BIOINFORMATICS APPLICATIONS NOTE Vol. 14 no. 6 1998
Pages 542–543

EUCLID: automatic classification of proteins in functional classes by their database annotations

Javier Tamames[1], Christos Ouzounis[2], Georg Casari[3], Chris Sander[2] and Alfonso Valencia[1]

[1]Protein Design Group, CNB-CSIC, Campus U. Autonoma, Cantoblanco, E-28049 Madrid, Spain, [2]EMBL-EBI, Cambridge CB10 1SD, UK and [3]Lion-AG, Heidelberg, Germany

Received on November 17, 1997; revised and accepted on March 3, 1998

Abstract

Summary: *A tool is described for the automatic classification of sequences in functional classes using their database annotations. The Euclid system is based on a simple learning procedure from examples provided by human experts.*

Availability: *Euclid is freely available for academics at http://www.gredos.cnb.uam.es/EUCLID, with the corresponding dictionaries for the generation of three, eight and 14 functional classes.*

Contact: *E-mail: valencia@cnb.uam.es*

Supplementary information: *The results of the EUCLID classification of different genomes are available at http://www.sander.ebi.ac.uk/genequiz/. A detailed description of the different applications mentioned in the text is available at http://www.gredos.cnb.uam.es/EUCLID/Full_Paper*

With the recent appearance of complete genomes, new challenges are emerging for computational biology. We describe here a tool for the automatic classification of sequences in functional classes using the detailed functional annotations provided by human experts or automatic systems.

In the field of genome analysis, the classification of proteins into a small set of functional classes is common. To mention only two, the human expressed sequence tags (ESTs) (Adams *et al.*, 1993) and the *Saccharomyces cerevisiae* genome (http://speedy.mips.biochem.mpg.de/mips/SC/) have been described in terms of functional classes. The classifications generally used correspond to variations of the cellular function classification originally proposed by Riley (1993).

The problem for deriving this classification automatically is that the annotations of current protein databases are very detailed, mostly biochemical, and far from being a general classification in classes of cellular function. We have approached the problem with the development of the Euclid system (Euclid: from the Greek ευ and κλειζ, meaning good key—paraphrasing the term keyword). Euclid is system based on training; it learns the relationship between functional classes and database annotations (keywords) from small sets of manually classified sequences, and applies these keyword–class relationships to the classification of new sequences. The process is iterated by re-extracting keywords from the sequences already classified, building new dictionaries and re-classifying as many sequences as possible (Figure 1). As for other text analysis systems, it cannot compete with the human capacity for generalization or accuracy, but it is faster, easier to test and able to handle large quantities of data in a reproducible fashion.

Euclid has been applied to the SwissProt database, which contains carefully annotated sequences and a fixed dictionary of 823 keywords (Bairoch and Apweiler, 1997). The algorithm itself is independent of the classes used; currently, different schemes of three, eight and 14 classes are implemented.

With a simple three-class dictionary (Energy, Information and Communication, defined as in Tamames *et al.*, 1996), the system was able to classify 81% of the SwissProt sequences containing some functional information. The accuracy and stability of the classification have been assessed with cross-validation tests and by direct comparison with human expert classification of new genomes. In general, Euclid is less effective than human experts in terms of the number of classified sequences, but sufficiently reliable in the assignment of proteins to classes. In the analysis of the *Mycoplasma genitalium* genome sequence by Fraser *et al.* (1995), the automatic system working with 14 functional classes was able to classify 52% of the sequences, while the original authors were able to classify 63% of the sequences. The lower coverage is compensated for by the reasonable accuracy of the classifications; 82% of the sequences were correctly classified, with better results in well-populated classes. The descriptions of this and other tests of the Euclid system are available as supplementary material.

We have already used the Euclid system in different areas of genome research, including: (i) the comparison of genomes by their composition in protein classes (Tamames *et al.*, 1996); (ii) the comparison of gene order by visual inspection, similar to other systems (Tatusov *et al.*, 1996); (iii) the analysis of the proximity of genes belonging to the same

