BIOINFORMATICS APPLICATIONS NOTE

Vol. 18 no. 2 2002
Pages 374–375

Interactive InterPro-based comparisons of proteins in whole genomes

A. Kanapin*, R. Apweiler, M. Biswas, W. Fleischmann, Y. Karavidopoulou, P. Kersey, E. V. Kriventseva, V. Mittard, N. Mulder, T. Oinn, I. Phan, F. Servant and E. Zdobnov

EMBL Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

Received on March 2, 2001; revised and accepted on September 27, 2001

ABSTRACT

Motivation: The SWISS-PROT group at the EBI has developed the Proteome Analysis Database utilizing existing resources and providing comprehensive and integrated comparative analysis of the predicted protein coding sequences of the complete genomes of bacteria, archaea and eukaryotes. The Proteome Analysis Database is accompanied by a program that has been designed to carry out interactive InterPro proteome comparisons for any one proteome against any other one or more of the proteomes in the database.

Availability: http://www.ebi.ac.uk/proteome/comparisons.html

Contact: alex@ebi.ac.uk; proteome@ebi.ac.uk

Genome sequencing is proceeding at an increasingly rapid rate and this has led to an equally rapid increase in predicted protein sequences entering the protein sequence databases. In this paper the term proteome is used to describe the protein equivalent of the genome. Most of these predicted protein sequences are without a documented functional role. One of the major challenges of the genome era is to predict molecular functions and biological roles for the predicted gene products.

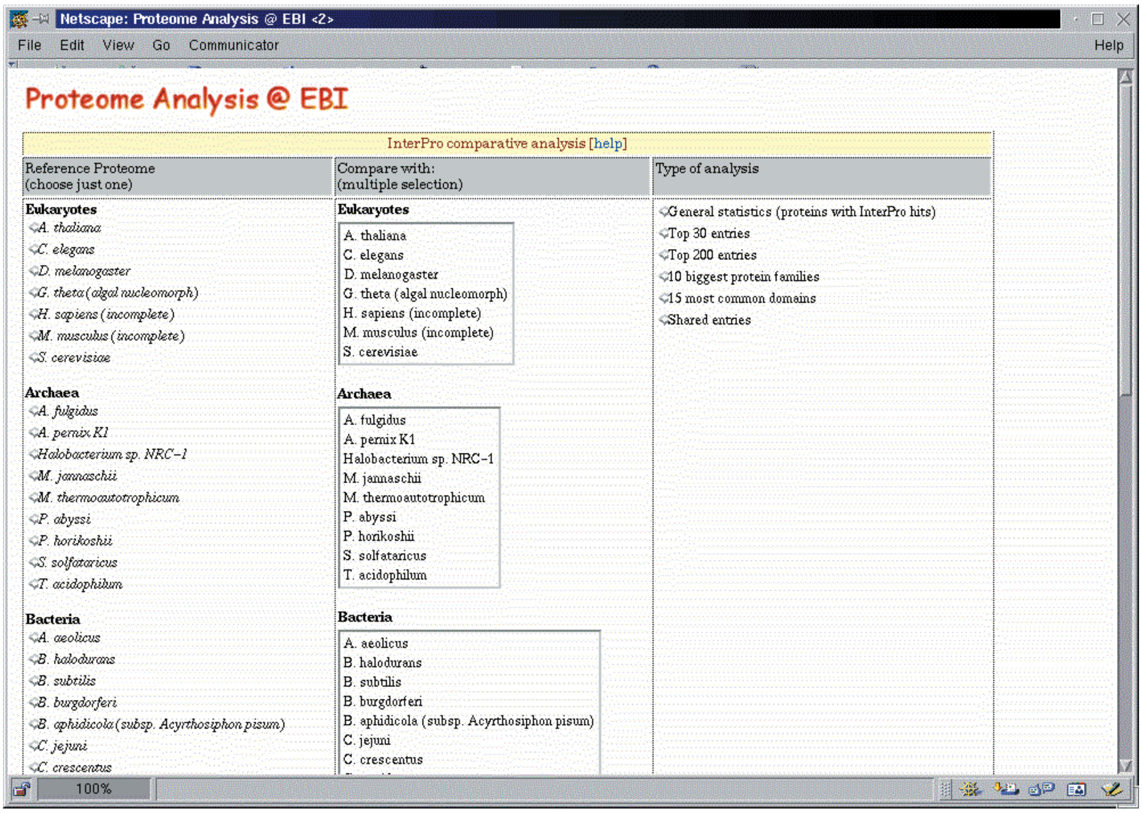
There are a number of existing databases that address some aspects of genome comparisons. The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a knowledge base for systematic analysis of gene functions, linking genomic information with higher order functional information (Kanehisa and Goto, 2000). The WIT Project attempts to produce metabolic reconstructions for sequenced (or partially sequenced) genomes (Overbeek *et al.*, 2000). A metabolic reconstruction is described as a model of the metabolism of the organism derived from sequence, biochemical, and phenotypic data. KEGG and WIT mainly address regulation and metabolic pathways although the KEGG scheme is being extended to include a number of non-metabolism-related functions. A phylogenetic classification of proteins encoded in more than 30 complete genomes can be found in the Clusters of Orthologous Groups of proteins (COGs; Tatusov *et al.*, 2000). COGs group together related proteins with similar but sometimes non-identical functions

InterPro (http://www.ebi.ac.uk/interpro/) is an integrated documentation resource of protein families, domains and functional sites that rationalizes the complementary efforts of the PROSITE, PRINTS, Pfam and ProDom database projects. InterPro is implemented as a relational database in Oracle and users have direct access via Java servlets. The InterPro database is distributed as XML-formatted flat files and as exports of the relational database. The InterPro database provides an integrated layer on top of the most commonly used signature databases to provide a user-friendly interface for text-based searches and sequence scans. In the present paper we describe an important tool that facilitates dynamic InterPro based comparative analysis of each proteome with other available proteomes allowing flexibility of access to the data and providing the means to generate custom-based comparisons.

