BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 16 2012, pages 2186–2188
doi:10.1093/bioinformatics/bts331

Structural bioinformatics

Advance Access publication June 11, 2012

ETAscape: analyzing protein networks to predict enzymatic function and substrates in Cytoscape

Benjamin J. Bachman[1,2,3,†], Eric Venner[1,2,†], Rhonald C. Lua[1], Serkan Erdin[1] and Olivier Lichtarge[1,2,3,*]

[1]Departments of Molecular and Human Genetics, [2]Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030 and [3]W. M. Keck Center for Interdisciplinary Bioscience Training, Houston, TX 77005, USA

Associate Editor: Anna Tramontano

ABSTRACT

Summary: Most proteins lack experimentally validated functions. To address this problem, we implemented the Evolutionary Trace Annotation (ETA) method in the Cytoscape network visualization environment. The result is the ETAscape plugin, which builds a structural genomics network based on local structural and evolutionary similarities among proteins and then globally diffuses known annotations across the resulting network. The plugin displays these novel functional annotations, their confidence, the molecular basis for individual matches and the set of matches that lead to a prediction.

Availability: The ETA Network Plugin is available publicly for download at http://mammoth.bcm.tmc.edu/networks/.

Contact: lichtarge@bcm.edu

Received on February 21, 2012; revised on May 11, 2012; accepted on May 30, 2012

1 INTRODUCTION

The Structural Genomics Initiative (SGI) generates abundant structural data (Erdin *et al.*, 2011; Valencia, 2005), but many of these structures lack annotation (Redfern *et al.*, 2008). Computational methods that match small structural motifs of functionally important residues (a template) and suggest a function when the geometry is close enough (Laskowski *et al.*, 2005; Redfern *et al.*, 2009) are an especially promising way to approach this problem. A template can be constructed from prior knowledge of functional residues and mechanisms, or it can be created *de novo* by Evolutionary Trace (ET) analysis, which predicts functionally relevant amino acids and pinpoints functional sites using evolutionary principles (Lichtarge *et al.*, 1996). ET accuracy has been thoroughly tested both experimentally (Adikesavan *et al.*, 2011; Rodriguez *et al.*, 2010) and computationally (Mihalek *et al.*, 2004; Res *et al.*, 2005).

Evolutionary Trace Annotation (ETA) first maps the evolutionary importance of each residue onto a structure and selects a cluster of important surface residues as the template. It then seeks a match that is similar both geometrically and evolutionarily in protein structures with known function. These ETA templates usually overlap with catalytic sites (Ward *et al.*, 2008) and identify function with 87% accuracy at 61% coverage (Kristensen *et al.*, 2008).

Using a network in which structures form the nodes and ETA matches form the edges helps to overcome limitations from sparse functional data. We make predictions by allowing Enzyme Commision (EC) numbers to 'diffuse' through the network according to the cost function:

$$\sum_i (y_i - f_i)^2 + \alpha \sum_{i,j} w_{ij}(f_i - f_j)^2 \quad (1)$$

where the elements of y are 1, 0 or -1 depending on whether a protein is known to have, known not to have or is unknown to have a particular EC number. After minimization, f contains the 'diffused' values. The first term reflects the desire to not lose known information, and penalizes nodes whose function differs before and after diffusion. The second term reflects the fact that we expect neighboring proteins in this network to have similar functions, and punishes neighbors where this is not the case according to the edge weight. Repeating this process for all ECs yields a prediction for each possible function at each node. By normalizing the prediction scores across all nodes with unknown function, we create a prediction confidence. In benchmarks, the accuracy of this ETA diffusion network was >97% at 50% coverage, allowing the prediction and experimental confirmation of the function of an unannotated *Staphylococcus aureus* protein (Venner *et al.*, 2010).

This method is now made widely available and more transparent by embedding it into the Cytoscape network visualization environment (Smoot *et al.*, 2011). The ETAscape plugin allows users to view ETA networks, add proteins, make novel predictions, as well as view annotations, ETA templates and protein structures, adding to a public suite of ET tools that make functional site analysis and function prediction transparent (Lua and Lichtarge, 2010; Ward *et al.*, 2009).

2 OVERVIEW OF ETASCAPE

The plugin is available from mammoth.bcm.tmc.edu/networks. All commands are available as menu options, and a manual and tutorial video are available from the download page. A starting network of ETA matches between a subset of the Protein Data Bank (PDB) filtered for 90% sequence identity (Berman *et al.*, 2000) is included with the plugin. Node colors are based on known enzymatic function and the layout clusters similar proteins. ETA networks subdivide into

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
[*]To whom correspondence should be addressed.

