Statistical Potentials for Prediction of Protein-Protein Interactions

A thesis submitted towards partial fulfilment of
BS-MS Dual Degree Programme

by

Abhilesh Sunil Dhawanjewar

under the guidance of

Dr. M. S. Madhusudhan

Associate Professor, IISER Pune

Indian Institute of Science Education and Research Pune

Certificate

This is to certify that this dissertation entitled "Statistical Potentials for Prediction of Protein-Protein Interactions" submitted towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research Pune represents original research carried out by "Abhilesh Dhawanjewar" at "IISER Pune", under the supervision of "Dr. M.S. Madhusudhan, Associate Professor, Biology" during the academic year 2014-2015.

Supervisor
Dr. M.S. Madhusudhan

Date: 25/03/2015

Declaration

I hereby declare that the matter embodied in the report entitled "Statistical Potentials for Prediction of Protein-Protein Interactions" are the results of the investigations carried out by me at the Department of Biology, Indian Institute of Science Education and Research, Pune, under the supervision of Dr. M.S. Madhusudhan and the same has not been submitted elsewhere for any other degree.

Student

Abhilesh Dhawanjewar

Date: 25/03/2015

Acknowledgements

I would like to offer my sincere thanks to Dr M.S. Madhusudhan, Associate Professor at the Indian Institute of Science Education and Research, Pune for providing me with the opportunity to work on this project. His invaluable guidance at each and every step of the project helped me develop a scientific outlook and his high energy levels kept me enthused throughout the term of the project and thereafter. I thoroughly enjoyed the countless debugging and brainstorming sessions that we had together. Thanks for being a great mentor.

I would also like to express my gratitude to Dr. Richa Rikhy, Assistant Professor at IISER Pune for the helpful comments she provided during the course of the project. Also, thanks for agreeing to be on my Thesis Advisory Committee.

I thank all the young researchers in the Madhusudhan Lab, Neelesh Soni, Sagar Gore, Kuan Pern Tan, Yogendra Ramtirtha, Parichit Sharma, Shipra Gupta, Neeladri Sen for the valuable discussions and support and also for the numerous cricket and dart games that we enjoyed together.

It would be wrong if I do not thank my friends at IISER Pune, who have provided me with fun and joy and have been very supportive (for all the good and the not so good deeds). Thanks for the countless memories we made together.

Finally, I wholeheartedly thank my parents and the rest of the family, who have been my true support. I thank them for believing in me and letting my follow my heart and study the natural world.

I would also like to thank Kalyanee Shirlekar and Sharvaree Vadgama for the beers they promised, given I put their names here.

Abstract

Protein-Protein Interactions are critical to life, playing crucial roles in a variety of cellular processes. Hence, prediction of protein-protein interactions would help in gaining insights into cellular processes so that we may be able to manipulate and control it. In this study, we have developed knowledge-based pairwise statistical potentials based on experimentally derived structures for the prediction of protein-protein complexes. Structures of protein dimers in the Protein Data Bank (PDB) were used for the construction of the statistical potentials. A total of 96 different pairwise potentials were constructed for different values of five parameters: distance threshold for interactions, interacting atom types, weight type, weighting scheme and reference state. The performance of these potentials was benchmarked using Receiver Operating Characteristics (ROC) curves and Rank-Ordering. The side chain-side chain pairwise potentials were the best performers keeping all other parameters constant. The best performing pairwise potential could discriminate native structures from a sequence-randomized background in a benchmark set of 296 structures with a false positive rate of 1.4% and a true positive rate of 98.6%. This result is an improvement over the MODTIE potential which had a false positive rate of 28.5% and a true positive rate of 71.5%. The pairwise potentials are also complementary to each other, in the sense that they are efficient on different subsets of the benchmark set. Hence, a combination of the different potentials could result in better prediction accuracy. An attempt towards the development of a 5-body potential based on the pairwise potential was also initiated. Two different versions, an unweighted and a weighted potential were developed. The weighted multi-body potentials performed better than the unweighted potential. These multi-body potentials will be further refined, which is a work in progress. This prediction system will be bundled into a web server in the near future.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

> *"You see, proteins, as I probably needn't tell you, are immensely complicated groupings of amino acids and certain other specialized compounds, arranged in intricate three-dimensional patterns that are as unstable as sunbeams on a cloudy day. It is this instability that is life, since it is forever changing it's position in an effort to maintain it's identity in the manner of a long rod balanced on an acrobat's nose."*
>
> - Isaac Asimov, Pebble in the Sky

1.1 Protein Interfaces

Proteins are generally referred to as Biology's Workforce, as they perform nearly every function required for life. Proteins are polypeptide chains, consisting of amino acids linked in a linear chain. Amino acids, the building blocks of proteins consist of an amino group, a carboxyl group and an amino acid specific side chain. The properties of different amino acids determine the kinds of interatomic interactions between them. To carry out its function, a protein needs to be folded in a specific three dimensional shape. The 3D structure of a protein is largely dependent on its amino acid sequence, as particular sequences of amino acids give rise to linear chains and other compact domains with specific structures.

Most cellular processes require proteins to often work in concert, forming complexes of varying shapes and sizes, transporting other proteins, modifying other proteins etc. Unsurprisingly, Protein - Protein Interactions underlie a range of cellular processes such as mediating signal transduction, translating energy to physical motion, regulating cellular metabolism, immunological response and enzymatic inhibition, hence playing a critical role in many biological pathways (Braun and Gingras, 2012). Some parts of the protein would need to interact with other proteins and hence would form the interface. Proteins inside a cell are diffusing randomly and colliding with one another all the time, but only a small fractions of these collisions result in biologically meaningful complexes and some chemistry (active - such as enzymatic activity, or pas-

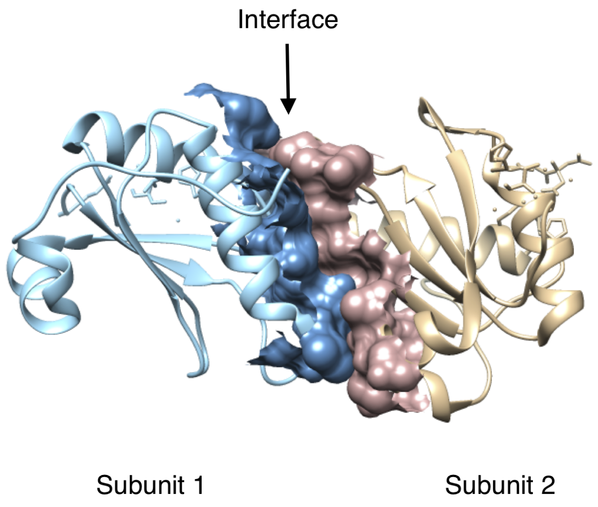


Figure 1.1: The interface between subunits of an RNA binding protein (PDB: 3S6E chains A & B) is shown in surface representation. The rest of the protein complex is in ribbon representation (Subunit A is coloured brown and Subunit B is coloured blue). This figure was rendered using Chimera (Pettersen et al., 2004).

sive - such as protein transport).Identifying the general rules behind protein-protein interactions is hence necessary for understanding the full repertoire of cellular pathways. The prediction of protein-protein interfaces can lead to advances in understanding disease pathways which involve aberrant protein-protein interactions such as cancer (Wong et al., 2003) and protein aggregate formations such as the Alzheimer's disease, Huntington's disease, Parkinson's disease, Creutzfeldt-Jakob disease and other Prion disorders (Kaytor and Warren, 1999) .

1.1.1 General Properties of Protein-Protein Interfaces

Protein-Protein interactions have been broadly categorized as homo- or hetero-oligomeric; obligate or non-obligate and transient or permanent (Nooren and Thornton, 2003). If identical proteins come together to form a complex, the resulting complex is termed as a homo-oligomer. Accordingly, assemblies of proteins with different subunits are termed as hetero-oligomeric. If the individual subunits of a complex can exist in solution independently, then the interaction between the subunits is a non-obligate one; in contrast, if the structure and function of the subunits is lost upon separation, it is an obligate interaction. Based on the lifetime of the interactions, protein associations are classified as transient (short-term interactions) or permanent (long-term interactions). These six different types of protein complexes differ in their amino-acid content and residue residue contact preferences (Ofran and Rost, 2003). Most protein complexes are a combination of these categories. The shape of a protein-protein interface (Figure 1.1) has been observed to be planar, globular and protruding, probably due to the symmetry involved in the associations (Argos, 1988, Jones and Thornton, 1996).

Earlier studies concerning protein folding ascribed hydrophobic effect as the major driving force behind protein folding (Dill, 1990). The folding of polypeptide chains buries the non-polar residues in the protein, minimising the number of thermodynamically unfavourable solute-solvent interactions. This burying of the hydrophobic residues resulting in the reduction of free energy also occurs during the aggregation of protein subunits and hence the hydrophobic effect is fundamental to the stabilisation of protein association as well (Chothia and Janin, 1975). However, contradictions between the measured values for enthalpy and entropy and the expected values for hydrophobic interactions have been noted for several protein association processes suggesting that it is not possible to account for the stability of protein associations on the basis of hydrophobic interactions alone (Ross and Subramanian, 1981). Analyses of multimeric protein structures in contemporary times have lead to the inclusion of electrostatic interactions (both long range coulombic interactions and short range hydrogen bonds and salt bridges) (Sheinerman et al., 2000, Xu et al., 1997), van der Waals forces, and hydrophobicity as major driving forces governing the association of proteins. Other forces such as aromatic stacking (Burley and Petsko, 1985), disulfide bonds, and cation-π interactions (Crowley and Golovin, 2005) also contribute to varying degrees.

1.1.2 The Protein Binding Phenomenon

The subunits in a protein complex are synthesised as separate proteins which then come together and bind in a particular orientation to give rise to the protein complex. The surfaces of the subunits in the monomeric state are completely hydrated. The hydrophilic amino acids residues on the protein surface make stabilising polar contacts and hydrogen bonds with the molecules of the solvent. Hence, for binding to take place between the subunits of a protein complex, the intermolecular interactions between the subunits must be more stabilising than the destabilisation caused by the desolvation of the subunit surfaces. The binding of a protein can be described as a two-step reaction:

$$A\ +\ B\ \rightleftharpoons A:B\ \rightleftharpoons AB \tag{1.1}$$

where A and B are the free proteins, $A:B$ is the intermediate complex (also known as the encounter complex) and AB is the bound protein complex (Selzer and Schreiber, 2001). The two subunits diffuse randomly in solution, their motions dictated by the dynamics of Brownian motion, until they reach an area, known as the *steering region*, the region where both the subunits are close enough to experience mutual electrostatic attraction. These aforementioned long-range electrostatic interactions cause the subunits to collide and form an encounter complex. At this stage, the short range electrostatic forces start acting at the interface of two proteins and contribute to the stabilisation of the encounter complex. Partial desolvation of the interface also contributes to a favourable entropy adding to the stability of the encounter complex (Ross and Subramanian, 1981). The electrostatic attractions between the two subunits hold the subunits associated to each other for a longer time, allowing them to achieve a proper orientation for binding (Sheinerman et al., 2000).

The interaction regions on proteins also contain binding motifs called *anchor residues*, that help stabilise protein complexes by reducing the kinetic costs associated with structural rearrangements at the protein binding sites (Rajamani et al., 2004). Molecular Dynamics simulations suggest that the side chains of these anchor residues frequently visit the conformations that are observed in the final bound state. They are also part of the complementary binding pockets often found on protein interfaces. Along with providing molecular recognition, these residues stabilise the encounter complexes that are in a near-native conformation. Further rearrangements in the side chains of amino acid residues, desolvation of the interface and the formation of non-covalent bonds lead to the final association in the stable complex. As a part of these events, certain *latch residues* present on the protein interface lock the subunits into the final stable conformation (Rajamani et al., 2004).

1.1.3 Experimental Determination of Protein Interfaces

Protein-protein interfaces can be experimentally determined using different methods. Some of the most commonly used methods are:

- X-ray crystallography: The three-dimensional coordinates of the atoms of a protein are estimated by analysing the diffracted angles and intensities of X-ray beams shone at a crystallised protein. Inherently, this method is unsuitable for determining the structures of proteins that are difficult to crystallize. This method also captures only a screenshot of the dynamic positions of the atoms of the protein. Despite these limitations, X-ray crystallography methods are the most popular to determine protein structure. Around 89 % of structures in the PDB are determined using X-ray Crystallography. However, only about 45 % of these structures depict protein-protein interactions.

- Nuclear Magnetic Resonance (NMR) spectroscopy: Determination of molecular structures using NMR spectroscopy measures the chemical shifts in the nuclei of the atoms in the protein, which are dependent on nearby atoms and their distances from each other, when the protein is placed in a strong magnetic field. This generates a list of constraints which can then be used to build a model of the protein describing the location of each atom. Since NMR spectroscopy is done on proteins in solutions, several models of the protein can be built, which can provide insight into the dynamics of the protein, unlike X-ray crystallography. A major limitation for this method is that it can only be used to determine the structure of smaller protein complexes. Currently, around 10 % of protein structures submitted in the PDB were solved using NMR spectroscopy. However, the number of interactions elucidated by NMR is much smaller.

- Electron Microscopy : Using a focused beam of accelerated electrons as the illumination source, electron microscopy is used to create images of large macromolecular structures. Proteins can be crystallized and then imaged by electron microscopy in a method similar to the one used in X-ray crystallographic methods of protein structure determination. Several images, providing different views may be taken for some symmetrical protein molecules. These images are then analysed and combined together to produce a three-dimensional map of the proteins atoms. This method is useful for producing low resolution maps of complex shapes but often cannot resolve the positions of individual amino acid residues.

- Chemical cross-linking followed by mass spectrometry: In this method, the protein complex is purified and tagged, its subunits are cross-linked by subjecting them to cross-linking reactions and then identified using mass spectrometry. This method is useful for producing low resolution structures of transient proteins. The cross-linking experiments are subject to several conditions and hence are error-prone processes.

Another set of experimental methods to detect protein interface residues exist such as the yeast two-hybrid method, which involves the construction of two plasmids and transforming them into a yeast strain. One of the plasmids encodes protein X with the DNA-binding domain of a transcription factor, while the other plasmid encodes the second protein Y in-frame with a transcription activation domain. Interactions between proteins X and Y reconstitutes an active transcription factor which binds upstream of the reporter genes and enables their expression (Causier and Davies, 2002). However, this method generates a lot of false positives due to non-specific interactions and often needs confirmations from other methods to reduce the false positive rates.

Mutagenesis experiments also aid in the detection of the protein interface residues. Amino acid residues in the protein subunits are systematically mutated and their effect on protein binding is studied with the use of protein expression assays. These experimental methods for the detection of protein-protein interfaces are labor-extensive and expensive, in addition to their general limitations. Hence, there is a need to develop fast and cost-effective computational methods that will enable us to generalize the principles of protein-protein associations and study protein interactions in greater detail.

1.1.4 Computational Methods for Studying Protein-Protein Interactions

Observations by Christian Anfinsen (Anfinsen, 1973) regarding the spontaneous refolding of an unfolded protein chain into its biologically active three-dimensional conformation led to the postulation of the Thermodynamic Hypothesis of Protein Folding. The Thermodynamic Hypothesis states that a native protein folds into a three-dimensional system in equilibrium, in which the state of the whole protein-solvent system corresponds to the global minimum of free energy (Xu et al., 2010). Based on this hypothesis, several computational studies concerning protein folding, protein-protein interactions and protein design depend on the derivation of a potential function to calculate the effective energy of a protein system. By matching the results of quantum mechanical calculations to the empirically determined thermodynamic properties of small molecules, parameters were derived for the development of potential functions (Sippl, 1993). These potential functions are then applied to macroscopic scales based on the assumption that properties of macroscopic states can be approximated by considering them as combinations of a large number of microscopic states. The potential functions developed through this inductive approach are termed as 'physics-based' or 'physical' potential functions. These physics-based potentials are based on atomic level models and hence are computationally very intensive.

Another set of potential functions are derived by extracting the parameters from a database of known structures (Sippl, 1993). These types of potentials follow the deductive approach and implicitly incorporate a variety of interactions. Therefore, these potentials do not represent true binding energies and hence are termed as 'knowledge-based' or 'pseudo-energy' potential functions. Though these methods do not reflect the true energies, they are algorithmically less intensive and have performed successfully. These potentials can be further divided into two cases. In one set, the knowledge-based potentials are derived by comparing the relative frequencies of interacting pairs in the database with that in a reference state (Miyazawa and Jernigan, 1996). In the other set, these potential functions are derived by optimisation with respect to certain criteria, e.g, by maximising the energy gap between the native conformations and the non-native conformations (Goldstein et al., 1992).

1.2 Knowledge-based Statistical Potentials

Knowledge-based statistical potentials are based on the *Boltzmann assumption*, that states frequently observed structural features correspond to low-energy states. Tanaka and Scheraga were the first to employ the above assumption to estimate pairwise amino acid interaction potentials by converting the observed frequencies of amino acid pairs into effective free energies (Tanaka and Scheraga, 1976). Since then many variants of pairwise amino acid potentials have extended this idea (Miyazawa and Jernigan, 1996, Sippl, 1993).

