Prediction of protein-protein interactions and the interaction site from sequence information - an extensive study of the co-evolution model

Jason Davis and Golan Yona
Department of Computer Science
Cornell University, Ithaca, NY 14853

Abstract

This paper studies new variants of the co-evolution model for prediction of protein-protein interactions from sequence information. Given two query proteins, the method uses information extracted from a database search to generate a multiple alignment and compute the likelihood that the two proteins interact. The model uses four elements, starting from the correlated divergence between proteins from different species, through potentials of correlated mutations, charge and hydrogen bonds. The significance of these measures is estimated from large populations of interacting and non interacting protein pairs. A variant over an EM algorithm is used to identify the subset of the database proteins (the homologs of the query proteins) that are more likely to interact, and a modified correlated mutations model is employed to maximize the strength of the signals.

The algorithm not only tries to suggest if two proteins interact, but also attempts to detect the localized binding box interaction region, information that is hardly ever available. We tune and test our model over a large set of protein interactions we extract from BIND.

1 Introduction

Protein-protein interactions are at the core of numerous basic reactions that make up complex biological processes. The detection of these interactions can help to better understand the molecular machinery of the cell, and expose biological processes and pathways that have not been characterized so far. Recently, several studies explored large networks of known protein interactions in search for structures, complexes and regular patterns [Schwikowski et al. 2000, Ho et al. 2002].

Traditionally, protein-protein interactions have been deduced via conventional wet-lab experimental methods, such as the yeast two-hybrid (Y2H) system [Fields & Song 1989] and mass spectrometry [Ho et al. 2002]. While these methods can detect protein interactions on a genomic scale, they tend to be very expensive and time consuming. Moreover, their error rate is high, especially with the Y2H method, with many false negatives (due to the cellular localization of the proteins, post-translational modifications, misfolding of the recruited proteins or steric hindrances) and high percentage of false positives [Uetz et al. 2000].

In this view, there is a strong interest in tools that can reliably predict the existence of protein-protein interactions. This problem has received a considerable attention recently, and several methods, summarized next, are described in the literature.

1.1 Related work

The field of predicting protein protein interactions via computational methods is relatively new and undeveloped. These methods can be mainly classified into three different approaches: studies that use structural information, relational data mining inference, and co-evolution analysis.

1.1.1 Protein structure-based approaches

Methods using protein tertiary structure have had some success [Sobolev et al. 1999, Gallet et al. 2000], but are relatively slow. Protein docking methods use known tertiary structure to try to predict the most probable binding site between two structures. As this problem is NP hard, conventional methods rely on local search heuristics. Other methods use known hydrogen bond potentials, charge potentials, and interaction site energy minimization techniques. Most of these approaches, however, do not take into consideration possible protein deformations that can change the binding properties at the time of interaction. Methods that do allow such flexible docking structures are sometimes computationally infeasible [Teodoro et al. 2001]. Furthermore, structure-based approaches fail to distinguish between two proteins that have the bio-chemical potential of interaction, and two proteins that physiologically interact. Lastly, these methods are limited to proteins of known structures, which account for less than 5% of all known proteins and 30% of the known protein families.

Protein threading methods can extend predictions to sequences of unknown structures. Lu et. al. use a multi-threading approach in their "Multiprospector" algorithm [Lu et al. 2002]. The method threads sequences in a library of monomer structures that are known to participate also as part of dimmer structures. If two sequences have significant signal with respect to two chains that are part of the same complex, the sequences are re-aligned (considering the other sequence in the template structure), the energy between the interacting residues is computed, and if the zscore of the dimmer is significant compared to that of the monomers, the sequences are predicted to interact. One of the main advantages of the method is that beyond predicting the interaction itself, it also predicts the interaction site. However, it is limited to solved complexes of interacting proteins, of which only a few are known in the protein data bank.

1.1.2 Relational data mining inference

These approaches are among the most successful in the field of computational protein-protein interaction inference. Kini and Evans' earlier work showed a correlation between interaction binding sites and proline residues: these residues are 2.5 times more likely in these areas [Kini & Evans. 1995]. Marcotte et. al. [Marcotte et al. 1999] note the relation between protein interaction and gene locality through a technique termed "domain fusion prediction". Sprinzak and Margalit [Sprinzak & Margalit 2001] provide a framework for identifying protein protein interactions through the analysis of over represented sequence signatures. Other methods, such as "InterPreTS", use a database of known interacting three-dimensional structures to determine possible interaction sites between two proteins [Alloy & Russel 2003]. Enright et. al present a refinement over typical domain fusion prediction methods suggesting that interacting proteins can also be the result of a 'composite' protein evolving into 'component' proteins [Enright et al. 1999]. These methods have had some success, but mostly for interactions of similar nature, either with the same signatures, or when there is an evidence of gene fusion/decomposition events in sequence databases. Clearly, such evolutionary circumstances are not the case for all protein protein interactions.

1.1.3 Co-evolution analysis

Co-evolution approaches include phylogenetic tree topology comparison, gene preservation correlation, and correlated mutation approaches. Phylogenetic tree topology methods compare homologs of interacting proteins (i.e. protein families): if the two trees are highly related (termed **mirror trees**), then the proteins have co-evolved and possibly interact [Goh et al. 2000, Pazos & Valencia 2001]. Gene preservation approaches are very simple: if two proteins interact to perform a vital biological function, then both proteins will be passed on during speciation [Enright et al. 1999]. Other gene preservation approaches consider locality constraints among protein sub-domains: sub-domains will tend to have the same relative position in interacting proteins [Dandekar et al. 1998]. A more localized position specific approach is the correlated mutations method introduced by Pazos and Valencia [Pazos et al. 1997]. This approach searches for correlated, compensating mutations in specific positions between two candidate interacting proteins. Their results presented on structural subdomains seem fairly strong [Valencia & Pazos 2002]. They later extended this algorithm to detect interacting partners based on the ratio between intra-protein correlations and inter-protein correlations [Pazos & Valencia 2002].

1.2 Our approach

In this paper we test the co-evolution model for inferring protein-protein interactions. We introduce new variants and a framework for integrating multiple sources of information or interaction potentials. Our basic assumption is similar to the co-evolution principle that was originally introduced in [Gobel et al. 1994]. Specifically, two proteins that interact will tend to co-evolve in a coordinated manner resulting in a higher evolutionary correlation between their corresponding homologs. The intuition behind this premise is fairly simple; if one partner in a protein interaction pair mutates, then its counterpart will have to adapt in order to preserve the interaction. Therefore, given two query proteins and their homologs one can theoretically predict an interaction if there is an evidence that the groups co-evolve.

Correlated evolution approaches offer several advantages over previous methods. First, there are several algorithms that can approximate a correlated evolution score very well and also run in polynomial time. Second, the idea of correlated evolution is a priori and fits in with basic biological principles. Sequence signatures, represent an a posterior approach; consequently, they can only identify protein interactions similar to those found in a training data set. Finally, co-evolution can detect proteins that physiologically interact.

Any co-evolution method for protein-protein interaction prediction relies on the knowledge of phylogenetic trees. However, in practice, the exact evolutionary path of a specific protein is unknown, therefore, one must infer a protein's phylogeny via careful examination of the differences between protein homologs located in different organisms. An ideal solution to this problem would be to reconstruct phylogenetic trees for each interacting protein partner and its homologs (herein referred to as a protein family) and then to compare their similarities. Unfortunately, phylogenetic tree reconstruction is provably an NP-complete problem, and existing measures for assessing co-evolution (as the pearson correlation coefficient) attempt to avoid this problem by considering all pairs of protein homologs. However, they are effective only if the signal is very strong.

