doi:10.1016/j.jmb.2005.07.005 J. Mol. Biol. (2005) 352, 1002–1015

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Assessing Protein Co-evolution in the Context of the Tree of Life Assists in the Prediction of the Interactome

Florencio Pazos[1]*†, Juan A. G. Ranea[2], David Juan[3] and Michael J. E. Sternberg[1]

[1]*Structural Bioinformatics Group, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ UK*

[2]*Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, London WC1E 6BT UK*

[3]*Protein Design Group National Centre for Biotechnology (CNB-CSIC), Madrid Spain*

**Corresponding author*

The identification of the whole set of protein interactions taking place in an organism is one of the main tasks in genomics, proteomics and systems biology. One of the computational techniques used by many investigators for studying and predicting protein interactions is the comparison of evolutionary histories (phylogenetic trees), under the hypothesis that interacting proteins would be subject to a similar evolutionary pressure resulting in a similar topology of the corresponding trees. Here, we present a new approach to predict protein interactions from phylogenetic trees, which incorporates information on the overall evolutionary histories of the species (i.e. the canonical "tree of life") in order to correct by the expected background similarity due to the underlying speciation events. We test the new approach in the largest set of annotated interacting proteins for *Escherichia coli*. This assessment of co-evolution in the context of the tree of life leads to a highly significant improvement ($P(N)$ by sign test $\sim$10E–6) in predicting interaction partners with respect to the previous technique, which does not incorporate information on the overall speciation tree. For half of the proteins we found a real interactor among the 6.4% top scores, compared with the 16.5% by the previous method. We applied the new method to the whole *E. coli* proteome and propose functions for some hypothetical proteins based on their predicted interactors. The new approach allows us also to detect non-canonical evolutionary events, in particular horizontal gene transfers. We also show that taking into account these non-canonical evolutionary events when assessing the similarity between evolutionary trees improves the performance of the method predicting interactions.

Keywords: protein interactions; phylogenetic tree; co-evolution; interactome; horizontal gene transfer

Introduction

The network of protein–protein interactions of a given organism ("interactome") contains extensive information about its biology, since protein interactions are involved in most cellular processes (structural macromolecular complexes, signalling cascades, metabolism, transcription control, etc.). Experimental techniques for the determination of protein interaction, mainly variations of the yeast two-hybrid system[1] and affinity purification,[2,3] have been applied in a high-throughput approach aiming to cover as much as possible of the interactome of a number of model organisms.[4–7] Knowledge of the complete interactome allowed some of the first studies of biological networks from a systems biology point of view, extracting important data on the topology, connectivity and evolution of the protein interaction network.[8–14]

Several computational techniques have emerged to complement these experimental approaches. These computational techniques can be used to guide experiments restricting the number of pairs to test experimentally,[15] to filter a set of experimental

Present address: F. Pazos, Protein Design Group, National Center for Biotechnology (CNB-CSIC), Campus UAM, Cantoblanco, 28049 Madrid, Spain.

Abbreviations used: HGT, horizontal gene transfer; 16 S rRNA, ribonucleic acid molecule in the small subunit of the ribosome; ORF, open reading frame; GO, gene ontology.

E-mail address of the corresponding author: pazos@cnb.uam.es

interactions and to combine it with other information in order to increase its accuracy,[16] or to predict interactions purely *in silico*. Computational methods for the prediction of protein interactions, despite being cheaper and faster than their experimental counterparts, have been shown to have similar (or even higher) levels of accuracy than those, when combined under certain circumstances.[17] These methods are based on genomic or sequence features related with interaction: conservation of gene neighbouring across genomes,[18,19] domain fusion events,[20,21] comparison of phylogenetic distributions (patterns of presence/absence of genes in a set of genomes "phylogenetic profiles"),[22,23] correlated mutations,[24] and similarity of phylogenetic trees,[25,26] among others.[27–29]

The fact that interacting (or functionally related) proteins have similar phylogenetic trees had been observed qualitatively for some families of ligand–receptor pairs,[30,31] and later quantified and tested in large data sets of proteins and protein domains.[25,26] The concepts behind this approach are that interacting proteins bear a similar evolutionary pressure (since they are involved in the same cellular process), and that they are forced to adapt to each other, both resulting in similar evolutionary histories. This co-evolution has been observed not only in the sequences but also in the gene expression levels of the interacting proteins.[32]

The Goh & Cohen[25] and Pazos & Valencia[26] methods for predicting interactions are based on the comparison of protein distance matrices (using a linear correlation coefficient) instead of phylogenetic trees themselves. Since the exact comparison of phylogenetic trees is a complex and not fully solved problem, the comparison of distance matrices has been demonstrated to be a convenient shortcut useful for the special case of predicting protein interactions. This approach (*mirrortree*) has been subsequently applied to many protein families and followed by those who developed different implementations and variations of it.[33–38]

The degree of similarity between the phylogenetic trees of two proteins can be influenced by many factors beside the co-adaptation of the proteins, the main one being the underlying speciation process. Consequently, the trees of two proteins have a certain "background" similarity between themselves (regardless of whether the two proteins interact), and with the canonical "tree of life". On the other hand, non-standard evolutionary events (in particular horizontal gene transfer, HGT) leave landmarks in the trees of the proteins resulting in species far from where they should be in the canonical tree and close to species not related with them by the canonical phylogeny. Both factors affect the similarity between trees and its application for prediction of interactions.

Methods for detecting HGT look for abnormal signals of transferred genetic material into the Darwinian genomic background of species.[39–41] Most of the methods can be classified into three main categories based on the signal that they analyse: DNA composition bias such as base-pair frequencies or codon usage;[40,42] abnormal sequence similarities between distantly related species;[43] and phylogenetic incongruence of gene trees or species distributions.[44–46] While phylogenetic methods are more accurate to detect ancient transfer between distant species, those methods based on DNA composition bias are more appropriate for detecting and dating of more recent HGT events.[40]

The detection of these HGT events is also important for the evolution-based prediction of protein interactions: (i) a protein predicted to have undergone HGT is not expected to have overall similar trees with their interaction partners (except those that have undergone the same HGT event (see next paragraph)); (ii) correcting by the canonical tree of life (see below) is not appropriate for proteins that have undergone HGT, since their underlying speciation events have not followed that tree. For these reasons, it is important to detect HGT events prior to any evolution-based prediction of protein interactions.

A special class of HGT is a group of genes being transferred together.[47] Having undergone HGT means that they form a self-contained functional module (with limited dependence on other genes); and having been transferred together suggests that they are involved in the same (modular) function. Groups of genes with such behaviour include genes related to antibiotic resistance and single metabolic functions. This is related to the concept of "selfish operon".[47,48]

Here, we present a new approach for the co-evolution-based prediction of protein interactions which takes into account the information of the canonical tree of life (the one derived from the 16 S rRNA sequences) when assessing the similarity of evolutionary histories. The similarity between the trees of two proteins is corrected by this background similarity. The new method also allows us to concomitantly detect features related to non-standard evolutionary events, like HGT and modular cassettes of related genes, in order to take them into account when predicting interactions. We tested the method in the largest repository of annotated *Escherichia coli* interacting proteins available, statistically showing that it is considerably better than the previous approach on predicting protein interactions. We apply this new method in a blind test to the whole *E. coli* proteome, and propose functions for some hypothetical proteins based on their predicted interactors. We also demonstrate the applicability of the method for the detection of non-standard evolutionary events, and that this yields a better performance in interaction prediction.

