Chapter 5 (preliminary version)

Protein interaction network based prediction of domain-domain and domain-peptide interactions

Katia S. Guimarães [1,2] and Teresa M. Przytycka [1]

[1] National Center of Biotechnology Information, NLM / NIH

[2] Center of Informatics, Federal University of Pernambuco, Recife, Brazil

email: guimarak@mail.nih.gov, przytyck@mail.nih.gov

Abstract. Protein-protein interaction networks provide important clues about cell function. However, the picture offered by protein interaction alone is incomplete, because techniques for determining interactions at genome scale lack details as to how they are mediated. Stable protein interactions are thought to be largely mediated by interactions between protein domains while transient interactions occur often between small globular domains and short protein peptides, the so called linear motifs. Recently a number of computational methods to predict interactions between two domains and between a domain and a (possibly modified) peptide have been proposed. In this chapter we review representative computational methods focusing on those that use high throughput protein interaction networks to uncover domain-domain and domain-peptide interactions.

1 Introduction

Information that can be extracted from protein–protein interaction networks has a growing impact on molecular biology. It facilitates, for example, prediction of protein function (see Chapter 8) and provides insights into the organization and the evolution of protein interaction networks (Chapters 7 and 9). However protein interaction data lacks details on how these interactions are mediated. Full understanding of interaction details would provide a powerful weapon for studying diseases and for designing drug targets. The knowledge of domain interactions and protein domain composition can also be used for prediction of protein-protein interaction (Lander, Linton et al. 2001; Sprinzak and Margalit 2001; Wojcik and Schachter 2001; Deng, Mehta et al. 2002; Shmulevich, Dougherty et al. 2002; Nguyen and Ho 2006; Singhal and Resat 2007).

There are several levels of detail for describing how protein interactions are mediated: from delineating interacting domains to atomic level description of binding

sites (Chapters 3 and 6). On the highest level, protein interactions are thought to be largely mediated by interactions between domains or between a domain and a peptide (Pawson and Nash 2003). Isolated interacting domains can usually fold independently and are readily incorporated into larger multi-domain proteins.

Domain interactions are quite versatile. Some domain-domain interactions are ***general*** (also called promiscuous (Riley, Lee et al. 2005)) meaning that if one protein contains one of the domains and another protein contains the other domain then the two proteins are highly likely to interact. However, many domain interactions, especially the ones involved in cell regulatory systems are highly ***specific*** where in a specific interaction, domains can interact or not, depending on a broader context, like cycle-dependent expression, localization in the cell, specific amino-acid sequence features, etc. For example, the interaction between Cyclin C and Pkinase is specific, since the corresponding domains are present in a large number of non-interacting protein pairs (Riley, Lee et al. 2005). Some domains interact only with other domains, others interact with peptides, but some domains (e.g. PDZ) can interact with a domain or a peptide (Pawson and Nash 2003).

Because of importance of the information on binding details for understanding protein interactions, prediction of interacting domains pairs and domain-peptide interactions receive a significant amount of attention in computational biology research. In this chapter, we discuss representative paradigms which are based on high-throughput protein interaction networks.

2 Predicting domain interactions from protein interaction networks

Most proteins contain two or more domains (Apic, Gough et al. 2001) and a protein interaction typically involves binding between two or more specific domains. Interacting domain pairs are often reused within the interactome of an organism and many of them are evolutionarily conserved from prokaryotes to eukaryotes. The relevance of this observation is even more significant in view of recent reports suggesting that domain interactions among several organisms may be more conserved than the protein interactions themselves (Itzhaki, Akiva et al. 2006).

In this section, we discuss methods that directly use the interaction network to predict domain-domain interaction. As representative methods, we selected Association, Maximum Likelihood Estimation, Domain Pair Exclusion Analysis, Parsimonious Explanation, and an integrative method. Other approaches that also decipher interacting domains from protein interaction networks include support vector machines (Bock and Gough 2001) (supervised learning methods are reviewed in Chapter 2), probabilistic network modeling (Gomez and Rzhetsky 2002), and lowest p-

value method (Nye, Berzuini et al. 2005). Obviously, protein interaction network is by no means the only source of information that can be used to predict interacting domains. For example, the gene fusion method (Marcotte, Pellegrini et al. 1999), discussed in Chapter 4, can be applied to detect domain interactions (Ng, Zhang et al. 2003). Similarly, Pagel and colleagues constructed a domain interaction map based on phylogenetic profiling (Pagel, Wong et al. 2004). More recently, Jothi and colleagues proposed mirror tree based approach (see Chapter 4) to identify interacting domain pairs (Jothi, Cherukuri et al. 2006; Kann, Jothi et al. 2007).

