Discrimination of permanent and transient heterodimers: The real challenge in the classification of protein-protein interactions

Eva Vennmann, Nadine Schneider, Matthias Rarey
Center for Bioinformatics, University of Hamburg

Abstract: The classification of protein-protein interactions into permanent, transient and crystal artifacts is still an unsolved problem. Current methods are often either too inaccurate or more complicated than necessary. Due to this we have calculated several descriptors concerning the characteristics of protein-protein interactions and analyzed diverse variations with a support vector machine. Using only two descriptors – the hydrophobicity of the interface and the proportion of the interface ratios – we achieved an accuracy of 92.5 % in our training dataset. Applying our generated model on an independent test set we obtained an accuracy of 77.1 %.
Analyzing the composition of available datasets, it becomes clear, that they are often biased by large portions of permanent binding homodimers. The real challenge, however, is the discrimination of permanent from transient heterodimers. Due to this, we have constructed a dataset containing only permanent and transient heterodimers and applied our most promising descriptors. The usage of interface hydrophobicity as the only descriptor for the discrimination revealed the best accuracy of 80.2 %.

Introduction

Protein-Protein interactions play a significant role in all biochemical pathways and signaling cascades. It is important to differentiate between biological complexes (permanent and transient) and crystal artifacts. In addition, it is of great interest to discriminate between permanent and transient protein-protein complexes since latter are potential drug targets [1]. Consequently, it is of particular importance to comprehend the mechanisms of these interactions. One aspect of these protein-protein complexes is their kind of interaction. Analyzing a single X-ray structure, they can be divided into permanent and transient complexes as well as crystal artifacts. Permanent protein-protein complexes, also called folding complexes or two-state complexes, are characterized by their high stability, since they cannot exist in their monomeric state [2 - 3]. In contrary, transient complexes, also called recognition complexes or three-state complexes, are stable in their monomeric form and able to fulfill biological functions in this state as well as in the dimeric state [1 – 3]. The third interaction type, crystal artifacts, are artificially formed upon crystallization [4, 5].
These three different types of protein-protein complexes have been investigated theoretically for years. Several characteristics have been determined for the diverse complex types. The amino acid composition of the protein-protein interface displays one characteristic. Transient interactions have a high percentage of polar amino acids, permanent complexes are in contrast more hydrophobic whereas crystal artifacts have a great contribution of charged residues [6 - 9]. Permanent protein-protein interactions appear in homomers as well as in heteromers while only a few transient interactions are known to be homodimeric [10, 11]. The interface of permanent interactions is often large and highly interfering, whereas transient interactions are generally smaller and more planar [3, 10]. Crystal contact interfaces tend to be smaller and less tightly packed than the former interactions [4, 5].
During the last years many computational methods have been evolved for the classification of permanent, transient and crystal protein-protein interactions. The first method was developed by Ponstingl et al. [12] and differentiated monomers from homodimers. The usage of the solvent accessible surface area as a descriptor resulted in a correct classification of 85 % of the 96 monomeric and 76 homodimeric structures, whereas the atom pair potential, a measurement of intermolecular contacts in the crystal, reached an accuracy of 87.5 %. Another classification method for the differentiation of monomers and homodimers as well as transient and permanent complexes was developed by Mintseris and Weng [13]. They used a kernel discriminant analysis with an atomic contact vector containing 171 types of atom pairs for the classification. An accuracy of 93 % was

achieved on the data set constructed by Ponstingl et al. [12]. Zhu et al. [14] trained a support vector machine (SVM), using a test set of 243 permanent and transient complexes as well as crystal artifacts. As descriptors they used the interface area, interface area ratio and amino acid composition and achieved a total classification accuracy of 91.8 %. They obtained an accuracy of 80 % for the discrimination of biological and crystal complexes on a test set, consisting of 380 complexes taken from Bahadur et al. [15]. Kottha and Schroeder [16] achieved a very high accuracy of 97 % for the differentiation of permanent and transient protein-protein interactions. Their test set consists of 403 permanent and transient protein-protein complexes. They used four descriptors, the molecular weight, the accessible surface area, hydrophobic contacts and the number of crystallographically determined water molecules within the interface. The usage of the molecular weight as the only descriptor achieved an accuracy of 80 %, without the need of crystal structures. Additionally they have analyzed the performance of their method for discriminating permanent and transient heterodimeric interactions. Here, they obtained an accuracy of 73 % using only the molecular weight. Table 1 contains a comparison of additional classification methods.

