Protein Engineering vol.14 no.9 pp.609–614, 2001

Similarity of phylogenetic trees as indicator of protein–protein interaction

Florencio Pazos and Alfonso Valencia[1]

Protein Design Group, CNB–CSIC, Cantoblanco, E-28049 Madrid, Spain

[1]To whom correspondence should be addressed.
E-mail: valencia@cnb.uam.es

Deciphering the network of protein interactions that underlines cellular operations has become one of the main tasks of proteomics and computational biology. Recently, a set of bioinformatics approaches has emerged for the prediction of possible interactions by combining sequence and genomic information. Even though the initial results are very promising, the current methods are still far from perfect. We propose here a new way of discovering possible protein–protein interactions based on the comparison of the evolutionary distances between the sequences of the associated protein families, an idea based on previous observations of correspondence between the phylogenetic trees of associated proteins in systems such as ligands and receptors. Here, we extend the approach to different test sets, including the statistical evaluation of their capacity to predict protein interactions. To demonstrate the possibilities of the system to perform large-scale predictions of interactions, we present the application to a collection of more than 67 000 pairs of *E.coli* proteins, of which 2742 are predicted to correspond to interacting proteins.

Keywords: bioinformatics/co-evolution/phylogenetic tree/protein interaction/proteomics

Introduction

The reconstruction of the network of protein–protein interactions is essential for the study of the dynamic properties of cellular systems. Such an interaction network would include key systems such as metabolic pathways, signaling cascades and transcription control networks. New and powerful experimental techniques, such as the Yeast Two-Hybrid System, are already tackling this problem systematically (Mendelsohn and Brent, 1999). Indeed, the first genome-scale results are already available: between 183 and 280 experimentally determined interactions in yeast (Ito *et al.*, 2000; Uetz *et al.*, 2000) and 261 in *Helicobacter pylori* (Rain *et al.*, 2001).

In parallel with these developments, a number of computational techniques have been designed for predicting protein interactions from the information contained in sequences and genomes (Dandekar *et al.*, 1998; Enright *et al.*, 1999; Marcotte *et al.*, 1999a,b; Pellegrini *et al.*, 1999). These computational techniques still have a limited range of applicability; for example, Enright *et al.* predicted a total of 64 interactions in three bacterial genomes (Enright *et al.*, 1999). The accuracy and coverage of these techniques were recently compared (Huynen *et al.*, 2000) (see Discussion).

It has been observed previously that phylogenetic trees of ligands and receptors, e.g. insulin and insulin receptors (Fryxell, 1996), were more similar to what could be expected from a general divergence between the corresponding species under the standard molecular clock hypothesis (Zuckerkandl, 1987). The similarity between the phylogenetic trees of interacting proteins was interpreted as an indication of their coordinated evolution and a direct consequence of the similar evolutionary pressure applied to all the members of a given cellular complex.

An extreme of co-evolution of two interacting proteins would be those cases in which both proteins are simultaneously lost in the same species, probably because one of them cannot perform its function without the other. One such example could be His5 (His synthesis) and TrpC (Trp synthesis). This observation is the base of the 'phylogenetic profiles' method (Pellegrini *et al.*, 1999).

In this work, we went one step beyond the binary information (presence/absence of the genes in different species) using the information contained in the full structure of the phylogenetic tree. We measured the similarity between trees as the correlation between the distance matrices used to build the trees, with a methodology similar to that recently published by Goh *et al.* (Goh *et al.*, 2000). In that work, they assessed the similarity of the trees in two examples, the two domains of phosphoglycerate kinase (PGK) and the chemokine–receptor system, quantifying the degree of symmetry between the corresponding trees. Here we extend the idea to a search for interaction partners in a large collection of possibilities. The results indicate that it is indeed possible to distinguish statistically a few true interactions among many possible alternatives, opening up the possibility of searching for interaction partners in large collections of proteins and complete genomes.

