DEG10150012; DEG10010151; DEG10070193; DEG10140203; DEG10250537; DEG10360015; DEG10200310; DEG10280411; DEG10180010; DEG10290317; DEG10130087; DEG10170184; DEG10380113; DEG10220457; DEG10350306; DEG10190005; DEG10370107; DEG10210110; DEG10050323; DEG10030078; DEG10320009; DEG10020156;
folD 126 bifunctional protein folD	20	DEG10290173; DEG10140091; DEG10150189; DEG10270593; DEG10160158; DEG10070147; DEG10130339; DEG10350282; DEG10320063; DEG10060008; DEG10310111; DEG10120105; DEG10360076; DEG10100532; DEG10330161; DEG10240209; DEG10190059; DEG10250662; DEG10110039; DEG10010172;
frr 147 Ribosome-recycling factor (RRF)	30	DEG10230126; DEG10130197; DEG10290153; DEG10270518; DEG10010131; DEG10330037; DEG10120044; DEG10050292; DEG10340098; DEG10150096; DEG10320036; DEG10200260; DEG10190031; DEG10240237; DEG10250559; DEG10070051; DEG10210145; DEG10280062; DEG10370059; DEG10060354; DEG10030447; DEG10100457; DEG10310023; DEG10160036; DEG10380056; DEG10140072; DEG10170158; DEG10360146; DEG10220391; DEG10180042;
ftsH 90 cell division protein	26	DEG10190184; DEG10080192; DEG10120163; DEG10200390; DEG10380004; DEG10360272; DEG10330229; DEG10250396; DEG10140307; DEG10160226; DEG10100569; DEG10110173; DEG10290109; DEG10240299; DEG10340120; DEG10130342; DEG10350095; DEG10350460; DEG10250704; DEG10370004; DEG10220007; DEG10230277; DEG10280402; DEG10020038; DEG10060369; DEG10030130;
ftsY 91 cell division protein FtsY	58	DEG10230049; DEG10210115; DEG10170147; DEG10240026; DEG10330294; DEG10280268; DEG10060239; DEG10070070; DEG10350024; DEG10110193; DEG10200459; DEG10280321; DEG10020124; DEG10350369; DEG10370131; DEG10080130; DEG10020123; DEG10160183; DEG10340351; DEG10380066; DEG10230083; DEG10240125; DEG10210140; DEG10180523; DEG10180407; DEG10320298; DEG10190240; DEG10120191; DEG10140264; DEG10130268; DEG10270524; DEG10170146; DEG10330185; DEG10340288; DEG10240293; DEG10120343; DEG10010122; DEG10010123; DEG10160290; DEG10190158; DEG10380142; DEG10100467; DEG10060035; DEG10050261; DEG10320217; DEG10220251; DEG10290384; DEG10200455; DEG10220042; DEG10070065; DEG10250568; DEG10030018; DEG10370067; DEG10140132; DEG10030135; DEG10030109; DEG10290133; DEG10250567;

ftsZ 184 Cell division protein ftsZ	31	DEG10340057; DEG10290360; DEG10350376; DEG10030475; DEG10070204; DEG10310088; DEG10140186; DEG10230203; DEG10160021; DEG10190018; DEG10110013; DEG10200339; DEG10370157; DEG10100322; DEG10210062; DEG10380170; DEG10330022; DEG10240115; DEG10050413; DEG10120032; DEG10010109; DEG10080167; DEG10360220; DEG10060191; DEG10180024; DEG10280449; DEG10130478; DEG10250404; DEG10020110; DEG10320022; DEG10170131;
fusA 171 Elongation factor G	86	DEG10150286; DEG10020086; DEG10210198; DEG10100449; DEG10350413; DEG10180508; DEG10180509; DEG10340483; DEG10110170; DEG10340049; DEG10120051; DEG10060114; DEG10120365; DEG10360266; DEG10350405; DEG10350406; DEG10180471; DEG10340492; DEG10250133; DEG10250132; DEG10020174; DEG10140150; DEG10030135; DEG10230160; DEG10110190; DEG10010032; DEG10160222; DEG10180568; DEG10370036; DEG10280414; DEG10320291; DEG10120342; DEG10220335; DEG10160298; DEG10230195; DEG10300067; DEG10260053; DEG10140071; DEG10100100; DEG10130059; DEG10020051; DEG10020052; DEG10190182; DEG10170048; DEG10170049; DEG10010137; DEG10350216; DEG10220422; DEG10280185; DEG10010033; DEG10270119; DEG10130160; DEG10140160; DEG10330302; DEG10170166; DEG10270506; DEG10200381; DEG10100099; DEG10380031; DEG10050188; DEG10220060; DEG10020137; DEG10380190; DEG10290114; DEG10320249; DEG10130313; DEG10050638; DEG10060366; DEG10330225; DEG10130146; DEG10370177; DEG10200009; DEG10240293; DEG10370071; DEG10070012; DEG10290030; DEG10060071; DEG10290087; DEG10230154; DEG10020099; DEG10270120; DEG10220276; DEG10360202; DEG10190231; DEG10210136; DEG10250549;
gap 122 glyceraldehyde-3-phosphate dehydrogenase	26	DEG10060242; DEG10230289; DEG10100232; DEG10270261; DEG10020194; DEG10280266; DEG10220016; DEG10340137; DEG10370037; DEG10250274; DEG10170077; DEG10160091; DEG10130309; DEG10210197; DEG10050001; DEG10180303; DEG10190123; DEG10380032; DEG10360021; DEG10130176; DEG10140018; DEG10320124; DEG10310185; DEG10330093; DEG10020070; DEG10070229;
glmS 87 glucosamine--fructose-6-phosphate	27	DEG10100542; DEG10150333; DEG10370138; DEG10120122; DEG10250670; DEG10020242; DEG10180545; DEG10200027; DEG10290391; DEG10190260; DEG10070011; DEG10070113; DEG10170302; DEG10240015; DEG10280494; DEG10130187; DEG10380152; DEG10010067; DEG10360327; DEG10210196; DEG10250154; DEG10350017; DEG10270602; DEG10270149; DEG10310182; DEG10100131; DEG10320312;
gltA 125 Citrate synthase	6	DEG10270203; DEG10130349; DEG10250165; DEG10240376; DEG10350491; DEG10280364;
gltX1 129 glutamyl-tRNA synthetase	49	DEG10050487; DEG10130234; DEG10030209; DEG10270537; DEG10230068; DEG10230162; DEG10240227; DEG10340318; DEG10050111; DEG10190138; DEG10150119; DEG10280365; DEG10220425; DEG10320075; DEG10210203; DEG10150191; DEG10360112; DEG10160067; DEG10370032; DEG10380029; DEG10160144; DEG10020041; DEG10070232; DEG10180117; DEG10140266; DEG10130461; DEG10330147; DEG10360074; DEG10100479; DEG10290169; DEG10170036; DEG10330068; DEG10120137; DEG10320192; DEG10120039; DEG10250585; DEG10350228; DEG10030428; DEG10010021; DEG10310173; DEG10180366; DEG10280246; DEG10350266; DEG10340494; DEG10220100; DEG10200249; DEG10060373; DEG10110046; DEG10190068;
glyA 251 Serine hydroxymethyltransferase	17	DEG10350437; DEG10350340; DEG10360253; DEG10250204; DEG10120284; DEG10130263; DEG10150275; DEG10280131; DEG10010253; DEG10020234; DEG10270194; DEG10240161; DEG10160056; DEG10140131; DEG10060325; DEG10330057; DEG10180392;
gmk 99 guanylate kinase	27	DEG10130451; DEG10230101; DEG10010111; DEG10120168; DEG10340379; DEG10180540; DEG10140269; DEG10380181; DEG10060088; DEG10080056; DEG10250261; DEG10200213; DEG10240176; DEG10210164; DEG10320306; DEG10290064; DEG10100223; DEG10330282; DEG10050642; DEG10360324; DEG10160278; DEG10350326; DEG10280276; DEG10220294; DEG10370166; DEG10170134; DEG10190251;
greA 82 Transcription elongation factor GreA	8	DEG10170220; DEG10250200; DEG10070073; DEG10140210; DEG10310029; DEG10020176; DEG10080151; DEG10060232;

groEL 105 Chaperonin	26	DEG10170284; DEG10060323; DEG10010077; DEG10360216; DEG10350338; DEG10270080; DEG10340356; DEG10380223; DEG10320342; DEG10030742; DEG10180584; DEG10100059; DEG10250085; DEG10120323; DEG10110220; DEG10290080; DEG10330342; DEG10100537; DEG10240163; DEG10230091; DEG10210039; DEG10220290; DEG10160337; DEG10370214; DEG10070105; DEG10200093;
groEL1 125 Chaperonin GroEL	26	DEG10170284; DEG10060323; DEG10010077; DEG10360216; DEG10350338; DEG10270080; DEG10340356; DEG10380223; DEG10320342; DEG10030742; DEG10180584; DEG10100059; DEG10250085; DEG10120323; DEG10110220; DEG10290080; DEG10330342; DEG10100537; DEG10200093; DEG10230091; DEG10210039; DEG10220290; DEG10160337; DEG10370214; DEG10070105; DEG10240163;
guaA 142 GMP synthase	24	DEG10170020; DEG10270006; DEG10050500; DEG10280368; DEG10270597; DEG10020024; DEG10250005; DEG10360157; DEG10120208; DEG10100221; DEG10350093; DEG10050104; DEG10250666; DEG10350246; DEG10100534; DEG10250259; DEG10130295; DEG10240249; DEG10130018; DEG10080069; DEG10220125; DEG10280158; DEG10240302; DEG10280051;
guaB 93 Inosine-5'-monophosphate dehydrogenase	19	DEG10240247; DEG10150086; DEG10220288; DEG10070111; DEG10020023; DEG10240037; DEG10270599; DEG10100536; DEG10080148; DEG10250667; DEG10360158; DEG10280044; DEG10250668; DEG10010005; DEG10270598; DEG10130475; DEG10310139; DEG10350247; DEG10200324;
gyrA 158 DNA gyrase subunit A	50	DEG10060003; DEG10200196; DEG10290228; DEG10320187; DEG10170177; DEG10120126; DEG10230048; DEG10170005; DEG10330078; DEG10200197; DEG10270005; DEG10150117; DEG10180351; DEG10160208; DEG10190132; DEG10180449; DEG10020004; DEG10140142; DEG10370111; DEG10110131; DEG10280495; DEG10280027; DEG10210120; DEG10210122; DEG10250004; DEG10290331; DEG10130327; DEG10140048; DEG10060172; DEG10330211; DEG10340287; DEG10030489; DEG10150298; DEG10220082; DEG10350321; DEG10010149; DEG10380137; DEG10220187; DEG10010004; DEG10130032; DEG10230219; DEG10370127; DEG10160075; DEG10120318; DEG10240181; DEG10320241; DEG10020150; DEG10380117; DEG10360286; DEG10360128;
gyrB 121 DNA gyrase subunit B	57	DEG10030490; DEG10340055; DEG10170176; DEG10060002; DEG10170004; DEG10270004; DEG10120327; DEG10190175; DEG10050551; DEG10160209; DEG10200287; DEG10350001; DEG10230200; DEG10220076; DEG10370110; DEG10020003; DEG10130002; DEG10140141; DEG10250003; DEG10020149; DEG10370077; DEG10210123; DEG10290006; DEG10290333; DEG10030004; DEG10320307; DEG10240353; DEG10280496; DEG10060171; DEG10350069; DEG10330281; DEG10150299; DEG10130487; DEG10220341; DEG10010148; DEG10380116; DEG10160277; DEG10210096; DEG10120150; DEG10110200; DEG10010003; DEG10380078; DEG10230175; DEG10070149; DEG10050180; DEG10340543; DEG10180452; DEG10200030; DEG10330212; DEG10360003; DEG10070042; DEG10280291; DEG10320242; DEG10190253; DEG10360287; DEG10100002; DEG10140293;
hemE 180 uroporphyrinogen decarboxylase	17	DEG10030053; DEG10350416; DEG10330260; DEG10240355; DEG10270485; DEG10320333; DEG10200480; DEG10180574; DEG10290069; DEG10130299; DEG10150309; DEG10280295; DEG10160257; DEG10250521; DEG10120367; DEG10110214; DEG10360302;
hisD 98 histidinol dehydrogenase	5	DEG10250321; DEG10130112; DEG10280286; DEG10270301; DEG10100258;
hisF 79 Imidazole glycerol phosphate synthase subunit	10	DEG10360119; DEG10100263; DEG10100262; DEG10050161; DEG10250325; DEG10280279; DEG10280280; DEG10130468; DEG10130467; DEG10250326;
hisG 73 ATP phosphoribosyl transferase	6	DEG10100317; DEG10250397; DEG10050158; DEG10130111; DEG10280287; DEG10270370;
hisS 151 histidyl-tRNA synthetase	30	DEG10130095; DEG10270475; DEG10060024; DEG10220456; DEG10020182; DEG10340359; DEG10240276; DEG10160060; DEG10370221; DEG10210142; DEG10140129; DEG10320199; DEG10330061; DEG10170228; DEG10190144; DEG10200424; DEG10120363; DEG10210212; DEG10080218; DEG10100416; DEG10230092; DEG10380230; DEG10350115; DEG10250504; DEG10280507;

		DEG10180381; DEG10290282; DEG10360160; DEG10010193; DEG10110144;
ileS 129 isoleucyl-tRNA synthetase	91	DEG10330352; DEG10290306; DEG10160148; DEG10240062; DEG10230300; DEG10340272; DEG10280301; DEG10130006; DEG10290105; DEG10170264; DEG10210040; DEG10030147; DEG10060284; DEG10120199; DEG10160079; DEG10070203; DEG10220258; DEG10180006; DEG10320184; DEG10310102; DEG10010110; DEG10020186; DEG10120108; DEG10240137; DEG10170240; DEG10360173; DEG10320349; DEG10210063; DEG10010218; DEG10190066; DEG10380176; DEG10270449; DEG10130366; DEG10120038; DEG10030503; DEG10080214; DEG10360249; DEG10020210; DEG10060222; DEG10280409; DEG10370028; DEG10110126; DEG10360169; DEG10010199; DEG10290288; DEG10170133; DEG10330151; DEG10380169; DEG10050502; DEG10160347; DEG10350060; DEG10320072; DEG10220171; DEG10370155; DEG10210162; DEG10250479; DEG10060273; DEG10350223; DEG10330003; DEG10150271; DEG10060013; DEG10230038; DEG10190293; DEG10140090; DEG10350361; DEG10100396; DEG10320005; DEG10370162; DEG10140258; DEG10200097; DEG10180599; DEG10220095; DEG10010010; DEG10160003; DEG10310141; DEG10380024; DEG10050332; DEG10180113; DEG10330081; DEG10280125; DEG10270288; DEG10200161; DEG10140025; DEG10020112; DEG10340148; DEG10180342; DEG10190129; DEG10130395; DEG10110002; DEG10250305; DEG10150075;
ilvA 72 Threonine dehydratase	14	DEG10130192; DEG10100268; DEG10130038; DEG10130102; DEG10250330; DEG10270228; DEG10100194; DEG10250233; DEG10280282; DEG10070217; DEG10270292; DEG10250310; DEG10270307; DEG10050519;
ilvC 117 ketol-acid reductoisomerase	8	DEG10100483; DEG10350074; DEG10240080; DEG10130393; DEG10270541; DEG10200308; DEG10280099; DEG10250588;
ilvD 89 Dihydroxy-acid dehydratase	7	DEG10350063; DEG10240068; DEG10280502; DEG10100014; DEG10300056; DEG10250028; DEG10270032;
infA 75 translation initiation factor IF-1	25	DEG10170312; DEG10050175; DEG10010058; DEG10100549; DEG10060144; DEG10320086; DEG10250679; DEG10330142; DEG10210027; DEG10030338; DEG10190071; DEG10240318; DEG10120192; DEG10130081; DEG10150146; DEG10340459; DEG10160139; DEG10080250; DEG10020256; DEG10140219; DEG10370020; DEG10290248; DEG10280500; DEG10200327; DEG10220399;
infB 144 translation initiation factor IF-2	100	DEG10150286; DEG10370177; DEG10020086; DEG10210198; DEG10270026; DEG10100449; DEG10350413; DEG10270156; DEG10180508; DEG10340483; DEG10280309; DEG10100036; DEG10110170; DEG10340049; DEG10120051; DEG10060114; DEG10120365; DEG10360266; DEG10250009; DEG10270346; DEG10350111; DEG10350405; DEG10180471; DEG10340492; DEG10250133; DEG10250132; DEG10020174; DEG10140150; DEG10250685; DEG10250684; DEG10030135; DEG10230160; DEG10110190; DEG10280185; DEG10180568; DEG10370036; DEG10270679; DEG10280414; DEG10320291; DEG10240038; DEG10120342; DEG10220335; DEG10270010; DEG10160298; DEG10230195; DEG10200381; DEG10320249; DEG10250057; DEG10140071; DEG10100100; DEG10130059; DEG10020051; DEG10020052; DEG10190182; DEG10170048; DEG10170049; DEG10010137; DEG10350216; DEG10220422; DEG10010032; DEG10010033; DEG10200394; DEG10270119; DEG10270610; DEG10270611; DEG10270612; DEG10160222; DEG10130160; DEG10140160; DEG10330302; DEG10170166; DEG10270506; DEG10100099; DEG10270627; DEG10220276; DEG10220060; DEG10020137; DEG10380190; DEG10290114; DEG10260053; DEG10130313; DEG10060366; DEG10330225; DEG10050236; DEG10270048; DEG10200009; DEG10240293; DEG10370071; DEG10070012; DEG10290030; DEG10060071; DEG10290087; DEG10270059; DEG10230154; DEG10020099; DEG10350030; DEG10270120; DEG10380031; DEG10210136; DEG10250549;
infC 121 translation initiation factor IF-3	27	DEG10030595; DEG10100277; DEG10060164; DEG10010207; DEG10220197; DEG10380102; DEG10360093; DEG10080017; DEG10280284; DEG10200124; DEG10340212; DEG10130386; DEG10070155; DEG10170246; DEG10160094; DEG10290211; DEG10230016; DEG10050473; DEG10120269; DEG10020191; DEG10190120; DEG10140092; DEG10210135; DEG10330096; DEG10180288; DEG10250342; DEG10320128;
katA 74 catalase	1	DEG10260019;

ksgA 122 Dimethyladenosine transferase	7	DEG10290316; DEG10050176; DEG10270183; DEG10030079; DEG10220031; DEG10310225; DEG10300018;
ldh 77 L-lactate dehydrogenase	4	DEG10050433; DEG10280153; DEG10020293; DEG10200456;
lepA 82 GTP-binding protein LepA	89	DEG10150286; DEG10020086; DEG10210198; DEG10100449; DEG10350413; DEG10270224; DEG10180508; DEG10180509; DEG10250229; DEG10340483; DEG10110170; DEG10340049; DEG10100190; DEG10120051; DEG10060114; DEG10120365; DEG10360266; DEG10350405; DEG10350406; DEG10180471; DEG10340492; DEG10250133; DEG10250132; DEG10020174; DEG10020137; DEG10030135; DEG10230160; DEG10110190; DEG10280185; DEG10010033; DEG10180568; DEG10370036; DEG10320291; DEG10120342; DEG10220335; DEG10160298; DEG10230195; DEG10200381; DEG10260053; DEG10140071; DEG10100100; DEG10130059; DEG10020051; DEG10020052; DEG10190182; DEG10170048; DEG10170049; DEG10010137; DEG10350216; DEG10220422; DEG10010032; DEG10160222; DEG10270119; DEG10130160; DEG10140160; DEG10330302; DEG10170166; DEG10270506; DEG10300067; DEG10100099; DEG10220276; DEG10050188; DEG10220060; DEG10140150; DEG10380190; DEG10290114; DEG10320249; DEG10130313; DEG10050638; DEG10060366; DEG10330225; DEG10050236; DEG10130146; DEG10370177; DEG10200009; DEG10240293; DEG10370071; DEG10070012; DEG10290030; DEG10060071; DEG10290087; DEG10230154; DEG10020099; DEG10270120; DEG10380031; DEG10360202; DEG10190231; DEG10210136; DEG10250549;
leuB 89 3-isopropylmalate dehydrogenase	12	DEG10240059; DEG10100480; DEG10130080; DEG10330110; DEG10270538; DEG10280489; DEG10350057; DEG10250586; DEG10260017; DEG10160108; DEG10050347; DEG10110075;
leuC 80 3-isopropylmalate dehydratase large subunit	16	DEG10260002; DEG10120348; DEG10310107; DEG10270536; DEG10280512; DEG10240369; DEG10290068; DEG10360072; DEG10200458; DEG10250584; DEG10270272; DEG10100247; DEG10050348; DEG10350493; DEG10130078; DEG10250287;
leuS 143 leucyl-tRNA synthetase	41	DEG10160148; DEG10330151; DEG10170264; DEG10060273; DEG10380169; DEG10250186; DEG10230300; DEG10200472; DEG10380024; DEG10320072; DEG10220171; DEG10050332; DEG10180113; DEG10280301; DEG10290105; DEG10370155; DEG10120199; DEG10360173; DEG10210040; DEG10210063; DEG10010218; DEG10240137; DEG10130366; DEG10350361; DEG10060222; DEG10140025; DEG10270182; DEG10270013; DEG10230180; DEG10020210; DEG10120117; DEG10340547; DEG10370028; DEG10100151; DEG10130395; DEG10070203; DEG10250012; DEG10190066; DEG10100005; DEG10210150; DEG10280074;
lysA 84 diaminopimelate decarboxylase	9	DEG10280304; DEG10100192; DEG10080050; DEG10250231; DEG10280503; DEG10270226; DEG10340035; DEG10130332; DEG10200058;
lysC 91 Aspartate kinase	16	DEG10050040; DEG10270648; DEG10360039; DEG10130174; DEG10030465; DEG10290290; DEG10280491; DEG10200111; DEG10240225; DEG10350268; DEG10150095; DEG10360147; DEG10250720; DEG10100583; DEG10080227; DEG10220008;
metG 119 methionyl-tRNA synthetase	53	DEG10160148; DEG10330151; DEG10320184; DEG10170025; DEG10380169; DEG10010199; DEG10010010; DEG10020186; DEG10100151; DEG10140185; DEG10380024; DEG10320072; DEG10220171; DEG10170240; DEG10050332; DEG10180113; DEG10330081; DEG10280301; DEG10130122; DEG10240137; DEG10290105; DEG10190066; DEG10370053; DEG10010218; DEG10140025; DEG10020210; DEG10370155; DEG10250186; DEG10200182; DEG10050455; DEG10380050; DEG10120117; DEG10360173; DEG10270182; DEG10340547; DEG10060013; DEG10200472; DEG10230180; DEG10120199; DEG10220048; DEG10160079; DEG10180342; DEG10210040; DEG10350361; DEG10370028; DEG10190129; DEG10130395; DEG10110126; DEG10020031; DEG10170264; DEG10210150; DEG10290247; DEG10280074;
metK 121 S-adenosylmethionine synthase	30	DEG10340012; DEG10130250; DEG10120331; DEG10290091; DEG10080031; DEG10230184; DEG10320238; DEG10190171; DEG10160205; DEG10250264; DEG10330208; DEG10170266; DEG10220389; DEG10010219; DEG10380159; DEG10240011; DEG10100226; DEG10030087; DEG10350014; DEG10060034; DEG10280374; DEG10020211; DEG10270253; DEG10180439; DEG10070146;

		DEG10200012; DEG10210133; DEG10370145; DEG10110166; DEG10140274;
miaA 92 tRNA dimethylallyltransferase	12	DEG10270492; DEG10130278; DEG10050031; DEG10330347; DEG10180589; DEG10100432; DEG10120240; DEG10280097; DEG10250530; DEG10200305; DEG10160342; DEG10290077;
murA 100 UDP-N-acetylglucosamine	25	DEG10260011; DEG10080106; DEG10290344; DEG10200439; DEG10190189; DEG10340375; DEG10170295; DEG10270240; DEG10360236; DEG10320254; DEG10250246; DEG10070059; DEG10100209; DEG10320092; DEG10020231; DEG10010252; DEG10030507; DEG10120111; DEG10220236; DEG10160232; DEG10130110; DEG10230099; DEG10330235; DEG10200328; DEG10280119;
ndk 162 nucleoside diphosphate kinase	9	DEG10150082; DEG10350113; DEG10030162; DEG10290209; DEG10240274; DEG10280006; DEG10200220; DEG10340412; DEG10180383;
nth 73 Endonuclease III	1	DEG10050614;
nusA 73 Transcription elongation protein	28	DEG10190183; DEG10010136; DEG10350217; DEG10080305; DEG10370178; DEG10330226; DEG10160223; DEG10110171; DEG10070137; DEG10250550; DEG10170163; DEG10340048; DEG10270507; DEG10220059; DEG10240294; DEG10060112; DEG10230194; DEG10100450; DEG10120366; DEG10360267; DEG10200010; DEG10130058; DEG10020136; DEG10380191; DEG10140070; DEG10290113; DEG10310049; DEG10030134;
nusG 121 Transcription anti-termination protein NusG	20	DEG10200385; DEG10330266; DEG10030046; DEG10350412; DEG10270112; DEG10280259; DEG10240346; DEG10340490; DEG10310052; DEG10050240; DEG10250122; DEG10290021; DEG10230159; DEG10130047; DEG10190275; DEG10160263; DEG10360211; DEG10120340; DEG10080222; DEG10220333;
obgE 122 GTPase ObgE	30	DEG10200195; DEG10190185; DEG10170235; DEG10240122; DEG10350373; DEG10210085; DEG10020183; DEG10200048; DEG10160228; DEG10140057; DEG10250476; DEG10110174; DEG10280434; DEG10070058; DEG10060316; DEG10270446; DEG10380157; DEG10130308; DEG10120388; DEG10290318; DEG10220166; DEG10180201; DEG10180473; DEG10100391; DEG10370143; DEG10020018; DEG10330231; DEG10070001; DEG10010194; DEG10140136;
pgi 73 glucose-6-phosphate isomerase	16	DEG10210205; DEG10340391; DEG10020081; DEG10290311; DEG10380026; DEG10260101; DEG10070233; DEG10200034; DEG10220238; DEG10250171; DEG10230110; DEG10120162; DEG10170096; DEG10140040; DEG10100140; DEG10370030;
pgk 194 phosphoglycerate kinase	23	DEG10130237; DEG10100233; DEG10010240; DEG10290093; DEG10140086; DEG10160203; DEG10230087; DEG10320237; DEG10340353; DEG10370204; DEG10170078; DEG10070030; DEG10330206; DEG10210041; DEG10220080; DEG10050168; DEG10060241; DEG10030088; DEG10180437; DEG10270262; DEG10250275; DEG10020071; DEG10190170;
pheA 69 Prephenate dehydratase	5	DEG10310031; DEG10280509; DEG10130259; DEG10050414; DEG10250755;
pheS 121 phenylalanyl-tRNA synthetase subunit alpha	29	DEG10060162; DEG10380088; DEG10010204; DEG10100278; DEG10360091; DEG10050468; DEG10200121; DEG10280242; DEG10190118; DEG10230086; DEG10130383; DEG10340352; DEG10320131; DEG10020101; DEG10290190; DEG10170120; DEG10330100; DEG10120202; DEG10160098; DEG10150140; DEG10210098; DEG10220370; DEG10350220; DEG10270315; DEG10080066; DEG10250344; DEG10030248; DEG10370087; DEG10140134;
pnp 176 Polyribonucleotide nucleotidyltransferase	30	DEG10100275; DEG10330136; DEG10130272; DEG10200437; DEG10290226; DEG10350318; DEG10200007; DEG10180469; DEG10250540; DEG10370112; DEG10260064; DEG10250339; DEG10190076; DEG10210121; DEG10030375; DEG10240184; DEG10270500; DEG10320095; DEG10160134; DEG10340147; DEG10360265; DEG10380118; DEG10050436; DEG10020139; DEG10050181; DEG10350077; DEG10240084; DEG10360126; DEG10130071; DEG10290116;
polA 73 DNA polymerase I	16	DEG10100274; DEG10380025; DEG10290388; DEG10370029; DEG10270313; DEG10340018; DEG10300115; DEG10160269; DEG10330273; DEG10060219; DEG10250338; DEG10130376; DEG10020195; DEG10140287; DEG10110207;

		DEG10210008;
ppa 76 inorganic pyrophosphatase	16	DEG10200011; DEG10130033; DEG10330351; DEG10180595; DEG10050058; DEG10140175; DEG10120212; DEG10030510; DEG10360179; DEG10320347; DEG10250708; DEG10060290; DEG10160346; DEG10270637; DEG10190292; DEG10240074;
prfA 75 Peptide chain release factor 1	51	DEG10340132; DEG10160089; DEG10230286; DEG10340070; DEG10320161; DEG10360257; DEG10100500; DEG10050564; DEG10170297; DEG10220446; DEG10250614; DEG10130286; DEG10200238; DEG10220263; DEG10350279; DEG10320232; DEG10060216; DEG10330201; DEG10070056; DEG10140028; DEG10370075; DEG10170071; DEG10070038; DEG10020065; DEG10360152; DEG10150279; DEG10100198; DEG10270559; DEG10120325; DEG10380135; DEG10110165; DEG10230217; DEG10020235; DEG10010255; DEG10270231; DEG10120035; DEG10030423; DEG10210093; DEG10160198; DEG10370126; DEG10200114; DEG10250236; DEG10380074; DEG10210114; DEG10010242; DEG10240308; DEG10280423; DEG10330091; DEG10290326; DEG10130472; DEG10190105;
proA 83 gamma-glutamyl phosphate reductase	5	DEG10220146; DEG10240139; DEG10130093; DEG10350360; DEG10280229;
proS 74 prolyl-tRNA synthetase	44	DEG10030594; DEG10240124; DEG10030179; DEG10220196; DEG10010134; DEG10340293; DEG10220203; DEG10360094; DEG10350370; DEG10160093; DEG10160049; DEG10340211; DEG10380220; DEG10130123; DEG10110078; DEG10050493; DEG10180054; DEG10120270; DEG10290274; DEG10170161; DEG10120305; DEG10270508; DEG10230053; DEG10320127; DEG10230015; DEG10280231; DEG10210193; DEG10100451; DEG10360042; DEG10290210; DEG10150236; DEG10350221; DEG10190121; DEG10050246; DEG10320048; DEG10250551; DEG10020134; DEG10190044; DEG10330095; DEG10180289; DEG10200267; DEG10370210; DEG10330050; DEG10070124;
prsA 143 Ribose-phosphate pyrophosphokinase	28	DEG10120236; DEG10240029; DEG10060045; DEG10160085; DEG10170027; DEG10250190; DEG10130356; DEG10010013; DEG10350025; DEG10380006; DEG10340271; DEG10220013; DEG10330087; DEG10210006; DEG10290330; DEG10100154; DEG10230034; DEG10050578; DEG10270187; DEG10360260; DEG10370006; DEG10180204; DEG10080122; DEG10190101; DEG10020034; DEG10320165; DEG10200079; DEG10140039;
purA 128 Adenylo succinate synthetase	10	DEG10250063; DEG10130178; DEG10350117; DEG10270064; DEG10240279; DEG10100041; DEG10030539; DEG10280506; DEG10360284; DEG10220310;
purD 81 Phosphoribosylamine--glycine ligase	13	DEG10130292; DEG10330246; DEG10150294; DEG10050320; DEG10270137; DEG10250149; DEG10120148; DEG10030037; DEG10320260; DEG10100125; DEG10160243; DEG10360280; DEG10310164;
purE 84 Phosphoribosyl amino imidazole carboxylase	4	DEG10340513; DEG10220210; DEG10100525; DEG10280351;
purF 72 amidophosphoribosyltransferase	33	DEG10250758; DEG10100542; DEG10100607; DEG10150333; DEG10290391; DEG10120122; DEG10250670; DEG10020242; DEG10180545; DEG10270676; DEG10200027; DEG10370138; DEG10250431; DEG10190260; DEG10070011; DEG10100131; DEG10170302; DEG10240015; DEG10280494; DEG10130187; DEG10380152; DEG10100345; DEG10010067; DEG10360327; DEG10210196; DEG10270602; DEG10350017; DEG10250154; DEG10270149; DEG10310182; DEG10070113; DEG10320312; DEG10270393;
purH 80 bifunctional	8	DEG10130291; DEG10270172; DEG10250177; DEG10260100; DEG10220182; DEG10340450; DEG10280056; DEG10100145;
pyk 205 Pyruvate kinase	13	DEG10140083; DEG10300074; DEG10270310; DEG10060183; DEG10070152; DEG10380153; DEG10370139; DEG10170252; DEG10050566; DEG10020196; DEG10210090; DEG10250333; DEG10100271;
pyrB 106 aspartate carbamoyltransferase	9	DEG10130179; DEG10130195; DEG10280190; DEG10100218; DEG10250256; DEG10100284; DEG10280094; DEG10270250; DEG10240054;

pyrD 85 Dihydroorotate dehydrogenase 2	6	DEG10260012; DEG10130186; DEG10350211; DEG10250401; DEG10300053; DEG10270373;
pyrH 124 uridylate kinase	31	DEG10230125; DEG10130196; DEG10270519; DEG10290152; DEG10120043; DEG10330036; DEG10320035; DEG10360039; DEG10150095; DEG10070153; DEG10350258; DEG10110017; DEG10190030; DEG10240236; DEG10210146; DEG10370058; DEG10060353; DEG10340100; DEG10050387; DEG10310170; DEG10100458; DEG10030448; DEG10160035; DEG10380055; DEG10140073; DEG10200261; DEG10360147; DEG10220392; DEG10250560; DEG10180041; DEG10170157;
recA 105 recombinase A	5	DEG10020142; DEG10280079; DEG10160189; DEG10080023; DEG10330191;
rho 105 Transcription termination factor Rho	54	DEG10200477; DEG10030559; DEG10100394; DEG10380085; DEG10190182; DEG10310006; DEG10060328; DEG10170196; DEG10200416; DEG10350417; DEG10080206; DEG10120357; DEG10140097; DEG10350216; DEG10360316; DEG10180471; DEG10210080; DEG10240272; DEG10220347; DEG10380087; DEG10290395; DEG10340350; DEG10130028; DEG10230082; DEG10250245; DEG10140165; DEG10280104; DEG10250477; DEG10280102; DEG10290065; DEG10240356; DEG10270447; DEG10100207; DEG10060330; DEG10230195; DEG10120308; DEG10100196; DEG10270239; DEG10070182; DEG10020238; DEG10270230; DEG10210078; DEG10070184; DEG10360329; DEG10350110; DEG10140095; DEG10160253; DEG10250235; DEG10140079; DEG10330256; DEG10370086; DEG10370084; DEG10320314; DEG10030039;
rnc 81 Ribonuclease III	11	DEG10180397; DEG10270525; DEG10100468; DEG10020121; DEG10320206; DEG10010120; DEG10190150; DEG10250569; DEG10080111; DEG10050006; DEG10290130;
rpe 80 Ribulose-phosphate 3-epimerase	15	DEG10360029; DEG10070087; DEG10290060; DEG10050179; DEG10220096; DEG10130116; DEG10150022; DEG10200023; DEG10070227; DEG10180517; DEG10170139; DEG10280217; DEG10270256; DEG10250268; DEG10080276;
rplA 115 50S ribosomal protein L1	20	DEG10290023; DEG10340488; DEG10230158; DEG10170041; DEG10380054; DEG10120338; DEG10060064; DEG10240344; DEG10030048; DEG10360209; DEG10150031; DEG10370057; DEG10250124; DEG10210147; DEG10020044; DEG10220331; DEG10130049; DEG10280261; DEG10140007; DEG10010025;
rplB 126 50S ribosomal protein L2	30	DEG10010038; DEG10360199; DEG10030534; DEG10160303; DEG10370012; DEG10190226; DEG10180503; DEG10280180; DEG10330307; DEG10240333; DEG10120056; DEG10130429; DEG10050268; DEG10230151; DEG10110187; DEG10220417; DEG10380011; DEG10020275; DEG10060125; DEG10350402; DEG10270126; DEG10340478; DEG10200138; DEG10100106; DEG10290035; DEG10250137; DEG10210013; DEG10170331; DEG10140238; DEG10320286;
rplC 124 50S ribosomal protein L3	30	DEG10130432; DEG10310224; DEG10060122; DEG10220420; DEG10030537; DEG10010035; DEG10370010; DEG10160300; DEG10190229; DEG10340481; DEG10280183; DEG10180506; DEG10290032; DEG10140241; DEG10330304; DEG10240336; DEG10120053; DEG10230153; DEG10110189; DEG10350404; DEG10380010; DEG10200135; DEG10360201; DEG10270124; DEG10020278; DEG10320289; DEG10210010; DEG10100103; DEG10250135; DEG10170334;
rplD 113 50S ribosomal protein L4	30	DEG10130431; DEG10280182; DEG10060123; DEG10310223; DEG10030536; DEG10010036; DEG10370011; DEG10080262; DEG10160301; DEG10190228; DEG10340480; DEG10180505; DEG10330305; DEG10140240; DEG10250136; DEG10240335; DEG10120054; DEG10230152; DEG10110188; DEG10200136; DEG10020277; DEG10350403; DEG10210011; DEG10360200; DEG10270125; DEG10220419; DEG10320288; DEG10100104; DEG10290033; DEG10170333;
rplE 123 50S ribosomal protein L5	27	DEG10240326; DEG10150044; DEG10120065; DEG10230149; DEG10360196; DEG10050277; DEG10020266; DEG10210020; DEG10340469; DEG10290044; DEG10100115; DEG10320277; DEG10200147; DEG10140229; DEG10130420; DEG10170322; DEG10010047; DEG10060134; DEG10110185; DEG10030525; DEG10350400; DEG10080256; DEG10160312; DEG10190217; DEG10280171; DEG10220409; DEG10330316;
rplF 107 50S ribosomal protein	31	DEG10240324; DEG10170319; DEG10130417; DEG10180496; DEG10360194; DEG10230148; DEG10020263; DEG10120068; DEG10110184; DEG10340466;

L6		DEG10220406; DEG10370016; DEG10210022; DEG10270131; DEG10140226; DEG10320274; DEG10280168; DEG10100117; DEG10350399; DEG10150046; DEG10290047; DEG10060137; DEG10050280; DEG10010050; DEG10030522; DEG10380015; DEG10250143; DEG10160315; DEG10200150; DEG10190214; DEG10330319;
rplI 126 50S ribosomal protein L9	5	DEG10030065; DEG10020006; DEG10140197; DEG10060075; DEG10010265;
rplJ 100 50S ribosomal protein L10	23	DEG10170042; DEG10120337; DEG10240343; DEG10160261; DEG10060295; DEG10330264; DEG10350410; DEG10180569; DEG10290024; DEG10020045; DEG10140108; DEG10320329; DEG10200082; DEG10280380; DEG10030049; DEG10250127; DEG10100090; DEG10010026; DEG10310054; DEG10210112; DEG10360208; DEG10130050; DEG10190276;
rplK 118 50S ribosomal protein L11	25	DEG10170040; DEG10120339; DEG10240345; DEG10060063; DEG10150030; DEG10290022; DEG10280260; DEG10160262; DEG10080221; DEG10100089; DEG10340489; DEG10350411; DEG10330265; DEG10020043; DEG10250123; DEG10200088; DEG10130048; DEG10320328; DEG10310053; DEG10140006; DEG10050166; DEG10030047; DEG10360210; DEG10220332; DEG10210148;
rplL 111 50S ribosomal protein L7/L12	25	DEG10030050; DEG10060296; DEG10170043; DEG10120336; DEG10240342; DEG10150032; DEG10220329; DEG10160260; DEG10110211; DEG10330263; DEG10340486; DEG10290025; DEG10140107; DEG10020046; DEG10310055; DEG10200083; DEG10100091; DEG10010027; DEG10360207; DEG10210113; DEG10180570; DEG10130051; DEG10190277; DEG10320330; DEG10280379;
rplM 121 50S ribosomal protein L13	28	DEG10240144; DEG10290341; DEG10100544; DEG10130371; DEG10340425; DEG10060339; DEG10320258; DEG10360234; DEG10350355; DEG10080010; DEG10020248; DEG10330240; DEG10280497; DEG10310194; DEG10050527; DEG10250673; DEG10180481; DEG10190194; DEG10170306; DEG10140190; DEG10010064; DEG10220339; DEG10150258; DEG10120287; DEG10160237; DEG10210189; DEG10200175; DEG10030117;
rplN 112 50S ribosomal protein L14	24	DEG10290042; DEG10240328; DEG10020268; DEG10120063; DEG10050275; DEG10180500; DEG10100113; DEG10200145; DEG10320279; DEG10130422; DEG10170324; DEG10150042; DEG10060132; DEG10010045; DEG10160310; DEG10030527; DEG10340471; DEG10220411; DEG10080258; DEG10280173; DEG10140231; DEG10330314; DEG10210019; DEG10190219;
rplO 102 50S ribosomal protein L15	26	DEG10290051; DEG10170315; DEG10240320; DEG10130413; DEG10010054; DEG10180493; DEG10060140; DEG10360192; DEG10030518; DEG10230146; DEG10340462; DEG10220402; DEG10320270; DEG10350397; DEG10140223; DEG10280164; DEG10330323; DEG10150049; DEG10120072; DEG10050284; DEG10110181; DEG10160319; DEG10080253; DEG10020259; DEG10200153; DEG10190210;
rplP 102 50S ribosomal protein L16	28	DEG10290039; DEG10360197; DEG10120060; DEG10150039; DEG10030530; DEG10050272; DEG10370015; DEG10160307; DEG10250140; DEG10100110; DEG10060129; DEG10190222; DEG10200142; DEG10340474; DEG10170327; DEG10240330; DEG10130425; DEG10010042; DEG10110186; DEG10380013; DEG10220413; DEG10020271; DEG10270128; DEG10140234; DEG10210017; DEG10280176; DEG10320282; DEG10330311;
rplQ 115 50S ribosomal protein L17	22	DEG10150053; DEG10330330; DEG10060149; DEG10290057; DEG10080248; DEG10250675; DEG10030511; DEG10160326; DEG10190204; DEG10240312; DEG10170307; DEG10130407; DEG10360186; DEG10010063; DEG10050289; DEG10120079; DEG10020251; DEG10370024; DEG10140214; DEG10210030; DEG10320264; DEG10200159;
rplR 98 50S ribosomal protein L18	24	DEG10240323; DEG10170318; DEG10130416; DEG10010051; DEG10020262; DEG10120069; DEG10220405; DEG10340465; DEG10210023; DEG10140225; DEG10320273; DEG10100118; DEG10280167; DEG10150047; DEG10060138; DEG10330320; DEG10290048; DEG10180495; DEG10050281; DEG10030521; DEG10310216; DEG10160316; DEG10200151; DEG10190213;
rplS 114 50S ribosomal protein	19	DEG10130446; DEG10050095; DEG10280270; DEG10240192; DEG10010126; DEG10160179; DEG10100463; DEG10060360; DEG10330181; DEG10020127;

L19		DEG10180403; DEG10210125; DEG10030113; DEG10140172; DEG10200031; DEG10290137; DEG10170151; DEG10320213; DEG10190155;
rplT 110 50S ribosomal protein L20	23	DEG10030597; DEG10150139; DEG10060166; DEG10010205; DEG10220199; DEG10020189; DEG10360092; DEG10340214; DEG10200122; DEG10190119; DEG10130384; DEG10080019; DEG10170244; DEG10320130; DEG10160096; DEG10290213; DEG10050475; DEG10120267; DEG10140094; DEG10210134; DEG10250343; DEG10180287; DEG10330098;
rplU 92 50S ribosomal protein L21	19	DEG10190187; DEG10220350; DEG10160230; DEG10320252; DEG10130364; DEG10170238; DEG10100393; DEG10240120; DEG10060197; DEG10280432; DEG10340142; DEG10200050; DEG10330233; DEG10120165; DEG10010196; DEG10020185; DEG10030076; DEG10290320; DEG10140127;
rplV 101 50S ribosomal protein L22	24	DEG10220415; DEG10060127; DEG10150037; DEG10030532; DEG10050270; DEG10080261; DEG10370013; DEG10160305; DEG10190224; DEG10330309; DEG10200140; DEG10130427; DEG10120058; DEG10010040; DEG10170329; DEG10210015; DEG10020273; DEG10340476; DEG10290037; DEG10140236; DEG10280178; DEG10250138; DEG10320284; DEG10100108;
rplW 94 50S ribosomal protein L23	23	DEG10130430; DEG10060124; DEG10310222; DEG10280181; DEG10010037; DEG10160302; DEG10190227; DEG10180504; DEG10030535; DEG10330306; DEG10240334; DEG10120055; DEG10050267; DEG10200137; DEG10020276; DEG10340479; DEG10220418; DEG10290034; DEG10100105; DEG10210012; DEG10170332; DEG10140239; DEG10320287;
rplX 81 50S ribosomal protein L24	25	DEG10240327; DEG10180499; DEG10120064; DEG10050276; DEG10020267; DEG10200146; DEG10320278; DEG10100114; DEG10290043; DEG10130421; DEG10150043; DEG10170323; DEG10060133; DEG10010046; DEG10310219; DEG10380014; DEG10030526; DEG10340470; DEG10080257; DEG10220410; DEG10160311; DEG10280172; DEG10140230; DEG10330315; DEG10190218;
rpmA 86 50S ribosomal protein L27	21	DEG10190186; DEG10200049; DEG10160229; DEG10170236; DEG10060199; DEG10210107; DEG10220349; DEG10320251; DEG10240121; DEG10100392; DEG10290319; DEG10150274; DEG10280433; DEG10340141; DEG10120166; DEG10140126; DEG10330232; DEG10010195; DEG10020184; DEG10030077; DEG10080051;
rpmE 78 50S ribosomal protein L31	11	DEG10240271; DEG10120084; DEG10200400; DEG10050257; DEG10010256; DEG10340354; DEG10080086; DEG10100197; DEG10280069; DEG10170298; DEG10020237;
rpmH 68 50S ribosomal protein L34	17	DEG10160274; DEG10240350; DEG10210201; DEG10020302; DEG10200108; DEG10080287; DEG10010271; DEG10190256; DEG10120010; DEG10140052; DEG10330278; DEG10060376; DEG10030002; DEG10050353; DEG10320310; DEG10290003; DEG10170351;
rpoA 177 DNA-directed RNA polymerase subunit alpha	30	DEG10290056; DEG10060148; DEG10100546; DEG10230142; DEG10250676; DEG10030512; DEG10210029; DEG10160325; DEG10350394; DEG10190205; DEG10110177; DEG10240313; DEG10130408; DEG10120307; DEG10170308; DEG10330329; DEG10360187; DEG10010062; DEG10380019; DEG10050288; DEG10120078; DEG10340455; DEG10020252; DEG10270605; DEG10140215; DEG10370023; DEG10220394; DEG10280159; DEG10320265; DEG10200158;
rpoB 182 DNA-directed RNA polymerase subunit beta	31	DEG10170044; DEG10030051; DEG10120335; DEG10240341; DEG10150033; DEG10220328; DEG10310056; DEG10110212; DEG10330262; DEG10380021; DEG10270116; DEG10140206; DEG10020047; DEG10290026; DEG10250129; DEG10200084; DEG10060281; DEG10050165; DEG10010028; DEG10280378; DEG10230156; DEG10100093; DEG10360206; DEG10350409; DEG10370026; DEG10210032; DEG10160259; DEG10180571; DEG10130052; DEG10190278; DEG10320331;
rpoC 148 DNA-directed RNA polymerase subunit beta	28	DEG10170045; DEG10030052; DEG10140205; DEG10120334; DEG10240340; DEG10180572; DEG10220327; DEG10110213; DEG10380022; DEG10330261; DEG10270117; DEG10290027; DEG10020048; DEG10200085; DEG10060280; DEG10100094; DEG10010029; DEG10230155; DEG10360205; DEG10350408; DEG10370027; DEG10160258; DEG10210033; DEG10130053; DEG10070226;

		DEG10250130; DEG10190279; DEG10320332;
rpsA 111 30S ribosomal protein S1	33	DEG10100275; DEG10330136; DEG10130272; DEG10200437; DEG10290226; DEG10350318; DEG10200007; DEG10180469; DEG10250540; DEG10370112; DEG10260064; DEG10250339; DEG10190076; DEG10170095; DEG10210121; DEG10030375; DEG10290116; DEG10270500; DEG10310113; DEG10320095; DEG10160134; DEG10110224; DEG10340147; DEG10360265; DEG10380118; DEG10050436; DEG10020139; DEG10050181; DEG10350077; DEG10240084; DEG10360126; DEG10130071; DEG10240184;
rpsB 138 30S ribosomal protein S2	27	DEG10140202; DEG10290150; DEG10120041; DEG10100460; DEG10330034; DEG10200263; DEG10030450; DEG10340423; DEG10210207; DEG10180039; DEG10150094; DEG10050330; DEG10240234; DEG10060052; DEG10010129; DEG10130265; DEG10220337; DEG10230259; DEG10080317; DEG10380224; DEG10190028; DEG10160033; DEG10020131; DEG10370216; DEG10320033; DEG10250562; DEG10170155;
rpsC 118 30S ribosomal protein S3	30	DEG10330310; DEG10360198; DEG10030531; DEG10150038; DEG10060128; DEG10050271; DEG10080260; DEG10160306; DEG10370014; DEG10190223; DEG10200141; DEG10130426; DEG10240331; DEG10120059; DEG10010041; DEG10170328; DEG10230150; DEG10380012; DEG10320283; DEG10340475; DEG10350401; DEG10270127; DEG10020272; DEG10220414; DEG10140235; DEG10210016; DEG10280177; DEG10250139; DEG10290038; DEG10100109;
rpsD 128 30S ribosomal protein S4	28	DEG10290055; DEG10370222; DEG10140283; DEG10230143; DEG10060252; DEG10250677; DEG10030513; DEG10020203; DEG10160324; DEG10350395; DEG10190206; DEG10100547; DEG10240314; DEG10200336; DEG10130409; DEG10010215; DEG10330328; DEG10120077; DEG10050287; DEG10210213; DEG10340456; DEG10360188; DEG10270606; DEG10380231; DEG10280132; DEG10220395; DEG10320266; DEG10170258;
rpsE 127 30S ribosomal protein S5	30	DEG10240322; DEG10170317; DEG10010052; DEG10130415; DEG10360193; DEG10230147; DEG10020261; DEG10250144; DEG10220404; DEG10370017; DEG10340464; DEG10210024; DEG10270132; DEG10320272; DEG10140224; DEG10100119; DEG10280166; DEG10350398; DEG10060139; DEG10330321; DEG10180494; DEG10290049; DEG10050282; DEG10120070; DEG10030520; DEG10110183; DEG10080254; DEG10160317; DEG10200152; DEG10190212;
rpsF 101 30S ribosomal protein S6	21	DEG10020019; DEG10130287; DEG10120223; DEG10050174; DEG10030062; DEG10240257; DEG10220127; DEG10180591; DEG10210049; DEG10200208; DEG10290338; DEG10320346; DEG10170017; DEG10340400; DEG10250015; DEG10080236; DEG10310142; DEG10010269; DEG10160343; DEG10360283; DEG10330348;
rpsG 118 30S ribosomal protein S7	28	DEG10170047; DEG10210199; DEG10150034; DEG10280186; DEG10110191; DEG10130145; DEG10370035; DEG10010031; DEG10340484; DEG10140161; DEG10290029; DEG10320292; DEG10220423; DEG10030060; DEG10240338; DEG10120050; DEG10060070; DEG10330301; DEG10160297; DEG10080219; DEG10100098; DEG10050189; DEG10380030; DEG10360203; DEG10200382; DEG10190232; DEG10350407; DEG10020050;
rpsH 111 30S ribosomal protein S8	26	DEG10240325; DEG10130418; DEG10180497; DEG10120067; DEG10360195; DEG10050279; DEG10020264; DEG10290046; DEG10250142; DEG10220407; DEG10210021; DEG10270130; DEG10140227; DEG10280169; DEG10320275; DEG10100116; DEG10200149; DEG10010049; DEG10170320; DEG10060136; DEG10340467; DEG10310217; DEG10030523; DEG10160314; DEG10190215; DEG10330318;
rpsI 129 30S ribosomal protein S9	22	DEG10130372; DEG10240145; DEG10050526; DEG10380217; DEG10250672; DEG10360233; DEG10020247; DEG10320257; DEG10290340; DEG10060338; DEG10170305; DEG10190193; DEG10180480; DEG10010065; DEG10160236; DEG10340424; DEG10220338; DEG10140191; DEG10200176; DEG10120288; DEG10030118; DEG10330239;
rpsJ 119 30S ribosomal protein S10	23	DEG10130433; DEG10030538; DEG10060121; DEG10150035; DEG10220421; DEG10010034; DEG10280184; DEG10340482; DEG10180507; DEG10330303; DEG10320290; DEG10140242; DEG10240337; DEG10120052; DEG10080263; DEG10160299; DEG10200134; DEG10190230; DEG10020279; DEG10370009;