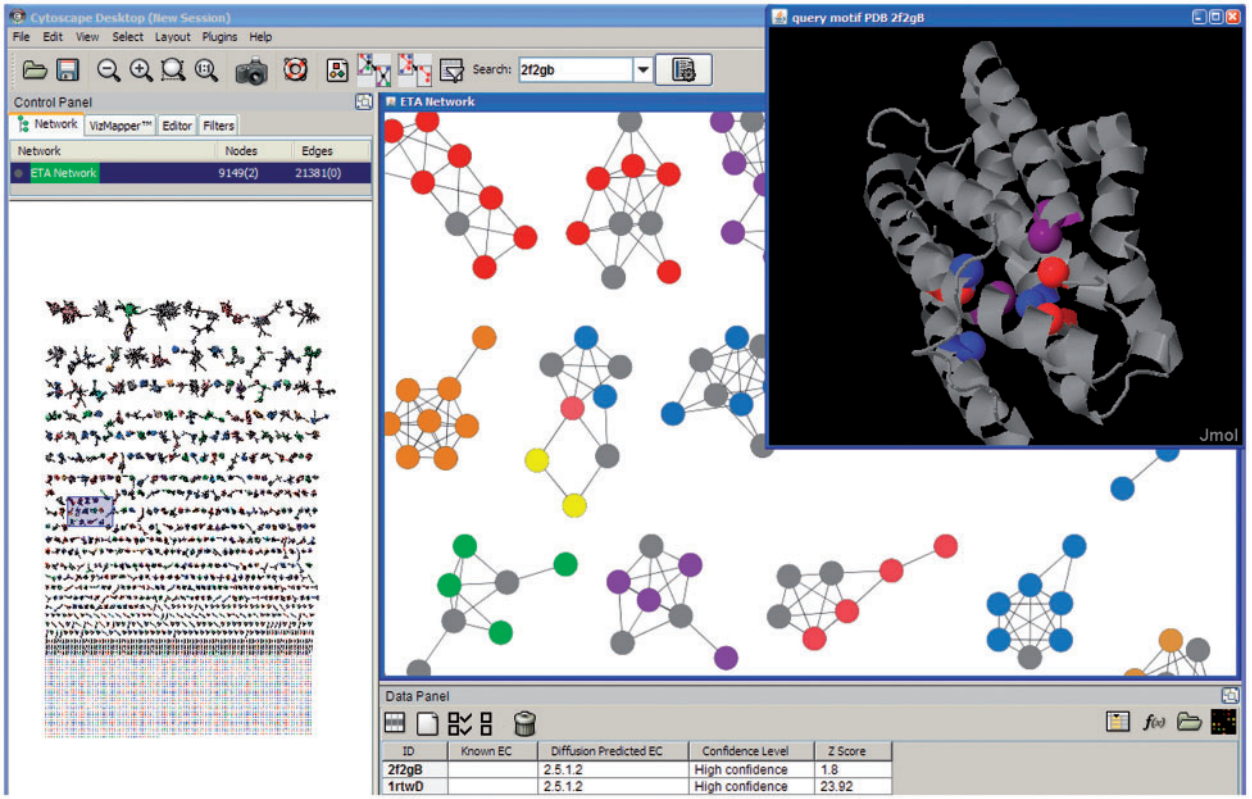

Fig. 1. Screen capture of the plugin. Nodes are proteins, with colors encoding the first two EC numbers. Edges are ETA template matches and indicate local structural and evolutionary similarity. Two unknown protein structures (yellow nodes in center) are highlighted and discussed in the Overview of ETAscape section below. A Jmol window (top right) shows the protein structure for 2f2gB and its mutual template match to 1rtwD (2f2gB's template is shown in blue and 1rtwD's is shown in red, purple indicates shared residues)

a large number of small networks due to the specificity filters. Right-clicking a node provides links to PDBsum (Laskowski, 2009) and the ET Server.

The 'Add new node to Network' menu option queries the ETA Server (Ward *et al.*, 2009), which opens in a browser and suggests an ET template that the user may customize. This template is then matched against proteins in the network and matches are filtered as described previously (Ward *et al.*, 2008). Modified networks may be saved and later reloaded.

