Bioinformatics

A User's Perspective

Naftali Kaminski

Lung Biology Center, University of California San Francisco, San Francisco, California; and Functional Genomics Unit, Sheba Medical Center, Tel-Hashomer, Israel

This review provides an overview of bioinformatics from the user's point of view. Bioinformatics, defined as the application of computers, databases, and computational methods to the management of biologic information, is essential for almost every aspect of data management in modern biology. The rapid accumulation of genomic sequence information together with the wide availability of new technologies that analyze global gene expression patterns have created an information overload. Molecular biology labs are increasingly dependent on computers, large-capacity databases, search and analysis tools, and high-quality Internet connections. Currently available bioinformatics tools are discussed and a general approach is outlined. Using the resources and approaches in this review, readers should be able to form their own view of bioinformatics and tailor the solutions to the information overload according to their needs.

Information management is definitely not an invention of the last decade. Many of the advances of humans are related to developments in information management. It is not by coincidence that printing is considered by many the most important invention of the last millennium. This tool for massive spread and storage of information changed the history of civilization. In biology, the realization that cellular information transfer is crucial at every functional level has grown slowly since the 1950s until it reached its full current blossom. To many of us, the notion that most of our basic science study involves the ways biologic systems manage and exchange information is almost trivial. Cells manage and exchange information in multiple ways, including receptor-mediated signaling, secretion of cellular effectors, and translation of stored genomic information to new proteins and phenotypes. The study of genomic information has especially influenced biology and biology-related fields. The growth in DNA sequence data available to researchers is unparalleled. GenBank, a major public database where DNA sequences are stored, doubles in size approximately every year. At the beginning of the year 2000, Genbank contained over 3.8 million sequence records, and this collection grows at a rate of more than a million nucleotides per day.

Since the publication of the first complete bacterial genome (*Haemophilus influenzae*) in 1995 (1), the complete genomes of three more eukaryotic model systems were published (the budding yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans*, and the fruit fly *Drosophila melanogaster*) (2). Two more are near completion (the flowering plant *Arabidopsis thaliana* and the fission yeast *Schizosaccharomyces pombe*) (2). The human genome project is also on its way to completion, with a working draft completed this year and a fully cited edited version expected by 2002. Novel technologies for large-scale expression analysis are already in use by many groups. These technologies use the information derived from the various sequencing projects to create comprehensive gene expression profiles and result in immensely large sets of data (3). On top of that, the biomedical literature is experiencing a citation explosion, with approximately 400,000 biomedical publications cited every year in PubMed, the National Library of Medicine's biomedical citation search service that provides access to biomedical literature databases. The huge amount of information that is currently available and is continuously generated cannot be analyzed manually. Computers, large-capacity databases, search and analysis tools, and high-quality Internet connections are now essential tools for biologists. Bioinformatics, defined as the application of computers, databases, and computational methods to the management of biologic information, is essential for almost every aspect of data management in modern biology.

The aim of this review is to provide the reader an overview of the current available Internet-based bioinformatics tools. The tools and databases discussed range from traditional tools for sequence analysis to the new and rapidly developing tools for analysis of microarray data. The discussion that follows distinguishes between tools and databases. This distinction is, of course, artificial because databases are useless without search tools, and vice versa. Further, many of the sites offer a wide array of services, tools, databases, and useful links to other sites. Naturally, because this is a rapidly changing field, a comprehensive guide to all the bioinformatics sites and applications may become outdated even before publication. However, the reader should be able to use the outlined approach and the sites mentioned to find additional sources and design an individual bioinformatics approach.

Tools for Sequence Analysis

Bioinformatics is relevant for any biologic field, including genetics, biochemistry, and medicine. However, much of the publicity and emphasis that bioinformatics has received in the last few years has been through its importance in DNA and protein-sequence analysis.

Sequence Alignment

Sequences obtained through sequencing or a database search are compared with multiple other sequences that are in databases using sequence alignment algorithms. Se-

(*Received in original form July 11, 2000*)

Address correspondence to: N. Kaminski, M.D., Functional Genomics, Institute of Respiratory Medicine, Sheba Medical Center, Tel-Hashomer 52621, Israel. E-mail: kamins@itsa.ucsf.edu

Abbreviations: expressed sequence tag, EST; self-organizing map, SOM.