Comparative analysis data is presented in two different versions in the proteome analysis database (Apweiler *et al.*, 2001a,b) as static and dynamic HTML pages. The static HTML pages contain a few of the most obvious proteome comparisons and these are listed in the section below. This comparison is run through the proteome comparison program and the data is updated weekly together with all other proteome analysis updates. The dynamic HTML pages are generated interactively using the InterPro proteome comparisons program.

Dynamic InterPro-based comparisons can be made using the InterPro proteome comparisons program to select the proteomes of the organisms to be compared and the type of comparative analysis to be carried out (http://www.ebi.ac.uk/proteome/comparisons.html). The program is a

*To whom all correspondence should be addressed.



Proteome Analysis @ EBI

InterPro comparative analysis [help]

Reference Proteome
(choose just one)

Eukaryotes
○ *A. thaliana*
○ *C. elegans*
○ *D. melanogaster*
○ *G. theta (algal nucleomorph)*
○ *H. sapiens (incomplete)*
○ *M. musculus (incomplete)*
○ *S. cerevisiae*

Archaea
○ *A. fulgidus*
○ *A. pernix K1*
○ *Halobacterium sp. NRC-1*
○ *M. jannaschii*
○ *M. thermoautotrophicum*
○ *P. abyssi*
○ *P. horikoshii*
○ *S. solfataricus*
○ *T. acidophilum*

Bacteria
○ *A. aeolicus*
○ *B. halodurans*
○ *B. subtilis*
○ *B. burgdorferi*
○ *B. aphidicola (subsp. Acyrthosiphon pisum)*
○ *C. jejuni*
○ *C. crescentus*

Compare with:
(multiple selection)

Eukaryotes
A. thaliana
C. elegans
D. melanogaster
G. theta (algal nucleomorph)
H. sapiens (incomplete)
M. musculus (incomplete)
S. cerevisiae

Archaea
A. fulgidus
A. pernix K1
Halobacterium sp. NRC-1
M. jannaschii
M. thermoautotrophicum
P. abyssi
P. horikoshii
S. solfataricus
T. acidophilum

Bacteria
A. aeolicus
B. halodurans
B. subtilis
B. burgdorferi
B. aphidicola (subsp. Acyrthosiphon pisum)
C. jejuni
C. crescentus

Type of analysis

○ General statistics (proteins with InterPro hits)
○ Top 30 entries
○ Top 200 entries
○ 10 biggest protein families
○ 15 most common domains
○ Shared entries

Fig. 1. The top section of the form used for setting up interactive InterPro-based comparisons of proteins in whole genomes.

web application with a client/server architecture that runs on a UNIX server and is accessed via a web browser (e.g. Netscape) from any computer that has access to the server via internet. Users can set up the dynamic comparison using a simple input form (Figure 1). The Java servlet creates an SQL script for the selected proteomes and analysis type (as selected by the user) and runs it over the Oracle database. The results are formatted and presented as web pages, similar to the static HTML pages.

Comparisons that can be made include general statistics, top 30 entries, top 200 entries, 15 most common protein families, 15 most common domains and 15 most common protein repeats. An additional feature is the option to compute a list of shared InterPro entries that are common to all the selected proteomes (this is similar in concept to the overlapping region of a Venn diagram).

REFERENCES

Apweiler,R., Biswas,M., Fleischmann,W., Kanapin,A., Karavidopoulou,Y., Kersey,P., Kriventseva,E.V., Mittard,V., Mulder,N., Phan,I. and Zdobnov,E. (2001a) Proteome analysis database: online application of InterProand CluSTr for the functional classification of proteins in wholegenomes. *Nucleic Acids Res.*, **29**, 44–48.

Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D.R., Durbin,R., Falquet,L., Fleischmann,W., Gouzy,J., Hermjakob,H., Hulo,N., Jonassen,I., Kahn,D., Kanapin,A., Karavidopoulou,Y., Lopez,R., Marx,B., Mulder,N.J., Oinn,T.M., Pagni,M., Servant,F., Sigrist,C.J.A. and Zdobnov,E.M. (2001b) The InterPro database, an integrated documentation resourcefor protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 29–34.

Overbeek,R., Larsen,N., Pusch,G.D., D'Souza,M., Selkov,E. Jr, Kyrpides,N., Fonstein,M., Maltsev,N. and Selkov,E. (2000) WIT: integrated system for high-throughput genome sequenceanalysis and metabolic reconstruction. *Nucleic Acids Res.*, **28**, 123–125.

Rubin,G.M., Yandell,M.D., Wortman,J.R., Gabor Miklos,G.L., Nelson,C.R., Hariharan,I.K., Fortini,M.E., Lip,W., Apweiler,R., Fleischmann,W. *et al.* (2000) Comparative genomics of the eukaryotes. *Science*, **287**, 2204–2215.

Snel,B., Bork,P. and Huynen,M. (2000) Genome evolution: gene fusion versus gene fission. *Trends Genet.*, **16**, 9–11.

Tatusov,R.L., Galperin,M.Y, Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of proteinfunctions and evolution. *Nucleic Acids Res.*, **28**, 33–36.