Integrative Approach for Computationally Inferring Protein Domain Interactions

See-Kiong Ng Zhuo Zhang Soon-Heng Tan
Laboratories for Information Technology
21 Heng Mui Keng Terrace
Singapore 119613

ABSTRACT

Motivation: The current need for high-throughput protein interaction detection has resulted in interaction data being generated *en masse*, using experimental methods such as yeast-two-hybrids and protein chips. Such data can be errorful and they often do not provide adequate functional information for the detected interactions; it is therefore useful to develop an *in silico* approach to further validate and annotate the detected protein interactions.

Results: Given that protein-protein interactions involve physical interactions between protein domains, domain-domain interaction information can be useful for validating, annotating, and even predicting protein interactions. However, large-scale experimentally determined domain-domain interaction data do not exist; as such, we describe an integrative approach to computationally derive putative domain interactions from multiple data sources, including rosetta stone sequences, protein interactions, and protein complexes. We show the usefulness of such an integrative approach by applying the derived domain interactions to predict and validate protein-protein interactions.

Contact: skng@lit.a-star.edu.sg

1. INTRODUCTION

The genome era has produced extensive lists of genes and their encoded proteins for many living organisms (Benson *et al.*, 2002; Hubbard *et al.*, 2002). However, simply knowing the parts list of genes and proteins does not tell us much about how life's many biological processes work. The cellular machinery is a complex dynamic system with hundreds of thousands of bio-molecules interacting with one another to execute life's many functions. To fully understand the genetic program of life, a comprehensive description of protein-protein interactions is required.

Historically, scientists have been studying individual protein interactions with *top-down, hypothesis-driven* approaches, designing focused experiments to derive detailed information for testing hypotheses about each interaction studied. Today, the need for high-throughput interaction detection has resulted in large quantities of protein interaction data being generated at an unprecedented rate, using methods such as two-hybrid systems (Ito *et al.*, 2001; Ito *et al.*, 2000; Uetz *et al.*, 2000) and protein chips (Zhu *et al.*, 2001). However, these detection systems often provide only mere detection of the physical molecular interactions. Such a paradigm shift to *bottom-up, data-driven* approaches has resulted in a lack of information for understanding the interactions that are now detected *en masse*. In addition, the prevalent focus on quantity may have also resulted in a compromise on the quality of the interaction data, as high error rates have been detected in interaction data generated by current high throughput methods (von Mering *et al.*, 2002). This calls for a need to validate the detected protein-protein interactions with other means. It is a key bioinformatic and experimental challenge now to explore methods that can characterize and validate the large quantities of detected protein interactions in a reliable and efficient manner.

This paper attempts to address this problem by focusing on domain-domain interactions. As protein-protein interactions involve physical interactions between the proteins' subunits or *domains*, domain-domain interactions can be useful for validating, annotating, and even predicting protein interactions. However, unlike protein-protein interaction detection where large-scale experiments have been performed to elucidate the map of various species' interactomes (Ito *et al.*, 2001; McCraith *et al.*, 2000; Rain *et al.*, 2001; Uetz *et al.*, 2000), high-throughput experimental results for domain-domain interactions remained unavailable. Deriving such information with bioinformatic means must be considered.

We describe an integrative approach to computationally infer putative domain-domain interactions from heterogeneous data sources ranging from Rosetta stone sequences to protein interactions and complexes. We use a confidence scoring system to integrate interaction information derived from multiple data sources. We show that such an integrative approach, which draws from multiple data sources and methods, can provide higher confidence predictions and better coverage than a non-integrative approach. We study the strengths of using domain interactions as evidential supports for protein interactions, and illustrate how they are best used for validating detected protein interactions and complexes.

2. BACKGROUND

Domains are modules of amino acid sequence within the proteins themselves with specific evolutionarily conserved motifs that have structural or functional implications. Protein domains are therefore reusable sequence units that can be found in multiple protein contexts, and all proteins can in principle be characterized by combinations of domains. The domains form the structural or functional units of proteins that partake in intermolecular interactions. The existence of certain domains in proteins can therefore suggest the propensity for the proteins to interact or form a stable complex to bring about certain biological functions. The analysis of many protein-protein interactions can thus be reduced to understanding the underlying domain-domain interactions between the proteins.

