Universidade Federal de Minas Gerais
Departamento de Bioquímica e Imunologia
Programa de Doutorado em Bioinformática

Laboratório de Biodados

Tese de Doutorado

Mineração de texto, agrupamento de seqüências e integração de dados para o desenvolvimento da Plant Defense Mechanisms Database.

ADRIANO BARBOSA DA SILVA

Orientador: Prof. Dr. José Miguel Ortega - UFMG
Co-orientador: Dr. Reinhard Schneider - EMBL

ADRIANO BARBOSA DA SILVA

"Mineração de texto, agrupamento de seqüências e integração de dados para o desenvolvimento da Plant Defense Mechanisms Database"

Tese apresentada ao Programa de Pós-graduação em Bioinformática da Universidade Federal de Minas Gerais como requisito parcial à obtenção do título de Doutor em Bioinformática.

ÁREA DE CONCENTRAÇÃO: ANÁLISE DE SEQÜÊNCIAS, MINERAÇÃO DE LITERATURA E INTEGRAÇÃO DE DADOS.

Orientador: Dr. José Miguel Ortega
Co-orientador: Dr. Reinhard Schneider

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa de Pós-graduação em Bioinformática
Departamento de Bioquímica e Imunologia
Belo Horizonte – MG
Fevereiro de 2008

Ao meu pai, Sr. Antonio Barbosa da Silva (*in memoriam*).

AGRADECIMENTOS

Agradeço inicialmente à minha mãe por ter me incentivado desde criança a não deixar de estudar, posso dizer que ela é a grande fomentadora da minha crença que a educação é um dos principais pilares de sustentação de uma sociedade mais justa e humana.

Agradeço a Lília por ter estado ao meu lado durante boa parte das principais etapas do meu doutorado, e ter me dado forças para superar os momentos difíceis dessa etapa, pelo seu amor, cuidado, carinho e respeito, e por ter vivido coisas incríveis comigo em vários lugares do Brasil e da Europa.

Agradeço a família da Lília: Sr. Antônio, D. Célia, Liliane, Luciene e Ana Elisa por terem me adotado como filho nos últimos 3 anos e por terem sido minha família durante esse tempo aqui em Belo Horizonte.

Agradeço aos meus professores do segundo grau: Dulcineide que me fez tomar gosto por Biologia e por todos os desafios que nós cientistas enfrentamos ao longo da nossa carreira para estudar os mais diversos fenômenos da natureza; ao meu professor de Física, Ivson, que me fez tomar gosto pela investigação científica, acho que ele foi o cara que me fez descobrir o quão prazeroso é ser curioso com os fenômenos que ocorrem ao nosso redor; e por fim ao grande filósofo das ciências exatas, Valdson, grande professor de matemática que sempre me incentivou pela ciência e sobretudo pela Biologia. Embora eu não mantenha atualmente contato com nenhuma dessas pessoas, sem dúvida lá na escola, bem no começo de tudo, eles foram os responsáveis por eu ter escolhido minha profissão que eu tanto prezo.

Agradeço à minha orientadora da iniciação científica, Profa. Dra. Ana Maria Benko-Iseppon, que me adotou como seu filho científico e que nunca deixou de me incentivar como cientista, mesmo nos momentos mais difíceis como aqueles que eu passei por ocasião da seleção mal-sucedida ao Mestrado em Genética da UFPE, um dos momentos mais tristes da minha carreira científica até o prezado momento. Tal evento me fez acreditar que as coisas não ocorrem por acaso, e que mesmo que não saibamos o porquê dos acontecimentos, eles sempre têm um motivo peculiar que somente ao longo da vida iremos descobrir sua razão.

Agradeço ao Miguelito (Prof. Dr. José Miguel Ortega), meu amigo e orientador da presente tese, por ter me acolhido em Belo Horizonte, ter permitido o meu ingresso em seu laboratório quatro anos atrás quando pouco do que está sendo apresentado nesta Tese existia, por sua confiança em minha teimosia por diversas vezes, e por ter

sido compreensivo nos momentos difíceis que eu certamente vivi ao seu lado. Por outro lado, por ter sido o Miguelito das 22:00 às 05:00, um cara maluco que eu adorei de ter convivido mais fora do que dentro do laboratório, ou sei lá: nos dois lugares! Valeu Miguelito!

Ao Dr. Reinhard Schneider, do EMBL, por ter me recebido de braços abertos em seu laboratório na Alemanha e por ter discutido diversos aspectos dos trabalhos apresentados nesta Tese. E por ter deixado eu trabalhar apenas meio expediente durante a Copa do Mundo de 2006, desde que eu torcesse pela Alemanha caso esta viesse a jogar contra o Brasil, ainda bem que isso não ocorreu.

Agradeço à professora Dra. Glória Franco (UFMG) por todo incentivo dado em diversos momentos ao longo do curso, e por ter facilitado em vários momentos minha vida durante o doutorado, pelo seu companheirismo e ótima convivência ao longo desses quatro anos.

Aos professores: Sérgio Campos, Alfredo Góes, Alessandra Campos, Paulo Beirão, Cristiano Gontijo, Adriano Pimenta, Ronaldo Nagem, Jader Cruz, Marcelo Santoro, Mauro Teixeira, Carlos Renato, Álvaro Eiras e Fabrício Santos pela boa convivência no departamento e fora dele.

Aos membros do Laboratório de Biodados: (Antigos) Maurício, Alessandra, Daniela, João, Estevam, Rosana, Saulo, Denize, (novos) Elisa, Dudu, Lin, Lucas, Gabriel, Bruno, Bellinha, Chico Prós, Igor e Rafael.

Aos amigos e amigas: Cécile, Chico Tosco, Chico Lobo, Cristina Ribeiro, Ciça, Maurício Mudado, Quelé, Cláudia Hollatz, Fê Kedhy, Fê Caldas, Jonny (esquisito), Fernanda Caldas, Ceará, Lúcio, Rodrigo Guabiraba, Rodrigo Ribeiro, Sávio.

Aos amigos do EMBL: Venkata, Georgios, Theo, Vangeli, Anna, Amoolya, Mani, Samuel e Michelle, por terem tornado Heidelberg um lugar *cool* durate meu doutorado sanduíche.

As minhas três mulheres alemães e o cubano: Caro, Ida, Kersting e Joel, meus *mitbewohnen* em Heidelberg.

Aos colegas da UFMG que sempre torceram por mim.

Aos colegas do futebol que sempre tornaram a segunda-feira o melhor dia da semana.

Ao Buteco da Biologia, reduto salvador da sexta-feira.

Enfim, a todos aqueles que de maneira direta ou indireta me ajudaram a concluir mais esta importante etapa na minha vida,

Obrigado!

ÍNDICE

LISTA DE ARTIGOS

No.	Título	Autores	Status/ *Revista*	Pg
1	LAITOR - Literature Assistant for Identification of Terms co-Occurrences and Relationships.	**Barbosa-Silva, A.,** Soldatos, T., Fiorini-Magalhães, I.L., Schneider, R., Ortega, J.M.,	Pubmicado *BMC Bioinformatics*	30
2	Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence.	**Barbosa-Silva, A.,** Satagopam, V., Schneider, R., Ortega, J.M.	Publicado *BMC Bioinformatics*	41
3	Development of SRS.php, a Simple Object Access Protocol-based library for data acquisition from integrated biological databases.	**Barbosa-Silva, A.,** Pafilis, E., Ortega, J.M., Schneider, R.	Publicado *Gen Mol Res*	56
4	Plant Defense Mechanisms Databases	**Barbosa-silva, A.,** Ortega J.M.	Submetido *Bioinformatics*	66
5	*In silico* survey of disease resistance (R) genes in *Eucalyptus* transcriptome	**Barbosa-da-Silva, A.,** Wanderley-Nogueira, A. C., Silva, R. M. R., Berlarmino, L. C., Soares-Cavalcanti, N. M., Benko-Iseppon, A. M.	Publicado *Gen Mol Biol*	72

LISTA DE TABELAS

LISTA DE FIGURAS

...continuação.

Número	Nome	Localização	Identificação	Página
20	Figure 9	Artigo 2	Neighbor joining trees of merged clusters	51
21	Figure 10	Artigo 2	Simulation of a novel and unfinished genome	52
22	Figure 1	Artigo 3	Function performIcarusQueryandGetNumberOfResult	60
23	Figure 2	Artigo 3	Function performQueryAndGetLoaderDefinedFields	61
24	Figure 3	Artigo 3	Function performQueryAndGetSpecificFields	61
25	Figure 4	Artigo 3	Function PerformLinkingQuery	62
26	Figure 5	Artigo 3	Using the SRS.php library	63
27	Fig. 1	Artigo 4	PDM Database	67
28	Figura 8	Resultados e Discussões	Comparação entre métricas de anotação entre as bases PDM e nr	69
29	Figure 1	Artigo 5	Representation of main R-genes classes considering the presence and position of conserved domains from literature data	75
30	Figure 2	Artigo 5	Graphic representation of the distribution of conserved domains against class-grouped clusters	77
31	Figure 3	Artigo 5	Subcellular prediction for each class of analyzed R genes in Eucalyptus transcriptome	78

SIGLAS E ABREVIATURAS

Sigla/Abreviatura	Significado
BBH	*Bidirectional Best Hit*
BLAST	*Basic Local Alignment Search Tool*
BLP	*Biomedical Language Processing*
Cel	*Caenorhabditis elegans*
COG	*Cluster of Orthologous Groups*
Dme	*Drosophila melanogaster*
EST	*Expressed Sequence Tags*
FTP	*File Transfer Protocol*
Hsa	*Homo sapiens*
HTML	*Hypertext Markup Language*
HTTP	*Hypertext Transfer Protocol*
IA	Inteligência Artificial
IE	*Information Extraction*
IR	*Information Retrieval*
LAITOR	*Literature Assistant for Identification of Terms co-Occurrences and Relationships*
MC	*Manual Clusters*
MCL	*Markov Cluster*
NLP	*Natural-Language Processing*
PAM	*Point Accepted Matrix*
PDM	*Plant Defense Mechanisms*
PIR	*Protein International Resource*
RC	*Rebuilt Clusters*
RPC	Chamadas de Procedimento Remoto
SGBD	Sistema Gerenciador do Banco de Dados
SOAP	*Simple Object Access Protocol*
SRS	*Sequence Retrieval System*
TM	*Trans-membrane domains*
TOG	*Tentative Ortholog Groups*
TOGA	*TIGR Orthologous Gene Alignments*
UDDI	*Universal Discovery Description and Integration*
WSDL	*Web Service Description Language*
WWW	*World Wide Web*
XML	*eXtensible Markup Language*

RESUMO

Este trabalho visa descrever as tecnologias utilizadas para o desenvolvimento da base de dados Plant Defense Mechanisms, uma base de dados sobre mecanismos de defesa em plantas contra estresse biótico e abiótico. Para isso desenvolvemos o programa LAITOR para identificar as co-ocorrências de nomes de proteínas e estímulos abióticos (*biointities*) na literatura científica juntamente com termos indicativos de uma ação biológica (*bioactions*), validado aquelas co-ocorrências na mesma frase apenas. A ferramenta NLPROT foi usada para a marcação inicial das *bioentities* que foram *a posteriori* validadas pelo LAITOR. Em seguida, para aqueles termos protéicos pertencentes a base de dados NCBI Gene que possuíam um registro correspondente na base de dados UniProtKB, foi realizado agrupamento de seqüências relacionadas nos outros organismos pertencentes a mesma base de dados, para isso desenvolvemos o software Seed Linkage. Este software explora as ligações múltiplas diretas e indiretas das seqüências desses outros organismos para com a seed inicialmente determinada. Encontramos os parâmetros de escore 400 (bruto) e 0.3 (relativo) como sendo os que maximizam a inclusão de seqüências corretas em clusters manualmente inspecionados. Depois de identificarmos 780 termos protéicos a partir da análise de 7.306 resumos científicos com o programa LAITOR, 1.390 identificadores únicos do UniProtKB foram utilizados para agrupar 15.669 seqüências nos 611 grupos que compõem a PDM. Desenvolvemos uma biblioteca, denominada SRS.php, para adquirir as informações referentes a cada umas destas proteínas a partir do servidor SRS instalado no EMBL utilizando a tecnologia de Web Services. Com o uso desta biblioteca, um cliente SOAP acessa o servidor e recupera, de maneira programática, os dados lá depositados. Depois de efetuarmos a análise de mineração de texto com o programa LAITOR, o agrupamento das seqüências através do método Seed Linkage e a aquisição subseqüente dos dados usando o protocolo SOAP, todas essas informações foram disponibilizadas num servidor HTML no sítio http://www.biodados.icb.ufmg.br/pdm. Neste sítio, os usuários podem efetuar uma busca utilizando palavras-chaves bem como busca por similaridade de seqüência pelo método BLAST. Após terem os registros desejados visualizados, um link é criado para as co-ocorrências dos termos protéicos na análise de mineração de texto, bem como para uma árvore filogenética das proteínas presentes em cada agrupamento da PDM. Além disso, implementamos o servidor SOAP da PDM, que faz com que seus dados sejam distribuídos por meio de Web Services. Criamos um método, denomidado query_pdm, onde todos os registros da base de dados podem ser consultados via SOAP. Em suma, apresentamos uma série de métodos implementados como componentes de softwares e programas propriamente ditos, que podem ser utilizados em aplicações semelhantes aquelas da PDM, sendo, todos eles, distribuídos gratuitamente a comunidade científica interessada nessas técnicas.

ABSTRACT

This work aims to describe the technologies used for the Plant Defense Mechnaisms Database development, a database about the defense mechanisms against biotic and abiotic types of stresses in plants. For this purpose we have developed the program LAITOR, this is used in order to identify in the scientific literature the protein terms and names of biotic and abiotic stimuli (bioentities) along with terms indicating of a biological action (bioaction), nevertheless, validating those occurrences in the same sentence only. The tool NLPROT has been used for the initial bioentities tagging which were validated a posteriori by LAITOR. Later, for those protein terms which belong to the NCBI Gene database and with a corresponding record in the UniProtKB database, it was performed the clustering of sequences belonging to other organisms deposited in the same UniProtKB database, to achieve this aim we developed the Seed Linkage software. This software exploits direct and indirect multiple links from the sequences of these organisms to the initially determined seed. We found that the raw and relative scores of 400 and 0.3, respectively, are those which maximizes the inclusion of correct sequences in the rebuilding of a manually inspected clusters dataset. After the identification of 780 protein terms from the analysis of 7,306 scientific abstracts using the program LAITOR, 1,390 unique UniProtKB identifiers were used to cluster 15,669 sequences in the 611 clusters of the PubMed database. We have developed a software library, named SRS.php, to acquire the information referring to each of these proteins, using for this purpose the SRS server installed at the EMBL using the Web Services technology. With the usage of this library, a SOAP client accesses the server and retrieve, in a programmatic manner, the available data. After to perform the text mining analysis with the program LAITOR, the sequence clustering using the Seed Linkage software, and the subsequent data acquisition using the SOAP protocol, all these information were made available by a HTML server at http://www.biodados.icb.ufmg.br/pdm. In this website, users are able to perform a search using keywords or a BLAST-based similarity search. After the visualization of the retrieved records, a link is created for the co-occurrence of the protein terms in the text mining analysis, as well as for the phylogenetic tree of the proteins grouped in each PDM cluster. Furthermore, we have implemented the PDM SOAP server, which enables the distribution of PDM data through Web Services. We have created a method, named query_pdm, where any record deposited in this database can be accessed using SOAP. Summarizing, we present a set of methods implemented as software components, or programs in fact, which can be used in similar applications to PDM, being, therefore, freely available for the scientific community interested in such techniques.

1. INTRODUÇÃO

1.1 GÊNESIS DA BIOINFORMÁTICA

O uso de computadores como ferramenta nos projetos envolvendo Biologia Molecular já era um fato comum uma década antes do seqüenciamento de DNA tornar-se parte do cotidiano dos laboratórios de pesquisa (Boguski, 1998). Entretanto, o termo Bioinformática não foi prontamente atribuído ao conjunto de técnicas que utilizavam meios computacionais, matemáticos e estatísticos para o estudo de importantes eventos biológicos. A idéia de que macromoléculas como DNA, assim como proteínas, serem as portadoras da informação relativa acerca de um sistema biológico (Hagen, 2000) criou um cenário propício para o desenvolvimento e consolidação da Bioinformática. Esta poderia fornecer um conjunto de ferramentas para abordar diferentes aspectos das ciências da vida.

1.1.1 A OBTENÇÃO DAS SEQÜÊNCIAS DE AMINOÁCIDOS

Embora possa parecer contraditório, face ao elevado número atual de seqüências determinadas diretamente para nucleotídeos, um dos fatores que proporcionou o desenvolvimento da Bioinformática em seus primórdios, foi a existência de um elevado número de seqüências protéicas. Desde que a primeira proteína foi inteiramente seqüenciada pelo bioquímico Frederick Sanger, agraciado com o prêmio Nobel em Química do ano de 1958 (Nobelprize.org, 2010), a idéia de que a informação biológica estaria codificada na seqüência de aminoácidos tornou-se o principal foco do estudo de proteínas na época (Sanger, 1959). Desde a insulina com seus 51 aminoácidos seqüenciados - semente para o mundo digital das seqüências protéicas - diversas outras proteínas foram seqüenciadas através de métodos refinados mais rápidos do que o empregado pela equipe Sanger em sua empreitada inicial. A partir desse momento, com o conceito de que a informação biológica estaria contida na estrutura primária das proteínas, e com a relativa facilidade com a qual os novos métodos permitiam que esses dados pudessem ser explorados, o número de proteínas seqüenciadas e de seqüências depositadas em coleções digitais aumentasse exponencialmente (Hagen, 2000).

1.1.2 A DESCOBERTA DA ESTRUTURA TRIDIMENSIONAL DAS PROTEÍNAS

Outro tipo de informação biológica contida na estrutura primária é aquela necessária para a definição da maneira altamente específica que as proteínas adotam quando são dobradas no espaço tridimensional, ou seja quando adotam suas estruturas secundárias e terciárias (Anfinsen, 1973). Entretanto, esse tipo de informação ainda não pôde ser completamente acessado analisando-se unicamente a seqüência de aminoácidos de uma proteína. Apesar disso, a seqüência de aminoácidos é imprescindível para a correta interpretação dos dados gerados a partir dos experimentos de difração de raios X, experimento conduzido para determinação espacial de

cada aminoácido na estrutura tridimensional das proteínas. Esses conceitos puderam ser estabelecidos a partir dos trabalhos de John Kendrew e Max Perutz, que, usando as técnicas apresentadas acima, desvendaram as estruturas tridimensionais da mioglobina e da hemoglobina nos anos 80 (Olby, 1985; Perutz, 1985). Pode-se dizer que esse foi o início da era digital das estruturas protéicas.

1.1.3 A PRIMEIRA LINGUAGEM DE PROGRAMAÇÃO USADA EM BIOINFORMÁTICA

Com os avanços alcançados pelos grupos direcionados aos estudos de estruturas protéicas, sejam elas primárias, secundárias ou terciárias, a verdade é que, durante a década de 80, a quantidade de informação disponível para proteínas já era considerável. Esta riqueza bioinformática primordial tem como precursor, cerca de duas décadas antes, o fato de que os computadores tornaram-se muito mais freqüentes nos grupos de pesquisa, e não eram mais vistos como apenas uma ferramenta para tratar a riqueza numérica derivada dos experimentos de Bioquímica e Biologia Molecular. Naquela época, a primeira linguagem de programação de alto nível fora criada: a FORTRAN (*formula translation*) (Hagen, 2000).

1.1.4 O PRIMEIRO BANCO DE DADOS

Com a riqueza de dados biológicos, e com a relativa facilidade que a FORTRAN permitia que os novos biólogos computacionais pudessem criar soluções para as diversas demandas bioinformáticas, as agências de fomento daquela época incentivavam cada vez mais o envolvimento de cientistas no desenvolvimento de programas computacionais voltados para a área das Ciências Biológicas. Um exemplo brilhante da contribuição das agências de fomento para com o desenvolvimento da Bioinformática moderna é o financiamento direcionado aos projetos conduzidos por Margareth O. Dayhoff nas décadas de 60 e 70 (Hunt, 1983). Financiada pela *National Biomedical Research Foundation* (NBRF), *National Institute of Health* (NIH), *National Science Foundation* (NSF), *National Aeronautics and Space Administration* (NASA) e, finalmente, pela *IBM corporation*, Dayhoff empenhou-se no desenvolvimento de programas computacionais para determinar a seqüência de aminoácidos a partir de moléculas protéicas (Dayhoff, 1965a). Avanços importantes foram obtidos nesse sentido (Dayhoff, 1969, 1974), tais como o desenvolvimento de programas capazes de determinar com elevada precisão em poucos minutos a seqüência de pequenas proteínas (tal como a ribonuclease). A mesma tarefa seria executada por uma equipe de cientistas em vários meses de dedicação.

Com o advento dos seqüenciadores automáticos de proteínas, o número de seqüências disponíveis elevar-se-ia naturalmente, com isso Dayhoff e seus colaboradores decidiram criar a primeira base de dados conhecida na história da biologia molecular: o *Atlas of Protein Sequences and Structures* (Dayhoff, 1965b), uma publicação anual que visava catalogar os dados obtidos para todas as seqüências conhecidas de aminoácidos. Finalmente, em 1984, o primeiro banco de dados

on-line para armazenar seqüências de aminoácidos a partir dos dados do *Atlas* fora criado: o *Protein International Resource* (PIR, http://pir.georgetown.edu).

Inicialmente, a base de dados PIR era composta por apenas um número restrito de seqüências que na sua maioria constituía variantes interespecíficas de um pequeno conjunto de proteínas. Apesar de ser verdade, essa aparente pobreza atribuída ao PIR no quesito diversidade protéica, foi o que possibilitou o estabelecimento dos fundamentos das atuais análises comparativas de proteínas. Por exemplo, um subconjunto de proteínas presentes no PIR nos anos 60 tratava-se do Citocromo C, pigmento respiratório encontrado nas células aeróbicas; o número de seqüências disponíveis para esta proteína nas espécies amostradas pelo PIR era tão grande, que era quase impossível se cogitar que estas não eram derivadas do mesmo ancestral comum, ou seja, tidas como homólogas. Através do uso dessas seqüências, foi possível construir uma árvore filogenética com a distribuição das espécies extremamente semelhante àquela obtida a partir da análise de caracteres taxonômicos tradicionais (Fitch & Margoliash, 1967).

1.1.5 AS PRIMEIRAS ANÁLISES FILOGENÉTICAS

Logicamente, a imensa quantidade de dados disponíveis tornava complexa a análise filogenética. Embora esta pudesse ser realizada manualmente, devido a crescente complexidade dos dados, aliado à disponibilidade dos programas desenvolvidos pela equipe de Margareth Dayhoff, as análises filogenéticas tornaram-se computadorizadas (Dayhoff, 1969). Entretanto, as primeiras técnicas utilizadas não visavam solucionar completamente o problema filogenético. No início, as análises eram baseadas no cálculo da distância mutacional entre as proteínas envolvidas na filogenia, e no cálculo de quantas mutações seriam necessárias para que uma proteína se convertesse em outra. Tratava-se de um processo passo-a-passo: inicialmente os cientistas escolhiam um conjunto de três seqüências (do Inglês *branches* - ramos) por árvore, e então se adicionava um novo ramo a cada ciclo, de modo que a nova árvore minimizasse a distância entre cada ramo (distância mutacional). Depois que todas as seqüências eram usadas, o processo todo se reiniciava a partir de um conjunto diferente de três seqüências. Ao final, caso a nova árvore gerada fosse menos satisfatória do que a previamente estabelecida, esta era descartada (Hagen, 2000).

1.2 ALINHAMENTO DE SEQÜÊNCIAS.

Métodos eficientes de seqüenciamento de DNA facilitaram a obtenção de dados acerca da seqüência de proteínas em maior ordem de grandeza do que a análise direta de suas estruturas ou funções. Proteínas homólogas podem divergir bastante ao longo do tempo, entretanto suas estruturas e funções ainda se apresentam relativamente conservadas (Altschul, 1998). Dessa maneira muito pode ser inferido sobre a função de proteínas não caracterizadas quando uma

similaridade estatisticamente significativa é detectada contra seqüências de proteínas que são bem conhecidas.

No contexto acima diversos algoritmos para alinhamento de seqüências foram desenvolvidos, neles, as seqüências de proteínas são analisadas do ponto de vista de sua composição, onde valores de similaridade são atribuídos entre elas. Os métodos de alinhamento, sejam eles globais ou parciais (locais), caracterizam as modificações mais freqüentes ocorridas entre as proteínas: as substituições, inserções e deleções. Um esquema de pontuação para cada substituição ou *indel* (inserção-deleção) observado entre os diferentes aminoácidos foi criado com base nas substituições ocorridas em famílias protéicas bastante conservadas depositadas em bases de dados. Uma das mais conhecidas matrizes de substituição foi criada por Margareth Dayhoff pela análise de proteínas conservadas depositadas na base de dados PIR: a PAM (*Point Accepted Matrix*) (Dayhoff, 1978). A ocorrência de *indels*, representado por um espaçamento ou vão (do Inglês *gap*) no alinhamento, é tratada por um sistema de penalidades que foi criado para que a existência de um espaçamento no alinhamento seja relativamente mais relevante do que a extensão de um espaçamento existente, ou seja, número de resíduos que um espaçamento contém. Atualmente, existem métodos onde a matriz de substituição utilizada muda de acordo com as proteínas que compartilham alta similaridade e que foram identificadas na análise, sendo, portanto, mais sensíveis a encontrarem membros distantes da família protéica em questão (Altschul et al., 1997).

Com base nos esquemas de pontuação descritos acima, uma série de algoritmos escritos em linguagem de programação dinâmica foi desenvolvida para encontrar o alinhamento ótimo dado um par de seqüências. O algoritmo de Needleman-Wunsch foi criado para alinhamentos globais (Needleman & Wunsch, 1970), enquanto que o algoritmo de Smith-Waterman foi criado para analisar alinhamentos locais ótimos entre duas seqüências (Smith & Waterman, 1981). Embora nenhum dos dois métodos permitisse inicialmente a ocorrência de espaçamento nos alinhamentos, essa característica foi implementada ao custo de uma redução na velocidade de execução dos algoritmos. Os programas de alinhamento local BLAST (Altschul et al., 1990) e FASTA (Pearson, 1995) foram criados para encontrar os melhores alinhamentos locais entre uma seqüência e uma base de dados de seqüências a partir da extensão de regiões altamente similares presentes nas seqüências comparadas. Embora esta estratégia fosse bem mais rápida do que o algoritmo Smith-Waterman, ela podia superestimar segmentos relacionados ao acaso entre as duas proteínas. Parâmetros especiais existentes nesses programas permitem o ajuste de algumas de suas características heurísticas, o que pode influenciar seus níveis de velocidade e sensibilidade. A estatística de determinação do E-value posteriormente permitiu uma estimativa da rejeição da hipótese de homologia entre as seqüências comparadas pelo BLAST, onde o E-value representa o número de alinhamentos locais iguais ou melhores que o obtido pelo programa sem que houvesse uma relação biológica entre as seqüências (Karlin & Altschul, 1990).

1.3 HOMOLOGIA

A crescente disponibilidade de proteomas seqüenciados tornou possível a inferência sobre a história evolutiva dos organismos baseada nos dados de suas seqüências de aminoácidos (Alexeyenko et al., 2006). Um dos métodos mais confiáveis para se obter anotação funcional de proteínas consiste em estabelecer a relação de ortologia entre elas. Para isso, foram desenvolvidos diversos métodos baseados em buscas recíprocas, a maioria usando o método BLAST.

Entretanto, antes que seja prosseguido um detalhamento sobre os principais métodos de busca por proteínas que possam compartilhar uma descendência evolutiva, assim como uma funcionalidade comum, torna-se necessária a descrição das nomenclaturas sobre os diferentes tipos de homologia abordados na literatura.

Inicialmente, foi proposto um modelo sobre a maneira que as proteínas se aparentam do ponto de vista evolutivo (Fitch, 1970). Fitch tentou relacionar proteínas contemporâneas de acordo com suas descendências e definiu como parálogas aquelas proteínas que se originaram por duplicação enquanto as ortólogas seriam aquelas derivadas por especiação. É importante notar que mesmo proteínas atualmente presentes em organismos diferentes ainda assim podem ser consideradas parálogas, desde que o evento original ocorrido entre elas no cenário evolutivo tenha sido o processo de duplicação (Jensen, 2001). Dessa maneira, é possível que eventos tardios de duplicação linhagem-específica possam levar a presença de mais de um gene ortólogo entre duas espécies sem que um seja 'mais ortólogo' do que o outro (Figura 1A).

Os termos ortologia e paralogia têm sido mal interpretados, causando uma série de discórdias na comunidade científica; talvez a causa principal desse problema tenha sido o fato de que eles terem sido criados em tempos em que a Genômica ainda não era prática rotineira nos laboratórios. Um exemplo do mau entendimento dos termos pode ser encontrado numa recente revisão publicada na revista *Genome Biology* por Gerlt e Babbitt (2000). Neste artigo, os autores definem como ortólogas as proteínas presentes em espécies diferentes e que desempenham o mesmo papel, e como parálogas, aquelas proteínas homólogas presentes na mesma espécie e que, por divergirem após a especiação, não desempenham o mesmo papel biológico (Gerlt & Babbitt, 2000). Embora de uso às vezes discutível, os termos são extremamente úteis para caracterizar os genes homólogos presentes em espécies diferentes (ortólogos) e homólogos que resultam de expansão gênica em uma espécie (parálogos).

1.3.1 SEQÜÊNCIA E FUNÇÃO

Roy A Jensen comentou recentemente que a Genômica deveria evitar complicações semânticas com relação à nomenclatura atribuída a genes homólogos, e que deveriam ser focadas as relações de seqüência, estrutura e função entre proteínas, e que são necessárias ao entendimento das origens estruturais da função biológica, bem como ao estudo das bases

moleculares de suas divergências funcionais (Jensen, 2001). E que nós que estudamos as relações entre seqüência, estrutura e função deveríamos evitar os termos 'ortologia' e 'paralogia' a menos que estivéssemos interessados em destacar os eventos de especiação e duplicação gênica: eventos responsáveis pela produção das diversidades funcionais observadas entre proteínas homólogas. Para isso sugeriu os termos hetero-(iso-) funcionais e hetero-(iso-) específicos para classificar proteínas com a mesma ou diferentes funções no mesmo ou em diferentes organismos, respectivamente (Figura 1B).

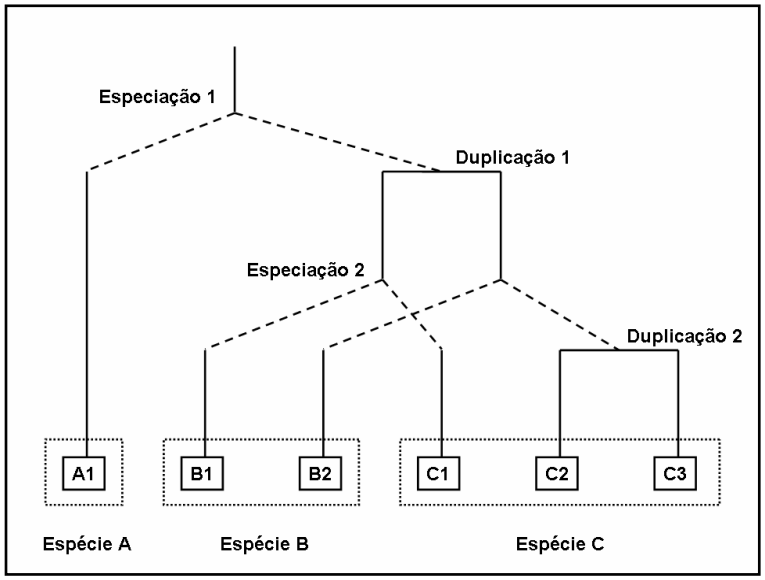


Figura 1: Relações filogenéticas. Na figura são representadas as relações evolutivas entre seqüências (A1, B1-2 e B1-3) de três espécies diferentes (A, B e C); eventos de especiação são representados por linhas tracejadas enquanto eventos de duplicação intra-específica por linhas horizontais; eventos recentes são mostrados próximos à base da árvore, enquanto aqueles tardios são mostrados mais próximos ao topo. (a) Para Fitch (1970) a proteína A1 possui três ortólogas na espécie C (C1-3) relacionadas pela Especiação 1; a proteína B1 possui C1 como ortóloga relacionada pela Especiação 2; já a proteína B2 possui duas ortólogas na Espécie C (C2 e C3) ambas relacionadas pela Especiação 2. Enquanto isso, B2 e C1 seriam parálogas uma vez que o evento mais antigo que relacionam elas duas na árvore é a Duplicação 1. Da mesma maneira C1, C2 e C3 compartilham a Duplicação 1, logo seriam consideradas parálogas. (b) Jensen (2001) utiliza o princípio da função biológica para classificar as proteínas em questão. Para ele C2 e C3, caso apresentem a mesma função biológica por terem divergido recentemente na Espécie C, poderiam ser consideradas isoespecíficas isofuncionais. As proteínas B1-B2 e C1-C2 (ou C1-C3) já se diferenciam um pouco mais tardiamente na árvore, podem ter acumulado diferenças significativas ao ponto de desempenharem funções diferentes nas respectivas espécies, nesse cenário, estes pares seriam considerados isoespecíficos heterofuncionais. Considerando agora o par B1-C1, estes divergiram recentemente pela Especiação 2, logo caso ainda desempenhem a mesma função (mesmo em espécies diferentes) seriam considerados heteroespecíficos isofuncionais. Porém, B1 em relação a C3, caso efetuem funções diferentes, seriam considerados heteroespecíficos heterofuncionais. (c) Por outro lado, adotando os conceitos propostos por

Sonnhammer e Koonin explicados adiante (Sonnhammer & Koonin, 2002), a árvore seria interpretada da seguinte maneira: B2 e C1 seriam parálogos externos, por se relacionarem pela Duplicação 1 e estarem presentes em espécies diferentes. C1, C2 e C3 seriam parálogos internos enquanto todas as proteínas da Espécie C seriam co-ortólogas da proteína A1 na espécie A, uma vez que estas se relacionam pela Especiação 1. Modificado de (Fitch, 1970).

1.3.2 NOVAS SOLUÇÕES PARA VELHOS PROBLEMAS

Para lidar com essa variação semântica no tocante ao evento de duplicação, Sonnhammer e Koonin propuseram uma nomenclatura para diferenciar os subtipos de parálogos (Sonnhammer & Koonin, 2002). Para eles, são consideradas 'paralogas internas' ("in-paralogs") aquelas proteínas parálogas que surgiram por eventos de duplicação linhagem específica e que a linhagem considerada ainda não foi sujeita ao evento de especiação. Por outro lado, 'parálogas externas' ("out-poaralogs") seriam as proteínas que surgiram por um evento de duplicação tardio antes do evento de especiação e o que levaria a uma linhagem diferente da linhagem considerada. Nesse caso, analisando-se a Figura 1B todas as proteínas parálogas internas de uma linhagem C qualquer são consideradas co-ortólogas a uma proteína A1 de uma linhagem A, desde que o último evento evolutivo comum entre a proteína A1 e a proteína progenitora das demais proteínas da linhagem C tenha sido uma especiação. É importante notar que parálogos externos caminham para uma divergência tal que os permite evoluírem para genes não mais relacionados funcionalmente, uma vez que a função original é supostamente executada por um deles. Assim, parálogos externos não são tão almejados por algoritmos de agrupamentos quanto parálogos internos e ortólogos.

1.4 ALGORITMOS

A detecção automática de proteínas ortólogas e parálogas internas, leia-se ortólogas, pela definição de Sonnhammer e Koonin (2002), é um problema de extrema importância, porém de difícil execução (Remm et al., 2001). A maneira mais natural de se detectar proteínas ortólogas, por definição que compartilham a mesma história evolutiva, seria a análise de árvores filogenéticas. Entretanto, os passos necessários para gerar tais árvores são de difícil automação. Para realizar tal análise considerando dois ou mais organismos, seria necessário agrupar as proteínas relacionadas, gerar um alinhamento múltiplo para cada grupo de domínios homólogos, construir uma árvore filogenética para cada grupo e finalmente extrair os grupos de homólogos destas árvores.

Cada um dos passos mencionados acima está sujeito a erros, o que poderia levar ao agrupamento de proteínas não relacionadas. Para reduzir essa vulnerabilidade, foram desenvolvidos diversos métodos que aplicam técnicas de alinhamento todos-contra-todos em

diversos grupos de organismos (Chervitz et al., 1998; Mushegian et al., 1998; Wheelan et al., 1999; Rubin et al., 2000).

1.4.1 COG

Partindo de relações entre pares de organismos, os estudos sobre genômica comparativa foram estendidos a grupos múltiplos, dada a disponibilidade de diversos organismos com genoma completamente seqüenciado. A base de dados COG (*Cluster of Orthologous Groups*) (Tatusov et al., 2000) talvez seja a mais conhecida iniciativa para agrupar as proteínas homólogas entre diferentes organismos. Esta base de dados utiliza métodos de alinhamento baseados em BLAST de todas as proteínas de um organismo *query* contra todas as proteínas de organismos *subject*. Sempre que estabelecida uma relação de melhor alinhamento bidirecional (do Inglês, *bidirectional best-hit*, BBH) entre três membros de três espécies diferentes, um grupo COG é formado; tal arranjo é denominado pelos autores da base de dados COG como 'triangulação'. A partir da trinca formada, novas proteínas de outros organismos, ou proteínas dos mesmos organismos utilizados na triangulação são incluídas ao grupo desde que respeitem a regra de BBH seja ele bidirecional (simétrico) ou indireto (assimétrico) (Figura 2).

A base de dados COG, bem como sua expansão para organismos eucarióticos KOG (*Eukariotic Clusters of Orthologous Groups*) (Tatusov et al., 2003), não representa uma análise filogenética concisa. Entretanto, fornece um recurso para agrupar as seqüências que mais provavelmente poderiam ser consideradas ortólogas. A regra de triangulação, porém, não permite pesquisadores estabelecerem relações de ortologia entre proteínas de dois outros organismos quaisquer não incluídos na base de dados COG. Para lidar com este problema foi criado o método InParanoid (Remm et al., 2001).

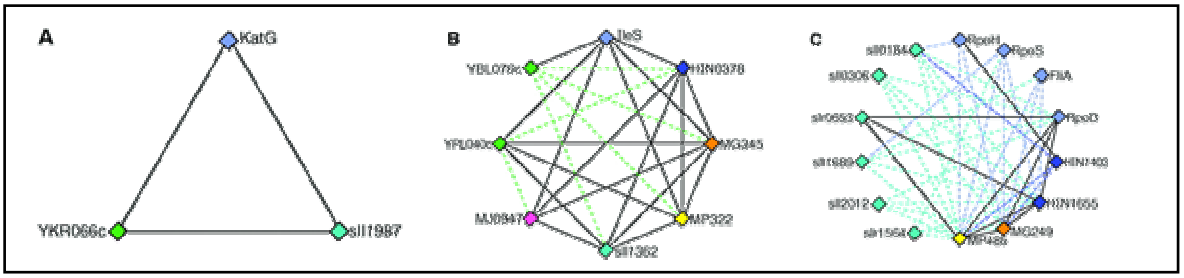


Figura 2. Esquema do algoritmo COG. Linhas sólidas mostram os melhores alinhamentos bidirecionais (ou BBH simétricos) enquanto as linhas tracejadas mostram os melhores alinhamentos indiretos (ou BBH assimétricos). Os genes comuns à mesma espécie são mostrados adjacentes e coloridos com a mesma cor. **(A)** estrutura do COG único, triangulação de três proteínas em três organismos diferentes. **(B)** COG simples com a ocorrência de uma paralogia

para o organismo representado em verde (levedura). **(C)** COG complexo onde múltiplos parálogos são exibidos. Modificado de (Tatusov et al., 1997).

1.4.2 INPARANOID

O método InParanoid identifica ortólogos e parálogos (internos) entre quaisquer pares de organismos. Entretanto, este método reduz a inclusão indesejável de seqüências parálogas externas. Esta metodologia pode ser considerada como uma extensão do método todos-contra-todos adotado pelo COG. A diferença fundamental é que o InParanoid adota regras especiais para a análise dos agrupamentos ("clusters") com o intuito de identificar todos os parálogos internos como membros do agrupamento (Remm et al., 2001). A premissa básica deste algoritmo é que, com relação a um grupo externo adotado, as seqüências que se apresentam como BBH uma da outra, precisam ter um *escore de alinhamento* maior entre si, do que com seqüências do grupo externo (Figura 3). Uma vez estabelecido o par BBH, homólogos adicionais em cada espécie (parálogos internos) são recrutados nestas espécies. A regra básica para o agrupamento de parálogos internos é que, em uma dada espécie, o par ortólogo principal (um dos pares BBH) é mais similar aos seus parálogos internos do que qualquer outra seqüência de outra espécie. Em seguida, regras são aplicadas para remover casos de ambigüidade entre os grupos formados.

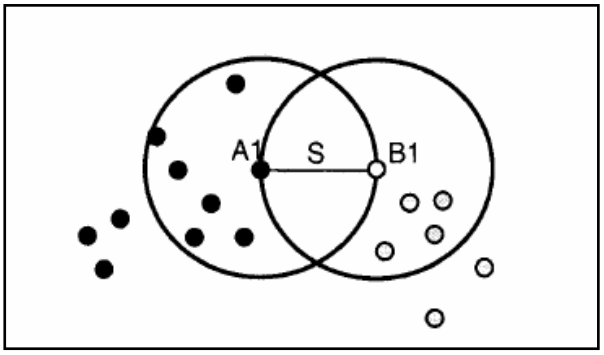
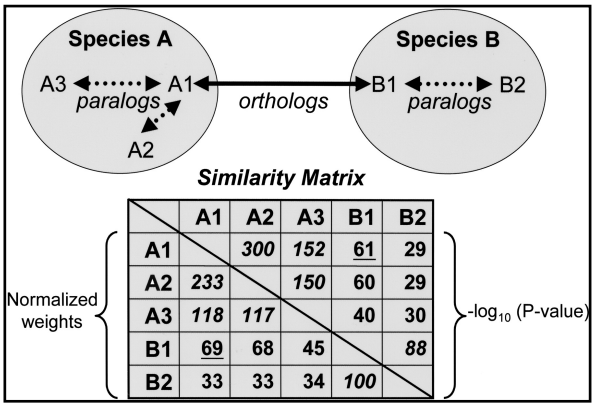


Figura 3: **Esquema do algoritmo InParanoid**. Agrupamento de parálogos internos adicionais. Os círculos representam as seqüências das espécies A (preto) e B (branco). Os ortólogos principais entre as espécies A e B estão destacados como A1 e B1. O InParanoid assume que a distância (inverso da similaridade, S), representada pela linha horizontal, entre cada ortólogo principal e os parálogos internos em cada organismo, é menor do que a distância entre o ortólogo e qualquer seqüência de outro organismo. A imagem mostra que todas as seqüências parálogas internas com distância menor que S então dentro da circunferência com diâmetro S. Dessa maneira as seqüências de ambas as espécies que estão fora da circunferência são consideradas parálogas externas e não são agrupadas. Modificado de (Remm et al., 2001).

