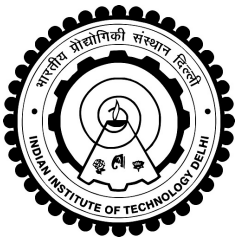
THERMODYNAMICS, KINETICS, AND SOLUTION STRUCTURE ENSEMBLE OF THERAPEUTIC PROTEINS USING METADYNAMICS

RICHA SINGH



DEPARTMENT OF CHEMICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI

OCTOBER 2019

THERMODYNAMICS, KINETICS, AND SOLUTION STRUCTURE ENSEMBLE OF THERAPEUTIC PROTEINS USING METADYNAMICS

by

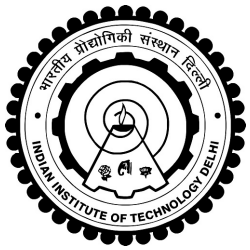
Richa Singh

Department of Chemical Engineering

Submitted

in fulfilment of the requirements of the degree of Doctor of Philosophy

to the



Indian Institute of Technology Delhi

October 2019

Dedicated to my family ...

"The most beautiful thing we can experience is the mysterious. It is the source of all true art and science."

-Albert Einstein

Certificate

This is to certify that the thesis titled **"Thermodynamics, Kinetics, and Solution Structure Ensemble of Therapeutic Proteis Using Metadynamics"** being submitted by **Ms. Richa Singh** in the Department of Chemical Engineering, Indian Institute of Technology Delhi for the award of the degree of **Doctor of Philosophy**, is a record of bona-fide research work carried out by her under my guidance and supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree. The results contained in this thesis have not been submitted for the award of any other degree, associateship or similar title of any university or institution.

Dr. Gaurav Goel
Assistant Professor
Department of Chemical Engineering
Indian Institute of Technology Delhi

Acknowledgements

I sincerely pay regards to my thesis adviser **Dr. Gaurav Goel** for his limit less patience and encouragement in guiding me through different learning curves of research. His idea of exploring the solution to a problem was never limited to trivial answers which constantly motivated me to go beyond my potential. The discussions I had with him were truly enlightening both in terms of- understanding the subject as well as inculcating work ethics in me. His drive to solve the problems along, determination, proficiency, and insights have enabled me to work through even the most difficult challenges in my research. It is under his counsel that I have learned the role of perseverance, discipline and focus not only in achieving academic goals but also to move through the obstacles in life.

I also thank my research review committee members Prof. Anurag Singh Rathore, Prof. Sudip Pattanayek and Prof. Tapan Chaudhury for their timely and expedient inputs that really helped me to upgrade the quality of my research. The critical yet encouraging assessment of my work by the committee has guided me to fill in the gaps and improve the dearth of knowledge of the subject.

I especially want to acknowledge Parvez, my senior for being an inordinate source of support and help throughout. It is worth mentioning the names of my juniors- Rit Pratik, Julie and Akshay who never failed to assist me even with the smallest things I asked for. I also want to thank Nidhi, who has been a very good friend, a mentor and an important part of the memorable time I spent at IIT Delhi. A special mention to another friend and colleague Karan, for being extremely considerate in times of need. I am also thankful to my dearest friends- Hari, Namrata, Snigdha, Rajesh, Rami and Jawad who were always there to uplift my spirit even through the toughest times I had in my research.

I am grateful to the entire staff of the department of Chemical Engineering at IIT Delhi for their constant support at various levels. I am also thankful to IIT Delhi and the Department of Science and Technology for providing fellowship. Special acknowledgement to the central facility at IIT Delhi for dispensing computational

resources for my research and the entire HPC team for their regular and extremely valuable guidance on the operation of HPC.