Table 1: Summary of different methods separating protein-protein interactions

Study	Classification Criteria	Number of Descriptors	Method of Classification	Accuracy [%]	Dataset size
Bahadur et al. [15]	M - H	22	statistical methods	~ 94	310
Kottha et al. [16]	T - P	4	SVM	97	403
Bradford et al. [19]	T - P	14	SVM	~ 75.5	180
Ofran et al. [20]	6 interaction types*	20	statistical methods	~ 77.2	1812
Zhu et al. [14]	T - P - C	22	SVM	91.8	243
Block et al. [18]	B - C T - P	26 62	DT + GA	94.8 93.6	517
Ponstingl et al. [12]	M - H	153	pair frequency SF	88.9	172
Park et al. [21]	4 interaction types#	157	ARBC	47.6 - 99.9‡	147
Mintseris et al. [13]	M - H T - P	171	KDA	93 91	517
Liu et al. [17]	B - C T - P	213	SVM	96.7 87.9	243
Rueda et al. [22]	M - H T - P	324	LDR	88.7 80.3	461

B - biological complexes
P - permanent complexes
T - transient complexes
C - crystal complexes
H - homodimers
M - monomers
DT - decision tree
GA - genetic algorithm
KDA - kernel discriminant analysis
SF - scoring function
LDR - linear dimensionality reduction
ARBC - association rule based classification

* Interfaces within one structural domain (1), between different domains of one chain (2), between permanently (3), or transiently (4) interacting identical chains, between permanently (5) or transiently (6) interacting different chains

Enzyme inhibitor comnplexes, non enzyme-inhibitors, hetero-oligomers, homo-oligomers

‡ Accuracies for different classification methods DT, Random Forest, K Nearest Neighbor, SVM, Naïve Bayes

In this paper we introduce a new classification model for the three different types of protein-protein interactions. Furthermore, a detailed analysis of available datasets used for training of the classification methods mentioned above is performed. We have tested several descriptors and combinations of them using a SVM to generate a novel classification model. Finally, we achieved a total accuracy of 92.5 % for the separation of transient, permanent and crystal protein-protein

complexes using only two descriptors, the hydrophobicity and the proportion of the interface ratios. The dataset contained 254 complexes, which we have taken from the dataset used in the study of Zhu et al. [14]. We obtained an accuracy of 95.3 % for the discrimination of crystal artifacts and biological complexes. We classify 89.2 % correct considering the determination of permanent and transient interactions. Existing methods already achieved satisfying results considering the separation of biological interactions and crystal artifacts [14, 17, 18]. However, the identification of transient and permanent complexes is the more severe problem. A detailed analysis of the available data shows that the discrimination of heterodimers in permanent and transient interactions is the real challenge in this field.

Methods

Descriptors for classification

We have calculated several diverse descriptors based on the characteristics described in the theoretical studies [3 - 11] of protein-protein interactions as well as descriptors used in former methods [14, 16, 19]. Our descriptors consider the energetics of the different protein-protein interactions, geometric properties of the diverse complexes as well as the amino-acid composition of the interfaces. Most of the descriptors were calculated on the interface region (6.5 Å around the atoms comprising the interface). All descriptors are calculated using in-house tools and software.

Energetics

Concerning the stability of the different interactions, we have calculated the free energy of binding ($\Delta G_{Binding}$) using the HYDE scoring function [23] which is developed for the estimation of protein-ligand complexes. There is no difference in the treatment of protein and ligand atoms in HYDE, consequently HYDE is also well suited for estimating the strength of protein-protein interactions. HYDE is based on HYdration and DEsolvation terms to estimate the energy of a complex. It describes consistently the favorable energy of hydrogen bonds and the hydrophobic effect, as well as destabilizing contributions to the binding energy arising from the desolvation of hydrophilic atoms in the interface. The descriptors we have calculated using the HYDE scoring function are: the ΔG_{HYDE} energy which results upon binding of the protein subunits and the energy emerging only of the hydrophobic effect ($\Delta G_{Hydrophobic}$) between the two protein subunits.

