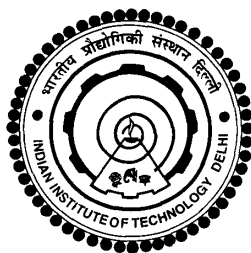


**AUTOMATION OF COMPUTER-AIDED GENOMES TO
HITS DISCOVERY PIPELINE AND ITS APPLICATION
VIA CASE STUDIES**

RUCHIKA BHAT



**DEPARTMENT OF CHEMISTRY
INDIAN INSTITUTE OF TECHNOLOGY DELHI
OCTOBER 2019**

© Indian Institute of Technology Delhi (IITD), New Delhi, 2019

**AUTOMATION OF COMPUTER-AIDED GENOMES TO
HITS DISCOVERY PIPELINE AND ITS APPLICATION
VIA CASE STUDIES**

by

RUCHIKA BHAT

DEPARTMENT OF CHEMISTRY

submitted

in fulfillment of the requirement of the degree of Doctor of Philosophy

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

OCTOBER 2019

Dedicated to my respected parents and in-laws

&

my inspiring brother; Mohit

&

my beloved husband; Ankur

Certificate

This is to certify that the thesis entitled, “**Automation of computer-aided genomes to hits discovery pipeline and its application via case studies**”, being submitted by **Ms. Ruchika Bhat** to the Indian Institute of Technology Delhi for the award of the degree of **Doctor of Philosophy** in Chemistry, is a record of bonafide research work carried out by her. Ruchika Bhat has worked under my guidance and supervision and has fulfilled the requirements for the submission of this thesis, which to my knowledge has reached the requisite standard.

The results contained in this dissertation have not been submitted in part or full to any university or institute for the award of any degree or diploma.

Prof. B. Jayaram
Professor
Department of Chemistry
Indian Institute of Technology Delhi

Date:

Acknowledgements

My Ph.D. journey has been a life-changing experience for me. It helped me grow as a better individual, not only professionally but personally also, and all this would not have been possible without the help of many people.

Foremost, I am deeply indebted to my supervisor, **Prof. B. Jayaram**, for his immense support and guidance all throughout my PhD tenure. His need for perfection, dedication and passion for new and fundamental discoveries, unflinching courage and conviction has always inspired me to do more. My sincere thanks to my supervisor who always guided and motivated me to keep marching ahead even in the low dips during PhD. I would also like to express my sincere gratitude towards **Mrs. Lakshmi Jayaram** (my supervisor's wife) for her unconditional love and blessings.

The major share of my success, till date, I owe to the blessings and constant support of my family members. My parents **Mr. Ramesh Bhat** and **Mrs. Nisha Bhat** who shaped me into a being I am today, my parents-in-law **Mr. Ashok Raina** and **Mrs. Anita Raina** for their unconditional love, support, care and blessings. My dearest brother, **Mohit Bhat**, who has always been the source of my strength and without him I would be nothing! His constant support and inspirational talks have always made me look forward in life with positivity. I owe a big time to him and his immense faith in me which kept me motivated always! Most importantly, I owe a deep gratitude towards a special person, my loving and caring husband, **Ankur Raina**, which cannot be sufficed in words to express. His words, "I am with you" would always fill me with the energy to solve any problem in life as a cakewalk. I am grateful to him for being a constant

pillar of support & encouragement. I consider myself the luckiest in the world to have such a loving and caring family.

I am also grateful to my dearest friends **Nanda, Manali, Bhavana** and **Shalini** who have been there no matter what came up in life. I have always found them being a constant source of energy that filled life with happiness and hope.

I greatly appreciate the efforts from my lab-friends **Akhilesh, Pradeep, Amita** with whom I shared the ups and downs of these 5 years. With them every problem became 1/4th and every joy increased 4 folds. I would thank my seniors **Dr. Abhilash, Dr. Ankita** and **Dr. Rahul** for guiding during initial stages of PhD and keeping my morale high. Special thanks to **Manpreet Singh** and **Puneeta Ma'am** for their technical helps and unconditional support whenever required.

I am also thankful to **Dr. Manidipa Banerjee, Dr. Vivekanandan Perumal, Dr. Ashok Patel** and **Prof. Ravi P Singh** for all the scientific discussions I had with them which always led to fruitful results. The collaborations I had with them have created a lifetime experience and beautiful memories for me.

I am also grateful to my SRC members; **Prof. Nalin Pant, Prof. D. Sundar** and **Prof. S. Khare** for their time to time guidance and keeping me posted with latest science updates and discoveries related to my area. I am grateful to all supporting staff at the Chemistry department, IIT Delhi for their kind help and cooperation. I acknowledge the financial support from DST, Govt. of India for my **INSPIRE fellowship**.

I am grateful to the immense support my IIT friends gave me during the ups and down of life and kept me going with knowledge and wisdom. A sincere thanks to my Art of Living family

and friends, esp. **Banani Didi, Subhinoy Dada, Tanushree Ma'am, Venketesh Bhaiya, Sakshi, Prashant, Anu, Ankit** and all my dear friends.

Lastly, I would like to thank **Almighty** for **everything** I have received in life so far and would be achieving ahead!