Anything said would be less to praise my husband Ketan for his consistent support, strength and love. He stood by me through all the crests and troughs, constantly motivating me to work harder. I thoroughly admire him for being my buffer against stress throughout the time of research. I am truly indebted to my parents for their belief in me that culminated the necessary confidence to move forward and achieve my goals. No words can express my gratitude for them who provided me with infinite affection and liberty to choose. I also want to thank my brothers Avinash and Abhay and my sister-in-law Reeta for always standing with me as pillars of support and love. My journey at IIT Delhi would not have been easier with out the bright moments I had shared with my lovely niece and nephew Arya and Shivansh. I have always strove to seek their delightful company for relieving myself of work pressure. I gratefully acknowledge my parents-in-law for their patience and faith that encouraged me to carry forward my academic interest. They definitely have a huge contribution to my achievements. I feel blessed to have as my partners in fun, my brother-in-law Chandan and his wife Aakriti. All our moments together while my short breaks from research hold a place very close to my heart. Credits to the little miracle, my daughter Ira who arrived into our lives timely enough, may be to get acknowledged in my thesis. She has been a wonderful child and her beautiful smile has always comforted me in the final days of my work.

This infinite world and its constituents, the balance of everything with its counter part, like that of failure with success is not possible without the supervision of something that is beyond infinity. I thank that ineffable supervisor 'God' for all his favors and blessings in every part of my existence.

New Delhi **Richa Singh**

Abstract

Knowledge of the solution ensemble of proteins, consisting of many short-lived intermediate conformations, can help in developing strategies to address two important protein-aggregation related challenges - (i) stability of liquid protein formulations and (ii) refolding yields of proteins. Experimental techniques used to characterize the solution ensemble provide important, but partial information as extensive time and ensemble averaging of data makes identification of individual short-lived intermediates difficult. Enhanced sampling molecular dynamics techniques, like metadynamics, have emerged as powerful computational tools that enable determination of the free energy landscape of protein folding as well as some other bio-molecular processes at a high spatial and temporal resolution. In metadynamics, the underlying hamiltonian is modified by adding a history-dependent potential that puts a penalty on already sampled regions of protein conformational space defined in terms of one or more collective variables (CVs), making possible the sampling of high free energy states. In this thesis, we have implemented well-tempered bias exchange metadynamics (WT-BEMD), a variant of metadynamics, to determine the solution structure ensemble and associated thermodynamics and kinetics of two important therapeutic proteins, namely insulin and transforming growth factor (TGF)-β3. Both proteins are known to aggregate under thermal and mechanical stress, thereby affecting their therapeutic applications. While sufficient spectroscopic and crystallographic information is available for insulin, TGF-β3 in turn, is a less investigated protein. Low solubility of TGF-β3 has long complicated its spectroscopic studies, although it is known that the active solution structure of TGF-β3 has several important differences with respect to its X-ray crystal structure.

WT-BEMD simulation of insulin was performed under high thermal and chemical stress replicating the amyloidogenic conditions used in several spectroscopic studies of this protein. A bin-based clustering method in-combination with a Markovian kinetic transition network model was employed on the BEMD trajectories to construct a thermodynamic and kinetic model of insulin folding. Major fraction (72 %) of the

obtained solution structure ensemble of insulin was observed to retain the native fold in form of the R-, R^f- and T-states of this protein. A comparison of α-proton chemical shifts of the obtained folded state ensemble (F) with those of the reported spectroscopic structure of insulin under amyloidogenic conditions provided an excellent correlation. The second largest fraction of structures corresponded to a compact denatured state ensemble (D) and the remaining structures belonged to an ensemble of partially folded states (PF) of insulin. Important differences in structure of the denatured and the partially folded intermediates were identified in terms of the local environments of amino acids. Moreover, in correspondence with the earlier findings, the breakdown of cooperativity in folding was observed for insulin where one of its partially folded states (folding intermediate) was found to connect the folded and the denatured states on the lowest free energy path (LFEP). These partially folded intermediates were compared and contrasted with spectroscopic, crystallographic, and calorimetric measurements during early stages of insulin aggregation. Further, a folding time of 0.46 ms for denatured to the folded state transition and an order of magnitude lower timescale for the partially folded to the folded state transition was identified using simulated annealing and kinetic Monte Carlo (KMC) algorithms.

