Bioinformatics Advance Access published July 17, 2008

Message-Passing Algorithms for the Prediction of Protein Domain Interactions from Protein-Protein Interaction Data

Mudassar Iqbal [1,*], Alex A. Freitas [1], Colin G. Johnson [1] and Massimo Vergassola [2]

[1] Computing Laboratory and Centre for BioMedical Informatics, University of Kent, Canterbury, U.K.
[2] Institut Pasteur, Unit In Silico Genetics; CNRS, URA2171, F-75015, Paris, France.

Associate Editor: Dr. Trey Ideker

ABSTRACT

Motivation: Cellular processes often hinge upon specific interactions among proteins, and knowledge of these processes at a system level constitutes a major goal of proteomics. In particular, a greater understanding of protein-protein interactions can be gained via a more detailed investigation of the protein domain interactions that mediate the interactions of proteins. Existing high throughput experimental techniques assay protein-protein interactions, yet they do not provide any direct information on the interactions among domains. Inferences concerning the latter can be made by analysis of the domain composition of a set of proteins and their interaction map. This inference problem is non-trivial, however, due to the high level of noise generally present in experimental data concerning protein-protein interactions. This noise leads to contradictions, i.e. the impossibility of having a pattern of domain interactions compatible with the protein-protein interaction map.

Results:

We formulate the problem prediction protein domain interactions in a form that lends itself to the application of belief propagation, a powerful algorithm for such inference problems, which is based on message-passing. The input to our algorithm is an interaction map among a set of proteins, and a set of domain assignments to the relevant proteins. The output is a list of probabilities of interaction between each pair of domains. Our method is able to effectively cope with errors in the protein-protein interaction dataset and systematically resolve contradictions. We applied the method to a dataset concerning the budding yeast *Saccharomyces cerevisiae* and tested the quality of our predictions by cross-validation on this dataset, by comparison with existing computational predictions, and finally with experimentally available domain interactions. Results compare favourably to those by existing algorithms.

Availability: A *C* language implementation of the algorithm is available upon request.

Contact: mi26@kent.ac.uk

1 INTRODUCTION

Protein complexes and interactions are major players in cellular life (Alberts, 1998; Eisenberg *et al*., 2000). High-throughput experimental methods, such as yeast two-hybrid (Uetz *et al*., 2000; Ito *et al*., 2001; Rual *et al.*, 2005) and mass spectroscopy methods (Gavin *et al*., 2002; Ho *et al*., 2002; Gavin *et al*., 2006; Krogan *et al*., 2003), assay those interactions and the structure of complexes. Information provided by these different techniques currently appears to be largely complementary, as witnessed by the scanty overlap between respective interaction maps (von Mering *et al.*, 2005). The weak overlap and the relatively high level of noise generally present in the data call for extensive post-processing of the experimental interaction data using computational methods, which constitute an important and active area of research.

A major goal of computational approaches is to predict yet unknown protein-protein interactions on the basis of currently available information (Shoemaker *et al*., 2007a,b; Valencia *et al*., 2002). A first approach to the problem employs one or more genomic features related to the protein pairs as predictor attributes. For example, Bock *et al*. (2001; 2003) developed a machine learning system (Support Vector Machine) trained to recognise potential interactions based on the primary structure and the associated physico-chemical properties of the proteins. Another well-known method is the so-called Rosetta Stone Method (Marcotte *et al*., 1999), which exploits the observation that some pairs of interacting proteins have homologs in other organisms fused into a single protein chain. Many methods use a single type of proxy to predict protein interactions, e.g, methods based on the similarity in phylogenetic profiles (Galperin and Koonin, 2000), gene fusion methods (Marcotte *et al*., 1999; Enright *et al*., 1999), co-evolution of interacting partners (Goh *et al*., 2000, 2002). Other methods integrate different genomic features using a variety of machine learning methods, see e.g. Yamanishi *et al*. (2004); Jansen *et al*. (2003); Rhodes *et al*. (2005).

Information highly relevant to the prediction of protein-protein interactions comes from their domain structures. This is quite sensible, both evolutionarily and structurally, as domains are often evolutionarily conserved sequence units and they constitute the building blocks of protein structures, largely accounting for the reciprocal interactions among the proteins to which they belong. Namely, a pair of proteins is thought to physically interact if at least one among their constituent domain pairs interacts. A vast majority of proteins in well-studied organisms like *S. cerevisiae* are assigned one or more domains and these data can be combined with experimentally determined protein interaction datasets.

*to whom correspondence should be addressed

A few methods have already been developed to use these combinations of data in order to predict domain interactions (Riley *et al*., 2005; Deng *et al*., 2002; Lee *et al*., 2006; Li *et al*., 2006; Sprinzak *et al*., 2001). The strategy common to all these methods is to find potential domain interactions from existing protein-protein interaction datasets and then exploit that information to predict unknown protein-protein interactions. In other words, the idea is to infer domain-domain interactions from protein-protein interactions and then use these inferred domain interactions to predict new interactions between proteins, given their domain structure. For example, Sprinzak *et al*. (2001) developed an association method which finds correlated sequence signatures (domains) occurring together more often than by chance. They used a log-odds measure to quantify the frequencies of occurrence of domains in interacting proteins. Another method developed by Deng *et al*. (2002) uses the Maximum Likelihood method to estimate domain-domain interaction probabilities consistent with protein interaction data in which they occur, and also takes into account potential errors in the measurement of protein-protein interactions. Lee *et al*. (2006) estimate domain interaction probabilities in a very similar way as Deng *et al*. (2002), but they consider more protein interaction data from different organisms and also integrate other genomic features related to domains using a Bayesian approach. The Domain Pair Exclusion Analysis (DPEA) method (Riley *et al*., 2005) extends the Maximum Likelihood formulation used by Deng *et al*. (2002) and also includes protein interaction data from multiple organisms.

