High-confidence prediction of global interactomes based on genome-wide coevolutionary networks

David Juan*, Florencio Pazos†, and Alfonso Valencia*‡

*Structural Bioinformatics Group, Spanish National Cancer Research Centre, Melchor Fernández Almagro 3, 28029 Madrid, Spain; and †Computational Systems Biology Group, National Centre for Biotechnology (CNB–CSIC), Darwin 3, Cantoblanco, 28049 Madrid, Spain

Edited by Michael Levitt, Stanford University School of Medicine, Stanford, CA, and approved November 29, 2007 (received for review October 11, 2007)

Interacting or functionally related protein families tend to have similar phylogenetic trees. Based on this observation, techniques have been developed to predict interaction partners. The observed degree of similarity between the phylogenetic trees of two proteins is the result of many different factors besides the actual interaction or functional relationship between them. Such factors influence the performance of interaction predictions. One aspect that can influence this similarity is related to the fact that a given protein interacts with many others, and hence it must adapt to all of them. Accordingly, the interaction or coadaptation signal within its tree is a composite of the influence of all of the interactors. Here, we introduce a new estimator of coevolution to overcome this and other problems. Instead of relying on the individual value of tree similarity between two proteins, we use the whole network of similarities between all of the pairs of proteins within a genome to reassess the similarity of that pair, thereby taking into account its coevolutionary context. We show that this approach offers a substantial improvement in interaction prediction performance, providing a degree of accuracy/coverage comparable with, or in some cases better than, that of experimental techniques. Moreover, important information on the structure, function, and evolution of macromolecular complexes can be inferred with this methodology.

coevolution | interaction | mirrortree

Coevolution is a well characterized process that takes place at all biological levels, from ecosystems to molecules. Coevolution between interacting protein families had been proposed for some cases based on the qualitatively observed similarity of their phylogenetic trees (1, 2). This tree similarity was later quantified and statistically demonstrated to be related to protein interactions in large datasets of interacting families (3, 4). This "mirrortree" approach has been followed by many authors, who have developed different extensions of the method. Many of these extensions have been aimed at correcting factors that influence tree similarity but that are not related with the interaction, thereby affecting the predictive performance of this technique. For example, an obvious extension has been the inclusion of information on the phylogeny of the organisms involved to correct for the "background similarity" expected for any pair of trees resulting from the underlying speciation events (5, 6).

Still, there are many other factors affecting the relationship between interactions and tree topology. Maybe one of the most important is related to the fact that a protein is coevolving with many interactors simultaneously. This would make it difficult to separate the effect of each of them on the topology of the tree. Nevertheless, all of the methods developed to date consider the pairs as isolated when evaluating their coevolution. Moreover, methods for predicting protein interactions based on tree similarities are prone to errors from several sources (e.g., problems in detecting orthologs, multiple sequence alignment errors, etc.). The paradigm- and methodology-related limitations have reduced the potential application of protein-interaction prediction based on coevolution. Here, we have introduced qualitative changes to the paradigm by moving from the limited pairwise observations toward a complete cellular "coevolutionary context." We propose use of the information contained in the whole "coevolutionary network" of an organism (the network containing all of the pairwise tree similarities) to gain information on the "coherence" or robustness of a given coevolutionary signal.

By using this coevolutionary context information, we predicted the interactome of *Escherichia coli* with a degree of accuracy and coverage comparable with that of the high-throughput experimental techniques. We evaluated the predictive performance of this method in large datasets representing different types of physical and functional relationships between proteins. We also discuss in detail the predictions for some particular systems, showing that this approach is able not only to detect the interactions within these systems but also to provide additional information on their substructure and functioning.

Results

Improving Coevolution-Based Prediction of Protein Interactions. To assess the improvements attributable to the new methodology, we used as our baseline the results of the original mirrortree method (4) that take into account only pairs of interactions and not the global coevolutionary landscape [see *Network of Raw Tree Similarities (Coevolutionary Network)* in *Methods*]. The accuracy of this method for the different test sets is shown in Fig. 1*A*. It was clear that this simple approach could capture part of the coevolutionary signal related to protein interactions, as is particularly evident for the manually curated EcoCyc complexes, where a level of confirmation close to 40% was obtained for the top 500 predictions.

