CHARACTERIZING PROTEIN SURFACES FOR BINDING ASSOCIATIONS

A thesis submitted in partial fulfillment of the requirement
of the degree of Doctor of Philosophy
by

Neeladri Sen

Registration No. – 20132006

Department of Biology
Indian Institute of Science Education and Research Pune
India - 411008

To

Ma and Baba

DECLARATION

I declare that this written submission represents my idea in my own words and where others' ideas have been included; I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Neeladri Sen

Date: 30/03/2020

Neeladri Sen
20132006
IISER Pune

CERTIFICATE

Certified that the work incorporated in the thesis entitled “Characterizing Protein Surfaces for Binding Associations”, submitted by Neeladri Sen was carried out by the candidate, under my supervision. The work presented here or any part of it has not been included in any other thesis submitted previously for the award of any degree or diploma from any other university or institution.

Date: 30/03/2020

Dr. M. S. Madhusudhan
Associate Professor
IISER Pune

Acknowledgments

The time in IISER was the longest I had spent anywhere after school and home. It had the perfect combination of friendship, attachment, liveliness, personal space, care and love. I never referred IISER as home but deep down it had taken the place of home, mostly because of the people I met here and the friendly environment that IISER provides to all its students. Because it is thesis acknowledgments, let me thank the people who directly contributed to my thesis, before thanking the ones who might made IISER home.

The supervisor - Madhu is however the common thread as to why IISER is home and why this acknowledgements exists. I still remember our first interaction during the first Bioconclave talks. I was an Int PhD fresher of the year 2013, and then this man comes and sits beside me, who looked like a young man whom I thought was either an old PhD or new post-doc (definitely not a faculty). He asked me to stop shaking my legs, which I did not adhere to, and from there beings not adhering to things he would say. I then joined his lab in the summer of 2014 for rotation, finally joining as a PhD student in Dec 2014. He has been the most terrific mentor, a caring person, a friend. He taught me how to think critically, analyze data, question ideas, to write, to present my work. Whatever I am presently in my professional life, is all because of him. He spent enormous amount of time and energy in papers and presentations, trying to improve its quality. He was always there with his treasure trove of ideas and always open to discussing my ideas, hearing it, questioning it, going to the root of each problem, even at points whenever I would want to give up. I could always text him with my problems at any point and he would be there to help. He too was always a really fun person to work with. Though sometimes his ideas and him not understanding that I do not want to further peruse a problem got on my nerves, yet he would lighten the mood with the bad humor. Your idealistic ideas are really great only if the world was an ideal place. Also, am sorry for the numerous amount of time I was rude to you. Thank you for tolerating me through these 5 long years, which I guess a lot of other mentors would not be able to do. I hope we keep discussing ideas with each other in the years to come.

Lab members- Next comes my labmates (in the order of the time stamp when we first interacted) – Neelesh, Sagar, Yogendra, KP, Minh, Binh, Abhilesh, Akash, Sanjana, Shipra, Nida, Parichit, Golding, Tejashree, Ankit, Swastik, Dayal, Kaustubh, Gulzar, Shreyas, Mukundan and Atreyi. I would like to thank all of them for making the lab such a fun and intellectually stimulating environment to work in, allowing me to grow. The lab meets were the times when everyone would help pick up problems in my thought process or work and suggest means to rectify them, also helping me in getting out of problems if and when required. The lab was always ready to help whenever I faced problems related to my thesis work. A long due thanks to Parichit, the scientific programmer in the lab because of which the lab kept running easily in the past. He has helped me immensely in the early years with problems related to installation, scripts, getting programs to run etc. He was followed by Gulzar who too helped in various installation and software related problems. Neelesh, was the next go to person to fix bugs and problems, discuss ideas and get inputs. He was the first PhD in the IISER lab and he did help a lot in getting me accustomed to programming. Yogendra was always ready to help me wherever he could, and is a treasure trove of information. He was like an easy search tool who was ready with an answer for a scientific query. Sanjana would go out of her way to help me. I remember once after lab meet she sat and helped me debug a script, I was really touched by the gesture. Tejashree, the person with whom I spoke the most in the lab, someone ready to help and scold in case you are not working. Kaustubh is the playboy of the lab. I had pulled his leg throughout, sometimes irritating and enraging him. I think, he thinks of me as a good friend of his, and yes the feeling might be mutual. Also, he is really hardworking. Ankit, the most logical person I had seen. His presence in the lab, prevented me from doing anything illogical in my PhD. Thanks are due to Golding too for being a quiet, fun presence in the lab. Also thanks to Atreyi and Mukundan for reading and commenting on the thesis

Nida, a very dear friend of mine, again another common link between my thesis and IISER being called home. She is a really hard working, intelligent, logical person whose life goal is to do science. We had numerous chats throughout my PhD and it was always fun being around you. I shall always cherish the time we spent together discussing random stuff such as my and your phobias, life, philosophies, sadness, problems etc. She is the reason for my first publication, thanks for that.

Other faculties- I would like to thank all my RAC members – Prof. Raghavan Varadarajan, Dr. Saikrishnan Kayarat, and Dr. Jeet Kalia. We met at yearly intervals to discuss my progress and they helped a lot by questioning my techniques, giving me new ideas to implement, helped me looking at various problems that I might have overlooked. I would like to thank Sai for introducing me to computation structural biology during my lab rotation, because of which I thought of trying dry lab with Madhu. Working with him too was a great learning experience and helped me develop critical thinking. I would also like to thank Sudha, with whom I did my first Int PhD rotation, who actually introduced me to scientific thinking and work. She has really been sweet, friendly and caring. I had rushed to her related to various non-academic issues and she was always there to listen to me, talk and help fix problems. I would also like to thank Akanksha for allowing me to work in her lab during the rotation.

Collaborators- I would also like to acknowledge my collaborators – Prof. Maya Topf, Prof. Chandra Verma and Prof. Ted. Hupp. I also had the chance to work with Maya at Birkbeck for 4 months. Working with her was a really enriching experience. She was a friendly person, always active with her inputs towards the project. I would also like to thank Josh from Maya's lab, who helped me during my visit to the lab.

Adiministration- I would like to thank the IT team, Neeta and Nisha, who were there to help in solving problems related to Rosalind and knl clusters and taking care of it. They took care of the cluster, helped in installations and proper functioning.

Also credits are due to the current and previous dean doctoral studies, Dr. Srabanti and Dr. Girish, who listened to me and helped me in solving various academic office related problems, stipend queries, fees payments etc. I cannot not thank Tushar and Sayalee from the academic office. They are the most efficient person I came across in IISER, always helping students whenever required wrt to form submissions, NOCs, signatures, fellowships. I would also like to thank Dr. Rao for running after CSIR for fellowships. Mrinalini and Mahesh from the Bio office too were extremely helpful.

Seniors- Thanks are due to the sweet Int PhD 2012 seniors – Mukul, Ankitha, Sukrut and Tomin. Mukul and Ankitha were someone who would be always ready to give you important life advices. They are really sweet and helpful people who would again go out of their way to help people out. I was really lucky to have worked with the two of you. Sukrut was my go to person whenever I got bored from my lab. Thank you for all the sweets and gift. Please get married. Tomin, is more of a batchmate than senior and the first paragraph for the batch holds true for you too.

IISER was home, thanks to Int PhD batch 2013- IISER Pune 2013 Int PhD batch – Sandip, Anish, Ron, Bharat, Swati, Shivani, Mehak, Jay, Dhriti, Amar, Akhila, Deepak, Aditi, Divya, Charu, Adarsh, Anshul, Harpreet and Kashyap. All the trips, lunches, dinners, outings, mid night tea, new year parties, pubs, long chats, karavaans, restaurants, birthdays etc. with you all is the reason IISER was home. Thank you for being the tolerating and fun loving batch you are/were. I know that I could not have asked for a better, more helpful, more fun loving batch other than this. Also I know that many a times you all might have done whatever I wanted, thanks for that.

Sandip, Anish and Ron, you three had been the greatest friends and people around. Thank you for tolerating me, listening to me, follow whatever I wanted. Thank you for allowing me to choose whatever I wanted most of the times. Thank you for the amazing, thoughtful birthday gifts. Thank you for taking all the criticisms from me and also listening me whine about everything in life, and my thoughts that I have a disease. The three of you were always there beside me whenever I needed, helped me in whatever way you people could. The lunches together and the long chats, few trips, Saturday night dinners etc will always be fresh in my mind. Thanks for reading and commenting on my thesis. Sandip was the one who tolerated me maximum, the one person who was tortured the most by me, the one person whose would just try to do whatever I wished so that I do not be sad. Sandip was the roommate who turned to family, my cooking partner, thanks for all the mallu food you (hardly) cooked. Life in IISER would have been this interesting without you. Anish, is really the nicest person I know. He is just a simple man who is happy with his simple needs in life. A lot can be learnt from him, as to how to be helpful, friendly, thoughtful and nice. Anish was always excited about going out of the way for birthday plans, which I did actually pull down (Sorry about that). Please think that you are grown up now (and stop telling when I would grow up). Ron, I do not know what to write about him - his stupid smile, bad haircuts, face etc. But yes, I think he is the most holy and helpful person I know. He would actually go way out of his way to help someone in need (people whom he would barely know). I do not know if you sing well, but best for your future with music. And please stop wasting money and be less dramatic.

Bharat has been a very irritating person throughout my PhD, but he is like a necessary evil. His room will always be open even in the middle of the night to chat. Dude, you really are a great friend of mine and yes get married, I need to see an overpriced big fat Indian wedding. Also, how can I forget, he is a really entertaining person, I guess, people will never get bored in his presence with the amount of shit talk he is capable of doing. Shivani, is like a mother in my head, who would sacrifice things for others. She too is fun person to be with and helpful. And yes, I really like the rajma made by her. Thank you for that. Saturday night dinners were fun with the two of you. Jay, the person with the straight face, who would watch random videos, and laugh. You have been a really great friend and also a motivation for me to run. Swati, the person who changed the most during the PhD from a chirpy bird to a brahmakumari. She was actually very selfless to me and helped me a lot, and yes thank you for all the trips. Mehak, again a really good friend, who does a lot of things for her friends. I really feel lucky that I have you as my friend. Thanks for all the tea. Dhriti, is a really entertaining human being to hang around as she is a treasure trove of gossip with all the drama, and is again a very helpful person. Amar, Akhila, the two of you were really fun to hang around and the restaurants and trips would have elements missing had the two of you not been there. Feel really bad that we had to miss the wedding. Aditi, thank you for the fun memories of the Rishikesh trip and after that. Deepak, thank you for all the sudden jump scares. Divya, thank you for getting married, giving us one more fun batch trip. Adarsh thank you for answering and helping with UK visa. Charu, Harpreet, Anshul, Adarsh and Kashyap, thanks for the small chit chats all along. I would really miss all of you and hope we do meet in the future here and there and keep the iPhD 2013 batch alive and as fun loving as the last 7 years were.

Family- The acknowledgements would be incomplete without thanking my family – Ma and Baba and Ankur. They had been my pillar of strength and had always supported me through all the decisions taken by me in my life. Thank you for being the understanding, helpful, non-judgmental, non-questioning and supportive family that you people had always been. Nothing I write would be enough to express my gratitude toward to you all. To end thank you for everything.

Funding- Money is important. I would like to thank CSIR for my SPMF fellowship and British Council and DBT for Newton Bhabha PhD placement fellowship to work in Birkbeck. I would also like to thank CSIR and Infosys for travel grants for attending conferences overseas.

- Neeladri

Contents

Chapter 1 - Overview

1. **Synopsis**
2. **Thesis organization**

Synopsis

Proteins interact with one another and other biomolecules to carry out their functions. These interactions are mediated via amino acids on the surface of proteins. In this thesis, we structurally characterized protein-protein and protein-small molecule interfaces and studied the residue environments at protein-protein interfaces. Characterization of these interface residues can help identify binding modes of proteins with one another and small molecules.

To characterize the protein-protein interfaces structurally, we created a database of protein-protein/domain-domain interfaces in the PDB and clustered them by their geometric similarity. We examined how proteins belonging to the same fold can utilize the same/different interface geometries to interact with one another. Further analysis, on specific proteins, showed that the geometry at the interface could be structurally similar irrespective of the fold that the protein belongs to. Further characterization of the protein-protein and domain-domain interfaces showed that though amino acid pairing across interface residues are similar, yet protein-protein interfaces have a higher self-amino acid pairing compared to domain-domain interfaces.

While this library can be used to model protein complexes of varying geometry, in this thesis, we specifically explored coiled-coil interfaces. We built a random forest based scoring scheme to predict if two coiled-coils would interact in a particular orientation. This algorithm was used to identify native coiled-coil interactions at the interface from non-native interactions. Along with scoring, we also predicted the interactions between the coiled-coiled domains of JC virus agnoprotein with Rab11B and p53 and built a model of agnoprotein with Rab11B.

Further, we studied amino acid environments at protein-protein interfaces. We examined how different interface residues transition from one environment in a monomer to another in a complex. The residue environment was characterized using residue depth, which is the distance of the residue from the nearest bulk solvent. We noticed that the hydrophobic amino acids have higher propensities of getting buried on complex formation compared to hydrophilic ones. We developed a depth based scoring potential for protein-protein interface residues, which was used to distinguish near-native interfaces from non-native interfaces.

Along with the characterization of interfaces, we also identified hotspot residues, which are important for mediating the interactions. We compared three properties of the hotspot and non-hotspot residues – depth change on complex formation, conservation and interaction potential (how favorable are the interactions of a residue with other residues from a different protein). These three properties were different for hotspot and non-hotspot residues and were used to build an empirical decision tree based classifier. Our method was shown to be robust across the different tested datasets and comparable to if not better than other state of the art methods.

Not only did we characterize protein-protein interfaces, we also studied protein-small molecule (drug molecule) interfaces. Small molecule drugs bind to target proteins based on structural and physicochemical complementarity. There is a likelihood that other proteins have binding sites structurally similar to the binding pocket of the target protein. We designed a general structural similarity based method to identify binding pockets of small molecules on proteins. We tested this methodology in identifying the alternate binding partners of a small molecule drug Nutlin, which is known to bind Mdm2. Our predictions were validated both computationally (molecular mechanics score, molecular dynamics simulations, docking scores) and experimentally (thermal shift assay for one of the predictions).

Further, we used this method to identify alternate drug binding pockets, in experimentally validated off-target proteins, for several other known drugs. We noticed that several of the predicted off-target binding sites had a bound drug/ligand in the crystal structures. This provides higher confidence in our predictions as previous literature shows that a variety of ligands can bind to the same binding site in a protein. This methodology can be used for drug repurposing or predicting off-target effects of drugs.

We also designed/predicted inhibitors against the different proteins of the Nipah virus and computationally analyzed their stability. To begin with, we modeled the proteome of the Nipah virus and designed 4 peptide inhibitors against 3 of its proteins. We then docked small drug like molecules onto Nipah proteins using Autodock and Dock. Using molecular dynamics simulations and molecular mechanics calculations, we analyzed the stability of protein-peptide/small drug like inhibitors. All the predicted/designed inhibitors will plausibly be effective against different strains of the Nipah virus.

While characterizing protein-protein and protein-small molecule interfaces, we characterized the different environments in proteins using residue depth. Different residues prefer to be at different depth levels. We calculated the effect of depth on residue substitutions and created three depth dependent amino acid substitution matrices. These matrices were then successfully used to predict deleterious mutations in proteins.

To conclude, this thesis captures various aspects of protein-protein and protein-small molecule interfaces, which can be used for scoring protein-protein interfaces, modeling protein complexes, identification of coiled-coiled interfaces, prediction of hotspot residues, prediction of off-target effects of drugs, drug repurposing and other applications.

Thesis Organization

Chapter 1 – Overview

The synopsis to the thesis and thesis organization.

Chapter 2 – Introduction to techniques

A brief introduction to various techniques used in the study – residue depth, CLICK, structure modeling, docking, molecular dynamics simulations, molecular mechanics, knowledge based potential and measures of accuracy.

Chapter 3 – Structural study of protein-protein interfaces

Created and structurally clustered a library of all known protein-protein and domain-domain interfaces. Showed how interfaces can be structurally similar despite topological differences in the respective proteins.

Chapter 4 – Prediction and modeling of coiled-coil protein-protein interfaces

Developed a scoring scheme to score coiled-coil interfaces, which was then used to identify binding partners of coiled-coil protein and also used to model coil-coil protein interfaces.

Chapter 5 – Study of residue environments at protein-protein interfaces to score protein complexes

Developed a knowledge based statistical potential derived from the different tendencies of the amino acids to get buried on complex formation. Utilized these potentials to identify near-native binding mode of protein-protein complexes.

Chapter 6 – Classification of interface residues into hotspot and non-hotspot residues

Studied the properties of hotspot and non-hotspot residues and utilized these properties to classify the interface residues.

Chapter 7 – Structural study of protein-small molecule interfaces: prediction of off-target effects of Nutlin

Created a structural similarity based search method to predict off-target binding sites of the small molecule drug Nutlin. These predictions were analyzed both computationally and experimentally.

Chapter 8 – Prediction of binding sites of drugs on off-target proteins

Utilized the methodology developed in the previous chapter to predict binding pockets of drugs on already known off-target proteins.

Chapter 9 – Predicting and designing therapeutics against the Nipah virus

Modeled the Nipah proteome and utilized it to design peptide inhibitors and predict small drug like molecules against the viral proteins.

Chapter 10 – Characterizing residue environments in proteins to develop environment dependent substitution matrices

Studied residue environments in proteins using depth. Created depth dependent amino acid substitution matrices and utilized it to predict deleterious mutations.

Chapter 11 – Conclusions and Future Prospects

A summary of what has been achieved in this thesis and what can be explored in the future. A detailed analysis of the results has been provided in the respective chapters.

Chapter 2 - Introduction to techniques

1. **Residue environment and Residue Depth**
2. **Binding site prediction**
3. **CLICK**
4. **Knowledge based potentials**
5. **Machine learning**
6. **Docking**
7. **Molecular Dynamics simulations**
8. **Binding free energy calculations**
9. **Mathematical measures**

1. Residue environment and Residue Depth

Proteins are molecular machines that are made of amino acids, which fold in a particular 3D shape. The environment of an amino acid depends on where it is located in a protein and the properties of the amino acid vary depending on its environment. For instance, the relative permittivity experienced by a polar chemical group in a solvent is 80, whereas in a protein surface it ranges between 20-30, whereas in the interior of the protein it is between 2-4 [1]. Besides, hydrogen bonds can be more than 1 kcal/mol stronger in non-polar environments as compared to polar environments [2,3]. The hydrophobic interactions too play an important role in protein stability [4]. The hydrophobic interactions are predominant in the protein interior, and hence it becomes important to quantify the environment of the residue based on the degree of the burial of different amino acids.

Traditionally, solvent accessible surface area (SASA) [5] was one of the ways in which the degree of burial was categorized. SASA values were classified into levels such as buried, intermediate and exposed. Residues in the hydrophobic core of a globular protein were typically buried while the polar residues that constituted the periphery of the protein were exposed. This classification, however, is rather coarse (Figure 2A) and does not stratify the interior of the protein adequately. Similarly, another way of quantifying residue environment is by calculating the number of atoms in contact within 4.5Å [6]. This parameter too has been used to predict the SASA of residues. A somewhat more concise description of the residue environment is provided by the depth measure [7]. Residue (or atom) depth is defined as the distance of a residue (or atom) to the closest molecule of bulk solvent (Figure 1). A solvent molecule is not defined as bulk if it has less than two neighboring water molecules in a sphere of radius 4.2Å (1.5 hydration shells). Hence any trapped water molecule or those present in cavities are not considered as the bulk solvent. DEPTH program for predicting residue depth also mimics solvent dynamics by repeatedly solvating the protein in different orientations [8]. The number of water molecules to define bulk solvent and the number of iterations for solvating a protein molecule can be changed based on the user. Residue depth offers a more stratified description of the protein interior (Figure 2B) and hence can be used to describe protein microenvironments.

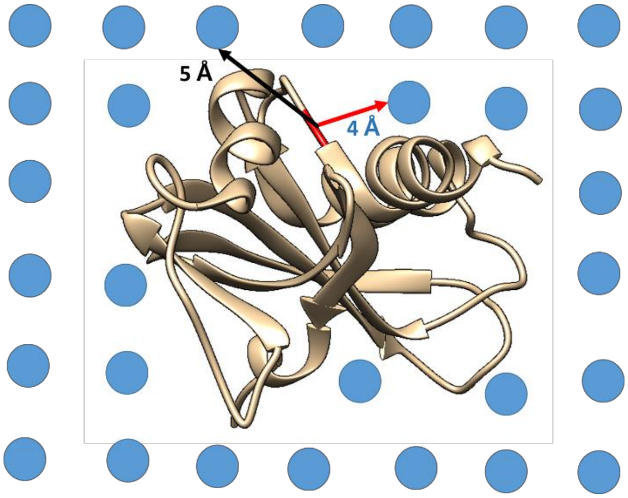


Figure 1- A protein molecule (shown in yellow ribbon) in a box of solvent molecules (represented as blue dots). The residue depth (for the residue in red) being predicted as the one from the nearest bulk solvent (shown in red arrow), as compared to the one in black arrow.

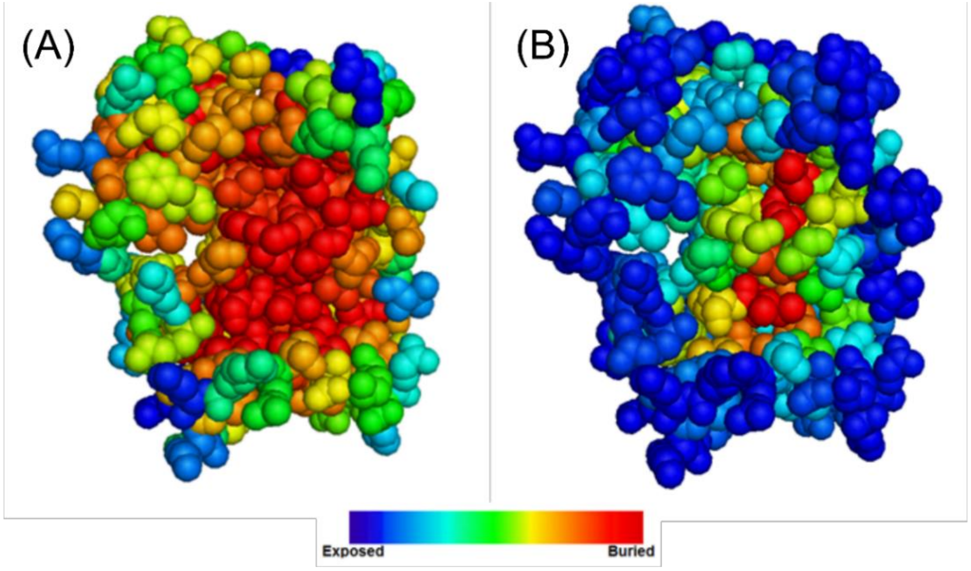


Figure 2- A cross-section of a protein (human dihydrofolate reductase, PDB- 1MVT) stratifying microenvironments by (a) SASA (b) Depth. All atoms of the protein are rendered in sphere representation and are colored according to SASA and depth using PyMol [9]*. Blue represents exposed residues and red represents buried residues.*

Residue depth is an apt descriptor of protein microenvironments is further evidenced from the many uses of depth. Residue depth correlates better with hydrogen-deuterium exchange data than SASA [7]. It is also a vital feature in the detection of post translational modification sites [10,11]. In conjunction with SASA, depth has been used to predict small molecule ligand binding sites and cavities in proteins [8,12]. Combining depth with SASA, electrostatic and hydrogen bonding interactions has been shown to effectively predict the pK$_a$ of ionizable groups in proteins [12]. Residue depth has been efficiently combined with hydrophobicity and hydrophobic moment derived from the primary sequence of the protein to predict temperature sensitive mutations [13]. In combination with evolutionary sequence profiles and SASA, depth could be used to recognize native protein folds [14,15]. In each of the applications mentioned above the key aspect has been the ability of depth to describe the immediate neighborhood of amino acid residues.

2. Small molecule ligand binding site prediction

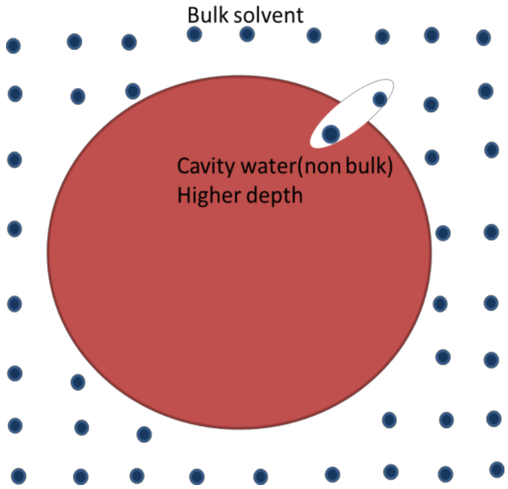


Figure 3 – Schematic of a protein (shown in red sphere) in a box of solvent molecule (solvent molecules as blue spheres). A cavity though solvent accessible still has high depths as the water molecule do not qualify as bulk solvent

Residue depth has been previously used for multiple applications [7,8,12], one is the prediction of small molecule ligand binding sites of proteins. The binding sites are accessible to water and hence have a high solvent accessible surface area, however because of the sizes of the binding site, the water in the binding sites does not qualify as bulk solvent (do not have 2 neighbors within 1.5 hydration shells) (Figure 3). Hence the residues at binding sites have a higher depth. Besides, binding site residues are more conserved as compared to the rest of the protein surface [16,17]. Hence, DEPTH program predicts a region on the protein surface as a binding site if the residues have a high solvent accessible surface area, high depth and are conserved [12]. In addition, all residues with at least one atom within 6.5 Å of the cavity water were listed as binding site residues candidates. For details about the algorithm of binding site residue prediction please refer to Tan et al. [12].

3. CLICK

Several computational tools have been developed for the identification of structural similarities between the 3-dimensional structures of proteins or parts of proteins [18–25]. The program CLICK [18] can be used to match 3-D structures of proteins. CLICK compares two constellations of points irrespective of their chain connectivity. The CLICK program creates small cliques of points (3-7 in number) from representative atoms (user defined criterion, C^{α} atoms or combination of C^{α}, C^{β} atoms etc.) of spatially proximal amino acid residues (Figure 4). These cliques are then superimposed by a 3D least-squares fit. To guide the matching of cliques, other features such as solvent accessibility, secondary structure and residue depth can also be used.

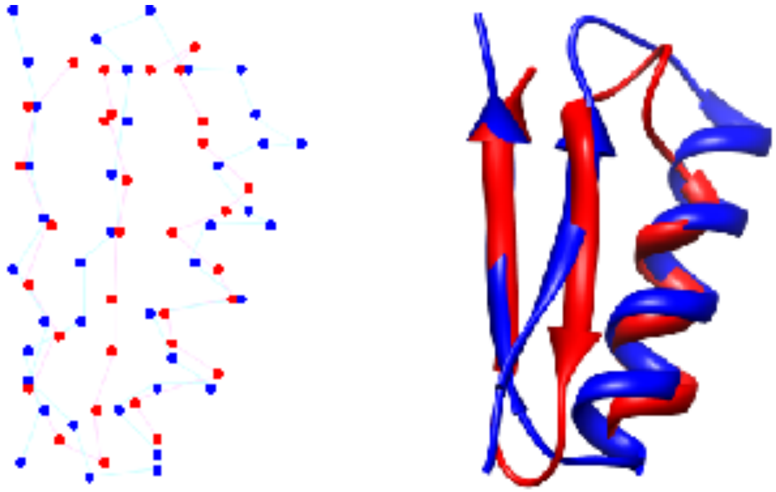


Figure 4 – Structural alignment of the C-terminal domain of Arginine repressor (PDB-1XXA) (red) and C-terminal domain of Translation Initiation Factor (PDB-1TIG) (blue). The left portion shows the representation of C^{α} atoms which are superimposed on each

other even though the two proteins have different topologies (right). (Adapted from http://cospi.iiserpune.ac.in/click/Design/img/Click.png)

The structural superimposition produced by CLICK is associated with two values - RMSD and structure overlap. RMSD is the root mean square deviation between the aligned representative atoms after the structures are superimposed on each other. Structure overlap is the percentage of representative atoms (of the smaller protein) that are within a cut-off distance of 2.5 Å from corresponding atoms of the other protein after structural superimposition.

The functionality of a protein depends on the spatial orientation of the residues, irrespective of the topology of the overall protein [26–28]. Such functional matches will be missed if we match protein structures with constraints based on their connectivity. Hence a topology independent protein structure matching tool such as CLICK, can help in identifying such structural similarities. CLICK has been successfully used for multiple structural comparisons [18,29–31].

4. Knowledge based potentials

Knowledge based potentials (or statistical potentials) are energy functions derived from analysis of already existing protein structures in the PDB [32]. These potentials rely on the assumption that the native structure generally has the lowest free energy as compared to other states [33,34]. These potentials involve statistical analysis of different features of proteins as seen in their high resolution crystal structures. Examples of such features involve amino acid pair preferences at protein interfaces [35], the distance between pairs of protein atoms [33], number of neighbors in contact with a residue [34] etc. Various knowledge based potentials have been developed for protein-protein, protein-ligand and protein-DNA interactions to predict either the binding free energy or feasibility of binding [36,37]. These potentials have also found applications in protein structure prediction and design [38,39].

Most of these potentials follow Sippl's formulation [40]. The potential of a feature r under study is defined as

$$U(r) = -kT\ (N_r^{obs})/N_r^{exp}$$

, where *k* is the Boltzmann constant, *T* is the absolute temperature N_r^{Obs} is the number of observed events of the feature r in the PDB and N_r^{exp} is the number of expected events of the feature *r*, under a random scenario.

5. Machine Learning

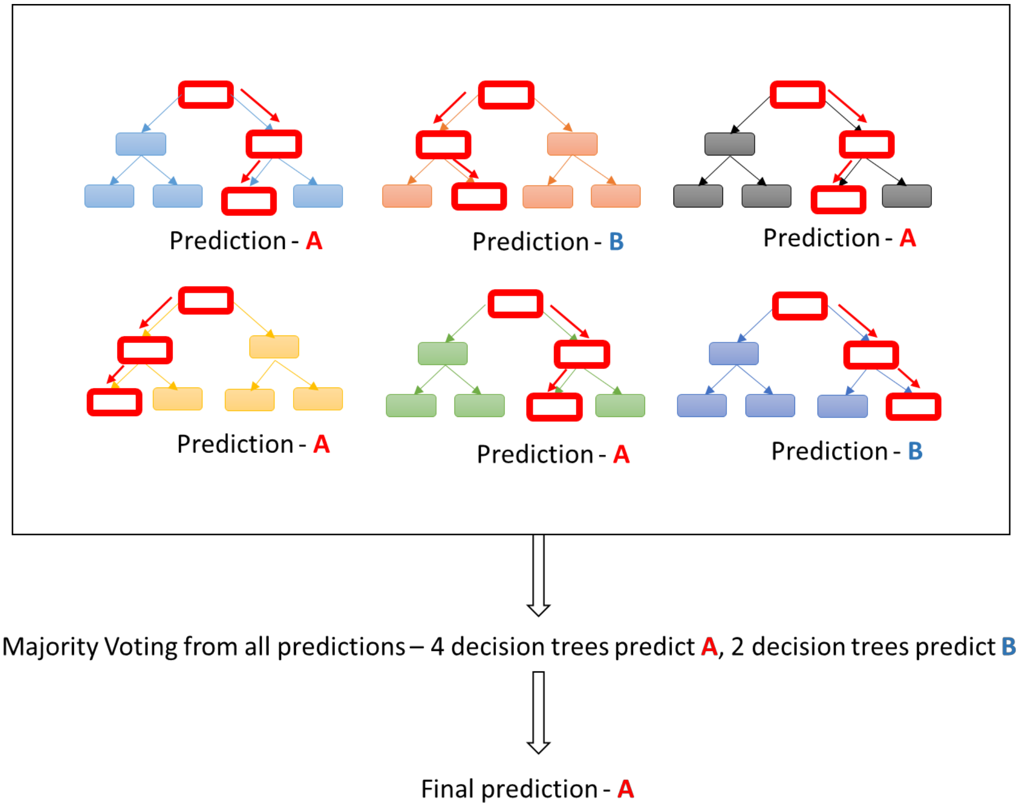


Figure 5 – Schematic showing random forest based classification. 6 decision trees shown in different colors make predictions as A or B. The decision making process made by each tree for an input has been highlighted. The box with red outline and red arrows show the flow of decision making process. Based on the voting, the prediction that gets the maximum vote is provided as the final prediction (here A).

Machine learning is a set of computer algorithms that learns from patterns from a training set of parameters and utilizes these patterns to build a model to make a prediction [41]. Machine learning tools can be classified as supervised or unsupervised. Supervised machine learning tools build a mathematical model based on the training set that contains both the training parameters and the desired output [42]. Supervised machine learning can be carried out using various tools like support vector machine, neural network, decision tree, random forest etc. A detailed review of supervised machine learning tools can be found elsewhere [42]. Unsupervised learning contains only the parameters of the dataset and does not have an output. These tools learn from the data and are used for the clustering of data points [43].

One example of a supervised machine learning tool is the Random forest, which creates several decision trees. Decision trees make predictions of a particular type based on recursive splitting of the dataset based on different sets of conditions [44]. The different decision trees make their own predictions. Random forest makes the final prediction taking the mode of the predictions made by its constituent decision tree (Figure 5) [45].

6. Docking

Molecular docking refers to the prediction of the orientation of a biomolecule with respect to another. Docking can be done for protein-protein, protein-small molecule ligand, protein-nucleic acid complexes. The shape of the biomolecule and their interactions help predict their orientations. The Coulombic, hydrogen bonds and van der Waals interactions between the individual components are computed and used to score a conformation [46]. Depending on the computational resources various degrees of flexibility of the bonds are provided during docking, which allows it to sample various conformations before making a final prediction. Rigid body docking is a faster docking scheme (compared to ones where protein residues too are flexible) wherein the protein is considered as a rigid body and the small molecule ligand is allowed to sample all the 6 translational and rotational degrees of freedom. The docking tools may allow the user to mention the site on the protein where the docking simulation is to be carried out [46–48]. Some of the commonly used docking tools are - HADDOCK [49], pyDock [50], SwarmDock [51], ZDock [52], RossettaDock [53], Autodock [54] etc. A more detailed analysis of docking tools can be found elsewhere [55–57].

7. Molecular dynamics simulations

Proteins are dynamic molecules that undergo various kinds of motion, which are important for their functioning. These dynamics can be studied using molecular dynamics simulations that predict the trajectory of atoms in biomolecules by solving Newton's equations of motions [58]. The forces between these atoms (or particles) are calculated by interatomic potentials or force fields, which can be of different types such as AMBER, CHARMM, OPLS, GROMOS etc. [59–62]. These force fields vary from each other in their description of the various parameters of atoms or the forces between them. Molecular dynamics simulations for non-membrane bound water soluble proteins involve solvating the biomolecule in a box of water, followed by neutralizing the system by the addition of counter ions. This is followed by the energy minimization of the system to get rid of steric clashes or inappropriate geometry. The solute-solvent system is then equilibrated by varying pressure till it reaches some pre-desired density and the temperature is stabilized. The next step involves the equilibration of the pressure. After stabilizing the temperature and the pressure, the system is ready for data collection regarding its trajectory. Details about various advances and applications of molecular dynamics simulations can be found elsewhere [58,63–66].

8. Binding free energy calculations

Binding free energies are estimated as a combination of molecular mechanics energies with Poisson-Boltzmann (MM/PBSA) [67,68] or generalized Born and surface area continuum solvation (MM/GBSA) [67]. These estimates are typically based on molecular dynamics simulations. The binding free energy is calculated as

$$G = G_{bond} + G_{el} + G_{vdW} + G_{pol} + G_{non-pol} - TS$$

, where G_{bond} is the molecular mechanic's energy term for bonded interactions (bond, angle, dihedral), G_{el} refers to the energy term for electrostatics, G_{vdW} for the energy term for van der Waals interactions. G_{pol} and $G_{non-pol}$ are the polar and non-polar contributions to the solvation free energy. The G_{pol} can be calculated either by Poisson-Boltzmann or generalized Born and surface area continuum solvation. The $G_{non-pol}$ is calculated using linear regression to solvent accessible surface area. *T* refers to the temperature and *S* to the entropy. The entropy change upon ligand binding is considered to be negligible and hence not used for relative binding free energy calculations [69]. Details about binding free energy calculations can be found elsewhere [70–72].

9. Mathematical measures

The goodness of a binary classification system can be given by the following measures

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$f1 = \frac{2 * TP}{2 * TP + +FP + FN}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

where, *TP* refers to true positive, *TN* to true negative, *FP* to false positive, *FN* to false negative. Out of all these measures, the Matthews Correlation coefficient (*MCC*) is the most balanced as it has all 4 terms – *TP*, *TN*, *FP* and *FN* [73].

The linear dependence of 2 variables *x* and *y* belonging to a sample size of n, was calculated using the Pearson's correlation coefficient (or correlation coefficient- $r_{x,y}$)

$$r_{x,y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

We have also used percentile rank of a score, which refers to the percentage of score that is lesser than or equal to the score. In case multiple entries had the same score the average percentage ranking of the scores were assigned to all the entries.

Chapter 3 - Structural study of protein-protein interfaces

1. **Creation of chain-chain and domain-domain interface library**
2. **Structure based clustering of the interface library**
3. **Interfaces can be structurally similar irrespective of the fold of the proteins**
4. **Amino acid pair preference at chain-chain and domain-domain interfaces**
5. **Small molecule binding sites at chain-chain interfaces**

1. Introduction

Proteins are the workforce of the cell that interact with one another and other biomolecules to carry out its functions. It has been estimated that ~80% of the proteins work in complexes [74]. Protein-protein interactions play important role in various biological processes [75,76]. Identifying these interactions can help explain the functioning of various proteins and the basis of various diseases [77,78]. These interactions can be identified using high throughput proteomics based experimental procedures [79], but these experiments do not provide any structural information. Various databases like Database of Interacting Proteins (DIP) [80], Biomolecular Interaction Network Database (BIND) [81,82], Molecular Interaction Database (MINT) [83], Interactome3D [84] etc. contain a list of experimentally validated interactions. However most of these databases, except Interactome3D lack structural information.

The 3D structure of these complexes can be identified by structural studies such as X-ray crystallography, NMR, cryo-EM. The 3D structure of these complexes can help explain its mechanism and functioning and is required to understand the repertoire of cellular pathways [85–88]. However, the experiments are costly, labor intensive and time consuming [89]. Though the number of structures of proteins are increasing over the years, still challenges are faced in crystallizing large protein (multi-domain proteins) or complexes [90]. Hence computational techniques can be used to model complexes of proteins [91–97]. These techniques face two challenges – sampling and scoring. The sampling involves the construction of plausible models of the protein complexes, whereas the scoring involves scoring these complexes to identify near-native complexes from non-native complexes.

Large proteins contain multiple domains, which are defined as an independent folding, evolving and structural units in proteins. Two proteins chain can fuse because of gene fusion leading to the formation of two protein domains in a single protein. Similarly, two domains of the same protein can break during evolution into independent chains [98]. During these events, a chain-chain interface can convert to a domain-domain interface or vice versa. Hence the interface across different domains can be structurally similar to that between different chains. The domain definitions/boundaries of individual proteins have been characterized in SCOP/SCOPe [99,100] or CATH [101–103], which can be utilized to identify interfaces between different protein domains.

Multiple libraries have been developed in the past to characterize protein-protein/domain-domain interfaces such as 3DID [104–106], PIBASE [107], SCOPPI [108], SNAPPI-DB [109], SCOWLP [110,111], ProtCID [112], QSbio [113]. A large number of these databases classify and cluster the interfaces to show similarities between different interfaces or study specific properties of the interfaces such as conservation, the importance of water etc. Techniques such as PRISM [91], InterComp [114] utilizes the interfaces as templates to model protein complexes.

Multiple studies have characterized protein-protein interfaces based on structure, packing, chemical complementarity, conservation etc. [115–120]. The previous study by Aloy *et al.* suggests that close homologs interact in a similar manner [121]. However, another study by Mika *et al.* suggests that such protein interaction prediction based on sequence homology holds only when the sequences are ~80% similar and if they belong to the same species [122]. Protein interfaces across different families can also be topologically different within the same superfamily [123]. Close homologues sometimes utilize different interfaces for interaction as shown in lectins [124], ASSP proteins that regulate apoptosis [125], bacterial chemotaxis proteins [126] etc. However, recent literature indicates that the structural repertoire of protein interfaces are degenerate and close to complete [127,128] and nature reuses similar interfaces across different proteins. Hence a composite library of such observed protein-protein/domain-domain interfaces will be useful in understanding and modeling protein complexes. Another way of modeling protein complexes involves docking one protein onto another, however, template based modeling of protein complexes has been shown to improve the predictions as compared to docking [129–131].

We created a library of all known interfaces between different proteins (chain-chain interface) or domains (domain-domain interface). We combined domain-domain and chain-chain interfaces as protein domains might break into different chains or different protein chains might fuse to form domains in a protein. Hence the domain-domain and chain-chain interfaces may be similar. We compared the SCOPe and CATH based domain definition during the process of library creation and utilized the CATH based definition because of its larger coverage of the PDB. We also compared the amino acid pair preference between the domain-domain and chain-chain interfaces. We clustered interfaces belonging to the same fold based on structural similarity, to identify how interfaces interact. Fold independent clustering was not possible given the volume of data. With certain examples, we showed how topologically different folds could have a structurally similar interface. Previous studies pointed towards the usage of a topologically independent structural match for a search of templates to model protein complexes [114]. In addition, previous and ongoing work in our lab has shown that proteins irrespective of their topology can be structurally similar at the small molecule ligand binding site [23] or DNA binding site (unpublished data). Hence a topology independent structural comparison (a structural match without taking into consideration the secondary structure) to identify templates for modeling protein complexes will help identify templates (to model protein complexes) for a larger number of proteins. However, the modeling of such complexes is beyond the scope of the study of this thesis and will be dealt with in the future.

2. Methods

2.1. Database of interfaces

All multi-chain and multi-domain (based on CATH/SCOPe domain definition) complexes were extracted from the PDB. In order to remove protein-peptide interfaces, we only considered those interfaces whose individual chain had at least 50 residues. The accessible surface area for the individual protein chains/domains and all possible binary protein chain/domain complexes were calculated using MODELLER [132]. Interfaces with greater than 400 Å^2 change in solvent accessible surface area were retained. The cut off was used to filter crystallographic artefacts from biologically relevant interfaces in lines with the PQS server [133–136]. There might be a few crystal artefacts with whose change in solvent accessibility >400 Å^2 [137]. A higher cut off for change in solvent accessible surface area would have removed artefacts better, but it would have come at the cost of losing many true interfaces. A study by Zhu *et al.* [138] has shown that the percentage of crystal artefacts reduce substantially (~30%) around 400-500 Å^2. We hope that the misleading contributions due to these artefacts would eventually be picked by the scoring schemes during evaluations. All the residues having at least one atom within 8 Å of another atom from a different chain/domain were used to create the library of the residues at the interface. The interface library was created such that the number of chains/domains for a chain-chain/domain-domain interface was 2. In case of oligomeric interfaces (>2 chains), all the possible combinations of homo/hetero-dimeric interfaces were created which follow the criterion for interface selection as described above.

Both SCOPe and CATH were used to define the domain boundaries for domain-domain interfaces. However, unless otherwise mentioned a domain-domain interface refers to domains based on CATH domain definition.

2.2. Clustering of interfaces

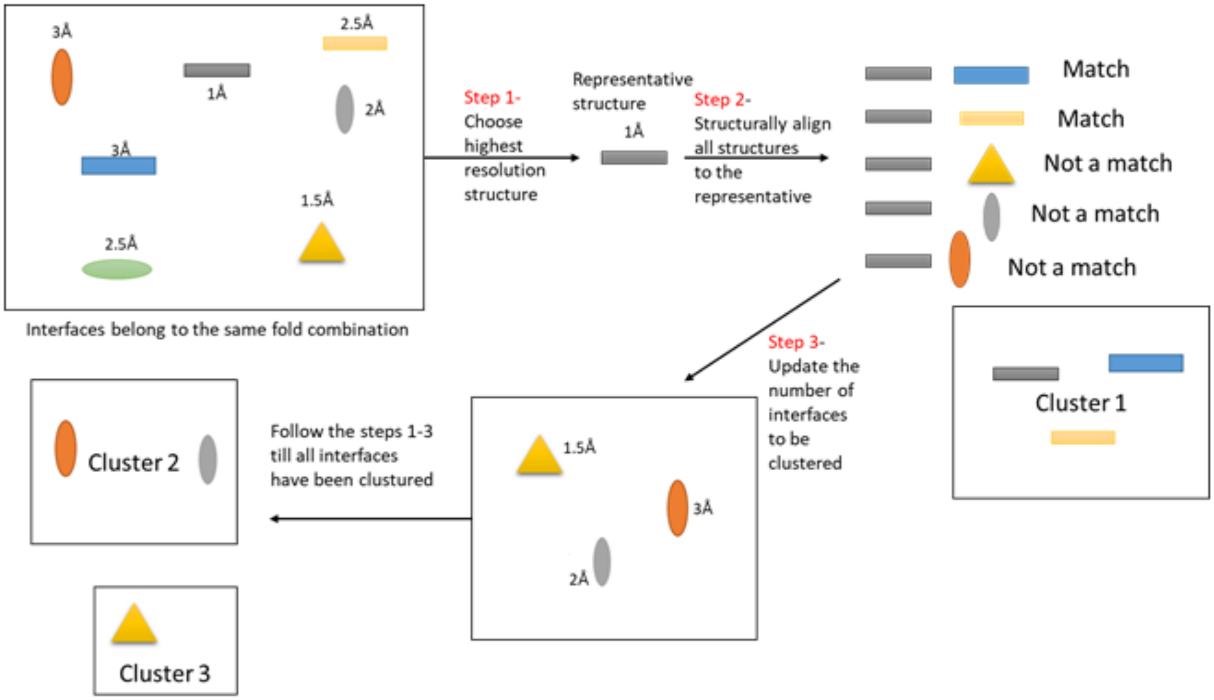

3Å
2.5Å
1Å
2Å
3Å
1.5Å
2.5Å
Interfaces belong to the same fold combination
Step 1-
Choose highest resolution structure
Representative structure
1Å
Step 2-
Structurally align all structures to the representative
Match
Match
Not a match
Not a match
Not a match
Step 3-
Update the number of interfaces to be clustered
Cluster 1
1.5Å
3Å
2Å
Follow the steps 1-3 till all interfaces have been clustured
Cluster 2
Cluster 3

Figure 1 – Flowchart explaining the clustering of the interface library. The different interfaces belonging to the same fold combination are shown in oval, triangle and rectangle in different colors. The resolution of the structure has been mentioned

alongside. The highest resolution structure i.e. grey rectangle is the first representative structure, all other interfaces are structurally aligned to it and considered a match based on a specific criterion. The interfaces, which did not structurally align forms the new set, from which the representative structure is selected, and the steps repeated till all interfaces have been clustered.

An interface containing two chains A and B belonging to fold c and d respectively is said to belong to the fold combination c-d. All interfaces made of the same combination of folds (called fold combination) were clustered together hierarchically such that the first chosen representative had the highest resolution. All interfaces were compared to the representative using CLICK [18] (a topology independent structural superimposition tool) with C^{α} and C^{β} atoms as representatives for the superimposition. The interface was clustered with the representative if the structure overlap was >80% and RMSD was <1.5 Å. A new representative was chosen from the remaining unclustered interfaces and the same procedure was repeated. This was iterated till all the interfaces were assigned to a cluster or only one interface was left, which then forms the only member of its cluster.

In cases where the number of PDBs in each fold combination was greater than 1000, then fold combination was broken into smaller sets of a maximum of 600 interfaces each, to reduce the number of structural comparisons to perform during clustering.

2.3. Pair preferences of the amino acid residues at protein-protein verses domain-domain interfaces

A residue-residue interaction profile was calculated for all side chain-side chain interactions using the same statistical potential developed for PIZSA [35,37]. For details about the methods used for calculation of these statistical potentials, refer to Dhawanjewar *et al.* [37] . The scoring scheme used was the ratio of the observed probability of the interface residue pair to that of the expected probability of the interface residue pair. To prevent overrepresentation of certain sequences, the PDBs were culled using PISCES [139] such that the maximum sequence identity was 40% and the resolution was 4 Å.

2.4. Prediction of the binding site at protein-protein interfaces

The binding site for all the protein chains was calculated using DEPTH [12] such that the number of water molecules for bulk solvent description was at least 4. Evolutionary information too was used during the computation (Refer to Chapter 2 Section 2 for details).

3. Results

3.1. Comparison of SCOPe and CATH domain definition

The number of folds in SCOPe (version – SCOPe-2.07-stable) and CATH (version - b.20180915) are 1,457 and 1,391 respectively. Out of these 377 and 282 folds (for SCOPe and CATH respectively) have no other domain interacting partner in the PDB. Only 31,063 PDBs had domain definitions (other than C terminal tag defined by tag l.1.1) according to SCOPe, whereas 114,839 PDBs had domain definitions according to CATH. The number of assigned multidomain PDBs according to SCOP is 17,303 while that according to CATH is 43,784. SCOPe assigns 57 proteins wherein a single domain is defined such that it spans multiple chains, whereas CATH has no such anomalous cases. Out of 112,043 chain-chain interfaces as observed in the PDB, 49,888 chain-chain interfaces did not have a fold assigned to at least 1 chain according to SCOPe definition, while 22,110 chain-chain interfaces did not have a fold classification according to CATH. Hence, because of the larger coverage of the domain definition in CATH for the PDB as compared to SCOPe, the CATH domain definition has been used for all future purposes.

3.2. Interface library

The number of chain-chain interfaces are 112,043 (belonging to 42,254 PDBs) and that of domain-domain interfaces are 66,442 (belonging to 28,085 PDBs). Out of the 112,043 chain-chain interfaces, 89,933 interfaces had the interacting residues assigned to 2,444 fold combination. The remaining 22,110 chain-chain interfaces did not have one or both its chains assigned to a CATH fold. Out of these 89,933 chain-chain interfaces, 55,352 (62%) interfaces are between chains belonging to the same fold, indicating a predominant homo-oligomeric association between different chains. These 55,352 interfaces belonged to 514 CATH fold combinations.

The 66,442 domain-domain interfaces interacted with each other in a 1,135 fold combination. Of these, 47,839 (72%) domain-domain interfaces were between domains belonging to different folds, indicating a predominant hetero-oligomeric association between different protein domains. The remaining 18,603 interfaces that belong to the same fold belongs to 107 CATH fold combinations.

Domain-domain interfaces may either contain residues which belong to the junction between 2 domains (i.e. the C-terminal of a domain is sequential neighbor of the N-terminal of another domain) or it may contain residues where they are not sequential neighbors (i.e the two interacting domains are either separated by other domain/s or separated by a stretch of residues with no domain annotation). The domains in the former case may come in contact because of the interface residues between the two domains being sequential neighbors. To check for the latter i.e. domains that have come in contact even though the constituent interface residues were not sequential neighbors we considered an empirical cut-off of 10 residues between the interfaces of the two constituent domains. Out of all the 66,442 domain-domain interfaces, only 6,684 interfaces (~10%) followed this criterion.

3.2.1. Number of interfaces per fold combination

Both the chain-chain and domain-domain interfaces interacted with each other in 3,065 unique fold combinations with 514 fold combinations containing both domain-domain and chain-chain interfaces. Out of the 3,065 fold combinations, 585 fold combinations had only 1 interface (1 chain-chain/domain-domain interface in 1 fold combination) and 2,502 (82%) fold combinations had <=30 interfaces (Figure 2). However, 21 fold combinations had >1000 interfaces in it, which were broken into subsets of 600 interfaces or less. Rossmann fold had the maximum number of interfaces (14,844 interfaces), followed by Immunoglobulin-like fold (6,528 interfaces) and Glutamine Phosphoribosylpyrophosphate, subunit 1, domain 1 fold (6,318 interfaces).

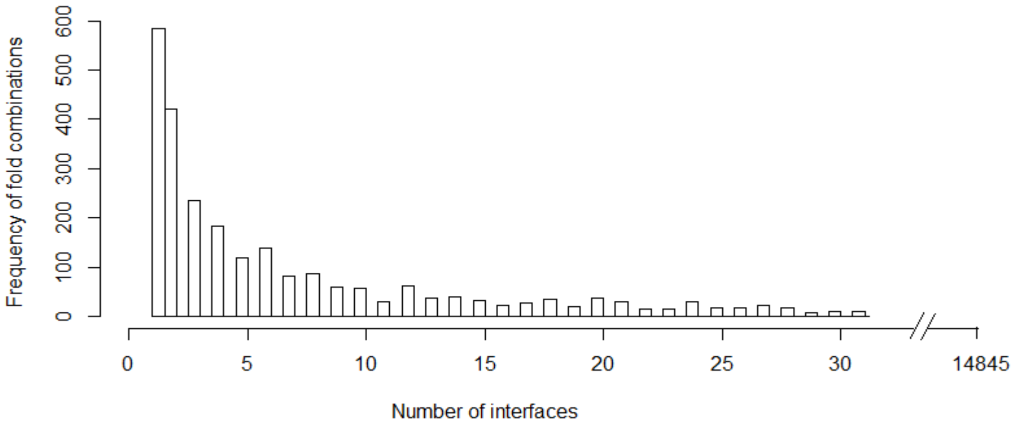


Figure 2 – Histogram showing the number of fold combinations (y-axis) having a particular number of interfaces (x-axis).

3.2.2. Interactions of a fold with other folds

A CATH fold can interact with anything between 1 to 183 different CATH folds (considering the fold is interacting with 1 fold at an interface). According to the present representation in PDB, only 1109 folds (out of 1,391 CATH folds) interact with one another, of which 1092 folds interact with <30 other folds. Rossman fold (CATH ID- 3.40.50), for instance, interacts with 283 folds in different orientations (Table 1). Single alpha-helices involved in interacting with coiled-coil or other helix-helix interfaces (CATH ID- 1.20.5) are seen to interact with 90 different folds. This can be because of the diversity of the sequences that can take up these particular folds [140]. However certain chain-chain interfaces have more than 2 folds forming the interface. In such cases, a particular fold can be present in combinations with 1 to 388 other folds.

Table 1 – Folds that interact with more than 30 other folds in a binary interaction

CATH ID	**Fold Name**	**Number of interacting folds in a binary interaction scenario**
3.40.50	Rossmann fold	283
2.60.40	Immunoglobulin fold	93

1.20.5	Single alpha helix involved in coiled-coil or other helix-helix interactions	90
3.30.70	Alpha-beta plaits	81
1.10.287	Helix hairpins	79
1.10.10	Arc-repressor mutants, subunit A	78
3.20.20	TIM barrels	66
2.60.120	Jelly rolls	57
1.25.40	Ser-Thr protein phosphatase 5, Tetratricopeptide repeat	54
1.20.58	Methane monooxygenase hydrolase Chain G, Domain 1	54
2.40.50	OB-fold (Dihydrolipoamide acetyltransferase. E2P)	48
1.20.120	4 helix bundle (Hemierythrin Met Subunit A)	40
3.10.20	Ubiquitin like rolls	38
3.30.420	Nucleotidyl transferase domain 5	37
2.30.30	SH3 type barrel	37

3.3. Clustering of the interface library

All the 156,375 (89,933 chain-chain and 66,442 domain-domain) interfaces were clustered into 27,317 clusters, such that all the interfaces in a cluster have structure overlap >=80% and interface RMSD <=1.5 Å (unless otherwise mentioned RMSD in the chapter refers to the interface RMSD) with respect to that of the representative PDB. Of these 27,317 clusters, 100 clusters had both chain-chain and domain-domain representatives. 12,877 (~47%) clusters only contain 1 PDB (Figure 3), which might be a result of the stringent RMSD and structure overlap criterion used during clustering. 25,951 (~95%) clusters contain less than 20 PDB per cluster (Figure 3). Certain folds like hemagglutinin-ectodomain chain B had 484 out of 567 interfaces clustered together with cluster representative 4gxx_BD. The other interfaces fell into different clusters because of the stringent criterion used during clustering (Figure 4). The other PDBs that did not cluster with 4gxx_BD had a structural overlap ranging between 70-80% and RMSD between 1.5 Å – 2.4 Å.

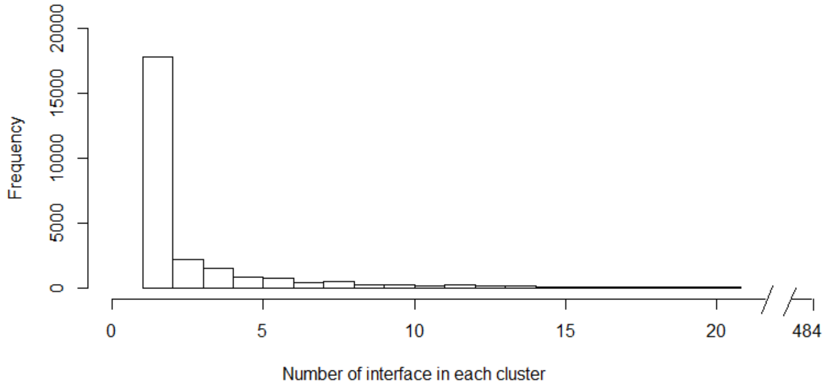


Figure 3 – Histogram showing the frequency of clusters (y-axis) having a particular number of interface in each cluster (x-axis).

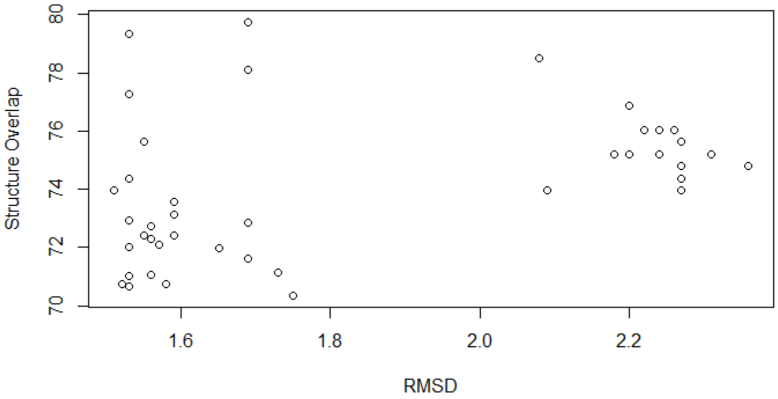


Figure 4 – RMSD vs Structure Overlap of the interfaces that were not put in the same cluster as the cluster representative 4gxx_BD.

3.3.1. Number of interface clusters per fold combinations

The number of clusters per fold combination ranges from 1 to 7,809. 1,623 fold combinations (out of 3,065 combinations) had only 1 cluster (585 fold combinations had only 1 interface and hence only had 1 cluster). 2,941 (96%) fold combinations had <20 clusters (Figure 5). 3043 (99%) fold combinations were clustered into less than 100 clusters.

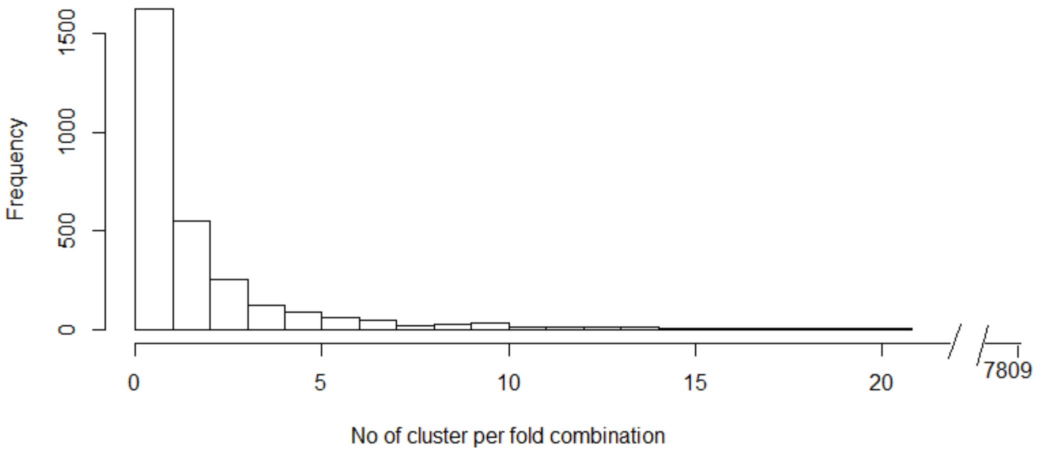


Figure 5 – Histogram showing the frequency of the number of clusters per fold combination.

14,844 interfaces of the Rossmann fold clustered into 7,809 clusters, 6,528 interfaces of the immunoglobulin-like fold clustered into 1,675 clusters and 3,598 interfaces belonging to TIM barrel fold clustered into 984 clusters. The higher number of clusters in these fold combinations was because of the stringency used in cut off selected for clustering, the diverse ways the same fold could interact and because multiple non-homologous proteins taking up these folds [141]. This results in multiple modes of interactions. Most of the interfaces (73%) have a high RMSD (>2 Å) and low structure overlap (<70%) when compared to each other (Figure 6). Hence, they interact differently with different types of proteins even though they belong to the same fold combination.

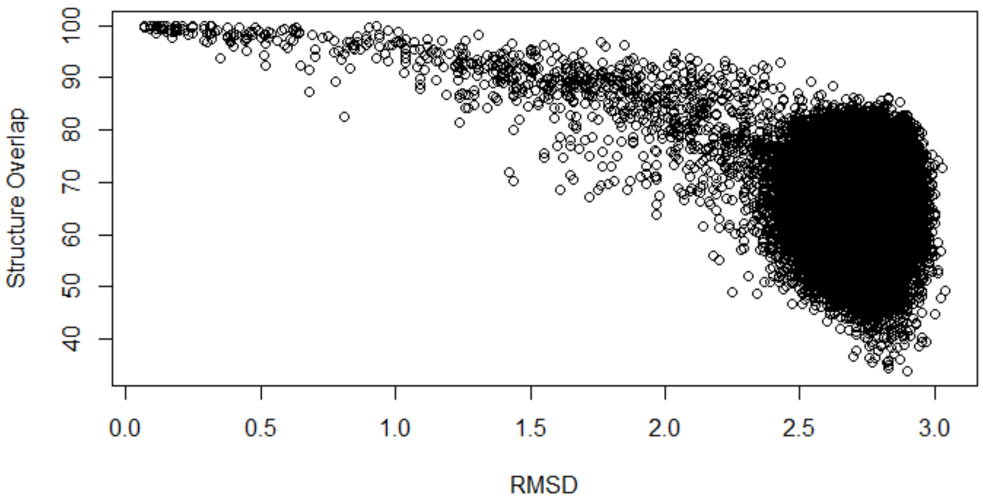


Figure 6 – Structure overlap (%) vs RMSD (Å) of the structural match between different interfaces belonging to the Rossmann fold. Most of the interfaces have high RMSD and low structure overlap with each other indicating multiple ways of interactions between Rossmann folds.

3.3.2. Certain chain-chain interfaces and domain-domain interfaces are structurally similar

514 fold combinations had both chain-chain and domain-domain interfaces. Out of the 66,442 domain-domain interfaces, 55,696 interfaces belong to these 514 folds. From these, 100 clusters had both chain-chain and domain-domain interface clustered together. One example is that of the domain-domain interface of Giardia dicer superimposed onto a chain-chain interface of the Nuclease domain of ribonuclease 3; with a structure overlap of 86% and an RMSD of 1.33 Å (Figure 7 A). The two proteins share a sequence identity of 23% and belong to the Ribonuclease iii N terminal endonuclease domain, Chain A fold. Another such example is that of the superimposition of the chain-chain interface of AVA_4353 protein onto the domain-domain interface of PhuS protein with a structure overlap of 91% and RMSD of 1.28 Å (Figure 7 B). The two proteins belong to the heme utilizing iron like fold and share no significant sequence similarity. Hence, sequentially unrelated proteins can have a structurally similar chain-chain and domain-domain interface.

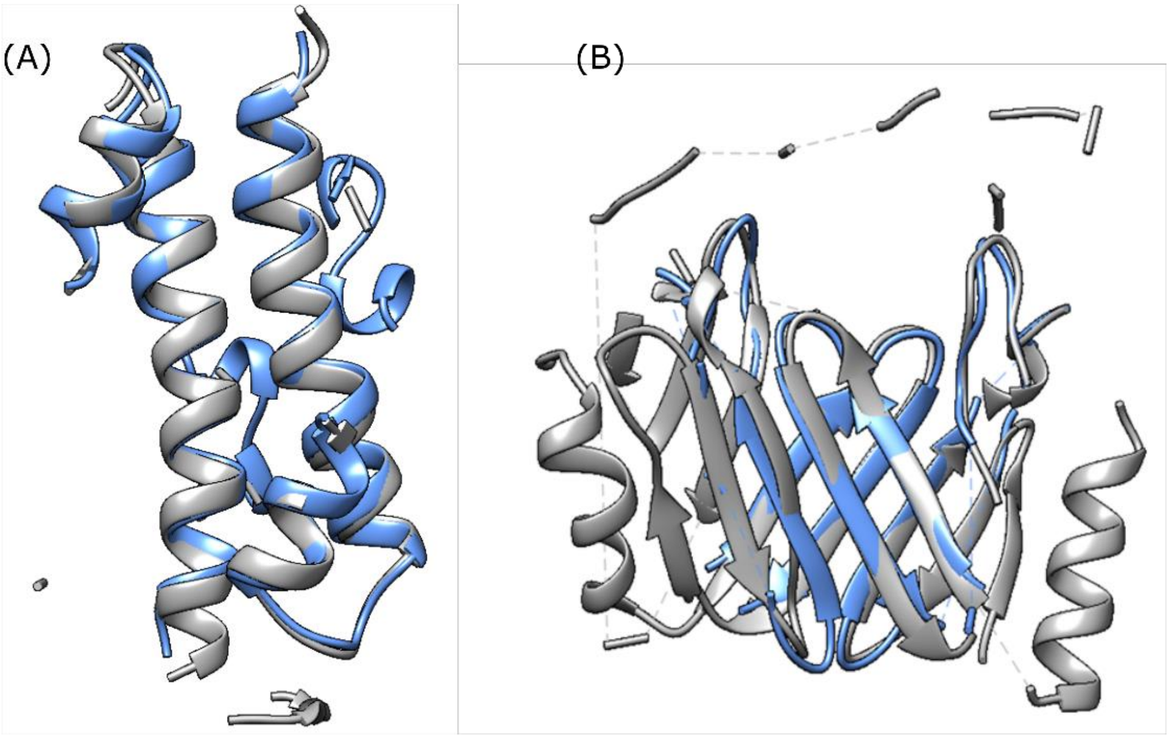


Figure 7 – (A) Chain-chain interface of Nuclease domain of ribonuclease3 (PDB – 3o2r_CD) in grey ribbons superimposed on the domain-domain interface of Giardia dicer (PDB – 2qvw_B) in blue ribbons (B)Chain-chain interface of AVA_4353 protein (PDB – 3FM2_AB) shown in blue ribbons superimposed on the domain-domain interface of PhuS protein (PDB – 4IMH_B) shown in grey ribbons

3.4. Examples of structurally similar protein-protein interfaces from different folds

The clustering of the interface library was limited to proteins belonging to the same fold combinations, as an all against all comparison of all the interfaces irrespective of their folds is computationally intensive and impossible with our present computation power. However, we compared few interfaces across different folds to check if there exists structural similarity of the interface irrespective of the fold the protein chains/domain belong to.

3.4.1. A fold interacting with different folds using the same geometry

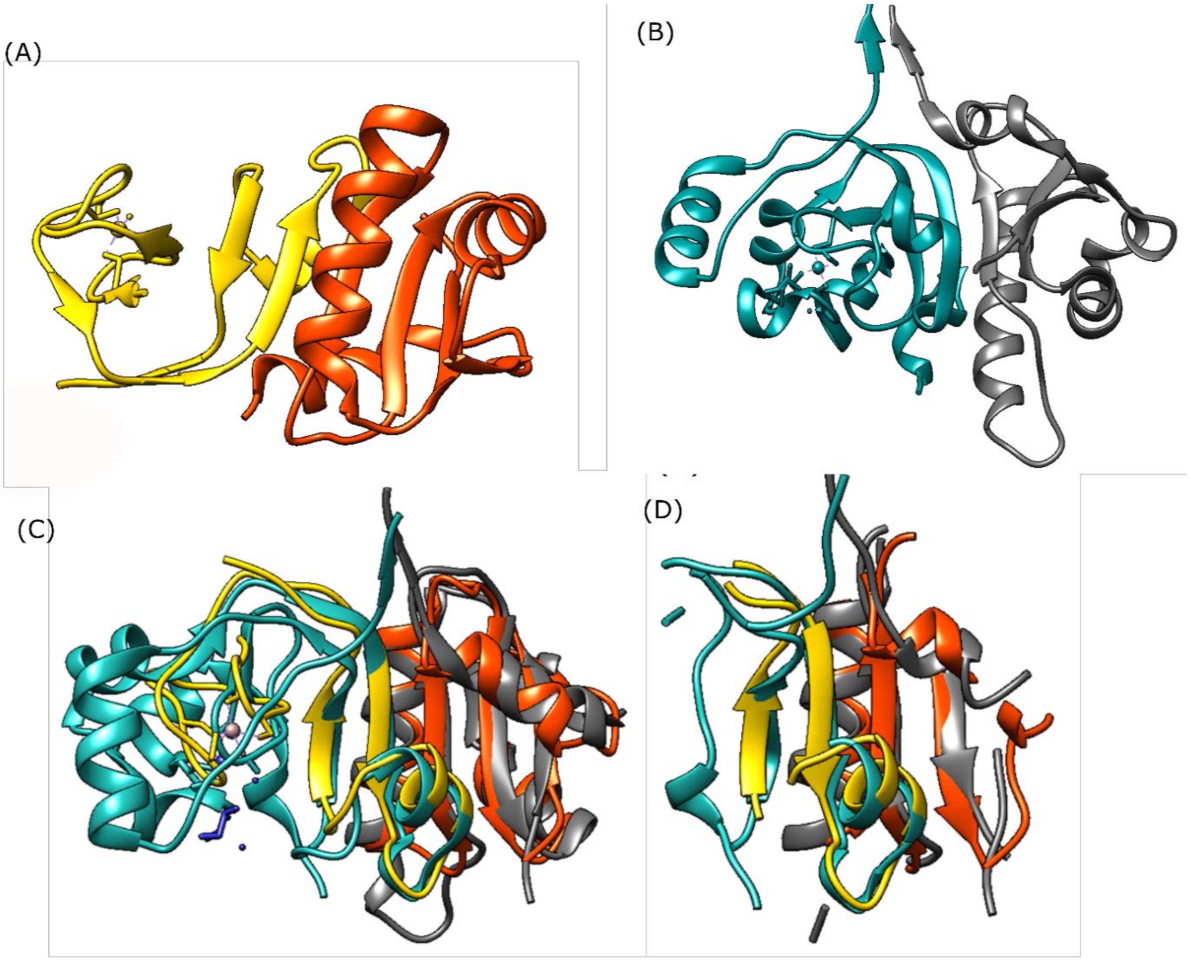


Figure 8 – (A) Complex of Nus G protein (PDB – 3LPE_G) (in orange ribbons) and DNA dependent RNA polymerase E (PDB – 3LPE_H) (in yellow ribbons) (B) Complex of SPT5 (PDB – 3H7H_B) (in grey ribbons) and SPT5 and SPT4 (PDB – 3H7H_A) (in cyan ribbons) (C) Superimposition of the complex of Nus G protein and DNA dependent RNA polymerase E onto the complex of (D) The interface residues from the same complexes following the same color scheme shows the structural similarity of the interface residues.

