BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 15 2010, pages 1905–1906
doi:10.1093/bioinformatics/btq306

Sequence analysis

Advance Access publication June 6, 2010

COMA server for protein distant homology search

Mindaugas Margelevičius, Mindaugas Laganeckas and Česlovas Venclovas*

Institute of Biotechnology, Graičiūno 8, LT-02241 Vilnius, Lithuania

Associate Editor: Burkhard Rost

ABSTRACT

Summary: Detection of distant homology is a widely used computational approach for studying protein evolution, structure and function. Here, we report a homology search web server based on sequence profile–profile comparison. The user may perform searches in one of several regularly updated profile databases using either a single sequence or a multiple sequence alignment as an input. The same profile databases can also be downloaded for local use. The capabilities of the server are illustrated with the identification of new members of the highly diverse PD-(D/E)XK nuclease superfamily.

Availability: http://www.ibt.lt/bioinformatics/coma/

Contact: venclovas@ibt.lt

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

Received on April 19, 2010; revised and accepted on June 2, 2010

1 INTRODUCTION

The concept of homology (common evolutionary origin) is at the heart of most studies dealing with protein sequence, structure and function. In the absence of protein structure, inference of homology usually has to rely exclusively on sequence data. At present, most sensitive sequence-based methods use comparison of multiple sequence alignments (MSA) represented as either Hidden Markov Models (Madera, 2008; Söding, 2005) or sequence profiles (Sadreyev and Grishin, 2003; Wang *et al*., 2009). New sequence-based methods may help in revealing previously unappreciated evolutionary relationships even for proteins with known structures (Alva *et al*., 2010).

With the goal of improving distant homology detection, we have recently developed a new method (COMA; *C*omparison *O*f *M*ultiple *A*lignments), based on sequence profile–profile comparison (Margelevičius and Venclovas, 2010). The new method has at least two major features distinguishing it from other profile–profile comparison methods. The first feature is position-specific variable gap penalties that are more biologically relevant than the fixed ones. The second one is a global score system leading to improved estimation of statistical significance of detected similarities. The standalone software implementing the COMA method provides all the tools necessary for profile search and construction of custom profile databases, making it suitable for large-scale studies. However, the use of standalone tools requires a certain level of expertise and substantial computer resources. Thus, to make this method accessible for larger biological community, we have developed a web server. The server is based on the latest version of the COMA algorithm, which is almost 2-fold faster than the original one without sacrifice in sensitivity. This has been achieved by the introduction of simple heuristics to filter out clearly random matches by requiring several high scoring segments on the same diagonal of the dynamic programming matrix. The server includes a number of add-ons for easier analysis and interpretation of homology search results such as merging alignments of multiple hits and structure modeling. The portability of the standalone software has also been improved by changing representation of profile databases from binary to the text format. Thus, the regularly updated server profile databases can be downloaded and used locally on different platforms.

Here, we describe major features of the COMA server and provide an example of its use for uncovering distant homology.

2 FEATURES AND USAGE

The main server window features a simple and intuitive user interface for setting-up homology search jobs. The input for the server is either a single protein sequence or a user-supplied MSA. If the input is a single protein sequence, the server will use it to initiate a PSI-BLAST (Altschul *et al*., 1997) search against one of the specified sequence databases to collect homologs and to generate a query-based MSA. If the input is MSA, the user can choose whether to use the alignment to jump start a PSI-BLAST search or to use it as is for the subsequent construction of the query sequence profile. The PSI-BLAST run can be controlled by several parameters. Alternatively, the user may entrust the server to handle a PSI-BLAST run. In such case, the server will attempt to collect as many related sequences as possible, while keeping PSI-BLAST results free from contamination by unrelated sequences.

The actual homology search is done by comparing query-based sequence profile constructed from the corresponding multiple sequence alignment against the selected profile database (at present SCOP, PDB and PFAM). The user can choose the *E*-value cutoff and the maximum number of resulting alignments to be displayed in the output. In addition, there are a number of advanced options for controlling the construction and comparison of profiles. These options may be useful for expert users.

Instead of waiting for homology search results in the interactive mode the user may bookmark the link to the web-page, where the results will be displayed, or may choose to receive the html link to the results by e-mail upon completion of the specified job. Every completed job is stored for a certain period of time (presently one month), during which the user has unlimited access to the results.

The structured output of the homology search (see Supplementary Fig. 1) is provided on the web page and includes (i) summary of the input data and job parameters, (ii) an overview of the profile hits in color-coded graphical representation and (iii) query-hit alignments. The output gives a concise information regarding structure, function

*To whom correspondence should be addressed.

and other relevant information for reported hits. Hits are linked to original databases (PDB, SCOP or PFAM) and to the search results within the collection of PubMed articles using annotation of hits as keywords.

The output is enriched with additional functionality. The user may extract the combined alignment between the query and any number of hits for further analysis. If the search was performed against PDB or SCOP, the server provides a possibility to generate a 3D model of the query based on the corresponding alignment.

3 EXAMPLE OF NOVEL HOMOLOGY DETECTION

To illustrate the utility of the COMA server for distant homology detection, we describe the discovery of new evolutionary links to the highly diverse PD-(D/E)XK nuclease superfamily.

