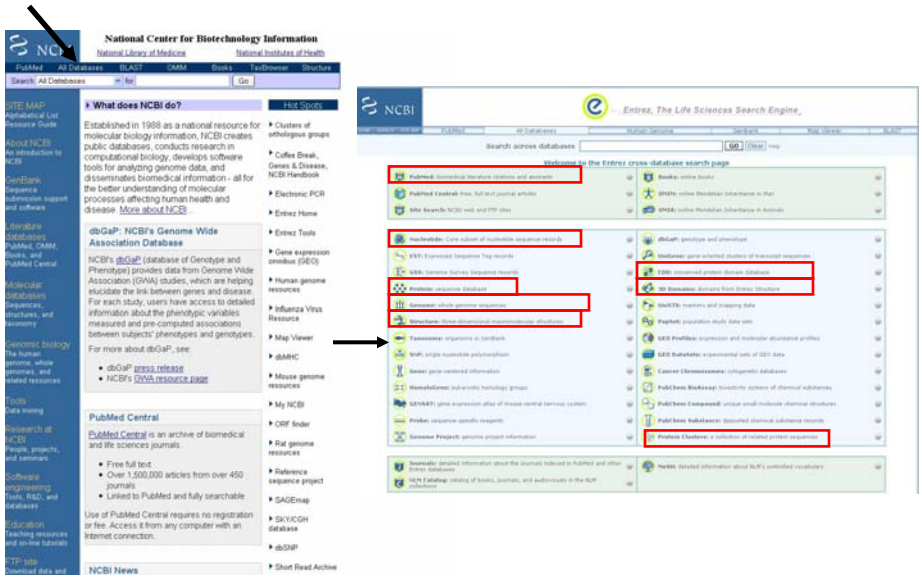
Web-based Bioinformatics Applications in Proteomics

Chiquito Crasto
ccrasto@genetics.uab.edu
January 30, 2009

NCBI (National Center for Biotechnology Information)

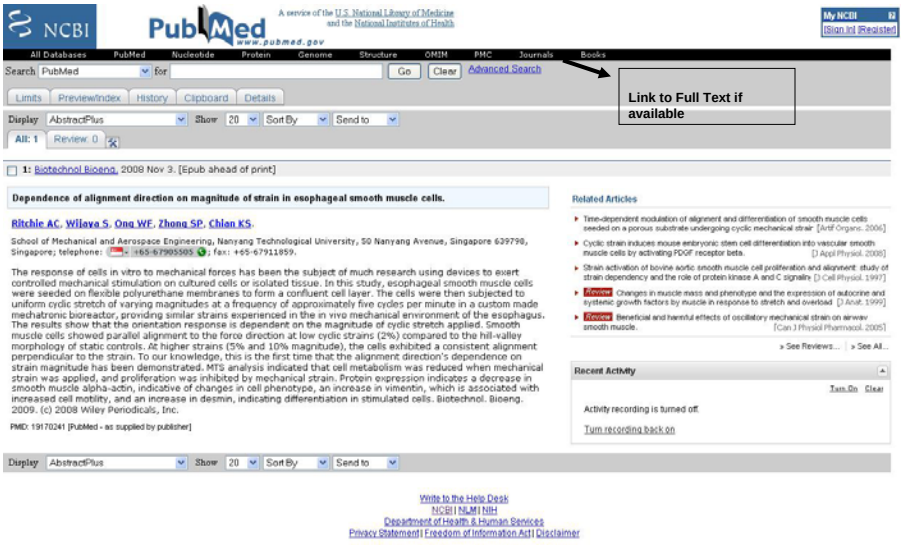
http://www.ncbi.nlm.nih.gov/



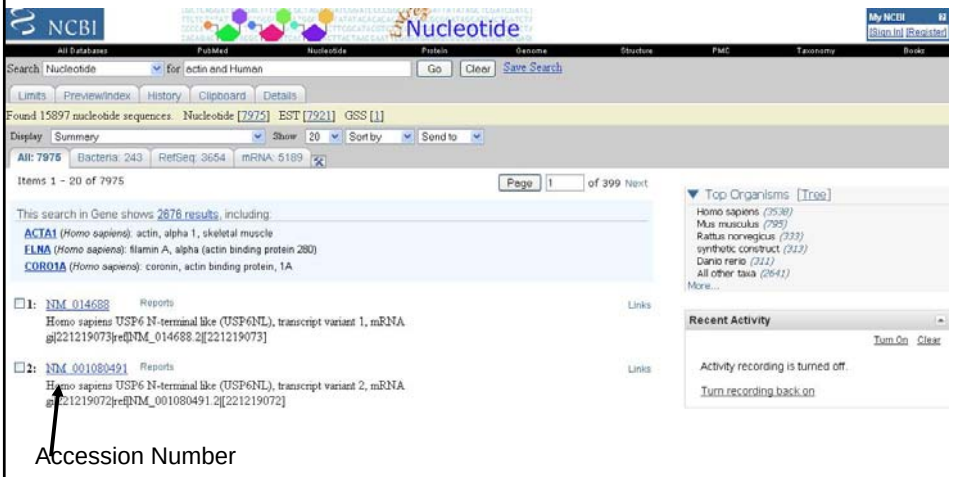
Pubmed—repository of biomedical abstracts

Information in Pubmed is available in several formats.
Abstracts can be downloaded 500 at a time
Abstracts can be specified in terms of date of publication, author lists, etc
If subscriptions are available, a user can access the full text of articles
NCBI has made several utility tools available to automatically download abstracts

