Structural Descriptors of Protein-Protein Binding Sites

Oliver Sander*, Francisco S. Domingues, Hongbo Zhu, Thomas Lengauer, Ingolf Sommer

Max-Planck-Institute for Informatics,
Stuhlsatzenhausweg 85
66123 Saarbrücken
Germany
E-mail: osander@mpi-sb.mpg.de

Structural bioinformatics provides new tools for investigating protein-protein interactions at the molecular level. We present two types of structural descriptors for efficiently representing and comparing protein-protein binding sites and surface patches. The descriptors are based on distributions of distances between five types of functional atoms, thereby capturing the spatial arrangement of physicochemical properties in 3D space. Experiments with the method are performed on two tasks: (1) detection of binding sites with known similarity from homologous proteins, and (2) scanning of the surfaces of two non-homologous proteins for similar regions.

Keywords: Structural Descriptor; Protein-Protein Interactions; Binding Sites.

1. Introduction

Throughout the life of a cell protein-protein interactions (PPI) are the driving force behind many molecular functions and cellular activities. Protein-protein interactions also drive many of the processes related to diseases, such as host-pathogen interactions including the immune response to invading pathogens and disease-related protein misfolding and aggregation.

Experimental high-throughput techniques such as yeast-two-hybrid screens, tandem affinity purification and co-immunoprecipitation can afford a comprehensive view of the network of protein-protein interactions inside a cell. The wealth of interaction data that are generated with these methods is further increased by predicting interactions computationally based on homology. However, these data suffer from severe limitations. First, experimentally derived interactions show inaccuracies,[31] which are then propagated by homology-based annotation. The discrepancies between different experimental and predicted data sets are considerable.[21] Second, binary interaction data lack molecular details about where and when, in which relative positioning, and how strongly proteins interact. This information is vital for assessing the effect of mutations on binding sites and for the development of inhibitors of PPIs.[2]

The spectrum of questions and methods in the field of protein-protein interactions is

[*] to whom correspondence should be addressed

wide. Appropriate classification of interactions e.g., as permanent or transient interactions,[17,18,37] is of high relevance in the cellular context but requires an in-depth study of structural features. In their paper from 2006,[1] Aloy and Russell proclaim structural analysis of protein interactions, interfaces, and binding sites as a crucial step towards understanding interactions at a system level. Temporal dynamics of interactions, spatial organizations of assemblies, locations of interactions and types of interactions need to be understood to place single interactors in their cellular or systems context.

In this work we study the similarities between binding sites of proteins. Global similarity of two proteins is neither necessary, nor sufficient for sharing similar binding partners. On the one hand, proteins from the same family can exhibit different binding specificities due to subtle changes in their binding sites. On the other hand, mimicking of binding sites enables two proteins with different global folds to bind to the same partner, such as the viral M3 protein imitating a chemokine homodimeric binding site[11] or the mimicking of CD4 by scorpion toxin.[12]

To study these phenomena purely data-driven analysis as well as similarity-based methods have been applied. The idea underlying data-driven analysis is that if complexes AB and A'C involving the domains A, A', B, and C were observed, A and A' being from the same homologous family, an alignment of A and A' can be used to analyze whether B and C are binding at equivalent binding sites to A and A', respectively. This indicates that the binding sites of B and C are likely to share some properties, as they are able to bind to similar partners. Henschel *et al.*[11] studied similar binding sites by extracting them from known complex structures using this concept.

In contrast to this data-driven analysis, similarity-based methods have been used to detect similarities between binding sites, despite global dissimilarity between the respective proteins. Similarity-based methods use either combinatorial matching procedures to find common properties or “structural descriptors” to capture the essential characteristics of a binding site. We use the notion “structural descriptor” here to mean an abstract representation allowing for efficient comparison, in contrast to methods like geometric hashing or clique search on correspondence graphs, which use simpler representations, but more complex combinatorial matching procedures.

