

TSEMA: interactive prediction of protein pairings between interacting families

José M. G. Izarzugaza, David Juan, Carles Pons¹, Juan A. G. Ranea², Alfonso Valencia¹ and Florencio Pazos*

National Center for Biotechnology (CNB-CSIC), C/Darwin, 3 Cantoblanco, 28049 Madrid, Spain, ¹National Bioinformatics Institute (INB), Barcelona Supercomputer Centre, C/Jordi Girona, 29 08034 Barcelona, Spain and ²Department of Biochemistry and Molecular Biology, Biomolecular Structure and Modelling Unit, University College London, London WC1E 6BT, UK

Received February 13, 2006; Revised March 1, 2006; Accepted March 9, 2006

ABSTRACT

An entire family of methodologies for predicting protein interactions is based on the observed fact that families of interacting proteins tend to have similar phylogenetic trees due to co-evolution. One application of this concept is the prediction of the mapping between the members of two interacting protein families (which protein within one family interacts with which protein within the other). The idea is that the real mapping would be the one maximizing the similarity between the trees. Since the exhaustive exploration of all possible mappings is not feasible for large families, current approaches use heuristic techniques which do not ensure the best solution to be found. This is why it is important to check the results proposed by heuristic techniques and to manually explore other solutions. Here we present TSEMA, the server for efficient mapping assessment. This system calculates an initial mapping between two families of proteins based on a Monte Carlo approach and allows the user to interactively modify it based on performance figures and/or specific biological knowledge. All the explored mappings are graphically shown over a representation of the phylogenetic trees. The system is freely available at <http://pdg.cnb.uam.es/TSEMA>. Standalone versions of the software behind the interface are available upon request from the authors.

INTRODUCTION

The prediction of protein interactions from sequence and genomic features is helping in the functional interpretation

of the massive amounts of genomic information. The context of a protein in the 'interactome' of a given organism provides important information on its biological role. Computational techniques for the prediction of protein interactions based on sequence and genomic features (context-based methods) provide information which is orthogonal and complementary to the traditional methods based on sequence similarity (similarity-based methods) (1).

One of these computational methods is based on the fact that interacting families of proteins (i.e. a family of ligands and their corresponding receptors) tend to have similar phylogenetic trees. This was first observed qualitatively (2) and later quantitatively evaluated in large sets of interacting proteins (3,4). A hypothesis for explaining this relationship between interaction and tree similarity states that interacting partners are forced to adapt to each other. This process of co-adaptation should lead to correlated evolutionary histories, which in turn should be reflected in a tree similarity higher than the expected background similarity due to the underlying speciation process (5). One practical way of quantifying this similarity is to calculate the correlation between the two sets of distances extracted from the two trees (3,4). This methodology has been followed by many authors who developed different implementations and variations of it (5–11).

The relationship between tree similarity and protein interaction has predictive power in two directions. It can be applied to evaluate whether or not two sets of proteins, for which the mapping (links between the leaves of both trees) is known, interact. For example, the eventual interaction between two sets of orthologues for which the mapping is provided by the organisms themselves, can be investigated. This allows, among other things, to predict pairs of interacting proteins on a genomic scale by evaluating the similarity of trees for all pairs of proteins within a genome (pairs of groups of orthologues, actually) (4). On the other hand, we can start from two families known to interact and predict the mapping based on the idea

*To whom correspondence should be addressed. Tel: +34 915854669; Fax: +34 915854506; Email: pazos@cnb.uam.es

that the real mapping would be the one maximizing the similarity between the trees (7). Predicting the mapping between the members of two interacting protein families (i.e. which receptor within one family interacts with which ligand within the other) is very important especially in eukaryotic organisms. In these organisms, large families of interacting paralogues exist for which only one or a small number of pairs of interacting proteins have been experimentally determined (i.e. Ras/Ras effectors, chemokines/chemokine receptors). Most biologists working with eukaryotic proteins face the problem that there is more than one paralogue for their protein and for its interactors and that, in many cases, the unknown network of interactions between the members of these families is crucial for explaining their biological role.

The exhaustive exploration of all possible mappings between two sets of proteins (in the search for the one maximizing the tree similarity) is unfeasible due to the combinatorial nature of the problem: the number of possible mappings between two sets of n elements is $n!$. For this reason, current approaches for finding these mappings use heuristic techniques to avoid the exhaustive exploration of all the possibilities. Ramani and Marcotte (7) developed a method which uses a Monte Carlo approach to perform a 'guided' exploration of the space of solutions in the search for the best one. The search space can be reduced even more by avoiding mappings incompatible with certain characteristics of phylogenetic trees, like automorphism (11). Because of their intrinsic heuristic nature, these methods do not ensure that the best solution is found but they may find a sub-optimal solution (trapped within a local minimum). For a user interested in the interactions between the members of two families of proteins, it is worth further exploring the (eventually sub-optimal) solution proposed by these heuristic approaches. This exploration can be driven by expert knowledge (forcing pairs of proteins suspected or known to interact) and/or by performance figures indicative of the reliability of the proposed links.