The general definition of a database-driven statistical potential as in (Sippl, 1990) is:

$$E(r) = -kT\,\text{ln}[f(r)] \tag{1.2}$$

where,

r = a protein structural parameter (eg. interatomic distance)

$E(r)$ = the energy at r

k = *Boltzmann's* constant

T = absolute temperature

Apart from r, the potential for a particular residue pair also depends upon the nature of atoms involved in the interaction and s, the separation of the respective amino acids in the amino acid sequence. At $s \geq 10$, the atoms can be considered as free particles and then by the Boltzmann approximation :

$$E^{obs}(r) = -kTln[f^{obs}(r)] \tag{1.3}$$

where, $f^{obs}(r)$ is approximated by the relative frequencies observed in the database.

Since these general potentials incorporate all interaction types between the atoms (electrostatic interactions, hydrogen bonds, van der Waals etc.) and also the influence of the surrounding medium on the interactions, they contain redundant information. In order to isolate the specific information in different potentials, we need to strip the redundant information from the general potentials. This redundant information can be defined in terms of a reference state. A suitable reference for intramolecular protein interactions is (Sippl, 1990):

$$E^s(r) = -kT\,\mathrm{ln}[f^s(r)] \tag{1.4}$$

where,

$$f^s(r) = \sum abf^{obs}(r) \tag{1.5}$$

which is averaged over all atom and residue types. Subtracting this redundant term from the general potentials, we get:

$$\Delta E^{obs}(r) = E^{obs}(r) - E^s(r) = -kT\left[\frac{f^{obs}(r)}{f^s(r)}\right] \tag{1.6}$$

The term $f^{obs}(r)$ comes from the database, whereas the term $f^s(r)$ is calculated as defined in the reference state. Hence, this potentials have a large dependence on the choice of reference state used.

1.3 Previous Related Work

Several researchers have attempted the prediction of protein-protein interactions using knowledge-based potentials in the past, and some of these methods have also been able to garner experimental evidence for their predictions.

Yasuda et. al., while working on the extracellular activation of tryptase ϵ used computational docking approaches to understand how tryptase ϵ selectively recognizes the activation sequence in pro-uPA. A lysine residue on loop A of tryptase ϵ (K20A) was predicted to be involved in recognizing the processing site of pro-uPA. Consistent with this prediction, they were able to show that K20A tryptase ϵ mutants failed to convert pro-uPA to uPA (Yasuda et al., 2005).

The PrePPI web server (`https://bhapp.c2b2.columbia.edu/PrePPI/`), set up by Honig lab at Columbia University, combines structural and non-structural cues in a bayesian framework to predict protein-protein interactions. The algorithm used in PrePPI generates structural representatives for two query protein sequences. Complexes formed by the structural neighbours of the representatives are then retrieved from the PDB to serve as interaction models. These interaction models are evaluated using five different scores, some of which are statistically derived. The researchers also tested nineteen PrePPI predictions of human interactions using Co-immunoprecipitation (Co-IP) experiments. Fifteen of these predictions were validated using the Co-IP experiments (Zhang et al., 2012).

Another example where knowledge-based bioinformatic predictions were experimentally validated was the predictions of new substrates for Aurora A kinase. The predictions were made by analysing the available data on Aurora A kinase and their phosphorylation sites and then using distinct types of biological information to generate a ranked list of potentials Aurora A kinase substrates. These predictions were validated by using $in\ vitro$ kinase assays and mass spectrometry analyses (Sardon et al., 2010).

1.4 Classifier Methods

Diagnostic decision making is an important process involved in the prediction of protein-protein complexes. In order to determine the threshold parameters for diagnosis, we need statistical methods to gauge which of the thresholds gives the most accurate predictions. One such method is the use of Receiver - Operating Characteristic (ROC) curves, which ensures that the number of true cases predicted does not come at the cost of an unreasonable number of false positives (Swets et al., 2000).

A classifier is a mapping that connects the instances to the predictions. Given a classifier and an instance, there are four possible outcomes. If the instance is positive and it is predicted as positive, it is termed *true positive*; if predicted negative, it is termed as *false negative*. If the instance is negative and it is predicted as positive, it is counted as a *false positive*; if predicted negative, it is a *true negative* (Fawcett, 2004). The two positive rate and the false positive rates of a classifier are defined below:

$$True\ positive\ rate \approx \frac{Positives\ correctly\ classified}{Total\ positives} \tag{1.7}$$

$$False\ positiverate \approx \frac{Negatives\ incorrectly\ classified}{Total\ negatives} \tag{1.8}$$

The True Positive Rate (TPR) is also referred to as *Sensitivity* and (1 - False Positive Rate) is also known as *Specificity*.

ROC curves are two-dimensional graphs in which FPR is plotted on the X-axis and TPR is plotted along the Y-axis. An ROC curve depicts the trade-off between the True Positives and the False Positives. Several points on the ROC curve are important. The point (0,0) never issues any false positives but it also does not return any true positives, whereas, the point (1,1) returns positives indiscriminately. The perfect classifier is represented by the point (0,1). At this point, all positives returned are True positives and none are False positives. Hence, the closer the ROC curve is to this point, the better the performance of the classifier. On the other hand, a random classifier lies on the $x = y$ line, as it is expected to return half the instances with positive predictions and the other half with negative predictions (Fig 1.2). To compare between different classifiers, the ROC curve performances are often reduced to a single scalar value. The Area Under the ROC Curve (AUC) is one such metric which is used to compare classifiers. Since, the AUC is a portion of the unit square, it's value always lies between 0 and 1. The random classifier is represented by a diagonal passing through the points (0,0) and (1,1), which corresponds to an AUC of 0.5, hence any real world classifier should not have an AUC value of less than 0.5.

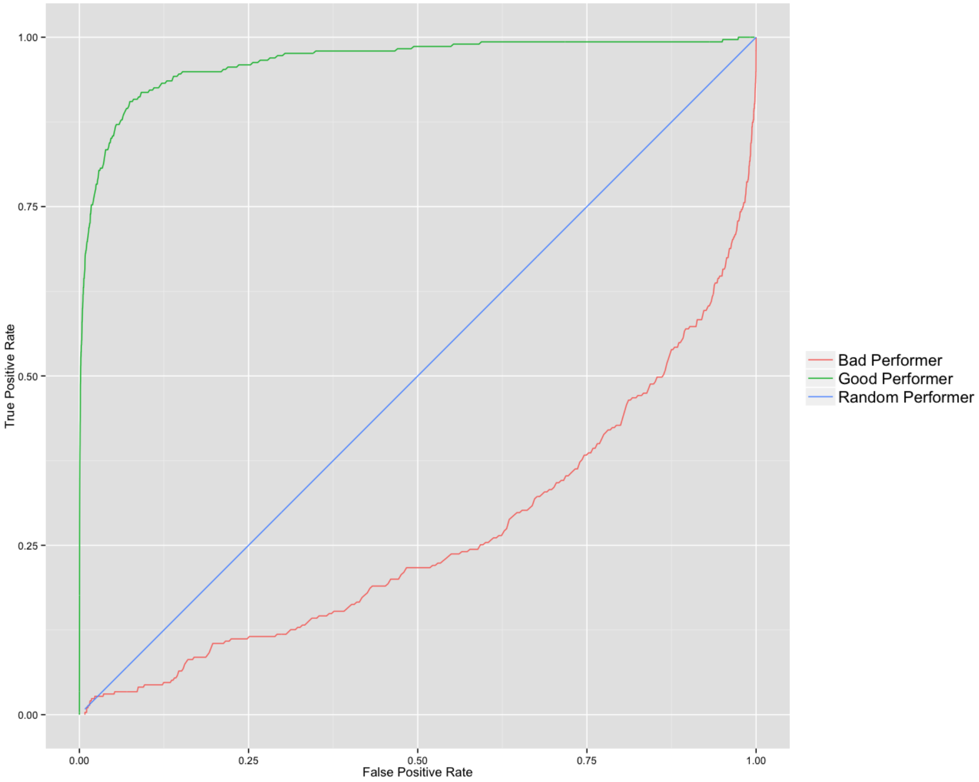


Figure 1.2: Sample ROC curves to illustrate the performance of different classifiers. The green line represents a good performing classifier $(AUC \approx 0.9)$. The blue line represents a random classifier $(AUC \approx 0.5)$ whereas the red line corresponds to a bad classifier $(AUC \approx 0.3)$

Chapter 2

Methods

2.1 Construction of the dataset

Three dimensional structures for protein dimers were retrieved from Protein Data Bank (Berman et al., 2000). The search for proteins with 2 chains (Asymmetrical Unit) returned 32871 structures. In order to remove redundancy, the above dataset was culled using the PISCES web server (Wang and Dunbrack, 2005) with the parameters given in Table 2.1. This resulted in a non-redundant dataset comprising of 6870 structures. Further filtering based on Buried Surface Area (BSA) was done. Buried Surface Area is defined as:

$$BSA = \sum_{n=1}^{N_{subunits}} ASA\,_{free}^{S_n} - ASA\,_{Complex} \tag{2.1}$$

where, $ASA\,_{free}^{S_n}$, is the solvent accessible surface area of the unbound subunits of the protein complex and $ASA\,_{Complex}$ is the solvent accessible surface area for the bound complex. BSA for the dimers was computed as per Eq. 2.1 using MODELLER (Sali and Blundell, 1993) and structures satisfying 400 Å^2 $\leq BSA \leq 2500$ Å^2 were taken to construct the final dataset. The lower bound on the BSA was put to remove false positives from crystal contact artifacts whereas the upper limit excluded structures with intertwined subunits. The final dataset comprising of 4060 protein dimers was divided into two sets: a training set of 3764 dimers, which were used for constructing the potential and a testing set of 296 dimers, used for benchmarking the statistical potentials. In order to make accurate predictions using statistical potentials, the number of samples in the training set should be large while keeping a reasonable number of samples in the testing set. Hence the division of the dimer set was made such that the testing set is $\sim$ 10 % of the training set. The PDB codes of the structures comprising the training and the testing set are listed in Appendix 1.

Sequence Percentage Identity	<= 40 %
Resolution	0.0 $\sim$ 3.0
R-Factor	0.3
Sequence Length	40 $\sim$ 10000
Non X-ray entries	Excluded
CA-only entries	Excluded
Cull PDB by	Entry
Cull chains within entries	No

Table 2.1: Parameters used for removing redundancy of the PDB dataset

2.2 Construction of Statistical Potentials

A series of Statistical Potentials were constructed using the protein dimers from the training dataset constructed above. Inter-atomic distances at different thresholds were computed for each structure using the 'cell list' implementation (borrowed from Neelesh Soni). Two amino acid residues were defined as interacting if any relevant atom of residue A of type i was within the distance threshold of any relevant atom of residue B of type j. Residue A and Residue B belong to different subunits of the protein complex. 96 different potentials were built using different values for five parameters : the contacting atom types (main chain-main chain, main chain-side chain, side chain-side chain or all), the weighing scheme for assigning weights to distinct residue interactions (cifa potential vs ipa potential), nature of the weights (derived at a single distance (norm) vs averaged over multiple distance (cmpd)), weights in the reference state (avg vs no_avg) and the distance threshold for contact participation (4, 6, or 8 Å). The combination of the different values for these five parameters gave rise to $4 \times 2 \times 2 \times 2 \times 3 \ = \ 96$ different potentials.

2.2.1 Two-Body Potentials

2.2.1.1 The cifa potential

$$\mathrm{S}_{i,j} = -\log \left[\left(\sum_{\forall \text{ interfaces}} \frac{ \dfrac{f^{\,int}_{\,ij}}{\sum\limits_{\forall ab} cifa^{\,int}_{\,ab}} \times \dfrac{cifa^{\,int}_{\,ij}}{\max(cifa^{\,int}_{\,ij})} }{ \dfrac{f_i}{N_m}\dfrac{f_j}{N_n} \times \langle cifa^{\,int}_{\,ij} \rangle } \right) \div N_{total} \right] \tag{2.2}$$

where,

f_{ij}^{int} = frequency of $i - j$ residue pairs across the interface

$cifa_{ij}^{int}$ = $\text{min}\left[\frac{\text{interacting atoms}_{\,i}}{\text{total atoms}_{\,i}}, \frac{\text{interacting atoms}_{\,j}}{\text{total atoms}_{\,j}}\right]$

$cifa\,_{ab}^{int}$ = frequency of any residue pair $a - b$ weighted by their respective $cifa$

$\frac{f_i}{N_m}$ = frequency of residues of type i in the subunit m

$\frac{f_j}{N_n}$ = frequency of residues of type j in the subunit n

N_m, N_n = Number of subunits in subunits m and n respectively

$\langle cifa\,_{ij}^{int} \rangle$ = average value of $cifa$ observed in the dataset for $i - j$ pairs across the interface

N_{total} = total number of protein complexes in the dataset

The observed probability for residue pairs of type i and j that belonged to different subunits and occurred within a distance threshold in a protein complex was weighted by $cifa$, the minimum of the fraction of the total atoms in each residue that were within the distance threshold. This weight was further normalised by max ($cifa_{ij}$), the maximum $cifa$ value for the i, j residue pair observed in the dataset. The probability of the occurrence of an amino acid pair of the type i, j was computed based on the occurrences of the residues i and j in their respective subunits. This probability weighted by the average value of $cifa$ observed in the dataset for the residue pair i, j forms the expected probability for a residue pair of the type i, j.

2.2.1.2 The ipa potential

$$\text{S}_{i,j} = -\,\text{log}\left[\left(\sum_{\forall\ \text{interfaces}} \frac{\dfrac{f\,_{ij}^{int}}{\sum\limits_{\forall ab} \alpha\,_{ab}} \times \alpha_{ij}}{\dfrac{f_i}{N_m}\dfrac{f_j}{N_n} \times \langle \alpha_{ij} \rangle}\right) \div N_{total}\right] \tag{2.3}$$

where,

$$\alpha_{ij} = \frac{ipa_{ij}^{int}}{\text{max}\ (ipa_{ij}^{int})}$$

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

In the second potential, the observed probability for residue pairs of type i and j that occurred within a distance threshold in a protein was weighted by ipa, the total number of interacting pairs of atoms between two residues. Similar to the first potential, this weight was further normalised by max (ipa_{ij}), the maximum ipa value observed for the i, j residue pair in the dataset. The reference state for this potential was similar to the reference state in the $cifa$ one, with the average value of ipa observed in the dataset for the residue pair i, j as the weight for the expected probability.

As Glycines lacks a side chain, they were handled in the following three ways in the side chain-side chain potentials. In the first scenario, assuming that all atom potential values should be representative of interactions concerning Glycine residues in the side chain-side chain case, the potential values for side chain-side chain interactions involving Glycine were borrowed from the corresponding all atom potentials. In the second scenario, following the assumption that side chain interactions are the major drivers for specificity in protein-protein interactions, the Glycine interactions were given a positive, hence unfavourable value of 1.38. For the third scenario, the potential value for all Glycine interactions was set to 0, the assumption underlying this scenario was that the occurrence of Glycine on protein-protein interfaces is random and hence the log odds of the observed probability against the expected probability of Glycine pairs is 1. The performance of the potential values for the three different scenarios were tested on the benchmark test by considering the number of native structures that were ranked the best against the randomised scores.

2.2.2 Multibody Potentials

Pairwise statistical potentials consider a protein-protein interface to be comprised of isolated residue pairs and hence devoid of any structural context. In a bid to include the structural neighbourhood of an amino acid residue while constructing the potential, 5-body statistical potentials were constructed, following the formulation of the two-body potentials. The interface of each protein complex was decomposed into 5-body amino acid cliques based on the interatomic distances between the residues. In graph theory, a clique is a special graph in which every vertex is connected to every other vertex in the graph. Two different distance thresholds, the intra-domain distance threshold and the inter-domain distance threshold, were used to define the connections in the clique, (i) the intra-domain threshold of 5 Å and (ii) the inter-domain threshold of 8.5 Å. We define two amino acids to be connected if any atom of residue A lies within a distance threshold R_0 of any atom of residue B (Fig 2.1). For these definitions, two cases were tested, Case (I) (the unweighted case), where the potentials were computed according to the formulation given in Eq. 2.4 and Case (II) (the weighted case), where the potentials in Eq. 2.4 were weighted using the average pairwise cifa values (borrowed from the two-body potentials) for the residue pairs constituting the cliques.

