*In silico* analysis of host-pathogen interaction: a case study with molecular mimicry phenomena

Thesis Submitted to the University of Delhi

* 1. **for the Degree of**
  2. **DOCTOR OF PHILOSOPHY  
     IN  
     BIOPHYSICS**
  3. **2020**
  4. 

**ANJALI GARG**

**DEPARTMENT OF BIOPHYSICS**

**UNIVERSITY OF DELHI, SOUTH CAMPUS**

**NEW DELHI - 110021**

**INDIA**

|  |
| --- |
| **DEPARTMENT OF BIOPHYSICS**  **UNIVERSITY OF DELHI, SOUTH CAMPUS**  Benito Juarez Road, New Delhi-110021  **UDSC** |

**CERTIFICATE**

This is to certify that ANJALI GARG has worked on the research project entitled, **“*In silico* analysis of host-pathogen interaction: a case study with molecular mimicry phenomena”** under my supervision and guidance. The contents of the thesis, being submitted to the Department of Biophysics, University of Delhi, South Campus for the award of the Degree of Doctor of Philosophy in Biophysics, are original and have been carried out by candidate himself. This thesis has not been submitted in full or part for the award of any degree or diploma to any other University.

**Anjali Garg**

(Research Scholar)

Enrolment Number: DR-562/08

|  |  |  |
| --- | --- | --- |
| **Dr. Manish Kumar**  (Supervisor)  Department of Biophysics  University of Delhi, South Campus  New Delhi - 110021 |  | **Head of the Department**  Department of Biophysics  University of Delhi, South Campus  New Delhi - 110021 |

|  |
| --- |
| **DEPARTMENT OF BIOPHYSICS**  **UNIVERSITY OF DELHI, SOUTH CAMPUS**  Benito Juarez Road, New Delhi-110021  **UDSC** |

Date:

**CERTIFICATE OF ORIGINALITY**

The research work embodied in this thesis entitled **“*In silico* analysis of host-pathogen interaction: a case study with molecular mimicry phenomena”** has been carried out by me at Department of Biophysics, University of Delhi, South Campus, New Delhi, India. The thesis has been subjected to the plagiarism checked by ‘**Turnitin’** software. The work submitted for the consideration of the award of Ph.D. is original.

**Anjali Garg**

|  |
| --- |
| **DEPARTMENT OF BIOPHYSICS**  **UNIVERSITY OF DELHI, SOUTH CAMPUS**  Benito Juarez Road, New Delhi-110021  **UDSC** |

**Supervisor’s Certificate for Exclusion of Self-Published work**

**The contents of the**

1. Part of the chapter 3 entitled “Ion Channel Collective Behaviour and Noise Analysis” has been published in Shrivastava, R., Malik, C., & Ghosh, S. (2016). Open channel current noise analysis of S6 peptides from KvAP channel on bilayer lipid membrane shows bimodal power law scaling. *Physica A: Statistical Mechanics and its Applications*, *451*, 533-540.

2. Part of the chapter 5 entitled “Role of Ion Channel Collective Behaviour in Generation of Action Potential” has been published in Talukdar, S., Shrivastava, R., & Ghosh, S. (2019). Modeling activity-dependent reduction in after hyper-polarization with Hodgkin-Huxley equation of action potential. *Biomedical Physics & Engineering Express*, *5*(4), 047001.

**These published works have been included in this thesis and has not been submitted for any degree to any University/Institute.**

**Signature of Student Signature of Supervisor**

Place: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**STUDENT APPROVAL FORM**

|  |  |
| --- | --- |
| **Name of the Author** | Anjali Garg |
| **Department** | Biophysics |
| **Degree** | Doctor of Philosophy |
| **University** | University of Delhi |
| **Guide** | Dr. Manish Kumar |
| **Thesis Title** | ***In silico* analysis of host-pathogen interaction: a case study with molecular mimicry phenomena** |
| **Year of Award** | 2020 |

**Agreement**

1. I hereby certify that, if appropriate, I have obtained and attached hereto a written permission/statement from the owner(s) of each third party copyrighted matter to be included in my thesis/dissertation, allowing distribution as specified below.

2. I hereby grant to the university and its agents the non-exclusive license to archive and make accessible, under the conditions specified below, my thesis/dissertation, in whole or in part in all forms of media, now or hereafter known. I retain all other ownership rights to the copyright of the thesis/dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis, dissertation, or project report.