Our approach for assessing co-evolution and prediction of protein-protein interactions addresses these issues. The model is composed of several factors; the first two are the overall correlated divergence, and the position-specific correlated mutations signals under a reduced tree model (the **star model**). This model reduces the phylogenetic tree to the set of leaf nodes and their closely related species (in the form of stars). Using this model we revise the definitions of correlated divergence and correlated mutations scores. Detection of co-evolution is only one element of our system. In addition, we also consider other interaction factors, such as the ionic charge potential and hydrogen-bonding potential. EM-like approaches are employed to improve the method sensitivity and robustness, and a probabilistic framework is used to assess the likelihood of an interaction and to localize the interaction site.

Our method presents more than just a refinement over previous co-evolution methods. Current methods merely provide evidence of the correlated mutation signal in some cases: our method provides a prediction framework. Moreover, it extends to include other elements that are indicative of protein-protein interactions, using a probabilistic model (the model can be further extended to include other features, such as sequence signatures and the existence of interacting homologs). Finally, previous methods were never attempted on a large scale. Here, we test our methods extensively on a large set of known protein-protein interactions to evaluate the potential of each of the individual measures, as well as of the combined model.

This paper is organized as follows. We first review the different potentials used in our system and the EM-like approaches. We then test our algorithms on interaction we extract from BIND [Bader et al. 2001]. Finally, a few predictions are attempted and compared to the known binding sites.

2 Methods

2.1 Definitions

We are given two query proteins q_1, q_2 and a multiple alignment of each with other, related, database sequences. The groups of database sequences are referred to as protein families F_1 and F_2. The families consist of n_1 (n_2) sequences originated from N_1 (N_2) organisms (note that usually $N_i < n_i$). Denote the organism sets by O_1 and O_2 respectively. We hypothesize that the query sequences interact and therefore there is an evidence of co-evolution among the two protein families.

The first step in our co-evolutionary protein-protein interaction algorithm is to select the subset of organisms that are common to both protein families, $O_{common} = O_1 \cap O_2$. In each organism $o \in O_{common}$, there exist at least one protein in F_1 and one protein in F_2. Each such pair is a candidate interacting pair. Initially, we assume that of all possible candidate pairs in an organism, at least one interacts in a similar way as the two query proteins, therefore the core set of interacting proteins consists of at least O_{common} interacting pairs. Assuming that all proteins from a family are involved in the hypothesized interaction, the maximal size of the set may be larger. Both these assumptions are revised later. With a little abuse of notation we revise our definitions of F_i, n_i and N_i to the groups of proteins that originate from the (smaller) organism set O_{common}. Denote $n = |O_{common}|$. The set of proteins from family F_i that is found in organism o is denoted by $F_{i,o}$.

Our model consists of four elements: the correlated divergence of F_1 and F_2, the correlated mutations potential, the charged potential and the hydrogen-bonding potential, described next. This general framework can be extended to include other elements such as sequence signatures and other types of prior information.

2.2 Correlated divergence

The assumption of co-evolution for interacting proteins holds only for the interaction site, however, this information is clearly an unknown parameter in our model, where the two query proteins are only hypothesized to interact. Even for known interactions, the binding box is usually unknown. Out of over 10,000 protein interactions in the BIND database, only about 100 have detailed binding site information. However, one would expect interacting proteins to have similar divergence patterns overall. Although not necessarily constrained in a correlated manner outside of the binding box, one could expect a similar rate of mutations for both interacting proteins. This is not true for different protein families in general, as different families have different *clocks*, some exhibiting faster mutation rates than others, where rapidly changing protein might undergo significant changes in conformation. However, since the general structure of the protein has to be preserved to maintain a structurally and functionally active interaction site, it is less likely that the two interacting proteins will evolve in significantly different paces.

Our first measure attempts to detect signals of **correlated divergence** and is similar to the MirrorTree approach [Goh et al. 2000]. To estimate the divergence rate we compute the total number of amino acid mutations between all protein pairs within a given family. If two protein families co-evolve, these mutation levels should correlate, indicating similar protein clocks. For each organism pair $o_1, o_2 \in O_{common}$ within a protein family F_i we define the mutation level (ML) to be the average number of mutations over of all protein pairs in the cross product of F_{i,o_1} and F_{i,o_2}. In order to prevent biases of different length proteins, the mutation level is normalized per 100 residues.

This measure was chosen as a starting point due to its simplicity and relative ease of implementation and computational speed. Moreover, this measure qualifies as a distance metric, as it is symmetric, non-negative and it satisfies the triangle inequality. This property is necessary to accurately estimate distances between sites that are evolutionary more distant. It is also a necessary condition for constructing a phylogenetic topology.

The correlated divergence is estimated by computing the Pearson correlation coefficient r between the mutations levels of organism pairs in family 1 and family 2:

$$r = \frac{\sum_{i=1}^{n}\sum_{j=i}^{n}(ML_1(o_i,o_j))(ML_2(o_i,o_j))}{\sqrt{\sum_{i=1}^{n}\sum_{j=i}^{n} ML_1(o_i,o_j)^2}\sqrt{\sum_{i=1}^{n}\sum_{j=i}^{n} ML_2(o_i,o_j)^2}}$$

The correlation coefficient is ranging in value from -1 (anti-correlation) to 1 (high correlation). Value of 0 indicates no correlation. In practice, this raw correlation score is not very robust and normalization measures are necessary when assessing the true correlation, as discussed in the next section.

An even more stressing problem is that the method does not scale well, as the number of pairs grows quadratically with the number of organisms. If the actual subset of organisms with interacting pairs ($O_{interact}$) is much smaller than the total number of organisms (i.e. $O_{interact} << O_{common}$), then the ratio of interacting pairs to all pairs, $\frac{|O_{interact}|^2}{|O_{common}|^2}$ will be close to zero, making it almost impossible to detect correlation signals. This problem will be addressed in section 2.5.

2.2.1 Estimating background distribution

The Pearson Correlation Coefficient is not a very robust test statistic. The score can be heavily influenced by a few outliers. Consider the scatterplots of the mutation levels (and their counterpart, the conservation levels) of two different interactions (Figure 1). The plot on the left shows signs of co-evolution, as the conservation scores are widely distributed and fit the linear regression model quite nicely. The plot on the right, however, is heavily influenced by five outliers pointed to by arrows. These points correspond to almost identical protein pairs (i.e. proteins that hardly mutated during speciation). Although these points represent only a small fraction of the total number of points, their exclusion causes the correlation score to decrease from .8519 to .2714.

