WWW DATA RESOURCES

Tom Oldfield

A proliferation of web sites provides different views, slices or means of access to data.

An increasingly dense network of these sites provides links among databanks and information-retrieval engines.

These links provide useful avenues to applications; but they also provide routes for propagation of errors in raw or immature data. Subsequently corrected in the databanks, the corrections are not propagated to the annotation.

EMBL-EBI Bioinformatics databases on the WWW

- Loose definition of database here
  - Most are “databanks”

- Vary widely in terms of offerings, data, tools and specialization

- Vary widely in terms of data collection methodologies
  - Some data is curated, some just “collect” data
  - Little validity information is provided

Can Databases really be used to answer questions ?

Depends what you want ?

- **Will this drug kill my patient ?**
  - **Needs to look up all the databases :Sequence, structure, expression, bio-pathways/cell biology, QSAR analysis, drug libraries, chemistry, medical…**
  - **Provide statistical analysis of validity of results : return dosage result.**
- **Can’t do this : Not even close !**

Databases

- Tell me about Kinase binding
  - Closer to this type of thing, providing a summary from multiple databases
  - But we still can’t do this !
  - You still need to go to each DB
  - Combine your own results

- Comparative analysis
  - All have Atlas pages (single summary)
  - Should produce comparisons

Protein Databases

- Protein sequence collections
- Clustering of protein data into families
  - Sequence
  - Structure
  - function
- Specialized protein sites
  - Organism
  - Function
  - Large variety of enzymes

Classifications

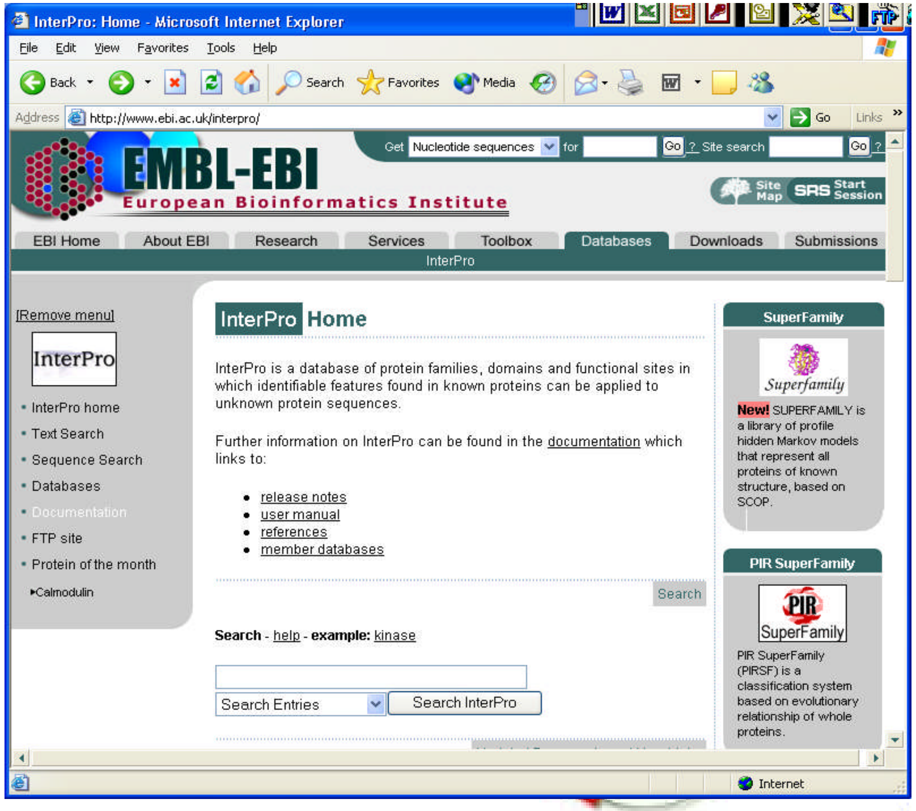
- **Classifying proteins based on five types of data:**
  - **Their domain structures**
  - **Protein-protein interactions**
  - **Genetic interactions**
  - **Co-participation in protein complexes**
  - **Cell cycle gene expression measurements**

EMBL-EBI

Protein Databases: InterPro

- a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences
- amalgamating the major protein signature databases, data have been manually integrated and *curated.*
  - PROSITE
  - Pfam
  - PRINTS
  - ProDom
  - SMART
  - TIGRFAMs

http://www.ebi.ac.uk/interpro/



InterPro Home

InterPro is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences.

Further information on InterPro can be found in the documentation which links to:

- release notes
- user manual
- references
- member databases

Search - help - **example:** kinase

SuperFamily

New! SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure, based on SCOP.

PIR SuperFamily

PIR SuperFamily (PIRSF) is a classification system based on evolutionary relationship of whole proteins.