Geometry and surface area

We analyzed the interface geometry of the diverse protein-protein interactions since, as mentioned above, planarity and interface size are important characteristics of different interaction types. The planarity of the interfaces was calculated as root mean square deviation of the interface C_α-atoms from the interface plane. This plane was constructed by performing a principal component analysis of the interface C_α-atom coordinates. Using the resulting principal components we are also able to calculate the structure tensors of the interface, a measurement which stems from the image processing field [24]. These structural tensors allow us to draw conclusions about the shape of the interface.

We calculated the molecular surface area (Connolly surface) of the complex interfaces and also the ratio of the interface area compared to the surface of the minimal subunit. This was done analogous to descriptors Zhu et al.[14] used in their study. Additionally, we estimated the ratio of each interface area to the surface area of the respective protein subunit to get the exact proportion of the interface area. Further, we extracted only the apolar surface area within the interface and compared this to the apolar surface area of the protein-protein complex.

Amino acid composition

We investigated the amino acid composition of the interfaces of the different protein-protein interactions and calculated this composition for each amino acid type, leading to 20 different descriptors (for more details see [14]). In order to get a clearer signal we grouped amino acids with similar characteristics together: polar (C, H, N, Q, S, T, W, Y), hydrophobic (A, F, G, I, L, M, P, V) and charged (D, E, K, R). The composition of the interface is calculated again using this raw classification of amino acids.

Interface/subunit symmetry

During our study we observed that the symmetry of the interfaces is also an important characteristic to discriminate between the diverse interaction types. Consequently, we developed descriptors which exactly consider the symmetry aspect of protein-protein interactions. Firstly, since the HYDE score is an atom-based score we were able to calculate the ratio of the hydrophobic effect of one subunit compared to the other subunit of the protein-protein interaction. A second symmetry descriptor is the quotient of the interface area ratios (IF-quotient) which we have mentioned above. Thirdly, we have computed the difference of the molecular weights of each protein subunit. This descriptor was also used in the study of Kottha et al. [16], however, they did not mention using this as a description of interface/subunit symmetry.

Classification method and validation measures

For the automatic classification of the three different types of protein-protein interactions we employed a support vector machine (SVM). We used two types of SVM classifiers: a multi-class SVM to discriminate the three interaction types in one step and a two-stage SVM to separate firstly the crystal artifacts from the biological meaningful complexes and in a second step the transient from the permanent interactions. We trained both with several combinations of the developed descriptors. We used the SVM implementation in the R package *e1071* [25].

The discrimination performance of our generated model was evaluated by calculating the accuracy. The accuracy is defined as the ratio of correctly predicted complexes divided by the sum of all predicted complexes.

Dataset retrieval

We extracted the protein-protein complexes to construct our dataset from several resources [9, 10, 14, 15, 26, 27] and extracted the structures from the PDB [28]. During the data collection it strikes out that some of the protein-protein interactions were annotated as transient in one dataset while they are classified as permanent in others. These complexes were excluded from our dataset (< 1 % of all complexes). Furthermore, complexes with more than two interacting chains were also removed from the dataset. Finally we attained a dataset which consists of 133 permanent, 121 transient and 152 crystal artifacts. Out of these data we have constructed different smaller datasets for manifold analysis purposes. We also inspected our dataset concerning the homo- and heterodimeric nature of the protein-protein complexes. Here, we used an in-house alignment tool (based on the edit distance algorithm by E. Ukkonen [29]) to align the two protein chains of the different interaction types. We calculated the alignment considering the whole chains as well as using only the amino acids which comprise the interface region. Table 2 gives an overview of the distribution of the different interactions types into the homo- and heterodimeric classification.

	Whole Chain		Interface Region	
	Homodimers	**Heterodimers**	**Homodimeric**	**Heterodimeric**
Permanent	76	57	73	60
Transient	5	116	5	116
Crystal artifact	118	34	22	130
Sum	199	207	100	306

Table 2: Distribution of heterodimers and homodimers within the dataset. Whole chains were considered as homodimers when exceeding a similarity of 96 % otherwise as heterodimers. The interface region is classified as homodimeric if the similarity index is above 85 %.

