

FEDERAL UNIVERSITY OF MINAS GERAIS
INSTITUTE OF BIOLOGICAL SCIENCES
DEPARTMENT OF GENERAL BIOLOGY
GRADUATE PROGRAM IN GENETICS



PhD THESIS

**Modelome Derived Intra-Species Broad Spectrum Drug and
Vaccine Targets Identification in the Animal Pathogen
*Corynebacterium pseudotuberculosis***

PHD STUDENT: **Syed Shah Hassan**

SUPERVISOR: **Prof. Dr. Vasco Ariston de Carvalho Azevedo**

CO-SUPERVISOR: **Profa. Dra. Rafaela Salgado Ferreira**

BELO HORIZONTE

March – 2013

Syed Shah Hassan

**Modelome Derived Intra-Species Broad Spectrum Drug and
Vaccine Targets Identification in the Animal Pathogen
*Corynebacterium pseudotuberculosis***

Thesis presented to the Post-Graduation Program in Genetics, Department of General Biology, Institute of Biological Sciences, Federal University of Minas Gerais as a partial requirement for obtaining the degree of Doctor of Philosophy in Genetics.

PhD. Student: **Syed Shah Hassan**

Supervisor: **Prof. Dr. Vasco Ariston de Carvalho Azevedo**

Co-supervisor: **Profa. Dra. Rafaela Salgado Ferreira**

FEDERAL UNIVERSITY OF MINAS GERAIS

INSTITUTE OF BIOLOGICAL SCIENCES

BELO HORIZONTE - MG

March – 2013

*In the name of Allah, the Beneficent, the Merciful,
"He it is, Who fashioneth you in the wombs as pleaseth Him.
There is no God, but Him, the Almighty, the Wise.
(Al-Quran 3:6)*

DEDICATION

Although the golden principles of honesty, commitment, effort and dedication were the basic elements for the accomplishment of this doctoral dissertation, yet I dedicate this work to God (the Almighty Allah (SWT) for giving me this wonderful opportunity to have been my Lord, my friend, my hope, and indeed everything to me, who has made possible this task for me. I dedicate the work to my Family members and friends, especially to my beloved parents and my late brother Syed Shah Hussain (may his soul rest in Peace, Amen), who are always been my sources of inspiration, courage and moral support throughout my academic career. I love you all. Finally, dedicated to the whole humanity, who for some reasons have not had enough resources and access to be enlightened by the great power of knowledge.

Eventually, appreciativeness's to the solitary inspiration persona,
Dr. A. Q. Khan

ACKNOWLEDGEMENTS

Primarily many thank to The Almighty Allah because He is the creator, beneficent and kind.

Cordially, all sorts of respects and regards to the last and true messenger of God, the Holly Prophet Muhammad PBUH for his continuation of guiding the mankind to the right path of life, by bringing them back under the light of Islam, to make them closer and respectable to their Lord, their Divine, the Almighty Allah.

Thank to my parents, for their eminent efforts to have risen, educated and guided me on the right paths, and much gratitude's to them for I got the position where I am.

I am speechless to mention the sort of kindness, efforts, commitment, wisdom, understanding of my supervisor Prof. Dr. Vasco Azevedo Ariston de Carvalho. His friendship and all academic and moral support, and encouragement, is incredible.

To co-supervisor Prof. Dr. Rafaela Salgado Ferreira, Prof. Dr. Anderson Miyoshi and Prof. Dr. Artur Silva for their full-time availability, suggestions and positive criticism and guidance during understanding and development of this work.

To Prof. Dr. Eduardo M. Tarazona Santos (Coordinator), and Prof. Dr. Ana Lúcia Brunialti Godard (Sub-coordinator) for providing me academic and financial assistance in for attending in and outdoor seminars and courses.

To Sheila Santana, Marcia Natalia Dutra Ramalho e Silva, Mary das Graças Santos (Office Secretaries) who have always been kind enough in resolving academic and administrative issues.

Gratitude to all dear colleagues from the Laboratory of Cellular and Molecular Genetics (LGCM) and LPDNA-UFPB, by making exceptional support and guidance in various parts of this work.