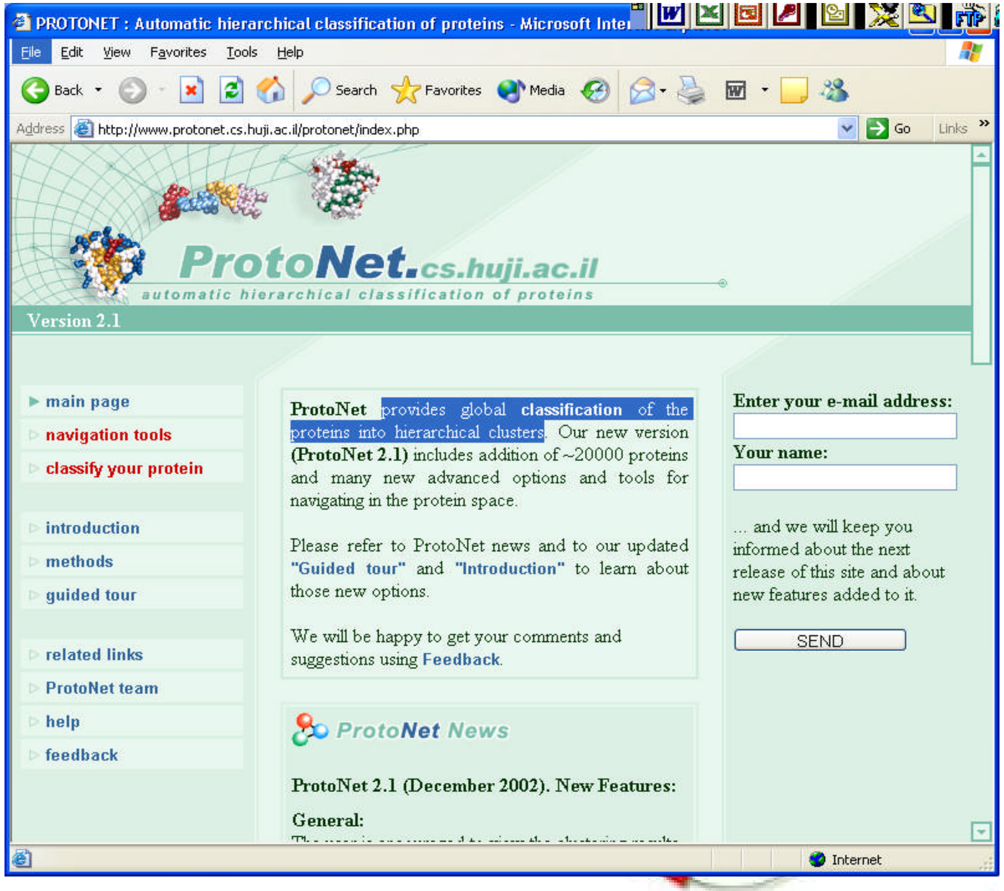
EMBL-EBI

Protein Databases: ProtoNet

- **provides global classification of the proteins, from the SWISS-PROT (UNIPROT) database into hierarchical clusters**

- **clustering is based on an all-against-all BLAST similarity search**

http://www.protonet.cs.huji.ac.il/


PROTONET : Automatic hierarchical classification of proteins - Microsoft Inter
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.protonet.cs.huji.ac.il/protonet/index.php
Go Links
ProtoNet.cs.huji.ac.il
automatic hierarchical classification of proteins
Version 2.1
main page
navigation tools
classify your protein
introduction
methods
guided tour
related links
ProtoNet team
help
feedback
Internet

ProtoNet provides global **classification** of the proteins into hierarchical clusters. Our new version **(ProtoNet 2.1)** includes addition of ~20000 proteins and many new advanced options and tools for navigating in the protein space.

Please refer to ProtoNet news and to our updated **"Guided tour"** and **"Introduction"** to learn about those new options.

We will be happy to get your comments and suggestions using **Feedback**.

Enter your e-mail address:

Your name:

... and we will keep you informed about the next release of this site and about new features added to it.

SEND

ProtoNet News

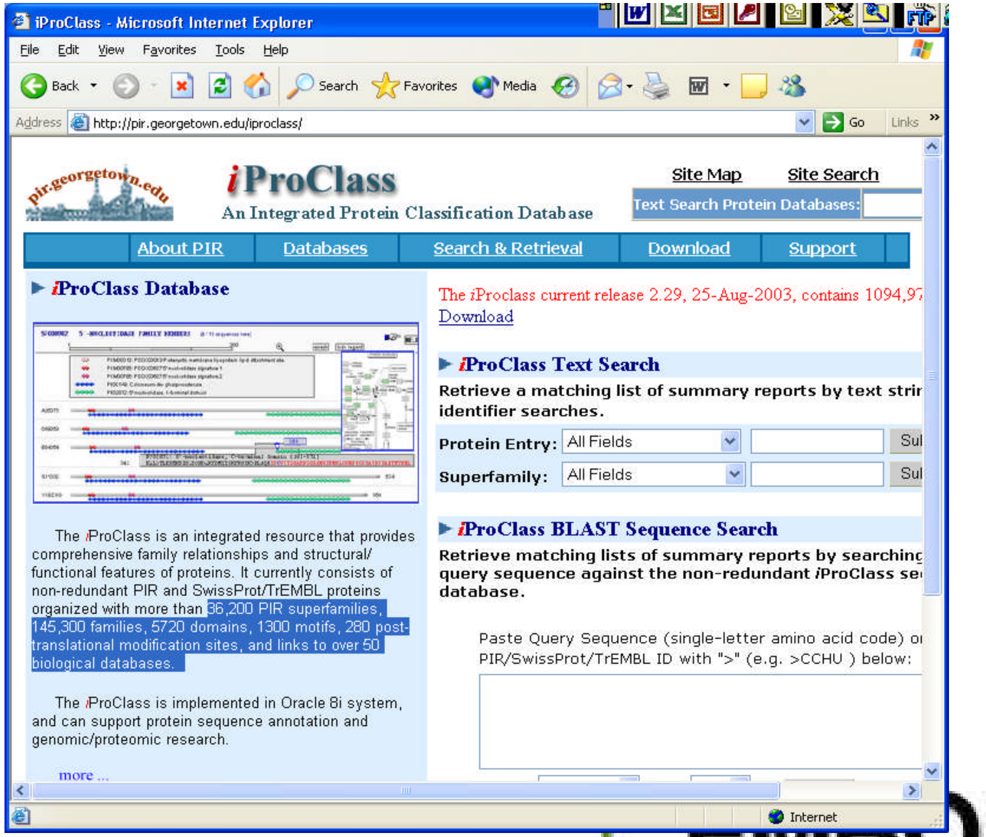
ProtoNet 2.1 (December 2002). New Features:

General:

Protein Databases: iProClass

- an integrated resource that provides comprehensive family relationships and structural/functional features of proteins
- currently consists of non-redundant PIR and SwissProt/TrEMBL (UNIPROT) proteins
  - 36,200 PIR superfamilies
  - 145,300 families
  - 5720 domains
  - 1300 motifs
  - 280 post-translational modification sites
  - links to over 50 biological databases.

http://pir.georgetown.edu/iproclass/



iProClass
An Integrated Protein Classification Database

iProClass Database

The iProclass current release 2.29, 25-Aug-2003, contains 1094,97

Download

The iProClass is an integrated resource that provides comprehensive family relationships and structural/ functional features of proteins. It currently consists of non-redundant PIR and SwissProt/TrEMBL proteins organized with more than 36,200 PIR superfamilies, 145,300 families, 5720 domains, 1300 motifs, 280 post-translational modification sites, and links to over 50 biological databases.

