THEORETICAL STUDIES ON THE STRUCTURE AND DYNAMICS OF SYMMETRIC NUCLEIC ACIDS AND RECOGNITION PATTERNS IN DNA-LIGAND/PROTEIN SYSTEMS

PRADEEP PANT

DEPARTMENT OF CHEMISTRY

INDIAN INSTITUTE OF TECHNOLOGY DELHI

APRIL 2019

THEORETICAL STUDIES ON THE STRUCTURE AND DYNAMICS OF SYMMETRIC NUCLEIC ACIDS AND RECOGNITION PATTERNS IN DNA-LIGAND/PROTEIN SYSTEMS

by

PRADEEP PANT

DEPARTMENT OF CHEMISTRY

submitted

in fulfillment of the requirements of the degree of doctor of philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

APRIL 2019

Dedicated to my beloved parents and my loving brothers;

Gaurav and Chetan

Certificate

This is to certify that the thesis entitled, "Theoretical studies on the structure and dynamics of symmetric nucleic acids and recognition patterns in DNA-ligand/protein systems", being submitted by **Mr. Pradeep Pant** 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 him. Pradeep Pant 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.

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

Dated

Acknowledgements

Ever since my academic career started, I have been crossing paths with many great people who inspired me & nudged me further towards my goals. It indeed is difficult to jot down my gratitude in words towards them but I should definitely be making a genuine effort in doing so.

I express my sincere gratitude towards Prof. B. Jayaram for his constant support and guidance who not only helped me become a better researcher but also made me grow as a person. I would like to thank him for trusting me by providing multiple opportunities to work independently. Without his encouragements, insights, and attention, I would never have been able to initiate or to end my thesis work. I again thank him wholeheartedly for providing a great work environment for active research.

I am extremely grateful to my parents and my brothers for making me believe in myself even when the times were rough. They stood by me in all the odds and always have encouraged me to do better in life. It goes without saying that their contribution in my life (both personal & academic) is way beyond the pages of this thesis. I thank my sister-in-law (Mrs. Ganga), my niece (Garima) and nephew (Saurav) for their sweet presence in my life.

I would like to thank Prof. Nidhi Jain, with whom I have worked in my master's dissertation. She not only ensured grooming me as a researcher but also inspired me to pursue Ph.D. I take pleasure to thank all the seniors (Dr. Chanchal, Dr. Ananya, Dr. Mahesh, Poonam, and Preeti) in her lab for making my stay memorable.

I am extremely glad to thank all my friends and colleagues for their immense support so far. I recall Ishwar S Nayal (for endless fun-filled moments and a source of real dedication), Ajay Josh (a true friend whose constant support helped me to improve), Leena Aggarwal (for always been there to help me out), Sweety Arora (a great friend who always went an extra mile for me and inspires me to do hard work), The TRIO - Ruchika Bhat, Amita Pathak and Akhilesh Mishra (for helping me in getting acquainted with the

bioinformatics stuff, for creating enjoyable-friendly environment in lab, for taking care of me in tough times and for all that fun we had together), Debarati DasGupta (for her sincere attitude towards me and her time to time advices), Amit Kumar (for helping me understanding the concepts of molecular biology) and Puneeta Jaiswal (for her inspirational discussions and easing out my paperwork related issues). I thank past lab members, especially Sheeba Malik, Neha Shivhare, Sahil, Shruti and Neetu for their help in programming and constant support. I thank Ankita Singh (for introducing me to LINUX during my initial stages of Ph.D.) and Shashank Shekhar (for his prompt help whenever I encountered any server related issues).

With all the issues evolving in my stressed-up mind, I have used music to get relieved with all my tensions and music has made my regular commute relaxing. I would like to thank my favorite singers, Taylor Swift and Arijit Singh, for their godly and blissful music.

I am thankful to my SRC members (Prof. Nalin Pant, Prof. N. D. Kurur, and Prof. V. Perumal), faculties and staff members of Department of Chemistry, Indian Institute of Technology Delhi for their help and support during my Ph.D. program. I thank CSIR for financial support.