Am. J. Respir. Cell Mol. Biol. Vol. 23, pp. 705–711, 2000
Internet address: www.atsjournals.org

quence similarities (homologies) can be used to infer the function of newly found sequences, to construct functional and structural protein families, to analyze control elements, and to gain insight into evolutionary relationships.

The most commonly used program for sequence alignment is Basic Local Alignment Search Tool (BLAST) (4). This is a set of similarity search programs designed to explore all available sequence databases regardless of whether the query is a protein or DNA. BLAST programs are designed for speed with minimal sacrifice of sensitivity. The scores provided with BLAST are statistically sound and simple to use.

The easiest way to use BLAST is through the Web at the site of the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/BLAST). The site contains different types of searches and several specialized data sets made available by different sequencing centers. BLAST can also be run locally, from a network server or through e-mail (all the details are at the BLAST home page). A probability value E is calculated for each sequence comparison, giving an estimate of the probability that the match happened by chance (the lower this number, the better the match). For the best matches, the similarity is shown directly using an alignment of the two sequences.

The underlying assumption is that the sequences are similar because they are homologous; that is, they are derived from a common ancestral sequence or they share common functional aspects. The best way to compare sequences is to align them, inserting gaps if necessary so that they match as much as possible. This is useful to identify regions of common function as previously mentioned. Two useful new BLAST applications are VecScreen, for BLAST-based detection of vector contamination, and IgBLAST, a tool developed at NCBI to facilitate analysis of immunoglobulin (Ig) sequences in GenBank.

Protein sequences can be analyzed similarly. Some common types of analyses are database similarity searching (to identify protein-sequence database entries similar to a given protein) and sequence comparison (for example, to align two protein sequences and identify common regions). More sophisticated ways to analyze protein sequences are motif searching and three-dimensional structure prediction.

Motif Searching

Motif searching is very important in trying to predict the function of a protein using its sequence. Motifs are sequences of amino acids in a protein that are associated with a known function or structural feature. A number of automated databases of protein motifs such as PROSITE (http://www.expasy.ch/prosite/) and PFAM (http://www/sanger.ac.uk/Software/Pfam/index.html) have been created either from literature surveys or directly from sequence databases, for the purpose of identifying proteins or domains or particular functional sites. The databases provide analytic tools that recognize specific amino-acid patterns with functional significance and allow development of hypotheses regarding the function of novel protein proteins and genes as well as assignment to functional and structural families.

Structure Prediction

The function of a protein is strongly dependent upon its three-dimensional structure, and one of the main challenges in bioinformatics is to predict the structure of a protein from its sequence. Methods for predicting the secondary structure of a protein (arrangement of α-helices and β-sheets) have been available for many years. Cn3D is a helper application that allows viewing of three-dimensional structures from NCBI's Entrez retrieval service. The new version of Cn3D simultaneously displays structure, sequence, and alignment; allows the user to set display styles for features of interest; and can be downloaded from Entrez (http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml). Another useful site that provides multiple protein-modeling services is the EMBL Biocomputing home page (http://www.sander.embl-heidelberg.de/future/other.html).

Tools for Analysis of Gene Expression Data

Gene expression information is important to the understanding of many aspects of cellular and whole-organ function. The advent of microarray technology allows the analysis of the expression of thousands of genes simultaneously (5), thus creating a comprehensive transcriptional profile of a condition studied. Microarrays have been used in a variety of experimental settings, including analysis of the cellular response to stimuli (6, 7), to distinguish specific transcriptional programs in whole organs of genetically modified animals (8) and to characterize human cancers (9). Typically, these techniques generate large amounts of data that require combined bioinformatics approaches and very advanced tools for analysis (10). In general, these tools use computational methods to group (cluster) genes or experiments with similar profiles of changes in expression levels. The assumption is that by distinguishing genes that behave similarly it is possible to gain insight into shared regulatory aspects, shared functions, or roles in the process studied. The tools most commonly used are:

- *Hierarchical clustering*. Pairwise gene distance matrices (used originally for molecular phylogeny using sequence alignments) can be used to identify genes sharing a similar expression pattern in multiple experiments with complementary DNA (cDNA) or oligonucleotide arrays. In one such method, all values are paired and modified Pearson correlations are calculated for each possible pairwise combination and used in distance matrices. This allows hierarchical clustering of groups of genes that behave most similarly. Cluster, an application developed by Eisen and colleagues (11), can be downloaded from http://rana.Stanford.EDU/software. It was originally designed for analysis of cDNA arrays but can easily be used to analyze oligonucleotide arrays. The resulting clusters can then be visualized using Treeview (available from the same site). The use of this method allows both the recognition of distinct groups of genes that behave similarly (clustering of genes), such as functionally related genes (8); and of unique transcriptional features of a condition (clustering of groups), for instance, for disease classification (9). Cluster now contains multiple tools for analysis of gene expression data, including self-organizing maps (SOMs).

