A HIGH-THROUGHPUT COMPUTATIONAL MODEL FOR PREDICTION AND MITIGATION OF BIO-THERAPEUTIC PROTEIN AGGREGATION

by

RIT PRATIK MISHRA

DEPARTMENT OF CHEMICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

DECEMBER 2023

A HIGH-THROUGHPUT COMPUTATIONAL MODEL FOR PREDICTION AND MITIGATION OF BIO-THERAPEUTIC PROTEIN AGGREGATION

by

RIT PRATIK MISHRA

DEPARTMENT OF CHEMICAL ENGINEERING

Submitted in fulfillment of requirements of degree of Doctor of Philosophy

INDIAN INSTITUTE OF TECHNOLOGY DELHI

December 2023

Dedicated to,

My beloved family, without whose endless love and support, I could not achieve this.

CERTIFICATE

This is to certify that the thesis entitled "A High-Throughput Computational Model for Prediction and Mitigation of Bio-Therapeutic Protein Aggregation", submitted by Mr. Rit Pratik Mishra to the Indian Institute of Technology Delhi for the award of the degree of Doctor of Philosophy, is a record of the original bonafide research work carried out by him. He has worked under my supervision and has fulfilled the requirements, which to my knowledge, has reached the requisite standard for the submission of this thesis. The results in this thesis have not been submitted in part or whole to any University or institute to award any degree or diploma.

Date: **27.12.2023**

Place: **Delhi**

Prof. Gaurav Goel
Professor
Department of Chemical Engineering
Indian Institute of Technology Delhi
New Delhi, India

Acknowledgments

I would like to express my sincere gratitude and appreciation to the following individuals and institutions for their invaluable support and contributions throughout the course of my research. First and foremost, I am deeply grateful to my supervisor, Professor Gaurav Goel, for his unwavering guidance, patience, and expertise. His insightful feedback, continuous encouragement, and dedication to my academic growth have been instrumental in shaping this thesis. I would like to extend my heartfelt thanks to Professor Anurag Singh Rathore, who served as the Chairman of my review committee. His valuable insights, critical evaluation, and constructive suggestions have greatly enriched the quality of this research work. I am also indebted to Professor Sudip Kumar Pattnayak, the department expert in my committee, for his valuable inputs, discussions, and guidance throughout this journey. His expertise and thorough review of my work have significantly contributed to its improvement. I extend my gratitude to Professor Tapan Kumar Chaudhuri, the external department expert, for his insightful comments and valuable suggestions during the evaluation process. His expertise and feedback have been immensely valuable in shaping the outcomes of this research.

I am deeply indebted to my family: my father, Rajendra Kumar Mishra, my mother, Gitanjali Mishra, and my brother, Rajpratik Mishra. Their unwavering love, support, and understanding have been the pillars of my strength, motivation, and perseverance throughout this endeavor. I am grateful to my dear friends, and my fiancé Rinnie, for their constant support, motivation, and unwavering presence throughout this journey. Their friendship, love, and encouragement have been invaluable in navigating the challenges encountered during this research. I would like to express my gratitude to my seniors, Richa Singh Rathore and Avinash Mishra, for their valuable advice, encouragement, and willingness to share their knowledge and experiences. Their insights have been invaluable in navigating the challenges encountered during this research.

I would like to acknowledge the financial support I received from the MHRD scholarship provided by the Indian Institute of Technology (IIT) Delhi. This scholarship played a crucial role in supporting my research and enabling me to pursue my academic goals. Additionally, I would like to express my gratitude to the HPC facility at IIT Delhi for granting me access to their high-performance computing resources. The computational resources provided by the facility were instrumental in performing complex simulations and data analysis, enhancing the quality and depth of my research. I would also like to extend my appreciation to our lab assistant, Krishna Kumar, for his invaluable assistance and support throughout my research. His technical expertise, willingness to help, and efficient management of laboratory resources have greatly facilitated my experiments and data collection.

To all those mentioned above and to anyone else who has contributed in ways big or small, I offer my sincerest thanks. This thesis would not have been possible without your support, encouragement, and belief in my abilities.

