BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 6 2001
Pages 571–572

The HIB database of annotated UniGene clusters

Birgitta Geier[1, 2], Gabi Kastenmüller[1], Matthias Fellenberg[1, 2], Hans-Werner Mewes[1] and Burkhard Morgenstern[1,*]

[1]MIPS Instituti für Bioinformatik GSF—Forschungszentrum für Umwelt und Gesundheit, GmbH Ingolstädter Landstraße, 1 85764 Neuherberg, Germany and [2]Biomax Informatics AG, Lochhamer Straße 11, 82152 Martinsried, Germany

Received on November 20, 2000; revised and accepted on February 2, 2001

ABSTRACT

Summary: The HumanInfoBase (HIB) is a database of putative human gene transcripts. UniGene clusters are assembled, and the resulting consensus sequences are submitted to the PEDANT software system (Frishman,D., Albermann,K., Hani,J., Heumann,K., Metanomski,A., Zollner,A. and Mewes,H.-W., 2001, *Bioinformatics*, **17**, 44–57) for fully automatic sequence analysis and annotation. Predicted transcripts are classified using a variety of functional and structural categories, and hyperlinks to various databases are provided for additional information. A WWW-based graphical user interface represents the assembly process as well as functionally important sites in the putative transcripts.

Availability: http://mips.gsf.de/proj/human/

Contact: morgenst@mips.gsf.de

The large number of Expressed Sequence Tags (ESTs) that are now available in public databases are representing a significant portion of all human genes. To exploit the information contained in these databases it is, however, necessary to reduce their redundancy and to increase the information content and quality of the EST raw data. We created a database called the HumanInfoBase (HIB) of automatically annotated putative human transcripts together with a functional classification based on systematic homology searches and pattern analysis.

Our primary source of data are non-redundant clusters of GenBank sequences from the UniGene database (Schuler, 1997). The current version of HIB comprises 1 415 603 UniGene sequences that are divided into 92 931 clusters. 56 697 of these clusters contain multiple sequences while the remaining 36 234 clusters consist of one single sequence each. Clusters consisting of more than one sequence are processed in several steps. First, the CAP3 assembly program (Huang, 1996) is applied to build one or several consensus sequences for every cluster. Clusters that cannot be assembled with CAP3 are submitted to the PHRAP program (Gordon *et al.*, 1998). A graphical overview of the resulting assembly is provided that shows the position of each UniGene sequence within the respective consensus sequence. Information from the source databases can be retrieved through links to EMBL entries. If there is more than one consensus sequence for a UniGene cluster, this may indicate alternative splicing.

Next, the BLASTX program (Altschul *et al.*, 1990) is applied to search every consensus sequence against a non-redundant protein database that has been assembled from PIR (Barker *et al.*, 2000), SWISSPROT (Bairoch and Apweiler, 2000) and the SWISSPROT supplement TrEMBL. For each cluster, the most significant match to a protein sequence is identified and, provided a threshold E-value of 0.01 is exceeded and no frame shift occurs, the matching region is extended to obtain a maximal Open Reading Frame (ORF) containing the best possible match to a known protein. Consensus sequences with no BLASTX hits below the threshold, or with frame shifts within the matching regions, are submitted to the ESTScan software (Iseli *et al.*, 1999). If these methods fail to identify ORFs, the longest possible ORF in the consensus sequence is extracted, instead. All ORFs that can be found in this way are translated and submitted to the PEDANT sequence analysis system (Frishman *et al.*, 2001) for detailed functional and structural characterization. PEDANT is a software system for automatic and exhaustive analysis of protein sequences incorporating many of the currently available bioinformatics tools.