Our aim here is to show that the problem of predicting domain-domain interactions from protein-protein interaction data can be recast in a form that lends to the application of Belief Propagation (BP), a very powerful and widely used inference method (see MacKay (2000); Pearl (1998)). BP belongs to the class of so-called message-passing algorithms as they share the common feature of sending messages among neighbouring nodes in the graphical model of the system, until convergence is reached (Mezard, 2007). Convergence and exact inferences are rigorously guaranteed when the underlying graphical model is loop-free. In the presence of loops, convergence is not guaranteed; nonetheless it was first observed (in the context of decoding) that convergence can still hold (Gallager, 1963), and similar observations have been later made in a number of other applications. A rationalization of these observations was recently obtained in Yedidia *et al*. (2005), showing that BP solutions, even in the presence of loops, extremize the so-called Bethe free energy. Furthermore, Chertkov and Chernyak (2006) showed that the solutions obtained by Belief Propagation in the presence of loops contain enough information as to allow *a priori* the calculation of the exact result. Belief Propagation and message-passing algorithms have proved their relevance in a wide range of inference problems (Mezard *et al*, 2002; Yedidia *et al*., 2005). A recent biological application is the clustering method developed by Frey and Dueck (2007).

The paper is organized as to present first the Methods, which contain the specific formulation of the problem together with the algorithm and its derivation. We shall then discuss applications to protein-protein interactions for the budding yeast *S. cerevisiae*, followed by comparisons with existing methods and conclusions.

2 METHODS

2.1 Belief Propagation Algorithm for Prediction of Protein Domain Interactions

We consider a set of P proteins containing a number of domains (generally different for each protein) from a list of D possible types. I protein pairs are known to interact and constitute the positive dataset but we have no information (worst possible case) as to which domains are driving the interactions. N protein pairs are known not to interact. Our goal is to infer the interaction profiles among the domains, i.e. tell for a pair of domains whether or not it interacts. The inference is based on the fact that two proteins P_1 and P_2 interact if at least one of their domain pairs (one domain belonging to P_1, the other to P_2) interact and are non-interacting otherwise.

Let us define σ_{ij}, a binary variable equal to unity if the two domains i and j interact and zero otherwise. The indices i and j run over all possible D domains and links are undirected, i.e. we have $D(D+1)/2$ independent σ's. Any *a priori* information on domain interactions can be exploited as a prior on the value of the σ_{ij}. In its absence (worst possible case), we shall suppose that all Boolean variables σ's have the same *a priori* probability β to be equal to unity. The complementary probability for the σ's to vanish is $1-\beta$ and a compact expression for the two probabilities reads $1-\beta+\sigma(2\beta-1)$.

The likelihood (partition function; Z) for our system is defined to be the sum over all states of the unknown variables (σ's) compatible with the interaction map that we are handed as input:

$$Z = \sum_{\{\sigma_{ij}\}} \prod_{(ij)} (1-\beta+\sigma_{ij}(2\beta-1)) \times \prod_{p=1}^{I} \theta\left(\sum_{c_p} \sigma_{c_p}\right) \prod_{q=1}^{N} \left(1-\theta\left(\sum_{c_q} \sigma_{c_q}\right)\right) . \quad (1)$$

Here, the indices p and q run over all pairs of proteins in the positive and negative dataset, respectively, while the indices c_p and c_q run over all the pairs of domains for each one of those protein pairs. In other words, if we have two proteins P_1 and P_2 among the set I which interact, the index c_p will run over all possible domain pairs composed of one domain belonging to P_1 and the other to P_2. The Heaviside θ-functions (defined as vanishing if the argument of the function is zero and unity if the argument is positive) ensure the constraints stemming from the protein-protein interaction map. Indeed, if two proteins interact, at least one of their domain pairs should interact and the argument of the corresponding θ function should be positive. Conversely, if two proteins belong to the non-interacting dataset, all domain pairs should be non-interacting and the argument of their θ functions should vanish.

Since experiments generally contain some noise, we should take into account the possibility that information about protein-protein interactions that we are handed is not correct. As an extreme case, some errors might even lead to contradictions and to the impossibility of having any solution for the observed interaction data, as shown in Fig. 1. A convenient way to deal with this problem is to "soften" the θ functions in the function nodes as

$$\theta_S(\sigma) = \begin{cases} \epsilon & \text{if } \sigma = 0, \\ 1-\epsilon & \text{if } \sigma > 0. \end{cases} \quad (2)$$

The parameter ϵ (which runs from zero to unity) represents the degree of reliability of the interaction datasets available for the inference. Full trust corresponds to $\epsilon = 0$, while the most noisy case corresponds to $\epsilon = 1/2$, when the interaction datum is irrelevant ($\theta_S \equiv 1/2$ irrespective of its argument). Values larger than $1/2$ correspond to the (rather unlikely) situation when input data tend to contradict reality. In particular, $\epsilon = 1$ corresponds to the case when the data are systematically reversed.

