

***IN SILICO* IDENTIFICATION OF LEAD MOLECULES
FOR ER AND NSP2 PROTEINS AND INTERCALATORS
FOR DNA: SOME METHODOLOGICAL ADVANCEMENTS**

ANJALI SONI



**DEPARTMENT OF CHEMISTRY
INDIAN INSTITUTE OF TECHNOLOGY DELHI
AUGUST 2017**

©Indian Institute of Technology Delhi (IITD), New Delhi, 2017

***IN SILICO* IDENTIFICATION OF LEAD MOLECULES
FOR ER AND NSP2 PROTEINS AND INTERCALATORS
FOR DNA: SOME METHODOLOGICAL ADVANCEMENTS**

by

Anjali Soni

Department of Chemistry

Submitted

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

to the



Indian Institute of Technology Delhi

August 2017

Certificate

This is to certify that the thesis entitled, “***In silico* Identification of Lead Molecules for ER and nsP2 Proteins and Intercalators for DNA: Some Methodological Advancements**”, being submitted by **Ms. Anjali Soni** 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. Anjali Soni 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 other university or institute for the award of any degree or diploma.

Date

Dr. B. Jayaram

Professor

Department of Chemistry

Indian Institute of Technology Delhi

Acknowledgements

My Ph.D. journey is a life-changing experience and would not have been possible without the help of many people. Foremost, I am deeply indebted to my supervisor, Prof. B. JAYARAM, for introducing me to the most fascinating field of drug discovery. I would like to express whole-heartedly my sincere gratitude and appreciations for the effort he has made to groom me as a researcher since the day I joined him as a graduate student. He is one of the most intellectual and smartest people I know. I am amazed by the enthusiasm and drive that he has towards science and his presence has motivated me to love science. I hope that I could be as lively, enthusiastic and energetic as Prof. Jay and to someday be able to command an audience as well as he does. I thank him for his scientific advices, insightful discussions and suggestions about the research and for everything he taught me over past six years.

I am also grateful to my SRC members, head and faculties of Department of Chemistry and biological sciences, Indian Institute of Technology Delhi for helping and supporting me during my research career and for providing the necessary facilities in the department. 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 DBT & DST, Govt. of India.

I am thankful to Dr. Nasimul Hoda, Prof. N. G. Ramesh, Dr. Ashok Patel, Prof. A. Ramanan and Dr. Seema Bhatnagar for all the scientific discussions I had with them which led to fruitful results. I wish to express my gratitude for my seniors and colleagues at the SCFBio for making the environment cordial and for helping me whenever required. I am really grateful to my dear friends Dr. Mrinmoyee, Dr. Surojit, Dr. Bharat, Sandeep, Dr. Sweety,

Soumen, Dr. Kasinath, Vandana, Arabinda, Zeba and Satya for their great support and for making my stay at IIT Delhi fun and memorable.

I owe everything to my parents, sisters and brother. With their efforts and support I have reached this stage. I found them always standing by my side supporting me and loving and caring unconditionally.

I greatly appreciate the efforts from my husband and great friend Dr. Gohil for being a constant pillar of support and for his unconditional love, care, moral support and encouragement. Lastly I thank almighty for everything.

Anjali Soni

Abstract

This thesis focuses on *in silico* development and identification of novel inhibitors against estrogen receptor (ER) and non-structural protein 2 (nsP2). ER and nsP2 protein are well explored protein targets against breast cancer and chikungunya fever respectively, for identification of effective therapeutic agents. Additionally, this thesis focuses on development of novel methodologies for binding mode identification and free energy predictions against the DNA-ligand and protein-ligand complexes. The proposed algorithms are developed and validated on large diversified set of complexes.

This thesis comprises six chapters: Chapter 1 discusses the general overview of computer-aided drug design/discovery (CADD) processes, its approaches and applications in drug discovery pipeline. This chapter highlights the key features of CADD and its integral role in modern drug discovery approach. It also introduces briefly the concerns raised in drug design while targeting proteins and DNA as drug targets.