		DEG10100102; DEG10210009; DEG10290031;
rpsK 129 30S ribosomal protein S11	26	DEG10150052; DEG10290054; DEG10100548; DEG10060147; DEG10130410; DEG10310204; DEG10030514; DEG10210028; DEG10160323; DEG10190207; DEG10110178; DEG10240315; DEG10280160; DEG10330327; DEG10170309; DEG10120076; DEG10050286; DEG10010061; DEG10340457; DEG10360189; DEG10020253; DEG10140216; DEG10370022; DEG10220396; DEG10320267; DEG10200157;
rpsL 174 30S ribosomal protein S12	28	DEG10120049; DEG10170046; DEG10370034; DEG10060069; DEG10030059; DEG10010030; DEG10280187; DEG10080220; DEG10220424; DEG10270118; DEG10130144; DEG10050190; DEG10340485; DEG10140162; DEG10310057; DEG10290028; DEG10020049; DEG10320293; DEG10240339; DEG10100097; DEG10160296; DEG10330300; DEG10210200; DEG10360204; DEG10200383; DEG10190233; DEG10250131; DEG10180510;
rpsM 117 30S ribosomal protein S13	27	DEG10170310; DEG10150051; DEG10290053; DEG10050285; DEG10180491; DEG10130411; DEG10360190; DEG10250678; DEG10030515; DEG10160322; DEG10060146; DEG10240316; DEG10190208; DEG10330326; DEG10120075; DEG10340458; DEG10380018; DEG10010060; DEG10270607; DEG10020254; DEG10280161; DEG10370021; DEG10220397; DEG10200156; DEG10140217; DEG10230144; DEG10320268;
rpsN 102 30S ribosomal protein S14	20	DEG10180498; DEG10010048; DEG10130419; DEG10150045; DEG10170321; DEG10320276; DEG10290045; DEG10120066; DEG10220408; DEG10340468; DEG10210218; DEG10160313; DEG10190216; DEG10280170; DEG10140228; DEG10050278; DEG10330317; DEG10200148; DEG10030524; DEG10020265;
rpsO 125 30S ribosomal protein S15	18	DEG10010138; DEG10340521; DEG10150285; DEG10180470; DEG10050538; DEG10240083; DEG10170168; DEG10160220; DEG10200008; DEG10140125; DEG10060344; DEG10080177; DEG10290115; DEG10130070; DEG10330223; DEG10030137; DEG10120152; DEG10220366;
rpsP 103 30S ribosomal protein S16	21	DEG10050096; DEG10320216; DEG10160182; DEG10030110; DEG10010124; DEG10190157; DEG10020125; DEG10170148; DEG10120330; DEG10330184; DEG10060362; DEG10130449; DEG10210127; DEG10380104; DEG10180406; DEG10140174; DEG10310092; DEG10240189; DEG10290134; DEG10340315; DEG10150088;
rpsQ 93 30S ribosomal protein S17	25	DEG10160309; DEG10020269; DEG10120062; DEG10310220; DEG10050274; DEG10250141; DEG10180501; DEG10190220; DEG10100112; DEG10200144; DEG10170325; DEG10130423; DEG10290041; DEG10140232; DEG10150041; DEG10060131; DEG10010044; DEG10030528; DEG10220412; DEG10340472; DEG10270129; DEG10280174; DEG10320280; DEG10210018; DEG10330313;
rpsR 111 30S ribosomal protein S18	18	DEG10280019; DEG10030064; DEG10170019; DEG10120222; DEG10050172; DEG10240259; DEG10340399; DEG10060074; DEG10330350; DEG10130288; DEG10080234; DEG10010267; DEG10200207; DEG10020021; DEG10210051; DEG10160345; DEG10290336; DEG10190290;
rpsS 108 30S ribosomal protein S19	23	DEG10010039; DEG10060126; DEG10150036; DEG10030533; DEG10290036; DEG10160304; DEG10180502; DEG10190225; DEG10330308; DEG10240332; DEG10120057; DEG10130428; DEG10050269; DEG10020274; DEG10220416; DEG10340477; DEG10200139; DEG10210014; DEG10100107; DEG10140237; DEG10170330; DEG10280179; DEG10320285;
rpsT 68 30S ribosomal protein S20	17	DEG10340056; DEG10200004; DEG10100385; DEG10150273; DEG10290309; DEG10080009; DEG10320003; DEG10180004; DEG10120017; DEG10140254; DEG10020175; DEG10010184; DEG10130204; DEG10280535; DEG10240055; DEG10310208; DEG10030144;
ruvA 70 Holliday junction DNA helicase subunit RuvA	6	DEG10350352; DEG10170234; DEG10220179; DEG10030355; DEG10300026; DEG10070007;
ruvB 83 Holliday junction DNA helicase subunit RuvB	12	DEG10170233; DEG10160083; DEG10350353; DEG10250510; DEG10340021; DEG10330085; DEG10210007; DEG10070009; DEG10080186; DEG10110113; DEG10300027; DEG10220302;

secA 103 Preprotein translocase subunit SecA	29	DEG10270340; DEG10030473; DEG10010243; DEG10150247; DEG10120164; DEG10230302; DEG10360217; DEG10160023; DEG10250638; DEG10370198; DEG10200375; DEG10130107; DEG10170070; DEG10140027; DEG10020064; DEG10060054; DEG10330024; DEG10290357; DEG10240117; DEG10100514; DEG10270575; DEG10250368; DEG10380210; DEG10220292; DEG10190021; DEG10180026; DEG10020297; DEG10320024; DEG10210052;
secY 134 Preprotein translocase subunit secY	30	DEG10290052; DEG10170314; DEG10130412; DEG10180492; DEG10010055; DEG10060141; DEG10360191; DEG10230145; DEG10030517; DEG10220401; DEG10340461; DEG10250145; DEG10160320; DEG10210025; DEG10270133; DEG10350396; DEG10140222; DEG10240319; DEG10190209; DEG10280163; DEG10330324; DEG10120073; DEG10380016; DEG10110180; DEG10080252; DEG10020258; DEG10370019; DEG10200154; DEG10320269; DEG10100121;
serS 144 seryl-tRNA synthetase	32	DEG10030229; DEG10220348; DEG10370181; DEG10330138; DEG10100605; DEG10170006; DEG10200296; DEG10270673; DEG10010006; DEG10250754; DEG10020005; DEG10190075; DEG10230295; DEG10240174; DEG10070134; DEG10290216; DEG10150149; DEG10320091; DEG10310039; DEG10130362; DEG10340140; DEG10160136; DEG10060004; DEG10350327; DEG10360084; DEG10120157; DEG10380193; DEG10080295; DEG10180145; DEG10280108; DEG10140016; DEG10210174;
smpB 97 SsrA-binding protein/SmpB superfamily	8	DEG10060046; DEG10220298; DEG10340345; DEG10300083; DEG10170083; DEG10140135; DEG10050343; DEG10020075;
sodA 162 Manganese superoxide dismutase	7	DEG10120355; DEG10130191; DEG10250756; DEG10270675; DEG10350056; DEG10150244; DEG10360215;
thrS 160 Threonyl-tRNA synthetase	29	DEG10030594; DEG10060308; DEG10120270; DEG10220196; DEG10360094; DEG10380064; DEG10100421; DEG10080016; DEG10340211; DEG10130387; DEG10250516; DEG10210142; DEG10110078; DEG10050493; DEG10170248; DEG10290210; DEG10160093; DEG10200074; DEG10230015; DEG10140296; DEG10350221; DEG10020192; DEG10190121; DEG10070202; DEG10270481; DEG10330095; DEG10180289; DEG10370065; DEG10320127;
thyA 90 Thymidylate synthase	23	DEG10230105; DEG10150011; DEG10210109; DEG10270498; DEG10360014; DEG10070139; DEG10240193; DEG10280412; DEG10030143; DEG10290127; DEG10180428; DEG10250538; DEG10340384; DEG10130088; DEG10170185; DEG10220458; DEG10350305; DEG10320229; DEG10120280; DEG10160195; DEG10200309; DEG10330198; DEG10020157;
tig 148 trigger factor Tig	1	DEG10300030;
tkt 90 transketolase	20	DEG10220364; DEG10180438; DEG10170174; DEG10240134; DEG10350363; DEG10360020; DEG10130437; DEG10330207; DEG10050368; DEG10010170; DEG10100237; DEG10270265; DEG10290092; DEG10250277; DEG10200450; DEG10160204; DEG10020147; DEG10140193; DEG10010146; DEG10280265;
topA 81 DNA topoisomerase I	22	DEG10240001; DEG10380140; DEG10220157; DEG10250713; DEG10020129; DEG10320157; DEG10050491; DEG10110077; DEG10290253; DEG10010128; DEG10130083; DEG10100575; DEG10180221; DEG10200333; DEG10360108; DEG10190108; DEG10060097; DEG10070067; DEG10210116; DEG10140176; DEG10270641; DEG10170154;
tpiA 120 triosephosphate isomerase	20	DEG10070195; DEG10120349; DEG10360269; DEG10380070; DEG10230264; DEG10060350; DEG10100234; DEG10220136; DEG10010239; DEG10170079; DEG10050225; DEG10270263; DEG10130057; DEG10140169; DEG10250276; DEG10340409; DEG10020072; DEG10240085; DEG10210091; DEG10200247;
trmD 136 tRNA (guanine-N(1)-)-methyltransferase	24	DEG10100464; DEG10060361; DEG10380105; DEG10320214; DEG10240191; DEG10160180; DEG10020126; DEG10330182; DEG10210126; DEG10190156; DEG10180404; DEG10110150; DEG10130447; DEG10120328; DEG10270523; DEG10010125; DEG10350309; DEG10370101; DEG10030112; DEG10140173; DEG10070040; DEG10290136; DEG10170150; DEG10250565;
trpC2 68 Indole-3-glycerol phosphate synthase	5	DEG10100267; DEG10280366; DEG10250329; DEG10270306; DEG10130297;

trpD 88 Anthranilate phosphoribosyl transferase	5	DEG10100339; DEG10280367; DEG10130296; DEG10050501; DEG10250424;
trpE 72 anthranilate synthase	15	DEG10100265; DEG10250633; DEG10100378; DEG10100150; DEG10250185; DEG10130114; DEG10110107; DEG10270181; DEG10280233; DEG10280391; DEG10130045; DEG10050115; DEG10250328; DEG10250463; DEG10270305;
trpG 68 Anthranilate synthase component II	13	DEG10270597; DEG10080069; DEG10130295; DEG10130018; DEG10270006; DEG10050500; DEG10250005; DEG10280368; DEG10250666; DEG10050421; DEG10280051; DEG10360157; DEG10100534;
trpS 146 tryptophanyl-tRNA synthetase	27	DEG10190234; DEG10100529; DEG10060101; DEG10340429; DEG10250658; DEG10360235; DEG10350275; DEG10220010; DEG10370227; DEG10170101; DEG10380240; DEG10260068; DEG10290061; DEG10120321; DEG10130345; DEG10030541; DEG10020084; DEG10140295; DEG10080239; DEG10210217; DEG10200016; DEG10270590; DEG10230255; DEG10280420; DEG10180516; DEG10070244; DEG10010091;
truA 127 tRNA pseudouridine synthase A	4	DEG10250674; DEG10100545; DEG10050599; DEG10060153;
truB 102 tRNA pseudouridine synthase B	2	DEG10340009; DEG10050460;
trxA1 72 Thioredoxin	10	DEG10220004; DEG10060099; DEG10030730; DEG10010202; DEG10150321; DEG10180400; DEG10130381; DEG10110205; DEG10290070; DEG10170122;
trxB 97 Thioredoxin reductase	19	DEG10100609; DEG10060083; DEG10240290; DEG10270683; DEG10010230; DEG10130147; DEG10200359; DEG10050440; DEG10240172; DEG10180527; DEG10220259; DEG10140281; DEG10140297; DEG10010241; DEG10020096; DEG10170073; DEG10250766; DEG10070179; DEG10210167;
tsf 160 elongation factor Ts	29	DEG10290151; DEG10010130; DEG10120042; DEG10330035; DEG10340422; DEG10370217; DEG10210206; DEG10380225; DEG10350259; DEG10230260; DEG10320034; DEG10110016; DEG10240235; DEG10270520; DEG10060352; DEG10130264; DEG10220336; DEG10310156; DEG10100459; DEG10030449; DEG10160034; DEG10190029; DEG10020132; DEG10200262; DEG10140201; DEG10250561; DEG10360148; DEG10180040; DEG10170156;
tuf 144 Elongation factor Tu	88	DEG10150286; DEG10020086; DEG10210198; DEG10100449; DEG10350413; DEG10270224; DEG10180508; DEG10180509; DEG10250229; DEG10340483; DEG10110170; DEG10340049; DEG10100190; DEG10120051; DEG10060114; DEG10120365; DEG10360266; DEG10350405; DEG10350406; DEG10180471; DEG10340492; DEG10250133; DEG10250132; DEG10020174; DEG10020137; DEG10030135; DEG10230160; DEG10110190; DEG10280185; DEG10160222; DEG10180568; DEG10370036; DEG10280414; DEG10320291; DEG10120342; DEG10220335; DEG10160298; DEG10230195; DEG10200381; DEG10320249; DEG10140071; DEG10290030; DEG10130059; DEG10020051; DEG10020052; DEG10190182; DEG10170048; DEG10170049; DEG10010137; DEG10350216; DEG10220422; DEG10010032; DEG10010033; DEG10270119; DEG10130160; DEG10140160; DEG10330302; DEG10170166; DEG10270506; DEG10100099; DEG10220276; DEG10050188; DEG10220060; DEG10140150; DEG10380190; DEG10290114; DEG10260053; DEG10130313; DEG10060366; DEG10330225; DEG10050236; DEG10130146; DEG10370177; DEG10200009; DEG10240293; DEG10370071; DEG10070012; DEG10100100; DEG10060071; DEG10290087; DEG10230154; DEG10020099; DEG10270120; DEG10380031; DEG10360202; DEG10190231; DEG10210136; DEG10250549;
tyrS 133 tyrosyl-tRNA synthetase	29	DEG10120232; DEG10340017; DEG10160102; DEG10310168; DEG10140182; DEG10020205; DEG10060368; DEG10280202; DEG10180273; DEG10230185; DEG10270322; DEG10350354; DEG10380020; DEG10190114; DEG10200237; DEG10070234; DEG10250355; DEG10030129; DEG10290124; DEG10170260; DEG10010216; DEG10330105; DEG10050579; DEG10320142; DEG10220064; DEG10370025; DEG10100291; DEG10210031; DEG10130004;
upp 74 uracil phosphoribosyltransferase	3	DEG10060021; DEG10020233; DEG10140121;

uvrB 119 UvrABC system protein B	10	DEG10050448; DEG10060055; DEG10250631; DEG10270219; DEG10050452; DEG10020066; DEG10270573; DEG10050321; DEG10020228; DEG10010112;
uvrC 90 UvrABC system protein C	6	DEG10060174; DEG10100231; DEG10250271; DEG10020104; DEG10270258; DEG10180295;
valS 185 valyl-tRNA synthetase	95	DEG10330352; DEG10290306; DEG10160148; DEG10240062; DEG10230300; DEG10340272; DEG10280301; DEG10130006; DEG10270288; DEG10370053; DEG10170264; DEG10210040; DEG10030147; DEG10220095; DEG10060284; DEG10120199; DEG10020031; DEG10070203; DEG10220258; DEG10180006; DEG10170025; DEG10310102; DEG10010110; DEG10020186; DEG10120108; DEG10170240; DEG10360173; DEG10320349; DEG10210063; DEG10250186; DEG10240137; DEG10380176; DEG10270449; DEG10130366; DEG10120038; DEG10030503; DEG10080214; DEG10360249; DEG10230180; DEG10020210; DEG10060222; DEG10280409; DEG10370028; DEG10360169; DEG10020112; DEG10290288; DEG10280074; DEG10330151; DEG10380169; DEG10050502; DEG10160347; DEG10350060; DEG10320072; DEG10220171; DEG10370155; DEG10210162; DEG10250479; DEG10060273; DEG10350223; DEG10330003; DEG10010218; DEG10150271; DEG10270182; DEG10060013; DEG10230038; DEG10190293; DEG10140090; DEG10380050; DEG10350361; DEG10100396; DEG10320005; DEG10370162; DEG10140258; DEG10200097; DEG10180599; DEG10200472; DEG10150075; DEG10100151; DEG10160003; DEG10310141; DEG10380024; DEG10050332; DEG10180113; DEG10280125; DEG10290105; DEG10200161; DEG10190066; DEG10140025; DEG10010199; DEG10340148; DEG10340547; DEG10130395; DEG10110002; DEG10170133; DEG10250305;
ychF 89 GTP-binding protein YchF	30	DEG10190185; DEG10170235; DEG10240122; DEG10350373; DEG10210085; DEG10020183; DEG10200048; DEG10160228; DEG10270446; DEG10250476; DEG10110174; DEG10280434; DEG10070058; DEG10060316; DEG10220166; DEG10380157; DEG10290318; DEG10120388; DEG10130308; DEG10020018; DEG10280250; DEG10180201; DEG10180473; DEG10100391; DEG10370143; DEG10140057; DEG10330231; DEG10070001; DEG10010194; DEG10140136;

References:

1 Zhang, R., Ou, H. Y. & Zhang, C. T. DEG: a database of essential genes. *Nucleic acids research* **32**, D271-D272 (2004).

3.1.8.14 – Alignment output for 181 essential proteins agains five hosts

Supplementary Material S14: Alignment output for 181 essential proteins against five hosts.

Este material contém o resultado do alinhamento feito pelo programa Blastp das 181 proteínas essenciais contra os hospedeiros naturais *Ovis aries, Capra hircus, Bos taurus, Equus caballus* e *Homo sapiens;* disponibilizado no CD que acompanha esta tese.

Supplementary Material

Supplementary Table S15: Blast result of 181 *Corynebacteria pseudotuberculosis* essential protein against the hosts *Ovis aries, Capra hircus, Bos taurus, Equus caballus* and *Homo sapiens* from non-redundant database of protein sequences (nr). The identity and coverage values were calculated by blastp program (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins). The following parameters were used: Max sequene target = 100, Expected threshold = 10, Matrix = BLOSUM62, Low complexity regions filter = false. The coverage and identity values represent the better alignment agaist host protein. Yellow rows: proteins with identity alignment value less than 30% against host. Green rows: non-host homologous proteins, proteins with low identity and low coverage alignment value against host. Green rows withou values: non-host homologous, proteins without significant hits against host.

Corynebacterium Pseudotuberculosis biovar ovis			*Ovis aries*		*Capra Hircus*		*Bos Taurus*		*Equus Caballus*		*Homo Sapiens*	
Gene	**Product**	**Degree**	**Coverage**	**Identity**	**Coverage**	**Identity**	**Coverage**	**Identity**	**Coverage**	**Identity**	**Coverage**	**Identity**
ackA	acetate kinase	96	35%	28%							14%	33%
adk	adenylate kinase	172	94%	33%	100%	36%	100%	35%	100%	36%	100%	38%
alaS	alanyl-tRNA synthetase	106	97%	35%	97%	35%	97%	35%	98%	34%	98%	35%
apt	Adenine phosphoribosyltransferase	72	84%	42%	55%	44%	84%	41%	61%	41%	83%	42%
argC	N-acetyl-gamma-glutamyl-phosphate reductase	86	15%	31%	25%	27%	15%	31%	15%	31%	15%	31%
argF	ornithine carbamoyltransferase	115	93%	36%	93%	35%	92%	36%	87%	36%	93%	36%
argG	argininosuccinate synthase	77	97%	36%	97%	36%	97%	36%	97%	36%	97%	37%
argS	arginyl-tRNA synthetase	112	94%	24%	94%	25%	94%	24%	94%	24%	94%	24%
aroB	3-dehydroquinate synthase	86			20%	34%					16%	30%
aroC	Chorismate synthase	125							20%	35%	13%	32%
asd	Aspartate-semialdehyde dehydrogenase	98	38%	26%	38%	26%	11%	44%	17%	31%	38%	25%
aspS	Aspartyl-tRNA synthetase	101	97%	41%	97%	41%	97%	41%	97%	41%	99%	40%
atpA	ATP synthase subunit alpha	127	89%	54%	89%	54%	89%	54%	89%	54%	89%	54%
atpD	ATP synthase subunit beta	111	96%	62%	96%	62%	97%	61%	97%	61%	97%	61%
atpG	ATP synthase subunit gamma	85	97%	23%	97%	23%	97%	23%	97%	23%	97%	23%
carA	carbamoyl-phosphate synthase small chain	93	88%	39%	82%	38%	88%	39%	88%	39%	88%	38%
cysK	cysteine synthase	78	97%	41%	97%	42%	97%	41%	97%	40%	97%	41%
dapA	Dihydrodipicolinate synthase	74	93%	27%	93%	27%	92%	26%	93%	27%	93%	27%
dapB	Dihydrodipicolinate reductase	72					20%	37%				
dnaB	Replicative DNA helicase	90			10%	33%	13%	32%	13%	32%	14%	28%
dnaG	DNA primase	109					16%	29%				
dnaK	Chaperone protein DnaK	239	95%	55%	94%	50%	95%	55%	95%	55%	95%	55%
efp	elongation factor P	133					47%	26%			32%	33%
engA	GTP-binding protein EngA	105	66%	28%	54%	29%	60%	28%	31%	37%	48%	30%
eno	enolase	232	99%	54%	99%	54%	99%	54%	99%	55%	99%	55%
ffh	Signal recognition particle protein	123	52%	29%	52%	29%	52%	29%	52%	29%	95%	33%
fmt	Methionyl-tRNA formyltransferase	116	72%	33%	72%	33%	72%	33%	91%	30%	73%	30%
folA	Dihydrofolate reductase	71	58%	32%	58%	32%	58%	33%	58%	33%	57%	34%

folD	bifunctional protein folD	126	87%	44%	87%	44%	87%	44%	93%	44%	87%	45%
frr	Ribosome-recycling factor (RRF)	147	98%	27%	98%	27%	98%	30%	98%	31%	98%	30%
ftsH	cell division protein	90	65%	48%	66%	47%	65%	48%	73%	45%	72%	46%
ftsY	cell division protein FtsY	91	74%	27%	74%	27%	57%	29%	57%	29%	47%	31%
ftsZ	Cell division protein ftsZ	184										
fusA	Elongation factor G	171	97%	44%	97%	44%	97%	44%	98%	39%	95%	44%
gap	glyceraldehyde-3-phosphate dehydrogenase	122	99%	48%	98%	48%	99%	48%	99%	48%	99%	49%
glmS	glucosamine--fructose-6-phosphate	87	99%	35%	96%	35%	99%	35%	96%	34%	99%	34%
gltA	Citrate synthase	125	85%	28%	85%	28%	85%	28%	85%	28%	85%	29%
gltX1	glutamyl-tRNA synthetase	129	97%	32%	96%	31%	97%	32%	97%	32%	97%	31%
glyA	Serine hydroxymethyltransferase	251	89%	45%	89%	45%	89%	45%	90%	45%	95%	43%
gmk	guanylate kinase	99	97%	40%	97%	40%	97%	40%	97%	41%	97%	41%
greA	Transcription elongation factor GreA	82	31%	30%	31%	30%	31%	30%	44%	29%	31%	33%
groEL	Chaperonin	105	97%	44%	97%	44%	97%	44%	97%	44%	97%	44%
groEL1	Chaperonin GroEL	125	96%	46%	96%	46%	96%	46%	95%	23%	96%	46%
guaA	GMP synthase	142	97%	36%	97%	36%	98%	36%	97%	37%	97%	36%
guaB	Inosine-5'-monophosphate dehydrogenase	93	93%	42%	90%	42%	91%	43%	93%	42%	91%	43%
gyrA	DNA gyrase subunit A	158	36%	26%	36%	26%	37%	26%	36%	25%	36%	24%
gyrB	DNA gyrase subunit B	121	89%	24%	89%	24%	89%	24%	89%	25%	89%	25%
hemE	uroporphyrinogen decarboxylase	180	94%	35%	84%	34%	94%	35%	94%	35%	94%	35%
hisD	histidinol dehydrogenase	98	11%	39%			18%	31%	31%	23%	21%	29%
hisF	Imidazole glycerol phosphate synthase subunit	79	59%	23%	59%	23%	59%	23%	59%	23%		
hisG	ATP phosphoribosyl transferase	73							18%	35%	25%	30%
hisS	histidyl-tRNA synthetase	151	98%	23%	98%	22%	98%	23%	92%	23%	99%	25%
ileS	isoleucyl-tRNA synthetase	129	93%	41%	93%	41%	93%	41%	93%	42%	93%	41%
ilvA	Threonine dehydratase	72	73%	29%	73%	29%	77%	29%	77%	29%	73%	30%
ilvC	ketol-acid reductoisomerase	117	19%	34%	19%	34%	19%	34%	18%	26%	19%	34%
ilvD	Dihydroxy-acid dehydratase	89							29%	23%	15%	32%
infA	translation initiation factor IF-1	75	74%	29%	74%	29%	61%	32%	61%	30%	61%	30%
infB	translation initiation factor IF-2	144	52%	41%	52%	41%	52%	41%	52%	41%	52%	41%
infC	translation initiation factor IF-3	121	100%	25%	91%	26%	97%	25%	64%	33%	57%	28%
katA	catalase	74	95%	43%	95%	43%	95%	43%	95%	43%	95%	43%
ksgA	Dimethyladenosine transferase	122	86%	32%	87%	32%	87%	31%	76%	34%	86%	31%
ldh	L-lactate dehydrogenase	77	97%	41%	97%	41%	97%	40%	97%	41%	97%	40%
lepA	GTP-binding protein LepA	82	96%	46%	96%	46%	96%	45%	96%	45%	97%	45%
leuB	3-isopropylmalate dehydrogenase	89	91%	33%	73%	36%	91%	33%	91%	33%	91%	33%
leuC	3-isopropylmalate dehydratase large subunit	80	79%	29%	79%	28%	79%	29%	79%	29%	79%	29%
leuS	leucyl-tRNA synthetase	143	98%	28%	98%	29%	98%	28%	98%	28%	98%	27%
lysA	diaminopimelate decarboxylase	84	83%	23%	77%	25%	83%	23%	81%	23%	81%	22%
lysC	Aspartate kinase	91										

metG	methionyl-tRNA synthetase	119	85%	33%	85%	33%	85%	33%	85%	34%	85%	34%
metK	S-adenosylmethionine synthase	121	94%	49%	94%	49%	94%	50%	94%	51%	94%	50%
miaA	tRNA dimethylallyltransferase	92	83%	28%	83%	29%	54%	36%	86%	29%	54%	35%
murA	UDP-N-acetylglucosamine	100										
ndk	nucleoside diphosphate kinase	162	95%	56%	94%	50%	95%	56%	94%	51%	94%	54%
nth	Endonuclease III	73	69%	28%	65%	29%	65%	28%	65%	29%	65%	29%
nusA	Transcription elongation protein	73	24%	30%					24%	30%	37%	37%
nusG	Transcription anti-termination protein NusG	121										
obgE	GTPase ObgE	122	50%	42%	50%	41%	58%	42%	50%	40%	65%	38%
pgi	glucose-6-phosphate isomerase	73	99%	52%	88%	54%	99%	52%	98%	51%	99%	50%
pgk	phosphoglycerate kinase	194	93%	43%	93%	43%	93%	43%	93%	43%	93%	43%
pheA	Prephenate dehydratase	69	17%	41%	19%	37%	18%	37%	20%	36%	13%	49%
pheS	phenylalanyl-tRNA synthetase subunit alpha	121	65%	31%	65%	31%	65%	31%	65%	31%	65%	32%
pnp	Polyribonucleotide nucleotidyltransferase	176	92%	36%	92%	36%	92%	36%	92%	36%	92%	36%
polA	DNA polymerase I	73	54%	32%	55%	25%	54%	31%	59%	29%	62%	29%
ppa	inorganic pyrophosphatase	76	77%	25%	77%	25%	77%	25%	77%	26%	81%	26%
prfA	Peptide chain release factor 1	75	76%	47%	76%	46%	76%	47%	76%	47%	87%	43%
proA	gamma-glutamyl phosphate reductase	83	92%	35%	92%	35%	92%	36%	92%	36%	92%	35%
proS	prolyl-tRNA synthetase	74	67%	40%	67%	40%	63%	40%	67%	41%	65%	39%
prsA	Ribose-phosphate pyrophosphokinase	143	95%	44%	95%	43%	96%	44%	95%	44%	95%	44%
purA	Adenylo succinate synthetase	128	97%	46%	97%	47%	97%	47%	90%	46%	97%	47%
purD	Phosphoribosylamine--glycine ligase	81	98%	40%	98%	40%	98%	41%	98%	41%	98%	42%
purE	Phosphoribosyl amino imidazole carboxylase	84	66%	30%	66%	29%	66%	30%	66%	32%	66%	31%
purF	amidophosphoribosyltransferase	72	84%	42%	84%	42%	84%	42%	88%	40%	88%	40%
purH	bifunctional	80	98%	32%	84%	36%	98%	32%	84%	36%	98%	33%
pyk	Pyruvate kinase	205	97%	38%	97%	37%	97%	37%	97%	36%	97%	36%
pyrB	aspartate carbamoyltransferase	106	97%	37%	97%	37%	97%	37%	97%	37%	97%	37%
pyrD	Dihydroorotate dehydrogenase 2	85	83%	45%	83%	44%	83%	43%	83%	44%	83%	43%
pyrH	uridylate kinase	124										
recA	recombinase A	105	55%	22%	37%	28%	53%	27%	53%	28%	53%	28%
recN	DNA repair protein recN	75									59%	29%
rho	Transcription termination factor Rho	105					29%	26%				
rnc	Ribonuclease III	81	81%	27%	81%	27%	81%	27%	81%	27%	81%	27%
rpe	Ribulose-phosphate 3-epimerase	80	95%	43%	81%	44%	95%	43%	80%	44%	95%	43%
rplA	50S ribosomal protein L1	115										
rplB	50S ribosomal protein L2	126	64%	43%	64%	41%	64%	43%	64%	42%	64%	42%
rplC	50S ribosomal protein L3	124	94%	31%	64%	36%	94%	31%	93%	29%	93%	33%
rplD	50S ribosomal protein L4	113	84%	37%	46%	48%	84%	35%	84%	38%	84%	36%
rplE	50S ribosomal protein L5	123	60%	27%	60%	27%	60%	27%			60%	28%
rplF	50S ribosomal protein L6	107	89%	26%	48%	30%	47%	30%	30%	38%	87%	27%

rplI	50S ribosomal protein L9	126	32%	52%	44%	28%	32%	52%	97%	32%	41%	48%
rplJ	50S ribosomal protein L10	100							36%	31%		
rplK	50S ribosomal protein L11	118	97%	40%	77%	38%	97%	40%	94%	42%	97%	40%
rplL	50S ribosomal protein L7/L12	111			94%	31%	94%	29%	56%	41%	94%	30%
rplM	50S ribosomal protein L13	121	91%	36%	91%	34%	91%	34%	82%	35%	78%	35%
rplN	50S ribosomal protein L14	112	82%	43%	82%	43%	82%	43%	95%	31%	82%	43%
rplO	50S ribosomal protein L15	102	43%	34%	43%	34%	43%	34%			43%	33%
rplP	50S ribosomal protein L16	102	97%	31%	97%	31%	97%	31%	84%	32%	84%	31%
rplQ	50S ribosomal protein L17	115	65%	38%	65%	37%	65%	38%	65%	38%	65%	36%
rplR	50S ribosomal protein L18	98	34%	29%							34%	27%
rplS	50S ribosomal protein L19	114	76%	36%	76%	37%	76%	37%	76%	36%	76%	36%
rplT	50S ribosomal protein L20	110	78%	34%	44%	38%	78%	35%	44%	38%	74%	36%
rplU	50S ribosomal protein L21	92	100%	25%	100%	25%	100%	25%			70%	29%
rplV	50S ribosomal protein L22	101	70%	27%	43%	38%	70%	27%			70%	27%
rplW	50S ribosomal protein L23	94	77%	25%	85%	34%	85%	35%	85%	34%	85%	34%
rplX	50S ribosomal protein L24	81	95%	38%			95%	38%	95%	36%	95%	37%
rpmA	50S ribosomal protein L27	86	95%	39%	95%	39%	75%	42%	95%	39%	75%	41%
rpmE	50S ribosomal protein L31	78										
rpmH	50S ribosomal protein L34	68	82%	38%	82%	38%	82%	38%	82%	38%	82%	41%
rpoA	DNA-directed RNA polymerase subunit alpha	177					20%	28%				
rpoB	DNA-directed RNA polymerase subunit beta	182	77%	29%	33%	29%	69%	29%	69%	29%	36%	29%
rpoC	DNA-directed RNA polymerase subunit beta	148	60%	26%	60%	25%	60%	26%	60%	25%	60%	26%
rpsA	30S ribosomal protein S1	111	77%	22%	70%	23%	70%	22%	67%	24%	71%	41%
rpsB	30S ribosomal protein S2	138	82%	28%			82%	28%			82%	29%
rpsC	30S ribosomal protein S3	118					57%	26%	69%	25%	69%	25%
rpsD	30S ribosomal protein S4	128	22%	38%	22%	33%					22%	40%
rpsE	30S ribosomal protein S5	127	52%	31%	64%	28%	64%	28%	64%	28%	87%	26%
rpsF	30S ribosomal protein S6	101									89%	32%
rpsG	30S ribosomal protein S7	118	90%	33%			90%	33%	90%	33%	90%	32%
rpsH	30S ribosomal protein S8	111	100%	23%	100%	23%	100%	25%	100%	26%	100%	23%
rpsI	30S ribosomal protein S9	129	88%	36%	88%	35%	88%	35%	87%	35%	88%	35%
rpsJ	30S ribosomal protein S10	119	99%	26%	65%	31%	99%	26%			99%	26%
rpsK	30S ribosomal protein S11	129	98%	39%	98%	39%	98%	39%	98%	39%	98%	37%
rpsL	30S ribosomal protein S12	174	90%	50%	90%	50%	90%	50%	69%	56%	89%	46%
rpsM	30S ribosomal protein S13	117									75%	25%
rpsN	30S ribosomal protein S14	102	95%	30%	95%	30%	95%	30%	95%	31%	95%	30%
rpsO	30S ribosomal protein S15	125	91%	27%	75%	31%	91%	28%	75%	31%	75%	37%
rpsP	30S ribosomal protein S16	103	50%	32%			50%	32%	50%	35%	50%	31%
rpsQ	30S ribosomal protein S17	93			83%	29%	83%	28%	83%	28%	83%	28%
rpsR	30S ribosomal protein S18	111	91%	36%	91%	36%	91%	38%	91%	38%	91%	36%

rpsS	30S ribosomal protein S19	108	53%	39%	53%	39%	53%	39%			54%	38%
rpsT	30S ribosomal protein S20	68			62%	31%	49%	45%				
ruvA	Holliday junction DNA helicase subunit RuvA	70									79%	21%
ruvB	Holliday junction DNA helicase subunit RuvB	83	51%	26%	51%	28%	51%	26%	17%	34%	51%	27%
secA	Preprotein translocase subunit SecA	103	10%	26%			8%	27%	7%	33%		
secY	Preprotein translocase subunit secY	134	88%	20%			88%	21%	64%	20%	64%	20%
serS	seryl-tRNA synthetase	144	79%	39%	75%	40%	71%	41%	96%	32%	71%	41%
smpB	SsrA-binding protein/SmpB superfamily	97							54%	28%		
sodA	Manganese superoxide dismutase	162	99%	59%	99%	59%	99%	59%	99%	56%	99%	58%
thrS	Threonyl-tRNA synthetase	160	96%	34%	96%	34%	96%	34%	91%	35%	94%	35%
thyA	Thymidylate synthase	90	98%	49%	98%	49%	98%	49%	98%	49%	98%	50%
tig	trigger factor Tig	148			36%	23%						
tkt	transketolase	90	84%	26%	82%	25%	84%	26%	92%	26%	95%	26%
topA	DNA topoisomerase I	81	50%	28%	50%	28%	50%	28%	50%	27%	50%	27%
tpiA	triosephosphate isomerase	120	94%	43%	71%	49%	94%	43%	94%	42%	94%	42%
trmD	tRNA (guanine-N(1)-)-methyltransferase	136	53%	26%			53%	25%			46%	45%
trpC2	Indole-3-glycerol phosphate synthase	68							20%	30%	43%	26%
trpD	Anthranilate phosphoribosyl transferase	88	31%	25%	31%	25%	31%	25%	11%	39%	30%	24%
trpE	anthranilate synthase	72					9%	32%			36%	27%
trpG	Anthranilate synthase component II	68	62%	26%	62%	26%	62%	26%	59%	27%	62%	26%
trpS	tryptophanyl-tRNA synthetase	146	95%	42%	95%	42%	95%	42%	95%	43%	96%	44%
truA	tRNA pseudouridine synthase A	127	80%	31%	80%	27%	80%	31%	80%	28%	80%	31%
truB	tRNA pseudouridine synthase B	102	78%	39%	78%	38%	78%	38%	72%	38%	72%	39%
trxA1	Thioredoxin	72	86%	41%	86%	41%	86%	41%	86%	42%	86%	41%
trxB	Thioredoxin reductase	97	58%	30%	57%	30%	58%	31%	67%	25%	67%	25%
tsf	elongation factor Ts	160	96%	28%	96%	29%	96%	28%	95%	27%	94%	28%
tuf	Elongation factor Tu	144	98%	55%	98%	55%	98%	55%	98%	55%	99%	54%
tyrS	tyrosyl-tRNA synthetase	133	91%	35%	91%	35%	91%	35%	93%	35%	93%	34%
upp	uracil phosphoribosyltransferase	74	90%	29%	63%	25%	90%	29%	63%	25%	90%	30%
uvrB	UvrABC system protein B	119	15%	35%	15%	35%	15%	36%	15%	34%	15%	34%
uvrC	UvrABC system protein C	90	7%	38%	8%	37%	8%	37%	8%	39%	8%	39%
valS	valyl-tRNA synthetase	185	98%	37%	74%	41%	96%	38%	98%	37%	98%	36%
ychF	GTP-binding protein YchF	89	99%	42%	99%	42%	99%	41%	99%	42%	99%	42%

3.2 - Label-free proteomic analysis to confirm the predicted proteome of *Corynebacterium pseudotuberculosis* under nitrosative stress mediated by nitric oxide

Wanderson M. Silva, Rodrigo D. Carvalho, Siomar C. Soares, Isabela F. S. Bastos, Edson Luiz Folador, Gustavo H. M. F. Souza, Yves Le Loir, Anderson Miyoshi, Artur Silva e Vasco Azevedo

No trablalho experimental de proteômica comparativa conduzido pelo Dr. Wanderson M. Silva, quando comparado uma amostra da linhagem 1002 de *C. pseudotuberculosis* submetida a estresse nitrosativo com uma amostra controle, foram identificadas proteínas diferencialmente expressas.

Em posse das redes de interação, neste trabalho foi criado uma subrede, contendo as interaçãoes entre um conjunto específico de proteínas. Assim, a rede de interação parcial para a linhagem 1002 foi formada pela interação entre as proteínas diferencialmente expressas somadas às proteínas exclusivamente expressas na condição de estresse.

A rede de interação propiciou uma visão sistêmica das proteínas envolvidas na resposta ao estresse nitrosativo e, junto com outros experimentos, auxiliou na interpretação dos mecanismos biológicos que permite a resistência e sobrevivência de *C. pseudotuberculosis* quando exposta à condição de stresse.

O artigo referente a este trabalho foi publicado na revista BMC Genomics em dezembro de 2014, tendo DOI número 10.1186/1471-2164-15-1065, estando também disponível no endereço eletrônico http://www.biomedcentral.com/1471-2164/15/1065.

Silva et al. BMC Genomics 2014, 15:1065
http://www.biomedcentral.com/1471-2164/15/1065

BMC Genomics

RESEARCH ARTICLE Open Access

Label-free proteomic analysis to confirm the predicted proteome of *Corynebacterium pseudotuberculosis* under nitrosative stress mediated by nitric oxide

Wanderson M Silva[1,4,5], Rodrigo D Carvalho[1], Siomar C Soares[1], Isabela FS Bastos[1], Edson L Folador[1], Gustavo HMF Souza[3], Yves Le Loir[4,5], Anderson Miyoshi[1], Artur Silva[2] and Vasco Azevedo[1*]

Abstract

Background: *Corynebacterium pseudotuberculosis* biovar *ovis* is a facultative intracellular pathogen, and the etiological agent of caseous lymphadenitis in small ruminants. During the infection process, the bacterium is subjected to several stress conditions, including nitrosative stress, which is caused by nitric oxide (NO). *In silico* analysis of the genome of *C. pseudotuberculosis ovis* 1002 predicted several genes that could influence the resistance of this pathogen to nitrosative stress. Here, we applied high-throughput proteomics using high definition mass spectrometry to characterize the functional genome of *C. pseudotuberculosis ovis* 1002 in the presence of NO-donor Diethylenetriamine/nitric oxide adduct (DETA/NO), with the aim of identifying proteins involved in nitrosative stress resistance.

Results: We characterized 835 proteins, representing approximately 41% of the predicted proteome of *C. pseudotuberculosis ovis* 1002, following exposure to nitrosative stress. In total, 102 proteins were exclusive to the proteome of DETA/NO-induced cells, and a further 58 proteins were differentially regulated between the DETA/NO and control conditions. An interactomic analysis of the differential proteome of *C. pseudotuberculosis* in response to nitrosative stress was also performed. Our proteomic data set suggested the activation of both a general stress response and a specific nitrosative stress response, as well as changes in proteins involved in cellular metabolism, detoxification, transcriptional regulation, and DNA synthesis and repair.

Conclusions: Our proteomic analysis validated previously-determined *in silico* data for *C. pseudotuberculosis ovis* 1002. In addition, proteomic screening performed in the presence of NO enabled the identification of a set of factors that can influence the resistance and survival of *C. pseudotuberculosis* during exposure to nitrosative stress.

Keywords: *Corynebacterium pseudotuberculosis*, Caseous lymphadenitis, Proteomics, Label-free proteomics, Nitrosative stress, Nitric oxide

Background

Corynebacterium pseudotuberculosis is a Gram-positive, facultative, intracellular pathogen belonging to the *Corynebacterium*, *Mycobacterium*, *Nocardia*, or CMN, group. This group belongs to the phylum Actinobacteria. The defining characteristics of the CMN group are a specific cell wall organization, consisting of peptidoglycan, arabinogalactan, and mycolic acids, and a high chromosomal G + C content [1]. *C. pseudotuberculosis ovis* is the etiological agent of the chronic infectious disease caseous lymphadenitis, which affects small ruminants worldwide. As a result, *C. pseudotuberculosis ovis* is responsible for significant economic losses in the goat and sheep industries, mainly stemming from decreased meat, wool, and milk production, reproductive disorders, and carcass contamination [1,2]. Bacterial factors that contribute to the virulence of *C. pseudotuberculosis* include phospholipase D [3], toxic cell wall lipids [4], and the iron transporter *fagABC* complex [5].

* Correspondence: vasco@icb.ufmg.br
[1]Depto de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
Full list of author information is available at the end of the article

In silico analysis of the genome of *C. pseudotuberculosis ovis* 1002 [6], as well as the pan-genome analysis of 15 other strains of *C. pseudotuberculosis* [7], identified genes involved in the response of this pathogen to different types of stress. Recently, the functional genome of *C. pseudotuberculosis ovis* 1002 was evaluated at the transcriptional level following exposure to different types of abiotic stress, including heat, osmotic, and acid stresses [8]. This allowed the characterization of several genes involved in distinct biological processes that favor the survival of the pathogen under the given stress condition.

However, during the infection process, *C. pseudotuberculosis* encounters nitrosative stress, caused by nitric oxide (NO), in the macrophage intracellular environment. A reactive nitrogen species (RNS) found in mammalian systems, NO is produced from L-arginine by NO synthases (NOS), and is present in three isoforms: endothelial NOS, neuronal NOS, involved in blood pressure control and neural signaling, and inducible NOS, associated with host defenses [9,10]. The NO produced during bacterial infection has antimicrobial properties, killing pathogens by causing damage to DNA, RNA, and proteins [11]. However, several pathogens contain pathways that allow bacterial survival under nitrosative stress conditions, including NO-sensitive transcriptional regulators [12], DNA and protein repair systems [13], and antioxidant systems [14].

Currently, little is known about the factors involved in the resistance of *C. pseudotuberculosis* to nitrosative stress. Pacheco et al. [15] showed that the alternative sigma (σ) factor, σ^E, plays a role in the survival of *C. pseudotuberculosis* in the presence of RNS. A σ^E null strain showed increased susceptibility to nitric oxide compared with the wild-type, and, in an *in vivo* assay, was unable to persist in mice. However, in iNOS-deficient mice, the mutant strain maintained its virulence [15]. In the same study, the extracellular proteome of *C. pseudotuberculosis* was analyzed in response to nitrosative stress, allowing the characterization of proteins that contribute to the adaptive processes of the pathogen in this environment [15].

To complement the results obtained in previous studies, and to identify factors involved in the survival of *C. pseudotuberculosis* under nitrosative stress conditions, we applied high-throughput proteomics using an liquid chromatograph high definition mass spectrometry (LC-HDMS^E) (data-independent acquisition, in ion mobility mode) approach to evaluate the global expression of the functional genome of *C. pseudotuberculosis ovis* 1002 at the protein level under nitrosative stress conditions.

Methods

Bacterial strain and growth conditions

C. pseudotuberculosis biovar *ovis* strain 1002, isolated from a goat, was maintained in brain heart infusion broth (BHI; HiMedia Laboratories Pvt. Ltd., Mumbai, India) at 37°C. For stress-resistance assays, strain 1002 was cultivated in a chemically-defined medium (CDM), containing $Na_2HPO_4.7H_2O$ (12.93 g/l), KH_2PO_4 (2.55 g/l), NH_4Cl (1 g/l), $MgSO_4.7H_2O$ (0.20 g/l), $CaCl_2$ (0.02 g/l), 0.05% (v/v) Tween 80, 4% (v/v) MEM vitamin solution (Invitrogen, Gaithersburg, MD, USA), 1% (v/v) MEM amino acid solution (Invitrogen), 1% (v/v) MEM non-essential amino acid solution (Invitrogen), and 1.2% (w/v) glucose, at 37°C [16].

Nitric oxide assay and preparation of whole bacterial lysates

Diethylenetriamine/nitric oxide adduct (DETA/NO) resistance of *C. pseudotuberculosis* was characterized as previously described [15]. When strain 1002 reached exponential growth phase (OD_{600} = 0.6) in the chemically-defined medium, the culture was divided into two aliquots (control condition, strain 1002_Ct; NO exposure, strain 1002_*DETA/NO*), and DETA/NO was added to the appropriate aliquot to a final concentration of 0.5 mM. The growth of strain 1002 in the presence of DETA/NO was then evaluated for 10 h. For proteomic analysis, protein was extracted after 1 h of exposure to DETA/NO. Both the control and DETA/NO cultures were centrifuged at 4,000 × *g* for 10 min at 4°C. The cell pellets were washed in phosphate buffered saline and then resuspended in 1 ml of lysis buffer (7 M urea, 2 M thiourea, 4% (w/v) CHAPS, and 1 M dithiothreitol (DTT)). The cells were then sonicated using five 1-min cycles on ice. The resulting lysates were centrifuged at 14,000 × *g* for 30 min at 4°C. The extracted proteins were then submitted to centrifugation at 13,000 × *g* for 10 min using a spin column with a threshold of 10 kDa (Millipore, Billerica, USA). Proteins were denatured with (0.1% (w/v) *Rapi*GEST SF surfactant at 60°C for 15 min (Waters, Milford, CA, USA), reduced using 10 mM DTT for 30 min at 60°C, and alkylated with 10 mM iodoacetamide in a dark chamber at 25°C for 30 min. Next, the proteins were enzymatically digested with 1:50 (w/w) trypsin at 37°C for 16 hours (sequencing grade modified trypsin; Promega, Madison, WI, USA). The digestion process was stopped by adding 10 μl of 5% (v/v) Trifluoroacetic acid (TFA) (Fluka, Buchs, Germany). Glycogen phosphorylase was added to the digests to a final concentration of 20 fmol/μl as an internal standard for normalization prior to each replicate injection. Analysis was carried out using a two-dimensional reversed phase (2D RP-RP) nanoUPLC-MS (Nano Ultra Performance Liquid Chromatography) approach, using multiplexed HDMS^E label-free quantitation as described previously [17].

LC-HDMS^E analysis and data processing

Qualitative and quantitative by 2D nanoUPLC tandem nanoESI-HDMS^E (Nano Electrospray High Definition Mass Spectrometry) experiments were conducted using a 1-h reversed phase (RP) acetonitrile (0.1% v/v formic

acid) gradient (7–40% (v/v)) at 500 nl/min on a nanoACQUITY UPLC 2D RP × RP Technology system [18]. A nanoACQUITY UPLC High Strength Silica (HSS) T3 1.8 μm 75 μm × 15 cm column (pH 3) was used in conjunction with a RP XBridge BEH130 C18 5 μm 300 μm × 50 mm nanoflow column (pH 10). Typical on-column sample loads were 250 ng of the total protein digests for each of the five fractions (250 ng/fraction/load). For all measurements, the mass spectrometer was operated in resolution mode, with a typical effective *m/z* conjoined ion-mobility resolving power of at least 1.5 M FWHM, an ion mobility cell filled with nitrogen gas, and a cross-section resolving power at least 40 Ω/ΔΩ. All analyses were performed using nano-electrospray ionization in the positive ion mode nanoESI (+), and a NanoLockSpray (Waters) ionization source. The lock mass channel was sampled every 30 s. The mass spectrometer was calibrated with a MS/MS spectrum of [Glu^1]-fibrinopeptide B (Glu-Fib) human solution (100 fmol/μl) delivered though the reference sprayer of the NanoLockSpray source. The double-charged ion ($[M + 2H]^{2+} = 785.8426$) was used for initial single-point calibration, and MS/MS fragment ions of Glu-Fib were used to obtain the final instrument calibration. Multiplexed data-independent scanning with added specificity and selectivity of a non-linear "T-wave" ion mobility ($HDMS^E$) experiments were performed using a Synapt G2-S HDMS mass spectrometer (Waters). The mass spectrometer was set to switch automatically between standard MS (3 eV) and elevated collision energies $HDMS^E$ (19–45 eV) applied to the transfer "T-wave" collision-induced dissociation cell with argon gas. The trap collision cell was adjusted for 1 eV using a millisecond scan time adjusted based on the linear velocity of the chromatography peak delivered though nanoACQUITY UPLC, to obtain a minimum of 20 scan points for each single peak at both low-energy and high-energy transmission, followed by an orthogonal acceleration time-of-flight from 50–2000 *m/z*. The radio frequency (RF) offset (MS profile) was adjusted so that the nanoUPLC-$HDMS^E$ data were effectively acquired from an *m/z* range of 400–2000, which ensured that any masses observed in the high energy spectra of less than 400 *m/z* arose from dissociations in the collision cell.

Data processing

Protein identification and quantitative data packaging were generated using dedicated algorithms [19,20], and by searching against a *C. pseudotuberculosis* database with default parameters for ion accounting [21]. The databases were reversed "on-the fly" during the database query searches, and appended to the original database to assess the false positive rate of identification. For proper processing of spectra and database searching conditions, ProteinLynxGlobalServer v.2.5.2 (PLGS) with Identity^E and Expression^E informatics v.2.5.2 (Waters) were used. UniProtKB (release 2013_01) with manually-reviewed annotations was also used, and the search conditions were based on taxonomy (*C. pseudotuberculosis*), maximum missed cleavages by trypsin allowed up to one, and variable carbamidomethyl, acetyl N-terminal, phosphoryl, and oxidation (M) modifications [21,22]. The Identity^E algorithm with Hi3 methodology was used for protein quantitation. The search threshold for accepting each individual spectrum was set to the default value, with a false-positive value of 4%. Biological variability was addressed by analyzing each culture three times. Normalization was performed using the Expression^E tool with a housekeeping protein that showed no significant difference in abundance across all injections. The proteins obtained were organized by the PLGS Expression^E tool algorithm into a statistically significant list corresponding to increased and decreased regulation ratios among the different groups. The quantitation values were averaged over all of the samples, and the quoted standard deviations at $p \leq 0.05$ in the Expression^E software refer to the differences between biological replicates. Only proteins with a differential expression $\log_2$ ratio between the two conditions greater than or equal to 1.2 were considered [23].

Bioinformatics analysis

The identified proteins were analyzed using the prediction tools SurfG+ v1.0 [24], to predict sub-cellular localization, and Blast2GO, to predict gene ontology functional annotations [25]. The PIPS software predicted proteins present in pathogenicity islands [26]. The protein-protein interaction network was constructed using interolog mapping methodology and metrics according to Rezende et al. [27]. A preview of the interaction network was generated using Cytoscape version 2.8.3 [28], with a spring-embedded layout. CMRegNet was used to predict gene regulatory networks [29].

Results

Effects of nitric oxide on the growth of *C. pseudotuberculosis*

In this study, we examined the exponential growth of *C. pseudotuberculosis* strain 1002 under nitrosative stress. The growth and cell viability of strain 1002 was monitored for 10 h with and without DETA/NO supplementation (Figure 1). The control culture reached stationary phase by 5 h post-inoculation, while the culture containing DETA/NO did not reach stationary phase until approximately 10 h post-inoculation. However, these results showed that although DETA/NO (0.5 mM) affected the growth rate, *C. pseudotuberculosis* likely contains factors that promote survival in the presence of RNS.

