Nucleic Acids Research, 2003, Vol. 31, No. 1 207–211
DOI: 10.1093/nar/gkg005

The PEDANT genome database

Dmitrij Frishman[1,*], Martin Mokrejs[1], Denis Kosykh[1], Gabi Kastenmüller[1], Grigory Kolesov[1], Igor Zubrzycki[1], Christian Gruber[2], Birgitta Geier[2], Andreas Kaps[2], Kaj Albermann[2], Andreas Volz[2], Christian Wagner[2], Matthias Fellenberg[2], Klaus Heumann[2] and Hans-Werner Mewes[1,3]

[1]Institute for Bioinformatics, GSF - National Research Center for Environment and Health, Ingolstädter Landstraße 1, 85764 Neueherberg, Germany, [2]Biomax Informatics AG, Lochhamer Straße 11, 82152 Martinsried, Germany and [3]Department of Genome-oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, 85350 Freising, Germany

Received August 13, 2002; Revised and Accepted September 12, 2002

ABSTRACT

The PEDANT genome database (http://pedant.gsf.de) provides exhaustive automatic analysis of genomic sequences by a large variety of established bioinformatics tools through a comprehensive Web-based user interface. One hundred and seventy seven completely sequenced and unfinished genomes have been processed so far, including large eukaryotic genomes (mouse, human) published recently. In this contribution, we describe the current status of the PEDANT database and novel analytical features added to the PEDANT server in 2002. Those include: (i) integration with the BioRS™ data retrieval system which allows fast text queries, (ii) pre-computed sequence clusters in each complete genome, (iii) a comprehensive set of tools for genome comparison, including genome comparison tables and protein function prediction based on genomic context, and (iv) computation and visualization of protein–protein interaction (PPI) networks based on experimental data. The availability of functional and structural predictions for 650 000 genomic proteins in well organized form makes PEDANT a useful resource for both functional and structural genomics.

OVERVIEW AND STATUS OF THE PEDANT DATABASE IN 2003

When the first version of the PEDANT genome database was launched in 1996 (1) it provided a computational analysis of the five first completely sequenced genomes available at that time using a limited set of algorithms and with results stored as static HTML pages. In the past seven years, the PEDANT genome analysis software has matured (2): it is now based on an efficient relational database schema compatible with both MySQL™ and Oracle™ database management systems, employs a broad range of modern bioinformatics methods to analyze sequence data, and offers an extensive user interface. In parallel, the database content was explosively growing following the fast pace of genome sequencing projects. However, the main concept of the database has not changed since the first day of its existence. Since in-depth manual annotation of all genomic sequences pouring into the databases is virtually impossible our goal has been to provide exhaustive functional and structural characterization of publicly available genomes by automatic means in a timely fashion. Being fully aware of the pitfalls of automatic sequence analysis (3) we use reasonably stringent recognition parameters to avoid excessive false positive rates, and at the same time not only provide search and prediction results in digested form, but also store the raw output of bioinformatics methods, enabling the annotator or the biologist using the database to make his own judgement on the significance of the results presented.

At the time of writing the total of 177 genomes are available on-line. The database consists of three major sections:

1. Genomes which undergo careful in-depth analysis by the MIPS biologists using the subsystem for manual annotation available in the PEDANT software suite. This section currently includes *Neurospora crassa*, *Thermoplasma acidophilum*, and *Arabidopsis thaliana*.
2. Completely sequenced and published genomes. The main source of sequence data for this section, including DNA contigs and ORF nomenclature, is the genomes division of GenBank (4), although in some cases we obtain data directly from sequencing centres. Whenever possible we use data manually curated by NCBI staff (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). If a curated version is not available, original data as submitted by the authors (ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria) is processed. This section contains 5 eukaryotic, 84 eubacterial, and 16 archaebacterial datasets.