The iProClass is implemented in Oracle 8i system, and can support protein sequence annotation and genomic/proteomic research.

iProClass Text Search

Retrieve a matching list of summary reports by text strir identifier searches.

Protein Entry: All Fields

Superfamily: All Fields

iProClass BLAST Sequence Search

Retrieve matching lists of summary reports by searching query sequence against the non-redundant iProClass se database.

Paste Query Sequence (single-letter amino acid code) o PIR/SwissProt/TrEMBL ID with ">" (e.g. >CCHU) below:

Protein Databases: Others

- Nuclear Protein Database – Proteins localized in the nucleus
- PLANT-PIs – Plant protease inhibitors
- UNIPROT– Curated protein sequences
- SENTRA – Sensory signal transduction proteins
- Ribonuclease P Database

Protein Sequence Motifs

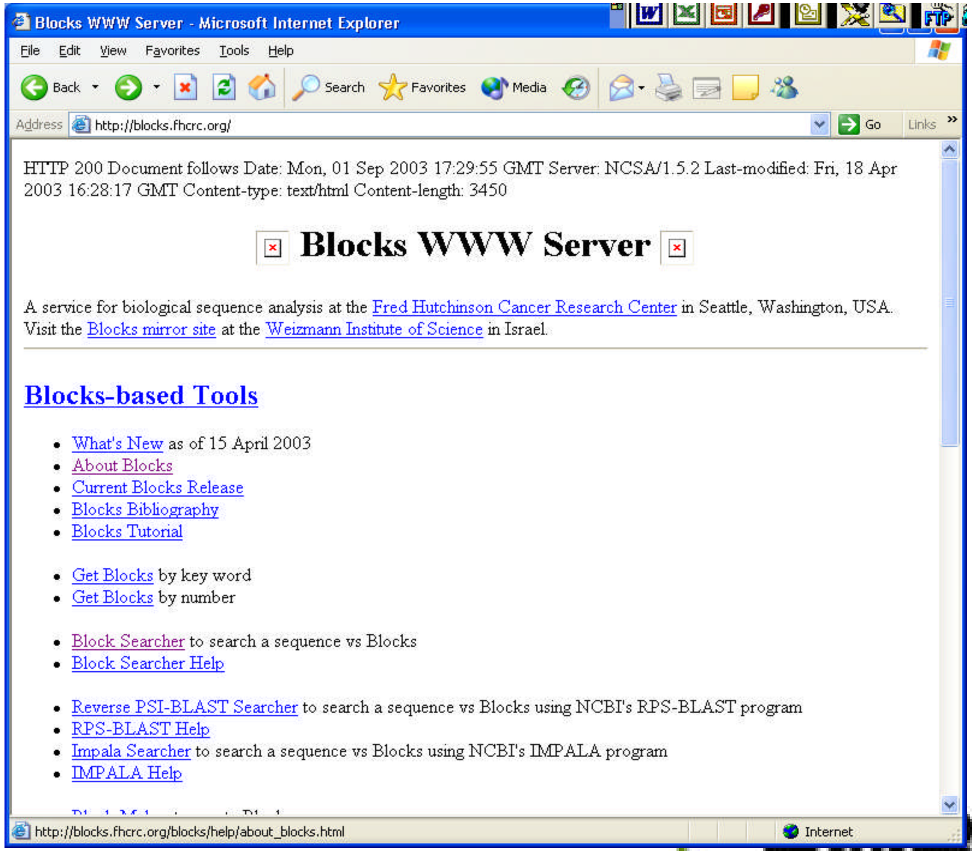
- Alignment of protein sequences
- Organization of proteins into families

EMBL-EBI

Protein Sequence Motifs: BLOCKS

http://blocks.fhcrc.org/

- Multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins
- Tools:
  - Block Searcher -- compare a protein or DNA sequence to a database of protein blocks
  - Get Blocks -- retrieve blocks
  - Block Maker -- create new blocks


Blocks WWW Server - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address http://blocks.fhcrc.org/
Go
Links
http://blocks.fhcrc.org/blocks/help/about_blocks.html
Internet

HTTP 200 Document follows Date: Mon, 01 Sep 2003 17:29:55 GMT Server: NCSA/1.5.2 Last-modified: Fri, 18 Apr 2003 16:28:17 GMT Content-type: text/html Content-length: 3450

Blocks WWW Server

A service for biological sequence analysis at the Fred Hutchinson Cancer Research Center in Seattle, Washington, USA. Visit the Blocks mirror site at the Weizmann Institute of Science in Israel.

Blocks-based Tools

- What's New as of 15 April 2003
- About Blocks
- Current Blocks Release
- Blocks Bibliography
- Blocks Tutorial

