19

Computational Tools and Databases for the Study and Characterization of Protein Interactions

Jose Ramon Blas[1], Joan Segura[2] and Narcis Fernandez-Fuentes[2,3]
[1]*Universidad de Castilla-La Mancha*
[2]*University of Leeds*
[3]*Aberystwyth University*
[1]*Spain*
[2,3]*United Kingdom*

1. Introduction

One of the most pressing challenges in the post genomic era is the characterization and charting of protein-protein interactions (PPIs) in living organisms, as these are essential in the shaping of normal and pathological behaviours in cells. It is for this reason that unravelling the nature of PPIs has been the pursuit of many experimental techniques, ranging from high-throughput to high-detail approaches (Shoemaker and Panchenko 2007), as well as a wide spectrum of computational prediction methods. Current estimations of human interactome size range from 100,000 to more than 600,000 interactions (Bork et al. 2004; Stelzl and Wanker 2006; Stumpf et al. 2008; Venkatesan et al. 2009). Experimental strategies have reached their best at describing around 50,000 interactions by collating a large number of small and very focused experiments with high-throughput ones, such as massive yeast two-hybrid (Rual et al. 2005; Stelzl et al. 2005), or mass spectrometry coupled to affinity purification experiments (Ewing et al. 2007; Hubner et al. 2010). The smallest gap between experimentally validated and theoretically predicted PPIs amounts to around 50% of total interactions, being probably much higher. When it comes to to studying PPIs in other species on which, even having been sequenced, experimental data is even more scarce, the need for PPI-map completeness is even more notorious. Computational prediction and characterization of PPIs, with its drawbacks, successes and challenges, constitutes a valuable aid in the way to a complete description of interactomes, hence being a promising research field that has enriched our image of living cells for some time now.

Computational tools can provide useful information at different levels of resolution and this chapter seeks to present an up-to-date and comprehensive review of these. The first part of the chapter presents the theoretical basis of computational tools designed to predict PPIs. The main aim of these tools is to predict whether two proteins A and B can interact, either directly or indirectly (functional associations), but without dwelling on the molecular details of the interaction, i.e. which proteins interact. These predictions are useful as complement to large-scale experimental analyses, either to confirm observed interactions or discard false

positives, and also to uncover novel interactions. The second part of the chapter is devoted to the computational methods developed to predict protein interfaces. At this level, predictions identify specific regions and residues of the protein that are likely to mediate PPIs. Thus, these methodologies uncover a higher level of detail, i.e. *how* proteins interact, and have a number of applications in experimental work such as guiding the mapping of protein interfaces by mutagenesis or structural modelling of protein complexes. A special emphasis will be given to a novel and highly accurate tool: VORFFIP(Segura et al. 2011). The concluding part of the chapter describes computational tools developed to predict the important regions or *hot spots* in protein interfaces. Recent successes in the quest for finding new therapeutic agents to modulate PPIs have been aided by the realization, following the pioneering work by Clackson and Wells(Clackson and Wells 1995), that the binding energy of many PPIs can be ascribed to a small and complementary set of interfacial residues: a *hot spot* of binding energy. Thus, identifying these critical residues by computational means has clear applications in drug discovery and in some aspects of protein design. PCRPi(Assi et al. 2010), a novel and highly precise tool will be discussed.

2. Prediction of protein-protein interactions

With the aim of detailing a complete protein interaction map that agglutinates the rising amount of genomic data, high-throughput experimental techniques have walked in parallel with computational approaches. There are six basic computational approaches to predict PPIs depending on the nature of the information used for the prediction. These include PPIs inferred from: (i) genomic context including phylogenetic profiles, gene neighbouring analyses and gene fusion events; (ii) co-evolution events; (iii) protein domain co-occurrence (or signatures) between pair of proteins; (iv) text mining; (v) transference of annotation between species: protein-protein interologs; and (vi) structural annotation including homology-based or *ab initio*. Figure 1 depicts an overall diagrammatic description of these basic approaches and tables 1 and 2 compile a number of on line databases and computational tools respectively.

2.1 Genomic context methods

Biological processes subjected to evolutionary pressure tend to cluster together all interrelated molecular actors in single units to simplify control mechanisms and thus avoid the lost of any essential component. This principle, which operates from maintaining bacterial operon systems to more sophisticated co-regulation strategies found in eukaryotes, is on the basis of genomic context based methods for the detection of functional PPIs.

The first group of genome context methods are based in the comparison of phylogenetic profiles. A phylogenetic profile is the presence or absence of a given gene across N species that can be expressed as an N-dimensional array of ones and zeroes. Originally, functional relationship was assumed if having similar phylogenetic profiles(Pellegrini et al. 1999); however, positive results were limited to very strong interactions and many relations between analogous proteins were missing. Further improvements were made by discarding overlaps given by chance(Wu et al. 2003), using protein domains instead of full length proteins(Pagel et al. 2004), through a concurrent search of multiple independent phylogenetic events of gain/loss of pairs of genes to discard spurious correlated

patterns(Barker and Pagel 2005), or the use of enhanced representation of phylogenetic trees(Ta et al. 2011). The second group of genome context methods is based on gene closeness among different genomes, considering closeness as a sign of functional relatedness. After some initial successes(Koonin et al. 2001; Evguenieva-Hackenberg et al. 2003) and despite some improvements such as allowing for changes in gene order and orientation(Szklarczyk et al. 2011), large-scale predictions should be considered cautiously. Finally, gene fusion approaches are a group of computational tools based on the evidence that some interacting proteins have orthologous where both proteins appear fused in a single protein. Thus, it has been observed that many of these pairs, fused in single proteins in other organisms, correspond to binding partners or at least functionally related proteins(Marcotte et al. 1999; Yanai et al. 2001). Rosetta method(Marcotte et al. 1999) and other implementations(Enright et al. 1999) exploit gene fusion events as predictors of PPIs.

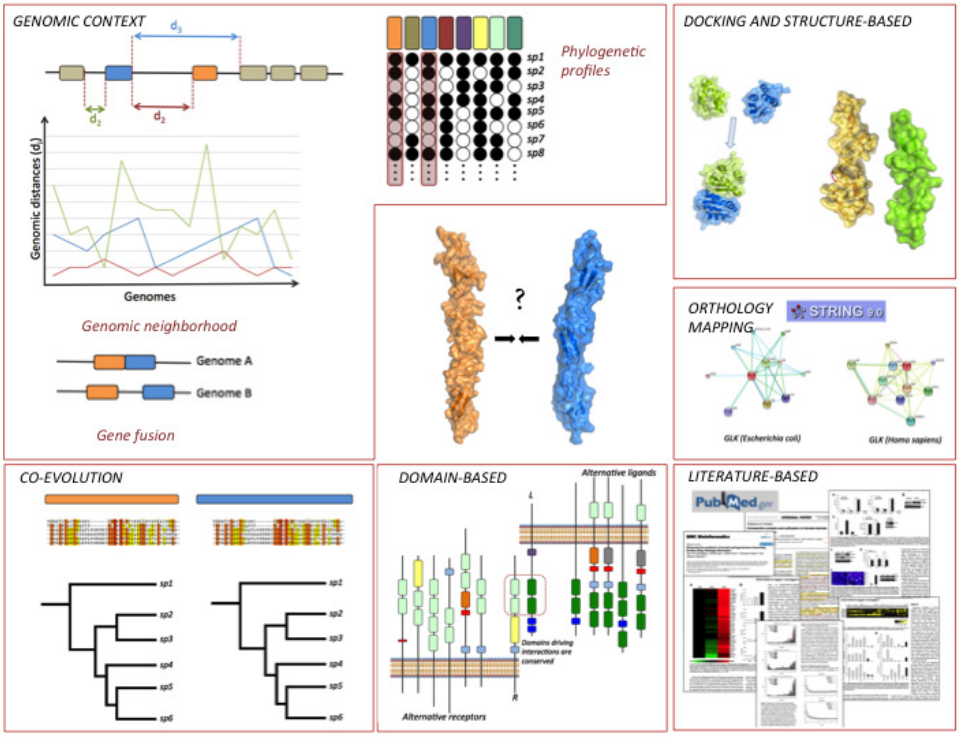

Fig. 1. Graphical description of the main strategies of PPIs prediction described in sections 2.1 to 2.6.

2.2 Co-evolution methods

Two proteins that share a functional relationship, either through direct interaction or functional association, may present evidences of co-evolution. Since the seminal work of Altschuh identifying correlated amino acid changes in Tobacco mosaic virus(Altschuh et al. 1987), studies recognizing co-evolution as an indicative, albeit subtle, signal of PPIs have

being reported in the literature (Travers and Fares 2007; Chao et al. 2008; Presser et al. 2008). Co-evolutionary information may be divided into three groups: the simultaneous loss or gain of orthologous genes(Marcotte et al. 1999), correlated changes affecting both interacting partners at whole sequence level (explored by *mirrortree*-based approaches)(Goh and Cohen 2002; Hakes et al. 2007; Juan et al. 2008) or single amino acids changes(Mintseris and Weng 2005; Madaoui and Guerois 2008).

In the case of mirrortree-based methods (e.g.(Ochoa and Pazos 2010)), the likelihood of interaction is measured as a correlation value between the phylogenetic trees of two families of proteins. Although these approaches have been successfully applied in PPIs prediction(Labedan et al. 2004; Dou et al. 2006; McPartland et al. 2007; Juan et al. 2008), it is still a major problem distinguishing between co-evolution arising from a direct PPI, what has been termed as *co-adaptation* (Pazos and Valencia 2008), from non-specific changes and thus not necessary driven by a functional relatedness(Lovell and Robertson 2010). Recent advances in this area include MatrixMatchMaker algorithm(Tillier and Charlebois 2009) and a faster implementation suitable for large-scale analyses(Rodionov et al. 2011).