The NusG (Transcription antitermination protein) and the Transcription elongation factor SPT5 belong to the same fold of alpha-beta plaits and are 34% sequentially identical to each other. The NusG protein interacts with DNA dependent RNA polymerase E, which belongs to Ruberythrin Domain 2 fold whereas the SPT5 interacts with the Transcription elongation factor SPT4 belongs to Herpes Virus 1 fold. Even though, the interacting proteins (DNA dependent RNA polymerase and SPT4) to the two proteins (NusG and SPT5 respectively) belong to different folds and are only 29% identical sequentially the interacting interface is similar with a structure overlap of 93% and RMSD of 1.72 Å (Figure 8).

3.4.2. Interfaces belonging to different folds utilize the same geometry

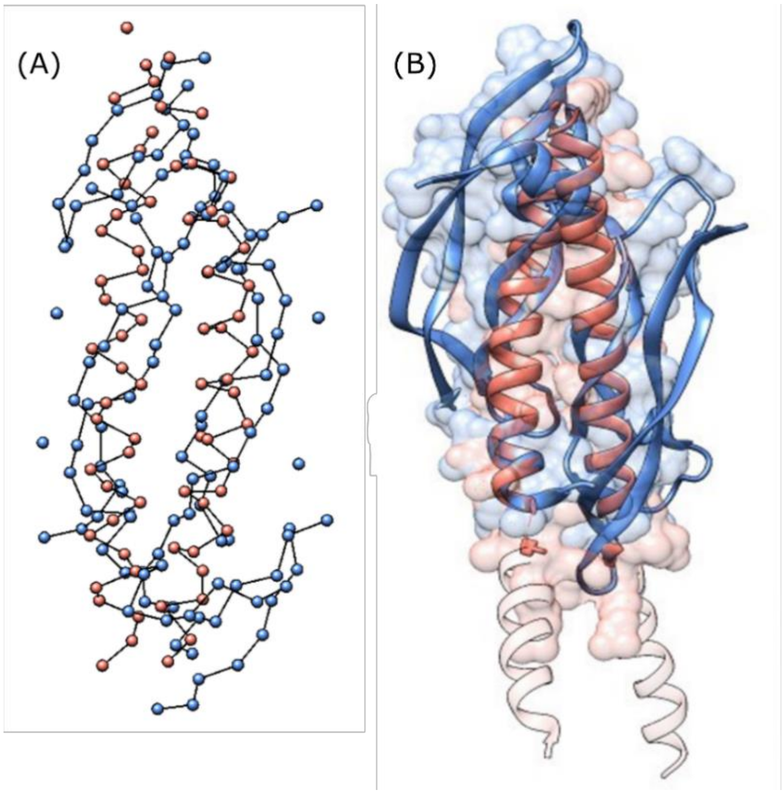


Figure 9 – (A) Superimposition of the interface of vascular endothelial growth factor A (PDB – 1mkk_AB) (in a blue ball and stick model of the C^{α} atoms) and AP1-c fos (PDB – 1s9k_ED) (in a salmon ball and stick model of the C^{α} atoms) (B) Superimposition of the two protein complexes (shown in ribbons) onto each other following the same color scheme.

The vascular endothelial growth factor A belongs to the cysteine knot cytokine protein whereas the complex between Transcription factor AP1 and c-fos belongs to the single alpha-helices involved in coiled-coil or other helix-helix interface fold. The interacting interface of one involves β-sheets (vascular endothelial growth factor A) whereas that of others involves α-helices (AP1-c fos). However, irrespective of the topology of the interface and 0% sequence identity, the two interfaces are structurally similar with a structure overlap of 78% and RMSD of 2.48 Å (Figure 9). In principle, we would want to learn about all such cases, but we are limited by computational resources.

3.5. Pair preference of the amino acid residues at protein-protein vs domain-domain interfaces

The amino acid preferences for a domain-domain interface, chain-chain interface, protein surface (residue depth<5) [137,142] were compared to each other (Figure 10). The protein surface has a higher preference for polar amino acids (Asp, Glu, Lys, Arg, Asn, Gln) while having a lower preference for non-polar amino acids (Phe, Met, Cys, Trp, Tyr, Ile, Leu) as compared to the chain-chain/domain-domain interface amino acid preference (Figure 10). The chain-chain and domain-domain interfaces have similar amino acid preferences (Figure 10).

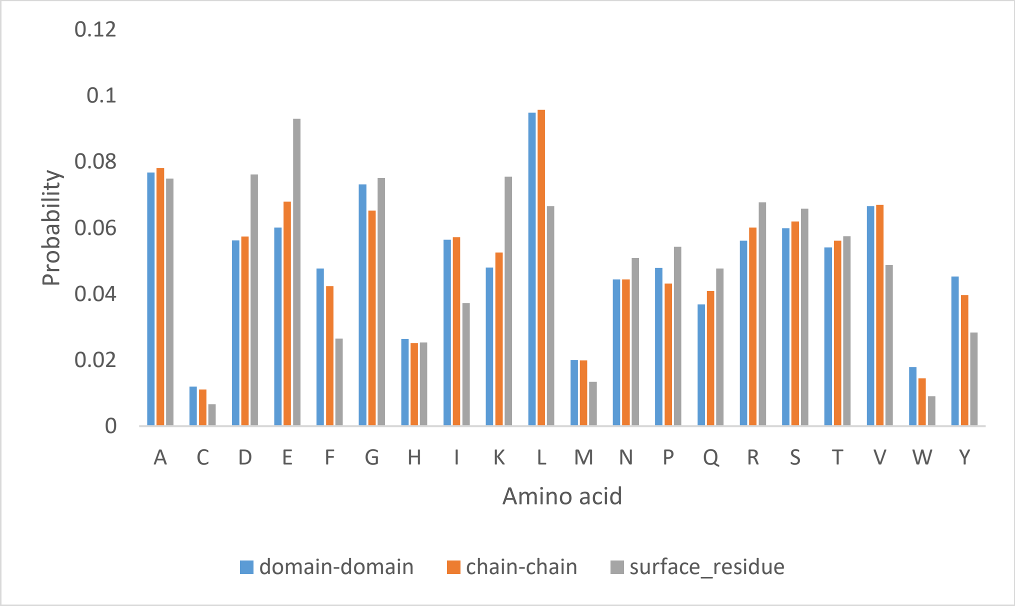
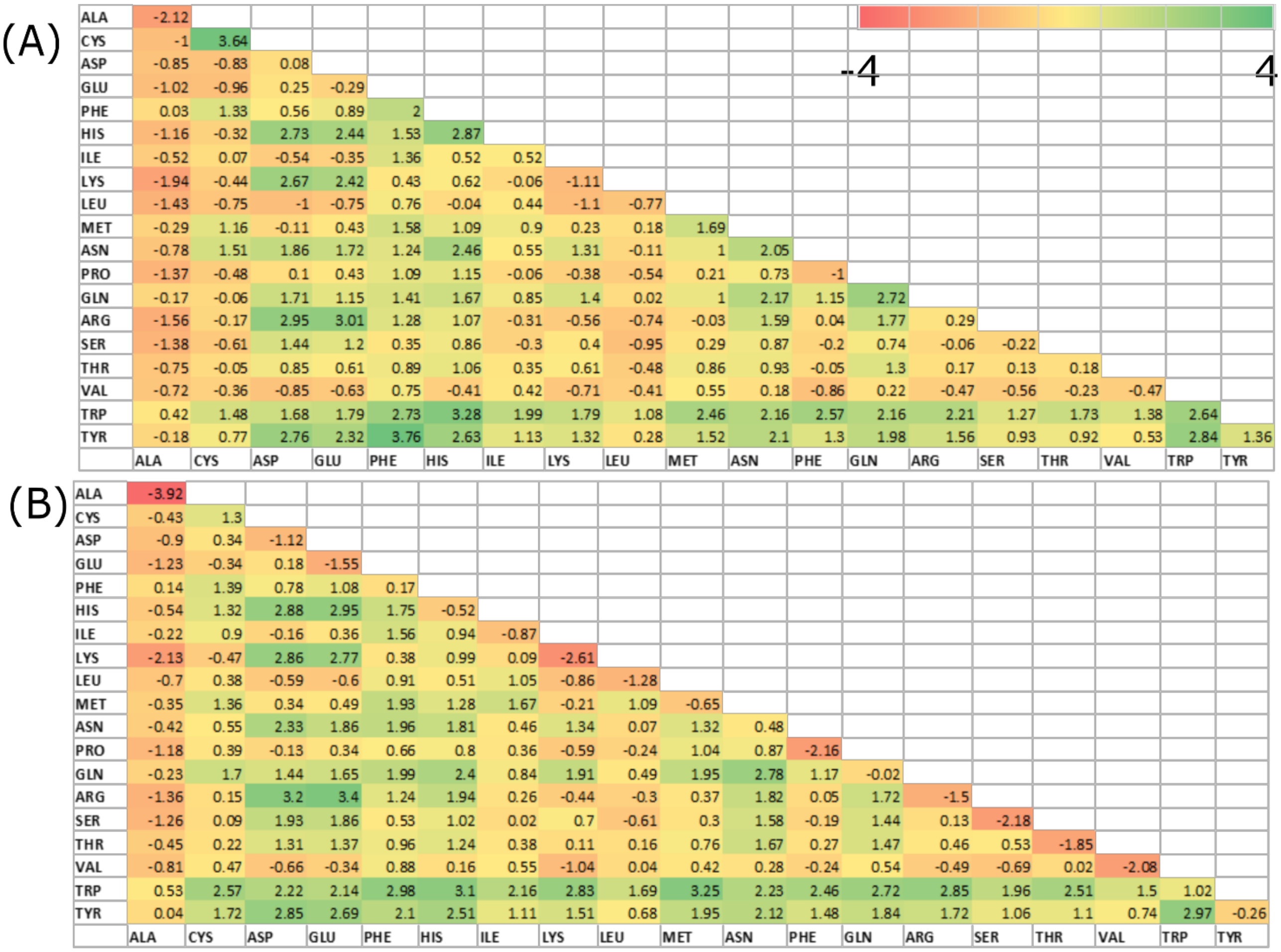


Figure 10 – Amino acid preference at domain-domain interfaces (blue bars), chain-chain interfaces (orange bars) and protein surface (grey bars)



(A)

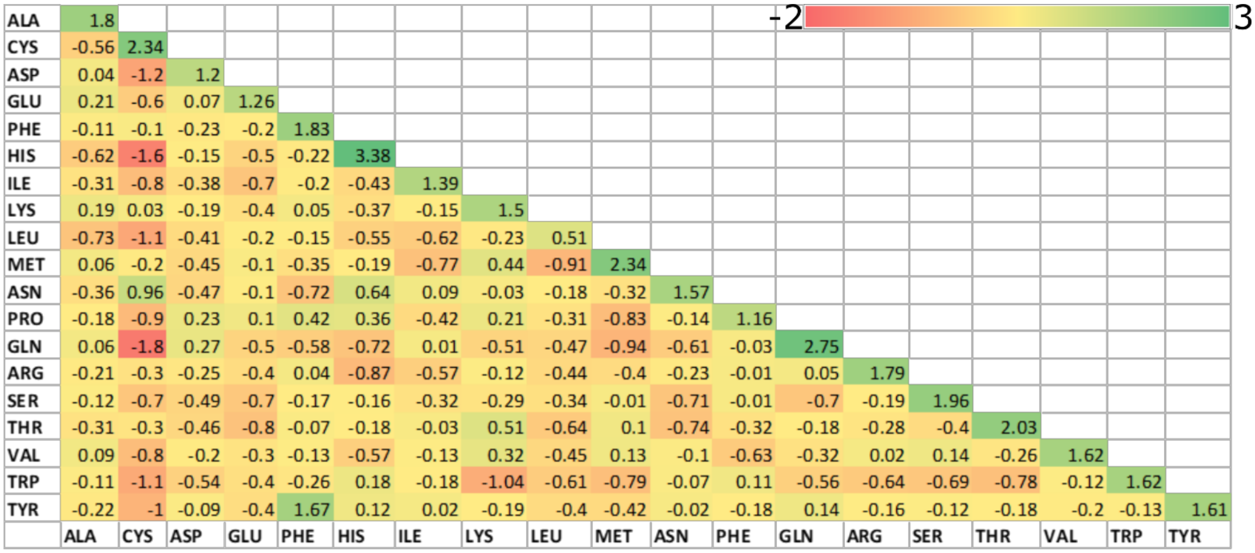
ALA	-2.12																		
CYS	-1	3.64																	
ASP	-0.85	-0.83	0.08																
GLU	-1.02	-0.96	0.25	-0.29															
PHE	0.03	1.33	0.56	0.89	2														
HIS	-1.16	-0.32	2.73	2.44	1.53	2.87													
ILE	-0.52	0.07	-0.54	-0.35	1.36	0.52	0.52												
LYS	-1.94	-0.44	2.67	2.42	0.43	0.62	-0.06	-1.11											
LEU	-1.43	-0.75	-1	-0.75	0.76	-0.04	0.44	-1.1	-0.77										
MET	-0.29	1.16	-0.11	0.43	1.58	1.09	0.9	0.23	0.18	1.69									
ASN	-0.78	1.51	1.86	1.72	1.24	2.46	0.55	1.31	-0.11	1	2.05								
PRO	-1.37	-0.48	0.1	0.43	1.09	1.15	-0.06	-0.38	-0.54	0.21	0.73	-1							
GLN	-0.17	-0.06	1.71	1.15	1.41	1.67	0.85	1.4	0.02	1	2.17	1.15	2.72						
ARG	-1.56	-0.17	2.95	3.01	1.28	1.07	-0.31	-0.56	-0.74	-0.03	1.59	0.04	1.77	0.29					
SER	-1.38	-0.61	1.44	1.2	0.35	0.86	-0.3	0.4	-0.95	0.29	0.87	-0.2	0.74	-0.06	-0.22				
THR	-0.75	-0.05	0.85	0.61	0.89	1.06	0.35	0.61	-0.48	0.86	0.93	-0.05	1.3	0.17	0.13	0.18			
VAL	-0.72	-0.36	-0.85	-0.63	0.75	-0.41	0.42	-0.71	-0.41	0.55	0.18	-0.86	0.22	-0.47	-0.56	-0.23	-0.47		
TRP	0.42	1.48	1.68	1.79	2.73	3.28	1.99	1.79	1.08	2.46	2.16	2.57	2.16	2.21	1.27	1.73	1.38	2.64	
TYR	-0.18	0.77	2.76	2.32	3.76	2.63	1.13	1.32	0.28	1.52	2.1	1.3	1.98	1.56	0.93	0.92	0.53	2.84	1.36
	ALA	CYS	ASP	GLU	PHE	HIS	ILE	LYS	LEU	MET	ASN	PHE	GLN	ARG	SER	THR	VAL	TRP	TYR

(B)

ALA	-3.92																		
CYS	-0.43	1.3																	
ASP	-0.9	0.34	-1.12																
GLU	-1.23	-0.34	0.18	-1.55															
PHE	0.14	1.39	0.78	1.08	0.17														
HIS	-0.54	1.32	2.88	2.95	1.75	-0.52													
ILE	-0.22	0.9	-0.16	0.36	1.56	0.94	-0.87												
LYS	-2.13	-0.47	2.86	2.77	0.38	0.99	0.09	-2.61											
LEU	-0.7	0.38	-0.59	-0.6	0.91	0.51	1.05	-0.86	-1.28										
MET	-0.35	1.36	0.34	0.49	1.93	1.28	1.67	-0.21	1.09	-0.65									
ASN	-0.42	0.55	2.33	1.86	1.96	1.81	0.46	1.34	0.07	1.32	0.48								
PRO	-1.18	0.39	-0.13	0.34	0.66	0.8	0.36	-0.59	-0.24	1.04	0.87	-2.16							
GLN	-0.23	1.7	1.44	1.65	1.99	2.4	0.84	1.91	0.49	1.95	2.78	1.17	-0.02						
ARG	-1.36	0.15	3.2	3.4	1.24	1.94	0.26	-0.44	-0.3	0.37	1.82	0.05	1.72	-1.5					
SER	-1.26	0.09	1.93	1.86	0.53	1.02	0.02	0.7	-0.61	0.3	1.58	-0.19	1.44	0.13	-2.18				
THR	-0.45	0.22	1.31	1.37	0.96	1.24	0.38	0.11	0.16	0.76	1.67	0.27	1.47	0.46	0.53	-1.85			
VAL	-0.81	0.47	-0.66	-0.34	0.88	0.16	0.55	-1.04	0.04	0.42	0.28	-0.24	0.54	-0.49	-0.69	0.02	-2.08		
TRP	0.53	2.57	2.22	2.14	2.98	3.1	2.16	2.83	1.69	3.25	2.23	2.46	2.72	2.85	1.96	2.51	1.5	1.02	
TYR	0.04	1.72	2.85	2.69	2.1	2.51	1.11	1.51	0.68	1.95	2.12	1.48	1.84	1.72	1.06	1.1	0.74	2.97	-0.26
	ALA	CYS	ASP	GLU	PHE	HIS	ILE	LYS	LEU	MET	ASN	PHE	GLN	ARG	SER	THR	VAL	TRP	TYR

Figure 11 – Statistical potential for the pairwise interaction between 2 amino acids from different chains for (A) chain-chain (B) domain-domain interfaces. Each of the pair preference is colored based on their pair preference score.

Along with checking for the amino acid preferences at a chain-chain vs domain-domain interface, we also checked if the amino acid at the two types of interfaces had different pair preferences. The preference of an amino acid pair to interact with one another (such that each amino acid from a chain/domain has at least 1 atom within 4 Å of the interacting chain/domain) in a chain-chain and domain-domain interface was calculated and a statistical potential was computed using the same formulation as PIZSA (Figure 11). The main chain of Gly might form important main chain-main chain interactions or main chain (from Gly)-side chain (from interacting amino acid) interactions. A detailed study of these statistical potentials for chain-chain interfaces at 3 different distance cut-offs of 4 Å, 6 Å and 8 Å for side chain-side chain, side chain-main chain and main chain-main chain interactions can be found at Dhawanjewar *et al.* [37]. However, for this study, we limited ourselves to studying side chain-side chain amino acid pair preferences. Since Gly lacks a side chain, no statistical potential values were computed for pairs containing Gly. For the calculation of the statistical potential for a chain-chain interface a total of 5,571 PDBs (forming 10,836 interfaces) were used, while for the domain-domain interface 2,241 PDBs (forming 2,839 interfaces) were used (sequence being culled at 40% identity). The pairwise statistical potential for a domain-domain interface ranges between -3.9 to 3.4 whereas that of the chain-chain interface range between -2.1 to 3.8. A negative value indicates that the residue pair is noticed less than expected by random chance (indicating an undesirable interaction) whereas a positive value indicates that the residue pair is noticed greater than by random chance (indicating a favored interaction). The overall trends for the amino acid pair preferences at a chain-chain interface and domain-domain interface look similar (Figure 10), however, a Wilcox paired test show that the two substitution trends are dissimilar with a p-value of 2.2e-16.



ALA	1.8																		
CYS	-0.56	2.34																	
ASP	0.04	-1.2	1.2																
GLU	0.21	-0.6	0.07	1.26															
PHE	-0.11	-0.1	-0.23	-0.2	1.83														
HIS	-0.62	-1.6	-0.15	-0.5	-0.22	3.38													
ILE	-0.31	-0.8	-0.38	-0.7	-0.2	-0.43	1.39												
LYS	0.19	0.03	-0.19	-0.4	0.05	-0.37	-0.15	1.5											
LEU	-0.73	-1.1	-0.41	-0.2	-0.15	-0.55	-0.62	-0.23	0.51										
MET	0.06	-0.2	-0.45	-0.1	-0.35	-0.19	-0.77	0.44	-0.91	2.34									
ASN	-0.36	0.96	-0.47	-0.1	-0.72	0.64	0.09	-0.03	-0.18	-0.32	1.57								
PRO	-0.18	-0.9	0.23	0.1	0.42	0.36	-0.42	0.21	-0.31	-0.83	-0.14	1.16							
GLN	0.06	-1.8	0.27	-0.5	-0.58	-0.72	0.01	-0.51	-0.47	-0.94	-0.61	-0.03	2.75						
ARG	-0.21	-0.3	-0.25	-0.4	0.04	-0.87	-0.57	-0.12	-0.44	-0.4	-0.23	-0.01	0.05	1.79					
SER	-0.12	-0.7	-0.49	-0.7	-0.17	-0.16	-0.32	-0.29	-0.34	-0.01	-0.71	-0.01	-0.7	-0.19	1.96				
THR	-0.31	-0.3	-0.46	-0.8	-0.07	-0.18	-0.03	0.51	-0.64	0.1	-0.74	-0.32	-0.18	-0.28	-0.4	2.03			
VAL	0.09	-0.8	-0.2	-0.3	-0.13	-0.57	-0.13	0.32	-0.45	0.13	-0.1	-0.63	-0.32	0.02	0.14	-0.26	1.62		
TRP	-0.11	-1.1	-0.54	-0.4	-0.26	0.18	-0.18	-1.04	-0.61	-0.79	-0.07	0.11	-0.56	-0.64	-0.69	-0.78	-0.12	1.62	
TYR	-0.22	-1	-0.09	-0.4	1.67	0.12	0.02	-0.19	-0.4	-0.42	-0.02	-0.18	0.14	-0.16	-0.12	-0.18	-0.2	-0.13	1.61
	ALA	CYS	ASP	GLU	PHE	HIS	ILE	LYS	LEU	MET	ASN	PHE	GLN	ARG	SER	THR	VAL	TRP	TYR

Figure 12 – Difference in amino acid pair potential (Domain domain interface score – Chain chain interface score) colored based on the difference.

The difference between the domain-domain score and the chain-chain scores were computed (Figure 12). Self-pairs are seen to be preferred in chain-chain interfaces as compared to domain-domain interfaces. Such self amino acid pairing at chain-chain interfaces are important for packing and its absence would lead to a hollow channel running through the 2 fold axis of the homodimer, leading to void between the interfaces [143]. We notice self-amino acid pairing (other than Cys, Asn, Phe and Trp) are not favored at the domain-domain interface i.e. have negative values. However, most self-amino acid pairing other than Ala, Lys, Leu, Phe and Val (Glu and Ser to some extent) are favored at the chain-chain interface. This can predominantly be because ~62% of the chain-chain interfaces are homo-oligomers as compared to ~28% of the domain-domain interfaces being homo-oligomeric (Results Section 3.2). We notice that Cystine interactions in domain-domain interfaces have lesser scores than that of the chain-chain interfaces, probably because of its lower natural abundance, and fewer number of domain-domain interfaces (2839 domain-domain interfaces compared to 10836 chain-chain interfaces).

3.6. Small molecule binding site at protein-protein interfaces

The small molecule binding site of the individual chains that form the protein-protein interfaces was predicted using DEPTH [12]. The overlap of the residues constituting the binding site and the interface was calculated (Figure 13). Out of the 112,043 chain-chain interfaces, 74,849 interfaces had at least one chain with a 30% overlap between the predicted binding site and the interface residue. The predicted binding site (using DEPTH) can be used to dock/predict small molecules that could bind at the interface, hence disrupting the formation of the complex [144–147].

The binding site on the Nipah virus glycoprotein had 30% overlap with that of the interface residues with the ephrin B2 receptor of human (PDB – 2VSM). Autodock [47] and DOCK [48] were used to predict the small drug like molecules that would go and bind the predicted binding site of the glycoprotein, hence preventing its interactions with ephrin-B2 receptor. Details about the methods and the computational stability of this complex are mentioned in Chapter 9 Methods Section 2.2, 2.3 and Results section 3.3. Similarly, the interfaces that have an overlap with that of the predicted binding site can be targeted to prevent the association.

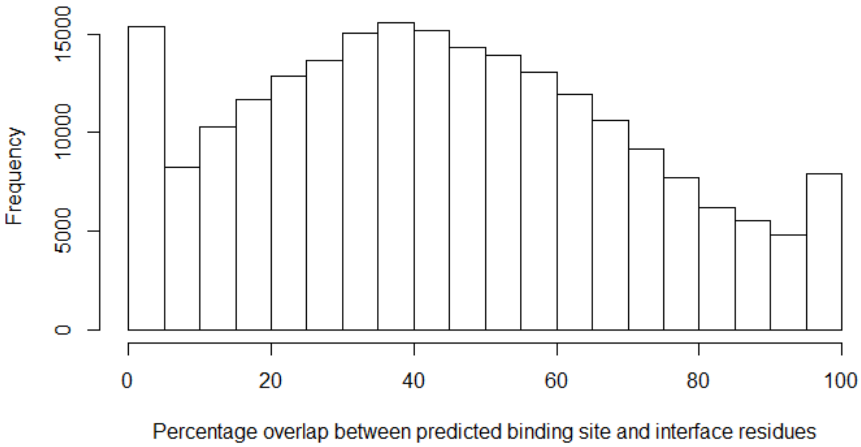


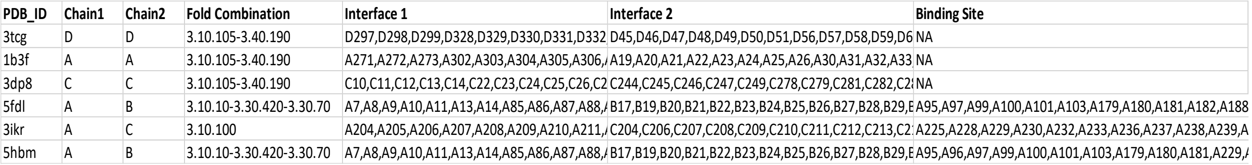
Figure 13 – Histogram showing the percentage overlap between the predicted small molecule binding sites with that of the interface residues

3.7. Database

The database containing the information about all the chain-chain and domain-domain interfaces can be accessed at -

http://www.iiserpune.ac.in/~madhusudhan/Neeladri_Sen_Thesis/interface_library.tsv.gz

This database contains the PDB ID, Chain ID(s), fold combination of the interface (CATH ID), residues at the interface, if it is a chain-chain or domain-domain interface. The last column contains the overlap between the interface and the binding site. It is set as NA for domain-domain interfaces (Figure 14). No analysis was done on chain-chain interfaces without any CATH definition, hence it has NA for fold combination and binding site columns in the table.



PDB_ID	Chain1	Chain2	Fold Combination	Interface 1	Interface 2	Binding Site
3tcg	D	D	3.10.105-3.40.190	D297,D298,D299,D328,D329,D330,D331,D332	D45,D46,D47,D48,D49,D50,D51,D56,D57,D58,D59,D6	NA
1b3f	A	A	3.10.105-3.40.190	A271,A272,A273,A302,A303,A304,A305,A306,	A19,A20,A21,A22,A23,A24,A25,A26,A30,A31,A32,A33	NA
3dp8	C	C	3.10.105-3.40.190	C10,C11,C12,C13,C14,C22,C23,C24,C25,C26,C2	C244,C245,C246,C247,C249,C278,C279,C281,C282,C2	NA
5fdl	A	B	3.10.10-3.30.420-3.30.70	A7,A8,A9,A10,A11,A13,A14,A85,A86,A87,A88,	B17,B19,B20,B21,B22,B23,B24,B25,B26,B27,B28,B29,	A95,A97,A99,A100,A101,A103,A179,A180,A181,A182,A188
3ikr	A	C	3.10.100	A204,A205,A206,A207,A208,A209,A210,A211,	C204,C206,C207,C208,C209,C210,C211,C212,C213,C2	A225,A228,A229,A230,A232,A233,A236,A237,A238,A239,A
5hbm	A	B	3.10.10-3.30.420-3.30.70	A7,A8,A9,A10,A11,A13,A14,A85,A86,A87,A88,	B17,B19,B20,B21,B22,B23,B24,B25,B26,B27,B28,B29,	A95,A96,A97,A99,A100,A101,A103,A179,A180,A181,A229,

Figure 14 – Snapshot of the table containing information about the various interfaces in the interface library. For a domain-domain interface the two chain IDs columns have the same chain ID, while for a chain-chain interface the two columns have different chain IDs.

Interface 1 and 2 refers to the residues belonging to the individual component of the interface. The binding site residues were not applicable for domain-domain interfaces and hence depicted with NA.

4. Discussions

An interface library of interacting protein chains and domains were created from the crystal structures deposited in the PDB. The CATH definition of domains was used as it has four times larger coverage of PDB as compared to SCOPe domain definition. There was a total of 112,043 (88,933 interfaces had an assigned CATH domain definition) chain-chain interfaces and 66,442 domain-domain interfaces interacting with each other. Out of the 1,391 folds, as defined by CATH, only 1109 folds had interacting partners (either with itself or other fold). 15 of these folds have interacting partners from 30 or more folds. Rossmann fold; for instance; interacts with 283 other folds. Besides, certain fold combinations such as Rossmann fold, Immunoglobulin fold, α helices fold has >1000 interfaces. These could be because of various non-homologous proteins folding into the same fold yet carrying out varied functions and interactions. Out of the 282 folds that do not interact with others ~38% of the folds belong to the orthogonal bundle (CATH ID – 1.10) or irregular architecture (CATH ID – 4.10). ~29% belong to a 2-layer sandwich (CATH ID – 3.10), alpha-beta complex (CATH ID – 3.90) and up-down bundle (CATH ID -1.20) architecture.

The protein data bank only has 3065 fold combinations from 1391 folds. A simple dimeric fold association between all folds would have a minimum of 966,745 fold associations, indicating either the fold combinations are either absent in the PDB or a large number of folds do not interact with each other.

The 155,375 interfaces (88,933 chain-chain interfaces+66,442 domain-domain interfaces) with an assigned domain definition clustered into 27,317 clusters based on structural similarity of the interface. The Rossmann fold, TIM barrel and Immunoglobulin fold again had >900 clusters predominantly because of the various ways they can interact with each other because of the diversity of sequences making up the fold. Approximately

73% of the TIM barrel interfaces were dissimilar (Structure overlap <70% and RMSD>2 Å) from each other as observed during clustering.

100 clusters had both chain-chain and domain-domain interfaces together, irrespective of the sequence similarity. This can indicate gene fusion leading to the formation of a domain-domain interface from a chain-chain interface or gene splitting leading to the formation of a chain-chain interface from the domain-domain interface. Because domain-domain interfaces and chain-chain interfaces are sometimes structurally similar, the library can provide an increased number of templates to model multi-domain protein whose individual domains have been crystallized separately.

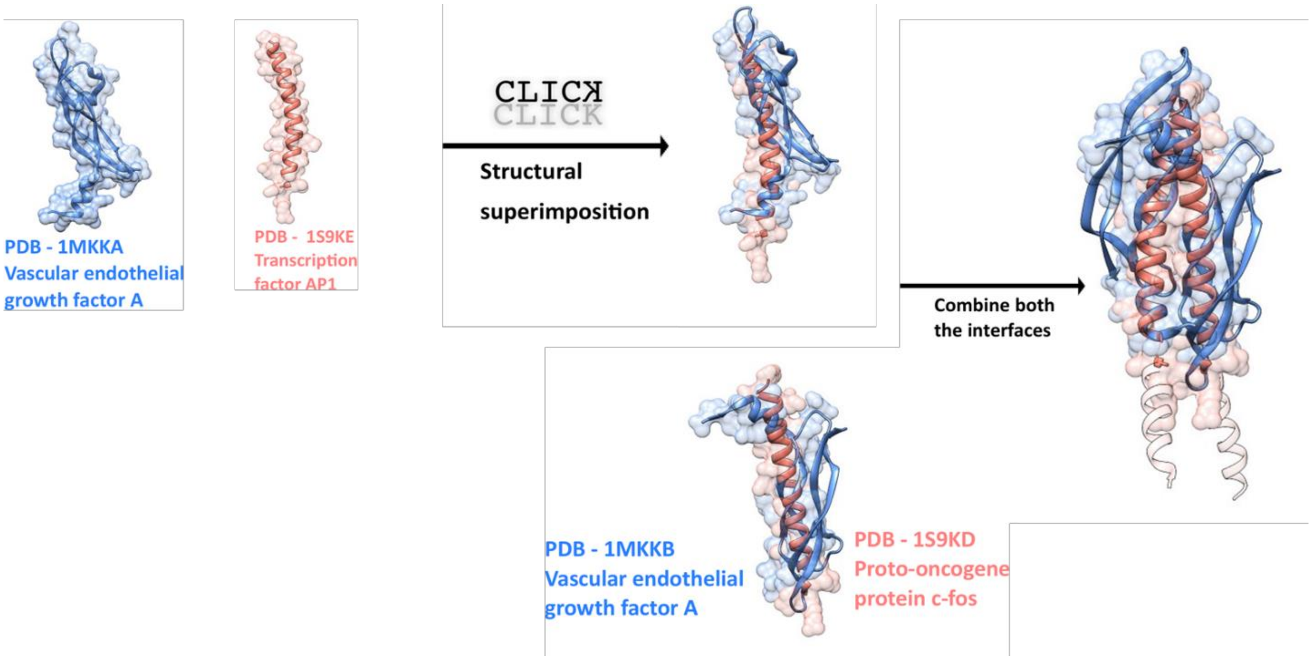

PDB - 1MKKA
Vascular endothelial growth factor A
PDB - 1S9KE
Transcription factor AP1
CLICK
Structural superimposition
Combine both the interfaces
PDB - 1MKKB
Vascular endothelial growth factor A
PDB - 1S9KD
Proto-oncogene protein c-fos

Figure 15 – Illustration of utilizing the interface library to model the protein complexes using a topology independent match as a template. The dimeric complex of vascular endothelial growth factor A (depicted in blue ribbons with surface representation) can be modeled using the interface between AP1 and c-fos (depicted in salmon ribbons and surface representation). A topology independent structural superimposition can be done utilizing CLICK. The two chains of vascular endothelial growth factor A interact in a structurally similar manner as c-fos with AP1.

Ideally, the clustering of the interface library should be done by an all against all structural superimposition, but it is computationally expensive and non-tractable with our present resources. In addition, we also did not utilize a non-redundant set for the library creation, as previous reports state that homologous proteins can interact with each other using structurally different interfaces as seen in lectins, bacterial chemotaxis proteins, ASPP proteins etc. However, structural comparisons showed how dissimilar folds can use the same geometry at the interface to interact with a certain protein fold. In addition, we noticed that irrespective of the topology of the interacting partners of a protein complex the interface can be structurally similar. Previous studies have pointed towards proteins utilizing similar geometry at the interfaces and the structural repertoire of the interface being close to complete [127]. Hence, the interface library can serve as a powerful tool to model protein complexes.

The interface library can be used to model complexes of proteins using a topology independent structural match (Figure 15) using the tool CLICK. For example, if we wanted to predict the dimeric complex of vascular endothelial growth factor (consisting of beta sheets), we would compare all the interfaces with the protein of interest (here vascular endothelial growth factor A). Say we find a match with transcription factor AP1 (which is topologically dissimilar as it contains alpha helices while the target protein containing beta sheets). We can then superimpose the other protein chain (here vascular endothelial growth factor A) onto the binding partner of AP1 (here proto-oncogene c-fos). A model of the complex (here dimer of vascular endothelial growth factor A) can be further built using the complex of the interacting proteins (here AP1 and c-fos) as a template. A topology independent structural match might increase the number of plausible templates for protein complex modeling as compared to that of a topology dependent structural match. Hence, using a topology independent match, we might be able to model more protein complexes.

We also computed the amino acid pair preference at a chain-chain and domain-domain interface to calculate a statistical potential. The amino acid pairs overall had a similar trend of being favored/unfavored at both the chain-chain/domain-domain interfaces. However self-amino acid pairs were generally favored at the chain-chain interface when compared to the domain-domain interface. This could be because ~62% of chain-chain interfaces are homo-oligomers whereas only ~28% of the domain-domain interfaces are homo-oligomers.

Approximately 67% of the protein-protein interfaces had at least a 30% overlap between the interface residues and predicted small molecule binding site residues. This resource can become helpful in docking/predicting small molecules that would bind at the interface region, hence inhibiting complex formation.

In conclusion, we have created a library of protein-protein and domain-domain interfaces, which in the future can be utilized to model complexes of proteins with each other and model multi-domain proteins whose individual domains have been crystallized separately. However, we chose to study coiled-coil interfaces to identify and model complexes of coiled-coil proteins, which have been described in the next chapter.

Chapter 4 - Prediction and modeling of coiled-coil protein-protein interfaces

1. **Coiled-coil interfaces in the database of protein-protein interfaces**
2. **Sequence based scoring scheme to detect coiled-coil interfaces**
3. **Prediction accuracy of the scoring scheme**

This study was in collaboration with Dr. Neelesh Soni, who helped in developing the scoring scheme.

1. Introduction

Coiled-coil is structural motifs formed by alpha helices winding around each other forming a supercoil [148]. Around 10% of all proteins contain coiled-coil motifs and these motifs help facilitate various functions such as molecular transport, structural rigidity, DNA binding, cell growth etc. [149–151]. In addition, they also facilitate protein-protein interactions [152]. Identification of accurate coiled-coil partners can help in protein-protein interaction predictions and design.

Coiled-coil motifs contain 2-7 helices (homo/hetero oligomer) arranged in parallel/anti-parallel orientation [153]. These motifs are made up of sequential repeats of amino acids that follow a canonical heptad or non-canonical hendecad repeat [154]. Heptad repeats are labeled from *a* to *g* in the protein sequence (Figure 1A). Hydrophobic residues are predominantly present in a and d positions that form the hydrophobic core of the coiled-coil whereas e and g positions are occupied by charged groups that helps in forming inter helical salt bridge interactions (Figure 1A). For a coiled-coil dimer, the positions *b*, *c* and *f* are also occupied by polar residues to interact with solvent molecules. These coiled-coil motifs are stabilized by hydrophobic contacts at heptad positions *a/d*, stabilized/destabilized by electrostatic attraction/repulsion at positions e and g and by interactions of water with polar groups at positions *b*, *c* and *f* [155]. The specificity, orientation and oligomerization are determined by the variants of these interactions.

Since coiled-coils follow a regular pattern, it makes it a perfect system to study structure-sequence relations [156,157], and hence many computational methods can be developed to predict coiled-coil properties from the sequence. Hence, we studied the properties of coiled-coil interfaces (a subset from the interface library explained in the previous chapter) to score and model coiled-coil proteins. Several sequence based prediction tools such as LOGICOIL [158], Multicoil2 [159], SCORER2 [160], RFCoil [161] etc. aim to predict the oligomerization state of the coiled-coils. Other sequence based coiled-coil domain prediction tools such as COILS [162], PCOILS [163], MARCOIL [164], MULTICOIL2 [159], CCHMM_PROF [165] has also been developed. Deepcoil is the latest technique based on neural networks that predict coiled-coil motifs based on a sequence or sequence profile [166]. A structure based coiled-coil prediction tool SOCKET [152] uses a knob into a hole packing to identify coiled coils [167,168]. The knob is formed by 1st and 4th residue that fits into a “hole” created by 4 resides on another helix. 3 out of 4 residues of the “hole” function as “knobs” themselves, resulting in an interlocking structure of coiled-coils.

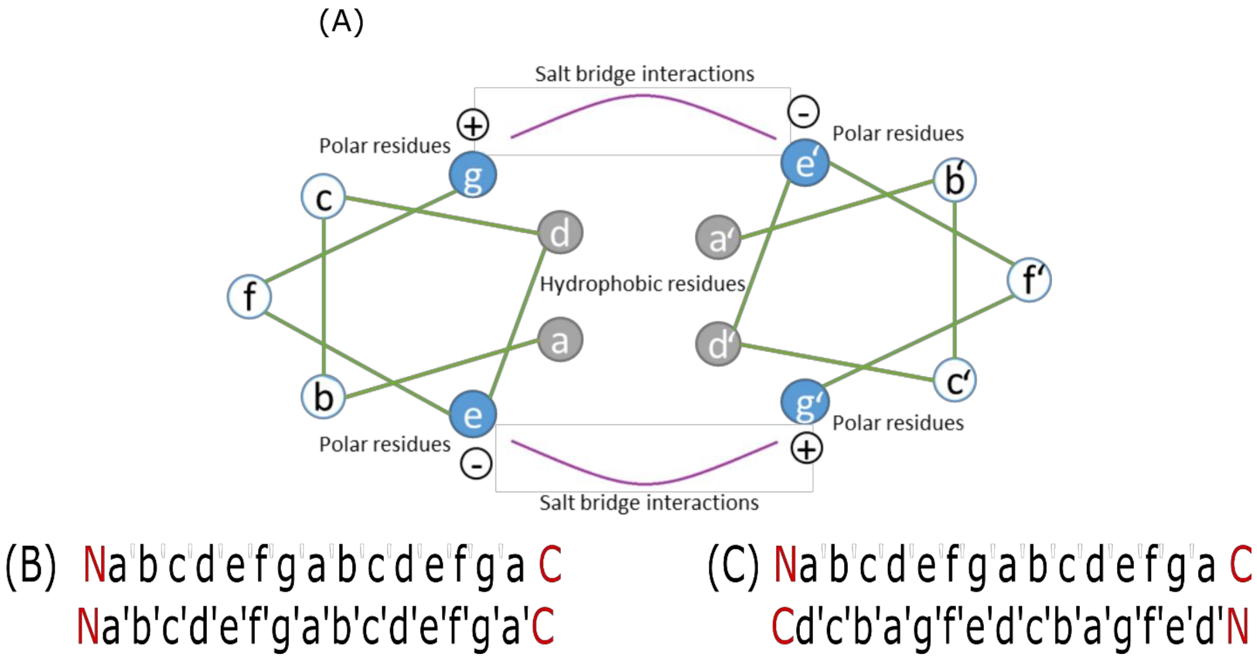


Figure 1 – (A) Schematic representation of heptad geometry in coiled-coils. The residues are labeled from a to g or a’ to g’; for the complimentary helix. The green straight lines indicate the covalent bond between the sequences. The purple curved line indicates salt bridge interactions. Heptad repeat alignment showing the amino acid pairing as seen in (B) parallel coiled-coil dimer (C) antiparallel coiled-coil dimers. The N and the C termini of the sequence have been mentioned in red by the letters N and C respectively.

Prediction of viable interactions between two helices that form a coiled-coil protein can help in understanding their function and to design coiled-coils. To the best of our knowledge, there are no known tools that predict if two protein sequences would interact in a coiled-coil geometry. In this chapter, we describe a sequence based scoring scheme that can be used to predict if two coiled-coil proteins would interact with one another in a particular orientation. The features involved in the scoring scheme are the amino acid pair preference at *a-a’, d-d’, a-g’* and *d-e’* positions (’ depicts amino heptad position from complementary strand) for parallel dimers and *a-d’, a-e’* and *d-g’* for antiparallel dimers.

These sequence derived features were used to train a random forest model to predict if two sequences would interact with one another in a particular orientation. The trained model was then used to score coiled-coil interfaces. The input for the random forest involves the two aligned sequences and their associated heptad pairing. The random forest scores the pairs of the amino acids in the input based on the amino acid pair preferences. The scoring scheme was also used to check the rank of a native coiled-coil interaction among non-native coiled-coil interactions. Besides, the scoring scheme was used to predict coiled-coil binding mode of JC virus agnoprotein with Rab11B and p53 proteins.

2. Methods

2.1. Datasets

2.1.1. Dataset for the training of random forest model for coiled-coil motif prediction

The coiled-coil dimers were identified from the PDB using the structure based coiled-coil predictor SOCKET [152]. The dataset for developing a random forest model (to predict if two sequences would interact in a coiled-coil geometry - described in Section2.3) consisted of the coiled-coil dimers from the PDB (as predicted by SOCKET) such that – (a) sequence redundancy was at 50% (b) heptad register from both the interacting coiled-coils are of equal length (c) the amino acid in neither of the strand was depicted with 'X' (X indicated the amino acid type was not determined) (d) the sequence length was at least 12 amino acids. The resultant dataset (as obtained from the PDB) had 2002 dimers of which 445 were parallel and 1557 were antiparallel. The resultant dataset was broken randomly into training and testing set such that the training set had 400 parallel and 1401 antiparallel dimers whereas the testing set had 45 parallel and 156 antiparallel dimers. 10 fold cross-validation was done using this dataset by varying the training and testing sets.

2.1.2. Coiled-coil interfaces

Coiled-coil geometry was predicted for the proteins that were already a part of the interface library (as mentioned in the previous chapter) using SOCKET. We restricted our studies to coiled-coil dimers. The total number of interfaces having a dimeric coiled-coil geometry was 2201 from 1251 proteins. There were a total of 1179 parallel coiled-coil interfaces and 1022 anti-parallel coiled-coil interfaces.

2.1.3. Creation of decoy set of coiled-coil interfaces

A random forest based scoring scheme was created to predict if two coiled-coils would interact with each other (described in Section 2.3). Decoy sets containing native coiled-coil dimer and non-native dimers were created. These decoy sets can help check if the score of the native binding partner of a coiled-coil interface helix is higher than non-native binding partners (that we presume would not bind). The decoy sets of coiled-coil interfaces (containing a native coiled-coil interface) was created such that – sequence identity <20%; the length of the heptad register is at least 12 residues long; amino acid sequence has no residue ‘X’ (‘X’ is usually present in residue positions of a PDB file where the identity of the residue is unknown); each of the decoy set should have native coiled-coil interface and at least 1 non-native coiled-coil interface. For a single helix of a coiled-coil interface, the non-native interactor(s) was/were chosen such that it - belongs to the set of observed coiled-coil helix belonging to a 20% redundant dataset with no amino acid labeled as ‘X’; does not have the same sequence as the native interactor; has the same heptad register as the native interactor. This creates a stringent decoy set as the non-native interactions were from coiled-coil proteins that have a heptad repeat, as compared to a decoy set created by random amino acid sequences.

2.2. Features for scoring scheme

The heptad positions *a/d* form the hydrophobic core of the coiled-coil and is important in keeping the helices together [148,155,169,170]. All dimeric helices from the CC+ database [171] (a database of coiled-coil proteins from PDB) were used to generate a pairwise distribution of amino acids at *a-a’* and *d-d’* for parallel coiled-coils and a-d’ for anti-parallel coiled-coils. In addition, the interactions between *a-g’* and *d-e’* for parallel coiled-coils and *a-e’* and *d-g’* have also been shown to help maintain the stability of the coiled-coils [148]. Hence these amino acid pair frequencies were calculated and used as a feature set to train a random forest model for the prediction and scoring of coiled-coil proteins.

2.3 Random forest based scoring of the training set

During the training (dataset mentioned in section 2.1.1.), the input contains the dimer sequences that interact with one another, its heptad register and the orientation of interaction (parallel or anti-parallel). Depending on the pairing of the amino acids and the orientation, the model is trained based on the features described in Section 2.2 (amino acid pair preferences). The random forest module of the scikit-learn package [172] of python was used to train the model (mentioned in Section 2.1.1) to predict if the interaction and orientation are feasible. Each of the random forest models contains 100 decision trees, where each tree node partitions the data using the entropy criterion. The parallel dimers were weighted 3.5 times higher than the antiparallel dimers to compensate for the imbalanced dataset (anti-parallel dimers were ~3.5 times more than parallel dimers). The model outputs a probability of how feasible the interaction is in the specific orientation. The output probability is dependent on the outputs predicted from each of the decision trees. The trained model was used to predict on a testing set (mentioned in section 2.1.1.) if the two amino acid sequence dimers (along with the heptad register) would interact or not in the mentioned orientation. A 10 fold cross-validation was done for the same and a ROC curve plotted.

2.4. Test case for prediction of coiled-coil interfaces

Human polyomavirus 2 (JC virus) is a type of human polyomavirus, that contains a protein called agnoprotein, which plays a role in the viral replication cycle. The agnoprotein interacts with a large number of proteins that contain a coiled-coil motif [173]. For the study, we restricted ourselves to two of the interacting partners of angiogenin – p53 and Rab11b. We predicted the coiled-coil domains of the agnoprotein and p53 and Rab11b using Deepcoil [166]. We then predicted if it was feasible for the 2 proteins to form coiled-coil interactions based on our scoring scheme. All possible modes of parallel and antiparallel interactions between the partners were created (by sliding one protein over another in both directions to capture both parallel and antiparallel binding modes) and scored. A model of Rab11b interacting with angiogenin was modeled using MODELLER.

3. Results

3.1. Features for the scoring scheme

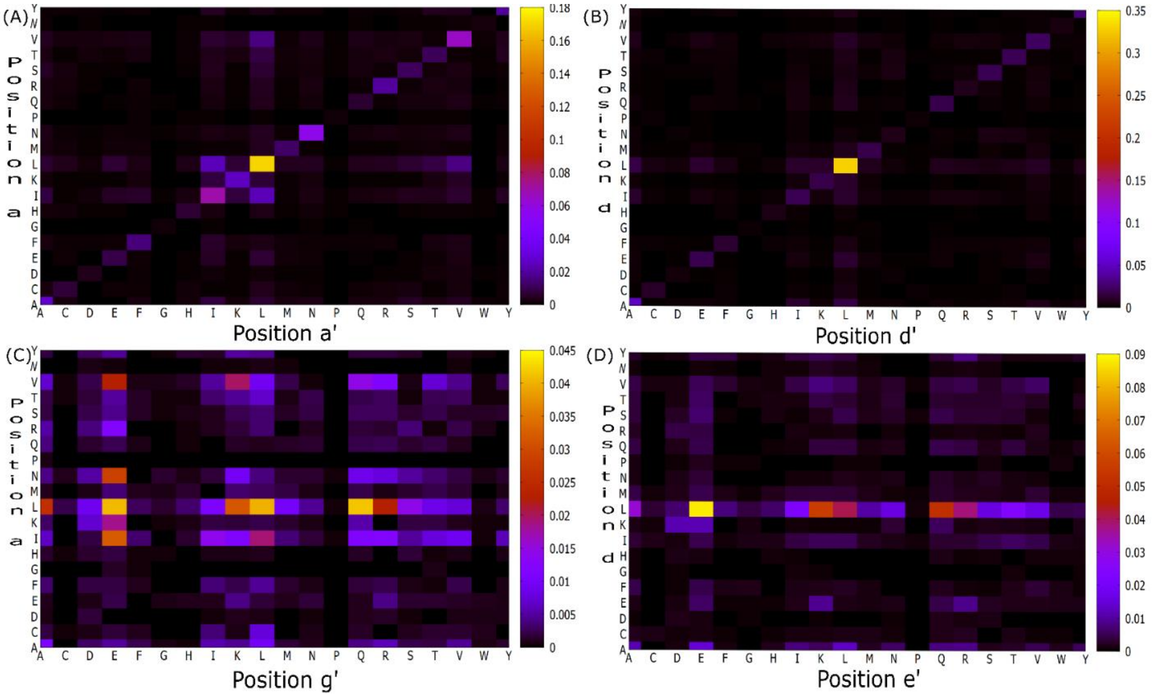


Figure 2 – Amino acid pair preference for parallel coiled-coil for (A) a-a’ (B) d-d’ (C) a-g’ (D) d-e’ positions. All the amino acid pair preference sum to 1. The color code is based on the calculated probability of the amino acid pairing. The reddish-yellow shades show higher probability compared to bluish-black shade.

The amino acid pair preferences were calculated for *a-a’, d-d’, a-g’* and *d-e’* positions for parallel coiled-coils. Leu-Leu pair was most preferred for a-a’ and d-d’ pairing. In addition, self-amino acid pairing involving hydrophobic amino acids or polar amino acids with a long hydrophobic chain such as Ile, Val, Asn, Ala, Lys Arg, Tyr, Phe pairing were favored at *a-a’* position. Also, the hydrophobic pairing between Leu-Ile and Val-Leu was favored at *a-a’* position. Similarly, a self-amino acid pairing of Ala, Tyr, Val, Thr, Ser, Ile, Glu, Gln, Met, Lys are favored at *d-d’* positions. These amino acids (expect Ser and Thr) are either hydrophobic or have a long hydrophobic chain. The general trend of preferred residue pairs at *a-a’* and *d-d’* seem to be similar. For the *a-g* pairing the predominant pairs involve hydrophobic amino acids such as Leu, Ile, Val and Asn at heptad position *a*, with residues Leu, Lys, Gln, Arg, Glu, Asp, Ile at *g’* position, which are either hydrophobic or have a long hydrophobic chain. For the d-e’ pairing heptad position *d* is occupied by Leu, Ile and Val (all three of which are hydrophobic) whereas position *e’* contain residues Lys, Leu, Gln, Arg, Glu (either hydrophobic or have a long hydrophobic chain). These amino acid preferences were used for scoring parallel orientations of coiled-coil helices.

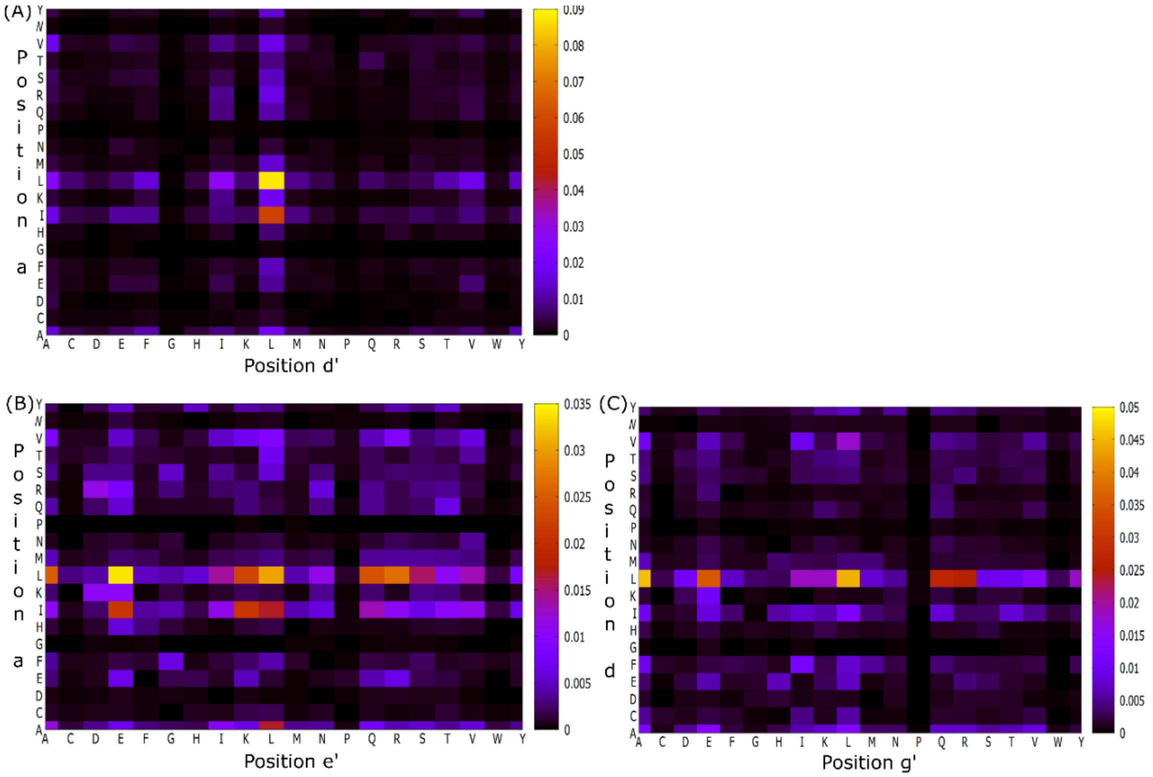


Figure 3 – Amino acid pair preference for antiparallel coiled-coil at (A) a-d’ (B) a-e’ (C) d-g’ positions. All the amino acid pair preference sum to 1. The color code is based on the calculated probability of the amino acid pairing. The reddish-yellow shades show higher probability compared to bluish-black shade.

The parallel and antiparallel orientations of coiled-coils have different amino acid pairing as shown in Figure 1B (for parallel coiled-coil dimers) and Figure 1C (for antiparallel coiled-coil dimers). Hence, for antiparallel coiled coils the *a-d', a-e'* and *d-g'* pair preferences were calculated as they form the interacting pairs. The *d-g', a-e'* (for antiparallel) and *d-e', a-g'* (for parallel) interactions are between the hydrophobic residues at *a* and *d* heptad positions with amino acids having long hydrophobic side chains at *e* and *g* heptad positions of the complementary chain. The *a-d'* pairing predominantly involved pairing between hydrophobic amino acids such as Leu-Leu, Ile-Leu, Leu-Ala, Lys-Leu, Val-Ala, Leu-Val, Ile-Ala, Val-Leu amino acid pairs. The *a-e'* pairing contains Leu and Ile at heptad position *a*, with residues Glu, Ile, Lys, Leu, Gln, Arg, Ser, Val, Ala at *e'* positions. The same hold true for *d-g'* pairing with d position containing hydrophobic amino acids and *g* position containing hydrophilic amino acids with a long chain. The *d-g'* pairing involves Leu, Ile at *d* position with residues Glu, Ala, Ile, Lys, Leu, Gln, Arg at *g'* positions.

3.2. ROC characteristics of the random forest model during 10 fold cross-validation

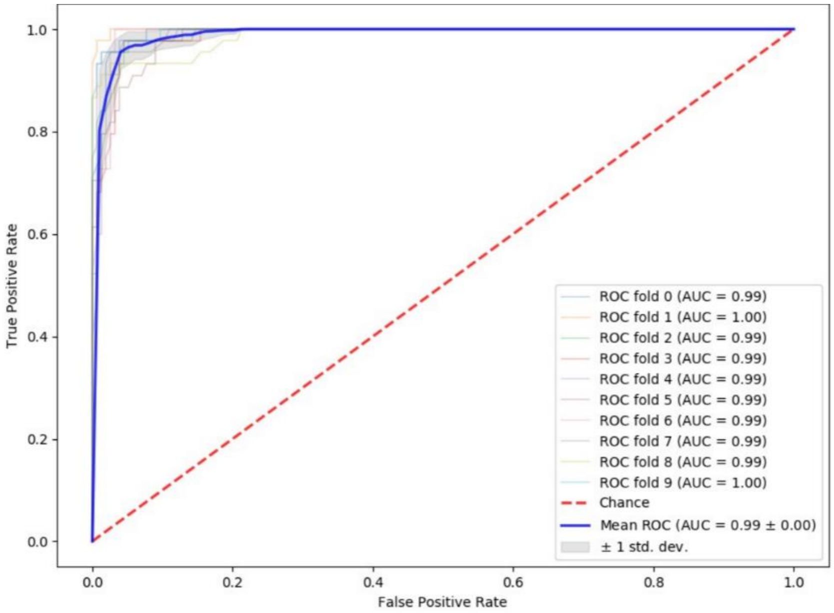


Figure 4 – ROC curve during 10 fold cross-validation. The blue curve is the mean value of ROC and the red dashed line indicate the ROC for a random chance event. The ROC

during the individual runs have been indicated by different colors as denoted by the legend in the graph.

The random forest model to predict if a coiled-coil interaction is feasible or not in a particular orientation was used to generate Receiver Operating Characteristic (ROC) curves [174] during a 10 fold cross-validation (mentioned in Section 2.1.1. and 2.3). The coiled-coil dimeric sequences were converted to the pair preferences for the heptad positions (as mentioned in Section3.1) and the random forest scoring scheme was used to make the prediction. The average area under the curve (AUC) during the cross-validation was 0.99+/-0.00 (Figure 4). The MCC across the 10 models is 0.9 and the prediction accuracy is 96%.

3.3. Prediction accuracy for coiled-coil interface decoy set

215 decoy sets of coiled-coil interfaces were created each containing between 1 to 49 decoys each (Figure 5). The 215 decoy sets contain decoys for 138 parallel dimers and 77 antiparallel dimers. The number of residues in the decoy set varied between 12 to 61 (Figure 6).

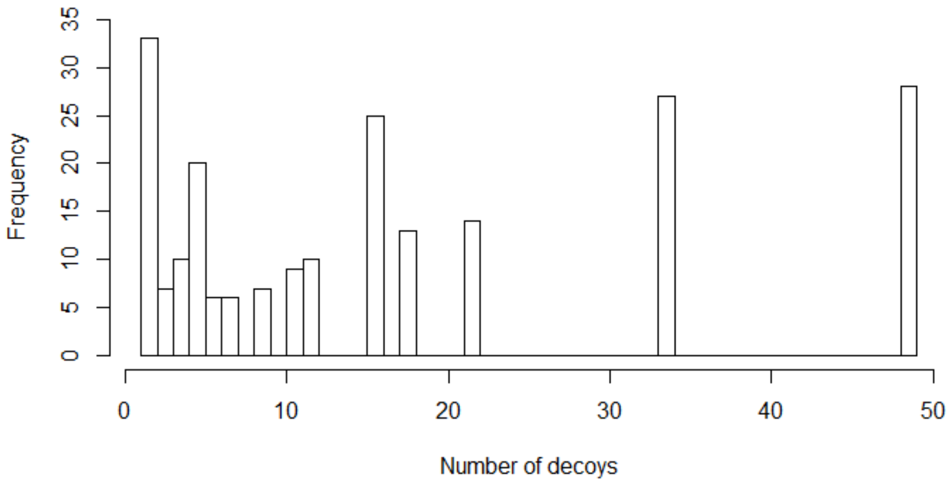


Figure 5 – Histogram showing the number of datasets (y-axis) each containing a particular number of decoys (x-axis).

The trained random forest model was used to predict the native interactor of coiled-coil interface helix from the non-native interactors (using the decoy set as explained in 2.1.3). The input for the predictions were 2 sequences of amino acid (to be predicted as interacting or not) along with the register of heptad repeat and the orientation. The amino acid sequence and its register are used by the random forest model to calculate probabilities of interaction, in a specific orientation, to be feasible. For each of the decoy set (containing multiple interactions – native/non-native), the probability score for each of the interaction is converted into a percentile score. The percentile scores for the native dimers is higher than that of the non-native dimers in most cases (Figure 7). Only 9% of non-native parallel dimers and 10% of non-native antiparallel dimers had a percentile rank of >=90. 109 out of 138 parallel dimers (78%) were predicted in the top 90 percentile ranks. Out of 77 antiparallel dimers, 56 (74%) were predicted in the top 90 percentile ranks.

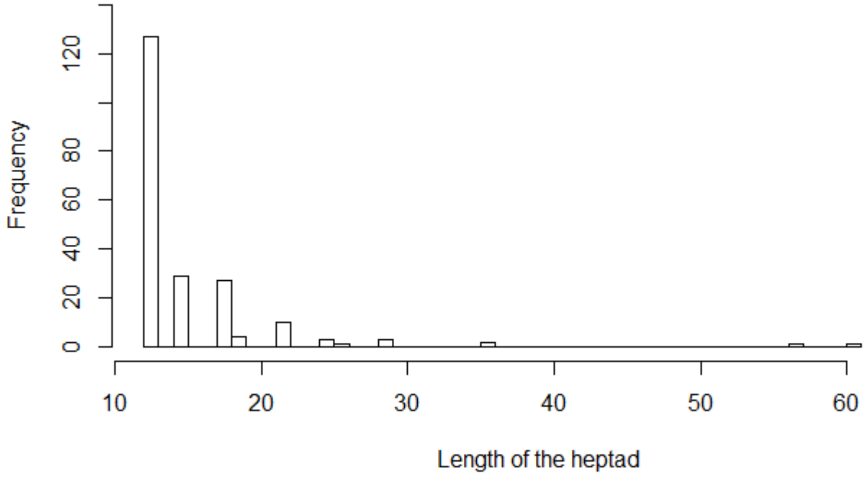


Figure 6 – Histogram showing the number of sequences (y axis) having a particular length of heptad repeats (number of residues) (x-axis) for the coiled-coil interface decoy set. Most interfaces is of length 12.

The coiled-coil interface decoy sets were created from proteins in the PDB. The interface can hence entirely be coiled-coil, or the coiled-coil might just be a part of a larger protein-protein interface. For the parallel decoy sets, that we failed to predict in the top 90 percentile the mean percentage of the total interface that was coiled-coil is 38%.

Whereas, the ones that were predicted in the top 90 percentile had a mean percentage of 57% (Figure 8B). However, the antiparallel decoy sets both the ones that we predicted in the top 90 percentile and those which we failed to predict only had an average of ~35% of the coiled-coil region forming the interface (Figure 8B).

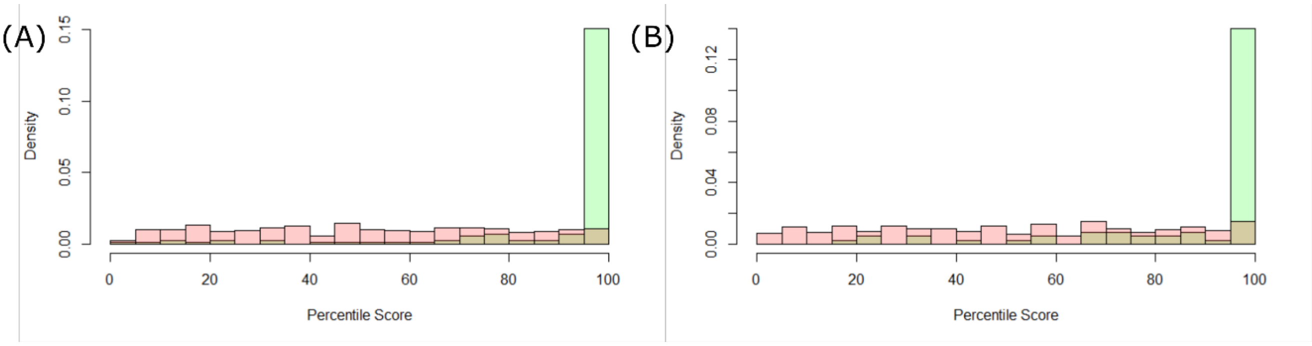


Figure 7 – Percentile score of (A) parallel (B) antiparallel coiled-coils, where green represents the scores for native dimers and red represents the score for non-native dimers.

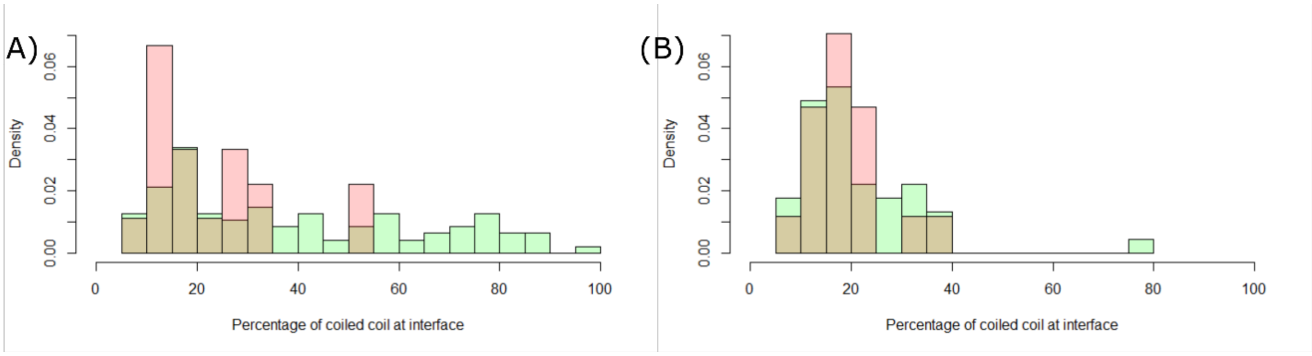


Figure 8 – Percentage of the interface that follows the coiled-coil geometry for (A) parallel (B) anti-parallel coiled-coils. The green bars represent the coiled-coils that were predicted in the top 90% percentile, while the remaining not predicted are represented in red bars.

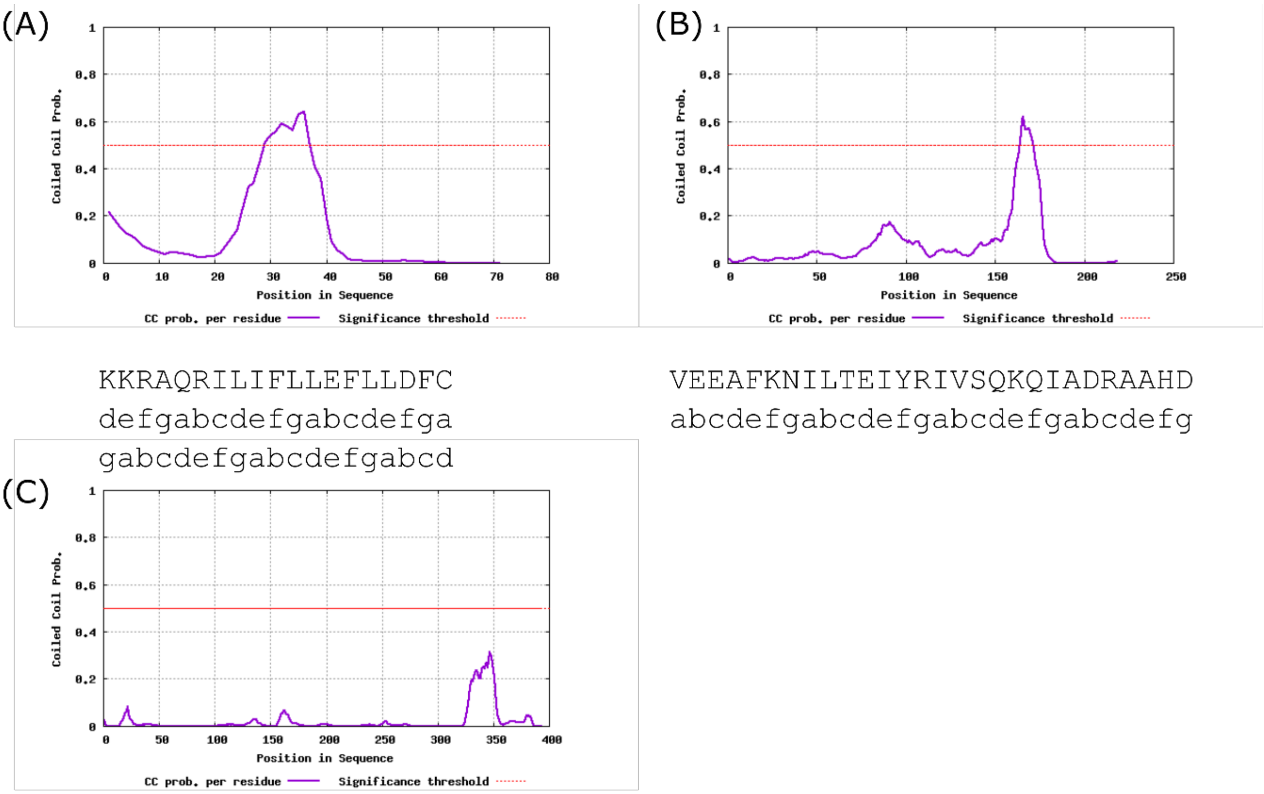
3.4. Test case for prediction of coiled-coil interface

Along with scoring coiled-coil motifs in proteins and protein-protein interfaces, we also predicted coiled-coil interactions between proteins that are known to interact with each other. For this study, we predicted the interactions of JC virus agnoprotein with its interacting partners – p53 and Rab11B. JC virus infects humans, affecting the central nervous system causing progressive multifocal leukoencephalopathy [175] and granule cell neuropathy [176]. The virus encodes proteins namely large T antigen, small T antigen, T', agnoprotein, VP1, VP2 and VP3 [176,177]. Among these proteins, the 71 residue long agnoprotein plays an important role in the viral life cycle [178], such as transcription, replication [179], virion formation [180], functioning as viroprotein [181] and deregulation of cell cycle progression [182]. The agnoprotein contains 2 helices namely minor helix (residue no. 7-12) and major helix (residue no. 22-39) with a loop connecting the two. Most of the molecular targets of agnoprotein, as revealed by proteomics data [173], showed coiled-coil motifs. The agnoprotein itself undergoes dimerization using its major alpha-helix [183]. The prediction of the coiled-coil region using Deepcoil [166] on agnoprotein predicts the region of the helix 29-37 as a coiled-coil (according to their probability threshold of 0.5). However, the regions 22-39 had a higher coiled-coil probability (>0.05) than the rest of the protein (Figure 9A), and hence for further predictions the entire stretch of residue no 22-39 was assigned with heptad repeats. We used our scoring scheme to predict viable interactions between 2 monomers of agnoprotein following the coiled-coil geometry.