- Get Blocks by key word
- Get Blocks by number

- Block Searcher to search a sequence vs Blocks
- Block Searcher Help

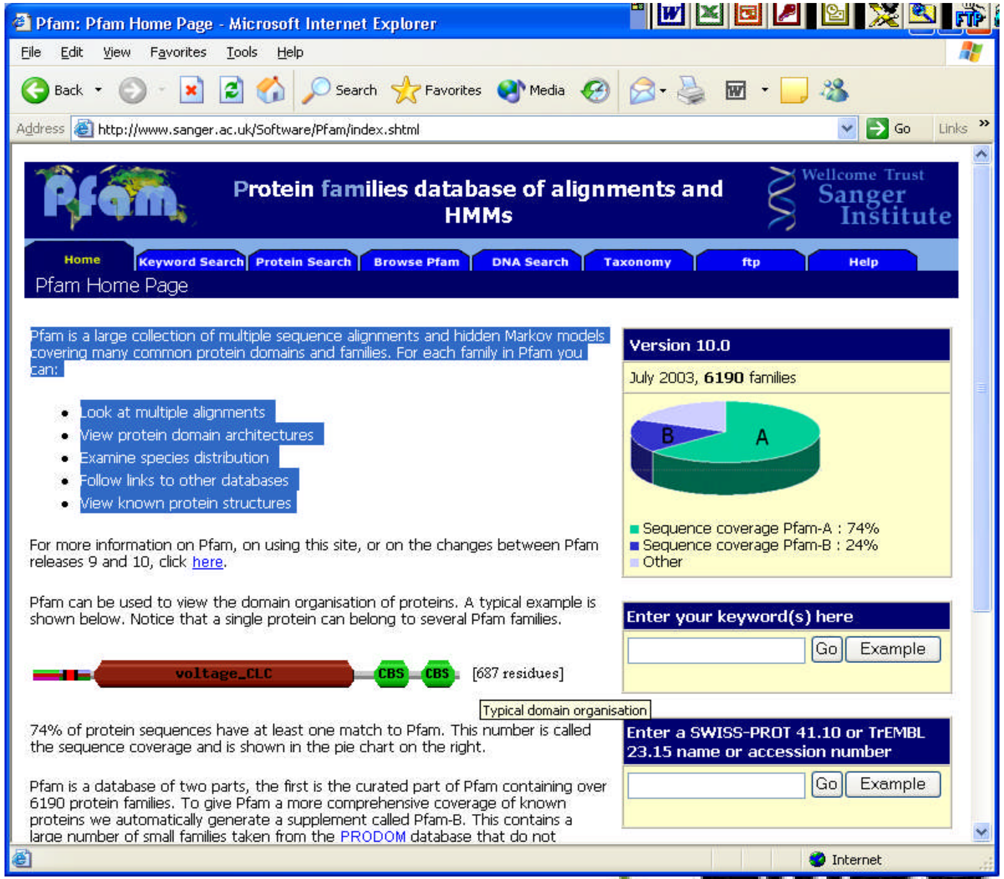
- Reverse PSI-BLAST Searcher to search a sequence vs Blocks using NCBI's RPS-BLAST program
- RPS-BLAST Help
- Impala Searcher to search a sequence vs Blocks using NCBI's IMPALA program
- IMPALA Help

EMBL-EBI

Protein Sequence Motifs:Pfam

- **Large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families.**
- **For each family in Pfam you can:**
  - **Look at multiple alignments**
  - **View protein domain architectures**
  - **Examine species distribution**
  - **Follow links to other databases**
  - **View known protein structures**

http://www.sanger.ac.uk/Software/Pfam/



http://www.sanger.ac.uk/Software/Pfam/index.shtml

Pfam — Protein families database of alignments and HMMs — Wellcome Trust Sanger Institute

Pfam Home Page

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families. For each family in Pfam you can:

- Look at multiple alignments
- View protein domain architectures
- Examine species distribution
- Follow links to other databases
- View known protein structures

For more information on Pfam, on using this site, or on the changes between Pfam releases 9 and 10, click here.

Pfam can be used to view the domain organisation of proteins. A typical example is shown below. Notice that a single protein can belong to several Pfam families.

[687 residues]

Typical domain organisation

74% of protein sequences have at least one match to Pfam. This number is called the sequence coverage and is shown in the pie chart on the right.

Pfam is a database of two parts, the first is the curated part of Pfam containing over 6190 protein families. To give Pfam a more comprehensive coverage of known proteins we automatically generate a supplement called Pfam-B. This contains a large number of small families taken from the PRODOM database that do not

Version 10.0

July 2003, **6190** families

🟩 Sequence coverage Pfam-A : 74%
🟦 Sequence coverage Pfam-B : 24%
▢ Other

Enter your keyword(s) here

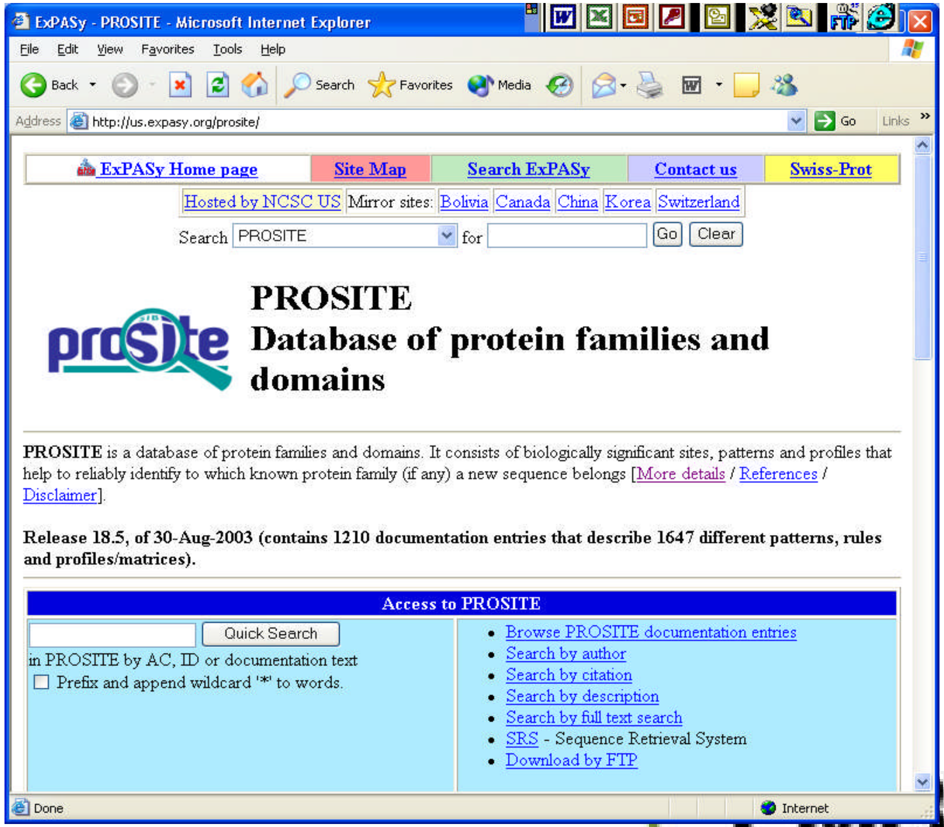
Enter a SWISS-PROT 41.10 or TrEMBL 23.15 name or accession number

EMBL-EBI

Protein Sequence Motifs:PROSITE

- Database of protein families and domains.
  - biologically characterized sites
  - patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs
- Currently contains patterns and profiles specific for more than a thousand protein families or domains.
- Each of these signatures comes with documentation on the structure and function of these proteins

http://us.expasy.org/prosite/


ExPASy - PROSITE - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address http://us.expasy.org/prosite/
Go
Links
ExPASy Home page
Site Map
Search ExPASy
Contact us
Swiss-Prot
Hosted by NCSC US
Mirror sites: Bolivia Canada China Korea Switzerland
Search PROSITE for
Go
Clear
prosite
Done
Internet

PROSITE
Database of protein families and domains

PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs [More details / References / Disclaimer].

