Chapter 13

Internet Resources and Public Databases

13.1 A Rapidly Changing Set of Resources

It is well known that resources available on the Internet are changing faster than almost everything else in the world of information processing. This also holds true for the dedicated tools available for biological sequence analysis. New tools are constantly becoming available, while others that are still available are getting obsolete. It is not easy to follow the state of the art in the many specialized areas of bioinformatics, where computational analysis is a powerful alternative to significant parts of the experimental investigation one may carry out.

Many of the tools offered the Internet are made available not by large organizations and research groups but by individual researchers many of whom may be actively involved in the field for only a shorter period. The funding situation, even for some of the major computational services, may change from year to year. This means that links are not updated regularly and that many servers may not be kept running 24 hours per day. If a service gets popular, the server behind it often will be upgraded sufficiently only after some delay. However, in many cases this is counterbalanced by mirror servers established by federal organizations, such as the NCBI in Washington, D.C., the EBI in Hinxton, U.K., and DDJB in Japan.

One highly confusing feature of the "open bioinformatics market" is that the same type of service can be available from many different sites based on different implementations. This is, for example, the case for protein secondary structure prediction, gene finding, and intron splice site prediction. The as-

signment of solvent exposure to amino acids in proteins is another type of prediction that is available from numerous sources. Since these methods most often have been constructed and tested with different sets of data, it can be hard even for specialists to assess objectively which method one should prefer. Often it may be disadvantageous to try to single out one particular method; instead following the statement from statistics that "averaging is better than voting" and using many methods in concert may lead to a more robust and reliable result.

It is notoriously hard to make benchmarks because benchmark sets of sequences often will overlap strongly with the sequences that went into the construction of some of the algorithms. Some approaches will be created with an inherent ability to "remember" the training data, while others are designed to extract only the average and generalizable features. For such methods the performance on the training set will only about reach the performance on a test set.

As described in Chapter 1 (Section 1.2), the amount of sequence data grows exponentially. Fortunately, the computing power in a typical PC or workstation also grows exponentially and, moreover, is available at ever-decreasing cost. For a long time computers have been getting twice as fast whenever the cost has been reduced roughly by a factor of two. This means that every six to ten months it gets twice as expensive, in terms of the economical cost, to perform the same search against the public databases using a query sequence or a regular expression. This means also that algorithms should constantly be redesigned in order to maintain the status quo.

13.2 Databases over Databases and Tools

In the area of biological sequence analysis there is a long tradition of creating databases over databases as a means for establishing an overview as well as for managing access to the vast number of resources. One of the earliest ones was the LiMB database (Listing of Molecular Biology databases), which has been published in hard copy [353]. Today, the only reasonable medium is the more flexible World Wide Web (WWW). Links can be followed and updated instantly. LiMB contains information about the contents and details of maintenance of databases related to molecular biology. It was created to facilitate the process of locating and accessing data sets upon which the research community depends.

The following sections contain lists of links to databases over databases, to major public sequence databases, and to selected prediction servers. Realistically, these lists should be updated on a daily basis, and the goal has not been to provide a nearly complete guide to the WWW. Rather, this material