Researchers have recently begun to investigate the use of domain-domain interactions for *in silico* prediction of protein-protein

interactions. Wojcik and Schächter (Wojcik *et al.*, 2001) have shown that using domain profile pairs can provide better prediction of protein interactions than using full-length sequences. Gomez and Rzhetsky (Gomez *et al.*, 2002) explored the use of domain interaction with network topology to predict protein-protein interactions statistically, while Deng *et al.* (Deng *et al.*, 2002) recently devised a maximum likelihood approach to infer domain-domain interactions and then used the inferred domain-domain interactions to predict protein interactions. The domain interaction information in these related works were either implicitly or explicitly derived solely from known protein-protein interactions.

Alternative computation means for predicting protein-protein interactions using protein domains have also been considered. The gene fusion method, also known as the "Rosetta Stone" method, has been used to predict protein-protein interactions (Enright *et al.*, 1999; Marcotte & Pellegrini & Ng *et al.*, 1999), as well as in combination with other non-homology methods to computationally assign functional links between proteins (Marcotte & Pellegrini & Thompson *et al.*, 1999). By focusing on domains instead of genes, the modified domain fusion method can also be used to infer domain-domain interactions from sequences in different species.

Results from previous works have shown that domain-domain interactions are good common denominators for protein interaction prediction (Deng *et al.*, 2002; Sprinzak *et al.*, 2001; Wojcik *et al.*, 2001). In these previous works, domain interactions were inferred solely from known protein-protein interactions. Here, we adopt an integrative approach that uses multiple data sources, including experimentally derived protein interactions, inter-molecular relationships in protein complexes, and computationally predicted Rosetta stone sequences, to collectively infer putative domain-domain interactions. Such an integrative approach should provide better coverage and enhance prediction reliability, as interactions independently derived from multiple data sources and methods are more likely to be genuine than those derived from a single data source or method. With a database of high quality putative domain-domain interactions in terms of coverage and reliability, better global analysis of protein-protein interactions can then be achieved.

3. MATERIALS AND METHODS

Our integrative approach uses multiple data sources for inferring interaction information. Currently, we use three different data sources: protein interactions, protein complexes, and domain fusions. Of course, this approach allows additional methods and data sources to be incorporated for even higher coverage and better quality predictions.

Once domain-domain interactions are inferred from the various data sources, they are integrated into a common database and sorted with a confidence scoring system that assigns higher scores to domain interactions that are more certain and multiply derived. This database of putative domain-domain interactions can then be used for validating, annotating, and predicting protein-protein interactions.

3.1 Domain Characterization of Proteins

The very first step is to characterize the input proteins by their respective protein domains, reducing protein-protein interactions to domain-domain interactions. We refer to the Pfam database (Bateman *et al.*, 2002) for pre-defined protein-domain relationships instead of deriving our own domain profiles such as in (Wojcik *et al.*, 2001). Pfam contains a large collection of multiple sequence alignments and profile hidden Markov models (HMM) covering the majority of protein domains. Proteins not listed by the Pfam database can be aligned with a profile HMM constructed from the seed alignment using the HMMER2 software (http://hmmer.wustl.edu) (Durbin *et al.*, 1998).

The Pfam database consists of two classes of domains: Pfam-A and Pfam-B. The domains from Pfam-A are manually curated and functionally assigned, whereas domains from Pfam-B are automatically generated by programs based on the ProDom database (Corpet *et al.*, 2000). Results from previous works have shown that it is advantageous to use a larger set of domains to ensure sufficient coverage; for example, the assignment of domains to the proteins was found to be a major limitation in (Sprinzak *et al.*, 2001), where their usable data were reduced by 50% because many of the interacting proteins cannot be assigned with a recognizable domain. In our case, Pfam-A also covers only 52.8% of our training proteins. As our main objective is to use domain-domain interactions to validate protein interactions, it is important to use as large a set of domains as possible to ensure coverage. Futhermore, although the overall quality of the Pfam-B domains may not be as good as the manually curated Pfam-A domains, the Pfam-B domains that emerge eventually in domain-domain interactions are quite likely to be genuine domains even though they are yet to be manually curated. We therefore used both Pfam-A and Pfam-B to characterize the interacting proteins in our training set.