The detection of site-specific co-evolution events reflecting PPIs, despite being more intuitive and informative, is even more challenging given the complexity of the mixed evolutionary-structural scenario involved. A single point mutation might ease or complicate each imaginable path of mutation at any other position in the complex, regardless of its distance from the interface(Lovell and Robertson 2010). In fact, co-evolution events have been detected affecting sites that are distant structurally (Gobel et al. 1994; Clarke 1995; Gloor et al. 2005; Fares and McNally 2006). On the other hand, the probability of correlated amino acid changes is closely related to the chemical nature of changes. In this sense, volume variations seem to strongly affect fitness, and so they are frequently balanced by evolution machinery (up to almost 50% of the cases)(Williams and Lovell 2009). Moreover, interface residues in obligate complexes evolve at a slower rate than those in transient interactions(Mintseris and Weng 2005).Taken together, all these particulars illustrate the challenges encountered when looking for site-specific co-evolution events related to PPIs. Recent developments have looked at improving the discrimination between direct and indirect correlations(Burger and van Nimwegen 2010), or including amino acid background distribution information and the mutual information of residues physicochemical properties(Gao et al. 2011). However, new, more discriminative, approaches are required to better understand co-evolution at residue-centred level.

2.3 Domain-based methods

There are strong evidences supporting the idea that the range of different PPIs can be accounted for by considering a more reduced set of specific domain-domain interactions, domain signatures, that are even conserved across different species(Finn et al. 2006; Itzhaki et al. 2006; Stein et al. 2011). Thus, the basis of domain-based methods is presence/absence of given domain signatures between pairs of proteins that can be used to infer interaction. An early method exploiting domain signatures was an association method where domain interactions were assumed if the frequency of association was higher than the expected frequency(Kim et al. 2002). Further improvements have been devised to improve predictions including the domain pair exclusion analysis, which implemented a new scoring

scheme(Riley et al. 2005), the use of Random Forest ensemble classifiers to deal with the pairing of multi-domain proteins(Chen and Liu 2005) or the use of Gene Ontology(Lee et al. 2006) or co-evolution data(Jothi et al. 2006).

2.4 Literature-based data mining methods

Numerous research efforts have been focused on automatically extracting and analysing information from the scientific literature in order to infer putative PPIs(Blaschke et al. 2001; Fundel et al. 2007; Airola et al. 2008). These include, the search for the co-occurrence of terms(Blaschke et al. 2001) or the presence of similar Gene Ontology terms(Pesquita et al. 2009) or kernel-based methods including subsequence kernels, tree kernels, shortest path kernels and graph kernels(Tikk et al. 2010). The most recent approaches use multiple kernels to maximize the information extracted from scientific papers(Kim et al. 2008; Miwa et al. 2009), the combination of multiple kernels and machine learning algorithms to improve the scoring(Yang et al. 2011), or the more recent neighbourhood hash graph kernels that are substantially faster than previous text-mining approaches(Zhang et al. 2011).

Name	URL	Reference
STRING	http://string-db.org	(Szklarczyk et al. 2011)
BioGRID	http://thebiogrid.org/	(Stark et al. 2011)
IntAct	http://www.ebi.ac.uk/intact/	(Aranda et al. 2010)
HPRD	http://www.hprd.org/	(Prasad et al. 2009)
HitPredict	http://hintdb.hgc.jp/htp/	(Patil et al. 2011)
DIP	http://dip.doe-mbi.ucla.edu/dip	(Salwinski et al. 2004)
MINT	http://mint.bio.uniroma2.it/mint/	(Chatr-aryamontri et al. 2007)
TAIR	www.arabidopsis.org/portals/proteome/	(Swarbreck et al. 2008)
iPFAM	http://ipfam.sanger.ac.uk/	(Finn et al. 2005)
3DID	http://3did.irbbarcelona.org/	(Stein et al. 2011)
DIMA 3.0	http://webclu.bio.wzw.tum.de/dima/	(Luo et al. 2011)
DOMINE	http://domine.utdallas.edu/cgi-bin/Domine	(Yellaboina et al. 2011)
GWIDD	http://gwidd.bioinformatics.ku.edu	(Kundrotas et al. 2010)
IsoBase	http://isobase.csail.mit.edu/	(Park et al. 2011)
I2D	http://ophid.utoronto.ca/ophidv2.201	(Brown and Jurisica 2007)
DroID	http://www.droidb.org	(Murali et al. 2011)
HCPIN	http://nesg.org:9090/HCPIN	(Huang et al. 2008)
HIV1,HPID	http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions	(Fu et al. 2009)
MPIDB	http://www.jcvi.org/mpidb/about.php	(Goll et al. 2008)

Table 1. List of major databases compiling experimentally determined or computationally predicted PPIs.

2.5 Orthology mapping (Interologs) methods

The basis of these methods is the transference of annotated interactions between organisms; hence the term *interologs* to refer to predicted homologous interactions(Walhout et al. 2000; Shoemaker and Panchenko 2007; Lewis et al. 2010). Interolog annotations have been successfully applied to transfer experimentally known interactions in yeast to predicted ones in worm(Matthews et al. 2001) and between mouse and human(Huang et al. 2007). Although some improvement has been devised such as scoring schemes that depend on the sources of experimental data(Jonsson and Bates 2006), the applicability of orthology mapping is limited. Firstly, accurate predictions require high sequence similarities between interologs (~70%)(Mika and Rost 2006) thus limiting its range of applicability. Secondly, even at high sequence identity level, in some cases small variations in protein sequence at the interface have been shown to dramatically change PPI specificity, thus redefining complex protein networks and leading to important phenotypic differences(Panni et al. 2002; Kiemer and Cesareni 2007).

2.6 Structure-based methods

A final category of computational methods includes those based in structural information. The structure of a protein complex formed by two or more proteins can be modelled using the structure of a known protein complex as template either by homology modelling or threading(Lu et al. 2002; Aloy et al. 2004; Hue et al. 2010). Even in the absence of a suitable template, the structure of the complex can be modelled by using protein docking(Wass et al. 2011) and selecting the protein complex based on predicted binding energy, i.e. *ab initio* modeling. Despite being a promising strategy, and without considering the high computational cost, the correlation between predicted and experimentally measured binding affinities, such as K_d, is very low thus greatly impairing its predictive power (Kastritis and Bonvin 2010; Stein et al. 2011). Other strategies combine structural data, docking and evolutionary conservation(Tuncbag et al. 2011).

Name	Methodology	URL	Reference
MirrorTree	Co-evolution	http://csbg.cnb.csic.es/mtserver/	(Ochoa and Pazos 2010)
MatrixMatchMaker	Co-evolution	http://www.uhnresearch.ca/labs/tillier/	(Tillier and Charlebois 2009)
iHOP	Text mining	http://www.ihop-net.org/	(Hoffmann and Valencia 2004)
PathBLAST	Orthology	http://www.pathblast.org/	(Kelley and Ideker 2005)
InterPreTS	Structure-based	http://www.russelllab.org/	(Aloy and Russell 2003)
IBIS	Structure-based	http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi	(Shoemaker et al. 2010)

Table 2. List of on line resources for the prediction of PPIs.

3. Prediction of protein binding sites

As indicated by its name, binding site prediction methods seek to define the regions in proteins that are more likely to mediate PPIs. The level of resolution is therefore higher and the starting point is either the sequences or structures of proteins that are known to interact (i.e. experimental evidence) but for which no structural details of the interaction are known.

3.1 Distinctiveness of interface residues

Large-scale analyses of the structures of protein complexes have shown that residues located in interfaces present a number of differential physicochemical and structural qualities. In general, hydrophobic residues are overrepresented in the interfaces of permanent complexes(Lo Conte et al. 1999; Glaser et al. 2001) and charge residues, Arg in particular, are also commonly found in interfaces and often define the lifetime of complexes(Zhou and Shan 2001). A higher accessibility to the solvent than exposed residues not located in interfaces is also a differential trait of interface residues(Chen and Zhou 2005), being the most effective feature to predict interfaces in homodimeric complexes(Jones and Thornton 1997). On the other hand and in agreement with earlier observations that found interface residues have lower crystallographic B-factors(Neuvirth et al. 2004), the side chains of interface residues are less likely to sample alternative rotamers, i.e. more rigid, to decrease the entropic cost upon complex formation(Cole and Warwicker 2002; Fleishman et al. 2011). Sequence conservation has also proved to be a predictor(Lichtarge et al. 1996; Wang et al. 2006), although it remains a contentious issue as some works have shown that interfaces are not more conserved than the rest of the protein(Grishin and Phillips 1994; Caffrey et al. 2004). Finally, it has been shown that interfaces are richer in β-strands and long loops while α-helical conformations are disfavoured(Neuvirth et al. 2004).

3.2 Prediction methods

Prediction methods rely on sequence and/or structural information that is unique to interface residues (see before). Hence, prediction methods can be divided into two groups: sequence-based methods, which rely only on the primary sequence of the protein and structure-based methods that require the three-dimensional structure of the protein. Table 3 compiles a list of on line computational tools to predict protein binding sites.

