doi:10.1016/S0022-2836(03)00114-1 *J. Mol. Biol.* (2003) **327**, 273–284

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Exploiting the Co-evolution of Interacting Proteins to Discover Interaction Specificity

Arun K. Ramani[1] and Edward M. Marcotte[1,2*]

[1]*Institute for Cellular and Molecular Biology Center for Computational Biology and Bioinformatics University of Texas at Austin Austin, TX 78712, USA*

[2]*Department of Chemistry and Biochemistry University of Texas at Austin Austin, TX 78712, USA*

**Corresponding author*

Protein interactions are fundamental to the functioning of cells, and high throughput experimental and computational strategies are sought to map interactions. Predicting interaction specificity, such as matching members of a ligand family to specific members of a receptor family, is largely an unsolved problem. Here we show that by using evolutionary relationships within such families, it is possible to predict their physical interaction specificities. We introduce the computational method of matrix alignment for finding the optimal alignment between protein family similarity matrices. A second method, 3D embedding, allows visualization of interacting partners *via* spatial representation of the protein families. These methods essentially align phylogenetic trees of interacting protein families to define specific interaction partners. Prediction accuracy depends strongly on phylogenetic tree complexity, as measured with information theoretic methods. These results, along with simulations of protein evolution, suggest a model for the evolution of interacting protein families in which interaction partners are duplicated in coupled processes. Using these methods, it is possible to successfully find protein interaction specificities, as demonstrated for >18 protein families.

Keywords: protein interactions; bioinformatics; phylogeny; co-evolution; interaction specificity

Introduction

Protein interaction specificity is vital to cell function, but the maintenance of such specificity requires that it persist even through the course of strong evolutionary change, such as the duplication and divergence of genes. Binding specificities of duplicate genes (paralogs) often diverge, such that new binding specificities are evolved. Given that such paralogous gene families abound, such as the >560 serine-threonine kinases in the human genome,[1] predicting interaction specificity can be difficult, especially when paralogs exist for both interaction partners. In these cases, the number of potential interactions grows combinatorially. This ambiguity can easily complicate the matching of ligands to specific receptors, and for such reasons, identification of ligands for orphan receptors is an important, but largely unsolved, problem.[2–4]

Computational methods for discovering specific protein interactions fall into three broad categories: (i) the identification of specific protein sequence or structural features indicative of protein interaction partners, such as sequence signatures,[5] correlated mutations,[6,7] and surface patches;[8,9] (ii) the use of genomic context[10] to identify interaction partners, exploiting information such as gene order,[11,12] gene fusions,[13,14] and phylogenetic profiles;[15] and (iii) the use of phylogenetic trees to account for the co-evolution of interacting proteins.[16–20]

Of these three classes, the third is of specific interest: the hypothesis underlying these approaches is that interacting proteins often exhibit coordinated evolution, and therefore tend to have similar phylogenetic trees. Goh *et al.*[17] demonstrated this by showing that chemokines and their receptors have very similar phylogenetic trees, as do individual domains of a single protein such as phosphoglycerate kinase. Detailed phylogenetic studies of the two-component signal transduction system[18] show that a phylogenetic tree constructed from two-component sensor proteins has a similar structure to that from two-component regulator proteins.

Here, we exploit this tendency for interacting proteins to have similar phylogenetic trees, and present a general computational method for the identification of specific interaction partners in

E-mail address of the corresponding author: marcotte@icmb.utexas.edu

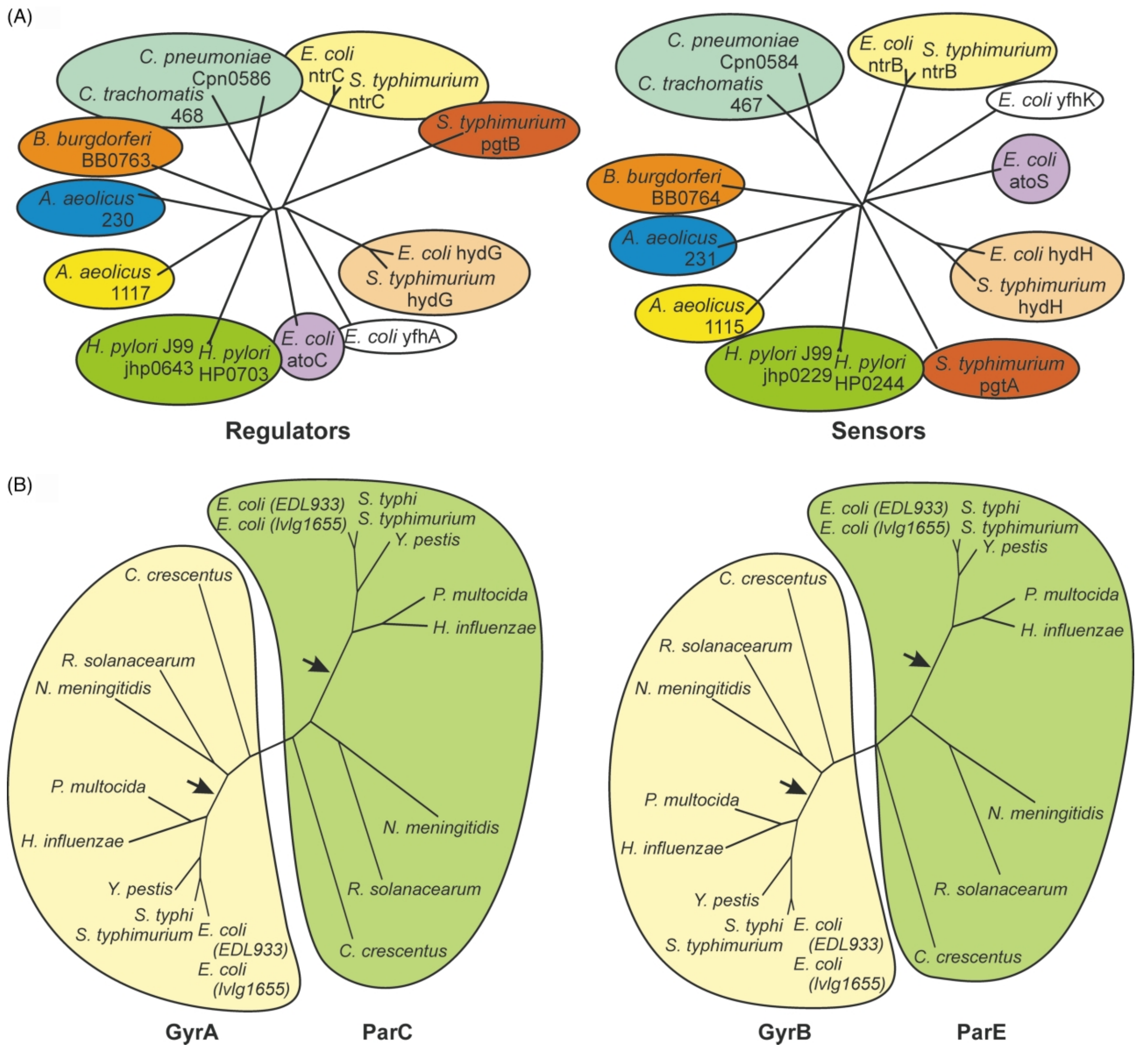

Figure 1. (A) A comparison of the phylogenetic trees of Ntr-family two-component sensor histidine kinases and their corresponding regulators. Circles enclose orthologous genes. Interacting proteins, colored similarly, sit in similar positions in the two trees. (B) A comparison of the phylogenetic tree of the GyrA and ParC proteins with the tree of their corresponding interaction partners, GyrB and ParE, colored as in (A). Bold arrows indicate an example of differing branch lengths, which help to distinguish the Gyr and Par subtrees.

such protein families. We provide an information-theoretic interpretation of when the method is appropriate, and present a model that emerges for the evolution of interacting proteins.

