BIOINFORMATICS ORIGINAL PAPER

Vol. 22 no. 1 2006, pages 40–49
doi:10.1093/bioinformatics/bti723

Phylogenetics

Inferring functional information from domain co-evolution

Yohan Kim[1], Mehmet Koyutürk[2], Umut Topkara[2], Ananth Grama[2] and Shankar Subramaniam[1,3,*]

[1]Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, CA 92093, USA, [2]Department of Computer Sciences, Purdue University, West Lafayette, IN 47907, USA and [3]Department of Bioengineering, University of California at San Diego, La Jolla, CA 92093, USA

Received on June 23, 2005; revised on October 7, 2005; accepted on October 16, 2005
Advance Access publication November 13, 2005
Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Co-evolution is a powerful mechanism for understanding protein function. Prior work in this area has shown that co-evolving proteins are more likely to share the same function than those that do not because of functional constraints. Many of the efforts founded on this observation, however, are at the level of entire sequences, implicitly assuming that the complete protein sequence follows a single evolutionary trajectory. Since it is well known that a domain can exist in various contexts, this assumption is not valid for numerous multi-domain proteins. Motivated by these observations, we introduce a novel technique called Coevolutionary-Matrix that captures co-evolution between regions of two proteins. Instead of using existing domain information, the method exploits residue-level conservation to identify co-evolving regions that might correspond to domains.

Results: We show that the Coevolutionary-Matrix method can detect greater number of known functional associations for the *Escherichia coli* proteins when compared with earlier implementations of phylogenetic profiles. Furthermore, co-evolving regions of proteins detected by our method enable us to make hypotheses about their specific functions, many of which are supported by existing biochemical studies.

Contact: shankar@sdsc.edu

1 INTRODUCTION

Identifying interacting pairs of proteins encoded in a genome is an important step towards understanding how a cell works. Towards this eventual goal, several computational and experimental techniques have been developed in recent years. For instance, recent developments in high-throughput experiments yielded protein interaction data on a very large scale (Uetz *et al*., 2000; Ito *et al*., 2001; Gavin *et al*., 2002; Ho *et al*., 2002; Giot *et al*., 2003). High-throughput methods, however, are prone to errors in terms of both false negatives and positives (von Mering *et al*., 2002). Hence, computational methods must be developed in parallel to complement experimental techniques. Indeed, integrating *in silico* analysis with experimental information provides more comprehensive and reliable understanding of functional association between proteins (Lee *et al*., 2004).

Computational methods that predict protein interactions have gained impetus from recently available databases of complete genome sequences. Using genome data, researchers have inferred functions of numerous proteins by comparing genomes across species (Dandekar *et al*., 1998; Pellegrini *et al*., 1999; Overbeek *et al*., 1999; Enright *et al*., 1999). One way of exploiting evolutionary pressure to understand function is quantifying the conservation of gene neighborhoods across genomes, which has been shown to correlate with their function (Dandekar *et al*., 1998; Overbeek *et al*., 1999). Another approach is comparing protein phylogenetic profiles, where each profile is a vector indicating presence or absence of a protein across genomes (Pellegrini *et al*., 1999). The similarity of two phylogenetic profiles, which captures the degree of co-evolution between the two corresponding proteins, has been shown to correlate with their functions (Pellegrini *et al*., 1999). Subsequent work has shown that many of the known pathways can be reconstructed using such methods (von Mering *et al*., 2003; Date and Marcotte 2003).

In earlier work, we presented a simple extension to a method based on protein phylogenetic profiles (Kim and Subramaniam, 2005). By taking into account the multi-domain nature of proteins, our method detected several known interactions missed by earlier methods. This extension, referred to as the Multiple-Profile method, simply partitioned a protein sequence into overlapping segments (e.g. 30 residues) of fixed length (e.g. 120 residues) and constructed separate phylogenetic profiles for each of these segments. Because large and fixed-length segments are used, boundaries of domains with different evolutionary histories cannot be cleanly resolved. Consequently it is difficult to assess whether these co-evolving segments correspond to true domains. In addition, there exists a possibility of introducing false phylogenetic profiles as artifacts of segmentation. This may occur when one of the segments covers two domains having different evolutionary histories.

This paper improves on existing techniques by using a novel method for identifying co-evolving regions precisely, thus reducing the number of false phylogenetic profiles. With this new tool, we show a number of examples from the *Escherichia coli* proteome

*To whom correspondence should be addressed.

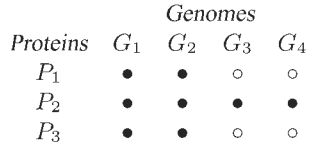

Fig. 1. An example illustrating binary phylogenetic profiles. Closed and open circles indicate presence and absence, respectively, of a protein in a genome.

where the identified co-evolving regions correspond to biochemically characterized and functionally associated domains.

2 COMPUTATIONAL METHODS AND ALGORITHMS

Our method, Coevolutionary-Matrix, is designed to assign phylogenetic similarity scores to each pair of proteins under consideration (e.g. all *E.coli* proteins) to predict functional associations between these proteins. Similar to other phylogenetic-profile-based interaction prediction methods, our method uses the amino acid sequences of proteins and a set of completely sequenced genomes belonging to different species. The method consists of three major steps:

(1) constructing detailed phylogenetic profiles for all proteins,

(2) using these profiles, constructing coevolutionary matrices for all protein pairs and

(3) assigning phylogenetic similarity scores to all protein pairs based on these matrices.

The following sections describe each of these steps in detail.

2.1 Constructing phylogenetic profiles

2.1.1 Protein phylogenetic profiles A phylogenetic profile of a protein is a vector, where each entry quantifies the existence of the protein in a genome. An example for phylogenetic profiles is shown in Figure 1. In this example, closed and open circles are used to indicate the presence or absence of a protein in a genome, respectively. Each row in the figure is the binary phylogenetic profile of the respective protein. Observe that the proteins P_1 and P_3 in the figure are likely to share a particular function as their phylogenetic profiles suggest that they have followed a similar evolutionary trajectory.