To simplify notation and conform to those commonly employed in graphical models, we recast (1) in the general and compact form:

$$Z = \sum_{\{\sigma\}} \left[\prod_k \psi_k(\sigma_k) \prod_\alpha f_\alpha\left(\{\sigma\}_\alpha\right) \right] , \quad (3)$$

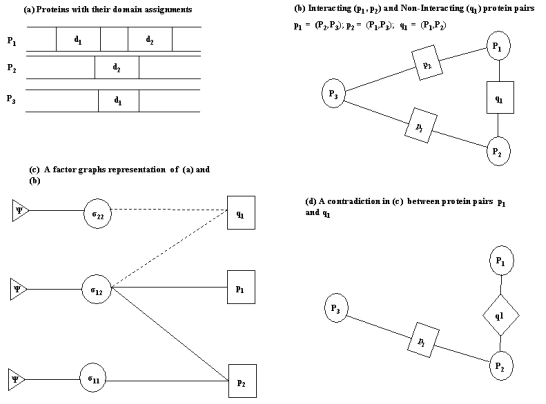

Fig. 1. A graphical illustration of a simple instance of protein and domain pair interactions. (a) shows the list of proteins together with their corresponding domains. (b) gives the list of the interactions between protein pairs and their graphical representation. In (c) we display the factor graph corresponding to the interactions in (b) where circles represent the domain pairs (variable nodes) while squares and diamonds represent the interacting and non-interacting protein pairs (function nodes) respectively. The ψ's on the left represent the priors on the variables, chosen here to be identical for all of the variables and controlled by the parameter β. Finally, (d) presents a simple example of pattern of interactions leading to a contradiction.

where the index k runs over all possible domain pairs, the index α runs over all proteins pairs present in the interaction datasets (both positive and negative), ψ_k is the local evidence (polarization) for the variable nodes σ_k and f_α denotes the so-called function nodes. The ensemble of variables $\{\sigma\}_\alpha$ denote the set of all the variables σ_{ij} for the pair of proteins α. A factor graph representation (with protein and domain pairs as function and variable nodes respectively) of the model is illustrated in Fig. 1. In our case, the local evidence is uniform, i.e. does not depend on the variable node:

$$\psi_k(\sigma) = \psi(\sigma) = 1 - \beta + \sigma\,(2\beta - 1)\ . \tag{4}$$

Function nodes take two different forms depending on whether the protein pair belongs to the dataset of interacting or non-interacting pairs:

$$f_\alpha(\{\sigma\}_\alpha) = \begin{cases} \theta\left(\sum_{\sigma\in\{\sigma\}_\alpha}\sigma\right) & \text{interacting} \\ 1-\theta\left(\sum_{\sigma\in\{\sigma\}_\alpha}\sigma\right) & \text{non interacting}\,, \end{cases} \tag{5}$$

Having recast the problem in the general form of graphical models, Belief Propagation equations associated to (3) follow from textbook derivations (see, e.g. MacKay (2000) page 336):

$$M_{\alpha\to k}(\sigma_k) \quad\propto\quad \sum_{\{\sigma\}_\alpha\neq\sigma_k} f_\alpha\left(\{\sigma\}_\alpha\right) \prod_{k'\in\{k\}_\alpha\neq k} M_{k'\to\alpha}\left(\sigma_{k'}\right) \tag{6}$$

$$M_{k\to\alpha}(\sigma_k) \quad\propto\quad \psi_k(\sigma_k) \prod_{\alpha'\in\{\alpha\}_k\neq\alpha} M_{\alpha'\to k}(\sigma_k)\,. \tag{7}$$

Messages $M_{\alpha\to k}$ are sent from function to variable nodes, while messages $M_{k\to\alpha}$ are sent in the opposite directions. The proportionality sign is meant to stress that, in the presence of loops, it is more appropriate to work with normalized equations to increase stability and facilitate convergence. Messages are exchanged among nodes until convergence is reached. The partition function Z is estimated as described in the next Section.

2.2 Bethe Free Energy and Belief Propagation

As stated earlier, beliefs calculated by (6) and (7) are exact when the underlying graph has no loops. Since message-update rules do not directly depend on the topology of the underlying graph, the iterative scheme (6) and (7) might be run on graphs with loops and the quality of the results might be assayed empirically (Frey *et al*., 1997). In this spirit, Belief Propagation (BP) has been successfully applied to many practical problems with loops (Gallager, 1963; Frey *et al*., 1997; Yedidia *et al*., 2002). A reason for these successful applications on graphs with loops has been put forward in (Yedidia *et al*., 2002, 2005) showing that BP solutions are extrema of an approximation to the original partition function Z of the model. The approximation to $F \equiv -logZ$ is known as Bethe free energy and the one associated to (3) takes the form :

$$F(\{b_\alpha\},\{b_k\}) = -\sum_\alpha \sum_{\{\sigma\}_\alpha} b_\alpha \ln f_\alpha - \sum_k \sum_{\sigma_k} b_k \ln \psi_k + \sum_\alpha \sum_{\{\sigma\}_\alpha} b_\alpha \ln b_\alpha - \sum_k \sum_{\sigma_k} (q_k - 1)\, b_k \ln b_k \tag{8}$$

