universität
wien

Magisterarbeit

Titel der Magisterarbeit

Computer-aided research in natural sciences

Verfasserin

Bakk.rer.soc.oec. Bakk.techn. Iris Studeny

angestrebter akademischer Grad

Magistra der Sozial- und
Wirtschaftswissenschaften (Mag.rer.soc.oec.)

Wien, 2010

Studienkennzahl lt. Studienblatt: A 066 922
Studienrichtung lt. Studienblatt: Magisterstudium Informatikmanagement
Betreuer: Ao.Univ.Prof. Dipl.-Ing. Dr.sc.med. Dr.techn. Dr.rer.nat. Frank Rattay

Thank to Prof. Rattay for advising my diploma thesis!
Thank you, my parents for everything!
Thank to Claus D. Volko, BSc for reviewing.

Contents

Part I

The Internet

Chapter 1

Early Beginnings of the Internet

1.1 Memex

The primary aim of the web was to be a platform for science. Ted Nelson contrived Xanadu, a universal hypermedia system which he described in his book "Literary machines". He described hypertext as non-sequential writing and as a literary medium. Vannevar Bush suggested a theoretical system adequate to a hypertext system of nowadays. Bush also saw the necessity of a machine supporting scientists to stay informed in their own disciplines and called this system Memex. He depicted Memex as

> *"a sort of mechanized private file and library".* [Hall et al., 2008]

The main idea was to reproduce the trails through which researchers found their results.

A venture of Douglas Engelbart was the "Augment project" in 1963 at Stanford University. Engelbart was interested in "augmenting the human intellect". This means people should be enabled to solve complex problems in a fast way. Practical inventions based on Engelbart's work are word processing, screen windows and the mouse etc.

(cf. [Hall et al., 2008], [Engelbart, 1962])

1.2 CERN

Every Apple Macintosh computer in the late eighties had a release of Hypercard familiarizing the concepts of hypertext and in this way it became a widespread technique. The fields of application of hypertext systems in distributed and large scale environments were analyzed in the nineties, e.g. Tim Berners-Lee started contriving a distributed hypertext environment for physicists at European Council for Nuclear Research (CERN). The open protocols hypertext transfer protocol (HTTP) and hypertext mark-up language (HTML) were developed. The first web-server of CERN could be reached using telnet from every

computer in the world. Users could generate HTML-documents by editors. The first graphical browser was Mosaic from National Center for Supercomputing Applications at the University of Illinois at Urbana-Champagne (NCSA). It was this browser that brought the quantum jump for the web.

(cf. [Hall et al., 2008])

1.3 Summary - Beginnings of the Internet

The Internet is an important part of all fields of life. It is an overwhelming means for spreading knowledge. There is nearly no constraint in acquiring knowledge caused by restricted access to it, like many centuries ago. There is another problem - most of the content in the Internet is not usable for research. It is not as scientific and not as academic as it was planned to be in the early beginnings of the Internet. The staggering amount of websites, data and information call for a tool to filter the content and to receive only a selection of relevant sites. These tools are called search engines. Some of them are specialized in scientific information and are created for an academic target group.

Chapter 2

Search Engines

The web expanded and so it became more and more complicated to find information by simply clicking through hypertext links. So the era of search engines started.

2.1 Google

Larry Page and Sergey Brin created the search engine BackRub for Stanford University in 1996. BackRub had been running on Stanford University's servers for more than one year. Finally BackRub needed too much bandwidth. Google got its name in 1997 when Larry Page and Sergey Brin decided to rename it from BackRub to Google. The name "Google" is a reference to "googol" that means 10^{100} to show the huge amount of information provided by Google. Google's famous success story started in 1998 beginning with a check for $ 100,000 and a report about Google in "PC Magazine". Google got its first "Webby Award" in 2000, released its first versions in other languages and started its famous growth. Google Adword and Google Toolbar enriched the product family. Google entered into a partnership with Yahoo, became its default search provider and declared to have an index of about one billion Uniform Resource Locators (URLs). This means, Google was then the largest search engine in the world. The Image search came up in 2001. Google Print, a forerunner of Google Book Search, was founded in 2003. Google entered social networking, offered Google Desktop Search and Google Scholar and Google Earth in 2004. Google Maps, Google Mail (Gmail) and code.google.com were launched in 2005. Google Analytics was created. A takeover of YouTube took place. The service "Patent Search" for the U.S. went online. Google Health was launched in 2008. Google Squared, Google Translator Kit and the "Moon" in GoogleEarth were new projects in 2009.

(cf. [Hall et al., 2008], [Google, 2010])

2.2 Search Engines for Academic Matter

Search engines for academic matter are:

- Bielefeld Academic Search Engine (BASE)
- CiteSeerX
- Eccellio Science
- iLib2
- Scirus
- ScienceSequere
- SweetSearch

These search engines provide scientific results like publications.

(cf. [BASE, 2010a], [of Information Sciences and Technology, 2010], [eccellio, 2010], [scirus, 2010], [sequere, 2010a], [Dulcinea Media, 2010b])

2.2.1 BASE

Bielefeld Academic Search Engine (BASE) is an academic search engine with a German interface. The target of BASE is to enable the user to search the Internet for open access scientific documents. It allows to search for different grammar cases which is important in German. Another feature is the possibility to search for plural forms or in 21 European languages.

(cf. [BASE, 2010a], [BASE, 2010b])

2.2.2 CiteSeerX

CiteSeerX is a search engine. It is also a scientific literature digital library. CiteSeerX was mainly created for the area of computer and information science.

(cf. [of Information Sciences and Technology, 2010])

2.2.3 Eccellio Science

Eccellio Science is one part of Eccellio. The parts of Eccellio are:

- Eccellio Arts
- Eccellio Movies
- Eccellio Science
- Eccellio Sports
- Eccellio Web

Eccellio also offers a browser search plugin for Firefox. The target group of Eccellio Science are students and researchers. The results contain academic matter as well as Wikipedia entries but it mainly focuses on scientific content.

(cf. [eccellio, 2010])

2.2.4 iLib2

ILib2 is a Chinese search engine that redelivers scientific abstracts and the journal where the corresponding publication can be found. The links are `http://scholar.ilib.cn/` and `http://www.ilib.cn/`.

2.2.5 Scirus

Scirus is a search engine for scientific information. Therefore it searches 410 million web pages. Scirus searches sites with medical, scientific content, articles, journals, patents and reports.

(cf. [scirus, 2010])

2.2.6 ScienceSequere

ScienceSequere is a search engine for scientific content. It finds articles, abstracts, patents and publications. It allows to search in countries, regions and subjects.

(cf. [sequere, 2010a], [sequere, 2010b])

2.2.7 SweetSearch

SweetSearch is a search engine. Its target group are students. It belongs to Dulcinea Media. Dulcinea Media offers the following services:

- encontrandoDulcinea
- findingDulcinea
- findingEducation
- SweetSearch
- SweetSearch Biographies
- SweetSearch SweetSites
- SweetSearch2Day
- SweetSearch4Me

SweetSearch searches 35,000 evaluated websites.

(cf. [Dulcinea Media, 2010b], [Dulcinea Media, 2009], [Dulcinea Media, 2010a])

2.3 Summary - Search Engines

Search engines have become necessary because the Internet has transcended every frontier of clarity. They help to pick special kinds of information out of the Internet. Meanwhile, several types of search engines serving several types of target groups have developed. The development ranges from specialized byproducts of large search engines like `books.google.com` or `scholar.google.com` to single scientific search engines.

2.4 Points of Criticism - Search Engines

One of the main problems is still the availability of information. Search engines enable the user to get a more or less overwhelming amount of websites - but most of them are either American or European sites - this means that the results are English, few of them French or German. Very few Japanese sites are represented. It is quite difficult to get worldwide results - even when searching for them. Some search engines do not offer an advanced search option. It can be difficult to search for results from special countries if search engines do not offer a special option for that. Specialized scientific search engines cover very small market niches competing with byproducts of their large competitors, encyclopedias or simple governmental services like special browser plug-ins. Most of them offer additionally semi-scientific results for pupils and students, like Wikipedia. These results are not qualitatively high enough for being cited in academic articles but they can serve as a good starting point if they contain links to high quality resources. Some search engines are enhanced by Google.

Google is a very special case because of its market power. Google had a share of the US market in advertising and search area of approximately 65.4 %. Microsoft Bing had a share of approximately 9.9 %, its partner Yahoo! had approximately 18 % and Microsoft planned to include results of WolframAlpha. Ask.com had a share of approximately 3.9 %, AOL.com had approximately 2.9 %. The sum of these declarations is 100.1 %. So all the other search engines have to share the rest of the market - the rounding difference of the market shares of the large search engines.

(cf. [APF, 2009])

Part II

Politics

Chapter 3

International Politics

3.1 CODATA

Committee on Data for Science and Technology (CODATA) is an interdisciplinary Committee that was founded in 1966 by the International Council for Science (ICSU). It is sponsored by the Canada Institute for Scientific and Technical Information (CISTI). The goal of CODATA is to enhance accessibility to international scientific data and to raise quality of research data in this way. This data can be calculations, experiments or observations.

(cf. [CODATA, 2010])

3.2 DataCite

DataCite is a consortium. It acts internationally. The goal of DataCite is to raise acceptance of data published in the Internet. Thererfore, the DataCite consortium promotes data archiving, easier global access to research data on the Internet and to raise acceptance of this data as being legitimate. DataCite also supports improved protection of research invest.

Members and associated members of the DataCite are:

- Australian National Data Service
- California Digital Library
- Canada Institute for Scientific and Technical Information
- Digital Curation Centre
- ETH Zurich
- German National Library of Medicine
- German National Library of Science and Technology
- GESIS - Leibniz Institute for Social Science
- Institute for Scientific and Technical Information

- Microsoft Research
- Purdue University Libraries
- Technical Information Center of Denmark
- The British Library
- TU Delft Library

(cf. [DataCite, 2010b], [DataCite, 2010a])

3.3 Research Data Strategy Working Group

The Research Data Strategy Working Group is a Canadian project. The main goal of the Research Data Strategy Working Group is to manage access to Canadian research data and to maintain it. Members are granting agencies, institutes, libraries, universities and researchers. Therefore a system is necessary. This system should also implement access control as Canadian research data should be protected.

(cf. [data donnees.gc.ca, 2008])

3.4 Summary International Politics

Politics describes consortia and committees. They are installed by different countries to promote science and scientific progress. These consortia and committees are responsible for funding, platforms, programs and the acquiring of data and state. This serves as base for the following steps. Cooperations and collaborations concern the EU, some countries within the EU or they are international. As there is a close connection between politics and its output, many topics concerning it are described in the chapter “Portals”. “International Politics” concentrates on programs outside the EU.

3.5 Points of Criticism - International Politics

The concentration of all sources to an interlinked group of scientific organizations is advantageous to all participants. A focusing of academic sources is good for simplification and standardization. An interesting point are the different laws in the participating countries.

Chapter 4

European Politics

4.1 E-MeP

European Membrane Protein Consortium (E-MeP) is a consortium funded by the Sixth Framework Programme. The budget was at about 10,350,000 €. E-MeP started in 2004. The target of the E-MeP is to promote the exploration of the structures of membrane proteins including proteomes and constituting pharmacological targets. Therefore the E-MeP research platform has been installed. The partners are:

- Aston University (United Kingdom)
- Bio-Xtal (United Kingdom)
- Bristol University (United Kingdom)
- Centre national de la recherche scientifique (France)
- École Polytechnique Fédérale de Lausanne (Switzerland)
- European Molecular Biology Laboratory - European Bioinformatics Institute (EBML-EBI) (United Kingdom, Germany)
- Imperial College (United Kingdom)
- Leeds University (United Kingdom)
- Martin-Luther-Universität Halle-Wittenberg (Germany)
- Max-Planck-Institut für Biophysik (Germany)
- Medical Research Council (United Kingdom)
- Philipps-Universität Marburg (Germany)
- Stockholms Universitet (Sweden)
- Université de la Méditerranée (France)
- University Medical Centre Nijmegen (The Netherlands)

- University of Glasgow (United Kingdom)
- University of Gothenburg (Sweden)
- University of Groningen (The Netherlands)

The main research interests of the E-MeP Consortium are:

- Membrane proteins of prokaryotes and especially eukaryotes
- Analyzation of structures
- Technological methods for high-throughput

(cf. [EMBL-EBI, 2010b], [EMeP, 2009], [E-Mep, 2009a], [E-Mep, 2009b])

4.2 ELIXIR

The aim of European Life Sciences Infra Structure for Biological Information (ELIXIR) is to arrange a European life science research infrastructure not only to support biological research but also to implement the results in related fields like environmental matter, food, plants, industry, medicine, pharma business and society. ELIXIR is also a project for establishing pan-European international capability in:

- Bio-compute centers
- Data resources
- Infrastructure for:
  - Integration of biological data
  - Services
  - Software
  - Tools
- Services concerning:
  - Standards
  - Training

The ELIXIR Steering Committee is responsible for the creation and definition of the infrastructure as a “Hub and Nodes concept” and especially the interoperability of services, to find duplications and to detect deficiencies.

(cf. [EMBL-EBI, 2010b], [ELIXIR, 2009a], [ELIXIR, 2009b])

4.3 EMBRACE

The European Model for Bioinformatics Research and Community Education (EMBRACE) Network of Excellence is a project having the target to assemble a group of European experts in biomolecular sciences. The target of this Network is to improve informatics and information gathering out of data and actual data access by biological scientific community as many national data warehouses were out-dated. EMBRACE service interfaces shall be used for proprietary data and tools. This should be possible for both academic and and commercial scientists. This aim shall be reached by implementing several tools and resources. These tools resources are:

- Databases
- Tools for bioinformatics
- webservices
- Grid services

(cf. [EMBL-EBI, 2010b], [EMBRACE, 2010], [EMBRACE, 2009])

4.4 EMERALD

Empowering the Microarray-Based European Research Area to Take a Lead in Development and Exploitation (EMERALD) is an European Bioinformatics Institute (EBI) hosted project. The aim of EMERALD is to constitute and to propagate metrics, quality standards and best laboratory practices for microarray research in the European scientific community. The improvement of the quality of microarray data should affect further array-based technologies. The quality of data and meta-data is essential for microarray data analysis and generation basing on this data. The quality of data is important for the amount of information that can thereof be extracted. Hence the Functional Genomics Data (FGED) Society (former old name of FGED (MGED) Society) introduced Minimum Information About a Microarray Experiment (MIAME) - guidelines for the description of experiments. FGED initiatives primary concentrate on data context and included data context. Further EMERALD partners are:

- Biological Research Center (BRC) (Hungary)
- Centro de Investigación Príncipe Felipe (CIPF) (Spain)
- EBI (United Kingdom)
- Institute for Reference Materials and Measurements (IRMM) (Belgium)
- Laboratory of the Government Chemist (LGC) (United Kingdom)
- Norges Teknisk-Naturvitenskapelige Universitet (NTNU)/Norwegian Microarray Consortium (NMC) (Norway)
- Uppsala University (UU) (Sweden)
- VIB (Belgium)

(cf. [EMBL-EBI, 2010b], [QUALITY, 2009a], [QUALITY, 2009b])

4.5 Summary European Politics

European and extra-European research politics concentrate on similar fields. Political initiatives are closely interlinked with their results, this means "Portals". America is one of the leaders in offering freely available, peer-reviewed publishings because of PubMed and related services. European services do not offer the same quality of assortment even though they are related to PubMed. There are large overlappings. On the other hand international services use European services.

4.6 Points of Criticism - European Politics

European services are related to extra-Europeans and they have overlappings in their services. One problem for future could be juridical questions and resulting career questions of researchers - not in respect of copyrights etc. - but in view of different laws in participating countries restricting searchers in their fields of research or not. The question is how far they may go in their research, briefly, which results can researchers achieve within their legal framework and which results can other researchers achieve using their results in other legal frameworks.

Part III

The Internet as Source of Knowledge

Chapter 5

International Portals

5.1 The NCBI

5.1.1 NCBI General

The National Center for Biotechnology Information (NCBI) was established in 1988 as a division of the National Library of Medicine (NLM). The NLM had great know-how in creating and managing biomedical databases and as it belonged to the National Institutes of Health (NIH) it was able to build up an internal research program in computational molecular biology. In this way the world's largest biomedical research facility was formed. The NCBI is organized in three branches - the Computational Biology Branch, the Information Engineering Branch and the Information Resources Branch. The NCBI is a national resource for molecular biology information. The focus lies in molecular biology, genetics and biochemistry. The purpose of the NCBI is knowledge processing. This means it is responsible for storing and collecting national as well as international information about these topics. Another target is to offer the possibility to analyze existing information for the research community. This means the NCBI has to conceive automated systems therefor. For that reason the NCBI offers among other things guidance of research, fosters collaboration with governmental agencies and institutes, is engaged in gaining researchers. The NCBI focuses on software, tools and databases which it develops, distributes and supports. The NCBI also develops and supports standards for databases, storage and communication of data. The gene-centered resources of the NCBI are Gene, Gene Expression Nervous System Atlas (GENSAT), Gene Expression Omnibus (GEO), HomoloGene and UniGene.

The NCBI offers training and tutorials at `http://www.ncbi.nlm.nih.gov/guide/training-tutorials/`.

(cf. [NCBI, 2004], [NCBI, 2010p], [Entrez, 2005a])

5.1.2 NCBI Databases

There are databases, tools, maps, collaborative cancer research, FTP sites and resource statistics. The main databases are:

- Literature Databases
- Molecular Databases
- Genomes

(cf. [NCBI, 2007a], [NCBI, 2010q])

NCBI Literature Databases

Literature Databases consist of

- Books
- Journals
- Medical Subject Headings (MeSH)
- NLM Catalog
- Online Mendelian Inheritance in Animals (OMIA)
- Online Mendelian Inheritance in Man (OMIM)
- PubMed
- PubMed Central (PMC)

(cf. [NCBI, 2007a], [NCBI, 2010q], [NCBI, 2010c])

Books, Bookshelf, Journals and the NLM Catalog In cooperation with publishers as well as authors, the Bookshelf contains a list of biomedical books that can be fully searched, but not all books can be browsed. The books available are digitalized versions of textbooks, monographs and text-based databases. The books of the NLM and the NCBI are browsable. Some books can be downloaded in PDF-format. Links to related subjects and animated tutorials exist in some cases and are auto-generated. These links can refer to other Entrez databases like Gene, Online Mendelian Inheritance in Man (OMIM), PubChem, PubMed Central (PMC) and PubMed too. Bookshelf is a part of Entrez.

Journal articles are not included in full-text in PubMed. Many of them can be obtained freely. A "Full free text filter" helps to find only results of free available articles.

The NLM has collected bibliographic data of print media, audio-visual media, software, electronic resources etc. The NLM Catalog provides access to this data via Entrez.

A list of Bookshelf books is available at `http://www.ncbi.nlm.nih.gov/bookshelf/`.

(cf. [NCBI, 2010c], [Bookshelf, 2010], [PubMed,], [NCBI, 2006b])

MeSH is the checked vocabulary database of the NLM. The purpose of Medical Subject Headings (MeSH) is the indexing of the articles of Medical Literature Analysis and Retrieval System Online (MEDLINE)/PubMed. The search terms of MeSH can be imported in the PubMed search. To do so, the concept is entered into the MeSH search. MeSH returns the medical subject heading that is used for the citations' indexing and a short definition of the term. It is possible to declare a number of concepts and restrictions and to give details of them. In this manner searchers specify their search and reduce their results. There are also some short videos available that show how to use MeSH. Concepts are updated every week.

MeSH also contains a browser, the MeSH browser. The search terms or fragments are typed into the MeSH browser that delivers the terms that are used for indexing the articles. The MeSH browser is used to get qualifiers, descriptors and additional concepts.

MeSH vocabulary can be downloaded in XML, ASCII and USMARC authority format.

(cf. [MeSH, 2010], [MeSH, 2009b], [MeSH, 2009a])

OMIM and OMIA - Catalogs of Genes Online Mendelian Inheritance in Animals (OMIA) is a gene database also containing information about inherited disorders, animal traits, textual information, references and links to related services of PubMed. Here, animals are meant as all animals except humans and mice.

OMIM was founded in 1985 as an online version of the still existing Mendelian Inheritance in Man (MIM), a book edition started in 1966, is a comprisal of human genes and genetic phenotypes and focuses on the connection between genotype and phenotype. It is a database of genes including information about all known genetic disorders and over 12,000 genes that is updated every day and offers links to related topics. The target group of OMIM are professionals concerned with genetic disorders.

(cf. [Entrez, 2010o])

PubMed The database PubMed is a free service of the NCBI at the NLM that belongs to the NIH. PubMed contains several millions citations, links and abstracts of biomedical literature of MEDLINE, that is part of PubMed, of journals, books, full-text articles and websites and other related resources. PubMed is an index of abstracts. It aims to be the database of choice for scientists and clinicians to look for relevant information. MEDLINE is a web-based database that is searchable freely. It keeps references to more than 17 million published biomedical articles of over 5,000 journals.

The Entrez service LinkOut allows publishers to present their citations and to make their full-text articles available. MeSH helps to specify the searching concepts.

(cf. [NCBI, 2010c], [PubMed,], [MEDLINE, 2008], [MeSH, 2010], [PMC, 2008], [NIH, 2010a], [NLM, 2010])

PubMed Central which is an index of full-text papers belongs to the NIH, more exactly to the NCBI in the NLM. The function of PMC is to be the leading digital library. The NLM also comprises a huge amount of printed journals. The PMC is the digital counterpart thereof. Some of the printed journals have been digitalized and inserted into PMC. The access to the materials is free as the NLM is confident that this strategy guarantees durability of the service. Another aim of PMC is not only to store data but to store it in a common format and in a central repository. This means data can be found, processed and combined much more easily.

Publishers are free to take part in this project although they have to serve special editorial standards. It would be desirable if publishers publicized their whole content and not only a few samples of their articles. A publishing on PMC does not change any copyright of the publisher and does not lead to any cost.

Many researchers who get research funding from the NIH publish in journals that do not take part in PMC. This means that these articles are not accessible to the public. In this case authors should deliver their final peer-reviewed and ready for printing manuscript to PMC during 12 month after publication. PMC has file submission specifications that ensure the quality of the article and its pictures.

PubMed Central Canada (PMC Canada) exists since 2009. PMC Canada describes itself as

> *free digital archive of full-text, peer-reviewed health and life sciences literature* [Canada, 2010a].

PMC Canada demands CHIR grant recipients to make their peer-reviewed publications freely available within six months after publication.

(cf. [PMC, 2008], [PMC, 2005], [NIH, 2010a], [NIH, 2010b], [PMC, 2010], [PMC, 2009], [Canada, 2010a])

Molecular Databases

are

- Genes
- Gene Expression
- Nucleotide Sequences
- Protein Sequences
- Structures
- Taxonomy

(cf. [NCBI, 2007a], [NCBI, 2010q])

Genes consists of:

- Consensus CoDing Sequence, Consensus CDS (CCDS)
- (Entrez) Gene
- HomoloGene
- UniGene

UniGene, Gene, HomoloGene, Expressed Sequence Tags (dbEST) and TraceArchive are related databases.

(cf. [UniGene, 2010d], [NCBI, 2010q])

CCDS : The ambition of Consensus CoDing Sequence, Consensus CDS (CCDS) is to identify protein coding regions of humans and mice. These sequences are stable enough to find identical gene arrangements. The data is published and supported by the genome browsers of the collaborating members. The participants in the CCDS are the EBI, the NCBI, the University of California, Santa Cruz (UCSC) and the Wellcome Trust Cancer Institute (WTSI).

(cf. [NCBI, 2010a])

Entrez Gene is the successor of LocusLink. It is an Entrez database and as one of the gene-centered resources it contains data records of annoted genes. Gene was created to serve genomic sequencing projects to identify and characterize genes by acting as connection between expression, function, genomic map, protein structure, sequence and homology data. Gene contains two types of genes - known and predicted. These genes can be defined by their map position or their nucleotide sequence. The complexity of the gene pool covers NCBI's Reference Sequences (RefSeq) and NIH's Mammalian Gene Collection. Gene offers a unique GeneID to find information about genes and loci.

(cf. [Entrez, 2010b], [Entrez, 2010c], [Entrez, 2005a])

HomoloGene serves to calculate genealogical trees. The input consists of proteins of the concerning organisms. HomoloGene automatically finds homogenous regions in completely sequenced eucaryotic genomes. At first, the closer related organisms are assembled. Then the other organisms are added to the tree. HomoloGene gets information from Clusters of Orthologous Groups (COG), FlyBase, Mouse Genome Informatics (MGI), OMIM, Saccharomyces Genome Database (SGD) and Zebrafish Information Network (ZFIN).

(cf. [HomoloGene, 2010a], [HomoloGene, 2010b])

UniGene comprises nucleotide sequences. It utilizes peptide databases for its search that do not include mitochondrial proteins. These peptide databases contain information about:

- Arabidopsis thaliana
- Caenorhabditis elegans
- Drosophila melanogaster
- Escherichia coli
- Homo sapiens
- Mus musculus
- Rattus norvegicus
- Saccharomyces cerevisiae

Analogies in sequence are discovered by Blastx. Blastx checks both strands of the result of a query of a six-frame conceptual translation product against the protein sequence database. An UniGene entry contains of a set of transcript sequences including additional information. These transcript sequences are assumed to originate from the same transcription locus. UniGene also stores Expressed Sequence Tag (EST) sequences. UniGene offers two types of build:

- Genome based build procedure:
  Several procedures calculate sets of transcript sequences that equate to genes or transcript loci.
- Transcriptome based build procedure:
  The subset sequences are clustered to find the ones that belong together. The similarity score is calculated by Boolean links. If the score passes a certain level the sequences are expected to belong together.

