BIOINFORMATICS

Vol. 17 no. 1 2001
Pages 44–57

Functional and structural genomics using PEDANT

Dmitrij Frishman[1,], Kaj Albermann[2], Jean Hani[2], Klaus Heumann[2], Agnes Metanomski[2], Alfred Zollner[2] and Hans-Werner Mewes[1]*

[1]GSF-Forschungszentrum für Umwelt und Gesundheit, Munich Information Center for Protein Sequences (MIPS) am Max-Planck-Institut für Biochemie, Am Klopferspitz 18, 82152 Martinsried, Germany and [2]Biomax Informatics AG, Lochhamer Straße 11, 82152 Martinsried, Germany

Received on April 28, 2000; revised and accepted on June 23, 2000

ABSTRACT

Motivation: Enormous demand for fast and accurate analysis of biological sequences is fuelled by the pace of genome analysis efforts. There is also an acute need in reliable up-to-date genomic databases integrating both functional and structural information. Here we describe the current status of the PEDANT software system for high-throughput analysis of large biological sequence sets and the genome analysis server associated with it.

Results: The principal features of PEDANT are: (i) completely automatic processing of data using a wide range of bioinformatics methods, (ii) manual refinement of annotation, (iii) automatic and manual assignment of gene products to a number of functional and structural categories, (iv) extensive hyperlinked protein reports, and (v) advanced DNA and protein viewers. The system is easily extensible and allows to include custom methods, databases, and categories with minimal or no programming effort. PEDANT is actively used as a collaborative environment to support several on-going genome sequencing projects.

The main purpose of the PEDANT genome database is to quickly disseminate well-organized information on completely sequenced and unfinished genomes. It currently includes 80 genomic sequences and in many cases serves as the only source of exhaustive information on a given genome. The database also acts as a vehicle for a number of research projects in bioinformatics. Using SQL queries, it is possible to correlate a large variety of pre-computed properties of gene products encoded in complete genomes with each other and compare them with data sets of special scientific interest. In particular, the availability of structural predictions for over 300 000 genomic proteins makes PEDANT the most extensive structural genomics resource available on the web.

Availability: The PEDANT genome analysis server is available at http://pedant.mips.biochem.mpg.de.

Contact: Genome sequencing centres interested in inclusion of their sequences in the PEDANT database should contact Dmitrij Frishman (frishman@mips.biochem.mpg.de).

*To whom correspondence should be addressed.

INTRODUCTION

Distilling meaningful information from billions of A, C, G and T characters generated by genome sequencing projects has become a formidable task for bioinformatics, imposing the need for more efficient, sensitive and reliable data analysis tools. It has also developed into a major stimulating factor for the research in computational molecular biology and led to appearance of totally novel scientific problems. Those include whole-genome gene prediction, cataloguing biological functions for a given organism, cross-genome comparisons, and support for structural genomics efforts, to name just a few. An increasingly important requirement is high productivity of data analysis and automation of possibly a large number of operations in order to allow the experts to concentrate on creative tasks requiring human attention. Last but not least, it has become evident that coping with large volumes of genome data poses a challenging technical problem. Processing of complete genomes with bioinformatics tools requires considerable computer resources, storing and retrieving tens of gigabytes of data necessitate the utilization of mature database management systems, and representing the results in an easily comprehensible form makes advanced visualization tools a must.

An early step from traditional, case-oriented sequence analysis work to automated large-scale genome crunching was made by Scharf *et al.* (1994) who applied their GeneQuiz system to the first complete yeast chromosome

sequenced (Bork *et al.*, 1992). Several other systems followed, each with design specifics reflecting the purposes and scientific interests pursued by the authors as well as their background (Sonnhammer and Durbin, 1994; Gaasterland and Sensen, 1996; Medigue *et al.*, 1999; Harris, 1997; Walker and Koonin, 1997; Bailey *et al.*, 1998; Andrade *et al.*, 1999; Saqi *et al.*, 1999). While comparing and classifying the existing genome analysis programs is difficult and beyond the scope of the present contribution, it is possible to state that the differences between them typically lie in the relative weighting of protein oriented versus DNA oriented analysis and interactive work versus command-line operation as well as in the spectrum of bioinformatics tools applied, the sophistication of the user interface, and the presence or absence of conveniency features, such as project management and data editors. However, the most important parameter—the fidelity of the results produced—is hard to measure, and no comparative benchmarks have been published so far. Moreover, creating such benchmarks would be complicated by the fact that the objectives of different systems may vary in terms of the chosen balance between the sensitivity and selectivity of the analyses.

Our goal was to create a versatile, easily expandable, and powerful software system to address a possibly wide spectrum of tasks in genome scale sequence analysis. We wanted to use it as (i) a workhorse for general bioinformatics research, (ii) a common framework for a number of genome analysis projects, (iii) a complete database of annotated genomes, and (iv) a tool for routine automatic analysis of large amounts of genomic contigs and ESTs (expressed sequence tags) generated in the public domain as well as in the industrial environment. With these objectives in mind, we have developed PEDANT (Protein Extraction, Description, and ANalysis Tool), a software suite for high-throughput analysis of bio-molecular data. The first version of the PEDANT web site was made available over internet in mid-1997 (Frishman and Mewes, 1997). In this communication, we describe the second version of PEDANT and its web site as well as some of the scientific results obtained with its help.

SYSTEM ARCHITECTURE

Overview

PEDANT consists of three major parts (Figure 1): (i) the database module serves for storing, modifying and accessing data, (ii) the processing module actually carries out bioinformatics computations, and (iii) the user interface allows communication with the system through a web-based mechanism. In addition, a collection of external tools collectively referred to as 'Input module' is used for data formatting, preliminary data analysis steps (e.g. identification of genetic elements in DNA), and population of the database. Individual bioinformatics programs and the respective databases that they are using are also external to the system and have to be properly installed.

Data access

The data access mechanism implemented in PEDANT is based on a standard RDBMS and the SQL language. At the present time the freely available MySQL DBMS is being used. However, all SQL calls are encapsulated and implemented using the universal PERL module (DBI) which also supports a number of other free and commercial systems, including Oracle. A CORBA interface for PEDANT is also available (A.Kaps, personal communication).

As seen in Figure 1, PEDANT supports two major types of SQL tables. Primary tables are used to store raw data, such as DNA and protein sequences as well as the results of individual bioinformatics applications (e.g. BLAST output). These results are subsequently parsed and stored in secondary data tables such that each individual piece of evidence can be retrieved, deleted, or updated.

A simplified PEDANT database schema is depicted in Figure 2. The name convention for primary tables is 'name_data', where 'name' can be 'prot', 'contig', 'blast', etc. All primary tables have the same structure. They refer to the special data table called *contig_data* via the foreign ID *contig_data_id*. The *contig_data* table contains contig sequences (bacterial contigs, genomes, ESTs). For example, if there are several bacterial contigs in a given PEDANT database, all data entries containing proteins from contig STY556 with the ID 12 in the *contig_data* table will have *contig_data_id*=12 and *contig_data_code*="STY556". All further protein-related results corresponding to these proteins will have the same *contig_data_id* and *contig_data_code*. This approach allows to establish unambiguous relations between contigs and genetic elements they contain, which is especially important for visualization. If no DNA data is available, all proteins will be associated with *contig_data_id*=0 by default. The *code* field typically refers to the protein code (e.g. P78996, NTB_ECOLI). It is assumed to be unique within each contig. The actual data blob (e.g. blast output) is stored in the *dat* field.

