BIOINFORMATICS APPLICATIONS NOTE
Vol. 20 no. 2 2004, pages 291–292
DOI: 10.1093/bioinformatics/btg414

AnaGram: protein function assignment

A. J. Pérez[1]*, G. Thode[1] and O. Trelles[2]

[1]Genetics Department and [2]Computer Architecture Department, University of Málaga, 29071 Málaga, Spain

Received on April 30, 2003; revised on July 10, 2003; accepted on August 6, 2003

ABSTRACT

Summary: AnaGram is a web service for protein function assignment based on identity detection of small significant fragments (*protomotifs*) that can act as modular pieces in peptide construction. The system is able to assign function by finding correlations between *protomotifs* and functional annotations contained in SWISS-PROT and Medline databases. In addition, function ontologies are used for hierarchical organization of the predicted functions. Extensive tests have been carried out to evaluate the accuracy and performance of the system.

Availability: http://jaguar.genetica.uma.es/anagram.htm

Contact: antoniojperez@uma.es

INTRODUCTION

Automatic knowledge discovering in genome information content is one of the most exciting challenges in the post-genomic era. Searching for homologies and evolutionary relationships between sequences is by far the most frequently used strategy for assigning functions to new sequences. However, when working with query sequences that have no clear homologues in the sequence databases the functional annotation process is especially difficult (Bork *et al*., 1998).

Several methods have been proposed for addressing this question (Rigoutsos and Floratos, 1998; Hoersch *et al*., 2000). Most of them are based on conventional similarity comparisons. In Pérez *et al*. (2002), an alternative data mining strategy was proposed based on by-identity detection of small significant fragments (Thode *et al*., 1996), which resemble strongly conserved signal and that can act as modules in peptide construction. The overall algorithm is divided into two distinct successive steps: first, subtle amino acid patterns (called *protomotifs*) are searched; and second, these *protomotifs* are associated with functional annotations obtained from the original SWISS-PROT entries (Boeckmann *et al*., 2003) that gave rise to them, and thus they can be used for assigning functions to the analysed sequence (Fig. 1a).

In this note, an application is presented that makes the procedure available to all through the web in an user-friendly manner. This strategy provides information for assigning function to a query protein, through information on domains or important punctual sites for the query protein. The system has been extensively tested (Pérez *et al*., 2002) with both known and singleton or without-homologue proteins, producing in all cases a satisfactory result and obtaining at least positive clues about the function of the query sequence.

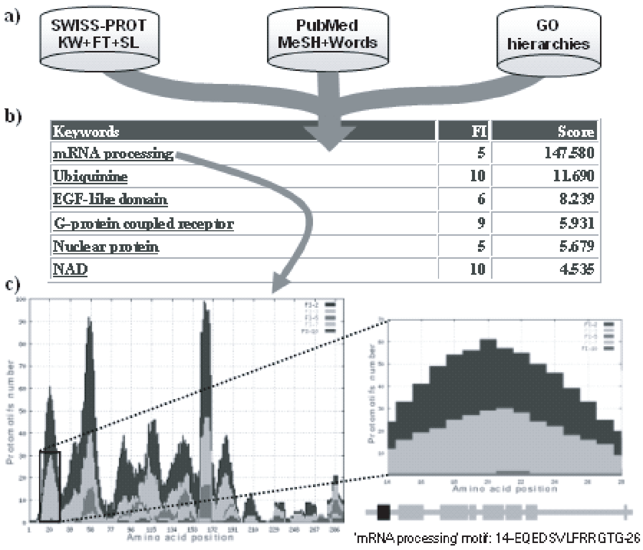

Fig. 1. Data and results from AnaGram server. **(a)** The three databases used for extracting functional annotations for function prediction. Above is the database name and below the used specific fields in this work (KW = keywords, FT = features, SL = subcellular location in the CC field). **(b)** Predicted keywords from the databases sorted by score. Second column (FI) refers to the *protomotifs*' significance. **(c)** Keyword accumulation profile (KAP). This is a histogram of the *protomotif* frequency, i.e. the number of database sequences that contain the *protomotif* in a given position of the query sequence. This is similar to the *protomotif* accumulation profile (PAP) but here only the *protomotifs* linked with one defined keyword are represented. In this histogram, peaks represent conserved zones of the protein, and valleys the less conserved zones; the latter could therefore represent transition zones between different domains. To the right is the zoom of the first peak and beneath it a diagram for locating the motif in the global sequence, and the amino acid sequence both of them from the 'mRNA processing' motif.