For methods that are based on protein-protein interaction network, some domain-domain interactions are more difficult to discover than others. An obvious limitation is the number of experiments which report interactions between proteins mediated by a given domain pair. Additional difficulty arises when a domain pair occurs predominantly in the context of interacting proteins that have multiple ***potential domain contacts***, that is, domain pairs that can potentially mediate a given interaction. In contrast, an interacting domain pair may have one or more ***witnesses,*** that is, interacting single-domain protein pairs in which one protein contains one interacting domain while the second protein contains the other domain. In other words, a witness to a domain interaction is an interacting protein pair which, under the assumption that protein interaction is mediated by domain interaction, can only be explained by interaction between the given domains. Obviously, if an interacting domain pair has enough witnesses to compensate for unreliability of high throughput protein interaction data, discovering such pair is trivial. To separate the trivial predictions from more difficult ones, Riley and colleagues (Riley, Lee et al. 2005) associate with each domain a measure called ***modularity***, which is equal to the average number of domains in proteins containing the given domain. A non-trivial prediction of interacting domain pairs would then involve at least one domain, out of the pair, with modularity above some threshold (in their study 2.0). High modularity, however, does not exclude the possibility that a given domain pair has witnesses, and even an isolated occurrence of a domain in a protein with a large number of domains increases the modularity of the domain significantly, without necessarily making the prediction process more difficult. Therefore, Guimarães and colleagues (Guimaraes, Jothi et al. 2006) adopt a more stringent partition into easy and difficult predictions. A domain-domain interaction is considered to be ***difficult*** to predict (from the underlying protein-protein interaction network) if it does not have witnesses and otherwise it is considered easy.

2.1 Association method

Association methods detect over-represented domain pairs in interacting protein pairs. In particular, the method proposed by Sprinzak and Margalit scores each domain pair by the log ratio of the frequency of occurrences in interacting proteins to the frequency of independent occurrences of those domains(Sprinzak and Margalit

2001). That is, if P_i is the observed frequency of domain i in the interaction network and P_{ij} is the observed frequency of domain pair (i, j) as a potential domain contact in interacting protein pairs, then

$$Association\_Score(i, j) = \log \frac{P_{ij}}{P_i P_j}$$

A similar but more sophisticated score has been used by Ng and colleagues (Ng, Zhang et al. 2003) in the construction of the domain interaction database InterDom. In their scoring formula they take into account that interactions between proteins with a smaller number of potential domain contacts provide a stronger evidence for domain interactions than interaction between multi-domain proteins, so the interactions are weighted accordingly. The score is computed as:

$$InterDom\_subScore(i, j) = \frac{\sum_{k=1}^{N} \#ex_k \cdot \frac{1}{n_k} \cdot n_k^{ij}}{\sum_{k=1}^{N} \#ex_k \cdot \frac{1}{n_k} \cdot (2 \cdot P_i \cdot P_j)},$$

where N is the number of edges in the protein-protein interaction network, $\#ex_k$ is the number of distinct experiments in the network detecting protein interaction k, n_k is the number of potential domain contacts in protein interaction k, n_k^{ij} is the number of potential domain contacts between pair (i, j) in protein interaction k, and P_i is, as before, the observed frequency of domain i in the proteins of the network. A similarly defined score is computed from protein complexes. The full score for an interaction between domains includes, in addition to the two aforementioned terms, an additive term set to 2.0 if a domain pairs is related by fusion event (see Chapter 4), and 0.0 otherwise.

2.2 Maximum Likelihood Estimation (MLE)

The main idea of the Maximum Likelihood Estimation (MLE) approach (Deng, Mehta et al. 2002) is to estimate, for each domain pair, the probability of interaction between domains so that the likelihood of the interaction network is maximized. An important feature of this method is that it allows that the false positives and false negatives of the high-throughput data that constitutes the protein interaction network be explicitly factored in. Here, protein-protein interactions and domain-domain interactions are treated as random variables denoted by P_{AB} and D_{ij}, respectively. $P_{AB} = 1$ if proteins A and B interact, and $P_{AB} = 0$ otherwise. In a similar manner, $D_{ij} = 1$ if domains i and j interact, and $D_{ij} = 0$ otherwise.

Under the assumption that two proteins *A* and *B* interact if and only if at least one of their potential domain contacts (*i*, *j*) interacts, the probability of interaction between two proteins A and B is obtained as:

$$\Pr(P_{AB}=1) = 1.0 - \prod_{D_{ij} \in P_{AB}} (1-\lambda_{ij}), \quad \textbf{(1)}$$

where $\lambda_{ij} = \Pr(D_{ij} = 1)$ denotes the probability that domain *i* interacts with domain *j* and $D_{ij} \in P_{AB}$ is the set of potential domain contacts in the protein pair (A, B).

