Bioinformatics Explained

Bioinformatics explained: Biological Databases

February 12, 2008

CLC bio
Gustav Wieds Vej 10 · 8000 Aarhus C · Denmark
Telephone: +45 70 22 55 09 · Fax: +45 70 22 55 19
www.clcbio.com · info@clcbio.com

Bioinformatics explained: Biological databases

Large scale research projects are becoming part of the research in more and more laboratories as both technical and scientific knowledge allows faster and more efficient research. As the amount of data gained from research projects all over the world is heavily increasing these years there is a growing need for databases storing and handling the biological data. Databases and the ability to organize data are needed in order to keep research efficient and to get optimal output and information from data obtained in the lab.

You can find databases related to different data types and subjects as e.g. nucleotides, proteins, genomes and taxonomy. Within each field several databases present their content in different ways, different file formats and with different purposes. Some databases are large and contain global data collections maintained and kept up to date by the responsible organization, while others are small and local and maybe only maintained for a limited period of time while a specific project is going on. This paper concentrates on describing some of the possibilities to be aware of when searching data on the Internet. The purpose is to give a short overview of a few large and well-known public databases.

Introduction

Starting out with any research project it is required to gain information on the problem to be investigated. This can be through books, fellow researchers or on the Internet. Unfortunately, many researches often struggle to find relevant information in the vast amount of data found in online resources. It is tedious and demanding to navigate through tens of website pages with varying degrees of relevant information. Depending on the nature of the project one may have to collect information from several databases. The Internet has become a very important and time-saving help in basic research but at the same time it is very demanding in terms of the knowledge which is required to filter and interpret the information.

Entire issues of journals have been dedicated to web servers and databases. Each January *Nucleic Acid Research* publishes an entire issue on publicly available databases and a July issue on webservers. Check the resources section below for additional information.

What is biological data?

Biological data comes in many different flavors depending on the research project. Most researches work with a number of different formats even though they may not at first hand realize this. Often you have to look at sequence data, interpret gel analyses, look for related topics on PubMed and finally write everything together in a paper or a report. Below is briefly listed some of the data which can be found when researching any biological question. Some of these data in the databases are partly overlapping and referring to each other.

- **Text.** Examples of text databases are PubMed and OMIM containing textual information and references related to biological data.
- **Sequence data.** GenBank and UniProt exemplifies biological databases containing DNA and protein sequences, respectively.
- **Protein structure.** You can also find databases specifically related to protein structure files as e.g. the PDB, SCOP and CATH databases.

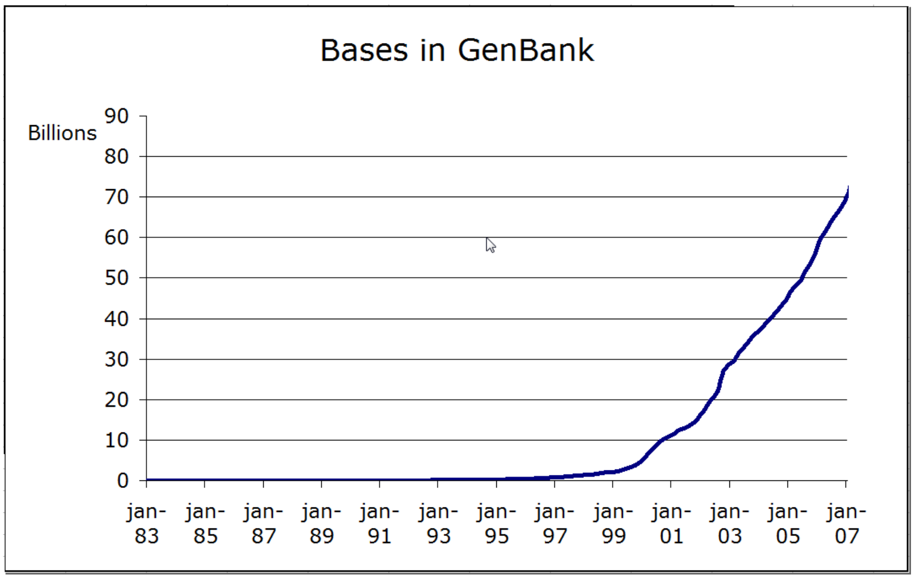

Figure 1: *Growth of the GenBank database (source:* `ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt`*).*

- **Links.** Most databases contain information on sequence data within a specific field or subject. A different type of database is e.g. the InterPro database consisting of a collection of links from protein domains and families to other databases providing related resources.

- **Images.** In the field of 2D gel and microscopic images you can also find various databases containing data e.g. identified on reference gel images.

- **Numerical data.** Gene expression data as well as other microarray data are also accessible from a number of databases. An example is the ArrayExpress database of the European Bioinformatics Institute, EBI.

- **Biological matter.** Frozen bacterial strains, vectors etc. are also to be found in databases collecting information on each of these specific biological matters, e.g. the UniVec database hosted by NCBI.

