A2-D2

Usage of Bioinformatics Tools and Identification of Information Sources

Project title:	Reasoning on the Web with Rules and Semantics
Project acronym:	REWERSE
Project number:	IST-2004-506779
Project instrument:	EU FP6 Network of Excellence (NoE)
Project thematic priority:	Priority 2: Information Society Technologies (IST)
Document type:	D (deliverable)
Nature of document:	R (report)
Dissemination level:	PU (public)
Document number:	IST506779/Dresden/A2-D2/D/PU/b1
Responsible editors:	Pedro Barahona
Reviewers:	Michael Schroeder and Tim Furche
Contributing participants:	Bucarest, Dresden, Edinburgh, Jena, Linköping, Lisbon, Paris, Rostock (corresponding partner), Skövde
Contributing workpackages:	A2
Contractual date of deliverable:	28 February 2005
Actual submission date:	7 March 2005

Abstract

Bioinformatics is an important application area for semantic web technologies as much of the data is online and accessible in XML format, as some sites already support web services, and as ontologies are widely used to annotate data. In this deliverable, we give a survey over 18 of the most important bioinformatics resources and discuss their availability and accessibility, which are two of the main criteria for these resources to act as bases for later demonstrators.

Keyword List

Bioinformatics, rules, reasoning, ontologies, sequence, structure, networks, data sources, XML, web services

Project co-funded by the European Commission and the Swiss Federal Office for Education and Science within the Sixth Framework Programme.

Usage of Bioinformatics Tools and Identification of Information Sources

Rolf Backofen[Jen], **Liviu Badea**[Buc], **Pedro Barahona**[Lis], **Albert Burger**[Edi], **Gihan Dawelbait**[Dre], **Andreas Doms**[Dre], **Franois Fages**[Par], **Anca Hotaran**[Buc], **Vaida Jakonienė**[Lin], **Ludwig Krippahl**[Lis], **Patrick Lambrix**[Lin], **Kenneth McLeod**[Edi], **Steffen Möller**[Ros], **Werner Nutt**[Edi], **Bjorn Olsson**[Sko], **Michael Schroeder**[Dre], **Sylvain Soliman**[Par], **He Tan**[Lin], **Doina Tilivea**[Buc], **Sebastian Will**[Jen]

[Buc] National Institute for Research and Development in Informatics, Bucharest, Romania, [Dre] Technische Universität Dresden, Germany, [Edi] Harriot-Watt University/MRC Human Genetics Unit, Edinburgh, UK, [Jen] Friedrich-Schiller-Universität Jena, Germany, [Lin] Linköpings universitet, Sweden, [Lis] Universidade Nova de Lisboa, Portugal, [Par] INRIA Rocquencourt, France, [Ros] Universität Rostock, Germany, [Sko] University of Skovde, Sweden

7 March 2005

Abstract

Bioinformatics is an important application area for semantic web technologies as much of the data is online and accessible in XML format, as some sites already support web services, and as ontologies are widely used to annotate data. In this deliverable, we give a survey over 18 of the most important bioinformatics resources and discuss their availability and accessibility, which are two of the main criteria for these resources to act as bases for later demonstrators.

Keyword List

Bioinformatics, rules, reasoning, ontologies, sequence, structure, networks, data sources, XML, web services

Contents

1 Introduction

The objective of the working group on Bioinformatics is to identify opportunities for applying Semantic Web technology and to create a testbed for contributions from other working groups. Bioinformatics is a particularly promising area for this purpose as

- most of the data in molecular biology is online,
- there is a huge demand for data and system integration,
- most of the data is available in XML,
- some of the data is accessible web service and
- ontologies are commonly used to annotate data.

This report seeks to give an overview of some of the most important bioinformatics data sources, thus addressing the goal stated in the specification of deliverable A2-D2:

> To understand the work of biologists and bioinformaticians, information on existing technologies, tools, algorithms, and databases will be collected giving an insight into the most important information sources. They will be evaluated regarding their use to underpin the envisaged bioinformatics semantic Web. Specifically, ontologies in bioinformatics will be surveyed and evaluated. Additionally, structured and unstructured data sources will be identified and rated.

We have grouped the existing data sources in six categories, according to the type of information they provide:

1. **Sequence-based resources,** containing DNA and protein sequences in general or for specific genomes;
2. **Structure resources,** containing data on protein structures and derived data such as structure classifications;
3. **Pathway and interaction resources,** containing information on metabolic pathways and protein interaction networks;
4. **Gene expression resources,** containing information on microarray and other experiments determining which genes are expressed in which tissue;
5. **Literature resources,** containing full articles in the life sciences or pointers to scientific literature;
6. **Ontology resources,** which contain structured vocabularies for the annotation of data.

The first four categories reflect the main subareas of biological research. Here, the most fundamental task is to identify the sequential structure of the genome and the proteins encoded by it. To understand the behaviour of proteins and other biological macromolecules it is crucial to know the shape that such molecules take in three-dimensional space and related structural properties. Molecules interact with each other in complex processes, which are for instance investigated by research in metabolic pathways. Finally, genetic information within

an organism is only activated at certain times and in specific tissues, which is the topic of gene expression research. In addition to being published in the literature, results in these four areas are also made available by means of information resources on the Web. The last two categories, literature and ontology resources, are concerned with making the research results accessible to the community.

For this survey, we selected 18 of the prime bioinformatics resources to which A2 members are contributing or that they are using. The emphasis of the selection is on sequence, structure, and pathway resources, reflecting research interests of the members in alignment, structure prediction, and pathway analysis.

For each resource we first describes its subject and content. Then we collect technical information on

- the ways in which it can be accessed (e.g. HTML, FTP, web service),
- its data formats (e.g. flat file, database, XML, RDF),
- the size of the data,
- query facilities,
- update policies of the publishers of the data,
- the way it is linked to other resources.

From the this one can see how to which degree the area of bioinformatics has already adopted technologies that are relevant for a Semantic Web. This information is also needed to decide about possible scenarios for a testbed.

It turned out that most of the discussed resources provide data in the form of XML documents, some of them (such as Kegg, PubMed, dbSNP, EnsEMBL, MSD) can be accessed as web services and one resource (UniProt) is available as RDF.

Other aspects of bioinformatics tools have been discussed already in deliverable A2-D1, namely algorithms for structure prediction, sequence and structure alignment, and algorithms for metabolic pathways. A comprehensive set of ontologies such as the GeneOntology, anatomies of human and model organisms such as mouse, fruit fly, zebra fish, and C. elegans, as well as EcoCyc, MBO, and Tambis have also been covered already in A2-D1. Consequently, we will limit ourselves to the main ontologies, the GeneOntology and anatomy ontologies (the latter is listed under gene expression resources as it is part of the Edinburgh Mouse Atlas).

The main purpose of this document is to create a common reference point for A2 members for later decisions on which resources to include in demonstrators. The reader should have some familiarity with bioinformatics concepts and resources.