Predicted proteins are classified according to several distinct criteria. Each putative transcript is assigned to a category in the MIPS *Functional Catalogue* that has been established for analysis of the yeast genome (Mewes *et al.*, 1999, http://www.mips.biochem.mpg.de/proj/yeast/catalogues/funcat/index.html). In addition, transcripts are functionally classified according to the Enzyme Commission (EC) Numbers (Enzyme Nomenclature, 1992), and finally, they are categorized depending on species to which the most significant BLASTX matches were found. All these categories can be accessed through

*To whom correspondence should be addressed.

a WWW interface. This way, the user can easily identify, for example, all putative transcripts belonging to the *Functional Catalogue* category 'Signal Transduction', select transcripts with an EC Number for hydrolases or retrieve all transcripts with strong similarity to a known protein from *C.elegans*. In addition, it is possible to search for keywords, PIR superfamilies, PROSITE patterns (Hofmann *et al.*, 1999), Pfam domains (Bateman *et al.*, 2000), SCOP classifications (Murzin *et al.*, 1995) and structural features. A graphical representation indicates the location of BLAST matches, PROSITE patterns or Pfam domains in the respective sequences. Hyperlinks to the source databases, to OMIM (Online Mendelian Inheritance in Man, 2000) and to GeneCards (Rebhan *et al.*, 1998) provide additional information on the predicted transcripts.

For the future, we are planning to add further functional annotation to the human genes represented in the HIB database. Different data types will be integrated such as gene expression data (DeRisi *et al.*, 1997), information about metabolic pathways (Fellenberg and Mewes, 1999) or protein–protein interactions (Fellenberg *et al.*, 2000) in order to provide more information about the biological context of individual genes. Algorithms and data models that we have recently developed (e.g. Fellenberg *et al.*, 2000; Kastenmüller and Mewes, 1999) will be employed to provide further information about the biological function of previously uncharacterized genes.

ACKNOWLEDGEMENTS

We would like to thank Dmitrij Frishman, Ole Bents, Norman Strack and Grigory Kolesov for their help with PEDANT and other software systems.

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bairoch,A. and Apweiler,R. (2000) The SWISSPROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Barker,W.C., Garavelli,J.S., Huang,H., McGarvey,P.B., Orcutt,B.C., Srinivasarao,G.Y., Xiao,C., Yeh,L.L., Ledley,R.S., Janda,J.F., Pfeiffer,F., Mewes,H.-W., Tsugita,A. and Wu,C. (2000) The Protein Information Resource (PIR). *Nucleic Acids Res.*, **28**, 41–44.

Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and geneticcontrol of gene expression on a genomic scale. *Science*, **278**, 680–686.

Enzyme Nomenclature (1992) *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes, NC-IUBMB*. Academic Press, New York.

Fellenberg,M. and Mewes,H.-W. (1999) Interpreting clusters of gene expression profiles in terms of metabolic pathways. In *Proceedings of German Conference on Bioinformatics 1999*, pp. 185–187.

Fellenberg,M., Albermann,K., Zollner,A., Mewes,H.-W. and Hani,J. (2000) Integrative analysis of protein interaction data. In *Proceedings of 8th International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Menlo Park, CA, pp. 152–161.

Frishman,D., Albermann,K., Hani,J., Heumann,K., Metanomski,A., Zollner,A. and Mewes,H.-W. (2001) Functional and structural genomics using PEDANT. *Bioinformatics*, **17**, 44–57.

Gordon,D., Abajian,C. and Green,P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.*, **8**, 195–202.

Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.

Huang,X. (1996) An improved sequence assembly program. *Genomics*, **33**, 21–31.

Iseli,C., Jongeneel,C.V. and Bucher,P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. In *Proceedings of 9th International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Menlo Park, CA, pp. 138–148.

Kastenmüller,G. and Mewes,H.W. (1999) An object-oriented data model for the dynamic modelling of metabolic pathways. In *Proceedings of German Conference on Bioinformatics 1999*, pp. 206–207.

Mewes,H.-W., Heumann,K., Kaps,A., Mayer,K., Pfeiffer,F., Stocker,S. and Frishman,D. (1999) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **27**, 44–48.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Online Mendelian Inheritance in Man (2000) OMIM (TM) McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), http://www.ncbi.nlm.nih.gov/omim/

Rebhan,M., Chalifa-Caspi,V., Prilusky,J. and Lancet,D. (1998) GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, **14**, 656–664. http://www.bioinfo.weizmann.ac.il/cards/

Schuler,G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.*, **75**, 694–698. http://www.ncbi.nlm.nih.gov/UniGene/