Geometric hashing[35] and other combinatorial matching techniques have been applied numerous times for the comparison of binding sites (i.e., the binding residues of one interaction partner) as well as protein interfaces (i.e., the binding residues from both interaction partners).[24,26] Keskin *et al.*[15] identified similar binding behaviour by structural alignment of interface residues. In contrast to the data-driven approach by Henschel *et al.*[11] outlined above, this procedure requires less data, but relies heavily on the structural alignment method, making it difficult to differentiate between differences in binding sites and methodological artefacts. For the comparison of enzymatic active sites the software packages TESS[32] and JESS[3] were developed by the Thornton group. The Klebe group developed clique search and geometric hashing approaches for the comparison of small-ligand binding sites.[23,33]

The concept of using structural descriptors for representing functional sites or structural arrangements has been described previously. Stahl *et al.*[29] have used distance-based descriptions for comparing active sites of enzymes based on chemical and geometric properties. Subsequently the sites were clustered and visualized using a self-organizing-map (SOM) approach. For the analysis of protein-protein interaction interfaces, Mintseris and Weng[16] have proposed atomic contact vectors which consist of contact counts derived from thresholded distance matrices. Distributions of atomic distances have been used successfully in structure comparison.[6,7] In protein structure prediction, distributions of distances have been applied in the form of knowledge-based potentials for evaluating the fit of a sequence to a structure.[28] Bock *et al.*[4,5] use spin-image representations to represent the arrangement of neighboring residues around a residue of interest. Via *et al.*[30] provide a recent review of methods for detecting of protein surface similarities. Several of the methods, such as distance distributions and spin-image representations, stem from the computer vision research field.[19,25]

In our recent work[22] we demonstrated the applicability of structural descriptors to the specific task of predicting HIV coreceptor usage based on properties of the viral V3 loop region. Here, we examine their applicability to the more general task of binding site comparison.

We propose a method for representing and comparing protein-protein binding sites. The structural descriptor is based on distributions of pairwise distances between functional atoms. Thereby, the descriptor encodes the spatial arrangement of those physico-chemical properties in a vector of fixed length.

We evaluate two modes of analysis: (1) using the structural descriptor to describe a whole binding site, i.e. the set of all residues binding to the partner in a protein-protein interaction, and (2) to describe a set of surface residues as defined by a sphere of radius r around the Cα-atom of a given central residue. The first mode can be used for comparing predefined protein-protein binding sites, whereas the second mode can be used to scan the surfaces of two proteins for similarities, if the binding patches are not known *a priori*.

This article is organized as follows. In Section 2 we describe the details of the distance-based descriptor and methodological variants and the nearest-neighbor prediction protocol. In Section 3 we present results of the performance evaluation on a data set of protein kinases and a case study on scanning protein surfaces for similar patches.

2. Comparison of protein binding sites and surface patches

Here, we introduce two variants of structural descriptors SD_{bsite} and SD_{patches}. SD_{bsite} describes the spatial arrangement of physico-chemical properties for given set of residues for a predefined binding site. In contrast, SD_{patches} provides a representation for several small surface patches and computes a combined match score. See Figure 1 for a schematic overview of the two methods.

2.1. *Structural descriptors of protein-protein binding sites*

The structural descriptor SD_{bsite} takes a set R of binding site residues as input and encodes their relative positioning in three-dimensional space. Residues losing more than 1 Å^2 of solvent accessible surface area upon complexation with the binding partner[14] are defined