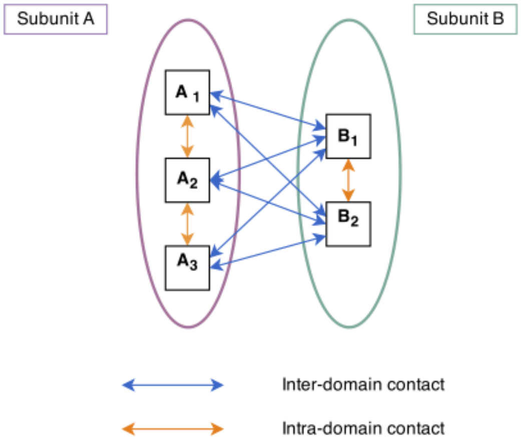


Figure 2.1: Schematic representation of a 5-body clique. Residues A_1, A_2 and A_3 belong to Subunit A of the protein complex whereas Residues B_1 and B_2 belong to Subunit B of the protein complex. The contacts between residues from the same subunit are termed as intra-domain contacts (shown by orange arrows) and the contacts between residues from different subunits are termed as inter-domain contacts (depicted by the blue arrows).

$$\mathsf{S}_{A_1A_2A_3B_1B_2} = -\log\left[\left(\sum_{\forall \text{ interfaces}} \frac{\dfrac{f_{A_1A_2A_3B_1B_2}}{\sum\limits_{\forall\, \alpha,\beta,\gamma,\delta,\epsilon}\left(f_{\alpha\beta\gamma\delta\epsilon}\right)}}{\dfrac{f^M_{A_1}}{\sum\limits_{x=1}^{20} f^M_x}\,\dfrac{f^M_{A_2}}{\sum\limits_{x=1}^{20} f^M_x}\,\dfrac{f^M_{A_3}}{\sum\limits_{x=1}^{20} f^M_x}\,\dfrac{f^N_{B_1}}{\sum\limits_{x=1}^{20} f^N_y}\,\dfrac{f^N_{B_2}}{\sum\limits_{x=1}^{20} f^N_y}}\right) \div N_{total}\right] \qquad (2.4)$$

where,

A_1, A_2, A_3 are residues that belong to subunit M

B_1, B_2 are residues that belong to subunit N

$f_{A_1A_2A_3B_1B_2}$ $=$ frequency of the clique $A_1A_2A_3B_1B_2$ across the interface

$f_{\alpha\beta\gamma\delta\epsilon}$ $=$ frequency of any $5-$ body clique $\alpha\beta\gamma\delta\epsilon$ across the interface

$\frac{f^M_{A_1}}{\sum f^M_x}$ $=$ frequency of the residues of type A_1 in the subunit $M;$ similarly for $\frac{f^M_{A_3}}{\sum f^M_x}$ and $\frac{f^M_{A_3}}{\sum f^M_x}$

$\frac{f^M_{B_1}}{\sum f^N_x}$ $=$ frequency of the residues of type B_1 in the subunit $N;$ similarly for $\frac{f^N_{B_2}}{\sum f^N_x}$

N_{total} $=$ total number of protein complexes in the dataset

The observed probability for a clique $A_1A_2A_3B_1B_2$ was obtained by dividing the number of occurrences of the clique $A_1A_2A_3B_1B_2$ by the number of all 5-body cliques observed in the protein. Considering, the choice of each amino acid in a 5-body clique as an independent event, the expectation term was obtained by multiplying the probabilities of picking amino acids A_i from their respective protein subunits.

2.3 Benchmarking of statistical potentials

The performance of the statistical potentials was tested on a benchmark set of 296 randomly selected dimers that were excluded during the construction of the potentials. The potential scores for the native structures were obtained by the addition of the potential scores for the individual residue pairs observed across the interface in the native structure. To distinguish the score of a native structure from the score of any non-interactant, these scores were compared against a randomised background set. There are two ways by which such a randomised background set could be obtained for each native structure: (i) physical models are built for the protein subunit by placing the two subunits of a protein complex in different relative orientations. The scores of these physical models then serve as the randomised background. (ii) keeping the structure of the protein complex unaltered, the sequence of the subunits is scrambled, which gives us randomised pairwise interactions across the interface for different scramblings. A number of such scramblings then constitute the randomised set.

The above methods of obtaining the background set are equivalent, we have chosen the latter method for the generation of the background set as it is less time consuming and algorithmically cleaner and easier to implement. For each of the 296 dimers in the benchmark set, 1000 decoy (non-interactants constituting the randomised background) confirmations were built by randomly scrambling the amino acid sequence of the dimers, followed by the computations of statistical potential scores for each of the decoy structures. The scrambling of the amino acid sequence was achieved by replacing each residue on the interface by another residue randomly chosen from the corresponding subunit. To access the significance of the raw statistical potential score, a Z-score was calculated based on the mean and standard deviation of the statistical potential scores for the decoy sets for each dimer (Eq. 2.5).

$$Z \;=\; \frac{x \;-\; \mu}{\sigma} \tag{2.5}$$

where,

x = raw score of the native structure

μ = mean of the raw scores of decoy structures

σ = standard deviation of raw scores of decoy structures

Receiver-operator Characteristics (ROC) curves are used to describe the observed false positive and true positive rates at different Z-score thresholds. ROC graphs are two dimensional graphs with Sensitivity or True Positive Rate (Eq. 2.10) plotted along the y-axis and (1 - Specificity) or the False Positive Rate (Eq. 2.11) plotted along the x-axis. For the construction of the ROC curves, the various definitions are given in Eq. 2.6 - 2.9. The Z-score thresholds for the ROC curves ranged from the minimum observed Z-score to the maximum observed Z-score for each potential along with an increment of 0.01. To compare the different potentials, the ROC curves were integrated to calculate the area under the curve (AUC). The AUC represents the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative instance, with 0.5 corresponding to a random prediction, and 1 to a perfect classifier (Fawcett, 2004). The optimal Z-score threshold for the best performing potential was taken as the Z-score where a tangent of slope 1 intersects the ROC curve.

$$\begin{aligned} True\ Positives\ (TP) \quad &= \quad No.\ of\ native\ structures\ with\ scores \\ &\quad\ \ lower\ than\ the\ threshold\ z-score \end{aligned} \tag{2.6}$$

$$\begin{aligned} False\ Positives\ (FP) \quad &= \quad No.\ of\ decoy\ structures\ with\ scores \\ &\quad\ \ lower\ than\ the\ threshold\ z-score \end{aligned} \tag{2.7}$$

$$\begin{aligned} True\ Negatives\ (TN) \quad &= \quad No.\ of\ decoy\ structures\ with\ scores \\ &\quad\ \ higher\ than\ the\ threshold\ z-score \end{aligned} \tag{2.8}$$

$$\begin{aligned} False\ Negatives\ (FN) \quad &= \quad No.\ of\ native\ structures\ with\ scores \\ &\quad\ \ higher\ than\ the\ threshold\ z-score \end{aligned} \tag{2.9}$$

$$True\ Positive\ Rate\ (TPR) \quad = \quad \frac{[TP]}{[TP\ +\ FN]} \tag{2.10}$$

$$False\ Positive\ Rate\ (FPR) \quad = \quad 1\ -\ \frac{[TN]}{[TN\ +\ FP]} \tag{2.11}$$

Chapter 3

Results

3.1 Dataset Generation

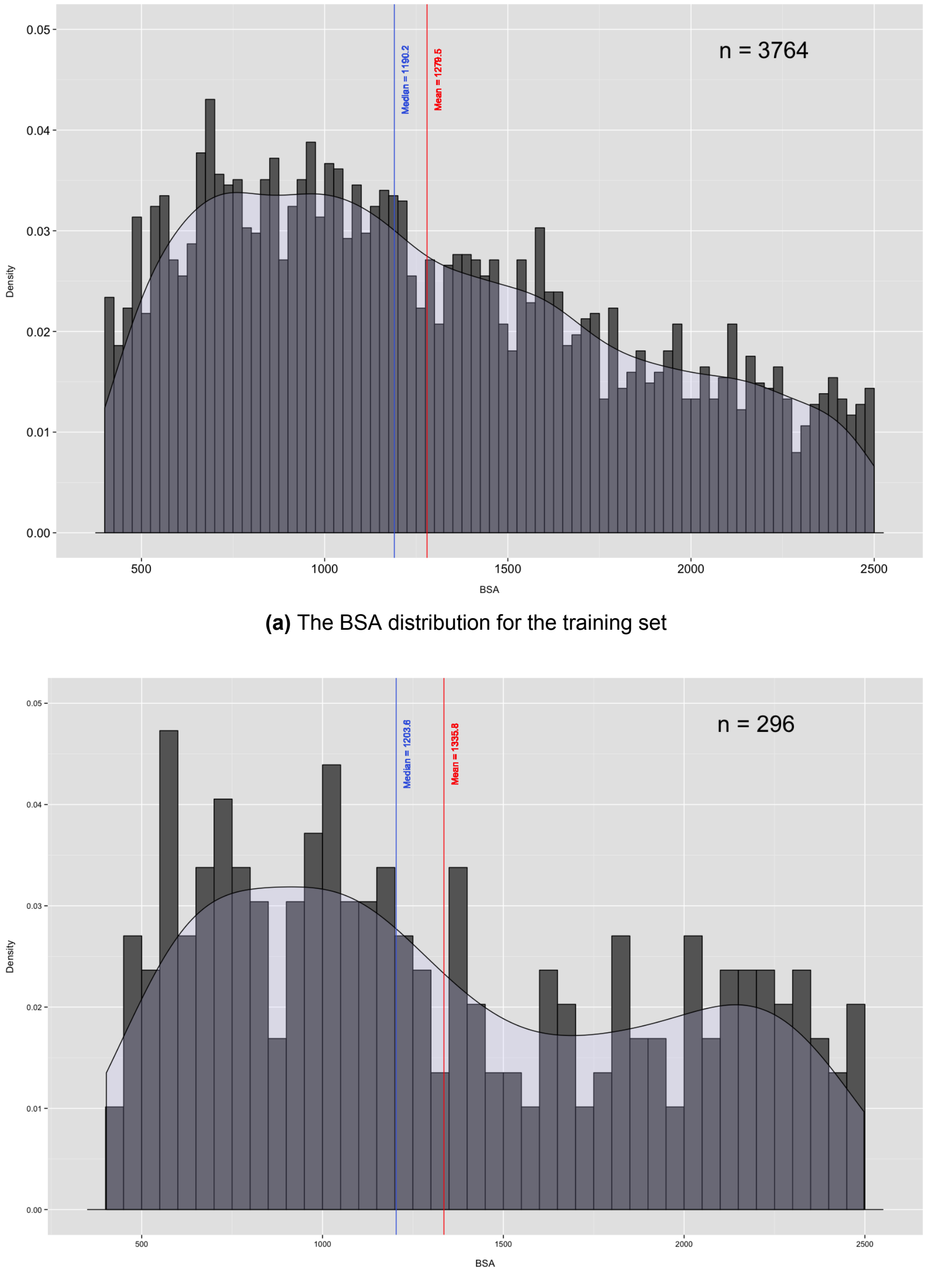
For accurate benchmarking of the potentials, the benchmark set used to test the performance of the potentials should be similar to the dataset the statistical potentials were trained on. Since the Buried Surface Area (BSA) of the complex was used as a filtration parameter, the frequency distributions of BSA were compared in the two sets to check for similarity (Figure 3.1). The BSA values for the 3764 structures in the training set and the 296 structures in the benchmark are similarly distributed. The Mean and Median values for the training set were 1279.5 and 1190.2 respectively, whereas the Mean and Median values for the benchmark set were 1279.5 and 1203.67 respectively.

Among the six ways of classifying protein interactions mentioned in the Introductions section (Sec 1.1.1), four categories (obligate, non-obligate, transient and permanent) pertain to the dynamics of protein complexes and it is not possible for us to retrieve this information from the crystal structures of proteins (though some of the studies may include information about the kind of interface, overall such studies are sparse). Concerning the oligomeric state of the protein complexes, we find that 90 % (3389 out of 3764) of the structures in the training set are homodimers. Similarly, 88 % (264 out of 300) of the structures in the testing set are homodimers.

3.2 Benchmarking of the Two-Body Potentials

The performance of the different statistical potentials was compared using two different methods.

1. Receiver-Operating Characteristic Curves for different Z-score thresholds
2. Rank - Ordering the scores of the native structures against scores from a randomized background set.



(a) The BSA distribution for the training set

(b) The BSA distribution for the testing set

Figure 3.1: Histograms depicting the distribution of the Buried Surface Area (BSA) for protein complexes in both the training and the testing set. The number of structures in the training and the testing sets are 3764 and 296 respectively. The binwidth used for plotting the distribution was 25. The red and blue lines on the plot represent the mean and the median of the distributions respectively. The black curve with grey shading is the kernel density estimation for the distribution.

3.2.1 Benchmarking using ROC curves

The testing of the 96 different statistical potentials was done on a benchmark set of 296 dimers and their performance was compared using ROC curves. The potentials showed a diverse range of performances as can be seen in Figure 3.2. 11 of the 96 potentials had an Area Under the Curve of their ROC curves greater than 0.90 (shown in the inset of Fig 3.2). All eleven of these potentials were variants of the side chain-side chain potentials that were weighted by $cifa$. The main chain-main chain potentials were the worst performers of all with some of the potentials performing worse than a random classifier.

The highest power of discrimination between the native and non-native interfaces was achieved by the statistical potential built from side chain-side chain interactions across the interface at the threshold distance of 4 Å (4.ss.norm.cifa.avg in Figure 3.2). The weighting parameter for this potential was $cifa$, calculated at a single distance of 4 Å and the reference state was weighted by the average weight for residue pairs in the dataset. The area under the curve (AUC) for the ROC curve for this potential was 0.9622. The true positive rates and the false positive rates at the optimal Z-score of -0.7 were 97.8% and % respectively.

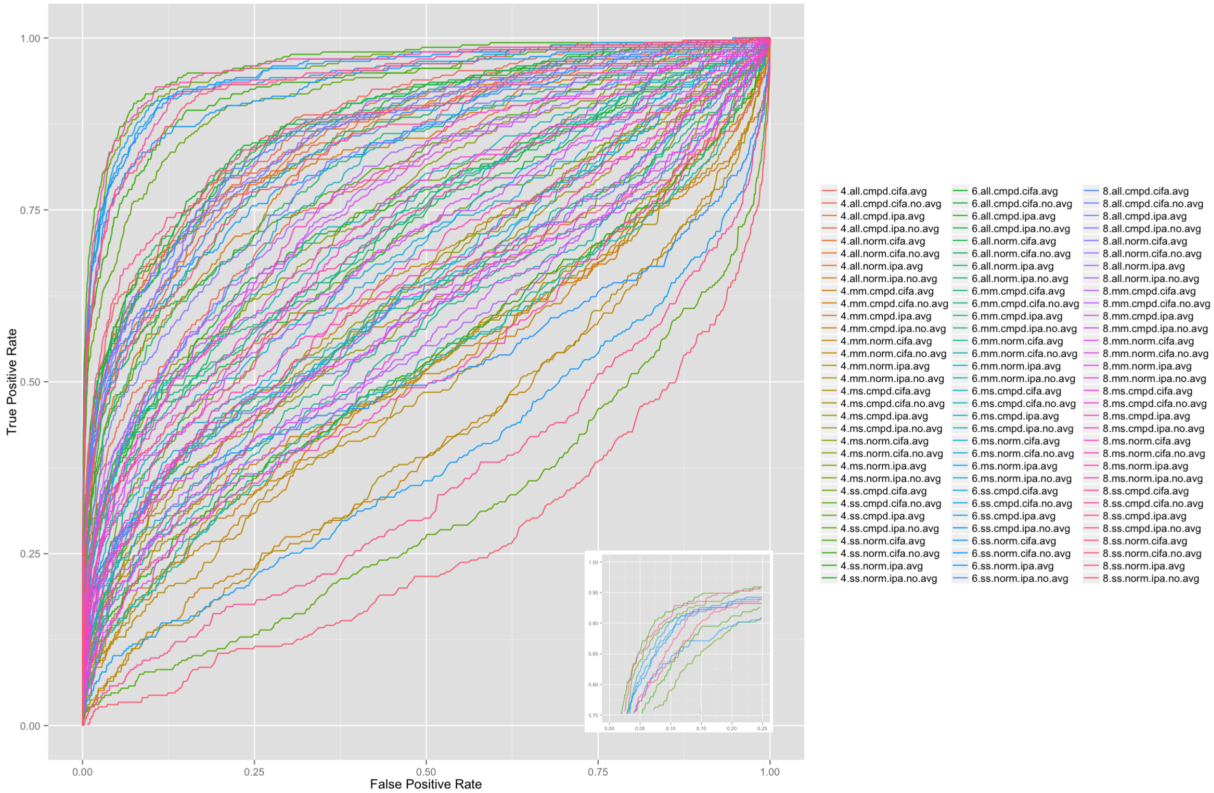


Figure 3.2: A comparison of the performances of the 96 different potentials as represented by their Receiver-Operator curves. **Inset**: Zoomed in version for the best performing potentials.

3.2.2 Benchmarking using Rank Ordering

After rank ordering the scores of the native and the decoy confirmations, the number of cases where the native confirmation had the best score was also used to compare the performance of the different potentials. On the basis of the interacting atoms, the performance of the potentials was in the order: side chain-side chain > all atoms > main chain-side chain > main chain-main chain (Fig 3.3). The performance of the potentials based on the nature of the weights (i.e. whether they were computed at a single distance (norm) or computed as an average from three different distances (avg)) was comparable across the different potentials. Only in the case of the potentials constructed at the distance threshold of 4 Å side chain-side chain potential, is the norm potential better than cmpd potential.