3.2 Inference of Domain-Domain Interactions

Three different data sources are currently used in our integrative approach for inferring domain-domain interaction information: experimentally derived protein-protein interactions, inter-protein relationships in detected protein complexes, and predicted domain fusion events.

3.2.1 Protein-protein interactions

The conventional data source for deriving domain-domain interactions is from pairwise protein-protein interactions. This was the method used in previous works (Bock *et al.*, 2001; Deng *et al.*, 2002; Gomez *et al.*, 2002; Sprinzak *et al.*, 2001; Wojcik *et al.*, 2001). Given two proteins that are known to bind to one another (e.g. in yeast-two-hybrid experiments), we can infer that some domains from the two sets of domains from the proteins could potentially interact with one another. For example, if two proteins P_r and P_s are known to bind to each other, then we infer that the domain $d_{r,i}$ potentially interacts with domain $d_{s,j}$ with a minimal probability of $\frac{1}{m_r m_s}$, where m_r and m_s are the number of domains in proteins P_r and P_s respectively, and $d_{r,i}$ and $d_{s,j}$ are the i^{th} and j^{th} domains of proteins P_r and P_s respectively.

Here, we used the protein-protein interaction data from the DIP (Xenarios *et al.*, 2002) database, a comprehensive curated catalog of about 18,000 experimentally determined interactions between proteins from over 110 organisms. For evaluation, we used only the 9,708 yeast interactions in DIP, and derived 38,524 possible interacting domain-domain combinations. Of course, many of

these domain-pairs could be chanced occurrences; we will be using a probabilistic scoring system to weed out these spurious predictions.

3.2.2 Protein complexes

Most biological functions involve the formation of protein complexes; several proteins can come together to form a multi-protein complex. Domain interaction information can be inferred from the inter-molecular relationships in these protein complexes.

Suppose proteins P_1, …, P_n are known to form an *n*-protein complex, , we can infer that domain $d_{r,i}$ potentially interacts with domain $d_{s,j}$ with a minimal probability of $\binom{n}{2}^{-1} \cdot \frac{1}{m_r m_s}$, where m_r and m_s are the number of domains in proteins P_r and P_s respectively, and $d_{r,i}$ and $d_{s,j}$ are the i^{th} and j^{th} domains of proteins P_r and P_s respectively. We currently used a set of 232 yeast protein complexes that comprises of an average of 11.5 proteins per complex from the Cellzome database (McCraith *et al.*, 2000), together with 7,451 complexes from the PDB (Westbrook *et al.*, 2002) that have at least 2 chains and no more than 5 different proteins, as a second additional data source to derive domain-domain interactions. A total of 11,102 putative interacting domain pairs were inferred from this second data source.

3.2.3 Domain fusions

The previous two data sources were both experimentally determined. For a third data source, we employed one that was computationally predicted. Scientists have observed that some pairs of interacting proteins have homologs in another organism that are fused into a single protein chain. For instance, separate genes encoding two interacting proteins in the yeast genome might be found as a single gene encoding a longer fused protein in the human genome. This observation can be used as a basis for predicting protein-protein or domain-domain interactions: if proteins, or protein domains, disparate in one organism are fused together in a second organism, it suggests that they may function or interact together in the first organism. The fused protein sequence P_r-P_s is called Rosetta Stone Sequence (Enright *et al.*, 1999; Marcotte & Pellegrini & Ng *et al.*, 1999).

The domain fusion method therefore looks for protein domains that are separate in one organism but fused together in another to postulate potential interactions between the domains. Scanning the SWISS-PROT (Bairoch *et al.*, 2000) database that contained proteins from over 7,000 species, the domain fusion method yielded 4,792 putative domain-domain interactions[+].

