BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 16 2009, pages 2076–2077
doi:10.1093/bioinformatics/btp346

Sequence analysis

PROCAIN server for remote protein sequence similarity search

Yong Wang[1], Ruslan I. Sadreyev[2] and Nick V. Grishin[2,3,*]

[1]Biomedical Engineering Program, University of Texas Southwestern Medical Center, [2]Howard Hughes Medical Institute and [3]Department of Biochemistry, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390-9050, USA

Received on April 22, 2009; revised on May 28, 2009; accepted on May 29, 2009

Advance Access publication June 3, 2009

Associate Editor: Burkhard Rost

ABSTRACT

Sensitive and accurate detection of distant protein homology is essential for the studies of protein structure, function and evolution. We recently developed PROCAIN, a method that is based on sequence profile comparison and involves the analysis of four signals—similarities of residue content at the profile positions combined with three types of assisting information: sequence motifs, residue conservation and predicted secondary structure. Here we present the PROCAIN web server that allows the user to submit a query sequence or multiple sequence alignment and perform the search in a profile database of choice. The output is structured similar to that of BLAST, with the list of detected homologs sorted by *E*-value and followed by profile–profile alignments. The front page allows the user to adjust multiple options of input processing and output formatting, as well as search settings, including the relative weights assigned to the three types of assisting information.

Availability: http://prodata.swmed.edu/procain/

Contact: grishin@chop.swmed.edu

1 INTRODUCTION

Protein similarity detection and sequence alignment is a significant branch of bioinformatics. It is widely used for prediction of protein structure and function and in protein evolution studies (Kinch *et al*., 2003). To increase the accuracy of these applications, homology detection sensitivity and alignment quality is crucial. However, despite significant research efforts, it is still difficult to accurately detect homologs with relatively low sequence similarity.

BLAST (Altschul *et al*., 1990) and FASTA (Pearson and Lipman, 1988) are the first generation of sequence similarity search programs. Based on sequence-sequence comparison, these programs perform very well for proteins with high sequence identity. PSI–BLAST (Altschul *et al*., 1997), a method based on the comparison of sequence to multiple sequence alignment (MSA), brings the sensitivity of similarity search to another level, since MSA incorporates information about the query protein family. COMPASS (Sadreyev and Grishin, 2003) is based on protein MSA–MSA comparison and improves similarity search further, especially for remote homologs. The latest version of COMPASS employs an advanced statistical model (Sadreyev and Grishin, 2008) to help increase the accuracy of remote similarity detection. HHsearch (Soding, 2005), another MSA–MSA comparison method, uses the formalism of hidden Markov models (HMMs), which allows for position-specific rather than fixed affine gap penalties. HHsearch also incorporates predicted or observed secondary structure (SS) in the process of alignment construction and statistical significance estimation. These two characteristics of HHsearch contribute to more accurate remote similarity detection.

Recently we developed PROCAIN, a method for protein MSA comparison with assisting information (Wang *et al*., 2009). PROCAIN combines residue substitution constraints at individual sequence positions with sequence motif matching, residue conservation and SS scoring. PROCAIN also incorporates an empirical method for the estimation of statistical significance that is based on the comparison of non-homologous proteins from a calibration database (for query sequence) and a database used for search (for subject sequence). This method produces more realistic *E*-values and improves the ranking of hits. Benchmarked together with COMPASS 3.0 and HHsearch (version 1.5), PROCAIN shows better remote homology inference and alignment quality (Figure 1).

2 FEATURES AND USAGE

The main page of the PROCAIN web server consists of an input box and several option sets. The user can paste a protein sequence or alignment into the input box or upload them using the browse button. The user can choose a protein database to search: the server features SCOP, PDB and PFAM databases. The user can access the results interactively in the current window or choose to receive an html link to results by email after the search is completed.

There are three option sets: input processing options, search options and output formatting options. Following the link provided by the name of each option will lead the user to a help page with a brief explanation of the option. Input option set includes options for running PSI–BLAST and further processing of the resulting alignment of detected homologs, such as the number of iterations, cutoff *E*-values, etc. Output formatting options include the upper-bound *E*-value to truncate the list of hits at, significance threshold and the maximum number of alignments the user wants to be shown. The output example button in the upper right corner of the input box will show the user a typical result page.

The search options are probably the most important for an experienced user. These include the values of affine gap penalties (costs of gap opening and gap extension) that allow for the adjustment of the coverage of produced alignments. Increasing gap penalties will decrease coverage and vice versa. PROCAIN

*To whom correspondence should be addressed.