Results

Prediction of interactions by matrix alignment

Figure 1(A) presents the phylogenetic trees of two families of interacting proteins, the Ntr-type two-component sensors and their corresponding regulators. There is striking similarity in the relative placement of interacting protein pairs across the two trees: The ntrC proteins from *Escherichia coli* and *Salmonella typhimurium* are adjacent in the regulator tree, as are their interaction partners (ntrB) in the sensor tree. Likewise, the ntrC proteins are roughly equidistant in the regulator tree from the hydG regulator proteins; this relationship is maintained by their interacting partners in the sensor tree. Many details of the overall tree structure are shared between the ligand and receptor tree, as noted previously for two-component sensor/regulators[18] and for chemokines/chemokine receptors.[17]

Figure 1(B) presents the simplest such case of interaction partners, in which each interacting protein (e.g. GyrA and GyrB) has a single paralog (e.g. ParC and ParE, respectively, which interact specifically with each other). Again, the trees of the interacting partners are notably similar. In fact, even the halves of the trees specific to each paralog

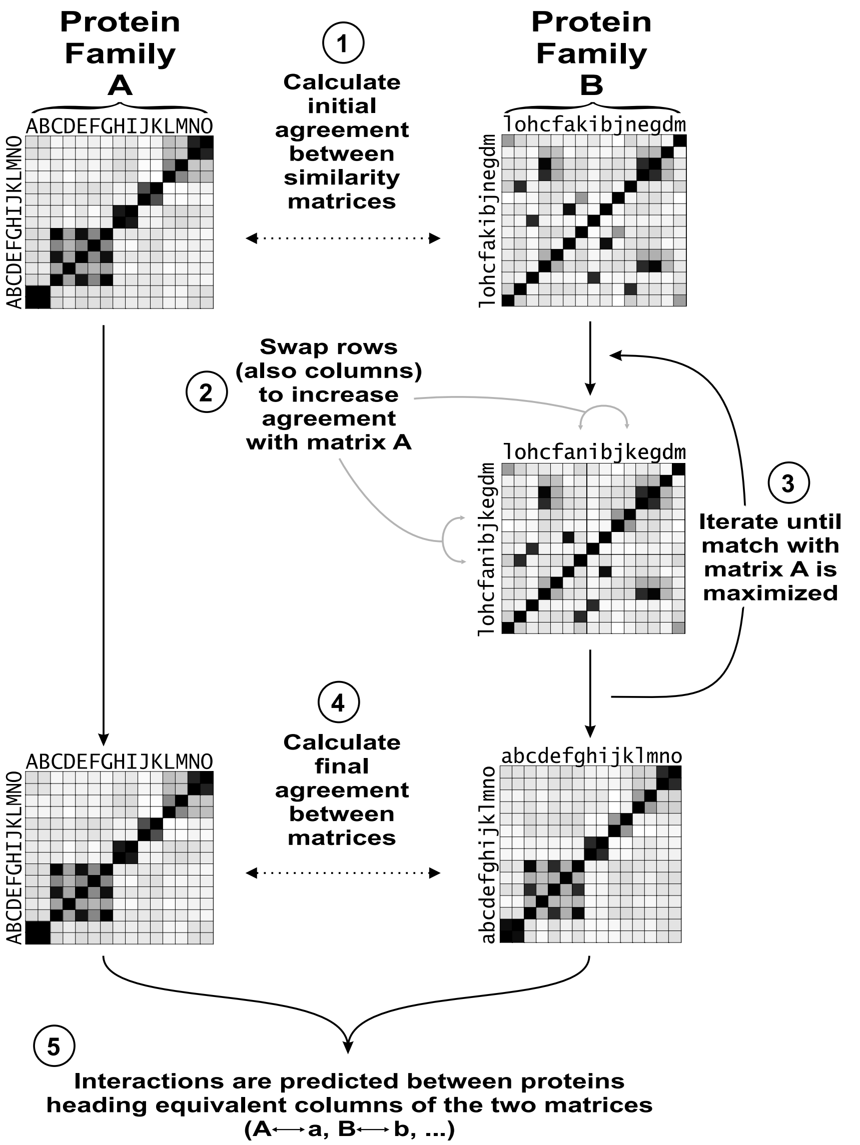
Protein Family A
Protein Family B
1
Calculate initial agreement between similarity matrices
ABCDEFGHIJKLMNO
lohcfakibjnegdm
2
Swap rows (also columns) to increase agreement with matrix A
lohcfanibjkegdm
3
Iterate until match with matrix A is maximized
4
Calculate final agreement between matrices
abcdefghijklmno
5
Interactions are predicted between proteins heading equivalent columns of the two matrices (A⟷a, B⟷b, ...)

Figure 2. The matrix alignment method for predicting protein interaction specificity. Proteins in family A interact with those in family B. In each family, a similarity matrix summarizes the proteins' evolutionary relationships. The algorithm uses the similarity matrices to pair up the genes in the two families. Columns of matrix B are re-ordered (along with their corresponding rows in the matrix) such that the B matrix agrees maximally with matrix A, judged by minimizing the root mean square difference (r.m.s.d.) between elements in the two matrices. Interactions are then predicted between proteins heading equivalent columns of the two matrices.

are similar, as the GyrA half strongly resembles both the GyrB and ParE halves. However, a careful examination of branch lengths indicates subtle differences between the halves, such as is indicated by the arrows in Figure 1(B), such that the correct interaction partners (GyrA with GyrB, and ParC with ParE) have the most similar subtrees.

In order to exploit the evolutionary information contained in such interacting protein families, we developed an algorithm that is conceptually equivalent to superimposing the phylogenetic trees of the two protein families. This approach, which we term matrix alignment and which is implemented in the program MATRIX, is diagrammed schematically in Figure 2.

Rather than directly compare the phylogenetic trees, the corresponding similarity matrices are compared to each other, each matrix summarizing the evolutionary relationships between the proteins within one sequence family. One matrix is shuffled, maintaining the correct relationships between proteins but simply re-ordering them in the matrix, until the two matrices maximally agree, minimizing the root mean square difference between elements of the two matrices. Interactions are then predicted between proteins heading equivalent columns of the two matrices. For matrix alignment, MATRIX currently applies a stochastic simulated annealing-based algorithm.

