W104–W109 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl171

TreeDomViewer: a tool for the visualization of phylogeny and protein domain structure

Blaise T. F. Alako[1,2], Daphne Rainey[3], Harm Nijveen[1] and Jack A. M. Leunissen[1,*]

[1]Laboratory of Bioinformatics, Wageningen University and Research Centre, PO Box 8128, 6700 ET Wageningen, The Netherlands, [2]Centre for BioSystems Genomics, PO Box 98, 6700 AB Wageningen, The Netherlands and [3]KEYGENE NV, PO Box 216 6700 AE Wageningen, The Netherlands

Received February 14, 2006; Revised and Accepted March 20, 2006

ABSTRACT

Phylogenetic analysis and examination of protein domains allow accurate genome annotation and are invaluable to study proteins and protein complex evolution. However, two sequences can be homologous without sharing statistically significant amino acid or nucleotide identity, presenting a challenging bioinformatics problem. We present TreeDomViewer, a visualization tool available as a web-based interface that combines phylogenetic tree description, multiple sequence alignment and InterProScan data of sequences and generates a phylogenetic tree projecting the corresponding protein domain information onto the multiple sequence alignment. Thereby it makes use of existing domain prediction tools such as InterProScan. TreeDomViewer adopts an evolutionary perspective on how domain structure of two or more sequences can be aligned and compared, to subsequently infer the function of an unknown homolog. This provides insight into the function assignment of, in terms of amino acid substitution, very divergent but yet closely related family members. Our tool produces an interactive scalar vector graphics image that provides orthological relationship and domain content of proteins of interest at one glance. In addition, PDF, JPEG or PNG formatted output is also provided. These features make TreeDomViewer a valuable addition to the annotation pipeline of unknown genes or gene products. TreeDomViewer is available at http://www.bioinformatics.nl/tools/treedom/.

INTRODUCTION

The past years have seen the rapid sequencing of genomes from many different organisms. Sequencing itself is no longer the bottleneck in genome studies; the bottleneck is a reliable annotation of new genes. Information from widely studied model species included in comparative annotation genomics has greatly aided in these annotation efforts, and proofed to be extremely valuable in contributing to the understanding of protein evolution (1). Sometimes homologous gene products have strong sequence similarities so that the inference of homology is straightforward. However, accumulation of multiple substitutions in the course of divergent evolution can make homologous sequences as dissimilar as any two proteins chosen randomly from a database (2).

Several bioinformatics approaches have been developed to identify remote homology in the absence of pairwise similarity, one of the popular ones being protein fold recognition (FR) (3). Briefly, FR detects homology based on the combination of evolutionary criteria and structural considerations. FR differs from traditional sequence homology database searches insofar as the databases to be searched by FR contain only proteins with experimentally determined structures rather than all protein sequences. Hence, the availability of a related structure in the Protein Data Bank is an essential but not sufficient prerequisite for the success of FR-based identification of homologs (4). However, homology is defined on the basis of evolution rather than function. Homologues can fulfill different functions and share only very general similarities; even orthologs may fulfill non-identical roles (5).

Moreover, homology is not necessarily a one-to-one relationship, because a single gene in one genome may correspond to a whole family of paralogs in another genome, which may be functionally diverse. Hence one pitfall is often, correctly defining orthologs when annotating unknown protein or gene function by homology, using either simple or sophisticated existing bioinformatics tools (4).

Currently there is a multitude of tools available for the visualization of information contained within a protein sequence such as signal peptides (6), transmembrane domains (7,8) and functional domains [e.g. InterProScan (9)]. The latter currently comprises 15 domain prediction methods.

However, until now there is no tool available combining in one view protein sequence analysis with orthology information, thereby essentially combining protein information with phylogeny [see e.g. (10)] independent of the available 3D structure in databases.

*To whom correspondence should be addressed. Tel: +31 317 482 036; Fax: +31 317 483 584; Email: jack.leunissen@wur.nl

In this paper, we present a more convenient way of identifying putative family members based on their evolutionary history. We examine the conservation of structural and functional domains which, unlike amino acid substitution, occurs at a slower rate throughout evolution. The domains examined are often predicted by robust HMMs, which allow definition of a domain to remain stable with multiple amino acid substitutions, thus giving a more accurate analysis on the presence of this domain.