Ruchika Bhat

Abstract

Since the deposition of first genome sequence on 9th June 1982 in the NCBI repository, the number of genome sequences has been exponentially increasing (~3,52,904 genome sequences deposited in NCBI covering over 93,618 organisms as per July 12, 2019 RefSeq Release 95). In the current scenario, on an average each approved drug takes about 14 years from conception to market and costs approximately \$2.6 billion. If computational methods become mature enough, then one expects significant decrease in cost and time as well as acceleration in the drug approval process in turn. Advances in computer aided drug design along with rise in the biological data enables to aim for an automated computer aided drug discovery pipeline, which can generate a list of hit molecules against any target protein of any pathogen (Chapter 1).

This thesis describes the automation of one such drug discovery pipeline named *Dhanvantari* (http://scfbio-iitd.res.in/software/dhanvantari_new/Home.html) which connects all the dots from ‘genome’ to ‘hit molecule’ identification computationally (Chapter 2). The pipeline starts with input of genome sequence of any pathogen and utilizing the central dogma of life predicts the corresponding protein sequences. After sequential steps of predicting a good drug target, its true binding pocket/active site, the pipeline helps in identification of potential hit molecules against the pathogen. Alternatively one can input gene sequence, protein sequence or tertiary structure of targeted protein, if available, directly in the pipeline to identify hit molecules against them. The pipeline works on pathogenic genomes. However, from gene level the pipeline can work on all organisms including humans. The pipeline is well validated on experimental data from FDA via *in silico* testing on 33 protein drug targets from protein level and 3 pathogens from genome level. The pipeline was able to bracket most of the FDA approved drugs available

against the test organisms and drug targets which builds confidence in the working of the pipeline.

Further, to validate the working efficacy of the pipeline, two individual case studies are conducted, one against Hepatitis B Virus and another against Hepatitis A Virus. In the case of both these viruses, only preventive vaccines are available till date and no FDA approved small molecule drug is available (although, a few nucleoside analogs are available in case of HBV). The pipeline has been used for finding anti-HBV and anti-HAV small molecule inhibitors against their respective target proteins. The computational findings of the pipeline were validated experimentally to demonstrate its applicability in solving existing drug discovery problems.

For HBV, two protein targets HBeAg and HBsAg have been targeted (Chapter 3). To date no small molecule inhibitor is reported to inhibit the *production* of HBsAg protein. Ours is first such study. A few other reported inhibitors against HBsAg are able to inhibit its secretion only. Eight compounds were identified using *Dhanvantari* protocol as promising hits against HBV infections. These compounds showed low micro molar scale inhibition in experimental assays providing a proof of concept.

In case of HAV, which belongs to *Picornaviridae* family of viruses, the pipeline was utilized to design generic inhibitors against this family of virus (Chapter 4). Here, 3C protease has been targeted as this protein is present in most of the viruses of this family. Seven compounds were identified using *Dhanvantari* protocol as promising hits against HAV infections. Also, two molecules are designed taking isatin as a parent scaffold to develop anti-HAV properties. All the nine compounds showed low micro molar scale inhibition in experimental assays providing a proof of concept.

To accelerate the application of the computational ‘genome to hit’ pipeline for identifying list of hit molecules against viral targets, a database having information of the structural and functional annotation of viral proteins is needed. Hence, a comprehensive database of 5,930 human host viral proteins is generated, named *viHumans* (<http://scfbio-iitd.res.in/viHumans/>) (Chapter 5). The database covers 1,116 human host viruses comprising 1,116,019 viral proteins. Information about the genome accession numbers of these viruses, gene and protein names, protein sequences, modelability index, sequence level annotations, pathways involved, taxonomic lineages etc. are provided in the database. The database consists of both available as well predicted information about structures, functions, mechanisms/pathways involved and functional sites of these proteins along with their known and predicted ligands. Therefore, *viHumans* database acts as a comprehensive repository for all the information required to initiate structure based drug discovery of 5930 viral proteins.

The overall focus of this thesis work has been to put together a complete drug design software suite with experimental validations using various case studies at different levels (Chapter 6).

शोध-सार

NCBI रिपॉजिटरी में 9 जून 1982 को पहले जीनोम अनुक्रम के चित्रण के बाद से, जीनोम अनुक्रमों की संख्या में तेजी से वृद्धि हुई है (~ जुलाई 12, 2019 RefSeq रिलीज 95 के अनुसार 93,664 जीवों को कवर करने वाले NCBI में जमा 3,52,904 जीनोम अनुक्रम हैं)। वर्तमान परिदृश्य में, औसतन प्रत्येक अनुमोदित दवा के गर्भाधान से बाजार तक लगभग 14 साल लगते हैं और इसकी लागत लगभग \$ 2.6 बिलियन है। यदि कम्प्यूटेशनल विधियां पर्याप्त रूप से परिपक्व हो जाती हैं, तो लागत और समय के साथ-साथ दवा अनुमोदन प्रक्रिया में तेजी से कमी की उम्मीद है। जैविक डेटा और कम्प्यूटेशनल पावर में वृद्धि, एक स्वचालित कंप्यूटर एडेड ड्रग डिस्कवरी पाइपलाइन बनाने के लक्ष्य को सक्षम बनाता है, जो किसी भी रोगजनक के किसी भी लक्ष्य प्रोटीन (अध्याय 1) के खिलाफ हिट अणुओं की एक सूची उत्पन्न कर सकता है।

इस थीसिस ने 'धन्वंतरी' (http://scfbio-iitd.res.in/software/dhanvantari_new/Home.html) नाम की एक ऐसी दवा खोज पाइपलाइन के स्वचालन का वर्णन किया है, जो कम्प्यूटेशनल रूप से 'जीनोम' से सभी एंटी पॉइंट्स को 'हिट अणु' तक जोड़ता है (अध्याय 2)। पाइपलाइन किसी भी रोगजनक कीटाणु के जीनोम अनुक्रम के इनपुट से शुरू होती है और

