Classifying permanent and transient protein interactions

Samatha Kottha and Michael Schroeder
Biotec and ZIH, TU Dresden, {samatha.kottha,michael.schroeder}@tu-dresden.de

Abstract: Currently much research is devoted to the characterization and classification of transient and permanent protein-protein interactions. From the literature, we take data sets consisting of 161 permanent (65 homodimers, 96 heterodimers) and 242 transient interactions. We collect over 300 interface attributes relating to size, physiochemical properties, interaction propensities, and secondary structure elements.

Our major discovery is a surprisingly simple relationship not yet reported in the literature: interactions with the same molecular weight or very big interfaces are permanent and otherwise transient. We train a support vector machine and achieve the following results: Molecular weight difference alone achieves 80% success rate. Together with the size of the buried surface the success rate improves to 89%. Adding water at the interface and the number of hydrophobic contacts we achieve a success rate of 97%.

1 Introduction

Protein-protein interactions are fundamental to most cellular processes such as recognition of foreign molecules, host response to infection, transport machinery across various biological membranes, packaging of chromatin, the network of sub-membrane filaments, muscle contraction, signal transduction, and regulation of gene expression. Aberrant or lack of certain protein-protein interactions leads to the neurological disorders such as Alzheimer's disease. The forces that are responsible for these interactions include electrostatic forces, hydrogen bonds, van der Waals forces, and hydrophobic effects. The understanding of these interactions will provide the clues to their biological function. Several groups have been analyzing protein-protein interactions by categorizing them as homo-complexes, homo-oligomers, hetero-complexes, hetero-oligomers, obligate and non-obligate complexes, transient and permanent complexes, folding type and recognition type complexes (10; 14; 17; 13; 5; 2; 3; 1; 4).

A fundamental distinction in the nature of protein-protein interfaces is the separation into permanent and transient interfaces which are also called two-state and three-state complexes, respectively (21). Folding and binding are inseparable for two-state complexes. However, in case of three-state complexes, proteins fold independently and then bind. It is widely believed that permanent interactions can occur in homomers and heteromers, and transient interactions mostly in heteromers. However, Nooreen et al. and and Schreiber et al. collected 13 experimentally validated homodimers with transient interactions (15; 20).

Several studies analyze protein-protein interactions using interface properties like size, shape, residue and atomic contact propensities, hydrophobicity, hydrogen bonds, and sec-

ondary structure (10; 17; 13; 15; 5; 2; 3). Not a single feature analyzed in these studies differentiates permanent interactions from transient interactions or vice versa. As Nooreen and Thornton point out (15), it is difficult to discriminate, especially the strong transient from permanent interactions or the weak permanent from transient interactions. Mintseris and Weng propose atomic contact vectors to tackle this difficult problem and achieve a 91% success rate (13). However, they use 171 features to classify 340 interactions.

In this paper, we derive a data set of transient and permanent interactions from literature and initially capture over 300 attributes for the interfaces. We analyse the most predictive attributes in detail and show that the four attributes of molecular weight difference of the chains, size of the buried surface, number of water molecules at the interface, and number of hydrophobic contacts achieve a classification success rate of 97% - to our knowledge the best success rate reported. Moreover, the difference in molecular weight of the two interacting chains is the single most predictive attribute, which achieves a success rate of 80% on its own. This is particularly remarkable, as it can be derived from sequence information only.

2 Materials and Methods

We use five datasets introduced in (13; 20; 15; 1; 4). Even though all these datasets are generated by applying stringent criteria, some of them are contradicting each other. For example, the transferase 1d09 A:B is classified as permanent in (13) and transient in (20) and the toxin 1bun A:B is classified as permanent in (4) and as transient in (20). We carefully examine all the interactions with contradicting classification and label them according to the literature. Overall, only 9 out of over 400 interactions are affected.

To obtain a non-redundant dataset, all the interacting chains' sequences are clustered using BLASTCLUST (`ftp.ncbi.nih.gov/blast/`). The interactions which have both interacting chains with $\geq$25% sequence identity are clustered together and one interaction from each cluster is selected. As a result, we have 161 permanent and 242 transient interactions in our dataset. For these two classes, it is important to cover both homo- and heterodimers. This is indeed the case for our dataset, as the breakdown below shows:

	transient	permanent	sum
homo	13	65	81
hetero	229	96	322
sum	242	161	403

Feature Collection. We collect over 300 attributes about the interacting chains, residues, interfaces, and secondary structure elements and categorize them into the following four sets:

Size. Number of residues per chain, molecular weight and Accessible Surface Area (ASA) of each interacting chain, molecular weight difference, interface area ΔASA, number of residues at interface compared to individual chains, number of residues at interface compared to total residues, contact surface area, contact volume, total number of residue con-

tacts, number of residues at interface.