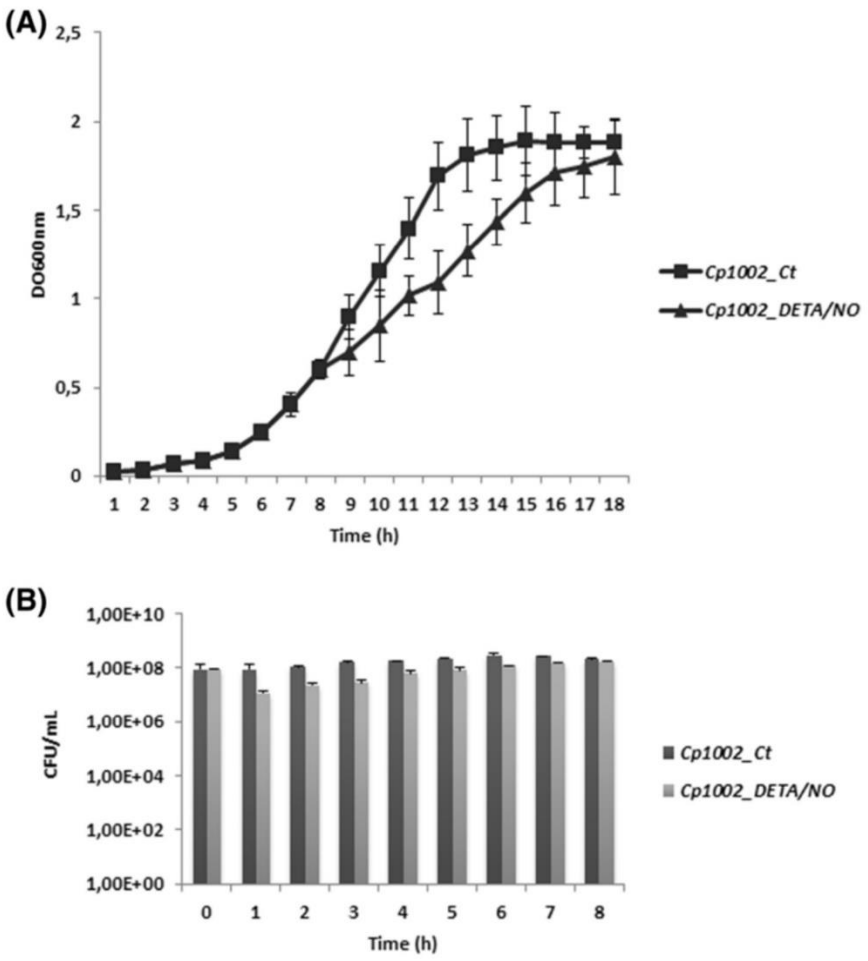


Figure 1 Growth and survival profile of *C. pseudotuberculosis* during NO exposure. (A) Growth of *C. pseudotuberculosis* after 10 h exposure to 0.5 mM DETA/NO. **(B)** Survival of *C. pseudotuberculosis* evaluated by colony forming units. The results shown in A and B represent an average of three independent experiments.

Label-free proteomic analysis of *C. pseudotuberculosis* grown under nitrosative stress conditions

Total proteome digests from three biological replicates of each individual condition were subjected to LC/MS^{E}. In total, we identified more than 31,000 peptides, with a normal distribution of 10 ppm error of the total identified peptides. Peptides as source fragments, peptides with a charge state of at least $[M + 2H]^{2+}$, and the absence of decoys were factors considered to increase data quality. A combined total of 2,063 proteins were present in at least two of the three biological replicates for the two conditions tested, with an average of 15 peptides per protein, and a false discovery rate (FDR) of 0% when decoy detection was set at agreement of two out of three replicates. The proteins referred to as exclusive to one condition or another was only identified in one condition within the detection limits of the experiment (LOD). The dynamic range of the quantified proteins is about 3 logs, and proteins unique to one condition or another were only observed above the LOD of the experiment, which was determined by the sample normalization prior to injection. Therefore, in our study, all samples were normalized using "scouting runs" taking into account the stoichiometry between the intensity and molarity proportion prior to the replicate runs per condition. The dynamic range was similar for each sample, and the total amount of sample used in fmol was nearly the same. We generate a graph of protein amounts of the identified proteins from all samples against protein ranks (Figure 2A).

After, analysis by PLGS v2.5.2 software, the 2,063 proteins originally identified in two out of three replicates were narrowed down to 699 proteins with $p \leq 0.05$. Among these proteins, 44 were up-regulated in the presence of DETA/NO, while 14 proteins were down-regulated (Table 1, Figure 2B and C). The remaining 641 proteins with $p \leq 0.05$ and $\log_2 < 1.2$ that were common to

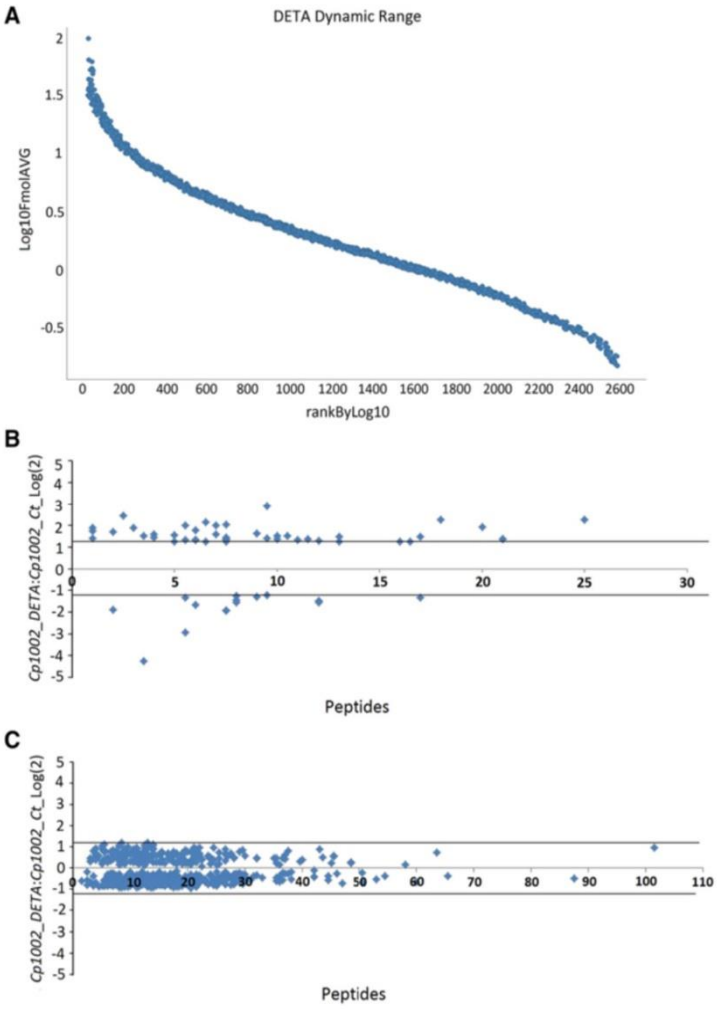


Figure 2 2D nanoUPLC HDMSE analysis showing: (A) Dynamic range of the method based on protein abundance estimates, data points derived from LC-HDMS^E analysis. (B and C) Proteins that were significantly differentially-regulated during NO exposure. The distribution of identified proteins with $p < 0.05$, and differentially-regulated proteins with an I:C $\log_2$ ratio < 1.2 in relation to the number of peptides identified for each protein. **(B)** Proteins with $p < 0.05$ and an I:C $\log_2$ ratio < 1.2. **(C)** Proteins with $p < 0.05$ and an I:C $\log_2$ ratio > 1.2.

the two treatments are summarized in Additional file 1. In addition to the 699 identified proteins that were present under both control and stress conditions, 34 proteins were exclusively expressed under the control conditions, and 102 proteins were exclusively expressed in response to DETA/NO stress (Additional files 2 and 3). Thus, our final list of proteins is composed of 835 proteins from *C. pseudotuberculosis*.

In silico analysis of LC-HDMS^E data

The 835 proteins were then analyzed using the SurfG+ tool to predict sub-cellular localization. According with SurfG+, our data set included approximately 41% of the predicted proteome of strain 1002 (Figure 3A). In addition, we characterized proteins belonging to the following cell fractions: cytoplasmic (CYT) (668 proteins), membrane (MEM) (59 proteins), potentially surface-exposed (PSE) (69 proteins), and secreted (SEC) (39 proteins) (Figure 3B).

To evaluate whether the proteins identified in our proteomic analysis could represent a protein set expressed by *C. pseudotuberculosis* during exposure to nitrosative stress, we correlated our proteomic data with the predicted core-genomes of 15 *C. pseudotuberculosis* strains [7]. Of the open reading frames (ORFs) coding for the differentially-regulated proteins and exclusive proteome of DETA/NO-exposed cells, 86% (50/58 proteins) and 82% (84/102 proteins) were identified, respectively, in the core-genome of *C. pseudotuberculosis* (Figure 3C and D). In addition, of the 835 total proteins identified from the proteome of strain 1002 following exposure to nitrosative stress, 83% (696 proteins) of the ORFs coding for these proteins were present in the core-genome of *C. pseudotuberculosis*,

this result correspond approximately 46% of the predicted core-genome of *C. pseudotuberculosis* (Figure 3E).

Functional classification of the proteome of *C. pseudotuberculosis* expressed under exposure to nitrosative stress

The strain 1002 proteome was functionally classified using the Blast2Go tool [24]. A large proportion of the differentially-regulated proteins and those exclusive to one condition were identified as hypothetical proteins. According to the biological function prediction, 18 biological processes were classified as differentially regulated (Figure 4A). In addition, the analysis of the exclusive proteome of each condition revealed 12 common processes between the control and stress conditions (Figure 4B). However, seven biological processes were identified only in stress-exposed cells. These processes were antibiotic metabolism (six proteins), nucleotide metabolism (five proteins), oxidative phosphorylation (three proteins), translation (three proteins), glycolysis pathways (one protein), iron-sulfur clusters (one protein), and starch and sucrose metabolism (one protein). Among all processes identified, DNA synthesis and repair proteins (14 proteins) were most common. An overview of the *C. pseudotuberculosis* response to nitrosative stress according with the proteins identified is shown in Figure 5.

The proteins that were grouped into of transcriptional process were evaluated by CMRegNet and among regulators identified; we identified the GntR- family regulatory protein (D9Q5B7_CORP1), genes regulated by GntR-type regulators are usually involved in carbohydrate metabolism. The CMRegNet analysis showed that of the four genes under the control of this regulator, the N-acetylglucosamine kinase (D9Q5B6_CORP1) protein was highly expressed by *C. pseudotuberculosis* in response to DETA/NO. We identified other regulator the LexA repressor (D9Q8W2_CORP1) that was down regulated in the DETA/NO condition. According with CMRegNet, two proteins regulated by this repressor were detected in the DETA/NO proteome specific, pyridoxal biosynthesis lyase (PdxS; D9Q5T9_CORP1) and DNA translocase (D9Q 8Z6_CORP1). Others proteins under the control of this repressor was detected, however not presented significant differential regulation like RecA protein

Protein-protein interaction network

To investigate the interactions among the proteins identified as exclusive and differentially regulated in cells exposed to DETA/NO, we generated a protein interaction network using Cytoscape. The interactome analysis revealed 67 protein-protein interactions (Figure 6). DnaB/DNA helicase (D9Q578_CORP1), identified in the exclusive proteome for strain 1002_*DETA/NO*, and PyrE/orotate phosphoribosyltransferase (D9Q4S2_CORP1), which was down-regulated in strain 1002_*DETA/NO*, showed the greatest number of interactions with other proteins (eight interactions each). Moreover, both of these proteins interact with proteins that are involved in metabolic processes, DNA processes, antibiotic metabolism, cell cycling, and translation.

Discussion

C. pseudotuberculosis is exposed to different forms of oxidative and nitrosative stress during the infection process. A previous study showed that *C. pseudotuberculosis* resists nitrosative stress generated by the NO-donor DETA/NO, and that a low concentration of DETA/NO (100 μM) induces a change in the extracellular proteome this pathogen [15]. To better understand the physiology of *C. pseudotuberculosis* in response to nitrosative stress, we analyzed the proteome of whole bacterial lysates of *C. pseudotuberculosis* in response to exposure to DETA/NO (0.5 mM).

The strain 1002 proteome under nitrosative stress reveals proteins involved in bacterial defense against DNA damage

Proteomic analysis identified proteins involved in DNA repair systems in both the exclusive proteome of DETA/NO-exposed cells and in the differentially-regulated proteome. We detected the proteins formamidopyrimidine-DNA glycosylase (Fpg) (D9Q598_CORP1), RecB (D9Q8 C9_CORP1), and methylated-DNA-protein-cysteine methyltransferase (Ada) (D9Q923_CORP1), the genes for which were previously identified in a transcriptome analysis of strain 1002 in response to different abiotic stresses [8]. Activation of these proteins in response to nitrosative stress confirms that they belong a group of general stress-response proteins in *C. pseudotuberculosis*.

The expression of Fpg was up-regulated in response to acid stress [8]. We also identified endonuclease III (Endo III) (D9Q615_CORP1), which, in addition to Fpg, is involved in the base excision repair (BER) system of various bacteria. This system cleaves N-glycosidic bonds from damaged bases, allowing their excision and replacement. In *Salmonella enterica* serovar Typhimurium, the BER system repairs DNA damaged by exposure to NO. In addition, an *S.* Typhimurium strain defective in Fpg demonstrated reduced virulence in a murine model [30]. Our interactome analysis showed that Endo III had one of the highest numbers of interactions with other proteins, including interactions with proteins involved in DNA replication such zinc metalloprotease (D9Q378_CORP1) and DNA translocase (D9Q8Z6_CORP1), suggesting that this protein could play an important role in the defense pathway against RNS.

The Ada and RecB protein were up-regulated in response to osmotic stress [8]. Ada is involved in the repair of DNA-methylation damage, this protein have plays important in the pathway DNA damage [31]. RecB is a component of the RecBC system, which is part of

Table 1 Proteins identified as differentially-expressed following exposure to nitrosative stress

Uniprot access	Proteins	Score	Peptides	log_2 DETA: CT[(a)]	*p*-value[(a)]	Subcellular localization[(c)]	Gene name	Genome[(b)]
Transport								
F9Y2Z3_CORP1	Cell wall channel	5321.88	4	1.42	1	CYT	*porH*	Shared
Cell division								
D9Q7G2_CORP1	Hypothetical protein	2417.8	21	1.34	1	CYT	*Cp1002_0716*	Core
DNA synthesis and repair								
D9Q5V6_CORP1	Nucleoid-associated protein	2327.08	5	1.52	1	CYT	*ybaB*	Core
D9Q923_CORP1	Methylated-DNA-protein-cysteine methyltransferase	6332.83	8	1.22	1	CYT	*ada*	Core
D9Q4P0_CORP1	7,8-dihydro-8-oxoguanine-triphosphatase	1640.23	8	−1.97	0	CYT	*mutT*	Core
Transcription								
D9Q8W2_CORP1	LexA repressor	800.31	6	−1.37	0.04	CYT	*lexA*	Shared
D9Q5L4_CORP1	ECF family sigma factor k	364.82	8	−1.58	0	CYT	*sigK*	Core
Translation								
D9Q753_CORP1	Fkbp-type peptidyl-prolyl cis-trans isomerase	7113.34	3	2.43	1	CYT	*fkbP*	Core
D9Q830_CORP1	50S ribosomal protein L35	2271.66	1	1.36	1	CYT	*rpmI*	Core
D9Q7W1_CORP1	Aspartyl glutamyl-tRNA amidotransferase subunit C	3100.8	7	1.24	0.99	CYT	*gatC*	Core
D9Q582_CORP1	50S ribosomal protein L9	41082.46	10	−1.25	0	CYT	*rplI*	
D9Q6H6_CORP1	30S ribosomal protein S8	45333.23	9	−1.34	0	CYT	*rpsH*	Core
Cell communication								
D9Q559_CORP1	Hypothetical protein	1402.27	6	1.99	1	PSE	*Cp1002_2005*	Core
D9Q5U9_CORP1	Thermosensitive gluconokinase	2068.35	7	1.96	0.99	CYT	*gntK*	Core
D9Q668_CORP1	Sensory transduction protein RegX3	2540.92	13	1.45	1	CYT	*regX3*	Core
Detoxification								
D9Q7U6_CORP1	Thioredoxin	1835.7	11	1.50	1	CYT	*trxA*	Core
D9Q4E5_CORP1	Glutathione peroxidase	1426.27	10	1.47	1	CYT	*Cp1002_1731*	Core
D9Q5T5_CORP1	Glyoxalase bleomycin resistance protein dihydroxybiphenyl dioxygenase	2417.77	11	1.28	1	CYT	*Cp1002_0124*	Shared
D9Q5N2_CORP1	NADH dehydrogenase	7030.94	12	1.25	1	CYT	*noxC*	Shared
D9Q680_CORP1	Glutaredoxin-like domain protein	292.69	2	−1.91	0	CYT	*Cp1002_0272*	Core
Glycolysis pathways								
D9Q5B6_CORP1	N-Acetylglucosamine kinase	228.69	6	1.74	0.98	CYT	*nanK*	Core
D9Q4U9_CORP1	Alcohol dehydrogenase	236.02	17	1.22	1	CYT	*adhA*	Shared
Iron-sulfur clusters								
D9Q7L6_CORP1	Ferredoxin	36927.57	7	2.10	1	CYT	*fdxA*	Core
Antibiotic resistance								
D9Q827_CORP1	Metallo-beta-lactamase superfamily protein	657.33	6	−2.95	0	CYT	*Cp1002_0937*	Core

Table 1 Proteins identified as differentially-expressed following exposure to nitrosative stress *(Continued)*

Amino acid metabolism								
D9Q622_CORP1	Phosphoserine phosphatase	949.15	9	1.58	0.99	PSE	*serB*	Core
D9Q4N1_CORP1	Carboxylate-amine ligase	205.54	16	1.24	1	CYT	*Cp1002_1819*	Core
D9Q6H4_CORP1	L-serine dehydratase I	284.11	17	−1.37	0	MEM	*sdaA*	Core
Lipid metabolism								
D9Q520_CORP1	Glycerophosphoryl diester phosphodiesterase	2417.8	21	1.34	1	PSE	*glpQ*	Core
Oxidative phosphorylation								
D9Q8I5_CORP1	Cytochrome aa3 controlling protein	676.2	6	1.28	1	MEM	*Cp1002_1095*	Core
Specific metabolic pathways								
D9Q5M9_CORP1	Inositol-3-phosphate synthase	7473.38	18	2.25	1	CYT	*ino1*	Core
D9Q721_CORP1	Hypothetical protein	4602.9	17	1.44	1	SEC	*Cp1002_0573*	Core
D9Q689_CORP1	3-Hydroxyisobutyrate dehydrogenase	2137.24	12	1.34	1	CYT	*mmsB*	Core
D9Q4X1_CORP1	Urease accessory protein UreG	1532.39	12	−1.6	0	CYT	*ureG*	Core
Nucleotide metabolism								
D9Q4S2_CORP1	Orotate phosphoribosyltransferase	2618.52	8	−1.26	0	CYT	*pyrE*	Core
Unknown function								
D9Q6Y9_CORP1	Hypothetical protein	491.89	10	2.87	1	CYT	*Cp1002_0540*	Core
D9Q6C7_CORP1	Hypothetical protein	689.6	25	2.25	1	PSE	*Cp1002_0320*	Core
D9Q3P3_CORP1	Hypothetical protein	5703.38	3	1.87	1	CYT	*Cp1002_1474*	Core
D9Q5V4_CORP1	Hypothetical protein	994.52	1	1.7	1	CYT	*Cp1002_0143*	Core
D9Q610_CORP1	Hypothetical protein	27217.36	2	1.67	1	CYT	*Cp1002_0202*	Core
D9Q8D8_CORP1	Hypothetical protein	2324.12	7	1.57	0.98	CYT	*Cp1002_1048*	Shared
D9Q6W1_CORP1	Hypothetical protein	9303.91	4	1.54	1	CYT	*Cp1002_0512*	Core
D9Q6V5_CORP1	Hypothetical protein	1346.2	4	1.5	0.99	CYT	*Cp1002_0506*	Core
D9Q5R7_CORP1	Hypothetical protein	2090.7	8	1.42	1	CYT	*Cp1002_0105*	Core
D9Q917_CORP1	Hypothetical protein	555.89	10	1.37	1	PSE	*Cp1002_1281*	Core
D9Q3P5_CORP1	Hypothetical protein	1121.7	6	1.29	1	SEC	*Cp1002_1476*	Core
D9Q7U5_CORP1	Hypothetical protein	517.06	8	1.28	1	CYT	*Cp1002_0852*	Core
D9Q7L1_CORP1	Hypothetical protein	15693.97	6	1.28	1	SEC	*Cp1002_0766*	Core
D9Q3P6_CORP1	Hypothetical protein	1729.59	5	1.22	1	CYT	*Cp1002_1477*	Core
D9Q6Z7_CORP1	Hypothetical protein	1835.7	13	1.22	1	CYT	*Cp1002_0548*	Core
D9Q8V8_CORP1	Hypothetical protein	293.23	8	−1.48	0		*Cp1002_1221*	Core
D9Q6C8_CORP1	Hypothetical protein	413.31	12	−1.52	0	PSE	*Cp1002_0321*	Core
D9Q5H0_CORP1	Hypothetical protein	12376.2	6	−1.71	0	CYT	*Cp1002_0007*	Core
D9Q4D5_CORP1	Hypothetical protein	10161.64	4	−4.29	0	CYT	*Cp1002_1721*	Shared
Others								
D9Q5N5_CORP1	Iron-regulated MEM protein	992.54	8	2.01	0	PSE	*piuB*	Core
D9Q922_CORP1	CobW/HypB/UreG, nucleotide-binding	1771.22	20	1.88	1	CYT	*Cp1002_1286*	Core

Table 1 Proteins identified as differentially-expressed following exposure to nitrosative stress *(Continued)*

D9Q8C4_CORP1	Prokaryotic ubiquitin-like protein Pup	2194.86	1	1.84	1	CYT	*pup*	Core
D9Q7B8_CORP1	Ribosomal-protein-alanine n-acetyltransferase	2791.1	10	1.34	1	CYT	*rimJ*	Shared
D9Q7K9_CORP1	Arsenate reductase	5147.54	8	1.32	1	CYT	*arsC*	Core

(a) Ratio values to: strain 1002_*DETA/NO*:strain 1002_*Ct*, Log(2) Ratio > 1.5, $p > 0.95$ = up-regulation, $p < 0.05$ = down-regulation.
(b) Core-genome analysis of 15 strains of *C. pseudotuberculosis*: shared = present in two or more strains; core = present in 15 strains of *C. pseudotuberculosis*.
(c) CYT =cytoplasmic, MEM = membrane, PSE = potentially surface-exposed, SEC = secreted.

the SOS response the more regulatory network encoded by prokaryotic involved in DNA repair [32]. The RecBC system acts in the recombination or degradative repair of arrested DNA replication forks. Studies in *S.* Typhimurium showed that *recBC* mutant strains are more attenuated than *recA* mutants in a murine model of infection [33]. In addition, unlike *recA* mutants, *recBC* mutants were susceptible to RNS [34], indicating that RecBC is highly important in the bacterial response to nitrosative stress. The LexA repressor (D9Q8W2_CORP1), which forms part of the general SOS system along with RecA [35], was down-regulated in *C. pseudotuberculosis* cells exposed to DETA/NO. We also detected the RecA protein (D9Q8Y3_CORP1); however, despite having a *p*-value <0.05, the fold-change of −0.50 showed that this protein was not activated under the experimental conditions. Studies performed in *Mycobacterium tuberculosis* showed that *recA* was not induced until cells had been exposed to DETA/NO (0.5 mM) for 4 h, but that hydrogen peroxide induced the immediate expression of *recA* [36], suggesting

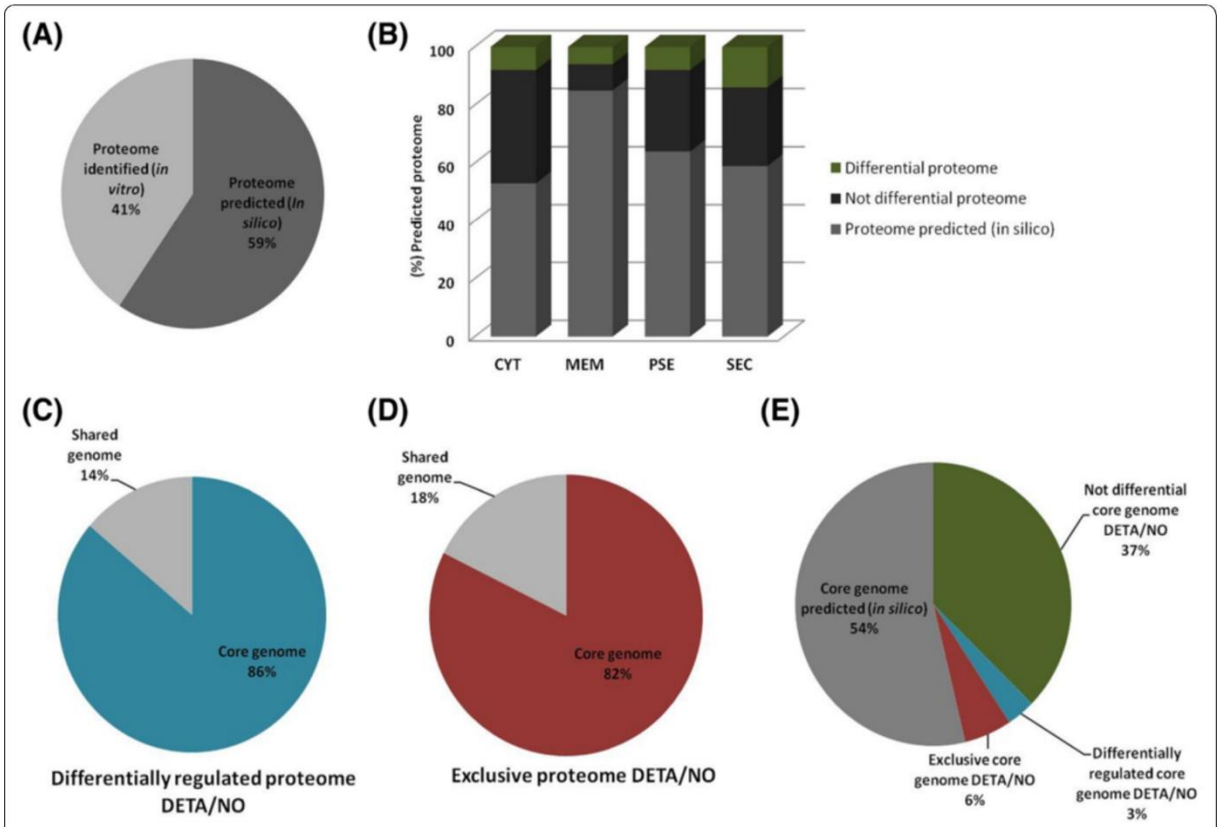


Figure 3 Correlation of *in silico* predicted data with proteome results. (A) Percentage of coverage of the *C. pseudotuberculosis* 1002 *in silico* proteome. **(B)** Prediction of the subcellular localization of the proteins identified by LC/MS. **(C)** Analysis of the differentially-regulated proteins of cells exposed to DETA/NO in relation to the core genome of *C. pseudotuberculosis* (shared genome: present in only two strains; core genome: present in 15 strains of *C. pseudotuberculosis*). **(D)** Analysis of the exclusive proteome of cells exposed to DEA/NO in relation to the core-genome of *C. pseudotuberculosis* (shared genome: present in only two strains; core genome: present in 15 strains of *C. pseudotuberculosis*). **(E)** Percent coverage of the core-genome of *C. pseudotuberculosis* in relation to the characterized proteome *in vitro*.

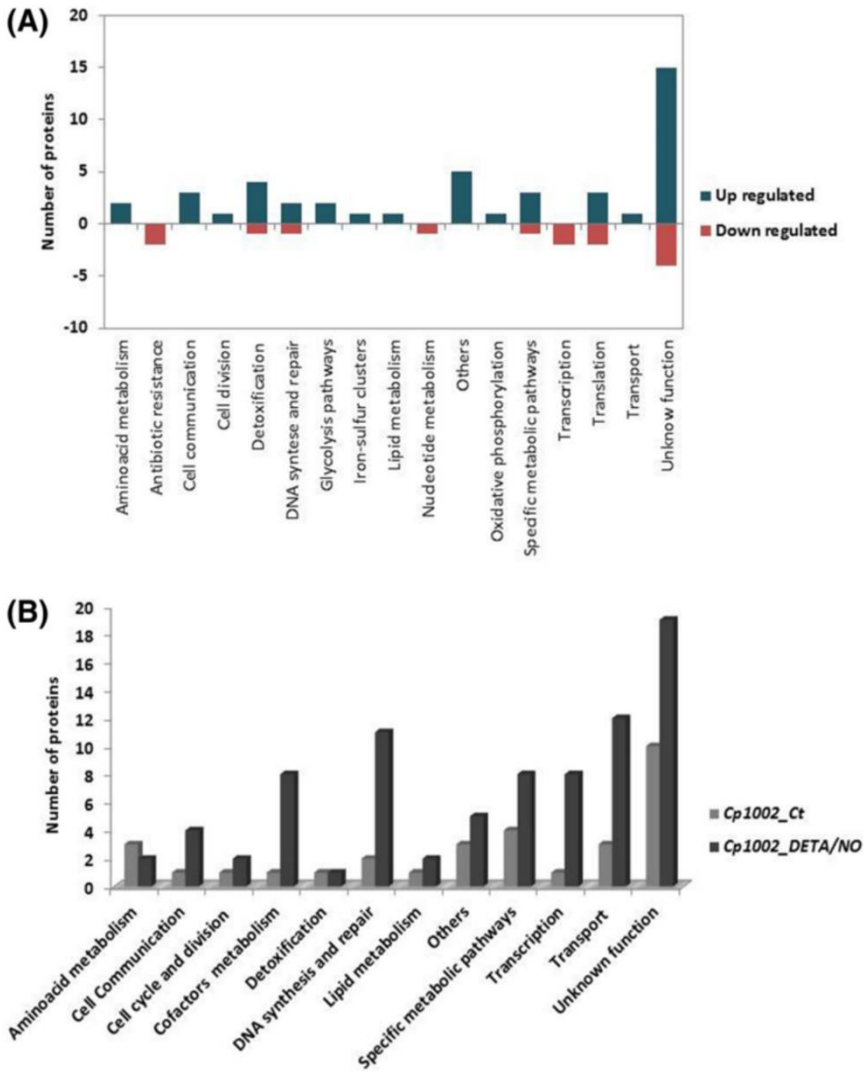


Figure 4 Comparison of biological processes between control and DETA/NO conditions. A representation of the biological processes in relation to a set list of proteins identified as **(A)** differentially-regulated in DETA/NO-stressed cells and **(B)** comparison of exclusive biological process between the two test conditions.

that RecA is involved in the later stages of the nitrosative stress response. Nevertheless, CMRegNet analysis identified other proteins that are regulated by LexA in the DETA/NO-specific proteome, including pyridoxal biosynthesis lyase (PdxS; D9Q5T9_CORP1) and DNA translocase (D9Q8Z6_CORP1).

NO-sensitive transcriptional regulators are activated in the presence of NO

To activate these DNA repair systems, it is essential that bacteria can detect ROS and RNS, and concomitantly activate the transcriptional regulators needed for the expression of genes involved in protection against these compounds. In the DETA/NO-specific proteome, we detected the transcription factor WhiB (D9Q6Y2_CORP1). The WhiB transcriptional family is composed of iron-sulfur (Fe-S) cluster proteins. These proteins are O_2- and NO-sensitive, and allow the sensing of both external environmental signals and the redox state for intracellular bacteria [37,38]. In *M. tuberculosis*, the reaction of the iron-sulfur cluster of WhiB3 with NO generates a dinitrosyl iron complex (DNIC), which activates a sensing mechanism in response to the NO, consequently activating a system of defense against nitrosative stress [12]. In addition, other *in vivo* and *in vitro* studies have also demonstrated that WhiB regulators play a role in the adaptation and survival of *M. tuberculosis* during exposure to redox environments [12,39-41].

We identified other regulators that are activated in response to environmental stimuli, such as a MerR-family transcriptional regulator (D9Q889_CORP1) and a LysR-

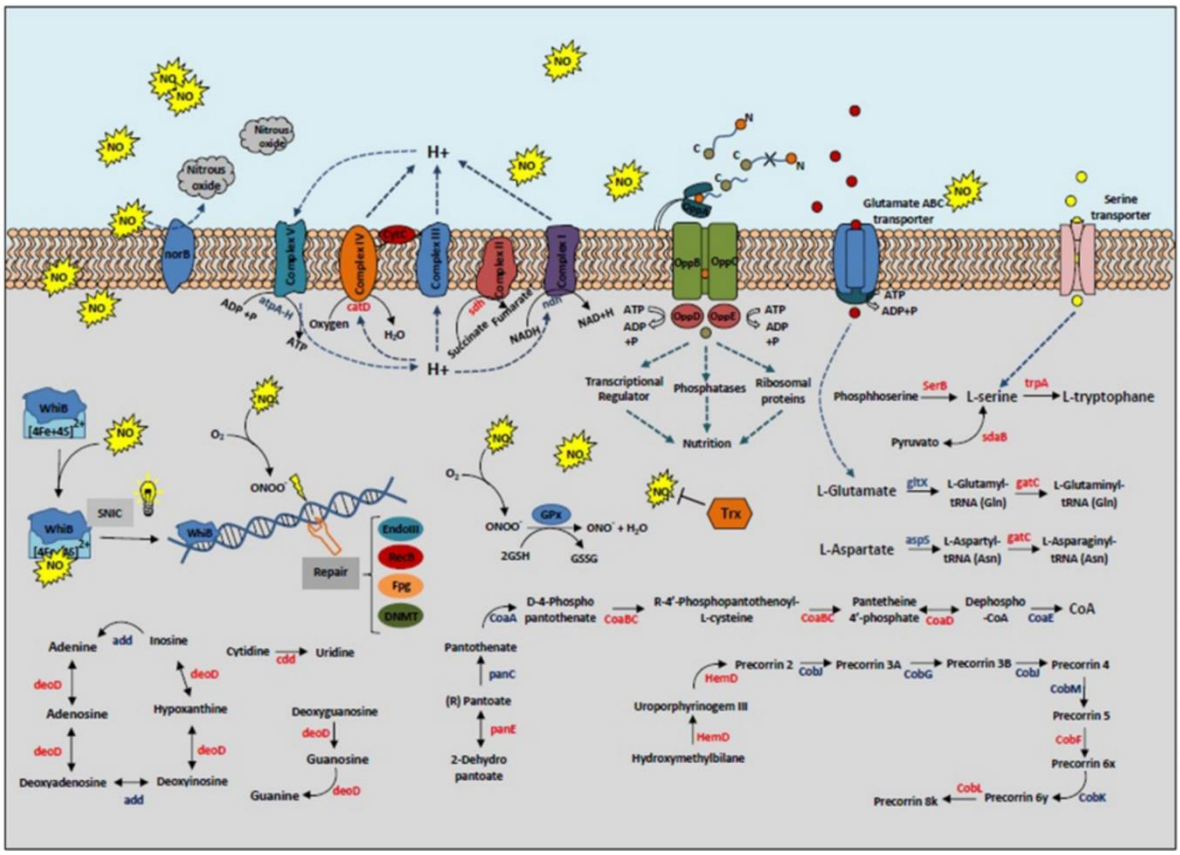


Figure 5 Overview of *C. pseudotuberculosis* response to nitrosative stress. All proteins detected by proteomic analysis are marked in red (differentially-regulated proteins or exclusive to the proteome of DETA/NO-stressed cells).

type transcriptional regulator (LTTR) (D9Q7H8_CORP1). This regulator was also highly expressed in the transcriptional response of *C. pseudotuberculosis* 1002 to acid stress [8]. MerR-type regulators have been described in the detoxification of toxic metal in several pathogenic and non-pathogenic bacteria [42]. Other studies have shown that this class of regulator plays a role in bacterial resistance to oxidative and nitrosative stress [43,44]. LTTRs are associated with the regulation of several biological processes, as well as in the adaptive response of bacteria to different types of stress [45]. In *Vibrio cholerae*, LTTRs are associated with efflux pump regulation, which contribute to antimicrobial resistance, and are involved in colonization of the human host [46]. In pathogens like *E. coli* [47], *Enterococcus faecalis* [48], *S. enterica* [49], and *Pseudomonas aeruginosa* [50], LTTRs are involved in resistance to oxidative stress.

The detoxification pathways of *C. pseudotuberculosis* following NO exposure

Our proteomic analysis identified proteins specifically expressed by cells exposed to DETA/NO that are involved in the detoxification process. Two of these proteins were thioredoxin (*trxA*) (D9Q7U6_CORP1) and glutathione peroxidase (D9Q4E5_CORP1). The thioredoxin and glutathione systems play major roles in thiol and disulfide balance, respectively [14]. In pathogens such as *Helicobacter pylori*, *Streptococcus pyogenes*, and *M. tuberculosis*, this system is of great importance in combating the presence of ROS/RNS [36,51,52]. A glyoxalase/dioxygenase (D9Q5T5_CORP1) was identified in the differential proteome of cells exposed to DETA/NO. This protein was previously detected in the proteome of *C. pseudotuberculosis* strain 1002 in response to 0.1 mM DETA/NO [15]. The presence of this protein suggests that glyoxalase/dioxygenase plays a role in the resistance of this pathogen to nitrosative stress.

Nevertheless, unlike *P. aeruginosa*, which contains a complete denitrification pathway [53], the predicted genome of *C. pseudotuberculosis ovis* 1002 revealed a truncated denitrification pathway. However, we detected the nitric-oxide reductase cytochrome b (NorB) (D9Q5T6_CORP1) in the exclusive proteome of DETA/NO-stressed cells. *norB*, which codes for this nitric-oxide reductase, is organized into the *norCBQDEF* operon in *Paracoccus*

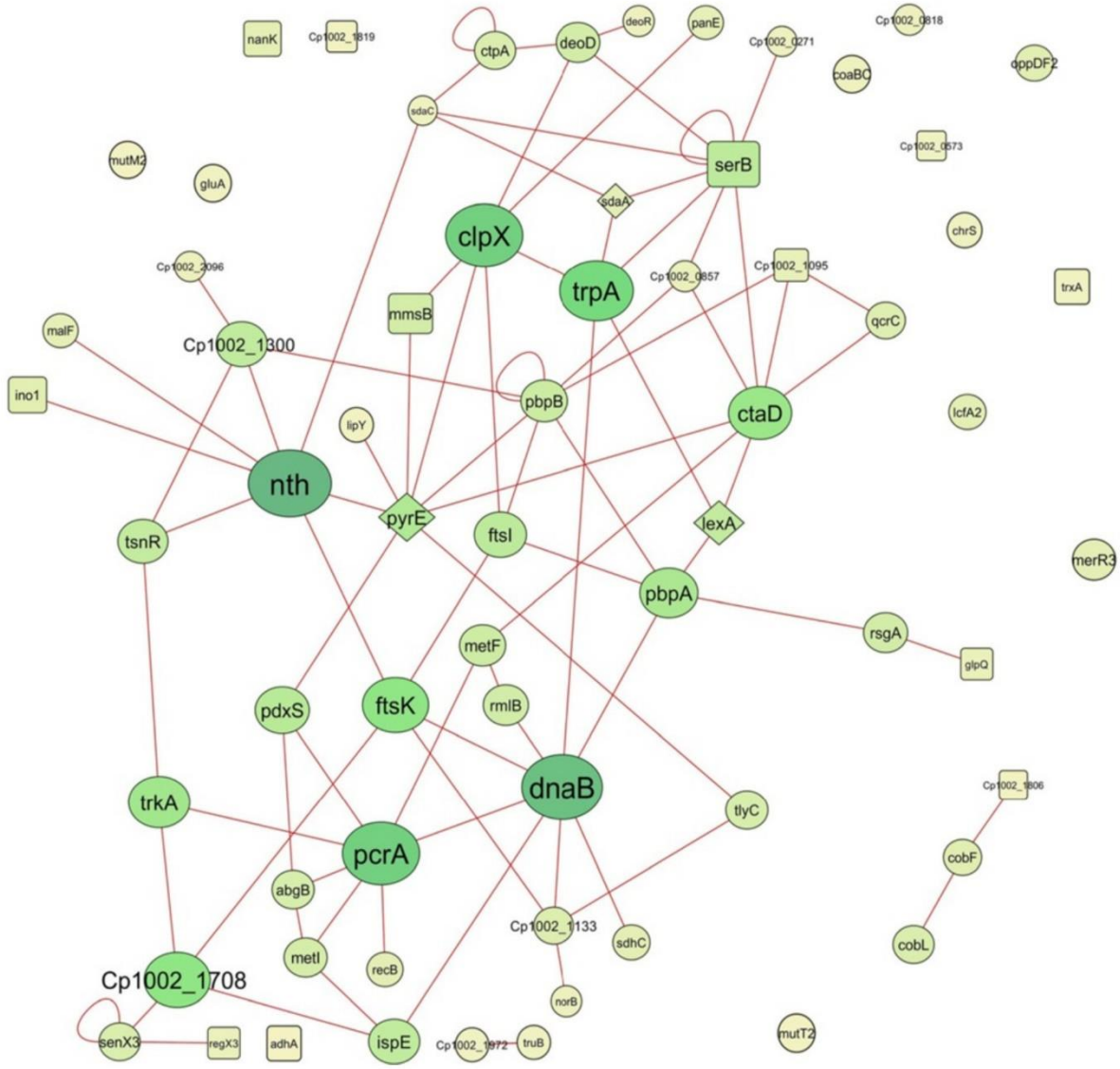


Figure 6 Protein-protein interactions. Protein-protein interactions of the proteins identified in DETA/NO-exposed cells. Exclusive proteome, circle; up-regulated, square; and down-regulated, rhombus. The sizes of the nodes represent the degree of interaction for each gene/protein; the major nodes demonstrate greater interactions. The colors of nodes and lines are in an increasing gradient scale from yellow to green to blue. The networks were visualized using Cytoscape.

denitrificans [54], and into the *norCBD* operon in *P. aeruginosa* [55]. The *C. pseudotuberculosis* genome was predicted to only contain *norB*. Moreover, *norB* is located in the *Cp1002PiCp12* pathogenicity island, suggesting horizontal acquisition of the gene by this pathogen. Nitric-oxide reductase is an important protein in the denitrification process of some bacteria [56]. In *P. aeruginosa*, NorB plays a role in both the growth of the pathogen in the presence of NO, and in its survival in macrophages [55]. The flavohemoglobin Hmp is involved in the NO detoxification pathway in *S.* Typhimurium, and levels of Hmp are increased approximately two-fold in macrophages [57]. Interestingly, in *N. meningitidis,* NorB levels are increased ten-fold in macrophages [58], demonstrating the great power of this protein in the detoxification process.

Metabolic profile of *C. pseudotuberculosis* in response to nitrosative stress

In addition to the presence of proteins involved in bacterial defense and detoxification pathways, strain 1002 needs to undergo metabolic adaptation to favor bacterial survival. We observed a metabolic readjustment in this pathogen in the proteomic analysis. Of the proteins

involved in central carbohydrate metabolism, we detected only phosphoglycerate mutase (D9Q533_CORP1) and N-acetylglucosamine kinase (D9Q5B6_CORP1) in the proteome of DETA/NO-exposed cells. Other essential proteins involved in glycolysis (the Embdem-Meyerhof pathway), the pentose phosphate pathway, and the citric acid cycle were not detected. Similar results were found in a metabolomic study of *V. cholerae* in response to nitrosative stress [59].

However, we hypothesized that *C. pseudotuberculosis* uses oxidative phosphorylation to obtain energy. This is supported by the presence of cytochrome C oxidase polypeptide I (D9Q486_CORP1), succinate dehydrogenase cytochrome b556 subunit (D9Q650_CORP1), and ubiquinol-cytochrome C reductase cytochrome C subunit (D9Q3J7_CORP1) in the exclusive proteome of DETA/NO-stressed cells, and by the up-regulation of the cytochrome oxidase assembly protein (D9Q8I5_CORP1) under the same conditions. However, this oxidative phosphorylation may be associated with the bacterial culture conditions used in this work, in which *C. pseudotuberculosis* was cultivated in the presence of DETA/NO under aerobic conditions. Studies have shown that growing *M. tuberculosis* in a low concentration of NO with low levels of O_2 can induce anaerobic respiration as a result of the inhibition of the respiratory proteins cytochrome *c* oxidase and NADH reductase by irreversible ligation of NO. The ligation of NO to the respiratory proteins is an effect that may be both short-term reversible and long-term irreversible [60]. Thus, we suggest that activation of the oxidative phosphorylation system may be a more effective pathway for this pathogen to obtain energy [61].

Another metabolic adjustment was observed in relation to amino acid biosynthesis. Transporters and enzymes involved in the synthesis of methionine, tryptophan, and serine were identified. However, the presence of these proteins can be associated with the bioavailability of these amino acids during exposure to NO. In addition, we detected two oligopeptide transport ATP-binding proteins (OppD) (D9Q6G5_CORP1/D9Q3X0_CORP1) that compose the oligopeptide permease system (Opp). This complex is associated with the internalization of peptides from the extracellular environment to be used as a source of carbon and nitrogen in bacterial nutrition [62]. We also identified proteins that are cofactors of metabolism, such as CoaBC (D9Q8L2_CORP1), phosphopantetheine adenylyltransferase (D9Q809_CORP1), and 2-dehydropantoate 2-reductase (D9Q7J9_CORP1). The presence of these proteins demonstrates activity in pantothenic acid metabolism and the biosynthesis of coenzyme A (CoA). Studies performed in species such as *Corynebacterium diphtheriae* [63], *Streptococcus haemolyticus* [64], and *M. tuberculosis* [65] showed that pantothenic acid and CoA could have an important role in the growth and viability of these pathogens.

Conclusions

In this work, we applied high-throughput proteomics to characterize the proteome of *C. pseudotuberculosis ovis* 1002 following exposure to NO. Our proteomic analysis generated two profiles, which together validated findings from previous *in silico* analyses of *C. pseudotuberculosis ovis* 1002. The proteomic profile generated after the addition of the NO-donor, DETA/NO (0.5 mM), revealed a set of proteins that are involved in distinct biological process. We detected proteins related to both the general stress response and to a more specific nitrosative stress response, which together form a network of factors that promote the survival of this pathogen under stress conditions. However, more detailed studies are needed to assess the true role of these proteins in response to nitrosative stress in *C. pseudotuberculosis*. In conclusion, this functional analysis of the genome of *C. pseudotuberculosis* shows the versatility of this pathogen in the presence of NO. Moreover, the results presented in this study provide insights into the processes of resistance of *C. pseudotuberculosis* during exposure to nitrosative stress.

Additional files

Additional file 1: Table S1. Complete list of proteins identified as significantly altered ($p < 0.05$).

Additional file 2: Table S2. Unique proteins identified in strain 1002_*DETA/NO*.

Additional file 3: Table S3. Unique proteins identified in strain 1002 control condition.

Abbreviations

G: Guanine; C: Citosine; NO: Nitric oxide; RNS: Reactive nitrogen species; NOS: Nitric oxide synthases; LC-HDMS^E: Liquid chromatograph high definition mass spectrometry; LC/MS: Liquid chromatograph mass spectrometry; CDM: Chemically-defined médium; DETA/NO: Diethylenetriamine/nitric oxide adduct; DTT: Dithiothreitol; 2D-RP: Two-dimensional reversed phase; nanoUPLC: Nano Ultra performance liquid chromatography; nanoESI-HDMS: Nano electrospray high definition mass spectrometry; HSS: High strength silica; PLGS: Protein lynx global server; FDR: False discovery rate.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

WMS, RDC, and IFSB performed microbiological analyses and sample preparation for proteomic analysis. GHMFS and WMS conducted the proteomic analysis. SCS and ELF performed bioinformatics analysis of the data. YLL and AM contributed substantially to data interpretation and revisions. VA and AS participated in all steps of the project as coordinators, and critically reviewed the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Genomics and Proteomics Network of the State of Pará of the Federal University of Pará, the Amazon Research Foundation (FAPESPA), the National Council for Scientific and Technological

Development (CNPq), the Brazilian Federal Agency for the Support and Evaluation of Graduate Education (CAPES), the Minas Gerais Research Foundation (FAPEMIG), and Waters Corporation, Brazil. Yves Le Loir is the recipient of a PVE grant (71/2013) from Programa Ciências sem Fronteiras.

Author details
[1]Depto de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [2]Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, Pará, Brazil. [3]Waters Corporation, MS Applications and Development Laboratory, São Paulo, Brazil. [4]Institut National de la Recherche Agronomique - INRA, UMR1253 STLO, Rennes 35042, France. [5]Agrocampus Ouest, UMR1253 STLO, Rennes 35042, France.