In this work we present TSEMA, 'the server for efficient mapping assessment'. This system is intended not only to provide the user with a predicted mapping based on an heuristic search, but to allow her/him to interactively explore and modify it through a web interface. This interactive process can be used to find better solutions not explored by the heuristic approach due to intrinsic limitations.

METHODS AND WEB INTERFACE

First step: generating the initial mapping

The initial input for the system are the two families of homologous proteins for which the user wants to predict the mapping (which protein within one family interacts with which one within the other). The user can submit this information through a plain web interface (Figure 1). The only compulsory fields (apart from the information on the families) are a name for the job and an email address where to send forthcoming messages and results. There is a set of advanced options to control the Monte Carlo algorithm (see below) which are intentionally blurred unless the user decides to enable them (Figure 1). For the two families of proteins, the user can either submit two phylogenetic trees (in Newick format) or two multiple sequence alignments in a format

compatible with ClustalW (12). In the second case, the system generates the phylogenetic trees from the alignments using the neighbour joining algorithm implemented in ClustalW. It is highly recommended that the user submits her/his own tree generated with more sophisticated techniques (parsimony or Bayesian trees). In the next step, distances for all pairs of proteins within both families are extracted from the phylogenetic trees by summing the length of the branches separating each pair of proteins in the trees.

These two sets of distances are used to run a modified implementation of Ramani and Marcotte's Monte Carlo Metropolis method (7) for finding the mapping between the two sets of proteins which maximizes the matching between these two sets of distances. This implementation, written in C, includes the possibility of running two sets with different number of proteins, which is the situation for most real interacting families, owing to promiscuity, pseudogenes and the like. In this implementation it is also possible to use different scores for evaluating the matching between the distance matrices, including linear correlation and root mean square deviation. Owing to its stochastic nature, the Monte Carlo step is repeated many times (500 by default) to obtain an estimation of the consistency of the results. For each one of the 500 runs, the system explores up to 1 000 000 solutions. The complexity of both trees is also calculated as the entropy of the distribution of distances. Trees with low complexity are expected to produce worse results since there is not enough information to distinguish between mappings (7).

The results of this first step include the overall best mapping obtained through all the Monte Carlo runs, the best mapping obtained in each one of the runs, and a contingency table which shows in how many of these runs a given pair of proteins is linked. This raw file can be further processed by the user to implement her/his own analysis, or it can be submitted to the interactive analysis part of the server ('New Analysis' button).

Second step: interactive analysis and modification of the proposed mapping

The single raw file with the results produced in the first step is the only input required for the analysis, although the user can additionally provide an email address and a job identifier to facilitate tracking of the jobs.

The interactive analysis interface (Figure 1) shows the list of predicted interacting pairs of proteins corresponding to the best mapping found (the one with the best score through all the runs). For each pair, four scores are shown: 'reliability', representing the percentage of mappings where that pair is present, and 'segregation' which gives an idea of the difference between the reliability of that pair and the second best reliability value. The reliability for pair AB could be different from pair BA, since A and B might be confronted with a different number of proteins. Hence, there are two values of reliability and segregation for each pair. The coincidence matrix can also be accessed from this interface. These scores are coloured from red (bad) to blue (good). The entropies of the two trees (see above) are also shown at the top of the lists.

The interface also displays a graphical representation of the two trees showing the predicted interacting pairs of proteins corresponding to the current mapping (Figure 1). The colour of