जीवन की केंद्रीय हठधर्मिता का उपयोग करके संबंधित प्रोटीन अनुक्रमों की भविष्यवाणी करती है। एक अच्छा दवा लक्ष्य, इसकी असली सक्रिय साइट की भविष्यवाणी के अनुक्रमिक चरणों पर निर्भर है। जिसके बाद पाइपलाइन रोगजनक के खिलाफ संभावित हिट अणुओं की पहचान करने में मदद करता है। यदि उपलब्ध है, तो पाइपलाइन में वैकल्पिक रूप से एक जीन अनुक्रम, प्रोटीन अनुक्रम या लक्षित प्रोटीन की तृतीयक संरचना का इनपुट दिया जा सकता है, हिट अणुओं की पहचान करने के लिए उनके खिलाफ। पाइपलाइन रोगजनक जीनोम पर काम करती है। हालांकि, जीन स्तर से पाइपलाइन मानव सहित सभी जीवों पर काम कर सकती है। प्रोटीन स्तर से 33 प्रोटीन दवा लक्ष्य और जीनोम स्तर से 3 रोगजनकों पर कम्प्यूटेशनल परीक्षण के माध्यम से एफडीए से प्रयोगात्मक डेटा पर पाइपलाइन अच्छी तरह से मान्य है। पाइपलाइन परीक्षण के जीवों और ड्रग लक्ष्यों के खिलाफ उपलब्ध एफडीए द्वारा अनुमोदित अधिकांश दवाओं को ब्रैकेट करने में सक्षम थी, जो पाइपलाइन के काम में विश्वास पैदा करती है।

इसके अलावा, पाइपलाइन की कार्यशील प्रभावकारिता को मान्य करने के लिए, दो अलग-अलग मामले के अध्ययन किए जाते हैं, एक हेपेटाइटिस बी वायरस के खिलाफ और दूसरा हेपेटाइटिस ए वायरस के खिलाफ। इन दोनों वायरस के मामले में, केवल निवारक टीके आज तक उपलब्ध हैं और एफडीए द्वारा अनुमोदित छोटे अणु दवा उपलब्ध नहीं है (हालांकि, एचबीवी के मामले में कुछ न्यूक्लियोसाइड एनालॉग उपलब्ध हैं)। पाइपलाइन का उपयोग एंटी-एचबीवी

और एंटी-एचएवी छोटे अणु अवरोधकों को उनके लक्षित प्रोटीन के खिलाफ खोजने के लिए किया गया है। मौजूदा दवा खोज समस्याओं को हल करने में इसकी प्रयोज्यता को प्रदर्शित करने के लिए पाइपलाइन के कम्प्यूटेशनल निष्कर्षों को प्रयोगात्मक रूप से मान्य किया गया था।

HBV के लिए, दो प्रोटीन लक्ष्य HBeAg और HBsAg को लक्षित किया गया है (अध्याय 3)। अभी तक किसी भी छोटा अणु अवरोधक HBsAg प्रोटीन के उत्पादन को बाधित करने की सूचना नहीं है। हमारा पहला ऐसा अध्ययन है। HBsAg के खिलाफ कुछ अन्य रिपोर्ट किए गए अवरोधक केवल इसके स्राव को रोकने में सक्षम हैं। एचबीवी संक्रमण के खिलाफ आशाजनक हिट के रूप में 'धनवंतरी' प्रोटोकॉल का उपयोग करके आठ अणुओं की पहचान की गई थी। इन अणुओं ने अवधारणा के प्रमाण प्रदान करने वाले प्रयोगात्मक अस्सेस (assays) में sub-micromolar पैमाने पर अवरोध दिखाया, जो प्रोग की दृष्टि से बेहद अच्छा है।

एचएवी के मामले में, जो वायरस के पिकोर्नवीरिडे परिवार से संबंधित है, पाइपलाइन का उपयोग वायरस के इस परिवार के खिलाफ जेनेरिक इनहिबिटर डिजाइन करने के लिए किया गया था (अध्याय 4)। यहां, 3 सी प्रोटीज को लक्षित किया गया है क्योंकि यह प्रोटीन इस परिवार के अधिकांश वायरस में मौजूद है। 'धनवंतरी' प्रोटोकॉल का उपयोग करते हुए सात अणुओं की पहचान की गई जो एचएवी संक्रमणों के खिलाफ आशाजनक हिट थे। इसके अलावा, दो अणुओं के एंटी-एचएवी गुणों को विकसित करने के लिए डिज़ाइन किया गया है। सभी नौ

अणुओं ने अवधारणा के प्रमाण प्रदान करने वाले प्रयोगात्मक assays में कम सूक्ष्म दाढ़ (sub-micromolar) पैमाने पर अवरोध दिखाया।