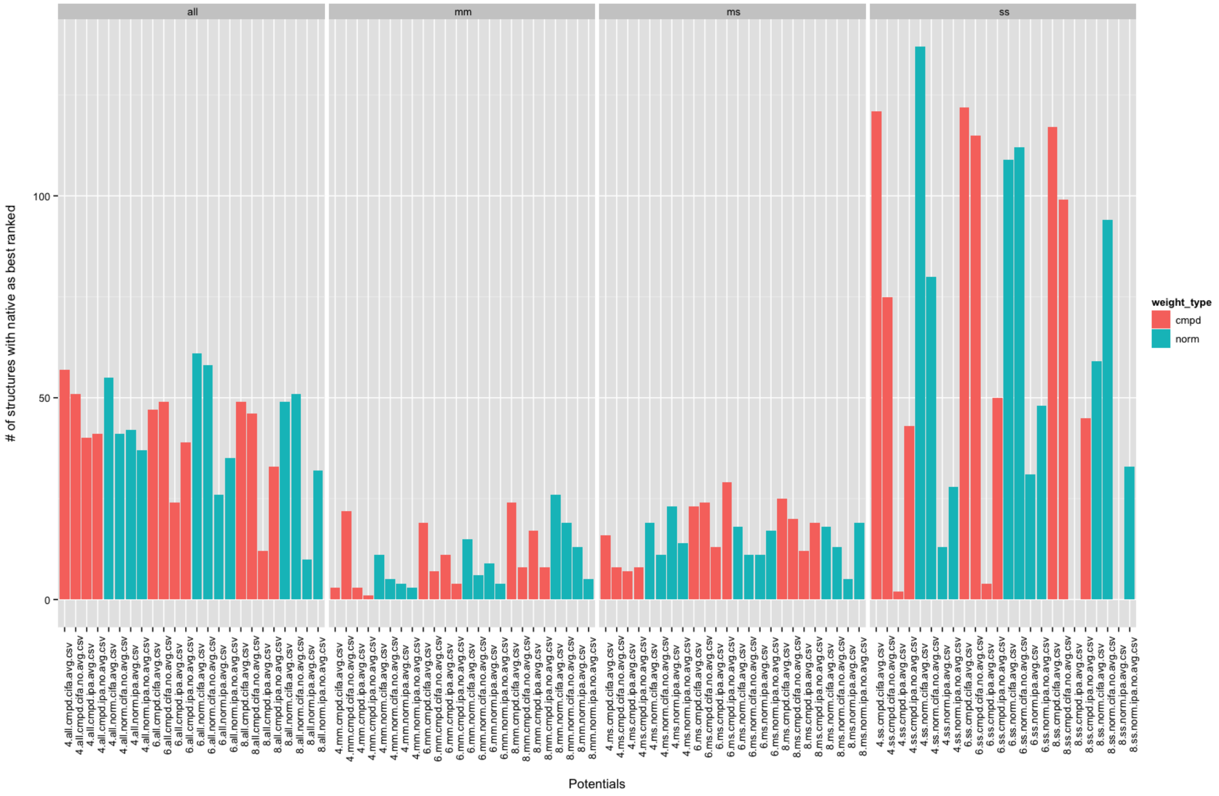


Figure 3.3: A comparison of the performances of the different potentials as measured by the number of native structures that were ranked 1 against a set of decoy structures. The different facets in the figure describe the performances of the potentials according to the interacting atoms type (all - all atom, mm - main chain-main chain, ms - main chain-side chain and ss - side chain-side chain). The colors differentiate between the potentials based on the nature of the weights, red for cmpd (weight computed as an average at three distances) and cyan for norm (weight computed at a single distance)

The performances of the potentials followed a similar pattern when dissected according to the distance threshold (Fig 3.4). The $cifa$ potential performed better than the ipa potential in all cases. Here again, the side chain-side chain potential at 4 Å with $cifa$ as the weighting parameter computed at a single distance (4.ss.norm.cifa.avg) was the best performer. 137 out of 296 native structures were best ranked when compared against their randomised backgrounds and 240 structures out of 296 have their native structures ranked under 25.

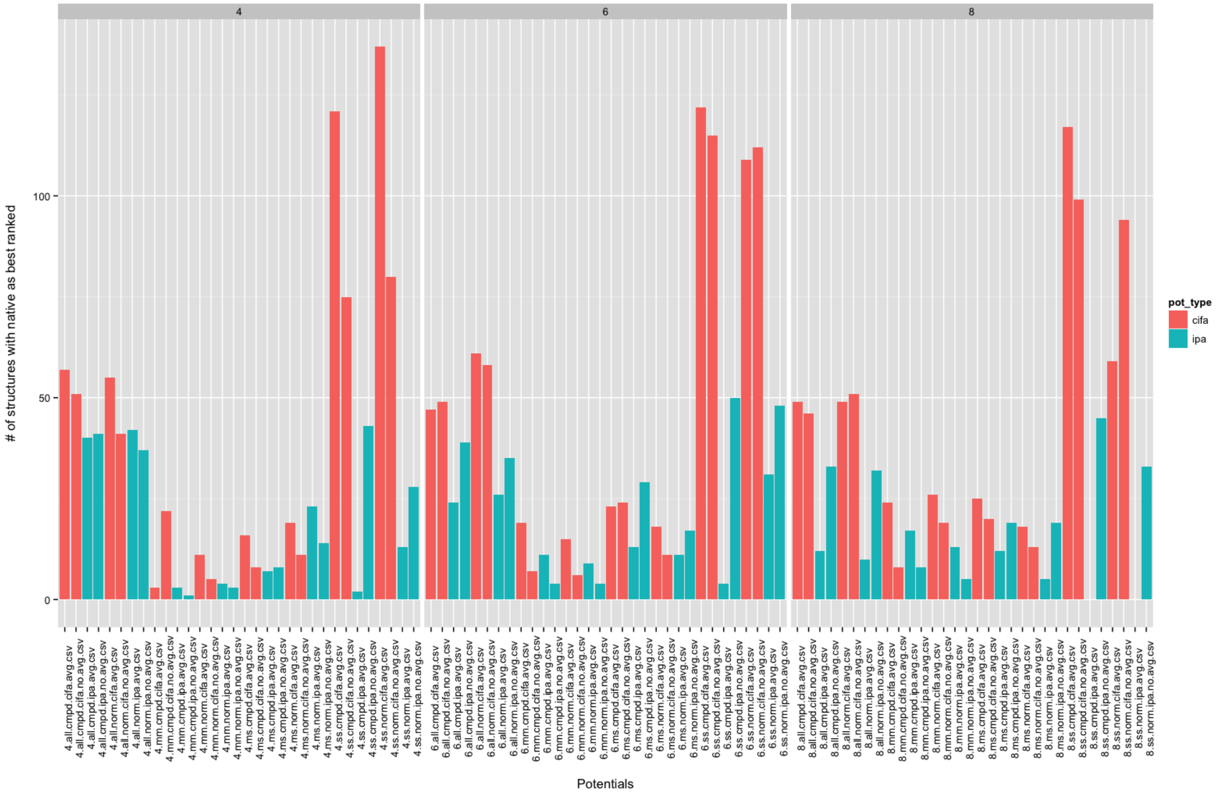


Figure 3.4: A comparison of the performances of the different potentials as measured by the number of native structures that were ranked 1 against a set of decoy structures. The graph is dissected based on threshold interaction distance and the different potential type; cyan represents the ipa potential whereas red stands for the $cifa$ potential.

The potentials with an Area Under the ROC Curve (AUC) greater than 0.90 were checked for complementarity in terms of protein complex prediction (whether different protein complexes were ranked best by different potentials). The results are summarised in Table 3.1. The union of the best ranked sets (set of structures whose native structures were ranked 1 against a randomised background) for the different potentials was greater than the best ranked set for the best performing statistical potential (4.ss.norm.cifa.avg).

cut_off rank	best performing potential (4.ss.norm.cifa.avg)	Union of AUC $\geq$ 0.9
$\leq$ 0	137	169
$\leq$ 5	190	217
$\leq$ 10	216	235
$\leq$ 15	224	247
$\leq$ 20	233	252
$\leq$ 25	239	261

Table 3.1: Complementarity between the different potentials at different rank cut-offs. The number of structures that were ranked 1 against their respective backgrounds for the two cases are given in the two columns.

There is a preference for same-interaction pairs and complementarity between opposite charges (eg Lysine pairing up favourably with Glutamate and Aspartate) is also observed in the log-odds ratio matrix (Fig 3.5). Cysteine-Cysteine and Histidine-Histidine are among the best scored residue residue contact pairs. The contact preference scores for the hydrophobic amino acids are overall favourable though any specific preferences are not observed.

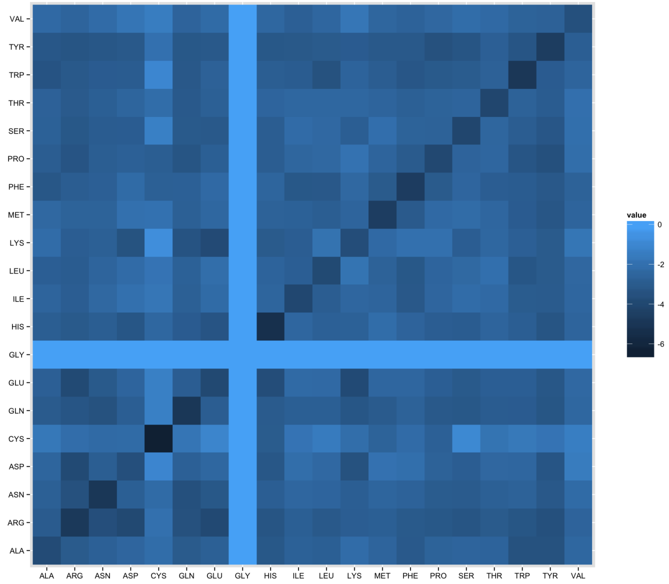


Figure 3.5: Log odds ratio of residue pair preferences across protein - protein interfaces for the best two-body potential. The darker the shade, the higher the preference

3.2.3 Testing Potential Values for GLY-GLY pairs

Since, Glycine lacks a side chain, the potential values for GLY-GLY pairs for the side chain-side chain potentials were tested as described in Methods. For the side-chain potentials at 4 and 8 Å , assuming that the occurrence of Glycines on the interface is random (column no-effect in Table 3.2), gave the best performance. For the potential at 6 Å, however, the assumption that GLY residues are unfavourable worked the best.

potentials	all atom potentials	unfavourable score (1.38)	no-effect (0.00)
4.ss.norm.cifa.avg	109	132	136
6.ss.norm.cifa.avg	100	131	113
8.ss.norm.cifa.avg	91	99	110

Table 3.2: Testing of side chain-side chain potential values for GLY-GLY pairs: Native confirmation scores were rank ordered against decoy conformation scores and the number of structures with native confirmations as best ranked was noted

3.2.4 Performance on the testing set

With a Z-score threshold of -0.7 for the best pairwise potential (4.ss.norm.cifa.avg), 284 out of the 295 native structures testing set had a z-score below the threshold, which corresponds to a true prediction. Among the 11 structures which had a z-score greater than the threshold, 7 structures were incorrectly submitted as dimers in the PDB. The biological assemblies for these structures (PDB codes: 3PNA, 1IFQ, 3MTX, 1PL3, 3QL9, 4CMP, 2XRW) is a monomeric entity, as given in the Protein Data Bank. These false classifications in the PDB may be a result of crystallization artefacts. Since, our potentials could successfully distinguish crystal artefacts from true interactions, these 7 structures were considered as correct predictions. Hence, our potentials could correctly identify 291 out of 295 structures, which translates to a prediction accuracy of 98.6 %.

3.2.5 Comparison with MODTIE

The performance of the best performer was compared with MODTIE (Davis et al., 2006) (Fig 3.6). Benchmarking for both the potentials was done on the same benchmark set. The Area under the curve for the MODTIE potential was 0.9445. The true positive rate and the false positive rate at the Z-score threshold of -1.7 was 71.5 % (211/295) and 28.5 % (84/295) respectively.

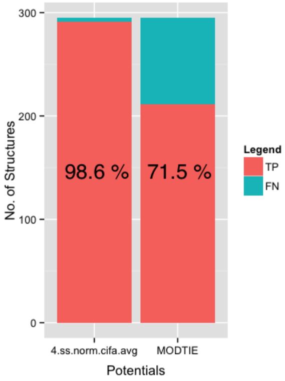


Figure 3.6: Performance of the two body potential in comparison to MODTIE. The red part and the cyan part of the plot depict the no. of True Positives and the no. of False Negatives respectively.

3.2.6 Multibody Potentials

The performance of the multi body potentials was accessed in a manner similar to the two-body potentials. For case (i), with the intra-domain distance threshold as 5 Å and the intra-domain distance threshold as 8.5 Å, a total of 280114 distinct cliques were observed, out of 323400 distinct possibilities. The Receiver Operating Curve is shown in Fig 3.7. The Area Under the Curve for the ROC of the potential without any weights was 0.3089, whereas the Area Under the Curve for the potential with the weights was 0.41006. Both the potentials performed worse than a random classifier.

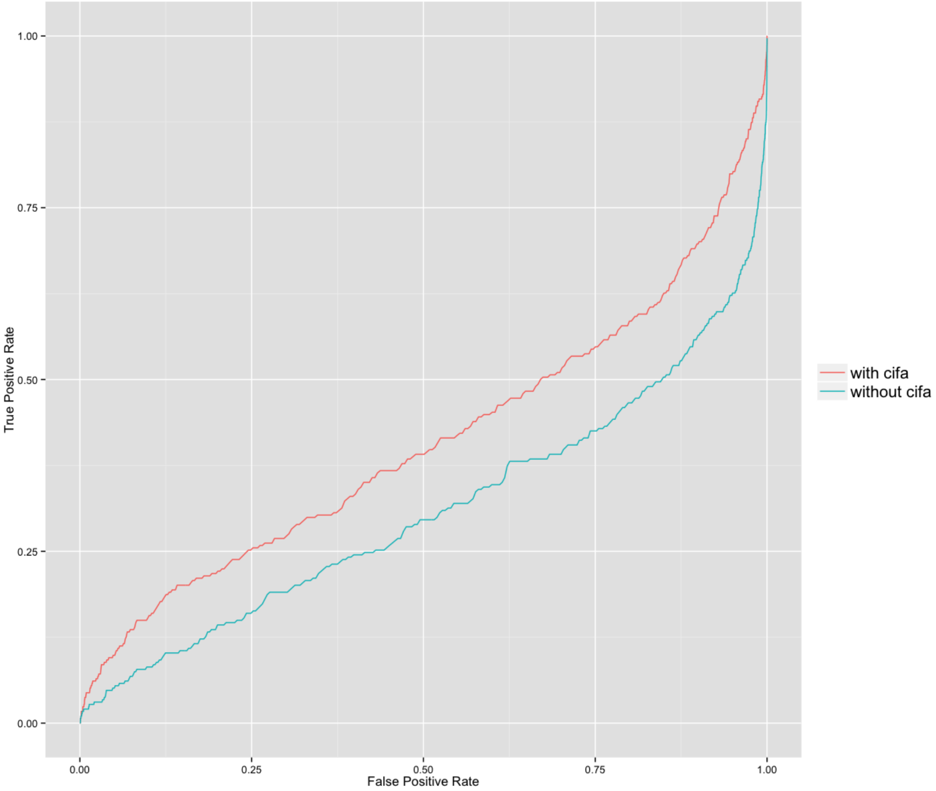


Figure 3.7: Performance of the multibody potentials as assessed by the ROC curves. The cyan curve represents the ROC for the unweighted 5-body potential whereas the red curve depicts the ROC for the weighted 5-body potential.

3.3 Validation

The potential 4.ss.norm.cifa.avg was tested for its prediction power on the Ral-GEF system. Six variants of GEF were tested for binding with Ral. Experimental evidence shows that four of these variants (RGL1, RGL2, RGL3 and RALGDS) bind Ral in a particular mode, while the other two variants weakly interact with Ral, binding in a different mode. The Z-scores for all the GEFs were below the threshold and hence all six variants are predicted to bind.

Based on the statistical potential scores, we predicted the following hotspot residues, SER:173:A, GLU:34:A, LYS:370:B, ARG:322:B, ARG:42:A and ARG:74:A, in RGL1-Ral complex that upon mutation would weaken the interaction between RGL1 and Ral. These hotspot residues lie in complementary clusters (Fig 3.9) and hence mutating them to other residues would lead to unfavourable interactions, thereby weakening the interaction between RGL1 and Ral.

GEFs	Z-scores
RALGDS	-3.59
RALGPS1	-3.39
RALGPS2	-3.72
RGL1	-3.21
RGL2	-3.11
RGL3	-2.72

Table 3.3: The predictions regarding the binding of Ral to GEF variants using the pairwise statistical potentials.