1.4.3 ORTHOMCL

O método OrthoMCL fornece um método escalar para reconstruir grupos ortólogos entre múltiplos taxa eucarióticos, usando um algoritmo de *Markov Cluster* (MCL) para agrupar prováveis pares de ortólogos e parálogos (Li et al., 2003). Este método funciona similarmente ao InParanoid

quando dois genomas são analisados (Figura 4). Inicialmente prováveis ortólogos entre duas espécies são identificados a partir de pares de melhores hits recíprocos. Em seguida, parálogos "recentes" (chamados de parálogos internos pelo InParanoid) são identificados para cada espécie, adotando a premissa deles serem mais similares ao membro do par ortólogo identificado para a espécie em questão, do que às seqüências de outros organismos. Após a identificação de todas as seqüências, pesos são criados com base no valor de escore do alinhamento entre cada par de seqüências envolvida na análise. Adicionalmente, é criada uma segunda pontuação para cada par, esta pontuação é normalizada pela média dos pares de ortólogos inicialmente identificados. Os valores finais são organizados como uma matriz simétrica onde um algoritmo de MCL é aplicado (Enright et al., 2002). O método MCL considera de maneira global e simultânea todas as relações entre as seqüências agrupadas, fornecendo uma maneira de separar parálogos divergentes, ortólogos distantes selecionados por baixos valores de escore durante o alinhamento recíproco e, presumidamente, seqüências com diferentes arquiteturas de domínios.



	A1	A2	A3	B1	B2
A1		*300*	*152*	61	29
A2	*233*		*150*	60	29
A3	*118*	*117*		40	30
B1	69	68	45		*88*
B2	33	33	34	*100*	

Figura 4: **Esquema do algoritmo OrthoMCL.** Um par de ortólogos prováveis (A1 e B1) é designado por melhor hit recíproco entre as espécies A e B. Parálogos recentes (A2, A3 e B2) originados por duplicação dentro das espécies são identificados. Na metade superior direita da matriz são representados os pesos iniciais (w_{ij}) baseados nos valores de alinhamentos de todas as seqüências entre si. Na metade inferior esquerda são representados os pesos (w_{ij}) divididos por W_{ij}/W, onde W representa a média entre o peso da distância entre todos os ortólogos (números sublinhados) e os parálogos recentes (números em itálico) e W_{ij} representa a média do peso entre todos os ortólogos das espécies i e j. Modificado de (Li et al., 2003).

1.4.4 TOGA

O Método TOGA (*TIGR Orthologous Gene Alignments*) tem como objetivo identificar ortólogos a partir da análise de seqüências de DNA ao invés de proteínas sendo bastante conservativo nos critérios escolhidos para identificação de ortólogos (Lee et al., 2002). A premissa é de que se os ortólogos são bem conservados ao nível protéico, eles devem também se apresentar conservados o suficiente ao nível de DNA, a ponto de serem identificados pela análise dos *matches* transitivos e altamente limitados entre três ou mais espécies ao limite de similaridade BLASTn com um *E*-value de 1e-5 (Figura 5). A inclusão de Eucariotos primitivos (como levedura) na análise efetuada pelo TOGA, bem como a existência de seqüências do mesmo gene que não

apresentam sobreposição pela baixa amostragem do método EST (*expressed sequence tags*) potencializaram a ocorrência de seqüências pertencentes a diferentes TOGs (*tentative ortholog groups*). Tal evento foi minimizado pela adoção de regras de desambiguação, onde TOGs que compartilham mais de dois terços das seqüências são agrupados no mesmo TOG.

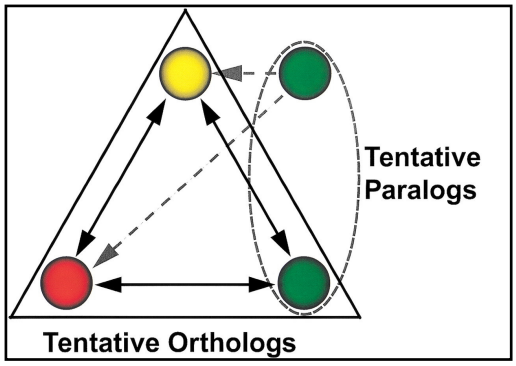


Figura 5: **Esquema do algoritmo TOGA**. As seqüências de consensos tentativos de 28 bases de dados do TIGR Gene Index são alinhadas todas-contra-todas. Os melhores hits transitivos que alinham três ou mais espécies definem um grupo ortólogo tentativo (*Tentative Ortholog Group* – TOG). Os outros *hits* não recíprocos definem os parálogos tentativos. Modificado de (Lee et al., 2002).

1.4.5 MULTIPARANOID

O método MultiParanoid foi criado a partir da aplicação do InParanoid na análise de múltiplos genomas. Basicamente o MultiParanoid lê as saídas criadas pelo InParanoid e cria agrupamentos multi-específicos a partir destas (Alexeyenko et al., 2006).

Em termos gerais o MultiParanoid constitui uma abordagem baseada em *single-linkage*. Por exemplo, consideradas três espécies A, B e C, e as tabelas de agrupamentos parciais formados entre as espécies A-B, B-C e A-C, o MultiParanoid procura a presença dos pares ortólogos de agrupamentos iniciais (*seeds*) da tabela A-B nas tabelas A-C e B-C, caso presentes, todos os membros (parálogos internos) nas tabelas A-C e B-C correspondentes são incluídos no agrupamento *seed*, sendo o processo repetido até que todos os pares de ortólogos sejam processados. Em casos excepcionais onde seqüências pertencem a diferentes agrupamentos, o MultiParanoid aplica regras de desambiguação, dessa maneira as seqüências que não são o ortólogo *seed* em nenhum dos agrupamentos são associadas ao cluster ao qual possui o maior escore, sendo, portanto, removida dos outros. Caso a seqüência seja a *seed* de um agrupamento, ela é mantida neste agrupamento e removida dos outros, evitando que o processo seja interrompido.

Além dos métodos citados acima, outros métodos foram criados para agrupar proteínas ortólogas em diferentes organismos (Tabela 1). Cada método possui suas peculiaridades e vários foram comparados no tocante as suas características (Chen et al., 2007). Da mesma maneira alguns desses métodos foram aplicados a organismos modelos. Bases de dados sobre proteínas ortólogas derivadas destes métodos foram desenvolvidas e encontram-se disponíveis em diversos sítios (Tabela 2).

Tabela 1: Alguns métodos adicionais disponíveis para agrupamento de proteínas ortólogas.

Métodos	Estratégia	Parâmetro analisado	Referência
RIO	Filogenia	Limiar de *bootstrap* para ortologia.	(Zmasek & Eddy, 2002)
Orthostrapper	Filogenia	Limiar de *bootstrap* para ortologia.	(Storm & Sonnhammer, 2002)
RSD	Distância	Limiar de E-value no BLASTP, Limiar de divergência.	(Wall et al., 2003)
RBH	BLASTP	Limiar de E-value no BLASTP.	(Hirsh & Fraser, 2001)
TribeMCL	Homologia	Limiar de E-value no BLASTP, índice de inflação MCL.	(Enright et al., 2002)

Tabela 2: **Algumas bases de dados disponíveis sobre homologia de seqüências.**

Base de dados	URL*
COG/KOG	http://www.ncbi.nlm.nih.gov/COG/
EGO	http://www.tigr.org/tdb/tgi/ego/
HomoloGene	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene
InParanoid/MultiParanoid	http://www.inparanoid.cgb.ki.se/ http://www.multiparanoid.cgb.ki.se/
HOPS	http://www.pfam.cgb.ki.se/HOPS/
KEGG	http://www.genome.jp/kegg/
OrthoMCL	http://www.orthomcl.cbil.upenn.edu/cgi-bin/OrthoMclWeb.cgi
PhIGs	http://www.phigs.jgi-psf.org/
Ensembl Compara	http://www.ensembl.org
MGD	http://www.informatics.jax.org/searches/homology_form.shtml
HOGENOM	http://www.pbil.univ-lyon1.fr/databases/hogenom.html
HOVERGEN	http://www.pbil.univ-lyon1.fr/databases/hovergen.html
INVHOGEN	http://www.bi.uni-duesseldorf.de/~invhogen/invhogen.html
TreeFam	http://www.treefam.org

* Consultado em 13/03/2008

Os exemplos anteriores mostram a riqueza de dados que passou a ser gerada no âmbito da Biologia Molecular, concomitantemente aliada aos diversos métodos para análise de seqüências. Estratégias para identificação de proteínas bem conhecidas em organismos pouco estudados permitiram que estes pudessem ser caracterizados 'a priori' com base no conhecimento acumulado. Logicamente, vários aspectos puderam ser inferidos acerca destes novos organismos, o que pode ser evidenciado pela quantidade crescente de novos artigos descrevendo as peculiaridades desses novos proteomas. Essa explosão de trabalhos científicos passou a dificultar

a vida do cientista interessado em se aprofundar em determinado domínio do conhecimento, para tal, um conjunto de técnicas conhecidas conjuntamente por Mineração de Texto (do Inglês, "*text mining*"), surgiu para facilitar o trabalho. Alguns tópicos relevantes sobre esta área são apresentados a seguir.

1.5 MINERAÇÃO DE TEXTO

Os avanços das técnicas biológicas para experimentação em larga-escala, aliados ao surgimento recíproco de diversas ferramentas bioinformáticas para análise dos dados produzidos, aceleraram o passo em que novas informações biológicas relevantes são geradas (Krallinger & Valencia, 2005). Na ciência atual, a quantidade de informação pode ser considerada como o número de artigos e revistas científicas que são acumulados em determinado período de tempo (Jensen et al., 2006). Dessa maneira, o crescimento da informação científica é refletido pelo crescente número de publicações científicas observados em determinada área do conhecimento.

A base de dados MEDLINE, principal base onde a ferramenta PubMed (http://www.ncbi.nlm.nih.gov/PubMed/) realiza suas consultas, armazena atualmente cerca de 17 milhões de artigos científicos (Abril de 2008). Nos últimos 20 anos, a informação contida na MEDLINE apresentou um taxa de crescimento de 4,2% ao ano, por outro lado o número de novas publicações na MEDLINE cresceu num ritmo aproximado de 3,1% ao ano (Hunter & Cohen, 2006). Isso mostra que a literatura científica disponível (leia-se, quantidade de informação) cresce em ritmos exponenciais (Figura 6). Esse tsunami de informação está tornando quase impossível a cientistas se manterem atualizados em toda literatura relevante até mesmo quando o universo de artigos é restringido à pesquisa desenvolvida em sua área de atuação (Jensen et al., 2006) .

Os últimos anos presenciaram os novos resultados da pesquisa sobre o processamento da linguagem biomédica (*biomedical language processing* – BLP). Em um universo onde cerca de 80% dos artigos (depositados no MEDLINE) contém informação em texto livre, estruturado de uma maneira que facilita a leitura por humanos, porém difícil de ser interpretada automaticamente por computadores (Hale, 2005), a BLP visa extrair automaticamente a informação implícita na literatura científica. À BLP compreende também os métodos e ferramentas computacionais para tratar os textos gerados por humanos como *input* e realizar tarefas como recuperação de informação, classificação de documentos, extração de informação, ou descoberta baseada em literatura. Num contexto mais biológico, as técnicas de extração de informações mais específicas tais como redes de interação proteína-proteína (Buckingham, 2005), estabelecimento de funções protéicas e associações genes-proteínas ou proteína-fenótipo (Gonzalez et al., 2007) são coletivamente chamadas de mineração de texto (*text mining*).

1.5.1 RECUPERAÇÃO DE INFORMAÇÕES

Os métodos utilizados em mineração de texto aplicados na literatura biológica tornaram-se essenciais aos pesquisadores, eles possibilitam a identificação dos artigos relevantes para determinada área de pesquisa, num processo conhecido como recuperação de informação (*information retrieval* – IR). Em IR, geralmente o conjunto de artigos a ser pesquisado em busca daqueles relevantes encontra-se depositado em bases de dados de conhecimento como a MEDLINE. O sistema que gera a interface com o usuário para recuperar a informação dessa base de dados é o PubMed, que é usado aqui para exemplificar as duas estratégias principais de recuperação de informações:

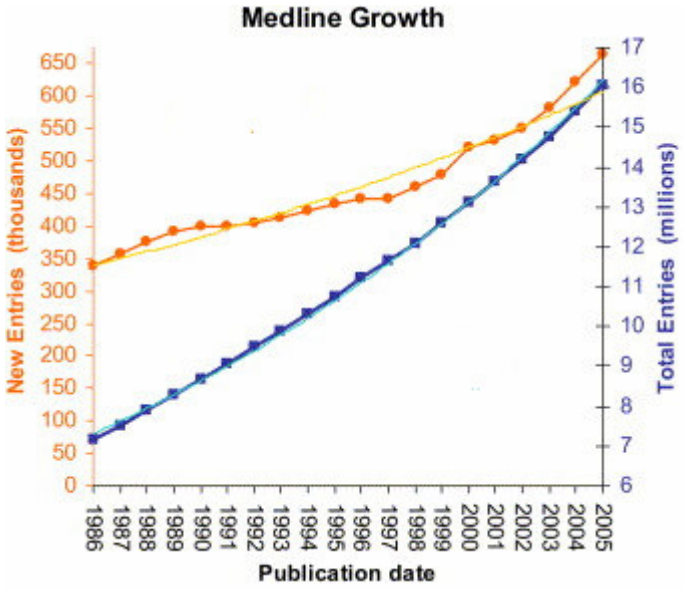


Figura 6: **Crescimento da literatura biomédica publicada entre 1986 e 2005**. Os círculos laranjas representam o número de novos artigos indexados no MEDLINE em determinado ano. Enquanto isso, os quadrados azuis mostram o número total de artigos indexados ao final de cada ano. Modificado de (Hunter & Cohen, 2006).

Modelo booleano. Trata-se de usar uma combinação de termos conhecida como 'pergunta' (em Inglês, *query*) para recuperar os artigos que apresentam todos ou parte dos termos dessa combinação em determinado campo de sua estrutura (seja ela título, corpo, resumo, etc.) ou ao longo de todo o artigo científico.

Modelo vetorial. Nesse modelo o elemento de busca na base de dados não é mais uma frase usada como *query*, e sim um artigo ou um conjunto de artigos pré-definido pelo usuário.

Nesse modelo, cada documento é tratado como um vetor (conjunto) de termos (palavras), e um escore para cada termo é gerado com base na freqüência em que esses termos ocorrem ao longo do universo analisado. Ao final, aqueles artigos-alvo (*subject*) que apresentam um escore similar ao(s) artigo(s) *query* inicial(is) são selecionados (em Inglês, *match*) (Wilbur & Yang, 1996).

Os métodos vetoriais são muitas vezes utilizados na classificação de documentos usando técnicas de inteligência artificial (IA) (Fynbo et al., 2006), onde um software aprende a diferenciar artigos relevantes daqueles irrelevantes com base na sua composição de termos (Marcotte et al., 2001). Técnicas de IA comumente utilizadas em mineração de texto são a *Support Vector Machine* (SVM) (Krallinger et al., 2005; Pahikkala et al., 2005) e as redes neurais artificiais .

Os modelos booleano e vetorial usados no PubMed utilizam recursos que facilitam a interpretação da terminologia usada durante a busca nas bases de dados. Por exemplo: remoção de palavras muito comuns na literatura (tais como preposições) que não representam de fato informação; sufixos em palavras como "*-s*" e "*-ing*" permitem que os radicais sejam interpretados como o mesmo termo, entre outros (Jensen et al., 2006). Um recurso conhecido como *thesauri* (Schijvenaars et al., 2005) é utilizado para expandir a *query* através da utilização dos sinônimos dos termos utilizados, dessa maneira a *query* "tomato AND disease AND resistance" seria estendida para "(tomato OR Solanum lycopersion) AND disease AND resistance".

1.5.2 RECONHECIMENTO DE ENTIDADES BIOLÓGICAS

Pode-se dizer que uma das etapas cruciais ao bom funcionamento das ferramentas de mineração de texto é a correta identificação dos termos que representam uma entidade biológica tais como genes e proteínas.

O reconhecimento das entidades nomeadas (*named entity recognition* – NER) consiste na triagem das palavras que se referem às entidades biológicas (genes, proteínas, organismos, doenças, etc.) seguido da atribuição de identificadores para tais entidades (Hirschman et al., 2002). Inicialmente os termos biológicos eram filtrados da literatura com base em regras simples, tais como avaliação das terminações das palavras, padrões identificados nas palavras ou até mesmo na observação das palavras adjacentes aos termos a serem analisados (Fukuda et al., 1998; Tanabe & Wilbur, 2002). Entretanto, a crescente disponibilidade dos bancos de dados de literatura científica onde as entidades biológicas são marcadas no texto (*tagged*) possibilitou que metodologias baseadas em inteligência artificial (IA) pudessem ser desenvolvidas (Zhou et al., 2005). Com o uso destes bancos, conhecidos como *corpus* (Hao et al., 2005), programas de IA conseguem identificar os nomes com base em seus padrões característicos observados em suas ocorrências em frases do *corpus*.

Em paralelo aos métodos baseados em IA, existem aqueles que são baseados em dicionários de genes e proteínas (Leonard & Duffy, 2002; Chang et al., 2004; Crim et al., 2005). Nesse enfoque, uma lista de termos, contendo todas as variações sinônimas para cada

gene/proteína é comparada ao conjunto de textos a ser analisado. Sempre que há uma ocorrência, o sistema não só reconhece a entidade biológica, como também atribui um identificador a ela, tudo com base no dicionário de termos adotado. Um dos grandes desafios do reconhecimento de entidades é lidar com a falta de padronização da literatura científica (Bajic et al., 2005; Mons, 2005). Muitas vezes genes são referidos na literatura por nomes pouco informativos e ainda pior, por nomes comuns e não necessariamente biológicos (Chen et al., 2005). Isso leva aos programas de NER serem às vezes pouco sensíveis. Para lidar com esse problema, cada vez mais é incentivado o uso de nomenclatura padronizada para descrever e nomear os genes a serem citados nos artigos científicos. As chamadas ontologias são atualmente o melhor recurso para se padronizar os termos científicos (Yoo et al., 2007). Essa prática não está somente restrita aos nomes de genes e proteínas (Ashburner et al., 2000), como também existem bases de dados de ontologia para descrever termos diversos, tais como anatômicos ou celulares (Ilic et al., 2007).

1.5.3 EXTRAÇÃO DE INFORMAÇÕES

A extração de informação (*information extraction* – IE) permite que fatos predefinidos sejam recuperados da literatura (Jensen et al., 2006). Os fatos se referem aos relacionamentos que diferentes entidades biológicas apresentam entre si. Essas relações são extraídas por diversas premissas, a primeira delas é o estudo da co-ocorrência dos termos (Srinivasan & Hristovski, 2004; Cohen et al., 2005). Caso a co-ocorrência das entidades biológicas seja freqüente num determinado conjunto de artigos, é possível que essas entidades estejam de fato biologicamente relacionadas. Esta correlação é medida por um esquema de pontuação baseado em freqüência de co-ocorrência. Contudo, usando apenas análises de co-ocorrência não é possível determinar a natureza da relação encontrada entre os pares, por exemplo, se uma determinada proteína A identificada no texto atua diretamente numa proteína B ou não, ou seja, se este efeito é dependente, digamos, de uma proteína C ou não. Outra característica que também não é detectada é o sentido da interação, ou seja, se a proteína A atua sobre a B ou vice-versa. Esses dados devem ser analisados por curadores da informação extraída pelo sistema de mineração de texto.

Outra abordagem de também visa IE trata-se do processamento natural de linguagem (*Natural-Language Processing*, NLP) (Libbus & Rindflesch, 2002). A NLP realiza uma análise ordenada da maneira em que as palavras são colocadas para formar as frases (sintaxe) e busca recuperar a idéia da relação que é explicitada da frase (semântica). O texto a ser analisado é simbolizado (do Inglês, *tokenized*) para identificar as bordas das sentenças informativas; em seguida um indicador das partes do discurso (do Inglês, *part-of-speech tag*) é atribuído às palavras pertencentes à frase simbolizada. Depois, métodos de IE ou dicionários servem para identificar em meio às palavras marcadas as entidades biológicas e as palavras de interesse (como, por exemplo, verbos que indicam uma ação biológica). For fim, um conjunto de regras é utilizado para

extrair as relações das árvores de sintaxe e das palavras com função semântica atribuída (Rindflesch et al., 2000; Jensen et al., 2006). Para obterem essa performance, os programas usados em NLP são treinados para atuarem de forma a reconhecerem a estrutura dos textos biológicos com base na análise de frases nos artigos depositados em bancos do tipo *corpus* (Craven & Kumlien, 1999).

1.6 INTEGRAÇÃO DE DADOS

Da mesma maneira que os últimos avanços científicos em ciências grou a imensa quantidade de dados refletida pela base de dados de proteínas PIR e pela base de dados de literatura PubMed, discutidas nas sessões anteriores, diversas outras bases de dados biológicas surgiram para lidar com essa diversidade de dados. Descrevemos abaixo algumas das comumente utilizadas.

A base de dados UniProt é um recurso central para armazenar e interconectar informações advindas de diversas fontes sobre seqüências de proteínas (Apweiler et al., 2004). O Uniprot é composto por quatro divisões otimizadas para diversas finalidades: a *UniProt Knowledgebase* (UniProtKB) é uma divisão composta por duas sessões, a *UniProt/Swiss-Prot*, contendo informações curadas com links para literatura específica (Boutet et al., 2007), enquanto a *Uniprot/TrEMBL* contém registros computacionalmente analisados enriquecidos com classificação e anotação automática (Bairoch & Apweiler, 1996) das proteínas nelas depositadas. A *UniProt Archive* (UniParc) (Leinonen et al., 2004) é a divisão de armazenamento dos dados e reflete o histórico das seqüências depositadas. A *UniProt Reference Clusters* (UniRef) é a divisão que reúne as seqüências relacionadas do ponto de vista de similaridade de seqüência (Suzek et al., 2007); esta possui as sessões UniRef 100%, 90% e 50%, dependendo da similaridade mínima apresentada pelas seqüências pertencentes aos agrupamentos. Finalmente a *UniProt Metagenomic and Environmental Sequences* (UniMES) que foi desenvolvido especificamente para dados derivados de análises metagenômicas.

Dentre outras bases de dados sobre informações biologias específicas, podem ser destacadas a base de dados KEGG - *Kyoto Encyclopedia od Genes and Genomes* (http://www.genome.jp/kegg/) que possui um acervo sobre diversos aspectos biológicos, tais como mapas globais de funções celulares e orgânicas (KEGG Atlas); mapas e módulos metabólicos (KEEG Pathway); hierarquias funcionais e ontologias (KEGG Brite); genomas, genes, proteínas e grupos ortólogos (KEGG Genes) além de compostos químicos, drogas, glicanos e ligações (KEGG Ligand). Outra base de dados especificamente relacionada ao agrupamento de seqüências em famílias protéicas é o PFam (Sonnhammer et al., 1998). Nesta base de dados, cada família protéica é representada por dois alinhamentos múltiplos de seqüências, dois *profiles-HMM* (profile-Hidden Markov Model) e um arquivo de anotação.

A existência de inúmeros recursos virtuais, cada um com seu modelo de distribuição de dados, torna necessário o desenvolvimento de metodologias para integrar as informações presentes nessas diversas fontes, tornando-as exploráveis de uma maneira conjunta e eficiente (Stein, 2003). Além disso, é imprescindível que as ferramentas de navegação entre as diversas fontes sejam fáceis de serem usadas pelos cientistas usuários destas bases (Davidson et al., 1995).

Para se desenvolver sistemas de integração faz-se necessária a compreensão da arquitetura padrão dos diversos tipos de recursos biológicos disponíveis na Internet. As bases de dados biológicas são construídas principalmente por três níveis organizacionais (Figura 7): na base do sistema, encontra-se o sistema gerenciador do banco de dados (SGBD) que armazena as informações características da base; e no topo encontra-se o sistema que interage com os usuários, conhecido como navegador web. Entre estes dois níveis, encontram-se os softwares que fazem a comunicação entre o usuário e o conteúdo da base de dados, que, na maioria das vezes e composto por um servidor web e por um software de acesso a base de dados, que, embora sejam transparentes aos usuários da base, são de extrema importância para que as solicitações dos usuários sejam transformadas como *queries* a serem efetuadas no banco de dados pelo SGBD, e retornadas ao navegador web como páginas “hiperlincáveis” (HTML) (Stein, 2003).

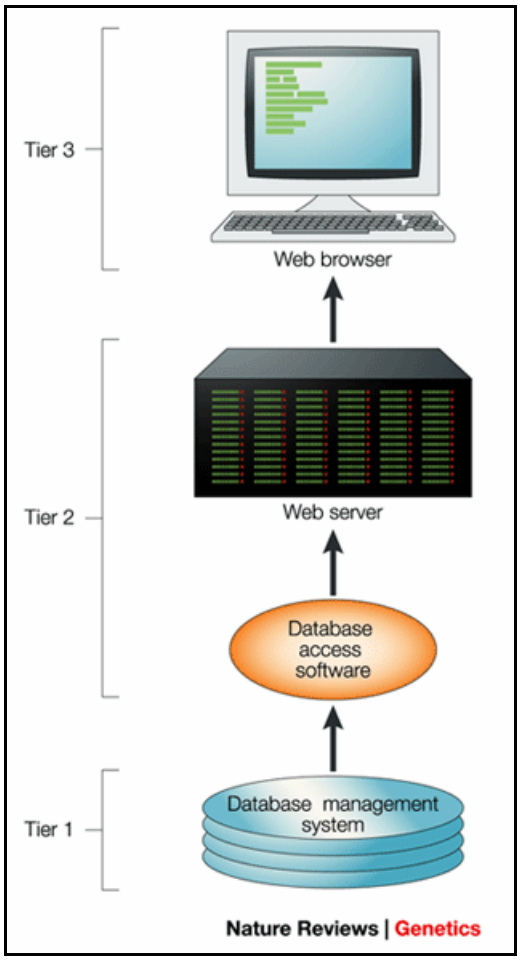


Figura 7: Arquitetura básica das bases de dados biológicas. No primeiro nível (Tier 1) encontra-se o sistema gerenciador do banco de dados, já no segundo nível (Tier 2) são posicionados os softwares para acesso as informações depositadas na base e o servidor web onde as páginas oferecidas ao usuário serão armazenadas. Por último, no terceiro nível (Tier 3), encontra-se o navegador web (browser) onde as informações são mostradas aos usuários das bases biológicas. Retirado de (Stein, 2003)

A necessidade da integração aparece quando diferentes bases de dados que armazenam diferentes tipos de informações acerca da mesma entidade biológica usam modelos diferentes para a representação da entidade na base ou, por exemplo, para armazenamento e disponibilização dos dados, e não pelo fato de cada uma armazenar um tipo diferente de informação (Markowitz, 1995; Haferkamp et al., 2002). A individualidade de cada base de dados reflete os diferentes interesses que cada grupo de pesquisa possui para com diversas áreas de pesquisa, e a existência de inúmeras bases não deve ser encarada como um problema, e sim como um incentivo à existência de técnicas que possam permitir a integração dos diferentes conjuntos de dados (Venkatesh & Harlow, 2002).

1.7 ESTRATÉGIAS PARA INTEGRAÇÃO DE DADOS

Diversas estratégias foram criadas para que diferentes bases de dados pudessem ser integradas (Stein, 2003); estas estratégias obviamente foram herdadas de propriedades previamente existentes no ambiente WWW (*world wide web*), algumas delas serão brevemente discutidas abaixo:

HIPERLINK

A integração baseada em links é uma das mais comumente utilizadas para se criar um ambiente de bases de dados integradas. Este tipo de integração busca fazer com que um determinado registro numa base de dados local X esteja vinculado a uma base de dados remota Y, porém os recursos de cada base de dados são armazenados individualmente. Trata-se de um modelo onde as bases de dados estão associadas pelo hiperlink, porém ambas mantêm sua individualidade e não necessariamente precisam trocar dados entre si, apenas apontar o registro correspondente entre elas (Stein, 2003).

Este modelo de integração é bastante simples e funcional, porém depende de algumas premissas que muitas vezes não são mantidas pelos gerenciadores das bases de dados. Uma delas é o fato dos nomes dos registros mudarem com elevada freqüência, o que faz com que uma referência feita por um dos links na base de dados deixe de existir. Para que esse sistema possa funcionar bem, gerenciadores de bases de dados que vinculam sua informação a fontes externas, precisam conhecer as regram adotadas para geração de links por essa base-alvo, reduzindo assim o problema da quebra de vínculo. Um problema extremamente difícil de se abordar pelo lado das bases de dados que estão sendo vinculadas (destino do hiperlink) é localizar e informar todas as bases de dados que as apontam, sobre modificação de determinado registro.

INTEGRAÇÃO VISUAL

Este sistema, pouco utilizado, visa integração de diferentes bases de dados do ponto de vista visual, dessa maneira, os dados apresentados ao usuário permanecem cada um nos diferentes servidores, entretanto, é gerado um ambiente onde a pergunta, originalmente feita na base de dados local, é processada e distribuída por softwares de controle a distintas bases de dados remotas. Depois que cada base responde especificamente às perguntas distribuídas, o ambiente gerado pelas bases de dados remotas é recriado como se fosse parte da resposta do sistema local, entretanto trata-se de uma visualização das respostas de cada base de dados conjuntamente, por um único servidor (Stein, 2003).

GERENCIAMENTO DE DADOS

O gerenciamento de dados (em inglês, data warehousing) é uma estratégia que cria uma base de dados que contém o conteúdo de diversas outras bases de dados remotas. Para que os dados sejam captados, é necessário que softwares de importação dos dados sejam escritos para cada uma das diferentes fontes, uma vez que cada uma apresenta seu modelo próprio de

armazenamento (Schonbach et al., 2000; Stein, 2003). O segundo passo nessa estratégia é transformar os dados importados no formato da base de dados local, segundo obviamente as regras de armazenamento impostas pelo sistema gerenciador do banco de dados ou SGBD. Em seguida, deve haver um sistema que carrega as informações adquiridas no banco de dados da base local, dessa maneira informações de diversas fontes podem ser integradas na mesma base de dados.

Apesar de parecer bastante promissora, esta estratégia encara o problema da atualização dos dados. Os sistemas de gerenciamento de dados devem estar permanentemente cientes da existência de *updates* em cada uma das fontes de dados, e sempre que constatada uma atualização, os novos dados precisam ser prontamente re-importados. Outro ponto que precisa ser considerado pelos sistemas de importação é no tocante à flexibilidade da estrutura das bases de dados que são consultadas, uma vez que caso haja uma mudança drástica, é possível que o sistema deixe de funcionar para determinada fonte (Stein, 2003).

1.7.1 WEB SERVICES

A necessidade de se criar um ambiente de integração de integração de dados advindos de fontes diferentes, fez com que os gerenciadores de bases de dados passassem a distribuir seus serviços não mais apenas pelo tradicional método de consultas via formulários HTML, onde usuários entram com suas *queries* em busca de registros da base de dados como respostas, mas também passaram a fornecer um acesso programático as informações contidas na base (Neerincx & Leunissen, 2005).

Para que seja implementado o acesso programático a uma base de dados, isto é, de computador para computador, uma série de especificações baseadas em Web Services precisam ser definidas. Um Web Service é um sistema de software (programa) na qual interfaces públicas (outros programas) e contratos são definidos e escritos em linguagem XML (*eXtensible Markup Language*). Estas definições podem ser descobertas por outros sistemas de software (UDDI, *Universal Discovery Description and Integration Protocol*). Estes sistemas podem então interagir com o Web Service em uma maneira definida pela sua definição (WSDL, *Web Service Description Language*), usando mensagens baseadas em XML e transportadas por protocolos da Internet (HTTP, *Hypertext Transfer Protocol*).

EXTENSIBLE MARKUP LANGUAGE (XML)

Atualmente é quase impossível para alguém que tenha entrado recentemente na comunidade científica imaginar como a pesquisa biológica poderia ser realizada sem os recursos da Internet. Em partes, o que gerou o sucesso obtido pela WWW foi a linguagem desenvolvida para obtenção, submissão, navegação e análise dos dados publicados. Entretanto a HTML tem diversas limitações principalmente por ser dedicada a navegação por humanos. Ela é uma linguagem pouco arbitrária e que não informa muito sobre a semântica dos documentos por ela

representados (Achard et al., 2001). Por causas dessas e de outras desvantagens, a linguagem HTML torna difícil a extração de informações de maneira programática. Para lidar com essa dificuldade foi criada a linguagem XML (Achard et al., 2001). A linguagem XML é uma linguagem de baixo nível dedicada para a Internet derivada de uma linguagem mais complexa chamada SGML. A diferença entre HTML e XML é que XML contém marcadores (do inglês, *tags*) que criam um formato de dados auto-descritivo, enquanto HTML apenas codifica como o dado deve ser formatado. Por exemplo, usando-se HTML pode-se especificar se um texto será apresentado em negrito ou itálico, enquanto que o XML poderia especificar se o texto trata de um nome de um gene ou de uma proteína (Neerincx & Leunissen, 2005).

SIMPLE OBJECT ACCESS PROTOCOL (SOAP)

No ambiente web diversos protocolos são utilizados para transferência de informações, por exemplo, o protocolo FTP é utilizado para transferência de arquivos na Internet. Para que Web Services sejam utilizados é necessário definir o protocolo em as informações estruturadas serão trocadas em um ambiente descentralizado. O SOAP é o protocolo de comunicação para os Web Services, ele especifica o formato das mensagens XML trocadas (W3C, 2008a). Existem outras especificações SOAP que descrevem como representar os dados do programa em XML e como usar o SOAP para fazer as chamadas de procedimento remoto (RPC). Estas partes opcionais da especificação são usadas para implementar as aplicações no estilo RPC, onde a mensagem SOAP, contendo a chamada e os parâmetros da função a ser executada, é enviada pelo cliente e o servidor retorna uma mensagem com os resultados da função executada.

WEB SERVICES DESCRIPTION LANGUAGE (WSDL)

É a linguagem de descrição do Web Service com a finalidade de documentar as mensagens que o Web Service aceita e gera como resposta. A notação que o arquivo WSDL usa para descrever o formato das mensagens é baseada em XML, o que significa que é uma linguagem de programação neutra e baseada em padrões, o que a torna adequada para escrever as interfaces dos Web Services, que são acessadas por uma grande variedade de plataformas e linguagens de programação. Além de descrever o conteúdo das mensagens trocadas com o Web Service, o WSDL define onde o serviço está disponível e quais os protocolos de comunicação que são usados para se comunicar com o serviço. Dessa maneira, o WSDL tem por função definir tudo o que é necessário para escrever um programa que utilize o XML do Web Service (W3C, 2008b).

UNIVERSAL DISCOVERY, DESCRIPTION AND INTEGRATION (UDDI) PROTOCOL

Para que os Web Services sejam localizados, é necessário que haja um mecanismo de busca. Os protocolos UDDI são usados para criar um diretório onde programas e aplicações encontram e usam dinamicamente um Web Service. Um diretório UDDI é um arquivo XML que descreve o serviço por meio de registros. A forma pela qual os serviços são definidos no

documento UDDI é chamado *Type Model* ou *tModel*; em muitos casos, o tModel contém o arquivo WSDL que descreve a interface SOAP do Web Service (UDDI, 2008).

2. OBJETIVOS

2.1 OBJETIVO GERAL

Desenvolver métodos de mineração de texto, agrupamento de seqüências e integração de dados para desenvolvimento de uma base de dados sobre proteínas relacionadas aos mecanismos de defesa em plantas.

2.2 OBJETIVOS ESPECÍFICOS

- Aplicar métodos de recuperação de informação e identificação de identidades biológicas na base de dados PubMed para extrair nomes de proteínas relacionadas aos mecanismos de defesa em plantas contra estresses bióticos e abióticos em resumos depositados nessa base;
- Aplicar técnicas de extração de informações nos resumos de artigos científicos do PubMed, visando obter a co-ocorrência das proteínas mencionadas no texto juntamente com outras proteínas e estímulos relevantes ao tema em estudo;
- Criar uma rede de relacionamento entre os termos extraídos a partir da análise de co-ocorrência para *input* em programas de visualização de dados biológicos, notadamente o Arena3D.
- Recuperar, quando possível, a seqüência de aminoácidos das proteínas identificadas na mineração de texto, pela consulta direta na base de dados UniProtKB.
- Desenvolver uma metodologia para se encontrar as proteínas parálogas internas e aquelas (co-)ortólogas a- cada uma das seqüências obtidas acima, através da análise de múltiplos links à uma proteína *seed* qualquer, e agrupá-las em respectivos *clusters*.
- Desenvolver um software para aquisição automática de dados biológicos relevantes, em diferentes bancos de dados, correspondentes às proteínas presentes nos clusters descritos acima, usando técnicas de integração de dados baseadas em *webservices*.
- Desenvolver uma base de dados com interface web e acesso programático baseado em XML-SOAP, nomeada *Plant Defense Mechanisms Database*, para armazenar e permitir a consulta das informações extraídas nos passos anteriores.

3. JUSTIFICATIVA

3.1 ANÁLISES DE MINERAÇÃO DE TEXTO PARA IDENTIFICAÇÃO DE PROTEÍNAS RELACIONADAS AOS MECANISMOS DE DEFESA EM PLANTAS.

A imensa quantidade de artigos científicos publicados nos últimos anos e o índice crescente de novas publicações disponibilizadas em bases de dados como a MEDLINE torna humanamente impossível a extração de todas as relações entre entidades biológicas presentes na literatura. Para facilitar o árduo trabalho de ler e associar os termos de importância biológica, diversas técnicas baseadas em mineração de literatura tem sido adaptadas para a mineração de textos biológicos. Embora tais recursos tenham sido aplicados com determinado sucesso ao estudo de genes e proteínas humanas ou de microorganismos, a inespecificidade dos dicionários utilizados torna difícil a correta extração de informações para proteínas de plantas. Uma das poucas estratégias de mineração de texto que existem na área de Biologia de Plantas, o *Dragon Plant Biology Explorer* (DPBE), leva em consideração a ocorrência de termos biológicos na área vegetal baseado em nove dicionários gerais, bem como sua co-ocorrência juntamente com outros termos ao longo de todo o resumo científico, o que torna a análise seja muito superficial, superestimando a predição de correlação biológica a partir da co-ocorrência espúria ao longo do artigo, muitas vezes com os termos presentes em sessões que não são relacionadas. Além disso o DBPE não permite ser localmente instalado o que dificulta análises em larga escala bem como a integração com outras metodologias para mineração de texto.

Propomos um método integrado à ferramenta NLPROT (Mika & Rost, 2004) para identificar nomes de proteínas em resumos obtidos pelo PubMed, validar esses nomes pela comparação com um dicionário de nomes específicos de proteínas/genes de plantas, identificar termos representativos de estímulos biológicos presentes em dicionários personalizados, e correlacionar estes termos quando se apresentarem intercalados por um ou mais termos indicativos de interação biológica na mesma frase, aumentando as chances dos termos filtrados estarem de fato relacionados ao mesmo evento biológico. O método é implementado como um script desenvolvido em PHP chamado LAITOR (*Literature Assistant for Identification of Terms co-Occurrences and Relationships*).

3.2 AGRUPAMENTO DE SEQÜÊNCIAS SIMILARES A PARTIR DE LINKS MULTIPLOS A UMA SEQÜÊNCIA FUNDADORA (SEED)

Não existe atualmente um recurso que monte grupos de proteínas similares a uma proteína *seed* específica do usuário adotando as regras de melhor hit recíproco como fazem os métodos COG e MultiParanoid. Além do mais, esses métodos mencionados restringem a análise aos organismos de proteoma completo, sem explorar o potencial das seqüências dos organismos com proteoma incompleto. O que propomos nesta parte do trabalho é criar um grupo de proteínas similares com as características do método MultiParanoid iniciando a busca com apenas uma seqüência *seed*, focando a análise apenas na proteína de interesse do usuário e dispensando tempo de processamento adicional para que todos os proteomas dos organismos de interesse sejam analisados para se obter "clusters" para apenas uma seqüência alvo. Para isso criamos o método Seed Linkage.

3.3 AQUISIÇÃO DE INFORMAÇÕES VIA WEB SERVICES PARA AS PROTEÍNAS PRESENTES NA BASE PDM.

Os métodos Seed Linkage e LAITOR citados acima, apesar de serem de uso universal, foram criados com o propósito de se estabelecer uma base de dados sobre proteínas relacionadas aos mecanismos de defesa em plantas (Plant Defense Mechanisms - PDM - Database) que está sendo desenvolvida. Desta maneira as proteínas mencionadas na literatura relacionada a essa área, bem como suas co-ocorrências junto a outras proteínas ou a estímulos bióticos ou abióticos, foram extraídas usando-se o LAITOR e a identificação das proteínas correspondentes a estas em outros organismos foi inferida pelo método Seed Linkage. Ao final, espera-se que haja um agrupamento de seqüências protéicas para cada proteína relevante que foi identificada na literatura.

Para se obter informações acerca dessas proteínas, foi criado um cliente baseado na tecnologia web service para comunicação com o sistema SRS instalado no EMBL em Heidelberg, por ocasião da recente disponibilidade da interface SOAP neste servidor remoto, o que facilita a aquisição de dados dessa base além de integrar suas informações em diversos sistemas. Usar web services para obter informações acerca das proteínas depositadas na PDM Database evita que uma cópia de todas as bases de dados em questão seja mantida na máquina servidora da PDM Database, uma vez que a PDM Database representa uma pequena parcela dessas bases maiores, evitando desperdício de recurso computacional além de facilitar a atualização do nosso sistema.

4. MATERIAIS E MÉTODOS

4.1 VERSÕES DOS SOFTWARES UTILIZADOS

- APACHE Version 2.0
- Arena3D Beta version
- BLAST Version 2.2.13
- NLPROT Version 1.0
- NuSOAP Version 0.7.3
- PHP Version 5.X
- TMHMM Version 2.0

4.2 SISTEMA OPERACIONAL

Todos os scripts foram criados e executados, respectivamente, em ambiente Linux, principalmente na distribuição CentOS 5.1 ou Fedora (versão 6 ou superior). Análises numéricas e figuras foram elaboradas usando os recursos do pacote MS Office em sistema operacional Windows.

4.3 BANCOS DE DADOS

O sistema gerenciador de bancos de dados usado para armazenar todas as informações geradas ao longo das diversas etapas apresentadas nessa tese foi o MySQL em ambiente Linux.

4.4 RECURSOS COMPUTACIONAIS DO LABORATÓRIO DE BIODADOS - UFMG

- axe: Athlon X2 2GHz, 1Gb Memória DDR2 800MHz, CentOS 5.1
- maracatu: Core2Duo Quad 2.4GHz, 4GB Memória DDR2 1066MHz, CentOS 5.1
- xote: Athlon X2 3GHz, 2Gb Memória DDR2 800MHz, CentOS 5.1
- baiao: Athlon X2 2,5GHz, 2Gb Memória DDR2 800MHz, CentOS 5.1

4.5 BASES DE DADOS CONSULTADAS

- NCBI Gene http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
- NCBI PubMed http://www.ncbi.nlm.nih.gov/pubmed/
- NCBI Taxonomy http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy
- Multiparanoid http://www.sbc.su.se/~andale/multiparanoid/html/index.html
- SRS-EMBL http://srs.ebi.ac.uk/
- UniProtKB http://www.uniprot.org/

5. RESULTADOS E DISCUSSÕES

5.1 DESENVOLVIMENTO DO PROGRAMA LAITOR: *LITERATURE ASSISTANT FOR IDENTIFICATION OF TERMS CO-OCCURRENCES AND RELATIONSHIPS.*

Apresentamos neste artigo uma metodologia baseada em mineração de texto (*text mining*) para a identificação de termos biológicos, tais como nomes de proteínas e estímulos bióticos e abióticos, bem como suas co-ocorrências em frases de resumos de artigos científicos, buscando identificar associações implícitas na literatura, daí seu nome **LAITOR**: ***L****iterature* ***A****ssistant for* ***I****dentification of* ***T****erms co-****O****ccurrence and* ***R****elationships*.

