PAIRPRED: PARTNER-SPECIFIC PREDICTION OF INTERACTING RESIDUES IN PROTEIN COMPLEXES

Fayyaz-ul-Amir Afsar Minhas[1], Brian J. Geiss[2] and Asa Ben-Hur[1,]*

[1]Department of Computer Science, [2]Department of Microbiology, Immunology & Pathology, Colorado State University, USA. [*]asa@cs.colostate.edu.

We present a partner-specific Support Vector Machine (SVM) based predictor of interacting residues called PAIRpred (Partner Aware Interacting Residues PREDictor). PAIRpred is designed to predict whether two residues belonging to two members of a complex of unknown bound structure interact or not using either sequence information alone or in conjunction with features derived from the unbound structures of the two proteins. PAIRpred offers state of the art accuracy with an AUC score of 87.2.

INTRODUCTION

Partner-specific (or partner-aware) prediction of interacting residues in protein complexes refers to the problem of predicting whether a residue a in protein A interacts with residue b in protein B given that A and B form a complex. Most existing methods for predicting binding sites do not produce partner-specific predictions, and as a consequence, ignore the fact that the binding propensity of a residue in one protein is also dependent on the nature of its interaction partner in the target protein. Partner-specific predictions offer a finer level of detail in understanding protein interactions for in comparison to partner-independent predictions. Yet, only a handful of partner-specific methods are available, and their accuracy is rather low.

METHODS

We have used Docking Benchmark Dataset[1] (DBD) version 3.0, which provides both the bound and unbound structures of proteins in 124 non-redundant protein complexes. We define two residues belonging to two different proteins in a complex as interacting if the inter-atomic distance between them is less than 6 Å.

We extract both sequence and structure features at the residue level from the unbound structure of each protein. Structure based features include: Relative Accessible Surface Area (rASA), Residue Depth (RD), Protrusion Index (PI) and a novel feature called Half Sphere Amino Acid Composition (HSAAC). HSAAC captures amino acid composition in the neighborhood of a residue in the direction of the side chain of a residue and in the direction opposite to the side chain. Sequence based features are extracted from PSI-BLAST profiles. When only sequence information is available, we use rASA predictions from sequence.

In our problem, a classification example i is a pair of residues (a_i, b_i) from two proteins in a complex with an associated label. The use of an SVM for this problem requires a pairwise kernel[2] of the form $K\left((a_i, b_i)(a_j, b_j)\right)$ that needs to be constructed from a kernel over residue features.

Performance was evaluated using leave-one-complex-out cross validation with area under the ROC curve (AUC) as the quality metric for comparison with existing work.

RESULTS & CONCLUSIONS

As expected, a combination of sequence and structure features performs much better than sequence features alone (AUC of 87.2 vs. 80.9 for DBD 3.0). Even with sequence features alone, PAIRpred's performance is better than that of the best existing partner-specific predictor called PPiPP[3], which has an AUC of 72.9 with the same evaluation criteria and dataset. PAIRpred provides an AUC score of 87.0 for DBD 4.0 (176 complexes). An error analysis of PAIRpred reveals that the AUC of a complex is inversely related to the degree of conformational change that complex undergoes upon complex formation (see figure 1). This degradation is much less severe than that for PPiPP. We also found that PAIRpred's top predictions usually lie very close to a known interaction, e.g., in 50% of complexes, at least one of the top 7 predictions is within 2 residues of a true interaction.

We applied PAIRpred to the prediction of interacting residues between human ISG15 and NS1 from Influenza A virus[4], which is not part of DBD. The top-most prediction by PAIRpred is a true positive (ISG15:L10, NS1:L88), and 18 of the 34 true interactions lie within the top 100 predictions (PAIRpred AUC: 92.4, PPiPP AUC: 67.2). We verified the importance of residues L88 and F34 in NS1 for this interaction by replicating the mutagenesis experiment performed by Guan et al.[4] *in silico*. A significant decrease in PAIRpred's prediction score is observed when either residue is mutated to an Alanine.

We also investigated the specific binding of NS1 to ISG15 from human and non-human primates by observing location dependent differences in PAIRpred prediction scores for (human ISG15, NS1) and (mouse ISG15, NS1) interactions using predicted structures. Interesting locations on ISG15 that can cause such selective binding (in order of decreasing magnitude of change in predictions scores) are 76, 77, 72, 74 and 49. This strengthens the claim made in by Guan et al[4]. These results clearly illustrate the power of pairwise analysis possible with PAIRpred.

In the future, we plan on adding features to capture shape complementarity, co-evolution, and flexibility to improve predictions even further.

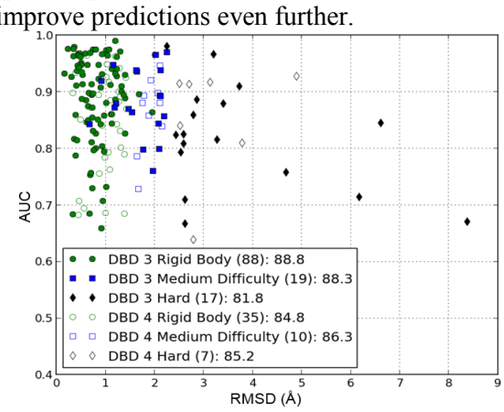

***FIGURE 1**. AUC score vs. the conformational change at the interface for different complexes in DBD 3.0 and 4.0.*

REFERENCES

1. Hwang, H., et al. Proteins: Structure, Function & Bioinformatics **78**, 3111–3114 (2010).
2. Brunner, C., et al. *Journal Of Machine Learning Research* **13**, 2279-2292 (2012).
3. Ahmad, S., & Mizuguchi, K. PloS one, **6(12)**, e29104, (2011).
4. Guan, R., et al. Proceedings of the National Academy of Sciences **108**, 13468–13473 (2011).