Integrative approach for computationally inferring protein domain interactions

See-Kiong Ng*, Zhuo Zhang and Soon-Heng Tan

Laboratories for Information Technology, 21 Heng Mui Keng Terrace, Singapore 119613

Received on June 9, 2002; revised and accepted on September 11, 2002

ABSTRACT

Motivation: The current need for high-throughput protein interaction detection has resulted in interaction data being generated en masse through such experimental methods as yeast-two-hybrids and protein chips. Such data can be erroneous and they often do not provide adequate functional information for the detected interactions. Therefore, it is 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. Here, we describe an integrative approach to computationally derive putative domain interactions from multiple data sources, including protein interactions, protein complexes, and Rosetta Stone sequences. We further prove the usefulness of such an integrative approach by applying the derived domain interactions to predict and validate protein–protein interactions.

Availability: A database of putative protein domain interactions derived using the method described in this paper is available at http://interdom.lit.org.sg.

Contact: skng@lit.org.sg

1 INTRODUCTION

The genome era has given rise to the explosive discovery of new genes and proteins at an exponential rate for various living organisms (Benson et al., 2002; Hubbard et al., 2002). However, the cellular machinery is a complex dynamic system with a multitude of biomolecular interactions. Simply knowing the existence of genes and proteins does not tell us much about the biological processes in which they participate. A comprehensive description of protein–protein interactions is necessary to understand the genetic program of life.

Historically, scientists have been studying individual protein interactions by a top-down, hypothesis-driven approach, designing focused experiments to test their hypotheses about each interaction. Today, the need for high-throughput interaction detection has resulted in the generation of large quantities of protein interaction data through the use of such methods as two-hybrid systems (Ito et al., 2000, 2001; Uetz et al., 2000) and protein chips (Zhu et al., 2001). However, 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 through other means. As such, this presents a computational and experimental challenge to explore various 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 remain unavailable. We describe an integrative approach to computationally infer putative domain–domain interactions from heterogeneous data sources, such as protein interactions and complexes and Rosetta Stone sequences. We used a confidence scoring system to integrate interaction information derived from multiple data sources and showed that such an integrative approach can provide higher confidence predictions and better coverage than a non-integrative approach. We studied the strengths of using domain interactions as evidential support for protein interactions and illustrate how they can best be used for validating detected protein interactions and complexes.
2 BACKGROUND

Protein domains are modules of amino acid sequence with specific evolutionarily conserved motifs. These protein domains are the structural or functional units that participate in intermolecular interactions. The existence of certain domains in proteins might therefore suggest the probability for two proteins to interact and/or form a stable complex to elicit a biological response. 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 (2001) have shown that the use of domain profile pairs can provide better prediction of protein interactions than the use of full-length protein sequences. Gomez and Rzhetsky (2002) explored the use of domain interaction with network topology to predict protein–protein interactions statistically, while Deng et al. (2002) recently devised a maximum likelihood approach to infer domain–domain interactions that was then used 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 computational 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 in predicting protein–protein interactions (Enright et al., 1999; Marcotte et al., 1999a) and in assigning functional links between proteins (Marcotte et al., 1999b). By focusing on domains instead of genes, the ‘Rosetta Stone’ method can also be modified to infer domain–domain interactions from sequences in different species.

Results from previous works have shown that domain–domain interactions are good indicators of protein–protein interactions (Deng et al., 2002; Sprinzak and Margalit, 2001; Wojcik and Schächter, 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, intermolecular relationships in protein complexes, and computationally predicted Rosetta Stone sequences, to collectively infer putative domain–domain interactions from sequences in different species. This integrative approach should provide better coverage and enhance prediction reliability, since interactions elucidated independently from multiple data sources and methods are more likely to be accurate than those 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. Additional methods and data sources can be incorporated later 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 multiply and confidently 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 and Schächter (2001). The Pfam database 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).