Two proteins Rab11B and p53 were predicted to bind to agnoprotein. We used Deepcoil to predict the coiled-coil regions of both these proteins. For Rab11B the residues 161-188 forms a helix, with the residues 164-171 being predicted as coiled-coil (with probability >0.5). However, the entire stretch (residue no. 161-188) has a coiled-coil forming propensity (Figure 9B). For p53 the region between the residues 335-355 exists in the form of a helix with a propensity to form coiled-coil (0.02-0.3) (Figure 9C). In addition, it was shown that residue number 1-36 of agnoprotein is important for interactions with p53 [182], of which residue number 22-36 have propensities of coiled-coil formation.

We then predicted the feasibility of the interactions between Agno with Rab11B and p53 using a coiled-coil motif. The heptad repeats for all 3 proteins were manually assigned. Alternate heptad assignments were also done for the sequences of agnoprotein and p53 (Figure 9).

All possible parallel and antiparallel orientations of Agno-Agno, Agno-Rab11b, Agno-p53 were created and scored by the random forest model. Our scoring scheme predicted that Agno homodimer could interact in an antiparallel orientation (Figure 10A) (Score = 0.975). The Agno-Rab11b oligomer was predicted to bind in an antiparallel conformation (Score = 0.97). 2 alternate antiparallel coiled-coil modes of binding were predicted with the same score (Figure 10B). The interactions of p53 with Agno was also predicted by an antiparallel coiled-coil binding mode with a score of 0.99 (Figure 9C).

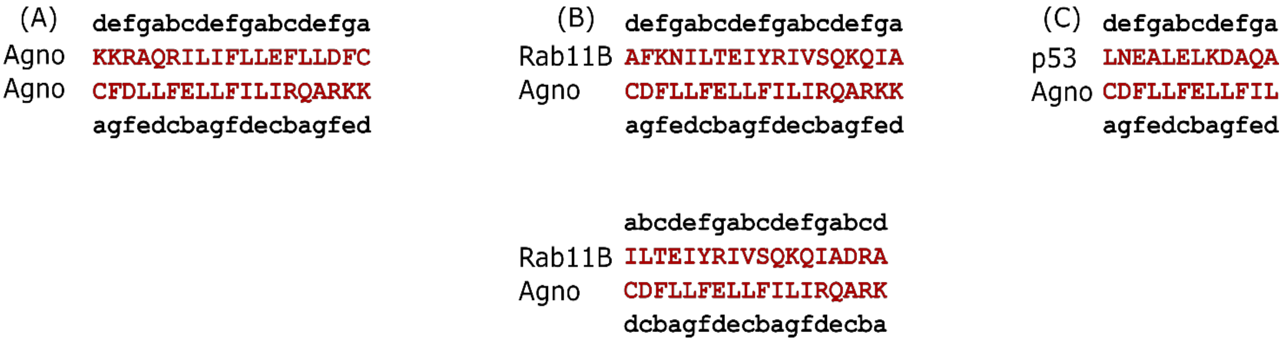


KKRAQRILIFLLEFLLDFC
defgabcdefgabcdefga
gabcdefgabcdefgabcd

VEEAFKNILTEIYRIVSQKQIADRAAHD
abcdefgabcdefgabcdefgabcdefg

RERFEMFRELNEALELKDAQA
bcdefgabcdefgabcdefga
fgabcdefgabcdefgabcde

Figure 9 – The prediction of the coiled-coil propensity by Deepcoil for (A) agnoprotein (B) Rab11B (C) p53. The red line indicates the threshold of 0.5 above which Deepcoil predicts the region as coiled-coil. The sequence that is used as coiled-coil and the assigned heptad repeat is mentioned below each plot.



```
(A)   defgabcdefgabcdefga        (B)   defgabcdefgabcdefga        (C)  defgabcdefga
Agno  KKRAQRILIFLLEFLLDFC     Rab11B  AFKNILTEIYRIVSQKQIA        p53  LNEALELKDAQA
Agno  CFDLLFELLFILIRQARKK     Agno    CDFLLFELLFILIRQARKK        Agno CDFLLFELLFIL
      agfedcbagfdecbagfed             agfedcbagfdecbagfed             agfedcbagfed

                                      abcdefgabcdefgabcd
                              Rab11B  ILTEIYRIVSQKQIADRA
                              Agno    CDFLLFELLFILIRQARK
                                      dcbagfdecbagfdecba
```

Figure 10 – Heptad pairing for the best predicted model of (A) agnoprotein homodimer (B) agnoprotein-Rab11B (C) agnoprotein-p53. The amino acid sequence is shown in red and the assigned heptad register in black.

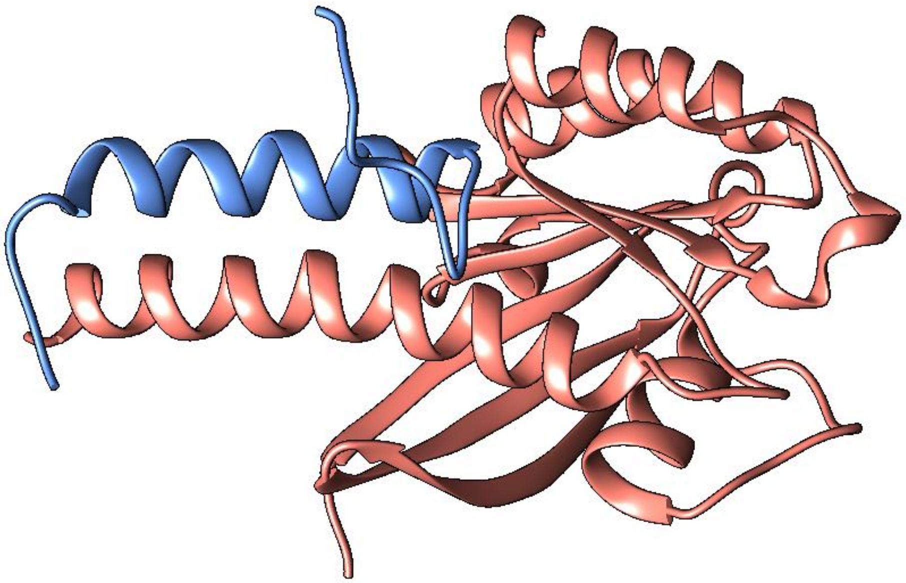


Figure 11 – Model of Agno (blue ribbon) with Rab11B (salmon ribbon) based on the prediction of their interacting interface.

The model of the agnoprotein-Rab11B complex (Figure 11) was modeled using MODELLER using Aicar Transfomylase-IMP Cyclohydrase (PDB - 1G8M_AB) as a template. The template was chosen such that the heptad register of both the interacting domains was the same as that predicted interface. The sequence alignment was manually done to ensure appropriate matching of the heptad register between the target (agnoprotein-Rab11B) and template (PDB – 1GMP_AB). Of the two interacting modes, the model was built using the interaction between the residues ILTEYIRIVSQKQIADRA (from Rab11B) and CDFLLFELLFILIRQARK (from agnoprotein).

4. Discussions

Coiled-coil motifs are found in about 10% of all proteins and these play important roles in cell growth, DNA binding etc. These motifs have also been seen in protein-protein interfaces. Prediction of the viability of such coiled-coil geometry between sequences that can take up coiled-coil architecture can help identify interacting partners of proteins. To the best of our knowledge, there are no tools that predict if two protein sequences having a heptad repeat would interact with one another in a particular orientation. Here, we build a random forest based model to predict if two sequences having a heptad repeat would interact with each other to form coiled-coils in a specific orientation (parallel/anti-parallel).

The input for the random forest model is two sequences with their assigned heptad repeats along with the orientation. The sequences and heptad information are converted into amino acid pair frequency scores for the heptad pairings *a-a', d-d', a-g', d-e'* (for parallel dimers) and *a-d', a-e', d-g'* (for antiparallel dimers). The output of the random forest model is the probability of the sequence to form a parallel/antiparallel dimer. The AUC for the predictions during 10-fold cross-validation was 0.99+/-0, with an MCC of 0.9 and an accuracy of 96%.

We then used the trained random forest model to predict if two sequences at a protein interface following a heptad architecture would interact with each other or not. For each of the 215 native interfaces, we created non-native decoys that most probably would not interact with one another. Each of the 215 decoy sets was scored using the above mentioned model and the percentile score for each dimer sequence in the decoy set was determined. The model predicted 78% and 74% of parallel and antiparallel coiled coils respectively in the top 90 percentile. We might have predicted more parallel cases as compared to antiparallel ones because in antiparallel dimers the dipole moment of the helix is in the opposite direction. Hence, providing additional stability to the interface, in addition to that of the amino acid pairwise interactions. The parallel coiled coils that were not predicted in the top 90 percentile had a mean of 38% of the interface as coiled-coil, whereas the ones predicted in the top 90 percentile had a mean of 57%. Hence, indicating that there might be interactions (that might come from the non-coiled-coil part of the interface) other than the coiled-coil interactions for the protein-protein interfaces, which failed to be predicted in the top 90 percentile. For the antiparallel coiled-coil interface there was no difference (both had ~35% of the interface as coiled-coils) between the cases that were predicted in the top 90 percentile verses that were not predicted. Overall, the parallel coiled-coils formed had a mean of 54% of the interface as coiled coils, whereas the antiparallel coiled-coils had a mean of 35% of the interface as coiled coils. This indicates that the antiparallel coiled-coil interfaces mostly had additional interactions (belonging to non-coiled-coil regions of the interface) stabilizing the interface that was not considered while making the prediction.

Along with utilizing the model to score and rank coiled-coils, we also used it to predict homo and hetero-oligomeric coiled-coil complexes of agnoprotein with Rab11B and p53. Agnoprotein undergoes homodimerization and is seen to interact with proteins containing a coiled-coil motif. The plausible coiled-coil regions for agnoprotein, Rab11B and p53 were predicted using Deepcoil and the heptad repeats were manually assigned. All possible binding modes of agnoprotein with itself and other proteins were scored. The model predicted all 3 complexes to form stable antiparallel coiled-coils. The model of agnoprotein with Rab11B was further modeled using MODELLER.

In the future, sequence based coiled-coil motif predictor such as Deepcoil [166] can be used to annotate coiled-coil regions in all proteins. This can be followed by the assignment of heptad repeats to the ones on the surface of the protein (surface residues will be involved in interactions). Following this, all possible parallel or antiparallel modes of interactions between these proteins can be scored by the random forest based scoring schemes. We might over-predict the actual number of feasible interactions (~9% of the non-native interactions were given scores higher than native interactors in the decoy set). However, building a model of the complex, using a coiled-coil template would get rid of the unfeasible interactions because of steric clashes.

To conclude, we have developed a random forest based tool that predicts if two sequences with a heptad repeat will interact using a coiled-coil motif in a particular orientation (parallel or antiparallel). This resource can be developed into a web server that can be of broader help to the scientific community to predict and model complexes of proteins that interact with each other using a coiled-coil motif.

In this chapter, we scored and predicted coiled-coil protein complexes. In the upcoming chapter, we shall describe how we can use residues environments at protein-protein interfaces to score protein complexes.

Chapter 5 - Study of residue environments at protein-protein interfaces to score protein complexes

1. **Creation of depth dependent scoring potentials for protein-protein interface**
2. **Validation of the scoring scheme on different decoy sets**
3. **Comparison of the technique to other technique - PIZSA**

This study was done by a collaboration with Prof. Maya Topf, Birkbeck. One of the decoy sets used in this study was developed in collaboration with Dr. Bullock from Prof. Topf's lab.

1. Introduction

Various computational methods have been developed to predict structures of protein complexes [86,95–97]. These methods face two challenges – sampling and scoring. Sampling involves the construction of the plausible models for these protein complexes [95] (discussed in Chapter 3), which is followed by scoring techniques that identify the near-native complexes from all the sampled conformations.

The scoring schemes are of three types - physics based, knowledge based and machine learning based [184]. Physics based energy functions involve a weighted combination of various energy terms like van der walls interactions, hydrogen bonds, Lennard Jones potential etc. Various docking schemes such as HADDOCK [49], pyDock [50], SwarmDock [51], ZDock [52], RossettaDock [53] etc. utilize these physics based energy calculations to score and identify protein complexes. The knowledge based potentials extract information from known structures such as residue-residue/atom-atom contacts to score protein complexes [185,186]. Tools such as PIZSA [35], CIPS[187] etc. utilize these potentials to score complexes. Certain scoring schemes combine both physics based and knowledge based potentials [188,189]. The machine learning based schemes involve a non-linear combination of various energy terms, physicochemical and geometric features to score protein complexes [190–192]. iScore is one such scheme that utilizes a machine learning based scoring scheme to identify protein complexes [193].

Most knowledge based scoring schemes involve a ratio of the observed number of interactions to that of the expected number of interactions. A higher value of this ratio indicates a favored interaction in nature over what is expected from a random sample [40]. These knowledge based potentials (PIZSA, CIPS etc.) usually utilize the contact propensities of the residues with each other for scoring protein complexes, to identify near-native protein complexes [35,185–187]. Along with the utilization of observed/expected ratios, currently developed potentials employ various other features such as Lennar-Jones potentials, solvation effects, number of atoms in contact, conservation etc. [35,194,195], to score protein complexes [196–198].

The blind prediction experiment, Critical Assessment of Prediction of Interactions (CAPRI) [199–202] for identification of interactions between proteins provides the competitors (tools that score protein complexes) with the structures of protein complexes (native, near-native and non-native) to evaluate and rank them. Despite the progress made in scoring protein complexes, CAPRI experiments have shown there is a need for the development of better schemes for the scoring of the protein complexes [203].

In Chapter 2 Section 1, we have explained how residue depth (the distance of the residue from the nearest bulk solvent) can be used to characterize microenvironments in proteins. Depth has also been used for the prediction of binding sites in proteins, cavity, pKa, temperature sensitive and deleterious mutations etc. [12,142]. In this chapter, we utilized depth to characterize the residue microenvironments at protein-protein interfaces. The residues present at the interface between proteins get buried upon complex formation i.e. there is a change in depth from their monomeric form to their oligomeric form. Burial of polar amino acids is destabilizing as compared to non-polar amino acids and the extent of instability depends on the amino acid type [204,205]. Hence, an interface should be enriched in residues that can undergo burial. The probability of a residue to transition from one depth in a monomer to another in a complex (as seen in the PDB), was used to develop a scoring scheme, called MODP to distinguish near-native complexes from non-native complexes. The utility of the scoring scheme was accessed on 3 different datasets (Refer to Section 2.5) to rank native structure and identify near-native structures.

2. Methods

2.1. Dataset for scoring scheme

The dataset for the creation of the depth based potentials involved structures from the PDB such that (a) sequences were non-redundant at 30% sequence identity, (b) resolution was <3 Å (c) R-factor < 0.3 (d) had more than 1 protein chains. These resulted in 8,498 structures.

2.2. Identification of interface residues

Any residue having at least 1 atom within 4 Å [37] from another atom belonging to a different chain is identified as an interface residue. A total of 1,395,179 residues were identified as interface residues from 8,498 PDBs.

2.3. Calculation of residue depth

Residue depth was calculated with the DEPTH program using default parameters (2 water molecules for bulk solvent definition and 25 iterations). The depth of the proteins in both their complex and monomeric forms were calculated. The oligomeric depth of the residues was computed using the protein complex. To calculate the depth of the residues in an individual monomer, the protein complex was separated into their chains, and their depths calculated. This was done under the assumption, that there was no induced-fit during the complex formation (the conformational changes during the complex formation was within the standard deviation of depth).

2.4. Computation of knowledge based statistical potentials

From the protein complexes, we computed the probability ($\mathrm{P}_{\mathrm{R}}^{\mathrm{Di}\rightarrow\mathrm{Df}}$) where Di is initial depth, Df is the depth in the complex and R is one of the 20 naturally occurring amino acids. This probability was calculated based on the dataset described in Section 2.1. The probability value was computed the number of residues. The probability value was computed using the number of residues of type R at a depth $\mathrm{D_i}$ in a monomeric form ($N^{D_i}{}_{R}$) and calculating how many of these residues (residues at depth $\mathrm{D_i}$ in a monomer) had moved to a depth D_f in a protein complex ($N^{D_i \rightarrow D_f}{}_{R}$) (equation below) (Figure 1). The residue depths were binned at 0.5 Å bins. For each of the 20 amino acids the values of D_i range between 2.5 to 10 Å, while that of D_f range between 2.5 to 20 Å. For each of the amino acids, the probabilities for all combinations of D_i and D_f was calculated using the formulae shown below.

$$P_{R}^{Di \to Df} = \frac{N_{R}^{Di \to Df}}{N_{R}^{Di}}$$

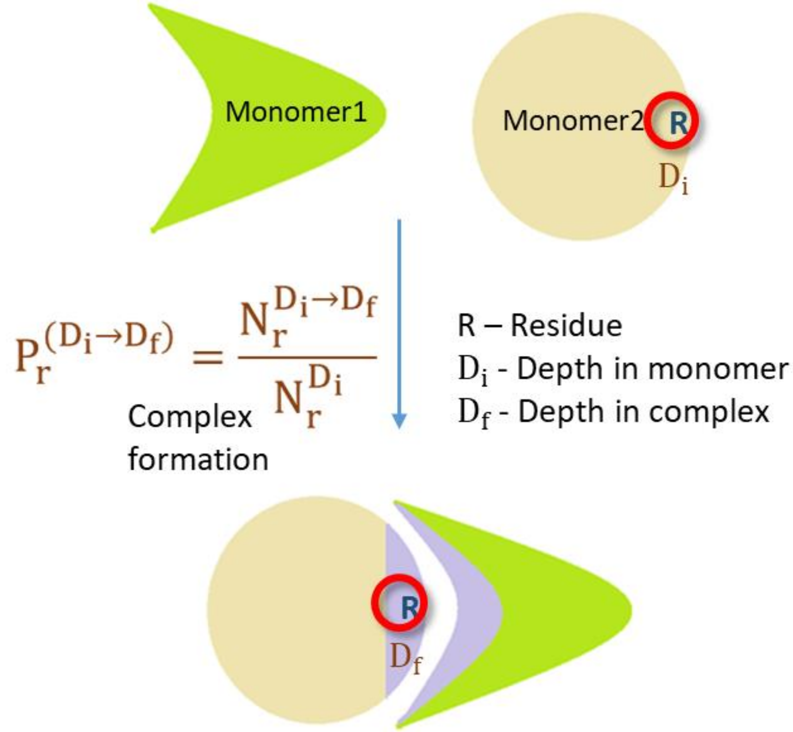


Figure 1 – Schematic showing 2 monomers (monomer 1 in green and 2 in yellow), with monomer 2 having a residue R (shown in blue, enclosed with a red circle) at D_i, which undergoes a depth change to D_f upon complex formation. The interface regions are shown in purple.

2.5. Monomer Oligomer Depth Potential (MODP)

The protein complexes were scored to identify near-native complexes from non-native complexes. The scoring scheme is called Monomer Oligomer Depth Potential (MODP). We computed the interface residues and calculated the residue depths in a monomer and the protein complex. We then retrieved the scores of the probability of the residues at the interface to move from a monomeric depth to complex depth (as computed in Section 2.4). The interface was scored as the summation of all the probability scores of all the interface residues.

$$Score = \sum_{i}^{n} P_{R}^{Di \rightarrow Df}$$

The z-score of the interface was then computed by comparing the score of the interface to a random background of 1000 scrambled interfaces. The z-score is computed using the mean (μ) and standard deviation (σ) of the 1000 scrambled interface scores. These scramble interfaces during each score computation were created by randomly scrambling the interface residues, keeping the amino acid composition for a protein interface conserved. These random decoys hence generated retained the monomeric and complex depth for the position but had a different (or same) amino acid. Sequence randomization has been shown to compare well with a physical model involving structure sampling as seen for fold assessment [206]. Sequence randomization has been previously used to calculate z-scores for protein interfaces [134].

$$Z - score = \frac{x - \mu}{\sigma}$$

2.6. Datasets of decoys of protein-protein complexes

These depth based scoring potentials were trained and tested on different decoy sets containing near-native and non-native models of protein complexes. The definition of near-native and non-native was used as defined by the dataset. Throughout the chapter, near-native complex would also include native conformation unless otherwise mentioned.

Bullock's dataset - The dataset mentioned in Bullock *et al.* [207] will be called Bullock's dataset. It contained 76 protein decoy sets each containing 101 decoys (near-native/non-native) each. The decoy set contained 75 dimeric and 1 oligomeric protein complex. The decoy set was broken into 2 parts – a training and testing set (60 proteins) and a validation set (16 proteins) for computation of z-score cut off to distinguish between near-native and non-native structures. The training and testing set had 48 proteins for training and 12 proteins for testing. 5 fold cross-validation was done on the training and testing set to optimize the z-score cut off for prediction. The decoys were defined as near-native or non-native based on mean RMSD and fnat. For the mean RMSD calculation, each chain of the decoy is iteratively superimposed on the native structure chain and the C^{α} RMSD is computed. The mean RMSD is the average of all the RMSDs calculated. The mean RMSD values that were provided in the dataset were used. Fnat refers to the fraction of the native contacts. It is computed as the ratio of the number of contacts in the decoy that are present in the native structure and the total number of contacts in the native structure. Near-native decoy structures are those that have a mean RMSD of <=4Å and fnat >=0.3. 53 out of 77 protein decoy set have 15 near-native structures each. The other decoy sets have around 5 to 45 decoys each (Table 1A). The ratio of near-native structures to non-native structures is 0.17.

Dockground dataset 1 - The dockground decoy set 1 contains 61 protein complexes each containing 100 non-native decoys and between 2-11 near-native decoys [208] (Table 1B). The near-natives are defined as the decoys that have an L-RMSD <=5Å. L-RMSD is defined as the RMSD of the ligand (smaller protein) after superimposing the receptor (larger protein). Even though the decoy set contained oligomers, the decoys were constructed such that 1 or more chains are clubbed together as a receptor and the remaining chains (mostly 1) are defined as a ligand. The L-RMSD values as indicated in the dataset was used in the study. The dataset contains 566 near-native and 6100 non-native structures. The total number of decoys (including the native structure) is 6666. The ratio of near-native structures to the non-native structure is 0.09.

CAPRI score_set - The CAPRI score_set [209] contained 13 complexes, 11 of which are dimers. The number of decoys in each set ranges between 600 to 2146. The classification of the decoys was based on fnat and i-RMSD, which is the RMSD of the interface residues [201]. The decoys were considered as incorrect (non-native) if the i-RMSD>4Å and fnat<0.1, acceptable if 2Å<i-RMSD<=4Å and fnat>=0.1, medium if 1Å<i-RMSD<=2Å and fnat>=0.1 and high if i-RMSD<=1Å and fnat>=0.5. A decoy was considered as a near-native if the decoy was acceptable/medium/high. For the study, we used the classification as mentioned in the dataset. The number of near natives ranges between 1 and 592 (Table 1). The dataset contains 1465 near-native and 16311 non-native structures. The total number of decoys (including the native structure) is 17776. The ratio of near-native structures to non-native structures is 0.09.

Table 1 – Number of near-native and non-native decoys for (A) BULLOCK's decoy set (B) Dockground decoy set 1 (C) CAPRI Score_set. The near-native count also contains the native structure. If the near native column has 1 for any complex it means, that the set only has a native structure.

(A)	PDB	Near-native	Non-native	(B)	PDB	Near-native	Non-native	(C)	PDB	Near-native	Non-native
	1AVX	15	62		1a2k_AB:C	3	100		Target29	98	1919
	1BUH	15	59		1a2y_AB:C	11	100		Target30	1	1119
	1CLV	15	66		1akj_AB:DE	11	100		Target32	4	596
	1D6R	15	62		1avw_A:B	11	100		Target37	64	1432
	1DFJ	15	78		1bth_LH:P	2	100		Target38	1	888
	1E6E	15	73		1bui_A:C	11	100		Target39	4	1383
	1E96	15	63		1bui_B:C	11	100		Target40	440	1706
	1EWY	14	68		1bvn_P:T	11	100		Target41	174	1006
	1F05	5	97		1cho_E:I	11	100		Target46	1	1640
	1FFW	14	63		1dfj_E:I	10	100		Target47	592	460
	1FQJ	15	57		1e96_A:B	11	100		Target50	45	1403
	1GHQ	15	44		1ewy_A:C	11	100		Target53	39	1362
	1GL1	15	66		1ezu_AB:C	11	100		Target54	2	1397
	1GLA	13	51		1f51_AB:E	11	100				
	1GPW	15	59		1f6m_A:C	11	100				
	1GXD	15	51		1fm9_A:D	11	100				
	1H9D	14	54		1g20_AB:EF	11	100				
	1HE1	15	68		1g6v_A:K	9	100				
	1IRI	22	63		1gpq_A:D	11	100				
	1J2J	13	60		1gpw_A:B	11	100				
	1JEQ	45	57		1he1_A:C	11	100				
	1JTG	15	62		1he8_A:B	2	100				
	1KAC	13	62		1hxy_AB:D	3	100				
	1KTZ	14	52		1jps_LH:T	11	100				
	1KXP	15	76		1ku6_A:B	11	100				
	1KXQ	14	59		1l9b_LMH:C	11	100				
	1MAH	15	69		1ma9_A:B	11	100				
	1OC0	15	59		1nbf_A:D	11	100				
	1OPH	15	48		1ook_AB:G	5	100				
	1OYV	14	70		1oph_A:B	11	100				
	1PPE	15	61		1p7q_AB:D	5	100				

	1PVH	14	55		1ppf_E:I	11	100				
	1QA9	15	64		1r0r_E:I	11	100				
	1R0R	15	61		1r4m_AB:I	2	100				
	1S1Q	15	59		1s6v_A:B	5	100				
	1SBB	14	50		1t6g_A:C	11	100				
	1T6B	15	55		1tmq_A:B	11	100				
	1U8F	22	79		1tx6_A:I	11	100				
	1UDI	15	65		1u7f_A:B	11	100				
	1UJZ	15	88		1uex_AB:C	2	100				
	1US7	15	46		1ugh_E:I	11	100				
	1XD3	15	78		1w1i_A:F	5	100				
	1Z0K	15	57		1wej_LH:F	11	100				
	1Z5Y	15	54		1wq1_R:G	11	100				
	1ZHH	15	63		1xd3_A:B	11	100				
	1ZHI	15	63		1xx9_A:CD	3	100				
	2A1A	14	68		1yvb_A:I	11	100				
	2A5T	15	54		1zy8_AB:K1	11	100				
	2A9K	15	61		1zy8_AB:K2	11	100				
	2AJF	14	55		2a5t_A:B	2	100				
	2AYO	15	83		2bkr_A:B	11	100				
	2B42	14	69		2bnq_AB:DE	2	100				
	2BTF	15	59		2btf_A:P	11	100				
	2FJU	15	51		2ckh_A:B	11	100				
	2GTP	14	54		2fi4_E:I	11	100				
	2HLE	15	66		2goo_A:C	11	100				
	2HQS	15	65		2kai_AB:I	11	100				
	2I25	15	60		2sni_E:I	11	100				
	2O8V	15	61		3fap_A:B	11	100				
	2OOB	15	59		3pro_A:C	11	100				
	2PSN	15	84		3sic_E:I	11	100				
	2UUY	15	72								
	2VDB	15	60								
	2X9A	15	62								
	2YVJ	15	67								
	3A4S	15	67								
	3BIW	15	54								
	3D5S	14	59								
	3DFQ	23	80								
	3H2V	15	61								
	3K75	15	43								
	3Q6M	16	87								
	3VLB	15	66								
	4H03	14	61								

	4M76	15	52								
	7CEI	15	65								

2.7. Identification of near-native complexes

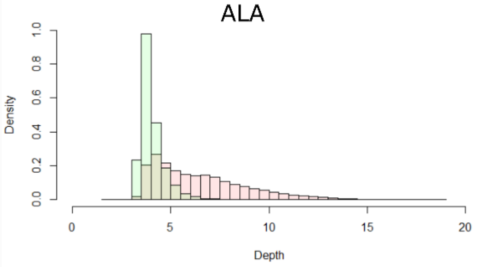
MODP was trained on the training set (Section 2.6 Bullock's dataset), by varying the z-scores between 0.8 to 3.0 in steps of 0.2. The z-score having the maximum MCC (in case of a tie in MCC, the one having the max f1 is chosen) for the 5 iterations was used to compute the average optimal MCC value. The same z-score was used on the testing and validation sets, Dockground decoy set 2 and CAPRI score_set to make the predictions.

3. Results

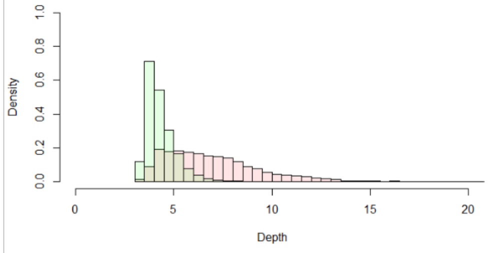
3.1. Construction of depth based scoring scheme – Monomer Oligomer Depth Potential (MODP)

In a monomer, the hydrophilic amino acids (Lys, Arg, Asp, Glu, Asn, Gln, Pro) prefer to be on the surface (~94% of the residues have depth<5), whereas the hydrophobic amino acids (Val, Ile, Leu, Met, Phe, Trp, Cys, Tyr) can be buried (compared to hydrophilic residues only 81-88% of the residues have depth <5) (Figure 2). In an oligomeric form, the distribution of the amino acids across different depths change, however, the peak of the distribution remains on the lower depth levels. The hydrophobic amino acids (Val, Ile, Leu, Met, Phe, Trp, Cys, Tyr) at the interface of an oligomeric complex gets buried (24-28% of the residues have depth<5) (Figure 2). However, the hydrophilic amino acids (Lys, Arg, Asp, Glu, Asn, Gln) and Pro seldom get buried (48-66% of the residues have depth<5) (Figure 2). Ala though has 93% of the residues with depth<5 in a monomer but in an oligomer it gets buried (35% residues having depth<5). The other amino acids (Gly, Ser, Thr, His) have 92-95% of the residues with depth <5, when in a monomer, wherein in an oligomer these residues have 40-43% with depth<5. (Figure 2).

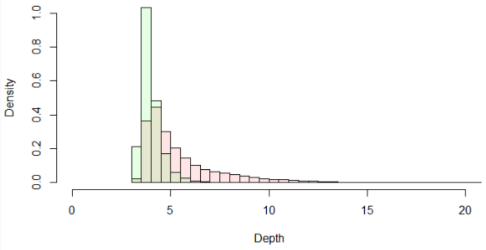
ALA



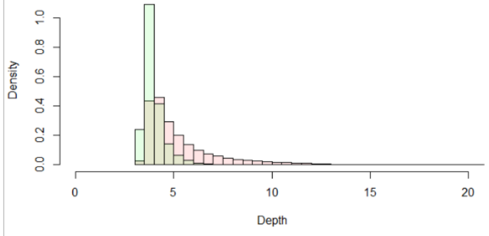
CYS



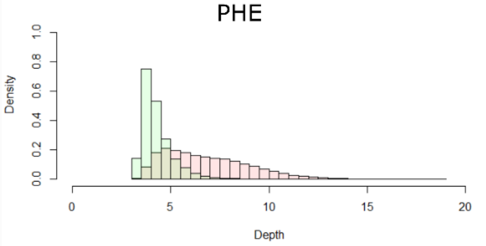
ASP



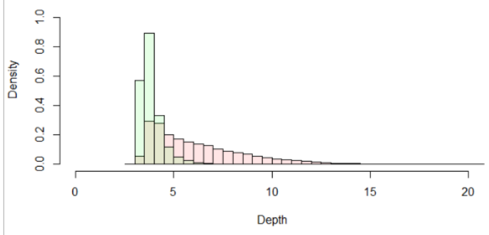
GLU



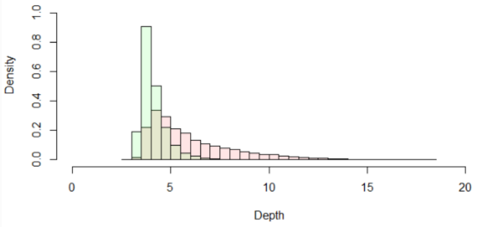
PHE



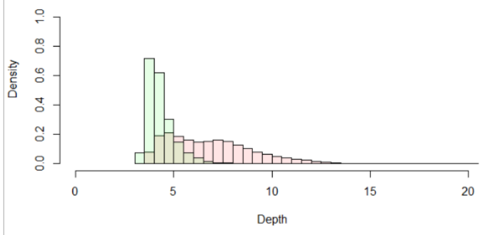
GLY



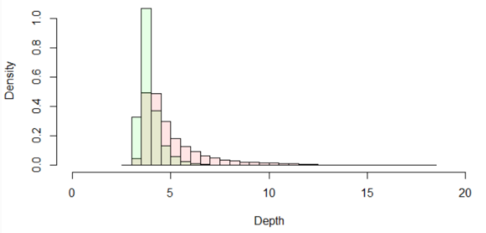
HIS



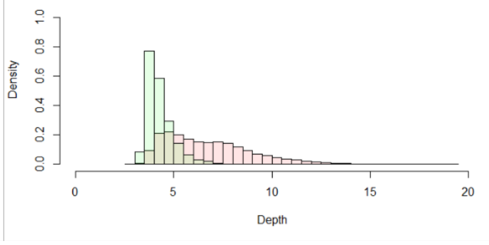
ILE



LLYS



MET



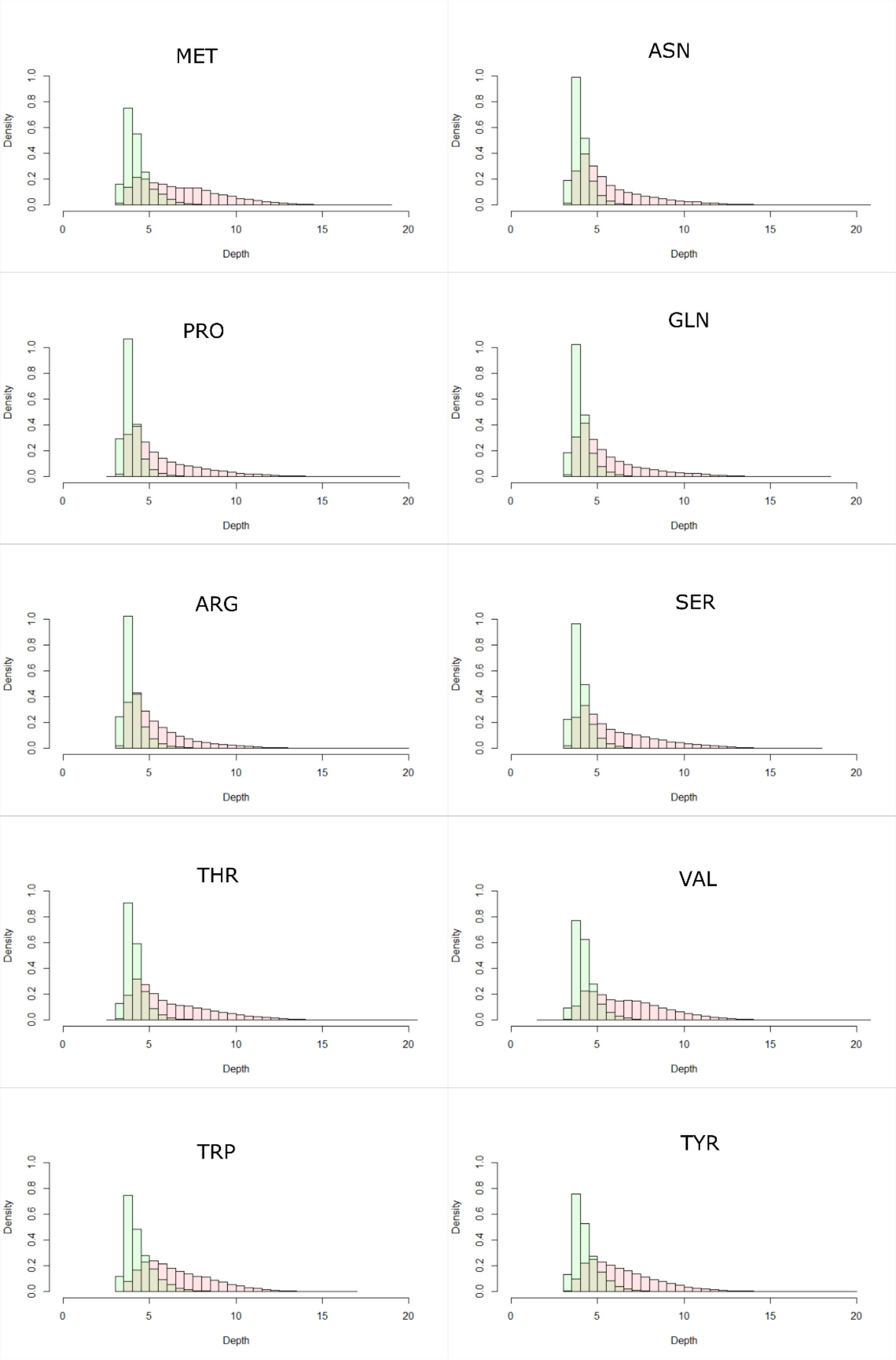
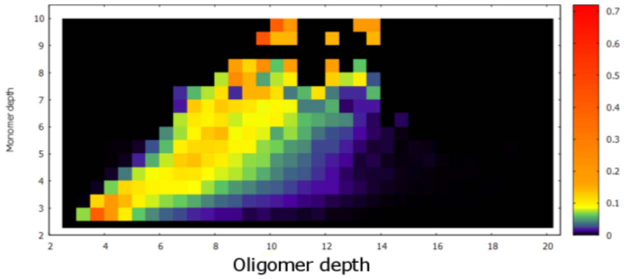
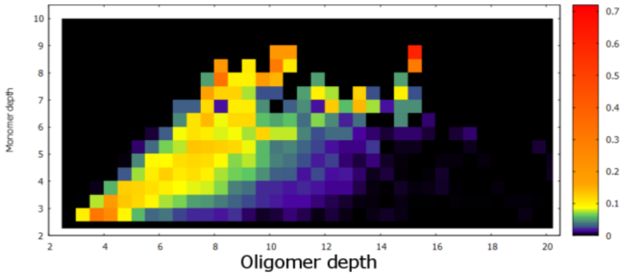


Figure 2 – Depth (in Å) distribution of the amino acids in monomeric form (green bars) and oligomeric form (red bars).

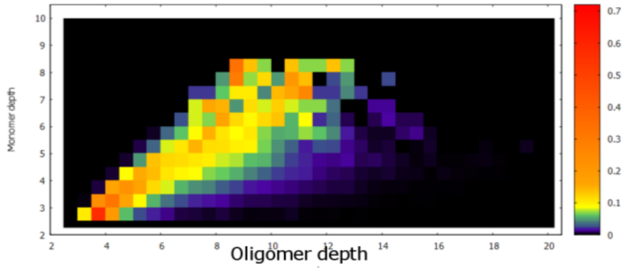
ALA



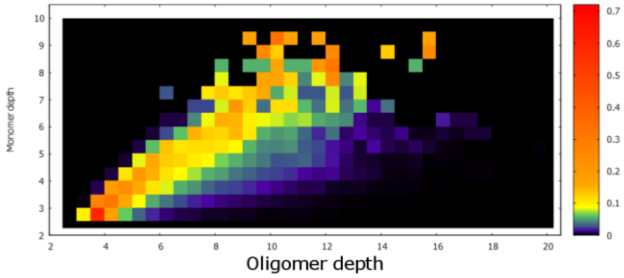
CYS



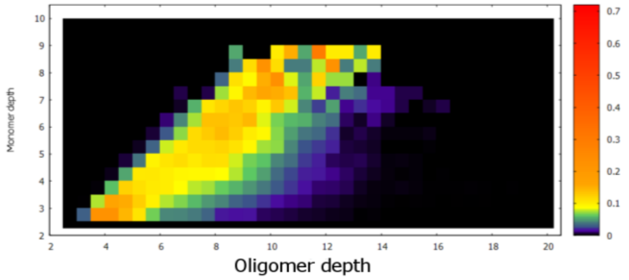
ASP



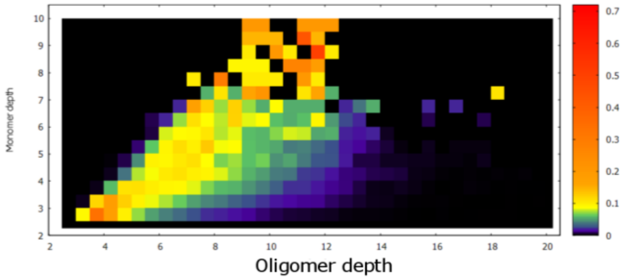
GLU



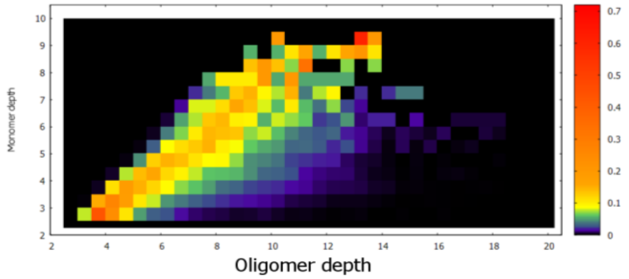
PHE



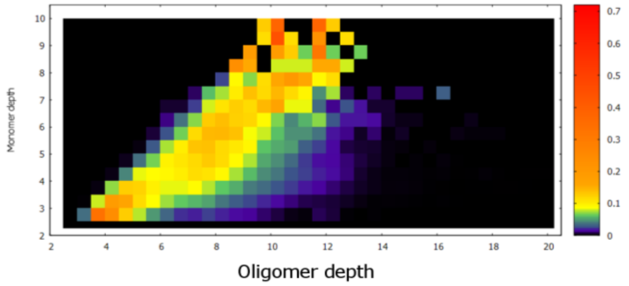
GLY



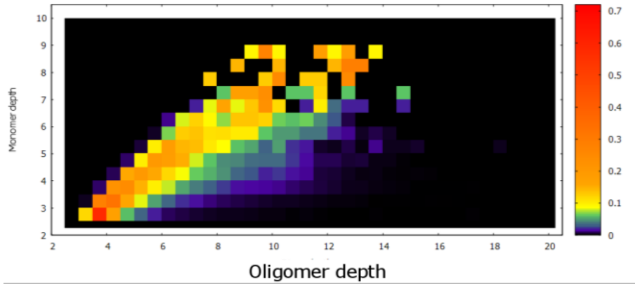
HIS



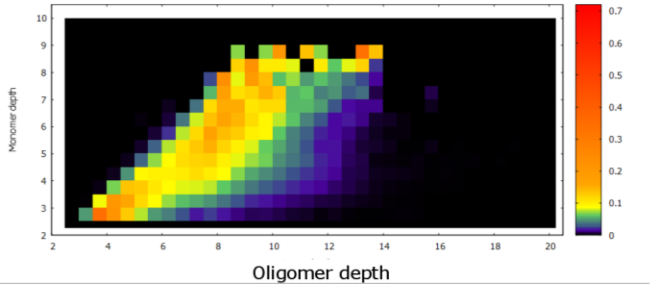
ILE



LYS



MET



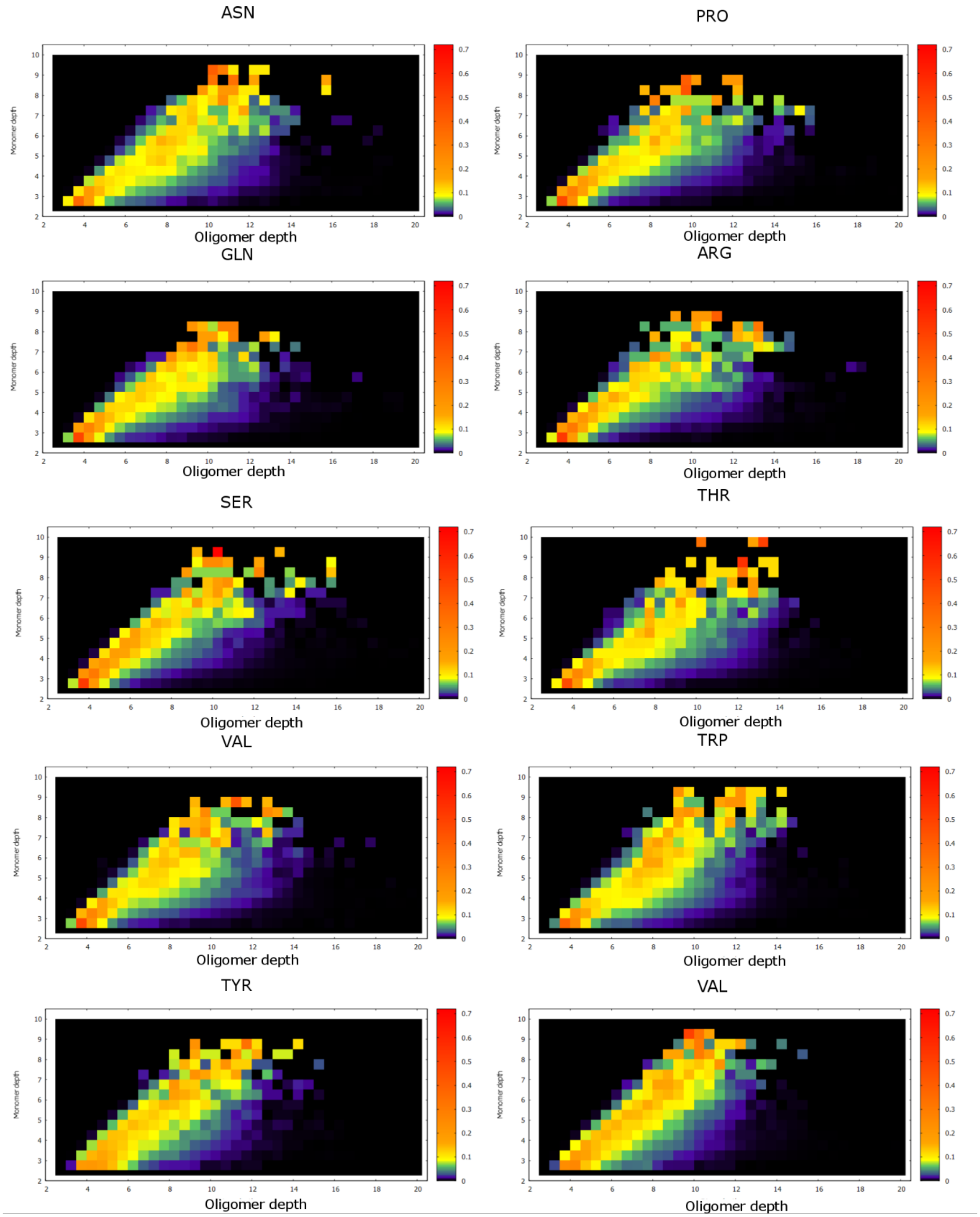


Figure 3 – The probability of the different residues to change depth (Å) upon complex

formation. The x-axis represents the depth of the residue in a complex and y-axis the depth of the residue in a monomer. The color code is based on the probability of the amino acid to transition from a monomeric depth to oligomeric depth upon complex formation. The reddish-yellow shades show higher probability compared to bluish-black shade.

Depending on the hydrophobicity of the residues, they have different propensities of movement from one depth in a monomer to another in a protein complex. The hydrophobic amino acids (such as Val, Leu, Ile, Trp, Met, Phe, Ala, Cys) undergoes larger changes in depth upon complex formation when compared with hydrophilic amino acids (such as Glu, Asp, Lys, His, Gln, Arg, Ser) (Figure 2). Ala and Gly allows higher changes in depth upon complex formation as compared to all other amino acids (Figure 3). These probability values are used to score protein-protein interfaces (Section 2.5). Hence, a true interface would predominantly contain residues that are more susceptible to getting buried upon complex formation.

3.2. Classification of the different decoys sets into near-native and non-native decoys and comparison to PIZSA

The ability of the technique MODP to predict native and non-native structures were analysed by 3 tests on 3 decoy sets. MODP was compared to PIZSA [37], which has itself compared to other techniques as CIPS [187], iScore [193] and was shown to be either at par or outperform the other techniques. The ranking of the decoys was done using the z-score (Section 3.2.1 and 3.2.2). The z-score cut off as determined from the training set was only used for the classification between near-native and non-native structure (Section 3.2.3).

3.2.1. The rank of the native structure among all its decoys

Bullock's dataset - On Bullock's dataset, MODP predicted 26, 42 and 59 native structures, whereas PIZSA predicted 16, 54 and 76 native structures in the top 5, 10 and 20 predictions respectively (Table 2A). Out of the 76 complexes, PIZSA ranked the native better than MODP in 37 complexes, while MODP ranked the native better than PIZSA in

33 complexes. In 6 of the decoy sets both the techniques had the same rank. The decoy sets where MODP outperformed PIZSA the difference in the rank was a maximum of 15. However, in 14 out of the 37 decoys where PIZSA outperformed MODP, the difference in rank between the 2 techniques was greater than 15.

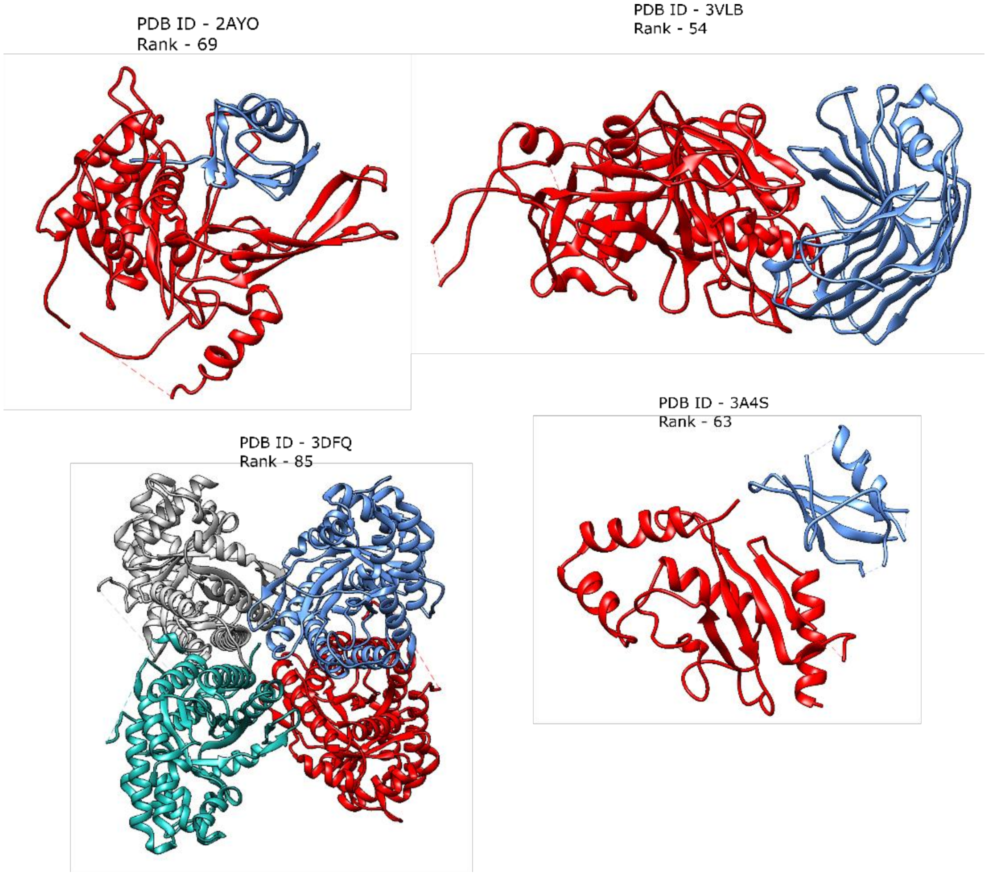


Figure 3 – Native PDBs in Bullock's decoy set with the rank of native >50.

4 of the native structures, PDB 3DFQ, 2AYO,3VLB and 2AS4 predicted by MODP had the native rank of >50 (Figure 3). PDBs 2AYO and 3VBL had a pocket like structure where another chain fits in. Hence, the structural complementarity of the pocket might be stabilizing the interface. PDB 2AS4 had a very small interface to score (23 residues as compared to an average of ~41 interfaces in other PDBs). One of the subunits (Chain A) has only 8 residues at the interface, hence the scrambled interfaces do not have differences in score w.r.t. that chain (Chain A) (the difference between the score of the native and mean score of the scrambled interface is 1 for the complex), resulting in low standard deviations of 1.3, hence the z-score of the interface is low. In the case of PDB

3DFQ, which is a homotetramer, the standard deviation of the score is 3.7, whereas the difference with the mean score of the scrambled interfaces is 0.8. The high standard deviation might be because the interface across the 4 chains contains 226 residues, leading to large variations in the scrambled interface, resulting in a low z-score.

Dockground decoy set 1 - MODP predicted 13, 22 and 30 native structures, whereas PIZSA predicted 55, 57 and 59 out of 61 native structures in the top 5, 10 and 20 ranks respectively (Table 2B). In 56 of the decoy sets PIZSA outperformed MODP, whereas MODP outperformed in only 3 decoy sets. 2 of the decoy sets had the same rank for the native structure by both the techniques. The cases where MODP outperformed PIZSA, the maximum difference in rank was 25.

CAPRI Score_set - CAPRI score set contains 13 decoy sets each containing 600 to 2146 decoys each. The rank in the case of this set was normalized to 100. MODP predicted 3, 4 and 9 native structures, whereas PIZSA predicted 6, 7 and 12 native structures in the top 5%, 10% and 20% ranks (Table 2C). PIZSA outperformed in the ranking in 8 of the decoy sets whereas MODP outperformed the ranking in 5 of the sets. All the cases where MODP outperformed PIZSA, the difference in rank was not greater than 16. However, in 2 cases where PIZSA outperformed MODP, the difference in rank was greater than 30 and 58.

Table 2 – Rank of the native structure among all the decoys in (A) BULLOCK's dataset (B) Dockground decoy set 1. Lower the rank, the better is the prediction. (C) Rank out of 100 for native structure among all structures in CAPRI Score_set.

(A)	PDB	MODP	PIZSA	(B)	PDB	MODP	PIZSA	(C)	Target	MODP	Pizsa
	1F05	13	3		1a2k_AB:C	32	1		Target29	23	22
	1JEQ	5	2		1a2y_AB:C	12	1		Target30	62	4
	1UJZ	6	3		1akj_AB:DE	25	1		Target32	3	2
	2PSN	7	9		1avw_A:B	12	1		Target37	31	1
	3Q6M	4	5		1bth_LH:P	23	19		Target38	13	5
	3DFQ	85	2		1bui_A:C	47	1		Target39	11	3
	1IRI	4	6		1bui_B:C	6	1		Target40	22	6
	1U8F	4	4		1bvn_P:T	93	1		Target41	4	20
	1AVX	12	12		1cho_E:I	82	1		Target46	17	19
	1BUH	10	10		1dfj_E:I	8	1		Target47	12	16

	1CLV	15	6		1e96_A:B	22	1		Target50	5	20
	1D6R	19	4		1ewy_A:C	5	1		Target53	20	16
	1DFJ	4	6		1ezu_AB:C	10	1		Target54	7	18
	1E6E	5	9		1f51_AB:E	11	36				
	1E96	9	9		1f6m_A:C	88	1				
	1EWY	7	11		1fm9_A:D	6	1				
	1FFW	5	11		1g20_AB:EF	3	1				
	1FQJ	10	8		1g6v_A:K	15	1				
	1GHQ	5	10		1gpq_A:D	6	1				
	1GL1	33	5		1gpw_A:B	27	1				
	1GLA	4	9		1he1_A:C	6	1				
	1GPW	8	7		1he8_A:B	3	1				
	1GXD	2	17		1hxy_AB:D	16	9				
	1H9D	7	9		1jps_LH:T	73	1				
	1HE1	4	4		1ku6_A:B	15	1				
	1J2J	15	8		1l9b_LMH:C	5	1				
	1JTG	2	6		1ma9_A:B	45	1				
	1KAC	18	9		1nbf_A:D	7	1				
	1KTZ	1	9		1ook_AB:G	6	19				
	1KXP	11	4		1oph_A:B	63	2				
	1KXQ	3	13		1p7q_AB:D	23	8				
	1MAH	15	15		1ppf_E:I	3	1				
	1OC0	34	7		1r0r_E:I	38	1				
	1OPH	25	12		1r4m_AB:I	2	1				
	1OYV	4	12		1s6v_A:B	19	1				
	1PPE	40	10		1t6g_A:C	53	1				
	1PVH	8	6		1tmq_A:B	3	1				
	1QA9	7	20		1tx6_A:I	22	1				
	1R0R	25	5		1u7f_A:B	2	1				
	1S1Q	3	4		1uex_AB:C	26	29				
	1SBB	17	9		1ugh_E:I	70	1				
	1T6B	13	8		1w1i_A:F	71	1				
	1UDI	17	9		1wej_LH:F	103	1				
	1US7	2	7		1wq1_R:G	80	1				
	1XD3	4	8		1xd3_A:B	1	1				
	1Z0K	2	6		1xx9_A:CD	16	1				
	1Z5Y	2	14		1yvb_A:I	25	1				
	1ZHH	1	3		1zy8_AB:K1	52	1				
	1ZHI	21	11		1zy8_AB:K2	38	1				
	2A1A	6	13		2a5t_A:B	4	1				
	2A5T	4	8		2bkr_A:B	5	1				
	2A9K	12	13		2bnq_AB:DE	25	1				
	2AJF	38	7		2btf_A:P	5	1				

	2AYO	69	15		2ckh_A:B	29	3				
	2B42	29	4		2fi4_E:I	34	1				
	2BTF	6	8		2goo_A:C	1	1				
	2FJU	8	9		2kai_AB:I	24	4				
	2GTP	7	11		2sni_E:I	30	1				
	2HLE	11	7		3fap_A:B	94	4				
	2HQS	20	7		3pro_A:C	24	1				
	2I25	3	7		3sic_E:I	9	1				
	2O8V	25	8								
	2OOB	7	6								
	2UUY	14	9								
	2VDB	2	10								
	2X9A	15	5								
	2YVJ	24	17								
	3A4S	63	15								
	3BIW	6	11								
	3D5S	5	6								
	3H2V	2	11								
	3K75	19	12								
	3VLB	54	14								
	4H03	38	11								
	4M76	33	10								
	7CEI	35	5								

3.2.2. Identification of native/near-native in top 10 predictions

The prediction ability of the MODP to identify native/near-native structure was evaluated based on CAPRI protocols i.e. the number of decoy sets having at least one near-native structure in the top 10 predictions. We also computed the number of near-native predictions in the top 10 predictions.

Bullock’s decoy set - 70 predictions by MODP and 72 predictions by PIZSA on BULLOCK’s decoy sets had at least 1 prediction in the top 10 predictions for the decoy set (Table 3A). In the top 10 predictions by MODP, at least 5 near-native structures were present in 45 decoy sets. However, PIZSA had only 10 decoy sets with at least 5 near-native structures.

10 of the decoy sets had an equal number of near-native structures in the top 10 predictions by both techniques. 51 decoy sets had MODP having a larger number of near-native predictions as compared to PIZSA, whereas 22 sets had a larger number of near-native predictions by PIZSA as compared to MODP in the top 10 predictions.

Dockground decoy set 1 - 40 predictions by MODP and 59 predictions by PIZSA on Dockground decoy set 1 had at least 1 prediction in the top 10 predictions for the decoy set (Table 3B). The top 10 predictions by MODP and PIZSA had 11 and 12 decoy sets (out of 61) respectively with at least 5 near-native structures.

9 of the decoy sets had an equal number of near-native predictions in the top 10 predictions. 19 decoy sets had MODP having a larger number of near-native predictions as compared to PIZSA, whereas 33 sets had a larger number of near-native predictions by PIZSA as compared to MODP in the top 10 predictions.

CAPRI Score_set - Both PIZSA and MODP predicted 4 decoy sets (out of 13) with at least 1 near-native structure in the top 10 predictions (Table 3C). 1 of these decoy sets had 5 near-native structures predicted by MODP in the top 10 predictions. However, for the top 10 predictions by PIZSA on these decoy sets, there were 2 decoy sets with at least 5 near-native structures. There were 3 decoy sets each where MODP had a larger number of near-native structures in the top 10 scores compared to PIZSA and vice versa.

Table 3 – Number of near-native structures in the top 10 predictions for (A) Bullock's decoy set (B) Dockground decoy set 1 (C) CAPRI score_set.

(A)	PDB	MODP	PIZSA	(B)	PDB	MODP	PIZSA	(C)	PDB	MODP	PIZSA
	1F05	2	1		1a2k_AB:C	0	2		Target29	1	0
	1JEQ	10	4		1a2y_AB:C	1	7		Target30	0	0
	1UJZ	4	4		1akj_AB:DE	5	8		Target32	0	1
	2PSN	10	4		1avw_A:B	0	3		Target37	0	1
	3Q6M	6	6		1bth_LH:P	0	0		Target38	0	0
	3DFQ	0	5		1bui_A:C	0	4		Target39	0	0
	1IRI	9	2		1bui_B:C	2	3		Target40	1	5
	1U8F	10	6		1bvn_P:T	0	1		Target41	2	0
	1AVX	2	1		1cho_E:I	0	2		Target46	0	0
	1BUH	6	4		1dfj_E:I	5	1		Target47	7	5
	1BVN	0	2		1e96_A:B	6	2		Target50	0	0

	1CLV	1	3		1ewy_A:C	6	5		Target53	0	0
	1D6R	2	5		1ezu_AB:C	10	3		Target54	0	0
	1DFJ	6	3		1f51_AB:E	1	4				
	1E6E	5	3		1f6m_A:C	0	2				
	1E96	4	4		1fm9_A:D	5	2				
	1EFN	3	4		1g20_AB:EF	3	6				
	1EWY	5	4		1g6v_A:K	2	2				
	1F34	2	0		1gpq_A:D	3	2				
	1FFW	6	3		1gpw_A:B	0	4				
	1FQJ	4	4		1he1_A:C	1	2				
	1GCQ	2	5		1he8_A:B	2	2				
	1GHQ	7	3		1hxy_AB:D	0	3				
	1GL1	2	3		1jps_LH:T	0	5				
	1GLA	5	3		1ku6_A:B	2	3				
	1GPW	7	4		1l9b_LMH:C	3	8				
	1GXD	7	2		1ma9_A:B	5	4				
	1H9D	9	3		1nbf_A:D	3	7				
	1HE1	8	4		1ook_AB:G	3	1				
	1J2J	4	4		1oph_A:B	2	2				
	1JTG	5	3		1p7q_AB:D	0	4				
	1KAC	2	4		1ppf_E:I	4	3				
	1KTZ	8	3		1r0r_E:I	2	1				
	1KXP	1	3		1r4m_AB:I	2	1				
	1KXQ	9	1		1s6v_A:B	2	2				
	1MAH	3	1		1t6g_A:C	2	1				
	1OC0	2	3		1tmq_A:B	2	3				
	1OPH	0	1		1tx6_A:I	0	2				
	1OYV	7	0		1u7f_A:B	4	2				
	1PPE	1	2		1uex_AB:C	0	0				
	1PVH	6	4		1ugh_E:I	0	2				
	1QA9	6	3		1w1i_A:F	1	1				
	1R0R	0	3		1wej_LH:F	0	9				
	1S1Q	8	4		1wq1_R:G	0	1				
	1SBB	7	2		1xd3_A:B	7	2				
	1T6B	4	3		1xx9_A:CD	2	1				
	1UDI	1	2		1yvb_A:I	0	6				
	1US7	7	3		1zy8_AB:K1	2	6				
	1XD3	7	2		1zy8_AB:K2	0	5				
	1Z0K	5	5		2a5t_A:B	1	1				
	1Z5Y	7	3		2bkr_A:B	1	3				
	1ZHH	7	4		2bnq_AB:DE	0	2				
	1ZHI	6	1		2btf_A:P	7	4				
	2A1A	9	2		2ckh_A:B	6	7				

	2A5T	6	2		2fi4_E:I	0	3				
	2A9K	9	0		2goo_A:C	3	1				
	2AJF	1	3		2kai_AB:I	0	4				
	2AYO	1	3		2sni_E:I	1	1				
	2B42	3	3		3fap_A:B	0	2				
	2BTF	8	3		3pro_A:C	3	1				
	2FJU	6	2		3sic_E:I	8	2				
	2GTP	5	3								
	2HLE	3	2								
	2HQS	2	3								
	2I25	5	3								
	2O8V	3	3								
	2OOB	7	5								
	2OUL	8	1								
	2UUY	1	3								
	2VDB	7	2								
	2X9A	5	4								
	2YVJ	7	1								
	3A4S	0	0								
	3BIW	5	2								
	3D5S	6	4								
	3H2V	8	5								
	3K75	4	2								
	3PC8	4	3								
	3SGQ	3	3								
	3VLB	0	2								
	4H03	1	2								
	4M76	3	5								
	7CEI	1	5								

3.2.3. Accuracy of classification of near-native structures from non-native structures

Bullock's decoy set - The training and the testing set contain 60 decoy sets (48 for training and 12 for testing). The validation set contain 16 decoy set (1GXD, 7CEI, 1DFJ, 2X9A, 1CLV, 2PSN, 2B42, 1KAC, 1KXP, 1FFW, 1EWY, 1JEQ, 2AYO, 2GTP, 2UUY, 1ZHI). A grid search was done in the range of 0.8 to 3 in an interval of 0.2. The optimal z-score was calculated as the average z-score (corresponding to the maximum MCC) over 5 iterations. All 5 iterations had the maximum MCC at a z-score cut off of 1.4. Hence, the average z-score cutoff, used on the testing and the validation sets, Dockground decoy set 1 and CAPRI Score_set was 1.4. The average MCC of MODP on the training, testing and validation sets are 0.36+/-0.01, 0.36+/-0.06 and 0.32 respectively (Table 3). The average MCC of PIZSA is 0.17+/-0.01, 0.16+/-0.03 and 0.16 on the training, testing and validation sets respectively (Table 4).

Dockground decoy set 1 – The MCC of MODP is 0.16 whereas that of PIZSA is 0.07 (Table 4) on dockground decoy set 1.

CAPRI Score_set - The MCC of MODP is 0.18 whereas that of PIZSA is 0.06 (Table 4).

Across the different decoy sets, MODP has a higher sensitivity, precision and f1 as compared to PIZSA. PIZSA though has a better specificity as compared to MODP

Table 4 – Measures of the accuracy of MODP and PIZSA in the classification of near-native from non-native structures in the different datasets. The cell having the higher value among the two techniques have been highlighted in red.

Dataset	**Sensitivity**		**Specificity**		**Precision**		**Accuracy**		**F1**		**MCC**	
	MODP	**PIZSA**	**MODP**	**PIZSA**	**MODP**	**PIZSA**	**MODP**	**PIZSA**	**MODP**	**PIZSA**	**MODP**	**PIZSA**
Bullock's dataset - Training	0.56 +/- 0.03	0.29 +/- 0.01	0.85 +/- 0.01	0.87 +/- 0.00	0.41 +/- 0.01	0.30 +/- 0.01	0.80 +/- 0.01	0.78 +/- 0.01	0.48 +/- 0.01	0.30 +/- 0.01	0.36 +/- 0.01	0.17 +/- 0.01
Bullock's dataset - Testing	0.56 +/- 0.10	0.29 +/- 0.03	0.85 +/- 0.02	0.87 +/- 0.02	0.41 +/- 0.03	0.30 +/- 0.04	0.80 +/- 0.02	0.77 +/- 0.01	0.48 +/- 0.05	0.29 +/- 0.03	0.36 +/- 0.06	0.16 +/- 0.03
Bullock's dataset - Validation	0.55	0.25	0.81	0.89	0.37	0.32	0.77	0.79	0.45	0.28	0.32	0.16
Dockground decoy set 1	0.49	0.31	0.76	0.80	0.16	0.12	0.74	0.76	0.24	0.18	0.16	0.07

CAPRI Score_set	0.74	0.50	0.58	0.60	0.14	0.10	0.59	0.60	0.23	0.17	0.18	0.06

4. Discussions

Different amino acids have different propensities to be in different protein environments [142]. The polar amino acids prefer to be on the surface, whereas non-polar amino acids prefer to be buried. The amino acids that are at the protein-protein interface, are predominantly solvent exposed in their monomeric form and get buried upon complex formation. These amino acids have different propensities to get buried on complex formation. The hydrophilic amino acids (Asp, Glu, Asn, Gln, Lys etc.) at the interface have lower propensity to get buried upon complex formation. However, the hydrophobic amino acids (Phe, Val, Ile, Leu, Met, Cys etc.) have a higher propensity to get buried upon complex formation. We used the probability of the different amino acids to shift from one depth in a monomer to another in a complex to score protein complex models. We calculated how different is the computed score against a random background of scrambled interface scores, called the z-score. We used these z-scores to classify the near-native structures from non-native structures.