This phylogenetic visualization tool allows a rapid 'first pass' quality screening of search results from InterProScan and others [e.g. the EMBOSS package (11)]. One of its strengths is the forthright generation of a publication-quality graphical output. TreeDomViewer is available as a Perl-based web interface that accepts a multiple sequence file in any common format as input and produces a phylogenetic tree with the corresponding protein domain information projected onto the multiple sequence alignment next to it. Although a powerful tool by itself, TreeDomViewer is obviously dependent on the quality of the analysis tools and multiple alignments.

IMPLEMENTATION AND DESIGN

Data preparation and processing

The minimal input required by TreeDomViewer is a set of aligned or unaligned sequences. In case where the input file is a sequence file solely, ClustalW (12) is used to align the sequences and a tree description is calculated subsequently using ClustalW's built-in neighbor-joining option (13).

By default TreeDomViewer combines the output from several programs, i.e. a multiple alignment (in any common sequence format, such as FASTA or Clustal), a phylogenetic tree [in standard Newick or PHYLIP format (14)] and domain predictions (in InterProScan's 'raw' format).

The ability to upload precalculated files makes the tool very flexible, as the user may want to upload the output

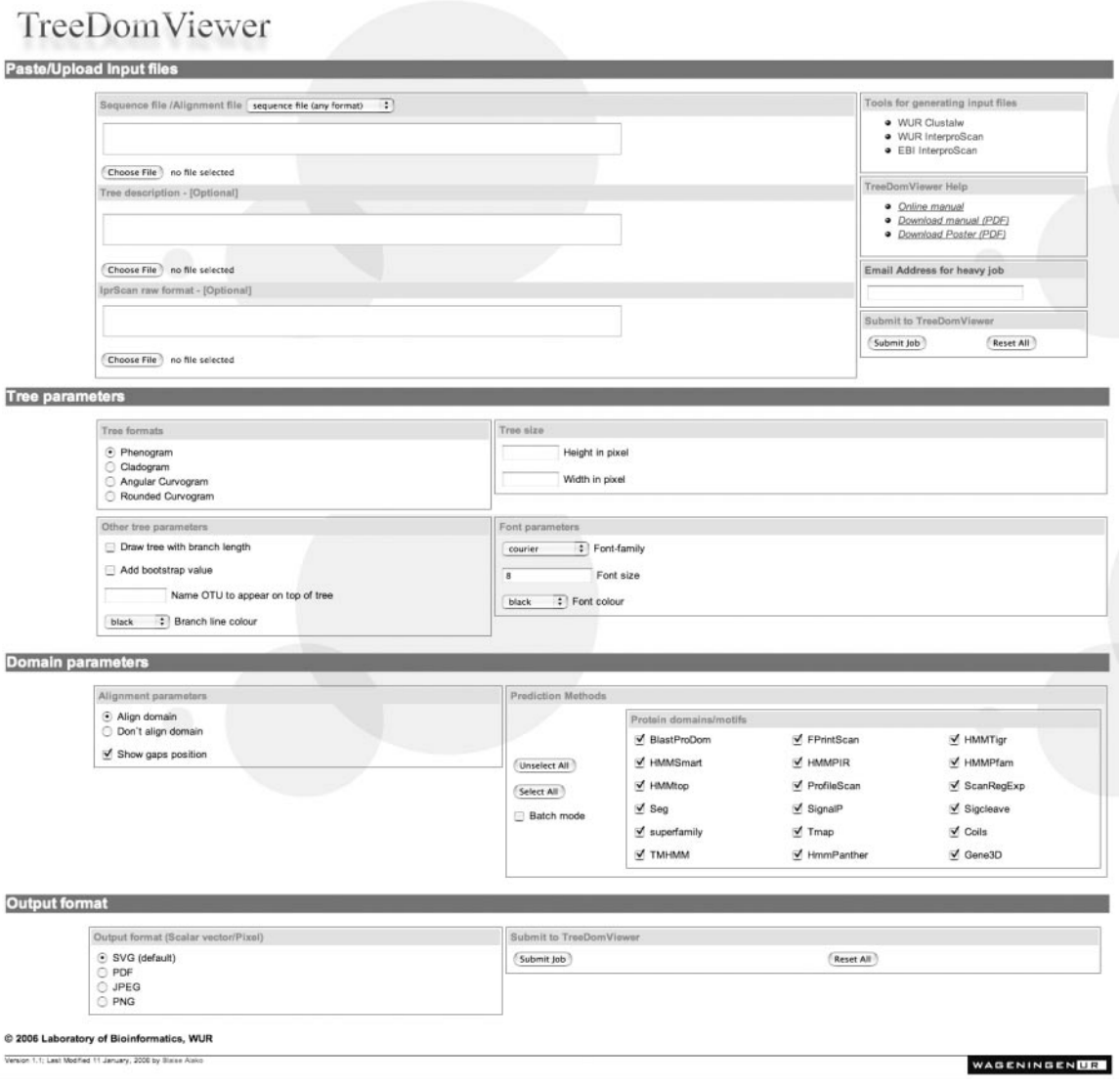

Figure 1. TreeDomViewer web-based interface. Alternative means of generating the input file are provided on the top-right panel.

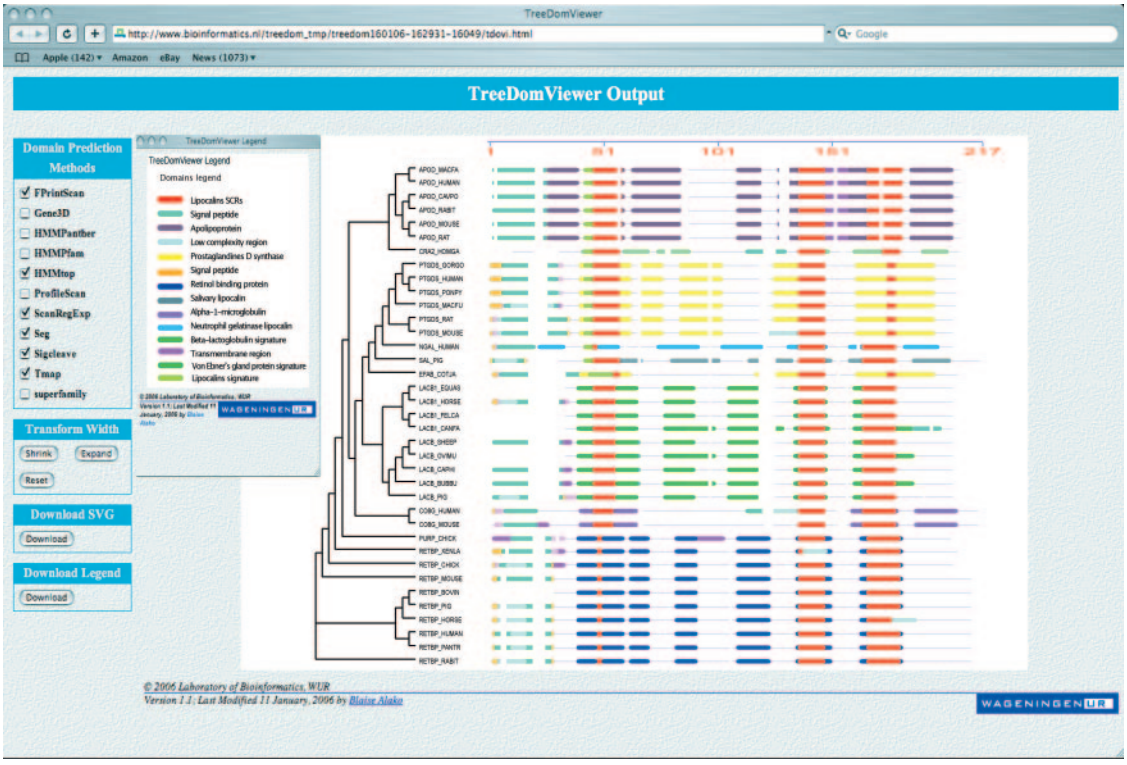

Figure 2. This figure illustrates the default SVG output of 37 lipocalin family members from different species. Shown in red are the main Structurally Conserved Residues (SCRs) that characterize the lipocalins. Inset shows TreeDomViewer domain legend (which appears as a separate pop-up).

from another program for alignment or phylogeny construction than the ones provided by TreeDomViewer.

There are two possibilities to run TreeDomViewer, either interactive, where the user uploads the sequence and/or the alignment, tree description file and the InterProScan analysis file, or in batch mode: the user uploads either the sequence or multiple alignment file but not the InterProScan file. He/she will receive links to the result via email upon job completion and get the option of saving input files as this will save time for future runs of the same dataset. The tool is sufficiently sophisticated to decide which prediction method is most time consuming and may automatically switch to batch mode.