3.3 An Integrative Scoring System

A weighted scoring system was devised to integrate the interactions derived from the heterogeneous data sources in a systematic way, assigning higher confidence to domain interactions that are more certain and are derived from multiple sources.

[+] Only Pfam-A domains are available in the third-party data source that we used (see Acknowledgements).

3.3.1 Scoring inferences from protein interactions

The putative interacting domain pairs were generated by combination between the sets of domains in the interacting proteins, which means that many of the inferred domain pairs were random occurrences. To detect those that are more likely to genuine, we compare the observed weighted frequencies of domain pairs against the corresponding expected frequencies of domain pairings by random occurrence. The greater the observed frequency is over the expected frequency, the more confident we can be about the inferred domain interactions.

We compute the observed and expected weighted frequencies of an inferred interacting domain pair $<d_x,d_y>$ as follows:

$$O_{\text{int}}(x,y) = \sum_{i=1}^{N_{\text{int}}} w_i^{\text{evidence}} \cdot w_i^{\text{domain}} \cdot \lambda_i(x,y)$$

$$E_{\text{int}}(x,y) = \sum_{i=1}^{N_{\text{int}}} w_i^{\text{evidence}} \cdot w_i^{\text{domain}} \cdot 2f(x)f(y)$$

where

N_{int} = number of protein-protein interactions used for training

w_i^{evidence} = weight of evidence supporting the *i*-th protein-protein interaction
= total number of distinct experiments detecting the *i*-th protein interaction[*]

w_i^{domain} = weight of domain pair being responsible for the *i*-th protein-protein interaction
= minimal probability $\frac{1}{m_{i,r} m_{i,s}}$, as described previously

$\lambda_i(x,y)$ = total number of occurrences of the domain pair $<d_x,d_y>$ in the *i*-th protein interaction[#]

$f(x)$ = frequency of domain d_x found in the interacting proteins of the training set

We define the confidence score for a derived interacting domain pair $<d_x,d_y>$ as the number of times it was observed more than it was expected as a random occurrence:

$$S_{\text{int}}(x,y) = O_{\text{int}}(x,y) / E_{\text{int}}(x,y)$$

This scoring scheme is based on odd-ratios; as such, domain-domain interactions that are derived from multiple protein interactions and are less likely to be chanced occurences would be favored. The probablistic weighting scheme allows the assignment of higher scores to those inferred from interacting proteins with fewer domains (hence, more certain).

[*] Here we are treating protein-protein interactions that have been independently observed from multiple experiments as equivalent to being separate interactions for inference.

[#] Although the domain pairing relation $<d_x,d_y>$ is symmetric, $<d_x,d_y>$ and $<d_y,d_x>$ are counted separately, as are multiple occurrences of the same domain in either proteins.

3.3.2 Scoring inferences from protein complexes

For protein complexes, we can assign confidence scores in a similar fashion. Here, the observed and expected weighted frequencies are:

$$O_{\text{cplx}}(x,y) = \sum_{i=1}^{N_{\text{cplx}}} w_i^{\text{evidence}} \sum_{j=1}^{M_i} w_{i,j}^{\text{domain}} \cdot \lambda_{i,j}(x,y)$$

$$E_{\text{cplx}}(x,y) = \sum_{i=1}^{N_{\text{cplx}}} w_i^{\text{evidence}} \sum_{j=1}^{M_i} w_{i,j}^{\text{domain}} \cdot 2f(x)f(y)$$

where

N_{cplx} = number of protein complexes in the training set

M_i = number of possible protein-protein pairs in the i-th protein complex
= $\binom{n_i}{2}$ where n_i is the number of proteins in the i-th protein complex

W_i^{evidence} = weight of evidence supporting the i-th protein complex
= total number of distinct experiments detecting the i-th protein complex