WT-BEMD simulation of TGF-β3 was performed under typical re-folding conditions. Bin based clustering of the obtained BEMD trajectories provided four metastable state ensembles of this protein. These ensembles were exhaustively characterized based on several structural metrics (tertiary contact maps, residue based - solvent accessible surface area (SASA), root mean square deviation (RMSD), R_g, secondary structure content and α-proton chemical shifts) and the nature of stabilizing interactions (hydrophobic, H-bonds, salt bridges, cation-π). A comparison of the entire solution ensemble constituted by these metastable state ensembles, with the reported spectroscopic structure of this protein provided excellent correspondence. Further, it was also found that the longest lived ensemble (constituting 50 % of the overall solution ensemble) provided the closest match with the reported details of solution structure of TGF-β3 and therefore, it was identified as the folded state of this protein in solution. This folded state was found connected to another long lived metastable ensemble containing denatured states, through a partially folded state on the lowest free energy path. Subsequently, the role of a zwitterionic detergent, CHAPS, in modulating the free energy landscape of TGF-β3 was studied by quantifying the preferential interaction coefficient of the obtained ensembles with CHAPS. The calculations revealed significant differences in solvation of native-like and partially folded / denatured states. Interestingly, two of these metastable ensembles showed signifi-

cant structural changes in short molecular dynamics simulations performed without CHAPS.

सारांश

प्रोटीन के घोल की जानकारी, जिसमें कई अल्पकालिक मध्यवर्ती संरचनाएं शामिल हैं, प्रोटीन एकत्रीकरण से संबधित दो मुख्य चुनौतिओं का हल निकालने के लिए रणनीति बनाने में सहायता कर सकती है – (i) तरल प्रोटीन योगों की स्थिरता और (ii) प्रोटीन की कार्यपरक उत्पादन मात्रा। प्रोटीन के घोल का गुणात्मक परीक्षण करने क लिए उपयुक्त प्रायोगिक तकनीकें महत्वपूर्ण, किंतु आंशिक जानकारी प्रदान करती हैं क्यूंकि इनसें उत्पन्न डेटा के व्यापक सामयिक और गुणात्मक औसत होने के कारण अल्पकालिक मध्यवर्ती प्रोटीन संरचनाओं की पहचान करना जटिल हो जाता है। इनहैंस्ड सैंप्लिंग मोलिक्युलर डायनमिक्स तकनीकें, जैसे मेटाडायनमिक्स, शक्तिशाली कंप्युटेश्नल उपकरण के रूप में उभरी हैं जो प्रोटीन तह और साथ ही कुछ अन्य जैव आंणविक प्रक्रियाओं के मुक्त उर्जा परिदृश्य को उच्च स्थानीय व सामयिक स्पष्टता से निर्धारण करने में सक्षम हैं। मेटाडायनमिक्स में, अंतर्निहित हैमिलटोनिअन को इतिहास–निर्भर पोटेंषिअल के संचय द्वारा संशोधित किया जाता है जिससे पहले से निर्धारित प्रोटीन संरचना क्षेत्र जो कलेक्टिव वेरिएबल्स (CVs) से परिभाषित होता है, के पुनः निर्धारण पर जुर्माना लगता है। इस थीसिस में, वेल–टेंपर्ड बायस एक्सचेंज मेटाडायनमिक्स (WT-BEMD) जो कि मेटाडायनमिक्स का एक प्रकार है, का उपयोग करके दो प्रमुख चिकित्साविधान प्रोटीनों, इंसुलिन और TGF-β3 की घोल में पाई जाने वाली संरचनाओं के समूह तथा उससे संबंधित ऊष्मागतिकी एवं बलगतिकी का निर्धारण किया गया है। दोनों प्रोटीन थर्मल और यांत्रिक तनाव से एकत्रित होने के लिए जाने जाते हैं, जिससे उनके चिकित्सकीय उपयोग पर असर पड़ता है। जहां इंसुलिन के लिए पर्याप्त स्पेट्रोस्कोपिक एवं क्रिस्टलोग्राफ़िक जानकारी प्राप्य है, TGF-β3 इसके इतर, एक कम जांचा गया प्रोटीन है। TGF-β3 की कम घुलनशीलता ने लंबे समय से इसके स्पेट्रोस्कोपिक अध्ययनों को जटिल बनाया है, हालांकि यह ज्ञात है कि TGF-β3 की घोल में पाई जाने वाली सक्रिय संरचना इसकी X-ray क्रिस्टल संरचना से काफ़ी अलग है।