Received: 4 September 2014 Accepted: 24 November 2014
Published: 4 December 2014

References

1. Dorella FA, Pacheco LG, Oliveira SC, Miyoshi A, Azevedo V: ***Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence.** *Vet Res* 2006, **37:**201–218.
2. Baird GJ, Fontaine MC: ***Corynebacterium pseudotuberculosis* and its role in ovine caseous lymphadenitis.** *J Comp Pathol* 2007, **137:**179–210.
3. Hodgson AL, Bird P, Nisbet IT: **Cloning, nucleotide sequence, and expression in *Escherichia coli* of the phospholipase D gene from *Corynebacterium pseudotuberculosis*.** *J Bacteriol* 1990, **172:**1256–1261.
4. Hard GC: **Comparative toxic effect of the surface lipid of *Corynebacterium ovis* on peritoneal macrophages.** *Infect Immun* 1975, **12:**1439–1449.
5. Billington SJ, Esmay PA, Songer JG, Jost BH: **Identification and role in virulence of putative iron acquisition genes from *Corynebacterium pseudotuberculosis*.** *J Bacteriol* 2002, **180:**3233–3236.
6. Ruiz JC, D'Afonseca V, Silva A, Ali A, Pinto AC, Santos AR, Rocha AA, Lopes DO, Dorella FA, Pacheco LG, Costa MP, Turk MZ, Seyffert N, Moraes PM, Soares SC, Almeida SS, Castro TL, Abreu VA, Trost E, Baumbach J, Tauch A, Schneider MP, McCulloch J, Cerdeira LT, Ramos RT, Zerlotini A, Dominitini A, Resende DM, Coser EM, Oliveira LM, *et al*: **Evidence for reductive genome evolution and lateral acquisition of virulence functions in two *Corynebacterium pseudotuberculosis* strains.** *PLoS One* 2011, **18:**e18551.
7. Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, Ali A, Santos AR, Pinto AC, Diniz C, Barbosa EG, Dorella FA, Aburjaile F, Rocha FS, Nascimento KK, Guimarães LC, Almeida S, Hassan SS, Bakhtiar SM, Pereira UP, Abreu VA, Schneider MP, Miyoshi A, Tauch A, Azevedo V: **The pan-genome of the animal pathogen *Corynebacterium pseudotuberculosis* reveals differences in genome plasticity between the biovar *ovis* and *equi* strains.** *PLoS One* 2013, **8:**e53818.
8. Pinto AC, de Sá PH, Ramos RT, Barbosa S, Barbosa HP, Ribeiro AC, Silva WM, Rocha FS, Santana MP, de Paula Castro TL, Miyoshi A, Schneider MP, Silva A, Azevedo V: **Differential transcriptional profile of *Corynebacterium pseudotuberculosis* in response to abiotic stresses.** *BMC Genomics* 2014, **9:**14.
9. Marletta MA: **Nitric oxide synthase: aspects concerning structure and catalysis.** *Cell* 1994, **23:**927–930.
10. Griffith OW, Stueh DJ: **Nitric oxide synthases: properties and catalytic mechanism.** *Annu Rev Physiol* 1995, **57:**707–736.
11. Nathan C, Shiloh MU: **Reactive oxygen and nitrogen intermediates in the relationship between mammalian hosts and microbial pathogens.** *Proc Natl Acad Sci USA* 2000, **1:**8841–8848.
12. Singh A, Guidry L, Narasimhulu KV, Mai D, Trombley J, Redding KE, Giles GI, Lancaster JR Jr, Steyn AJ: ***Mycobacterium tuberculosis* WhiB3 responds to O_2 and nitric oxide via its [4Fe-4S] cluster and is essential for nutrient starvation survival.** *Proc Natl Acad Sci USA* 2007, **10:**11562–11567.
13. Ehrt S, Schnappinger D: **Mycobacterial survival strategies in the phagosome: defence against host stresses.** *Cell Microbiol* 2009, **11:**1170–1178.
14. Lu J, Holmgren A: **The thioredoxin antioxidant system.** *Free Radic Biol Med* 2013, **8:**75–87.
15. Pacheco LG, Castro TL, Carvalho RD, Moraes PM, Dorella FA, Carvalho NB, Slade SE, Scrivens JH, Feelisch M, Meyer R, Miyoshi A, Oliveira SC, Dowson CG, Azevedo V: **A role for sigma factor $\sigma^{(E)}$ in *Corynebacterium pseudotuberculosis* resistance to nitric oxide/peroxide stress.** *Front Microbiol* 2012, **3:**126.
16. Moura-Costa LF, Paule BJA, Freire SM, Nascimento I, Schaer R, Regis LF, Vale VLC, Matos DP, Bahia RC, Carminati R, Meyer R: **Chemically defined synthetic medium for *Corynebacterium pseudotuberculosis* culture.** *Rev Bras Saúde Prod An* 2002, **3:**1–9.
17. Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ: **Absolute quantification of proteins by LC/MS^E: a virtue of parallel MS acquisition.** *Mol Cell Proteomics* 2006, **5:**144–156.
18. Gilar M, Olivova P, Daly AE, Gebler JC: **Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions.** *J Sep Sci* 2005, **28:**1694–1703.
19. Silva JC, Denny R, Dorschel CA, Gorenstein M, Kass IJ, Li GZ, McKenna T, Nold MJ, Richardson K, Young P, Geromanos S: **Quantitative proteomic analysis by accurate mass retention time pairs.** *Anal Chem* 2005, **77:**2187–2000.
20. Geromanos SJ, Vissers JP, Silva JC, Dorschel CA, Li GZ, Gorenstein MV, Bateman RH, Langridge JI: **The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS.** *Proteomics* 2009, **9:**1683–1695.
21. Li GZ, Vissers JP, Silva JC, Golick D, Gorenstein MV, Geromanos SJ: **Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures.** *Proteomics* 2009, **9:**1696–1719.
22. Curty N, Kubitschek-Barreira PH, Neves GW, Gomes D, Pizzatti L, Abdelhay E, Souza GH, Lopes-Bezerra LM: **Discovering the infectome of human endothelial cells challenged with *Aspergillus fumigatus* applying a mass spectrometry label-free approach.** *J Proteomics* 2014, **31:**126–140.
23. Levin Y, Hadetzky E, Bahn S: **Quantification of proteins using data-independent analysis (MSE) in simple and complex samples: a systematic evaluation.** *Proteomics* 2011, **11:**3273–3287.
24. Barinov A, Loux V, Hammani A, Nicolas P, Langella P, Ehrlich D, Maguin E, van de Guchte M: **Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other Gram-positive bacteria.** *Proteomics* 2009, **9:**61–73.
25. Conesa A, Gotz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **15:**3674–3676.
26. Soares SC, Abreu VA, Ramos RT, Cerdeira L, Silva A, Baumbach J, Trost E, Tauch A, Hirata R Jr, Mattos-Guaraldi AL, Miyoshi A, Azevedo V: **PIPs: pathogenicity island prediction software.** *PLoS One* 2012, **7:**e30848.
27. Rezende AM, Folador EL, Resende Dde M, Ruiz JC: **Computational prediction of protein-protein interactions in *Leishmania* predicted proteomes.** *PLoS One* 2012, **7:**e51304.
28. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498–2504.
29. Pauling J, Röttger R, Tauch A, Azevedo V, Baumbach J: **CoryneRegNet 6.0 -Updated database content, new analysis methods and novel features focusing on community demands.** *Nucleic Acids Res* 2012, **40:**D610–614.
30. Richardson AR, Soliven KC, Castor ME, Barnes PD, Libby SJ, Fang FC: **The base excision repair system of *Salmonella enterica* serovar *typhimurium* counteracts DNA damage by host nitric oxide.** *PLoS Pathog* 2009, **5:**e1000451.
31. Sedgwick B: **Repairing DNA-methylation damage.** *Nat Rev Mol Cell Biol* 2004, **5:**148–157.
32. Baharoglu Z, Mazel D: **SOS, the formidable strategy of bacteria against aggressions.** *FEMS Microbiol Rev* 2014, **38:**1126–1145.
33. Cano DA, Pucciarelli MG, García-del Portillo F, Casadesús J: **Role of the *recBCD* recombination pathway in *Salmonella* virulence.** *J Bacteriol* 2002, **184:**592–595.
34. Koskiniemi S, Andersson DI: **Translesion DNA polymerases are required for spontaneous deletion formation in *Salmonella typhimurium*.** *Proc Natl Acad Sci U S A.* 2009, **23:**10248–10253.
35. Butala M, Zgur-Bertok D, Busby SJ: **The bacterial LexA transcriptional repressor.** *Cell Mol Life Sci* 2009, **66:**82–93.
36. Voskuil MI, Bartek IL, Visconti K, Schoolnik GK: **The response of *Mycobacterium tuberculosis* to reactive oxygen and nitrogen species.** *Front Microbiol* 2011, **13:**105.
37. Green J, Paget MS: **Bacterial redox sensors.** *Nat Rev Microbiol* 2004, **2:**954–966.
38. Green J, Rolfe MD, Smith LJ: **Transcriptional regulation of bacterial virulence gene expression by molecular oxygen and nitric oxide.** *Virulence* 2014, **4:**5(4).
39. Singh A, Crossman DK, Mai D, Guidry L, Voskuil MI, Renfrow MB, Steyn AJ: ***Mycobacterium tuberculosis* WhiB3 maintains redox homeostasis by regulating virulence lipid anabolism to modulate macrophage response.** *PLoS Pathog* 2009, **5:**e1000545.
40. Chawla M, Parikh P, Saxena A, Munshi M, Mehta M, Mai D, Srivastava AK, Narasimhulu KV, Redding KE, Vashi N, Kumar D, Steyn AJ, Singh A:

Mycobacterium tuberculosis WhiB4 regulates oxidative stress response to modulate survival and dissemination *in vivo*. *Mol Microbiol* 2012, **85**:1148–1165.
41. Larsson C, Luna B, Ammerman NC, Maiga M, Agarwal N, Bishai WR: **Gene expression of *Mycobacterium tuberculosis* putative transcription factors WhiB1-7 in redox environments.** *PLoS One* 2012, **7**:e37516.
42. Hobman JL: **MerR family transcription activators: similar designs, different specificities.** *Mol Microbiol* 2007, **63**:1275–1278.
43. McEwan AG, Djoko KY, Chen NH, Couñago RL, Kidd SP, Potter AJ, Jennings MP: **Novel bacterial MerR-like regulators their role in the response to carbonyl and nitrosative stress.** *Adv Microb Physiol* 2011, **58**:1–22.
44. Brown NL, Stoyanov JV, Kidd SP, Hobman JL: **The MerR family of transcriptional regulators.** *FEMS Microbiol Rev* 2003, **27**:145–163.
45. Maddocks SE, Oyston PC: **Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins.** *Microbiology* 2008, **154**:3609–3623.
46. Chen S, Wang H, Katzianer DS, Zhong Z, Zhu J: **LysR family activator-regulated major facilitator superfamily transporters are involved in *Vibrio cholerae* antimicrobial compound resistance and intestinal colonisation.** *Int J Antimicrob Agents* 2013, **41**:188–192.
47. Gonzalez-Flecha B, Demple B: **Role for the *oxyS* gene in regulation of intracellular hydrogen peroxide in *Escherichia coli*.** *J Bacteriol* 1999, **181**:3833–3836.
48. Verneuil N, Rincé A, Sanguinetti M, Posteraro B, Fadda G, Auffray Y, Hartke A, Giard JC: **Contribution of a PerR-like regulator to the oxidative-stress response and virulence of *Enterococcus faecalis*.** *Microbiology* 2005, **151**:3997–4004.
49. Lahiri A, Das P, Chakravortty D: **The LysR-type transcriptional regulator Hrg counteracts phagocyte oxidative burst and imparts survival advantage to *Salmonella enterica* serovar Typhimurium.** *Microbiology* 2008, **154**:2837–2846.
50. Reen FJ, Haynes JM, Mooij MJ, O'Gara F: **A non-classical LysR-type transcriptional regulator PA2206 is required for an effective oxidative stress response in *Pseudomonas aeruginosa*.** *PLoS One* 2013, **8**:e54479.
51. Comtois SL, Gidley MD, Kelly DJ: **Role of the thioredoxin system and the thiol-peroxidases Tpx and Bcp in mediating resistance to oxidative and nitrosative stress in *Helicobacter pylori*.** *Microbiology* 2003, **149**:121–129.
52. Brenot A, King KY, Janowiak B, Griffith O, Caparon MG: **Contribution of glutathione peroxidase to the virulence of *Streptococcus pyogenes*.** *Infect Immun* 2004, **72**:408–413.
53. Kalkowski I, Conrad R: **Metabolism of nitric oxide in denitrifying *Pseudomonas aeruginosa* and nitrate-respiring *Bacillus cereus*.** *FEMS Microbiol Lett.* 1991, **15**:107–111.
54. de Boer AP, van der Oost J, Reijnders WN, Westerhoff HV, Stouthamer AH, van Spanning RJ: **Mutational analysis of the *nor* gene cluster which encodes nitric-oxide reductase from *Paracoccus denitrificans*.** *Eur J Biochem* 1996, **15**:592–600.
55. Kakishima K, Shiratsuchi A, Taoka A, Nakanishi Y, Fukumori Y: **Participation of nitric oxide reductase in survival of *Pseudomonas aeruginosa* in LPS-activated macrophages.** *Biochem Biophys Res Commun* 2007, **6**:587–591.
56. Hendriks J, Oubrie A, Castresana J, Urbani A, Gemeinhardt S, Saraste M: **Nitric oxide reductases in bacteria.** *Biochim Biophys Acta* 2000, **15**:266–273.
57. Stevanin TM, Poole RK, Demoncheaux EA, Read RC: **Flavohemoglobin Hmp protects *Salmonella enterica* serovar Typhimurium from nitric oxide-related killing by human macrophages.** *Infect Immun* 2002, **70**:4399–4405.
58. Stevanin TM, Moir JW, Read RC: **Nitric oxide detoxification systems enhance survival of *Neisseria meningitidis* in human macrophages and in nasopharyngeal mucosa.** *Infect Immun* 2005, **73**:3322–3329.
59. Stern AM, Liu B, Bakken LR, Shapleigh JP, Zhu J: **A novel protein protects bacterial iron-dependent metabolism from nitric oxide.** *J Bacteriol* 2013, **195**:4702–4708.
60. Brown GC: **Regulation of mitochondrial respiration by nitric oxide inhibition of cytochrome C oxidase.** *Biochim Biophys Acta* 2001, **1**:46–57.
61. Kadenbach B: **Intrinsic and extrinsic uncoupling of oxidative phosphorylation.** *Biochim Biophys Acta* 2003, **5**:77–94.
62. Payne JW, Smith MW: **Peptide transport by micro-organisms.** *Adv Microb Physiol* 1994, **36**:1–80.
63. Mueller JH, Klotz AW: **Pantothenic acid as a growth factor for the diphtheria bacillus.** *J Am Chem Soc* 1938, **60**:3086–3087.
64. McIlwain H: **Pantothenic acid and the growth of *Streptococcus haemolyticus*.** *Br J Exp Pathol* 1939, **20**:330–333.
65. Sassetti CM, Boyd DH, Rubin EJ: **Genes required for mycobacterial growth defined by high density mutagenesis.** *Mol Microbiol* 2003, **48**:77–84.

doi:10.1186/1471-2164-15-1065
Cite this article as: Silva *et al.*: Label-free proteomic analysis to confirm the predicted proteome of *Corynebacterium pseudotuberculosis* under nitrosative stress mediated by nitric oxide. *BMC Genomics* 2014 **15**:1065.

4 - Discussão Geral

Como resultado do trabalho desenvolvido nesta tese, obtivemos dois resultados principais. O primeiro resultado foi a validação de uma metodologia genérica para a predição de interação proteína-proteína, descrito no capítulo de metodologia. O segundo resultado foi obtido com a aplicação desta metodologia validada para a predição das redes de interação para nove linhagens de *Corynebacterium pseudotuberculosis* biovar *ovis*.

No primeiro trabalho, objetivamos identificar e validar métricas, extraídas dos valores dos alinhamentos feito pelo BLASTp, que pudessem ser usadas para diferenciar interações falsas e positivas. Para isto, usamos a base de dados pública DIP, contendo interações experimentais e curadas, como padrão ouro. Usamos também as bases de dados públicas (pDB) String, Intact e Psibase para mapearmos as interações. Assim, usando o programa BLASTp e as sequências de aminoácidos cada interação em formato FASTA, fizemos o alinhamento recíproco, mapeamos e transferimos as interações encontradas nas pDB para DIP. Sendo DIP nosso padrão ouro, contabilizamos estatisticamente as interações falsas e verdadeiras. Como DIP contém somente interações verdadeiras, o conjunto de interações negativas foi criado com identificadores da mesma base de dados, contendo em proporção de cinco vezes a quantidade de interações positivas, criadas aleatoriamente.

Para isto, geramos dois conjuntos de dados distintos para serem avaliados, ambos contendo os alinhamentos recíprocos entre as pDB e DIP, gerados pelo BLASTp. No primeiro conjunto de alinhamentos, somente o primeiro alinhamento do BLASTp foi considerado, justificado pela maior probabilidade de ser uma proteína homóloga. No segundo conjunto de alinhamentos, foram considerados os 20 primeiros alinhamentos do BLASTp, visando assim, identificar outros alinhamentos entre proteínas homólogas. Para ambos conjuntos de dados, os valores dos alinhamentos retornados pelo BLASTp foram recuperados, sendo eles o *score, e-value, bitscore,* similaridade, identidade e cobertura. Adicionalmente, geramos subconjuntos com combinações dos valores obtidos dos alinhamentos feitos com o BLASTp. Assim, no total foram gerados 42 subconjuntos distintos de predições a serem avaliados (dois conjuntos de dados com sete métricas para três pDB).

Cada subconjunto ou combinação destes foram submetidos a avaliação com a curva *Receiver Operaing Characteristic* (ROC), visando identificar a métrica com maior *Area Under Curve* (AUC) que pudesse melhor diferenciar as interações verdadeiras das falsas. Assim, nós identificamos, para cada pDB, os valores retornados do alinhamento feito pelo BLASTp que melhor contribuem para as predições.

A combinação dos valores de identidade e cobertura extraídos dos alinhamentos compuseram a melhor métrica, correspondendo a um AUC de 0,96 para pDB individual e um AUC de 0,93 para a combinação de pDB. O ponto de corte de 0,70 para a métrica identidade vezes cobertura, corresponde à especificidade de 0,95 e sensibilidade de 0,90, demostrando que nosso método prediz eficientemente as interações proteína-proteína.

Adicionalmente, em vez de usarmos somente o primeiro alinhamento do BLASTp, nós usamos os 20 primeiros alinhamentos, aumentando a quantidade de pares de interação preditos e a cobertura na rede de interação. Consequentemente, aumentamos também exponencialmente a quantidade de alinhamentos e pares de interação para serem manipulados e tratados. Ao usar mais que um alinhamento do BLASTp, gera-se redundância de pares de interação preditos entre as pDB e entre as proteínas homólogas contidas dentre os 20 alinhamentos do BLASTp. Sob o ponto de vista tecnológico esta quantidade de dados não útil pode gerar problemas, exigindo computadores mais potentes ou algoritmo mais eficiente para o processamento.

No segundo trabalho, aplicamos esta metodologia com as métricas validadas para gerar as redes de interação para nove linhagens do biovar *ovis* de *C. pseudotuverculosis* (Cp). Assim, seguindo a metodologia, executamos o alinhamento recíproco entre as nove linhagens de Cp contra as pDBs, identificamos os pares de interação e usamos os valores de identidade vezes cobertura extraídos dos alinhamentos do BLASTP para calcular a métrica e gerar as redes de interações.

Como resultado, foram preditos aproximadamente 16.000 pares de interação para cada linhagem de Cp, sendo ~99% mapeado do gênero *Corynebacterium*, ou seja, de um organismo filogeneticamente próximo, aumentando biologicamente a probabilidade que as interações preditas realmente ocorram em Cp. Destes pares de interação preditos, 15.495 são conservados entre as nove linhages do biovar *ovis* de Cp. Este conjunto de interações conservadas foi usado para fazer análise dos clusteres e identificação de proteínas essenciais.

Antes, porém, nos preocupamos em validar as redes de interação preditas e verificar se possuíam características de redes biológicas. Submetemos então as redes de interação preditas para validação quanto a menor caminho (*Shortest Path*) e verificar se o grau de interação seguia uma distribuição livre de escala (*Scale Free*) com aproximação à lei de poderes (*Power Law*). Ambas análises topológicas sugerem que todas as redes de interação preditas possuem característica pertencentes às redes biológicas.

Adicionalmente, foi verificado se as redes de interação preditas tinham alguma chance de serem geradas aleatoriamente. Assim, submetemos as redes de interação geradas ao teste de distribuição normal denominado Shapiro-Wilk teste, qual descartou definitivamente a probabilidade que as redes de interação tivessem uma distribuição normal, obtendo um p-value < 2.2e-16 (Shapiro e Wilk, 1965). Ainda, comparamos as redes de interação preditas contra redes de interação geradas aleatoriamente. Nesta comparação, os valores do Coeficiente de Clusterização, Correlção e R^2 obtidos são extremamente diferentes entre os dois tipos de redes, sugerindo que as redes preditas não foram formadas por interações expúrias ou aleatórias, possuindo um viés biológico, possivelmente devido à pressão evolucionária exercida sobre estas interações no organismo. Em tempo, o alto valor do Coeficiente de Clusterização sugere uma auto organização nas célula de Cp motivada pelas interações (Galeota *et al.*, 2015).

Seguros de estarmos analisando redes de interação biológicas, procedemos com a análise dos clusteres de proteínas e das proteínas essenciais. Dentre os clustes encontrados, selecionamos cinco com maior quantidade de proteínas para serem analisados com suporte da literatura, sendo eles principalmente formados por proteínas Ribossomais e de RNA Polimerase, Sistema de transporte de Oligopeptídeos, Biosintese de Cobalamina, Aquisição de Ferro e regulação intracelular e, Divisão celular e biossíntese da parede celular.

Ao analisar os clusters, o viés biológico exercido sobre estes e as interações, é identificado e apoiado pela descrição na literatura e caracterização por métodos experimentais, mesmo que em outros organismos filogeneticamente próximos. Este conhecimento a nível de biologia de sistemas, obtidos na literatura, pode então ser transferido, via rede de interação, para Cp, possibilitanto melhor entendimento do organismo. Da mesma forma, a falta de informação na literatura sobre algumas interações, faz das redes de interação proteína-proteína uma importante ferramenta para melhor analisar e entender o comportamento celular de Cp, permitindo levantar novas hipóteses e direcionar novos experimentos em laboratório, visando testar a drogabilidade e essencialidade destas proteínas e interações.

Entre as 15.495 interações conservadas nas nove redes de interação preditas para Cp, considerando principalmente o grau de interação, 181 proteínas essenciais foram identificadas (Khuri e Wuchty, 2015); participando principalmente no metabolismo de carbono, envelope celular e síntese da parede celular, biossíntese de nucleotídeos, enovelamente, translocação, formação do ribossomo, fatores de transcrição, síntese de tRNA, metabolismo de RNA e, via metabólica respiratória. Dentre estas proteínas essencias, somente a *DNA repair* (RecN) não foi identificada como essencial na base de dados DEG.

Enquanto a maioria das proteínas essenciais possuem mais proteínas em mais de 20 organismos de DEG, outras três proteínas essenciais em Cp tiveram homologia com apenas um organismo de DEG: *Catalase* (KatA), *Endonuclease III* (Nth) and *Trigger factor Tig* (Tig). Isto pode ser explicado pelo fato de que a essencialidade nem sempre é conservada entre as espécies (Caufield *et al.*, 2015). Dentre as proteínas essenciais 41 não tiveram homologia contra seus hospedeiros, sendo boas candidatas para uso em diagnóstico ou alvos para drogas.

Além da identificação de clusteres e proteínas essenciais, as redes de interação podem ser usadas em conjunto com outras técnicas experimentais para auxiliar na interpretação dos resultados. Assim, em posse da rede de interação proteína-proteína gerada para a linhagem 1002 de *C. pseudotuberculosis*, foram identificadas as interações entre as proteínas com baixa e alta expressão, bem como as proteínas exclusivamente expressas, quando submetidas a stresse nitrosativo. A visão sistêmica das proteínas envolvidas na condição de estresse, propiciada pela rede de interação, auxiliou na interpretação dos resultados do experimento de proteômica comparativa.

Ao analisar as redes de interação com mais atenção aos detalhes e considerando os resultados obtidos durante o desenvolvimento desta tese, é perceptível que muitos outros trabalhos derivados ou somados às redes de interação poderão ser desenvolvidos, sejam eles de natureza experimental ou computacional.

5 - Conclusão e Perspectivas

Neste trabalho, analisamos e validamos um conjunto de métricas capaz de mapear com eficiência interações ortólogas de bases de dados públicas, aumentando inclusive a cobertura em uma rede de interação. Pela primeira vez usamos esta metodologia validada para mapear as interação proteína-proteína para nove linhagens do biovar *ovis* de *C. pseudotuberculosis*. Adicionalmente, geramos a rede de interação dos genes diferencialmente e exclusivamente expressos para auxiliar na interpretação dos resultados gerados por experimento de proteomica comparativa.

Mais importante que a validação estatística aplicada sobre as redes preditas, evidenciando que possuem características de redes biológicas, são as evidências biológicas encontradas, apoiadas pela literatura, na análise dos clusteres e proteínas essenciais. Assim, o método para predição de redes de interação proteína-proteína se mostra uma importante ferramenta para biólogos estudarem e entenderem os organismos de interesse a nível de biologia de sistemas, bem como, uma valiosa ferramenta para a predição de proteínas essenciais, com potencial uso em diagnóstico ou como alvos para drogas.

Neste trabalho, além das 181 proteínas essenciais preditas, existem aproximadamente 15.000 interações conservadas entre as nove linhagens para ser exploradas experimentalmente e gerar trabalhos futuros.

Dentre algumas perpectivas de trabalhos, experimentais ou computacionais, podemos citar:

- Estudar os clusteres e interações identificadas que tenham relevância biológica visando entender melhor *C. pseudotuberculosis* e sua patogenicidade, direcionando novos trabalhos em laboratório (Marsh *et al.*, 2013);

- Testar experimentalmente as proteínas essenciais identificadas;

- Re-anotar as proteínas hipotéticas baseado na função de seus parceiros de interação encontrados na rede (Peng *et al.*, 2014; Hao *et al.*, 2015);

- Desenvolver uma base de dados pública e disponibilizar as interações preditas para *C. pseudotuberculosis*, bem como uma forma eficiente e amigável para sua visualização;

- Aplicar a metodologia desenvolvida na predição das redes de interação proteína-proteína de outros organismos de interesse biotecnológico;

- Considerando as montagens geradas para os novos genomas sequenciados do biovar *equi*, fazer a predição de interação proteína-proteína e comparar as diferenças e semelhanças com o biovar *ovis*.

- Cruzar as redes de interação com os dados gerados por experimentos de RNA-Seq, SNPs, proteomica ou outros experimentos biológicos, visando extrair informação e entender como estas proteínas interagem e cooperam nas condições testadas.

Neste sentido, em colaboração com o Dr. Wanderson Marques Silva, esta em adamento um trabalho para caracterizar o proteoma total das linhagens 1002 (biovar *ovis*) e 258 (biovar *equi*) e explorar as diferenças entre os dois biovares que possam fornecer dados a respeito da biologia deste patógeno. Experimentos com proteômica já foram feitos e foram caracterizadas aproximadamente 1.321 proteínas de *C. pseudotuberculosis.* Estas proteínas serão analisadas nas redes de interação considerando o nível de expressão e se pertencem ao interactoma central ou específico de cada biovar.

Bibliografia

ABEBE, D.; SISAY TESSEMA, T. Determination of Corynebacterium pseudotuberculosis prevalence and antimicrobial susceptibility pattern of isolates from lymph nodes of sheep and goat at organic export abattoir, Modjo, Ethiopia. **Letters in Applied Microbiology**, 2015. ISSN 1472-765X.

ADÉKAMBI, T.; DRANCOURT, M.; RAOULT, D. The rpoB gene as a tool for clinical microbiologists. **Trends in microbiology,** v. 17, n. 1, p. 37-45, 2009. ISSN 0966-842X.

ALLEN, C. E.; SCHMITT, M. P. Novel hemin binding domains in the Corynebacterium diphtheriae HtaA protein interact with hemoglobin and are critical for heme iron utilization by HtaA. **Journal of bacteriology,** v. 193, n. 19, p. 5374-5385, 2011. ISSN 0021-9193.

ANH, N. H. et al. Discovery of pathways in protein-protein interaction networks using a genetic algorithm. **Data & Knowledge Engineering**, 2015. ISSN 0169-023X.

ASSENOV, Y. et al. Computing topological parameters of biological networks. **Bioinformatics,** v. 24, n. 2, p. 282-284, 2008. ISSN 1367-4803.

BAIRD, G. J.; FONTAINE, M. C. Corynebacterium pseudotuberculosis and its Role in Ovine Caseous Lymphadenitis. **Journal of comparative pathology,** v. 137, n. 4, p. 179-210, 2007. ISSN 0021-9975.

BARABÁSI, A. L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. **Nature Reviews Genetics,** v. 5, n. 2, p. 101-113, 2004. ISSN 1471-0056.

BARH, D. et al. Conserved host–pathogen PPIs Globally conserved inter-species bacterial PPIs based conserved host-pathogen interactome derived novel target in C. pseudotuberculosis, C. diphtheriae, M. tuberculosis, C. ulcerans, Y. pestis, and E. coli targeted by Piper betel compounds. **Integrative Biology,** v. 5, n. 3, p. 495-509, 2013.

BETUL, K.; ERIC, A. Experimental evolution of protein-protein interaction networks. **Biochemical Journal,** v. 453, n. 3, p. 311-319, 2013. ISSN 1470-8728.

BRAIBANT, M.; GILOT, P. The ATP binding cassette (ABC) transport systems of Mycobacterium tuberculosis. **FEMS microbiology reviews,** v. 24, n. 4, p. 449-467, 2000. ISSN 1574-6976.

BRAUN, P.; GINGRAS, A. C. History of protein–protein interactions: From egg-white to complex networks. **Proteomics,** v. 12, n. 10, p. 1478-1498, 2012. ISSN 1615-9861.

BROWN, S. D. et al. Molecular dynamics of the Shewanella oneidensis response to chromate stress. **Molecular & Cellular Proteomics,** v. 5, n. 6, p. 1054-1071, 2006. ISSN 1535-9476.

BUSS, J. et al. A Multi-layered Protein Network Stabilizes the Escherichia coli FtsZ-ring and Modulates Constriction Dynamics. 2015. ISSN 1553-7404.

BUTLER, W.; AHEARN, D.; KILBURN, J. High-performance liquid chromatography of mycolic acids as a tool in the identification of Corynebacterium, Nocardia, Rhodococcus, and Mycobacterium species. **Journal of clinical microbiology,** v. 23, n. 1, p. 182-185, 1986. ISSN 0095-1137.

CAMACHO, C. et al. BLAST+: architecture and applications. **BMC bioinformatics,** v. 10, n. 1, p. 421, 2009. ISSN 1471-2105.

CARBALLIDO-LÓPEZ, R.; ERRINGTON, J. A dynamic bacterial cytoskeleton. **Trends in cell biology,** v. 13, n. 11, p. 577-583, 2003. ISSN 0962-8924.

CASTRO-ROA, D.; ZENKIN, N. In vitro experimental system for analysis of transcription–translation coupling. **Nucleic acids research,** v. 40, n. 6, p. e45-e45, 2012. ISSN 0305-1048.

CAUFIELD, J. H. et al. Protein Complexes in Bacteria. **PLOS Computational Biology,** v. 11, n. 2, 2015. ISSN 1553-734X.

CERDEIRA, L. T. et al. Whole-genome sequence of Corynebacterium pseudotuberculosis PAT10 strain isolated from sheep in Patagonia, Argentina. **Journal of bacteriology,** v. 193, n. 22, p. 6420-6421, 2011. ISSN 0021-9193.

COENYE, T.; VANDAMME, P. Organisation of the S10, spc and alpha ribosomal protein gene clusters in prokaryotic genomes. **FEMS microbiology letters,** v. 242, n. 1, p. 117-126, 2005. ISSN 0378-1097.

COLOM-CADENA, A. et al. Management of a caseous lymphadenitis outbreak in a new Iberian ibex (Capra pyrenaica) stock reservoir. **Acta Veterinaria Scandinavica,** v. 56, n. 1, p. 83, 2014. ISSN 1751-0147.

CONTRERAS, H. et al. Heme uptake in bacterial pathogens. **Current opinion in chemical biology,** v. 19, p. 34-41, 2014. ISSN 1367-5931.

CORRENTI, C.; STRONG, R. K. Mammalian siderophores, siderophore-binding lipocalins, and the labile iron pool. **Journal of Biological Chemistry,** v. 287, n. 17, p. 13524-13531, 2012. ISSN 0021-9258.

CROFT, M. T. et al. Algae acquire vitamin B12 through a symbiotic relationship with bacteria. **Nature,** v. 438, n. 7064, p. 90-93, 2005. ISSN 0028-0836.

CUI, T.; HE, Z.-G. Improved understanding of pathogenesis from protein interactions in Mycobacterium tuberculosis. **Expert review of proteomics**, n. 0, p. 1-11, 2014. ISSN 1478-9450.

CUTLER, R. G. Oxidative stress and aging: catalase is a longevity determinant enzyme. **Rejuvenation research,** v. 8, n. 3, p. 138-140, 2005. ISSN 1549-1684.

DAI, Q.-G. et al. CPL: Detecting Protein Complexes by Propagating Labels on Protein-Protein Interaction Network. **Journal of Computer Science and Technology,** v. 29, n. 6, p. 1083-1093, 2014. ISSN 1000-9000.

DALL, H. P. et al. Omics profiles used to evaluate the gene expression of Exiguobacterium antarcticum B7 during cold adaptation. **BMC genomics,** v. 15, n. 1, p. 986, 2014. ISSN 1471-2164.

DE LAS RIVAS, J.; FONTANILLO, C. Protein–protein interaction networks: unraveling the wiring of molecular machines within the cell. **Briefings in Functional Genomics**, 2012. ISSN 2041-2649.

DEN BLAAUWEN, T.; ANDREU, J. M.; MONASTERIO, O. Bacterial cell division proteins as antibiotic targets. **Bioorganic chemistry,** v. 55, p. 27-38, 2014. ISSN 0045-2068.

DEUERLING, E. et al. Trigger factor and DnaK cooperate in folding of newly synthesized proteins. **Nature,** v. 400, n. 6745, p. 693-696, 1999. ISSN 0028-0836.

DORELLA, F. A. et al. Corynebacterium pseudotuberculosis: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. **Veterinary research,** v. 37, n. 2, p. 201-218, 2006. ISSN 0928-4249.

EISEN, J. A.; HANAWALT, P. C. A phylogenomic study of DNA repair genes, proteins, and processes. **Mutation Research/DNA Repair,** v. 435, n. 3, p. 171-213, 1999. ISSN 0921-8777.

EL ZOEIBY, A.; SANSCHAGRIN, F.; LEVESQUE, R. C. Structure and function of the Mur enzymes: development of novel inhibitors. **Molecular microbiology,** v. 47, n. 1, p. 1-12, 2003. ISSN 1365-2958.

ERRINGTON, J.; DANIEL, R. A.; SCHEFFERS, D.-J. Cytokinesis in bacteria. **Microbiology and Molecular Biology Reviews,** v. 67, n. 1, p. 52-65, 2003. ISSN 1092-2172.

ESTRADA, E. Virtual identification of essential proteins within the protein interaction network of yeast. **Proteomics,** v. 6, n. 1, p. 35-40, 2006. ISSN 1615-9861.

FLÓREZ, A. et al. Protein network prediction and topological analysis in Leishmania major as a tool for drug target selection. **BMC bioinformatics,** v. 11, n. 1, p. 484, 2010. ISSN 1471-2105.

FOLADOR, E. L. et al. An improved interolog mapping-based computational prediction of protein-protein interactions with increased network coverage. **Integrative Biology**, 2014.

FRANCESCHINI, A. et al. STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. **Nucleic acids research,** v. 41, n. D1, p. D808-D815, 2013. ISSN 0305-1048.

FRANKENBERG, N.; MOSER, J.; JAHN, D. Bacterial heme biosynthesis and its biotechnological application. **Applied microbiology and biotechnology,** v. 63, n. 2, p. 115-127, 2003. ISSN 0175-7598.

GALEOTA, E. et al. The hierarchical organization of natural protein interaction networks confers self-organization properties on pseudocells. **BMC Systems Biology,** v. 9, n. Suppl 3, p. S3, 2015. ISSN 1752-0509.

GARMA, L. et al. How Many Protein-Protein Interactions Types Exist in Nature? **PloS one,** v. 7, n. 6, p. e38913, 2012. ISSN 1932-6203.

GONZALEZ, M. W.; KANN, M. G. Protein interactions and disease. **PLoS computational biology,** v. 8, n. 12, p. e1002819, 2012. ISSN 1553-7358.

GOWTHAMAN, R.; LYSKOV, S.; KARANICOLAS, J. DARC 2.0: Improved Docking and Virtual Screening at Protein Interaction Sites. **PloS one,** v. 10, n. 7, p. e0131612, 2015. ISSN 1932-6203.

GÓRSKA, A.; SLODERBACH, A.; MARSZAŁŁ, M. P. Siderophore–drug complexes: potential medicinal applications of the ‘Trojan horse’strategy. **Trends in pharmacological sciences,** v. 35, n. 9, p. 442-449, 2014. ISSN 0165-6147.

HADDADIN, F. A. T.; HARCUM, S. W. Transcriptome profiles for high-cell-density recombinant and wild-type Escherichia coli. **Biotechnology and bioengineering,** v. 90, n. 2, p. 127-153, 2005. ISSN 1097-0290.

HAN, J.-D. J. et al. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. **Nature,** v. 430, n. 6995, p. 88-93, 2004. ISSN 0028-0836.

HAO, T. et al. Function Annotation of Proteins in Eriocheir sinensis Based on the Protein-Protein Interaction Network. The Proceedings of the Third International Conference on Communications, Signal Processing, and Systems, 2015, Springer. p.831-837.

HARIHARAN, H. et al. Serological detection of caseous lymphadenitis in sheep and goats using a commercial ELISA in Grenada, West Indies. 2014.

HASSAN, S. S. et al. Complete genome sequence of Corynebacterium pseudotuberculosis biovar ovis strain P54B96 isolated from antelope in South Africa obtained by Rapid Next Generation Sequencing Technology. **Standards in genomic sciences,** v. 7, n. 2, p. 189, 2012.

_____. Proteome scale comparative modeling for conserved drug and vaccine targets identification in Corynebacterium pseudotuberculosis. **BMC genomics,** v. 15, n. Suppl 7, p. S3, 2014. ISSN 1471-2164.

HELDT, D. et al. Aerobic synthesis of vitamin B12: ring contraction and cobalt chelation. **Biochemical Society Transactions,** v. 33, n. 4, p. 815-819, 2005. ISSN 0300-5127.

HERMJAKOB, H. et al. IntAct: an open source molecular interaction database. **Nucleic acids research,** v. 32, n. suppl 1, p. D452-D455, 2004. ISSN 0305-1048.

HIRON, A. et al. Only one of four oligopeptide transport systems mediates nitrogen nutrition in Staphylococcus aureus. **Journal of bacteriology,** v. 189, n. 14, p. 5119-5129, 2007. ISSN 0021-9193.

HÄUSER, R. et al. A Second-generation Protein–Protein Interaction Network of Helicobacter pylori. **Molecular & Cellular Proteomics,** v. 13, n. 5, p. 1318-1329, 2014. ISSN 1535-9476.

HÉMOND, V. et al. Lymphadénite axillaire à< i> Corynebacterium pseudotuberculosis</i> chez une patiente de 63 ans. **Médecine et maladies infectieuses,** v. 39, n. 2, p. 136-139, 2009. ISSN 0399-077X.

IKEDA, M. Towards bacterial strains overproducing L-tryptophan and other aromatics by metabolic engineering. **Applied microbiology and biotechnology,** v. 69, n. 6, p. 615-626, 2006. ISSN 0175-7598.

IVANOVIĆ, S. et al. Caseous lymphadenitis in goats. **Biotechnology in Animal Husbandry,** v. 25, n. 5-6-2, p. 999-1007, 2009. ISSN 1450-9156.

JEONG, H. et al. Lethality and centrality in protein networks. **arXiv preprint cond-mat/0105306**, 2001.

JONES, M. M. et al. Role of the Oligopeptide Permease ABC Transporter of Moraxella catarrhalis in Nutrient Acquisition and Persistence in the Respiratory Tract. **Infection and immunity,** v. 82, n. 11, p. 4758-4766, 2014. ISSN 0019-9567.

JUNG, B. Y. et al. Serology and clinical relevance of Corynebacterium pseudotuberculosis in native Korean goats (Capra hircus coreanae). **Tropical Animal Health and Production**, p. 1-5, ISSN 0049-4747.

_____. Serology and clinical relevance of Corynebacterium pseudotuberculosis in native Korean goats (Capra hircus coreanae). **Tropical animal health and production,** v. 47, n. 4, p. 657-661, 2015. ISSN 0049-4747.

KHURI, S.; WUCHTY, S. Essentiality and centrality in protein interaction networks revisited. **BMC Bioinformatics,** v. 16, n. 1, p. 109, 2015. ISSN 1471-2105.

KOHL, M.; WIESE, S.; WARSCHEID, B. Cytoscape: software for visualization and analysis of biological networks. In: (Ed.). **Data Mining in Proteomics**: Springer, 2011. p.291-303. ISBN 1607619865.

KUNKLE, C. A.; SCHMITT, M. P. Analysis of a DtxR-regulated iron transport and siderophore biosynthesis gene cluster in Corynebacterium diphtheriae. **Journal of bacteriology,** v. 187, n. 2, p. 422-433, 2005. ISSN 0021-9193.

KÖSTER, W. ABC transporter-mediated uptake of iron, siderophores, heme and vitamin B 12. **Research in microbiology,** v. 152, n. 3, p. 291-301, 2001. ISSN 0923-2508.

LAGE, K. Protein-protein interactions and genetic diseases: The Interactome. **Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease**, 2014. ISSN 0925-4439.

LI, H. et al. A Computational Method to Identify Druggable Binding Sites That Target Protein-Protein Interactions. 2014.

LI, M. et al. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. **BMC systems biology,** v. 6, n. 1, p. 15, 2012. ISSN 1752-0509.

LIU, Z.-P. et al. Inferring a protein interaction map of Mycobacterium tuberculosis based on sequences and interologs. **BMC bioinformatics,** v. 13, n. Suppl 7, p. S6, 2012. ISSN 1471-2105.

LO, Y. et al. Reconstructing genome-wide protein-protein interaction networks using multiple strategies with homologous mapping. **PloS one,** v. 10, n. 1, p. e0116347, 2015. ISSN 1932-6203.

LOPES, T. et al. Complete Genome Sequence of Corynebacterium pseudotuberculosis Strain Cp267, Isolated from a Llama. **Journal of bacteriology,** v. 194, n. 13, p. 3567-3568, 2012. ISSN 0021-9193.

LUO, H. et al. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. **Nucleic acids research,** v. 42, n. D1, p. D574-D580, 2014. ISSN 0305-1048.

LUTKENHAUS AND, J.; ADDINALL, S. Bacterial cell division and the Z ring. **Annual review of biochemistry,** v. 66, n. 1, p. 93-116, 1997. ISSN 0066-4154.

MARSH, J. A. et al. Protein complexes are under evolutionary selection to assemble via ordered pathways. **Cell,** v. 153, n. 2, p. 461-470, 2013. ISSN 0092-8674.

MARTÍN, J. F. et al. Ribosomal RNA and ribosomal proteins in corynebacteria. **J. Biotechnol,** v. 104, p. 41-53, 2003.

MCGARY, K.; NUDLER, E. RNA polymerase and the ribosome: the close relationship. **Current opinion in microbiology,** v. 16, n. 2, p. 112-117, 2013. ISSN 1369-5274.

MILSE, J. et al. Transcriptional response of Corynebacterium glutamicum ATCC 13032 to hydrogen peroxide stress and characterization of the OxyR regulon. **Journal of biotechnology,** v. 190, p. 40-54, 2014. ISSN 0168-1656.

MIRA, C. et al. Epidemiological and Histopathological Studies on Caseous Lymphadenitis in Slaughtered Goats in Algeria. **lung,** v. 6, p. 26.5, 2014.

MITRA, A. Biology, Genetic Aspects, and Oxidative Stress Response of Streptomyces and Strategies for Bioremediation of Toxic Metals. **Microbial Biodegradation and Bioremediation**, p. 287, 2014. ISSN 0128004827.

MONNET, V. Bacterial oligopeptide-binding proteins. **Cellular and Molecular Life Sciences CMLS,** v. 60, n. 10, p. 2100-2114, 2003. ISSN 1420-682X.

MOORE, S.; WARREN, M. The anaerobic biosynthesis of vitamin B12. **Biochemical Society Transactions,** v. 40, n. 3, p. 581, 2012. ISSN 0300-5127.

MORA, A.; DONALDSON, I. M. Effects of protein interaction data integration, representation and reliability on the use of network properties for drug target prediction. **BMC bioinformatics,** v. 13, n. 1, p. 294, 2012. ISSN 1471-2105.

MORAES, P. M. et al. Characterization of the Opp Peptide Transporter of Corynebacterium pseudotuberculosis and Its Role in Virulence and Pathogenicity. **BioMed research international,** v. 2014, 2014. ISSN 2314-6133.

MORRIS, J. H. et al. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. **BMC bioinformatics,** v. 12, n. 1, p. 436, 2011. ISSN 1471-2105.

MOSCA, R. et al. Towards a detailed atlas of protein–protein interactions. **Current opinion in structural biology,** v. 23, n. 6, p. 929-940, 2013. ISSN 0959-440X.

MULDER, N. J. et al. Using biological networks to improve our understanding of infectious diseases. **Computational and Structural Biotechnology Journal**, 2014. ISSN 2001-0370.

NAIDER, F.; BECKER, J. M. Multiplicity of oligopeptide transport systems in Escherichia coli. **Journal of bacteriology,** v. 122, n. 3, p. 1208-1215, 1975. ISSN 0021-9193.

NELSON, D.; COX, M. Lehninger, Princípios de Bioquímica. **Sarvier,** v. 3ª edição, São Paulo, p. 202, 2002.

ORCHARD, S. et al. Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. **Nature methods,** v. 9, n. 4, p. 345-350, 2012. ISSN 1548-7091.

OREIBY, A. et al. Caseous lymphadenitis in small ruminants in Egypt. **Tierärztliche Praxis Großtiere,** v. 42, n. 5, p. 271-277, 2014. ISSN 1434-1220.

OSMAN, A. Y. et al. Caseous Lymphadenitis in a Goat: A Case Report. **International Journal of Livestock Research,** v. 5, n. 3, p. 128-132, 2015.

PARK, H.-S. et al. Transcriptomic analysis of Corynebacterium glutamicum in the response to the toxicity of furfural present in lignocellulosic hydrolysates. **Process Biochemistry**, 2014. ISSN 1359-5113.

PELAY-GIMENO, M. et al. Structure-Based Design of Inhibitors of Protein–Protein Interactions: Mimicking Peptide Binding Epitopes. **Angewandte Chemie International Edition**, 2015. ISSN 1521-3773.

PENG, W. et al. Improving protein function prediction using domain and protein complexes in PPI networks. **BMC systems biology,** v. 8, n. 1, p. 35, 2014. ISSN 1752-0509.

PETHICK, F. E. et al. Complete Genome Sequences of Corynebacterium pseudotuberculosis Strains 3/99-5 and 42/02-A, Isolated from Sheep in Scotland and Australia, Respectively. **Journal of Bacteriology,** v. 194, n. 17, p. 4736-4737, 2012. ISSN 0021-9193.

PINTO, A. C. et al. Differential transcriptional profile of Corynebacterium pseudotuberculosis in response to abiotic stresses. **BMC genomics,** v. 15, n. 1, p. 14, 2014. ISSN 1471-2164.

RESENDE, B. et al. DNA repair in Corynebacterium model. **Gene,** v. 482, n. 1, p. 1-7, 2011. ISSN 0378-1119.

REZENDE, A. M. et al. Computational Prediction of Protein-Protein Interactions in Leishmania Predicted Proteomes. **PloS one,** v. 7, n. 12, p. e51304, 2012. ISSN 1932-6203.

RODIONOV, D. A. et al. Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. **Journal of Biological Chemistry,** v. 278, n. 42, p. 41148-41159, 2003. ISSN 0021-9258.

RONDON, M. R.; TRZEBIATOWSKI, J. R.; ESCALANTE-SEMERENA, J. C. Biochemistry and molecular genetics of cobalamin biosynthesis. **Progress in nucleic acid research and molecular biology,** v. 56, p. 347-384, 1996. ISSN 0079-6603.

ROSTAS, K. et al. Nucleotide sequence and LexA regulation of the Escherichia coli recN gene. **Nucleic acids research,** v. 15, n. 13, p. 5041-5049, 1987. ISSN 0305-1048.

ROTH, J.; LAWRENCE, J.; BOBIK, T. Cobalamin (coenzyme B12): synthesis and biological significance. **Annual Reviews in Microbiology,** v. 50, n. 1, p. 137-181, 1996. ISSN 0066-4227.

ROYSTON, J. An extension of Shapiro and Wilk's W test for normality to large samples. **Applied Statistics**, p. 115-124, 1982. ISSN 0035-9254.

RUIZ, J. C. et al. Evidence for reductive genome evolution and lateral acquisition of virulence functions in two Corynebacterium pseudotuberculosis strains. **PLoS One,** v. 6, n. 4, p. e18551, 2011. ISSN 1932-6203.

SAHBANI, S. K. et al. The relative contributions of DNA strand breaks, base damage and clustered lesions to the loss of DNA functionality induced by ionizing radiation. **Radiation research,** v. 181, n. 1, p. 99-110, 2014. ISSN 0033-7587.

SAITO, Y. et al. Characterization of endonuclease III (nth) and endonuclease VIII (nei) mutants of Escherichia coli K-12. **Journal of bacteriology,** v. 179, n. 11, p. 3783-3785, 1997. ISSN 0021-9193.

SAKMANOĞLU, A. et al. Identification and antimicrobial susceptibility of Corynebacterium pseudotuberculosis isolated from sheep. **Eurasian Journal of Veterinary Sciences,** v. 31, n. 2, p. 116-121, 2015. ISSN 1309-6958.

SANTAROSA, B. P. et al. MENINGOENCEFALITE SUPURATIVA POR Corynebacterium pseudotuberculosis EM CABRA COM LINFADENITE CASEOSA: RELATO DE CASO. **Veterinária e Zootecnia,** v. 21, n. 4, p. 537-542, 2015. ISSN 2178-3764.

SCHALK, I. J. Innovation and Originality in the Strategies Developed by Bacteria To Get Access to Iron. **Chembiochem,** v. 14, n. 3, p. 293-294, 2013. ISSN 1439-7633.

SCOTT, A.; ROESSNER, C. Biosynthesis of cobalamin (vitamin B (12)). **Biochemical Society Transactions,** v. 30, n. 4, p. 613-620, 2002. ISSN 0300-5127.

SELIM, S. Oedematous skin disease of buffalo in Egypt. **Journal of Veterinary Medicine, Series B,** v. 48, n. 4, p. 241-258, 2001. ISSN 1439-0450.

SERAFINI, D. M.; SCHELLHORN, H. E. Endonuclease III and endonuclease IV protect Escherichia coli from the lethal and mutagenic effects of near-UV irradiation. **Canadian journal of microbiology,** v. 45, n. 7, p. 632-637, 1999. ISSN 0008-4166.

SEYFFERT, N. et al. High seroprevalence of caseous lymphadenitis in Brazilian goat herds revealed by< i> Corynebacterium pseudotuberculosis</i> secreted proteins-based ELISA. **Research in veterinary science,** v. 88, n. 1, p. 50-55, 2010. ISSN 0034-5288.

SHANNON, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. **Genome research,** v. 13, n. 11, p. 2498-2504, 2003. ISSN 1088-9051.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, p. 591-611, 1965. ISSN 0006-3444.

SHARAN, R. et al. Conserved patterns of protein interaction in multiple species. **Proceedings of the National Academy of Sciences of the United States of America,** v. 102, n. 6, p. 1974-1979, 2005. ISSN 0027-8424.

SHELDON, J. R.; HEINRICHS, D. E. Recent developments in understanding the iron acquisition strategies of gram positive pathogens. **FEMS microbiology reviews**, p. fuv009, 2015. ISSN 1574-6976.

SHENG, C. et al. State-of-the-art strategies for targeting protein–protein interactions by small-molecule inhibitors. **Chemical Society Reviews**, 2015.

SILVA, A. et al. Complete genome sequence of Corynebacterium pseudotuberculosis I19, a strain isolated from a cow in Israel with bovine mastitis. **Journal of bacteriology,** v. 193, n. 1, p. 323-324, 2011. ISSN 0021-9193.

SILVA, W. M. et al. Label-free proteomic analysis to confirm the predicted proteome of Corynebacterium pseudotuberculosis under nitrosative stress mediated by nitric oxide. **BMC genomics,** v. 15, n. 1, p. 1065, 2014. ISSN 1471-2164.

SMID, E. J.; PLAPP, R.; KONINGS, W. Peptide uptake is essential for growth of Lactococcus lactis on the milk protein casein. **Journal of bacteriology,** v. 171, n. 11, p. 6135-6140, 1989. ISSN 0021-9193.

SMITH, J. L. The physiological role of ferritin-like compounds in bacteria. **Critical reviews in microbiology,** v. 30, n. 3, p. 173-185, 2004. ISSN 1040-841X.

SOARES, S. C. et al. The pan-genome of the animal pathogen Corynebacterium pseudotuberculosis reveals differences in genome plasticity between the biovar ovis and equi strains. **PloS one,** v. 8, n. 1, p. e53818, 2013. ISSN 1932-6203.

SONGER, J. G. et al. Biochemical and genetic characterization of Corynebacterium pseudotuberculosis. **American journal of veterinary research,** v. 49, n. 2, p. 223-226, 1988. ISSN 0002-9645.

STELZL, U. et al. Ribosomal proteins: role in ribosomal functions. **eLS**, 2001. ISSN 047001590X.

SUKHODOLETS, M. V.; GARGES, S. Interaction of Escherichia coli RNA polymerase with the ribosomal protein S1 and the Sm-like ATPase Hfq. **Biochemistry,** v. 42, n. 26, p. 8022-8034, 2003. ISSN 0006-2960.

TANG, Y. et al. CytoNCA: A cytoscape plugin for centrality analysis and evaluation of protein interaction networks. **Biosystems**, 2014. ISSN 0303-2647.

TAYLOR, I. W.; WRANA, J. L. Protein interaction networks in medicine and disease. **Proteomics,** v. 12, n. 10, p. 1706-1716, 2012. ISSN 1615-9861.

TEIXEIRA, D. et al. The tufB–secE–nusG–rplKAJL–rpoB gene cluster of the liberibacters: sequence comparisons, phylogeny and speciation. **International Journal of Systematic and Evolutionary Microbiology,** v. 58, n. 6, p. 1414-1421, 2008. ISSN 1466-5026.

TROST, E. et al. The complete genome sequence of Corynebacterium pseudotuberculosis FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. **BMC genomics,** v. 11, n. 1, p. 728, 2010. ISSN 1471-2164.

VAN DONGEN, S. A cluster algorithm for graphs. **Report-Information systems**, n. 10, p. 1-40, 2000. ISSN 1386-3681.

VILLOUTREIX, B. O. et al. Drug-Like Protein□ Protein Interaction Modulators: Challenges and Opportunities for Drug Discovery and Chemical Biology. **Molecular informatics,** v. 33, n. 6-7, p. 414-437, 2014. ISSN 1868-1751.

VOIGT, K. et al. Eradication of caseous lymphadenitis under extensive management conditions on a Scottish hill farm. **Small Ruminant Research**, 2012. ISSN 0921-4488.

VOLLMER, W.; BLANOT, D.; DE PEDRO, M. A. Peptidoglycan structure and architecture. **FEMS microbiology reviews,** v. 32, n. 2, p. 149-167, 2008. ISSN 1574-6976.

WALTER, B. M. et al. The LexA regulated genes of the Clostridium difficile. **BMC microbiology,** v. 14, n. 1, p. 88, 2014. ISSN 1471-2180.

WANDERSMAN, C.; DELEPELAIRE, P. Bacterial iron sources: from siderophores to hemophores. **Annu. Rev. Microbiol.,** v. 58, p. 611-647, 2004. ISSN 0066-4227.

WANG, J. et al. Recent advances in clustering methods for protein interaction networks. **BMC genomics,** v. 11, n. Suppl 3, p. S10, 2010. ISSN 1471-2164.

WETIE, A. G. N. et al. Protein–protein interactions: switch from classical methods to proteomics and bioinformatics-based approaches. **Cellular and Molecular Life Sciences,** v. 71, n. 2, p. 205-228, 2014. ISSN 1420-682X.