Here, q_k denotes the number of function nodes which have the k-th variable as input. The b's are beliefs for the probability distributions of individual and node variables, computed from the messages as follows :

$$b_\alpha(\{\sigma\}_\alpha) \propto f_\alpha(\{\sigma\}_\alpha) \prod_{k\in\{\sigma\}_\alpha} M_{k\to\alpha}(\sigma_k)\,; \tag{9}$$

$$b_k(\sigma_k) \propto \psi_k(\sigma_k) \prod_{\alpha\in\{\alpha\}_k} M_{\alpha\to k}(\sigma_k)\,. \tag{10}$$

The proportionality signs indicate that beliefs should be normalized (in agreement with the fact that they represent estimates of marginal probability distributions). BP estimates are consistent under marginalization, i.e. $\sum_{\{\sigma\}_\alpha\neq\sigma_k} b_\alpha(\{\sigma\}_\alpha) = b_k(\sigma_k)$. This follows from (6) and (7).

To demonstrate that solutions of our BP equations indeed extremize the free energy (8) one can proceed as in (Yedidia *et al*., 2005), introducing Lagrange multipliers to enforce normalization of beliefs and consistency under marginalization. The condition that derivatives with respect to b_α and b_k vanish is thus shown to coincide with equations (6) and (7). Details of the derivation can be found in (Yedidia *et al*., 2005).

The Bethe free energy is extremely useful for our purposes as we have two unknown parameters in our model (the prior parameter β and the noise parameter ϵ). We shall then run BP equations to convergence and choose the values of the parameters β and ϵ that correspond to the minimum of the Bethe free energy (maximum of the partition function).

Numerical implementation

Starting with initial values of unity for all of the messages, we iterate the BP equations (6 and 7) for given values of β and ϵ in equation (2). BP iterations are stopped after the changes in all the messages are below a threshold, set equal to 10^{-2}. Results do not change if the threshold is set smaller. In order to reach convergence, a standard trick employed to reduce oscillations is to use a damping factor λ so that each message is updated as λ times its value from previous iteration plus $1-\lambda$ times its current value. For example, the message $M^{(n+1)}_{\alpha\to k}(\sigma_k)$ is updated as $(1-\lambda)\sum_{\{\sigma\}_\alpha\neq\sigma_k} f_\alpha\left(\{\sigma\}_\alpha\right)\prod_{k'\in\{k\}_\alpha\neq k} M^{(n)}_{k'\to\alpha}\left(\sigma_{k'}\right) + \lambda M^{(n)}_{\alpha\to k}(\sigma_k)$ (compare to (6)). After some numerical experiments, we chose a damping factor $\lambda = 0.5$ in all the runs of the algorithm.

When iterations are run at very small ϵ, errors in experimental data makes that for some domain pairs no solution is found, i.e., beliefs are all zero (or extremely small). On the other hand, these configurations are not very interesting as they have a huge Bethe free energy. We therefore decided to circumvent this numerical problem by working with a small, yet nonzero, predefined precision of 10^{-10}.

Prediction of protein-protein interactions

Predictions of domain-domain interactions can be exploited to predict protein-protein interactions. As an example of this approach, we performed a cross-validation analysis on avalaible protein-protein interactions. Knowing the composition in domains of a protein pair α, the probability Pr_α of their

interaction is estimated from beliefs $b(\sigma_{ij})$ of interaction between domains i and j as

$$Pr_\alpha = 1 - \prod_{\sigma_{ij} \in \{\sigma\}_\alpha} (1 - b(\sigma_{ij})) \ , \quad (11)$$

where the ensemble of variables $\{\sigma\}_\alpha$ denotes the set of all domain pair variables σ_{ij} for the pair of proteins α.

3 MATERIALS

Domain assignments

We obtained domain assignments for *S. cerevisiae* genome from the SUPERFAMILY database (Gough *et al*., 2001; Madera *et al*., 2004) (web site www.supfam.org). This database is a library of HMMs modelling all proteins of known structure. These models are used to annotate the sequence of over 50 genomes. For *S. cerevisiae*, there exist 3346 sequences with at least one domain assignment, which is about 50% of total sequences. In total, 4681 domains are assigned and there are 685 superfamily domains with at least one assignment.

Positive Interaction Dataset

We obtained the *S. cerevisiae* interaction dataset from DIP (Database of Interacting Proteins) (Salwinski *et al*., 2004; Xenarios *et al*., 2002). We obtained nearly 5000 high confidence positive interactions from CORE, which is a subset of the total number of reported protein interactions in DIP. Furthermore, since there are some proteins which do not have any significant domain assignment, we only kept those proteins which have at least one domain assignment in the superfamily database. This process reduces our interactions to 3070 pairs, which constitute our dataset of positive interactions.