Thanks to all friends in Brazil and abroad who have always supported and encouraged my decisions and always illuminated me with precious advises. Love you all !!!

I would like to thank the Federal University of Minas Gerais by excellent training, the Academy of Sciences for Third World Countries (T'WAS) for scientific career making opportunity and the National Council for Scientific and Technological Development (CNPq), by indispensable financial support toward all costs of my doctoral studies.

Affectionately thanks to everyone in the scientific circle.

CONTENTS

LIST OF ABBREVIATIONS AND ACRONYMS.....	IX
LIST OF FIGURES.....	XII
Abstract.....	13
Resumo.....	14
I. PRESENTATION.....	15
I.I Collaborations.....	16
I.II Manuscript's Delineation.....	18
I.III Justifications.....	19
II. OBJECTIVES.....	20
II.I General Objectives.....	21
II.II Specific Objectives.....	21
INTRODUCTION 1	
<i>Corynebacterium pseudotuberculosis</i>, Caseous lymphadenitis (CLA) and Genomic Sequencing.....	22
1.1 The Species <i>Corynebacterium pseudotuberculosis</i>	23
1.2 Caseous Lymphadenitis (CLA): Transmissions and Laboratory Diagnosis	24
1.3 Mechanism of Bacterial Infection/Pathogenesis.....	25
1.4 Treatment, Control Measures and Vaccine Development.....	25
PART 1	
Genomic Sequencing and the Genome Project of <i>Corynebacterium pseudotuberculosis</i>.....	27
1.5 Genomic Sequencing; Completely Sequenced Genomes, New Generation Sequencers (NGS), the Massive Data Production and their Utilization.....	28
1.6 The Genome Project of <i>Corynebacterium pseudotuberculosis</i> Species.....	29
1.6.1 Research Article.....	31
Complete genome sequence of <i>Corynebacterium pseudotuberculosis</i> biovar <i>ovis</i> strain P54B96 isolated from antelope in South Africa obtained by rapid next generation sequencing technology. Standards in Genomic Sciences 2012	
1.6.2 Research Article	43
Genome sequence of <i>Corynebacterium pseudotuberculosis</i> biovar <i>equi</i> strain 258 and prediction of antigenic targets to improve biotechnological vaccine production. Journal of Biotechnology 2012	
1.6.3 Research Article.....	51
Whole-genome sequence of <i>Corynebacterium pseudotuberculosis</i> strain Cp162, isolated from camel. Journal of Bacteriology 2012	

1.6.4 Research Article.....	55
Complete genome sequence of <i>Corynebacterium pseudotuberculosis</i> strain 1/06-A, isolated from a horse in North America. Journal of Bacteriology 2012	
1.6.5 Research Article.....	58
Complete genome sequences of <i>Corynebacterium pseudotuberculosis</i> strains 3/99-5 and 42/02-A, isolated from sheep in Scotland and Australia, respectively. Journal of Bacteriology 2012	
1.6.6 Research Article.....	62
Complete genome sequence of <i>Corynebacterium pseudotuberculosis</i> strain Cp267, isolated from a llama. Journal of Bacteriology 2012	
1.6.7 Research Article.....	66
The pan-genome of the animal pathogen <i>Corynebacterium pseudotuberculosis</i> reveals differences in genome plasticity between the biovar <i>ovis</i> and <i>equi</i> strains. Plos One 2013	
1.6.8 Overall Conclusion PART 1.....	81
INTRODUCTION 2	
Large-Scale Genomic Data and Protein Modelling; Structure Based Drug and Vaccine Targets Identification.....	
	82
2.1 Bioinformatics and Computational Biology.....	83
2.2 3D Structure Prediction and Drug Designing; A Computational Approach.....	83
2.3 Application of Comparative Homology Modeling.....	87
PART 2	
Core-Modelome of <i>C. pseudotuberculosis</i> Species.....	
	89
2.4 MHOLine: A High Throughput Workflow for Comparative Homology Modelling..	90
2.5 Research Article	93
Modelome derived intra-species broad-spectrum drug and vaccine targets identification in the animal pathogen <i>Corynebacterium pseudotuberculosis</i> . (In submission process to PROTEINS: STRUCTURE, FUNCTION & BIOINFORMATICS ISSN: 1097-0134)	
2.5.1 Conclusion.....	132
2.6 Research Article.....	133
Globally conserved inter-species bacterial PPIs based conserved host-pathogen interactome derived novel target in <i>C. pseudotuberculosis</i> , <i>C. diphtheriae</i> , <i>M. tuberculosis</i> , <i>C. ulcerans</i> , <i>Y. pestis</i> , and <i>E. coli</i> targeted by <i>Piper betel</i> compounds. INTEGRATIVE BIOLOGY 2013	
2.6.1 Conclusion.....	149
2.7 Final Conclusion.....	150