The names of secondary tables are derived from the names of primary tables by removing the '_data' part. For example, the table containing the parsed blast results (stored in the dat field of the 'blast_data' table) will be called 'blast'. Each secondary table contains obligatory fields relating each entry to a particular contig and a particular protein or gene via the fields *contig_data_id* and *prot_data_id*, respectively. The fields *conf* and *manual* are used to flag manual modifications and assign confidence levels to them. The rest of the fields are specific for each secondary table. There may be several entries corresponding to the same *prot_data_id*. For example, a protein can

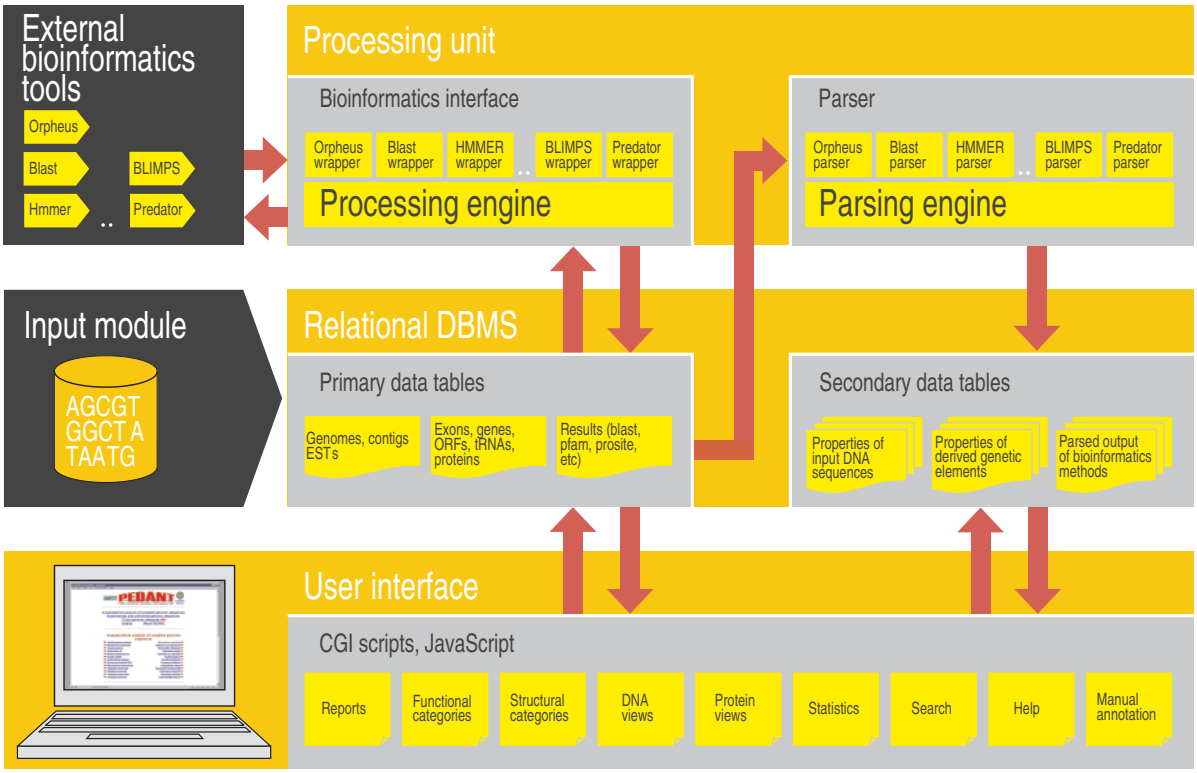

Fig. 1. PEDANT architecture. The three main parts of the system are (a) relational DBMS, (b) processing unit, and (c) user interface. After pre-processing in the input module, sequence data (DNA contigs, proteins, genetic elements, exons, genes, etc.) are loaded into input primary tables. The processing unit automates the application of various bioinformatics methods (e.g. BLAST searches, secondary structure predictions) to each data element; results of the calculations are saved in the output primary tables. Results are subsequently parsed and stored in secondary tables where each piece of information (e.g. local BLAST alignments, *E*-values, secondary structure elements, etc.) can be individually accessed. The user interface allows to access the data using a standard WWW browser. See text for more explanations.

have several different blast hits, and several alignments with the same blast hit.

In addition, there are a number of special tables that have different structure. These tables hold blast indices, update information, dataset- and user-specific information, etc.

Operation in command line mode

There are three principal functions that can be effected in command line mode: (a) applying bioinformatics methods to sequences, (b) parsing the data tables, and (c) querying the resulting database. The processing engine of PEDANT (see Figure 1) reads each entry from a source primary table (e.g. prot_data), subjects it to a given bioinformatics method (e.g. blast), and writes the result in the appropriate destination primary table (e.g. blast_data). This table, in its turn, can serve as a source table for another method. For example, blast alignments can serve as input for secondary structure prediction. Computation will go on until all entries in the source table have been processed. If a multi-processor computer or a farm of workstations is available, it is possible to start many parallel jobs operating on the same output table, or many jobs running different methods. Table locking prevents individual processes from conflicting with each other.

After a given primary table has been filled, it can be parsed into a secondary table. A large variety of queries (e.g. produce the list of all functional categories, produce the list of all ORFs attributed to the functional category ‘glycolysis’, etc.) can be performed both on primary and secondary tables. All parameters used to control the execution of the bioinformatics programs as well as for parsing of the tables and querying the database (location of the databases to be searched, search thresholds) are stored in a single configuration file and can be easily adjusted.

Web interface

Upon parsing all essential tables a given PEDANT dataset is immediately viewable using the genome browser. No static HTML pages are required. The DNA and protein viewers make direct access to the SQL tables.

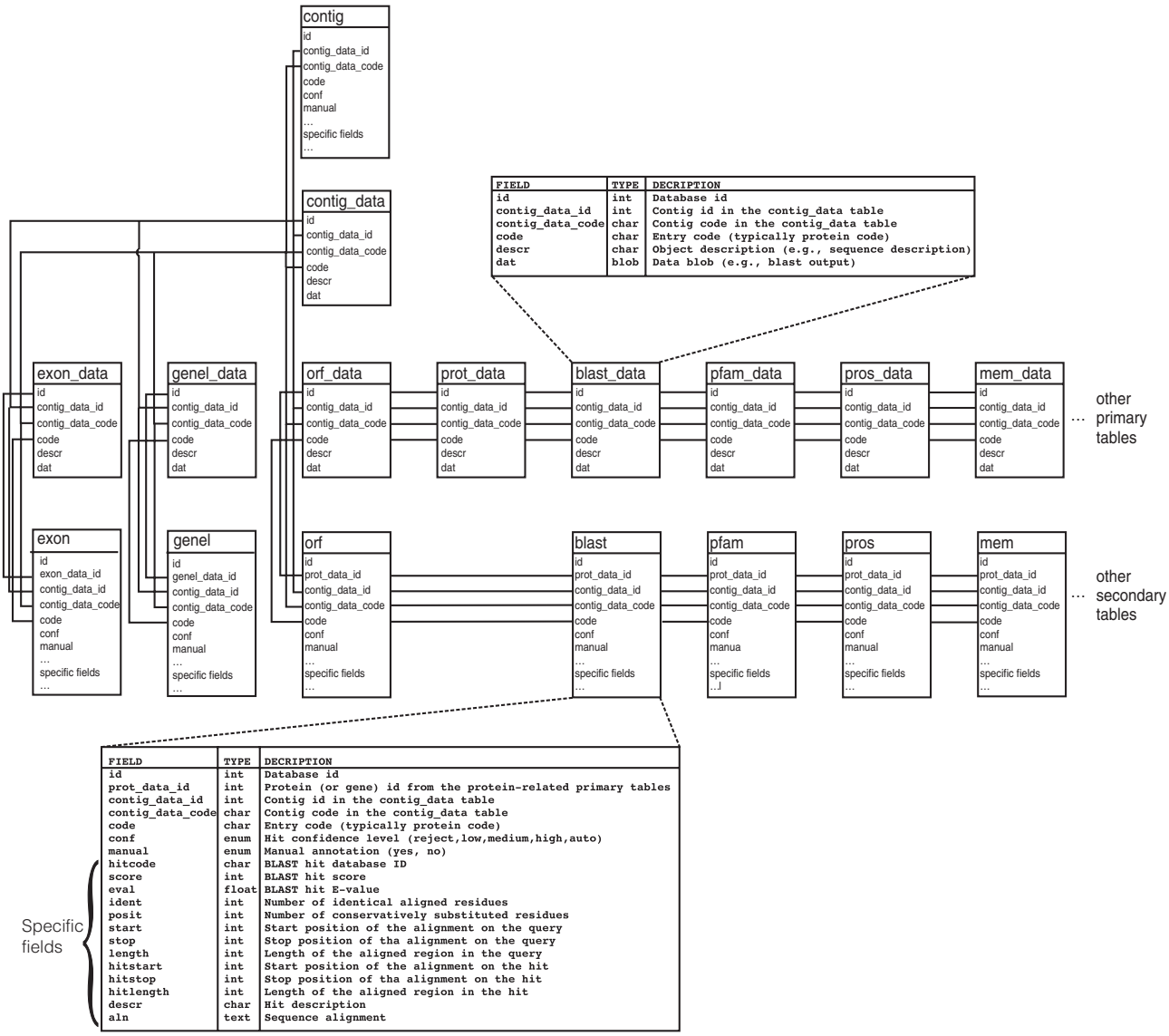

Fig. 2. PEDANT relational schema (simplified). See text for detailed explanation.