Physiochemical properties. Isoelectric Point of each interacting chain, hydrophobicity of the interface, normalized hydrophobicity by the interface size, hydrogen bonds, salt bridges, disulfide bonds and hydrogen bonds per 100 ASA in interface, water at interface, interaction strength, number of aromatic, charged, polar, hydrophobic, hydrophilic, hydro-neutral residues in interface, and the contacting residues pairs properties like aromatic_aromatic etc.

Amino acid propensities. Counts of residues A,C, . . . , Y and contacts A-A, A-C, . . . , Y-Y at interface.

Secondary structure elements. The absolute and normalized counts of interacting residues' secondary structure elements (helix, strand, coil, and turn).

The above attributes range from very general attributes like the number of hydrophobic-hydrophobic contacts to very special ones like the individual residue pair propensities including all pairs of hydrophobic residues, which appears redundant. However, the objective behind collection of both specific and general attributes is that all of them may play a role. If permanent interactions have large interfaces, there should be hydrophobic cores and hence hydrophobic-hydrophobic contacts could be important. Residue propensities vary strongly for different pairs and hence individual counts of residue-residue interactions may also be important. In the end, all of these attributes are collected, so that the algorithm can select the most predictive ones.

The molecular weight and the isoelectric points are calculated using the bioperl module with the EMBOSS value set. Accessible surface areas and ΔASA are determined using NACCESS (`wolf.bms.umist.ac.uk/naccess/`). The contact surface area and volume are derived by computing convex hulls of interaction interfaces (7). A novel, experimentally determined Stephen-White hydrophobicity scale (9) is used to calculate hydrophobicity. It does not lead to different results compared to the Kyte–Doolittle scale (12). The number of hydrophobic contacts is computed at residue level (F, A, I, M, L, V, C are hydrophobic) and if a residue participates in several hydrophobic-hydrophobic contacts, all of them are counted. While hydrophobic-hydrophobic contacts are a count, hydrophobicity is the sum of all interface residues' hydrophobicity according to (9).

Different types of bonds between two chains are determined using WHATIF (22). The interaction strength is calculated based on the bonds formed between two chains. The bond strength is measured by the amount of energy required to break the bond. Although the strength of a bond depends on the environment, a covalent bond is nearly 90 times stronger than a single hydrogen bond in water. Therefore, we consider disulfide bridges with a strength of 90, salt bridges with 3 and hydrogen bonds with 1.

Water at the interface is the number of water molecules which are $\leq 5\mathring{A}$ distance to both interacting chains.

The absolute and normalized counts of all amino acids in the interface are considered along with the contacting residue pairs. The two residues are said to be in contact if their atoms are within or equal to 5 $\mathring{A}$ distance.

Using STRIDE (8), the secondary structure elements of the interacting residue pairs are

determined. We consider both the absolute and the normalized counts.

Algorithms. We have 161 instances of permanent and 242 instances of transient interactions each with a vector of over 300 attributes in the training set. To identify the most relevant attributes for the classification task, we use relief estimation (11), which ranks the most predictive features independent of any learning algorithm. For the classification of permanent and transient interactions we use decision trees (C4.5) (18) to derive specific rules and support vector machines (SVM) to carry out an overall classification. For the SVM we use the LIBSVM library (6). We use a Radial Basis Function (RBF) kernel to map data into a higher dimensional space. We perform a grid search on internal parameters C and γ using cross validation and the value set with the best cross validation accuracy is picked. To avoid the problem of overfitting we use stratified 10-fold cross validation for both, the SVM and C4.5 algorithms.

Evaluation. In the results section we apply support vector machines to compute the overall success rate for a set of attributes, as well as sensitivity and specificity of built model and decision trees to derive intuitive classification rules. For these rules we report accuracy and support. Accuracy assesses how good the rule's classification is and support assesses to how many examples in the data set the rule applies.

The success rate is defined as the number of correctly predicted interfaces divided by all interactions: Success rate = Correct predictions / All interactions i.e. the success rate assesses the overall percentage of correct predictions. The sensitivity = TP / TP+FN and the specificity = TN / TN+FP .