2 Sequence Resources

Genetic information is encoded in sequences of nucleic acids in the DNA of organisms. Moreover, certain sections of DNA are transcribed into sequences of amino acids, thus giving rise to proteins. While variations among the DNA of individuals of a single species are minimal, similar sequences can also be found in the genome of very different species. An important branch of biological research aims at identifying the sequential structure of DNA and proteins

in specific species together with the possible variations of these structures and the similarities across species.

In this section, we discuss sources with information resulting from this research. *EnsEMBL* is a database with the genome of currently 17 different species. *UniProt* is a primal resource of protein information. *FlyBase* makes available both DNA and protein data of the fruit fly Drosophila, which is one of the best studied model organisms. *dbSNP* is a resource collecting data on genetic variations such as mutations. *Pfam* contains data on the similarity of various proteins. *TRANSFAC* has information on the transcription of nucleic sequences into proteins.

2.1 EnsEMBL: A Genome Browser

EnsEMBL is a joint project between EMBL-EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on metazoan (organism has more than a single cell) genomes. The complete automation of the process led to an integration of many tools for this process. However, the annotation process is linear with no special thinking about anything because of the homogeneity of the organisms being annotated. They are mostly mammals plus a few insects, fishes and birds.

2.1.1 Accessibility

- **URL:** http://www.ensembl.org/

Accessible via web, database queries, FTP and DAS (Distributed Annotation System by Lincoln Stein), which includes web service access.

2.1.2 Query Facilities

Via (and constrained by) web interface. Arbitrary SQL queries when accessed directly. Perl and Java APIs, and Java-based EnsEMBL Mart (every problem will have its own algorithm. EnsEMBL is mainly a repository).

2.1.3 Data Formats

Text files (sequences only in FASTA format, EMBL format), MySQL databases.

2.1.4 Size of Data

Most queries are very small, particularly when asked over the web. The size of the files is 6.3G (multi-species_27.1/) and 5.8G (mart_27.1/). As for the databases their sizes are 9.9G (ensembl-mart-27_1/), 9.9G (ensembl-compara_27_1/) and 1.7G (ensembl-go_27_1/). The above does not include a table on DNA sequences. Furthermore, each organism has its own specialist database. Expect $>>$ 100GB for a full installation.

2.1.5 Update Policy

New releases monthly, with a total substitution. Changes are announced via email.

2.1.6 Links to other Sources

Protein domain databases, sequence databases (UniProt, RefSeq, EST), Affymetrix, OMIM. Arbitrary sources added to the system by DAS. Presentation of results from tools accepted for annotation (TMHMM, GenScan, Wise).

2.2 UniProt: Universal Protein Resource

UniProt (Universal Protein Resource) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR. UniProt is comprised of three components, each optimized for different uses. The UniProt Knowledgebase (UniProt) is the central access point for extensive curated protein information, including function, classification, and cross-reference. The UniProt Non-redundant Reference (UniRef) databases combine closely related sequences into a single record to speed searches. The UniProt Archive (UniParc) is a comprehensive repository, reflecting the history of all protein sequences.

2.2.1 Accessibility

- **URL:** http://www.uniprot.org/

Accessible via HTTP and FTP (mirrored throughout the world).

2.2.2 Query Facilities

Queries may be done with Unix tools combined with Perl scripts. There are specialised programs that understand the UniProt syntax and some of its semantics (e.g. SRS - cf. below). Queries may also be via sequence similarity programs like BLAST, FASTA, Smith-Waterman. A number of specialised algorithms help with some types of searches (sequence based search, gene ontology-based selection, ambiguous selection on the basis of chromosomal location or probes on microarray chip), and interpretation of the sequences (Sequence similarity, threading of sequence into presumed similar structures, ab initio structure prediction, accessibility to proteases, susceptibility to post-translational modification or other sequence patterns, likelihood to attract antibodies (epitope prediction), and many others.

2.2.3 Data Formats

Text files (full text and sequences only in FASTA format), XML. A relational database exists, but is generated from the text file. An experimental RDF version is available, too.

2.2.4 Size of Data

The gzipped full text SWISS-PROT (= human curated) has 128,154KB. The FASTA and XML SWISS-PROT have 28,222KB and 149,372KB, respectively. The full text TrEMBL (= automated annotation) has 492,937KB. The FASTA and XML TrEMBL have 28,222KB and 149,372KB, respectively.

2.2.5 Update Policy

Update Policy: Internal updates weekly. Irregular external updates. Updates consist of additions (substitutions of whole entries) with no further notion on changes (Entries are separated by double slashes and the fields within an entry are inter-dependent. Most essential is the link from the sequence to the FT lines of the entry that describe the sequence). The DT lines of an entry mention a change in the protein sequence or the protein annotation.

2.2.6 Links to other Sources

Linked from many sites like

- **EnsEMBL** (http://www.ensembl.org),
- **SRS** (http://srs.ebi.ac.uk),
- **InterPro** (http://www.ebi.ac.uk/interpro).

The SRS environment has a summary on the most important outgoing links (check in http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+LibInfo+-id+2Jb751PF567+-lib+UNIPROT). UniProt links to the Protein Domain/Family databases (Pfam, PRINTS, Prosite, InterPro), Medline, disease information (Omim), structure information (HSSP, PDB), Protein Interaction (IntAct), expression data (SWISS-2Dpage but not RNA expression data), nucleotide database (EMBL nucleotide database containing the originally submitted DNA sequence(s) from which the reported protein sequence may be derived.

2.3 FlyBase: A Database of the Drosophila Genome

The FlyBase database contains genetic and molecular data for Drosophila, primarily Drosophila melanogaster, including information on genes and mutant alleles, expression and properties of transcripts and proteins, functions of gene products, images that illustrate Drosophila anatomy and development terms, and a bibliography of Drosophila citations (for full content list see their web site). FlyBase is maintained by a consortium of researchers from Harvard University, University of Cambridge, Indiana University, University of California, and the European Bioinformatics Institute.

2.3.1 Accessibility

- **URL:** http://flybase.bio.indiana.edu

Primary access to FlyBase is through its web pages, which provide a range of query facilities.

2.3.2 Query Facilities

Queries are encoded as URLs. For automation of any query, it is recommended to carry out the query once "by hand", i.e. through the web pages, and then reuse the generated URL for further processing. URLs need to be edited to specify the data format required. Specific purpose viewers, such as GBrowse for molecularly and cytologically mapped data and Apollo for annotations and large genomic regions, are also available. Access to data for further computational processing is supported through bulk data retrieval, where queries are answered with data sets in one of several formats, e.g. tabbed, XML, etc. (see above).

2.3.3 Data Formats

FlyBase uses relational databases, currently PostgreSQL, to store its data. It includes data files, documents, indices, forms, reports and images. Through their bulk download feature, data are available in various formats, including tabbed, comma separated, Acode, HTML, plain text and XML. FlyBase is currently developing a new database structure, an implementation of Chado, the GMOD modular schema (www.gmod.org/schema).

2.3.4 Size of Data

No information available.

2.3.5 Update Policy

Regular updates take place, primarily corrections and additions. New releases are indicated by release numbering, e.g. from 4.0 to 4.1.

2.3.6 Links to other Sources

Links to others sources: FlyBase links to a number of relevant external databases. For a list of these databases, please see

- http://flybase.bio.indiana.edu/allied-data/extdb/ExternalLinks.htm.