III. PERSPECTIVES.....	151
Future Perspectives.....	152
Short Paper/Overview.....	152
Homology modeling of sortases from <i>C. diphtheriae</i> , putative targets for computational drug designing. Metabolites ISSN 2218-1989...in progress)	
IV. APPENDIX.....	153
IV.I Research Article.....	154
Exoproteome and secretome derived broad-spectrum novel drug and vaccine candidates in <i>Vibrio cholerae</i> targeted by <i>Piper betel</i> derived compounds. Plos One 2013	
IV.II Research Article.....	165
Genome sequence of <i>Exiguobacterium antarcticum</i> B7, isolated from a biofilm in Ginger Lake, King George Island, Antarctica. Journal of Bacteriology 2012	
IV.III Research Article.....	169
<i>Campylobacter fetus</i> subspecies: Comparative genomics and prediction of potential virulence targets. Gene 2012	
V. CURRICULUM VITAE.....	182
VI. BIBLIOGRAPHY.....	189

LIST OF ABBREVIATIONS AND ACRONYMS

Å	Angstrom = 1.0×10^{-10} meters
ACT	Artemis Comparison Tool
A+T	Adenine + Thymine content
BAC	Bacterial Artificial Chromosome
BATS	Blast Automatic Targeting for Structures
BLASTn	Basic Local Alignment Search Tool (nucleotide)
BLASTp	Basic Local Alignment Search Tool (protein)
BLOSUM	BLOCK SUBstitution Matrix
BHI	Brain Heart Infusion
BISTIC	Biomedical Information Science and Technology Initiative Consortium
BLAST	Basic Local Alignment Search Tool
CMNR	<i>Corynebacterium, Mycobacterium, Nocardia e Rhodococcus</i>
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
°C	Degree Celsius
C α	Alpha Carbon
CDS	Coding Sequences
CLA	Caseous Lymphadenitis
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico ("National Counsel of Technological and Scientific Development")
COGs	Protein Database of Clusters of Orthologous Groups
<i>Cp</i>	<i>Corynebacterium pseudotuberculosis</i>
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
3D	Three-Dimensional
DDBJ	DNA Data Bank of Japan
DEG	Database of Essential Genes
DNA	Deoxyribonucleic Acid
dsDNA	double-stranded DNA
EC	Enzyme Commission
EBI	European Bioinformatics Institute
ECF	Extracytoplasmic Function
EDTA	Ethylenediamine Tetraacetic Acid
EMBL	European Molecular Biology Laboratory
E-value	Expected Value
ExpASY	Expert Protein Analysis System
FAPEMIG	Fundação de Amparo a Pesquisa do Estado de Minas Gerais

(Research Support Foundation of the State of Minas Gerais)

FAPESPA	Fundação Amazônia Paraense (Research Support Foundation of the State of Pará)
G+C	Guanine + Cytosine Content
GEBA	Genome Encyclopedia of Bacteria and Archaea Genomes
GO	Gene Ontology
GOLD	Genome Online Database
GSS	Genomic Survey Sequence
HCl	Hydrochloric Acid
HGT	Horizontal Gene Transfer
HMMs	Hidden Markov Models
ICEX	Instituto de Ciências Exatas
INSDC	International Nucleotide Sequence Databases Collaboration
IDA	Inferred from Direct Assay
IIOAB	Institute of Integrative Omics and Applied Biotechnology
IMG	Integrated Microbial Genomes
IMP	Integral Membrane Protein
InterPro	Integrative Protein Signature Database
LGCM	Laboratório de Genética Celular e Molecular (Laboratory of Cellular and Molecular Genetics)
LPDNA	Laboratório de Polimorfismo do DNA (Laboratory of DNA Polymorphism)
LVI	Length Variation Index
MD	Molecular Dynamics
µg	Micrograms
MG	Minas Gerais
MIGS	Minimum Information About a Genome Sequence
µM	Micrometer
mM	Mili Molar
µL	Micro Liter
MLST	Multilocus Sequence Typing
NaCl	Sodium Chloride
NAS	Non-traceable Author Statement
NCBI	National Center of Biotechnology Information
NGS	Next-Generation Sequencing
NIH	National Institute of Health
NMR	Nuclear Magnetic Resonance