When we placed side-by-side these results and those obtained by correlating the coevolutionary profiles (see *Calculation of the Coherence of Evolutionary Similarity Based on The Whole Coevolutionary Network* in *Methods*), we observed a drastic improvement up to the first 500 predictions, obtaining an accuracy of 100% for the top 100 predictions when evaluated against EcoCyc (Fig. 1*B*). If we went down the first 500 top predictions, the accuracy (for EcoCyc) was similar to that of the original mirrortree. This improvement at the top of the list was attributable to a dramatic reduction in the number of false positives produced previously by the original mirrortree method. This improvement was a direct consequence of using the evolutionary information of the whole proteome for confirming the coevolution of a given pair of proteins. The best accuracies were obtained for the datasets related to manually annotated com-

Author contributions: D.J. designed research; D.J. performed research; D.J., F.P., and A.V. analyzed data; and D.J., F.P., and A.V. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

‡To whom correspondence should be addressed. E-mail: valencia@cnio.es.

This article contains supporting information online at www.pnas.org/cgi/content/full/0709671105/DC1.

www.pnas.org/cgi/doi/10.1073/pnas.0709671105

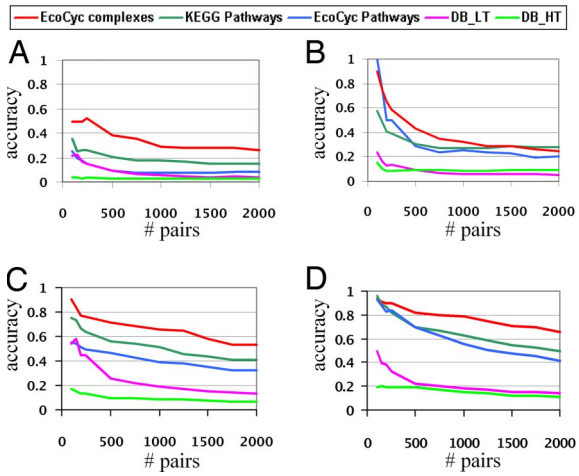

Fig. 1. Confirmation of the predictions for different steps of the method and for different interaction datasets. (*A*) Mirrortree. (*B*) Profile–profile correlation. (*C*) Partial correlation, 1st level. (*D*) Partial correlation, 10th level. The *x* axes represent the number of top predictions (pairs with highest scores), and the *y* axes represent the confirmation according to the different datasets of protein interactions and relationships.

plexes, followed by metabolic pathways. These datasets were among the more reliable ones.

A substantial improvement was obtained when the most stringent partial correlation of each pair of proteins was used to score that pair (level 1; see *Methods* and Fig. 1*C*). In this step, a large proportion of the false positives introduced by the "broad" evolutionary trends are removed, such as those attributable to the speciation process (5). Indeed, the improvement observed at this point was impressive (Fig. 1*C*), roughly doubling the accuracy of the previous step for the first 1,000 predictions (from 31% to 65% for manually curated complexes).

Finally, we focused on recovering those protein coevolutionary relationships that involve more than two proteins. We did this by relaxing the partial correlation criteria considering the *n*th most stringent partial coevolution for each protein pair (steps 4 and 5 in Fig. 4). Fig. 1*D* shows the results for the 10th level of partial correlation. These results were even slightly better than those obtained for the most stringent level (previous point), indicating that some valuable relationships are masked when filtering for the influence of third proteins. The 100 top predictions produced an accuracy of 93–96% with the datasets derived from KEGG and EcoCyc (Fig. 1*D*). Globally, this 10th level produced the best predictions (data not shown), although better results could be obtained for particular proteins with different specificities of partial correlation (depending on whether or not they are involved in functionally related broad coevolutionary trends).

Additional representations of the results for the EcoCyc dataset by using receiver operator curves are described in the supporting information (SI) *Text* and SI Fig. 5.

We also performed some tests with the yeast proteome. A detailed description of the methodology, datasets, and results for this organism are available in the *SI Text*. For the profile–profile correlations (see *Methods*), the accuracies for these two organisms were similar (SI Fig. 6*B*). However, when we considered the partial correlations (see *Methods*), no improvement was obtained for yeast (18% accuracy for the 2,000 top-scoring pairs in the KEGG dataset), contrary to what occurred in *E. coli* (51%) (Fig. 1*D* and SI Fig. 6*D*). This is probably attributable to factors such as the difficulty in obtaining clean sets of orthologs and the lower number of (eukaryotic) organisms used to build the alignments/trees. All of these factors led to a reduction in the number of possible pairs to build the coevolutionary network and hence in the choices for finding third proteins that could explain the observed coevolution for a given pair. Additional results are available in the *SI Text*.

Performance of the Different Sets of Protein Relationships. The coevolutionary information seems to be differentially related to the different types of protein associations (Fig. 1). Furthermore, these differences are relatively consistent for the different steps of our protocol. In all cases, coevolution seems to be strongly related to the protein associations represented by the manually curated complexes. These complexes are well studied, stable macromolecular machines with a strong functional dependence, and, in most cases, they are conserved between organisms. All of these features make them particularly apt to display coevolutionary behavior. Lower accuracies were obtained when comparing against datasets representing weaker or "human-imposed" relationships (i.e., metabolic pathways). All of these findings suggest that the strength of the functional/physical association is directly related to the level of coevolution.