Nosso método inicia-se pela aquisição de artigos relevantes que porventura descrevam a atuação de proteínas de plantas contra patógenos ou contra condições ambientais adversas. Para isso utilizamos um sistema baseado em *support vector machine* (Soldatos, et al. dado não publicado) treinado com artigos manualmente classificados como relevantes e irrelevantes ao tema de estudo: proteínas envolvidas com os mecanismos de defesa em plantas. Depois de aplicar o SVM no conjunto de artigos publicado nos últimos cinco anos para *Arabidopsis* na base de dados PubMed, 230 resumos foram selecionados para as análises posteriores.

Criamos três dicionários para a análise conforme descritos abaixo:

1. **Dicionário de nomes de proteínas**: nomes de proteínas e seus sinônimos foram obtidos da base de dados NCBI Gene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene), considerando apenas os genes de plantas (Viridiplantae). Em seguida, para enriquecer o dicionário de nomes de proteínas, executamos uma busca Booleana usando o sistema PubMed. Todavia, nossa busca foi limitada apenas aos artigos diretamente ligados a organismos da divisão Viridiplantae do banco de dados NCBI Taxonomy. Utilizamos a *query "(resistan* OR toleran*) AND ((high light OR high-light) disease OR cold OR pathogen* OR drought OR salinity OR "oxidative stress*" OR (high temperature OR high-temperature)) AND plant*" para selecionar os resumos dos artigos de interesse, que depois foram salvos localmente no formato XML. A expansão se deu pela inclusão de nomes de proteínas citados em campos específicos dos arquivos XML correspondentes aos artigos acima recuperados, desde que estes termos já não estivessem representados como nome ou sinônimo de um registro da base de dados Gene.

2. **Dicionário de termos indicativos de interação biológica**: as biointerações, como designamos estes termos nos artigos, foram criadas manualmente pela inspeção de um subconjunto de verbos previamente listados na literatura (Hoffmann & Valencia, 2005) como sendo os mais relevantes e freqüentes nos artigos científicos. Ademais, incluímos termos especialmente usados na área de Biologia Vegetal, principalmente na área de interação patógeno-hospedeiro ou resposta contra estresse, manualmente.

3. **Dicionário de estímulos bióticos e abióticos**: assim chamamos o terceiro dicionário utilizado em nossa análise. Esse dicionário é composto por uma série de termos que

representam hormônios vegetais (etileno, jasmonatos, ácido salicílico, etc.), estímulos ambientais (seca, salinidade, frio, etc.) e compostos químicos que geram respostas fisiológicas nas plantas quando submetidas a estresse (peróxido de hidrogênio, etc.).

Todos os dicionários utilizados são compostos por um nome e seus respectivos sinônimos ou variações tipográficas. O que permite uma busca mais ampla pelos termos que representam uma determinada entidade biológica de interesse.

Determinados os dicionários, o passo seguinte foi identificar estes nomes nos resumos selecionados. Nessa etapa, utilizamos o programa NLPROT (Mika & Rost, 2004) para identificar os nomes das proteínas. Porém, o NLPROT possui um dicionário interno que faz com que ele identifique proteínas de organismos não-vegetais, logo, um termo que representa o nome de uma proteína animal, por exemplo, pode não se referir a uma proteína presente em vegetais; dessa maneira, interações errôneas poderiam sem filtradas. Para lidar com esse problema, todas as proteínas marcadas pelo NLPROT nos resumos foram conferidas contra o dicionário de proteínas estabelecido. Os nomes de termos de biointerações e de estímulos bióticos e abióticos foram marcados no texto usando expressões regulares definidas.

Sempre que dois nomes de proteínas ou o nome de uma proteína e de um estímulo co-ocorrem na mesma frase, esta frase é armazenada para a verificação de um termo de biointeração entre elas, caso este termo seja encontrado, a co-ocorrência é validada e inserida na rede de co-ocorrência.

Depois que todos os artigos são analisados, e as co-ocorrências extraídas, o sistema conta, para cada par estabelecido, o número total de linhas em que este foi identificado e o número de artigos nas quais estas linhas foram encontradas. Estas informações são apresentadas como uma tabela no formato texto, uma página web com as co-ocorrências para cada termo identificado na análise, e também como um arquivo de entrada para o programa Arena3D (Pavlopoulos, et al. dado não publicado), um aplicativo que está sendo desenvolvido no EMBL para visualização de redes biológicas complexas e que utilizou os dados do LAITOR como estudo de caso em seu desenvolvimento.

O programa LAITOR ainda está em fase de desenvolvimento, porém os scripts individuais que compõem o método podem ser adquiridos em nosso servidor na página http://biodados.icb.ufmg.br/laitor bem como o resultado das análises efetuadas com os abstracts classificados pelo sistema SVM acima apresentada. Esperamos desenvolver um aplicativo em Java plataforma-independente que possa permitir o usuário fazer a "*query*" diretamente no PubMed, recuperar os resumos de interesse no formato pertinente, criar ou importar os dicionários biológicos relevantes à determinada área de pesquisa, e realizar as análises de reconhecimento de entidades e extração de informação (redes de co-ocorrência) de maneira integrada usando um único recurso.

O manuscrito a seguir descreve a análise dos 230 resumos classificados como relevantes pelo sistema de SMV para *Arabidopsis* e visando criar uma rede integrada sobre os mecanismos de resposta contra estímulos bióticos e abióticos, usando os dicionários acima descritos. Este trabalho foi publicado na sessão de software da revista *BMC Bioinformatics* com o título: **LAITOR: Literature Assistant for Identification of Terms co-Occurrences and Relationships.**

As seqüências de aminoácidos das proteínas correspondentes aos nomes identificados nesses resumos serão obtidas no banco UniProtKB para diferentes organismos e incluídas na base de dados PDM. Por sua vez, essas seqüências servirão como *seeds* do método Seed Linkage, que visa o agrupamento de proteínas similares do ponto de vista de seqüência entre diversos organismos e é apresentado na próxima sessão dos resultados.

Barbosa-Silva et al. BMC Bioinformatics 2010, 11:70
http://www.biomedcentral.com/1471-2105/11/70

BMC Bioinformatics

SOFTWARE Open Access

LAITOR - Literature Assistant for Identification of Terms co-Occurrences and Relationships

Adriano Barbosa-Silva[1,2,3], Theodoros G Soldatos[3,4], Ivan LF Magalhães[2], Georgios A Pavlopoulos[3], Jean-Fred Fontaine[1], Miguel A Andrade-Navarro[1], Reinhard Schneider[3], J Miguel Ortega[2*]

Abstract

Background: Biological knowledge is represented in scientific literature that often describes the function of genes/proteins (bioentities) in terms of their interactions (biointeractions). Such bioentities are often related to biological concepts of interest that are specific of a determined research field. Therefore, the study of the current literature about a selected topic deposited in public databases, facilitates the generation of novel hypotheses associating a set of bioentities to a common context.

Results: We created a text mining system (LAITOR: *Literature Assistant for Identification of Terms co-Occurrences and Relationships*) that analyses co-occurrences of bioentities, biointeractions, and other biological terms in MEDLINE abstracts. The method accounts for the position of the co-occurring terms within sentences or abstracts. The system detected abstracts mentioning protein-protein interactions in a standard test (BioCreative II IAS test data) with a precision of 0.82-0.89 and a recall of 0.48-0.70. We illustrate the application of LAITOR to the detection of plant response genes in a dataset of 1000 abstracts relevant to the topic.

Conclusions: Text mining tools combining the extraction of interacting bioentities and biological concepts with network displays can be helpful in developing reasonable hypotheses in different scientific backgrounds.

Background

The richness of information generated by different research groups is sometimes focused on issues that lack explicit connection with those generated by colleagues from other groups. However, currently, there are available literature mining techniques that permit to connect the knowledge generated by distinct groups and improve the understanding of some key points of their research [1]. Text mining machines have been created to mine the biological information in a trial to establish new biological concepts from previous knowledge [2-4]. These machines were proven to be reliable in extracting biological facts either analyzing full text [5,6] or just condensed information present in the abstracts of scientific papers [7,8] as stored in the MEDLINE database.

Text mining techniques for information-retrieval comprise some basic steps: to find relevant articles in the research field of interest; to identify the biological entities cited in the text, as well as to disambiguate confuse bioentity names (i.e. genes and proteins) within and among distinct species; to infer putative relationships between bioentities based on co-occurrence of biological terms in the same article, abstract, sentence or phrase [2]. Recently, AliBaba has been developed to graphically visualize information on associations between biological entities extracted from PubMed using pattern matching and co-occurrence filtering (http://alibaba.informatik.hu-berlin.de/, [9]). Later, a system called NetSynthesis [10] has been developed to permit the controlled building of biomolecular networks by users, where the searching criteria on PubMed are customized by using parse tree query language [11]. However, these systems do not permit the integration of customized dictionaries on their algorithm.

We present here a system called LAITOR (*Literature Assistant for Identification of Terms co-Occurrences and Relationships*). This software was developed to normalize the bioentities names tagged in the abstracts to a user defined protein dictionary; as well as to extract their co-occurrence, along with other protein or important biotic/abiotic stimuli terms, the later implemented

* Correspondence: miguel@icb.ufmg.br
[2]Laboratório de Biodados, Dpto. de Bioquímica e Imunologia, ICB - UFMG. 31270-901, Belo Horizonte - MG, Brazil

Barbosa-Silva et al. BMC Bioinformatics 2010, 11:70
http://www.biomedcentral.com/1471-2105/11/70

as a customized concept dictionary. Such co-occurrences are extracted taking into consideration the presence of terms in the same sentence of scientific abstracts and adopting a set of rules to filter bioentity pairs that occur in several sentence structures (see details in Implementation). The software performed as a greatly precise method. Here, it has been used to mine protein co-occurrences related to green plant-pathogen interactions.

Implementation

Abstracts retrieval

In order to retrieve scientific abstracts related to green plants that would be related to defense mechanisms, we have used the system MedlineRanker [12]. Two MeSH http://www.nlm.nih.gov/mesh/ terms (Host-Pathogen Interactions AND Plants) have been used as "training dataset" to rank 10,000 recently-published abstract from the whole MEDLINE database. After the MedlineRanker analysis we retrieved the top 1,000 PubMed IDs from the generated rank to be loaded as "application dataset" for the next steps of our analysis [Additional file 1].

Protein tagging

LAITOR is optimized to work by analyzing tagged scientific abstracts. For this purpose, we adopted the NLPROT [13] program as LAITOR's protein tagger. The plain text format (-f txt) must be chosen for the NLPROT output file, where bioentity names present in the text are tagged between "<n>" and "</n>" tags. The tagged protein names are filtered according to a user-defined bioentity dictionary, in our case study: a plant protein name and synonym dictionary.

Protein Dictionaries

Two protein dictionaries have been generated for the development of LAITOR. The first (named human proteins dictionary) created for the evaluation of LAITOR performance (explained below) in the BioCreative II Interaction Article Subtask (IAS) [14]. The second (named plant protein dictionary) has been used in the identification of co-occurring of green-plant protein pairs retrieved for abstracts related to host-pathogen interactions.

The human protein dictionary has been created by using all the protein records deposited for *Homo sapiens* [NCBI Taxonomy id: 9606] in the UniProt-SwissProt-TrEMBL (UP-SP-TR) database. In this dictionary, the definition(s) and synonym(s) for all human UP-SP-TR proteins are included. Furthermore, for each record, the corresponding NCBI Gene symbol and synonyms were used to enrich the representative terms of said protein. At the end, the human proteins dictionary is composed by 87,537 records (IDs), comprising a total of 112,686 distinct protein terms, which have been completed by the addition of 40,234 supplementary terms from the NCBI Gene database.

Additionally, specific genes names and synonyms for every organism deposited in the NCBI Taxonomy database that have gene records in the NCBI Gene database have been used to create LAITOR readable dictionaries. To use these dictionaries, users must inform the taxonomy identification number (Taxonomy ID) for the preferred organism followed by the extension ".dictionary" (e.g. "9606.dictionary" for "Homo sapiens" genes) during set up, as explained at LAITOR's documentation file.

For the plant dictionary, the complete Gene tab-delimited database from Entrez website has been downloaded (5,317,958 records), which comprises 505,403 different organisms (Taxonomy IDs - TAXIDs). To filter only those records related to green-plant proteins, we used the NCBI Taxonomy database to select from the Gene table only those records with a TAXID corresponding to Viridiplantae organisms, which included 99,488 different records. At the end, the plant protein dictionary contained 148 plants organisms (0.02% of total organisms) and a total of 237,077 Gene records (4.45%), which included 217,224 distinct protein symbols and 62,521 synonyms (see one example for the Gene PR1 of *Arabidopsis thaliana* [GenBank: 815949] in Additional file 2).

The resulting table displays two columns: one for the bioentity names, and the second with their respective synonyms so that it can exist as lines (records) as synonyms for each bioentity name (Additional file 2).

Name ambiguity

Another aspect explored by LAITOR, is how to handle gene name ambiguity. The strategy of using the Taxonomy database to limit the number of used entries reduced the possibility of inclusion of names of other organisms which would cause ambiguity among terms. However there are terms that commonly occur for more than one organism, or different proteins from the same organism that share the same name or synonym. To cope with this, LAITOR creates a tag file in which the ambiguous terms identified in the analysis are normalized to the same name in the protein dictionary. Such terms that match multiple protein names or that are synonyms of multiple protein names are marked in the LAITOR output. This warns users about the possibility of misinterpretation for such a term.

Concepts Dictionary

In order to check the co-occurrence and likely involvement of plant proteins names along with biotic and abiotic stimuli names, a list of previously known stimuli and their synonyms has been provided as Concept Dictionary (for example: Jasmonic Acid, Jasmonate and JA were included as the same concept). Both, Protein and

Barbosa-Silva et al. BMC Bioinformatics 2010, 11:70
http://www.biomedcentral.com/1471-2105/11/70

Concept Dictionaries are available as additional material [Additional files 3 and 4].

Additionally, in order to attend different contexts, we have populated all the sub-headings of NCBI's Medical Sub Headings (MeSH) Trees (available at http://www.nlm.nih.gov/mesh/trees.html) as LAITOR's concepts dictionaries, as explained at LAITOR's documentation.

Biointeractions Dictionary

A list representing the different types of interactions or relationships between proteins was generated based on previously published list [4,15]. It is composed by 76 terms, which have been included together with a total of 886 synonyms as seen in Additional file 5, Table S2. Considering all terms, the biointeraction dictionary in its entirety is composed of 963 different words.

Co-occurrence analysis

Once the abstracts to be analyzed had been retrieved and tagged for protein and gene names, biointeractions and concepts, LAITOR was used to perform a co-occurrence analysis [see Additional file 6].

At the sentence level, each line of the tagged abstracts was divided at every full stop (".") punctuation sign. We paid special attention to the presence of these full stop marks in alternative positions that did not indicate the end of the period, as in the case of species names (for example: *A. thaliana*) or protein names (for example: PDF1.2 protein).

Initially the whole abstract is screened to store the occurrence of all bioentity names. After storage of all names, each protein name is checked for its occurrence in each of the separated sentences. If a bioentity term is found, let us name this term as "Pair 1", the script checks the occurrence of a second bioentity name, "Pair 2", different from Pair 1 in the same sentence. To avoid redundancy, the script checks on-the-fly if Pair 2 is a synonym of the previously identified Pair 1 and discards such cases.

It has been previously published that 90% of the biointeractions among proteins documented in the literature adopts the pattern "Protein-Biointeraction-Protein" [16], this pattern being chosen by approaches like iHOP [15] and HomoMINT [17]. Nevertheless, we adjusted LAITOR to identify other patterns of Protein-Protein or Protein-Concept co-occurrence, as explained below.

The co-occurrences identified by LAITOR are classified into four types. From the most to the least stringent, these types are:

Type 1: Both co-occurring protein names/synonyms must not refer to the same protein (common for all types of co-occurrences), they must be present in the same sentence of the abstract and, additionally, it is required that a term from the Biointeractions Dictionary occurs in between the considered terms. An extra optional step is the identification of a biological stimuli (represented as a term from the Concepts Dictionary) term anywhere in the sentence, which is then associated to the interacting pair;

Type 2: Same as Type 1, except that the biointeraction may occur anywhere in the sentence;

Type 3: Same as Type 1, except that the occurrence of a biological term in the sentence is not required;

Type 4: All the pairs of co-occurring protein names/synonyms mentioned in the abstract are considered, whether they are in the same sentence or not.

Thus, when LAITOR performs under type 4, the other co-occurrence types are included.

Multiple co-occurrences of type 1, 2 and 3, might happen in a given sentence. To cope with this, our system was adapted to perform an overlapped search. This means that in cases where two proteins (A and B) occur along with the same biointeraction, like in the sentence "A and B regulate C", the pairs "A-regulate-C" and "B-regulate-C" are identified as type 1 co-occurrences. Note that the co-occurring pair "A-B" will be assigned type 2. Moreover, in more complex sentences such as "A is regulated by B and activates C", the system will retrieve as co-occurrences of type 1 "A-regulated-B", "A-regulated, activates-C', and "B-activates-C" (together with type 2 "A-regulated, activates-B" and type 2 A-regulated, activates-C) thus over predicting the number of different bio-interactions between the A, B and C proteins. However such complex sentences may not be very frequent. In order to determine if they are a serious problem, we performed a series of manual evaluations of the results of LAITOR's analysis on several abstract datasets.

Performance evaluation

Protein term co-occurrences at sentence level of scientific abstracts might be potentially useful for the prediction of literature-based protein-protein interactions. Therefore, we have tested the performance of LAITOR to find protein-protein interaction data in abstracts. For this purpose, we have used the BioCreative II test dataset for the Interaction Article Subtask (IAS) as gold standard [14]. This "performance evaluation dataset" is composed of relevant (3,529) and irrelevant (1,957) abstracts for the curation of protein-protein interactions present in the MINT and IntAct databases [18]. Once LAITOR identifies a co-occurring protein pair in an abstract, this is considered to be positively (relevant) classified. After the classification of all gold standard abstracts the precision and recall are calculated for each of the four co-occurrence types (1-4), and the performance compared to methods participating in the BioCreative II challenge. A receiver operating curve (ROC) was created by using the package ROCR [19]. Positive and negative performance

Barbosa-Silva et al. BMC Bioinformatics 2010, 11:70
http://www.biomedcentral.com/1471-2105/11/70

evaluation datasets are provided as additional material [Additional file 7].

Network representation

A protein and stimuli co-occurrence analysis created by LAITOR from PubMed abstracts is parsed from a general output file into a tab-delimited text file (extension .co) that is used as input by most network visualization software. As default, LAITOR generate inputs for two of these programs: EMBL Medusa [20] and EMBL Arena3D [21], which provide networks in one- and multi-dimensional charts, respectively, enabling the complex output generated by LAITOR to be efficiently handled.

Results and Discussion

LAITOR's developmental pipeline

LAITOR has been developed by combining a flexible rule-based method together with a pre-defined vocabulary match approach. Figure 1 illustrates the pipeline for LAITOR's development, which is explained in detail in the following sections.

LAITOR uses as input a set of scientific abstracts as stored in the records of the MEDLINE database. Abstracts are analyzed individually for co-occurrences, which are extracted and classified into four types according to the rules described in Implementation section. Additional file 8, Figure S1 exemplifies a tagged sentence extracted from the PubMed article identified by PMID 19061405. The co-occurrence analysis starts by (i) the creation of a list with the occurring bioentities (proteins or genes, [see Additional file 2, Table S2]) and stimuli names present in precompiled dictionaries (see Implementation), for the whole abstract. In the example the names detected were: HSP90, RAR1 and SGT1. (ii) Further, each sentence is queried for the co-occurrences of different bioentity names establishing pairs. In this

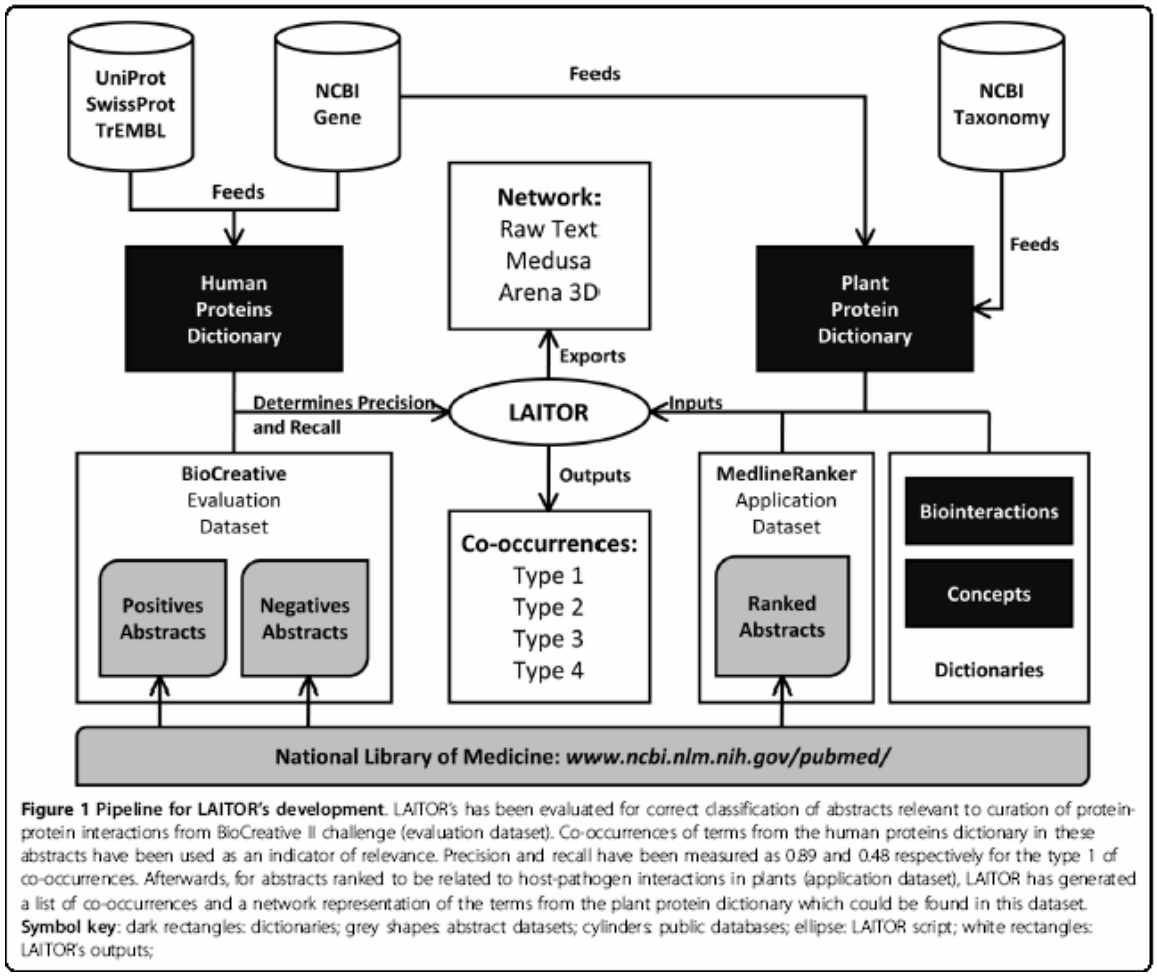


Figure 1 Pipeline for LAITOR's development. LAITOR's has been evaluated for correct classification of abstracts relevant to curation of protein-protein interactions from BioCreative II challenge (evaluation dataset). Co-occurrences of terms from the human proteins dictionary in these abstracts have been used as an indicator of relevance. Precision and recall have been measured as 0.89 and 0.48 respectively for the type 1 of co-occurrences. Afterwards, for abstracts ranked to be related to host-pathogen interactions in plants (application dataset), LAITOR has generated a list of co-occurrences and a network representation of the terms from the plant protein dictionary which could be found in this dataset. **Symbol key**: dark rectangles: dictionaries; grey shapes: abstract datasets; cylinders: public databases; ellipse: LAITOR script; white rectangles: LAITOR's outputs;

Barbosa-Silva et al. BMC Bioinformatics 2010, 11:70
http://www.biomedcentral.com/1471-2105/11/70

example the co-occurrences of the types 1, 2 and 3 are defined as follows.

Type 1: the pairs HSP90 and RAR1, as well as, HSP90 and SGT1 were both extracted with the interleaved biointeraction term "interact" associating the members of each pair (see Additional file 5, Table S2 for example of a Biointeraction term representation in the Biointeraction Dictionary).

Type 2: the pair RAR1 and SGT1 was extracted, with the occurrence of the biointeraction "interact" in the same sentence, however not interleaved.

Type 3: Other co-occurrences of the protein terms (HSP90, RAR1 and SGT1) found in the same sentence were considered as co-occurrences of type 3.

Furthermore, the combinations of all the bioentity names identified in the abstract, except synonyms, are considered as co-occurrences of type 4 (see Implementation for explanation).

Evaluation against BioCreative II

LAITOR was compared to the Interaction Article Subtask (IAS) of BioCreative II text mining challenge [14]. Table 1 shows that LAITOR could predict abstracts considered relevant for the curation of protein-protein interaction (evaluation dataset) with a maximum precision of 0.89 and a corresponding recall of 0.48 considering type 1 co-occurrences (bioentities co-occur within the same sentence, and they are interleaved by some biointeraction term; see Implementation for a detailed description). Among the 19 evaluated methods for the IAS task, LAITOR's predictions (considered to be a non SVM-based prediction) demonstrated to be the second most precise method keeping a reasonable sensitivity (recall) index. In predictions using the co-occurrence types 2-4, which do not require the presence of a biointeraction term, LAITOR produced results with a precision ranging from 0.82 to 0.85, a recall ranging from 0.61 to 0.70 and a F-score ranging from 0.60 to 0.72 (See Table 1 for values for each type). This implies that LAITOR's detection of protein co-occurrences with biointeraction terms improves precision that the expense of a small reduction of recall and therefore increases the likelihood of filtered protein pairs from such abstracts will indeed display biologically relevant fact.

Table 1 LAITOR evaluation against BioCreative II IAS subtask.

Type	Precision	Recall	F-score	Accuracy
1	0,89	0,48	0,63	0,63
2	0,85	0,61	0,71	0,68
3	0,83	0,62	0,72	0,68
4	0,81	0,70	0,60	0,60

Manual examination of some false-positive abstracts showed that although the biointeraction was not correctly identified, the selected sentences described a relevant biological interaction. For example, this sentence: "*Taken together, these results suggest that loss of RPA1 activates the Chk2 signaling pathway in an ATM-dependent manner*" (PMID: 15620706), was interpreted as RPA1 activates Chk2 because the term "activates" was found between the protein names RPA [Entrez Gene id: 6117] and Chk2 [Entrez Gene id: 11200]. The sentence actually indicates a different relation but it is informative in terms defining a functional relation between these two proteins.

In further comparison of LAITOR's performance with other methods from the BioCreative II challenge in order to correctly classify the IAS gold standard abstracts, we scored LAITOR's prediction of these abstracts with a score S = 5-T where T, that is the type of co-occurrence, ranges from 1 to 4, according to the presence of at least one sentence displaying a co-occurrence of types 1 to 4 (adopting S = 0 when no co-occurrence is detected in the abstract). Then, we calculated the area under the receiver operating curve (AROC), corresponding to 0.74 (Figure 2).

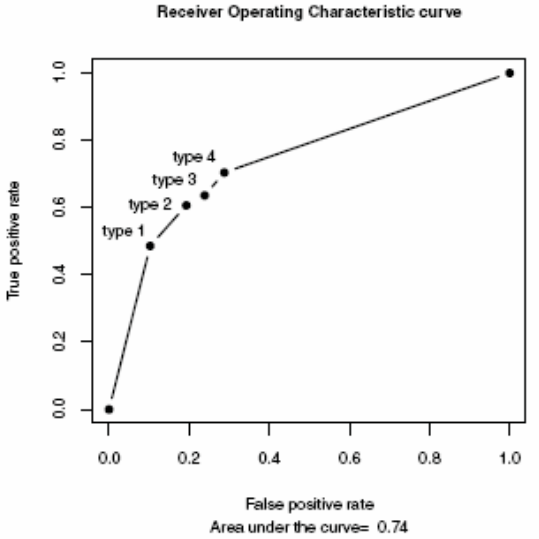


Figure 2 Receiver Operating Characteristic (ROC) curve for LAITOR predictions. The corresponding area under the curve (AROC) is 0.74, calculated using the four types of interactions found in such an abstract as a measurement of its overall predictive power. Note that higher types include the lowest ones.

Barbosa-Silva et al. BMC Bioinformatics 2010, 11:70
http://www.biomedcentral.com/1471-2105/11/70

Case-study: co-occurrence analysis of terms related to a plant-pathogen interaction dataset

We performed a case study by applying LAITOR to generate a list of green plant's protein co-occurrences related to host-pathogen interactions. Plants respond to diverse environmental stimuli, biotic and abiotic, by mobilizing specific protein networks used to identify its source and to activate the cellular mechanisms to surpass changes caused by stressful conditions. Commonly, the adaptative responses found in plants are flexible and the same subset of proteins/genes can be activated by different types of stimuli, including defense against pathogens or tolerance under severe environmental conditions [22]. Therefore, a system like LAITOR used in this context should be expected to be useful in suggesting novel roles for known protein interactions.

Moreover, this topic is important for plant biotechnological and physiological studies, since (i) diverse economically important crops are attacked by several phytopathogens in the field, which is prejudicial for agricultural practices along the world [23], and (ii) cultivated lands are often affected, for instance, by severe abiotic conditions such as high salinity [24], drought [25], over-flooding [26] or extreme cold [27]. As a result of this interest, during the last few decades several efforts have been dedicated to characterize these mechanisms, which resulted in a fair amount of related publications deposited in MEDLINE. These data comprises proteins or entire protein networks that are used by plants, as well as chemicals identified to have a key role in the signaling pathways that establish the plant adaptative responses. Jasmonic acid (JA) [28], ethylene (ET) [29] and salicylic acid (SA) [30,31] are examples of phytohormones employed by plants that act as signaling molecules in diverse defense response networks [32]. This wealth of data facilitates a text mining procedure such as LAITOR.

A total of 1,000 abstracts on the topic of green plant's host-pathogen interactions were retrieved with MedlineRanker [26] (application dataset) and analysed with LAITOR, of which 79 displayed at least one filtered co-occurrence. From the total 9,823 parsed sentences (including titles), 116 provided co-occurrences of the different types and pairs of bioentities (Table 2). A total of 263 pairs were retrieved from the application dataset.

In this dataset, a total of 68 different biointeraction terms could be identified among the co-occurring pairs, considering that the co-occurrences of type 3 do not restrict the filtering of biointeraction terms in the sentences. The top 10 most-common biointeraction terms and their frequencies within the application dataset are shown in Additional file 9, Table S3.

Table 2 Survey of sentences and pairs extraction using the LAITOR algorithm on application dataset.

Type	Sentences	Pairs
1	25	52
2	35	66
3	24	27
4	N. A.*	21
Total	116	263

*N.A.: not applicable, as LAITOR does not consider sentences to extract co-occurrences of type 4.

Network visualization

LAITOR generates a network file relating the co-occurrences extracted. The nodes represent bioentities and the edges their co-occurrences in the set of abstracts used as input. Each edge is annotated by the type of co-occurrence from strictest (type 1) to least strict (type 4).

As an example we generated a network for a total of 51 nodes and 143 edges found in the application dataset only representing the co-occurrences of type 1, in order to reduce the complexity of the network [Additional file 10, Figure S2]. We illustrate the relevance for the analysis of using the dictionary of concepts in Additional file 11, Figure S3. It can be noticed that the displayed subnetwork with 9 proteins (Additional file 11, Figure S3A; this is one of the subnetworks of the network represented in Additional file 10, Figure S2) gained two more members (catalase and SOD) when the concepts "oxidative stress" and "jasmonic acid" were also considered [see Additional file 11, Figure S3B]. The top 10 most-common terms present in the concept dictionary and their observed frequencies within the application dataset are shown in Additional file 12, Table S4.

Hypothesis generation example

One of the most interesting applications of a co-occurrence based text mining analysis is the support given to new hypothesis generation [33,34]. Here we explore this functionality in LAITOR by examining the involvement of a common member of the photosystem response and disease signaling in *Arabidopsis* [see Additional file 13, Figure S4].

Accessing the abstracts analyzed by LAITOR and listed in Additional file 13, Figure S4B we observe that the *Arabidopsis thaliana* gene *RPS4* (RESISTANT TO *P. SYRINGAE* 4 [Entrez GeneID: 834561]) confers resistance to the bacterial pathogen *Pseudomonas syringae* carrying the avirulence gene *avrRps4* [Entrez GeneID: 3555344, PMID: 8589423]. We can use LAITOR to find genes that could be hypothetically involved in resistance mechanisms regulated by *RPS4*. LAITOR associates this gene to several other genes. In

Barbosa-Silva et al. BMC Bioinformatics 2010, 11:70
http://www.biomedcentral.com/1471-2105/11/70

the topic of resistance against pathogens *EDS1* stands out: we can see that RPS4 requires the gene *EDS1* (ENHANCED DISEASE SUSCEPTIBILITY1 [Entrez GeneID: 823964]) to confer *avrRps4*-independent resistance in tomato plants transiently expressing *RPS4* [PMID: 15447648]. Using LAITOR we can see that there is another pathogen resistance gene that, similarly to *RPS4*, also requires EDS1, although in a different context [see Additional file 13, Figure S4A]. This is *PAD4* (PHYTOALEXIN DEFICIENT4 [Entrez GeneID: 824408]), which confers resistance against the phloem-feeding green peach aphid (GPA) infesting *Arabidopsis*, and also requires its signaling and stabilizing partner *EDS1* [PMID: 17725549].

Now, LAITOR shows that *PAD4* is related to three genes: *LSD1* [Entrez GeneID: 827786], *SIZ1* [Entrez GeneID: 836163], and *WIN3* [Entrez GeneID: 831173]. In more detail, a win3-T *Arabidopsis* (*WIN3*) mutant shows greatly reduced resistance to the bacterial pathogen *Pseudomonas syringae* carrying the avirulence gene *avrRpt2* and expression of this gene at an infection site partially requires *PAD4* [PMID:17918621]. The small Ubiquitin-like Modifier E3 Ligase (encoded by the gene *SIZ1*) interacts epistatically with PAD4 to regulate pathogenesis related gene expression and disease resistance [PMID: 17163880]. Finally, the disease resistance signaling components EDS1 and PAD4 are essential regulators of the cell death pathway controlled by LSD1 in *Arabidopsis* [PMID: 11595797].

Given the fact that both RPS4 and PAD4 require EDS1, one could explore weather or not these three known targets of PAD4 (SIZ1, WIN3, LSD1) could also be targets of RPS4, a fact not represented in the literature as evidenced by the absence of matches for the PubMed query "RPS4 AND (SIZ1 OR WIN3 OR LSD1)" [see Additional file 13, Figure S4C]. This example highlights the potential of LAITOR to unearth undiscovered public knowledge [35] using the condensed information of abstracts [36]. Thus, the system is able to extract precise information from the sentences in abstracts that can be used to generate new hypotheses.

Current limitations of LAITOR

The main limitations of the system can be classified as those producing false positives and those producing false negatives co-occurring pairs. False negatives are mainly due to terms not recognized to be gene/protein names, and to failure to recognize a biointeraction. The first problem can be solved by improving the tagging mechanism and the underlying dictionaries. We approach the second by manually adding to the dictionary of biointeractions those that we find to be common. Some false positives co-occurrences are caused due to misrecognition of gene/protein names and/or biointeractions. The current tagging is conservative and therefore does not increase false identification of gene/protein names (see Material and Methods); it actually constitutes the slower step of the method. This ensures that the identified biointeractions actually point to relevant sentences. Most falsely identified biointeractions were originating from sentences with large numbers of genes. We are considering adding an option to dismiss sentences with more than two gene/proteins as a choice for users requiring greater accuracy.

Comparison to other similar systems specialized in co-occurrence extraction

LAITOR is, as far as we know, the only method of co-occurrence detection along with customized that has been designed as standalone software to be included as part of other systems. However, LAITOR has some methodological particularities that merit comparison to recently developed systems that apply biological term co-occurrence as part of their functionalities.

STRING [37] is a web resource focused on a pre-compiled list of protein-protein interactions extracted by different methods. STRING uses Natural Language Processing [38] to search for statistically relevant co-occurrences of gene names, and also extract a subset of semantically specified interactions. Similarly, iHOP [15] is focused on the navigation of the scientific literature using biological term co-occurrence networks as a natural way of accessing PubMed abstracts. iHOP's text mining approach retrieves and ranks all the sentences for a given gene according to significance, impact factor of published journal, publication date or syntax structures where the gene occurs (i.e. gene-biointeraction-gene pattern). Furthermore, iHOP uses MeSH terms as source for information about gene function, what could be comparable to LAITOR's concepts search. Similarly to iHOP, co-occurrence methods have been developed for plant-directed literature analysis using *Arabidopsis thaliana* as a model [39]. This system, called PLAN2L, also classifies the extracted terms and co-occurrences as being related to physical and regulatory events for developmental processes, as well as with sub-cellular context, for that PLAN2L uses from co-occurrence to syntactic/semantic rule-based algorithms and supervised machine learning methods.

Although being designed for different purposes, we compared the features among LAITOR, STRING and iHOP (Table 3), once that these systems use biological term co-occurrences as part of their text mining strategies.

The main novelty of LAITOR in comparison to previous published software, besides the implementation of the concepts search, is the possibility to customize the dictionaries to be considered in the co-occurrence analysis (bioentities and biointeractions).

Reflecting this flexibility, we have included in the current LAITOR's distribution package a set of genes

Barbosa-Silva et al. BMC Bioinformatics 2010, 11:70
http://www.biomedcentral.com/1471-2105/11/70

Table 3 Comparison of features between LAITOR, STRING and iHOP.

Features	LAITOR	STRING	iHOP
Software type	Command-line script	Website application	Website application
Information sources	Any type of text loaded by the user (e.g. PubMed, OMIM, Wikipedia)	PubMed, SGD, OMIM, The Interactive Fly.	PubMed
Text limit	Any type of tagged text	Only abstracts	Only abstracts
Protein name tagging	Depends of external software (NLPROT), confers against loaded dictionary	YES, filtered by selected organism	YES, filtered by selected organism
List of used synonyms	Flexible user-based dictionary input	Variety of pre-compiled dictionaries	Entrez Gene, FlyBase, UniProt and HUGO Nomenclature Committee
Explores biological concepts	YES, finds user loaded concepts linked to a co-occurring pair at sentence level.	NOT	YES, searches species names, MeSH and compound terms
Extracts co-occurrences among proteins	YES, considering whole text and isolated sentences	YES, limited to the whole abstract	YES, at sentence level only
Extracts interactions among proteins	YES, considering a biointeractions dictionary defined by the user	NOT	YES, considering a pre-compiled biointeractions dictionary
Terms co-occurrences	YES, extracts terms mentioned in the full text or in isolated sentences at different structures which are scored differently	YES, extract terms mentioned together in abstracts, more often than what would be expected by chance based on their overall occurrence	YES, extracts terms mentioned in isolated sentences
Semantic understanding	YES, extracts the biointeractions and concepts linked to an extracted pair at sentence level in different co-occurrence types	NOT, only checks co-occurrences of terms	YES, extracts the biointeractions and concepts linked to an extracted pair at sentence level
Co-occurrence frequency report	YES, displays the frequency that a pair co-occurred in general sentences, and for each found biointeraction	YES, only the number of times that a pair co-occurred in each abstract	NOT
Outputs network	YES, in tabular format and in pre-compiled formats for third-part applications (ARENA3D, MEDUSA)	YES, displays the network in the browser from selected abstracts	YES, users can build a network by adding a set of nodes per time by selecting desired abstracts

symbols/synonyms dictionaries pre-compiled from GeneDB records and divided by all the organisms deposited NCBI's Taxonomy Database http://www.ncbi.nlm.nih.gov/Taxonomy, in addition to the green plants dictionary used in the test case described above, making it possible to use LAITOR virtually for any species with gene data. Furthermore, in order to provide users with a wide set of relevant dictionaries for the concepts search, we compiled LAITOR's concepts dictionaries for each of the NCBI's Medical Subject Headlines (MeSH) main tree structures http://www.nlm.nih.gov/mesh/trees2008.html. The information about how to use these dictionaries is available in the documentation file of LAITOR.

Conclusions

We presented here a new text mining software component called LAITOR, which performs co-occurrence analysis of scientific abstracts where biological entities are filtered from the tagged text using a user defined bioentity dictionary as support. Subsequently, a rule based system is used to detect the co-occurrence of such names along with biointeraction and, optionally, other biological terms provided by the Concepts Dictionary (such as stimuli), in scientific abstracts. We provide here an example of knowledge discovery by applying LAITOR to a subset of abstracts published about defense mechanisms in *Arabidopsis*. In this example, genes from different contexts (light and pathogen responses) have been placed together. Additionally, we have explored a new feature in biological text mining, which is the application of a user pre-defined concept dictionary in order to mine the literature and gather facts previously not reported together. Here, we have evidenced that the inclusion of the concept "oxidative stress" in the analysis conducted for *Arabidopsis* abstracts has brought two new members to a predicted gene network thought to be related to "jasmonic acid" signaling pathway.

Taken together, our results suggest that LAITOR is very precise in identifying abstracts of scientific literature mentioning interactions between genes and proteins. LAITOR is able to extract very variable types of

protein co-occurrences, no matter how they have been cited in the abstract. In our future work, we intend to adapt LAITOR components to an on-line tool, in which users, as well as computers (using the web services technology) will be able to load their desired literature and perform a LAITOR-based co-occurrence analysis that, integrated with other databases (for example, KEGG [40]), will provide a flexible framework for literature mining-based knowledge discovery.

Availability and requirements

LAITOR is distributed under the General Public License (GPL). Access http://laitor.sourceforge.net to obtain LAITOR's repository and its documentation from SourceForge.net.

LAITOR requires Linux as operating system, PHP version 5.3.2 or superior, MySQL version 5.0.45 or superior to run. Additional information is found on-line in the LAITOR documentation file.

Additional file 1: Application dataset.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S1.TXT]

Additional file 2: Table S1: Example of a protein term and its synonyms representation in the Protein Dictionary.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S2.DOC]

Additional file 3: Plant protein dictionary.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S3.TXT]

Additional file 4: Concepts dictionary.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S4.TXT]

Additional file 5: Table S2: Example of a biointeraction term represented in the Biointeraction Dictionary.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S5.DOC]

Additional file 6: LAITOR co-occurrence pipeline.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S6.PPT]

Additional file 7: Performance evaluation dataset.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S7.ZIP]

Additional file 8: Figure S1: Example of a tagged phrase output.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S8.TIFF]

Additional file 9: Table S3: Top-10 biointeraction terms most cited in the green plants application analysis.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S9.DOC]

Additional file 10: Figure S2: Full network created by LAITOR from application dataset.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S10.TIFF]

Additional file 11: Figure S3: Co-occurrence sub-networks generated by LAITOR.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S11.TIFF]

Additional file 12: Table S4: Top-10 concepts terms mostly cited in the co-occurrence analysis.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S12.DOC]

Additional file 13: Figure S4: Hypothesis generation supported by LAITOR output.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-11-70-S13.TIFF]

Abbreviations

PMID: PubMed Identifier; IAS: Interaction Article Subtask.