Materials and methods

Data sets

Structural domains. The first data set was composed of 13 proteins of known structure for which two structural domains in close interaction are clearly visible (Pazos *et al.*, 1997). These proteins were used to produce a collection of domains. The multiple sequence alignments were taken from the HSSP database (Sander and Schneider, 1993), March 2000 version. The calculation of the similarity of phylogenetic trees was carried out for those pairs of domains with at least 11 sequences from the same species (see below). The final set contained 133 pairs of domains including 13 pairs of truly interacting domains, that is, pairs in which the two domains belong to the same original protein of known structure.

Proteins. A second set was build with 53 *Escherichia coli* proteins extracted from a set previously analyzed by Dandekar *et al.* (Dandekar *et al.*, 1998). The multiple sequence alignments for those proteins were made searching with BLAST (Altschul *et al.*, 1990) using a cut-off value of $P(N) < 1\times10^{-5}$ and aligning with ClustalW (Higgins *et al.*, 1992) the homologous sequences in 14 fully sequenced microbial genomes (*M.tuberculosis*, *Rhizobium* sp., *E.coli*, *H.pylori*, *Synechocystis* sp., *M.thermoautotrophicum*, *A.aeolicus*, *B.burgdorferi*,

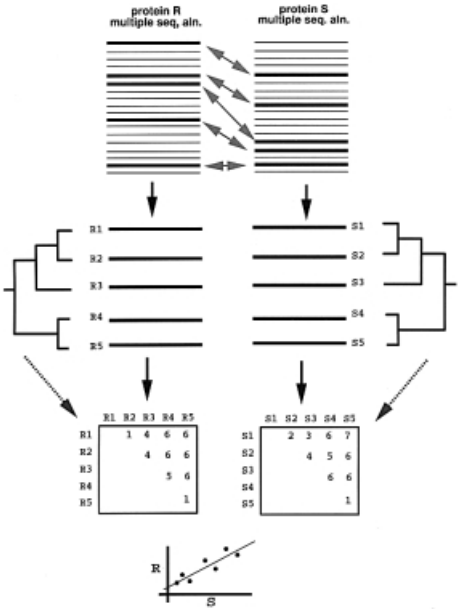

Fig. 1. Scheme of the *mirror tree* method. The initial multiple sequence alignments of the two proteins are reduced, leaving only sequences of the same species and consequently the trees constructed from these reduced alignments would have the same number of leaves and the same species in the leaves. From the reduced alignments, the matrices containing the average homology for every possible pair of proteins are constructed. Such matrices contain the structure of the phylogenetic tree. Finally, the similarity between the data sets of the two matrices and implicitly the similarity between the two trees are evaluated with a linear correlation coefficient.

P.horikoshii, *T.pallidum*, *B.subtilis*, *M.jannaschii*, *H.influenzae*, *A.fulgidus*). As in the previous case, only pairs of proteins for which it was possible to collect more than 11 sequences from the same species were analyzed, leading to a final number of 244 pairs, that included eight pairs of known interactions and eight pairs of possible interactions (see below).

Genomes. A whole genome experiment was performed by collecting alignments for 4300 *E.coli* proteins and combining them in 67 209 pairs of matrices for the analysis. The alignments were collected as in the set above, from 14 complete genomes by searching with BLAST and aligning the homologous sequences with ClustalW. The results of the searches are available at http://www.pdg.cnb.uam.es/mirrortree.

Methodology

For each pair of proteins, the two initial multiple sequence alignments were refined by selecting the sequences that correspond to common species, producing two trees with the same number of leaves (Figure 1). In the cases in which one species contained more than one homologous sequences of a given protein (paralogous sequences), only one of them was selected. We used a simple criterion for the selection, choosing the sequence more similar to the *E.coli* protein or to the HSSP master sequence in the cases of the *E.coli* and structural domains test sets.

We imposed an additional restriction on the minimum size of the protein families by selecting only those cases in which it was possible to collect at least 11 sequences from the same species for both proteins. For example, the pair of proteins A and B is analyzed only if it is possible to find 11 or more species containing both proteins (proteins A and B of *E.coli*, A and B of *H.pylori*, etc.). This minimum limit was set empirically as a compromise between being sufficiently small to provide enough cases and large enough for the matrices to contain sufficient information.