There is almost complete disagreement between our predictions and high-throughput pull-down experiments. These experiments were aimed at detecting stable protein complexes that would strongly overlap with manually curated ones. To obtain insight into the reasons for this disagreement, we evaluated the "accuracy" of the protein pairs derived from the high-throughput set that were also predicted by our method by using the EcoCyc complexes and KEGG pathways as "gold standards." We compared this accuracy with that obtained for the high-throughput pairs alone and for our predicted pairs alone. There was little confirmation of the experimental high-throughput pairs (≈5% with EcoCyc complexes and ≈11% with KEGG pathways; SI Fig. 7 *A* and *B*), which may reflect a large set of previously unexpected stable interactions or, more probably, a large proportion of false positives. In any case, when our method for selecting a subset of these data was used, the agreement increased up to ≈55% for complexes and ≈70% for pathways. Although the agreement of this subset was lower than that obtained by using predictions alone (SI Fig. 7 *A* and *B*), this must be interpreted as the value of integrating experimental information, because these "new" positive pairs were extracted from predictions with lower scores (SI Fig. 7 *C* and *D*). This suggests that there is room for improvement when combining these two sources of information (experimental and computational).

Finally, the predictions evaluated against the pairs derived from low-throughput experiments were also worse than expected (Fig. 1), except perhaps for those derived from very restrictive partial correlation criteria (Fig. 1*C*). This may be related to one of the main limitations of the coevolution-based approaches. As suggested previously (7), coevolution does not seem to occur (or it occurs to a much lesser extent) between transiently interacting proteins. This is probably related to the fact that the evolutionary (coevolutionary) pressure for these pairs is not as strong because of their nonpermanent nature. Even though they may be fundamental for the regulation of cellular processes, they are not as mutually dependent as more stable interactions.

Interesting Examples of Coevolution Specificity. Some examples of interaction networks predicted with this new methodology are shown (Fig. 2), where the colors of the links in Fig. 2 represent predicted pairs with the 1st, 5th, or 10th best partial correlation value >0.6 (red, blue, and black, respectively). We can see how different levels of coevolution specificity affected our predictions. Very specific coevolution (not related with third proteins) arose at the first level (no external influence allowed), and as we went to higher levels, this specificity was relaxed as clusters of coevolving proteins were detected.

One group of proteins that coevolve in a very tight way is that

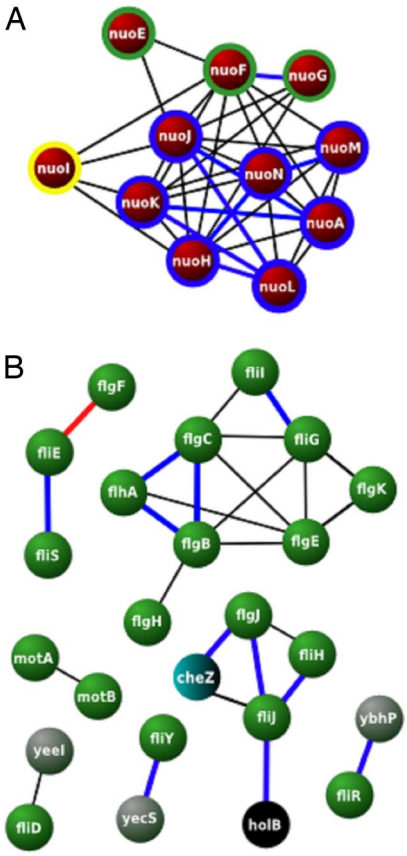

Fig. 2. Examples of predicted clusters of related proteins. (*A*) Proteins related to the NADH oxidoreductase complex. (*B*) Flagellar assembly proteins. The link colors represent levels of coevolutionary specificity: 1st level (red), 5th level (blue), and 10th level (black). The colors of the nodes represent those belonging to the same complex/pathway. (Gray is used to represent unknown/hypothetical proteins, and black is used to represent a false positive.) For the NADH oxidoreductase example, the colors of the surrounding circles represent different structural and functional modules of the complex. More examples are shown in SI Fig. 8.