The following sections give a more detailed description of some of the databases mentioned above.

Sequence databases

With the current speed of sequencing projects a lot of work is required to store and organize sequence data. Most sequence databases store additional information along with the sequence. That could be references to the original research papers stored in PubMed, information about annotated regions, regions were conflicting residues have been published, information on species and much more. So far, a common standard for handling all of this information has not been created. Thus every database has its own standard on how to store the data. Most data is, however, stored in a plain text format (flatfile) and can thus be opened in standard software like Word, Notepad etc. However, large amounts of plain text may not be easy comprehensible. Another problem by storing in a flatfile format is the size of the database. Databases with thousands of sequence entries may become too large to handle on a normal PC for most users.

An alternative approach used by most websites with large databases is to store all the information in a relational database.

Relational databases have connections or pointers to additional data in other databases or tables. Thus one can easily and very fast retrieve a large amount of information on one particular sequence.

There are many sequence databases but here we will only focus on what we think are the four major sequence databases. One of the characteristics of these databases is that they are maintained and kept up to date on a regular basis.

- **GenBank.** A US-based comprehensive collection of various biological data.
- **EMBL.** The main European resource of nucleotide sequence data.
- **DDBJ** The DNA Data Bank of Japan.
- **UniProt** The universal protein resource.

The four databases are described below.

GenBank at NCBI

Hosted at http://www.ncbi.nlm.nih.gov/, the National Institute of Health has achieved a dominant position in collecting biological data of almost any kind [Wheeler et al., 2006]. In addition to storing sequence data, NCBI stores almost all kinds of biological sequence related data. PubMed is probably the mostly used service that NCBI offers to theirs users together with BLAST, an option for searching for homologous sequences in the entire database [Altschul et al., 1990]. Moreover, the NCBI staff provides software tools for handling sequence data.

EMBL

The EMBL Nucleotide Sequence Database is hosted by EBI - the European Bioinformatics Institute, at the European Molecular Biology Laboratory (EMBL), http://www.ebi.ac.uk/embl/.

DNA and RNA sequences are directly submitted to the EMBL nucleotide sequence database by individual researchers as well as by genome sequencing projects and patent applications, and the database is produced and maintained collaborating with both GenBank and the DNA Data Bank of Japan (DDBJ). The international collection of sequence data is exchanged between EMBL, GenBank and DDBJ on a daily basis and a knowledge of global sequence information can be retrieved from any of the three entries.

DNA Data Bank of Japan

DDBJ (DNA Data Bank of Japan) is a nucleotide database hosted in Japan and is accepting DNA submission from mainly Japanese researchers. They work in close collaboration with GenBank and EMBL and the three databases store almost identical data. DDBJ also provides various search and analysis tools through the website http://www.ddbj.nig.ac.jp/.

UniProt

UniProt is the universal protein resource, and as stated on its website the database intend to be both comprehensive and of high quality:

"The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information."

At the UniProt website, `http://www.uniprot.org/`, data has been divided into three classifications - core data, supporting data and information - and you can search information of your sequence in these three categories. You can also do BLAST searches and create alignments; a couple of other services are also provided. The UniProt Knowledgebase, UniProtKB, contains translations of the coding sequences submitted to EMBL, GenBank and DDBJ and the UniProtKB contains all publicly available protein sequences [Consortium, 2007].

Other valuable databases and resources

PubMed

PubMed gives you biological data in text format and this service provided by the U.S. National Library of Medicine links to more than 17 million resources from different journals within the field of life science.

A relatively new functionality at the NCBI website is the possibility to sign up for an account at My NCBI which is a service offering a customized and automated PubMed update. After registration at My NCBI you can save your searches and set up automated searches alerting you by e-mail. You can also customize e.g. filtering options on the searches.

PubMed can be accessed at `http://www.ncbi.nlm.nih.gov/pubmed/`.

Ensembl

Ensembl is a project developing software for automatic annotation of eukaryotic genomes. EMBL - EBI and the Sanger Institute are behind the project and at the website, `http://www.ensembl.org/index.html`, you can search within all the data from the Ensemble project divided into species.

EBI

One of the larger European bioinformatic centers, European Bioinformatics Institute (EBI), hosts a number of databases and a lot of methods to help analyze all this data.

The EBI website `http://www.ebi.ac.uk/` also stores the EMBL nucleotide sequence database (`http://www.ebi.ac.uk/embl/`) [Kulikova et al., 2007].

InterPro

Most of the databases mentioned above also provide links to related information in other databases. Nevertheless, the InterPro database try to a larger extent to link from one protein domain or family to a number of different databases which individually contains a lot of relevant information [Mulder et al., 2007]. The InterPro database does not contain any sequence information but is largely a mesh of hyperlinks to various other resources.