Conventional methods (Pellegrini *et al*., 1999; Date and Marcotte, 2003), hereon referred to as Single-Profile methods, rely on a single phylogenetic profile associated with each protein. Given a set of proteins $\mathbf{P} = \{P_1, P_2, \ldots, P_n\}$ and genomes $\mathbf{G} = \{G_1, G_2, \ldots, G_m\}$, the phylogenetic profile ψ_i for protein P_i is a vector defined as

$$\psi_i(j) = \frac{-1}{\log(E_{ij})}, \quad 1 \le j \le m, \tag{1}$$

where E_{ij} is the minimum (i.e. most significant) BLAST (Altschul *et al*., 1997) *E*-value of local alignments between P_i and G_j. Each profile element is thus a real value that quantifies our confidence of knowing whether a protein exists in a genome. To avoid the logarithm-induced artifacts, the maximum value that a phylogenetic profile element can take is set to 1, indicating the absence of the protein in the corresponding genome. This corresponds to an *E*-value cutoff of 0.5 if $\log_2$ is used. This threshold was used to faithfully replicate the method of Date and Marcotte (2003) so that our method can be compared with a well-known implementation of the Single-Profile method. As was noted in the same study, using real values instead of booleans for profile elements offers the advantage of capturing degrees of sequence divergence, providing greater information than booleans.

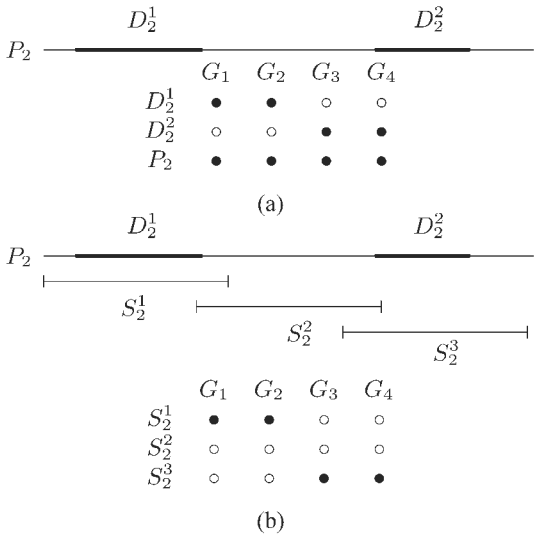

Fig. 2. (a) An example illustrating that the Single-Profile method does not capture domain-level evolutionary histories. Protein P_2 contains two domains, shown by thick lines on its sequence. While domain D_2^1 on protein P_2 follows an evolutionary trajectory similar to that of proteins P_1 and P_3 of Figure 1, the phylogenetic profile of P_2 does not reveal this information as it combines the independent evolutionary histories of D_1^2 and D_2^2. (b) Dividing P_2 into fixed-size segments, we can capture the phylogenetic similarity between proteins P_1 and P_2 since $\mu_M(P_1, P_2) = \max_s I(\psi_1, \psi_2^s) = I(\psi_1, \psi_2^1) = 1$.

For assessing the similarity between two phylogenetic profiles, mutual information provides a useful measure that takes into account co-existence and co-absence of proteins together. Indeed, it has been shown to be reliable and used successfully for predicting protein interactions (Date and Marcotte, 2003). The mutual information $I(X, Y)$ of a pair of random variables X and Y is defined as follows:

$$I(X, Y) = H(X) + H(Y) - H(X, Y), \tag{2}$$

where $H(X)$ is the Shannon entropy of X, which is defined as

$$H(X) = \sum_{x \in \mathcal{X}} p_x \log(p_x). \tag{3}$$

Here, $\mathcal{X}$ is the set of possible values taken by X and $p_x = \Pr\{X = x\}$. Similarly, $H(X, Y)$ is the joint entropy of X and Y.

A probability distribution for a phylogenetic profile ψ_i is computed by quantizing profile elements into a certain number of bins and estimating the relative frequency of each bin. Then, the phylogenetic similarity between ψ_i and ψ_j is computed as

$$\mu_S(P_i, P_j) = I(\psi_i, \psi_j) \tag{4}$$

by the Single-Profile method. In the example of Figure 1, the mutual information between the profiles of P_1 and P_3 is 1, while it is 0 between P_1 and P_2. Intuitively, as P_2 exists in all genomes, its co-existence with P_1 in some genomes does not provide any information on the functional association of these proteins.

2.1.2 Segment phylogenetic profiles While providing a useful computational method for predicting interactions between proteins, Single-Profile methods may miss many existing interactions. This is because domains within a single protein may have followed very different evolutionary trajectories. Since there are numerous multi-domain proteins in both prokaryotes and eukaryotes, such occurrence may be quite frequent. This point is illustrated by a simple example in Figure 2a. To capture

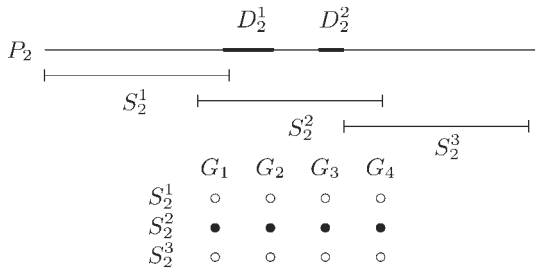

Fig. 3. A scenario where the Multiple-Profile method fails to identify domain-level co-evolution. Since the two separate domains D_2^1 and D_2^2 are covered together by segment S_2^2, their individual phylogenetic profiles do not appear in the segment phylogenetic profiles. As no a priori information is available on the size and location of the domains, it is not possible to avoid such situations using fixed-size segments.

domain-level co-evolution, the Multiple-Profile method (Kim and Subramaniam, 2005) chops each protein P_i into overlapping segments $S_i^1, S_i^2, \ldots, S_i^k$ of fixed size and computes the phylogenetic similarity between two proteins as

$$\mu_M(P_i, P_j) = \max_{s,t} I\left(\psi_i^s, \psi_j^t\right), \tag{5}$$

where ψ_i^sdenotes the phylogenetic profile for segment S_i^s of a protein P_i. The Multiple-Profile method, illustrated in Figure 2b, is shown to perform better than the Single-Profile method in identifying functional associations between proteins accurately.

2.1.3 Residue phylogenetic profiles While the Multiple-Profile method can detect known interactions missed by the Single-Profile methods by emphasizing on domain-level co-evolution, it still has flaws in capturing the underlying domain information. A scenario where the Multiple-Profile method fails to accurately identify co-evolving domains is shown in Figure 3. The figure illustrates that the Multiple-Profile method may miss potentially informative domains because segments have fixed lengths and their placements are pre-determined.

In this study, to capture the underlying domain information accurately, we further extend the phylogenetic profile-based methods by computing residue phylogenetic profiles for each protein. Our approach relies on the fact that a significant local alignment between two proteins corresponds to the unusual similarity between two contiguous portions of the two proteins rather than entire sequences. Therefore, while aligning a protein with a genome, instead of regarding a significant local alignment as the indicator of existence of the entire protein, we attribute this existence to the residues that are covered in the alignment. This allows fine-grain analysis of sequence conservation at the domain level.