Results

Prediction of interaction partners

We tested the method for the detection of interacting pairs of proteins in the whole set of *E. coli* annotated interacting pairs in DIP.[49] For each

protein we obtained a list of pairs sorted by the interaction score (r_{AB}) (see Materials and Methods). We evaluated the accuracy of the method by evaluating the percentage of false positives, that is, the percentage of proteins in the sorted list that score higher than the highest scoring real interaction partner. The lower this parameter, the better the prediction: a perfect method locating the true interaction at the top of the list would produce 0% false positives, whereas a random method would produce around 50% false positives (true hit in the middle of the list, on average). Additionally, we calculated the receiver operator characteristics (ROC) area derived from the list of each protein (see below).

Figure 1 shows the results for the test set (118 proteins). For most of the proteins, the real interactions are found at the top of the list. The average fraction of false positives for the whole set is 14.9% (median 6.4%). The number of pairs tested for each protein is also indicated in the Figure (187 on average).

To evaluate the improvement with respect to the previous version of the *mirrortree* approach[26] we applied that method to the same dataset obtaining a level of false positives of 23.4%. Thus, the fraction of false positives is reduced by 8.5% with this new method. An "intermediate" method, which uses the distances extracted from the phylogenetic trees (instead of percentages of sequence identity extracted directly from the multiple sequence alignments, as the old method does), but without correcting by the 16 S rRNA distances, yields 21.9% false positives. So, the improvement comes both from using distances extracted directly from the trees (substitutions/site) and from the correction with the 16 S rRNA distances, this last factor yielding the largest contribution to the improved accuracy (Table 1A) (see below).

For 23 of the 118 proteins (19.5%), the real interactor is the highest score (0% false positives) among the 166 (on average) pairs tested for each one (leftmost in Figure 1(a)). For 45% of the proteins, the level of false positives is lower than 5%. The score of the top hit can be used as a measure of confidence. If we restrict to proteins where that top score is $\geqq 0.97$ (70 out of the 118 cases), the level of false positives goes down to 12.4%. For the more restrictive cutoff of 0.98 (35 cases) that figure is 11.4%.

As an example of a prediction, the method is able to locate the right interaction between P00575 and P00577 (β and β′ chains of DNA-directed RNA polymerase; SwissProt accession numbers) as the first hit in the lists where both proteins were tested, composed of 232 and 222 proteins, respectively.

Only seven of the 118 proteins have scores worse than random (50%). The worst case (rightmost column in Figure 1(a)) is P15046 (acetate kinase) whose real interactor (P08839, phosphoenolpyruvate-protein phosphotransferase) is very low in the sorted list. The way in which we are evaluating false negatives is quite restrictive: the fact that a given interaction is not annotated in DIP does not mean that it is necessarily false. One example is P08374 (RNA polymerase omega), whose annotated interaction with RNA polymerase alpha is down in the list, which produces a high percentage of false positives, but which has a number of plausible interactors at the top of the list, including P00583 (DNA polymerase III beta), P16921 (transcription antitermination protein nusG) and other transcription factors.

The percentage of false positives provides a simple and intuitive way of evaluating the performance of the method. A more general and more formal way of quantifying this performance for methods producing a sorted list of scores is to calculate the area under an ROC curve. Those curves represent the relationship between sensitivity and specificity (also known as true-positive and false-positive rates, respectively) of a prediction method. They give an idea of the distribution of true and false hits in a list sorted by a score. ROC curves were calculated for the 118 proteins. A random method, applied to a set of proteins, would produce an average ROC area of 0.5 (regardless of the number of true hits in the lists), whereas a hypothetical perfect method would produce an ROC area of 1.0. For lists with only one positive (most of the cases, see Materials and Methods), the ROC area and the fraction of false positives provide the same measure (since ROC_area =1.0-frac._false_pos.). As we have more positives in the list, the probability of some of them to be high in the list increases by chance and hence the fraction of false positives decreases. This is why we introduce the ROC area, which is independent of the number

Table 1. Overview of *tol-mirrortree* performance and comparison with previous methods

A. *Average values of percentage of false positives and ROC area*

	% False positives	ROC area
mirrortree	23.4	0.71
mirrortree with tree distances	21.9	0.73
tol-mirrortree	14.9	0.79

Results are shown for the old *mirrortree* method; for the same method using distances extracted from the phylogenetic tree (instead of sequence identities); and for *tol-mirrortree*, which uses distances extracted from trees and corrects by the 16 S rRNA tree

B. *For each pair of methods, P-values of the one-sided sign test for the null hypothesis that there is no real differences in the performances between both methods and that the observed differences are happening by chance, against the alternative that method A is really better than method B*

	B		
A	*mirrortree*	*mirrortree* tree dist.	*tol-mirrortree*
mirrortree			
mirrortree (tree dist.)	0.276		
tol-mirrortree	5.60×10^{-6}	1.91×10^{-5}	