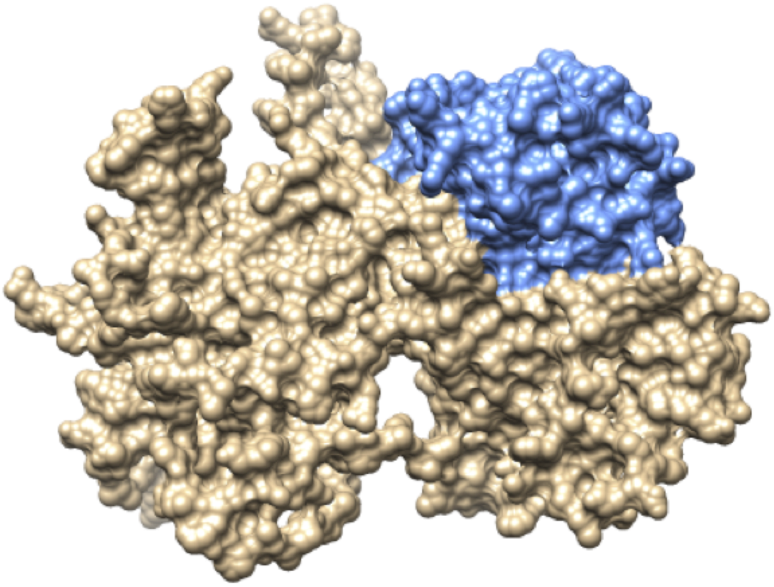
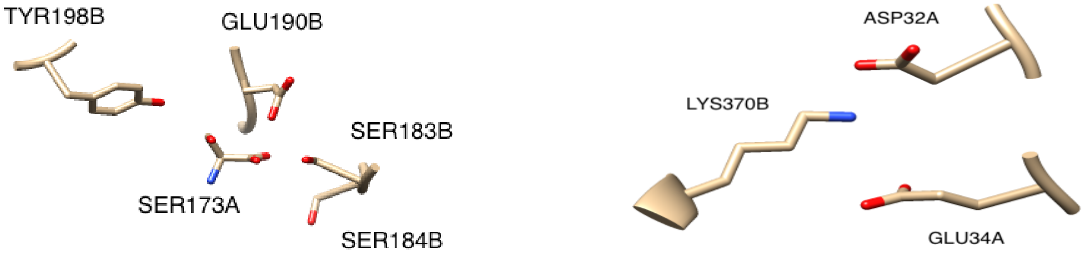


Figure 3.8: The RalGDS-Ral complex. The Ral subunit is shown in blue in surface representation whereas the RalGDS is depicted in brown, also in surface representation. Image rendered using Chimera (Pettersen et al., 2004)



(a) Interaction cluster of SER173 of Ral

(b) Interaction cluster of LYS370 of RGL1

Figure 3.9: Hotspot residues in the RGL1-Ral Complex. Subunit A is Ral and Subunit B is RGL1. Image rendered using Chimera (Pettersen et al., 2004)

Chapter 4

Discussions

4.1 Statistical Potentials

Due to the labor-intensive and expensive nature of the experimental methods for validation of protein-protein associations, computational methods for the prediction of protein-protein interactions have become quite popular. Between the two different types of computational methods for the prediction of protein associations, namely the physics-based methods and the knowledge-based statistical methods, we have used the latter to develop a way to determine protein binding. We chose statistical potentials because they are algorithmically and computationally much more feasible than the physics-based model. Another limitation of the physics-based models is their heavy dependence on the accuracy of the structure of the protein. A small discrepancy in the atomic coordinates will lead to a significant deviation in the estimation of the energies, when computed using the physics-based models. Statistical potentials are robust enough that such minor discrepancies do not affect the estimates of potentials by a significant amount.

Statistical potentials help us portray a picture of how interactions between proteins are mediated and can be used as stand-ins for binding free energies. They work on the principle that the most frequently observed amino acid residue pairs are energetically more preferred than the pairs less frequently observed. However, because statistical potentials do not discriminate between interaction types and their strengths (for eg, the strength of a hydrogen bond vs that of a van der Waal's interaction), the statistical potential scores do not correlate perfectly with the binding affinities. To build a statistical potential for predicting binding affinities, known structures will have to be subsetted according to their binding affinities and then statistical potentials built for each subset of the dataset. However, the dearth of data on experimental binding affinities prevents the construction of a meaningful statistical potential. Based on observations made on an experimental dataset, statistical potentials allow us to derive approximate functions which can be used to predict the energy of an unknown system.

4.2 Pairwise Potentials

We tested 96 different pairwise statistical potentials for their ability to predict protein-protein interactions. In both the methods for benchmarking the performance of the potentials, the side chain-side chain potentials were significantly much better than the other potentials. The performance of the other potentials based on the type of interacting atoms follows the order: all atom > main chain-side chain > main chain-main chain. Since, protein associations require specific interactions between the atoms of the constituting amino acid residues, these specificities are provided by the properties of the different side chains. Consistent with this, the side chain-side chain potential has the best power for discriminating between native and non-native associations, whereas main chain-main chain (which lack any specificity) potentials are the worst performers.

We constructed two statistical potentials using two different weighting schemes $cifa$ and ipa. Between the two different potential types $cifa$ and ipa, we observed that the $cifa$ performed better than the ipa potential when all other parameters are kept constant. The difference between the two different weights is that while $cifa$ aims to capture the contribution of each residue to the interaction between two residues and then considers the contribution of just one residue towards the weighting, the ipa potential weighs the different residue pairs based on the number of interatomic interaction pairs between two different residues.

Since Glycines lack a side chain and are present abundantly on the interface, we need to incorporate them in the side chain-side chain potentials. Three different scenarios were tested in this regard. The potential values for the GLY interactions in the side chain-side chain potentials were derived using a semi-optimisation approach. Of the three different scenarios tried, the case which assumes the distribution of Glycines on the interface is random (the ratio of the observed frequency of GLY pairs and the expected frequency of the GLY pairs is 1) gave the best results. Since Glycine lacks a side chain, in our potential, we have considered that it does not discriminate between amino acid residues and interacts with any residue the same way.

Cysteine-Cysteine pairs have the best scores for any residue pair. This observation previously reported by Glaser (Glaser et al., 2001), is expected since the sulphurs in Cysteine have been observed to form disulphide bonds which may play an important role in the stability of protein complexes. Cysteine-Cysteine pairs along with Histidine-Histidine pairs are also found in metal coordination sites across the interface (eg. zinc finger domain). These may be the reasons why Cysteine-Cysteine and Histidine-Histidine residue pairs have high scores. Other residue pairs with favourable contact scores are the oppositely charged residues (for eg. Lysine and Arginine (with positively charged side chains) with Glutamate and Aspartate (with negatively charged side chains)). These residue pairs form salt bridges across the interface and help strengthen the interaction. Also, since the burial of charged amino acid residues is energetically unfavourable they are often observed to be paired with oppositely charged amino acids.

The non-specific van der Waal's force is the major interaction force between the hydrophobic amino acids

(Leucine, Isoleucine, Alanine, Valine, Proline, Methionine, Phenylalanine and Tryptophan). Given the non-specific nature of this interaction, the hydrophobic residues clump together showing no particular residue pair preferences. As seen in the contact potential matrix, any hydrophobic - hydrophobic residue pair gets a favourable score without showing any particular preferences, except in the case of Tryptophan-Tryptophan pairs which get a higher score than the other hydrophobic pairs.

In the log odds ratio matrix for the pairwise potential, the self-interaction scores between residues are high scoring. This means that like charged residue pairs (eg. Arginine-Arginine pairs) which are expected to get unfavourable scores are assigned favourable scores. A significant proportion of the dimer structures solved are homodimers and our dataset is also comprised of mostly homodimers. Because of the symmetric nature of the homodimers, it is likely that similar residues come closer more often and hence, they have high favourable scores in our score matrices. However, such like charge interactions have been the focus of other studies (Magalhaes et al., 1994, Pednekar et al., 2009) which find that such like charged pairs do occur in protein-protein interactions if the interaction between them is mediated through a water molecule (Heyda et al., 2010). Magalhaes et. al. (Magalhaes et al., 1994) provides examples several where Arginine-Arginine pairs are found in close proximity. Since water molecules cannot be reliably captured in low resolution X-ray crystal structures and also since information about the presence of water in the protein structures in our training set is missing, we cannot explore this possibility. An alternative hypothesis behind this observation might be that at the 4 Å level, there might be significant main chain-main chain interactions which might contribute to the favourable scores for the diagonal elements. Further investigation is needed to pin down the reason behind this observation.

4.2.1 Testing the performance of pairwise potentials

Benchmarking by rank ordering is one of the most robust ways to test the performance of a potential as it imposes the stringent constraint that the native conformation must have the lowest score when compared with 1000 non-native confirmation scores. The results from this benchmark echo the ones observed using the ROC analysis. When this test was applied to compare the performance of a union of best performing potentials versus the performance of any one of these potentials, it was observed that the union of potentials performed better than the best performing potential. This seems to suggest that different potentials are more efficient at discriminating certain types of protein complexes than the other potentials. As an example, a protein from *Enterococcus faecalis* (PDB Code: 3NAT) was ranked 462 out of 1000 when a side chain-side chain potential was used. However, when a main chain-main chain potential was used on the same protein, it was ranked 1. This suggests that, in this protein, main chain-main chain interactions are more important at the interface than side chain-side chain interactions and hence, a main chain-main chain potential gave us better predictions.

Since, pairwise potentials interpret protein-protein interfaces in terms of isolated residue pair interactions, these potentials ignore the structural context of an amino acid residue in a protein. Often, the surrounding amino acid residues of a particular residue may be important for bringing that residue in a particular confirmation to facilitate the interaction with the other subunit. This absence of contextual awareness might explain why these pairwise potentials do not predict protein complex formation perfectly.

4.3 Multibody Potentials

Statistical Potentials built using extended stretches of amino acid residues would solve the problem mentioned in the previous section. By taking into account the structural neighbourhood of an amino acid residue during the construction of the potential, we look for clusters of residues. Following the same assumption as in the pairwise potential, that the most frequently observed clusters of amino acid residues correspond to the energetically favourable states, attempts were made at constructing 5-body statistical potentials.

The structural definition of a multi-body clique across a protein interface is more complicated than the simplistic definition used for defining interactions in the pairwise potential case. Two different distance thresholds are now required for the definition, an intra-domain interaction distance and an inter-domain interaction distance. The most optimal values for these parameters are not easy to determine, a smaller, stringent distance threshold will take into account the strongest interactions but we may not sample enough distinct cliques, which would affect the performance of the potential. However, setting a liberal distance threshold, we may be able to get a larger number of multi-body cliques but only at the expense of picking up some false interactions. The problem of weighing the different interactions suffers in a similar way (the definition used in the pairwise case - any atom of residue A lies within any atom of residue B; gives rise to a lot of false interactions when looking at cliques). Our results demonstrate that an appropriate weighting scheme can improve the prediction results significantly (Fig 3.7) . All these problems need substantial sampling to gauge the best definitions for a multi-body cliques. These refinements are being incorporated in the next iterations of multi-body potentials.

4.4 Validation on the RalGEF-Ral system

The RalGEF-Ral system is an important signalling pathway involved in oncogenesis. The ability of Ral to bind to six different variants of RalGEFs (RalGDS, RalGPS1, RalGPS2, RGL1, RGL2, RGL3) was tested using our pairwise statistical potentials. All six variants of RalGEFs were predicted to bind to Ral. Four of these variants (RGL1, RGL2, RGL3, RalGDS) are found to bind Ral experimentally, whereas the other two variants might be weakly interacting. Since the statistical potentials make predictions on binding events of two proteins, we suspect that if all the different GEFs are put in a $in\ vitro$ setting, they will bind to Ral. However, in a cellular context, the bindings may not be strong, with proteins out-competing each other. Also, as statistical potentials are based on average properties of residue-residue interactions, they do not correlate well with binding affinities and hence fail to determine the RalGEF variants which bind Ral more strongly than the other variants.

We predicted some hotspot residues which upon mutations to other residues will weaken the interactions between RGL1-Ral complex. These hotspots residues sit in complementary clusters and hence the mutation of these residues to residues which disrupt the interaction (oppositely charged residues in case of polar residues) will lead to unfavourable interactions that would destabilise the complex. This observation shows that along with complex level prediction of protein-protein binding, our statistical potential can also help predict important interactions at the residue level.

4.5 Applications

Apart from predicting whether two protein subunits would form a stable association or not, these potentials can be applied to a variety of problems such as the prediction of binding hot spot residues, protein design etc.

Experimental alanine scanning is one of the best methods to determine the contribution of individual residues to the stabilisation of a protein-protein interface. However, this method is very labor-intensive as it involves systematically mutating all the residues in a protein to alanine and measuring the effect of the mutation on the binding of the complex. Statistical potentials such as the one presented in this thesis can be used as an alternative method to predict hot spot residues across protein interfaces. $In\ silico$ mutagenesis experiments are conducted on the protein of interest and then physical models of the protein are built. The resulting models are then scored using the statistical potential and the scores compared with the native model. Binding Hot Spot residues are then defined as those residues that lead to a large destabilisation in the final score of the protein.

These potentials can also aid in protein design processes. Given a protein structure, a favourable, complementary surface can be designed and optimised using these potentials which would ensure binding. Starting with a generic protein surface, residues on this surface can be tweaked to ensure complementarity with the target protein of interest. This method can also be employed to design novel antibodies.

The pairwise statistical potentials prediction system will be bundled in a web server in the near future, so that researchers can submit their protein complexes and make use of this facility. An ultimate test for the multibody potentials would be to test it on the solutions submitted by computational biologists for the target structures in CAPRI (Critical Assessment of PRediction of Interactions) (Janin, 2002). CAPRI is a community-wide, blind test experiment which tests the ability of protein-protein docking algorithms to predict modes of association between two proteins based on their three-dimensional structures.

References

Anfinsen, C B. Principles that govern the folding of protein chains. *Science (New York, N.Y.)*, 181(4096): 223--230, 1973. ISSN 0036-8075. doi: 10.1126/science.181.4096.223.

Argos, Patrick. An investigation of protein subunit and domain interfaces. *Protein Engineering, Design and Selection*, 2(2):101--113, 1988. ISSN 17410126. doi: 10.1093/protein/2.2.101.

Berman, H M; Westbrook, J; Feng, Z; Gilliland, G; Bhat, T N; Weissig, H; Shindyalov, I N, and Bourne, P E. The Protein Data Bank. *Nucleic acids research*, 28(1):235--242, 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235.

Braun, Pascal and Gingras, Anne Claude. History of protein-protein interactions: From egg-white to complex networks. *Proteomics*, 12:1478--1498, 2012. ISSN 16159853. doi: 10.1002/pmic.201100563.

Burley, SK and Petsko, GA. Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science*, 229(4708):23--28, 1985.

Causier, Barry and Davies, Brendan. Analysing protein-protein interactions with the yeast two-hybrid system. *Plant Molecular Biology*, 50(1989):855--870, 2002.

Chothia, C and Janin, J. Principles of protein-protein recognition. *Nature*, 256:705--8, 1975. ISSN 0028-0836. URL `http://www.ncbi.nlm.nih.gov/pubmed/1153006`.

Crowley, Peter B and Golovin, Adel. Cation-pi interactions in protein-protein interfaces. *Proteins*, 59 (February):231--239, 2005. ISSN 1097-0134. doi: 10.1002/prot.20417.

Davis, Fred P; Braberg, Hannes; Shen, Min-Yi; Pieper, Ursula; Sali, Andrej, and Madhusudhan, M S. Protein complex compositions predicted by structural similarity. *Nucleic acids research*, 34(10):2943--52, January 2006. ISSN 1362-4962. doi: 10.1093/nar/gkl353. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1474056&tool=pmcentrez&rendertype=abstract`.

Dill, K a. Dominant forces in protein folding. *Biochemistry*, 29(31):7133--7155, 1990. ISSN 0006-2960. doi: 10.1021/bi00483a001.

Fawcett, Tom. ROC Graphs : Notes and Practical Considerations for Researchers. *ReCALL*, 31:1--38, 2004. ISSN 08997667. doi: 10.1.1.10.9777. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.10.9777&amp;rep=rep1&amp;type=pdf`.

Glaser, Fabian; Steinberg, David M; Vakser, Ilya a; Ben-tal, Nir, and E-mail, Israel. Residue Frequencies and Pairing Preferences at Protein – Protein Interfaces protein – protein interfaces of known high-resolu- residue – residue contact preferences . The residue statistical strength of the data set . Differences between amino acid dist. *Interfaces*, 102(November 2000):89 --102, 2001.

Goldstein, R a; Luthey-Schulten, Z a, and Wolynes, P G. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 89(October):9029--9033, 1992. ISSN 0027-8424. doi: 10.1073/pnas.89.19.9029.

Heyda, Jan; Mason, Philip E., and Jungwirth, Pavel. Attractive interactions between side chains of histidine-histidine and histidine-arginine-based cationic dipeptides in water. *Journal of Physical Chemistry B*, 114: 8744--8749, 2010. ISSN 15206106. doi: 10.1021/jp101031v.

Janin, Joël. Welcome to CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins: Structure, Function and Genetics*, 47(August 1999):257, 2002. ISSN 08873585. doi: 10.1002/prot.10111.

Jones, S and Thornton, J M. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 93(January):13--20, 1996. ISSN 0027-8424. doi: 10.1073/pnas.93.1.13.

Kaytor, Michael D. and Warren, Stephen T. Aberrant protein deposition and neurological disease, 1999. ISSN 00219258.

Magalhaes, a.; Maigret, B.; Hoflack, J.; Gomes, J. N F, and Scheraga, H. a. Contribution of unusual Arginine-Arginine short-range interactions to stabilization and recognition in proteins. *Journal of Protein Chemistry*, 13(2):195--215, 1994. ISSN 02778033. doi: 10.1007/BF01891978.

Miyazawa, Sanzo and Jernigan, Robert L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of molecular biology*, 256:623--644, 1996. ISSN 0022-2836. doi: 10.1006/jmbi.1996.0114.

Nooren, I. M a and Thornton, Janet M. Diversity of protein-protein interactions. *EMBO Journal*, 22(14): 3486--3492, 2003. ISSN 02614189. doi: 10.1093/emboj/cdg359.