Let $\mathbf{A}(P_i, G_j)$ be the set of significant local alignments between a protein P_i and a genome G_j. Each alignment $A \in \mathbf{A}(P_i, G_j)$ is associated with a contiguous interval $T(A) = [r_b, r_e]$ of residues on P_i and a BLAST E-value $E(A)$. Then, for each amino acid residue r on P_i, we define phylogenetic profile ψ_i^r as follows:

$$\psi_i^r(j) = \min_{A \in \mathbf{A}_r} -\frac{1}{\log(E(A))}, \quad 1 \le j \le m. \tag{6}$$

Here, $\mathbf{A}_r = \{A \in \mathbf{A}(P_i, G_j): r \in T(A)\}$ is the set of local alignments that contain r. In Equation (6), the most significant of E-values for a residue was chosen because we want to know whether the region of a protein covering the residue is present in a genome. Choosing less significant E-values would mean dampening the signals needed to detect the presence of this region of a protein.

Note that the phylogenetic profile of a single residue does not correspond to its conservation since the alignment can contain mismatches and gaps. However, analyzing residue-level phylogenetic profiles defined in this way provides information on the conservation of a particular portion of the protein. Specifically, if the phylogenetic profiles of a contiguous group of residues are similar, this group might indeed correspond to a conserved domain on the protein. In terms of the co-evolution of two proteins, this corresponds to the co-evolution of such contiguous regions on each protein. In the following sections, we discuss how residue profiles can be used to identify these co-evolved regions.

2.2 Computing coevolutionary matrices

To capture the co-evolution of proteins at the domain level, we construct a co-evolutionary matrix for each pair of proteins. For a pair of proteins P_i and P_j let l_i and l_j denote their respective lengths. The co-evolutionary matrix M_{ij} of P_i and P_j is an $l_i \times l_j$ rectangular matrix, where each entry corresponds to the mutual information score between a pair of residues each from one protein, i.e.

$$M_{ij}(r, s) = I\left(\psi_i^r, \psi_j^s\right), \tag{7}$$

for $1 \le r \le l_i$ and $1 \le s \le l_j$. Each entry of the matrix quantifies the residue-level co-evolution between the two proteins. If the proteins contain a co-evolved domain, this appears as a contiguous block of high mutual information scores. Sample co-evolutionary matrices for the *E.coli* proteins that are shown in Figures 8 and 9 illustrate this point.

Note that the computation of full co-evolutionary matrices might be infeasible in practice. Given a set of n proteins and m genomes, it is necessary to compute $O(n^2)$ matrices. If the longest protein consists of l residues, the overall time complexity is $O(ml^2 n^2)$. Since conserved regions are usually fairly long, considering all pairs of residues on them is redundant. Therefore, by downsampling the co-evolutionary matrix, we can avoid the complexity penalty without significantly impacting the sensitivity of the algorithm. Using a downsampling factor of f, the size of the largest co-evolutionary matrix is reduced to l^2/f^2. In general, f can set to be large enough so that l/f is bounded by a constant. Note that the complexity of an algorithm that does not consider individual residues is $O(mn^2)$. In this manner, the simplification reduces the overhead of residue-profile-based algorithm to a constant factor, l^2/f^2.

2.3 Deriving phylogenetic similarity scores

A co-evolutionary matrix contains information about which regions from two proteins have co-evolved. It is important to note that there might be spurious (large) entries in the matrix due to artifacts created while compiling BLAST outputs. To identify co-evolved regions accurately, we use a filtering scheme. Our algorithm is based on the intuition that co-evolved regions of the two proteins must be sufficiently large to be considered as significant ones. In terms of the co-evolutionary matrix, there must be a sufficiently large submatrix such that all entries in that submatrix are consistently high. Clearly, the submatrix with the maximum consistently high mutual information score provides the degree of co-evolution between the two proteins. Hence, we formulate the phylogenetic similarity between proteins P_i and P_j as follows:

$$\mu_C(P_i, P_j) = \max_{\substack{1 \le r \le l_i \\ 1 \le s \le l_j}} \min_{\substack{r \le a < r+W \\ s \le b < s+W}} M_{ij}(a, b). \tag{8}$$

Here, W is the window parameter that quantifies the sufficiency of the size of a region on a protein to be considered as a conserved domain. The overall algorithm for computing the co-evolutionary-matrix-based phylogenetic similarity between each pair of proteins is shown in Figure 4.

3 RESULTS

We implemented the proposed method and tested on 4311 *E.coli* proteins. We used 152 genomes to construct phylogenetic profiles.

Algorithm Compute Coevolutionary-Matrix
Input: Protein sequences $\mathbf{P} = \{P_i\}$
Input: Genomes $\mathbf{G} = \{G_j\}$
Output: Phylogenetic similarity scores $\mu_C(P_i, P_j)$ for each protein pair
for each $P_i \in \mathbf{P}$
  for each $G_j \in \mathbf{G}$
    $\mathbf{A} \leftarrow$ BLAST(P_i, G_j)
    ▷ $\mathbf{A}$ is the set of all local alignments $A = \{$Interval
    ▷ $T(A)$, E-value $E(A)\}$
    for each residue r on P_i
      $E_r(j) \leftarrow \min_{A \in \mathbf{A} | r \in T(A)} E(A)$
      ▷ $E_r(j)$ is the E-value of most significant
      ▷ alignment that contains r
      $\psi_i^r(j) \leftarrow -1/\log(E_r(j))$
      ▷ ψ_i^r is the phylogenetic profile of residue r of
      ▷ protein P_i
for each $P_i, P_j \in \mathbf{P} \times \mathbf{P}$
  $M_{ij} \leftarrow [I(\psi_i^r, \psi_j^s) | r \in P_i, s \in P_j]$
  ▷ M_{ij} is the coevolutionary matrix of all residue
  ▷ pairs in $P_i \times P_j$
  ▷ In the implementation, these entries are downsampled
  for each $r \in P_i, s \in P_j$
    $m(r, s) \leftarrow \min_{r \le a < r+W, s \le b < s+W} M_{ij}(a, b)$
    ▷ $m(r, s)$ is the mutual information of contiguous
    ▷ regions of size W starting at residues r on P_i
    ▷ and s on P_j
  $\mu_C(P_i, P_j) \leftarrow \max_{r \in P_i, s \in P_j} m(r, s)$

Fig. 4. Algorithm for computing Coevolutionary-Matrix.

Although some genomes are redundant in the sense that they share a large fraction of their proteins, our collection of genomes is diverse enough to cover the three branches of life (131 Bacteria, 17 Archaea and 4 Eukaryota). The complete list of genomes is at http://genome.ucsd.edu/CoevolutionaryMatrix/list-152.txt. Using a default setting, we ran BLAST (i.e. blastp program) for each one of 4311 *E.coli* proteins against each one of 152 genomes. For the Single-Profile, only the most significant *E*-value was kept for each protein. For the co-evolutionary matrix, the same BLAST run was carried out except that all matched region information and corresponding *E*-values meeting the threshold were kept.