Acknowledgements

We are grateful to Venkata Satagopam, Evangelos Pafilis and RS for the training given to ABS at EMBL-Heidelberg during his external Ph.D training in Germany. This work has been developed as part of ABS Ph.D thesis which has been sponsored by Foundation for Research Support of Minas Gerais State (FAPEMIG), Brazilian Ministry of Education (CAPES/ME) and Brazilian Ministry of Science and Technology (CNPq/MCT). This work was supported by grants from Germany's National Genome Research Network (Bundesministerium für Bildung und Forschung) and from The Helmholtz Alliance on Systems Biology (Helmholtz-Gemeinschaft Deutscher Forschungszenten).

Author details

[1]Computational Biology and Data Mining Group, Max-Delbrück Center for Molecular Medicine. Robert-Rössle-Strasse. 10, D-13125, Berlin, Germany. [2]Laboratório de Biodados, Dpto. de Bioquímica e Imunologia, ICB - UFMG. 31270-901, Belo Horizonte - MG, Brazil. [3]European Molecular Biology Laboratory, EMBL-Heidelberg, Meyerhofstrasse 1, 69117, Heidelberg, Germany. [4]LIFE Biosystems GmbH. Poststrasse 34, D-69115, Heidelberg, Germany.

Authors' contributions

ABS created the main idea of the article. ILFM and TGS helped in the development and initial discussion of LAITOR algorithm. ABS and ILFM developed the prototype scripts. ABS developed the final scripts. TGS and RS provided the biointeraction dictionaries. GAP idealized the graph outputs. ABS performed the evaluation and application experiments. JFF and MAAN idealized the concept search and helped in the evaluation experiment. ABS wrote the article. JMO, MANN and RS corrected the article. JMO and RS supervised the initial development of LAITOR. JMO and MAAN supervised the final development of LAITOR. All authors read and approved the final version of the article.

Received: 7 August 2009
Accepted: 1 February 2010 Published: 1 February 2010

References

1. Andrade MA, Bork P: **Automated extraction of information in molecular biology.** *FEBS Lett* 2000, **476**:12-17.
2. Krallinger M, Valencia A: **Text-mining and information-retrieval services for molecular biology.** *Genome Biol* 2005, **6**:224.

3. Kostoff RN, DeMarco RA: **Extracting information from the literature by text mining.** *Anal Chem* 2001, **73**:370A-378A.
4. Blaschke C, Andrade MA, Ouzounis C, Valencia A: **Automatic extraction of biological information from scientific text: protein-protein interactions.** *Proc Int Conf Intell Syst Mol Biol* 1999:60-67.
5. Natarajan J, Berrar D, Dubitzky W, Hack C, Zhang Y, DeSesa C, Van Brocklyn JR, Bremer EG: **Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line.** *BMC Bioinformatics* 2006, **7**:373.
6. Yu H, Hatzivassiloglou V, Friedman C, Rzhetsky A, Wilbur WJ: **Automatic extraction of gene and protein synonyms from MEDLINE and journal articles.** *Proc AMIA Symp* 2002, 919-923.
7. Schuemie MJ, Weeber M, Schijvenaars BJ, van Mulligen EM, Eijk van der CC, Jelier R, Mons B, Kors JA: **Distribution of information in biomedical abstracts and full-text publications.** *Bioinformatics* 2004, **20**:2597-2604.
8. Rodriguez-Penagos C, Salgado H, Martinez-Flores I, Collado-Vides J: **Automatic reconstruction of a bacterial regulatory network using Natural Language Processing.** *BMC Bioinformatics* 2007, **8**:293.
9. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U: **AliBaba: PubMed as a graph.** *Bioinformatics* 2006, **22**:2444-2445.
10. Tari L, Hakenberg J, Gonzalez G, Baral C: **Querying parse tree database of Medline text to synthesize user-specific biomolecular networks.** *Pac Symp Biocomput* 2009:87-98.
11. Thu PH, Baral C, Gonzales G: **Generalized text extraction from molecular biology text using parse tree database querying.** Technical Report TR-08-004, Arizona State University 2008.
12. Fontaine JF, Barbosa-Silva A, Schaefer M, Huska MR, Muro EM, Andrade-Navarro MA: **MedlineRanker: flexible ranking of biomedical literature.** *Nucleic Acids Res* 2009, **37**:W141-146.
13. Mika S, Rost B: **NLProt: extracting protein names and sequences from papers.** *Nucleic Acids Res* 2004, **32**:W634-637.
14. Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A: **Overview of the protein-protein interaction annotation extraction task of BioCreative II.** *Genome Biol* 2008, **9(Suppl 2)**:S4.
15. Hoffmann R, Valencia A: **Implementing the iHOP concept for navigation of biomedical literature.** *Bioinformatics* 2005, **21(Suppl 2)**:ii252-258.
16. Blaschke C, Valencia A: **The potential use of SUISEKI as a protein interaction discovery tool.** *Genome Inform* 2001, **12**:123-134.
17. Persico M, Ceol A, Gavrila C, Hoffmann R, Florio A, Cesareni G: **HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms.** *BMC Bioinformatics* 2005, **6(Suppl 4)**:S21.
18. Chatr-aryamontri A, Kerrien S, Khadake J, Orchard S, Ceol A, Licata L, Castagnoli L, Costa S, Derow C, Huntley R, Aranda B, Leroy C, Thorneycroft D, Apweiler R, Cesareni G, Hermjakob H: **MINT and IntAct contribute to the Second BioCreative challenge: serving the text-mining community with high quality molecular interaction data.** *Genome Biol* 2008, **9(Suppl 2)**:S5.
19. Sing T, Sander O, Beerenwinkel N, Lengauer T: **ROCR: visualizing classifier performance in R.** *Bioinformatics* 2005, **21**:3940-3941.
20. Hooper SD, Bork P: **Medusa: a simple tool for interaction graph analysis.** *Bioinformatics* 2005, **21**:4432-4433.
21. Pavlopoulos GA, O'Donoghue SI, Satagopam VP, Soldatos TG, Pafilis E, Schneider R: **Arena3D: visualization of biological networks in 3D.** *BMC Syst Biol* 2008, **2**:104.
22. Fujita M, Fujita Y, Noutoshi Y, Takahashi F, Narusaka Y, Yamaguchi-Shinozaki K, Shinozaki K: **Crosstalk between abiotic and biotic stress responses: a current view from the points of convergence in the stress signaling networks.** *Curr Opin Plant Biol* 2006, **9**:436-442.
23. Rommens CM, Kishore GM: **Exploiting the full potential of disease-resistance genes for agricultural use.** *Curr Opin Biotechnol* 2000, **11**:120-125.
24. Tuteja N: **Mechanisms of high salinity tolerance in plants.** *Methods Enzymol* 2007, **428**:419-438.
25. Seki M, Umezawa T, Urano K, Shinozaki K: **Regulatory metabolic networks in drought stress responses.** *Curr Opin Plant Biol* 2007, **10**:296-302.
26. Jackson MB, Colmer TD: **Response and adaptation by plants to flooding stress.** *Ann Bot (Lond)* 2005, **96**:501-505.
27. Sharma P, Sharma N, Deswal R: **The molecular biology of the low-temperature response in plants.** *Bioessays* 2005, **27**:1048-1059.
28. Wasternack C: **Jasmonates: an update on biosynthesis, signal transduction and action in plant stress response, growth and development.** *Ann Bot (Lond)* 2007, **100**:681-697.
29. Broekaert WF, Delaure SL, De Bolle MF, Cammue BP: **The role of ethylene in host-pathogen interactions.** *Annu Rev Phytopathol* 2006, **44**:393-416.
30. Loake G, Grant M: **Salicylic acid in plant defence–the players and protagonists.** *Curr Opin Plant Biol* 2007, **10**:466-472.
31. Pieterse CM, van Loon LC: **Salicylic acid-independent plant defence pathways.** *Trends Plant Sci* 1999, **4**:52-58.
32. Kachroo A, Kachroo P: **Salicylic acid-, jasmonic acid- and ethylene-mediated regulation of plant defense signaling.** *Genet Eng (N Y)* 2007, **28**:55-83.
33. Kell DB, Oliver SG: **Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era.** *Bioessays* 2004, **26**:99-105.
34. Ananiadou S, Kell DB, Tsujii J: **Text mining and its potential applications in systems biology.** *Trends Biotechnol* 2006, **24**:571-579.
35. Swanson DR: **Fish oil, Raynaud's syndrome, and undiscovered public knowledge.** *Perspect Biol Med* 1986, **30**:7-18.
36. Ding J, Berleant D, Nettleton D, Wurtele E: **Mining MEDLINE: abstracts, sentences, or phrases?.** *Pac Symp Biocomput* 2002, 326-337.
37. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C: **STRING 8–a global view on proteins and their functional interactions in 630 organisms.** *Nucleic Acids Res* 2009, **37**:D412-416.
38. Saric J, Jensen LJ, Ouzounova R, Rojas I, Bork P: **Extraction of regulatory gene/protein networks from Medline.** *Bioinformatics* 2006, **22**:645-650.
39. Krallinger M, Rodriguez-Penagos C, Tendulkar A, Valencia A: **PLAN2L: a web tool for integrated text mining and literature-based bioentity relation extraction.** *Nucleic Acids Res* 2009, **37**:W160-165.
40. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 1999, **27**:29-34.

doi:10.1186/1471-2105-11-70
Cite this article as: Barbosa-Silva *et al.*: LAITOR - Literature Assistant for Identification of Terms co-Occurrences and Relationships. *BMC Bioinformatics* 2010 **11**:70.

5.2 SEED LINKAGE: UM PROGRAMA PARA AGRUPAMENTO DE PROTEÍNAS COGNATAS EM GENOMAS DISTINTOS A PARTIR DE LIGAÇÕES MÚLTIPLAS A UMA SEQÜÊNCIA FUNDADORA.

Apresentamos neste artigo um novo método para criar agrupamentos de proteínas relacionadas do ponto de vista de similaridade de seqüência a partir de uma "seed" original (seqüência fundadora).

Conforme descrevemos inicialmente, existem diversos métodos de se detectar proteínas que provavelmente compartilham a mesma função através da similaridade de seqüência apresentada por elas. Esta estratégia permite a anotação "a priori" de proteínas desconhecidas a partir daquela designada para proteínas bem estudadas. Ao mesmo tempo, seqüências similares mostram indicativos de descendência evolutiva comum, e que é considerado pelo nosso método adotando o sistema de classificação proposto por Koonin e Sonnhammer.

Alguns métodos desempenham uma função similar à proposta neste artigo, como o Multiparanoid e o COG. Esses métodos, entretanto, foram desenvolvidos para agrupar proteínas similares quando proteomas completos das espécies envolvidas são analisados. Tais estratégias criam grupos de proteínas co-ortólogas e parálogas internas para cada organismo. Contudo, foi mostrado que na base de dados COG a chance de inclusão de proteínas parálogas externas, indesejáveis em tais agrupamentos, é maior do que na base Multiparanoid-DB.

Em ambas estratégias, caso o usuário esteja interessado nas proteínas similares que possam compartilhar uma função semelhante a uma determinada proteína de interesse, ele teria que localizar qual agrupamento pré-definido para as espécies analisadas contém esta proteína, e limitar a análise dentro das espécies utilizadas pelos supracitados métodos, sem ter a chance de estender a análise para outros organismos de interesse com os mesmos critérios em que os agrupamentos foram criados (métodos KOG e Multiparanoid).

Através do Seed Linkage, método apresentado neste artigo, o usuário pode analisar para todos os organismos presentes na base de dados, quais seriam as proteínas que se agrupam com uma "seed" original, sem ter que para isso utilizar como "input" da análise todas as proteínas de todos os organismos envolvidos, um processo que demanda tempo e intensa atividade computacional.

O Seed Linkage baseia-se na estratégia fundamental adotada pelos métodos COG e Multiparanoid: a existência de melhor hit bi-direcional (bidirectional best hit, BBH) entre os pares de ortólogos mais prováveis entre pelo menos dois organismos diferentes (três no caso do COG). Em nosso caso, é requerida pelo menos identidade e percentual de alinhamento mínimos de 50% contra a "seed", para daí proceder com a captura de seqüências parálogas internas em cada um dos organismos considerados. Entretanto, um conjunto de regras é aplicado para tornar mais rigorosa a inclusão de novas seqüências nos grupos. Por exemplo, a presença de um BBH da seqüência "seed" com qualquer organismo da base de dados estabelece um limite para inclusão

de seqüências parálogas internas em ambos organismos sujeitos a BBH; tanto no organismo "seed" (aquele da seqüência original) quanto nos organismos candidatos (aqueles sujeitos ao BBH). As seqüências parálogas internas só são incluídas se apresentarem melhor hit contra o par do BBH correspondente no seu próprio organismo e com escore de alinhamento maior ou igual ao escore entre o par do BBH ao par correspondente no outro organismo. Em casos excepcionais de duplicação linhagem específica recente, a seqüência candidata a paráloga interna que não possui o par do BBH mas sim uma terceira seqüência como melhor hit nesse alinhamento só é incorporada ao "cluster" se esta terceira seqüência já houver sido previamente incorporada como paráloga interna. Ao final da busca por parálogos internos no organismo "seed", estas seqüências são elevadas à condição de "seed" para iniciar todo o processo contra os organismos candidatos, o que reduz o viés para escolha aleatória de seqüências arbitrárias a serem utilizadas como seed original.

Para tratar casos onde a seqüência original utilizada como "seed" não estabelece BBH contra nenhum organismo da base de dados, nós estudamos alguns parâmetros baseados em limites de escore bruto e relativo para inclusão de seqüências parálogas internas mesmo com a ausência de um BBH. Estes resultados mostraram que, para o conjunto controle adotado, o limite de escore relativo (o percentual do valor do escore da "seed" contra ela mesma) que uma paráloga interna precisa apresentar contra a "seed" deve ser maior ou igual a 30%. Nota-se que este parâmetro é razoável, uma vez que naturalmente, para todos os casos, limites de escore mínimo e cobertura de alinhamento precisam ser previamente superados.

Embora tenha sido desenvolvido inicialmente para criar agrupamentos a partir de uma única seqüência "seed" original, o método Seed Linkage pode ser aplicado para múltiplas "seeds" através da desambigüação dos agrupamentos formados para cada uma das "seeds" individuais. Dessa maneira, o Seed Linkage pôde ser comparado ao Multiparanoid, uma vez que este é o método que se sobressai na definição co-ortólogos e parálogos internos nos agrupamentos de proteínas para múltiplos organismos.

Outras duas avaliações foram realizadas no método Seed Linkage. A primeira com relação à capacidade do método ser aplicado a seqüências de organismos com proteoma incompleto; para isso conduzimos simulações através da utilização de agrupamentos onde os todos os organismos envolvidos apresentavam mais de quatro seqüências agrupadas. Nesses agrupamentos, de uma a três seqüências foram removidas, e as demais seqüências dos outros organismos foram usadas como "seed" na tentativa de reconstruir o grupo. A segunda avaliação testou o poder de agrupamento do Seed Linkage quando apenas um round de interação foi permitido, o que torna o método similar ao método RBH (reciprocal best hit), dessa maneira apenas um parálogo interno é considerado no organismo "seed" e os organismos candidatos estão sujeitos apenas a um BBH (o provável ortólogo da "seed").

O método Seed Linkage diferencia-se dos previamente publicados na literatura pela possibilidade de focalizar o agrupamento de proteínas similares entre organismos distintos com base numa única proteína inicial de interesse, o que simplifica a identificação de proteínas que possam desempenhar um papel biológico semelhante à proteína "seed" escolhida. O método utiliza as definições propostas por Koonin e Sonnhammer para classificar proteínas relacionadas como co-ortólogas ou parálogas internas, caso a comparação seja feita entre organismos diferentes ou dentro do próprio organismo, respectivamente. Como o nosso interesse não é estabelecer as relações filogenéticas entre as seqüências estudadas, e sim posicionar no mesmo agrupamento seqüências protéicas similares que possam também compartilhar a mesma função biológica, usamos as definições acima indicadas apenas para facilitar a implementação e o controle do nosso algoritmo. Dessa maneira, dados dois organismos A e B num mesmo agrupamento, todas as seqüências do organismo A são consideradas co-ortólogas a todas as seqüências do organismo B, assim como as seqüências do organismo A são consideradas parálogas internas entre si, o mesmo para as seqüências do organismo B.

Diferentemente dos métodos anteriores, o Seed Linkage permite que a relação de BBH seja indireta, proporcionando que "seeds" escolhidas arbitrariamente pelo usuário não criem um viés na análise. Para isso, mesmo que o melhor hit de uma proteína de interesse num determinado organismo candidato não corresponda à seqüência BBH, caso esta seqüência estabeleça um BBH com um parálogo interno previamente selecionado, este é considerado um BBH, e habilitado a iniciar a busca por seus parálogos internos no organismo candidato. Da mesma maneira, durante a busca pelos parálogos internos, seja no organismo "seed" ou nos candidatos, o melhor hit das seqüências candidatas a parálogas internas pode ser tanto a "seed" original (ou o BBH) ou qualquer outra seqüência previamente agrupada como paráloga interna.

Além de permitir o agrupamento de seqüências relacionadas a uma "seed" desejável, o método Seed Linkage pode também ser aplicado na criação de agrupamentos múltiplos, sem que para isso todo o proteoma precise ser analisado. Neste caso, basta o usuário fornecer uma lista com todas as proteínas originais a serem utilizadas como "seed". Em seguida, depois que os agrupamentos individuais forem estabelecidos, regras de desambiguação são criadas para resolver casos de redundância entre os grupos. Dessa maneira, é possível através do método Seed Linkage focalizar nas proteínas de interesse e evitar demasiado processamento computacional para que as seqüências de todo o proteoma sejam analisadas.

O método Seed Linkage é implementado em PHP e documentado de maneira tal que possa ser distribuído gratuitamente aos membros da comunidade científica interessados no estudo de algoritmos para agrupamento de seqüências de proteínas, ou apenas àqueles interessados em utilizar os recursos do sistema de uma maneira fácil e rápida. Este método foi publicado na revista *BMC Bioinformatics* com o título: "**Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence**" (Barbosa-Silva et al., 2008).

BMC Bioinformatics

BioMed Central

Software

Open Access

Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence

Adriano Barbosa-Silva[1,2], Venkata P Satagopam[2], Reinhard Schneider[2] and J Miguel Ortega*[1]

Address: [1]Laboratório de Biodados, Dep. Bioquímica e Imunologia, Instituto de Ciências Biológicas, UFMG, Av. Antônio Carlos 6627, Belo Horizonte, MG, Brasil and [2]European Molecular Biology Laboratory, EMBL-Heidelberg, Meyerhofstr 69117, Heidelberg, Germany

Email: Adriano Barbosa-Silva - barbosa@embl.de; Venkata P Satagopam - venkata.satagopam@embl.de; Reinhard Schneider - schneider@embl.de; J Miguel Ortega* - miguel@icb.ufmg.br

* Corresponding author

Published: 5 March 2008

BMC Bioinformatics 2008, **9**:141 doi:10.1186/1471-2105-9-141

Received: 31 October 2007
Accepted: 5 March 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/141

Abstract

Background: Modern proteomes evolved by modification of pre-existing ones. It is extremely important to comparative biology that related proteins be identified as members of the same cognate group, since a characterized putative homolog could be used to find clues about the function of uncharacterized proteins from the same group. Typically, databases of related proteins focus on those from completely-sequenced genomes. Unfortunately, relatively few organisms have had their genomes fully sequenced; accordingly, many proteins are ignored by the currently available databases of cognate proteins, despite the high amount of important genes that are functionally described only for these incomplete proteomes.

Results: We have developed a method to cluster cognate proteins from multiple organisms beginning with only one sequence, through connectivity saturation with that Seed sequence. We show that the generated clusters are in agreement with some other approaches based on full genome comparison.

Conclusion: The method produced results that are as reliable as those produced by conventional clustering approaches. Generating clusters based only on individual proteins of interest is less time consuming than generating clusters for whole proteomes.

Background

Modern proteomes are generated from ancestral ones by modifications that occur at the DNA level of the corresponding coding genomes. Such modifications (termed genetic variations) have different sources, among them: mutations, genetic recombination and alternative splicing (the last occurring at the RNA level). All generate variability in the protein repository present in one population. As a result, after isolation and speciation events, populations carrying closely-related proteomes can produce highly related protein sets.

However, a great part of the cognate proteins encoded by distinct proteomes is strictly similar, at the sequence level, to their counterparts in related species. This similarity is more than structural, often reflecting also in the function of these proteins in the biological system. Proteins derived from a common ancestor are termed homologs.

BMC Bioinformatics 2008, **9**:141 http://www.biomedcentral.com/1471-2105/9/141

There are different subtypes of homology relationships attributed to proteins based on an evolutionary point of view. Among these, we highlight those proposed by Sonnhammer and Koonin [1] which defines orthologs, inparalogs and out-paralogs as subtypes of homolog protein/genes, which can be operationally used by bioinformatics tools.

Several approaches have been designed to cluster related sequences from different organisms into the same ortholog group. Some of them use either all-versus-all alignments among different species [2] or pair-wise alignment among target organisms [3] as well. Each of these techniques is based on information deposited for fully sequenced genomes, and generate distinct ortholog groups using customized algorithms or thresholds in their searches [4-6].

Despite the prosperity of methods for definition of ortholog proteins among complete proteomes, there are few developments when it is desired to define such groups when poorly sequenced or unfinished proteomes are included in the search. Furthermore, most of the available methods are used in a high-throughput way, considering the whole protein dataset.

Here, we propose a methodology for finding highly-related groups of proteins to one single desired Seed sequence, classifying each of them as potential orthologs or in-paralogs into complete and or unfinished proteomes.

Methods

Algorithm

Input dataset

Seed Linkage runs using as input a fasta formatted database and a MySQL [7] table that contains taxonomic information (obtained from NCBI Taxonomy database) for all sequences contained in the database. The fasta file is expected to contain seed sequences and all candidate sequences with which one aims to establish a clustering relationship. Using the access to the MySQL database the program automatically recognizes which sequence belongs to which organism, a procedure that facilitates the setup and limits BLAST searches to the set of interest (the ongoing hits). We have tested and established the best thresholds that allows a sequence to be selected as a correct member of a group, which allows the possibility of working with sequences from organisms with incompletely-sequenced genomes.

Alignment details

Alignments between sequences are obtained with NCBI BLAST: BLASTp program [8], 10^{-10} E-value cutoff, low complexity filter off (-F f), tabular output (-m 8). BLAST parameters specified in the Seed Linkage configuration file can be altered by the user. To minimize the problem of fused genes/domains we defined 50% as the cut-off for both minimum identity and alignment coverage (with the Seed sequence), as explained by Remm *et al*. [3]; this too can be customized.

The main trait that distinguishes the Seed Linkage method from other approaches is the manner by which the alignments are conducted. The seed protein from the Seed Organism is used in a BLAST search against the full database, and the best hit from each organism is used as secondary query sequence. The secondary query is considered a bidirectional best hit subject (BBHsj) for its organism when the original Seed is its best hit from the Seed Organism. The BBHsj is considered as a putative ortholog in that organism; from the information contained in these alignments inparalogs are gathered as described below.

Inparalog search into Seed Organism

The Seed-to-BBHsj score from the best scoring organism is used as a threshold to limit inparalog inclusion for proteins encoded by the Seed Organism. This means that only inparalog candidates that are more (or equally) similar to the seed sequence than the highest scoring BBHsj will be grouped with the Seed (Figure 1). Besides surpassing the thresholds, an inparalog candidate is grouped only if it shows a BBH relationship with either the Seed or with an already-grouped inparalog.

If the Seed does not establish any BBH relationship with sequences from other organisms, the *inparalog retrieval score limit* in the Seed Organism is set to a minimum value for the parameter called SEED-Inparalog$_{relative_score}$ given by the formula:

$$\text{SEED-Inparalog}_{relative\_score} = \text{Score\_inparalog\_vs\_SEED} / \text{Score\_SEED\_vs\_SEED} \quad (1)$$

This parameter is set to a default value of 0.3 (see 'Results') but can be customized during setup. This means that all inparalog candidates have to present a score that is at least 30% of the score of the Seed against itself. Similarly, for other searches, an inparalog candidate is grouped only if it shows a BBH relationship with either the Seed or with an already grouped inparalog (Figure 1, black or upstream grey diamonds).

Inparalog search for Candidate Species

Candidate Species are organisms where a BBHsj was found by the initial alignment to the Seed. The score of the BBHsj against the Seed is used to limit the inparalog search within this respective species, as shown in Figure 1. Moreover, an inparalog candidate is grouped only if it

BMC Bioinformatics 2008, 9:141 http://www.biomedcentral.com/1471-2105/9/141

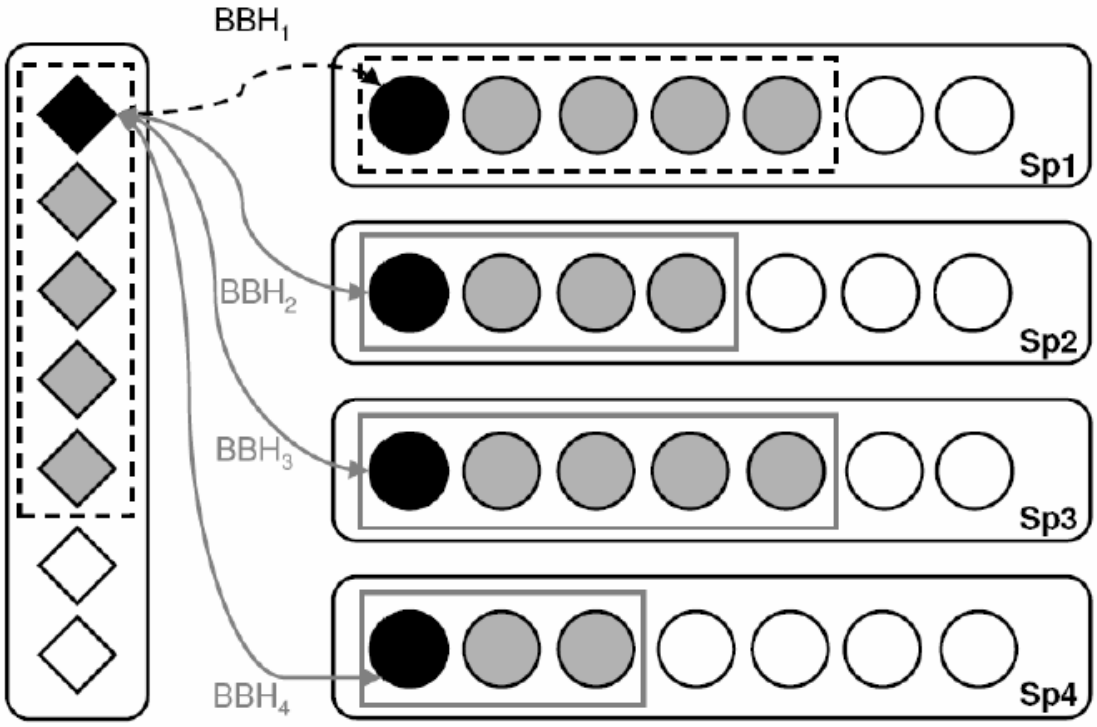


Figure 1
BBH algorithm adopted for Seed Linkage. The algorithm starts by aligning the Seed sequence from the Seed Organism (black diamond) to sequences from all other organisms in the database (circles in Candidate Species), searching for a BBH for each Species. The score BBH_1 between the Seed and the highest-scoring sequence (black circle in Sp1) defines the *inparalog retrieval score limit* in Seed Organism. The inparalogs for the Seed Organism are those sequences whose alignment score between the Seed and the potential inparalog (grey diamonds) exceeds BBH_1 (dashed boxes within Seed Organism). The BBH scores (BBH_{1-4}) are used to filter potential inparalogs (grey circles) from the respective Candidate Species (Sp1-4, respectively) when the BBHsj from each species (black circles) are used as secondary queries against proteins from the Candidate Species genome. These thresholds aim to avoid the inclusion of additional spurious sequences in clusters (white diamonds and circles). Inclusion requires a BBH relationship between candidates (grey symbols to be incorporated) and already grouped sequences (black and grey symbols) within the respective Candidate Species.

shows a BBH relationship with either the BBHsj or with an already grouped inparalog.

Iteration

All inparalogs from the Seed Organism, but not the grouped proteins from the Candidate Species (BBHsj and its inparalogs), are brought to the condition of Seed and the process of clustering is repeated either until it converges or by a limited number of rounds, set by the parameter "r" in the script. The default is r = 10, but this can also be customized during input. That means that all inparalogs from the Seed Organism are allowed to gather additional inparalogs, plus orthologs and their respective inparalogs in Candidate Species, until at the most the 10th search for inparalogs from the Seed Organism is done. However, iterations will always respect the *inparalog retrieval score limit* defined by the original Seed in the Seed Organism; and the score between the added inparalog used as Seed against its BBHsj as threshold to limit the inparalog inclusion from the Candidate Species.

Search of almost identical inparalog candidates

The tree diagram in Figure 2 exemplifies a case of what we term 'hidden inparalogy'. Consider sequence A1 from organism A (as either Seed or BBHsj), and two other sequences A2 and A3 generated by a lineage-specific duplication event so they are more similar to each other than to sequence A1. Then, A2 and A3 will not match A1

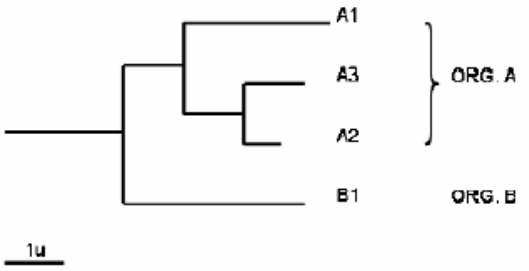


Figure 2
Distance tree showing very similar inparalogs A2 and A3. The best hit of sequence A3 (Seed), within organism B, is B1. However, when B1 is reciprocally aligned against sequences from organism A, A2 is the best hit, not A3, since A2 and A3 are very similar inparalogs. B1 is considered to form a BBH with A3 (Seed) if A3 appears at the 2nd position in the reciprocal alignment and A2 has a BBH relationship with A3. Another 'hidden inparalogy' case occurs with inparalog searches beginning with A1 being A2 and A3 inparalog candidates. A2 is considered to form a BBH with A1 (Seed) if A1 appears at the 2nd position in the reciprocal alignment and A2 has a BBH relationship with A3.

reciprocally; instead, they will match each other. To solve this problem, when an already-grouped protein (black and grey symbols in Figure 1, e.g. A1) is 2nd in the returned alignment, the candidate is considered to have a reciprocal match, given that the pair A2-A3 does establish a BBH.

Search initiated by a Seed with an almost identical inparalog

Similar to the previous case, when a sequence A3 from organism A is used as Seed in a search for BBHsj from organism B, often sequence B1 matches reciprocally A3 in the organism A. However, consider sequence A2, an inparalog of A3, being more ancient in Organism A than A3. In this scenario, B1 will match reciprocally A2 instead of A3. To cope with this event, when Seed (e.g. A3) is 2nd in the returned alignment, the candidate is considered to have a reciprocal match, given that the pair A2-A3 actually establishes a BBH.

Batch search

Though not aiming to generate clusters starting with multiple seed sequences, the Seed Linkage approach permits users to create such clusters by initially defining individual clusters through a batch search executed by the program, followed by the application of disambiguation rules. Thus, the user does not need to verify whether or not Seed sequences may be inparalogs.

Cluster disambiguation

Considering two clusters *i* and *j*, with size (number of sequences) N*i* and N*j*, we defined four relationships between clusters *i* and *j*:

(1) *i* is contained in *j*: in this case, if all the sequences of cluster *i* also belongs to cluster *j*, then cluster *i* is deleted.

(2) *i* is identical to cluster *j*: in this case, just one of the clusters is maintained.

(3) *i* is completely distinct from *j*: in this case both clusters are kept separately.

(4) *i* has elements in common with cluster *j*: in this case, if cluster *j* is larger than cluster *i*, and more than 50% of the sequences in cluster *i* are present in cluster *j*, than clusters *i* and *j* are merged, otherwise they are maintained separately.

Implementation

Core scripts

The algorithm described above was implemented as a Linux command line script written in PHP command line interface [9]. The script is connected to a MySQL database where the information regarding to the sequence source database is stored as simple tables. Furthermore, it is also necessary to set a fasta sequence database and the path of the alignment software must be edited in the script. To facilitate database formatting we have developed a configuration file, in which the parameters pertaining to the script can be easily adjusted.

The main package consists of three files, clearly documented. Additional scripts to parse the model databases are provided within the package as Additional file 1.

Algorithm evaluation

Manually curated database and non-related sequences

To validate the Seed Linkage approach a manually-curated dataset of 1363 trans-membrane proteins [10] previously grouped into 221 reference clusters was used as Seed against the proteome sequences of the following organisms: *C. elegans* (Cel), *D. melanogaster* (Dme) and *H. sapiens* (Hsa), comprising 19099, 14100 and 35118 sequences, respectively. The rebuilt clusters (RCs) were disambiguated and the resultant clusters were compared to the original reference clusters.

Results

To develop and verify a procedure that results in clusters of proteins linked to seed sequences, we have chosen to assay the same manually-curated database of trans-membrane proteins that was used to verify clusters using the Inparanoid procedure. Briefly, Inparanoid drives auto-

BMC Bioinformatics 2008, 9:141 http://www.biomedcentral.com/1471-2105/9/141

matic clustering of orthologs and inparalogs shared by different organisms with completely-sequenced genomes [3,11] or even by multiple proteomes from large taxonomic groups [4]. The manually-curated database contains 1363 sequences that are expected to constitute 221 manual clusters (MCs). The Seed Linkage procedure was applied to each sequence as Seed. As a challenge to correctly form clusters without including additional sequences, the BLAST database also included the complete proteomes from worm, fly and man, adding up to 66,954 entries. Grouped inparalog candidates were then iteratively treated as new Seeds and the resultant clusters were disambiguated.

Clustering

The batch search script provided by the Seed Linkage package was used to run the 1363 individual processes. As detailed in "Methods," only the cutoff for the parameter SEED-Inparalog$_{relative_score}$ was omitted. Moreover, for the primary analysis (Figures 3, 4, 5, 6, 7, 8), cluster disambiguation was not performed. Of the generated clusters, 114 Seeds (8.36%, belonging to 17 curated clusters) remained as clusters of size 1 (actually not forming clusters), while the remaining 1249 sequences formed clusters whose size ranged from 2 up to 103 sequences (Figure 3). A lesser number of Seeds (173 sequences, 12.6%) participated in the largest clusters (>11 members). Sequences grouped by Seed Linkage always require a BBH relationship with either a Seed or previously grouped sequences so, in this experiment, a total of 7289 BBH events occurred, with 5582 events (76.6%) composed of sequences from the original trans-membrane dataset, and 1707 events involving additional sequences (737 distinct sequences). Thus, additional filtering seemed to be necessary to reduce the chance of mis-inclusion of spurious sequences; some of these might be out-paralogs, which could have diverged significantly to acquire new functionalities [1].

Inparalog threshold evaluation

The thresholds (similarity = 50, alignment coverage = 50 and E-value < 10^{-10}) used in BLAST alignments, together with the requirement of a BBH relationship for the sequence to be grouped, were not enough to limit the inclusion of additional sequences. In clusters lacking a reference BBHsj, this could be even more relevant since the search for inparalogs from the Seed Organism might tend to include outparalogs, which iteratively could include undesirable sequences. We decided to investigate two variables potentially useful in this task: raw and relative scores.

Raw score

We graphed the number of sequences from the original (trans-membrane) and additional (complementary proteomes) datasets that were recruited into the RCs against the raw score between each accepted inparalog candidate and either Seed (in Seed Organism) or BBHsj (in Candidate Species). Due to the expected possibility that the same sequence is gathered by different query sequences into distinct clusters, we reported all inclusion events for both original and additional sequences. The analysis was applied to clusters initiated either with or without a BBHsj reference hit (+BBHsj, -BBHsj). Figure 4 shows the raw score distribution in the batch search. For clusters initiated with a BBHsj (Figure 4a), the distribution of raw score resulting in the recruitment of additional sequences (filled symbols) resembles, although in lower proportion, that of original sequences (open symbols). This suggests that additional filtering would not significantly avoid gathering undesirable proteins. In these cases, the score Seed/BBHsj might have acted as a natural *inparalog retrieval score limit* for inparalog inclusion. However, the distribution of raw scores presented during the recruitment of additional sequences in the absence of a BBHsj (Figure 4b, filled symbols) was remarkably concentrated at the low raw scores, indicating that an *inparalog retrieval score limit* could be set to a value such as 400 bits, since 58% of the gathering events for additional sequences were concentrated below this range, as opposed to 19% for original sequences.

Relative score

The second analyzed parameter was the relative score. The ratio between the score of each gathering event and the score of Seed self-alignment was recorded. Again, in the presence of a BBHsj, the distribution of relative scores for the additional sequences is not distinct from those for the original ones (Figure 5a). However, Figure 5 shows that 57% of the gathering events of additional sequences occur in relative score ranges less than or equal to 0.3 (Figure 5b, filled symbols). In the same range, only 8% of gathering events of original sequences happen.

Alternative inparalogs thresholds definition

Considering the results above, we tested raw score 400 and relative score 0.3 as thresholds to minimize the inclusion of additional sequences to the MCs during Seed Linkage rebuilding. Data in Figure 6a shows that gathering events of original sequences represent 34% and 46% of the total alignments in clusters with and without a BBHsj, while gathering events of additional sequences represent 3% and 16%, respectively.

Adopting a threshold for inclusion of sequences with relative score higher than 0.3 (Figure 6b), the percentage of gathering events of original sequences decreased very little (to 33% and 42%, in clusters with and lacking a BBHsj, respectively). Similar behavior was observed for gathering events of additional sequences in clusters with a BBHsj

BMC Bioinformatics 2008, 9:141 http://www.biomedcentral.com/1471-2105/9/141

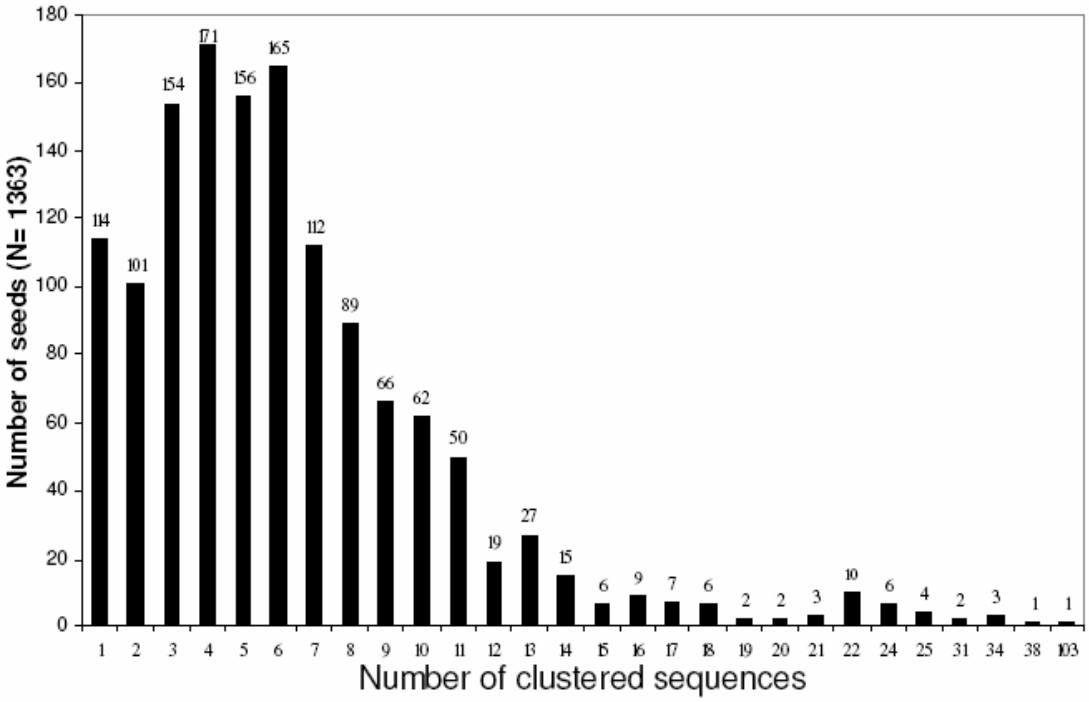


Figure 3
Distribution of number of sequences clustered by Seeds.

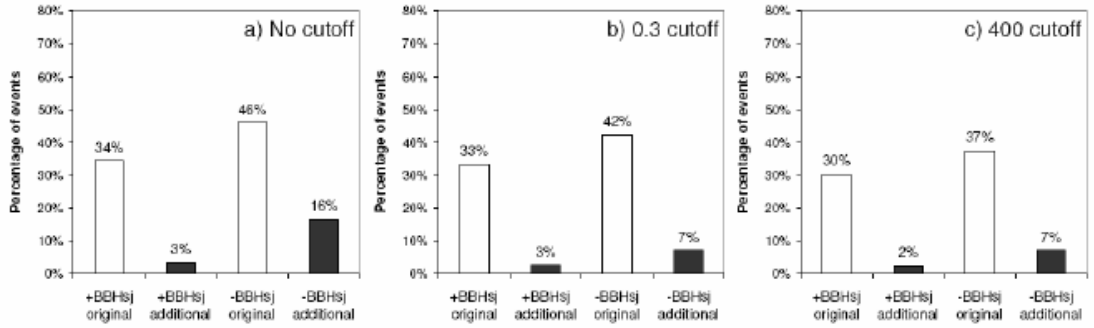


Figure 6
Effects of relative and raw score in the percentage of sequences included into rebuild clusters. Relative score of 0.3 (b) or raw score cutoff of 400 bits (c) were applied. Legend: '+BBHsj', clusters with BBHsj reference sequence;'-BBHsj', clusters lacking a BBHsj reference sequence.

