Nucleic Acids Research Advance Access published May 13, 2009

Nucleic Acids Research, 2009, 1–6
doi:10.1093/nar/gkp339

INTREPID: a web server for prediction of functionally important residues by evolutionary analysis

Sriram Sankararaman[1], Bryan Kolaczkowski[2] and Kimmen Sjölander[2,3,*]

[1]Department of Computer Science, [2]Department of Bioengineering and [3]Department of Plant and Microbial Biology, University of California, Berkeley, USA

Received March 3, 2009; Revised April 13, 2009; Accepted April 21, 2009

ABSTRACT

We present the INTREPID web server for predicting functionally important residues in proteins. INTREPID has been shown to boost the recall and precision of catalytic residue prediction over other sequence-based methods and can be used to identify other types of functional residues. The web server takes an input protein sequence, gathers homologs, constructs a multiple sequence alignment and phylogenetic tree and finally runs the INTREPID method to assign a score to each position. Residues predicted to be functionally important are displayed on homologous 3D structures (where available), highlighting spatial patterns of conservation at various significance thresholds. The INTREPID web server is available at http://phylogenomics.berkeley.edu/intrepid.

INTRODUCTION

Predicting the functional residues in a protein is an important and challenging problem in bioinformatics. Prediction methods provide a useful starting point to understand the functions of proteins and serve to prioritize site-directed mutagenesis experiments. A variety of approaches have been developed, including those that primarily or exclusively exploit sequence information [e.g. (1–6)] and those that make use of both sequence and structure information [e.g. (7–9)]. INTREPID (Information-theoretic Tree Traversal for Protein Functional Site Identification) falls into the first class of sequence-only methods, and uses phylogenetic analysis of a family of homologous sequences to identify positions that are conserved at different levels of an evolutionary tree. It has been shown to outperform other sequence-only methods at detecting catalytic sites in proteins (10).

Structural information clearly provides a significant boost in prediction accuracy, but is available for only a small fraction of proteins. For this reason, sequence-based methods play a key role in bioinformatics prediction of functional residues. Sequence-only methods for functional residue prediction are based on the assumption that mutations disrupting function will not be tolerated by evolution; i.e. we can exploit nature's mutagenesis experiments to reveal positions playing critical roles. Such positions are identified using multiple sequence alignment (MSA) analysis: positions that display a high degree of conservation against a backdrop of divergence across related sequences can be posited as functionally (or perhaps structurally) important.

Notably, while sequence-based methods depend on the evolutionary context to predict functional residues, many methods include only moderately divergent homologs in an alignment. This restriction may be designed to limit the changes in function and structure that accumulate with evolutionary distance (11), to reduce the likelihood of alignment errors (12,13) or for reasons of computational efficiency, but effectively reduces the total possible information available to a method.

INTREPID is designed to make full use of the information in a protein family containing many distantly related sequences through the use of tree traversal: INTREPID computes an information-theoretic score for each position in the sequence at each subtree encountered on a path from the root to the leaf corresponding to the sequence of interest. The score for each position is set to the maximal score obtained on the path. Positions that are conserved across the entire family obtain their maximum score at the root of the tree, while other positions will achieve maximum scores at nodes corresponding to one of the nested subtrees. This tree traversal gives INTREPID the ability to detect subtle evolutionary patterns that other methods might miss. For instance, positions that are critical for function for one subfamily but variable across the family as a whole may remain undetected by prediction methods that do not use tree

*To whom correspondence should be addressed. Tel: +1 510 642 9932; Fax: +1 510 642 5835; Email: kimmen@berkeley.edu

traversal to extract the evolutionary signal. In contrast, since INTREPID computes the importance of each position within all subtrees containing the sequence of interest, it suffices for a position to become conserved within some subtree for it to be detectable. In analysis of enzymes from the manually curated section of the Catalytic Site Atlas (14), INTREPID has been shown to achieve significantly higher levels of recall and precision at catalytic residue prediction than other sequence-based methods, including those that use evolutionary tree analysis, primarily due to the use of tree traversal to mine the information contained in highly divergent datasets (10).