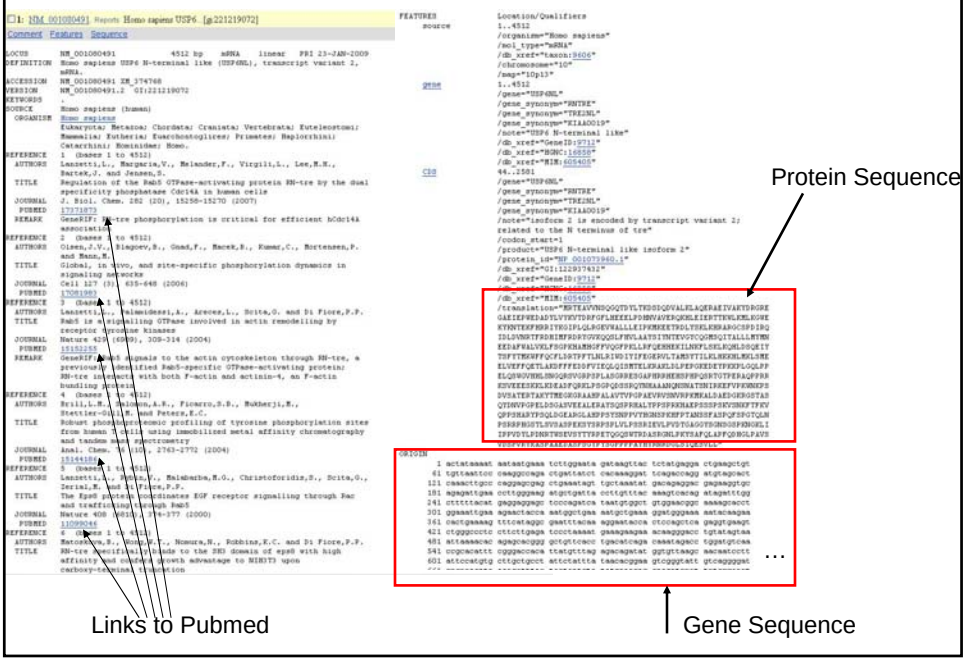
A single Abstract in Pubmed



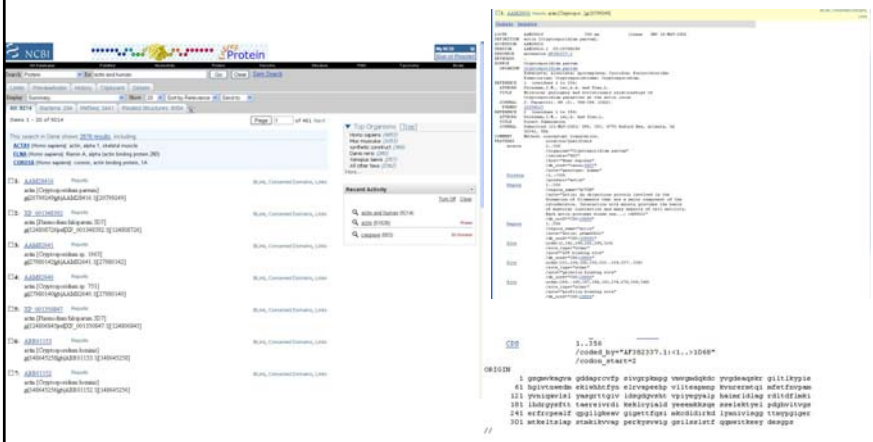
Nucleotide-Genbank's gene repository



A Nucleotide Entry in Genbank



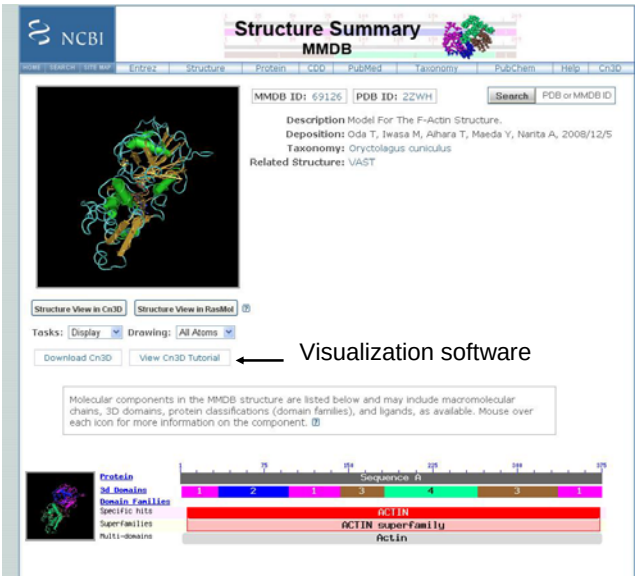
Protein Sequence in Genbank (based on the search “Actin and Human:



Note that the protein sequence and the rest of the entries are formatted similar to that of the nucleotide sequences in Genbank.
This is a database architecture issue.