Chapter 2 involves *de novo* design and development of some novel inhibitors against ER, a target for breast cancer. Biphenyl scaffolds are chosen for designing the molecules as they are regarded as surrogates of the steroidal backbone. Guided by molecular docking, molecular dynamics simulations and free energy analyses, a few potent lead molecules are identified. These lead molecules identified are synthesized and assayed against breast cancer cell lines in collaboration. The molecules showed sub-micromolar activities building a synergy between computations and experiments.

Chapter 3 deals with identification of lead molecules against nsP2 protein of chikungunya virus. The nsP2 protein is crucial for the survival of virus due to the proteolytic activity residing within. Drug repurposing is employed to identify some potential drug

molecules targeting nsP2 protease from the library of ~3000 FDA approved drugs. Utilizing computer-aided molecular modelling, a few molecules are identified which are under procurement for performing experimental testing.

In chapter 4, *Bappl+* methodology is proposed for estimating the binding free energies of non-metallo and metallo protein-ligand complexes. *Bappl+* methodology is an extended version of our previous scoring functions (*Bappl* and *Bappl-Z*), and works very well in comparison to most of the reported state-of-the-art scoring functions. The methodology is validated against various datasets suggesting its generality and wider applicability.

Chapter 5 of the thesis describes a novel methodology (*Intercalate*) for an efficient prediction of ligand binding modes and binding free energies for DNA-ligand complexes. The methodology is designed to handle DNA-intercalators complexes binding non-covalently. A large number of complexes are used to develop and validate the *Intercalate* methodology. Given a DNA sequence and intercalation site information, *Intercalate* generates the 3D structure of DNA, creates the intercalation site, performs docking at the intercalation site and evaluates DNA–intercalator binding energy in an automated way. This drug-DNA intercalation methodology works with high accuracy and should prove useful in the discovery of potential intercalators for their use as anticancer compounds, antibacterials or antivirals.

Chapter 6 presents a summary and perspective of the work carried out in this thesis.

Overall, the thesis work through *Bappl+* (Chapter 4) and *Intercalate* (Chapter 5) accomplishes the task of upgrading and broadening the applicability of *Sanjeevini* software suite (<http://www.scfbio-iitd.res.in/sanjeevini/sanjeevini.jsp>) further validating its utility (Chapters 2 & 3) in lead molecule discovery.

यह शोध सिल्लिको विकास और एस्ट्रोजेन रिसेप्टर (ईआर) और गैर-संरचनात्मक प्रोटीन 2 (एनएसपी 2) के खिलाफ उपन्यास अवरोधकों की पहचान में केंद्रित है। प्रभावी चिकित्सीय एजेंटों की पहचान के लिए ईआर और एनएसपी 2 प्रोटीन का क्रमशः स्तन कैंसर और चिकनगुनिया बुखार के खिलाफ प्रोटीन लक्ष्य का पता लगाया गया है। इसके अतिरिक्त, यह थीसिस बाध्यकारी मोड पहचान और डीएनए-लिंगेड और प्रोटीन-लैंगेड परिसरों के खिलाफ मुफ्त ऊर्जा भविष्यवाणियों के लिए उपन्यास पद्धतियों के विकास पर केंद्रित है। प्रस्तावित एल्गोरिदम बड़े पैमाने पर परिसर के विविध सेट पर विकसित और मान्य हैं।

इस थीसिस में छह अध्याय शामिल हैं: अध्याय 1 कंप्यूटर-सहायता प्राप्त दवा डिजाइन / खोज (सीएडीडी) प्रक्रियाओं, उसके दृष्टिकोण और आवेदनों की दवा की खोज पाइपलाइन में सामान्य अवलोकन के बारे में चर्चा करता है। इस अध्याय में सीएडीडी की प्रमुख विशेषताओं और आधुनिक दवा की खोज के दृष्टिकोण में इसकी अभिन्न भूमिका पर प्रकाश डाला गया है। दवा के लक्ष्य के रूप में प्रोटीन और डीएनए को लक्षित करते हुए यह ड्रग डिजाइन में उठाए गए चिंताओं को संक्षेप में प्रस्तुत करता है।