Link to InterPro: http://www.ebi.ac.uk/interpro/

Pfam

A very useful database for finding protein domains is the Pfam database [Finn et al., 2006]. Pfam currently stores information on more than 9000 protein families. When working on an unknown protein it is often very valuable to retrieve information of the actual protein family as identification of functional domains within a protein sequence can benefit your knowledge about the role and function of the protein.

You can access the Pfam database at http://pfam.janelia.org/.

Structure databases

Information about protein structure is not developing as fast as sequence data information due to slower pace in solving 3 dimensional structures of proteins.

The RCSB Protein Data Bank hosted at http://www.pdb.org holds slightly more than 48000 structures. At the website you can download structure files and you are provided a number of tools for structure studies.

SCOP: Structural Classification of Proteins is accessible at http://scop.berkeley.edu/. The SCOP database describes structural and evolutionary relationships between all known protein structures and also provides a number of links to other on-line resources related to protein structure and to sequence databases in general.

CATH Protein Structure Classification. The CATH database hosted at http://www.cathdb.info/ classifies protein structures from the PDB according to a four-level hierarchy.

Species-specific databases

In addition to all of the databases mentioned above it is possible to find a number of species-specific databases. Such databases usually hold very detailed information about only one particular species. A few examples are:

- **Colibase** An *E. coli* database http://colibase.bham.ac.uk/.
- **Flybase** A drosophila database http://flybase.bio.indiana.edu/.
- **Wormbase** A database for *C. elegans* and other nematodes http://www.wormbase.org/.
- **TAIR** The Arabidopsis Information Resource http://www.arabidopsis.org/.

Retrieving information

Unfortunately, there are no shortcuts for retrieving all information on one sequence in only one step. The vast abundance of data on the Internet has made it impossible to cope with all related information to one specific DNA or protein sequence. Nevertheless, the authors of one web page is trying to retrieve a lot of related information simply by displaying information from various databases on one single web-page: http://harvester.embl.de/.

Another retrieval system is the well-known SRS system from Lion Bioscience. Using the SRS system one can put together advanced text queries across a number of different databases.

Many research groups are also trying to develop methods and algorithms able to find relations between genes or proteins. Much information is available in the research papers found in PubMed, and advanced text-mining tools can be used to filter out information.

Erroneous data

With an exploding amount of data submitted to the databases there is an increased possibility of finding erroneous data. One emerging problem is that many computational prediction methods are trained on data extracted from databases in which they are later used to annotate sequences. In this way, they end up predicting based on their own predictions. Unfortunately, such errors are very hard to find and require a lot of labor and manual work.

Acknowledgements

We want to excuse the exclusion of many highly useful databases in this paper. Many researches have used weeks, months or even years to build and create databases for specific usages. We cannot mention all of them but simply acknowledge that they exist and more is coming as new methods are created which may yield new types of data. Many of these databases can be found in the semi annual database issue in the journal *Nucleic Acid Research*.

Other useful resources

Wikipedia on Biological Databases

http://en.wikipedia.org/wiki/Biological_database

Issues of Nucleic Acid Research

Year 2007

- **Database issue** http://nar.oxfordjournals.org/content/vol35/suppl_1/index.dtl
- **Webserver issue** http://nar.oxfordjournals.org/content/vol34/suppl_2/index.dtl

Year 2008

- **Database issue** http://nar.oxfordjournals.org/content/vol36/suppl_1/index.dtl

Creative Commons License

References

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

[Consortium, 2007] Consortium, T. U. (2007). The universal protein resource (uniprot). *Nucleic Acids Res*, 35(Database issue):D193–D197.

[Finn et al., 2006] Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. L., and Bateman, A. (2006). Pfam: clans, web tools and services. *Nucleic Acids Res*, 34(Database issue):D247–D251.

[Kulikova et al., 2007] Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Hoad, G., Kanz, C., Lee, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Nardone, F., Pastor, M. P. G., Plaister, S., Sobhany, S., Stoehr, P., Vaughan, R., Wu, D., Zhu, W., and Apweiler, R. (2007). Embl nucleotide sequence database in 2006. *Nucleic Acids Res*, 35(Database issue):D16–D20.

[Mulder et al., 2007] Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., Courcelle, E., Das, U., Daugherty, L., Dibley, M., Finn, R., Fleischmann, W., Gough, J., Haft, D., Hulo, N., Hunter, S., Kahn, D., Kanapin, A., Kejariwal, A., Labarga, A., Langendijk-Genevaux, P. S., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Nikolskaya, A. N., Orchard, S., Orengo, C., Petryszak, R., Selengut, J. D., Sigrist, C. J. A., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. (2007). New developments in the interpro database. *Nucleic Acids Res*, 35(Database issue):D224–D228.

[Wheeler et al., 2006] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Helmberg, W., Kapustin, Y., Kenton, D. L., Khovayko, O., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Pruitt, K. D., Schuler, G. D., Schriml, L. M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Suzek, T. O., Tatusov, R., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2006). Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 34(Database issue):D173–D180.