- *SOMs* are ideally suited for exploratory data analysis. They are considered superior to hierarchical clustering when analyzing "messy data" that contains outliers, irrelevant variables, and nonuniform densities (10). The basic concept is that you impose a partial structure on the data and then adjust the structure iteratively according to the data. The input data are the raw expression values (not ratios) obtained from the array experiments. The output is a series of SOMs represented by their similar patterns. GeneCluster is an application that produces and displays the SOMs. It was developed by Tamayo and associates at the Whitehead Institute (12) and applied successfully to the study of cancer classification (13). The program can be downloaded from http://waldo.wi.mit.edu/MPR/software.html.

Databases and Useful Websites

Genome Maps

The study of genomic information is vital to our understanding of genetic diseases and evolution. At the highest level of genetic organization, the order of markers along a chromosome can be elucidated using genetic linkage analysis, in which the frequency of recombination between markers is used as an indicator of the distance between these markers. Higher-resolution maps of the location of clones and polymerase chain reaction fragments can be obtained by molecular techniques. Particularly interesting and useful sites are the Genome Channel (http://compbio.ornl.gov/channel/) and the Entrez genome site (http://www3.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome). The CEPH-Généthon website (http://www.cephb.fr/bio/ceph-genethon-map.html) and the Cooperative Human Linkage Center (http://lpg.nci.nih.gov/CHLC/) provide genome maps and are relatively easy to use. Another useful site for both genomic and gene expression data is the Whitehead Institute for Biomedical Research/MIT Center for Genome Research website (http://www-genome.wi.mit.edu/).

Nucleotide Sequence Databases

There are three major databases that collect publicly available nucleotide sequences.

- Genbank is maintained by NCBI in the United States (http://www.ncbi.nlm.nih.gov/Genbank/index.html).
- EMBL is maintained by the European Bioinformatics Institute in the United Kingdom (http://www.ebi.ac.uk/).
- DNA Database of Japan is based in Japan (http://www.ddbj.nig.ac.jp/fromddbj-e.html).

These databases contain the same sequences, but use different formats for their annotations. All three offer tools for sequence and structure prediction. They contain both genomic and cDNA sequences.

When analyzing a genomic DNA sequence it is important to determine where the protein coding regions are located. Coding regions account for only a small fraction of the human genome (1 to 2%), thus gene identification is not an easy undertaking.

One strategy for analyzing coding sequences in the genome is by sequencing cDNAs and creating large libraries of expressed sequence tags (ESTs). ESTs are single-pass sequences obtained from the 3′ or 5′ ends of cDNAs. Their quality may be low, but the volume of EST sequences produced makes these sequences a great source for new gene sequences. DbEST (http://www.ncbi.nlm.nih.gov/dbEST/) and Unigene (http://www.ncbi.nlm.nih.gov/UniGene/index.html) are databases dedicated to ESTs. UniGene is an experimental system for automatically partitioning GenBank sequences into a nonredundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information, such as the tissue types in which the gene has been expressed and the map location. The information collected in Unigene has been extremely useful as a resource for gene discovery and for probe design for cDNA and oligonucleotide arrays.

Protein Databases and Websites

Databases dedicated to proteins often include sequence information as well as analysis tools. The most used database is SWISS PROT (http://www.expasy.ch/sprot/), a curated protein-sequence database that provides a high level of annotations (such as the function of a protein, its domains structure, post-translational modifications, variants, etc.) Several sites provide PFAM, PROSITE, and other tools for protein sequence analysis. The Sanger Center (http://www.sanger.ac.uk/) provides access and maintains PFAM and multiple other useful links and genomic tools, including three-dimensional protein structure prediction (http://genomic.sanger.ac.uk/123D/123D.shtml). Another very useful site that contains links to multiple protein databases is ExPASy (Expert Protein Analysis System) (http://www.expasy.ch/), the proteomics server of the Swiss Institute of Bioinformatics. In addition to access to databases it provides multiple protein analysis tools, such as PROSITE (http://www.expasy.ch/prosite/) and Swiss-3Image (http://www.expasy.ch/sw3d/).

