D122–D124 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh109

DBSubLoc: database of protein subcellular localization

Tao Guo, Sujun Hua, Xinglai Ji and Zhirong Sun*

Institute of Bioinformatics, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing 100084, China

Received August 14, 2003; Revised September 29, 2003; Accepted October 14, 2003

ABSTRACT

We have built a protein subcellular localization annotation database, the DBSubLoc database, which is available at http://www.bioinfo.tsinghua.edu.cn/dbsubloc.html. Annotations were taken from primary protein databases, model organism genome projects and literature texts, and then were analyzed to dig out the subcellular localization features of the proteins. The proteins are also classified into different categories. Based on sequence alignment, non-redundant subsets of the database have been built, which may provide useful information for subcellular localization prediction. The database now contains >60 000 protein sequences including ~30 000 protein sequences in the non-redundant data sets. Online download, search and Blast tools are also available.

INTRODUCTION

Subcellular localization is one of the key features of a protein, since it is closely related to biological function (1). During translation or later, proteins will be transported into different regions such as cytoplasm, membrane system, nuclear region, mitochondrion, etc., or may be secreted out of the cell. As high-throughput genome sequencing projects have produced an enormous amount of raw protein sequence data, it is very important to annotate their functional features, including subcellular localization.

Most known protein subcellular localizations are determined by experimental methods and some others can be obtained based on very high sequence similarities. Now some bioinformatics methods have been developed to predict the subcellular location of proteins, which make use of either the sorting signals (2), or amino acid composition in the sequences (3–9). There are two factors that have an effect on the prediction. One is the number of protein sequences for training in the artificial intelligence method; and the other is the number of target subcellular locations covered in the data set, which determine the capability of the prediction method.

We have built the DBSubLoc database to collect and manage information related to subcellular localization, and to make it into an integrated platform to improve both the amount and the quality of the data sets, which may provide useful information for prediction methods, and also for research into functional relations of proteins. Sequence analyses were also performed to produce high-quality non-redundant sub-datasets.

The DBSubLoc database has been built on annotations from primary protein sequence databases: Swiss-Prot/TrEMBL (10) and the Protein Information Resource (PIR) (11). Annotations were also provided by model organism genome projects such as the SGD (*Saccharomyces cerevisiae*) (12), TAIR (*Arabidopsis thaliana*) (13), FlyBase (*Drosophila melanogaster*) (14) and MGD (15) (*Mus musculus*). We selected only full-length and unambiguous proteins to build the DBSubLoc database. Repetitive sequences and short sequences of <20 amino acids were excluded. In the selected sequences, there are two types of subcellular location annotations. One type is annotated in natural language that is easy for humans to understand, but hard for programs to process, and the other is cross-referenced to the Gene Ontology (GO) term database (16), which is ideal for further processing. Most of the text annotations in natural language are converted to certain GO cellular component by automatic keyword recognition or manual identification. For other annotations that give very complex descriptions to the subcellular localization features, or describe proteins that are localized into multiple cellular components, their GO cross-references are determined manually. Some subcellular localization features are not determined by experimental methods directly; these low-quality annotations, marked with 'by similarity', 'probable' or 'potential', are also collected into the DBSubLoc database, but most of them are eliminated in the non-redundant data sets. Therefore, in the DBSubLoc, each protein entry is cross-referenced to at least one GO cellular component term that indicates its subcellular location. The proteins annotated in model organism genomes are cross-referenced to the NCBI Taxonomy database (http://www.ncbi.nlm.nih.gov/Taxonomy/tax.html), and are also categorized based on their taxonomy class (i.e. virus, archaea, bacteria, eukaryote, etc.)

Based on the DBSubLoc database, we have also provided subsets of the database that are composed of non-redundant protein sequences for each taxonomy class. Using Blast, protein sequences were compared with each other and grouped according to their sequence similarity. In each non-redundant subset, the similarities between two protein entries are <60%.

*To whom correspondence should be addressed. Tel/Fax: +86 10 62772237; Email: sunzhr@mail.tsinghua.edu.cn

Table 1. Brief statistical information of the full DBSubLoc database

		Eukaryotes						
Full data set	Bacteria	Total	Fungi	Plants	Animals	Viruses	Archaea	Total
Total	19664	38275	5608	5468	25192	3472	2640	64051
Nucleus	0	5444	971	412	3995	170	168	5782
Cytoplasm	3254	2463	423	113	920	0	417	6134
Membrane	6604	12877	1599	1329	9479	972	654	21107
Extracellular	936	3554	126	229	3173	15	0	4505
Mitochondrion	0	1726	393	211	1048	0	0	1726
Chloroplast	0	1789	0	1293	0	0	0	1789
Ribosome	2625	1000	271	257	365	0	788	4413
Other	6245	9422	1825	1624	6212	2315	613	18595

The number of entries is listed in each cell.