अध्याय 2 में ईआर के खिलाफ कुछ उपन्यास अवरोधकों का विकास और विकास शामिल है, जो स्तन कैंसर के लिए एक लक्ष्य है। बिफेनील स्काॅफॉल्ड्स अणुओं को डिजाइन करने के लिए चुना जाता है क्योंकि उन्हें स्टेरायडल रिढ़ की रक्षा के रूप में माना जाता है। आणविक डॉकिंग, आणविक गतिशीलता सिमुलेशन और निःशुल्क ऊर्जा विश्लेषण द्वारा निर्देशित, कुछ शक्तिशाली लीड अणुओं की पहचान की जाती है। इन प्रमुख अणुओं को संश्लेषित और सहयोग में स्तन कैंसर सेल लाइनों के खिलाफ assayed हैं पहचान की। अणुओं ने कम्प्यूटेशन और प्रयोगों के बीच तालमेल बनाकर उप-माइक्रोलोरर गतिविधियों का प्रदर्शन किया।

अध्याय 3 चिकनगुनिया विषाणु की एनएसपी 2 प्रोटीन के खिलाफ लीड अणुओं की पहचान के साथ संबंधित है। एनएसपी 2 प्रोटीन, वायरस के अस्तित्व के लिए महत्वपूर्ण है, जो अंदर स्थित प्रोटियोलेटीक गतिविधि है। ~ 3000 एफडीए अनुमोदित दवाओं के पुस्तकालय से एनएसपी 2 प्रोटेज़ को लक्षित करने के लिए कुछ संभावित दवा अणुओं की पहचान करने के लिए ड्रग का पुनः प्रयोग किया जाता है। कंप्यूटर सहायता प्राप्त आणविक मॉडलिंग का उपयोग करना, कुछ अणुओं की पहचान की जाती है जो प्रयोगात्मक परीक्षण करने के लिए खरीद के अधीन हैं।

अध्याय 4 में, गैर-मेटलो और मेटलो प्रोटीन-लैंगेड परिसरों के बाध्यकारी मुक्त ऊर्जा के आकलन के लिए बैपल+ पद्धति का प्रस्ताव है। बैपल+ पद्धति हमारे पिछले स्कोरिंग कार्यों (बैपल और बैपल-जेड) का एक विस्तारित संस्करण है, और अधिकांश रिपोर्टिंग राज्य-के-कला स्कोरिंग कार्यों की तुलना में बहुत अच्छी तरह से काम करता है। इसकी व्यापकता और व्यापक प्रयोज्यता का सुझाव देने वाले विभिन्न डेटासेट के खिलाफ पद्धति को मान्य किया गया है।

थीसिस के अध्याय 5 में लिंगेड बाइंडिंग मोड के प्रभावी पूर्वानुमान और डीएनए-लैंगेड कॉम्प्लेक्स के लिए निःशुल्क ऊर्जा बंधन के लिए एक उपन्यास पद्धति (इंटरकेटेट) का वर्णन किया गया है। इस पद्धति को डीएनए-इंटरकैलेटर कॉम्प्लेक्स को गैर-गहन रूप से बंधन रखने के लिए डिज़ाइन किया गया है। इंटरकैलेट पद्धति

को विकसित और मान्य करने के लिए बड़ी संख्या में परिसरों का उपयोग किया जाता है। एक डीएनए अनुक्रम और अंतर साइट की जानकारी को देखते हुए, इंटरैलेट डीएनए की 3 डी संरचना तैयार करता है, इंटरसीलेशन साइट बनाता है, इंटरकनेशन साइट पर डॉकिंग करता है और स्वचालित तरीके से डीएनए-इंटरसेलेटर बाध्यकारी ऊर्जा का मूल्यांकन करता है। यह दवा-डीएनए अंतरण पद्धति उच्च सटीकता के साथ काम करती है और एंटीकैंसर यौगिकों, एंटीबायोटिक्स या एंटीवायरल के रूप में उनके उपयोग के लिए संभावित इंटरकैटर्स की खोज में उपयोगी साबित होनी चाहिए।

अध्याय 6 इस थीसिस में किए गए कार्य के सारांश और परिप्रेक्ष्य को प्रस्तुत करता है।

