BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 11 2009, pages 1426–1427
doi:10.1093/bioinformatics/btp160

Structural bioinformatics

Evolutionary Trace Annotation Server: automated enzyme function prediction in protein structures using 3D templates

R. Matthew Ward[1,2,3], Eric Venner[1,2], Bryce Daines[1], Stephen Murray[1], Serkan Erdin[1,3], David M. Kristensen[1,2,3] and Olivier Lichtarge[1,2,3,4,5,*]

[1] Departments of Molecular and Human Genetics, [2]Program in Structural and Computational Biology and Molecular, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, [3]W. M. Keck Center for Interdisciplinary Bioscience Training, Houston, TX 77005, [4]Department of Pharmacology and [5]Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

Received on October 16, 2008; revised on February 06, 2009; accepted on March 16, 2009

Advance Access publication March 23, 2009

Associate Editor: Thomas Lengauer

ABSTRACT

Summary: The Evolutionary Trace Annotation (ETA) Server predicts enzymatic activity. ETA starts with a structure of unknown function, such as those from structural genomics, and with no prior knowledge of its mechanism uses the phylogenetic Evolutionary Trace (ET) method to extract key functional residues and propose a function-associated 3D motif, called a 3D template. ETA then searches previously annotated structures for geometric template matches that suggest molecular and thus functional mimicry. In order to maximize the predictive value of these matches, ETA next applies distinctive specificity filters—evolutionary similarity, function plurality and match reciprocity. In large scale controls on enzymes, prediction coverage is 43% but the positive predictive value rises to 92%, thus minimizing false annotations. Users may modify any search parameter, including the template. ETA thus expands the ET suite for protein structure annotation, and can contribute to the annotation efforts of metaservers.

Availability: The ETA Server is a web application available at http://mammoth.bcm.tmc.edu/eta/.

Contact: lichtarge@bcm.edu

1 INTRODUCTION

As the number of protein structures mushrooms, in large part due to structural genomics (SG) efforts, a detailed knowledge of their biological roles remains elusive (Redfern *et al*., 2008). Thus most Protein Data Bank (PDB) (Berman *et al*., 2000) annotations are computationally rather than experimentally derived, and still 28% of the 2191 SG proteins solved last year were labeled 'unknown' or 'hypothetical' as of September 2008.

Annotation transfer among homologs identified by PSI-BLAST (Altschul *et al*., 1997) or similar tools remains the most popular and useful method. The problem is that homology does not guarantee functional equivalence, as often divergence yields proteins of different functions (Gerlt and Babbitt, 2000). Even at 65% sequence identity, 10% of protein pairs already have different 4-digit Enzyme Commission (EC) functions, and at 45% identity 10% differ in the less specific 3-digit functions (Tian and Skolnick, 2003). This leads to errors that propagate, dramatically decreasing the effectiveness of future predictions (Brenner, 1999). Increasing annotation specificity is therefore paramount.

To this end, an orthogonal approach relies on 3D templates: small structural motifs built from key amino acid functional determinants that suggest functional similarity when matched geometrically in unannotated proteins (Wallace *et al*., 1997). Two such methods are in the popular ProFunc metaserver: Enzyme Active Sites and Reverse Templates (Laskowski *et al*., 2005). Because 3D templates are local and narrowly focus on the molecular basis of function, they can remain accurate even when overall similarity becomes unreliably low, or when it remains so high as to obscure a key functional site variation. However, 3D template annotations also have weaknesses: a lack of known functional determinants from which to build them on a large scale, and low specificity when derived heuristically (Kristensen *et al*., 2006).

To build templates without any prior knowledge of the catalytic mechanism, the Evolutionary Trace Annotation (ETA) (Kristensen *et al*., 2006; 2008) server heuristically selects residues based on Evolutionary Trace (ET) predictions of functional sites in protein structures (Lichtarge *et al*., 1996). These predictions were extensively validated experimentally (Onrust *et al*., 1997; Ribes-Zamora *et al*., 2007; Sowa *et al*., 2001) and computationally (Mihalek *et al*., 2004; Res *et al*., 2005; Yao *et al*., 2003). Moreover, ETA templates either overlap catalytic residues (78%), or lie in their immediate vicinity (22%) (Ward *et al*., 2008).

To raise specificity, ETA filters geometric template matches (i) by ET rank similarity (Kristensen, *et al*., 2006); (ii) by match reciprocity back to the original protein (Ward *et al*., 2008); and (iii) by the extent that a plurality of matches point to the same function (Kristensen *et al*., 2008). In 1218 SG control enzymes, ETA made 527 predictions, i.e. 43% prediction coverage, of which 478 were true, for 92% positive predictive value (PPV). ETA's performance improves on the Enzyme Active Site and Reverse Template methods from ProFunc (Ward *et al*., 2008). ETA also proved complementary to sequence-based methods (Kristensen *et al*., 2008). If needed, prediction coverage can be raised to 77%

*To whom correspondence should be addressed.

(934/1218) by including non-reciprocal matches, but PPV then decreases to 82% (769/934) PPV.

