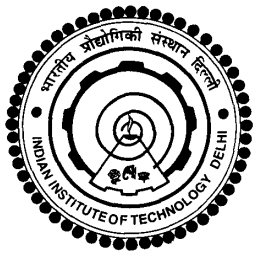
AUTOMATION OF COMPUTER-AIDED GENOMES TO HITS DISCOVERY PIPELINE AND ITS APPLICATION VIA CASE STUDIES

RUCHIKA BHAT



DEPARTMENT OF CHEMISTRY

INDIAN INSTITUTE OF TECHNOLOGY DELHI

OCTOBER 2019

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 $1/4^{th}$ 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 Baneerjee**, **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 *vi*Humans (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, *vi*Humans 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 है)। 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 रोगज़नक़ 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 एंट्री पॉइंट्स को 'हिट अणु' तक जोड़ता है (अध्याय 2)। पाइपलाइन किसी भी रोगज़नक़ 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 रोगज़नक़ 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। 'धन्वंतरि' 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

List of Figures

List of Tables