Conditions:

|  |  |
| --- | --- |
| 1. Release the entire work for access worldwide | **No** |
| 2. Release the entire work for ‘My University’ only for  1 Year  2 Year  3 Year  and after this time release the work for access worldwide. | Release the entire work for  ‘My University’ only for  **2 Years** and after this time  release the work for access worldwide. |
| 3. Release the entire work for ‘My University’ only  while at the same time releasing the following parts of the work (e.g. because other parts relate to publications) for worldwide access.  a) Bibliographic details and Synopsis only.  b) Bibliographic details, synopsis and the following  chapters only.  c) Preview/Table of Contents/24 page only. | Table of Contents Only |
| 4. View Only (No Downloads) (worldwide) | View Only (No Downloads) (worldwide) |

Signature of the Scholar Signature and seal of the Supervisor

**Anjali Garg Dr. Manish Kumar**

Place: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Dedicated***

***to***

***My Parents***

**Acknowledgements**

The peaceful environment of South Campus of University of Delhi and Department of Biophysics is an excellent place for learning growth and development. I would like to thank University of Delhi for providing me a platform for Ph.D. work. I am grateful for state-of-the-art facilities, expert scientists and technical staff available, who directly or indirectly helped me in my Ph.D. work. Certainly, there are a group of people who supported and made it possible, I would like to extend my gratitude towards them.

First and foremost, a special thanks to my supervisor Dr. Manish Kumar for his guidance, support and motivation, without which it would have been impossible to complete my thesis. He provided requisite facilities and always stood by me. His interdisciplinary approach towards science, with strong emphasis to fundamentals, helped me address the scientific problems with new perspective. Most importantly the freedom to explore new ideas and methods benefited enormously in development of my technical and theoretical understanding of the subject.

I would also like to thank the members of my PhD advisory committee, Dr. Manisha Goel, Dr. Yogender Pal Khasa and Dr. Jagreet Kaur for their insightful comments and suggestions. I am obliged to Prof. Alo Nag (HOD, Department of Biophysics) for his comprehensive advice, support and guidance.

I am obliged to Prof. Subhendu Ghosh (Former HOD, Department of Biophysics) for his comprehensive advice, support and guidance. I would like to express my heartfelt gratitude to Dr. Neelja Singhal for guidance, advice and encouragement without which this thesis would not be possible.

Special thanks go out to all the lab members, current and former, Dr. Ravindra Kumar, Vandana Chaurasia, Sohni Singh Jain, Abhishikha Srivastava, Deeksha Pandey, Manisha Aswal, Aakriti Jain and Govinda Rao Dabburu for their support and motivation. It has been a pleasure to work with a great group of colleagues, who made the work environment enjoyable and productive. I wish all of you the best of luck with your future accomplishments.

I cannot forget to mention here all the researchers at the Department of Biophysics, Shobha Kumari, Ved Vrat Verma, Dr. Shikha Rani, Poonam Sharma, Vineeta Kaushik, Anchal Deswal, Archana Sharma, Dr. Chetan Mallik, Sandipan Talukdar, Rajan Srivastava, Shikha Srivastava, Spandan Kumar, Bhanu Sharma, Shumaila Iqbal Siddiqui, Nikita, Dr. Sunil, Katiki Madhusudhanarao and Daniel T. Tuikhang for their suggestions and help. I would like to acknowledge lab staff Mr. Pramod Kumar and Rajesh Kumar for his help and technical support.

I owe my gratitude to my parents, and to my siblings for their never-ending love, comfort and support. They have provided invaluable advices throughout all stages of my life and have always been motivating me to be curious and to follow my dreams.

Last but not the least, I take this opportunity to express my appreciation to Indian Council of Medical Research which have provided me with financial assistance in my PhD study.

At last but not least, I would like to thank University Grant Commission (UGC), Council of Scientific and Industrial Research (CSIR) and Department of Science and Technology (DST) for providing me financial assistance.