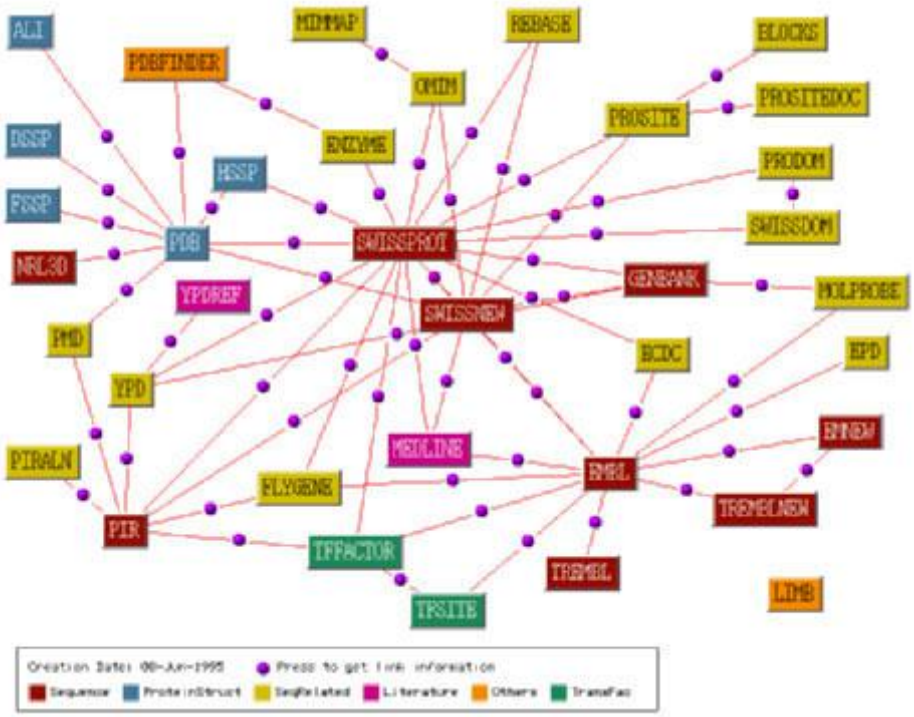

Figure 13.1: Some of the Databases Available over the World Wide Web.

should be seen as examples of the kinds of tools that can be useful for serious analysis of experimental data. It is recommended that the metadatabases be browsed regularly and that the common WWW search engines be used to spot the most recent material. Most of the links listed below come from the page started by Jan Hansen (http://www.cbs.dtu.dk/biolink.html) at the Center for Biological Sequence Analysis in Denmark. The links indicated below focus on sequence and annotation retrieval. Dedicated sites for sequence submission have not been included.

13.3 Databases over Databases in Molecular Biology

SRS Sequence Retrieval System (network browser for databanks in molecular biology)
http://www.embl-heidelberg.de/srs5/

Survey of Molecular Biology Databases and Servers
http://www.ai.sri.com/people/pkarp/mimbd/rsmith.html

BioMedNet Library
http://biomednet.com

DBGET Database Links
http://www.genome.ad.jp/dbget/dbget.links.html

Harvard Genome Research Databases and Selected Servers
http://golgi.harvard.edu

Johns Hopkins Univ. OWL Web Server
http://www.gdb.org/Dan/proteins/owl.html

Index of Biology Internet Servers, USGS
http://info.er.usgs.gov/network/science/biology/index.html

Listing of Molecular Biology Databases (LiMB)
gopher://gopher.nih.gov/11/molbio/other

WWW Server for Virology, UW-Madison
http://www.bocklabs.wisc.edu/Welcome.html

UK MRC Human Genome Mapping Project Resource Centre
http://www.hgmp.mrc.ac.uk/

WWW for the Molecular Biologists and Biochemists
http://www.yk.rim.or.jp/~aisoai/index.html

Links to other Bio-Web servers
http://www.gdb.org/biolinks.html

Molecular Modelling Servers and Databases
http://www.rsc.org/lap/rsccom/dab/ind006links.htm

EMBO Practical Structural Databases
http://xray.bmc.uu.se/embo/structdb/links.html

Web Resources for Protein Scientists
http://www.faseb.org/protein/ProSciDocs/WWWResources.html

ExPASy Molecular Biology Server
http://expasy.hcuge.ch/cgi-bin/listdoc

The Antibody Resource Page
http://www.antibodyresource.com

Bioinformatics WWW Sites
http://biochem.kaist.ac.kr/bioinformatics.html

Bioinformatics and Computational Biology at George Mason University
http://www.science.gmu.edu/~michaels/Bioinformatics/

INFOBIOGEN Catalog of Databases
http://www.infobiogen.fr/services/dbcat/

National Biotechnology Information Facility
http://www.nbif.org/data/data.html

Human Genome Project Information
http://www.ornl.gov/TechResources/Human_Genome

Archives for biological software and databases
http://www.gdb.org/Dan/software/biol-links.html

Proteome Research: New Frontiers in Functional Genomics (book contents)
http://expasy.hcuge.ch/ch2d/LivreTOC.html

13.4 Sequence and Structure Databases

13.4.1 Major Public Sequence Databases

EMBL WWW Services
http://www.EMBL-heidelberg.de/Services/index.html

GenBank Database Query Form (get a GenBank entry)
http://ncbi.nlm.nih.gov/genbank/query_form.html

Protein Data Bank WWW Server (get a PDB structure)
http://www.rcsb.org

European Bioinformatics Institute (EBI)
http://www.ebi.ac.uk/

EBI Industry support
http://industry.ebi.ac.uk/

SWISS-PROT (protein sequences)
http://www.expasy.ch/sprot/sprot-top.html

PROSITE (functional protein sites)
http://expasy.hcuge.ch/sprot/prosite.html

Macromolecular Structures Database
http://BioMedNet.com/cgi-bin/members1/shwtoc.pl?J:mms

