BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 1 2006, pages 103–105
doi:10.1093/bioinformatics/bti749

Data and text mining

BioThesaurus: a web-based thesaurus of protein and gene names

Hongfang Liu[1], Zhang-Zhi Hu[2], Jian Zhang[2] and Cathy Wu[2]

[1]Department of Information Systems, University of Maryland at Baltimore County, 1000 Hilltop Circle, MD 21250, USA and [2]Department of Biochemistry and Molecular Biology, Georgetown University Medical Center, Washington DC, USA

Received on May 31, 2005; revised on October 24, 2005; accepted on October 27, 2005
Advance Access publication November 2, 2005
Associate Editor: Alex Bateman

ABSTRACT

BioThesaurus is a web-based system designed to map a comprehensive collection of protein and gene names to protein entries in the UniProt Knowledgebase. Currently covering more than two million proteins, BioThesaurus consists of over 2.8 million names extracted from multiple molecular biological databases according to the database cross-references in iProClass. The BioThesaurus web site allows the retrieval of synonymous names of given protein entries and the identification of protein entries sharing the same names.

Availability: BioThesaurus is accessible for online searching at http://pir.georgetown.edu/iprolink/biothesaurus

Contact: hfliu@umbc.edu

Supplementary information: Supplementary data are accessible at http://pir.georgetown.edu/iprolink/biothesaurus/supplement/

INTRODUCTION

The accelerated expansion of biological research, reduction in computing costs and widespread access of the Internet have created a diverse and overwhelming volume of knowledge stored in various databases. One such database is the UniProt Knowledgebase (UniProtKB) (Bairoch *et al*., 2005), a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL and PIR-PSD (Wu *et al*., 2003). Systems that integrate and provide access to information stored in these various databases can significantly assist researchers in the interpretation of experimental data and in the discovery of new knowledge (Sujansky, 2001; Bry and Kröger, 2003). The iProClass database (Wu *et al*., 2004) provides comprehensive descriptions of all UniProtKB proteins with rich links to over 90 molecular databases and serves as a framework for data integration in a distributed networking environment.

Recently, natural language processing (NLP) techniques have been explored to facilitate sequence annotation and improve the quality of biological databases through effective text mining (Hirschman *et al*., 2002; Shatkay and Feldman, 2003). One prerequisite for knowledge extraction from scientific literature is to accurately recognize and map biological entity names in free text to corresponding entries in biological databases. One important component of such a task is a named entity dictionary that maps between names and records in the databases. Several groups have developed terminological resources for genes and proteins. GENA (Koike *et al*., 2003) automatically gathers official gene symbols, official full names and synonyms from several databases, such as Swiss-Prot, FlyBase (Drysdale and Crosby, 2005) and MGD (Eppig *et al*., 2005). ProMiner (Hanisch *et al*., 2003) extracted gene symbols, alias names and full names from HUGO (http://www.gene.ucl.ac.uk/nomenclature), Swiss-Prot and TrEMBL. Both GENA and ProMiner were developed primarily for biological named entity tagging systems and names were not directly linked to records in biological databases.

Mapping protein and gene names to the corresponding entries in UniProtKB is important for researchers to explore rich information stored in UniProtKB as well as iProClass. Additionally, it is a necessary component for using NLP techniques to facilitate protein annotation and to improve the quality of the databases. In this paper, we present a web-based system BioThesaurus that maps a thesaurus of protein and gene names extracted from multiple molecular biological databases to all known protein sequences. With its comprehensiveness and database association, BioThesaurus can be used for many applications: (1) dictionary-based biological named entity tagging of text; (2) literature mining by query expansion using synonymous names; (3) name mapping service for searching synonymous names and resolving name ambiguities and (4) name standardization and nomenclature using more broadly accepted names, and by selecting abbreviations or gene symbols that do not conflict with names of other entities.