The multiple sequence alignments were used for building matrices containing the distances between all possible protein pairs. Distances were calculated as the average value of the residue similarities taken from the McLachlan amino acid homology matrix (McLachlan, 1971) (Figure 1). Finally, the linear correlation coefficient (r) between the data of these two matrices was calculated according to the standard equation (Press *et al.*, 1992):

$$r = \frac{\sum_{i=1}^{n} (R_i - \bar{R})\,(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n} (R_i - \bar{R})^2}\quad\sqrt{\sum_{i=1}^{n} (S_i - \bar{S})^2}}$$

where n is the number of elements of the matrices, that is, $n = (N^2 - N)/2$, N is the number of sequences in the multiple sequence alignments (Figure 1), R_i are the elements of the first matrix (the distances among all the proteins in the first multiple sequence alignment), S_i is the corresponding value for the second matrix and $\bar{R}$ and $\bar{S}$ are the means of R_i and S_i, respectively.

It is important to note that the method does not require the construction of the phylogenetic trees and only the underlying distance matrices are analyzed, which makes this approach independent of any given tree-construction method.

Results

Interactions between structural domains

Table I contains the full list of correlation values for the 133 pairs of domains analyzed. The positions of the 13 real interactions are highlighted. It can be seen that most of them correspond to high correlation values (nine out of 13 have correlation values better than 0.77).

The representation of these data in Figure 2 shows how the true positives separate well from the bulk of negatives and how correlation values are good indicators of interaction. In this test most of the false positives are produced by two of the proteins: metallothionein (PDB code 4mt2) and cytochrome *c* (2c2c). The wrong predictions produced by the metallothionein could be related to its sequence composition, rich in Cys, since composition bias influences very negatively the quality of multiple sequence alignments (Wootton and Federhen, 1996). We do not have a clear *a posteriori* interpretation for the ubiquitous presence of cytochrome *c* interactions.

There are also a few false negatives, including the two domains of a β-lactamase (PDB code 3blm) and adenylate kinase (3adk), that present low correlation values (0.60 and 0.55, respectively), which makes them undetectable by this method.

This experiment with the structural domains could be considered an 'easy' test, since the interaction partners are domains of the same protein and therefore likely to be subject to stronger evolutionary pressure and co-adaptation. However, it is still an interesting test since it provides an upper threshold for the correlation value of true interactions. The average value of the true interactions (Table I) is 0.78, very similar to the value obtained by Goh *et al.* in the two-domain protein assessed by them (Goh *et al.*, 2000) (i.e. $r = 0.79$ for the two domains of PGK).