Molecules R Us (search and view a protein molecule)
http://cmm.info.nih.gov/modeling/net_services.html

PIR-International Protein Sequence Database
http://www.gdb.org/Dan/proteins/pir.html

SCOP (structural classification of proteins), MRC
http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.1.html

HIV Sequence Database, Los Alamos
http://hiv-web.lanl.gov/

HIV Molecular Immunology Database, Los Alamos
http://hiv-web.lanl.gov/immuno/index.html

TIGR Database
http://www.tigr.org/tdb/tdb.html

The NCBI WWW Entrez Browser
http://www.ncbi.nlm.nih.gov/Entrez/index.html

Cambridge Structural Database (small-molecule organic and organometallic crystal structures)
http://www.ccdc.cam.ac.uk

Gene Ontology Consortium
http://genome-www.stanford.edu/GO/

13.4.2 Specialized Databases

ANU Bioinformatics Hypermedia Server
(virus databases, classification and nomenclature of viruses)
http://life.anu.edu.au/

O-GLYCBASE (a revised database of O-glycosylated proteins)
http://www.cbs.dtu.dk/OGLYCBASE/cbsoglycbase.html

Genome Sequence Database (GSDB) (relational database of annotated DNA sequences)
http://www.ncgr.org

EBI Protein topology atlas
http://www3.ebi.ac.uk/tops/ServerIntermed.html

Database of Enzymes and Metabolic Pathways (EMP)
http://www.empproject.com/

MAGPIE (multipurpose automated genome project investigation environment)
http://www.mcs.anl.gov/home/gaasterl/magpie.html

E.coli database collection (ECDC) (compilation of DNA sequences of *E. coli* K12)
http://susi.bio.uni-giessen.de/ecdc.html

Haemophilus influenzae database (HIDC) (genetic map, contigs, searchable index)
http://susi.bio.uni-giessen.de/hidc.htm

EcoCyc: Encyclopedia of *Escherichia coli* Genes and Metabolism
http://www.ai.sri.com/ecocyc/ecocyc.html

Eddy Lab snoRNA Database
http://rna.wustl.edu/snoRNAdb/

GenProtEc (genes and proteins of *Escherichia coli*)
http://www.mbl.edu/html/ecoli.html

NRSub (non-redundant database for Bacillus subtilis)
http://pbil.univ-lyon1.fr/nrsub/nrsub.html

YPD (proteins from *Saccharomyces cerevisiae*)
http://www.proteome.com/YPDhome.html

Saccharomyces Genome Database
http://genome-www.stanford.edu/Saccharomyces/

LISTA, LISTA-HOP and LISTA-HON (compilation of homology databases from yeast)
http://www.ch.embnet.org/

FlyBase (*Drosophila* database)
http://flybase.bio.indiana.edu/

MPDB (molecular probe database)
http://www.biotech.ist.unige.it/interlab/mpdb.html

Compilation of tRNA sequences and sequences of tRNA genes
http://www.uni-bayreuth.de/departments/biochemie/trna/index.html

Small RNA database, Baylor College of Medicine
http://mbcr.bcm.tmc.edu/smallRNA/smallrna.html

SRPDB (signal recognition particle database)
http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html

RDP (the Ribosomal Database Project)
http://rdpwww.life.uiuc.edu/

Structure of small ribosomal subunit RNA
http://rrna.uia.ac.be/ssu/index.html

Structure of large ribosomal subunit RNA
http://rrna.uia.ac.be/lsu/index.html

RNA modification database
http://medlib.med.utah.edu/RNAmods/

HAMSTeRS (haemophilia A mutation database) and factor VIII mutation database
http://europium.csc.mrc.ac.uk/usr/WWW/WebPages/main.dir/main.htm

Haemophilia B (point mutations and short additions and deletions)
ftp://ftp.ebi.ac.uk/pub/databases/haemb/

Human p53, hprt and lacZ genes and mutations
http://sunsite.unc.edu/dnam/mainpage.html

PAH mutation analysis (disease-producing human PAH loci)
http://www.mcgill.ca/pahdb

ESTHER (cholinesterase gene server)
http://www.ensam.inra.fr/cgi-bin/ace/index

IMGT (immunogenetics database)
http://www.ebi.ac.uk/imgt/

p53 mutations in human tumors and cell lines
ftp://ftp.ebi.ac.uk/pub/databases/p53/