Negative Interaction Datasets

Information on negative protein-protein interactions, i.e. pairs of proteins which are not interacting in the experimental conditions of assay, was hard to find. Reasons for this, and remarks upon the importance of negative datasets, are presented in the Conclusions. In this section we describe the motivation for and construction of two negative datasets. Both of them are built upon data concerning the localization of proteins in cellular compartments.

- A first dataset is built by sampling from all pairs of proteins that are localized in different compartments of the cell. We will refer to this dataset as *NonCoLoc_Neg*, i.e., Non-CoLocalized Negative dataset. This type of dataset has been used by many researchers in this field (e.g. Jansen *et al*. (2003) and Rhodes *et al*. (2005)). There are many hundreds of thousand of protein pairs which are not co-localized, a huge amount compared with the number of positives. The standard procedure, which we followed as well, is to randomly sample from this pool of possible negatives. We also imposed the constraint that proteins ought to have at least one domain assignment in the superfamily database. We thus ended up sampling a total of 3070 negative interactions between pairs of proteins, as many as the positive ones.
- The biological motivation for the previous choice of the negative dataset, even though employed in the literature, is not quite clear. Indeed, potentially harmful interactions between two proteins located in different compartments of the cell are already largely prevented by their different localization. The two proteins can therefore afford to have domains that would be interacting if they were brought in contact. This motivated us to compare results obtained using the previous *NonCoLoc_Neg* with those using *CoLoc_Neg*, i.e., Co-Localized Negative dataset. To generate the latter, we collected localization data from MIPS (Mewes *et al*., 2002), built a sample of pairs of proteins having the same cellular localization and classified them as negatives if they are not reported in DIP-CORE set of positive interactions. We further kept only those pairs which have at least one domain assignment in SUPERFAMILY database and ended up with a subset of 3740 pairs, constituting the ensemble of negative interactions for the dataset *CoLoc_Neg*.

4 RESULTS AND DISCUSSION

Figs. 2 and 3 show the Bethe free energy for the experimental datasets of non-co-localized (*NonCoLoc_Neg*) and co-localized (*CoLoc_Neg*) proteins, constructed as described in the Methods section. Bethe free energies, as defined in the Methods section, are shown as a function of the noise parameter ϵ for different values of the prior parameter β. In both cases the minimum of the Bethe free energy is reached at $\beta = 0.2$ and at comparable small values of ϵ. However, the value of the minimum of the Bethe free energy for non-colocalized proteins *NonCoLoc_Neg*, i.e., the dataset where negative interactions are obtained from proteins appearing in different localization classes, is sizeably higher than for the other dataset *CoLoc_Neg*. The difference is quantitatively substantial since one should remember that the partition function Z and the free energy F are related as $Z = e^{-F}$. Furthermore, *CoLoc_Neg* contains more negative data, i.e. corresponding value of Z should a priori be smaller and the Bethe free energy should be higher (for a fixed quality of the dataset). The fact that *CoLoc_Neg* has a lower minimal free energy than *NonCoLoc_Neg* is therefore highly significant and signals that the former is a better sample of negative interactions as compared to the latter. Biological consequences of this result are postponed to the Conclusions. Note that these results stress the importance of having a good gold standard of negative interactions in order to have a robust inference of domain interactions.

Note that contradictions in the experimental data, which were mentioned in the Methods, are indeed present and relevant. At $\epsilon = 0$, i.e. when interaction data are taken at face value without any possible modification, the number of contradictory interactions in the positive and negative (Colocalized) datasets are 1025 and 1020, respectively (over a total of 3070 and 3740). At the minimum of the Bethe free energy ($\beta = 0.2$ and $\epsilon = 0.04$), contradictions are sizeably reduced as the number of positive and negative interactions that remain unchanged is 2667/3070 and 3420/3740, respectively.

4.1 Cross-validation: Predicting Protein Interactions

We performed a 10-fold cross-validation analysis, predicting domain interactions from training data and using them to predict protein-protein interactions on test data. We performed these cross validation analyses for *CoLoc_Neg* since this data was shown to be more effective in minimizing the Bethe free energy. For each computational experiment, we divided the data (for both positive and negative classes separately) randomly into ten equal folds. Each time we used nine out of ten folds as training and the remaining one fold as a test. This process was repeated ten times, each time using a different fold as the test set. Protein pairs in test data which do not contain any domain pair from the training data were removed. For each of the 10 iterations of the cross-validation procedure, we inferred the normalized beliefs of domain pairs from the training set using the belief propagation procedure, as described above. We then did the experiments corresponding to a range of values of ϵ and β and predicted protein-protein interactions for the test fold as described in the Methods section.

We calculated the prediction accuracies for each value of ϵ and β comparing the prediction to the experimental assignment. Note

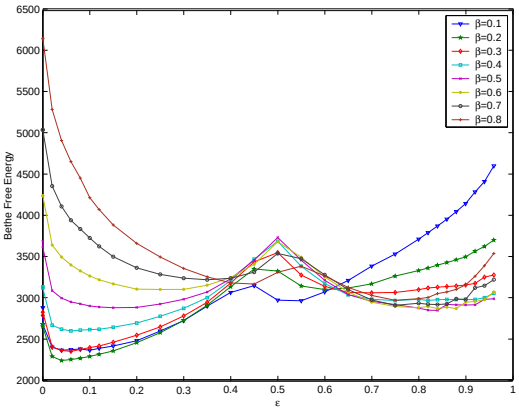

Fig. 2. Bethe free energy as a function of the parameter ϵ (quantifying the amount of noise and incorrect data in the experimental dataset), for different values of the parameter β, controlling the prior on the expected number of positive interactions among protein domains. Curves refer to the dataset (*NonCoLoc_Neg*) where negative protein-protein interactions are constructed from pairs of proteins having different cellular localizations.

