BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 22 2009, pages 3020–3025
doi:10.1093/bioinformatics/btp522

Databases and ontologies

Comparative analysis and unification of domain–domain interaction networks

Patrik Björkholm* and Erik L. L. Sonnhammer

Stockholm Bioinformatics Center, Albanova, Stockholm University, 10691 Stockholm, Sweden

Received on May 4, 2009; revised on August 12, 2009; accepted on August 26, 2009

Advance Access publication August 31, 2009

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Certain protein domains are known to preferentially interact with other domains. Several approaches have been proposed to predict domain–domain interactions, and over nine datasets are available. Our aim is to analyse the coverage and quality of the existing resources, as well as the extent of their overlap. With this knowledge, we have the opportunity to merge individual domain interaction networks to construct a comprehensive and reliable database.

Results: In this article we introduce a new approach towards comparing domain–domain interaction networks. This approach is used to compare nine predicted domain and protein interaction networks. The networks were used to generate a database of unified domain interactions, UniDomInt. Each interaction in the dataset is scored according to the benchmarked reliability of the sources. The performance of UniDomInt is an improvement compared to the underlying source networks and to another composite resource, *Domine*.

Availability: http://sonnhammer.sbc.su.se/download/UniDomInt/

Contact: Erik.Sonnhammer@sbc.su.se

1 INTRODUCTION

Proteins are social molecules that network with other proteins. If we are ever to properly understand the processes that make up life, it is essential that these interactions and networks can be reliably predicted and understood. High-throughput techniques exist for experimentally identifying protein–protein interactions (PPI), but these methods have a tendency to produce a high rate of false positives and false negatives (Mrowka *et al*., 2001). This creates a need to be able to accurately predict and assess PPI in a reliable manner.

A good beginning for prediction of PPI is to accurately predict domain–domain interactions. Domain interactions are often conserved across species (Itzhaki *et al*., 2006). Several approaches for predicting domain–domain interactions have been developed. The first approach was the association method, where domains were mapped to interacting proteins. In this method, the domain interactions selected are those were the frequency of the domain interaction exceeds the number of expected interactions given the domains' abundance in the proteome (Kim *et al*., 2002). Extensions to this approach include 'domain pair exclusion analysis' (Riley *et al*., 2005) and random forest optimization (Chen and Liu, 2005). In 'domain pair exclusion analysis' a new score was introduced, based on the log ratio of two domains interacting over not interacting. The random forest optimization method explores all possible domain interactions over all available domains using predicted PPI. An advantage of using random forest optimization is that it considers domains as a protein feature and can therefore estimate the effect of multi-domain combinations on interactions. Phylogenetic profiling has also been used for predicting domain–domain interactions, i.e. by inference from co-occurrence of domains across various species (Pagel *et al*., 2004).

It is also possible to integrate different types of data for predicting domain interactions, for instance by combining gene ontology functional annotation with protein interaction data as done in ME using a Bayesian approach (Lee *et al*., 2006). Also co-evolutionary analysis is used for generating domain networks. In this approach the co-evolution between domains is estimated by analysing structure and sequence (Jothi *et al*., 2006). Thus, a wide variety of ideas have been developed in order to predict domain-domain interactions, with different degrees of reliability.

Domine is a composite protein domain interaction resource, generated by combining eight predicted networks (Raghavachari *et al*., 2008). Each predicted interaction is assigned to one of three confidence levels (high, medium and low) using a system based on three rules. A prediction by *ME* or by two methods gets high confidence. If not, it is assigned low confidence, or medium if both domains have the same GO terms.

We here introduce a benchmarking strategy for obtaining a continuous confidence score. A standard approach to evaluate domain–domain interactions is to calculate the overlap of interactions in a reference database. The standard reference databases are *iPfam* and *3DID* (Finn *et al*., 2005, Stein *et al*., 2005). The sources of these interactions are experimentally derived three-dimensional structures in the Protein Data Bank (Westbrook *et al*., 2002). To use these resources as references is a sensible choice as the structure-derived interactions are the closest thing to true domain–domain interactions. There is however a potential danger that this approach is biased due to the limitations of reaching certain structure spaces of the proteome with NMR or X-ray crystallography (Mrowka *et al*., 2001).