Table I. Results of the structural domains test set

Pair	*r*	Pair	*r*	Pair	*r*
2c2c_2–4mt2_2	0.959	3trx_1–3pgk_2	0.577	3trx_2–2c2c_1	0.281
9pap_1–9pap_2	***0.907***	4tnc_1–4mt2_1	0.556	1sgt_1–2c2c_2	0.270
3pgk_1–3pgk_2	***0.901***	3adk_1–3pgk_1	0.554	1alc_1–4mt2_2	0.268
2c2c_1–4mt2_2	0.901	3adk_1–3dfr_2	0.554	1sgt_1–2c2c_1	0.268
4mt2_1–4mt2_2	***0.898***	***3adk_1–3adk_2***	***0.552***	2c2c_1–1rnd_1	0.263
3trx_1–3trx_2	***0.894***	2c2c_1–9pap_2	0.549	2c2c_2–3adk_2	0.254
4tms_1–4tms_2	***0.854***	3trx_2–3pgk_2	0.544	9pap_1–3adk_2	0.254
2c2c_2–4mt2_1	0.849	3adk_1–3pgk_2	0.539	3adk_2–3pgk_1	0.251
1rnd_1–1rnd_2	***0.817***	1rnd_1–4mt2_1	0.513	1sgt_1–1rnd_1	0.238
2c2c_1–4mt2_1	0.813	3trx_2–3pgk_1	0.499	3adk_2–3pgk_2	0.238
1alc_1–1alc_2	***0.801***	2c2c_2–9pap_2	0.486	9pap_2–3adk_2	0.221
4tnc_1–4tnc_2	***0.794***	3adk_2–4tnc_2	0.486	1sgt_2–1alc_2	0.219
2c2c_1–2c2c_2	***0.773***	3trx_1–3adk_1	0.479	2c2c_2–1alc_1	0.203
3pgk_1–4tms_1	0.756	3trx_2–3adk_1	0.469	9pap_1–4tnc_1	0.202
3pgk_1–4tms_2	0.731	3adk_1–4tnc_1	0.466	1sgt_2–1rnd_1	0.191
2c2c_1–3adk_1	0.726	3adk_2–4tnc_1	0.465	1sgt_1–1alc_2	0.178
3pgk_2–4tms_1	0.723	1alc_1–4mt2_1	0.462	3trx_2–3adk_2	0.175
2c2c_2–3pgk_1	0.715	9pap_2–3adk_1	0.459	1sgt_1–1rnd_2	0.168
1alc_1–1rnd_1	0.712	9pap_1–3adk_1	0.455	2pf2_2–1alc_1	0.160
2c2c_2–3pgk_2	0.698	4tnc_2–4mt2_1	0.453	2c2c_1–1alc_1	0.155
1alc_2–1rnd_1	0.697	1sgt_2–4mt2_2	0.452	9pap_1–4tnc_2	0.149
1sgt_1–1sgt_2	***0.693***	4tnc_1–4mt2_2	0.448	2c2c_2–1rnd_2	0.146
3pgk_2–4tms_2	0.691	1alc_2–4mt2_2	0.446	4tms_2–3dfr_1	0.130
3adk_2–3dfr_2	0.675	9pap_2–4tnc_1	0.446	3trx_1–3adk_2	0.128
1sgt_2–2pf2_2	0.673	1sgt_2–4mt2_1	0.433	2c2c_2–1rnd_1	0.125
3dfr_1–3dfr_2	*0.672*	3adk_1–4tnc_2	0.421	2c2c_1–1rnd_2	0.113
2c2c_2–9pap_1	0.658	4tnc_2–4mt2_2	0.405	1sgt_2–1rnd_2	0.073
2c2c_1–3pgk_1	0.648	1rnd_1–4mt2_2	0.405	2c2c_2–4tnc_2	0.050
3trx_2–9pap_1	0.646	2c2c_1–3adk_2	0.401	3trx_1–4tnc_1	0.033
1sgt_1–2pf2_2	0.646	1sgt_2–2c2c_1	0.399	2pf2_2–1alc_2	0.028
2c2c_2–3adk_1	0.631	4tms_2–3dfr_2	0.394	3trx_2–4tnc_1	0.024
3trx_1–9pap_1	0.627	3adk_1–3dfr_1	0.390	4tms_1–4tnc_2	0.024
2c2c_2–1alc_2	0.626	1sgt_2–2c2c_2	0.381	2c2c_1–4tnc_1	0.021
3trx_2–9pap_2	0.620	3adk_2–3dfr_1	0.372	2c2c_2–4tnc_1	0.008
2c2c_1–3pgk_2	0.620	1sgt_2–1alc_1	0.371	2c2c_1–4tnc_2	–0.008
1rnd_2–4mt2_1	0.619	4tms_1–3dfr_2	0.358	4tms_1–4tnc_1	–0.014
1alc_2–1rnd_2	0.607	1sgt_1–4mt2_1	0.343	3trx_1–4tnc_2	–0.067
1rnd_2–4mt2_2	0.606	1sgt_1–4mt2_2	0.336	3trx_2–4tnc_2	–0.123
3blm_1–3blm_2	***0.603***	9pap_2–4tnc_2	0.331	4tms_2–4tnc_2	–0.149
1alc_1–1rnd_2	0.599	4tms_1–3dfr_1	0.327	3pgk_1–4tnc_1	–0.158
3trx_1–3pgk_1	0.595	3trx_1–2c2c_2	0.319	3pgk_1–4tnc_2	–0.169
3trx_1–9pap_2	0.589	3trx_1–2c2c_1	0.312	3pgk_2–4tnc_1	–0.178
1alc_2–4mt2_1	0.588	1sgt_1–1alc_1	0.312	3pgk_2–4tnc_2	–0.181
2c2c_1–1alc_2	0.587	3trx_2–2c2c_2	0.287	4tms_2–4tnc_1	–0.217
2c2c_1–9pap_1	0.581				