*To whom correspondence should be addressed. Tel: +49 89 31874201; Fax: +49 89 31873585; Email: d.frishman@gsf.de

3. Unfinished genomic sequences. Gene prediction is conducted by ORPHEUS (5) in a completely automatic fashion, usually allowing for large overlaps between ORFs. This leads to many over-predicted ORFs, but ensures that fewer real ORFs are missed. In many cases, the PEDANT database is the only source of annotation for such datasets. In recent time, this section of the database was growing slower then before because we chose to commit our processing capacity to the quickly growing number of completely sequenced genomes recently published, including all publicly available eukaryotic genomes. This section contains 15 eukaryotic, 51 eubacterial, and 3 archaebacterial datasets.

Among the most significant recent additions to the database is mouse genome data obtained from http://genome.cse.ucsc.edu. The mouse database contains 20 chromosome contigs with 37 793 genes predicted using the Fgenesh++ software (www.softberry.com).

For each of the roughly 650 000 protein sequences processed so far the following pre-computed analyses are available:

(A) Protein function

- BLAST (6) similarity searches against the complete non-redundant protein sequence database.
- Motif searches against the Pfam (7), BLOCKS (8), and PROSITE (9) databases. InterPro (10) calculations are in preparation.
- Predictions of cellular roles and functions based on high-stringency BLAST searches against protein sequences with manually assigned functional categories according to the FunCat Functional Catalogue developed by MIPS and Biomax Informatics AG. The FunCat catalogue covers a broad range of biological concepts, including cellular processes, systemic physiology, development and anatomy for prokaryotes and unicellular eukaryotes, plants and animals. In addition, genomes annotated with other vocabularies (such as Gene Ontology) can be mapped to FunCat annotations and thus integrated into the similarity search, as already done for the genomes of *Drosophila melanogaster* and *Caenorhabditis elegans*. At present, we use proteins with manually assigned functional categories of the following species: plant *A.thaliana*, fungi *Saccharomyces cerevisiae*, eubacterium *Listeria monocytogenes EGD* and archaebacterium *T.acidophilum*. More species-specific catalogues are in preparation and will be available shortly (e.g. bacteria *Bacillus subtilis*, *Helicobacter pylori*, *N. crassa*).
- Similarity-based predictions of enzyme nomenclature (EC numbers).
- Similarity-based extraction of keywords and superfamily assignments from the PIR-International sequence database (11).
- Assignment of sequence to known clusters of orthologous groups [COGS, (12)].

(B) Protein structure

- Sensitive similarity-based identification of known 3D structures and structural domains. For this purpose, we are using the IMPALA software (13) which allows comparison of each gene product with a collection of position specific scoring matrices, or profile library, representing sequences with known three dimensional structure from the PDB database (14) and sequences of structural domains from the SCOP database (15). CATH (16) domain predictions are being currently added to the database.
- Prediction of transmembrane regions using the TMHMM software (17).
- Identification of local low similarity regions and entire non-globular domains based on the SEG algorithm (18).
- Prediction of coiled coil motifs (19).
- Prediction of protein structural classes (all-α, all-β, α/β).

In some cases, further analyses may be available. For example, for cDNA collections we conduct BLASTN searches against relevant taxonomic subdivisions of the EMBL database (20). Several additional methods to predict protein features, such as localization or presence of signal peptides are implemented, but not systematically used due to high error rates.

Perhaps the most characteristic feature of the PEDANT user interface, available since its conception, is the automatic assignment of gene products to various functional and structural categories. There are two types of such categories:

- Individual categories, such as sequences with homologues. Selecting this category immediately leads to the list of sequences possessing a BLAST hit, sorted by significance. Further categories of this type are: sequences without homology, non-identical closest homologues, sequences with predicted transmembrane segments, coiled coils, low complexity and non-globular regions.
- Group categories, such as sequence and structure motifs. Selecting such category first leads to the list of all groups of a given type actually identified in a particular genome. In a second step, the user selects an item of interest, e.g., a Pfam domain, and gets the list of sequences that are predicted to possess this domain. Categories of this type are: Pfam, BLOCKS, and PROSITE motifs, functional categories, EC numbers, PIR keywords and superfamilies, SCOP and CATH domains, COGs, as well as sequence clusters (see below). In addition, BLAST similarity hits are classified based on their taxonomic origin; additional categories in the taxonomy section—superkingdom, kingdom, phylum, class, and species—allow the user to obtain the lists of respective taxonomic divisions and then select sequences that have at least one BLAST hit in a given division.