THE INTREPID WEB SERVER

The INTREPID web server provides a pipeline for homolog selection and alignment, phylogenetic tree construction, identification of homologous 3D structures and calculation of INTREPID scores. It is available at http://phylogenomics.berkeley.edu/intrepid/.

INPUT

The input to the web server is a protein sequence (seed) in FASTA format. The user can supply an email address to which results will be sent, or bookmark the page displayed after clicking 'Submit'.

PROCESSING STEPS

Initial processing steps

The INTREPID functional residue prediction server is a multi-stage automated analysis pipeline. We start by gathering homologs from the UniProt protein sequence database using PSI-BLAST. Since INTREPID performance improves as sequence divergence in the dataset increases (10), we use four iterations of PSI-BLAST with an *E*-value inclusion cutoff of $1e-04$. We retrieve the full-length proteins for all hits found by PSI-BLAST and align these sequences using MUSCLE (15). The MSA is edited to remove columns corresponding to a gap in the seed, followed by removing sequences that are composed entirely of gaps. A Neighbor-Joining phylogenetic tree is estimated using the QuickTree software (16) and rooted at the midpoint of the longest span. The tree and MSA are then given as input to the INTREPID algorithm. We run additional analyses such as identification of PFAM domains and prediction of transmembrane helices, retrieve GO annotations and biological literature, identify homologous 3D structures and gather other data from various external resources. These are not used by INTREPID, but are provided to the user to assist in interpreting INTREPID program output.

INTREPID scoring

Given an MSA, a corresponding phylogenetic tree and a specified seed sequence, INTREPID examines each subtree containing the seed (i.e. corresponding to the subtrees encountered on the path from the root to the seed sequence leaf). For each position in the seed and for each subtree independently, INTREPID computes the Jensen–Shannon divergence between the amino acid distribution at the position and the background distribution in the subtree. This produces a set of scores for each position in the sequence, one for each subtree encountered in the tree traversal. Upon reaching the leaf, the maximal score for each position is identified. The scores are then normalized and sorted to produce a rank-ordered list. Details on the algorithm are available in (10).

Identification of homologous 3D structures

We search for homologous 3D structures using an HMM constructed from the MSA using the UCSC SAM w0.5 software (17). Note that INTREPID does not make use of structural information; the PDB structures are simply provided to enhance user interpretation of results.

Time to complete

The bulk of the INTREPID computational complexity is due to phylogenetic tree construction (in contrast, computing the INTREPID scores is very fast once an MSA and tree are available, and typically takes <5 min). The time required from input to final results can range from under 15 min to many hours, depending primarily on the number of sequences retrieved.

OUTPUT

Two links are provided, either sent by email or on the web page bookmarked by the user. The first link gives the INTREPID score results (Figure 1). The second link is to a web page displaying the alignment, phylogenetic tree and additional data for the protein family and includes results from a second functional site prediction protocol (Figure 2).

INTREPID results

Results include a normalized score for each residue in the input protein sequence. Scores can be viewed on the output page or downloaded in CSV format. If a homologous PDB structure can be detected, these scores will be used to highlight top-ranked residues on the structure; if not, we report, 'This sequence has no evident homology to any PDB structure.' Homologous PDB structures are displayed using the Jmol Java-based structure viewer (18). The top 10% of residues are colored by default. A scale on the right of the Jmol display shows the range of INTREPID scores and allows the user to select an appropriate threshold. A link is also provided to allow users to view scores for all residues and to sort the results based on residue position, residue type and INTREPID score (Figure 1).