Implementation and system requirements

The core of PEDANT is written in Perl5 programming language. The only exception is the graphical viewer which is implemented in C++. PEDANT was extensively tested on COMPAQ, SGI, Hewlett–Packard, and Linux computers and should be easily portable to any other UNIX system. The client part of the system can also be used on a computer running MS-Windows.

Performance

The CPU time needed to process one protein sequence is practically equal to the sum of the times required by each individual method applied. Analysing a typical protein sequence from a bacterial genome takes approximately 3 min on a standard workstation. For shorter ORFs, e.g. ORFs extracted from EST sequences, less time is required for protein-related analyses. However, the total time may be longer if additional analyses on the DNA level (e.g. similarity searches in nucleic sequence databases) are performed.

Using a multi-processor computer system, processing can be accelerated due to the parallel capabilities built in the PEDANT system. At MIPS the LSF batch system (Platform Computing Corporation) is used to run PEDANT jobs and to balance the load between the 20 DEC-Alpha CPUs available. This allows to conduct

automatic annotation of an average size bacterial genome in just one day.

BIOINFORMATICS METHODS

Overview of the PEDANT processing pipeline

In a most typical situation, any number of genomic contigs or ESTs can be submitted to PEDANT through the input module (Figure 1) which supports user and dataset management and allows to choose various analysis parameters. Dependent on the type of DNA data, appropriate algorithms for identification of coding regions and various genetic elements will be first applied. Extracted gene products are subjected to exhaustive bioinformatics analysis, including homology searches, detection of protein motifs, prediction of secondary structure and other protein features, as well as sensitive fold recognition. Proteins are also automatically attributed to pre-defined functional categories. We sought to select a set of computational techniques and sources of information that would be complementary to each other. This set is highly dynamic and is frequently updated, reflecting the progress in the bioinformatics field. A full list of the computational methods and databanks used by PEDANT is available at http://pedant.mips.biochem.mpg.de/about.html.

Prediction of genes and other genetic elements

Dependent on the source and nature of the nucleic acid sequence submitted for analysis, an appropriate method to extract protein coding regions will be applied as detailed in Table 1. The user has the option to choose one of the 15 genetic codes to be used for the analysis (Jukes and Osawa, 1993; Osawa *et al.*, 1992, http://www.ncbi.nlm.nih.gov/htbin-post/Taxonomy/wprintgc?mode=c), otherwise the standard genetic code will be assumed. For short bacterial genomic contigs full-scale gene prediction procedure can not be applied since there is not enough data to derive reliable coding potential information and ribosome-binding site consensus; six-frame translation is used instead.

Functional and structural categories

The main distinctive feature of the PEDANT system is its ability to assign proteins to automatically derived structural and functional categories. The categorization system is multidimensional in that each sequence can be assigned to many different categories, and each category can contain any number of gene products.

The main vehicle for similarity searches is the PSI-BLAST algorithm developed at the National Center for Biotechnology Information, Bethesda (Altschul *et al.*, 1997). This method is used for general-purpose searches against the full non-redundant protein sequence databank as well as searches against a number of special datasets, including the MIPS functional categories (see below) and the COG database (Tatusov *et al.*, 1997). In addition, detection of PROSITE (Hofmann *et al.*, 1999), PFAM (Bateman *et al.*, 2000), and BLOCKS (Henikoff *et al.*, 1999) sequence motifs is performed. For those sequences that have significant matches in the PIR-International Protein Sequence Database (Barker *et al.*, 2000), the annotation of the respective entries is analysed and keywords, enzyme classification, and superfamily information is extracted.

Structural categorization of gene products involves PSI-BLAST searches against the sequences with known 3D structure as deposited in the PDB databank (Berman *et al.*, 2000). If a significant relationship exists, the secondary structure assignment of the respective three-dimensional structures as defined by the STRIDE software (Frishman and Argos, 1995) is inserted into the PEDANT structural summary in upper case. Otherwise, secondary structure information predicted by PREDATOR (Frishman and Argos, 1997) is shown in lower case. Other predicted structural features include low complexity regions (Wootton and Federhen, 1993), membrane regions (Klein *et al.*, 1985), coiled coils (Lupas and van Stock, 1991), and signal peptides (Nielsen *et al.*, 1997). Highly sensitive comparison of each predicted protein with the SCOP database of known structural domains (Lo *et al.*, 2000; Brenner *et al.*, 2000) is carried out using the novel IMPALA software (see Section **PEDANT as a structural genomics resource**).

The computational methods described above form the core of the PEDANT processing pipeline. In addition, any other similarity searches against user-supplied datasets can be conducted, and their results appropriately visualized, without the need to modify the program code. Due to the open architecture of PEDANT, novel techniques can be easily added to the system, requiring only a minor coding effort.

Yeast biological role categories

The principal raison d'être of genome sequencing is to describe the pathways existing in a given organism and, consequently, to understand its physiology. To achieve this goal, the specific function(s) of each individual gene product should be inferred as precisely as possible based on the evidence available. On a less detailed level, it has proven to be extremely instrumental to categorize genes according to their function. The system of biological role categories was first developed by Riley (1993) to describe the genes of *E.coli* known at that time. This system was later adapted for other bacterial genomes (e.g. Fleischmann *et al.*, 1995; Kunst *et al.*, 1997).

During the yeast genome sequencing project, an advanced hierarchical functional catalogue was designed at MIPS based on the Riley scheme to address a much broader and complex spectrum of functions

Table 1. Methods used to extract coding regions and genetic elements from DNA contigs

Sequence source	Sequence type	Procedure	Program/Method	Reference
Eukaryotes	Genomic DNA	Gene prediction	GenScan	Burge and Karlin (1997)
Prokaryotes	Genomic DNA	Gene prediction	Orpheus	Frishman *et al.* (1998)
All	EST	Prediction of the most probable ORF	Six-frame translation with subsequent verification through BlastX and BlastN searches. If no significant hits are found, the longest ORF is taken.	Altschul *et al.* (1997)
Human	EST	Prediction of the most probable ORF	Same as above, but optionally consideration of ESTScan predictions is possible.	Iseli *et al.* (1999)
All	Genomic DNA	tRNA prediction	tRNAScan	Lowe and Eddy (1997)
All	Genomic DNA	Prediction of other non-protein coding genetic elements (rRNAs, scRNAs, snRNAs, misc. RNAs, origin of replication, ARS, CEN and LTRs)	DDS search against a selection of genetic element sequences from the EMBL database	Huang *et al.* (1997)
All	All	Six-frame translation	Orpheus	Frishman *et al.* (1998)

present in this eukaryotic organism (Mewes *et al.*, 1997, http://www.mips.biochem.mpg.de/proj/yeast). The novel aspect of the MIPS catalogue is its multidimensionality—a gene product can be attributed to several functional categories. This feature allows for efficient handling of multi-domain proteins as well as multi-functional domains. The catalogue has a hierarchical structure. Each of the 15 main classes (e.g. metabolism, energy) contains three to four subclasses, with the total number of functional categories exceeding 200. Nearly 4000 yeast genes could be ascribed to at least one functional category based on careful manual analysis of extrinsic evidence (similarity to known proteins, presence of indicative sequence patterns) as well as experimental data from the literature.