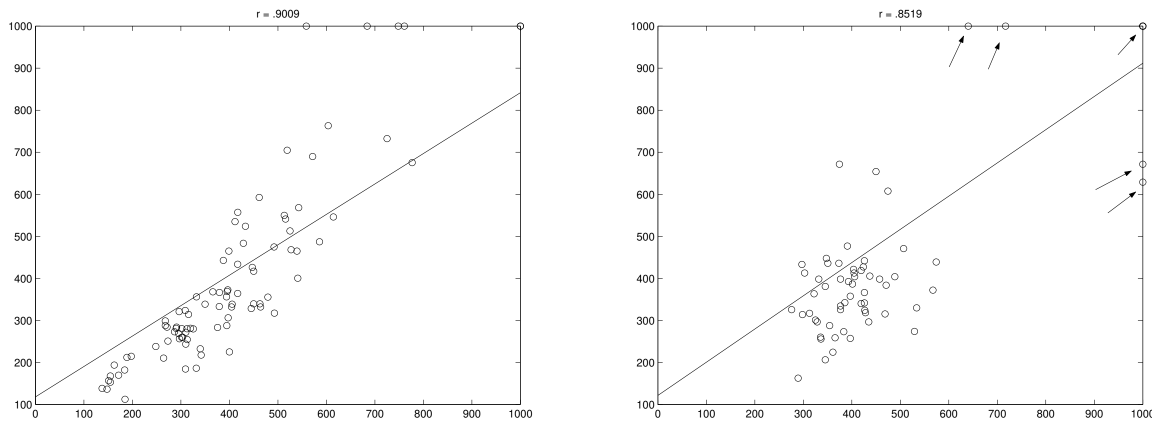

Figure 1: **Instability of the Pearson Correlation Coefficient**. Scatterplots of conservation levels ($1000 - mutations$) of two interactions: **(a)** enoyl-CoA isomerase (GI: 6323314) and Delta(3,5)-delta(2,4)-dienoyl-CoA isomerase (GI: 6324754). **(b)** bacterial seryl-tRNA synthases (GI: 6321799) and Hypothetical ORF (GI: 6325261). In both cases the correlation is high but only the left shows signs of co-evolution.

There are several approaches one can take to handle these kind of situations. One possibility is to iteratively exclude outliers (above a certain threshold), however, there is no rigorous was of determining the threshold, and one usually uses an arbitrary one. In the specific case in hand, one can weigh the entries based on their relative similarity, to decrease the contributions of highly similar proteins. Here we carry a simple simulation to create a background distribution for the co-evolution correlation scores of non-corresponding proteins and normalize the raw Pearson score accordingly. This is done by randomly permuting the organism order of one of the families and recomputing r as defined above. From this distribution, the mean Pearson correlation score, $\bar{r}$, and standard deviation, $\bar{\sigma}$, are calculated. Along with the aligned correlation score, r^*, the normalized z-score is calculated as:

$$z = \frac{r^* - \bar{r}}{\sigma}$$

When performed on the above scatterplots with 50 randomly permuted organism sets (and hence 50 background distribution points), the procedure resulted in $\bar{r} = .5926$ and $\sigma = .0597$ for the first interaction. As expected, the normalized score for this interaction, $z = 5.1580$, is very high. When this procedure was run on the plot on the right, however, $\bar{r} = .8306$ and $\sigma = .0252$. As expected, the normalized score, $z = .8423$, is insignificant.

Finally, we validate that the co-evolution confirmed by the normalized z-score is causally related to the protein-protein interaction and not the result of correlation due to some other form of co-adaptation. This is achieved by performing the above procedure on a set of non-interacting protein pairs. The two populations are then compared in terms of the distributions of their z-scores. The results are given in section 3.

2.3 Correlated mutations

Every protein-protein interaction occurs within some specified subregion of the protein. This subregion is known as the interaction binding site. Revisiting our initial premise, we can conclude that only these binding sites should show relatively strong signs of adaptation and coevolution. While our first measures seeks a generally similar divergence levels across species for both protein families, our second measure focuses on the correlation between amino acid mutations that occur within the interaction site. Our assumption is that both phenomena should happen to preserve the viability of an interacting protein pair.

Our first measure in a series of measures is similar to one described in [Pazos et al. 1997] with adaptation to multiple proteins per organism. Consider the co-evolution between the i^{th} amino acid position in the first family, a_i, and the j^{th} amino acid position in the second family, b_j. Refining the correlated divergence model, we use a conventional log-odds based scoring matrix (BLOSUM62 [Henikoff & Henikoff 1992]) to determine the evolutionary distance between residues a and b (instead of the McLachlan scoring matrix that was used originally). We calculate this co-evolution between residue position i of family 1 and position j of family 2 using the Pearson Correlation Coefficient:

$$r_{ij} = \frac{\sum_{o_1=1}^{n}\sum_{o_2=o_1+1}^{n} sim(a_i(o_1), a_i(o_2)) * sim(b_j(o_1), b_j(o_2))}{\sqrt{\sum_{o_1=1}^{n}\sum_{o_2=o_1+1}^{n} sim(a_i(o_1), a_i(o_2))^2}\sqrt{\sum_{o_1=1}^{n}\sum_{o_2=o_1+1}^{n} sim(b_j(o_1), b_j(o_2))^2}}$$

where $a_i(o)$ is the set of amino acids at position i of proteins in organism o and $sim(a_i(o1), a_i(o2))$ is the average pairwise similarity between the two sets.

This correlation score is then calculated for each pair of positions to create a correlation matrix of size $(p \times q)$, where p is the alignment length of family 1, and q is the alignment length of family 2. Protein-protein interactions should take on the form of a 'hot spot' in this matrix: the binding site areas should have significantly higher correlation scores than non-binding sites.

This method introduces two elements that are missing from the correlated divergence method: the interacting domain is localized, and the use of a log-odds substitution matrix accounts for the non-uniform amino acid substitution probabilities. On the other hand, this method suffers from the same scalability problem as correlated divergence (the problem is addressed in section 2.5) and it also introduces two drawbacks. First, the runtime of this method is considerably longer: $O(m^2n^2)$, m = length of the longer of the aligned protein families, $n = |O_{common}|$. Secondly, the log-odds substitution matrices that are used here are not distance metrics. Although they can easily be scaled to satisfy the non-negativity constraint, no algebraic manipulation exists to satisfy the triangle inequality.

2.3.1 Estimating background distribution

Coupled with the non-robustness of the Pearson Correlation Coefficient, the raw score in this method also suffers from the fact that many residue positions within a protein family are not well preserved and might not exist in certain organisms. Therefore, separate background distributions need to be calculated for each position in the correlated mutation matrix. This is done in the same way as in the correlated divergence method. The organism order is shuffled for one of the families, r is recomputed, and a z-score is associated with the raw correlation score based on that distribution of random scores.

To test our premise that the interaction sites should show higher correlation than non-interacting sites, we must again compare the z-score observed for interacting sites vs. those observed for non-interacting sites. If the exact amino acid binding pairs were known, then this could most easily be accomplished. In practice, however, this information is rarely known. More commonly known is the interaction site domain for each protein: a range of amino acids over which the interaction occurs. These two interaction site domains form what is known as an interaction bounding box within the correlated mutation matrix.

Clearly, not all positions in the bounding box co-evolve. Volume and stereo-chemical constraints allow each amino acid to be spatially close to only a very few other amino acids. Therefore, given a bounding box of size $L \times L$ it is likely that the interaction interface involves only $O(L)$ amino-acid pairs. In that view, averaging the correlated mutations score over the bounding box will diminish any signal that might exist in the bounding box.

To assess the correlated mutations potential of a bounding box we use two measures. The first is the number of significant positions in the box $M(B)$, where a significant position is defined as a position with a correlated mutations zscore > 3 (a parameter that can be optimized for performance). The second is the average zscore of the L highest scoring positions in the box $A(B)$. The significance of both measures is assessed by characterizing the distribution of scores over a large population of bounding boxes. We characterize the distribution both empirically and based on a simple model.