To reduce the time and memory requirements associated with the filtering algorithm, we downsampled the co-evolutionary matrix by a factor of $f = 30$. For two proteins with l_i and l_j amino acid residues, the dimensions of their co-evolutionary matrix is $(l_i/30) \times (l_j/30)$. In addition, the parameter $W = 2$ was chosen. The use of the downsampling factor f of 30 and W of 2 translate to dividing proteins into overlapping segments that are 60 residues long. Since an average domain size is around 100 residues, current values for the f and W are reasonable.

Using this implementation of the Coevolutionary-Matrix method and an implementation of the Single-Profile method proposed by Date and Marcotte (2003), we compared their performances. Since homologous proteins should have similar phylogenetic profiles and thus have high mutual information scores, we excluded them from our analysis. To compare mutual information scores under the two methods, we converted them into *p*-values. Here, the *p*-value of a protein pair is defined as the fraction of non-homologous protein pairs in *E.coli* that have higher mutual information score than the one in question. In other words,

$$p\big(\mu(P_i, P_j)\big) = \frac{|\{(P_a, P_b) \in \mathbf{N} : \mu(P_a, P_b) > \mu\big(P_i, P_j\big)\}|}{|\mathbf{N}|}, \quad (9)$$

where $\mathbf{N}$ is the set of all non-homologous protein pairs. Here, μ denotes the phylogenetic similarity score assigned by the Single-Profile (μ_S) or Coevolutionary-Matrix (μ_C) method.

We used a set of reference protein interactions that we derived from the KEGG database (Kanehisa *et al*., 2004) to test and compare the Single-Profile and Coevolutionary-Matrix methods. We use the term 'interactions between proteins' to imply a broad range of interactions, from physical binding to functional association. In this respect, proteins participating in different steps of a biochemical pathway are considered interacting. Consequently, we define a reference interaction as a pair of proteins that share a KEGG pathway assignment. To generate this set of reference interactions, for each *E.coli* pathway retrieved from KEGG, we formed a 'clique' of proteins that participate in the corresponding pathway. The final reference set consists of 1282 proteins and 43 331 interacting protein pairs derived from these proteins after excluding homologous pairs (BLAST *E*-value <1.0).

3.1 Comparison of Coevolutionary-Matrix and Single-Profile methods

Both the Coevolutionary-Matrix and Single-Profile methods are used to predict interactions between *E.coli* proteins by setting a threshold on the phylogenetic similarity score. In other words, proteins P_i and P_j are predicted as interacting partners if $\mu(P_i, P_j) > \mu^*$. For each value of μ^*, coverage is defined as the sum of true positives (TP) and false positives (FP). Both are numbers of protein pairs that meet the threshold. Furthermore, proteins in each pair are represented in the KEGG dataset. The difference is that TP protein pairs are interacting in the KEGG dataset but not FP pairs. In addition, positive predictive value (PPV) is defined as TP/(TP + FP).

PPV versus coverage plots for the Coevolutionary-Matrix and Single-Profile methods are shown in Figure 5. A similar plot for the Multiple-Profile method is also shown for comparison. It is evident from the figure that the Coevolutionary-matrix method has about 1.5-fold greater coverage at PPV of 0.7 than that of the Single-Profile method. ROC curves for the methods, which plot sensitivity against (1 − specificity), also indicate that the Coevolutionary-Matrix performs better than the Single-Profile, although the difference is rather small (figure not shown). Sensitivity is defined as TP/(TP + FN), and Specificity is TN/(FP + TN), where TN and FN are true negatives and false negatives, respectively. Both TN and FN are numbers of protein pairs that do not meet the threshold. In addition, proteins in each pair are represented in the KEGG dataset. Their difference is that TN protein pairs are not interacting in the KEGG dataset while FN pairs are.

To have a closer look at the performances of the two methods, we show PPV, specificity and sensitivity for the three different sets of predicted pairs by each method in Table 1. At same number of predicted pairs, the Coevolutionary-Matrix method is again

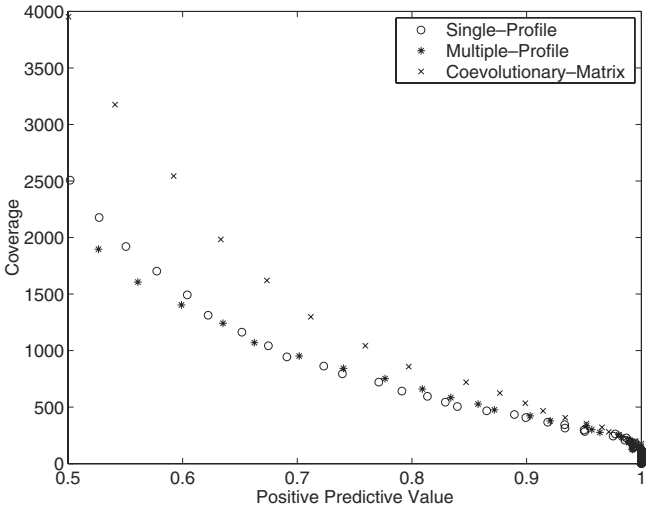

Fig. 5. PPV versus coverage plots for the Single-Profile, Multiple-Profile, and Coevolutionary-Matrix methods using the KEGG interaction dataset. As mutual information score threshold is varied, coverage and PPV at that threshold are plotted. Each dot represents such pair.

Table 1. Number of predicted interactions at various mutual information score thresholds

MIS	No. of PP	Covg.	TP	PPV	Specificity	Sensitivity
Single-Profile						
0.68702	15340	1239	789	0.637	0.99942	0.01821
0.72673	8549	855	620	0.725	0.99970	0.01431
0.76461	4971	617	499	0.809	0.99985	0.01152
Coevolutionary-Matrix						
0.60500	15339	1620	1091	0.673	0.99932	0.02518
0.64350	8548	1043	792	0.759	0.99968	0.01828
0.68200	4970	720	610	0.847	0.99986	0.01408

MIS, mutual information score threshold; no. of PP, number of predicted protein pairs; Covg., coverage; TP, true positives; PPV, positive predictive value.

shown to perform better than the Single-Profile both in terms of PPV and sensitivity. In Table 2, we also show PPV for both overlapping and non-overlapping areas between the sets of predicted pairs in Table 1. In section A in Table 2, percent overlap between the two sets is 55% and the PPV for this overlap is 0.75, which is higher than any one of the methods alone. Furthermore, PPV of the Coevolutionary-Matrix method alone is higher than that of the Single-Profile method alone (0.57 versus 0.33). Similar observations are made for the sections B and C in Table 2. These results indicate that the Coevolutionary-Matrix method predicts a significantly different set of interactions from those of the Single-Profile with greater number of TP.