Pradeep Pant

Abstract

Since the deposition of first DNA structure in RCSB PDB site (PDB ID: 1ZNA), on March 18, 1981, the number of DNA and associated complexes has registered a significant increase (~6600, as of October 2018, in structural repositories such as RCSB, NDB, etc.). Several FDA approved drugs are now known to exert their biological action by interacting with DNA as a target and there are over ~400 DNA-drug complexes hosted by structural repository NDB. Despite advances in genomics and structural biology, the drug discovery pipeline is drying up due to increasing cost and time. Thus, there is a pressing need to develop computational methodologies in order to efficiently generate DNA binders which could be of considerable importance in DNA targeted drug discovery. Further, protein-DNA recognition, which has been witnessed in ~4743 crystal structures (available in NDB), is a routine process crucial for several biomolecular pathways and for maintaining cellular integrity. Several attempts notwithstanding, no rules for specific recognition have hitherto been identified. This necessitates the exploration of protein-DNA systems from a new perspective.

The aim of this thesis work has been towards understanding the structure and dynamics of some modified DNA molecules as well as exploring DNA-ligand/protein interactions using computational approaches. This thesis is divided into six chapters. Chapter 1 gives an overview of the structural repositories for obtaining experimentally determined DNA associated structures and computational protocols for generating DNA structures using customized helicoidal parameters, performing molecular dynamics simulations and analyzing the results. The

importance of DNA and its important associates (DNA-water, DNA-drug, and DNA-proteins) has also been discussed.

Directionality (5'→3') associated with DNA is crucial for the processes of transcription and replication. Chapter 2 of the thesis examines DNA directionality. It is discovered that the directionality of each strand of the DNA can be removed by several ways to create *in silico* models of directionally symmetric nucleic acids. The symmetric nucleic acids designed are characterized by some state of the art all-atom explicit solvent molecular dynamics simulations (ns to μs scale). Some of the symmetric molecules designed maintained B-DNA type structure as measured by helicoidal parameters and hydrogen bond strength. These modified nucleic acid molecules could have importance in the fields of antiviral development and antisense therapy for targeting mRNAs.

DNA is a target for several drug molecules underscoring the importance of computational studies for obtaining insights into interactions and specificity associated with DNA-drug complexes. However, screening a million compound library to target a specified DNA sequence is computationally forbidding due to the compute-intensive nature of docking protocols. Development of a computational tool that can provide fast and accurate estimates of binding free energies by performing rapid screening of molecules against a given DNA molecule is of significant importance. Chapter 3 focuses on the development of a rapid throughput screening protocol (RASDD) for screening a given DNA sequence against a million compound repository for the estimation of binding free energies of DNA-ligand molecule without actually docking the molecule in the minor groove of the DNA by utilizing physicochemical and structural properties of DNA as well as that of candidate molecules. The methodology, when tested on 25 DNA-drug complexes,

showed a good correlation (R=0.84) between estimated binding free energies and experimental binding free energies. RASDD can scan a million compounds against a given DNA sequence in ~18 seconds. Chapter 4 utilizes the RASDD protocol developed in the preceding chapter. A complete pipeline starting from the DNA sequence to the generation of some minor groove binders is reported (PSDDF). For any given DNA sequence, in the first step, a list of minor groove binders is obtained by virtual screening. These molecules are then docked in the minor groove using "DNA-ligand docking" software followed by the binding free energy estimations. The methodology is demonstrated on *Mycobacterium tuberculosis* H37Rv wherein two potential candidate lead molecules have been obtained.

DNA-protein recognition happens routinely by utilizing the characteristic structural and chemical features offered by both DNA and proteins. There have been several attempts to quantify these interactions. However, this complex code of recognition still remains an enigma. Chapter 5 examines protein-DNA recognition from a new perspective by utilizing the concepts of atomic clusters and cluster pairs at the interfaces of the complexes. This methodology is able to efficiently capture hydrogen bond signatures in protein-DNA complexes. Several important atomic cluster pairs have been identified which may serve as a platform to gain a better understanding of DNA-protein recognition.

In summary, this thesis addresses structural, dynamic and energetic properties of symmetric nucleic acids, recognition patterns in DNA-ligand and DNA-protein systems. This has been summarized with scope for future work in Chapter 6.