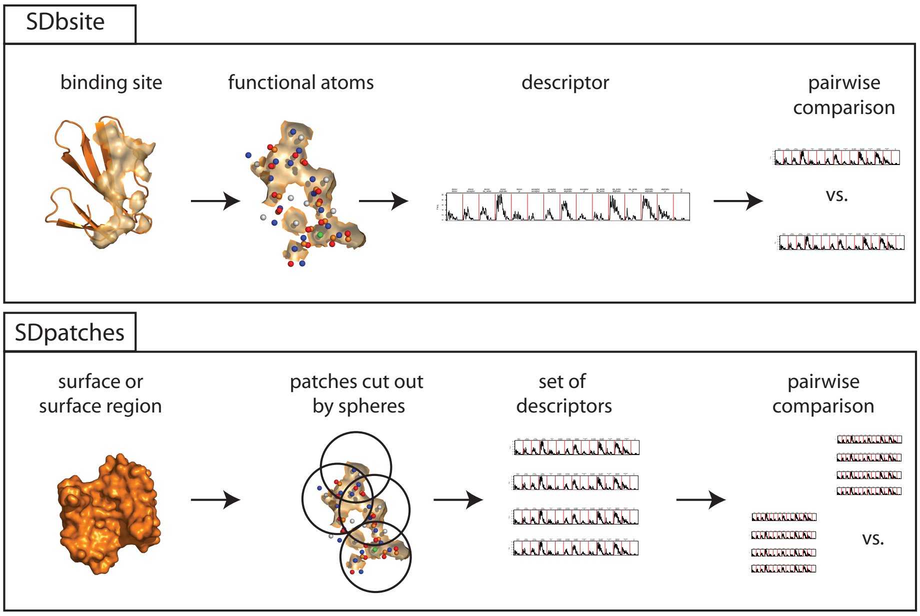


Fig. 1. Schematic overview of the structural descriptors SD_{bsite} and SD_{patches}

as binding site residues. The solvent-accessible surface areas for the single domains as well as their complexes are computed using NACCESS.[13]

Following Schmitt *et al.*[23] we represent the side chains using five functional atom types, namely, hydrogen-bond donor, acceptor, ambivalent donor/acceptor, aliphatic, or aromatic ring. Amino acids R, N, Q, K, and W are classified as donors. Acceptors are N, D, Q, and E. Ambivalent donor/acceptors comprise H, S, T, and Y. As aliphatic amino acids we consider A, R, C, I, L, K, M, P, T, and V. Pi-stacking centers are H, F, W, and Y. Pseudo-atoms for donors, acceptors, and ambivalent donor/acceptor interaction centers are placed at the respective nitrogen or oxygen atoms. For aliphatic and aromatic interaction centers all involved atom positions were averaged per residue to compute a pseudo-atom. We used the unweighted average of carbons to determine the center of aliphatic side chains, and do not consider backbone atoms as pi-stacking interaction centers.

To derive the structural descriptor, the spatial arrangement of these functional pseudo-atoms is encoded by distance distributions. For each of the 15 combinations of functional atom types (i.e. donor-donor, donor-acceptor, etc.) pairwise Euclidean distances between the respective pseudo-atoms in the residue set R are calculated. Note that the number of these distances depends on the number of pseudo-atoms in the two respective groups. From these distance matrices we derive distance distributions using a kernel density estimate with a Gaussian kernel and a smoothing kernel bandwidth of 1 Å. The density estimates are then discretized by uniform sampling at intervals of 1 Å from 1Å to 10Å, resulting in a 15 (distance distributions for atom type combinations) times 10 (sample points) dimensional vector. The resulting vector is used as a structural descriptor for a given set of binding site residues R. Distance distributions are representations of protein structure invariant under translation and rotation. The smoothing kernel bandwidth as well as the sampling intervals for the distance-based descriptors have been set to values based on empirical observations. Variations within a reasonable range did not result in significant changes of the performance.

2.2. *Comparison and retrieval of structural descriptors*

The structural descriptor is a vector of fixed length. The length only depends on the parameters of the method, not on the size or number of residues of the binding site to be described. By using a vectorial representation of a binding site or surface patch multivariate analysis and statistical learning techniques can directly be applied to the descriptors. Here we use simple nearest-neighbor classification, but in principle, kernel-based discriminative methods can be applied directly.

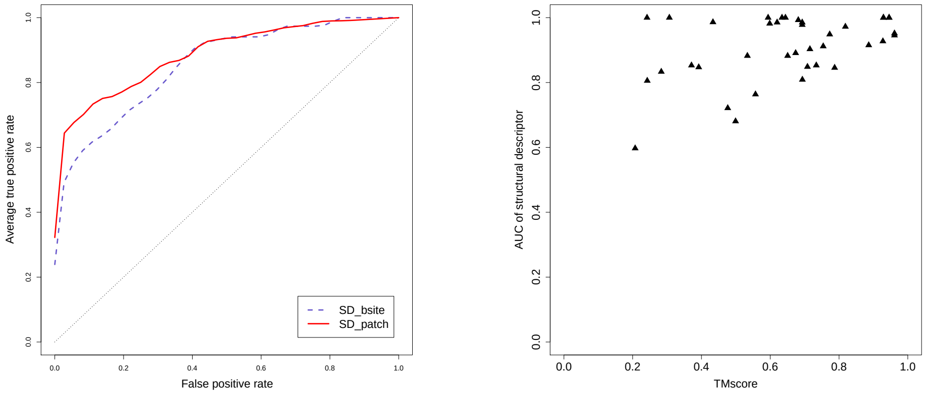
A wide variety of distance functions, Minkowski norms like the Euclidean or Manhattan metric, information theoretic measures like the Kullback-Leibler distance or Jensen-Shannon divergence, or other statistical approaches like χ^2-test, dot-products, or cosine distance can be used to compare two descriptor vectors. On the tasks and data sets studied here, the cosine and Euclidean measures provide very good performance. While on the rather small data sets used here for evaluation we used pairwise distance computations to determine nearest neighbors, spatial indexing methods like kd-trees[8] can be used to speed up the retrieval of nearest neighbors from a massive set of hundred thousands or millions of descriptors.