Release 18.5, of 30-Aug-2003 (contains 1210 documentation entries that describe 1647 different patterns, rules and profiles/matrices).

Access to PROSITE

Quick Search
in PROSITE by AC, ID or documentation text
□ Prefix and append wildcard '*' to words.

- Browse PROSITE documentation entries
- Search by author
- Search by citation
- Search by description
- Search by full text search
- SRS - Sequence Retrieval System
- Download by FTP

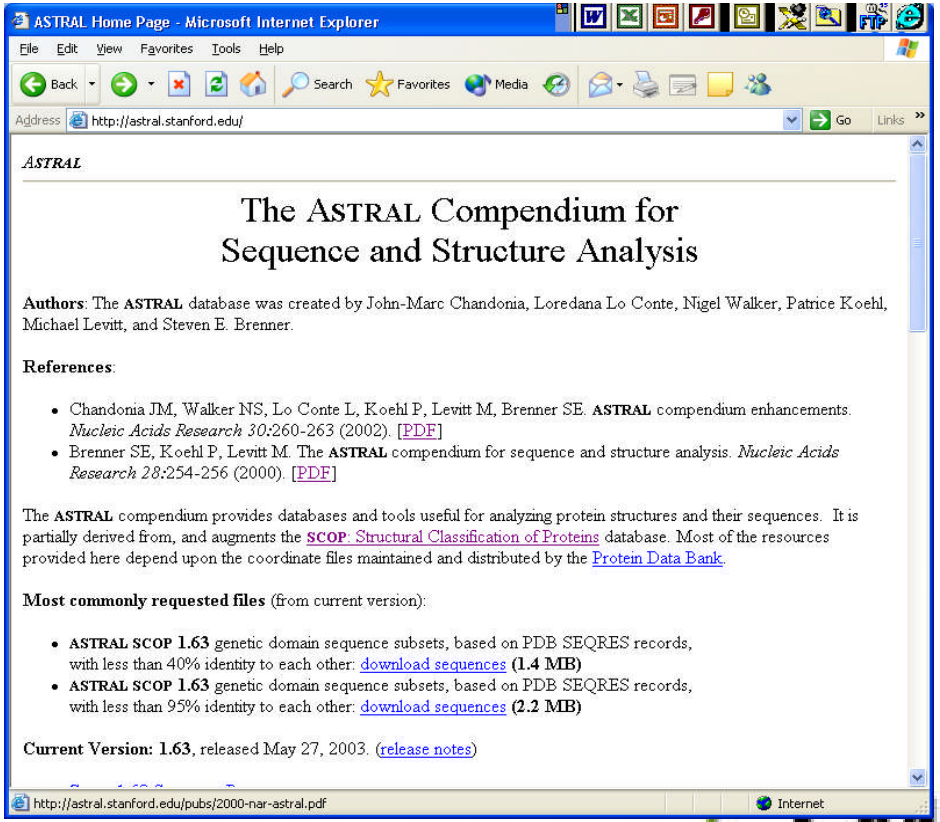
Protein Sequence Motifs: Others

- ASC – Active Sequence Collection – Biologically active oligopeptides
- ClusTr – Automatic classification of SWISS-PROT and TrEMBL proteins
- TMPDB – Experimentally-characterized transmembrane topology
- O-GLYCBASE – O- and C- linked glycosylation sites in proteins

Structure: ASTRAL

- ➢ **Provides databases and tools useful for analyzing protein structures and their sequences**
- ➢ **Partially derived from the SCOP database (Structural Classification of Proteins)**

http://astral.berkeley.edu/


ASTRAL Home Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address http://astral.stanford.edu/
Go
Links
http://astral.stanford.edu/pubs/2000-nar-astral.pdf
Internet

ASTRAL

The ASTRAL Compendium for Sequence and Structure Analysis

Authors: The **ASTRAL** database was created by John-Marc Chandonia, Loredana Lo Conte, Nigel Walker, Patrice Koehl, Michael Levitt, and Steven E. Brenner.

References:

- Chandonia JM, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. **ASTRAL** compendium enhancements. *Nucleic Acids Research 30:*260-263 (2002). [PDF]
- Brenner SE, Koehl P, Levitt M. The **ASTRAL** compendium for sequence and structure analysis. *Nucleic Acids Research 28:*254-256 (2000). [PDF]

The **ASTRAL** compendium provides databases and tools useful for analyzing protein structures and their sequences. It is partially derived from, and augments the **SCOP**: Structural Classification of Proteins database. Most of the resources provided here depend upon the coordinate files maintained and distributed by the Protein Data Bank.