We tested these z-scores to identify and classify near-native interfaces from non-native interfaces on 3 different datasets – Bullock's dataset, Dockground decoy set 1 and CAPRI Score_set. Bullock's decoy set had ~15% of the decoys as near-native whereas the other two decoy sets had ~8% (0-56%) of the decoys as near-native on average. Across all the decoy sets PIZSA outperformed MODP in identifying the native structure. This can be because of PIZSA being a fine-grained scoring scheme developed to identify near-native structures. PIZSA involves both the residue pairing preference at the protein-protein interface along with a preferred number of shared atomic contacts. Hence the scoring scheme goes to an atomic level granularity, which makes it ideal for the identification of native structures. MODP, however, is a coarse-grained scheme with the usage of the probability of the amino acid to undergo depth changes upon complex formation.

Both the techniques had a similar number of decoy sets with at least 1 near-native structure prediction in the top 10 decoys for Bullock's and CAPRI decoy set. However, PIZSA outperformed MODP (59 predictions by PIZSA vs 40 predictions by MODP) in Dockground decoy set 1. We could not attribute any specific reason for PIZSA outperforming MODP in Dockground decoy set 1. MODP outperformed PIZSA in having a larger number of near-native structures in the top 10 predictions in Bullock's decoy set, whereas both of them had similar performance in CAPRI Score_set.

We compared MODP to PIZSA in classifying near-native structures from non-native structures. Across all the 3 datasets MODP outperformed PIZSA with higher f1 and MCC values. The accuracy and specificity of both the techniques are similar, however, MODP has a higher sensitivity and precision as compared to PIZSA. This indicates a large number of true positives were predicted by MODP without increasing the number of false positives. Hence, MODP was able to predict a larger number of near-native configurations as binders as compared to PIZSA. This again could be attributed to the fact that PIZSA contains residue and atomic propensity terms making the scoring scheme fine-grained as compared to MODP. Hence PIZSA, identifies native structures well, but MODP identifies a larger number of near-native structures from the decoy sets. Hence the two techniques can serve as complementary to one another and a scoring scheme involving both these techniques may outperform the individual scoring schemes.

Various changes can be made to MODP, to check for improvement in the ranking of the native structure. We can redefine the way z-score is computed i.e. instead of scrambling the interface, the entire protein could be scrambled. In the present technique, the residues at the interface are scrambled, which are the ones that will get buried at the interface. Hence a better scrambled interface might be one where the entire protein is scrambled, hence removing the bias towards residues that would get buried at the interface. The interface residues can also be weighted based on the type of amino acids or based on the depth change of the residue in an oligomer. This would involve giving greater weights to the residues that undergo burial on the complex formation or the ones whose depth has changed more compared to others (say depth change>1 Å). These might help identify the patch of the residues that form the core of the interface.

To conclude, we developed a depth based scoring scheme to identify near-native structures from among non-native decoys. This technique was shown to outperform the state of art technique PIZSA. However, PIZSA outperforms MODP in the identification of native from the decoy sets. PIZSA can hence be useful in the identification of native structures, whereas MODP can be used in the identification of near-native structures, a key towards the identification of protein complex models. The depth dependent potential can in the future be modified suitably to predict patches on the surface of the protein that are suitable to undergo oligomerization. This would involve identification of regions on the protein surface that contain residues that allow changes in depth upon complex formation.

This chapter described the characterization of protein-protein interfaces based on residue environments and scoring of poses of protein complexes. The next chapter will describe the identification of interface residues that are important for binding (hotspot residues).

Chapter 6 - Classification of interface residues into hotspot and non-hotspot residues

1. **Parameters to differentiate between hotspot and non-hotspot residues**
2. **Creation of a measure to predict hotspot residues**
3. **Comparison to pre-existing techniques**

.

1. Introduction

The interface residues do not contribute equally to the binding free energy of the interactions between proteins [210]. A small subset of the interface residues called hotspot residues contributes predominantly to the binding free energy of proteins [211–214]. A hotspot residue is precisely defined as residue whose mutation to alanine reduces the binding free energy by at least 2 kcal/mol [215]. The hotspot residues are experimentally determined by alanine scanning wherein the interface residues are mutated to alanine, one residue mutation at a time, and the change in binding free energy is calculated. An alanine mutation is conjectured to retain the secondary structure of the region of the interface while disrupting the interactions with the binding partner [216], hence determining the contribution of the residue in the binding event. Hotspots from experimental alanine scanning mutagenesis experiments are deposited in Alanaine Scanning Energetics Database (ASEdb) [215]. Another database, Binding Interface Database (BID) [217] contains the experimentally verified interface hotspot residues from literature. However experimental alanine scanning is time consuming and expensive, hence computational determination of hotspot residues is a viable alternative.

Computational methods for prediction of hotspot residues are of three types – molecular dynamics based methods, physical/knowledge based methods and machine learning based methods. Molecular dynamics based methods [218–220] estimate the change in binding free energy by simulating alanine mutations. But these methods cannot be used on a large scale as they are time consuming and computationally intensive. Empirical energy functions that are calibrated using experimental data can be used to detect hotspots. One such energy function, Robetta [221] provides a simple physical model for the prediction of hotspot residues using solvation, hydrogen bonds and packing of residues. The energetic contribution of the residues at the interface can also be estimated using non-covalent interactions and can be used to predict hotspot residues [222]. Hotspot residues can also be predicted utilizing knowledge based interaction potentials and solvation of residues [223]. Various machine learning tools (such as neural network, decision trees support vector machine, Bayesian network etc.) have been developed, which learn various features related to the properties of amino acid, conservation patterns, solvation etc. from a training set of hotspot and non-hotspot residues [224–231]. These machine learning tools can then be used to predict hotspot residues. However, a lot of machine learning approaches overfit data, wherein the machine learning tools have a high accuracy for their datasets but fail in other datasets [232]. Besides, the datasets are imbalanced as the number of non-hotspot residues is much more than the hotspot residues, adding to the disadvantage of using machine learning based methods [233].

Certain residues like Tyr, Arg and Trp predominantly occur as hotspot residues because it can form both hydrogen bonding and hydrophobic interactions [211]. Solvent occlusion serves as an important factor for the energetics of interactions [234,235], hence hotspot residues are surrounded by residues that shield them from bulk solvent. As a result, hotspot residues are buried compared to other interface residues. Previously, it has been shown that the change in depth upon complex formation can be used to identify residues that have the largest contribution to the binding free energy [7]. Depth change can also be used to quantify the ‘O ring’ model of Bogan and Thorn [211]. Also, protein interfaces are found to be more conserved as compared to the rest of the protein [118,236–238]. Conservation of residues at the protein interfaces can hence be used to identify hotspot residues in protein interfaces [237]. In addition to conservation and burial of residues, the pair-potential of residues (the propensity of an interface to interact with another residue) were found to be important determinants for predicting hotspot residues [35,239]. Knowledge based pair potentials are extracted from frequencies of contact between the residues as seen on the protein interface. These pair potentials can hence be used as a feature to identify hotspot residues.

We developed a technique, DepthCon for prediction of hotspot residues using an empirical decision tree based approach using the residue depth, conservation and residue pair potential. Various combinations of one or more of these properties were empirically used to predict hotspot residues. We compared our technique with existing techniques such as Hotpoint, KFC2 [228] and PredHS [231]. Hotpoint is a tool similar to our prediction scheme, which makes use of contact potential and solvent accessible surface area to predict hotspot residues. The other two tools KFC and PredHS use machine learning. These tools performed well in their own training/testing sets but their

MCC and f1 (for a definition of MCC and f1 refer to Chapter 2 Section 9) dropped when utilized to make predictions on different datasets, whereas our prediction scheme provided a stable MCC and f1 values across different training and testing sets.

2. Methods

2.1. Dataset of hotspot residues

2.1.1 Training set

The Hotpoint test set was used as a training set for the prediction scheme named DepthCon. It consisted of 112 residues (belonging to 19 PDBs) on 25 monomers containing 54 hotspot and 58 non-hotspot residues. This dataset was created from the BID database with sequences at a redundancy of <35%. The ones labeled with “strong” interactions were called hotspot residues while others were called non-hotspot residues.

2.1.2. Testing set

Test set 1 – This testing set was the Hotpoint training set containing 150 residues (from 23 interfaces belonging to 14 PDBs) of which 58 were hotspot and the remaining 92 non-hotspot residues. The set was created from the ASEdb with a sequence identity <35%. The residues with binding free energy ≥ 2 kcal/mol were termed hotspot residues. Residues with binding free energy < 0.4 kcal/mol were termed non-hotspot residues.

Test set 2 – This testing dataset was created from protein sequences with <35% sequence identity from ASEdb. The dataset contained 199 residues (from 23 interfaces belonging to 14 PDBs) of which 58 were hotspot residues and 141 were non-hotspot residues. The hotspot residues had binding free energy of ≥ 2 kcal/mol whereas the rest were termed as non-hotspot residues.

Test set 3 – This dataset contained the residues from the training set of PredHS but was not used to train KFC. The dataset contained 25 hotspot residues and 56 non-hotspot residues (from 18 interfaces belonging to 11 PDBs). This dataset was created from ASEdb.

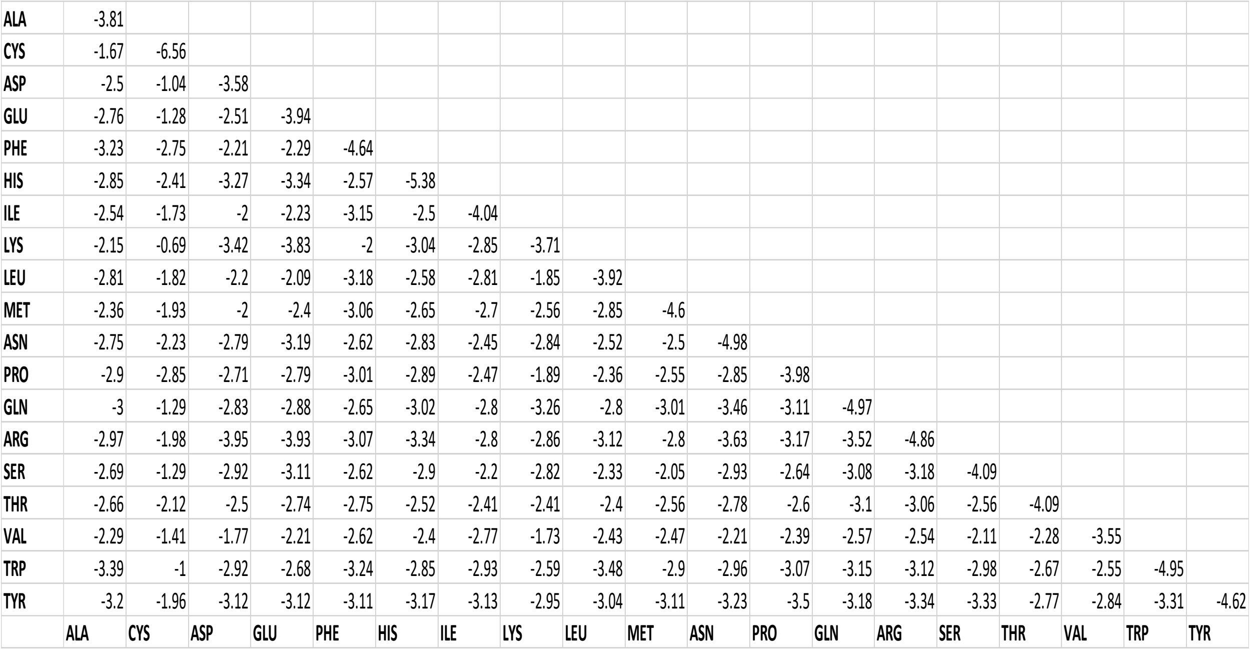
Test set 4 – This dataset contains the test set as used by PredHS and KFC. This dataset contained 39 hotspot residues and 88 non-hotspot residues (belonging to 24 interfaces from 18 proteins). This dataset was created from the BID database.

2.1. Features used for classifying hotspot and non-hotspot residues

2.1.1. Residue depth change on complex formation

The depth of a residue in a protein complex (*depth complex*) was calculated using the DEPTH program [12] using default parameters of the minimum number of neighborhood atoms as 2 and the number of solvation cycles of 25. The depth of the residues in individual monomers (*depth monomer*) was also calculated using the default parameters by separating the chains. This was done under the assumption that there were no conformational changes in the individual monomers during complex formation. The change in depth of a residue upon complex formation was then calculated as the difference between the depth of the residue in a complex and a monomer.

2.1.2. Contact potential



ALA	-3.81																		
CYS	-1.67	-6.56																	
ASP	-2.5	-1.04	-3.58																
GLU	-2.76	-1.28	-2.51	-3.94															
PHE	-3.23	-2.75	-2.21	-2.29	-4.64														
HIS	-2.85	-2.41	-3.27	-3.34	-2.57	-5.38													
ILE	-2.54	-1.73	-2	-2.23	-3.15	-2.5	-4.04												
LYS	-2.15	-0.69	-3.42	-3.83	-2	-3.04	-2.85	-3.71											
LEU	-2.81	-1.82	-2.2	-2.09	-3.18	-2.58	-2.81	-1.85	-3.92										
MET	-2.36	-1.93	-2	-2.4	-3.06	-2.65	-2.7	-2.56	-2.85	-4.6									
ASN	-2.75	-2.23	-2.79	-3.19	-2.62	-2.83	-2.45	-2.84	-2.52	-2.5	-4.98								
PRO	-2.9	-2.85	-2.71	-2.79	-3.01	-2.89	-2.47	-1.89	-2.36	-2.55	-2.85	-3.98							
GLN	-3	-1.29	-2.83	-2.88	-2.65	-3.02	-2.8	-3.26	-2.8	-3.01	-3.46	-3.11	-4.97						
ARG	-2.97	-1.98	-3.95	-3.93	-3.07	-3.34	-2.8	-2.86	-3.12	-2.8	-3.63	-3.17	-3.52	-4.86					
SER	-2.69	-1.29	-2.92	-3.11	-2.62	-2.9	-2.2	-2.82	-2.33	-2.05	-2.93	-2.64	-3.08	-3.18	-4.09				
THR	-2.66	-2.12	-2.5	-2.74	-2.75	-2.52	-2.41	-2.41	-2.4	-2.56	-2.78	-2.6	-3.1	-3.06	-2.56	-4.09			
VAL	-2.29	-1.41	-1.77	-2.21	-2.62	-2.4	-2.77	-1.73	-2.43	-2.47	-2.21	-2.39	-2.57	-2.54	-2.11	-2.28	-3.55		
TRP	-3.39	-1	-2.92	-2.68	-3.24	-2.85	-2.93	-2.59	-3.48	-2.9	-2.96	-3.07	-3.15	-3.12	-2.98	-2.67	-2.55	-4.95	
TYR	-3.2	-1.96	-3.12	-3.12	-3.11	-3.17	-3.13	-2.95	-3.04	-3.11	-3.23	-3.5	-3.18	-3.34	-3.33	-2.77	-2.84	-3.31	-4.62
	ALA	CYS	ASP	GLU	PHE	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR

Figure 1 – Amino acid pairwise score for residues at the interface

We used a contact potential developed by Dhawanjewar *et al.* [240] using residue pair preferences between the side chains of residues based on the formula as mentioned below. The score for each pair of amino acids is given in Figure 1. The scores vary in the range -0.69 to -6.59, with lower the value, indicating more favorable the interaction

$$\mathrm{S}_{i,j} = -\log\left[\left(\sum_{\forall\text{ interfaces}} \frac{\dfrac{f_{ij}^{int}}{\sum_{\forall ab} cifa_{ab}^{int}} \times \dfrac{cifa_{ij}^{int}}{\max(cifa_{ij}^{int})}}{\dfrac{f_i}{N_m}\dfrac{f_j}{N_n} \times \langle cifa_{ij}^{int}\rangle}\right) \div N_{total}\right]$$

where,

$$\begin{aligned}
f_{ij}^{int} &= \text{frequency of } i-j \text{ residue pairs across the interface}\\
cifa_{ij}^{int} &= \min\left[\frac{\text{interacting atoms}_i}{\text{total atoms}_i}, \frac{\text{interacting atoms}_j}{\text{total atoms}_j}\right]\\
cifa_{ab}^{int} &= \text{frequency of any residue pair } a-b \text{ weighted by their respective } cifa\\
\frac{f_i}{N_m} &= \text{frequency of residues of type } i \text{ in the subunit } m\\
\frac{f_j}{N_n} &= \text{frequency of residues of type } j \text{ in the subunit } n\\
N_m, N_n &= \text{Number of subunits in subunits } m \text{ and } n \text{ respectively}\\
\langle cifa_{ij}^{int}\rangle &= \text{average value of } cifa \text{ observed in the dataset for } i-j \text{ pairs across the interface}\\
N_{total} &= \text{total number of protein complexes in the dataset}
\end{aligned}$$

For calculation of the contact potential, the residues were considered to be in contact if the distance between the centroid of the side chain of the residues was less than 7 Å [223]. The *contact potential* for a residue *i* is calculated by summing over all the contacts of the residue *i* with residues from the interacting chain. The absolute value of the contact potential was used. Higher values of the contact potential of a residue would indicate more favorable interactions.

$$Contact\ potential = abs\left(\sum_{j=1}^{j=n} S_{i,j}\right)$$

2.1.3. Conservation

The homologs of the protein subunits were annotated using 5 psiblast runs [241] with an e-value cut off of 0.0001. The conservation for each residue was calculated using the following formula similar to Shannon entropy [242]

$$\text{Conservation} = \sum_{i=0}^{20} P^i * logP^i$$

, where P^i is the probability of residue i at that position. Higher values indicate higher conservation of the position.

2.3. Prediction of hotspot residues

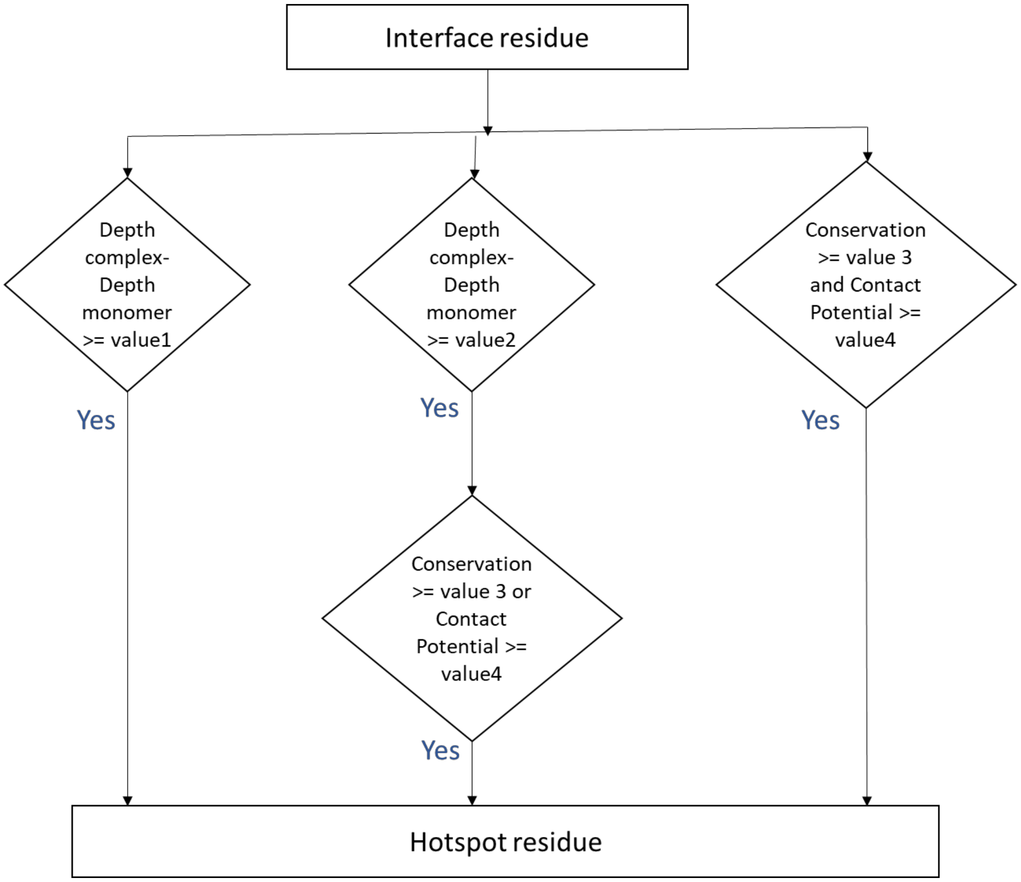

Interface residue
Depth complex-Depth monomer >= value1
Depth complex-Depth monomer >= value2
Conservation >= value 3 and Contact Potential >= value4
Yes
Yes
Yes
Conservation >= value 3 or Contact Potential >= value4
Yes
Hotspot residue

Figure 2 – Flowchart showing the prediction of the hotspot residues using various combinations of values of depth change on complex formation, conservation of the residue and contact potential of the residue.

The depth change on complex formation, conservation of residues and the contact potentials of interface residue were calculated as mentioned earlier in section 2.1. The hotspot residues were predicted depending on the flowchart (Figure 2). For a residue to be predicted as a hotspot it has to fulfil at least one of the following condition – (a) large changes in depth upon complex formation (>=value1) (b) If the depth change is not large (>=value2), then it either has to be conserved (>=value3) or have a high contact potential (>=value4) (c) the residue needs to be both conserved (>=value3) and have a high contact potential (>=value4). If any residue fails to satisfy at least one of the criterion, the residue is labelled as non-hotspot residue.

A grid search was done over different values of value 1, value 2, value 3 and value 4 (Table 1) to obtain the value, which gives the maximum MCC on the training set.

Table 1 – The range and interval over which the grid search was done for obtaining the optimal value for each of the variables

Variable	**Range**	**Interval**
Value1	3.0-6.0 Å	1.0
Value2	0.8-5.0 Å	0.2
Value3	0.1-0.4	0.05
Value4	5.0-11.0	0.5

3. Results

3.1. Comparisons of parameters between hotspot and non-hotspot residues

We compared the properties of hotspot and non-hotspot residues- a) depth change on complex formation b) contact potential c) conservation on the training set. Wilcox signed rank was carried out to identify if these properties were different between the hotspot and non-hotspot residues as observed in the training set. The hotspot residues are more buried as compared to non-hotspot residues (Figure 3 A) with a p-value of 0.0004. The hotspot residues are more conserved as compared to non-hotspot residues (Figure 3 B) with a p-value of 0.007. The hotspot residues have a higher contact potential compared to that of non-hotspot residues (Figure 3 C) with a p-value of 0.017. The correlation between depth of the residue in complex and conservation between the different datasets is 0.17+/-0.04.

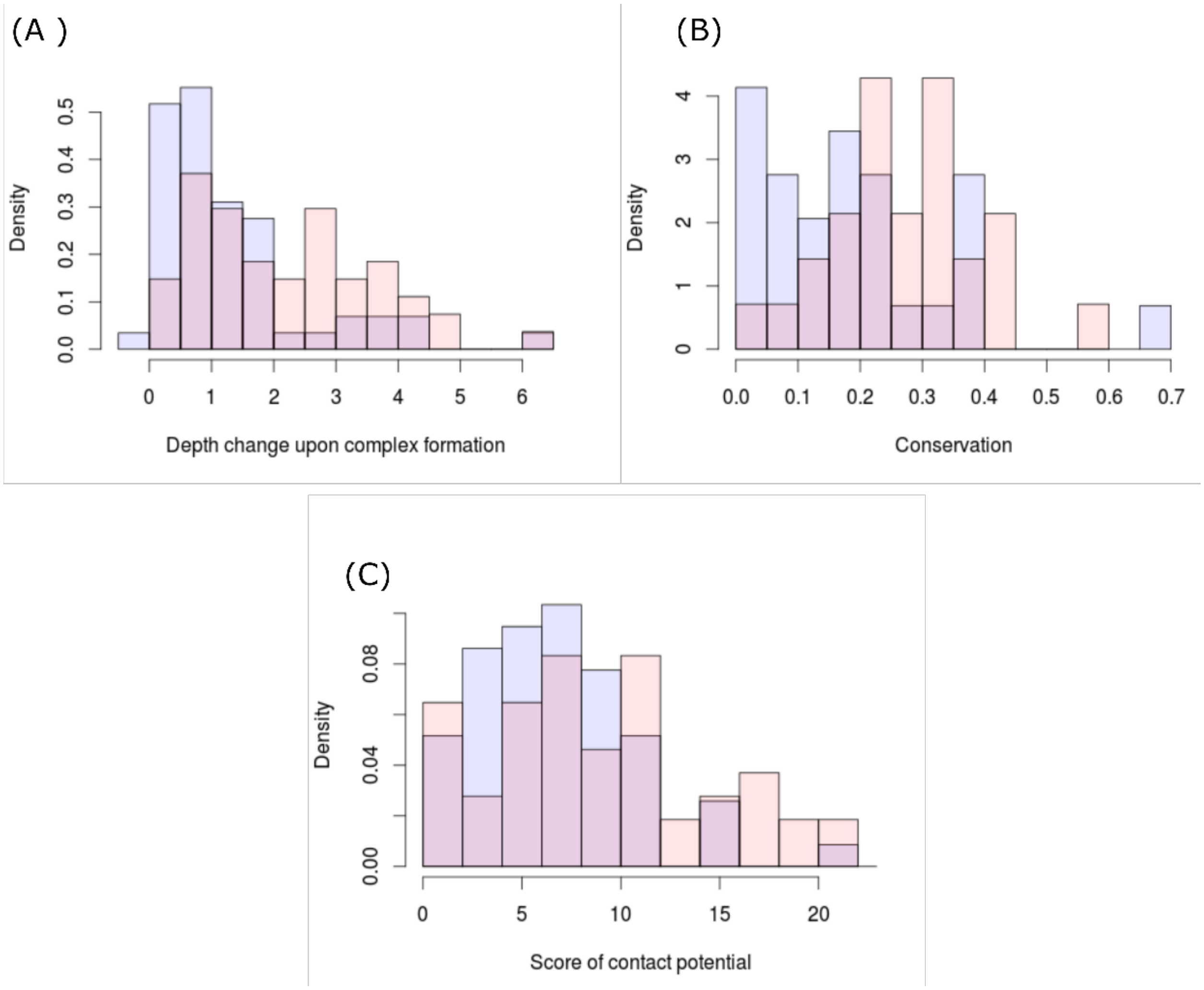


Figure 3 – Histogram showing the (A) depth change upon complex formation (Å) (B) Conservation of residues (C) Score of contact potential between hotspot residues (pink) and non-hotspot residues (blue)

3.2. Prediction of hotspot residues

A grid search across value 1, value 2, value 3, value 4 on the training set gave a highest MCC of 0.46 for DepthCon when value 1 (larger depth change) is 3, value 2 (smaller depth change) is 1.2, value 3 (conservation) is 0.2 and value 4 (contact potential) is 6.5. The MCC across the various test sets ranges between 0.44-0.48. The technique was compared to various hotspot prediction tools - Hotpoint, KFC2a and KFC2b, PredHS-SVM and PredHS-Ensemble. The MCC of Hotpoint, KFC2a, KFC2b, PredHS-SVM and PredHS-Ensemble ranges between 0.28-0.39, 0.28-0.41, 0.34-0.47, 0.29-0.57 and 0.26-0.60 respectively (Table 2). Our technique outperformed the average MCC and f1 for Hotpoint, KFC2a and KFC2b (Table 3). However, when compared to PredHS-SVM and PredHS-Ensemble, the average MCC and f1 for our technique DepthCon were either at par or a bit lower (Table 3). However, the standard deviation of MCC across the different sets was higher (0.13 standard deviation for MCC) for PredHS-Ensemble and PredHS-SVM as compared to DepthCon (0.01 standard deviation for MCC). This is because PredHS-Ensemble and PredHS-SVM had larger variations in MCC across different datasets.

Table 2 – Sensitivity, specificity, precision, accuracy, f1 and MCC of the technique across training and testing sets. The highest value for each column in each set is highlighted.

Technique	Sensitivity	Specificity	Precision	Accuracy	f1	MCC
Training set						
Hotpoint	59	79	73	70	0.65	0.39
KFC2a	59	71	65	65	0.62	0.30
KFC2b	46	**87**	**78**	68	0.58	0.38
PredHS-SVM	59	71	65	65	0.62	0.30
PredHS-Ensemble	**79**	45	57	61	0.67	0.26
DepthCon	74	72	71	**73**	**0.73**	**0.46**
Test Set 1						
Hotpoint	52	81	64	70	0.57	0.35
KFC2a	62	77	63	71	0.63	0.39

KFC2b	43	88	69	70	0.53	0.35
PredHS-SVM	66	**89**	**79**	**80**	0.72	0.57
PredHS-Ensemble	**90**	71	67	78	**0.76**	**0.60**
DepthCon	88	60	59	71	0.70	0.48
Test set 2						
Hotpoint	52	77	48	69	0.50	0.28
KFC2a	62	77	63	71	0.63	0.39
KFC2b	43	88	69	70	0.53	0.35
PredHS-SVM	66	**89**	**72**	**82**	0.68	0.56
PredHS-Ensemble	**90**	75	60	79	**0.72**	**0.59**
DepthCon	88	62	49	69	0.63	0.46
Test set 3						
Hotpoint	52	79	52	70	0.52	0.31
KFC2a	56	73	48	68	0.52	0.28
KFC2b	48	84	**57**	**73**	0.52	0.34
PredHS-SVM	40	**86**	56	72	0.47	0.29
PredHS-Ensemble	80	64	50	69	0.62	0.41
DepthCon	**88**	59	49	68	**0.63**	**0.44**
Test set 4						
Hotpoint	59	74	50	69	0.53	0.31
KFC2a	74	74	56	74	0.64	0.41
KFC2b	59	87	68	79	0.63	0.47
PredHS-SVM	59	**93**	**79**	**83**	**0.68**	**0.57**
PredHS-Ensemble	74	80	63	79	**0.68**	0.53
DepthCon	**92**	54	47	66	0.63	0.44

Table 3 – Average and standard deviation of f1 and MCC of the different hotspot prediction tools across 1 training set and 4 testing sets. The least standard deviations have been highlighted

Technique	**f1**	**MCC**
Hotpoint	0.55+/-0.05	0.33+/-0.04
KFC2a	0.61+/-**0.04**	0.35+/-0.05
KFC2b	0.56+/-**0.04**	0.38+/-0.05
PredHS-SVM	0.63+/-0.09	0.46+/-0.13
PredHS-Ensemble	0.69+/-0.05	0.48+/-0.13
DepthCon	0.66+/-**0.04**	0.46+/-**0.01**

In order to find the importance of each of the parameters [(A) depth change>=value1 (B) depth change>=value2 and either conservation>=value3 or contact potential>=value4 (C) conservation>=value3 and contact potential>=value4], we neglected the condition A, B or C, one at a time and using the predetermined values (value 1-4) the MCCs for the different datasets calculated (Table 4). All the datasets, except Test set 2 had a lower MCC as compared to DepthCon when any of the conditions of DepthCon was neglected. Test Set 2, however, had a higher MCC when condition B was neglected. In this dataset condition B increased the number of false positive predictions by 22, whereas the true positive predictions increase by 9. Neglecting condition A lead to a maximum reduction in MCC for Test 3 by 0.05. Neglecting condition B lead to the maximum reduction in MCC for Training set, Test Set 3 and Test Set 4 by 0.16, 0.05 and 0.07 respectively. Neglecting condition C lead to the maximum reduction in MCC for Test Set 2 by 0.07. Hence across different datasets, a cumulative output by all the three conditions (A, B, C) leads to uniform MCC.

Table 4 – MCC on the training and different testing sets when a condition of DepthCon is neglected. The last row represents the MCC of DepthCon with all conditions.

Condition neglected	**Training set**	**Test set 1**	**Test set 2**	**Test set 3**	**Test set 4**
depth change>=value1	0.43	0.48	0.45	0.39	0.42

depth change>=value2 and either conservation>=value3 or contact potential>=value	0.30	0.48	0.48	0.39	0.37
conservation>=value3 and contact potential>=value	0.41	0.41	0.42	0.44	0.33
None	0.46	0.48	0.46	0.44	0.44

4. Discussions

All residues at the interface do not contribute equally to the binding affinity between the interacting proteins. The residues that contribute predominantly to binding i.e. the hotspot residues serve as important targets to modulate interaction affinities between proteins. Experimental prediction of protein hotspot is expensive and time consuming. Hence computational techniques can assist in the prediction of hotspot residues. The available machine learning tools, although promising, fail to replicate their accuracies, MCC, f1 etc. in different datasets.

Here we present an empirically created decision tree based technique DepthCon. These uses input features - depth change upon complexation, conservation of the residue and residue pairwise contact potential in various combinations to predict hotspot residues. According to our technique, a residue is predicted as a hotspot residue if either of the following holds - a) the residue undergoes a depth change of >3 Å upon complex formation b) the residue undergoes a depth change of >1.2 Å and has a conservation score of >0.2 or a contact potential of >6.5 c) the residue has a conservation score of >0.2 and contact potential of >6.5.

Though we do not outperform the machine learning based prediction tools in all the testing datasets, our prediction scheme maintains a stable MCC of 0.46+/-0.01 and f1 of 0.66+/-0.04. It should be noted that certain tools (KFC2a, KFC2b and Hotpoint) had lower MCC (<=0.38) and f1 (<=0.61) when compared to our method, whereas other tools showed large variations in their MCC (0.13 for PredHS-SVM and PredHS-Ensemble) and f1 values (0.05 for PredHS-SVM and 0.09 for PredHS-Ensemble). Hence the prediction accuracy of PredHS-SVM and PredHS-Ensemble depended on the dataset being used to make the prediction. This may be either because the feature set used to train the tools was not able to capture the important features for the prediction or because of overfitting of data to the training set.

Although, DepthCon has low specificity (54-71%) and precision (47-71%), when compared to other methods, the sensitivity of our technique is consistently high (74-92%). It should be noted that DepthCon over predicts the number of hotspot residues, thereby giving a large number of false positives. We are currently unable to pinpoint, which parameters (depth, conservation and contact potentials) lead to over predictions. Across the different datasets, the false positives did not seem to have consistent depth, conservation, contact potential values or residue type. Neglecting any of the conditions of DepthCon reduced the MCC values in almost all cases (except 1). Hence, all the conditions are important in maintaining robust MCC across different datasets.

The addition of parameters like hydrogen bonding might not help in improving the prediction scheme. The number of hotspot residues involved in side-chain hydrogen bonds with the partner chain (in the training set) is 7 (out of 54 residues), while the number of non-hotspot residues having side-chain hydrogen bonds is 6 (out of 58 residues). Hence, this might lead to increased false positive predictions without a significant increase in true positive predictions.

To conclude, our study shows that a simple combination of pairwise contact potential, depth change upon complex formation and conservation of residues can be reliably and effectively used to distinguish hotspot residues from non-hotspot residues. Further developments in the scoring scheme would involve figuring out ways to reduce the number of false positives without reducing true positive predictions.

Along with studying protein-protein interfaces we also studied protein-small molecule interfaces which have been dealt in the next two chapters.

Chapter 7 - Structural study of protein-small molecule interfaces: prediction of off target effects of the drug Nutlin

1. **Predicting proteins having similar binding pocket as Nutlin binding pocket of Mdm2**
2. **Computational prediction of the stability of the bound complexes**
3. **Docking of Nutlin onto the predicted binding site using Autodock Vina**
4. **Comparison of our technique to others**

The search for binding site similar to that of Nutlin was done by Dr. Minh N. Nguyen. The binding free energy of the predicted complexes was calculated by Lin Meiyen with help from Thomas Leonard Joseph. The ingenuity pathway, KEGG and Reactome analysis were carried out by Candida Vaz and Vivek Tanavde. The experimental validation was done by Luke Way. The analysis of the predicted binding site, molecular dynamics simulations and the docking of Nutlin onto the predicted binding site were carried out by Neeladri Sen. The comparisons of our method to other preexisting method was done by Neeladri Sen and Minh N. Nguyen.

Published - Nguyen M.N.*, **Sen N.***, Meiyin L.*, Joseph T.L., Vaz C., Tanavde V., Way L., Hupp T., Verma C. and Madhusudhan M.S. (2019), Discovering putative protein targets of small molecules: A study of the p53 activator Nutlin**,** Journal of Chemical Information and Modeling, doi: 10.1021/acs.jcim.8b00762 [* equal contributions]

1. Introduction

The drugs bind selectively to a pocket on the targets because of complementarity in shape and physicochemical properties [243]. However, the diversity of protein shapes is limited and similar binding pockets could likely be found in other proteins [243,244]. The binding of the drug to these off-target proteins could either lead to adverse drug reactions [245] or indicate an alternate use of the drug [246]. A relevant observation to support this claim would be that of protein kinases, which have structurally similar ATP binding pockets [247]. A drug designed against the ATP binding pocket of one kinase often also binds to other kinases [248], making the drug less specific. An efficient drug discovery effort would be strengthened with the identification of putative binding pockets on off-target proteins. Current drug discovery efforts usually identify one or few protein target(s) and attempt to inhibit these using small molecules/peptides. Most inhibitors/drugs are identified using various computational/experimental screens followed by rounds of rational manipulation and extensive experimental validation [249].

Our tool CLICK [18] (Refer to Chapter 2 Section 3) can compare the 3D structures or even sub-structures of molecules. CLICK is capable of aligning structures with dissimilar topologies, conformations, or even molecular types. These unique properties make CLICK particularly well suited for comparing protein substructures, such as ligand binding sites [18]. Though we can use CLICK to compare binding site for any ligand, however, in this study, we tested the efficacy of CLICK in identifying similar binding pockets of the small molecule Nutlin.

Nutlin is known to bind MDM2, a negative regulator of the tumor suppressor protein p53 [250,251]. Upregulated activities of MDM2 in several cancers result in increased degradation of p53 and hence is being pursued as a potential therapeutic target [252]. The interactions between MDM2 and Nutlin have been explored using several experimental techniques including crystallography (PDB ID: 1RV1/4J3E) [251,253,254], which show that Nutlin occupies a hydrophobic pocket in the N-terminal domain of MDM2 and mimics key residues of the N-terminal region of p53 which occupy the same pocket (PDB ID: 1YCR) [255] (Figure 1A). In cancer cells with upregulated levels of MDM2 [256], the abolition of the interactions between MDM2 and p53 by peptides or small molecules such as Nutlin is demonstrably sufficient to induce activation of p53[251]. Thus, Nutlin-like molecules could be potential drugs for such cancers; several of which are currently in clinical trials [257]. However, there is some evidence of toxicity by this approach [258] and hence there is a need to develop robust methods to pre-screen potential drugs for potential adverse reactions.

The current study provides a tentative list of proteins that may be targets of Nutlin in addition to MDM2. If validated, this technique has the potential of adding specificity filters in drug design for detecting off target effects of small molecule compounds resulting in cost savings. We have used the program CLICK to predict potential protein targets of Nutlin other than MDM2. Further, we computed the binding free energy of Nutlin with these proteins using the Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) protocol of Amber11. We also compared our binding poses with the poses predicted by docking Nutlin onto the CLICK predicted binding pocket using AutoDock Vina. Molecular dynamics simulations were carried out on 4 of the putative proteins complexed to Nutlin to probe the stability of these complexes. Finally, thermal shift assay was done to validate the binding of Nutlin to one of the predicted off target proteins. Though the pilot study involved predicting binding pockets of Nutlin in off target proteins, this methodology can be used for predicting binding pocket of any small molecule ligand based on structural similarity.

2. Methods

2.1. Nutlin binding sites on MDM2

The binding site residues within 6 Å of Nutlin-2 (Figure 2A) and Nutlin-3a (Figure 2B) were extracted from the structures of their complexes with MDM2 (PDB- 1RV1:B for Nutlin-2 and PDB- 4J3E:A for Nutlin-3a). Henceforth, Nutlin will be used to refer to both Nutlin-2 and Nutlin-3a.

To account for binding site flexibility, Nutlin binding site residues were also extracted from 5 snapshots from a molecular dynamics (MD) simulation trajectory of MDM2, at 3, 6, 9, 12 and 15ns [253].

2.2. Dataset of representative protein structures

To search for putative non-MDM2 proteins that could bind Nutlin, 4239 crystal structures of proteins were selected from the PDB using the program PISCES [139], such that the proteins a) were all human proteins, b) were resolved at resolutions higher than 3 Å, c) had R-factor<0.3, d) were not more than 95% sequentially identical to one another, and e) had a length greater than 40 residues. The complete list of human proteins used for the study can be found at - http://cospi.iiserpune.ac.in/click/Download/Human.txt.

2.3. CLICK searching

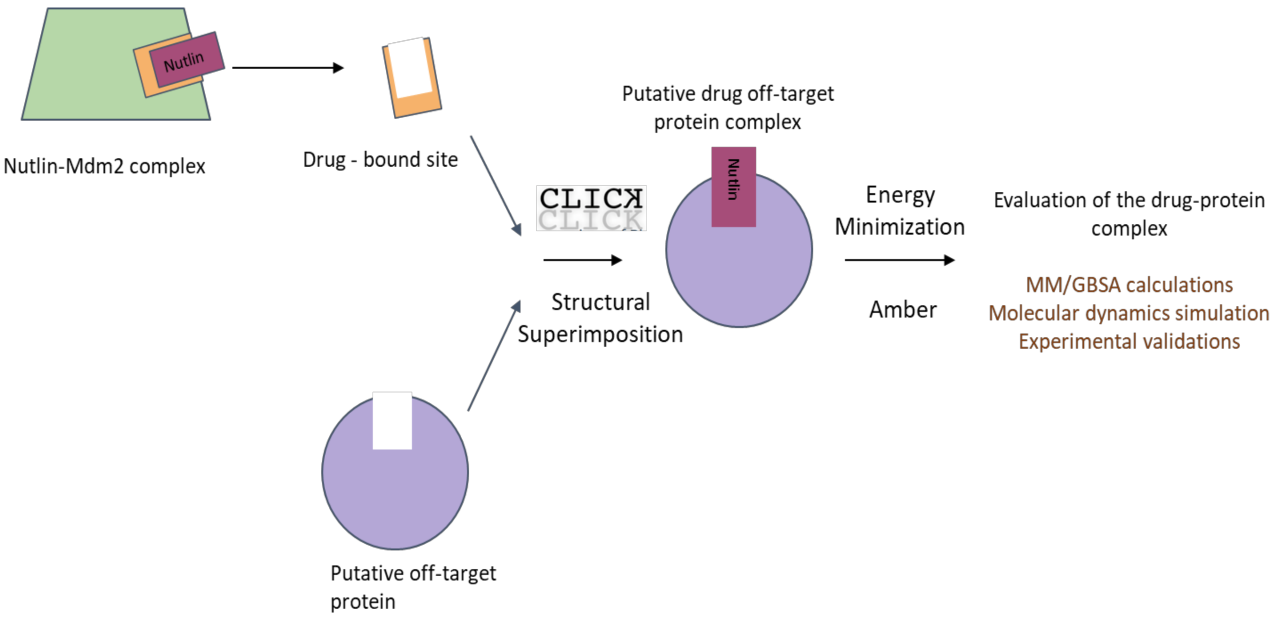

Nutlin
Nutlin-Mdm2 complex
Drug - bound site
CLICK
Structural Superimposition
Putative drug off-target protein complex
Nutlin
Energy Minimization
Amber
Evaluation of the drug-protein complex
MM/GBSA calculations
Molecular dynamics simulation
Experimental validations
Putative off-target protein

Figure 1 – Schematic showing the prediction of off target binding site for Nutlin. The drug bound protein (Mdm2) shown is shown in green with the drug molecule (Nutlin) shown in dark purple and the binding site with orange. The binding site is searched on another protein (shown in light purple) using the structural superimposition tool CLICK. The model of the drug-putative off target is built, which is energy minimized and evaluated.

The Nutlin binding site(s) were structurally superimposed on the entire proteins in our dataset using CLICK (*http://cospi.iiserpune.ac.in/click*) (Figure 1). Our CLICK program superimposes two molecular structures, even if they are topologically dissimilar, by a 3D least-squares fit of their representative atoms. In this case, the C^{α} and C^{β} atoms of the residues were chosen as representative atoms for structural superimposition. A clique of points is made with the representative atoms such that no pair of atoms within a clique is separated by more than a distance threshold of 10 Å as earlier optimized [23]. The clique size was earlier optimized to contain 7 residues [23]. To ensure that equivalent residues occupy similar environments in their respective proteins a match was only made if the residue depth difference was less than 2.25 Å. Residue depth is defined as the closest distance of the residue from the bulk solvent[7]. The CLICK program produces a Z-score for the reliability of match and we had previously established that a score of 2 and above was indicative of a significant comparison. For each of the human protein, the region of the protein having the highest structural overlap with the Nutlin binding site (such that the residue depth difference was less than 2.25 Å) is produced as an output by CLICK. The CLICK program creates small cliques of points (3-7 in number) from representative atoms (user defined criterion, C^{α} atoms or combination of C^{α}, C^{β} atoms etc.) of spatially proximal amino acid residues. These cliques are then superimposed by a 3D least-squares fit.

The predicted binding site was further processed as mentioned in the sections below. The objective of these comparisons was to match regions on proteins that structurally resembled the Nutlin binding pocket on MDM2.

2.4. Eliminating hits clashing with Nutlin

In our protocol, we superimposed the proteins from the database (section 2.2) onto the Nutlin binding sites. Proteins that had regions that matched the Nutlin binding sites with significant Z-scores were termed as ‘hits’. Both Nutlin-2 and Nutlin-3a were then independently transferred as rigid bodies onto the hit protein to form a complex. The complex was energy minimized using Amber11 [259]. Steric hindrances in the complexes (with either of the Nutlins) was quantified by a clash score. A clash results when the intermolecular atomic distance between two non-hydrogen atoms of the hit and Nutlin is less than 2.0 Å. Ideally, we would want no short contacts between the atoms of the protein and the ligand. However, we tolerated a few short contacts, empirically set to 5 short contacts involving the protein side chains and 1 short contact involving the protein main chain. Our tolerance levels were decided upon following the logic that short contacts with the side chain could be more easily resolved (moving individual side chains) as opposed to making conformational changes to the main chain.

2.5. Validation by scoring the poses of Nutlin

2.5.1. Single point binding energy calculations and hydrophobicity of the binding pocket

A single point binding energy of the observed/predicted complex was computed using the Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) method with the GB module of Amber11 (Figure 1). The binding free energy of Nutlin was also calculated on the human analogs of MDM2, Hdm2 and MDM4, which are structurally similar to MDM2. The binding energy, as computed here, is essentially the Enthalpy change (Δ*H)* as a result of binding. The more negative Δ*H* is, the tighter Nutlin binds to the protein target. The shortlisted hits were rank ordered by binding energies (Table 1) calculated using Amber11.

2.5.2 Docking of Nutlin-3a onto the target protein using AutoDock Vina

To validate the binding site and binding pose, AutoDock Vina [54] was used to dock Nutlin-3a onto the energy minimized structures of the target proteins obtained from Section2.4. The AutoDock exercise was not carried out with the crystal structures of these proteins as in many of the cases the binding pocket could not accommodate the ligand before conformational changes. Polar hydrogens were added and charges were assigned to atoms of both the target protein and Nutlin-3a. A 30*30*30 Å^3 box (dimensions chosen considering the size of Nutlin) for docking of Nutlin-3a was centered on its N1 atom from the CLICK predicted structure. The binding free energy and the corresponding RMSD (calculated between the central 5 membered ring atoms – N, N1, C10, C11, C18) to the CLICK computed binding pose were calculated for the best AutoDock pose (Table 1).

2.6. Identifying functions/pathways of putative Nutlin targets

The UniProt (Universal Protein Resource, http://www.uniprot.org) database was mined for information relating to the biological role and function of the hit protein, its interactions with other proteins, binding sites, and post-translational modifications [260]. The Ingenuity Pathway Analysis (IPA) software [261], KEGG (*http://www.kegg.jp*), and Reactome (*http://www.reactome.org*) were used to identify the pathways that the hit proteins were involved in.

2.7. Molecular dynamics (MD) simulations of Nutlin targets

MD simulations were carried out on a subset of the proteins found to bind Nutlin, namely Gamma-glutamyl hydrolase (GGH) (PDB ID: 1L9X:A), steryl sulfatase (PDB ID: 1P49:A) and interferon-gamma (PDBID: 1FYH:A) and Human dead box RNA helicase (PDB ID: 3DKP:A). The rationale for choosing only these systems for MD simulations studies is mentioned in the results section. Simulations (for all systems) were carried out in triplicate on the unliganded native structures of the proteins as well as on their modeled complexes with Nutlin-3a. The simulations were carried out using Gromacs [262,263] with the Amber99SB-ILDN force field [264] using spc/e water model [265]. Parameters for Nutlin-3a were obtained using antechamber [59,266]. Each system was solvated in a cubic water box whose sides were at a minimum distance of 25 Å from any protein atom. Charge neutrality was achieved by adding sodium or chloride counterions. The particle mesh Ewald sum method was used for treating electrostatic interactions, LINCS [267] was used to constrain the hydrogen bond lengths, enabling a time step of 2fs. Initially, the whole system was minimized for 5000 steps or until the maximum force was < 1000kJ/mol/nm. The system was then heated to 300K in an NVT ensemble simulation for 100ps using a Berendsen thermostat [65]. The system was subsequently equilibrated in an NPT ensemble simulation for 100ps to stabilize the pressure using a Parrinello-Rahman barostat [268]. Finally, each system was simulated for a maximum of 100ns and structural snapshots were captured every 10ps. Simulations were stopped when the distance between Nutlin-3a and the geometric center of the protein increased by 10 Å compared to the starting structure. The 10 Å distance was empirically chosen as indicative of the ligand irreversibly leaving the binding pocket. The temperature, potential energy and kinetic energies were monitored during the simulation to check for anomalies.

3. Results

3.1. Nutlin binding pocket

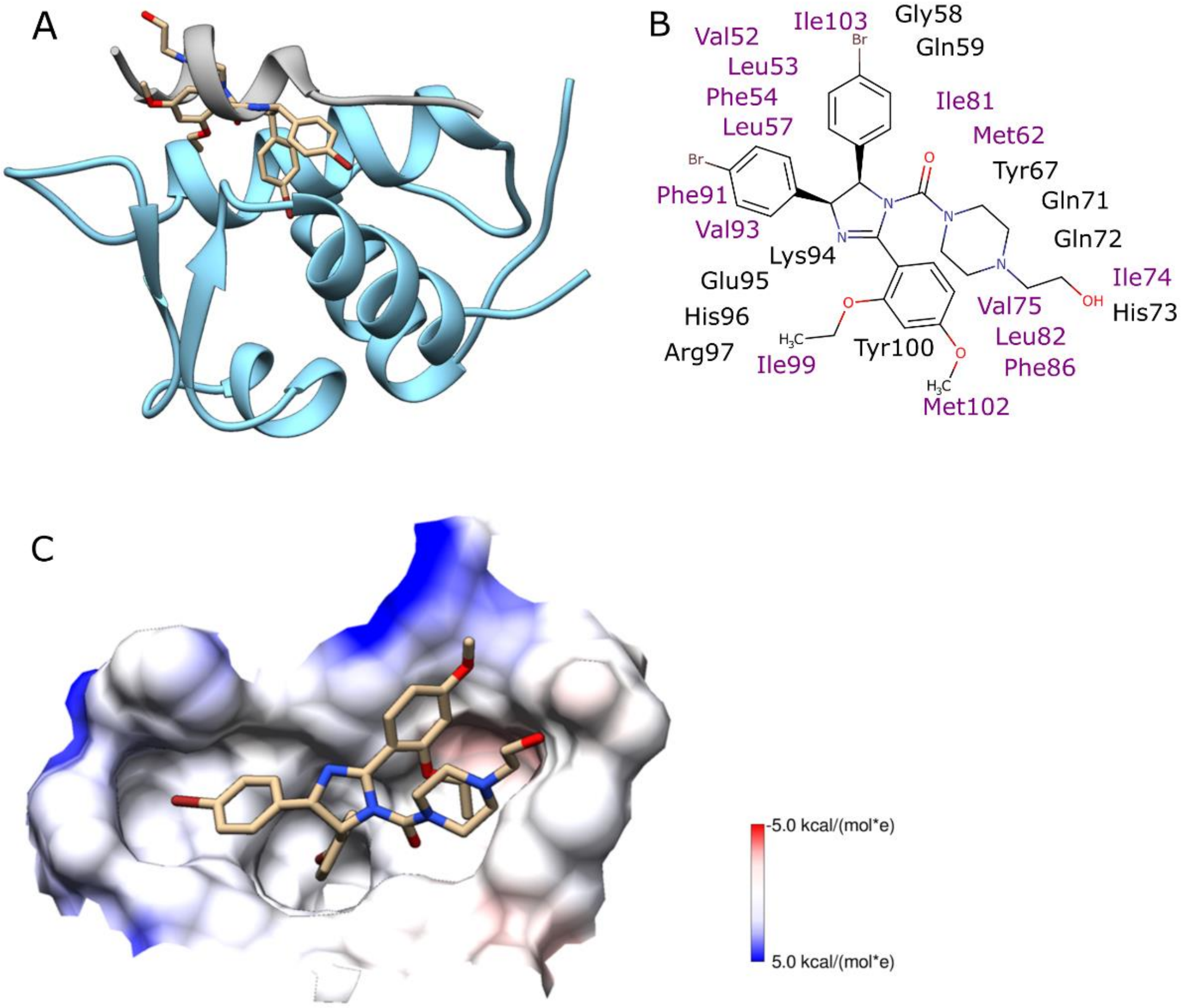


Figure 2- A) The superimposition of the crystal structure complexes of MDM2 (blue ribbons) with Nutlin-2 (tan sticks, PDB ID: 1rv1) and p53 (grey ribbons, PDB ID: 1ycr) B) 2D representations of the interactions, within 6 Å, of Nutlin 2 with residues of MDM2 from

the crystal structures and MD snapshots. Hydrophobic residues are colored in purple. C) A surface representation of the binding pocket residues shown in [B] along with the bound Nutlin (tan sticks). The binding pocket is colored as per the Chimera[41] rendered coulombic charge representation, where shades of blue and red represent positively and negatively charged regions respectively.

We have studied two variants of Nutlin – Nutlin-2 and Nutlin-3 (Figure 3). Both variants are similarly structured with a central 5 membered imidazole ring, 3 of whose atoms are connected to 6 membered aromatic rings. Of these 3 aromatic rings, 2 are halogenphenyls (bromophenyl in Nutlin-2 and chlorophenyl in Nutlin-3). The other aromatic ring of Nutlin-2 being ethoxy methoxy phenyl whereas Nutlin-3 being methoxy-2-(propan-2-yloxy) phenyl group. Another atom of the central imidazole ring is connected to a pipirazine ring. The pipirazine group in Nutlin-2 is hydroxyl ethyl piprizine whereas in Nutlin-3 it is piprazin-2-one. Nutlin-3 exists in 2 enantiomeric form Nutin-3a and Nutlin-3b where Nutlin-3b is 150 fold less potent inhibitor of MDM2 than Nutlin-3a [269].

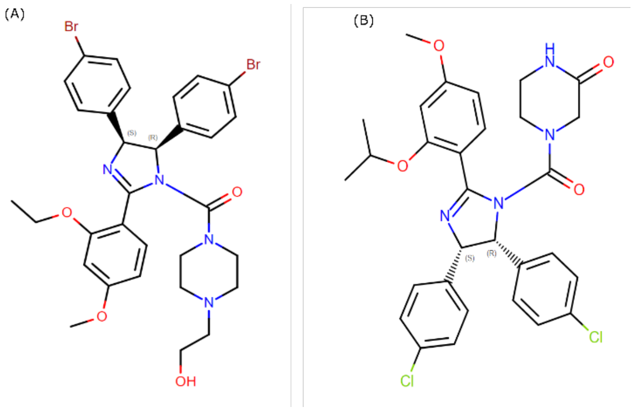


Figure 3 – Structure of (A) Nutlin-2 (B) Nutlin-3a

The binding sites of Nutlin-2 and Nutlin-3a on MDM2 are almost identical and predominantly hydrophobic. In the crystal structures of Nutlin-2 and Nutlin-3a bound to MDM2 (PDB codes 1rv1 and 4j3e respectively), there are 24 and 25 residues within a distance of 6 Å from Nutlin-2 and Nutlin-3a, respectively (Figure 2B). Though the side chains are important in receptor ligand interactions, we only considered C^{α} and C^{β} atoms of the residues to constitute our binding site descriptor, to get a description of the binding pocket and an approximate orientation of the side chains. Atoms in the side chains, especially in the solvent exposed regions are flexible and in their apo-structures may not be positioned appropriately for ligand binding. In order to account for the dynamics of the protein, snapshots from MD trajectories of Nutlin bound MDM2 were also considered[253]. The number of binding site residues from the MD snapshots varies between 23 to 25 but retains their predominantly hydrophobic characteristic (Table 1) with 18 of these residues being hydrophobic (Figure 2C). While there are a few polar and charged amino acids in the pocket, their side chains are often pointed away from Nutlin. Sometimes, such as in the MD snapshot after 3ns, the Nutlin binds deeper inside the cavity and hydrogen bonds with the side-chain of Gln72 of MDM2.

3.2. Identification of putative Nutlin binding proteins

We used our CLICK program to identify proteins from amongst a set of 4239 human proteins that had regions structurally similar to the Nutlin binding site of MDM2. Structural overlap of 70% or above using C^{α} and C^{β} and a Z-score of 2 or greater were empirically chosen thresholds (previously optimized, unpublished data) to determine meaningful matches (Table 1). For each of these hits, a model was constructed with the Nutlin in its new putative binding site. To begin with, the model is simply the coordinates of the Nutlin transferred onto the new hit after superimposing with the MDM2 binding site. This complex is energy minimized and the resultant structure is examined for steric clashes. Models with severe clashes (as described in the methods) are discarded. This search protocol for alternate binding partners yielded 49 human proteins (Table 1). Only 2 of the 49 hits, MDM4 and Hdm2 (52 and 96% sequence identity respectively), are homologs of MDM2. The other predicted targets of Nutlin are unrelated to MDM2.

In 16 of these 49 predicted target sites, a putative binding site residue (within 6 Å of Nutlin) is involved in protein function either as an active site residue or one that undergoes post-translational modification such as glycosylation (Table 1). The functions of these proteins are likely to be affected by binding Nutlin. The other 33 predicted target sites, while viable for ligand binding by our predictions are not close to any known functional site of the protein. While the binding of Nutlin to these sites can have indirect functional consequences, we focused only on some of the hits where the functional consequences could be directly affected. The hit proteins play a role in various biosynthetic pathways including endocytosis, protein folding, metabolism, apoptosis, signaling, cell migration, immune system, transport of ions, proteolysis etc. (Table 1).

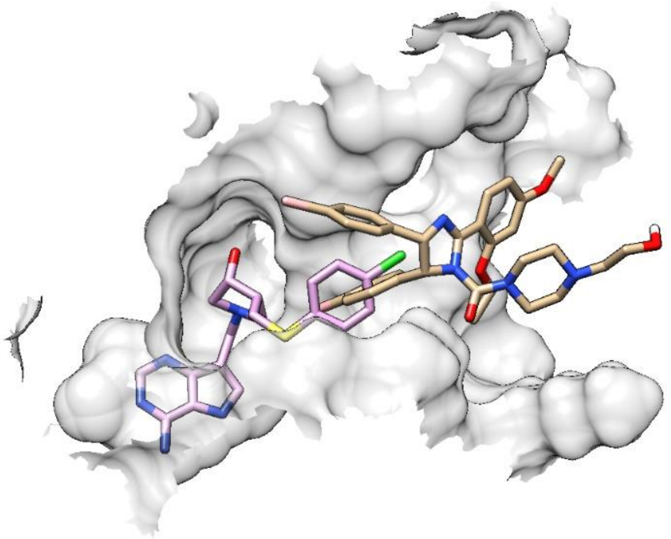


Figure 4- Nutlin-3a (shown in tan sticks) bound to the predicted binding pocket of 5'-deoxy-5'-methyllthioadenosine phoshorylase. The ligand 4CT (shown in magenta sticks) binds to a binding pocket close to the predicted binding pocket of Nutlin-3a (PDB ID: 3OZC:A). One of the benzene rings of both Nutlin-3a and 4CT superimpose in the binding pocket.

Only one of the predicted target sites, in 5'-deoxy-5'-methylthioadenosine phosphorylase (PDBID: 1CB0:A/3OZC:A), had a bound ligand pCl-phenylthioDADMeImmA (PDBID:3OZC:A). Interestingly, a part of this ligand, an aromatic ring, superimposes on one of the aromatic rings of Nutlin-3a (Figure 4).

One of the 16 proteins whose function is likely to be affected on Nutlin binding is Gamma glutamyl hydrolase. Overexpression of this protein is associated with several cancers including breast and bladder, and rheumatoid arthritis[270]. In principle, Nutlin could be repurposed to serve as a drug to combat the above diseased conditions.

*Table 1: Details of the actual and predicted human proteins that bind Nutlin. # RMSD between central imidazole ring of Nutlin-3a as predicted by CLICK (after Amber11 energy minimization) and as predicted by AutoDock Vina * Column informs if the putative binding site has been identified using the crystal structure or the MD snapshot of Nutlin (3 ns, 6 ns, 9 ns, 12 ns, 15 ns) $ The sequence identity refers to the identity between the aligned residues of the predicted off-target protein and the Nutlin-2 binding site as obtained from the crystal structure. In cases with 0 sequence identity, none of the aligned residues was identical.*

Name of actual/predicted Nutlin binding proteins	**PDB code**	**Binding energy with Nutlin-2 kcal/mol**	**Binding energy with Nutlin-3a kcal/mol**	**AutoDock Vina binding energy with Nutlin-3a kcal/mol**	**RMSD (Å) #**	**Template ***	**CLICK Z-score**	**CLICK SO (%)**	**CLICK RMSD (Å)**	**Number of residues within 6 A of Nutlin**	**Seq Identity (%) $**	**Number of aligned residues in Active/ functional sites**	**Pathway protein involved**	**Nutlin binding sites with superimposed residues in their active/functional sites**
AP-2 Complex Subunit Beta	2g30: A	-75.57	-57.08	-11.4	2.3	3ns	6.85	100.00	1.95	48	7.41	0 res	Endocrine and other factor-regulated calcium reabsorption; Endocytosis; Huntington's disease; Synaptic vesicle cycle	No
Glutataryl-CoA Dehydrogenase, mitochondrial	1siq: A	-70.65	-55.99	-11.0	0.54	15ns	6.85	97.06	1.94	53	7.41	0 res	Fatty acid metabolism; Lysine degradation; Metabolic pathways; Tryptophan metabolism	No
Gamma-glutamyl hydrolase	1l9x: A	-67.93	-57.58	-10.9	0.56	3ns	6.85	100.00	1.97	46	11.11	1 res	Folate biosynthesis	Yes

Steryl-sulfatase	1p49: A	-64.35	-56.49	-10.0	0.33	6ns	7.58	100.00	1.67	37	3.70	1 res	Steroid hormone biosynthesis	Yes
Peptidylprolyl Isomerase domain and WD Repeat Containing Protein 1	2a2n: A	-61.82	-53.03	-4.0	1.35	Crystal	2.24	85.12	2.71	48	0.00	0 res	Protein folding	No
Stromelysin-1	1hy7: A	-61.00	-63.05	-10.6	0.51	Crystal	2.24	84.30	2.77	47	3.70	8 res	Rheumatoid arthritis	Yes
MDM2 (3ns)	-	-55.31	-50.83			3ns				25				Yes
Interferon-gamma	1fyh: A	-48.92	-39.89	-8.6	5.16	Crystal	2.24	84.30	2.70	27	7.41	0 res	African trypanosomiasis; Amoebiasis; Antigen processing and presentation; Chagas disease (American trypanosomiasis); Cytokine-cytokine receptor interaction; Influenza A; Jak-STAT signaling pathway; Leishmaniasis; Malaria; Measles; Natural killer cell mediated cytotoxicity; Osteoclast differentiation; Proteasome; Regulation of autophagy; Salmonella infection; T cell receptor signaling pathway; TGF-beta signaling pathway; Toxoplasmosis; Tuberculosis; Type I diabetes mellitus	No
MDM2 (6ns)	-	-48.44	-44.93			6ns				24				Yes
Human Dead box RNA helicase DDX52	3dkp: A	-47.92	-39.60	-8.8	0.52	6ns	6.61	96.67	1.97	32	11.11	0 res	RNA helicase	No

MDM2 (crystal)	1rv1: B/ 4j3e: A	-47.21	-45.02	-8.4	0.48	Crystal	13.41	100.00	0.00	24/25	100.0			Yes
MDM2 (12ns)	-	-45.18	-43.72			12ns				24				Yes
MDM2 (15ns)	-	-44.66	-43.11			15ns				24				Yes
HDM2	2axi: A	-42.74	-41.60			Crystal	11.79	100.00	0.61	29	95.8			Yes
Hexokinase-2	2nzt: A	-39.83	-34.20	-6.7	4.71	6ns	6.61	96.67	1.91	22	7.41	0 res	Amino sugar and nucleotide sugar metabolism; Butirosin and neomycin biosynthesis; Carbohydrate digestion and absorption; Fructose and mannose metabolism; Galactose metabolism; Glycolysis / Gluconeogenesis; Insulin signaling pathway; Metabolic pathways; Starch and sucrose metabolism; Type II diabetes mellitus	No
Multiple PDZ domain protein	2o2t: A	-39.46	-38.01	-7.4	6.58	Crystal	3.69	79.17	2.19	21	7.41	0 res	Tight junction	No
MDM4	3fea: A	-38.62	-39.32	-6.1	0.73	Crystal	10.64	100.00	1.00	24	51.85			Yes
Enhancer of MRNA-Decapping protein 3	2vc8: A	-35.55	-33.67	-5.8	0.79	Crystal	3.94	87.50	2.09	22	14.81	0 res	Gene Expression; Metabolism of RNA	No

Ketohexokinase	2hlz:A	-34.48	-28.51	-7.7	6.23	Crystal	3.94	83.33	2.00	22	3.70	1 res	Fructose and mannose metabolism; Metabolic pathways	Yes
Programmed cell death protein 10	3ajm:A	-31.54	-30.85	-6.8	2.41	Crystal	6.12	87.50	1.76	23	7.41	5 res	Apoptotic pathways	Yes
Endothelial Nitric-oxide synthase	1m9m:A	-30.89	-29.16	-6.8	2.32	Crystal	5.64	83.33	1.84	23	7.41	0 res	Arginine and proline metabolism; Calcium signaling pathway; Metabolic pathways; VEGF signaling pathway	No
Glucocorticoid receptor 2	3gn8:A	-30.13	-28.72	-7.5	3.60	Crystal	5.15	87.50	2.11	26	3.70	0 res	Developmental biology; Metabolism	No
INTERLEUKIN-5	1hul:A	-30.01	-29.51	-7.0	4.67	Crystal	4.91	75.00	1.87	27	3.70	0 res	Allograft rejection; Asthma; Autoimmune thyroid disease; Cytokine-cytokine receptor interaction; Fc epsilon RI signaling pathway; Hematopoietic cell lineage; Intestinal immune network for IgA production; Jak-STAT signaling pathway; T cell receptor signaling pathway	No
CHITINASE-3 like protein 1	1hjx:A	-29.92	-26.64	-6.9	4.20	Crystal	3.69	75.00	1.94	25	7.41	0 res	Amino sugar and nucleotide sugar metabolism	No
RecQ-mediated genome instability protein 1	3mxn:A	-29.63	-26.77	-6.2	3.43	Crystal	3.94	87.50	2.08	22	3.70	0 res	Fanconi anemia pathway	No
GTP-binding protein Di-Ras1	2gf0:A	-28.64	-28.58	-6.2	2.66	Crystal	4.18	79.17	1.99	19	11.11	0 res	GTPase activity	No

Dual-specificity Phosphatase DUPD1	2y96:A	-28.09	-29.93	-6.3	2.31	Crystal	3.94	79.17	2.19	23	0.00	0 res	Dephosphorylation activity	No
Cytochrome P450 1B1	3pm0:A	-28.07	-26.67	-7.7	3.72	Crystal	5.64	83.33	1.89	23	3.70	2 res	Metabolism of xenobiotics by cytochrome P450; Steroid hormone biosynthesis; Tryptophan metabolism	Yes
Ubiquitin thioesterase ZRANB1	3zrh:A	-27.61	-26.70	-6.6	3.68	Crystal	3.94	83.33	2.11	22	3.70	0 res	Cell migration,hydrolysis of ester, thioester, amide, peptide, isopeptide	No
Leukotriene C4 synthase	3pcv:A	-27.47	-26.89	-6.8	3.29	15ns	6.85	91.18	1.59	21	0.00	1 res	Arachidonic acid metabolism; Metabolic pathways	Yes
Golgi reassembly-stacking protein 2	3rle:A	-26.99	-23.67	-6.8	3.17	Crystal	3.69	79.17	2.12	21	7.41	2 res	Assembly and golgi stacking of golgi cisternae	Yes
Phosphopantothenoyl-cysteine synthetase	1p9o:A	-26.49	-22.20	-7.0	3.26	Crystal	3.69	79.17	2.13	24	0.00	0 res	Metabolic pathways; Pantothenate and CoA biosynthesis	No
Serine/threonine-protein kinase/endoribonuclease IRE1	3p23:A	-25.84	-25.49	-7.0	3.97	Crystal	4.66	79.17	1.77	21	7.41	0 res	Alzheimer's disease; Protein processing in endoplasmic reticulum	No
Cytochrome P450 2C9	1r9o:A	-25.67	-22.20	-8.0	3.70	Crystal	4.66	79.17	1.78	24	0.00	0 res	Arachidonic acid metabolism; Drug metabolism - cytochrome P450; Linoleic acid metabolism; Metabolic pathways; Metabolism of xenobiotics by cytochrome P450; Retinol metabolism	No

C-type lectin domain family 4 member K	3p7g: A	-25.57	-24.70	-6.8	3.60	Crystal	4.42	75.00	1.69	21	0.00	1 res	Immune system	Yes
Chloride intracellular channel protein 2	2r4v: A	-24.84	-20.66	-6.4	3.78	Crystal	4.42	79.17	1.89	20	3.70	0 res	Chloride transmembrane transport, glutathione peroxidase activity	No
Lanosterol 14-alpha demethylase	3ld6: A	-24.76	-23.48	-7.1	3.25	Crystal	4.18	83.33	2.16	28	3.70	0 res	Metabolic pathways; Steroid biosynthesis	No
Heparin-binding growth factor 1	3o3q: A	-24.54	-24.07	-5.4	5.64	Crystal	3.94	79.17	2.10	24	7.41	0 res	MAPK signaling pathway; Melanoma; Pathways in cancer; Regulation of actin cytoskeleton	No
Hypothetical ubiquitin-conjugating enzyme LOC55284	2a7l: A	-24.44	-23.86	-6.8	3.22	Crystal	5.39	87.50	1.86	24	11.11	0 res	Ubiquitin mediated proteolysis	No
Protein-Glutamine Gamma-Glutamyltransferase KNo	2xzz: A	-24.12	-25.41	-6.6	2.45	Crystal	2.97	79.17	2.30	21	0.00	0 res		No
General vesicular transport factor p115	2w3c: A	-23.75	-29.05	-6.0	3.58	Crystal	4.42	75.00	1.88	20	0.00	0 res		No
Protein SET	2e50: A	-23.59	-22.18	-8.1	5.43	Crystal	2.97	70.83	1.95	22	3.70	0 res	Gene expression; Metabolism of RNA	No

Receptor tyrosine-protein kinase erbB-4	2r4b:A	-23.48	-18.18	-5.8	11.45	Crystal	4.42	75.00	1.77	15	0.00	2 res	Tyrosine kinase activity,	Yes
Glutathione Transferase Zeta	1fw1:A	-23.12	-17.80	-6.2	6.88	Crystal	3.45	79.17	2.09	21	0.00	0 res	Drug metabolism - cytochrome P450; Glutathione metabolism; Metabolic pathways; Metabolism of xenobiotics by cytochrome P450; Tyrosine metabolism	No
INTERLEUKIN-17A	2vxs:A	-22.33	-21.15	-7.0	3.34	Crystal	3.45	75.00	2.09	20	3.70	0 res	Cytokine-cytokine receptor interaction; Rheumatoid arthritis	No
Serine-pyruvate aminotransferase	3r9a:A	-21.59	-19.08	-6.7	4.63	Crystal	3.69	83.33	2.06	23	0.00	0 res	Alanine, aspartate and glutamate metabolism; Glycine, serine and threonine metabolism; Metabolic pathways; Peroxisome	No
Receptor tyrosine-protein kinase erbB-2	3pp0:A	-21.55	-21.60	-6.3	5.58	Crystal	4.42	75.00	1.75	17	0.00	1 res	Calcium signaling pathway; Endocytosis; ErbB signaling pathway	Yes
Proto-oncogene Tyrosine protein Kinase Receptor RET	2ivs:A	-21.54	-21.77	-6.6	12.51	Crystal	5.64	79.17	1.38	17	7.41	1 res	Endocytosis; Pathways in cancer; Thyroid cancer	Yes
Sulfotransferase 1C2	2gwh:A	-21.39	-20.20	-5.9	4.68	Crystal	5.15	79.17	1.83	22	7.41	1 res	Metabolism	Yes
Fibroblast growth factor	1q1u:A	-21.18	-19.79	-6.0	2.98	Crystal	2.97	70.83	1.95	22	0.00	0 res	MAPK signaling pathway; Melanoma; Pathways in cancer; Regulation of actin cytoskeleton	No

homologous factor 1														
Pigment Epithelium-Derived factor	1imv:A	-20.41	-16.13	-6.0	2.56	Crystal	3.69	75.00	1.94	20	3.70	1 res	Inhibition of angiogenesis	Yes
Activin receptor type IIB	2qlu:A	-19.77	-16.61	-6.1	4.60	Crystal	3.21	70.83	1.71	20	7.41	0 res	Cytokine-cytokine receptor interaction; TGF-beta signaling pathway	No
5'-Deoxy 5'-Methylthioadenosine phosphorylase	1cb0:A	-18.41	-24.24	-7.5	9.03	Crystal	4.42	75.00	1.82	20	0.00	2 res	Cysteine and methionine metabolism; Metabolic pathways	Yes
Alkaline phosphatase, placental type	3mk1:A	-17.82	-14.82	-5.7	6.93	Crystal	3.69	79.17	1.97	27	3.70	0 res	Folate biosynthesis; Metabolic pathways	No
Tumor necrosis factor receptor superfamily member 1B	3alq:R	-17.17	N/A	-4.6	4.56	Crystal	3.21	70.83	1.83	18	0.00	0 res	Adipocytokine signaling pathway; Amyotrophic lateral sclerosis (ALS); Cytokine-cytokine receptor interaction	No
Dual specificity mitogen-activated protein kinase kinase 4	3aln:A	-16.81	-19.05	-6.4	6.17	Crystal	3.69	75.00	1.93	19	14.81	1 res	Chagas disease (American trypanosomiasis); Epithelial cell signaling in Helicobacter pylori infection; ErbB signaling pathway; Fc epsilon RI signaling pathway; GnRH signaling pathway; HTLV-I infection; Influenza A; MAPK signaling pathway; Toll-like receptor signaling pathway	Yes

3.3. Computational measurement of Nutlin-protein complex stability

We measured the strengths of association between Nutlin and the putative hit through direct and indirect computations. We directly measured the binding free energies using single point molecular mechanics computations and using the AutoDock Vina energy function. Indirectly, we assessed the strength of the complex by subjecting it to MD simulations and determined if the association was stable.