2.4 dbSNP: Single Nucleotide Polymorphism

dbSNP is a central repository developed by the National Human Genome Research Institute and the National Center for Biotechnology Information (NCBI) in the United States for both single base nucleotide substitutions (SNPs) as well as other classes of genetic variation, such as:

- microsatellite repeats (also called short tandem repeats or STRs)
- small insertion/deletion polymorphisms (also called deletion insertion polymorphisms or DIPs).

dbSNP uses the term "SNP" in the much looser sense of "minor genetic variation," so there is no requirement or assumption about minimum allele frequencies for the polymorphisms in the database. Thus the scope of dbSNP includes disease causing clinical mutations as well as neutral polymorphisms.

2.4.1 Accessibility

- **URL:** http://www.ncbi.nlm.nih.gov/SNP/
- **URL:** ftp://ftp.ncbi.nih.gov/snp/
- **URL:** http://www.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html)

Access via web interface, ftp or NCBI Web service which enables access to Entrez Utilities (EGQuery, EFetch, EInfo, ELink, ESearch, and ESummary) via the Simple Object Access Protocol (SOAP). The service has been tested with Java (Apache Axis 1.2 RC1), MS SOAP Toolkit 3.0 and MS Word Visual Basic, and C# in MS Visual Studio .NET. The service has not been tested with clients in other languages but it should work with any language with Web services support.

It is possible to create a local copy of dbSNP, which is a relational database with about 100 tables. NCBI deploys dbSNP in both MSSQL and Sybase environments, and the public can download the full contents of the database from the dbSNP FTP site. Also, dbSNP is available for downloading in other various formats: ASN.1 binary, ASN.1 flat file, XML, and FASTA.

2.4.2 Query Facilities

There are two means of doing batch queries:

- **Searching batches submitted by individual laboratories:** This method is used to search groups, or batches, of SNPs submitted by individual laboratories
- **Submitting a batch of requests:** This method would allow a user to submit a batch of queries, or requests. The results would then be returned to the user's email account in ASN.1, FASTA, XML, chromosome report, or text flatfile format. The batch of requests can be submitted as an upload file or entered using a web interface.

In addition to batch queries, dbSNP can be searched both via other NCBI resources or directly in six different ways:

- **Using NCBI's Entrez system:** using the same approach as the other Entrez databases (search by SNP accession number, submitter SNP ID, NCBI Assay ID, or genome SNP ID); (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp)
- **By submitter:** search by submitter handle;
  (http://www.ncbi.nlm.nih.gov/SNP/snp_tableList.cgi?type=submitter)
- **New Batches:** search by local batch ID;
  (http://www.ncbi.nlm.nih.gov/SNP/snp_newbatch.cgi?searchType=newBatch)
- **Method:** search by method used by the submitter to identify the SNP;
  (http://www.ncbi.nlm.nih.gov/SNP/snp_tableList.cgi?type=method)
- **Population:** search by the type of population studied;
  (http://www.ncbi.nlm.nih.gov/SNP/snp_tableList.cgi?type=pop)
- **Publication:** search by publication title;
  (http://www.ncbi.nlm.nih.gov/SNP/snp_tableList.cgi?type=pub)
- **Chromosome Report:** reports of SNPs with STS mapping information, sorted by cRay distance where possible.
  (http://www.ncbi.nlm.nih.gov/SNP/get_html.cgi?whichHtml=./maplists/maplist)

dbSNP is also cross-linked for browsing from other NCBI resources:

- **By gene name/nomenclature association:** query results from the LocusLink database will show a purple "S" button in SNP records have been mapped to the gene. Clicking on the S will take you to a list of the reference SNP records for any gene in the LocusLink database;
- **by map location:** dbSNP is currently being integrated to GeneMap99 and the integrated physical maps that are being constructed at NCBI. When integration is completed, the maps may be browsed for SNP content in user specified regions of the map.

dbSNP can also be searched with specialised algorithms (BLAST), via a BLAST operation on dbSNP using a candidate sequence: the sequences in dbSNP are currently being formatted to be searched by BLAST. Users are able to submit a query sequence to BLAST, and receive a list of any SNPs in the database that hit the sequence.
(http://www.ncbi.nlm.nih.gov/SNP/snpblastByChr.html)

2.4.3 Data Formats

Due to the integration in the Entrez system, results can be provided in a variety of formats, including XML.

2.4.4 Size of Data

The species represented in the database are mostly Homo sapiens (10,079,771 SNPs), Gallus gallus (3,291,672), Mus musculus (581,577), Anopheles Gambiae (1,136,268) and only few entries for other species. For the Human 10,079,771 RefSNP clusters (rs#), 5,007,794 are validated, 1,045,322 contain frequency information and 1,822,844 have associated genotypes.

2.4.5 Update Policy

dbSNP is in a state of growth, both in terms of the rate of submissions, and in terms of the relational schema used by NCBI to efficiently represent both the submitted data and the results of post-submission computation. In general, the database grows at a rate of about 90 SNPs per month. It is expected to grow in erratic jumps for the next few years. Dumps are now refreshed weekly, usually on Sunday nights. A number of different "flavors" of submissions (insertions) are possible to dbSNP:

- simultaneous submission of SNP and STS data;
- submission of a new SNP;
- submission of individual genotypes for a new SNP;
- submission of genotypes for a SNP in the database;
- submitting new population frequency information on a SNP already in the database.

Records can subsequently be updated but can not be completely deleted from the database. A record can be marked as "withdrawn," however, so that a query of that SNP will indicate that the submitter has chosen to withdraw that data.

Announcements for the release of new builds and notification of corrections to existing database content are posted to a public mail list. To receive these notifications it is necessary to subscribe at http://www.ncbi.nlm.nih.gov/mailman/listinfo/dbsnp-announce.

2.4.6 Links to other Sources

Connections exist between refSNP clusters and other NCBI resources: LocusLink, UniSTS, UniGene, PubMed, dbMHC, GeneBank.

2.5 PFam: Protein Families Database of Alignments and HMMs

PFam (Protein families database of alignments and HMMs) is a database of multi-sequence alignments and profile hidden Markov models for protein sequences. This database groups amino acid sequences within proteins into families of domains. There are two databases in PFam: PFam-A, which is a curated and annotated database of domain families, and PFam-B which is similar to PFam-A but generated automatically without human intervention. PFam is a useful tool for identifying domains in a protein sequence, and finding known structures that may make good homology models for these domains, as it includes species information that help find proteins from related organisms.

2.5.1 Accessibility

- **URL:** http://www.sanger.ac.uk/Software/Pfam/

PFam is available as an Internet service.

2.5.2 Query Facilities

The user provides a sequence of IDs. The server can provide information on domain family classifications and phylogenetics.

2.5.3 Data Formats

The data format is flat text (either protein Uniprot ID or sequence). Output is HTML.

2.5.4 Size of Data

Typically $\approx$ 10KB per query.

2.5.5 Update Policy

New releases of Pfam approximately once every three months. Updates include insertion of new families, removal of families, and reclassification of families from PFam-B to PFam-A.