Three-dimensional structure views in Genbank

Structure of Actin—Genbank Structure View

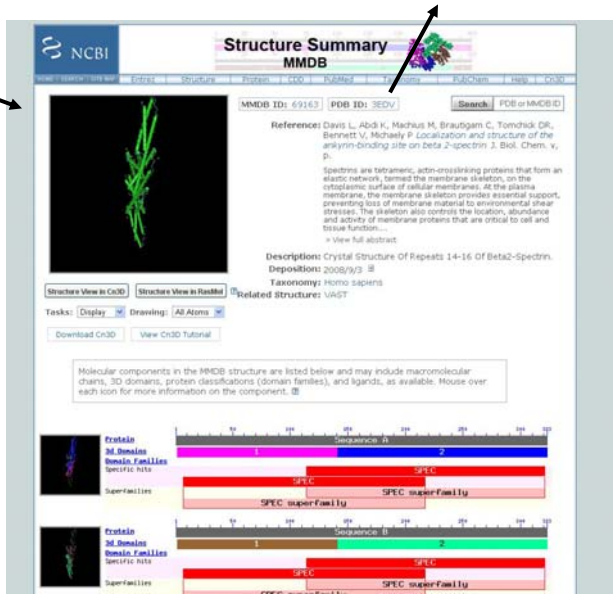


Visualization software

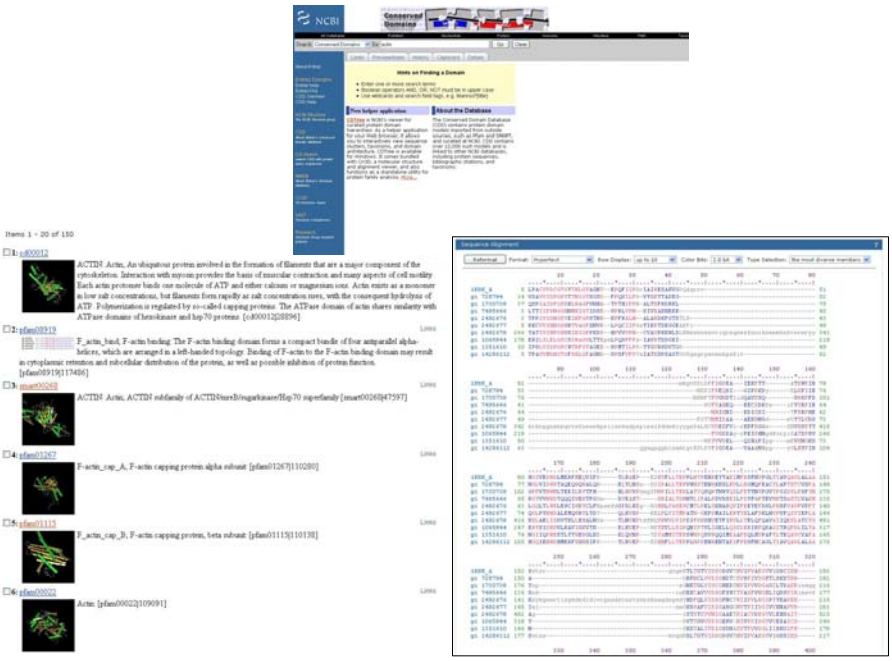
Structure of Domains in Genbank

List of domains related to or associated with Actin

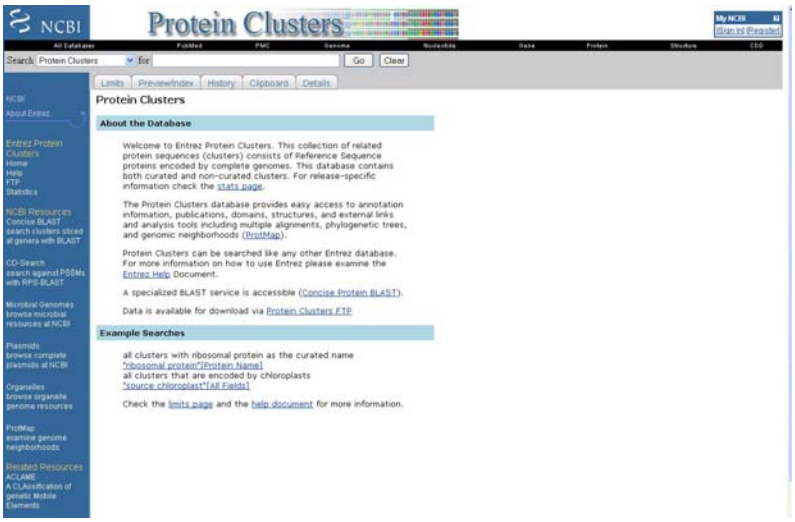
Link to Protein Databank



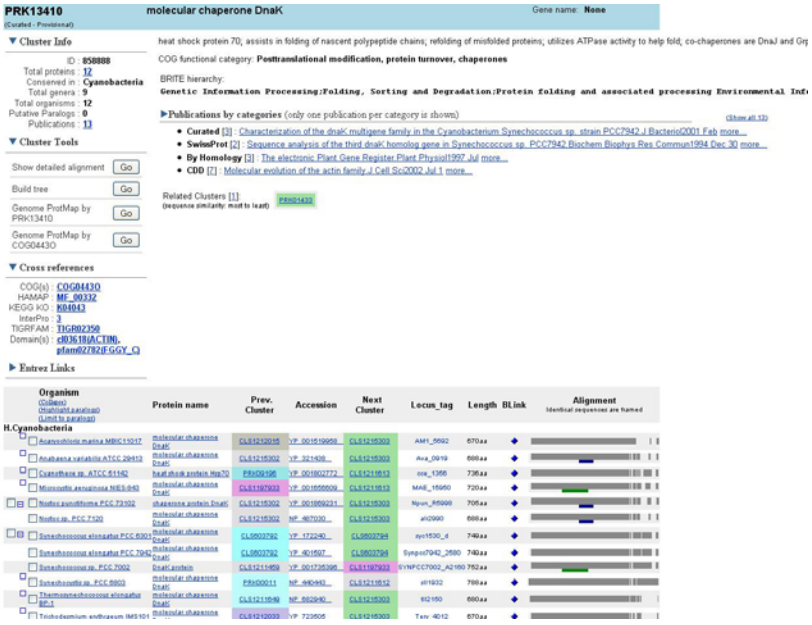
Conserved domain database (CDD) in Genbank



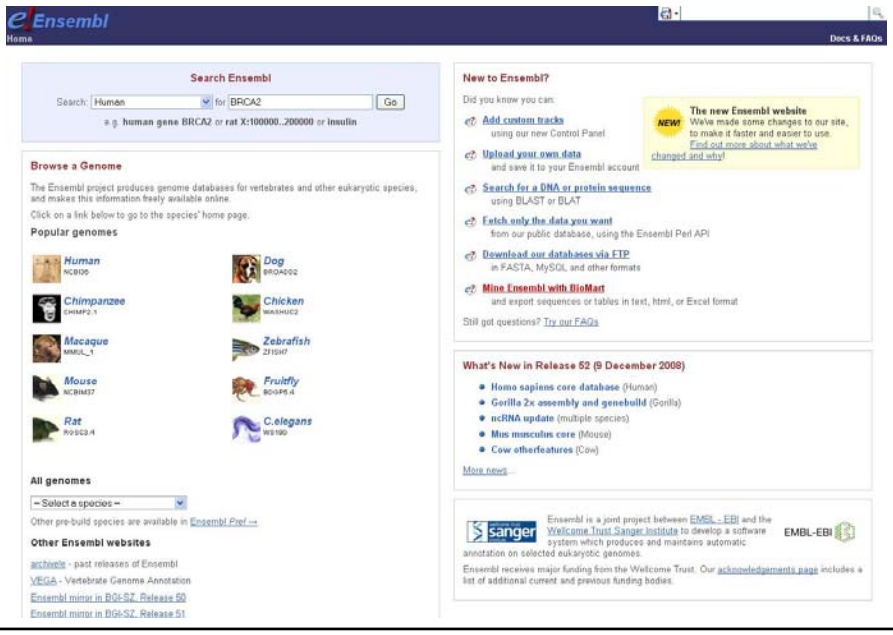
Clustering Proteins in terms of Sequence Similarities--Genbank