2.3. *Structural descriptors of protein surface patches*

While the SD_{bsite} descriptor relies on a predefined set of residues, here for SD_{patches}, we drop this prerequisite. In contrast to predefined protein-protein binding sites the comparison of two proteins for similar surface regions does not provide a defined set of residues to be described by the structural descriptor. SD_{patches} describes the surface of a protein or parts of it by a set of patches. Each patch is composed of the residues within a sphere of radius r around the Cα atom of a given central residue. In the current implementation we use one sphere per surface residue. A multi-resolution approach can be implemented by using spheres of different radii and combining the matches appropriately in the subsequent p-value computation.

Each surface patch is represented by a distance distribution like with SD_{bsite}. Thus the comparison of two protein surfaces turns into the comparison of two sets of descriptors. From the raw descriptor match scores described above, we compute p-values. This is done by generating a background distribution of similarity scores of unrelated pairs of descriptors. For efficient lookup of p-values the cumulative distribution function of the top 5% scores in the distribution is smoothed by a cubic spline with 4 knots and fitted by a piecewise linear function. p-values above a threshold of 5% are set to 1, to avoid the accumulation of spurious similarities.

To compare two sets of patches, each patch in the first set receives the score of the



(a) (b)

Fig. 2. Retrieval of similar binding sites: (a) ROC curve, (b) comparison of AUC performance of the structural descriptor against TMscore

respective best hit in the second set and the p-values of all hits get accumulated by multiplication (assuming statistical independence). To avoid numerical instabilities $-log_{10}(p)$ scores are computed and accumulated by summation.

3. Experiments & Evaluation

Experiments are performed on a set of binding sites from Pkinases and their respective binding partners. In addition to the quantitative evaluation, a case study on an instance of viral mimicking is presented.

3.1. *Retrieving Similar Binding Sites - Kinases*

We analyzed and evaluated the structural binding site descriptors SD_{bsite} and SD_{patches} on a set of protein kinase binding sites and the binding sites of their respective partners. This data sets consists of binding sites derived from domain interfaces from 25 Pkinase complexes comprising 50 binding sites. For the selection of these binding sites we used the SCOPPI database.[34] SCOPPI provides an extensive data set of domain-domain interactions for all SCOP domains in the protein data bank PDB. In addition to the preprocessed list of pairwise interactions, SCOPPI supplies a categorization of binding sites into face types. The binding sites of all domains within a specific family are compared on the basis of how many residues they share that are matched in an alignment of the two protein families. Based on this criterion strongly overlapping binding sites on equivalent regions of the domain surface are classified into the same face type.

Complexes with redundant entries (i.e., using the same binding faces in both interactions) were removed if they exhibited a sequence identity level of at least 90% with already included complexes. From the resulting set of 50 binding sites we removed one of the two binding sites in each symmetric homo-dimeric complex. Due to symmetry, these pairs are highly similar and would be trivial to find in the subsequent matching experiment. The resulting data set consists of a set of 38 binding sites. Each of these binding sites is labelled with a four-tuple: SCOP family of the domain of the binding site, SCOPPI face type of the binding site, SCOP family of the binding partner, face type of the partnering binding site.