To empirically estimate the significance of a score S we utilize a natural normalization metric: create a background distribution $P_0(S)$ by selecting random bounding boxes of the same size, located outside of the true interacting binding box, and computing $M(B)$ and $A(B)$ for each one.

One can also assume a simple model to estimate $P_0(S)$. Specifically, we are given a complete mutation matrix of size $N = L_1 * L_2$ with total of T significant correlated mutations and a putative binding box B of size $n = l_1 * l_2$ with t significant correlated mutations. The null hypothesis is that the correlated mutations are randomly scattered in the matrix, in other words the probability that a position will be deemed significant is $p = n/N$. The number of significant positions within a random box of size n is expected to follow the binomial distribution

$$P(t\ correlated\ mutations) = \binom{n}{t} p^t (1-p)^{T-t}$$

Given the observed number of correlated mutations (or the type $(t_1, T_1 - t_1)$) one can estimate the significance by computing the probability to observe types that are less probable (see [Bejerano 2003]).

This method also leads to a more natural approach to *predicting* interaction bounding boxes: a 2-dimensional function could be estimated to locate the maximally correlated fixed-size bounding box. After the optimal position is found, the bounding box size could then be tuned accordingly (see section 3.4).

Finally, as a validation step, we create a population of non-interacting protein pairs with randomized bounding boxes. These bounding boxes are created in a random manner such that their positions are chosen uniformly with respect to the size of the interaction matrix, and their dimensions drawn from inferred interacting bounding boxes distributions. We repeat the process described above on this set of these non-interacting pairs. Given our hypothesis, we expect the normalized bounding box scores to be higher in the interaction set than in our non-interaction set.

2.4 Charge and hydrogen-bonding potentials

Vital to any protein-protein interaction is a bio-chemically feasible binding site. In order for two proteins to interact, they must be able to physically bond. We define the bonding potential between amino acid pairs within potentially interacting proteins. The two types of bonds we consider are ionic bonds (salt-bridges) and hydrogen bonds. An ionic bond can occur if one binding amino acid has a positive charge, and the other has a negative charge. A hydrogen bond can take place when one residue has the capability of donating a hydrogen atom (i.e. a single proton) and the other has the ability to accept protons. While some amino acids have the ability to either donate *or* accept protons, we do not assign residues to donor or acceptor but rather calculate their potential. Table 1 shows charge and hydrogen bonding potentials for each of the 20 amino acids.

For each pair of potential protein interactions between the two families, both the charge and hydrogen bond potentials are calculated for all amino acid pairings. The interaction potential, p, between position i of family 1 and position j of family 2 is defined as:

$$p = \frac{1}{|O_{ij}|} \sum_{o \epsilon O_{ij}} potential(a^1_{i_o}, b^2_{j_o})$$

where O_{ij} is the set of organisms that are preserved in position i of family 1 and position j of family 2 and $potential(a^1_{i_o}, b^2_{j_o})$ is the average potential between amino acids in position i and j of proteins found in organism o and families 1 and 2, respectively. Note that potentials are only calculated for proteins within a given organism.

This method must also makes use of the interaction bounding box. As with correlated mutations, the interaction bounding box significance is estimated based on the background distribution of potentials for non-interacting bounding boxes and using the model described in section 2.3.1.

2.5 Expectation Maximization approaches

Although the conceptual framework presented in the preceding methods is very clean and simple, in practice, many assumptions made are in need of revision. The EM-like approaches outlined in this section are primarily

Amino Acid	Charge	HB
Alanine	0	Neither
Arginine	+	Acceptor
Asparagine	0	Acceptor
Aspartic acid	-	Donor
Cysteine	0	Acceptor, Donor
Glutamine	0	Donor
Glutamic acid	-	Donor
Glycine	0	None
Histidine	+	Acceptor, Donor
Isoleucine	0	None
Leucine	0	None
Lysine	+	Acceptor
Methionine	0	Donor
Phenylalanine	0	None
Proline	0	None
Serine	0	Acceptor, Donor
Threonine	0	Acceptor, Donor
Tryptophan	0	Acceptor
Tyrosine	0	Acceptor, Donor
Valine	0	None

Table 1: **Charge and Hydrogen Bond Potentials**. Opposite charges or donnor/acceptor pairs are scored +1. Similarly charged or pairs of donnor/donnor and acceptor/acceptor are assigned a score of -1. Otherwise the score is set to 0.

concerned with determining the true protein-protein interaction subset. Our approach consists of three main steps: (1) determining the subset of interacting proteins within an organism (2) determining the subset of relevant organisms (3) predicting the binding box.

2.5.1 Protein subset selection

It is not uncommon to find in a genome multiple genes that belong to the same protein family. Indeed, for many of the interactions we studied, we collected multiple instances from the same genome o for each one of the interacting protein families. There are two evolutionary phenomena that can explain why these sets have size greater than one. In both cases the phenomena is driven by duplication events. The difference lies in the timing

1. At the time of speciation of o, only one protein from each family existed. Duplication events then took place in o, possibly in coordination (e.g. when the interacting genes are located physically close to each other), forming multiple homologous genes and possibly multiple interacting pairs.

2. These multiple homologous genes were in existence before speciation occurs for o. When speciation occurred, these homologs were passed on to o.

Given two sets of proteins, $F_{1,o}$ and $F_{2,o}$, in organism o from families F_1 and F_2, respectively, our methods currently take the average of the $|F_{1,o}| \times |F_{2,o}|$ protein pairs between them. However, even if one of these gene pairs is known to interact, it is unclear if all homologous gene pairs can from an interaction. We contend that this protein multiplicity will only weaken our co-evolutionary signal, as many of the pairs considered are not interacting, and therefore are likely to evolve without explicit co-evolution constraints. This affects both the divergence and the correlated-mutations measures: Consider a protein set, $F_{i,o}$ such that $|F_{i,o}| > 1$. If the protein homologs are a result of case 1, then it is impossible to deduce their *co-evolution*. Without a speciation barrier, one cannot determine a protein's evolutionary path. On the other hand, if this multiplicity is a result of case 2, then case 1 must have occurred for some ancestor, o_a of o, in which case it is impossible to determine the protein *divergence* for proteins in o_a.

It should be noted that there is an overwhelming evidence that biological systems employ fail-safe, redundancy-based mechanisms thus suggesting that many of these pairs are actually interacting (Dirks and Yona, unpublished results [Dirks & Yona 2003]). Nevertheless, the maximal number of expected interactions is of the order of $O(n)$ while the actual number of pairs considered in this analysis is of the order of $O(n^2)$. Thus the majority of pairs is only indirectly constrained. In this view it is clear that one should consider only the truly interacting pairs in the analysis, excluding all other homologs, even those are significantly similar. However, without experimentation it is hard to determine in advance which pairs interact and which are not.

To address this problem we propose a two steps EM-like algorithm (a third one is proposed for future work). In our case the hidden variables are the indicator variables that specify if the gene interacts with another gene or not. The maximization step selects from each organism the pair that maximizes the co-evolution signal, given the existing set. The expectation step is essentially the averaging over all pairs in all other genomes[1].

Our first EM approach consists of reducing the set of interacting proteins in each organism to a single interaction, i.e. $|F_{i,o}| = 1$ for both families and all organisms o. As discussed above, it is very unlikely that all possible pairs interact, however, we are almost guaranteed that at least one pair interacts, and we target that gene pair. Our final algorithm (shown below), first iterates through all protein organism sets and then continues these epochs until the subset remains unchanged and the correlation scores stop improving.