*To whom correspondence should be addressed.

MAIN FEATURES

The initial analysis proceeds to obtain *protomotifs* from a query sequence: fixed length fragments contained in both the query and a given database sequence, are registered by query sequence position forming the *protomotif* accumulation profile (PAP) (Fig. 1c). In this histogram (the *protomotif* frequency at a given position of the query sequence) peaks represent conserved zones of the protein associated with functional information in the database sequences (i.e. keyword from the *KW* field in SWISS-PROT) that share the *protomotif*. A significant keyword list is also provided with the sorted predicted functions (Fig. 1b). These functions have a keyword accumulation profile (KAP, similar to PAP) with the associated keyword histogram. By applying significance levels, the system can delineate the putative function associated with a given accumulation, and thus delimit zones or motifs of the query (Fig. 1c).

Additional sources of information can be incorporated into the analysis for fine-tuning the prediction or for narrowing or broadening the focus of it:

Multidomain information: The analysis incorporates domain and post-translational modification information from the F T field in SWISS-PROT, making easier the identification of protein domains with different functions.

Subcellular location: When known, the subcellular location (very important for protein molecular function) can be incorporated into the analysis leading to more specific results.

Medline bibliography: The Medline references from protein sequence entries can be used to incorporate keywords from PubMed abstracts in addition to the SWISS-PROT ones. These keywords can be MeSH (key words in the abstracts from the National Library of Medicine—Schulman, 2001) and/or Words (significant nouns in the abstracts).

Keyword hierarchies: The hierarchy of keywords can be used to broaden or narrow the focus of analysis. The hierarchy of SWISS-PROT keywords can be organized using Gene Ontology (GO) (The Gene Ontology Consortium, 2000). MeSH terms can also be organized by using their own hierarchy. The main point of the hierarchies is to join specific keywords in an upper or general level grouping syntactically different but semantically or functionally related keywords, so corroborating and strengthening the support of predictions. Therefore, the score for a GO term comes from the addition of the keywords grouped in this term.

The core of *AnaGram* is a PERL-CGI library of algorithms and visualization methods embedded in a platform-independent web tool for interactive analysis of sequences. The design of *AnaGram* is suited for remote and multi-user operation. Since it follows the HTML standard, its portability is very high across web browsers. Online help and user-manual facilities are provided for explanation of the options and algorithm parameters. The user can readily import the query sequence or browse a file with the *protomotifs* for saving CPU-time for previously analysed sequences. HTML output is provided as a compressed file becoming available for the user.

CONCLUSIONS

AnaGram web-service offers a good alternative to currently existing software as an aid in function prediction and delimitation of protein domains, especially when a function cannot be assigned by the traditional methods, e.g. in experiments of function definition, site-directed mutagenesis, drugs design, etc. The system has already been tested with a broad sequence set (Pérez *et al*., 2002), and has been shown to be effective and accurate.

ACKNOWLEDGEMENTS

The authors would like to thank Drs Miguel Andrade and Carolina Perez-Iratxeta for their help in launching this server and Jorge Garcia de la Nava for his useful comments. In addition, special thanks are due to Silvia Cano and all in the Department of Genetics at the University of Málaga without whom this work would not have been possible. We acknowledge economic support from the Junta de Andalucía in Spain through its predoctoral program.

REFERENCES

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Bork,P., Dandekar,T., Diaz-Lazcoz,Y., Eisenhaber,F., Huynen,M. and Yuan,Y. (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.*, **283**, 707–725.

Hoersch,S., Leroy,C., Brown,N.P., Andrade,M.A. and Sander,C. (2000) The GeneQuiz Web server: protein functional analysis through the Web. *Trends Biochem. Sci.*, **25**, 33–35.

Pérez,A.J., Rodríguez,A., Trelles,O. and Thode,G. (2002) A computational strategy for protein function assignment which addresses the multidomain problem. *Comp. Funct. Gen.*, **3**, 423–440.

Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics*, **14**, 55–67.

Schulman,J.L. (2001) What's new for 2001 MeSH. *NLM Tech. Bull.*, Nov-Dec, (317): e4.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Thode,G., Garcia-Ranea,J.A. and Jiménez,J. (1996) Search of ancient patterns in protein sequences. *J. Mol. Evol.*, **42**, 224–233.