Transcription Factor Databases

In transcription factor databases you can search for protein-binding sites in DNA sequences, especially transcription factor binding sites. One such database is Matinspector Public Domain (http://genomatix.gsf.de/cgi-bin/matinspector/mantinspector.pl). Alternative programs are Proscan and Sigscan, offered at the Bioinformatics & Molecular Analysis Section (BIMAS) site (http://bimas.dcrt.nih. gov/molbio/index.html). Because these databases do not overlap it is often useful to use more than one in the analysis of a given sequence.

Signaling Pathways

Signaling pathway databases connect many proteins together and provide a context for the study of signaling-relevant gene expression at the RNA and protein levels. Signaling pathway databases combine information about substrates, reactions, enzymes, and regulatory mechanisms. Some examples are:

- G-Protein Coupled Receptor Database (http://www.gcrdb.uthscsa.edu)
- Protein Kinase Resource (http://www.sdsc.edu/projects/Kinases/pk_home.html)
- Cell Signaling Networks Database (http://geo.nihs.go.jp/csndb/)

Immunologic Databases and Websites

These databases provide focused, detailed, and updated information about genes relevant to the immune response.

- Cytokine Family cDNA Database (dbCFC) (http://cytokine.medic.kumamoto-u.ac.jp/) is a collection of EST records of cytokines deposited in GenBank. A related database is the Cytokine Receptor Family Database (http://crf.medic.kumamoto-u.ac.jp/), which contains similar information for cytokine receptors.
- CD Guides (http://www.ncbi.nlm.nih.gov/prow/guide/45277084.htm) is a very useful database that contains definitions of all molecules assigned CD numbers, as well as information about their function and cellular expression.
- The Kabat Database of Sequences of Proteins of Immunological Interest (http://immuno.bme.nwu.edu/) contains sequences that are classified according to antigen specificity.
- The Immunogenetics database (IMGT) (http://imgt.cnusc.fr:8104/) focuses on Igs, T-cell receptors, and major histocompatibility complex molecules.

Gene Expression Databases

Several sites offer useful protocols, downloadable software, and complete sets of experimental data. The Brown Lab website (http://cmgm.stanford.edu/pbrown/) provides a "do-it-yourself" manual for making and using cDNA arrays, in addition to complete sets of gene expression data. The Whitehead/MIT Center for Genome research website (http://waldo.wi.mit.edu/MPR) contains data sets and downloadable software. So far there is no central repository for gene expression data, however several groups are concentrating their efforts on creating standards that will enable the creation of such a central repository. One such effort is GeneX, a Web-based database for gene expression data being developed at the National Center for Genome Resources (NCGR) (http://www.ncgr.org/).

General Databases

Several general databases contain information from multiple sources, thus making them extremely useful for data mining and informatics.

- GeneCards, human genes proteins and diseases (Weizman Institute) (http://nciarray.nci.nih.gov/cards/index.html), contains automatically extracted information from multiple sites with links to NCBI, Unigene, OMIM, and SWISS PROT.
- OMIM (Online Mendelian Inheritance in Man) is a catalog of human genes and genetic disorders. The database contains textual information, pictures, and reference information. It also contains copious links to NCBI's Entrez database of MEDLINE articles and sequence information (http://www.ncbi.nlm.nih.gov/Omim/).
- Kyoto Encyclopedia of Genes and Genomes (KEGG) presents the current knowledge of molecular and cellular biology in terms of information pathways, consisting of interacting molecules or genes. It also provides multiple links to gene catalogs produced by genome sequencing projects (http://www.genome.ad.jp/kegg/).
- NCGR (http://www.ncgr.org/) provides access to multiple bioinformatics tools, some including Gepasi (a biochemical kinetics simulator), PathDB (a database of biochemical and metabolic pathways), and GeneX (a Web-based gene expression database).
- The Mouse Genome Informatics (MGI) site (http://www.informatics.jax.org/) provides integrated access to various sources of information on the genetics and biology of the laboratory mouse. The MGI resource comprises the Mouse Genome Database (MGD), the Gene Expression Database (GXD), and related resources, including the Mouse Tumor Biology database, the Rat Data resource, Michael Festing's Listing of Inbred Strains of Mice and Rats, and MouseBLAST (http://www.informatics.jax.org/userdocs/aboutMGI.shtml).