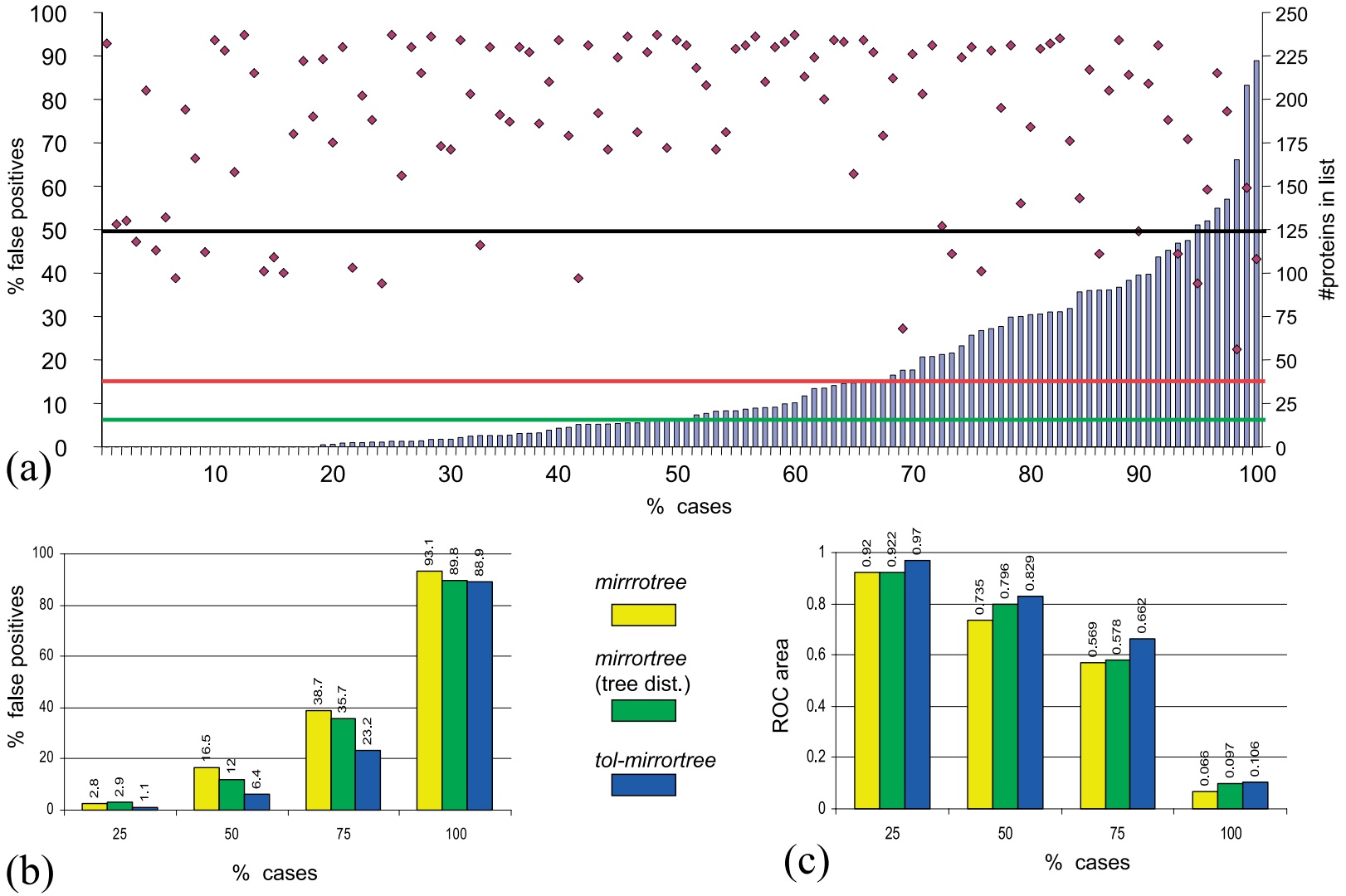

Figure 1. Results for the DIP dataset of *E. coli* interacting proteins. (a) Each of the 118 proteins is represented by a bar. The height of the bar represents the percentage of false positives, that is, the fraction of proteins with score higher than any annotated interactor. The random fraction of false positives (50%) is indicated with a black line. The average (14.9%) and the median (6.4%) are indicated with a red line and a green line, respectively. The number of pairs calculated for each protein is indicated with a dot. (b) Percentage of false positives a user has to accept (*Y* axis) in order to find the real interactor for a given fraction of the dataset (*X* axis). The equivalent numbers for the old *mirrortree* method, and for this old method with distances extracted from the phylogenetic tree are also shown. (c) Same representations as in (b) using the ROC area as a measure of accuracy.

of positives in the list, as an additional measure to consider these cases, whilst maintaining also the percentage of false positives as a more intuitive parameter. The new *tol-mirrortree* method produces an average ROC area of 0.79, whereas the corresponding figure for the old *mirrortree* is 0.71 (Table 1A): 12% of proteins produce an ROC area of 1.0 (perfect predictions); 93% have ROC areas higher than 0.5 (random).

A potential user of this method who aims to identify interactors for a large set of proteins (interactome) has to decide on a balance between accuracy of the prediction and number of cases predicted (coverage). Figure 1(b) shows the percentage of false positives the user has to accept in order to find a real interactor for a given fraction of the dataset. For example, in order to find a real interactor for half of the cases, a level of 6.4% false positives has to be accepted. The equivalent figure for the old *mirrortree* method is 16.5%. Figure 1(c) show the equivalent numbers using the ROC area as a measure of performance.

Although all the results presented here are protein-based (performance is evaluated for the list of each protein independently), we also constructed global ROC curves for both methods combining all the 19,991 pairs together regardless of the protein they come from (Figure 2(a)). These curves show that *tol-mirrortree* has a higher discriminative power than *mirrortree*, especially in the more important region of high scores (top of the lists). Interestingly, *mirrortree* shows a slightly better discriminative power, recovering positives down in the lists (low scores). The global ROC areas are 0.74 and 0.72, respectively. We also attached correlation values to some points in these curves, which allows one to obtain values of specificity and sensitivity from the raw scores of the methods in a hypothetical large-scale experiment. Figure 2(b) shows the same curves but using, instead of the correlation values themselves, the *z*-scores they have with respect to the rest of the pairs calculated for a given protein. The ROC areas in this case are 0.74 (*tol-mirrortree*) and 0.70 (*mirrortree*).

These case-by-case and global experiments illustrate two possible scenarios for the application of this method. First, when one is testing a protein of interest against a relatively small number of possible partners that include the real interactor(s). In this case, the user would expect the figures of percentage of false positives, etc. reported in Figure 1 and Table 1. The second scenario is closer to a high-throughput all-against-all experiment not focused on a given protein. Here, the user can obtain from the ROC curves of Figure 2 the expected values of specificity and sensitivity cutting the whole list of calculated pairs by a given correlation value, and play with this threshold depending on the number of positives he/she wants to recover and the negatives he/she can tolerate.

To assess whether the differences in the performances of *mirrortree* and *tol-mirrortree* are statistically significant or due to chance we performed a sign test[50] for each pair of methods based on the number of cases where one of the methods outperforms the other. To determine whether one method outperforms the other for each protein in turn, we calculated the average rank of the positive cases on the list sorted by score (see above) and we consider the wining method for this protein the one with highest average rank (positive cases closer to the top of the list). Since we are comparing, for a given protein, two lists with the same number of elements and the same number of positives, the average rank becomes a good figure for determining whether one method is better than the other for this protein. Table 1B shows the sign test *P*-values for the null hypothesis (i.e. there is no real difference between the two methods and the observed differences are happening by chance) against the alternative that one method is really better than the other. Those values show that *tol-mirrortree* is clearly better than the other two methods ($P(N)$ of the order of 10^{-5}–10^{-6}). The difference between the old *mirrortree* approach and the same method with tree distances is not statistically supported ($P(N) \sim 10^{-1}$).

Interaction-based function prediction

Predicted interaction partners can be used to assign function to hypothetical proteins. The concept behind this is that interacting proteins would have a similar function or would be involved in the same cellular process. This approach has been demonstrated to be able to assign function with accuracies ranging from 70% to almost 90%.[51,52] These works describe very sophisticated algorithms for inferring functions from interactions. Here, we used a simple implementation (see Materials and Methods).

Table 2 shows some results obtained for the blind test with hypothetical proteins described in Materials and Methods. In this case, gene ontology (GO)[53] terms shared between three or more of the top four predicted interactors of a given protein are assigned to it. Some GO terms are highly unspecific and do not provide interesting information about the protein (like GO:0016020, "membrane"). But for some other hypothetical proteins, the associated GO terms provide a clear picture of their possible cellular roles, like Q46920. This can be done by looking for shared GO terms in different fractions of top hits (see Table SI in Supplementary Data). We assessed some of the blind predictions presented in Table 2 by looking for functional suggestions about the proteins in public databases and resources. For four of the five cases with predicted specific functional features, the clues found in other resources would agree with these predictions. The results are given in Supplementary Data (Table SIV).