To define the accuracy and support of a rule, let us denote the correct predictions of the rule as TP (True Positives) and the incorrect predictions as FP (False Positives). Then, the accuracy of a rule's prediction is defined as the percentage of correctly predicted examples for the rule: Accuracy = TP / TP+FP. The support indicates how general a rule is, i.e. to how much of the data it applies to: Support = TP+FP / All interactions. Generally, we wish to define rules with high accuracy and support.

3 Results

Molecular weight difference achieves 80% classification success rate. The ten most highly predictive attributes (in descending order) are molecular weight difference, ΔASA, hydrophobic-polar contacts, hydrophobic-hydrophobic contacts, water at interface, no. alanine-lysine contacts, no. isoleucine-tyrosine contacts, no. helix-helix contacts, no. methionine at interface, and no. leucine-serine contacts. The difference in molecular weights is the most outstanding feature separating permanent from transient interactions - both for homo- and heterodimers. Consider the scatterplot in Fig. 1a. Most permanent interactions are located on or close to the diagonal, i.e. both chains are of (nearly) equal molecular weight. This is not surprising for homodimers, but the majority (96 out of 161) of permanent interactions in the data set are actually heterodimers. Using a support vector machine (see materials and methods), the molecular weight difference alone can classify 80% of interactions correctly with a sensitivity of 71% and specificity of 86%. A closer exami-

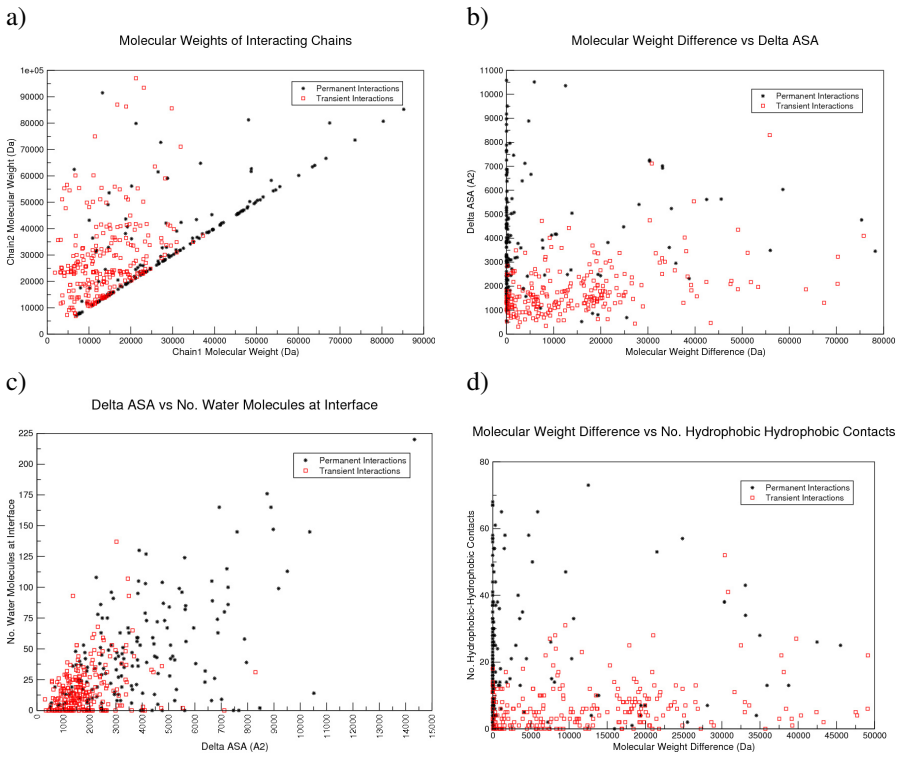

Figure 1: **a)** Scatterplot for molecular weight difference of interacting chains. Permanent interactions are close to the diagonal as they have similar weights. This is particularly remarkable as 96 out of 161 permanent interactions are heterodimers. Transient interactions mostly involve a lighter and a heavier chain. **b)** Scatterplot for molecular weight difference of interacting chains against ΔASA. Permanent complexes loose more surface accessible surface area upon complexation than the transient ones. Permanent interactions with more than 5 kDa molecular weight difference have mostly large interface of greater than 2000 $\mathring{A}^2$. **c)** Scatterplot for absolute counts of water at the interface plotted against ΔASA. There is some correlation (0.486) between the two attributes. **d)** Scatterplot for the number of hydrophobic contacts plotted against molecular weight difference. The plot shows that permanent interfaces have more hydrophobic contacts and are therefore a useful additional feature in the classification task.