कुल मिलाकर, बैपल+ (अध्याय 4) और इंटरकैलेट (अध्याय 5) के माध्यम से थीसिस का काम संजीवनी सॉफ्टवेयर सट (<http://www.scfbio-iitd.res.in/sanjeevini/sanjeevini>) के प्रयोज्यता के उन्नयन और विस्तार का कार्य पूरा करता है। जेएसपी) आगे की उपयोगिता (अध्याय 2 और 3) को लीड अणु डिस्कवरी में मान्यता प्रदान करता है

Contents

<i>Certificate</i>	i
<i>Acknowledgements</i>	ii
<i>Abstract</i>	iv
<i>List of Figures</i>	ix
<i>List of Tables</i>	xii
Chapter 1 Introduction.....	1-33
1.1 Computer-aided drug design: an overview.....	2
1.1.1 Structure-based drug design (SBDD).....	5
1.1.2 Ligand-based drug design (LBDD).....	8
1.1.3 Virtual Screening.....	9
1.1.4 Docking.....	14
1.1.5 Scoring functions.....	17
1.1.6 Molecular Dynamics (MD) simulation.....	20
1.2 Proteins – as drug targets.....	23
1.3 DNA – as a drug target.....	25
1.4 Scope of the thesis.....	26
1.5 References.....	27
Chapter 2 Design of subtype selective biphenyls against estrogen receptor.....	34-51
2.1 Introduction.....	35
2.2 Materials and Methods.....	37
2.2.1 Docking studies.....	37
2.2.2 MD simulation studies.....	38
2.3 Results and Discussion.....	39
2.4 Conclusions.....	48

2.5	References.....	50
Chapter 3 Lead molecule identification against nsP2 protease of chikungunya virus through drug repurposing.....52-70		
3.1	Introduction.....	53
3.2	Materials and Methods.....	56
3.2.1	Flow chart of the methodology.....	56
3.2.2	Molecular dynamics (MD) simulation protocol.....	57
3.3	Results and Discussion	57
3.3.1	Active site identification of nsP2 protease of CHIKV	57
3.3.2	Virtual screening and molecular docking	58
3.3.3	MD simulations based analyses.....	60
3.4	Conclusions.....	64
3.5	References.....	65
<i>Annexure</i>		68
Chapter 4 Improving the binding free energy estimations for non-metallo and metallo protein-ligand complexes.....71-105		
4.1	Introduction.....	72
4.2	Materials and Methods.....	74
4.2.1	Dataset preparation.....	74
4.2.2	Quantum calculations on metallo-PL complexes.....	76
4.2.3	Scoring function methodology.....	77
4.3	Results and Discussion.....	82
4.4	Conclusions.....	87
4.5	References.....	88
<i>Annexure</i>		92
Chapter 5 Methodology development for prediction of structures and energetics of DNA-intercalator complexes.....106-160		

5.1	Introduction.....	107
5.1.1	Principal modes of DNA-binding molecules.....	107
5.1.2	Modes of DNA-Intercalation.....	108
5.2	Materials and Methods.....	110
5.2.1	Creation of DNA structure with an intercalation site(s).....	110
5.2.2	Dataset preparation.....	113
5.2.3	MD simulations protocol.....	114
5.2.4	Docking methodology.....	115
5.2.5	Scoring methodology.....	117
5.3	Results and Discussion	122
5.3.1	Performance of Intercalate methodology on creating DNA structure with intercalation site(s).....	122
5.3.2	Performance of Intercalate methodology on docking experiments.....	123
5.3.3	Performance of Intercalate methodology on binding energy estimations.....	127
5.3.4	Free energy analyses of DNA-intercalation complexes.....	129
5.4	Conclusions.....	133
5.5	Intercalate methodology as a webserver.....	134
5.5.1	Workflow.....	134
5.5.2	Input.....	135
5.5.3	Output.....	135
5.6	References.....	136
	<i>Annexure</i>	141
	Chapter 6 Summary & Perspectives.....	161-164
6.1	Summary.....	162
6.2	Perspectives and score for future work.....	164
	<i>List of Publications & Patents</i>	165
	<i>Curriculum Vitae of Author</i>	166