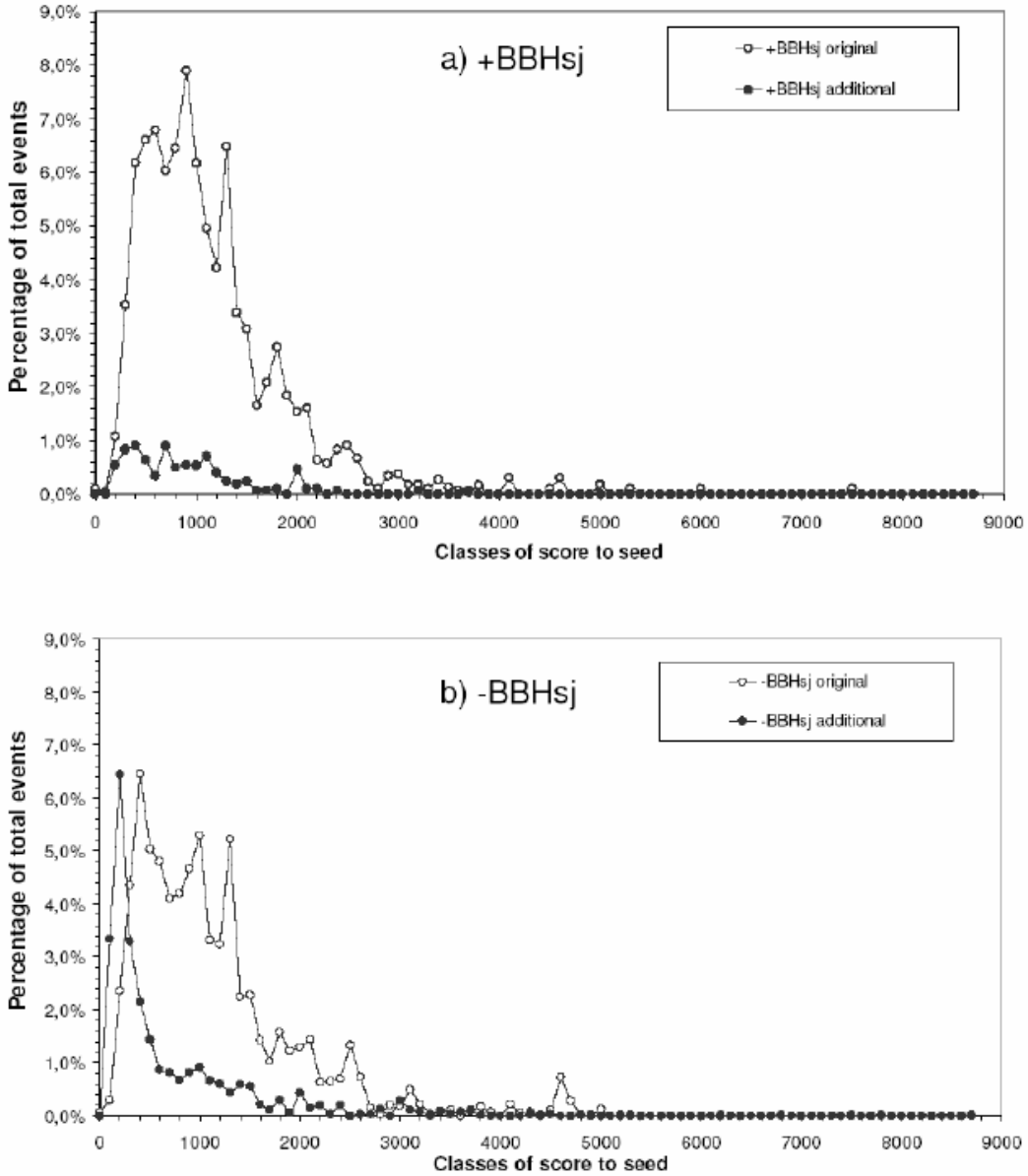


Figure 4
Raw score distribution in clusters with (a) or lacking (b) a BBHsj reference sequence. The alignment score to Seed for each gathering event was saved and the percentage of events within each class of score interval (binned in 50-bit increments) was determined. Legend: '+BBHsj original' and '+BBHsj additional', gathering events involving sequences respectively present or absent in manual clusters, in clusters with BBHsj reference sequence; '-BBHsj original' and '-BBHsj additional', gathering events involving sequences respectively present or absent in manual clusters, in clusters lacking a BBHsj reference sequence.

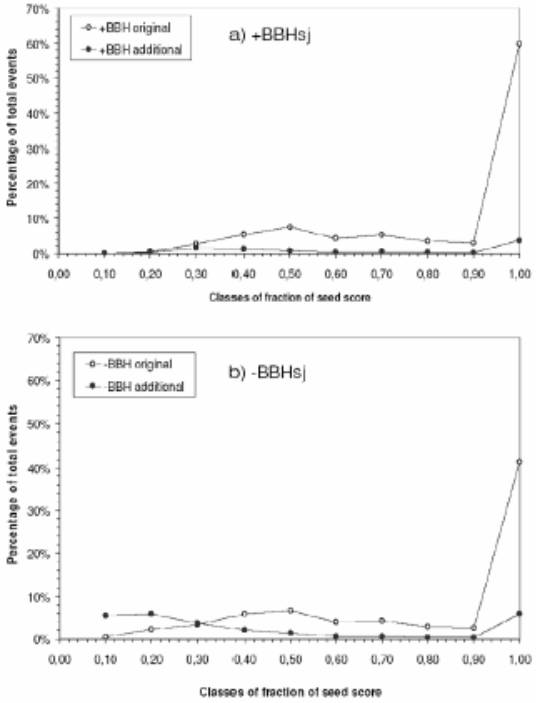


Figure 5
Relative score distribution in clusters with (a) or lacking (b) a BBHsj sequence. The alignment score to Seed divided by the score of Seed aligned to itself for each gathering event was saved and the percentage of events within each class of fraction of seed score interval (binned in .1 intervals) was determined. Legend: '+BBHsj original' and '+BBHsj additional', gathering events involving sequences respectively present or absent in manual clusters, in clusters with BBHsj reference sequence; '-BBHsj original' and '-BBHsj additional', gathering events involving sequences respectively present or absent in manual clusters, in clusters lacking a BBHsj reference sequence.

(maintained at 3%). Remarkably, these events decreased to 7% in clusters without a BBHsj. Analyzing the effect of a raw score cutoff of 400 bits (Figure 6c), it is observed that the proportion of gathering events of original sequences decreased slightly more than using the relative score cutoff (Figure 6b), reaching 37% in the clusters lacking a BBHsj, without an effect on the gathering events of additional sequences in absence of BBHsj. Thus, adoption of a cutoff based on relative score 0.3 appears to be more efficient than using a raw scrore 400 cutoff. Furthermore, the adoption of an extra *inparalog retrieval score limit* does not appear to be necessary in when a BBHsj is established. A direct comparison between the two approaches – raw and relative score – is presented by ROC curves shown in Figure 7. In the presence of a BBHsj (open symbols) the percentage of accumulated original sequences increases linearly as the cutoff is made less stringent. However, for the clustering initiated without a BBHsj (solid symbols) a cutoff less than either 400 (Figure 7a) or 0.3 (Figure 7b) drives the procedure to recruit relatively more additional sequences than original sequences. The use of the relative cutoff appears to be advantageous, so it has been adopted as default.

Post-filtering analysis of remaining additional sequences

Having applied the 0.3 relative score cutoff to clusters initiated both with and without a BBHsj, an analysis of the order in which the sequences were gathered reveals a curious phenomenon: the inclusion of original sequences after the inclusion of additional ones. Data in Figure 8 indicates that this scenario is frequent ("+Posteriors"), 52% of all events. These events comprise a total of 275 additional sequences, whereas 149 out of them (54.2%) have been gathered before an original one at least once. This might suggest that some gathered additional sequences were not recruited inappropriately. In fact, an analysis of the structural presence of trans-membrane domains (TM) shows that most of the additional sequences gathered with manual sequences *a posteriori* display at least two TM segments (133 out of 149, 89.3%). From the 126 additional sequences gathered at the last position by the algorithm, an additional 20 sequences also displayed at least two TM segments. Thus, possibly 133 plus 20 out of the 275 additional sequences (55.6%) might have had been suitably gathered.

Cluster disambiguation and comparison to manually curated dataset

After applying the 0.3 threshold to filter all inparalog sequences below or equal to this level in clusters initiated without a BBHsj, the resultant clusters were disambiguated using the rules described in "Methods." After disambiguation, a total of 1638 unique sequences remained in the dataset. This represents an increase of 20% over the initial number of Seeds used, corresponding to the inclusion of 275 additional sequences to the original 1363 ones. As suggested above, some of them (153) might have not been inappropriately gathered, what would represent an increase of just 9.3% in the initial universe of 1363 original sequences.

With disambiguation, the initial 1249 clusters formed by the 1363 seed sequences were reduced to 263 clusters and 38 singlets (which belong to 19 MCs), which approximates the expected number of 221 MCs. Some MCs (59 out of 221, 26.7%) were split into more than one RC, while the remaining 162 MCs (73.3%) were represented by only one RC each. Conversely, 248 out of the 263 RCs

BMC Bioinformatics 2008, 9:141 http://www.biomedcentral.com/1471-2105/9/141

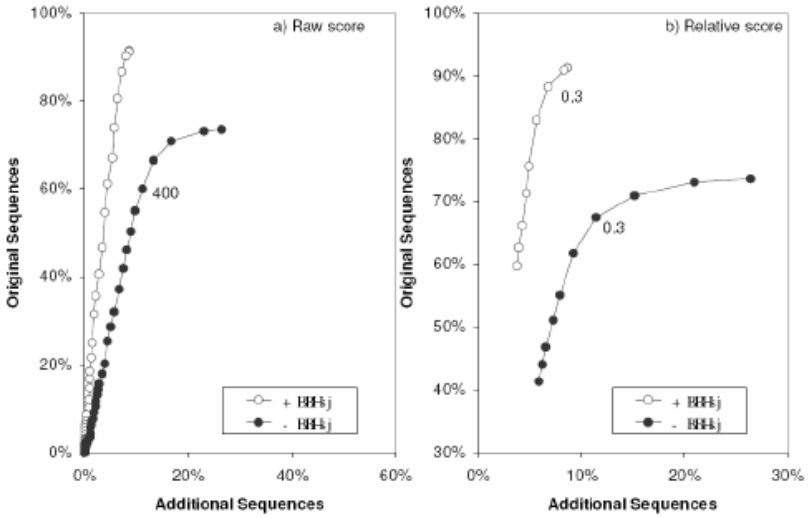


Figure 7
ROC curves comparing the raw and relative score approach. Each point represents cumulative recruitment over a given cutoff – highest percentages correspond to no cutoff; 400 raw score and 0.3 relative score are indicated.

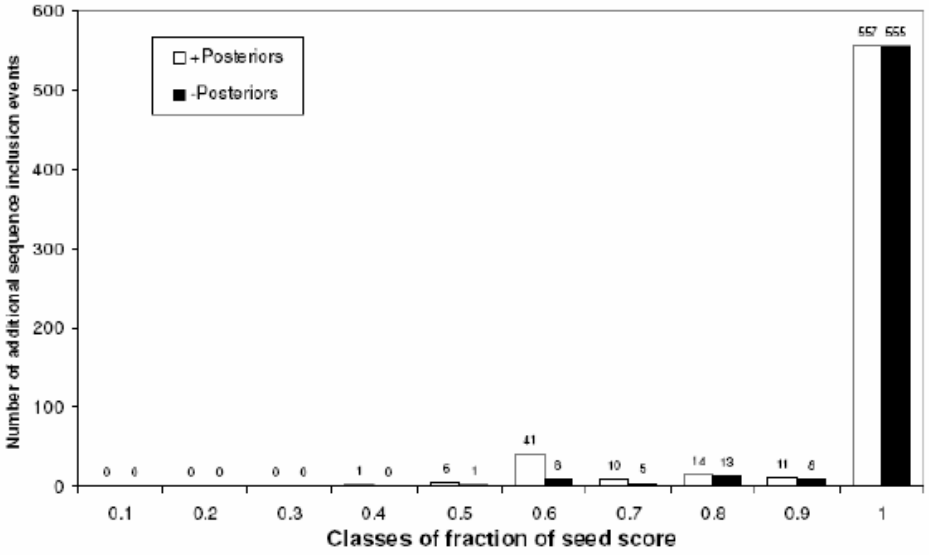


Figure 8
Additional sequences recruited before original ones. Additional sequences are shown per relative score range with (white bars) or without (black bars) original sequences being recruited subsequently.

(94.3%) were built with sequences from only a single MC, and within these, just 118 RCs (47.5% of 248) have gathered additional sequences. Indeed, in 100 of these 118 RCs, all additional sequences display more than two TM segments.

Seed Linkage has merged sequences from 25 MCs to produce 15 of the 263 disambiguated RCs. This is unlikely to be explained by mixing of unrelated proteins since all alignments are required to surpass a 10^{-10} E-value cutoff. From these 15 RCs, 12 result from the merging of all elements of the involved MCs. As an example, the neighbor joining (NJ) tree for the sequences present in RC1219 is shown in Figure 9a, wherein three MCs plus two additional sequences have been grouped. Moreover, the three remaining merged clusters represent exclusive cases also illustrated in Figure 9: RC254 has mixed all sequences from MC117 (size 9) together with one sequence from MC118 (size 2), the other sequence from MC118 has not built a cluster when used as Seed (represented on Figure 9b as a dashed branch). Note that, when used as Seed, sequence gi7293823 (MC118) was able to gather 5 out of 7 sequences from MC118 (labeled with asterisks in Figure 9b), what is the likely reason for all sequences being merged in RC254. In Figure 9c it is shown the merging of all sequences from MC248, divided in two main branches, with one sequence (gi7297676) from MC249 (size 3), closely placed between them. The two remaining sequences from MC249 have built a duplet with each other (represented as the dashed branch) and none of them included gi7297676 when they used as Seed. The last case of partial merging of clusters occurred for RC12 that merged all sequences from MC4 with four out of five sequences from MC2, while the remaining sequence from MC2, that was not included in RC12, has built the RC11 by recruiting two additional sequences. All these examples illustrate the consistence of clusters generated by Single Linkage that merge proteins judged as distantly related by manual curation.

Comparison to MultiParanoid approach

We compared our approach to another method, MultiParanoid, that has been shown to be very efficient in defining inparalog and ortholog clusters among multiple proteomes [4]. The result of the comparison of Seed Linkage versus MultiParanoid using the manually curated clusters as reference is shown in Table 1. While MultiParanoid produces a slightly reduced number of clusters (214) as compared to the manual curation (221), Seed Linkage produced 263 RC. This might indicate that Seed Linkage is more stringent for propagating information between clustered sequences since Seed Linkage yielded a larger number of clusters (59 clusters are a subset of the manual ones and 15 MC are split into distinct RC). Furthermore, Seed Linkage has rebuilt clusters with a larger number of seed sequences, since the number of singlets is rather small (only 38 against 179) as compared to MultiParanoid. Addition of sequences not included in curation was not restricted to Seed Linkage given that MultiParanoid presented a compatible number. As mentioned above, 153 of them provide evidence of being appropriately gathered as judged by the criteria used during curation – detection of over two TM domains. Agreement with manual curation seems weak at first glance, since only 13 RC are a perfect match as compared to 132 MultiParanoid clusters; however, in 100 RC all additional sequences display over two TM domains, thus 113 RC compare better with these 132 MultiParanoid clusters.

As mentioned above, 15 Seed Linkage RC merge MC, but this is not restricted to our procedure, since in nine events a MultiParanoid is split in two or more Manual [4].

Usage with partial proteomes

We selected from the 263 RC generated by Seed Linkage those clusters containing four or more sequences from organisms which we would artificially deplete to simulate an unfinished genome. For worm (Cel), fly (Dme) and human (Hsa), the respective set of 21, 15 and 88 clusters contained 115, 83 and 524 sequences. For each test, we randomly deleted 0 – 3 sequences from each of the 21 (Cel test), 15 (Dme test), or 88 (Hsa test) clusters, in turn, then dissolved the clusters. We then attempted to rebuild them using only the sequences of the remaining two organisms as Seeds (any remaining sequences from the test organism were present but not used as Seeds). The results are shown in Figure 10. With zero deletions from the test organism (X axis value at 100%), about 80% of the sequences were clustered by the Seeds from the other two organisms. This indicates that the clustering of the remaining 20% sequences from the test organism depends, at least in part, on paralogs from that test organism. Deletion of 1, 2, or 3 sequences per cluster yielded a linear recovery rate. For example, for fly, when three sequences per cluster were removed (retaining 34% of sequences in the recruitable set), 30% of all fly sequences (93% of the recruitable set) were clustered (triangles). The average clustering for each depletion test was 82% (that is, 82% of the clusterable sequences did indeed cluster) indicating that Seed Linkage can be applied to unfinished proteomes as well.

Usage with a single iteration

Besides being used up to convergence, Seed Linkage can be used in single-run mode (r parameter = 0). A comparison of performance is shown in Table 2, together with an execution that limited the gathering of only one Reciprocal Best Hit (RBH) per Seed per organism (one inparalog in the Seed organism and one ortholog in Other Species). Clustering with the default usage produced clusters of

BMC Bioinformatics 2008, **9**:141 http://www.biomedcentral.com/1471-2105/9/141

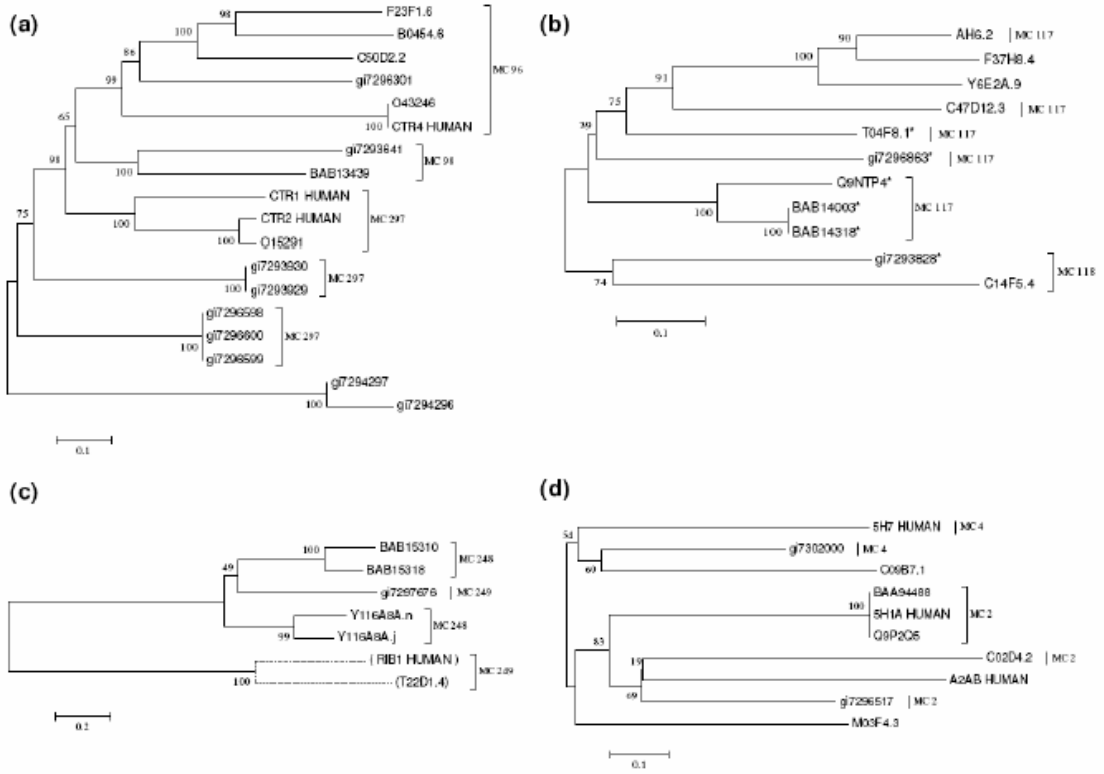


Figure 9

Neighbor joining trees of merged clusters. (a) Rebuilt Cluster RC1219 merges completely the three manually curated clusters MC96, MC98 and MC297 (represented in brackets) and additional sequences. **(b)** RC524 merges partially the clusters MC117 and MC118 and the remaining sequence from MC118 (C14F5.4, represented in the tree as the dashed branch), forms a singlet when used as Seed. When used as Seed, sequence gi72933828 gathers five sequences from MC117 (assigned by asterisks). **(c)** RC 1171 merges all sequences from MC248 with one sequence from MC 249. It is also represented the cluster RC1176, rebuilt by the remaining two sequences from MC249 (dashed branches). **(d)** RC12 gathers sequences from MC2 and MC4 have been merged. However, the remaining sequence from MC 2 (gi7296517) has formed the RC11 together with two additional sequences (not associated to the MC brackets).

mean size similar to the manual ones, while RBH grouped 1449 sequences in 402 clusters. The recall rate (sensitivity) that represents the amount of manual sequences that were clustered (not singlets) was comparatively high (97%) for both single iteration and convergence. Determination of Sensitivity with the raw results favors RBH execution; however, if one considers the recruited additional sequences that bear two or more trans-membrane domains (TM) as valid gathering events, then the Specificity* favors both single iteration and convergence. Thus, the advantage of using the convergence method seems to be increasing the linkage between clusters, resulting in a closer approximation of their number and mean size to the manual set.

Discussion

Seed Linkage was developed as an application to enrich the knowledge of similar proteins in species other than the Seed Organism. The present large size of proteome databases such as UniProt [12] suggests that a BLAST search involving all sequences against themselves would require an excessive amount of processing. Seed Linkage simplifies the search, focusing on the subjects of the Seed.

Similar to the approach used by Inparanoid, particular importance is given to the score between Seed and the best scoring Seed subject that establishes with Seed a Bidirectional Best Hit (BBH) relationship. Using a manually-curated dataset of trans-membrane proteins from worm,

BMC Bioinformatics 2008, 9:141 http://www.biomedcentral.com/1471-2105/9/141

Table 1: Comparison of Seed Linkage *versus* MultiParanoid using manually curated clusters as a reference.

	Seed Linkage	MultiParanoid
Number of RCs	263	214
Additional sequences	275 (153[a])	224
Singlets (non clustered manual sequences)	38	179
RC = MC (perfect match)	13	132
RC = MC + additional sequences	118 (100[a])	28
RC = MC - some manual sequences[b]	59	17
MC is split in two or more RC[b]	15	9
Two or more MC are merged in RC[b]	15	9

[a]additional sequences display two or more TM segments; [b]including or not additional sequences, MC: manually curated clusters, RC: rebuilt clusters.

fly and man, we found that 37% of all gathering events (Figure 6a) were initiated by finding a BBH subject (BBHsj) in a species different from the Seed Organism. When this happens, the *inparalog retrieval score limit* is made equal to the score obtained by the alignment of Seed and BBHsj, both within the Seed Organism and in the Candidate Species as well (Figure 1). Most of the grouped sequences when a BBHsj is attained corresponded to the ones listed in the manually curated database used as reference. However, it is possible that additional sequences are actually correct recruitments (Figure 8 and Table 1).

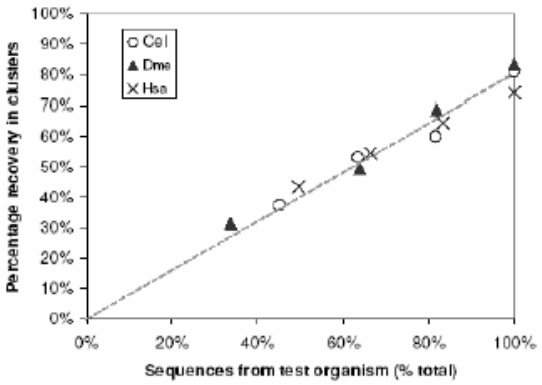


Figure 10
Simulation of a novel and unfinished genome. Seed Linkage rebuilt clusters containing four or more proteins from the indicated (test) organism were selected and 0 – 3 paralogs were artificially removed. Sequences of the other two complementary organisms were used as Seed and the percentage of recovery of the sequences from the indicated organism in clusters is shown. Complete recovery of paralogs would be represented by a diagonal line from zero up to 100% in both axes.

Seed linkage then continues the search for related proteins by two means, aiming to group inparalogs both in Seed Organism, using Seed as bait, and in Candidate Species, using the BBH relationship established by a recruited inparalog in the Seed Organism. We tested the performance of Seed Linkage and found better results when the iterative searches of sequences in Candidate Species that are initiated by a recruited inparalog is limited by the initial score between Seed and its BBHsj (data not shown), which makes the procedure more robust since one intended use of Seed Linkage is to identify circumstances that warrant propagation of information associated with Seed.

Inparalogs from the Seed Organism often are able to simultaneously recruit additional inparalogs in the Seed Organism, by establishing an additional BBH relationship with a Candidate Species. For example, from all gathering events in a cluster initiated by a Seed-BBHsj match, 11% do not involve the Seed itself.

Special attention was given to cases when a BBHsj was not found, to establish an *inparalog retrieval score limit*. For the studied curated database, such events occurred often (58%, Figure 5a, adding the last two bars). Two possible limits were investigated: a raw score and a relative score. The relative score was chosen to reflect a proportion of the score that the Seed shows when aligned to itself, thus incorporating the information of its size. Analysis of the distribution of the scores during the event of recruitment allowed us to find an empirical value for the *inparalog retrieval score limit*, so we could attain results that are similar to the processes initiated in the presence of a BBH relationship between Seed and BBHsj. The greatest performance was yielded by setting the value to 0.3 (Figure 5b), which reduced the recruitment of additional sequences to an acceptable rate. In our implementation, we provide support for users who wish to apply more stringent limits based on inspection of Figure 5. In the absence of a BBHsj, the *inparalog retrieval score limit* of 0.3 might benefit of an adjustment *a posteriori* to adapt to each

Table 2: Comparison of Seed Linkage usage under different iterations.

	Manual	RBH	One iteration	Convergence
Sequences	1363	1449	1605	1638
Clusters	221	402	302	263
Mean cluster size	6.2	3.6	5.3	6.2
Singlets	0	158	44	38
Originals in clusters	1363	1205	1319	1325
Sensitivity[a]	100%	88%	97%	97%
Additionals	0	87	242	275
Additionals < 2 TM	0	6	28	36
Specificity[b]	100%	93%	84%	83%
Specificity*[b,c]	100%	89%	96%	96%

[a]Sensitivity = Originals in clusters/Total of Manual Sequences; [b]Specificity = True Positives/Total of Sequences; [c]Specificity* was determined considering sequences with 2 or more TM domains as True Positives.

protein family. Further studies in this respect are envisaged.

Many databases group similar proteins and propagate the information amongst the members. Two examples of such databases are GOA (Gene Ontology Annotation, by EBI [13]) and KOG (the eukaryotic version of COG, by NCBI [6]). If we consider a database as a table composed of a column for each organism and a row for each protein, it might be noticed that GOA prioritizes the enlargement of the columns, which will contain very different number of gene entries per organism. Conversely, databases such as KOG strictly target the completion of the rows, listing all genes with similar function in the constituting organisms. Several approaches tend to focus more on these rows (proteins) than to enlarge the columns (organisms), although they might actually work on between those goals, such as Inparanoid/Multiparanoid, OrthoMCL [14], Kegg Orthology [15], and EGO [16], amongst others. Seed Linkage joins these efforts with a declared option for grouping cognate proteins from multiple organisms beginning with only one sequence, through connectivity saturation with that Seed sequence. As an example, using a protein from a dicot plant as Seed, it might be able to gather similar proteins in monocot plants, which in turn can act as a better reference sequence for similarity searches in the monocot Species. Moreover, Seed Linkage adds two relevant functionalities: (i) it does not require the Candidate Species to have completed genome and (ii) it saves computing time since it does not require alignment of all sequences to each other.

Certainly the recruitment yielded by Seed Linkage is amenable to additional approaches to validate the clustering such as literature support [17], mapping of conserved domains [18], alignment of secondary structure, etc. Indeed, a comparative analysis of Seed Linkage and Multiparanoid did not yield the same results (Table 1), although a similar number of clusters and performance were obtained. However, Seed Linkage recruits, with an acceptable level of confidence, a significant number of candidates, maximizing the search on all available proteomes while minimizing computing time. The software is made available for the research community, and a web service dedicated to Seed Linkage is currently under construction. Seed Linkage is also currently being used to construct a Database for Protein Defense Mechanisms in plants.

Conclusion

The Seed Linkage software was produced with the aim of clustering cognate proteins from multiple organisms beginning with a single sequence through connectivity saturation with that Seed sequence. The method results were comparable to conventional clustering approaches. Generating clusters based only on a protein of interest is less time consuming than generating clusters for whole proteomes, and can be applied to establish members from unfinished proteomes as well.

Availability and requirements

- **Project name:** Seed Linkage clustering of related protein sequences;
- **Project home page:** http://biodados.icb.ufmg.br/seedlinkage;
- **Operating system:** Linux;
- **Programming language:** PHP;
- **Other requirements:** PHP 5 or higher, MySQL 5 or higher, NCBI BLAST package;
- **Any restrictions to use by non-academics:** License needed.
- **NCBI Taxonomy database:** http://ftp.ncbi.nih.gov/pub/taxonomy

Authors' contributions

AB-S created, implemented and tested the proposed algorithm; created and conducted the pilot tests and wrote the paper. VPS helped in the algorithm discussion, development and implementation. RS and JMO advised in the algorithm implementation, discussed the pilot tests and coordinated the method development. JMO created the paper's main idea and supervised the writing of the paper. All authors read and approved the final manuscript.

Additional material

Additional file 1

Seed Linkage Package. Compressed core of scripts written in PHP necessary to run the Seed Linkage procedure. Also contains a documentation file about each sub item.

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-141-S1.tar]

Acknowledgements

We would like to sincerely thank Darren A. Natale (PIR) for critically reviewing the manuscript and to Mauricio de Alvarenga Mudado (FUNED) for indispensable comments and suggestions, besides his unquestionable technical support. We thank especially to VPS and RS for the valuable training given to AB-S during his special external doctorate training in the European Molecular Biology Laboratory (EMBL Heidelberg) – Germany. This work was developed as part of AB-S PhD thesis, which has been sponsored by the Brazilian Ministry of Education (CAPES) and Foundation for Research Support of Minas Gerais State (FAPEMIG).

References

1. Sonnhammer EL, Koonin EV: **Orthology, paralogy and proposed classification for paralog subtypes.** *Trends Genet* 2002, **18(12):**619-620.
2. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Res* 2001, **29(1):**22-28.
3. Remm M, Storm CE, Sonnhammer EL: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.** *J Mol Biol* 2001, **314(5):**1041-1052.
4. Alexeyenko A, Tamas I, Liu G, Sonnhammer EL: **Automatic clustering of orthologs and inparalogs shared by multiple proteomes.** *Bioinformatics* 2006, **22(14):**e9-15.
5. Chen F, Mackey AJ, Stoeckert CJ Jr., Roos DS: **OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups.** *Nucleic Acids Res* 2006, **34(Database issue):**D363-8.
6. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4:**41.
7. **MySQL** [http://www.mysql.com]
8. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25(17):**3389-3402.
9. **PHP** [http://www.php.net]
10. Remm M, Sonnhammer E: **Classification of transmembrane protein families in the Caenorhabditis elegans genome and identification of human orthologs.** *Genome Research* 2000, **10(11):**1679-1689.
11. O'Brien KP, Remm M, Sonnhammer ELL: **Inparanoid: a comprehensive database of eukaryotic orthologs.** *Nucleic Acids Research* 2005, **33:**D476-D480.
12. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang HZ, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LSL: **UniProt: the Universal Protein knowledgebase.** *Nucleic Acids Research* 2004, **32:**D115-D119.
13. Camon E, Barrell D, Lee V, Dimmer E, Apweiler R: **The Gene Ontology Annotation (GOA) Database--an integrated resource of GO annotations to the UniProt Knowledgebase.** *In Silico Biol* 2004, **4(1):**5-6.
14. Li L, Stoeckert CJ Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13(9):**2178-2189.
15. Mao X, Cai T, Olyarchuk JG, Wei L: **Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary.** *Bioinformatics* 2005, **21(19):**3787-3793.
16. Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai J, Parvizi B, Cheung F, Antonescu V, White J, Holt I, Liang F, Quackenbush J: **Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA).** *Genome Res* 2002, **12(3):**493-502.
17. Miotto O, Tan TW, Brusic V: **Supporting the curation of biological databases with reusable text mining.** *Genome Inform* 2005, **16(2):**32-44.
18. Sankar N, Machado J, Abdulla P, Hilliker AJ, Coe IR: **Comparative genomic analysis of equilibrative nucleoside transporters suggests conserved protein structure despite limited sequence identity.** *Nucleic Acids Res* 2002, **30(20):**4339-4350.

5.3 DESENVOLVIMENTO DA BIBLIOTECA SRS.PHP, UM RECURSO BASEADO EM SIMPLE OBJECT ACCESS PROTOCOL (SOAP) PARA AQUISIÇÃO DE DADOS ORIUNDOS DE BASES DE DADOS INTEGRADAS.

Nesta sessão descrevemos o desenvolvimento da biblioteca SRS.php, um conjunto de funções específicas para comunicação via SOAP com o SRS (Sequence Retrieval System) instalado no EMBL (European Molecular Biology Laboratory) em Heidelberg – Alemanha.

A biblioteca SRS.php foi criada com o intuito de se obter dinamicamente os diversos dados depositados nas bases biológicas que compõem o SRS do EMBL. Essa aquisição de dados é feita usando a tecnologia de Web Services que, conforme anteriormente descrita, permite a troca de informações em um ambiente descentralizado, independentemente do sistema usado pelos computadores presentes na network.

Exploramos a classe PHP chamada nuSOAP para desenvolver a biblioteca SRS.php que possui 4 funções específicas para explorar a maioria dos recursos do SRS, entre eles a facilidade de conectar registros entre qualquer base que compõe o sistema.

Para se comunicar com o SRS, a biblioteca acessa o arquivo de descrição dos web services (WSDL) deste sistema disponível em http://srs.embl.de/axis/services/SrsWrapper?wsdl. Inicialmente foram implementados 4 métodos, que permitem acessos a mais de 90 bases de dados diferentes. Os métodos permitem:

1. Listar o número de registros em qualquer base de dados que apresentam o termo usado como “query” no campo indicado;
2. Retornar o conteúdo de um registro específico das bases de acordo com as informações fornecidas pelo “loader” invocado.
3. Acessar campos específicos dos registros de qualquer base de dados.
1. Conectar registros entre quaisquer pares de bases de dados potencialmente conectáveis

A finalidade prática desses métodos na nossa linha de pesquisa é coletar informações específicas oriundas do SRS-EMBL sobre as proteínas identificadas na literatura como relacionadas aos mecanismos de defesa em plantas pelo sistema de “text mining” e de agrupamento de seqüências previamente apresentados.

Os métodos são implementados como funções em PHP agrupados na biblioteca SRS.php; tal biblioteca pode ser adquirida gratuitamente no site http://www.biodados.icb.ufmg.br/srs_php. Este trabalho foi publicado na revista *Genetics and Molecular Research* com o título: “**Development of SRS.php, a Simple Object Access Protocol-based library for data acquisition from integrated biological databases**” (Barbosa-Silva et al., 2007).

Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Development of SRS.php, a Simple Object Access Protocol-based library for data acquisition from integrated biological databases

A. Barbosa-Silva[1,2], E. Pafilis[2], J.M. Ortega[1] and R. Schneider[2]

[1]Departamento de Bioquímica e Imunologia,
Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil
[2]Structural and Computational Biology Unit,
European Molecular Biology Laboratory, Heidelberg, Germany
Corresponding author: J.M. Ortega
E-mail: miguel@icb.ufmg.br

Genet. Mol. Res. 6 (4): 1142-1150 (2007)
Received August 30, 2007
Accepted November 20, 2007
Published December 11, 2007

ABSTRACT. Data integration has become an important task for biological database providers. The current model for data exchange among different sources simplifies the manner that distinct information is accessed by users. The evolution of data representation from HTML to XML enabled programs, instead of humans, to interact with biological databases. We present here SRS.php, a PHP library that can interact with the data integration Sequence Retrieval System (SRS). The library has been written using SOAP definitions, and permits the programmatic communication through webservices with the SRS. The interactions are possible by invoking the methods described in WSDL by exchanging XML messages. The current functions available in the library have been built to access specific data stored in any of the 90 different databases (such as UNIPROT, KEGG and GO) using the same query syntax format. The inclusion of the described functions in the source of scripts written in PHP enables them as webservice clients to the SRS server. The functions permit one to query the whole content of any SRS database, to list specific records in these databases, to get specific fields from the records, and to link any record among any pair of linked databases. The case study presented exemplifies the library usage to retrieve information regarding registries of a Plant Defense Mechanisms database. The Plant Defense Mechanisms database is currently being developed, and the proposal of SRS.php library usage is

release their stored data for users (e-Utilities; Miyazaki et al., 2004; NCICB). One of these resources is the EMBL-SRS Universe (http://srs.embl.de) (Zdobnov et al., 2002). It is possible to query and navigate through several databases available in the server by using the Sequence Retrieval System (SRS) of this machine. Additionally, the users can exploit the link between the databases to precisely extract the information from multiple sources. The aim of the present study was to describe SRS.php, a computational resource designed to integrate dynamic scripts written in Hypertext Preprocessor (PHP) to the SRS available by the EMBL-Heidelberg server. This is done by using SOAP-based webservice requests that communicate directly with the system by changing messages based in XML. This resource is released as an open source library: a set of functions that can be embedded in the source code of scripts written to perform specific tasks. The utility of the library is illustrated by the usage of SRS.php in the acquisition of specific information from multiple sources for sequences deposited in a bioDB developed by us, the Plant Defense Mechanisms Database.

MATERIAL AND METHODS

Features of target Sequence Retrieval System

The aim of the library proposed in this study was to interact with the SRS made available by the EMBL-Heidelberg; the server is composed of approximately 90 different biological databases (http://srs.embl.de/srs/databanklist.do). The databases have a description webpage where precise information about the features of each of them can be accessed. The information contains the database description, the field list for each record, the database loaders, and the target databases that the selected database links in the same system.

Record features

The records deposited in a database under an SRS have a peculiar syntax that deserves comment before the explanation of how it is accessed through SOAP. Each item is deposited as a table in which the rows are the attributes of such record. The fields are abbreviated by a short name that facilitates the database querying. An example of a typical record of the database UNIPROT (Apweiler et al., 2004) is illustrated briefly in Table 1.

Table 1. Brief representation of the UNIPROT database installed in a Sequence Retrieval System.

Database name UNIPROT		
Field name	Short name	Content
Identifier	id	Q9XET3_SOLLC
Prim. Accession	pac	Q9XET3
Description	des	Disease resistance protein I2
Organism	org	*Solanum lycopersicum* (Tomato) (*Lycopersicon esculentum*)
NCBI Taxonomy ID	txi	4081

Query syntax

One who wants to get the information of a record deposited in UNIPROT database from an SRS running on a Linux platform, for example, should do it from the command

line using the program getz (see SRS documentation available at http://srs.embl.de/srs/doc/srsbooks/srs_user_guide/21_1.html). This program receives several parameters, but in order to describe the usage of our library, we will focus on just some of them.

Listing records from some database

To display all the records deposited for the organism exemplified in Table 1, one should submit the getz query in Formula 1, this would retrieve all the records belonging to the organism identified by the NCBI Taxonomy id (txi) 4081 in the UNIPROT database.

[srs_server]$ getz "[UNIPROT-txi:4801]" (Formula 1)

Displaying records from some database

In case the user wants to display a specific record from some database, it is necessary to use some optional flags in the getz program. The following command uses the flag '-e' in order to invoke the loader CompleteEntry which accesses and displays the whole record identified by the accession number Q9XET3.

[srs_server]$ getz "[UNIPROT-acc:Q9XET3]" –e (Formula 2)

Access specific fields

Another command flag of the getz program allows users to retrieve from the command line the specific fields of some record. In Formula 3, the operator '-vf' is used for this purpose. In the example, the Medline database is queried and the field 'authors' (aut) is chosen to be displayed for the article identified by the Medline ID (id) 8208723.

[srs_server]$ getz "[medline-id:8208723]" –vf aut (Formula 3)

Exploring the link universe

The SRS feature used to link the records among the distinct databases can be explored using the getz program as well. For this, the operator ">" is used in Formula 4; with the illustrated command, the article from the Medline database identified by the ID 15371431 is linked to its corresponding record in the UNIPROT database.

[srs_server]$ getz "[medline-id: 15371431]>UNIPROT" (Formula 4)

Implementation of SRS.php library

The SRS.php library was written in PHP 5 and has been tested in a system running the same or superior version. However, the command line interface has been implemented in PHP since version 4.3.0, for this reason it is expected that the library should also work with versions as old as this. In some machines, it is possible that some warnings or error messages could be displayed on the screen; to avoid this, users must set the *error_reporting* variable in the php.ini configuration file

for "E_ALL & ~E_NOTICE | E_STRICT" for instance. Another feature that can affect the library performance is the resource limits of the PHP installed in the machine; in order to avoid problems in regard to this issue, the variables related to these limits should be set to compatible values.

The library is expected to enable scripts that include it on the code to act as SOAP clients to the EMBL-SRS server. That is why it uses the SOAP features available through the class nuSOAP, which can be freely downloaded from the developers' webpage (nuSOAP).

RESULTS AND DISCUSSION

Description of the library

The initial core of request methods present in the SRS.php library consists of 4 functions that allow the data integration between the SRS database engine and PHP dynamic scripts. These functions interact with SRS based on information found on its WSDL description file. This file is accessed using the *soapclient* nuSOAP function, which receives the WSDL file location as parameter (http://srs8.embl.de:8989/axis/services/SrsWrapper?wsdl); this is the common feature present in all functions which are described below.

Acquisition of number of results for queries

To access the number of results for one request for any of the SRS databases, we created the function *performIcarusQueryAndGetNumberOfResults* (Figure 1). This function receives as parameter one string corresponding to the getz query in Formula 1, and returns a one-dimensional array with the number of records that matched the request in SRS (Figure 5A, 1). As an example, one can obtain the present number of entries in Uniprot for a given organism.

```
function performIcarusQueryAndGetNumberOfResults($query){

  // Use the server URL
  global $server;

  // Create instance to server
  $client = new soapclient($server);

  // Fill the parameters for the service
  $param=array("query"=>$query);

  // Submit the query to the web-service
  $results = $client -> call("performIcarusQueryAndGetNumberOfResults",$param);

  return($results);
}

#EXAMPLE
#$num=performIcarusQueryAndGetNumberOfResults("[UNIPROT-txi:3702]");
#echo $num[0];
```

Figure 1. Function performIcarusQueryAndGetNumberOfResults.

Accessing records using loaders

The second function, *performQueryAndGetLoaderDefinedFields*, was created to allow the access to the full records deposited in the SRS databases, either directly or using the Link Universe

facility as well. As parameter, one must supply the getz query (like in Formula 2) and the name of the loader, a tool to extract specific information from each SRS entry, to be used (Figure 2). The function returns a one-dimensional array with the content of the loader result (i.e., defined attributes for one SRS entry) as output (Figure 5A, 2). For instance, it is possible from a UNIPROT accession number to retrieve the complete entry of a record using the loader CompleteEntry.

```
function performQueryAndGetLoaderDefinedFields($query,$loader){

  // Use the server URL
  global $server;

  // Create instance to server
  $client=new soapclient($server);

  // Fill the parameters for the service
  $param=array("query"=>$query,
               "loader"=>$loader,
               );

  // Submit the query to the web-service
  $results = $client -> call("performQueryAndGetLoaderDefinedFields",$param);

  // Result from the function
  return($results);
}
#EXAMPLE
#$result=performQueryAndGetLoaderDefinedFields("[UNIPROT-acc:Q9XET3]","CompleteEntry");
#echo $result[0];
```

Figure 2. Function performQueryAndGetLoaderDefinedFields.

Accessing specific fields in the records

The content of specific attributes for one SRS record can be accessed using the third function, *performQueryAndGetSpecificFields* (Figure 3). This function needs four parameters: i) the target database to be queried, ii) the target field, iii) the query term, and iv) the fields to be displayed from the accessed record. The function returns the specific fields that matched the term (Figure 5A, 3). The operation performed by this function is equivalent to the getz query of Formula 3.

```
function performQueryAndGetSpecificFields($queryTargetDb,$queryTargetField,$queryTerm,$fields){

  // Use the server URL
   global $server;

  // Create instance to server
  $client=new soapclient($server);

  // Fill the parameters for the service
  $param=array("queryTargetDb"=>$queryTargetDb,
               "queryTargetField"=>$queryTargetField,
               "queryTerm"=>$queryTerm,
               "fields"=>$fields
               );

  // Submit the query to the web-service
  $results = $client -> call("performQueryAndGetSpecificFields",$param);

  // Removes the inicial field of the getz query
  $results[0]=preg_replace("/\s+/","#",$results[0]);
  $results=explode("#",$results[0]);
  array_shift($results);
  $field=implode(" ",$results);

  // Result from the function
  return($field);
}

#EXAMPLE
#$field=performQueryAndGetSpecificFields("medline","id","8208723","aut");
#echo $field;
```

Figure 3. Function performQueryAndGetSpecificFields.