The standard overlap approach is here further developed for domain interaction network comparisons into a measure called *weighted overlap*. This expanded approach compares overlapping interactions with possible overlapping interactions. The advantage

*To whom correspondence should be addressed.

of this approach is that the result is no longer as affected by the difference in size between networks and their domain composition. The developed measure was used to compare nine predicted networks. The networks were then scored and combined in a weighted way into the Unified Domain Interaction database (UniDomInt).

2 MATERIALS AND METHODS

2.1 Networks and domain space

We define I as the set of protein–domain interactions in a domain network. The domain space D is defined as the set of domains in I, meaning that all domains in D must be present in at least one interaction in I. From a network point of view the domains can be seen as nodes and the interactions as edges between them.

2.2 Shared domain space and potential number of shared interactions

For two networks I_a and I_b and their domain spaces D_a and D_b, the common domain space D_{ab} is $D_a \cap D_b$. For I_a, the number of potentially shared interactions, $I_{a \to b}$, is the number of interactions where both interacting domains belong to D_{ab}, and likewise $I_{b \to a}$ is the number of potentially shared interactions for I_b.

2.3 Measures for comparing networks

To assess the level of similarity between two networks, it is necessary to know the potential number of shared interactions given the shared domain space (see Fig. 1). The *weighted overlap, W*o, is the number of common interactions in the two networks, divided by the number of potential interactions given both networks:

$$W_o = \left(\frac{2(I_a \cap I_b)}{I_{a \to b} + I_{b \to a}} \right)$$

2.4 Reference set

A reference set has been generated to assess the predicted interaction networks. The reference data-set was generated by merging together the structure-based domain interaction resources *3DID* and *iPfam* (*iPfam* version 21.0 and *3DID* August 2005 version) (Finn *et al*., 2005; Stein *et al*., 2005). This combined network contains 4349 PDB-derived domain interactions, of which 2064 are homo-domain interactions (self-interacting domains).

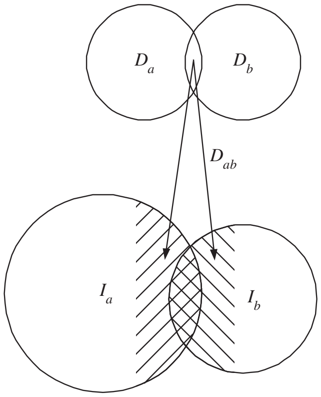

Fig. 1. Venn diagram explaining the weighted overlap score W_o. Given two domain interaction networks I_a and I_b, their shared domain space D_{ab} defines the number of potentially shared interactions, $I_{a \to b}$ and $I_{b \to a}$, shown as cross-hatched areas of the interaction networks. By calculating Wo as the networks' intersection relative to $I_{a \to b}$ and $I_{b \to a}$, the overlap becomes independent of the network sizes.

2.5 Source network comparisons

An all versus all comparison was done for all predicted networks using the *weighted overlap* score and domain-space similarity. The pairwise Wo scores and shared domain-space fractions between all networks were arranged into distance matrices that were used to generate UPGMA trees for visualizing their relationships. The application Belvu was used for this purpose (Sonnhammer and Hollich, 2005).

2.6 Accuracy score

An accuracy score was calculated for each source network, as the *weighted overlap* score against the reference set made up of the combined *3DID* and *iPfam* networks.

2.7 Domain mapping

The networks chosen as sources to generate UniDomInt, were first converted into Pfam-A if necessary. Two of the networks (*P-value* and *HIMAP*) use different domain identifiers, SCOP and Interpro. The SCOP domains were converted to Pfam using SGD (http://www.yeastgenome.org/) and the Interpro domains were converted to Pfam using the match table from the Interpro website (http://www.ebi.ac.uk/interpro/ISearch?mode=db&query=H). Domains not available in Pfam release 23.0 (Finn *et al*., 2008) were removed from the source networks. The number of interactions and domains remaining in each network can be seen in Table 1.

2.8 Reliability score

Each interaction in UniDomInt receives a reliability score between 0 and 1. This is calculated as the sum of the accuracy scores of the networks containing the interaction, divided by the sum of the accuracy scores for all source networks.