**Anjali Garg**

Place: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Table of Contents

1. Introduction1
   1. Introduction2
   2. Aims and objective2
   3. Structure of thesis2
2. Review of Literature4
   1. Innate and adaptive immunity5
      1. Innate immune system: Fast and broadly effective6
      2. The adaptive immune system: Precision and a long memory6
      3. T lymphocytes or T cells6
      4. B lymphocytes or B cells6
   2. “Molecular mimicry”: an evolving concept5
      1. Type of molecular mimicry6
      2. Molecular mimicry and cross-reactivity: what is necessary? 6
   3. The initiation of autoimmunity: lighting the match5
      1. Type of autoimmune diseases6
      2. Autoimmunity can predispose to infection diseases 6
      3. Therapeutics options for autoimmunity-associated infection diseases 6
   4. Mycobacterium tuberculosis (Mtb): a case study5
3. **miPepBase: A database of experimentally verified peptides involved in molecular mimicry**4
   1. Introduction5
   2. Material and Methods5
      1. Data collection and compilation6
      2. Web interface and database architecture 6
      3. Database accessibility6
      4. Tools integrated in miPepBase 6
   3. Results 5
   4. How to search query into miPepBase?5
   5. Discussion5
   6. Comparison with other available database of antigenic peptides5
   7. Limitations and future prospects5
   8. Database update5
   9. Accessibility and data download5
4. **Structural and functional charcteristics of mimicry proteins and peptides of bacteria, viruses and host**4
   1. Introduction5
   2. Material and Methods5
      1. Retrieval of experimentally validated mimicry proteins from miPepBase6
      2. Structural characterization of host- and path-proteins5
         1. Order/disorder propensity of amino acids in the host- and path-proteins 6
         2. Investigating the IDPRs – MoRFs, SliMs and LCRs 6
      3. Functional enrichment analysis6
   3. Results 5
      1. Benchmarking dataset 5
      2. Structural order/disorderliness in the host and pathogen mimicry proteins5
      3. MoRFs, SLiMs, and LCRs in mimicry protein/peptides 5
      4. Functional characterization of the host and pathogen mimicry proteins 5
   4. Discussion5
5. **Using molecular-mimicry-inducing pathways of pathogens as novel drug targets**4
   1. Introduction5
   2. Material and Methods5
      1. Data extraction6
      2. Protein-protein interaction search5
      3. Pathway mapping and determination of chokepoints in mycobacterial metabolic pathway6
      4. Authentication of chokepoint targets and druggability of selected targets6
      5. Drug-target interaction6
      6. Drug repurposing of M. tuberculosis 6
   3. Results And discussion5
      1. Identification of interacting proteins of mimicry proteins 5
      2. Identification and validation of the metabolic chokepoints5
      3. Drug molecule for metabolic chokepoints5
      4. Drug repurposing for M. tuberculosis 5
   4. Prospects of the current approach5
   5. concluding remarks5
6. **Using molecular-mimicry-inducing pathways of pathogens as novel drug targets**4
   1. Introduction5
   2. Material and Methods5
      1. Retrieval of experimentally validated MAP mimicry proteins involved in autoimmune diseases6
      2. Protein-protein interaction studies5
      3. Removal of human homologs of MAP proteins6
      4. Determining the chokepoint(s) of the metabolic pathways of MAP6
      5. Validation of the essentiality of chokepoint proteins 6
      6. Identification and validation of drug molecules against the chokepoint proteins6
      7. Analysis of drug-target interactions 6
   3. Results And discussion5
      1. Identification of interacting partners of MAP mimicry proteins 5
      2. Identification and validation of the metabolic chokepoints of MAP5
      3. Drug molecule for metabolic chokepoints5
      4. Quality assessment of the modeled metabolic chokepoint protein structures 5
      5. Protein-ligand interactions between the MAP proteins and DrugBank molecules5
      6. Prediction of off-target binding 5
7. **Mtb etablishmnet**4
   1. Introduction5
   2. Material and Methods5
      1. 6
   3. Results And discussion5
      1. 5
8. **mRNALoc: a novel machine-learning based *in-silico* tool to predict mRNA subcellular localization** 4
   1. Introduction5
   2. Material and Methods5
      1. Collection of subcellular location annotated dataset of mRNA 6
      2. Redundancy removal 5
      3. Construction of training and independent dataset and training methodology 6
      4. Conversion of a nucleotide sequence into machine learning input feature 6
      5. Hybrid Feature Vectors (HFV)6
      6. Fragmented Sequence Encoding 6
      7. Performance Evaluation Strategies 6
   3. Overview of mRNALoc5
   4. result and discussion5
      1. Performance on Complete Sequence Information 6
      2. All Features Helped, but the Combination Performed Best and Most Robust 5
      3. Receiver Operating Characteristics (ROC) Plot and Area Under ROC Curve (AUC) analysis 6
      4. Benchmarks on independent mRNA datasets 6
   5. Comparison with existing mRNA subcellular localization prediction methods 5
   6. Description of the webserver 5
      1. Implementation of mRNALoc 6
      2. The output of mRNALoc 5
   7. Conclusions and future prospects5