There are two classes of domains in the Pfam database: 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, protein domain assignment was found to be a major limitation in Sprinzak and Margalit (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. Domain coverage is important in our proposed application of domain–domain interactions in validating protein interactions. Although the overall quality of the Pfam-B domains is not as good as the manually curated Pfam-A domains, the Pfam-B domains that emerge eventually in the high quality domain–domain interactions are quite likely to be genuine domains. 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 pair-wise protein–protein interactions. This was the method used in previous works (Bock and Gough, 2001; Deng et al., 2002; Gomez and Rzhetsky, 2002; Sprinzak and Margalit, 2001; Wojcik and Schächter, 2001). Given two proteins that are known to bind to each other (e.g. in yeast-two-hybrid experiments), we infer that certain domains from one protein potentially interact with those from the other protein. In other words, if two proteins \( P_i \) and \( P_j \) are known to bind to each other, we infer that 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.

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 9708 yeast interactions in DIP and derived 38 524 possible interacting domain–domain combinations. Of course, many of these domain-pairs could be chance occurrences; we will be evaluating, we used only the 9708 yeast interactions in DIP and derived 38 524 possible interacting domain–domain combinations. Of course, many of these domain-pairs could be chance occurrences; we will be evaluating the expected frequencies of such 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–domain interactions.

Suppose proteins \( P_1, \ldots, P_n \) 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 \( \left( \frac{2}{m} \right)^{-1} \cdot \frac{1}{m_r m_s} \), 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.

Here, we 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 another set of 7451 complexes from the PDB (Westbrook et al., 2002) that have at least two chains and no more than five different proteins, as an additional data source to derive domain–domain interactions. A total of 11 102 putative interacting domain pairs were inferred from this second data source.

3.2.2 Protein complexes. Most biological functions involve the formation of protein complexes; two or more proteins can come together to form a multi–protein complex. Domain interaction information can be inferred from the intermolecular relationships in these protein complexes.

Suppose proteins \( P_1, \ldots, P_n \) 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 \( \left( \frac{2}{m} \right)^{-1} \cdot \frac{1}{m_r m_s} \), 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.

Here, we 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 another set of 7451 complexes from the PDB (Westbrook et al., 2002) that have at least two chains and no more than five different proteins, as an 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. Scientists have observed pairs of interacting proteins having homologs in another organism that were fused into a single protein chain: separate genes encoding two interacting proteins, say 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 interactions: if proteins disparate in one organism are fused together in a second organism, it may suggest that they 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 et al., 1999a).

The domain fusion method looks for protein domains that were separate in one organism but fused together in another to postulate potential interactions between the domains. Scanning the SWISS-PROT (Bairoch and Apweiler, 2000) database that contained proteins from over 7000 species, this method has yielded 4792 putative domain–domain interactions†. Note that the data from the previous two data sources were both experimentally determined. Here, we have used data that were computationally predicted.

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. The scoring system was designed to assign higher confidence to domain interactions that are more certain (e.g. single-domain interactions) and those that are derived from multiple sources.

3.3.1 Scoring inferences from protein interactions. As described above, putative interacting protein domain pairs were generated by exhaustive combination between the sets of domains in the interacting proteins. As such, many of the inferred domain pairs were probably random occurrences. To identify those that are more likely to be genuine interactions, we can compare the observed weighted frequencies of domain pairs against the corresponding expected frequencies of such 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–domain interactions.

We compute the observed and expected weighted frequencies of an inferred interacting domain pair \( \langle d_x, d_y \rangle \) as

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

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

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

where

- \( N_{\text{cplx}} \) = number of protein–protein interactions used for training
- \( w_i^{\text{evidence}} \) = weight of evidence supporting the occurrence of the \( i \)th protein–protein interaction
- \( w_i^{\text{domain}} \) = weight of the domain pair being responsible for the \( i \)th protein–protein interaction
- \( \lambda_i(x, y) \) = total number of occurrences of the domain pair \( \langle dx, dy \rangle \) in the \( i \)th protein interaction
- \( f(x) \) = frequency of domain \( dx \) found in the proteins of the training set.