Matching two-component sensors to regulators

As a first test of matrix alignment, we examined the Ntr-type two-component sensor and regulator families of Figure 1. Binding partners were assigned according to the KEGG pathway database[21] resulting in a set of 14 interactions, spanning genes from eight organisms. Matrix alignment was performed, testing specifically whether or not the genes from one genome (for example, the four *E. coli* regulators) could be matched to their correct binding partners (here, the four *E. coli* sensor proteins).

Table 1. The prediction of protein interactions between interacting protein families by the method of matrix alignment

Ntr-type Regulators

Ntr-type Sensors	*E. coli* ntrC	*S. typhimurium* ntrC	*E. coli* hydG	*S. typhimurium* hydG	*E. coli* atoC	*E. coli* yfhA	*S. typhimurium* pgtA	*H. pylori* 26695 HP0703	*H. pylori* J99 jhp064	*C. trachomatis* CT468	*C. pneumoniae* CPn058	*B. burgdorferi* BB0763	*A. aeolicus* aq_111	*A. aeolicus* aq_230
E. coli ntrB	0.23	0.57	0.04	0.15					0.01					
S. typhimurium ntrB	0.57	0.23	0.15	0.04				0.01						
E. coli hydH	0.07	0.12	0.59	0.22										
S. typhimurium hydH	0.12	0.07	0.22	0.59										
E. coli atoS					0.95	0.02						0.02	0.01	
E. coli yfhK						0.84	0.03					0.12	0.01	
S. typhimurium pgtB						0.02	0.97					0.01		
H. pylori 26695 HP0244		0.01						0.6	0.39					
H. pylori J99 jhp022	0.01							0.39	0.6					
C. trachomatis CT467										0.2	0.8			
C. pneumoniae CPn058										0.8	0.2			
B. burgdorferi BB0764					0.04	0.03						0.73	0.13	0.07
A. aeolicus aq_111						0.06						0.06	0.1	0.78
A. aeolicus aq_231					0.01	0.03						0.06	0.75	0.15

CKR Chemokine Receptors

CKR Chemokines	Human CKR1	Mouse CKR1	Human CKR2	Mouse CKR2	Rat CKR2	Human CKR3	Mouse CKR3	Rat CKR3	Human CKR4	Mouse CKR4	Human CKR6	Mouse CKR6	Human CKR7	Mouse CKR7	Human CKR8	Mouse CKR8	Human CKR9	Mouse CKR9
Human SY03	0.22	0.46	0.04	0.1	0.1		0.02	0.02	0.02	0.04								
Mouse SY03	0.46	0.22		0.1	0.1	0.04	0.02	0.02	0.04	0.02								
Human SY02	0.02	0.06	0.48			0.44	0.02											
Mouse SY02	0.1	0.1		0.14	0.2		0.26	0.22										
Rat SY02	0.1	0.1		0.2	0.14		0.22	0.26										
Human EOTA	0.06	0.02	0.44			0.48		0.02										
Mouse EOTA	0.02	0.04		0.3	0.18		0.18	0.3										
Rat EOTA	0.04	0.02		0.18	0.3		0.3	0.18										
Human SY22									0.48	0.2		0.02	0.06		0.1	0.12	0.04	
Mouse SY22									0.2	0.48	0.02			0.06	0.12	0.1		0.04
Human SY20											0.02	0.08	0.3	0.28			0.1	0.24
Mouse SY20											0.08	0.02	0.28	0.3			0.24	0.1
Human SY21									0.02	0.06	0.06	0.06	0.2	0.16	0.04		0.28	0.14
Mouse SY21									0.06	0.02	0.06	0.06	0.16	0.2		0.04	0.14	0.28
Human SY01						0.06			0.06	0.14					0.46	0.3		
Mouse SY01			0.06						0.14	0.06					0.3	0.46		
Human SY25											0.48	0.3	0.02				0.18	0.04
Mouse SY25											0.3	0.48		0.02			0.04	0.18

The top table indicates the predicted interactions between Ntr-type two-component sensors and regulators, and the bottom table indicates the predicted interactions between CKR-type chemokines and chemokine receptors. The diagonal of each matrix represents the correct known interacting pairs based on the assignments of the KEGG database (top) or measured binding affinities (bottom). Each Table entry represents the fraction of matrix alignment runs in which a given interaction was predicted. Filled boxes represent the predicted interaction partners observed in the highest fraction of the runs, while broken line boxes represent the interaction partners predicted when allowing interactions between orthologs. There is an ambiguity in the interaction partners of the chemokine/chemokine receptors, indicated by bold broken boxes, leading to either two correct or two incorrect predictions.

The results following 100 runs of simulated annealing are presented in Table 1 (and later summarized in Figure 4(A)). Diagonal entries in the table correspond to the correct binding partners, and the values reported in each table cell indicate the fraction of simulated annealing runs in which the corresponding proteins were predicted to be binding partners. For example *E. coli* atoS is paired correctly with *E. coli* atoC 95% of the time (in 95 of the 100 runs); as this match outscores any other match to atoS or atoC, these are predicted to be interaction partners. In a typical run, the starting

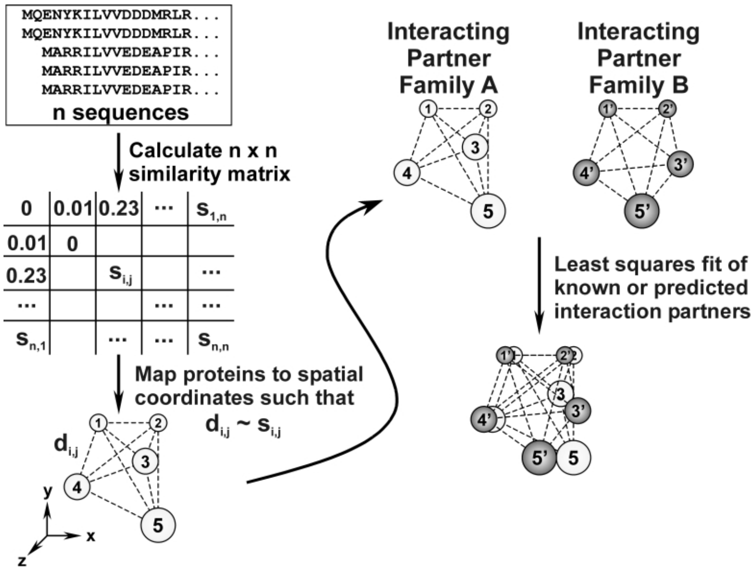
MQENYKILVVDDDMRLR...
MQENYKILVVDDDMRLR...
MARRILVVEDEAPIR...
MARRILVVEDEAPIR...
MARRILVVEDEAPIR...
n sequences
Calculate n x n similarity matrix
0
0.01
0.23
...
$s_{1,n}$
$s_{i,j}$
$s_{n,1}$
$s_{n,n}$
Map proteins to spatial coordinates such that
$d_{i,j} \sim s_{i,j}$
$d_{i,j}$
y
x
z
Interacting Partner Family A
Interacting Partner Family B
Least squares fit of known or predicted interaction partners

Figure 3. To visualize protein families, proteins are plotted in 3D space such that each protein is separated from other proteins in its family by distances d_{ij} proportional to the evolutionary similarities s_{ij} in the family's similarity matrix. To visualize interactions between two protein families (labeled A and B), the families are superimposed by rigid-body least-squares fit of the predicted interaction partners onto each other.

r.m.s.d. between the sensor and regulator similarity matrices was ~0.242; following application of the algorithm, it was ~0.207. For comparison, the correct pairing corresponded to an r.m.s.d. of 0.181, indicating that the algorithm typically found a solution that efficiently minimized the r.m.s.d. but still did not find the global optimum from among the 14!, or $\sim 10^{11}$, possible solutions.