**Summary & future prospects 4**

**Bibliography** 4

**Annexure**

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure 3.1.** | Architecture of miPepBase. |  |
| **Figure 3.2.** | **Data statistics** (A) Based on pathogen taxonomic group  (B) Based on autoimmune disease. |  |
| **Figure 3.3.** | **Process of stepwise data retrieval and analysis in miPepBase**. The user can search query with following options: (A) Keyword search, (B) Search by disease, (C) Search by host and pathogen taxonomic group, (D) Search by host and pathogen name. The search from (A–D) options display search result table and from that user can select the entry/displayed result for further detailed analysis. Sequence based search can also be searched by (E) BLAST search option and each hit is further linked to its details information page. The detail of result obtained from search options (A–E) is displayed by corresponding small case (the number indicates step number). From example a1–a6 denotes the results that can be obtained using keyword search option (A). |  |
| **Figure 4.1.** | Gene Ontology based functional annotation of bacterial, viral and host mimicry proteins. (a) Biological Process, (b) Molecular Function and (c) Cellular Component |  |
| **Figure 5.1.** | **The scheme of drug repurposing proposed against *Mtb***. In the figure we have explained the complete process which is clustered into three major sections: (1) *Interactome analysis* includes protein data retrieval; collection of interacting proteins and removal of path-proteins which are homologous to human protein(s). (2) *Filter potential target(s*) that include mapping of mycobacterial nHIPPP dataset into their metabolic pathway(s) and search of possible chokepoint protein(s). If chokepoint proteins could able to pass through filters namely part of core proteome (A) or essential proteins (B). All chokepoint protein that crosses either filter is moved to the third step. (3) *During drug repurposing* chokepoints proteins were searched for effective ligand(s) and their interaction was analyzed after docking. In the last step *Mtb* homolog was searched for each chokepoint protein (C). List of databases & servers used during the whole process is 1: miPepBase, 2: STRING, 3: KEGG, 4: UniProtKB, 5: DEG, 6: DrugBank, 7: PatchDock, 8: LigPlot+ v.1.4 |  |
| **Figure 6.1.** | The workflow adopted for the identification of novel drug targets for treating *M. avium* subsp. *paratuberculosis*-associated autoimmune disorders. |  |
| **Figure 8.1.** | Systematic workflow of mRNAloc training and functioning. |  |
| **Figure 8.2.** | **Overall schema of mRNALoc.** mRNALoc predicts five subcellular locations viz., mitochondria, cytoplasm, nucleus, endoplasmic reticulum and extracellular. Initially it removes non-standard nucleotides from the sequence, generates combined features from pseudo K-tuple nucleotide composition which is further used as input for Support Vector Machine (SVM) prediction. |  |
| **Figure 8.3.** | Performance achieved during five-fold cross-validation with different values of K-mer. |  |
| **Figure 8.4.** | The ROC curves for (a) extracellular region (b) endoplasmic reticulum (c) cytoplasm (d) mitochondria (e) nucleus. A ROC cure plots the true positive rate (i.e. sensitivity) against the false positive rate. For each location performance on complete and splitted sequence is shown at different Kmer values. The performance on independent data is also shown. |  |
| **Figure 8.5.** | Screenshots of mRNALoc webserver. |  |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **Table 4.1.** | Categorization of mimicry proteins of bacteria, viruses and host on the basis of PDR (percentage of disordered residues in protein) |  |
| **Table 4.2.** | Distribution of MoRFs, SLiMs and LCRs in the mimicry proteins and mimitopes of bacteria, viruses and host |  |
| **Table 5.1.** | List of *Mycobacterium. spp*. protein that are involved in molecular mimicry (source: Garg, A., et al. (2017) miPepBase: A Database of Experimentally Verified Peptides Involved in Molecular Mimicry. *Front Microbiol*. 8, 2053). |  |
| **Table 5.2.** | List of pathogen mimicry protein, its interaction partners (IPPP), name of human homolog (if present), KEGG pathway ID to which IPPP belongs and chokepoint proteins. The table shows information related to pathogen protein involved in molecular mimicry (Column 2), IPPP collected from STRING database at default parameters (Column 3), HIPPP among the IPPP (column 4), IPPP which couldn’t be mapped on KEGG (column 5), KEGG pathway ID in which IPPP mapped (Column 6) and chokepoints proteins found after manual survey of KEGG pathways ID listed in column 6 (Column 7). |  |
| **Table 5.3.** | Drug target validation. The table shows information of path-proteins (Column 1), potential chokepoint found in KEGG metabolic network (Column 2), chokepoint proteins which were part of essential genes (Column 3) and core proteins (Column 4), homologous of chokepoint proteins in Mtb proteome (Column 5) and chokepoint protein listed as drug target in DrugBank database (Column 6). Column 7 has potential drug molecule as per DrugBank target protein and Column 8 contained the drugs, which qualified the filter of drug candidate filter. |  |
| **Table 6.1.** | The list of MAP mimicry proteins, interacting partners of MAP mimicry proteins (IPMMP), name of human homolog of IPMMP (if present), Non-Human Homolog of IPMMP (nHIPMMP), KEGG pathway ID to which non-human homolog of IPMMP belongs (KEGG id is present in parenthesis), and chokepoint proteins. |  |
| **Table 6.2.** | Drug and their target details. |  |
| **Table 6.3.** | Details of drugs proposed against MAP associated autoimmunity. |  |
| **Table 6.4.** | Interaction pattern between drug and target proteins. |  |
| **Table 6.5.** | Quality assessment scores of *in silico* protein models. |  |
| **Table 7.** |  |  |
| **Table 8.1.** | Estimation of the performance metrics for mRNA location identification under different combination of K-mer features. Values indicated in italics and bold font show the highest prediction score of training, and only in bold font show the performance on independent dataset. (Sen: Sensitivity, Spe: Specificity, ACC: Accuracy, MCC: Mathew’s correlation coefficient, TN: True negative, FN: False negative, TP: True positive, FP: False positive, THR: Threshold, and AUC: Area under ROC curve) |  |
| **Table 8.2.** | Brief comparison of the advantages and limitations among mRNALoc, RNATracker and iLoc-mRNA. |  |
| **Table 8.3.** | Comparative evaluation of mRNALoc and iLoc-mRNA. In extracellular region and mitochondria no human mRNA was present, hence these two locations were not included in the evaluation. |  |