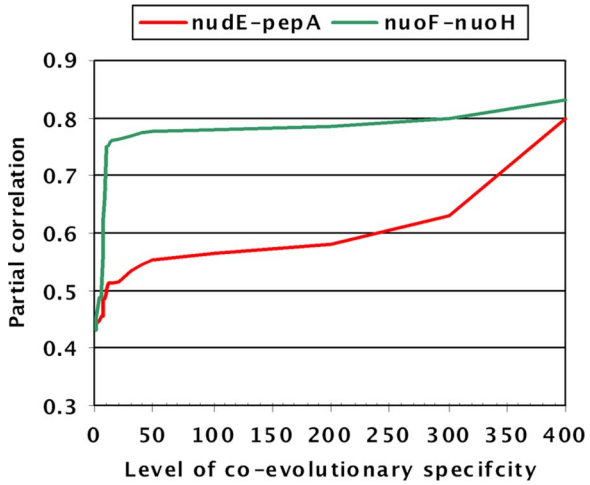

Fig. 3. Coevolutionary specifity. Different partial correlation specificity levels for two protein pairs. Partial correlation values of different levels of specificity for an unrelated pair of proteins (NudE–PepA, red line) and a positive case involving two proteins, the NADH oxidoreductase complex (NuoF–NuoH, green line).

of the 12 proteins forming the NADH–quinone oxidoreductase complex (Fig. 2*A*). Three different structural and functional modules have been described in this complex. The N module oxidizes NADH, the Q module reduces ubiquinone or menaquinone, and the P module translocates the protons across the membrane (8). Although no associations were detected at the first level of specificity, more relaxed levels showed increasing degrees of evolutionary dependence related to the biological functioning of this complex. The progress of the partial correlation values for different levels of specificity for the NuoF/NuoH pair is compared with a negative case, two proteins that do not interact physically or functionally, NudE/PepA (Fig. 3). For the very first levels, both pairs had very similar values (although at the first level the negative case had a slightly higher value). However, after removing the effect of other cocomplexed partners, the values for NuoF/NuoH rose dramatically, whereas the negative case displayed a much milder increase. We found that at a midlevel of specificity, all of the connections were intramodular (six for the P and one for the N module; Fig. 2*A*), whereas at the 10th level (relaxed specificity), a high level of coordinated evolution between all of the members of the complex was observed (intermodule). The predictions for the members of this complex up to level 10 did not contain any false positives. It is important to note that only two members of the complex were not detected with the thresholds defined above: NuoB and NuoCD. NuoB shows significant relationships ($P < 10^{-6}$) with NuoE, NuoF, NuoG, NuoH, NuoI, NuoJ, and NuoK but with partial correlation values below the established threshold (ranging from 0.43 to 0.55). Moreover, NuoB did not show any other significant relationship. In the case of NuoCD, no significant relationships were found regardless of the threshold. This is because only 18 orthologs were detected for this protein with the methodology used here. This low number of orthologs hides the evolutionary signal for this sequence, making it more difficult to detect its evolutionary dependences. Thus, this method is not only able to relate most of the members of the NADH–quinone oxidoreductase complex with high sensitivity and specificity, but it also distinguished between highly specific relationships (intramodule) and less specific broad ones (intermodule), providing additional information on the detailed structure and functioning of the complex.

Another interesting example is the machinery for flagellar assembly (Fig. 2*B*). For this very complex machine, we are able to obtain predictions for 19 proteins that display a high degree of connectivity among them. This case is particularly difficult because of the large number of proteins (37 proteins according to KEGG). This large number of proteins would suppose a level of specificity less than ten, because we would expect to have to remove evolutionary patterns related to this process. Furthermore, the large number of relationships increases the difficulty of obtaining an accurate ranking of partial correlations. Even so, we found that most of the linked proteins indeed participated in this process, and only three (apparent) false positives and three predictions involving unknown proteins were identified. The first two false positives involve CheZ, a protein known to participate in chemotaxis, a process that is related to the regulation of flagellar rotation (9). CheZ has a very specific coevolutionary signal with FliJ, one of three soluble components of the flagellar export system and an association of intermediate specificity with one of the other two, FilH. The other false positive was the δ subunit of the DNA polymerase III (HolB). Although replication is a completely different process, and there is insufficient evidence to suggest a functional relationship between HolB and FliJ, it is interesting to note that several studies have identified a close regulatory relationship between the assembly of the flagellum and the DNA replication in *Caulobacter crescentus* (10). With regard to the unknown proteins, there is no infor-

mation for two of them (YeiI and YbhP), whereas for the third, YecS, there is indirect evidence linking it to FliY. Both proteins seem to be part of a cysteine ABC transporter, and they are predicted to be linked by other "context-based" computational methods such as gene fusion, gene neighborhood, and gene cooccurrence in STRING (11). Finally, there is a larger cluster composed of eight highly connected proteins (all of them part of the flagellum) that probably represents the ancestral core of the machinery. This is consistent with the observation that if we relaxed the specificity level up to 20, another five proteins would become attached to this cluster without adding any further false positives (FlgG, FlgI, FlgL, FliF, and FliM; data not shown). In the *SI Text*, we discuss in detail the predictions for some other complexes.