UniGene offers its data at `ftp://ftp.ncbi.nih.gov/repository/UniGene/`.

(cf. [UniGene, 2010d], [UniGene, 2010c], [UniGene, 2010a], [UniGene, 2010b])

Gene Expression is built of:

- Entrez GEO DataSets
- Entrez GEO Profiles
- GENSAT
- GEO

(cf. [NCBI, 2010q])

Entrez GEO DataSets is a database that saves gene expression DataSets, original series and platform records in the GEO storage. It queries complete experiments. Entrez GEO DataSets can be used to delve into series and DataSets.

(cf. [Entrez, 2010i], [Entrez, 2010j])

Entrez GEO Profiles is a database that saves gene expression profiles in the GEO storage and delivers individual gene expression profiles. The GEO staff rearranges the submitters' records to comparable DataSets. These DataSets are then loaded in Entrez GEO Profiles to ensure this function.

(cf. [Entrez, 2010k], [Entrez, 2010j])

GENSAT is a database that is funded by the National Institute of Neurological Disorders and Stroke (NINDS). It contains in-situ hybridization data and Bacterial Artificial Chromosomes (BAC) transgenic data. The in-situ hybridization data comes from St. Jude Children's Research Hospital. The BAC transgenic data originates from The Rockefeller University. GENSAT holds a gene expression atlas of the central nervous system of the mouse. It shows the expression of genes in its brain during the stages of development. For this reason GENSAT keeps histological data from transgenic mouse lines at embryonic day 15.5, postnatal day 7 and adult to trace the gene expression in the brain of a mouse. Information about GENSAT at The Rockefeller University is available at `http://www.gensat.org/index.html`. The modified mouse lines data can be searched at the Mutant Mouse Regional Resource Center (MMRRC), `http://www.mmrrc.org/`.

(cf. [NCBI, 2010m], [NCBI, 2010j])

GEO is a public repository that gets its data from the scientific community. GEO contains microarrays, next-generation sequencing and other high-throughput functional genomic data. It offers web-based interfaces and applications for querying and downloading experiments and gene expression patterns. MIAME delivers the minimum guidelines to describe a microarray experiment at GEO and many journals require data fitting MIAME guidelines.

GEO offers two types of navigation - browsing and querying. A query consists of:

- DataSets
- GEO accession display tool
- GEO BLAST
  is a tool that queries Entrez GEO profiles
- GEO Profiles

The design of Browse is:

- DataSet Browser (`http://www.ncbi.nlm.nih.gov/sites/GDSbrowser/`)
- GEO accession
  (`http://www.ncbi.nlm.nih.gov/geo/query/browse.cgi`); allows to browse:
  - Platform
  - Samples
  - Series

DataSets can be clustered and visualized by GEO DataSet Cluster Analysis. Columns correspond to Samples and rows to genes. The results of clustering are shown as "heat maps" with a two-color spectrum. GEO DataSet Cluster Analysis allows selection, enlargement, download, charting as line-plot and a view in Entrez GEO Profiles of interesting cluster data portions.

The clustering processes are:

- Distance metrics
- Hierarchical clustering
  - Average linkage/UPGMA
  - Complete linkage
  - Single linkage
- K-means/K-median clustering

(cf. [GEO, 2010a], [GEO, 2010d], [NCBI, 2010n], [GEO, 2010b], [Entrez, 2010k], [GEO, 2010c])

Nucleotide Sequences consists of:

- dbEST
- Genome Survey Sequence (dbGSS)
- MHC database (dbMHC)
- Single Nucleotide Polymorphisms (dbSNP)
- Sequence Tagged Sites (dbSTS)
- GenBank
- Nucleotide
- PopSet
- Probe
- RefSeq
- Trace Archive
- unified, non-redundant view of sequence tagged sites (UniSTS)
- UniVec
- Whole Genome Shutgun Sequences (WGS)

(cf. [NCBI, 2010q])

dbEST is a unit of GenBank. It keeps sequence data, information on "single path" cDNA sequences and EST. The dbEST release 052110 contains 65,587,756 EST entries of organisms.

(cf. [NCBI, 2010i], [NCBI, 2010d])

dbGSS sequences are embedded into the Genome Survey Sequence (GSS) database that belongs to GenBank. The majority of its sequences are genomic in origin. Genome Survey Sequence (dbGSS) contains more explicit information about data. It keeps over 28,000,000 public entries.

(cf. [NCBI, 2003], [NCBI, 2010f])

dbMHC is a database that serves as a platform for DNA and Major Histocompatibility Complex (MHC). It is a cooperation project with the Medical University Graz, Austria (`http://www.meduni-graz.at/`).

Parallel Resources are:

- dbLRC
  is a database offering a public platform for DNA and Leukocyte Receptor Complex (LRC). It is a cooperation project with the Medical University Graz, Austria, too.

- dbRCB
  is a database that is operating as a publicly accessible platform for human Red Blood Cells (RBC). It is another cooperation project with the Medical University Graz, Austria.

(cf. [NCBI, 2010g])

dbSNP is a database containing single nucleotide substitutions and short deletion and insertion polymorphisms (SNP)s of a species, any part of a genome. The SNPs in the database are linked to ESTs. (cf. [NCBI, 2006c])

dbSTS is a part of GenBank. It keeps sequence data for short genomic landmark sequences or Sequence Tagged Sites. Sequence Tagged Sites (STS) data can be searched by Entrez or Basic Local Alignment Search Tool (BLAST).

(cf. [NCBI, 2010h])

GenBank The GenBank project originally belonged to the Los Alamos National Laboratory (LANL). GenBank is a genetic sequence database that offers DNA sequences as its name implies from over 100,000 organisms and the associated protein translations. It doubles its size every ten months. GenBank together with the DNA Data Bank of Japan (DDBJ), the EBI and the European Molecular Biology Laboratory (EMBL) built the International Nucleotide Sequence Database Collaboration in the mid-1990s. Nucleotide data can be submitted to GenBank by scientists. GenBank's submission tools are BankIt, used for simple submission data, and Sequin for more complex entries. The NCBI does not constrict the use or distribution of the GenBank data but it does not guarantee that submitters of the data do not have any patent, copyright etc.

(cf. [NCBI, 2010k], [Ilene Mizrachi, 2007], [Entrez, 2010n])

Nucleotide is an Entrez database for nucleotides and proteins. It holds STS, Whole Genome Shutgun Sequences (WGS), nucleotide sequences and RefSeqs.

(cf. [Monica Romiti, 2010])

PopSet contains a set of DNA sequences, nucleotide and protein sequence data. It serves to reconstruct the evolutionary relationship of a population. The sequence set can stem from studies concerning populations, phylogeny or mutations.

(cf. [Entrez, 2010f], [Monica Romiti, 2010])

Probe is a database for biomedical research. Probe contains nucleic acid reagents and associated information. It keeps applications for:

- Gene Expression
- Gene Silencing
- Genome Mapping
- Genotyping
- SNP Discovery
- Variation Analysis

(cf. [Probe, 2010b], [Probe, 2010a])

RefSeq is intended to deliver high-quality data of sequences, including genomic DNA, proteins and transcripts for diversity, functional and medical studies. It offers RefSeqs of of taxonomically diverse organisms.

(cf. [NCBI, 2010s])

Trace Archive contains sequence data sets of WGS sequencing. Trace Archive offers these sets for download but the number of records delivered per request is restricted to 40,000.

(cf. [NCBI, 2010u])

UniSTS is a database that provides STS data. This STS data originates from experiments like the STS-based map.

(cf. [Entrez, 2010u])

UniVec is a database for identifying segments within nucleic acid sequences that may have their source in a vector. Redundant sequences have been removed from the data and so UniVec comprises only one copy of each sequence segment. The majority of its entries stem from GenBank.

(cf. [NCBI, 2010v])

WGS sequencing projects describe incomplete genome data or chromosome data. The contigs in WGS projects are overlapping reads without any gap. There are several types of data in WGS projects:

- Nucleotide data
- Nucleotide data from environmental projects
- Protein data

Nucleotide data gets in the BLAST wgs database. If the nucleotide data originates in environmental projects it can be imported into the BLAST env_nt database too and protein data gets into the BLAST nr database.

(cf. [NCBI, 2010w])

Protein Sequences The protein sequences are:

- Conserved Domains Database (CDD)
- RefSeq (→ Molecular Databases/Nucleotide Sequences)
- Protein Clusters
- Proteins

(cf. [NCBI, 2010q])

CDD is a resource for protein annotation. The Conserved Domains Database (CDD) is a compilation of multiple sequence alignment models concerning ancient domains and full-length proteins, receivable as position-specific score matrices (PSSMs). The purpose is the quite fast detection of conserved domains of protein sequences. Some sources offer 3D-structure information.

(cf. [CDD, 2010a], [CDD, 2010b])

Protein Clusters is a database that offers access to a collection of related protein clusters. It contains RefSeqs that are encoded by genomes. Protein Clusters also grants access to additional information like analysis tools, annotations, domains, genomic neighborhood, phylogenetic trees, structures and links.

(cf. [Entrez, 2010p])

Proteins is a database that contains protein sequences originating in different sources among other things originating in GenBank or RefSeq.

(cf. [Protein, 2010])

Structures consists of:

- 3D Domains
- Molecular Modeling Database (MMDB)
- PubChem BioAssay
- PubChem Compound
- PubChem Substance

(cf. [NCBI, 2010q])

3D Domains describe compact structural domains. 3D Domains are detected automatically in the Molecular Modeling Database (MMDB).

(cf. [Entrez, 2010q])

MMDB belongs to Entrez. The MMDB is a database for macromolecular 3D structures. It contains homologs of protein sequences. The MMDB is also called Entrez Structure.

(cf. [Entrez, 2010q], [Entrez, 2010r])

PubChem BioAssay is a database that offers screens of chemical substances of PubChem Substance as well as descriptions which can be searched through.

(cf. [PubChem, 2010a])

PubChem Compound is a database that provides a description for PubChem Substance. Identity and similarity groups serve as means for cross-referencing and pre-clustering.

(cf. [PubChem, 2010b])

PubChem Substance is a database that provides descriptions of chemical samples and links.

(cf. [PubChem, 2010c])

Taxonomy is built of:

- Entrez Taxonomy
- TaxBrowser

(cf. [NCBI, 2010q])

Entrez Taxonomy is a database where the names of all organisms whose data is stored in genetic databases can be found. Entrez Taxonomy belongs to GenBank. GenBank also contains extinct species.

(cf. [Entrez, 2010l], [Entrez, 2010s], [Entrez, 2010h])

TaxBrowser allows to search for the organisms in Taxonomy respectively GenBank. Some sources are grouped by categories. Organisms which are often used in research projects can be accessed by following a link.

(cf. [Entrez, 2010t], [Entrez, 2010m])

Genomes

projects are:

- Cancer Chromosomes
- dbGAP (Genotypes and Phenotypes)
- Entrez Genomes
- Entrez Genome Project
- Map Viewer
- SKY/M-FISH & CGH-Database

(cf. [NCBI, 2010q])

Cancer Chromosomes is a part of Entrez databases. Cancer Chromosomes consists of three subdatabases. These are the NCI Mitelman Database of Chromosome Aberrations in Cancer, the NCI/NCBI SKY/M-FISH & CGH-Database and the NCI Recurrent Aberrations in Cancer.

(cf. [NCBI, 2009])

dbGaP aims to publish results of concerning studies about the interaction between genotypes and phenotypes. As database of Genotypes and Phenotypes (dbGaP) contains sensitive and non-sensitive data there are "open" and "controlled" as levels of access. The data is not protected by any intellectual property patents. The "Open-access-data" can be grouped into:

- Genotype-Phenotype Analyses
- Phenotypic Variables
- Studies
- Study Documents

The “Controlled Data” can be divided into:

- De-identified phenotypes and genotypes
- Pedigrees
- Pre-computed associations between genotype and phenotype

(cf. [NCBI, 2010e])

Entrez Genomes is a database that offers views for genetic data. This data consists of chromosomes, genomes, genetic maps, physical maps and sequence maps. The database is structured into:

- Archaea
- Bacteria
- Eucaryotae
- Plasmids
- Viruids
- Viruses

Entrez Genomes also offers information about chromosomes, genome assemblies and organelles.

(cf. [Entrez, 2010d])

Entrez Genome Project is a database that provides sequencing and mapping information of cellular organisms. The Entrez Genome Project is subdivided into organism-specific overviews. It serves as a portal delivering all information concerning an organism and external links. Projects that only contain data of organelles, phages, plasmids and viruses are not enclosed. The Entrez Genome Project and Entrez Genomes are companion databases.

(cf. [Entrez, 2010e], [Entrez, 2005b])

Map Viewer is a browser for a part of the organisms represented in Entrez Genomes. The superclasses of organisms represented are:

- Fungi
- Human
- Invertebrates
- Plants
- Protozoa
- Vertebrates

Map Viewer serves for searching and viewing complete genomes, chromosome maps and sequence details. It offers different types of view:

- Genome View
- Home Page of organisms
- Map View
- Sequence View

(cf. [Entrez, 2004a])

SKY/M-FISH & CGH-Database is a platform for molecular cytogenetic data. The name comes from Spectral Karyotyping (SKY)/Multiplex Fluorescence In Situ Hybridization (M-FISH) and Comparative Genomic Hybridization (CGH) which are all complementary fluorescent molecular cytogenetic techniques. SKY and M-FISH allow to show the human and mouse chromosomes simultaneously whereas CGH works with hybridization for visualizing DNA in tumor genomes.

(cf. [NCBI, 2010t])

5.1.3 NCBI Tools

The main tools are:

- Data Analysis Tools
- Entrez Tools
- FTP
- Programming Tools

(cf. [NCBI, 2010q])

Data Analysis Tools

consist of

- BLAST
  is used to compare sequences of genes or proteins with each other
- Gene Expression Tools
  are GEO, Serial Analysis of Gene Expression Tag to Gene Mapping (SAGEmap), Cancer Genome Anatomy Project (CGAP) and UniGene DDD.
- Genome Analysis Tools
  consist of COGs, Entrez Genomes, Map Viewer and SKY/M-FISH & CGH-Database
- Molecular Structure Analysis Tools
  are Conserved Domain Search (CDSearch), Cn3D and VASTSearch.

- Nucleotide Sequence Analysis Tools
  are composed of BLAST, Electronic PCR, Entrez Gene, Model Maker, ORF Finder, Organism specific resources, SAGEmap, Spidey, Splign, VecScreen and Viral Genotyping Tool.
- Protein Sequence Analysis and Proteomics Tools
  are composed of BLAST, BLink, CD Search, CDART, OMSSA and TaxPlot.

(cf. [NCBI, 2010q], [NCBI, 2007b])

Entrez Tools

consist of

- Batch Citation Matcher
  delivers 100 citations or less.
- Batch Entrez Nucleotides and Batch Entrez Proteins
  are used for uploads to different types of databases.
- Citation Matcher
  is used to get PubMed citations.
- Cubby
  is the log-in area.
- Entrez Data Model
  shows a view of all Entrez databases and their connections.
- Entrez Utilities
  aim to provide access to Entrez data outside the web query interface.
- LinkOut
  is an Entrez service for linking from PubMed to Entrez databases.
- Query All Entrez Databases
  is the cross-database search base of Entrez.

(cf. [NCBI, 2010q], [NCBI et al., 2010], [Entrez, 2010a], [NCBI, 2010r], [Entrez, 2004b], [Entrez, 2009], [NCBI, 2006a], [Entrez, 2010g])

FTP

provides links to many resources like data and software.

(cf. [NCBI, 2010b])

Programming Tools

consist of

- Entrez Programming Utilities
- Information Engineering Branch (IEB) Tools and the NCBI ToolBox
  where NCBI's software and databases are designed and built. The ToolBox is made of Data Encoding, Data Model and Programming Libraries and strongly used within NCBI pipelines and tools whereby some tools are available for the public. Extensible Markup Language (XML) is used for simpler views of the data and conversions are possible either by Full Data Conversion or by Targeted DTDs for end users.
- NCBI Toolkit
  provides modules of the overall framework.

(cf. [NCBI, 2010q], [Entrez, 2009], [IEB, 2002], [IEB, 2001], [IEB, 2003], [IEB, 2010], [NCBI, 2010o])

Education - Bookshelf

Bookshelf consists of:

- Coffeebreak
- Genes and Diseases

(cf. [NCBI, 2010q])

Coffeebreak shows biological reports on discoveries with usage of NBCI tools, which results in interactive tutorials. The NBCI tools and resources that have been used are highlighted. The articles presented is targeted on clinicians, molecular biologists, colleges and students.

(cf. [Dean and Johanna McEntyre, 2010])

Genes and Diseases is a miscellany website offering articles concerning genes and resulting diseases.

(cf. [NCBI, 2010l])

5.2 NRC-CISTI Strategic Initiatives

The NRC-CISTI focuses on three types of strategic initiatives:

- Cooperation with federal health departments and the scientific community to provide health, scientific and technical information
- Services to provide access to Canadian research data
- Services to provide access to international research data

(cf. [NRC-CISTI, 2010a])

5.2.1 Cooperations

Cooperations are:

- Canadian Virtual Health Library
- Federal Science eLibrary
- Health Canada-CISTI partnership for shared library service

(cf. [NRC-CISTI, 2010a])

Canadian Virtual Health Library

The Canadian Virtual Health Library is a bilingual, national network of health libraries. It is a service for professionals. The purpose of this project is to advance health care and development, education and research. Access to health information and services should be guaranteed nationally.

(cf. [CIHR, 2009])

eLibrary

eLibrary provides seamless access to full text scientific, technical and medical (STM) articles, e-journals and e-indexes to more than 24,000 federal governmental decision makers, policy makers and researchers to be competitive to university researchers. The eLibrary initiative is headed by Agriculture and Agri-Food Canada, Environment Canada, Fisheries and Oceans Canada, Health Canada, Natural Resources Canada, NRC-CISTI. The aim of this service are in science based programs and services.

(cf. [CIHR, 7 17], [CIHR, 2010a])

Health Canada-CISTI partnership for shared library service

The Health Canada-CISTI partnership for shared library service is a new service that started on April 1^{st} 2010.

(cf. [NRC-CISTI, 2010a])

5.2.2 Canadian research data

Canadian research data can be found at:

- NPArC ($\rightarrow$ The NRC Information Source Library)
- PubMed Central Canada ($\rightarrow$ The NRC Information Source Library)

(cf. [NRC-CISTI, 2010a])

5.2.3 International research data

International Data is offered at:

- CODATA Canada (→ Politics)
- DataCite (→ Politics)
- Gateway to Scientific Data (→ The NRC Information Source Library)
- Research Data Strategy Working Group (→ Politics)
- WorldWideScience.org

(cf. [NRC-CISTI, 2010a])

WorldWideScience.org

WorldWideScience.org serves as a global science gateway for searching Canadian and international databases and portals. It was launched in 2007. This service was established by a multilateral partnership. 65 nations like China, Russia, France and some Latin American countries are involved. WorldWideScience.org provides 400 million pages of science.

WorldWideScience.org has implemented the Microsoft Translator technology as a multilingual translation tool that allows to search and translate international literature in real-time as a $WorldWideScience.org^{BETA}$ version. Results can be translated into Chinese, English, French, German, Japanese, Korean, Portuguese, Spanish or Russian. Further languages are in progress. The target is to break the language barrier and to install research networks around the globe.

(cf. [NRC, 3 17], [CIHR, 2010b], [OSTI, 6 11], [of Energy, 2010])

5.2.4 The NRC Information Source Library

The National Research Council (NRC) Information Source Library offers the following services:

- Catalogue
- Discover
- NPArC
- Gateway to Scientific Data
- PubMed Central Canada

(cf. [NRC, 3 25])

Catalogue

Catalogue is a service that allows to search the National Research Council's Canada Institute for Scientific and Technical Information (NRC-CISTI) collection as well as the Canadian Agriculture Library (CAL) collection. Items can be ordered after registration with Infotrieve. The NRC-CISTI collection provides internationally published information about engineering, health sciences, physical and life sciences and technology. It contains books, reports, proceedings and journals. CAL provides information about Agriculture, Agri-Food Canada and portfolio partners like the Canadian Food Inspection Agency. It contains over 1,000,000 volumes including topics like dairying, veterinary medicine, entomology, pesticides, plant diseases, horticulture and soil science.

(cf. [NRC, 2010a], [NRC-CISTI, 3 29], [CAL, 2010])

Discover

Discover is a service of NRC-CISTI for buying articles. It contains more than 20 million STM articles, mainly published after 1993. Open Access articles are also available. Bought articles can be viewed and printed one time only but the files cannot be saved to the desktop. A special plug-in, that ensures this, is needed to use the service.

(cf. [NRC-CISTI, 2010b])

NPArC

NRC Publications Archive (NPArC) is an online repository containing NRC-authored articles of NRC projects. As NPArC is no full text database, not all documents are represented in full text form. There is no possibility to look only for full text documents.

(cf. [NRC, 2010c], [NRC, 2010d], [NRC, 3 29])

Gateway to Scientific Data

Gateway to Scientific Data is an approach to improve access to research data. The data consists of:

- Data Management and Curation
- Data Sets

Gateway to Scientific Data also offers resources for data management activities.

Data Management and Curation contains listings of policies and best practices for data management activities.

Data Sets comprise the topics aerospace, agriculture, astronomy and astrophysics, biochemistry, biological sciences, biotechnology, chemistry, climatology, construction engineering, crystallography, environment, forestry, fuel cells and hydrogen, genomics, geosciences, health sciences, information communication technology, magnetic resonance, manufacturing technologies, marine sciences,

materials, meteorology, metrology, nanotechnology, nutrition sciences, oceanography, physics, plant sciences, thermodynamics and water quantity data but not all areas contain contents.

(cf. [NRC, 3 31], [NRC, 2010b], [NRC, 3 21])

PubMed Central Canada

PMC Canada describes itself as:

> *a free archive of life science journals* ([Canada, 2010b])

It is based on the US PMC.

PMC Canada originates in a partnership between the NRC-CISTI, the Canadian Institutes of Health Research (CIHR) and the NLM. The PMC Canada contains PMC, PMC of UK (UKPMC) as well as Canadian peer-reviewed articles.

Scientists who receive funding from the CIHR have to ensure that their publications are freely accessible within six month of publication since January 1^{st} 2008. There are journals that make the results freely accessible at once, journals which will do this within six month and other journals that allow to publish them in a central or institutional repository within the set period of time.

SHERPA/RoMEO is a database that supports researchers to find journals adhering with CIHR policy. Researchers also have to store their data into the right database when publishing. The original datasets have to be kept for five years.

The purpose of open access publications is the possibility of new discoveries for Canadian or foreign researchers. Another function is to increase the impact of the results and to enlarge the circle of audience.

(cf. [Canada, 2010b], [CIHR, 4 29])

5.3 Summary International Portals

"International Portals" describe extra-European portals. The manifestation of the process of implementing portals offering freely available scientific content to the public is in quite different states in different countries. The U.S.A., Canada and the United Kingdom established their versions of PMC. These portals provide scientific and peer-reviewed articles, abstracts, data, annotations, tools etc. The structure of the NCBI services is not consistent so that many high quality services are difficult to locate. The terms of use leave very much scope of interpretation for the provider. It can happen to the scientist that the access to services is denied for his IP-address for an indefinite period of time without warning. Most sources and tools are external. The mainly used format for scientific data is the FAST-ALL (FASTA) format. Chinese scientists publish in English language journals, many of them are American journals.

Chapter 6

European Portals

6.1 The United Kingdom

6.1.1 UK PubMed Central

UKPMC is an online biomedical and health research resource that can be accessed freely. It contains at about 20 million peer reviewed, published articles, abstracts, clinical guidelines, PhD theses and provides access to databases. The concerning databases are:

- Agricola
- Chinese Biological Abstracts
- Citeseer
- European Patent Office Patents

UKPMC offers a Journal List where each journal can be searched. The Journal List also contains information about free access, participation level and volumes.

UKPMC was designed in cooperation with the NCBI, which is the funder of PubMed and PMC. Its purpose is to offer published, peer-reviewed biomedical and health research abstracts and articles that can be accessed freely. The UKPMC search function is grounded on EBI's CiteXplore database so not only UKPMC will be searched but also:

- Agricola
- Chinese Biological Abstracts
- Citeseer
- Patents
- PubMed

UKPMC can be searched for:

- All Clinical Trials
- All Practical Guidelines
- Biological Patents
- Meta-Analyses
- Ran. Controlled Trials
- Reviews
- UK Clinical Guidelines
- UK Research Reports
- Theses

A Grant Lookup Tool is also integrated.

(cf. [UKPMC, 2010d], [UKPMC, 2010c], [UKPMC, 2010a], [UKPMC, 2010b])

6.1.2 NERC Open Research Archive

The National Environment Research Council (NERC) repository, NERC Open Research Archive (NORA), is an open research archive containing mainly peer-reviewed, published materials by researchers funded by the Natural Environment Research Council. The concerning research centres are:

- British Antarctic Survey (BAS)
- British Geological Survey (BGS)
- Centre for Ecology & Hydrology (CEH)
- Proudman Oceanographic Laboratory (POL)
- Swindon Office

Metadata can be accessed freely to NORA. The usage of metadata for non-profit-purposes is also free. NORA contains research results in following areas:

- Atmospheric science
- Earth science
- Earth observation
- Environmental maths and statistics
- Marine
- Polar
- Science based archaeology

- Technology and e-science
- Terrestrial and freshwater science

It is possible to browse by:

- Centre
- NERC Author
- Programme
- Subject
- Year

(cf. [NERC, 2010a], [NERC, 2010d], [NERC, 2010b], [NERC, 2010c])

6.2 EMBL-EBI

European Molecular Biology Laboratory - European Bioinformatics Institute (EBML-EBI) is a part of EMBL. EMBL's mission is to conduce to biologic research and to provide access to bioinformatic services and data. EMBL is divided into:

- EMBL-EBI
- EMBL Grenoble
- EMBL Heidelberg (headquarters)
- EMBL Hamburg
- EMBL Monterotondo

The EBML-EBI has resources as well as expertises to provide services to support researchers in their work. It started as EMBL Nucleotide Sequence Data Library, that is now called EMBL-Bank, in 1980 at the EMBL laboratories in Heidelberg in Germany. The EMBL Nucleotide Sequence Data Library was the first nucleotide sequence database in the world and its aim was to be a central DNA sequence database. In its early days, information was extracted from literature. Later, data could be submitted directly to the database. As the need for research, providing services and collaboration with partners grew, the EBI moved towards research institutes. So, EBML-EBI is located in the Wellcome Trust Genome Campus in Cambridgeshire where also the Wellcome Trust Sanger Institute is situated since 1995. EBI is responsible for two databases - EMBL-Bank and Universal Protein Resource (UniProt), hosts EU projects and controls EU-funded networks.