Within the PEDANT system, the MIPS classification is being used for automatic assignment of functional categories to gene products based on significant homology to one or many functionally characterized yeast genes. This approach is certainly not perfect due to the differences in function specificity between the proteins from different organisms, and between paralogous proteins from the same organism. However, it allows to create a useful first approximation which can be subsequently refined by manual annotation. As seen in Figure 3, on the highest classification level (broad categories such as metabolism, energy, etc.), the differences in physiology between different organisms are reasonably captured. For example, it is easy to see that 'parasitic' organisms (e.g. *T.pallidum* and *B.burgdorferi*) show only limited metabolic capacities, compared to free living organisms as for example *B.subtilis*. In contrast, higher multicellular eukaryotic organisms, like *C.elegans*, have a much higher fraction of proteins related to cellular communication processes compared to lower unicellular eukaryotes such as *S.cerevisiae*.

Visualization

The PEDANT genome browser provides access to contigs, ORFs, and their annotation in a variety of ways. It allows to select individual functional and structural categories and conduct text searches in annotation and BLAST searches against the sequences belonging to the dataset. For each ORF in the dataset an integrated, hypertext-linked protein report is provided showing analysis results according to dynamically set thresholds (Figure 4). 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 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

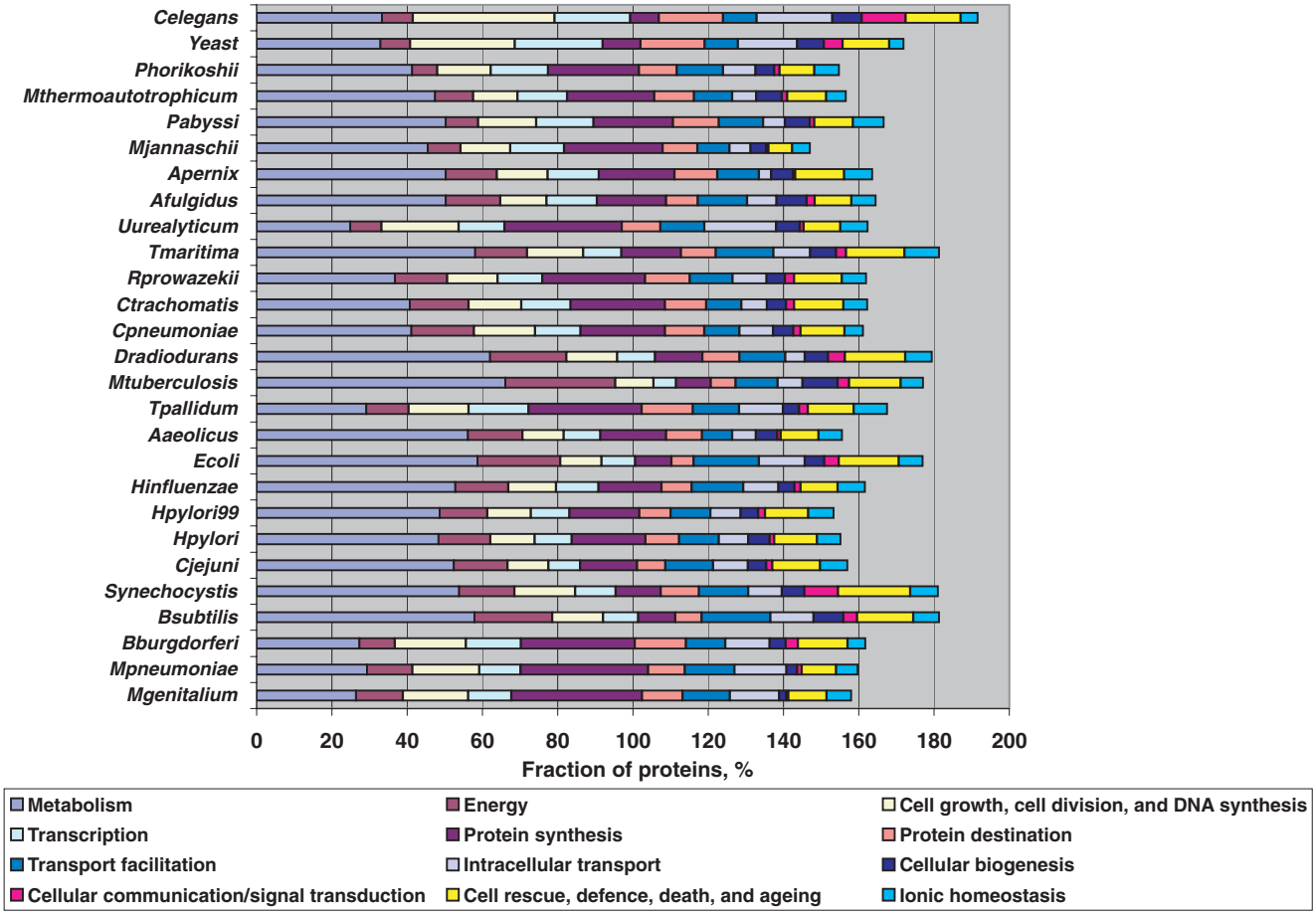

Fig. 3. Distribution of ORFs in completely sequenced genomes over the high-level functional categories of the MIPS functional catalogue derived through high-stringency PSI-BLAST searches against yeast sequences representing respective categories. Note that due to multidimensionality of the functional catalogue, gene products may be attributed to several categories; hence the sum of fractions for each particular genome is typically greater than 100%.

predicted protein features, e.g. PROSITE motifs and PFAM domains. This is especially useful for judging on the domain structure of the homology hits.

Automatic versus manual annotation

The number of PEDANT users complaining about excessively optimistic functional assignments is approximately equal to the number of those who consider the default settings too conservative. It is clear that any automatically produced sequence analysis implies a reasonable compromise between sensitivity and selectivity, and that no ideal recognition threshold exists that would allow for perfect separation of true and false similarities. In addition, in spite of the continuous improvement in the overall quality of bioinformatics methods, a number of complications in gene functional assignment can hardly be addressed in a completely automatic fashion. Most notably, the problem of error propagation in databases (Bork and Bairoch, 1996) is intrinsically unsolvable without human intervention. Once introduced in a public database, protein sequences corrupted due to sequencing artefacts or derived from wrong gene models as well as erroneous annotation of database entries caused by human error or insufficient knowledge threaten to influence subsequent annotation efforts. Other typical sources of false annotations (Galperin and Koonin, 1998) include spurious similarity hits caused by compositionally biased protein sequences and failure to take into consideration multi-domain organization of proteins. Only a limited improvement can be achieved through the application of filtering algorithms and taking into account the domain structure of the similarity hits.

Thus, the genome analysis produced by any automatic system should be considered as a useful first