The rate-limiting step in TreeDomViewer is the computation of the structural domains using InterProScan. By running these calculations in parallel on 10 nodes of a small Linux cluster, turn-around times are still acceptable. For example, the analysis of 60 protein sequences of 1000 amino residues each is performed in <3 min.

Output description

By default TreeDomViewer provides scalar vector graphics (SVG) output of the tree and domain information. The user's web browser needs to be SVG-enabled in order to view the output. Conveniently, the viewer first checks the web browser to clarify whether it is SVG-enabled or not, and if needed, initiates the installation of the Adobe SVG plug-in.

The user can change parameters for the tree plotting such as tree format, set to phenogram as default, and many more features as shown in Figure 1. Links to individual protein analysis tools are also provided. It is noteworthy that TreeDomViewer does not execute protein analysis on its own, but instead provides an interface to InterProScan and other programs as shown in the prediction method section of its interface. The domains are sorted by size front-to-back, to prevent large domains obscuring any smaller domains in the same region.

There are several interactive features such as zoom-in and zoom-out, mouse-over access to information on each domain, references to techniques used to produce the domain, and on-the-fly switching on and switching off of domain prediction through the left control panel (Figure 2 as well as an accompanying legend of the graphic). Alternative formats such as PDF, JPG and PNG are also provided.

Although TreeDomViewer was designed for protein analysis, nucleotide sequences can be handled as well. Moreover, TreeDomViewer is able to generate the output of any domain prediction tool, making it the visualization tool of choice at any level of functional or phylogenetic study. Tools such as Adobe Illustrator can be used to manipulate domain colors of the TreeDomViewer SVG file.

In order to illustrate our approach we analyzed a subset of the lipocalin family members. Lipocalins are a superfamily of proteins that carry hydrophobic prosthetic groups. Lipocalins share a very low sequence similarity, hence it can be expected to be a cumbersome affair to infer homology with the conventional sequence similarity or identity techniques. To further our illustration a subset of the lipocalins was selected manually in accordance with those reported by Ganfornina *et al.* (15). We chose this family to illustrate the features of TreeDomViewer because of their strong divergent protein sequence, denoting a rapid rate of molecular evolution,