EM Protein Subset Selection Algorithm

```
f̂1^0 = f1;          // f̂1^i, f̂2^i are the reduced protein families from epoch i
f̂2^0 = f2;
epoch = 0;
loop until convergence:(f̂1^epoch == f̂1^(epoch−1)) and (f̂2^epoch == f̂2^(epoch−1))
     f̂1^epoch = f̂1^(epoch−1);
     f̂2^epoch = f̂2^(epoch−1);
     foreach (organism o, in interaction)
          r* = ComputeCorrelatedDivergence(f̂1^epoch, f̂2^epoch);
          foreach (protein p1 in f1, p2 in f2)
               f̄1 = f̂1^epoch, s.t.|P_(f̂1^epoch, o)| = p1;
               f̄2 = f̂2^epoch, s.t.|P_(f̂2^epoch, o)| = p2;
               r̄ = ComputeCorrelateDivergence(f̄1, f̄2);
               if (r̄ > r*)
                    f̂1^epoch = f̄1;
                    f̂2^epoch = f̄2;
                    r* = r̄;
               end
          loop
     loop
loop
```

2.5.2 Organism subset selection

Our assumption that $O_{common} = O_{interact}$ is, in practice, not always true. After or during speciation a pair of interacting proteins might loose its ability to interact. In section 2.2, we conclude that if $O_{interact} << O_{common}$, then co-evolution signals from interacting proteins can quite easily be overwhelmed by data from the numerous non-interacting pairs, therefore we would like to minimize this set to the smallest possible one, $O_{interact}$. To solve this problem one can apply an iterative algorithm that in each iteration eliminates the organisms that reduce the correlation score and adds the organisms that improve the score. An alternative simple, greedy heuristic is to continually remove an organism o_i from the current set of organisms, O, such that:

$$o_i = \arg \max_{i \epsilon O_{common}} (ComputeCorrelatedDivergence(F_1 - F_{1,o_i}, F_2 - F_{2,o_i}))$$

[1] Formally, this procedure is not exactly an Expectation Maximization algorithm, however it is inspired by the EM algorithm and is therefore referred to as an EM-like procedure.

Repeat until the correlation coefficient r stops improving, or the size of O becomes too small to feasibly detect co-evolution.

2.5.3 Revising the evolutionary model for correlated mutations

In the standard model of correlated mutations (as in sections 2.2 and 2.3), all protein pairs are considered when evaluating co-evolution. However, even with the minimal set of organisms and proteins (as in sections 2.5.1 and 2.5.2), this model is contradictory to the fundamental protein co-evolution assumptions. Given a phylogenetic tree topology representing $O_{interact}$, each organism o in this set is only directly related to a small subset of organisms. Although indirect relationships to all other organisms can be established by transitivity, the correlated mutations signal decreases exponentially with the evolutionary distance between species, to practically undetectable levels for even relatively small distances. Assume that the probability for a random correlated mutation is p (a function of the evolutionary pressure). Given a pair of organisms that are k evolutionary units apart, their correlated mutations signal will be weaker than a direct random mutation if $p^k < 1 - p$. For even very high correlated mutation probabilities $p = 0.8$, this is satisfied for already $k = 7$. Therefore, when analyzing co-evolution within a protein family, one has to adjust the model and compare only proteins whose organisms are phylogenetically closely related.

This approach presents various challenges. As discussed in the introduction, phylogenetic tree reconstruction is an extremely difficult problem. Currently, complete phylogenetic species trees that span across higher order domains (i.e. Eubacteria, Eukaryote, and Archea) are largely unavailable. To approximately define the set of relationships that are induced by the true phylogenetic tree we use the given set of proteins from the two interacting families. However, we do not try to recover the complete topology of the tree, since we are only interested in organism pairs that are closely related. When determining co-evolution, we consider only the k most similar relations, where k is optimized by maximizing the evolutionary divergence (maximizing the statistical distance between the distributions of divergence scores for interacting and non-interacting pairs). we refer to this model as the **Star Model**. Specifically, for each organism we compile the list of neighboring organisms using the k-nearest neighbor algorithm, utilizing the smith-waterman [Smith & Waterman 1981] local alignment algorithm. Proteins with higher similarity are assumed to be more closely related (since the proteins are already known to be related, the local alignment similarity scores can quite reliably indicate which organisms are closer than others to the query sequence).

3 Results

The majority of protein-protein interaction data was based on the *Saccharomyces cerevisiae* (yeast) data set. The functions of most proteins in the yeast genome are known, thus making this data set very suitable for training and testing our models. Over 4,000 interacting pairs were used in our analysis. All of the interactions used can be found at www.bind.ca. This database is one of the most extensive resources of information on protein-protein interactions. We also considered other databases such as DIP [Xenarios et al. 2001] and BRITE [BRITE]. However we could not find information on the interaction bounding box in these other databases.

For each interacting pair the analysis starts by generating multiple alignments. To generate the alignments, each protein was searched against the Swissprot & TrEMBL database, using BLAST [Altschul et al. 1997] with evalue threshold of 0.01. The homologous sequences were then aligned using CLUSTALW [Higgins et al. 1996] and iterative PSIBLAST [Altschul et al. 1997], using the query sequence as the seed. The multiple alignments were then analyzed as described in section 2.

3.1 Signals of correlated divergence

The first ingredient of our model is the correlated divergence score. Figure 2 plots the distribution of z-scores for 4,233 interacting pairs and 4,707 non-interacting pairs. For each of these pairs, $O_{common} \geq 5$. This threshold was arbitrarily determined. For each interacting pair, a background distribution of 50 data points was created by the randomization algorithm described in section 2.2, and results were normalized accordingly.

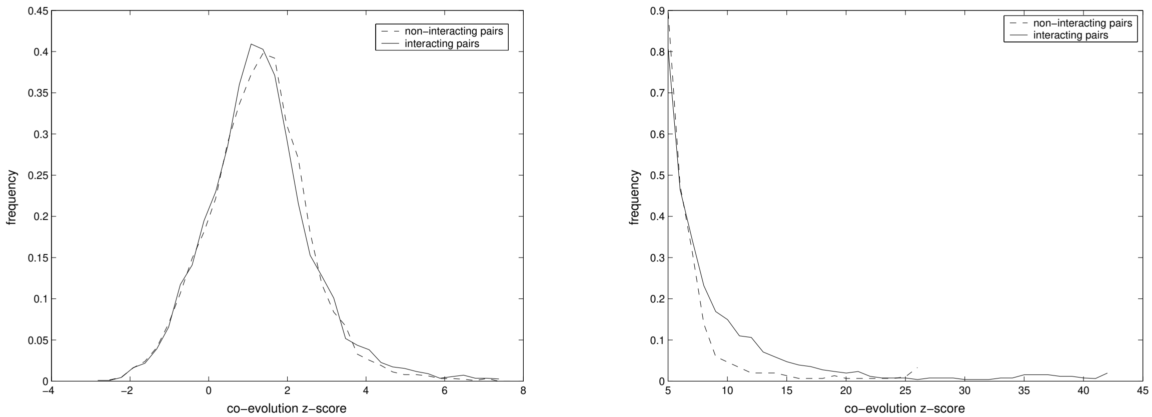

Figure 2: **Distribution of divergence scores for interacting vs. non-interacting pairs**. While both distributions are very similar (a), the long tail of the first distribution suggests this method can be very effective in detecting strong co-evolution signals among some protein interactions. The outliers present in the distribution tails are magnified in (b)

	μ	σ	95% confidence interval
Interacting Pairs	1.6609	2.4037	[1.5885, 1.7334]
Interacting Pairs (outliers removed)	1.4597	1.3672	[1.4182, 1.5012]
Non-interacting Pairs	1.4213	1.2599	[1.3852, 1.4573]

Table 2: **Correlated divergence statistical results**

Although both populations appear to be normally distributed, there are 135 outliers with z-scores > 5 in the interacting pairs distribution, compared with only 54 such pairs among the non-interacting set. These outliers have a significant effect on the distribution's standard deviation. Excluding these points, the standard deviation drops by more than half to 1.193. This relatively small number of outliers suggests that this approach may only be effective in inferring a subset of interactions. Nevertheless, the signal, when exists, seems to be quite reliable, as the false positive rate is less than 15% with z-score > 8. Figure 3 shows the similarity distributions for two of these outlying interaction pairs.