3.2.1 Structure-based prediction methods

One of the first structure-based prediction methods, later updated(Murakami and Jones 2006), was based on surface patch analysis(Jones and Thornton 1997). Surface patches were defined by grouping neighbouring exposed residues that were subsequently ranked using a scoring function that included the solvation potential, interface propensity, hydrophobicity, protrusion and accessible surface area of each of the residues within the patch. A probabilistic approach, ProMate, also based on patch analysis, was developed for heteromeric transient protein complexes by combining secondary structure content, hydrophobicity and crystallographic B-factors information(Neuvirth et al. 2004). The combination of ProMate's predictions and a parametric scoring function based of sequence conservation and structural features resulted in an improvement of the accuracy of the predictions(de Vries et al. 2006). Other implementations of prediction methods include an

empirical scoring function composed of side chain energy score, residue conservation and interface propensity(Liang et al. 2006), the search of structural interaction templates extracted from protein complexes(Chang et al. 2006) and a clustering algorithm that identifies residues with a high propensity of being located in interfaces(Negi et al. 2007).

In order to combine and integrate heterogenous data, i.e. sources of information of a different nature (e.g. hydrophobicity indexes and solvent accessibility surface) into a common and coherent scoring framework, a number of machine learning methods have been proposed including Neural Networks (NN)(Fariselli et al. 2002; Chen and Zhou 2005; Porollo and Meller 2007), Support Vector Machines (SVM)(Bradford and Westhead 2005), Random Forests (RF)(Sikic et al. 2009; Segura et al. 2011) and Bayesian Networks (BN)(Bradford et al. 2006; Ashkenazy et al. 2010). Thus, the commonality of these approaches is the use of a machine-learning algorithm (NN, SVM, RF or BN) to combine a set of sequence- and structural-based measures into an unified score or probability. The nature of the combined features used by the prediction methods includes: evolutionary conservation and surface disposition(Fariselli et al. 2002); sequence conservation, electrostatic potentials, SASA, hydrophobicity, protusion and interface propensity(Bradford and Westhead 2005; Bradford et al. 2006); properties taken from the AAIndex database(Kawashima et al. 2008) (e.g. expected number of contacts within 14 Å sphere), multiple sequence alignment-derived features (e.g. amino acid frequency), and structural features(Porollo and Meller 2007); structure-based, energy terms, sequence conservation and crystallographic B-factors(Segura et al. 2011); structural features, sequence and secondary structure(Sikic et al. 2009); or more complex approaches that combine several prediction methods in the form of a meta-prediction(Qin and Zhou 2007; Ashkenazy et al. 2010).

3.2.2 Sequence-based prediction methods

Even if the structure of the protein is not available, there are still a number of prediction methods that are based solely on sequence information. Early examples of approaches in this category include a NN (Ofran and Rost 2003) that uses local sequence information, which was subsequently improved by including a post-neural network filtering step(Ofran and Rost 2007). Other approaches include SVMs that combine sequence profiles and other sequence-based information such as spatially neighbouring residues(Koike and Takagi 2004; Res et al. 2005; Chen and Li 2010), a RF that integrates physicochemical properties of residues, evolutionary conservation and amino acid distances(Chen and Jeong 2009), and a naive Bayesian classifier trained to integrate position-specific scoring matrix and predicted accessibility(Murakami and Mizuguchi 2010). Finally, other sequence-based methods have been developed to improve prediction by tacking issues such as the problem of unbalanced data in protein sets(Yu et al. 2010), i.e. the interface accounts for a small proportion of the exposed residues so the number of negative cases (non-interface residues) is much larger than the number of positive cases (interface residues) or improving the sampling(Engelen et al. 2009) in evolutionary trace-based(Lichtarge et al. 1996) methodologies.

3.3 VORFFIP, a holistic approach to predict protein binding sites in protein structures

VORFFIP is a novel, structure-based, method that integrates a wide range of residue-based features and environment information using a 2-step Random Forest ensemble

classifier(Segura et al. 2011). Residue-based features include structural-based, energy terms, evolutionary conservation and crystallographic B-factors information. VORFFIP implements a novel definition of local environment by means of Voronoi Diagrams (see next and Fig. 2) that complements residue-based information improving the accuracy of predictions.

Residue-based information characterizes individual residues. Structure-based features account from 16 different features and define the local geometry of the protein at residue level. Structural features include, among others, the absolute and relative accessibility surface area, the protrusion index that is a measure of the local concavity/convexity and a deepness index(Vlahovicek et al. 2005). The energetic state of exposed residue is characterized by 10 energy terms including electrostatic potential, solvent exposure energy, entropy and hydrogen bond energy among others(Guerois et al. 2002). The sequence conservation of residues consist of the regional conservation score that defines the conservation for each residue and its neighbourhood in the 3D space(Landgraf et al. 2001) and a sequence positional score calculated from multiple sequence alignment profiles(Pei and Grishin 2001). Finally, crystallographic B-factors, which are a measure of thermal motion, are converted to Z-score as described previoulsy(Yuan et al. 2003).

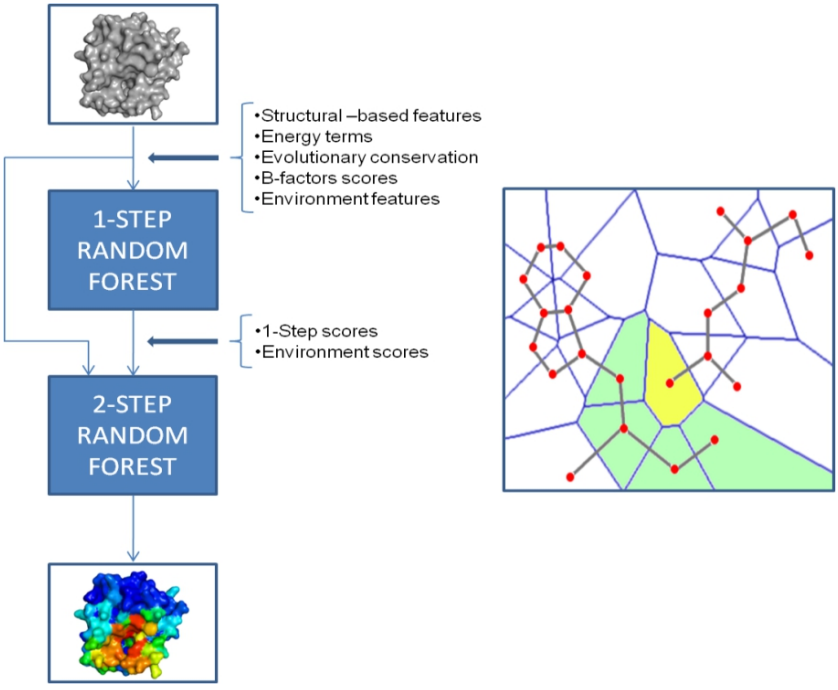

Fig. 2. Overview of prediction process in VORFFIP and a Voronoi Diagram of a interacting pair. The left side of the figure illustrates the 2-step prediction approach in VORFFIP. The right side of the figure shows the Voronoi Diagram of two neighbouring residues; heavy atoms are represented by red dots and coloured cells illustrate interaction between atoms of neighbouring residues.

Environment-based information accounts the local structural environment of residues. Interfaces tend to form contiguous patches on the surface and thus, the environment of a residue can provide valuable information for predictions. Several methods have been used to account for the local environment of residues including sliding window (e.g.(Ofran and Rost 2003)) and Euclidian distances (e.g. (Porollo and Meller 2007)). VORFFIP however uses a novel definition of environment by means of Voronoi Diagrams (VD). VD is computed using the heavy atoms coordinates as seeds and as a result the 3D space is partitioned into polyhedral cells where each single cell contains one of the atoms (Barber et al. 1996). Atoms sharing a common facet in the VD are said to be in contact or neighbours, i.e. part of the local environment. Figure 2 shows a 2D representation of a VD diagram depicting the interaction between atoms of two neighbouring residues. The number of contacts between

Name	Input	Method	URL	Reference
VORFFIP	Structure	RF	http://www.bioinsilico.org/VORFFIP	(Segura et al. 2011)
ProMate	Structure	Scoring function	http://bioinfo.weizmann.ac.il/promate	(Neuvirth et al. 2004)
ISIS	Sequence	NN	http://rostlab.org/cms/resources/web-services/	(Ofran and Rost 2007)
WHISCY	Structure	Scoring function	http://nmr.chem.uu.nl/Software/whiscy	(de Vries et al. 2006)
PPI-pred	Structure	SVM	http://www.bioinformatics.leeds.ac.uk/ppi_pred	(Bradford and Westhead 2005)
SPPIDER	Structure	NN	http://sppider.cchmc.org	(Porollo and Meller 2007)
PINUP	Structure	Scoring function	http://sparks.informatics.iupui.edu	(Liang et al. 2006)
meta-PPISP	Structure	Meta-server	http://pipe.scs.fsu.edu/meta-ppisp.html	(Qin and Zhou 2007)
Protemot	Structure	Scoring function	http://bioinfo.mc.ntu.edu.tw/protemot	(Chang et al. 2006)
InterProSurf	Structure	Scoring function	http://curie.utmb.edu	(Negi et al. 2007)
cons-PPISP	Structure	NN	http://pipe.scs.fsu.edu/ppisp.html	(Chen and Zhou 2005)
PSIVER	Sequence	BN	http://tardis.nibio.go.jp/PSIVER/	(Murakami and Mizuguchi 2010)
SHARP	Structure	Scoring function	http://www.bioinformatics.sussex.ac.uk/SHARP2	(Murakami and Jones 2006)

Table 3. List of online resources for protein binding site prediction.

neighbouring residues is used to derive weights that will be then used to normalize residue-based features among residues within the local environment. The advantage of using VD over other definition of local environment is that there are no requirements with regards cut-off to define the local environment (e.g. a distance cut-off) and that a weighting system can be easily implemented based on the number of interactions (i.e. neighbouring residues) in the VD. When the performance of VORFFIP was assessed in term of type of methods used to define local environment, VD were superior to Euclidean distances and sliding window approaches(Segura et al. 2011).