Protein family page

This page includes an MSA, phylogenetic tree, predicted subfamilies, hidden Markov models for the family and subfamilies, GO annotations, matching PFAM domains and homologous PDB structures,

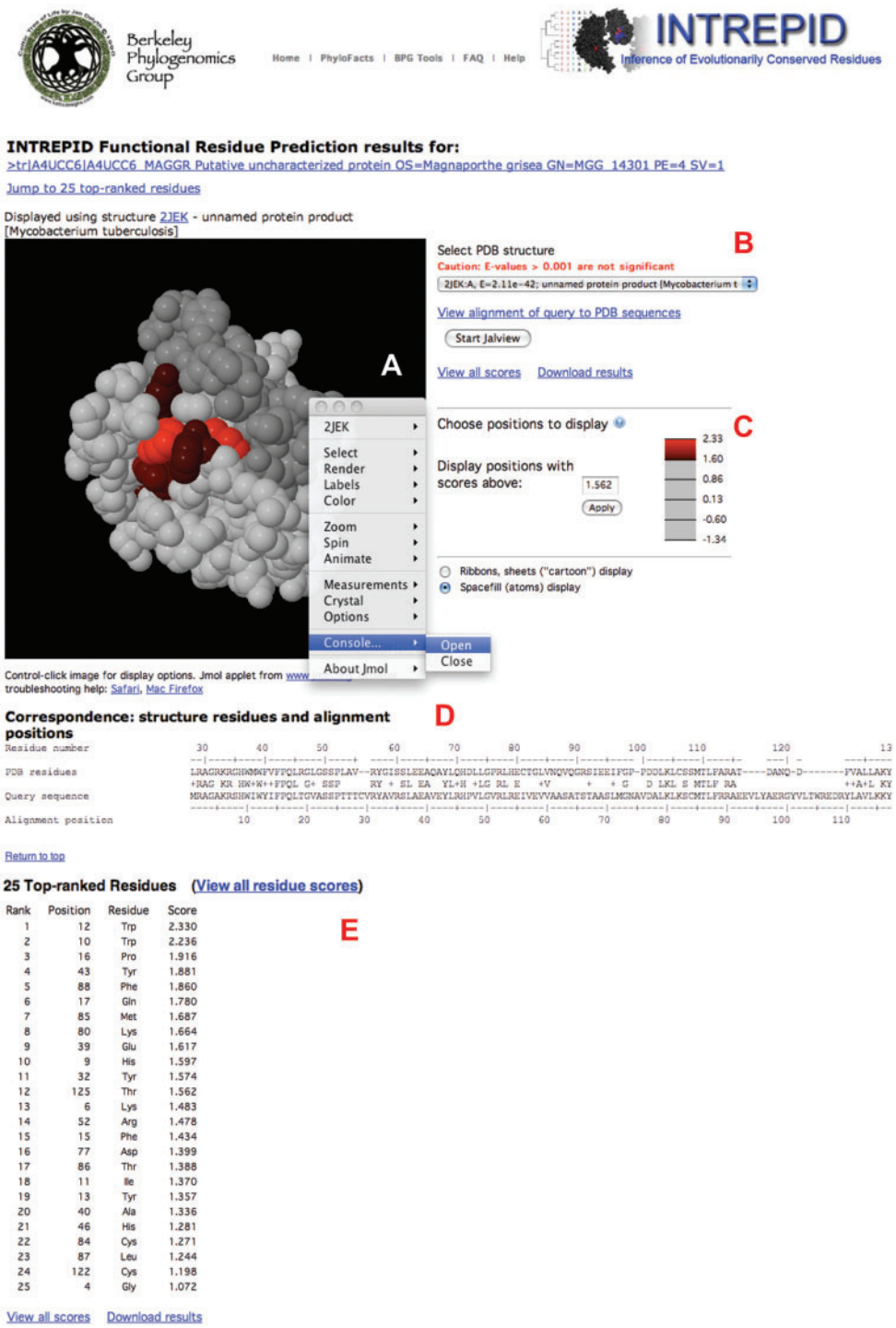

Figure 1. INTREPID program output. This figure shows the first result returned for input sequence A4UCC6 from the UniProt database. (**A**) Homologous structure(s) displayed using the Jmol structure viewer. Clicking on the structure pane (Mac: Ctrl-click; PC: Right click) brings up an interactive mode; selecting the console permits users to highlight or modify the display of individual residues or regions in the structure. (**B**) Users can select different PDB structures for viewing using the pull-down menu; scores displayed are the HMM-based *E*-values. (**C**) Users can change the score cutoff used to select residues for highlighting. (**D**). The pairwise alignment of the selected PDB structure and the query, based on alignment to the HMM constructed for the family. (**E**) The top 25-ranked residues are displayed. Users can also view all residue scores; score files can be downloaded by following the link at the bottom of the page.