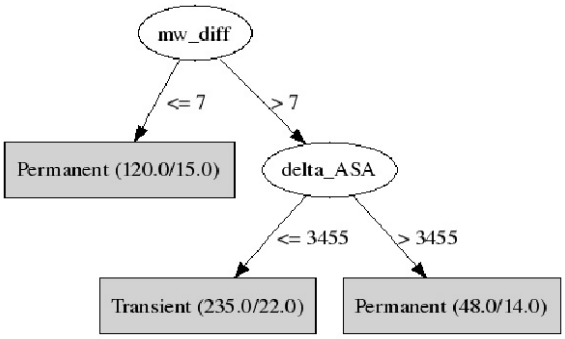

Figure 2: Decision tree with molecular weight difference and ΔASA. The boxes contain the predicted class. The total number of interactions and the number of incorrectly classified examples are in brackets. The ovals are the decision points defined by the algorithm. It identifies more or less than 7 Da molecular weight difference as main separating feature for transient and permanent interactions. It also automatically separates very big interfaces from other interfaces.

nation of the distribution also reveals that the interaction between chains with less than 7 Da weight difference are mostly permanent (88% accuracy and 30% support), while interactions between chains with more than 10 kDa molecular weight difference are usually transient (83% accuracy and 40% support).

The interesting aspect of these two rules is that they do not require any structural information and as only 65 out of 161 permanent interactions are homodimers.

Molecular weight difference and buried surface achieve 89% classification success rate.. As stated above, ΔASA is the second most predictive feature. The scatterplot in Fig. 1b shows that permanent complexes loose more solvent accessible surface area than transient complexes. In particular, nearly all permanent interactions with more than 5 kDa molecular weight difference have interfaces bigger than 2000 $\mathring{A}^2$, while most transient interactions have smaller interfaces.

To quantify this observation, we trained a support vector machine (see materials and methods) for molecular weight difference and ΔASA and achieved a classification success rate of 89% (sensitivity 84% and specificity 93%). In order to capture intuitive rules for this classification task, we also generated a decision tree (see materials and methods) shown in Fig. 2. The decision tree procedure automatically derives cut-off values. For ΔASA, it distinguishes very big ($3455\mathring{A}^2$) or not and for molecular weight differences small (≤ 7 Da) or not. Overall, the decision tree consists of three rules as shown in Fig. 3, which can be summarized as follows: Interactions with very small molecular weight difference (≤ 7 Da) or very big interfaces ($\geq 3455\mathring{A}^2\Delta$ASA) are permanent, otherwise they are transient. This single rule on its own achieves accuracy of 87% and a support of 100%.

No	Weight Difference		ΔASA		Class.	Acc.	Supp.
1	Very small	$\leq$7 Da	Does not matter		Permanent	88	30
2	Not small	>7 Da	Very big	$> 3455\mathring{A}^2$	Permanent	71	12
3	Not small	>7 Da	Not very big	$\leq 3455\mathring{A}^2$	Transient	91	58

Figure 3: Classification rules derived from a decision tree with their accuracy (Acc.) and support (Supp.). Rule 1 and 3 have the biggest support, i.e. they capture a large portion of the data set. Rule 1 states that if the molecular weight difference is very small the interaction is permanent. Rule 3 states that a difference in molecular weights and an interface, which are not very big, imply a transient interaction.

Adding hydrophobic contacts and water achieves 97% classification success rate.. To further improve the classification results we added two more features: water at the interface, which is a feature for transient interfaces (16), and the number of hydrophobic contacts, which is important for permanent interactions. As stated above, the number of hydrophobic-polar contacts is the third most predictive feature. However, molecular weight difference, ΔASA, water at the interface and hydrophobic-hydrophilic contacts are performing slightly worse (96.03%) than hydrophobic-hydrophobic contacts (97.27%). Both features achieve roughly similar results as they are highly correlated (0.8), but hydrophobic-hydrophobic contacts are slightly less correlated to water at the interface (0.35) than hydrophobic-hydrophilic contacts are (0.43). It is also established that large interfaces have hydrophobic cores (see e.g. (10)), so that the better performance of hydrophobic-hydrophobic contacts and its role in large interfaces led us to choose it over hydrophobic-hydrophilic contacts. So, the attributes molecular weight difference, ΔASA, water at the interface, and hydrophobic-hydrophobic contacts could classify 97% of interactions (sensitivity 95% and specificity 99%) correctly.