**List of Abbreviations**

|  |  |
| --- | --- |
| ACC | Accuracy |
| ACE | atomic contact energy |
| AUC | Area Under ROC Curve |
| BLAST | Basic Local Alignment Search Tool |
| BLOSUM | BLOcks SUbstitution Matrix |
| BP | Biological Process |
| CC | Cellular Component |
| CD | Crohn’s Disease |
| DEG | Database of Essential Genes |
| DNA | Deoxyribonucleic acid |
| FDA | Food and Drug Administration |
| FN | false negatives |
| FP | false positives |
| GO | Gene Ontology |
| HIPMMP | Homologous interacting partners of MAP mimicry proteins |
| HIPPP | Homologous interaction partners of pathogen’s proteins |
| IDPRs | intrinsically disordered protein regions |
| IDRs | Intrinsically Disordered Regions |
| IPMMP | Interacting partners of MAP mimicry proteins |
| IPPP | Interaction partners of pathogen’s proteins |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| LCRs | low complexity regions |
| MAP | *Mycobacterium avium* subspecies *paratuberculosis* |
| MCC | Matthew’s correlation coefficient |
| MF | Molecular Function |
| miPepBase | Mimicry Peptide Database |
| MoRFs | Molecular recognition features |
| mRNALoc | mRNA Localization |
| MS | Multiple Sclerosis |
| Mtb | *Mycobacterium tuberculosis* |
| nHIPMMP | Non-homologous interacting partners of MAP mimicry proteins |
| nHIPPP | Non-homologous interaction partners of pathogen’s proteins |
| PDB | Protein Data Bank |
| PDR | percentage of disordered residues |
| PPI | protein-protein interactions |
| PseKNC | Pseudo K-tuple nucleotide composition |
| QMEAN | Qualitative Model Energy Analysis |
| RMSD | root-mean-square deviation |
| RNA | Ribonucleic acid |
| ROC | receiver operating characteristics |
| SLiMs | short linear motifs |
| T1DM | Type 1 diabetes mellitus |
| TB | Tuberculosis |
| TN | true negatives |
| TNF | Tumor necrosis factor |
| TP | true positives |
| TTD | Therapeutic Target Database |

**Abstract**

Abstract