obtained using a slightly modified version of the PhyloBuilder pipeline (19). These data are displayed on the web page using the same formatting as PhyloFacts family 'books' (20). The structure icon is linked to a page presenting results from a different method for predicting functional sites (using family and subfamily conservation patterns) and is provided as a complementary analysis to the INTREPID scores. Interactive viewers

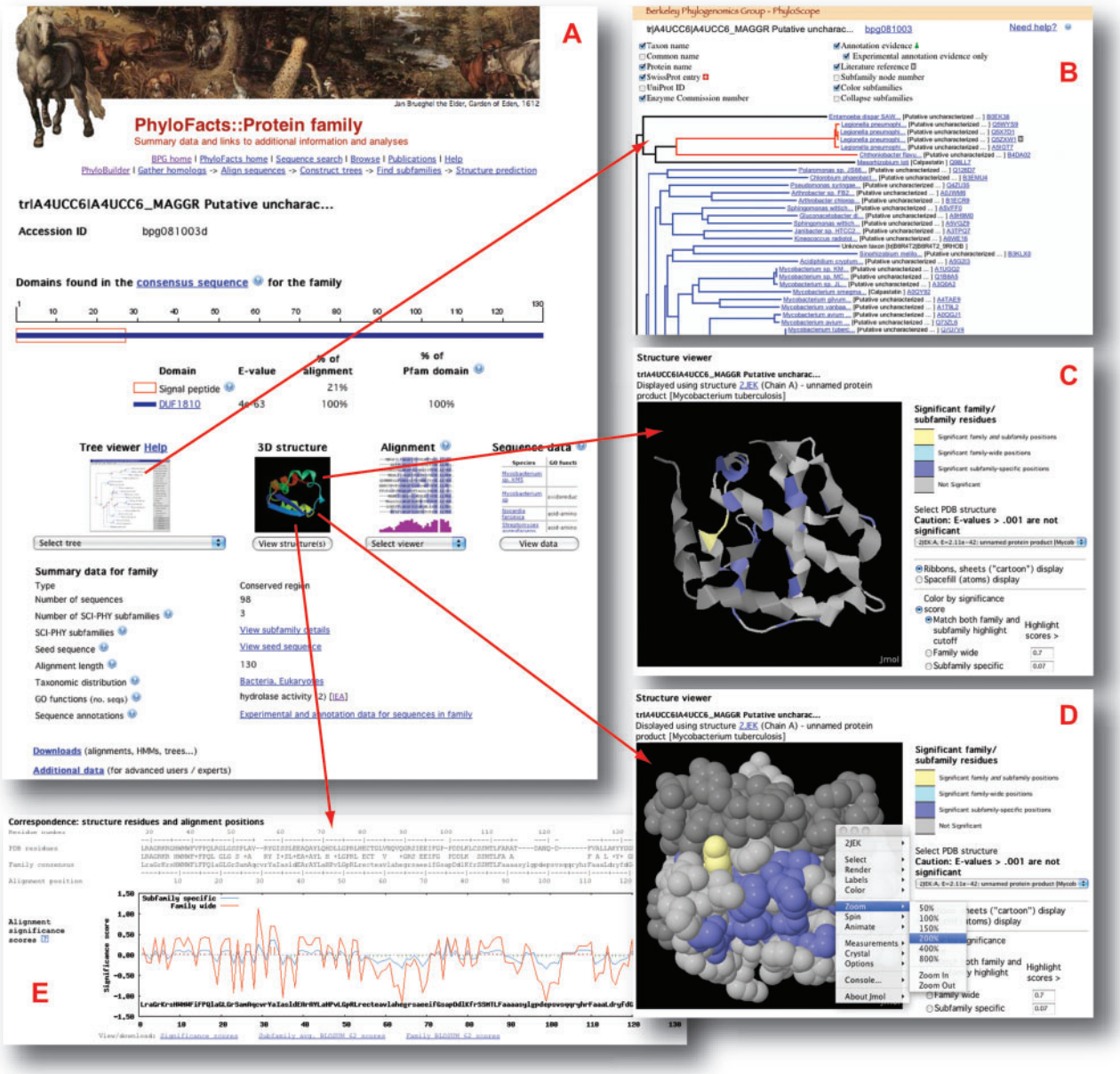

Figure 2. Result of PhyloBuilder family analysis. (**A**) The PhyloBuilder 'book', showing a variety of data for the family. *Top*: homologous PFAM domains. *Middle*: phylogenetic tree, homologous structures, multiple sequence alignment (viewed with Jalview or in hypertext), and a spreadsheet of data. *Bottom*: Summary information for the family, including GO annotations and evidence codes, taxonomic distribution and predicted subfamilies using the SCI-PHY algorithm. The MSA, phylogenetic tree and HMMs for the family and subfamily can be downloaded from the Downloads link at bottom. (**B**) The phylogenetic tree for the family, displayed using the PhyloScope viewer. Subtrees of different colors indicate subfamilies predicted by SCI-PHY. Icons are linked to data from external resources: Swiss flags indicate manually curated SwissProt sequences; green flasks indicate experimental support for assigned functions; page icons indicate one or more publications are available. (**C**) Homologous structures are displayed using Jmol. Positions are colored according to their conservation pattern (e.g. light blue means family wide conservation and dark blue means subfamily specific conservation). (**D**) Space-fill view of the same structure. Clicking on the structure pane (Mac: Ctrl-click; PC: Right click) brings up an interactive mode, allowing the user to select different display options, including zoom. (**E**) Plot of conservation patterns at the family (red) and subfamily specific (blue) levels. The alignment of the PDB structure to the family consensus is displayed (derived from aligning the PDB sequence to the HMM for the family).

are provided for the phylogenetic tree, structure and MSA. We provide two phylogenetic tree viewers: the Java-based ATV software (21) and a new Javascript phylogenetic tree viewer, PhyloScope. The alignment, tree and HMMs can be downloaded from the web site (Figure 2).