A special mention goes to my pet cat, Billoo, who provided unconditional love and companionship during the challenging times of the COVID-19 lockdown in Delhi. His presence brought comfort and solace, making the difficult days more bearable. Lastly, I would like to express my gratitude to Lord Hanuman, whose divine blessings and guidance have provided me with strength and determination, especially during times of self-doubt and difficulty. His light has shown me the way and kept my spirits high.

Date: December 2023
Place: Bangalore

Rit Pratik Mishra
Roll No.: 2016CHZ8292

Abstract

Protein aggregation has been identified as one of the main problems in the manufacture and storage of biotherapeutics. Mechanistic understanding and quantitative predictions of aggregation kinetics would enable design of effective mitigation strategies. A comprehensive understanding of the aggregation process at the molecular level is essential to gain better structural insights and a quantitative understanding of protein-protein interactions (PPIs) and to predict kinetic parameters involved in each stage (folding/unfolding, nucleation, and fibrillation). However, studying the entire process at a single resolution is not feasible due to the different time and length scales at which these stages occur. In the present thesis, we have developed a multiscale computational model to accurately predict protein aggregation kinetics parameters, provide insights into molecular determinants of PPIs involved in aggravating aggregation, and quantify the effects of various stress factors (monomer concentration, excipients, temperature) on the aggregation kinetics. Two representative biotherapeutic proteins were studied: insulin, a small globular protein, and rituximab, an IgG1 antibody, as representative biotherapeutic proteins. These proteins are of commercial significance and necessitate stable formulation designs. However, due to the substantial disparity in size (with rituximab being approximately 25 times larger than insulin), computational complexities arise, particularly in terms of computational resource requirements and achieving convergence in sampling. The designed predictive model was used successfully to make accurate predictions for insulin aggregation kinetics and discover FDA approved inactives to inhibit its aggregation. Finally, we developed a protocol to sample aggregation prone regions (APRs) of rituximab, which can be extended to other classes of monoclonal antibodies.

Even at its lowest complexity, understanding protein aggregation requires comprehensive knowledge of single molecule conformational flexibility and bio-molecular assembly over a broad range of timescales. For insulin and mAbs, this spans timescales on the order of femtoseconds (bond vibrations) to seconds and hours (aggregation nucleation). We combined kinetic, thermodynamic, and structural information from single molecule (protein folding) and two molecules (association) explicit-solvent simulations for the determination of kinetic parameters in protein aggregation nucleation with insulin as a model protein. We observed this "early-events" enough to accurately predict the kinetic parameters up to the nucleation event. A structural bioinformatics approach was developed to account for heterogeneity of aggregation-prone species with the transition complex theory applicable

in modelling association kinetics involving non-native species. Specifically, the kinetic pathway for formation of aggregation-prone oligomeric species was found to contain a structurally specific dominant binding mode, making the kinetic process similar to native protein association. The kinetic parameters thus obtained were used in a population balance model and accurate predictions for aggregation nucleation time varying over two orders of magnitude with changes in either insulin concentration or an aggregation-inhibitor ligand concentration were obtained, while an empirical parameter set was not found to be transferable for prediction of ligand effects. Further, this physically determined kinetic parameter set provided several mechanistic insights, such as the identification of the rate-limiting step in aggregation nucleation and a quantitative explanation for the switch from Arrhenius to non-Arrhenius aggregation kinetics around the melting temperature of insulin.

Establishing structural and interaction specificity in insulin self-association along the non-native aggregation pathway enabled us to use structure-based, computer aided methods for identification of ligands with high-specificity for inhibiting protein-protein interactions (PPIs) implicated in early-stage aggregation nucleation. We proceeded to identify lead molecules from the FDA inactives database that can inhibit insulin aggregation using a multi-level virtual screening protocol. Some key challenges specific to the present problem, when compared with native protein association, include structural heterogeneity of the protein species involved, multiple association pathways, and relatively higher probability of conformational rearrangement of the association complex. A first-level screening of the inactives database was based on dominant pharmacophore features of three short peptides and an organic fluorogen, both shown to significantly inhibit insulin aggregation nucleation by binding to aggregation-prone conformations of insulin. We then performed ensemble docking of various low-energy ligand conformations on aggregation-prone conformers of insulin followed by molecular dynamics simulations and binding affinity calculations on a subset of docked complexes to identify a subset of lead molecules with a high probability for inhibiting insulin aggregation nucleation. We then applied our designed aggregation nucleation kinetics model to account for protein-ligand interaction details and conformational rearrangement effects during protein oligomerization to determine ligand concentration for effective aggregation inhibition. The effect of these select ligands (lead compounds) aggregation inhibition was extensively investigated by incubating insulin at various concentrations under aggregation-prone aqueous buffer conditions (low pH, high temperature). Herein, aggregation kinetics were quantified using size exclusion chromatography and the secondary structure was determined using circular dichroism spectroscopy to quantify effect of ligand on native structure content after co-incubation with insulin under aggregation-prone conditions. Specifically, two lead compounds (Riboflavin and Aspartame) showed significant retardation of insulin aggregation at formulation concentration when used at a 2:1 ligand: protein molar concentration.