Androgen receptor gene mutations database
ftp://www.ebi.ac.uk/pub/databases/androgen/

Glucocorticoid receptor resource
http://nrr.georgetown.edu/GRR/GRR.html

Thyroid hormone receptor resource
http://xanadu.mgh.harvard.edu//receptor/trrfront.html

16SMDB and 23SMDB (16S and 23S ribosomal RNA mutation database)
http://www.fandm.edu/Departments/Biology/Databases/RNA.html

MITOMAP (human mitochondrial genome database)
http://www.gen.emory.edu/mitomap.html

SWISS-2DPAGE (database of two-dimensional polyacrylamide gel electrophoresis)
http://expasy.hcuge.ch/ch2d/ch2d-top.html

PRINTS (protein fingerprint database)
http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html

KabatMan (database of antibody structure and sequence information)
http://www.bioinf.org.uk/abs/

ALIGN (compendium of protein sequence alignments)
http://www.biochem.ucl.ac.uk/bsm/dbbrowser/ALIGN/ALIGN.html

CATH (protein structure classification system)
http://www.biochem.ucl.ac.uk/bsm/cath/

ProDom (protein domain database)
http://protein.toulouse.inra.fr/

Blocks database (system for protein classification)
http://blocks.fhcrc.org/

HSSP (homology-derived secondary structure of proteins)
http://www.sander.embl-heidelberg.de/hssp/

FSSP (fold classification based on structure-structure alignment of proteins)
http://www2.ebi.ac.uk/dali/fssp/fssp.html

SBASE protein domains (annotated protein sequence segments)
http://www.icgeb.trieste.it/~sbasesrv/

TransTerm (database of translational signals)
http://uther.otago.ac.nz/Transterm.html

GRBase (database linking information on proteins involved in gene regulation)
http://www.access.digex.net/~regulate/trevgrb.html

ENZYME (nomenclature of enzymes)
http://www.expasy.ch/enzyme/

REBASE (database of restriction enzymes and methylases)
http://www.neb.com/rebase/

RNaseP database
http://jwbrown.mbio.ncsu.edu/RNaseP/home.html

REGULONDB (database on transcriptional regulation in E. coli)
http://www.cifn.unam.mx/Computational_Biology/regulondb/

TRANSFAC (database on transcription factors and their DNA binding sites)
http://transfac.gbf.de/

MHCPEP (database of MHC-binding peptides)
http://wehih.wehi.edu.au/mhcpep/

Mouse genome database
http://www.informatics.jax.org/mgd.html

Mouse knockout database
http://BioMedNet.com/cgi-bin/mko/mkobrwse.pl

ATCC (American type culture collection)
http://www.atcc.org/

Histone sequence database of highly conserved nucleoprotein sequences
http://www.ncbi.nlm.nih.gov/Baxevani/HISTONES

3Dee (database of protein structure domain definitions)
http://barton.ebi.ac.uk/servers/3Dee.html

InterPro (integrated resource of protein domains and functional sites)
http://www.ebi.ac.uk/interpro/

NRL_3D (sequence-structure database derived from PDB, pictures and searches)
http://www.gdb.org/Dan/proteins/nrl3d.html

VBASE human variable immunoglulin gene sequences
http://www.mrc-cpe.cam.ac.uk/imt-doc/public/INTRO.html

GPCRD (G protein-coupled receptor data)
http://www.gpcr.org/7tm/

Human Cytogenetics (chromosomes and karyotypes)
http://www.selu.com/bio/cyto/human/index.html

Protein Kinase resource
http://www.sdsc.edu/projects/Kinases/pkr/pk_info.html#Format

Carbohydrate databases
http://www.boc.chem.ruu.nl/sugabase/databases.html

Borrelia Molecular Biology Home Page
http://www.pasteur.fr/Bio/borrelia/Welcome.html

Human papillomaviruses database
http://HPV-web.lanl.gov/

Human 2-D PAGE databases for proteome analysis in health and disease
http://biobase.dk/cgi-bin/celis

DBA mammalian genome size database
http://www.unipv.it/~webbio/dbagsh.htm

DOGS database Of Genome Sizes
http://www.cbs.dtu.dk/databases/DOGS/index.html

U.S. patent citation database
http://cos.gdb.org/repos/pat/

13.5 Sequence Similarity Searches

Sequence similarity search page at EBI
http://www.ebi.ac.uk/searches/searches.html