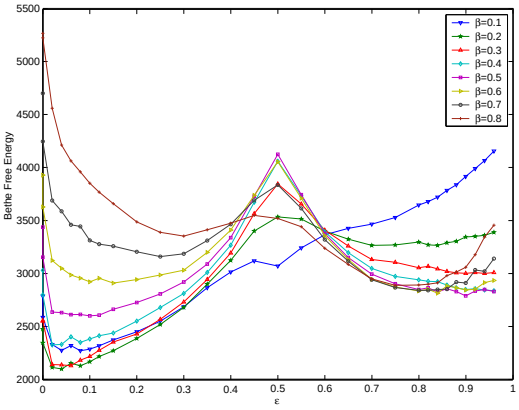

Fig. 3. The same curves as in Figure 2, for the the dataset (*CoLoc_Neg*) where negative protein-protein interactions are constructed from protein pairs not appearing in the list of interacting proteins *and* having the same cellular localizations.

that the presence of noise in the experimental data makes that we should not expect the accuracy to be optimal at the same values as the minimum of the Bethe free energy. Some of the data are indeed likely to be incorrect and, since our method is built so as to reverse them, we expect that the values of ϵ will be comparable yet not quite identical. Indeed, Fig. 4 shows the ratio of true positive rate over the false positive rate for the test set predictions, for different values of ϵ and β. True Positive Rate (TPR) or Sensitivity is defined as the number of true positives over total number of positives and False Positive Rate (FPR) is defined as the number of false positives over total number of negatives in the data. We can see that this ratio is overall maximum for predictions corresponding to $\beta = 0.2$, i.e. the same value which gives the minimum free energy in all folds as well as the full data as shown in fig. 3. On the other hand, the previous ratio peaks at a value of ϵ which is comparable, yet larger than the one giving the minimum of the Bethe free energy.

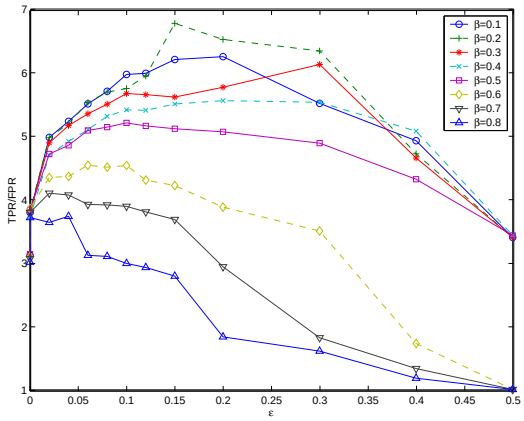

Fig. 4. Average values of true positive rate over false positive rate for different values of ϵ and β.

The average prediction accuracy values over ten folds corresponding to the parameter (ϵ and β) values which minimized the Bethe free energy is 82% and the corresponding values of sensitivity and specificity are 79% and 85%.

4.2 Comparison With Other Domain Interaction Prediction Methods

To compare the results obtained here with those by methods previously appeared in the literature, we found it very useful the database DOMINE constructed by Raghavachari *et al*. (2007), compiling a set of 20513 predicted domain-domain interactions from experimental sources as well as from existing computational methods. Among the experimental sources, they used the *iPfam* database, which contains domain interactions observed in PDB entries (Berman *et al*., 2000), and the *3did* database, which contains domain interactions among the proteins with known high resolution structure (Stein *et al*., 2005). Other domain interactions included in the DOMINE database are from 8 computational methods using different approaches to uncover underlying domain interactions in the experimental data of protein-protein interactions. Some of the methods also use other genomic features along with the assignment of domains to proteins. For example, Lee *et al*. (2006) use domain-domain interactions predicted using Maximum Likelihood method from protein-protein interaction data in multiple organisms and use a Bayesian data integration scheme to combine these data with gene ontology and domain fusion information.

Since all computational methods reported in DOMINE use Pfam-A (Finn *et al*., 2006) domain definitions, in order to make a comparison we created a dataset of positive and negative interactions as described in the Materials section while using domain assignments according to Pfam-A definitions. We used 2642 positive and 3123 negative protein interactions in this experiment and run our Belief Propagation algorithm to extract the results of domain interactions corresponding to the minimum value of the Bethe free energy, as described in the Methods.

We compared these results to those by other computational methods in DOMINE and also to the experimental gold standard set of domain interactions, which is the union of interactions from

Table 1. Comparison of Percentage overlap of BP with experimental gold standard interactions with respect to other computational methods in the DOMINE database

ME	RDCP	P-value	Fusion	NetOpt	RDFF	PP	BP
52.9	12.9	9.6	11.9	10.9	4.8	1.1	14.6

iPfam and *3did* databases. It is important to mention here that comparisons in DOMINE are made only for positive domain interactions while in our method we also predict non-interactions as well. It is also worth noting that the various methods are not predicting the same set of interactions. For each of the given 20513 domain pairs in the DOMINE database, our method has three kind of predictions, i.e, the pair is predicted either positive or negative or we do not have any predictions because that particular pair was not in our dataset (as it is the case with all other methods). For those domain pairs where we have a prediction and there is a prediction in the gold standard (*ipfam+ 3did*) as well, we find 133 matching predictions out of 198 total cases.