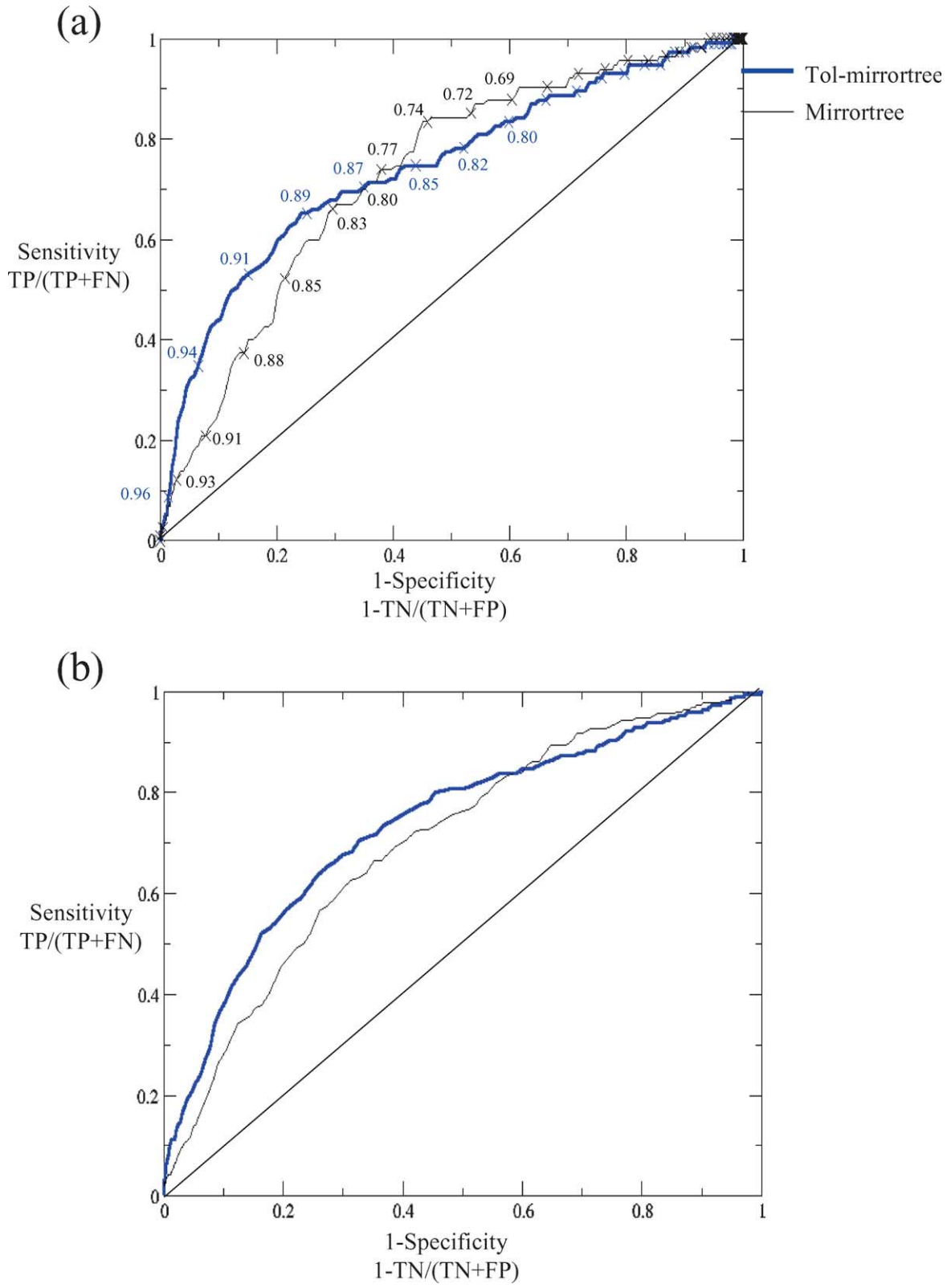

Figure 2. Global ROC curves for *mirrortree* and *tol-mirrortree*. The curves are constructed using the whole list of pairs (19,991). (a) The whole list of pairs is sorted according to the score of the methods (correlation coefficient). This coefficient is indicated for some points in the curves. (b) Curves based on *z*-scores, instead of raw correlation values. "Sensitivity" and "1-Specificity" can also be interpreted as true-positive and false-positive rates, respectively. TP, true positives; FN, false negatives; TN, true negatives; FP, false positives.

Table 2. Predicted functions for hypothetical proteins based on their predicted interaction partners

Protein	Four top predicted interactors	Shared GO terms (> =3)
P45395	**P52648 P25888 P31219 P13030**	GO:0005524 ATP binding
P37596	**P07813 P04805 P00954 P06710**	GO:0005524 ATP binding
Q46920	**P76182 P77611 P76181** P13652	GO:0009399 nitrogen fixation GO:0006118 electron transport
Q46827	P32703 **P31060 P77265 P09980**	GO:0005524 ATP binding
P77645	**P76181 P77179 P75706** P05852	GO:0016020 membrane
P77481	**P77265 P07671 P07025** P77415	GO:0005524 ATP binding
P77481	P77265 **P07671 P07025 P77415**	GO:0003677 DNA binding
P76258	**P15044** P37631 **P52648 P09833**	GO:0005524 ATP binding
P75959	**P43672** P45465 **P37345 P17888**	GO:0005524 ATP binding
P45528	**P77179 P04983 Q46821** P39099	GO:0016020 membrane
P39874	P00373 **P37053 P76270 P77334**	GO:0005554 molecular_function unknown
P39290	**P76181 P76182** P27248 **P00837**	GO:0016020 membrane
P37596	**P07813 P04805 P00954** P06710	GO:0006418 tRNA aminoacylation for protein translation GO:0004812 tRNA ligase activity
P37027	**P21499** P17580 **P08576 P33398**	GO:0003676 nucleic acid binding
P36880	P30870 **P27300 P20082 P36879**	GO:0005524 ATP binding
P31805	P77475 **P31060 P00804 P23886**	GO:0016020 membrane
P28634	**P11880 P07862** P25539 **P22188**	GO:0005737 cytoplasm
P22186	P39341 **P23859 P23858 P23860**	GO:0016020 membrane GO:0006810 transport
P09997	**P06609** P02918 **P77173 P75958**	GO:0016020 membrane

E. coli hypothetical proteins for which three or more of the top four predicted interaction partners share some GO term. Proteins are labelled with their SwissProt accession numbers. The top four predicted interactors are indicated. The predicted interactors sharing that these term(s) are marked in bold type. The shared GO term(s) and their descriptions are also shown.

Detection of horizontal transfer and implications for interaction prediction

We illustrate the ability of the method to detect HGT events using the prototypical and well-studied case of the aminoacyl-tRNA synthetases and ribosomal proteins. Both protein sets are involved in the translation machinery and perform essential and universal functions in bacteria. However, while ribosomal protein phylogenies are usually consistent with the accepted overall phylogeny of species, tRNA synthetases typically violate the accepted taxonomic structure.[54–56] Distances extracted from the protein trees of the 24 *E. coli* tRNA synthetases and the 56 ribosomal proteins were plotted against the distances extracted from the 16 S rRNA tree. The analysis of those plots reveals that a number of tRNA synthetases probably underwent horizontal transfer whereas most ribosomal genes were vertically inherited. For example, the prolyl-tRNA synthetase (P16659) distance plot shows a clear HGT signature (Figure 3(a)), reflected in a low linear correlation coefficient (0.53). This HGT event is supported by independent observations.[56] The upper cloud of dots in the plot, caused by five species only, indicates a transfer from "outside" the sample of bacteria used in this work. A BLAST search[57] confirms that the prolyl-tRNA synthetases in these five organisms are more similar to the corresponding eukaryotic and archaeal tRNA synthetases (E-values of the order of 10^{-100}) than to those of the bacteria used in this work ($\sim 10^{-10}$). On the other hand, ribosomal proteins show a higher correlation with the expected phylogenetic distances (for example, ribosomal protein L36, P21194, Figure 3(b), $r = 0.72$) and comparatively a lower mutational rate (compare y-axis in Figure 3(a) and (b)).

Detection of HGT cases before evolution-based interaction prediction is important, since these proteins are expected to produce bad results in these methods (see Introduction). Indeed, the level of false positives in the protein interaction dataset (previous section) increases from 15% to 25% for the 12 proteins whose correlation with the 16 S rRNA is lower than 0.5 (possible HGT cases in the context of this method), whereas this figure decreases to 13.7% for the remaining 106 proteins (possible non-HGT cases).