इंसुलिन का WT-BEMD सिमुलेशन उच्च तापीय और रासायनिक तनाव में किया गया था जो कि इस प्रोटीन के कई स्पेट्रोस्कोपिक अध्ययनों में इस्तेमाल अमाइलॉइडोजेनिक स्थितियों की नकल है। इंसुलिन की तह के ऊष्मागतिकी एवं बलगतिकी मॉडल के निर्माण हेतु BEMD प्रक्षेपवक्रों पर बिन–आधारित समूहीकरण और मार्कोविअन काइनेटिक ट्रांज़ीशन नेटवर्क मॉडल का उपयोग किया गया था। प्राप्त इंसुलिन घोल संरचना समूह के मुख्य

अंश (७२ प्रतिशत) को इस प्रोटीन की R-, Rf- और T- अवस्थाओं में पाया गया है। प्राप्त की गई सक्रिय संरचना के समूह और इंसुलिन की अमाइलॉइडोजेनिक स्थितियों में कथित स्पेट्रोस्कोपिक संरचना की α-प्रोटोंन केमिकल शिफ़्ट की तुलना करने पर उतक्रिष्ट सहसंबंध मिलता है। संरचनओं का दूसरा सबसे मुख्य अंश एक सघन विक्रित संरचना समूह (D) के अनुरूप था और शेष संरचनाएं आंशिक रूप से बने संरचना (PF) समूह से संबंधित थीं। विक्रित और आंशिक रूप से बनीं मध्यवर्ती संरचनाओं में अमीनो ऐसिड के स्थानीय वातावरण के संदर्भ में महत्वपूर्ण अंतर पाया गया है। इसके अलावा, पहले के निष्कर्शो से मेल करता इंसुलिन तह प्रक्रिया में सहकारिता का टूटना अवलोकित किया गया है जहां इसकी एक आंशिक रूप से बनीं अवस्था (मध्यवर्ती तह संरचना) को न्यूनतम मुक्त उूर्जा पथ (LFEP) पर सक्रिय और विक्रित संरचना के मध्य पाया गया है। इन आंशिक रुप से बनी मध्यवर्ती संरचनाओं की तुलना इंसुलिन एकत्रीकरण के शुरुआती चरणों में लिए गए स्पेट्रोस्कोपिक, क्रिस्टलोग्राफ़िक और कैलोरीमीट्रिक मापों से की गई है। इसके अलावा, सिमुलेटेड अनीलिंग और काइनेटिक मोंटे कार्लो (KMC) कलन विधि के प्रयोग से विक्रित से सक्रिय संरचना रूपांतरण का समय 0.46 ms और आंशिक रूप से बनी संरचना से सक्रिय संरचना रूपांतरण का समय एक परिमाण कम निर्धारित किया गया है।

