

# STRUCTURAL FEATURES AND EVOLUTION OF PROTEIN-PROTEIN INTERACTIONS

JOACHIM VON EICHBORN      STEFAN GÜNTHER  
joachim.eichborn@charite.de      stefan.guenther@charite.de

ROBERT PREISSNER  
robert.preissner@charite.de

*Structural Bioinformatics Group, Institute for Physiology, Charité–University  
Medicine, Arnimallee 22, 14195 Berlin, Germany*

Solved structures of protein-protein complexes give fundamental insights into protein function and molecular recognition. Although the determination of protein-protein complexes is generally more difficult than solving individual proteins, the number of experimentally determined complexes increased conspicuously during the last decade. Here, the interfaces of 750 transient protein-protein interactions as well as 2,000 interactions between domains of the same protein chain (obligate interactions) were analyzed to obtain a better understanding of molecular recognition and to identify features applicable for protein binding site prediction.

Calculation of knowledge-based potentials showed a preference of contacts between amino acids having complementary physicochemical properties. The analysis of amino acid conservation of the entire interface area showed a weak but significant tendency to a higher evolutionary conservation of protein binding sites compared to surface areas that are permanently exposed to solvent. Remarkably, contact frequencies between outstandingly conserved residues are much higher than expected confirming the so-called “hot spot” theory. The comparisons between obligate and transient domain contacts reveal differences and point out that structural diversification and molecular recognition of protein-protein interactions are subjected to other evolutionary aspects than obligate domain-domain interactions.

*Keywords:* protein interactions; domain interactions; conservation; interaction hot spots.

## 1. Introduction

Proteins interact quickly and specifically with other biomolecules, especially other proteins [18]. Compared to the genomic data of complex organism that increased explosively during the last decades, knowledge about the human interactome is still very fragmentary [21]. The analysis of solved structures of protein-protein complexes give fundamental insights into molecular recognition and may help to obtain protein properties that can be used to predict specific interaction partners by *in silico* methods. The properties of protein-protein interaction sites have been studied since the 1970s [1, 3, 4, 12, 13] and results were integrated in some very useful tools for protein-protein docking [5, 7, 10]. Such tools can provide evidence for the right binding mode of two proteins. Docking is usually achieved by maximizing the shape

and physicochemical complementarity through generation of large sets of possible conformations. However, sampling of relevant conformations and the discrimination of native-like configurations from the large number of non-native alternatives remain challenging.

To determine the relevance of a certain protein-protein interaction model (decoy), the residue contacts between proteins in the model can be evaluated. In 1995 a pioneering work of Clackson and Wells was published that has shown that just a small and complementary set of cooperative contact residues, termed “hot spots” maintains the largest part of the binding affinity [6]. Clackson and Wells distinguished between functionally important residues and surrounding residues that are less important. However, their analysis is based on a single complex structure of the human growth hormone and its receptor. In 2007 a similar line of reasoning was taken up by Shulman-Peleg et al. [20] based on a more extensive data set of 71 complex structures. They came to similar results, claiming that a few highly conserved residues are crucial for protein interactions.

Physical interactions between proteins are mostly controlled by their domains, the fundamental units in the evolution of genes and functions. Beside domain-domain contacts that develop temporarily during specific protein-protein interactions (transient interactions) crystallographic information on numerous domain-domain contacts between domains of the same multidomain protein are available in the Protein Data Bank (PDB) [2]. Such obligate interactions would provide a valuable resource for analysing molecular recognition, if they show similar properties of molecular recognition as transient interactions. An existing analysis addressing this question was performed 2005 and is based on a dataset consisting of 212 transient and 115 obligate protein complexes [16]. They come to the conclusion that some differences in evolutionary pressure exist between the two types of interactions. Obligate complexes evolve at a relatively slower rate, allowing them to coevolve with their interacting partners. Nevertheless the analysis is based only on selected domain family-family pairs representing only a small part of the structural information available today.

Here, we have analyzed characteristics of protein-protein interaction sites with the aim of describing them and to obtain properties that are suitable for improving protein docking. For this purpose we have performed a large-scale analysis based on a data set of 2,000 obligate and 750 transient domain-domain interfaces. Differences between surface regions and interaction sites as well as differences between obligate and transient interactions are pointed out. The hot spot hypothesis is rechecked and specific contacts between residues are systematically evaluated. Knowledge-based residue contact maps are provided for evaluation of protein-protein decoys and protein models.