$W_{i,j}^{\text{domain}}$ = weight of a domain pair being responsible for the j-th pairing of proteins in the i-th protein complex
= minimal probability $\binom{n_i}{2}^{-1} \cdot \frac{1}{m_{j,r} m_{j,s}}$, or $\frac{1}{M_i m_{j,r} m_{j,s}}$, as described previously

$\lambda_{i,j}(x,y)$ = total number of occurrences of the domain pair $<d_x,d_y>$ in the j-th pairing of proteins in the i-th protein complex

$f(x)$ = frequency of domain d_x found in all the protein components of the training set

Again, the confidence score for a domain interaction $<d_x,d_y>$ inferred from protein complex data is:

$$S_{\text{cplx}}(x,y) = O_{\text{cplx}}(x,y) / E_{\text{cplx}}(x,y)$$

3.3.3 Scoring inferences from domain fusions

For domain-domain interactions $<d_x,d_y>$ inferred from predicted domain fusion events, instead of using a probablistically weighted odd-ratio scheme similar to those described above, we currently assign a standard scoring of $S_{\text{fus}}(x,y) = 2$ (to indicate a non-chanced occurrence). This is because we have obtained our domain fusion data from a third party (see Acknowledgements) and the background data for deriving a probabilistic scoring were unavailable at the time of writing.

3.3.4 Putting it together

For each independently inferred domain-domain interaction $<d_x,d_y>$, we compute a combined weighted confidence score as follows:

$$score(x,y) = w_{\text{int}} S_{\text{int}}(x,y) + w_{\text{cplx}} S_{\text{cplx}}(x,y) + w_{\text{fus}} S_{\text{fus}}(x,y)$$

Although this scoring scheme allows giving more weights to inferences from selected data sources that are found to be more reliable than others, we use equal weighting, $w_{\text{int}} = w_{\text{cplx}} = w_{\text{fus}} = 1$, for all three data sources in the current system. Different weights can be used later when we have established the relative usefulness of the different data sources.

4. SYSTEM

We have developed an automated interacting domain discovery system, InterDom, based on this integrative approach. The InterDom system was implemented in a UNIX environment, and the data are stored in a relational database in mySQL for scalability. Automated methods for searching the various databases and for dynamically displaying the selected tables and domain interaction graphs to the users were built with a combination of Perl, PHP, Java, and HTML

The InterDom database is accessible on the World Wide Web (http://InterDom.lit.org.sg). The site provides a useful web interface for validating and annotating detected protein-protein interactions and complexes that are computationally predicted or experimentally detected. For example, user can enter a list of two or more molecule names that have suspected interaction relationships, and the system will validate the hypothesis by linking the input molecules with potential domain-domain interactions between them. The resulting structure is laid out graphically in a java applet for easy viewing and navigation, as shown in Figure 1.

5. EVALUATION

As large-scale experimentally determined domain-domain interaction data do not exist, we could not directly assess the accuracy of our inferred domain-domain interactions. Instead, we evaluate the usefulness of our integrative approach by applying the domain-domain interactions inferred to validate experimental protein-protein interaction data. Quality of the inferred domain-domain interactions is evaluated by measuring the efects of more data sources on the true positive rates on positive protein interaction data, as well as the false positive rates on negative protein interaction data.

For true positive rates, we performed a 20-fold cross-validation on the 9,708 yeast protein interaction data from DIP. A *true positive* is a protein interaction that can be validated with at least one domain-domain interaction inferred from the data sources used. For false positive rates, because sufficient experimentally validated non-interacting protein pair data are currently unavailable, we generate 20 sets of 485 *putative* non-interacting protein pairs each by randomly pairing the proteins from the 20-fold cross-validation, excluding, of course, any actual interacting pairs. An "estimated" *false positive*, in this case, is a protein pair from a negative set that can be validated with an inferred domain-domain interaction.

We evaluated the quality of domain-domain interactions inferred from one data source (protein interactions), two data sources (protein interactions plus complexes), and three data sources (protein interactions, complexes, and domain fusions). The resulting average true positive and false positive rates on the yeast protein testing data set are shown in Table 1.