(cf. [EMBL, 2010r], [EMBL, 2010b], [EMBL-EBI, 2010b])

6.2.1 Networks

The EBI controls three EU-funded Networks of Excellence:

- BioSapiens
- EMBRACE
- Experimental Network for Functional INtegration (ENFIN)

(cf. [EMBL, 2010b], [EMBL, 2010m])

6.2.2 EU Projects

The EBI hosts the following EU projects:

- 1,000 Genomes
- BioCatalogue
- BioSapiens
- E-MeP (no online resources and services; politics)
- European Genome-phenome Archive (EGA) (a database)
- ELIXIR (politics)
- EMBRACE (politics)
- EMERALD (politics)
- Experimental Network for Functional INtegration (ENFIN)
- Free European Life-science Information and Computational Services (FELICS)
- IMproving Protein Annotation through Coordination and Technology (IMPACT)
- International Nucleotide Sequence Database Collaboration (INSDC)
- Locus-Reference-Genomic (LRG)
- Microbial Pathway Genomics (MICROME)
- Serving Life-science Information for the Next Generation (SLING)
- Structural Proteomics in Europe (SPINE)
- Systems biology analysis of fungal pathogen interactions with the human immune system (SYBARIS)
- Synergies in Medical Informatics and Bioinformatics (SYMBIOmatics)

(cf. [EMBL-EBI, 2010b])

1000 Genomes

1,000 Genomes data can be accessed by the NCBI and the EBI. The 1,000 Genomes project was created to sequence genomes of a vast number of people to get a base of human genetic variation and to find genetic frequencies represented in a minimum of 1 % of the population surveyed. The data is stored in freely accessible databases or can be downloaded from ftp servers.

(cf. [EMBL-EBI, 2010b], [Genomes, 2010a], [Genomes, 2010b])

BioCatalogue

The BioCatalogue project acts as a registry of biological web Services. One of the main problems of researchers is the lack of overview of services and sources as there is no central catalogue. Another problem is the existence of services - but the absence of suitable documentation. The aim of BioCatalogue is to function as a central info point for web Services including a single registration point for providers and to provide a single search site. BioCatalogue is thought to procure a quality of service standard for Life Science web Services. These services shall be classified and annoted and their quality like availability or reliability shall be rated. This approach is intended to lead to more overview of already existing sources and services. The BioCatalogue can also be indexed and searched by search engines.

(cf. [EMBL-EBI, 2010b], [BioCatalogue, 2010a], [BioCatalogue, 2010b])

BioSapiens

BioSapiens is a EU funded network for bioinformatic research providing genome data by European laboratories as well as tools. Another goal of the project is to form a European Virtual Institute for Genome Annotation.

(cf. [EMBL-EBI, 2010b], [biosapiens, 2010])

ENFIN

Experimental Network for Functional INtegration (ENFIN) is a European Network of Excellence. ENFIN Partners are:

- Centre for Research & Technology Hellas (CERTH)
- Centro Nacional de Investigaciones Oncológicas (Spanish National Cancer Research Centre) (CNIO)
- Technical University Denmark (DTU)
- EMBL
- EBML-EBI
- GENOSCOPE
- Centre for Integrative Bioinformatics VU - vrije Universiteit amsterdam (IBIVU)

- Ludwig Institute for Cancer Research (LICR)
- Max-Planck Institute for Molecular Genetics (MPIMG)
- Medical Research Council Mammalian Genetics Unit (MRCMGU)
- QURETEC
- Swiss Institute of Bioinformatics (SIB)
- Technische Universität Braunschweig (TUBS)
- Humboldt University Berlin (UBER)
- University College London (UCL)
- University of Helsinki (UH)
- University of Basel (UNIBAS)
- University of Rome Tor Vergata (UNITOR)
- University of Dundee (UNIVDUN)
- Weizmann Institute of Science (WI)

ENFIN delves into an EU-wide integration of computational projects concerning system biology. The ENFIN network contains the set of analysis tools ENSuite and the data integration platform EnCORE.

(cf. [EMBL-EBI, 2010b], [ENFIN, 2010a], [ENFIN, 2010b])

FELICS

The goal of the Free European Life-science Information and Computational Services (FELICS) project is to organize and release all biomolecular information needed by European life science research. This project comprises databases and support for patent-related queries.

(cf. [EMBL-EBI, 2010b], [EMBL-EBI, 2010c])

IMPACT

IMproving Protein Annotation through Coordination and Technology (IMPACT) is an EU-funded project. The goal of IMPACT is to use existing technologies to improve already existing projects. Genomes and proteomes can be searched for protein signatures by different tools. All the different sources should be joined into one single source to improve and to simplify research - the InterPro database. The IMPACT project was also designed for providing related services like software and other databases.

(cf. [EMBL-EBI, 2010b], [IMPACT, 2008])

INSDC

The International Nucleotide Sequence Database Collaboration (INSDC) is an international sequence database. This project is the result of a collaboration between

- DDBJ
- European Nucleotide Archive (ENA)
- GenBank (NCBI)

since 1986 respectively 1987.

(cf. [EMBL-EBI, 2010b], [INSDC, 2010])

LRG

The Locus-Reference-Genomic (LRG) project provides the opportunity to store mutations at a stable genomic DNA framework. The LRG is the result of a collaboration between sequence providers, knowledge centers, databases providers and laboratories. The received data contains human genomic sequences as a reference standard for disease-causing human gene variants.

(cf. [EMBL-EBI, 2010b], [LRG, 2009a], [LRG, 2009b])

MICROME

Microbial Pathway Genomics (MICROME) was created as an EU Framework Programme 7 Collaborative Project. It is a scalable portal containing and producing different pathways like reference pathways, species-specific pathway assemblies, variants and global metabolic models. It enables to get comparative, evolutionary and functional details on the microbes.

(cf. [EMBL-EBI, 2010b], [microme, 2010])

SLING

Serving Life-science Information for the Next Generation (SLING) is a three-year project that is funded by the European Commission within Research Infrastructures of the FP7 Capacities Specific Program. The aim of SLING is to provide biomolecular information to expedite European science. SLING encompasses bioinformatic tools, databases, software tools and web services from its partners. The partners are:

- BRaunschweig ENzyme DAtabase (BRENDA) database at the Technische Universität Braunschweig (TUBS)
- EBML-EBI
- Enzymeta GmbH
- European Patent Office (EPO)
- Swiss Institute of Bioinformatics (SIB)

(cf. [EMBL-EBI, 2010b], [SLING, 2010])

SPINE

Structural Proteomics in Europe (SPINE) is a European research framework for the development of high-throughput structural biology.

(cf. [EMBL-EBI, 2010b])

SYBARIS

Systems biology analysis of fungal pathogen interactions with the human immune system (SYBARIS) is an EU-granted EU Framework Programme 7 Collaborative Project. The goal of this project is to explore drug resistance in pathogenic fungi and appendant cell-mediated response molecular at level and to find biomarkers of resistance and sensitivity. Therefore, wet-lab work is joined with bioinformatics and high-throughput computing. The resulting data is integrated in a knowledgebase of drug resistance genes.

(cf. [EMBL-EBI, 2010b], [SYBARIS, 2010a], [SYBARIS, 2010b])

SYMBIOmatics

The SYMBIOmatics Specific Support Action (SSA) project was created to get the current state of European bioinformatics and the medical area of informatics, to find synergies between them, to utilize them and to detect future fields of activity.

(cf. [EMBL-EBI, 2010b], [Rebholz-Schuhman et al., 2007])

6.2.3 EMBL Databases

Databases can be grouped into:

- Biological Ontology Databases
- Database Browsing & Entry Retrieval
- Literature Databases
- Microarray Databases
- Nucleotide Databases
- Pathway & Network Databases
- Protein Databases
- Proteomic Databases
- Small Molecule Databases
- Structure Databases

(cf. [Institute, 2010g], [Institute, 2010a], [Institute, 2010p], [Institute, 2010q], [Institute, 2010r], [Institute, 2010s], [Institute, 2010t], [Institute, 2010u], [Institute, 2010v])

Biological Ontology Databases

Biological Ontology Databases encompass:

- Chemical Entities of Biological Interest (ChEBI)
- Experimental Factor Ontology (EFO)
- Gene Ontology (GO)
- Ontology Lookup
- QuickGO
- Systems Biology Ontology (SBO)
- Taxonomy via UniProt
- UniProtKB-GOA (→ Protein Databases)

(cf. [Institute, 2010a])

ChEBI Chemical Entities of Biological Interest (ChEBI) is a free molecular and chemical lexicon.

(cf. [EMBL, 2010m], [EMBL, 2010h], [Institute, 2010a])

EFO Experimental Factor Ontology (EFO) is an ontology modeling application for experimental factors in ArrayExpress and to raise the number of annotations. Another goal of EFO is to integrate external data and to create mappings to domain specific ontologies.

(cf. [EMBL, 2010m], [EMBL, 2010v], [Institute, 2010a])

GO The responsibility of Gene Ontology (GO) is a controlled vocabulary for molecular, genetic and cellular matter within collaborating databases. The goal is to submit queries across the services.

(cf. [EMBL, 2010m], [EBI, 2010c], [Institute, 2010a])

Ontology Lookup The Ontology Lookup Service (OLS) is a spin-off of the PRoteomics IDEntifications database (PRIDE) project. The OLS queries ontologies which are available online whereas each of them comes with its own interface and output format.

(cf. [EMBL, 2010m], [EBI, 2010m], [Institute, 2010a])

QuickGO QuickGO is a browser. It is used for annotations and Gene Ontology terms of the UniProtKB-GOA group. It is possible to search for biological process terms, cellular component terms, molecular function terms, protein identifiers and names and GO terms.

(cf. [EMBL, 2010m], [EBI, 2010s], [EBI, 2010t], [Institute, 2010a])

SBO The Systems Biology Ontology (SBO) belongs to the BioModels project and contains a set of System Biology specific vocabulary.

(cf. [EMBL, 2010m], [EBI, 2010x], [Institute, 2010a])

Taxonomy via UniProt The Taxonomy database provides taxons of the tree structure of the relationships of organisms whereby the taxonomy data of UniProt Knowledgebase (UniProtKB) is manually curated.

(cf. [EMBL, 2010m], [Consortium, 2010b], [Institute, 2010a])

Database Browsing & Entry Retrieval

Database Browsing & Entry Retrieval comprises:

- ArrayExpress (→ Microarray Databases)
- BioMart
- EB-eye Search
- European Nucleotide Archive (ENA)
- FetchTools
- Integr8
- QuickGO (→ Biological Ontology Databases)
- UniProt DAS
- UniProt Search

(cf. [Institute, 2010g])

BioMart BioMart was designed as a query-orientated data management system by EBI and the Ontario Institute for Cancer Research (OICR). BioMart supports searching for complex descriptive data. The "out-of-the-box" website can be adapted by the user. BioMart offers graphical and text based applications.

(cf. [EMBL, 2010m], [EMBL and OICR, 2010], [Institute, 2010g])

EB-eye Search EB-eye Search is a search tool for getting biomolecular information from databases, literature and EBI-websites.

(cf. [EMBL, 2010m], [EMBL, 2010s], [Institute, 2010g])

ENA The European Nucleotide Archive (ENA) provides nucleotide sequencing information. The ENA contains annotations, assembled sequences and raw data.

(cf. [EMBL, 2010m], [EMBL, 2010u], [EMBL, 2010c], [Institute, 2010g])

EMBL-SVA The ENA Sequence Version Archive is a repository. It contains all entries of the EMBL-Bank Sequence Database.

(cf. [EMBL, 2010m], [EMBL, 2010z])

FetchTools The FetchTools consist of:

- Database fetch (dbfetch)
- emblfetch
- medlinefetch
- WSDbfetch

(cf. [Institute, 2010g])

dbfetch, emblfetch and medlinefetch Database fetch (dbfetch) was designed for getting entries from EBI databases or from the MEDLINE database. Usage from a browser or from within a script is possible. Database fetch (Dbfetch) is CGI based.

(cf. [EMBL, 2010m], [EMBL, 2010n])

WSDbfetch WSDbfetch is the SOAP based equivalent to dbfetch. It enables to find entries from natural sciences databases.

(cf. [Institute, 2010g], [Institute, 10 a])

Integr8 Integr8 is an ancestor of Ensembl Genomes. It is a web portal for deciphered genomes and corresponding proteomes. It includes data of other services, like:

- CluSTr
- EMBL Nucleotide Sequence Database
- Ensembl
- Genome Reviews
- InterPro
- International Protein Index (IPI)
- UniProt Knowledgebase
- UniProtKB-GOA

(cf. [EMBL, 2010m], [EBI, 2010g], [Institute, 2010g])

UniProt DAS The distributed annotation system (DAS) is a client-server system making International Protein Index (IPI), UniProt Archive (UniParc) and UniProt data available. Thereby one client supplies the user with data from multiple servers. DAS clients are:

- Dasty2
- Ensembl DAS client
- SPICE

(cf. [EMBL, 2010m], [EBI, 2010z], [Institute, 2010g])

UniProt Search UniProt Search is a tool for searching UniProt sections. These sections are:

- Core Data
- Information
- Supporting Data

(cf. [EMBL, 2010m], [Consortium, 2010c], [Institute, 2010g])

EMBL Literature Databases

Literature Databases are:

- CiteXplore
- MEDLINE (→ NCBI - Literature Databases - PubMed)
- OMIM (→ NCBI - Literature Databases)
- Patent Abstracts

(cf. [Institute, 2010p])

CiteXplore CiteXplore is a tool joining literature search as well as text mining.

(cf. [EMBL, 2010m], [EMBL, 2010j], [Institute, 2010p])

Patent Abstracts Patent Abstracts is as the name implies a tool for searching for patent abstracts.

(cf. [EMBL, 2010m], [EBI, 2010w], [Institute, 2010p])

Microarray Databases - ArrayExpress

The Microarray Databases are represented in EMBL by ArrayExpress. ArrayExpress in an international public archive for data from array-based platforms. It contains functional genomics experiments including gene expression. ArrayExpress allows to query and to download data. It is built of two parts:

- ArrayExpress Experiments Archive
- Gene Expression Atlas

ArrayExpress acts as an archive for data for publications and as a service for high-quality data put in a standard-format. NCBI GEO data, especially GEO experiments of GEO DataSets, are imported every week.

(cf. [EMBL, 2010m], [EMBL, 2010e], [EMBL, 2010f], [EMBL, 2010a], [Institute, 2010g])

ArrayExpress Experiments Archive The ArrayExpress Experiments Archive is a database responsible for the storage of complete data supporting publications.

(cf. [EMBL, 2010a])

Gene Expression Atlas The Gene Expression Atlas is a database using meta-analysis based summary statistics. The Atlas has semantic enrichments and is grounded on a subset of the ArrayExpress data. It is a statistically robust cross-platform framework for gene expression experiment results. The Gene Expression Atlas contains condition-specific gene expression patterns, genes and samples. It substitutes the ArrayExpress Data Warehouse.

(cf. [EMBL, 2010a], [EMBL, 2010d], [Institute, 2010g])

Nucleotide Databases

The Nucleotide Databases contain:

- Database of Genomic Variants Archive (DGVa)
- European Genome-phenome Archive (EGA)
- ENA (→ Database Browsing & Entry Retrieval)
- ENA Genomes Server
- Ensembl
- Ensembl Genomes
- Genome Reviews
- HUGO Gene Nomenclature Committee (HUGO Gene Nomenclature Committee (HGNC))
- IMGT/HLA

- IMGT/LIGM
- Immuno Polymorphism Database (IPD)
- Karyn's Genomes
- Ligand-Gated Ion Channel database (LGICdb)
- Parasites
- Patent Data Resources

(cf. [Institute, 2010q])

DGVa The Database of Genomic Variants Archive (DGVa) is a repository storing genomic variants at an individual level.

(cf. [EMBL, 2010m], [Institute, 2010q])

EGA The European Genome-phenome Archive (EGA) is a repository for genotype experiments and studies. There also exist Copy Number Variation (CNV) and SNP genotypes. The access of data depends on the study. It can be publicly available or limited.

(cf. [EMBL, 2010m], [EMBL, 2010t], [Institute, 2010q])

ENA Genomes Server The Genomes Server provides access to complete genomes and WGSs. It contains genomes of:

- Archaea
- Archaeal viruses
- Bacteria
- Eukaryota
- Organelles
- Phages
- Plasmids
- Viroids
- Viruses

(cf. [EMBL, 2010m], [EMBL, 2010x], [Institute, 2010q])

Ensembl Ensembl is a eukaryote genome database, especially for vertebrates. Ensembl contains special subunits for metazoa, plants, fungi, protists, bacteria and archaea.

(cf. [EMBL, 2010m], [EMBL, 2010p], [EMBL, 2010w], [Institute, 2010q])

Ensembl Genomes Ensembl Genomes consists of

- Ensembl Bacteria
- Ensembl Fungi
- Ensembl Metazoa
- Ensembl Plants
- Ensembl Protists

(cf. [EMBL, 2010m], [EMBL, 2010q], [Institute, 2010q])

Genome Reviews The Genome Review Database is used for a genomic sequence view of organisms with complexly deciphered genomes. It comprises genomes of:

- Archaea
- Bacterias
- Bacteriophages
- Eukaryota

Genome Reviews are receivable as MySQL database or flat file format.

(cf. [EMBL, 2010m], [EBI, 2010b], [Institute, 2010q])

HGNC The goal of the HUGO Gene Nomenclature Committee (HGNC) is to find one unique gene name and symbol for each human gene and to store accepted symbols in the HGNC database. This nomenclature is also used for the Human Genome Project. The same applies to pseudogenes, non-coding RNA, phenotype and genotype features. This approach was made to simplify communication and data exchange, especially in electronic form.

(cf. [EMBL, 2010m], [EMBL-EBI, 2010a], [Institute, 2010q])

IMGT/HLA The IMGT/HLA is a database storing information about sequences of the human major histocompatibility complex. It uses the nomenclature asserted by the WHO.

(cf. [EMBL, 2010m], [EBI, 2010e], [Institute, 2010q])

IMGT/LIGM The IMGT/LIGM Database is as its name says a database of the ImMunoGeneTics (IMGT). It was desigend for vertebrate immunoglobulin and the T cell receptor nucleotide sequences. It was contrived in 1989 by Laboratoire d'ImmunoGénétique Moléculaire (LIGM). It offers selective query with five search types:

- Annotation labels
- Catalogue

- Keywords
- References
- Taxonomy

(cf. [Institute, 2010q], [IMGT, 2010], [Guidicelli et al., 2006])

IPD The Immuno Polymorphism Database (IPD) was created as a centralized system to research polymorphism in immune system genes.

(cf. [EMBL, 2010m], [EBI, 2010i], [Institute, 2010q])

Karyn's Genomes Karyn's Genomes is a collection. It contains sequenced genomes, links and descriptions. Karyn's Genomes consists of four parts:

- Archaea
- Bacteria
- Eucaryotes
- Viruses

(cf. [Institute, 2010q], [EBInstitute, 2010e])

LGICdb The Ligand-Gated Ion Channel database (LGICdb) is a database that stores nucleic and proteic sequences of ion channel subunits.

(cf. [EMBL, 2010m], [EBI, 2010l], [Institute, 2010q])

Parasites Parasites is a parasite genome database. It allows to search parasite specific nucleotide sequence databases.

(cf. [EMBL, 2010m], [EBI, 2001], [Institute, 2010q])

Patent Data Resources Patent Data Resources at the EBI can be grouped into:

- Patent Abstracts
- Patent Chemical Compounds
- Patent Equivalent Data
- Patent Sequences

These sources can be accessed and searched by:

- ChEBI
- ChEBI Advanced
- CiteXplore

- EB-eye
- Non-redundant patent sequences
- Patent Nucleotides
- Patent Proteins
- SRS

(cf. [EMBL, 2010m], [EBI, 2010n], [Institute, 2010q])

Pathway & Network Databases

Pathway & Network Databases are composed of:

- BioModels
- Interactions (IntAct)
- Reactome
- Rhea

(cf. [Institute, 2010r])

BioModels BioModels is a database for mathematical models concerning biological matter and links to relevant stuff.

(cf. [EMBL, 2010m], [EMBL, 2010y], [Institute, 2010r])

IntAct Interactions (IntAct) was created to work with protein interaction data. Thererfore it provides analysis tools.

(cf. [EMBL, 2010m], [EBI, 2010f], [Institute, 2010r])

Reactome Reactome is a biological pathway database of human biology but it is not only possible to search Homo sapiens but also other species. The concerning species are:

- *Arabidopsis thaliana*
- *Bos taurus*
- *Caenorhabditis elegans*
- *Canis familiaris*
- *Danio rerio*
- *Dictyostelium discoideum*
- *Drosophila melanogaster*
- *Escherichia coli*
- *Gallus gallus*

- *Homo sapiens*
- *Mus musculus*
- *Mycobacterium tuberculosis*
- *Oryza sativa*
- *Plasmodium falciparum*
- *Rattus norvegicus*
- *Saccharomyces cerevisiae*
- *Schizosaccharomyces pombe*
- *Staphylococcus aureus N315*
- *Sus scrofa*
- *Taeniopygia guttata*
- *Xenopus tropicalis*

[Reactome, 2010b]

It also allows to start cross-species comparisons.
Reactome offers cross-reference to standard databases and services like:

- ChEBI small molecule databases
- Ensembl
- GO
- HapMap Genome Browser
- NCBI Entrez Gene
- KEGG Compound
- PubMed
- UCSC

(cf. [EMBL, 2010m], [Reactome, 2010a], [Reactome, 2010b], [Institute, 2010r])

Rhea Rhea is a database of chemical reactions. It contains manual annotations and was developed in cooperation with the Swiss Institute of Bioinformatics (SIB).

(cf. [EMBL, 2010m], [EBI, 2010v], [Institute, 2010r])

Protein Databases

Protein Databases are:

- CluSTr
- Catalytic Site Atlas (CSA)
- UniProtKB/Swiss-Prot Human Proteome Initiative (HPI)

- Integrated relational Enzyme database (IntEnz)
- InterPro
- IPI
- LGICdb (→ Nucleotide Databases)
- Protein and Associated NucleotideDomains with Inferred Trees (PANDIT)
- Patent Data Resources (→ Nucleotide Databases)
- UniProt
- UniProtKB-GOA
- UniProtKB Sequence/Annotation Version Archive (UniSave)

(cf. [Institute, 2010s])

CluSTr CluSTr is a database. It clusters IPI and UniProt Knowledgebase proteins into groups of related proteins by pairwise comparison of protein sequences. CluSTr also offers links to InterPro.

(cf. [EMBL, 2010m], [EMBL, 2010k], [Institute, 2010s])

CSA The Catalytic Site Atlas (CSA) is an enzyme database. CSA is responsible for recording catalytic residues and active sites in 3D structure.

(cf. [EMBL, 2010m], [EMBL, 2010g], [Institute, 2010s])

HPI The UniProtKB/Swiss-Prot Human Proteome Initiative (HPI) is an ancestor annotation programme of Core Data for Chordata (CD2). The HPI programme stopped in 2008. The goal of this programme is to annote all human protein sequences and their mammal equivalents. Quality standards are already specified by UniProtKB/Swiss-Prot.

(cf. [EMBL, 2010m], [SIB, 2010b], [Institute, 2010s])

IntEnz Integrated relational Enzyme database (IntEnz) is an enzyme nomenclature database. It is a collaboration project with the SIB also resulting in the ENZYME nomenclature database.

(cf. [EMBL, 2010m], [EBI, 2010h], [SIB, 2010a], [Institute, 2010s])

InterPro InterPro is a database for a prediction, classification and annotation of proteins at three levels:

- Superfamily
- Family
- Subfamily

(cf. [EMBL, 2010m], [Institute, 2010s])

IPI IPI offers protein sequence information of proteomes of higher eukaryotic organisms and manages cross-references between data sources. IPI gets its data from:

- Ensembl
- H-InvDB
- RefSeq
- The Arabidopsis Information Resource (TAIR)
- UniProtKB/Swiss-Prot
- UniProtKB/Translations of EMBL (TrEMBL)
- Vega

(cf. [EMBL, 2010m], [EBI, 2010j], [EBI, 2010k], [Institute, 2010s])

PANDIT Protein and Associated NucleotideDomains with Inferred Trees (PANDIT) stores sequence alignments and phylogenetic trees including protein domains. The project has been frozen in 2008 due to a shortage in funding support.

(cf. [EMBL, 2010m], [EBI, 2008], [Institute, 2010s])

UniProt UniProt stores information about proteins concerning structure and function. It consists of:

- UniProt Metagenomic and Environmental Sequences (UniMES)
- UniParc
- UniProtKB
  - UniProtKB/Swiss-Prot
  - UniProtKB/Translations of EMBL (TrEMBL)
- UniProt Reference Clusters (UniRef)

(cf. [EMBL, 2010m], [EBI, 2010y], [Consortium, 2010a], [Institute, 2010s])

UniProt/UniMES UniProt Metagenomic and Environmental Sequences (UniMES) is a database for metagenomic and environmental data.