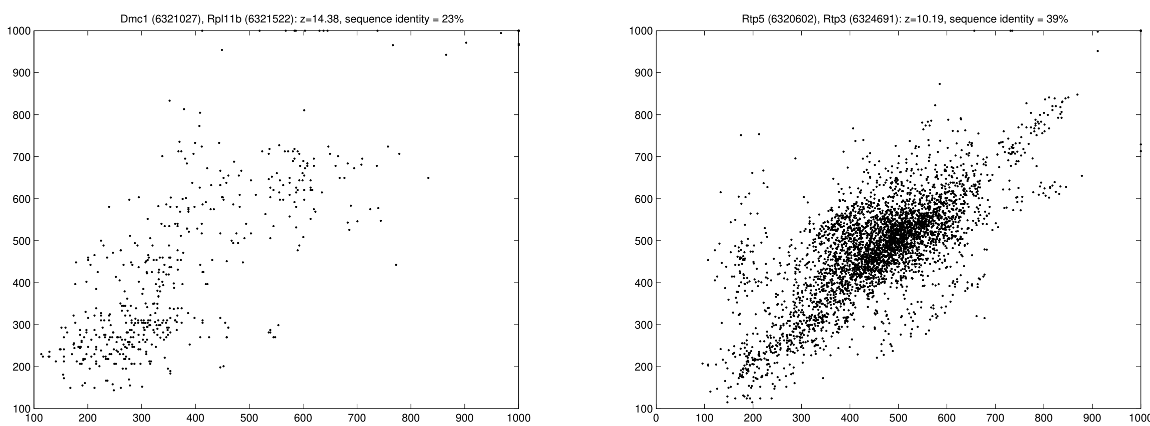

Figure 3: **Co-evolution similarity plots for interacting pairs with strong correlated divergence scores.** Gene names (with GenBank IDs in parenthesis) and the zscore are given on top of each figure.

It should be noted that there is a strong correlation ($r = .6677$) between $|O_{common}|$ and z-scores, as expected (Figure 4a). Statistically significant co-evolution is more difficult to detect with smaller organism subsets, and relatively large organisms subsets (i.e. > 20) are necessary in order to use this approach. The results also seem to be somewhat sensitive to the alignment algorithm used. With the exception of two strongly correlated outliers, the correlated divergence algorithm performed quite differently when run using CLUSTALW generated alignments vs. when run with PSIBLAST generated alignments. A scatterplot comparing output can be found in Figure 4b. Excluding the two outliers, the pearson correlation score is .2359.

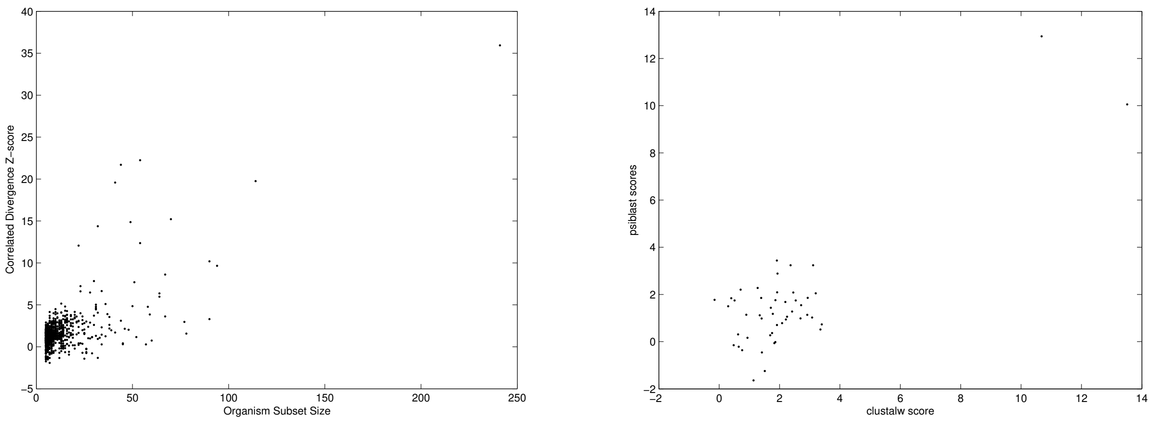

Figure 4: **(a) Correlated divergence z-scores vs. organizm subset size. (b) Correlated divergence scores resulting from PSIBLAST and CLUSTALW multiple alignments**

Table 3 shows the z-scores and corresponding interacting partners for the top 20 interactions from the correlated divergence algorithm.

z-score	Protein 1 [Genbank ID]	Protein 2 [Genbank ID]
39.3929	alpha-4-beta-4 subunit of isocitrate dehydrogenase 1 [6324291]	NAD+-dependent isocitrate dehydrogenase [6324709]
29.0567	Ribosomal protein L23B (L17aB) (YL32) [6320963]	Ribosomal protein L23A (L17aA) (YL32) [6319384]
23.9282	GTP-binding protein, RAS superfamily [6323576]	probable purine nucleotide-binding protein [6322866]
22.2419	Ribosomal protein P1B (L44') (YP1beta) (Ax) [6320073]	Ribosomal protein P2B (YP2beta) (L45) [6320590]
19.9891	60S ribosomal protein L11B (L16B) (rp39B) (YL22) [6321522]	meiosis-specific protein related to RecA and Rad51p [6321027]
17.4560	60S ribosomal protein P0 (A0) (L10E) [6323371]	Ribosomal protein P1B (L44') (YP1beta) (Ax) [6320073]
14.5120	60S ribosomal protein P0 (A0) (L10E) [6323371]	Ribosomal protein P2B (YP2beta)(L45) [6320590]
12.5256	MEK homolog [6320042]	MAP protein kinase homolog [6321477]
12.5202	CTP synthase [6322563]	CTP synthase, highly homologus to URA8 CTP synthase [6319432]
12.3582	phosphofructokinase alpha subunit [6321679]	phosphofructokinase beta subunit [6323861]
11.4844	serine/threonine MAP kinase [6321819]	protein kinase [6325117]
11.2995	Actin assembly factor [6324755]	Proline-rich protein verprolin [6323370]
11.0076	20S proteasome subunit beta3_sc [6320941]	putative proteasome subunit [6324731]
10.8943	26S protease regulatory subunit [6324691]	ATPase component of the 26S proteasome complex [6320602]
10.8319	Acidic ribosomal protein P1A (YP1alpha) (A1) [6320122]	60S ribosomal protein P0 (A0) (L10E) [6323371]
10.6663	member of 70 kDa heat shock protein family [6320950]	Mitochondrial isoform of citrate synthase [6325257]
10.6316	similar to formate dehydrogenases [6324964]	3-phosphoglycerate dehydrogenase [6320925]
10.2146	similar to a C.elegans ZK632.3 protein [6324693]	Ribosomal protein S1A (rp10A) [6323474]
9.6602	meiosis specific protein, E.coli MutS protein [6321109]	MutS homolog involved in chromosome exchange [6320047]
9.5135	myosin I [6323756]	formin homology domains; homologous to BNR1 [6324058]