2.5.6 References

- "The Pfam Protein Families Database, Alex Bateman, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik L. L. Sonnhammer, David J. Studholme, Corin Yeats and Sean R. Eddy, Nucleic Acids Research (2004), Database Issue 32:D138-D141
- "The Pfam Protein Families Database Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002) Nucleic Acids Research 30(1):276-280.

2.6 TRANSFAC: Transcription Factor DataBase

TRANSFAC is a commercial (500 Euro per year for an "Academic Research Group License") database on eukaryotic transcription factors and their binding sites in genes. Its internal structure is a relational database consisting of 80 tables which are condensed to 7 flat files. Users obtain information via a web interface, which enables queries for diverse fields of the database tables. It is also possible to install TRANSFAC locally which requires running an own web server. TRANSFAC is integrated in the assembly of biological databases of Biobase in which TRANSFAC constitutes the most important part. Other Biobase products are TRANSCompel (add-on to TRANSFAC, a specialized database module on composite regulatory elements), TRANSPATH (signal transduction networks).

2.6.1 Accessibility

- **URL:** http://www.biobase.de/cgi-bin/biobase/transfac/start.cgi

Accessible via web, or download of the complete database (executable with a installed web server and the same cgi-scripts).

2.6.2 Query Facilities

Query facilities: In the form of http requests via an HTML form (combined search in different fields of a single table, conjunctive and/or disjunctive combination of search terms). The output consists of dynamically created html-files with links to the matches and free arbitrary output fields for each match. There are also a number of algorithms provided to access data, namely

- *MATCH*: Search for binding sites in an input sequence by using the available MATRIX entries
- *PATCH*: Search for binding sites in an input sequence by using individual known sites

2.6.3 Data Formats

Seven text-files (EMBL-like, but not compatible to EMBL).

2.6.4 Size of Data

About 40.4 MB, for 18,349 entries divided as follows:

- SITE (4,406 entries in 11.7 MB): individual transcription factor binding sites
- FACTOR (5,711 entries in 13.4 MB): transcription factors
- MATRIX (735 entries in 1.2 MB): nucleotide distribution matrices (PSSM) of aligned binding sequences for single transcription factors.
- GENE (4,144 entries in 7.9 MB): genes for which information (binding sites ...) is contained in TRANSFAC.
- CLASS (53 entries in 304 KB): some background information about the transcription factor classes

- CELL (1,650 entries in 743 KB): brief information about the cellular source of proteins that have been shown to interact with these sites
- REFERENCE (1,650 entries in 5.2 MB): literature references

2.6.5 Update Policy

New releases with a quarterly frequency. Updates are mostly insertion and correction of entries, but there are also changes in the user interface and in the algorithms and tools provided. Changes are announced via email (for registered users).

2.6.6 Links to other Sources

- literature references (PubMED links),
- hyperlinks to other databases (EMBL, SwissProt, EPD, PIR, Flybase, PDB, TRANSCompel, PathoDB, SMARtDB, TRANSPATH, PROSITE, CLDB).

3 Structure Resources

The chemical properties and the possible functions of a biological macromolecule depend crucially on its 3-dimensional structure. The main database that collects 3D structure information, obtained from crystallographic experiments, is the Protein Data Bank *PDB*. Other resources build upon the information in PDB. For instance, *SCOP* classifies proteins with known structure, while *STRIDE* and *DSSP* determine a variety of other, so-called secondary, structures of molecules.

3.1 PDB: Protein Data Bank

The PDB is the single worldwide repository for the processing and distribution of 3-D structure data of large molecules of proteins and nucleic acids. New structures are released each Wednesday by 1:00am Pacific time. Details about the history, function, progress, and future goals of the PDB can be found in the PDB Annual Reports and PDB Newsletters. The information contained in a PDB entry is: the protein which is subject to the PDB ID and the species it came from, the solved structure, and references to publications describing the structure determination, experimental details about the structure determination, including information related to the general quality of the result such as resolution of X-ray structure determination and stereo chemical statistics, the amino acid sequence. Additionally molecules appear in the structure, including cofactors, inhibitors, and water molecules are listed as well as assignments of secondary structure like helixes and sheets and the atomic coordinates.

3.1.1 Accessibility

- **URL:** http://www.pdb.org/

Accessible via web, download of the flat files, or querying the database via a search dialogue. The derived database MSD is also available as a web service.

3.1.2 Query Facilities

Search for a PDB ID like "a1b1" results in a summary of all information. Some web sites offer a picture of the structure (e.g. PDBsum). Keyword search for IDs, sequence alignment and multiple sequence alignment for sequences, possibly structural alignment, (which is a refinement of sequence alignment as an additional step). More information in

- **URL:** http://www.rcsb.org/pdb/info.html#Press_Releases/

3.1.3 Data Formats

Text Files (PDB has its own flat file format). There are also XML and SQL versions of the same data.

3.1.4 Size of Data

About 30GB, for about 150,000 entries at the moment, of different quality.

3.1.5 Update Policy

Weekly updates (new structures are released each Wednesday by 1:00am Pacific time). Inserts, Updates. (A list of files for structures that have become obsolete and for structures re-released for any reason during the previous quarter is included in update releases so users can update their entire set of structures. Index files in the pub/resource sub-directory include all structures in the current PDB FTP site as of the release date). Subscribers will automatically be sent their preferred format each quarter.

3.2 SCOP: Structural Classification of Proteins

The SCOP database aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known, including all entries in the Protein Data Bank (PDB). It is available as hypertext pages that offer a panoply of representations of proteins, including links to PDB entries, sequences, references, images and interactive display systems. The SCOP classification of proteins has been constructed manually by visual inspection and comparison of structures, but with the assistance of tools to make the task manageable and help provide generality. The different major levels in the hierarchy are Family (clear evolutionary relationship), Superfamily (probable common evolutionary origin), Fold (major structural similarity).

3.2.1 Accessibility

- **URL:** http://scop.mrc-lmb.cam.ac.uk/scop/
- **URL:** http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.html/

Accessible via web interface.

3.2.2 Query Facilities

Queries by key search using web server, or by hierarchy browsing through clicking on links. Answers provided as text and HTML pages. There is a glossary of terms used in the fold classification and references to fold classification methods and fold definitions. Cf. respectively

- **URL:** http://scop.mrc-lmb.cam.ac.uk/scop/gloss.html
- **URL:** http://scop.mrc-lmb.cam.ac.uk/scop/refs.html

3.2.3 Data Formats

Text files which can be easily parsed and stored as a relational database.

3.2.4 Size of Data

The size of the data files (release 1.65) is:

- **scop.cla**: 4.5 MB,
- **scop.des**: 4.3 MB,
- **scop.hie**: 1.3 MB,
- **scop.com**: 1.1 MB.

3.2.5 Update Policy

SCOP has been updated (except for 2004) almost twice a year (the latest releases were 1.65 in Dec/03, 1.63 in May/03, 1.61 in Nov/02, 1.59 in May/02, 1.57 in Jan/02 and 1.55 in Jul/01). Updates include insertions of new entries and the editing of existing entries, namely their reclassification. Updates are announced through the web page news and via email.