List of Figures

Fig. 1.1 Some FDA approved drugs which came out of CADD.....	3
Fig. 1.2 A schematic representation of CADD approaches.....	4
Fig. 1.3 A cartoon representation of the binding mode of phosphotyrosine inhibitor identified using structure-based/molecular modeling approaches.....	6
Fig. 1.4 A schematic picture of the ligand binding pocket of estrogen receptor bound to an inhibitor (PDB-3ERT).....	7
Fig. 1.5 A schematic of virtual screening approach.....	10
Fig. 1.6 Types of scoring functions.....	17
Fig. 1.7 A molecular mechanics potential function representing various interactions in a molecule.....	20
Fig. 1.8 A simplified schematic of MD simulations protocol.....	22
Fig. 1.9 Proportion of drug targets and drug molecules targeting human proteins according to a recent study.....	24
Fig. 2.1 Biphenyl derivatives of interest.....	37
Fig. 2.2 A comparative depiction of interactions of biphenyl compounds 3(a-d) with the active site residues of ER α . The violet color represents various biphenyl compounds. Dashed bonds represent the hydrogen bonds along with the distances whereas arcs represent VDW and hydrophobic contacts. Circled residues represent hydrophobic side chain interactions with the corresponding ligands which are also observed in ER α complexed with tamoxifen. All the snapshots for the above analyses are taken from the last 10ns of the MD trajectories.....	40
Fig. 2.3 (A) RMSD patterns seen (with reference to initial structure) during the course of MD simulations with compounds 3(a-d) bound to ER α . (B) RMSD patterns of compound 3b bound to ER α and ER β together with RMSD of the control ER α -Tam.....	42
Fig. 2.4 A comparison of per-residue binding free energy for key residues between ER α -3b and ER β -3b.....	46
Fig. 2.5 (A) A superposition of complex ER α -3b (cyan) with ER α -Tam (PDB: 3ERT) (magenta). (B) A superposition of complex ER β -3b (cyan) with ER β -THC (PDB: 1L2J) (magenta). For clarity all hydrogens are removed and only a few important amino acids interactions are displayed to highlight selectivity. Yellow dotted bonds represent hydrogen bonds.....	47

Fig. 3.1 The lifecycle of CHIKV in the host cell.....	54
Fig. 3.2 Depicts the active site of nsP2 protease of CHIKV (PDB: 3TRK) containing the cysteine-histidine dyad, the interchangeable serine and tryptophan.....	55
Fig. 3.3 Steps involved in the identification of drug molecules against CHIKV nsP2 protease.....	57
Fig. 3.4 Molecular structural formulas of the nsP2 inhibitors chosen for further experimental studies.....	59
Fig. 3.5 RMSDs of the simulated complexes during production run.....	60
Fig. 3.6 Interactions of the inhibitors with the surrounding active site residues of nsp2 protease. Hydrogen bonds are shown as red-dashed bonds.....	62
Fig. 4.1 A computational flowchart adopted in <i>Bappl+</i> for estimating binding free energies.....	75
Fig. 4.2 Correlation plots between the experimental and predicted binding affinities for all the test datasets (A) 2007 core dataset (B) 2013 core dataset and (C) 2016 core dataset.....	85
Fig. 4.3 Correlation plot between experimental and predicted binding affinities for a series of ligands against (A) Trypsin, (B) ER, (C) UPA, (D) BACE-1, (E) HIV.....	86
Fig. 5.1 A cartoon representation of modes of drug intercalation to DNA. 3FT6 and 224D are the protein data bank (PDB) entries of DNA-intercalator complexes which illustrate proflavine binding at CpG as a classic (parallel) intercalator and nogalamycin binding at TpG as a threading (perpendicular) intercalator.....	109
Fig. 5.2 A representation of the thermodynamic cycle adopted for the computation of DNA-intercalator binding free energies.....	121
Fig. 5.3 Root mean square deviations (RMSDs) of 58 MD DNA and canonical BDNA structures with respect to their crystal structures.....	122
Fig. 5.4 A schematic representation of the intercalation site(s) created for MD DNA (green) and canonical BDNA (magenta) which are superimposed on crystal DNA (orange) for PDB - 282D and 1D12. (A) 282D (having one intercalation site) - MD DNA is closer to the crystal DNA (RMSD 1.0 Å) than canonical (RMSD 1.2 Å). (B) 1D12 (having two intercalation sites) - MD DNA (RMSD 1.0 Å) is better than canonical (RMSD 1.3 Å).....	123
Fig. 5.5 (A) Local RMSDs of best docked poses for various DNA models within 5 Å distance from the centre of mass of the planar ligand chromophore are represented. RCSB DNA dockings contain a few additional structures comprising modified residues. (B) RMSDs of ligand for the best docked structures with respect to their crystal structures.....	125