WETIE, N. et al. Investigation of stable and transient protein–protein interactions: Past, present, and future. **Proteomics**, 2013. ISSN 1615-9861.

WILLIAMSON, P.; NAIRN, M. E. Lesions caused by Corynebacterium pseudotuberculosis in the scrotum of rams. **Australian Veterinary Journal,** v. 56, n. 10, p. 496-498, 1980. ISSN 1751-0813.

WINDSOR, P. A. Control of caseous lymphadenitis. **Veterinary Clinics of North America: Food Animal Practice,** v. 27, n. 1, p. 193-202, 2011. ISSN 0749-0720.

XENARIOS, I. et al. DIP: the database of interacting proteins. **Nucleic acids research,** v. 28, n. 1, p. 289-291, 2000. ISSN 0305-1048.

YIN, L.; BAUER, C. E. Controlling the delicate balance of tetrapyrrole biosynthesis. **Philosophical Transactions of the Royal Society of London B: Biological Sciences,** v. 368, n. 1622, p. 20120262, 2013. ISSN 0962-8436.

ZHANG, R.; OU, H. Y.; ZHANG, C. T. DEG: a database of essential genes. **Nucleic acids research,** v. 32, n. suppl 1, p. D271-D272, 2004. ISSN 0305-1048.

ZHANG, X.; XU, J.; XIAO, W.-X. A New Method for the Discovery of Essential Proteins. **PloS one,** v. 8, n. 3, p. e58763, 2013. ISSN 1932-6203.

ZHOU, H. et al. Stringent homology-based prediction of H. sapiens-M. tuberculosis H37Rv protein-protein interactions. **Biol Direct,** v. 9, n. 5, 2014.

ZORAGHI, R.; REINER, N. E. Protein interaction networks as starting points to identify novel antimicrobial drug targets. **Current opinion in microbiology,** v. 16, n. 5, p. 566-572, 2013. ISSN 1369-5274.

Anexos

I - *C. pseudotuberculosis Phop* confers virulence and may be targeted by natural compounds

Sandeep Tiwari, Marcília Pinheiro da Costa, Sintia Almeida, Syed Shah Hassan, Syed Babar Jamal, Alberto Oliveira, Edson Luiz Folador, Flavia Rocha, Vinícius Augusto Carvalho de Abreu, Fernanda Dorella, Rafael Hirata, Diana Magalhães de Oliveira, Maria Fátima da Silva Teixeira, Artur Silva, Debmalya Barh e Vasco Azevedo

Após a construção de uma rede de interação proteína-proteína, seja por método experimental ou computacional, diversas análises podem ser executadas. Dentre estas análises, podemos citar a comparação entre duas ou mais redes de interação, a análise de um conjunto específico de proteínas como um cluster, a análise de uma via metabólica de interesse ou mesmo análise de interação entre proteínas específicas.

Neste trabalho, foi gerada a rede de interação parcial para as proteínas codificadas por dois genes específicos de interesse: *phoP* e *phoR.* A rede de interação, contendo do primeiro até o terceiro nível de interação do sistema *phoPR*, permitiu o planejamento de experimentos em laboratório para verificar como a expressão destes dois genes poderiam regular a expressão de outras proteínas. Após submissão do artigo, visto que haviam evidências experimentais comprovando os resultados, a pedido dos revisores, a imagem da rede de interação foi retirada (Figure 45).

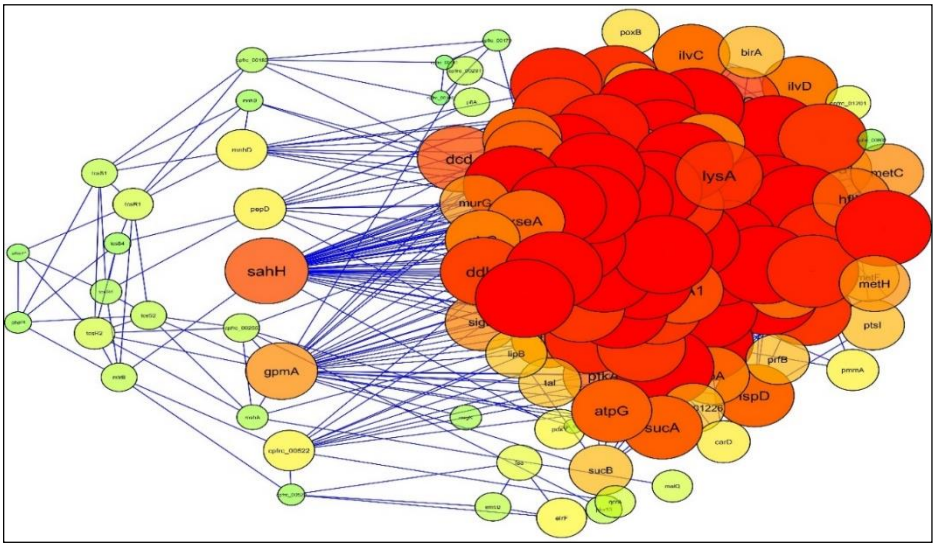


Figure 45. Rede de interação parcial das proteínas codificadas pelos genes phoPR.

Este trabalho desenvolvido em colaboração com o MSc. Sandeep Tiwari e foi publicado em setembro de 2014 pela revista Integrative Biology com DOI número 10.1039/C4IB00140K, disponível em http://pubs.rsc.org/en/content/articlehtml/2014/ib/c4ib00140k.

Integrative Biology

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/c4ib00140k

C. pseudotuberculosis Phop confers virulence and may be targeted by natural compounds†

Sandeep Tiwari,‡[a] Marcília Pinheiro da Costa,‡[b] Sintia Almeida,[a] Syed Shah Hassan,[a] Syed Babar Jamal,[a] Alberto Oliveira,[a] Edson Luiz Folador,[a] Flavia Rocha,[c] Vinícius Augusto Carvalho de Abreu,[a] Fernanda Dorella,[a] Rafael Hirata,[d] Diana Magalhaes de Oliveira,[b] Maria Fátima da Silva Teixeira,[b] Artur Silva,[e] Debmalya Barh[f] and Vasco Azevedo*[ac]

Received 18th June 2014,
Accepted 22nd August 2014

DOI: 10.1039/c4ib00140k

www.rsc.org/ibiology

The bacterial two-component system (TCS) regulates genes that are crucial for virulence in several pathogens. One of such TCS, the PhoPR system, consisting of a transmembrane sensory histidine kinase protein (PhoR) and an intracellular response regulator protein (PhoP), has been reported to have a major role in mycobacterial pathogenesis. We knocked out the *phoP* in *C. pseudotuberculosis*, the causal organism of caseous lymphadenitis (CLA), and using a combination of *in vitro* and *in vivo* mouse system, we showed for the first time, that the *PhoP* of *C. pseudotuberculosis* plays an important role in the virulence and pathogenicity of this bacterium. Furthermore, we modeled the *PhoP* of *C. pseudotuberculosis* and our docking results showed that several natural compounds including Rhein, an anthraquinone from *Rheum undulatum*, and some drug-like molecules may target *PhoP* to inhibit the TCS of *C. pseudotuberculosis*, and therefore may facilitate a remarkable attenuation of bacterial pathogenicity being the CLA. Experiments are currently underway to validate these *in silico* docking results.

Insight, innovation, integration

Here, we report for the first time the importance of *PhoP* protein of the PhoPR system of *C. pseudotuberculosis* that plays a vital role in the virulence and pathogenicity of this bacterium. For this, we developed a mutant *phoP* gene of *Corynebacterium pseudotuberculosis* Cp1002 strain and subsequently evaluated the consequences of this mutation *via in vitro* and *in vivo* analyses. Furthermore, by extrapolating the results of our analyses, we modeled this *PhoP* protein of *C. pseudotuberculosis* and computational analyses were performed. Our docking results showed that several natural compounds, including Rhein from *Rheum undulatum* and some other drug-like compounds acquired from public drug database/s, targeted the N-terminal response regulator domain of the *PhoP* protein to inhibit the TCS of *C. pseudotuberculosis* and might facilitate a remarkable attenuation of the pathogenicity of this important veterinary pathogen.

Introduction

The Gram-positive bacterium *Corynebacterium pseudotuberculosis* of the class *Actinobacteria* is the etiological agent of caseous lymphadenitis (CLA), or cheesy gland disease, an illness that affects small ruminants (sheep and goats) worldwide.[1] CLA is mainly characterized by abscess formation in superficial and internal lymph nodes.[2] In some cases, the visceral organs,

[a] *PG Program in Bioinformatics, Laboratory of Cellular and Molecular Genetics (LGCM), Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil. E-mail: sandip_sbtbi@yahoo.com, sintiaalmeida@gmail.com, hassan_chemist@yahoo.com, Jamal-syedbabar.jamal@gmail.com, afojunior@gmail.com, edson.folador@gmail.com, vini.abreu@gmail.com, fernandadorella@gmail.com, vasco@icb.ufmg.br*

[b] *Center for Genomics and Bioinformatics, State University of Ceará, Fortaleza, CE, Brazil. E-mail: marciliacosta@hotmail.com, Diana.magalhaes@uece.br, mfteixeira@hotmail.com*

[c] *Department of General Biology, PG Program in Genetics, Laboratory of Cellular and Molecular Genetics (LGCM), Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil. E-mail: flasouz@yahoo.com.br, vascoariston@gmail.com; Fax: +55-31 3409-2610; Tel: +55-31 3409-2610*

[d] *Universidade do Estado do Rio de Janeiro, UERJ, Brazil. E-mail: rhiratajunior@gmail.com*

[e] *Laboratory of DNA polymorphism (LPDNA), Federal University of Pará, Belém, PA, Brazil. E-mail: asilva@ufpa.br*

[f] *Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, WB, India. E-mail: dr.barh@gmail.com*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c4ib00140k

‡ These authors contributed equally to this work.

such as the lungs, kidneys, liver and spleen, are also infected.[3] The development of CLA compromises these animals and generates significant economic losses due to reduced wool, meat and milk yields, decreased reproductive efficiency and increased condemnation of carcasses and skins in abattoirs.[2,4] Once disseminated, CLA eradication is difficult due to the inefficacy of currently available drug therapies.[2,3] The insufficiency of available control procedures and the lack of knowledge regarding the molecular mechanisms of the virulence of this disease require an aggressive search for efficient treatment strategies.[5] Recently, novel virulence determinants of *C. pseudotuberculosis* Cp1002 have been identified; however, little is known about this complex pathogen. Phospholipase D and mycolic acids are the most important virulence factors identified in *C. pseudotuberculosis* Cp1002.[6]

The role of two-component signal transduction systems (TCS) in bacteria and certain eukaryotes, such as protozoa, is to detect and respond to the milieu vicissitudes.[7] Many studies have reported the involvement of the versatile TCS systems in the regulation of a variety of processes, including cell division, nutrient acquisition, regulation of osmolarity, redox potential, nitrogen fixation, phosphate uptake, sporulation, pilus formation, adhesion, drug resistance and expression of virulence factors.[8–11] The PhoPR system consists of a transmembrane sensory histidine kinase protein (PhoR) that phosphorylates the receiver domain of the response regulator protein (PhoP) of this system. Phosphorylation induces a conformational change in the response regulator, which activates the effectors domain and thereby triggering the cellular response.[12]

More than 4000 TCSs have been identified in bacterial genomes, demonstrating a broad range of interactions and environmental adaptation of bacteria. Interestingly, the PhoPR system from *M. tuberculosis* has shown a high degree of similarity to PhoPQ from a diverse range of intracellular bacterial pathogens such as *Salmonella* sp., *Shigella* sp. and *Yersinia* sp.[13] The development and application of specific inhibitors against TCS systems operate differently from conventional antibiotics. It is possible to transform such inhibitors into new drugs, which perform selectively against various drug-resistant bacteria.[14,15] Furthermore, inhibiting the function of TCSs that control the expression of virulence factors may attenuate the virulence of the pathogenic bacteria.[16]

PhoP gene is the response regulator of the PhoPR two-component systems, which plays a crucial role in the virulence. Progresses in the understanding of *M. tuberculosis* biology have promoted the rational development of attenuated strains, and, currently, the *phoP* mutant is used as a main vaccinal strategy against tuberculosis.[17] Recent studies have demonstrated that this mutant has shown a variety of changes, including the loss of expression of lipids of the cell envelope, impaired intracellular growth in infected macrophages, which persists in *in vitro* cultured-macrophages and in mouse organs, lack of secretion of the major T-cell antigen ESAT-6, virulence attenuation and protective efficacy against tuberculosis.[18,19]

In this work, we have attempted to study the possible role of the PhoP protein of the two-component signal transduction systems in *C. pseudotuberculosis* Cp1002 through the generation of mutant strain by disrupting the *phoP* gene by simple homologous recombination. After the *ΔphoP* genotype analysis, tests were performed to confirm the role of PhoP in the virulence and pathogenicity of *C. pseudotuberculosis*. This mutant strain was also assessed for adhesion and viability in macrophages and for virulence in mice to analyze the persistence of infection and mortality rates. After confirming the possible role of PhoP protein in virulence, we followed an *in silico* approach with a focus on the PhoP protein of the TCS of animal pathogen *C. pseudotuberculosis* Cp1002 for identifying putative therapeutic inhibitors.

Materials and methods

Bacterial strains and culture conditions

All bacterial strains and plasmids used in this study are listed in Table 1. *Escherichia coli* DH5α was grown in Luria-Bertani broth (LB, Difco Laboratories, Detroit, USA) at 37 °C under agitation for 18 h and on 1.5% (w/v) LB agar plates at 37 °C for 18 h. Plasmid-containing transformants were selected by adding ampicillin (Amp) (Invitrogen, San Diego, CA) and X-Gal (Invitrogen, San Diego, CA) to the media. The supplement concentrations were ampicillin (100 μg ml^{-1}) and X-Gal (40 μg ml^{-1}), respectively. *C. pseudotuberculosis* biovar *ovis* strain 1002 was used as the wild type parental. Bacteria were aerobically grown in brain heart infusion broth (BHI, Acumedia Manufacturers, Inc., Baltimore, MD, USA) and on 1.5% (w/v) BHI agar plates at 37 °C for 72 h.[20] *C. pseudotuberculosis phoP* mutant (*ΔphoP*) was aerobically grown in brain heart infusion broth (BHI) and on BHI agar plates at 37 °C for 72 h. For the selection of recombinants, kanamycin (Km) 50 μg ml^{-1} was added to the media.

Genomic DNA extraction and the construction of *C. pseudotuberculosis ΔphoP* strain

Genomic DNA extraction was performed according to a previously described protocol.[20] For the amplification of a segment

Table 1 Strains and plasmids used in this work

Strain or plasmid	Description and relevant characteristics	Source
Escherichia coli	DH5α/[*sup*44Δ*lac*U169 (φ 80 *lac*ZΔM15) *hsd*R17 *rec*A1 endA1 *gyr*A96 *thi*-1 *rel*A1]	Invitrogen
Corynebacterium pseudotuberculosis[a]	1002 strain (wild-type)	UFBA
Corynebacterium pseudotuberculosis	Δ*phoP* strain (*phoP* mutant)	This study
pCR®2.1-TOPO®	Cloning vector/Am^r-Km^r/pUC ORI	Invitrogen
pCR®2.1-TOPO®:*phoP*	Cloning vector pCR®2.1-TOPO® containing the *C*p 436 bp fragment from the ORF *phoP*	This study

[a] Virulent *C. pseudotuberculosis* biovar *ovis* strain isolates from caprines; obtained from the Universidade Federal da Bahia, UFBA, Brazil.

of phoP ORF of 436 bp of *C. pseudotuberculosis* Cp1002 the following primers were used: sense (5′-GGATCCGATGGAAGG CGTGAACGAG-3′) antisense (5′- GCGTAAGATCGGCGTACACC-3′). The PCR assays were carried out in a final volume of 50 μl, containing 50 ng of genomic DNA, 1 pmol ml^{-1} of each primer, 0:25 mM dNTPs, 0.1 units of Taq DNA polymerase (Invitrogen), 2 mM $MgCl_2$, and 1× concentrated enzyme buffer (Invitrogen). Amplification was performed using the thermal cycler (PTC-100, MJ Research, Inc.) as follows: first denaturation at 95 °C for 4 min, 30 cycles followed by denaturation at 95 °C for 30 s, 58 °C annealing for 30 s, extension at 72 °C for 1.5 min; and final extension for 5 min at 72 °C. The DNA fragment was purified from bands in 1.0% (w/v) agarose gels using the Consert TM Rapid Gel Extraction System kit (Gibco-BRL, Gaithersburg, MD, USA). The retrieved fragment of *phoP* was then ligated into the pCR2.1-TOPO vector, as described in the manufacture's protocol. The recombinant plasmid pCR 2.1-TOPO/*phoP* was then introduced into the competent *E. coli* DH cellα and single recombinant colonies were selected. The presence of insert DNA fragment was confirmed by colony PCR of seven clones selected using the same primers described above. In this reaction, for final volume of 10 L, 1 pmol ml^{-1} of each primer was used; 0.25 mM dNTPs; 0.1 unit of Taq DNA polymerase (Invitrogen); 2 mM $MgCl_2$ buffer and 1× enzyme concentrate (Invitrogen). After the identification of *E. coli* clone in the genomic library that contained the *phoP* fragment cloned into pCR2.1-TOPO (Invitrogen), plasmid DNA extraction was performed using the Wizard Plus Maxipreps DNA Purification System (Promega). The extracted plasmid was directly transformed into *C. pseudotuberculosis* strain 1002 according to Dorella *et al.*[1] The selection of the phoP mutant clones was performed in BHI medium supplemented with 50 μg mL^{-1} kanamycin. For the construction of *C. pseudotuberculosis* Δ*phoP* by simple homologous recombination,[21] a clone of *C. pseudotuberculosis* genome library was used, which contained a fragment of the open reading frame (ORF) of the cloned *phoP* gene. To confirm *phoP* inactivation by the insertion of the suicide vector, polymerase chain reaction (PCR) was performed using primers aligned with *phoP* as well as m13 and km. All other molecular biology techniques were performed according to Sambrook *et al.*[51]

The constructed suicide plasmid was then introduced into competent *C. pseudotuberculosis* cells. This plasmid did not replicate in *C. pseudotuberculosis* and it should integrate into the genome by homologous recombination between the *phoP* gene on the chromosome and the cloned sequence on the plasmid (Fig. 3). The resistant strain of *C. pseudotuberculosis* was confirmed by mPCR. The disruption of *phoP* gene of this strain was evaluated by PCR using pair primers listed in Table 2. Amplicons could also be observed in Table 3. Compared with the parent strain Cp1002, mutant (Δ*phoP*) strain was confirmed by the detection of *phoP* sequence in *C. pseudotuberculosis*. A 5.2 kb fragment was obtained in mutant *phoP*, indicating the presence of plasmid into genomic DNA (Fig. 4a, lane 15), while the wild type yielded a 0.8 kb product (Fig. 4a, lane 14).

Table 2 Primers used to confirm the deletion of *phoP* gene of *C. pseudotuberculosis* strain

Primer designation	Orientation	Sequence (5′→ 3′)
phoPF	Forward	GGATCCGATGGAAGGCGTGAACGAG
phoPR	Reverse	AAGCTTTTACGTATTCCGAGGCTTAC
phoPRfrag	Reverse	GCGTAAGATCGGCGTACACC
phoPFext	Forward	CTTTACTAAGAGATCGGGG
phoPRext	Reverse	GCTCATCTACTACTTTCTGC
KmF	Forward	ATGATTGAACAAGATGGATTG
KmR	Reverse	TTAATAATTCAGAAGAACTC
M13F	Forward	GTAAAACGACGGCCAG
M13R	Reverse	CAGGAAACAGCTATGAC

Generation and characterization of *phoP* mutant of *C. pseudotuberculosis*

High concentrated plasmid DNA was isolated from cells using the Wizard® Plus Maxipreps DNA Purification System kit (Promega, Madison, WI). The recombinant plasmid (pCR®2.1-TOPO®/*phoP fragment*) was then introduced into competent *C. pseudotuberculosis* cells.[22] Plating recognized strains with homologous recombination of recombinant plasmid at the phoP locus on medium containing kanamycin. Resistant colonies appeared after 72 h at 37 °C, and single recombinant colony was selected and characterized.

The chosen mutant strain of *C. pseudotuberculosis* Cp1002 was evaluated by mPCR, as described by Pacheco *et al.*, 2007.[20] To compare the growth rate of *C. pseudotuberculosis* wild type and phoP mutant, a growth curve was made in BHI T80 (0.05% Tween 80), and the OD_{600} and serial plate agar dilution were measured at 0, 30, 60, 120, 180, 240, 300 and 360 min. The difference between the bacterial size and colony morphology were also observed between the wild type and mutant strains.

The proper integration of the plasmid was further determined by PCR using the primers of Table 2. PCR primer pairs, melting temperature and amplicon lengths of each PCR reaction can be seen in Table 3.

In vitro assays of *C. pseudotuberculosis* strains in macrophage cells

The J774 murine macrophage cells were cultured in suspension in Dulbecco's Modified Eagle's medium (DMEM, Sigma Chemical Co., St. Louis, MO, USA) supplemented with gentamicin (50 μg ml^{-1}), fungizone (2.5 μg ml^{-1}) and 5% fetal calf serum (Gibco BRL, NY, USA) at 37 °C in an atmosphere of 5% (v/v) CO_2. Confluent monolayers were trypsinized at two day intervals with saline containing 0.2% trypsin (w/v) and 0.02% EDTA (w/v).

Table 3 Initiator combinations and melting temperatures used in each PCR reaction, and fragments produced by amplification

Primer pair	Mt (°C)	Amplicon (bp)
phoPF and *phoPR*	58	714
phoPF and *phoPRfrag*	58	436
M13F and *phoPRfrag*	58	548
M13F and *phoPR*	58	825
phoPF and *M13R*	58	525
KmF and *KmR*	56	795
phoPR and *KmR*	56	2532
phoPFext and *phoPRext*	57	5152

After 3 min of interaction, the culture medium was removed and the cells were incubated in DMEM, counted and diluted to 1×10^6 cells per ml. For adherence assay, an average of 500 μL of cell suspension cultures ($\sim 5 \times 10^5$ cells per well) were cultured in 24-well plates. Then, the cells were incubated for 48 h in a 5% CO_2 incubator to allow the cells to grow to about 95% confluency. *C. pseudotuberculosis* wild type and *phoP* mutant strains were cultivated in BHI and incubated for 24 h at 37 °C. Microorganisms were washed three times with Dulbecco's phosphate-buffered saline solution, resuspended in DMEM and diluted to a concentration of 10^7 colony-forming units (CFU) ml^{-1}. The suspension was diluted to 1:10 ($\sim 5 \times 10^6$ CFU ml^{-1}) with DMEM (500 μL per well). After interaction times (1, 3 and 6 h), the supernatants were removed, diluted and counted. Then, the monolayers were washed six times with Dulbecco's phosphate-buffered saline solution and resuspended with 500 μL of lysis buffer (0.1% Triton X-100 (Sigma) in Dulbecco's phosphate-buffered saline solution). The lysates were diluted and cultivated in BHI. Inoculated cells corresponded to the number of viable bacterial cells in supernatant plus the number of viable bacteria (intracellular plus extracellular) associated with cell monolayers. The total number of adhered cells was expressed as the percentage of the inoculum recovered from monolayers after 1, 3 and 6 h incubation. For the determination of intracellular viable bacteria after incubation times (1, 3 and 6 h), monolayers were washed six times with PBS, and treated with 150 μg ml^{-1} gentamicin sulphate (Sigma) for 1 h. The number of intracellular bacteria was determined by viable counts in BHI after the lysis of monolayers with 0.5 ml of 0.1% Triton X-100 (Sigma) in PBS. The percentage of intracellular bacteria was deduced from J774 cell-associated bacteria.

Immunization assay, challenge and easement of protection level

The standardization of parameters such as animal model, calculating the lethal dose (LD50) to be employed in the immunization studies, the volume of the cultures to be inoculated in animals, most appropriate route of inoculation, intervals between immunizations and challenges, and use of Glanvac™ 3 (P-fazer) as control of immune response were already performed by Dorella (2009)[52] and Moraes (2014).[23] All procedures with animals were carried out according to the regulations of the Ethics Committee for Animal Experimentation of the Federal University of Minas Gerais, Brazil. For experiments evaluating virulence through experimental infection, 6–8 weeks-old BALB/c mice were divided into four groups of 10 animals each. All groups of animals used in immunization trials were submitted to blood sampling on days 0 and 14 starting from the time of immunization. The blood was collected from the retro-orbital vein plexus with the help of Pasteur pipettes. After coagulation, the blood was centrifuged at 3000 rpm for 10 min. The serum was removed and stored at −20 °C until the determination of specific immunoglobulins by ELISA. In pilot immunization experiments, mice were grouped into four (Group 1, 2, 3 and 4) groups with each group consisting of ten animals. On day 0 ($t = 0$), the Group 1 mice were immunized with a commercial vaccine Glanvac™ 3 (P-fazer) subcutaneously in a volume of 300 μL. Four weeks later, on day 28 ($t = 28$), the same group of mice received a second dose of the vaccine (re-vaccination) under the same conditions as the initial dose. Group 2 were vaccinated with 0.9% saline (negative control) intraperitoneally in a volume of 100 μL ($t = 0$). Group 3 was immunized with the parental strain *C. pseudotuberculosis* Cp1002 intraperitoneally in a volume of 100 μL containing 10 CFU per ml ($t = 0$). The fourth group was immunized intraperitoneally with *C. pseudotuberculosis* mutant strain PhoP in a volume of 100 μL containing 10 CFU per ml ($t = 0$). The animals were challenged intraperitoneally 14 days after immunization with 10 CFU per ml of the virulent strain of *C. pseudotuberculosis* MIC-6 strain. The protection conferred by the immunization process was evaluated by comparing the survival of immunized animals to those inoculated with the commercial vaccine Glanvac™ and the wild strain of *C. pseudotuberculosis*. Mice were evaluated for four weeks after the challenge, and the entire experiment was performed in triplicate (Fig. 1).

Detection of specific IgG, IgG1 and IgG2a antibodies

Serum samples were taken 14 days after immunization. Blood samples were collected through retro-orbital bleeding. After coagulation, the blood was centrifuged at 3000 rpm for ten minutes. The serum samples were collected and stored at −20 °C. These samples were analyzed using an enzyme-linked immuno-sorbent assay (ELISA) to measure the total levels of specific IgG, IgG1 and

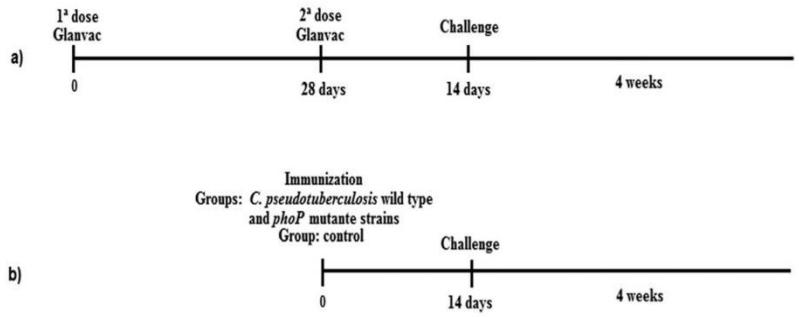


Fig. 1 Vaccination plan. (a) Schematic diagram illustrating the timelines of the vaccination of Glanvac™ 3 group in two doses, infectious challenge and monitoring time after challenge; (b) Illustration of the immunization of other groups, infectious challenge and monitoring time after challenge.

IgG2a antibodies. ELISA was performed according to a previously describe protocol by Ribeiro *et al.*[24]

Statistical analysis

All data were expressed as means ± standard deviation (S.D) and analyzed using GraphPad Prism (v4.03, GraphPad Software, San Diego, CA). Statistical differences among groups were identified using one-way ANOVA. A one-tailed Student's *t*-test was used to determine if there were any significant differences between the experimental and control groups. A P value of 0.05 or less was considered to be significant.

Homology modeling and protein model validation

The amino acid sequences of *C. pseudotuberculosis* Cp1002 PhoP protein were retrieved from the UniProt (http://www.uniprot.org) that has 237 amino acids (Accession number: D8KPL7). The query sequence was searched for identity analysis using the Basic Local Alignment Search Tool (BLAST)[25] against Protein Data Bank (PDB) for the corresponding template structure. The structure of PhoP protein from *Mycobacterium Tuberculosis* (PDB ID: 3R0J-A) was found to be the best template for the PhoP protein with an identity of 64% and a sequence similarity of 80%. PhoP protein was modeled in 2012 by Moraes *et al.*[26] where the identity was 39% and the sequence similarity was 60% with the selected template from *T. maritima* (PDB ID: 1KGS); up to 2012, the crystal structure of PhoP protein from *Mycobacterium Tuberculosis* (PDB ID: 3R0J-A) was not present in the PDB database. The computational 3D (three-dimensional) structure of *C. pseudotuberculosis* for PhoP was generated by comparative homology modeling using 3R0J-A template structures by through SWISS-MODEL.[27] The protein structure was validated using various bioinformatics tools such as Procheck[28] and ANOLEA (Atomic Non-Local Environment Assessment).[29] The best model structure was then compared with the template protein structure by superimposing both the structures *via* the Chimera program.[30]

Ligand library preparation

A selection of active antimicrobial compounds isolated from different natural sources was carried out following a literature review. The structures of these compounds were retrieved from PubChem (http://pubchem.ncbi.nlm.nih.gov/) and the Drug Bank databases (http://www.drugbank.ca/). Calculation for various drugs properties, such as mutagenic, tumorigenic, irritant nature and an adverse effect of the compounds on the reproductive system, was performed using Molinspiration (http://www.molinspiration.com/cgi-bin/properties) and Osiris Property Explorer (http://www.organic-chemistry.org/prog/peo/). The tools also give the drug-likeness and total drug-score of the compounds based on the "Lipinski rule of five". The overall drug score of a compound is the drug-likeness, *i.e.*, $c\log P$ (logarithm of the partition coefficient between 1-octanol and water, for the hydrophilicity of the compound, where low $\log P$ refers to high absorption or permeation, value less than 5), $\log S$ (a unit stripped logarithm (base 10) for the aqueous solubility of a compound in mol l^{-1}), molecular weight and toxicity risk were calculated by Osiris Property Explorer for each compound, thus having an overall potential to qualify for a drug. Furthermore, a second set of the ligand library was developed from the ZINC database,[31] by extracting 11 000 drug-like molecules with *Tanimoto* cut-off level of 60%, and was subsequently screened against the Phop protein for the identification of putative inhibitors.

Active site residues identification and docking analysis

The 3D structure of the PhoP protein from *C. pseudotuberculosis* Cp1002 was further checked for their active site amino acid residues. For this, the published structural data related to the PhoP protein from *C. pseudotuberculosis* Cp1002 and the template structure of the target protein was also checked to confirm a key role of these potential residues in substrate recognition. The docking analyses was performed using Molegro Virtual Docker.[32]

Result and discussion

Construction and characterization of *phoP* mutant

To study the importance of the *phoP* regulatory gene in the regulation of virulence and pathogenicity in *C. pseudotuberculosis*, a *phoP* mutant was generated. We investigated the consequences of this mutation using a combination of *in vitro* and *in vivo* assays. The construction of *C. pseudotuberculosis* strain carrying chromosomal deletion for the *phoP* gene was obtained by the insertion of a plasmid containing a part of the *phoP* gene into the chromosomal copy of *phoP* by homologous recombination. To do this, we first amplified a fragment of the *phoP* gene of 436 bp by PCR using primers listed in Table 2. Subsequently, the gene sequence was cloned in the pCR®2.1-TOPO® plasmid (Fig. 2). The presence or absence of the insert was confirmed by restriction enzyme digestion with *Eco*RI (Fig. 2).

The wild type and mutant strains observed under optical microscopy showed no obvious difference in the shape and size of the bacteria (data not shown). No change was verified in the morphology of colony compared with both the strains (data not shown). No difference was observed in the *in vitro* growth curves of the parent and mutant strains, implying that the deletion of the *phoP* gene had no significant influence on the growth of *C. pseudotuberculosis* (Fig. 5).

Macrophages adhesion and intracellular viability of *C. pseudotuberculosis*

Intracellular pathogens, such as *C. pseudotuberculosis*, have the ability to adhere to host cells, internalize, survive and replicate within infected cells, thereby causing the illness. Thus, to understand the changes caused by mutation in *phoP* gene of *C. pseudotuberculosis*, we studied the adhesion and invasion of wild type and mutant strains in J774 murine macrophage cells. The bacterial adherence and the invasion rates of wild type and mutated strains are given in Table 4.

The results correlated with the percentage of bacteria that adhered to macrophage cells are shown in Fig. 6a. In the adherence assay, in three times post-infection (1, 3 and 6 h),

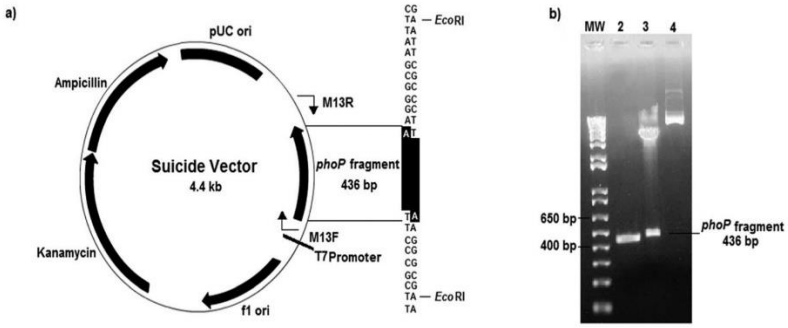


Fig. 2 (a) Schematic representation of recombinant plasmid pCR®2.1-TOPO®/*phoP* fragment with restriction sites. Plasmid was used to clone the *C. pseudotuberculosis phoP* fragment linked to promoter T7. (b) Plasmid isolated, purified and then visualized on 1% agarose gel. The 1 kb Ladder plus molecular marker (Invitrogen) is indicated as MW (first lane); lane 2: amplification of the *C. pseudotuberculosis phoP* fragment (436 bp amplicon); lane 3: recombinant plasmid pCR®2.1-TOPO®/*phoP* fragment, after digestion with *Eco*RI; lane 4: recombinant plasmid pCR®2.1-TOPO®/*phoP* fragment did not undergo digestion. Molecular masses are indicated.

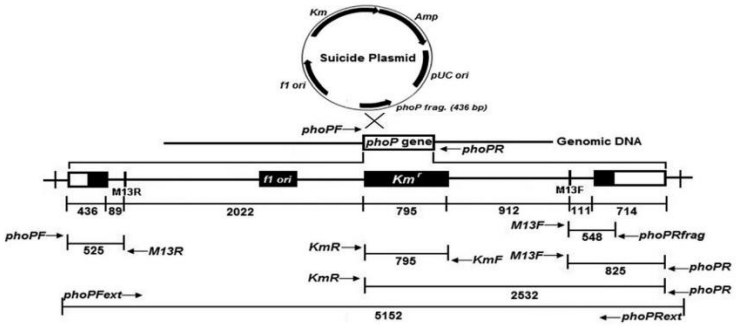
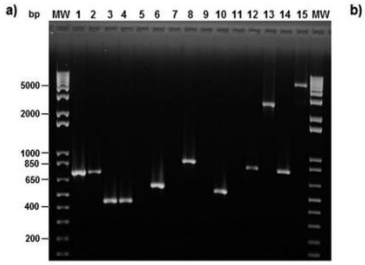


Fig. 3 Integration of suicide plasmid (pCR®2.1-TOPO®/*phoP* fragment) disrupted the *phoP* gene. Homologous recombination (X) to give the fragments indicated on the diagram. The solid boxes represent fragments from suicide plasmid. The size of portions and primers also are indicated.



b)

Lane	Strain	Primers	Amplicon (bp)
1	Wild-type (*Cp*1002)	*phoPF + phoPR*	714
2	*phoP* Mutant	*phoPF + phoPR*	714
3	Wild-type (*Cp*1002)	*phoPF + phoPRfrag*	436
4	*phoP* Mutant	*phoPF + phoPRfrag*	436
5	Wild-type (*Cp*1002)	*M13F + phoPRfrag*	—
6	*phoP* Mutant	*M13F + phoPRfrag*	548
7	Wild-type (*Cp*1002)	*M13F + phoPR*	—
8	*phoP* Mutant	*M13F + phoPR*	825
9	Wild-type (*Cp*1002)	*M13R + phoPF*	—
10	*phoP* Mutant	*M13R + phoPF*	525
11	Wild-type (*Cp*1002)	*KmF + KmR*	—
12	*phoP* Mutant	*KmF + KmR*	795
13	*phoP* Mutant	*phoPR + KmR*	2532
14	Wild-type (*Cp*1002)	*phoPFext + phoPR ext*	787
15	*phoP* Mutant	*phoPFext + phoPR ext*	5152

Fig. 4 Confirmation of *phoP* gene mutation in *C. pseudotuberculosis*. (a) Analysis of PCR products amplified from genomic DNA of the wild type and mutant strains visualized on 1% agarose gel electrophoresis. The 1 kb Ladder Plus Molecular Marker (Invitrogen) is indicated as MW (first and last lanes). Molecular masses are indicated. (b) Detail of the amplicons of the agarose gel. Lanes in the gel, strains, pair primers in each PCR reaction and the length of amplification fragments are shown.

a statistically significant difference was observed between the wild type and mutated strains. At all times, the adherence rate remained low in the mutated strain. After evaluating the percentage of viable intracellular bacteria in both the strains at two points in time (1 and 3 h post-infection), it was possible to observe a significant increase in the number of *phoP* mutant viable in macrophages, compared with wild-type, while no difference was noted at 6 h (Fig. 6b). Intracellular viability was low for the wild-type bacteria at 1 and 3 h post-infection. The number of cells of mutant strain adhered at 1 and 3 h post-infection (35.44% and 60.37%, respectively) and the high percentage of internalized cells (65.54% and 74.28%, respectively)

cxci

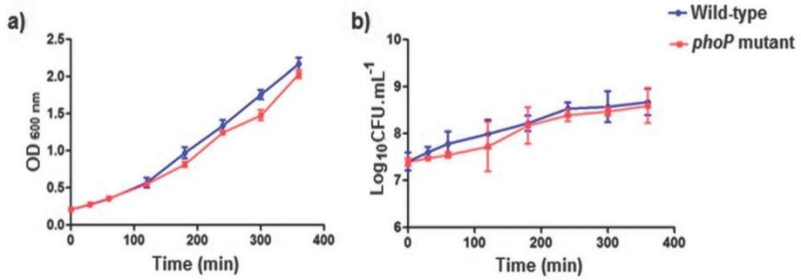


Fig. 5 Growth curves of wild type and mutant *C. pseudotuberculosis* strains. Cultures were grown in BHI medium supplemented with Tween 80 (0.05%). Growth profiles of both strains in (a) OD_{600nm} and (b) CFU per milliliter were measured during 360 min. No significant differences could be detected between both the groups. Mean values are the result of three independent experimental data ± S.D.

Table 4 Percentages of adhesion and internalization of *C. pseudotuberculosis* strains in J774 macrophage cells during three different time intervals. Each value represents the mean ± S.D. from triplicate measurements

	Interaction time (h)		
Wild-type strain of *C. pseudotuberculosis*	**1**	**3**	**6**
Adhered bacteria (%)	72.77 ± 3.65	90.1 ± 1.21	99.49 ± 0.06
Viable intracellular bacteria (%)	30.67 ± 2.80	48.83 ± 3.98	16.9 ± 0.79
	Interaction time (h)		
Δ*phoP* strain of *C. pseudotuberculosis*	**1**	**3**	**6**
Adhered bacteria (%)	35.44 ± 3.70	60.37 ± 6.70	74.13 ± 2.41
Viable intracellular bacteria (%)	65.54 ± 14.83	74.28 ± 4.38	18.48 ± 4.00

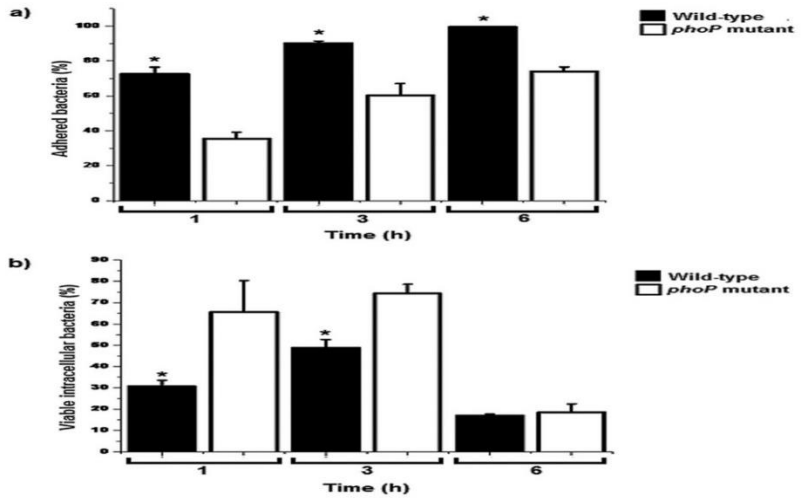


Fig. 6 Percentage of (a) adhered bacteria and (b) viable intracellular bacteria of wild-type and *phoP* mutant strains in J774 murine macrophage after three times. Mean ± S.D. for three independent replicated experiments are shown. Asterisks (*) indicate statistically significant differences between wild type and *phoP* mutant groups. Statistical analyses were performed using Student's *t*-test; $p < 0.05$ considered to be significant.

suggested that the reduced cell-adhesion rates were due to the rapid recognition and internalization of the mutant by macrophages. These results are in accordance with a study in *M. tuberculosis phoP* mutant, which reported that deficiency in some forms of cell envelope lipids affects the surface properties of bacteria and results in enhanced interaction with host cells.[33] In *Salmonella typhimurium*, phoP-regulated gene products decrease the processing and presentation of antigens because activated macrophages process *phoP* mutant with greater efficiency than wild type.[34] Therefore, the molecular basis for infecting macrophages by *C. pseudotuberculosis* still remains poorly understood.

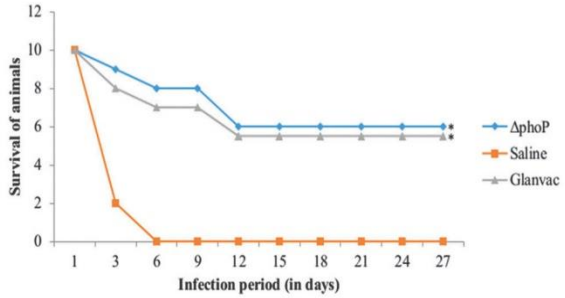


Fig. 7 Graphic: % survival or Number of survived animal *vs.* time after challenge (in days)

Immunization assay, challenge and easement of protection level

To assess the protection against CLA in mice, immunization assays were performed by intra-peritoneal inoculation with doses of *C. pseudotuberculosis* wild-type 1002 and *phoP* mutant strain, Glanvac, as well as a saline control. This was done by the determination of mice survival, after a challenge dose with the wild-type *C. pseudotuberculosis* MIC-6 strain.

All the mice vaccinated with the wild-type strain showed clinical signs of morbidity, and most mortality happened before the challenge. In this group, animals resistant to infection died within 2 days after the challenge. Mice immunized with saline died between 1 and 2 days after the challenge. Mice vaccinated with Glanvac presented a survival rate of 55% (Fig. 7). The survival rate of mice given the *ΔphoP* strain was the highest among the vaccine groups (60% survival) and was significantly different than that of the saline group. A single immunization with *phoP* mutant strain as a live vaccine conferred a significant protection against the bacterial infection in mice. These data provide the evidence that the deletion of *phoP* gene may have caused an attenuation of virulence in mice.

Detection of specific IgG, IgG1 and IgG2a antibodies

To verify the production of specific IgG antibodies, the serum samples from mice immunized with *C. pseudotuberculosis ΔphoP* were compared with saline (control) and Glanvac immunized mice by ELISA. IgG1 and IgG2a were investigated separately because IgG1 is related to a Th2 cellular immune response whereas IgG2a is related to a Th1 response in the same species. The preliminary results from the immunological assays showed that mice immunized with *C. pseudotuberculosis ΔphoP strain* developed high levels of IgG antibodies when compared to the control group. Similar substantial levels were observed in animals immunized with the vaccine Glanvac™ 3 (Fig. 8a).

Immunization with the phoP mutant strain also induced the production of IgG1 (Fig. 8b) in all the immunized mice. Analyzing the results shown in the graph indicate that IgG1 production rates in the group vaccinated with Glanvac™ 3 was significantly higher relative to the control group. This increase was also observed in the group inoculated with the phoP mutant showing a significant increase in the production of

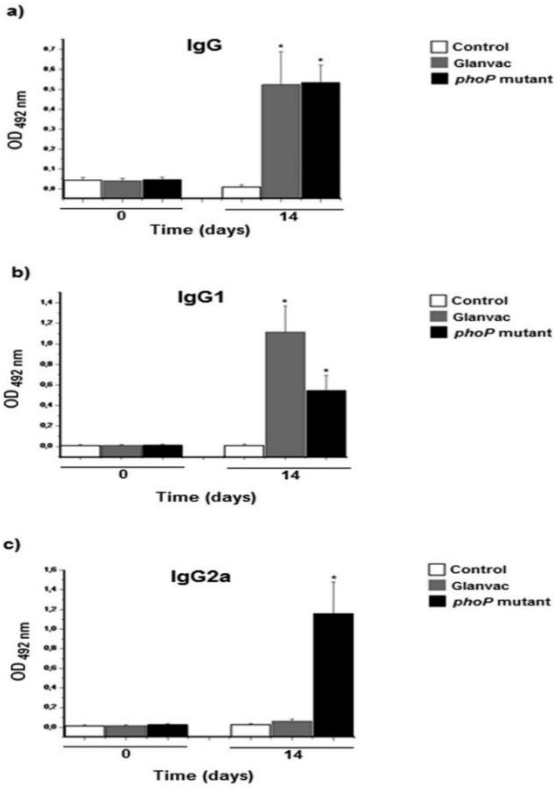


Fig. 8 Profile of total IgG antibody response after immunization in mice. (a) IgG antibody titer; (b) IgG1 iso-type titer; (c) IgG2a iso-type titer. Groups of mice were immunized as follows: intra-peritoneal injection of saline 0.9% (control), *phoP* mutant strain and Glanvac. Data are expressed as mean ± S.D. values. Results are representative of $n = 5$. Statistically significant differences between *phoP* mutant and Glanvac groups and control mice are denoted with an asterisk ($p < 0.05$).

this immunoglobulin; thus exhibiting a Th2 pattern immune response. Regarding the levels of IgG2a induced after immunization (Fig. 8c), inoculation with phoP mutant induced quite significant immunoglobulin response compared to the control group. The production of IgG2a in mice is characteristic of a Th1 type of immune response, which is responsible for the elimination of intracellular pathogens *C. pseudotuberculosis*,[35] suggesting that the protection type induced by mutant PhoP response is consistent with that expected in combating bacteria. Our results showed that the *C. pseudotuberculosis phoP* mutant conferred with a single dose partial protection against CLA. 60% of mice vaccinated with *phoP* mutant strain survived the lethal challenge with *C. pseudotuberculosis* MIC-6 strain. In contrast, other groups of animals (control, wild-type) were not protected against the same dose. These findings indicate an attenuated phenotype conferred by mutation in the *phoP* gene of *C. pseudotuberculosis*. Supporting this argument, other studies have also shown virulence attenuation in pathogens such as *M. tuberculosis* and *S. typhimurium*.[36] Vaccination with the *phoP* mutant induced significant production of mutant-specific IgG, IgG1 and IgG2, producing a mixed Th1/Th2, which is necessary for the effective protection against *C. pseudotuberculosis*.

The protection conferred by *C. pseudotuberculosis ΔphoP* strain may be associated with changes conferred in phoP gene mutation, as observed in *M. tuberculosis.*[37] Absence of the expression of important virulence genes may be the main factor justifying partial protection observed after challenge with the virulent strain of *C. pseudotuberculosis*. It has been reported in previous studies that TCS systems are involved in the regulation of a variety of processes, including cell division, nutrient acquisition, regulation of osmolarity, redox potential, nitrogen fixation, phosphate uptake, sporulation, pilus formation, adhesion, drug resistance and expression of virulence factors.[38,39] Therefore, we further expanded our analyses to find out some inhibitors for this protein obtained from natural sources and ZINC database and performed docking analysis.

PhoP protein homology modeling and structure validation

The PhoP protein 3D model was generated using SWISS-MODEL server and the best models were selected based on the PROCHECK and QMEAN6 scores analyses. PROCHECK is an analysis tool that provides an idea of the stereo-chemical quality of all the protein chains of a given PDB structure and generates a Ramchandran plot, as shown in ESI† (Fig. S1). The atomic empirical mean force potential analysis tool ANOLEA was used to assess the packing quality of the modeled structures. The *Y*-axis of the plot represents the energy for each amino acid of the protein chain. Negative energy values (in green) represent favorable energy environment whereas positive energy values (in red) represent unfavorable energy environment for a given amino acid. The ANOLEA results for PhoP protein are shown in Fig. S3, (ESI†). Inside the SWISS-MODEL Workplace, the QMEAN6[40] score evaluated the generated models. QMEAN6 is a reliability score for the entire model, which can be used to compare and rank alternative models for the same target protein structure. The structure quality ranges between values of 0 to 1 with higher values representing better models. The QMEAN6 score for the PhoP protein structure (0.689) is shown in Fig. S3 (ESI†). To observe the structural quality, we compared the template-query protein structures in Chimera using root mean square deviation metric (RMSD = 0.78), which measures the difference in the positions of corresponding Carbon α-atoms between two structures. The smaller the deviation, the better is the spatial arrangement of the two protein structures. The superimposition of protein structures was performed using the Chimera program, and the comparison is shown in Fig. S2 (ESI†).

Active site residues identification and docking analysis

As mentioned before, the two-component signal transduction system PhoPR of *C. pseudotuberculosis* consists of the Histidine kinases (PhoR protein) and the response regulator (PhoP protein). These systems are conserved in many bacterial species because they play an vital role in regulating cell cycle progression and development, which is vital for bacterial survival as well as for their adaptation to environmental changes.[41] The active site residues of the N-terminal receiver domain of the template crystallographic structure (PDBID: 3R0J-A from *M. tuberculosis*)[42] for PhoP protein were identified through the structural comparison of the model structure with the template protein and the

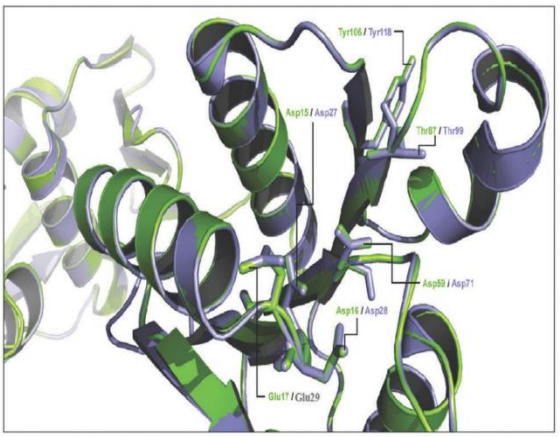


Fig. 9 Conserved residues of the PhoP protein (from N-terminal receiver domain) based on comparison with the corresponding template structure (PDB ID: 3R0J-A). PhoP Protein model is shown in limone and the template is shown in light blue.

residues are labeled in the target protein *via* the PyMol program (www.pymol.org/). Based on this comparison, the conserved residues found in the N terminal receiver domain of PhoP protein are: Asp15, Asp16, Asp59, Glu17, Thr87 and Tyr106 (Fig. 9). These amino acid residues were considered for docking analyses.