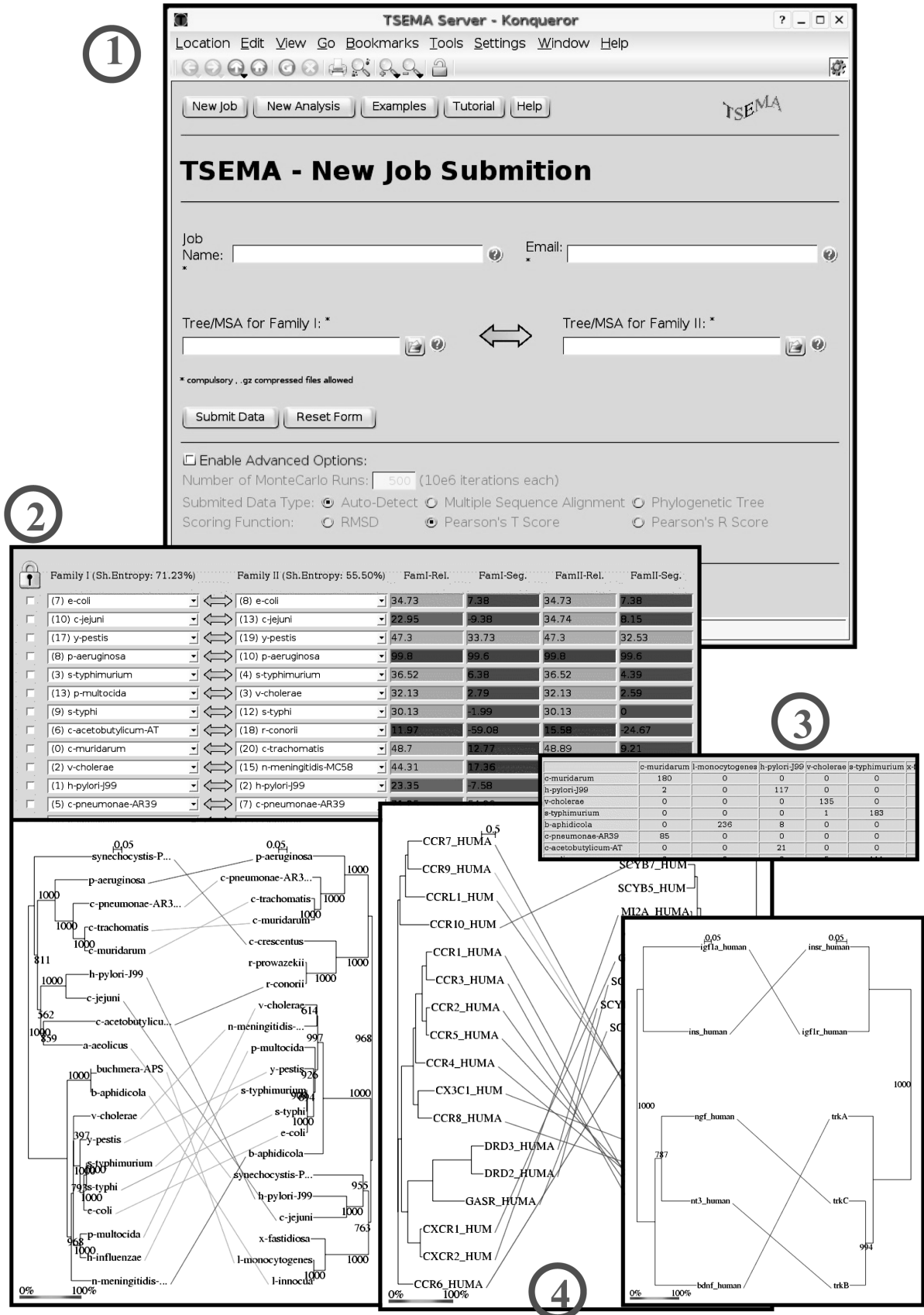


Figure 1. Screenshots of TSEMA's web interface. (1) Initial screen for submitting a new job. (2) List of predicted links (mapping). (3) Coincidence matrix. (4) Examples of representations of the phylogenetic trees together with the predicted mappings.

the links correspond to the AB reliability score of the list above. If there is bootstrap information in the trees provided by the user it is included in the representation. If the user submits multiple sequence alignments, the system generates bootstrap trees. Many wrong pairings are associated to internal nodes with low bootstrap support (data not shown). The initial layouts of the trees are calculated with NJPlot (13).

The last section of the interface shows the distance correlation plots corresponding to the current and other mappings. The figure on the left shows the correlation plot of the current mapping superposed with that of the previous mapping, while the one on the right shows the correlation plot of the current mapping compared with that of the original mapping. These plots can be used to assess whether a given change in the mapping affects many distances, or whether a given mapping produces an overall good score but there are some outliers. The plots are generated with GNUPlot (www.gnuplot.info).

In this interactive interface, the user can start changing links in the list of predicted pairs and assess how these changes affect the scores (reliability and segregation). Anytime the user changes a link, the new mapping incorporating that change is generated and shown in the tree representation and in the correlation plots. The user can undo the changes to the previous mapping or load the original (first) mapping. The links the user is more confident about can be 'fixed' to avoid changing them.

This interface allows the user to interactively explore alternative mappings by applying some changes and assessing the quality of the new mappings graphically and by the scores. The coincidence matrix (Figure 1) is a good starting point for guessing possible changes in the mappings. A given pair which is found in most of the mappings generated by the different runs (high reliability in the coincidence matrix) might not be present in the overall highest scoring tree. In this case, it would be worth forcing that pair in the mapping. The user can also incorporate expert information in this process, e.g. by forcing some pairs suspected to interact.

The system also has a help page and a guided tutorial for the user to get familiar with the interface and its functioning. There are also some precomputed examples the user can play with.