Results and Discussion

Evaluation of the descriptors

Diverse combinations of the developed descriptors were tested with the multi-class SVM. Previously, we have divided our dataset in two independent sets, one for the training of the SVM and the other for external validation of our model. Our training set contained 254 complexes (74 permanent, 60 transient and 120 crystal artifacts) which we have taken from the dataset used in the study of Zhu et al. [14]; the remaining 152 complexes (59 permanent, 61 transient and 32 crystal artifacts) were included in our test set. We obtained the best results with the usage of only two descriptors - the hydrophobic score and the quotient of the interface ratios. An accuracy of 91.7 % was achieved with the multi-class approach in a leave-one-out cross-validation. These two promising descriptors were also used for model generation in the two-stage SVM which resulted in a total accuracy of 92.5 % in the leave-one-out cross-validation. The separation of crystal artifacts and biological complexes reaches an accuracy of 95.3 % while the discrimination of permanent and transient complexes achieves an accuracy of 89.7 % (Table 3a). Additionally, we performed a ten-fold cross-validation for the multi-class SVM as well as for the two-stage SVM to test the stability of the generated model. This validation resulted in an accuracy of 91.3 % for the multi-class SVM, as well as of 93.7 % for the first step and of 88.8 % for the second step of the two-stage SVM. We have applied the generated two-stage SVM model on our test set for external validation. Within this dataset we achieved a total accuracy of 77.1 % (Table 3b).

a)	**Predicted**			
	Permanent	**Transient**	**Crystal artifact**	**Total**
Permanent	71	1	2	74
Transient	6	51	3	60
Crystal artifact	3	4	113	120
Total	80	56	118	254
b)	**Predicted**			
	Permanent	**Transient**	**Crystal artifact**	**Total**
Permanent	41	16	2	59
Transient	7	47	7	61
Crystal artifact	3	0	29	32
Total	51	63	38	152

Table 3: Two-stage SVM results using two descriptors: $\Delta G_{Hydrophobic}$ and IF-quotient. a) Leave-one-out cross-validation results on the training set. b) Performance results of the test set.

Refinement of the model – using only one descriptor for classification

We further investigated the two promising descriptors mentioned above considering their predictivity. Figure 1 shows the distribution of these in the constructed training dataset. It is directly obvious that the first descriptor, the $\Delta G_{Hydrophobic}$, is perfectly suited to discriminate between crystal artifacts and biological complexes. The second descriptor, the quotient of the interface ratios, on the other hand is able to distinguish permanent from transient interactions.