Discussion

Many features characterizing living systems can only be understood by considering the complex network of relationships between cellular components. Biological systems are the prototype of complex systems, where "the whole is more than the sum of the parts" (12, 13).

The relationship between protein coevolution and interactions has been repeatedly demonstrated by many authors (see the Introduction). Hypotheses for explaining such a relationship include the similar evolutionary pressure and the mutual coadaptation of interacting proteins. The coadaptation hypothesis is a very challenging one, and there are results in favor and against it (7, 14). It is important to keep in mind that the practical utility of the method is totally independent of this hypothesis to be true or not, and it only depends on the demonstrated relationship between tree similarity and interaction. It is reasonable to consider coevolutions resulting from coadaptations at the molecular level as more specific, involving pairs of proteins (or very small groups), and to consider coevolutions not resulting from specific coadaptations (but from similar evolutionary pressures) as more general, involving large macromolecular complexes and groups of proteins (pathways, etc.). In this context, we think that the method presented here, which is able to separate specific from broad coevolutions, could help in clarifying this issue. In any case, coadaptation is interesting as a working hypothesis, and it drove some of the improvements of the method.

The full network of molecular interactions in a cell can be seen as a coevolving system in which the individual properties of the components depend on the interactions with others. The complex network of coevolutionary relationships between proteins cannot be easily split into the individual pairwise coevolutions because the fact that proteins has to adapt to many different interactors makes these pairwise coevolutions highly dependent to each other. Therefore, it makes sense to study the properties of coevolution in the complete system.

The idea of coevolution of the entire interacting system can be implemented and tested in various possible ways. The method we propose here falls within the context of those based on the comparison of protein family phylogenetic trees (mirrortree) for the prediction of interactions. Our approach is completed in two steps: in the first, we evaluated the similarity of the coevolutionary patterns of the two proteins (patterns of coevolution with all of the other available family trees); and in the second step, we assessed the influence of other proteins in the coevolution of a given pair by calculating the corresponding partial correlation of their family trees.

The results presented here demonstrate that the predictions of protein interactions are clearly better when the "evolutionary context" is taken into account. The inclusion of this context information corrects in a natural way many factors that affect the performance of the method, including the background similarity attributable to the underlying speciation process addressed previously by other authors (5, 6). We observe that the predictions obtained with this new approach are more closely related with the relationships represented by EcoCyc complexes and metabolic pathways (KEGG and EcoCyc). Conversely, they appear to be less closely related to the interactions detected by high-throughput techniques and the low-throughput ones annotated in curated databases. The low agreement with the high-throughput data could be explained by the poor quality of these data. Furthermore, we show that although significant improvements can be made by discarding nonspecific coevolutions, further improvements can be obtained by establishing a less stringent level of specificity. Actually, our results show that on average, better results are obtained for the 10th level of partial correlation. This can be easily explained by the fact that some nonspecific coevolutions are important. In some complexes, all of the proteins coevolve with one another. Rather than an artifact, this nonspecific global coevolution is intrinsic to their function, and, hence, it should not be corrected. This highlights the potential value of detecting the optimal level of specificity for each system.

Together with evaluating the implementation in global terms of prediction accuracy, we show how the analysis of the coevolutionary relationships can be used to explore the functional topology and evolution of macromolecular complexes. This was clearly demonstrated for a number of cases, including the NADH–dehydrogenase complex. Hence, this method can be used not only for predicting interactions with high confidence but also to gain insight into the function and structure of the macromolecular complexes by using sequence information alone.

There are other methods for predicting protein interactions and functional relationships that use genomic and sequence features intuitively related with protein interactions. For a recent review, see ref. 15. Nevertheless, all of these methods consider a given pair of proteins as totally independent from the others. The method presented here is the first that uses information on the whole proteome for assessing the possible interaction of two proteins.

Our approach requires a minimum amount of evolutionary information about a given protein from a nonredundant set of organisms to build congruent protein alignments of related species. It is conceivable that the unceasing increase in the number of genomes sequenced will substantially increase the possibilities of applying methods based on coevolution models, such as the one presented in this work.

Methods

The ContextMirror method is depicted in Fig. 4. An initial coevolutionary network containing raw tree similarities for all protein pairs is calculated. An optimized measure of coevolution between two given proteins is then obtained by comparing their patterns of coevolution with all of the others. Finally, the influence of third proteins on the coevolution of a given pair is evaluated.