CONSTRUCTION OF BIOTHESAURUS

An overview of BioThesaurus construction is shown in Figure 1. The thesaurus was designed to provide comprehensive protein and gene names for all protein entries in UniProtKB. The underlying knowledgebase used by BioThesaurus was extracted from multiple online resources based on the cross-references provided by iProClass (Wu, 2004). A total of 13 underlying data sources was used to construct BioThesaurus: (1) protein databases maintained by UniProt, including Swiss-Prot, TrEMBL and PIR-PSD, (2) gene and protein resources at NCBI (Wheeler *et al*., 2005), including Entrez Gene, RefSeq and GenPept (GenBank translation), (3) genome databases of model organisms, such as MGD (Eppig *et al*., 2005), SGD (Christie *et al*., 2004), RGD (de la Cruz *et al*., 2005), FlyBase (Drysdale and Crosby, 2005), and WormBase (Chen *et al*., 2005), and (4) a few other databases, such as the HUGO human gene nomenclature database, the EC enzyme nomenclature (Gegenheimer, 2000; Tipton and Boyce, 2000), and the OMIM

*To whom correspondence should be addressed.

database of human genes and genetic disorders. Table 1 summarizes the annotation fields and number of names from each data source. The primary sources are UniProtKB and Entrez Gene. Names are also extracted from PIR-PSD because PSD names are not incorporated into UniProtKB. Due to redundant information in NCBI's Entrez Gene, RefSeq, and GenPept and their various degrees of annotation quality, only the 'definition' field of RefSeq is used and only unique names in the GenPept 'Features' not already in Entrez Gene or RefSeq are used.

Software was developed to automatically gather names from the underlying sources and parse individual names from annotation fields that contain multiple names separated by parentheses or other delimiters such as semicolons or commas. A raw thesaurus was then compiled, associating names with the corresponding UniProtKB entries. The raw thesaurus was further filtered to remove highly ambiguous and nonsensical names. The 'name filter' (Table 2) was compiled based on frequency counts of names in UniProtKB entries and by curator judgment as 'nonsensical.' Examples of filtered names include novel protein, fragment, and hypothetical protein. We also mapped names in the thesaurus to names in the UMLS (Unified Medical Language System) (Bodenreider, 2004), a biomedical knowledge source that is popularly used in the medical domain and includes proteins and their related concepts (Table 3). Finally, names were grouped to include textual variants caused by case difference (e.g. MIG-5 versus mig-5), punctuation (e.g. TIMP3 versus TIMP-3), or syntactic variants (tissue inhibitor of metalloproteinase 3 versus tissue inhibitor of metalloproteinases 3). One name from a group was chosen for display and a frequency count for each group was calculated, totaling the number of distinct source databases (Table 1) from which these names were derived, such that a source with more than one text variant is counted only once. The count indicates the relative 'popularity' of names in a synonymous list for a given UniProtKB entry, and may suggest more broadly adopted or official names and reveal potential misnomers with incorrect annotations.

BIOTHESAURUS WEB SITE

The BioThesaurus currently (release 1.0, August 01, 2005) consists of over 2.8 million distinct protein and gene names, or 2.3 million names after combining text variants, covering over 2.0 millions of UniProtKB entries (Table 1). It will be automatically updated monthly to maintain consistency with the underlying data sources. The BioThesaurus is accessible for online searching at http://pir.georgetown.edu/iprolink/biothesaurus. The web interface supports search options for (1) report retrieval using UniProtKB identifier (accession number or ID), and (2) record retrieval using text searches for three fields protein/gene name, sequence unique identifier and source organism. To facilitate fast retrieval, we provide multiple indexes to the thesaurus, including indexing of all textual variants using a text indexing system as described in (Huang *et al*., 2004).