Co-occurred HGT events can point to pairs of related proteins involved in "transferable" or modular functions. It is known that functionally related groups of genes (like operons) are prone to be transferred together.[47] We tried to detect those cases by looking for pairs of proteins with high correlation between their tree topologies and, at the same time, low correlation with the canonical tree (see Materials and Methods). The 40 pairs with highest scores were selected for further analysis (see Table SII in Supplementary Data). The presence in some cases of the same protein in more than one pair allowed us to group the pairs into 14 related clusters (see Table SIII in Supplementary Data). Some of the proteins are annotated as hypothetical, and for some cases the function characterization is not clear. A total of 20 of the 40 selected HGT pairs present, at least, one membrane transport protein involved, amongst other functions, in sugar and ion fluxes, or in bacterial drug resistance. One interesting case is the large cluster where the principal hub corresponds to the chaperone *hscA*. Although the function of this chaperone is unknown, it has been

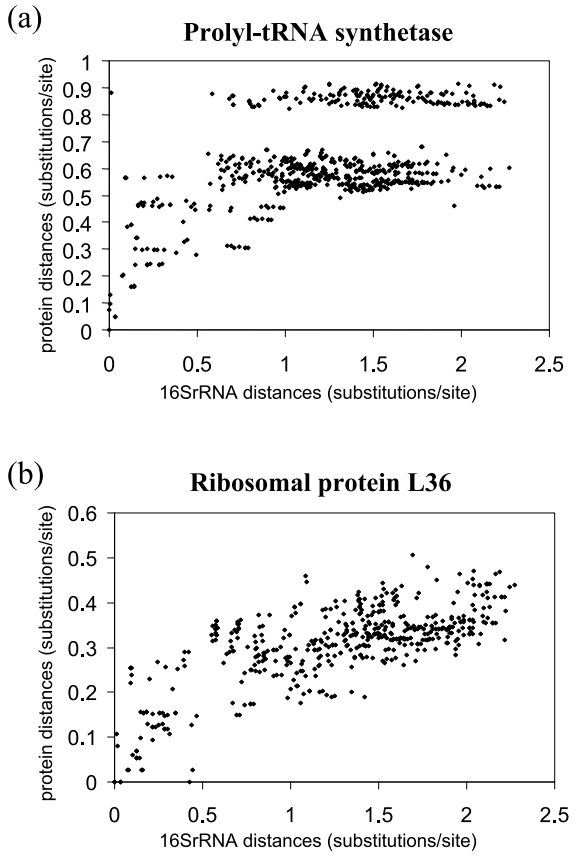

Figure 3. Comparing the distances extracted from the protein trees with the distances extracted from the 16 S rRNA tree to detect HGT. Plots are shown for prolyl-tRNA synthetase (SwissProt acc. no. P16659) and ribosomal protein L36 (P21194). In the first case there is a clear indication of HGT that had already been reported by other authors.[56]

suggested that it could be involved in the assembly of proteins containing iron-sulphur centers (Fe/S).[58] In this cluster *hscA* is related to seven different protein partners all involved in translation (ribosomal proteins, tRNA synthetases, etc.). These results could yield new suggestions about a possible functional role of the chaperone *hscA* in translation. For some of these clusters, their members are adjacent or close in the genome of *E. coli* (see Figure S1 in Supplementary Data), which would support the relationship between those co-occurred HGT events and "selfish operons"[47]).

Discussion

Here, we propose a new method for the prediction of protein interactions based on the detection of similar evolutionary histories. This method incorporates information on the canonical speciation tree (the one based on the 16 S rRNA sequences) to correct by the background similarity expected between any two proteins due to the underlying speciation events, allowing also to concomitant detection of non-standard evolutionary events.

Even with the simple correction introduced (subtracting phylogenetic distances), the results of this new version on predicting interactions are significantly better than the previous method, as statistically demonstrated using as test set the largest repository of *E. coli* annotated interactions available. Using distances extracted from the phylogenetic trees (substitutions/site) instead of percentages of sequence identity extracted directly from the multiple sequence alignments also contributes to the improvement, although to a lesser extent.

In the test set used, we can be confident about the positives (annotated interactions) but not about the negatives (the remaining pairs), since many of them can be real interactions not yet discovered or annotated. Thus, possibly the accuracy figure obtained is a lower limit of the true value. Another consequence of working with this dataset is that we could only evaluate proteins for which we have at least one interactor. Hence, the accuracy figures reported here are applicable to only those cases. In other words, we have to know *a priori* that a given protein is interacting with anything before applying this method expecting the reported accuracy values. Estimations for the yeast proteome suggest that this is the case for almost all proteins (to be involved in one or more interactions).[59]

One advantage of this method with respect to other approaches for predicting protein interactions, is that it does not require fully sequenced genomes to work, as other evolution-based approaches do (e.g. phylogenetic profiles[22]).

The incorporation of information on the standard tree of life allows the automatic detection of non-standard evolutionary events in a concomitant way with the prediction of interactions. In detecting HGT events, our method falls in the category of approaches based on "phylogenetic incongruence". A method for detecting HGT has been recently porposed by Farahi *et al.*,[46] which has some similarities to ours, since it also relies on matrices of evolutionary distances. The main difference is that the Farahi *et al.* method uses ribosomal protein trees, instead of the more standard 16 S rRNA tree, in order to establish a vertical inheritance model to compare against.

Pairs of proteins predicted to have undergone HGT and, at the same time, correlated to each other are candidates to form part of independent transferable functional modules or selfish operons.[47,48] We detected some of these cases, including proteins related with drug resistance. Drug resistance is the prototypical case of modular transferable independent function. This function requires a minimum number of genes (for example, the subunits of a membrane pump) that are independent and do no interfere with other cellular processes, and hence can be transferred and "accepted" as a whole. To detect these modules is

important not only because of the eventual practical applications (antibiotic resistance) but because of the theoretical implications as well ("isolated" functional modules, highly independent on others, and hence easier to model).

The detection of these non-canonical evolutionary histories, in addition to being important by itself, is also critical when assessing interactions by evolution-based methods, as discussed in the Introduction. Indeed, our results are better when discarding proteins automatically predicted to have suffered HGT (see Results). This method is not intended to compete with the many existing methods specifically designed to locate HGT (see Introduction), but to detect it in a simple and convenient way, naturally integrated with the interaction prediction methodology.

Finding the right interaction partner among the 15% top scores (15% false positives) could be considered insufficient for a researcher interested in a given protein. However, we consider that this performance is useful for two main reasons: (i) it implies an important reduction in the number of pairs to test experimentally; and (ii) for certain studies of the interactome that deal with global properties such as topology or connectivity,[8–14] a high rate of false positives can be tolerated, since these studies are statistical in nature and deal with global properties not markedly affected by the detailed description of the individual components. Indeed, high-throughput experimental sets of interactions have a very high degree of error[17] and they are still used for many studies.[8–14] So, this method can help, together with other computational and experimental techniques, in the *in silico* reconstruction of protein interaction networks.

Another application of the predicted interaction partners is the prediction of function for open reading frames (ORFs) by simply transferring function from their interactors.[51,52] This is an approach orthogonal and complementary to the standard function transfer by sequence homology.