Table 2. Brief statistical information of the non-redundant DBSubLoc database

		Eukaryotes						
NR data set	Bacteria	Total	Fungi	Plants	Animals	Viruses	Archaea	Total
Total	9801	18143	4201	1367	12575	1182	1231	30357
Nucleus	0	3049	802	181	2066	73	83	3205
Cytoplasm	1446	1203	327	25	629	0	186	2835
Membrane	4230	6223	1247	304	4517	363	412	11228
Extracellular	508	1281	61	76	1096	10	0	1799
Mitochondrion	0	963	291	91	581	0	0	963
Chloroplast	0	331	0	331	0	0	0	331
Ribosome	642	376	116	78	182	0	287	1305
Other	2975	4717	1357	281	3504	736	263	8691

The number of entries is listed in each cell.

DATABASE CONTENT

In the DBSubLoc database, each entry is composed of several records that describe one protein. Each entry contains the following information: the unique integer identity of the entry in the database; the name and text description of the protein; the taxonomic name of its source organism; the text annotations of its subcellular location; the amino acid sequence and cross-references to another database. Each cross-reference record indicates one link to other databases including Swiss-Prot, the GO term database, the NCBI Taxonomy database, PubMed, etc. The cross-reference record provides the referenced database name and the unique identifier in that database. As the database grows, more cross-references will be appended to existing entries.

The DBSubLoc database and the non-redundant sub-datasets are released as plain text file. The format is similar to that of a Swiss-Prot data file. Each line in the file is one record of an entry in the 'KEY VALUE' format, for example, the 'ID 10000001' record means that the unique identity of this entry is 10000001. The cross-reference records begin with a 'CX' key, each of the value data contain one cross-reference record in the 'Reference Database: Reference ID' format, for example, the 'CX GO: 0005737' record means that the protein entry is linked to the GO term database 0005737 entry. The sequence of the protein may be spread into several records with the 'SQ' key. A detailed description of the format can be found on the web page. Tables 1 and 2 give brief statistical information on full data set and non-redundant data sets. The number of entries are listed in the tables. More statistical information is available at the website.

DATABASE ACCESS

We provide free access to the DBSubLoc database for education and research users. The website is available at http://www.bioinfo.tsinghua.edu.cn/dbsubloc.html. Users can download the database release file or smaller taxonomy-categorized files. Users can also search the database with protein name, protein identity or cross-referenced database identity. An online sequence alignment service is also provided at the website. Users can submit one protein sequence to search for homologous sequences in the complete DBSubLoc database or in one of its non-redundant subsets. With the development of the DBSubLoc database, new database releases, more services and tools will be provided.

FUTURE DEVELOPMENT

We aim to collect more data and information, and classify and purify them into an efficient relational data model for further research. The non-redundant data sets are to be tested in developing new prediction methods. Because of the complexity of cellular components, more work is needed to make data set purification and partition better. More annotations will be appended from the literature, other database and by prediction.

ACKNOWLEDGEMENTS

This work was funded by the National Key Foundational Research Grant in China (863-2002AA23/03/, 2002AA234041, 973-2003CB715900, NSFC-90303017).

REFERENCES

1. Eisenhaber,F. and Bork,P. (1998) Wanted: subcellular localization of proteins based on sequence. *Trans. Cell Biol.*, **8**, 169–170, 911.
2. Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.*, **8**, 581–599.
3. Gardy,J.L., Spencer,C., Wang,K., Ester,M., Tusnady,G.E., Simon,I., Hua,S., deFays,K., Lambert,C., Nakai,K. and Brinkman,F.S. (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res.*, **31**, 3613–3617.
4. Hua,S. and Zhirong,S. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.
5. Nakashima,H. and Nishikawa,K. (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.*, **238**, 54–61.
6. Reinhardt,A. and Hubbard,T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.*, **26**, 2230–2236.
7. Yuan,Z. (1999) Prediction of protein subcellular locations using Markov chain models. *FEBS Lett.*, **451**, 23–26.
8. Feng,Z.P. (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. *Biopolymers*, **58**, 491–499.
9. Feng,Z.P. (2002) An overview on predicting the subcellular location of a protein. *In Silico Biol.*, **2**, 291–303.
10. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
11. Wu,C.H., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z.-Z., Ledley,R.S., Lewis,K.C., Mewes,H.-W., Orcutt,B.C. *et al.* (2002) The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res.*, **30**, 35–37.
12. Dwight,S.S., Harris,M.A., Dolinski,K., Ball,C.A., Binkley,G., Christie,K.R., Fisk,D.G., Issel-Tarver,L., Schroeder,M., Sherlock,G. *et al.* (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.
13. Huala,E., Dickerman,A., Garcia-Hernandez,M., Weems,D., Reiser,L., LaFond,F., Hanley,D., Kiphart,D., Zhuang,J., Huang,W. *et al.* (2001) The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis and visualization system for a model plant. *Nucleic Acids Res.*, **29**, 102–105.
14. The FlyBase Consortium (2003) The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.*, **31**, 172–175.
15. Blake,J.A., Richardson,J.E., Bult,C.J., Kadin,J.A., Eppig,J.T. and the members of the Mouse Genome Database Group. 2003. MGD: The Mouse Genome Database. *Nucleic Acids Res.*, **31**, 193–195.
16. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.