वायरल लक्ष्यों के खिलाफ हिट अणुओं की सूची की पहचान करने के लिए कम्प्यूटेशनल the “जीनोम टू हिट” पाइपलाइन के आवेदन में तेजी लाने के लिए, वायरल प्रोटीन के संरचनात्मक और कार्यात्मक एनोटेशन की जानकारी रखने वाले डेटाबेस की आवश्यकता है। इसलिए, 5,930 मानव मेजबान वायरल प्रोटीन का एक व्यापक डेटाबेस उत्पन्न होता है, जिसका नाम *vi*Humans (<http://scfbio-iitd.res.in/viHumans/>) (अध्याय 5) है। डेटाबेस में 1,116,019 वायरल प्रोटीन वाले 1,116 मानव मेजबान वायरस शामिल हैं। डेटाबेस में इन वायरस, जीन और प्रोटीन के नाम, प्रोटीन अनुक्रम, मॉडलनेबिलिटी इंडेक्स, अनुक्रम स्तर एनोटेशन, रास्ते शामिल, टैक्सोनोमिक वंशावली आदि के जीनोम परिग्रहण संख्या के बारे में जानकारी प्रदान की जाती है। डेटाबेस में प्रोटीनों की संरचना, कार्य, तंत्र / मार्ग के बारे में और उनके ज्ञात और अनुमानित लिगेंड के साथ कार्यात्मक साइट के बारे में अनुमानित जानकारी दोनों उपलब्ध हैं। इसलिए, *vi*Humans डेटाबेस 5930 वायरल प्रोटीन की संरचना आधारित दवा खोज शुरू करने के लिए आवश्यक सभी जानकारी के लिए एक व्यापक भंडार के रूप में कार्य करता है।

इस थीसिस कार्य का समग्र फोकस विभिन्न स्तरों पर विभिन्न केस स्टडीज (अध्याय 6) का उपयोग करके प्रयोगात्मक मान्यताओं के साथ एक पूर्ण ड्रग डिजाइन सॉफ्टवेयर सूट का अविष्कार है।

Contents

<i>Certificate</i>	i
<i>Acknowledgements</i>	ii
<i>Abstract in English</i>	v
<i>Abstract in Hindi</i>	viii
<i>List of Figures</i>	xvii
<i>List of Tables</i>	xxiii

Chapter 1	Introduction	1-29
1.1.	Human diseases	1
	1.1.1. Viruses	2
	1.1.2. Bacteria	2
	1.1.3. Fungi	3
	1.1.4. Protozoa	3
	1.1.5. Others	4
1.2.	Introduction to computer aided drug design	4
	1.2.1. CADD at nucleotide (DNA/RNA) level	5
	1.2.2. CADD at proteome level	5
1.3.	Structure Based Drug Design	6
1.4.	Ligand Based Drug Design	7
1.5.	Biological Databases	8
	1.5.1. DNA Sequence Databases	8
	1.5.2. Protein Sequence Databases	9
	1.5.3. Protein Function Related Databases	9
	1.5.4. DNA/Protein Structure Related Databases	10
	1.5.5. Metabolic Pathways Related Databases	10
	1.5.6. Protein-Ligand Complex Related Databases	11
	1.5.7.Small Molecules Databases	11

1.6.	Available tools and works in CADD	12
1.7.	Scope of thesis	15
	1.7.1. Dhanvantari: A comprehensive web tool connecting genome to hits identification	16
	1.7.2. Successful Applications	17
	1.7.3. Viral Database	17
	1.7.4. Improvements by implementing <i>in silico</i> toxicity predictions	18
1.8.	References	20
Chapter 2	Automation of computer aided drug discovery pipeline from genomes to hits	30-75
2.1.	Introduction	30
2.2.	Materials and Methods	32
	2.2.1. Five different entry points within the pipeline	32
	2.2.2. Working of pipeline	34
	2.2.3. Web-front of the pipeline	37
2.3.	Results and Discussion	38
	2.3.1. Validation from genome to hits	38
	2.3.2. Validation from proteins to hits	44
	2.3.3. Case study available at <i>Dhanvantari</i> webserver	50
2.4.	Conclusions	51
2.5.	References	52
Chapter 3	Identification of hit molecules against precore (HBeAg) and surface (HBsAg) proteins of Hepatitis B virus	76-103
3.1.	Introduction	76
3.2.	Materials and Methods	78
	3.2.1. Protein target identification	79
	3.2.2. Generation, optimization of 3D structures and identification of active sites	80
	3.2.3. Screening of million compounds	80
	3.2.4. Docking and scoring	80
	3.2.5. Molecular dynamics simulations	81

	3.2.6. Binding free energy calculations	82
3.3.	Results and Discussion	82
	3.3.1. Protein target identification	82
	3.3.2. 3D Structure of HBeAg and HBsAg and identification of active sites	83
	3.3.3. Screening of million compounds	86
	3.3.4. Docking and scoring	86
	3.3.5. Molecular dynamics simulations	86
	3.3.6. Experimental validation of computational predictions	89
	3.3.7. Binding mode analysis of molecules 6 and 7	90
	3.3.8. Comparison of molecules 6 and 7 with other known inhibitors	93
3.4.	Conclusions	94
3.5.	References	96
Chapter 4	Identification of hit molecules against 3C protease protein of Hepatitis A virus	104-146
4.1.	Introduction	104
4.2.	Materials and Methods	107
	4.2.1. Protein target identification	108
	4.2.2. 3D Structure of 3C protease and identification of active site	109
	4.2.3. Screening of million compounds	109
	4.2.4. Docking and scoring	110
	4.2.5. Molecular dynamics simulations	110
	4.2.6. Binding free energy calculations	111
4.3.	Results and Discussion	112
	4.3.1. Protein target identification	112
	4.3.2. 3D Structure of 3C protease and identification of active site	112
	4.3.3. Screening of million compounds	113
	4.3.4. Docking and scoring	113
	4.3.5. Molecular dynamics simulations	117
	4.3.6. Binding mode analysis of best leads	123