Network of Raw Tree Similarities (Coevolutionary Network). The starting point of the method is the generation of the coevolutionary network that contains all of the raw tree similarities for all of the possible pairs of proteins within the *E. coli* genome. The protocol for calculating these tree similarities is similar to that described previously (5). The first step is the generation of alignments of orthologs for all of the proteins in the *E. coli* genome. Orthologs for all of the proteins from *E. coli* are detected in a set of 116 fully sequenced genomes by using the standard BLAST "best bidirectional hit" method, with a cutoff P value of 10^{-5} and requiring an alignment of at least 70% of the protein length. These 116 genomes were obtained from the set of all fully sequenced prokaryotic genomes (218 in February 2005) by taking only one representative from each clade at the first level of the National Center for Biotechnology Information taxonomic tree (the one with the largest genome), which means taking only one strain for each species. The sequences within these ortholog sets were aligned with MUSCLE (16), and phylogenetic trees were obtained from these alignments with the "neighbor-joining" (NJ) algorithm implemented in ClustalW (17). The mirrortree method is based on the comparison

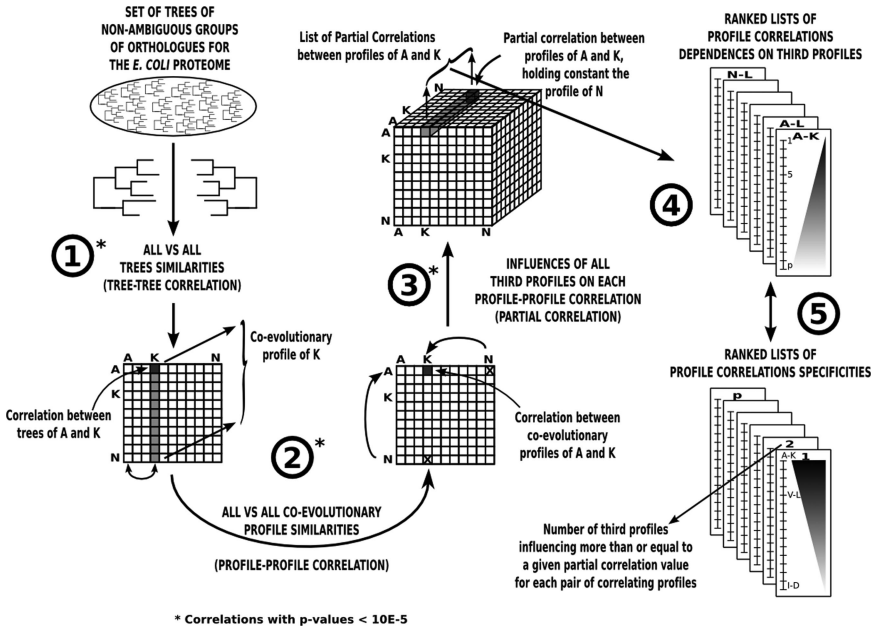

Fig. 4. Schema of the ContextMirror method. An initial coevolutionary network containing raw tree similarities for all protein pairs is calculated (step 1). The similarity between coevolutionary patterns (vectors containing all of the tree similarities) is calculated for all pairs of proteins (step 2). The specificity of the coevolution between two proteins is evaluated by calculating their partial correlation given all of the others (step 3). The list of partial correlations for each pair of proteins is sorted (step 4). Levels of partial correlation specificity for all of the protein pairs are obtained and ranked (step 5). In all of the steps, only pair relationships with a *P* value of $<10^{-5}$ were considered.

of protein distance matrices rather than of the phylogenetic trees themselves. A distance matrix for each family (set of orthologs) is obtained from the phylogenetic tree described above by summing the lengths of the branches separating each pair of proteins.

Finally, the tree similarity between two families is calculated as the correlation between their distance matrices. Only the distances between species included in the multiple sequence alignment of both families can be used for this calculation. A minimum of 15 common species are required to test a pair of families (105 distance values). The mirrortree score between A and B (r_{AB}) is calculated as in ref. 5. Significant correlation values are selected based on significance cutoffs of 10^{-5} (tabulated *P* values). In this step, we obtain significant correlation values for 1,089,362 pairs comprising 2,077 proteins.