In addition, the following searches can be performed interactively against protein sequences as well as DNA sequences or ORFs and contigs of a particular genome:

- BLAST search with a user query sequence
- Sequence pattern search using the PROSITE regular expression language

As soon as an ORF of interest has been selected from a given category or based on an interactive search, an integrated, hyperlinked protein report is provided showing analysis results according to dynamically set thresholds. All evidence available is summarized in the report, including a number of calculated parameters, such as molecular weight, pI value, position of the ORF on the contig, homology-derived data, as well as

predicted structural features. A navigation toolbar in the upper part of the report page allows access to the protein and DNA sequence of a given ORF and the raw results of individual computational methods. Those are also equipped with Web links and can be used as reference for further manual annotation. An advanced DNA viewer represents contigs in graphical form and allows one to navigate, zoom, produce six-frame translation, and show DNA features such as restriction sites and genetic elements (genes, ORFs, exons, tRNAs, etc.). The protein viewer visualizes information about similarity to entries in the protein databases used and predicted protein features, e.g. sequence motifs and secondary structure elements. This is especially useful for judging on the domain structure of the homology hits.

The public PEDANT database server has been upgraded in terms of CPU speed, RAM memory and disk space. In order to improve the performance of the public MySQL database server, a separate server is utilized to conduct computations and prepare the data. When newly created datasets pass extensive quality tests and a substantial number of new databases have been accumulated, a new release of the PEDANT database is made. At the time of writing the version of the database is 1.0.2.

SEARCHING AND DATA MINING IN THE PEDANT GENOME DATABASE USING THE BioRS™ INTEGRATION AND RETRIEVAL SYSTEM

In order to enable users to take full advantage of the exhaustive genome annotation available in the PEDANT database, fast and efficient data mining and search capabilities must be provided. However, given the enormous amount of pre-computed bioinformatics analyses stored in MySQL tables this requirement is not easy to meet. Although MySQL is arguably the fastest relational database currently available a simple text search for the word 'kinase' in only one 500 mB table containing BLAST results for the *A.thaliana* genome takes more than a minute to complete, and composite queries in such large datasets are all but impossible.

To enhance the data-mining capabilities of the PEDANT Genome Database its latest release has been integrated with the BioRS Integration and Retrieval System developed by Biomax Informatics AG (www.biomax.de). The BioRS system is able to integrate and search flat-file databases as well as relational databases (at present, MySQL, Oracle and DB2). Additional index data structures are generated, allowing queries to be processed on the index for enhanced query performance. The original data source is accessed only when the user requests the entire entry or when indexing is performed. Because the open Common Object Request Broker Architecture (CORBA) is used as platform-independent middleware, indexing and querying processes can be distributed over as many CPUs as are available, facilitating timely updates of the indices.

The PEDANT GUI now provides an HTML-based search form which allows one to specify complex search terms (using wildcards) and apply them selectively to different parts of the annotation, e.g. to search only in Pfam motifs, functional categories or known 3D structures. Several instances of such pairs of attributes and search values are provided and can be combined by Boolean operators. Additional criteria for searching include sequence length, number of transmembrane regions, pI range and percentage of low complexity sequence. After clicking the 'Search' button, a CGI program is initiated to translate the values of the HTML search form into the BioRS Query Language. The query is executed by the BioRS core using search daemons and the results are returned to the PEDANT client which then generates an HTML-based table including hyperlinks to the corresponding protein reports. Due to the use of pre-calculated indices search results are returned essentially instantly, allowing interactive exploration of the information contained in the PEDANT database. For example, a search for *A.thaliana* proteins having the word 'transcription' in functional categories, the word 'floral' in BLAST search results, the word 'mads' anywhere in the annotation, and pI in the range from 4 to 8 finds 12 hits in the 11 gB annotation of the genome in just a few seconds.