The table contains the full list of pairs of domains constructed merging the domains of 13 two-domain proteins and where the multiple sequence alignments of the two domains contain 11 or more sequences in common. The pairs of domains are named pdbid1_domain1-pdbid2_domain2. The correlation coefficient, indicator of tree similarity, is shown for every pair. The table is sorted by that value. 'True interactions' corresponding to the two structural domains of the same protein are highlighted in underlined bold italic type.

Interactions between proteins

The second test was carried out on the 244 pairs of proteins derived from the Dandekar *et al.*'s set (Dandekar *et al.*, 1998) (see Materials and methods). This set contains eight true interactions between well-known proteins and a small number of other possible interactions, e.g. different ribosomal proteins which form part of the same macromolecular complex even though they may not interact directly.

As in the previous test, most of these pairs of truly interacting proteins have high correlation values (Figure 3) and there is a clear correlation between interaction index and true interactions with eight out of eight true interations and seven out of eight possible interactions with correlation values better than 0.8. The pair with the highest correlation value corresponds to the known interaction between the α and β subunits of the membrane ATP synthase and the first 'false positive' corresponds to the pair formed by the chaperonin GroEl and the ribosomal protein S7.

Interactions in the E.coli genome

We carried out a fully automatic prediction of protein interactions at the genomic scale with the aim of obtaining a significant number of predictions. We generated alignments for 4300 *E.coli* proteins which allowed the study of 67 209 possible interaction pairs. This number is still far from the total of 9.2×10^6 possible pairs between *E.coli* proteins, of which about 20 000 true interactions are expected if we consider the average of interactions per protein detected in *H.pylori* by Yeast Two-Hybrid screening (Rain *et al.*, 2001). In our case, the main limitation for building a larger data set was the use of a relatively small set of 14 genomes for constructing the alignments.