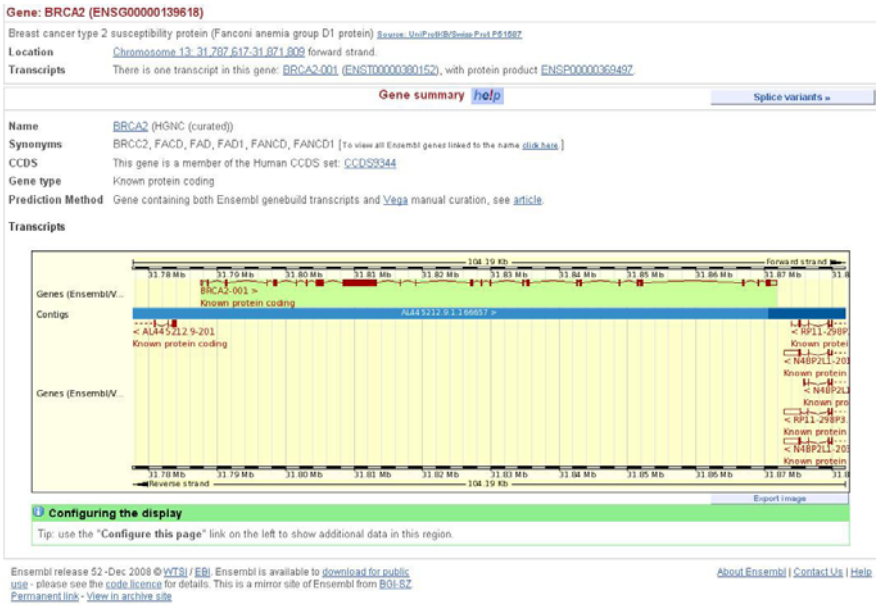
Clustering Proteins in terms of Sequence Similarities--Genbank



ENSEMBL—European version of Genbank—now focused exclusively on genome wide applications



Sample Ensembl Result—Chromosomal location and other features for downloading information



SWISSPROT--http://www.expasy.ch/

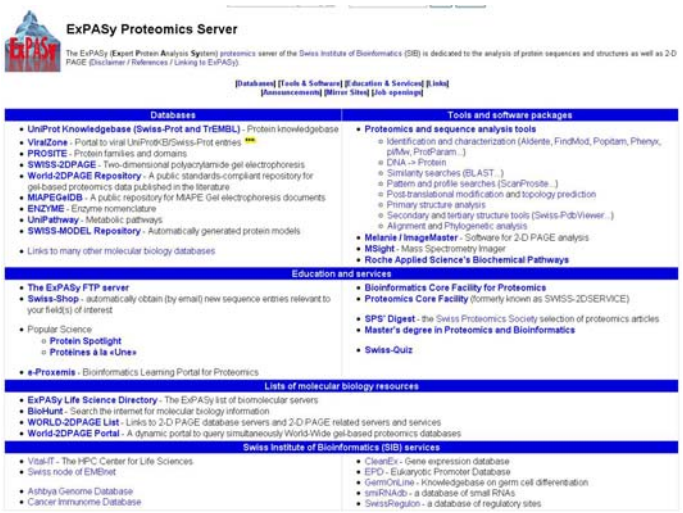
– UniProt combines SwissProt and TrEMBl

"UniProtKB/TrEMBL (unreviewed) contains protein sequences associated with computationally generated annotation and large-scale functional characterization. UniProtKB/Swiss-Prot (reviewed) is a high quality manually annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions" --http://www.uniprot.org/help/uniprotkb

In the very near future, UniProt will replace SwissProt

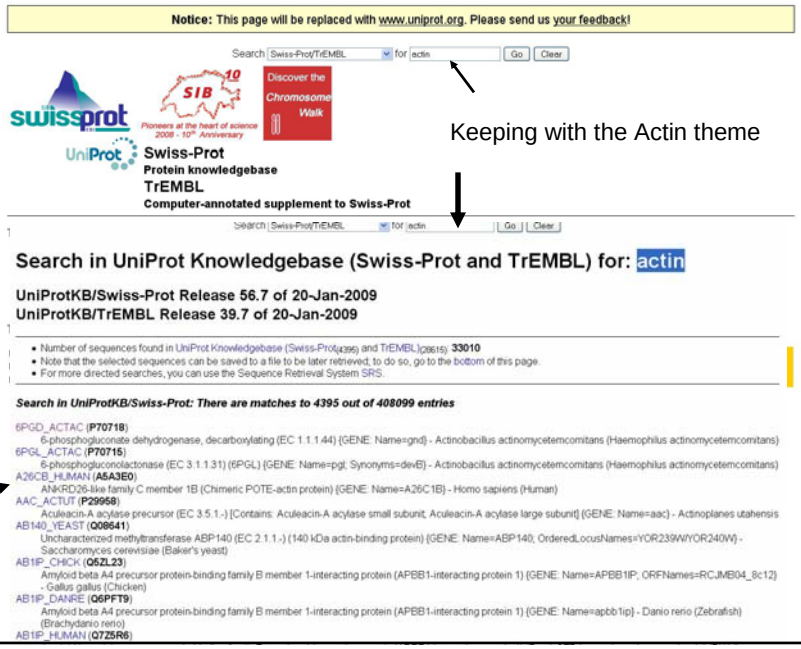
Mirro Sites
Switzerland: **http://www.expasy.org/** at Swiss Institute of Bioinformatics, Geneva
Australia: **http://au.expasy.org/** at Australian Proteome Analysis Facility, Sydney
Brazil: **http://br.expasy.org/** at Laboratório Nacional de Computação Científica, Petrópolis
Canada: **http://ca.expasy.org/** at Canadian Bioinformatics Resource, Halifax
China: **http://cn.expasy.org/** at Peking University
Korea: **http://kr.expasy.org/** at Yonsei Proteome Research Center, Seoul