As for other computational methods, we can just compare the overlap of positive predictions with the available reference gold standard domain interactions. Table 1 shows the percentage overlap of positive domain interactions predicted by different computational methods (including BP) against the gold standard data of experimental domain interactions. Belief Propagation (BP) results have over 14.5% overlap with the gold standard data of positive domain interactions which is second to only one method out of 8 (in fact 7 in total since in DOMINE database, two methods are combined into one due to their similarity). In fact, the method (Lee *et al*., 2006) that has maximum overlap is using protein interaction maps from multiple species and then integrate the information gained from them about domain interactions with other genomic features as well as *ipfam* in the training of the method itself.BP inference about predicting domain interactions from protein interaction data is, therefore highly competitive in this comparative setting.

We extended the comparison proceedings as in (Raghavachari *et al*., 2007), i.e. calculating the percentage overlap between predictions of our method (BP) with different computational methods reported in DOMINE, as shown in Table 2. This overlap is quite variable with respect to individual methods, but $\sim 98\%$ of positive interactions predicted by BP are also predicted by at least one other method.

Table 2. Percentage overlap of BP predictions with other computational methods

ME	RDCP	P-value	Fusion	NetOpt	RDFF	PP
16.3	17.7	10.0	6.3	29.0	72.7	1.1

Finally, the DOMINE database features a list of 55 high-confidence domain interactions which are predicted by at least 4 different computational methods. We checked them against our predictions, and found about 83% correctly predicted by our method, which again compares favourably with other methods (Raghavachari *et al*., 2007).

5 CONCLUSION

We have addressed the problem of inferring domain interactions from large-scale protein-protein interaction data. The problem was recast as a factor graph model lending to the use of Belief Propagation (BP). This powerful message-passing inference method was employed to estimate the probability of interaction between domains. The Bethe free energy of the corresponding BP solutions provides a systematic way to quantify the amount of noise in the experimental dataset and pinpoint those data which are the most problematic, e.g. because they lead to contradictions in the pattern of domain-domain interactions. This specific feature of our method has a double interest: first, it allows extracting reliable predictions from noisy datasets and, second, it can be used as a guide for further experimental verifications to correct false data and increase the quality of interaction datasets.

A major reason of interest in domain-domain interactions is that they can be exploited to improve the quality of predictions for protein-protein interactions. As an example, we successfully used the domain interactions predicted by our BP method on a test dataset using a standard cross-validation procedure. Furthermore, the domain interaction predictions of our method were compared against the set of experimentally available gold standard set of domain interactions and also with other known computational methods. Comparative results indicate that Belief Propagation is a very effective method to attack the domain-interaction inference problem.

An interesting biological remark that emerged from our analysis is related to the importance and the nature of negative protein-protein interactions. What we have shown here is that protein pairs localized in the same cellular compartments *and* not appearing in the interaction datasets seem to provide for a better sample of negative interactions than protein pairs in different compartments of the cell. The latter type of dataset was previously used in the literature. Preventing noxious, e.g. for their potential toxicity, interactions is quite a sensible issue from a biological point of view and examples of potentially toxic products are quite common in metabolic pathways. As a matter of fact, the necessity to run chemical reactions in specific conditions and keep some of the products physically separated to avoid their cross-reactions constitute a major drive towards the compartmentalization of the cell. Our results point at the importance of similar prevention effects for protein-protein interactions as well. Finally, data on negative interactions, i.e. pairs of proteins not interacting in physiological conditions, are unfortunately hardly found in the literature. One of the reasons has probably to do with the negative character of the datum. The other reason has to do with experiments themselves, as it is particularly difficult to check whether an observed absence of interaction is real or due to a problem in the experimental procedure. The effort is quite worthwhile, though, as our results show that the quality of domain interaction inferences can be strongly improved by a proper dataset of negative interactions. We hope that the results shown here will stimulate future experiments in these directions.

ACKNOWLEDGEMENTS

Mudassar Iqbal, Alex Freitas and Colin Johnson acknowledge the financial support from EPSRC under grant GR/T11265/01. Mudassar Iqbal also acknowledges further financial support from the Computing Laboratory, University of Kent.

REFERENCES

Alberts B. (1998) The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists, *Cell* **92**, 291-294.

Berman H. M. *et al*. (2000) The protein data bank, *Nucleic Acids Res.* **28**,235242.

Bock J.R., Gough D.A. (2001) Predicting protein-protein interactions from primary structure, *Bioinformatics***17(5)**,455-460.

Bock J.R., Gough D.A. (2003) Whole proteome interaction mining, *Bioinformatics* **19(1)**,125-135.

Chertkov M., Chernyak V.Y., (2006) Loop series for discrete statistical models on graphs *Journal of Statistical Mechanics-Theory and Experiment*, **P06009**.

