PANAL: an integrated resource for Protein sequence ANALysis

Kevin A. T. Silverstein*, Alan Kilian, John L. Freeman, James E. Johnson, Ihab A. Awad and Ernest F. Retzel*

Computational Biology Centers, Academic Health Center, University of Minnesota, Mayo Mail Code 43, 420 Delaware St SE Minneapolis, MN 55455-0312, USA

Received on May 31, 2000; revised on August 21, 2000; accepted on August 23, 2000

Abstract

Summary: We present PANAL, an integrated resource for protein sequence analysis. The tool allows the user to simultaneously search a protein sequence for motifs from several databases, and to view the result as an intuitive graphical summary.

Availability: PANAL can be freely accessed on our World Wide Web server: http://mgd.ahc.umn.edu/panal/. Source code is available upon request to non-profit organizations.

Contact: panal@ahc.umn.edu

Sequence-similarity programs such as BLAST (Altschul et al., 1990) and FASTA (Pearson and Lipman, 1988) provide useful insights on the general function of new protein sequences. However, several methods exist with higher sensitivity and reliability (Bork and Gibson, 1996; Eddy, 1996; Hofmann, 1998). Known collectively as family-based methods, these programs extract features from a multiple sequence alignment of related sequences, and use a model of the conserved features to find new members (Bork and Gibson, 1996; Hofmann, 1998). Among the models, in increasing complexity, are: (i) regular-expression patterns, used by PROSITE (Hofmann et al., 1999); (ii) groups of potentially far-separated ungapped position specific scoring matrices (PSSMs), employed by PRINTS (Attwood et al., 2000) and Blocks (Henikoff et al., 2000); (iii) profiles, which are essentially PSSMs with gaps, also used by PROSITE; and (iv) Hidden Markov Models (HMMs), the statistical model used by Pfam (Bateman et al., 2000).

Each of these family-based methods has its own merits. Pfam is well-designed to detect distantly-related family members. PRINTS is capable of identifying subtle subfamily relationships. Each method covers a slightly different, but overlapping, set of families. Considering this complementarity, it is best to scan sequences of interest using each method in turn. This approach has its own complications: (1) A user must manually input sequences at each web site; (2) the results from the sites are in a variety of incompatible formats; and (3) consequently, no integrated summary of the results from all methods can be provided.

We have created PANAL to address these issues. PANAL is an integrated resource that searches for motifs from PROSITE (both patterns and profiles), PRINTS, Blocks, and Pfam concurrently. The following public programs are used, respectively, for the searches: prosite, pfscan, fingerPRINTScan, blimps/blkprob, and hmmpfam. In addition to the text reports provided by each program, we have developed a consistent graphical summary of the results of all methods.

The main PANAL webpage allows the user to enter a sequence into a form, and to select motif databases to search against. The user may filter out matches based on statistical significance. Once the user submits the sequence, a Tcl script runs each selected program on our servers. The only non-default parameters sent to the programs are the $E$-value cutoff provided by the user and the current primary sequence database size used in the computation of $E$-values. $E$-values reported from all programs are re-normalized, for consistency, to the same database size (see http://mgd.ahc.umn.edu/panal/e_values.html). Status indicators notify the user which programs are still in progress. As each program completes, a java applet dynamically integrates a summary of the results into the display.

An example of the graphical summary for the sequence GTPA_HUMAN is provided in Figure 1. More motifs are identified in this sequence than in a typical query sequence. However, this sequence is useful to illustrate the differences among the programs.

In the display, a horizontal line representing the user’s input sequence stretches across the top, along with vertical lines that indicate amino acid position in the sequence. Glyphs for each matched motif are shown below the user’s sequence. These glyphs are color- and shape-coded...
Fig. 1. A snapshot of PANALs graphical summary report for the sequence GTPA_HUMAN. Each motif-matching program is represented by a different color and shape. Results are sorted by expected value. Additional details are provided in the text.

according to the searching program they represent, and are sorted vertically by expect value.

Databases that match multiple motifs in a family signature, such as PRINTS and Blocks, are displayed as a connected set of shapes. A horizontal black bar indicates the extent of the complete family signature. Matches to individual motifs in the signature have been labeled strong/weak for quick visual inspection. The fingerPRINTScan program (Scordis et al., 1999) defines strong matches as those with a score greater than 30. For other programs that generate multiple motifs, we define strong matches as those whose $E$-value exceeds the user-specified $E$-value cutoff. Strong matches to one of these motifs are indicated by a rectangle in front of the connecting bar; weak matches appear as rectangles behind the bar; missed matches appear as ovals behind the bar. The shape boundaries indicate the start and end amino acid positions reported by each matching program. All glyphs are annotated by a family descriptor and an expect value, where applicable.

Mirror software is used to ensure that all databases and programs are kept up-to-date. Details on all programs, databases and procedures are provided at http://mgd.ahc.umn.edu/panal/databases. A subset of the features and databases in PANAL may also be found on the InterPro server (Apweiler et al., 2000).

Acknowledgements

This work was supported by an SCA with the USDA-ARS, grants from NSF (9872565, 9872633), DOE (DE-FC07 97ID13550), USDA (58-3625-8-117), and the University of Minnesota Academic Health Center. Some hardware used in this work was supplied by an AEG from Sun Microsystems Inc.

References