Another factor that could affect the similarity of trees of interacting proteins is the difference in evolutionary rates. A certain relation has been found between this parameter and the number of interactors a protein has to co-evolve with,[9] being highly connected proteins with slightly lower evolutionary rates. We will explore the possibility of extending the *tol-mirrortree* method in order to take this factor into account.

Although very fast and convenient, "neighbour joining" is not the state-of-the-art technique for constructing phylogenetic trees. We plan to assess the performance of the method using phylogenetic trees obtained with more sophisticated techniques. Moreover, although used by many investigators as a very convenient and practical shortcut, which has been shown to achieve a good performance in predicting interactions (see Introduction), the usage of a correlation formulation to compare distance matrices is not very robust from the mathematical point of view, since the values (distances) are not independent. We plan to study more exhaustive, albeit practical, ways to compare evolutionary histories. This dependence also makes it impossible to associate tabulated *P*-values (which would be easier to interpret) to the correlation scores, since these *P*-values are based on a null hypothesis, which involves the independence of the data. It would be useful to explore the possibility of constructing background distributions, which do take into account the intrinsic dependence of distance matrices data (i.e. from random trees) to extract *P*-values from them.

The evolutionary assumptions made by *mirrortree*-like approaches (i.e. co-evolution of interacting proteins; see Introduction) were the basis for generating and improving the methodologies. But the methods and their performances do not depend on these assumptions to be true. The fact is that similarity of phylogenetic distances is related to interaction. This observation can be used to derive putative interactions, irrespective of the underlying evolutionary assumptions. Nevertheless, the value of working with these assumptions is that they can lead to future improvements and produce scientific knowledge beyond just a black-box predictive tool.

The method presented here allows the user to study and compare evolutionary histories in an integrated framework, for predicting protein interactions, non-standard evolutionary histories and protein function. Since it is fully automatic and requires only sequence information to work, it can be coupled to the continuous stream of new sequences coming from the many whole-genome sequencing projects.

Materials and Methods

A schema of the *tol-mirrortree* (tree-of-life-mirrortree) method is shown in Figure 4. In summary, phylogenetic trees for all *E. coli* proteins are constructed based on multiple sequence alignments of orthologous sequences. The trees are then converted into distance matrices. The standard 16 S rRNA tree is similarly converted into a distance matrix. Comparison of these three distance matrices gives information on the evolutionary histories of the proteins. New matrices for the proteins are constructed corrected by the 16 S rRNA distances. These correcting matrices are compared to assess the possible interaction between the two proteins.

Generation of the multiple sequence alignments, phylogenetic trees and distance matrices

Multiple sequence alignments for all *E. coli* K12 proteins were generated looking for their orthologues in 43 fully sequenced prokaryotic genomes and aligning them with ClustalW[60] (default parameters). Orthologues were detected using the standard bi-directional BLAST[57] procedure with an *E*-value cut-off of 10^{-5}. That is, a given *E. coli* protein is BLASTed against the sequences in another genome, and the top hit is taken as the orthologue of the original *E. coli* protein in that genome only if the *E*-value is above the cut-off and a "reverse" BLAST of that

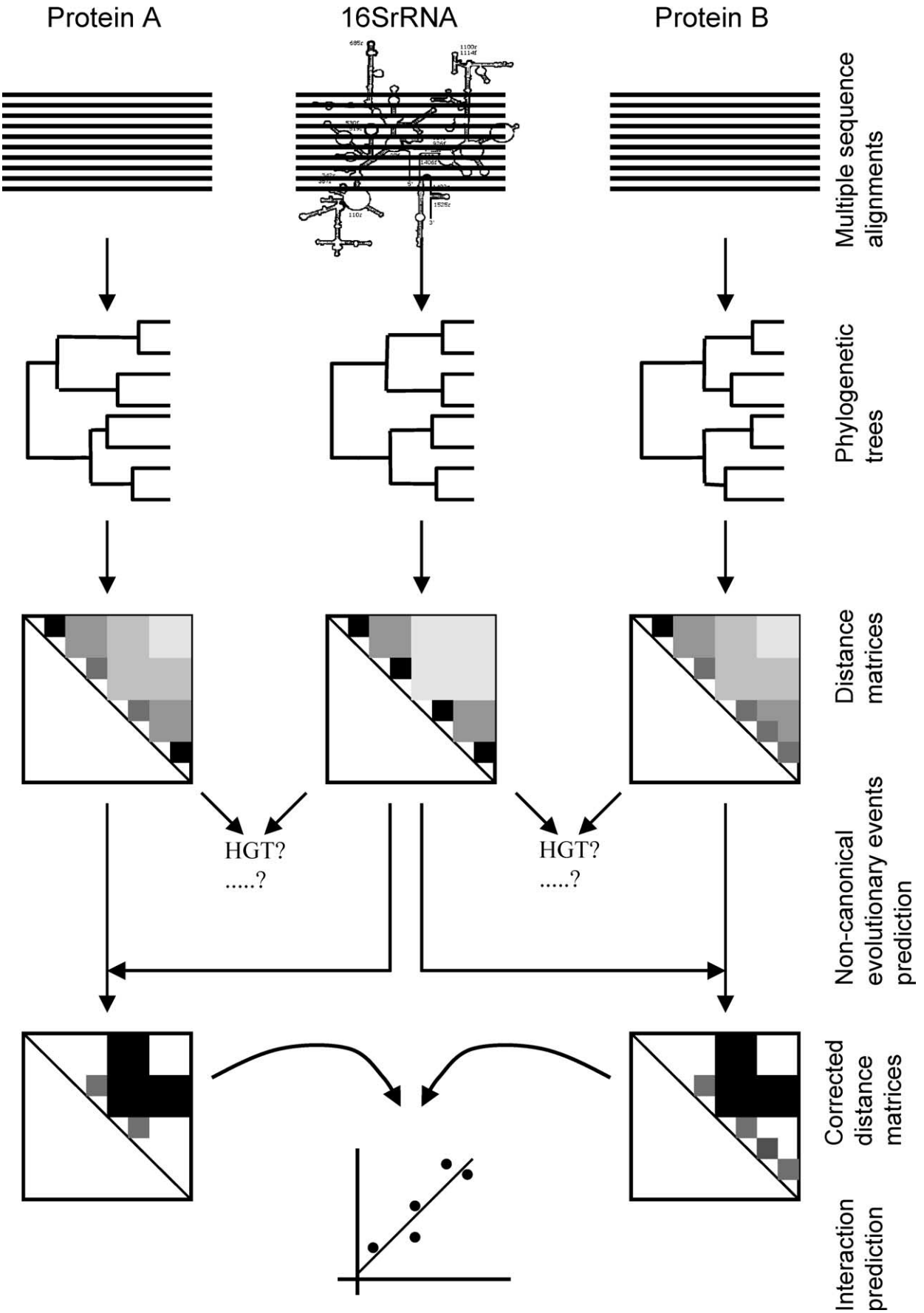

Figure 4. Schema of the *tol-mirrortree* method. Phylogenetic trees are built from the multiple sequence alignments of the proteins. Matrices containing the distances between species are extracted from those trees. The distances between the

protein back against *E. coli* also finds the original protein as the top hit (again with an *E*-value above the cut-off).