The BioThesaurus report for a given UniProtKB protein entry can be retrieved using the web search form or, alternatively, using a URL search string with UniProtKB identifier, as in http://pir.georgetown.edu/cgi-bin/biothesaurus_search.pl?id=P48032. The report lists all protein and genes names with hypertext links to the underlying database entries, together with frequency counts and a brief summary of protein information. The hypertext link on each name provides online text search for all UniProtKB entries sharing the same name. The synonymous names in the report can be used for query expansion to uncover literature citations that may otherwise be missed. For example, PubMed search using 'tissue inhibitor of metalloproteinase 3' returned 432 citations, using *TIMP3* returned 131 citations, while searching using both names returned 509 citations.

Text search allows retrieving BioThesaurus record based on protein or gene names, sequence unique identifier in each of the underlying data sources and source organism name. The results are displayed in a summary table with information on UniProtKB identifier, UniProtKB protein name, source organism, protein classification and the matched field, along with hypertext links to full BioThesaurus reports. The result of text search reveals interesting relationships among entities sharing the same name. The protein classification and source organism information can help distinguish different types of name ambiguities, such as the same name resulting from protein homology in different organisms and name overloading from gene symbols that represent different proteins.

CONCLUSION

BioThesaurus is a full-scale thesaurus that maps an extensive collection of names to all known proteins in UniProtKB. BioThesaurus can be used by researchers to recognize protein names for protein data exploration and query expansion and to seek protein name standardization. With its comprehensiveness and database association, it can also be used by automated applications that require the mapping between names and UniProtKB records, for example, information extraction or information retrieval systems.

ACKNOWLEDGEMENTS

The project was supported by grant IIS-0430743 from the National Science Foundation and in part by grant U01-HG02712 from the National Institutes of Health (for UniProt) and grant DBI-0138188 from the National Science Foundation (for iProClass).

REFERENCES

Bairoch,A. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33** (Database issue), D154–D159.

Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*., **32** (Database issue), D267–D270.

Bry,F.K. and Kröger,P. (2003) a computational biology database digest: data, data analysis, and data management. *Distributed Parallel Databases*, **13**, 7–42.

Chen,N. *et al*. (2005) WormBase: a comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Res*., **33** (Database issue), D383–D389.

Christie,K.R. *et al*. (2004) *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces* cerevisiae and related sequences from other organisms. *Nucleic Acids Res*., **32** (Database issue), D311–D314.

de la Cruz,N. *et al*. (2005) The Rat Genome Database (RGD): developments towards a phenome database. *Nucleic Acids Res*., **33** (Database issue), D485–D491.

Drysdale,R.A. and Crosby,M.A. (2005) FlyBase: genes and gene models. *Nucleic Acids Res*., **33** (Database issue), D390–D395.

Eppig,J.T. *et al*. (2005) The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucleic Acids Res*., **33** (Database issue), D471–D475.

Gegenheimer,P. (2000) Enzyme nomenclature: functional or structural? *RNA*, **6**, 1695–1697.

Hanisch,D. *et al*. (2003) Playing biology's name game: identifying protein names in scientific text. *Pac. Symp. Biocomput*., 403–414.

Hirschman,L. *et al*. (2002) Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, **18**, 1553–1561.

Huang,H. *et al*. (2004) The PIR integrated protein databases and data retrieval system. *Data Sci. J*., **3**, 163–174.

Koike,A. *et al*. (2003) Kinase pathway database: an integrated protein-kinase and NLP-based protein-interaction resource. *Genome Res*., **13**, 1231–1243.

Shatkay,H., Feldman,R. *et al*. (2003) Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol*., **10**, 821–855.

Sujansky,W. (2001) Heterogeneous database integration in biomedicine. *J. Biomed. Inform*., **34**, 285–298.

Tipton,K. and Boyce,S. (2000) History of the enzyme nomenclature system. *Bioinformatics*, **16**, 34–40.

Wheeler,D.L. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33** (Database issue), D39–D45.

Wu,C.H. *et al*. (2003) The Protein Information Resource. *Nucleic Acids Res*., **31**, 345–347.

Wu,C.H. *et al*. (2004) The iProClass integrated database for protein functional analysis. *Comput. Biol. Chem*., **28**, 87–96.