DISCUSSION

The relationship between protein interactions and similarity of phylogenetic trees has been extensively used for assessing the possible interaction between two proteins, and to predict the mapping between the members of two families known to interact. The server presented here is intended to be used in this second scenario. It allows the user to predict a mapping between two sets of proteins using an heuristic approach and to interactively refine and improve it.

The accuracy of Monte Carlo-based methods like the one implemented here has been quantified for some cases (7,11). It is very difficult to obtain a large enough set of examples for testing these methods: interacting families for which the mapping is known, with enough members and so on. This is something that remains to be done and that will allow to obtain performance figures for our modified method (obviously without the interactive part) and, more importantly, to relate that performance with parameters like tree complexity,

number of sequences, tree similarity of the right mapping, bootstrap values and the like. While that quantification is not available, the user has to qualitatively assess these parameters. For example, if the complexity (entropy) of one or the two trees is low, it means that there is not enough topological information in the trees to distinguish the right mapping. Even if the overall entropy of the trees is not bad, there could be 'local' low-entropy regions producing bad results. For instance, for two proteins which are exactly at the same distance from their ancestral node, the mappings involving one or the other would have exactly the same score and hence they would be indistinguishable. Special instances of this would be pairs of identical proteins (distance = 0). Similarly, we have observed many wrong predictions involving clades in the tree for which the bootstrap support is low.

In summary, the results of the method are totally dependent on the tree quality. This is why it is desirable for the user to provide a manually-curated tree, obtained with state-of-the-art methodologies (like Bayesian trees) instead of relying on the neighbour-joining tree automatically generated by the server. The generation of Bayesian trees is very CPU-demanding and, to some extent, a manual process. We are working on incorporating this methodology in future versions of the server.

As more genomes continue to be sequenced in a high-throughput way, the number of interacting families for which the mapping is unknown will grow too, specially for eukaryotic genomes. For many families of paralogues with biomedical interest, the differential interaction between their members is a crucial aspect for explaining their functioning (Ras, chemokines, G-proteins and so on). The server presented here can help in elucidating these complex networks of interactions by interactively assessing the landmarks the evolution left on them.

ACKNOWLEDGEMENTS

We would like to acknowledge Diego Díez (IIB-CSIC) and the members of the Protein Design Group (CNB-CSIC) for discussion and suggestions. We are specially grateful to Ana M. Rojas for her help on Bayesian trees, Michael Tress for critical reading of the manuscript, and Eduardo Andres and Angel Carro for technical assistance. F.P. is the recipient of a 'Ramón y Cajal' contract from the Spanish Ministry for Education and Science. C.P. work is supported by a grant from 'Genoma España' to the National Institute for Bioinformatics. This work has been partially funded by the GeneFun (LSHG-CT-2004-503567) and BioSapiens (LSHC-CT-2003-505265) EU projects and a grant from the 'Fundación BBVA'. Standalone versions of the software behind the interface are available upon request from the authors. Funding to pay the Open Access publication charges for this article was provided by the BioSapiens EU project (LSHC-CT-2003-505265).

Conflict of interest statement. None declared.

REFERENCES

1. Huynen, M., Snel, B., Lathe, W. and Bork, P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, **10**, 1204–1210.

2. Fryxell, K.J. (1996) The coevolution of gene family trees. *Trends Genet.*, **12**, 364–369.
3. Goh, C.-S., Bogan, A.A., Joachimiak, M., Walther, D. and Cohen, F.E. (2000) Co-evolution of proteins with their interaction partners. *J. Mol. Biol.*, **299**, 283–293.
4. Pazos, F. and Valencia, A. (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng.*, **14**, 609–614.
5. Pazos, F., Ranea, J.A.G., Juan, D. and Sternberg, M.J.E. (2005) Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. *J. Mol. Biol.*, **352**, 1002–1015.
6. Gertz, J., Elfond, G., Shustrova, A., Weisinger, M., Pellegrini, M., Cokus, S. and Rothschild, B. (2003) Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics*, **19**, 2039–2045.
7. Ramani, A.K. and Marcotte, E.M. (2003) Exploiting the co-evolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.*, **327**, 273–284.
8. Kim, W.K., Bolser, D.M. and 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.
9. Tan, S., Zhang, Z. and Ng, S. (2004) ADVICE: automated detection and validation of interaction by co-evolution. *Nucleic Acids Res.*, **32**, W69–W72.
10. Sato, T., Yamanishi, Y., Kanehisa, M. and Toh, H. (2005) The inference of protein–protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. *Bioinformatics*, **21**, 3482–3489.
11. Jothi, R., Kann, M.G. and Przytycka, T.M. (2005) Predicting protein–protein interaction by searching evolutionary tree automorphism space. *Bioinformatics*, **21**, i241–i250.
12. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. and Thompson, J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
13. Perrière, G. and Gouy, M. (1996) WWW-Query: an on-line retrieval system for biological sequence banks. *Biochimie*, **78**, 364–369.