Exploring the Sequence Retrieval System link universe

This special characteristic of the SRS was included in the library by the fourth function, *performLinkingQuery* (Figure 4). This function, like the query in Formula 4, receives as parameters the database in which the query term is deposited, the parameter (field) to which the query refers, the query term and the database to which the query should be linked. The function above returns the record IDs in the linked database related to the term provided by the function (Figure 5A, 4). For example, all the INTERPRO domain signatures of a UNIPROT sequence could be retrieved by using this function, through the SRS Link Universe resource.

```
function performLinkingQuery($queryTargetDb,$queryTargetField,$queryTerm,$linkingTargetDb){

 // Use the server URL
 global $server;

 // Create instance to server
 $client=new soapclient($server);

 // Fill the parameters for the service
 $param=array("queryTargetDb"=>$queryTargetDb,
              "queryTargetField"=>$queryTargetField,
              "queryTerm"=>$queryTerm,
              "linkingTargetDb"=>$linkingTargetDb
              );

 // Submit the query to the web-service
 $results = $client -> call("performLinkingQuery",$param);

 // Result from the function
 return($results);
}

#EXAMPLE
#$link=performLinkingQuery("medline","id","15371431","UNIPROT");
#for($i=0;$i<sizeof($link);$i++) print "$link[$i].<BR>";
```

Figure 4. Function performLinkingQuery.

Figure 5A summarizes the functions present in the SRS.php library. A case study using the functions mentioned above is reported in the next session.

Case study: data warehousing for Plant Defense Mechanisms database

One of us (A.B.-S.) is developing a database which collects proteins that are involved in Plant Defense Mechanisms (PDM). Currently, the database contains sequences from UNIPROT database (seed sequences), and collected similarity-related sequences (putative orthologs) to the seed sequences through a Seed Linkage clustering approach (Barbosa-Silva A, Satagopam VP, Schneider R and Ortega JM, unpublished results) based on bidirectional best-hit strategy.

To improve the annotation of the data deposited in the PDM database, we have integrated the information from diverse set of databases.

Using the SRS.php library, we first queried for each plant represented in the PDM database, about its total number of entries deposited in the following databases: UNIPROT, RefSeq, UniRef (100, 90 and 50), and PIR (Figure 5B).

In a second step, we used the SRS.php library to access the annotations in the UNIPROT, GO, PFamA, INTERPRO, and Prosite databases, which can be related to the PDM sequences

using their UNIPROT ID as query (Figure 5C). Interested users can download the script used for these tasks, retrieve_srs.php, from the supplementary material (http://www.biodados.icb.ufmg.br/seed_linkage).

The usage of SRS.php functions in the annotation of PDM database sequences was a very straightforward approach, since the acquisition of further information on demand for the PDM sequences was very fast, compared to manual annotation, considering the existence of only the UNIPROT ID as starting point. Hitherto, we have added several forms of useful information that increased undoubtedly the content and quality of the PDM entries.

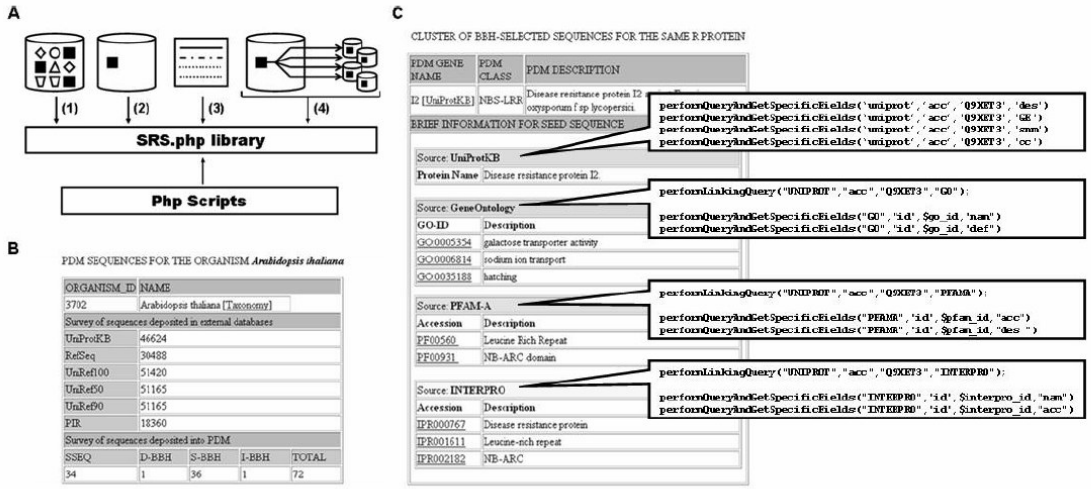


Figure 5. A. Using the library SRS.php, the user can (1) count the number of matches (small black squares) in one Sequence Retrieval System (SRS) database (DB1), (2) retrieve specific records (black square), and (3) retrieve specific attributes of each deposited record (different lines). Additionally (4), the user can link the record (small black square) to the universe of SRS databases. **B.** Number of records deposited for *Arabidopsis thaliana* in different databases. **C.** Retrieved annotations for gene I2 (UNIPROT accession: Q9XET3) of *Lycopersicon esculentum* in Plant Defense Mechanisms (PDM) from different databases (UNIPROT, GO, PFam-A, INTERPRO); the screenshots represent examples of the SRS.php information retrieval for PDM records.

CONCLUSIONS

Integration of biological databases is a current issue addressed by several data providers in life sciences (Stein, 2002). Currently, webservices is the default technology chosen for data integration not only in biology, but also for other internet resources when the information is not centered on just one source. A unified language that permits biological information to be shared in this scenario is XML (Achard et al., 2001). Hitherto, an increasing number of open source applications have been developed to explore the facilities of webservices. We provide SRS.php as an open source library that gives support for the development of new dynamic scripts that can act as true data retrievers through a programmatic access to several bioDBs under the SRS hosted by EMBL. In a study case, information integration provided automated maintenance of Protein Defense Mechanism Database with information contained in several bioDBs such as UNIPROT, INTERPRO, PFam, etc. In the future, following the development of new services provided by SOAP interface of SRS, plus the feedback received from users,

we aim to improve the content and functionalities of SRS.php library.

AKNOWLEDGMENTS

The authors are thankful to Theodorus Soldatos, Georgios Pavlopoulos and Venkata Satagopam for providing training at EMBL to A. Barbosa-Silva and CAPES for providing the PDEE fellowship to develop this research in association with EMBL; to EMBL for providing scientific resources that enabled the execution of this study, and to FAPEMIG for supporting our research.

REFERENCES

Achard F, Vaysseix G and Barillot E (2001). XML, bioinformatics and data integration. *Bioinformatics* 17: 115-125.

Apweiler R, Bairoch A, Wu CH, Barker WC, et al. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* 32: D115-D119.

Curcin V, Ghanem M and Guo Y (2005). Web services in the life sciences. *Drug Discov. Today* 10: 865-871.

e-Utilities - Entrez Utilities Web Service. http://www.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html. Accessed August 23, 2007.

HTML. HyperText Markup Language. http://www.w3.org/TR/html401/. Accessed August 23, 2007.

Miyazaki S, Sugawara H, Ikeo K and Gojobori T (2004). DDBJ in the stream of various biological data. *Nucleic Acids Res.* 32: D31-34.

NCICB. National Cancer Institute Web Services. http://ncicb.nci.nih.gov/infrastructure/cacore_overview. Accessed August 23, 2007.

nuSOAP. http://dietrich.ganx4.com/nusoap/. Accessed August 23, 2007.

PHP. Hypertext Preprocessor. http://www.php.net. Accessed August 23, 2007.

Romano P, Marra D and Milanesi L (2005). Web services and workflow management for biological resources. *BMC Bioinformatics* 6 (Suppl 4): S24.

SOAP. Simple Object Access Protocol Specifications. http://www.w3.org/TR/soap/. Accessed August 23, 2007.

Stein L (2002). Creating a bioinformatics nation. *Nature* 417: 119-120.

Stein LD (2003). Integrating biological databases. *Nat. Rev. Genet.* 4: 337-345.

UDDI. Universal Description, Discovery and Integration. http://www.uddi.org. Accessed August 23, 2007.

Wang L, Riethoven JJ and Robinson A (2002). XEMBL: distributing EMBL data in XML format. *Bioinformatics* 18: 1147-1148.

WSDL. Web Services Description Language. http://www.w3.org/TR/wsdl. Accessed August 23, 2007.

XML. Extensible Markup Language. http://www.w3.org/XML/. Accessed August 23, 2007.

Zdobnov EM, Lopez R, Apweiler R and Etzold T (2002). The EBI SRS server - new features. *Bioinformatics* 18: 1149-1150.

5.4 ESTRATÉGIAS PARA O DESENVOLVIMENTO DA PLANT DEFENSE MECHANISMS DATABASE

Para concluir o conteúdo apresentado na tese, descrevemos a utilização das estratégias implementadas nas sessões anteriores para o desenvolvimento da Plant Defese Mechanisms Database.

O programa LAITOR foi aplicado a um conjunto de 7.306 abstracts recuperados na base de dados PubMed utilizando a query: "(resistan* OR toleran*) AND ((high light OR high-light) disease OR cold OR pathogen* OR drought OR salinity OR "oxidative stress*" OR (high temperature OR high-temperature)) AND plant", os termos de proteínas que apresentaram co-ocorrência proteínas nesta análise foram associados a identificadores do UniProtKB e tiveram suas seqüências recuperadas.

1.390 das seqüências recuperadas acima foram utilizadas como seed num processo de agrupamento utilizando para isso o programa Seed Linkage, previamente descrito. Após o agrupamento, um total de 15.669 foram agrupadas em 611 clusters desambígüos. Cada uma das seqüências presentes nestes clusters tiveram seus arquivos XML diretamente baixados do UniProtKB usando a tecnologia de SOAP-based web services implementados na biblioteca SRS.php também descrita nas sessões anteriores.

Um website contendo as informações da PDM foi construído no domínio http://www.biodados.icb.ufmg.br/pdm. Nesse site, é possível visualizar as informações dos clusters PDM, bem como de cada umas das seqüências componentes, além de uma árvore filogenética designada para cada agrupamento.

Finalmente, um servidor SOAP foi desenvolvido para a base de dados PDM, inicialmente, o método query_pdm permite que, a partir de um identificador UniProtKB de uma seqüência depositada na base, as mesmas informações acessadas na interface web sejam distribuídas via SOAP, tornando, desta maneira, a base de dados PDM um Web Service com acesso programático. A descrição do Web Service está documentada em seu WSDL disponível em http://www.biodados.icb.ufmg.br/pdm/soap/.

Escrevemos um manuscrito do tipo Application Notes da categoria Database and ontologies que foi submetido para publicação na revista Bioinformatics com o título: "Protein Defense Mechanisms Database", o qual é apresentado em seguida.

BIOINFORMATICS APPLICATIONS NOTE

Databases and ontologies

Plant Defense Mechanisms Database

Adriano Barbosa-Silva[1] and J. Miguel Ortega[1,*]

[1]Laboratório de Biodados, Universidade Federal de Minas Gerais, Av. Antonio Carlos 6627, Belo Horizonte, MG, Brazil.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: Plant Defense Mechanisms (PDM) Database stores clusters of protein sequences involved in defense responses in plants. This new resource allows human as well as programmatic access to its contents using both, HTML or SOAP protocol, respectively. We present here PDM building strategies and its features.

Availability: http://biodados.icb.ufmg.br/PDM

Contact: miguel@icb.ufmg.br

Supplementary Material: http://biodados.icb.ufmg.br/PDM

1 INTRODUCTION

Resistance proteins are important economical traits found in plant and explored in an economic basis (Jones, 2001). Hitherto restrict databases directed to plant defense mechanisms (PDM) provide a SOAP-based web service interface for acquisition of information for defense protein sequences (Schoof et al., 2004). Moreover, it is noteworthy that several research groups would benefit of either local or web served databases on specific issues. The building of PDM is based on a pipeline of software developed by our group that consorts to an automated building of the resource. Here we describe the building and distribution of Plant Defense Mechanism Database and the process of its creation using underlined software.

2 TEXT MINING

In order to identify new proteins mentioned in the literature as related to defense mechanisms in plants, we have used the program LAITOR (Barbosa-Silva et al., in preparation) to find associations over the retrieved abstracts. Initially a Boolean search was performed in the PubMed website (http://www.ncbi.nlm.nih.gov/pubmed/) where the matching abstracts were retrieved in the XML format. The query used was: "". After, the retrieved abstracts were loaded in the NLPROT program (Mika and Rost, 2004) in order to tag the protein names for further analysis with the program LAITOR. After analysis, those proteins terms identified in the abstracts together with another protein or stimuli term were associated to UniProtKB unique identifiers and loaded in the clustering program Seed Linkage (Barbosa-Silva et al., 2008) as seeds (see below). The co-occurring pairs along with their citations in the abstracts were stored and are displayed as web-pages in the Text Mining session of the PDM website. Clusters of sequences are available for download or can be searched within PDM website by means of BLAST searches.

3 SEQUENCE CLUSTERING

All protein names identified in the above step were linked to a unique identifier of the UniProtKB when possible. This enabled us to use these sequences as seeds of the Seed Linkage program (Barbosa-Silva et al., 2008), using all proteins from the Viridiplantae division as database. Since different proteins included as seeds are expected to create redundant clusters, we disambiguated each individual cluster using Seed Linkage internal rules for disambiguation. Additionally, for the created clusters, we generated a Neighbor Joining-based phylogenetic tree by using the programs available in the package Phylip (http://evolution.genetics.washington.edu/phylip.html) which is also available for visualization in the PDM website

4 DATA ACQUISITION

It has been performed a round of SOAP-based data acquisition for each sequence present in the clusters mentioned above by using the SRS.php library (Barbosa-Silva et al., 2007) connected to the Web Services available by EBI described by its WSDL file at http://www.ebi.ac.uk/Tools/webservices/wsdl/WSDbfetch.wsdl.
The fetchBatch method has been used to retrieve the XML records for each entry present in the PDM clusters.

5 DATABASE CONTENT

Currently PDM database is composed by a total of 780 protein terms, identified in the text mining step from the analysis of a total of 7,306 abstracts retrieved by the above mentioned query. From these terms it was possible to retrieve 1,390 UniprotKB unique identifiers which were used to clusters a set of 15,669 sequences grouped into 611 disambiguated clusters.
Each sequence from in the PDM database is presented in a website (Figure 1) as follows:
(1) Selected sequence: displays the sequence together with the associated PDM cluster.
(2) Sequence details: displays the information retrieved from the UniprotKB database using the SOAP protocol. Here it is informed the Protein and Gene names for such sequence; the organism source of the sequence; the comments about the record, such as: functions, subunits, subcellular location, domains and similarities, directly retrieved from its acquired XML record. Finally, for each

*To whom correspondence should be addressed.

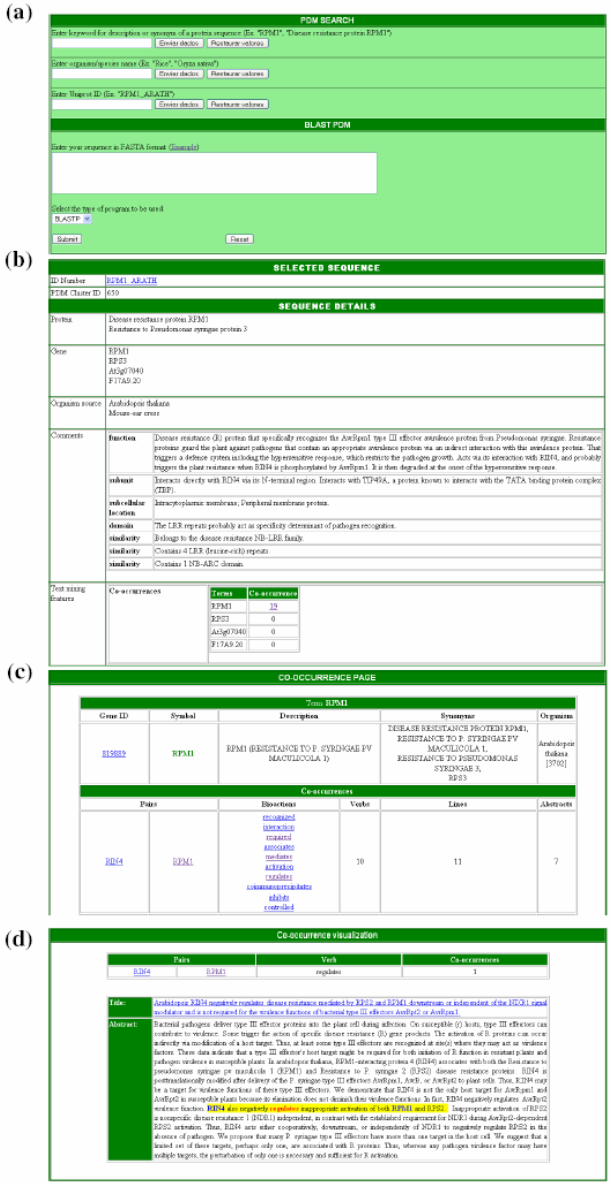


Fig. 1. PDM Database. (A) Database search: users can either perform a term search by using protein terms, organism names or UniProtKB unique identifiers; as well as a BLAST alignment to PDM sequences. (B) PDM database record RPM1_ARATH. In this page are found the Sequence Details, the Text-mining Features, the other cluster members and the phylogenetic tree of the sequence's cluster (not shown). (C) Co-occurrence session. For each term/synonym of the referred sequence, it is shown the co-occurring pair RPM1-RIN4 together with biointeraction terms; by clicking on each of these terms, (D) it is displayed a web-page with the abstract highlighted at the phrase where the co-occurrence has been found (not shown).

sequence, it is shown the text mining feature of it by showing the number of co-occurrences of each name attributed to the selected sequence

(3) PDM Cluster Details: displays all the sequences from the selected cluster together with a phylogenetic tree for the cluster members.

(4) The database also possesses a search webpage where users can query the database in several ways, including BLAST search using PDM as database.

6 DATA DISTRIBUTION

PDM Database is distributed through SOAP-based web services. We have created a method called query_pdm in which the same information exhibited in the HTML format at the PDM website is converted to a simple XML tagged document. This allows that client softwares implemented using the web services definitions can interact with the PDM SOAP interface described at http://biodados.icb.ufmg.br/pdm/pdm_soap_server.php and request the desired information. We provide a SOAP client written in PHP using the nuSOAP class (http://dietrich.ganx4.com/nusoap/) as Supplementary Material. Novel methods will be released at the PDM SOAP Server website whilst they become available.

7 CONCLUSION

A Plant Defense Mechanism Database was built by using a literature search followed by extraction of gene names from the abstracts. Thus, the various aspects of the phenomena of interest are connected to specific literature. With Seed Linkage software, the involved proteins were able to recruit a significative number of other sequences that enrich the database, improving its use as a source of annotation. Automated updating of information on these sequences grant the correct propagation of information, which is conducted by machine automated services. PDM distribution itself is offered using webservices. In conclusion, a informative Database linking protein sequences to phenomena and interactions between proteins was built and PDM building process are promptly applicable to desired similar databases using the procedure communicated here.

ACKNOWLEDGEMENTS

This work was developed as AB-S PhD thesis which has been supported by CAPES/MEC. We thank FAPEMIG and CNPQ for the research support given to Biodados laboratory as well.

REFERENCES

Barbosa-Silva A, Pafilis E, Ortega JM, Schneider R (2007) Development of SRS.php, a Simple Object Access Protocol-based library for data acquisition from integrated biological databases. Genet Mol Res 6: 1142-1150

Barbosa-Silva A, Satagopam VP, Schneider R, Ortega JM (2008) Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence. BMC Bioinformatics 9: 141

Jones JD (2001) Putting knowledge of plant disease resistance genes to work. Curr Opin Plant Biol 4: 281-287

Mika S, Rost B (2004) NLProt: extracting protein names and sequences from papers. Nucleic Acids Res 32: W634-637

Schoof H, Ernst R, Nazarov V, Pfeifer L, Mewes HW, Mayer KF (2004) MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource for plant genomics. Nucleic Acids Res 32: D373-376

3.4 ANOTAÇÃO MANUAL DE GENES DE RESISTÊNCIA EM EUCALIPTO E VALIDAÇÃO DA ANOTAÇÃO EFETUADA PELA PDM.

Para concluir o conteúdo apresentado na tese, descrevemos a utilização das estratégias implementadas nas sessões anteriores para o desenvolvimento da Plant Defese Mechanisms Database.

A pesquisa de genes de resistência em plantas pode ser acelerada com a utilização da base de dados PDM. Nela, os genes de resistência estão agrupados segundo critérios ditados pelo programa Seed Linkage. As entradas são escolhidas por co-ocorrências na literatura. Portanto, sua formulação pode conter lacunas, dado que genes de resistência poderiam não ter sido encontrados em co-ocorrência com a ênfase escolhida para alimentar o programa LAITOR. Nesta sessão compararemos a composição da PDM com uma lista de candidatos a genes de resistência em eucalipto que foi gerada por anotação manual dos contigs desta planta. Evidenciamos que alguns genes não participam da PDM construída tendo como base as consultas processadas pelo LAITOR e, como esperado, não foram adicionados quando da amplificação dos agrupamentos pelo Seed Linkage.

A lista de candidatos a gene R de eucalipto foi publicada no trabalho "In silico survey of resistance (R) genes in Eucalyptus transcriptome" (Barbosa-Silva et al., 2005) e os detalhes da anotação das seqüências estão descritos no trabalho apresentado a seguir. Como é uma fonte confiável de anotação manual das seqüências, a lista apresenta-se como um bom controle para a capacidade de análise automática feita pela PDM. Lembramos que a base PDM, como está apresentada atualmente, não sofreu nenhuma adição de seqüências por interesse especial.

Inicialmente as seqüências de eucalipto foram utilizadas em buscas de similaridade com as seqüências constituintes da PDM utilizando-se BLASTx e um valor de cutoff de e-value de 1e-10. É esperado que a PDM realize anotação correta quando a proteína homóloga mais próxima evolutivamente estiver na base PDM. Caso isso não ocorra, dado que genes de resistência compartilham similaridades estruturais a ponto de serem agrupados em classes de resistência (Tabela 3), é provável que a PDM efetue uma especulação de anotação. Durante o desenvolvimento deste trabalho de tese colaboramos com a construção de uma ferramenta de anotação denominada "Protein Classification Tool" (PCT) acessada em http://biodados.icb.ufmg.br/pct. Para evitar problemas de especulação pelas bases secundárias, uma pesquisa de similaridade na base de dados não redundante do Entrez Protein (nr) foi realizada. Assim, caso o alinhamento com uma seqüência da base nr apresentasse escore mais

alto que o que ocorreu com a base secundária, o usuário era informado. Portanto, para avaliar a anotação feita pela PDM, nós comparamos os melhores alinhamentos obtidos com a base PDM e com a base nr. Utilizamos também como fonte de todo o universo protéico a base de dados UniProtKB, todavia como os resultados foram equivalentes aos obtidos com nr, faremos referência somente a esta última.

A Figura 8A mostra a diferença de escore (eixo Y) obtida pela pesquisa de similaridade com BLASTx feita com todas as seqüências de eucalipto. As seqüências de eucalipto foram ordenadas no eixo X de forma a mostrar as diferenças de escore nr – PDM de forma crescente. Percebe-se claramente que a maioria das seqüências apresenta diferenças de escore entre zero e 150 e que os resultados são aparentemente contínuos até o escore 150, quando uma população de resultados começa a se destacar desta tendência. Estudos posteriores devem ser realizados para avaliar se o limiar de 150 pode expressar a expectativa de tratar-se de uma seqüência que se apresente mais semelhante e esteja exclusivamente ausente da PDM. Por isso resolvemos analisar outras duas métricas para o mesmo efeito.

A Figura 8B representa experimento similar, todavia a métrica escolhida foi a razão entre o escore nr sobre o escore PDM. A ordenação dos resultados de forma crescente (eixo X) não é a mesma que a obtida para o experimento da Figura 8A (não mostrado), todavia é bastante aproximada. O limiar determinado por essa análise é igual a dois. Acima desta razão, é esperado que o alinhamento com a PDM não seja confiável.

Por fim, a figura 8C apresenta a mesma análise feita com uma terceira métrica, a diferença de identidades (coluna 3 do resultado tabular de BLAST) obtidas com nr ou PDM. Encontramos o valor 15 como um limiar para a anotação supostamente correta.

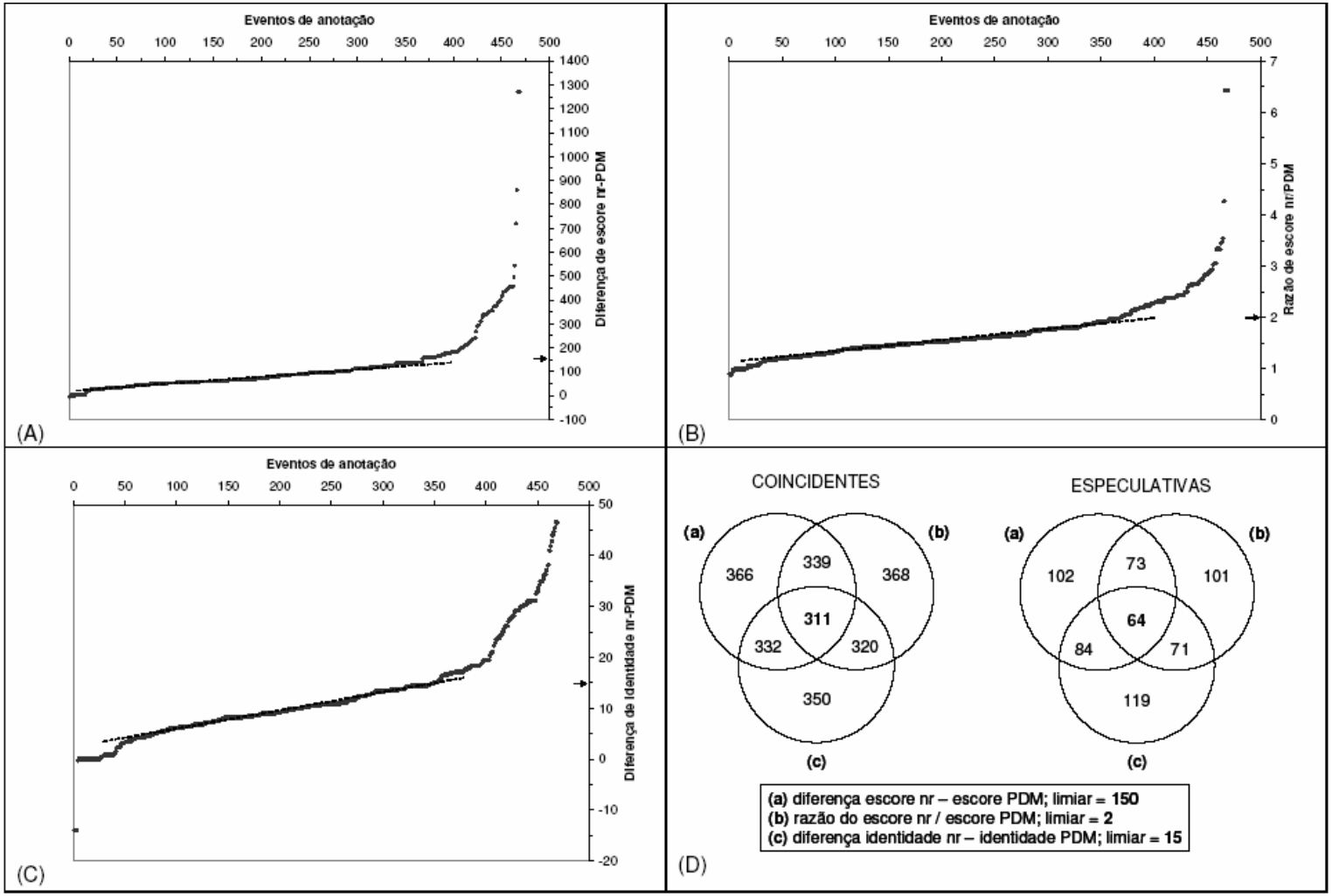


Figura 8: **Comparação entre métricas de anotação entre as bases PDM e nr. A)** Diferença de escore entre as bases, limiar (seta) 150. **B)** Razão do escore das bases, limiar 2. **C)** Diferença de identidade entre as bases, limiar 15. **D)** distribuição das anotações coincidentes e especulativas entre as bases.

A Figura 8D mostra a distribuição de COINCIDENTES (supostamente anotados tão bem por PDM quanto por nr) e de ESPECULATIVAS (provavelmente não bem anotados pela PDM por falta das seqüências apropriadas na base de dados). Escolhemos avaliar as seqüências selecionadas triplamente como COINCIDENTES ou ESPECULATIVAS.

Dos 469 alinhamentos de seqüências de eucalipto com membros da PDM, 311 foram classificados como positivos (66%), mostrando que os critérios de seleção definidos pelo LAITOR cobriram uma boa parcela do universo. Devemos lembrar que, contrabalanceando esta perda, os dados de co-ocorrência adicionados à PDM pelo LAITOR adiciona informação não facilmente disponível pela curadoria manual.

Dentre as seqüências ESPECULATIVAS, encontramos 29 ocorrências de seqüências de eucalipto anotadas como PTO, possivelmente pela ausência de um cluster contendo a própria seqüência PTO na base de dados PDM. Todavia, três seqüências anotadas como PTO apresentavam igual anotação na PDM e em nr. A análise manual dessas seqüências aponta para a semelhança estrutural de membros protéicos da família kinase, que alta similaridade de seqüência, e que justificariam a divergência na anotação desses clusters de eucalipto pelas duas bases consideradas.

Por outro lado, 64 seqüências de eucalipto sugerem a necessidade de inclusão de entradas de gene de resistência na PDM. As seqüências que denunciam falta de homólogos na PDM haviam sido anotadas como similares aos genes HRT, I2, R1, Rp1, Sw5, Xa1, Bs2, Gpa2, Mi1, Rx1, Rx2, Cf5, Cf4 e Cf9. O que chamou atenção para o desenvolvimento de novas regras de detecção de co-ocorrência do LAITOR, a serem implementadas a partir desse experimento. Para isso criaremos um filtro para validação de co-ocorrências notadamente com nomes de patógenos além das já existentes. Dado que embora determinadas proteínas conferirem resistência contra patógenos, esta não necessariamente co-ocorra na literatura científica com um dos tipos de termos previamente utilizados.

Considerando-se apenas os alinhamentos abaixo dos limiares descritos (COINCIDENTES), as quais devem portanto possuir homólogos significativamente próximos na PDM, pode-se agrupar as seqüências de eucalipto em 16 agrupamentos como mostrado na Tabela 3. Vários agrupamentos capturam seqüências com anotação manual similar, como o 548 e o 1326. Similares a estes são os agrupamentos 949 e 1149, os quais adicionam os genes P e RPS4.

Tabela 3: Anotação automática por cluster PDM das seqüências manualmente designadas como genes de resistência em eucalipto.

Cluster	Anotação manual para Gene R	Classe de resistência			
		KINASE	LRR	NBS	TIR
472	PTO	X			
518	PTO	X			
979	PTO	X			
1090	PTO	X			
1088	PTO CF4 CF5 CF9 XA21	X	X		
980	CF4 CF5 CF9 XA21	X	X		
982	CF4 CF5 CF9 XA21	X	X		
1339	CF4 CF5 CF9 XA21	X	X		
1156	CF4		X		
548	L M N RPM1 RPM4 RPM5		X	X	X
650	I2 rp1 mla rpm1 GPA2 HRT pib pita r1 RPP13 RPP8 RX2 SW5 XA1		X	X	X
949	L M N P RPP1 RPP4 RPP5 RPS4		X	X	X
1059	RPP5 RPS5		X	X	X
1149	L M N P RPP1 RPP4 RPP5 RPS4		X	X	X
1326	L M N RPM1 RPM4 RPM5		X	X	X
903	HR3 RPP8		X	X	X

Concluindo, a determinação experimental de limiares de confidência de anotação pela PDM como feita nesta sessão nos leva a dois resultados importantes: (i) o usuário pode avaliar, pela comparação de performance nr e PDM, se a classificação no agrupamento PDM é aceitável; (ii) conseguimos identificar casos onde é necessária inclusão de seqüências na PDM por uma vertente diferente ou complementar àquela oferecida pelo LAITOR. Essas novas seqüências, como por exemplo representantes de PTO bem documentados na literatura, podem ser incorporadas facilmente com o programa Seed Linkage e a série de programas que geram a parte gráfica da PDM fará a atualização das tabelas.

Genetics and Molecular Biology, 28, 3 (suppl), 562-574 (2005)

www.sbg.org.br

Research Article

In silico survey of resistance (*R*) genes in *Eucalyptus* transcriptome

Adriano Barbosa-da-Silva, Ana C. Wanderley-Nogueira, Raphaela R.M. Silva, Luiz C. Berlarmino, Nina M. Soares-Cavalcanti and Ana M. Benko-Iseppon

Universidade Federal de Pernambuco, Centro de Ciências Biológicas, Departamento de Genética, Laboratório de Genética e Biotecnologia Vegetal, Recife, PE, Brazil.

Abstract

A major goal of plant genome research is to recognize genes responsible for important traits. Resistance genes are among the most important gene classes for plant breeding purposes being responsible for the specific immune response including pathogen recognition, and activation of plant defence mechanisms. These genes are quite abundant in higher plants, with 210 clusters found in *Eucalyptus* FOREST database presenting significant homology to known R-genes. All five gene classes of *R*-genes with their respective conserved domains are present and expressed in *Eucalyptus*. Most clusters identified (93) belong to the LRR-NBS-TIR (genes with three domains: Leucine-rich-repeat, Nucleotide-binding-site and Toll interleucine 1-receptor), followed by the serine-threonine-kinase class (49 clusters). Some new combinations of domains and motifs of *R*-genes may be present in *Eucalyptus* and could represent novel gene structures. Most alignments occurred with dicots (94.3%), with emphasis on *Arabidopsis thaliana* (Brassicaceae) sequences. All best alignments with monocots (5.2%) occurred with rice (*Oryza sativa*) sequences and a single cluster aligned with the gymnosperm *Pinus sylvestris* (0.5%). The results are discussed and compared with available data from other crops and may bring useful evidences for the understanding of defense mechanisms in *Eucalyptus* and other crop species.

Key words: serine-threonine kinase, nucleotide binding site, leucine-rich repeats, gene-for-gene interaction.

Received: May 28, 2004; Accepted: March 28, 2005.

Introduction

Pathogen attack can severely affect crop production, with losses that can achieve 80% of the production especially in tropical countries. At the global level, losses have been estimated to accomplish around 12% of the world crop production (James *et al.*, 1990). The most important group of genes that has been used by breeders for disease control is the plant resistance (*R*) genes: single determinant of an effective and specific resistance that can often be characterized by localized necrosis at attempted infection sites (Rommens and Kishore, 2000).

It is proposed that pathosystems are usually highly specific, with a matching *R*-gene on vegetal cell that recognizes elicitor proteins (called Avr-effector) of each infective pathogen. Plant will be resistant and the growth of the pathogen will be arrested only when both genes, *R* and *Avr*, are present (Ellis *et al.*, 2000a). So, for each *R*-gene a correspondent *Avr* gene co-exists: this is the basis of the gene-for-gene concept, suggested by Flor (1956, 1971).

Send correspondence to Ana Maria Benko-Iseppon, Universidade Federal de Pernambuco, Centro de Ciências Biológicas, Departamento de Genética, Laboratório de Genética e Biotecnologia Vegetal, Av. Prof. Moraes Rego s/ n., 50732-970 Recife, PE, Brazil. E-mail: celisep@hotlink.com.br.

Avirulence gene products actually described do not comprise a defined family of related proteins, since no sharing similar motifs or domains could be found. On the opposite, *R*-gene products are separated into distinct but related protein classes, according to their conserved structural domains. Conserved domain function identified for R proteins suggests two fundamental mechanisms during pathogenic infection: (I) the pathogen recognition, conducted mainly by leucine-rich repeats (LRR) regions, which play a direct role in protein-protein specific recognition event; and (II) signaling of pathogen presence in order to activate defense related genes (Richter and Ronald, 2000).

The TIR (Toll interleukine 1-receptor) and CC (coiled coil) regions are involved in signal transduction during many cell processes (Martin *et al.*, 2003), while the NBS (Nucleotide Biding Site) usually signalizes for programmed cell death in animal cells (van der Biezen and Jones, 1998). Additionally, a kinase catalytic region is present in some *R*-genes. This domain plays a direct role in both signaling processes and pathogen effectors. Additionally the NBS region contains not only the three motifs involved in nucleotide binding but additional motifs as well. This extended region of homology is referred to as the NB-ARC domain (Richter and Ronald, 2000). Sometimes this do-

main contains a distinct predicted nucleoside triphosphatase (NTPase) domain known as NACHT, common in animal, fungal and bacterial proteins, implicated with apoptosis induction and transcription activation (Koonin and Avarind, 2000).

Resistance genes are members of a very large multigene family, are highly polymorphic and have diverse recognition specificities. They are commonly clustered in the genome, often in tandem direct repeats, what is consistent with the theory that they originated through gene duplication and that they are continuously evolving through unequal exchange (Song *et al.*, 1997).

Most of the resistance genes that have been cloned and characterized resemble components involved in signal transduction. These can be classified into five categories based on their predicted protein structure (Song *et al.*, 1997, Ellis and Jones, 1998).

The first class is represented by the *Pto* gene of tomato, which encodes a protein with a catalytic serine-threonine kinase (ser-thre-kinase) and a myristoylation motif in his amino terminal region (Martin *et al.*, 1993).

The second class comprises many proteins that present a region rich in repetitions of leucine (LRR, Leucine-rich repeats), a Nucleotide Binding Site (NBS) and a leucine zipper (LZ) or a coiled-coil (CC) sequence. Many genes encode proteins of this class: *I2* (Ori *et al.*, 1997), *Mi* (Milligan *et al.*, 1998) and *Sw5* (Brommonschenkel *et al.*, 2000) from tomato; *RPM1* (Grant *et al.*, 1995), *RPP8* (McDowell *et al.*, 1998), *RPS2* (Mindrinos *et al.*, 1994) and *RPP13* (Bittner-Eddy *et al.*, 2000) from *Arabidopsis thaliana*; *Pib* (Wang *et al.*, 1999), *Pi-ta* (Bryan *et al.*, 2000) and *Xa1* (Yoshimura *et al.*, 1998) from *Oryza sativa* (rice); *Gpa2* (Van der Vossen *et al.*, 2000), *Hero* (Ernst *et al.*, 2002), *R1* (Ballvora *et al.*, 2002), *Rx1a* (Bendahmane *et al.*, 1995) and *Rx2* (Bendahmane *et al.*, 2000) from potato; *Rp1* from maize (Collins *et al.*, 1999); *Mla* from barley (Halterman *et al.*, 2001) and *Dm3* from lettuce (Meyers *et al.*, 1998).

The third class includes similar proteins as described for class II, presenting a toll receptor for interleukine-I (IL-1R) instead of a CC sequence at the amino terminal region (Meyers *et al.*, 1999). This class is referred as TIR-NBS-LRR, including the genes *L* (Lawrence *et al.*, 1995), and *P* (Dodds *et al.*, 2001) of flax; *RPP1* (Botela *et al.*, 1998), *RPP4* (van der Biezen *et al.*, 2002), *RPP5* (Parker *et al.*, 1997) and *RPS4* (Gassmann *et al.*, 1999) of *A. thaliana* and *N* (Whithan *et al.*, 1996) of tobacco. This class (also present in animals) is supposed to be absent in monocotyledonous plants (Ellis and Jones, 1998), being present in all dicotyledonous taxa actually studied.

The proteins encoded by the three classes of genes previously cited do not present a transmembrane sequence and are therefore classified as intracellular *R*-proteins (Martin *et al.*, 2003).

The fourth class of resistance genes belongs to the tomato *Cf*-family, encoding similar proteins with an extracellular LRR and a short cytoplasmatic tail, but no NBS or any further recognizable domain (Dixon *et al.*, 1996). Member of this family are *Cf-2* (Dixon *et al.*, 1998), *Cf-4* (Joosten *et al.*, 1994; Thomas *et al.*, 1997), *Cf-5* (Dixon *et al.*, 1998) and *Cf-9* (Jones *et al.*, 1994).

The fifth class includes a single gene, *Xa21* from rice that presents an extracellular LRR, a transmembrane region (TM) and a cytoplasmatic ser-thre-kinase. Thus, the structure of *Xa21* indicates an evolutionary link between different classes of plant disease resistance genes (Song *et al.*, 1997).

There is still a sixth class that presents genes with no conserved domains, as described for the previous five classes. This group comprises the gene *Hm1* from maize, a reductase that confers resistance to the fungus *Cochliobolus carbonum* (Johal and Briggs, 1992); *Mlo* from barley, a putative regulator of defense against *Blumenaria graminis* (Piffanelli *et al.*, 2002) possibly associated to the plasma membrane (Buschges *et al.*, 1997); and *RPW8* from *A. thaliana*, that confers non-specific resistance to the fungus *Erysyphe chicoracearum* (Xiao *et al.*, 2001).

Due to its qualities as high level of adaptability, fast growing capacity and wood quality, *Eucalyptus* plantations are carried out in all tropical areas in diverse continents. *Eucalyptus* is the most widely used tree for delivering raw material for the paper industry used in the production of cellulose and to regenerate degraded areas. Over the past 50 years large-scale planting of fast growing exotic *E. grandis*, *E. urophyla*, *E. saligna* and many hybrids (particularly *grandis* x *urophyla*) has occurred in Brazil aiming to reforest some regions and to create an adequate supply of wood, timber and fuel for different purposes (McNabb, 2002). In the late 2001s growing areas reached 138.132 ha, generating more than 7,398 direct employments (BRACELPA, 2004).

The advance of plantations to hot and humid areas resulted in favourable conditions to the development of diseases especially in young individuals that are often severely attacked by fungal (*e.g. Mycosphaerella cryptica*, *Dichomera versiformis*, *Cylindrocladium* spp. and *Phaeophleospora epicpccoides*) and bacterial pathogens (Barber *et al.*, 2003, Mafia and Alfenas, 2003).

Eucalyptus Genome Sequencing Consortium (FOREST) aimed to identify over 15,000 expressed genes from 100,000 sequenced EST from 19 libraries from specific tissues and stages.

The present work aimed to perform a data mining-based identification of plant disease *R*-genes in FOREST database, by using well known *R*-genes sequences as template, comparing the identified sequences with known *R*-genes deposited in public DNA and protein databases.

Materials and Methods

Amino-acid sequences of known genes have been used as query in the search for *R*-gene homologues and analogs in *Eucalyptus* transcriptome database. Accession numbers at NCBI (National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov) of sequences used are shown in Table 1, together with sequences features and accession numbers. They are grouped according to the conserved domains previously described. Members of the sixth class (reductases and other *R*-genes with no recognizable conserved domains) have not been included in the present evaluation.

All *Eucalyptus* sequences used during this work were obtained from FOREST project and derived from cDNA libraries specific to different tissues, organs or conditions of growth from the species *E. grandis*, *E. globulus*, *E. saligna* and *E. urophylla*. For detailed information see https://forests.esalq.usp.br/Librariesinfo.html.

Reverse alignments were realized on 'FOREST EG_Clusters' database using the program TBLASTN (Altschul *et al.*, 1990), the e-value cutoff adopted was $1e^{-23}$. Matching clusters to query sequences were then annotated on a local database called 'non-redundant' made with aid of the Microsoft Access® program. Cluster name was adopted as primary key in order to prevent data redundancy regarding clusters aligning with more than one query sequence. In the few cases when this occurred the name of both queries has been also annotated for the respective cluster.