Most commonly requested files (from current version):

- **ASTRAL SCOP 1.63** genetic domain sequence subsets, based on PDB SEQRES records, with less than 40% identity to each other: download sequences **(1.4 MB)**
- **ASTRAL SCOP 1.63** genetic domain sequence subsets, based on PDB SEQRES records, with less than 95% identity to each other: download sequences **(2.2 MB)**

Current Version: 1.63, released May 27, 2003. (release notes)

Structure: Other Examples

- CADB – conformation angles of protein structures, with associated crystallographic data
- Database of Macromolecular Movements
- DSDBase – Disulfide Bonds in proteins
- PSSH – alignment between sequences and tertiary structures
- SUPERFAMILY – Assignments of proteins to structural superfamilies

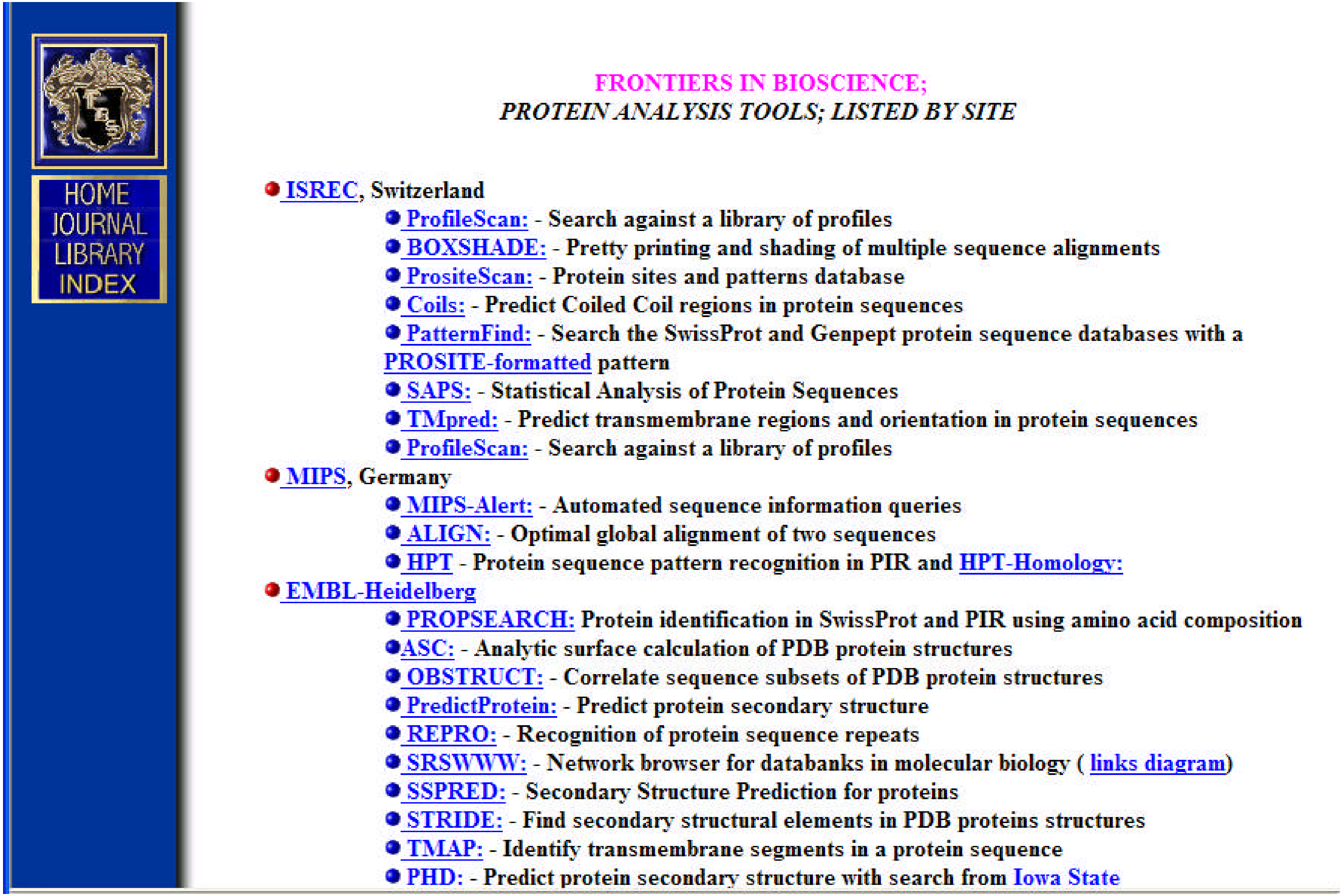
Other Databases: Intermolecular Interactions

- BIND – Molecular interactions, complexes and pathways
- DIP (Database of Interacting Proteins) – Experimentally determined protein-protein interactions
- KDBI – Kinetic data on biomolecular interactions

Metabolic Pathways and Cellular Regulation

- **KEGG – Kyoto Encyclopedia of Genes and Genomes**
- **MetaCyc – Metabolic Pathways and Enzymes from Various organisms**
- **PathDB**
- **EcoCyc – E. coli K-12 genome and pathway data**
- **PRODORIC – gene regulation and regulatory networks in prokaryotes**