The analysis of the possible interactions leads to the proposal

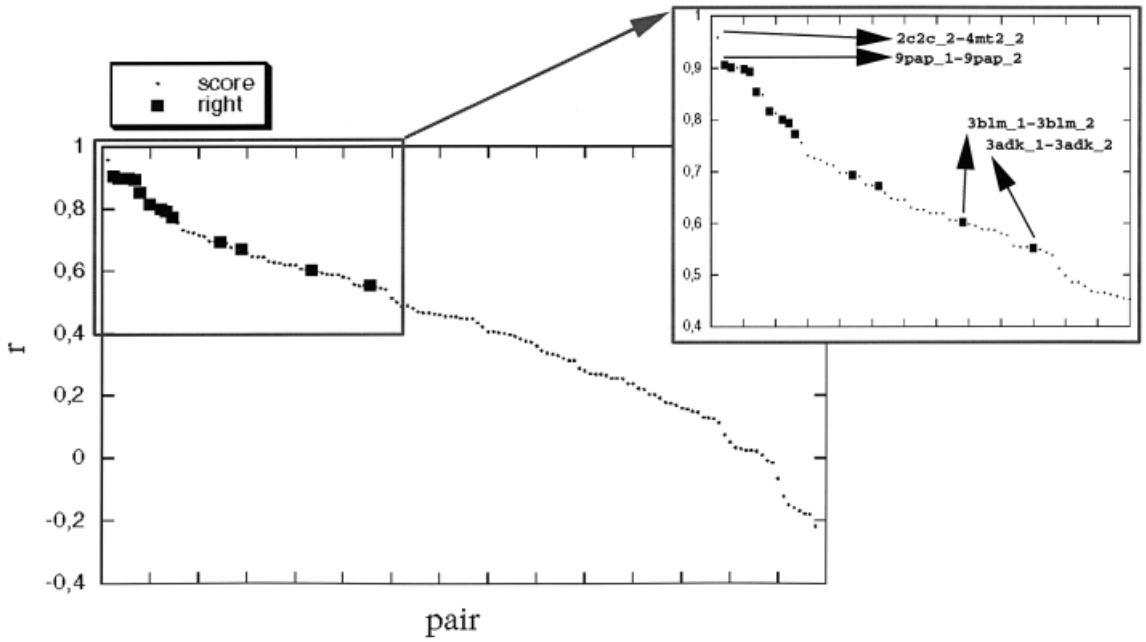

Fig. 2. Representation of the results for the structural domains data set. The data in Table I are plotted representing the correlation value for the 133 pairs of domains. Pairs representing 'true interactions' – the two structural domains of the same protein – are marked with a filled square. Some of the pairs are labeled with the pair name as in Table I.

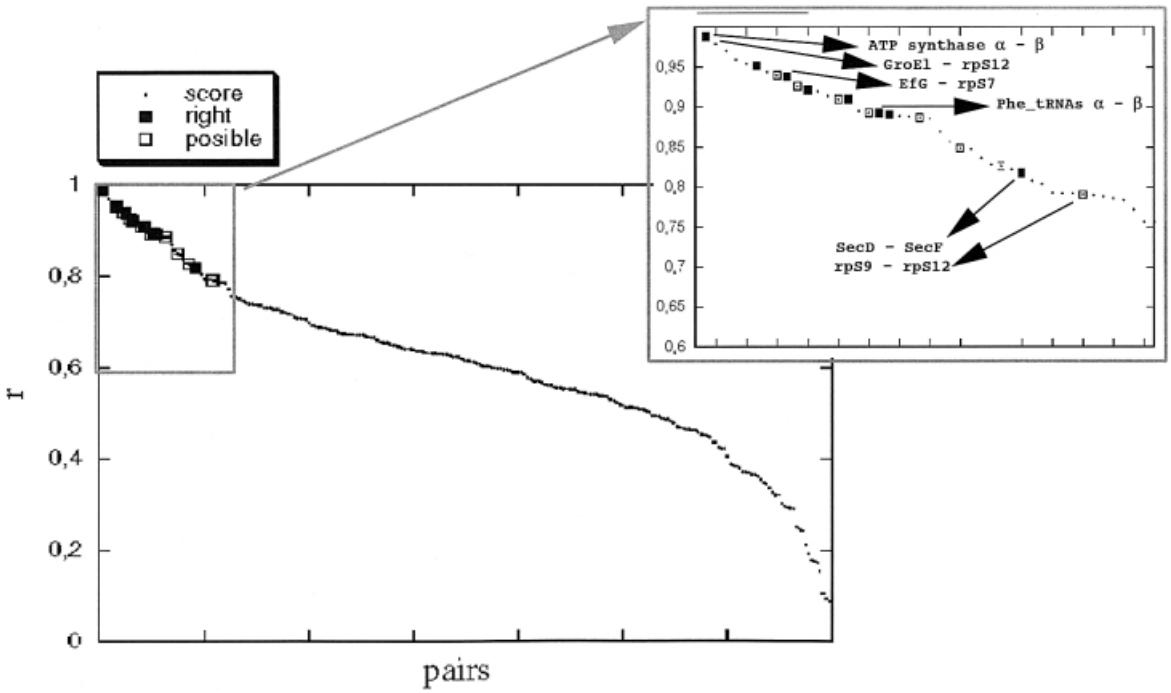

Fig. 3. Representation of the results for Dandekar *et al.*'s data set (Dandekar *et al.*, 1998). The correlation value for the 244 pairs is shown. True interactions are marked with a filled square and possible ones (i.e. pairs of ribosomal proteins) with open squares. Representative pairs are labeled with the name of the corresponding proteins.

of more than 2700 pairs of proteins that were found to have scores better than 0.8 (Figure 4). Well-known interactions are included among the stronger predictions, including proteins such as ATP synthase α and β, elongation factors Tu and G, and ribosomal proteins S2–S10 and S2–S11. Among the pairs with interaction predictions better than 0.8, there are 460 proteins labeled as hypothetical. For example, the protein of unknown function YHBZ_ECOLI is predicted to interact with the ribosomal protein S4 and YFGK_ECOLI is predicted to interact with the polynucleotide phosphorylase PNP_ECOLI. For these proteins these predictions are the first clue about their possible function.