In the retrieval experiment we aimed at efficiently recovering similar binding sites, as defined by the label described above. Both compared methods, $SD_{\rm bsite}$ and $SD_{\rm patches}$, use a set of binding site residues as input. While $SD_{\rm bsite}$ represents the binding site globally, $SD_{\rm patches}$ represents the binding site by a set of smaller local patches, as described in Section 2.1. In order to assess the predictive performance of the structural descriptors we performed leave-one-out cross-validation. Evaluation of predictive performance was done using ROCR.[27]

The measure used for evaluation of predictive performance is the area under the ROC curve (AUC). The AUC is calculated by adding the area of trapezoid strips under the ROC curve. This is equal to the value of the Wilcoxon-Mann-Whitney test statistic and also to the probability that the classifier will score a randomly drawn positive sample higher than a randomly drawn negative sample.[10] A ROC curve is computed for each sample in the data set, quantifying, how well similar sites are being retrieved.

In Figure 2 (a) a vertically averaged ROC curve is shown for each of the two descriptors $SD_{\rm bsite}$ and $SD_{\rm patches}$. $SD_{\rm patches}$ clearly outperforms $SD_{\rm bsite}$ on the retrieval task of the 38 kinase binding sites. This is due to the sensitivity of $SD_{\rm bsite}$ to small changes in the binding site definitions. For example, augmenting a highly similar binding site by a small terminal tail changes the descriptor considerably. The AUC values (one per query binding site) have a mean of 0.9078 and a median of 0.9364 for $SD_{\rm bsite}$ and a mean of 0.9236 and a median of 1.0000 for $SD_{\rm patches}$. Thus, for at least half of the 38 binding sites $SD_{\rm patches}$ is able to perform a perfect classification, i.e. all similar binding sites from the same class are ranked above binding sites from other classes.

While the AUC quantifies the overall ability to rank samples with the same label higher than samples with another label, the accuracy at the top-rank or in the top-k ranks quantifies the fraction of training samples for which a similar site could be detected. $SD_{\rm bsite}$ is able to find for 68.42%, 73.68%, and 81.58% of the 38 binding sites a similar binding site with the same label at the top-rank, in the top-3, and in the top-5, respectively. $SD_{\rm patches}$ finds for 71.05%, 81.58%, and 89.47% of the 38 binding sites a similar binding site with the same label at the top-rank, in the top-3, and in the top-5, respectively.

Figure 2 (b) shows a scatter plot of the AUC of $SD_{\rm bsite}$ per binding site versus the TMscore[36] between that binding site and its closest binding site from the same family. TMscore is the structural similarity measure provided by the TM-align program. Although this program performs the structural alignment respecting sequence ordering, it can be applied here, as the binding sites labelled to be similar are from the same families. It can be observed that the variability in performance of $SD_{\rm bsite}$ depends on the TMscore: a high TMscore implies a high AUC performance, whereas lower TMscores might result in worse AUCs. This means that $SD_{\rm bsite}$ performs very well on binding sites with a structurally similar closest hit. With decreasing similarity of the best hit the worst-case AUC performance decreases linearly, but for some dissimilar binding sites good performance is still possible.

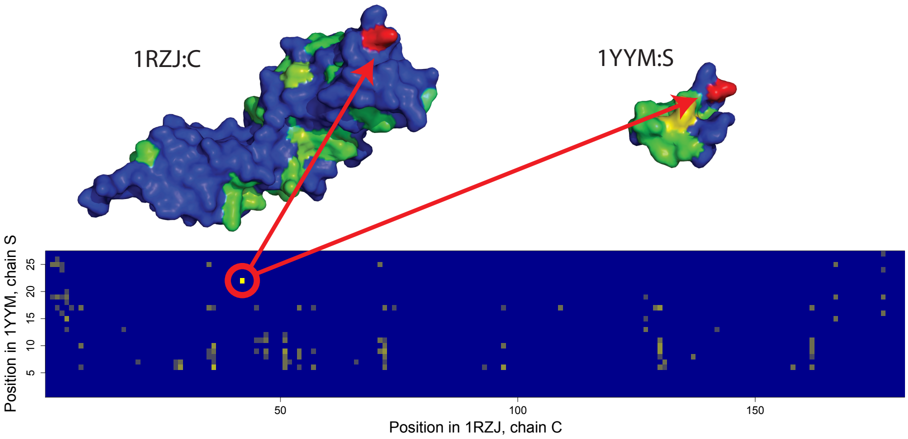


Fig. 3. Using SD_{patches} to compare CD4 (1RZJ:C) with its mimicking scorpion toxin protein (1YYM:S)

3.2. *Scanning for similar surface patches*

Huang *et al.*[12] analyze the mimicking of CD4 by a small scorpion toxin fold. The scorpion toxin is a 31-amino-acid protein consisting of two beta strands and an α-helix, held together by disulphide bonds. It has been designed to mimick the binding site of CD4 to the viral protein gp120.