Structures and ETA templates can be opened in Jmol (Hanson, 2010) windows by selecting nodes and running the 'Show Templates' menu option. The 'Run Diffusion' menu option predicts the function of unannotated proteins with our diffusion model (Venner *et al.*, 2010). Novel annotations and prediction confidences are available in the node attribute browser. The Show Influencing Proteins menu command shows proteins with the largest influence on the predicted function of the selected protein, often including nodes with strong indirect connections. After making predictions, users can export them to a file.

As an example, the plugin predicts that a protein expressed from gene locus At3g16990 in *Arabidopsis thaliana* (PDB ID 2f2gB), an SGI protein of unknown function, is a thiamine pyridinylase with Enzyme Classification number EC 2.5.1.2. (Fig. 1). Although the direct matches lack functional annotation, the software arrives at this prediction by diffusing the function across the intermediate links. There is one other function present in the subnetwork in proximity to the query protein (Aminopyrimidine aminohydrolase, EC 3.5.99.2). Interestingly, even though they are distinct reactions, both functions share the substrate Thiamine, possibly explaining the detected template similarity. 2f2gBs direct matches are well below the reliable homology range with sequence identities of 16% with 1rtwD and 14% with 1z72B. As observed previously (Ward *et al.*, 2008), the direct matches share overall structural similarity: many of the proteins in this cluster belong to the CATH heme oxygenase superfamily.

3 CONCLUSIONS

The ETAscape plugin extends an existing suite of protein function annotation tools to infer functional residues, identify functional sites and predict protein function (Lua and Lichtarge, 2010; Ward *et al.*, 2009). This tool pairs state-of-the-art network analysis with network visualization, putting the ability to generate novel predictions into the hands of researchers. Perhaps more importantly, it provides insight into the basis for those predictions.

ACKNOWLEDGEMENTS

Funding: The authors gratefully acknowledge grant support from the National Institute of Health, the National Science Foundation

and National Library of Medicine. National Institute of Health, NIH GM079656 and GM066099; National Science Foundation, NSF, CCF 0905536 and DBI 1062455; and from the National Library of Medicine NLM T15LM007093 through the Gulf Coast Consortia's Keck Center.

Conflict of Interest: none declared.

REFERENCES

Adikesavan,A.K. *et al.* (2011) Separation of recombination and SOS response in escherichia coli RecA suggests LexA interaction sites. *PLoS Genet*., **7**, e1002244.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Erdin,S. *et al.* (2011) Protein function prediction: towards integration of similarity metrics. *Curr. Opin. Struct. Biol.*, **21**, 180–188.

Hanson,R.M. (2010) Jmol a paradigm shift in crystallographic visualization. *J. Appl. Crystallogr*., **43**, 1250–1260.

Kristensen,D.M. *et al.* (2008) Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics*, **9**, 17.

Laskowski,R.A. *et al.* (2005) ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.*, **33**, W89–W93.

Laskowski,R.A. (2009) PDBsum new things. *Nucleic Acids Res.*, **37**, D355–D359.

Lichtarge,O. *et al.* (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.

Lua,R.C. and Lichtarge,O. (2010) PyETV: a PyMOL evolutionary trace viewer to analyze functional site predictions in protein complexes. *Bioinformatics*, **26**, 2981–2982.

Mihalek,I. *et al.* (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, **336**, 1265–1282.

Redfern,O.C. *et al.* (2008) Exploring the structure and function paradigm. *Curr. Opin. Struct. Biol.*, **18**, 394–402.

Redfern,O.C. *et al.* (2009) FLORA: a novel method to predict protein function from structure in diverse superfamilies. *PLoS Comput. Biol*., **5**, e1000485.

Res,I. *et al.* (2005) An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics*, **21**, 2496–2501.

Rodriguez,G.J. *et al.* (2010) Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. *Proc. Nat. Acad. Sci.*, **107**, 7787–7792.

Smoot,M.E. *et al.* (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.

Valencia,A. (2005) Automatic annotation of protein function. *Curr. Opin. Struct. Biol.*, **15**, 267–274.

Venner,E. *et al.* (2010) Accurate protein structure annotation through competitive diffusion of enzymatic functions over a network of local evolutionary similarities. *PLoS ONE*, **5**, e14286.

Ward,R.M. *et al.* (2008) De-orphaning the structural proteome through reciprocal comparison of evolutionarily important structural features. *PloS ONE*, **3**, e2136.

Ward,R.M. *et al.* (2009) Evolutionary trace annotation server: automated enzyme function prediction in protein structures using 3D templates. *Bioinformatics*, **25**, 1426–1427.