The second biotherapeutic of interest, rituximab, is a IgG1 monoclonal antibody used in the treat chronic lymphocytic leukemia. Linear response driven

molecular dynamics simulations, based on accelerating conformational transition along the slow normal modes, were used to efficiently and extensively explore the configuration space, and thus, identify partially-folded aggregation-prone conformations of rituximab at acidic conditions. A structural bioinformatic protocol based on combining information on residue hydrophobicity, solvent accessible surface area, β-propensity, and protein tertiary structure (3D structure) was used to identify potential aggregation hotspots in these partially-folded rituximab conformers. A set of putative non-native dimeric complexes, totaling 57 complexes, was generated through a multi-step approach involving rigid body docking with APRs as the target and subsequent molecular dynamics simulations for conformational rearrangement. MMPBSA calculations were performed on these complexes to evaluate the contribution of APRs to the non-native binding free energy. Based on the ranking obtained from these calculations, we identified nine critical APRs (six in Fc, two in Fab, and one in Hinge) that play a key role in driving the aggregation of rituximab. The APRs identified in the Fc region were explicitly localized within the C_{H2} domain, which aligns with the findings from Differential scanning calorimetry (DSC) investigations conducted on IgG1 antibodies under different pH conditions. Additionally, the presence of the identified APR in the hinge region has already been substantiated through mutation studies. Furthermore, the APRs in the Fab region were found to reside within the complementarity-determining region (CDR) loop, with their sequences overlapping with residues involved in antigen binding. These findings provide valuable insights into the underlying mechanisms of rituximab aggregation and its potential implications in bio-pharmaceuticals development.

सार

प्रोटीन एकत्रिति को निर्माण और भंडारण में मुख्य समस्याओं में से एक के रूप में पहचाना गया है। एकैकिक अवलोकन और मात्रात्मक पूर्वानुमान की मृदुता के साथ एकैकिक अवलोकन ने प्रोटीन–प्रोटीन संवादों में शामिल एकद्वितीयता के बेहतर संरचना अंदाज और प्रत्येक चरण (फोल्डिंग/अनफोल्डिंग, न्यूक्लिएशन, और फाइब्रिलेशन) में शामिल होने वाले किनेटिक पैरामीटरों की यथात्मक समझ प्राप्त करने की संवादिक समझ होना आवश्यक है। हालांकि, इन चरणों का एक ही संकलन में अध्ययन करना योग्य नहीं है क्योंकि इन चरणों को प्राप्त करने के लिए अलग–अलग समय और लंबाई की स्थायिता होती है। वर्तमान थीसिस में, हमने प्रोटीन एकत्रिति की गतिकी पैरामीटरों का सटीक अनुमान लगाने, एकत्रिति को बढ़ाने वाले प्रोटीन–प्रोटीन संवादों के मौलिक निर्धारकों के बारे में अंदाज़ा देने, और विभिन्न तनाव कारकों (मोनोमर संघटन, अल्पका, तापमान) के प्रभाव को न्यूनतम करने के प्रभाव को मात्रात्मक रूप से मापन करने के लिए एक बहुमानक संगणन मॉडल विकसित किया है। दो प्रतिनिधित्वकर्ता बायोथेरेप्यूटिक प्रोटीनों का अध्ययन किया गया: इंसुलिन, एक छोटे गोल मौलिक प्रोटीन, और रिटक्सिमैब, एक IgG1 एंटीबॉडी, जो व्यापारिक महत्व के हैं और स्थिर रूप से निर्माण डिज़ाइन की आवश्यकता है। हालांकि, आकार में विशाल अंतर (रिटक्सिमैब लगभग इंसुलिन के 25 गुना बड़ा होने के बावजूद), विशेष रूप से संगणन संसाधन आवश्यकताओं और सैम्पलिंग में संघटन प्राप्त करने के प्रक्रियाकलित मुद्दे उत्पन्न होते हैं। डिज़ाइन किया गया पूर्वानुमानित मॉडल सफलतापूर्वक इंसुलिन एकत्रिति की गतिकी पैरामीटरों के लिए सटीक पूर्वानुमान बनाने के लिए उपयोग किया गया और उसकी एकत्रिति को रोकने के लिए FDA द्वारा मंजूर किए गए असक्रिय तत्वों की खोज की। आखिरकार, हमने रिटक्सिमैब के एकत्रिति प्रवृत्ति क्षेत्रों (एएपीआर) को सैम्पल करने के लिए प्रोटोकॉल विकसित किया, जो अन्य प्रकार की मोनोक्लोनल एंटीबॉडियों के लिए बढ़ाया जा सकता है।