The single point binding energies of Nutlin to its original targets, MDM2, including its MD snapshots, MDM4, and Hdm2 all lie in the range of around -55 to -39 kcal/mol. The binding energy among these targets is the lowest for the 3ns snapshot of MDM2 to Nutlin-2/Nutlin-3a (-55 and -51 kcal/mol respectively) where the Nutlin binds deep inside the cavity and hydrogen bonds to Gln72. The binding free energy for the 49 hits for Nutlin binding site ranges from -76 to -15 kcal/mol (Table 1). 8 (AP-2 Complex subunit beta, mitochondrial Glutaryl-CoA dehydrogenase, Gamma glutamyl hydrolase, Streyl sulfatase, Stromelysin-1, Interferon-gamma, Human dead box RNA helicase DDX52, Peptidylprolyl Isomerase domain and WD Repeat Containing Protein 1) of the 49 proteins appear to bind Nutlin better than the original targets (marked in table 1).

Nutlin-3a was docked onto the CLICK discovered binding pockets of the 49 putative alternate target proteins using AutoDock Vina. The RMSD to the CLICK-predicted pose and the binding energies were computed for the best bound complexes (Table 1). The best pose was the one that had the least AutoDock energy. All the AutoDock binding energies were in the range of -11.4 to -4.0 kcal/mol, indicative of favorable binding events. The binding energy of Nutlin-3a onto MDM2 was -8.4 kcal/mol with a binding pose RMSD of 0.48 Å. 7 proteins (AP-2 Complex subunit beta, mitochondrial Glutaryl-CoA dehydrogenase, Gamma glutamyl hydrolase, Streyl sulfatase, Stromelysin-1, Interferon-gamma, Human dead box RNA helicase DDX52) had better AutoDock binding energies than MDM2. All of these also had better single point energy scores than MDM2 as described above. Interestingly, both methods compute the binding energy between Nutlin and the AP-2 Complex Subunit Beta (PDB ID- 2g30: A) as the best scoring interaction.

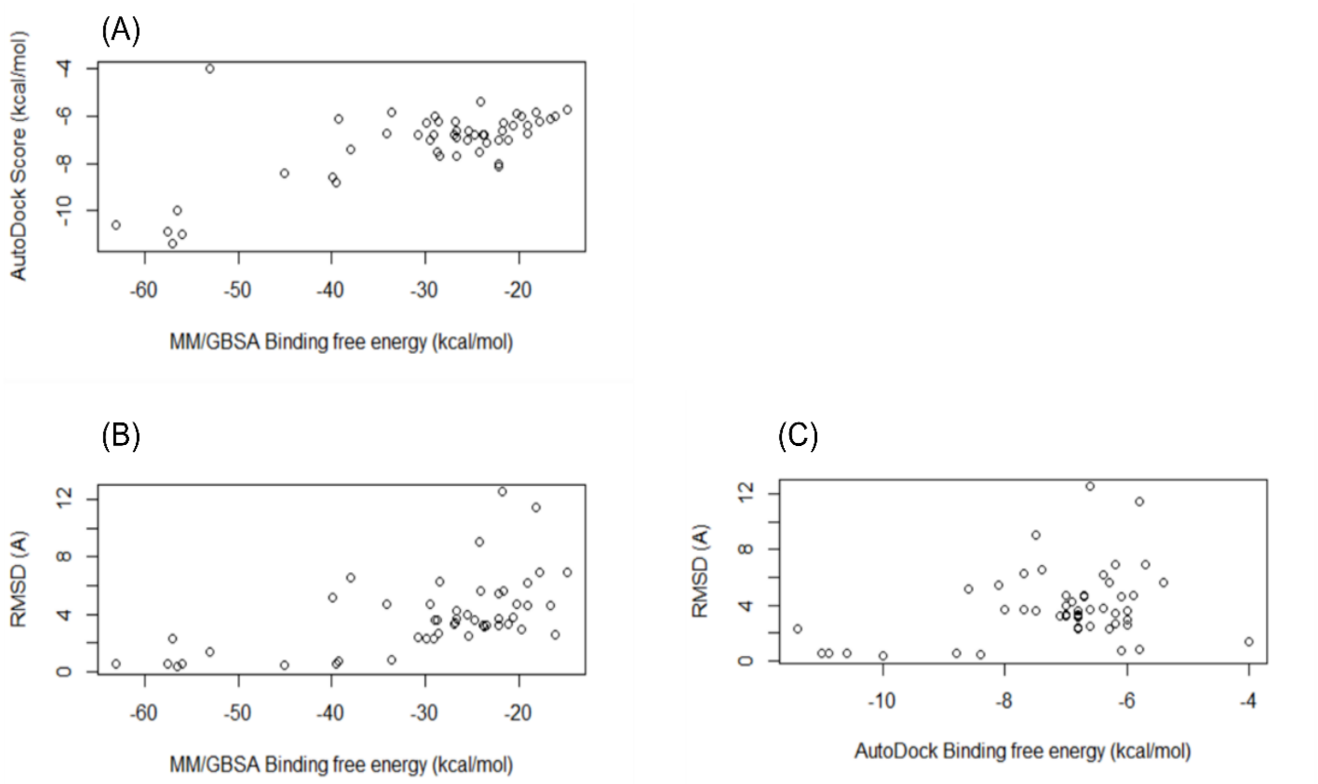


Figure 5- (A) Plot of binding energies predicted by Autodock Vina vs that predicted by MM/GBSA (Correlation coefficient 0.69). (B) Binding free energy calculated by MM/GBSA vs the RMSD between the central 5 membered ring as predicted by CLICK (after energy minimization) and the best pose as predicted by Autodock Vina (Correlation coefficient 0.57) (C) Binding free energy calculated by Autodock Vina vs the RMSD between the central 5 membered ring as predicted by CLICK (after energy minimization) and the best pose as predicted by Autodock Vina (Correlation coefficient 0.34)

Overall, the single point energy scores show a similar trend as the AutoDock scores. Protein-Nutlin complexes that score well with one measure also do so with the other. The correlation between the single point scores and the AutoDock scores was 0.69 (Figure 5A). We also compared the AutoDock poses to the CLICK-predicted poses. Here again, the trends show that the larger the deviation (higher RMSD) from the CLICK pose, the less favorable is the energy (Table 1, Figure 5B, 5C).

To test the binding mode of Nutlin-3a on one of the target proteins, Gamma glutamy hydrolase, Nutlin-3a was docked onto its crystal structure (PDB ID: 1L9X:A) using Autodock Vina. The ligand was allowed to search the conformational space around the

CLICK predicted binding site. The residues within 6 Å of Nutlin-3a, as predicted after energy minimization of the CLICK predicted complex was chosen as flexible during the docking procedure. Autodock Vina predicted Nutlin-3a to bind outside the CLICK predicted binding site, on the surface of the protein (Figure 6A) with predicted binding energy of -7.1 kcal/mol.

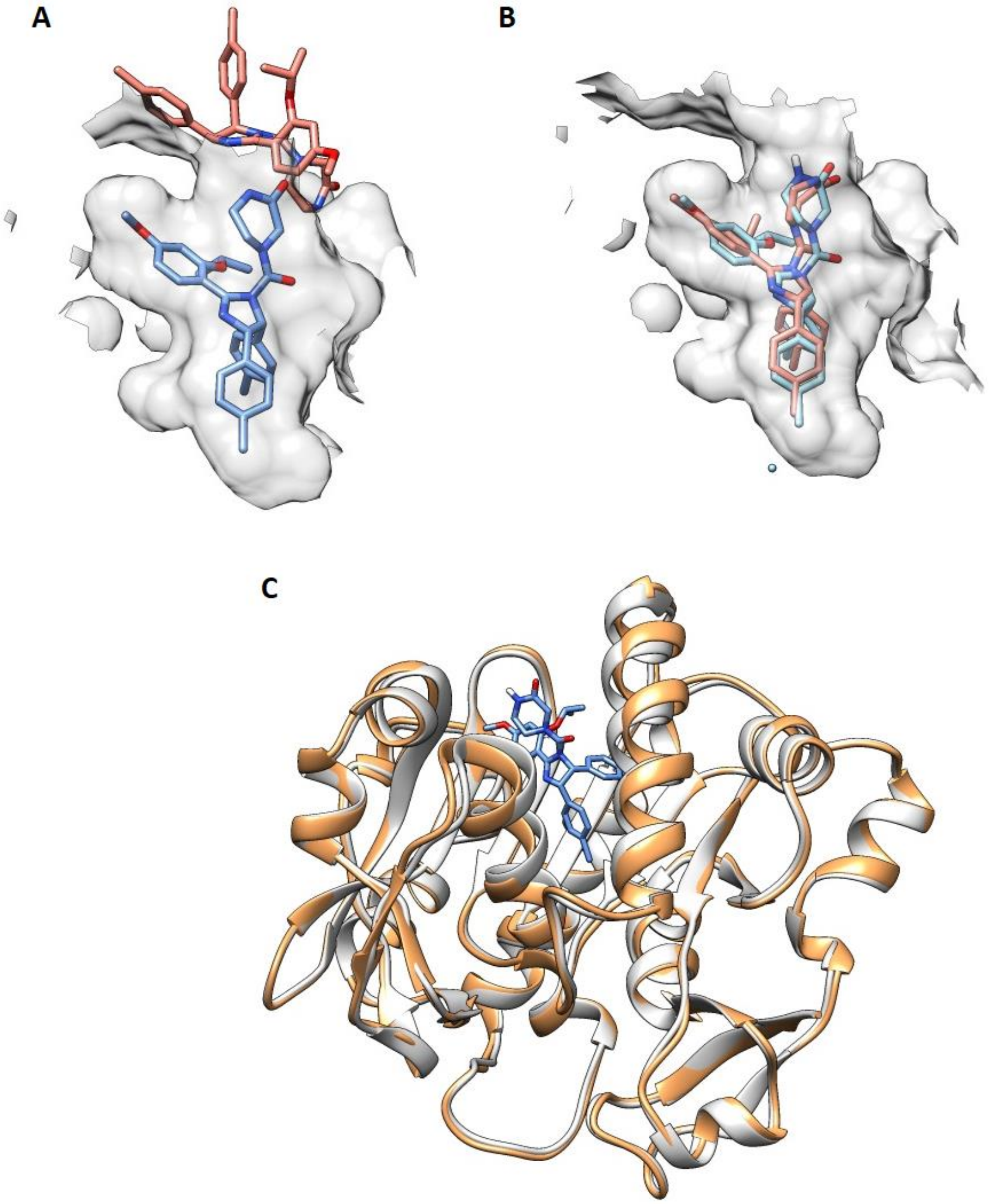


Figure 6- (A) Complex of Gamma glutamyl hydrolase binding site (in surface representation PDB ID-1L9X:A) with Nutlin-3a containing the predicted Nutlin posed by

CLICK (blue stick representation) and as predicted by Autodock-Vina (salmon stick representation) [when the docking was done on the crystal structure] (B) Complex of Gamma glutamyl hydrolase binding site (in surface representation PDB ID-1L9X:A) with Nutlin-3a containing the predicted Nutlin pose by CLICK (grey stick representation) and as predicted by Autodock-Vina (salmon stick representation) [when docking was performed on the structure obtained after Amber11 relaxation of Nutin-3a-GGH complex] (C) Superimposition of the crystal structure of Gamma glutamyl hydrolase (grey ribbon) and the structure after Amber11 relaxation with Nutlin-3a (brown ribbon). Nutlin-3a being depicted in blue sticks.

Nutlin-3a was then docked onto the structure of Gamma-glutamyl hydrolase obtained after energy minimizing the PDB with Nutlin-3a using Amber11 (Section 2.4). Autodock Vina predicted the top binding pose similar to that predicted by CLICK with the 3 aromatic rings of Nutln-3a superimposing onto each other (Figure 6B) with a predicted binding free energy of -10.9 kcal/mol. Autodock Vina, only allows side-chain movements of the specified flexible residues, while thermal fluctuations can also bring about main chain movements leading to opening up of pocket to fit the ligand. Energy minimization using Amber11 after transferring the ligand to the CLICK predicted binding site allowed sufficient movement of the main chain atoms to fit in the ligand (Figure 6C).

Though we have computed binding energies in 2 different ways, these may not necessarily be indicative of favorable (or unfavorable) binding. These energy/scoring functions are inexact and do not always capture the surface chemistry accurately. In this case, the ligand is hydrophobic and we believe that the binding surface of its receptor should similarly be hydrophobic, as seen in the Nutlin-MDM2 complex crystal structure. The hydrophobicity of the 8 predicted binding pockets from proteins that had better single point binding energies than MDM2 were examined by manual inspection (using Columbic coloring in Chimera). 2 of the 8 proteins, including gamma glutamyl hydrolase and human deadbox RNA helicase, showed predominantly hydrophobic pockets and were expected to bind stably to Nutlin. The other 6 proteins had polar patches in their binding pockets or had polar pocket peripheries. Either one of these characteristics was deemed as destabilizing towards binding Nutlin (Figure 7).

To test the validity of the hydrophobicity conjecture proposed above, 4 of these 8 proteins and their Nutlin bound complexes were subjected to triplicate 100 ns MD simulations. Of the 4 hits, two had hydrophobic pockets (Gamma glutamyl hydrolase and human deadbox RNA helicase) while the other two had some polar residues lining the pockets (inteferon gamma) and/or the pocket periphery (steryl sulphatase) (Figure 7).

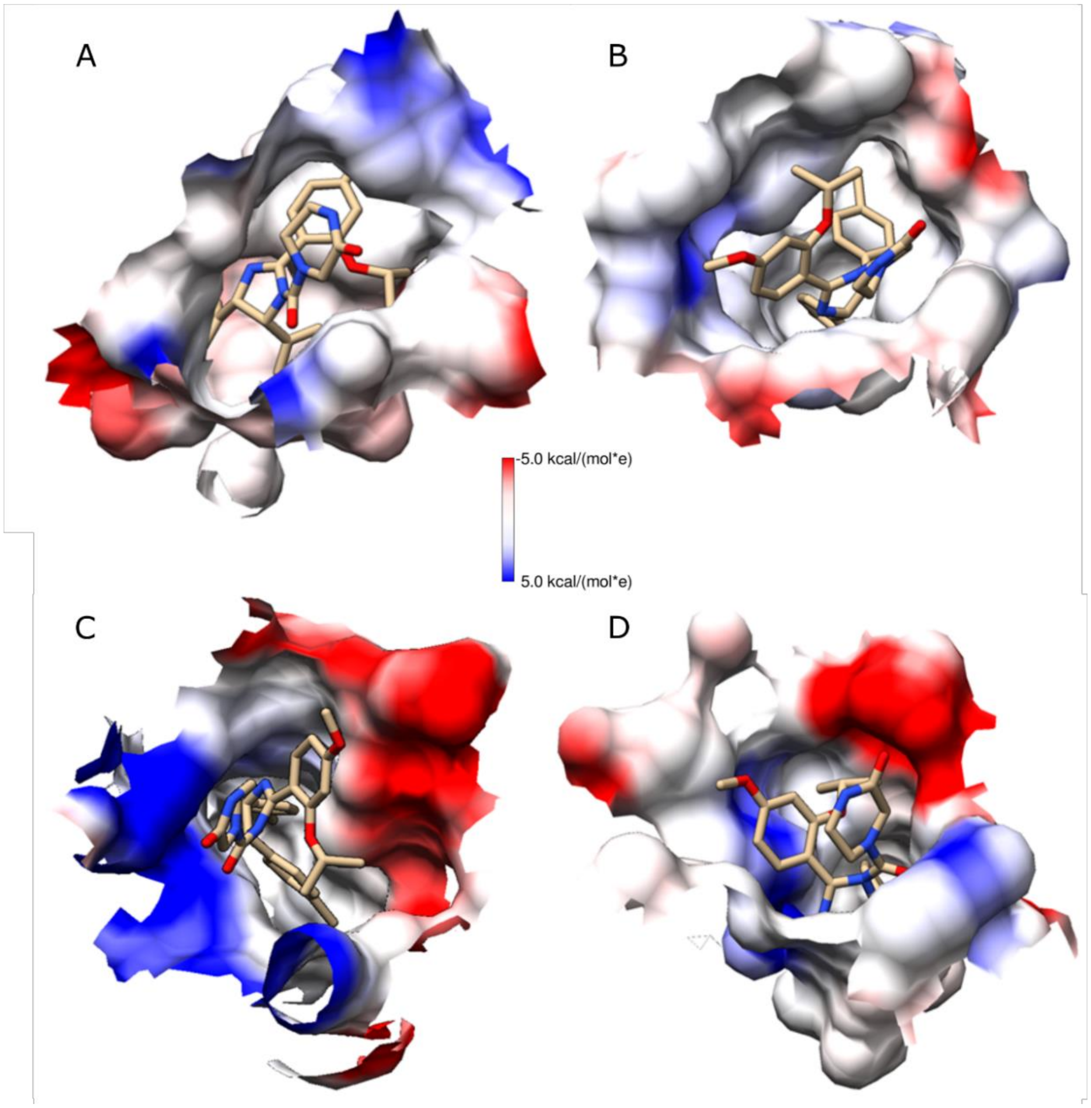


Figure 7- Surface representation of the predicted binding pocket of Nutlin (within 6 Å) in (A) Gamma-glutamyl hydrolase (1L9X:A), (B) Human dead box RNA helicase (3DKP:A) (C) interferon-gamma (1FYH:A) (D) steryl sulfatase (1P49:A). The binding pocket is coloured as per the Chimera [271] *rendered columbic charge representation.*

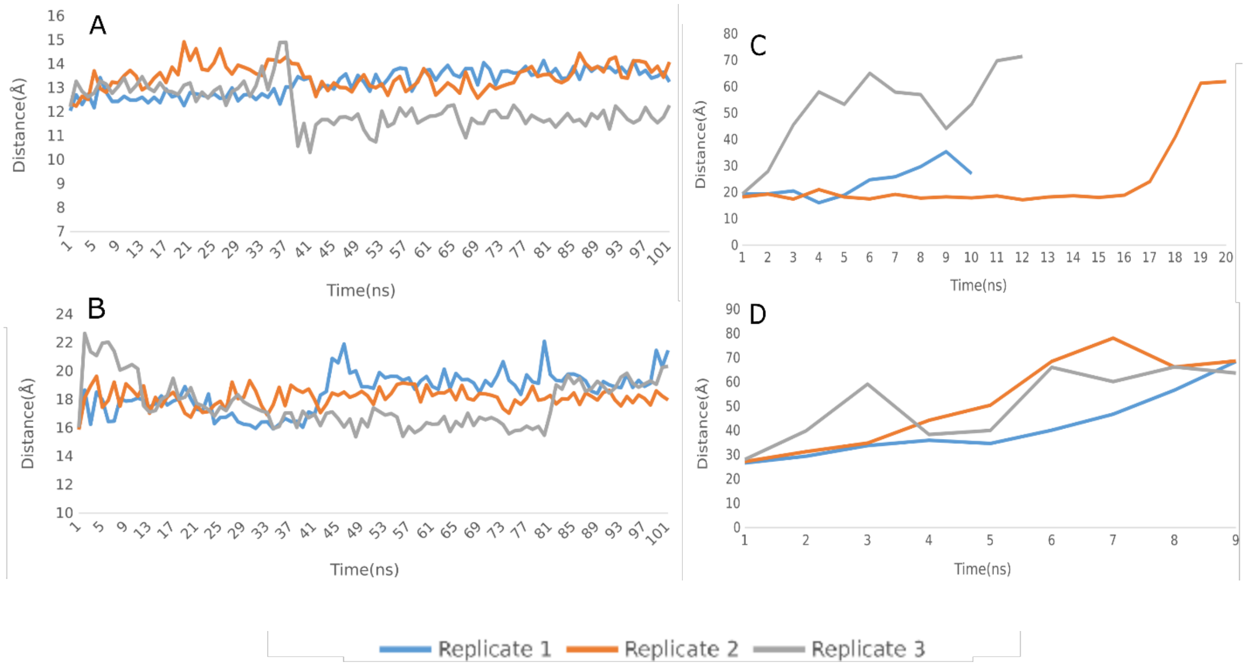


Figure 8- Distance trajectories of the center of the protein from the center of the Nutlin-3a for (A) Gamma-glutamyl hydrolase (1L9X:A), (B) Human dead box RNA helicase (3DKP:A) (C) interferon-gamma (1FYH:A) (D) steryl sulfatase (1P49:A). The three different trajectories from the triplicate simulations are depicted in different colors.

We measured the stability of the Nultin-bound protein complexes by analyzing the trajectories of the distances of the center of Nutlin-3a from the center of the protein during the MD simulations (Figure 8). Nutlin-3a remained in the predicted binding site for the two proteins (Gamma glutamyl hydrolase and Human dead box RNA helicase) (in all the triplicate simulations) with a hydrophobic pocket and rim throughout the course of the simulation. The average distances between Nutlin-3a and Gamma glutamyl hydrolase and Human dead box RNA helicase were 13.0 Å (+/-0.8 Å) and 18.1 Å (+/-1.3 Å) respectively. The distances between the centers of the Nutlin to that of the protein in complexes with a hydrophilic binding site/periphery (Interferon Gamma and Steryl Sulfatase) increased to greater than 10 Å of the initial value (in all the triplicate simulations). At this stage the simulation was stopped. Such large deviations from the initial position are indicative of an irreversible dissociation event. In order to check if the Nutlin bound complex showed unusual fluctuations of their residues, an average root mean square fluctuations (RMSF) of the residues with respect to the average position of the residues during the simulations were calculated. Nutin bound and Nutlin unbound complexes of Gamma glutamyl hydrolase and Human dead box RNA helicase show a similar RMSF, indicating the stability of the complex (Figure 9). The stability of two of the simulations with a predominantly hydrophobic rim and pocket indicates that along with matching 3D structural environment, the physicochemical properties should match for efficient binding.

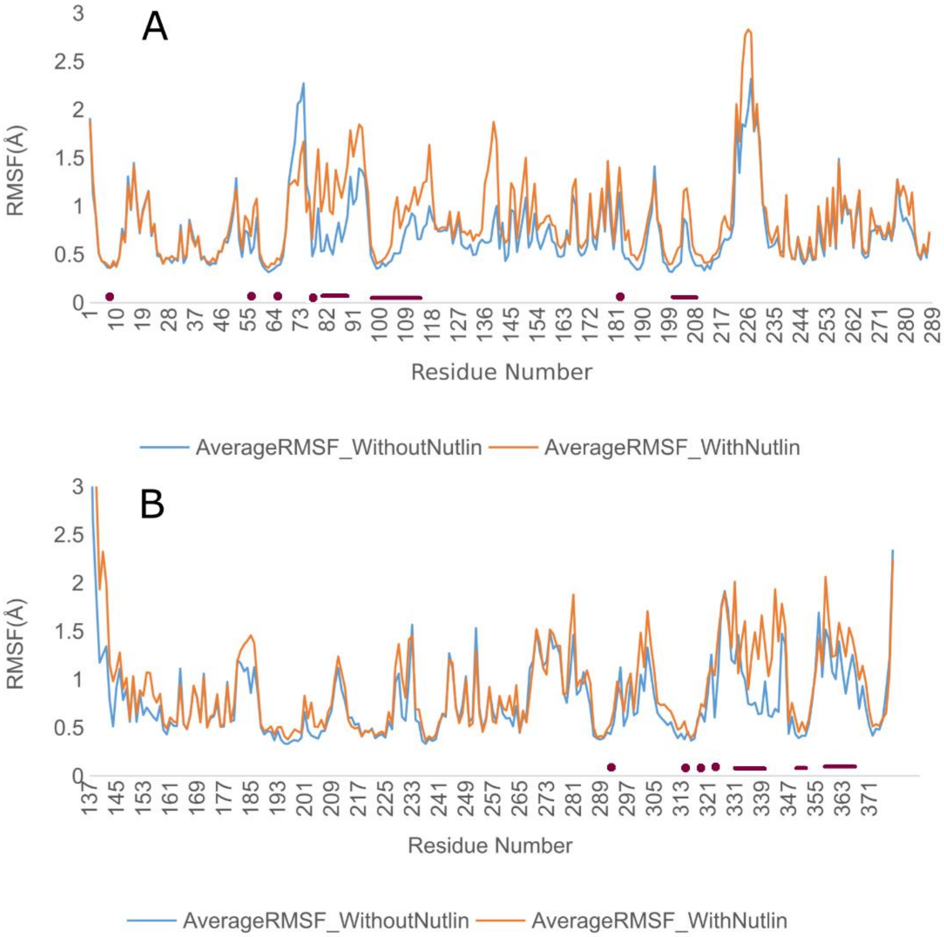


Figure 9- RMSF of individual residues with respect to the average structure generated from the MD simulation for the Nutlin-3a bound (maroon) and unbound (blue) structures of (A) Gamma-glutamyl hydrolase (1L9X:A), (B) Human dead box RNA helicase (3DKP:A). The regions in the protein that are within 6 Å of Nutlin-3a are depicted by purple dots or lines on the x-axis.

3.4. Thermal shift assay

Our computational analysis predicts a stable association between Nutlin-3a and Gamma glutamyl hydrolase. Not only does it have favorable binding energy, but its predicted binding pocket is also hydrophobic and the bound Nutlin-3a does not dissociate during the triplicate 100ns MD runs. Gamma glutamyl hydrolase has an AutoDock score of -10.9 kcal/mol and an RMSD of 0.54 Å between the AutoDock relaxed posed and the CLICK binding pose. The Nutlin-3a binding to this protein was hence chosen for experimental validation. Gamma glutamyl Hydrolase (100 ng in PBS buffer) (obtained from Abcam) was diluted in PBS buffer containing DMSO at 0.1% final concentration (25 µl) with or without Nutlin-3a at a concentration of 1 µM, 2.5 µM, 5 µM, 10 µM, 20 µM, 40 µM. The dose titration (through 1 µM, 2.5 µM, 5 µM, 10 µM, 20 µM, 40 µM of Nutlin-3a and DMSO) at 42°C revealed that at least 40 µM Nutlin-3a was required to protect the enzyme Gamma glytamyl hydrolase from thermal denaturation.

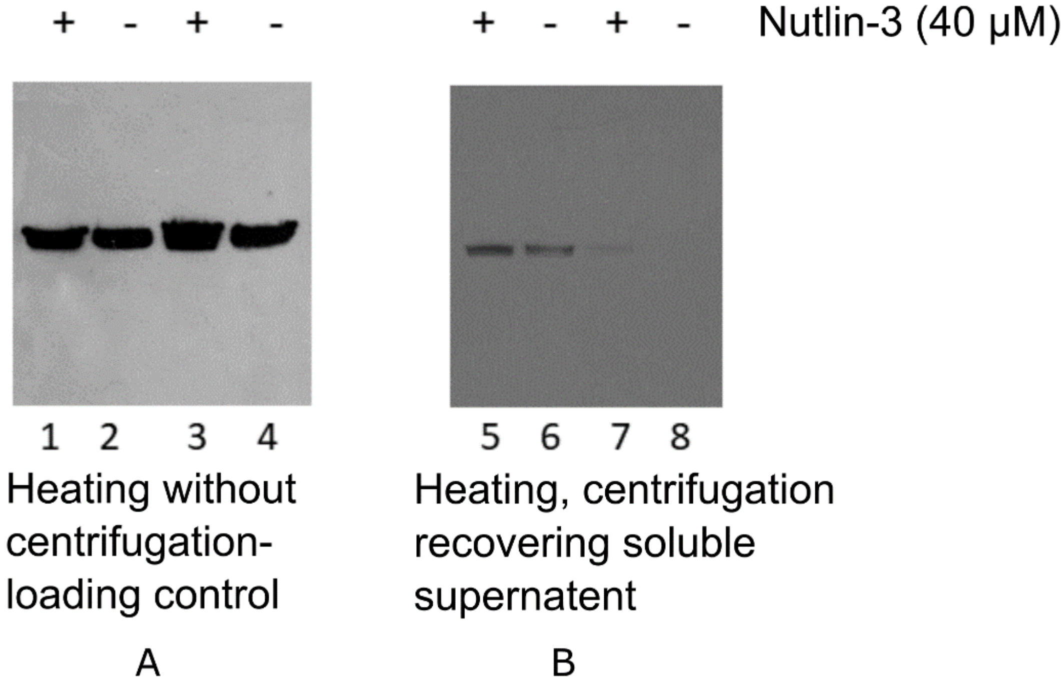

+ - + -
+ - + -
Nutlin-3 (40 µM)
1 2 3 4
5 6 7 8
Heating without centrifugation-loading control
Heating, centrifugation recovering soluble supernatent
A
B

Figure 10- Data showing that incubation of Gamma glutamyl hydrolase (100 ng in 25 µl PBS containing DMSO at 0.1% final concentration) with Nutlin-3a can (40 µM containing DMSO at 0.1% final concentration) stimulates the protection of the protein from heat induced aggregation/denaturation, which is suggestive of binding as modeled. The right

four lanes show the effects on Gamma glutamyl hydrolase stability as a function of temperature and Nutlin-3a. (+) indicates the presence of Nutlin-3a whereas (-) indicates its absence. A) The left four lanes are loading controls that were processed without centrifugation and (B) after centrifugation in the soluble fraction. Samples in lanes 1, 2, 5 and 6 are incubated at 37℃ while samples at lane 3, 4, 7, 8 are incubated at 42℃.

Following incubation of the enzyme with (+) and without (-) Nutlin-3a at a concentration of 40 μM (containing DMSO at 0.1% final concentration) at 37℃ (Figure 10, lanes 1, 2, 5, 6) or 42℃ (Figure 10, lanes 3, 4, 7, 8) for 30 minutes, the samples were either centrifuged (Figure 10B, lanes 5-8) or not centrifuged as controls (Figure 10A, lanes 1-4). Following this procedure, the soluble supernatant was recovered. The data (Figure 10B lanes 5 and 7) show that Nutlin-3a can promote enhanced thermal protection of Gamma glutamyl Hydrolase from heat aggregation/denaturation. These data suggest that, in principle, our screens can identify novel functional modes of binding for the small molecule Nutlin-3a.

3.5. Comparison of CLICK with other methods for identifying putative Nutlin binding site

In order to check the efficacy of CLICK in identifying putative binding sites for Nutlin, it was compared to other methods for investigating similarities in Nutlin binding sites. To the best of our knowledge, there are currently 2 methods/servers for doing this task, SMAP [272] and idTarget [273]. We tested the predictive ability of these servers for prospective Nutlin binders. The SMAP server (http://smap.nbcr.net; this site is currently unreachable) lists 5 best hits with significant p-values (<1.0e-4) (Table 2). It searches for regions in other proteins that are structurally similar to the Nutlin binding site on MDM2. SMAP is used for the comparison and the similarity search of protein 3D motifs independent of sequence order and has been applied for predicting drug side effects and to repurpose existing drugs for new indications. All the 5 best hits of SMAP – 1z1mA (62), 2qagC (4), 3dzuA (20), 2qagB (10) and 1lv2A (43) have clashes with Nutlin (as mentioned by the number in parenthesis), and their MM/GBSA binding energies could not be calculated.

Table 2: The results from SMAP server to find Nutlin binding pockets in human protein structures

PDB ID	Number of steric clashes
1z1m:A	62
2qag:C	4
3dzu:A	20
2qag:B	10
1lv2:A	43

The second method, idTarget, predicts possible binding targets of a small chemical molecule via a divide-and-conquer docking approach, using the scoring scheme from AutoDock4. idTarget has been shown to be able to reproduce known off-targets of drugs or drug-like compounds. The idTarget server (*http://idtarget.rcas.sinica.edu.tw*), detected 34 hits with Z-score<-0.5 and using the same clash tolerance as described in section 2.4. We have made a simplistic assumption in this study that binding to Nutlin should also mean similarity in the binding site. In all the putative targets identified by CLICK, the predicted binding sites overlap with the MDM2 site by more than 70%. Only three of the idTarget hits have an overlap of 70.21% with the MDM2 site. The overlaps of the predicted hits by idTarget range between 23 – 70% (Table 3). Though, idTarget used AutoDock scoring schemes to score the predictions and are all shown to have favorable interaction energies with Nutlin, the binding free energies calculated using MM/GBSA of Amber 11 show positive values. 2r4v (Chloride intracellular channel protein 2) is commonly identified as a target both by CLICK and idTarget, but identify different regions on the protein as binding sites. The 7 alternative target identified by CLICK that has lower single point energy and AutoDock binding affinity scores than MDM2, was neither predicted by SMAP nor by idTarget.

Table 3: The results from idTarget server to find Nutlin binding pockets in human protein structures.

PDB ID	**Number of steric clashes**	**Binding energy (kcal/mol)**	**Number of binding site residues in the predicted binding site within 6 Å**	**Structure Overlap with respect to the number of residues in Nutlin-2 binding site of Mdm2**
3kjd	2	965.05	32	61.70
3l0l	0	590.71	38	70.21
3i28	1	6927.15	34	68.09
2zb4	1	395.69	39	68.09
3inm	0	18.91	23	51.06
3g2f	1	127.59	32	70.21
2x7g	0	64.62	32	65.96
1wb0	0	5.54	7	23.40
2vuw	1	95.31	35	63.83
2h7c	1	596.55	20	48.93
3e7e	0	127.37	29	61.70
2ipx	2	88.84	28	53.19
2wax	0	250.74	13	42.55
3bgv	0	1003.23	28	51.06
3c0i	0	472.41	35	70.21
1z70	0	1788.47	20	44.68
2jc9	1	66.97	25	61.70
2wef	1	167.99	34	65.96

2r4v	0	99.25	19	59.57
3bpt	0	257.11	28	55.32
3epy	0	126.83	15	51.06
2pez	0	59.56	26	51.06
3ebb	1	3945.18	18	48.94
2ql9	1	232.35	7	29.79
1s1d	0	160.35	23	59.57
2vfk	1	32.01	21	48.94
2wm1	0	448.78	15	42.55
3iai	0	46.17	30	68.09
1wl4	0	524.28	22	51.06
3ijj	2	602.19	15	42.55
1d4a	1	84.94	20	53.19
3e9k	1	155.08	21	57.45
2zg1	2	203.13	16	46.81
1elv	1	153.43	16	48.94

4. Discussion

Broadly speaking, for the productive binding/interaction of biomolecules, there needs to be complementarity in geometry and chemistry. In this study, we have showcased the utility of our CLICK software in detecting protein sub-structures (binding sites) with similar geometry. We had previously shown that CLICK could detect ATP binding sites by structural similarity. Here, we have used CLICK to detect putative binding sites on proteins that are structurally similar to the Nutlin binding pocket on MDM2. This was effected by mining a non-redundant structural database of human proteins for regions of proteins that are structurally similar to the Nutlin binding sites obtained from the MDM2-Nutlin complex crystal structure and snapshots from its MD simulations. We found 49 human proteins that have regions that are structurally similar to the Nutlin binding site on MDM2. To ensure that the geometric similarities were significant, we ensured that at least 70% of the residues in the putative hits overlapped. Additionally, when placing the Nutlin in these predicted pockets there were less than 1 main chain and 5 side-chain clashes. We believe that such stringent criteria would exclude some of the known Nutlin binders, such as Bcl-X_L [274], but help in minimizing false positives. In future applications, we are exploring the use of sub-optimal matches, which would have predicted Bcl-X_L as one of the putative binders.

Having satisfactorily obtained similarly structured pockets, we next evaluated the chemistry of interaction or simply the binding energies of the predicted Nutlin-protein complex. These computations were done using an MM/GBSA scoring scheme as well as the AutoDock energy scores. These putative alternate targets of Nutlin also had favorable energies of binding as computed/predicted by single point energy calculations and by the AutoDock energy function. The AutoDock computed energies for 7 Nutlin-protein complexes had lower values than the native Nutlin-MDM2 complex. Consistently, all 7 of these complexes also have better MM/GBSA scores than Nutlin-MDM2, in addition to Peptidylprolyl Isomerase domain and WD Repeat Containing Protein 1. The rank ordering of the two scoring schemes of the putative alternate targets was also similar (correlation coefficient of 0.69). On comparing the RMSD of the two poses and the single point energies, we again found a good correlation (coefficient 0.57). Interestingly, the RMSDs between the two predicted poses were smaller with more favorable binding energies (by both methods).

Despite the good agreement between our single point energy scores and AutoDock evaluations, we are aware that these molecular mechanics and empirical scoring schemes are often not very accurate. We evaluated the physicochemical nature of the ligand and its receptor site. Nutlin is predominantly hydrophobic and in its complex with MDM2, it is bound in a hydrophobic site. We chose 2 predicted binders each with hydrophobic and non-hydrophobic pockets and subjected them to triplicate MD simulations. The Nutlin bound to Interferon gamma and Steryl sulfatase did not remain bound to the predicted non-hydrophobic pocket. Whereas, Nutlin remained bound to the hydrophobic pockets of Gamma glutamyl hydrolase and Human dead box RNA helicase in triplicate 100ns MD trajectories.

In order to assess whether the binding of the ligand/drug would alter/affect the functioning of the putative hits, we attempted to correlate functional information to the amino acids that constitute the binding site. Function is associated with all amino acids in the active site of enzymes and at sites of post-translational modifications. Simplistically, we have assumed that any binding in the proximity to these functional sites would impair protein activity. We found 16 predicted targets whose functions are likely to be affected on Nutlin binding. It is possible that (some of them) other predicted target sites may affect protein function through allostery.

Given our somewhat modest resources and the inhibitive cost of Nutlin, we experimentally validated the binding of Nutlin-3a to Gamma glutamyl hydrolase. This enzyme has a hydrophobic binding site, does not dissociate with Nutlin in MD simulations and we predict would have its function (glycosylation) affected on binding Nutlin. We showed that Nutlin-3a can protect Gamma glutamyl hydrolase from thermal denaturation.

An important implication of our study is that this procedure can be used not just to discover alternate binding sites for known ligands/inhibitors/drugs, but it could serve as a platform to repurpose known drugs. For instance, the levels of Gamma glutamyl hydrolase have been implicated in several disease conditions including several cancers and arthritis, and perhaps the binding of Nutlin could influence favorable therapeutic outcomes. In this study, in conjunction to the energy scores computed by two different methods, we felt it was necessary to manually look into ligand-receptor specific properties, in this case, hydrophobicity. For a larger more general application of this method, an automated classification of the ligand and receptor/binding site would be required. While we are working towards that end, it is beyond the scope of this study.

Another positive aspect of using the CLICK software was that we identified hits that are not readily identified by other docking procedures. For instance, when we relaxed/scored the ligand-protein complexes with AutoDock Vina, we had to make use of our complex structures as a starting point and could not begin with the crystal structures. This is because our predicted models after energy minimization had opened to slot in the ligand while the binding sites on the crystal structures were seldom in a conformation conducive to ligand binding.

We also compared the performance of our method to two other methods, SMAP and idTarget. Except for one target, none of the proteins identified by our method was predicted by the other methods. All the targets identified by the SMAP server had several clashes with the Nutlin, and hence the single point energies could not be calculated. The MM/GBSA energies for the hits identified by idTarget were consistently unfavorable. Even if this assessment may not be completely accurate, we also noticed that the structural overlap between the idTarget hits to the MDM2 binding site was seldom, if at all, as high as our predictions. We used a threshold of 70% similarity to filter our CLICK identified hits. idTarget predictions had structural overlaps in the range of 23-70%.

In conclusion, the program CLICK has been used to identify the possible proteins that Nutlin can bind to. The participation of these proteins in different biological pathways hints at likely off-target effects such as toxicity. Experimental techniques such as CETSA [275] that have the potential to identify the drug target but in general are time consuming and/or expensive. In contrast, CLICK can exhaustively and quickly search large sets of protein structures to identify best target candidates, and can hence reduce the experimental tests to a limited number of proteins, resulting in a reduction in time and cost efficiency. Hence CLICK can be used as an initial screening tool for cost effective toxicology studies of drugs. In this chapter, we present a list of proteins that could potentially bind to Nutlin, which can be used to validate their binding and off target effects. The best hits presented in this chapter are only a partial list of targets for Nutlin binding as not all proteins have known 3D structures. A limitation of our method is that it is dependent on the availability of experimentally determined 3D structures of proteins. We believe that a larger study could be envisioned by utilizing homologous structures or models. However, that is also beyond the scope of this current study.

In this chapter, we described the prediction of binding pockets of small molecule drug Nutlin based on structural similarity. Though the pilot study was done only with one ligand, in principal the technique can be generalized to predict binding site of any small molecule ligand based on structural similarity. We utilized the same strategy to identify binding pockets in already known alternate targets of drugs, which has been described in the next chapter.

Chapter 8 - Prediction of binding sites of drugs on off-target proteins

1. Prediction of the binding site of drugs on experimentally validated alternate targets

The work was done in collaboration with Kaustubh Amritkar.

1. Introduction

Small molecules bind to proteins based on complementarity in shape and physicochemical properties. Different proteins can have a similar binding site. Hence a small molecule can bind multiple targets. In the previous chapter, we developed a structural match based method to predict off target effects of the small molecule drug Nutlin. This method can be used to predict binding pockets of small molecules such as drugs in other proteins. In this chapter, we predicted the binding pockets of drugs in proteins that are experimentally validated to be off targets for the drug. We first extracted the binding site of the drug from its target protein and then searched for a structurally similar binding pocket in other experimentally validated off-targets. This tool can be used for the repurposing of an existing drug molecule and can also be used to recognize potential off target effects of drugs.

2. Methods

2.1. Dataset of off target proteins of drugs

Table 1 – List of the tested drug (drug predicted by Campillos et al. to bind to the same target) and the reference drug that shares the common off target protein. The original target of the tested drug has also been mentioned [277].

Tested Drug	Original target of tested frug	Reference Drug	Off target protein
Donepezil	Acetylcholine esterase enzyme	Venlafaxine	Serotonine Transporter (5HTT)
Fluoxetine	Sodium dependent serotonine receptor	Rabeprazole	Dopamine Receptor D3 (DRD3)
Rabeprazole	Cytochrome P450 2C19	Zolmitriptan	5-Hydroxytryptamine receptor 1D (HTR1D)
Rabeperazole	Cytochrome P450 2C19	Pergolide	Dopamine Receptor D3 (DRD3)
Paroxetine	Cytochrome P450 2D6	Rabeprazole	Dopamine Receptor D3 (DRD3)

Zaleplon	Gamma-aminobutyric acid receptor subunit alpha-1	Mirtazapine	Histamine receptor H1 (HRH1)
Disopyramide	Sodium channel protein type 5, subunit alpha	Maprotiline	Histamine receptor H1 (HRH1)
Clomiphene	Estrogen receptor alpha	Cetirizine	Histamine receptor H1 (HRH1)
Loratadine	Histamine receptor H1	Estazolam	Translocator protein BZRP
Raloxifene	Estrogen receptor alpha and beta	Tegaserod	5-Hydroxytryptamine receptor 1D (HTR1D)
Acitretin	Retinoic acid receptor	Cetirizine	Histamine receptor H1 (HRH1)
Doxorubicin	DNA topoisomerase 2-alpha	Ziprasidone	Histamine receptor H1 (HRH1)
Ketoconazole	Lanosterol 14-alpha demethylase	Prochlorperazine	Serotonin Receptor

Campillos *et al.* [246] used phenotypic side-effect similarity to predict if two drugs share a common target. This scheme was applied to 746 marketed drugs, which lead to 1018 side effect driven drug relations. Out of these, 261 chemically dissimilar drugs were identified. 20 out of these drug-off target relations were experimentally tested, of which 13 were validated by in vitro binding assays. The binding site of these 13 drugs on their off-target protein (as predicted by Campillos *et al.*) was predicted.

The experimentally determined structures of the drug bound protein complex (for each of the drugs tested by Campillos *et al.*) (Table 1) were extracted from PDB [278]. The binding site of the drug was extracted using a distance cut-off of 6 Å from the drug molecule.

2.2. Prediction of structurally similar binding sites

The binding sites on the off target protein for a drug (as predicted by Campillos *et. al.*) (Table 1) were predicted using the program DEPTH [12]. During this computation, the minimum number of water molecules for bulk solvent calculations was set to 4. Evolutionary information was also included during the computation. The probability threshold of 0.8 was used as a cutoff for binding site prediction.

As mentioned in the previous chapter, CLICK, a topology independent structure comparison tool was used for superimposition and comparison between the extracted drug bound site (from the crystal structure) and the DEPTH predicted binding site on the off target protein. The approach used in the previous chapter (Chapter 7) has some difference compared to this approach. In the previous chapter, the search for the structurally similar binding pocket was done on the entire protein. However, in this chapter, we restricted the search to the DEPTH predicted binding site to identify structurally similar binding pocket. The drug molecule is then transferred as a rigid body on the structurally similar binding site predicted by CLICK to obtain the off target protein-drug complex (Refer to Chapter 7 Figure 1). The number of clashes (empirically chosen distance <2 Å) between a non-hydrogen atom from the target protein and drug molecule is determined. These are classified into two classes – main chain (MC) and side chain (SC) clashes. BLOSUM62 substitution matrix is used to check for the significance of the structural alignment obtained by CLICK. In case the drug was present in a complex with multiple PDBs, all the individual PDBs were used as templates for the structural match onto the DEPTH predicted binding site.

2.2.1. Energy Minimization of the drug-protein complexes using Molecular Mechanics

The protein structures obtained from the PDB database are a snapshot of the protein biomolecule without the drug bound to it. To get a more accurate estimate of the predicted interaction, energy minimization was carried out on the set of off target protein-drug complexes using Gromacs [262,263] with OPLS-AA [61,279] force field. Antechamber [266] was used for the addition of hydrogen atoms to the drug and protein structure files.

Parameter for the drug molecules was obtained using LigParGen [280]. Charge neutrality was achieved by adding sodium or chloride counterions. Steric hindrances in a complex were quantified by the number of clashes between the drug molecule and the target protein atoms. Ideally, we expect no or very few clashes after energy minimization.

2.3. Docking of drugs onto off-target proteins

To validate the binding site and binding pose, the docking tool Autodock4.2 [47] was used to dock a drug molecule onto the energy minimized target protein obtained from the previous step. A 40x40x40 $Å^3$ box was considered for docking of all drug molecules centered at the geometric center of the CLICK predicted drug molecule binding pose. The binding free energies (as obtained from Autodock4.2) and corresponding ligand RMSD (distance between the centroid of the drug molecule in the CLICK computed poses and the ones after docking) were calculated.

3. Results

3.1. Dataset of drug-protein complexes

5 different proteins (5HTT, DRD3, HTR1D, HRH1 and BZRP) were experimentally validated as off targets of 13 drugs. Out of these 5 proteins, only two proteins: Dopamine Receptor D3 DRD3 (PDB - 3PBL) and Histamine H1 receptor HRH1 (PDB - 3RZE) had available structures in the PDB. Out of 13 unique drug effects tested by Campillos *et al.* only 6 drugs (Cetirizine (2 stereoisomers LCR, CZE), Disopyramide (DP0), Doxorubicin (DM2), Fluoxetine (RFX), Paroxetine (8PR) and Rabeprazole (RZX)) had a structure of drug-protein complexes in the PDB. In case a drug was present in multiple structures in the PDB, all the binding sites from the different crystal structures were extracted and used for comparison. There was a total of 26 drug-protein complexes as extracted from PDB - 1 each for LCR/CZE, RZX, 2 each for DP0 and RFX, 9 for 8PR and 10 for DM2**.**

3.2. Structural similarity between binding pockets for the same drug bound to different proteins

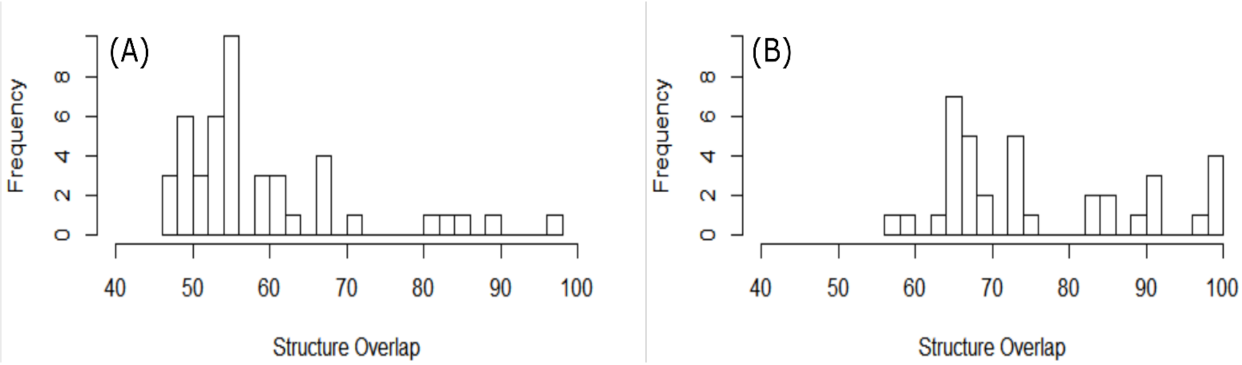


Figure 1- A histogram depicting the structural overlap (%) between binding pockets for the same drug bound to different proteins. (A) Doxorubicin (B) Paroxetine

Multiple drug bound protein complexes were available for Doxorubicin (10 structures) and Paroxetine (9 structures). An all against all CLICK superimposition between the binding sites (extracted using a distance cut off of 6 Å from the drug molecule) were performed. Hence for a drug molecule with n structures, the total number of comparisons is n*(n-1)/2. On performing a CLICK superimposition among these binding sites, the structure overlap between the binding sites varied between 46%-97% and 56%-100% for Doxorubicin and Paroxetine respectively (Figure 1). This shows that the same drug can bind to a structurally similar or a dissimilar binding site [281]. However, in this method, we limit ourselves to making predictions for the drugs that bind a structurally similar binding pocket compared to their known binding sites.

3.3. Prediction of structurally similar off-target binding sites

Gromacs with the OPLS-AA force field was used to energy minimize the CLICK predicted drug-protein complex structures. Out of the 26 complexes, parameters for 3 drug molecules could not be created using LigParGen hence energy minimization was performed for only 23 complexes. All of the 23 potential drug-protein binding sites had zero clashes after energy minimization, suggesting the structurally similar CLICK binding site is viable.

A structure overlap ranging between 67% to 88% is observed among the CLICK superimpositions between the binding site from the crystal structure and the predicted binding site. The BLOSUM62 scores (Table 2) for comparing sequence similarity in the structural alignments done by CLICK is low for most cases (ranging between -56 to 9).

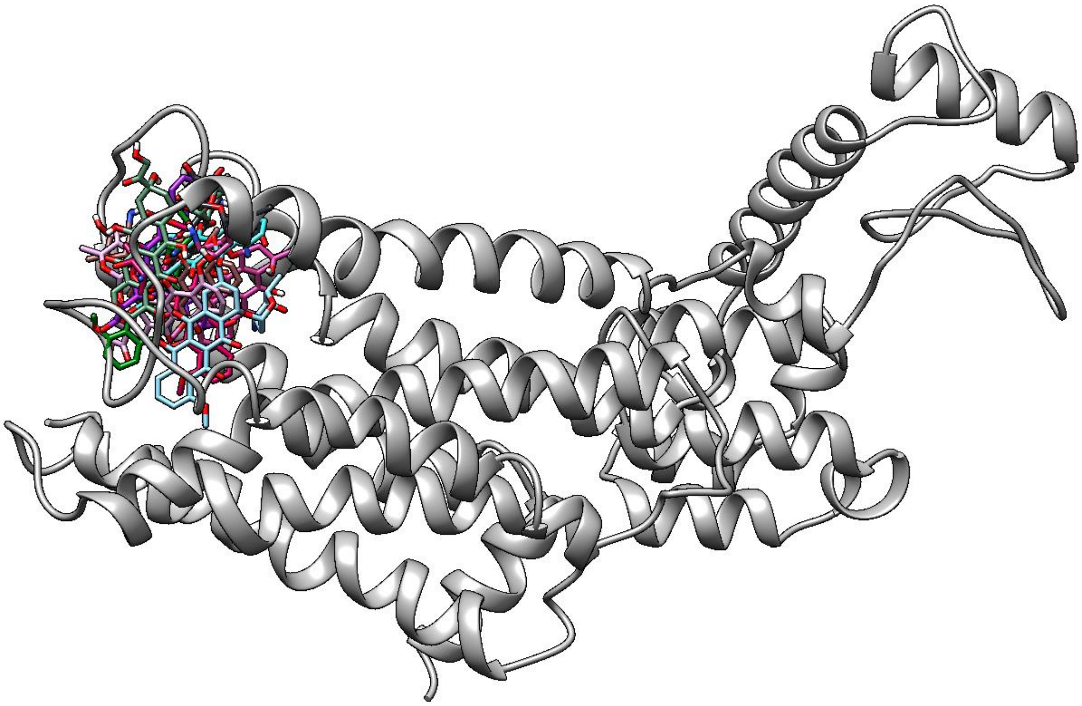


Figure 2 – Predictions of the binding pose of DM2 (shown in colored sticks) onto HRH1 (shown in grey ribbons) (after energy minimization) depict similar regions of binding. DM2 was present in complex in 10 PDB structures, each of which was used as a template to search for the structurally similar binding pocket. The final predicted binding pose from these 10 templates has been shown in different colored sticks [276].

Predictions made using all the 10 sites from the Doxorubicin bound protein complexes structurally superimposed on the same binding pocket on HRH1 (Figure 2). In some cases, the CLICK predicted binding site for a drug molecule (Cetirizine and Paroxetine) on a target protein differed based on the binding pocket used for a structural similarity search (template). Out of the 9 binding sites for Paroxetine (from crystal structures), 7 predicted a common binding pocket, while 2 others predicted a different binding pocket (Figure 3). Similarly, the two predictions for Cetirizine by CLICK were on two different pockets. This could happen because CLICK provides the best superimposed hit as the output and in certain cases, the best superimposed hit might not be the correct binding site.

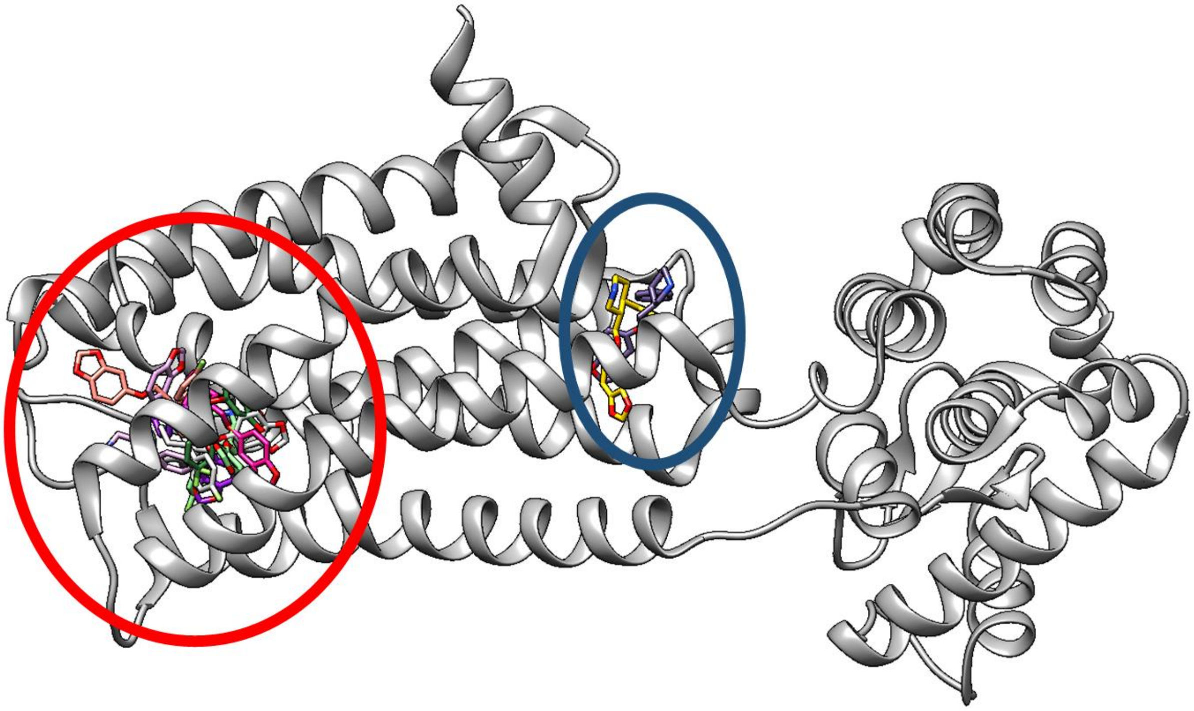


Figure 3 – Structurally similar binding site of a drug 8PR predicted from 9 different protein complexes on the target protein DRD3 (shown in grey ribbons). The different binding poses as predicted from the 9 complexes are shown in stick representation in different colors. 2 templates predicted the binding pockets, on the center of the protein (the binding pose of 8PR shown in yellow and purple sticks and the binding site highlighted by blue circle). The other 7 templates predicted the binding site on the left side of the protein (8PR conformations shown in stick representation of varying color, binding site highlighted in red circle) [276].

Table 2 - Superimposition data for drug bound protein binding site onto the off target protein predicted binding site. ['-' indicates .pdbqt file required for docking couldn't be created, '' indicates drug parameters could not be created from LigParGen]. The CLICK RMSD and Structure overlap are CLICK generated parameters. The MC and SC clashes refer to the clashes before energy minimization. The ligand RMSD is the RMSD of the ligand centroid predicted between the CLICK predicted pose after energy minimization and the docked pose. The binding free energy is predicted by Autodock4.2.*

Drug	Drug PDB	Target	CLICK RMSD (Å)	Structure Overlap (%)	MC clash	SC clash	BLOSUM62 Score	Ligand RMSD (Å)	Binding free Energy (kcal/mol)
Cetirizine	5dqf-CZE	HRH1	1.99	78.38	0	0	-1	*	*
Cetirizine	5dqf-LCR	HRH1	2.21	72.50	4	17	-36	*	*
Disopyramide	3apw-A	HRH1	2.45	71.70	0	23	-17	1.31	-5.65
Disopyramide	3apw-B	HRH1	2.47	71.70	2	32	-16	-	-
Doxorubicin	1i1e	HRH1	2.55	75.00	2	38	-14	0.21	-13.35
Doxorubicin	2dr6	HRH1	2.42	70.00	6	25	-22	0.38	-13.73
Doxorubicin	4dx7-A1	HRH1	2.51	66.67	14	13	-48	0.21	-13.04
Doxorubicin	4dx7-A2	HRH1	2.29	66.67	9	25	-47	0.83	-11.76
Doxorubicin	4dx7-B	HRH1	2.31	73.53	9	24	-24	1.72	-12.79
Doxorubicin	4zvm-A	HRH1	2.47	70.00	2	42	-25	1.45	-11.56
Doxorubicin	4zvm-B	HRH1	2.49	70.00	35	47	-42	0.86	-14.17
Doxorubicin	5mra	HRH1	2.17	81.48	35	11	-12	*	*
Doxorubicin	5om7	HRH1	2.24	74.00	12	26	-29	0.29	-15.5
Doxorubicin	6ftp	HRH1	2.69	76.60	6	19	-6	0.59	-13.68
Fluoxetine	3gwv	DRD3	2.24	88.24	0	16	-29	3.78	-5.77
Fluoxetine	4mm8	DRD3	2.34	77.78	0	12	-32	2.28	-6.91
Paroxetine	3v5w	DRD3	2.90	76.92	0	19	-31	4.01	-8.28
Paroxetine	4jlt	DRD3	2.60	87.50	7	37	-56	-	-
Paroxetine	4l9i-A	DRD3	2.69	78.26	1	8	-38	2.44	-8.22
Paroxetine	4l9i-B	DRD3	2.35	76.09	2	12	-32	3.78	-9.61
Paroxetine	4mm4-A	DRD3	2.44	77.05	1	18	-39	2.97	-8.86
Paroxetine	4mm4-B	DRD3	2.28	73.33	4	11	-30	1.92	-7.33
Paroxetine	5i6x	DRD3	2.38	79.31	1	33	-33	1.82	-7.52
Paroxetine	6awn	DRD3	2.32	79.31	4	23	9	-	-

Paroxetine	6f6i	DRD3	2.33	88.24	1	34	-30	-	-
Rabeprazole	3pgl	DRD3	2.48	87.50	0	21	-12	1.42	-7.01

3.3.1. Presence of ligand in the CLICK predicted binding site

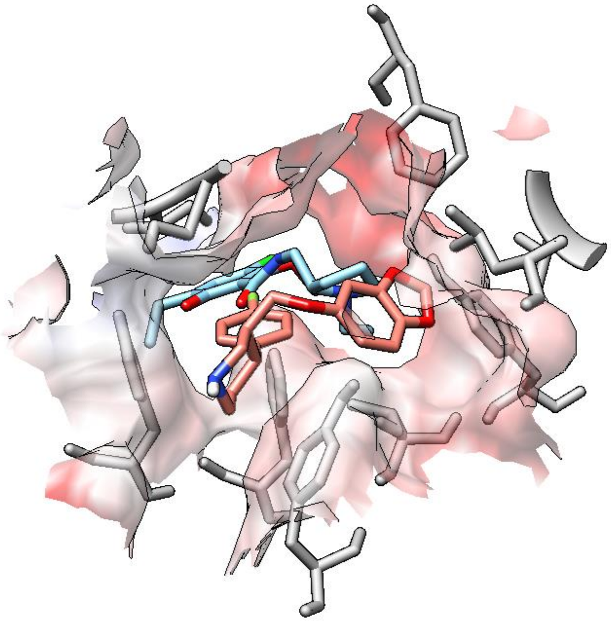


Figure 4 - Surface representation of the predicted binding pocket of Paroxetine (drug in salmon sticks) (surface in salmon surface and binding site residues in grey sticks). There is a ligand (Eticlopride on blue sticks) occupying the same binding pocket in the off target protein DRD3 (PDB ID - 3PBL).

In multiple cases, the CLICK predicted binding pocket for a drug molecule in a target protein was occupied by another ligand in the crystal structure, hence a common binding site is observed for different ligands. Paroxetine was present in a complex with 9 proteins. 7 out of these 9 complexes predicted the binding site of Paroxetine on DRD3 (PDB - 3PBL) onto an Eticlopride bound site (as seen in the crystal structure) (Figure 4). Similarly, the binding site predicted (from both the templates) for Fluoxetine on DRD3 was the same as the pocket with bound Eticlopride. Similarly, all the binding pocket predicted (from 10 templates) for Doxorubicin onto HRH1 (PDB - 3RZE), already had the ligand Doxepin bound (in the crystal structure). The predicted binding site of Disopyramide (from both the templates) on HRH1 was the site with bound ligand Doxepin (in the crystal structure). 1 stereoisomer of Cetrezine (LCR) was predicted to bind the site with bound Doxepin. Hence ~80% of the sites predicted by CLICK had an already bound ligand. This is consistent with previous literature where multiple ligands have been shown to bind to same/similar binding pockets [243,282].

3.4. Docking of drugs onto off-target proteins

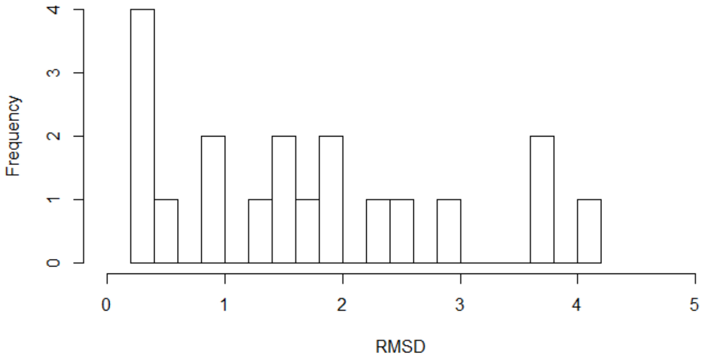


Figure 5 – Distribution of ligand RMSD (Å) between the Autodock4.2 predicted pose and the pose after energy minimization of the CLICK predicted pose.

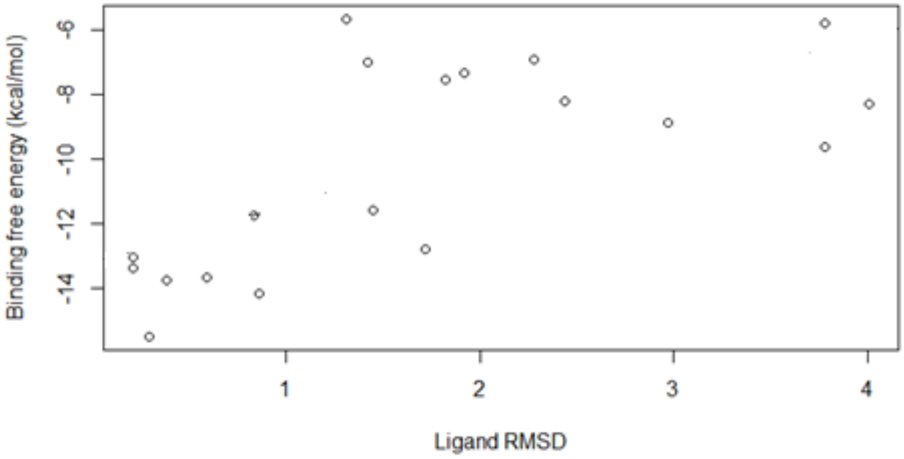


Figure 6 – Plot showing the Ligand RMSD (Å) versus the binding free energy as predicted by Autodock4.2. The straight line is the linear regression line.