3.2.6 Links to other Sources

- **SSM:** Structural similarity search of SCOP
  (http://www.ebi.ac.uk/msd-srv/ssm/ssmstart.html)
- **CE:** Combinatorial Extension method for structural comparison
  (http://cl.sdsc.edu/ce.html)
- **PALI:** Pairwise and multiple alignments of SCOP families
  (http://pauling.mbu.iisc.ernet.in/"pali)
- **SUPFAM:** Structure/sequence relationships and structural similarity search of SCOP using 3dSearch
  (http://pauling.mbu.iisc.ernet.in/"supfam)
- **SA:** Structural alignment of SCOP sequences (database + server)
  (http://bioinfo.mbb.yale.edu/align/scop)
- **PINTS:** Patterns In Non-homologous Tertiary Structures
  (http://www.russell.embl.de/pints)

- **FPS:** For sequence similarity searching of SCOP
- **CATH:** Structural classification
(http://www.biochem.ucl.ac.uk/bsm/cath)
- **Dali:** Structural comparison and FSSP structural classification
(http://www.ebi.ac.uk/dali/index.html)
- **PDB:** Protein Data Bank
(http://www.rcsb.org/pdb)
- **PDB**at a Glance
(http://cmm.info.nih.gov/modeling/pdb at a glance.html)
- **3Dee:** Protein Domain Definition
(http://www.compbio.dundee.ac.uk/3Dee)
- **MSD**: Macromolecular Structure Database
(http://www.ebi.ac.uk/msd)
- **NDB:** Nucleic Acid Database
(http://ndbserver.rutgers.edu)
- Swiss-Model
(http://swissmodel.expasy.org//SWISS-MODEL.html)
- Macromolecular Motions Database
(http://www.molmovdb.org/molmovdb)
- The **PRESAGE** Database for Structural Genomics
(http://www.molmovdb.org/molmovdb)
- Genome Census
(http://bioinfo.mbb.yale.edu/genome)
- Function assignment and metabolic models
(http://wit.mcs.anl.gov/WIT2)

3.3 STRIDE: Protein Secondary Structure Assignment

STRIDE (Protein Secondary Structure Assignment from Atomic Coordinates) assigns secondary structure elements to amino acid residues in proteins from a PDB file. In addition to the H-bond patterns required by IUPAC rule 6.3, STRIDE also uses dihedral angle information to assign secondary structure elements, thus providing an alternative to DSSP.

3.3.1 Accessibility

- **URL:** http://bioweb.pasteur.fr/seqanal/interfaces/stride.html

Access via web interface.

3.3.2 Query Facilities

The query facilities are minimal. User provides the structure of the protein, and the server returns a report assigning secondary structure elements.

3.3.3 Data Formats

Data Formats: The output is a flat text file of the secondary structure assignments. The input of the algorithm is a PDB structure file.

3.3.4 Size of Data

40MB database. Input $\approx$ 100KB (structure file), Output $\approx$ 10KB.

3.3.5 Update Policy

Not applicable.

3.3.6 Links to other Sources

It uses as input PDB files.

3.3.7 References

The algorithm for assigning secondary structures is detailed in the following two articles:

- "Knowledge-based protein secondary structure assignment." Frishman D., Argos P., Proteins, Dec 1995, 23(4):566-79
- "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.", Frishman D., Argos P., Biopolymers, Dec 1983, 22(12):2577-637.

3.4 DSSP: Definition of Secondary Structure of Proteins

DSSP (Define Secondary Structure of Proteins) determines secondary structure in proteins from a PDB file, according to IUPAC rule 6.3. This program provides a standard assignment of secondary structure elements. The algorithm assigns secondary structure motifs to amino acid residues by identifying patterns of hydrogen bonds and by measuring the changes in orientation of the protein main chain.

3.4.1 Accessibility

- **URL:** http://bioweb.pasteur.fr/seqanal/interfaces/dssp-simple.html (server)
- **URL:** http://swift.cmbi.kun.nl/swift/dssp/ (database)

DSSP can be used locally, from an Internet service, or by downloading the database of the DSSP assignments for the Protein Data Bank files

3.4.2 Query Facilities

Query facilities are minimal. User provides the structure of the protein, and the server returns a report assigning secondary structure elements.

3.4.3 Data Formats

The output is a flat text file of the secondary structure assignments. The input of the algorithm is a PDB structure file.

3.4.4 Size of Data

40MB database. Input $\approx$ 100KB (structure file), Output $\approx$ 10KB

3.4.5 Update Policy

Insertion of new records in the database, but no updates seem to have been made recently.

3.4.6 Links to other Sources

It uses as input PDB files.

3.4.7 References

The algorithm is detailed in

- "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.", Kabsch W, Sander C., Biopolymers, Dec 1983, 22(12):2577-637.

4 Pathway and Interaction Resources

In this section we survey four sources that collect information on the roles that various molecules play in the processes taking place in cells, such as cell development or metabolism. It is characteristic for such processes that organic molecules undergo specific transformations and conversions, which are summarised in so-called pathways.

4.1 KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG aims to link lower-level information (genes, proteins, enzymes, reaction molecules, etc.) with higher-level information (interactions, enzymatic reactions, pathways, etc.). The basis for KEGG are three main databases:

- *GENES*, containing collections of genes for completed and partially completed genomes;
- *PATHWAY*, containing graphs representing metabolic, signaling and transport pathways;
- *LIGAND*, containing information on compounds, enzymatic molecules and enzymatic reactions.

Additional data have recently been added to these three basic databases, but they still constitute the "backbone" of KEGG.

4.1.1 Accessibility

- **URL:** http://www.genome.jp/kegg/

Web interface at the KEGG web site. Also accessible through the DBGET system. In addition, KEGG offers APIs for accessing the data over the web from your application program as a web service. This is available in Perl, Ruby, Python and Java.

4.1.2 Query Facilities

Two main procedures allow the retrieval of a KEGG entity.

- Querying via the DBGET system, allowing full text search of components in any of the KEGG databases.
- Browsing all the databases which are organised hierarchically, by organism, type of pathway (metabolism, genetic information processing, environmental information processing, etc.), involved enzyme/compound, etc.

No specialized algorithms are provided, except the ability to use common tools (like BLAST) on the retrieved objects. Complex entities like pathways are only "displayed".

4.1.3 Data Formats

Text files for basic data. GIF files for pathways, which are manually drawn, but KEGG also offers an XML interface in the form of KGML, which is a "KEGG Markup Language", designed for representation and exchange of pathway representations. Not all KEGG pathways are available as KGML, however all KGML ones, and even some other KEGG pathways are available in another XML language, namely SBML (http://www.sbml.org/) using tools such as KEGG2SBML (http://systems-biology.org/001/).

4.1.4 Size of Data

The number of entries in the October 2004 release (rel 32) was 18,812 pathways, 841,693 genes, 243 genomes, 11,881 compounds and 6,357 reactions. There are additional data categories, but the above covers most of the contents.

4.1.5 Update Policy

Daily updates, with major releases quarterly (Jan, Apr, Jul, Oct). Insertions and modifications are made by the KEGG team directly. They are summarized on the web (cf. http://www.genome.jp /kegg/docs/upd_pathway.html). Propagation is left at the entire responsibility of all linked databases.