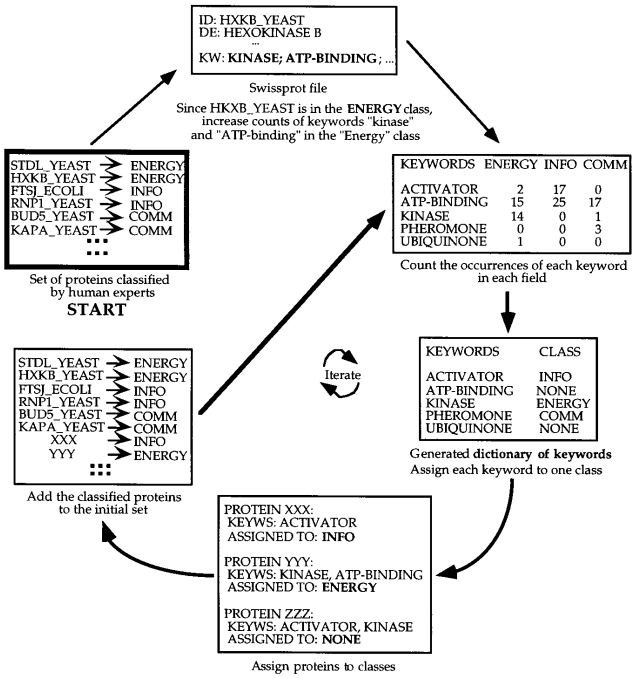

Fig. 1. Scheme of the iterative method used to classify sequences in three functional classes. The example shows how (**a**) hxkb_yeast, hexokinase from yeast, is assigned to the ENERGY class by human experts, (**b**) a dictionary is constructed, keywords associated with hexokinase are registered in the ENERGY class, and (**c**) the dictionary is used for the classification of other sequences. The process is iterated until no more keywords are assigned to classes. An additional set of restrictions is applied during the classification: (i) only keywords appearing in the keyword field are considered for the construction of the dictionaries; (ii) keywords with no functional information are disregarded, e.g. hypothetical protein; (iii) during the iterations, those keywords that are not concentrated in a single functional class are excluded; (iv) the sequences are classified in the class in which the majority of their associated keywords are found; (v) sequences without information in the keyword field are classified if they contain words similar to the keywords of the dictionaries in other fields.

functional class (Tamames *et al.*, 1997). The Euclid system has also been instrumental in the prediction of functional class by the neighboring relationships along the chromosome (A.Valencia *et al.*, unpublished). The Euclid system is also included as one of the modules of the Genequiz package (Scharf *et al.*, 1994; Casari *et al.*, 1996).

The main areas in which Euclid is being improved technically are as follows: (i) increasing the amount of functional input information for each sequence; to circumvent the scarcity of functional annotations in sequence databases, we have developed tools for extracting keywords directly from MEDLINE abstracts (Andrade and Valencia, 1997); (ii) extending the size of the input set of manually classified sequences, including the available new genomes that have been classified into functional classes by their authors; (iii) including the information about homologous sequences after careful assessment of a safe degree of similarity indicative of pertaining to the same functional class; (iv) implementation of a more elaborate weighting scheme to overcome the problem created by difference in size between classes.

Acknowledgements

The work described here was carried out when J.T. and A.V. were at CNB-CSIC Madrid, C.S. at EMBL-EBI, G.C. at EMBL-Heidelberg and C.O. at SRI International, Menlo Park, as part of the collaborative EC-TMR project GENEQUIZ (ERB 4061 PL 95-0315). The Protein Design Group is supported by grant no. BIO9496 CICYT, Spain.

References

Adams,M.D., Kerlavage,A.R., Fields,C. and Venter,J.C. (1993) 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nature Genet.*, **4**, 256–267.

Andrade,M. and Valencia,A. (1997) Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system. *ISMB*, **5**, 25–32.

Bairoch,A. and Apweiler,R. (1997) The SwissProt protein sequence data bank and its new supplement TrEMBL. *Nucleic Acids Res.*, **24**, 21–25.

Casari,G., Ouzounis,C., Valencia,A. and Sander,C. (1996) GeneQuiz II: automatic function assignment for genome sequence analysis. In Hunter,L. and Klein,T.E. (eds), *First Annual Pacific Symposium on Biocomputing*. World Scientific, Hawaii, pp. 707–709.

Fraser,F.C. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–403.

Riley,M. (1993) Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.*, **57**, 862–952.

Scharf,M., Schneider,R., Casari,G., Bork,P., Valencia,A., Ouzounis,C. and Sander,C. (1994) GeneQuiz: a workbench for sequence analysis. In Altman,R., Brutlag,D., Karp,P., Lathrop,R. and Searls,D. (eds), *Intelligent Systems for Molecular Biology 1994*. AAAI Press, Stanford, CA, pp. 348–353.

Tamames,J., Ouzounis,C., Sander,C. and Valencia,A. (1996) Genomes with distinct functional composition. *FEBS Lett.*, **389**, 96–101.

Tamames,J., Ouzounis,C., Casari,G. and Valencia,A. (1997) Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.*, **44**, 66–73.

Tatusov,R.L., Mushegian,A.R., Bork,P., Brown,N.P., Hayes,W.S., Borodovsky,M., Rudd,K.E. and Koonin,E.V. (1996) Metabolism and evolution of *Haemophilus influenzae* deduced from a whole genome comparison to *Escherichia coli*. *Curr. Biol.*, **6**, 279–291.