2.9 Generating UniDomInt

In this article, nine predicted networks were evaluated, and if a network was divergent enough it was merged into the database UniDomInt. The networks were merged one at a time, recording the sources and reliability score for each interaction

Table 1. A summary of the networks used to create the UniDomInt database after converting them to Pfam-A (release 23.0) and removing redundant interactions. The table also contains the number of homo-domain (self) interactions in each network.

	No. of interactions	No. of homo-domain interactions	No. of domains
DIMA	3783	259	2007
Interdom	2768	0	1405
RDFF	2475	90	630
ME	2391	758	1235
DPEA	1811	215	1025
LP	1213	191	728
RCDP	994	122	484
P-value	469	21	339
HIMAP	270	32	165
UniDomInt	13166	1158	3562
Reference set	4349	2064	2948

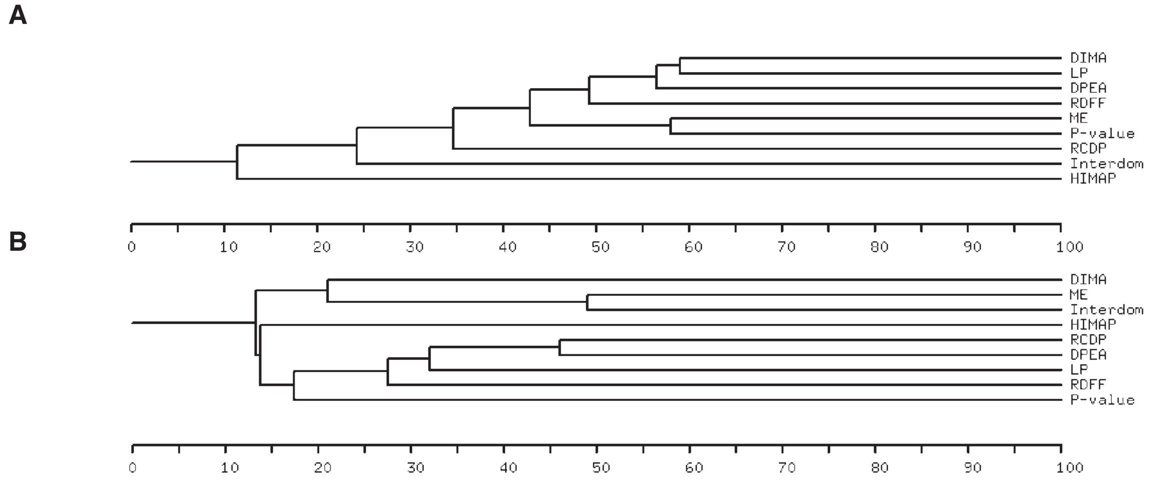

Fig. 2. (**A**) Relationships between nine different domain-domain interaction networks, in terms of *Weighted Overlap* score Wo. (**B**) Network relatedness based on shared domain space, where percentages between two networks were calculated as the number of overlapping domains divided by the size of the smaller set. The trees were built using UPGMA in Belvu (Sonnhammer and Hollich, 2005).

2.10 Source networks

The source networks chosen for the UniDomInt database were obtained from published articles. They are: *Interdom, DPEA, P-value, RDFF, RCDP, ME, DIMA, LP* and *HIMAP.* All networks use Pfam domain identifiers except *P-value* and *HIMAP*.

Interdom is a database of compiled domain–domain interactions that uses an integrated approach to predict potential interactions. *Interdom* uses four different sources, domain fusions, PPI, protein complexes and scientific literature (Bader *et al*., 2003, Mack and Hehenberger, 2002; Marcotte *et al*., 1999; Ng and Wong, 1999; Westbrook *et al*., 2002; Xenarios *et al*., 2000). The genome fusion dataset was used from this network.

Domain pair exclusion analysis (*DPEA*) is a statistical approach that predicts domain–domain interactions from incomplete interactomes from different species. This is done by creating likelihood ratios using an expectation maximization algorithm (Dempster *et al*., 1977; Riley *et al*., 2005).