The pairs of proteins with highest similarities of phylogenetic trees that correspond to new predictions of interaction are two GTP-binding proteins LEPA and YCHF, the chaperone GroEl with the ribosomal protein S15 and glutamyl tRNA synthetase

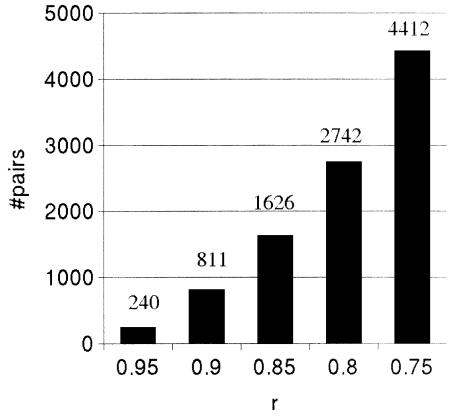

Fig. 4. Number of predicted interactions in *E.coli* depending on the cut-off value considered.

with a GMP synthetase. The validity of these predictions would have to be confirmed experimentally.

Discussion

The prediction of protein interaction partners with bioinformatics methods has become a topic of increasing interest in recent years. Different methods have emerged based on the study of conservation of gene order (Dandekar *et al.*, 1998), the presence/absence of pairs of proteins in full genomes (Gaasterland and Ragan, 1998; Pellegrini *et al.*, 1999) or the presence of proteins assembled in multi-domain proteins in other genomes (Enright *et al.*, 1999; Marcotte *et al.*, 1999).

We based the study presented here on the common previous observation of the similarity between the trees of interacting proteins. Examples of systems in which this relation was previously observed are insulin and dockerin and their corresponding receptors (Fryxell, 1996; Pages *et al.*, 1997). Recently, the relation between the corresponding phylogenetic trees was carefully quantified for multiple sequence alignments corresponding to the two domains of phosphoglycerate kinase (Goh *et al.*, 2000).

Here we present a systematic extension of these ideas about the similarity in the evolutionary history of complementary proteins ('mirror trees') by applying it to the large-scale detection of interacting protein partners. The results obtained in different systems demonstrate that the similarity between phylogenetic trees can be used as a predictor of protein interaction, with >66% of true positives detected at correlation values better than 0.8. This value of 0.8 seems to be a good empirical cut-off to discriminate between true and false interactions (Figures 2 and 3) in accordance with Goh *et al.* (Goh *et al.*, 2000). This cut-off is probably a stringent one, since it has been derived from domains of the same protein that are likely to be in permanent (not transient) interaction. It is possible that for free independent proteins the pressure for interaction would be smaller, the signals left by their interaction in the corresponding trees probably weaker and correspondingly the correlation values smaller.

One of the more interesting properties of the current approach is its capacity to cover a significant number the potential interactions. For example, in the set of proteins derived from the study of Dandekar *et al.* (Dandekar *et al.*, 1998), it was possible to study the interaction of 244 pairs, accounting for 18% of the total number of 1378 possible pairs. In the genome-based experiment the number of possible interactions explored was as large as 67 209, a substantial number even if it is still a small fraction of the possible 9.2×10^6 pairs. In this case the number of predictions of interaction was of 2742, which is clearly higher than the 64 interactions predicted from the information about domain arrangements by Enright *et al.* for three genomes (Enright *et al.*, 1999) or the 749 predicted by Marcotte *et al.* for *E.coli* (Marcotte *et al.*, 1999). The coverage of these techniques was compared using the genome of *M.genitaliun* for benchmarking (Huynen *et al.*, 2000) and it ranges from 6% for the techniques based on gene fusion events to 37% for those based on the conservation of gene order. A separate issue is how accurate these predictions would be.