Bioinformatics: The User's Perspective

From the previous discussion it is clear that there is not only a current overload of information but also an overload of bioinformatics tools. There is an increase in the availability of user-friendly, intuitive tools for bioinformatics, and there is no doubt that these tools are taking their place in the armamentarium of every molecular biology laboratory. However, even with the current wealth of available informatics tools, the rapid growth of biologically relevant information by far outweighs the computational capacity of a molecular biology lab. For instance, when a molecular biology starts using microarrays it immediately becomes an information-intensive lab. This means that the lab is flooded with massive volumes of information at a very rapid rate. Obtaining optimal value from the data requires both a preplanned databasing effort and the use of "intelligent" sieving systems for automatic detection of relevant information with a minimum of human input.

These technologies not only are expensive but also carry with them a hidden economic burden: the price of manipulating information. The costs of computers, software packages, and infrastructure are high, and the need for frequent upgrades generates a continuing expense. The lack of trained personnel as well as the nature of scientific curiosity cause scientists to invest days in learning and implementing analysis tools—tasks that did not exist in the recent past. These new tasks and costs require rethinking the traditional laboratory personnel structure as well as the implementation of institutional solutions for personnel training and software license acquisition. Several simple suggestions that can help in making the transition to information intensive environment easier are:

- When you purchase a computer system, find the one that most analysis platforms fit.
- Dedicate workstations to informatics—do not allow the loading of unnecessary software. This will prevent unnecessary system slowing and/or crashes.
- Make sure that you have fast and stable Internet access.
- Buy only the minimum necessary software licenses. Many tools are in the public domain; it is not always true that the commercial packages are better.
- When you are considering a software package, always try it before purchase. Most companies provide trial

"demo" software and it is important to know beforehand that the package actually fits your needs.
- Set up standards for your lab database. File names should contain dates and information that identifies the experiment. Keep an independent index of these names. This database should also contain gene information with automated links. Preferably it should be designed so that data from multiple sources can be automatically retrieved. Ideally, this would be an institutional effort.
- Set up the database so that components can be published as Web-based complementary datasets to published papers.

This review provides an overview of bioinformatics tools and approaches for the molecular biologist and pulmonary scientist. By using the resources and the approaches outlined here the reader should be able to form an individual approach to bioinformatics and tailor the solutions to the information overload according to individual needs. A list of useful bioinformatics websites (Appendix 1) and a glossary of bioinformatics terms (Appendix 2) are available at http://medicine.ucsf.edu/divisions/lbc/.

Acknowledgments: The author thanks Dr. Dean Sheppard and Dr. David Erle for their useful discussions and suggestions. Dr. Renu Heller and Dr. June Lee were also extremely helpful in the preparation of this manuscript.

References

1. Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, *et al*. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd [see comments]. *Science* 269:496–512.
2. Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology [In Process Citation]. *Nat. Genet.* 25:25–29.
3. Huang, S. 1999. Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. *J. Mol. Med.* 77:469–480.
4. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.
5. Brown, P. O., and D. Botstein. 1999. Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* 21:33–37.
6. Spellman, P. T., G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* 9:3273–3297.
7. Boehm, U., T. Klamp, M. Groot, and J. C. Howard. 1997. Cellular responses to interferon-gamma. *Annu. Rev. Immunol.* 15:749–795.
8. Kaminski, N., J. D. Allard, J. F. Pittet, F. Zuo, M. J. Griffiths, D. Morris, X. Huang, D. Sheppard, and R. A. Heller. 2000. Global analysis of gene expression in pulmonary fibrosis reveals distinct programs regulating lung inflammation and fibrosis. *Proc. Natl. Acad. Sci. USA* 97:1778–1783.
9. Alizadeh, A. A., M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T, Moore, J. Hudson, Jr., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, L. M. Staudt, *et al.* 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling [see comments]. *Nature* 403:503–511.
10. Claverie, J. M. 1999. Computational methods for the identification of differential and coordinated gene expression [In Process Citation]. *Hum. Mol. Genet.* 8:1821–1832.
11. Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95:14863–14868.
12. Tamayo, P., D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentation. *Proc. Natl. Acad. Sci. USA* 96:2907–2912.
13. Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531–537.