Using COMA to query individual PFAM families against a custom-made profile database of PD-(D/E)XK representatives, we noticed that the Rai1 family (PF08652) has produced intriguing results. The absolutely conserved D and EXK motifs within the Rai1 family were aligned to the active site motifs of PD-(D/E)XK representatives hinting at possible evolutionary relationship. Using recently solved structures for Rai1 family members (*Schizosaccharomyces pombe* Rai1 and mouse Dom3Z) (Xiang *et al*., 2009) we could directly assess this hint. DaliLite (Holm *et al*., 2008) searches with Rai1/Dom3Z structures against PDB indeed found PD-(D/E)XK proteins as the best matches with significant *Z*-scores (e.g. λ exonuclease was detected with *Z*-score = 4.7). The analysis also revealed that Rai1/Dom3Z have the structurally conserved core characteristic of PD-(D/E)XK nucleases including all three active site motifs (Fig. 1). To explore this relationship further, we constructed MSA for the conserved Rai1/Dom3Z sequence region (e.g. 181–345 in mouse Dom3Z; Supplementary Material) and used it as is for the COMA server search against structure databases (SCOP and PDB; Supplementary Figs 1 and 2). Server not only detected λ exonuclease as a top hit with the significant *E*-value ($<10^{-4}$), but also aligned all three active site motifs in agreement with DaliLite. Taken together, these results make a convincing case for homology between the Rai1 family and PD-(D/E)XK nucleases. The novelty of this finding is underscored by the fact that the determined Rai1 family structures were considered to represent a new fold (Xiang *et al*., 2009). The newly established homology also has practical implications. Xiang *et al*. showed for the first time that Rai1 has the RNA 5′ pyrophosphohydrolase activity and identified three acidic residues as part of the Rai1 active site. However, since Xiang *et al*. have not linked Rai1/Dom3Z with the PD-(D/E)XK superfamily, the conserved Lys (K241 in *S.pombe* Rai1), part of the EXK motif (Fig. 1), has been overlooked as the putative active site residue.

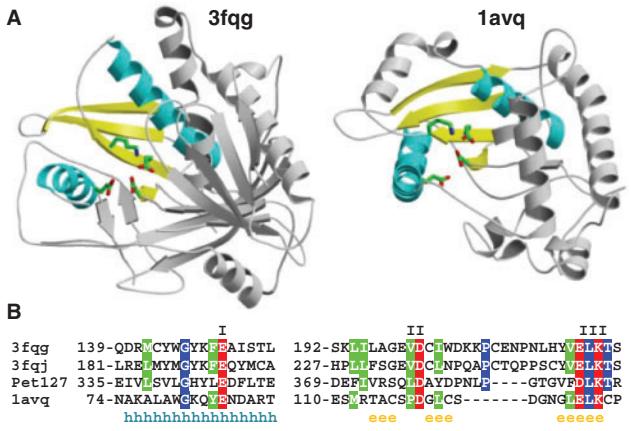

Fig. 1. Structure (**A**) and sequence (**B**) comparison of newly identified members of the PD-(D/E)XK superfamily with λ exonuclease (PDB id: 1avq). (A) Structures of Rai1 (3fqg) and λ exonuclease are oriented similarly in respect to the conserved core (colored). Side chains of corresponding active site residue are also shown. (B) Three active site regions (I, II and III) of λ exonuclease are aligned with the corresponding sequence fragments of Rai1, Dom3Z (3fqj) and Pet127. Red background indicates corresponding active site positions, blue and green denotes identical and similar residues, respectively. Secondary structure of λ exonuclease is shown below the alignment (h, helix; e, strand).

We next asked whether the inferred homology could be extended to proteins of unknown structure. To this end, we used the COMA server to search PFAM with the same Rai1/Dom3Z MSA. We readily identified three other PFAM families related (*E*-values $<10^{-6}$) to Rai1 (Supplementary Fig. 3). These include PF10505 (NMDA receptor-regulated gene protein 2; NARG2), PF05091 (Eukaryotic translation initiation factor 3 subunit 7; eIF3d) and PF08634 (Mitochondrial protein Pet127). While the PD-(D/E)XK active site seems to be absent in NARG2 and not conserved in eIF3d, Pet127 has all three conserved elements (Fig. 1). Consistent with the inferred active site, Pet127 has been found to possess the 5′ RNA exonuclease activity (Fekete *et al*., 2008). The involvement of both Rai1/Dom3Z and Pet127 in 5′ RNA processing raises a possibility that binding to the 5′-end of RNA might also be among potential molecular functions of NARG2 and eIF3d.

The above example illustrates how the COMA server may help in inferring structure and function of specific protein families.

Funding: Lithuanian State Science and Studies Foundation; Howard Hughes Medical Institute.

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.

Alva,V. *et al.* (2010) A galaxy of folds. *Protein Sci.*, **19**, 124–130.

Fekete,Z. *et al.* (2008) Pet127 governs a 5′ -> 3′-exonuclease important in maturation of apocytochrome b mRNA in Saccharomyces cerevisiae. *J. Biol. Chem.*, **283**, 3767–3772.

Holm, L. *et al.* (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics*, **24**, 2780–2781.

Madera,M. (2008) Profile Comparer: a program for scoring and aligning profile hidden Markov models. *Bioinformatics*, **24**, 2630–2631.

Margelevičius,M. and Venclovas,Č. (2010) Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison. *BMC Bioinformatics*, **11**, 89.

Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.

Söding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Wang,Y. *et al.* (2009) PROCAIN: protein profile comparison with assisting information. *Nucleic Acids Res.*, **37**, 3522–3530.

Xiang,S. *et al.* (2009) Structure and function of the 5′–>3′ exoribonuclease Rat1 and its activating partner Rai1. *Nature*, **458**, 784–788.