BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 8 2005, pages 1743–1744
doi:10.1093/bioinformatics/bti235

Data and text mining

GPSDB: a new database for synonyms expansion of gene and protein names

Violaine Pillet[1], Marc Zehnder[1], Alexander K. Seewald[2], Anne-Lise Veuthey[1] and Johann Petrak[2,*]

[1]Swiss Institute of Bioinformatics, CMU—Rue Michel-Servet 1, 1211 Geneva 4, Switzerland and [2]Austrian Research Institute for Artificial Intelligence, Freyung 6/6, A-1010 Vienna, Austria

Received on November 1, 2004; accepted on December 15, 2004
Advance Access publication December 21, 2004

ABSTRACT

Summary: We present a new database, GPSDB (Gene and Protein Synonyms DataBase) which collects gene/protein names, in a species specific way, from 14 main biological resources. A web-based search interface gives access to the database: given a gene/protein name, it retrieves all synonyms for this entity and queries Medline with a set of user-selected terms.

Availability: GPSDB is freely available from http://biomint.oefai.at/

Contact: johann@oefai.at

INTRODUCTION

Although guidelines exist for naming gene and protein entities, many authors describe the latter in scientific texts using their own terms. Furthermore, before such nomenclatures existed, authors could freely choose the names for the genes and proteins they were studying. As a result, there may be numerous ways (full name, symbol, synonym) of describing the same entity. For instance, almost 30 different terms are assigned to the *antennapedia* gene in *Drosophila melanogaster*. Moreover, an identical term may relate to two or more separate entities within a single species, or among different species—this is the problem of homonymy. For example, the *ACS3* term simultaneously designates the human *FACL3* and *twist* genes. In this paper, we describe a new resource—GPSDB (Gene and Protein Synonyms DataBase)—which enables an easy navigation through the jungle of gene/protein names.

GPSDB was constructed using the main current biological resources of gene/protein names. The database is accessible through a web interface which, given a gene/protein name, retrieves a list of synonyms for this entity and queries Medline through PubMed. This enables the recovery of a maximum of publications describing a particular gene/protein. GPSDB was created in the framework of BioMinT (www.biomint.org), a European project that aims to develop a generic text-mining tool that will—in a semi-automatic way—assist Swiss-Prot (Bairoch *et al*., 2004) and PRINTS (Attwood *et al*., 2004) curators in their protein annotation activity. A similar gene/protein name resource—GENA—has also been provided for the development of a dictionary-based name recognition tool (Koike and Takagi, 2004), but this resource is limited to eukaryotic model organisms.

*To whom correspondence should be addressed.

ARCHITECTURE

The first step for constructing GPSDB consisted in identifying the main resources where gene and protein names were available. In order to populate the database, 14 such resources were used: LocusLink and Swiss-Prot for multispecies; GDB, HUGO and OMIM for Human; MGD for Mouse; RGD and Ratmap for Rat; Flybase for *Drosophila*; SGD for *Saccharomyces cerevisiae*; TAIR for *Arabidopsis thaliana*; WormBase for *Caenorhabditis elegans*; SubtiList for *Bacillus subtilis*; and EcoGene for *Escherichia coli*. From each database, specific fields were extracted (official name, symbol name, synonyms, database cross-reference links, species name, entry ID, etc.).

In order to retrieve a complete list of synonyms for a given gene/protein, all entries from the databases above relating to a same entity were merged. The identification and connection of such entries were achieved by making use of the (transitive and symmetric) database cross-references that link these entries together. Moreover, this procedure makes the distinction between homonyms possible.

Database entries, corresponding to pseudogenes or non-protein encoding genes, were ignored since our focus is on proteins and protein-encoding genes. On the other hand, some terms present in these databases but scarcely mentioned in the literature, like accession numbers resulting from various sequencing project (e.g. KIAA cDNA clones), were also discarded. Similarly, terms consisting of one letter or only digits were excluded because of their irrelevance for searching Medline. Finally, the content of some entries was modified: additional information such as comments or special characters were removed. Regular expressions were used for this cleaning-up process.

The resulting database contains 532 970 different synonyms describing 319 386 protein entities. The total number of species/subspecies taken into account exceeds 7000. GPSDB is updated every three months.