Appendix 1: List of Useful Websites

Introductory Websites

National Center for Biotechnology Information
http://www3. ncbi.nlm.nih.gov/
European Bioinformatics Institute
http://www.ebi.ac.uk/
Whitehead Institute/MIT Center for Genome Research
http://www-genome.wi.mit.edu/
KEGG: Kyoto Encyclopedia of Genes and Genomes
http://www.genome.ad.jp/kegg/
Weizmann Institute Bioinformatics unit
http://bioinformat ics.weizmann.ac.il/
National Center for Genome resources
http://www.ncgr.org/

Genome Databases

Genome Channel
http://compbio.ornl.gov/channel/
Entrez Genome
http://www3.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome.
CEPH-Généthon
http://www.cephb.fr/bio/ceph-genethon-map.html
CHLC Cooperative Human Linkage Center
http://lpg.nci. nih.gov/CHLC/
Hugo
http://www.gene.ucl.ac.uk/hugo/
Whitehead Institute/MIT Center for Genome Research
http://www-genome.wi.mit.edu/
TIGR Human Gene Index
http.tigr.org/tdb/hgi/hgi.html
Mouse Genome Database MGI
http://www.informatics. jax.org/
Rat Genome Database
http://www.informatics.jax.org/rat

Nucleotide and Sequence Databases

Genbank
http://www.ncbi.nlm.nih.gov/Genbank/index.html
DDBJ
http://www.ddbj.nig.ac.jp/fromddbj-e.html
DbEST
http://www.ncbi.nlm.nih.gov/dbEST/
Unigene
http://www.ncbi.nlm.nih.gov/UniGene/index.html

Nucleotide and Protein Sequence Analysis

BLAST
http://www.ncbi.nlm.nih.gov/BLAST
PROSITE
http://www.expasy.ch/prosite/
PFAM
http://www.sanger.ac.uk/Software/Pfam/index.html

Protein Structure Prediction

CN3D
http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml

Swiss-3Image
http://www.expasy.ch/sw3d/
ExPASy Expert Protein Analysis System
http://www.expasy.ch/
Sanger Center
http://www.sanger.ac.uk/

Transcription Factors

Matinspector Public Domain
http://genomatix.gsf.de/cgi-bin/matinspector/matinspector.pl
Transfac
http://bioinformatics.weizmann.ac.il/transfac/.
Proscan and Sigscan
http://bimas.dcrt.nih.gov/molbio/index.html/.

Signaling

G-Protein Coupled Receptor Database
http://www.gcrdb.uthscsa.edu/
Protein Kinase Resource
http://www.sdsc.edu/projects/Kinases/pk_home.html
Cell Signaling Networks Database
http://geo.nihs.go.jp/csndb/

Immunology

Kabat Database
http://immuno.bme.nwu.edu/
The Immunogenetics Database (IMGT)
http://imgt.cnusc.fr:8104/
Cytokine Receptor Family Database
http://crf.medic.kumamoto-u.ac.jp/
Cytokine Family cDNA Database dbCFC
http://cytokine.medic.kumamoto-u.ac.jp/
CD Guides
http://www.ncbi.nlm.nih.gov/prow/guide/45277084.htm

Expression Data and Analysis

The Brown Lab
http://cmgm.stanford.edu/pbrown/
Stanford Genome Center
http://genome-www.stanford.edu/
The Microarray Project at NHGRI
http://www.nhgri.nih.gov/DIR/LCG/15K/HTML
Whitehead/MIT Center for Genome Research
http://waldo.wi.mit.edu/MPR/
Array Papers at Rockefeller University
http://linkage.rockefeller.edu/wli/microarray/
Leming Shi's DNA Microarray
http://www.gene-chips.com/
GenMapp
http://gladstone.ucsf.edu/labs/conklin/GenMAPP/entry.html
Bodymap
http://bodymap.ims.u-tokyo.ac.jp/
Sheppard Lab
http://medicine.ucsf.edu/divisions/lbc/

Proteomics

Danish Center for Human Genome Research
http://biobase.dk/cgi-bin/celis
EXPASY
http://www.expasy.ch/

General

GeneCards
http://nciarray.nci.nih.gov/cards/index.html
SWISS PROT
http://www.expasy.ch/sprot/
OMIM–Online Mendelian Inheritance in Man
http://www.ncbi.nlm.nih.gov/Omim/
The Virtual Genome Center
http://alces.med.umn.edu/VGC.html
Genes and Diseases
http://www.ncbi.nlm.nih.gov/disease/
Geneatlas
http://www.citi2.fr/GENATLAS/welcome.html
The Mouse Genome Informatics MGI
http://www.informatics.jax.org/
Hugo Directory
http://www.gdb.org.hugo
Human SNP Database
http://www-genome.wi.mit.edu/SNP/human/index.html
Gene Ontology Consortium
http://www.geneontology.org/