NCBI: BLAST notebook
http://www.ncbi.nlm.nih.gov/BLAST/

BLITZ ULTRA Fast Search at EMBL
http://www.ebi.ac.uk/searches/blitz_input.html

EMBL WWW services
http://www.embl-heidelberg.de/Services/index.html#5

Pattern scan of proteins or nucleotides
http://www.mcs.anl.gov/compbio/PatScan/HTML/patscan.html

MEME (motif discovery and search)
http://meme.sdsc.edu/meme/website/

CoreSearch (dentification of consensus elements in DNA sequences)
http://www.gsf.de/biodv/coresearch.html

The PRINTS/PROSITE scanner (search motif databases with query sequence)
http://www.biochem.ucl.ac.uk/cgi-bin/attwood/SearchPrintsForm.pl

DARWIN system at ETH Zurich
http://cbrg.inf.ethz.ch/

PimaII find sequence similarity using dynamic programming
http://bmerc-www.bu.edu/protein-seq/pimaII-new.html

DashPat find sequence similarity using a hashcode comparison with a pattern library
http://bmerc-www.bu.edu/protein-seq/dashPat-new.html

PROPSEARCH (search based on amino acid composition, EMBL)
http://www.embl-heidelberg.de/aaa.html

Sequence search protocol (integrated pattern search)
http://www.biochem.ucl.ac.uk/bsm/dbbrowser/protocol.html

ProtoMap (automatic hierarchical classification of all swissprot proteins)
http://www.protomap.cs.huji.ac.il/

GenQuest (Fasta, Blast, Smith Waterman; search in any database)
http://www.gdb.org/Dan/gq/gq.form.html

SSearch (searches against a specified database)
http://watson.genes.nig.ac.jp/homology/ssearch-e_help.html

Peer Bork search list (motif/pattern/profile searches)
http://www.embl-heidelberg.de/~bork/pattern.html

PROSITE Database Searches (search for functional sites in your sequence)
http://www.ebi.ac.uk/searches/prosite.html

PROWL—Protein Information Retrieval at Skirball Institute
http://mcphar04.med.nyu.edu/index.html

CEPH genotype database
http://www.cephb.fr/cephdb/

13.6 Alignment

13.6.1 Pairwise Sequence and Structure Alignment

Pairwise protein alignment (SIM)
http://expasy.hcuge.ch/sprot/sim-prot.html

LALNVIEW alignment viewer program
ftp://expasy.hcuge.ch/pub/lalnview

BCM Search Launcher (pairwise sequence alignment)
http://searchlauncher.bcm.tmc.edu/seq-search/alignment.html

DALI compare protein structures in 3D
http://www2.ebi.ac.uk/dali/

DIALIGN (aligment program without explicit gap penalties)
http://www.gsf.de/biodv/dialign.html

13.6.2 Multiple Alignment and Phylogeny

ClustalW (multiple sequence alignment at BCM)
http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html

PHYLIP (programs for inferring phylogenies)
http://evolution.genetics.washington.edu/phylip.html

Other phylogeny programs, a complication from PHYLIP documentation
http://expasy.hcuge.ch/info/phylogen.sof

Tree of Life Home Page (information about phylogeny and biodiversity)
http://phylogeny.arizona.edu/tree/phylogeny.html

Links for Palaeobotanists
http://www.uni-wuerzburg.de/mineralogie/palbot1.html

Phylogenetic analysis programs (the tree of life list)
http://phylogeny.arizona.edu/tree/programs/programs.html

Cladistics
http://www.kheper.auz.com/gaia/biosphere/systematics/cladistics.htm

Cladistic software (a list from the Willi Hennig Society)
http://www.cladistics.org/education.html

BCM search launcher for multiple sequence alignments
http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html

AMAS (analyse multiply aligned sequences)
http://barton.ebi.ac.uk/servers/amas_server.html

Vienna RNA Secondary Structure Package
http://www.tbi.univie.ac.at/~ivo/RNA/

WebLogo (sequence logo)
http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi

Protein sequence logos using relative entropy
http://www.cbs.dtu.dk/gorodkin/appl/plogo.html

RNA structure-sequence logo
http://www.cbs.dtu.dk/gorodkin/appl/slogo.html

RNA mutual information plots
http://www/gorodkin/appl/MatrixPlot/mutRNA/

13.7 Selected Prediction Servers

13.7.1 Prediction of Protein Structure from Sequence

PHD PredictProtein server for secondary structure, solvent accesibility, and transmembrane segments
http://www.embl-heidelberg.de/predictprotein/predictprotein.html