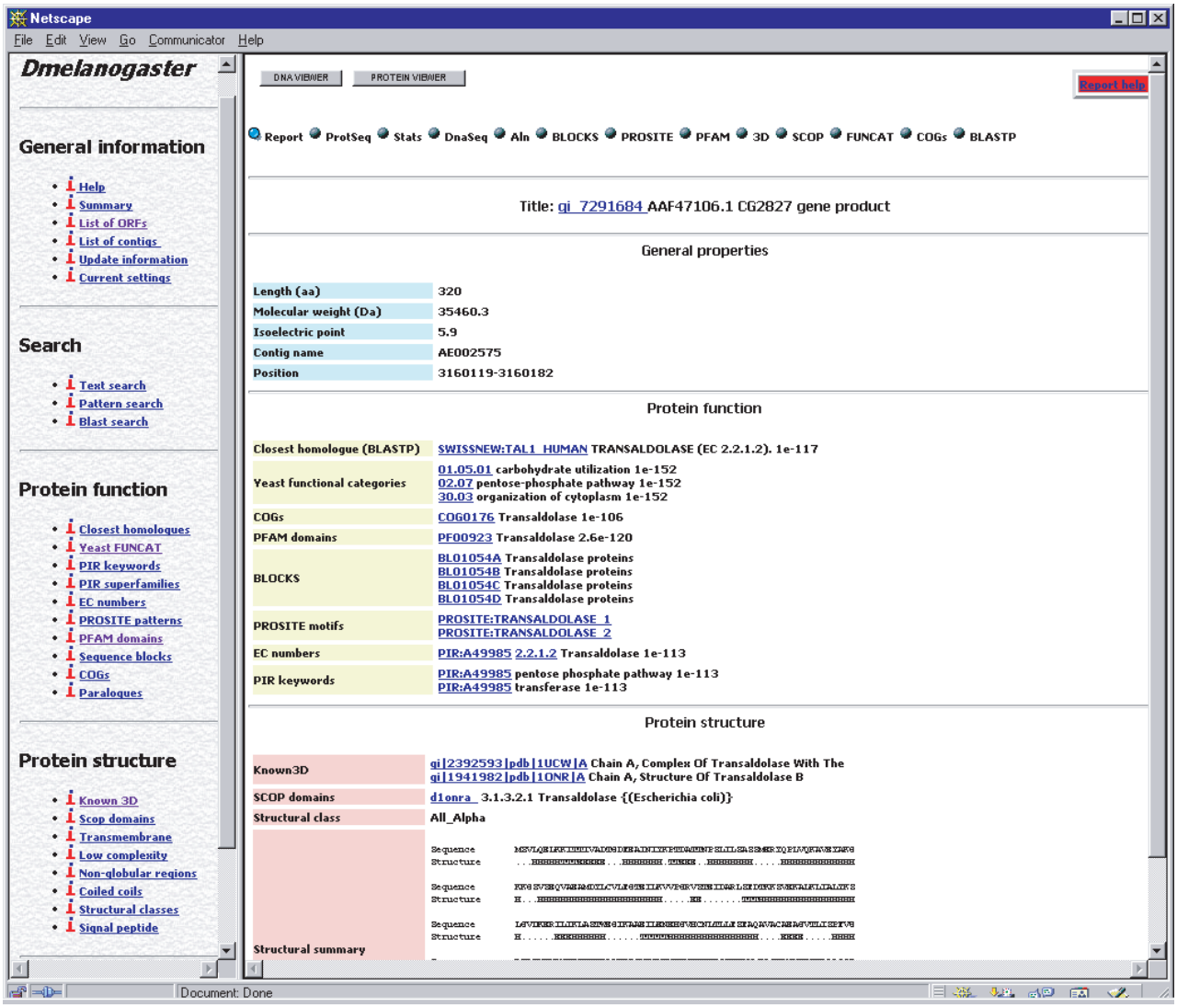

Fig. 4. Protein report page. The buttons in the upper part of the page launch the DNA viewer and the protein viewer. In the next row, links are provided to raw results of the bioinformatics calculations, e.g. BLAST output. Three main sections of the report provide a summary of general features of the protein, functional information, and structural assignments. Any number of additional, user-specified fields can be introduced in the course of manual annotation.

approximation. Further improvement of the data quality requires involvement of human experts. To address this problem, the PEDANT genome browser now includes a comprehensive environment for manual sequence annotation which allows to modify, delete and add ORFs, introduce arbitrary data fields, and assign ORFs to custom categories.

In particular, for the manual functional categorization of the genes, a catalogue independent from the automatic characterization via similarity to yeast genes was established. This catalogue was extended to include functions specific for plants and bacteria (e.g. secondary metabolism and special pathways found only in bacteria). In contrast to the yeast categories that are assigned to protein sequences automatically via similarity searches, categories from this extended catalogue are assigned by the annotators manually. We discovered that the great advantage of our functional catalogue is its flexibility. It can be adapted to every organism, but has retained its fundamental structure since its first version. Whenever the knowledge, time or money prevents finer categorization, a protein can first be quickly placed into a certain higher category of the catalogue; in a later annotation step it is possible to move the protein into finer categories. At the moment the functional

catalogue has 528 categories in total. It is divided into 20 main categories, 143 second level categories, 160 third level categories, 128 forth level categories, 62 fifth level categories and 12 sixth level categories.

Among the other standard data fields forming the manual annotation are 'Title', 'Gene ID', 'Classification' (e.g. known protein, strong similarity to known protein, similarity to unknown protein, etc.), 'PubMed ID', 'Cellular localization', as well as free text 'Remarks' and 'Comments'. In addition, with a single PEDANT command, it is possible to create dataset-specific annotation fields of the following types: pull-down menu, text box, link, and selection. The genome browser also allows to create ORF groups; such groups can be selectively visualized or exported.

Another important feature of the manual annotation module is the possibility to assign confidence levels ('reject', 'low', 'medium', 'high') to any piece of automatically generated evidence, which by default has the confidence level 'auto'. Thus, if the top scoring similarity hit for a given query sequence is an experimentally uncharacterized or hypothetical protein, it can be rejected and will not appear in the annotation any more, while the next best hit will be used. By setting manually chosen confidence levels to BLAST hits against various databases as well as to PFAM, PROSITE and BLOCKS domains found, the overall quality of the annotation can be improved.

The decision to subject a given genome to careful manual annotation influences the strategy of the automatic processing steps. In contrast to the case when contigs are analysed in an automatic fashion only, larger overlaps between ORFs are allowed. In addition, less stringent recognition thresholds are used for similarity searching. These measures lead to a significantly greater number of false positives both at the gene prediction and at the functional assignment steps, but reduce the chances to miss important evidence. They also increase the productivity of manual annotation since deleting false assignments is generally faster and easier than adding missed bits of information.

A further advantage of the manual annotation subsystem is that it enables a group of users, possibly at different locations, to co-operate on a number of annotation projects. At MIPS, the genome of *T.acidophilum* (Ruepp *et al.*, 2000) was analysed in cooperation with SmithKlineBeecham Pharmaceuticals, while the genome of *H.salinarum* is currently being annotated by a large consortium of scientists from several labs in Europe and the USA.

Data release management

Genome sequencing projects are typically carried out under time pressure, often caused by competition with other sequencing labs or commercial companies. In order to win time, it is thus mandatory to begin the annotation process at the very early stages of sequencing, when only unordered DNA contigs are available. As new contigs, and eventually the final complete genome sequence become available, they have to be intelligently merged with the existing data pool. While re-calculating the automatic part of the sequence annotation is only a matter of CPU resources, the labour-intensive part of the analysis conducted by human experts must be retained. This leads to the trivial but technically difficult problem of transferring the manual annotation between subsequent data releases.

As illustrated in Figure 5, there are several circumstances that complicate the annotation transfer: (i) separate genes on the same contig may become fused due to the slightly changed sequence, e.g. as a result of a frameshift correction, (ii) separate genes on two different contigs may become fused if the new contig incorporates both old contigs, (iii) gene boundaries are subject to change if the contigs are extended in length or their sequence is altered and: (iv) new genes may appear in the new portions of the sequence. It is therefore clear that the annotation transfer can not be perfect and in some cases existing annotation must be reconsidered, e.g. if a gene product 'acquires' a new domain. However, most of the manually introduced data can be automatically mapped on the revised sequence data.

In PEDANT, all manually editable tables contain a special field 'manual' which can be set to just two values—'yes' or 'no'. For all automatically generated data (e.g. blast hits), the value of 'manual' is initially set to 'no'; for all manually introduced or automatically calculated and subsequently altered data it is set to 'yes'. Each new release of sequenced data is first subjected to the full cycle of automatic annotation, including gene prediction and PEDANT analysis. The annotation transfer process starts with running a high-stringency BLAST search with each predicted gene product of the new release against the set of proteins from the previous release. Then all data fields in the old release with *manual*='yes' are transferred to the corresponding ORFs of the new release based on sequence similarity. Thus, a PFAM domain identified in a certain ORF in the new release will first have the attributes *manual* and *conf* set to 'no' and 'auto', respectively. If during the annotation transfer process the corresponding ORF in the previous release will be found to have the same domain rejected by the annotators, the attributes *manual* and *conf* in the ORF from the new release will be set to 'yes' and 'reject', respectively. Consequently, the PFAM domain involved will not appear in the annotation, although it will still be present in the raw PFAM output.