SwissProt--http://www.expasy.org/

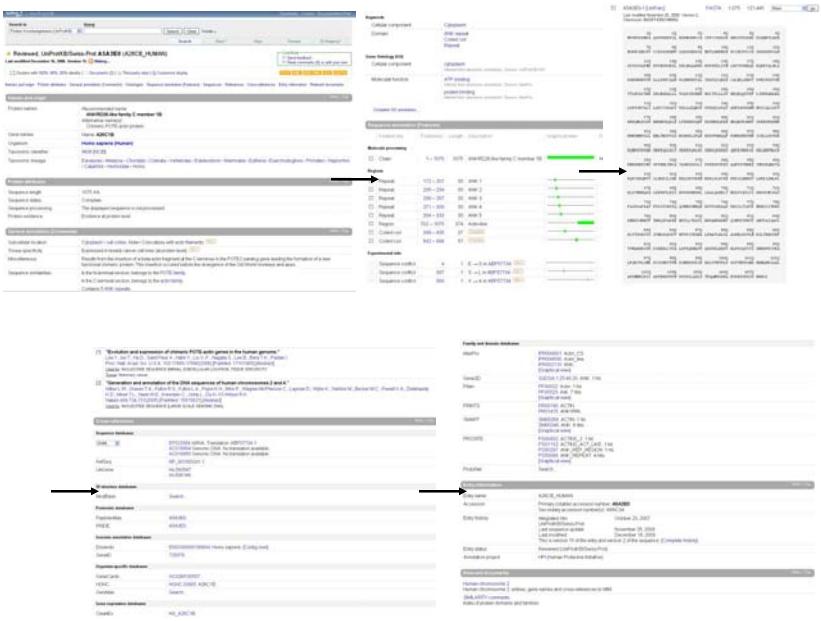


The links here are the set of tools for various proteomic analysis. Some of these are replicated in Genbank.

SwissProt—search for Proteins



Swissprot-Actin (Results in a Single Page)



Swiss 2DPage

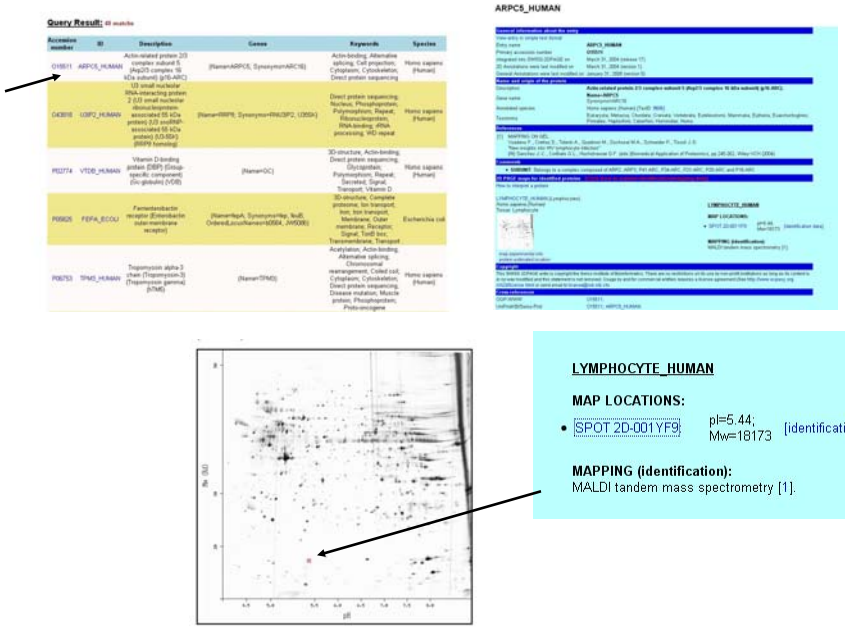
SWISS-2DPAGE

Search by description (DE), entry name (ID), gene name (GN) or UniProtKB/Swiss-Prot keywords (KW)