Ofran, Yanay and Rost, Burkhard. Analysing six types of protein-protein interfaces. *Journal of Molecular Biology*, 325(02):377--387, 2003. ISSN 00222836. doi: 10.1016/S0022-2836(02)01223-8.

Pednekar, Deepa; Tendulkar, Abhijit, and Durani, Susheel. Electrostatics-defying interaction between arginine termini as a thermodynamic driving force in protein-protein interaction. *Proteins: Structure, Function and Bioinformatics*, 74:155--163, 2009. ISSN 08873585. doi: 10.1002/prot.22142.

Pettersen, Eric F.; Goddard, Thomas D.; Huang, Conrad C.; Couch, Gregory S.; Greenblatt, Daniel M.; Meng, Elaine C., and Ferrin, Thomas E. UCSF Chimera - A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25:1605--1612, 2004. ISSN 01928651. doi: 10.1002/jcc.20084.

Rajamani, Deepa; Thiel, Spencer; Vajda, Sandor, and Camacho, Carlos J. Anchor residues in protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 101(31):11287--11292, 2004. ISSN 0027-8424. doi: 10.1073/pnas.0401942101.

Ross, P D and Subramanian, S. Thermodynamics of protein association reactions: forces contributing to stability. *Biochemistry*, 20:3096--3102, 1981. ISSN 0006-2960. doi: 10.1021/bi00514a017.

Sali, A and Blundell, T L. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, 234:779--815, 1993. ISSN 0022-2836. doi: 10.1006/jmbi.1993.1626.

Sardon, Teresa; Pache, Roland a; Stein, Amelie; Molina, Henrik; Vernos, Isabelle, and Aloy, Patrick. Uncovering new substrates for Aurora A kinase. *EMBO reports*, 11(12):977--984, 2010. ISSN 1469-221X. doi: 10.1038/embor.2010.171. URL `http://dx.doi.org/10.1038/embor.2010.171`.

Selzer, Tzvia and Schreiber, Gideon. New Insights into the Mechanism of Protein – Protein Association. *Change*, 198(May):190 --198, 2001. ISSN 0887-3585. doi: 10.1002/2001. URL `http://onlinelibrary.wiley.com/doi/10.1002/prot.1139/full`.

Sheinerman, Felix B.; Norel, Raquel, and Honig, Barry. Electrostatic aspects of protein-protein interactions. *Current Opinion in Structural Biology*, 10:153--159, 2000. ISSN 0959440X. doi: 10.1016/S0959-440X(00)00065-8.

Sippl, M J. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins., 1990. ISSN 0022-2836.

Sippl, M J. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *Journal of computer-aided molecular design*, 7:473--501, 1993. ISSN 0920-654X. doi: 10.1007/BF02337562.

Swets, J a; Dawes, R M, and Monahan, J. Better decisions through science. *Scientific American*, 283:82--87, 2000. ISSN 0036-8733. doi: 10.1038/scientificamerican1000-82.

Tanaka, S and Scheraga, H a. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, 9(6):945--50, 1976. ISSN 0024-9297. doi: 10.1021/ma60054a013. URL `http://www.ncbi.nlm.nih.gov/pubmed/1004017`.

Wang, Guoli and Dunbrack, Roland L. PISCES: Recent improvements to a PDB sequence culling server. *Nucleic Acids Research*, 33(10):94--98, 2005. ISSN 03051048. doi: 10.1093/nar/gki402.

Wong, Johnson M S; Ionescu, Daniela, and Ingles, C James. Interaction between BRCA2 and replication protein A is compromised by a cancer-predisposing mutation in BRCA2. *Oncogene*, 22:28--33, 2003. ISSN 09509232. doi: 10.1038/sj.onc.1206071.

Xu, D; Tsai, C J, and Nussinov, R. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein engineering*, 10(9):999--1012, 1997. ISSN 0269-2139. doi: 10.1093/protein/10.9.999.

Xu, Y.; Xu, D., and Liang, J. *Computational Methods for Protein Structure Prediction and Modeling: Volume 2: Structure Prediction*. Biological and Medical Physics, Biomedical Engineering. Springer, 2010. ISBN 9780387688251. URL `https://books.google.co.in/books?id=nVyitwIJ4QwC`.

Yasuda, Shinsuke; Morokawa, Nasa; Wong, G. William; Rossi, Andrea; Madhusudhan, Mallur S.; Šali, Andrej; Askew, Yuko S.; Adachi, Roberto; Silverman, Gary a.; Krilis, Steven a., and Stevens, Richard L. Urokinase-type plasminogen activator is a preferred substrate of the human epithelium serine protease tryptase ϵ/PRSS22. *Blood*, 105(10):3893--3901, 2005. ISSN 00064971. doi: 10.1182/blood-2003-10-3501.

Zhang, Qiangfeng Cliff; Petrey, Donald; Deng, Lei; Qiang, Li; Shi, Yu; Thu, Chan Aye; Bisikirska, Brygida; Lefebvre, Celine; Accili, Domenico; Hunter, Tony; Maniatis, Tom; Califano, Andrea, and Honig, Barry. Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature*, 490:556--560, 2012. ISSN 0028-0836. doi: 10.1038/nature11503. URL `http://dx.doi.org/10.1038/nature11503`.

Appendix A

Supplementary Data

A.1 Training Set PDB codes

1A25	1A2O	1A6J	1AAZ	1AC6	1ADU	1AMU	1AOE	1AOH	1AOR	1AQU	1AT3	1ATZ	1AU1
1AYO	1AZT	1AZW	1B3U	1B43	1B88	1BC5	1BCM	1BEH	1BF6	1BIN	1BJA	1BMT	1BQU
1BXT	1BYF	1C3R	1C94	1C9O	1CI4	1CI9	1CJA	1CKU	1COL	1COZ	1CP2	1CQ3	1CRU
1CZP	1D0Q	1D2O	1DBW	1DDV	1DEB	1DEK	1DJ7	1DJT	1DLE	1DM9	1DNP	1DOW	1DQE
1DVK	1DWU	1DYN	1DYO	1DYS	1DZK	1E0B	1E30	1E5R	1E6C	1E8C	1E9G	1EAJ	1ECE
1EDM	1EEJ	1EEO	1EGA	1EI7	1EJD	1EJF	1EK6	1EKE	1EM9	1EPA	1EQ9	1EUJ	1EUV
1EXT	1EYV	1EZG	1F08	1F0K	1F0L	1F1C	1F35	1F39	1F46	1F6B	1F7D	1F86	1F9M
1FIW	1FJ2	1FJR	1FM0	1FMT	1FN8	1FN9	1FNN	1FOC	1FP3	1FQT	1FS5	1FSG	1FUU
1G61	1G71	1G8Q	1GEQ	1GG4	1GGG	1GHE	1GIQ	1GL4	1GNW	1GNX	1GOI	1GPE	1GQA
1GTD	1GU2	1GU7	1GUD	1GVE	1GVF	1GVK	1GVU	1GXM	1GYG	1GYO	1H03	1H1O	1H2B
1H32	1H3F	1H3L	1H4P	1H4R	1H4X	1H6G	1H7S	1H80	1H8G	1H8P	1H97	1H9O	1HEK
1HKQ	1HLC	1HPL	1HRU	1HSL	1HST	1HY5	1I19	1I31	1I3Z	1I4J	1I4N	1I4U	1I7K
1IHN	1II2	1IJY	1IN0	1IO7	1IOO	1IPS	1IQ4	1IRD	1IRX	1ISI	1IT2	1ITH	1ITV
1IWM	1IX9	1IXC	1IYB	1IZ5	1J0W	1J1N	1J2X	1J3M	1J6R	1J71	1J7J	1J83	1JAT
1JEK	1JET	1JFL	1JH6	1JHF	1JI1	1JIH	1JL0	1JL9	1JMK	1JMT	1JO0	1JR2	1JS8
1JVN	1JYA	1K07	1K0E	1K38	1K3S	1K4Z	1K66	1K68	1K6D	1K8Q	1KAG	1KAP	1KCF
1KHV	1KJN	1KMT	1KNQ	1KOL	1KPT	1KRH	1KU1	1KUG	1KUT	1KWA	1KXI	1KXJ	1KYF
1L1E	1L1J	1L4I	1L5J	1L6R	1L7A	1L7M	1L8R	1L9M	1LB6	1LEH	1LF6	1LK0	1LKK
1LM4	1LM5	1LM7	1LNZ	1LQT	1LWJ	1LXD	1LYQ	1M0Z	1M1F	1M1Z	1M2D	1M45	1M48
1M55	1M6U	1M8A	1MBY	1MI1	1MIW	1MJH	1MK4	1MKI	1MKZ	1MOL	1MPG	1MQS	1MQV
1MY7	1MZG	1N08	1N0S	1N1B	1N2Z	1N45	1N46	1N7H	1N8V	1NBQ	1NCN	1ND4	1NNW
1NO7	1NOW	1NPE	1NQ7	1NQJ	1NS5	1NSZ	1NTV	1NU0	1NU4	1NUB	1NUL	1NUU	1NXM
1O0W	1O12	1O5U	1O63	1O7I	1O81	1O8B	1OAI	1OBB	1OBO	1OBX	1OCU	1ODZ	1OF3
1OFZ	1OH0	1OHU	1OIZ	1OJ5	1OMZ	1ON2	1OOH	1OQJ	1OW4	1P0K	1P1X	1P4U	1P5T
1P7W	1P9L	1P9Y	1PAM	1PBW	1PD3	1PE9	1PFB	1PGU	1PKH	1PP3	1PP4	1PQ4	1PQH
1PS1	1PT6	1PUI	1PX5	1PXY	1PZL	1PZX	1Q1A	1Q3O	1Q67	1Q77	1Q7F	1Q8Y	1QAH
1QDL	1QEX	1QF8	1QFT	1QGR	1QH5	1QJC	1QJJ	1QJS	1QKR	1QKS	1QLS	1QO2	1QOZ
1QSD	1QUP	1QW9	1QWT	1QYA	1QYR	1R12	1R1D	1R77	1R7A	1R7L	1R9D	1RD5	1REG
1RG8	1RHF	1RHY	1RIF	1RKI	1RKQ	1RP0	1RRL	1RRM	1RW0	1RYL	1RZU	1RZX	1S0P
1S4K	1S4N	1S5P	1S98	1S9R	1SEI	1SFD	1SFL	1SH0	1SH8	1SJ1	1SMO	1SMX	1SQJ
1SQU	1SUL	1SW6	1SWV	1SZ0	1SZH	1SZW	1T0I	1T0P	1T1V	1T2L	1T3G	1T4O	1T6F
1T6T	1T7R	1T92	1TBX	1TDQ	1TE2	1TE5	1TH8	1THT	1TIQ	1TL9	1TLT	1TOA	1TR8
1TVF	1TVN	1TW4	1U00	1U07	1U19	1U5K	1U5U	1U7B	1UAX	1UC7	1UCG	1UCR	1UEB

1UG3 1UJ2 1UJN 1UJW 1UKC 1UMU 1UMZ 1UOC 1UPK 1UPS 1UQT 1URH 1URJ 1URS
1UTI 1UV7 1UWW 1UWZ 1UXZ 1UZ3 1V1A 1V1P 1V37 1V47 1V74 1V8H 1V96 1V9K
1VA6 1VBK 1VC1 1VC4 1VCD 1VDR 1VDW 1VH5 1VHX 1VI2 1VIA 1VIO 1VJ7 1VJL
1VJQ 1VJU 1VKI 1VL4 1VM7 1VMA 1VMO 1VP2 1VPV 1VQQ 1VQU 1VS3 1VYB 1VZY
1W32 1W5R 1W94 1W9C 1W9P 1W9S 1WB4 1WB7 1WC3 1WDU 1WDV 1WEH 1WKO 1WKR
1WLG 1WMH 1WMS 1WMX 1WN1 1WOQ 1WPN 1WQ6 1WR8 1WRA 1WSC 1WSR 1WUF 1WV2
1WVG 1WWL 1WWM 1WWP 1WZ9 1WZD 1X2I 1X6I 1X7O 1X9Z 1XAH 1XCR 1XFS 1XG2
1XGS 1XHK 1XI3 1XIY 1XJU 1XK9 1XM7 1XM8 1XOC 1XOF 1XQA 1XQR 1XRP 1XRS
1XSZ 1XTN 1XVI 1XVS 1XVW 1XYZ 1XZO 1Y0U 1Y1M 1Y1P 1Y3T 1Y44 1Y4T 1Y5H
1Y71 1Y7Y 1Y9Z 1YAC 1YBX 1YC0 1YC5 1YCD 1YDY 1YF2 1YGA 1YLM 1YLQ 1YLX
1YMT 1YNP 1YOC 1YOD 1YOZ 1YPF 1YPQ 1YPY 1YQ1 1YQ5 1YQD 1YQH 1YRK 1YRR
1YZ4 1YZH 1YZY 1Z1Y 1Z2W 1Z2Z 1Z3E 1Z6U 1Z72 1Z85 1Z96 1ZB1 1ZC6 1ZEE
1ZH8 1ZHH 1ZJ8 1ZKC 1ZKD 1ZKI 1ZLP 1ZPL 1ZQ9 1ZSO 1ZTD 1ZUO 1ZUY 1ZVT
1ZY4 1ZY7 1ZYS 1ZZW 2A0S 2A1K 2A2M 2A2R 2A35 2A5L 2A6A 2A6P 2A70 2A8N
2A9D 2AB5 2ABQ 2ABW 2ACV 2AE2 2AEE 2AFB 2AFC 2AFW 2AG4 2AHF 2AHX 2AIB
2AJA 2AMX 2ANX 2APO 2AQ6 2AQP 2AR0 2ARC 2AS9 2ASU 2AUW 2AVN 2AYT 2AZ4
2B0R 2B1L 2B2N 2B3R 2B3Y 2B4M 2B6C 2B82 2B8N 2B97 2B9D 2B9H 2B9R 2BBA
2BCO 2BGH 2BHG 2BJD 2BJN 2BKL 2BKM 2BLF 2BLN 2BM5 2BON 2BPH 2BPO 2BPS
2BRW 2BRY 2BSJ 2BT6 2BU3 2BV4 2BVF 2BWF 2BWR 2BYC 2BZ9 2C0G 2C3I 2C40
2C5U 2C77 2C8J 2C95 2CAR 2CAY 2CB8 2CC0 2CC3 2CFA 2CFO 2CGK 2CI5 2CIA
2CJ4 2CJP 2CKD 2CN3 2CO5 2CU3 2CUN 2CV8 2CVH 2CVI 2CX6 2CX7 2CXD 2CY9
2D1G 2D1H 2D42 2D4G 2D5C 2DB0 2DB7 2DBS 2DC0 2DC3 2DC4 2DEB 2DFJ 2DFY
2DI4 2DOK 2DPR 2DPS 2DPY 2DQ4 2DQA 2DQL 2DQW 2DS5 2DSJ 2DTC 2DUR 2DXU
2DYJ 2E12 2E2E 2E3P 2E5Y 2E85 2E8G 2E8Y 2EAB 2EAV 2EAY 2EBE 2EBJ 2EEN
2EF8 2EG4 2EGG 2EGJ 2EGZ 2EHP 2EIH 2EIS 2EIX 2EJA 2EJN 2EJQ 2EKC 2ERV
2EUC 2EX0 2EXV 2F20 2F23 2F25 2F31 2F37 2F3O 2F4E 2F4M 2F51 2F5J 2F5Y
2F7L 2F8M 2F8Y 2F9H 2F9S 2FAE 2FAO 2FAZ 2FCO 2FCT 2FCW 2FEA 2FFG 2FFI
2FFU 2FH5 2FHP 2FHQ 2FHZ 2FIA 2FJR 2FK5 2FLU 2FN0 2FNA 2FNO 2FP1 2FPR
2FSH 2FSK 2FT0 2FTR 2FTX 2FU4 2FV7 2FVU 2FYX 2FZF 2G09 2G3W 2G58 2G6T
2G7Z 2G8L 2GA1 2GAI 2GAK 2GCL 2GCO 2GD9 2GDQ 2GEC 2GF3 2GF4 2GFF 2GGS
2GGZ 2GHA 2GHV 2GIY 2GJ3 2GLZ 2GMF 2GMQ 2GN4 2GOM 2GOP 2GP4 2GPY 2GPZ
2GRR 2GRU 2GS9 2GSO 2GSV 2GT1 2GV9 2GVY 2GZ6 2GZB 2GZX 2H1C 2H1E 2H1Y
2H34 2H3H 2H7Z 2H98 2HAL 2HB0 2HBA 2HDI 2HDV 2HEK 2HEV 2HF1 2HF2 2HF9
2HFS 2HI0 2HIH 2HIN 2HJ3 2HJV 2HKE 2HLC 2HLS 2HNL 2HP4 2HPL 2HQ4 2HQ9
2HQY 2HRA 2HRV 2HSI 2HTA 2HU9 2HW6 2HWY 2I02 2I0E 2I1S 2I1Y 2I27 2I2O
2I4R 2I4S 2I58 2I5G 2I6H 2I6K 2I6L 2I74 2I9X 2IA1 2IAB 2IB0 2IBN 2IC2
2ICH 2ID1 2IDL 2IEP 2IEW 2IG3 2IM8 2IMZ 2IN5 2INW 2IQJ 2IRP 2IRU 2ISM
2ITB 2ITM 2IUY 2IXN 2IXO 2IXS 2IYG 2IYK 2IZ6 2J16 2J1V 2J4D 2J5B 2J5Y
2J8I 2J9W 2JBV 2JBX 2JCB 2JD4 2JDA 2JDJ 2JE8 2JEM 2JEP 2JF7 2JFZ 2JGB
2JHN 2JIG 2JIK 2JJ7 2JK9 2JKG 2JKH 2MSB 2NLI 2NLV 2NOG 2NRV 2NS9 2NTE
2NTT 2NTX 2NUJ 2NV0 2NW0 2NYU 2NZ5 2O16 2O1E 2O1K 2O1Q 2O2K 2O2T 2O30
2O3B 2O3I 2O5A 2O5H 2O5N 2O62 2O6L 2O6P 2O7G 2O8S 2OA9 2OAF 2OB3 2OB9
2OD0 2OD4 2ODA 2ODM 2OEE 2OER 2OFC 2OFP 2OFY 2OG1 2OGI 2OJL 2OKC 2OKG
2OL7 2OLW 2OM6 2OOC 2OOI 2OPI 2OQ1 2OQA 2OQB 2OQC 2OQQ 2ORV 2ORW 2OTN
2OUS 2OVS 2OWA 2OWL 2OXC 2OXL 2OY9 2OYK 2OZ5 2OZJ 2OZV 2OZZ 2P08 2P0M
2P11 2P12 2P13 2P1A 2P1G 2P35 2P38 2P3P 2P4P 2P4Z 2P62 2P6C 2P6H 2P6X
2P7I 2P8J 2P9R 2PA2 2PBF 2PD2 2PF6 2PFI 2PHK 2PIE 2PIF 2PK3 2PKE 2PKF
2PLG 2PLR 2PMA 2PNZ 2POF 2PPT 2PPW 2PQG 2PQV 2PR7 2PR8 2PRV 2PRX 2PSP
2PW0 2PX6 2PYG 2PZ0 2PZE 2PZI 2Q03 2Q0N 2Q24 2Q2B 2Q2G 2Q3F 2Q3G 2Q3X
2Q5C 2Q5W 2Q6O 2Q7T 2Q7X 2Q83 2Q8X 2Q9O 2QA9 2QAI 2QB7 2QCQ 2QCU 2QCX
2QDQ 2QDR 2QE8 2QEB 2QF4 2QF9 2QG3 2QH5 2QH9 2QHQ 2QJ3 2QJ8 2QJV 2QJZ
2QKH 2QKL 2QMW 2QN4 2QND 2QOS 2QRR 2QS8 2QSJ 2QSQ 2QSX 2QTY 2QV0 2QV5
2QX5 2QXX 2QXY 2QY1 2QY6 2QYC 2QYV 2QZ7 2QZA 2QZC 2R15 2R19 2R25 2R2A
2R5X 2R6J 2R6O 2R6Z 2R76 2R85 2R8B 2R8Q 2R8R 2RA4 2RAD 2RB6 2RBD 2RBG