## 2. Methods

### 2.1. Data Set

For this study the protein interaction sites provided by the JAIL-database [11] are used. JAIL contains an exhaustive set of 190,000 protein interaction sites. They are computed using the protein structures available from the PDB.

The data sets used for this study are based on all interactions between SCOP-defined protein domains [17] that are part of the JAIL database in the version from June 2009. We constructed two sets of interactions. The first one contains interactions between domains that are part of different proteins (transient). In contrast, the interactions in the second set have to be between domains on one amino amino acid chain (obligate).

Afterwards, both sets were filtered, so that interactions between domains of any two different SCOP-classes are represented by at most one interaction. Thereby, nonredundant sets of protein class interactions are guaranteed.

After filtering we obtained 750 transient interactions and 2,000 obligate ones.

### 2.2. Interface Definition

The residues that participate in a protein interaction are referred to as the “interface” in the following text. We derived the interface definition used by JAIL. According to this definition a complete residue is part of an interface if at least one atom of the amino acid is located within a range of 4.5 Å to any atom of the interacting domain. These residues are considered to be in contact with each other (Fig. 1). Additionally, one part of an interface must consist of at least 5 C $\alpha$  atoms.

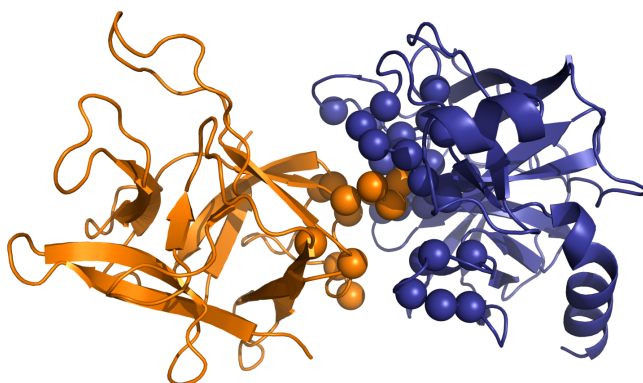


Fig. 1: Interface between two SCOP-domains (SCOP-Ids d1avxa\_ and d1avxb\_). The two interacting domains are shown in orange and blue. C $\alpha$  atoms of residues that are in contact with residues of the other domain are shown as spheres.

### 2.3. Surface Definition

To determine the residues on the protein surface, we used calc-surface [8]. If only one atom of a residue is on the surface, the whole residue is considered to be part of the surface.

### 2.4. Contact Frequency Normalization

The preferences and aversions towards residue contacts are given as multiples of the expected values to correct for different frequencies of the participating amino acids. These multiples are computed by equation 1.  $x$  and  $y$  denote the interacting amino acids and  $C(x, y)$  the number of contacts between them.  $count(i)$  is the number of occurrences of an amino acid  $i$  in interfaces and  $count(contacts)$  is the total number of contacts that exist.

$$M(x, y) = \frac{C(x, y)}{\frac{count(x)}{\sum_{z \in AA}^{20} count(z)} \cdot \frac{count(y)}{\sum_{z \in AA}^{20} count(z)} \cdot count(contacts)} \quad (1)$$

The normalization of the contact values between residues of specific conservation grades is done analogously.

### 2.5. Conservation Score

The conservation information is taken from the ConSurf-DB [9]. The ConSurf-DB contains precomputed conservation scores for all structures in the PDB and is updated on a monthly basis. The conservation scores listed in the ConSurf-DB are computed using the Rate4Side algorithm [19] based on a multiple sequence alignment of homologue sequences.

Positive scores indicate variable positions and negative scores conserved ones. Afterwards, the scores are normalized, such that the average over all residues is zero and the standard deviation is one. Finally the scores are binned in the integer range of one to nine. These binned values are used in this study.

## 3. Results

We investigated the frequencies of the different amino acids in interfaces and on surfaces. Fig. 2 displays the results for the transient interfaces, the results for obligate interfaces are almost similar (not shown). The most remarkable difference between the groups is, that Glycine is nearly as frequent in obligate interfaces as it is on the surfaces whereas for transient interfaces Glycine is more frequent on the surfaces.