4.3.7.	Inhibition of HRV 3C by compounds 1-9 is comparable to that seen for HAV 3C	126
4.3.8.	Inhibition of 3Cproteases of other Picornaviridae family members	128
4.4.	Conclusions	135
4.5.	References	137
Chapter 5	Development of viral protein database for human host viruses	147-173
5.1.	Introduction	147
5.2.	Materials and Methods	150
	5.2.1. Basic and taxonomic lineage data collection	151
	5.2.2. Sequence level annotations	152
	5.2.3. Function level predictions and compilations	154
	5.2.4. Structure level prediction and compilations	155
	5.2.5. Binding pocket level annotation	157
5.3.	Results and Discussion	159
	5.3.1. How to use/Search options and routes to access database	159
	5.3.2. Download options and predicted information	162
	5.3.3. <i>Post facto</i> database annotations	162
5.4.	Conclusions	164
5.5.	References	166
Chapter 6	Summary & Perspectives	174-176
6.1	Summary	174
	APPENDIX	177-194
	<i>Publications</i>	195
	<i>Bio- data</i>	196-200

List of Figures

- Fig. 1.1. Pie chart shows the percentage contribution of each class of pathogen towards mortality rate due to human infectious diseases^{3,4}. The percentage is calculated by including all the reported number of deaths per year by WHO for diseases caused by each class of pathogen. 1
- Fig. 1.2. Distribution of macromolecular drug targets based on a review published by Santos *et al.* in 2016 *Nature Review Drug Discovery*³⁴. Proteins show majority in being drug targets under different categories such as G-protein coupled receptors GPCRs, enzymes, ion channel proteins, nuclear receptors and transporters with only small percentage of share for nucleic acids. 6
- Fig. 1.3. Steps involved in identifying a drug molecule using SBDD process. The steps mentioned in dark blue tiles cover SBDD process and the light blue tiles cover experimental steps for converting the drug candidate to an approved drug. 7
- Fig. 1.4. Steps involved in identifying a drug molecule using LBDD process. The steps mentioned in dark blue tiles cover LBDD process and the light blue tiles cover experimental steps for converting the drug candidate to an approved drug. 8
- Fig. 2.1. A) Bar graph showing the number of genomes sequenced per year (Source NCBI^{2,20}). B) Pie chart depicting the difference between known protein sequences and their structures followed by the difference between known versus unknown active sites among all the known protein structures (Source: PDB³ and BioLiP Database). C) Bar graph showing less than 34 FDA approvals per year (Source: DrugBank²¹ and FDA). D) Bar graph highlighting the exponential difference between the number of diseases versus FDA approved drugs against them (Source: FDA). 31
- Fig. 2.2. A schematic representation of the work-flow (methodology) for the automated “genome to hit” computational pipeline 33
- Fig. 2.3. Webfront of Dhanvantari pipeline available at http://www.scfbio-iitd.res.in/software/dhanvantari_new/Home.html 37
- Fig. 2.4. Pictorial representation of a case study against *S. aureus*. Four antibiotics viz. Moxifloxacin, Ciprofloxacin, Norfloxacin and Novobiocin are targeted from the genomic level via the pipeline. As seen from the blue and red boxes on the right side of the figure, the blue box depicts that the FDA drug was filtered as a hit molecule against the particular cavity among the top 10 (namely S1 to S10). The true cavity identified was cavity number S1. 39
- Fig. 2.5. (A) Histogram showing total number of FDA approved drugs against some major diseases and the number of true positives (hits) and true negatives (non-hits) obtained via the *Dhanvantari* pipeline, at a -5 kcal/mol threshold for predicted binding energy values. (B) List of 44

- some major life-threatening diseases and their protein targets used for validation of *Dhanvantari* pipeline.
- Fig. 3.1. HBV genome and four ORFs, namely P, S, C and X along with their location within the genome. The 'S' ORF is shown in green color spanning from 2850-837 bp, forming three proteins large-, middle- and small-surface proteins. The 'P' ORF is shown in red color spanning from 2309-1625 bp, forming the polymerase protein. The 'P' ORF is the longest ORF in HBV. The 'X' ORF is shown in light blue color spanning from 1376-1840 bp, forming HBxAg protein. The 'C' ORF is shown in purple color spanning from 1816-2454 bp, forming two proteins HBeAg and HBcAg. S domain is common in all surface proteins. Large surface protein has PreS1, PreS2 and S domains. Middle surface protein has PreS2 and S domains. Small surface protein (HBsAg) has S domain. Polymerase protein has RT domain. HBeAg protein has PreC domain. 77
- Fig. 3.2. Methodology employed for the identification of inhibitors against HBeAg and HBsAg proteins. The protocol starts from identification of protein sequences of the target proteins, followed by their model generation, active site identification, screening, docking and molecular dynamics simulation studies leading to eight molecules. All the eight proposed molecules were tested and two best molecules were selected on the basis of experimental findings of multiple protein inhibition. 79
- Fig. 3.3. (A) Overlap of the crystal structure of HBcAg protein, 1QGT (magenta) and modeled HBeAg protein (blue) (B) Validation of modeled structure of HBeAg protein using ProtSAV tool. (C) Active site residues shown in stick model (deep green) for HBeAg protein (shown as cartoon in purple). The residues common to HBsAg protein are labeled in red and the residues similar to binding pocket of HBsAg protein are labeled in blue. 83
- Fig. 3.4. Active Site residues of core antigen HBsAg protein are shown in this figure. The binding pockets of proteins HBcAg and HBeAg are same with just the residue number change. The residues common to HBsAg protein are labeled in red and the residues similar to binding pocket of HBsAg protein are labeled in blue. 84
- Fig. 3.5. (A) The superimposed structures are of initial modeled HBsAg protein (blue) on optimized HBsAg protein (green). (B) ProtSAV score has improved from orange (5-8 Å RMSD range) for initial modeled structure of HBsAg protein to yellow (2-5 Å RMSD) range indicating that the optimized structure is now within acceptable limits. (C) Active site residues shown in stick model (green) for HBsAg protein (shown as cartoon in yellow). (D) Ramachandran Plot of HBsAg protein showing allowed and disallowed amino acid percentage within the optimized structure. 85
- Fig. 3.6. A 2D representation of the interaction profiles for molecules 1-8 with HBeAg protein generated using LigPlot+⁵¹. Two interaction poses are 87

