PEDANT genome database: 10 years online

M. Louise Riley¹, Thorsten Schmidt², Irena I. Artamonova¹, Christian Wagner³, Andreas Volz³, Klaus Heumann³, Hans-Werner Mewes¹,² and Dmitrij Frishman¹,²,*

¹Institute for Bioinformatics, GSF-National Research Center for Health and Environment, Ingolstädter Landstrasse 1, 85764 Neuherberg, Germany, ²Department of Genome-oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, 85350 Freising, Germany and ³Biomax Informatics AG, Lochhamer Strasse 9, 82152 Martinsried, Germany

Received September 15, 2006; Accepted October 30, 2006

ABSTRACT
The PEDANT genome database provides exhaustive annotation of 468 genomes by a broad set of bioinformatics algorithms. We describe recent developments of the PEDANT Web server. The all-new Graphical User Interface (GUI) implemented in Java™ allows for more efficient navigation of the genome data, extended search capabilities, user customization and export facilities. The DNA and Protein viewers have been made highly dynamic and customizable. We also provide Web Services to access the entire body of PEDANT data programmatically. Finally, we report on the application of association rule mining for automatic detection of potential annotation errors. PEDANT is freely accessible to academic users at http://pedant.gsf.de.

INTRODUCTION
The PEDANT genome database was first announced in a short note entitled ‘PEDANTic genome analysis’ which appeared in Trends in Genetics in 1997 (1) and reported computational analysis of seven completely sequenced and two partial genomes available at that time. From the very beginning the main mission of PEDANT was defined as filling the gap between manually curated high quality protein sequence databases, such as UniProt/Swiss-Prot (2), and the enormous amounts of other protein sequences produced by genome sequencing projects at an ever increasing pace. Over the past decade the PEDANT genome database was produced by systematically applying an automatic annotation pipeline to genome data released in the public domain. These efforts resulted in one of the most comprehensive curations of sequence databases, such as PEDANT3 genome, the analysis window is opened displaying a list of genes and their best-blast hits. The left-hand panel contains a contig and (iv) genetic element report (where available): information about a selected nonprotein coding genetic element. On selection of a PEDANT3 genome, the analysis window is opened displaying a list of genes and their best-bl...
location, General properties and Export) and lists the analysis results for a particular gene product, which are also available for download in different formats. The DNA and Protein viewers have been completely re-designed. They can be extensively customized and are now based on dynamic panel architecture. The protein viewer displays several panels for various types of evidence that can be configured by selecting ‘preferences’ in the top panel. For example certain panels can be hidden, resized and the color scheme can be changed. The DNA viewer allows for easy navigation along the chromosomes and is capable of displaying an unlimited number of features at different zoom levels.

Any PEDANT3 dataset can be searched using sequence IDs, sequences, PROSITE-like patterns or free text as the query option.

**PEDANT Web Services**

Web Services technology is becoming increasingly popular within the bioinformatics community as a means to exploit the large amounts of data, software programs and computing power available at various institutions (4,5). According to the World Wide Web Consortium (W3C) a Web Service is a software system designed to support interoperable machine-to-machine interactions over a network (http://www.w3.org/2002/ws/). This technology is based on the eXtensible Markup Language (XML) and open standards, and is platform and programming language independent. This enables clients for a particular service to be written in many languages, such as Java or Perl, irrespective of the language the service was written in.

A Web Service has an interface that is described in a machine processable format using the XML based Web Services Description Language (WSDL). WSDL provides a format for the description of a Web Service interface, including parameters and data types in sufficient detail for a programmer to write a client application for that service. Tools are available for various programming languages to generate the required client classes, such as Apache Axis’s WSDL2Java (http://ws.apache.org/axis/java/user-guide.html). The client...
Improving PEDANT annotation quality by data mining

Table 1. Methods available in the Data Retrieval Service

| Methods                        | Description                                                                 |
|--------------------------------|------------------------------------------------------------------------------|
| getDatabases                   | Returns a string array of the pedant2 databases available for data retrieval |
| getProteinSequencesbyDb        | Takes the database name as a parameter and returns all protein codes and sequences in the db as a 2D string array |
| getReportbyDbCodeandContig      | Takes the database name, protein code and contig as parameters and returns the report data as a ReportItem bean |
| areCodesUniqueWithinDb          | Takes the database name as a parameter and returns a Boolean |
| getRawHitsbyDbCodeandMethod     | Takes the database name, protein code and method as parameters and returns the code and raw output for that method as a 2D string array |
| getProteinCodesbyDb             | Takes the database name as a parameter and returns the protein codes as a string array |
| getProteinCodesandContigsbyDb   | Takes the database name as a parameter and returns the protein codes and contigs as a 2D string array |
| getMethodsbyDb                 | Takes the database name as a parameter and returns the available methods as a string array |

programs interact with the Web Service using messages based on the Simple Object Access Protocol (SOAP). As with WSDL, SOAP messages are XML based, permitting the interoperability of Web Services. For the transport layer itself, Web Services typically use the Hypertext Transfer Protocol (HTTP), preventing problems sending the SOAP messages through firewalls.