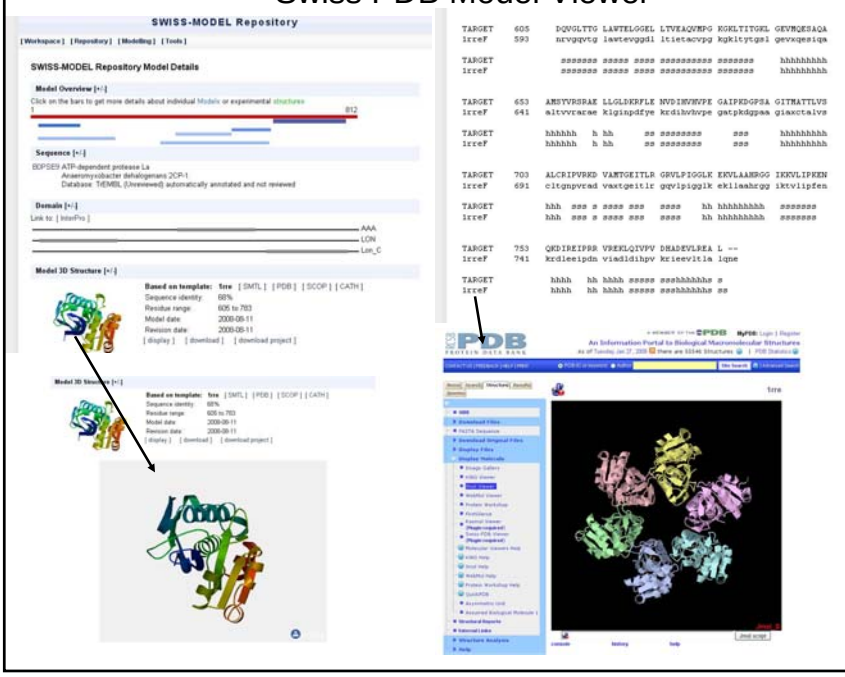
Enter search keywords: actin

Limit to: ◉ All fields ○ DE ○ ID ○ GN ○ KW

Swiss 2DPAGE --Actin

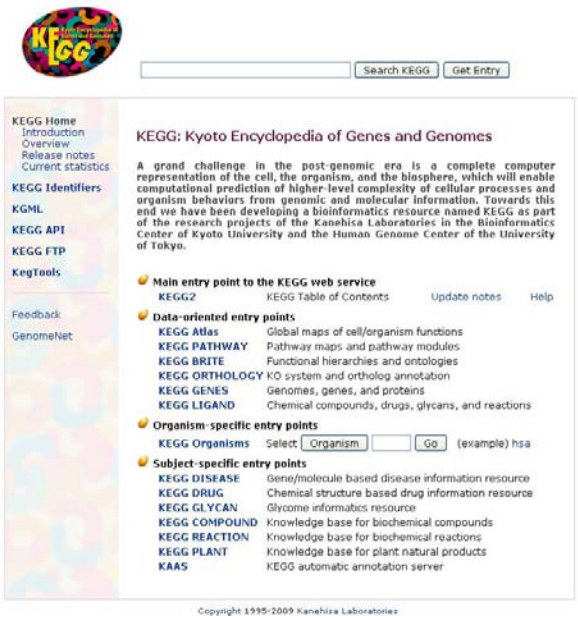


Swiss PDB Model Viewer

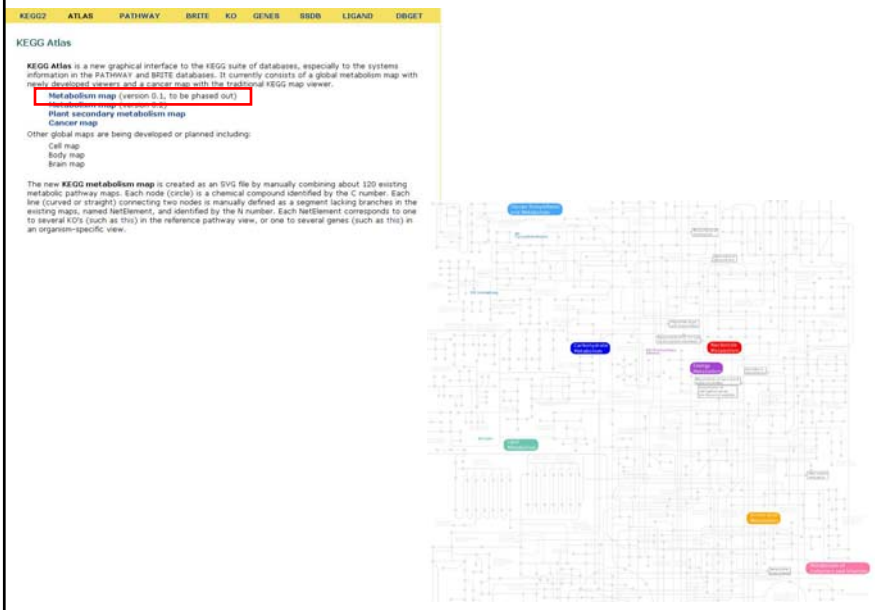


KEGG (Kyoto Encyclopedia of Genes and Genomes)