To assess the accuracy of the interaction prediction, two values were examined: the stringent accuracy, defined as the accuracy of exact matches of known binding partners, and the effective accuracy, which was evaluated by accepting matches to orthologous protein family members (such as correctly matching ntrB to ntrC, but with the match occurring between the *E. coli* protein and the *S. typhimurium* protein, rather than *E. coli* with *E. coli*.) Because the species is known in every case, we can typically increase the accuracy by considering the orthologs. For the Ntr-type two-component regulator/sensor case, the stringent accuracy was 57% while the effective accuracy was 86%. All four *E. coli* proteins were correctly matched to their interaction partners, as were the *S. typhimurium* proteins. Thus, inherent information exists in the phylogenetic trees of the two families that can be automatically extracted to predict protein interaction partners.

Visualization of protein interaction partners by 3D embedding

In order to summarize in a clear manner the many evolutionary relationships and interactions, we developed a method, termed 3D embedding and diagrammed in Figure 3, for effectively visualizing the aligned similarity matrices and predicted protein interaction partners: coordinates in three-dimensional space are assigned to proteins in a sequence family such that the spatial separation of the proteins is proportional to the evolutionary distances between the proteins described in the similarity matrix. Protein interaction partners can then be visualized by assigning coordinates to each protein in the two protein families that interact with each other, followed by superposition of one family onto the other by least squares minimization of the distance between interacting partners. During this superposition, the relative distances between the proteins of a sequence family are unchanged. Instead, only the orientation of the resulting "constellation" of proteins in one family is changed relative to the proteins of the other family, as shown in Figure 3.

Figure 4(A) shows the application of 3D embedding to the Ntr regulator/sensor proteins. In this example, the proteins are aligned such that the distances between the predicted interaction partners are minimized. As can be seen in the Figure, proteins cluster in distinct regions in space, mirroring the adjacent placement of orthologs in the phylogenetic trees of Figure 1. Interacting protein partners generally sit close to each other in space. Orthologs appear to exhibit little apparent preference for their precise positions within a particular spatial cluster, consistent with the tendency of the matrix alignment algorithm to assign interactions to orthologous protein sequences rather than the sequences of the correct species. From Figure 4(A), it is obvious that matrix alignment succeeds in finding quite complex relationships that successfully satisfy the many constraints, such as matching yfhA to yfhK, rather than the potentially closer hydH, in order that both *S. typhimurium* and *E. coli* hydH interactions could be predicted.

Figure 4(B) shows the application of 3D embedding to the simpler problem of matching interaction partners given the right pair and a homologous pair as competition. The solution demonstrates the extreme robustness of matrix alignment for such simple cases. Here, interactions

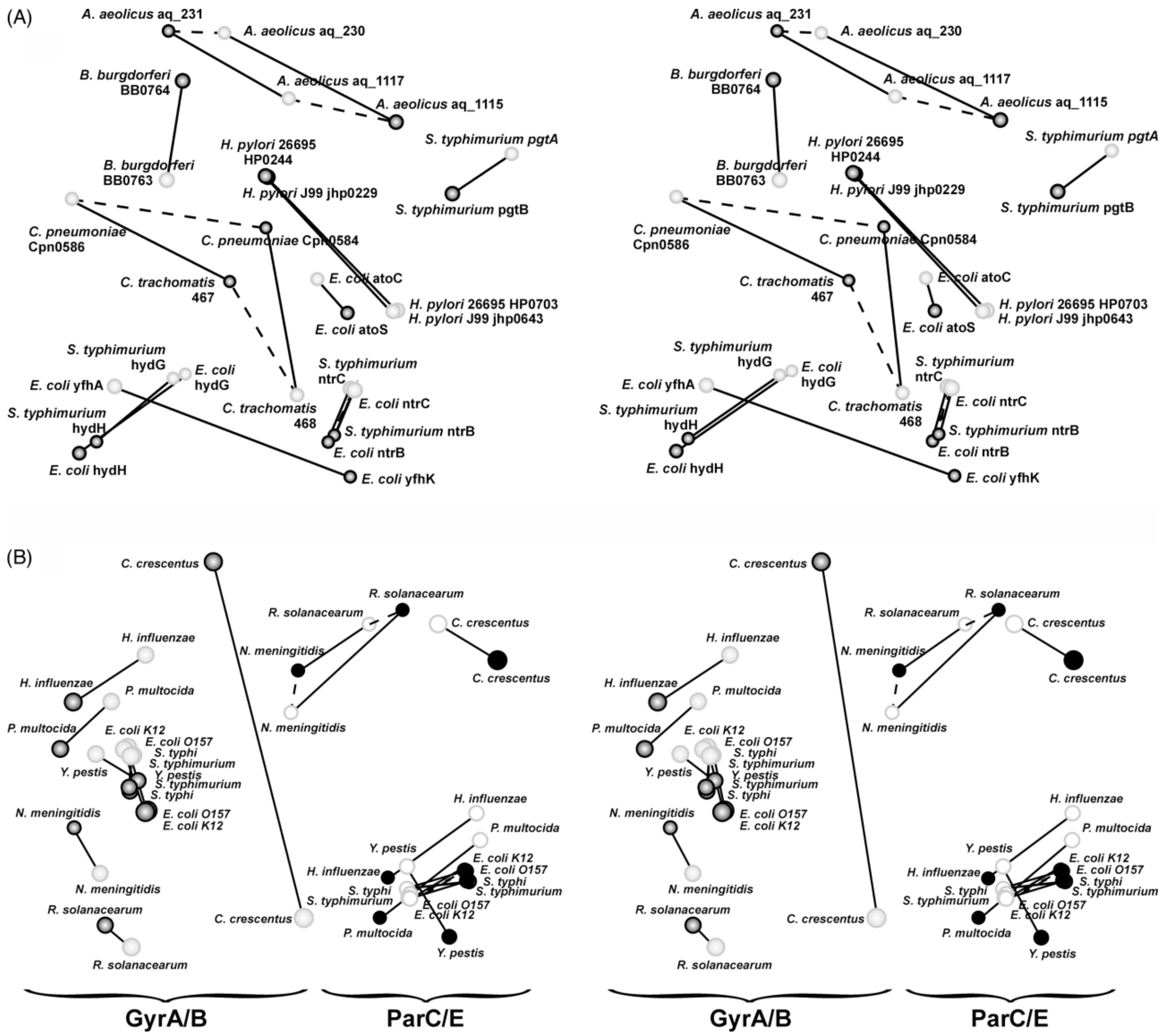

Figure 4. (A) A side-by-side stereo diagram representing the predicted and known interactions between Ntr-type two-component sensors (dark spheres) and regulators (light spheres). For both A and B continuous lines indicate interactions predicted by matrix alignment and broken lines indicate known interaction partners for cases with incorrect predictions. 12 out of 14 interactions are correctly predicted; if predictions to orthologous proteins are allowed, only the predictions for *A. aeolicus* are incorrect. (B) Stereo diagram of the interactions between GyrA (dark gray spheres) and its homolog ParC (black spheres) with their respective interaction partners GyrB (light gray spheres) and its homolog ParE (white spheres). The Gyr and Par proteins are separated into distinct spatial regions in the process of 3D embedding. With the exception of the *C. crescentus* proteins, interaction partners consistently sit adjacent to one another in space.