The docking analyses were performed using two set of ligands, one ligand library was developed using compounds isolated from different natural sources and the other ligand library of 11 000 drug-like compounds obtained from ZINC database were further considered for docking analysis. Natural compounds have already been reported to show antimicrobial activities and have fulfilled molecular and drug-like properties in accordance with the "Lipinski Rule of Five" and the result obtained from Molinspiration and Osiris scores (Table 5). The first library contained 30 compounds while the second library contained 11 000 compounds that were docked against the aforementioned identified conserved residues for PhoP protein. From our docking studies with these 30 compounds, we made an attempt to find out computationally the antibacterial compounds that showed strong binding affinities toward the aforementioned identified active site residues of the target PhoP protein. The compounds showing the best MolDock scores, their structures and interactions with the respective target protein residues are shown in Table 6 and Fig. 10. According to our observations, the compound Rhein (CID 10168), which is an anthraquinone substance obtained from various plant sources (*Rheum undulatum, Rheum palmatum*[43] and *Cassia reticulate*[44]), ranked as the best inhibitor molecule. It has already been established experimentally that this compound possesses antibacterial activity.[44] Thus, we consider Rhein to be the best inhibitor molecule with a drug-likeness of 0.18 and a drug-score of 0.61, showing good MolDock score and considerable binding interactions with the negatively charged acidic residues Asp15, Asp16, Asp59 and Glu17 of the receiver domain (Fig. 10 and Table 6). It has been reported that the aspartate receptor of the activation signal is

Table 5 Molinspiration and Osiris results of compounds selected for further docking analyses

Compounds	*c* log *P*	Solubility	Mol. weight	Drug-likeness	Drug-score
Liriodenin (CID 10144)	4	−6.25	275	−2.39	0.28
Pinostrobin (CID 73201)	3	−3.24	270	2.05	0.8
Axisonitrile-3 (CID 181226)	4.5	−3.47	231	−9.44	0.36
Ileabethoxazole (CID 9549062)	5.47	−5.77	325	−1.3	0.26
Rhein (CID 10168)	2.44	−4.15	284	0.18	0.61
Pilocarpine (CID 5910)	0.4	−1.2	208	1.17	0.85
Voacangine (CID 73255)	2.97	−3.86	368	1.8	0.7
Texalin (CID 473253)	2.8	−4.9	266	0.56	0.58
Araguspongine (CID 5276744)	5.98	−4.22	478	−1.7	0.24
Jacarandic acid (CID 73645)	3.88	−5.36	488	0.19	0.37
Leptophyllin B (CID 10447482)	4.06	−3.66	299	−8.03	0.37

Table 6 Docking results of the best-predicted natural compound (Rhein) that showed good interaction in our *in silico* analysis, with their MolDock Score and the interacting residues

Compound name	MolDock score	No. of H-bonds/ residues-compound interactions
Rhein (CID 10168)	−59.8635	**3** Asp15, Asp16, Asp59

the Asp59 residue (phosphorylation is absolutely conserved)[26] that is located at the N-terminal receiver domain of the response regulator of the PhoP protein, which along with Asp15 has an important role in Mg^{+2} binding.[42] In addition, after the virtual screening of 11 000 compounds, the top ranked 200 compounds were analyzed in Chimera for shape complementarity and hydrogen bond interactions, leading to the selection of a final set of 5 compounds (Table 7) for PhoP target protein, demonstrating that these compounds are binding to some of the same residues

Table 7 ZINC codes, MolDock scores and predicted hydrogen bonds for the five compounds selected among the top ranking 100 molecules against PhoP protein

ZINC IDs	MolDock score	No. of H-bonds/ residues-compound interactions
ZINC02127223	−85.678	**4** Asp59, Asp16, Asp15, Lys109
ZINC00078435	−84.6183	**3** Asp15, Asp16, Met61
ZINC01423999	−97.227	**4** Asp16, Asp59, Ays109
ZINC00406666	−95.2886	**2** Asp15, Asp59
ZINC01408034	−85.216	**3** Asp16, Asp59, Lys109

such as the natural antimicrobial Rhein compound. For example, the compound ZINC ID: 02127223 showed almost similar binding affinity with the target protein residues as Rhein (Fig. 11). The considerable binding affinities of Rhein compound isolated from natural sources and of other drug like compounds from the ZINC database support the poly pharmacological idea of promiscuous target protein. Because Rhein showed good interactions with the conserved residues of the PhoP receiver domain, compounds from the ZINC database might also act as putative drug molecules in accordance with the aforementioned drug likeness properties, which may possibly hinder the normal phosphorylation process of the target protein and inhibit the normal functioning of the PhoPR signaling cascade. It is obvious from our *in silico* analyses that the putative TCS inhibitors targeted the receiver domain of the response regulator protein and they might interfere with the

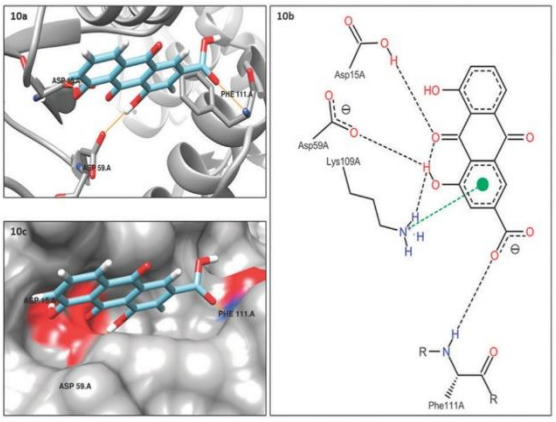


Fig. 10 (a) 3D cartoon representation of the docking analyses for PhoP protein structure with Rhein compound (CID 10168). (b) 3D surface representation of the docking analyses for the structures of Rhein compound with Phop receiver domains. (c) Two-dimensional representation of Rhein compound using PoseView[45] interacting with the conserved residues of the N-terminal receiver domain of the PhoP protein. The residues are negatively charged acidic Asp15, Asp59, Phe111 and Lys109.

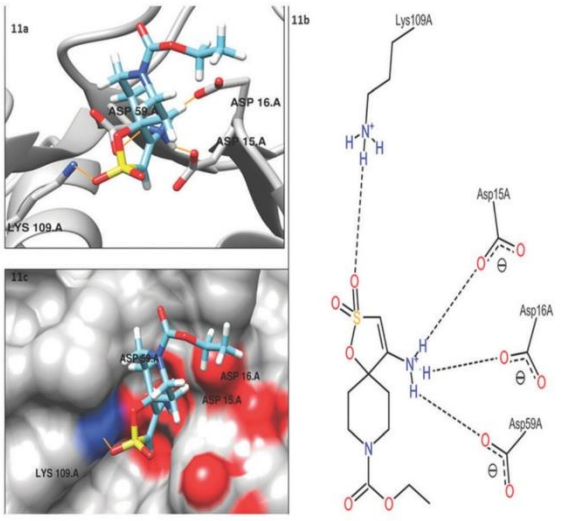


Fig. 11 (a) 3D cartoon representation of the docking analyses for PhoP protein structure with compound ZINC ID: 02127223. (b) 3D surface representation of the docking analyses for the structures of compound ZINC ID: 02127223 with Phop receiver domains. (c) Two-dimensional representation of compound ZINC ID: 02127223 using PoseView interacting with the conserved residues of the N-terminal receiver domain of the PhoP protein. The residues are negatively charged acidic Asp15, Asp16, Asp59 and Lys109.

domain normal function. Here, after verifying a possible role of the PhoP protein in the virulence of *C. pseudotuberculosis*, an additional computational approach has been taken to identify new TCSs specific inhibitors with potential drug-like properties for future experimental efforts (Fig. 10 and 11).

Conclusion

An effective prophylaxis for CLA requires the development of novel vaccine and drug candidates capable of generating an adequate immune response and protection against the disease.

Several virulence-attenuated mutants of intracellular pathogens have been used with promising vaccine strategy, including mutants of the gene *phoP*.[46–49] Alternatively, due to the importance of the TCS system, the same PhoP protein could be an attractive target for designing rational and target specific molecular inhibitors. In this context, we have initially described, in the present report, the development and use of a *C. pseudotuberculosis* Cp1002 vaccine strain based on the disruption of *phoP* gene. At a later stage, the 3D model of the PhoP protein has been subjected to detailed computational analyses *via in silico* approaches as a therapeutic target for the identification of novel inhibitors.

Our results confirm that the *phoP* mutated *C. pseudotuberculosis* lacked certain properties or functions associated with PhoPR system, and were affected by the absence of the PhoP protein. Under normal conditions, no significant differences were detected between the growth rate of *C. pseudotuberculosis phoP* mutant and wild-type strains. Similar result was found in *M. tuberculosis phoP* mutant during logarithmic and stationary phases of the growth curve.[13] Bacterial adhesion and invasion to the host cell are important steps in bacterial infection. Here, we reported the adhesion and invasion abilities of both *C. pseudotuberculosis* strains (wild-type and *phoP* mutant) in macrophages. Possible changes in the lipid composition of the cell envelope may have affected the interaction of *phoP* mutant strain with macrophages. Intracellular viability was related with the adhesive ability of strains. Our analyses suggest that the disruption of *phoP* gene might be involved in the alteration of some surface receptors. The elucidation of this bacterial strategy for *in vitro* infection is an important subject for understanding the dynamics of gene regulation during host interactions and for the development of a novel attenuated vaccine. The protective immunity of the vaccine based on *phoP* mutant of *C. pseudotuberculosis* was evaluated by determining the survival rates and the serum antibody titers of the mice. Our data indicate that immunization with *C. pseudotuberculosis phoP* mutant promoted cellular immune response and generated partial protection in mice possibly due to reduction of the virulence; however, further studies are required to understand the mechanisms of attenuation.

The overall importance and a promiscuous role of the PhoP protein brought us to follow some *in silico* approaches for the development of novel drugs. In this context, comparative studies are very important to predict new target genes. Based on this identification strategy, many bacterial genes have been revealed, and the genes of the PhoP regulatory locus, initially studied in Enterobacteria, are among them. Based on our observations, we have proposed that targeting the *PhoP* protein of the *PhoPR* system of the *C. pseudotuberculosis*, which have a similar role in regulating the expression of several genes to the closely related *M. tuberculosis*,[47,50] might help in attenuating pathogen growth in disease condition. The relationship of virulence-associated two-component PhoPR system in phylogenetically distant bacteria also supports our hypothesis. However, further experimental studies are required to discover the relationship between *C. pseudotuberculosis* PhoPR system and a possible virulence effect. Furthermore, our docking results have revealed that Rhein from *Rheum undulatum* and *Rheum palmatum* might be good therapeutic molecules for future wet lab studies. This might potentially inhibit the response regulator domain of the two-component signal transduction system, resulting in decreased pathogen virulence in *in vitro* experimentation. Because the PhoP protein is homologous and the residues are conserved across a broad range of closely related pathogenic bacteria (keeping in mind the non-pathogenic as well), the proposed drugs could target as many TCSs in the pathogenic bacteria as possible. In this context the PhoP inhibitors could be regarded as the broad-spectrum inhibitors and not only for *C. pseudotuberculosis*. The work presented here is the first ever conducted on the PhoP protein of *C. pseudotuberculosis* Cp1002 and optimistically the availability of these *in silico* information might serve as a basis for further future investigation.

Competing interests

The authors declare no conflict of interest.

Financial disclosure

We would like to gratefully acknowledge the help of all the team members & the financing agencies. Sandeep Tiwari acknowledges the receipt of fellowship from "TWAS-CNPq Postgraduate Fellowship Programme" for doctoral studies. This work was partially executed by Rede Paraense de Genômica e Proteômica supported by FAPESP (Fundação de Amparo à Pesquisa do Estado do Pará), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brasil), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil) and FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais, Brasil).

References

1 F. A. Dorella, M. S. Fachin, A. Billault, E. Dias Neto, C. Soravito, S. C. Oliveira, R. Meyer, A. Miyoshi and V. Azevedo, *GMR, Genet. Mol. Res.*, 2006, **5**, 653–663.
2 L. H. Williamson, *The Veterinary clinics of North America: Food animal practice*, 2001, **17**, 359–371.
3 R. G. Batey, *Aust. Vet. J.*, 1986, **63**, 269–272.
4 F. A. Dorella, L. G. Pacheco, S. C. Oliveira, A. Miyoshi and V. Azevedo, *Vet. Res.*, 2006, **37**, 201–218.

5 G. J. Baird and M. C. Fontaine, *J. Comp. Pathol.*, 2007, **137**, 179–210.
6 S. C. McKean, J. K. Davies and R. J. Moore, *Microbiology*, 2007, **153**, 2203–2211.
7 G. M. Pao and M. H. Saier Jr., *J. Mol. Evol.*, 1995, **40**, 136–154.
8 A. G. Blanco, M. Sola, F. X. Gomis-Ruth and M. Coll, *Structure*, 2002, **10**, 701–713.
9 J. A. Hoch, *Curr. Opin. Microbiol.*, 2000, **3**, 165–170.
10 I. Lopez-Goni, C. Guzman-Verri, L. Manterola, A. Sola-Landa, I. Moriyon and E. Moreno, *Vet. Microbiol.*, 2002, **90**, 329–339.
11 M. Matsushita and K. D. Janda, *Bioorg. Med. Chem.*, 2002, **10**, 855–867.
12 A. M. Stock, V. L. Robinson and P. N. Goudreau, *Annu. Rev. Biochem.*, 2000, **69**, 183–215.
13 E. Perez, S. Samper, Y. Bordas, C. Guilhot, B. Gicquel and C. Martin, *Mol. Microbiol.*, 2001, **41**, 179–187.
14 L. E. Ulrich, E. V. Koonin and I. B. Zhulin, *Trends Microbiol.*, 2005, **13**, 52–56.
15 D. Beier and R. Gross, *Curr. Opin. Microbiol.*, 2006, **9**, 143–152.
16 Y. Gotoh, Y. Eguchi, T. Watanabe, S. Okamoto, A. Doi and R. Utsumi, *Curr. Opin. Microbiol.*, 2010, **13**, 232–239.
17 P. J. Cardona, J. G. Asensio, A. Arbues, I. Otal, C. Lafoz, O. Gil, N. Caceres, V. Ausina, B. Gicquel and C. Martin, *Vaccine*, 2009, **27**, 2499–2505.
18 M. L. Chesne-Seck, N. Barilone, F. Boudou, J. Gonzalo Asensio, P. E. Kolattukudy, C. Martin, S. T. Cole, B. Gicquel, D. N. Gopaul and M. Jackson, *J. Bacteriol.*, 2008, **190**, 1329–1334.
19 J. Gonzalo-Asensio, S. Mostowy, J. Harders-Westerveen, K. Huygen, R. Hernandez-Pando, J. Thole, M. Behr, B. Gicquel and C. Martin, *PLoS One*, 2008, **3**, e3496.
20 L. G. Pacheco, R. R. Pena, T. L. Castro, F. A. Dorella, R. C. Bahia, R. Carminati, M. N. Frota, S. C. Oliveira, R. Meyer, F. S. Alves, A. Miyoshi and V. Azevedo, *J. Med. Microbiol.*, 2007, **56**, 480–486.
21 V. S. Kalogeraki and S. C. Winans, *Gene*, 1997, **188**, 69–75.
22 F. A. Dorella, E. M. Estevam, P. G. Cardoso, B. M. Savassi, S. C. Oliveira, V. Azevedo and A. Miyoshi, *Vet. Microbiol.*, 2006, **114**, 298–303.
23 P. M. Moraes, N. Seyffert, W. M. Silva, T. L. Castro, R. F. Silva, D. D. Lima, R. Hirata Jr., A. Silva, A. Miyoshi and V. Azevedo, *BioMed. Res. Int.*, 2014, **2014**, 489782.
24 D. Ribeiro, S. Rocha Fde, K. M. Leite, C. Soares Sde, A. Silva, R. W. Portela, R. Meyer, A. Miyoshi, S. C. Oliveira, V. Azevedo and F. A. Dorella, *Vet. Res.*, 2014, **45**, 28.
25 S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, *J. Mol. Biol.*, 1990, **215**, 403–410.
26 G. Moraes, V. Azevedo, M. Costa, A. Miyoshi, A. Silva, V. da Silva, D. de Oliveira, M. F. Teixeira, J. Lameira and C. N. Alves, *J. Mol. Model.*, 2012, **18**, 1219–1227.
27 K. Arnold, L. Bordoli, J. Kopp and T. Schwede, *Bioinformatics*, 2006, **22**, 195–201.
28 R. A. Laskowski, M. W. Macarthur, D. S. Moss and J. M. Thornton, *J. Appl. Crystallogr.*, 1993, **26**, 283–291.
29 F. Melo and E. Feytmans, *J. Mol. Biol.*, 1998, **277**, 1141–1152.
30 E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin, *J. Comput. Chem.*, 2004, **25**, 1605–1612.
31 J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad and R. G. Coleman, *J. Chem. Inf. Model.*, 2012, **52**, 1757–1768.
32 R. Thomsen and M. H. Christensen, *J. Med. Chem.*, 2006, **49**, 3315–3321.
33 N. L. Ferrer, A. B. Gomez, O. Neyrolles, B. Gicquel and C. Martin, *PLoS One*, 2010, **5**, e12978.
34 E. A. Groisman, *J. Bacteriol.*, 2001, **183**, 1835–1842.
35 A. L. Hodgson, K. Carter, M. Tachedjian, J. Krywult, L. A. Corner, M. McColl and A. Cameron, *Vaccine*, 1999, **17**, 802–808.
36 Y. Li, M. Jiang, W. Liu, L. Zhang, S. Zhang, X. Zhao, R. Xiang and Y. Liu, *Mol. Cell. Probes*, 2010, **24**, 68–71.
37 S. Gupta, A. Sinha and D. Sarkar, *FEBS Lett.*, 2006, **580**, 5328–5338.
38 J. F. Barrett, R. M. Goldschmidt, L. E. Lawrence, B. Foleno, R. Chen, J. P. Demers, S. Johnson, R. Kanojia, J. Fernandez, J. Bernstein, L. Licata, A. Donetz, S. Huang, D. J. Hlasta, M. J. Macielag, K. Ohemeng, R. Frechette, M. B. Frosco, D. H. Klaubert, J. M. Whiteley, L. Wang and J. A. Hoch, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 5317–5322.
39 R. Utsumi and M. Igarashi, *Yakugaku zasshi*, 2012, **132**, 51–58.
40 P. Benkert, M. Kunzli and T. Schwede, *Nucleic Acids Res.*, 2009, **37**, W510–W514.
41 E. J. Capra and M. T. Laub, *Annu. Rev. Microbiol.*, 2012, **66**, 325–347.
42 S. Menon and S. Wang, *Biochemistry*, 2011, **50**, 5948–5957.
43 L. Hoerhammer, H. Wagner and I. Koehler, *Arch. Pharm. Ber. Dtsch. Pharm. Ges.*, 1959, **292/64**, 591–601.
44 M. Anchel, *J. Biol. Chem.*, 1949, **177**, 169–177.
45 K. Stierand and M. Rarey, *ACS Med. Chem. Lett.*, 2010, **1**, 540–545.
46 H. S. Garmory, K. A. Brown and R. W. Titball, *FEMS Microbiol. Rev.*, 2002, **26**, 339–353.
47 J. Gonzalo Asensio, C. Maia, N. L. Ferrer, N. Barilone, F. Laval, C. Y. Soto, N. Winter, M. Daffe, B. Gicquel, C. Martin and M. Jackson, *J. Biol. Chem.*, 2006, **281**, 1313–1316.
48 C. Martin, A. Williams, R. Hernandez-Pando, P. J. Cardona, E. Gormley, Y. Bordat, C. Y. Soto, S. O. Clark, G. J. Hatch, D. Aguilar, V. Ausina and B. Gicquel, *Vaccine*, 2006, **24**, 3408–3419.
49 M. V. Mendes, S. Tunca, N. Anton, E. Recio, A. Sola-Landa, J. F. Aparicio and J. F. Martin, *Metab. Eng.*, 2007, **9**, 217–227.
50 S. B. Walters, E. Dubnau, I. Kolesnikova, F. Laval, M. Daffe and I. Smith, *Mol. Microbiol.*, 2006, **60**, 312–330.
51 J. Sambrook, E. F. Fritsch and T. Maniatis, *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA, 2nd edn, 1989.
52 F. A. Dorella, L. G. Pacheco, N. Seyffert, R. W. Portela, R. Meyer, A. Miyoshi and V. Azevedo, *Expert Rev. Vaccines*, 2009, **8**, 205–213.

II - Outros resultados

Aqui, serão apresentados cinco trabalhos publicados na forma de artigo cujo resultados não possuem relação direta com redes de interação proteína-proteína, mas que foram desenvolvidos durante o período de doutorado. Estas atividades, por serem diferentes do tema principal desenvolvido na tese, complementam o conhecimento na área de Bioinformática, sendo estes momentos de colaboração uma grande oportunidade para novos aprendizados.

Se tratando de montagem, anotação e curadoria de genomas, este aprendizado é extrapolado ainda mais, pois, além das técnicas e ferramentas usadas no processo de montagem e anotação, a atividade de curadoria, apesar de ser uma tarefa “manual” e trabalhosa, conduz a uma reflexão biológica sobre o organismo, viabilizando conhecer melhor os genes, proteínas e sua organização. Apesar de pouco valorizada cientificamente, o trabalho de montagem, anotação e curadoria de genomas é extremamente relevante, pois, é a base para o desenvolvimento de futuros trabalhos científicos, inclusive para predições *in silico* de interação proteína-proteína, como desenvolvido nesta tese.

Adicionalmente à curadoria manual de genoma mas ainda relacionados a esta atividade, foram desenvolvidos dois *scripts* na linguagem de programação Perl com as seguintes finalidades: (i) corrigir a posição de *start* e *stop* códon dos elementos estruturais após curadoria de genomas fragmentado e distribuído para vários curadores, situação que ocorre principalmente após correções de *frame-shifts* gerados por regiões de homopolímeros, quando as coordenadas dos elementos estruturais do genoma se alteram, consequentemente modificando as coordenadas subsequente do genoma curado por outro pesquisador, necessitando ser corrigida e; (ii) transferir automaticamente a anotação de um genoma já curado para outro genoma em processo de anotação. Estes scripts não foram desenvolvidos com intuito de gerar publicação, mas sim de serem utilizados pelo grupo para agilizar o processo de anotação automática e curadoria de genomas, dentre os quais, alguns dos quais eu tive oportunidade de participar.

A seguir estão relacionados quatro artigos científicos publicados nos quais colaborei principalmente nas etapas de anotação funcional e curadoria de genoma. No quinto artigo publicado, as atividades de colaboração se resumem principalmente na execução de programas de bioinformática e análises dos resultados retornados.

genomeA
Journals.ASM.org

Genome Sequence of *Lactococcus lactis* subsp. *lactis* NCDO 2118, a GABA-Producing Strain

Letícia C. Oliveira,[a] Tessália D. L. Saraiva,[a] Siomar C. Soares,[a]* Rommel T. J. Ramos,[b] Pablo H. C. G. Sá,[b] Adriana R. Carneiro,[a] Fábio Miranda,[b] Matheus Freire,[b] Wendel Renan,[b] Alberto F. O. Júnior,[a] Anderson R. Santos,[a]* Anne C. Pinto,[a] Bianca M. Souza,[a] Camila P. Castro,[a] Carlos A. A. Diniz,[a] Clarissa S. Rocha,[a] Diego C. B. Mariano,[a] Edgar L. de Aguiar,[a] Edson L. Folador,[a] Eudes G. V. Barbosa,[a] Flavia F. Aburjaile,[a] Lucas A. Gonçalves,[a] Luís C. Guimarães,[a] Marcela Azevedo,[a] Pamela C. M. Agresti,[a] Renata F. Silva,[a] Sandeep Tiwari,[a] Sintia S. Almeida,[a] Syed S. Hassan,[a] Vanessa B. Pereira,[a] Vinicius A. C. Abreu,[a] Ulisses P. Pereira,[a]* Fernanda A. Dorella,[c] Alex F. Carvalho,[c] Felipe L. Pereira,[c] Carlos A. G. Leal,[c] Henrique C. P. Figueiredo,[c] Artur Silva,[b] Anderson Miyoshi,[a] Vasco Azevedo[a]

Laboratory of Cellular and Molecular Genetics, Institute of Biologic Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil[a]; Institute of Biologic Sciences, Federal University of Pará, Belém, PA, Brazil[b]; AQUACEN, National Reference Laboratory for Aquatic Animal Diseases, Ministry of Fisheries and Aquaculture, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil[c]

* Present address: Siomar C. Soares, AQUACEN, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil; Anderson R. Santos, Federal University of Uberlândia, Minas Gerais, MG, Brazil; Ulisses P. Pereira, Federal University of Uberlândia, Minas Gerais, MG, Brazil.

***Lactococcus lactis* subsp. *lactis* NCDO 2118 is a nondairy lactic acid bacterium, a xylose fermenter, and a gamma-aminobutyric acid (GABA) producer isolated from frozen peas. Here, we report the complete genome sequence of *L. lactis* NCDO 2118, a strain with probiotic potential activity.**

Received 21 August 2014 **Accepted** 26 August 2014 **Published** 2 October 2014

Citation Oliveira LC, Saraiva TDL, Soares SC, Ramos RTJ, Sá PHCG, Carneiro AR, Miranda F, Freire M, Renan W, Júnior AFO, Santos AR, Pinto AC, Souza BM, Castro CP, Diniz CAA, Rocha CS, Mariano DCB, de Aguiar EL, Folador EL, Barbosa EGV, Aburjaile FF, Gonçalves LA, Guimarães LC, Azevedo M, Agresti PCM, Silva RF, Tiwari S, Almeida SS, Hassan SS, Pereira VB, Abreu VAC, Pereira UP, Dorella FA, Carvalho AF, Pereira FL, Leal CAG, Figueiredo HCP, Silva A, Miyoshi A, Azevedo V. 2014. Genome sequence of *Lactococcus lactis* subsp. *lactis* NCDO 2118, a GABA-producing strain. Genome Announc. 2(5):e00980-14. doi:10.1128/genomeA.00980-14.

Address correspondence to Vasco Azevedo, vasco@icb.ufmg.br.

Lactic acid bacteria (LAB), in general, acquire energy from the conversion of sugars into lactic acid (1) and are used for production of many fermented products, such as cheese, yogurt, butter, and wine. Food conservation is due to the medium acidification and production of molecules that inhibit the growth of undesirable microbiota, contributing to the development of desirable organoleptic properties in the final product (2). Moreover, some specific LAB strains produce bioactive molecules such as gamma-aminobutyric acid (GABA) (3), a product of glutamate decarboxylation by the glutamic acid decarboxylase (GAD) enzyme. Usually, GABA acts by modulating the central nervous system, contributing to smooth muscle relaxation and presenting hypotensor activity (4). Also, GABA can immunomodulate the immune system (5). Therefore, GABA-producing bacteria generally present probiotic properties (6). *Lactococcus lactis* NCDO 2118 is a nondairy strain, a xylose fermenter (a common trait of plant-associated strains), and a GABA producer isolated from frozen peas (6, 7).

L. lactis NCDO 2118 was sequenced three times, due to assembling complexity. First, the genome was decoded with the SOLiD 5500 platform with mate-paired libraries, generating a total of 5,133,057,360 bp, (coverage of 2,053 times). The reads were subjected to a Phred 20 quality filter using Quality Assessment software (8) and assembled with the CLC Genomics Workbench, generating a total of 1,641 overlapping sequences. These sequences were removed with the Simplifier (9), ordered and oriented based on the reference *L. lactis* KF147 genome sequence (a plant-associated strain, accession number CP001834). Then manual curation was performed using Artemis (10), and SSPACE (11) and Gapfiller (12) were used to generate the scaffold and resolve gaps, respectively. At the end of curation and sequence assembly, a total of 409 scaffolds (2,874,854 bp) were obtained.

L. lactis NCDO 2118 was then decoded with the Ion PGM platform with fragment libraries generating a total of 187,303,001 bp (coverage of ~71 times). Genome assembly was performed using Mira 3.9 (13), and the assembled genome sequence was reference aligned with CONTIGuator (14). The redundant overlapping sequences were removed with "in-house scripts," closing the remnant gaps. Annotation and frameshifts curation were then performed using Artemis and CLC, reducing the initial 1821 pseudogenes to 480.

Finally, the DNA was sequenced using the Ion Torrent PGM with fragment libraries, yielding a total of ~1,249,154,478 bp (coverage of 474 times). Assembly was performed with Mira 4.0.1 and Newbler 2.9 (15). We used CONTIGuator and FGAP 1.7 (16) to perform the alignment and gap closure steps, respectively. We followed the same previously explained steps for annotation and frameshift curation, reducing the pseudogenes to 52.

The complete genome of *L. lactis* NCDO 2118 consists of a single circular chromosome of 2,554,693 bp, containing 2,386 coding sequences (CDS), which had 52 pseudogenes, 66 tRNA genes, and 6 rRNA operons, with a G+C content of 34.9%. There is one plasmid, pNCDO2118 (37,571 bp), with 48 CDS, from which 4 are pseudogenes with a G+C content of 32.33%.

Nucleotide sequence accession numbers. The *Lactococcus lactis* NCDO 2118 chromosome and the plasmid were deposited at DDBJ/EMBL/GenBank under the accession numbers CP009054 and CP009055, respectively.

ACKNOWLEDGMENTS

This work was supported by Rede Paraense de Genômica e Proteômica, Ministério da Pesca e Aquicultura (MPA), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG). We also acknowledge the support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

REFERENCES

1. **Carr FJ, Chill D, Maida N.** 2002. The lactic acid bacteria: a literature survey. Crit. Rev. Microbiol. **28:**281–370. http://dx.doi.org/10.1080/1040-840291046759.
2. **van de Guchte M, Ehrlich SD, Maguin E.** 2001. Production of growth-inhibiting factors by *Lactobacillus delbrueckii*. J. Appl. Microbiol. **91:**147–153. http://dx.doi.org/10.1046/j.1365-2672.2001.01369.x.
3. **Zareian M, Ebrahimpour A, Bakar FA, Mohamed AK, Forghani B, Ab-Kadir MS, Saari N.** 2012. A glutamic acid-producing lactic acid bacteria isolated from Malaysian fermented foods. Int. J. Mol. Sci. **13:**5482–5497. http://dx.doi.org/10.3390/ijms13055482.
4. **Inoue K, Shirai T, Ochiai H, Kasao M, Hayakawa K, Kimura M, Sansawa H.** 2003. Blood-pressure-lowering effect of a novel fermented milk containing gamma-aminobutyric acid (GABA) in mild hypertensives. Eur. J. Clin. Nutr. **57:**490–495. http://dx.doi.org/10.1038/ sj.ejcn.1601555.
5. **Jin Z, Mendu SK, Birnir B.** 2013. GABA is an effective immunomodulatory molecule. Amino Acids **45:**87–94. http://dx.doi.org/10.1007/s00726-011-1193-7.
6. **Mazzoli R, Pessione E, Dufour M, Laroute V, Giuffrida MG, Giunta C, Cocaign-Bousquet M, Loubière P.** 2010. Glutamate-induced metabolic changes in *Lactococcus lactis* NCDO 2118 during GABA production: combined transcriptomic and proteomic analysis. Amino Acids. **39:**727–737. http://dx.doi.org/10.1007/s00726-010-0507-5.
7. **Siezen RJ, Starrenburg MJ, Boekhorst J, Renckens B, Molenaar D, van Hylckama Vlieg JE.** 2008. Genome-scale genotype-phenotype matching of two *Lactococcus lactis* isolates from plants identifies mechanisms of adaptation to the plant niche. Appl. Environ. Microbiol. **74:**424–436. http://dx.doi.org/10.1128/AEM.01850-07.
8. **Ramos RT, Carneiro AR, Baumbach J, Azevedo V, Schneider MP, Silva A.** 2011. Analysis of quality raw data of second generation sequencers with Quality Assessment Software. BMC Res. Notes **4:**130. http://dx.doi.org/10.1186/1756-0500-4-130.
9. **Ramos RT, Carneiro AR, Azevedo V, Schneider MP, Barh D, Silva A.** 2012. Simplifier: a web tool to eliminate redundant NGS contigs. Bioinformation **8:**996–999. http://dx.doi.org/10.6026/97320630008996.
10. **Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944–945. http://dx.doi.org/10.1093/bioinformatics/16.10.944.
11. **Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W.** 2011. Scaffolding preassembled contigs using SSPACE. Bioinformatics **27:**578–579. http://dx.doi.org/10.1093/bioinformatics/btq683.
12. **Nadalin F, Vezzi F, Policriti A.** 2012. GapFiller: a *de novo* assembly approach to fill the gap within paired reads. BMC Bioinformatics **13:**S8. http://dx.doi.org/10.1186/1471-2105-13-S14-S8.
13. **Chevreux B, Wetter T, Suhai S.** 1999. Genome sequence assembly using trace signals and additional sequence information, p 45–56. Comput. Sci. Biol.: Proc. German Conference on Bioinformatics GCB'99 GCB. Hannover, Germany.
14. **Galardini M, Biondi EG, Bazzicalupo M, Mengoni A.** 2011. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. Source Code Biol. Med. **6:**11. http://dx.doi.org/10.1186/1751-0473-6-11.
15. **Ohnishi N, Maruyama F, Ogawa H, Kachi H, Yamada S, Fujikura D, Nakagawa I, Hang'ombe MB, Thomas Y, Mweene AS, Higashi H.** 2014. Genome sequence of a *Bacillus anthracis* outbreak strain from Zambia, 2011. Genome Announc. **2**(2):e00116-14. http://dx.doi.org/10.1128/genomeA.00116-14.
16. **Piro VC, Faoro H, Weiss VA, Steffens MB, Pedrosa FO, Souza EM, Raittz RT.** 2014. FGAP: an automated gap closing tool. BMC Res. Notes **7:**371. http://dx.doi.org/10.1186/1756-0500-7-371.

genomeA
Journals.ASM.org

Genome Sequence of *Corynebacterium pseudotuberculosis* MB20 bv. equi Isolated from a Pectoral Abscess of an Oldenburg Horse in California

Rafael A. Baraúna,[a] Luís C. Guimarães,[b] Adonney A. O. Veras,[a] Pablo H. C. G. de Sá,[a] Diego A. Graças,[a] Kenny C. Pinheiro,[a] Andreia S. S. Silva,[a] Edson L. Folador,[b] Leandro J. Benevides,[b] Marcus V. C. Viana,[b] Adriana R. Carneiro,[a] Maria P. C. Schneider,[a] Sharon J. Spier,[c] Judy M. Edman,[c] Rommel T. J. Ramos,[a] Vasco Azevedo,[b] Artur Silva[a]

Institute of Biological Sciences, Federal University of Pará, Belém, PA, Brazil[a]; Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil[b]; Department of Medicine and Epidemiology, School of Veterinary Medicine, University of California Davis, California, USA[c]

The genome of *Corynebacterium pseudotuberculosis* MB20 bv. equi was sequenced using the Ion Personal Genome Machine (PGM) platform, and showed a size of 2,363,089 bp, with 2,365 coding sequences and a GC content of 52.1%. These results will serve as a basis for further studies on the pathogenicity of *C. pseudotuberculosis* bv. equi.

Received 28 August 2014 **Accepted** 8 October 2014 **Published** 13 November 2014

Citation Baraúna RA, Guimarães LC, Veras AAO, de Sá PHCG, Graças DA, Pinheiro KC, Silva ASS, Folador EL, Benevides LJ, Viana MVC, Carneiro AR, Schneider MPC, Spier SJ, Edman JM, Ramos RTJ, Azevedo V, Silva A. 2014. Genome sequence of *Corynebacterium pseudotuberculosis* MB20 bv. equi isolated from a pectoral abscess of an Oldenburg horse in California. Genome Announc. 2(6):e00977-14. doi:10.1128/genomeA.00977-14.

Address correspondence to Rafael A. Baraúna, rabarauna@ufpa.br.

Recent advances have been made in the genomic analysis of *Corynebacterium pseudotuberculosis*, a species of veterinary and biotechnological interest. Molecular diagnosis of several diseases caused by this species was achieved using multiplex PCR (PCR) (1), and a description of its pangenome was published using 15 strains of *C. pseudotuberculosis* bv. ovis and equi (2). *C. pseudotuberculosis* bv. equi comprises strains that infect horses and cattle and cause infection of cutaneous lymphatic vessels, termed ulcerative lymphangitis, which is characterized by the development of multiple waxy ulcerative lesions. The incidence of this disease in horses has been reported in the literature since the 1910s (3) and remains prevalent in animals worldwide (4, 5); moreover, this infection is likely underreported and has been characterized as a neglected zoonosis (6).

The host-pathogen interaction in this disease has been studied using omics approaches (7). For instance, the differential gene expression of a *C. pseudotuberculosis* bv. ovis strain was analyzed using RNA-seq, and genes involved in the molecular responses of the bacterium to different stresses during infection were identified (8). A study of reverse vaccinology reported by Soares et al. (9) identified 49 possible antigens from the genome of the *C. pseudotuberculosis* bv. equi strain 258 that may serve as targets for the development of effective vaccines. In addition, a new *C. pseudotuberculosis* bv. equi strain was isolated and sequenced, which will aid in future broader studies. These new data combined with those already reported will serve as a basis for the development of studies aimed at a better understanding of the pathogenic potential of *C. pseudotuberculosis* bv. equi.

The MB20 strain was isolated from a pectoral abscess of a 4-year-old horse of the breed Oldenburg, raised in the city of Vacaville, CA, USA. Genomic DNA was sequenced from a fragment library on a 318 chip of the Ion Torrent Personal Genome Machine (PGM) platform (Life Technologies). A total of 2,331,864 reads were generated with an average length of 420 bp, which were used for genome assembly using the software Mira (10). The contigs generated with Mira were analyzed using the SeqMan Pro tool of the software Lasergene 11 Core Suite (DNASTAR) to remove redundant sequences. This approach resulted in 3 contigs, which were sorted with the Artemis Comparison tool (11) using the genome of *C. pseudotuberculosis* 316 as a reference. The scaffold produced at the end of the assembly was 2,363,089 bp in size and underwent automatic annotation using Rapid Annotation using Subsystem Technology (RAST) (12). As a result, 2,365 coding sequences (CDSs), 11 rRNA genes, 51 tRNA genes and a 52.1% GC content were identified. Of the 2,365 CDSs, 790 (33.4%) were classified as hypothetical proteins.

Nucleotide sequence accession numbers. The genomic sequence obtained in this study was deposited in the DDBJ/EMBL/GenBank under accession number JPUV00000000. The version described in this paper is version JPUV01000000.

ACKNOWLEDGMENTS

This study was conducted by the Rede Paraense de Genômica e Proteômica, with support from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and the Fundação de Amparo a Pesquisa do Estado do Pará (FAPESPA).

REFERENCES

1. **Pacheco LGC, Pena RR, Castro TLP, Dorella FA, Bahia RC, Carminati R, Frota MNL, Oliveira SC, Meyer R, Alves FSF, Miyoshi A, Azevedo V.** 2007. Multiplex PCR assay for the identification of *Corynebacterium pseudotuberculosis* from pure cultures and for rapid detection of this pathogen in clinical samples. J. Med. Microbiol. **56**:480–486. http://dx.doi.org/10.1099/jmm.0.46997-0.
2. **Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, Ali A, Santos AR, Pinto AC, Diniz C, Barbosa EG, Dorella FA, Aburjaile F, Rocha FS, Nascimento KK, Guimarães LC, Almeida S, Hassan SS, Bakhtiar SM,**

cci

Pereira UP, Abreu VA, Schneider MP, Miyoshi A, Tauch A, Azevedo V. 2013. The pan-genome of the animal pathogen *Corynebacterium pseudotuberculosis* reveals differences in genome plasticity between the biovar ovis and equi strains. PLoS One **8:**e53818. http://dx.doi.org/10.1371/ journal.pone.0053818.

3. **Hall IC, Stone RV.** 1916. The diphtheroid bacillus of Preisz-Nocard from equine, bovine, and ovine abscesses: ulcerative lymphangitis and caseous lymphadenitis. J. Infect. Dis. **18:**195–208. http://dx.doi.org/10.1093/infdis/18.2.195.
4. **Hepworth-Warren KL, Sponseller BT, Wong DM, Kinyon JM.** 2014. Isolation of *Corynebacterium pseudotuberculosis* biovar equi from a horse in central Iowa. Case Rep. Vet. Med. **2014:**1–3. http://dx.doi.org/10.1155/2014/436287.
5. **Spier SJ, Leutenegger CM, Carroll SP, Loye JE, Pusterla JB, Carpenter TE, Mihalyi JE, Madigan JE.** 2004. Use of a real-time polymerase chain reaction-based fluorogenic 5′ nuclease assay to evaluate insect vectors of *Corynebacterium pseudotuberculosis* infections in horses. Am. J. Vet. Res. **65:**829–834. http://dx.doi.org/10.2460/ajvr.2004.65.829.
6 **Join-Lambert OF, Ouache M, Canioni D, Beretti JL, Blanche S, Berche P, Kayal S.** 2006. *Corynebacterium pseudotuberculosis* necrotizing lymphadenitis in a twelve-year-old patient. Pediatr. Infect. Dis. J. **25:**848–851.
7. **Dorella FA, Gala-Garcia A, Pinto AC, Sarrouh B, Antunes CA, Ribeiro D, Aburjaile FF, Fiaux KK, Guimarães LC, Seyffert N, El-Aouar R, Silva R, Hassan SS, Castro TLP, Marques WS, Ramos R, Carneiro A, Sá P, Miyoshi A, Azevedo V, Silva A.** 2013. Progression of "OMICS" methodologies for understanding the pathogenicity of *Corynebacterium pseudotuberculosis*: the Brazilian experience. Comput. Struct. Biotechnol. J. **6** e201303013. http://dx.doi.org/10.5936/csbj.201303013.
8. **Pinto AC, Sá PH, Ramos RT, Barbosa S, Barbosa HP, Ribeiro AC, Silva WM, Rocha FS, Santana MP, Castro TL, Miyoshi A, Schneider MP, Silva A, Azevedo V.** 2014. Differential transcriptional profile of *Corynebacterium pseudotuberculosis* in response to abiotic stresses. BMC Genomics **15:**14. http://dx.doi.org/10.1186/1471-2164-15-14.
9. **Soares SC, Trost E, Ramos RT, Carneiro AR, Santos AR, Pinto AC, Barbosa E, Aburjaile F, Ali A, Diniz CA, Hassan SS, Fiaux K, Guimarães LC, Bakhtiar SM, Pereira U, Almeida SS, Abreu VA, Rocha FS, Dorella FA, Miyoshi A, Silva A, Azevedo V, Tauch A.** 2013. Genome sequence of *Corynebacterium pseudotuberculosis* biovar equi strain 258 and prediction of antigenic targets to improve biotechnological vaccine production. J. Biotechnol. **167:**135–141. http://dx.doi.org/10.1016/j.jbiotec.2012.11.003.
10. **Chevreux B, Pfisterer, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S.** 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. **14:**1147–1159. http://dx.doi.org/10.1101/gr.1917404.
11. **Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J.** 2005. ACT: the Artemis comparison Tool. Bioinformatics **21:**3422–3423. http://dx.doi.org/10.1093/bioinformatics/bti553.
12. **Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B, Shukla M, Vonstein V, Wattam AR, Xia F, Stevens R.** 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. **42:**D206–D214. http://dx.doi.org/10.1093/nar/gkt1226.

genomeA
Journals.ASM.org

Genome Sequence of *Corynebacterium ulcerans* Strain 210932

Marcus Vinicius Canário Viana,[a] Leandro de Jesus Benevides,[a] Diego Cesar Batista Mariano,[a] Flávia de Souza Rocha,[a] Priscilla Carolinne Bagano Vilas Boas,[a] Edson Luiz Folador,[a] Felipe Luiz Pereira,[b] Fernanda Alves Dorella,[b] Carlos Augusto Gomes Leal,[b] Alex Fiorini de Carvalho,[b] Artur Silva,[c] Siomar de Castro Soares,[b] Henrique Cesar Pereira Figueiredo,[b] Vasco Azevedo,[a] Luis Carlos Guimarães[a]

Department of General Biology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil[a]; AQUACEN, National Reference Laboratory for Aquatic Animal Diseases, Ministry of Fisheries and Aquaculture, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil[b]; Department of Genetics, Federal University of Pará, Belém, Pará, Brazil[c]

In this work, we present the complete genome sequence of *Corynebacterium ulcerans* strain 210932, isolated from a human. The species is an emergent pathogen that infects a variety of wild and domesticated animals and humans. It is associated with a growing number of cases of a diphtheria-like disease around the world.

Received 15 October 2014 **Accepted** 21 October 2014 **Published** 26 November 2014

Citation Viana MVC, de Jesus Benevides L, Batista Mariano DC, de Souza Rocha F, Bagano Vilas Boas PC, Folador EL, Pereira FL, Alves Dorella F, Gomes Leal CA, Fiorini de Carvalho A, Silva A, de Castro Soares S, Pereira Figueiredo HC, Azevedo V, Guimarães LC. 2014. Genome sequence of *Corynebacterium ulcerans* strain 210932. Genome Announc. 2(6):e01233-14. doi:10.1128/genomeA.01233-14.

Address correspondence to Vasco Azevedo, vasco@icb.ufmg.br.

Corynebacterium ulcerans is a toxogenic zoonotic agent and Gram-positive bacterium that belongs to the *Actinobacteria* class, which includes the genera *Corynebacterium*, *Mycobacterium*, *Nocardia*, and *Rodococcus* and is referred to as a CMNR group. Studies using the 16S rRNA gene showed that *Corynebacterium pseudotuberculosis* and *Corynebacterium diphtheriae* are closely related to *C. ulcerans*. The species is facultative anaerobic, non–spore forming, nonmotile, catalase positive, and nitrate and oxidase negative. It differs from other species of the genus by fermentation of glycogen and starch (1).

The species can infect a variety of wild and domesticated animals and humans (2). It causes bovine mastitis and other infections in cats, dogs, monkeys, squirrels, otters, orcas, camels, lions, pigs, and goats. In humans, it causes diphtheria-like disease, pharyngitis, sinusitis, tonsillitis, pulmonary nodules, and skin ulcers (3). Contaminations in humans have been associated with raw milk and derivatives and contact with cattle and infected domestic pets (4). *C. ulcerans* is considered an emergent pathogen because the number of cases of infection in humans has been constantly increasing in the last two decades in the United States, Brazil, Western Europe, and Japan (5).

This species has a varied set of virulence factors, including *diphtheriae*-like toxin, phospholipase D, neuraminidase H, endoglycosidase EndoE, and a novel type of ribosome-binding protein with structural similarity to Shiga-like toxins. The sequencing of more *C. ulcerans* genomes, both toxigenic and non-toxigenic, will help in the identification of distinctive features of strains from human and animal sources, as well as in describing the zoonotic transmission in more detail (6). In addition, the data generated by newly sequenced genomes is helpful in identifying antibiotic and vaccine targets by comparative analysis (7). To date, only three complete genomes of *C. ulcerans* and two drafts have been deposited in the NCBI database.

Herein, we present the complete genome sequence of *Corynebacterium ulcerans* strain 210932, isolated from a human. Its genome sequencing was performed by the Ion Personal Genome Machine (PGM) System, using a fragment library. A total of 1,606,464 genomic reads were filtered by quality using the software FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), and *de novo* assembling was done using Mira software version 3.9.18. The assembling step generated 12 contigs with a mean coverage of 129.23× and an N_{50} of 487,508. The contigs were scaffolded using the *C. ulcerans* strain 0102 as reference. The gaps were closed using CONTIGuator software (http://contiguator.sourceforge.net/) via the web tool SIMBA (SImple Manager for Bacterial Assemblies) (http://lgcm.icb.ufmg.br/simba/). CLC Workbench version 7 was used for manual curation of homopolymers, generating a final assembled genome with 2,484,335 bp.

An automatic annotation was done by RAST (http://rast.nmpdr.org/), followed by manual curation using Artemis software (http://www.sanger.ac.uk/resources/software/artemis/) and the Uniprot database (http://www.uniprot.org/). The genome has 2,282 coding sequences (from which 654, or 28.65%, were annotated as "hypothetical proteins"), 12 rRNAs, 51 tRNAs, and a G+C content of 53.32%.

Nucleotide sequence accession number. This whole-genome shotgun project has been deposited in GenBank under the accession number CP009500.

ACKNOWLEDGMENTS

This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Ministério da Pesca e Agricultura.

REFERENCES

1. **Riegel P, Ruimy R, de Briel D, Prévost G, Jehl F, Christen R, Monteil H.** 1995. Taxonomy of *Corynebacterium diphtheriae* and related taxa, with rec-

ognition of *Corynebacterium ulcerans* sp. nov. nom. rev. FEM Microbiol. Lett. **126:**171–176.
2. **Tiwari TS, Golaz A, Yu DT, Ehresmann KR, Jones TF, Hill HE, Cassiday PK, Pawloski LC, Moran JS, Popovic T, Wharton M.** 2008. Investigations of 2 cases of diphtheria-like illness due to toxigenic *Corynebacterium ulcerans*. Clin. Infect. Dis. **46:**395–401. http://dx.doi.org/10.1086/525262.
3. **Bernard K.** 2012. The genus *Corynebacterium* and other medically relevant coryneform-like bacteria. J. Clin. Microbiol. **50:**3152–3158. http://dx.doi.org/10.1128/JCM.00796-12.
4. **Wagner KS, White JM, Crowcroft NS, De Martin S, Mann G, Efstratiou A.** 2010. Diphtheria in the United Kingdom, 1986–2008: the increasing role of *Corynebacterium ulcerans*. Epidemiol. Infect. **138:**1519–1530. http://dx.doi.org/10.1017/S0950268810001895.
5. **Dias AA, Santos LS, Sabbadini PS, Santos CS, Silva Júnior FC, Napoleão F, Nagao PE, Villas-Bôas MH, Hirata Júnior R, Guaraldi AL.** 2011. *Corynebacterium ulcerans* diphtheria: an emerging zoonosis in Brazil and worldwide. Rev. Saude Publica **45:**1176–1191. http://dx.doi.org/10.1590/s0034-89102011000600021.
6. **Trost E, Al-Dilaimi A, Papavasiliou P, Schneider J, Viehoever P, Burkovski A, Soares SC, Almeida SS, Dorella FA, Miyoshi A, Azevedo V, Schneider MP, Silva A, Santos CS, Santos LS, Sabbadini P, Dias AA, Hirata R, Jr, Mattos-Guaraldi AL, Tauch A.** 2011. Comparative analysis of two complete *Corynebacterium ulcerans* genomes and detection of candidate virulence factors. BMC Genomics **12:**383. http://dx.doi.org/10.1186/1471-2164-12-383.
7. **Barbosa EG, Aburjaile FF, Ramos RT, Carneiro AR, Le Loir Y, Baumbach J, Miyoshi A, Silva A, Azevedo V.** 2014. Value of a newly sequenced bacterial genome. World J Biol Chem. **5:**161–168. http://www.wjgnet.com/1949-8454/full/v5/i2/161.htm.

genomeA
Journals.ASM.org

Genome Sequence of *Corynebacterium ulcerans* Strain FRC11

Leandro de Jesus Benevides,[a] Marcus Vinicius Canário Viana,[a] Diego César Batista Mariano,[a] Flávia de Souza Rocha,[a] Priscilla Carolinne Bagano,[a] Edson Luiz Folador,[a] Felipe Luiz Pereira,[b] Fernanda Alves Dorella,[b] Carlos Augusto Gomes Leal,[b] Alex Fiorini Carvalho,[b] Siomar de Castro Soares,[b] Adriana Carneiro,[c] Rommel Ramos,[c] Edgar Badell-Ocando,[d] Nicole Guiso,[d] Artur Silva,[c] Henrique Figueiredo,[b] Vasco Azevedo,[a] Luis Carlos Guimarães[a]*

Laboratory of Cellular and Molecular Genetics (LGCM), Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil[a]; National Reference Laboratory for Aquatic Animal Diseases (AQUACEN), Ministry of Fisheries and Aquaculture, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil[b]; Laboratory of Polymorphic DNA (LPDNA), Federal University of Para, Belém, Brazil[c]; Institut Pasteur, Unité de Prévention et Thérapies Moléculaires des Maladies Humaines, National Centre of Reference of Toxigenic Corynebacteria, Paris, France[d]

* Present address: Luis Carlos Guimarães, Departamento de Biologia Geral, ICB/UFMG, Pampulha, Belo Horizonte, Minas Gerais, Brazil.

Here, we present the genome sequence of *Corynebacterium ulcerans* strain FRC11. The genome includes one circular chromosome of 2,442,826 bp (53.35% G+C content), and 2,210 genes were predicted, 2,146 of which are putative protein-coding genes, with 12 rRNAs and 51 tRNAs; 1 pseudogene was also identified.

Received 30 January 2015 **Accepted** 4 February 2015 **Published** 12 March 2015

Citation Benevides LDJ, Viana MVC, Mariano DCB, Rocha FDS, Bagano PC, Folador EL, Pereira FL, Dorella FA, Leal CAG, Carvalho AF, Soares SDC, Carneiro A, Ramos R, Badell-Ocando E, Guiso N, Silva A, Figueiredo H, Azevedo V, Guimarães LC. 2015. Genome sequence of *Corynebacterium ulcerans* strain FRC11. Genome Announc 3(2):e00112-15. doi: 10.1128/genomeA.00112-15.

Address correspondence to Vasco Azevedo, vasco@icb.ufmg.br.

Corynebacterium ulcerans is a bacterium that presents catalase-positive, nitrate-negative, and urease-positive biochemical proprieties (1). This bacterium belongs to the *Actinobacteria* class, which includes the genera *Corynebacterium*, *Mycobacterium*, *Nocardia*, and *Rodococcus*, collectively termed the CMNR group. This is a very heterogeneous group; however, most of the species share particular characteristics, such as (i) a specific organization of the cell wall, which is mainly composed of peptidoglycans, arabinogalactans, and mycolic acids, and (ii) high G+C content (2–4).

Although *C. ulcerans* has increasing medical and veterinary importance, little is known about its lifestyle and associated virulence factors (5). The sequencing of more *C. ulcerans* genomes of both toxigenic and nontoxigenic strains will help in the identification of distinctive features of strains from human and animal sources (6). In addition, the data generated by newly sequenced genomes are helpful for identifying antibiotic and vaccine targets by way of a comparative analysis (7).

Nowadays, only seven complete genomes and two drafts are available in the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov/genome/). This scenario shows that more genomic knowledge is required in order to better characterize the virulence mechanisms of this emergent pathogen.

In the current study, we present the genome sequence of *C. ulcerans* strain FRC11, isolated from a 74-year-old human a with leg ulcerans infection in Toulouse, France. This strain was first identified as *Corynebacterium pseudotuberculosis* (8), but recent analysis shows that it belongs to *C. ulcerans*.