Calculation of the Coherence of Evolutionary Similarity Based on the Whole Coevolutionary Network. Despite having been repeatedly shown to detect protein interactions with reasonable accuracy, these raw tree similarities still display some degree of noise that produces many false positives and negatives (see the Introduction and references therein). Part of this noise may be attributable to the interdependence between these similarities not taken into account so far, although it may also be related to intrinsic methodological limitations, such as the automatic nature and the assumptions of the method (orthology detection method, NJ tree, etc.). We try to reduce these sources of noise by evaluating the coherence of the coevolutionary signals by using all of the information in the coevolutionary network. The idea is based on a ''conservative witnesses consensus opinion'' principle that could be phrased as, ''Because I cannot believe you, I will ask all those that know you.'' Imagine how we could ensure that two people who claim to know each other are not lying. If we investigate their friends, and we find that they share many of them, we have additional indirect proof of their friendship. In this case, we ''believe'' the coevolutionary signal between two families only if their patterns of coevolution with all of the others are also similar. That is, if their coevolutionary contexts are also similar. This rationale is implemented in a very simple way. We represent the correlation values of the significant pairs obtained in the previous step ($P < 10^{-5}$) in a matrix. A row/column in this matrix (correlation vector) contains the correlation values for a given protein with all of the others. We then calculate the Pearson's correlation for every pair of correlation vectors (Fig. 4). Thus, the correlations of protein A with all of the other proteins (r_{Ai}) and those of protein B (r_{Bi}), both calculated as above, were used to calculate a new correlation coefficient between these two proteins (r'_{AB}):

$$r'_{AB} = \frac{\sum_{i=1}^{N} (r_{Ai} - \overline{r_{Ai}}) \cdot (r_{Bi} - \overline{r_{Bi}})}{\sqrt{\sum_{i=1}^{N} (r_{Ai} - \overline{r_{Ai}})^2} \cdot \sqrt{\sum_{i=1}^{N} (r_{Bi} - \overline{r_{Bi}})^2}},$$

where *N* is the number of proteins in the genome for which the correlation values (see above) with both A and B could be calculated. In this way, we reassess the evolutionary similarity between A and B by evaluating the coherence of their coevolution with all of the other proteins. In other words, we consider that two proteins coevolve not only if their trees are similar but also if their coevolutionary behaviors with respect to all of the other proteins are also similar. This additional restriction allows us to optimize the measure of coevolution. In this step, we obtain significant correlations ($P \leq 10^{-5}$) for 574,997 pairs comprising 1,942 proteins.

Assessment of the Influence of Third Proteins on the Coevolution of a Given Pair. The fact that the evolution of two proteins is coordinated, either when measured directly or by evolutionary context (see above), does not ensure that this coevolution is ''specific'' or ''particular'' to these proteins. Rather, it may be attributable to a ''general'' evolutionary tendency involving more proteins, which is reflected in similar pairwise coevolution for all of them. These broad coevolutions can be very informative in some cases, i.e., macromolecular complexes with constituents that are subject to similar evolutionary pressures. However, often they are far from reflecting protein interactions or functional relationships [i.e., similar coevolution resulting from the speciation process, ribosomal proteins, etc. (5, 6)]. In contrast, specific coevolution (not attributable to third proteins) is intuitively more closely related with functional relationships.

To separate these two types of coevolution, we calculated the partial correlation coefficient between every pair of proteins given each one of the others. The partial correlation between proteins A and B given protein N is calculated as:

Table 1. Composition of the protein interaction datasets used

	LT experiments	HT experiments	EcoCyc complexes	KEGG pathways	EcoCyc pathways
Pairs, *n*	3,965	53,002	1,354	78,532	4,491
Proteins, *n*	812	2,842	591	1,339	719

LT, low throughput; HT, high throughput.

$$\rho'_{AB.N} = \frac{r'_{AB} - r'_{AN} \cdot r'_{BN}}{\sqrt{(1 - r'^2_{AN}) \cdot (1 - r'^2_{BN})}},$$

where $\rho'_{AB.N}$ is the partial correlation between the coevolutionary profiles of proteins A and B, holding constant the coevolutionary profile of protein N, and r'_{AB}, r'_{AN}, and r'_{BN} are the Pearson's correlations between the coevolutionary profiles of A and B, A and N, and B and N, respectively (see above). For this work, only partial correlations with P values of $\leq 10^{-6}$ were considered significant.

Therefore, for each pair of proteins (A and B), we produce an ascending-sorted list of partial correlations with all other proteins. Going down this list (from low to high partial correlations), we move from highly specific coevolutions (first levels) to more relaxed coevolutions that can be partially explained by third proteins. This specificity relax procedure allows to retrieve those coevolutionary patterns shared by small groups of proteins (i.e., protein complexes), independent of the rest of the proteome evolution.

The sorted lists in Fig. 4 illustrate the final set of results that we obtained. Each list represents a pair of proteins, and each row represents a level of partial correlation. From these lists, we can extract the number of third proteins influencing our pair for a given partial correlation cutoff, which we call the ''partial correlation level.'' With the P value threshold mentioned above, we obtain significant partial correlations for 17,256 pairs in the 10th level (comprising 1,390 proteins).

