BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 8 2011, pages 1190–1191
doi:10.1093/bioinformatics/btr101

Databases and ontologies

A comprehensive protein-centric ID mapping service for molecular data integration

Hongzhan Huang[1,†,*], Peter B. McGarvey[2,†], Baris E. Suzek[2], Raja Mazumder[2], Jian Zhang[2], Yongxing Chen[1] and Cathy H. Wu[1,2]

[1]Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE 19711 and [2]Protein Information Resource, Georgetown University Medical Center, Washington, DC 20007, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Identifier (ID) mapping establishes links between various biological databases and is an essential first step for molecular data integration and functional annotation. ID mapping allows diverse molecular data on genes and proteins to be combined and mapped to functional pathways and ontologies. We have developed comprehensive protein-centric ID mapping services providing mappings for 90 IDs derived from databases on genes, proteins, pathways, diseases, structures, protein families, protein interaction, literature, ontologies, etc. The services are widely used and have been regularly updated since 2006.

Availability: www.uniprot.org/mapping and proteininformation-resource.org/pirwww/search/idmapping.shtml

Contact: huang@dbi.udel.edu

Received on January 14, 2011; revised on February 10, 2011; accepted on February 17, 2011

1 INTRODUCTION

A central difficulty in integrating and comparing molecular data is finding and maintaining correspondences of identifiers (IDs) for genes, proteins and higher level functional groupings like pathways and Gene Ontology (GO) terms. Various research groups and data repositories use different data sources, IDs and names for nucleotide and proteins sequences in their analysis. The first step to integrate diverse datasets is to map the data to a common set of biological objects. In our experience working with research labs, we are often presented with 3–4 different types of database IDs for the same types of biological object, whether proteins or genes. In addition, many of these IDs are redundant in that several actually point to the same gene/protein.

Several groups have developed tools to address this common problem. Many take a gene-centric approach useful for genomic and microarray work (Berriz and Roth, 2008; Bussey *et al.*, 2003; Diehn *et al.*, 2003; Draghici *et al.*, 2006; Huang Da *et al.*, 2008; Waegele *et al.*, 2009). Others take a more protein-centric approach (Cote *et al.*, 2007; Mudunuri *et al.*, 2009), which is more useful for proteomics work, and also can be useful for systems biology as proteins are often the active biological object that interacts with other objects and pathways.

We have developed comprehensive protein-centric ID mapping tools and services built on the iProClass protein data warehouse (Huang *et al.*, 2003), which provides mappings between UniProtKB (UniProt Consortium, 2009) and currently supports IDs from 90 data sources. The services allow individual and batch mappings as well as programmatic access and FTP downloads. The mappings are updated monthly.

2 METHODS

The data sources for ID mapping consist of UniProKB cross-references and additional sequence and cross-reference information from the iProClass database, which contains information derived from over 100 data sources in records for over 14 million protein sequences. A complete list of data sources is found at http://proteininformationresource.org/cgi-bin/iproclass_stat. ID mappings can be one of the three types: (i) between similar biological objects, for example, an NCBI GI number and UniProtKB accession for a protein; (ii) between related biological objects Gene > mRNA > Protein, for example, UniProt Accession to ENSEMBL gene; and (iii) mapping from objects to their properties, such as a GenBank/EMBL/DDBJ accession to a KEGG pathway ID or GO ID. Combining all three types provides a powerful network from which biological knowledge can be extracted. Because of data source heterogeneity, mapping between database IDs can be complex. The basic approaches we use for establishing a relationship are as follows: (i) use database cross-references from well-curated databases, such as UniProtKB, which includes cross-references in each protein entry. Most of the cross-references in UniProtKB are actively maintained via collaborations between UniProt and the cross-referenced databases to ensure quality. (ii) Use other database IDs as a bridge (transitive mappings). For example, one can use a GenBank/EMBL/DDBJ accession referenced by UniProtKB and also by an NCBI GI number to make a correspondence. (iii) Use sequence identity to establish the relationship between two database IDs. Usually, a 100% sequence identity for the same taxon is required for our mapping.

The system is built upon the iProClass protein-centric data warehouse. IDs are mapped to UniProtKB or UniParc accession numbers using the approaches outlined above. Database tables are built for storing the ID mapping data and Apache Lucene indexes are generated for database search and retrieval. We have two implementations for the public: the ID mapping services on the UniProt web site (Jain *et al.*, 2009) and on the PIR web site (proteininformationresource.org). The system design is illustrated in Figure 1. We currently update our data warehouse and the ID mapping database every 4 weeks in conjunction with UniProtKB.