(cf. [EBI, 2010y])

UniParc UniParc is a repository for sequences and sequence identifiers. It stores only protein sequences. These proteins are linked to other sources. Additional data is retrieved from other sources by cross-references. UniParc contains extracted data from:

- EMBL
- EMBL WGS
- Ensembl
- European Patent Office proteins
- FlyBase
- H-Invitational Database
- IPI
- Japan Patent Office proteins
- Protein Data Bank (PDB)
- PIR-PSD
- RefSeq
- TROME Database
- UniProtKB/Swiss-Prot
- UniProtKB/Translations of EMBL (TrEMBL)
- United States Patent and Trademark Office (USPTO) proteins
- WormBase

(cf. [EBI, 2010y], [Institute, 2010w])

UniProtKB UniProtKB is the central access point for protein data. It has two subunits - UniProtKB/Swiss-Prot and UniProtKB/Translations of EMBL (TrEMBL).

The UniProtKB contains important subunits:

- UniProtKB/Swiss-Prot is the manually annotated and reviewed section of UniProtKB. The UniProtKB/Swiss-Prot Knowledgebase contains protein annotations.
  (cf. [EMBL, 2010m], [Institute, 2010x], [Institute, 2010s], [EBI, 2010y])
- UniProtKB/TrEMBL is the automatically annotated and non-reviewed section of UniProtKB. The UniProtKB/TrEMBL Knowledgebase is a database storing computer-annotations for proteins.
  (cf. [EMBL, 2010m], [Institute, 2010x], [Institute, 2010s], [EBI, 2010y])

(cf. [EBI, 2010y])

UniProt/UniRef UniProt Reference Clusters (UniRef) is a database for covering the sequence space by combining UniProtKB and UniParc data. UniRef is a database for sequence merging and getting better and faster results when searching UniProt.

(cf. [EMBL, 2010m], [Institute, 2010y], [Institute, 2010s], [EBI, 2010y])

UniProtKB-GOA The UniProtKB-GOA is the Gene Ontology Database. It supports the UniProtKB and the IPI with protein annotations. The UniProtKB-GOA serves as a central dataset for multi-species databases.

(cf. [EMBL, 2010m], [Institute, 2010l], [Institute, 2010a])

UniSave UniProtKB Sequence/Annotation Version Archive (UniSave) acts as repository for UniProtKB/Swiss-Prot and UniProtKB/TrEMBL.

(cf. [EMBL, 2010m], [Institute, 2010z], [Institute, 2010s])

Proteomic Databases - PRIDE

The Proteomic Database is PRIDE. PRIDE acts as a database for proteomics data. It contains protein and peptide identifications. Protein accessions can be received from:

- Ensembl
- IPI
- NCBI gi numbers
- RefSeq
- UniProtKB
- UniProtKB IDs

(cf. [EMBL, 2010m], [EBI, 2010r], [Institute, 2010t])

Small Molecule Databases

Small Molecule Databases are:

- ChEBI (→ Biological Ontology Databases)
- ChEMBL Database
- EuroCarbDB
- PDBeChem (→ Structure Databases)
- RESID

(cf. [Institute, 2010u])

ChEMBL Database The ChEMBL Database concentrates on small, drug-like molecules, their 2D structure, abstracted bioactivities and calculated properties.

(cf. [EMBL, 2010m], [EMBL, 2010i], [Institute, 2010u])

EuroCarbDB EuroCarbDB is a database. As its name says it contains data about carbohydrate structures.

(cf. [Institute, 2010u])

RESID RESID is a protein modification database. It contains the concerning annotations and structures therefor.

(cf. [EMBL, 2010m], [EBI, 2010u], [Institute, 2010u])

Structure Databases

Structure Databases are:

- CSA (→ Protein Databases)
- database of secondary structure assignments for all of the entries in the Protein Data Bank (DSSP)
- Electron Microscopy Data Bank (EMDB)
- families of structurally similar proteins (FSSP)
- homology-derived structures of proteins (HSSP)
- Protein Data Bank in Europe (PDBe)
- PDBe NMR
- PDBeChem
- PDBeFold
- PDBeMotif
- PDBe Protein Interfaces, Surfaces and Assemblies (PDBePISA)
- PDBeView
- PDBsum
- Protein Function (ProFunc)

(cf. [Institute, 2010v])

DSSP The database of secondary structure assignments for all of the entries in the Protein Data Bank (DSSP) is a database that was made for standardizing secondary structure assignments at the Protein Data Bank (PDB).

(cf. [EMBL, 2010m], [EMBL, 2010l], [Institute, 2010v])

EMDB The Electron Microscopy Data Bank (EMDB) stores electron microscopy density maps of macromolecular complexes as well as subcellular structures.

(cf. [Institute, 2010v], [EBInstitute, 2010b])

FSSP The families of structurally similar proteins (FSSP) is a database containing additional information of the PDB. The FSSP stores structural alignments of proteins and extended structural families for protein chains.

(cf. [EMBL, 2010m], [EBInstitute, 2010c], [Institute, 2010v])

HSSP The homology-derived structures of proteins (HSSP) is a derived database. It hybridizes structural information and sequence information. Structural information is comprised of 2D and 3D data, sequence information of 1D data. It stores sequence homologues for every protein in the PDB whereby homologues are assumed to have an equal 3D structure to the PDB protein. The HSSP also provides implied secondary and tertiary structures.

(cf. [EMBL, 2010m], [EBI, 2010d], [Institute, 2010v])

PDBe The Protein Data Bank in Europe (PDBe) stores data of biological macromolecular structures. There are cooperations with other protein databases.

(cf. [EMBL, 2010m], [EBI, 2010a], [Institute, 2010v])

PDBe NMR NMR entries are stored and analyzed at the PDBe.

(cf. [Institute, 2010v], [EBInstitute, 2008])

PDBeChem PDBeChem is a PDBe service. It is a ligand library containing chemical component descriptions of PDBe entries.

(cf. [Institute, 2010v], [EBInstitute, 2009a])

PDBeFold PDBeFold is a tool for multiple and pairwise comparison and 3D alignment of protein structures.

(cf. [Institute, 2010v], [EBInstitute, 2010i])

PDBeMotif PDBeMotif is a very fast PDBe Java search tool joining chemical structure, protein sequence and 3D data. PDBeMotif was created to analyze characteristics of the binding sites of proteins or protein classes. It works on Linux, Mac, Solaris and Windows platforms.

(cf. [Institute, 2010v], [EBInstitute, 2009b])

PDBePISA PDBe Protein Interfaces, Surfaces and Assemblies (PDBePISA) works interactive. It is a tool created for querying the PDB archive, the Protein Interfaces, Surfaces and Assemblies (PISA) database, analyzing macromolar interfaces, receiving structures and precalculated results.

(cf. [Institute, 2010v], [EBInstitute, 2010d])

PDBeView PDBeView is the search tool for the PDBe allowing to search by text terms or the ID.

(cf. [Institute, 2010v], [EBInstitute, 2010a])

PDBsum The PDBsum offers an overview of PDB macromolecule structure data in the form of schematic molecule diagrams.

(cf. [EMBL, 2010m], [EBI, 2010o], [Institute, 2010v])

ProFunc Protein Function (ProFunc) is a server. The goal of the project is to derive likely biochemical functions of a protein from its 3D structure. In order to achieve this sequence- as well as structure-based methods are used to analyze PDB data.

(cf. [EMBL, 2010m], [EBI, 2010p], [EBI, 2010q], [Institute, 2010v])

6.2.4 EMBL Tools

EMBL provides a wide range of tools and services which are maintained by EMBL Heidelberg in Germany. This range of tools comprises:

- ID Mapping
- Literature
- Microarray Analysis
- Protein Function Analysis
- Sequence Analysis
- Similarity & Homology
- Structural Analysis
- Tools Miscellaneous

(cf. [Institute, 2010i])

ID Mapping

The ID Mapping tools at EBML-EBI are represented by Protein Identifier Cross-Reference (PICR).

(cf. [EBInstitute, 2010j])

PICR PICR is a web service. It uses a mapping-algorithm and its data bases on UniParc which serves as data warehouse. PICR enables protein cross-references and uses at about 90 source databases. The query can be limited by the species name and the mapping databases.

(cf. [EMBL, 2010o], [2010, 2010k])

Literature

The Literature tools are:

- EBI Web service for Medline information retrieval (EBIMed)
- Protein Corral
- Whatizit

(cf. [EBInstitute, 2010j])

EBIMed EBI Web service for Medline information retrieval (EBIMed) is a web service. It finds and dissects abstracts of MEDLINE to give an overview on connections between Drugs, GO, Species and UniProt data. Whatizit is a related application.

(cf. [2010, 2010a])

Protein Corral Protein Corral is a web service. It finds MEDLINE abstracts, examines them and shows an overview of connections between UniProt protein and gene names. After that, the associations are grouped by their confidence.

(cf. [EMBL, 2010o])

Whatizit Whatizit is a system for text processing. There exist also streamed service and web service versions of Whatizit. It allows to place text queries, to define the type of text and to select a pipeline. The system retrieves and searches MEDLINE abstracts.

(cf. [EMBL, 2010o], [2010, 2010n])

Microarray Analysis

Microarray Analysis Tools are:

- Bioconductor
- Expression Profiler

(cf. [EBInstitute, 2010j], [2010, 2010i])

Bioconductor Bioconductor is an open source and open development software tool for bioinformatics. It was created for the analysis of high-throughput genomic data. It uses R. Bioconductor can be utilized in following fields:

- Annotation
- High Throughput Assays
- Microarrays
- Sequence data

(cf. [EMBL, 2010o], [Bioconductor, 2010])

Expression Profiler The Expression Profiler is an open web-based platform. It is extensible. This platform was created for data analysis, sequence analysis, microarray gene expression, pattern analysis, statistics, machine learning and clustering.

(cf. [2010, 2010i])

Protein Function Analysis

The Protein Function Analysis encompasses:

- Annotation of Domains (AnDom)
- CluSTr Search
- FingerPRINTScan
- GFP-cDNA
- Inquisitor
- InterProScan
- Phobius
- Polymorphism Phenotyping (PolyPhen)
- PPSearch
- Pratt
- Prediction Of Unknown Sub-types (PROUST)
- Rapid Automatic Detection and Alignment of Repeats (Radar)
- Simple Modular Architecture Research Tool (SMART)

(cf. [Institute, 2010i], [EBInstitute, 2010f], [EMBL, 2010o])

AnDom Annotation of Domains (AnDom) is a web tool. Its purpose is to correlate structural domains with protein sequences. After that, they are sorted compliantly with Structural Classification of Proteins (SCOP).

(cf. [Institute, 2010i], [Institute, 2004])

CluSTr Search CluSTr Search is the search tool for the CluSTr database. It enables simple and advanced search.

(cf. [2010, 2010m])

FingerPRINTScan FingerPRINTScan is a tool for protein sequence matching by scores. It uses family definitions of the PRINTS database to group proteins. This allows to find evolutionary relationships in mass genome data. FingerPRINTScan uses motifs to qualify families like a fingerprint. Searching against these "fingerprints" can lead to a detection of distant relatives.

(cf. [EMBL, 2010o], [2010, 2010c], [2010, 2010d], [2010, 2010b])

GFP-cDNA The GFP-cDNA project serves as a platform for projects for characterizing and finding protein products. The underlying human open reading frames (ORFs) are known.

(cf. [Institute, 2010i], [GFP-cDNA, 2005])

Inquisitor Inquisitor is a tool that checks a protein sequence against the UniProt Knowledgebase and the Integr8 proteomes. The delivered result says if there is a match or which sequence represents the next related result. Another output is an analysis of the examined sequence. Inquisitor also uses FASTA and InterProScan.

(cf. [EMBL, 2010o], [2010, 10 am])

InterProScan InterProScan is a search tool working with InterPro database entries. It joins diverse protein signature recognition methods to find relationships in proteins. InterProScan is used to discover distant relationships in protein sequences and to classify them. Another target of this tool is to derive the protein function.

(cf. [EMBL, 2010o], [2010, 10 an], [2010, 2010e])

Phobius Phobius is a server. The amino acid sequence of a protein is used to predict transmembrane topology and signal peptides. Therefore the predictor uses a hidden Markov model.

(cf. [EMBL, 2010o], [2010, 2010j], [2010, 2010f])

PolyPhen Polymorphism Phenotyping (PolyPhen) is a tool for prognosticating. It foretells the effect of amino acid replacement on function and structure of human proteins. This is important because human genetic variation that induces phenotypes is coded in SNPs.

(cf. [Institute, 2010i], [PolyPhen, 2010b], [PolyPhen, 2010a])

PPSearch PPSearch is a tool for finding protein motifs in protein sequences. Most proteins can be classified by resemblances in their sequences. The patterns are stored in a special pattern database - the Database of protein domains, families and functional sites (PROSITE) which stores protein families and domains.

(cf. [EMBL, 2010o], [2010, 2010g])

Pratt Pratt is a pattern-matching tool. It was created for searching for conserved patterns in protein sequences. Pratt can read FASTA format and Swiss-Prot format.

(cf. [EMBL, 2010o], [2010, 2010h])

PROUST Prediction Of Unknown Sub-types (PROUST) is as its name says a service for prediction of unknown sub-types in protein classification. The service also comprises a server. PROUST can handle different formats of protein alignments. The aim of this project is to get relevant information about proteins and their function.

(cf. [Institute, 2010i], [PROUST, 2002a], [PROUST, 2002b])

Radar Rapid Automatic Detection and Alignment of Repeats (Radar) is an algorithm for segmenting protein sequences into repeats. This process is grouped into three phases:

1. Specification of repeat length
2. Optimize repeat boarders
3. Confirmation of distant repeats

Radar finds repeats regardless of the complexity of the sequence. This means it finds repeats that are gapped approximately, short composition biased and complex in architecture. Radar detects these repeats automatically, in other words no manual input is needed.

(cf. [EMBL, 2010o], [2010, 2010l], [2010, 10 ak])

SMART Simple Modular Architecture Research Tool (SMART) is a tool for annotating and identifying genetically mobile domains and studying their structure. The annotations cover details on function, phyletic distribution, taxonomy and structures.

(cf. [Institute, 2010i], [SMART, 2010a], [SMART, 2010b])

Proteomic Services

Proteomic tools are:

- Dasty2
- Database on Demand (DOD)

- IntEnz (→ Protein Databases)
- PRIDE (→ Proteomic Databases)
- UniProt DAS (→ Database Browsing & Entry Retrieval → FASTA)

Dasty2 Dasty2 is a web client. It is JavaScript-based and uses AJAX. It was designed for the visualization of protein sequence feature information by DAS servers. Dasty2 combines the data it obtained from the servers and offers a unified view of sequence-annotated features.

(cf. [EMBL, 2010o], [Institute, 2010o])

DOD Database on Demand (DOD) is a web-based tool for database preprocessing. It produces sequence databases that are formated in FASTA style. The sources for this process are:

- IPI_ARABIDOPSIS
- IPI_CHICKEN
- IPI_COW
- IPI_HUMAN
- IPI_MOUSE
- IPI_RAT
- IPI_ZEBRAFISH
- UniProt/Swiss-Prot
- UniProt/TrEMBL

Enzymes can be selected or defined for the processing optionally. Predefined enzymes are:

- Arg-C
- Arg-C/P
- Chymotrypsin
- Lys-C
- Lys-C/P
- Lys-N
- Trypsin
- Trypsin/P

(cf. [EMBL, 2010o], [Institute, 2010h])

Sequence Analysis

Sequence Analysis tools are:

- Agadir
- Align
- Alignment tools
- CENSOR
- Kalign
- MAFFT
- MUSCLE
- Statistical Analysis of Protein Sequences (SAPS)
- T-COFFEE

(cf. [Institute, 2010i])

Agadir Agadir is an algorithm for predicting the helical behavior of monomeric peptides where it observes short range interactions. Input parameters are ionic strength, pH and temperature.

(cf. [Institute, 2010i], [Agadir, 2010])

Alignment tools

EMBOSS EMBOSS is a tool for sequence comparison using pairwise alignment algorithms. The two subtools of EMBOSS are:

- Needle
- Water

EMBOSS CpGPlot, CpGReport and Isochore are programs for discovering genomic regions that contain many CpG patterns. These regions are restistant to methylation and so CpG patterns usually refer to genes that are frequently switched on.

Needle uses the Needleman-Wunsch global alignment algorithm to compare two sequences. The Needleman-Wunsch algorithm is a dynamic programming algorithm. The needle tool aims to find the best of all possible alignments of the sequences.

Water is a tool basing on the Smith-Waterman local alignment algorithm to check one sequence against another. The Smith-Waterman algorithm belongs to the class of dynamic programming algorithms.

EMBOSS Pepinfo, Pepwindow and Pepstats are tools for discovering, exposing and reading protein sequences and metrics and returning graphs.

Transeq is a tool for the translation of nucleic acid sequences into peptide or protein sequences in several types of forward or reverse frames. Therefore it can use standard genetic or non-standard codes.

(cf. [EMBL, 2010o], [Institute, 2010k], [emboss, 2010b], [emboss, 2010a], [2010, 10 bb], [Institute, 2010j], [2010, 10 ag])

ClustalW2 ClustalW2 is a tool for analyzing multiple alignments of protein sequences or DNA. Its main purpose is to find conserved sequence regions and to support experiments concerning modification and function of proteins. ClustalW2 also shows differences, identities and similarities of the examined sequences. Cladograms and phylograms show the calculated evolutionary relationship.

(cf. [EMBL, 2010o], [Institute, 2010d], [Institute, 2010e])

CENSOR CENSOR is a web server. Its purpose is aligning queries against a reference collection of repeats. CENSOR derives its name from "censoring" homologous portions by replacing aligned portions by masking symbols.

(cf. [EMBL, 2010o], [Institute, 2010c], [Institute, 2010n])

Kalign Kalign is a sequence algorithm for nucleotide and protein sequences. It uses string matching to calculate sequence distances. Kalign is able to align 1,500 sequences in less than 10 seconds.

(cf. [EMBL, 2010o], [2010, 10 bg], [2010, 2010o])

MAFFT Multiple Alignment using Fast Fourier Transform (MAFFT) is a sequence alignment tool. It is much faster than ClustalW2 and more than 100 times faster than T-COFFEE. It uses fast Fourier transform and a scoring system. MAFFT was created for Unix derivates but there exist also two versions for windows - a version using Cygwin and an all-in-one version.

(cf. [EMBL, 2010o], [2010, 10 ao], [2010, 10 ah], [Katoh, 2009], [Katoh, 2010])

MUSCLE MUltiple Sequence Comparison by Log-Expectation (MUSCLE) is an alignment tool. It can process protein sequence alignments faster than ClustalW2 and T-COFFEE. MUSCLE supports FASTAindexFASTA format as input format. Possible output formats are:

- ClustalW2
- ClustalW2 Strict
- FASTA
- HTML
- MSF

- Phylip interleaved
- Phylip Sequental

(cf. [EMBL, 2010o], [2010, 10 ai])

SAPS Statistical Analysis of Protein Sequences (SAPS) is a tool for assessing protein sequence properties by statistical criteria. Properties are:

- Amino acid types
- Clusters
- Compositonal biases
- Motifs
- Repetitive structures
- Spacings

The results are interesting regions proposed for experimental studies, protein applications or protein groupings. Supported species are:

- Bacillus subtilis
- Chicken
- Drosophila melanogaster
- Escherichia coli
- Frog
- Human
- Mouse
- Rat
- Saccharomyces cerevisiae

(cf. [EMBL, 2010o], [2010, 10 al])

T-COFFEE T-COFFEE is a multiple sequence alignment tool that enables the user to join results. These results can arise from other alignment programs.

(cf. [EMBL, 2010o])

Wise2 Wise2 consists of principle Wise2 programs and other Wise2 programs. Wise2 principle programs are:

- estwise
- estwisedb
- genewise
- genewisedb

Other Wise2 programs are:

- DNA Block Aligner (DBA)
- Protein Smith-Waterman (psw)
- Protein Smith-Waterman Database searching (pswdb)

(cf. [2010, 10 aq])

estwise, genewise, estwisedb and genewisedb Estwise and genewise are programs for the comparison of protein sequences or profile Hidden Markov Model (HMM) to DNA sequences. Estwisedb and genewisedb are the database searching versions belonging to estwise respectively genewise.

(cf. [2010, 10 aq])

GeneWise - DBA The DNA Block Aligner (DBA) is a tool for sequence comparison. Two sequences are submitted to the DBA form. The DBA got its name from aligning these two sequences that both have several collinear blocks of conservation. The model for doing so is a probabilistic finite state machine. These blocks can be disjoined by a variable length of DNA. They may also include one or two gaps. The blocks can be graduated into 65, 75, 85 or 95 percent identity. One block is estimated at about 1 % of the DNA sequence. The gaps are assumed to be linear gaps and are modeled at a probability of 0.05. These conserved blocks are suspected to be responsible for the regulation of genes.

(cf. [EMBL, 2010o], [Institute, 2010m])

psw and pswdb Protein Smith-Waterman (psw) is a program for smith-waterman alignments. Protein Smith-Waterman Database searching (pswdb) is the database for psw.

(cf. [2010, 10 aq])

PromoterWise PromoterWise is a query service for comparing DNA sequences. Inversions and translocations can be factored into the query.

(cf. [EMBL, 2010o], [2010, 10 ar])

Similarity & Homology

Similarity & Homology tools comprise:

- BLAST
- ENA (→ Database Browsing & Entry Retrieval)
- FASTA
- GGSEARCH
- GLSEARCH
- PSI-Search
- SSEARCH
- Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)

(cf. [Institute, 2010i], [EBInstitute, 2010g])

BLAST BLAST is used for protein and sequence data. It is a tool for checking a new sequence against just known ones. The purpose of BLAST is to deliver regions containing similar sequences. These regions are indicative for function and structure of this sequence. EBML-EBI provides access to the following BLAST services:

- Ensembl Genomes BLAST Services
- Ensembl WU-BLAST2 Programs
- NCBI-BLAST2 EVEC
- NCBI-BLAST2 Nucleotide
- NCBI-BLAST2 Protein
- Pattern Hit Initiated BLAST (PHI-BLAST)
- Position specific iterative BLAST (PSI-BLAST)
- WU-BLAST2 ASD
- WU-BLAST2 Nucleotide
- WU-BLAST2 Parasites
- WU-BLAST2 Protein

Allowed sequence formats are:

- GCG
- FASTA
- free text - raw sequence
- NBRF

- Phylip
- Protein Information Resource (PIR)
- UniProtKB/Swiss-Prot

(cf. [EMBL, 2010o], [Institute, 2010b], [2010, 10 aj])

PHI-BLAST Pattern Hit Initiated BLAST (PHI-BLAST) is a search tool. It regards one pattern that appears twice in the query sequence as two independent sequences. PHI-BLAST delivers a statistical analysis at the end of the search.

(cf. [2010, 10 aj])

PSI-BLAST Position specific iterative BLAST (PSI-BLAST) is an analysis tool. It can detect remote relative protein members of families whose connection cannot be identified by straight sequence comparison. PSI-BLAST can also be used for finding the function of unnotated hypothetical proteins. It is a search tool that compares a protein sequence against a database.

(cf. [EMBL, 2010o], [2010, 10 aj])

FASTA FASTA is a tool for fast similarity searching. It got its name because of its rapidity in protein comparison and nucleotide comparison.

General FASTA programs are:

- FASTA-Nucleotide
- FASTA-Protein

These programs search against the corresponding database - the nucleotide respectively protein database. Specialized FASTA programs are:

- FASTA-ASD server
- FASTA-GENOME server
- FASTA-LGIC Nucleotide server
- FASTA-LGIC Protein server
- FASTA Proteome server
- FASTA-SNP server
- FASTA-WGS server
- GGSEARCH
- GLSEARCH
- PSI-Search
- SSEARCH

(cf. [EMBL, 2010o], [2010, 10 ae])

FASTA-Nucleotide FASTA-Nucleotide is a tool for similarity searching that uses FASTA programs. It searches against databases of nucleotides.

(cf. [2010, 10 ae], [2010, 10 ad])

FASTA-Protein FASTA-Protein is a tool for similarity searching of proteins. It uses FASTA as well as GGSEARCH, GLSEARCH and SSEARCH programs to search against protein databases.

(cf. [2010, 10 ae], [2010, 10 af])

FASTA-ASD server FASTA Alternative Splicing Database (FASTA-ASD) server was created for sequence similarity search against databases. These databases are nucleotide or protein databases. An advantage of FASTA is that it can identify long regions with few similarities. Sequence similarity search against genome or proteome databases is also possible.

(cf. [2010, 10 ae], [2010, 10 aa])

FASTA-GENOME server FASTA-Genome server is a sequence similarity search tool. It allows to search against genome databases by FASTA tools.

(cf. [2010, 10 ae], [2010, 10 ab])

FASTA-LGIC Nucleotide server FASTA Ligand Gated Ion Channel Database Nucleotide (FASTA-LGIC Nucleotide) server is a FASTA tool for performing sequence similarity searches against the Ligand Gated Ion Channel Database.

(cf. [2010, 10 ae], [2010, 10 ac])

FASTA-LGIC Protein server FASTA Ligand Gated Ion Channel Database Protein (FASTA-LGIC Protein) server is a tool for sequence similarity searching. It enables the user to execute searches against the Ligand Gated Ion Channel Database.

(cf. [2010, 10 ae], [2010, 10 bc])

FASTA Proteome server The FASTA Proteome server is a tool for performing similarity sequence searches against proteome databases.