We use the structural descriptor SD_{patches} to compare the surface of CD4 against the surface of the mimicking scorpion toxin. Figure 3 shows the pairwise similarities between patches in CD4 (1RZJ:C) and the scorpion toxin (1YYM:S). Patches are colored by significance of their similarity. The most similar red surface patches in both proteins show a p-value of $10^{-4.223}$ for similarity assessed with SD_{patches} and, in fact, they correspond to the loops mimicking each other. The matrix in Figure 3 shows the pairwise similarity p-values for all patches in both proteins, ordered along the sequences of both proteins. The highest match is highlighted. The second highest similarity (p-value of $10^{-2.742}$) is clearly less pronounced. The structural descriptor SD_{patches} is able to pick the binding site from the scorpion toxin mimicking the CD4 binding site, despite the global dissimilarities of the two proteins.

4. Conclusion and Outlook

The proposed structural descriptor is an efficient and accurate method for describing binding sites and surface patches. The major remaining problem is the evaluation, as data with annotation is scarce. There is no clear notion of "non-trivial" similarity that should be detected by methods focusing on local similarity of proteins. Even if the proteins have the same functions or bind to the same partner at the same respective site, it is not guaranteed that they share some detectable similarities. Further directions are (1) relating local similarities of surface patches to protein function, and (2) comparison of the described approach against other descriptor based methods like ACVs, spin-image representations and combinatorial matching approaches like geometric hashing or clique search.

Acknowledgement

We would like to thank Christoph Winter for providing a flatfile version of the SCOPPI database. Analysis of the results and prediction was performed using the statistical language R[20] with the package ROCR.[27] Protein structure visualizations were created using PyMOL.[9]

References

1. Patrick Aloy and Robert B Russell. Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol*, 7(3):188–197, Mar 2006.
2. Michelle R Arkin and James A Wells. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat Rev Drug Discov*, 3(4):301–317, Apr 2004.
3. Jonathan A Barker and Janet M Thornton. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics*, 19(13):1644–9, Sep 2003.
4. Mary Ellen Bock, Guido M. Cortelazzo, Carlo Ferrari, and Concettina Guerra. Identifying similar surface patches on proteins using a spin-image surface representation. In A. Apostolico, M. Crochemore, and K. Park, editors, *CPM 2005, LNCS 3537*, pages 417–428, Heidelberg, 2005. Springer-Verlag Berlin.
5. Mary Ellen Bock, Claudio Garutti, and Concettina Guerra. Discovery of similar regions on protein surfaces. *Journal of Computational Biology*, 14(3):285–299, 2007.
6. Stefan Canzar and Jan Remy. Shape distributions and protein similarity. In *Proceedings of the German Conference on Bioinformatics (GCB '06)*, pages 1–10, 2006.
7. Oliviero Carugo and Sandor Pongor. Protein fold similarity estimated by a probabilistic approach based on C(alpha)-C(alpha) distance comparison. *J Mol Biol*, 315(4):887–898, 2002.
8. Mark de Berg, Marc de Kreveld, and Mark Overmars. *Computational Geometry. Algorithms and Applications*. Springer, Berlin, 2000.
9. Warren L. DeLano. The PyMOL molecular graphics system, 2002. DeLano Scientific. San Carlos, CA, USA. http://www.pymol.org.
10. Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.
11. Andreas Henschel, Wan Kyu Kim, and Michael Schroeder. Equivalent binding sites reveal convergently evolved interaction motifs. *Bioinformatics*, 22(5):550–555, Mar 2006.
12. Chih-Chin Huang, Francois Stricher, Loic Martin, Julie M Decker, Shahzad Majeed, Philippe Barthe, Wayne A Hendrickson, James Robinson, Christian Roumestand, Joseph Sodroski, Richard Wyatt, George M Shaw, Claudio Vita, and Peter D Kwong. Scorpion-toxin mimics of CD4 in complex with human immunodeficiency virus gp120 crystal structures, molecular mimicry, and neutralization breadth. *Structure*, 13(5):755–768, May 2005.
13. Simon J. Hubbard and Janet M. Thornton. Naccess. Computer Program, Department of Biochemistry and Molecular Biology, University College London., 1993.
14. Susan Jones and Janet M. Thornton. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*, 93(1):13–20, Jan 1996.
15. Ozlem Keskin and Ruth Nussinov. Similar binding sites and different partners: implications to shared proteins in cellular pathways. *Structure*, 15(3):341–354, Mar 2007.
16. Julian Mintseris and Zhiping Weng. Atomic contact vectors in protein-protein recognition. *Proteins*, 53(3):629–639, 2003.