2REE	2REK	2RFM	2RG4	2RG8	2RI9	2RJI	2RJW	2RKK	2RL8	2RMP	2SCP	2SQC	2UVF
2UWI	2UXT	2V1Q	2V1Y	2V25	2V27	2V2F	2V33	2V3T	2V3Z	2V5C	2V5E	2V6U	2V6V
2V8P	2V94	2V9B	2V9T	2VA8	2VCY	2VD3	2VE3	2VGX	2VH1	2VH3	2VHA	2VHF	2VK7
2VKP	2VLQ	2VLU	2VNG	2VOB	2VOK	2VOZ	2VP8	2VPN	2VPP	2VPV	2VQ3	2VQH	2VQQ
2VRN	2VVE	2VVG	2VVW	2VXB	2VXG	2VYI	2VZC	2W00	2W1J	2W1K	2W2G	2W3G	2W3Y
2W50	2W53	2W56	2W59	2W5F	2W5Z	2W7A	2W7V	2W7Z	2W8D	2W8M	2W9J	2W9T	2W9X
2WB7	2WB9	2WCR	2WD6	2WE8	2WEE	2WEK	2WFH	2WFV	2WG7	2WHN	2WIV	2WJ9	2WKF
2WNH	2WNY	2WOD	2WOK	2WP4	2WPX	2WRZ	2WTP	2WUQ	2WVQ	2WZ1	2X02	2X03	2X0K
2X1Q	2X32	2X3J	2X4D	2X4K	2X5Q	2X61	2X6R	2X7X	2X8S	2X98	2X9J	2X9Q	2XCJ
2XE4	2XEP	2XES	2XET	2XEX	2XFA	2XFV	2XGG	2XGU	2XHA	2XHF	2XHS	2XI8	2XI9
2XMJ	2XMO	2XMX	2XOC	2XOL	2XOT	2XPP	2XQX	2XR1	2XSS	2XSW	2XT2	2XTL	2XTM
2XTY	2XUA	2XUS	2XVC	2XVM	2XXN	2XYI	2XZ4	2XZ8	2XZI	2Y1H	2Y2X	2Y43	2Y4J
2Y7E	2Y7I	2Y7S	2Y8E	2Y8U	2Y9M	2YB7	2YCH	2YEQ	2YFQ	2YG2	2YHN	2YJ6	2YJG
2YKT	2YMY	2YN1	2YN5	2YN7	2YNA	2YOA	2YOC	2YOR	2YQY	2YQZ	2YR1	2YV9	2YVR
2YWW	2YXD	2YXE	2YXO	2YXW	2YY6	2YYB	2YYS	2YYV	2Z0U	2Z22	2Z26	2Z5B	2Z5D
2Z64	2Z73	2Z8F	2Z8G	2ZAY	2ZBI	2ZC2	2ZCA	2ZFU	2ZGY	2ZKT	2ZMV	2ZOS	2ZOU
2ZSI	2ZTB	2ZU9	2ZVD	2ZVR	2ZWA	2ZWI	2ZWR	2ZX2	2ZXD	2ZYR	2ZZ8	2ZZV	3A07
3A0Y	3A1D	3A1S	3A21	3A24	3A35	3A43	3A45	3A4M	3A4R	3A4T	3A54	3A5I	3A6S
3A9F	3A9L	3AAG	3AAY	3AB1	3ABG	3ADR	3AEH	3AEI	3AFF	3AFM	3AGX	3AHN	3AIH
3AJ6	3AJA	3AJR	3AKJ	3AL3	3AMI	3AMN	3ANW	3AOF	3APQ	3APR	3APT	3APU	3APZ
3AQ9	3AQG	3AQL	3AS5	3ASL	3ATY	3AU4	3AVR	3AWU	3AXA	3AXD	3AYC	3AZD	3AZO
3B0F	3B0P	3B4N	3B4Q	3B5E	3B5I	3B5Q	3B6H	3B73	3B7S	3B85	3BA3	3BBD	3BBZ
3BDV	3BE3	3BEU	3BF7	3BFV	3BGA	3BGE	3BGH	3BGY	3BH4	3BHD	3BHW	3BIT	3BJ4
3BMX	3BNW	3BO6	3BOH	3BOO	3BP3	3BQ9	3BQO	3BQP	3BRN	3BRS	3BS6	3BS7	3BTP
3BUS	3BVO	3BVP	3BW1	3BWS	3BWV	3BXP	3BXW	3BYP	3BZB	3BZY	3C0G	3C0U	3C1A
3C3R	3C4N	3C4S	3C4V	3C57	3C5N	3C7M	3C8C	3C8L	3C9F	3C9G	3C9H	3C9Q	3CB2
3CBW	3CEG	3CEI	3CEU	3CEX	3CFU	3CG6	3CG7	3CHH	3CIO	3CIT	3CJP	3CK1	3CKC
3CLK	3CNH	3CNR	3CNY	3COB	3COK	3COL	3COV	3CP7	3CPT	3CQB	3CQC	3CQL	3CRN
3CT6	3CU2	3CU5	3CUC	3CV0	3CWC	3CWF	3CWV	3CX3	3CYG	3CZ1	3CZB	3CZH	3D21
3D34	3D37	3D3B	3D3Q	3D4J	3D59	3D5J	3D5P	3D6I	3D6R	3D6W	3D7A	3D8C	3D8D
3D8U	3D9Y	3DA5	3DAD	3DB0	3DBA	3DBG	3DC6	3DCD	3DDE	3DDL	3DEP	3DEU	3DGP
3DKA	3DLQ	3DME	3DNF	3DNS	3DNT	3DO8	3DOH	3DR2	3DRF	3DRN	3DRW	3DS2	3DSK
3DTB	3DTN	3DUP	3DWM	3DWV	3DXO	3DXQ	3DXR	3DYJ	3DYN	3DZV	3E0X	3E11	3E2V
3E3R	3E48	3E4W	3E57	3E58	3E7H	3E7J	3E9C	3E9G	3EA0	3EAE	3EAG	3EB8	3EB9
3EC3	3EC4	3EC9	3ECN	3ECQ	3ECR	3EDN	3EDP	3EDV	3EE6	3EEA	3EEF	3EEQ	3EFP
3EGR	3EHD	3EIP	3EJW	3ELK	3EMX	3EN9	3ENC	3ENP	3EO6	3EOP	3EOQ	3EOZ	3EP0
3EPS	3EQX	3EQZ	3ERP	3ERX	3ESA	3ESL	3ETC	3ETO	3ETQ	3ETZ	3EU7	3EUS	3EVI
3EVY	3EWI	3EWL	3EWM	3EWO	3EYY	3EZH	3F08	3F0P	3F13	3F1P	3F42	3F4A	3F5H
3F66	3F69	3F6C	3F6I	3F6K	3F6O	3F70	3F7E	3F7Q	3F8B	3F95	3F9U	3FB9	3FBG
3FCD	3FCG	3FCM	3FD4	3FD7	3FDI	3FDW	3FDX	3FE3	3FE4	3FF1	3FF5	3FG7	3FGV
3FHW	3FID	3FIL	3FJV	3FLA	3FLE	3FLT	3FM2	3FM3	3FN1	3FN5	3FNC	3FO3	3FO5
3FPK	3FPN	3FPR	3FQD	3FQM	3FSO	3FUT	3FVD	3FVV	3FVW	3FW3	3FYF	3FZY	3G12
3G1E	3G1J	3G1P	3G23	3G2F	3G2M	3G3R	3G3S	3G46	3G48	3G4D	3G4E	3G5J	3G68
3G8K	3GAE	3GAX	3GAZ	3GBV	3GBY	3GD4	3GDI	3GF5	3GF6	3GFF	3GFV	3GGN	3GHD
3GI7	3GID	3GJ0	3GKN	3GKX	3GLV	3GME	3GMG	3GNL	3GO6	3GOC	3GPK	3GPV	3GQS
3GRA	3GRD	3GRI	3GRN	3GRO	3GRZ	3GUD	3GUE	3GUU	3GV4	3GVE	3GWB	3GWL	3GWR
3GXH	3GYC	3GYZ	3GZ5	3GZA	3H05	3H1Q	3H2B	3H2S	3H30	3H3A	3H3N	3H5J	3H5L
3H7J	3H7O	3H8Q	3H8V	3HA2	3HAM	3HBW	3HCS	3HCW	3HCY	3HDF	3HDT	3HEB	3HFH
3HGT	3HHF	3HHI	3HIS	3HJ4	3HJ6	3HJG	3HK0	3HKL	3HKS	3HKV	3HL1	3HL6	3HLK
3HM4	3HME	3HMT	3HN0	3HN5	3HO6	3HOB	3HPE	3HPK	3HQR	3HRQ	3HRS	3HS3	3HU5
3HV1	3HV2	3HWJ	3HWO	3HWP	3HY0	3HYJ	3I00	3I1A	3I1I	3I3Q	3I3W	3I41	3I4O
3I57	3I5O	3I5R	3I5W	3I6D	3I6S	3I7J	3I83	3I8N	3I9F	3IA1	3IA8	3IAU	3IB3
3IBW	3IBX	3IC5	3ICY	3ID9	3IDF	3IE4	3IE5	3IEG	3IGE	3IHS	3IHT	3IHV	3IIC
3IJL	3IJM	3IJW	3IKB	3INO	3IO1	3IOL	3IPF	3IPJ	3IPO	3IQ0	3IQ2	3IQC	3IQU

3IR9 3IRB 3IS6 3ITE 3ITQ 3ITW 3IU1 3IUK 3IUO 3IUP 3IUS 3IUW 3IUY 3IV7
3IVL 3IVV 3IWF 3IWG 3IX1 3IX3 3IX7 3IX9 3JQ1 3JR7 3JRR 3JRU 3JSB 3JSL
3JT0 3JTN 3JUU 3JW8 3JWI 3JXF 3JXO 3K0Z 3K1R 3K1W 3K2A 3K2N 3K2O 3K2Z
3K51 3K5O 3K6F 3K6O 3K7B 3K85 3K8G 3K8R 3K9V 3KA5 3KB1 3KB2 3KBQ 3KBY
3KD3 3KD4 3KD6 3KDG 3KEA 3KEP 3KEW 3KEZ 3KF6 3KFA 3KG8 3KG9 3KGK 3KHE
3KHN 3KK7 3KKS 3KLQ 3KM5 3KMA 3KMI 3KMR 3KNB 3KPE 3KS9 3KSM 3KSU 3KTZ
3KUZ 3KWR 3KY9 3KYJ 3KZY 3L01 3L0Q 3L0R 3L12 3L15 3L18 3L32 3L46 3L50
3L6I 3L6T 3L6V 3L6X 3L7O 3L81 3L8C 3L8E 3L8M 3L9J 3LAE 3LAG 3LAZ 3LB2
3LET 3LF5 3LFR 3LFT 3LG3 3LGB 3LHN 3LHX 3LID 3LIF 3LIU 3LJB 3LJS 3LKB
3LKL 3LLM 3LLP 3LLZ 3LM2 3LMH 3LMN 3LNN 3LNY 3LQ9 3LQM 3LS1 3LS8 3LST
3LUO 3LVC 3LWE 3LX4 3LX6 3LXQ 3LY0 3LYP 3LYX 3M33 3M6Z 3M7A 3M8J 3M8T
3MAB 3MAL 3MAZ 3MB4 3MBC 3MC9 3MCA 3MCB 3MCF 3MCS 3MCW 3MD1 3MD9 3MDF
3ME4 3ME7 3MEA 3MER 3MFD 3MG1 3MGD 3MGG 3MH9 3MIT 3MIZ 3MJQ 3MK4 3MKL
3MNL 3MOZ 3MPC 3MPD 3MQ2 3MR0 3MTI 3MTK 3MTR 3MUQ 3MUX 3MVC 3MVP 3MWB
3MWX 3MX3 3MXO 3MYU 3MYV 3MYX 3MZ2 3N01 3N08 3N10 3N1E 3N4I 3N6Y 3N72
3N89 3N9B 3N9V 3NA5 3NBC 3NCE 3NCX 3NDA 3NDO 3NEK 3NEQ 3NFH 3NFQ 3NGF
3NHM 3NI7 3NIQ 3NJ2 3NJE 3NK6 3NKL 3NKU 3NME 3NMW 3NNG 3NNN 3NNS 3NO8
3NPF 3NPP 3NQN 3NR1 3NRF 3NRH 3NRL 3NRX 3NT8 3NTK 3NTX 3NUF 3NW0 3NWP
3NY3 3NYI 3NZE 3NZN 3NZZ 3O0A 3O0L 3O0Q 3O0X 3O14 3O2U 3O53 3O5Y 3O60
3O66 3O6W 3O7A 3O83 3O8Q 3OAJ 3OBE 3OBF 3OBH 3OBL 3OBQ 3OBY 3OCO 3OCP
3OG5 3OG6 3OG7 3OGN 3OHE 3OIQ 3OKW 3OKZ 3OL3 3OMD 3OMT 3ON3 3ON9 3ONM
3OOV 3OOX 3OP6 3OPE 3OQI 3OT2 3OTN 3OVP 3OWC 3OWG 3OXP 3OY2 3OYO 3OYY
3OZD 3OZI 3OZX 3P09 3P0U 3P1U 3P2E 3P3Q 3P3V 3P5R 3P69 3P6A 3P9X 3PAF
3PC6 3PD7 3PE5 3PES 3PET 3PF8 3PG7 3PGG 3PGS 3PH9 3PHG 3PHX 3PIV 3PJP
3PM6 3PMC 3PMG 3PNR 3PQH 3PQU 3PRB 3PSM 3PSQ 3PT3 3PT8 3PU8 3PVE 3PWX
3Q18 3Q1I 3Q2J 3Q49 3Q6K 3Q6V 3Q72 3Q7R 3Q87 3QAT 3QAX 3QB8 3QC2 3QC4
3QE2 3QEE 3QEK 3QF2 3QHB 3QHP 3QHQ 3QI7 3QIJ 3QIS 3QN9 3QPI 3QR2 3QR7
3QRC 3QSL 3QSZ 3QT5 3QTA 3QTG 3QTM 3QU5 3QUF 3QVL 3QVM 3QW9 3QWG 3QX1
3QYE 3QYY 3QZ4 3QZM 3QZR 3R07 3R0J 3R15 3R1J 3R27 3R41 3R42 3R4R 3R4S
3R5Z 3R62 3R6A 3R7A 3R7G 3RA5 3RAO 3RAU 3RB5 3RBY 3RC4 3RDK 3RE1 3RE4
3RFS 3RGC 3RGH 3RH0 3RHY 3RHZ 3RI0 3RJT 3RK1 3RKC 3RLS 3RMH 3RNQ 3RO3
3ROI 3ROT 3RP2 3RPJ 3RQ9 3RS1 3RUX 3RV6 3RY0 3RY3 3S0R 3S0T 3S2X 3S4K
3S5B 3S5F 3S5W 3S63 3S6E 3S7D 3S8I 3S8K 3S8P 3S93 3S95 3S9U 3SAF 3SAO
3SCZ 3SD4 3SEI 3SEO 3SFW 3SG8 3SGH 3SHP 3SIM 3SIT 3SJ5 3SKV 3SLU 3SLZ
3SO6 3SOJ 3SOK 3SON 3SOV 3SP1 3SP4 3SPE 3SQF 3SQJ 3SRI 3STY 3SUB 3SUK
3SWH 3SYL 3SZ6 3T0P 3T13 3T1O 3T47 3T4L 3T5G 3T5X 3T6K 3T8B 3T9G 3T9K
3TB6 3TBH 3TC8 3TCA 3TCN 3TCR 3TCV 3TDN 3TDQ 3TDV 3TE8 3TEB 3TEJ 3TEK
3TEV 3TFG 3TII 3TIQ 3TKF 3TL1 3TLQ 3TM8 3TOD 3TOV 3TP2 3TP9 3TQF 3TQW
3TRB 3TSA 3TSJ 3TSM 3TTM 3TUF 3TV1 3TVA 3TVT 3TWD 3TWE 3TWF 3TWK 3TX3
3TYQ 3TZG 3TZT 3U0H 3U0J 3U1D 3U1U 3U1X 3U21 3U23 3U3B 3U4T 3U4Y 3U4Z
3U7R 3U7Z 3U80 3U8V 3U96 3U9J 3U9Q 3UAN 3UC4 3UEC 3UES 3UGF 3UHA 3UID
3UIW 3UL3 3ULJ 3ULL 3ULT 3ULY 3UMZ 3UN7 3UO3 3UP1 3UP3 3UPV 3UR8 3URR
3USH 3USS 3USY 3UT4 3UUG 3UV0 3UV1 3UXN 3UY7 3UYJ 3V0D 3V1E 3V30 3V33
3V3L 3V43 3V48 3V67 3V69 3V8D 3V8I 3V97 3V98 3VAS 3VAY 3VCC 3VCF 3VDH
3VEJ 3VF1 3VFZ 3VHS 3VJA 3VJE 3VJP 3VK5 3VKG 3VMT 3VO2 3VOQ 3VPP 3VPS
3VRC 3VTA 3VTH 3VTX 3VU2 3VU4 3VU9 3VUP 3VUS 3VV1 3VV3 3VV5 3VX3 3VX4
3VYP 3VZI 3W08 3W0E 3W0K 3W19 3W1O 3W2Y 3W3W 3W4S 3W57 3W5F 3W5S 3W6P
3W7T 3W9S 3W9V 3WA4 3WA8 3WAE 3WAS 3WDF 3WDW 3WE2 3WE5 3WEA 3WEU 3WFI
3WH9 3WHT 3WI7 3WJ9 3WKY 3WL2 3WL4 3WL6 3WMD 3WMG 3WMI 3WNO 3WOL 3WPW
3WQO 3WUR 3WV4 3WWN 3WX1 3WYD 3ZBD 3ZBO 3ZD2 3ZFI 3ZG6 3ZGJ 3ZH5 3ZHO
3ZIH 3ZIL 3ZIT 3ZIU 3ZJE 3ZK9 3ZL1 3ZME 3ZMR 3ZO9 3ZPY 3ZQS 3ZRG 3ZTP
3ZWF 3ZXC 3ZXF 3ZXN 3ZY7 3ZYG 3ZYL 3ZYR 3ZYW 4A0E 4A0Z 4A2O 4A37 4A48
4A6F 4A6V 4A7U 4A7W 4A8H 4AAZ 4ACV 4ADN 4ADT 4ADY 4ADZ 4AE4 4AEE 4AEF
4AGG 4AHC 4AJW 4AKL 4AKM 4ALF 4AM6 4AMJ 4APX 4AQN 4ARV 4ASR 4AU9 4AUC