are mapped between the homologs GyrA and ParC (from ten organisms, as shown in Figure 1(B)) with their respective interaction partners GyrB and ParE. In the Figure, the Gyr proteins are spatially well-separated from the Par proteins, illustrating the ability of 3D embedding to separate members of a protein family into their functional subtypes. In all cases, GyrA proteins are paired with GyrB proteins, while ParC proteins are paired with ParE proteins. As with Figure 4(A), the interacting partners tend to be clustered in space. In all, 14 out of the 20 interactions are predicted correctly; when matches to orthologs are allowed, all 20 interactions (100%) are correctly predicted.

The effects of phylogenetic tree structure on inferring protein interactions

Since phylogenetic relationships and tree structure form the foundation of this approach, we investigated the importance of tree structure to the method's success. For example, we expect pairs of proteins in a tree that are highly similar to each other to be difficult to distinguish when assigning interaction partners, as in the case of the *E. coli*/*S. typhimurium* ntrC/ntrB proteins of Figure 1(A) that are incorrectly paired up in Table 1. Several such pairs of similar proteins can even lead to alternate, equally scoring solutions, as is the case

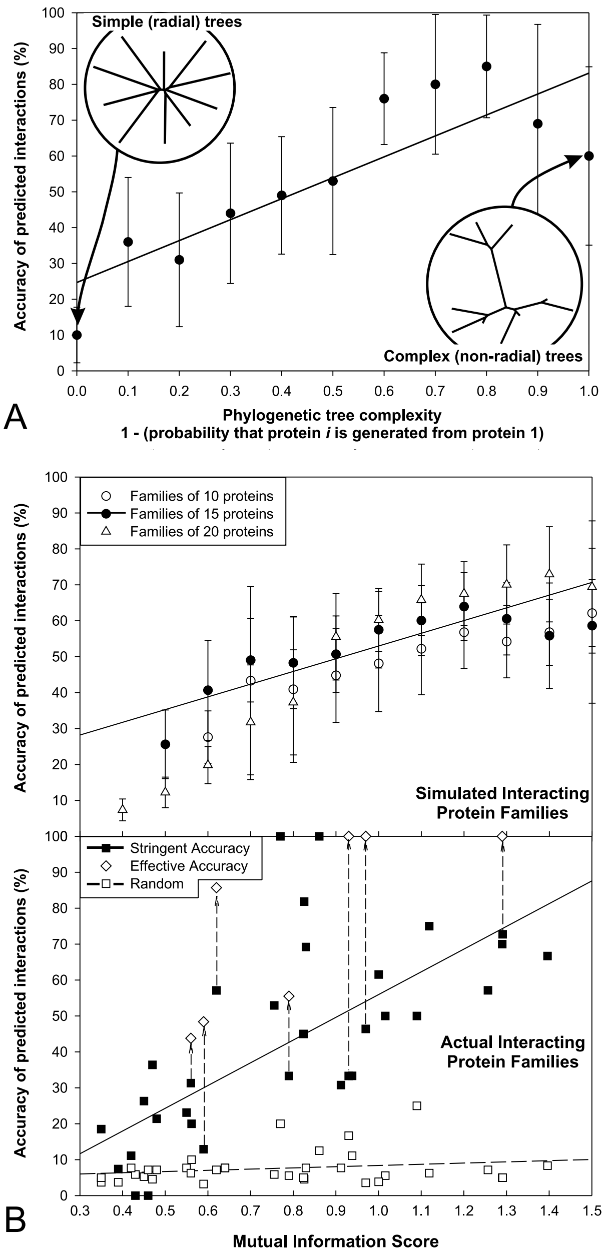

Figure 5. The accuracy of matrix alignment depends strongly on the complexity of the phylogenetic trees. (A) Simulations of the evolution of interacting proteins indicate that the tree complexity, measured by constraining simulated trees to be more or less radial, limits the accuracy of matrix alignment. As tree complexity increases, accuracy increases. This relationship is exploited in (B) (top panel), which shows that mutual information of similarity matrices correlates with prediction accuracy. Results from simulations involving pairs of protein families of different sizes indicate that as the mutual information of the similarity matrices increases, interaction prediction accuracy increases. Mutual information values are calculated in bins of width 0.1 ((B), bottom panel). This trend is confirmed in 34 actual interacting protein families, listed in Table 2. By allowing matches to orthologous proteins, the effective accuracy of the algorithm (white diamonds) is considerably higher than the stringent accuracy from exact matches (black squares). Matrix alignment significantly outperforms random choices of interaction partners (white squares).

for the CKR-type chemokines and their receptors in Table 1. In this example, the mouse/rat EOTA chemokines are predicted to bind the mouse/rat CKR2 and CKR3 receptors with equal confidence, so the precise binding partners are obscured by this underlying symmetry in the phylogenetic trees.

In order to systematically test the relationship between tree structure and matrix alignment, protein phylogenetic trees with differing complexities were created by simulating the evolution of a single protein into a protein family. Pairs of trees, representing co-evolved interaction partners, were created in coupled simulations and were analyzed by matrix alignment. By systematically varying the complexity of the trees created, the contribution of tree complexity to the effectiveness of matrix alignment could be examined.

For a given simulation of one protein (the progenitor protein) evolving into a family, tree complexity was controlled by specifying the frequency at which the progenitor protein was duplicated as compared to other proteins in the growing tree. Each new protein was added to the family by duplicating, with mutation, an existing protein under the following rule: the progenitor protein was duplicated with probability p_0, and a different protein in the family (chosen at random) was duplicated with probability $1 - p_0$. In this way, trees generated with $p_0 \sim 1$ are composed only of direct duplications of the progenitor protein, with all proteins approximately the same evolutionary distance from each other. These trees are quite simple and approximately radial in structure, as illustrated in the inset in the top panel of Figure 5. In contrast, trees generated with $p_0 \sim 0$ are more complex in structure, since lifting the requirement to duplicate the progenitor protein allows more complex patterns of duplications to occur and produces more diverse evolutionary relationships between the proteins.

To simulate the evolution of protein interaction partners, two families were "evolved" in a coupled fashion from two initial seed sequences, generated randomly as described in Materials and Methods, with the choice of protein to be duplicated at each step forced to be equivalent for the two families. For example, if in protein family A, the second protein was duplicated to create the third, then the second protein would be duplicated to create the third in family B as well. In this manner, the trees would be similar, though not identical, as stochastic mutations were introduced with each duplication as described in Materials and Methods.