The final stage of the method is the integration of residue- and environment-based features using a machine learning approach: a 2-steps RF ensemble classifier (Fig. 2), which is also a novel feature as most machine learning methodologies use a single step classifier. In the first-step RF, residue and residue-environment features are calculated and used as input variables. The scores yielded by the first-step RF are then decomposed into a number of new input variables including VD-derived environment scores. Residue and environment scores together with the previously calculated features form the new set of input variables to the second-step RF that will output the final scores. The logic behind using a second-step RF relates to the observation that residues belonging to the same interface tend to form contiguous patches on the surface, i.e. high scoring residues are expected to be neighbouring mainly high scoring residues unless located at the boundaries of the interface. Thus, the second-step RF harmonizes outliers and generates more homogenous scores for interface residues resulting in better predictions as shown by the competitive results obtained(Segura et al. 2011) when comparing to other methods(de Vries et al. 2006; Porollo and Meller 2007; Sikic et al. 2009).

4. Prediction and charting of hot spots in protein interfaces

The final part of the chapter describes the current state in computational prediction of hot spots in protein interfaces. The goal of these methods is the prediction of the region of a given interface that contributes the most to the binding energy of the complex, i.e. the hot spot of the interaction. These methods are a good complement to highly intensive and costing experimental techniques, in particular in large-scale analyses, and have clear applications in drug discovery and protein engineering.

4.1 Distinctiveness of hot spot residues

As in the case of interface residues, hot spot residues present a number of structural and physicochemical properties unique to them and these are exploited by the prediction methods. The first is the type of residues that are commonly found in hot spots: while the proportion of Trp, Arg and Tyr is higher, Leu, Ser and Val are disfavoured(Bogan and Thorn 1998). Likewise, Asn and Asp are more commonly found in hot spots than chemically comparable (but bulkier) Gln and Glu(Bogan and Thorn 1998). Hot spot residues are optimally packed, structurally conserved and usually located in the central part of the interface(Keskin et al. 2005; Yogurtcu et al. 2008). One more characteristic of hot spot residues is that they are often located in complemented pockets, i.e. hot spot residues in one protein interact with hot spot residues of cognate protein(s)(Li et al. 2004). Finally, hot spot residues usually have a higher evolutionary conservation than the rest of the residues in the interface(Guharoy and Chakrabarti 2005).

4.2 Prediction algorithms

A number of computational methods have been developed for the prediction of hot spots in protein interfaces. An important part of these is represented by energy-based methods that predict changes in binding energy upon mutations, i.e. *in silico* alanine scanning. These methodologies range from scoring function derived from simple physical models(Guerois et al. 2002; Kortemme and Baker 2002; Kruger and Gohlke 2010) to more complex, time consuming atomistic simulations to model effect of mutations in the binding energy(Almlof et al. 2006; Lafont et al. 2007; Moreira et al. 2007; Benedix et al. 2009; Diller et al. 2010). Other methods exploit individual features (or combination of them) that are characteristic to hot spots such as solvent accessibility(Landon et al. 2007; Tuncbag et al. 2009; Xia et al. 2010; Li et al. 2011), atomic contacts(Li et al. 2006), structural conservation(Li et al. 2004), restricted mobility(Yogurtcu et al. 2008), relative location of residues in the interface(Keskin et al. 2005), sequence conservation(Hu et al. 2000; Ma and Nussinov 2007) and pattern mining(Hsu et al. 2007). Other examples include a number of machine learning approaches (Darnell et al. 2007; Ofran and Rost 2007; Cho et al. 2009; Lise et al. 2009; Assi et al. 2010) such as PCRPi (see next) that integrate a range of structural- and sequence-based information and a docking-based approach(Grosdidier and Fernandez-Recio 2008).

4.3 PCRPi: *P*resaging *C*ritical *R*esidues in *P*rotein *i*nterfaces, a novel and highly accurate prediction algorithm

While the attributes described in section 4.1 have predictive power, it has been found that individually cannot unambiguously define hot spot residues(DeLano 2002). To overcome this limitation, PCRPi(Assi et al. 2010), a novel computation tool for the prediction of hot spots residues, integrates seven different variables that account for structural, evolutionary conservation and predicted binding energy (Fig.3).

The structural information of interface residues is described by two different variables: the interaction engagement (IE) and the topographical (TOP) indexes. The IE index gauges for the number of inter-chain atomic interactions of the given residue normalized by total number of atoms that can potentially interact. An IE index of 1.0 would indicate that all atoms are actively engaged in atomic interactions with groups of cognate protein(s). The TOP index describes the structural environment of residues and is ratio between the number of neighbouring residues of cognate proteins and the average number neighbouring residues. Neighbouring residues are any residues of cognate protein(s) whose carbon alpha is enclosed in a sphere of 10 Angstroms of radius centered on the carbon alpha of the residue of interest. Thus, TOP index quantifies whether residues are intimately interacting with cognate proteins or are located in a more flat or unprotected region.

The second group of variables used by PCRPi relates to evolutionary conservation. Evolutionary conservation is quantified by looking at the sequence conservation and the 3D regional conservation (i.e. structural conservation of patches) in both target (ANCCON and ANC3DCON) and cognates proteins (CON and 3DCON). To calculate ANCCON and CON values, sequence profiles are derived as described(Fernandez-Fuentes et al. 2007). Next, ANCCON corresponds to conservation scores as calculated by al2co(Pei and Grishin 2001) and the CON variable is the ratio between residues with and al2co scores above 1.0 an the number of cognate residues in the interface. Likewise, the ANC3DCON and 3DCON values

are calculate but instead of using al2co scores, the normalize regional conservation scores as defined by Landgrad et al(Landgraf et al. 2001) are used. The last input used by PCRPi is the BE index, which represents the predicted binding energy change upon mutation, i.e. *in silico* Alanine scanning, as calculated using FoldX(Guerois et al. 2002).

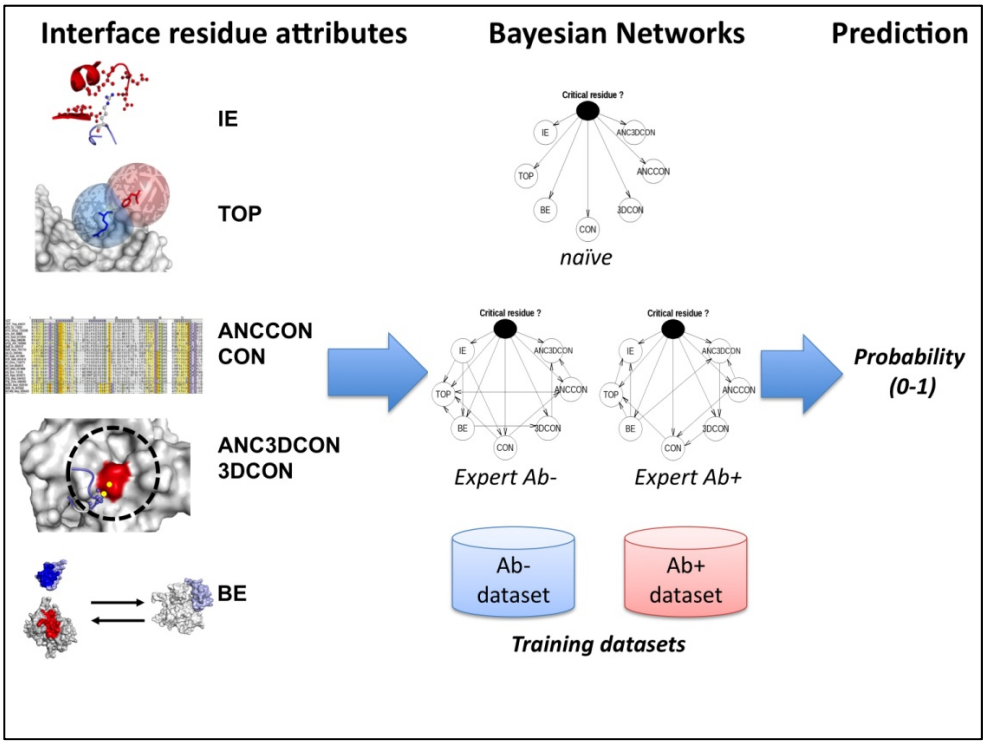

Fig. 3. Overview of the prediction process. PCRPi integrates seven features characterizing interface residues that are used as input variables to three different Bayesian networks, two experts and one naïve, that can be trained with protein complexes including (Ab+) or excluding (Ab-) Antigen-Antibodies complexes. PCRPi outputs a probability where the higher the probability the more likely the residues to be critical, i.e. hot spot residues, for the interaction.