4.1.6 Links to other Sources

There are lots of mappings between KEGG entries and other databases, either through the DBGET system (for instance for PDB) or directly in the entry description (UniProt, Colibri, RegulonDB, NCBI, ChEBI, PubChem, etc.).

4.2 IntAct: Protein Interaction Data

IntAct is a protein interaction database and analysis system. It provides a query interface and modules to analyse interaction data. It consists of the following main objects: Experiment, Interaction and Interactor. To ensure data consistency, IntAct makes intensive use of controlled vocabularies. All controlled vocabularies are available in Gene Ontology DAG-Edit format and HTML format.

4.2.1 Accessibility

- **URL:** http://www.ebi.ac.uk/intact/index.jsp/
- **URL:** ftp://ftp.ebi.ac.uk/pub/databases/intact/current/

4.2.2 Query Facilities

Queries by searching using Gene name, IntAct Ac: EBI-141, SPTR Ac:Q08162, SPTR Id: rr44 yeast, InterPro Ac, GO Id: GO, or PubMed Id. Answers are provided as HTML pages.

4.2.3 Data Formats

PSI MI XML format (http://psidev.sourceforge.net/mi/xml/doc/user/) which is a data exchange format for protein-protein interaction, not a proposed database structure.

4.2.4 Size of Data

41 files with 55,949 KB (release 30 Nov 2004)

4.2.5 Update Policy

IntAct is released monthly, normally on the first working day of each month. Updates include insertions of new objects and the date of the object that was updated last. Updates are announced through the web page news.

4.2.6 Links to other Sources

- **EBI:** European Bioinformatics Institute, Cambridge, United Kingdom (http://www.ebi.ac.uk/index.html)
- **MPI-MG:** Max-Planck-Institut fuer Molekulare Genetik, Berlin, Germany (http://www.molgen.mpg.de/)
- **SIB:** Swiss Institute of Bioinformatics, Geneva, Switzerland (http://www.isb-sib.ch/)
- **SDU:** Odense University, Denmark (http://www.sdu.dk/)
- **UB1:** University of Bordeaux, France
- **CNB-CSIC:** Centro Nacional de Biotecnologia, Madrid, Spain (http://www.cnb.uam.es/)
- **HUJI:** The Hebrew University of Jerusalem, Israel (http://www.huji.ac.il/)

- **GSK:** GlaxoSmithKline plc, Harlow, United Kingdom (http://www.gsk.com/index.htm)
- **MINT:** University Tor Vergata, Rome, Italy (http://mint.bio.uniroma2.it/mint/)

4.2.7 References

- "IntAct—an open source molecular interaction database." H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstor, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, R. Apweiler. Nucleic Acids Research 2004 32: D452-D455.

4.3 BIND: Biomolecular Interaction Network Database

The Biomolecular Interaction Network Database (BIND) is a collection of records documenting molecular interactions. The contents of BIND include high-throughput data submissions and hand-curated information gathered from the scientific literature. BIND is an interaction database with three classifications for molecular associations: molecules that associate with each other to form interactions, molecular complexes that are formed from one or more interaction(s) and pathways that are defined by a specific sequence of two or more interactions. A BIND record represents an interaction between two or more objects that is believed to occur in a living organism. A biological object can be a protein, DNA, RNA, ligand, molecular complex, gene, photon or an unclassified biological entity. BIND records are created for interactions which have been shown experimentally and published in at least one peer-reviewed journal. A record also references any papers with experimental evidence that support or dispute the associated interaction. Interactions are the basic units of BIND and can be linked together to form molecular complexes or pathways.

4.3.1 Accessibility

- **URL:** http://bind.ca/

Access via web interface. The database can be imported into a relational data base.

4.3.2 Query Facilities

Queries by browsing the contents of the BIND database, or searching BIND data by using an identifier (such as PubMed Id, GenInfo Id, PDB Id, GO Id) or using a a simple text query, or building a field specific query. Answers as HTML pages

4.3.3 Data Formats

Flat files, XML.

4.3.4 Size of Data

Over 36 MB (citations 2.2MB, complex2ints 198.5KB, complex2matrix 4.8MB, complex2spoke 815.7KB, complex2subunits 812.3 KB, complexes 384.1KB, complexes2refs 37.3KB, ints 8.7MB, labels 7.4MB, nrints 5.5MB, redundant gi 902.6KB, refs 4.6MB, taxon 37.7KB) in the Nov 2004 release.

4.3.5 Update Policy

Frequently updated (almost twice a month) through insertion of new records and editing of previous records. Updates are announced on the news of the web page.

4.3.6 Links to other Sources

- **BRITE:** Biomolecular Relations in Information Transmission and Expression (http://www.genome.jp/brite/)
- **Curagen:** Pathcalling (http://curatools.curagen.com/)
- **DIP:** Database of Interacting Proteins (http://dip.doe-mbi.ucla.edu/)
- **DPInteract:** DNA-protein interactions http://arep.med.harvard.edu/dpinteract/)
- **DRC:** Database of Ribosomal Crosslinks (http://www.mpimg-berlin-dahlem.mpg.de/ ag_ribo/ag_brimacombe/drc/)
- Dynamic Signaling Maps (http://www.hippron.com/hippron/)
- **ICBS:** Inter-Chain Beta-Sheets—A database of protein-protein interactions mediated by interchain beta-sheet formation (http://www.igb.uci.edu/servers/icbs/)
- **JenPep:** Immunology MHC-peptide database (http://www.jenner.ac.uk/JenPep/)
- **Khon:** Kohn Molecular Interaction Maps (http://discover.nci.nih.gov/kohnk/interaction_maps.html)
- **MHCPEP:** A database of MHC binding peptides (http://wehih.wehi.edu.au/mhcpep/)
- **MINT:** A database of Molecular INTeractions (http://160.80.34.4/mint/)
- **PATIKA:** Pathway Analysis Tool for Integration and Knowledge Acquisition (http://www.patika.org/)
- **PIM:** Protein Interaction Map (http://pim.hybrigenics.com/pimriderext/common/)
- **PIMdb:** Drosophila Protein Interaction Map database (http://proteome.wayne.edu/PIMdb.html)
- **Relibase:** A program for searching protein-ligand databases (http://relibase.ebi.ac.uk/reli-cgi/rll?/reli-cgi/general_layout.pl+home)
- **SPIN-PP:** Surface Properties of INterfaces — Protein Protein Interfaces
- **SYFPEITHI:** A Database of MHC Ligands and Peptide Motifs (http://syfpeithi.bmi-heidelberg.com/)
- **TRANSFAC:** The Transcription Factor Database (http://www.gene-regulation.com/)

Genome Databases:

- **COMPEL:** Composite Regulatory Elements (http://compel.bionet.nsc.ru/new/index.html)