2 ETA SERVER OVERVIEW

The ETA Server provides functional annotations of enzyme activity. A web interface lets users pick a protein. The server then automatically creates a template, identifies matches to annotated structures, applies specificity filters, and reports likely functions. Backtracking is possible, and users can alter the template. In-line help is available, as well as a manual with a complete walkthrough example.

2.1 Template creation

Users select a protein by PDB code and chain (e.g. 1yvwA). The server then either retrieves a cached ET analysis or runs one anew for this protein (∼5 min). The user may also submit custom ET data as a zip file from the ET Wizard (Morgan *et al*., 2006), allowing full control of the ET analysis, or use of novel structures.

Next, ETA builds a template of C_α atoms from the six best-ranked residues in a cluster of 10 surface ET residues (Kristensen *et al*., 2008). A PyMOL (DeLano, 2002) image of the protein structure displays the template so the user may see and revise the residue choices, triggering image updates. A PyMOL session can also be downloaded to study the template interactively.

For a given template, the server displays the amino acid types that it can match in another protein, chosen from cognate residues in homologs. All choices are customizable.

2.2 Geometric search and annotation

The residue numbers and types form a complete template that is searched against proteins in the 2006 PDB_SELECT_90 (Hobohm *et al*., 1992). A support vector machine (Ward *et al*., 2008) classifies the most relevant matches, using geometric (least root mean squared deviation, RMSD) and evolutionary similarity features (difference in ET score) (Kristensen *et al*., 2006). Reciprocally, templates from each protein in the PDB_SELECT_90 are also searched back against the query structure. Matches are grouped by function and whether they are reciprocal.

Annotations fall in two classes: those exclusively from reciprocal matches, which are the most reliable; and those that also rely on one-way matches, which are more sensitive but less specific. In both cases, the enzymatic function with a plurality of matches is listed first, followed by possible alternatives. These functions—three-digit EC numbers—are linked to their definitions. Matches to non-enzymes and unannotated proteins are also displayed, as they may still provide useful information.

Each match that supports a given prediction is listed, with a link to the relevant PDB structure, a list of matched residues, their RMSD, and their ET similarity. Images of the template and match can also be generated to review them visually. All the raw ET and ETA data can be downloaded.

3 CONCLUSIONS

The ETA server expands the ET suite for protein structure annotation (Mihalek *et al*., 2006; Morgan *et al*., 2006) by predicting enzymatic functions of protein structures without prior knowledge of functional sites or mechanisms. In reciprocal mode, it is biased to minimize misannotations by maximizing PPV (92%) at the expense of prediction coverage (43%). In all-match mode, prediction coverage is better (77%), but then PPV is lower (82%). The interface allows customized searches, displays predicted functions, and provides supporting evidence and raw data. Eventually, upgrades should add non-enzymatic function predictions as well. Feedback and suggestions are welcome at etaserver@bcm.edu.

ACKNOWLEDGEMENTS

We thank Ms. Deepti Karandur for testing the server, and Dr. Cindy Ly for proof reading the manuscript.

Funding: National Science Foundation (DBI-0547695 to O.L.); National Institute of Health (R01-GM066099, R01-GM079656); March of Dimes (1-FY06-371); Keck Center for Interdisciplinary Bioscience Training (National Library of Medicine grant no. 5T15LM07093, training fellowships to R.M.W., S.E. and D.M.K.).

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.

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

Brenner,S.E. (1999) Errors in genome annotation. *Trends Genet.*, **15**, 132–133.

DeLano,W.L. (2002) *The PyMOL Molecular Graphics System*. DeLano Scientific, Palo Alto, CA.

Gerlt,J.A. and Babbitt,P.C. (2000) Can sequence determine function? *Genome Biol.*, **1**, REVIEWS0005.

Hobohm,U. *et al*. (1992) Selection of representative protein data sets. *Protein Sci.*, **1**, 409–417.

Kristensen,D.M. *et al*. (2006) Recurrent use of evolutionary importance for functional annotation of proteins based on local structural similarity. *Protein Sci.*, **15**, 1530–1536.

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.

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

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

Mihalek,I. *et al*. (2006) Evolutionary trace report_maker: a new type of service for comparative analysis of proteins. *Bioinformatics*, **22**, 1656–1657.

Morgan,D.H. *et al*. (2006) ET viewer: an application for predicting and visualizing functional sites in protein structures. *Bioinformatics*, **22**, 2049–2050.

Onrust,R. *et al*. (1997) Receptor and betagamma binding sites in the alpha subunit of the retinal G protein transducin. *Science*, **275**, 381–384.

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

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

Ribes-Zamora,A. *et al*. (2007) Distinct faces of the Ku heterodimer mediate DNA repair and telomeric functions. *Nat. Struct. Mol. Biol.*, **14**, 301–307.

Sowa,M.E. *et al*. (2001) Prediction and confirmation of a site critical for effector regulation of RGS domain activity. *Nat. Struct. Biol.*, **8**, 234–237.

Tian,W. and Skolnick,J. (2003) How well is enzyme function conserved as a function of pairwise sequence identity? *J. Mol. Biol.*, **333**, 863–882.

Wallace,A.C. *et al*. (1997) TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.*, **6**, 2308–2323.

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

Yao,H. *et al*. (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.*, **326**, 255–261.