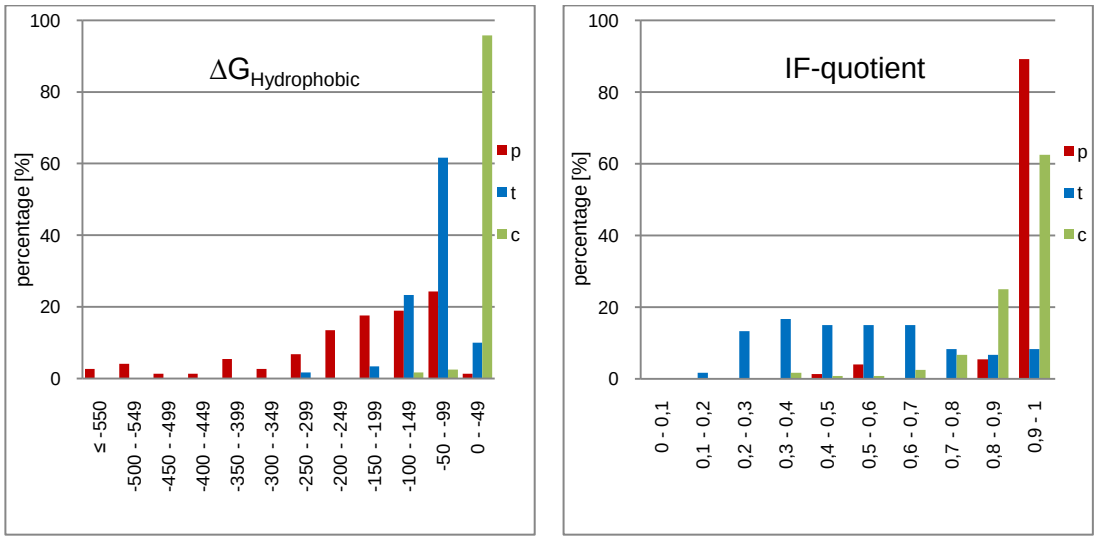

Figure 1: Distribution of the two descriptors ($\Delta G_{Hydrophobic}$, IF-quotient) for the three different types of protein-protein interactions of the training set; p = permanent, t = transient, c = crystal artifact.

Due to this observation, we performed the two-stage SVM using only one descriptor for each separation step. We include the molecular weight difference as an additional descriptor in this analysis. This is done to allow a comparison of our "single" descriptor approach to that accomplished by Kottha and Schroeder [16]. They achieved an accuracy of 80 % in the discrimination of transient and permanent interactions in their dataset using a SVM. Table 4 summarizes our results using these three descriptors on the training set as well as on the test set. The separation of biological complexes and crystal artifacts works extremely well employing the $\Delta G_{Hydrophobic}$ as the only descriptor. In both datasets we received accuracies of 95.2 % and 91.5 % respectively. On the other hand, using either the quotient of interface ratios or the molecular weight difference as the only descriptor shows poor performance in the differentiation of biological complexes and crystal artifacts. The quotient of interface ratios as well as the molecular weight difference, describe symmetry aspects of protein-protein interactions. Consequently, permanent and crystal artifacts which have similar homodimeric structures cannot be discriminated by these descriptors. However, the differentiation of biological relevant and non-relevant complexes also worked with former methods [14, 17, 18]. The more severe problem is the correct classification of permanent and transient protein-protein interactions. Herein the molecular weight difference and the interface quotient perform better in both datasets. Our descriptor, the interface quotient, exceeds the molecular weight difference by 4 – 7 %. An important observation in this analysis is the strong dependency of the achieved accuracies on the composition of the datasets. This aspect is investigated in the next section of the results.

Descriptors	Training dataset		Test dataset	
	B - C[a]	P - T[b]	B - C	P - T
$\Delta G_{Hydrophobic}$	95.2	72.4	91.5	73.0
IF-quotient	64.9	89.5	all classified as B	68.3
ΔMW[c]	70.1	85.8	all classified as B	61.6

Table 4: Two-stage SVM results using a single descriptor in each step. [a] Discrimination of biological complexes (B) and crystal artifacts (C). [b] Discrimination of permanent (P) and transient (T) interaction types. [c] Molecular weight difference of the subunits compromising the interface.

Analysis of the composition of the datasets concerning homodimers and heterodimers

Having a closer look at the composition of the datasets it becomes clear that most datasets are biased with respect to the distribution of homo- and heterodimeric structures (Table 5 and datasets used in [14, 16, 18, 22]). Within the training dataset we found no transient homodimers at all and within the test set only five were included. The training dataset Kottha and Schroeder used for their classification also contained only 13 transient homodimers (3.2 %). Their first differentiation method, using only the molecular weight difference, did not perform well for permanent heterodimers; only 52 % of those were classified correctly. Due to this result and our observations the separation of permanent and transient heterodimers seems to be the toughest problem related to the classification of protein-protein interactions. To our knowledge this problem is not been formulated yet and is probably overlooked due to the bias in the dataset. Consequently, most classification methods rely rather on the differentiation of homodimers and heterodimers than on the true classification of permanent and transient complexes.

	Type	Permanent	Transient	Crystal artifact	Sum
Training set	**Homodimers**	53	0	92	145
	Heterodimers	21	60	28	109
Test set	**Homodimers**	23	5	26	54
	Heterodimers	36	56	6	98
Sum		133	121	152	406

Table 5: Composition of the datasets concerning homodimers and heterodimers. The classification of homo- or heterodimer is based a sequence identity of the whole protein chains. Protein complexes achieving a chain identity of more than 96 % were classified as homodimers.