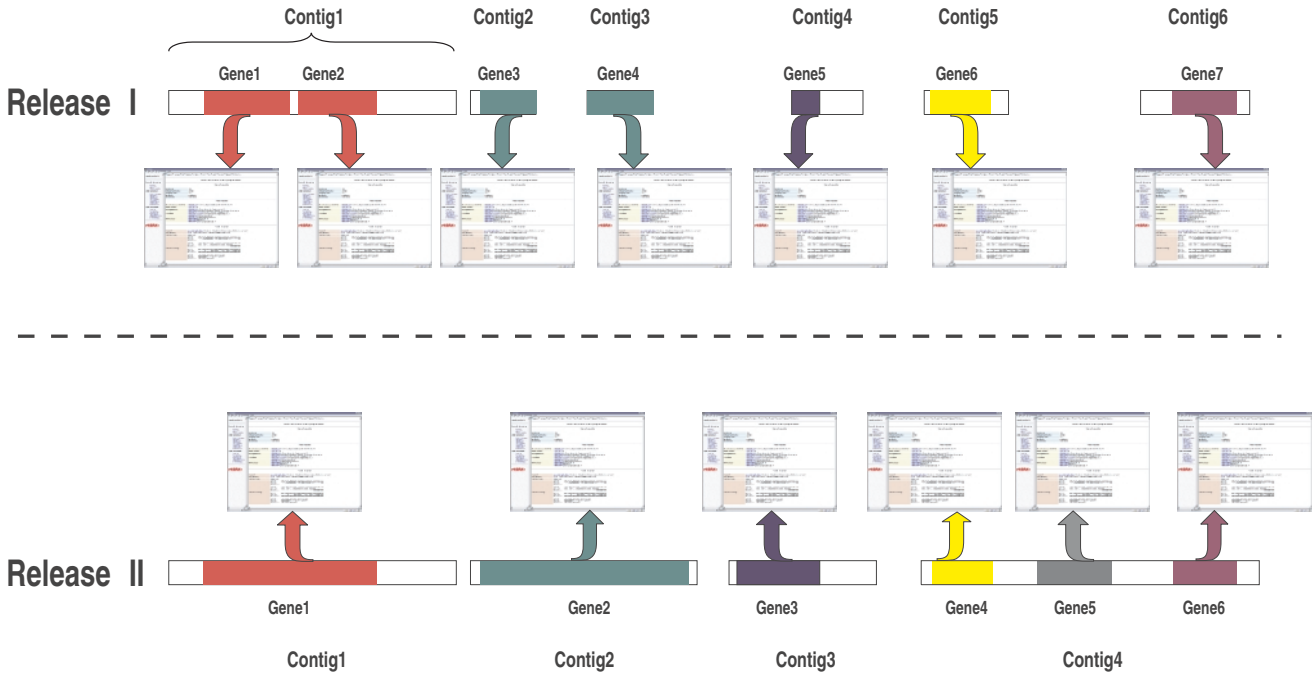

Fig. 5. Manual annotation transfer between two subsequent genome releases. Some typical problem cases are shown, including gene fusion as a result of sequence correction or contig merging, modification of gene boundaries caused by altered contig sequences, and the appearance of new genes in the newly sequenced portions of the genome.

THE PEDANT GENOME DATABASE

Annotation of publicly available completely sequenced and unfinished genomes

Over the past three years, the PEDANT system was systematically applied to analyse genomic sequences available in the public domain. The result is a comprehensive database which currently provides computational analysis of 80 genomes, with the total amount of data managed by the RDBMS approaching 100 gB. The PEDANT web site is split in three major divisions:

(a) Genomes that are being annotated and published by MIPS. This section currently includes *A.thaliana*, *N.crassa*, and *T.acidophilum*; the genome of *H.salinarum* is in preparation and will be added shortly. These datasets include extensive manual annotation.

(b) Completely sequenced and published genomic sequences. In most of the cases the sequence data and ORF nomenclature as provided by the NCBI genomes division are employed, and the ORF descriptions supplied by the original authors are preserved.

(c) Unfinished and/or unpublished genomic sequences. Gene prediction is conducted by ORPHEUS (Frishman *et al.*, 1998) in a completely automatic fashion, usually allowing for large overlaps between ORFs. This leads to many overpredicted ORFs, but ensures that fewer real ORFs are missed. In many cases, the PEDANT database is the only source of annotation for such datasets.

PEDANT as a structural genomics resource

Structural genomics is an emerging area of biological research aimed at solving the complete representative set of protein structures through the application of high-throughput structure determination techniques. Particular areas of work include exhaustive structural analysis of all proteins from a number of model organisms with completely sequenced genomes and the class-based approach focusing on protein groups of special medical or biotechnological interest (Terwilliger *et al.*, 1998). Computational molecular biology created the rationale for structural genomics by deriving the general principles of protein structure organization and by providing a tentative upper boundary for the total number of existing protein folds, efficient ways of their prediction and classification. Comparative protein sequence and structure analysis is a major cost-saving factor in high-throughput structure determination leading to optimal, most economic selection of targets for x-ray crystallography or NMR

studies. The cornerstone computational approach used in structural genomics is similarity-based fold recognition in completely sequenced genomes.

Constant progress in bioinformatics software tools parallels the increase of data volumes and complexity and allows to routinely obtain results which were previously the domain of experts. The National Center for Biotechnology Information (NCBI, Bethesda) has recently released a new software tool for sensitive similarity searches called IMPALA (Schäffer *et al.*, 1999). This program allows to compare a query protein sequence with a collection of position specific scoring matrices generated by BLAST and is thus perfectly suitable for similarity-based fold recognition. Our current approach to genomic fold recognition involves the following steps: (i) create a non-redundant protein sequence database with proteins possessing predicted membrane regions, coiled coils, and low complexity regions eliminated, (ii) run a PSI-BLAST search with ten iterations with each SCOP domain against the non-redundant protein sequence database (prepared with the *nrdb* program, W.Gish, unpublished) and save the resulting profiles, (iii) construct a SCOP profile library using the IMPALA software suite, and (iv) run an IMPALA search with each genomic sequence against the SCOP library. The same procedure is applied to the non-redundant collection of complete PDB sequences. The performance of IMPALA in terms of the percentage of genomics proteins assigned to folds (Figure 6) as well as its selectivity is comparable to many advanced threading techniques published so far (Frishman, in preparation).

In spite of significant advances in assigning proteins to known structures, a majority of gene products encoded in complete genomes are still 'structural orphans'. These proteins can only be structurally characterized on a very coarse level using a variety of prediction techniques. Secondary structure prediction can help to attribute a given protein to one of the major folding classes and evaluate the significance of the database hits and their domain arrangement. Membrane region predictions and detection of coiled coils can be very informative for functional classification, while detection of signal peptides is instrumental in judging on protein cellular localization. The PEDANT genome browser allows to select sets of proteins predicted to have a given structural feature. Predicted and homology-derived structural features are also visualized on each protein report page.

To summarize, structural assignments and predictions for over 300 000 genomic proteins are available in the PEDANT database at the time of writing, which makes it the most comprehensive resource of this kind on the web.

Cross-genome comparison

As described in Section **Data access**, the PEDANT relational scheme allows to handle multiple contigs within one database. This property can be utilized to create, without any technical modification of the software, cross-genome datasets, in which each genome is treated as an individual contig. All queries that are typically made on individual genomes (e.g. find all ORFs assigned to a given functional category) can thus be easily performed on the full set of genomes available at the PEDANT site. At present, the cross-genome dataset is implemented for 44 genomes and several most important types of queries: functional categories, PIR keywords, superfamilies, and EC numbers, PROSITE, PFAM and BLOCKS motifs as well as SCOP structural domains.

In addition, all-against-all comparison of the protein complements for the completely sequenced genomes was conducted, and the corresponding BLAST similarity hits are visualized on each protein report page; hyperlinks allow to navigate from one genome to another.