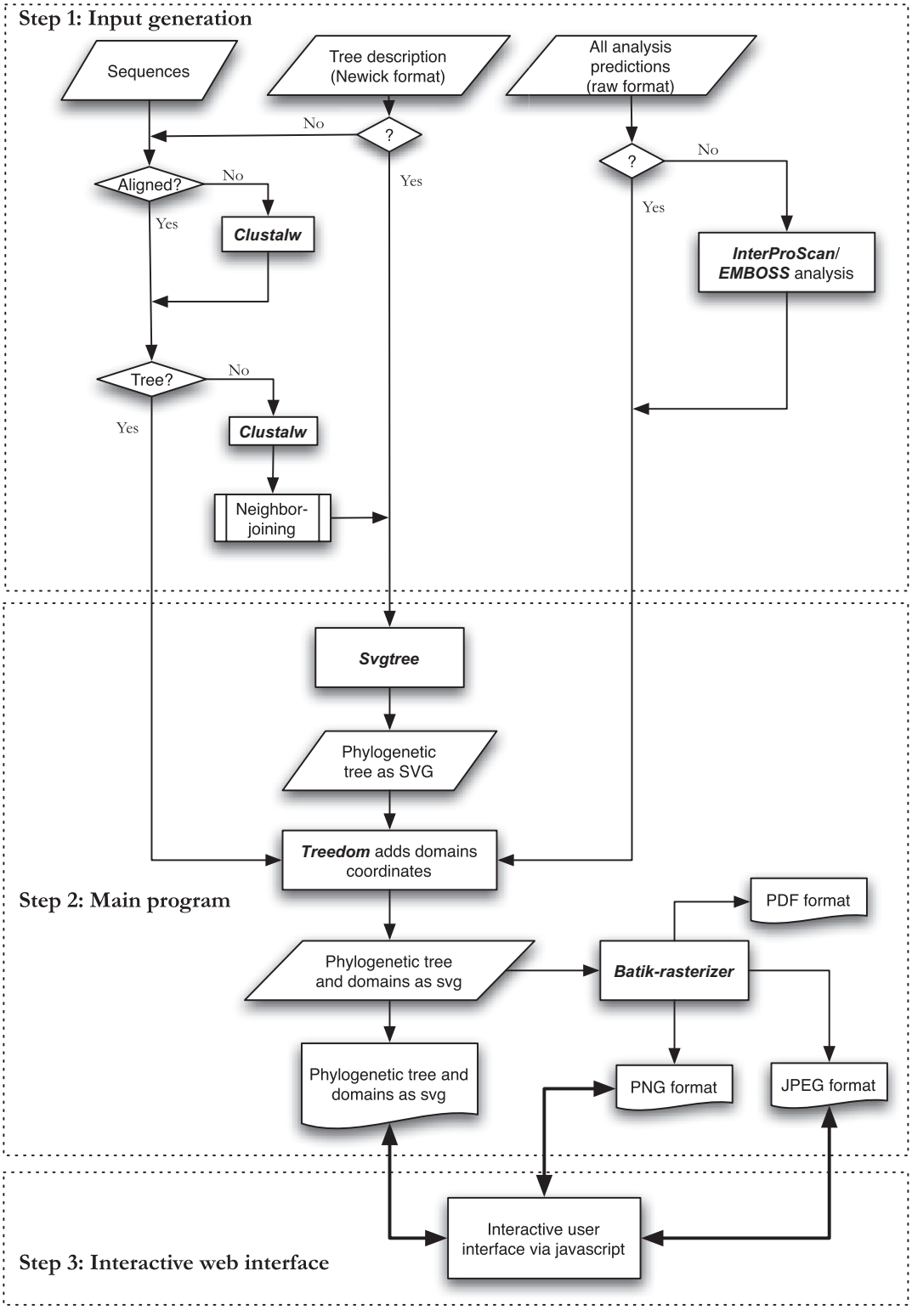

Figure 3. Flowchart of TreeDomViewer illustrating sequence of application implemented. Software tools used are in boldface. Three types of data input are processed and domain information is coordinated with the alignment and phylogenetic tree information to produce an interactive SVG output.

Moreover, the evolutionary history of the lipocalins is rich in gene duplication events, which increases the difficulty of obtaining an understanding of orthologous relationships. There are three conserved sequence motifs called structurally conserved regions (SCRs, denoted in red in Figure 2) that have been proposed by Flower *et al.* (16) as a prerequisite for a protein to be considered a lipocalin.

Although our tool places no restriction on the number of sequences to be used in the analysis, the user's web browser and hardware could be a limiting factor to visualize large

SVG output files. TreeDomViewer was used to visualize a set of 530 Receptor-Like Proteins (RLP) obtained from the arabidopsis genome-wide survey of RLPs without any problem on a standard PC or Mac (data not shown).

Design overview

TreeDomViewer is implemented in Perl as a web-based service running on an Apache 2.0 web server on a Linux platform (SuSE linux Enterprise Server 9). The core application consists of three main programs: *Svgtree*, *Treedom* and *Clustalw*. The first two programs are full command line tools written in-house in C and Perl, respectively, and can be used as plug-in for other applications. A web interface was built on top of these programs via a Perl CGI script (Figure 1). This preserves platform independence across multiple operating systems and allows the user to interact with the different *TreeDomViewer* programs without computer programming or (shell) scripting skills. A global overview of TreeDomViewer workflow is presented in Figure 3. Full explanation of the tool's mode of action is available as an online or downloadable (PDF) manual at the website.

The software was developed on a Linux platform (SuSE 8.2 and SuSE linux Enterprise Server 9) and most of its modules were written from scratch to prevent dependency issues when migrating to newer versions of Linux or Perl.

The TreeDomViewer web interface was tested on Windows XP, Mac OS X and several flavors of Linux OS browsers with good results. Some JavaScript event handling problems for interacting with the SVG output were encountered on Mac OS X and Linux OS. This can be attributed to the web browsers used (konqueror, Mozilla, Opera), as at the moment no browser supports SVG to its full extent. Currently most browsers still require an SVG plug-in, downloadable from the Adobe site. However, the latest version of the Mozilla Firefox browser (version 1.5) has already native (built-in) SVG support and it is to be expected that more browsers will soon follow.