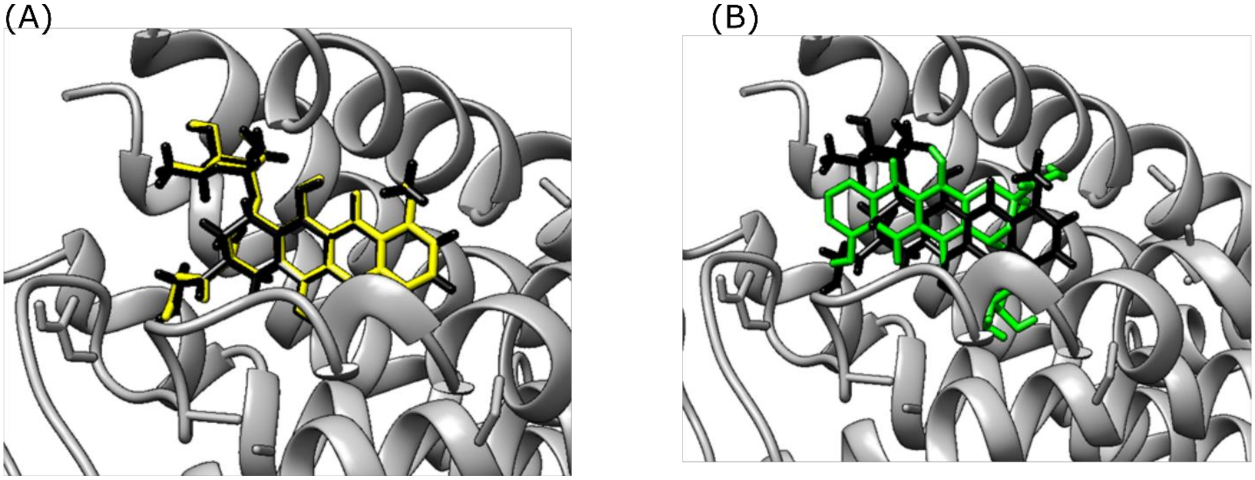


Figure 7 - Comparison between the CLICK predicted (black) (after energy minimization) drug and different docking poses of a drug Doxorubicin (DM2) onto the off target protein (HRH1) (shown in grey ribbons). (A) Docked pose (yellow) is similar to the CLICK predicted pose with a RMSD of 0.29 Å (B) DM2 is flipped in the docked pose (green) compared to the CLICK predicted pose, hence a RMSD of 2.52 Å.

The drug molecules were docked onto the CLICK predicted binding pocket for all 23 pairs using Autodock4.2. For 4 out of the 23 pairs, Autodock was unable to create the drug 'pdbqt' files (file containing the partial charges of the ligand atoms) essential for docking. Docking was also not done for the cases where the parameter file for energy minimization could not be generated by LigParGen (Section 3.3). Docking was hence limited to 19 complexes. The Autodock binding energies for the best binding pose (one with the lowest energy) were in the range of -15.5 to -4.3 kcal/mol (Table 1), indicative of favorable binding events.

We also compared the Autodock predicted poses to the CLICK predicted poses of the drug molecules. The ligand RMSD ranges from 0.21 – 4.66 Å (Figure 5). Out of the 19 poses, 7 poses had the ligand RMSD<1 Å and 13 had ligand RMSD<2 Å. The larger the deviation (higher RMSD) from the CLICK pose, the less favorable is the energy (Table 2) (correlation coefficient of 0.67) (Figure 6). For certain cases, the orientation of the drug (for example Doxorubicin) was different between the docked pose and CLICK predicted pose (after energy minimization) giving higher values for the RMSD (2.52 Å) (Figure 7).

4. Discussion

In this study, we have examined the method for predicting off-target binding sites for a given drug molecule based on structural similarities with already known binding sites. A dataset of 13 experimentally validated drug-proteins pairs with unknown binding sites is used as the validation set for the method. 2 proteins (DDR3 and HRH1) served as an alternate binding site for 6 drug molecules (Cetirizine, Disopyramide, Doxorubicin, Fluoxetine, Paroxetine and Rabeprazole). We restricted our analysis to these pairs either because of the absence of crystal structure of the off target protein or because of the absence of a drug bound protein complex. None of these 6 drugs had a crystal structure with the off target proteins.

For efficient binding of drug molecules and protein, structural and physicochemical complementarity is essential. We matched the structural complementarity by superimposing the drug's binding site (from a PDB structure) and with DEPTH predicted binding site from the off-target protein.

The two proteins that we analyzed in this study (DRD3 and HRH1) are membrane receptors (having an extracellular, transmembrane and intracellular part). Most of the predictions (~80%) onto the off target proteins were on a site with a bound ligand in the crystal structure. Previous studies have shown multiple unrelated ligands bind to the same binding pocket, hence increasing the confidence in the CLICK predicted binding site. In addition, all these sites (with the bound ligand in the crystal structure) were on the extracellular side of the protein, giving higher confidence in our predictions, as most ligands bind to extracellular regions of membrane receptors [283]. 13 out of the 19 CLICK predicted drug protein complexes (excluding the ones for whom docking or energy minimization could not be performed) were similar to that predicted by docking tool Autodock4.2 (<2 Å Ligand RMSD). In certain cases, though the predicted pose looked similar, however, the binding orientation was different leading to high ligand RMSD.

CLICK predicted different binding sites on the same off target protein for Paroxetine and Cetirizine, when the drug binding sites (templates for the match) were extracted from different PDB structures for the structural match. This could be because CLICK predicts a single structural match having the highest structural overlap. The other binding pocket has a lower structural overlap and hence was not predicted. Hence sub-optimal alignments might help improve the prediction but might also lead to increased false predictions.

To conclude, this method can be used for the prediction of off target effects of drugs with known structures of the complex. This technique, however, considers only the structural similarity of the binding pocket while making a match, but along with the structure, the physicochemical complementarity of the binding site with the ligand is important. The physicochemical complementarity of the predicted binding pocket can be scored using various physics/knowledge based scoring potentials. Such potentials can be directly used from the various docking tools, which have their own potentials to score the protein-ligand complex. We can also improve CLICK such that during the structural match the chemistry of the binding pocket can also be matched. Currently, various docking tools exist to predict the binding pose of a ligand on a protein, however, the scoring of the different poses might not be accurate, leading to a prediction of binding sites, which might not be feasible. Improvement of the scoring schemes to score the various sampled posed might help in improving the predictions of the binding pocket of ligands.

Along with characterizing protein-protein and protein-small molecule interfaces, we also predicted and designed inhibitors that would bind to different surface patches on the Nipah proteins, which has been described in detail in the next chapter.

Chapter 9 - Predicting and designing therapeutics against the Nipah virus

1. **Modeling the Nipah proteome**
2. **Designing peptide inhibitors against the Nipah proteins**
3. **Predicting small molecule inhibitors against the Nipah proteins**
4. **Computational analysis of the stability of the inhibitors**
5. **Viability of the inhibitors among different strains of Nipah**

The homology modeling of the Nipah proteome was done by Ankit A. Roy and Kaustubh Amritkar. The *ab initio* modeling was done by Neeladri Sen and Kaustubh Amritkar. The peptide inhibitors against the F and M proteins were designed and analyzed by Neelesh Soni and is a part of his thesis. The peptide inhibitors against the G proteins were designed and analyzed by Neeladri Sen. The binding free energy calculations for the peptide inhibitors were calculated by Shreyas Supekar. The docking studies were done by Tejashree R. Kanitkar. The analysis and binding free energy calculations for the small molecule inhibitors were done by Neeladri Sen and Tejashree R. Kanitkar. The viability of the drugs against the different strains was analyzed by Neeladri Sen and Ankit A. Roy. The modeling of protein-protein interactions was carried out by Sanjana Nair. The web service was designed by Gulzar Singh.

Published - **Sen N.***, Roy A.A.*, Kanitkar T.R.*, Soni N., Amritkar K., Supekar S., Nair S., Singh G., Madhusudhan M.S., Predicting and designing therapeutics against the Nipah virus (2019), Plos Neglected Tropical Diseases, https://doi.org/10.1371/journal.pntd.0007419 [* equal contributions]

1. Introduction

The May 2018 outbreak of the Nipah Virus (NiV) in Kerala, India, claimed the lives of 21 of the 23 infected people [284,285]. This zoonotic pathogen was first detected to infect humans in an outbreak in Malaysia in 1998 [286]. Since then, the mortality rate, especially in the Indian subcontinent has been high with Bangladesh and India reporting 72% and 86% fatalities respectively [287–289]. Though the overall number of fatalities linked with each outbreak has never been more than 105, NiV poses a deadly threat and could potentially become pandemic [290–292]. Considering its high mortality and transmission rates, NiV features in the WHO R&D Blueprint list of epidemic threats that need immediate R&D action [287]. In light of this, the Coalition for Epidemic Preparedness Innovations (CEPI) has extended US$ 25 million support to Profectus BioSciences, Inc. and Emergent BioSolutions Inc. for the development of vaccines against NiV in 2018 [293][294]. NiV is currently classified as a Biosafety Level 4 (BSL-4) pathogen [295] with no licensed drugs or vaccines. Ribavirin and 4-Azidocytidine have been investigated as putative Hepatitis C viral therapeutics [296,297]. However, the efficacy of ribavirin against NiV is unclear [298]. During the 1998-1999 Malaysian outbreak, it showed a 36% reduction in mortality compared to the control group [296]. The control group, however, consisted of patients who were admitted prior to the availability of ribavirin and hence did not necessarily follow the same treatment regimen which could have contributed to higher mortality. It was also administered to patients during the Kerala outbreak and as post-exposure prophylaxis to medical professionals. None of the medical personnel who were administered prophylactic ribavirin acquired the disease. The only two survivors were given ribavirin, although it is not clear how many others also received it as 6 fatalities had been reported before confirmation of disease etiology [284,293]. While ribavirin efficacy *in vivo* is uncertain, 4-azidocytidine trials against Hepatitis C Virus and DenV were halted due to low efficacy and extreme toxicity [298–300]. The drug favipiravir [301] protects against lethal doses of NiV in hamster models and is in Phase II of clinical trials (for influenza, which like NiV is a member of the Paramyxoviridae family). However, *in vitro* studies have shown the emergence of resistance to this drug among members of the influenza family

[302]. A monoclonal antibody, m102.4 [303] acts against the G protein of the virus has been shown to be effective on animal models but human trials are yet to be conducted, though preliminary indications appear promising [304]. In principle, structure based rational design of therapeutics and drugs could help combat the disease and also address the concerns of drug resistance.

The NiV genome encodes six structural proteins viz. Glycoprotein (G), Fusion protein (F), Matrix protein (M), Nucleoprotein (N), RNA-directed RNA polymerase (L), Phosphoprotein (P) and three non-structural proteins named W, C and V [305]. The G protein helps in viral attachment to host cell ephrin receptors and the F protein mediates viral fusion [306–308]. The P protein binds to the N protein and maintains it in a soluble form and increases its specificity towards viral RNA instead of non-specific cellular RNA. The N-P protein complex binds the viral RNA forming the nucleocapsid [309]. This nucleocapsid coated viral RNA acts as a template for viral polymerase L to replicate itself and the host machinery is then utilized to translate its proteins [310]. After replication, the M protein homodimerizes and the dimers form arrays at the plasma membrane. These dimer-dimer interactions induce curvature in the membrane that enables the budding/release of new viral particles [311,312]. The non-structural proteins W, V, and C act against interferon signaling to escape the host immune response [313]. All these proteins are potential targets for rational drug design. Some studies in the recent past have targeted epitopes of these viral proteins [314,315]. However, to the best of our knowledge, the whole proteome modeling of NiV for drug discovery has not been attempted.

In this study, we have used the experimentally determined structures of the NiV proteins and built models for the remaining proteins in trying to find putative lead compounds against the virus. Four proteins (F, G, N and P proteins) have structural data available in the Protein Data Bank (PDB) [278] with varying degrees of structural coverage (Table 1). Using homology based methods; we have extended the structural coverage of these proteins and built models for four of the remaining proteins using either homology modeling or threading/*ab initio* methods. We designed peptide inhibitors targeting interacting sites on G protein-human ephrin-B2 receptor, F protein trimer and M protein dimer. Binding stability of inhibitory peptides was assessed with molecular dynamics (MD) simulations. In addition, to quantify the binding affinities, binding free energies of the designed peptide inhibitors to their respective targets were also evaluated, based on conformations from MD simulations. We have predicted putative drug like molecules using molecular docking that could bind to NiV proteins. The stability of a few of our top docked protein-inhibitor complexes was evaluated based on MD simulations and binding free energy calculations. Our proposed inhibitors should potentially bind to viral proteins and hinder their function thereby preventing viral life-cycle progression. Finally, we have compared the proteomes of Malaysian, Bangladesh and Indian NiV isolates for sequence variations and mapped them onto their protein structures. This enables us to delineate the consequences (if any) of sequential variation among strains on the efficacy of proposed drugs.

2. Methods

2.1. Protein structure modeling

At the time of modeling, the sequence of the Indian strain was not available and so all the modeling was carried out using the Malaysian strain (AY029768.1) [316]. From our experience, using one strain over another would only minimally affect the computed models (Refer Results Section 4 for details on sequence conservation). Monomeric structures of the proteins were built using the homology modeling pipeline ModPipe-2.2.0 [317] and their multimeric complexes were built using MODELLER v9.17 [318,319]. Protein domains/regions that could not be reliably modeled by MODELLER (either greater than zero Normalized DOPE score or with less than 50% structural coverage) were rebuilt using meta-threading and *ab initio* methods on the I-TASSER web server [320]. Models built using I-TASSER were assessed with Normalized DOPE scores along with their C-scores, predicted TM scores and RMSD scores provided by the webserver.

2.2. Prediction of putative small molecules that can bind to NiV proteins

Docking was used to identify putative small molecules that can potentially bind and inhibit the activities of the NiV proteins (G, N, F, P and M proteins). The screening library consisted of 22685 ligands that were the 70% non-redundant set of ~13 million clean drug like molecules of the ZINC database [321,322]. The binding pockets for docking on the targets were predicted using the DEPTH server [8]. Docking was performed using Autodock4 [47], and DOCK6.8 [48,323]. The final energies reported by DOCK6.8 were used for the evaluation and selection of the putative leads.

2.3. Accessing the stability of inhibitory peptides and small molecules against the NiV proteins

One peptide inhibitor was computationally designed against each of the F and M proteins while 2 inhibitors were designed against the G protein. Additionally, 13 small molecules were predicted with high confidence to bind different NiV proteins. Details of the procedures for modeling/predicting peptide/small molecule inhibitors are stated in the results section. MD simulations were carried out in triplicates for all four predicted protein-peptide inhibitor complexes. The simulations were carried out using Gromacs [262,263] with the Amber99SB-ILDN force field [264] using spc/e water model [265]. Parameters for the small molecules were generated using Antechamber [59,266]. The Amber99SB-ILDN force field has been used for the MD simulations of protein-peptide and protein-ligand complexes extensively. In the cases where the small molecule ligand got free of the binding site, we re-simulated the system using the CHARMM27 force field [60], another popularly used molecular mechanics package. We did the second simulation to ascertain that binding was indeed weak. Parameters for the small molecules in the CHARMM27 simulations were generated using SwissParam [324].

A water box whose sides were at a minimum distance of 1.2 nm from any protein atom was used for solvating each of the systems. Sodium or chloride counter ions were added to achieve charge neutrality. Electrostatic interactions were treated using the particle mesh Ewald sum method [325] and LINCS [267] was used to constrain hydrogen bond lengths. A time step of 2 fs was used for the integration. The whole system was minimized for 5000 steps or till the maximum force was less than 1000 kJ/mol/nm. The system was then heated to 300K in an NVT ensemble simulation for 100 ps using a Berendsen thermostat [65]. The pressure was stabilized in an NPT ensemble simulation for 100 ps using a Berendsen barostat. The systems were simulated (NPT) for a maximum of 100 ns (for protein-peptide inhibitor complexes) or for 50 ns (for protein-small molecule inhibitor complexes) where pressure was regulated using the Parrinello-Rahman barostat [268]. Structures were stored after every 10ps. The temperature, potential energy and kinetic energy were monitored during the simulation to check for anomalies.

Free energy of binding of the putative peptide inhibitors/small molecules provides an important quantitative description of its efficacy. In this study, the extensive MD simulations of protein-inhibitor complexes were post-processed to obtain binding free energy estimates using the molecular mechanics Poisson-Boltzmann surface area (MM/PBSA) approach [67,68]. The MM/PBSA method employs an implicit solvation model to estimate the free energy of binding by evaluating ensemble averaged classical interaction energies (MM) and continuum solvation free energies (PBSA) of the protein-ligand complex conformations from the MD trajectories. Snapshots of protein-peptide complexes were obtained at every 100 ps from the last 50 ns of the MD trajectories, thus totaling 500 snapshots. The last 50 ns of protein-peptide inhibitors were selected for MM/PBSA treatment to ensure sampling of equilibrium conformations for appropriate MM/PBSA energy evaluations (Supporting Figures 2, 3, 4 and 5 for RMSD and distance between the center of peptide and protein). The MM/PBSA calculations of the protein-small molecule inhibitors were calculated based on the last 40 ns trajectory with snapshots obtained after every 1000 ps, totaling to 40 snapshots. The MD snapshots were energy minimized for 2000 steps before evaluation of interaction and solvation free energies. The protein and solvent were modeled with dielectric constants of ε =2 and ε =80, respectively. APBS suite [326] and GMXPBSA [327] were used for implicit solvent calculations. In this study, we attempted to calculate the entropic estimate of binding using the interaction entropy formalism [328]. However, converged entropic values with reasonable error estimates for protein-peptide trajectories could not be obtained, which is often the case when evaluating entropic contributions from molecular simulations. We, therefore neglected entropic contributions to the binding free energies, as estimated entropy change upon binding is often negligible and can be ignored for relative binding free energies calculations [69]. The enthalpies of binding obtained from MM/PBSA calculations are reported as binding energies for the protein-peptide complexes.

2.4. Mapping strain variants onto the structure

Protein sequences of 15 different NiV isolates, 7 from Malaysia (AY029768.1,AJ564621.1, AJ627196.1, AY029767.1, AJ564622.1, AJ564623.1, AF212302.2), 3 from Bangladesh (AY988601.1, JN808857.1, AY988601.1) [329] and 5 from India (MH523641.1, MH523642.1, MH396625.1, MH523640.1, FJ513078.1) were retrieved from their translated genomes deposited in the NCBI nucleotide database [330] and were used to identify sequence variations in proteins. We also verified that the translated protein sequences of the Malaysian strain matched with those of the protein sequences deposited in SwissProt [331]. Multiple sequence alignments of the sequences obtained from the 15 isolates were performed with MUSCLE [332]. Positions with amino acid variations were mapped onto the structures. Amino acid variations within 5Å at inhibitor binding sites were identified.

3. Results

3.1. Structural coverage of the NiV proteome

In this study, we first focused on characterizing the structures of the NiV proteins. Partial structures for 4 (F, G, N and P protein) of the 9 NiV proteins are available in the PDB (Table 1). Computationally, we attempted to extend the structural coverage of these 4 proteins and to build models for the remaining 5 proteins using homology modeling (with MODELLER), *ab initio* modeling and threading (with I-TASSER). Model accuracies were carefully scrutinized before attempting to design/predict inhibitors against all possible proteins in the proteome. In this section, we only present the results of homology modeling as all models built using I-TASSER resulted in structures that were not favorably assessed (Normalized DOPE > 0) (Table 2).

The structure of only the pre-fusion state of the NiV F protein (Class I fusion protein) has been determined experimentally (PDB id: 5EVM) (Figure 1A). We modeled the post-fusion state using the structure of the human Parainfluenza Virus 3 (PDB ID: 1ZTM) as a template since it is also a class I fusion protein (Figure 1B). The details about the modeling of the post Fusion F protein and its inhibitor design can be found elsewhere [333,334].

Table 1- List of NiV proteins with their lengths, PDB codes of crystal structures along with their resolution in parenthesis, coverage of crystal structures, coverage of models, additional coverage obtained by the models and the overall sequence coverage. In cases where models have increased the coverage over existing crystal structures, the original coverage is in parentheses.

Protein	**Length**	**X-ray structures (Resolution)**	**X-ray coverage**	**Model coverage**	**Additional coverage**	**Overall coverage (%)**
Pre-fusion F protein	546	5EVM (3.4Å), 1WP7 (2.2Å), 3N27 (1.8 Å)	27 - 482	27 - 482	0	84
Post-fusion F protein		-	-	72-418	347	64
G protein	602	2VSM (1.8 Å), 2VWD (2.25 Å), 3D11 (2.3 Å), 3D12 (3.0 Å)	176 - 602	98 - 597	79	84 (71)
N protein	532	4CO6 (1.7 Ā)	32 - 371	39 - 414	44	72 (64)
P protein	709	4CO6 (1.7 Å), 4GJW (3.0 Å), 4N5B (2.2 Å), 6EB8 (2.5 Å), 6EB9 (1.9 Å)	1 - 38 471 - 578	655 - 709	55	37 (29)
M protein	352	-	-	45 - 352	308	88
L protein	2244	-	-	1814 - 2024	210	9
V protein	456	-	-	1 - 38 87 - 243 313 - 414	297	65
W protein	450	-	-	1 - 38 87 - 243 321 - 391	266	59
C protein	166	-	-	-	-	-

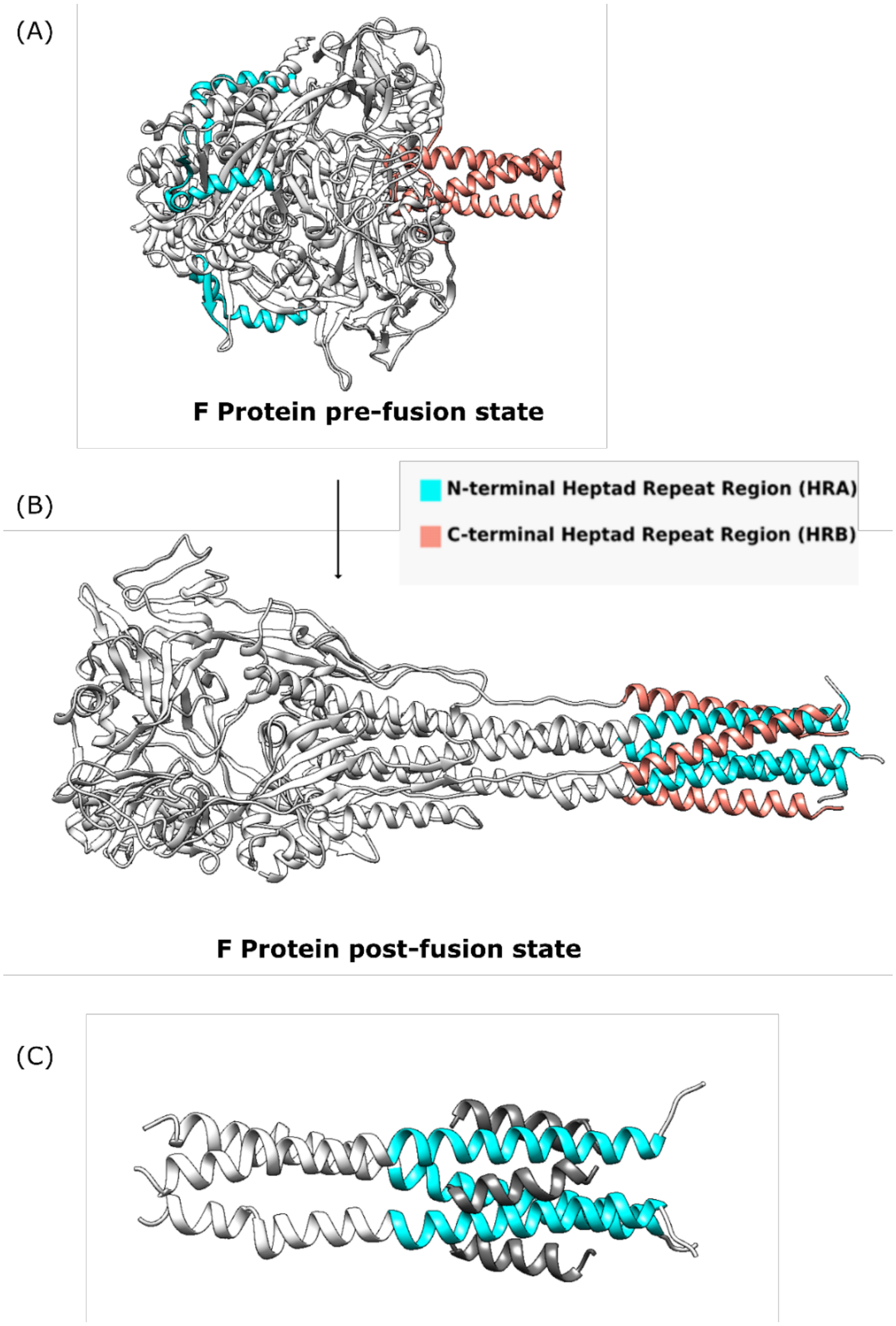

(A)
F Protein pre-fusion state
(B)
N-terminal Heptad Repeat Region (HRA)
C-terminal Heptad Repeat Region (HRB)
F Protein post-fusion state
(C)

Figure 1 – (A) Pre-fusion F protein (PDB ID – 5EVM) (B) Modeled Post-fusion F protein shown in white ribbon. The N-terminal heptad repeats are shown in cyan ribbon, while the C terminal heptad repeats are shown in salmon ribbon. (C) The designed inhibitor against the F protein (shown in grey ribbon) bound to the N-terminal heptad repeat (shown

in cyan ribbon) that would inhibit the transition of the F protein from the pre-fusion state to the post fusion state [333].

Multiple models were constructed for each of the proteins using all available templates. All proteins, except C, had at least one model with a normalized DOPE score of less than or equal to zero. All models built for proteins with existing X-ray structure conferred additional sequence coverage except for the F protein (Table 1). The structural coverage of the N, P and G proteins increased by 8-13% after modeling (Table 1). Overall, we increased the structural coverage of the NiV proteome by 90%, from ~23% (1364 residues) to ~43% (2623 residues). The increased coverage of the proteome helped in designing and predicting inhibitors against the M and F proteins. M protein did not have a structure and all inhibitor studies i.e. designing and docking were carried out on the homology model while the post fusion F protein was used for an inhibitor design.

Table 2- Model quality evaluation of the protein structures built using I-TASSER web server. The best model predicted by I-TASSER (based on their C-Score) has their Normalized DOPE scores and C-scores in bold. TM-scores and RMSDs are only calculated for the best models. L protein was divided into three domains, indicated by their residue numbers in parentheses, and modeled separately.

Protein	**Normalized DOPE**	**C-score**	**Predicted TM-score$^{\$}$**	**Predicted RMSD (Å)**
V	**2.09**, 1.70, 1.99, 1.23, 1.27	**-0.79**, -1.82, -0.32, -3.56, -2.69	0.61	8.9
W	**1.45**, 0.77, 0.75, 1.56, 0.80	**-1.42**, -1.73, -3.07, -4.34, -3.30	0.54	10.4
C	**0.49**, -0.08, -1.53, -0.33, -0.88	**-3.68**, -3.67, -3.29, -4.16, -4.09	0.31	13.6
L$^{\#}$ (14 - 1177)	**0.29**, 0.31, -0.04, 0.01, -0.16	**0.07**, -0.30, -1.87, -1.05, -0.83	0.72	9.1
L$^{\#}$ (1191 - 1435)	**0.52**	**1.1**	0.86	3.6

L[#] (1553 - 1859)	**0.21**, 0.82, 0.95, 2.69, 0.44	**-2.61**, -4.24, -4.52, -4.76, -5.00	0.41	12.4

[#] The protein was built domain wise because I-TASSER has a maximum size limit of 1500 residues. [$] Although models built for V, W proteins and two of the Polymerase L domains had a TM-scores greater than 0.5, none of these models had a Normalized DOPE score less than or equal to zero and therefore were not used further in the study

3.2. Design and stability of protein peptide inhibitor complexes

One peptide inhibitor was designed against the F protein to prevent its transition from pre-post fusion complex. M protein inhibitor was designed to prevent the dimerization of the M protein. Peptide inhibitors can be designed such that they mimic the natural interactions of the proteins. Hence, co-crystals of target proteins has been used to design inhibitors against them [335,336]. The crystal structures of the G protein with other proteins or its homologous proteins have been used to design peptide inhibitors against it. The details about the modelling of post fusion F protein and inhibitor design (Figure 1C) can be found at Sen *et al*., 2019 [333,334].

3.2.1. Peptide Inhibitors of G protein-ephrin interaction

The NiV infection is initiated by the binding of the G protein to the ephrin receptors on the host cell [337] (PDB id: 2VSM). Inhibiting this protein-protein interaction could prevent viral entry. In this study, we have tested the feasibility of using 2 peptides to inhibit the G-protein – ephrin interaction. One peptide (FSPNLW) is the part of the ephrin-B2 receptor that interacts with the G-protein [338]. The other peptide (LAPHPSQ) is a part of a monocolonal antibody, m102.3, that binds [304] to both NiV and Hendra virus. A crystal structure of the antibody bound to Hendra virus G protein (PDB id: 6CMG) was used as a template (79% target-template sequence identity) to construct the antibody-NiV G protein complex. 3D structural models of the speculated G-protein—peptide interactions were also constructed using MODELLER v9.17.

3.2.2. Computational prediction of the stability of the protein-inhibitor complexes

Table 3- Mean and standard deviation of the energy, distance of the center of the FSPNLW inhibitor with the center of the G protein, RMSD of the inhibitor and the protein-peptide binding energies obtained from the three 100 ns MD simulations of G protein-FSPNLW inhibitor complex.

Run	Energy (kJ/mol)		Protein-peptide distance (nm)		RMSD (nm)		Binding energies (kJ/mol)	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
1	-1140578	1550	1.65	0.05	0.12	0.04	-107.2	10.8
2	-1140385	1714	1.66	0.05	0.17	0.04	-96.0	10.8
3	-1140696	1708	1.64	0.03	0.13	0.03	-93.9	12.2
Mean	**-1140553**		**1.65**		**0.14**		**-99.0**	

Three independent MD simulations of 100 ns each were performed to assess the stability of each of the four protein-peptide complexes. The results of the MD simulations of G protein inhibitors have been further discussed. The peptide inhibitors designed against the G protein bind to a predominantly hydrophobic pocket. For each of the trajectories, the total potential energy, the distance between the center of the protein and peptide, RMSD and RMSF of the peptide after superimposition of protein were analyzed and found to be consistent across independent runs (Figure 2-3 and Table 3-4). The protein-peptide complex was stable during the simulation as can be inferred by the peptide RMSDs, peptide RMSFs and the distances between the protein and peptide. The distance of the center of the protein to that of the peptide fluctuated with a standard deviation of 0.03-0.05 nm (Table 3-4 and Figure 2-3) around the average distance. While these measures are all indicative of tight binding, we used the trajectories to determine the binding energy of association using the MM/PBSA protocol. The inhibitors of the F and M proteins bind tightly (~110 kJ/mol) to their targets. However, in the case of G protein inhibitors, the inhibitors FSPNLW and LAPHPSQ bind the G protein with ~-100 and ~-60 kJ/mol, respectively, suggesting that ephrin-B2 receptor based design binds 40 kJ/mol stronger. This trend is also reflected in RMSD/RMSF values (Figure 2-3).

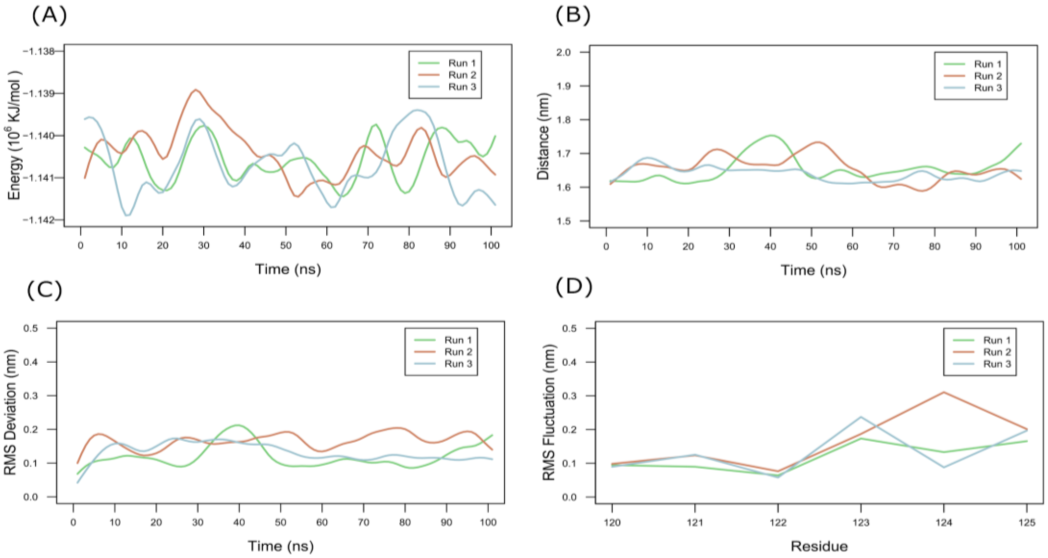


Figure 2- A) Energy of the G protein-FSPNLW inhibitor complex during 100 ns of MD simulation B) Distance of the center of the inhibitor from the center of the G protein during the simulation C) RMSD # of the designed inhibitor during the simulation D) RMSF # of the inhibitory peptide during the simulation. Each of the simulations was run in triplicate, each run being color-coded as red, green and blue. (# RMSD and RMSF were calculated for the inhibitor by superimposing the protein molecule)

Table 4- Mean and standard deviation of the energy, distance of the center of the LAPHPSQ inhibitor with the center of the G protein, RMSD of the inhibitor and the protein-peptide binding energies obtained from the three 100 ns MD simulations of G protein-LAPHPSQ inhibitor complex.

Run	**Energy (kJ/mol)**		**Protein-peptide distance (nm)**		**RMSD (nm)**		**Binding energies (kJ/mol)**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
1	-1124416	1628	1.50	0.03	0.19	0.03	-61.0	12.9
2	-1124513	1695	1.72	0.05	0.21	0.05	-58.9	11.4
3	-1124582	1578	1.54	0.05	0.17	0.03	-65.1	12.2
Mean	**-1124504**		**1.59**		**0.19**		**-61.7**	

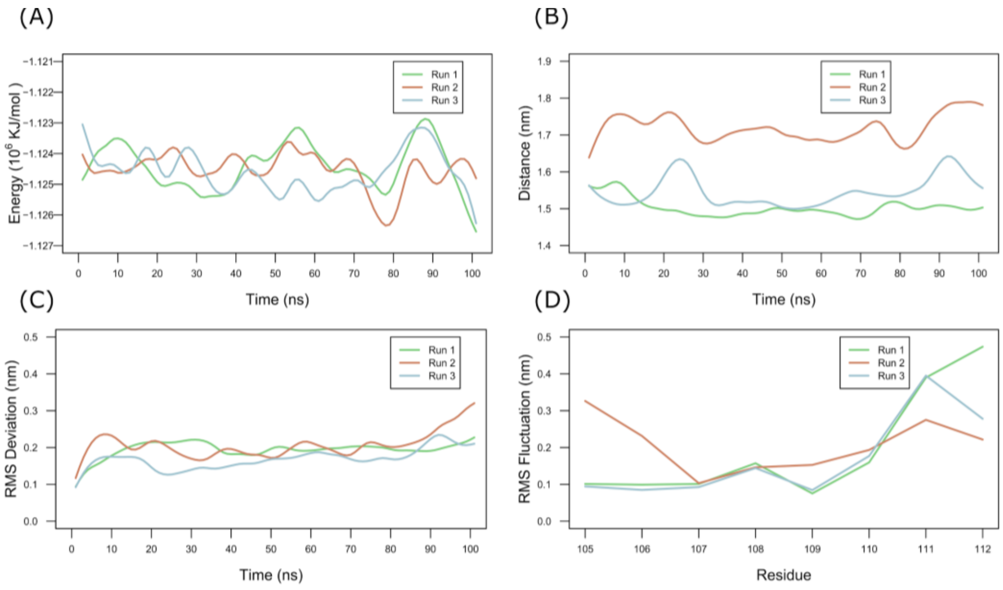


*Figure 3- A) Energy of the G protein-LAPHPSQ inhibitor complex during 100 ns of MD simulation B) Distance of the center of the inhibitor from the center of the G protein during the simulation * C) RMSD # of the designed inhibitor during the simulation D) RMSF # of the inhibitory peptide during the simulation. Each of the simulations was run in triplicate, each run being color-coded as red, green and blue. (* The distance between the inhibitor and the protein was consistently 0.2 nm higher for one of the replicates. This is because we considered the center of the protein as the all-atom center and fluctuations in the side chains can explain the deviation. # RMSD and RMSF were calculated for the inhibitor by superimposing the protein molecule)*

3.3. Prediction of putative small molecules that can bind to NiV proteins

The crystal structures of the G, N, P and F proteins were used in docking studies to find plausible small molecule inhibitors. First, we predicted the plausible binding pockets on each of the proteins using the DEPTH server. All 12 binding sites (as predicted by DEPTH) were used to screen 22685 drug like molecules from the 70% nonredundant ZINC database of clean drug like molecules using two different docking tools, DOCK6.8 and Autodock4. To corroborate our predictions, we measured the RMSD between the same ligand (in the common list) as docked by the two different tools (top 5 poses predicted by Autodock4 were compared to the top pose predicted by DOCK6.8), after superimposing the proteins. This measure is referred to as RMSD_lig. 10 unique drug like molecules had an RMSD_lig of less than 0.15 nm between their docked poses and were the top 100 predicted ligands by both the docking tools (based on the energy calculations by the docking tool).

There are however 3 molecules that are of interest despite their relatively large RMSD_lig values. The molecule ZINC91252717 is predicted as the best binder to the P protein by Autodock4 (binding energy of -14 kcal/mol) and the second best binder by DOCK6.8 (grid score of -71). These scores were among the best achieved during this docking exercise. We selected ZINC00814199 that was docked onto the M protein and was similar to ZINC01725633, which in turn formed 14 and 8 hydrogen bonds with Autodock4 and Dock6.8 respectively. ZINC00814199 was within the top 14 ranked compounds by both methods. Lastly, the hydrophobic molecule ZINC63411510 is predicted to bind the G protein on its ephrin-B2 binding interface. Though both docking methods identify this site, the docking poses are different (RMSD_lig of 0.8 nm). We hypothesize that the hydrophobic nature of the binding pocket and its size is contributing to the difference in docked poses.

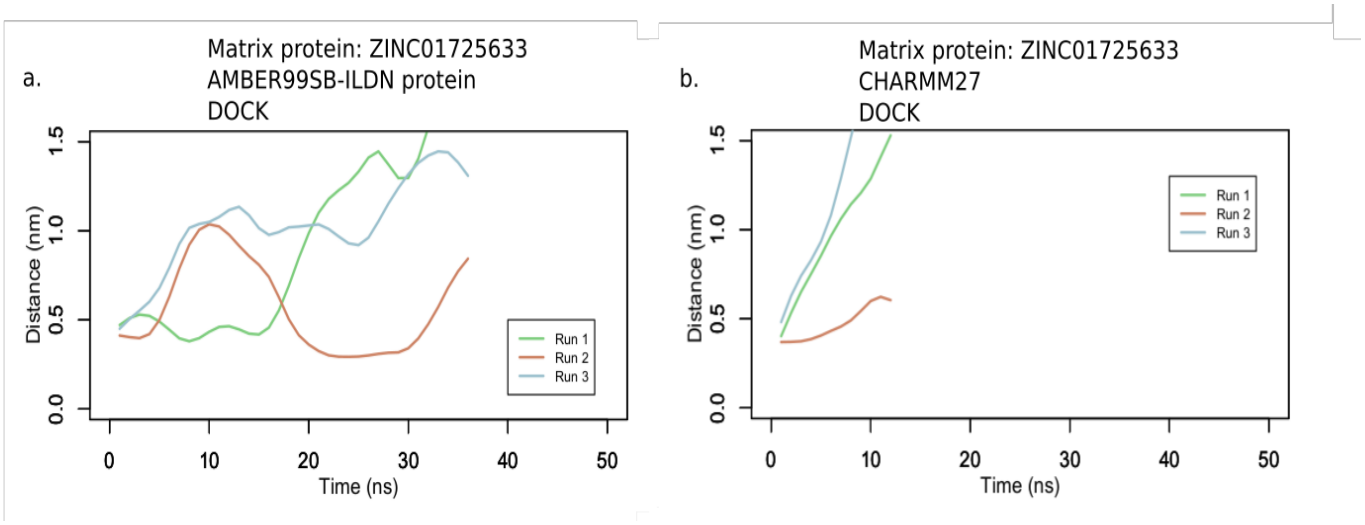
Interestingly, a known drug (ZINC04829362), an antiasthmatic and antipsoriatic among other uses, binds to a pocket of the N protein with RMSD_lig of 0.085 nm. Another drug (ZINC12362922) used in the treatment of depression and Parkinson's disease (101) also binds the N protein with RMSD_lig < 0.15 nm.

3.3.1. Computational prediction of the stability of the protein-inhibitor complexes

To assess the stability of the 13 protein-small molecule ligand complexes, we carried out three independent MD simulations of 50 ns each, using the AMBER99SB-ILDN force field [264]. 10 of the 13 ligands have RMSD_lig values of less than 0.15 nm and were in the top 100 scoring models as predicted by both the docking tools. For these ligands, the simulations were carried out starting with the DOCK6.8 predicted pose. For each of the trajectories, the distance of the center of the small molecule ligand to the center of the binding pocket (based on the starting structure after NPT equilibration) was monitored (Figure 4-5). The triplicate MD simulations were terminated if this distance in 2 of the 3 trajectories exceeds 1 nm from its starting value and these complexes were then re-simulated using the CHARMM27 force field (this was restricted to cases where RMSD_lig < 0.15 nm). This happened in 5 of the 10 cases. For the 3 ligands with RMSD_lig > 0.15 nm simulations were carried out starting with both the DOCK6.8 and Autodock4 predicted poses.

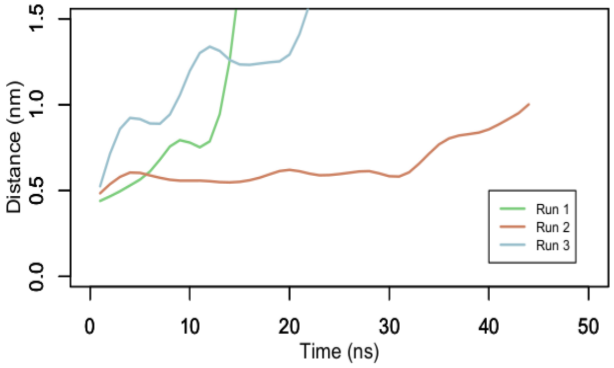
We computed binding energies for the protein-ligand complexes using MM/PBSA (as mentioned in Methods section 3). 9 of the binding energies were computed to be negative in at least one of the replicates (3 for N protein, 4 for P protein, 1 for G protein and 1 for M protein). In one case (P protein-ZINC7262705 ligand), the binding energy with the CHARMM force field (after the AMBER simulation was terminated) was computed to have positive energy. In 3 cases (1 for N, P and M protein each) the ligand did not remain bound to the protein in both CHARMM and AMBER simulations (Table 5-6).

The two known drugs, ZINC04829362 and ZINC12362922 remained bound to the N protein in all 3 replicates with negative binding energies in at least 2 of the trajectories. The important druggable site on the G protein (that recognizes the ephrin receptor on the host), the ligand remained bound in all 3 replicates when starting with the Autodock4 bound pose with negative binding energies.



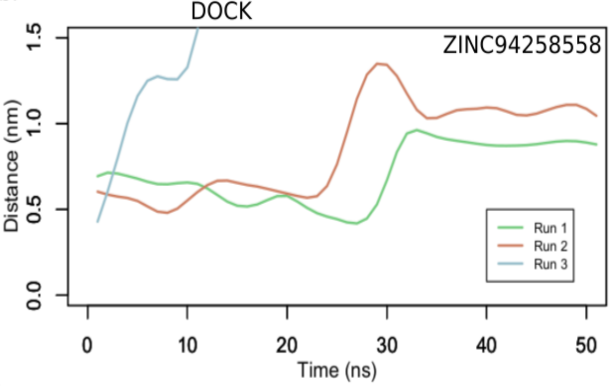
a.

Nucleoprotein: ZINC94258558
AMBER99SB-ILDN protein
DOCK



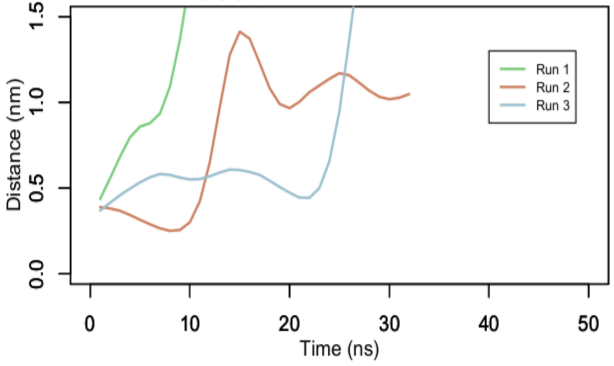
b.

Nucleoprotein: ZINC94258558
CHARMM27
DOCK



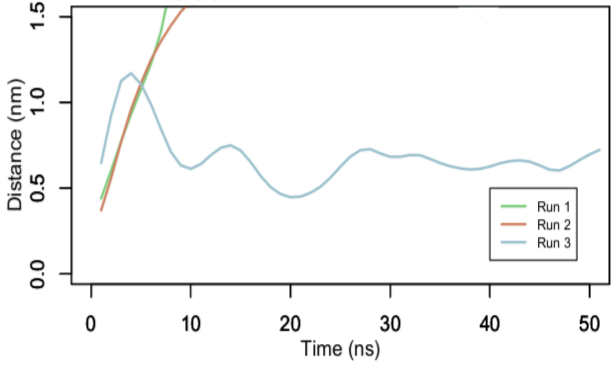
a.

Nucleoprotein: ZINC73641145
AMBER99SB-ILDN protein
DOCK



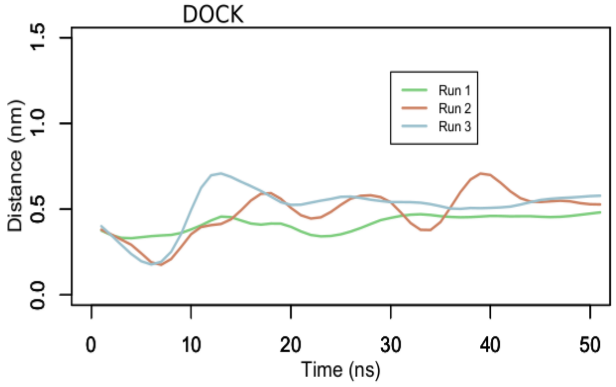
b.

Nucleoprotein: ZINC73641145
CHARMM27
DOCK



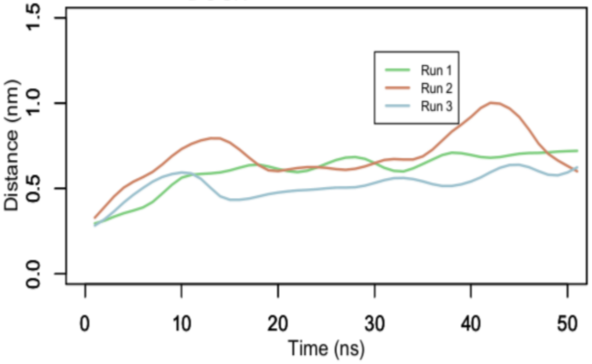
a.

Nucleoprotein: ZINC1236222
AMBER99SB-ILDN protein
DOCK



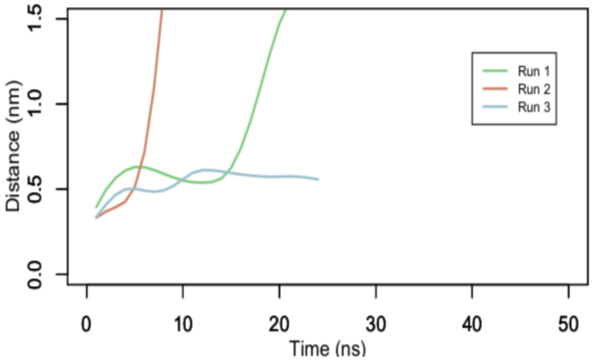
a.

Nucleoprotein: ZINC94258558
AMBER99SB-ILDN protein
DOCK



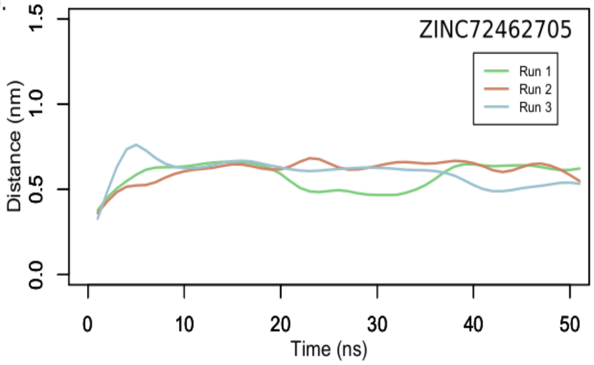
a.

Phosphoprotein: ZINC72462705
AMBER99SB-ILDN protein
DOCK



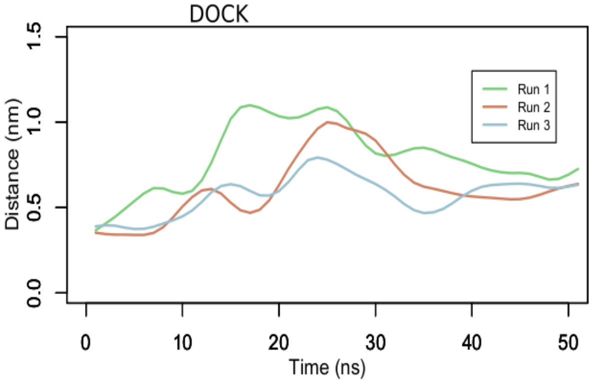
b.

Phosphoprotein: ZINC72462705
CHARMM27
DOCK



a.

Phosphoprotein: ZINC86098248
AMBER99SB-ILDN protein
DOCK



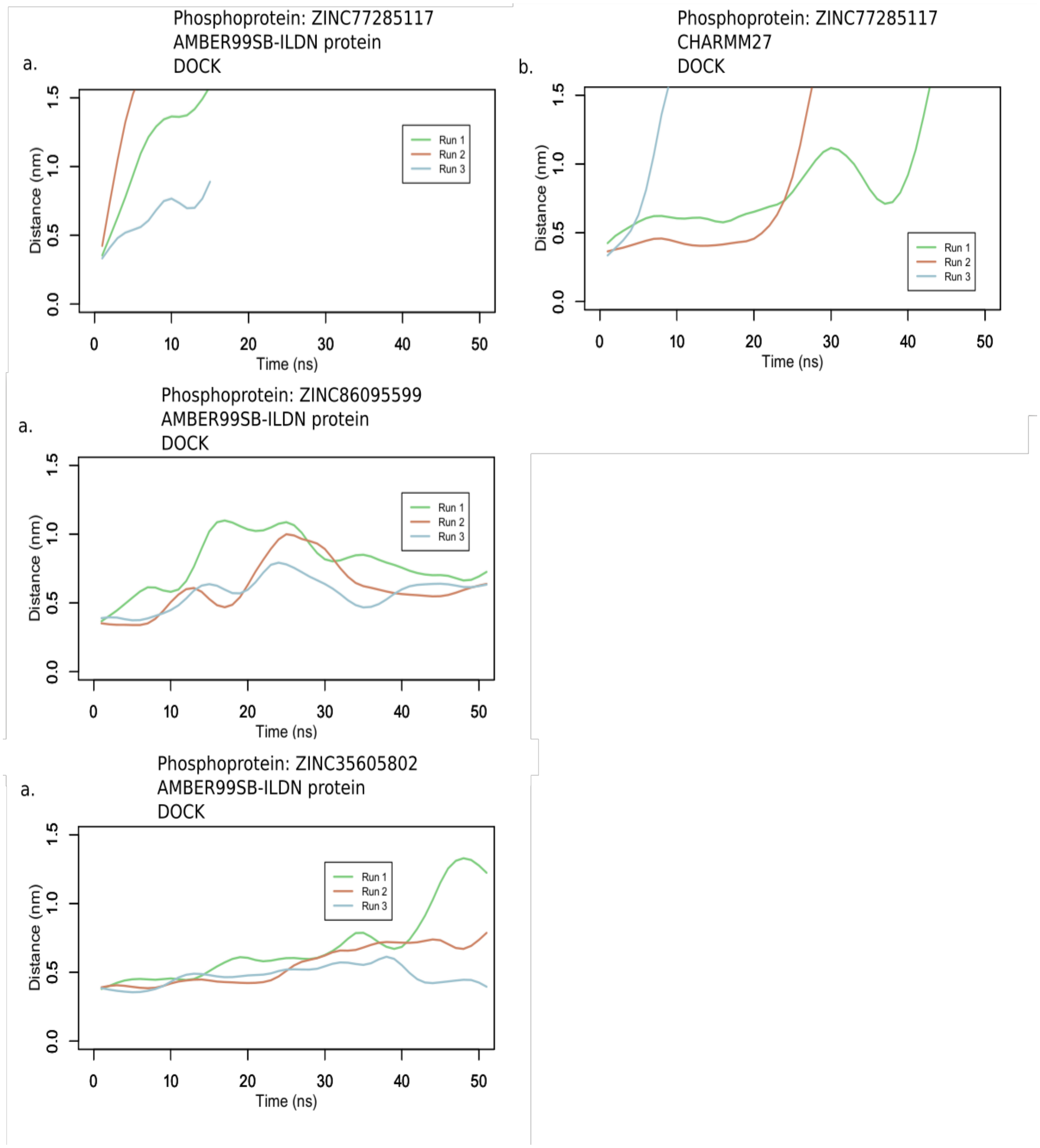


Figure 4 - Distance of the center of the ligand from the center of the binding site (calculated based on the residues within 5Å of the first snapshot after NPT equilibration) during the simulation. The identity of the ligand, force field and docking strategy used and the target protein has been indicated above each plot.

Table 5 – Binding free energy as predicted using MM/PBSA calculations from molecular dynamics simulations carried out using AMBER and CHARMM force fields for top 10 ligands predicted against N, P and M proteins. The binding free energies were not calculated (depicted by -) when the ligand left the binding site in at least 2 out of 3 replicates. CHARMM was only used to run molecular dynamics simulations when the ligand left the binding pocket in AMBER simulations.

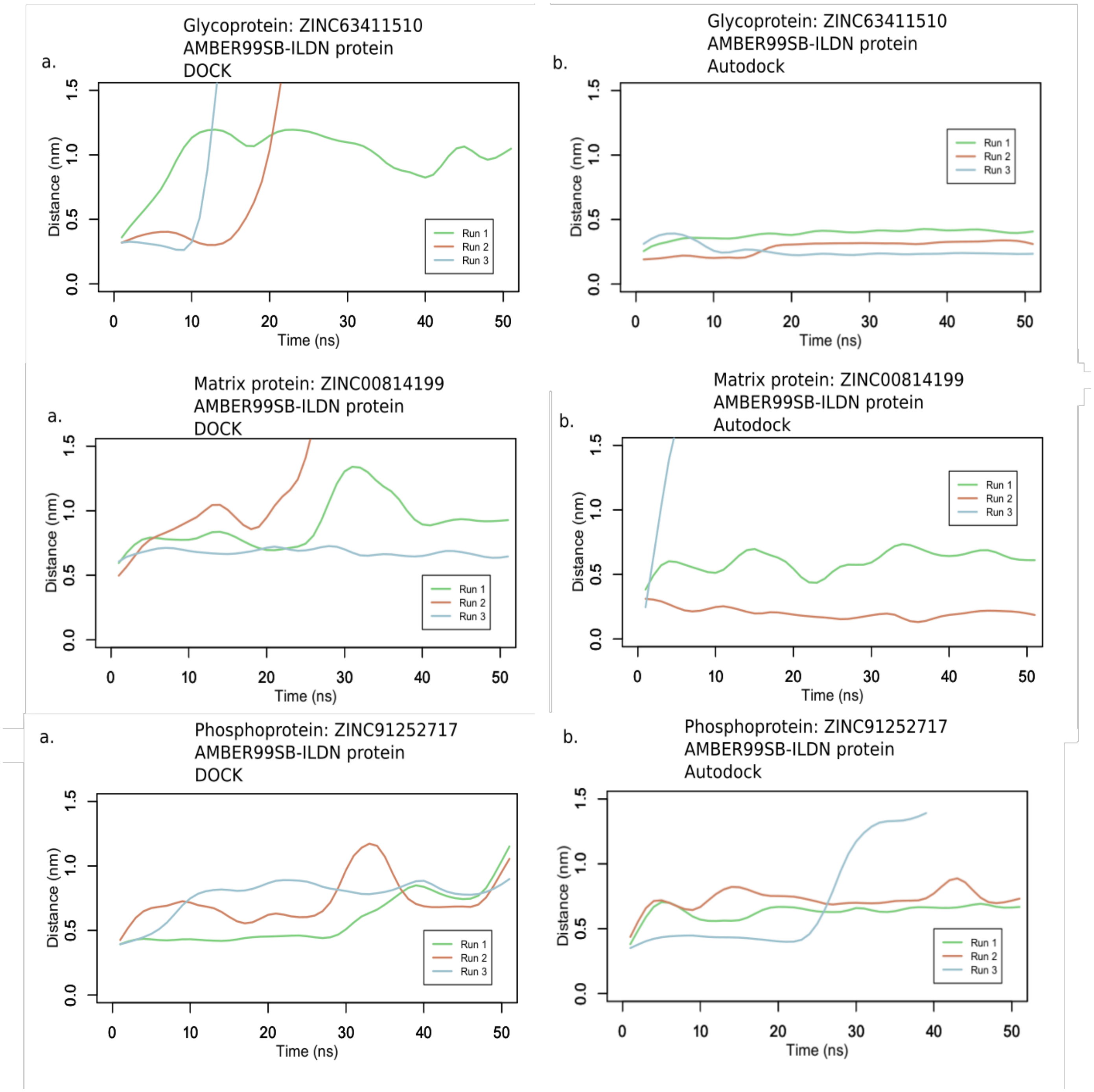
ZINC ID	**Protein**	**Replicate**	**Binding free energy as predicted during**	
			AMBER simulation (kJ/mol)	**CHARMM simulation (kJ/mol)**
ZINC94258558	N	1	-	-114+/-10
		2	-	-86+/-4
		3	-	-
ZINC73641145	N	1	-	-
		2	-	-
		3	-	-
ZINC12362922	N	1	-96+/-8	
		2	-100+/-10	
		3	-69+/-6	
ZINC04829362	N	1	-37+/-7	
		2	86+/-7	
		3	-101+/-8	
ZINC72462705	P	1	-	106+/-4
		2	-	86+/-4
		3	-	98+/-5
ZINC86098248	P	1	39+/-5	
		2	-65+/-7	
		3	37+/-3	
ZINC77285117	P	1	-	-
		2	-	-

		3	-	-
ZINC86095599	P	1	-196+/-7	
		2	-149+/-10	
		3	-153+/-14	
ZINC35605802	P	1	-14+/-0.5	
		2	1+/-1	
		3	-98+/-8	
ZINC01725633	M	1	-	-
		2	-	-
		3	-	-

Table 6 - Binding free energy as predicted using MM/PBSA calculations from molecular dynamics simulations carried out using AMBER force fields for 3 ligands predicted against G, M and P proteins for both the predicted DOCK and Autodock pose. The binding free energies were not calculated (depicted by -) when the ligand left the binding site in at least 2 out of 3 replicates.

ZINC ID	**Protein**	**Replicate**	**Binding free energy as predicted from (kJ/mol)**	
			DOCK pose (kJ/mol)	**Autodock pose (kJ/mol)**
ZINC00814199	M	1	-153+/-6	-119+/-8
		2	-	-184+/-3
		3	-203+/-6	-
ZINC63411510	G	1	-	-44+/-4
		2	-	-79+/-4
		3	-	-59+/-4

ZINC91252717	P	1	-158+/-9	-187+/-8
		2	-256+/-10	-196+/-8
		3	-251+/-7	-



a.

Glycoprotein: ZINC63411510
AMBER99SB-ILDN protein
DOCK

b.

Glycoprotein: ZINC63411510
AMBER99SB-ILDN protein
Autodock

a.

Matrix protein: ZINC00814199
AMBER99SB-ILDN protein
DOCK

b.

Matrix protein: ZINC00814199
AMBER99SB-ILDN protein
Autodock

a.

Phosphoprotein: ZINC91252717
AMBER99SB-ILDN protein
DOCK

b.

Phosphoprotein: ZINC91252717
AMBER99SB-ILDN protein
Autodock

Figure 5 - Distance of the center of the ligand from the center of the binding site (calculated based on the residues within 5Å of the first snapshot after NPT equilibration) during the simulation. The identity of the ligand, force field and docking strategy used and the target protein has been indicated above each plot.

3.4. Sequence variations in NiV isolates

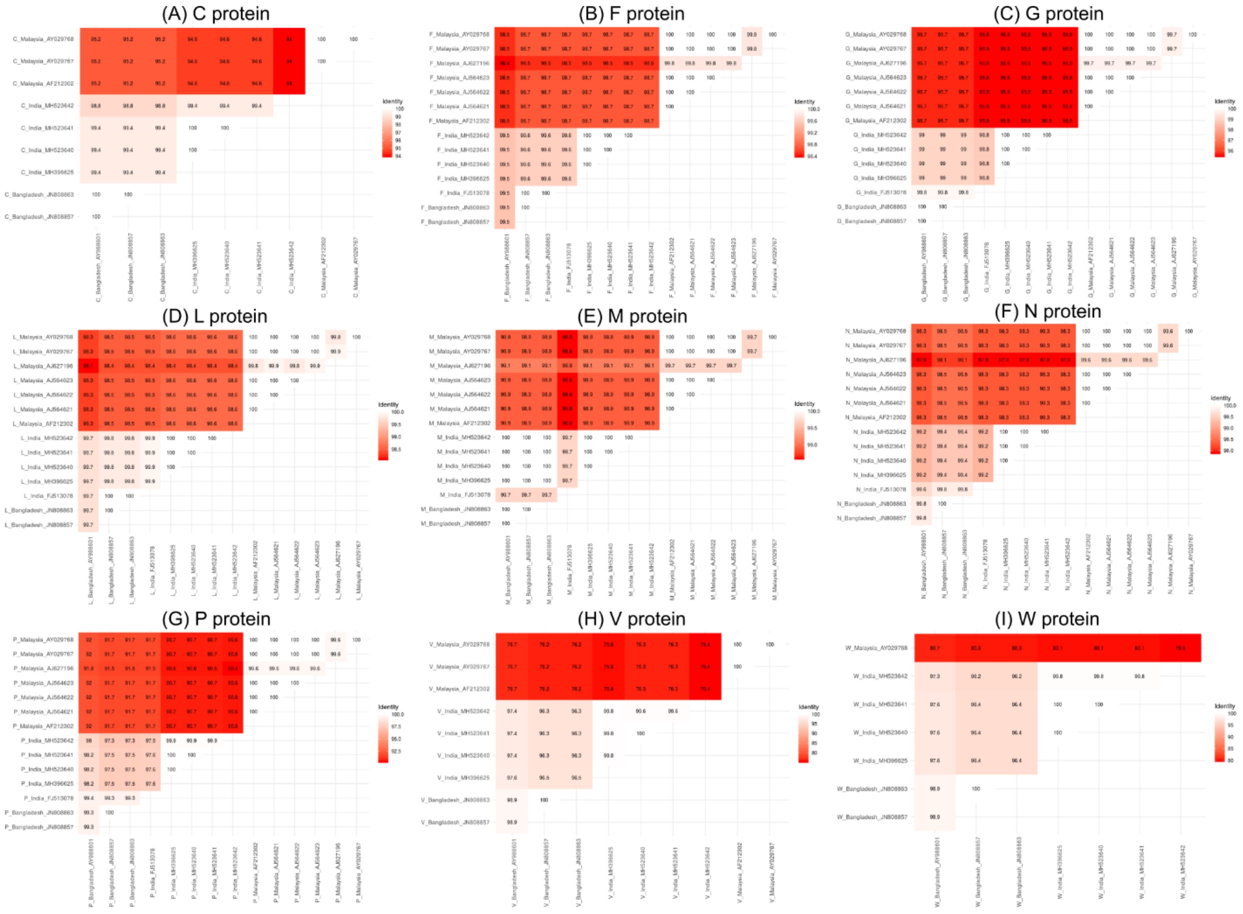


Figure 6 - Heatmap showing the sequence conservation between the different strains of NiV for (A) C protein (B) F protein (C) G protein (D) L protein (E) M protein (F) N protein (G) P protein (H) V protein (I) W protein. The color gradient represents sequence conservation where white indicates 100% conservation and redder shades indicate lesser sequence conservation. The labelling convention is Protein_Country_Genome-accession code.

At the time of modeling the NiV proteins, the sequence data from the 2018 outbreak was not available (77). Hence, all the modeling was done by considering that sequence of the Malaysian strain. We rationalized that as the Malaysian and Bangladeshi/Indian strains shared a high degree (79-99 %) of sequence similarity, structural models using sequences of one strain would be applicable to the other, which is the basis of comparative modeling. However, we wanted to assess whether the efficacy of the designed/proposed therapeutic molecules would be affected by observed sequence variations between the different strains (7 Malaysian, 3 Bangladeshi and 5 Indian) of NiV. The amino acid variations were mapped onto their respective structures. All protein sequences are of equal length except the V protein whose length varies between the different strains. The V and W protein have the least sequence conservation (~79%) while the M protein is the most conserved (98.6%). A general observation is that the Bangladeshi and Indian strains are more similar to one another than they are to the Malaysian sequences (Figure 6).

We mapped the sequence variations onto all the protein structures/models that were used for peptide inhibitor design and drug docking. No variations in the sequence were found close to the peptide inhibitor binding sites on the F, M and G proteins. We found 1 (Lys236Arg), 2 (Asp188Glu, Gln211Arg), 1 (Asp252Gly) and 1 (Ile331Val) variations close to the docking sites on G, N, F and M protein respectively. All the mutations (except for Asp252Gly on F protein) on the binding site were conservative (similar physicochemical properties and BLOSUM62 score >= 0) and hence are conjectured not to affect the interactions between the protein and the inhibitor. Though there is a non-conservative change (ASP252Gly) in one of the drug/inhibitor binding sites of the F protein, this position is not involved in H-bonding with the ligand. Hence the binding of the inhibitor to the protein is probably not going to be affected. Among the top 13 shortlisted ligands, ZINC04829362 and ZINC12362922 bound to N protein and ZINC63411510 bound to G protein were within 0.5 nm of the amino acids that showed variations. No single sequence variant we have studied appears to show that the drug binding would be directly affected.

3.5. Web service and Database

We have archived all structures/models of NiV proteins and their inhibitor bound complexes in a consolidated database at http://cospi.iiserpune.ac.in/Nipah. The data at this site lists details of modeling, docking features and multiple sequence alignments (between the various NiV strains) such as template PDB code, target-template sequence identity, model quality assessment score, docking energies, docking rank and the RMSD_lig between the docking poses. The data from the webservice can be used by general public in further studies related to NiV.

4. Discussion

NiV is a deadly zoonotic virus with a mortality rate of 72% and 86% in Bangladesh and India respectively. There are no approved drugs/therapeutics against NiV. The overarching aim of this study is to computationally design inhibitors and predict small molecule drugs against NiV proteins. To design/predict therapeutic molecules to act against NiV, we characterized all of its proteins. As a part of this effort, we constructed partial models of 5 NiV proteins viz., M, L, V, W proteins along with the post-fusion conformation of the F protein. The structure of the post-fusion conformation of the F protein is modeled for the first time in this study. Our model is based on the post-fusion structures of another class I fusion protein from Human Parainfluenza virus 3.

Our efforts have increased the coverage of existing structures of the G, N and P proteins (by 13%, 8% and 8% respectively) by modeling a fraction of their unresolved residues. No reliable models could be generated for the C protein. Effectively, we doubled the number of amino acids in the NiV proteome that were structurally characterized. While we aim to use these models to predict/design inhibitors, we believe that many of our models are by themselves quite insightful. They could serve as templates for future structure-guided drug designing efforts against members of the Paramyxoviridae family. We attempted to build complexes of the viral and host protein (host cathepsin-L with NiV F protein and host AP3-B1 with NiV M protein) to target the interactions for inhibitor design. However, we were unsuccessful in making reliable models of host-pathogen protein-protein interaction complexes. With improvements to protein-protein docking methods, the quality of such models of complexes could be improved, which in turn would help in better targeting host-viral interactions.

We next used these models to design 4 peptide inhibitors against the F, M and G proteins. The inhibitor against F protein would putatively prevent the pre to post-fusion transition of the F protein, a crucial step for viral entry (not described in this chapter). Our model of the post-fusion conformation of the F protein was crucial in designing this inhibitor. Another inhibitor against the M protein was designed such that it would prevent the dimerization of the protein, hence preventing the budding process (not described in this chapter). The two inhibitors against the G proteins were selected such that they bind to the ephrin receptor binding pocket, preventing viral attachment to the host cell. The peptides here mimic the ephrin-B2 protein and an antibody (m102.3) that are bound at the same site. We conjectured that these peptides would competitively inhibit the G protein from binding the host ephrin receptors. All of these protein-peptide systems were subjected to triplicate runs of 100 ns MD simulations to assess interaction strengths. The distance of the center of the inhibitor and the peptide fluctuates with a standard deviation of 0.03-0.09 nm from the mean distance, indicative of the inhibitor remaining in the binding pocket. The inhibitors against the F and M proteins also had stable hydrogen bond associations in the MD trajectories. Binding affinity calculations suggest that three of the designed putative inhibitors bind tightly (~100 kJ/mol) to their targets, making them promising leads against NiV proteins.

We screened a set of drug like molecules in a docking exercise to identify potential small molecule inhibitors of NiV. The screen consisted of 22685 compounds of the 70% non-redundant set of clean drug like molecules of the ZINC library. The docking onto the NiV proteins was done using two different docking programs, Autodock4 and Dock6.8. Empirically, we chose the top 150 ligands from each of the two methods and selected those that were common between them. This resulted in 146 compounds that bound the G, N, P, F and M proteins of NiV. As a more stringent test, we whittled down this list to only include those molecules that were docked in similar poses (empirically chosen RMSD of 0.15 nm or smaller) on the same binding site and were in the top 100 scored models by both docking schemes. Hence, we predicted 10 compounds that would inhibit the N (5), P (4) and M (1) proteins of NiV. In addition, we also included 3 drugs to the list that did not clear the criteria explained above. These drugs include one that binds the G protein on its ephrin binding interface and two others which bind to P and M proteins. The 13 ligand bound protein complexes were subjected to triplicate MD simulations (50 ns each) to gauge the stability of the association. In 9 of the complexes, at least one of the trajectories was evaluated to have favourable (negative) binding energy. While the simulations and the energy calculations that follow are not to be construed as indicators of binding strength, they do provide the same general trends and give pointers and/or boost our confidence in the binding efficacy of the ligand-protein complex. Only 3 of the 13 ligands consistently moved away from the original predicted binding pocket even when the simulations were repeated using a different force field. In one other case, though the protein-ligand complex remained conformationally stable throughout the course of the triplicate trajectories, our energy estimates of this interaction were unfavorable (positive energy). In the absence of experimental validation, which we seek to do next, these MD simulations serve as indicators of the viability of the ligands to bind the viral proteins.