NTMs	<i>Non-tuberculosis Mycobacteria</i>
PCR	Polymerase Chain Reaction
ORF	Open Reading Frame
PDB	Protein Data Bank
PDF	Probability Density Function
PGM	Personal Genome Machine
PGDM	Pathway/Genome Database
Pfam	Database of Protein Families
PLD	Phospholipase D
rDNA	ribosomal DNA
RCSB	Research Collaboratory for Structural Bioinformatics
RFLP	Restriction Fragment Length Polymorphism
RGMG	Rede Genoma de Minas Gerais
RMSD	Root Mean Square Deviation
RNA	Ribonucleic Acid
rRNA	ribosomal RNA
RPGP	Rede Paraense de Genômica e Proteômica
RPM	Rotation per Minute
<i>rpoB</i> gene	β Subunit of RNA Polymerase
TAS	Traceable Author Statement
SIGS	Standard in Genomics
TDM	Trehalose Dimycolate
TMM	Trehalose Monomycolate
TWAS	The Academy of Sciences for the Developing World, Trieste, Italy
UFMG	Universidade Federal de Minas Gerais (Federal University of Minas Gerais)
UFPA	Universidade Federal do Pará (Federal University of Pará)
UniProt	Universal Protein Resource
WGS	Whole Genome Shotgun

LIST OF FIGURES

FIGURE 1:28

Graph showing the accelerated number of genome projects in the Genome OnLine Database (GOLD).

FIGURE 2:84

Protein Data Bank showing the yearly growth of total searchable structures.

FIGURE 3:.....86

Accuracy and application of protein structure models by comparative homology modelling, threading and *de novo* structure prediction techniques.

FIGURE 4:.....90

A schematic representation of **MHOLline** biological workflow, which combines a specific set of programs using multi fasta files of amino acids as input data for processing.

Abstract

Corynebacterium pseudotuberculosis (Cp) is the etiological agent of several infectious and contagious chronic diseases, including caseous lymphadenitis (CLA), ulcerative lymphangitis, mastitis, and edematous skin disease thus affecting a broad spectrum of animal hosts, including ruminants that cause a huge decrease in wool production and carcass quality, thus threatening economic and dairy industries worldwide. In this work, first, manual annotation was performed using Artemis: the annotation tool, for two *C. pseudotuberculosis* strains. For Cp162, isolated from a camel in UK and completely sequenced using the SOLiD v3 Plus NGS platform, a genome of 2,293,464 bp in size, having 52.17% GC content, 2,002 coding sequences, 4 rRNA operons, 49 tRNA genes, and 87 pseudogenes, was obtained. For strain P54B96, isolated from an antelope in South Africa and sequenced using the Ion Torrent PGM NGS platform, a genome of 2,337,657 bp in size, having 52.2% GC content, 2,084 coding sequences, 4 rRNA operons, 49 tRNA genes, and 62 pseudogenes, was obtained. They were deposited in the NCBI GenBank database under accession numbers CP003652 and CP003385, respectively. Motivated by an increasing demand for new drug and/or vaccine targets, secondly, a novel strategy using a high throughput workflow, called MHOLline, was followed to construct the modelome (3D models) of 15 *C. pseudotuberculosis* strains from various hosts and countries. Only Very High, High, Good and Medium to Good quality sequences were used from MHOLline classified groups (G2). Using a locally installed BALST NCBI tool, a set of 331 conserved proteins with 3D structures was selected showing 95-100% intra-species sequence similarity. The host proteomes considered in this study were human, horse, cow and sheep. Further filtering this core-modelome for essential and non-host homologous proteins, resulted in a final set of 10 proteins. Among these, only 4 proteins were identified as essential and non-host homologous and were considered as putative drug and vaccine targets, subjected to virtual screening and docking analyses. The druggability score, a drug target prioritization parameter, among others, was also calculated for all these proteins, allowing the prediction of druggable protein cavities. We further extrapolated our research to the other 6 essential host homologous proteins. A deep structure analysis at their residue level confirmed some conserved active site residues either exhibiting different conformations or even completely different residues. A multiple sequence alignment proved the residues conservation in these targets. Further, the role of these targets in different bacterial metabolic pathways, pathogenicity and virulence were also determined. It was proposed that the 6 essential host homologous proteins might also provide an extended choice for therapeutic targets. It is expected that our data will facilitate selection of *C. pseudotuberculosis* proteins for successful designing of new drugs and vaccines for a broad range of hosts.