Table 3: **Correlated Divergence results of top 20 interactions**

3.2 Signals of correlated mutations

Unfortunately, the number of protein-protein interactions with known binding sites is very small. Of the 12000+ interactions listed in BIND, only 150 have detailed binding site information. Among these 150 interactions, the following problems trimmed this set down to 55 interactions:

1. $O_{common} < 5$: If the number of total organisms is too small, co-evolution cannot be inferred

2. The interaction bounding box is too large (covering all or almost all of the interacting proteins) making it impossible to accurately sample non-interacting regions.

3. Database errors - several interactions had invalid bounding boxes (i.e. the bounding box position overran the length of the protein).

The correlated mutation score of each pair of positions was normalized by a background distribution of 50 randomized samples (shuffled). Positions which were completely preserved were ignored. In this case, the standard deviation is zero and normalization is impossible. Positions in which the number of aligned, non-gapped amino acids was very small (i.e. < 5) were also ignored. This threshold was also arbitrarily determined.

Preliminary results for our basic correlated mutation method do not show strong signals. However, we hypothesized from our correlated divergence results, that perhaps only a relatively small percentage of protein interactions will show strong signs of co-evolution. With approximately only 50 data points, we were not able to draw any such conclusions in our correlated mutations data set. Table 4 gives a few examples of known protein-protein interactions which resulted in relatively high z-scores.

BIND ID	z-score	Protein 1	length	binding site	Protein 2	length	binding site
12534	2.2806	plexin-B3	1909	1278-1909	glutamate transport modulator GTRAP48	1522	14-375
12535	2.1731	plexin-B3	1909	1260-1891	atrophin-1	1278	226-603
12505	1.8071	neurexin I-alpha	1542	1487-1542	myosin head and IQ CaM binding motifatrophin-1	1045	281-662

Table 4: **Correlated Mutation results (EM Algorithm)**

3.2.1 Protein subset selection

To improve the correlated mutation scores, a protein subset selection algorithm was implemented as described in section 2.5.1. This algorithm tends to pick proteins that are more similar to the multiple alignment seed protein (see Figure 5). When applied to the same data set as in section 3.1, the algorithm improved the separation between interacting and non-interacting pairs, but the improvement over the basic method is marginal, with the bounding box z-scores for both interacting and non-interacting sets appear to be normally distributed and centered about zero.

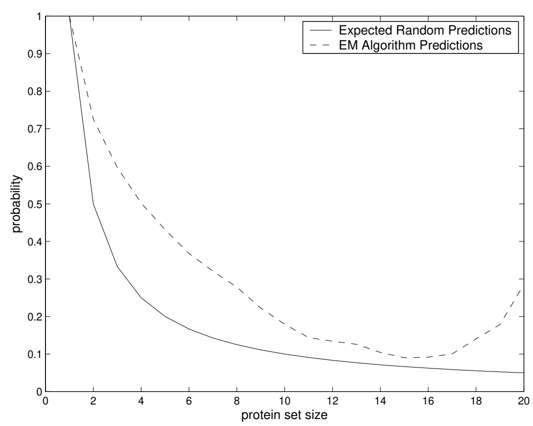

Figure 5: The Protein Subset Selection EM algorithm tends to pick proteins more likely to seed proteins (maximum similarity score to the seed protein). The expected random baseline is simply $\frac{1}{protein\ set\ size}$. For any given protein set size, the EM algorithm tends to choose the more similar protein with higher than expected probability.

3.2.2 The reduced evolutionary model

Our kNN-based algorithms (the star model) showed slight improvements over the basic correlated mutations algorithm and the protein subset selection EM algorithm. We ran the kNN algorithm described in the methods section on the protein subsets determined in section 2.5.3. For each protein, we find the k-closest other proteins. Since the proteins in these data sets are mapped one-to-one with organisms, the k closest proteins can be thought of as the k closest organisms. As expected, we got better performance with smaller values of k, however, the signals were still too weak and noisy (given the small size of the data).

3.3 Charge and HB potentials

Like the correlated mutation approach, these approaches also need protein-protein interaction bounding box details. Therefore, the data set available is identical to that used for correlated mutations. The following histogram (Figure 6) shows the normalized z-score frequency for both charge and hydrogen bonding potentials, drawn from 49 samples. Although the population is too small to draw any definitive conclusions, the results are promising. Table 5 outlines the statistics for the two populations.

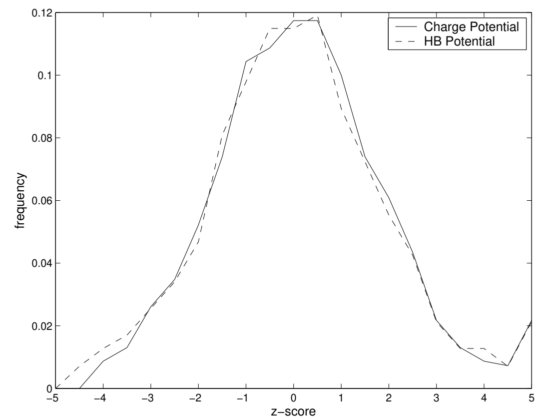

Figure 6: Histogram plot of charge and hydrogen bonding potentials within specified interaction bounding boxes

.

	μ	σ	95% confidence interval.
Charge Potential	0.3646	1.8009	[-0.1557, .8851]
HB Potential	0.2948	1.7438	[-.2037, .7933]

Table 5: Charge and HB Potential Statistical Results

3.4 Integrating the potentials into a single probabilistic framework

Even when the correlated divergence signal is weak, a hybrid model may still predict an interaction if the other components support this hypothesis. Each one of the measures described thus far contains some information about the potential of forming an interaction. However the formation of an interaction site usually requires the combination of all these factors. The likelihood of an interaction can be estimated based on the significance or the probability that one could have observed these scores at random.

Formally, we are given two protein families, F_1 and F_2 with divergence correlation $D_{1,2}$, and a candidate interaction site $I[x_1, y_1, x_2, y_2]$ (where x_1, y_1 and x_2, y_2 are the coordinates of the corners) with correlated mutations potential $M(I)$, charge potential $C(I)$ and hydrogen-bonding potential $H(I)$. The significance of the interaction can be estimated by

$$P_0(D > D_{1,2})P_0(M > M(I))P_0(C > C(I))P_0(H > H(I))$$

where we already employed the independence assumption.

A simple prediction procedure naturally follows: given two protein families compute their divergence score first as in section 2.2. Then proceed to search for the most significant bounding box. Specifically, given the bounding box size $l \times l$ scan the scoring matrix using a sliding box with a shift of 1. For each putative binding box B compute the number of

- M(B) correlated mutations (columns with correlated mutations zscore $\leq T$)
- C(B) number of salt-bridges
- H(B) number of hydrogen bonds

then compute the probability $P_0(M > M(B))P_0(C > C(B))P_0(H > H(B))$. Output the box that minimizes the probability P_0.