Let the random variable O_{AB} describe the experimental observation of an interaction between proteins *A* and *B*; $O_{AB} = 1$ if an interaction between proteins *A* and *B* is observed and $O_{AB} = 0$ otherwise. Denoting false positive and negative rates respectively by *fp* and *fn* we have

$$\Pr(O_{AB}=1) = \Pr(P_{AB}=1)(1-fn) + (1-\Pr(P_{AB}=1))\,fp. \quad \textbf{(2)}$$

The goal of the MLE method is to estimate parameters λ_{ij} to maximize the likelihood function *L* given by

$$L = \prod_{(A,B)|O_{AB}=1} \Pr(O_{AB}=1) \prod_{(A,B)|O_{AB}=0} 1-\Pr(O_{AB}=1). \quad \textbf{(3)}$$

Hence, denoting by λ the vector composed of all λ_{ij}, the likelihood *L* is a function of λ, *fp*, and *fn*. Deng and colleagues estimated *fp* and *fn* to be *fp* = 2.5E-4 and *fn* = 0.80. The values λ_{ij} are computed using expectation maximization (EM) that maximizes *L*. In each iteration, *t*, values of λ_{ij}^{t-1} are used to compute $\Pr(O_{AB}=1 \mid \lambda^{t-1})$ using equations (1) and (2), and update the parameters using the following Expectation and Maximization steps:

Expectation Step: $$E(D_{ij}^{AB}) = \frac{\lambda_{ij}^{t-1}(1-fn)^{O_{AB}}\, fn^{(1-O_{AB})}}{\Pr(O_{AB}=1 \mid \lambda^{t-1}, f_n, f_p)}$$

Maximization Step: $$\lambda_{ij}^{t} = \frac{1}{N_{ij}} \sum_{A,B} E(D_{ij}^{AB}),$$

where $E(D_{ij}^{AB})$ is the expectation that domain pair (i, j) negotiates the interaction between proteins A and B, and N_{ij} is the number of protein pairs in the network that have (i, j) as a potential domain pair.

2.3 Domain Pair Exclusion Analysis (DPEA)

One limitation of the MLE method is its difficulty in detecting specific domain interactions. Indeed, if the interaction between domains i and j is highly specific then λ_{ij} is likely to be small. It has also difficulty in recovering interacting domains which have high modularity. To overcome these weaknesses, Riley and colleagues proposed an alternative domain interaction prediction method, Domain Pairs Exclusion Analysis (DPEA) (Riley, Lee et al. 2005). The underlying idea behind this method is the assumption that the maximum likelihood score of a network is, in some sense, a measure of how well the probabilities assigned to putative domain interactions explain the network. Thus, if domain pair (i, j) indeed mediates some protein-protein interactions, then excluding such domain pair as a possible mediator (by fixing the parameter corresponding to λ_{ij} in the MLE method to zero) should decrease the likelihood of interactions between these proteins. This change is measured by value E_{ij} defined as:

$$E_{ij} = \sum_{\substack{\text{protein pairs } (A, B) \text{ such that} \\ (i, j) \text{ is a potential domain contact}}} \log \frac{\Pr(O_{AB} = 1)}{\Pr(O_{AB} = 1 \mid \lambda_{ij} = 0)} \quad \textbf{(4)}$$

where λ_{ij} is the probability of interaction between domains i and j estimated in a way similar to the one used in the MLE method but without including the reliability of the protein interaction network as a component of the likelihood score. Thus, the numerator is the probability that proteins A and B interact, given that domains i and j might interact. The denominator is the probability that proteins A and B interact, given that domains i and j do not interact (also estimated by the expectation maximization procedure where λ_{ij} is set to zero).

The 3,005 domain pairs with E_{ij} at least 3.0 were considered predicted to interact with high-confidence. The DPEA method was able to recover significantly more modular interactions (Riley, Lee et al. 2005) confirmed by iPFAM than the MLE method.

2.4 Parsimonious Explanation (PE)

The idea of recovering interacting domains by examining how well the potential domain contacts explain the protein interaction network has been developed further by Guimarães and colleagues (Guimaraes, Jothi et al. 2006). Based on the hypothesis that protein interactions evolved in a most parsimonious way, they proposed the *Parsimonious Explanation* (*PE*) method which finds a smallest weighted set of domain interactions that can explain the protein interaction network. This model is formalized as an optimization problem and solved with a Linear Programming procedure. The variables of the linear program represent the potential domain contacts derived from the protein interaction network, and the constraints are given by each protein-protein interaction (edge) in the given network as described below. Those variables can take real values between 0 and 1. The constraint imposed by a given protein interaction enforces that the values of the variables representing the potential domain pairs of that interaction add up to at least 1.0. The construction is illustrated in Figure 1.

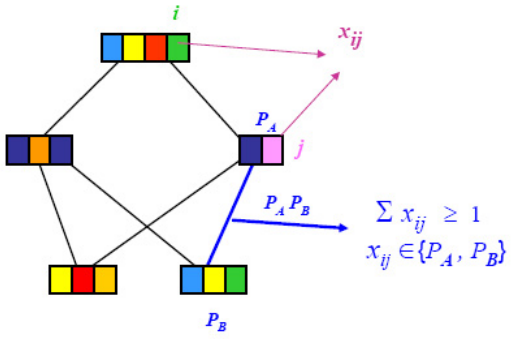

Figure 1 Construction of a Linear Program from a given protein interaction network.