प्रोटीन एकत्रिति को समझने के लिए उसके सबसे न्यून संघटन में भी, एक मोलेक्यूल के आकार बदलने की लचीलापन और बाइओ–मोलेक्यूलर समूह को बड़े संविदान घड़ी के स्तर पर जानकारी की आवश्यकता होती है। इंसुलिन और mAbs के लिए, इसमें फेम्टोसेकंड्स (बॉन्ड वाइब्रेशन) से सेकंड और घंटों (एग्रीगेशन न्यूक्लिएशन) के आदेश के कार्यकाल परिसर होते हैं। हमने इंसुलिन को मॉडल प्रोटीन के रूप में प्रोटीन एग्रीगेशन न्यूक्लिएशन में किनेटिक पैरामीटर्स की निर्धारण के लिए सिंगल मोलेक्यूल (प्रोटीन फोल्डिंग) और दो मोलेक्यूल (संघटन) स्पष्ट–सोल्वेंट सिमुलेशन से किनेटिक, थर्मोडायनमिक, और संरचनात्मक जानकारी को मिलाया। हमने इस "प्रारंभिक घटनाओं" को काफी समय तक देखा ताकि हम न्यूक्लीएशन घटना तक किनेटिक पैरामीटर्स की सटीक पूर्वानुमान कर सकें। हमने एक संरचनात्मक जैव–सूचना विज्ञान दृष्टिकोण विकसित किया ताकि एकजुटता प्रोटीन संघटन प्रवृत्ति के विभिन्न प्रजातियों को मनमानी रूप से विचार किया जा सके, जो गैर–प्राकृतिक प्रजातियों को शामिल करने के लिए पर्याप्त था। विशेष रूप से, संघटन–प्रवृत्ति घनिष्ठ प्रजातियों के गठन के किनेटिक पथ में एक संरचनात्मक विशिष्ट प्रमुख बाइंडिंग मोड होने का पता चला, जिससे किनेटिक प्रक्रिया मूल प्रोटीन संघटन के समान हो गई। इस तरीके से प्राप्त किनेटिक पैरामीटर सेट का उपयोग एक जनसंख्या संतुलन मॉडल में किया गया और इंसुलिन की चिकित्सा में परिवर्तन के साथ दो आदेशों की विभिन्नता

वाले समय के अनुसार एकत्रिति न्यूक्लीएशन का समय का सटीक पूर्वानुमान मिला, जबकि एक प्राकृतिक पैरामीटर सेट का प्रशंसा परमीकरण के पूर्वानुमान के लिए स्थायी नहीं मिला। इसके अलावा, यह भौतिक रूप से निर्धारित किनेटिक पैरामीटर सेट ने कई यांत्रिक अंशों की समझ प्रदान की, जैसे कि एकत्रिति न्यूक्लीएशन में गति सीमित करने वाली कदम की पहचान और इंसुलिन के पिघलाने के तापमान के चारों ओर अरेनियस से गैर–अरेनियस एकत्रिति किनेटिक्स की ओट देने के लिए एक मात्रबादी पैरामीटर सेट के बदलाव के लिए क्वांटिटेटिव व्याख्या।