The clusters frame of the TBLASTN alignment was used to predict the Open Reading Frames (ORFs) for each searched cluster. For this purpose, the Expasy Translate Tool (bo.expasy.org/tools/dna.html) was used, which predicts the correct ORF for a DNA sequence in the corresponding amino acid FASTA sequence. The obtained ORFs were subsequently submitted to a Reverse Position Specific BLAST (RPS-BLAST) against Conserved Domain Database (Marchler-Bauer *et al.*, 2002) aiming to identify patterns or motifs in predicted cluster products.

Reciprocal alignments were conducted for ORFs by downloading the nr databank and stand alone BLAST package from NCBI ftp site for local use at our server (Laboratório de Genética e Biotecnologia Vegetal, UFPE) performing a high-throughput alignment approach. Matched sequences were annotated for latter comparison.

Predictions of subcellular localization have been inferred by using TargetP program available at CBS (Center for Biotechnology Sequence Analysis) Prediction Servers

Table 1 - Classification and features of *R*-genes used as query against the FOREST database. The used genes are grouped in five *R*-gene classes (**I**: Kinase; **II**: LRR+NBS; **III**: LRR+NBS+TIR; **IV**: only LRR; **V**: LRR+Kinase) with respective accession number at NCBI, source species, gene name and domain range (in amino-acids).

Class of *R*-gene	Accession number	Source species	Gene name	Sequence size	Domain range (initial-last aa)							
					LRR		Kinase		NBS		TIR	
				(aa)	Start	End	Start	End	Start	End	Start	End
I	2112354A	*Lycopersicon esculentum*	*Pto*	321	-	-	41	236	-	-	-	-
	AF234174_1	*Arabidopsis thaliana*	*HRT*	909	579	868	-	-	150	460	-	-
	NP_172686.1	*Arabidopsis thaliana*	*Rps5*	889	540	636	-	-	140	444	-	-
II	AF118127_1	*Lycopersicon esculentum*	*I2*	1266	578	1231	-	-	154	457	-	-
	AAG31014.1	*Lycopersicon esculentum*	*Sw5*	1246	-	-	-	-	519	818		
	BAA25068.1	*Oryza sativa*	*Xa1*	1802	771	1773	-	-	283	593	-	-
	AAP81262.1	*Zea mays*	*Rp1*	1269	596	1228	-	-	145	457	-	-
	AAC72977.1	*Arabidopsis thaliana*	*RPP1*	1189	668	1011	-	-	226	505	54	184
	RP13_ARATH	*Arabidopsis thaliana*	*RPP13*	835	-	-	-	-	147	453	14	148
	AF440696_1	*Arabidopsis thaliana*	*RPP4*	1135	642	1053	-	-	185	441	15	145
	AAF08790.1	*Arabidopsis thaliana*	*RPP5*	1361	643	1151	-	-	188	465	14	148
	RPP8_ARATH	*Arabidopsis thaliana*	*RPP8*	908	577	867	-	-	149	459	15	145
	BAB11393.1	*Arabidopsis thaliana*	*Rps4*	1232	663	889	-	-	198	473	21	149
III	AAP41025.1	*Lactuca serriola*	*RGC2*	352	49	235	-	-	-	-	21	149
	AF093649_1	*Linum usitatissimum*	*L*	1294	607	1277	-	-	220	521	63	195
	T18548	*Linum usitatissimum*	*M*	1305	744	1288	-	-	235	534	78	210
	AF310960_2	*Linum usitatissimum*	*P*	1211	693	1023	-	-	205	238	23	153
	AF202179_1	*Capsicum chacoense*	*Bs2*	905			-	-	152	439	63	195
	A54810	*Nicotiana glutinosa*	*N*	1144	597	908	-	-	172	447	14	147
	AF195939_1	*Solanum tuberosum*	*Gpa2*	912	561	863	-	-	119	422	14	147
	CAA61264.1	*Solanum tuberosum*	*Rx1*	248			-	-	-	-	23	153
	CAB56299.1	*Solanum tuberosum*	*Rx2*	938	561	859	-	-	138	422	78	210
	CAD29728.1	*Solanum tuberosum*	*HERO*	1283	-	-	-	-	504	811	54	184
IV	T07015	*Lycopersicon esculentum*	*Cf4*	855	81	758	-	-	-		-	-
	AAC78591.1	*Lycopersicon esculentum*	*Cf5*	968	96	855	-	-	-	-	-	-

site (http://www.cbs.dtu.dk/services/). Additionally, transmembrane helix segments were inferred with aid of the TMHMM program as well.

Results

After the TBLASTN alignments performed at FOREST EG_Clusters database, a total of 478 clusters aligned with the diverse *R*-genes (Table 1) used as query (data not showed). These clusters were, as described in section 'Material and Methods', inserted on a local database called 'non-redundant'. This procedure generated a set of 210 non-redundant clusters which have been annotated for one or more than one *R*-gene (data summarized in Figure 1 and Tables 2 and 3).

Clusters representing exclusive *R*-gene classes were: (I) serine-threonine kinase (here named KINASE): 49; (II) LRR+NBS: 21; (III) LRR+NBS+TIR: 93; (IV) Only LRR + Transmembrane (LRR+TM): 17 and (V) LRR+TM+ Kinase: 8 (Figure 1).

Regarding the sequence identity of the best alignment, 22 clusters showed equally significant similarity to two different classes of *R*-genes. From these, 18 included LRR plus LRR-Kinase here called MIX I (sequence data presented in Table 3); three included NBS-LRR plus TIR-NBS-LRR (called MIX II) and one LRR plus Kinase (called MIX III).

Sizes of *Eucalyptus* clusters aligned to *R*-genes varied from 3,316 (cluster EGEQRT3301C03 classified to group MIX-III) to 520 nucleotides. The prediction of clusters coding regions revealed that ORFs were coded in both forward and reverse reading frames, with an average of 304 amino acids (aa) in length. ORF sizes varied from 990 (cluster EGEQRT3301C03 of the LRR-KINASE class) to 134aa. Regarding the average ORF length in each *R*-gene class, we observed 417aa for KINASE, 276aa for NBS, 238aa for TIR-NBS-LRR, 247aa for LRR-TM, 352aa for LRR-KINASE, 372aa for MIX I, 343aa for MIX II and 990aa for MIX III class.

The search for conserved domains (CD-Search) revealed conserved regions (Figure 1, Table 1) in 166 of the 210 here analyzed clusters. A total of 40 clusters presented the kinase domain, 37 of them matched to *Pto* gene (class I) after the TBLASTN alignment, with only three grouping into KINASE-LRR (two of them) and MIX III (one of them) classes. These two classes also showed associated LRR segments as well. Regarding the LRR domains, these could be identified in 67 different clusters in all classes (except KINASE class I, represented by *Pto*) with a total of 442 occurrences. This number is higher than the number of clusters due to their occurrence in tandem repetitions. Sometimes these sequences are imperfect and may be difficult to recognize with available *in silico* tools, so it is possible that a larger number may be identified manually.

Twenty clusters showed the NB-ARC domain. In a specific case, this domain occurred associated to a different TIR domain as was cited above. Additionally, a NACHT domain (closed-related to NB-ARC) was identified exclu-

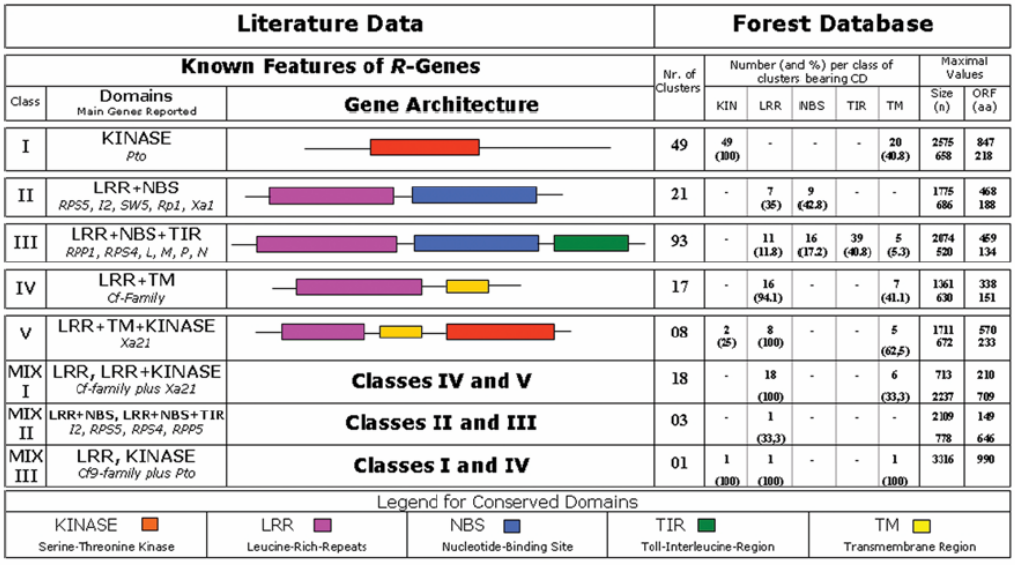


Figure 1 - Representation of main *R*-genes classes considering the presence and position of conserved domains from literature data, as compared with *Eucalyptus* clusters from FOREST database. For each class the data about significant alignments to *R*-genes is given, including following information: number of clusters identified for each class (clusters aligning with more than one class are not included), number and percentage of clusters per class bearing indicated conserved domains, size range (maximal and minimum) of sequence in nucleotides (n) and of ORF in amino-acids (aa). Abbreviation: CD = Conserved domains.

Table 2 - Blast results and sequence evaluation of *Eucalyptus* R genes, including the best matches of each *R* gene and MIX classes: (I) data about the query: gene class and name, NCBI gi|-number, species and family. (II) Features and evaluation results of *Eucalyptus* clusters related to R-genes: cluster number, cluster size in nucleotides (n), ORF (Open Reading Frame) size in amino-acids (aa), e-value; score and frame.

(I) Query Information				(II) Cluster features and evaluation					
Gene class & expected domain	Gene name	NCBI gi|-nr.	Plant species and family	*Eucalyptus* cluster n.	Size (n)	ORF (aa)	E-value	Score and frame	
Class I KINASE	*Pto*	27754635	*Arabidopsis thaliana*	EGEQRT3100D07	2460	722	0.0	1036,6	2
	Pto	15235204	*Arabidopsis thaliana*	EGEQRT3104A12	2575	847	0.0	909,8	1
	Pto	18418211	*Arabidopsis thaliana*	EGUTFB1098H02	2511	616	0.0	904,0	3
	Pto	10177052	*Arabidopsis thaliana*	EGCBRT3133E11	1728	575	0.0	738,0	3
	Pto	25405628	*Arabidopsis thaliana*	EGMCRT3148C12	1705	568	0.0	709,9	1
Class IV LRR	*Cf5*	14626935	*Gossypium hirsutum*	EGEQSL5001G09	1223	321	$2.00e^{-139}$	496,1	2
	Cf5	15240263	*Arabidopsis thaliana*	EGCCRT3339F06	922	307	$3.8e^{-83}$	309,3	2
	Cf5	15239124	*Arabidopsis thaliana*	EGCBST2063A06	697	232	$2.0e^{-61}$	236,5	-2
	Cf4, Cf5	27754637	*Arabidopsis thaliana*	EGACRT3321G06	1361	338	$1.4e^{-60}$	234,6	3
	Cf4, Cf9	14269077	*Lycopersicom esculentum*	EGJMCL1299H10	682	226	$1.9e^{-53}$	209,9	3
Class V LRR KIN	*Xa21*	9651941	*Glycine max*	EGRFRT3357D01	1584	527	0.0	869,4	3
	Xa21	15239540	*Arabidopsis thaliana*	EGEQCL1200B12	1711	570	0.0	658,7	1
	Xa21	19881587	*Oryza sativa*	EGJEST2023F09	716	238	$5.6e^{-51}$	201,8	1
	Xa21	15218385	*Arabidopsis thaliana*	EGSBCL1280C05	725	241	$7.8e^{-48}$	191,4	3
	Xa21	15218385	*Arabidopsis thaliana*	EGSBCL1280C05	725	241	$3,00e^{-04}$	191.4	3
Class II NBS LRR	*Hrt, I2, Sw5, Xa1, Rp1, R1*	18652501	*Oryza sativa*	EGUTRT3110A12	1041	346	$1.2e^{-62}$	241,5	3
	I2, Xa1, Rpm1, Rp1, R1, Pib, Mi1	28300299	*Manihot esculenta*	EGJFSL4202E08	876	291	$5.3e^{-55}$	215,7	3
	Bs2, Gpa2, I2, Rx1, Rx2, Sw5, Xa1, Rp1Mi1	15487949	*Theobroma cacao*	EGEQCL1001F08	934	311	$2.1e^{-52}$	207,2	1
	Gpa2, Rx2, Rpm1, R1, Pib, I2	28300299	*Manihot esculenta*	EGJERT3026C12	804	267	$7.7e^{-50}$	198,4	2
	Gpa2, Hrt, Rpp13, Rpp8, Rx2, Sw5, Rpm1R1, Pi-Ta, Pib	22775643	*Oryza sativa*	EGCECL1282E03	779	231	$1.1e^{-48}$	194,1	1
Class III TIR NBS LRR	*L, M, N, P, Rpp1, Rpp4, Rpp5, Rps4*	7488903	*L. usitatissimum*	EGJMFB1107C10	1395	445	$1.7e^{-83}$	311.2	1
	L, M, N, P, Rpp1, Rpp4, Rpp5, Rps4	9965103	*Glycine max*	EGMCLV2264D03	1155	329	$7.4e^{-75}$	282,0	3
	L, M, N, P, Rpp1, Rpp4, Rpp5, Rps4	12056928	*Glycine max*	EGJEST2234G10	1270	420	$8.8e^{-74}$	278,9	-1
	L, M, N, P, Rpp1, Rpp4, Rpp5, Rps4	27764536	*Glycine max*	EGJMST6019E06	1155	351	$3.0e^{-69}$	263,5	1
	L, M, N, P, Rpp1, Rpp5, Rps4	23477203	*Populus balsamifera*	EGCBRT6029A01	1227	378	$2.6e^{-66}$	253,8	1
MIX I (LRR and LRR-KIN)	*Cf4, Cf5, Cf9, Xa21*	25287710	*Arabidopsis thaliana*	EGBMRT3129F10	2129	709	0.0	705,3	2
	Cf4, Cf5, Cf9, Xa21	21391894	*Lycopersicon peruvianum*	EGCEST2256F04	1687	561	0.0	704,1	3
	Cf4, Cf5, Cf9, Xa21	15240215	*Arabidopsis thaliana*	EGUTFB1136E01	1911	595	$5.00e^{-172}$	605,9	2
	Cf4, Cf5, Cf9, Xa21	15240528	*Arabidopsis thaliana*	EGEZST2207A10	2237	465	$2.00e^{-143}$	510,8	1
	Cf5, Xa21	15230539	*Arabidopsis thaliana*	EGEQRT3201E07	1229	376	$4.00e^{-127}$	455,7	1
MIX II TIR-NBS-LRR and NBS-LRR	*Rpp5, Rps4, I2*	15218365	*Arabidopsis thaliana*	EGEZRT3006B12	2109	646	$1.1e^{-50}$	203,0	3
	Rpp5, Rps5	15221252	*Arabidopsis thaliana*	EGEQST6001H02	778	234	$8.9e^{-33}$	141,4	1
	Rpp5, Rps5	15487963	*Theobroma cacao*	EGJECL1208G03	871	149	$5.8e^{-24}$	110,5	1
MIX III KINASE and LRR	*Cf5, Cf9, Pto*	26450791	*Arabidopsis thaliana*	EGEQRT3301C03	3316	990	0.0	1293,9	1

sively in two TIR-NBS-LRR related clusters (EGCCCL1328B05.g and EGSBRT3118H01).

Most of the 44 clusters with no conserved domains presented shorter ORFs (262 aa in average), with four of them presenting a putative transmembrane region.

A graphic representation of the distribution of conserved domains as compared with class-grouped clusters is presented in Figure 2.

Considering the best matches to the 210 clusters identified, 198 were from plants of Dicotyledonous families, with emphasis on *A. thaliana*. From monocots only rice (*O. sativa*) sequences appeared as best matches (11 clusters). One of the sequences from MIX III group aligned with *Pinus silvestris* (Gymnosperm), the only non-Angiosperm included in the present study. A comprehensive inventory of all species that aligned with *Eucalyptus* with their taxonomic affiliation and habit (herbaceous or woody) is presented in Table 4.

The post-translational inferences carried out for cluster products (TargetP program) revealed a large number of

Table 3 - FOREST clusters classified in the MIX I group, resembling to genes which belong to LRR and LRR-KINASE classes, including: respective templates (query sequences), cluster number and size in nucleotides (n), ORF-size in amino-acids (aa), range of LRR domain after CD-search, identity and results of the best alignment (BLASTp) in NCBI (GI number, species, score and e-value).

Template	Cluster	Size (n)	ORF (aa)	LRR-domain		GI	Description	Score	E-value
				Start	End				
	EGBMRT3129F10.g	2129	709	139	620	25287710	*Arabidopsis thaliana*	705.3	0.0
	EGUTFB1136E01.g	1911	595	91	523	15240215	*Arabidopsis thaliana*	605.9	$5e^{-172}$
Cf4	EGEZST2207A10.g	2237	465	22	262	15240528	*Arabidopsis thaliana*	510.8	$2e^{-143}$
Cf5	EGEQST2201G12.g	2049	412	36	223	25402587	*Arabidopsis thaliana*	360.9	$1.7e^{-98}$
Cf9	EGUTRT3368G02.g	799	265	38	254	15225805	*Arabidopsis thaliana*	323.2	$2e^{-87}$
Xa21	EGCEST2256F04.g	1687	561	22	501	21391894	*Lycopersicon peruvianum*	704.1	0.0
	EGUTSL4018B05.g	1384	447	96	432	3894385	*Lycopersicon esculentum*	257.3	$3e^{-67}$
Cf5, Cf9	EGSBRT3314G03.g	1263	412	118	380	15223460	*Arabidopsis thaliana*	408.3	$9e^{-113}$
Xa21	EGBMRT3131G11.g	1155	384	2	336	15237312	*Arabidopsis thaliana*	349.4	$4.7e^{-95}$
	EGEQRT3201E07.g	1229	376	106	348	15230539	*Arabidopsis thaliana*	455.7	$4e^{-127}$
	EGABST2047C09.g	773	210	13	180	15225805	*Arabidopsis thaliana*	249.6	$1.9e^{-65}$
	EGBMSL4023G05.g	729	242	4	219	15237426	*Arabidopsis thaliana*	226.9	$1.7e^{-58}$
	EGCBST6013F02.g	808	265	32	254	18700171	*Arabidopsis thaliana*	214.5	$1.0e^{-54}$
Cf5	EGCESL5078H03.g	771	257	27	245	15237426	*Arabidopsis thaliana*	211.1	$1.1e^{-53}$
Xa21	EGBGLV3221H06.g	734	235	24	213	3894383	*Lycopersicon esculentum*	200.7	$1.2e^{-50}$
	EGCBRT6048F01.g	713	237	13	227	3641252	*Malus* X *domestica*	327.4	$8.9e^{-89}$
	EGEZST2003B08.g	1586	353	107	346	21952787	*Oryza sativa* (cv. *japonica*)	304.7	$1.2e^{-81}$
	EGEPSL4003G09.g	857	278	59	251	12054894	*Pinus sylvestris*	240.4	$1.9e^{-62}$

predictions (Figure 3). The reliability class (RC), which is a confidence measure for the prediction, showed that only 11 sequences were defined into RC1 (higher than 80%), and 53 for RC2 (higher than 60%) class. Most of the sequences are predicted to be located at unspecific subcellular localization (133 sequences) while 35, 20 and 19 were predicted to contain mitochondrial targeting, signal and chloroplast transit peptides, respectively (Figure 3).

After evaluation with the TargetP program, sequences with motifs specific for transmembrane anchoring could be identified in 44 of all analyzed sequences. From these 19 belonged to LRR or LRR-KINASE-related sequences and, unexpectedly, five showed to be TIR-NBS-LRR and 20 to be KINASE-related sequences.

Discussion

The reverse alignment (TBLASTN) strategy (Altschul *et al.*, 1997) adopted by our group identified a set of 210 clusters similar to the major classes of disease *R*-genes in the current version of the FOREST database,

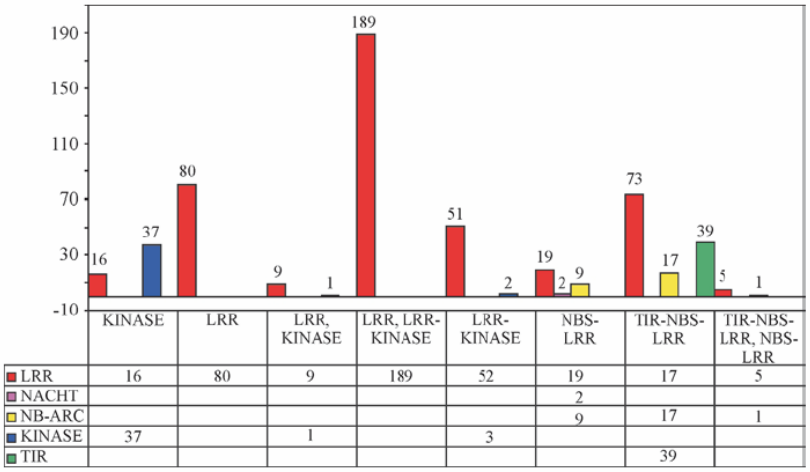


Figure 2 - Graphic representation of the distribution of conserved domains against class-grouped clusters. Values on the base after each domain indicate the number of clusters of each class presenting the indicated domain (also represented in the corresponding columns). Abbreviations: LRR = Leucine-rich-repeats; NB-ARC = Nucleotide-binding-site and additional motifs; NACHT = NB-ARC related domain, including an NTPase implicated in apoptosis and MHC transposition activation.

Table 4 - Inventory of the organisms that appeared as best alignment to each of the 210 here identified *Eucalyptus* clusters related to known resistance genes. The organisms are grouped by gene class (I to V and MIX I to III), taxonomic affiliation (class, subclass, family and species) and habit (herbaceous or woody). Numbers in parenthesis indicate amount of gene members in each taxonomic group or species.

Gene class	Higher taxonomic affiliation	Family	Species	Habit
I KINASE	Dicots (46)	Brassicaceae (43)	*Arabidopsis thaliana* (42)	H
			Brassica napus (1)	H
		Curcubitaceae (1)	*Cucumis melo* (1)	H
		Salicaceae (1)	*Populus nigra* (1)	W
		Solanaceae (2)	*Capsicum anuum* (1)	H
			Nicotiana tacabum (1)	H
	Monocots (2)	Poaceae (2)	*Oryza sativa* (2)	H
II LRR-NBS	Dicots (14)	Asteraceae (1)	*Lactuca sativa* (1)	H
		Brassicaceae (6)	*Arabidopsis thaliana* (6)	H
		Euphorbiaceae (3)	*Manifot esculenta* (3)	W
		Leguminosae (2)	*Glycine max* (1)	H
			Phaseolus vulgaris (1)	H
		Stercullariaceae (2)	*Theobroma cacao* (2)	W
	Monocots (7)	Poaceae (7)	*Oryza sativa* (7)	H
III NBS-LRR-TIR	Dicots (93)	Brassicaceae (10)	*Arabidopsis thaliana* (10)	H
		Curcubitaceae (5)	*Cucumis melo* (5)	H
		Leguminosae (10)	*Glycine max* (10)	H
		Asteraceae (14)	*Helianthus annuus* (14)	H
		Linaceae (34)	*Linum usitatissimum* (34)	H
		Euphorbiaceae (1)	*Manihot esculenta* (1)	W
		Salicaceae (12)	*Populus balsamifera* (10)	W
			Populus tremula (2)	W
		Solanaceae (7)	*Lycopersicon esculentum* (1)	H
			Solanum tuberosum (6)	H
IV LRR	Dicots (17)	Brassicacea (7)	*Arabidopsis thaliana* (7)	H
		Leguminosae (1)	*Glycine max* (1)	H
		Malvaceae (1)	*Gossypium hirsutum* (1)	W
		Solanaceae (8)	*Lycopersicon esculentum* (3)	H
			Lycopersicon hirsutum (2)	H
			Nicotiana tabacum (1)	H
			Petunia X *hybrida* (1)	H
			Solanum tuberosum (1)	H
V LRR-KINASE	Dicots (7)	Brassicaceae (6)	*Arabidopsis thaliana* (6)	H
		Leguminosae (1)	*Glycine max* (1)	H
	Monocot (1)	Poaceae (1)	*Oryza sativa* (1)	H
MIX I	Dicots (16)	Brassicaceae (12)	*Arabidopsis thaliana* (12)	H
		Solanaceae (3)	*Lycopersicon esculentum* (2)	H
			Lycopersicon peruvianum (1)	H
		Rosaceae (1)	*Malus* X *domestica* (1)	W
	Monocot (1)	Poaceae (1)	*Oryza sativa* (1)	H
	Gymnosperm (1)	Pinaceae (1)	*Pinus sylvestris* (1)	W
MIX II	Dicots (3)	Brassicaceae (2)	*Arabidopsis thaliana* (2)	H
		Stercullariaceae (2)	*Theobroma cacao* (1)	W
MIX III	Dicot 1	Brassicaceae (1)	*Arabidopsis thaliana* (1)	H

Synopsis regarding features of aligned species

		N.	%
Grouped by taxonomic affiliation	Dicots	198	94,3
	Monocots	11	5,2
	Gymnosperm	1	0,5
Grouped by habit	Herbaceous	187	89,0
	Woody	23	10,9

what comprises 0.63% of the actually generated clusters. This approach allowed the identification of a large set of candidate sequences by using various representative genes per class, while some recent works employed few genes (Koczyk and Chelkowski, 2003). Using several previously described and sequenced *R*-genes as template was a useful and low-time consuming strategy in the search for *R*-genes candidates in plants. In this approach it was expected that some similar genes grouped at the same class should cause some level of redundancy (Meyers *et al.*, 1999). The strategy of generating a local database (called non-redundant) by adopting the cluster number as a primary key register was very effective in the solution of this problem. Additionally, this approach was useful in the identification of the respective *R*-gene class for each *Eucalyptus* cluster.

The number of *R*-genes here identified is quite high, especially considering that none of the 19 libraries were obtained under pathogen stress condition. By the other hand, when additional ESTs are generated especially under infection by pathogen, many of the identified clusters may be united in larger clusters of *R*-genes that may include more domains.

Evidences have shown that *R*-genes are quite abundant in higher plants, but the most functionally defined *R*-genes belong to the supergene LRR-NBS family. After completing the whole genome sequencing of the model plant *A. thaliana* a total of 85 TIR-NBS-LRR have been identified (The Arabidopsis Genome Initiative, 2000), less than the number of clusters (93) actually identified in *Eucalyptus*. Especially genes containing NBS-LRR domains were estimated to be in number of ca.166 for *A. thaliana* and ca.600 for rice (*O. sativa*) by Richly *et al.* (2002), but this later number is still not confirmed.

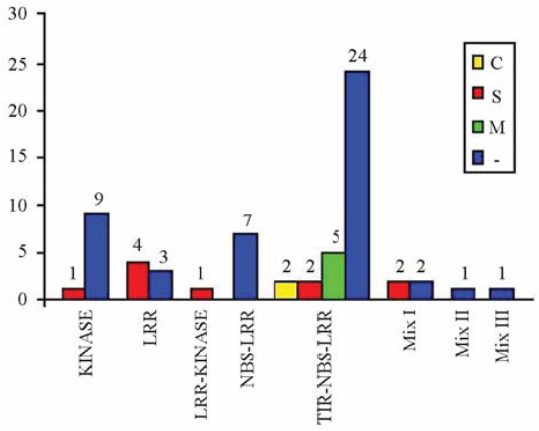


Figure 3 - Subcellular prediction for each class of analyzed *R* genes in *Eucalyptus* transcriptome, considering the predictions whit RC1 and RC2 of Target-P program. Legend: C = Chloroplast transit peptides; S = Secretory pathway; M = Mitochondrial targeting; '-' = No specific localization; LRR = Leucine-rich-repeat; NBS = Nucleotide-binding-site; TIR = Toll-interleucine-1 receptor.

A recent work reevaluated and reannotated all NBS-LRR encoding genes in *A. thaliana* genome database, revealing 149 genes of this class (including 94 TIR and 55 non-TIR sequences) in the genome of *A. thaliana* (Meyers *et al.*, 2003). In our evaluation of FOREST database we found 114 clusters (93 and 21, respectively) of this class. It is interesting to note that in the evaluation of Meyers *et al.* (2003) not only the presence of the TIR or of the CC motif was determinant for the grouping of both distinct classes. Also the NBS-LRR domains co-evolved and were determinant in the divergent evolution of the two groups, with the CC-bearing sequences forming four subgroups and the TIR-bearing sequences forming eight subgroups, regarding the size, composition and order or introns and exons.

Pan *et al.* (2000) compared tomato and *Arabidopsis* sequences of this class by systematically amplifying the tomato genome using a variety of primer pairs based on ubiquitous NBS motifs, generating 70 sequences, from which 10% were putative pseudogenes. The sequences were also used in mapping approaches, revealing a clustering *R*-gene homologues between tomato and potato (*Solanum tuberosum*, also from the Solanaceae family). Clustering of *R*-genes was also detected in *A. thaliana*, with most of the genes located in chromosomes 1 (49) and 5 (55), confirming the initial hypothesis that these genes are clustered in few chromosomes (The Arabidopsis Genome Initiative, 2000). This fact was also observed in other crops, as chickpea (*Cicer arietinum*; Benko-Iseppon *et al.*, 2003). In this last case, with some synteny and colinearity within this species and *Arabidopsis*. The clustering of *R*-genes in specific chromosomes and the existence of conserved domains have allowed the establishment of interesting strategies for identification, mapping and breeding directed to the incorporation of such genes from wild relatives. Considering the number of genes from this group in this last species, it is to expect that they are also clustered in *Eucalyptus*, what can also be valuable for the establishment of *Eucalyptus* breeding strategies in the future, especially considering the previous existence of mapping populations for this crop.

Overall annotation revealed that *Arabidopsis* also carries homologues of other *R*-gene classes, including 174 genes encoding LRR-kinases (*Xa21* group), but many of which are likely to play a role in development rather than defense (Jones, 2001). The present work revealed only eight clusters with significant homology to *Xa21* but this number can increase if only the kinase sequence is used as template, since the LRR may be quite variable between rice and *Eucalyptus*. Exceptional *R*-genes have proven to provide durable disease control, due to the fast evolving pathogen genome that breaks resistance. The *Xa21* gene is an important exception to this rule that reveals the full potential of *R*-genes for breeding purposes (Rommens and Kishore, 2000). This may be very valuable especially considering the possibility of pyramidization of such genes in important crops, increasing the potentiality of an effective specific *R-Avr* intection.

Another abundant family of *R*-genes in plants is the ser-thr-kinase with about 50 genes in *Arabidopsis* encoding protein kinases that are strongly homologous to tomato's *Pto* gene (Jones, 2001). In *Eucalyptus* we found almost the same number (49) of clusters also with high homology to the *Pto* sequence.

Regarding *R*-gene classes identified in *Eucalyptus*, an interesting phenomenon was observed in the present work: *R*-genes pertaining to different classes were able to align significantly to the same cluster on *Eucalyptus* database. This can be explained by the evidences that known *R*-genes combine a limited number of related functional domains (Ellis *et al.*, 1999, 2000a). Then, similar motifs would be present in different *R*-genes, and it is possible that a gene resembling to a determined class may search another belonging to a different class by local similarity at the site of the conserved motif. But in the practice, previous works do not speculate this possibility, once that the genes identified for specific *R*-genes are directly assigned to its own class as shown by evidences raised from works previously reported (Ronald, 1997; Jones, 2001; Romeis, 2001).

The MIX class one (MIX I) included 18 clusters resembling to genes which belong to both LRR and LRR-KINASE classes. These clusters were searched basically by using Cf (Jones *et al.*, 1998) and Xa21 (Song *et al.*, 1995) amino acid sequences as queries. In this case, the most plausible explanations would be the presence of the LRR domain, common to both classes, being responsible for the alignment and grouping of some clusters in both classes. By the other hand, LRRs are referred as fast evolving sequences and are in some cases quite imperfect, making manual annotation necessary. Often their amino-acid sequences are quite specific to their gene group (Dixon *et al.*, 1998; Ellis *et al.*, 1999). For example, using the LRR of *Xa21* against GenBank database will reveal significant alignments only to *Xa21* genes of rice (and some other Poaceae) and less significantly to *Arabidopsis*, but no sequence including other gene classes align significantly. A similar approach to the present work was used for the analysis of SUCEST (Sugarcane EST project, also running in Brazil) database (Morais, 2003) with no similar results. Song *et al.*, (1997) suggested that the structure of *Xa21* (here referred as class V) itself indicates an evolutionary link between different classes (I and IV) of plant disease resistance genes. May this be the case of this cluster that present a new link between two classes and can represent a new gene for Angiosperms?

Another surprising result was obtained by analyzing the unique cluster with both domains LRR and KINASE. It would be expected to find both domains in genes resembling *Xa21* but this cluster (EGEQRT3301C03.g) showed itself similar to both *Pto* (class I, described by Martin *et al.*, 1993) and *Cf* (Class IV, described by Jones *et al*, 1994)

genes. This double similarity occurred on different motifs. The *Pto* gene is known to encode a ser-thre-kinase protein (Martin *et al.*, 1993) and it was at this motif that the cluster showed similarity to this gene. On the other hand, *Cf* genes encode extracellular LRRs and it was at the LRR motif that the similarity was found. This cluster could be grouped in the LRR-KINASE class. So, why did it not align with *Xa21*, the single known gene with both LRR-KINASE domains? It should be answered by analyzing the KINASE-related clusters. Despite of the conservation of this region (Romeis, 2001), none of the *Pto* (KINASE) or *Xa21* (LRR and a receptor-KINASE) related clusters were mixed (aligned together) during the annotation process. This shows that the kinase segment is less-redundant than LRR at least during our *in silico* gene prediction, once that the kinase CD is present in both *Pto* and *Xa21* genes, they do not caused the mixture of their matching clusters on a mixed class.

The last case of mixture occurred to MIX class II including the motif TIR-NBS-LRR. Two of the three clusters pertaining to this mixed class (EGEQST6001H02.g and EGJECL1208G03.g) were searched at the FOREST database by the genes *RPP5* (TIR-NBS-LRR; Parker *et al.*, 1997) and *RPS5* (NSB-LRR; Noel *et al.*, 1999). The third cluster (EGEZRT3006B12.g) was obtained through search using *RPP5* and *RPS4* (both TIR-NBS-LRR; Gassmann *et al.*, 1999) and *I2* (NBS-LRR; Simmons *et al.*, 1998) queries. We initially supposed that the redundancy was due to the presence of NB-ARC (NBS) conserved motif. However, the first two clusters did not show any motif after *in silico* CD-search and, again, the region that apparently caused the mixture of the classes was the LRR motif, once that it was predicted in cluster EGEZRT3006B12.g.

In view of the results discussed above, could we speculate that *Eucalyptus* bears some new classes of *R*-genes? Before taking further conclusions and in order to solve the questions raised by the present work, we intend to evaluate these groups of clusters in regard to their domain and interdomain structure and organization, evaluating also the clusterization process, before taking further conclusions.

The conserved domains (CDs) identified during our investigation showed that most of the *Eucalyptus* predicted sequences possess the same motifs shared by disease *R*-genes. The CD with the higher level of sampling was LRR, which was present in all classes (except KINASE class I, represented by *Pto*) with a total of 442 occurrences. The other frequent domain shared by *R*-genes, the NB-ARC, was observed in 27 sequences, notably in TIR-NBS-LRR and NBS-LRR predicted clusters. This motif is commonly found in such sequences, and it is proposed that NB-ARC plays a role in activation of downstream effectors (van der Biezen and Jones, 1998) by their sequence similarity to mammalian CED-4 and APAF-1 proteins which are involved in apoptosis (Chinnaiyan *et al.*, 1997). In plants the TIR motif is found only associated to NBS regions of dicotyledones, being possibly absent in monocotyledones (Meyers *et al.*, 1999). In *Eucalyptus* (a eudicot genus of the Myrtaceae family) TIR domains were quite abundant, as expected, being found in 39 clusters (all from TIR-NBS-LRR-class).

Another very common motif present in two classes of disease *R*-genes is the kinase domain. This motif is shared by *Pto* (ser-thre-kinase) and *Xa21* (receptor-kinase) genes, members of the KINASE and LRR-KINASE classes, respectively. We found that all kinase domains found were associated to the classes KINASE, LRR-KINASE and MIX III. As commented here, despite of its conservation, this domain generally does not cause redundancy while searching in databases.

Transmembrane motifs were found only in 44 of all analyzed sequences. Of these clusters five TM were, unexpectedly, found in TIR-NBS-LRR-related sequences (a group of *R*-genes that acts at the intracellular level), while the remaining 19 were as expected LRR or LRR-KINASE-related sequences.

Information regarding the localization of disease resistance proteins in plant cells is still scarce (Martin, 1999). Spatial organization is usually variable among distinct gene classes and tissues affected, and there are no strong evidences in favor of conserved correspondence between R and Avr products spatial occurrence (Bonas and Lahaye, 2002). However, immunocytochemistry approaches allowed the subcellular localization of some Avr and R components (Boyes *et al.*, 1998). Here, we adopted an *in silico* approach which uses neural network-based methods to predict the topology (*i.e.* localization) of protein sequences of the selected clusters. In spite of the large number of predictions obtained, only 11 sequences were defined into RC1 (reliability class 1 $\geq$ 80%), and 53 for RC2 ($\geq$ 60%). Of these significant predictions, we observed that neural network was able to predict the localization of only a small number of proteins (29.62%) compared to the total sample of *Eucalyptus R*-genes. This percentage of representation is much lower than the 80% obtained for plant test sets carried out by Emanuelsson *et al.* (2000) with the same approach. It is important to note that these predictions are based on the N-terminal information available for sequences. Thus, this low number of predictions can be explained by the fact that the FOREST database was obtained from expressed sequence tags, an approach that usually do not include N-termini for many EST generated.

Our *Eucalyptus* transcriptome cDNA sequence analysis revealed that there are 210 clusters with significant alignment to major classes of plant disease *R*-genes. Differentially from the other genomic efforts, as *O. sativa* (Goff *et al.*, 2002) we used a redundant set of well described *R*-genes to screen for RGAs (Resistance Genes Analogs) on FOREST database. This proved to be a very sensitive approach, since best matches in NCBI present sometimes annotation mistakes and we also observed during the present work that

some of the best GenBank matches to *Eucalyptus R*-clusters presented no conclusive description of function. This was also the case also of the first annotation of *Arabidopsis* genome sequences, as pointed out by Meyers *et al.*, (2003). After reannotation of NBS-LRR sequences a total of 56 of the *A. thaliana R*-genes had to be corrected from earlier evaluations on GenBank (Meyers *et al.*, 2003). These results show how important procedures as annotation and detailed evaluation of generated sequences are. These evidences bring to reflections about the strategic design of many genome and transcriptome projects, considering that the data mining is not expensive (normally only fellowships are needed) but still receive few investments from financing agencies, diminishing the final impact of the results.

The comparison of our results regarding the number (and maybe the organization) of identified *Eucalyptus* clusters was mainly with *A. thaliana*, especially due to the lack of open databases for other plant species with EST projects. Many differences considering the here analyzed *R*-related sequences can be explained by using diverse arguments: (i) The larger genome of *Eucalyptus* (e.g. *E. grandis* with 640 Mbp; Myburg *et al.*, 2003) in contrast with the small and "compact" genome of *A. thaliana* (120 Mbp) (ii) The distant taxonomic position: both are dicots, but distantly related families (Brassicaceae and Myrtaceae) and finally (iii) the different levels of complexity: *Eucalyptus* is a wood perennial plant species and *Arabidopsis* is an annual herb. Herbaceous species are often regarded as faster evolving than woody species considering different morphological and genetic aspects (Bennet, 1972, Enrendorfer, 1982, Morawetz 1984, 1986, Bennet and Leitch, 1995, 2000).

Considering these evidences we observed that most of the information regarding *R*-genes available in databases refer to herbaceous (not woody) crop plants (few wild plants), maybe because most identified and sequenced *R*-genes were consequence of mapping approaches that are very time consuming in woody plants and difficult to realize in open pollinated species. The larger number of sequences from *A. thaliana* representing best alignments to *Eucalyptus* does not represent a higher similarity to this plant species, moreover it reflects the large number of sequences of this model plant deposited in GenBank. In our evaluation, only 23 woody species appeared as best matches for the clusters studied, including 22 species from different dicotyledonous families and one Gymnosperm species (*Pinus sylvestris*). This may justify some of the surprising results obtained in the present work and suggest that identification of *R*-genes in a larger number of taxonomic groups may be a very promissory approach to understand the natural evolution of these sequences when not affected by the influence of man. Regarding the actual knowledge of R-gene structure and diversity, some authors suggested that this gene class evolves faster than other genes (Ellis *et al*., 2000b) what should be evaluated in a larger number of taxonomic entities including wild species and also primitive taxa.

Concluding Remarks

Using bioinformatic tools it was possible to identify classify and verify the actually sequenced *R*-genes in *Eucalyptus* transcriptome. No previous sequences of this type could be found in protein or nucleotide databases for this crop. The identified sequences will be valuable resources for the development of markers for molecular breeding and identification of RGAs (resistance gene analogs) in *Eucalyptus* and other related species. The identified clusters constitute also excellent probes for physical mapping of genes in this species, giving support to genetic mapping programs and synteny studies. Considering the size of some clusters, they may also be used for fluorescent *in situ* hybridization (FISH) on *Eucalyptus* chromosomes, helping also in the comparison of different parental species and the respective hybrids.

The present work on *Eucalyptus*, based on FOREST database brought some light to the existing *R*-gene group in this important crop species and also regarding resistance response in higher plants, leading to the following conclusions:

• All five gene classes of *R*-genes with their respective conserved domains are present and expressed in *Eucalyptus*.

• Some new combinations of domains and motifs of *R*-genes may be present in *Eucalyptus* and could represent novel *R*-gene structures, what should be analyzed in detail.

• Despite the lack of libraries from tissues ellicitated by pathogens a high number of *R*-genes was found in different libraries of FOREST project. This may suggest, that the identified clusters are expressed constitutively but also leads to the supposition that a higher number of *R*-genes may be present in *Eucalyptus* under other experimental conditions.

Besides the detailed analysis of different groups of genes and domains we intend to evaluate the expression of the selected clusters in the different libraries of the project. Furthermore, some additional efforts may be necessary to complete some sequences of *R*-genes, especially considering that their size vary between 321 (in case of *Pto*) and 1802 amino-acids (in case of *Xa1* gene) and many identified sequences possibly present incomplete domains.

Further *in silico*, *in vitro* and *in vivo* evaluations of *Eucalyptus* genome may be a very promissory approach. Manipulation of the expression of these genes in economically important woody plant species aiming to improve disease resistance is necessary. Despite of the challenge that this mission may represent, some reports indicate that this strategy is feasible.

Acknowledgements

The present authors thank Ms. David Anderson de Lima Morais and Dr. Valdir Queiroz Balbino for interesting discussions and instructions about some of the pro-

grams and tools used in the present work. To Dr. Reginaldo de Carvalho and Claudete Maria Marques da Silva we thank for valuable technical support. We thank also CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for the concession of a fellowship to the last author (Grant no. 478895/2003).