The PhyloScope phylogenetic tree viewer

The PhyloScope viewer is designed to enable biologists to predict the function of proteins in a phylogenomic context, by overlaying experimental data, biological literature and other information on the phylogenetic tree. Large trees are collapsed to functional subfamilies

identified using SCI-PHY (22). Subtrees can be collapsed or expanded individually by clicking on internal nodes (or collapsed terminal nodes). The availability of experimental data, literature and external resources with additional data is indicated by icons that are hyperlinked to these resources (Figure 2b).

INTREPID program dependencies

INTREPID depends on the availability of homologs to predict functional positions. If fewer than four sequences are retrieved using PSI-BLAST, the program terminates with the message 'Functional Residue Prediction – terminated. Too few homologs were found to perform phylogenomic analysis.'

DISCUSSION

The INTREPID server predicts functional residues based on sequence information alone. The tree traversal enables INTREPID to identify sites that are important for the query sequence even if they are variable across the family as a whole. As a result, INTREPID is robust to functional divergence in specific lineages of the family and to alignment and phylogeny errors. The computational efficiency of the tree traversal allows INTREPID to scale to very large and divergent protein families, enabling it to exploit the information in these families to boost its predictive power. In fact, we have shown in (10) that the accuracy of INTREPID improves as we increase the evolutionary divergence of the input data.

The INTREPID web server takes as input a protein sequence, retrieves and aligns homologs, constructs a phylogenetic tree and searches for homologous 3D structures. The time required to complete these analyses can vary from a few minutes to several hours, dependent primarily on the number of homologs retrieved. Results are plotted on homologous 3D structures, when available, and can be downloaded or viewed online. Additional outputs of the INTREPID web server include a web page containing an MSA, phylogenetic tree, a spreadsheet of data, predicted subfamilies, homologous 3D structures, Gene Ontology annotations with evidence codes, biological literature and hidden Markov models for the protein family as a whole.