सार

आरसीएसबी पीडीबी साइट (पीडीबी आईडी: 1ZNA) में पहले डीएनए संरचना के चित्रण के बाद से, 18 मार्च 1981 को, डीएनए और संबंधित परिसरों की संख्या में उल्लेखनीय वृद्धि दर्ज की गई है (~ 6600, अक्टूबर 2018 तक, संरचनात्मक रिपॉजिटरी जैसे आरसीएसबी, एनडीबी, आदि)। एफडीए द्वारा अनुमोदित कई दवाएं अब लक्ष्य के रूप में डीएनए के साथ जैविक कार्रवाई को बढ़ाने के लिए जानी जाती हैं और संरचनात्मक रिपॉजिटरी एनडीबी द्वारा होस्ट किए गए ~ 400 से अधिक डीएनए-ड्रग कॉम्प्लेक्स हैं। जीनोमिक्स और संरचनात्मक जीव विज्ञान में प्रगति के बावजूद, दवा की खोज पाइपलाइन बढ़ती लागत और समय के कारण सूख रही है। इस प्रकार, डीएनए बाइंडरों को कुशलता से उत्पन्न करने के लिए कम्प्यूटेशनल कार्यप्रणाली विकसित करने के लिए एक दबाव की आवश्यकता होती है जो डीएनए लक्षित खोज में काफी महत्व का हो सकता है। इसके अलावा, प्रोटीन-डीएनए समूह, जिसे ~ 4743 क्रिस्टल संरचनाओं (एनडीबी में उपलब्ध) में देखा गया है, कई बायोमॉलीकुलर मार्गों के लिए और सेलुलर अखंडता को बनाए रखने के लिए महत्वपूर्ण एक नियमित प्रक्रिया है। कई प्रयासों के बावजूद, विशिष्ट मान्यता के किसी भी नियम की पहचान नहीं की गई है। यह प्रोटीन-डीएनए प्रणालियों के नए दृष्टिकोण से अन्वेषण की आवश्यकता है।

इस थीसिस कार्य का उद्देश्य कुछ संशोधित डीएनए अणुओं की संरचना और गतिशीलता को समझने के साथ-साथ कम्प्यूटेशनल दृष्टिकोणों का उपयोग करके डीएनए-लिगैंड / प्रोटीन

इंटरैक्शन की खोज करना है। यह थीसिस छह अध्यायों में विभाजित है। अध्याय 1, प्रयोगात्मक रूप से निर्धारित डीएनए से संबंधित संरचनाओं को प्राप्त करने और अनुकूलित हेलिकॉइडल मापदंडों का उपयोग करके डीएनए संरचनाओं के निर्माण के लिए संरचनात्मक रिपॉजिटरी का अवलोकन देता है, आणविक गतिशीलता सिमुलेशन का प्रदर्शन करता है और परिणामों का विश्लेषण करता है। डीएनए और उसके महत्वपूर्ण सहयोगियों (डीएनए-पानी, डीएनए-ड्रग और डीएनए-प्रोटीन) के महत्व पर भी चर्चा की गई है।

ट्रांसक्रिप्शन और प्रतिकृति की प्रक्रियाओं के लिए डीएनए से जुड़ी दिशात्मकता (5 '→ 3') महत्वपूर्ण है। थीसिस का अध्याय 2 डीएनए की दिशात्मकता की जांच करता है। यह पता चला है कि डीएनए के प्रत्येक स्ट्रैंड की दिशात्मकता को सीधे सममित न्यूक्लिक एसिड के सिलिको मॉडल में बनाने के कई तरीकों से हटाया जा सकता है। डिजाइन किए गए सममित न्यूक्लिक एसिड को कला के कुछ राज्य द्वारा विशेषता दी जाती है, जो स्पष्ट रूप से विलायक आणविक गतिशीलता सिमुलेशन है। डिजाइन किए गए कुछ सममित डीएनए को हेलिकॉइडल मापदंडों और हाइड्रोजन बांड ताकत द्वारा मापा गया बी-डीएनए प्रकार की संरचना। ये संशोधित न्यूक्लिक एसिड अणु एंटी-वायरल विकास और एंटीजन थेरेपी के क्षेत्रों में mRNAs को लक्षित करने के लिए महत्व दे सकते हैं।

डीएनए कई दवा अणुओं के लिए एक लक्ष्य है, जो डीएनए-दवा परिसरों से जुड़ी बातचीत और विशिष्टता में अंतर्दृष्टि प्राप्त करने के लिए कम्प्यूटेशनल अध्ययन के महत्व को रेखांकित करता