To determine which KEGG pathways are most represented in top-scoring protein pairs using each method, we took top 2000 pairs out of all non-homologous pairs for the *E.coli* proteome and counted the number of those that belong to each pathway (data not shown). Phylogenetic similarity score thresholds used to generate these sets are 0.835 ($p < 2.2 \times 10^{-4}$) for the Single-Profile and 0.741 ($p < 2.2 \times 10^{-4}$) for the Coevolutionary-Matrix. These thresholds are considered to be strict and hence should yield high-confidence predictions.

Top-scoring protein pairs predicted by both methods are from a wide range of pathways, whose rankings based on the number of protein pairs falling into each are similar (36 pathways for the Coevolutionary-Matrix and 29 for the Single-Profile method out of a total of 131 KEGG pathways). Some of the highly populated pathways shared by both methods include flagellar assembly, phosphotransferase system, ABC transporters, oxidative phosphorylation, ubiquinone biosynthesis and histidine metabolism. In the set of 2000 pairs for the Single-Profile method, there are 339 pairs that share at least a KEGG pathway. For the Coevolutionary-Matrix method, there are 391 pairs with 281 of them overlapping with those of the Single-Profile method. Despite this relatively high number of shared pairs, the overall percent overlap between the two sets of 2000 pairs is 61.5%.

In Figure 6, we show mutual information score distributions of sets of interacting and random protein pairs calculated with each method. At zero mutual information score, the Single-Profile method shows a peak for the distribution of interacting protein pairs while the Coevolutionary-Matrix method does not have this peak. Based on this observation, it appears that a significant portion of the interacting protein pairs with very low mutual information scores under the Single-Profile gained higher (and potentially meaningful) scores under the Coevolutionary-Matrix. To investigate this further, Figure 7 shows how *p*-values of mutual information scores of the interacting protein pairs are correlated between the two methods. Although there is a rough correlation, there are many outliers. Some of these have *p*-value differences as high as four orders of magnitude. The presence of these outliers indicate that the two methods can make very different predictions for some proteins.

3.2 Examples of domain co-evolution

In this section, we show three examples of domain-level co-evolution from the *E.coli* cellular systems. We then hypothesize how co-evolved domains detected by the coevolutionary-matrix method fit with existing biochemical data. Finally, top interacting partners predicted by the two methods for these proteins are compared.

3.2.1 Phosphotransferase system The phosphotransferase system (PTS) is the major pathway through which translocation of sugars across the bacterial inner membrane is coupled with phosphorylation (Tchieu *et al*., 2001). The cytoplasmic protein IIAB transfers a phosphoryl group from the cytoplasmic proteins I and HPr to substrates through interactions with the membrane proteins IIC and IID in the case of mannose-specific PTS (Tchieu *et al*., 2001). The co-evolving region detected for the proteins IIAB and IIC is shown in Figure 8.

Figure 8 clearly captures two different regions of IIAB. In fact, these regions correspond to the domains IIA (residues 1–170) and IIB (residues 170–320). Figure 8 also captures the notion that the domain IIB co-evolved with IIC instead of the domain IIA. This can be explained in light of how the task assigned to IIAB as a whole is divided between IIA and IIB. Within the PTS, domain IIA has the role of receiving the phosphoryl group from proteins I and HPr and passing it to domain IIB. Domain IIB then passes the phosphoryl group to membrane proteins (in this case, IIC and IID) and then to sugars. Physical interaction between the domain IIB and protein IIC

Table 2. Number of true interactions in the overlapping and non-overlapping sets of predicted interactions between the Single-Profile (SP) and Coevolutionary-Matrix (CM) methods.

	A				B				C			
	No. of PP	Covg.	TP	PPV	No. of PP	Covg.	TP	PPV	No. of PP	Covg.	TP	PPV
SP Λ (∼CM)	6932	340	111	0.33	3731	233	94	0.40	2073	160	83	0.52
CM Λ (∼SP)	6931	721	413	0.57	3730	421	266	0.63	2072	263	194	0.74
SP Λ CM	8408	899	678	0.75	4818	622	526	0.85	2898	457	416	0.91

Symbols Λ and ∼ indicate 'logical AND' and 'logical negation,' respectively; no. of PP, number of predicted protein pairs; Covg, coverage; TP, true positives; PPV, positive predictive value.

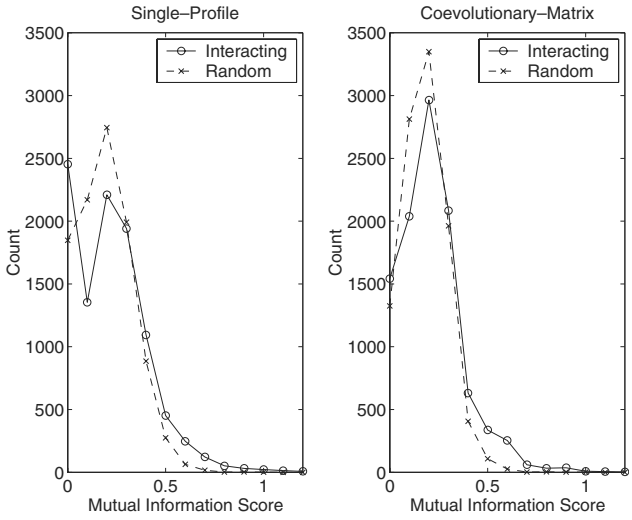

Fig. 6. Mutual information score distributions under the Single-Profile and Coevolutionary-Matrix methods. From the KEGG interaction set, 10 000 interacting proteins are randomly selected. Then another 10 000 protein pairs, without requiring them to be interacting, are randomly selected from the proteins present in the KEGG interaction set.

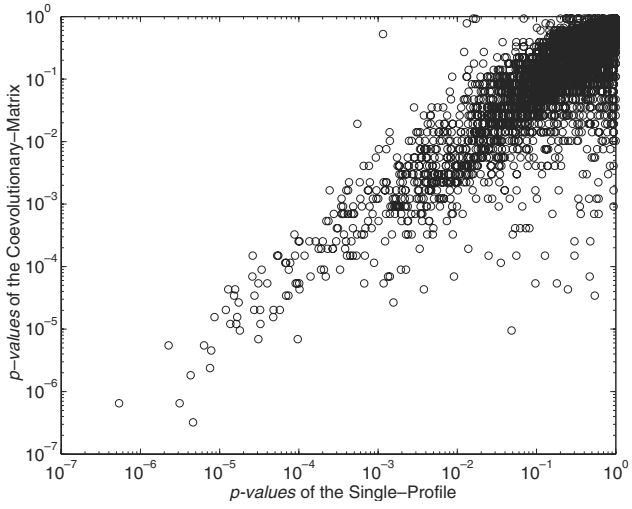

Fig. 7. *p*-values of mutual information scores of interacting proteins under the Single-Profile method versus those under the Coevolutionary-Matrix method. The same set of 10 000 interacting protein pairs used in Figure 6 is used here. Each circle represents two proteins that are known to interact.

