Response to Queries

I would like to thank my TAC examiner for her valuable comments. The following are my responses to the comments, in the same order in which they have been given.

1 A section mentioning some of the previous studies using statistical potentials for the prediction of protein-protein interactions with experimental validations has been added to the introduction section (Section 1.3 on page 7).

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).

2 This question has been addressed in the Results section (Section 3.1 on page 18).

Among the six ways of classifying protein interactions mentioned in the Introduction 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 Relevant information has been added in Section 2.1 (page 10)

In order to make accurate predictions using statistical methods, 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.

4 All typographical errors have been corrected

5 A more comprehensive description of Figure 3.5 is given in Section 4.2 in the Discussions section (page 28, 29)

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 several examples 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.

6 Comments have been added regarding this question in Section 4.1 (page 27)

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.

Apart from the responses to the TAC queries, the following sections were added after the initial submission of the Thesis.

In the results section, on page 24.

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 %.

In the discussions section, on page 29

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.