Lung-Related Databases

Asthma Gene Database
http://cooke.gsf.de/asthmagen/main.cfm
Yale Pulmonary Medicine Internet Resources
http://www.med.yale.edu/library/sir/select.php3?prof_subject=Medicine~Pulmonary+Medicine
American Thoracic Society
http://www.thoracic.org/index. html

Appendix 2: Glossary of Bioinformatics

BLAST (Basic Local Alignment Search Tool): A set of similarity search programs designed to explore all available sequence databases.
Browser: Client software that facilitates navigation on the Internet.
Client: A program or computer that is able to share the resources (printers, files, programs) of another program or computer, called a server. In client/server computing, the client requests services from the server.
Database: Data (or information) organized in a structure designed for easy retrieval, updating, and deleting.
Database engine: A database management system that provides tools for creating the database (database design) and managing the information stored in the database.
Database query: A statement that requests data from the database. Most database engines have simple interfaces to perform queries and do not require programming.
Data mining tools: Software applications that allow for sifting through large amounts of data to distinguish (to mine) the most meaningful and valuable information. These include visualization tools and data management tools.
FTP (File Transfer Protocol): A standard that allows files to be transferred between computers of different architectures and on different networks.

Gene expression: A highly specific process in which a gene is switched on at a certain time and begins production of its protein.
Gene mapping: Determining the relative positions of genes on a chromosome and the distance between them.
Genome: The sum of all the genetic information of an organism, usually meaning the entire DNA contained in an organism or a cell, including both nuclear and mitochondrial DNA.
Homologue: A gene similar in a major part of its sequence to another gene. A common ancestral origin is usually hypothesized as well as a relatively similar function.
HTML (Hypertext Mark-up Language: The standard coding language in which World Wide Web documents are written.
HTTP (Hypertext transfer protocol): The standard language used by World Wide Web client and servers.
Hyperlink: Hyperlinks are features of an electronic document that point the reader to another place in the same document or to a completely different document. These are the most important component of hypertext systems. Often just called "links."
Microarray technology: A new way of studying the simultaneous expression of thousands of genes. These methods use advanced technologies of printing (cDNA arrays) or photolithography (oligonucleotide arrays) to attach multiple (thousands of) cDNA clones or olignucleotides to a slide. The next steps involve hybridization with a labeled sample and scanning and analysis using specialized software. The level of fluorescence is supposed to represent the level of expression of a specific gene in the sample.
Physical mapping: A map of identifiable landmarks on DNA regardless of inheritance (e.g., restriction sites, genes). Distance is measured in base pairs (bp). For the human genome the lowest-resolution physical map is the banding patterns on the chromosomes; the highest would be the complete sequence.
Proteome: The entire protein repertoire contained in an organism or a cell, including all its components.
PFAM: A semiautomatic protein family database, it contains a collection of protein families and domains, multiple protein alignments, and profiles of these families.
PROSITE: An application used to determine the function of uncharacterized proteins translated from genomic or cDNA sequences, it consists of a database of biologically significant sites and patterns and computational tools that allow the user to rapidly and reliably identify the known family of protein (if any) to which the new sequence belongs.
SAGE (Serial Analysis of Gene Expression): A method for comprehensive analysis of gene expression patterns. Three principles underlie the SAGE methodology: (*1*) A short sequence tag (10 to 14 bp) contains sufficient information to uniquely identify a transcript, provided that the tag is obtained from a unique position within each transcript. (*2*) Sequence tags can be linked together to form long serial molecules that can be cloned and sequenced. (*3*) The number of times a particular tag is observed provides the expression level of the corresponding transcript.
Sequencing DNA: Determining the exact order of the bp in a segment of DNA.
Sequencing Protein: Determining the exact order of the amino acids in a protein.
URL (Unique Resource Locator): A standard protocol for providing a unique address for different documents and services on the World Wide Web. URLs include websites, FTP sites, e-mail servers, etc.
Website: A place on the Internet. Every Web page has a location where it resides, called a site. This location is stated by the URL. The statement usually begins with "http://.... ."