The final part of the prediction is the integration of the data, i.e. IE, TOP, ANCCON, CON, ANC3DCON, 3DCON and BE, into a common probabilistic framework by using BN. PCPRi features three different types BN, two experts and one naïve (Fig. 3). The difference between them is the relationship of dependence between input variables; while naïve BN assumes independence, an expert BN allows conditional dependence between variables (Fig. 3). Both expert and naïve BNs are trained using two specific sets of protein complexes: Ab+ and Ab- (Fig. 3). The Ab+ set corresponds to protein complexes that can include non-evolutionary related complexes such as Antigen-Antibodies complexes while Ab- does not include the latter. The reason being is the lack of sequence conservation in the complementary determining regions of Antibodies, i.e. regions that mediate interaction, which renders

evolutionary information meaningless for prediction purposes and thus special BNs were devised to cope with this problem. In terms of performance, PCRPi delivers highly consistent and competitive predictions as shown in the study of the protein complex formed by RAS and VH-HRAS antibody(Tanaka et al. 2007) and a comprehensive comparative study(Assi et al. 2010). Moreover, PCPRi is a central part of a database that compiles and annotates hot spot in protein interfaces: PCRPI-DB (Segura and Fernandez-Fuentes 2011).

Name	Input	Method	URL	Reference
PCRPi	Structure	Machine learning	http://www.bioinsilico.org/PCRPi	(Assi et al. 2010)
Robetta	Structure	Energy-based	http://robetta.bakerlab.org	(Kortemme and Baker 2002)
FoldX	Structure	Energy-based	http://foldx.crg.es	(Guerois et al. 2002)
DrugScorePPi	Structure	Energy-based	http://cpclab.uni-duesseldorf.de/dsppi	(Kruger and Gohlke 2010)
CC/PBSA server	Structure	Energy-based	http://ccpbsa.biologie.uni-erlangen.de/ccpbsa	(Benedix et al. 2009)
KFC	Structure	Machine learning	http://kfc.mitchell-lab.org	(Darnell et al. 2008)
HotPoint	Structure	Scoring function	http://prism.ccbb.ku.edu.tr/hotpoint	(Tuncbag et al. 2009)
ISIS	Sequence	Neural Network	http://rostlab.org/cms/resources/web-services/	(Ofran and Rost 2007)

Table 4. List of online resources for prediction of hot spots.

5. Conclusions and outlook

During the last years, scientists aiming at understanding living organisms at a molecular level have seen their benches become swapped with the sheer amount of information and this burst of data being mirrored by the development of a wide and miscellaneous set of computational tools designed to unveil biologically relevant information from the noisy background. PPIs are among the most crucial events that define the behaviour of a living system and that explains the rise of research efforts and strategies to describe the nature of PPIs. This chapter presents a summary and extensive view on computational methods devoted to predict which proteins participate in PPIs (section 2), which are the regions involved in the interaction (section 3) and which are the most important regions or residues in the interaction (section 4).

In general the prediction tools achieve a high rate of prediction success and are important tools for scientists. However, there are still a number of unmet needs and challenges to be solved. In the case of prediction of PPIs, genome context approaches would benefit from improved definitions of phylogenetic profiles and the masking effect of gene fusion events. Text-mining approaches require further development to reduce false positive rates and increase efficiency. A deeper understanding of the complex interlink between (bio)chemistry,

structure and genetics that governs the evolution of protein interfaces would certainly benefit co-evolution-based methods. The correct detection of remote homology between interologs is a major challenge as is the lack of correlation between predicted and observed binding affinities in structure-based methods.

Protein binding site prediction methods have also their own limitations and challenges. The physical forces and chemical properties that drive the interaction between proteins are not fully understood and thus current models do not reflect the binding process accurately. However, the increasing amount of experimental data that is being generated in an important factor that plays in favour of developing novel and more accurate computational tools. Some specific challenges in the field are the prediction of binding sites in proteins that recognize multiple partners (hub proteins) and the distinction between each of the interfaces that are relevant to each of the interacting partners. Current methods cannot properly handle binding events that involve conformational changes in any of the intervening components, including those mediated by intrinsic disordered regions, and thus future efforts need to be directed to tackle this very important question. Finally, the main challenge in the prediction of hot spots is the development of new approaches to bridge the gap between highly computationally expensive methods and those based on simplified models by finding the right balance between the accuracy of the former and the speed of the latter.

6. Acknowledgments

NFF acknowledges support from the Research Councils United Kingdom (RCUK) under the Academic Fellowship scheme. JS acknowledges support from the Leeds Institute of Molecular Medicine (PhD scholarship). JR is supported by a postdoctoral grant awarded by the Consejeria de Educacion y Cultura of the Junta de Comunidades de Castilla La Mancha anb by the European Social Fund. NFF also thanks Dr Gendra for critical reading and insightful comments to the manuscript, and Ms Martina and Ms Daniela G Fernandez for continuing inspiration and motivation. Publication costs were funded by The Biomedical and Health Research Centre (BHRC).

7. References

Airola, A., S. Pyysalo, J. Bjorne, T. Pahikkala, F. Ginter &T. Salakoski (2008). "All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning." *BMC Bioinformatics* 9 Suppl 11: S2.

Almlof, M., J. Aqvist, A. O. Smalas &B. O. Brandsdal (2006). "Probing the effect of point mutations at protein-protein interfaces with free energy calculations." *Biophys J* 90(2): 433-42.

Aloy, P., B. Bottcher, H. Ceulemans, C. Leutwein, C. Mellwig, S. Fischer, A. C. Gavin, P. Bork, G. Superti-Furga, L. Serrano &R. B. Russell (2004). "Structure-based assembly of protein complexes in yeast." *Science* 303(5666): 2026-9.

Aloy, P. &R. B. Russell (2003). "InterPreTS: protein Interaction Prediction through Tertiary Structure." *Bioinformatics.* 19(1): 161.

Altschuh, D., A. M. Lesk, A. C. Bloomer &A. Klug (1987). "Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus." *Journal of Molecular Biology* 193(4): 693-707.

Aranda, B., P. Achuthan, Y. Alam-Faruque, I. Armean, A. Bridge, C. Derow, M. Feuermann, A. T. Ghanbarian, S. Kerrien, J. Khadake, J. Kerssemakers, C. Leroy, M. Menden, M. Michaut, L. Montecchi-Palazzi, S. N. Neuhauser, S. Orchard, V. Perreau, B. Roechert, K. van Eijk &H. Hermjakob (2010). "The IntAct molecular interaction database in 2010." *Nucleic Acids Res* 38(Database issue): D525-31.

Ashkenazy, H., E. Erez, E. Martz, T. Pupko &N. Ben-Tal (2010). "ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids." *Nucleic Acids Res* 38(Web Server issue): W529-33.

Assi, S. A., T. Tanaka, T. H. Rabbitts &N. Fernandez-Fuentes (2010). "PCRPi: Presaging Critical Residues in Protein interfaces, a new computational tool to chart hot spots in protein interfaces." *Nucleic Acids Res* 38(6): e86.

Barber, C. B., D. P. Dobkin &H. Huhdanpaa (1996). "The Quickhull algorithm for convex hulls." *ACM TRANSACTIONS ON MATHEMATICAL SOFTWARE* 22(4): 469-483.

Barker, D. &M. Pagel (2005). "Predicting functional gene links from phylogenetic-statistical analyses of whole genomes." *PLoS Comput Biol* 1(1): e3.

Benedix, A., C. M. Becker, B. L. de Groot, A. Caflisch &R. A. Bockmann (2009). "Predicting free energy changes using structural ensembles." *Nat Methods* 6(1): 3-4.

Blaschke, C., R. Hoffmann, J. C. Oliveros &A. Valencia (2001). "Extracting information automatically from biological literature." *Comp Funct Genomics* 2(5): 310-3.

Bogan, A. A. &K. S. Thorn (1998). "Anatomy of hot spots in protein interfaces." *J Mol Biol* 280(1): 1-9.

Bork, P., L. J. Jensen, C. von Mering, A. K. Ramani, I. Lee &E. M. Marcotte (2004). "Protein interaction networks from yeast to human." *Curr Opin Struct Biol* 14(3): 292-9.

Bradford, J. R., C. J. Needham, A. J. Bulpitt &D. R. Westhead (2006). "Insights into protein-protein interfaces using a Bayesian network prediction method." *J Mol Biol* 362(2): 365-86.

Bradford, J. R. &D. R. Westhead (2005). "Improved prediction of protein-protein binding sites using a support vector machines approach." *Bioinformatics* 21(8): 1487-94.

Brown, K. R. &I. Jurisica (2007). "Unequal evolutionary conservation of human protein interactions in interologous networks." *Genome Biol* 8(5): R95.

Burger, L. &E. van Nimwegen (2010). "Disentangling direct from indirect co-evolution of residues in protein alignments." *PLoS computational biology* 6(1): e1000633.

Caffrey, D. R., S. Somaroo, J. D. Hughes, J. Mintseris &E. S. Huang (2004). "Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?" *Protein Sci* 13(1): 190-202.

Chang, D. T., Y. Z. Weng, J. H. Lin, M. J. Hwang &Y. J. Oyang (2006). "Protemot: prediction of protein binding sites with automatically extracted geometrical templates." *Nucleic Acids Res* 34(Web Server issue): W303-9.

Chao, J. A., Y. Patskovsky, S. C. Almo &R. H. Singer (2008). "Structural basis for the coevolution of a viral RNA-protein complex." *Nature Structural and Molecular Biology* 15(1): 103-105.

Chatr-aryamontri, A., A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli &G. Cesareni (2007). "MINT: the Molecular INTeraction database." *Nucleic Acids Res* 35(Database issue): D572-4.