Fig. 5.6 A schematic representation of two docked configurations (top ranked and best pose) of ligands superimposed on the crystal ligand pose for PDB 1Z3F (A-D) and 1D54 (E-H). Green shows the crystal ligand configurations whereas blue and pink show top ranked and best pose of ligands respectively. (A & E) RCSB DNA docking (side view); (B & F) RCSB DNA docking (top view); (C & G) MD DNA docking (top view); (D& H) Canonical BDNA docking (top view). In (C, D, G & H) dockings, the DNA structures are generated explicitly from the DNA sequences and hence two DNA cartoons can be observed.....126

Fig. 5.7 A correlation between the experimental ($\Delta G^{\circ}_{\text{exp}}$) and predicted binding energies ($\Delta G^{\circ}_{\text{pred}}$) for 43 complexes ($R^2 = 0.69$; $R = \sim 0.83$). All binding energies are in kcal/mol. The experimental binding energies ($\Delta G^{\circ}_{\text{exp}}$) are extracted from the binding constant data by applying $\Delta G^{\circ} = -RT \ln K$128

Fig. 5.8 Histograms representing a consensus outlook of the energetics of 43 DNA-intercalator complexes through various energy components contributing to binding free energy using MM-GB/PBSA and NMODE. (A) Direct energy components; (B) Net energy components.....132

Fig. 5.9 The flow chart of Intercalate webserver.....135

List of Tables

Table 1.1 List of some popular protein tertiary structure prediction tools/servers.....	5
Table 1.2 A list of some widely used binding pocket prediction softwares.....	8
Table 1.3 Some widely used small molecule databases.....	11
Table 1.4 Lists some of the most common biological databases.....	11
Table 1.5 A list of some <i>in silico</i> packages available for predicting ADMET profiles.....	13
Table 1.6 Properties considered in Lipinski's rule of five.....	13
Table 1.7 A list of some of the popular docking softwares.....	16
Table 1.8 Information on some marketed drugs targeting major families of proteins along with their uses.....	24
Table 1.9 Information on some marketed drugs targeting DNA along with their uses.....	25
Table 2.1 Calculated binding free energies (in kcal/mol) for ER α and ER β with compounds 3(a-d) using ParDOCK.....	39
Table 2.2 Binding free energies (in kcal/mol) calculated using MM-GBSA and nmode method for ER α and ER β	44
Table 3.1 List of 5 best molecules selected for experimental testing based on molecular interactions.....	59
Table 3.2 Binding free energies (in kcal/mol) calculated using MM-GB/PBSA and nmode method for nsP2 protease of CHIKV.....	64
Table 4.1 A list of charge and vdW parameters assigned to the metal ions for binding affinity calculation.....	77
Table 4.2 A description of 22 atom types used in the hydrophobicity calculation along with their regression coefficients.....	80
Table 4.3 A comparative evaluation of the performance of <i>Bappl+</i> and 22 other scoring functions on PDBbind 2007 core dataset comprising 195 complexes.....	82
Table 4.4 A comparative evaluation of the performance of <i>Bappl+</i> and 26 other scoring functions on PDBbind 2013 core data set comprising 195 complexes.....	84
Table 5.1 Set of reference library parameters adopted to create the DNA structure with an intercalation site.....	112

Table 5.2 Local RMSDs of the simulated 6 complexes (with known structures and energetics) with respect to experimental structures.....	114
Table 5.3 Docking accuracies carried with various DNA models.....	124