BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 10 2010, pages 1370–1371
doi:10.1093/bioinformatics/btq137

Sequence analysis

Advance Access publication March 30, 2010

Studying the co-evolution of protein families with the Mirrortree web server

David Ochoa and Florencio Pazos*

National Centre for Biotechnology, Computational Systems Biology Group (CNB-CSIC), c/ Darwin, 3. Cantoblanco, 28049 Madrid, Spain

Associate Editor: Burkhard Rost

ABSTRACT

Summary: The Mirrortree server allows to graphically and interactively study the co-evolution of two protein families, and investigate their possible interactions and functional relationships in a taxonomic context. The server includes the possibility of starting from single sequences and hence it can be used by non-expert users.

Availability and Implementation: The web server is freely available at http://csbg.cnb.csic.es/mtserver. It was tested in the main web browsers. Adobe Flash Player is required at the client side to perform the interactive assessment of co-evolution.

Contact: pazos@cnb.csic.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 21, 2010; revised on March 23, 2010; accepted on March 26, 2010

1 INTRODUCTION

A lot of biological knowledge is hidden in the complex networks of relationships of different nature between molecular entities. In the case of proteins, their biological roles can only be fully understood in the context of their interaction with others. This importance in deciphering as much as possible of the complex network of interactions and functional relationships between proteins has led to the development of specific experimental (Shoemaker and Panchenko, 2007a) and computational (Shoemaker and Panchenko, 2007b) techniques for this task. One family of these computational techniques is based on the observed relationship between protein interactions and co-evolution [similarity of evolutionary histories as represented by phylogenetic trees; see Pazos and Valencia (2008) and references herein]. This approach, termed *mirrortree*, has been applied not only to look for interaction partners in large datasets of proteins (e.g. Juan *et al*., 2008), but also to study in depth the co-evolution and interactions in particular pairs of protein families (e.g. Dou *et al*., 2006; Labedan *et al*., 2004; McPartland *et al*., 2007). Many authors developed variations and different implementations of this approach [e.g. see references in Pazos and Valencia (2008)], but none of them are intended to be operated by non-experts users. They are either very specific for certain needs or are distributed as non-interactive command-line programs or require a complex preparation of the input data (e.g. generation of the multiple sequence alignments (MSAs) and/or phylogenetic trees). This precludes these techniques from being used by most molecular biologists.

In this work, we present the Mirrortree server, an automatic system for the interactive assessment of co-evolutionary features between two protein families. The system only requires as input the sequence of a single representative of each family to start, which allows it to be used by non-bioinformaticians. All the subsequent steps (search for homologues, localization of orthologues, generation and filtering of MSAs and trees, and tree comparison) are fully automatic. Nevertheless, expert users have the possibility of providing their (manually curated) MSAs or trees. Moreover, the tree comparison is done in an interactive interface that allows users to study in depth the co-evolution of their families and investigate their interactions in a taxonomic context.

2 WORKFLOW

Supplementary Material 1 contains an exhaustive description of the server workflow. What follows is a short description. Each one of the two input sequences is BLASTed (Altschul *et al*., 1997) against the Integr8 database of fully sequenced genomes (Kersey *et al*., 2005). The list of putative homologues is filtered to discard fragments, divergent sequences, etc. The remaining sequences are aligned with Muscle (Edgar, 2004). The resulting MSA is filtered again (see Supplementary Material 1 for details) and only one homologue per species is retained as the putative orthologue (the one with highest similarity to the master). The final MSA of putative orthologues is used to construct a phylogenetic tree with the 'neighbour-joining' (NJ) algorithm implemented in ClustalW (Chenna *et al*., 2003). Expert users can bypass these steps by providing their own MSAs or phylogenetic trees (i.e. generated with more sophisticated techniques than NJ). The computationally expensive steps are delegated to a computer cluster. As an example, running the whole process for two families of around 800 residues long with 120 species in common takes 10 min.

3 INTERFACE

When the process is completed, the user receives an e-mail containing a link to the interactive Flash-based visualization of the trees of the two families (Fig. 1), as well as files with useful intermediate results (MSAs and trees for the two families, static graphical representations of the mirroring trees, etc). Organisms present in both families are connected by lines in this representation. Tree branches can be swapped in order to confront matching clades between the two trees and obtain a better representation. The tree

*To whom correspondence should be addressed.