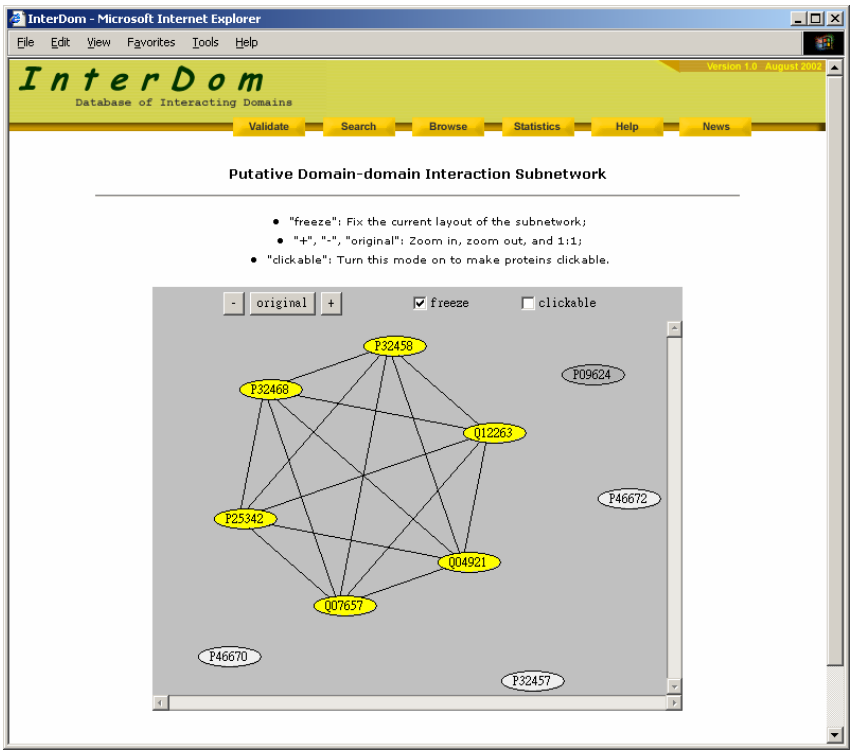

Figure 1. The InterDom system. Here, six septin-complex molecules together with other non-septin proteins were input to the system. The system found putative domain-domain interaction links between the six septin members, while the non-septin members remained unlinked.

Table 1. Average true positive and false positive rates for protein interaction validation based on inferred domain-domain interaction information from various data sources.

	Protein interactions only	**Protein interactions and protein complexes**	**Protein interactions, complexes, and domain fusions**
True positive	38.92%	58.28%	58.97%
False postive	8.49%	11.51%	12.51%

Table 1 shows that an integrative approach that uses multiple data sources for protein interaction validation is advantageous. By introducing an additional data source of protein complexes, the true postive rate was vastly improved without greatly affecting the false positive rate with the additional inferred domain-domain interactions. While the addition of the third data source (namely, domain fusions) only slightly improves the true positive rate, it also did not compromise on the false positive rate. The quality of inferred domain-domain interactions should improve as more different data sources are integrated.

Table 2 shows the overlap of the inferred domain-domain interaction from the three data sources. Currently, nearly a quarter of the inferred domain-domain interactions were independently derived from both protein-protein interaction and protein complex data sources, while a lesser degree of overlap occurs between domain fusions and interactions and complexes. The latter may be due to the fact that only Pfam-A domains had been used for domain fusion inferrence, and also that the testing data set has been restricted to only yeast proteins, as the effect of the domain fusion method may be more pronounced with multi-species testing data. Nevertheless, this result suggests the vastness of the domain interaction space. Given the existing limited coverage of the various data sources, it is therefore essential to adopt an integrative approach to combine more data sources and approaches to achieve comprehensive coverage.

Table 2. Percentage of domain-domain interactions inferred from disparate data sources.