SEQUENCE CLUSTERING AND PARALOGOUS GENE FAMILIES

One of the important aspects of genome annotation involves evaluation of gene duplication and the analysis of paralogous gene families. Within each completely sequenced genome we conduct an all against all comparison of proteins by PSI-BLAST, with low complexity sequence regions masked. Sequences possessing sufficient degree of similarity in a reciprocal fashion (BLAST similarity score greater than 45 bits) are joined into single-linkage groups. In cases where reciprocal BLAST comparisons produce only one local alignment between two sequences in each direction, this hit is made symmetrical by taking into account only the longer alignment. Additionally, results of sensitive recognition of Pfam domains through HMMER searches (21) are taken into account. If two or more proteins in a genome display similarity to the same Pfam domain with a significant E-value (typically 0.001), it may be safely assumed that the corresponding protein sequence spans are similar to each other, even if BLAST fails to recognize such relationships. Correspondingly, by selecting the 'sequence clusters' category on the PEDANT launch panel the user is presented with a list of sequence clusters found in the given genome, with the number of sequences in each cluster and the cluster name indicated. The latter is automatically derived from the description lines of the cluster sequences, with informative description lines given priority over those containing the words 'unknown', 'putative', and the like. For each cluster the list of sequences can be displayed. In addition, a graphical representation of the cluster is available in form of a circular diagram, visualizing the structure of the BLAST and Pfam hits as well as the structural information available for the cluster proteins (22).

COMPARATIVE GENOMICS

Starting from the year 2002 an exhaustive all-on-all BLAST comparison of all protein sequences in completely sequenced genomes is conducted for each major release of the PEDANT database; the current version encompasses 165 000 proteins in 70 genomes. After selecting the 'intergenome comparison'

category on the launch panel the user may choose up to 10 genomes to be compared and obtain a table of similarity relationships between a query genome and the selected target genomes. Similarity hits are coloured according to their BLAST score and equipped with links to respective genome datasets. In addition, on each report page of proteins involved in the cross-genome comparison a link 'compare genomes starting from this gene' appears, leading to the appropriate page of the genome comparison table. Such table is a very convenient tool for quickly assessing the distribution of a given gene across selected representatives of main taxonomic groups or most important model organisms. Since chromosomal coordinates of genes are also provided it is also possible to estimate the conservation of genomic context around a given gene of interest.

For more in-depth exploration of gene context we have developed a novel computational method called SNAP [Similarity-Neighbourhood APproach; (23)]. A Similarity-Neighbourhood Graph (SN-Graph) is built that involves chains of alternating S- and N-relationships. The former represent BLAST similarity hits between putative orthologues in different genomes while the latter involve neighbouring genes on the same genome. An SN-Graph can thus be thought of as a walk across many genomes which begins with a particular gene in genome A and proceeds to its orthologue in genome B. The walk then continues to encompass a given number of neighbours of this orthologue on each side. Subsequently, orthologues of these neighbours are found in other genomes, their neighbours identified, and so on. Closed paths on an SN-graph, that we call SN-cycles, are strongly non-random and have the tendency to join functionally related genes involved in the same biochemical process. A specialized Web server, Snapper, has been developed which allows one to submit a protein sequence for a SNAP analysis [http://pedant.gsf.de/snapper; (24)]. This server takes full advantage of the PEDANT functional annotation and provides links to PEDANT entries. Conversely, a Snapper session can be launched from any PEDANT database report page by pressing the 'submit this sequence for SNAP analysis' button.

Yet another way to establish functional links between gene products in a similarity-free fashion is through phylogenetic profiling which involves finding genes with correlated occurrence in different genomes (25). We have incorporated a feature-rich implementation of this method (Wong *et al.*, in preparation) into the PEDANT server. In this case, too, the user can invoke a profiling analysis for a gene of interest directly from the PEDANT report page.