The sequencing, assembly, and annotation of this strain were performed by the teams from the Laboratory of Cellular and Molecular Genetics (LGCM) and the National Reference Laboratory for Aquatic Animal Diseases (AQUACEN), both located at the Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, and the Laboratory of Polymorphic DNA (LPDNA) at the Federal University of Pará, Belém, Pará, Brazil.

The platform used for sequencing was the Ion Torrent Personal Genome Machine (PGM) system (Life Technologies), using a fragment library. The quality of the raw data was analyzed using the Web tool FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The assembly was done using the Simple Manager for Bacterial Assemblies (SIMBA) interface (http://ufmg-simba.sourceforge.net). The reads with good quality were assembled using a *de novo* strategy with the software MIRA 4.0 (9).

The assembly produced a total of 30 contigs, with a coverage of 179.14× and an N_{50} contig length of 236.335. Additionally, a scaffold was created using the CONTIGuator 2 software (10), using the genome sequence of *C. ulcerans* strain 0102 (accession no. NC_018101.1) (11) as a reference. The gap closure was performed automatically using SIMBA and manually using the CLC Genomics Workbench 7 software.

The genome was automatically annotated using Rapid Annotations using Subsystems Technology (RAST) (12). The manual curation of the annotation was performed using the Artemis software (13) and the UniProt database (http://www.uniprot.org). The CLC Genomics Workbench 7 software was used to correct indel errors in the regions of homopolymers.

The genome includes one circular chromosome of 2,442,826 bp (53.35% G+C content), and 2,210 genes were predicted, 2,146 of which are putative protein-coding genes, with 12 rRNAs and 51 tRNAs; 1 pseudogene was also identified.

Nucleotide sequence accession number. This genome has been deposited in GenBank under the accession no. CP009622.

CCV

ACKNOWLEDGMENTS

This work was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Ministério da Pesca e Aquicultura, and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG). We also acknowledge support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Rede Paraense de Genômica e Proteômica.

REFERENCES

1. **Riegel P, Funke G.** 2000. Corynebacterium et bactéries apparentées, p 993–1019. *In* Freney J, Renaud F, Hansen W (ed), Précis de Bacteriologie Clinique. Edtions Alexandre Lacassagne, Paris, France.
2. **Dorella FA, Pacheco LG, Oliveira SC, Miyoshi A, Azevedo V.** 2006. *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet Res **37:**201–218. http://dx.doi.org/10.1051/vetres:2005056.
3. **Hassan SS, Guimarães LC, Pereira UDP, Islam A, Ali A, Bakhtiar SM, Ribeiro D, Rodrigues Dos Santos A, Soares SDC, Dorella F, Pinto AC, Schneider MP, Barbosa MS, Almeida S, Abreu V, Aburjaile F, Carneiro AR, Cerdeira LT, Fiaux K, Barbosa E, Diniz C, Rocha FS, Ramos RTJ, Jain N, Tiwari S, Barh D, Miyoshi A, Müller B, Silva A, Azevedo V.** 2012. Complete genome sequence of *Corynebacterium pseudotuberculosis* biovar ovis strain p54b96 isolated from antelope in South Africa obtained by rapid next generation sequencing technology. Stand Genomic Sci **7:**189–199. http://dx.doi.org/10.4056/sigs.3066455.
4. **Soares SC, Silva A, Trost E, Blom J, Ramos R, Carneiro A, Ali A, Santos AR, Pinto AC, Diniz C, Barbosa EGV, Dorella Fa, Aburjaile F, Rocha FS, Nascimento KKF, Guimarães LC, Almeida S, Hassan SS, Bakhtiar SM, Pereira UP, Abreu VA, Schneider MP, Miyoshi A, Tauch A, Azevedo V.** 2013. The pan-genome of the animal pathogen *Corynebacterium pseudotuberculosis* reveals differences in genome plasticity between the biovar ovis and equi strains. PLoS One **8:**e53818. http://dx.doi.org/10.1371/journal.pone.0053818.
5. **LaPointe P, Wei X, Gariépy J.** 2005. A role for the protease-sensitive loop region of Shiga-like toxin 1 in the retrotranslocation of its A1 domain from the endoplasmic reticulum lumen. J Biol Chem **280:**23310–23318. http://dx.doi.org/10.1074/jbc.M414193200.
6. **Trost E, Al-Dilaimi A, Papavasiliou P, Schneider J, Viehoever P, Burkovski A, Soares SC, Almeida SS, Dorella Fa, Miyoshi A, Azevedo V, Schneider MP, Silva A, Santos CS, Santos LS, Sabbadini P, Dias AA, Hirata R, Jr, Mattos-Guaraldi AL, Tauch A.** 2011. Comparative analysis of two complete *Corynebacterium ulcerans* genomes and detection of candidate virulence factors. BMC Genomics **12:**383. http://dx.doi.org/10.1186/1471-2164-12-383.
7. **Barbosa EG, Aburjaile FF, Ramos RT, Carneiro AR, Le Loir Y, Baumbach J, Miyoshi A, Silva A, Azevedo V.** 2014. Value of a newly sequenced bacterial genome. World J Biol Chem **5:**161–168.
8. **Guiso N.** 2008. Rapport Annuel d'activité 2008. Institut Pasteur, Paris, France. http://www.pasteur.fr/ip/resource/filecenter/document/01s-00004f-0q7/ra-cnr-coryne-2008.pdf.
9. **Chevreux B, Wetter T, Suhai S.** 1999. Genome sequence assembly using trace signals and additional sequence information, p 45–56. *In* Computer science and biology. Proceedings of the German Conference on Bioinformatics, GCB '99. GCB, Hannover, Germany.
10. **Galardini M, Biondi EG, Bazzicalupo M, Mengoni A.** 2011. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. Source Code Biol Med **6:**11. http://dx.doi.org/10.1186/1751-0473-6-11.
11. **Sekizuka T, Yamamoto A, Komiya T, Kenri T, Takeuchi F, Shibayama K, Takahashi M, Kuroda M, Iwaki M.** 2012. *Corynebacterium ulcerans* 0102 carries the gene encoding diphtheria toxin on a prophage different from the *C. diphtheriae* NCTC 13129 prophage. BMC Microbiol **12:**72. http://dx.doi.org/10.1186/1471-2180-12-72.
12. **Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O.** 2008. The RAST server: Rapid Annotations using Subsystems Technology. BMC Genomics **9:**75. http://dx.doi.org/10.1186/1471-2164-9-75.
13. **Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944–945. http://dx.doi.org/10.1093/bioinformatics/16.10.944.

II.V - Proteome scale comparative modeling for conserved drug and vaccine targets identification in *Corynebacterium pseudotuberculosis*

Syed Shah Hassan[1], **Sandeep Tiwari**[1], **Luís Carlos Guimarães**[1], **Syed Babar Jamal**[1], **Edson Folador**[1], **Neha Barve Sharma**[45], **Siomar de Castro Soares**[1], **Síntia Almeida**[1], **Amjad Ali**[1], **Arshad Islam**[6], **Fabiana Dias Póvoa**[2], **Vinicius Augusto Carvalho de Abreu**[1], **Neha Jain**[45], **Antaripa Bhattacharya**[5], **Lucky Juneja**[45], **Anderson Miyoshi**[1], **Artur Silva**[3], **Debmalya Barh**[5], **Adrian Gustavo Turjanski**[7], **Vasco Azevedo**[1] and **Rafaela Salgado Ferreira**[2*]

- * Corresponding author: Rafaela S Ferreira rafaelasf@gmail.com

Author Affiliations
[1] Laboratory of Cellular and Molecular Genetics, Department of General Biology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[2] Departament of Biochemistry and Immunology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[3] Institute of Biological Sciences, Federal University of Pará, Belém, Para, Brazil
[4] School of Biotechnology, Devi Ahilya University, Khandwa Road Campus, Indore, MP, India
[5] Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal, India
[6] Department of Chemistry, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[7] Structural Bioinformatics Group, Institute of Physical Chemistry of Materials, Environment and Energy, University of Buenos Aires, Argentine

BMC Genomics 2014, **15**(Suppl 7):S3 doi:10.1186/1471-2164-15-S7-S3
The electronic version of this article is the complete one and can be found online at:
http://www.biomedcentral.com/1471-2164/15/S7/S3

II.V.I - Abstract

Corynebacterium pseudotuberculosis (Cp) is a pathogenic bacterium that causes caseous lymphadenitis (CLA), ulcerative lymphangitis, mastitis, and edematous to a broad spectrum of hosts, including ruminants, thereby threatening economic and dairy industries worldwide. Currently there is no effective drug or vaccine available against Cp. To identify new targets, we adopted a novel integrative strategy, which began with the prediction of the modelome (tridimensional protein structures for the proteome of an organism, generated through comparative modeling) for 15 previously sequenced *C. pseudotuberculosis* strains. This pan-modelomics approach identified a set of 331 conserved proteins having 95-100% intra-species sequence similarity. Next, we combined subtractive proteomics and modelomics to reveal a set of 10 Cp proteins, which may be essential for the bacteria. Of these, 4 proteins (tcsR, mtrA, nrdI, and ispH) were essential and non-host homologs (considering man, horse, cow and sheep as hosts) and satisfied all criteria of being putative targets. Additionally, we subjected these 4 proteins to virtual screening of a drug-like compound library. In all cases,

molecules predicted to form favorable interactions and which showed high complementarity to the target were found among the top ranking compounds. The remaining 6 essential proteins (adk, gapA, glyA, fumC, gnd, and aspA) have homologs in the host proteomes. Their active site cavities were compared to the respective cavities in host proteins. We propose that some of these proteins can be selectively targeted using structure-based drug design approaches (SBDD). Our results facilitate the selection of *C. pseudotuberculosis* putative proteins for developing broad-spectrum novel drugs and vaccines. A few of the targets identified here have been validated in other microorganisms, suggesting that our modelome strategy is effective and can also be applicable to other pathogens.

II.V.II - Background

Antimicrobial resistance involving a rapid loss of effectiveness in antibiotic treatment and the increasing number of multi-resistant microbial strains pose global challenges and threats. Thereby, efforts to find new drug and/or vaccine targets to control them are becoming indispensible. *Corynebacterium pseudotuberculosis* (Cp) is a pathogen of great veterinary and economic importance, since it affects animal livestock, mainly sheep and goats, worldwide, and its presence is reported in other mammals in several Arabic, Asiatic, East and West African and North and South American countries, as well as in Australia [1]. *C. pseudotuberculosis* is a Gram-positive, facultative intracellular, and pleomorphic organism; it is non-motile, although presenting fimbriae [2]. Based on *rpoB* gene (a β subunit of RNA polymerase), it shows a close phylogenetic relationship with other type strains of CMNR (*Corynebacterium, Mycobacterium, Nocardia* and *Rhodococcus*), a group that comprises genera of great medical, veterinary and biotechnological importance [1,3]. A recent study showed that phylogenetic analysis for the identification of *Corynebacterium* and other CMNR species based on *rpoB* gene sequences are more accurate than analyses based on 16S rRNA [4]. Its pathogenicity and biological impact have already led to the sequencing of various strains of this pathogen from a wide range of hosts [3]. The pathogen causes several infectious diseases in goat and sheep population (biovar *ovis*), including caseous lymphadenitis (CLA), a chronic contagious disease characterized by abscess formation in superficial lymph nodes and in subcutaneous tissues. In severe cases, biovar *equi* infects the lungs, kidneys, liver and spleen, thereby threatening the herd life of the infected animals [2,5]. The disease has been rarely reported in humans, as a result of occupational exposure, with symptoms similar to lymphadenitis abscesses [6-8]. The bacteria can survive for several weeks in soil in adverse conditions, what seems to contribute to its resistance and disease transmission [9,10]. Direct contact to infectious secretions or contaminated materials are the primary sources of pathogen transmission between animals, but most frequently the infection occurs through exposed skin lacerations [5]. Given the medical importance of Cp and a lack of efficient medicines, in this study we applied a computational strategy to search for new molecular targets from this bacterium.

Recently, computational approaches such as reverse vaccinology, differential genome analyses [11], subtractive and comparative microbial genomics have become popular for rapid identification of novel targets in the post genomic era [12], [13]. These approaches were used to identify targets in various human pathogens, like *Mycobacterium tuberculosis* [14], *Helicobacter pylori* [15], *Burkholderia pseudomalleii* [16], *Neisseria gonorrhea* [17], *Pseudomonas aeruginosa* [18] and *Salmonella typhi* [19]. In general, such approaches follow the principle that genes/proteins must be essential to the pathogen and preferably have no homology to the host proteins [20]. Nevertheless, essential targets that are homologous to their corresponding host proteins may also be molecular targets for structure-based selective inhibitors development. In this case, the targets must show significant differences in the active sites or in other druggable pockets, when pathogenic and host proteins are compared [21-23].

Once a molecular target is chosen, the conventional experimental methods for drug discovery consist of testing many synthetic molecules or natural products to identify lead compounds. Such practices are laborious, time consuming and require high investments [24,25]. On the other hand, computational methods for structure-based rational drug design can expedite the process of ligand identification and molecular understanding of interactions between receptor and ligand [26]. Such approaches are dependent on the availability of the structural information about the target protein. Considering the availability of experimental structures in PDB (Protein Data Bank) only for a low percentage of the known protein sequences, comparative modeling is frequently the method of choice for obtaining 3D coordinates for proteins of interest [27] for the development of specific drugs and docking analyses [28,29].

In this work, we used a modelomic approach for the predicted proteome of *C. pseudotuberculosis* species. This served to bridge the gap between raw genomic information and the identification of good therapeutic targets based on the three dimensional structures. The novelty of this strategy relies in using the structural information from high-throughput comparative modeling for large-scale proteomics data for inhibitor identification, potentially leading to the discovery of compounds able to prevent bacterial growth. The predicted proteomes of 15 *C. pseudotuberculosis* strains were modeled (pan-modelome) using the MHOLline workflow. Intra-species conserved proteome (core-modelome) with adequate 3D models was further filtered for their essential nature for the bacteria, using the database of essential genes (DEG). This led to the identification of 4 essential bacterial proteins without homologs in the host proteomes, which were employed in virtual screening of compound libraries. Furthermore, we investigated a set of 6 essential host homologs proteins. We observed residues of the predicted bacterial protein cavities that are completely different from the ones found in the homologous domains, and therefore could be specifically targeted. By applying this computational strategy we provide a final list of predicted putative targets in *C. pseudotuberculosis*, in biovar *ovis* and *equi*. They could provide an insight into designing of peptide vaccines, and identification of lead, natural and drug-like compounds that bind to these proteins.

II.V.III - Materials and methods

II.V.III.I - Genomes selection

Proteomes predicted based on the genomes of fifteen *C. pseudotuberculosis* strains, including both biovar *equi* and biovar *ovis* (Table 1**)** were used in this study. Most of these genomes were sequenced by our group and are available at NCBI. We downloaded the genome sequences in gbk format from the NCBI server (ftp://ftp.ncbi.nih.gov/genomes/Bacteria webcite) and the corresponding protein sequences (curated CDSs) were exported using Artemis Annotation Tool [30] for further analyses.

Table 1. Strains of *C. pseudotuberculosis* employed in the pan-modelome study, and their respective information regarding genomes statistics, disease prevalence and broad-spectrum hosts.

II.V.III.II - Pan-modelome construction

A high throughput biological workflow, MHOLline (http://www.mholline.lncc.br webcite), was used to predict the modelome (complete set of protein 3D models for the whole proteome) for each Cp strain. MHOLline uses the program MODELLER [31] for protein 3D structure prediction through comparative modeling. Furthermore, the workflow includes BLASTp (Basic Local Alignment Search Tool for Protein) [32], HMMTOP (Prediction of transmembrane helices and topology of proteins) [33], BATS (Blast Automatic Targeting for Structures), FILTERS, ECNGet (Get Enzyme Commission Number), MODELLER and PROCHECK [34] programs. The protocol used here was modified accordingly from the original work by Capriles et al., 2010 [35]. Briefly, the input files of protein sequences were used in FASTA format for all strains because the MHOLline accepts only .faa format files for the whole process. Firstly, MHOLline selected the template structures available at the Protein data Bank (PDB) via BLASTp (version 2.2.18), using the default parameters (e-value $\leq 10e^{-5}$). Secondly, the program BATS refined the BLASTp search for template sequence identification into different groups namely G0, G1, G2 and G3. Only the protein sequences in the group G2, which are characterized by an e-value $\leq 10e^{-5}$, Identity ≥ 0.25 and LVI ≤ 0.7 (where LVI is a length variation index of the BATS program for sequence coverage, the lower the LVI value, the higher the sequence coverage and vice versa) were selected. Among the MHOLline output files, the group G2 contained the largest number of protein sequences ($\geq$ 50% for each input file). Subsequently, the "Filter" tool classified the group G2 sequences into seven distinct quality models groups, from "Very High" to "Very Low" depending on the quality of the template structure for a given query protein sequence. The program MODELLER then modeled all these groups in an automated manner. The number of sequences in the group G2 varies for each *C. pseudotuberculosis* strain. Only the first four distinct quality model groups of G2 were taken into consideration in this study, these were: 1- Very High quality model sequences (identity $\geq$ 75%) (LVI $\leq$ 0.1), 2- High quality model sequences (identity $\geq$ 50%) and < 75%) (LVI $\leq$ 0.1), 3- Good quality model sequences (identity $\geq$ 50%) (LVI > 0.1 and $\leq$

0.3) and 4- Medium to Good quality models (identity ≥ 35% and < 50%) (LVI ≤ 0.3) (http://www.mholline.lncc.br webcite). The percentage of identity represents identity between query and template sequences, a LVI ≤ 0.1 is equivalent to coverage of more than 90%, while LVI ≤ 0.3 corresponds to coverage of more than 70%. Therefore, all protein 3D models considered in this study were built from sequences for which there existed a template with identity ≥ 35% and LVI coverage over 70%. Later on, the ECNGet tool assigned an Enzyme Commission (EC) number to each sequence in G2, according to the best PDB template. The MODELLER (v9v5) program performed the automated global alignment and 3D protein model construction. Finally, the program PROCHECK (v3.5.4) evaluated the constructed models based on their stereo-chemical quality. Additionally, transmembrane regions in the input protein sequences were predicted by HMMTOP, for putative vaccine and drug targets identification.

II.V.III.III - Identification of intra-species conserved genes/proteins

The words genes and proteins are interchangeably used here but they refer to the same protein target of the pathogen. For the identification of highly conserved proteins with 3D models in all Cp strains (≥ 95% sequence identity), the standalone release of NCBI BLASTp+ (v2.2.26) was acquired from the NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ webcite), installed on a local machine and a search was performed for all strains using Cp1002 as a reference genome. The highly conserved proteins were selected using a comparative genomics/proteomics approach using an all-against-all BLASTp analysis with cut off values of *E = 0.0001* [12,17,20,36].

II.V.III.IV - Analyses of essential and non-host homologous (ENH) proteins

To select conserved targets that were essential to the bacteria, a subtractive genomics approach was followed [20]. Briefly, the set of core-modelome proteins from *C. pseudotuberculosis* were subjected to the Database of Essential Genes (DEG) for homology analyses. DEG contains experimentally validated essential genes from 20 bacteria [37]. The BLASTp cutoff values used were: *E-value* = 0.0001, *bit score* ≥100, *identity* ≥ 35% [20].

Furthermore, the pool of essential genes was subjected to NCBI-BLASTp (*E-value = 0.0001, bit score ≥100, identity ≥ 35%*) against (human, equine, bovine and ovine proteomes) to identify essential non-host homologs targets [12]. The set of essential non-host homologous proteins were further crosschecked with the NCBI-BLASTp PDB database using default parameters to find any structural similarity with the available host homologs protein structures, keeping cutoff level to ≤ 15% for query coverage. These proteins were checked for their biochemical pathway using KEGG (Kyoto Encyclopedia of Genes and Genomes) [38], virulence using PAIDB (Pathogenicity island database) [39], functionality using UniProt (Universal Protein Resource) [40], and cellular localization using CELLO (subCELlular LOcalization predictor) [41]. The final list of targets was based on 12 criteria as described previously [20].

II.V.III.V - Analyses of essential and host homologous (EH) proteins

We have extrapolated our analyses and also considered protein targets that were predicted as essential to bacterial survival but showed homology to host proteins. This was based on the possibility to find differences between bacterial and host proteins to rationally design inhibitors. The pool of essential protein targets that showed cut off values equal or higher than those for essential non-host homologs through NCBI-BLASTp was treated as host homologous proteins. These were also analyzed for pathway involvement, virulence, functional annotation and cellular localization like essential non-host homologous proteins. To verify the presence of significant residue differences in druggable protein cavities, a structural comparison was performed for each pathogen and their corresponding host protein through the molecular visualization program PyMOL (v1.5, Schrodinger, LLC) (http://www.pymol.org webcite). The related published data of each template structure for each host homolog was also crosschecked for information about these residues, based on the PDB code of each template structure as input in the PDBelite server [42]. Catalytic Site Atlas (CSA) was also consulted to get robust information of the active site residues for the druggable enzyme targets [43]. CSA is a database documenting enzyme active sites and catalytic residues in enzymes of 3D structure and has 2 types of entry, original hand-annotated entries with literature references and homologous entries, found by PSI-BLAST alignment to an individual original entry, using an *e-value* cut-off of 0.00005. CSA can be accessed via a 4-letter PDB code. The equivalent residue that aligns in the query sequence to the catalytic residue found in the original entry is documented. Though the DoGSiteScorer predicts the druggable protein cavities, the host homologous proteins were further subjected to CASTp (Computed Atlas of Surface Topography of Proteins) [44], Pocket-Finder and Q-SiteFinder [45] to get more reliable and robust results about the druggable cavities of the target proteins.

II.V.III.VI - Prediction of druggable pockets

3D structure information and druggability analyses are important factors for prioritizing and validating putative pathogen targets [46,47]. As aforementioned, for druggability analyses, the final list of essential non-host and host homologous protein targets in PDB format, were subjected to DoGSiteScorer [48], an automated pocket detection and analysis tool for calculating the druggability of protein cavities. For each cavity detected the program returns the residues present in the pocket and a druggable score ranging from 0 to 1. The closer to 1 the obtained values are, the more druggable the protein cavity is predicted to be, i.e. the cavities are predicted to be more likely to bind ligands with high affinity [48]. The DoGSiteScorer also calculates volume, surface area, lipophilic surface, depth and other related parameters for each predicted cavity.

II.V.III.VII - Virtual screening and docking analyses

The ligand library was obtained from the ZINC database, containing 11,193 drug-like molecules, with Tanimoto cutoff level of 60% [49]. Proteins were inspected for structural errors such as missing atoms or erroneous bonds and protonation states in MVD (Molegro Virtual Docker) [50]. The cavities predicted with DogSiteScorer (druggability ≥ 0.80) for all protein targets, were compared with the cavities detected by MVD. The most druggable cavity, according to DogSiteScorer, was subjected to virtual screening. MVD includes three search algorithms for molecular docking namely MolDock Optimizer [50], MolDock Simplex Evolution (SE), and Iterated Simplex (IS). In this work the MolDock Optimizer search algorithm, which is based on a differential evolutionary algorithm, was employed. The default parameters used for the guided differential evolution algorithm are a) population size = 50, b) crossover rate = 0.9, and c) scaling factor = 0.5. The top ranked 200 compounds for each protein were analyzed in Chimera for shape complementarity and hydrogen bond interactions, leading to the selection of a final set of 10 compounds for each target protein.

II.V.IV - Results and discussion

II.V.IV.I - Modelome and common targets in *C. pseudotuberculosis* species

Here we report the identification of common putative targets among 15 strains of *C. pseudotuberculosis* species based on the construction of genome scale protein three-dimensional structural models. Structural information of target proteins can aid in drug and/or vaccine design and in the discovery of new lead compounds [51]. The approach employed here generated high-confidence structural models through the MHOLline workflow **(**Figure 1**)** from orthologous protein. To identify the common conserved proteins with a sequence similarity of 95-100%, a comparative genomics approach was performed where all the BATS classified G2 sequences from "Very High" to "Medium to Good" quality, from 14 Cp strains, were aligned to the G2 sequences of Cp1002, assumed as a reference genome for this study. In total, a set of 331 protein sequences was selected, being conserved in all strains. An overview of the different steps involved in this computational approach for genome scale modelome and prioritization of putative drug and vaccine targets is given in Figure 2a-b.

Figure 1. High-throughputness (efficiency) of the MHOLline biological workflow for genome-scale modelome (3D models) prediction. Predicted proteomes from the genomes of 15 *C. pseudotuberculosis* strains were fed to the MHOLline workflow in FASTA format. The blue line represents the number of input data, according to the left-hand side y-axis. The bars show the number in the form of MHOLline output data (according to the right-hand side y-axis) of: not aligned sequences (G0, green bars); sequences for which there is a template structure available at RCSB PDB (yellow bars); sequences with acceptable template structures that where modeled in the MHOLline workflow (G2, red bars); sequences with predicted transmembrane regions (HMMTOP, purple bars) and the number of sequences that were predicted as enzymes in each genome and were assigned an EC number (ECNGet, gray bars). The x-axis represents the *C. pseudotuberculosis* genomes used in this study.

Figure 2. Overview of different computational steps employed in the identification of putative essential targets (non-host homologous and host homologous) for drugs and vaccines from the core-proteome of 15 *C.* pseudotuberculosis strains. **Figure 2b**. Intra-species subtractive modelomics workflow for conserved targets identification in *C.* pseudo tuberculosis species. The table (from left to right) represents the total number of protein sequences as an input data in fasta format fed to the MHOLline workflow (upper forward arrow). The remaining columns show the output data of group G2 (upper backward arrow), first by BATS and then by Filter tools of the MHOLline workflow respectively. Columns 4th-7th constitute the number of protein sequences of different qualities of all 15 Cp strains, where the sequences of 14 Cp strains were compared using BLASTp, to the sequences of Cp1002 strain as reference, for the identification of conserved protein targets (core-modelome). The funnel shows how this workflow processes and filters a large quantity of genomic data for putative drug and vaccine targets identification of a pathogen.

II.V.IV.II - Identification of ENH and EH proteins as putative drug and/or vaccine targets

To identify essential proteins as putative therapeutic targets in *C. pseudotuberculosis*, from the set of core-modelome, these were compared to the Database of Essential Genes (DEG). Based on this filter, the number of selected targets was reduced drastically to a final set of only 10 targets. These were compared to the aforementioned corresponding host proteomes, leading to the identification of 4 essential non-host homologous proteins (ENH, Table 2) and 6 essential host homologous proteins (EH, Table 3).

Table 2. Drug and/or vaccine targets prioritization parameters and functional annotation of the four essential non-host homologous putative targets.

Table 3. Drug and/or vaccine targets prioritization parameters and functional annotation of the six essential host homologous putative targets.

Among the ENH proteins, two targets were selected from a bacterial unique pathway, the two component signaling system. These targets are tcsR (two-component response regulator) and mtrA (two component sensory transduction transcriptional regulatory protein). While the tcsR is a novel protein target, as it is has not been described so far as a target in any organism, mtrA has been already reported as a target in *Mycobacterium* [52] and provides multidrug resistance to *Mycobacterium avium* [53]. Therefore, targeting mtrA in *C. pseudotuberculosis* may also be effective in controlling the infection of CLA. The remaining ENH protein targets, nrdI and ispH, also participate in biochemical pathways. NrdI (ribonucleoside-diphosphate reductase alpha chain) is a flavodoxin which contains a diferric-tyrosyl radical cofactor and it is involved in nucleotide metabolism in *E. coli* [54]. It has been reported as a putative target in several pathogens including *C. pseudotuberculosis, Corynebacterium diphtheriae* and *Mycobacterium tuberculosis* [20]. The target ispH (4-hydroxy-3-methylbut-2-enyl diphosphate reductase; EC 1.17.1.2) is an essential cytoplasmic enzyme in *Escherichia coli* [55]. This iron-sulfur protein plays a crucial role in terpene metabolism of various pathogenic bacteria [56,57] and it is a predicted target in *Salmonella tyhpimurium* [58] and *Plasmodium falciparum* [59]. It should be noted that according to the cut off threshold for NCBI-BLASTp

that we have followed, ispH shows homology only to the human host. So, if human is not considered as a possible host, ispH can also be considered as a common putative target. The roles of these proteins in different metabolic pathways was confirmed from KEGG [38] and METACYC [60] databases.

II.V.IV.III - Prioritization parameters of drug and/or vaccine targets

Previous studies have shown several factors that can aid in determining the suitability of therapeutic targets [46]. The availability of 3D structural information, the main approach of our study, is very helpful in drug development. Other important factors for drug targets include preferred low MW and high druggability. On the other hand, for vaccine targets the information about subcellular localization is important and proteins that contain transmembrane motifs are preferred [36,46,61,62]. We have determined most of these prioritizing properties for the 10 essential proteins (Table 2 &3). Interestingly, according to the target-prioritizing criterion, all targets have a low MW, and are predicted to be localized in the cytoplasmic compartment of the Cp. Druggability evaluation with DoGSiteScorer [48] for all conserved targets allowed the prediction of numerous druggable cavities with at least one druggable cavity for each Cp target. For the 4 ENH proteins tcsR, mtrA, nrdI, and ispH, 3, 5, 5 and 2 cavities with score ≥ 0.80 were observed respectively. For each protein, the cavity that exhibited the highest druggability score was selected for docking analyses. For 6 EH targets, adk, gapA, glyA, fumC, gnd, and aspA, 1, 3, 3, 2, 8 and 6 cavities were observed respectively according to the aforementioned druggability score criteria (Table 2 &3). Here, in each case, the most druggable predicted cavity was structurally compared with the cavities in respective host proteins.

II.V.IV.IV - Virtual screening and molecular docking analyses of ENH targets

For each ENH target protein (mtrA, ispH, tcsR and nrdl), the top 200 drug-like molecules from virtual screening were visually inspected to select 10 molecules that showed favorable interactions with the target. The biological importance of each target and an analysis of the predicted protein-ligand interaction are described below. ZINC codes and MolDock scores of selected ligands, the number of hydrogen bonds as well as protein residues involved in these interactions, are shown in a table for each target protein (Tables 4, 5, 6, 7. Figures showing the predicted binding mode for one of the 10 selected ligands are also shown for each target (Additional files 1, 2, 3, 4, 5).

Table 4. ZINC codes, MolDock scores and predicted hydrogen bonds for the ten compounds selected among the top ranking 200 molecules against **Cp1002_0515** (**MtrA**, DNA-binding response regulator).

Table 5. ZINC codes, MolDock scores and predicted hydrogen bonds for the ten compounds selected among the top ranking 200 molecules against **Cp1002_0742** (**IspH**, 4-hydroxy-3-methyl but-2-enyl diphosphate reductase).

Table 6. ZINC codes, MolDock scores and predicted hydrogen bonds for the ten compounds selected among the top ranking 200 molecules against **Cp1002_1648** (**TcsR,** Two component transcriptional regulator).

Table 7. ZINC codes, MolDock scores and predicted hydrogen bonds for the ten compounds selected among the top ranking 200 molecules against **Cp1002_1676** (**NrdI**).

Additional file 1. Docking representation of the best drug-like compound **ZINC75109074** in the most druggable protein cavity of **Cp1002_0515** (**MtrA**, DNA-binding response regulator). Three hydrogen bonds were observed with Thr73, Asp48 and Arg116.

Additional file 2. Docking representation of compound **ZINC00510419** in the most druggable protein cavity of **Cp1002_0742** (**IspH**, 4-hydroxy-3-methyl but-2-enyl diphosphate reductase). Residues Cys39, Thr225, Ser250, His68 and Asn252 are predicted to make seven hydrogen bonds to this ligand.

Additional file 3. Docking representation of the best drug-like compound **ZINC00510419** in the most druggable protein cavity of **Cp1002_1648** (**TcsR,** Two component transcriptional regulator). Hydrogen bonds were observed with residues Val76, Gln185 and Asn193.

Additional file 4. Docking representation of the best drug-like compound **ZINC04721321** in the most druggable protein cavity of **Cp1002_1676** (**NrdI** protein). Hydrogen bonds were observed with residues Ser8, Thr13 and Leu116.

Additional file 5 (a-f). Comparison among the most druggable cavities from essential bacterial and the respective host homologue proteins. Protein structures are shown as cartoon (green for the bacterial protein and gray for *Ovis aries* host protein). Other host proteins are not shown for simplicity, but the same substitutions were present in all host proteins analyzed. Residues that differ in the bacterial and host cavity are highlighted in sticks and labeled (bacterial labels in green and host labels in black). a) **Cp1002_0692** (Glyceralderayde 3-phosphate dehydrogenase); b) **Cp1002_0385** (adenylate kinase); c) **Cp1002_0728** (serine hydroxymethyltransferase); d) **Cp1002_0738** (fumarate hydratase class II) the site shown is formed by three monomers, which are represented in green, blue and orange. No residues are highlighted, since the active sites are identical between bacteria and host; e) **Cp1002_1005** (6-phosphogluconate dehydrogenase); f) **Cp1002_1042** (aspartate ammonia-lyase). Figures were prepared with the PyMol.

Cp1002_0515 (**MtrA**, DNA-binding response regulator) is part of the two-component signal transduction system consisting of the sensor kinase (Histidine protein kinases, HKs) and the response regulator, MtrB and MtrA respectively. This system is highly conserved in *Corynebacteria* and *Mycobacteria* and it is essential for their survival to adapt to environmental changes. Homologs of MtrA and MtrB are present in many species of the genera *Corynebacterium, Mycobacterium, Nocardia, Rhodococcus* (CMNR), and others like *Thermomonospora, Leifsonia, Streptomyces, Propionibacterium,* and *Bifidobacterium* [63]. MtrA represents the fourth family member of the OmpR/PhoB family of response regulators. Like other family members, MtrA has been reported to be essential in *M. tuberculosis* [64]. It possesses an N-terminal regulatory domain and a C-terminal helix-turn-helix DNA-binding

domain, already indicating that this response regulator functions as a transcriptional regulator, with phosphorylation of the regulatory domain modulating the activity of the protein [65]. Based on a comparison with a crystallographic structure of the MtrA template (2GWR, MtrA from *M. tuberculosis*), the active site residues involved in H-bond interactions with the crystallographic ligand are Val145, Gln151, Ile152 and Leu154. Although none of these residues is predicted to form hydrogen bonds with the ten selected docked ligands, these molecules were predicted to interact with other residues in the pocket. Table 4 shows the 10 selected ligands according to their minimum energy values and number of hydrogen bond interactions. **ZINC75109074** (N-benzyl-N-[[2-(2-thienyl)-1H-imidazol-4-yl] methyl] prop-2-en-1-amine) is shown here as the top scoring ligand **(**Additional file 1**)**.

Cp1002_0742 (**IspH**, 4-hydroxy-3-methylbut-2-enyl diphosphate reductase) is an iron-sulfur oxidoreductase enzyme that plays a key role in the metabolism of terpenes in several pathogens. Terpenes constitute a large class of natural compounds. Their biosynthesis initiates with the building blocks isopentenyl-diphosphate (IPP) and dimethylallyldiphosphate (DMAPP), and differs in bacteria and mammals [57]. In bacteria and other pathogenic microorganisms the enzyme IspH catalyzes the last step in the production of IPP and DMAPP. The three structural units of the enzyme harbor a cubic iron-sulfur cluster at their center, enabling the enzyme to accomplish a challenging reaction by converting an allyl alcohol to two isoprene components. The iron-sulfur proteins normally participate in electron transfers. The IspH enzyme, thereby, in a similar fashion, binds the substrate directly to the iron-sulfur cluster [57]. In the template crystal structure of IspH (PDB 3KE8), it has been shown that His41, His74, His124, Thr167, Ser225, Ser226, Asn227 and Ser269 are the active site residues that are involved in hydrogen bond interactions with the ligand 4-hydroxy-3-methylbutyldiphosphate (EIP). Also, Cys12, Cys96, Cys197 and EIP have been shown to make metal interaction with the Fe_4S_4 (Iron/Sulfur Cluster). Although the ten selected drug-like compounds (Table 5**)** did not show any interaction with the aforementioned IspH residues, they are predicted to make very good hydrogen bond interactions with other surrounding residues of the predicted cavity. The predicted binding mode of the best scoring compound, **ZINC00510419** is shown in Additional file 2. Good shape complementarity and 6 hydrogen bond interactions are observed in this complex.

Cp1002_1648 (**TcsR,** Two component transcriptional regulator) is a novel target without host homologs proteins. Differently from MtrA and IspH, in this case the template structure from *Escherichia coli* for TcsR did not contain any ligand (PDB 1A04), and no reported information was found about the ligand-residues interactions in their cavities. Therefore, among the cavities identified by MVD, the best cavity for virtual screening analysis was simply chosen based on the highest druggability score by the DogSiteScorer. Compound **ZINC00510419** (Additional file 3**)** was the top-ranking compound, forming a network of 3 hydrogen bonds with Val76, Gln185 and Asn193. Table 6 lists the 10 compounds selected for this target.

Cp1002_1676 (**NrdI,** protein) belongs to the nrdI protein family, a unique group of metalloenzymes that are essential for cell-proliferation [66]. It is classified as a ribonucleotide reductase (RNR), an iron-dependent enzyme that belongs to class Oxidoreductases (EC 1.17.4.1) acting on CH or CH_2 groups with a disulfide as acceptor [67]. The class Ia enzyme supplies deoxynucleotides during normal aerobic growth. The class Ib RNR plays a similar role although its function in *E. coli* is not clear, but it is reported to be expressed under oxidative stress and iron-limited conditions [68]. Class I RNR enzymes have two homodimeric subunits, α2 (NrdE), where nucleotide reduction takes place, and β2 (NrdF) containing an unidentified metallocofactor for initiating nucleotide reduction in α2. Although the exact function of NrdI within RNR has not yet been fully characterized, it is found in the same operon as NrdE and NrdF, and encodes an unusual flavodoxin, a bacterial electron-transfer protein that includes a flavin mononucleotide that has been proposed to be involved in metallocofactor biosynthesis and/or maintenance. It has also been proposed that NrdI plays an important role in *E. coli* class Ib RNR cluster assembly. Recent *in vitro* studies have shown that a stable diferric-tyrosyl radical (FeIII2-Y·) and dimanganese (III)-Y· (MnIII2-Y·) cofactors are active in nucleotide reduction [69]. The first one can be formed by self-assembly from FeII and O_2 while the later cofactor can be generated from MnII-2-NrdF, but only in the presence of O_2 and NrdI protein [54,69]. RNR is responsible for the *de novo* conversion of ribonucleoside diphosphates into deoxyribonucleoside diphosphates and it is essential for DNA synthesis and repair [70]. The active site residues of RNR, in the template structure of NrdI protein (PDB 3N3A), include Ser8, Ser9, Ser11, Ser48, Asn13, Asn83, Thr14, Tyr49, Ala89 and Gly91, all of which are involved in a hydrogen bond network with the cofactor flavin mononucleotide isoalloxazine ring (FMN, PDB 3N3A) [71]. Interestingly, two of these residues, Ser8 and Tyr49, were predicted to make hydrogen bonds with all 10 selected ligands (Table 7**)**. The interaction between the top scoring compound **ZINC01585114** (5-nitro-3, 4-diphenyl-2-furamide) and the residues from the predicted target cavities are shown in Additional file 4.

Furthermore, the drug-like molecule **ZINC00510419** (3,4-bis (5-methylisoxazole-3-carbonyl)-1,2,5-oxadiazole 2-oxide was among the top ten selected molecules for three of the pathogen target proteins, showing good H-bond interactions. It ranked first against the targets Cp1002_0742 (MolDock score = -151.376, no. of H-bonds = 7) and Cp1002_1648 (MolDock score = -167.633, no. of H-bonds = 3) and ranked fourth against the target Cp1002_1676 (MolDock score = -154.064, no. of H-bonds = 4).

II.V.IV.V - Essential host homologous as putative targets

To compare the predicted EH protein targets to their host homologs, two approaches were taken. First, ClustalX (v2.1, http://www.clustal.org webcite), a multiple sequence alignment program, was used to find different residues between bacterial and host proteins. As expected, a high percentage of residues was found to be conserved, but significant differences were also observed. Most percentage identities are between 35 and 50 (Table 8), except for fumarate

hydratase, which shows 54% sequence identity to human and equine homologous proteins, but no hits in bovine and ovine proteomes.

Table 8. Percentage of sequence identity between *C.* pseudotuberculosis and host homologous proteins.

Next, to determine if the observed differences could be exploited in rational design of ligands selective to bacterial proteins, we focused on the predicted druggable cavities. A structural alignment to the host homologous proteins was performed and the cavities were compared in PyMol. In most cases, the DogSiteScorer predicted more than one cavity for each input Cp protein structure. The number of residues in the bacterial predicted cavity that differ from the residues in the cavity of the host protein, for all druggable pockets, varied from zero to seven (Table 9).

Table 9. Comparison of the residues from druggable cavities in *C. pseudotuberculosis* proteins and the corresponding residues in structurally aligned host protein cavities.

For conserved host-homologous targets Cp1002_0385 (adk, Adenylate kinase), Cp1002_0692 (gapA, Glyceraldehyde 3-phosphate dehydrogenase), Cp1002_0728 (glyA, Serine hydroxymethyltransferase), Cp1002_0738 (fumC, Fumarate hydratase class II/fumarase), Cp1002_1005 (gnd, 6-Phosphogluconate dehydrogenase) and Cp1002_1042 (aspA, Aspartate ammonia-lyase/aspartase), three, four, five, zero, seven and three different residues were observed, respectively. Then, a more detailed analysis was performed for the predicted highest druggable cavity for each protein. The results are described below, together with information about the biological importance of each target protein.

Cp1002_0692 (**GapA,** Glyceraldehyde 3-phosphate dehydrogenase, GAPDH/G3PDH, EC 1.2.1.12) catalyzes the sixth step of glycolysis. In addition, GAPDH has recently been shown to be involved in several non-metabolic processes, including transcription activation, initiation of apoptosis [72] fast axonal or axoplasmic transport and endoplasmic reticulum to Golgi vesicle shuttling [73,74]. This enzyme has been reported as an anti-trypanosomatid and anti-leishmania drug target in structure-based drug design efforts [21-23]. Furthermore, it has been shown as an interesting putative drug and vaccine target in malaria pathogenesis [75]. Comparison of protein cavities reveals significant differences between bacterial and host proteins, with replacement of bacterial Lys157, Arg229 and Asn311 by Asp, Thr and Ala, respectively. Such differences result in a more basic cavity in bacteria, making it possible to rationally design selective ligands, especially negatively charged molecules, which interact with Lys157 and Arg229, or compounds able to form hydrogen bond to Asn311 (Additional file 5).

Nucleoside monophosphate kinases vitally participate in sustaining the intracellular nucleotide pools in all living organisms. **Cp1002_0385** (**Adk,** Adenylate kinase, EC 2.7.4.3) is a ubiquitous enzyme, which catalyzes the reversible $Mg2^{+}$-dependent transfer of the terminal phosphate group from ATP to AMP, releasing two molecules of ADP [76]. Only one

highly druggable cavity was predicted for adenylate kinase, with a druggability score = 0.81. Three residues in the bacteria cavity were different from the hosts: Leu, Met and Val in the hosts replaced Phe35, Ile53 and Thr64, respectively (Additional file 5). These differences impact the cavity volume, since aromatic and bulky Phe is replaced by Leu, and the ability to make hydrogen bonds, through the replacement of a Thr by a Val. Therefore; the bacterial cavity is smaller and more hydrophilic, making it possible to envision rational design of selective ligands that interact with Thr64.

Cp1002_0728 (GlyA, Serine hydroxymethyltransferase EC 2.1.2.1) is an enzyme that plays an important role in cellular one-carbon pathways by catalyzing the reversible, simultaneous conversions of L-serine to glycine (retro-aldol cleavage) and tetrahydrofolate to 5,10-methylenetetrahydrofolate [77]. In Plasmodium, serine hydroxymethyltransferase (SHMT) has been reported as an attractive drug target [78]. For this protein 3 residues were observed different between bacteria and host: Ala99 and Ala101 replaced two Ser residues while Trp177 replaced Thr (Additional file 5). At first glance these changes could have a big impact in the active site, generating a considerably more hydrophilic pocket in the hosts. However, careful inspection of the pocket reveals that the side chains of these residues are not turned towards the pocket, in such a way that these differences probably would not allow rational design of selective ligands.

Cp1002_0738 (**FumC,** Fumaratehydratase class II/fumarase EC 4.2.1.2) catalyzes the reversible hydration/dehydration of fumarate to S-malate during the ubiquitous Krebs cycle, through the aci-carboxylate intermediate subsequent to olefin production [79]. There are two classes of fumarases; Class I fumarases, composed of heat-labile, iron-sulfur (4Fe-4S) homodimeric enzymes, only found in prokaryotes; and Class II fumarases, made of thermostable homotetrameric enzymes [80] found in both prokaryotic and eukaryotic mitochondria. Class II belongs to a superfamily that also includes aspartate-ammonia lyases, arginino-succinatases, d-crystallins and 3-carboxy-cis, cis-muconate lactonizing enzymes. All these enzymes release fumarate from different substrates, ranging from adenylosuccinate to malate [81-84]. FumC of *Escherichia coli* is the first member of class II fumarases family whose structure has been solved and provided most of the structural information [85]. Inhibition of fumarase in the tricarboxylic acid cycle (TCA) has been reported as a potential molecular target of bismuth drugs in *Helicobacter pylori* [86]. Comparison of the active site cavity of this protein, which is formed in the interface of three monomers, revealed no differences between bacteria and hosts (additional file 5).

Cp1002_1005 (**Gnd,** 6-Phosphogluconate dehydrogenase EC 1.1.1.44) is an enzyme from the pentose phosphate pathway. It forms ribulose 5-phosphate from 6-phosphogluconate. The enzyme 6-phosphogluconate dehydrogenase is a potential drug target for the parasitic protozoan *Trypanosoma brucei*, the causative organism of human African trypanosomiasis [87]. Three druggable sites with score > 0.80 were detected in this protein. As opposed to the observation for other proteins, the most druggable predicted cavity (score = 0.88) was not the active site. Leu, Lys and Val residues in the hosts replace residues Met94, Gln96 and Ile148

in the bacterial cavity, respectively (Additional file 5). The most significant of these differences is the replacement of Gln by Lys, which could make binding of negative molecules more favorable to the host proteins.

Cp1002_1042 (**AspA,** Aspartate ammonia-lyase/aspartase EC 4.3.1.1) catalyzes the deamination of aspartic acid to form fumarate and ammonia [88]. Recent progresses to prepare enantiopure l-aspartic acid derivatives, highly valuable tools for biological research and chiral building blocks for pharmaceuticals and food additives, make it a target of interest for industrial applications. On the other hand, the important role that it plays in microbial nitrogen metabolism makes it a putative drug target in overcoming bacterial pathogenesis [89]. Based on the sequence alignment for this protein, two significant differences in residues are observed in the most druggable pocket: bacterial His447 and Ile428 are replaced by Leu and Lys in host proteins. Such differences should allow rational ligand design. It is interesting to note that additional differences in the position of helices that contain these residues increase the difference between the active sites (Additional file 5).

Based on the above-mentioned analyses, we conclude that it would be difficult to rationally design selective ligands for **Cp1002_0738** (**FumC,** Fumaratehydratase class II), since no residue differences were observed in the most druggable cavity, and for **Cp1002_0728 (GlyA**, Serine hydroxymethyltransferase), where the side chains of differing residues are not turned toward the druggable pocket. On the other hand, for putative essential and homologous targets that include **Cp1002_0692** (**GapA,** Glyceraldehyde 3-phosphate dehydrogenase), **Cp1002_0385** (**Adk,** Adenylate kinase), **Cp1002_1005** (**Gnd,** 6-Phosphogluconate dehydrogenase) and **Cp1002_1042** (**AspA,** Aspartate ammonia-lyase), significant differences were observed in druggable pockets, suggesting that despite the existence of a host homologous protein they could be good targets for the design of ligands, selective only to the bacterial proteins.

II.V.V - Conclusion

Here, for the first time, the genomic information was used to determine the conserved predicted proteome of 15 strains of *C. pseudotuberculosis*, along with their three-dimensional structural information. Even though the structural information discussed is fully computationally predicted, and could therefore deviate from eventually solved experimental structures, we have been careful to concentrate on the analysis of protein models for which there were good templates which provided high quality models, minimizing this concern. The data presented here can effectively contribute in guiding further research for antibiotics and vaccines development. The final dataset can provide valuable information in designing molecular biology and immunization experiments in animal models for validating the targets of a pathogen, as well as in experimental structure determination protocols.

The criterion for target selection in *C. pseudotuberculosis* was stringent, resulting in a small set of prioritized putative drug and vaccine targets, of which four are essential and non-

homologous and six are essential and host homologous proteins. For the latter, a detailed structural comparison between the residues of the predicted cavities of host and pathogen proteins has been performed, showing in most cases the potential for the development of selective ligands. Therefore, we suggest that the whole set can be considered for antimicrobial chemotherapy, especially the four essential non-host homologous targets.

The *in silico* approaches followed in this study might aid in the development of novel therapeutic drugs and vaccines in a broad-spectrum of hosts at intraspecies level against *C. pseudotuberculosis*. Furthermore, the strategy described here could also be applied to other pathogenic microorganisms.

II.V.VI - Authors' contributions

Coordinated entire work: SSH RSF VA DB. Performed all *in silico* analyses: SSH RSF ST SBJ NBS FDP LCG. Cross-analyzed genome contents, pan-modelome construction, conserved pan-modelome, subtractive modelome approach, virtual screening & docking analyses and residue level structural comparison: SSH RSF ST FDP AI SCS SA DB AGT. Provided timely consultation and reviewed the manuscript: VA AI SCS SA DB NBS LCG AA AM AS VACA AGT. Read and approved the final manuscript: RSF SSH ST AI SCS SBJ SA DB NBS LCG AGTAA AM AS VA. Conceived and designed the work: SSH RSF VA DB. Analyzed the data: SSH RSF ST AI SCS SBJ SA DB NBS LCG AA AB LJ AGTAM AS VA. Wrote the paper: SSH RSF ST.

II.V.VII - Conflict of interest

The authors declare that they have no competing interests.

II.V.VIII - Acknowledgements

We acknowledge financial support from the funding agencies CNPq, CAPES and FAPEMIG. Hassan S.S acknowledges the receipt of fellowship under "TWAS-CNPq Postgraduate Fellowship Program" for doctoral studies.

This article has been published as part of *BMC Genomics* Volume 15 Supplement 7, 2014: Proceedings of the 9th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-Meeting 2013). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/15/S7.