P-value is a statistical method for predicting interactions between pairs of SCOP super families. The *P-value* is a measure for the strength of the evidence for the interaction (Nye *et al*., 2005). In *P-value,* interactions were predicted between SCOP domains for yeast proteins.

RDFF uses a domain-based Random Decision Forest Framework to predict domain interactions. This is done by using Pfam domains in the proteins as features of proteins (Chen and Liu, 2005). Predictions were made from *Saccharomyces cerevisiae* PPI data (Deng *et al*., 2002; Xenarios *et al*., 2000).

Relative Co-evolution of Domain Pairs (*RCDP*) is a method that studies PPI in F1-ATPase, Sec23p/Sec24p, DNA directed RNA polymerase and nuclear pore complexes (Jothi *et al*., 2006). The domains were analysed from co-evolutionary perspective using both structures and sequences. The results were used to predict domain–domain interactions from the yeast interactome.

ME uses a Bayesian approach to predict domain–domain interactions by integrating gene ontology and protein interaction data from yeast, worm, fruitfly and humans. This method was used to predict domain–domain interactions in *Helicobacter pylori* (Lee *et al*., 2006).

Domain interaction map (DIMA) contains domain interactions predicted by phylogenetic profiling or by the DPEA algorithm. All interactions available from DIMA were used (Pagel *et al*., 2004, 2007).

Linear Programming (LP) method uses a generalized parsimonious explanation (GPE) method for predicting domain interactions. LP looks for the smallest set of domain interactions to explain all protein interactions in a network (Guimarães and Przytycka, 2008; Guimarães *et al*., 2006).

HIMAP uses a semi-naïve Bayesian model to predict PPI by integrating heterogeneous evidences. One of these is domain interactions, that were derived by statistical analysis of domain pairs overrepresented in protein interactions from the HPRD (http://www.hprd.org) database. Interpro was used for domain mapping (Rhodes *et al*., 2005).

2.11 Cut-off values used for the networks

Due to the very different methods and data used for predicting domain interactions in the different networks, it is difficult to compare stringency across networks. Cut-off values used for the different source networks were chosen by the publishers of the original data.

3 RESULTS AND DISCUSSION

The aim of this article is to create a unified and reliable resource of predicted domain–domain interactions. For this purpose nine domain interaction networks were evaluated in terms of overlap and accuracy. Based on this analysis we defined a reliability score for each interaction in the merged dataset.

3.1 Network similarities

To investigate the similarity between the nine networks, we used a measure of overlap that compensates for the different domain content in each network. This measure only considers interactions between domains occurring in both networks being compared. The results were used to build a tree of network relatedness (Fig. 2A). The network that made the most divergent predictions was *HIMAP*, followed by *Interdom* and *RCDP*. A similar tree was built based on the level of shared domain space between the networks (Fig. 2B). As in the interaction tree, *HIMAP* is an outlier. For the rest, the trees are very different, i.e. there is generally no connection between shared domain space and shared interactions. The moderate overlap between the networks shows that each method is relatively unique.

Table 2. Quality benchmark of domain–domain interaction networks

	Accuracy (%)	Precision (%)
ME	57.88	52.91
LP	31.84	19.87
DIMA	31.61	13.90
DPEA	29.57	12.09
RCDP	20.67	12.98
Interdom	20.10	11.85
HIMAP	16.67	11.11
P-value	12.61	9.38
RDFF	9.35	4.85

All databases that were merged into UniDomInt were compared pairwise against a combined set of reference interactions from *iPfam* and *3DID*. Accuracy is measured as the *weighted overlap W*o with this dataset. Precision is the fraction of *iPfam* or *3DID* interactions present in the network.

3.2 Network reliability

The quality of each network was assessed by comparison to the structure-derived databases *iPfam* and *3DID*, taken together. These are commonly used as gold standards. The results are shown in Table 2. The predictions made by *ME* were closest to the reference datasets, with a precision of 52.91% and an accuracy of 57.88% [precision = True Positives/(True Positives + False Positives)]. Each network's accuracy can be used as a token for its interactions' reliability. The performance of the networks is dependent on the data used in the predictions, especially so when comparing networks predicted by similar approaches like DPEA and RDFF.