इंसुलिन स्व–संघटन में गैर–प्राकृतिक एकत्रिति मार्ग के साथ संरचनात्मक और इंटरएक्शन विशिष्टता स्थापित करने ने हमें इस्प्राकृतिक संघटन न्यूक्लीएशन के प्रारंभिक चरण में शामिल प्रोटीन–प्रोटीन इंटरएक्शन (PPIs) को निरोधित करने के लिए उच्च–विशिष्टता वाले लिगैंड्स की पहचान के लिए संरचना–आधारित, कंप्यूटर सहायित विधियों का उपयोग करने की संभावना दी। हम एफडीए अनैक्टिव्स डेटाबेस से वो नेतृत्वक अणुओं की पहचान करने के लिए एक मल्टी–स्तर वर्चुअल स्क्रीनिंग प्रोटोकॉल का उपयोग करके इंसुलिन एकत्रिति को निरोधित कर सकने वाले डिज़ाइन किया। पैरम्यूफोर फीचर्स के द्वारा तीन छोटे पेप्टाइड्स और एक आर्गेनिक फ्लोरोजेन की प्रमुख फार्माकोफोर फीचर्स के आधार पर इनैक्टिव्स डेटाबेस की पहली स्तर की स्क्रीनिंग पर आधारित थी, जिनमें इंसुलिन की एकत्रिति न्यूक्लीएशन को बाधित करने के लिए बाइंडिंग करते समय एकत्रिति–प्रवृत्ति संरचनाओं के साथ बाइंड करने के रूप में दिखाया गया था। फिर हमने इंसुलिन के एकत्रिति–प्रवृत्ति संरचनाओं पर विभिन्न लो–ऊर्जा लिगैंड संरचनाओं का एन्सेम्बल डॉकिंग किया, और मॉलेक्युलर डायनेमिक्स सिम्युलेशन और बाइंडिंग अफीनिटी गणनाएँ करने के बाद एकत्रिति न्यूक्लीएशन को निरोधित करने के लिए उपयोग किए जाने वाले नेतृत्वक अणु संरचनाओं पर विभिन्न लो–ऊर्जा लिगैंड संरचनाओं के डॉक करने के रूप में एक सबसेट की पहचान के साथ एक सबसेट की पहचान करने के लिए मॉलेक्युलर डायनेमिक्स सिम्युलेशन और बाइंडिंग अफीनिटी गणनाएँ की। फिर हमने प्रोटीन–लिगैंड इंटरएक्शन विवरण और प्रोटीन ओलिगोमरिज़ेशन के दौरान संरचनात्मक पुनर्व्यवस्थिति प्रभाव को शामिल करने के लिए हमारे डिज़ाइन किए गए एग्रीगेशन न्यूक्लीएशन किनेटिक्स मॉडल का अप्लाई किया ताकि प्रभावी एकत्रिति निरोधन के लिए लिगैंड गति का निर्धारण किया जा सके। इन चयनित लिगैंड्स (लीड कॉम्पाउंड्स) के एकत्रिति निरोधन पर इसके प्रभाव की खोज बड़े पैमाने पर की गई द्वारका, इंसुलिन को विभिन्न गतिसंघटन–प्रवृत्ति जलीय बफर स्थितियों (कम pH, उच्च तापमान) में विभिन्न गतिसंघटन की ताप संरेखण का उपयोग करके किया गया, यहां, आकार विच्छेदन क्रोमेटोग्राफी का उपयोग किया गया और सर्कुलर डिक्रोइज्म स्पेक्ट्रोस्कोपी का उपयोग किया गया था ताकि गतिसंघटन पैरामीटरों का प्रभाव लिगैंड के साथ को–इंक्यूबेशन के बाद मूल संरचना सामग्री पर कितना होता है। विशेष रूप से, दो लीड कॉम्पाउंड्स (राइबोफ्लेविन और एस्पारटेम) जब इन्सुलिन: प्रोटीन मौलर गति में 2:1 लिगैंड के अनुपात पर प्रयुक्त किए जाते हैं, तो इंसुलिन की एकत्रिति को प्रमुख रूप से विलंबित करते हैं।