FUTURE PLANS

Our future plans for this server include expanding the functionality to utilize information from 3D structure, where available, and providing for prediction of specificity-determining residues (e.g. substrate recognition) and positions involved in protein–protein interaction. We also plan to construct comparative (homology) models for user-submitted sequences to assist in interpretation of scores.

ACKNOWLEDGEMENTS

We thank Dr Tom Alber and anonymous referees for helpful comments on the manuscript and recommendations for future developments of the INTREPID web server.

FUNDING

Presidential Early Career Award for Scientists and Engineers (from the National Science Foundation, grant number 0238311 to K.S.); Microbial Genome Sequencing Program of the National Science Foundation (grant number 0732065 to K.S.); National Institutes of Health (grant number HG002769 to K.S.). Funding for open access charge: National Institutes of Health grant HG002769 and National Science Foundation grant 0732065 to KS.

Conflict of interest statement. None declared.

REFERENCES

1. Glaser,F., Pupko,T., Paz,I., Bell,R.E., Bechor-Shental,D., Martz,E. and Ben-Tal,N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**, 163–164.
2. Mihalek,I., Res,I. and Lichtarge,O. (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, **336**, 1265–1282.
3. Carro,A., Tress,M., de Juan,D., Pazos,F., Lopez-Romero,P., del Sol,A., Valencia,A. and Rojas,A.M. (2006) TreeDet: a web server to explore sequence space. *Nucleic Acids Res.*, **34**, W110–W115.
4. Feenstra,K.A., Pirovano,W., Krab,K. and Heringa,J. (2007) Sequence harmony: detecting functional specificity from alignments. *Nucleic Acids Res.*, **35**, W495–W498.
5. Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.
6. Kalinina,O.V., Mironov,A.A., Gelfand,M.S. and Rakhmaninova,A.B. (2004) Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci.*, **13**, 443–456.
7. Youn,E., Peters,B., Radivojac,P. and Mooney,S.D. (2007) Evaluation of features for catalytic residue prediction in novel folds. *Protein Sci.*, **16**, 216–226.
8. Gutteridge,A., Bartlett,G.J. and Thornton,J.M. (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.*, **330**, 719–734.
9. Petrova,N.V. and Wu,C.H. (2006) Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. *BMC Bioinformatics*, **7**, 312.
10. Sankararaman,S. and Sjölander,K. (2008) INTREPID—INformation-theoretic TREe traversal for Protein functional site IDentification. *Bioinformatics*, **24**, 2445–2452.
11. Brown,D. and Sjölander,K. (2006) Functional classification using phylogenomic inference. *PLoS Comput. Biol.*, **2**, e77.
12. Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.
13. Edgar,R.C. and Sjölander,K. (2004) A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics*, **20**, 1301–1308.
14. Porter,C.T., Bartlett,G.J. and Thornton,J.M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.
15. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
16. Howe,K, Bateman,A and Durbin,R. (2002) QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics*, **18**, 1546–1547.
17. Hughey,R., Karplus,K., Krogh,A., Diekhans,M., Grate,L., Barrett,C., Brown,M., Cline,C., Figel,T., Karchin,R. *et al.* (2000),

Sequence Alignment and Modeling (SAM) software. *Technical Report*, UC Santa Cruz.
18. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.Jmol.org/.
19. Glanville,J.G., Kirshner,D., Krishnamurthy,N. and Sjölander,K. (2007) Berkeley Phylogenomics Group web servers: resources for structural phylogenomic analysis. *Nucleic Acids Res.*, **35**, W27–W32.
20. Krishnamurthy,N., Brown,D.P., Kirshner,D. and Sjölander,K. (2006) PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. *Genome Biol.*, **7**, R83.
21. Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
22. Brown,D.P., Krishnamurthy,N. and Sjölander,K. (2007) Automated protein subfamily identification and classification. *PLoS Comput. Biol.*, **3**, e160.