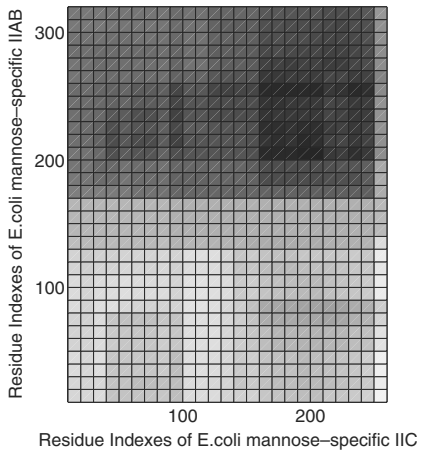

Fig. 8. Co-evolutionary matrix plot for the *E.coli* proteins mannose-specific IIAB and IIC (ManY). Darker color indicates higher mutual information score. The matrix shown here is the downsampled version of the original matrix. Each bin is 10 residues wide.

likely drives their co-evolution. This is shown for the mannitol-specific protein domains IIB and IIC, where domains IIA, IIB and IIC are fused to form a single protein (Robillard and Boos, 1999). Similar observation has been made by another group of researchers based on a different study (R.D.Barabote and M.H.Saier,Jr unpublished data).

In Table 3, top 20 predicted interacting partners of the protein IIAB under the Single-Profile and Coevolutionary-Matrix methods are shown. Although both methods pick out the proteins IIC (ManY) and IID (ManZ) as their top-scoring proteins, those under the Coevolutionary-Matrix show greater number of proteins that are involved in PTS. Four PTS proteins (ManY, ManZ, AgaC and AgaD) are found using the Single-Profile method and eight PTS proteins (ManY, ManZ, AgaD, AgaC, AgaW, CelC, CelB and CelA) are found with the Coevolutionary-Matrix method.

3.2.2 Chemotaxis Chemotaxis signaling pathway allows a bacterium to sense the state of its external environment and determine its swimming behavior accordingly. CheA is a multi-domain protein whose domains carry out different functions in this system (Falke *et al*., 1997). A plot of the co-evolutionary matrix for CheA and CheB, another chemotaxis component, is shown in Figure 9.

Figure 9 suggests that the N-terminus and C-terminus regions of CheA (residues 1–200 and 540–670, respectively) co-evolved with

Table 3. Top 20 interacting partners predicted by the Single-Profile and Coevolutionary-Matrix methods for the *E.coli* mannose-specific IIAB. Predicted interacting proteins are ranked based on their mutual information scores (MI) and shown here with their NCBI GenBank Identifiers (GI) and names. GI number of mannose-specific IIAB is 16129771. Each entry represents a single protein and more than one name is provided if available. Known PTS components are highlighted.

Rank	Single-Profile				Coevolutionary-Matrix			
	MI	pValue	GI	Name	MI	pValue	GI	Name
1	0.728	0.0009	16129772	ManY;PtsP	0.984	7e−06	16129772	ManY;PtsP
2	0.690	0.0016	33347589	ManZ;PtsM	0.928	2e−05	33347589	ManZ;PtsM
3	0.652	0.0028	33347619	GatR	0.909	3e−05	16131032	AgaD
4	0.642	0.0033	16131031	AgaC	0.891	3e−05	16131031	AgaC
5	0.623	0.0045	16131032	AgaD	0.703	0.0003	16131026	AgaW
6	0.608	0.0058	16131023	AgaR	0.591	0.0022	16129064	AscF
7	0.598	0.0069	16131615	Kup;TrkD	0.591	0.0022	16129690	CelC
8	0.594	0.0072	16129529	DicA	0.591	0.0022	16129691	CelB
9	0.588	0.0080	16129863	SdiA	0.591	0.0022	16132125	SgcC
10	0.582	0.0089	16129245	YciT	0.572	0.0022	16129090	PepT
11	0.581	0.0091	16129943	YeeS	0.553	0.0030	16132233	NirC
12	0.574	0.0101	16130367	IntZ	0.534	0.0056	16129692	CelA
13	0.569	0.0109	16131277	HslR	0.534	0.0056	16129951	DacD
14	0.565	0.0115	16131724	YihW	0.534	0.0056	16131590	BglF
15	0.565	0.0116	16130614	SrlR;GutR	0.516	0.0056	16130343	NagE;PstN
16	0.561	0.0124	16130605	OraA;RecX	0.516	0.0056	16131113	NanE
17	0.560	0.0126	16130559	YfjY	0.516	0.0056	16131297	GlpR
18	0.558	0.0130	16128921	FabA	0.516	0.0056	33347806	FrvB
19	0.555	0.0136	16129052	RpmF	0.516	0.0056	33347834	SgaT
20	0.555	0.0137	16131388	GadX	0.497	0.0103	16128871	FocA

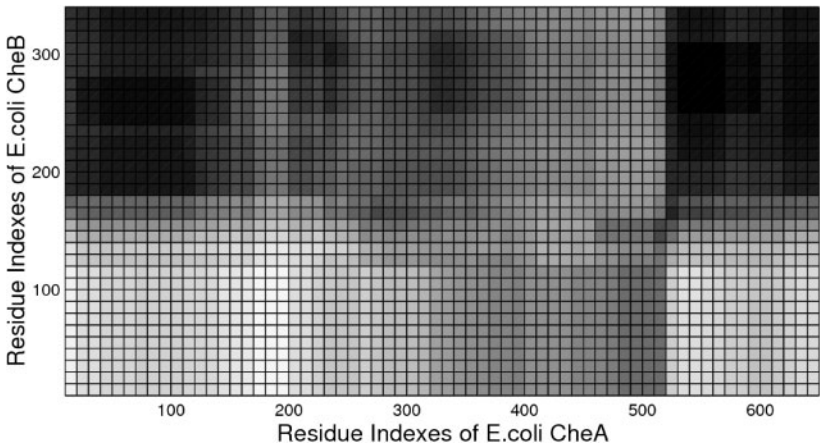

Fig. 9. Co-evolutionary matrix for the *E.coli* proteins CheA and CheB. Darker color indicates higher mutual information score. The matrix shown here is the downsampled version of the original matrix. Each bin is 10 residues wide.

the C-terminus region of CheB (residues 170–340). Although there is a biochemical evidence that CheB binds to the N-terminus region (Li *et al*., 1995), none exists for the binding of CheB to the C-terminus region. However, it is known that CheW, a chemotaxis component, binds to the C-terminus region (Gegner and Dahlquist, 1991). The sequence region from residues 200 to 350 of CheA, which shows weaker co-evolution, corresponds to the dimerization domain (Bilwes *et al*., 1999). Another region of CheA (residues 355–540) that does not seem to co-evolve with CheB corresponds to the kinase domain. The co-evolving regions of CheA identified from the matrix are essentially the same for Mcp's, CheW, CheR and CheB proteins.