3.3 Network coordination and merging

To compare the different networks a common coordinate system was necessary, and therefore all networks were converted to use Pfam identifiers. This also made a merger between the networks possible. Combining all accepted networks resulted in a merged network of 13 166 interactions and 3562 unique domains, called UniDomInt. This covers almost 35% of the available domain space in Pfam-A release 23. The combined dataset of interactions was visualized using Cytoscape (Shannon *et al.*, 2003), see Figure 3. Most interactions (∼89%) are part of a giant component that lacks prominent sub-clusters.

3.4 Interacting domains

From the merged database the most interacting domains were collected. Sixteen domains were connected with more than or equal to one hundred domains (Table 3). The most connected domain was PF00069, the protein kinase domain, that interacted with 480 other domains. Looking through the literature we found that the human 'kinome', i.e. all kinases, comprises about 518 proteins (Manning *et al.*, 2002). This shows that the scale of the predicted kinase domain interactions is plausible, especially considering that they are widely used in modular signalling systems for regulating a wide range of cellular processes.

3.5 Network accuracy

Each network's accuracy in the previous benchmark was used as a weighting factor when calculating the interaction reliability score in UniDomInt. This reliability makes it possible to rank the interactions by confidence. To assess the quality of UniDomInt with this scoring system, we evaluated it using the same *iPfam*/*3DID* benchmark as above. The difference is that we can now rank order the interactions by reliability score and display the results as a curve in a true/false positive plot (Fig. 4). In order to give some perspective on the accuracy of the predicted networks a random network of domain interactions was generated using the domain space of UniDomInt. A true/false positive curve was generated and can be seen at the bottom of Figure 4. The random network performed very poorly compared against all predicted networks. UniDomInt generally showed improved precision compared to the source networks. The UniDomInt curve grazes the true positive/false positive point of the best source network, *ME*. Yet, thanks to its scoring system, UniDomInt can be considerably more sensitive or specific than *ME*.

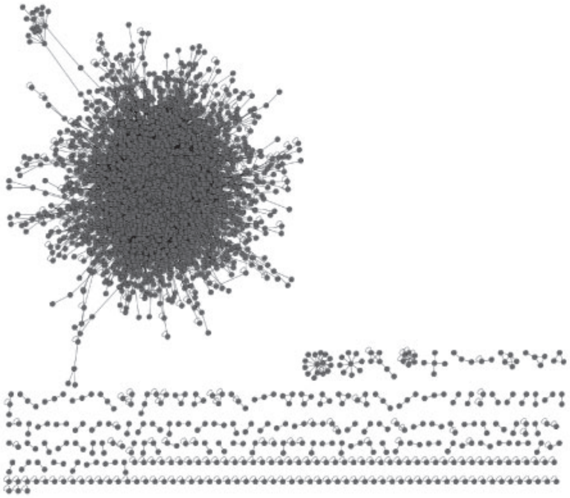

Fig. 3. A visualization of the domain interaction network UniDomInt presented in this article. The circles are domains and the lines are interactions between them.

Table 3. All domains in the UniDomInt network that interact with one hundred or more domains

Accession	Connectivity	Domain name
PF00069	480	Protein kinase domain
PF00076	289	RNA recognition motif
PF00400	221	WD domain, G-beta repeat
PF01423	195	LSM domain
PF00071	194	Ras family
PF00085	169	Thioredoxin
PF00271	158	Helicase conserved C-terminal domain
PF00096	148	Zinc finger, C2H2 type
PF00097	135	Zinc finger, C3HC4 type (RING finger)
PF00018	131	SH3 domain
PF00505	106	HMG (high mobility group) box
PF00515	106	Tetratricopeptide repeat
PF00013	105	KH domain
PF00382	102	Transcription factor TFIIB repeat
PF00270	100	DEAD/DEAH box helicase
PF00004	100	AAA domain