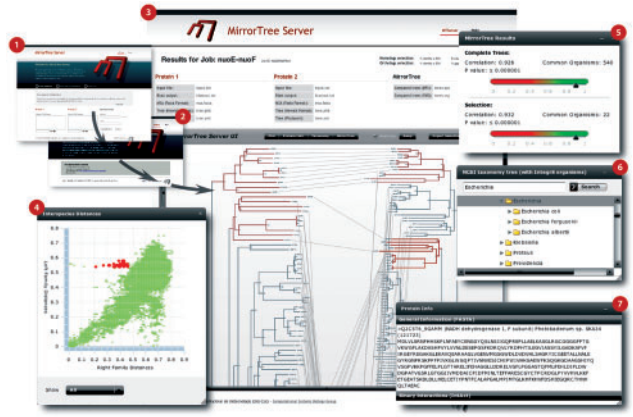

Fig. 1. Interface of the Mirrortree server. (**1**) Job submission page. (**2**) Job ID and status. (**3**) Main interface for viewing and manipulating the trees. The different panels can be shown/hidden and freely moved/resized in a windows-like manner. (**4**) Panel with the distance correlation plot. (**5**) Tree and sub-tree similarity scales and associated *P*-values. (**6**) Taxonomy browser. (**7**) Uniprot information for individual proteins.

representation can be zoomed and the user can select different proteins (leaves) or whole clades (internal nodes) in both trees in order to restrict the calculation of tree similarity to certain groups of organisms. Panels with additional tools and information are arranged on the top of this representation and can be shown/hidden and freely moved/resized in a windows-based interface (Fig. 1). One of these panels shows the similarity of the trees as calculated by *mirrortree* in a colour scale. The tree similarity for the current selection is also shown in this panel. Another panel shows information available for the selected proteins (leaves) in the Uniprot resource (Uniprot Consortium, 2009), such as protein name, sequence, organism and reported interactions. Organism selection can also be done by taxonomic criteria using the included taxonomy browser (Fig. 1), i.e. to evaluate the co-evolution in a certain kingdom or family. Selections in the tree are also shown in the taxonomy browser. The sub-alignment for the sequences in the current selection can be exported for further analysis. Finally a plot with a simplified representation of the correlation between the inter-protein distances in both families is also shown. This plot can show all the distances or only the ones involving the selected organisms. This plot is very useful to detect outliers: clouds of points far from the diagonal representing non-correlated distances that decrease the overall similarity of the trees. In many cases, these are related to non-standard evolutionary events such as horizontal gene transfer (Pazos *et al*., 2005). Selections of points in this plot cause the corresponding organisms/clades in the trees to be selected. The server has many other features extensively explained in a help file. There is also a guided tutorial for illustrating the kind of studies that can be performed with the server.

4 CONCLUSION

The Mirrortree server is the first system for interactively assessing the co-evolution between two protein families in order to evaluate their possible interactions in a taxonomic framework. There are related systems such as TSEMA (Izarzugaza *et al*., 2008) which, based on the same relationship between protein interactions and tree similarity, are nevertheless intended for predicting the mapping (connections between the leaves) between two families already known to interact. Moreover, that server does not include the possibility of automatically generating MSAs and hence it is more difficult to be used by non-experts.

An important requirement for a computational tool to be used by biologists is simplicity. That left most existing tools for studying co-evolution and predicting protein interactions out of their standard toolkit. The Mirrortree server was developed with the goal of being amenable to be used by non-experts, in such a way that any user can interactively study the co-evolution between his/her families of interest in a taxonomic context starting with single sequences.

ACKNOWLEDGEMENTS

We want to thank Octavio Diaz-Pines and the members of the CTI-CSIC for computer support. We also want to acknowledge Daniel Lopez (CNB-CSIC), David Juan and Alfonso Valencia (CNIO) for comments and suggestions.

Funding: BIO2006-15318 project of the Spanish Ministry for Science and Innovation (in part); PhD fellowship of the Basque Country Government (to D.O.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Chenna,R. *et al*. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.

Dou,T. *et al*. (2006) Co-evolutionary analysis of insulin/insulin like growth factor 1 signal pathway in vertebrate species. *Front. Biosci.*, **11**, 380–388.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Izarzugaza,J.M. *et al*. (2008) Enhancing the prediction of protein pairings between interacting families using orthology information. *BMC Bioinformatics*, **9**, 35.

Juan,D. *et al*. (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc. Natl Acad. Sci. USA*, **105**, 934–939.

Kersey,P. *et al*. (2005) Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res.*, **33**, D297–D302.

Labedan,B. *et al*. (2004) Using quaternary structures to assess the evolutionary history of proteins: the case of the aspartate carbamoyltransferase. *Mol. Biol. Evol.*, **21**, 364–373.

McPartland,J.M. *et al*. (2007) Coevolution between cannabinoid receptors and endocannabinoid ligands. *Gene*, **397**, 126–135.

Pazos,F. *et al*. (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.

Pazos,F. and Valencia,A. (2008) Protein co-evolution, co-adaptation and interactions. *EMBO J.*, **27**, 2648–2655.

Shoemaker,B.A. and Panchenko,A.R. (2007a) Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Comput. Biol.*, **3**, e42.

Shoemaker,B.A. and Panchenko,A.R. (2007b) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput. Biol.*, **3**, e43.

Uniprot Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.