According to the parsimony principle, the objective function aims to minimize the overall sum of the variables x_{ij}. Formally, if PDP is the set of the potential domain pairs found in the protein interaction network, and PPI is the set of protein-protein interactions in the given network, then the linear program is given by:

$$Minimize \sum_{(i,j)\in PDP} x_{ij}$$

$$Subject\ to: \forall (A,B) \in PPI\left(\sum_{(p,q)\in(A,B)} x_{pq}\right) \geq 1$$

The value assigned to the variable x_{ij} reflects the contribution of the domain pair (i, j) in explaining the network under the maximum parsimony principle. In the PE method, the false positives of the protein interaction network are modeled by performing a randomization process. In particular, 1000 instances of protein interaction network are constructed which are sub-networks of the input network where each edge is maintained with probability equal to the estimated reliability of the network (in (Guimaraes, Jothi et al. 2006) this value was set to 0.5). For each such randomized network, the corresponding linear program is constructed using the procedure described above, and solved. The reported score, called *LP-score*, of a given domain pair (i, j), is computed by the arithmetic average of the values x_{ij} returned by these 1000 linear programs.

In addition to the LP-score, the PE method offers the so called pw_score which quantifies the confidence in the LP-score. The pw_score of a domain pair (i, j) is computed as the minimum of two measures, the p-value of domain pair (i, j), computed from simulations, and a confidence estimation provided by the possible existence of witnesses. The combined witness and p-value score is expressed as:

$$pw\_score = \min(p\_value(i, j), (1-r)^{w(i,j)})$$

where r is the estimated reliability of the network and $w(i,j)$ is the number of witnesses of domain pair (i, j).

Unlike the previously discussed methods, the Parsimonious Explanation method was able to detect a significant number of difficult interactions confirmed by crystal structures in iPFAM.

2.5. Integrative approaches

With the exception of the scoring function of (Ng, Zhang et al. 2003), all methods discussed so far were based exclusively on protein interaction data and protein domain composition. More recently, Lee and colleagues (Lee, Deng et al. 2006) proposed a Bayesian approach that complements the protein interaction data with other information about domains; we call their method *Integrative Bayesian* (*IB*).

In the IB method, the expectation of the domain pair interaction is computed separately for each of four organisms, yeast, worm, fruit fly, and humans. The scores for the domain pairs are obtained using a method similar to MLE. The likelihood function is the same as the one used by the MLE method (Deng, Mehta et al. 2002), however, instead of using $\Pr(D_{ij} = 1)$ directly to score the domain interactions, the IB method scores each domain pair by the expectation of the domain pair interaction given by

$$E(\#D_{ij}) = N_{ij} \cdot \Pr(D_{ij} = 1),$$

where, as before, N_{ij} is the number of protein pairs in the network that have (i, j) as a potential domain contact.

The results obtained for the four networks are considered as four independent pieces of information and used as features in the integrative model. Two additional features considered are the number of times the two domains in the pair appear together, or co-exist, in one protein chain, and the information if the two domains belong to the same biological process as assessed by the Gene Ontology (GO) database (Harris, Clark et al. 2004). The scores of all domain pairs with respect to each distinct feature are binned. The likelihood score of a domain pair is computed based on the ratio of domain pairs confirmed by crystal structures to the number of domain pairs not confirmed by crystal structures in the bin containing the score of the given domain pair.

It is interesting to examine how the information which is not obtained based in protein interaction influences the prediction of this method. To elucidate this, Lee and colleagues (Lee, Deng et al. 2006) performed a comparison using the domain pairs in iPFAM as true positives, and the remaining domain pairs as true negatives. The results of that comparison are reported in Figure 4 of their paper, which shows the relationship between the false positive rate (FP / (FP+TN)) and the sensitivity (TP / (TP+FN)) of the predictions based on different combinations of information. By this evaluation standard, the Gene Ontology terms combined with the domain co-existence gives a better iPFAM pairs recovery than information obtained from the interaction networks using MLE type analysis (see also the discussion in the next section).

Very recently, Wang et al (Wang, Segal et al. 2007) introduced a different integrative method, InSite. In addition to the evidences used in the IB method described above they included Prosite (Hulo, Bairoch et al. 2006) motifs treating them in the same way as protein domains. Unlike previous methods, they score domain contacts in the context dependant manner. That is, the score of the same domain pair depends on the protein pair where a given domain pairs makes a potential contact. To obtain such score they use a method similar to the one proposed by Riley et. al (Riley, Lee et al. 2005) (see section 2.3 of this chapter, equation (4)). However, rather than looking at the effect of disallowing all interactions between a given domain pair, they consider the effect of disallowing single instance of such interaction as possible mediator of a particular protein interaction. This allowed them to measure how well given domain interaction explains the given protein interaction. That is they disallow the domain interaction “locally” rather than “globally”.