- **Ecocyc** (and Metacyc) (http://ecocyc.org/)
- **FlyBase** (http://flybase.bio.indiana.edu/)
- **GeneNet:** Gene networks (http://wwwmgs.bionet.nsc.ru/mgs/gnw/genenet/)
- **GeNet:** Gene Networks Database (http://www.csa.ru/Inst/gorb_dep/inbios/genet/genet.htm/)
- **HOX Pro db:** Homeobox Genes DataBase (http://www.iephb.nw.ru/labs/lab38/spirov/hox_pro/hox-pro00.html)
- **Indigo:** Gene network (http://195.221.65.10:1234/Indigo/)
- **KEGG:** Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/)
- **MIPS:** Yeast Genome Database (http://mips.gsf.de/genre/proj/yeast/index.jsp)
- **IFTI:** Mirage (http://www.ifti.org/)
- Mouse Genome Informatics (http://www.informatics.jax.org/)
- Rat Genome Database (http://rgd.mcw.edu/)
- **RegulonDB:** A DataBase On Transcriptional Regulation in E. Coli (http://www.cifn.unam.mx/Computational_Genomics/regulondb/)
- Saccharomyces Genome Database (http://www.yeastgenome.org/)
- **SELEX DB** (http://wwwmgs.bionet.nsc.ru/mgs/systems/selex/)
- **SoyBase** (http://soybase.ncgr.org/)
- **TRRD** (http://wwwmgs.bionet.nsc.ru/mgs/dbases/trrd4/)
- **WormBase:** Transcription Regulatory Regions Database (http://www.wormbase.org/)

Pathway Databases:

- **BBID:** Biological Biochemical Image Database (http://bbid.grc.nia.nih.gov/)
- **Biocarta** (http://www.biocarta.com/)
- **Biocyc:** Knowledge Library (http://www.biocyc.org/)
- **BioPathways :** Consortium (http://www.biopathways.org/)
- **CSNDB:** Cell Signaling Networks Database
- Dynamic Signaling Maps (http://www.hippron.com/hippron/)
- **NetBiochem:** Welcome Page (http://www.mcphu.edu/netbiochem/NetWelco.htm)
- **PFBP:** Protein Function and Biochemical Pathways (http://www.ebi.ac.uk/research/pfbp/)

- **Project SPAD:** Signaling Pathway Database
- **STKE:** Signal Transduction Knowledge Environment (http://stke.sciencemag.org/)
- **TRANSPATH:** Signal Transduction Browser (http://193.175.244.148/)

Enzyme Databases:

- Biocatalysis/Biodegradation Database (http://umbbd.ahc.umn.edu/)
- **BRENDA**

4.4 Biocarta: Charting Biological Pathways

BioCarta is an interactive web-based resource for exploring biological pathways. These are well integrated with gene-oriented and literature resources. Simultaneously, BioCarta offers an easy and dynamic forum for information exchange and collaboration between researchers, educators and students. BioCarta is an "open source" approach:

- online maps depict molecular relationships from areas of active research;
- constantly integrates emerging proteomic information from the scientific community;
- catalogs and summarizes important resources providing information for over 120,000 genes from multiple species;
- contains both classical pathways as well as current suggestions for new pathways;
- Biocarta is open to submissions from outside researchers.

The following categories of pathways are available: Adhesion, Apoptosis, Cell Activation, Cell Cycle Regulation, Cell Signalling, Cytokines/Chemokines, Developmental Biology Expression, Hematopoeisis, Immunology, Metabolism, Neuroscience.

4.4.1 Accessibility

- **URL:** http://www.biocarta.com/

Access via web interface.

4.4.2 Query Facilities

There are two broad categories of queries: gene-oriented and pathway-oriented queries.

- **Gene oriented queries:** allow the user to search for a gene in a variety of different organisms by gene symbol or gene name. Additional search parameters include the category of the pathway (e.g. adhesion, apoptosis) or the organism. The results are cross-linked with multiple databases for gene specific information (Omim, PubMed, LocusLink, UniGene, Kegg, Entrez/Protein, Swissprot, SNP, HomoloGene, as well as organism-specific databases such as WormBase or FlyBase) (http://www.biocarta.com/search/index.asp)

- **Pathway-oriented queries:** allow the user to search for pathways by title or category. Biocarta pathway records contain a graphical model in which the various molecules are cross-linked to other resources as in the case of gene queries above. The pathway record also contains a short description of the pathway and a machine-readable list of proteins involved in the pathway (mentioning the names of the proteins and the associated LocusID accession number) (http://www.biocarta.com/genes/index.asp).

4.4.3 Data Formats

Currently the content of the Biocarta database is accessible only in HTML format, so a specialized HTML wrapper is necessary in a Semantic Web context.

4.4.4 Size of Data

355 pathways, over 120,000 genes from multiple species.

4.4.5 Update Policy

The site is updated regularly, but updates are not explicitly announced or propagated. However, the size of the pathway database allows its frequent re-downloading.

4.4.6 Links to other Sources

Omim, PubMed, Locus, UniGene, Kegg, EntrezProtein, Swissprot, SNP, HomoloGene, WormBase, FlyBase, etc.

5 Gene Expression Resources

Genes in the DNA are only blueprints for biologically active molecules. Gene expression research deals with the question at which stages of an organism's development and in which of its anatomical parts genes become active. Since experiments to answer this question destroy the investigated animal, research relies on model organisms. Both the Gene Expression Database *GXD* and the Edinburgh Mouse Atlas *EMAP* collect gene expression data for the common mouse, which is the most widely used mammalian model organism. While GXD links its data to mouse tissues by an anatomy ontology alone, EMAP uses a combination of anatomical vocabulary and a 3-dimensional model of the mouse.

5.1 GXD: MGI Gene Expression Database

The GXD (Gene Expression Database) is part of the MGI (Mouse Genome Informatics) system at the Jackson Laboratory, USA. It accumulates, stores and makes accessible gene expression information from the laboratory mouse. This includes information about the expression profiles of transcripts and proteins over a range of mouse stains and mutants. It uses a hierarchically-structured anatomy ontology, similar to the one used by the EMAP project (genex.hgu.mrc.ac.uk).

5.1.1 Accessibility

- **URL:** http://www.informatics.jax.org/mgihome//GXD/aboutGXD.shtml

Primary access to GXD is through the GXD web pages.

5.1.2 Query Facilities

For computational access, the GXD's Sybase database can be accessed through an SQL interface. For specific queries, which need to be agreed on with the GXD group, daily reports in the form of tabulated flat files are generated and can be downloaded (via FTP) from their servers.

5.1.3 Data Formats

GXD uses relational database technology (Sybase) to store all of its data.

5.1.4 Size of Data

No information available.

5.1.5 Update Policy

Daily updates take place, mostly additions, but modifications and deletions are possible. There is currently no update propagation mechanism in place, but a mailing list-based approach is being considered.

5.1.6 Links to other Sources

GXD is closely integrated with other MGI databases, specifically mouse genome sequence data. Jointly with the EMAP and EMAGE databases at the MRC Human Genetics Unit in Edinburgh, UK, it forms the Mouse Gene Expression Information Resource (MGEIR). GXD also links to PubMed and Online Mendelian Inheritance in Man (OMIM).