दूसरी बायोथेराप्यूटिक जिस पर ध्यान दिया जा रहा है, रिटक्सीमैब, एक आईजीजी1 मोनोक्लोनल एंटीबॉडी है जो क्रॉनिक लिम्फोसिटिक लीकीमिया का इलाज करने में प्रयुक्त होता है। लीनियर प्रतिक्रिया द्वारा प्रेरित मॉलेक्युलर डायनेमिक्स सिमुलेशन, धीमे सामान्य मोड के साथ संरचनात्मक संवाद को गति देने के रूप में, कॉन्फ़िगरेशन स्पेस को प्रभावी और बड़े पैमाने पर अन्वेषित करने के लिए उपयोग किए गए थे, और इस तरह, रिटक्सीमैब के आम्लीय स्थितियों में आंशिक फोल्डेड एग्रीगेशन–प्रवृत्ति संरचनाओं की पहचान की। एक स्थानिक बायोइन्फोर्मेटिक प्रोटोकॉल, जिसमें अमीनो अम्लीयता, औद्योगिक पहुंची हुई सतह क्षेत्र, β–प्रवृत्ति, और प्रोटीन तृतीयक संरचना (3डी संरचना) की जानकारी को एक साथ मिलाने का उपयोग उन आंशिक–फोल्डेड रिटक्सीमैब संरचनाओं में एग्रीगेशन हॉटस्पॉट्स की संभावित पहचान के लिए किया गया था। एक पूरी तरह रिजिड बॉडी डॉकिंग के साथ आपर के रूप में आगंतुक को लक्ष्य के रूप में लेकर और फिर संरचनात्मक पुनर्व्यवस्थिति के लिए आपकल्पित गैर–नैटिव डाइमेरिक संरचनाओं की तैयारी के एक मल्टी–स्टेप दृष्टिकोण में द्विमिक्षासन किया गया।

इन संरचनाओं पर MMPBSA गणनाएँ की गईं ताकि गैर–नैटिव बाइंडिंग स्वतंत्र ऊर्जा के योगदान का मूल्यांकन किया जा सके। इन गणनाओं से प्राप्त रैंकिंग के आधार पर, हमने रिटक्सीमैब के एग्रीगेशन को ड्राइव करने में महत्वपूर्ण एपीआर्स की पहचान की (एफसी में छः, फैब में दो, और हिंज में एक)। एफसी क्षेत्र में पहचाने गए एपीआर्स की स्पष्ट स्थानिकता थी, जो आईजीजी1 एंटीबॉडियों के विभिन्न pH शर्तों के तहत डिफरेंशियल स्कैनिंग कैलोरिमेट्री (डीएससी) अन्वेषणों के परिणामों से मेल खाती है। इसके अलावा, फैब क्षेत्र में पाए गए एपीआर्स का पता चला कि वे संप्रेषकता–निर्धारण क्षेत्र (सीडीआर) लूप के भीतर बसे हुए थे, जिनके अंकों में एंटीजन बाइंडिंग में शामिल रहने वाले अमीनो अम्ल शामिल थे। ये फिंडिंग्स रिटक्सीमैब के एग्रीगेशन के पीछे के मेकेनिज़्म के अंदर की विचारों को प्रदान करती हैं और इसके जीव–फार्मास्यूटिकल विकास में संभावित प्रभावों की देन करती हैं।

Table of Contents

List of Figures

List of Tables

List of Symbols

χ	Rotational angle
κ_n	Relative fibril concentration at t_{nu}
$\mathrm{p}(\mathrm{E}_i)$	Boltzmann averaged energetic contribution of a pharmacophore feature to binding energy
$\mathrm{T}_{\mathrm{cof}}$	Tanimoto coefficient
A	Aggregation Prone Conformation
$cos\Theta$	Interface similarity index
D	Denatured conformation
f	Fractional monomer conversion
k_{aL}, k_{dL}	Ligand association/dissociation rate
k_a, k_d	Non−native association/dissociation rate
k_{fib}, k_{-fib}	Fibril growth/dissociation rate
k_{nu}, k_{-nu}	Nucleus growth and dissociation rate
k_r, k_{-r}	Conformation re−arrangement rate
k_u, k_f	Unfolding, Folding rate
N	Native conformation
N_c	Total number of inter−atomic contacts

N_c^*	Total number of inter–atomic contacts in the transition state
n_c^*	Critical nucleus size
P_i	Pharmacophore feature
T_m	Protein melting temperature
t_{nu}	Nucleation time