Deng M. *et al*. (2002) Inferring Domain-Domain Interactions From Protein-Protein Interactions,*Genome Res.*, **12**, 1540-1548.

Eisenberg D. *et al*. (2000) Protein function in the post-genomic era, *Nature*, **405**, 823-826.

Enright AJ *et al*. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86 90.

Finn R. D. *et al*. (2006) Pfam: clans, web tools and services, *Nucleic Acids Res.* **34**, D247D251.

Frey B. J. and Dueck D. (2007) Clustering by Passing Messages Between Data Points, *Science*, **315**, 972.

Frey B. J. and Mackay D. J. C. (1997) A revolution: Belief propagation in graphs with cycles, In M. Jordan *et al* (Eds.), *Adv. in Neural Information Processing Systems*, **10**, MIT Press.

Gallager R. G. (1963) Low Density Parity Check Codes, (MIT press, Cambridge, MA).

Galperin M.Y. Koonin E.V. (2000), Whos your neighbor? New computational approaches for functional genomics. *Nat Biotechnol* **18**, 609 613.

Gavin A.C. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141147.

Gavin A.C. *et al*. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631636.

Goh C. *et al*. (2000) Co-evolution of Proteins with their Interaction Partners. *J. Mol. Biol.*, **299**, 283-293.

Goh C. Cohen F.E. *et al*. (2002) Co-evolutionary Analysis Reveals Insights into ProteinProtein Interactions *J. Mol. Biol.*, **324**, 177-192.

Gough J. *et al*. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure, *J. Mol. Biol.*, **313(4)**, 903-19.

Ho Y. *et al*. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180183.

Ito T. *et al*. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 45694574.

Jansen R. *et al*. (2003) A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data, *Science*,**302**, 449-453.

Krogan N.J. *et al*. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*,**440**, 637-643.

Lee H. *et al*. (2006) An Integrated Approach to the Prediction of Domain-Domain Interactions, *BMC Bioinformatics*.,**7:269**.

Li X. *et al*. (2006) Improving domain-based protein interaction prediction using biologically-significant negative dataset, *International Journal of Data Mining and Bioinformatics*, **1(2)**,138-149.

Li S. *et al*. (2004) A Map of the Interactome Network of the Metazoan C. elegans, *Science*, **303**, 540-543.

MacKay D. J. C. (2003) *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press.

Madera M. *et al*. (2004) The SUPERFAMILY database in 2004: additions and improvements, *Nucleic Acids Res.*, **32(1)**, D235-9.

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

Mewes H. W. *et al*. (2002) MIPS: a database for genomes and protein sequences, *Nucleic Acids Res.* **30(1)**, 31-4.

Mezard M. (2007) Computer Science – Where are the exemplars?, *Science*, **315**, 949-951.

Mezard M. *et al* (2002) Analytic and Algorithmic Solution of Random Satisfiability Problems, *Science*, **297**, 812.

Pearl J., (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers.

Raghavachari B. *et al*. (2007) DOMINE: a database of protein domain interactions, *Nucleic Acids Research*, 1 6.

Rhodes D. R. *et al*. (2005) Probabilistic model of the human protein-protein interaction network, *Nature Biotechnology* **23(8)**, 951-959.

Riley R. *et al*. (2005) Inferring Protein Domain Interactions From Databases of Interacting Proteins,*Genome Biology*,**6:R89**.

Rual JF. *et al*. (2005) Towards a proteome-scale map of the human protein-protein interaction network *Nature* **437**, 1173-1178.

Salwinski L. *et al*. (2004) The Database of Interacting Proteins: 2004 update, *Nucleic Acids Res.* **32**, Database Issue: D449-51.

Shoemaker B. A., Panchenko A. R. (2007)Deciphering ProteinProtein Interactions. Part-I: Experimental Techniques and Databases, *PLoS Computational Biology* **3(3):e42**.

Shoemaker. B. A., Panchenko A. R. (2007) Deciphering Protein-Protein Interactions. Part-II: Computational Methods to Predict Protein and Domain Interaction Partners, *PLoS Computational Biology* **3(4):e43**.

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

Stein A. *et al*. (2005) 3did: interacting protein domains of known three-dimensional structure, *Nucleic Acids Res.* **33**, D413D417.

Uetz P. *et al*. (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*, *Nature*, **403(1)**, 623-627.

Valencia A., Pazos F. (2002) Computational methods for the prediction of protein interactions, *Current Opinions in Structural Biology* **12**,368-373.

von Mering C.*et al*. (2002), Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417(6887)**, 399403.

Xenarios I. *et al*. (2002) DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions, *NAR* **30**, 303-305.

Yamanishi Y. *et al*. (2004) Protein network inference from multiple genomic data: a supervised approach, *Bioinformatics* **20** Suppl.1,i363-i370.

Yedidia J.S. *et al*. (2005) Constructing Free-Energy Approximations and Generalized Belief Propagation Algorithms, *IEEE Transactions on Information Theory*, ISSN; 0018-9448, **51(7)**, 2282-2312.

Yedidia J.S. *et al*. (2002) Understanding Belief Propagation and its Generalizations, *Technical Report*, TR-2001-22. Mitsubishi Electric Research Labratories, Cambridge, Massachusetts.