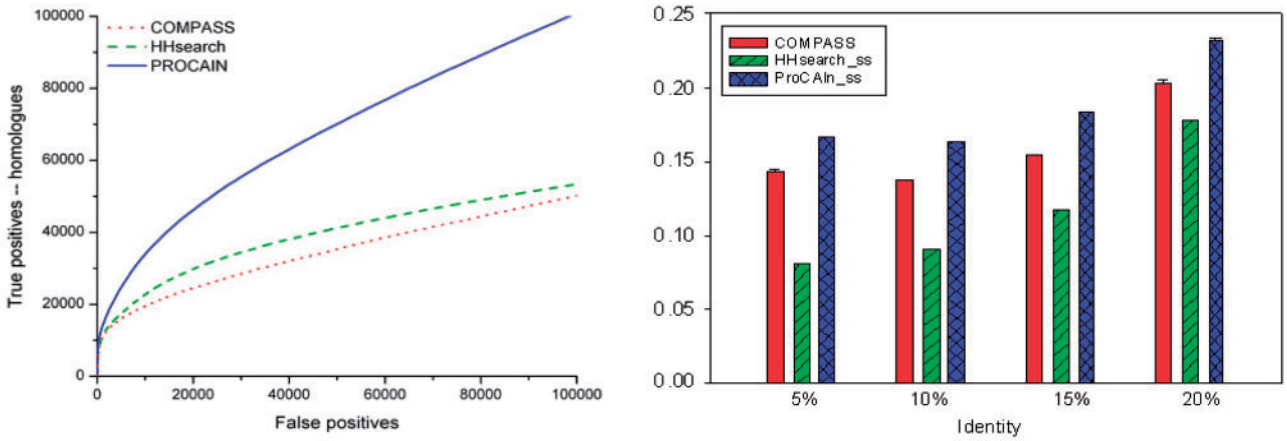

Fig. 1. PROCAIN's overall performance with respect to homology detection accuracy (left) and alignment quality (right). Homology detection quality is measured by overall structure similarity of detected protein pairs. If the detected protein pair has a GDT-TS score (normalized by query length) larger than 15 (on the scale from 0 to 100), then the hit is considered a true positive; false positive otherwise. The GDT-TS cutoff was determined previously (Qi *et al*., 2007) based on the observed GDT-TS distributions for homologs and non-homologs. Alignment quality is measured by the average GDT–TS–like score (vertical axis on the scale from 0 to 1) based on the alignments of homologous sequences.

generally produces long alignments with coverage of 40% larger than COMPASS and almost 200% larger than HHsearch. Such longer alignments are normally favored by the users attempting to predict the structure of the query protein, since they provide a more complete picture of the possible structure of the query protein.

PROCAIN constructs alignments based on the combination of four scores: sequence similarity scores, amino-acid conservation scores, sequence motif scores and SS scores:

$$s = s_{seq}(1 + w_c C) + \delta_m w_m s_m + w_{ss} s_{ss}$$

where s_{seq} is the sequence similarity score; C is the total conservation score in the two columns, normalized to the range [0–1]; s_m is the sum of the sequence similarity scores of the current position and its previous and next neighboring positions; s_{ss} is the SS similarity score. w is the weight parameter for each score. $\delta_m = 1$ if the sequence similarity scores of the current position and its two neighboring positions are all positive, $\delta_m = 0$ otherwise. The resulting all-positions to all-positions scores s are used to construct the optimal local alignment of the two profiles using Smith-Waterman algorithm (Smith and Waterman, 1981). The search options provide an opportunity to adjust this scoring function by changing the weight of the different components according to the user's experience and expectations. The default weights of score terms are the optimal values obtained on a subset of diverse SCOP domains (Wang *et al*., 2009). Although the composition of the training set (47.9% for α/β, 17.6% for all α, 9.6% for all β and 8.9% for $\alpha + \beta$) may reflect the overall composition of the SCOP database very well, it is dominated by α/β class. This composition bias leads to different homology detection accuracy in different protein classes. PROCAIN performance in the α/β class is very similar to the overall performance, whereas the other three classes show significant differences. These differences suggest that homology detection in all α, all β and $\alpha + \beta$ classes may benefit from individualized adjustment of weights in the scoring function. For example, decreasing the contribution of SS score and putting more emphasis on residue similarity can potentially improve the detection quality among all α and all β proteins, where types and boundaries of SS elements are less informative for alignment construction. Thus experienced users are encouraged to adjust the weights of the three additional scores according to the properties of the query protein.

ACKNOWLEDGEMENTS

The authors would like to thank Ming Tang for providing technical support with setting up the server.

Funding: National Institutes of Health (grant number GM67165 to N.V.G.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

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.

Kinch,L.N. *et al*. (2003) CASP5 assessment of fold recognition target predictions. *Proteins*, **53** (Suppl. 6), 395–409.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Qi,Y. *et al*. (2007) A comprehensive system for evaluation of remote sequence similarity detection, *BMC Bioinformatics*, **8**, 314.

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.

Sadreyev,R.I. and Grishin,N.V. (2008) Accurate statistical model of comparison between multiple sequence alignments, *Nucleic Acids Res.*, **36**, 2240–2248.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences, *J. Mol. Biol.*, **147**, 195–197.

Soding,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.