PhdThreader (fold recognition by prediction-based threading)
http://www.embl-heidelberg.de/predictprotein/phd_help.html

PSIpred (protein strcuture prediction server)
http://insulin.brunel.ac.uk/psipred/

THREADER (David Jones)
http://www.biochem.ucl.ac.uk/~jones/threader.html

TMHMM (prediction of transmembrane helices in proteins)
http://wwwcbs.dtu.dk/services/TMHMM/

Protein structural analysis, BMERC
http://bmerc-www.bu.edu/protein-seq/protein-struct.html

Submission form for protein domain and foldclass prediction
http://genome.dkfz-heidelberg.de/nnga/def-query.html

NNSSP (prediction of protein secondary sturcture by nearest-neighbor algorithms)
http://genomic.sanger.ac.uk/pss/pss.html

Swiss-Model (automated knowledge-based protein homology modeling server)
http://www.expasy.ch/swissmod/SWISS-MODEL.html

SSPRED (secondary structure prediction with multiple alignment)
http://www.mrc-cpe.cam.ac.uk/jong/predict/sspred.htm

SSCP (secondary structure prediction content with amino acid composition)
http://www.mrc-cpe.cam.ac.uk/jong/predict/sscp.htm

SOPM (Self Optimized Prediction Method, secondary structure) at IBCP, France.
http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopm.html

NNPREDICT (neural network for residue-by-residue prediction)
http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html

SSpro (secondary structure in 3 classes)
http://promoter.ics.uci.edu/BRNN-PRED/

SSpro8 (secondary structure in 8 classes)
http://promoter.ics.uci.edu/BRNN-PRED/

ACCpro (solvent accessibility)
http://promoter.ics.uci.edu/BRNN-PRED/

CONpro (contact number)
http://promoter.ics.uci.edu/BRNN-PRED/

TMAP (service predicting transmembrane segments in proteins)
http://www.embl-heidelberg.de/tmap/tmap_info.html

TMpred (prediction of transmembrane regions and orientation)
http://www.ch.embnet.org/software/TMPRED_form.html

MultPredict (secondary structure of multiply aligned sequences)
http://kestrel.ludwig.ucl.ac.uk/zpred.html

NIH Molecular Modeling Homepage (modelling homepage with links)
http://cmm.info.nih.gov/modeling/

BCM Search Launcher (protein secondary structure prediction)
http://searchlauncher.bcm.tmc.edu/seq-search/struc-predict.html

COILS (prediction of coiled coil regions in proteins)
http://www.ch.embnet.org/software/coils/COILS_doc.html

Coiled Coils
http://www.york.ac.uk/depts/biol/units/coils/coilcoil.html

Paircoil (location of coiled coil regions in amino acid sequences)
http://theory.lcs.mit.edu/ bab/webcoil.html

PREDATOR (protein secondary structure prediction from single sequence)
http://www.embl-heidelberg.de/argos/predator/predator_info.html

DAS (Dense Alignment Surface; prediction of transmembrane regions in proteins)
http://www.biokemi.su.se/~server/DAS/

Fold-recognition at UCLA-DOE structure prediction server
http://www.doe-mbi.ucla.edu/people/frsvr/frsvr.html

Molecular Modelling Servers and Databases
http://bionmr5.bham.ac.uk/modelling/model.html

EVA (automatic evaluation of protein structure prediction servers)
http://cubic.bioc.columbia.edu/eva/

13.7.2 Gene Finding and Intron Splice Site Prediction

NetGene (prediction of intron splice sites in human genes)
http://www.cbs.dtu.dk/services/NetGene2/

NetPlantGene (prediction of intron splice sites in *Arabidopsis thaliana*)
http://www.cbs.dtu.dk/services/NetPGene

GeneQuiz (automated analysis of genomes)
http://www.sander.embl-heidelberg.de/genequiz/

GRAIL interface (protein coding regions and functional sites)
http://avalon.epm.ornl.gov/Grail-bin/EmptyGrailForm

GENEMARK (WWW system for predicting protein coding regions)
http://genemark.biology.gatech.edu/GeneMark

GENSCAN Web Server: Complete gene structures in genomic DNA
http://gnomic.stanford.edu/~chris/GENSCANW.html

FGENEH Genefinder: Prediction of gene structure in human DNA sequences
http://mbcr.bcm.tmc.edu/Guide/Genefinder/fgeneh.html