In Table 4, top 20 predicted interacting partners of CheA using each method are shown. Under the Single-Profile method, only Mcp3 is known to participate in chemotaxis. In contrast, Mcp3, Mcp2, CheW, Mcp4, CheR and CheB are known to participate in chemotaxis under the Coevolutionary-Matrix method (Falke *et al*., 1997). In the same list, Aer is involved in aerotaxis; and FlgC, MotB, MotA, FlgF and FlgL are likely picked because the chemotaxis signaling pathway is coupled to the flagellar motor

Table 4. Top 20 interacting partners predicted by the Single-Profile and Coevolutionary-Matrix methods for the *E.coli* CheA

Rank	Single-Profile				Coevolutionary-Matrix			
	MI	*p*-value	GI	Name	MI	*p*-value	GI	Name
1	0.897	0.000097	16129380	Mcp3	0.984	6.9e−06	16129380	Mcp3
2	0.871	0.000136	16129902	YedQ	0.984	6.9e−06	16130967	Aer
3	0.844	0.000192	16129302	YdaM	0.966	1.21e−05	16129838	Mcp2
4	0.830	0.000228	16129338	Dos	0.966	1.21e−05	16129839	CheW
5	0.778	0.000451	16128300	YahA	0.947	1.21e−05	16129837	Mcp4
6	0.777	0.000456	16131386	YhiV	0.741	0.000193	16129836	CheR
7	0.770	0.000497	16128802	YliF	0.703	0.000325	16129835	CheB
8	0.769	0.000504	16130007	YegE	0.684	0.000536	33347519	YcfQ
9	0.765	0.000528	33347585	YeaP	0.684	0.000536	16129449	YddV
10	0.757	0.000591	16130498	RpoE	0.666	0.000536	16129037	FlgC
11	0.757	0.000594	16129449	YddV	0.666	0.000536	16129302	YdaM
12	0.756	0.000599	33347519	YcfQ	0.666	0.000536	16130498	RpoE
13	0.744	0.000716	16131140	YhdA	0.647	0.000696	16129448	Dos
14	0.742	0.000737	16129092	PhoQ	0.628	0.001225	16129841	MotB
15	0.740	0.000766	16130525	YfiN	0.628	0.001225	16129842	MotA
16	0.733	0.000841	16129185	NarX;NarR	0.609	0.001225	16128300	YahA
17	0.710	0.001163	16131154	AcrF	0.609	0.001225	16128731	ModA
18	0.709	0.001177	16128370	YaiC	0.609	0.001225	16129040	FlgF
19	0.709	0.001181	16131869	LexA;ExrA	0.609	0.001225	16129046	FlgL
20	0.702	0.001312	16129131	YcgG	0.609	0.001225	16129566	RstA;UrpT

Predicted interacting proteins are ranked based on their mutual information scores (MI) and shown here with their NCBI GenBank Identifiers (GI) and names. GI number of CheA is 16129840. Each entry represents a single protein and more than one name is provided if available. Known chemotaxis components are highlighted.

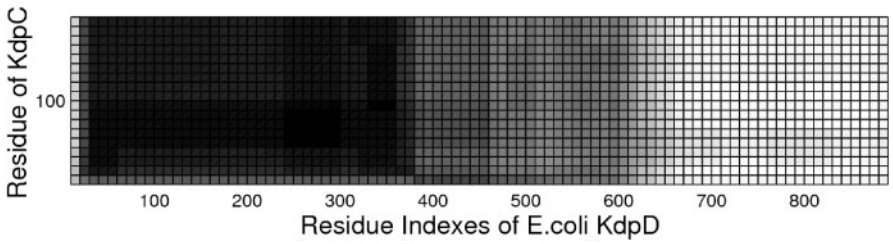

Fig. 10. Co-evolutionary matrix for the *E.coli* proteins KdpD and KdpC. Darker color indicates higher mutual information score. The matrix shown here is the downsampled version of the original matrix. Each bin is 10 residues wide.

system. As a note, Mcp2, Mcp3, Mcp4 and Aer are homologous (BLAST *E*-value < 1.0).

3.2.3 Kdp system In *E.coli*, KdpD and KdpE regulate expression of the kdpFABC operon, which encodes a high affinity K^+ transport ATPase (Walderhaug *et al*., 1992). KdpD is a multi-domain protein which consists of an N-terminal cytoplasmic domain (residues 1–395), four transmembrane domains and a cytoplasmic C-terminal transmitter domain (Heermann *et al*., 2003). The Coevolutionary-Matrix method clearly delineates three corresponding domains in Figure 10.

Figure 10 suggests that the N-terminal domain of KdpD co-evolved with KdpC. Supporting this hypothesis, a recent study has shown that this N-terminal domain alone triggers semi-constitutive expression of the kdpFABC operon through interactions with KdpE (Heermann *et al*., 2003). Interaction between KdpD and KdpC is therefore of functional dependence rather than physical. Top 10 interacting partners predicted by the Single-Profile include only KdpE from this system while those of the Coevolutionary-Matrix include KdpE, KdpA and KdpC. Mutual information scores of KdpC and KdpA with respect to KdpD using the Single-Profile method are 0.1829 and 0.2496, respectively. These very low mutual information scores suggest that the Single-Profile method cannot detect co-evolution between KdpC/KdpA and KdpD.

4 DISCUSSION

The results shown in this paper strongly suggest that co-evolution of proteins should be captured at the domain level. As indicated by the co-evolutionary matrices shown in Figures 8, 9 and 10, sequence regions with conflicting evolutionary histories can co-exist within a single protein. By representing protein co-evolution at the domain level, the Coevolutionary-Matrix method can assign very different phylogenetic similarity scores to proteins when compared with the Single-Profile method (Fig. 7). In turn, these differences have substantial effect on the performances of the two methods (Tables 3, 4 and 5).