3 RESULTS AND CONCLUSIONS

In December 2010, there were over 125 million IDs mapped to over 14 million UniProtKB and UniParc protein accessions. The set

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

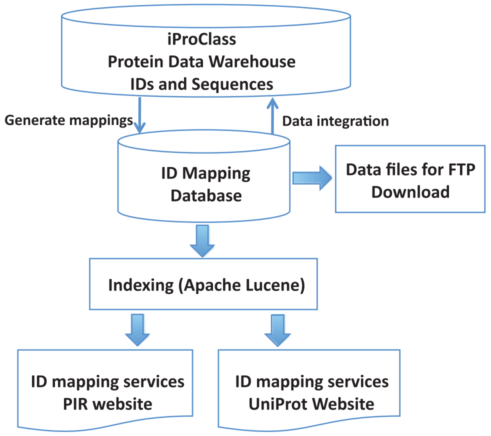

Fig. 1. System design.

can produce a very much larger combinatorial network of IDs pairs using UniProt accessions as a bridge. The mapping services currently support 90 types of IDs. All numbers will continue to grow.

Web-based ID mapping: this is available from both the UniProt and PIR web sites at http://uniprot.org/mapping and http://proteininformationresource.org/pirwww/search/idmapping.shtml, respectively. The two sites draw from identical mapping tables and provide batch mappings. The UniProt site provides UniProtKB centric mapping (i.e. one of the pairs must be a UniProt accession), while the PIR site allows transitive mappings between all IDs bridged by UniProtKB. For details on the options and output formats available, see the online help at each web site.

Programmatic access: this is provided on the UniProt web site via a REST style web service, for help and example code see http://www.uniprot.org/faq/28. On the PIR web site, help is at http://proteininformationresource.org/cgi-bin/idmapping_http_client.

FTP downloads: FTP downloads of ID mapping tables are available from ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping. The FTP directory contains two ID mapping data files, the idmapping.dat file with tab-delimited columns for: UniProtKB accession, ID type and ID and an idmapping_selected.tab file contains IDs that are most often requested by users. See the README file for more information.

The ID mapping services have been used extensively since 2006. The FTP files are downloaded on average 15 000 times a year from about 4500 unique IP addresses a year. Analysis of web usage shows 897 000 uses during the last year, an average of 74 700 a month. User logs suggest that most users are mapping small sets of accession/ID (including single proteins) to UniProtKB. A lesser number of power users map thousands at a time or download from the FTP site. Currently, the most popular mapping pair is from Entrez Gene ID to UniProtKB accession. In addition, other sites including NCBI and a few other ID mapping services redistribute these mappings from their sites and services.

The main strengths of our services compared to others are as follows: a large number of IDs are supported; all organisms are supported; frequent updates; mappings are available for download and well as from the web site; and, we interact directly with data provided through the UniProt consortium to ensure accuracy.

Selected published examples of use cases for the service include mapping of NCBI GI numbers to UniProtKB accession/IDs and finally to GO terms to identify functionally linked proteins (Cokus *et al.*, 2007); functional mapping of mass spectrometry data to pathways (Park *et al.*, 2008) and integration of omics data for pathogen–host interactions (McGarvey *et al.*, 2009).

ACKNOWLEDGEMENTS

The Authors would like to thank the UniProt Consortium for their help and advice in creating and evaluating these services.

Funding: National Human Genome Research Institute:National Institutes of Health (1U41HG006104-01).

Conflict of Interest: none declared.

REFERENCES

Berriz,G.F. and Roth,F.P. (2008) The Synergizer service for translating gene, protein and other biological identifiers. *Bioinformatics*, **24**, 2272–2273.

Bussey,K.J. *et al.* (2003) MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol.*, **4**, R27.

Cokus,S. *et al.* (2007) An improved method for identifying functionally linked proteins using phylogenetic profiles. *BMC Bioinformatics*, **8** (Suppl. 4), S7.

Cote,R.G. *et al.* (2007) The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics*, **8**, 401.

Diehn,M. *et al.* (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.*, **31**, 219–223.

Draghici,S. *et al.* (2006) Babel's tower revisited: a universal resource for cross-referencing across annotation databases. *Bioinformatics*, **22**, 2934–2939.

Huang Da,W. *et al.* (2008) DAVID gene ID conversion tool. *Bioinformation* **2**, 428–430.

Huang,H. *et al.* (2003) iProClass: an integrated database of protein family, function and structure information. *Nucleic Acids Res.*, **31**, 390–392.

Jain,E. *et al.* (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics*, **10**, 136.

McGarvey,P.B. *et al.* (2009) Systems integration of biodefense omics data for analysis of pathogen-host interactions and identification of potential targets. *PLoS ONE*, **4**, e7162.

Mudunuri,U. *et al.* (2009) bioDBnet: the biological database network. *Bioinformatics*, **25**, 555–556.

Park,D. *et al.* (2008) MassNet: a functional annotation service for protein mass spectrometry data. *Nucleic Acids Res.*, **36**, W491–W495.

UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.

Waegele,B. *et al.* (2009) CRONOS: the cross-reference navigation server. *Bioinformatics*, **25**, 141–143.