(cf. [2010, 10 ae], [2010, 10 bd])

FASTA-SNP server The FASTA-SNP server is a tool for searching SNP sequences. The target of this project is to provide a program for detecting the connection between genes and diseases and genes and drug responsiveness.

(cf. [2010, 10 ae], [2010, 10 be])

FASTA-WGS server The FASTA - Whole Genome Shotgun Sequence Similarity Search (FASTA-WGS) server is a tool for sequence similarity searching against the EMBL WGS database.

(cf. [2010, 10 ae], [2010, 10 bf])

GGSEARCH GGSEARCH is a tool that performs the Needleman-Wunsch algorithm. It compares a DNA or protein sequence against a protein database.

(cf. [2010, 10 ae], [2010, 10 af])

GLSEARCH GLSEARCH is a tool for assembling a DNA or protein sequence to a sequence database.

(cf. [EMBL, 2010o], [2010, 10 ae], [2010, 10 af])

PSI-Search PSI-Search joins PSI-BLAST and the Smith-Waterman algorithm. The aim of PSI-Search is to detect distantly related protein sequences by SSEARCH and BLASTpgp for the search iterations.

(cf. [EMBL, 2010o], [2010, 10 ae], [2010, 10 bh])

SSEARCH SSEARCH is a tool for searching databases whereby SSEARCH performs a Smith-Waterman search. SSEARCH provides:

- SSEARCH-Protein
- SSEARCH-Proteomes

These tools search against corresponding sequences.

(cf. [EMBL, 2010o], [2010, 10 ae], [2010, 10 ap])

STRING Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) is a database. It contains known and derived protein interactions.

(cf. [Institute, 2010i], [2010, 10 bi])

Structural Analysis

Structural Analysis tools contain:

- Coiled-coil prediction
- Distance matrix ALIgnment Lite (DaliLite)
- DisEMBL
- FOLD-X
- GlobPlot
- Interaction Prediction through Tertiary Structure (InterPreTS)

- MaxSprout
- PDBe Services
- Patterns In Non-homologous Tertiary Structures (PINTS)
- PROCOGNATE
- Structure Prediction Guide
- Tempura
- The ELM Server

(cf. [Institute, 2010i], [Institute, 2010f], [EBInstitute, 2010h])

Coiled-coil prediction Coiled-coil prediction is a tool that accepts FASTA format or raw sequences to predict coiled coils from protein sequences.

(cf. [2010, 10 ba], [Coil, 2010])

DaliLite Distance matrix ALIgnment Lite (DaliLite) is a tool for comparing protein structures. It was created for checking two structures against each other.

(cf. [EMBL, 2010o], [Institute, 2010f])

DisEMBL DisEMBL is a prediction tool for jumbled regions. These regions are situated within a protein sequence.

(cf. [DisEMBL, 2010])

FOLD-X FoldX is an algorithm for the calculation of the stability of proteins and protein complexes basing on interactions. FoldX is fast, qualitative and uses a full atomic description of the protein structure.

(cf. [FoldX, 2010])

GlobPlot GlobPlot is a web service. It was created to plot the tendency within a protein to order or disorder. GlobPlot also finds ordered regions and inter-domain segments containing motifs.

(cf. [Globplot, 2006])

InterPreTS Interaction Prediction through Tertiary Structure (InterPreTS) is a server. As its name indicates it was designed for "Interaction Prediction through Tertiary Structure". InterPreTS accepts FASTA format and uses BLAST for searching for homologues of structures. After that, it estimates the suitability of the templates for the interaction.

(cf. [InterPreTS, 2010])

MaxSprout MaxSprout is an algorithm for databases as web service. The aim of this algorithm is to calculate protein backbones and side chain co-ordinates.

(cf. [EMBL, 2010o], [2010, 10 bj])

PDBe Services PDBe services are:

- BIObar
- EMsearch
- NMR Representative
- OLDERADO
- PDBe Analysis
- PDBeChem (→ Structure Databases)
- PDBeChemSearch
- PDBeFold (→ Structure Databases)
- PDBeMapQuick
- PDBeMotif (→ Structure Databases)
- PDBePISA (→ Structure Databases)
- PDBe Template
- PDBeView (→ Structure Databases)
- Search OCA

(cf. [EMBL, 2010o], [2010, 10 ci])

BIObar The BIObar is a Firefox browser add-on. It is a toolbar for providing access to biological data resources. The BIObar allows to search and get data from more than 45 bioinformatics services.

(cf. [2010, 10 ci], [2010, 10 ce], [jswamin, 2010])

EMsearch EMsearch is an EMDB search tool.

(cf. [2010, 10 ci], [Velankar, 2009])

NMR Representative The NMR Representative is a tool for searching within a list of representatives NMR models by their PDB ID. These NMR models are supplied by OLDERADO.

(cf. [2010, 10 ci], [2010, 10 ck])

OLDERADO OLDERADO is a tool that provides clustering information for NMR entries. These entries are stored in the PDB. The PDB data has been compiled using NMRCORE and NMRCLUST.

(cf. [2010, 10 ci], [2010, 10 cf])

PDBe Analysis The PDBe Analysis is a service for academic and statistical analysis of the PDBe macromolecular structure database. The statistical analysis can be performed for data subsets, molecule based information and residue based information. PDBe Analysis also provides:

- Atom Statistics
- Database Browser
- Entry/Residue Statistics
- PDBemine
- Residue Statistics
- Structure Selection
- Structure Statistics

(cf. [2010, 10 ci], [2010, 10 cg])

PDBeChemSearch PDBeChemSearch is a tool for performing substructure search against more than 900,000 chemical compounds.

(cf. [2010, 10 ci], [2010, 10 ca])

PDBeMapQuick PDBeMapQuick is a cross-reference tool for protein structure data and databases, especially UniProt.

(cf. [2010, 10 ci], [2010, 10 cb])

PDBe Template PDBe Template is a service for searching the protein structure of small protein fragments. These structures can be fitted according to their chemical characteristics. PDBe Template bases on a database that contains protein fragments. These fragments have been detected by doing data mining of the protein databank.

(cf. [2010, 10 ci], [2010, 10 cc])

Search OCA Search OCA is a tool for searching OCA. OCA describes itself as "a browser-database for protein structure/function" [2010, 2006a]. OCA integrates the following external tools:

- G protein-coupled receptors DB (GPCRDB)
- Kyoto Encyclopedia of Genes and Genomes (KEGG)

- OMIM
- Orientations of Proteins in Membranes (OPM)
- PDB
- PDBselect
- Pfam
- PubMed
- SCOP
- Swiss-Prot

(cf. [2010, 10 ci], [2010, 2006a], [2010, 2006b])

PINTS Patterns In Non-homologous Tertiary Structures (PINTS) is a tool that returns amino acids that are located closely to each other in space. This does not mean that these amino acids lie close or co-linear in sequence. PINTS matches a protein structure against a database of patterns. It also compares a structural pattern, that contains more than 100 amino acids against a protein structures database.

(cf. [PINTS, 2003])

PROCOGNATE PROCOGNATE is a database. It contains cognate ligands for the domains of enzyme structures. The CATH, Pfam and SCOP databases classified the structures. To detect the ligands ENZYME and Kyoto Encyclopedia of Genes and Genomes (KEGG) database data was matched against PDB ligand. Then the chemical similarity was calculated.

(cf. [EMBL, 2010o], [2010, 2009])

Structure Prediction Guide The Structure Prediction Guide is a tool for foretelling a protein 3D structure. This prediction bases on the protein sequence.

(cf. [2010, 2002])

Tempura Tempura is a server. It enables the user to define the amino acids that are used to search by "reverse template" approach and to search against non-redundant samples of the PDB, single PDB code or an uploaded list of PDB codes. Tempura can be extended by RasMol. RasMol is a freely available molecular graphics software that shows molecules and their interactions in 3D.

(cf. [EMBL, 2010o], [2010, 10 cd])

The ELM Server The Eukaryotic Linear Motif resource server (ELMServer) is a tool for the prediction of functional sites in eukaryotic proteins. It finds functional sites by regular expressions respectively special patterns. It also includes context-based rules and logical filters to minimize the number of false positive results. The sequence that should be scanned or its Swiss-Prot/TrEMBL number can be entered. Context information is:

- Cell compartment(s)
- Species

The species can be input manually or selected from a predefined list. One or more cell compartments can only be selected from a list.

(cf. [ELM, 2010])

Tools Miscellaneous

Tools Miscellaneous include:

- Harvester
- pI
- Protein Colourer
- Readseq
- Shared Ortholog and Gene Order Tree Reconstruction Tool (SHOT)
- Simple Indexing and Retrieval System Web interface (SIRW)
- WebMol

(cf. [Institute, 2010i])

Harvester Harvester is a tool that crawls more than 50 sites of bioinformatical matter. It also crosslinks more than 40 bioinformatic data sources. Harvester also allows to search for:

- Arabidopsis
- Human
- Mouse
- Rat
- Zebrafish

(cf. [Harvester, 2010])

pI The pI is a service that determines the isoelectric point. The input is a sequence typed in uppercase letters. Lehninger Biochemie of 1979 serves as reference.

(cf. [2010, 1995])

Protein Colourer The Protein Colourer is a tool for coloring amino acid sequences. The coloration will orientate on the hydrophobicity of the amino acids or on their size. The coloration scheme can be changed.

(cf. [EMBL, 2010o], [2010, 10 ch])

Readseq Readseq is a tool for biosequence conversion.

(cf. [EMBL, 2010o], [2010, 10 cj])

SHOT Shared Ortholog and Gene Order Tree Reconstruction Tool (SHOT) is as its name says a "Shared Ortholog and Gene Order Tree Reconstruction Tool". It was created to calculate genome phylogenies. Organisms that can be selected are:

- Archaea
- Eubacteria
- Eukaryotae

(cf. [SHOT, 2002])

SIRW Simple Indexing and Retrieval System Web interface (SIRW) is a web interface for Simple Indexing and Retrieval System (SIR). SIRW is a tool for searching nucleotide and protein databases with keywords and patterns. The supported databases are:

- ENZYME
- Pdbfinder
- PROSITE
- RefSeq
- Sprot
- Sptrembl
- Swissnew
- Swiss-Prot

(cf. [SIRW, 2010])

WebMol WebMol is a tool for analyzation of structural information of molecules. It contains a graphical interface that enables different color codings for atom types, chain, secondary structure etc., animations and zoom.

(cf. [cmpharm, 1998])

6.3 Summary - European Portals

European portals and their services are all interlinked with the non-European ones. They usually cooperate with each other, offer a similar variety of services - but overlappings exist.

Part IV

Reconstruction of Organisms

Chapter 7

Arthropoda and non-biomineralized Fauna of the Herefordshire Lagerstätte

7.1 Gaining basic Data

The Herefordshire Lagerstätte keeps fossils of non-biomineralized fauna. This fauna originates from the Silurian. The fossils are preserved as 3D calcitic void infills in the form of calcareous nodules. This is of interest because petrifactions of other Konservat-Lagerstätten like Burgess Shale are normally only 2D or nearly 2D. The moulds show a carapace with well conserved details of the external surface but soft tissues like parts of the gut were also preserved very fine. Some of them have a dimension of less than a millimeter and others are a few centimeters large. The fauna mainly contains sponges, worms and arthropoda. Fossils do not show any signs of post-mortem violation or compression. The bigger part of the petrifactions can be identified as dark-colored sparry calcite shapes while the matrix is brighter. Some fossils keep organic material or siliceous skeletons. Splitting or cutting nodules only shows details of the organism along the rupture line or cut edge. This makes it difficult to interpret the structure of a fossil and its border to the matrix. Another approach to study the petrifaction is to isolate it from the matrix in a chemical way. This is possible if the fossil and the matrix are chemically different. A mechanical way to part the fossil from its matrix only works if the fossil is big and tough enough in structure. Serial slicing is a technique that allows to get cross-sections by clipping parallel planes. The gained data can be used to reconstruct 3D computer models. There are non-destructive approaches as well, like Computer Tomography (CT) and Magnetic Resonance Imaging (MRI). CT gives interesting results for 3D data on moulds while MRI is not an optimal solution for solid geological materials. The achieved information contains “slice-based” datasets like X-ray absorption and the extent of magnetically induced nuclear resonance. The fossils of Herefordshire could be scanned by CT but generally available

equipment is practical and attains adequate results for slice spacings and pixel sizes at about 0.3 mm to 0.5 mm. As special equipment for CT is too valuable the moulds of the Herefordshire Lagerstätte were not analyzed at high resolution. Data collection occurs by photography at regular intervals taken using a digital camera that is affixed to a microscope working at a resolution of 1,000 by 800 pixels in 24-bit color. The compound is fixed in resin. To make the specimens available for computer-aided studies, they were grounded layer by layer whereas each layer has a thickness of 30 μm. Between two grinding operations, the photos are taken and checked. Unnecessary structures can be removed by software. PhotoPaint was the software that was used for cropping, aligning and registration of the pictures partially using scripts. Before generating a video it is possible to adapt the photo data, e.g. reducing image size for reducing the size of the resulting video file. Another reason is that AVS only works with volumes up to 255 x 255 x 255 voxels maximum. AVSCONV requires a multiple of four as image width. This software is used to convert serially numbered data in Bitmap (BMP) format to AVS. These photos build the base for slice videos which consist of a series of the photos just mentioned. In this way a first primitive 3D video is generated.

7.2 Methods of Reconstruction

7.2.1 Surface Approach

The surface approach depends on a manual interpretation of the slices. Structures are marked by hand so that they are totally surrounded and form a 2D figure. These stacked 2D forms are put into 3D ones by algorithms by generating a mesh of trinangular facets. The advantage in this method is that it does not require high-quality data, e.g. photos. The disadvantage can be traced back to the fact that the fossils are interpreted manually by scientist. This work is not only costly in terms of time but means also a very subjective interpretation of the shapes. This method is suboptimal for Herefordshire fauna.

7.2.2 Volume-based Methods

3D images are composed of voxels. Voxels are regularly spaced volume elements and numbers that save the properties of the object at a special point. Medical applications usually use volume-based methods to get 3D data e.g. produced by Magnetic Resonance (MR) or CT scanners. In this way the mould is also interpreted. As the serial slice images produced using the grinding technique can be regarded as voxel data and structures are quite fine Herefordshire fauna is interpreted by volume-based methods.

Volume rendering subscribes a set of methods that base on analysis by ray. A computer computes a 2D image using information about the angels of the rays and their angles to the specimen. The 3D image results from mean-intensity volume rendering algorithms producing a virtual roentgenogram.

Another approach bases on calculating an isosurface using MR and CT data. An isosurface is a polygon mash surface. First the user selects a special intensity then the algorithm joins all according points. Then the algorithm builds the isosurface. Ray-tracing leads to the best results because it provides both -

objectivity of volume data and sharpness and so it was used for the Herefordshire fauna. Taking image data from sequential angles forms the basis of the computer video files showing the reconstructed organism.

(cf. [Siveter et al., 2007], [Sutton et al., 2001])

7.3 Summary - Reconstruction of Arthropoda

The Herefordshire Lagerstätten contain very well preserved arthropoda petrifactions in 3D structure. These fossils consist of calcitic infills. Slicing and CT are quite good methods for studying these fossils. Slicing is an invasive method. CT is non-invasive and so the method of choice for valuable and rare fossils. The data resulting from the analysis is used to create models of the arthropoda.

Chapter 8

Saurians

8.1 General Reconstruction of Saurians

Saurians are extinct animals, more exactly spoken vertebrates. Their bones are millions of years old and often not completely remained or in good condition e.g. deformed. They are valuable, very rare or unique so that it is interesting that they are not damaged or even destroyed for scientific research.

The reconstruction of organisms demands for fossils like skeletons. The better the skeletons of extinct organisms are preserved, the better the quality of reconstruction that can be achieved. Two main types of reconstruction exist:

- Physical reconstruction
- Digital reconstruction

There are a few techniques for usage of software and combinations of them for doing computer-aided research. There exist reconstructions of the whole body and of parts of it. If a digital 3D model of an extinct animal can be gained, this means that compared to creating a physical model a digital one has advantages in costs because it can be changed without repeating the whole reconstruction. This approach also saves costs and time. Variables and factors can be changed without rebuilding a completely new model and above all 3D reconstructions are much smaller than a model in original size.

(cf. [Bates et al., 2009])

8.2 3D Reconstructions of Saurians

The 3D reconstruction of saurians is a type of digital reconstruction. It requires Light Detection And Range (LiDAR), a laser scan system, software for Computer-Aided Design (CAD) and a fully mounted skeleton to be scanned. Additional information can be gained by CT scans.

Interesting points of the reconstruction of organisms are biomechanical descriptions like:

- Body mass
- Centre mass
- Inertial resistance
- Mass properties

for the body respectively its segments. This information is needed to get acceleration and translation movements and thus to get hints for diversification or extinction.

One approach of 3D reconstruction of a Tyrannosaurus rex bases on receiving mount coordinates by gridding the floor beneath it. Skeletal landmarks are then prepared using a CAD-package. Together with the CT-scan data of the lower part of the body this builds the base to generate a low resolution framework and to reconstruct cavities. A newer approach is laser scanning combined with computer modelling to get more detailed computer models of skeletons. Whole installed skeletons can be scanned and their spatial arrangement and articulation can also be rebuilt virtually. This allows an estimation of the animal's mass and its centers. The scanned data is processed by CAD to get body outline and density of body volume.

Detailed 3D reconstructions of body masses of whole animals can be achieved by LiDAR scanning - a type of laser scanning - and computer modeling techniques. LiDAR scanning makes it possible to attain a skeleton where all bones are at the right position with their joints. This reconstruction also delivers an insight into anatomical details like body cavities and organs and as a consequence the body segment masses, the center of mass and the moments of inertia. This process works without damaging the bones. LiDAR is a scanning method that allows to scan whole mounted skeletons and get subcentimetre data of surface geometry from a distance of up to 800 m as 3D geometrical data. As it is noninvasive and it allows to scan large objects it is ideal for palaeontology.

In study [Bates et al., 2009] comparing five non-avian dinosaurs of significant range in body size (two specimens of Tyrannosaurus, one specimen of Acrocanthosaurus atokensis, one specimen of Struthiominus sedens, one specimen of a juvenile Edmonotsaurus annectens) a RIEGL LMS-Z420i 3D terrestrial laser scan system was used. As this laser scan system is able to deliver adequately fine data and as it works with an eye safe near-infrared laser no extra safety measures were necessary. This facilitates its use in a museum. The laser scan system has an maximum error of 5 mm when capturing dense 3D point data and can be run using a 24V or 12V car battery and a laptop. The software used is RiSCAN PRO to manipulate the 3D data. To get full 3D data the museum gallery with the skeletons was scanned from different positions. To merge the point clouds of the skeletons received by the scanning procedures into one model the data was imported into the PolyWorks software. Then a minimum of three points in two overlapping scans are manually selected. Then a best-fit-algorithm calculates the alignment of the point cloud. Unwanted points are removed. The resulting skeleton is parted into its segments whose mass properties are elicited. An octree filter improves the quality of the point cloud and RiSCAN PRO generates a triangle mesh. The mesh and so the triangles are then reduced

to areas of low topographic variation. The body outline was created using the CAD software Maya whereby right side data is taken and copied to generate the whole skeleton. This is also a way to reproduce missing limbs. The body outlines were generated using Non-Uniform Rational B-Spline (NURBs). The lungs and airsacs, organs of low density, are created using the CAD programme by creating simple NURBs structures and remodelling them step by step until they fit. After the completion of the model mass and inertia properties are computed.

(cf. [Bates et al., 2009])

8.3 Summary Reconstruction of Saurians

The reconstruction of organisms was traditionally managed by building models. This is a quite practical way to study the appearance of it and it is a quite impractical way to study real organisms. A model is of course not constructed of bones and meat or of other materials that built the organism alive. The synthetic materials do not permit to calculate facts like mass centers, to give insights into organic structures or to simulate movements. A digitalized model enables a scientist to do all these things.

Part V

Appendix

Zusammenfassung

Diese Magisterarbeit beschäftigt sich mit den Möglichkeiten, die der Computer den Scientific Communities bietet. In der Mitte des letzten Jahrhunderts wurde das Internet als Unterstützung für Forscher im CERN erdacht. Heutzutage hat sich das Internet zu einem Werkzeug für den "normalen" Menschen entwickelt. Es enthält viel Inhalt jedoch nur ein sehr kleiner Teil davon ist wissenschaftlich. Nichtsdestotrotz eröffnet das Internet neue Möglichkeiten, da es derart verbreitet ist. Es gibt viele wissenschaftliche Quellen, die für die Öffentlichkeit frei zugänglich sind.

Suchmaschinen bieten eine Möglichkeit, diese Quellen zu erreichen, ist, danach zu suchen. Das kann durch Suchmaschinen realisiert werden. Suchmaschinen existieren um Webseiten, die von Interesse sind, zu finden. Eine spezielle Art von Suchmaschinen sind wissenschaftliche Suchmaschinen. Diese wissenschaftlichen Suchmaschinen sind Nischenprodukte. Sie haben einen äußerst geringen Marktanteil. Wissenschaftliche Suchmaschinen erlauben es, nach wissenschaftlichen Inhalten zu suchen. Einige wissenschaftliche Suchmaschinen liefern auch Wiki-Einträge oder semiwissenschaftliche Resultate. Der Grund für diesen Effekt ist, dass diese Suchmaschinen nicht nur für Forscher, sondern auch für Schüler und Studenten entwickelt wurden. Wissenschaftliche Suchmaschinen und semiwissenschaftliche Suchmaschinen sind oft auch Nebenprodukte von größeren Suchmaschinen.

Ein anderer Ansatz sind Portale. Portale bieten eine Reihe von wissenschaftlichen Werkzeugen in konzentrierter Form. Sie versorgen die Scientific Community mit frei zugänglichen akademischen Publikationen. Sie bieten auch Daten und annotierte Daten. Die meisten dieser Sourcen sind extern. Diese Sourcen werden von verschiedenen Typen von Partnern bereitgestellt. Diese Partner können Unversitäten, staatliche Institutionen, Forschungsinstitute etc. sein. Die Partner aller Kooperationen der Portale sind international. PMC-Portale sind in den USA, in Kanada und im Vereinigten Königreich von Grossbritannien implementiert. Ein anderes Portal ist das EBI-Portal. Alle Portale nützen internationale Services. Es gibt Überschneidungen in den Resourchen der Portale. Die Portale sind miteinander verbunden.

Wissenschaftlicher Fortschritt ist für Regierungen von Interesse. Entwicklungen können in der Industrie, im Umweltmanagement, in der Medizin oder im sozialen Bereich angewandt werden. Einige Projekte werden von Regierungsorganisationen finanziert. Die verantwortlichen staatlichen Organe sind Konsortien und Komiteen, die Daten zu den aktuellen Verhältnissen zusammentragen und die wissenschaftliche Entwicklung vorantreiben. Die gesponserten Sourcen sind Portale, Datenbanken, Werkzeuge, Publikationen etc.

Ein weiterer Teil von computergestützter Forschung sind Rekonstruktionen von Organismen, besonders von Fossilien. Es gibt destruktive und zerstörungsfreie Methoden. Destruktive Methoden werden nicht bei wertvollen und/oder seltenen Versteinerungen angewandt. Ein Beispiel einer destruktiven Methode ist das Splicing. Splicing wird für die Erforschung des Aufbaus von Arthropoda angewandt. Die Scheiben werden fotographiert und dann wird die 3D-Struktur generiert.

Wertvolle und äußerst seltene Fossilien erfordern zerstörungsfreie Methoden. Zerstörungsfreie Methoden entstammen der medizinischen Untersuchung. Es sind dies MRI, Röntgen und CT. Röntgen ist eine ältere Technik, die hauptsächlich am Ende des letzten Jahrhunderts eingesetzt wurde, z.B. für die Forschung an Mumien. Eine weitere nichtinvasive Me-

thode ist LiDAR. LiDAR erlaubt es große Objekte aus großer Entfernung zu scannen. Das ist optimal für Skelette großer Tiere, wie Dinosaurier. Neue Techniken erlauben es digitale Modelle an Computer zu erstellen. Diese Computermodelle sind sehr praktisch, wenn die Fossilien, z.B. die Skelette nicht vollständig erhalten sind. Eine Kopie des erhaltenen Teils kann gespiegelt und ins Modell eingefügt werden. Digitale Modelle bieten den Vorteil, dass sie geändert werden, für Simulationen und Berechungen genutzt werden können. Ein weiterer Vorteil von digitalen Modellen ist die Möglichkeit Strukturen innerhalb von Körperhöhlen, wie das Gehirn, zu erforschen.

Abstract

This master thesis concentrates on the possibilities for the scientific communities caused by the computer. In the middle of the last century the Internet was conceived as a support for researchers at CERN. Nowadays, the Internet has evolved into a tool for the "normal" human. It contains much stuff but only a very small part of it is academic. Nevertheless, the Internet opens up new possibilities because it is so common. There exist many academic sources that are accessible freely for the public.

One way to get these sources is to search for them. This is realizable by search engines. Search engines exist to find websites of interest. A special type of search engines are scientific search engines. These scientific search engines is niche products. They have a very small market share. Scientific search engines allow to search for academic topics. Some scientific search engines also deliver Wiki-entries or other semi-scientific results. The reason for this effect is, that these search engines are not only designed for researchers but also pupils and students. Scientific search engines or semi-scientific search engines are also often a side-product of larger search engines.

Another approach are portals. Portals provide a range of scientific tools in a concentrated manner. They also supply the scientific community with freely accessible academic publishings. They also offer data and annotated data. Most of these sources are external. These sources are provided by several types of partners. These partners can be universities, governmental institutions, research institutes etc. The partners of all cooperations of portals are international. PMC portals are implemented in the U.S.A., in Canada and in the United Kingdom. Another portal is the EBI portal. All portals use international services. Overlappings of resources of portals exist. The portals are interlinked.