Following each simulation, interactions between the two simulated interacting sequence families were predicted by matrix alignment. The results, plotted in Figure 5(A), indicate that tree complexity is strongly correlated with algorithm performance. Predictive accuracy increases with increasing tree complexity, consistent with our intuition that simple trees are ambiguous about relationships between proteins, and therefore are less useful for

Table 2. The performance of matrix alignment at predicting diverse protein interaction partners

Interacting protein families	No. of proteins[a]	Effective accuracy (%)	Stringent accuracy (%)	Mutual information
Chemokine/receptor—mouse/human/rat	31	48.4	12.9	0.59
Chemokine/receptor—human	13	NA	23.1	0.55
CKR-type chemokine/receptor—mouse/human/rat	18	55.5	33.3	0.79
CCR-type chemokine/receptor—mouse/human	6	100	33.3	0.93
Omp-type regulator/sensors—*E. coli*	14	NA	21.4	0.48
Omp-type regulator/sensors—*B. subtilis*	13	NA	7.7	0.64
Omp-type regulator/sensors—5 bacteria	16	43.8	31.3	0.56
Omp-type regulator/sensors—*E. coli/B. subtilis*	27	NA	18.5	0.35
Nar-type regulator/sensors—8 bacteria	22	36.4	36.4	0.47
Ntr-type regulator/sensors—8 bacteria	14	85.7	57.1	0.62
Cit-type regulator/sensors—*E. coli/B. subtilis*	5	100	100	0.77
Lyt-type regulator/sensors—*E. coli/B. subtilis*	4	50	50	1.09
Two component sensor/regulators—*E. coli*	27	NA	7.4	0.39
Lyt-, Ple-, and “other”-type regulator/sensors—8 bacteria	20	NA	5	0.35
CheA/CheY—11 bacteria	13	69.2	69.2	0.83
ABC transporter membrane protein 1/2—*E. coli*	19	NA	26.3	0.45
ABC transporter memb./binding prot.—*E. coli*	17	NA	0	0.43
ABC transporter membrane protein 1/2—*H. influenzae*	14	NA	0	0.46
ABC transporter memb./binding prot.—*H. influenzae*	13	NA	11.1	0.42
GyrA/B,ParC/E—α-proteobacteria	20	100	70	1.29
GyrA/B,ParC/E—Gram positive bacteria	28	100	46.4	0.97
Single interaction partners from multiple organisms				
CheA/CheB—bacteria	8	NA	100	0.86
Acetyl CoA carboxylase α/β Gram positive bacteria	9	NA	33.3	0.94
Acetyl CoA carboxylase α/β proteo bacteria	16	NA	75	1.12
Succinate CoA synthetase α/β proteo bacteria	22	NA	81.8	0.83
Succinate CoA synthetase α/β archaea	13	NA	30.8	0.91
GyrA/GyrB—α-proteobacteria	20	NA	72.7	1.29
GyrA/GyrB—Gram positive bacteria	18	NA	50	1.02
GyrA/GyrB—archaea	10	NA	20	0.56
Pyruvate dehydrogenase α/β—bacteria	17	NA	52.9	0.76
ParC/ParE—bacteria	26	NA	61.5	1.00
ParC/ParE—α-proteobacteria	12	NA	66.6	1.40
ParC/ParE—Gram positive bacteria	14	NA	57.1	1.26
DNA polymerase III E2/E3—bacteria	20	NA	45	0.82

[a] Number of proteins in a family of interacting proteins (e.g. number of columns in the corresponding similarity matrix).

predicting interactions in the manner we have described.

A score that quantitatively predicts the accuracy of matrix alignment

As simulations demonstrate a clear dependence of the success of matrix alignment upon the complexity of the phylogenetic trees, we asked if a measure of agreement between similarity matrices that also considered tree complexity would accurately predict the algorithm's performance. One such measure is the mutual information[22] of the similarity matrices, which is a function of both the entropy of the matrices, taking into account the phylogenetic tree complexity, and the agreement of the two similarity matrices with each other.

Interaction prediction accuracy was compared to the mutual information of the similarity matrices from simulations of pairs of co-evolving families of 10, 15, or 20 proteins of varying tree complexity. Results, plotted in Figure 5(B), (top) indicate that the mutual information correlates well with the prediction accuracy, with higher values of mutual information corresponding to higher prediction accuracy. No significant dependency of the measure on the size of the protein family was observed.

To extend this analysis to real data and test the general applicability of matrix alignment, we evaluated its performance on 34 sets of actual protein interaction partners, listed in Table 2, including the Omp, Nar, Cit, and Lyt-type two-component sensor/regulator proteins, the CKR and CCR-type chemokine/chemokine receptors, and membrane/substrate binding protein and interacting membrane protein components of ABC transporters. We tested simpler binary interactions, such as matching the paralogs GyrA and ParC with their specific partners, GyrB and ParE, respectively. Finally, we also tested the matching of phylogenetic trees composed of single interaction partners but from multiple species to see if they lent themselves to a similar analysis. Each set of interaction partners was analyzed by matrix alignment, and the prediction accuracy from the analyses (reported in Table 2) was compared to the mutual information of the corresponding sequence similarity matrices.

A plot of the mutual information values against the prediction accuracy (bottom panel of Figure

5(B)) shows a clear positive correlation ($R = 0.7$; accuracy $= (63.29 \times \mathrm{MI}) - 7.35$), significantly outperforming random expectations and indicating that mutual information can be used as an independent measure of the prediction accuracy. A mutual information value of 0.9 corresponds roughly with a stringent prediction accuracy of 50%; a mutual information value of 1.3 corresponds to ~75% accuracy. The effective accuracies consistently exceed these values. The trend line from the simulations agrees within error with the actual protein interactions examined, indicating that the mutual information measure correctly models both phylogenetic tree complexity and similarity, and is an appropriate measure for the prediction of protein interaction partners.

Discussion

Here, we present an automated method to predict protein interaction partners based upon similarity between the phylogenetic trees of interacting proteins. The method is effective, especially when combined with a quantitative score that correctly predicts the method's performance that arises from an information theoretic analysis of the complexity of the phylogenetic trees and their similarity to each other. Although we have specifically focused on interacting protein families of identical size, the method is easily generalized to families of different sizes by finding the subset of proteins in the larger family that best matches the proteins in the smaller family. Also, we have presented an approach based on optimization; it is reasonable to expect that methods of lower algorithmic complexity are available. Although we describe the hardest case for the algorithm, in which any protein can interact with any partner, in practice a branch-and-bound approximation is likely to greatly reduce the search space and improve the algorithm's performance. This improvement could be made by allowing similarity matrix columns to be exchanged only between proteins of the same species. However, for the case in which all proteins derive from one organism (for example, the human chemokines and receptors), such an improvement is ineffective, and algorithmic complexity will have to be reduced by other approaches.

Simulations of protein evolution indicate when the alignment of phylogenetic trees is expected to be informative. For low complexity trees, proteins are not uniquely different from each other; the consequence of this trend is that little information is stored in the tree that allows it to be oriented unambiguously to another tree. For complex phylogenetic trees, proteins have sufficiently unique patterns of similarity that alignments of such trees are unambiguous and more likely to lead to successful predictions, as shown in Figure 5.