II.V.IX - References

1. Hassan SS, Schneider MP, Ramos RT, Carneiro AR, Ranieri A, Guimaraes LC, Ali A, Bakhtiar SM, Pereira Ude P, dos Santos AR, *et al.*: **Whole-genome sequence of *Corynebacterium pseudotuberculosis* strain Cp162, isolated from camel.** *Journal of bacteriology* 2012, **194**(20)**:**5718-5719.
2. Dorella FA, Pacheco LG, Oliveira SC, Miyoshi A, Azevedo V: ***Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence.** *Veterinary research* 2006, **37**(2)**:**201-218.
3. Soares SC, Trost E, Ramos RT, Carneiro AR, Santos AR, Pinto AC, Barbosa E, Aburjaile F, Ali A, Diniz CA, *et al.*: **Genome sequence of Corynebacterium pseudotuberculosis biovar *equi* strain 258 and prediction of antigenic targets to improve biotechnological vaccine production.** *Journal of biotechnology* 2012.
4. Khamis A, Raoult D, La Scola B: **Comparison between rpoB and 16S rRNA gene sequencing for molecular identification of 168 clinical isolates of Corynebacterium.** *Journal of clinical microbiology* 2005, **43**(4)**:**1934-1936.
5. Williamson LH: **Caseous lymphadenitis in small ruminants.** *Vet Clin North Am Food Anim Pract* 2001, **17**(2)**:**359-371. vii
6. Peel MM, Palmer GG, Stacpoole AM, Kerr TG: **Human lymphadenitis due to Corynebacterium pseudotuberculosis: report of ten cases from Australia and review.** *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 1997, **24**(2)**:**185-191.
7. Luis MA, Lunetta AC: **[Alcohol and drugs: preliminary survey of Brazilian nursing research].** *Revista latino-americana de enfermagem* 2005., **13**Spec No:1219-1230
8. Mills AE, Mitchell RD, Lim EK: **Corynebacterium pseudotuberculosis is a cause of human necrotising granulomatous lymphadenitis.** *Pathology* 1997, **29**(2)**:**231-233.
9. Augustine JL, Renshaw HW: **Survival of Corynebacterium pseudotuberculosis in axenic purulent exudate on common barnyard fomites.** *American journal of veterinary research* 1986, **47**(4)**:**713-715.
10. Yeruham I, Friedman S, Perl S, Elad D, Berkovich Y, Kalgard Y: **A herd level analysis of a Corynebacterium pseudotuberculosis outbreak in a dairy cattle herd.** *Veterinary dermatology* 2004, **15**(5)**:**315-320.
11. Perumal D, Lim CS, Sakharkar KR, Sakharkar MK: **Differential genome analyses of metabolic enzymes in Pseudomonas aeruginosa for drug target identification.** *In silico biology* 2007, **7**(4-5)**:**453-465.
12. Barh D, Gupta K, Jain N, Khatri G, Leon-Sicairos N, Canizalez-Roman A, Tiwari S, Verma A, Rahangdale S, Shah Hassan S, *et al.*: **Conserved host-pathogen PPIs.** *Integrative biology : quantitative biosciences from nano to macro* 2013.
13. Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B, *et al.*: **Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing.** *Science* 2000, **287**(5459)**:**1816-1820.
14. Asif SM, Asad A, Faizan A, Anjali MS, Arvind A, Neelesh K, Hirdesh K, Sanjay K: **Dataset of potential targets for Mycobacterium tuberculosis H37Rv through comparative genome analysis.** *Bioinformation* 2009, **4**(6)**:**245-248.
15. Dutta A, Singh SK, Ghosh P, Mukherjee R, Mitter S, Bandyopadhyay D: **In silico identification of potential therapeutic targets in the human pathogen Helicobacter pylori.** *In silico biology* 2006, **6**(1-2)**:**43-47.
16. Chong CE, Lim BS, Nathan S, Mohamed R: **In silico analysis of Burkholderia pseudomallei genome sequence for potential drug targets.** *In silico biology* 2006, **6**(4)**:**341-346.
17. Barh D, Kumar A: **In silico identification of candidate drug and vaccine targets from various pathways in Neisseria gonorrhoeae.** *In silico biology* 2009, **9**(4)**:**225-231.
18. Sakharkar KR, Sakharkar MK, Chow VT: **A novel genomics approach for the identification of drug targets in pathogens, with special reference to Pseudomonas aeruginosa.** *In silico biology* 2004, **4**(3)**:**355-360.
19. Rathi B, Sarangi AN, Trivedi N: **Genome subtraction for novel target definition in Salmonella typhi.** *Bioinformation* 2009, **4**(4)**:**143-150.
20. Barh D, Jain N, Tiwari S, Parida BP, D'Afonseca V, Li L, Ali A, Santos AR, Guimaraes LC, de Castro Soares S, *et al.*: **A novel comparative genomics analysis for common drug and vaccine targets in Corynebacterium pseudotuberculosis and other CMN group of human pathogens.** *Chemical biology & drug design* 2011, **78**(1)**:**73-84.
21. Aronov AM, Verlinde CL, Hol WG, Gelb MH: **Selective tight binding inhibitors of trypanosomal glyceraldehyde-3-phosphate dehydrogenase via structure-based drug design.** *Journal of medicinal chemistry* 1998, **41**(24)**:**4790-4799.
22. Singh S, Malik BK, Sharma DK: **Molecular modeling and docking analysis of Entamoeba histolytica glyceraldehyde-3 phosphate dehydrogenase, a potential target enzyme for anti-protozoal drug development.** *Chemical biology & drug design* 2008, **71**(6)**:**554-562.
23. Suresh S, Bressi JC, Kennedy KJ, Verlinde CL, Gelb MH, Hol WG: **Conformational changes in Leishmania mexicana glyceraldehyde-3-phosphate dehydrogenase induced by designed inhibitors.** *Journal of molecular biology* 2001, **309**(2)**:**423-435.
24. Adams CP, Brantner VV: **Estimating the cost of new drug development: is it really 802 million dollars?** *Health affairs* 2006, **25**(2)**:**420-428.

25. Kola I, Landis J: **Can the pharmaceutical industry reduce attrition rates?** *Nature reviews Drug discovery* 2004, **3**(8)**:**711-715.
26. Congreve M, Murray CW, Blundell TL: **Structural biology and drug discovery.** *Drug discovery today* 2005, **10**(13)**:**895-907.
27. Baker D, Sali A: **Protein structure prediction and structural genomics.** *Science* 2001, **294**(5540)**:**93-96.
28. Cavasotto CN, Phatak SS: **Homology modeling in drug discovery: current trends and applications.** *Drug discovery today* 2009, **14**(13-14)**:**676-683.
29. Behera DK, Behera PM, Acharya L, Dixit A, Padhi P: **In silico biology of H1N1: molecular modelling of novel receptors and docking studies of inhibitors to reveal new insight in flu treatment.** *Journal of biomedicine & biotechnology* 2012, **2012:**714623.
30. Mural RJ: **ARTEMIS: a tool for displaying and annotating DNA sequence.** *Briefings in bioinformatics* 2000, **1**(2)**:**199-200.
31. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A: **Comparative protein structure modeling using MODELLER.** *Current protocols in protein science / editorial board, John E Coligan [et al]* 2007. Chapter 2:Unit 2 9
32. Mount DW: **Using the Basic Local Alignment Search Tool (BLAST).** *CSH protocols* 2007. 2007:pdb top17
33. Tusnady GE, Simon I: **The HMMTOP transmembrane topology prediction server.** *Bioinformatics* 2001, **17**(9)**:**849-850.
34. Laskowski RA, Macarthur MW, Moss DS, Thornton JM: **Procheck - a Program to Check the Stereochemical Quality of Protein Structures.** *J Appl Crystallogr* 1993, **26:**283-291.
35. Capriles PV, Guimaraes AC, Otto TD, Miranda AB, Dardenne LE, Degrave WM: **Structural modelling and comparative analysis of homologous, analogous and specific proteins from Trypanosoma cruzi versus Homo sapiens: putative drug targets for chagas' disease treatment.** *BMC genomics* 2010, **11:**610.
36. Abadio AK, Kioshima ES, Teixeira MM, Martins NF, Maigret B, Felipe MS: **Comparative genomics allowed the identification of drug targets against human fungal pathogens.** *BMC genomics* 2011, **12:**75.
37. Zhang R, Ou HY, Zhang CT: **DEG: a database of essential genes.** *Nucleic acids research* 2004, **32**(Database)**:**D271-272.
38. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic acids research* 2000, **28**(1)**:**27-30.
39. Yoon SH, Park YK, Lee S, Choi D, Oh TK, Hur CG, Kim JF: **Towards pathogenomics: a web-based resource for pathogenicity islands.** *Nucleic acids research* 2007, **35**(Database)**:**D395-400.
40. Magrane M, Consortium U: **UniProt Knowledgebase: a hub of integrated protein data.** *Database : the journal of biological databases and curation* 2011, **2011:**bar009.
41. Yu CS, Lin CJ, Hwang JK: **Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions.** *Protein science : a publication of the Protein Society* 2004, **13**(5)**:**1402-1406.
42. Velankar S, Alhroub Y, Best C, Caboche S, Conroy MJ, Dana JM, Fernandez Montecelo MA, van Ginkel G, Golovin A, Gore SP, *et al*.: **PDBe: Protein Data Bank in Europe.** *Nucleic acids research* 2012, **40**(Database)**:**D445-452.
43. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.** *Nucleic acids research* 2004, **32**(Database)**:**D129-133.
44. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J: **CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues.** *Nucleic acids research* 2006, **34**(Web Server)**:**W116-118.
45. Laurie AT, Jackson RM: **Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites.** *Bioinformatics* 2005, **21**(9)**:**1908-1916.
46. Aguero F, Al-Lazikani B, Aslett M, Berriman M, Buckner FS, Campbell RK, Carmona S, Carruthers IM, Chan AW, Chen F, *et al*.: **Genomic-scale prioritization of drug targets: the TDR Targets database.** *Nature reviews Drug discovery* 2008, **7**(11)**:**900-907.
47. Butt AM, Nasrullah I, Tahir S, Tong Y: **Comparative genomics analysis of Mycobacterium ulcerans for the identification of putative essential genes and therapeutic candidates.** *PloS one* 2012, **7**(8)**:**e43080.
48. Volkamer A, Kuhn D, Rippmann F, Rarey M: **DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment.** *Bioinformatics* 2012, **28**(15)**:**2074-2075.
49. Voigt JH, Bienfait B, Wang S, Nicklaus MC: **Comparison of the NCI open database with seven large chemical structural databases.** *Journal of chemical information and computer sciences* 2001, **41**(3)**:**702-712.
50. Thomsen R, Christensen MH: **MolDock: a new technique for high-accuracy molecular docking.** *Journal of medicinal chemistry* 2006, **49**(11)**:**3315-3321.
51. Hopkins AL, Groom CR: **The druggable genome.** *Nature reviews Drug discovery* 2002, **1**(9)**:**727-730.
52. Li Y, Zeng J, He ZG: **Characterization of a functional C-terminus of the Mycobacterium tuberculosis MtrA responsible for both DNA binding and interaction with its two-component partner protein, MtrB.** *Journal of biochemistry* 2010, **148**(5)**:**549-556.
53. Cangelosi GA, Do JS, Freeman R, Bennett JG, Semret M, Behr MA: **The two-component regulatory system mtrAB is required for morphotypic multidrug resistance in Mycobacterium avium.** *Antimicrobial agents and chemotherapy* 2006, **50**(2)**:**461-468.

54. Cotruvo JA, Stubbe J: **NrdI, a flavodoxin involved in maintenance of the diferric-tyrosyl radical cofactor in Escherichia coli class Ib ribonucleotide reductase.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**(38)**:**14383-14388.
55. McAteer S, Coulson A, McLennan N, Masters M: **The lytB gene of Escherichia coli is essential and specifies a product needed for isoprenoid biosynthesis.** *Journal of bacteriology* 2001, **183**(24)**:**7403-7407.
56. Eberl M, Hintz M, Reichenberg A, Kollas AK, Wiesner J, Jomaa H: **Microbial isoprenoid biosynthesis and human gammadelta T cell activation.** *FEBS letters* 2003, **544**(1-3)**:**4-10.
57. Span I, Wang K, Wang W, Zhang Y, Bacher A, Eisenreich W, Li K, Schulz C, Oldfield E, Groll M: **Discovery of acetylene hydratase activity of the iron-sulphur protein IspH.** *Nature communications* 2012, **3:**1042.
58. Plaimas K, Eils R, Konig R: **Identifying essential genes in bacterial metabolic networks with machine learning methods.** *BMC systems biology* 2010, **4:**56.
59. Vinayak S, Sharma YD: **Inhibition of Plasmodium falciparum ispH (lytB) gene expression by hammerhead ribozyme.** *Oligonucleotides* 2007, **17**(2)**:**189-200.
60. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, *et al*.: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.** *Nucleic acids research* 2010, **38**(Database)**:**D473-479.
61. Caffrey CR, Rohwer A, Oellien F, Marhofer RJ, Braschi S, Oliveira G, McKerrow JH, Selzer PM: **A comparative chemogenomics strategy to predict potential drug targets in the metazoan pathogen, Schistosoma mansoni.** *PloS one* 2009, **4**(2)**:**e4413.
62. Crowther GJ, Shanmugam D, Carmona SJ, Doyle MA, Hertz-Fowler C, Berriman M, Nwaka S, Ralph SA, Roos DS, Van Voorhis WC, *et al*.: **Identification of attractive drug targets in neglected-disease pathogens using an in silico approach.** *PLoS neglected tropical diseases* 2010, **4**(8)**:**e804.
63. Brocker M, Mack C, Bott M: **Target genes, consensus binding site, and role of phosphorylation for the response regulator MtrA of Corynebacterium glutamicum.** *Journal of bacteriology* 2011, **193**(5)**:**1237-1249.
64. Zahrt TC, Deretic V: **An essential two-component signal transduction system in Mycobacterium tuberculosis.** *Journal of bacteriology* 2000, **182**(13)**:**3832-3838.
65. Friedland N, Mack TR, Yu M, Hung LW, Terwilliger TC, Waldo GS, Stock AM: **Domain orientation in the inactive response regulator Mycobacterium tuberculosis MtrA provides a barrier to activation.** *Biochemistry* 2007, **46**(23)**:**6733-6743.
66. Lammers M, Follmann H: **The Ribonucleotide Reductases - a Unique Group of Metalloenzymes Essential for Cell-Proliferation.** *Struct Bond* 1983, **54:**27-91.
67. Nordlund P, Reichard P: **Ribonucleotide reductases.** *Annual review of biochemistry* 2006, **75:**681-706.
68. Monje-Casas F, Jurado J, Prieto-Alamo MJ, Holmgren A, Pueyo C: **Expression analysis of the nrdHIEF operon from Escherichia coli. Conditions that trigger the transcript level in vivo.** *The Journal of biological chemistry* 2001, **276**(21)**:**18031-18037.
69. Cotruvo JA, Stubbe J: **An active dimanganese(III)-tyrosyl radical cofactor in Escherichia coli class Ib ribonucleotide reductase.** *Biochemistry* 2010, **49**(6)**:**1297-1309.
70. Elledge SJ, Zhou Z, Allen JB: **Ribonucleotide reductase: regulation, regulation, regulation.** *Trends in biochemical sciences* 1992, **17**(3)**:**119-123.
71. Boal AK, Cotruvo JA, Stubbe J, Rosenzweig AC: **Structural basis for activation of class Ib ribonucleotide reductase.** *Science* 2010, **329**(5998)**:**1526-1530.
72. Tarze A, Deniaud A, Le Bras M, Maillier E, Molle D, Larochette N, Zamzami N, Jan G, Kroemer G, Brenner C: **GAPDH, a novel regulator of the pro-apoptotic mitochondrial membrane permeabilization.** *Oncogene* 2007, **26**(18)**:**2606-2620.
73. Zala D, Hinckelmann MV, Yu H, Lyra da Cunha MM, Liot G, Cordelieres FP, Marco S, Saudou F: **Vesicular glycolysis provides on-board energy for fast axonal transport.** *Cell* 2013, **152**(3)**:**479-491.
74. Bressi JC, Verlinde CL, Aronov AM, Shaw ML, Shin SS, Nguyen LN, Suresh S, Buckner FS, Van Voorhis WC, Kuntz ID, *et al*.: **Adenosine analogues as selective inhibitors of glyceraldehyde-3-phosphate dehydrogenase of Trypanosomatidae via structure-based drug design.** *Journal of medicinal chemistry* 2001, **44**(13)**:**2080-2093.
75. Pal-Bhowmick I, Andersen J, Srinivasan P, Narum DL, Bosch J, Miller LH: **Binding of aldolase and glyceraldehyde-3-phosphate dehydrogenase to the cytoplasmic tails of Plasmodium falciparum merozoite duffy binding-like and reticulocyte homology ligands.** *mBio* 2012., **3**(5)
76. Bellinzoni M, Haouz A, Grana M, Munier-Lehmann H, Shepard W, Alzari PM: **The crystal structure of Mycobacterium tuberculosis adenylate kinase in complex with two molecules of ADP and Mg2+ supports an associative mechanism for phosphoryl transfer.** *Protein science : a publication of the Protein Society* 2006, **15**(6)**:**1489-1493.
77. Appaji Rao N, Ambili M, Jala VR, Subramanya HS, Savithri HS: **Structure-function relationship in serine hydroxymethyltransferase.** *Biochimica et biophysica acta* 2003, **1647**(1-2)**:**24-29.
78. Sopitthummakhun K, Thongpanchang C, Vilaivan T, Yuthavong Y, Chaiyen P, Leartsakulpanich U: **Plasmodium serine hydroxymethyltransferase as a potential anti-malarial target: inhibition studies using improved methods for enzyme production and assay.** *Malaria journal* 2012, **11:**194.
79. Mechaly AE, Haouz A, Miras I, Barilone N, Weber P, Shepard W, Alzari PM, Bellinzoni M: **Conformational changes upon ligand binding in the essential class II fumarase Rv1098c from Mycobacterium tuberculosis.** *FEBS letters* 2012, **586**(11)**:**1606-1611.

80. Woods SA, Schwartzbach SD, Guest JR: **Two biochemically distinct classes of fumarase in Escherichia coli.** *Biochimica et biophysica acta* 1988, **954**(1)**:**14-26.
81. Sampaleanu LM, Vallee F, Slingsby C, Howell PL: **Structural studies of duck delta 1 and delta 2 crystallin suggest conformational changes occur during catalysis.** *Biochemistry* 2001, **40**(9)**:**2732-2742.
82. Yang J, Wang Y, Woolridge EM, Arora V, Petsko GA, Kozarich JW, Ringe D: **Crystal structure of 3-carboxy-cis,cis-muconate lactonizing enzyme from Pseudomonas putida, a fumarase class II type cycloisomerase: enzyme evolution in parallel pathways.** *Biochemistry* 2004, **43**(32)**:**10424-10434.
83. Toth EA, Yeates TO: **The structure of adenylosuccinate lyase, an enzyme with dual activity in the de novo purine biosynthetic pathway.** *Structure* 2000, **8**(2)**:**163-174.
84. Tsai M, Koo J, Yip P, Colman RF, Segall ML, Howell PL: **Substrate and product complexes of Escherichia coli adenylosuccinate lyase provide new insights into the enzymatic mechanism.** *Journal of molecular biology* 2007, **370**(3)**:**541-554.
85. Weaver TM, Levitt DG, Donnelly MI, Stevens PP, Banaszak LJ: **The multisubunit active site of fumarase C from Escherichia coli.** *Nature structural biology* 1995, **2**(8)**:**654-662.
86. Chen Z, Zhou Q, Ge R: **Inhibition of fumarase by bismuth(III): implications for the tricarboxylic acid cycle as a potential target of bismuth drugs in Helicobacter pylori.** *Biometals : an international journal on the role of metal ions in biology, biochemistry, and medicine* 2012, **25**(1)**:**95-102.
87. Ruda GF, Campbell G, Alibu VP, Barrett MP, Brenk R, Gilbert IH: **Virtual fragment screening for novel inhibitors of 6-phosphogluconate dehydrogenase.** *Bioorganic & medicinal chemistry* 2010, **18**(14)**:**5056-5062.
88. Shi W, Dunbar J, Jayasekera MM, Viola RE, Farber GK: **The structure of L-aspartate ammonia-lyase from Escherichia coli.** *Biochemistry* 1997, **36**(30)**:**9136-9144.
89. de Villiers M, Puthan Veetil V, Raj H, de Villiers J, Poelarends GJ: **Catalytic mechanisms and biocatalytic applications of aspartate and methylaspartate ammonia lyases.** *ACS chemical biology* 2012, **7**(10)**:**1618-1628.

II.VI - Curriculum Vitae

Edson Luiz Folador

Curriculum Vitae

Junho/2015

II.VI.I - Dados pessoais

Nome Edson Luiz Folador
Filiação Eloi Nelso Folador e Jadviga Kinga Folador
Nascimento 23/11/1972 - Cascavel/PR - Brasil
Identidade 19958749 PC - MG - 25/09/2012
CPF 528.696.521-00

II.VI.II - Formação acadêmica/titulação

2013 - Atual Doutorado em Bioinformática.
Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Brasil
Título: Predição e análise comparativa da rede de interação proteína-proteína para os biovares *ovis* e *equi* de Corynebacterium pseudotuberculosis
Orientador: Vasco Ariston de Carvalho Azevedo
Bolsista do(a): Conselho Nacional de Desenvolvimento Científico e Tecnológico

2006 - 2008 Mestrado em Tecnologia em Saúde.
Pontifícia Universidade Católica do Paraná, PUC/PR, Curitiba, Brasil
Título: GO-SIEVe: Software para determinar códigos de evidência em anotação gênica, Ano de obtenção: 2008
Orientador: Humberto Maciel França Madeira

2003 - 2004 Especialização em Desenvolvimento de Sistemas Web e Apoio a Decisão.
Universidade Paranaense, UNIPAR, Umuarama, Brasil
Título: Bancos de Dados Relacionais: Um Estudo da Viabilidade de utilização de Tabela Resumo
Orientador: Angelo Alfredo Sucolotti

1999 - 2002 Graduação em Sistemas de Informação.
Universidade Paranaense, UNIPAR, Umuarama, Brasil
Título: Desenvolvimento Sistema Controle Financeiro
Orientador: Angelo Alfredo Sucolloti

II.VI.III - Formação complementar

2014 - 2014 Curso de curta duração em PATRIC: Recursos integrados estudo patogenicidade.
Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Brasil
Bolsista do(a): Conselho Nacional de Desenvolvimento Científico e Tecnológico

2014 - 2014 Extensão universitária em Formação em Docência do Ensino Superior.
Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Brasil

2014 - 2014 Curso de curta duração em Practical Bioinformatics on Gene Functional Netwok.
Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Brasil
Bolsista do(a): Conselho Nacional de Desenvolvimento Científico e Tecnológico

2012 - 2012 Curso de curta duração em Montagem, anotação e extração dados transcriptoma.
Centro de Pesquisa René Rachou, CPQRR, Brasil

2012 - 2012 Curso de curta duração em RNAseq.
Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Brasil

2011 - 2011 Curso de curta duração em Técnicas para montagem e análise de genomas.
Universidade Estadual de Campinas, UNICAMP, Campinas, Brasil

2010 - 2010 Curso de curta duração em Curso de verão em bioinformática.
Universidade de São Paulo, USP, Sao Paulo, Brasil

2005 - 2005 Curso de curta duração em Formação de Tutores Moodle.
Universidade de Brasília, UNB, Brasília, Brasil

2002 - 2002 Curso de curta duração em Data Warehouse.
Universidade Paranaense, UNIPAR, Umuarama, Brasil

2002 - 2002 Curso de curta duração em Php.
Universidade Paranaense, UNIPAR, Umuarama, Brasil

2002 - 2002 Curso de curta duração em Montagem e Manutenção de Computadores.
Universidade Paranaense, UNIPAR, Umuarama, Brasil

2002 - 2002 Curso de curta duração em Interbase.
Universidade Paranaense, UNIPAR, Umuarama, Brasil

2001 - 2001 Curso de curta duração em Tcp Ip.
Universidade Paranaense, UNIPAR, Umuarama, Brasil

2001 - 2001 Curso de curta duração em Recursos Informática Aplicados Ensino de Biologia.
Universidade Paranaense, UNIPAR, Umuarama, Brasil

1999 - 1999 Curso de curta duração em Redes e Telecomunicações.
Universidade Paranaense, UNIPAR, Umuarama, Brasil

1999 - 1999 Curso de curta duração em Modelo de Arquitetura de Sistemas de Informação.
Universidade Paranaense, UNIPAR, Umuarama, Brasil

1999 - 1999 Curso de curta duração em Métricas Sobre Internet.
Universidade Paranaense, UNIPAR, Umuarama, Brasil

1998 - 1998 Curso de curta duração em Como Calcular Custo e Preço de Venda no Comércio.
Serviço Brasileiro de Apoio às Micro e Pequenas Empresas, SEBRAE, Brasília, Brasil

1997 - 1998 Língua Espanhola.
Centro de Línguas Estrangeiras Modernas, CELEM, Brasil

1993 - 1993 Curso de curta duração em Criatividade Em Vendas.
Serviço Nacional de Aprendizagem Comercial, SENAC, Brasil

1993 - 1993 Curso de curta duração em Como Implantar Os Controles Financeiros Básicos na.
Serviço Brasileiro de Apoio às Micro e Pequenas Empresas, SEBRAE, Brasília, Brasil

1993 - 1993 Curso de curta duração em Como Calcular Os Custos e Formar Preços de Venda.
Serviço Brasileiro de Apoio às Micro e Pequenas Empresas, SEBRAE, Brasília, Brasil

1992 - 1992 Curso de curta duração em Técnica de Atendimento e Motivação Em Vendas.
Serviço Nacional de Aprendizagem Comercial, SENAC, Brasil

II.VI.IV - Atuação profissional

1. Universidade Federal de Minas Gerais - UFMG

Vínculo institucional

2013 - Atual Vínculo: Bolsista, Enquadramento funcional: Analista em Bioinformática, Carga horária: 40, Regime: Dedicação exclusiva

Atividades

08/2014 - Atual Pesquisa e Desenvolvimento, Instituto de Ciências Biológicas
Linhas de pesquisa:
Predição e análise comparativa da rede de interação proteína-proteína para 15 linhagens dos biovares ovis e equi de Corynebacterium pseudotuberculosis

08/2013 - Atual Outra atividade técnico-científica, Instituto de Ciências Biológicas
Especificação:
Administração de Sistema de Gerenciamento de Banco de Dados, Curadoria e Anotação funcional de Genomas, Desenvolvimento de rotinas em linguagem PG/pgSQL, Desenvolvimentos de rotinas de computador em linguagem Bash ou Perl para solução de problemas em Bioinformática, Modelagem de Banco de Dados para predição de interação proteína-proteína e transferência de anotação genética

08/2013 - 07/2014 Pesquisa e Desenvolvimento, Instituto de Ciências Biológicas
Linhas de pesquisa:
Validação de metodologia computacional para predição de redes de interação proteína-proteína

2. Centro de Pesquisa René Rachou - CPQRR

Vínculo institucional

2012 - 2013 Vínculo: Bolsista, Enquadramento funcional: Bolsista, Carga horária: 40, Regime: Dedicação exclusiva

Atividades

03/2012 - 06/2012 Treinamento, LPCM
Especificação:
Lógica de programação para Bioinformática com exemplos práticos na linguagem de programação Perl

03/2012 - 05/2013 Serviço Técnico Especializado, LPCM
Especificação:
Administração e modelagem do bando de dados de predição de epítopos, Administração e modelagem dos bancos de dados do laboratório de Bioinformática, Desenvolvimento de rotinas de Bioinformática nas linguagens de programação C, Perl, Php

3. Instituto Nacional de Câncer - INCA

Vínculo institucional

2009 - 2011 Vínculo: Bolsista CNPQ DTI-1, Enquadramento funcional: Analista em Bioinformática, Carga horária: 40, Regime: Dedicação exclusiva

Atividades

11/2009 - 11/2009 Pós-graduação, Programa de Pós-Graduação em Oncologia (PPGO)
Disciplinas ministradas:
Introdução a Bioinformática (Módulo de Bando de Dados)

03/2009 - 06/2012 Serviço Técnico Especializado, Coordenação de Pesquisa, Laboratório de Bioinformática e Biologia Computacional (LBBC)
Especificação:
Desenvolvimento de aplicações e rotinas principalmente nas linguagens de programação Perl, PHP e HTML., Desenvolvimento de um Sistema de Gerenciamento de Informações para Laboratório (LIMS) de proteômica

03/2009 - 02/2012 Serviço Técnico Especializado, Coordenação de Pesquisa, Laboratório de Bioinformática e Biologia Computacional (LBBC)
Especificação:
Administração de Banco de Dados (DBA): instalação, configuração, gerenciamento e modelagem das bases de dados sob o Sistema de Gerenciamento de Banco de Dados (SGBD) Postgres.

4. Instituto de Estudos Avançados e Pós-Graduação - ESAP

Vínculo institucional

2006 - 2008 Vínculo: Celetista formal, Enquadramento funcional: Professor títular, Carga horária: 8, Regime: Parcial

Atividades

07/2008 - 12/2008 Graduação, Sistema de informação
Disciplinas ministradas:
Projeto e Análise de Algoritmos II

02/2008 - 07/2008 Graduação, Sistema de informação
Disciplinas ministradas:
Projeto e Análise de Algoritmos I

10/2007 - 12/2008 Direção e Administração, Curso Sistemas de Informação
Cargos ocupados:
Coordenador de Curso

08/2007 - 12/2007 Graduação, Sistema de informação
Disciplinas ministradas:
Banco de Dados I

02/2007 - 06/2007 Graduação, Sistema de informação
Disciplinas ministradas:
Banco de Dados II

02/2007 - 06/2007 Graduação, Administração
Disciplinas ministradas:
Recursos Computacionais II

07/2006 - 12/2006 Graduação, Sistema de informação
Disciplinas ministradas:
Engenharia de Software I

5. Universidade Estadual do Oeste do Paraná - UNIOESTE

Vínculo institucional

2005 - 2005 Vínculo: Colaborador, Enquadramento funcional: Colaborador em projeto de pesquisa, Carga horária: 2, Regime: Parcial
2003 - 2005 Vínculo: Colaborador, Enquadramento funcional: Professor titular, Carga horária: 24, Regime: Parcial

Atividades

07/2004 - 07/2004 Conselhos, Comissões e Consultoria, Conselho de Ensino, Pesquisa e Extensão
Especificação:
Banca Avaliadora Monitoria Disciplina Engenharia de Software

01/2004 - 12/2004 Graduação, Engenharia Agrícola
Disciplinas ministradas:
Processamento de Dados

01/2004 - 12/2004 Graduação, Engenharia Civil
Disciplinas ministradas:
Introdução a Computação

01/2004 - 12/2004 Graduação, Informática
Disciplinas ministradas:
Banco de Dados I

07/2003 - 12/2003 Graduação, Informática
Disciplinas ministradas:
Algoritmos e Estrutura de Dados, Engenharia de software

07/2003 - 12/2003 Graduação, Engenharia Civil
Disciplinas ministradas:
Introdução a Computação

6. União Panamericana de Ensino - UNIPAN

Vínculo institucional

2004 - 2007 Vínculo: Outro, Enquadramento funcional: Professor titular, Carga horária: 4, Regime: Parcial

Atividades

01/2007 - 07/2007 Graduação, Ciência da Computação
Disciplinas ministradas:
Pesquisa e Ordenação de Dados

01/2006 - 12/2006 Graduação, Ciência da Computação
Disciplinas ministradas:
Banco de Dados, Pesquisa e Ordenação de Dados

01/2005 - 12/2005 Graduação, Ciência da Computação
Disciplinas ministradas:
Estrutura, Pesquisa e Ordenação de Dados, Banco de Dados

03/2004 - 12/2004 Graduação, Ciência da Computação
Disciplinas ministradas:
Estrutura, Pesquisa e Ordenação de Dados - C

7. União Educacional do Médio Oeste Paranaense Ltda - UNIMEO

Vínculo institucional

2004 - 2004 Vínculo: Outro, Enquadramento funcional: Professor titular, Carga horária: 8, Regime: Parcial

Atividades

07/2004 - 10/2004 Graduação, Sistema de Informação
Disciplinas ministradas:
Pesquisa e Ordenação de Dados - C

02/2004 - 06/2004 Graduação, Sistema de Informação
Disciplinas ministradas:
Projeto e Análise de Dados Orientado a Objeto, Estrutura de Dados - C

8. Maxicon System Ltda - MAXICON

Vínculo institucional

2002 - 2003 Vínculo: Funcionário, Enquadramento funcional: Programador Sênior, Carga horária: 44, Regime: Dedicação exclusiva
2001 - 2002 Vínculo: Estagiário, Enquadramento funcional: Programador, Carga horária: 40, Regime: Integral

Atividades
07/2001 - 02/2003 Serviço Técnico Especializado, Desenvolviemnto de sistemas
Especificação:
Análise e desenvolvimento de sistema sob BD Oracle com Front End Forms 6.0 e Linguagem de programação PL/SQL

9. Salgado & Haddad Ltda - CDI

Vínculo institucional

1995 - 1996 Vínculo: Funcionário, Enquadramento funcional: Instrutor Informática, Carga horária: 20, Regime: Parcial

Atividades
08/1995 - 09/1996 Treinamento
Especificação:
Treinamento Aplicativo Word, Excel, Power Point

10. Comercial de Calçados Âncora Ltda - ÂNCORA

Vínculo institucional

1992 - 1995 Vínculo: Funcionário, Enquadramento funcional: Gerente, Carga horária: 44, Regime: Integral

Atividades
02/1992 - 03/1995 Direção e Administração
Cargos ocupados:
Gerente

11. Grisa & Grisa Ltda - GRISA

Vínculo institucional

1989 - 1991 Vínculo: Funcionário, Enquadramento funcional: Vendedor Interno, Carga horária: 44, Regime: Integral

Atividades

03/1989 - 02/1991 Serviço Técnico Especializado
Especificação:
Vendedor Balconista, Crediarista

II.VI.V - Linhas de pesquisa

1. Predição e análise comparativa da rede de interação proteína-proteína para 15 linhagens dos biovares *ovis* e *equi* de Corynebacterium pseudotuberculosis

2. Validação de metodologia computacional para predição de redes de interação proteína-proteína

II.VI.VI - Projetos

Projetos de pesquisa

2015 - Atual Estudo do interatoma e exossoma em Corynebacterium pseudotuberculosis para pesquisa de novos alvos terapêuticos
Descrição: Existe uma dificuldade na eliminação da C. pseudotuberculosis por macrófagos, e desvendar como ocorre a interação entre patógeno e hospedeiro, conhecer a cascata de resposta em nível transcricional, nos dois organismos simultaneamente, bem como elucidar o efeito do exossoma secretado na resposta imune do hospedeiro, abriria um leque de tentativas para busca de soluções eficazes contra este problema enfrentado. Tanto o patógeno quanto o hospedeiro buscam uma resposta rápida, adaptativa, eficaz para a própria sobrevivência. Assim, perceber a alteração no ambiente e transmitir a informação montando uma rede de resposta ideal é o ponto chave para entender todo o processo para manutenção dos organismos no ambiente. Chamada de projetos MEC/MCTI/CAPES/CNPq/FAPs nº 09/2014.
Situação: Em andamento Natureza: Projetos de pesquisa
Alunos envolvidos: Mestrado acadêmico (4); Doutorado (2);
Integrantes: Edson Luiz Folador; Adriana Ribeiro Carneiro (Responsável)

2013 - Atual Rede de cooperação acadêmica para o estudo e desenvolvimento de ferramentas para a genômica Estrutural e Funcional
Descrição: Fortalecer e ampliar o intercâmbio acadêmico entre os programas inter-unidades de Pós-Graduação em Bioinformática da UFMG (CAPES 6) e da USP (5), o de Biotecnologia da UFPA (CAPES 5) e o de Bioinformática da UFPR (CAPES 3) com a criação de uma rede voltada a aumentar a formação de recursos humanos em Biologia Computacional, em resposta à presente chamada. Edital nº 51/2013 BIOLOGIA COMPUTACIONAL.
Situação: Em andamento Natureza: Projetos de pesquisa
Alunos envolvidos: Mestrado acadêmico (7); Doutorado (6);

Integrantes: Edson Luiz Folador; HASSAN, SYED SHAH; TIWARI, SANDEEP; ALMEIDA, SINTIA; OLIVEIRA, ALBERTO; Diego Cesar Batista Mariano; Letícia C. Oliveira; Vinicius Augusto Carvalho de Abreu; Vasco Azevedo (Responsável); Rafaela Salgado Ferreira

---

II.VI.VII - Produção bibliográfica

Artigos completos publicados em periódicos

1. **FOLADOR EL**, OLIVEIRA, ALBERTO, TIWARI, SANDEEP, JAMAL, SYED BABAR, FERREIRA, R. S., BARH, D., Ghosh, P., SILVA, A., AZEVEDO, V.
In silico protein-protein interactions: avoiding data and method biases over sensitivity and specificity. Current Protein and Peptide Science., v.16, p.1 -, 2015.

2. **FOLADOR, EDSON LUIZ**, HASSAN, SYED SHAH, LEMKE, NEY, BARH, DEBMALYA, SILVA, ARTUR, FERREIRA, RAFAELA SALGADO, AZEVEDO, VASCO
An improved interolog mapping-based computational prediction of protein-protein interactions with increased network coverage. Integrative Biology., v.6, p.1080 - 1087, 2014.

3. SILVA, WANDERSON M, CARVALHO, RODRIGO D, SOARES, SIOMAR C, BASTOS, ISABELA FS, **FOLADOR, EDSON L**, SOUZA, GUSTAVO HMF, LE LOIR, YVES, MIYOSHI, ANDERSON, SILVA, ARTUR, AZEVEDO, VASCO
Label-free proteomic analysis to confirm the predicted proteome of Corynebacterium pseudotuberculosis under nitrosative stress mediated by nitric oxide. BMC Genomics., v.15, p.1065 -, 2014.

4. TIWARI, SANDEEP, DA COSTA, MARCÍLIA PINHEIRO, ALMEIDA, SINTIA, HASSAN, SYED SHAH, JAMAL, SYED BABAR, OLIVEIRA, ALBERTO, **FOLADOR, EDSON LUIZ**, ROCHA, FLAVIA, DE ABREU, VINÍCIUS AUGUSTO CARVALHO, DORELLA, FERNANDA, HIRATA, RAFAEL, DE OLIVEIRA, DIANA MAGALHAES, DA SILVA TEIXEIRA, MARIA FÁTIMA, SILVA, ARTUR, BARH, DEBMALYA, AZEVEDO, VASCO
C. pseudotuberculosis Phop confers virulence and may be targeted by natural compounds. Integrative Biology., v.9, p.1 - 12, 2014.

5. HASSAN, S. S., TIWARI, SANDEEP, GUIMARÃES, LUIS CARLOS, JAMAL, SYED BABAR, **FOLADOR, EDSON LUIZ**, SHARMA, N. B., SOARES, SIOMAR DE CASTRO, ALMEIDA, SINTIA, ALI, A., ISLAM, A., POVOA, F. D., ABREU, V. A. C., JAIN, N., BHATTACHARYA, A., JUNEJA, L., MIYOSHI, A., SILVA, A., BARH, D., TURJANSKI, A. G., AZEVEDO, V., FERREIRA, R. S.
Proteome scale comparative modeling for conserved drug and vaccine targets identification in Corynebacterium pseudotuberculosis. BMC Genomics., v.15, p.S3 -, 2014.

6. REZENDE, ANTONIO M., **FOLADOR, EDSON L.**, RESENDE, DANIELA DE M., RUIZ, J. C.
Computational Prediction of Protein-Protein Interactions in Leishmania Predicted Proteomes. Plos One., v.7, p.e51304 -, 2012.

7. BARAUNA, R. A., GUIMARAES, L. C., VERAS, A. A. O., DE SA, P. H. C. G., GRACAS, D. A., PINHEIRO, K. C., SILVA, A. S. S., **FOLADOR, E. L.**, BENEVIDES, L. J., VIANA, M. V. C., CARNEIRO, A. R., SCHNEIDER, M. P. C., SPIER, S. J., EDMAN, J. M., RAMOS, R. T. J., AZEVEDO, V., SILVA, A.
Genome Sequence of Corynebacterium pseudotuberculosis MB20 bv. *equi* Isolated from a Pectoral Abscess of an Oldenburg Horse in California. Genome Announcements., v.2, p.e00977-14 - e00977-14, 2014.

8. BENEVIDES, LEANDRO DE JESUS, VIANA, MARCUS VINICIUS CANÁRIO, MARIANO, DIEGO CÉSAR BATISTA, ROCHA, FLÁVIA DE SOUZA, BAGANO, PRISCILLA CAROLINNE, **FOLADOR, EDSON LUIZ**, PEREIRA, FELIPE LUIZ, DORELLA, FERNANDA ALVES, LEAL, CARLOS AUGUSTO GOMES, CARVALHO, ALEX FIORINI, SOARES, SIOMAR DE CASTRO, CARNEIRO, ADRIANA,

RAMOS, ROMMEL, BADELL-OCANDO, EDGAR, GUISO, NICOLE, SILVA, ARTUR, FIGUEIREDO, HENRIQUE, AZEVEDO, VASCO, GUIMARÃES, LUIS CARLOS
Genome Sequence of Corynebacterium ulcerans Strain FRC11. Genome Announcements., v.3, p.e00112-15 -, 2015.

9. VIANA, M. V. C., DE JESUS BENEVIDES, L., BATISTA MARIANO, D. C., DE SOUZA ROCHA, F., BAGANO VILAS BOAS, P. C., **FOLADOR, E. L.**, PEREIRA, F. L., ALVES DORELLA, F., GOMES LEAL, C. A., FIORINI DE CARVALHO, A., SILVA, A., DE CASTRO SOARES, S., PEREIRA FIGUEIREDO, H. C., AZEVEDO, V., GUIMARAES, L. C.
Genome Sequence of Corynebacterium ulcerans Strain 210932. Genome Announcements., v.2, p.e01233-14 - e01233-14, 2014.

10. OLIVEIRA, L C, SARAIVA, T D L, SOARES, S C, RAMOS, R T J, SA, P H C G, CARNEIRO, A R, MIRANDA, F, FREIRE, M, RENAN, W, JUNIOR, A F O, SANTOS, A R, PINTO, A C, SOUZA, B M, CASTRO, C P, DINIZ, C A A, ROCHA, C S, MARIANO, D C B, DE AGUIAR, E L, **FOLADOR, E L,** BARBOSA, E G V, ABURJAILE, F F, GONCALVES, L A, GUIMARAES, L C, AZEVEDO, M, AGRESTI, P C M, SILVA, R F, TIWARI, S, ALMEIDA, S S, HASSAN, S S, PEREIRA, V B, ABREU, V A C, PEREIRA, U P, DORELLA, F A, CARVALHO, A F, PEREIRA, F L, LEAL, C A G, FIGUEIREDO, H C P, SILVA, A, MIYOSHI, A, AZEVEDO, V
Genome Sequence of Lactococcus lactis subsp. lactis NCDO 2118, a GABA-Producing Strain. Genome Announcements., v.2, p.e00980-14 - e00980-14, 2014.

11. TAVARES, RAPHAEL, SCHERER, NICOLE DE MIRANDA, PAULETTI, BIANCA ALVES, ARAÚJO, ELÓI, **FOLADOR, EDSON LUIZ**, ESPINDOLA, GABRIEL, Ferreira, Carlos Gil, LEME, ADRIANA FRANCO PAES, DE OLIVEIRA, PAULO SERGIO LOPES, Passetti, Fabio
SpliceProt: a protein sequence repository of predicted human splice variants. Proteomics (Weinheim. Print)., v.14, p.181 - 185, 2014.

12. Santos, Paula F, Santos, Paula F, Ruiz, Jerônimo C, Soares, Rodrigo PP, Moreira, Douglas S, Rezende, Antônio M, **Folador, Edson L**, Oliveira, Guilherme C, Romanha, Alvaro J, Murta, Silvane MF, Oliveira, Guilherme C, Ruiz, Jerônimo C, Rezende, Antônio M, Soares, Rodrigo PP, Murta, Silvane MF, Moreira, Douglas S, Folador, Edson L, Romanha, Alvaro J
Molecular characterization of the hexose transporter gene in benznidazole resistant and susceptible populations of Trypanosoma cruzi. Parasites & Vectors., v.5, p.161 - 186, 2012.

13. WAJNBERG, G., BRAIT, M., **FOLADOR, E.L.**, PARRELLA, P., CAIMS, P., BARBANO, R., FERREIRA, C.G., PASSETTI, F., SIDRANSKY, D., HOQUE, M.O.
573 Copy Number Variation Analysis for Identification of Novel Disease-related Regions in Bladder Cancer. European Journal of Cancer., v.48, p.S136 -, 2012.

14. Renaud, Gabriel, Neves, Pedro, Folador, Edson L, Ferreira, Carlos Gil, Passetti, Fabio
Segtor: Rapid Annotation of Genomic Coordinates and Single Nucleotide Variations Using Segment Trees. Plos One., v.6, p.e26715 -, 2011.

15. BIDARRA, Jorge, Folador, Edson L, CAVASIN, Rodrigo José, MARCON, Marlon
xListas - Um léxico eletrônico para a Língua Portuguesa. Línguas & Letras (UNIOESTE)., v.1, p.6 - 6, 2005.

Capítulos de livros publicados

1. ABURJAILE, F. F., SANTANA, M. P., VIANA, M. V. C., SILVA, WANDERSON M, **FOLADOR EL**, SILVA, A., AZEVEDO, V.
Genomics In: A Textbook of Biotechnology.1 ed.Irving, TX 75039, USA : SM Online Publishers LLC, 2015, v.1, p. 32-50.

Trabalhos publicados em anais de eventos (resumo)

1. Folador, Edson L, Gomes, Renata B., Neves, Pedro, Renaud, Gabriel, Ferreira, Carlos Gil, Abdelhay, Eliane, Passetti, Fabio

pLIMS: an innovative approach to manage and analyze 2D/1D protein gel In: International Workshop on Genomic Databases - IWGD, 2010, Buzios.
IWGD'10 Abstracts book., 2010.

2. Folador, Edson L, Gomes, Renata B., Neves, Pedro, Renaud, Gabriel, Ferreira, Carlos Gil, Abdelhay, Eliane, Passetti, Fabio
PLIMS: A Bioinformatic tool for the 2D/1D protein gel electrophoresis experiments management and analysis In: X-meeting, 2009, Angra dos Reis.
X-meeting abstracts book 2009., 2009.

3. Folador, Edson L, SUCOLOTTI, Angelo A.
Estudo da Viabilidade do Uso de Tabelas Resumo em Banco de Dados Relacional In: III Encontro de iniciação Científica, III Fórum de Pesquisa, 2004, Umuarama.
3º Encontro de Iniciação Científica e Fórum de Pesquisa. Unipar - Umuarama - PR: DEGPP/Unipar, 2004. v.3. p.249 - 250

II.VI.VIII - Apresentação de trabalho e palestra

1. VIANA, M. V. C., BENEVIDES, L. J., MARIANO, D. C. B., ROCHA, FLAVIA, **FOLADOR, E. L.**, PEREIRA, F. L., DORELLA, F. A., LEAL, C. A. G., CARVALHO, A. F., SILVA, A., SOARES, S. C., FIGUEIREDO, H. C. P., AZEVEDO, V., GUIMARAES, L. C.
Complete genome sequence of Corynebacterium ulcerans strain 210932, 2014. (Congresso, Apresentação de Trabalho)

2. BENEVIDES, L. J., VIANA, M. V. C., MARIANO, D. C. B., ROCHA, FLAVIA, **FOLADOR, E. L.**, PEREIRA, F. L., DORELLA, F. A., CARVALHO, A. F., LEAL, C. A. G., SILVA, A., SOARES, S. C., FIGUEIREDO, H. C. P., AZEVEDO, V., GUIMARAES, L. C.
Complete genome sequence of Corynebacterium ulcerans 210931, 2014. (Seminário, Apresentação de Trabalho)

3. Mariano, D. C. B, OLIVEIRA, L. C., **Folador EL**, DE AGUIAR, E. L., BENEVIDES, L. J., PEREIRA, F. L., RAMOS, R. T. J., AZEVEDO, V.
SIMBA: A web tools for complete assembly of bacterial genomes, 2014. (Congresso, Apresentação de Trabalho)

4. Folador, Edson L, Gomes, Renata B., Neves, Pedro, Renaud, Gabriel, Ferreira, Carlos Gil, Abdelhay, Eliane, Passetti, Fabio
Current status of the pLIMS project: a Bioinformatics tool to promote collaborative 1D/2D-PAGE proteomics experiments, 2011. (Congresso, Apresentação de Trabalho)

5. Madeira, Humberto M. F., MAlucelli, Andreia, **Folador, Edson L**
GO-SIEVE - A method to aid the assignment of evidence codes in genome annotations, 2010. (Congresso, Apresentação de Trabalho)

6. Folador, Edson L, Gomes, Renata B., Renaud, Gabriel, Neves, Pedro, Ferreira, Carlos Gil, Passetti, Fabio
pLIMS: uma abordagem inovadora para gerenciamento e análise de experimentos em gel de eletroforeses 2D/1D de proteína para projetos colaborativos, 2010. (Congresso, Apresentação de Trabalho)

7. Folador, Edson L, Gomes, Renata B., Renaud, Gabriel, Neves, Pedro, Ferreira, Carlos Gil, Passetti, Fabio
pLIMS: Ferramenta de bioinformática para gerenciamento e análise de experimentos em gel de eletroforese 1D/2D, 2009. (Congresso, Apresentação de Trabalho)

8. Folador, Edson L, MAlucelli, Andreia, Madeira, Humberto M. F.

GO-SIEV – Software system for inferring annotation evidence from already annotated genes, 2007. (Congresso, Apresentação de Trabalho)

II.VI.IX - Programa de computador sem registro

1. Folador, Edson L, Passetti, Fabio
pLIMS: uma abordagem inovadora para gerenciamento e análise de experimentos em gel de eletroforeses 2D/1D para projetos colaborativos, 2009

2. **Folador, Edson L**
GO-SIEVe - Software para inferir códigos de evidência em anotação genética, 2008

3. **Folador, Edson L**
Sistema de Controle de Auto Peças, 2001

4. **Folador, Edson L**
Sistema de Cotrole para pedidos de Compras Bibliográficas, 2001

Demais produções técnicas

1. **Folador EL**
Introdução a Bioinformática, 2012. (Extensão, Curso de curta duração ministrado)

2. **Folador EL**
O uso de ferramentas de Bioinformática para a inovação científica em Oncologia, 2012. (Extensão, Curso de curta duração ministrado)

3. Passetti, Fabio, **Folador, Edson L**
I Curso prático de introdução à programação para Bioinformática, 2011. (Extensão, Curso de curta duração ministrado)

II.VI.X - Orientações e Supervisões

Orientações e supervisões concluídas

Trabalhos de conclusão de curso de graduação

1. Jeferson do Nascimento. **Aplicação de data mining na busca de padrões de dados referente à criminalidade no município de Cascavel**. 2006. Curso (Ciência da Computação) - União Pan-Americana de Ensino

II.VI.XI - Eventos

Participação em eventos

1. Apresentação de Poster / Painel no(a) **X-Meeting**, 2014. (Congresso)
SIMBA: A web tools for complete assembly of bacterial genomes.

2. **Publications Ethics and Optimizing your Chances of Acceptance in Journals**, 2014. (Seminário).

3. **X-Meeting**, 2013. (Congresso).

4. Apresentação de Poster / Painel no(a) **X-Meeting**, 2011. (Congresso)
Current status of the pLIMS project: a Bioinformatics tool to promote collaborative 1D/2D-PAGE proteomics experiments.

5. **III Fórum de Integração dos Alunos de Pós-Graduação**, 2011. (Encontro).

6. **Curso de Bioinformática - Algoritmos e técnicas computacionais para montagem e análise de genomas.**, 2011. (Seminário).

7. Apresentação de Poster / Painel no(a) **X-Meeting**, 2010. (Congresso)
GO-SIEVE - A METHOD TO AID THE ASSIGNMENT OF EVIDENCE CODES IN GENOME ANNOTATIONS.

8. Apresentação Oral no(a) **International Workshop on Genomic Databases - IWGD**, 2010. (Congresso)
pLIMS: uma abordagem inovadora para gerenciamento e análise de experimentos em gel de eletroforeses 2D/1D de proteína para projetos colaborativos.

9. **Curso de verão em bioinformática (USP)**, 2010. (Seminário).

10. Apresentação de Poster / Painel no(a) **X-meeting**, 2009. (Congresso)
pLIMS: Ferramenta de bioinformática para gerenciamento e análise de experimentos em gel de eletroforese 1D/2D.

11. **GE Day**, 2009. (Encontro).

12. Apresentação Oral no(a) **X-meeting**, 2007. (Congresso)
GO-SIEV - Software system for inferring annotation evidence from already annotated genes.

13. **II EPAC - Encontro Paranaense de Computação**, 2007. (Encontro).

14. **I EPAC - Encontro Paranaense de Computação**, 2005. (Encontro).

15. **3ª Semana de Informática**, 2003. (Encontro).

II.VI.XII - Organização de evento

1. Passetti, Fabio, **Folador, Edson L**
I Curso prático de introdução à programação para Bioinformática, 2011. (Outro, Organização de evento)

2. Kessler, Neivor, Oliveira, Lindomar S., Folador, Edson L, Santos, Vera B.
Empresa Destaque 2007, 2007. (Outro, Organização de evento)

II.VI.XIII - Participação em banca de trabalhos de conclusão

Graduação

1. Konopatzki, Angélica Lima, Gavioli, Alan, **Folador, Edson L**
Participação em banca de Susana Paula Saretto Ferronatto. **Mapeamento tecnológico dos estabelecimentos de ensino médio de Cascavel nas intituições públicas e privadas**, 2007
(Ciência da Computação) União Pan-Americana de Ensino

2. Konopatzki, Angélica Lima, Folador, Edson L, Wagner, Emerson
Participação em banca de Matheus de Lima Boza. **Mineração de dados para definição do perfil da saúde pública em Cascavel com relação às doenças crônicas não-transmissíveis**, 2007
(Ciência da Computação) União Pan-Americana de Ensino

3. Wagner, Emerson, Chrusciak, Daniele, **Folador, Edson L**
Participação em banca de Giancarlo E. C. Fiorenza. **Modelo para implantação de tecnologia da informação em prefeituras municipais de pequeno porte**, 2007
(Ciência da Computação) União Pan-Americana de Ensino

4. Antiquera, Paulo R. da Silva, Folador, Edson L, Chrusciak, Daniele
Participação em banca de Alexandre Magno Semmer. **Persistência em banco de dados relacional para sistemas web**, 2007
(Ciência da Computação) União Pan-Americana de Ensino

5. Piovesan, Suzan Lelly Borges, Gavioli, Alan, **Folador, Edson L**
Participação em banca de Jony Carlos Palaoro. **Protótipo de algoritmo genético para roteamento de rodovias**, 2007
(Ciência da Computação) União Pan-Americana de Ensino

II.VI.XIV - Participação em banca de comissões julgadoras

1. **Processo de Seleção de Monitores**, 2004
Universidade Estadual do Oeste do Paraná

II.VI.XV - Outras informações relevantes

1 Aprovado em 3º lugar no cuncurso público do CEFET/MG para a disciplina de Algoritmos e Programação de Computadores.
Edital geral Nº 149/2014 e Edital específico Nº 62/14.
http://www.jusbrasil.com.br/diarios/72348349/dou-secao-3-30-06-2014-pg-60.
http://pesquisa.in.gov.br/imprensa/servlet/INPDFViewer?jornal=3&pagina=60&data=30/06/2014&captchafield=firistAccess