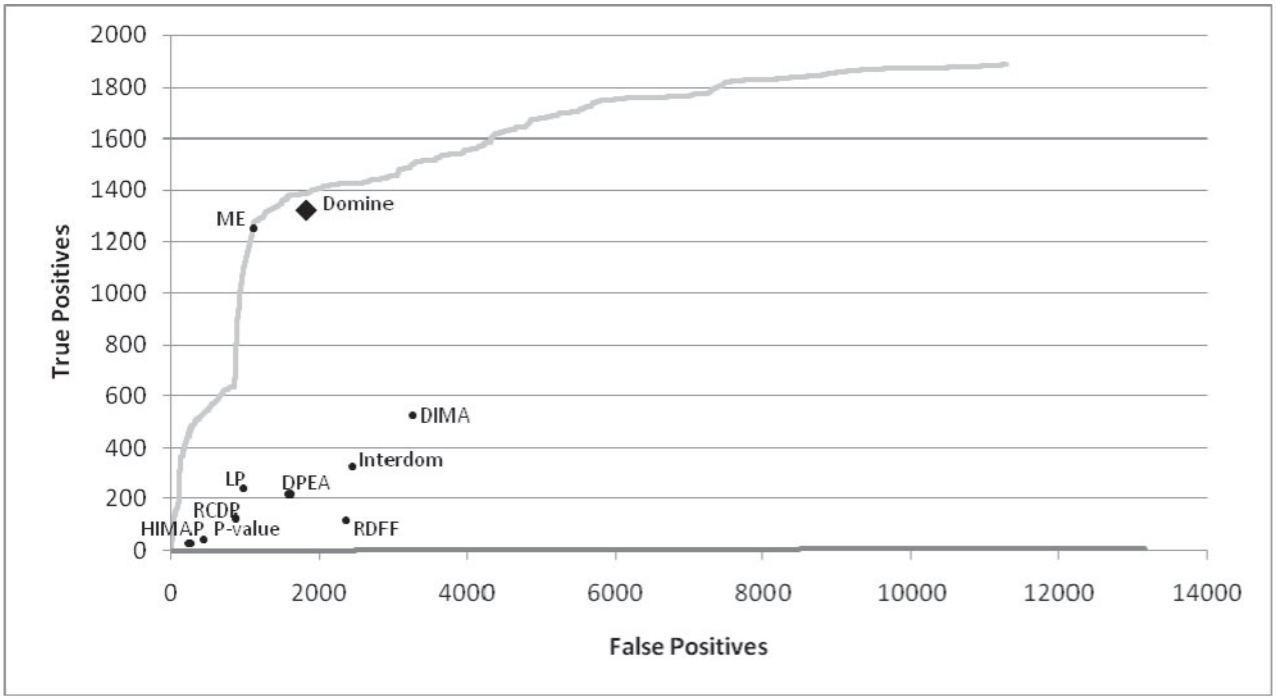

Fig. 4. Quality assessment of domain interaction networks. As UniDomInt provides a score for each domain–domain interaction, a curve can be drawn for it (light grey) based on hits in the reference dataset. The dots represent the performance of the networks used in generating UniDomInt. For comparison, the performance of the high-confidence part of *Domine* is shown as a diamond. UniDomInt generally achieves higher coverage of true positives (precision) at lower false positive rates than the underlying networks; the curve passes through the ME point. A random network of domain-interactions was generated using UniDomInts domain space. This is shown as a dark grey curve is (almost at the *x*-axis). All methods predict domain interactions significantly better than the random network.

3.6 Benchmark

For comparison a benchmark was done against another similar published resource, *Domine*. Although we used more source networks than *Domine*, we obtained in total 4615 fewer interactions. The reason for this discrepancy is that we used the latest versions, which in many cases have become smaller for example *DIMA* and *LP*. This reduction appears to be due to a decrease in the false positive rate. We benchmarked the true and false positive rate for *Domine*'s high-confidence network (22.6% of the entire dataset), using the same reference dataset as before. As can be seen in Figure 4, our network outperforms *Domine's* high-confidence network. The reason for UniDomInts improved performance is likely the use of a more elaborate scoring system, combined with use of more and updated networks. Another difference is that UniDomInt scores each interaction consistently, while *Domine* uses a rule-based score with only three levels.