Fig. 3 shows the amino acids' interface/surface preferences as ratios between the interface and surface frequencies from Fig. 2. A characteristic pattern of interface/surface preferences can be observed. For example Cysteine, Tyrosine,

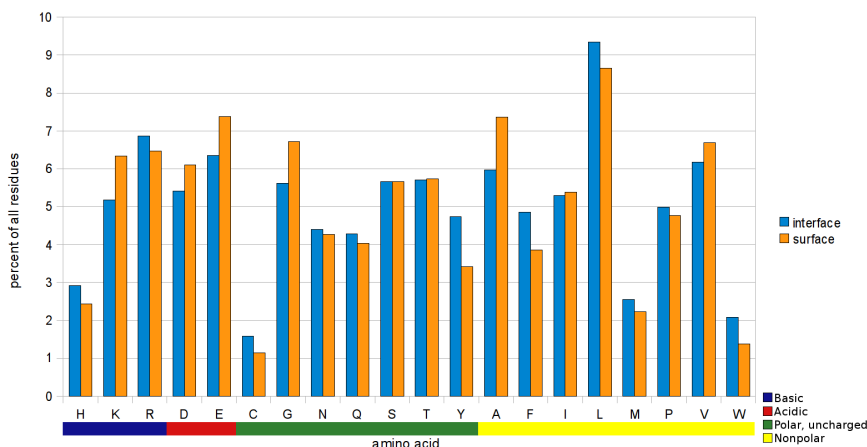


Fig. 2: Frequencies of the different amino acids in the interfaces and surfaces of transient interactions.

Phenylalanine and Tryptophan are considerably more frequent in interfaces than in surfaces.

In the following we examined the pairwise contact preferences of the different amino acid types. In the data sets used for this study each interface contains on average nearly 60 residue contacts. We created heat maps showing the preferences of all possible amino acid pairs to be in contact. To compute these values at first the absolute number of occurrences is counted for each amino acid pair. Those numbers are affected by the frequencies of the amino acids in the interfaces (the “interface” bars from Fig. 2). For example there will be a large number of contacts between two highly abundant amino acids, even if they have no preference for contact with each other. To deal with this, the observed number of contacts for each amino acid

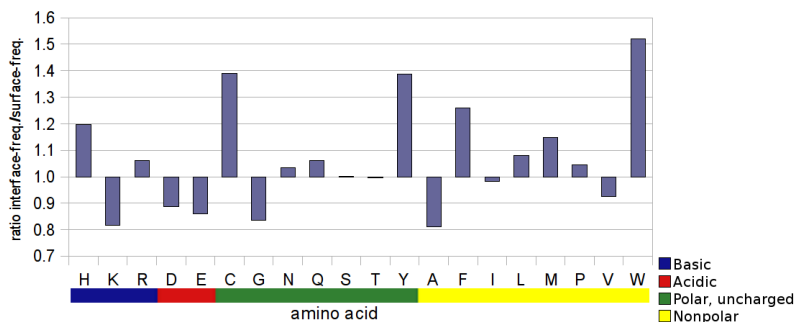


Fig. 3: Ratio of interface-frequency and surface-frequency for the different amino acids in transient interactions. Values  $>1$  indicate interface preferences, values  $<1$  surface preferences.

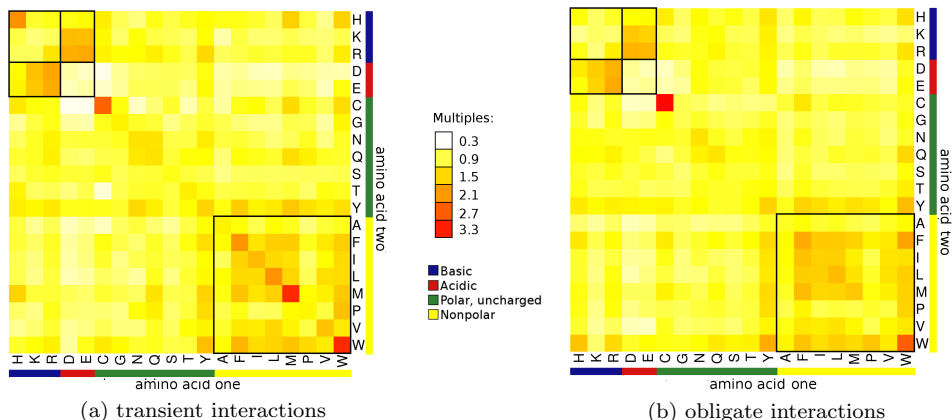


Fig. 4: Multiples of the expected values that were observed for all amino acid contacts. The amino acids are grouped according to their physicochemical properties, some groups are highlighted by black rectangles.

pair is normalized with the expected number of contacts between the participating amino acids as described in the methods section.