5.2 EMAP: Edinburgh Mouse Atlas Project

The Edinburgh Mouse Atlas consists of *EMAP*, the actual atlas, and *EMAGE*, the gene expression database. *EMAP* is a digital atlas of mouse development and serves as a framework for spatio-temporal data such as *in-situ* gene expression and cell lineage. It consists of 3D reconstructions (3D grey-level voxel images) of embryos at various stages of development, an anatomy ontology and mappings between the two. *EMAGE* holds *in-situ* gene expression data that has been mapped onto the atlas and complements the data that can be found in GXD (see above). The Mouse Atlas has been developed by the MRC Human Genetics Unit in Edinburgh (in collaboration with Edinburgh University).

5.2.1 Accessibility

- **URL:** http://genex.hgu.mrc.ac.uk

Accessibility: Access is primarily supported by the project web pages and specific 2D and 3D viewer programs. Programmatic access to data is available via a CORBA interface, though plans are now under way to provide more access through web services.

5.2.2 Query Facilities

Queries are primarily supported through the project's web interfaces, using textual as well as image-based query formulation.

5.2.3 Data Formats

EMAP uses the object-oriented database system ObjectStore, but efforts are now underway to move to a relational data model using IBM's DB2. Some of the data is available in XML format.

5.2.4 Size of Data

No information available.

5.2.5 Update Policy

Regular updates take place, primarily additions, but deletes and modifications can also happen. No update notification available at the moment. Email notification is planned for registered users.

5.2.6 Links to other Sources

The Mouse Atlas has close links to a number of other resources, particularly GXD and EnsEMBL. Various experiments are under way to use GRID software to link with other sources. These are currently internal projects and not publicly available.

6 Literature Resources

PubMed is the single most important resource that provides access to research literature in medicine and biology.

6.1 PubMed: Digital Archive of Biomedical and Life Sciences Journal Literature

The National Library of Medicine runs a search and retrieval system called Entrez, which integrates information from the National Center for Biotechnology databases. These databases include nucleotide sequences, protein sequences, macromolecular structures, genomes and MEDLINE articles, which stands for National Library of Medicine's database of indexed journal citations and abstracts from the 1966 Index Medicus forward. All this is accessible for the user through the web-based interface of PubMed. PubMed is a database of bibliographic information drawn primarily from the life sciences literature. The majority of the articles come from MEDLINE. Currently, the user can access more than 14 million citations for biomedical articles back to 1950. All MEDLINE citations have assigned MeSH terms and publication types from a controlled vocabulary. MeSH is the National Library of Medicine's controlled vocabulary used for indexing articles. The MeSH terminology provides a consistent way to retrieve information from sources that may use different terminology for the same concepts. The MeSH vocabulary is used for indexing journal articles from Index Medicus and MEDLINE and also for cataloging

books and audiovisuals. PubMed contains links to full-text articles at participating publishers' web sites as well as links to other third party sites. It also provides access and links to the integrated molecular biology databases maintained by the National Center for Biotechnology Information.

6.1.1 Accessibility

- **URL:** http:// www.pubmed.org/

Accessible via the Web, as a Web Service, via a local SQL database.

6.1.2 Query Facilities

Querying via the website, but also SOAP for web services can be used.

6.1.3 Data Formats

Text Files (either as own flat file format, as XML or as database dump).

6.1.4 Size of Data

About 40GB, which include 15.000.000 articles (only the abstract, possibly the link to the full-text).

6.1.5 Update Policy

Updated frequently with insertion of new articles and replacement of old or changed articles by a new ones). Updates are announced via email, subscribers can download changes via FTP (the IP address has to be registered for this).

6.1.6 Links to other Sources

PubMed has links to very many other databases such as Nucleotide DBs, Protein DBs, Structure DBs, Taxonomy DBs, Genome DBs, Expression DBs, Chemical DBs. Check in

- **URL:** http://www.ncbi.nlm.nih.gov/Database/index.html/

7 Ontologies

The Gene Ontology *GO* is the most widely used controlled vocabulary in bioinformatics resources.

7.1 GO: Gene Ontology

The use of ontologies in bioinformatics has grown drastically since database builders concerned with developing systems for different (model) organisms joined to create the Gene Ontology Consortium in 1998. Since then several other organizations have joined the consortium. (For a complete list of member organizations we refer to the Gene Ontology Consortium home page.) The goal of the Gene Ontology project is to produce a controlled vocabulary that can be

applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing. GO provides three structured networks (directed acyclic graphs) of defined terms to describe gene product attributes: biological process, molecular function and cellular component.

The biological process ontology deals with biological objectives to which the gene or gene product contribute. A process is accomplished via one or more ordered assemblies of molecular functions. The molecular function ontology deals with the biochemical activities of a gene product. It describes what is done without specifying where or when the event takes place. The cellular component ontology describes the places where a gene product can be active.

The GO ontologies have nowadays become a de facto standard and are used by many databases containing information about genes and proteins for annotation. GO is also one of the controlled vocabularies of the Open Biological Ontologies (OBO http://obo.sourceforge.net/).

7.1.1 Accessibility

- **URL:** http://www.geneontology.org/

The ontologies can be downloaded or accessed via the GO Consortium's home page.

7.1.2 Query Facilities

There are a number of tools that can be used to browse and query the GO ontologies. AmiGO and DAG-Edit are tools developed by the consortium. AmiGO is a tool that allows search and browsing of the GO Database. One can query based on GO terms or on gene names. The search can be restricted to predefined fields. DAG-Edit is a Java application that provides an interface to browse, query and edit GO ontologies or any other vocabulary that has a directed acyclic graph data structure In addition, several other tools for querying and browsing are developed by other groups and many of these are available via the GO home page. Several tools are available for searching and browsing the GO ontologies, for annotation of gene products using GO terms, and for using GO ontologies in gene expression and microarray analysis. Regarding the annotation of genes and gene products the GO Consortium aims to standardize information about the quality of the annotation using evidence codes. An evidence code indicates how annotation to a particular term is supported. Examples of evidence codes are IC (inferred by curator), IEA (inferred from electronic annotation), IEP (inferred from expression pattern) and TAS (traceable author statement).

7.1.3 Data Formats

The GO ontologies are available in the GO flat file format, the OBO flat file format, XML and the GO Database/MySQL format. The GO Database is built from the data publicly available as flat files from the main GO web site. The database is not used for data management, but only for querying, either with AmiGO, the go-db-perl modules or with MySQL

7.1.4 Size of Data

On December 13, 2004 the ontologies comprised 18,151 terms concerning molecular functions (7452 terms) biological processes (9195 terms) and cellular components (1504 terms).

7.1.5 Update Policy

The XML files and GO Database releases are built on a monthly basis. The publicly available FTP archive is updated every 30 minutes. A read-only publicly accessible CVS repository for the GO project is updated daily. Updates are made by curators. New terms can be added. Terms can be changed, marked obsolete, split and merged with other terms.

7.1.6 Links to other Sources

There exist mappings between GO terms and terms in other ontologies and databases (SWISS-PROT, EC numbers, TIGR, GenProtEC, InterPro, MIPS, MetaCyc, MultiFun, Pfam, ProDom, PRINTS, Prosite, SMART, Reactome, COG). Most of these are updated monthly, although some are not updated at all.