- overlapped, one is the initial docked pose (dark blue residues) and the other is the biggest cluster pose during the simulation (black residues). The red circles show the common residues involved in hydrophobic contacts. The green dotted lines show hydrogen bonds along with their bond lengths and the residues involved are labeled in green color.
- Fig. 3.7. Convergence plots of ligand rmsd (red) and C α backbone atoms of protein-ligand complex (black) of HBeAg protein show molecules 6 and 7 show lesser backbone and ligand fluctuations. The plots are generated using Origin 4.0. 89
- Fig. 3.8. Convergence plots of ligand rmsd (red) and C α backbone atoms of protein-ligand complex (black) of molecules 6 and 7 with HBsAg protein. The plots are generated using Origin 4.0. 91
- Fig. 3.9. The interaction poses for molecule 6 and 7 with HBsAg protein. The biggest cluster pose during the trajectory run shows molecule 6 tends to get stabilized further more than its initial docking conformation with an increase in one hydrogen bond with Gln54. The molecule 7 stays stable within the cavity keeping intact the hydrogen bond with a backbone O atom of Ala45. Both molecules have a favorable conformation within the active site cavity showing stable hydrophobic and hydrogen bonding during most of the 100 ns molecular dynamics simulations. 91
- Fig. 3.10. Plots show distance between donor and acceptor of residues and molecules interacting in formation of hydrogen bonds. (A) The hydrogen bond distance between heavy atoms of molecule 6 and Ser135, Trp131 and Tyr147 residues of HBeAg protein. (B) The hydrogen bond distance between heavy atoms of molecule 7 and Tyr147, Ser135 and Trp131 residues of HBeAg protein. (C) The hydrogen bond distance between heavy atoms of molecule 6 and Asn52 and Gln54 residues of HBsAg protein. (D) The hydrogen bond distance between heavy atoms of molecule 7 and Cys76 and Ala45 residues of HBsAg protein. The hydrogen bond distance is calculated between heavy atoms and is within the acceptable limits (cut off of 3 Å donor-acceptor heavy atoms distance and cut off of 120 degrees for bond angles) 53. 92
- Fig. 4.1. (A) Autoproteolytic cleavage of HAV polyprotein by its 3C protease (shown with red arrows), (B) Preferred cleavage site of HAV 3C after glutamine, (C) Cellular proteins cleaved by HAV 3C (PTBP- polypyrimidine tract binding protein, IKK γ - inhibitor of nuclear factor kappa-B kinase subunit gamma, PABP- polyA binding protein) and (D) Active site of HAV 3C comprising Cys172, His44 and Asp84 as a catalytic triad. 106
- Fig. 4.2. Methodology employed for screening, designing, synthesis and *in vitro* validation of inhibitors against HAV 3C protease. The protocol starts from obtaining 3D coordinates information from crystal structure, followed by active site-based ligand screening, docking and 108

- molecular dynamics simulation studies leading to seven proposed molecules from ZINC database. Simultaneous iterative modifications in isatin-based scaffold lead to identification of two best binders which were then synthesized. All the proposed molecules were tested against HAV 3C protease via enzyme assays.
- Fig. 4.3. ROC curves for four target enzymes (ACHE, ALR2, AMPC and HIVPR), show threshold of -10 kcal/mol is most significant to increase the robustness of docking protocol. The ParDOCK docking protocol was able to remove ~99% of the decoys when the threshold was set to -10 kcal/mol. The ligand and decoys set was taken from the DUD site ⁵⁰. The ROC curve shows that -10 kcal/mol is an optimum cut off where both the sensitivity and specificity of the docking protocol is better. 114
- Fig. 4.4. (A) Interaction patterns of initial docked pose and one frame among the biggest cluster of ligand from 100 ns long molecular dynamics simulations (complexed with compounds 1-9). The overlapping 2D poses show hydrophobic and hydrogen bond contacts with the active site residues of HAV 3C protease between docking versus post docking stable conformations. The red circles show the common interacting residues between docking and biggest cluster frame (BCF) of the respective protein-ligand complex. Blue color denotes hydrophobic residues of the initial docked pose. Red color residues (labeled in black) denotes hydrophobic residues of the BCF. Green color residues are the hydrogen bonding residues. Blue lines and green lines show hydrogen bonds in docked pose and BCF respectively. Most of the residues are found to be conserved maximum number of times during 100 ns molecular dynamics simulations. These plots have been generated using LIGPLOT software ⁵⁸ (B) 3D interactions of the most potent inhibitors (compounds 6, 8 and 9) in the catalytic site of HAV 3C protease. 118
- Fig. 4.5. Convergence plots of RMSD of ligands and backbones of HAV 3C proteases bound with compounds 1-9 for a run length of 100 ns in molecular dynamics simulation studies. The plots show an overall stability in the trajectories indicating that the protein-ligand complexes (protein backbone along with compounds) were in their energetically stable conformations throughout the molecular dynamics simulations. Compound 5 shows large ligand RMSD fluctuations but overall stable backbone RMSD. 119
- Fig. 4.6. Binding mode analyses of compound 5 versus compounds 7 and 9 in complex with HAV 3C protease. Distances between A) Cys172 and central atom of compound 5, B) Cys172 and central atom of compound 7 with an extension of 50 ns molecular dynamics simulations to check the stability after 90ns peak jump, C) Cys172 and central atom of compound 9 within the binding pocket throughout the 100 ns molecular dynamics simulations. Binding mode analyses of compound 5 versus compound 7 shows that the 122