GRAIL and GENQUEST (E-mail sequence analysis, gene assembly, and sequence comparison)
http://avalon.epm.ornl.gov/manuals/grail-genquest.9407.html

CpG islands finder
http://www.ebi.ac.uk/cpg/

Eukaryotic Pol II promoter prediction
http://biosci.umn.edu/software/proscan.html

Promoter prediction input form
http://www-hgc.lbl.gov/projects/promoter.html

Web Signal Scan Service (scan DNA sequences for eukaryotic transcriptional elements)
http://bimas.dcrt.nih.gov/molbio/signal/

Gene Discovery Page
http://konops.imbb.forth.gr/~topalis/mirror/gdp.html

List of genome sequencing projects
http://www.mcs.anl.gov/home/gaasterl/genomes.html

13.7.3 DNA Microarray Data and Methods

Cyber-T (DNA microarray data analysis server)
http://128.200.5.223/CyberT/

Brown Lab guide to microarraying
http://cmgm.stanford.edu/pbrown

Stanford Microarray Database
http://genome-www4.stanford.edu/MicroArray/SMD/

Stanford MicroArray Forum
http://cmgm.stanford.edu/cgi-bin/cgiwrap/taebshin/dcforum/dcboard.cgi

Brazma microarray page at EBI
http://industry.ebi.ac.uk/~brazma/Data-mining/microarray.html

Web resources on gene expression and DNA microarray technologies
http://industry.ebi.ac.uk/~alan/MicroArray/

Gene-X (array data management and analysis system)
http://www.ncgr.org/research/genex/

UCI functional genomics array tools and software
http://www.genomics.uci.edu/

Matern's DNA Microarray Page
http://barinth.tripod.com/chips.html

Public source for microarraying information, tools, and protocols
http://www.microarrays.org/

Weisshaar's listing of DNA microarray links
http://www.mpiz-koeln.mpg.de/~weisshaa/Adis/DNA-array-links.html

DNA microarray technology to identify genes controlling spermatogenesis
http://www.mcb.arizona.edu/wardlab/microarray.html

13.7.4 Other Prediction Servers

NetStart (translation start in vertebrate and A. thaliana DNA)
http://www.cbs.dtu.dk/services/NetStart/

NetOGlyc (O-glycosylation sites in mammalian proteins)
http://www.cbs.dtu.dk/services/NetOGlyc/

YinOYang (O-β-GlcNAc sites in eukaryotic protein sequences)
http://www.cbs.dtu.dk/services/YinOYang/

SignalP
(signal peptide and cleavage sites in gram+, gram-, and eukaryotic proteins)
http://www.cbs.dtu.dk/services/SignalP/

NetChop (cleavage sites of the human proteasome)
http://www.cbs.dtu.dk/services/NetChop/

NetPhos (serine, threonine and tyrosine phosphorylation sites in eukaryotic proteins)
http://www.cbs.dtu.dk/services/NetPhos/

TargetP (prediction of subcellular location)
http://www.cbs.dtu.dk/services/TargetP/

ChloroP (chloroplast pransit peptide prediction)
http://www.cbs.dtu.dk/services/SignalP/

PSORT (prediction of protein-sorting signals and localization from sequence)
http://psort.nibb.ac.jp/

PEDANT (prtein extraction, description, and analalysis tool)
http://pedant.mips.biochem.mpg.de/

Compare your sequence to COG database
http://www.ncbi.nlm.nih.gov/COG/cognitor.html

Prediction of HLA-binding peptides from sequences
http://www-bimas.dcrt.nih.gov/molbio/hla_bind/index.html

13.8 Molecular Biology Software Links

Visualization for bioinformatics
http://industry.ebi.ac.uk/ alan/VisSupp/

The EBI molecular biology software archive
http://www.ebi.ac.uk/software/software.html

The BioCatalog
http://www.ebi.ac.uk/biocat/e-mail_Server_ANALYSIS.html

Archives for biological software and databases
http://www.gdb.org/Dan/softsearch/biol-links.html

Barton group software (ALSCRIPT, AMPS, AMAS, STAMP, ASSP, JNET, and SCANPS)
http://barton.ebi.ac.uk/new/software.html

Cohen group software rotamer library, BLoop, QPack, FOLD, Match,
http://www.cmpharm.ucsf.edu/cohen/pub/

Bayesian bioinformatics at Wadsworth Center
http://www.wadsworth.org/res&res/bioinfo/