Scientific progress is interesting for the governments. Developments can be implemented in industry, ecology management, medicine or social matter. Several projects are funded by governmental organizations. The responsible governmental organs are consortia and committees that gather data of the actual condition and promote the scientific development. The sponsored sources are portals, databases, tools, publications etc.

Another part of computer-aided research is reconstructions of organisms, especially fossils. There exist destructive and non-destructive methods. Destructive methods are not used for valuable and/or rare petrifactions. An example for a destructive method is splicing. Splicing is used for exploring the construction of arthropoda. The splices are photographed and then a 3D structure is generated.

Valuable and very rare fossils call for non-destructive methods. Non-destructive methods stem from medical examination. They are MRI, roentgen and CT. Roentgen is an older technique that was mainly used at the end of the last century, e.g. for the study of mummies. Another non-invasive method is LiDAR. LiDAR allows to scan huge objects from large distances. This is optimal for skeletons of large animals like dinosaurs. New techniques allow to build digitalized models on the computer. These computer models are quite practical if the fossils, e.g. the skeletons are not completely conserved. A copy of the non-destroyed part can be mirrored and inserted into the model. Digitalized models have the advantage that they can be changed, used for simulations and calculations. Another advantage of digitalized models is the possibility of researching structures within spaces of the body like the brain.

Iris Emanuela Studeny - Curriculum vitae

Education

2007 – 2008	Bakk.rer.soc.oec. University of Vienna and Technical University of Vienna, Informatikmanagement
2002 – 2010	Bakk.techn. University of Vienna and Technical University of Vienna, Software & Information Engineering

Masterthesis

Title	Computer-aided research in natural sciences
Supervisors	Ao.Univ.Prof. Dipl.-Ing. Dr.sc.med. Dr.techn. Dr.rer.nat. Frank Rattay
Description	Internet, portals, politics, reconstructions

Experience

2010

Volunteer at the ICCHP, Technical University of Vienna, Vienna.
2010/07/14 – 2010/07/16 Volunteer at the International Conference on Helping People with Special Needs.
Detailed achievements:

- Access control;
- Support of presenters;
- Support of people with special needs;
- Video recording;
- Photography

Language

German	Native
English	Fluent
French	Very good

Computer Skills

Programming languages	C/C++, C#, Java, Lisp, Prolog, VB/VBA
Modeling	ArgoUML, Dia, MS Visio, StarUML
Internet	HTML, CSS, JavaScript, PHP, VBScript
IDE	Bloodshed, Eclipse, MS Visual Studio, NetBeans
CMS & E-Learning	Joomla, Moodle, Wiki
Data	MS SQL Management Studio Express, SQL, XML, XSLT
Operating Systems	Linux, Windows
Office & Documents	Latex (TeXnicCenter), MS Office, OpenOffice
Digital image processing	Gimp, Paint.Net
Other	Clementine (Data Mining), OpenProj (Project Management), Kompozer (Websites)

Interests

Macro photography	Insects, Snails
Reading	Biographies, Ibáñez, Kishon, Dahl
Traveling	Cultural tourism
My house plants	Orchids, Succulents

Vösendorf, 14th October 2010

Index of Keywords

Glossary

Arabidopsis thaliana

a plant also called “thale cress”, “common wallcress” or “mouse-ear cress” 32

BackRub

predecessor of Google 15

Bielefeld Academic Search Engine

16, 115

Bing

search engine of Microsoft 18

Caenorhabditis elegans

a model organism of genetics research 32

CiteSeerX

a scientific library and search engine 16

code.google.com

`http://code.google.com/` 15

Drosophila melanogaster

an animal also called “fruit fly” or “vinegar fly” 32

Eccellio Science

a search engine for scientific results 16

Escherichia coli

a species of bacteria 32

Firefox

a free browser 16

FlyBase

a genetics database of Drosophila 31, 68

Google

search engine 15, 18

Google Scholar

a search engine for qualitative results 15

Homo sapiens

human 32

iLib2

Chinese scientific search engine 16

Memex

memory extender 13

Mosaic

first graphical browser 13

Mus musculus

an animal also called “house mouse” 32

National Center for Biotechnology Information

National Center for Biotechnology Information [NCBI, 2004] 27, 117

OCA

Oca is the spanish word for goose. [2010, 2006a] 88

PhotoPaint

software 94

Rattus norvegicus

an animal also called “brown rat” or “Norway rat” 32

Saccharomyces cerevisiae

brewer’s yeast 32

ScienceSequere

a search engine 16, 17

Scirus

a search engine 16, 17

Silurian

Geologic era 94

List of abbreviations

AnDom	Annotation of Domains
BAC	Bacterial Artificial Chromosomes
BASE	Bielefeld Academic Search Engine
BLAST	Basic Local Alignment Search Tool
BMP	Bitmap
BRC	Biological Research Center
BRENDA	BRaunschweig ENzyme DAtabase
CAD	Computer-Aided Design
CAL	Canadian Agriculture Library
CCDS	Consensus CoDing Sequence, Consensus CDS
CDD	Conserved Domains Database
CDSearch	Conserved Domain Search
CD2	Core Data for Chordata
CERN	European Council for Nuclear Research
CERTH	Centre for Research & Technology Hellas
CGAP	Cancer Genome Anatomy Project
CGH	Comparative Genomic Hybridization
ChEBI	Chemical Entities of Biological Interest
CIHR	Canadian Institutes of Health Research
CIPF	Centro de Investigación Príncipe Felipe
CISTI	Canada Institute for Scientific and Technical Information
CNIO	Centro Nacional de Investigaciones Oncológicas (Spanish National Cancer Research Centre)
CNV	Copy Number Variation
CODATA	Committee on Data for Science and Technology
COG	Clusters of Orthologous Groups
CSA	Catalytic Site Atlas
CT	Computer Tomography
DaliLite	Distance matrix ALIgnment Lite
DAS	distributed annotation system
DBA	DNA Block Aligner
dbEST	Expressed Sequence Tags
Dbfetch	Database fetch
dbfetch	Database fetch

dbGaP	database of Genotypes and Phenotypes
dbGSS	Genome Survey Sequence
dbMHC	MHC database
dbSNP	Single Nucleotide Polymorphisms
dbSTS	Sequence Tagged Sites
DDBJ	DNA Data Bank of Japan
DGVa	Database of Genomic Variants Archive
DOD	Database on Demand
DSSP	database of secondary structure assignments for all of the entries in the Protein Data Bank
DTU	Technical University Denmark
EBI	European Bioinformatics Institute
EBIMed	EBI Web service for Medline information retrieval
EFO	Experimental Factor Ontology
EGA	European Genome-phenome Archive
ELIXIR	European Life Sciences Infra Structure for Biological Information
ELMServer	Eukaryotic Linear Motif resource server
EMBL	European Molecular Biology Laboratory
EBML-EBI	European Molecular Biology Laboratory - European Bioinformatics Institute
EMBRACE	European Model for Bioinformatics Research and Community Education
EMDB	Electron Microscopy Data Bank
E-MeP	European Membrane Protein Consortium
EMERALD	Empowering the Microarray-Based European Research Area to Take a Lead in Development and Exploitation
ENA	European Nucleotide Archive
ENFIN	Experimental Network for Functional INtegration
EPO	European Patent Office
EST	Expressed Sequence Tag
FASTA	FAST-ALL
FASTA-ASD	FASTA Alternative Splicing Database
FASTA-LGIC Nucleotide	FASTA Ligand Gated Ion Channel Database Nucleotide
FASTA-LGIC Protein	FASTA Ligand Gated Ion Channel Database Protein
FASTA-WGS	FASTA - Whole Genome Shotgun Sequence Similarity Search
FELICS	Free European Life-science Information and Computational Services
FGED	Functional Genomics Data
FSSP	families of structurally similar proteins

GENSAT	Gene Expression Nervous System Atlas
GEO	Gene Expression Omnibus
Gmail	Google Mail
GO	Gene Ontology
GPCRDB	G protein-coupled receptors DB
GSS	Genome Survey Sequence
HGNC	HUGO Gene Nomenclature Committee
HMM	Hidden Markov Model
HPI	UniProtKB/Swiss-Prot Human Proteome Initiative
HSSP	homology-derived structures of proteins
HTML	hypertext mark-up language
HTTP	hypertext transfer protocol
IBIVU	Centre for Integrative Bioinformatics VU - vrije Universiteit amsterdam
ICSU	International Council for Science
IEB	Information Engineering Branch
IMGT	ImMunoGeneTics
IMPACT	IMproving Protein Annotation through Coordination and Technology
INSDC	International Nucleotide Sequence Database Collaboration
IntAct	Interactions
IntEnz	Integrated relational Enzyme database
InterPreTS	Interaction Prediction through Tertiary Structure
IPD	Immuno Polymorphism Database
IPI	International Protein Index
IRMM	Institute for Reference Materials and Measurements
KEGG	Kyoto Encyclopedia of Genes and Genomes
LANL	Los Alamos National Laboratory
LGC	Laboratory of the Government Chemist
LGICdb	Ligand-Gated Ion Channel database
LICR	Ludwig Institute for Cancer Research
LiDAR	Light Detection And Range
LIGM	Laboratoire d'ImmunoGénétique Moléculaire
LRC	Leukocyte Receptor Complex
LRG	Locus-Reference-Genomic
MAFFT	Multiple Alignment using Fast Fourier Transform
MEDLINE	Medical Literature Analysis and Retrieval System Online
MeSH	Medical Subject Headings

M-FISH	Multiplex Fluorescence In Situ Hybridization
MGED	old name of FGED
MGI	Mouse Genome Informatics
MHC	Major Histocompatibility Complex
MIAME	Minimum Information About a Microarray Experiment
MICROME	Microbial Pathway Genomics
MIM	Mendelian Inheritance in Man
MMDB	Molecular Modeling Database
MMRRC	Mutant Mouse Regional Resource Center
MPIMG	Max-Planck Institute for Molecular Genetics
MR	Magnetic Resonance
MRCMGU	Medical Research Council Mammalian Genetics Unit
MRI	Magnetic Resonance Imaging
MUSCLE	MUltiple Sequence Comparison by Log-Expectation
NCBI	National Center for Biotechnology Information
NCSA	National Center for Supercomputing Applications at the University of Illinois at Urbana-Champagne
NERC	National Environment Research Council
NIH	National Institutes of Health
NINDS	National Institute of Neurological Disorders and Stroke
NLM	National Library of Medicine
NMC	Norwegian Microarray Consortium
NORA	NERC Open Research Archive
NPArC	NRC Publications Archive
NRC	National Research Council
NRC-CISTI	National Research Council's Canada Institute for Scientific and Technical Information
NTNU	Norges Teknisk-Naturvitenskapelige Universitet
NURBs	Non-Uniform Rational B-Spline
OICR	Ontario Institute for Cancer Research
OLS	Ontology Lookup Service
OMIA	Online Mendelian Inheritance in Animals
OMIM	Online Mendelian Inheritance in Man
OPM	Orientations of Proteins in Membranes
ORFs	open reading frames
PANDIT	Protein and Associated NucleotideDomains with Inferred Trees
PDB	Protein Data Bank
PDBe	Protein Data Bank in Europe
PDBePISA	PDBe Protein Interfaces, Surfaces and Assemblies

PHI-BLAST	Pattern Hit Initiated BLAST
PICR	Protein Identifier Cross-Reference
PINTS	Patterns In Non-homologous Tertiary Structures
PIR	Protein Information Resource
PISA	Protein Interfaces, Surfaces and Assemblies
PMC	PubMed Central
PMC Canada	PubMed Central Canada
PolyPhen	Polymorphism Phenotyping
PRIDE	PRoteomics IDEntifications database
ProFunc	Protein Function
PROSITE	Database of protein domains, families and functional sites
PROUST	Prediction Of Unknown Sub-types
PSI-BLAST	Position specific iterative BLAST
PSSMs	position-specific score matrices
psw	Protein Smith-Waterman
pswdb	Protein Smith-Waterman Database searching
Radar	Rapid Automatic Detection and Alignment of Repeats
RBC	Red Blood Cells
RefSeq	Reference Sequences
SAGEmap	Serial Analysis of Gene Expression Tag to Gene Mapping
SAPS	Statistical Analysis of Protein Sequences
SBO	Systems Biology Ontology
SCOP	Structural Classification of Proteins
SGD	Saccharomyces Genome Database
SHOT	Shared Ortholog and Gene Order Tree Reconstruction Tool
SIB	Swiss Institute of Bioinformatics
SIR	Simple Indexing and Retrieval System
SIRW	Simple Indexing and Retrieval System Web interface
SKY	Spectral Karyotyping
SLING	Serving Life-science Information for the Next Generation
SMART	Simple Modular Architecture Research Tool
SNP	single nucleotide substitutions and short deletion and insertion polymorphisms
SPINE	Structural Proteomics in Europe
SSA	SYMBIOmatics Specific Support Action
STM	scientific, technical and medical
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
STS	Sequence Tagged Sites

SYBARIS	Systems biology analysis of fungal pathogen interactions with the human immune system
SYMBIOmatics	Synergies in Medical Informatics and Bioinformatics
TAIR	The Arabidopsis Information Resource
TrEMBL	Translations of EMBL
TUBS	Technische Universität Braunschweig
UBER	Humboldt University Berlin
UCL	University College London
UCSC	University of California, Santa Cruz
UH	University of Helsinki
UKPMC	PMC of UK
UNIBAS	University of Basel
UniMES	UniProt Metagenomic and Environmental Sequences
UniParc	UniProt Archive
UniProt	Universal Protein Resource
UniProtKB	UniProt Knowledgebase
UniRef	UniProt Reference Clusters
UniSave	UniProtKB Sequence/Annotation Version Archive
UniSTS	unified, non-redundant view of sequence tagged sites
UNITOR	University of Rome Tor Vergata
UNIVDUN	University of Dundee
URLs	Uniform Resource Locators
USPTO	United States Patent and Trademark Office
UU	Uppsala University
WGS	Whole Genome Shutgun Sequences
WI	Weizmann Institute of Science
WTSI	Wellcome Trust Cancer Institute
XML	Extensible Markup Language
ZFIN	Zebrafish Information Network

Bibliography

[2010, 2002] 2010, E. (2002). A guide to structure prediction (version 2.1). Website. http://www.russelllab.org/gtsp/; visited on September 21st 2010.

[2010, 2006a] 2010, E. (2006a). Oca©, a browser-database for protein structure/function. Website. http://www.ebi.ac.uk/msd-srv/oca/oca-docs/oca-home.html; visited on September 22nd 2010.

[2010, 2006b] 2010, E. (2006b). Oca sources. Website. http://www.ebi.ac.uk/msd-srv/oca/oca-docs/sources.html; visited on September 22nd 2010.

[2010, 2009] 2010, E. (2009). Procognate. Website. http://www.ebi.ac.uk/thornton-srv/databases/procognate/; visited on September 20th 2010.

[2010, 2010a] 2010, E. (2010a). Website. http://www.ebi.ac.uk/Rebholz-srv/ebimed/index.jsp; visited on September 9th 2010.

[2010, 2010b] 2010, E. (2010b). 2can support portal - protein function. Website. http://www.ebi.ac.uk/2can/tutorials/function/FingerPRINTScan.html; visited on September 10th 2010.

[2010, 2010c] 2010, E. (2010c). Fingerprintscan. Website. http://www.ebi.ac.uk/Tools/printsscan/; visited on September 10th 2010.

[2010, 2010d] 2010, E. (2010d). Help - fingerprintscan. Website. http://www.ebi.ac.uk/Tools/printsscan/help.html; visited on September 10th 2010.

[2010, 2010e] 2010, E. (2010e). Help - interproscan help. Website. http://www.ebi.ac.uk/Tools/InterProScan/help.html; visited on September 10th 2010.

[2010, 2010f] 2010, E. (2010f). Help - phobius. Website. http://www.ebi.ac.uk/Tools/phobius/help.html; visited on September 10th 2010.

[2010, 2010g] 2010, E. (2010g). Help - ppsearch. Website. http://www.ebi.ac.uk/Tools/ppsearch/help.html; visited on September 10th 2010.

[2010, 2010h] 2010, E. (2010h). Help - pratt - pattern matching. Website. http://www.ebi.ac.uk/Tools/pratt/help.html; visited on September 10th 2010.

[2010, 2010i] 2010, E. (2010i). Microarray analysis. Website. http://www.ebi.ac.uk/Tools/microarrayanalysis.html; visited on September 10th 2010.

[2010, 2010j] 2010, E. (2010j). Phobius. Website. http://www.ebi.ac.uk/Tools/phobius/; visited on September 10th 2010.

[2010, 2010k] 2010, E. (2010k). Picr - protein identifier cross-reference service. Website. http://www.ebi.ac.uk/Tools/picr/; visited on September 8th 2010.

[2010, 2010l] 2010, E. (2010l). Radar. Website. http://www.ebi.ac.uk/Tools/Radar/; visited on September 11th 2010.

[2010, 2010m] 2010, E. (2010m). Search clustr. Website. http://www.ebi.ac.uk/clustr-srv/CSearch; visited on September 10th 2010.

[2010, 2010n] 2010, E. (2010n). What is it? Website. http://www.ebi.ac.uk/webservices/whatizit/info.jsf; visited on September 9th 2010.

[2010, 10 aa] 2010, E. (2010 aa). Fasta - asd nucleotide similarity search. Website. http://www.ebi.ac.uk/Tools/fasta/asd.html; visited on September 18th 2010.

[2010, 10 ab] 2010, E. (2010 ab). Fasta - genomes similarity search. Website. http://www.ebi.ac.uk/Tools/fasta33/genomes.html; visited on September 19th 2010.

[2010, 10 ac] 2010, E. (2010 ac). Fasta - lgic nucleotide similarity search. Website. http://www.ebi.ac.uk/Tools/fasta/lgicn.html; visited on September 19th 2010.

[2010, 10 ad] 2010, E. (2010 ad). Fasta - nucleotide similarity search. Website. http://www.ebi.ac.uk/Tools/fasta33/nucleotide.html; visited on September 18th 2010.

[2010, 10 ae] 2010, E. (2010 ae). Fasta @ ebi. Website. http://www.ebi.ac.uk/Tools/fasta/; visited on September 18th 2010.

[2010, 10 af] 2010, E. (2010 af). Fasta/ssearch/ggsearch/glsearch - protein similarity search. Website. http://www.ebi.ac.uk/Tools/fasta33/; visited on September 18th 2010.

[2010, 10 ag] 2010, E. (2010 ag). Help - emboss-transeq. Website. http://www.ebi.ac.uk/Tools/emboss/transeq/help.html; visited on September 12th 2010.

[2010, 10 ah] 2010, E. (2010 ah). Help - mafft. Website. http://www.ebi.ac.uk/Tools/mafft/help.html; visited on September 12th 2010.

[2010, 10 ai] 2010, E. (2010 ai). Help - muscle. Website. http://www.ebi.ac.uk/Tools/muscle/help.html; visited on September 12th 2010.

[2010, 10 aj] 2010, E. (2010 aj). Help - psi-blast. Website. http://www.ebi.ac.uk/Tools/blastpgp/help.html; visited on September 13th 2010.

[2010, 10 ak] 2010, E. (2010 ak). Help - radar. Website. http://www.ebi.ac.uk/Tools/Radar/help.html; visited on September 11th 2010.

[2010, 10 al] 2010, E. (2010 al). Help - saps. Website. http://www.ebi.ac.uk/Tools/saps/help.html; visited on September 12th 2010.

[2010, 10 am] 2010, E. (2010 am). Integr8: Inquisitor. Website. http://www.ebi.ac.uk/integr8/InquisitorPage.do; visited on September 10th 2010.

[2010, 10 an] 2010, E. (2010 an). Interproscan sequence search. Website. http://www.ebi.ac.uk/Tools/InterProScan/; visited on September 10th 2010.

[2010, 10 ao] 2010, E. (2010 ao). Mafft. Website. http://www.ebi.ac.uk/Tools/mafft/; visited on September 12th 2010.

[2010, 10 ap] 2010, E. (2010 ap). Similarity and homology - ssearch. Website. http://www.ebi.ac.uk/Tools/ssearch/; visited on September 18th 2010.

[2010, 10 aq] 2010, E. (2010 aq). Wise2 - documentation (version 2.1.20 stable). Website. http://www.ebi.ac.uk/Tools/Wise2/doc_wise2.html; visited on September 12th 2010.

[2010, 10 ar] 2010, E. (2010 ar). Wise2- promoterwise. Website. http://www.ebi.ac.uk/Tools/Wise2/promoterwise.html; visited on September 12th 2010.

[2010, 10 ba] 2010, E. (2010 ba). Coiled-coil predictions. Website. http://www.russell.embl.de/cgi-bin/coils-svr.pl; visited on September 19th 2010.

[2010, 10 bb] 2010, E. (2010 bb). Emboss pepinfo/pepwindow/pepstats. Website. http://www.ebi.ac.uk/Tools/emboss/pepinfo/; visited on September 11th 2010.

[2010, 10 bc] 2010, E. (2010 bc). Fasta - lgic protein similarity search. Website. http://www.ebi.ac.uk/Tools/fasta/lgicp.html; visited on September 19th 2010.

[2010, 10 bd] 2010, E. (2010 bd). Fasta - proteomes similarity search. Website. http://www.ebi.ac.uk/Tools/fasta33/proteomes.html; visited on September 19th 2010.

[2010, 10 be] 2010, E. (2010 be). Fasta - snp similarity search. Website. http://www.ebi.ac.uk/Tools/snpfasta3/; visited on September 19th 2010.

[2010, 10 bf] 2010, E. (2010 bf). Fasta - whole genome shotgun sequence similarity search. Website. http://www.ebi.ac.uk/Tools/fasta33/wgs.html; visited on September 19th 2010.

[2010, 10 bg] 2010, E. (2010 bg). Help - kalign. Website. http://www.ebi.ac.uk/Tools/kalign/help.html; visited on September 11th 2010.

[2010, 10 bh] 2010, E. (2010 bh). Psi-search. Website. http://www.ebi.ac.uk/Tools/psisearch/; visited on September 19th 2010.

[2010, 10 bi] 2010, E. (2010 bi). String - known and predicted protein-protein interactions. Website. http://string.embl.de/; visited on September 19th 2010.

[2010, 10 ca] 2010, E. (2010 ca). Chemical search. Website. http://www.ebi.ac.uk/pdbe-site/chemsearch/; visited on September 22nd 2010.

[2010, 10 cb] 2010, E. (2010 cb). Pdbemapping. Website. http://www.ebi.ac.uk/pdbe-as/pdbemapquick/; visited on September 22nd 2010.

[2010, 10 cc] 2010, E. (2010 cc). Pdbetemplate. Website. http://www.ebi.ac.uk/pdbe-as/pdbetemplate/; visited on September 22nd 2010.

[2010, 2010o] 2010, E. B. (2010o). Kalign. Website. http://www.ebi.ac.uk/Tools/kalign/; visited on September 11th 2010.

[2010, 10 bj] 2010, E. B. (2010 bj). Maxsprout. Website. http://www.ebi.ac.uk/Tools/webservices/services/maxsprout; visited on September 20th 2010.

[2010, 1995] 2010, E. I. (1995). Embl www gateway to isoelectric point service. Website. http://www3.embl.de/cgi/pi-wrapper.pl; visited on September 21st 2010.

[2010, 10 cd] 2010, E. I. (2010 cd). About tempura. Website. http://www.ebi.ac.uk/thornton-srv/databases/tempura/tempura_docs.html; visited on September 22nd 2010.

[2010, 10 ce] 2010, E. I. (2010 ce). Biobar - a toolbar for browsing biological data and databases. Website. http://www.ebi.ac.uk/pdbe/docs/biobar.html; visited on September 22nd 2010.

[2010, 10 cf] 2010, E. I. (2010 cf). Olderado. Website. http://www.ebi.ac.uk/pdbe/olderado/; visited on September 22nd 2010.

[2010, 10 cg] 2010, E. I. (2010 cg). Pdbeanalysis: Pdbe data analysis service list. Website. http://www.ebi.ac.uk/pdbe-as/pdbevalidate/; visited on September 22nd 2010.

[2010, 10 ch] 2010, E. I. (2010 ch). Protein colourer. Website. http://www.ebi.ac.uk/cgi-bin/proteincol/ProteinColourer.pl; visited on September 21st 2010.

[2010, 10 ci] 2010, E. I. (2010 ci). Protein data bank in europe : Online services. Website. http://www.ebi.ac.uk/pdbe/docs/Services.html; visited on September 22nd 2010.

[2010, 10 cj] 2010, E. I. (2010 cj). Readseq - biosequence conversion tool. Website. http://www.ebi.ac.uk/cgi-bin/readseq.cgi; visited on September 21st 2010.

[2010, 10 ck] 2010, E. I. (2010 ck). Representative nmr ensemble pdb search. Website. http://www.ebi.ac.uk/pdbe/pqs/pqs-nmr.html; visited on September 22nd 2010.

[Agadir, 2010] Agadir (2010). Agadir - an algorithm to predict the helical content of peptides. Website. http://agadir.crg.es/; visited on September 11th 2010.

[APF, 2009] APF (2009). Bing gains search market share, nears 10 percent. Website. http://www.google.com/hostednews/afp/article/ALeqM5iFIa_CCCc6QwmkYkicyfMLE_oGiQ; visited on October 2nd 2010.

[BASE, 2010a] BASE (2010a). Base: Hilfe zur suche. Website. http://base.ub.uni-bielefeld.de/de/help_search.php?menu=3; visited on September 26th 2010; original language: German.

[BASE, 2010b] BASE (2010b). Über base. Website. http://base.ub.uni-bielefeld.de/de/index.php; visited on September 26th 2010; original language: German.