We created these heat maps with normalized values for both kinds of interfaces (Fig. 4). The amino acids in the figures are grouped into basic, acidic, polar uncharged and nonpolar based on their physicochemical properties. In both maps contact preferences between some of these groups can be observed. The most striking examples are the groups of acidic and basic amino acids. Amino acids in these groups are only rarely in contact with other amino acids of the same group. However, they show a strong tendency to come into contact with amino acids of the other group. The only exception in these groups is Histidin, which does not prefer contacts to acidic amino acids, and shows a strong tendency to be in contact with itself in transient interfaces.

Another clear tendency is, that nonpolar amino acids are preferably in contact with other nonpolar amino acids.

As the contacts in the interfaces are crucial, an effect on the conservation of interface residues can be expected. Residue contacts in interfaces take place in close spatial proximity and preferably between residues that fit well together according to their physicochemical properties, as shown above. Because of these two factors the mutation rate can be expected to be lower in interfaces than on the surface.

As a matter of fact interfaces tend to be on average more conserved than surfaces. The average conservation of interfaces was 5.6 and 5.8 in the two interface groups. In contrast the average conservation of the surfaces was 5.2 and 5.1. So the overall conservation is higher in interfaces.

By looking at the ratio of the interface- and surface-frequencies of differentially conserved residues, the difference becomes even clearer (Fig. 5). Residues that are

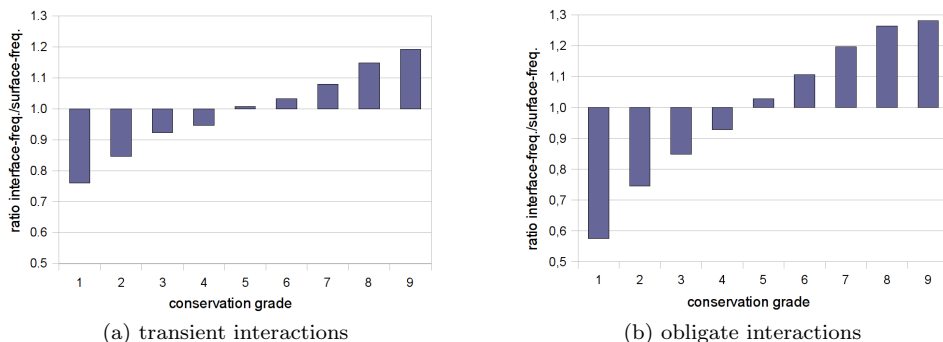


Fig. 5: Ratio of interface-frequency and surface-frequency for the residues of each conservation grade. Values  $>1$  indicate interface preferences, values  $<1$  surface preferences.

very variable (conservation grade 1) are more frequent in the surface than in the interface. In contrast, highly conserved residues (conservation grade 9) are more frequent in the interface. For the conservation grades between these extreme values the preference is shifting from surface to interface with rising conservation grade. The surface/interface preferences are stronger in obligate interfaces than in transient ones.

We analyzed, how the contact preferences of residues are affected by their conservation. In Fig. 6 the observed numbers of all possible combinations between residues of specific conservation grades are shown as multiples of the expected values. There is a clear tendency for residues, to be in contact with other residues that share the same conservation grade. Generally, a contact between residues becomes more and more unlikely with growing difference in their conservation grade.

Additionally there are peaks where very conserved and very variable residues interact with equally conserved/variable residues. The peak for very conserved residues interacting with each other supports previous findings that certain regions inside interfaces are crucial for interaction. The described pattern is even stronger for obligate interfaces than for transient ones.