Rasmol software and script documentation
http://scop.mrc-lmb.cam.ac.uk/std/rs/

MolScript
http://ind1.mrc-lmb.cam.ac.uk/external-file-copies/molscript.html

WHAT IF
http://www.hgmp.mrc.ac.uk/Registered/Option/whatif.html

Biosym (Discover)
http://ind1.mrc-lmb.cam.ac.uk/external-file-copies/biosym/
discover/html/Disco_Home.html

SAM software for sequence consensus HMMs at UC Santa Cruz
http://www.cse.ucsc.edu/research/compbio/sam.html

HMMER (source code for hidden Markov model software)
http://hmmer.wustl.edu/

ClustalW
http://www.ebi.ac.uk/clustalw/

DSSP program
http://www.sander.embl-heidelberg.de/dssp/

Bootscanning for viral recombinations
http://www.bio.net//hypermail/RECOMBINATION/recom.199607/0004.html

Blocking Gibbs sampling for linkage analysis in very large pedigrees
http://www.cs.auc.dk/~claus/block.html

ProMSED (protein multiple sequences editor for Windows)
ftp://ftp.ebi.ac.uk/pub/software/dos/promsed/

DBWatcher for Sun/Solaris
http://www-igbmc.u-strasbg.fr/BioInfo/LocalDoc/DBWatcher/

ProFit (protein least squares fitting software)
http://www.bioinf.org.uk/software/

Indiana University IUBIO software and data
http://iubio.bio.indiana.edu/

Molecular biology software list at NIH
http://bimas.dcrt.nih.gov/sw.html

ProAnalyst software for protein/peptide analysis
ftp://ftp.ebi.ac.uk/pub/software/dos/proanalyst/

DRAGON protein modelling tool using distance geometry
http://www.nimr.mrc.ac.uk/~mathbio/a-aszodi/dragon.html

Molecular Surface Package
http://www.best.com/~connolly/

Biotechnological Software and Internet Journal
http://www.orst.edu/~ahernk/bsj.html

MCell (Monte Carlo simulator of cellular microphysiology)
http://www.mcell.cnl.salk.edu/

HHMpro (HMM simulator for sequence analysis with graphical interface)
http://www.netid.com/html/hmmpro.html

13.9 Ph.D. Courses over the Internet

Biocomputing course resource list: course syllabi
http://www.techfak.uni-bielefeld.de/bcd/Curric/syllabi.html

Ph.D. course in biological sequence analysis and protein modeling
http://www.cbs.dtu.dk/phdcourse/programme.html

The Virtual School of Molecular Sciences
http://www.ccc.nottingham.ac.uk/vsms/sbdd/

EMBnet Biocomputing Tutorials
http://biobase.dk/Embnetut/Universl/embnettu.html

Collaborative course in protein structure
http://www.cryst.bbk.ac.uk/PPS/index.html

GNA's Virtual School of Natural Sciences
http://www.techfak.uni-bielefeld.de/bcd/Vsns/index.html

Algorithms in molecular biology
http://www.cs.washington.edu/education/courses/590bi/

ISCB education working group
http://www.sdsc.edu/pb/iscb/iscb-edu.html

13.10 Bioinformatics Societies

International Society for Computational Biology (ISCB)
http://www.iscb.org/

Society for Bioinformatics in the Nordic countries
http://www.socbin.org/

Japanese Society for Bioinformatics
http://www.jsbi.org/

13.11 HMM/NN simulator

A number of projects described in the book have been carried using the machine learning software environment for biological sequence analysis developed in collaboration by Net-ID, Inc. and employees at the Danish Center for Biological Sequence Analysis in Copenhagen.

The foundation for the software environment is based on NetLibs, an object-oriented library of C++ classes for graphical modeling, machine learning, and inference developed by Net-ID. The library supports the hierarchical and recursive implementation of any graphical model (NNs, HMMs, Bayesian Networks, etc.) together with general local message-passing algorithms for the propagation of information, errors, and evidence during inferential/learning processes and dynamic programming.

Net-Libs provides, among other things, the foundation, for an HMM simulator and an NN simulator for biological sequence analysis. The easy-to-use graphical interface for both simulators is in Java. The software environment runs both under Unix and NT platforms.

In addition, the software environment contains facilities for manipulating input/output sequences, databases, and files, as well as libraries of trained models. The libraries include HMMs for a number of protein families and DNA elements (promoters, splice sites, exons, etc.) and a number of NNs for the detection of particular structural or functional signals, both in protein and DNA sequences.

For more information please contact: admin@netid.com.

This page intentionally left blank