[Bates et al., 2009] Bates, K. T., Manning, P. L., Hodgetts, D., and Sellers, W. I. (2009). Estimating mass properties of dinosaurs using laser imaging and 3d computer modelling. *PLoS ONE*.

[BioCatalogue, 2010a] BioCatalogue (2010a). Biocatalogue wiki. Website. http://www.biocatalogue.org/wiki/; visited on August 25th 2010.

[BioCatalogue, 2010b] BioCatalogue (2010b). Biocatalogue wiki - faq. Website. http://www.biocatalogue.org/wiki/doku.php?id=public:faq; visited on August 25th 2010.

[Bioconductor, 2010] Bioconductor (2010). Bioconductor - open source software for bioinformatics. Website. http://www.bioconductor.org/; visited on September 10th 2010.

[biosapiens, 2010] biosapiens (2010). Biosapiens network of excellence - a european virtual institute for genome annotation. Website. http://www.biosapiens.info/page.php; visited on August 25th 2010.

[Bookshelf, 2010] Bookshelf, N. (2010). Bookshelf. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=books; visited on April 8th 2010.

[CAL, 2010] CAL (2010). Canadian agriculture library. Website. http://www4.agr.gc.ca/AAFC-AAC/display-afficher.do?id=1176485402230&lang=eng; visited on July 7th 2010.

[Canada, 2010a] Canada, P. (2010a). A free archive of life science journals. Website. http://pubmedcentralcanada.ca/; visited on April 7th 2010.

[Canada, 2010b] Canada, P. (2010b). Pubmed central canada. Website. http://pubmedcentralcanada.ca/; visited on June 25th 2010.

[CDD, 2010a] CDD, N. (2010a). Conserved domains and protein classification. Website. http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml; visited on June 18th 2010.

[CDD, 2010b] CDD, N. (2010b). Conserved domains database. Website. http://www.ncbi.nlm.nih.gov/sites/cdd; visited on June 18th 2010.

[CIHR, 7 17] CIHR (2007-07-17). Federal science elibrary. Website. http://safstl-asbstf.scitech.gc.ca/eng/index.html; visited on July 12th 2010.

[CIHR, 2009] CIHR (2009). Canadian virtual health library. Website. http://www.chla-absc.ca/nnlh/cvhl.htm; visited on September 5th 2010.

[CIHR, 2010a] CIHR (2010a). Federal science elibrary: Access to the world's scientific information through an interdepartmental initiative. Website. http://safstl-asbstf.scitech.gc.ca/eng/initiative.html; visited on July 12th 2010.

[CIHR, 2010b] CIHR (2010b). Worldwidescience.org - the global science gateway. Website. http://worldwidescience.org/; visited on July 13th 2010.

[CIHR, 4 29] CIHR (2010-04-29). Cihr policy on access to research outputs. Website. http://www.cihr-irsc.gc.ca/e/32005.html; visited on July 5th 2010.

[cmpharm, 1998] cmpharm (1998). Webmol - java pdb viewer. Website. http://www.cmpharm.ucsf.edu/~walther/webmol.html; visited on September 22nd 2010.

[CODATA, 2010] CODATA (2010). Welcome to the canadian national committee for codata. Website. http://www.codata.org/canada/about.shtml; visited on September 5th 2010.

[Coil, 2010] Coil, C. (2010). Predicting coiled coils from protein sequences. Website. http://www.ncbi.nlm.nih.gov/pubmed/2031185?dopt=AbstractPlus; visited on September 19th 2010.

[Consortium, 2010a] Consortium, U. (2010a). About uniprot. Website. http://www.uniprot.org/help/about; visited on August 22th 2010.

[Consortium, 2010b] Consortium, U. (2010b). Taxonomy. Website. http://www.uniprot.org/taxonomy/; visited on July 28th 2010.

[Consortium, 2010c] Consortium, U. (2010c). Welcome. Website. http://www.uniprot.org/; visited on July 28th 2010.

[data donnees.gc.ca, 2008] data donnees.gc.ca (2008). Research data strategy working group: Opening new pathways to canadian research data. Website. http://data-donnees.gc.ca/eng/about/backgrounder.html; visited on September 5th 2010.

[DataCite, 2010a] DataCite (2010a). Members. Website. http://www.tib-hannover.de/fileadmin/datacite/members.html; visited on September 5th 2010.

[DataCite, 2010b] DataCite (2010b). What is datacite? Website. http://www.tib-hannover.de/fileadmin/datacite/whatisdc.html; visited on September 5th 2010.

[Dean and Johanna McEntyre, 2010] Dean, L. and Johanna McEntyre, N. (2010). Coffee break - tutorials for ncbi tools. Website. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=coffeebrk; visited on June 25th 2010.

[DisEMBL, 2010] DisEMBL (2010). Abstract. Website. http://dis.embl.de/html/abstract.html; visited on September 19th 2010.

[Dulcinea Media, 2009] Dulcinea Media, I. (2009). Dulcinea media. Website. http://www.dulcineamedia.com/; visited on September 26th 2010.

[Dulcinea Media, 2010a] Dulcinea Media, I. (2010a). About sweetsearch. Website. http://www.sweetsearch.com/info/about; visited on September 26th 2010.

[Dulcinea Media, 2010b] Dulcinea Media, I. (2010b). Sweetsearch - a search engine for students. Website. http://www.sweetsearch.com/; visited on September 26th 2010.

[E-Mep, 2009a] E-Mep (2009a). About e-mep. Website. http://www.e-mep.org/about/index.htm; visited on September 23rd 2010.

[E-Mep, 2009b] E-Mep (2009b). Partners. Website. http://www.e-mep.org/about/partners.htm; visited on September 23rd 2010.

[EBI, 2001] EBI (2001). Parasite genome databases and genome research resources. Website. http://www.ebi.ac.uk/parasites/parasite-genome.html; visited on July 27th 2010.

[EBI, 2008] EBI (2008). Pandit: Protein and associated nucleotide domains with inferred trees. Website. http://www.ebi.ac.uk/goldman-srv/pandit/; visited on July 27th 2010.

[EBI, 2010a] EBI (2010a). Bringing structure to biology. Website. http://www.ebi.ac.uk/pdbe/#m=5&h=0&e=0&r=0&l=0&a=0&w=0; visited on July 27th 2010.

[EBI, 2010b] EBI (2010b). Genome reviews. Website. http://www.ebi.ac.uk/GenomeReviews/; visited on July 26th 2010.

[EBI, 2010c] EBI (2010c). Go @ ebi. Website. http://www.ebi.ac.uk/GO/; visited on July 26th 2010.

[EBI, 2010d] EBI (2010d). Hssp. Website. http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+LibInfo+-lib+HSSP; visited on July 27th 2010.

[EBI, 2010e] EBI (2010e). Imgt/hla database. Website. http://www.ebi.ac.uk/imgt/hla/; visited on July 27th 2010.

[EBI, 2010f] EBI (2010f). Intact home. Website. http://www.ebi.ac.uk/intact/main.xhtml; visited on July 27th 2010.

[EBI, 2010g] EBI (2010g). Integr8 : Access to complete genomes and proteomes. Website. http://www.ebi.ac.uk/integr8/EBI-Integr8-HomePage.do; visited on July 27th 2010.

[EBI, 2010h] EBI (2010h). Intenz - home. Website. http://www.ebi.ac.uk/intenz/; visited on July 27th 2010.

[EBI, 2010i] EBI (2010i). Ipd - the immuno polymorphism database. Website. http://www.ebi.ac.uk/ipd/; visited on July 27th 2010.

[EBI, 2010j] EBI (2010j). Ipi - international protein index. Website. http://www.ebi.ac.uk/IPI/IPIhelp.html; visited on July 27th 2010.

[EBI, 2010k] EBI (2010k). Ipi - international protein index - source databases for ipi. Website. http://www.ebi.ac.uk/IPI/Databases.html; visited on July 27th 2010.

[EBI, 2010l] EBI (2010l). Ligand-gated ion channel database. Website. http://www.ebi.ac.uk/compneur-srv/LGICdb/LGICdb.php; visited on July 27th 2010.

[EBI, 2010m] EBI (2010m). Ols - ontology lookup service. Website. http://www.ebi.ac.uk/ontology-lookup/; visited on July 27th 2010.

[EBI, 2010n] EBI (2010n). Patent data resources at the ebi. Website. http://www.ebi.ac.uk/patentdata/; visited on July 27th 2010.

[EBI, 2010o] EBI (2010o). Pdbsum. Website. http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/; visited on July 27th 2010.

[EBI, 2010p] EBI (2010p). Profunc - analysis of a protein's 3d structure to help identify its likely biochemical function. Website. http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/; visited on July 27th 2010.

[EBI, 2010q] EBI (2010q). Profunc - prediction of protein function from 3d structure. Website. http://www.ebi.ac.uk/thornton-srv/databases/profunc/index.html; visited on July 27th 2010.

[EBI, 2010r] EBI (2010r). Proteomics identifications database (pride). Website. http://www.ebi.ac.uk/pride/; visited on July 27th 2010.

[EBI, 2010s] EBI (2010s). Quickgo. Website. http://www.ebi.ac.uk/QuickGO/; visited on July 27th 2010.

[EBI, 2010t] EBI (2010t). Quickgo help. Website. http://www.ebi.ac.uk/QuickGO/help.html; visited on July 27th 2010.

[EBI, 2010u] EBI (2010u). Resid database at the ebi. Website. http://www.ebi.ac.uk/RESID/; visited on July 28th 2010.

[EBI, 2010v] EBI (2010v). Rhea - home. Website. http://www.ebi.ac.uk/rhea/; visited on July 28th 2010.

[EBI, 2010w] EBI (2010w). search patent abstracts. Website. http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+query+-libList+PATABS; visited on July 27th 2010.

[EBI, 2010x] EBI (2010x). Systems biology ontology. Website. http://www.ebi.ac.uk/sbo/; visited on July 28th 2010.

[EBI, 2010y] EBI (2010y). Uniprot - welcome to uniprot. Website. http://www.ebi.ac.uk/uniprot/; visited on July 28th 2010.

[EBI, 2010z] EBI (2010z). The uniprot protein das server. Website. http://www.ebi.ac.uk/uniprot-das/; visited on July 28th 2010.

[EBInstitute, 2008] EBInstitute (2008). Pdbe nmr home page. Website. http://www.ebi.ac.uk/pdbe/docs/NMR/main.html; visited on September 3rd 2010.

[EBInstitute, 2009a] EBInstitute (2009a). Ligands in pdb. Website. http://www.ebi.ac.uk/msd-srv/msdchem/cgi-bin/cgi.pl; visited on September 3rd 2010.

[EBInstitute, 2009b] EBInstitute (2009b). Motifs and sites. Website. http://www.ebi.ac.uk/pdbe-site/pdbemotif/; visited on September 3rd 2010.

[EBInstitute, 2010a] EBInstitute (2010a). Advanced search. Website. http://www.ebi.ac.uk/pdbe-srv/view/; visited on September 3rd 2010.

[EBInstitute, 2010b] EBInstitute (2010b). The electron microscopy data bank (emdb) at ebi. Website. http://www.ebi.ac.uk/pdbe/emdb/; visited on September 3rd 2010.

[EBInstitute, 2010c] EBInstitute (2010c). Fssp. Website. http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+LibInfo+-lib+FSSP; visited on July 26th 2010.

[EBInstitute, 2010d] EBInstitute (2010d). Interfaces and assemblies. Website. http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html; visited on September 3rd 2010.

[EBInstitute, 2010e] EBInstitute (2010e). Karyn's genomes. Website. http://www.ebi.ac.uk/2can/genomes/genomes.html; visited on September 3rd 2010.

[EBInstitute, 2010f] EBInstitute (2010f). Protein functional analysis. Website. http://www.ebi.ac.uk/Tools/protein.html; visited on September 8th 2010.

[EBInstitute, 2010g] EBInstitute (2010g). Similarity & homology. Website. http://www.ebi.ac.uk/Tools/similarity.html; visited on September 8th 2010.

[EBInstitute, 2010h] EBInstitute (2010h). Structural analysis. Website. http://www.ebi.ac.uk/Tools/structural.html; visited on September 8th 2010.

[EBInstitute, 2010i] EBInstitute (2010i). Structure similarity. Website. http://www.ebi.ac.uk/msd-srv/ssm/; visited on September 3rd 2010.

[EBInstitute, 2010j] EBInstitute (2010j). Toolbox at the ebi. Website. http://www.ebi.ac.uk/Tools/; visited on September 8th 2010.

[eccellio, 2010] eccellio (2010). eccellio excellent faceting search - search and simplicity with one additional single-click. Website. http://www.eccellio.com/about/discover.php; visited on September 24th 2010.

[ELIXIR, 2009a] ELIXIR (2009a). What is elixir? Website. http://www.elixir-europe.org/page.php; visited on August 26th 2010.

[ELIXIR, 2009b] ELIXIR (2009b). What is elixir? Website. http://www.elixir-europe.org/page.php?page=call; visited on August 26th 2010.

[ELM, 2010] ELM (2010). Elm - the eukaryotic linear motif resource for functional sites in proteins. Website. http://elm.eu.org/; visited on September 21st 2010.

[EMBL, 2010a] EMBL (2010a). About arrayexpress. Website. http://www.ebi.ac.uk/microarray/doc/; visited on July 24th 2010.

[EMBL, 2010b] EMBL (2010b). About the embl-ebi. Website. http://www.ebi.ac.uk/Information/About_EBI/about_ebi.html; visited on July 24th 2010.

[EMBL, 2010c] EMBL (2010c). About the european nucleotide archive. Website. http://www.ebi.ac.uk/ena/about/page.php?page=about; visited on July 26th 2010.

[EMBL, 2010d] EMBL (2010d). About the gene expression atlas. Website. http://www.ebi.ac.uk/gxa/help/AboutAtlas; visited on July 26th 2010.

[EMBL, 2010e] EMBL (2010e). Arrayexpress. Website. http://www.ebi.ac.uk/microarray-as/ae/; visited on July 24th 2010.

[EMBL, 2010f] EMBL (2010f). Arrayexpress faq. Website. http://www.ebi.ac.uk/microarray/doc/help/faq.html; visited on July 24th 2010.

[EMBL, 2010g] EMBL (2010g). Catalytic site atlas. Website. http://www.ebi.ac.uk/thornton-srv/databases/CSA/; visited on July 26th 2010.

[EMBL, 2010h] EMBL (2010h). Chebi hauptseite. Website. http://www.ebi.ac.uk/chebi/; visited on July 26th 2010.

[EMBL, 2010i] EMBL (2010i). Chembl database. Website. http://www.ebi.ac.uk/chembldb/; visited on July 26th 2010.

[EMBL, 2010j] EMBL (2010j). Citexplore literature searching. Website. http://www.ebi.ac.uk/citexplore/; visited on July 26th 2010.

[EMBL, 2010k] EMBL (2010k). Clustr. Website. http://www.ebi.ac.uk/clustr/; visited on July 26th 2010.

[EMBL, 2010l] EMBL (2010l). Dssp. Website. http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+LibInfo+-lib+DSSP; visited on July 26th 2010.

[EMBL, 2010m] EMBL (2010m). Ebi databases index - a to z. Website. http://www.ebi.ac.uk/Information/databases_sitemap.html; visited on July 24th 2010.

[EMBL, 2010n] EMBL (2010n). Ebi dbfetch. Website. http://www.ebi.ac.uk/cgi-bin/dbfetch; visited on July 26th 2010.

[EMBL, 2010o] EMBL (2010o). Ebi tools index - a to z. Website. http://www.ebi.ac.uk/Information/tools_sitemap.html; visited on July 24th 2010.

[EMBL, 2010p] EMBL (2010p). e!ensembl. Website. http://www.ensembl.org/index.html; visited on July 26th 2010.

[EMBL, 2010q] EMBL (2010q). e!ensemblgenomes. Website. http://www.ensemblgenomes.org/; visited on July 26th 2010.

[EMBL, 2010r] EMBL (2010r). Embl - european molecular biology laboratory. Website. http://www.embl.org/; visited on July 24th 2010.

[EMBL, 2010s] EMBL (2010s). Embl-ebi. Website. http://www.ebi.ac.uk/ebisearch/; visited on July 26th 2010.

[EMBL, 2010t] EMBL (2010t). The european genome-phenome archive. Website. http://www.ebi.ac.uk/ega/page.php; visited on July 26th 2010.

[EMBL, 2010u] EMBL (2010u). European nucleotide archive. Website. http://www.ebi.ac.uk/ena/; visited on July 26th 2010.

[EMBL, 2010v] EMBL (2010v). Experimental factor ontology. Website. http://www.ebi.ac.uk/efo/; visited on July 26th 2010.

[EMBL, 2010w] EMBL (2010w). Find a species. Website. http://www.ensembl.org/info/about/species.html; visited on July 26th 2010.

[EMBL, 2010x] EMBL (2010x). Genomes pages - at the ebi. Website. http://www.ebi.ac.uk/genomes/; visited on July 26th 2010.

[EMBL, 2010y] EMBL (2010y). http://www.ebi.ac.uk/biomodels-main/. Website. BioModelsDatabase-ADatabaseofAnnotatedPublishedModels; visited on July 26th 2010.

[EMBL, 2010z] EMBL (2010z). Sequence version archive. Website. http://www.ebi.ac.uk/cgi-bin/sva/sva.pl/; visited on July 26th 2010.

[EMBL and OICR, 2010] EMBL and OICR (2010). Biomart project. Website. http://www.biomart.org/index.html; visited on July 28th 2010.

[EMBL-EBI, 2010a] EMBL-EBI (2010a). About the hgnc. Website. http://www.genenames.org/aboutHGNC.html; visited on July 27th 2010.

[EMBL-EBI, 2010b] EMBL-EBI (2010b). Ebi-hosted eu project websites. Website. http://www.ebi.ac.uk/euprojects/; visited on August 25th 2010.

[EMBL-EBI, 2010c] EMBL-EBI (2010c). Ebi-hosted eu project websites. Website. http://www.ebi.ac.uk/euprojects/index.html; visited on August 26th 2010.

[emboss, 2010a] emboss (2010a). ebmoss water. Website. http://emboss.sourceforge.net/apps/cvs/emboss/apps/water.html; visited on August 23th 2010.

[emboss, 2010b] emboss (2010b). emboss needle. Website. http://emboss.sourceforge.net/apps/cvs/emboss/apps/needle.html; visited on August 23th 2010.

[EMBRACE, 2009] EMBRACE (2009). Introduction to information for bioinformaticians.

[EMBRACE, 2010] EMBRACE (2010). Embrace network of excellence - a european model for bioinformatics research and community education. Website. http://www.embracegrid.info/page.php; visited on September 23rd 2010.

[EMeP, 2009] EMeP (2009). European membrane protein consortium. Website. http://www.e-mep.org/; visited on August 25th 2010.

[ENFIN, 2010a] ENFIN (2010a). Enfin - enabling systems biology. Website. http://www.enfin.org/page.php; visited on August 26th 2010.

[ENFIN, 2010b] ENFIN (2010b). Enfin partners. Website. http://www.enfin.org/page.php?page=partners; visited on August 26th 2010.

[Engelbart, 1962] Engelbart, D. C. (1962). Augmenting human intellect: A conceptual framework. Website. http://www.dougengelbart.org/pubs/augment-3906.html; visited on October 3rd 2010.

[Entrez, 2004a] Entrez, N. (2004a). Entrez map viewer help document. Website. http://www.ncbi.nlm.nih.gov/projects/mapview/static/MapViewerHelp.html; visited on June 21st 2010.

[Entrez, 2004b] Entrez, N. (2004b). Ncbi - model of entrez databases. Website. http://www.ncbi.nlm.nih.gov/Database/datamodel/index.html; visited on April 11th 2010.

[Entrez, 2005a] Entrez, N. (2005a). Entrez gene: A directory of genes. Website. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=handbook&part=ch19; visited on June 2nd 2010.

[Entrez, 2005b] Entrez, N. (2005b). Genome project: A collection of genome specific information. Website. http://www.ncbi.nlm.nih.gov/genomes/static/gprj_help.html#introduction; visited on June 21st 2010.

[Entrez, 2009] Entrez, N. (2009). Entrez programming utilities. Website. http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html; visited on June 22nd 2010.

[Entrez, 2010a] Entrez, N. (2010a). Batch entrez. Website. http://www.ncbi.nlm.nih.gov/sites/batchentrez?db=Nucleotide; visited on June 22nd 2010.

[Entrez, 2010b] Entrez, N. (2010b). Entrez gene. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene; visited on June 2nd 2010.

[Entrez, 2010c] Entrez, N. (2010c). Entrez gene help: Integrated access to genes of genomes in the reference sequence collection. Website. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helpgene&part=EntrezGene; visited on June 2nd 2010.

[Entrez, 2010d] Entrez, N. (2010d). Entrez genome. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome; visited on June 21st 2010.

[Entrez, 2010e] Entrez, N. (2010e). Entrez genome project - connection, information, discovery. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=search&db=genomeprj; visited on June 21st 2010.

[Entrez, 2010f] Entrez, N. (2010f). Entrez popset. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=popset; visited on June 14th 2010.

[Entrez, 2010g] Entrez, N. (2010g). Entrez, the life science search engine. Website. http://www.ncbi.nlm.nih.gov/sites/gquery; visited on June 22nd 2010.

[Entrez, 2010h] Entrez, N. (2010h). Extinct organisms that are represented with sequence data at genbank. Website. http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=extinct; visited on June 21st 2010.

[Entrez, 2010i] Entrez, N. (2010i). Geo datasets. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=gds; visited on June 2nd 2010.

[Entrez, 2010j] Entrez, N. (2010j). Geo frequently asked questions. Website. http://www.ncbi.nlm.nih.gov/geo/info/faq.html; visited on June 2nd 2010.

[Entrez, 2010k] Entrez, N. (2010k). Geo profiles. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=geo; visited on June 2nd 2010.

[Entrez, 2010l] Entrez, N. (2010l). The ncbi entrez taxonomy homepage. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy; visited on June 20th 2010.

[Entrez, 2010m] Entrez, N. (2010m). The ncbi taxonomy homepage. Website. http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi; visited on June 21st 2010.

[Entrez, 2010n] Entrez, N. (2010n). Nucleotide - alphabet of life. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide; visited on June 14th 2010.

[Entrez, 2010o] Entrez, N. (2010o). Omia. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=omia; visited on May 1st 2010.

[Entrez, 2010p] Entrez, N. (2010p). Protein clusters. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=proteinclusters; visited on June 20th 2010.

[Entrez, 2010q] Entrez, N. (2010q). Structure. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=domains; visited on June 20th 2010.

[Entrez, 2010r] Entrez, N. (2010r). Structure - molecular modeling database (mmdb). Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=structure; visited on June 20th 2010.

[Entrez, 2010s] Entrez, N. (2010s). Taxonomy browser. Website. http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root; visited on June 21st 2010.

[Entrez, 2010t] Entrez, N. (2010t). Taxonomy resources. Website. http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=resources; visited on June 21st 2010.

[Entrez, 2010u] Entrez, N. (2010u). Unists - integrating markers and maps. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=unists; visited on June 16th 2010.

[FoldX, 2010] FoldX (2010). Foldx - a force field for energy calculations and protein design. Website. http://foldx.crg.es/; visited on September 20th 2010.

[Genomes, 2010a] Genomes, . (2010a). 1000 genomes - a deep catalog of human genetic variation. Website. http://www.1000genomes.org/page.php; visited on August 25th 2010.

[Genomes, 2010b] Genomes, . (2010b). About the 1000 genomes project. Website. http://www.1000genomes.org/page.php?page=about; visited on August 25th 2010.

[GEO, 2010a] GEO, N. (2010a). Geo. Website. http://www.ncbi.nlm.nih.gov/geo/; visited on June 8th 2010.

[GEO, 2010b] GEO, N. (2010b). Geo accession display tool. Website. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi; visited on June 8th 2010.

[GEO, 2010c] GEO, N. (2010c). Geo dataset cluster analysis. Website. http://www.ncbi.nlm.nih.gov/geo/info/cluster.html; visited on June 8th 2010.

[GEO, 2010d] GEO, N. (2010d). Geo faq. Website. http://www.ncbi.nlm.nih.gov/geo/info/faq.html; visited on June 8th 2010.

[GFP-cDNA, 2005] GFP-cDNA, E. (2005). Information. Website. http://gfp-cdna.embl.de/html/information.html; visited on September 10th 2010.

[Globplot, 2006] Globplot (2006). Abstract. Website. http://globplot.embl.de/html/abstract.html; visited on September 20th 2010.

[Google, 2010] Google (2010). Google milestones. Website. http://www.google.com/intl/en/corporate/history.html; visited on March 31st 2010.

[Guidicelli et al., 2006] Guidicelli, V., Duroux, P., Ginestoux, C., Folch, G., Jabado-Michaloud, J., Chaume, D., and Lefranc, M.-P. (2006). Imgt/ligm-db, the imgt comprehensive database of immunoglobulin and t cell receptor nucleotide sequences. *Nucleic Acids Research (Oxford Journals)*, 34.

[Hall et al., 2008] Hall, W., Roure, D. D., and Shadbolt, N. (2008). The evolution of the web and implications for eresearch. *The Royal Society*.

[Harvester, 2010] Harvester (2010). http://harvester.fzk.de/harvester/. Website. http://harvester.fzk.de/harvester/; visited on September 21st 2010.

[HomoloGene, 2010a] HomoloGene, N. (2010a). Homologene - discover homologs. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene; visited on June 2nd 2010.

[HomoloGene, 2010b] HomoloGene, N. (2010b). Homologene build procedure. Website. http://www.ncbi.nlm.nih.gov/HomoloGene/HTML/homologene_buildproc.html; visited on June 2nd 2010.

[IEB, 2001] IEB, N. (2001). Ncbi toolbox. Website. http://www.ncbi.nlm.nih.gov/IEB/ToolBox/index.cgi; visited on June 22nd 2010.

