Similarity of phylogenetic trees as indicator of protein–protein interaction

Florence Pazos and Alfonso Valencia

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

To whom correspondence should be addressed. E-mail: valencia@cnb.csic.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 $P(N) < 1 \times 10^{-5}$ 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.

**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 adenylyl 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).
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 interactions 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 $9.2 \times 10^6$ 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
was as large as 67,209, a substantial number even if it is still a small fraction of the possible $9.2 \times 10^6$ 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. genitalium 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
F. Pazos and A. Valencia


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