Chen, H. &H.-X. Zhou (2005). "Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data." *Proteins* 61(1): 21-35.

Chen, P. &J. Li (2010). "Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information." *BMC Bioinformatics* 11: 402.

Chen, X.-w. &J. C. Jeong (2009). "Sequence-based prediction of protein interaction sites with an integrative method." *Bioinformatics* 25(5): 585-591.

Chen, X. W. &M. Liu (2005). "Prediction of protein-protein interactions using random decision forest framework." *Bioinformatics* 21(24): 4394-400.

Cho, K. I., D. Kim &D. Lee (2009). "A feature-based approach to modeling protein-protein interaction hot spots." *Nucleic Acids Res* 37(8): 2672-87.

Clackson, T. &J. A. Wells (1995). "A hot spot of binding energy in a hormone-receptor interface." *Science* 267(5196): 383-386.

Clarke, N. D. (1995). "Covariation of residues in the homeodomain sequence family." *Protein Science* 4(11): 2269-2278.

Cole, C. &J. Warwicker (2002). "Side-chain conformational entropy at protein-protein interfaces." *Protein Sci* 11(12): 2860-70.

Darnell, S. J., L. LeGault &J. C. Mitchell (2008). "KFC Server: interactive forecasting of protein interaction hot spots." *Nucleic Acids Res* 36(Web Server issue): W265-9.

Darnell, S. J., D. Page &J. C. Mitchell (2007). "An automated decision-tree approach to predicting protein interaction hot spots." *Proteins* 68(4): 813-23.

de Vries, S. J., A. D. J. van Dijk &A. M. J. J. Bonvin (2006). "WHISCY: what information does surface conservation yield? Application to data-driven docking." *Proteins* 63(3): 479-89.

DeLano, W. L. (2002). "Unraveling hot spots in binding interfaces: progress and challenges." *Curr Opin Struct Biol* 12(1): 14-20.

Diller, D. J., C. Humblet, X. Zhang &L. M. Westerhoff (2010). "Computational alanine scanning with linear scaling semiempirical quantum mechanical methods." *Proteins* 78(10): 2329-37.

Dou, T., C. Ji, S. Gu, J. Xu, K. Ying, Y. Xie &Y. Mao (2006). "Co-evolutionary analysis of insulin/insulin like growth factor 1 signal pathway in vertebrate species." *Frontiers in bioscience : a journal and virtual library* 11: 380-8.

Engelen, S., L. A. Trojan, S. Sacquin-Mora, R. Lavery &A. Carbone (2009). "Joint Evolutionary Trees: A Large-Scale Method To Predict Protein Interfaces Based on Sequence Sampling." *PLoS Comput Biol* 5(1): e1000267.

Enright, A. J., I. Iliopoulos, N. C. Kyrpides &C. A. Ouzounis (1999). "Protein interaction maps for complete genomes based on gene fusion events." *Nature* 402(6757): 86-90.

Evguenieva-Hackenberg, E., P. Walter, E. Hochleitner, F. Lottspeich &G. Klug (2003). "An exosome-like complex in Sulfolobus solfataricus." *EMBO Rep* 4(9): 889-93.

Ewing, R. M., P. Chu, F. Elisma, H. Li, P. Taylor, S. Climie, L. McBroom-Cerajewski, M. D. Robinson, L. O'Connor, M. Li, R. Taylor, M. Dharsee, Y. Ho, A. Heilbut, L. Moore,

S. Zhang, O. Ornatsky, Y. V. Bukhman, M. Ethier, Y. Sheng, J. Vasilescu, M. Abu-Farha, J. P. Lambert, H. S. Duewel, Stewart, II, B. Kuehl, K. Hogue, K. Colwill, K. Gladwish, B. Muskat, R. Kinach, S. L. Adams, M. F. Moran, G. B. Morin, T. Topaloglou &D. Figeys (2007). "Large-scale mapping of human protein-protein interactions by mass spectrometry." *Mol Syst Biol* 3: 89.

Fares, M. A. &D. McNally (2006). "CAPS: Coevolution analysis using protein sequences." *Bioinformatics* 22(22): 2821-2822.

Fariselli, P., F. Pazos, A. Valencia &R. Casadio (2002). "Prediction of protein--protein interaction sites in heterocomplexes with neural networks." *Eur J Biochem* 269(5): 1356-61.

Fernandez-Fuentes, N., B. K. Rai, C. J. Madrid-Aliste, J. E. Fajardo &A. Fiser (2007). "Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments." *Bioinformatics* 23(19): 2558-65.

Finn, R. D., M. Marshall &A. Bateman (2005). "iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions." *Bioinformatics* 21(3): 410-2.

Finn, R. D., J. Mistry, B. Schuster-B√∂ckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. L. Sonnhammer &A. Bateman (2006). "Pfam: clans, web tools and services." *Nucleic Acids Res* 34(Database issue): D247-51.

Fleishman, S. J., S. D. Khare, N. Koga &D. Baker (2011). "Restricted sidechain plasticity in the structures of native proteins and complexes." *Protein Sci* 20(4): 753-7.

Fu, W., B. E. Sanders-Beer, K. S. Katz, D. R. Maglott, K. D. Pruitt &R. G. Ptak (2009). "Human immunodeficiency virus type 1, human protein interaction database at NCBI." *Nucleic Acids Res* 37(Database issue): D417-22.

Fundel, K., R. Kuffner &R. Zimmer (2007). "RelEx--relation extraction using dependency parse trees." *Bioinformatics* 23(3): 365-71.

Gao, H., Y. Dou, J. Yang &J. Wang (2011). "New methods to measure residues coevolution in proteins." *BMC Bioinformatics* 12: 206.

Glaser, F., D. M. Steinberg, I. A. Vakser &N. Ben Tal (2001). "Residue frequencies and pairing preferences at protein-protein interfaces." *Proteins* 43(2): 89.

Gloor, G. B., L. C. Martin, L. M. Wahl &S. D. Dunn (2005). "Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions." *Biochemistry* 44(19): 7156-7165.

Gobel, U., C. Sander, R. Schneider &A. Valencia (1994). "Correlated mutations and residue contacts in proteins." *Proteins: Structure, Function and Genetics* 18(4): 309-317.

Goh, C.-S. &F. E. Cohen (2002). "Co-evolutionary analysis reveals insights into protein-protein interactions." *Journal of Molecular Biology* 324(1): 177-192.

Goll, J., S. V. Rajagopala, S. C. Shiau, H. Wu, B. T. Lamb &P. Uetz (2008). "MPIDB: the microbial protein interaction database." *Bioinformatics* 24(15): 1743-4.

Grishin, N. V. &M. A. Phillips (1994). "The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences." *Protein Sci* 3(12): 2455-8.

Grosdidier, S. &J. Fernandez-Recio (2008). "Identification of hot-spot residues in protein-protein interactions by computational docking." *BMC Bioinformatics* 9: 447.

Guerois, R., J. E. Nielsen &L. Serrano (2002). "Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations." *J Mol Biol* 320(2): 369-87.

Guharoy, M. &P. Chakrabarti (2005). "Conservation and relative importance of residues across protein-protein interfaces." *Proc Natl Acad Sci U S A* 102(43): 15447-52.

Hakes, L., S. C. Lovell, S. G. Oliver &D. L. Robertson (2007). "Specificity in protein interactions and its relationship with sequence diversity and coevolution." *Proceedings of the National Academy of Sciences of the United States of America* 104(19): 7999-8004.

Hoffmann, R. &A. Valencia (2004). "A gene network for navigating the literature." *Nat Genet* 36(7): 664.

Hsu, C. M., C. Y. Chen, B. J. Liu, C. C. Huang, M. H. Laio, C. C. Lin &T. L. Wu (2007). "Identification of hot regions in protein-protein interactions by sequential pattern mining." *BMC Bioinformatics* 8 Suppl 5: S8.

Hu, Z., B. Ma, H. Wolfson &R. Nussinov (2000). "Conservation of polar residues as hot spots at protein interfaces." *Proteins* 39(4): 331-42.

Huang, T. W., C. Y. Lin &C. Y. Kao (2007). "Reconstruction of human protein interolog network using evolutionary conserved network." *BMC Bioinformatics* 8: 152.

Huang, Y. J., D. Hang, L. J. Lu, L. Tong, M. B. Gerstein &G. T. Montelione (2008). "Targeting the human cancer pathway protein interaction network by structural genomics." *Mol Cell Proteomics* 7(10): 2048-60.

Hubner, N. C., A. W. Bird, J. Cox, B. Splettstoesser, P. Bandilla, I. Poser, A. Hyman &M. Mann (2010). "Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions." *J Cell Biol* 189(4): 739-54.

Hue, M., M. Riffle, J. P. Vert &W. S. Noble (2010). "Large-scale prediction of protein-protein interactions from structures." *BMC Bioinformatics* 11: 144.

Itzhaki, Z., E. Akiva, Y. Altuvia &H. Margalit (2006). "Evolutionary conservation of domain-domain interactions." *Genome Biol* 7(12): R125.

Jones, S. &J. M. Thornton (1997). "Prediction of protein-protein interaction sites using patch analysis." *J.Mol.Biol.* 272(1): 133.

Jonsson, P. F. &P. A. Bates (2006). "Global topological features of cancer proteins in the human interactome." *Bioinformatics* 22(18): 2291-7.

Jothi, R., P. F. Cherukuri, A. Tasneem &T. M. Przytycka (2006). "Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions." *J Mol Biol* 362(4): 861-75.