To circumvent this problem, we have constructed a new dataset containing only heterodimeric complexes (57 permanent and 116 transient). Again, we analyzed the performance of the hydrophobic score, the quotient of the interface area ratios and the molecular weight difference. Table 6 includes the results of this analysis. The analysis shows that the classification of permanent and transient interactions using the molecular weight difference as only descriptor is impossible. In this case, all complexes are classified as transient interactions. In contrast, $\Delta G_{Hydrophobic}$ as well as the quotient of the interface area ratios perform reasonably well, whereas $\Delta G_{Hydrophobic}$ used as the only descriptor achieves the best results in the leave-one-out and in the 10-fold cross-validation, with 80.3 % and 80.2 % respectively. The combination of both descriptors obtains an even higher accuracy of 83.8 % in the leave-one-out cross-validation and 81.8 % in the ten-fold cross-validation. The classification of transient interactions works remarkably well, with no more than 10 out of 116 false classified complexes. Contrarily, the permanent heterodimers show higher false positive rates. A possible reason might be the often used, but dubious criterion for the selection of permanent complexes: a complex is

termed permanent, if the separate crystal structures for the subunits of the complex are not available [9, 20]. However, the crystallization of the individual subunits might not be possible yet, although they may exist in vivo.

Descriptor	Accuracy [%]		Interaction type	Predicted	
	LOOCV	10-fold CV		P[a]	T[b]
IF-quotient	75.1	73.3	**P**	24	33
			T	10	106
$\Delta G_{Hydrophobic}$	80.3	80.2	**P**	24	33
			T	1	115
ΔMW	all classified as T	67.4	**P**	0	57
			T	0	116
IF-quotient + $\Delta G_{Hydrophobic}$	83.8	81.4	**P**	32	25
			T	3	113

Table 6: Performance of the classification of permanent and transient heterodimers employing the second step of the two-stage SVM with different descriptors. Results of a leave-one-out (LOOCV) and ten-fold cross-validation (10-fold CV) are shown. [a] P = permanent. [b] T = transient.

Conclusion

In this study we have analyzed the performance of diverse descriptors for the discrimination of three different protein-protein interactions: permanent, transient and crystal artifacts. We achieved an accuracy of 92.5 % using two descriptors – the hydrophobicity ($\Delta G_{Hydrophobic}$) and the quotient of the interface area ratios. An important observation we made was the fact that the performance of the methods is highly dependent of the constitution of the datasets. We found a bias in the available datasets towards a composition of predominantly heterodimeric transient interactions and homodimeric permanent interactions. Consequently, the rearrangement of the datasets reveals the real challenge in this field: the correct discrimination of transient and permanent heterodimers. Many former methods may only differentiate between homodimers and heterodimers since most of the methods used these biased datasets for the training of their models. In our new constructed dataset which only consists of permanent and transient heterodimers we achieved a classification accuracy of 80.2 % in a 10-fold cross-validation using the hydrophobicity as the only descriptor. An even higher accuracy of 81.4 % can be obtained including the quotient of the interface ratios.
For further research it would be interesting to shed more light on the existence and ratio of homodimeric transient protein-protein complexes in nature.

Web server for classification of protein-protein interactions
A web server for the usage of our method is currently in progress and will be finished until September this year.

Acknowledgments
We would like to thank Gudrun Lange and Robert Klein from Bayer CropScience for helpful discussions and a long and fruitful cooperation during the development of the HYDE scoring function. We would like to thank Ingolf Sommer for providing the structures of the 120 crystal artifacts used in our dataset. We are grateful to Stefan Bietz for provision of the sequence alignment tool.