Test Sets. We used the well studied model organism *E. coli* to evaluate the new method, to compare it with existing methodologies, and to assess whether it could give additional information on the functioning of protein complexes. A variety of different datasets of protein interaction and relationship with different characteristics were used for this purpose. Table 1 shows the number of interactions and proteins included in each test set. In each dataset, the set of negative pairs is constructed by forming all of the possible pairs among all of the proteins involved in the positive (interacting) pairs. We also performed some tests for yeast, which are described in the *SI Text*.

Low-throughput physical interactions. Binary physical interactions for *E. coli* were obtained from DIP (18), BIND (19), MINT (20), and Intact (21). The final set contains 3,965 interactions among 812 proteins. We considered only interactions coming from manually curated databases and low-throughput experiments, resulting in a small but highly reliable set of physical interactions.

Protein complexes. To test the accuracy of our method in predicting pairs of proteins belonging to the same macromolecular complex, we used the set of well characterized complexes available at EcoCyc (22). EcoCyc includes only manually curated data, and it is an extremely reliable source of functional information for *E. coli*. We retrieved 245 complexes that involve 591 proteins, which we translated into 1,354 binary relationships (all against all). Furthermore, we used protein complexes coming from high-throughput pull-down experiments (23, 24), although these data were less reliable than the previous set.

Metabolic pathways. We also compared our predictions with a relatively complete set of functional relationships given by copresence of proteins in the same metabolic pathways. We retrieved all of the pathways from both EcoCyc and KEGG (25) and translated them into binary functional interactions by considering a link between any pair of proteins belonging to same pathway. This results in a more relaxed type of functional relationship, not always related to a direct physical interaction.

The *SI Text* contains information on the availability of the software and on the predictions.

ACKNOWLEDGMENTS. We thank the members of the Computational Systems Biology Group and the Structural Bioinformatics Group for interesting discussions and support. This work was funded, in part, by Spanish Ministry for Education and Science Grants BIO2006-15318 and PIE 200620I240 and European Union Sixth Framework Programme Grant LSHG-CT-2003-503265, BioSapiens Network of Excellence.

1. van Kesteren RE, Tensen CP, Smit AB, van Minnen J, Kolakowski LF, Meyerhof W, Richter D, van Heerikhuizen H, Vreugdenhil E, Geraerts WP (1996) *J Biol Chem* 271:3619–3626.
2. Fryxell KJ (1996) *Trends Genet* 12:364–369.
3. Goh C-S, Bogan AA, Joachimiak M, Walther D, Cohen FE (2000) *J Mol Biol* 299:283–293.
4. Pazos F, Valencia A (2001) *Protein Eng* 14:609–614.
5. Pazos F, Ranea JAG, Juan D, Sternberg MJE (2005) *J Mol Biol* 352:1002–1015.
6. Sato T, Yamanishi Y, Kanehisa M, Toh H (2005) *Bioinformatics* 21:3482–3489.
7. Mintseris J, Weng Z (2005) *Proc Natl Acad Sci USA* 102:10930–10935.
8. Brandt U (2006) *Annu Rev Biochem* 75:69–92.
9. Manson M, Armitage J, Hoch J, Macnab R (1998) *J Bacteriol* 180:1009–1022.
10. Muir RE, Gober JW (2001) *Mol Microbiol* 41:117–130.
11. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B (2003) *Nucleic Acids Res* 31:258–261.
12. Kitano H (2002) *Science* 295:1662–1664.
13. Nurse P (2003) *Nature* 424:883.
14. Hakes L, Lovell SC, Oliver SG, Robertson DL (2007) *Proc Natl Acad Sci USA* 104:7999–8004.
15. Shoemaker BA, Panchenko AR (2007) *PLoS Comput Biol* 3:e43.
16. Edgar RC (2004) *Nucleic Acids Res* 32:1792–1797.
17. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003) *Nucleic Acids Res* 31:3497–3500.
18. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D (2002) *Nucleic Acids Res* 30:303–305.
19. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW (2001) *Nucleic Acids Res* 29:242–245.
20. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G (2002) *FEBS Lett* 513:135–140.
21. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, *et al.* (2007) *Nucleic Acids Res* 35:D561–D565.
22. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD (2005) *Nucleic Acids Res* 33:D334–D337.
23. Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, *et al.* (2005) *Nature* 433:531–537.
24. Arifuzzaman M, Maeda M, Itoh A, Nishikata K, Takita C, Saito R, Ara T, Nakahigashi K, Huang HC, *et al.* (2006) *Genome Res* 16:686–691.
25. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) *Nucleic Acids Res* 32:D277–D280.