- distance between the catalytic site residue Cys172 increases with respect to time during the molecular dynamics simulations. D) 3D ball and stick pictorial representation of ligand in active site residues using PyMOL, shows the drift of compound 5 from docked position (shown in blue color) to most favorable (biggest cluster) location (shown in green color) during molecular dynamics simulations.
- Fig. 4.7. Plot showing number of hydrogen bonds formed during the 100 ns molecular dynamics simulations along with fluctuation of hydrogen bond distance as a function of time for the hydrogen bonds forming residues for the compounds 6, 8 and 9. The hydrogen bond distance is calculated between heavy atoms and is within the acceptable limits (cut off of 3 Å donor-acceptor heavy atoms distance and cut off of 120 degrees for bond angles) 61. 124
- Fig. 4.8. Specific sub-sites of HAV 3C protease 20. Residues colored in blue belong to HAV 3C protease. The sub-sites are named as has been reported 20 where hydrogen bonding interactions are reported for main chain N of Val144 with O of P4-Leu and main-chain O of Val144 with N of P2-Ala (or P2-Phe), the main-chain O of Gly194 with N of P3-Ala and the main-chain N of Gly194 with O of P3-Ala (or P3-Phe) and the main chain N of Gly170 with O of (P1-Gln). The work also mentions the S2 site to be formed by the side-chain atoms of His44, Phe48, Tyr143, His145 and Leu155 and the main-chain atoms of His44, Tyr143, Val144 and His145. (Yin et al. have shown in their work how these residues play roles at S1, S2, and S3 sub-sites in HAV 3C protease 20). 125
- Fig. 4.9. Superimposition of 3C proteases of picornaviruses showing similar active site residues (highlighted with blue circles). Red color residues belong to Hepatitis A virus (PDB ID: 2CXV³⁷), Dark blue color residues belong to Human Rhinovirus 14 (PDB ID: 2IN2⁶²), Green color residues belong to Foot and Mouth Disease Virus (PDB ID: 2BHG⁶³), Yellow color residues belong to Poliovirus (PDB ID: 1LIN⁶⁴), Cyan color residues belong to Coxsackievirus B3 (PDB ID: 2VB0³⁸). Stick model representation of compound 6 (interacting with HAV 3C protease). The figure is generated using PyMOL (<https://pymol.org/2/>). 127
- Fig. 4.10. Interaction patterns of the nine identified compounds 1-9 showing hydrophobic and hydrogen bond contacts with the active site residues of HRV14 3C protease (PDB ID: 2IN2). The interaction diagrams are of protein-ligand complexes obtained from the best binding docked poses. These plots have been generated using LIGPLOT software⁵⁸ 128
- Fig. 4.11 Interaction patterns of the three best proposed compounds 6, 8 and 9 showing hydrogen bonding and van der Waals contact network shown in 2D, depicting similar patterns of interactions within 3C proteases of intra-species of picornavirus family, namely Human Rhinovirus 14, Poliovirus, Coxsackievirus B3, Foot and Mouth Disease Virus, Enteroviruses EV-D68 and EV-A71 respectively. 131

	These plots have been generated using LIGPLOT software ⁵⁸ .	
Fig. 5.1.	A Venn diagram representation having unreviewed sequences in white background and reviewed proteins represented via grey circle. Different categories of information are represented with different colors. The number labeled under each category name is the total number of viral proteins, the labels on the outer side of each color circles represent the number of unreviewed proteins and inner circle labels show reviewed proteins having experimental information available for that category.	149
Fig. 5.2.	The work flow of database generation. The numbers in red highlight the number of viral proteins available/predicted at each step.	150
Fig. 5.3.	Number of predicted/modeled viral proteins falling under in three different score-bins of SDIndex obtained is shown.	153
Fig. 5.4.	Number of proteins predicted via I-TASSER and RaptorX individually as well as by both	156
Fig. 5.5.	Number of proteins for which active sites/binding pockets and ligands are predicted via I-TASSER and RaptorX. The prediction of active sites and ligands for experimental structures is also shown.	158
Fig. 5.6.	A depiction of information provided at <i>vi</i> Humans database	161
Fig. 5.7.	Post database generation structural, functional, binding pockets and hit molecules annotations have increased as seen by the pie charts above. The red color represents the experimentally known section, blue represents the predicted section and yellow denotes the section which was neither experimentally know nor could be computationally predicted. The panel shows percentage distribution of the 5930 viral proteins on the basis of four categories, namely 3D structure availability, active site(s) availability, function(s)/role(s) availability and small molecule binders/ligands availability.	163
Fig. 5.8.	Database schema of <i>vi</i> Humans. Primary keys are shown in blue highlighted boxes and the foreign keys (FK) are represented by italics and underlining.	164