Juan, D., F. Pazos &A. Valencia (2008). "High-confidence prediction of global interactomes based on genome-wide coevolutionary networks." *Proc Natl Acad Sci U S A* 105(3): 934-9.

Kastritis, P. L. &A. M. Bonvin (2010). "Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark." *J Proteome Res* 9(5): 2216-25.

Kawashima, S., P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama &M. Kanehisa (2008). "AAindex: amino acid index database, progress report 2008." *Nucleic Acids Res* 36(Database issue): D202-5.

Kelley, R. &T. Ideker (2005). "Systematic interpretation of genetic interactions using protein networks." *Nat Biotechnol* 23(5): 561-6.

Keskin, O., B. Ma &R. Nussinov (2005). "Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues." *J Mol Biol* 345(5): 1281-94.

Kiemer, L. &G. Cesareni (2007). "Comparative interactomics: comparing apples and pears?" *Trends Biotechnol* 25(10): 448-54.

Kim, S., J. Yoon &J. Yang (2008). "Kernel approaches for genic interaction extraction." *Bioinformatics* 24(1): 118-26.

Kim, W. K., J. Park &J. K. Suh (2002). "Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair." *Genome Inform* 13: 42-50.

Koike, A. &T. Takagi (2004). "Prediction of protein-protein interaction sites using support vector machines." *Protein Engineering Design and Selection* 17(2): 165-173.

Koonin, E. V., Y. I. Wolf &L. Aravind (2001). "Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach." *Genome Res* 11(2): 240-52.

Kortemme, T. &D. Baker (2002). "A simple physical model for binding energy hot spots in protein-protein complexes." *Proc Natl Acad Sci U S A* 99(22): 14116-21.

Kruger, D. M. &H. Gohlke (2010). "DrugScorePPI webserver: fast and accurate in silico alanine scanning for scoring protein-protein interactions." *Nucleic Acids Res* 38(Web Server issue): W480-6.

Kundrotas, P. J., Z. Zhu &I. A. Vakser (2010). "GWIDD: Genome-wide protein docking database." *Nucleic Acids Res* 38(Database issue): D513-7.

Labedan, B., Y. Xu, D. G. Naumoff &N. Glansdorff (2004). "Using quaternary structures to assess the evolutionary history of proteins: the case of the aspartate carbamoyltransferase." *Molecular biology and evolution* 21(2): 364-73.

Lafont, V., M. Schaefer, R. H. Stote, D. Altschuh &A. Dejaegere (2007). "Protein-protein recognition and interaction hot spots in an antigen-antibody complex: free energy decomposition identifies "efficient amino acids"." *Proteins* 67(2): 418-34.

Landgraf, R., I. Xenarios &D. Eisenberg (2001). "Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins." *J.Mol.Biol.* 307(5): 1487.

Landon, M. R., D. R. Lancia, Jr., J. Yu, S. C. Thiel &S. Vajda (2007). "Identification of hot spots within druggable binding regions by computational solvent mapping of proteins." *J Med Chem* 50(6): 1231-40.

Lee, H., M. Deng, F. Sun &T. Chen (2006). "An integrated approach to the prediction of domain-domain interactions." *BMC Bioinformatics* 7: 269.

Lewis, A. C. F., R. Saeed &C. M. Deane (2010). "Predicting protein-protein interactions in the context of protein evolution." *Molecular BioSystems* 6: 55-64.

Li, L., B. Zhao, Z. Cui, J. Gan, M. K. Sakharkar &P. Kangueane (2006). "Identification of hot spot residues at protein-protein interface." *Bioinformation* 1(4): 121-6.

Li, X., O. Keskin, B. Ma, R. Nussinov &J. Liang (2004). "Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking." *J Mol Biol* 344(3): 781-95.

Li, Z., L. Wong &J. Li (2011). "DBAC: a simple prediction method for protein binding hot spots based on burial levels and deeply buried atomic contacts." *BMC Syst Biol* 5 Suppl 1: S5.

Liang, S., C. Zhang, S. Liu &Y. Zhou (2006). "Protein binding site prediction using an empirical scoring function." *Nucleic Acids Res* 34(13): 3698-707.

Lichtarge, O., H. R. Bourne &F. E. Cohen (1996). "An evolutionary trace method defines binding surfaces common to protein families." *J.Mol.Biol.* 257(2): 342.

Lise, S., C. Archambeau, M. Pontil &D. T. Jones (2009). "Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods." *BMC Bioinformatics* 10: 365.

Lo Conte, L., C. Chothia &J. Janin (1999). "The atomic structure of protein-protein recognition sites." *J.Mol.Biol.* 285(5): 2177.

Lovell, S. C. &D. L. Robertson (2010). "An integrated view of molecular coevolution in protein-protein interactions." *Molecular Biology and Evolution* 27(11): 2567-2575.

Lu, L., H. Lu &J. Skolnick (2002). "MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading." *Proteins* 49(3): 350-64.

Luo, Q., P. Pagel, B. Vilne &D. Frishman (2011). "DIMA 3.0: Domain Interaction Map." *Nucleic Acids Res* 39(Database issue): D724-9.

Ma, B. &R. Nussinov (2007). "Trp/Met/Phe hot spots in protein-protein interactions: potential targets in drug design." *Curr Top Med Chem* 7(10): 999-1005.

Madaoui, H. &R. Guerois (2008). "Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking." *Proceedings of the National Academy of Sciences of the United States of America* 105(22): 7708-7713.

Marcotte, E. M., M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates &D. Eisenberg (1999). "Detecting protein function and protein-protein interactions from genome sequences." *Science* 285(5428): 751-3.

Matthews, L. R., P. Vaglio, J. Reboul, H. Ge, B. P. Davis, J. Garrels, S. Vincent &M. Vidal (2001). "Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs"." *Genome Res* 11(12): 2120-6.

McPartland, J. M., R. W. Norris &C. W. Kilpatrick (2007). "Coevolution between cannabinoid receptors and endocannabinoid ligands." *Gene* 397(1-2): 126-35.

Mika, S. &B. Rost (2006). "Protein-protein interactions more conserved within species than across species." *PLoS Comput Biol* 2(7): e79.

Mintseris, J. &Z. Weng (2005). "Structure, function, and evolution of transient and obligate protein-protein interactions." *Proceedings of the National Academy of Sciences of the United States of America* 102(31): 10930-10935.

Miwa, M., R. Saetre, Y. Miyao &J. Tsujii (2009). "Protein-protein interaction extraction by leveraging multiple kernels and parsers." *Int J Med Inform* 78(12): e39-46.

Moreira, I. S., P. A. Fernandes &M. J. Ramos (2007). "Computational alanine scanning mutagenesis--an improved methodological approach." *J Comput Chem* 28(3): 644-54.

Murakami, Y. &S. Jones (2006). "SHARP2: protein-protein interaction predictions using patch analysis." *Bioinformatics* 22(14): 1794-5.

Murakami, Y. &K. Mizuguchi (2010). "Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites." *Bioinformatics* 26(15): 1841-8.

Murali, T., S. Pacifico, J. Yu, S. Guest, G. G. Roberts, 3rd &R. L. Finley, Jr. (2011). "DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for Drosophila." *Nucleic Acids Res* 39(Database issue): D736-43.

Negi, S. S., C. H. Schein, N. Oezguen, T. D. Power &W. Braun (2007). "InterProSurf: a web server for predicting interacting sites on protein surfaces." *Bioinformatics* 23(24): 3397-9.

Neuvirth, H., R. Raz &G. Schreiber (2004). "ProMate: a structure based prediction program to identify the location of protein-protein binding sites." *J Mol Biol* 338(1): 181-99.

Ochoa, D. &F. Pazos (2010). "Studying the co-evolution of protein families with the Mirrortree web server." *Bioinformatics* 26(10): 1370-1.

Ofran, Y. &B. Rost (2003). "Predicted protein-protein interaction sites from local sequence information." *FEBS Lett* 544(1-3): 236-9.

Ofran, Y. &B. Rost (2007). "ISIS: interaction sites identified from sequence." *Bioinformatics* 23(2): e13-6.

Ofran, Y. &B. Rost (2007). "Protein-protein interaction hotspots carved into sequences." *PLoS Comput Biol* 3(7): e119.

Pagel, P., P. Wong &D. Frishman (2004). "A domain interaction map based on phylogenetic profiling." *J Mol Biol* 344(5): 1331-46.

Panni, S., L. Dente &G. Cesareni (2002). "In vitro evolution of recognition specificity mediated by SH3 domains reveals target recognition rules." *J Biol Chem* 277(24): 21666-74.

Park, D., R. Singh, M. Baym, C. S. Liao &B. Berger (2011). "IsoBase: a database of functionally related proteins across PPI networks." *Nucleic Acids Res* 39(Database issue): D295-300.

Patil, A., K. Nakai &H. Nakamura (2011). "HitPredict: a database of quality assessed protein-protein interactions in nine species." *Nucleic Acids Res* 39(Database issue): D744-9.

Pazos, F. &A. Valencia (2008). "Protein co-evolution, co-adaptation and interactions." *Embo J* 27(20): 2648-55.

Pei, J. &N. V. Grishin (2001). "AL2CO: calculation of positional conservation in a protein sequence alignment." *Bioinformatics* 17(8): 700-12.

Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg &T. O. Yeates (1999). "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." *Proc.Natl.Acad.Sci.U.S.A* 96(8): 4285.