है। हालांकि, डॉकिंग प्रोटोकॉल की गणना-गहन प्रकृति के कारण एक निर्दिष्ट डीएनए अनुक्रम को लक्षित करने के लिए एक मिलियन कंपाउंड लाइब्रेरी की स्क्रीनिंग करना अनिवार्य है। एक कम्प्यूटेशनल उपकरण का विकास जो किसी दिए गए डीएनए के खिलाफ अणुओं की तेजी से जांच करके बाध्यकारी ऊर्जाओं के तेजी से और सटीक अनुमान प्रदान कर सकता है, महत्वपूर्ण महत्व है। अध्याय 3 वा डीएनए-लिगंड के बंधन मुक्त ऊर्जा के आकलन के लिए एक मिलियन कंपाउंड रिपॉजिटरी के खिलाफ एक डीएनए अनुक्रम देने के लिए एक तेजी से थ्रूपुट स्क्रीनिंग प्रोटोकॉल (RASDD) के विकास पर केंद्रित है। डीएनए के भौतिक और संरचनात्मक गुणों के साथ-साथ उम्मीदवार अणुओं का उपयोग करके। 25 डीएनए-ड्रग कॉम्प्लेक्स पर परीक्षण किए जाने पर कार्यप्रणाली ने अनुमानित बाध्यकारी मुक्त ऊर्जा और प्रयोगात्मक बंधन मुक्त ऊर्जा के बीच एक अच्छा संबंध (R = 0.84) दिखाया। RASDD किसी दिए गए डीएनए अनुक्रम के खिलाफ ~ 18 सेकंड में एक लाख यौगिकों को स्कैन कर सकता है। अध्याय 4 पूर्ववर्ती अध्याय में विकसित RASDD प्रोटोकॉल का उपयोग करता है। कुछ छोटे बाइंडरों के लिए डीएनए अनुक्रम से शुरू होने वाली एक पूरी पाइपलाइन की रिपोर्ट की गई है (PSDDF)। किसी भी डीएनए अनुक्रम के लिए, पहले चरण में, वर्चुअल स्क्रीनिंग द्वारा सूची प्राप्त की जाती है। इन अणुओं को तब बंधन मुक्त ऊर्जा अनुमानों के बाद "डीएनए-लिगंड डॉकिंग" सॉफ्टवेयर का उपयोग करके खांचे में डॉक किया जाता है।

माइकोबैक्टीरियम ट्यूबरकुलोसिस H37Rv पर कार्यप्रणाली का प्रदर्शन किया गया है जिसमें दो संभावित उम्मीदवार लीड अणुओं को प्राप्त किया गया है।

डीएनए और प्रोटीन दोनों की पेशकश की विशेषता संरचनात्मक और रासायनिक सुविधाओं का उपयोग करके डीएनए-प्रोटीन मान्यता नियमित रूप से होती है। इन अंतःक्रियाओं को निर्धारित करने के लिए कई प्रयास किए गए हैं। हालांकि, मान्यता का यह जटिल कोड अभी भी एक पहेली बना हुआ है। अध्याय 5 परिसर के इंटरफेस में परमाणु समूहों और क्लस्टर जोड़े की अवधारणाओं का उपयोग करके एक नए दृष्टिकोण से प्रोटीन-डीएनए मान्यता की जांच करता है। यह पद्धति प्रोटीन-डीएनए परिसरों में हाइड्रोजन बांड को कुशलता से पकड़ने में सक्षम है। कई महत्वपूर्ण परमाणु क्लस्टर जोड़े की पहचान की गई है जो डीएनए-प्रोटीन मान्यता की बेहतर समझ हासिल करने के लिए एक मंच के रूप में काम कर सकते हैं।

सारांश में, यह थीसिस सममित न्यूक्लिक एसिड के संरचनात्मक, गतिशील और ऊर्जावान गुणों को संबोधित करती है, डीएनए-लिगैंड और डीएनए-प्रोटीन सिस्टम में मान्यता पैटर्न। इसे अध्याय 6 में भविष्य के काम के लिए गुंजाइश के साथ संक्षेप में प्रस्तुत किया गया है।

Contents

List of Figures

and at VDW distance (orange) is shown. Clusters specific to DNA bases, base + sugar, sugar, sugar + phosphate, phosphate and clusters that are present in base + sugar + phosphate (all) are shown. It is noticed that the percentage of DNA backbone based contacts (sugar + phosphate; COP, COC) increases significantly in cluster pairs and that of base-specific clusters decreases while scanning cluster pairs from H-bond distance to VDW distance

List of Tables