4AUP 4AVB 4AVR 4AWX 4AXK 4AXN 4AY0 4AYA 4AYG 4B0Z 4B1Y 4B2N 4B3B 4B45
4B5Q 4B61 4B6G 4B6M 4B6X 4B8B 4B8E 4B91 4B93 4B9G 4BD2 4BEG 4BFA 4BG2
4BGO 4BHR 4BI3 4BLU 4BND 4BOP 4BPG 4BPZ 4BQ4 4BQ9 4BQN 4BQU 4BRC 4BS6
4BSZ 4BUC 4BUU 4BVQ 4BVX 4BWO 4BWV 4BX8 4BXH 4C0R 4C16 4C1D 4C1L 4C1S
4C23 4C29 4C3D 4C76 4C7A 4C7D 4C8B 4C97 4C9Y 4CA1 4CB7 4CBP 4CDJ 4CEM
4CGR 4CGS 4CGY 4CHF 4CHH 4CI7 4CI8 4CJ0 4CJ9 4CK4 4CMR 4CQ8 4CRW 4CSD
4CU9 4CUA 4CXF 4CXV 4CZJ 4CZX 4D05 4D0O 4D0Y 4D2C 4D2O 4D8I 4D9I 4D9S
4DBG 4DCB 4DCZ 4DEY 4DGF 4DGH 4DHK 4DI8 4DIX 4DJB 4DJG 4DKC 4DKN 4DLH
4DLQ 4DM4 4DO4 4DO7 4DOI 4DOK 4DOO 4DOV 4DQ9 4DQZ 4DS2 4DSD 4DT5 4DTE
4DY0 4DYH 4DYN 4DYW 4DZM 4DZZ 4E15 4E19 4E1Y 4E3Y 4E57 4E5V 4E5W 4E6F
4E7S 4E8U 4E94 4E9J 4EBR 4ECO 4EDH 4EE6 4EEI 4EET 4EF0 4EFO 4EG0 4EGD
4EH1 4EHS 4EHU 4EI0 4EI7 4EIB 4EIR 4EIS 4EIV 4EJR 4EMT 4EP4 4EPP 4EQB
4EQQ 4ERC 4ERY 4ES8 4ETV 4ETZ 4EUK 4EUU 4EVQ 4EVU 4EVW 4EW5 4EWI 4EWL
4EYG 4EYZ 4EZG 4F0D 4F14 4F1J 4F27 4F3V 4F3Y 4F44 4F4F 4F7K 4F7O 4F82
4FCH 4FCZ 4FD4 4FD9 4FDI 4FDX 4FDY 4FEK 4FET 4FGQ 4FHR 4FID 4FKB 4FKZ
4FP1 4FPW 4FQ5 4FQD 4FRF 4FRX 4FXQ 4FYP 4FYT 4FZL 4FZP 4FZV 4G0I 4G0M
4G0S 4G1I 4G2B 4G2C 4G2U 4G37 4G3B 4G3C 4G3V 4G4K 4G4L 4G4M 4G6Q 4G6U
4G7X 4G8K 4G9M 4G9S 4GBF 4GBO 4GBS 4GC1 4GCN 4GCS 4GE6 4GEK 4GGG 4GHB
4GIW 4GKC 4GKF 4GKG 4GKM 4GKP 4GL6 4GMN 4GNE 4GNI 4GNS 4GNU 4GOF 4GQ6
4GUC 4GVB 4GVF 4GVO 4GXB 4GXL 4GYT 4H05 4H0A 4H0C 4H0K 4H2D 4H4D 4H5I
4H5S 4H61 4H6Q 4H7X 4H87 4H8F 4H8M 4HAP 4HBQ 4HC8 4HCE 4HCI 4HDH 4HEH
4HEO 4HEQ 4HFS 4HG2 4HH6 4HHV 4HI7 4HI8 4HIA 4HIL 4HJD 4HJZ 4HKE 4HKG
4HL0 4HL2 4HLS 4HN9 4HNE 4HNH 4HP8 4HQZ 4HR1 4HRZ 4HS5 4HSS 4HT3 4HU5
4HW8 4HWU 4HWV 4HY4 4HYJ 4HYL 4HYN 4HZR 4I1K 4I1U 4I2Z 4I3G 4I4K 4I4O
4I5T 4I6P 4I6R 4I82 4I84 4I86 4I93 4IB2 4ID2 4ID3 4IGA 4IGW 4IHE 4IHZ
4IJ5 4IJR 4IJZ 4IKN 4ILO 4ILV 4IMQ 4IN0 4IN9 4INA 4INE 4INO 4INZ 4IO2
4IU3 4IUP 4IX3 4IXA 4IXJ 4IXN 4IYB 4IZB 4IZK 4J05 4J0X 4J1Y 4J2G 4J2K
4J3H 4J5R 4J6O 4J73 4J7Q 4J8B 4J8C 4J8E 4J8S 4J9C 4JBS 4JCH 4JCW 4JDE
4JE1 4JE6 4JEM 4JES 4JF3 4JGG 4JGI 4JGP 4JGW 4JGX 4JIX 4JJ0 4JJH 4JK8
4JLI 4JMD 4JN3 4JOQ 4JPQ 4JQT 4JR6 4JT4 4JUI 4JVU 4JX0 4JXB 4JXD 4JXE
4JY3 4JZP 4JZQ 4JZZ 4K00 4K02 4K05 4K0D 4K12 4K1C 4K28 4K2W 4K35 4K3L
4K4K 4K5A 4K6J 4K7J 4K7K 4K8Y 4K9Q 4KBM 4KCE 4KDX 4KED 4KF8 4KFS 4KFW
4KGD 4KGH 4KH6 4KH7 4KH9 4KHO 4KJM 4KJR 4KMD 4KN8 4KNC 4KNK 4KP2 4KPO
4KQR 4KRG 4KRT 4KT1 4KT3 4KTW 4KUJ 4KUN 4KV2 4KV9 4KWY 4KX8 4KYU 4KYX
4L00 4L0R 4L3N 4L3R 4L3T 4L4W 4L51 4L5G 4L68 4L6S 4L6U 4L7A 4L7X 4L8I
4L9O 4L9U 4LA2 4LAS 4LBA 4LCI 4LE7 4LEB 4LEC 4LIR 4LJI 4LJL 4LK2 4LLD
4LN2 4LN9 4LNL 4LOW 4LP4 4LPS 4LQ8 4LQC 4LQX 4LS4 4LUB 4LV5 4LW8 4LWK
4LXO 4LXQ 4M0H 4M1A 4M1B 4M1Q 4M3O 4M4D 4M8R 4M91 4MAC 4MAE 4MAK 4MAL
4MDU 4ME9 4MES 4MF9 4MG3 4MH1 4MHV 4MIK 4MIX 4MJ2 4MJD 4MJG 4MJK 4MLM
4MLZ 4MM2 4MMG 4MN5 4MN7 4MNW 4MO1 4MOV 4MPB 4MPM 4MPS 4MQB 4MR0 4MTL
4MUV 4MVE 4MW0 4MY6 4MYA 4MYP 4MYV 4MZ3 4MZJ 4MZZ 4N01 4N04 4N06 4N0K
4N0R 4N0V 4N3P 4N3V 4N4U 4N6A 4N6C 4N6F 4N7F 4N7W 4N82 4N8O 4N8Y 4N9Z
4NC7 4NCR 4NE2 4NET 4NFC 4NHB 4NIR 4NJH 4NKT 4NN2 4NOF 4NOH 4NPL 4NQ8
4NSD 4NSV 4NTG 4NTQ 4NWO 4NX8 4NZV 4O1J 4O1S 4O2H 4O2I 4O2T 4O3V 4O42
4O5P 4O71 4O7H 4O7J 4O8V 4O9D 4O9K 4O9S 4OEV 4OF6 4OFK 4OFQ 4OH7 4OHJ
4OK9 4OKE 4OLK 4OLT 4OM7 4OMV 4ON1 4ONW 4ONY 4OO0 4OO4 4OPM 4OTE 4OUC
4OVS 4OVT 4OWI 4OX6 4OZE 4P0J 4P0T 4P2I 4P2L 4P32 4P3F 4P5E 4P5F 4P5N
4P7B 4P7C 4P7O 4P93 4PAB 4PAS 4PE0 4PFZ 4PH8 4PI3 4PIC 4PID 4PIV 4PKC
4PM4 4PMK 4PMO 4PN6 4PO6 4POW 4PQ1 4PQ9 4PR3 4PSF 4PSR 4PTB 4PUI 4PVC
4PXW 4PXY 4PYS 4PZ7 4Q14 4Q2T 4Q3H 4Q4K 4Q53 4Q5G 4Q60 4Q69 4Q6J 4Q6U
4Q7E 4Q7O 4Q7Q 4Q82 4Q88 4Q8L 4Q9A 4Q9B 4Q9T 4Q9W 4QAK 4QAM 4QAN 4QAS
4QBN 4QC6 4QE0 4QF3 4QGO 4QHJ 4QI0 4QJB 4QJI 4QM9 4QMI 4QO2 4QPM 4QPV
4QSE 4QT9 4QUV 4QWO 4QYB 4R01 4R1K 4R1S 4R23 4R7X 4R80 4R86 4R8O 4R8R
4R9X 4RD8 4RHA 4RHP 4RK6 4RK9 4RPC 4RS2 4TKR 4TL1 4TMX 4TQL 4TR6 4TR7

4TY0 4U13 4U4I 4U99 4U9C 4UNU 4UON 4UOP 4UP0 4UQW 4UQY 4URG 4USQ 4UUU

A.2 Testing Set PDB codes

3L6U 1Y2K 2WW4 1IFQ 2WNS 2F06 2WNW 2OFK 2BKX 4PFY 4FAJ 4GAI 4O89 3WPU
1AE9 2PMQ 2GAX 3T7Y 1VQ0 1P1C 3RPD 3LLH 1O7Z 1PL3 3ANO 2DST 3T7H 1OB8
1EP3 4HU4 3MTX 4PMZ 1N7K 3SNX 4NV5 2P97 4EFP 4NV0 1R1G 1XFF 3LJD 3PNA
4DFR 4QXD 4C27 1OO0 3NCV 3N8H 4AR9 3ORE 3H8K 1MSP 3POA 4JG9 1HDH 4KS9
2VIF 3L0S 1SE0 4A27 4IPV 3QIV 1XXL 4PU7 4KW3 1F74 2VLI 4BUB 3W42 4I9F
3KWS 3MJE 4TVY 3LYN 2F22 1LBV 4R33 1AQL 4IC3 1NRJ 4KXQ 3D3M 3D3O 1DBX
4JTM 4ALY 1QOR 4IGQ 2Q7S 1RKU 1TV8 2AVT 1EX2 1XO1 3CWR 2V6X 3D3W 2VT8
1OF5 2W70 4P3H 2QAS 3GWO 3KAL 3C3K 3NUA 4LZF 3ZXO 1XX6 2BE3 3GMX 4HFM
3AOS 2XZO 3CGG 2W9M 4B9F 3NEH 3IEV 3O7I 1Y89 3VB8 3P8A 3RC8 2NNC 3QR3
3TJ8 4JIF 3B0Z 3MMH 4K70 4CMP 3QKX 1JY5 2RE3 4GER 4AB5 4LEV 4MS4 4N65
4V24 4M8K 3VOT 2EPG 1Q6O 3W20 2V0P 2WH6 1QWR 3U4V 2OKF 2D4Y 3AJG 1Y6Z
4BE3 4TPW 4GUD 1I58 4HQM 3C8I 1VKY 1S1D 4PQH 3L41 2OOQ 1JR8 3BOF 3PQC
3C8Z 4WSO 4EAE 2IHY 3HZ4 3MES 3KOJ 4LHK 4PAG 4TWC 3H3H 2EK0 4N6J 3RG9
2WT9 2QZT 4GD5 3CB7 2HXR 3TG9 2FFY 2AC7 3RT9 1TW0 4GRJ 4ACY 1DMU 1KW2
3FPQ 3F52 1L8D 3MOL 4KE7 3B2Y 2C3V 3UOR 4EBG 2I5E 256B 3OTX 4ART 4OSE
1KZQ 4PSE 4CGX 4BQM 4UVJ 4PZ4 2CD9 3FHU 2IKK 1G4M 3POJ 1D4T 3NMR 3PP2
4L9D 4M66 4L9A 3KST 2FAW 3BFQ 3LMB 3OJI 2JI5 2JBA 2IUT 3CSX 2JBH 2QTW
2AJ7 1PCX 4GT4 4O6I 1GGP 2IFT 3R8C 3HVA 3TJ4 1AOC 4OBE 1PSR 4RCM 1OAH
1AOZ 2YPD 4I6G 3D8P 4B7L 3NAR 4PZA 2O6K 4K92 3FFV 3BQA 2XH3 2IU5 1VR9