Bioinformatics users can avoid keeping local copies of databases and software and use a client program instead to access remote databases and software via Web Services. The PEDANT Web Service allows the user to query the database in an automated way from client programs and workflows. We provide a number of data retrieval methods in our Data Retrieval Service (Table 1). For example, to fetch the functional and structural annotations of a particular protein, the client program can call the getReportbyDbCodeandContig method. All these methods are described in the WSDL file: http://mips.gsf.de/webservice/pedant2retrieval/services/DataRetrievalService?wsdl.

Currently this Web Service only retrieves data from PEDANT2 analyses; the PEDANT3 Web Service is in preparation. The raw (unparsed) output from various bioinformatics methods can be obtained using the methods getRawHitsbyDbCodeandMethod and getReportbyDbCodeandContig. For completely sequenced genomes, a protein can be uniquely identified with just the database name and protein code. For unfinished genomes, where the proteins were predicted using Orpheus (6), it is necessary to provide the database name, protein code and contig name to uniquely identify a protein. A list of the bioinformatics methods available for each genome can be obtained by calling the getMethodsbyDb method.

We have generated a java client program using the Apache Axis software (WSDL2Java) which has an example class demonstrating how to make calls to the DataRetrievalService. The jar file is available at http://pedant.gsf.de/webservices.jsp. We have also included PEDANT Web Service client functionality in our PROMPT workbench as a use case to demonstrate the various advantages of the Web Services. PROMPT is a standalone application which enables a user to compare protein sequence sets, revealing statistically significant differences in their annotation features (7); (http://webcu.bio.wzw.tum.de/prompt).

Improving PEDANT annotation quality by data mining

Genome annotation produced by automatic pipelines such as PEDANT is notoriously prone to various kinds of errors (8). While every effort is being made to select the most reliable, carefully benchmarked bioinformatics tools and to avoid spurious similarity hits by setting conservative similarity thresholds, automatically produced function predictions are still not on a par with the results of manual curation of sequence data in high-quality databases such as UniProt (2). We estimate that only 5% of all known proteins have been manually annotated and, given the progress in superfast sequencing technologies (9), it is becoming increasingly clear that the overwhelming majority of sequence data will not be processed by human experts.

We have recently developed a technology to improve the quality of automatically generated annotation using data mining techniques (10). The entire body of annotation available in the PEDANT database can be considered as a collection of records of variable length, one for each gene product, containing functional and structural attributes, such as predicted functional categories, domain assignments and the like. Using a large collection of PEDANT records as input, we apply an unsupervised learning algorithm called association rule mining (11) to derive protein features that occur frequently together. Specifically, implications in the form ‘A&B ⇒ C’ are derived which mean that the majority of proteins possessing features A and B also possess C. Some rules have the strength 1.0 which means that they are always fulfilled, while other rules may be true only in a certain percentage of cases. An exception from a reasonably strong rule will have features A and B, but not C, which may be caused either by over-annotation (features A or B ascribed erroneously), or by under-annotation (feature C missed).

We have shown that ~70% of exceptions from strong association rules found in PEDANT data point to incompatible or missing annotation and thus may be instrumental in identifying annotation errors. However, since strong association rules detect not annotation errors as such but only incompatible feature combinations, it is not possible to automatically correct errors. Instead, we highlight suspicious features and add the features which were putatively missed in a separate section at the bottom of the report page (Figure 1). In a recent pilot project we applied this approach to analyze the PEDANT annotation of 10 model genomes—Arabidopsis thaliana, Aeropyrum pernix, Bacillus subtilis, Escherichia coli, Helicobacter pylori, Mycobacterium tuberculosis, Parachlamydia, Saccharomyces cerevisiae, Synechocystis and Thermoplasma acidophilum—containing a total of 55,123 protein entries. In the course of 2007 we intend on providing rule-based correction for all PEDANT genomes.
ACKNOWLEDGEMENTS

The authors are grateful to Richard Gregory for assistance with the Web Services. Irena Artamonova was supported by the EU Network of Excellence ‘Biosapiens’. Funding to pay the Open Access publication charges for this article was provided by the GSF.

Conflict of interest statement. None declared.

REFERENCES

1. Frishman, D. and Mewes, H.W. (1997) PEDANTic genome analysis. Trends Genet., 13, 415–416.
2. Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. et al. (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res., 33, D154–D159.
3. 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.
4. Wilkinson, M., Schoof, H., Ernst, R. and Haase, D. (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlaNet exemplar case. Plant Physiol., 138, 5–17.
5. Pillai, S., Silventoinen, V., Kallio, K., Senger, M., Sobhany, S., Tate, J., Velankar, S., Golovin, A., Henrick, K., Rice, P. et al. (2005) SOAP-based services provided by the European Bioinformatics Institute. Nucleic Acids Res., 33, W25–W28.
6. 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.
7. Schmidt, T. and Frishman, D. (2006) PROMPT: a protein mapping and comparison tool. BMC Bioinformatics, 7, 331.
8. 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.
9. Metzker, M.L. (2005) Emerging technologies in DNA sequencing. Genome Res., 15, 1767–1776.
10. Artamonova, I., Frishman, G., Gelfand, M.S. and Frishman, D. (2005) Mining sequence annotation databanks for association patterns. Bioinformatics, 21, ii349–ii357.
11. Agrawal, R. and Srikant, R. (1994) Fast algorithms for mining association rules. In Jorge B. Bocca, Matthias Jarke and Carlo Zaniolo, Proceedings of the Twentieth International Conference on Very Large Data Bases, Morgan Kaufmann, pp. 487–499.