#### 4. Discussion

The described results show, that physicochemical properties of amino acids have a considerable influence on their contact preferences in interfaces. When amino acids are grouped by their physicochemical properties, a general pattern of preferences and aversions for contacts between members of the different groups can be observed. Additionally, there are some remarkable contacts between single amino acid pairs. The most striking example is Cysteine-Cysteine, which shows the highest preference of all contacts. We guess the reason for this is, that Cysteines are able to form disulfide bonds with each other. These bonds can help to strengthen

reversible and stable interactions. However, Cystein-Cystein contacts should be more frequent in obligate interfaces than in transient interfaces on the surfaces. Indeed there are 2.6 times more Cystein-Cysteine contacts than expected in the transient interfaces. In obligate interfaces this ratio rises to 3.5.

Another set of amino acids, that show a remarkable behavior, are the aromatic amino acids Tyrosine, Phenylalanine and Tryptophan. These amino acids show strong preferences to be in contact with other amino acids from the set. A plausible explanation for this is, that aromatic amino acids can interact via aromatic stacking. This may also be the reason for the high frequency of Histidin-Hisitdin contacts on transient interfaces.

The second focus of this work was to look for relations between the contacts in interfaces and evolutionary constraints. We observed, that conserved residues are more likely to be in an interface than on the rest of the surface. This preference gets weaker with decreasing conservation and turns into a surface preference for variable residues.

This observation is congruent with previous findings [15]. Interfaces are regions where the participating domains or proteins interact with each other specifically and selectively. Therefore, the interaction partners have to be able to identify each other by having complementary interaction sites. To achieve high levels of selectivity and specificity, the interacting sites have to fit together well. This enforces a high conservation of interaction sites as mutations may inhibit the binding of interaction partners.

The preferences of variable residues for the surface and of conserved ones for the interface are stronger in transient interfaces than in obligate ones. This supports previous findings by Mintseris and Weng [16], that obligate interfaces evolve slower than transient ones. Mintseris and Weng argument, that obligate interactions are

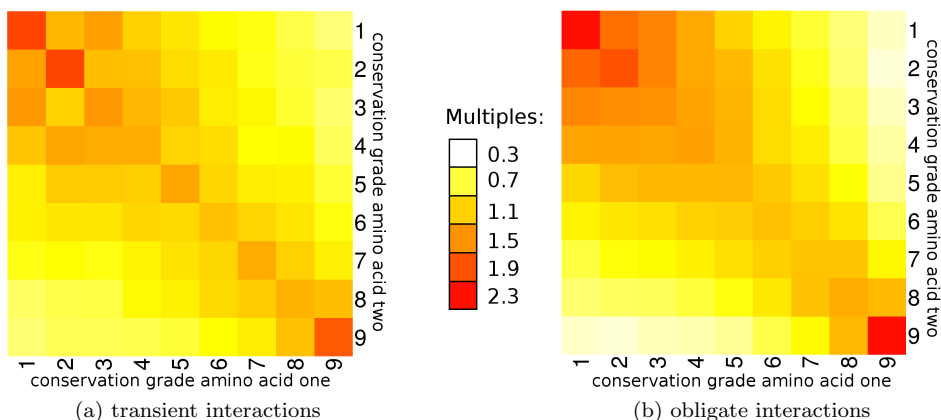


Fig. 6: Multiples of the expected values that were observed for contacts between residues of specific conservation grades.

less frequently mutating, giving the interaction partner the chance to coevolve. On the opposite transient interfaces have to be able to adapt to changes of the interaction partner rather fast, leading to a lower conservation.

Additionally, we examined, whether the residues' conservations are related to their probability of being in contact with each other. There is a very clear picture, that residues prefer to be in contact with other residues that have the same grade of conservation. That means, that conserved residues are not randomly distributed on both sides of the interface. Very variable and very conserved residues especially prefer to interact with equally conserved residues. This finding supports the theory, that a few residues in an interface are crucial for interaction [14]. These residues have to be present on both sides of the interface, therefore they have to be strictly conserved in order to maintain the interaction. On the other hand, very variable residues also show a strong preference for contact with each other. The argument for this follows from the one above. If a residue-residue contact is important, both participating residues have to be conserved. If one residue of a contact is very variable, the contact seems to be rather unimportant for the interaction. Therefore the other residue can be variable, too.

This also explains why residues, whose interaction grades are very different only interact about half as often as expected. This combination is not reasonable as the conservation of just one partner does not ensure that the interaction will always be working.

