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**<br>
Supervisor: **Prof. Dr. Vasco Ariston de Carvalho Azevedo**<br>
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***

# *ACKNOWLEDEMENTS*

*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-UFPA, 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 (TWAS) 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

**INTRODUCTION 2**

**PART 2**

## LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| Å | Angstrom = 1.0 x $10^{-10}$ 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 | *Corynebacterium, Mycobacterium, Nocardia e Rhodococcus* |
| CAPES | Coordenação de Aperfeiçoamento de Pessoal de Nível Superior |
| °C | Degree Celsius |
| Cα | Alpha Carbon |
| CDS | Coding Sequences |
| CLA | Caseous Lymphadenitis |
| CNPq | Conselho Nacional de Desenvolvimento Científico e Tecnológico<br>("National Counsel of Technological and Scientific Development") |
| COGs | Protein Database of Clusters of Orthologous Groups |
| *Cp* | *Corynebacterium pseudotuberculosis* |
| 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<br>(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<br>(Laboratory of Cellular and Molecular Genetics) |
| LPDNA | Laboratório de Polimorfismo do DNA<br>(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 | *Non-tuberculosis Mycobacteria* |
| 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 |
| *rpoB* 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<br>(Federal University of Minas Gerais) |
| UFPA | Universidade Federal do Pará<br>(Federal University of Pará) |
| UniProt | Universal Protein Resource |
| WGS | Whole Genome Shotgun |

## LIST OF FIGURES

## 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 *Cp*162, 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 *Cp*162, 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.