Most browsers handle SVG pictures quite well when standard shapes such as rectangles or lines were instructed to be drawn on the screen. In this matter TreeDomViewer takes it one step further by giving life to these shapes through JavaScript. As all browsers support and display JPEG (Joint Photographic Experts Group) and PNG (Portable Network Graphic), TreeDomViewer uses *batik-rasterizer* to provide alternative output formats besides PDF format, thereby circumventing the need for an SVG plug-in as noted above. *Batik-rasterizer* is part of the open source Apache Batik toolkit 1.6 (http://xml.apache.org/batik/).

Most of the SVG output features such as mouse-over events are retained except zoom-in and zoom-out. As we aimed at integrating as much information as possible within a single picture, domain predictions are linked to their source database where more information can be retrieved.

Future plans

We intend to broaden the scope of TreeDomViewer by incorporating more structural prediction algorithms in the visualization, as well as making it accessible as a BioMOBY web service (17). Furthermore, we plan to improve TreeDomViewer performance by expanding the distributed network of cluster mirrors.

CONCLUSION

TreeDomViewer is a biological web-based tool combining in one picture protein information on phylogenetic and structural information. As such it provides information about the relatedness of proteins and protein families, and thus adds support for inferring function of gene products, in particular when sequence identity is low. One feature of major importance in TreeDomViewer is the alignment of structural domains. This allows for quick checking of the alignment quality, easy inference of homology even when the sequence residue similarity is very low and support for the phylogeny based on functional characteristics evidences.

TreeDomViewer therefore helps in any phylogenetic analysis resolving both the relationship among different group members and the relationship between groups, based solely on the aligned domain structure of each participant.

ACKNOWLEDGEMENTS

The authors wish to thank Pieter Neerincx for testing the tool on Mac OS X and providing valuable tips for preparing the figures. This project was (co)financed by the Centre for BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research. Funding to pay the Open Access publication charges for this article was provided by Wageningen University and Research Centre.

Conflict of interest statement. None declared.

REFERENCES

1. Constantinesco,F., Forterre,P., Koonin,E.V., Aravind,L. and Elie,C. (2004) A bipolar DNA helicase gene, herA, clusters with rad50, mre11 and nurA genes in thermophilic archaea. *Nucleic Acids Res*., **32**, 1439–1447.
2. Bujnicki,J.M. (2004) Bioinformatics-guided identification and experimental characterization of novel RNA methyltransferases. In Gross,H.J. (ed.), *Nucleic Acids and Molecular Biology*, *1st edn*. Springer, Berlin, Heidelberg, Germany, Vol. 15, pp. 146–148.
3. Jones,D.T., Taylort,W.R. and Thornton,J.M. (1992) A new approach to protein fold recognition. *Nature*, **358**, 86–89.
4. Pevsner,J. (2003) *Bioinformatics and Functional Genomics. 1st edn*. Wiley-Liss, Hoboken, New Jersey.
5. Todd,A.E., Orengo,C.A. and Thornton,J.M. (2002) Sequence and structural differences between enzyme and nonenzyme homologs. *Structure*, **10**, 1435–1451.
6. Von Heijne,G. (1986) A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res*., **14**, 4683–4690.
7. Milpetz,F., Argos,P. and Persson,B. (1995) TMAP: a new email and WWW service for membrane-protein structural predictions. *Trends Biochem. Sci*., **20**, 204–205.
8. Tusnady,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.
9. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
10. Eisen,J.A. and Wu,M. (2002) Phylogenetic analysis and gene functional predictions: phylogenomics in action. *Theor. Popul. Biol*., **61**, 481–487.
11. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet*., **16**, 276–277.
12. Thompson,J., Higgins,D. and Gibson,T. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment

through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
13. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
14. Felsenstein,J. (2002) PHYLIP (Phylogeny Inference Package) Version 3.6a3 Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
15. Ganfornina,M.D., Gutierrez,G., Bastiani,M. and Sanchez,D. (2000) A phylogenetic analysis of the lipocalin protein family. *Mol. Biol. Evol.*, **17**, 114–126.
16. Flower,D.R., North,A.C.T. and Attwood,T.K. (1993) Structure and sequence relationships in the lipocalins and related proteins. *Protein Sci.*, **2**, 753–761.
17. Wilkinson,M.D. and Links,M. (2002) BioMOBY: an open source biological web services proposal. *Brief. Bioinform.*, **3**, 331–341.