TGF-β3 के WT-BEMD सिमुलेशन को प्रायः प्रयुक्त रीफ़ोल्डिंग परिस्थितियों के अंतर्गत किया गया है। प्राप्त BEMD प्रक्षेपवक्रों के बिन–आधारित समूहीकरण ने इस प्रोटीन के चार मेटास्टेबल संरचना समुह प्रदान किये। इन समूहों का कई संरचनात्मक मापदंडों (टरशियरी संपर्क मानचित्र, रेसिडियू–आधारित विलायक सुलभ क्षेत्र (SASA), रूट मीन स्क्वायर डेविएशन (RMSD), Rg, सेकेंडरी स्ट्रक्चर मात्रा और α-प्रोटोंन केमिकल शिफ़्ट) साथ ही स्थिरता प्रदान करने वाले संपर्को (हाइड्रोफ़ोबिक, H-बॉडः साल्ट ब्रिज, कैटायन–π) की प्रकृति के आधार पर विस्तृत रूप से विश्लेषण किया गया। इन मेटास्टेबल संरचना समूहों से बने संपूर्ण घोल समूह की इस प्रोटीन की कतिथ स्पेट्रोस्कोपिक संरचना से तुलना करने पर उत्क्रिष्ट समानता प्राप्त होती है। इसके अलावा, यह भी पाया गया है कि सबसे दीर्घ आयु वाला समूह (संपूर्ण घोल समूह का ५० प्रतिशत) TGF-β3 की कथित घोल संरचना के साथ सबसे निकटतम समानता दर्शाता है और इसीलिए, इसको इस प्रोटीन की घोल में सक्रिय संरचना माना गया है। इस सक्रिय संरचना को न्यूनतम मुक्त उूर्जा पथ पर एक और दीर्घायु मेटास्टेबल समूह जिसमें विक्रित संरचनायें हैं, क साथ एक आंशिक रूप से बनी संरचना के माध्यम से संपर्क में पाया गया है। तत्पश्चात, प्रिफ़रेंशियल इंटरेक्शन कोफ़ीशिएंट के मात्रात्मक ज्ञान से TGF-β3 के मुक्त उूर्जा परिदृश्य को संशोधित करने में एक जुइटरायनिक डिटर्जेंट, CHAPS की भूमिका का अध्ययन किया गया है। गणना से पता चलता है कि सक्रिय समरूप संरचना और आंशिक रूप से बनी या विक्रित संरचनाओं के विलायकयोजन में महत्वपूर्ण अंतर हैं। दिलचस्प बात यह है कि, इनमें से दो मेटास्टेबल समूहों नें CHAPS की अनुपस्थिति में की गई लघु आणविक गतिशीलता सिमुलेशन में महत्वपूर्ण संरचनात्मक बदलाव प्रदर्शित किये।

Table of contents

List of figures

List of tables

List of symbols

Symbol	Definition
ω	Gaussian potential deposition rate
γ	Bias factor
σ	CV width
τ_{G}	Gaussian potential deposition time
τ_{b}	Correlation time
$\Gamma^{\mathrm{i}*}_{\mathrm{XP}}$	Local preferential interaction coefficient as function of distance
Γ^{*}_{XP}	Global preferential interaction coefficient as function of distance
ds	Hypercube dimension
s	CV value
D	Diffusion coefficient
F	Free energy
f^{S}	1D Free energy obtained from re-scaled Gaussian heights in HILLS files
V_{G}	Bias potential
W	Gaussian potential height
A_{rel}	Relative solvent exposed surface area
$Q_{\mathrm{hydrophobic}}$	Fraction native hydrophobic contacts

List of abbreviations

ANS	8-Anilinonaphthalene-1-sulfonic acid
BEMD	Bias exchange metadynamics
CV	Collective variable
FRET	Fluorescence emission transfer
FF	Force field
HDX	hydrogen/deuterium exchange
HPC	High performance computing
MD	Molecular dynamics
MC	Monte-Carlo
MetaD	Metadynamics
NMR	Nuclear magnetic resonance
PDB	Protein data bank
PC	Principal component
QM/MM	Quantum Mechanics/ Molecular Mechanics
RDC	Residual dipolar coupling
REMD	Replica exchange metadynamics
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
SASA	Solvent Accessible Surface Area
SAXS	Small Angle X-ray Scattering
SMD	Steered Molecular Dynamics
UV CD	Ultra Violet Circular Dichroism
WT-BEMD	Well Tempered Bias Exchange Metadynamics