2.6 Evaluation of domain-domain interaction prediction methods

Due to a low coverage of experimentally confirmed domain-domain interactions, evaluation of the accuracy of genome scale methods to predict domain-domain and domain-peptide interactions poses a formidable challenge. One method used to evaluate the quality of predictions is by estimating how accurately one can reconstruct the protein interaction network based on the assigned domain-domain interaction scores (Deng, Mehta et al. 2002). However the quality of prediction of protein-protein interaction is not necessarily a good measure of correct prediction of domain-domain interaction. While domain pairs that make non-specific interactions are good predictors of protein interactions, the specific domain interactions are not.

An alternative method for assessing predictions was proposed by Nye et al. (Nye, Berzuini et al. 2005). The basic idea is to test if in each pair of interacting proteins, the domain pair with the highest score is correctly predicted as the domain pair mediating the interaction. The test set contains only interacting protein pairs with multiple potential domain contacts and at least one domain pair that is known to interact (e.g. based on the information from the iPFAM database). Guimarães and colleagues (Guimaraes, Jothi et al. 2006) used this method applied to 1780 protein interactions to compare the performances of several domain-domain prediction methods. In that assessment, the Association and the MLE methods achieved a positive predictive value (PPV = TP/(TP+FP)) around 11%, far below the 27% obtained if a potential domain contact had been chosen at random for each protein pair in the set. The DPEA and the PE methods achieved PPV values of 43% and 75%, respectively. That comparison used the Expectation Maximization scores of Riley and colleagues (Riley, Lee et al. 2005). Since, unlike the other methods compared, the IB method excludes PFAM-B as possible interacting domains, and its predictions were made based on a different data set, IB was not included in the above comparison. However, in a similar estimation including only 456 protein interactions whose potential domain contacts all have IB score above 0.0, the performance of the IB method is similar to that of PE approach (Guimaraes and Przytycka (unpublished data)). The InSite method has been published when this review was virtually completed. It uses a different data set and the scores for the domain pairs were not made available at this time so could not be included in the comparison.

Another method often used to evaluate the quality of domain interaction predictions estimates how well a given method recovers known domain-domain interactions. In this approach, known domain interactions (e.g. domain interactions from iPFAM (Finn, Mistry et al. 2006), 3DID (Stein, Russell et al. 2005) or CBM (Shoemaker, Panchenko et al. 2006)) are considered as true positive and all other domain pairs as true negatives. Under this assumption one can make false positive rate versus true positive rate (or similar) plots. Indeed, if a method is successful, then the corresponding curve should demonstrate a performance clearly better than ex-

pected by chance. Among the methods discussed, the highest percentage of iPFAM domains in the top 50 predicted interactions has the InSite method (Wang, Segal et al. 2007). However, the number of experimentally confirmed domain-domain interactions is very small relatively to the number of estimated domain-domain interactions. According to a recent study involving E. coli, yeast, worm, fly, and human data, conducted by Itzhaki and colleagues (Itzhaki, Akiva et al. 2006), the percentage of protein-protein interactions to which high-confidence domain-domain interactions from iPFAM or 3DID could be mapped is no more than 20% for any of the organisms. Therefore, any domain-domain prediction method that undertakes the task of explaining protein interactions through domain-domain interactions is expected to *correctly* recover domain pairs that are not in those high-confidence databases yet.

Riley et al. bypassed the above problem by selecting a set of true positives among known interacting domain pairs and a set of true negatives (of a similar size) as a set containing domain pairs which belong to interacting protein pairs but do not interact (as confirmed based on available crystal structures of protein complexes). Under this assumption, they tested how many of such true positives and true negatives have been correctly predicted. Using this criterion they estimated that the DPEA method has the specificity of 97% and the sensitivity of 6% (Riley, Lee et al. 2005).

Finally, while evaluating domain interaction prediction methods one has to be careful to avoid circularity. Due to a greater interest in some specific domains or functional roles, it is quite possible for some methods to be trained on one type of data and then be evaluated on data that is indirectly related to the one used for training, bringing up opportunity for an artificially inflated performance. Methods that use functional annotation data in particular risk for such circularity (Zhang, Wong et al. 2004; Suthram, Shlomi et al. 2006).

3 Predicting Domain-Peptide interactions from protein interaction networks

The methods to predict domain-domain interactions described in the previous section rely on the assumption that protein interactions are mediated by domain-domain interactions. This assumption is well supported for stable protein complexes. However much of the signaling, trafficking, and targeting is mediated by reversible interactions between small globular domains and short protein peptides, the so called *linear motifs*. One of the best studied examples is the SH3 domain which binds to Proline reach motif PxxP (where x represents any arbitrary amino-acid). A linear motif may, but does not have to be part of a globular domain. In fact, most of such motifs are not (Puntervoll, Linding et al. 2003; Neduva, Linding et al. 2005). Furthermore, domain-peptide interactions are often very specific, that is homologous domains often bind to different (although related) linear motifs. For example, PxxP

is the canonical binding motif for the SH3 domain while a motif for a subclass is often more specific (Toro, Thore et al. 2001). One of the first computational problems considered in the context of domain–peptide binding is that of identifying linear motifs that are recognized by a given binding domain (Reiss and Schwikowski 2004; Ferraro, Via et al. 2006; Lehrach, Husmeier et al. 2006). In these approaches experimentally determined SH3 domain-peptide interactions serve as a training set for discovering binding motifs of SH3 domains.