Pesquita, C., D. Faria, A. O. Falcao, P. Lord &F. M. Couto (2009). "Semantic similarity in biomedical ontologies." *PLoS Comput Biol* 5(7): e1000443.

Porollo, A. &J. Ç. Meller (2007). "Prediction-based fingerprints of protein-protein interactions." *Proteins* 66(3): 630-45.

Prasad, T. S., K. Kandasamy &A. Pandey (2009). "Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology." *Methods Mol Biol* 577: 67-79.

Presser, A., M. B. Elowitz, M. Kellis &R. Kishony (2008). "The evolutionary dynamics of the Saccharomyces cerevisiae protein interaction network after duplication." *Proceedings of the National Academy of Sciences of the United States of America* 105(3): 950-954.

Qin, S. &H. X. Zhou (2007). "meta-PPISP: a meta web server for protein-protein interaction site prediction." *Bioinformatics* 23(24): 3386-7.

Res, I., I. Mihalek &O. Lichtarge (2005). "An evolution based classifier for prediction of protein interfaces without using protein structures." *Bioinformatics* 21(10): 2496-2501.

Riley, R., C. Lee, C. Sabatti &D. Eisenberg (2005). "Inferring protein domain interactions from databases of interacting proteins." *Genome Biol* 6(10): R89.

Rodionov, A., A. Bezginov, J. Rose &E. R. Tillier (2011). "A new, fast algorithm for detecting protein coevolution using maximum compatible cliques." *Algorithms for molecular biology : AMB* 6: 17.

Rual, J. F., K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth &M. Vidal (2005). "Towards a proteome-scale map of the human protein-protein interaction network." *Nature* 437(7062): 1173-8.

Salwinski, L., C. Miller, A. Smith, F. Pettit, J. Bowie &D. Eisenberg (2004). "The Database of Interacting Proteins: 2004 update." *Nucleic Acids Res* 32: D449 - D451.

Segura, J. &N. Fernandez-Fuentes (2011). "PCRPi-DB: a database of computationally annotated hot spots in protein interfaces." *Nucleic Acids Res* 39(Database issue): D755-60.

Segura, J., P. F. Jones &N. Fernandez-Fuentes (2011). "Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams." *BMC Bioinformatics* 12: 352.

Shoemaker, B. &A. Panchenko (2007). "Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners." *PLoS Comput Biol* 3(4): 595 - 601.

Shoemaker, B. A. &A. R. Panchenko (2007). "Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners." *PLoS Comput Biol* 3(4): e43.

Shoemaker, B. A., D. Zhang, R. R. Thangudu, M. Tyagi, J. H. Fong, A. Marchler-Bauer, S. H. Bryant, T. Madej &A. R. Panchenko (2010). "Inferred Biomolecular Interaction Server--a web server to analyze and predict protein interacting partners and binding sites." *Nucleic Acids Res* 38(Database issue): D518-24.

Sikic, M., S. Tomic &K. Vlahovicek (2009). "Prediction of protein-protein interaction sites in sequences and 3D structures by random forests." *PLoS Comput Biol* 5(1): e1000278.

Stark, C., B. J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski &M. Tyers (2011). "The BioGRID Interaction Database: 2011 update." *Nucleic Acids Res* 39(Database issue): D698-704.

Stein, A., A. Ceol &P. Aloy (2011). "3did: identification and classification of domain-based interactions of known three-dimensional structure." *Nucleic Acids Res* 39(Database issue): D718-23.

Stein, A., R. Mosca &P. Aloy (2011). "Three-dimensional modeling of protein interactions and complexes is going -omics." *Current Opinion in Structural Biology* 21(2): 200-208.

Stelzl, U. &E. E. Wanker (2006). "The value of high quality protein-protein interaction networks for systems biology." *Curr Opin Chem Biol* 10(6): 551-8.

Stelzl, U., U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksoz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach &E. E. Wanker (2005). "A human protein-protein interaction network: a resource for annotating the proteome." *Cell* 122(6): 957-68.

Stumpf, M., T. Thorne, E. de Silva, R. Stewart, H. An, M. Lappe &C. Wiuf (2008). "Estimating the size of the human interactome." *Proceedings of the National Academy of Sciences of the United States of America* 105(19): 6959 - 6964.

Swarbreck, D., C. Wilks, P. Lamesch, T. Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz, A. Radenbaugh, S. Singh, V. Swing, C. Tissier, P. Zhang &E. Huala (2008). "The Arabidopsis Information Resource (TAIR): gene structure and function annotation." *Nucleic Acids Res* 36(Database issue): D1009-14.

Szklarczyk, D., A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen &C. von Mering (2011). "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored." *Nucleic Acids Res* 39(Database issue): D561-8.

Ta, H. X., P. Koskinen &L. Holm (2011). "A novel method for assigning functional linkages to proteins using enhanced phylogenetic trees." *Bioinformatics* 27(5): 700-6.

Tanaka, T., R. L. Williams &T. H. Rabbitts (2007). "Tumour prevention by a single antibody domain targeting the interaction of signal transduction proteins with RAS." *EMBO J* 26(13): 3250-9.

Tikk, D., P. Thomas, P. Palaga, J. Hakenberg &U. Leser (2010). "A comprehensive benchmark of kernel methods to extract protein-protein interactions from literature." *PLoS Comput Biol* 6: e1000837.

Tillier, E. R. M. &R. L. Charlebois (2009). "The human protein coevolution network." *Genome Research* 19(10): 1861 -1871.

Travers, S. A. A. &M. A. Fares (2007). "Functional coevolutionary networks of the Hsp70-Hop-Hsp90 system revealed through computational analyses." *Molecular Biology and Evolution* 24(4): 1032-1044.

Tuncbag, N., A. Gursoy &O. Keskin (2009). "Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy." *Bioinformatics* 25(12): 1513-20.

Tuncbag, N., A. Gursoy, R. Nussinov &O. Keskin (2011). "Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM." *Nat. Protocols* 6(9): 1341-1354.

Venkatesan, K., J. F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K. I. Goh, M. A. Yildirim, N. Simonis, K. Heinzmann, F. Gebreab, J. M. Sahalie, S. Cevik, C. Simon, A. S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R. R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M. E. Cusick, F. P. Roth, D. E. Hill, J. Tavernier, E. E. Wanker, A. L. Barabasi &M. Vidal (2009). "An empirical framework for binary interactome mapping." *Nat Methods* 6(1): 83-90.

Vlahovicek, K., A. Pintar, L. Parthasarathi, O. Carugo &S. Pongor (2005). "CX, DPX and PRIDE: WWW servers for the analysis and comparison of protein 3D structures." *Nucleic Acids Res* 33(Web Server issue): W252-4.

Walhout, A. J., R. Sordella, X. Lu, J. L. Hartley, G. F. Temple, M. A. Brasch, N. Thierry-Mieg &M. Vidal (2000). "Protein interaction mapping in C. elegans using proteins involved in vulval development." *Science* 287(5450): 116-22.

Wang, B., P. Chen, D.-S. Huang, J.-j. Li, T.-M. Lok &M. R. Lyu (2006). "Predicting protein interaction sites from residue spatial sequence profile and evolution rate." *FEBS Lett* 580(2): 380-4.

Wass, M. N., G. Fuentes, C. Pons, F. Pazos &A. Valencia (2011). "Towards the prediction of protein interaction partners using physical docking." *Mol Syst Biol* 7: 469.

Williams, S. G. &S. C. Lovell (2009). "The effect of sequence evolution on protein structural divergence." *Molecular Biology and Evolution* 26(5): 1055-1065.

Wu, J., S. Kasif &C. DeLisi (2003). "Identification of functional links between genes using phylogenetic profiles." *Bioinformatics* 19(12): 1524-30.

Xia, J. F., X. M. Zhao, J. Song &D. S. Huang (2010). "APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility." *BMC Bioinformatics* 11: 174.

Yanai, I., A. Derti &C. DeLisi (2001). "Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes." *Proc Natl Acad Sci U S A* 98(14): 7940-5.

Yang, Z., Y. Lin, J. Wu, N. Tang, H. Lin &Y. Li (2011). "Ranking support vector machine for multiple kernels output combination in protein-protein interaction extraction from biomedical literature." *Proteomics* 11(19): 3811-7.

Yellaboina, S., A. Tasneem, D. V. Zaykin, B. Raghavachari &R. Jothi (2011). "DOMINE: a comprehensive collection of known and predicted domain-domain interactions." *Nucleic Acids Res* 39(Database issue): D730-5.

Yogurtcu, O. N., S. B. Erdemli, R. Nussinov, M. Turkay &O. Keskin (2008). "Restricted mobility of conserved residues in protein-protein interfaces in molecular simulations." *Biophys J* 94(9): 3475-85.

Yu, C.-Y., L.-C. Chou &D. Chang (2010). "Predicting protein-protein interactions in unbalanced data using the primary structure of proteins." *BMC Bioinformatics* 11(1): 167.

Yuan, Z., J. Zhao &Z.-X. Wang (2003). "Flexibility analysis of enzyme active sites by crystallographic temperature factors." *Protein Eng* 16(2): 109-14.

Zhang, Y., H. Lin, Z. Yang &Y. Li (2011). "Neighborhood hash graph kernel for protein-protein interaction extraction." *J Biomed Inform*.

Zhou, H. X. &Y. Shan (2001). "Prediction of protein interaction sites from sequence profile and residue neighbor list." *Proteins* 44(3): 336-43.