Interactions + Complexes	**Interactions + Domain fusions**	**Complexes + Domain fusions**	**All**
23.03%	4.70%	4.06%	3.59%

6. CONCLUSIONS AND FUTURE WORK

We have presented an integrative approach for computationally inferring domain-domain interactions from heterogeneous data sources, using a probabilistic confidence scoring scheme. We have shown that by drawing from heterogeneous sources, ranging from experimentally determined protein interactions and complexes to computationally predicted domain fusion events, the

integrative approach's sensitivity in validating detected protein interactions improves as more data sources are integrated.

We plan to investigate the use of additional data sources and methods to derive domain interactions, so that we can arrive at better quality data in future. One possible data source for exploitation is the scientific literature. Scientific text mining is becoming an increasingly researched topic in post-genome bioinformatics (Mack *et al.*, 2002), as results of scientific research are still reported in scientific journals and conferences despite the proliferation of sequence and structure databases. We can use text mining approaches such as those described in (Ng *et al.*, 1999) to automatically extract domain-domain interactions, protein-protein interactions, and protein complex information from MEDLINE abstracts as additional sources of information for inferring domain-domain interactions.

The specificity of the domain-domain interaction approach for protein interaction validation could be further improved by exploring other factors that potentially underlie protein interactions, and incorporating these factors into the validation process. For example, some interactions between protein domains could be non-binary, and may depend on other non-domain factors. It is possible to employ machine learning methods to detect such complex domain-domain interactions. For example, Bock and Gough (Bock *et al.*, 2001) has used a support vector machine system to predict protein-protein interactions based on primary structures and physiochemical properties. It will be useful to explore the use of machine learning techniques to discover the more complex domain-domain interactions, as well as any other biological factors that affect domain-domain interactions. This will lead to useful knowledge for better *in silico* validation, annotation, and even prediction of protein-protein interactions.

7. ACKNOWLEDGMENTS

Our thanks to Lin Kui (Genome Institute of Singapore) for providing the domain fusion data, and all the other scientists who have made their data available on the world wide web. We are also grateful to the anonymous reviewers for their useful suggestions.

8. REFERENCES

Bairoch, A. and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*, **28,** 45-48.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res*, **30,** 276-280.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. *et al.* (2002) GenBank. *Nucleic Acids Res*, **30,** 17-20.

Bock, J.R. and Gough, D.A. (2001) Predicting protein--protein interactions from primary structure. *Bioinformatics*, **17,** 455-460.

Corpet, F., Servant, F., Gouzy, J. and Kahn, D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res*, **28,** 267-269.

Deng, M., Metah, S., Sun, F. and Chen, T. (2002).Inferring Domain-Domain Interactions from Protein-Protein Interactions. In *Proceedings of The ACM-SIGACT Sixth Annual International Conference on Computational Moledular Biology (RECOMB02)*, Washington D.C., 117-126.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) The Theory Behind Profile HMMs. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

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

Gomez, S.M. and Rzhetsky, A. (2002) Towards the prediction of complete protein--protein interaction networks. *Pac Symp Biocomput***,** 413-424.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res*, **30,** 38-41.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, **98,** 4569-4574.

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T. *et al.* (2000) Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A*, **97,** 1143-1147.

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

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. *et al.* (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285,** 751-753.

Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402,** 83-86.

McCraith, S., Holtzman, T., Moss, B. and Fields, S. (2000) Genome-wide analysis of vaccinia virus protein-protein interactions. *Proc Natl Acad Sci U S A*, **97,** 4879-4884.

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.

Rain, J.C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C. *et al.* (2001) The protein-protein interaction map of Helicobacter pylori. *Nature*, **409,** 211-215.

Sprinzak, E. and Margalit, H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, **311,** 681-692.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S. *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403,** 623-627.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G. *et al.* (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417,** 399-403.

Westbrook, J., Feng, Z., Jain, S., Bhat, T.N., Thanki, N. *et al.* (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res*, **30,** 245-248.

Wojcik, J. and Schachter, V. (2001) Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, **17 Suppl 1,** S296-305.

Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M. *et al.* (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, **30,** 303-305.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A. *et al.* (2001) Global analysis of protein activities using proteome chips. *Science*, **293,** 2101-2105.