Of the 13 binding tests, two are in interface regions, one in the M protein dimer interface and another on the ephrin receptor recognition site of protein G. When not bound to these two sites, the ability of the ligands to functionally impair the virus would only be known with experimental testing. The most important aspect of the docking study is that the molecular screen consists of known drugs or drug like compounds. The implication is that a few of our proposed inhibitors could be readily tested and repurposed. For instance, we have identified Cyclopent-1-ene-1,2-dicarboxylic acid (ZINC04829362) as an inhibitor of the NiV N protein. This compound is a known drug prescribed for antiasthmatic and antipsoriatic among other disorders. Another example is Bicyclo[2.2.1]hepta-2,5-diene-2,3-dicarboxylic acid (ZINC12362922) that we propose also inhibits the N protein, is a drug prescribed against depression and Parkinson's disease. Both these ligands have a negative binding free energy in at least 2 of the 3 replicates.

In all our computational predictions, an independent scoring scheme(s) was used to evaluate results. MD simulations were always carried out in triplicate and sometimes using different force fields. In short, we have taken care to ensure cross-validation of our computations to whatever extent practically possible. We cannot overemphasize the importance of these computational predictions, especially for swift acting potent viruses as NiV where mortality rates are high.

Finally, we assessed how effective our proposed inhibitors would be against different strains of the virus and assess the risk of the virus getting drug resistant. For this, we studied 3 Bangladeshi, 7 Malaysian and 5 Indian strains and inferred the variations between the various strains from their multiple sequences alignment. Further, we investigated whether such changes would affect inhibitor binding. Here, we narrowed the changes only to those residues that were in direct contact (< 0.5 nm) from the inhibitors. We precluded the possibility of allosteric interactions. None of the residues contacting the peptide inhibitors showed any variations in their sequence. Only 5 residue positions that were involved in binding the drug like inhibitors were changed between the different strains. 4 of these changes are conservative substitutions where the nature of the mutated residue is not deemed to change the binding property of the protein to its inhibitor. Only 1 amino acid change of Asp252Gly of the F protein is a non-conservative change, however, the Asp is not involved in hydrogen bonding with the ligand. We conclude that it is likely that the proposed inhibitors would be potent against all strains of the virus Nipah and other zoonotic viruses that pose a serious epidemic threat. Computational approaches can help identify/design inhibitors that could be rapidly tested or even deployed as they may be drugs previously licensed for other uses. Our study also has connotations for related viruses such as Hendra and other Paramyxoviruses. Importantly, our models and the web pages we have created could be modified to serve as a portal to study the epidemiology of the virus should there be further outbreaks.

The previous chapters dealt with the study and characterization of protein-protein and protein-small molecule interfaces. In this chapter, we describe designing peptide inhibitors and predicting small molecule inhibitors that would bind to Nipah proteins, inhibiting them. In addition to these, we also characterized residue environments in proteins, which we described in the next chapter.

Chapter 10 - Characterizing residue environments in proteins to develop environment dependent substitution matrices

1. **Different residues prefer different environments in protein**
2. **Creation of amino acid substitution matrices at different depth levels**
3. **Comparison of the different amino acid substitution matrices to check difference in substitution patterns in different environments**
4. **Utilizing the matrices to predict deleterious mutations**

This work was done in collaboration with Nida Farheen.

Published - Farheen N.*, **Sen N.***, Nair S., Tan K.P., Madhusudhan M.S. (2017), Depth dependent amino acid substitution matrices and their use in predicting deleterious mutations, Progress in Biophysics and Molecular Biology, doi: 10.1016/j.pbiomolbio.2017.02.004 [* equal contributions]

1. Introduction

The 3D structure of a protein is key to determining its function or biological role. The primary sequence of a protein folds into a particular 3D shape, given a particular set of conditions [339]. The number of shapes that proteins fold into is limited, and by various estimates, is of the order of 1,000 [340]. It is believed that the native fold of a protein is its minimum energy conformation [341]. This conformation is solely dependent on its amino acid sequence. Any changes to the amino acid chain would result in a perturbation of this 3D structure.

With respect to mutations of amino acids in proteins, one of the key questions to answer would be to determine if the mutations could change the conformation of the protein sufficiently enough to affect function. Note that the function of a protein could also be affected by mutations that need not necessarily change its 3D shape. For the purposes of this study, we are only interested in those mutations that affect the stability of the 3D structure of the protein. Our motivation arises from the fact that ~80% of the Mendelian-disease-associated single mutations are a consequence of protein destabilization [342].

The effects of a single point mutation in a protein sequence are felt most acutely by its immediate spatial neighbors. In essence, every single amino acid in a protein is embedded in its own characteristic microenvironment. In this study, we are going to utilize this feature of residue depth [7] to characterize residue microenvironment and to determine how the immediate neighborhood of an amino acid is affected by mutation.

An observation that is crucial to our study is that the amino acid abundance at different depth levels is markedly different (Figure 1). The depth preferences of some of the amino acids could be categorized based on the nature of their side chains. The polar amino acids (N, Q, H, K, R, D, E) show a sharp decline in their abundance with an increase in depth. The hydrophobic amino acids (V, I, L, M, F) have an increase in abundance with an increase in depth. The amino acids S, T and G also behave like the polar amino acids only that the decrease in abundance is not sharp. The amino acids A, Y and W have their maximum abundance in an environment that is neither deep nor shallow. Cysteines, though considered polar by some studies, show the same behavior as non-polar residues while Prolines, which is sometimes considered apolar, displays the same tendency as polar residues. It is clear that with stratification by depth, relative abundances of amino acids vary. We use this fact to compute the likelihood of amino acid substitutions. Note that these trends are best observed by parameterizing the protein environment using Depth as opposed to SASA [7].

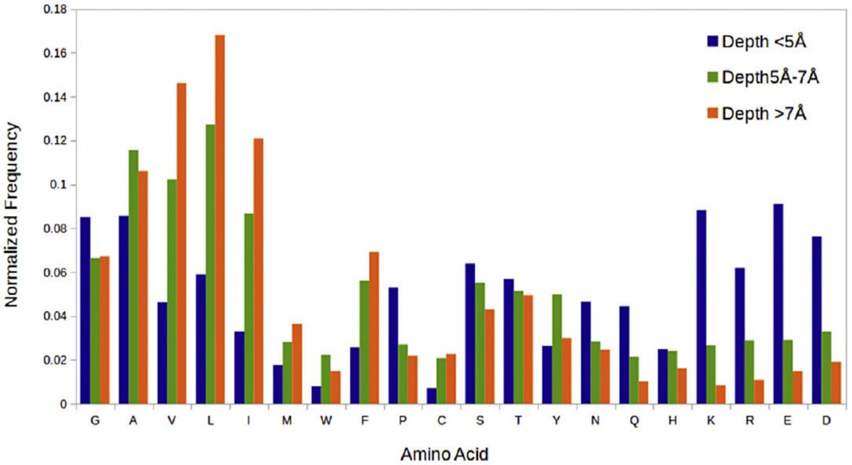


Figure 1- Histograms of the relative abundance of amino acids at different depth levels <5 Å (blue), 5-7 Å (green) and >7 Å (orange). Normalization of the abundance was done depth wise.

Computations of substitution likelihoods have been well documented [343,344] and widely used from aligning two sequences to one another to detecting homologous sequences [345,346]. The traditional substitution likelihoods bundled into the so called substitution matrices, such as PAM and BLOSUM, are however devoid of any context. In fact, amino acid substitutions involving a pair of residues is averaged over several different environmental and fold contexts. In the light of secondary structure prediction and 3D structure modeling/evaluation exercises, amino acid environments have to be described accurately. This implies that an amino acid substitution table that considered not just the likelihoods of pairwise substitutions but also the environmental context and/or protein categories would be best suited for the purpose [22,347–363]. As depth is a concise measure of amino acid environment, we developed depth dependent substitution matrices that can capture the substitution likelihoods in different environments. In this study, we have categorized amino acids into 3 distinct environments – residues that lie in depth ranges <5 Å, between 5 to 8 Å and > 8 Å. As described earlier, the relative abundance of the residues in these depth environments is different and hence it is likely that their substitution rates would also differ in the different contexts. A symmetric substitution matrix was computed for each of the depth environments considering a log-odds ratio of observed over expected frequencies.

The efficacy of the matrices was tested by using it to predict the destabilizing effects of single point mutations in protein sequences. Other computational methods that address this question use a combination of sequence and structural information to deduce the effect of the mutation. The approaches for predicting mutational stability could be divided into sequence-based and structure-based methods. Sequence based methods such as SIFT[364], Polyphen[365] and SuSPect[366] rely on multiple sequence alignments of proteins to extract substitution trends from sequence profiles. Polyphen and SuSPect also utilize structure features. SuSPect incorporates the extraction of information from protein domains, PSSM, protein-protein network interactions, position-specific known mutants and is one of the methods compared to in this study. Most structure-based methods are based on machine learning that fit a non-linear function to experimental data. We have compared ourselves to several such methods including I-Mutant[367], Automute [368], mCSM [369], SDM[370] and DUET[371]. I-Mutant incorporates pH, temperature and mutation type as features in its support vector machine. Automute is based on a multi-body statistical potential that combines energy-based and machine learning approaches. mCSM [369] uses a graph metric to summarize physiochemical interactions within a cutoff distance and train them with a Gaussian process regression model. SDM[370] is a statistical method that builds an environment dependent substitution matrix. DUET[371] is a meta-algorithm combining mCSM and SDM.

Predictions made by the depth dependent substitution matrices were benchmarked using saturation mutagenesis data available for T4 Lysozyme [372] and *E.coli* controller for cell death B (CcdB) protein (Adkar et al., 2012; Tripathia et al., 2016). The accuracy of our predictions were compared to those made by other methods described above.

2. Methods

2.1. Computation of residue depth

Depth is a concise descriptor of an amino acid residue environment [7,8,10–12]. It is defined as the average distance of the atoms of the residue to their nearest bulk solvent. In this study, residue depth was computed by previously described methods (Tan et al., 2013, 2011), using default parameters. Here, we have only considered protein structures that had only a few or no missing residues (see section 2.2). Missing residues could alter the distance to the closest molecule of bulk solvent and hence affect depth values.

2.2. Pairwise alignments for matrix creation

1607 structures were culled from the protein data bank (PDB) [278] using PISCES [139] and home grown scripts such that their a) sequences were non-redundant at 30% sequence identity, b) resolution was <3 Å with R-factor < 0.3 and c) structures were missing fewer than 6 contiguous residues. Missing stretches were modeled using the loop modeling [375] module of MODELLER [318]. Structures that had more than 6 missing residues were discarded, as errors in loop modeling could be significant enough to introduce errors in depth measurements.

BLAST [345] was used to identify the homologs of these 1607 proteins from the PDB. From this, 1426 homologs of 947 structures (from the initial 1607) were chosen such that the e-values were less than 0.001 and pair-wise sequence identities were less than 30%. From these 2383 (1426 + 947) structures, 3696 pairwise structure-structure alignments were constructed using SALIGN [22] such that the SALIGN quality score was >= 85% and the length difference between the 2 aligned proteins was <35 residues. These alignments gave us 800,558 residue substitutions.

2.3. Creation of depth dependent substitution matrices

Multiple substitution matrices were created from the pair-wise structure-structure alignments. All matrix values, $S_{i,j}^{d}$, were ratios of observed over expected residue substitution likelihoods and were computed using similar formulae used in BLOSUM[376].

$$S_{i,j}^{d} = 2 \cdot log_2\left(\frac{q_{i,j}^{d}}{e_{i,j}^{d}}\right)$$

where *i* and *j* are the residues that are being substituted to one another and d being the depth of the residue. Note that in these matrices, the substitution of *i* to *j* is considered equivalent to that of *j* to *i*. $q_{i,j}^{d}$ is the observed substitution probability and $e_{i,j}^{d}$ is the expected probability. The matrix values are scaled by a factor of $2 \cdot log_2$, similar to the BLOSUM62 matrix.

The observed probability is computed as

$$q_{i,j}^{d} = \frac{f_{i,j}^{d}}{\sum_{i=1}^{20}\sum_{j=1}^{20} f_{i,j}^{d}}$$

where $f_{i,j}^{d}$ is the number of substitutions of residue *i* to *j* (and vice versa) at depth range *d*. The denominator is the total number of observed residue substitutions.

The expected probability of residue substitution at the different depth ranges is given by

$$e_{i,j}^{d} = \begin{cases} p_i^d * p_j^d \text{when} i = j \\ 2 \cdot p_i^d * p_j^d \text{when} i \neq j \end{cases}$$

where, p_i^d is the probability of residue *i* at depth range *d* and is given by

$$p_i^d = q_{i,i}^{d} + \sum_{i \neq j} \frac{q_{i,j}^{d}}{2}$$

2.4. Database of single point mutants

Depth dependent substitution matrices were used to predict the effect of single point mutations in proteins. The predictions were trained on 1966 mutations of T4 Lysozyme [372], where 163 of the 164 amino acids of the protein were mutated to one of 13 different amino acids (A, C, E, F, G, H, K, L, P, Q, R, S, and T) after removal of the key catalytic residues (D10, E11, R145, and R148). The prediction training was done using a grid search over substitution values (searched in a range of -3 to 0.25 in steps of 0.25) in the three depth dependent matrices that could best discriminate between deleterious (destabilizing) and neutral mutations.

With the optimal parameters derived from the training set, the predictions were tested on another saturation mutagenesis set of 1534 mutants of the 101 residues long *E. coli* protein Controller of cell division or death B (CcdB) (Adkar et al., 2012; Tripathia et al., 2016) after removal of key catalytic residues (I24, I25, N95, F98, W99, G100, and I101). For the training and testing sets the crystal structures of T4 lysozyme (PDB code: 2LZM [377]) and CcdB (PDB code: 3VUB [378]) were used for depth computations. The experimental studies for T4 Lysozyme and CcdB ranked the severity of the mutant phenotype on a scale of 2-5 and 2-9 respectively. For both proteins, we considered level 2 to represent neutral (native like) mutations and all other levels to be destabilizing.

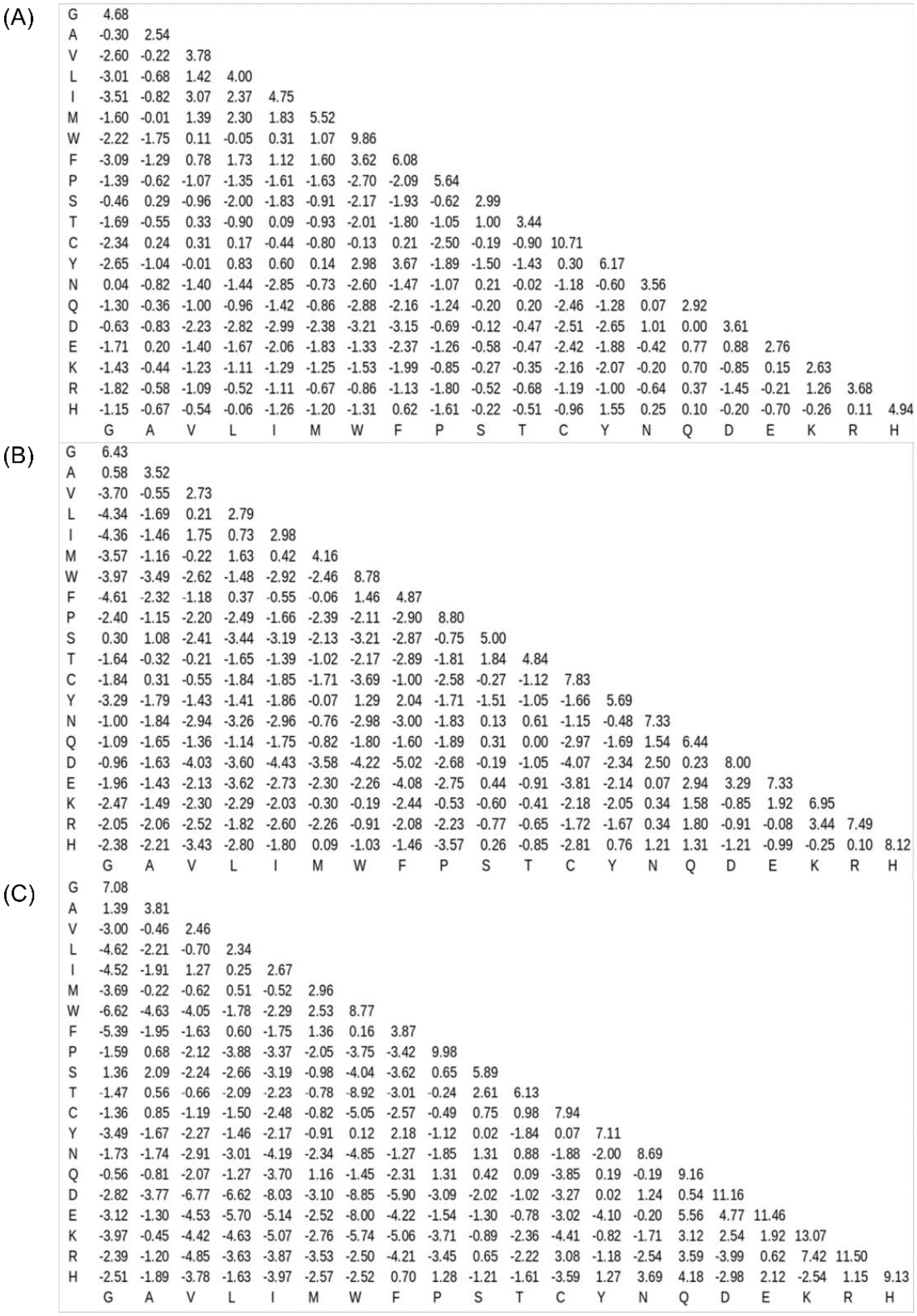
3. Results

3.1. Matrix creation and optimization

It was decided *apriori* to have a set of three 20 X 20 depth matrices, one each for exposed (E), intermediate (I) and buried (B) environments. We first determined the optimal ranges of depth values for these three matrices. As the computation of depth reports a mean value and an associated standard deviation, we decided that the minimum depth range for any of the 3 matrices should be 1.5 Å. The lower bound of the matrix corresponding to the exposed environment was set to a depth value of 2.5 Å. Its upper bound was tested in the range of 4.0 Å to 5.5 Å in steps of 0.5 Å. The lower bound of the intermediate matrix was the upper bound of the exposed matrix. It's upper bound was tested in the range of its lower bound + 1.5 Å to 8.0 Å in steps of 0.5 Å. The lower bound of the buried environment matrix was the upper bound of the intermediate matrix and had no upper bound. For each combination of the three ranges, an average root mean square distance, D_M, was computed between the matrices as

$$D_M = \langle \sqrt{\sum_{i,j} \frac{(X_{i,j} - Y_{i,j})^2}{210}} \rangle$$

where X and Y are either of the three matrices E, I or B (Figure 2) and $X_{i,j}$ is the score for substituting amino acid i for j in matrix X. The depth ranges with the highest D_M score (1.95) and hence considered optimal were (2.5-5.0 Å; 5.0-8.0 Å; > 8.0 Å). The D_M score averages over the root mean square distances of 1.44, 1.82 and 2.60 between the matrix pairs of exposed-intermediate, buried-intermediate and exposed-buried respectively.



(A)

G	4.68																			
A	-0.30	2.54																		
V	-2.60	-0.22	3.78																	
L	-3.01	-0.68	1.42	4.00																
I	-3.51	-0.82	3.07	2.37	4.75															
M	-1.60	-0.01	1.39	2.30	1.83	5.52														
W	-2.22	-1.75	0.11	-0.05	0.31	1.07	9.86													
F	-3.09	-1.29	0.78	1.73	1.12	1.60	3.62	6.08												
P	-1.39	-0.62	-1.07	-1.35	-1.61	-1.63	-2.70	-2.09	5.64											
S	-0.46	0.29	-0.96	-2.00	-1.83	-0.91	-2.17	-1.93	-0.62	2.99										
T	-1.69	-0.55	0.33	-0.90	0.09	-0.93	-2.01	-1.80	-1.05	1.00	3.44									
C	-2.34	0.24	0.31	0.17	-0.44	-0.80	-0.13	0.21	-2.50	-0.19	-0.90	10.71								
Y	-2.65	-1.04	-0.01	0.83	0.60	0.14	2.98	3.67	-1.89	-1.50	-1.43	0.30	6.17							
N	0.04	-0.82	-1.40	-1.44	-2.85	-0.73	-2.60	-1.47	-1.07	0.21	-0.02	-1.18	-0.60	3.56						
Q	-1.30	-0.36	-1.00	-0.96	-1.42	-0.86	-2.88	-2.16	-1.24	-0.20	0.20	-2.46	-1.28	0.07	2.92					
D	-0.63	-0.83	-2.23	-2.82	-2.99	-2.38	-3.21	-3.15	-0.69	-0.12	-0.47	-2.51	-2.65	1.01	0.00	3.61				
E	-1.71	0.20	-1.40	-1.67	-2.06	-1.83	-1.33	-2.37	-1.26	-0.58	-0.47	-2.42	-1.88	-0.42	0.77	0.88	2.76			
K	-1.43	-0.44	-1.23	-1.11	-1.29	-1.25	-1.53	-1.99	-0.85	-0.27	-0.35	-2.16	-2.07	-0.20	0.70	-0.85	0.15	2.63		
R	-1.82	-0.58	-1.09	-0.52	-1.11	-0.67	-0.86	-1.13	-1.80	-0.52	-0.68	-1.19	-1.00	-0.64	0.37	-1.45	-0.21	1.26	3.68	
H	-1.15	-0.67	-0.54	-0.06	-1.26	-1.20	-1.31	0.62	-1.61	-0.22	-0.51	-0.96	1.55	0.25	0.10	-0.20	-0.70	-0.26	0.11	4.94
	G	A	V	L	I	M	W	F	P	S	T	C	Y	N	Q	D	E	K	R	H

(B)

G	6.43																			
A	0.58	3.52																		
V	-3.70	-0.55	2.73																	
L	-4.34	-1.69	0.21	2.79																
I	-4.36	-1.46	1.75	0.73	2.98															
M	-3.57	-1.16	-0.22	1.63	0.42	4.16														
W	-3.97	-3.49	-2.62	-1.48	-2.92	-2.46	8.78													
F	-4.61	-2.32	-1.18	0.37	-0.55	-0.06	1.46	4.87												
P	-2.40	-1.15	-2.20	-2.49	-1.66	-2.39	-2.11	-2.90	8.80											
S	0.30	1.08	-2.41	-3.44	-3.19	-2.13	-3.21	-2.87	-0.75	5.00										
T	-1.64	-0.32	-0.21	-1.65	-1.39	-1.02	-2.17	-2.89	-1.81	1.84	4.84									
C	-1.84	0.31	-0.55	-1.84	-1.85	-1.71	-3.69	-1.00	-2.58	-0.27	-1.12	7.83								
Y	-3.29	-1.79	-1.43	-1.41	-1.86	-0.07	1.29	2.04	-1.71	-1.51	-1.05	-1.66	5.69							
N	-1.00	-1.84	-2.94	-3.26	-2.96	-0.76	-2.98	-3.00	-1.83	0.13	0.61	-1.15	-0.48	7.33						
Q	-1.09	-1.65	-1.36	-1.14	-1.75	-0.82	-1.80	-1.60	-1.89	0.31	0.00	-2.97	-1.69	1.54	6.44					
D	-0.96	-1.63	-4.03	-3.60	-4.43	-3.58	-4.22	-5.02	-2.68	-0.19	-1.05	-4.07	-2.34	2.50	0.23	8.00				
E	-1.96	-1.43	-2.13	-3.62	-2.73	-2.30	-2.26	-4.08	-2.75	0.44	-0.91	-3.81	-2.14	0.07	2.94	3.29	7.33			
K	-2.47	-1.49	-2.30	-2.29	-2.03	-0.30	-0.19	-2.44	-0.53	-0.60	-0.41	-2.18	-2.05	0.34	1.58	-0.85	1.92	6.95		
R	-2.05	-2.06	-2.52	-1.82	-2.60	-2.26	-0.91	-2.08	-2.23	-0.77	-0.65	-1.72	-1.67	0.34	1.80	-0.91	-0.08	3.44	7.49	
H	-2.38	-2.21	-3.43	-2.80	-1.80	0.09	-1.03	-1.46	-3.57	0.26	-0.85	-2.81	0.76	1.21	1.31	-1.21	-0.99	-0.25	0.10	8.12
	G	A	V	L	I	M	W	F	P	S	T	C	Y	N	Q	D	E	K	R	H

(C)

G	7.08																			
A	1.39	3.81																		
V	-3.00	-0.46	2.46																	
L	-4.62	-2.21	-0.70	2.34																
I	-4.52	-1.91	1.27	0.25	2.67															
M	-3.69	-0.22	-0.62	0.51	-0.52	2.96														
W	-6.62	-4.63	-4.05	-1.78	-2.29	2.53	8.77													
F	-5.39	-1.95	-1.63	0.60	-1.75	1.36	0.16	3.87												
P	-1.59	0.68	-2.12	-3.88	-3.37	-2.05	-3.75	-3.42	9.98											
S	1.36	2.09	-2.24	-2.66	-3.19	-0.98	-4.04	-3.62	0.65	5.89										
T	-1.47	0.56	-0.66	-2.09	-2.23	-0.78	-8.92	-3.01	-0.24	2.61	6.13									
C	-1.36	0.85	-1.19	-1.50	-2.48	-0.82	-5.05	-2.57	-0.49	0.75	0.98	7.94								
Y	-3.49	-1.67	-2.27	-1.46	-2.17	-0.91	0.12	2.18	-1.12	0.02	-1.84	0.07	7.11							
N	-1.73	-1.74	-2.91	-3.01	-4.19	-2.34	-4.85	-1.27	-1.85	1.31	0.88	-1.88	-2.00	8.69						
Q	-0.56	-0.81	-2.07	-1.27	-3.70	1.16	-1.45	-2.31	1.31	0.42	0.09	-3.85	0.19	-0.19	9.16					
D	-2.82	-3.77	-6.77	-6.62	-8.03	-3.10	-8.85	-5.90	-3.09	-2.02	-1.02	-3.27	0.02	1.24	0.54	11.16				
E	-3.12	-1.30	-4.53	-5.70	-5.14	-2.52	-8.00	-4.22	-1.54	-1.30	-0.78	-3.02	-4.10	-0.20	5.56	4.77	11.46			
K	-3.97	-0.45	-4.42	-4.63	-5.07	-2.76	-5.74	-5.06	-3.71	-0.89	-2.36	-4.41	-0.82	-1.71	3.12	2.54	1.92	13.07		
R	-2.39	-1.20	-4.85	-3.63	-3.87	-3.53	-2.50	-4.21	-3.45	0.65	-2.22	3.08	-1.18	-2.54	3.59	-3.99	0.62	7.42	11.50	
H	-2.51	-1.89	-3.78	-1.63	-3.97	-2.57	-2.52	0.70	1.28	-1.21	-1.61	-3.59	1.27	3.69	4.18	-2.98	2.12	-2.54	1.15	9.13
	G	A	V	L	I	M	W	F	P	S	T	C	Y	N	Q	D	E	K	R	H

Figure 2- The three symmetric 20X20 depth dependent substitution at depth ranges of (A) <5Å, (B) 5-8Å and (C) >8Å.

The three depth dependent substitution matrices are all distinctly different from one another as can be gauged from their pairwise D_M values. A closer look into the pairwise matrix comparison reveals that most of the substitution scores (in the integral version) are different (Figure 3A-C and Table 1). In all, 200 of the 210 substitution values change in the different matrices (see section 3.3 for detailed description of substitution trends).

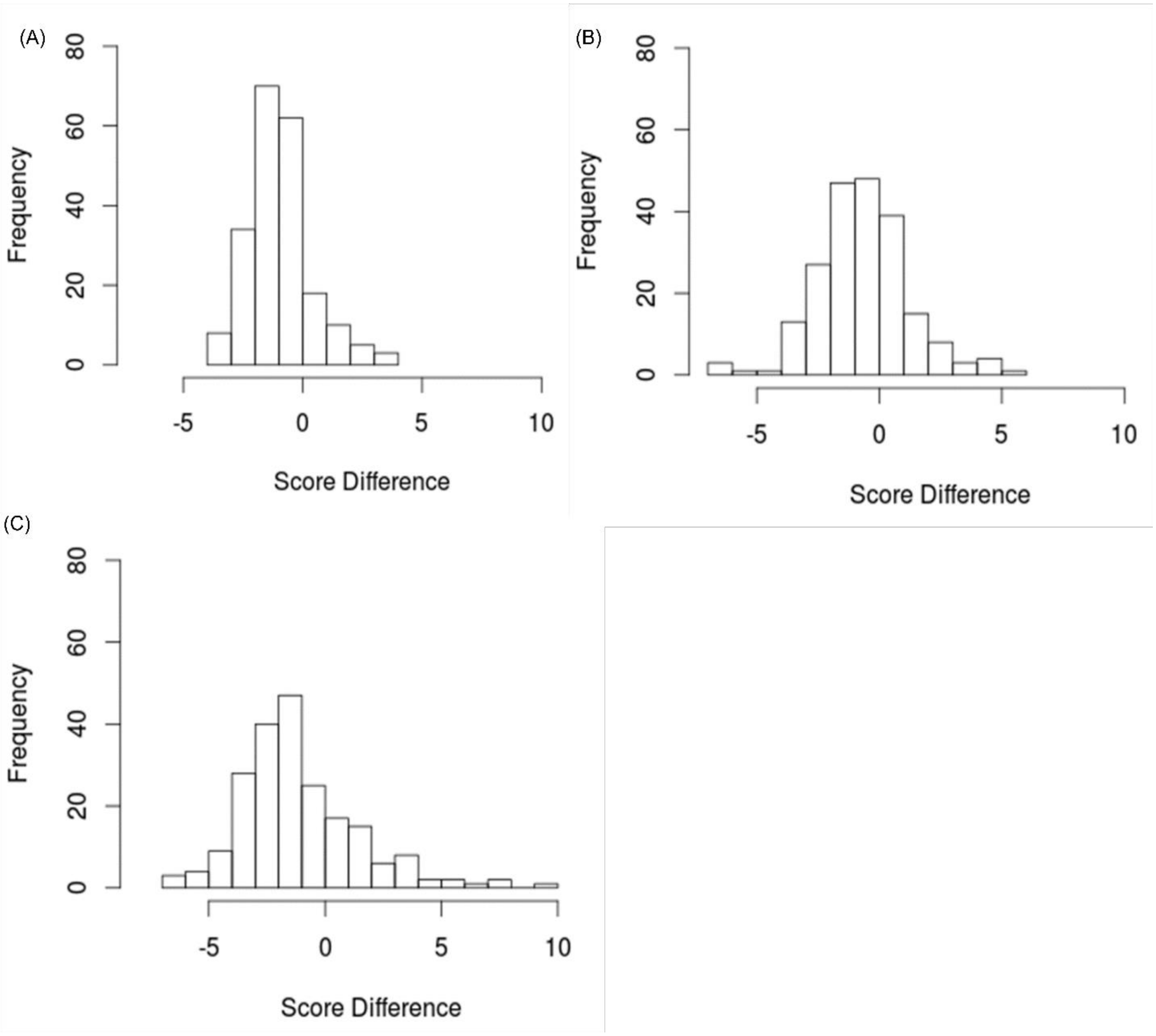


Figure 3- Histogram of score difference of matrices between (a) Intermediate and Exposed residues (b) Buried and Intermediate residues (c) Buried and Exposed residues.

Table 2- Frequency of residues having a score difference of 0, +/-1,+/-2 or >+/-2 between the matrices for exposed and intermediate residues, intermediate and buried residues

and exposed and buried residues.

Score difference	**0**	**+/-1**	**+/-2**	**> +/-2**
Exposed-Intermediate	62	88	44	16
Intermediate-Buried	48	86	42	34
Exposed-Buried	25	64	55	66

A test of accuracy of the matrix values was to create a regular substitution matrix. This composite matrix was created as described in the methods section, only this time without taking into consideration depth levels. When our composite matrix was compared to the BLOSUM62 matrix, it was identical for 49%of the substitution values and varied by +/- 1 unit in 47% of the substitution values. Here again, we believe that the differences of +/- 1 are mainly caused by rounding off the matrix values. This validates that the three depth dependent matrices are stratified versions of the regular substitution matrices.

In order to check the consistency of the dataset, regular substitution matrices were created 5 times with a random sample of 80% of the data. 183 out of 210 substitutions (87%) had the same score in all 5 times. The remaining scores for the 27 substitutions varied only by +/-1.

3.2 Depth conservation in alignments

The substitution matrices were created from pairwise alignments of protein structures. Over 90% of the aligned residues had depth differences of <1.5 Å, with 81% having differences of <1 Å (Figure 4). The depth differences were examined for different types of substitutions *i.e.* polar to polar, polar to non-polar (or vice versa), and non-polar to non-polar (Table 2). Polar residues substituted by other polar residue showed the least variation in depth as these residues are predominantly found in the outer layer of the protein. Non-polar to non-polar residue substitutions showed larger depth changes in comparison. A possible reason could be that non-polar residues are present in larger proportions and interchange with one another frequently at deeper depths. In this deep environment, changes in amino acid size precipitously change their depth values. The values of depth difference for non-polar to polar substitutions (or vice versa) lie in between the two values discussed above.

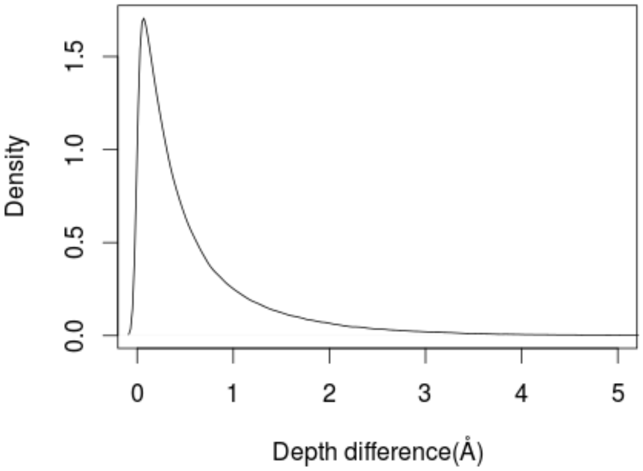


Figure 4- Density plot of the depth difference between the aligned residues

Table 2- Residue substitutions from polar to non-polar, polar to polar and non-polar to non-polar and the proportions that have depth difference of greater than 1Å and 1.5 Å.

Type of substitution	**Substitutions with depth difference >1Å (%)**	**Substitutions with depth difference >1.5Å (%)**
Non-polar to non-polar	24	13
Polar to non-polar	19	10
Polar to polar	11	5

3.3 Substitution trends

As discussed earlier, the relative abundances of the 20 amino acids at different depth levels are different from one another (Figure 1). It is reasonable to expect that their substitution rates would also vary accordingly. The depth dependent substitution matrices capture these variations (Figure 3). The substitution trends across depth levels show that the polar amino acids are easier to substitute at deeper depths while the hydrophobic amino acids show the opposite trend and get harder to substitute. It should be noted however that this higher/lower propensity of substitution is relative and at any depth, amino acid self-substitutions score the highest.

Some interesting information one can extract from the depth substitution matrices are the substitution trends across depth ranges. In the matrices derived in this study, there are six different types of substitution behaviors as we traverse from the outside of the protein (low depth) to the interior (high depth). Scores increase for 23 substitutions, decrease in 52 cases and remain the same in 10 substitutions. In addition to this, 90 substitutions have the same score for 2 consecutive depth ranges and their first/last value increase (25)/decrease (65). Some substitutions showed a trend where their values in the middle depth range had a lower or higher score as compared to scores in the other 2 ranges (35 out of 210, denoted as ∨ or ∧ in Figure 5).

A closer look at the substitutions show that by and large the score for substituting one polar (S,T,C,Y,N,Q,D,E,K,R,H) amino acid either increases or remains the same from exposed to buried environments. This is possibly because in deeper environments substituting one polar group by another would maintain charge-charge interactions and leave no unpaired charges buried. Cysteine mutations buck this trend and are generally less favorable to mutate in deeper environments. Threonine and Serine are also less likely to be substituted by any of the larger polar (charged or uncharged) groups in deep environments. The trends for hydrophobic (G, A, V, L, I, M, W, F, P) to hydrophobic substitutions is in some sense the opposite of what is seen in polar residues. The deeper one goes into the protein the less likely it gets to substitute a non-polar group by another. This is probably because the difference between the individual hydrophobic groups could contribute to substantial differences in hydrophobic packing. The trends for substituting non-polar groups by polar groups (or vice versa) get more unlikely in deeper environments. Again, there are exceptions to this - Serine, Threonine and Cysteine are more amenable to being substituted by small amino acids such as Alanine and Glycine with increasing depth. An unusual exception is the increased likelihood of substituting Tryptophan by Glutamine that gets less unfavorable in the hydrophobic core.

The substitutions of E-H, M-Q, H-P, R-S, A-T, P-S, P-Q, C-R, C-T, A-P AND D-K that are disfavored (negative score) in the protein exterior become favored (positive score) in the hydrophobic core. The substitutions F-H, F-I, I-Y, F-V, M-V, I-M and L-Y while favored in the exterior become disfavored in the interior. In addition to these extreme cases, there are several cases of neutral mutations becoming less or more favored in different environments and vice versa. Several of the anomalous substitution behavior could be explained away by sparse observations. Cysteine, Methionine and Tryptophan, for instance, have low abundances and hence the computation of the observed by expected substitution likelihood ratios could sometimes be erroneous. The matrix as a whole, we believe, is largely reflective of the real substitution rates between amino acids residues.

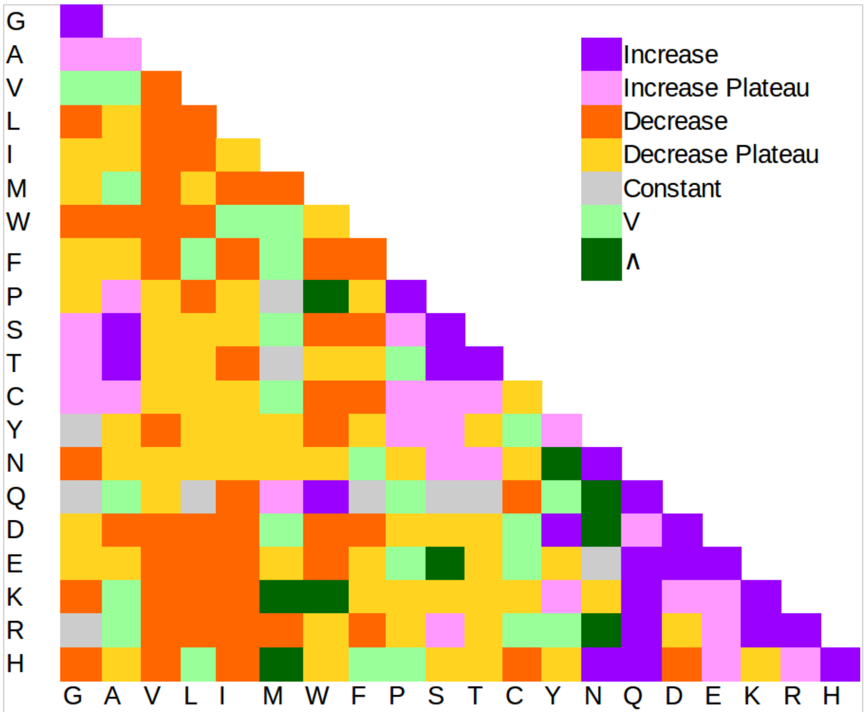


Figure 5 - Trends in substitution scores as noticed from the three depth dependent matrices. Substitutions colored purple increase in value as depth increases. Those that decrease in value with increasing depth are colored orange. Substitutions that have a constant value across all the depth environments are colored grey. Substitutions that increase and then plateau and decrease and then plateau are colored pink and yellow respectively. Substitutions that are colored light and dark green are those whose values decrease and then increase or increase and then decrease respectively.

An important trend that we have not explicitly considered here is the difference in substitution rates between Cysteine (free thiol) and Cystine (disulphide bridged). In the ~3,700 pairwise structure-structure alignments we have used for constructing the matrices we have very few Cystine substitutions (on average 25 substitutions of Cystine to other amino acids) and in some cases, substitutions are not observed at all. For this reason, the current matrices have not differentiated between Cysteine and Cystine.

3.4 Applications of the matrix to detect deleterious single point mutations

We used saturation mutagenesis data from two proteins, T4 Lysozyme and CcdB, to train and test a prediction schema for identifying destabilizing point mutations. The experimental data for both proteins described the mutagenesis data in terms of intensity of phenotype of the mutations. In the case of T4 Lysozyme the mutational sensitivity was scored on a scale of 2 to 5 while the range was 2 to 9 in the case of CcdB. For the computations described below we have taken a sensitivity score of 2 to imply neutral (or native like) mutations and all other scores to imply destabilizing mutations for both proteins. The datasets hence consist of 1362 (69%) and 1258 (82%) neutral mutations and 604 (31%) and 276 (18%) of destabilizing mutations in T4 Lysozyme and CcdB respectively. Our simplistic method consisted of finding threshold values in the depth substitution matrices using the training set data that would best distinguish between neutral and destabilizing mutations. The thresholds were found by a grid search that varied the threshold value in the range -3 to -0.25 in steps of 0.25 overall three matrices. The optimal threshold values (-3, -0.25 and -0.25 for the <5 Å, 5-8 Å and >8 Å depth matrices respectively) were then applied to evaluate the efficacy of the method over the testing set data. The accuracy of our binary classification method, and those of other methods compared to ours, was measured in terms of Sensitivity, Specificity, Precision, Accuracy, f1 and MCC [73][377].

Our depth dependent substitution matrix (FADHM) method was compared to other popular methods including Automute [368], DUET [371], I-mutant [367], SuSPect [366], msCSM [379] and SDM[370] which predict if mutations are destabilizing (Table 3). Our precision values (60% and 44% for T4 lysozyme and CcdB respectively) and specificity values (85% and 78% for T4 lysozyme and CcdB respectively) were either the highest or comparable to that of the other methods. Consistently, in both the training and testing sets FADHM has the best accuracy (75% and 78%), f1 (0.55 and 0.57) and MCC values (0.38 and 0.48). The next best methods for theT4 lysozyme and CcdB datasets have MCC values of 0.30 (I-mutant) and 0.40 (DUET) respectively. Predictions of destabilizing mutations by our simple method outperform other sophisticated algorithms.

*Table 3- Prediction performance comparison of different prediction techniques on (a) training set (T4 lysozyme) (b) testing set (CcdB). The maximum performance of each value measure is indicated in bold. * FADHM is Amino acid DeptH substitution Matrices*

Technique	Sensitivity(%)	Specificity(%)	Precision(%)	Accuracy(%)	F1	MCC
(A)						
I-mutant	56	75	50	69	0.53	0.30
SuSPect	**75**	53	41	60	0.53	0.26
Automute	70	54	41	59	0.52	0.23
mCSM	60	70	26	68	0.37	0.22
SDM	58	73	28	71	0.38	0.24
DUET	61	70	27	69	0.38	0.24
FADHM*	51	**85**	**60**	**75**	**0.55**	**0.38**
(B)						
I-mutant	64	78	44	73	0.52	0.36
SuSPect	**99**	21	22	35	0.36	0.21
Automute	76	49	30	55	0.43	0.21
mCSM	68	76	44	74	0.54	0.39
SDM	54	**81**	45	75	0.49	0.33
DUET	67	78	**46**	76	0.54	0.40
FADHM	80	78	44	**78**	**0.57**	**0.48**

To check the robustness of our results we repeated the accuracy computations 10 times for both the training and testing sets, this time considering only a randomly selected 40% subset of the data. These tests showed that the average MCC value for T4 Lysozyme and CcdB was 0.39 with a standard deviation of 0.03 and 0.48 with a standard deviation of 0.02 respectively.

4. Discussion

In earlier studies, we had established the utility of the residue depth measure to concisely describe the local environment. The depth measure has been successfully used for diverse applications including, but not exhaustive, finding small molecule binding sites on proteins, predicting what single point mutations would yield temperature sensitive mutations and estimating the pK_as of ionizable amino acids. In this study, we have used the depth measure in conjunction with the knowledge that the relative abundances of different amino acids differ in different protein environments. This suggests that the substitution rates of amino acids would also be different at different depths. The 3 depth dependent substitution matrices were hence created.

We arbitrarily chose to create a set of three matrices to represent the substitutions in expose, intermediate and buried environments. The depth values (5Å and 8Å) that demarcated the boundaries between these 3 classes were obtained by attempting to maximize the differences between the matrices. The resulting matrices are quite different from one another and show the difference in the substitution likelihoods in different environments. We observed 6 different substitution trends in pairwise residue substitutions across different environments. The patterns include substitutions whose values remained unchanged, increased or decreased monotonically, increased/decreased and then plateaued, increase and then decreased or vice versa. Only 10 of the 210 substitutions remained unchanged across all three environments. The matrices show many expected trends such as how replacing a hydrophobic residue with a polar one in the buried environments is generally unfavorable. There were many surprising substitutions trends where the intermediate region was the most favored in comparison to buried and exposed environments. Some of these trends could be artefacts of low abundance of residues such as Methionine, Cysteine and Tryptophan. The other such trends indicate that the matrices were able to capture some of the nuances of residue preferences and substitutions across different environments.

We tested the matrices for their ability to detect mutations that lead to protein instability. Saturation mutagenesis data from T4 Lysozyme and CcdB were used as the training and testing sets respectively. Mutations/substitutions were considered as destabilizing if the substitution score (native to mutant) was less than -3.00, -0.25 and -0.25 in exposed, intermediate and buried environments respectively. Our somewhat simplistic approach outperformed other popular methods, some of which use machine learning rigorously. Of the 276 deleterious mutations in the CcdB test set, we accurately identified 220 while the next best method identified only ~160. Our method, and the others, produce a large number of false positives and hence a somewhat modest overall performance (MCC of 0.38 on the training set and 0.48 on the testing set). In comparison to the others, our method has low sensitivity, is comparable in terms of specificity and precision but outperforms in accuracy, f1 and MCC values.

We believe that these depth dependent substitution matrices are important in describing the internal environments of proteins. Further developments of these potentials could include the creation of asymmetric substitution matrices as the relative abundances of different amino acids in the different environments vary. These matrices should be able to improve the accuracy in aligning distantly related homologues with one another. This is the first of what we expect to be a series of studies to learn from substitution likelihoods in different protein environments.

Chapter 11 - Conclusions and Future Prospects

Proteins are essential ‘workforce’ of the cell that interact with one another and other biomolecules to carry out its functions. The protein surfaces can be characterized structurally and physicochemically to study the binding interface on them.

To structurally describe the protein-protein interfaces we created a library of all protein-protein and domain-domain interfaces. Domain-domain interfaces were used in our study as during evolution protein chains can join together to form a larger protein, hence converting a chain-chain interface into a domain-domain interface or vice versa (when protein domains split into different chains). Hence, we assume that chain-chain interfaces and domain-domain interfaces would be structurally similar and can hence be studied together. The library was clustered by structural folds (CATH nomenclature), though an ideal structural clustering would involve an all against all comparison for the clustering. An all against all structural comparison was intractable given a large number of protein structures (178,485 structures leading to $178{,}485^{178{,}485}$ comparisons). As a result of foldwise clustering, we saw that certain domain-domain interfaces were structurally similar to chain-chain interfaces. Certain folds such as Rossmann fold, Immunoglobulin-like fold can interact with a large number of folds because of the sheer diversity in its sequences. Certain comparisons of interfaces across protein folds also show that interfaces can be topologically similar irrespective of the topology of the protein.

We also identified small molecule binding sites at protein-protein interfaces, which can be targeted by peptides/small molecules and hence inhibit the formation of that protein complex. This has immense use in the field of drug targeting and modulating protein associations. We also physicochemically characterized and compared protein-protein and domain-domain interfaces based on the pair preferences of amino acids. We noticed that both the interfaces (chain-chain and domain-domain) have similar pair preferences for amino acids. However, chain-chain interfaces, have a higher preference of self-pairing of amino acids. This is possibly because of the presence larger number of homo-oligomer (involving the same fold) for a chain-chain interface as compared to domain-domain interfaces.

This interface library can serve as an important repertoire to identify binding modes between proteins as nature reuses the geometry of the binding interface. We can also utilize this library to model proteins with multiple domains where individual domains have been crystallized separately. A combination of both chain-chain and domain-domain interfaces with a topology independent structural search technique can help identify a larger number of templates for modeling of protein complexes and multi-domain proteins. This library can be used as structural templates to sample binding modes of proteins, hence helping in modeling protein complexes.

The interface library contains a repertoire of structural binding modes between proteins, which can be used to model protein complexes of different types. However, we only studied coiled-coil interfaces. We developed a random forest based scoring scheme to predict if two coiled-coil proteins would interact with one another in a particular orientation. Features of the scoring scheme involve pair preferences of the interacting amino acid pairs in a coiled-coil motif. We utilized this scoring scheme to predict native coiled-coil interface interactions from non-native interactions. To showcase the utility of this scoring scheme, we identified the binding modes of JC virus agnoprotein to p53 and Rab-11B. We also modeled the complex of agnoprotein with Rab-11B based on the best scoring coiled-coil pose using a template coiled-coil protein. Several coiled-coil related scoring schemes exist, which aim to predict coiled-coil regions in proteins, oligomerization state, but to the best of our knowledge, this is the first scoring scheme to identify if two coiled-coil proteins would interact in a specific orientation. This technique will be converted into a webserver to identify binding partners of coiled-coil proteins and model the complexes of these proteins. This resource can also be developed to design inhibitory peptides following a heptad repeat pattern that would bind to a coiled-coil protein of interest with a higher affinity compared to its native interacting partner.

Along with the structural description of protein-protein interfaces, we also studied the residue environments at protein-protein interfaces. Not all residues at an interface prefer to change their environments upon complex formation. The hydrophilic amino acids do not prefer to get buried upon complex formation, whereas the hydrophobic amino acids have higher propensities to get buried. Hence the hydrophilic amino acids probably form the rim of the interface and the hydrophobic amino acids the core of the interface. We utilized these probabilities of amino acids moving from one depth level in a monomer to another depth level in a complex, to develop a scoring scheme (called MODP) to score protein complexes. We employed MODP to identify native protein complex structures and to distinguish near-native structures from non-native structures in 3 different protein complex decoy sets. We compared our technique to another state of the art technique PIZSA. PIZSA outperforms MODP in ranking the native structures among its decoys. However, MODP (MCC in the range of 0.16-0.36) outperforms PIZSA (MCC in the range of 0.07-0.17) in identifying near-native structures in all the three datasets. This can be because of inherent differences in the scoring schemes. PIZSA involves residue level and atomic level propensities in the scoring and is hence able to identify native structures well. MODP, however, is a coarse-grained technique and hence allows the identification of near natives better. We can hence use PIZSA and MODP, as complementary techniques to each other in the identification of near-native and ranking of native interfaces. MODP can also be further suitably modified to identify patches on the surface of the proteins that are likely to undergo oligomerization. This would involve scoring surface patches and identifying regions on the surface with hydrophobic amino acids, which are more likely to form the interacting interface. To the best of our knowledge, this method would be the first to predict interacting regions on protein surfaces.

We also studied the properties of protein-protein interface residues to identify the residues that are important for binding (hotspot residues). We developed an empirical decision tree based scoring scheme involving- a) depth change on complex formation, b) conservation and c) amino acid contact pair potential to classify hotspot residues from non-hotspot residues. We compared ourselves (DepthCon) with other state of the art scoring schemes. DepthCon yielded an MCC of 0.46+/-0.01 across 1 training and 4 testing sets. Whereas the best technique, PredHS-Ensemble had an MCC of 0.48+/-0.13. The higher standard deviation of PredHS-Ensemble indicates that PredHS-Ensemble performs well in certain cases whereas fails in certain others, whereas DepthCon has similar performances across different datasets. This technique can be converted to a web server for use by the scientific community to predict hotspot regions in proteins. Identification of these hotspot residues can help modulate the interaction strength of a complex by mutation of the hotspot residues.

Along with the characterization of protein-protein interfaces we also structurally studied protein-small molecule interfaces. We did a pilot study with the small molecule Nutlin to predict its off target proteins. Nutlin binds to Mdm2, which is upregulated in a lot of cancer and is under clinical trials for cancers caused by upregulated Mdm2. We carved the binding pocket of Nutlin on Mdm2 and utilized CLICK to identify structurally similar binding pockets in other proteins (off target). We identified 49 human proteins with a structurally similar binding pocket of Nutlin as on Mdm2. We also computationally accessed the binding strength of these complexes by MM/GBSA, docking scores and molecular dynamics simulations. We were able to predict binding pockets of Nutlin, which were inaccessible to docking tools. Most of the docking tools sample poses of the small molecule, keeping the protein rigid. Flexible docking, however, allows movement of the side chain residues, and is computationally intensive. However, the inherent flexibility of the structural match (non-zero RMSD between the 2 structures being matched) allows us to predict binding pockets, which might be inaccessible in the crystal structure but are accessible in its natural environment because of thermal fluctuations. The binding to one of the predicted off target protein, Gamma-glutamyl hydrolase was also experimentally verified by thermal shift assay. Gamma glutamyl hydrolase has been associated with arthritis and cancer and hence Nutlin might be a useful drug against these conditions (drug repurposing). Hence, this method can be used for prediction of off target effects of drugs and drug repurposing.

In addition to studying the off target effects of Nutlin, we also predicted the binding pockets of drugs with experimentally verified off targets. We predicted the binding pocket of 6 drugs (Cetirizine, Disopyramide, Doxorubicin, Fluoxetine, Paroxetine and Rabeprazole) on 2 proteins (DRD3 and HRH1). Both these proteins are transmembrane proteins and most of the predictions (~80%) made by us were on the extra-cellular region of the proteins, where, a large number of ligands were already known to bind. The predicted binding site for a most of the cases (~80%) was on a site with an already bound ligand in the crystal structure. This further validates our prediction as previous studies point towards multiple drugs binding to the same binding pocket. This technique developed by us can be useful in predicting off target effects of drugs with known complexes, on human proteins and also help in drug repurposing. This study can be made more extensive by inclusion of computational models of human proteins during the prediction. The CLICK based structural match, does not take into consideration the physicochemical properties of the binding site. Hence, it become important to develop appropriate scoring schemes to score the protein-small molecule complexes.

Along with studying protein-protein and protein-small molecule interfaces, we also designed and predicted inhibitors against all the 9 proteins of Nipah virus. We modeled the proteome of the virus and computationally designed 4 peptide inhibitors against 3 proteins (G, F and M proteins) and 146 small molecule inhibitors against 5 proteins (F, G, M, N and P proteins). We increased the structural coverage of the Nipah proteome by 90% of what was deposited in the PDB. We computationally accessed the stability of all 4 peptide inhibitors and 13 of the top shortlisted small molecules. MM/PBSA calculations were used to computationally analyze the stability of binding. All of the peptide inhibitors and 9 of the small molecule ligands provided favorable binding free energy from the MM/PBSA calculations. We also analyzed the drug resistance of the predicted/designed inhibitors against the different strains of the Nipah virus. None of the sequence variations, other than 1, in the binding pocket of the inhibitor, had a negative BLOSUM substitution score indicating conservative substitutions. The one mutation that had a non-negative substitution score had its side chain pointing away from the binding site. Hence, we assume that the predicted/designed inhibitors will be effective against the different strains of the Nipah virus. We tackled the prediction of therapeutics on a proteome wide scale and these molecules can serve as the starting points for drug development.

Besides studying protein surfaces for protein-protein and protein-small molecule interactions, we also characterized residue environments in proteins. Different amino acids prefer to be at different environments in a protein. We characterized the amino acid environments using a parameter called residue depth, which is the distance of the residue from the nearest bulk solvent molecule. The hydrophilic amino acids prefer to be on the surface (lower depth) while the hydrophobic amino acids prefer to be buried (higher depth). Since amino acids prefer to be in different environments, the rate of substitution will also be different depending on the environment. We computed the rates of substitution of the different amino acids at different depth levels (exposed, intermediate and buried) and created 3 depth dependent amino acid substitution matrices (FADHM). These 3 matrices captured the substitution trends in these different environments. We noticed that depending on the nature of amino acid substitution, their substitution trends vary in different depth levels. We utilized these matrices to predict deleterious mutations in proteins on a test set of CcdB mutations containing 1534 mutations, of which 276 mutations were deleterious. FADHM was able to predict 220 mutations (MCC 0.48) whereas the next best technique (DUET) predicted only 160 (MCC 0.40). The substitution matrices can be further improved and utilized to improve structure-sequence alignments of proteins, hence improving the quality of models predicted by homology based structure modeling. These matrices can be modified to asymmetric matrices or family based matrices, but we are unsure of the presence of adequate structural data for such studies.

This thesis encompasses various structural and residue environment based characterization of proteins with the specific aim of studying protein-protein and protein-small molecule interfaces. The work carried out in this thesis has multiple applications in the prediction of deleterious mutations in proteins, sampling and scoring protein-protein complex models, scoring and modeling coiled-coil proteins, identification of hotspot residues, predicting off target effects of drugs and drug repurposing.

Publications

1. **Sen N.***, Roy A.A.*, Kanitkar T.R.*, Soni N., Amritkar K., Supekar S., Nair S., Singh G., Madhusudhan M.S., Predicting and designing therapeutics against the Nipah virus (2019), Plos Neglected Tropical Diseases, https://doi.org/10.1371/journal.pntd.0007419 [* equal contributions]
2. Nguyen M.N.* ,**Sen N.***, Meiyin L.*, Joseph T.L., Vaz C., Tanavde V., Way L., Hupp T., Verma C. and Madhusudhan M.S. (2019), Discovering putative protein targets of small molecules: A study of the p53 activator Nutlin**,** Journal of Chemical Information and Modeling, doi: 10.1021/acs.jcim.8b00762 [* equal contributions]
3. Bullock J.M.A., **Sen N.**, Thalassinos K., Topf M. (2018), Modelling protein complexes using distance and solvent accessibility restraints from cross-linking mass spectrometry**,** Structure, doi: 10.1016/j.str.2018.04.016
4. Roy A., Nair S., **Sen N.**, Soni N.,Madhusudhan M.S. (2017), *In silico* methods for design of biological therapeutics, Methods, doi: 10.1016/j.ymeth.2017.09.008 (Review)
5. Farheen N.*, **Sen N.***, Nair S., Tan K.P., Madhusudhan M.S. (2017), Depth dependent amino acid substitution matrices and their use in predicting deleterious mutations, Progress in Biophysics and Molecular Biology, doi: 10.1016/j.pbiomolbio.2017.02.004 [* equal contributions]

In press

6. Kanitkar T.R., **Sen N.**, Soni N., Nair S., Amritkar K., Ramatirtha Y., Madhusudhan M.S., Methods for Molecular Modeling of Protein complexes, Methods in Molecular Biology : Structural Proteomics 3rd Edition, Springer (Book Chapter)

Manuscripts under preparation

7. **Sen N.***, Soni N.*, Madhusudhan M.S., Prediction and Modelling of coiled-coil protein-protein interfaces [* equal contributions] (in preparation)
8. **Sen N.** , Nguyen M.N., Madhusudhan M.S., A library of all known protein-protein and domain-domain interfaces to structurally characterize them (in preparation)

9. **Sen N.**, Topf M., Madhusudhan M.S., Amino acid depth dependent scoring potentials to discriminate between native and non-native protein complexes (in preparation)
10. **Sen N.**, Madhusudhan M.S., Prediction of hotspot residues at protein interfaces (in preperation)

References

1. Li L, Li C, Zhang Z, Alexov E. On the dielectric “constant” of proteins: Smooth dielectric function for macromolecular modeling and its implementation in DelPhi. J Chem Theory Comput. 2013; doi:10.1021/ct400065j
2. Gao J, Bosco DA, Powers ET, Kelly JW. Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins. Nat Struct Mol Biol. 2009; doi:10.1038/nsmb.1610
3. Pace CN. Energetics of protein hydrogen bonds. Nature Structural and Molecular Biology. 2009. doi:10.1038/nsmb0709-681
4. Pace CN, Fu H, Fryar KL, Landua J, Trevino SR, Shirley BA, et al. Contribution of hydrophobic interactions to protein stability. J Mol Biol. 2011; doi:10.1016/j.jmb.2011.02.053
5. Lee B, Richards FM. The interpretation of protein structures: Estimation of static accessibility. J Mol Biol. 1971;55. doi:10.1016/0022-2836(71)90324-X
6. Samanta U, Bahadur RP, Chakrabarti P. Quantifying the accessible surface area of protein residues in their local environment. Protein Eng. 2002; doi:10.1093/protein/15.8.659
7. Chakravarty S, Varadarajan R. Residue depth: a novel parameter for the analysis of protein structure and stability. Structure. 1999;7: 723–732. doi:http://dx.doi.org/10.1016/S0969-2126(99)80097-5
8. Tan KP, Varadarajan R, Madhusudhan MS. DEPTH: A web server to compute depth and predict small-molecule binding cavities in proteins. Nucleic Acids Res. 2011;39. doi:10.1093/nar/gkr356
9. DeLano WL. The PyMOL Molecular Graphics System. Schrödinger LLC wwwpymolorg. 2002;Version 1.: http://www.pymol.org. doi:citeulike-article-id:240061
10. Pintar A, Carugo O, Pongor S. Atom depth as a descriptor of the protein interior. Biophys J. 2003;84: 2553–2561. doi:10.1016/S0006-3495(03)75060-7
11. Pintar A, Carugo O, Pongor S. Atom depth in protein structure and function. Trends in Biochemical Sciences. 2003. pp. 593–597. doi:10.1016/j.tibs.2003.09.004

12. Tan KP, Nguyen TB, Patel S, Varadarajan R, Madhusudhan MS. Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. Nucleic Acids Res. 2013;41. doi:10.1093/nar/gkt503

13. Tan KP, Khare S, Varadarajan R, Madhusudhan MS. TSpred: A web server for the rational design of temperature-sensitive mutants. Nucleic Acids Res. 2014;42. doi:10.1093/nar/gku319

14. Liu S, Zhang C, Liang S, Zhou Y. Fold recognition by concurrent use of solvent accessibility and residue depth. Proteins Struct Funct Genet. 2007;68: 636–645. doi:10.1002/prot.21459

15. Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins Struct Funct Genet. 2005;58: 321–328. doi:10.1002/prot.20308

16. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol. 2009; doi:10.1371/journal.pcbi.1000585

17. Magliery TJ, Regan L. Sequence variation in ligand binding sites in proteins. BMC Bioinformatics. 2005; doi:10.1186/1471-2105-6-240

18. Nguyen MN, Tan KP, Madhusudhan MS. CLICK--topology-independent comparison of biomolecular 3D structures. Nucleic Acids Res. 2011;39: W24-8. doi:10.1093/nar/gkr393

19. Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM. MUSTANG: a multiple structural alignment algorithm. Proteins. 2006; doi:10.1002/prot.20921

20. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol. 1993; doi:10.1006/jmbi.1993.1489

21. Nussinov R, Wolfson HJ. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. Proc Natl Acad Sci U S A. 1991; doi:10.1073/pnas.88.23.10495

22. Braberg H, Webb BM, Tjioe E, Pieper U, Sali A, Madhusudhan MS. Salign: A web server for alignment of multiple protein sequences and structures. Bioinformatics. 2012;28: 2072–2073. doi:10.1093/bioinformatics/bts302

23. Nguyen MN, Madhusudhan MS. Biological insights from topology independent comparison of protein 3D structures. Nucleic Acids Res. 2011;39. doi:10.1093/nar/gkr348

24. Konc J, Janežič D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. Bioinformatics. 2010; doi:10.1093/bioinformatics/btq100

25. Štular T, Lešnik S, Rožman K, Schink J, Zdouc M, Ghysels A, et al. Discovery of Mycobacterium tuberculosis InhA Inhibitors by Binding Sites Comparison and Ligands Prediction. J Med Chem. 2016; doi:10.1021/acs.jmedchem.6b01277

26. Grishin N V. Fold change in evolution of protein structures. J Struct Biol. 2001; doi:10.1006/jsbi.2001.4335

27. Hou J, Sims GE, Zhang C, Kim SH. A global representation of the protein fold space. Proc Natl Acad Sci U S A. 2003; doi:10.1073/pnas.2628030100

28. Pascual-García A, Abia D, Ortiz ÁR, Bastolla U. Cross-over between discrete and continuous protein structure space: Insights into automatic classification and networks of protein structures. PLoS Comput Biol. 2009; doi:10.1371/journal.pcbi.1000331

29. Sahoo MR, Gaikwad S, Khuperkar D, Ashok M, Helen M, Yadav SK, et al. Nup358 binds to AGO proteins through its SUMO -interacting motifs and promotes the association of target mRNA with miRISC . EMBO Rep. 2017; doi:10.15252/embr.201642386

30. Doan DNP, Li KQ, Basavannacharya C, Vasudevan SG, Madhusudhan MS. Transplantation of a hydrogen bonding network from West Nile virus protease onto Dengue-2 protease improves catalytic efficiency and sheds light on substrate specificity. Protein Eng Des Sel. 2012; doi:10.1093/protein/gzs049

31. Banerjee S, Roy A, Madhusudhan MS, Bairagya HR, Roy A. Structural insights of a cellobiose dehydrogenase enzyme from the basidiomycetes fungus Termitomyces clypeatus. Comput Biol Chem. 2019; doi:10.1016/j.compbiolchem.2019.05.013

32. Berman HM. The Protein Data Bank. Nucleic Acids Res. 2000;28: 235–242. doi:10.1093/nar/28.1.235

33. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. Journal of molecular biology. 1990. pp. 859–883. doi:10.1016/S0022-2836(05)80269-4

34. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules. 1985;18: 534–552. doi:10.1021/ma00145a039

35. Roy AA, Dhawanjewar AS, Sharma P, Singh G, Madhusudhan MS. Protein Interaction Z Score Assessment (PIZSA): an empirical scoring scheme for evaluation of protein-protein interactions. Nucleic Acids Res. 2019; doi:10.1093/nar/gkz368

36. Zhang C, Liu S, Zhu Q, Zhou Y. A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. J Med Chem. 2005; doi:10.1021/jm049314d

37. Dhawanjewar AS, Roy AA, Madhusudhan MS. A knowledge-based scoring function to assess the stability of quaternary protein assemblies. bioRxiv. 2019; doi:10.1101/562520

38. Poole AM, Ranganathan R. Knowledge-based potentials in protein design. Current Opinion in Structural Biology. 2006. doi:10.1016/j.sbi.2006.06.013

39. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006;15: 2507–2524. doi:10.1110/ps.062416606

40. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol. 1995;5: 229–235. doi:10.1016/0959-440X(95)80081-6

41. Bishop CM. Pattern Recoginiton and Machine Learning. Information Science and Statistics. 2006.

42. Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: A review of classification and combining techniques. Artif Intell Rev. 2006; doi:10.1007/s10462-007-9052-3

43. Francis L. Unsupervised learning. Predictive Modeling Applications in Actuarial Science: Volume I: Predictive Modeling Techniques. 2014. doi:10.1017/CBO9781139342674.012

44. Loh WY. Classification and regression trees. Wiley Interdiscip Rev Data Min Knowl Discov. 2011; doi:10.1002/widm.8

45. Breiman L. Random forests. Mach Learn. 2001; doi:10.1023/A:1010933404324

46. van Zundert GCP, Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis PL, Karaca E, et al. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. J Mol Biol. Academic Press; 2016;428: 720–725. doi:10.1016/J.JMB.2015.09.014

47. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem. NIH Public Access; 2009;30: 2785–91. doi:10.1002/jcc.21256

48. Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, Thomas V, et al. DOCK 6: combining techniques to model RNA-small molecule complexes. RNA. Cold Spring Harbor Laboratory Press; 2009;15: 1219–30. doi:10.1261/rna.1563609

49. Vangone A, Rodrigues JPGLM, Xue LC, van Zundert GCP, Geng C, Kurkcuoglu Z, et al. Sense and simplicity in HADDOCK scoring: Lessons from CASP-CAPRI round 1. Proteins Struct Funct Bioinforma. 2017; doi:10.1002/prot.25198

50. Jiménez-García B, Pons C, Fernández-Recio J. pyDockWEB: A web server for rigid-body protein-protein docking using electrostatics and desolvation scoring. Bioinformatics. 2013. doi:10.1093/bioinformatics/btt262

51. Torchala M, Moal IH, Chaleil RAG, Fernandez-Recio J, Bates PA. SwarmDock: A server for flexible protein-protein docking. Bioinformatics. 2013; doi:10.1093/bioinformatics/btt038

52. Pierce BG, Wiehe K, Hwang H, Kim BH, Vreven T, Weng Z. ZDOCK server: Interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics. 2014; doi:10.1093/bioinformatics/btu097

53. Lyskov S, Gray JJ. The RosettaDock server for local protein-protein docking. Nucleic Acids Res. 2008; doi:10.1093/nar/gkn216

54. OLEG TROTT AJO, Schroer A. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and

Multithreading. J Comput Chem. 2010; doi:10.1002/jcc

55. Lengauer T, Rarey M. Computational methods for biomolecular docking. Curr Opin Struct Biol. 1996; doi:10.1016/S0959-440X(96)80061-3

56. de Ruyck J, Brysbaert G, Blossey R, Lensink MF. Molecular docking as a popular tool in drug design, an in silico travel. Advances and Applications in Bioinformatics and Chemistry. 2016. doi:10.2147/AABC.S105289

57. Pagadala NS, Syed K, Tuszynski J. Software for molecular docking: a review. Biophysical Reviews. 2017. doi:10.1007/s12551-016-0247-1

58. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. Nature Structural Biology. 2002. doi:10.1038/nsb0902-646

59. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general Amber force field. J Comput Chem. 2004; doi:10.1002/jcc.20035

60. Buck M, Bouguet-Bonnet S, Pastor RW, MacKerell AD, Jr. Importance of the CMAP Correction to the CHARMM22 Protein Force Field: Dynamics of Hen Lysozyme. Biophys J. The Biophysical Society; 2006;90: L36. doi:10.1529/BIOPHYSJ.105.078154

61. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc. 1996; doi:10.1021/ja9621760

62. Oostenbrink C, Villa A, Mark AE, Van Gunsteren WF. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. J Comput Chem. 2004; doi:10.1002/jcc.20090

63. Hospital A, Goñi JR, Orozco M, Gelpí JL. Molecular dynamics simulations: Advances and applications. Advances and Applications in Bioinformatics and Chemistry. 2015. doi:10.2147/AABC.S70333

64. Zhou R. Replica exchange molecular dynamics method for protein folding simulation. Methods Mol Biol. 2007;350: 205–223. doi:10.1016/S0009-2614(99)01123-9

65. Berendsen HJC, Postma JPM, Van Gunsteren WF, Dinola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys. 1984; doi:10.1063/1.448118