APPLICATIONS

Arabidopsis thaliana chromosome IV

The sequence of the *A.thaliana* chromosome IV was determined jointly by the European Union and US Arabidopsis Genome Sequencing Consortiums (Mayer *et al.*, 1999). 3744 protein coding genes were identified using a variety of gene prediction programs and considering extrinsic evidence available. PEDANT was used as the main annotation engine for the protein complement. Up to 90% of the proteins had significant BLAST matches in the protein sequence database. However, careful manual classification of the similarity data (Figure 7) demonstrates that only roughly 30% of the gene products are known proteins or strongly similar to known proteins. A sizeable portion of the similarity hits are either relatively weak or come from proteins with unknown function. In the course of the manual annotation, BLAST hits of every 5th protein were manually corrected. Thus, the resulting analysis represents a significant departure from the first pass automatic annotation.

Among the most important highlights of the chromosome IV analysis was the comparison of protein structural classes with other model organisms. It was revealed, for the first time, that multi-cellular organisms tend to have a higher fraction of all-alpha and a smaller fraction of mixed alpha/beta structural domains than unicellular species.

Assembled human transcripts

A large collection of human UniGene (Wheeler *et al.*, 2000) clusters was subjected to PEDANT analysis (Geier *et al.*, in preparation; http://www.mips.biochem.mpg.de/proj/human/pedant). The challenge here is primarily of a technical nature. The dataset comprises over 75 000 contigs, from which the most significant coding region is used for further scrutiny. The total amount of data in

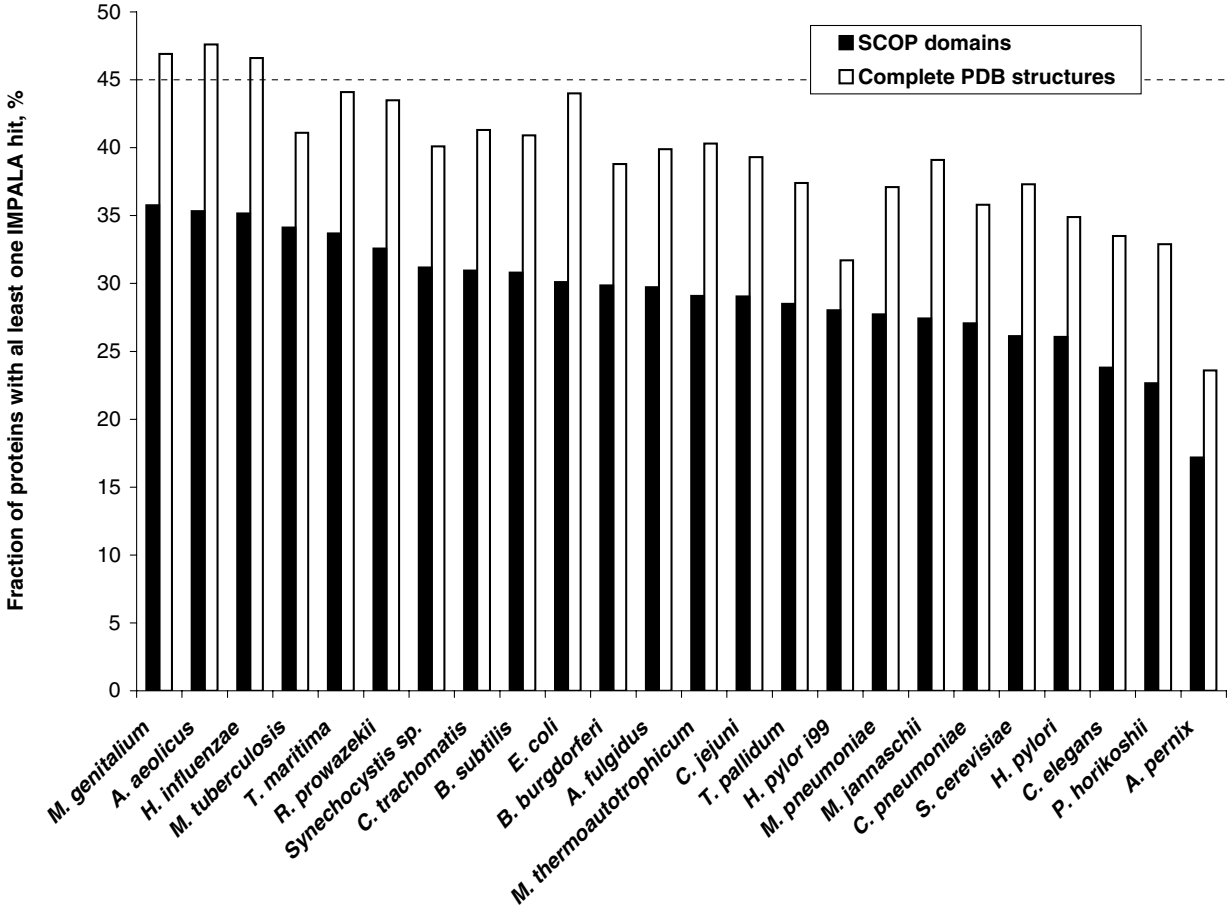

Fig. 6. Performance of the IMPALA algorithm in detecting SCOP structural domains and complete PDB structures in genomic proteins.

this particular MySQL database is close to 8 gB, and some of the MySQL tables contain over half a million lines. Due to appropriate optimization, queries in this large database are accomplished in acceptable time such that the PEDANT user interface can be used interactively to view the results. This demonstrates the suitability of the PEDANT relational scheme for supporting large-scale EST sequencing projects. One such project, involving analysis of over 400 000 EST sequences, is currently underway at Biomax Informatics AG.

Analysis of the GroEL substrates

The PEDANT system can be used as a general purpose bioinformatics tool and applied to a wide spectrum of research problems. One such investigation involved a computational analysis of the proteins interacting with the common *E.coli* chaperonin GroEL (Houry *et al.*, 1999, http://pedant.mips.biochem.mpg.de/GROEL/). Exact identities of 52 GroEL substrates were experimentally determined using immunoprecipitation and 2D-gel electrophoresis, and the corresponding protein sequences processed with PEDANT. The central question to be answered with bioinformatics means was that of a structural motif common for proteins relying on GroEL for folding *in vivo*. Comparison of their predicted structural classes with those of the full complement of soluble *E.coli* proteins indicates that GroEL substrates predominantly consist of two or more α/β domains involving buried β-sheets with large hydrophobic surfaces. Such proteins are especially prone to aggregation and therefore critically need the chaperonin for productive folding.

On a more general line, the availability of structural predictions and similarity-based domain assignments for all genomic proteins allows to conduct comparisons of any protein set of interest with the complete protein complement of a given genome and thus delineate its specific properties. This approach is highly compatible with the target-based structural genomics efforts (Terwilliger *et al.*, 1998), aimed at elucidation of structural features of proteins of special interest for biotechnology and medicine.