For each derived interacting domain pair \( \langle dx, dy \rangle \), we define its confidence score as the number of times the domain interaction was observed in the training data more than the number of times it was expected as a random occurrence:

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

This is a scoring scheme based on odd-ratios. Domain–domain interactions that are derived from multiple protein interactions and are therefore less likely to be chanced occurrences would be favored. The probabilistic weighting component allows higher scores to be assigned to those inferred from interacting proteins with fewer domains (hence, more certain than domain pairs inferred from interacting proteins with large numbers of domains).

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:

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

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

where

- \( N_{\text{cplx}} \) = number of protein complexes in the training set
- \( M_i \) = number of possible protein–protein pairs in the \( i \)th protein complex
- \( \lambda_i(x, y) \) = weight of the domain pair being responsible for the \( i \)th pairing of proteins in the \( i \)th protein complex
- \( f(x) \) = frequency of domain \( dx \) found in the proteins of the training set.

Again, the confidence score for a domain interaction \( \langle dx, dy \rangle \) inferred from complex data is computed as:

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

3.3.3 Scoring inferences from domain fusions. For domain–domain interactions \( \langle dx, dy \rangle \) inferred from predicted domain fusion events, instead of using a similar probabilistically weighted odd-ratio scoring scheme, 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 to us at the time of writing.

3.3.4 Putting it together. For each independently inferred domain–domain interaction \( \langle dx, dy \rangle \), we compute a overall confidence score as follows:

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

This weighted scoring scheme is designed to allow for more weights to be given to inferences from data sources that are found to be more reliable than others. However, in the current system, we used equal weighting, \( w_{\text{int}} = w_{\text{cplx}} = w_{\text{fus}} = 1 \), for all three data sources. Different weights can be used later for the different data sources when we have established their relative usefulness.
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. The derived data were stored in a relational database (mySQL) for scalability. Automated methods for searching the various databases and for dynamically displaying the selected tables and domain interaction graphs 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 either computationally predicted or experimentally detected. For example, a user can enter a list of two or more molecule names that have suspected interacting relationships; 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
Currently, large-scale experimentally-determined domain–domain interaction data do not exist. Thus, we cannot directly assess the accuracy of our inferred results of domain–domain interactions. Instead, we evaluated the usefulness of our integrative approach by applying the inferred results to validate experimental protein–protein interaction data. We also assessed the consequence of having 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. We performed a 20-fold cross-validation on the 9708 yeast protein interaction data from DIP to determine true positive rates. 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, 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 from our cross-validation study on the yeast protein testing data set are shown in Table 1.

The results in Table 1 show that an integrative approach that uses multiple data sources for protein interaction validation is advantageous. For instance, by introducing an additional data source of protein complexes, the true positive rate was vastly improved without greatly affecting the false positive rate. While the addition of the third data source (namely, domain fusions) only slightly improved the true positive rate, it also did not compromise on the false positive rate. The quality of inferred results would most likely improve as more data sources are integrated.

Table 2 shows the degree of 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 inference, and also that the testing data set has been restricted to only yeast proteins. This result suggests the potential vastness of the domain interaction space. Given the existing limitations in coverage of various data sources and experimental approaches (von Mering et al., 2002), it is therefore essential to adopt an integrative approach to combine more data sources from experimental and computational approaches to achieve comprehensive coverage.

## 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. One useful data source that we could exploit is scientific literature. As the results of scientific research are still reported primarily in scientific journals and conferences despite the proliferation of sequence and structure databases, scientific text mining has become an increasingly researched topic in post-genome bioinformatics (Mack and Hehenberger, 2002). Here, we can use text mining approaches such as those described by Ng and Wong (1999) to automatically extract domain–domain interactions, protein–protein interactions, and protein complex information from MEDLINE abstracts to provide a rich source of additional information for inferring domain–domain interactions.

For protein interaction validation applications, the specificity of the domain–domain interaction approach could be further improved by exploring other factors that potentially underlie protein interactions, and then incorporating these factors into the validation process. For example, certain interactions between protein domains could be non-binary, and may also 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 (2001) used a support vector machine system to predict protein–protein interactions based on primary structures and physicochemical properties. We hope 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.

## ACKNOWLEDGEMENTS

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, and our colleague Allen Chong for his help in improving this manuscript.

REFERENCES