http://www.bioscience.org/urllists/pranal1.htm

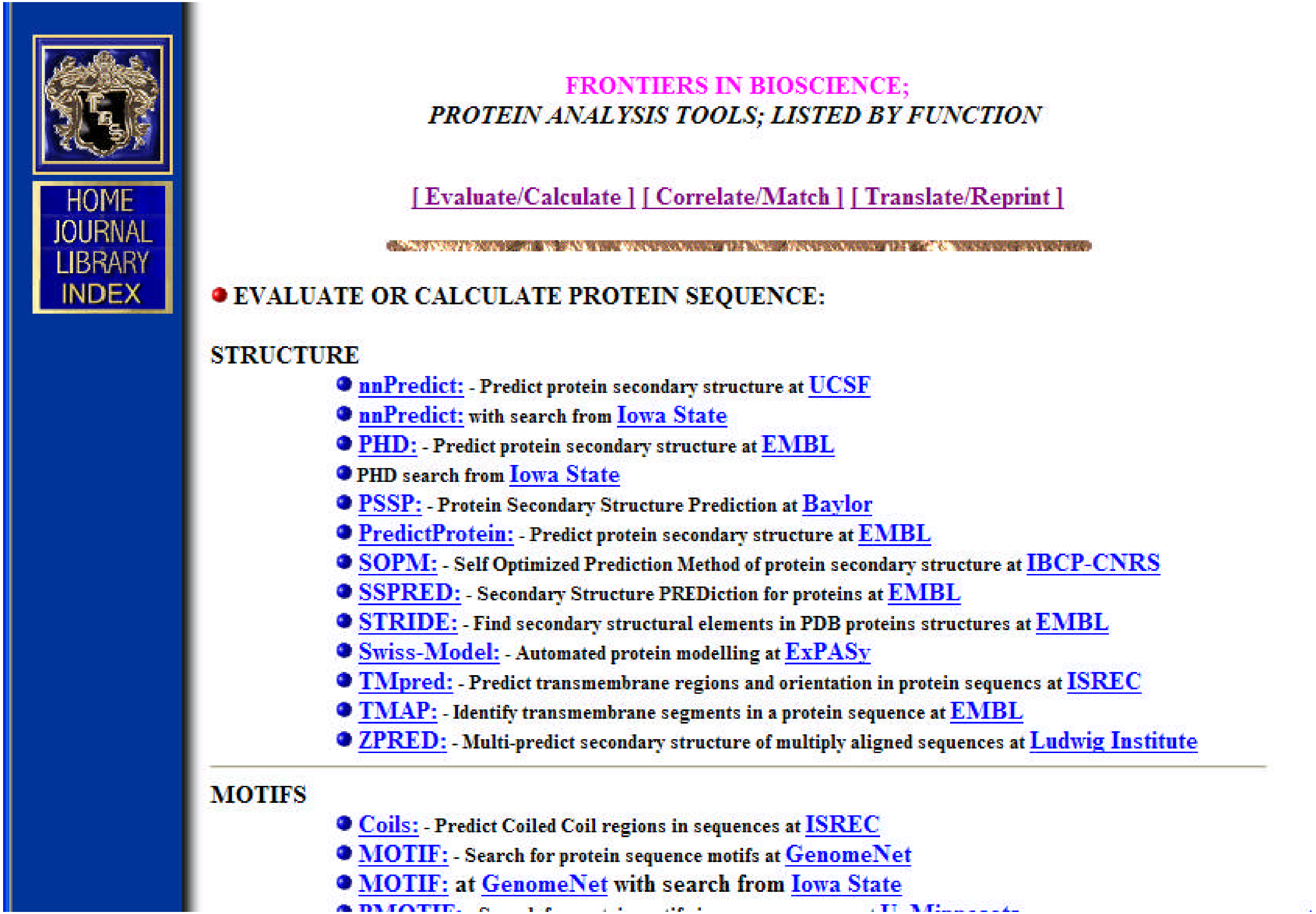


FRONTIERS IN BIOSCIENCE;

PROTEIN ANALYSIS TOOLS; LISTED BY SITE

- ISREC, Switzerland
  - ProfileScan: - Search against a library of profiles
  - BOXSHADE: - Pretty printing and shading of multiple sequence alignments
  - PrositeScan: - Protein sites and patterns database
  - Coils: - Predict Coiled Coil regions in protein sequences
  - PatternFind: - Search the SwissProt and Genpept protein sequence databases with a PROSITE-formatted pattern
  - SAPS: - Statistical Analysis of Protein Sequences
  - TMpred: - Predict transmembrane regions and orientation in protein sequences
  - ProfileScan: - Search against a library of profiles
- MIPS, Germany
  - MIPS-Alert: - Automated sequence information queries
  - ALIGN: - Optimal global alignment of two sequences
  - HPT - Protein sequence pattern recognition in PIR and HPT-Homology:
- EMBL-Heidelberg
  - PROPSEARCH: Protein identification in SwissProt and PIR using amino acid composition
  - ASC: - Analytic surface calculation of PDB protein structures
  - OBSTRUCT: - Correlate sequence subsets of PDB protein structures
  - PredictProtein: - Predict protein secondary structure
  - REPRO: - Recognition of protein sequence repeats
  - SRSWWW: - Network browser for databanks in molecular biology (links diagram)
  - SSPRED: - Secondary Structure Prediction for proteins
  - STRIDE: - Find secondary structural elements in PDB proteins structures
  - TMAP: - Identify transmembrane segments in a protein sequence
  - PHD: - Predict protein secondary structure with search from Iowa State

http://www.bioscience.org/urllists/protanal.htm



FRONTIERS IN BIOSCIENCE;

PROTEIN ANALYSIS TOOLS; LISTED BY FUNCTION

[Evaluate/Calculate] [Correlate/Match] [Translate/Reprint]