Recently it has been recognized that high throughput interaction networks also provide valuable resource in prediction of protein-peptide interactions. In this section we discuss briefly two methods that take advantage of this information.

3.1 Discovering domain-peptide interactions from protein interaction networks

The short length of linear motifs makes their reliable discovery computationally challenging. Recently, several related approaches have been developed that find statistically over-represented motifs in non-homologous sequences with a common property, for example that bind to a certain kinase or phosphatase (Neduva, Linding et al. 2005; Davey, Shields et al. 2006). We describe here the method of Neduva et al., since this method combines discovery of liner motifs with prediction of direct domain-peptide interaction based on high throughput protein interaction networks.

In the work of Neduva and colleagues (Neduva, Linding et al. 2005) the putative linear motifs are identified as sequence fragments observed sufficiently often in protein sequences after removing globular domains (identified as PFAM-A domains), trans-membrane segments, coiled-coils, collagen regions, and signal peptides. Furthermore, homologous sequences were also identified and removed. This preprocessing reduces the probability of detecting motifs shared due to evolutionary relationship or sequence motifs associated with structural motifs such as β-turns. In that approach, all non-overlapping motifs of 3-8 residues are identified using program TEIRESIAS (Rigoutsos and Floratos 1998). Common motifs are required, in particular, to agree perfectly on at least two positions and to occur in at least tree sequences in the set. The neighbors of each protein in the network are examined for occurrences of such common motifs. A common motif observed to be overrepresented among the interacting partners of a given protein is predicted to be the binding motif.

In addition to finding binding motifs of individual proteins, Neduva et al. also searched for the more general binding motifs of homology domains. To do this, they merged sets of binding partners of proteins containing a common domain. Such merged “domain sets” were then analyzed in the same manner as described above for individual proteins.

The analysis of the results obtained with this method is quite revealing. Despite the fact that the data in the protein-protein interaction networks is error-prone, the results were quite accurate, although the number of confidently predicted motifs was relatively small (11 in yeast, 26 in fly, 27 worm, and 112 in human). In all organisms under study, many of the known motifs were missed, as demonstrated by inspection, due to too few sequences with the correct motif to reach significance. The better results for the human network are attributed to the better quality of the data in hand-curated human interactions (Peri, Navarro et al. 2003) used in the study. The domain set approach was, in some instances, successful in detecting less specific motifs. For example, in the fly network the SH3 motif has been only identified on this level since there was not enough data to detect the more specific binding motifs. The authors have been able to confirm experimentally some of the predicted motifs.

3.2 Utilizing protein interaction network in discovering phosphorylation networks

Signal transduction is the primary means by which cells respond to external stimuli such as nutrients, growth factors, and stress. The dynamics of cell signaling pathways is, in large extend, governed by reversible phosphorylation (Krebs and Beavo 1979) of specific substrates performed by protein kinases. Thousands of in vivo phosphorylation sites have been discovered by targeted biochemical studies and, more recently, through spectrometry (Hjerrild, Stensballe et al. 2004). However our understanding of phosphorylation-dependent signaling networks remains incomplete. In particular, despite advances in in-vitro experiments (Ptacek, Devgan et al. 2005) it is not fully known which protein kinases are responsible for the phosphorylation of many known phosphorylation sites.

There are several computational approaches towards mapping phosphorylation sites to corresponding kinases which are based on identifying consensus sequence motifs recognized by specific kinases (Obenauer, Cantley et al. 2003; Hjerrild, Stensballe et al. 2004). However, these motifs alone are often insufficient for a unique identification of the kinases responsible for the phosphorylation of the corresponding sites. Specificity of kinase activity is also achieved through cellular localization, cell-cycle specific co-expression, binding to scaffold proteins, etc. Such information, termed by Linding et al. “contextual” (Linding, Jensen et al. 2007), if available, should also be used to enhance the accuracy of prediction of phosphorylation networks. Along these lines, a recent approach, NetworKIN, combines the motif based and contextual approach into one two-step algorithm.

During the first step on the NetworKIN algorithm, an experimentally determined phosphorylation site is mapped to a protein sequence. Then the protein family

that is likely to be responsible for the phosphorylation of the site is predicted based on the consensus motif approach. This is done by applying a neural network machine learning approach to obtain position specific scoring matrices (PSSMs) (Obenauer, Cantley et al. 2003; Hjerrild, Stensballe et al. 2004) describing biding motifs of all kinase families under study. Once the family (or families) of kinases whose members can potentially phosphorylate a given site is identified, the candidate proteins that could be responsible for the phosphorylation of the site are identified by BLAST search.