These trends reflect not the degree of co-evolution of the interacting partners, but rather the intrinsic ambiguities in matching up trees in this fashion. The mutual information calculation accounts for this trend, providing a quantitative measure of the trees' agreement with each other as well as their intrinsic complexity. With the mutual information scoring technique, the importance of tree structure can be exploited to improve predictions: the precise proteins included in an analysis, or the organisms from which they derive, can be chosen to maximize the phylogenetic trees' mutual information, thereby enhancing the accuracy of predicted interactions. Many of the 34 examples in Table 2 represent just such experiments. For example, matching all of the *E. coli* two-component sensors against all of the two-component regulators, produces a low mutual information score (0.39) and a low prediction accuracy (7%), but limiting the analysis to the Cit-type regulator/sensor subfamilies results in higher mutual information scores (0.77) and correspondingly higher accuracy (100%).

When the information content of the trees is high, the correct interaction partners might be easily predictable simply by examining the trees. In practice, manual tree comparisons are often non-trivial and provide no information about the confidence to be placed in the predictions, as illustrated by the Gyr/Par trees of Figure 1(B). The mutual information between these trees is quite high, even though the topologies of the Gyr/Par subtrees are identical to each other. Finding interaction partners by visual examination of the trees requires careful attention to subtle changes in the branch lengths. However, the matrix alignment method offers an objective, quantitative measure of the significance of the predicted interactions. Most important, the approach is automated, allowing it to be applied on a large-scale to many protein families.

Accompanying the matrix alignment algorithm is a new method, termed 3D embedding, for visualizing protein families and interactions between them. For one protein family, this method visually summarizes the evolutionary relationships among the proteins. For two interacting protein families, these 3D embeddings can be superimposed, and the potential interaction partners can be directly visualized. 3D embedding opens the possibility of rank-ordering predicted interaction partners, such as by their spatial distance from each other. The method potentially allows the least squares alignment of two families on the basis of known protein interactions, followed by the prediction of interactions between the proteins not specifically used to generate the alignment, allowing the analysis of protein families of unequal sizes, and possibly even proteins with multiple binding partners.

Finally, the 3D embedding method illustrates how matrix alignment sometimes proceeds in a surprising fashion. As an example, it correctly pairs the *C. crescentus* GyrA and GyrB proteins, in spite of the fact that the two proteins sit in quite dissimilar relationships to the rest of their respective families (Figure 4(B)). However, the interaction is presumably predicted between the *C. crescentus* proteins because all other protein pairs match better, thereby forcing the *C. crescentus* proteins together in spite of the poor fit.

A model for the evolution of interacting proteins

Proteins are constrained to maintain their interactions and therefore have to co-evolve with their interaction partners.[23] However, the fact that the method presented here works illustrates an additional aspect of the evolution of interacting proteins: Two models can be considered for the evolution of interacting proteins, which contrast in the degree of coupling between the evolution of protein interaction specificity and the ancestral genetic events producing protein families (specifically, we consider the case of paralogs). Both models begin with an ancestral pair of interacting proteins. In the first model, the progenitor proteins are duplicated, and the duplicated proteins (paralogs) are free to evolve new interaction partners, such as by mutation and selection. After multiple duplications and evolution of new interaction specificities, two families of interacting proteins result such that the correlation in position in the phylogenetic trees is lost between pairs of paralogs with their corresponding interaction partners. In short, when gene duplications precede the evolution of interaction specificity, the phylogenetic trees of the interaction partners are no longer alignable in the fashion of the trees examined here.

However, in an alternate model, interacting protein partners are duplicated in a correlated fashion through the course of evolution. The interaction specificity is maintained or created in a process tightly coupled to the process of gene duplication. Only in this case will the phylogenetic trees of the interacting protein families be similar. The data presented here support this second model, suggesting that interacting proteins in these families are not simply duplicated and freed to evolve new interaction partners, but rather that interacting partners are duplicated in coupled processes leading to a measurable association between the specificity of protein interaction partners and the genetic relationships of their corresponding genes.

Materials and Methods

Sequence alignments, similarity matrices, and phylogenetic trees

Sequences from SwissProt[24] were aligned using CLUSTALW1.7. Similarity matrices were calculated from the multiple sequence alignment using CLUSTALW.[25] Each similarity matrix entry s_{ij} represents the evolutionary distance between a pair of proteins in a sequence family after corrections for multiple mutations per amino acid residue.[26] Similarity matrices for pairs of interacting protein families were input to the MATRIX matrix alignment algorithm described below. Unrooted phylogenetic trees were calculated *via* neighbor joining using PHYLIP.[27] Chemokine interactions were defined as described by Oppenheim and Feldmann.[28] Other interactions were assigned according to the KEGG database, version 22.0.[21]

Optimal alignment of similarity matrices

Pairs of similarity matrices were compared by their root mean square difference (r.m.s.d.), calculated as:

$$\text{rmsd} = \sqrt{\frac{2}{n(n-1)} \sum_{j=2}^{n} \sum_{i=1}^{j-1} (a_{ij} - b_{ij})^2},$$

where a_{ij} and b_{ij} represent equivalent elements of the two similarity matrices, and n is the number of proteins in each family. Smaller r.m.s.d. indicates greater agreement between two matrices.

To align matrices, the order of the rows in one matrix (and therefore columns, as a matrix is symmetric) is optimized with simulated annealing[29] to minimize the r.m.s.d. between matrices: One similarity matrix (family A in Figure 2) remains unchanged. In the second similarity matrix (family B in Figure 2), pairs of rows (and their symmetric columns) are randomly chosen and their elements are swapped, evaluating the resulting change in r.m.s.d. If r.m.s.d. decreases, the swap is kept. If r.m.s.d. increases, the swap is kept with a probability p proportional to an external control variable T, such that $p = \exp(-\delta/T)$, where δ equals the increase in r.m.s.d. with the swap. The control variable T is initialized such that p is first set to 0.8; T is decreased linearly with each iteration ($T_{\text{new}} = 0.95T_{\text{old}}$). This process is iterated until the probability of accepting an increase is less than 10%.

Following simulated annealing, interactions are predicted between proteins heading the corresponding rows of the two similarity matrices. As the possible number of re-ordered matrices is factorial with the number of proteins in the matrix, this method does not guarantee the correct solution for large matrices (>15 proteins). In these cases, the protocol is repeated 100 times, and the frequency of occurrence of a given interacting protein pair is calculated and tabulated in order to test the reproducibility of the predictions. Interactions are then assigned between the most frequent protein pairings.

3D embedding of protein sequence families

Proteins were represented as mass-less points in space connected by springs whose equilibrium lengths were equal to the proteins' pair-wise similarities (s_{ij}). Each protein in a sequence family was initially assigned to a random position, then moved in an iterative fashion to minimize the action of spring forces. At equilibrium, the proteins are placed such that distances separating the proteins (d_{ij}) agree maximally with the similarities in the similarity matrix, except for the distortion inherent in mapping high-dimensional relationships into three-dimensional space. Pairs of interacting protein families visualized in this fashion were superimposed by rigid body least squares fit of one family onto the other using SwissPDBViewer,[30] minimizing the distance between predicted or known interaction partners. Note that the possibility exists for positioning a set of proteins in mirror-image embeddings, complicating alignment of interacting proteins. In practice, repeating the embedding to achieve compatible handedness with the interacting proteins can circumvent this problem.