● **EVALUATE OR CALCULATE PROTEIN SEQUENCE:**

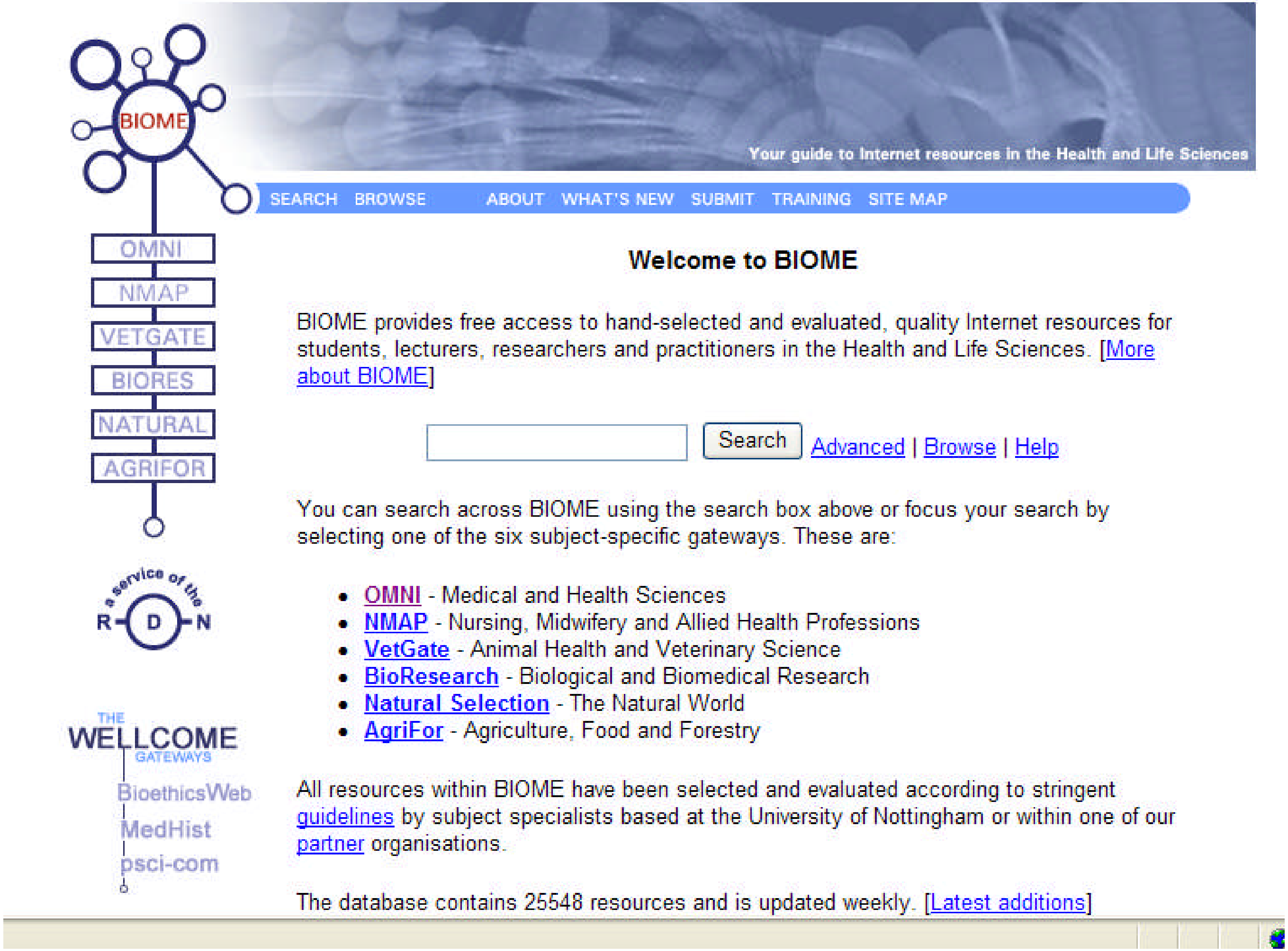
STRUCTURE

- nnPredict: - Predict protein secondary structure at UCSF
- nnPredict: with search from Iowa State
- PHD: - Predict protein secondary structure at EMBL
- PHD search from Iowa State
- PSSP: - Protein Secondary Structure Prediction at Baylor
- PredictProtein: - Predict protein secondary structure at EMBL
- SOPM: - Self Optimized Prediction Method of protein secondary structure at IBCP-CNRS
- SSPRED: - Secondary Structure PREDiction for proteins at EMBL
- STRIDE: - Find secondary structural elements in PDB proteins structures at EMBL
- Swiss-Model: - Automated protein modelling at ExPASy
- TMpred: - Predict transmembrane regions and orientation in protein sequencs at ISREC
- TMAP: - Identify transmembrane segments in a protein sequence at EMBL
- ZPRED: - Multi-predict secondary structure of multiply aligned sequences at Ludwig Institute

MOTIFS

- Coils: - Predict Coiled Coil regions in sequences at ISREC
- MOTIF: - Search for protein sequence motifs at GenomeNet
- MOTIF: at GenomeNet with search from Iowa State
- PMOTIF: - Search for protein motifs in sequences at U. Minnesota

http://biome.ac.uk/



Welcome to BIOME

BIOME provides free access to hand-selected and evaluated, quality Internet resources for students, lecturers, researchers and practitioners in the Health and Life Sciences. [More about BIOME]

You can search across BIOME using the search box above or focus your search by selecting one of the six subject-specific gateways. These are:

- OMNI - Medical and Health Sciences
- NMAP - Nursing, Midwifery and Allied Health Professions
- VetGate - Animal Health and Veterinary Science
- BioResearch - Biological and Biomedical Research
- Natural Selection - The Natural World
- AgriFor - Agriculture, Food and Forestry

All resources within BIOME have been selected and evaluated according to stringent guidelines by subject specialists based at the University of Nottingham or within one of our partner organisations.

The database contains 25548 resources and is updated weekly. [Latest additions]

General Search Sites

Search engine	Comments
http://biome.ac.uk/	A searchable catalogue of Internet sites and resources covering the health and life sciences.
http://www.expasy.ch/BioHunt/	Created especially for retrieving molecular biology sites, but some non-relevant results may appear.
http://teoma.com/	Gives "relevant" results list, suggests sites for narrowing your search, and has links to "experts and enthusiasts" collections.
http://www.google.com/	Google is one of the easiest search engines to use. It also offers an 'Advanced search' facility
http://www.searchenginewatch.com/	This site has up to date information on search engines and tips on how to search effectively.

Meta (Multi) Search Engines

Iquick (http://www.ixquick.com/)	Ranked as one of the best meta-search engines
Profusion (http://www.profusion.com/)	Provides a range of other facilities, including a general subject directory, and advanced search options
SurfWax http://www.surfwax.com/	Searches against major engines or provides those who open free accounts the ability to choose from a list of hundreds.

Most of the authoritative information accessible over the Internet is invisible to most popular search engines, such as AltaVista, HotBot and Google.

This invaluable information resides on the "The Invisible Web", which is largely comprised of content-rich databases from universities, libraries, associations, businesses and government agencies from around the world

Subject Directories and The Invisible Web

Digital Librarian
http://www.digital-librarian.com/
A librarian's choice of the best of the web.

BUBL Link
http://www.bubl.ac.uk/link/
BUBL provides access to selected Internet resources covering all academic subject areas. It does have a UK focus

Galaxy
http://www.galaxy.com/
This directory provides a large number of subject categories

INFOMINE
http://infomine.ucr.edu/
Scholarly internet resource collections.

Invisible Web Directory
http://www.invisible-web.net/
A directory of some of the best resources the Invisible Web has to offer.

Complete Planet
http://aip.completeplanet.com/
A comprehensive listing of "deep" Web searchable databases, search engines and sites.

A Collection of Special Search Engines
http://www.leidenuniv.nl/ub/biv/specials.htm
Good alternatives to the big search engines. Many are subject specific.

Summary

- Huge range of resources on the WWW
  - Some are more equal that others
  - Some are maintained – some not
  - Little information on data validity
  - Search engines can be used to find new ones