Resumo

Corynebacterium pseudotuberculosis (*Cp*) é o agente etiológico de diversas doenças infecciosas e contagiosas crônicas, incluindo linfadenite caseosa (LC), linfangite ulcerativa, mastite e doença edematosa da pele e afeta um amplo número de hospedeiros. Neste trabalho, primeiramente, foi realizada a anotação gênica manual em duas linhagens de *C. pseudotuberculosis* utilizando a ferramenta Artêmis. A linhagem Cp162, isolado de camelo no Reino Unido e a P54B96, isolada de antílope na África do Sul, as quais tem genomas de 2.293.464 pb e 2.337.657 pb, 52,17% e 52,2% de conteúdo GC, 2002 e 2084 sequências codificadoras, 87 e 62 pseudogenes respectivamente. Ambas possuem 49 genes de tRNA e 4 operons de rRNA. Os genomas foram depositadas no banco de dados do NCBI GenBank sob os números de acesso CP003652 e CP003385, respectivamente. No segundo momento foram usados 15 genomas de *C. pseudotuberculosis* para um estudo em larga escala das proteínas preditas a partir destas sequências genômicas, através de técnicas de bioinformática e modelagem comparativa, usando o workflow MHOLline. Somente as sequências de qualidade, categorizadas em muito alta, alta, boa e média a boa foram usadas a partir dos grupos classificados por MHOLline. Usando alguns *scripts* desenvolvido pela equipe do LGCM, foram selecionadas um conjunto de 331 proteínas conservadas, que apresentavam estruturas 3D e considerando 95-100% de similaridade entre sequências intra-espécies. Visando identificar proteínas essenciais homólogas e não homólogas aos hospedeiros humano, equino, bovino e ovino, o core-modeloma foi filtrado resultando em um conjunto final de 10 proteínas. Apenas quatro proteínas foram identificadas como essenciais e não homólogas e foram consideradas como alvos putativos para drogas e vacinas, submetidos a triagem virtual e análises de *docking*. Adicionalmente, entre outros, um parâmetro para avaliar a drogabilidade de todas as proteínas, também foi calculado, permitindo a predição de cavidades das proteínas. Pesquisamos as outras 6 proteínas essenciais e homólogas realizando uma análise profunda das suas estruturas (*Cp* e dos hospedeiros) em nível de resíduos confirmando alguns resíduos conservados nos sítios ativos com conformações diferentes ou resíduos completamente diferentes. A conservação desses resíduos foi validada através de um alinhamento múltiplo de sequências. Além disso, o papel destes alvos em diferentes vias metabólicas bacterianas, patogenicidade e virulência também foram determinados. Baseado nesses estudos, as 4 proteínas essenciais e não homólogas foram propostas como potenciais alvos para tratamento e as 6 proteínas essenciais e homólogas também podem ser utilizadas como possíveis alvos terapêuticos devido às diferenças observadas em resíduos conservados. Espera-se que os nossos dados possam ajudar na seleção de proteínas de *C. pseudotuberculosis* para o desenvolvimento bem sucedido de novas drogas e vacinas para uma ampla classe de hospedeiros.