References

1: Wells, J. A., McClendon, C. L. (2007). Reaching for high-hanging fruit in drug discovery at the protein-protein interfaces. Nature 450, 1001-1009
2: Tsai, C.-J., Nussinov, R. (1997). Hydrophobic folding units at protein-protein interfaces: Implications to protein folding and to protein-protein association. Prot. Sci. 6, 1426-1437
3: Jones, S., Thornton, J. M. (1996). Principles of protein-protein interactions. Proc. Natl. Acad. Sci. USA 93, 13-20
4: Janin, J., Rodier, F. (1995). Protein-protein interactions at crystal contacts. Proteins 23, 580-587
5: Carugo, O., Argos, P. (1997). Protein-protein crystal-packing contacts. Prot. Sci. 6, 2261-2263
6: Lo Conte, L., Chothia, C., Janin, J. (1999). The Atomic Structure of Protein-Protein Recognition Sites. J. Mol. Biol. 285, 2177-2198
7: Cieslik, M., Derewenda, Z. S. (2009). The role of entropy and polarity in intermolecular contacts in protein crystals. Acta Crystallogr. 65, 500-509
8: Lukman, S., Sim, K. (2007). Interacting amino acid preferences of 3D pattern pairs at the binding sites of transient and obligate protein complexes. WSPC 11, 14
9: De, S., Krishnadev, O., Srinivasan, N., Rekha, N. (2005). Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. BMC Struct. Biol. 5, 15
10: Nooren, I. M. A., Thornton, J. M. (2003). Structural Characterisation and Functional Significance of Transient Protein-Protein Interactions. J. Mol. Biol. 325, 991-1018
11: Neuvirth, H., Raz, R., Schreiber, G. (2004). Promate: a structure based prediction program to identify the location of protein-protein binding sites. J. Mol. Biol. 338, 181-99
12: Ponstingl, H., Henrick, K., Thornton, J. M. (2000). Discriminating Between Homodimeric and Monomeric Proteins in the Crystalline State. Proteins 41, 47-57
13: Mintseris, J., Weng, Z. (2003). Atomic Contact Vectors in Protein-Protein Recognition. Proteins 53, 629-639
14: Zhu, H., Domingues, F. S., Sommer I., Lengauer, T. (2006). NOXclass: prediction of protein-protein interaction types. BMC Bioinformatics 7, 24
15: Bahadur, R. P., Chakrabarti, P., Rodier, F., Janin, J. (2004). A Dissection of Specific and Non-specific Protein-Protein Interfaces. J. Mol. Biol. 336, 943-955
16: Kottha, S., Schroeder, M. (2006). Classifying permanent and transient protein interactions. GCB, 54-63
17: Liu, Q., Li, J. (2009). Propensity vectors of low-ASA residue pairs in the distinction of protein interactions. Proteins 78, 589-602
18: Block, P., Paern, J., Hüllermeier, E., Sanschagrin, P., Sotriffer, C. A., Klebe, G. (2006). Physicochemical Descriptors to Discriminate Protein-Protein Interactions in Permanent and Transient Complexes Selected by Means of Machine Learning Algorithms. Proteins 65, 607-622
19: Bradford, J. R., Westhead, D. R. (2005). Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics 8, 1487-1494
20: Ofran, Y., Rost, B. (2003). Analysing Six Types of Protein-Protein Interfaces. J. Mol. Biol. 325, 377-387
21: Park, S. H., Reyes, J.A., Gilbert, D. R., Kim, J. W., Kim, S. (2009). Prediction of protein-protein interaction types using association rule based classification. BMC Bioinformatics 10, 36
22: Rueda, L., Banerjee, S., Aziz, M.M., Raza, M. (2010). Protein-protein Interaction Prediction using Desolvation Energies and Interface Properties. IEEE, 17-22
23: Reulecke, I., Lange, G., Albrecht, J., Klein, R., Rarey, M. (2008). Towards an Integrated Description of Hydrogen Bonding and Dehydration: Decreasing False Positives in Virtual Screening with the HYDE Scoring Function. Chem. Med. Chem. 3, 885-897
24: Bigün, J., Granlund, G. H. (1987). Optimal Orientation Detection of Linear Symmetry. IEEE, 433–438
25: Dimitriadou, E., Hornik, K., Leisch, F., Meyer, F., Weingessel, A. (2011). Misc Functions of the Department of Statistics (e1071). Version 1.5-26
26: Chen, Y., Lim, C. (2008). Common physical basis of macromolecule-binding sites in proteins. Nuc. Acids Res. 36, 7078-7087
27: Madaoui, H., Guerois, R. (2008). Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. Proc. Natl. Acad. Sci. USA 22, 7708-7713
28: Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. Nuc. Acids Res. 28, 235-242
29: Ukkonen, E. (1985). Algorithms for Approximate String Matching. Information and Control. 64, 100-118