QUERYING GPSDB

An interface provides several options for querying GPSDB (Fig. 1). The user begins the search using a gene/protein name or a string of characters. Wildcard expressions are allowed. For example, if the query is '*hydrolase*' each entry containing this string will be retrieved. To limit the query, the user can specify one/several taxonomic ranges (e.g. Eukaryota, Mammalia, Viridiplantae),

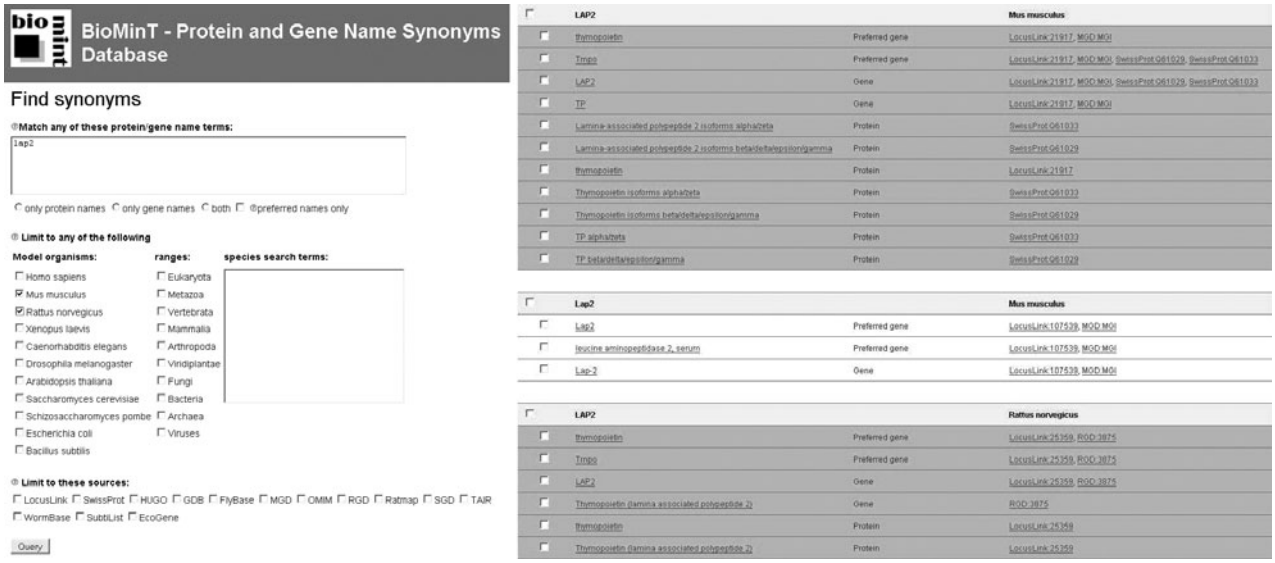

Fig. 1. GPSDB query interface and query output example.

one/several model organisms (e.g. *Homo sapiens*, *Mus musculus*, *A.thaliana*), or enter one/several species names not mentioned in the list. In addition, the query may be restrained to one/several source databases.

The resulting output presented in Figure 1 is a list of synonyms sorted by matching name, species and gene/protein entity in case of homonymy. For each synonym listed, the source database is mentioned with a direct link to the corresponding database entry. Clicking on a synonym retrieves PubMed statistics, namely how many documents are retrieved when searching Medline with this term, and thus gives an indication of its relevance for the query.

At this stage, the user can choose one/several/all synonyms to formulate a PubMed query. Alternatively, additional terms such as species name or other search words may be incorporated into the query. The latter is directly sent to the PubMed search engine, which then displays all the documents found.

FURTHER PROSPECT

We plan to enhance the coverage of GPSDB by adding new organism specific databases, as well as synonyms of family and domain names. We also envisage including ranking algorithms to sort the retrieved documents by species or information topics (protein function, subcellular location, associated disease, etc.). A preliminary study on these algorithms has been published elsewhere (Seewald, 2004).

ACKNOWLEDGEMENTS

The BioMinT project is funded by the European Commission, contract-no. QLRI-CT-2002-02770 under the RTD programme ‘Quality of Life and Management of Living Resources’.

REFERENCES

Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P., Uddin,A. and Zygouri,C. (2003) PRINTS and its automatic supplement, prePRINTS, *Nucleic Acids Res.*, **31**, 400–402.

Bairoch,A., Boeckmann,B., Ferro,S. and Gasteiger,E. (2004) Swiss-Prot: juggling between evolution and stability, *Brief. Bioinform.*, **5**, 39–55.

Koike,A. and Takagi,T. (2004) Gene/protein/family name recognition in biomedical literature. In *BioLINK: Linking Biological Literature, Ontologies, and Databases*, Boston, MA, pp. 9–16.

Seewald,A.K. (2004) Ranking for BioMinT: investigating performance, local search and homonymy recognition. *Proceedings of the Symposium on Knowledge Exploration in Life Science Informatics (KELSI 2004)* Milano, Italy *LNCS* **3303**, Springer Science+Business Media, Berlin.