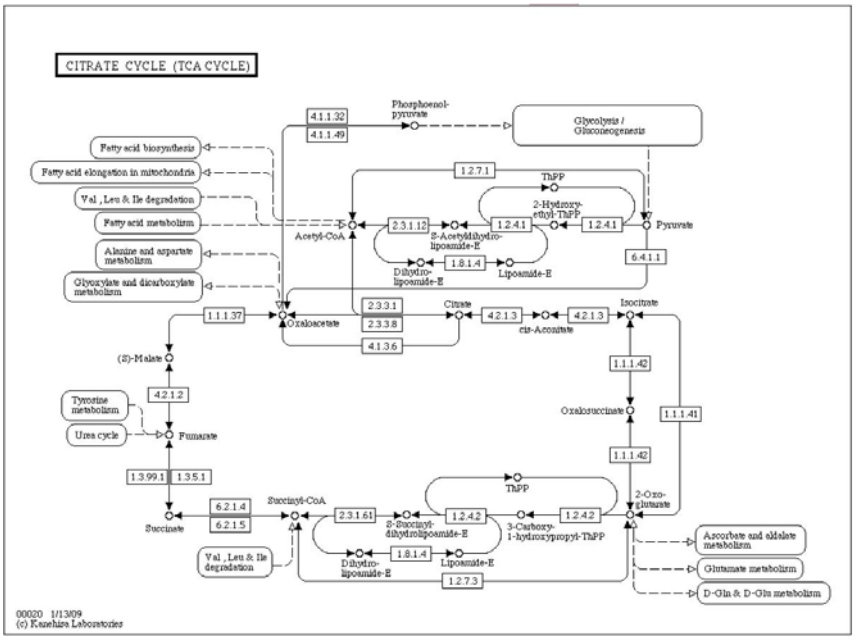
http://www.genome.jp/kegg/



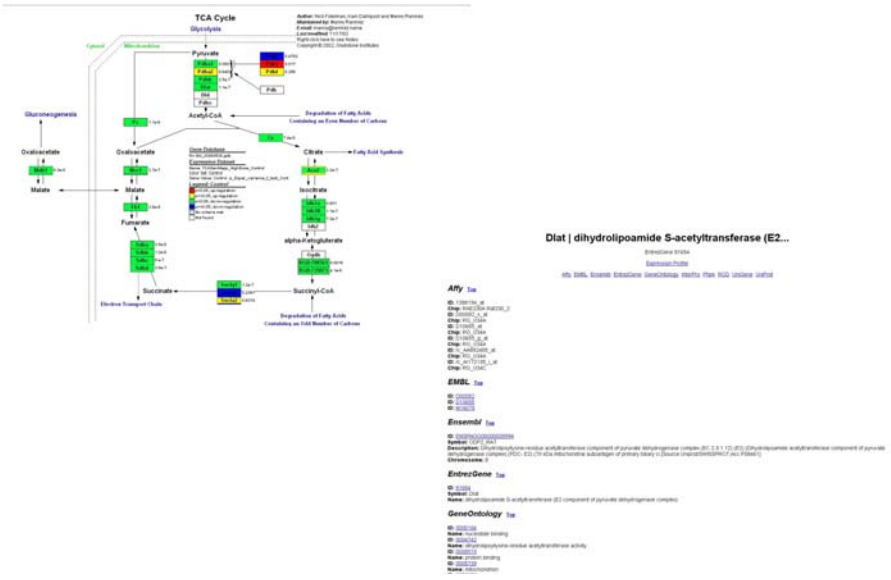
Kegg Atlas



KEGG Pathway

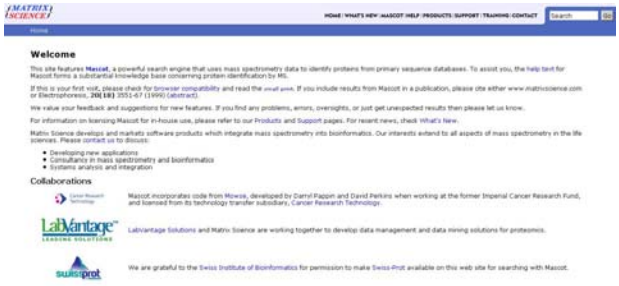


Center for Gene Nutrition Interactions in Cancer --UAB

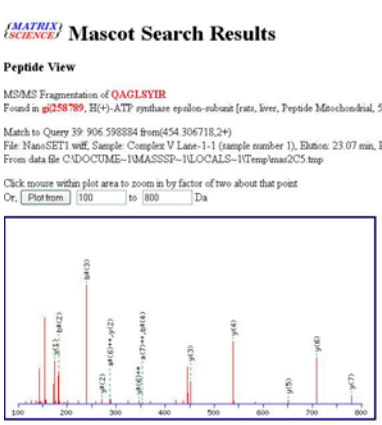


MASCOT—Protein Identification from Mass Spectroscopy Data

- Peptide Mass Fingerprinting
- Sequence Query
- MS/MS Ion Search



MASCOT Search Results



Mascot Search Results

Peptide View

MS/MS Fragmentation of QAGLSYIR
Found in gi|258789, H(+)-ATP synthase epsilon-subunit [rats, liver, Peptide Mitochondrial, 50 aa]

Match to Query 39: 906.598884 from(454.306718,2+)
File: NanoSET1.wiff, Sample: Complex V Lane-1-1 (sample number 1), Elution: 23.07 min, Period: 1, Cycle(s): 2586 (Experiment 4)
From data file C:\DOCUME~1\MASSSP~1\LOCALS~1\Temp\mas2C5.tmp

Click mouse within plot area to zoom in by factor of two about that point
Or, Plot from 100 to 800 Da

Monoisotopic mass of neutral peptide Mr(calc): 906.49
Ions Score: 47 **Expect:** 0.042
Matches (Bold Red): 14/84 fragment ions using 23 most intense peaks

#	a	a^{++}	a^*	a^{*++}	b	b^{++}	b^*	b^{*++}	Seq.	y	y^{++}	y^*	y^{*++}	#
1	101.07	51.04	84.04	42.53	129.07	65.04	112.04	56.52	Q					8
2	172.11	86.56	155.08	78.04	200.10	100.56	**183.08**	92.04	A	**779.44**	390.22	762.41	381.71	7
3	229.13	115.07	212.10	106.56	257.12	129.07	**240.10**	120.55	G	**708.40**	354.71	691.38	**346.19**	6
4	342.21	171.61	325.19	163.10	370.21	185.61	**353.18**	177.09	L	**651.38**	326.19	634.36	317.68	5
5	429.25	215.13	412.22	206.61	457.24	229.12	440.21	220.61	S	**538.30**	269.65	521.27	261.14	4
6	592.31	296.66	575.28	**288.14**	620.30	310.66	603.28	302.14	Y	**451.27**	226.14	434.24	217.62	3
7	705.39	**353.20**	688.37	344.69	733.39	367.20	716.36	358.68	I	**288.20**	144.61	**271.18**	136.09	2
8									R	**175.12**	88.06	158.09	79.55	1

Other Mass Spectrometry Analysis Software

- Phenyx--http://phenyx.vital-it.ch/pwi/login/login.jsp
- Aldente--http://www.genebio.com/aldente/index.html

Protein Data Bank-PDB

- http://www.rcsb.org/pdb/home/home.do
- **"A Resource for Studying Biological Macromolecules**

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the wwPDB, the RCSB PDB curates and annotates PDB data according to agreed upon standards.

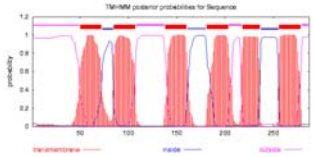
The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists."

Bioinformatics and Computational Biology as a Drive or Discovery

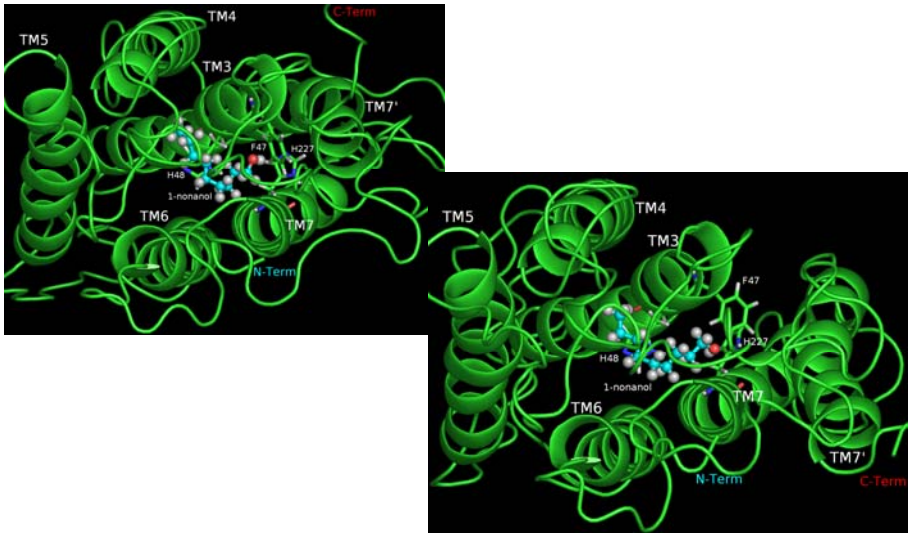
- **Recognition of the mechanism of olfaction depends upon understanding the sequence-structure-function relationships of olfactory receptors. We used bioinformatics methods to identify a new structural sub-class of olfactory receptors and GPCRs. We used available statistical methods to predict trans-membrane helical domains in olfactory receptor hOR17-210, a receptor that has been shown to be variably functional and pseudogenic in humans.**
- **TM domain identification was undertaken as a prelude to modeling this olfactory receptor in order to understand its interaction with ligands that have been experimentally shown to bind to this receptor. Our analyses revealed that there are only five typically observed TM regions in this protein with an additional orphan TM. The C-terminus is extra-cellular. This reversed polarity in the termini does not disrupt the positions of typical OR-motifs that initiate the signal transduction process at the membrane.**
- **Our observations are contrary to conventional structural knowledge about ORs and GPCRs. Preliminary sequence analysis studies have shown that such a structure is observed in a limited number of olfactory receptors distributed across different mammalian species.**
- **We also present a methodology for modeling such olfactory receptors. When combined with experimental data, we believe that this information will further our understanding of olfaction.**