Table 5. Top 10 interacting partners predicted by the Single-Profile and Coevolutionary-Matrix methods for the *E.coli* KdpD

Rank	Single-Profile				Coevolutionary-Matrix			
	MI	*p*-value	GI	Name	MI	*p*-value	GI	Name
1	0.62955	0.004117	16129184	NarL	0.815625	8.93e−05	16128672	KdpC
2	0.622577	0.004604	16128961	TorR	0.815625	8.93e−05	16128674	KdpA
3	0.621782	0.004664	16128670	KdpE	0.590625	0.002223	16128670	KdpE
4	0.616888	0.005055	16129861	UvrY	0.571875	0.002223	16130019	BaeR
5	0.605839	0.006033	16129196	Hnr	0.553125	0.003018	16128961	TorR
6	0.605499	0.006065	16131708	GlnG	0.553125	0.003018	16129861	UvrY
7	0.605108	0.006102	16130921	QseB	0.553125	0.003018	16130313	YpdB
8	0.598099	0.006833	16131539	UhpA	0.553125	0.003018	16131282	OmpR
9	0.596262	0.007033	16132215	CreB	0.553125	0.003018	16131939	BasR
10	0.595111	0.007166	16130479	YfhA	0.534375	0.005556	16128603	CitB

Predicted interacting proteins are ranked based on their mutual information scores (MI) and shown here with their NCBI GenBank Identifiers (GI) and names. GI number of KdpD is 16128671. Each entry represents a single protein. Known Kdp system components are highlighted.

Others have also noted the importance of including domain information when predicting protein interactions. By incorporating interaction profile of domains in their method, Wojcik and Schachter (2001) reported increased performance in inferring protein interaction of one organism from the interaction network of another. Similar in spirit to our approach, Pagel *et al*. (2004) improved upon the phylogenetic profiling method using domains defined with the Pfam database (Bateman *et al*., 2004). Although the coverage of their method is limited by that of the Pfam database, it has the advantage of requiring less computing time and having a simple update procedure as more genomes are used.

Interestingly, similar to databases such as Pfam (Bateman *et al*., 2004), the Coevolutionary-Matrix method can delineate 'domains' within a protein. Because of the way parameters were chosen, the co-evolving regions detected with our method are required to have sizes of at least 60 residues. The size requirement ensures that it is in the range of independently folding protein domains, excluding those of loops (i.e. less than 20 residues).

Motivated by the performance of the Coevolutionary-Matrix method, we explored the idea of whether co-evolving domains captured indeed are involved in interactions at the domain level. For the PTS proteins IIAB and IIC, physical interaction between the domain IIB and protein IIC seems plausible based on available evidence. However, for some proteins such as those involved in the chemotaxis pathway, it appears that much of the co-evolution between the domains was driven by their functional dependence. For example, the Coevolutionary-Matrix method identified that the N-terminus and C-terminus regions of CheA co-evolved with CheB. The method indicates that these same regions also co-evolved with Mcp's, CheW and CheR proteins. Most likely all these proteins do not physically interact with the same two regions of CheA. Likewise, the N-terminus domain of KdpD in the Kdp system does not physically interact with KdpC or KdpA but is needed to drive the expression of the latter two proteins.

5 CONCLUDING REMARKS

Since evolution and functions of proteins are coupled, greater understanding of the former can reveal much about the latter. By capturing co-evolution of proteins at the domain level, regions that are important for supporting both functional and physical interactions between these proteins are detected. With examples from the cellular systems of the *E.coli* bacterium, we showed that these regions correspond to biochemically characterized protein domains.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support by NIH grant GM068959. Funding to pay the Open Access publication charges for this article was provided by National Institutes of Health/Institute of General Medical Sciences, Grant No. RO1-GM068959.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Bateman,A. *et al*. (2004) The Pfam Protein Families Database. *Nucleic Acids. Res*., **32**, D138–D141.

Bilwes,A.M. *et al*. (1999) Structure of CheA, a signal-transducing histidine kinase. *Cell*, **96**, 131–141.

Dandekar,T. *et al*. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci*., **23**, 324–328.

Date,S.V. and Marcotte,E.M. (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol*., **21**, 1055–1062.

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

Falke,J.J. *et al*. (1997) The two-component signaling pathway of bacterial chemotaxis: a molecular view of signal transduction by receptors, kinases, and adaptation enzymes. *Ann. Rev. Cell Dev. Biol*., **13**, 457–512.

Gavin,A.C. *et al*. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Gegner,J.A. and Dahlquist,F.W. (1991) Signal transduction in bacteria: CheW forms a reversible complex with the protein kinase CheA. *Proc. Natl Acad. Sci. USA*, **88**, 750–754.

Giot,L. *et al*. (2003) A Protein Interaction Map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.

Heermann,R. *et al*. (2003) The N-terminal input domain of the sensor kinase KdpD of *Escherichia coli* stabilizes the interaction between the cognate response regulator KdpE and the corresponding DNA-binding Site. *J. Biol. Chem*., **278**, 51277–51284.

Ho *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

Ito,T. *et al*. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Kanehisa,M. *et al*. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–280.

Kim,Y. and Subramaniam,S. (2005) Locally defined protein phylogenetic profiles reveal previously missed functional relationships. *Proteins*, (in press)

Lee,I. *et al*. (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.

Li,J. *et al*. (1995) The response regulators CheB and CheY exhibit competitive binding to the kinase CheA. *Biochemistry*, **34**, 14626–14636.

Overbeek,R. *et al*. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.

Pagel,P. *et al*. (2004) A domain interaction map based on phylogenetic profiling. *J. Mol. Biol*., **344**, 1331–1346.

Pellegrini,M. *et al*. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.

Robillard,G.T. and Broos,J. (1999) Structure/function studies on the bacterial carbohydrate transporters, enzymes II, of the phosphoenolpyruvate-dependent phosphotransferase system. *Biochim. Biophys. Acta*., **1422**, 73–104.

Tchieu,J.H. *et al*. (2001) The complete phosphotransferase system in *Escherichia coli*. *J. Mol. Microbiol. Biotechnol*., **3**, 329–346.

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

von Mering,C. *et al*. (2002) Comparative assessment of large-scale datasets of protein–protein interactions. *Nature*, **417**, 365–470.

von Mering,C. *et al*. (2003) Genome evolution reveals biochemical networks and functional modules. *Proc. Natl Acad. Sci. USA*, **100**, 15428–15433.

Walderhaug,M.O. *et al*. (1992) KdpD and KdpE, proteins that control expression of the kdpABC operon, are members of the two-component sensor-effector of regulators. *J. Bacteriol*., **174**, 2152–2159.

Wojcik,J. and Schachter,V. (2001) Protein–protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, **17**, S296–S305.