Although the absolute number of water molecules at the interface correlates to some degree (0.486) with the interface size ΔASA (see Fig. 1c), it improves the classification success rate as shown below. As an additional feature relating to the role of water, we also checked water mediated contacts. These are contacts between two residues from different interacting chains, which are in contact through a single water molecule but not in direct contact ($> 5\mathring{A}$ distance).

However, water-mediated contacts do not play a role in this classification task, which is consistent with Rodier et al. (19), who found that water density at homodimeric interfaces and protein-protein complexes is the same. Note, that the number of water molecules at the interface and the number of water-mediated contacts are not highly correlated (only 0.424).

Besides water, we investigated hydrophobic contacts as it is widely believed that permanent interfaces are more hydrophobic than transient ones. For the analyses of hydrophobicity we used the Stephen-White hydrophobicity scale (9). Fig. 1d shows that the feature of hydrophobic contacts separates transient and permanent interfaces well.

As a final step, we trained a support vector machine (see materials and methods) with the four attributes molecular weight difference, ΔASA, number of water molecules at the

Molecular weight difference

	transient	permanent	sum
homo	0/13	65/65	65/78
hetero	207/229	50/96	257/325
sum	207/242	115/161	322/403

Molecular weight difference, ΔASA

	transient	permanent	sum
homo	8/13	58/65	66/78
hetero	217/229	77/96	257/325
sum	225/242	135/161	360/403

Weight diff., ΔASA, hydrophobic-hydrophobic contacts, water at interface

	transient	permanent	sum
homo	11/13	61/65	72/78
hetero	229/229	91/96	320/325
sum	240/242	152/161	392/403

Figure 4: Breakdown of correctly classified protein-protein interactions for transient homodimers, transient heterodimers, permanent homodimers, and permanent heterodimers. The overall success rates achieved are consistent with all these subclasses. Molecular weight difference alone classifies permanent homodimers and transient heterodimers very well and permanent heterodimers reasonably well. Adding the other three attributes, success rates for all these subclasses are in the 90s.

interface (within $5\mathring{A}$), and number of hydrophobic contacts. We achieve a classification success rate of 97% for over 400 interactions in the data sets taken from (13; 20; 15; 1; 4).

Heterodimers vs. Homodimers and Transient vs. Permanent.. To test whether the above results also hold for heterodimers only, we considered 96 transient and 96 permanent heterodimer interactions. Thus, a random predictor achieves an expected success rate of 50%. The four attributes considered above perform as follows: Molecular weight difference alone achieves 73%. Molecular weight difference and delta ASA achieve 84%. Molecular weight difference, delta ASA, water at the interface and hydrophobic contacts achieve 88%. These results are in line with the ones for hetero- and homodimers reported above, in particular as homodimer interactions are not always permanent and as our dataset contains 13 such transient homodimer interactions, which are difficult to classify.

Indeed, it is an interesting questions how the success rates for the classification of the full 403 interactions break down between the classes of homo-transient, hetero-transient, homo-permanent, and hetero-permanent. Figure 4 shows three tables with these success rates for the three combinations of the four attributes. The first table shows that molecular weight difference alone classifies permanent homodimers and transient heterodimers very well and permanent heterodimers reasonably well. It does not handle the transient homodimers well. Adding ΔASA, the success rates for transient homodimers and permanent heterodimers greatly increase. Finally, the third table in Fig. 4 shows that the overall success rate of 97% is consistently achieved in all subclasses of transient homodimers (85%), transient heterodimers (100%), permanent homodimers (94%), and permanent heterodimers

(95%). Also, homo- and heterodimers achieve consistent success rates (92% and 99%, respectively) and transient and permanent interactions, too (99% and 94%, respectively).

4 Conclusion

There is great interest in characterizing and classifying protein interactions as transient or permanent (10; 14; 17; 13; 5; 2; 3; 1; 4). In particular, Mintseris and Weng achieve 91% prediction success rate using their atomic contact model with 171 features to classify 340 interfaces (13).

In this paper, we have assembled a data set consisting of 161 permanent and 242 transient interactions taken from the literature (13; 20; 15; 1; 4). For the interfaces we collected over 300 attributes relating to the size, physiochemical properties, residue propensities, and secondary structure elements.

