Internet Sites of Interest - Bioinformatics (updated)

This updates my first Internet Sites of Interest column, which appeared in March 2005. The subject has continued to develop, and an update seemed timely.

I have checked the items in the original column, updating where necessary. I have added some new items, and rearranged the article somewhat.

I have used a number of sources in compiling this column, and they are listed as "further reading" at the end.

All links were checked on 13th February 2008.

What is bioinformatics?

The Online Medical Dictionary defines bioinformatics as:

"The use of computers in solving information problems in the life sciences, mainly, it involves the creation of extensive electronic databases on genomes, protein sequences, etc. Secondarily, it involves techniques such as the three-dimensional modeling of biomolecules and biologic systems." (http://cancerweb.ncl.ac.uk/cgi-bin/omd?query=bioinformatics, accessed 13th February 2008)

Cellular function depends on proteins. Proteins consist of chains of amino acids. Cells produce proteins, and the information that determines which proteins are produced is contained in DNA.

A molecule of DNA is made up of thousands of nucleotides. Each nucleotide is made up of a "base", plus a phosphate plus a sugar. The sequence of the bases specifies the order of the amino acids in a protein. A segment of DNA carrying information to encode (produce) an amino acid is a gene. The complete set of genetic information relating to an organism is its genome.

Information about the sequence of amino acids in a protein or the bases in nucleotides, or genes, is available in those "extensive databases". Because the information is stored digitally, it can be manipulated and compared with other data.

The column includes sites about bioinformatics, as well as some of those "extensive databases", tools for searching them and working with the data, and other related resources.

Information about bioinformatics

Bioinformatics

http://bioinformatics.oupjournals.org/

Peer reviewed journal published by Oxford University Press. Abstracts are available free of charge, but you will need a subscription to read full text, unless the article has been published under an open access model. Such papers are clearly indicated.

BMC Bioinformatics
http://www.biomedcentral.com/bmcbioinformatics/
An open access journal publishing peer reviewed papers.

European Bioinformatics Institute
http://www.ebi.ac.uk/
The institute manages databases of biological data, and conducts research.

National Center for Biotechnology Information (NCBI)
http://www.ncbi.nlm.nih.gov/
NCBI manages databases of biological data. NCBI resources are listed at http://www.ncbi.nlm.nih.gov/sites/gquery, and the NCBI Handbook provides a guide to each one, at http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook. The annual Database Issue of Nucleic Acids Research (see below) includes an update on NCBI resources (Wheeler et al., 2008).

Nucleic Acids Research
http://nar.oxfordjournals.org/
This has recently become an open access journal and is available via the publisher website and also through PubMed Central. The annual Database Issue contains articles about factual biological databases, and the annual Web Server Issue articles about web based software tools for analysing data.

Online Lectures on Bioinformatics
http://lectures.molgen.mpg.de/online_lectures.html
Web based tutorial, from the Max Planck Institute for Molecular Genetics in Berlin.

Protein Sequence Analysis: A Practical Guide
http://www.bioinf.manchester.ac.uk/dbbrowser/bioactivity/
Bioinformatics web practical, from the University of Manchester. It introduces a range of tools and databases.

Bioinformatics databases

Following Xiong (Xiong, 2006), I have divided databases into **primary**, **secondary**, and **specialized**. Primary databases contain biological data, secondary databases add some processing to that data, and specialized databases contain data relevant to a particular research interest.

The Molecular Biology Database Collection is a list maintained by Nucleic Acids Research. It can be sorted by name of database, or by category, and there are links to each database, a summary about each, and a link to any article in NAR itself. Access the collection at http://www.oxfordjournals.org/nar/database/a. The annual Database Issue of NAR includes an update paper on this collection (Galperin, 2008), as well as papers on individual databases.

Primary databases

The three listed here work in close collaboration with each other.

DNA Databank of Japan (DDBJ)
http://www.ddbj.nig.ac.jp/

EMBL Nucleotide Sequence Database
http://www.ebi.ac.uk/embl/index.html

GenBank
http://www.ncbi.nlm.nih.gov/Genbank/
GenBank is a database of genetic sequence data. It can be accessed through several routes, outlined on this page.

Secondary databases

Gene
http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
Gene is a searchable database of genes, including information about their chromosomal location and their function.

OMIM, the Online Mendelian Inheritance in Man
http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim
Catalogue of human genes and associated diseases, edited by Victor McKusick and colleagues.

PROSITE: database of protein domains, families and functional sites
http://us.expasy.org/prosite/
PROSITE is a secondary database, which analyses sequence data from primary databases.

Protein Information Resource (PIR): integrated protein informatics resource for genomic and proteomic research
http://pir.georgetown.edu/
Access to various primary databases of protein information.