17. Irene M A Nooren and Janet M Thornton. Diversity of protein-protein interactions. *EMBO J*, 22(14):3486–92, Jul 2003.
18. Yanay Ofran and Burkhard Rost. Analysing six types of protein-protein interfaces. *J Mol Biol*, 325(2):377–87, Jan 2003.
19. Robert Osada, Thomas Funkhouser, Bernard Chazelle, and David Dobkin. Shape distributions. *ACM Transactions on Graphics (TOG)*, 21(4):807–832, October 2002.
20. R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.
21. Fidel Ramirez, Andreas Schlicker, Yassen Assenov, Thomas Lengauer, and Mario Albrecht. Computational analysis of human protein interaction networks. *Proteomics*, 7(15):2541–52, 2007.
22. Oliver Sander, Tobias Sing, Ingolf Sommer, Andrew J Low, Peter K Cheung, P. Richard Harrigan, Thomas Lengauer, and Francisco S Domingues. Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. *PLoS Comput Biol*, 3(3):e58, Mar 2007.
23. Stefan Schmitt, Daniel Kuhn, and Gerhard Klebe. A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol*, 323(2):387–406, 2002.
24. Maxim Shatsky, Alexandra Shulman-Peleg, Ruth Nussinov, and Haim J. Wolfson. *Recognition of Binding Patterns Common to a Set of Protein Structures*, volume 3500, chapter Lecture Notes in Computer Science, pages 440 – 455. Springer-Verlag GmbH, May 2005.
25. Philip Shilane, Patrick Min, Michael Kazhdan, and Thomas Funkhouser. The Princeton shape benchmark. In *Shape Modeling International, Genova, Italy*, 2004.
26. Alexandra Shulman-Peleg, Ruth Nussinov, and Haim J. Wolfson. SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. *Nucleic Acids Research*, 33:W337–W341, 2005.
27. Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer. ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941, 2005.
28. Manfred J. Sippl. Knowledge-based potentials for proteins. *Curr Opin Struct Biol*, 5(2):229–235, 1995.
29. Martin Stahl, Chiara Taroni, and Gisbert Schneider. Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Protein Eng*, 13(2):83–88, 2000.
30. Allegra Via, Fabrizio Ferre, Barbara Brannetti, and Manuela Helmer-Citterich. Protein surface similarities: a survey of methods to describe and compare protein surfaces. *Cell Mol Life Sci*, 57(13-14):1970–1977, Dec 2000.
31. Christian von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, May 2002.
32. Andrew C. Wallace, Neera Borkakoti, and Janet M. Thornton. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci*, 6(11):2308–23, Nov 1997.
33. Nils Weskamp, Daniel Kuhn, Eyke Hüllermeier, and Gerhard Klebe. Efficient similarity search in protein structure databases by k-clique hashing. *Bioinformatics*, 20(10):1522–6, Jul 2004.
34. Christof Winter, Andreas Henschel, Wan Kyu Kim, and Michael Schroeder. Scoppi: a structural classification of protein-protein interfaces. *Nucleic Acids Res*, 34(Database issue):D310–D314, Jan 2006.
35. Haim J. Wolfson and Isidore Rigoutsos. Geometric hashing: an overview. *Computational Science and Engineering, IEEE*, 4(4):10–21, 1997.
36. Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*, 33(7):2302–2309, 2005.
37. Hongbo Zhu, Francisco S Domingues, Ingolf Sommer, and Thomas Lengauer. NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics*, 7:27, 2006.