Phylogenetic trees are derived from the multiple sequence alignments using the neighbour-joining algorithm implemented in ClustalW. These trees are converted into distance matrices by summing the length of the branches separating each pair of species. The canonical tree of life, constructed from the 16 S rRNA sequences, was obtained from the European rRNA Database.[61] This 16 S rRNA tree is converted into a distance matrix in the same way. New distance matrices for the proteins are obtained by subtracting from each value the distance between the corresponding species in the 16 S rRNA distance matrix. Due to the different scale of protein and RNA distances matrices, their values are rescaled before subtraction. For this rescaling, we need an equivalence between RNA and protein distances. That equivalence can be obtained from "molecular clock" proteins, proteins expected to reflect the same evolutionary history as the "standard tree" (16 S rRNA). For that, we took the proteins with the trees most similar to the 16 S rRNA (highest correlations). For these proteins the relation distance_protein/distance_RNA was, on average, 0.42/1. We rescaled the values of the matrices with these figures before subtraction. The final corrected matrices are expected to contain only the distances between orthologues that are not due to speciation but to other reasons related to function.

Prediction of interacting pairs

An interaction score is obtained by calculating the linear correlation coefficient between the two corrected matrices. In order for the matrices to have the same dimension and to be comparable, only the distances between species that are in the multiple sequence alignment of both proteins are considered. We require a minimum of ten species in common (45 distance values) to test a pair of proteins. So, for two proteins A and B with n species in common in their multiple sequence alignment, being dA_{ij} the distance between species i and j in the tree of protein A, dB_{ij} the distance in the tree of protein B (both rescaled as explained before), and dR_{ij} the distance between species i and j in the standard 16 S rRNA tree; the interaction score between A and B (r_{AB}) is calculated as:

$$r_{AB} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(dA'_{ij}-\overline{dA}')\cdot(dB'_{ij}-\overline{dB}')}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(dA'_{ij}-\overline{dA}')^2}\cdot\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(dB'_{ij}-\overline{dB}')^2}}$$

Where dA'_{ij} and dB'_{ij} are the values of the corrected matrices, that is $dA'_{ij} = dA_{ij} - dR_{ij}$; $dB'_{ij} = dB_{ij} - dR_{ij}$; and $\overline{dA}'$, $\overline{dB}'$ are the average values of those matrices. High values of that score reflect similar evolutionary histories not due to the overall speciation and hence are expected to be related to physical interaction or functional relationship. Hence, the two differences between *tol-mirrortree* and the old *mirrortree* method are: (i) the use of distances extracted from phylogenetic trees instead of sequence similarities; and (ii) the correction by the 16 S rRNA distances.

Protein interactions test set

We tested this method for predicting protein interactions in the whole set of *E. coli* interacting proteins annotated in the DIP database[49] (version of February 2004). This database contains interactions manually extracted from the literature and, except for organisms for which more specialised databases exist, it can be considered the current gold standard for protein interactions. There were 516 pairs of interacting pairs, comprising 512 different *E. coli* proteins. This test set contains only physically interacting proteins, not functionally related ones. In order to get a set of negative examples involving the same proteins, we generated all possible pairs between those proteins. This set of negative examples may actually erroneously include positive examples, since the fact a pair of proteins is not annotated in DIP does not guarantee that these two proteins do not interact. In other words, we can be sure of the positives but not of the negatives. Due to the limitation of requiring ten or more species in common in order for two proteins to be tested, our final test set contained 19,991 pairs, 115 of them being true interactions. There were 118 proteins for which at least one pair comprising a real interactor could be calculated. From these 118 cases, 80 (68%) have only one annotated interactor in the list, 20 (17%) have two, and the remaining 18 (15%) have three or more.

We acknowledge that this dataset is limited, since there is no experimental high-throughput protein interaction data for *E. coli*. This incompleteness creates some problems in evaluating the results (especially false positives) discussed elsewhere in the article. To use other organisms with a far larger coverage of experimental protein interaction data (such as yeast) would overcome these problems. There are two main reasons to restrict to prokarya and not use the available eukaryotic interaction datasets: (i) some characteristics of eukaryotic proteins (including multidomain and low-complexity regions) make the automatic generations of reliable multiple sequence alignments more difficult than for prokarya; (ii) while the phylogenetic tree for bacteria is more or less well established, the one involving eukarya is not so clear (especially the separation between the three kingdoms). Since the tree of life is an input for this method, we wanted to restrict to a well-established one. Moreover, concepts like operon or genome closeness used in this study make sense only for prokaryotic organisms. Additionally, for *E. coli* we can produce blind predictions, which could be useful for further investigation, since not all the interactions are known. For these reasons, we decided to use *E. coli* as a simpler model organism in spite of the relatively limited protein interaction data.

Interaction-based prediction of function

As a blind prediction, we ran all-against-all *E. coli* proteins. Due to the limitations described, we could make

same species are also extracted from the canonical tree of life, the one based on the 16 S rRNA sequences. Comparison of the protein trees (distance matrices) with the 16 S rRNA tree gives information on non-standard evolutionary events, like horizontal gene transfers (HGT). The distances in the protein matrices are corrected by those in the 16 S rRNA. These new corrected distances matrices are compared using a linear correlation criteria to assess the possible interaction between the two proteins.

calculations for 836,934 pairs. For the proteins annotated as "hypothetical", we took the top list of predicted interaction partners, extracted their functional annotations from the GO database[53] and assigned to the hypothetical protein those GO terms shared by its predicted interactors.

Detection of non-standard evolutionary events

Non-standard evolutionary events are detected by comparing the trees (distance matrices) of the two proteins, between them and with the 16 S rRNA tree. For this purpose we use the original distance matrices (before rescaling and correcting) (see Figure 4).

Proteins predicted to have undergone HGT events are those whose trees are different from the 16 S rRNA tree (poorly correlated, low r_{AR}). Modular cassettes of functionally related proteins are composed of proteins with a high degree of similarity between their trees but a low degree of similarity with the 16 S rRNA tree, thus indicating a possible joint HGT due to functional reasons. To quantify this we calculated, for all possible pairs in *E. coli* the difference between the correlation of the trees of the two proteins and the average correlation of these proteins with the 16 S rRNA tree ($r_{AB}-(r_{AR}+r_{BR})/2$). In this case correlation values (r) are calculated from the original matrices, before correcting with the 16 S rRNA distances (Figure 4).

Acknowledgements

We acknowledge Alfonso Valencia (CNB, CSIC), Michael Stumpf (Division of Molecular Biosciences, IC), Victor Lesk (SBG, IC) and Hernan Dopazo (CNIO) for interesting discussion and comments, Dr Elena Kulinskaya (Statistical Advice Service, IC) for help and Eduardo Andrés (CNB, CSIC) for computer assistance. We also acknowledge the maintainers of the databases and resources used in this work and one of the anonymous referees for fair and constructive comments. F.P. is supported by the DTI beacon project QCBB/C/012/00003. All computer programs and datasets discussed in this article are available upon request for academic use.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2005.07.005