Swiss-Prot
http://us.expasy.org/sprot/
A composite database, amalgamating data from various sources. TrEMBL, its computer annotated supplement, is also accessible from this site. (The site is about to be replaced with http://beta.uniprot.org/)

TIGR Genome Projects (J. Craig Venter Institute)
http://www.tigr.org/db.shtml
Suite of databases of DNA and protein sequences, gene expression, and other information, for humans and other organisms.

UniProt (http://www.ebi.ac.uk/uniprot/) combines Swiss-Prot, PIR, and TrEMBL.

Specialized databases

These include, among many others:

AceDB
http://www.acedb.org
Genome database for Caenorhabditis elegans (a nematode worm, used in genetics research).

FlyBase
http://flybase.bio.indiana.edu
A database of the genome of Drosophila – a fruit fly used in genetics research.

HIV Databases
http://www.hiv.lanl.gov/content/
Information on genetic sequences, immunological epitopes, drug-resistance associated mutations, and vaccine trials. Databases hosted by the Los Alamos National Laboratory, funded by the NIH.

TAIR
http://www.arabidopsis.org
Genetic and molecular biology information for *Arabidopsis thaliana*, a member of the mustard family used in genetics research.

Related resources

The annual Web Server issue of Nucleic Acids Research (see above) includes papers on individual search tools and software packages.

BLAST (Basic local search alignment tool)
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
BLAST compares local nucleotide or protein sequences to sequence databases, to look for similarities.

Ensembl
http://www.ensembl.org/
Software system for searching information on the genomes of various organisms including humans, rats and mosquitoes.

Entrez
http://www.ncbi.nlm.nih.gov/sites/gquery
The NCBI's life sciences search engine. This page links to a range of databases of biological data, and also to PubMed, PubMed Central, MeSH and other familiar things!

HAPMAP
http://www.hapmap.org/
A public resource to help researchers to find human genes associated with disease.

PubMed
http://www.ncbi.nlm.nih.gov/sites/entrez/
Bibliographic references in PubMed will link to related information in other NCBI resources. PubMed, as Claverie and Notredame point out (Claverie and Notredame, 2003), is an important way to locate published information on sequences.

References

CLAVERIE, J.-M. & NOTREDAME, C. (2003) *Bioinformatics for dummies,* Indianapolis, Wiley.

GALPERIN, M. Y. (2008) The Molecular Biology Database Collection: 2008 update. *Nucl. Acids Res.,* 36**,** D2-4.

WHEELER, D. L., BARRETT, T., BENSON, D. A., BRYANT, S. H., CANESE, K., CHETVERNIN, V., CHURCH, D. M., DICUCCIO, M., EDGAR, R., FEDERHEN, S., FEOLO, M., GEER, L. Y., HELMBERG, W., KAPUSTIN, Y., KHOVAYKO, O., LANDSMAN, D., LIPMAN, D. J., MADDEN, T. L., MAGLOTT, D. R., MILLER, V., OSTELL, J., PRUITT, K. D., SCHULER, G. D., SHUMWAY, M., SEQUEIRA, E., SHERRY, S. T., SIROTKIN, K., SOUVOROV, A., STARCHENKO, G., TATUSOV, R. L., TATUSOVA, T. A., WAGNER, L. & YASCHENKO, E. (2008) Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res.,* 36**,** D13-21.

XIONG, J. (2006) *Essential bioinformatics,* New York, Cambridge University Press.

Further reading

In addition to things cited above, here are some things I found useful while writing the old or new versions of this column. Some include information on the science, which I found useful as a non-biologist, and some include information on some of the databases and how to use them.

ATTWOOD, T. K. & PARRY-SMITH, D. J. (1999) *Introduction to bioinformatics,* Harlow, Longman.

BRADLEY, J., JOHNSON, D. & RUBENSTEIN, D. (2001) *Lecture notes on molecular medicine,* Oxford, Blackwell Science.

FOGEL, G. B. (2003) Internet resources for bioinformatics data and tools. IN FOGEL, G. B. & CORNE, D. W. (Eds.) *Evolutionary computation in bioinformatics.* Amsterdam, Morgan Kaufmann.

MOORE, J. H. (2007) Bioinformatics. *J Cell Physiol,* 213**,** 365-9. (this “minireview” discusses databases and data mining software)

Keith Nockels

Clinical Sciences Library
University of Leicester
RKCSB, Leicester Royal Infirmary
PO Box 65,
Leicester,
LE2 7LX
UK

Telephone: (0116) 252 3101
Fax: (0116) 252 3107
Email: khn5@le.ac.uk.

February 2008