PROTEIN–PROTEIN INTERACTIONS

Another novel feature of the PEDANT database introduced in 2002 is the incorporation of the data on protein–protein interactions (PPI). The information is directly imported from the MIPS PPI catalogue [(26); http://mips.gsf.de/proj/yeast/CYGD/interaction] which currently describes the total of 13 842 interactions for 4033 proteins from the *S.cerevisiae* genome. In particular, the catalogue includes the following two components: (i) the original PPI catalogue which was being built by a group of MIPS biologists since 1997 based on careful analysis of yeast literature (27). This 'classical' part of the catalogue contains information on 1889 proteins involved in 4924 interactions, classified into physical and genetic interactions, and (ii) recently published data from large-scale two-hybrid experiments [e.g., (28)]. After clicking on the category 'protein–protein interactions' on the PEDANT launch panel the user is presented with a list of individual experiments (for convenience the 'classic' catalogue is treated as one experiment although data come from hundreds of different publications). For each experiment, a table of interactions between pairs of ORFs is shown, interlinked to the corresponding protein reports. In addition, individual disjoint PPI networks can be delineated and visualized using a graphical Java applet. Direct incorporation of PPI data into PEDANT facilitates its efficient exploration in the context of functional annotation (29). At present, this feature is only available for the *S.cerevisiae* genome; data on other organisms will be added in the future.

STRUCTURAL GENOMICS

The rich set of structural and functional characteristics derived for each protein as well as the high degree of automation and advanced analytical features make the PEDANT database a useful tool for structural genomics. In particular, PEDANT can be used to facilitate the target selection process. Using the sequence clustering results described above it is easy to judge the domain structure of the protein families. Further, circular diagrams visualize available structural information on each cluster member (domains with known three-dimensional structure, transmembrane regions). Based on these pre-computed results we have created an efficient target selection tool called STRUDEL [STRucture DEtermination Logic; (22)]. A Web-based interface for this tool allowing PEDANT users to select structural targets of interest according to specified criteria is currently being developed.

REFERENCES

1. Frishman,D. and Mewes,H.W. (1997) PEDANTic genome analysis. *Trends Genet.*, **13**, 415–416.
2. 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.
3. Galperin,M.Y. and Koonin,E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico. Biol.*, **1**, 55–67.
4. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.
5. Frishman,D., Mironov,A., Mewes,H.W. and Gelfand,M. (1998) Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res.*, **26**, 2941–2947.
6. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
7. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
8. Henikoff,S., Henikoff,J.G. and Pietrokovski,S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**, 471–479.

9. Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
10. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
11. Barker,W.C., Garavelli,J.S., Huang,H., McGarvey,P.B., Orcutt,B.C., Srinivasarao,G.Y., Xiao,C., Yeh,L.S., Ledley,R.S., Janda,J.F. *et al.* (2000) The protein information resource (PIR). *Nucleic Acids Res.*, **28**, 41–44.
12. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
13. Schaffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST- constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
14. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
15. Lo,C.L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
16. Pearl,F.M., Martin,N., Bray,J.E., Buchan,D.W., Harrison,A.P., Lee,D., Reeves,G.A., Shepherd,A.J., Sillitoe,I., Todd,A.E. *et al.* (2001) A rapid classification protocol for the CATH Domain Database to support structural genomics. *Nucleic Acids Res.*, **29**, 223–227.
17. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
18. Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.
19. Lupas,A., Van Dyke,M. and Stock,J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.
20. Stoesser,G., Baker,W., van den,Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V. *et al.* (2002) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **30**, 21–26.
21. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
22. Frishman,D. (2002) Knowledge-based selection of targets for structural genomics. *Protein Eng.*, **15**, 169–183.
23. Kolesov,G., Mewes,H.W. and Frishman,D. (2001) Snapping up functionally related genes based on context information: a colinearity-free approach. *J. Mol. Biol.*, **311**, 639–656.
24. Kolesov,G., Mewes,H.W. and Frishman,D. (2002) SNAPper: gene order predicts gene function. *Bioinformatics*, **18**, 1017–1019.
25. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
26. Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Munsterkotter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.
27. Mewes,H.W., Albermann,K., Bahr,M., Frishman,D., Gleissner,A., Hani,J., Heumann,K., Kleine,K., Maierl,A., Oliver,S.G. *et al.* (1997) Overview of the yeast genome. *Nature*, **387**, 7–65.
28. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
29. Fellenberg,M., Albermann,K., Zollner,A., Mewes,H.W. and Hani,J. (2000) Integrative analysis of protein interaction data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 152–161.