In the second stage, the set of candidate kinases in narrowed down using contextual information. The contextual information is obtained from the STRING database (von Mering, Jensen et al. 2007). This data base integrates information from curated pathways, co-occurrence in abstracts of scientific articles, physical protein interactions, co-expression, and predicted interaction based on genomic context (gene fusion, gene neighborhood, and phylogenetic profiles). All scoring schemas for all evidences were benchmarked and calibrated on signaling and metabolic pathways from KEGG database (Kanehisa, Goto et al. 2006) resulting in probabilistic scores for all evidence types. Additionally, association from orthologous protein in other organisms were included using a Bayesian scoring scheme to combine the evidence. The resulting probabilistic association network is used to find kinases that are proximal to the substrate (the protein containing the given phosphorylation site). Namely, for every candidate kinase, the Floyd-Warshall algorithm (Cormen, Leiserson et al. 2001) is used to compute the most likely path connecting this kinase to the substrate. A set of kinases with the best paths are predicted as responsible for the phosphorylation.

The work of Linding et al. demonstrated that the network-based contextual information has a tremendous impact on the prediction accuracy of phosphorylation. The authors estimated that 80% of the predictive power of their approach comes from the contextual information.

4 Conclusions and future directions

In this chapter we focused exclusively on the methods to predict domain-domain and domain-peptide interactions that use, in a significant way, high-throughput protein interaction networks. Within this group of methods we selected a set (by no means exhaustive) of representative approaches. We demonstrated that, despite the fact that high-throughput interactions are inherently noisy, they provide extremely valuable resource for predicting domain and peptide interactions. The noise in the high-throughput protein interaction data dictates, however, that the methods that are based exclusively on the network information are only capable of predicting interactions occurring multiple times.

An important and not completely resolved problem is the issue of evaluation of prediction methods. A standard way to assess such methods is to test how well they predict known interactions. Yet, the set of currently known interactions is not only very limited but since PDB data is well known to be biased (Brenner, Chothia et al. 1997; Gerstein 1998; Peng K 2004; Mestres 2005; Xie and Bourne 2005) such biases are also likely to be inherited by iPFAM. For example, in the context of domain-domain interactions, the crystal structures favor stable and well studied protein complexes. Therefore, an important issue in prediction methods is an experimental validation of new predicted interactions.

References

Apic, G., J. Gough, et al. (2001). "An insight into domain combinations." Bioinformatics **17 Suppl 1**: S83-9.

Bock, J. R. and D. A. Gough (2001). "Predicting protein--protein interactions from primary structure." Bioinformatics **17**(5): 455-60.

Brenner, S. E., C. Chothia, et al. (1997). "Population statistics of protein structures: lessons from structural classifications." Curr Opin Struct Biol **7**(3): 369-76.

Cormen, T. H., C. H. Leiserson, et al. (2001). Introduction to Algorithms, MIT Press.

Davey, N. E., D. C. Shields, et al. (2006). "SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent." Nucleic Acids Res **34**(12): 3546-54.

Deng, M., S. Mehta, et al. (2002). "Inferring domain-domain interactions from protein-protein interactions." Genome Res **12**(10): 1540-8.

Ferraro, E., A. Via, et al. (2006). "A novel structure-based encoding for machine-learning applied to the inference of SH3 domain specificity." Bioinformatics **22**(19): 2333-9.

Finn, R. D., J. Mistry, et al. (2006). "Pfam: clans, web tools and services." Nucleic Acids Res **34**(Database issue): D247-51.

Gerstein, M. (1998). "Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census." Proteins **33**(4): 518-34.

Gomez, S. M. and A. Rzhetsky (2002). "Towards the prediction of complete protein--protein interaction networks." Pac Symp Biocomput: 413-24.

Guimaraes, K. S., R. Jothi, et al. (2006). "Predicting domain-domain interactions using a parsimony approach." Genome Biol **7**(11): R104.

Guimaraes, K. S. and T. M. Przytycka ((unpublished data)). "Study of the power of the parsimony principle for prediction of domain-domain interactions."

Harris, M. A., J. Clark, et al. (2004). "The Gene Ontology (GO) database and informatics resource." Nucleic Acids Res **32**(Database issue): D258-61.

Hjerrild, M., A. Stensballe, et al. (2004). "Identification of Phosphorylation Sites in Protein Kinase A Substrates Using Artificial Neural Networks and Mass Spectrometry." J. Proteome Res. **3**(3): 426-433.

Hulo, N., A. Bairoch, et al. (2006). "The PROSITE database." Nucleic Acids Res **34**(Database issue): D227-30.

Itzhaki, Z., E. Akiva, et al. (2006). "Evolutionary conservation of domain-domain interactions." Genome Biol **7**(12): R125.

Jothi, R., P. F. Cherukuri, 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**(4): 861-75.