List of Tables

Table 1.1.	List of software available under each category of CADD and their availability and usage is shown below.	13
Table 2.1.	Three case studies from genome to hits of the <i>Dhanvantari</i> pipeline, viz. <i>Staphylococcus aureus</i> , Human Immunodeficiency Virus and Influenza A virus. The Genome reference, Protein Ids, Protein name, RMSD values of the target proteins (experimental versus the modeled) and the cavity points identified compared with literature. The first ranked modeled proteins from the <i>Dhanvantari</i> pipeline were taken in each case.	40
Table 2.2.	Experimental and predicted binding free energy data for <i>Staphylococcus aureus</i> against available FDA drugs.	41
Table 2.3.	Experimental and predicted binding free energy data for HIV against available FDA drugs	42
Table 2.4.	Experimental and predicted binding free energy data for Influenza A Virus against available FDA drugs.	43
Table 2.5.	Compute resources and time consumption of each module of the genome to hit pipeline. The compute time (exclusive of the in queue timing) is noted for the jobs once they are in their running status	43
Table 2.6.	The validation of genome to hits pipeline on major deadliest (infectious and non-infectious) diseases including 33 protein targets and 111 FDA drugs. The FDA drugs known against each target were screened through the pipeline. Out of 111, 100 FDA drugs were filtered out as hits via pipeline.	45
Table 3.1.	Predicted free binding energies for all the eight Molecules against HBeAg protein calculated using ParDOCK (for docking score) and using AMBER (for average values during MD simulations). The comparison with other docking tools has been done to cross-validate the molecular binding efficacy.	88
Table 3.2.	IC ₅₀ values (μM) for lamivudine-sensitive Wild type (Wt) and lamivudine-resistant rtM204I mutant HBV encoded HBeAg and HBsAg.	90
Table 3.3.	Predicted free binding energies for molecule 6 and 7 against HBsAg protein calculated using ParDOCK (for docking score) and using AMBER (for average values during MD simulations). The comparison with other docking tools has been done to cross-validate the molecular binding efficacy.	90
Table 3.4.	Structural similarity based on Tanimoto coefficients between chemical compounds identified in published reports and molecule 6 and 7 identified in this study.	94
Table 4.1.	Some case studies (target enzymes) checked against DUD decoys dataset above and below threshold value of -5 kcal/mol, -8 kcal/mol, -10 kcal/mol and -12 kcal/mol. The decoy dataset and target information was taken from DUD download site ⁵⁰ . This check was	115

performed to validate the docking protocol. The specificity and sensitivity were calculated using the different True positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) for four target enzymes at various cut off values of threshold viz. -5 kcal/mol, -8 kcal/mol, -10 kcal/mol and -12 kcal/mol. The sensitivity and specificity at threshold of -10kcal/mol was found to be optimum.

Table 4.2.	Molecular structural formulas of compounds identified with molecular weights (in Daltons) and corresponding inhibition constants (K_i in μM , obtained experimentally) of HAV and HRV14 3C proteases.	116
Table 4.3.	IUPAC names and Zinc IDs of the compounds 1-9 identified against Hepatitis A Virus 3C protease.	117
Table 4.4.	Data on predicted binding free energies (in kcal/mol) of compounds 1-9 with HAV 3C protease using ParDOCK, AutoDock 59, SwissDock 60, MMBAPPL, MMGBSA and MMPBSA along with the experimentally obtained K_i values. The MMBAPPL, MMGBSA and MMPBSA scores are obtained from 100 ns molecular dynamics simulations run.	120
Table 4.5.	Active site residues and similarity scores for 3C proteases of 14 species of the <i>Picornaviridae</i> family predicted using PocketMatch ⁶⁵ .	129
Table 4.6.	Predicted binding free energies (in kcal/mol) of compounds 1-9 with 3C proteases from different members of picornaviruses, obtained after docking and scoring.	132
Table 4.7.	Computationally predicted ADME properties ⁶⁷ of the identified compounds 1-9.	133
Table 5.1.	Data on number of associated orders, families, genera, species and strains of human host viruses in case of reviewed and unreviewed protein sequences.	148
Table 5.2.	A brief summary of collection tools used in the <i>viHumans</i> database for gathering basic and taxonomic information of all the 1,116,019 viral proteins.	151
Table 5.3.	A brief summary of collection and identification/prediction tools used in the <i>viHumans</i> database for inhibitors/hit molecules information of all the 5930 viral proteins.	153
Table 5.4.	A brief description of function prediction software used for functional characterization.	155
Table 5.5.	A brief summary of the methods/tools used for features compilation/prediction at protein structure level along with their availability.	156
Table 5.6.	A brief summary of ligand binding site identification/prediction tools used in the <i>viHumans</i> database for ligand binding site prediction of all the 5248 modeled and 592 experimentally known structures.	157
Table 5.7.	A brief summary of collection and identification/prediction tools used in the <i>viHumans</i> database for inhibitors/hit molecules information of all the 5930 viral proteins.	158

Table 5.8. A summary of search keywords which can be used for a systematic and specific browsing of the database.

159