66. Cheng X, Ivanov I. Molecular dynamics. Methods in Molecular Biology. 2012. doi:10.1007/978-1-62703-50-2_11

67. Srinivasan J, Cheatham TE, Cieplak P, Peter A. Kollman, and David A. Case. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate−DNA Helices. J Am Chem Soc. American Chemical Society; 1998;120. doi:10.1021/JA981844+

68. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, et al. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. American Chemical Society; 2000; doi:10.1021/AR000033J

69. Yang T, Wu JC, Yan C, Wang Y, Luo R, Gonzales MB, et al. Virtual screening using molecular simulations. Proteins. NIH Public Access; 2011;79: 1940–1951. doi:10.1002/prot.23018

70. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opinion on Drug Discovery. 2015. doi:10.1517/17460441.2015.1032936

71. Gapsys V, Michielssens S, Peters JH ennin., de Groot BL, Leonov H. Calculation of binding free energies. Methods Mol Biol. 2015; doi:10.1007/978-1-4939-1465-4_9

72. Cournia Z, Allen B, Sherman W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. Journal of Chemical Information and Modeling. 2017. doi:10.1021/acs.jcim.7b00564

73. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. BBA - Protein Struct. 1975;405: 442–451. doi:10.1016/0005-2795(75)90109-9

74. Rao VS, Srinivas K, Sujini GN, Kumar GNS. Protein-Protein Interaction Detection: Methods and Analysis. Int J Proteomics. 2014;2014: 1–12. doi:10.1155/2014/147648

75. Yamada T, Bork P. Evolution of biomolecular networks lessons from metabolic and protein interactions. Nature Reviews Molecular Cell Biology. 2009. doi:10.1038/nrm2787

76. Alberts B. The cell as a collection of protein machines: Preparing the next generation of molecular biologists. Cell. 1998. doi:10.1016/S0092-8674(00)80922-8

77. Ryan DP, Matthews JM. Protein-protein interactions in human disease. Current Opinion in Structural Biology. 2005. doi:10.1016/j.sbi.2005.06.001

78. Kuzmanov U, Emili A. Protein-protein interaction networks: Probing disease mechanisms using model systems. Genome Medicine. 2013. doi:10.1186/gm441

79. Zhou M, Li Q, Wang R. Current Experimental Methods for Characterizing Protein-Protein Interactions. ChemMedChem. 2016. doi:10.1002/cmdc.201500495

80. Xenarios I. DIP: the Database of Interacting Proteins. Nucleic Acids Res. 2000; doi:10.1093/nar/28.1.289

81. Isserlin R, El-Badrawi RA, Badery GD. The biomolecular interaction network database in PSI-MI 2.5. Database. 2011; doi:10.1093/database/baq037

82. Bader GD, Betel D, Hogue CWV. BIND: The Biomolecular Interaction Network Database. Nucleic Acids Research. 2003. doi:10.1093/nar/gkg056

83. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, et al. MINT: The Molecular INTeraction database. Nucleic Acids Res. 2007; doi:10.1093/nar/gkl950

84. Mosca R, Céol A, Aloy P. Interactome3D: Adding structural details to protein networks. Nat Methods. 2013; doi:10.1038/nmeth.2289

85. Kiel C, Beltrao P, Serrano L. Analyzing Protein Interaction Networks Using Structural Information. Annu Rev Biochem. 2008; doi:10.1146/annurev.biochem.77.062706.133317

86. Aloy P, Russell RB. Structural systems biology: Modelling protein interactions. Nature Reviews Molecular Cell Biology. 2006. doi:10.1038/nrm1859

87. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature. 2012; doi:10.1038/nature11503

88. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, et al. A human protein-protein interaction network: A resource for annotating the proteome. Cell. 2005; doi:10.1016/j.cell.2005.08.029

89. Shoemaker BA, Panchenko AR. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. PLoS Computational Biology. 2007. doi:10.1371/journal.pcbi.0030042

90. Kuzu G, Keskin O, Gursoy A, Nussinov R. Constructing structural networks of signaling pathways on the proteome scale. Current Opinion in Structural Biology. 2012. doi:10.1016/j.sbi.2012.04.004

91. Baspinar A, Cukuroglu E, Nussinov R, Keskin O, Gursoy A. PRISM: A web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. Nucleic Acids Res. 2014; doi:10.1093/nar/gku397

92. Guerler A, Govindarajoo B, Zhang Y. Mapping monomeric threading to protein-protein structure prediction. J Chem Inf Model. 2013; doi:10.1021/ci300579r

93. Hosur R, Peng J, Vinayagam A, Stelzl U, Xu J, Perrimon N, et al. A computational framework for boosting confidence in high-throughput protein-protein interaction datasets. Genome Biol. 2012; doi:10.1186/gb-2012-13-8-r76

94. Hosur R, Xu J, Bienkowska J, Berger B. IWRAP: An interface threading approach with application to prediction of cancer-related protein-protein interactions. J Mol Biol. 2011; doi:10.1016/j.jmb.2010.11.025

95. Soni N, Madhusudhan MS. Computational modeling of protein assemblies. Current Opinion in Structural Biology. 2017. doi:10.1016/j.sbi.2017.04.006

96. Vangone A, Oliva R, Cavallo L, Bonvin AMJJ. Prediction of biomolecular complexes. From Protein Structure to Function with Bioinformatics: Second Edition. 2017. doi:10.1007/978-94-024-1069-3_8

97. Rodrigues JPGLM, Bonvin AMJJ. Integrative computational modeling of protein interactions. FEBS Journal. 2014. doi:10.1111/febs.12771

98. Enright AJ, Illopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. Nature. 1999; doi:10.1038/47056

99. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995; doi:10.1016/S0022-2836(05)80134-2

100. Fox NK, Brenner SE, Chandonia JM. SCOPe: Structural Classification of Proteins

- Extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 2014; doi:10.1093/nar/gkt1240

101. Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, et al. CATH: An expanded resource to predict protein function through structure and sequence. Nucleic Acids Res. 2017; doi:10.1093/nar/gkw1098

102. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, et al. The CATH domain structure database: New protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res. 2007; doi:10.1093/nar/gkl959

103. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH - A hierarchic classification of protein domain structures. Structure. 1997; doi:10.1016/s0969-2126(97)00260-8

104. Mosca R, Céol A, Stein A, Olivella R, Aloy P. 3did: A catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. 2014; doi:10.1093/nar/gkt887

105. Stein A, Russell RB, Aloy P. 3did: Interacting protein domains of known three-dimensional structure. Nucleic Acids Research. 2005. doi:10.1093/nar/gki037

106. Stein A, Céol A, Aloy P. 3did: Identification and classification of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. 2011; doi:10.1093/nar/gkq962

107. Davis FP, Sali A. PIBASE: A comprehensive database of structurally defined protein interfaces. Bioinformatics. 2005; doi:10.1093/bioinformatics/bti277

108. Winter C. SCOPPI: a structural classification of protein-protein interfaces. Nucleic Acids Res. 2006; doi:10.1093/nar/gkj099

109. Jefferson ER, Walsh TP, Roberts TJ, Barton GJ. SNAPPI-DB: A database and API of structures, iNterfaces and Alignments for Protein-Protein Interactions. Nucleic Acids Res. 2007; doi:10.1093/nar/gkl836

110. Teyra J, Doms A, Schroeder M, Pisabarro MT. SCOWLP: A web-based database for detailed characterization and visualization of protein interfaces. BMC Bioinformatics. 2006; doi:10.1186/1471-2105-7-104

111. Teyra J, Samsonov SA, Schreiber S, Pisabarro MT. SCOWLP update: 3D

classification of protein-protein, -peptide, -saccharide and -nucleic acid interactions, and structure-based binding inferences across folds. BMC Bioinformatics. 2011; doi:10.1186/1471-2105-12-398

112. Xu Q, Dunbrack RL. The protein common interface database (ProtCID)-A comprehensive database of interactions of homologous proteins in multiple crystal forms. Nucleic Acids Res. 2011; doi:10.1093/nar/gkq1059

113. Dey S, Ritchie DW, Levy ED. PDB-wide identification of biological assemblies from conserved quaternary structure geometry. Nat Methods. 2018; doi:10.1038/nmeth.4510

114. Mirabello C, Wallner B. Topology independent structural matching discovers novel templates for protein interfaces. Bioinformatics. 2018. doi:10.1093/bioinformatics/bty587

115. Hubbard SJ, Argos P. Cavities and packing at protein interfaces. Protein Sci. 1994; doi:10.1002/pro.5560031205

116. Conte L Lo, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol. 1999; doi:10.1006/jmbi.1998.2439

117. McCoy AJ, Chandana Epa V, Colman PM. Electrostatic complementarity at protein/protein interfaces. J Mol Biol. 1997; doi:10.1006/jmbi.1997.0987

118. Valdar WSJ, Thornton JM. Protein-protein interfaces: Analysis of amino acid conservation in homodimers. Proteins Struct Funct Genet. 2001; doi:10.1002/1097-0134(20010101)42:1<108::AID-PROT110>3.0.CO;2-O

119. Ofran Y, Rost B. Analysing six types of protein-protein interfaces. J Mol Biol. 2003; doi:10.1016/S0022-2836(02)01223-8

120. Jones S, Thornton JM. Principles of protein-protein interactions. Proceedings of the National Academy of Sciences of the United States of America. 1996. doi:10.1073/pnas.93.1.13

121. Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. J Mol Biol. 2003; doi:10.1016/j.jmb.2003.07.006

122. Mika S, Rost B. Protein-protein interactions more conserved within species than across species. PLoS Comput Biol. 2006; doi:10.1371/journal.pcbi.0020079

123. Rekha N, Machado SM, Narayanan C, Krupa A, Srinivasan N. Interaction interfaces of protein domains are not topologically equivalent across families within superfamilies: Implications for metabolic and signaling pathways. Proteins Struct Funct Genet. 2005; doi:10.1002/prot.20319

124. Prabu MM, Suguna K, Vijayan M. Variability in quaternary association of proteins with the same tertiary fold: A case study and rationalization involving legume lectins. Proteins Struct Funct Genet. 1999; doi:10.1002/(SICI)1097-0134(19990401)35:1<58::AID-PROT6>3.0.CO;2-A

125. Iosub Amir A, Van Rosmalen M, Mayer G, Lebendiker M, Danieli T, Friedler A. Highly homologous proteins exert opposite biological activities by using different interaction interfaces. Sci Rep. 2015; doi:10.1038/srep11629

126. Park SY, Beel BD, Simon MI, Bilwes AM, Crane BR. In different organisms, the mode of interaction between two signaling proteins is not necessarily conserved. Proc Natl Acad Sci U S A. 2004; doi:10.1073/pnas.0401038101

127. Gao M, Skolnick J. Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. Proc Natl Acad Sci U S A. 2010; doi:10.1073/pnas.1012820107

128. Verma R, Pandit SB. Unraveling the structural landscape of intra-chain domain interfaces: Implication in the evolution of domain-domain interactions. PLoS One. 2019; doi:10.1371/journal.pone.0220336

129. Kundrotas PJ, Zhu Z, Janin J, Vakser IA. Templates are available to model nearly all complexes of structurally characterized proteins. Proc Natl Acad Sci U S A. 2012; doi:10.1073/pnas.1200678109

130. Tuncbag N, Keskin O, Nussinov R, Gursoy A. Fast and accurate modeling of protein-protein interactions by combining template-interface-based docking with flexible refinement. Proteins Struct Funct Bioinforma. 2012; doi:10.1002/prot.24022

131. Kuzu G, Gursoy A, Nussinov R, Keskin O. Exploiting conformational ensembles in modeling protein-protein interactions on the proteome scale. J Proteome Res. 2013; doi:10.1021/pr400006k

132. Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial

restraints. J Mol Biol. 1993;234: 779–815. doi:10.1006/jmbi.1993.1626

133. Henrick K, Thornton JM. PQS: A protein quaternary structure file server. Trends Biochem Sci. 1998; doi:10.1016/S0968-0004(98)01253-5

134. Davis FP, Braberg H, Shen MY, Pieper U, Sali A, Madhusudhan MS. Protein complex compositions predicted by structural similarity. Nucleic Acids Res. 2006; doi:10.1093/nar/gkl353

135. Chen J, Sawyer N, Regan L. Protein-protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area. Protein Sci. 2013; doi:10.1002/pro.2230

136. Yan C, Wu F, Jernigan RL, Dobbs D, Honavar V. Characterization of protein-protein interfaces. Protein J. 2008; doi:10.1007/s10930-007-9108-x

137. Bahadur RP, Chakrabarti P, Rodier F, Janin J. A Dissection of Specific and Non-specific Protein-Protein Interfaces. J Mol Biol. 2004; doi:10.1016/j.jmb.2003.12.073

138. Zhu H, Domingues FS, Sommer I, Lengauer T. NOXclass: Prediction of protein-protein interaction types. BMC Bioinformatics. 2006; doi:10.1186/1471-2105-7-27

139. Wang G, Dunbrack RL. PISCES: A protein sequence culling server. Bioinformatics. 2003;19: 1589–1591. doi:10.1093/bioinformatics/btg224

140. Bhattacharyya M, Upadhyay R, Vishveshwara S. Interaction Signatures Stabilizing the NAD(P)-Binding Rossmann Fold: A Structure Network Approach. PLoS One. 2012; doi:10.1371/journal.pone.0051676

141. Nagano N, Orengo CA, Thornton JM. One fold with many functions: The evolutionary relationships between TIM barrel families based on their sequences, structures and functions. Journal of Molecular Biology. 2002. doi:10.1016/S0022-2836(02)00649-6

142. Farheen N, Sen N, Nair S, Tan KP, Madhusudhan MS. Depth dependent amino acid substitution matrices and their use in predicting deleterious mutations. Progress in Biophysics and Molecular Biology. 2017. doi:10.1016/j.pbiomolbio.2017.02.004

143. Saha RP, Bahadur RP, Chakrabarti P. Interresidue contacts in proteins and protein-protein interfaces and their use in characterizing the homodimeric

interface. J Proteome Res. 2005; doi:10.1021/pr050118k

144. Arkin MR, Tang Y, Wells JA. Small-molecule inhibitors of protein-protein interactions: Progressing toward the reality. Chemistry and Biology. 2014. doi:10.1016/j.chembiol.2014.09.001

145. Corbi-Verge C, Kim PM. Motif mediated protein-protein interactions as drug targets. Cell Communication and Signaling. 2016. doi:10.1186/s12964-016-0131-4

146. Skwarczynska M, Ottmann C. Protein-protein interactions as drug targets. Future Medicinal Chemistry. 2015. doi:10.4155/fmc.15.138

147. Higueruelo AP, Jubb H, Blundell TL. Protein-protein interactions as druggable targets: Recent technological advances. Current Opinion in Pharmacology. 2013. doi:10.1016/j.coph.2013.05.009

148. Mason JM, Arndt KM. Coiled coil domains: Stability, specificity, and biological implications. ChemBioChem. 2004. doi:10.1002/cbic.200300781

149. Lupas AN, Bassler J. Coiled Coils – A Model System for the 21st Century. Trends in Biochemical Sciences. 2017. doi:10.1016/j.tibs.2016.10.007

150. Faix J, Steinmetz M, Boves H, Kammerer RA, Lottspeich F, Mintert U, et al. Cortexillins, major determinants of cell shape and size, are actin- bundling proteins with a parallel coiled-coil tail. Cell. 1996; doi:10.1016/S0092-8674(00)80136-1

151. Kohler JJ, Schepartz A. Kinetic studies of Fos·Jun·DNA complex formation: DNA binding prior to dimerization. Biochemistry. 2001; doi:10.1021/bi001881p

152. Walshaw J, Woolfson DN. SOCKET: A program for identifying and analysing coiled-coil motifs within protein structures. J Mol Biol. 2001; doi:10.1006/jmbi.2001.4545

153. Liu J, Zheng Q, Deng Y, Cheng CS, Kallenbach NR, Lu M. A seven-helix coiled coil. Proc Natl Acad Sci U S A. 2006; doi:10.1073/pnas.0604871103

154. Hicks MR, Walshaw J, Woolfson DN. Investigating the tolerance of coiled-coil peptides to nonheptad sequence inserts. Journal of Structural Biology. 2002. doi:10.1006/jsbi.2002.4462

155. O'Shea EK, Lumb KJ, Kim PS. Peptide "Velcro": Design of a heterodimeric coiled

coil. Curr Biol. 1993; doi:10.1016/0960-9822(93)90063-T

156. Grigoryan G, Degrado WF. Probing designability via a generalized model of helical bundle geometry. J Mol Biol. 2011; doi:10.1016/j.jmb.2010.08.058

157. Szczepaniak K, Ludwiczak J, Winski A, Dunin-Horkawicz S. Variability of the core geometry in parallel coiled-coil bundles. J Struct Biol. 2018; doi:10.1016/j.jsb.2018.07.002

158. Vincent TL, Green PJ, Woolfson DN. LOGICOIL - Multi-state prediction of coiled-coil oligomeric state. Bioinformatics. 2013; doi:10.1093/bioinformatics/bts648

159. Trigg J, Gutwin K, Keating AE, Berger B. Multicoil2: Predicting coiled coils and their oligomerization states from sequence in the twilight zone. PLoS One. 2011; doi:10.1371/journal.pone.0023519

160. Armstrong CT, Vincent TL, Green PJ, Woolfson DN. SCORER 2.0: An algorithm for distinguishing parallel dimeric and trimeric coiled-coil sequences. Bioinformatics. 2011; doi:10.1093/bioinformatics/btr299

161. Li C, Wang XF, Chen Z, Zhang Z, Song J. Computational characterization of parallel dimeric and trimeric coiled-coils using effective amino acid indices. Mol Biosyst. 2015; doi:10.1039/c4mb00569d

162. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. Science (80-). 1991; doi:10.1126/science.252.5009.1162

163. Gruber M, Söding J, Lupas AN. REPPER - Repeats and their periodicities in fibrous proteins. Nucleic Acids Res. 2005; doi:10.1093/nar/gki405

164. Delorenzi M, Speed T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics. 2002; doi:10.1093/bioinformatics/18.4.617

165. Bartoli L, Fariselli P, Krogh A, Casadio R. CCHMM_PROF: A HMM-based coiled-coil predictor with evolutionary information. Bioinformatics. 2009; doi:10.1093/bioinformatics/btp539

166. Ludwiczak J, Winski A, Szczepaniak K, Alva V, Dunin-Horkawicz S. DeepCoil-a fast and accurate prediction of coiled-coil domains in protein sequences. Bioinformatics. 2019; doi:10.1093/bioinformatics/bty1062

167. O'Shea EK, Klemm JD, Kim PS, Alber T. X-ray structure of the GCN4 leucine

zipper, a two-stranded, parallel coiled coil. Science (80-). 1991; doi:10.1126/science.1948029

168. Crick FHC. The packing of α-helices: simple coiled-coils. Acta Crystallogr. 1953; doi:10.1107/s0365110x53001964

169. Tripet B, Wagschal K, Lavigne P, Mant CT, Hodges RS. Effects of side-chain characteristics on stability and oligomerization state of a de novo-designed model coiled-coil: 20 Amino acid substitutions in position “d.” J Mol Biol. 2000; doi:10.1006/jmbi.2000.3866

170. Akey DL, Malashkevich VN, Kim PS. Buried polar residues in coiled-coil interfaces. Biochemistry. 2001; doi:10.1021/bi002829w

171. Testa OD, Moutevelis E, Woolfson DN. CC+: A relational database of coiled-coil structures. Nucleic Acids Res. 2009; doi:10.1093/nar/gkn675

172. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;

173. Saribas AS, Datta PK, Safak M. A comprehensive proteomics analysis of JC virus Agnoprotein-interacting proteins: Agnoprotein primarily targets the host proteins with coiled-coil motifs. Virology. 2020; doi:10.1016/j.virol.2019.10.005

174. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006; doi:10.1016/j.patrec.2005.10.010

175. Ferenczy MW, Marshall LJ, Nelson CDS, Atwood WJ, Nath A, Khalili K, et al. Molecular biology, epidemiology, and pathogenesis of progressive multifocal leukoencephalopathy, the JC virus-induced demyelinating disease of the human brain. Clinical Microbiology Reviews. 2012. doi:10.1128/CMR.05031-11

176. Soleimani-Meigooni DN, Schwetye KE, Angeles MR, Ryschkewitsch CF, Major EO, Dang X, et al. JC virus granule cell neuronopathy in the setting of chronic lymphopenia treated with recombinant interleukin-7. J Neurovirol. 2017; doi:10.1007/s13365-016-0465-0

177. Frisque RJ, Bream GL, Cannella MT. Human polyomavirus JC virus genome. J Virol. 1984; doi:10.1128/jvi.51.2.458-469.1984

178. Sariyer IK, Saribas AS, White MK, Safak M. Infection by agnoprotein-negative mutants of polyomavirus JC and SV40 results in the release of virions that are

mostly deficient in DNA content. Virol J. 2011; doi:10.1186/1743-422X-8-255

179. Safak M, Barrucco R, Darbinyan A, Okada Y, Nagashima K, Khalili K. Interaction of JC Virus Agno Protein with T Antigen Modulates Transcription and Replication of the Viral Genome in Glial Cells. J Virol. 2001; doi:10.1128/jvi.75.3.1476-1486.2001

180. Sariyer IK, Akan I, Palermo V, Gordon J, Khalili K, Safak M. Phosphorylation Mutants of JC Virus Agnoprotein Are Unable To Sustain the Viral Infection Cycle. J Virol. 2006; doi:10.1128/jvi.80.8.3893-3903.2006

181. Suzuki T, Orba Y, Makino Y, Okada Y, Sunden Y, Hasegawa H, et al. Viroporin activity of the JC polyomavirus is regulated by interactions with the adaptor protein complex 3. Proc Natl Acad Sci U S A. 2013; doi:10.1073/pnas.1311457110

182. Darbinyan A, Darbinian N, Safak M, Radhakrishnan S, Giordano A, Khalili K. Evidence for dysregulation of cell cycle by human polyomavirus, JCV, late auxiliary protein. Oncogene. 2002; doi:10.1038/sj.onc.1205744

183. Coric P, Saribas AS, Abou-Gharbia M, Childers W, White MK, Bouaziz S, et al. Nuclear Magnetic Resonance Structure Revealed that the Human Polyomavirus JC Virus Agnoprotein Contains an -Helix Encompassing the Leu/Ile/Phe-Rich Domain. J Virol. 2014; doi:10.1128/jvi.00146-14

184. Pierce B, Weng Z. ZRANK: Reranking protein docking predictions with an optimized energy function. Proteins Struct Funct Genet. 2007; doi:10.1002/prot.21373

185. Huang SY, Zou X. An iterative knowledge-based scoring function for protein-protein recognition. Proteins Struct Funct Genet. 2008; doi:10.1002/prot.21949

186. Liu S, Vakser IA. DECK: Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking. BMC Bioinformatics. 2011; doi:10.1186/1471-2105-12-280

187. Nadalin F, Carbone A. Protein–protein interaction specificity is captured by contact preferences and interface composition. Bioinformatics. Oxford University Press; 2018;34: 459–468. doi:10.1093/bioinformatics/btx584

188. Zimmermann MT, Leelananda SP, Kloczkowski A, Jernigan RL. Combining

statistical potentials with dynamics-based entropies improves selection from protein decoys and docking poses. J Phys Chem B. 2012; doi:10.1021/jp2120143

189. Pons C, Talavera D, De La Cruz X, Orozco M, Fernandez-Recio J. Scoring by intermolecular pairwise propensities of exposed residues (SIPPER): A new efficient potential for protein-protein docking. J Chem Inf Model. 2011; doi:10.1021/ci100353e

190. Moal IH, Barradas-Bautista D, Jiménez-García B, Torchala M, Van Der Velde A, Vreven T, et al. IRaPPA: Information retrieval based integration of biophysical models for protein assembly selection. Bioinformatics. 2017; doi:10.1093/bioinformatics/btx068

191. Bourquard T, Bernauer J, Azé J, Poupon A. A collaborative filtering approach for protein-protein docking scoring functions. PLoS One. 2011; doi:10.1371/journal.pone.0018541

192. Bordner AJ, Gorin AA. Protein docking using surface matching and supervised machine learning. Proteins Struct Funct Genet. 2007; doi:10.1002/prot.21406

193. Geng C, Jung Y, Renaud N, Honavar V, Bonvin AMJJ, Xue LC. iScore: a novel graph kernel-based function for scoring protein-protein docking models. Bioinformatics. 2020; doi:10.1093/bioinformatics/btz496

194. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol. 1996; doi:10.1006/jmbi.1996.0114

195. Tovchigrechko A, Vakser IA. GRAMM-X public web server for protein-protein docking. Nucleic Acids Res. 2006; doi:10.1093/nar/gkl206

196. Xue LC, Jordan RA, Yasser EM, Dobbs D, Honavar V. DockRank: Ranking docked conformations using partner-specific sequence homology-based protein interface prediction. Proteins Struct Funct Bioinforma. 2014; doi:10.1002/prot.24370

197. Andreani J, Faure G, Guerois R. InterEvScore: A novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. Bioinformatics. 2013; doi:10.1093/bioinformatics/btt260

198. Tress M, De Juan D, Graña O, Gómez MJ, Gómez-Puertas P, González JM, et al.

Scoring docking models with evolutionary information. Proteins: Structure, Function and Genetics. 2005. doi:10.1002/prot.20570

199. Lensink MF, Wodak SJ. Docking and scoring protein interactions: CAPRI 2009. Proteins Struct Funct Bioinforma. 2010; doi:10.1002/prot.22818

200. Lensink MF, Wodak SJ. Docking, scoring, and affinity prediction in CAPRI. Proteins Struct Funct Bioinforma. 2013; doi:10.1002/prot.24428

201. Lensink MF, Méndez R, Wodak SJ. Docking and scoring protein complexes: CAPRI 3rd Edition. Proteins: Structure, Function and Genetics. 2007. doi:10.1002/prot.21804

202. Lensink MF, Velankar S, Wodak SJ. Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. Proteins Struct Funct Bioinforma. 2017; doi:10.1002/prot.25215

203. Moal IH, Moretti R, Baker D, Fernández-Recio J. Scoring functions for protein-protein interactions. Current Opinion in Structural Biology. 2013. doi:10.1016/j.sbi.2013.06.017

204. Zhou H, Zhou Y. Quantifying the Effect of Burial of Amino Acid Residues on Protein Stability. Proteins Struct Funct Genet. 2004; doi:10.1002/prot.10584

205. Loladze V V., Ermolenko DN, Makhatadze GI. Thermodynamic consequences of burial of polar and non-polar amino acid residues in the protein interior. J Mol Biol. 2002; doi:10.1016/S0022-2836(02)00465-5

206. Melo F, Sánchez R, Sali A. Statistical potentials for fold assessment. Protein Sci. 2009; doi:10.1002/pro.110430

207. Bullock JMA, Sen N, Thalassinos K, Topf M. Modeling Protein Complexes Using Restraints from Crosslinking Mass Spectrometry. Structure. 2018; doi:10.1016/j.str.2018.04.016

208. Kundrotas PJ, Anishchenko I, Dauzhenka T, Kotthoff I, Mnevets D, Copeland MM, et al. Dockground: A comprehensive data resource for modeling of protein complexes. Protein Sci. 2018; doi:10.1002/pro.3295

209. Lensink MF, Wodak SJ. Score_set: A CAPRI benchmark for scoring protein complexes. Proteins Struct Funct Bioinforma. 2014; doi:10.1002/prot.24678

210. Moreira IS, Fernandes PA, Ramos MJ. Hot spots - A review of the protein-protein

interface determinant amino-acid residues. Proteins: Structure, Function and Genetics. 2007. doi:10.1002/prot.21396

211. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol. 1998; doi:10.1006/jmbi.1998.1843

212. Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. Science (80-). 1995; doi:10.1126/science.7529940

213. Keskin O, Ma B, Nussinov R. Hot regions in protein-protein interactions: The organization and contribution of structurally conserved hot spot residues. J Mol Biol. 2005; doi:10.1016/j.jmb.2004.10.077

214. Kouadio JLK, Horn JR, Pal G, Kossiakoff AA. Shotgun alanine scanning shows that growth hormone can bind productively to its receptor through a drastically minimized interface. J Biol Chem. 2005; doi:10.1074/jbc.M502167200

215. Thorn KS, Bogan AA. ASEdb: A database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics. 2001. doi:10.1093/bioinformatics/17.3.284

216. Morrison KL, Weiss GA. Combinatorial alanine-scanning. Current Opinion in Chemical Biology. 2001. doi:10.1016/S1367-5931(00)00206-4

217. Fischer TB, Arunachalam K V., Bailey D, Mangual V, Barkhru S, Russo R, et al. The binding inteference database (BID): A compilation of amino acid hot spots in protein interfaces. Bioinformatics. 2003;19: 1453–1454. doi:10.1093/bioinformatics/btg163

218. Kortemme T, Kim DE, Baker D. Computational alanine scanning of protein-protein interfaces. Sci STKE. 2004; doi:10.1126/stke.2192004pl2

219. Grosdidier S, Fernández-Recio J. Identification of hot-spot residues in protein-protein interactions by computational docking. BMC Bioinformatics. 2008; doi:10.1186/1471-2105-9-447

220. Huo S, Massova I, Kollman PA. Computational alanine scanning of the 1:1 human growth hormone-receptor complex. J Comput Chem. 2002; doi:10.1002/jcc.1153

221. Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein-protein complexes. Proc Natl Acad Sci U S A. 2002; doi:10.1073/pnas.202485799

222. Gao Y, Wang R, Lai L. Structure-based method for analyzing protein-protein interfaces. J Mol Model. 2004; doi:10.1007/s00894-003-0168-3

223. Tuncbag N, Keskin O, Gursoy A. HotPoint: Hot spot prediction server for protein interfaces. Nucleic Acids Res. 2010; doi:10.1093/nar/gkq323

224. Ofran Y, Rost B. Protein-protein interaction hotspots carved into sequences. PLoS Comput Biol. 2007; doi:10.1371/journal.pcbi.0030119

225. Darnell SJ, Page D, Mitchell JC. An automated decision-tree approach to predicting protein interaction hot spots. Proteins Struct Funct Genet. 2007; doi:10.1002/prot.21474

226. Cho KIi, Kim D, Lee D. A feature-based approach to modeling protein-protein interaction hot spots. Nucleic Acids Res. 2009; doi:10.1093/nar/gkp132

227. Xia JF, Zhao XM, Song J, Huang DS. APIS: Accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. BMC Bioinformatics. 2010; doi:10.1186/1471-2105-11-174

228. Zhu X, Mitchell JC. KFC2: A knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. Proteins Struct Funct Bioinforma. 2011; doi:10.1002/prot.23094

229. Assi SA, Tanaka T, Rabbitts TH, Fernandez-Fuentes N. PCRPi: Presaging Critical Residues in Protein interfaces, a new computational tool to chart hot spots in protein interfaces. Nucleic Acids Res. 2009; doi:10.1093/nar/gkp1158

230. Wang L, Liu ZP, Zhang XS, Chen L. Prediction of hot spots in protein interfaces using a random forest model with hybrid features. Protein Eng Des Sel. 2012; doi:10.1093/protein/gzr066

231. Deng L, Zhang QC, Chen Z, Meng Y, Guan J, Zhou S. PredHS: A web server for predicting protein-protein interaction hot spots by using structural neighborhood properties. Nucleic Acids Res. 2014; doi:10.1093/nar/gku437

232. Hawkins DM. The Problem of Overfitting. Journal of Chemical Information and Computer Sciences. 2004. doi:10.1021/ci0342472

233. Visa S, Ralescu A. Issues in mining imbalanced data sets-a review paper. Proceedings of the sixteen midwest artificial intelligence and cognitive science conference. 2005.

234. Moreira IS, Fernandes PA, Ramos MJ. Hot spot occlusion from bulk water: A comprehensive study of the complex between the lysozyme HEL and the antibody FVD1.3. J Phys Chem B. 2007; doi:10.1021/jp067096p

235. Desrosiers DC, Peng ZY. A binding free energy hot spot in the ankyrin repeat protein GABPβ mediated protein-protein interaction. J Mol Biol. 2005; doi:10.1016/j.jmb.2005.09.045

236. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. Evolutionary rate in the protein interaction network. Science (80-). 2002; doi:10.1126/science.1068696

237. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci U S A. 2003; doi:10.1073/pnas.1030237100

238. Caffrey DR. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci. 2004; doi:10.1110/ps.03323604

239. Tuncbag N, Gursoy A, Keskin O. Identification of computational hot spots in protein interfaces: Combining solvent accessibility and inter-residue potentials improves the accuracy. Bioinformatics. 2009; doi:10.1093/bioinformatics/btp240

240. Dhawanjewar A. Statistical Potentials for Prediction of Protein-Protein Interactions [Internet]. 2015. Available: http://dr.iiserpune.ac.in:8080/xmlui/handle/123456789/466

241. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25: 3389–402.

242. Lin J. Divergence Measures Based on the Shannon Entropy. IEEE Trans Inf Theory. 1991;37: 145–151. doi:10.1109/18.61115

243. Gao M, Skolnick J. A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins. PLoS Comput Biol. 2013; doi:10.1371/journal.pcbi.1003302

244. Gold ND, Jackson RM. Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. J Mol Biol. 2006; doi:10.1016/j.jmb.2005.11.044

245. International drug monitoring: the role of national centres. Report of a WHO

meeting. World Heal Organ - Tech Rep Ser. 1972;

246. Campillos M, Kuhn M, Gavin A-C, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science. 2008; doi:10.1126/science.1158140

247. Cohen P, Tcherpakov M. Will the ubiquitin system furnish as many drug targets as protein kinases? Cell. 2010; doi:10.1016/j.cell.2010.11.016

248. Uitdehaag JCM, Verkaar F, Alwan H, De Man J, Buijsman RC, Zaman GJR. A guide to picking the most selective kinase inhibitor tool compounds for pharmacological validation of drug targets. British Journal of Pharmacology. 2012. doi:10.1111/j.1476-5381.2012.01859.x

249. Perry ME. The regulation of the p53-mediated stress response by MDM2 and MDM4. Cold Spring Harbor perspectives in biology. 2010. doi:10.1101/cshperspect.a000968

250. Momand J, Zambetti GP, Olson DC, George D, Levine AJ. The mdm-2 oncogene product forms a complex with the p53 protein and inhibits p53-mediated transactivation. Cell. 1992; doi:10.1016/0092-8674(92)90644-R

251. Brown CJ, Cheok CF, Verma CS, Lane DP. Reactivation of p53: From peptides to small molecules. Trends Pharmacol Sci. 2011; doi:10.1016/j.tips.2010.11.004

252. Hoe KK, Verma CS, Lane DP. Drugging the p53 pathway: Understanding the route to clinical efficacy. Nature Reviews Drug Discovery. 2014. doi:10.1038/nrd4236

253. Joseph TL, Madhumalar A, Brown CJ, Lane DP, Verma C. Differential binding of p53 and nutlin to MDM2 and MDMX: Computational studies. Cell Cycle. 2010; doi:10.4161/cc.9.6.11067

254. Vu B, Wovkulich P, Pizzolato G, Lovey A, Ding Q, Jiang N, et al. Discovery of RG7112: A small-molecule MDM2 inhibitor in clinical development. ACS Med Chem Lett. 2013; doi:10.1021/ml4000657

255. Kussie PH, Gorina S, Marechal V, Elenbaas B, Moreau J, Levine AJ, et al. Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. Science (80-). 1996; doi:10.1126/science.274.5289.948

256. Berberich SJ. RNAi knockdown of HdmX or Hdm2 leads to new insights into p53 signaling. Cell Cycle. 2010. doi:10.4161/cc.9.18.13255

257. Secchiero P, Bosco R, Celeghini C, Zauli G. Recent Advances in the Therapeutic Perspectives of Nutlin-3. Curr Pharm Des. 2011; doi:10.2174/138161211795222586

258. Burgess A, Chia KM, Haupt S, Thomas D, Haupt Y, Lim E. Clinical Overview of MDM2/X-Targeted Therapies. Front Oncol. 2016;6: 1–7. doi:10.3389/fonc.2016.00007

259. Case DA, Darden TA, Cheatham TE, Simmerling CL, Wang J, Duke RE, et al. AMBER 11. University of California, San Francisco. 2010. doi:10.1017/CBO9781107415324.004

260. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. The Universal Protein Resource (UniProt). Nucleic Acids Res. 2005;33: 154–159. doi:10.1093/nar/gki070

261. Jiménez-Marín Á, Collado-Romero M, Ramirez-Boo M, Arce C, Garrido JJ. Biological pathway analysis by ArrayUnlock and Ingenuity Pathway Analysis. BMC Proc. 2009; doi:10.1186/1753-6561-3-s4-s6

262. Berendsen HJC, van der Spoel D, van Drunen R. GROMACS: A message-passing parallel molecular dynamics implementation. Comput Phys Commun. 1995; doi:10.1016/0010-4655(95)00042-E

263. Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, et al. GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics. 2013; doi:10.1093/bioinformatics/btt055

264. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins Struct Funct Bioinforma. 2010; doi:10.1002/prot.22711

265. Berendsen HJC, Grigera JR, Straatsma TP. The missing term in effective pair potentials. J Phys Chem. 1987; doi:10.1021/j100308a038

266. Wang J, Wang W, Kollman PA, Case DA. Automatic atom type and bond type perception in molecular mechanical calculations. J Mol Graph Model. 2006; doi:10.1016/j.jmgm.2005.12.005

267. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A Linear Constraint Solver for molecular simulations. J Comput Chem. 1997; doi:10.1002/(SICI)1096-

987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H

268. Parrinello M, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method. J Appl Phys. 1981; doi:10.1063/1.328693

269. ElSawy KM, Verma CS, Lane DP, Caves LSD. On the origin of the stereoselective affinity of Nutlin-3 geometrical isomers for the MDM2 protein. Cell Cycle. 2013; doi:10.4161/cc.27273

270. Jerzy Świerkot, Ryszard Ślęzak, Paweł Karpiński JP, Leszek Noga JS, Wiland P. Associations between single-nucleotide polymorphisms of RFC-1, GGH, MTHFR, TYMS, and TCII genes and the efficacy and toxicity of methotrexate treatment in patients with rheumatoid arthritis. Pol Arch Med Wewn. 2015;

271. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera - A visualization system for exploratory research and analysis. J Comput Chem. 2004; doi:10.1002/jcc.20084

272. Ren J, Xie L, Li WW, Bourne PE. SMAP-WS: A parallel web service for structural proteome-wide ligand-binding site comparison. Nucleic Acids Res. 2010; doi:10.1093/nar/gkq400

273. Wang JC, Chu PY, Chen CM, Lin JH. idTarget: A web server for identifying protein targets of small chemical molecules with robust scoring functions and a divide-and-conquer docking approach. Nucleic Acids Res. 2012; doi:10.1093/nar/gks496

274. Shin JS, Ha JH, He F, Muto Y, Ryu KS, Yoon HS, et al. Structural insights into the dual-targeting mechanism of Nutlin-3. Biochem Biophys Res Commun. 2012; doi:10.1016/j.bbrc.2012.02.113

275. Jafari R, Almqvist H, Axelsson H, Ignatushchenko M, Lundbäck T, Nordlund P, et al. The cellular thermal shift assay for evaluating drug target interactions in cells. Nat Protoc. 2014; doi:10.1038/nprot.2014.138

276. Amritkar K. Assessing, predicting and designing peptide ligands for proteins [Internet]. Dept. of Biology. 2020. Available: http://dr.iiserpune.ac.in:8080/xmlui/handle/123456789/4782

277. Drugs - DrugBank [Internet].

278. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The

Protein Data Bank. Nucleic Acids Res. 2000;28: 235–242. doi:10.1093/nar/28.1.235

279. Robertson MJ, Tirado-Rives J, Jorgensen WL. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. J Chem Theory Comput. 2015; doi:10.1021/acs.jctc.5b00356

280. Dodda LS, De Vaca IC, Tirado-Rives J, Jorgensen WL. LigParGen web server: An automatic OPLS-AA parameter generator for organic ligands. Nucleic Acids Res. 2017; doi:10.1093/nar/gkx312

281. Barelier S, Sterling T, O’Meara MJ, Shoichet BK. The Recognition of Identical Ligands by Unrelated Proteins. ACS Chem Biol. 2015; doi:10.1021/acschembio.5b00683

282. Ma B, Shatsky M, Wolfson HJ, Nussinov R. Multiple diverse ligands binding at a single protein site: A matter of pre-existing populations. Protein Sci. 2009; doi:10.1110/ps.21302

283. Venkatakrishnan AJ, Deupi X, Lebon G, Tate CG, Schertler GF, Madan Babu M. Molecular signatures of G-protein-coupled receptors. Nature. 2013. doi:10.1038/nature11896

284. Arunkumar G, Chandni R, Mourya DT, Singh SK, Sadanandan R, Sudan P, et al. Outbreak Investigation of Nipah Virus Disease in Kerala, India, 2018. J Infect Dis. Narnia; 2019;219: 1867–1878. doi:10.1093/infdis/jiy612

285. Spiropoulou CF. Nipah Virus Outbreaks: Still Small but Extremely Lethal. J Infect Dis. Narnia; 2019;219: 1855–1857. doi:10.1093/infdis/jiy611

286. Looi L-M, Chua K-B. Lessons from the Nipah virus outbreak in Malaysia. Malays J Pathol. 2007;29: 63–7.

287. WHO | Nipah R&amp;D. WHO. World Health Organization; 2018;

288. Rajbari M, Rajbari N, Faridpur F. Morbidity and mortality due to Nipah or Nipah-like virus encephalitis in WHO South-East Asia Region Country: India. 2018; 2018.

289. WHO | Nipah virus infection. WHO. World Health Organization; 2018;

290. Thibault PA, Watkinson RE, Moreira-Soto A, Drexler JF, Lee B. Zoonotic Potential of Emerging Paramyxoviruses. Advances in virus research. 2017. pp. 1–55.

doi:10.1016/bs.aivir.2016.12.001

291. Simons RRL, Gale P, Horigan V, Snary EL, Breed AC. Potential for introduction of bat-borne zoonotic viruses into the EU: a review. Viruses. Multidisciplinary Digital Publishing Institute (MDPI); 2014;6: 2084–2121. doi:10.3390/v6052084

292. Luby SP. The pandemic potential of Nipah virus. Antiviral Research. 2013. doi:10.1016/j.antiviral.2013.07.011

293. Banerjee S, Niyas VKM, Soneja M, Shibeesh AP, Basheer M, Sadanandan R, et al. First experience of ribavirin postexposure prophylaxis for Nipah virus, tried during the 2018 outbreak in Kerala, India. J Infect. W.B. Saunders; 2019;78: 491–503. doi:10.1016/J.JINF.2019.03.005

294. CEPI Awards $25 Million Contract to Profectus BioSciences and Emergent BioSolutions to Develop Nipah Virus Vaccine | CEPI [Internet]. Available: http://cepi.net/news/cepi-awards-25-million-contract-profectus-biosciences-and-emergent-biosolutions-develop-nipah

295. Who, Searo. Fact Sheet - Nipah Virus Infection.

296. Chong H-T, Kamarulzaman A, Tan C-T, Goh K-J, Thayaparan T, Kunjapan SR, et al. Treatment of acute Nipah encephalitis with ribavirin. Ann Neurol. John Wiley & Sons, Ltd; 2001;49: 810–813. doi:10.1002/ana.1062

297. Hotard AL, He B, Nichol ST, Spiropoulou CF, Lo MK. 4′-Azidocytidine (R1479) inhibits henipaviruses and other paramyxoviruses with high potency. Antiviral Res. Elsevier; 2017;144: 147–152. doi:10.1016/J.ANTIVIRAL.2017.06.011

298. Nguyen NM, Tran CNB, Phung LK, Duong KTH, Huynh H le A, Farrar J, et al. A Randomized, Double-Blind Placebo Controlled Trial of Balapiravir, a Polymerase Inhibitor, in Adult Dengue Patients. J Infect Dis. 2013;207: 1442–1450. doi:10.1093/infdis/jis470

299. Nelson DR, Zeuzem S, Andreone P, Ferenci P, Herring R, Jensen DM, et al. Balapiravir plus peginterferon alfa-2a (40KD)/ribavirin in a randomized trial of hepatitis C genotype 1 patients. Ann Hepatol. 11: 15–31. Available: http://www.ncbi.nlm.nih.gov/pubmed/22166557

300. Roberts SK, Cooksley G, Dore GJ, Robson R, Shaw D, Berns H, et al. Robust antiviral activity of R1626, a novel nucleoside analog: A randomized, placebo-

controlled study in patients with chronic hepatitis C. Hepatology. 2008;48: 398–406. doi:10.1002/hep.22321

301. Dawes BE, Kalveram B, Ikegami T, Juelich T, Smith JK, Zhang L, et al. Favipiravir (T-705) protects against Nipah virus infection in the hamster model. Sci Rep. Nature Publishing Group; 2018;8: 7604. doi:10.1038/s41598-018-25780-3

302. Goldhill DH, Te Velthuis AJW, Fletcher RA, Langat P, Zambon M, Lackenby A, et al. The mechanism of resistance to favipiravir in influenza. Proc Natl Acad Sci U S A. National Academy of Sciences; 2018;115: 11613–11618. doi:10.1073/pnas.1811345115

303. Bossart KN, Zhu Z, Middleton D, Klippel J, Crameri G, Bingham J, et al. A Neutralizing Human Monoclonal Antibody Protects against Lethal Disease in a New Ferret Model of Acute Nipah Virus Infection. Basler CF, editor. PLoS Pathog. Public Library of Science; 2009;5: e1000642. doi:10.1371/journal.ppat.1000642

304. Xu K, Rockx B, Xie Y, DeBuysscher BL, Fusco DL, Zhu Z, et al. Crystal structure of the Hendra virus attachment G glycoprotein bound to a potent cross-reactive neutralizing human monoclonal antibody. PLoS Pathog. Public Library of Science; 2013;9: e1003684. doi:10.1371/journal.ppat.1003684

305. Rockx B, Winegar R, Freiberg AN. Recent progress in henipavirus research: Molecular biology, genetic diversity, animal models. Antiviral Res. Elsevier B.V.; 2012;95: 135–149. doi:10.1016/j.antiviral.2012.05.008

306. Bonaparte MI, Dimitrov AS, Bossart KN, Crameri G, Mungall BA, Bishop KA, et al. Ephrin-B2 ligand is a functional receptor for Hendra virus and Nipah virus. Proc Natl Acad Sci U S A. National Academy of Sciences; 2005;102: 10652–10657. doi:10.1073/pnas.0504887102

307. Diederich S, Dietzel E, Maisner A. Nipah virus fusion protein: Influence of cleavage site mutations on the cleavability by cathepsin L, trypsin and furin. Virus Res. 2009;145: 300–306. doi:10.1016/j.virusres.2009.07.020

308. Lee B, Ataman ZA. Modes of paramyxovirus fusion: A Henipavirus perspective. Trends in Microbiology. 2011. doi:10.1016/j.tim.2011.03.005

309. Omi-Furutani M, Yoneda M, Fujita K, Ikeda F, Kai C. Novel phosphoprotein-interacting region in Nipah virus nucleocapsid protein and its involvement in viral

replication. J Virol. American Society for Microbiology; 2010;84: 9793–9799. doi:10.1128/JVI.00339-10

310. Jordan PC, Liu C, Raynaud P, Lo MK, Spiropoulou CF, Symons JA, et al. Initiation, extension, and termination of RNA synthesis by a paramyxovirus polymerase. Rey FA, editor. PLOS Pathog. Public Library of Science; 2018;14: e1006889. doi:10.1371/journal.ppat.1006889

311. Battisti AJ, Meng G, Winkler DC, McGinnes LW, Plevka P, Steven AC, et al. Structure and assembly of a paramyxovirus matrix protein. Proc Natl Acad Sci U S A. National Academy of Sciences; 2012;109: 13996–14000. doi:10.1073/pnas.1210275109

312. Watkinson RE, Lee B. Nipah virus matrix protein: expert hacker of cellular machines. FEBS Lett. NIH Public Access; 2016;590: 2494–2511. doi:10.1002/1873-3468.12272

313. Yoneda M, Guillaume V, Sato H, Fujita K, Georges-Courbot M-C, Ikeda F, et al. The Nonstructural Proteins of Nipah Virus Play a Key Role in Pathogenicity in Experimentally Infected Animals. Masucci MG, editor. PLoS One. Public Library of Science; 2010;5: e12709. doi:10.1371/journal.pone.0012709

314. Mohammed AA, Shantier SW, Mustafa MI, Osman HK, Elmansi HE, Osman I-AA, et al. Epitope - based peptide vaccine against glycoprotein G of Nipah henipavirus using immunoinformatics approaches. bioRxiv. Cold Spring Harbor Laboratory; 2019; 678664. doi:10.1101/678664

315. Kamthania M, Sharma DK. Epitope-Based Peptides Prediction from Proteome of Nipah Virus. Int J Pept Res Ther. Springer Netherlands; 2016;22: 465–470. doi:10.1007/s10989-016-9526-8

316. Chan YP, Chua KB, Koh CL, Lim ME, Lam SK. Complete nucleotide sequences of Nipah virus isolates from Malaysia. J Gen Virol. 2001;82: 2151–2155. doi:10.1099/0022-1317-82-9-2151

317. Sánchez R, Sali A. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proc Natl Acad Sci U S A. 1998;95: 13597–602.

318. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993;234: 779–815. doi:10.1006/jmbi.1993.1626

319. Webb B, Sali A. Comparative protein structure modeling using MODELLER. Curr Protoc Bioinforma. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2016;2016: 5.6.1-5.6.37. doi:10.1002/cpbi.3

320. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008;9: 40. doi:10.1186/1471-2105-9-40

321. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: A Free Tool to Discover Chemistry for Biology. 2012; doi:10.1021/ci3001277

322. Irwin JJ, Shoichet BK. ZINC-A Free Database of Commercially Available Compounds for Virtual Screening.

323. Mukherjee S, Balius TE, Rizzo RC. Docking Validation Resources: Protein Family and Ligand Flexibility Experiments. J Chem Inf Model. American Chemical Society; 2010;50: 1986–2000. doi:10.1021/ci1001982

324. Zoete V, Cuendet MA, Grosdidier A, Michielin O. SwissParam: A fast force field generation tool for small organic molecules. J Comput Chem. 2011; doi:10.1002/jcc.21816

325. Darden T, York D, Pedersen L. Particle mesh Ewald: An N ·log(N) method for Ewald sums in large systems. J Chem Phys. American Institute of Physics; 1993;98: 10089–10092. doi:10.1063/1.464397

326. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci U S A. National Academy of Sciences; 2001;98: 10037–10041. doi:10.1073/pnas.181342398

327. Paissoni C, Spiliotopoulos D, Musco G, Spitaleri A. GMXPBSA 2.1: A GROMACS tool to perform MM/PBSA and computational alanine scanning. Comput Phys Commun. North-Holland; 2015;186: 105–107. doi:10.1016/J.CPC.2014.09.010

328. Duan L, Liu X, Zhang JZH. Interaction Entropy: A New Paradigm for Highly Efficient and Reliable Computation of Protein–Ligand Binding Free Energy. J Am Chem Soc. American Chemical Society; 2016;138: 5722–5728. doi:10.1021/jacs.6b02682

329. Harcourt BH, Lowe L, Tamin A, Liu X, Bankamp B, Bowden N, et al. Genetic characterization of Nipah virus, Bangladesh, 2004. Emerg Infect Dis. 2005;

doi:10.3201/eid1110.050513

330. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2007;36: D13–D21. doi:10.1093/nar/gkm1000

331. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. Methods Mol Biol. 2007;406: 89–112.

332. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; doi:10.1093/nar/gkh340

333. Soni N. Computational Modeling Of 3d Structure Of Keratin Intermediate Filament By Satisfaction Of Spatial Restraints. IISER Pune. 2020.

334. Sen N, Kanitkar TR, Roy AA, Soni N, Amritkar K, Supekar S, et al. Predicting and designing therapeutics against the Nipah virus. PLoS Negl Trop Dis. 2019; doi:10.1371/journal.pntd.0007419

335. Bullock BN, Jochim AL, Arora PS. Assessing helical protein interfaces for inhibitor design. J Am Chem Soc. 2011; doi:10.1021/ja206074j

336. Siegert TR, Bird MJ, Makwana KM, Kritzer JA. Analysis of Loops that Mediate Protein-Protein Interactions and Translation into Submicromolar Inhibitors. J Am Chem Soc. 2016; doi:10.1021/jacs.6b05656

337. Xu K, Rajashankar KR, Chan Y-P, Himanen JP, Broder CC, Nikolov DB. Host cell recognition by the henipaviruses: crystal structures of the Nipah G attachment glycoprotein and its complex with ephrin-B3. Proc Natl Acad Sci U S A. National Academy of Sciences; 2008;105: 9953–9958. doi:10.1073/pnas.0804797105

338. Bowden TA, Aricescu AR, Gilbert RJC, Grimes JM, Jones EY, Stuart DI. Structural basis of Nipah and Hendra virus attachment to their cell-surface receptor ephrin-B2. Nat Struct Mol Biol. 2008; doi:10.1038/nsmb.1435

339. Anfinsen CB. Principles that Govern the Folding of Protein Chains. Science (80-). 1973;181: 223–230. doi:10.1126/science.181.4096.223

340. Chothia C. One thousand families for the molecular biologist. Nature. 1992. doi:10.1038/357543a0

341. Wolynes PG, Onuchic JN, Thirumalai D. Navigating the folding routes. Science. 1995;267: 1619–1620. doi:10.1126/science.7886447

342. Wang Z, Moult J. SNPs, protein structure, and disease. Hum Mutat. 2001;17: 263–270. doi:10.1002/humu.22

343. Dayhoff M, Schwartz R. A Model of Evolutionary Change in Proteins. Atlas protein Seq Struct. 1978; 345–352. doi:10.1.1.145.4315

344. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992;89: 10915–10919. doi:10.1073/pnas.89.22.10915

345. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–10. doi:10.1016/S0022-2836(05)80360-2

346. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, et al. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 2003;31: 3497–3500. doi:10.1093/nar/gkg500

347. Goldman N, Thorne JL, Jones DT. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics. 1998;149: 445–458. doi:10.1093/molbev/msl086

348. Rice DW, Eisenberg D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol. 1997;267: 1026–38. doi:10.1006/jmbi.1997.0924

349. Lüthy R, McLachlan a D, Eisenberg D. Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. Proteins. 1991;10: 229–239. doi:10.1002/prot.340100307

350. Overington J, Johnson MS, Sali a, Blundell TL. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proceedings. Biological sciences / The Royal Society. 1990. pp. 132–145. doi:10.1098/rspb.1990.0077

351. Dimmic MW, Rest JS, Mindell DP, Goldstein RA. rtREV: An amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J Mol Evol. 2002;55: 65–73. doi:10.1007/s00239-001-2304-y

352. Abascal F, Posada D, Zardoya R. MtArt: A new model of amino acid replacement for Arthropoda. Mol Biol Evol. 2007;24: 1–5. doi:10.1093/molbev/msl136

353. Arvestad L. Efficient methods for estimating amino acid replacement rates. J Mol

Evol. 2006;62: 663–673. doi:10.1007/s00239-004-0113-9

354. Jones DT, Taylor WR, Thornton JM. A mutation data matrix for transmembrane proteins. FEBS Lett. 1994;339: 269–275. doi:10.1016/0014-5793(94)80429-X

355. Adachi J, Hasegawa M. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol. 1996;42: 459–468. doi:8642615

356. Koshi JM, Goldstein RA. Context-dependent optimal substitution matrices. Protein Eng Des Sel. 1995;8: 641–645. doi:10.1093/protein/8.7.641

357. Johnson MS, Overington JP, Blundell TL. Alignment and searching for common protein folds using a data bank of structural templates. J Mol Biol. 1993;231: 735–52. doi:10.1006/jmbi.1993.1323

358. Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 2004;21: 1095–1109. doi:10.1093/molbev/msh112

359. Thorne JL, Goldman N, Jones DT. Combining protein evolution and secondary structure. Mol Biol Evol. 1996;13: 666–673. doi:10.1093/oxfordjournals.molbev.a025627

360. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol. 2001;310: 243–257. doi:10.1006/jmbi.2001.4762

361. Madhusudhan MS, Webb BM, Marti-Renom MA, Eswar N, Sali A. Alignment of multiple protein structures based on sequence and structure features. Protein Eng Des Sel. 2009;22: 569–574. doi:10.1093/protein/gzp040

362. Mehta PK, Heringa J, Argos P. A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. Protein Sci. 1995;4: 2517–25. doi:10.1002/pro.5560041208

363. Wako H, Blundell TL. Use of AA env-dependent substitution tables and conf propensities in struc prediction from aligned sequences of homologous proteins. II. Secondary struc. J Mol Biol. 1994;238: 693–708. doi:10.1006/jmbi.1994.1330

364. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31: 3812–3814. doi:10.1093/nar/gkg509

365. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7: 248–249. doi:10.1038/nmeth0410-248

366. Yates CM, Filippis I, Kelley LA, Sternberg MJE. SuSPect: Enhanced prediction of single amino acid variant (SAV) phenotype using network features. J Mol Biol. 2014;426: 2692–2701. doi:10.1016/j.jmb.2014.04.026

367. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 2005;33. doi:10.1093/nar/gki375

368. Masso M, Vaisman II. AUTO-MUTE: Web-based tools for predicting stability changes in proteins due to single amino acid replacements. Protein Eng Des Sel. 2010;23: 683–687. doi:10.1093/protein/gzq042

369. Pires DE V, Ascher DB, Blundell TL. MCSM: Predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics. 2014;30: 335–342. doi:10.1093/bioinformatics/btt691

370. Worth CL, Preissner R, Blundell TL. SDM--a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res. 2011;39: W215-22. doi:10.1093/nar/gkr363

371. Pires DE V, Ascher DB, Blundell TL. DUET: A server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res. 2014;42. doi:10.1093/nar/gku411

372. Rennell D, Bouvier SE, Hardy LW, Poteete AR. Systematic mutation of bacteriophage T4 lysozyme. J Mol Biol. 1991;222. doi:10.1016/0022-2836(91)90738-R

373. Adkar B V., Tripathi A, Sahoo A, Bajaj K, Goswami D, Chakrabarti P, et al. Protein model discrimination using mutational sensitivity derived from deep sequencing. Structure. 2012;20: 371–381. doi:10.1016/j.str.2011.11.021

374. Arti Tripathia, Kritika Gupta, Shruti Kharea, Pankaj C. Jaina, Siddharth Patela, Prasanth Kumara, Ajai J. Pulianmackala, Nilesh Agheraa RV. Molecular determinants of mutant phenotypes, inferred from saturation mutagenesis data. Mol Biol Evol. 2016; 1–35. doi:10.1093/molbev/msw182

375. Fiser a, Fiser A, Do RK, Do RK, Sali A, Sali A. Modeling of loops in protein structures. Protein Sci. 2000;9: 1753–73. doi:10.1110/ps.9.9.1753

376. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992;89: 10915–10919. doi:10.1073/pnas.89.22.10915

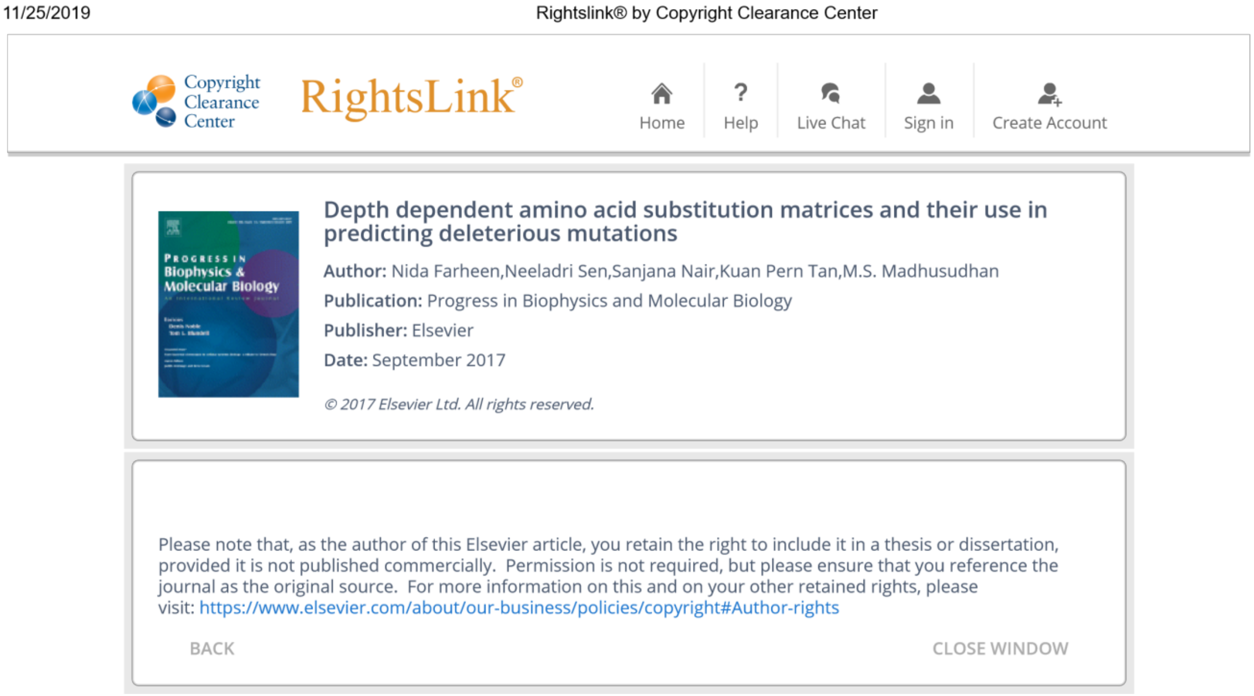
377. Weaver LH, Matthews BW. Structure of bacteriophage T4 lysozyme refined at 1.7 ?? resolution. J Mol Biol. 1987;193: 189–199. doi:10.1016/0022-2836(87)90636-X

378. Loris R, Dao-Thi MH, Bahassi EM, Van Melderen L, Poortmans F, Liddington R, et al. Crystal structure of CcdB, a topoisomerase poison from E. coli. J Mol Biol. 1999;285: 1667–1677. doi:10.1006/jmbi.1998.2395

379. Topham CM, Srinivasan N, Blundell TL. Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. Protein Eng. 1997;10: 7–21. doi:10.1093/protein/10.1.7

Copyright forms

11/25/2019 Rightslink® by Copyright Clearance Center



Depth dependent amino acid substitution matrices and their use in predicting deleterious mutations

Author: Nida Farheen,Neeladri Sen,Sanjana Nair,Kuan Pern Tan,M.S. Madhusudhan

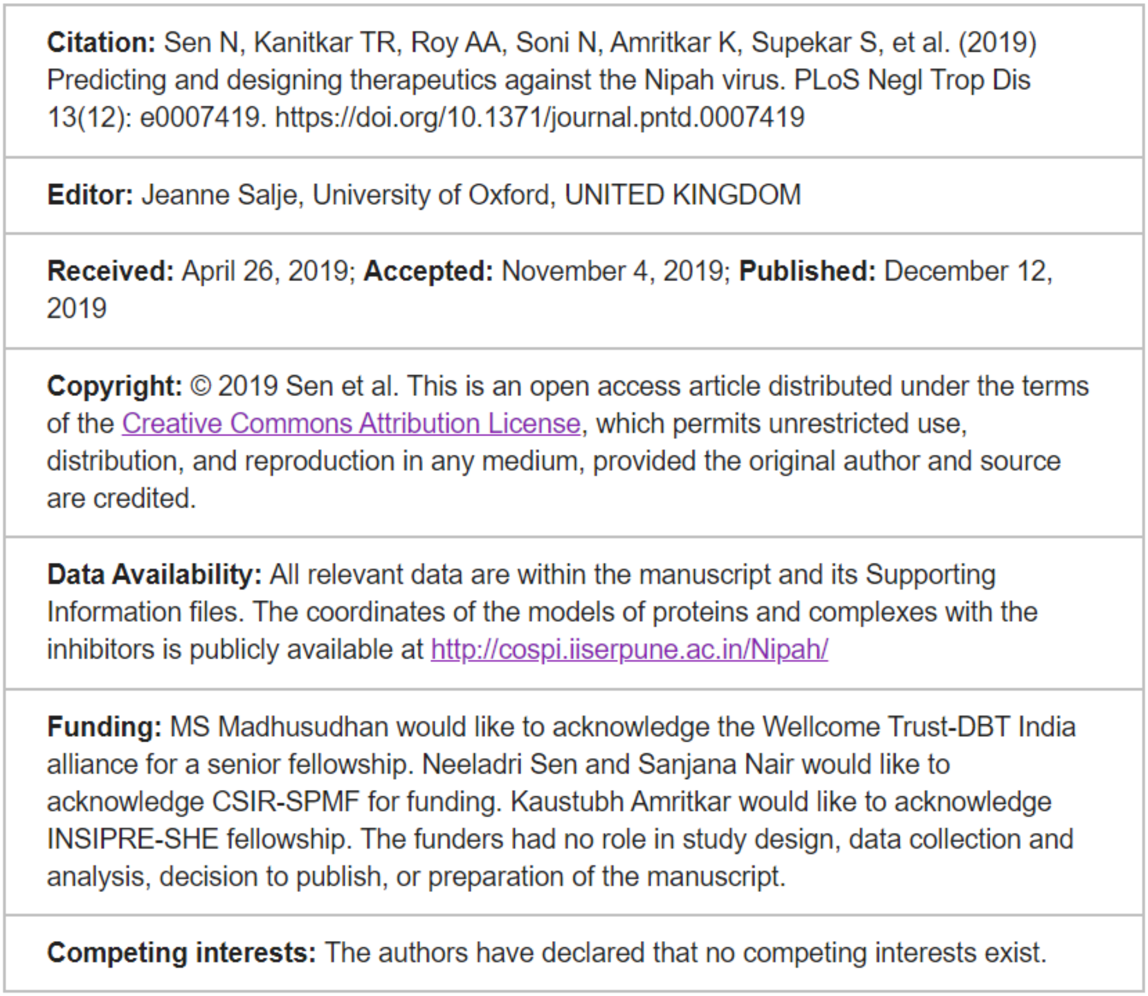
Publication: Progress in Biophysics and Molecular Biology

Publisher: Elsevier

Date: September 2017

© 2017 Elsevier Ltd. All rights reserved.

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights



Citation: Sen N, Kanitkar TR, Roy AA, Soni N, Amritkar K, Supekar S, et al. (2019) Predicting and designing therapeutics against the Nipah virus. PLoS Negl Trop Dis 13(12): e0007419. https://doi.org/10.1371/journal.pntd.0007419

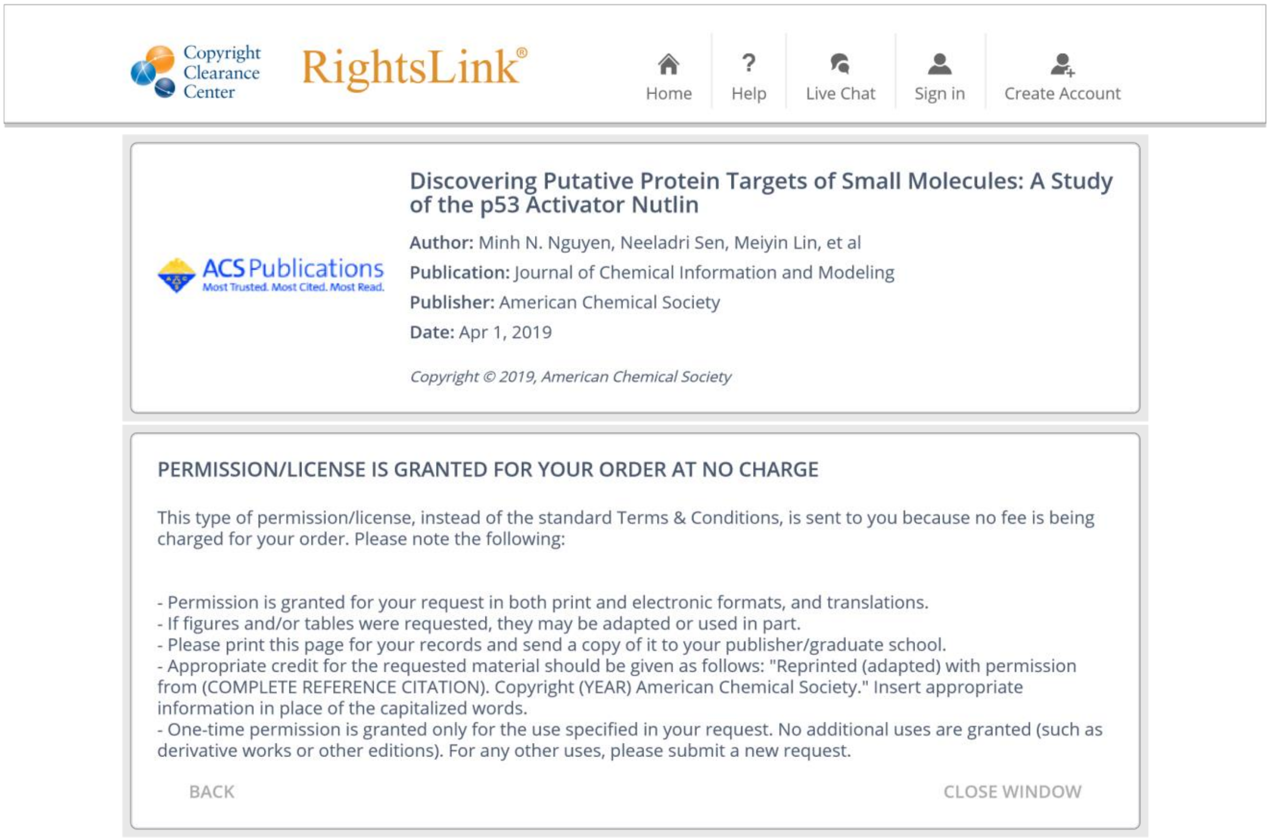
Editor: Jeanne Salje, University of Oxford, UNITED KINGDOM

Received: April 26, 2019; **Accepted:** November 4, 2019; **Published:** December 12, 2019

Data Availability: All relevant data are within the manuscript and its Supporting Information files. The coordinates of the models of proteins and complexes with the inhibitors is publicly available at http://cospi.iiserpune.ac.in/Nipah/

Funding: MS Madhusudhan would like to acknowledge the Wellcome Trust-DBT India alliance for a senior fellowship. Neeladri Sen and Sanjana Nair would like to acknowledge CSIR-SPMF for funding. Kaustubh Amritkar would like to acknowledge INSIPRE-SHE fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.



Discovering Putative Protein Targets of Small Molecules: A Study of the p53 Activator Nutlin

Author: Minh N. Nguyen, Neeladri Sen, Meiyin Lin, et al

Publication: Journal of Chemical Information and Modeling

Publisher: American Chemical Society

Date: Apr 1, 2019

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.