Alternatively, one can use the zscores obtained for each measure and take the maximal one or the product of the zscores as the final score (these measures are closely related to the pvalues, assuming uni-modal distributions). Establishing the distributions of these statistics in populations of non-interacting proteins can provide us with a measure of significance for the putative interacting proteins.

While we have not tested these procedures yet, we believe that they will improve the accuracy and sensitivity of predictions of protein-protein interactions and we are currently testing this hypothesis.

4 Discussion

In this paper we present a method to predict protein-protein interactions and the interaction bounding box using a hybrid system that combines signals of co-evolution with other binding potentials. The premise of co-evolution methods was originally revealed in [Gobel et al. 1994, Pazos et al. 1997, Goh et al. 2000], however without statistical assessment. Here we expand this model and test it extensively. As opposed to docking and binding site identification methods, our method does not require knowledge of the structure. Similar to the Rosetta stone method, it attempts to predict interactions from sequence. Our model can be extended to include prior information about sequence signatures, signals of gene fusion events and amino acid frequencies.

The underlying co-evolution model is very appealing at first: if two proteins interact one would expect to find some patterns of constrained, correlated mutations. Detecting such patterns in protein pairs that haven't been tested experimentally can suggest that the proteins interact. Indeed, previous studies that explored the potential of this method suggested that the co-evolution model might be very effective for detection of novel interactions. However, these studies focused only on a few specific examples that appear to behave according to this model.

Motivated by this premise we tested extensively this model and variants of the model that attempt to improve the signal-to-noise ratio. We also integrated this model with other elements that reflect the potential of forming an interaction between proteins. However, despite the extensions and enhancements we implemented we were disappointed to find out that only a relatively small number of interactions seem to follow the co-evolution model. There are several possible explanations:

- The data set is biased or skewed. We think this is unlikely since the data set we used is fairly large and was obtained from a high quality database (BIND).
- The metric used is not sensitive enough. We conclude that only very few interactions can be predicted with previous measures that were reported in the literature. While we tested extensions on the correlated mutations potential we have yet to explore variants of the divergence measure.

Detecting correlated mutations was even more challenging. Unfortunately, our biggest limitation is the lack of known interaction sites even for known protein-protein interactions. However, even with this small data set, it became clear that the binding sites seem to have only a weak signal of co-evolution pressure. Further refinements of the correlated mutations measure based on the star model and the EM algorithms might improve the signal. However, we believe that even then the performance would be weak since the interaction interface is usually small and the correlated mutations between sites, even if exist, might be dominated by other sites that are not constraint to the same extent. Moreover, the selective pressure might have eliminated any other possible variations that could have supported this hypothesis.

Our final conclusion is that the co-evolution model, in its current form, is not powerful enough to predict protein-protein interactions effectively in itself. Further enhancements (some are currently underway), integration with other sources of information and with other techniques, and larger data sets with refined information about the binding site may improve the performance in the future.

5 Acknowledgments

This work is supported by the National Science Foundation under Grants No. 0133311 and 0218521 to Golan Yona.

References

[Alloy & Russel 2003] Aloy, P. & Russell, R. (2003). InterPreTS: protein Interaction Prediction through Tertiary Structure. *Bioinformatics* **19**, 161-162.

[Altschul et al. 1997] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.

[Bader et al. 2001] Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F., Pawson, T. & Hogue, C. W. (2001). BIND - The Biomolecular Interaction Network Database. *Nucleic Acids Res.* **29**, 242-245.

[Bejerano 2003] Bejerano, G. (2003). Efficient exact p-value computation and applications to biosequence analysis. *In the proceedings of RECOMB 2003*, 38-47, ACM press.

[BRITE] BRITE website: http://www.genome.ad.jp/brite/

[Dandekar et al. 1998] Dandekar T, Snel B, Huynen M & Bork P. (1998). Conservation of Gene Order: A Fingerprint of Proteins that Physically Interact. *Trends Biochem Sci* **23**, 324-328.

[Dirks & Yona 2003] Dirks, W. & Yona, G. (2003). A comprehensive study of the notion of functional link between genes based on microarray data, promoter signals, protein-protein interactions and pathway analysis. *Submitted*.

[Enright et al. 1999] Enright AJ, Iliopoulos I, Kyrpides NC & Ouzounis CA. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90.

[Fields & Song 1989] Fields, S. & Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-246.

[Gallet et al. 2000] Gallet X, Charloteaux B, Thomas A & Brasseur R. (2000) A fast method to predict protein interaction sites from sequences. *J. Mol. Biol.* **302**, 917-926.

[Gobel et al. 1994] Gobel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309-317.

[Goh et al. 2000] Goh C, Bogan A, Joachimiak M, Walther D & Cohen F. (2000). Co-evolution of Proteins with their Interaction Partners. *J. Mol. Biol.* **299**, 283-293.

[Henikoff & Henikoff 1992] Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* **89**, 10915-10919.

[Higgins et al. 1996] Higgins, D. G., Thompson, J. D. & Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* **266**, 383-402.

[Ho et al. 2002] Ho Y., Gruhler, A., Heilbut, A., et al. (2003). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180-183.

[Kini & Evans. 1995] Kini RM & Evans HJ. (1995). A hypothetical structural role for proline residues in the flanking segments of protein-protein interaction sites. *Biochem. Biophys. Res. Commun.* **212**, 1115-1124.

[Lu et al. 2002] Lu, L., Lu, H. & Skolnick, J. (2002). MULTIPROSPECTOR: An Algorithm for the Prediction of Protein-Protein Interactions by Multimeric Threading. *Proteins* **49**, 350-364.

[Marcotte et al. 1999] Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO & Eisenberg D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751-753.

[Pazos et al. 1997] Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **271**, 511-523.

[Pazos & Valencia 2001] Pazos, F. & Valencia, A. (2001). Similarity of Phylogenetic Trees as Indicator of Protein-Protein Interaction. *Protein Eng* **14**, 609-614.

[Pazos & Valencia 2002] Pazos, F. & Valencia, A. (2002). In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **47**, 219-227.

[Schwikowski et al. 2000] Schwikowski, B., Uetz, P. & Fields, S. (2000). Abstract A network of protein-protein interactions in yeast. *Nat Biotechnol.* **18**, 1257-1261.

[Smith & Waterman 1981] Smith, T. F. & Waterman, M. S. (1981). Comparison of Biosequences. *Adv. App. Math.* **2**, 482-489.

[Sobolev et al. 1999] Sobolev V, Sorokine A, Prilusky J, Abola EE & Edelman M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics* **4**, 327-332.

[Sprinzak & Margalit 2001] Sprinzak, E. & Margalit, H. (2001). Correlated Sequence-signatures as Markers of Protein-Protein Interaction. *J. Mol. Biol.* **311**, 681-692.

[Teodoro et al. 2001] Teodoro, M, Phillips, G & Kavraki, L. (2001). Molecular Docking: A Problem with Thousands of Degrees of Freedom *IEEE International Conference on Robotics and Automation*.

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

[Valencia & Pazos 2002] Valencia, A. & Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.* **12**, 368-373.

[Xenarios et al. 2001] Xenarios, I., Fernandez, E., Salwinski, L., Duan, X.J., Thompson, M.J., Marcotte, E.M. & Eisenberg, D. (2001) DIP: The Database of Interacting Proteins: 2001 update. *Nucl. Acids Res.* **29**, 239-241.