Based on these data, we made a surprisingly simple discovery not yet reported in the literature: The difference in molecular weight between the interacting chains is the single most informative feature to distinguish transient from permanent interactions. Using this feature, 80% of interactions can be correctly classified. This is particularly important, as the molecular weight can be derived from sequence alone, so that no structural data is needed. Together with attributes known to play a role such as the size of the solvent accessible surface area lost upon complex formation, we can formulate the simple rule that interactions with small molecular weight difference or very big interfaces are permanent and otherwise they are transient. This simple rule achieves 87% success rate.

Finally, we added two more attributes known to be important, namely water at the interface and number of hydrophobicity contacts. Overall, we achieve a classification success rate of 97%, thus improving on other results previously published.

As next step, we wish to underpin our key insight that permanent interactions - like lasting marriages - require equal partners by developing physical models of the protein masses and moments, which can shed further light on this observation.

Acknowledgment: We gratefully acknowledge support of the EFRE Project CODI. We would like to thank Wan Kyu Kim, Joan Teyra, Gihan Dawelbait and Christoph Winter for helpful discussions and comments.

References

[1] S Ansari and V Helms. Statistical analysis of predominantly transient protein-protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 61(2):344–355, November 2005.

[2] R P Bahadur, P Chakrabarti, F Rodier, and J Janin. Dissecting subunit interfaces in homodimeric proteins. *PROTEINS: Structure, Function, and Genetics*, 53(3):708–719, November 2003.

[3] R P Bahadur, P Chakrabarti, R Rodier, and J Janin. A dissection of specific and non-specific protein-protein interfaces. *Journal of Molecular Biology*, 336(4):943–955, February 2004.

[4] J R Bradford and D R Westhead. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, 21(8):1487–1494, April 2005.

[5] P Chakrabarti and J Janin. Dissecting protein-protein recognition sites. *PROTEINS: Structure, Function, and Genetics*, 47(3):334–343, May 2002.

[6] C C Chang and C J Lin. *LIBSVM : A Library for Support Vector Machines (Version 2.6)*, 2004. www.csie.ntu.edu.tw/ cjlin/libsvm.

[7] P Dafas, D Bolser, J Gomoluch, J Park, and M Schroeder. Using convex hulls to extract interaction interfaces from known structures. *Bioinformatics*, 20(10):1486–1490, July 2004.

[8] M Heinig and D Frishman. Stride: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research*, 32:W500–502, July 2004.

[9] T Hessa, H Kim, K Bihlmaier, C Lundin, J Boekel, H Andersson, I Nilsson, S H White, and G von Heijne. Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, 433(7024):377–381, January 2005.

[10] S Jones and J M Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences USA*, 93(1):13–20, January 1996.

[11] I Kononenko. Estimating attributes: analysis and extensions of relief. In F Bergadano and L De Raedt, editors, *Proceedings of Machine Learning: ECML-94*, pages 171–182. Springer Verlag, 1994.

[12] J Kyte and R F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, May 1982.

[13] J Mintseris and Z Weng. Atomic contact vectors in protein-protein recognition. *Proteins*, 53(3):629–639, November 2003.

[14] I M A Nooren and J M Thornton. Diversity of protein-protein interactions. *The EMBO Journal*, 22(14):3486–3492, July 2003.

[15] I M A Nooren and J M Thornton. Structural characterisation and functional significance of transient protein-protein interactions. *Jorunal of Molecular Biology*, 325(5):991–1018, January 2003.

[16] R Nussinov, C J Tsai, and D Xu. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Engineering*, 10(9):999–1012, September 1997.

[17] Y Ofran and B Rost. Analyzing six types of protein-protein interfaces. *Journal of Molecular Biology*, 325(2):377–387, January 2003.

[18] J R Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann, San Francisco, 1993.

[19] F Rodier, R P Bahadur, P Chakrabarti, and J Janin. Hydration of proteinprotein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 60(1):36–45, July 2005.

[20] G Schreiber, R Raz, and H Neuvirth. Promate: A structure based prediction program to identify the location of protein-protein bindings. *Journal of Molecular Biology*, 338(1):181–199, April 2004.

[21] C J Tsai, D Xu, and R Nussinov. Protein folding via binding and vice versa. *Folding & Design*, 3(4):71–80, 1998.

[22] G Vriend. What if: A molecular modeling and drug design program. *Journal of Molecular Graphics*, 8(1):52–56, March 1990.