Kanehisa, M., S. Goto, et al. (2006). "From genomics to chemical genomics: new developments in KEGG." Nucleic Acids Res **34**(Database issue): D354-7.

Kann, M. G., R. Jothi, et al. (2007). "Predicting protein domain interactions from coevolution of conserved regions." Proteins **67**(4): 811-20.

Krebs, E. G. and J. A. Beavo (1979). "Phosphorylation-dephosphorylation of enzymes." Annu Rev Biochem **48**: 923-59.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

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

Lehrach, W. P., D. Husmeier, et al. (2006). "A regularized discriminative model for the prediction of protein-peptide interactions." Bioinformatics **22**(5): 532-40.

Linding, R., L. J. Jensen, et al. (2007). "Systematic discovery of in vivo phosphorylation networks." Cell **129**(7): 1415-26.

Marcotte, E. M., M. Pellegrini, et al. (1999). "Detecting protein function and protein-protein interactions from genome sequences." Science **285**(5428): 751-3.

Mestres, J. (2005). "Representativity of target families in the Protein Data Bank: impact for family-directed structure-based drug discovery." Drug Discov Today **10**(23-24): 1629-37.

Neduva, V., R. Linding, et al. (2005). "Systematic discovery of new recognition peptides mediating protein interaction networks." PLoS Biol **3**(12): e405.

Ng, S. K., Z. Zhang, et al. (2003). "Integrative approach for computationally inferring protein domain interactions." Bioinformatics **19**(8): 923-9.

Ng, S. K., Z. Zhang, et al. (2003). "InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes." Nucleic Acids Res **31**(1): 251-4.

Nguyen, T. P. and T. B. Ho (2006). "Discovering signal transduction networks using signaling domain-domain interactions." Genome Inform **17**(2): 35-45.

Nye, T. M., C. Berzuini, et al. (2005). "Statistical analysis of domains in interacting protein pairs." Bioinformatics **21**(7): 993-1001.

Obenauer, J. C., L. C. Cantley, et al. (2003). "Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs." Nucleic Acids Res **31**(13): 3635-41.

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

Pawson, T. and P. Nash (2003). "Assembly of cell regulatory systems through protein interaction domains." Science **300**(5618): 445-52.

Peng K, O. Z., Vucetic S. (2004). "Exploring bias in the Protein Data Bank using contrast classifiers." Pac Symp Biocomput.: 435–446.

Peri, S., J. D. Navarro, et al. (2003). "Development of human protein reference database as an initial platform for approaching systems biology in humans." Genome Res **13**(10): 2363-71.

Ptacek, J., G. Devgan, et al. (2005). "Global analysis of protein phosphorylation in yeast." Nature **438**(7068): 679-84.

Puntervoll, P., R. Linding, et al. (2003). "ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins." Nucl. Acids Res. **31**(13): 3625-3630.

Reiss, D. J. and B. Schwikowski (2004). "Predicting protein-peptide interactions via a network-based motif sampler." Bioinformatics **20 Suppl 1**: i274-82.

Rigoutsos, I. and A. Floratos (1998). "Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm." Bioinformatics **14**(1): 55-67.

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

Shmulevich, I., E. R. Dougherty, et al. (2002). "Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks." Bioinformatics **18**(2): 261-74.

Shoemaker, B. A., A. R. Panchenko, et al. (2006). "Finding biologically relevant protein domain interactions: conserved binding mode analysis." Protein Sci **15**(2): 352-61.

Singhal, M. and H. Resat (2007). "A domain-based approach to predict protein-protein interactions." BMC Bioinformatics **8**: 199.

Sprinzak, E. and H. Margalit (2001). "Correlated sequence-signatures as markers of protein-protein interaction." J Mol Biol **311**(4): 681-92.

Stein, A., R. B. Russell, et al. (2005). "3did: interacting protein domains of known three-dimensional structure." Nucleic Acids Res **33**(Database issue): D413-7.

Suthram, S., T. Shlomi, et al. (2006). "A direct comparison of protein interaction confidence assignment schemes." BMC Bioinformatics **7**: 360.

Toro, I., S. Thore, et al. (2001). "RNA binding in an Sm core domain: X-ray structure and functional analysis of an archaeal Sm protein complex." Embo J **20**(9): 2293-303.

von Mering, C., L. J. Jensen, et al. (2007). "STRING 7 - Recent developments in the integration and prediction of protein interactions." Nucleic Acids Research **35**(D358-D362).

Wang, H., E. Segal, et al. (2007). "InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale." Genome Biol **8**(9): R192.

Wang, H., E. Segal, et al. (2007). "InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale." Genome Biol **8**(9): R192.

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

Xie, L. and P. E. Bourne (2005). "Functional coverage of the human genome by existing structures, structural genomics targets, and homology models." PLoS Comput Biol **1**(3): e31.

Zhang, L. V., S. L. Wong, et al. (2004). "Predicting co-complexed protein pairs using genomic and proteomic data integration." BMC Bioinformatics **5**: 38.