References

Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Mille W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Ballvora A, Ercolano MR, Weiss J, Meksem K, Bormann CA, Oberhagemann P, Salamini F and Gebhardt C (2002) The *R1* gene for potato resistance to late blight (*Phytophthora infestans*) belongs to the leucine zipper/NBS/LRR class of plant resistance genes. Plant J 30:361-371.

Barber PA, Smith IW and Keane PJ (2003) Foliar diseases of *Eucalyptus* spp. grown for ornamental cut foliage. Austral Plant Pathol 32:109-111.

Bendahmane A, Kohn BA, Dedi C and Baulcombe DC (1995) The coat protein of potato virus X is a strain-specific elicitor of Rx1-mediated virus resistance in potato. Plant J 8:933-941.

Bendahmane A, Querci M, Kanyuka K and Baulcombe DC (2000) *Agrobacterium* transient expression system as a tool for the isolation of disease resistance genes: Application to the Rx2 locus in potato. Plant J 21:73-81.

Benko-Iseppon AM, Winter P, Huettel B, Staginnus C, Muehlbauer FJ and Kahl G (2003) Molecular markers closely linked to fusarium resistance genes in chickpea show significant alignments to pathogenesis-related genes located on *Arabidopsis* chromosomes 1 and 5. Theor Appl Genet 107:379-386.

Bennett MD (1972) Nuclear DNA content and minimum generation time in herbaceous plants. Proc R Soc Lond Bot 181:109-135.

Bennet MD and Leitch IJ (1995) Nuclear DNA Amounts in Angiosperms. Ann Bot 76:113-176.

Bennet MD and Leitch, IJ (2000) Variation in nuclear DNA amount (C-value) in monocots and its significance. In: Wilson KL and Morrison DA (eds) Monocots: Systematics and Evolution. 1st edition. CSIRO Publishers, Sydney, pp 137-146.

Bittner-Eddy PD, Crute IR, Holub EB and Beynon JL (2000) RPP13 is a simple locus in *Arabidopsis thaliana* for alleles that specify downy mildew resistance to different avirulence determinants in *Peronospora parasitica*. Plant J 21:177-88.

Bonas U and Lahaye T (2002) Plant disease resistance triggered by pathogen-derived molecules: Refined models of specific recognition. Curr Opin Microbiol 5:44-50.

Botella MA, Parker JE, Frost LN, Bittner-Eddy PD, Beynon JL, Daniels MJ, Holub EB and Jones JD (1998) Three genes of the *Arabidopsis* RPP1 complex resistance locus recognize distinct *Peronospora parasitica* avirulence determinants. Plant Cell 10:1847-1860.

Boyes DC, Nam J and Dangl JL (1998) The *Arabidopsis thaliana RPM1* disease resistance gene product is a peripheral plasma membrane protein that is degraded coincident with the hypersensitive response. Proc Natl Acad Sci USA 95:15849-15854.

BRACELPA (2004), Associação Brasileira de Celulose e Papel. Brazil. Available from World Wide Web: http://www.bracelpa.org.br, release date 20/March/2004, cited 25/April/2004.

Brommonschenkel SH, Frary A and Tanksley SD (2000) The broad-spectrum tospovirus resistance gene *Sw-5* of tomato is a homolog of the root-knot nematode resistance gene *Mi*. Mol Plant-Microbe Interact 13:1130-38.

Bryan GT, Wu KS, Farrall L, Jia Y, Hershey HP, McAdams SA, Faulk KN, Donaldson GK, Tarchini R and Valent B (2000) A single amino acid difference distinguishes resistant and susceptible alleles of the rice blast resistance gene *Pi-ta*. Plant Cell 12:2033-2046.

Buschges R, Hollricher K, Panstruga R, Simons G, Wolter M, Frijters A, van Daelen R, van der Lee T, Diergaarde P, Groenendijk J, Topsch S, Vos P, Salamini F and Schulze-Lefert P (1997) The barley *Mlo* gene: A novel control element of plant pathogen resistance. Cell 88:695-705.

Chinnaiyan AM, Chaudhary D, O'Rourke K, Koonin E and Dixit M (1997) Role of CED-4 in the activation of CED-3. Nature 388:728-729.

Collins N, Drake J, Ayliffe M, Sun Q, Ellis J, Hulbert S and Pryor T (1999) Molecular characterization of the maize Rp1-D rust resistance haplotype and its mutants. Plant Cell 11:1365-1376.

Dixon MS, Jones JGD, Keddie JS, Thomas CM, Harisson K and Jones JGD (1996) The tomato *Cf-2* disease resistance locus comprises two functional genes encoding leucine-rich repeat proteins. Cell 84:451-459.

Dixon MS, Hatzixanthis K, Jones DA, Harisson K and Jones JGD (1998) The tomato *Cf-5* disease resistance gene and six homologs show pronounced allelic variation in leucine-rich repeat copy number. Plant Cell 10:1915-1925.

Dodds P, Lawrence G and Ellis J (2001) Six amino acid changes confined to the leucine-rich repeat beta-strand/beta-turn motif determine the difference between the P and P2 rust resistance specificities in flax. Plant Cell 13:163-78.

Ehrendorfer F (1982) Speciation patterns in woody angiosperms of tropical origin. In: Barigozzi C (ed) Mechanisms of Speciation. Alan R. Liss. Inc., New York, pp 479-509.

Ellis J and Jones D (1998) Structure and function of proteins controlling strain-specific pathogen resistance in plants. Curr Opin Plant Biol 1:288-293.

Ellis JG, Lawrence GJ, Luck JE and Dodds N (1999) Identification of regions in alleles of the flax rust resistance gene *L* that determines differences in gene-for-gene specificity. Plant Cell 11:495-506.

Ellis J, Dodds P and Pryor T (2000a) The generation of plant disease resistance genes specificities. Trends Plant Sci 5:373-379.

Ellis J, Dodds P and Pryor T (2000b) Structure, function and evolution of plant disease resistance genes. Curr Opin Plant Biol 3:278-284.

Emanuelsson O, Nielsen H, Brunak B and von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300:1005-1016.

Ernst K, Kumar A, Kriseleit D, Kloos DU, Phillips MS and Ganal MW (2002) The broad-spectrum potato cyst nematode resistance gene (Hero) from tomato is the only member of a large gene family of NBS-LRR genes with an unusual amino acid repeat in the LRR region. Plant J 31:127-136.

Flor HH (1956) The complementary genetic systems in flax and flax rust. Adv Genet 8:29-54.

Flor HH (1971) Current status of the gene-for-gene concept. Annu Rev Plant Pathol 9:275-296.

Gassmann W, Hinsch ME and Staskawicz BJ (1999) The *Arabidopsis RPS4* bacterial-resistance gene is a member of the TIR-NBS-LRR family of disease-resistance genes. Plant J 20:265-277.

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A and Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92-100.

Grant MR, Godiard L, Straube E, Ashfield T, Lewald J, Sattler A, Innes RW and Dangl JL (1995) Structure of the *Arabidopsis RPM1* gene enabling dual specificity disease resistance. Science 269:843-846.

Halterman D, Zhou F, Wei F, Wise RP and Schulze-Lefert P (2001) The MLA6 coiled-coil, NBS-LRR protein confers AvrMla6-dependent resistance specificity to *Blumeria graminis* f. sp. *hordei* in barley and wheat. Plant J 3:335-348.

James WC, Teng PS and Nutter FW (1990) Estimated losses of crops from plant pathogens. In: Pimentel D (ed) CRC Handbook of Pest Management, CRC Press, Boca-Raton, pp 15-50.

Johal GS and Briggs SP (1992) Reductase activity encoded by the *HM1* disease resistance gene in maize. Science 158:958-987.

Jones DA, Thomas CM, Hammond-Kosac KE, Balint-Kurti J and Jones JGD (1994) Isolation of the tomato *Cf-9* gene for resistance to *Cladosporium fulvum* by transposon tagging. Science 266:789-793.

Jones DG (2001) Putting the knowledge of plant disease resistance genes to work. Curr Opin Plant Biol 4:281-287.

Jones JB, Stall RE and Bouzar H (1998) Diversity among xanthomonads pathogenic on pepper and tomato. Ann Rev Phytopathol 36:41-58.

Joosten MH, Cozijnsen TJ and De Wit PJ (1994) Host resistance to a fungal tomato pathogen lost by a single base-pair change in an avirulence gene. Nature 367:384-386.

Koczyk G and Chelkowski J (2003) An assessment of the resistance gene analogues of *Oryza sativa* ssp. *japonica*, their presence and structure. Cell Mol Biol Lett 8:963-972.

Koonin EV and Aravind L (2000) The NACHT family - A new group of predicted NTPases implicated in apoptosis and MHC transcription activation. Trends Biochem Sci 25:223-224.

Lawrence GJ, Finnegan EJ, Ayliffe MA and Ellis JG (1995) The *L6* gene for flax rust resistance is related to *Arabidopsis* bacterial resistance gene *RPP2* and tobacco viral gene *N*. Plant Cell 7:1195-1206.

Mafia RG and Alfenas AC (2003) Diferenciação sintomatológica de manchas foliares em *Eucalyptus* spp. causadas por patógenos fúngicos e bacterianos. Fitopatol Bras 28:688-688.

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY and Bryant SH (2002) CDD: A database of conserved domain alignments with links to domain three-dimensional structure. Nucl Acids Res 30:281-283.

Martin GB (1999) Functional analysis of plant disease resistance genes and their downstream effectors. Curr Opin Plant Biol 2:273-279.

Martin GB, de Vicente MC and Tanksley SD (1993) Hight resolution linkage analysis and physical characterization of the Pto bacterial locus in tomato. Mol Plant-Microbe Interact 6:26-34.

Martin GB, Bogdanove AJ and Sessa G (2003) Understanding the functions of plant disease resistance proteins. Annu Rev Plant Physiol Plant Mol Biol 54:23-61.

McDowell JM, Dhandaydham M, Long TA, Aarts MG, Goff S, Holub EB and Dangl JL (1998) Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the RPP8 locus of *Arabidopsis*. Plant Cell 10:1861-1874.

McNabb K (2002) Clonal propagation of *Eucalyptus* in Brazilian nurseries. In: Dumroese RK, Riley LE and Landis TD (eds) National Proceedings: Forest and Conservation Nursery Associations. USDA Forest Service, Rocky Mountain Research Station, Ogden, pp 165-168.

Meyers BC, Chin DB, Shen KA, Sivaramakrishnan S, Lavelle DO, Zhang Z and Michelmore RW (1998) The major resistance gene cluster in lettuce is highly duplicated and spans several megabases. Plant Cell 10:1817-32.

Meyers BC, Diekcman AW, Michelmore RW, Sivaramakrishnan S, Sobral BW and Young ND (1999) Plant disease resistance genes encodes members of an ancient and diverse protein family within the nucleotide-biding superfamily. Plant J 20:317-332.

Meyers BC, Kozik A, Griego A, Kuang H and Michelmore RW (2003) Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. Plant Cell 15:809-834.

Milligan SB, Bodeau J, Yaghoobi J, Kaloshian I, Zabel P and Williamson VM (1998) The root knot nematode resistance gene *Mi* from tomato is a member of the leucine zipper, nucleotide binding, leucine-rich repeat family of plant genes. Plant Cell 10:1307-19.

Mindrinos M, Katagiri F, Yu GL and Ausubel FM (1994) The *A. thaliana* disease resistance gene *RPS2* encodes a protein containing a nucleotide-biding site and leucine-rich repeats. Cell 78:1089-1099.

Morais DAL (2003) Análise bioinformática de genes de resistência a patógenos no genoma da cana-de-açúcar. Master Dissertation, Universidade Federal de Pernambuco, Recife.

Morawetz W (1984) How stable are genomes of tropical woody plants? Heterozygosity in C-banded Karyotypes of *Porcelia* as compared with *Annona* (Annonaceae) and Drymys (Winteraceae). Pl Syst Evol 145:29-39.

Morawetz W (1986) Remarks on karyological differentiation patterns in tropical woody plants. Pl Syst Evol 152:49-100.

Myburg AA, Griffin AR, Sederoff RR and Whetten RW (2003) Comparative genetic linkage maps of *Eucalyptus grandis*, *Eucalyptus globulus* and their F1 hybrid based on a double pseudo-backcross mapping approach. Theor Appl Genet 107:1028-1042.

Noel L, Moores TL, van Der Biezen EA, Parniske M, Daniels MJ, Parker JE and Jones JD (1999) Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of *Arabidopsis*. Plant Cell 11:2099-2112.

Ori N, Eshed Y, Paran I, Presting G, Aviv D, Tanksley S, Zamir D and Fluhr R (1997) The I2C family from the wilt disease resistance locus I2 belongs to the nucleotide binding, leucine-rich repeat superfamily of plant resistance genes. Plant Cell 9:521-532.

Pan Q, Liu YS, Budai-Hadrian O, Sela M, Carmel-Goren L, Zamir D and Fluhr R (2000) Comparative genetics of nucleotide binding site leucine-rich repeat resistance gene homologues in the genomes of two dycotyledons: Tomato and *Arabidopsis*. Genetics 155:309-322.

Parker JE, Coleman MJ, Dean C and Jones JGD (1997) The *Arabidopsis* downy mildew resistance gene *RPP5* shares similarity to the Toll and interleukin-1 receptors with N and L6. Plant Cell 9:879-894.

Piffanelli P, Zhou F, Casais C, Orme J, Jarosch B, Schaffrath U, Collins NC, Panstruga R and Schulze-Lefert P (2002) The barley MLO modulator of defense and cell death is responsive to biotic and abiotic stress stimuli. Plant Physiol 129:1076-1085.

Richly E, Kurth J and Leister D (2002) Mode of amplification and reorganization of resistance genes during recent *Arabidopsis thaliana* evolution. Mol Biol Evol 19:76-84.

Richter TE and Ronald PC (2000) The evolution of disease resistance genes. Plant Mol Biol 42:195-204.

Romeis T (2001) Protein kinases in the plant defense response. Curr Opin Plant Biol 4:407-414.

Rommens CM and Kishore GM (2000) Exploiting the full potential of disease resistance genes for agricultural use. Curr Opin Biotechnol 11:120-125.

Ronald PC (1997) The molecular basis of disease resistance in rice. Plant Mol Biol 35:179-186.

Simmons G, Groenendijk J, Wijbrandi J, Reijans M, Groenen J, Diergaarde van der Lee T, Bleeker M, Onstenk J, De Both M, Haring M, Mes J, Cornelissen B, Zabeau M and Vos P (1998) Dissection of the fusarium *I2* gene cluster in tomato reveals six homologs and one active gene copy. Plant Cell 10:1055-1068.

Song WY, Pi LY, Wang GL, Gardner J, Holsten T and Ronald PC (1997) Evolution of the rice *Xa21* disease resistance genes family. Plant Cell 9:1279-1287.

Song WY, Wang GL, Chen LL, Kim HS, Pi LY, Holsten T, Gardner J, Wang B, Zhai WX, Zhu LH, Fauquet C and Ronald PC (1995) A receptor kinase-like protein encoded by the rice disease resistance gene *Xa21*. Science 270:1804-1806.

The *Arabidopsis* Genome Iniciative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796-815.

Thomas CM, Jones DA, Parniske M, Harrison K, Balint-Kurti PJ, Hatzixanthis K and Jones JD (1997) Characterization of the tomato *Cf-4* gene for resistance to *Cladosporium fulvum* identifies sequences that determine recognitional specificity in *Cf-4* and *Cf-9*. Plant Cell 9:2209-2224.

van der Biezen EA and Jones JGD (1998) The NB-ARC domais: A novel signaling motif shared by plant resistance gene products and regulators of cell death in animals. Curr Biol 8:R226-R227.

van der Biezen EA, Freddie CT, Kahn K, Parker JE and Jones JD (2002) *Arabidopsis RPP4* is a member of the *RPP5* multigene family of TIR-NB-LRR genes and confers downy mildew resistance through multiple signaling components. Plant J 29:439-51.

van der Vossen EA, van der Voort JN, Kanyuka K, Bendahmane A, Sandbrink H, Baulcombe DC, Bakker J, Stiekema WJ and Klein-Lankhorst RM (2000) Homologues of a single resistance-gene cluster in potato confer resistance to distinct pathogens: A virus and a nematode. Plant J 23:567-576.

Wang ZX, Yano M, Yamanouchi U, Iwamoto M, Monna L, Hayasaka H, Katayose Y and Sasaki T (1999) The *Pib* gene for rice blast resistance belongs to the nucleotide binding and leucine-rich repeat class of plant disease resistance genes. Plant J 19:55-64.

Whithan S, McCormick S and Baker B (1996) The *N* gene of tobacco confers resistance to tobacco mosaic virus in transgenic tomato. Proc Natl Acad Sci USA 93:8776-81.

Xiao S, Ellwood S, Calis O, Patrick E, Li T, Coleman M and Turner JG (2001) Broad-spectrum mildew resistance in *Arabidopsis thaliana* mediated by RPW8. Science 291:118-20.

Yoshimura S, Yamanouchi U, Katayose Y, Toki S, Wang ZX, Kono I, Kurata N, Iwata N and Sasaki T (1998) Expression of *Xa1*, a bacterial blight-resistance gene in rice, is induced by bacterial inoculation. Proc Natl Acad Sci USA 95:1663-1668.

Associate Editor: Marcio de Castro Silva Filho

6. CONSIDERAÇÕES FINAIS

O desenvolvimento de ferramentas computacionais para tratar dados biológicos e facilitar a extração de novos conhecimentos a partir dos dados gerados obteve imensa importância nos últimos anos. Nesta tese procuramos seguir essa abordagem, desenvolvendo ferramentas e componentes de softwares que pudessem ajudar ao desenvolvimento da base de dados proposta inicialmente, a Plant Defense Mechanisms Database.

Através da ferramenta de mineração de texto LAITOR, distribuímos um recurso adicional que pode ser aplicado não apenas na análise de dados de plantas, mas sim aos diversos tipos de dados biológicos mencionados na literatura, sendo, para isso, apenas necessária a adaptação dos dicionários utilizados pelo programa.

O método Seed Linkage é um método inédito no tocante ao estabelecimento de clusters de proteínas relacionadas utilizando regras de melhor hit cruzado em genomas completos ou incompletos a partir de uma única seqüência fundadora, o que faz com que apenas clusters de uma determinada proteína de interesse seja considerado, ideal para a rápida identificação e expansão de proteínas relacionadas a mecanismos específicos, sem ter, para isso, que considerar todos os genomas envolvidos.

Para a obtenção de informações pertinentes a cada uma das seqüências agrupadas nos clusters acima descritos, nós desenvolvemos uma biblioteca que permite a atuação de scripts em PHP como clientes SOAP do servidor SRS instalado no EMBL, nosso colaborador nesse projeto.

Os três recursos acima foram possíveis graças à bolsa de doutorado sanduíche concedida ao aluno Adriano Barbosa pela CAPES. Durante a estadia no EMBL, foi possível que o aluno fosse treinado no tocantes as tecnologias de aquisição de dados via Web Services usando o protocolo SOAP, assim como o contato com os métodos de mineração de texto empregados pelo grupo do professor Reinhard Schneider, co-orientador desta tese. Além disso, Adriano pode desenvolver sua metodologia de agrupamento de seqüências nomeada Seed Linkage, a partir de diversas discussões entre os dois grupos de pesquisas: o Laboratório de Biodados na UFMG e o *Data Integration and Management Group* no EMBL, o que sem dúvida fortaleceu os laços entre estas duas instituições de ponta e pesquisa no Brasil e na Alemanha. Atualmente adriano mantém contato com diversos pesquisadores do EMBL, que por seu caráter dinâmico, em breve, novos grupos serão liderados por esses contatos o que viabilizará o intercâmbio científico entre os grupos a serem originados e o Brasil.

A base de dados PDM, tema inicial do projeto de doutorado de Adriano, está ainda em pleno desenvolvimento, porém as informações iniciais depositadas na base de dados foram frutos dos diversos métodos desenvolvidos e apresentados neste trabalho. Esperamos aprimorar o conteúdo da base de dados utilizando os métodos já descritos bem como os novos recursos a serem desenvolvidos daqui pra frente.

7. REFERÊNCIAS BIBLIOGRÁFICAS

Achard, F., Vaysseix, G., and Barillot, E. (2001). XML, bioinformatics and data integration. Bioinformatics (Oxford, England) **17,** 115-125.

Alexeyenko, A., Tamas, I., Liu, G., and Sonnhammer, E.L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics (Oxford, England) **22,** e9-15.

Altschul, S.F. (1998). Fundamentals of database searching. Trends Guide to Bioinformatics**,** 7-9.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecular biology **215,** 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research **25,** 3389-3402.

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. Science (New York, N.Y **181,** 223-230.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., and Yeh, L.S. (2004). UniProt: the Universal Protein knowledgebase. Nucleic acids research **32,** D115-119.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics **25,** 25-29.

Bairoch, A., and Apweiler, R. (1996). The SWISS-PROT protein sequence data bank and its new supplement TREMBL. Nucleic acids research **24,** 21-25.

Bajic, V.B., Veronika, M., Veladandi, P.S., Meka, A., Heng, M.W., Rajaraman, K., Pan, H., and Swarup, S. (2005). Dragon Plant Biology Explorer. A text-mining tool for integrating associations between genetic and biochemical entities with genome annotation and biochemical terms lists. Plant physiology **138,** 1914-1925.

Barbosa-Silva, A., Pafilis, E., Ortega, J.M., and Schneider, R. (2007). Development of SRS.php, a Simple Object Access Protocol-based library for data acquisition from integrated biological databases. Genet Mol Res **6,** 1142-1150.

Barbosa-Silva, A., Satagopam, V.P., Schneider, R., and Ortega, J.M. (2008). Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence. BMC Bioinformatics **9,** 141.

Boguski, M.S. (1998). Bioinformatics - a new era. Trends Guide Bioinformatics**,** 1-3.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A. (2007). UniProtKB/Swiss-Prot: The Manually Annotated Section of the UniProt KnowledgeBase. Methods in molecular biology (Clifton, N.J **406,** 89-112.

Buckingham, S.D. (2005). Data mining for protein-protein interactions in invertebrate model organisms. Invert Neurosci **5,** 183-187.

Chang, J.T., Schutze, H., and Altman, R.B. (2004). GAPSCORE: finding gene and protein names one word at a time. Bioinformatics (Oxford, England) **20,** 216-225.

Chen, F., Mackey, A.J., Vermunt, J.K., and Roos, D.S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE **2,** e383.

Chen, L., Liu, H., and Friedman, C. (2005). Gene name ambiguity of eukaryotic nomenclatures. Bioinformatics (Oxford, England) **21,** 248-256.

Chervitz, S.A., Aravind, L., Sherlock, G., Ball, C.A., Koonin, E.V., Dwight, S.S., Harris, M.A., Dolinski, K., Mohr, S., Smith, T., Weng, S., Cherry, J.M., and Botstein, D. (1998). Comparison of the complete protein sets of worm and yeast: orthology and divergence. Science (New York, N.Y **282,** 2022-2028.

Cohen, A.M., Hersh, W.R., Dubay, C., and Spackman, K. (2005). Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. BMC Bioinformatics **6,** 103.

Craven, M., and Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. Proceedings /. International Conference on Intelligent Systems for Molecular Biology; ISMB**,** 77-86.

Crim, J., McDonald, R., and Pereira, F. (2005). Automatically annotating documents with normalized gene lists. BMC Bioinformatics **6 Suppl 1,** S13.

Davidson, S.B., Overton, C., and Buneman, P. (1995). Challenges in integrating biological data sources. J Comput Biol **2,** 557-572.

Dayhoff, M.O. (1965a). Computer aids to protein sequence determination. Journal of theoretical biology **8,** 97-112.

Dayhoff, M.O. (1969). Computer analysis of protein evolution. Scientific American **221,** 86-95.

Dayhoff, M.O. (1974). Computer analysis of protein sequences. Federation proceedings **33,** 2314-2316.

Dayhoff, M.O. (1978). Atlas of Protein Sequence and Structure. (National Biomedical Research Foundation, Washington).

Dayhoff, M.O., Eck,R.V., Chang,M.A., Sochard,M.R. (1965b). Atlas of Protein Sequence and Structure. (National Biomedical Research Foundation, Silver Spring, Maryland).

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic acids research **30,** 1575-1584.

Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. Systematic zoology **19,** 99-113.

Fitch, W.M., and Margoliash, E. (1967). Construction of phylogenetic trees. Science (New York, N.Y **155,** 279-284.

Fukuda, K., Tamura, A., Tsunoda, T., and Takagi, T. (1998). Toward information extraction: identifying protein names from biological papers. Pacific Symposium on Biocomputing**,** 707-718.

Fynbo, J.P., Watson, D., Thone, C.C., Sollerman, J., Bloom, J.S., Davis, T.M., Hjorth, J., Jakobsson, P., Jorgensen, U.G., Graham, J.F., Fruchter, A.S., Bersier, D., Kewley, L., Cassan, A., Ceron, J.M., Foley, S., Gorosabel, J., Hinse, T.C., Horne, K.D., Jensen, B.L., Klose, S., Kocevski, D., Marquette, J.B., Perley, D., Ramirez-Ruiz, E., Stritzinger, M.D., Vreeswijk, P.M., Wijers, R.A., Woller, K.G., Xu, D., and Zub, M. (2006). No supernovae associated with two long-duration gamma-ray bursts. Nature **444,** 1047-1049.

Gerlt, J.A., and Babbitt, P.C. (2000). Can sequence determine function? Genome biology **1,** REVIEWS0005.

Gonzalez, G., Uribe, J.C., Tari, L., Brophy, C., and Baral, C. (2007). Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures. Pacific Symposium on Biocomputing**,** 28-39.

Haferkamp, I., Hackstein, J.H., Voncken, F.G., Schmit, G., and Tjaden, J. (2002). Functional integration of mitochondrial and hydrogenosomal ADP/ATP carriers in the Escherichia coli membrane reveals different biochemical characteristics for plants, mammals and anaerobic chytrids. Eur J Biochem **269,** 3172-3181.

Hagen, J.B. (2000). The origins of bioinformatics. Nature reviews **1,** 231-236.

Hale, R. (2005). Text mining: getting more value from literature resources. DDT **10,** 377-379.

Hao, Y., Zhu, X., Huang, M., and Li, M. (2005). Discovering patterns to extract protein-protein interactions from the literature: Part II. Bioinformatics (Oxford, England) **21,** 3294-3300.

Hirschman, L., Park, J.C., Tsujii, J., Wong, L., and Wu, C.H. (2002). Accomplishments and challenges in literature data mining for biology. Bioinformatics (Oxford, England) **18,** 1553-1561.

Hirsh, A.E., and Fraser, H.B. (2001). Protein dispensability and rate of evolution. Nature **411,** 1046-1049.

Hoffmann, R., and Valencia, A. (2005). Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics **21 Suppl 2,** ii252-258.

Hunt, L.T. (1983). Margaret O. Dayhoff 1925-1983. DNA (Mary Ann Liebert, Inc **2,** 97-98.

Hunter, L., and Cohen, K.B. (2006). Biomedical language processing: what's beyond PubMed? Molecular cell **21,** 589-594.

Ilic, K., Kellogg, E.A., Jaiswal, P., Zapata, F., Stevens, P.F., Vincent, L.P., Avraham, S., Reiser, L., Pujar, A., Sachs, M.M., Whitman, N.T., McCouch, S.R., Schaeffer, M.L., Ware, D.H., Stein, L.D., and Rhee, S.Y. (2007). The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant. Plant physiology **143,** 587-599.

Jensen, L.J., Saric, J., and Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. Nature reviews **7,** 119-129.

Jensen, R.A. (2001). Orthologs and paralogs - we need to get it right. Genome biology **2,** INTERACTIONS1002.

Karlin, S., and Altschul, S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A **87,** 2264-2268.

Krallinger, M., and Valencia, A. (2005). Text-mining and information-retrieval services for molecular biology. Genome biology **6,** 224.

Krallinger, M., Padron, M., and Valencia, A. (2005). A sentence sliding window approach to extract protein annotations from biomedical articles. BMC bioinformatics **6 Suppl 1,** S19.

Lee, Y., Sultana, R., Pertea, G., Cho, J., Karamycheva, S., Tsai, J., Parvizi, B., Cheung, F., Antonescu, V., White, J., Holt, I., Liang, F., and Quackenbush, J. (2002). Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). Genome research **12,** 493-502.

Leinonen, R., Diez, F.G., Binns, D., Fleischmann, W., Lopez, R., and Apweiler, R. (2004). UniProt archive. Bioinformatics (Oxford, England) **20,** 3236-3237.

Leonard, T., and Duffy, J.C. (2002). A Bayesian fixed effects analysis of the Mantel-Haenszel model applied to meta-analysis. Stat Med **21,** 2295-2312.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research **13,** 2178-2189.

Libbus, B., and Rindflesch, T.C. (2002). NLP-based information extraction for managing the molecular biology literature. Proceedings / AMIA. Annual Symposium**,** 445-449.

Marcotte, E.M., Xenarios, I., and Eisenberg, D. (2001). Mining literature for protein-protein interactions. Bioinformatics (Oxford, England) **17,** 359-363.

Markowitz, V.M. (1995). Heterogeneous molecular biology databases. J Comput Biol **2,** 537-538.

Mika, S., and Rost, B. (2004). NLProt: extracting protein names and sequences from papers. Nucleic Acids Res **32,** W634-637.

Mons, B. (2005). Which gene did you mean? BMC Bioinformatics **6,** 142.

Mushegian, A.R., Garey, J.R., Martin, J., and Liu, L.X. (1998). Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. Genome research **8,** 590-598.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology **48,** 443-453.

Neerincx, P.B., and Leunissen, J.A. (2005). Evolution of web services in bioinformatics. Briefings in bioinformatics **6,** 178-188.

Nobelprize.org (2010). "The Nobel Prize in Chemistry 1958". (URL: http://nobelprize.org/nobel_prizes/chemistry/laureates/1958/). Acessado em 21/11/2010.

Olby, R.C. (1985). The 'mad pursuit': X-ray crystallographers' search for the structure of haemoglobin. History and philosophy of the life sciences **7,** 171-193.

Pahikkala, T., Ginter, F., Boberg, J., Jarvinen, J., and Salakoski, T. (2005). Contextual weighting for Support Vector Machines in literature mining: an application to gene versus protein name disambiguation. BMC bioinformatics **6,** 157.

Pearson, W.R. (1995). Comparison of methods for searching protein sequence databases. Protein Sci **4,** 1145-1160.

Perutz, M. (1985). Early days of protein crystallography. Methods in enzymology **114,** 3-18.

Remm, M., Storm, C.E., and Sonnhammer, E.L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Journal of molecular biology **314,** 1041-1052.

Rindflesch, T.C., Tanabe, L., Weinstein, J.N., and Hunter, L. (2000). EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pacific Symposium on Biocomputing**,** 517-528.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., Cherry, J.M., Henikoff, S., Skupski, M.P., Misra, S., Ashburner, M., Birney, E., Boguski, M.S., Brody, T., Brokstein, P., Celniker, S.E., Chervitz, S.A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R.F., Gelbart, W.M., George, R.A., Goldstein, L.S., Gong, F., Guan, P., Harris, N.L., Hay, B.A., Hoskins, R.A., Li, J., Li, Z., Hynes, R.O., Jones, S.J., Kuehl, P.M., Lemaitre, B., Littleton, J.T., Morrison, D.K., Mungall, C., O'Farrell, P.H., Pickeral, O.K., Shue, C., Vosshall, L.B., Zhang, J., Zhao, Q., Zheng, X.H., and Lewis, S. (2000). Comparative genomics of the eukaryotes. Science (New York, N.Y **287,** 2204-2215.

Sanger, F. (1959). Chemistry of insulin; determination of the structure of insulin opens the way to greater understanding of life processes. Science (New York, N.Y **129,** 1340-1344.

Schijvenaars, B.J., Mons, B., Weeber, M., Schuemie, M.J., van Mulligen, E.M., Wain, H.M., and Kors, J.A. (2005). Thesaurus-based disambiguation of gene symbols. BMC bioinformatics **6,** 149.

Schonbach, C., Kowalski-Saunders, P., and Brusic, V. (2000). Data warehousing in molecular biology. Briefings in bioinformatics **1,** 190-198.

Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. Journal of molecular biology **147,** 195-197.

Sonnhammer, E.L., and Koonin, E.V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet **18,** 619-620.

Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A., and Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res **26,** 320-322.

Srinivasan, P., and Hristovski, D. (2004). Distilling conceptual connections from MeSH co-occurrences. Medinfo **11,** 808-812.

Stein, L.D. (2003). Integrating biological databases. Nat Rev Genet **4,** 337-345.

Storm, C.E., and Sonnhammer, E.L. (2002). Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics (Oxford, England) **18,** 92-99.

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics (Oxford, England) **23,** 1282-1288.

Tanabe, L., and Wilbur, W.J. (2002). Tagging gene and protein names in biomedical text. Bioinformatics (Oxford, England) **18,** 1124-1132.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997). A genomic perspective on protein families. Science (New York, N.Y **278,** 631-637.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic acids research **28,** 33-36.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., and Natale, D.A. (2003). The COG database: an updated version includes eukaryotes. BMC bioinformatics **4,** 41.

UDDI. (2008). Universal Description, Discovery and Integration. Acessado em: 12/03/2008.

Venkatesh, T.V., and Harlow, H.B. (2002). Integromics: challenges in data integration. Genome biology **3,** REPORTS4027.

W3C. (2008a). SOAP specifications (URL: http://www.w3.org/TR/soap). Acessado em: 12/03/2008.

W3C. (2008b). Web Services Description Language (URL: http://www.w3.org/TR/wsdl). Acessado em: 12/03/2008.

Wall, D.P., Fraser, H.B., and Hirsh, A.E. (2003). Detecting putative orthologs. Bioinformatics (Oxford, England) **19,** 1710-1711.

Wheelan, S.J., Boguski, M.S., Duret, L., and Makalowski, W. (1999). Human and nematode orthologs--lessons from the analysis of 1800 human genes and the proteome of Caenorhabditis elegans. Gene **238,** 163-170.

Wilbur, W.J., and Yang, Y. (1996). An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. Computers in biology and medicine **26,** 209-222.

Yoo, I., Hu, X., and Song, I.Y. (2007). Biomedical ontology improves biomedical literature clustering performance: a comparison study. Int J Bioinform Res Appl **3,** 414-428.

Zhou, G., Shen, D., Zhang, J., Su, J., and Tan, S. (2005). Recognition of protein/gene names from text using an ensemble of classifiers. BMC Bioinformatics **6 Suppl 1,** S7.

Zmasek, C.M., and Eddy, S.R. (2002). RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC bioinformatics **3,** 14.

8. PRODUÇÃO CIENTÍFICA DURANTE O DOUTORADO

* Trabalhos diretamente relacionados ao tema da tese encontram-se iniciados por asterisco.

8.1 ARTIGOS CIENTÍFICOS PUBLICADOS EM REVISTAS INTERNACIONAIS

*** Barbosa-Silva, A.**, Satagopam, V. P., Schneider, R., Ortega, J. M. (2008) Clustering of cognate proteins among distinct proteomes derived from multiple links to a single seed sequence. BMC Bioinformatics 9:141.

Nascimento, D.G., Rates, B., Santos, D.M., **Barbosa-Silva, A.**, Verano-Braga, T., Dutra, A.A.A., Biondi, I., Martin-Eauclaire, M.F., De lima, M.A., Pimenta, A.M.C. (2006) Moving pieces in a taxonomic puzzle: venom 2D-LC/MS and data clustering analyses to infer phylogenetic relationships in some scorpions from the Buthidae family (Scorpiones). Toxicon. 47(6):628-39.

Cestari, I. C., Haver, N. J., **Barbosa-Silva, A.**, Ramirez, M. I. (2006) PROTOGIM: A novel tool to search motifs and domains in hypothetical proteins of protozoan genomes. Parasitology Research 98(4):375-377.

8.2 ARTIGOS CIENTÍFICOS PUBLICADOS EM REVISTAS NACIONAIS

*** Barbosa-Silva, A.**, Ortega, J. M., Pafilis, E., Schneider, R. (2007) Development of SRS.php, a SOAP-based library for data acquisition from integrated biological databases. Genetics and Molecular Research 6(4): 1142-1150.

Wanderley-Nogueira, A. C., Soares-Cavalcanti, N. M., Morais, D. A. L., Berlarmino, L. C., **Barbosa-Silva, A.**, Benko-Iseppon, A. M. (2007) Abundance and Diversity of Resistance (R) Genes in the Sugarcane Transcriptome. Genetics and Molecular Research 6 (4): 866-889.

Barbosa-da-Silva, A., Wanderley-Nogueira, A. C., Silva, R. M. R., Berlarmino, L. C., Soares-Cavalcanti, N. M., Benko-Iseppon, A. M. (2005) In silico survey of resistance (R) genes in Eucalyptus transcriptome. Genetics and Molecular Biology 28(3):562-574.

8.3 ARTIGOS ACEITOS PARA PUBLICAÇÃO

Barbosa, D., Fernandes G., Prosdocimi, F., Pena, I., Santos, L., Mudado, M., Natale, D., Ortega, J. M., Junior, O. C., **Barbosa-Silva, A.**, Velloso, H. M.,. Aguiar, S. C. V. A procedure to recruit members to enlarge Protein Families Databases - the building of UECOG (UniRef-Enriched COG Database) as a model. Genetics and Molecular Research.

8.4 TRABALHOS APRESENTADOS EM CONGRESSOS

* **Barbosa-Silva, A**., Mudado, M., Ortega, J. M. (2005). Plant Defense Mechanisms Database (PDM): Building and Evaluation. In: Anals of the 1st International Conference of The Brazilian Association of Bioinformatics and Computational Biology (X- Meeting). Caxambu – MG.

* **Barbosa, A.**, Mudado, M. A., Ortega, J. M. (2005) PDM, A Plant Defense Mechanism Database Suitable For Automatic Annotation and Classification. In: Anais do 51º Congresso Nacional de Genética. Águas de Lindóia – SP.

Barbosa, A., Benko-Iseppon, A. M. (2004). In silico screening of tomato disease resistance pathway in sugarcane transcriptome. In: II ICOBICOBI - International Conference on Bioinformatics and Computational Biology. Angra dos Reis - RJ.

FERREIRA, R. N. ; RATES, B. ; MELO, M. N. ; CISCOTTO, P. H. C. ; **BARBOSA-SILVA, A.** ; SANCHEZ, E. F. ; DE LIMA, M. E. ; PIMENTA, A. M. C. Venomic analyses from Bothrops species: a new approach to determine taxonomical/phylogenetic relationships based in venom complexities.. In: IX Congresso da Sociedade Brasileira de Toxinologia, 2006, Fortaleza, CE. IX Congresso da Sociedade Brasileira de Toxinologia, 2006. v. 1.

NASCIMENTO, D. G. ; RATES, B. ; SANTOS, D. M. ; Verano-Braga, T ; **BARBOSA-SILVA, A.** ; DUTRA, A.A.A ; BIONDI, I. ; MARTIN-EAUCLAIRE, M. F. ; DE LIMA, M. E. ; PIMENTA, A. M. C. Moving pieces in a taxonomic puzzle: venom 2D-LC/MS anda data clustering analysis to infer phylogenetic relationships in some scorpions from the buthidae family (scorpiones). In: IX Congresso da Sociedade Brasileira de Toxinologia, 2006, Fortaleza, CE. IX Congresso da Sociedade Brasileira de Toxinologia, 2006. v. 1.

FERREIRA, R. N. ; RATES, B. ; CISCOTTO, P. H. C. ; MELO, M. N. ; **BARBOSA-SILVA, A.** ; DE LIMA, M. E. ; SANCHEZ, E. F. ; PIMENTA, A. M. C. . Venomic analyses of Bothrops species: a new approach to determine taxonomical/phylogenetic relationships based in venom complexities..

In: XXXV Reunião Anual da SBBq, 2006, Águas de Lindóia. Anais da XXXV Reunião Anual da SBBq, 2006. v. 1.

NASCIMENTO, D. G. ; RATES, B. ; SANTOS, D. M. ; Verano-Braga, T ; **BARBOSA-SILVA, A.** ; DUTRA, A.A.A ; BIONDI, I. ; MARTIN-EAUCLAIRE, M. F. ; DE LIMA, M. E. ; PIMENTA, A. M. C. . Moving pieces in a taxonomic puzzle: Venom 2D-LC/MS and data clustering analyses to infer phylogenetic relationships in some scorpions from the Buthidae family (Scorpiones).. In: XXXV Reunião Anual da SBBq, 2006, Águas de Lindóia, SP. Anais da XXXV Reunião Anual da SBBq, 2006. v. 1.

Rates, B., Nascimento, D.G., Santos, D.M., Verano-Braga, T., **Barbosa-Silva, A.**, Dutra, A.A.A., Biondi, I., Martin-Eauclaire, M.F., Lima, M.E., Pimenta, A.M.C. (2005). Phylogenetic inferences of Buthidae scorpions by venom 2D-LC/MS. In: Anals of 1st Brazilian Congress on Mass Spectrometry (BrMASS). Campinas, SP, Brazil.

Nascimento, D.G., Rates, B., Santos, D.M., Verano-Braga, T., Dutra, A.A.A., **Barbosa-Silva, A.**, Lima, M.E., Pimenta, A.M.C. (2005) Moving pieces in a taxonomic puzzle: venom 2D-LC/MS and data clustering analyses to infer phylogenetic relationships in some scorpions from the Buthidae family (Scorpiones). In: VIII Encontro de pesquisa do Instituto de Ciências Biológicas/ III Encontro de Anual Pesquisa em Bioquímica e Imunologia, 2005, Belo Horizonte - MG, Brazil.

Rates, B. **Barbosa-Silva, A.**, Nascimento, D. G., Santos, D. M., Braga, T. V., Dutra, A. A. A., Lima, M. E., Pimenta, A. M. C. (2005). Machine learning-based clustering analysis of venom 2D-LC/MS data to infer phylogenetic relationships in scorpions from the Buthidae family. In: Anals of the 1st International Conference of The Brazilian Association of Bioinformatics and Computational Biology (X- Meeting). Caxambu - MG.

Andrade, P.P., Andrade, C.R., Araújo, D.A.M., Balbino, V.Q., Barbosa, M.H.N., **Barbosa-da-Silva, A.** , et al. (2006) Survey of Leishmania chagasi transcriptome. In: Internatioanl Conference On Intelligent System for Molecular Biology. Fortaleza - CE, Brazil.

Casanova, F. M., Andrade, P. P., Kido, E. A., Cavalcanti, E. S., Rayol, C. A., **Barbosa, A.** (2005). Patterns of gene expression from Leishmania ProGeNE cDNA libraries. In: Anals of the 1st International Conference of The Brazilian Association of Bioinformatics and Computational Biology (X- Meeting). Caxambu – MG.

Barbosa, D. V. C., Faria-Campos, A. C., **Barbosa, A.**, Ortega, J. M. (2005). Identification of the divergent genes in group of paralogs from microorganisms with complete genome. In: Anals of the

1st International Conference of The Brazilian Association of Bioinformatics and Computational Biology (X- Meeting). Caxambu – MG.

Mudado, M. A., Pinto, S. A. P., **Barbosa, A. S.**, et al. (2004). A Picture of the Sampling of Gene Expression In Model Metazoa Using EST And KOG Proteins. In: II International Conference on Bioinformatics and Computational Biology. Angra dos Reis – RJ.

Bevitori, R., Lopes, D., **Barbosa, A.**, et al. (2004). Generation and analysis of Expressed Sequence Tags from Oryza sativa under aluminum stress.In: PLANT GENOMICS EUROPEAN MEETING., Lyon. Proceedings of the Plant Genomics European Meeting., 2004. v. 3. p. 235