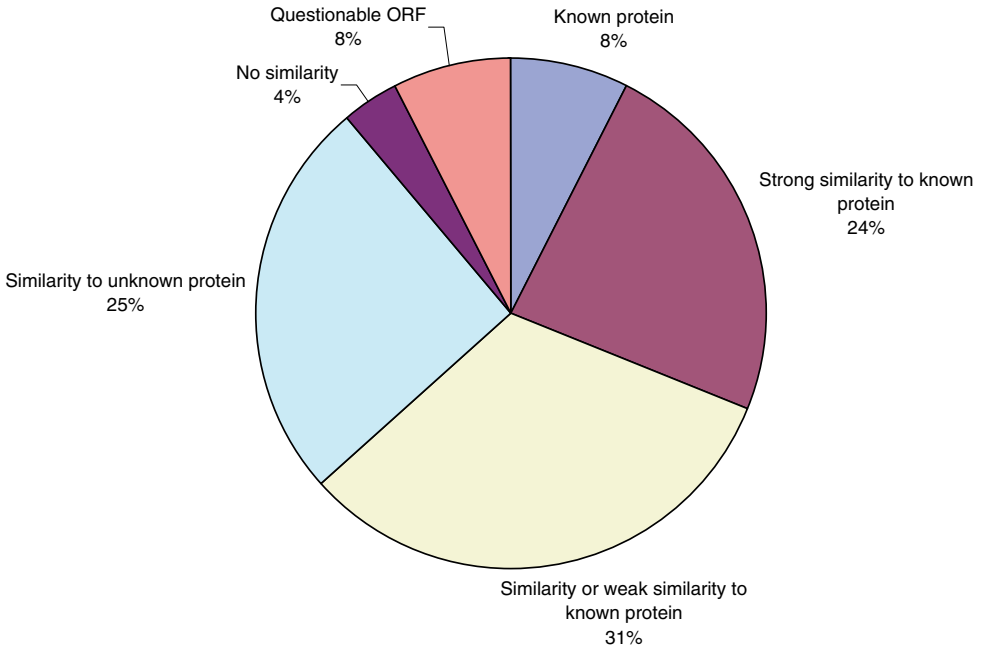

Fig. 7. Classification of predicted gene products in the chromosome IV of *A.thaliana* in terms of the degree of their homology to functionally characterized proteins based on BLAST scores.

OUTLOOK

Pitfalls of automated sequence analysis notwithstanding, the PEDANT software suite and the genome database associated with it have proved to be a useful tool for genome annotation and bioinformatics research. Due to the dynamic nature of the bioinformatics field, constant efforts have to be made to keep up-to-date the set of computational techniques and databases utilized. Even more importantly, better decision rules need to be employed in order to improve the quality of the automatic annotation and reduce the effort spent by human experts on manual annotation. For example, work is in progress to achieve better treatment of the spurious function assignments caused by multidomain proteins. Another upcoming enhancement is the incorporation of the similarity-free approach to function prediction which exploits functional coupling between genes located in adjacent positions on the chromosome (Overbeek *et al.*, 1999). Other planned developments include: new features in the genome viewers (e.g. representation of global DNA statistical tendencies), the possibility to manually annotate and manipulate predicted genetic elements (e.g. long terminal repeats), support of the Oracle^{TM} RDBMS, implementation of the automatic gene prediction pipeline for higher eukaryotes, improved interface for queries in the PEDANT database, interactive capabilities (e.g. re-processing of certain parts of data with user-modified parameters), and better update mechanisms.

ACKNOWLEDGEMENTS

We would like to thank Ole Bents and Norman Strack for their assistance with the system software and Birgitta Geier, Friedhelm Pfeiffer, Susanne Stocker and Christian Gruber for many valuable suggestions.

REFERENCES

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.

Andrade,M.A., Brown,N.P., Leroy,C., Hoersch,S., de Daruvar,A., Reich,C., Franchini,A., Tamames,J., Valencia,A., Ouzounis,C. and Sander,C. (1999) Automated genome sequence analysis and annotation. *Bioinformatics*, **15**, 391–412.

Bailey,L.C. Jr, Fischer,S., Schug,J., Crabtree,J., Gibson,M. and Overton,G.C. (1998) GAIA: framework annotation of genomic sequence. *Genome Res.*, **8**, 234–250.

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., 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.

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.

Bork,P., Ouzounis,C., Sander,C., Scharf,M., Schneider,R. and Sonnhammer,E.L.L. (1992) What's in a genome? *Nature*, **358**, 287.

Bork,P. and Bairoch,A. (1996) Go hunting in sequence databases but watch out for the traps. *Trends Genet.*, **12**, 425–427.

Brenner,S.E., Koehl,P. and Levitt,M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, **28**, 254–256.

Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.

Fleischmann,R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.

Frishman,D. and Mewes,H.W. (1997) PEDANTic genome analysis. *Trends Genet.*, **13**, 415–416.

Frishman,D. and Argos,P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–579.

Frishman,D. and Argos,P. (1997) Seventy-five percent accuracy in protein secondary structure prediction. *Proteins*, **27**, 329–335.

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.

Gaasterland,T. and Sensen,C.W. (1996) MAGPIE: automated genome interpretation. *Trends Genet.*, **12**, 76–78.

Galperin,M.Y. and Koonin,E.V. (1998) Sources of systematic errors in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement, and operon disruption. *In Silico Biol.*, **1**, 0007.

Harris,N.L. (1997) Genotator: a workbench for sequence annotation. *Genome Res.*, **7**, 754–762.

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.

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

Houry,W.A., Frishman,D., Eckerskorn,C., Lottspeich,F. and Hartl,F.U. (1999) Identification of *in vivo* substrates of the chaperonin GroEL. *Nature*, **402**, 147–154.

Huang,X., Adams,M.D., Zhou,H. and Kerlavage,A.R. (1997) A tool for analyzing and annotating genomic sequences. *Genomics*, **46**, 37–45.

Iseli,C., Jongeneel,C.V. and Bucher,P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Intell. Syst. Mol. Biol.*, 138–147.

Jukes,T.H. and Osawa,S. (1993) Evolutionary changes in the genetic code. *Comput. Biochem. Physiol.*, **106**B, 489–494.

Klein,P., Kanehisa,M. and DeLisi,C. (1985) The detection and classification of membrane-spanning proteins. *Biochim. Biophys. Acta*, **815**, 468–476.

Kunst,F. *et al.* (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**, 249–256.

Lo,C.L., Ailey,B., Hubbard,T.J., Brenner,S.E., Murzin,A.G. and Chothia,C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257–259.

Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.

Lupas,A.N. and van Stock,J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.

Mayer,K. *et al.* (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature*, **402**, 769–777.

Medigue,C., Rechenmann,F., Danchin,A. and Viari,A. (1999) Imagene: an integrated computer environment for sequence annotation and analysis. *Bioinformatics*, **15**, 2–15.

Mewes,H.W., Albermann,K., Baehr,M., Frishman,D., Gleissner,A., Hani,J., Heumann,K., Kleine,K., Maierl,A., Oliver,S.G., Pfeiffer,F. and Zollner,A. (1997) The yeast genome directory. *Nature*, **387**, 7–65.

Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.

Osawa,S., Jukes,T.H., Watanabe,K. and Muto,A. (1992) Recent evidence for evolution of the genetic code. *Microbiol. Rev.*, **56**, 229–264.

Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA*, **96**, 2896–2901.

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

Ruepp,A., Graml,W., Santos-Martinez,M.-L., Kosetke,K.K., Volker,C., Mewes,H.-W., Frishman,D., Stocker,S., Lupas,A.N. and Baumeister,W. (2000) The genome sequence of the thermoacidophilic scavenger **Thermoplasma acidophilum**. *Nature*, **407**, 508–513.

Saqi,M.A., Wild,D.L. and Hartshorn,M.J. (1999) Protein analyst-a distributed object environment for protein sequence and structure analysis. *Bioinformatics*, **15**, 521–522.

Scharf,M., Schneider,R., Casari,G., Bork,P., Valencia,A., Ouzounis,C. and Sander,C. (1994) GeneQuiz: a workbench for sequence analysis. *Intell. Syst. Mol. Biol.*, **2**, 348–353.

Schäffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) MPALA: matching a protein sequence against a collection of PSI-BLAST—constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.

Sonnhammer,E.L.L. and Durbin,R. (1994) A workbench for large-scale sequence homology analysis. *Comput. Appl. Biosci.*, **10**, 301–307.

Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Terwilliger,T.C., Waldo,G., Peat,T.S., Newman,J.M., Chu,K. and Berendzen,J. (1998) Class-directed structure determination: foundation for a protein structure initiative. *Protein Sci.*, **7**, 1851–1856.

Walker,D.R. and Koonin,E.V. (1997) SEALS: a system for easy analysis of lots of sequences. *Intell. Syst. Mol. Biol.*, **5**, 333–339.

Wheeler,D.L., Chappey,C., Lash,A.E., Leipe,D.D., Madden,T.L., Schuler,G.D., Tatusova,T.A. and Rapp,B.A. (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28**, 10–14.

Wootton,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.