References

1. Fields, S. & Song, O. (1989). A novel genetic system to detect protein–protein interactions. *Nature*, **340**, 245–246.
2. Gavin, A. C., Bösche, M., Krause, R., Grandi, P., Marcioch, M., Bauer, A. *et al*. (2002). Functional organisation of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
3. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L. *et al*. (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.
4. Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R. *et al*. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–631.
5. Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S. *et al*. (2001). The protein–protein interaction map of *Helicobacter pylori*. *Nature*, **409**, 211–215.
6. Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y. *et al*. (2003). A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.
7. Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M. *et al*. (2004). A map of the interactome network of the metazoan *C. elegans*. *Science*, **303**, 540–543.
8. Jeong, H., Mason, S. P., Barabási, A. L. & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
9. Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. & Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science*, **296**, 750–752.
10. Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H. *et al*. (2003). Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucl. Acids Res.* **31**, 2443–2450.
11. Kelley, B. P., Sharan, R., Karp, R. M., Sittler, T., Root, D. E., Stockwell, B. R. *et al*. (2003). Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl Acad. Sci. USA*, **100**, 11394–11399.
12. Qin, H., Lu, H. H., Wu, W. B. & Li, W. H. (2003). Evolution of the yeast protein interaction network. *Proc. Natl Acad. Sci. USA*, **100**, 12820–12824.
13. Wuchty, S., Oltvai, Z. N. & Barabasi, A. L. (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genet.* **35**, 176–179.
14. Yeger-Lotem, E. & Margalit, H. (2003). Detection of regulatory circuits by integrating the cellular networks of protein–protein interactions and transcription regulation. *Nucl. Acids Res.* **31**, 6053–6061.
15. Lappe, M. & Holm, L. (2004). Unraveling protein interaction networks with near-optimal efficiency. *Nature Biotechnol.* **22**, 98–103.
16. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S. *et al*. (2003). A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science*, **302**, 449–453.
17. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. *et al*. (2002). Comparative assessment of large scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.
18. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. (1999). Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol.* **1**, 93–108.
19. Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physicaly interact. *Trends Biochem. Sci.* **23**, 324–328.
20. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
21. Marcotte, E. M., Pellegrini, M., Ho-Leung, N., Rice,

D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
22. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein pylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
23. Date, S. V. & Marcotte, E. M. (2003). Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnol.* **21**, 1055–1062.
24. 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.
25. Goh, C.-S., Bogan, A. A., Joachimiak, M., Walther, D. & Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* **299**, 283–293.
26. Pazos, F. & Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.* **14**, 609–614.
27. Salwinski, L. & Eisenberg, D. (2003). Computational methods of analysis of protein–protein interactions. *Curr. Opin. Struct. Biol.* **13**, 377–382.
28. Valencia, A. & Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.* **12**, 368–373.
29. Huynen, M. A., Snel, B., von Mering, C. & Bork, P. (2003). Function prediction and protein networks. *Curr. Opin. Cell Biol.* **15**, 191–198.
30. Fryxell, K. J. (1996). The coevolution of gene family trees. *Trends Genet.* **12**, 364–369.
31. Pages, S., Belaich, A., Belaich, J. P., Morag, E., Lamed, R., Shoham, Y. *et al*. (1997). Species-specificity of the cohesin-dockerin interaction between *Clostridium thermocellum* and *Clostridium cellulolyticum*: prediction of specificity determinants of the dockerin domain. *Proteins: Struct. Funct. Genet.* **29**, 517–527.
32. Fraser, H. B., Hirsh, A. E., Wall, D. P. & Eisen, M. B. (2004). Coevolution of gene expression among interacting proteins. *Proc. Natl Acad. Sci. USA*, **101**, 9033–9038.
33. Gertz, J., Elfond, G., Shustrova, A., Weisinger, M., Pellegrini, M., Cokus, S. *et al*. (2003). Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics*, **19**, 2039–2045.
34. Goh, C. S. & Cohen, F. E. (2002). Co-evolutionary analysis reveals insights into protein–protein interactions. *J. Mol. Biol.* **324**, 177–192.
35. Ramani, A. K. & Marcotte, E. M. (2003). Exploiding the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.* **327**, 273–284.
36. Sato, T., Yamanishi, Y., Horimoto, K., Toh, H. & Kanehisa, M. (2003). Prediction of protein–protein interactions from phylogenetic trees using partial correlation coefficient. *Genome Informatics*, **14**, 496–497.
37. Kim, W. K., Bolser, D. M. & Park, J. H. (2004). Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics*, **20**, 1138–1150.
38. Tan, S., Zhang, Z. & Ng, S. (2004). ADVICE: automated detection and validation of interaction by co-evolution. *Nucl. Acids Res.* **32**, W69–W72.
39. Brown, J. R. (2003). Ancient horizontal gene transfer. *Nature Rev. Genet.* **4**, 121–132.
40. Lawrence, J. G. & Ochman, H. (2002). Reconciling the many faces of lateral gene transfer. *Trends Microbiol.* **10**, 1–4.
41. Philippe, H. & Douady, C. J. (2003). Horizontal gene transfer and phylogenetics. *Curr. Opin. Microbiol.* **6**, 498–505.
42. Nakamura, Y., Itoh, T., Matsuda, H. & Gojobori, T. (2004). Biased biological functions of horizontally transfered genes in prokaryotic genomes. *Nature Genet.* **36**, 760–766.
43. Clarke, G. D., Beiko, R. G., Ragan, M. A. & Charlebois, R. L. (2002). Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *J. Bacteriol.* **184**, 2072–2080.
44. Daubin, V., Moran, N. A. & Ochman, H. (2003). Phylogenetics and the cohesion of bacterial genomes. *Science*, **301**, 829–832.
45. Kunin, V. & Ouzounis, C. A. (2003). GeneTRACE-reconstruction of gene content of ancestral species. *Bioinformatics*, **19**, 1412–1416.
46. Farahi, K., Whitman, W. B. & Kraemer, E. T. (2003). RED-T: utilizing the ratios of evolutionary distances for determination of alternative phylogenetic events. *Bioinformatics*, **19**, 2152–2154.
47. Lawrence, J. G. (1997). Selfish operons and speciation by gene transfer. *Trends Microbiol.* **5**, 355–359.
48. Omelchenko, M. V., Makarova, K. S., Wolf, Y. I., Rogozin, I. B. & Koonin, E. V. (2003). Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. *Genome Biol.* **4**, R55.
49. Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M. & Eisenberg, D. (2002). DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucl. Acids Res.* **30**, 303–305.
50. Bland, M. (1987). *An Introduction to Medical Statistics*, Oxford Medical Publications, Oxford University Press, London.
51. Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nature Biotechnol.* **21**, 697–700.
52. Samanta, M. P. & Liang, S. (2003). Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc. Natl Acad. Sci. USA*, **100**, 12579–12583.
53. Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R. *et al*. (2004). The Gene Ontology (GO) database and informatics resource. *Nucl. Acids Res.* **32**, D258–D261.
54. Brown, J. R. & Doolittle, W. F. (1995). Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc. Natl Acad. Sci. USA*, **92**, 2441–2445.
55. Shiba, K., Motegi, H. & Schimmel, P. (1997). Maintaining genetic code through adaptations of tRNA synthetases to taxonomic domains. *Trends Biochem. Sci.* **22**, 453–457.
56. Woese, C. R., Olsen, G. J., Ibba, M. & Soll, D. (2000). Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev.* **64**, 202–236.
57. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. *et al*. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
58. Dougan, D. A., Mogk, A. & Bukau, B. (2002). Protein

folding and degradation in bacteria: to degrade or not to degrade? That is the question. *Cell. Mol. Life Sci.* **59**, 1607–1616.
59. Grigoriev, A. (2003). On the number of protein-protein interactions in the yeast proteome. *Nucl. Acids Res.* **31**, 4157–4161.
60. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. *et al*. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucl. Acids Res.* **31**, 3497–3500.
61. Wuyts, J., Perriere, G. & Van De Peer, Y. (2004). The European ribosomal RNA database. *Nucl. Acids Res.* **32**, D101–D103.

Edited by J. Thornton

(Received 20 January 2005; received in revised form 22 June 2005; accepted 4 July 2005)
Available online 15 August 2005