4 CONCLUSION

We present a new integrated domain–domain interaction network called the Unified Domain Interaction (UniDomInt) dataset. The advantage of UniDomInt is not only its high coverage, but also that each interaction is assigned a reliability score. This makes it possible for the user to choose the desired level of stringency. The UniDomInt network showed an increased precision when using our scoring system then the individual networks used to generate the database. It also performed better when compared against another similar resource *Domine*.

Using the approach that only shared domain space is useful in comparison of networks, we developed the *weighted overlap* score. This score was used to analyse the relationships between nine predicted domain-domain interaction networks. The network making the most unique predictions is *HIMAP*. We observed moderate levels of redundancy between the networks' interactions, and even less between their domain spaces. This means that the total domain space still increases in size for each new prediction method being released. However, a core domain space is shared so that the networks have a basis for comparison.

Funding: Pharmacia and the Swedish Research Council.

Conflict of Interest: none declared.

REFERENCES

Bader,G.D. *et al*. (2003) BIND: the biomolecular interaction network database. *Nucleic Acids Res*., **31**, 248–250.

Chen,X.W. and Liu,M. (2005) Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics*, **21**, 4394–4400.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc*., **39**, 1–38.

Deng,M. *et al*. (2002) Inferring domain-domain interactions from protein-protein interactions. *Genome Res*., **12**, 1540–1548.

Finn,R.D. *et al*. (2005) iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.

Finn,R.D. *et al*. (2008) The Pfam protein families database. *Nucleic Acids Res.*, Issue 36, D281–D288.

Guimarães,K.S. and Przytycka,T.M. (2008) Interrogating domain-domain interactions with parsimony based approaches. *BMC Bioinformatics*, **9**, 171.

Guimarães,K.S. *et al*. (2006) Predicting domain-domain interactions using a parsimony approach. *Genome Biol.*, **7**, R104.

Itzhaki,Z. *et al*. (2006) Evolutionary conservation of domain-domain interactions. *Genome Biol.*, **7**, R125.

Jothi,R. *et al*. (2006) Co-evolutionary analysis of domains in interacting proteins reveals insights into domain–domain interactions mediating protein–protein interactions. *J. Mol. Biol.*, **362**, 861–875.

Kim,W.K. *et al*. (2002) Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair. *Genome Inform.*, **13**, 42–50.

Lee,H. *et al*. (2006) An integrated approach to the prediction of domain-domain interactions. *BMC Bioinformatics*, **7**, 269.

Mack,R. and Hehenberger,M. (2002) Text-based knowledge discovery: search and mining of life-sciences documents. *Drug Discov Today*, **7**(Suppl. 11), S89–S98.

Manning,G. *et al*. (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.

Marcotte,E.M. *et al*. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.

Mrowka,R. *et al*. (2001) Is there a bias in proteome research? *Genome Res.*, **11**, 1971–1973.

Ng,S.K. and Wong,M. (1999) Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Inform. Ser. Workshop Genome Inform.*, **10**, 104–112.

Nye,T.M.W. *et al*. (2005) Statistical analysis of domains in interacting protein pairs. *Bioinformatics*, **21**, 993–1001.

Pagel,P. *et al*. (2004) A domain interaction map based on phylogenetic profiling. *J. Mol. Biol.*, **344**, 1331–1346.

Pagel,P. *et al*. (2008) DIMA 2.0—predicted and known domain interactions. *Nucleic Acids Res.*, **36**, D651–D655.

Raghavachari,B. *et al*. (2008) DOMINE: a database of protein domain interactions. *Nucleic Acids Res.*, **36**(Database Issue), D656–D661.

Riley,R. *et al*. (2005) Inferring protein domain interactions from databases of interacting proteins. *Genome Biol.*, **6**, R89.

Rhodes,D.R. *et al*. (2005) Probabilistic model of the human protein-protein interaction network. *Nat Biotech*., **23**, 951–959.

Shannon,P. *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*., **13**, 2498–2504.

Sonnhammer,E.L. and Hollich,V. (2005) Scoredist: a simple and robust protein sequence distance estimator. *BMC Bioinformatics,* **6**, 108.

Stein,A. *et al*. (2005) 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res*., **33**(Database issue), D413–D417.

Westbrook,J. *et al*. (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.*, **30**, 245–258.

Xenarios,I. *et al*. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res*., **28**, 289–291.