## 5. Conclusion

We presented properties that can help to identify protein binding sites on the surface of proteins. This information can be used to locate functionally important structural polymorphisms or to determine the near native model in a set of protein docking decoys. Additionally, characteristics of the residue contacts between domains were systematically analyzed and are shown as contact maps. They may be valuable for further refinement of protein-protein complex models.

Although obligate interactions are more conserved than transient interactions, they share most of their physicochemical properties. Therefore, obligate interactions can be used as models for protein-protein interactions.

## Acknowledgements

This work was supported by the International Research Training Group Boston-Kyoto-Berlin, funded by the German Research Foundation (DFG).

## References

- [1] Argos, P., An investigation of protein subunit and domain interfaces, *Protein Eng.*, 2(2):101–113, 1988.
- [2] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E., The protein data bank, *Nucleic Acids Res.*, 28(1):235–242, 2000.

- [3] Chothia, C., Janin, J., Principles of protein-protein recognition, *Nature*, 256(5520):705–708, 1975.
- [4] Chakrabarti, P., Janin, J., Dissecting protein-protein recognition sites, *Proteins*, 47(3):334–343, 2002.
- [5] Chen, R., Li, L., Weng, Z., Zdock: an initial-stage protein-docking algorithm, *Proteins*, 52(1):80–87, 2003.
- [6] Clackson, T., Wells, J. A., A hot spot of binding energy in a hormone-receptor interface, *Science*, 267:383–386, 1995.
- [7] Comeau, S. R., Gatchell, D. W., Vajda, S., Camacho, C. J., Cluspro: a fully automated algorithm for protein-protein docking, *Nucleic Acids Res.*, 32(Web Server issue):W96–99, 2004.
- [8] Gerstein, M., Tsai, J., Levitt, M., The volume of atoms on the protein surface: calculated from simulation, using voronoi polyhedra, *J. Mol. Biol.*, 249(5):955–966, 1995.
- [9] Goldenberg, O., Erez, E., Nimrod, G., Ben-Tal, N., The consurf-db: pre-calculated evolutionary conservation profiles of protein structures, *Nucleic Acids Res.*, 37(Database issue):D323–327, 2009.
- [10] Günther, S., May, P., Hoppe, A., Frömmel, C., Preissner, R., Docking without docking: Isearch–prediction of interactions using known interfaces. *Proteins*, 69(4):839–844, 2007.
- [11] Günther, S., von Eichborn, J., May, P., Preissner, R., Jail: a structure-based interface library for macromolecules, *Nucleic Acids Res.*, 37(Database issue):D338–341, 2009.
- [12] Hubbard, S. J., Argos, P., Cavities and packing at protein interfaces, *Protein Sci.*, 3(12):2194–2206, 1994.
- [13] Jones, S., Thornton, J. M., Analysis of protein-protein interaction sites using surface patches, *J. Mol. Biol.*, 272(1):121–132, 1997.
- [14] Ozlem Keskin, O., Ma, B., Nussinov, R., Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues, *J. Mol. Biol.*, 345(5):1281–1294, 2005.
- [15] Ma, B., Elkayam, T., Wolfson, H., Nussinov, R., Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces, *Proc. Natl. Acad. Sci. U.S.A.*, 100(10):5772–5777, 2003.
- [16] Mintseris, J., Structure, function, and evolution of transient and obligate protein-protein interactions, *Proceedings of the National Academy of Sciences*, 102(31):10930–10935, 2005.
- [17] Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C., Scop: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol.*, 247(4):536–540, 1995.
- [18] Pawson, T., Nash, P., Assembly of cell regulatory systems through protein interaction domains, *Science*, 300(5618):445–452, 2003.
- [19] Pupko, T., Bell, R. E., Mayrose, I., Glaser, F., Ben-Tal, N., Rate4site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues, *Bioinformatics*, 18 Suppl 1:S71–77, 2002.
- [20] Shulman-Peleg, A., Shatsky, M., Nussinov, R., Wolfson, H. J., Spatial chemical conservation of hot spot interactions in protein-protein complexes, *BMC Biol.*, 5:43, 2007.
- [21] Stumpf, M. P. H., From the cover: Estimating the size of the human interactome, *Proceedings of the National Academy of Sciences*, 105(19):6959–6964, 2008.