Simulations of the evolution of protein interactions

Pairs of amino acid sequences of length 300, representing ancestral interacting proteins (sequences 1A and 1B), were randomly generated using naturally occurring amino acid frequencies. The evolution of a sequence pair into two families of interacting paralogs was then modeled by successive duplication, with mutation, of a protein from family A and the corresponding protein from family B, forcing parallel duplications in the two families. Mutations were randomly introduced at each duplication with the amino acid substitution frequencies of a PAM25 substitution matrix,[31] which has the effect of mutating ~25% of the amino acid residues per protein per duplication. In this manner, the underlying pattern of duplications is held constant between two families, and point mutations in each sequence are modeled.

After a simulation, the family A sequences were aligned to each other, as were the family B sequences. The similarity matrix for each family was calculated (as for actual proteins) and matrix alignment performed. Correct predictions were assigned between equivalent proteins (e.g. pairing 1A to 1B, the first duplicate of 1A to the first duplicate of 1B, etc.). Simulations were repeated with a parameter p_0 controlling the choice of ancestor for each new paralog, as described in the text. In Figure 5(A), simulations were performed ten times per data point plotted for protein families of ten members; in Figure 5(B), 100 simulations per value of p_0 were performed for a given family size, sampling from $p_0 = 0.0$ to 1.0 in 0.1 increments.

Information theoretic-based measure of agreement between phylogenetic trees

The agreement between pairs of phylogenetic trees was calculated using an information theory[22]-based metric, mutual information, which accounts both for the similarity matrices' agreement as well as for their intrinsic information content. The information content of a similarity matrix is assessed as the entropy $H(x)$ of the distribution of values in the similarity matrix, calculated as:

$$H(x) = -\sum_{x} p(x)\log p(x),$$

where x represents bins of values drawn from a similarity matrix, and $p(x)$ represents the frequency with which those values are observed in the matrix. Given two similarity matrices, the relative entropy $H(x,y)$ represents the extent of their agreement, calculated as:

$$H(x,y) = -\sum_{x,y} p(x,y)\log p(x,y),$$

where x,y represents bins of pairs of values in equivalent positions of the two similarity matrices, and $p(x,y)$ represents the relative frequency with which pairs of values are observed in equivalent positions of the two matrices.

The mutual information (MI) between two matrices, representing their overall agreement, is calculated as:

$$\mathrm{MI} = H(x) + H(y) - H(x,y),$$

accounting both for the complexity of the phylogenetic trees (in the $H(x)$ and $H(y)$ terms, which are larger with more complex trees) and their similarity (in the $H(x,y)$ term, which is smaller given better agreement). A high mutual information score indicates a pair of complex and mutually consistent phylogenetic trees.

Acknowledgements

The authors acknowledge support from grant F-1515 of the Welch Foundation, the Texas Advanced Research Program, the National Science Foundation, and a Dreyfus New Faculty Award and Packard Fellowship for E.M.M.

References

1. Pruitt, K. D. & Maglott, D. R. (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucl. Acids Res.* **29**, 137–140.
2. Hsu, S. Y., Nakabayashi, K., Nishi, S., Kumagai, J., Kudo, M., Sherwood, O. D. & Hsueh, A. J. (2002). Activation of orphan receptors by the hormone relaxin. *Science*, **295**, 671–674.
3. Saito, Y., Nothacker, H. P., Wang, Z., Lin, S. H., Leslie, F. & Civelli, O. (1999). Molecular characterization of the melanin-concentrating-hormone receptor. *Nature*, **400**, 265–269.
4. Chambers, J., Ames, R. S., Bergsma, D., Muir, A., Fitzgerald, L. R., Hervieu, G. *et al.* (1999). Melanin-concentrating hormone is the cognate ligand for the orphan G-protein-coupled receptor SLC-1. *Nature*, **400**, 261–265.
5. Sprinzak, E. & Margalit, H. (2001). Correlated sequence-signatures as markers of protein–protein interaction. *J. Mol. Biol.* **311**, 681–692.
6. Pazos, F. & Valencia, A. (2002). In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins: Struct. Funct. Genet.* **47**, 219–227.
7. Lockless, S. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
8. Lichtarge, O., Bourne, H. & Cohen, F. (1996). An evoloutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
9. Jones, S. & Thornton, J. (1997). Prediction of protein–protein interaction sites using patch analysis. *J. Mol. Biol.* **272**, 133–143.
10. Huynen, M., Snel, B., Lathe, W. & Bork, P. (2000). Predicting protein function by genomic context: quantitive evaluation and qualitative inferences. *Genome Res.* **10**, 1204–1210.
11. Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998). Conservation of gene order: a finger print of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328.
12. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. (1999). The use of gene clusters to

infer functional coupling. *Proc. Natl Acad. Sci.* **96**, 2896–2901.

13. Enright, A., Iliopopulos, I., Kyrpides, N. & Ouzounis, C. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
14. Marcotte, E., Pellegrini, M., Ng, H., Rice, D., Yeates, T. & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.
15. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci.* **96**, 4285–4288.
16. Fryxell, K. J. (1996). The coevolution of gene family trees. *Trends Genet.* **12**, 364–369.
17. 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.
18. Koretke, K., Lupas, A., Warren, P., Rosenberg, M. & Brown, J. (2000). Evolution of two-component signal transduction. *Mol. Biol. Evol.* **17**, 1956–1970.
19. Pazos, F. & Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.* **14**, 609–614.
20. Hughes, A. L. & Yeager, M. (1999). Coevolution of the mammalian chemokines and their receptors. *Immunogenetics*, **49**, 115–124.
21. Kanehisa, M. (1996). Toward pathway engineering: a new database of genetic and molecular pathways. *Sci. Technol. Jpn*, **59**, 34–38.
22. Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423.
23. Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. (2002). Evolutionary rate in protein interaction network. *Science*, **296**, 750–752.
24. Bairoch, A. & Apweiler, R. (1997). The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucl. Acids Res.* **25**, 31–36.
25. Thompson, J., Higgins, D. & Gibson, T. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
26. Kimura, M. (1983). *The Natural Theory of Molecular Evolution*, Cambridge University Press.
27. Felsenstein, J. (1993). *PHYLIP 3.5c (Phylogeny Inference Package)*, University of Washington, Seattle.
28. Oppenheim, J. J. & Feldmann, M. (2001). Cytokine reference, a compendium of cytokines and other mediators of host defense. *Chemokine Reference*, Academic Press, San Diego.
29. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, **220**, 671–680.
30. Guex, N., Diemant, A. & Peitsch, M. C. (1999). Protein modelling for all. *Trends Biochem. Sci.* **24**, 364–367.
31. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., ed.), vol. 5, pp. 345–352, National Biomedical Research Foundation, Washington DC.

Edited by F. E. Cohen

(Received 17 September 2002; received in revised form 17 December 2002; accepted 10 January 2003)