Despite the promising results obtained in the different tests carried out, a number of problems are still present in the current approach. First, the number of possible interactions could have been increased by collecting sequences from a larger number of genomes or by improving the process of selection of the corresponding sequences from the same species in the corresponding pairs of protein families. It is to be expected that the continuous stream of new sequences and genomes would alleviate this problem, allowing us to increase the number of predictions easily. Second, the quality of the underlying alignments is a key factor and a number of false positives are introduced in the case such as the Cys-rich protein discussed in Results. Third, it is possible that some inaccuracies are introduced by comparing distance matrices instead of the real phylogenetic trees, since the distance matrices are not a perfect representation of the corresponding phylogenetic trees. Given that the comparison of phylogenetic trees is a difficult and not fully solved problem, we decided to short-cut the problem by comparing their underlying distance matrices. Finally, it is really difficult to assess definitively the accuracy of any of the protein interaction prediction methods in the absence of a well-accepted and large enough collection of annotated protein interactions.

Among the positive features of the mirror tree approach, it is interesting to mention that it does not require the presence of fully sequenced genomes, as other methods do, e.g. the 'phylogenetic profiles' method (Pellegrini *et al.*, 1999), since the mirror tree approach is based only on information about protein families whether they are coming from complete genomes or not.

This approach and the others commented upon here have different ranges of reliability and applicability. A prospect for the future is to combine them to obtain an *in silico* prediction of the interaction network of an organism.

Acknowledgements

We acknowledge A.de Daruvar (Univerity of Bordeaux) and members of the Protein Design Group for interesting discussions.

References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.

Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) *Trends Biochem. Sci.*, **23**, 324–328.

Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) *Nature*, **402**, 86–90.

Fryxell,K.J. (1996) *Trends Genet.*, **12**, 364–369.

Gaasterland,T. and Ragan,M.A. (1998) *Microb. Comp. Genomics*, **3**, 199–217.

Goh,C.S., Bogan,A.A., Joachimiak,M., Walther,D. and Cohen,F.E. (2000) *J. Mol. Biol.*, **299**, 283–293.

Higgins,D.G., Bleasby,A.J. and Fuchs,R. (1992) *Comput. Appl. Biosci.*, **8**, 189–191.

Huynen,M., Snel,B., Lathe,W. and Bork,P. (2000) *Genome Res.*, **10**, 1204–1210.
Ito,T., Tashiro,K., Muta,S., Ozawa,R., Chiba,T., Nishizawa,M., Yamamoto,K., Kuhara,S. and Sakaki,Y. (2000) *Proc. Natl Acad. Sci. USA*, **97**, 1143–1147.
Marcotte,E.M., Pellegrini,M., Ho-Leung,N., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999a) *Science*, **285**, 751–753.
Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999b) *Nature*, **402**, 83–86.
McLachlan,A.D. (1971) *J. Mol. Biol.*, **61**, 409–424.
Mendelsohn,A.R. and Brent,R. (1999) *Science*, **284**, 1948–1950.
Pages,S., Belaich,A., Belaich,J.P., Morag,E., Lamed,R., Shohan,Y. and Bayer,E.A. (1997) *Proteins*, **29**, 517–527.
Pazos,F., Helmer-Citterich,M., Ausiello,G. and Valencia,A. (1997) *J. Mol. Biol.*, **271**, 511–523.
Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B.P. (1992) *Numerical Recipes in C: the Art of Scientific Computing*. 2nd edn. Cambridge University Press, Cambridge.
Rain,J.C. *et al.* (2001) *Nature*, **409**, 211–215.
Sander,C. and Schneider,R. (1993) *Nucleic Acids Res.*, **21**, 3105–3109.
Uetz,P. *et al.* (2000) *Nature*, **403**, 623–631.
Wootton,J.C. and Federhen,S. (1996) *Methods Enzymol.*, **266**, 554–571.
Zuckerkandl,E. (1987) *J. Mol. Evol.*, **26**, 34–46.

Received January 1, 2001; revised May 25, 2001; accepted June 18, 2001