[IEB, 2002] IEB, N. (2002). Information engineering branch. Website. http://www.ncbi.nlm.nih.gov/IEB/; visited on June 22nd 2010.

[IEB, 2003] IEB, N. (2003). Xml at ncbi. Website. http://www.ncbi.nlm.nih.gov/IEB/ToolBox/XML/; visited on June 22nd 2010.

[IEB, 2010] IEB, N. (2010). Ncbi data in xml. Website. http://www.ncbi.nlm.nih.gov/IEB/ToolBox/XML/ncbixml.txt; visited on June 22nd 2010.

[Ilene Mizrachi, 2007] Ilene Mizrachi, N. (2007). Genbank: The nucleotide sequence database. Website. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=handbook&part=ch1; visited on May 29th 2010.

[IMGT, 2010] IMGT (2010). Welcome! to imgt/ligm-db. Website. http://imgt.cines.fr/cgi-bin/IMGTlect.jv?livret=0; visited on September 3rd 2010.

[IMPACT, 2008] IMPACT (2008). About impact. Website. http://www.ebi.ac.uk/impact/page.php?page=about; visited on August 26th 2010.

[INSDC, 2010] INSDC (2010). International nucleotide sequence database collaboration. Website. http://www.insdc.org/; visited on August 26th 2010.

[Institute, 2004] Institute, E. B. (2004). Using the andom-server). Website. http://coot.embl.de/AnDom/Usage.html; visited on August 31st 2010.

[Institute, 2010a] Institute, E. B. (2010a). Biological ontology databases. Website. http://www.ebi.ac.uk/Databases/ontology.html; visited on August 31st 2010.

[Institute, 2010b] Institute, E. B. (2010b). Blast @ ebi. Website. http://www.ebi.ac.uk/Tools/blast/; visited on August 25th 2010.

[Institute, 2010c] Institute, E. B. (2010c). Censor. Website. http://www.ebi.ac.uk/Tools/censor/; visited on August 30th 2010.

[Institute, 2010d] Institute, E. B. (2010d). Clustalw2. Website. http://www.ebi.ac.uk/Tools/clustalw2/; visited on August 30th 2010.

[Institute, 2010e] Institute, E. B. (2010e). Clustalw2. Website. http://www.ebi.ac.uk/Tools/clustalw2/faq.html; visited on August 30th 2010.

[Institute, 2010f] Institute, E. B. (2010f). Dalilite pairwise comparison of protein structures. Website. http://www.ebi.ac.uk/Tools/dalilite/; visited on August 30th 2010.

[Institute, 2010g] Institute, E. B. (2010g). Database browsing and entry retrieval. Website. http://www.ebi.ac.uk/Databases/service.html; visited on August 31st 2010.

[Institute, 2010h] Institute, E. B. (2010h). Database on demand. Website. http://www.ebi.ac.uk/pride/dod/pages/dodStart.jsf?sessionControlDisabled=true; visited on August 31st 2010.

[Institute, 2010i] Institute, E. B. (2010i). Embl computational services overview. Website. http://www.ebi.ac.uk/embl_services/; visited on August 31st 2010.

[Institute, 2010j] Institute, E. B. (2010j). Emboss cpgplot/cpgreport/isochore. Website. http://www.ebi.ac.uk/Tools/emboss/cpgplot/; visited on August 30th 2010.

[Institute, 2010k] Institute, E. B. (2010k). Emboss pairwise alignment algorithms. Website. http://www.ebi.ac.uk/Tools/emboss/align/; visited on August 23th 2010.

[Institute, 2010l] Institute, E. B. (2010l). Gene ontology annotation (uniprotkb-goa) database. Website. http://www.ebi.ac.uk/GOA/; visited on August 23th 2010.

[Institute, 2010m] Institute, E. B. (2010m). Genewise - dna block aligner. Website. http://www.ebi.ac.uk/Tools/Wise2/dbaform.html; visited on August 31st 2010.

[Institute, 2010n] Institute, E. B. (2010n). Help - censor. Website. http://www.ebi.ac.uk/Tools/censor/help.html; visited on August 30th 2010.

[Institute, 2010o] Institute, E. B. (2010o). Introduction. Website. http://www.ebi.ac.uk/dasty/index.php?cont=introduction; visited on August 30th 2010.

[Institute, 2010p] Institute, E. B. (2010p). Literature databases. Website. http://www.ebi.ac.uk/Databases/literature.html; visited on August 31st 2010.

[Institute, 2010q] Institute, E. B. (2010q). Nucleotide databases. Website. http://www.ebi.ac.uk/Databases/nucleotide.html; visited on August 31st 2010.

[Institute, 2010r] Institute, E. B. (2010r). Pathway and network databases. Website. http://www.ebi.ac.uk/Databases/pathways.html; visited on August 31st 2010.

[Institute, 2010s] Institute, E. B. (2010s). Protein databases. Website. http://www.ebi.ac.uk/Databases/protein.html; visited on August 31st 2010.

[Institute, 2010t] Institute, E. B. (2010t). Proteomic databases. Website. http://www.ebi.ac.uk/Databases/proteomic.html; visited on August 31st 2010.

[Institute, 2010u] Institute, E. B. (2010u). Small molecule databases. Website. http://www.ebi.ac.uk/Databases/smallmolecules.html; visited on August 31st 2010.

[Institute, 2010v] Institute, E. B. (2010v). Structure databases. Website. http://www.ebi.ac.uk/Databases/structure.html; visited on August 31st 2010.

[Institute, 2010w] Institute, E. B. (2010w). Uniparc at the ebi. Website. http://www.ebi.ac.uk/uniparc/; visited on July 28th 2010.

[Institute, 2010x] Institute, E. B. (2010x). Uniprot documentation. Website. http://www.ebi.ac.uk/uniprot/Documentation/index.html; visited on August 23th 2010.

[Institute, 2010y] Institute, E. B. (2010y). Uniref at the ebi. Website. http://www.ebi.ac.uk/uniref/; visited on August 22th 2010.

[Institute, 2010z] Institute, E. B. (2010z). Unisave. Website. http://www.ebi.ac.uk/uniprot/unisave/; visited on August 23th 2010.

[Institute, 10 a] Institute, E. B. (2010 a). Wsdbfetch (soap). Website. http://www.ebi.ac.uk/Tools/webservices/services/dbfetch; visited on September 2nd 2010.

[InterPreTS, 2010] InterPreTS (2010). Interprets - interaction prediction through tertiary structure. Website. http://www.russell.embl.de/cgi-bin/tools/interprets.pl; visited on September 20th 2010.

[jswamin, 2010] jswamin (2010). Biobar 2.0.1. Website. https://addons.mozilla.org/en-US/firefox/addon/169/; visited on September 22nd 2010.

[Katoh, 2009] Katoh, K. (2009). Mafft version 6 - multiple alignment program for amino acid or nucleotide sequences. Website. http://mafft.cbrc.jp/alignment/software/index.html; visited on September 12th 2010.

[Katoh, 2010] Katoh, K. (2010). Mafft version 6 - multiple alignment program for amino acid or nucleotide sequences. Website. http://mafft.cbrc.jp/alignment/software/windows.html; visited on September 12th 2010.

[LRG, 2009a] LRG (2009a). Locus reference genomic. Website. http://www.lrg-sequence.org/page.php; visited on August 26th 2010.

[LRG, 2009b] LRG (2009b). Lrg frequently asked questions (faqs). Website. http://www.lrg-sequence.org/page.php; visited on August 26th 2010.

[MEDLINE, 2008] MEDLINE (2008). Fact sheet medline. Website. http://www.nlm.nih.gov/pubs/factsheets/medline.html; visited on April 6th 2010.

[MeSH, 2009a] MeSH (2009a). Medical subject headings - files available to download. Website. http://www.nlm.nih.gov/mesh/filelist.html; visited on April 6th 2010.

[MeSH, 2009b] MeSH (2009b). Mesh browser. Website. http://www.nlm.nih.gov/mesh/mbinfo.html; visited on April 6th 2010.

[MeSH, 2010] MeSH (2010). Mesh. Website. http://www.ncbi.nlm.nih.gov/sites/entrez/query.fcgi?db=mesh; visited on April 6th 2010.

[microme, 2010] microme (2010). Organization overview. Website. http://www.microme.eu/organization; visited on August 29th 2010.

[Monica Romiti, 2010] Monica Romiti, N. E. (2010). Entrez help. Website. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helpentrez&part=EntrezHelp#EntrezHelp.Using_Limits; visited on May 25th 2010.

[NCBI, 2003] NCBI (2003). Genome survey sequences database. Website. http://www.ncbi.nlm.nih.gov/dbGSS/; visited on June 8th 2010.

[NCBI, 2004] NCBI (2004). Ncbi at a glance - our mission. Website. http://www.ncbi.nlm.nih.gov/About/glance/ourmission.html; visited on April 1st 2010.

[NCBI, 2006a] NCBI (2006a). Linkout - linking to a world of resources. Website. http://www.ncbi.nlm.nih.gov/projects/linkout/doc/linkoutoverview.html; visited on June 22nd 2010.

[NCBI, 2006b] NCBI (2006b). Nlm catalog help. Website. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helppubmed&part=pubmedhelp#pubmedhelp.PubMed_Quick_Start; visited on April 11th 2010.

[NCBI, 2006c] NCBI (2006c). Submission of snps to dbsnp. Website. http://www.ncbi.nlm.nih.gov/SNP/how_to_submit.html#QUICKSTART; visited on June 14th 2010.

[NCBI, 2007a] NCBI (2007a). Site map - about ncbi. Website. http://www.ncbi.nlm.nih.gov/About/sitemap.html; visited on May 28th 2010.

[NCBI, 2007b] NCBI (2007b). Tools for data mining. Website. http://www.ncbi.nlm.nih.gov/Tools/index.html; visited on June 21st 2010.

[NCBI, 2009] NCBI (2009). Cancer chromosomes. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=cancerchromosomes; visited on June 21st 2010.

[NCBI, 2010a] NCBI (2010a). Ccds database. Website. http://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi; visited on June 1st 2010.

[NCBI, 2010b] NCBI (2010b). Databases and tools - ftp site. Website.

[NCBI, 2010c] NCBI (2010c). Databases and tools - literature databases. Website.

[NCBI, 2010d] NCBI (2010d). dbest: database of "expressed sequence tags". Website. http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html; visited on June 1st 2010.

[NCBI, 2010e] NCBI (2010e). dbgap - genotypes and phenotypes. Website. http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html; visited on June 21st 2010.

[NCBI, 2010f] NCBI (2010f). dbgss: database of genome survey sequences. Website. http://www.ncbi.nlm.nih.gov/dbGSS/dbGSS_summary.html; visited on June 8th 2010.

[NCBI, 2010g] NCBI (2010g). dbmhc home. Website. http://www.ncbi.nlm.nih.gov/gv/mhc/main.cgi?cmd=init; visited on June 8th 2010.

[NCBI, 2010h] NCBI (2010h). dbsts: database of "sequence tagged sites". Website. http://www.ncbi.nlm.nih.gov/dbSTS/; visited on June 14th 2010.

[NCBI, 2010i] NCBI (2010i). Expressed sequence tags database. Website. http://www.ncbi.nlm.nih.gov/dbEST; visited on June 1st 2010.

[NCBI, 2010j] NCBI (2010j). Faq. Website. http://www.ncbi.nlm.nih.gov/projects/gensat/static/faq.shtml; visited on June 4th 2010.

[NCBI, 2010k] NCBI (2010k). Genbank overview. Website. http://www.ncbi.nlm.nih.gov/genbank/index.html; visited on May 29th 2010.

[NCBI, 2010l] NCBI (2010l). Genes and disease. Website. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=gnd; visited on June 25th 2010.

[NCBI, 2010m] NCBI (2010m). Gensat - methods. Website. http://www.ncbi.nlm.nih.gov/projects/gensat/static/methods_brief.shtml; visited on June 4th 2010.

[NCBI, 2010n] NCBI (2010n). Geo and miame. Website. http://www.ncbi.nlm.nih.gov/geo/info/MIAME.html; visited on June 8th 2010.

[NCBI, 2010o] NCBI (2010o). Introduction to the c++ toolkit. Website. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=toolkit&part=ch_intro; visited on June 22nd 2010.

[NCBI, 2010p] NCBI (2010p). Ncbi at a glance - organizational structure. Website. http://www.ncbi.nlm.nih.gov/About/glance/organizational.html; visited on April 1st 2010.

[NCBI, 2010q] NCBI (2010q). Ncbi sitemap. Website. http://www.ncbi.nlm.nih.gov/Sitemap/index.html; visited on May 30th 2010.

[NCBI, 2010r] NCBI (2010r). Pubmed single citation matcher. Website. http://www.ncbi.nlm.nih.gov/entrez/query/static/citmatch.html; visited on June 22nd 2010.

[NCBI, 2010s] NCBI (2010s). Refseq - genomes, transcripts, proteins. Website. http://www.ncbi.nlm.nih.gov/RefSeq/; visited on June 14th 2010.

[NCBI, 2010t] NCBI (2010t). Sky/m-fish & cgh database. Website. http://www.ncbi.nlm.nih.gov/sky/; visited on June 21st 2010.

[NCBI, 2010u] NCBI (2010u). Trace archive frequently asked questions. Website. http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=show&f=faq&m=main&s=faq; visited on June 14th 2010.

[NCBI, 2010v] NCBI (2010v). Vecscreen. Website. http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html; visited on June 18th 2010.

[NCBI, 2010w] NCBI (2010w). Whole genome shotgun submissions. Website. http://www.ncbi.nlm.nih.gov/genbank/wgs.html; visited on June 18th 2010.

[NCBI et al., 2010] NCBI, PubMed, and PMC (2010). Batch citation matcher. Website. http://www.ncbi.nlm.nih.gov/entrez/getids.cgi; visited on June 22nd 2010.

[NERC, 2010a] NERC (2010a). Nerc open research archive. Website. http://nora.nerc.ac.uk/; visited on July 23th 2010.

[NERC, 2010b] NERC (2010b). Nerc open research archive - browse repository. Website. http://nora.nerc.ac.uk/view/; visited on July 23th 2010.

[NERC, 2010c] NERC (2010c). Nerc open research archive - policies. Website. http://nora.nerc.ac.uk/policies.html; visited on July 23th 2010.

[NERC, 2010d] NERC (2010d). Nerc open research archive - simple search. Website. http://nora.nerc.ac.uk/cgi/search/simple; visited on July 23th 2010.

[NIH, 2010a] NIH (2010a). National institutes of health public access - include pmcid in citations. Website. http://publicaccess.nih.gov/citation_methods.htm; visited on April 7th 2010.

[NIH, 2010b] NIH (2010b). National institutes of health public access - overview. Website. http://publicaccess.nih.gov/; visited on April 7th 2010.

[NLM, 2010] NLM (2010). Question: What is the difference between medlineplus and medline/pubmed? Website. http://www.nlm.nih.gov/medlineplus/faq/difference.html; visited on September 3rd 2010.

[NRC, 2010a] NRC (2010a). Catalogue. Website. http://cisti-icist.nrc-cnrc.gc.ca/eng/services/cisti/catalogue/index.html; visited on July 6th 2010.

[NRC, 2010b] NRC (2010b). Gateway to scientific data - data sets. Website. http://data-donnees.cisti-icist.nrc-cnrc.gc.ca/gsi/ctrl?action=catba; visited on July 7th 2010.

[NRC, 2010c] NRC (2010c). Nrc publications. Website. http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?lang=en; visited on July 7th 2010.

[NRC, 2010d] NRC (2010d). Nrc publications. Website. http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/jsp/nparc_ab.jsp; visited on July 7th 2010.

[NRC, 3 17] NRC (2010-03-17). Worldwidescience.org launches multilingual translation tool. Website. http://cisti-icist.nrc-cnrc.gc.ca/eng/news/cisti/2010/multilingual-translation-tool.html; visited on July 13th 2010.

[NRC, 3 21] NRC (2010-03-21). Gateway to scientific data - data management and curation. Website. http://data-donnees.cisti-icist.nrc-cnrc.gc.ca/gsi/ctrl?action=resba; visited on July 7th 2010.

[NRC, 3 25] NRC (2010-03-25). Information sources. Website. http://cisti-icist.nrc-cnrc.gc.ca/eng/services/cisti/information-sources/index.html; visited on July 6th 2010.

[NRC, 3 29] NRC (2010-03-29). Nrc publications. Website. http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/jsp/nparc_fq.jsp; visited on July 7th 2010.

[NRC, 3 31] NRC (2010-03-31). Gateway to scientific data. Website. http://data-donnees.cisti-icist.nrc-cnrc.gc.ca/gsi/ctrl?lang=en; visited on July 7th 2010.

[NRC-CISTI, 2010a] NRC-CISTI (2010a). Notice - cisti update. Website. http://cisti-icist.nrc-cnrc.gc.ca/eng/news/cisti/2010/nrc-cisti-update.html; visited on July 12th 2010-03-17.

[NRC-CISTI, 2010b] NRC-CISTI (2010b). Nrc-cisti: Discover. Website. http://discover-decouvrir.cisti-icist.nrc-cnrc.gc.ca/dcvr/ctrl?action=faq; visited on July 7th 2010.

[NRC-CISTI, 3 29] NRC-CISTI (2010-03-29). What's in the collection? Website. http://cisti-icist.nrc-cnrc.gc.ca/eng/ibp/cisti/collection/definition.html; visited on July 6th 2010.

[of Energy, 2010] of Energy, U. D. (2010). Multilingual worldwidescience.org launch broadens access to global science. Website. http://www.science.doe.gov/News_Information/News_Room/2010/multilingual.htm; visited on July 13th 2010.

[of Information Sciences and Technology, 2010] of Information Sciences, T. C. and Technology (2010). Citeseerx beta. Website. http://citeseerx.ist.psu.edu/about/site; visited on September 26th 2010.

[OSTI, 6 11] OSTI (2010-06-11). Worldwidescience.org goes multilingual. Website. http://www.osti.gov/news/pressreleases/2010/june/wws.shtml; visited on July 13th 2010.

[PINTS, 2003] PINTS (2003). Pints - patterns in non-homologous tertiary structures. Website. http://www.russell.embl.de/pints/; visited on September 20th 2010.

[PMC, 2005] PMC (2005). Author manuscripts in pmc. Website. http://www.ncbi.nlm.nih.gov/pmc/about/authorms.html; visited on April 7th 2010.

[PMC, 2008] PMC (2008). Pmc overview. Website. http://www.ncbi.nlm.nih.gov/pmc/about/intro.html; visited on April 7th 2010.

[PMC, 2009] PMC (2009). Pmc file submission specifications. Website. http://www.ncbi.nlm.nih.gov/pmc/about/PMC_Filespec.html; visited on April 7th 2010.

[PMC, 2010] PMC (2010). New in pmc. Website. http://www.ncbi.nlm.nih.gov/pmc/about/new_in_pmc.html; visited on April 7th 2010.

[PolyPhen, 2010a] PolyPhen (2010a). Help contents. Website. http://genetics.bwh.harvard.edu/pph/pph_help.html; visited on September 10th 2010.

[PolyPhen, 2010b] PolyPhen (2010b). Polyphen: prediction of functional effect of human nssnps. Website. http://coot.embl.de/PolyPhen/; visited on September 10th 2010.

[Probe, 2010a] Probe, N. (2010a). Applications. Website. http://www.ncbi.nlm.nih.gov/projects/genome/probe/doc/Applications.shtml; visited on June 14th 2010.

[Probe, 2010b] Probe, N. (2010b). Probe - reagents of functional genomics. Website. http://www.ncbi.nlm.nih.gov/projects/genome/probe/doc/Overview.shtml; visited on June 14th 2010.

[Protein, 2010] Protein, N. (2010). Protein - translation of life. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=protein; visited on June 18th 2010.

[PROUST, 2002a] PROUST (2002a). About proust-ii and how to use the server. Website. http://www.russell.embl.de/proust2/help.html; visited on September 10th 2010.

[PROUST, 2002b] PROUST (2002b). Prediction of unknown sub-types (proust) ii. Website. http://www.russell.embl.de/proust2/; visited on September 10th 2010.

[PubChem, 2010a] PubChem, N. (2010a). Pubchem bioassay. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=pcassay; visited on June 20th 2010.

[PubChem, 2010b] PubChem, N. (2010b). Pubchem compound. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=pccompound; visited on June 20th 2010.

[PubChem, 2010c] PubChem, N. (2010c). Pubchem substance. Website. http://www.ncbi.nlm.nih.gov/sites/entrez?db=pcsubstance; visited on June 20th 2010.

[PubMed,] PubMed, N. Pubmed help.

[QUALITY, 2009a] QUALITY, M. (2009a). About us. Website. http://www.microarray-quality.org/about.html; visited on September 23rd 2010.

[QUALITY, 2009b] QUALITY, M. (2009b). Partners. Website. http://www.microarray-quality.org/partners.html; visited on September 23rd 2010.

[Reactome, 2010a] Reactome (2010a). About reactome. Website. http://www.reactome.org/about.html; visited on July 28th 2010.

[Reactome, 2010b] Reactome (2010b). Reactome - a curated knowledgebase of biological pathways. Website. http://www.reactome.org/; visited on July 28th 2010.

[Rebholz-Schuhman et al., 2007] Rebholz-Schuhman, D., Cameron, G., Clarc, D., van Mulligen, E., Coatrieux, J.-L., Barbolla, E. D. H., Martin-Sanchez, F., Milanesi, L., Porro, I., Beltrame, F., Tollis, I., and der Lei, J. V. (2007). Symbiomatics: Synergies in medical informatics and bioinformatics - exploring current scientific literature for emerging topics. *BioMed Central*.

[scirus, 2010] scirus (2010). About scirus... Website. http://www.scirus.com/srsapp/aboutus/; visited on September 26th 2010.

[sequere, 2010a] sequere (2010a). sequere.eu. Website. http://www.science.sequere.eu/; visited on September 26th 2010.

[sequere, 2010b] sequere (2010b). Welcome to your new homepage - sequere.eu all in one place! Website. http://www.sequere.eu/; visited on September 26th 2010.

[SHOT, 2002] SHOT (2002). Shot shared ortholog and gene order tree reconstruction tool. Website. http://coot.embl.de/~korbel/SHOT//; visited on September 21st 2010.

[SIB, 2010a] SIB (2010a). Enzyme - enzyme nomenclature database. Website. http://www.expasy.org/enzyme/; visited on July 27th 2010.

[SIB, 2010b] SIB (2010b). Hpi - the uniprotkb/swiss-prot human proteome initiative. Website. http://www.expasy.org/sprot/hpi/hpi_stat.html; visited on July 27th 2010.

[SIRW, 2010] SIRW (2010). Sirw help. Website. http://sirw.embl.de/sirw/new/help.html; visited on September 21st 2010.

[Siveter et al., 2007] Siveter, D. J., Fortey, R. A., Sutton, M. D., Briggs, D. E. G., and Siveter, D. J. (2007). A silurian 'marrellomorph' arthropod. *The Royal Society*.

[SLING, 2010] SLING (2010). Sling - serving life-science information for the next generation. Website. http://www.sling-fp7.org/; visited on August 29th 2010.

[SMART, 2010a] SMART (2010a). Smart - simple modular architecture research tool. Website. http://smart.embl-heidelberg.de/; visited on September 11th 2010.

[SMART, 2010b] SMART (2010b). Smart - simple modular architecture research tool. Website. http://smart.embl-heidelberg.de/help/smart_about.shtml; visited on September 11th 2010.

[Sutton et al., 2001] Sutton, M. D., Briggs, D. E. G., Siveter, D. J., and Siveter, D. J. (2001). Methodologies for the visualization and reconstruction of three-dimensional fossils from the silurian herefordshire lagerst"atte. *Palaentologia Electronica*.

[SYBARIS, 2010a] SYBARIS (2010a). Sybaris home. Website. `http://www.sybaris-fp7.eu/`; visited on August 29th 2010.

[SYBARIS, 2010b] SYBARIS (2010b). Sybaris home. Website. `http://www.sybaris-fp7.eu/organization`; visited on August 29th 2010.

[UKPMC, 2010a] UKPMC (2010a). Advanced search. Website. `http://ukpmc.ac.uk/advancesearch`; visited on July 23th 2010.

[UKPMC, 2010b] UKPMC (2010b). Grant lookup tool. Website. `http://ukpmc.ac.uk/GrantLookup/`; visited on July 23th 2010.

[UKPMC, 2010c] UKPMC (2010c). Ukpmc faq. Website. `http://ukpmc.ac.uk/FAQ`; visited on July 23th 2010.

[UKPMC, 2010d] UKPMC (2010d). Why use uk pubmed central? Website. `http://ukpmc.ac.uk/About`; visited on July 20th 2010.

[UniGene, 2010a] UniGene, N. (2010a). Unigene - build procedure-genome based. Website. `http://www.ncbi.nlm.nih.gov/UniGene/build2.html`; visited on June 4th 2010.

[UniGene, 2010b] UniGene, N. (2010b). Unigene - build procedure-transcriptome based. Website. `http://www.ncbi.nlm.nih.gov/UniGene/build1.html`; visited on June 4th 2010.

[UniGene, 2010c] UniGene, N. (2010c). Unigene: An organized view of the transcriptome. Website. `http://www.ncbi.nlm.nih.gov/unigene`; visited on June 4th 2010.

[UniGene, 2010d] UniGene, N. (2010d). Unigene faq. Website. `http://www.ncbi.nlm.nih.gov/UniGene/FAQ.shtml`; visited on June 4th 2010.

[Velankar, 2009] Velankar, S. (2009). Emdb search tool. Website. `http://www.ebi.ac.uk/pdbe-srv/emsearch/`; visited on September 22nd 2010.