- **Sequence Features for OR17-210**
- **This protein sequence for olfactory receptor OR17-210 appears as a pseudogene in the HORDE1 database**:
  - ATGATGAAGAAGAACCAAACCATGATCTCAGAGTTCCTGCTCCTGGGCCTTCCATCCAACCTGAGCAGCAGAATCTGTTCTATGCCT
    TGTTCTTGGCCGTGTATCTTACCACCCTCCTGGGGAACCTCCTCGTCATTGTCCTCATTCGACTGGACTCCCACCTCCACATGCCTA
    TGTATTTGTGTCTCAGCAACTTGTCCTTCTCTGACCTCTCTTTTCCTCGGTCACAATGCCCAAATTGCTGCAGAACATGCAGAGCCA
    AAACCCATCCATCCCCTTTGCGGACTGCCTGGCTCAGATGTACTTTCATCTGTTTTATGGAGTTCTGGAGAGCTTCCTCCTTGTGGT
    CATGGCTTATCACTGCTATGTGGCTATTTGCTTTCCTCTGCACTACACCACTATCATGAGCCCCAAGTGTTGCCTTGGTCTGCTGACA
    CTCTCCTGGCTGTTGACCACTGCCCATGCCACGTTGCACACCTTGCTTATGGCCAGGCTGTCCTTTTGTGCTGAGAATGTGATTCCT
    CACTTTTTCTGTGATACATCTACCTTGTTGAAGCTGGCCTGCTCCAACACGCAAGTCAATGGGTGGGTGATGTTTTTCATGGGCGGG
    CTCATCCTTGTCATCCCATTCCTACTCCTCATCATGTCCTGTGCAAGAATCGTCTCCACCATCCTCAGGGTCCCTTCCACTGGGGGC
    ATCCAGAAGGCTTTCTCCACCTGTGGCCCCCACCTCTCTGTGGTGTCTCTCTTCTATGGGACAATTATTGGTCTCTACTTGTGCCCA
    TTGACGAATCATAACACTGTGAAGGACACTGTCATGGCTGTGATGTACACTGGGGTGACCCACATGCTGAACCCCTTCATCTACAGC
    CTGAGGAACAGAGACATGAGGGGGAACCCTGGGCAGAGTCTTCAGCACAAAGAAAATTTTTTTGTCTTTAAAATAGTAATAGTTGGC
    ATTTTACCGTTATTGAAT
- **Intuitively Translated as**:
  - MMKKNQTMISEFLLLGL/PIQPEQQNLFYALFLAVYLTTLLGNLLVIVLIRLDSHLHMPMYLCLSNLSFSDLCFSSVTMPKLLQNMQSQNPSIP
    FADCLAQMYFHLFYGVLESFLLVVMAYHCYVAICFPLHYTTIMSPKCCLGLLTLSWLLTTAHATLHTLLMARLSFCAENVIPHFFCDTSTLLK
    LACSNTQVNGWVMFFMGGLILVIPFLLLIMSCARIVSTILRVPSTGGIQKAFSTCGPHLSVVSLFYGTIIGLYLCPLTNHNTVKDTVMAVMYT
    GVTHMLNPFIYSLRNRDMRGNPGQSLQHKENFFVFKIVIVGILPLLN
- **A two nucleotide frame shift however results in a functional protein with the following sequence**:
  - MPMYLCLSNLSFSDLCFSSVTMPKLLQNMQSQNPSIPFADCLAQMYFHLFYGVLESFLLVVMAYHCYVAICFPLHYTTIMSPKCCLGLLTL
    SWLLTTAHATLHTLLMARLSFCAENVIPHFFCDTSTLLKLACSNTQVNGWVMFFMGGLILVIPFLLLIMSCARIVSTILRVPSTGGIQKAFSTC
    GPHLSVVSLFYGTIIGLYLCPLTNHNTVKDTVMAVMYTGVTHMLNPFIYSLRNRDMRGNPGQSLQHKENFFVFKIVIVGILPLLN

- OR17-210 is an Atypical Olfactory Receptor
- OR17-210 begins with MPMY---. This sequence PMY is strongly conserved in most ORs. This sequence typically marks the beginning of the second transmembrane region. Hidden Markov Models2 have predicted that in OR17-210, this region is not a TM3. Furthermore, an HA-epitope tag experiment revealed this region of the protein to be extra-cellular.



- What is typically helix 3 in ORs is helix 1 in OR17-210. This TM has the MAYD(E)RY motif, which marks the intracellular side and (part of intracellular loop 2) of TM3 The directionality of this TM1 is extracellular to intracellular. This correctly positions the DRY region of the TM intracellulary—where structural changes following activation may be necessary for signal transduction in GPCRs4
- This allows only five typically observed in TMs in OR17-210. HMM strongly predicts that the cDNA sequence has an additional TM helix in the long C-terminus following what would be the seventh TM in most OR sequences. We call this the 7' TM. OR17-210 has a homolog in chimpanzee with greater than 95% sequence similarity. A BLAST search of the 7' sequence, "FVFKI VIVGILPLLN LVGVVKLI" does not return any matches in other ORs, GPCRs or any other protein sequence in GENBANK.
- TM 7' can then occupy either the position of the missing TM1 or TM2 in order to maintain the TM scaffold and protect the ligand and the binding pocket from the surrounding lipid layer
- If one follows the progression of N-terminus-TM1-IC1-TM2-EC1-TM3 .. etc, the C-terminus of this receptor is extra-cellular



Frame shifts, stop and start codons

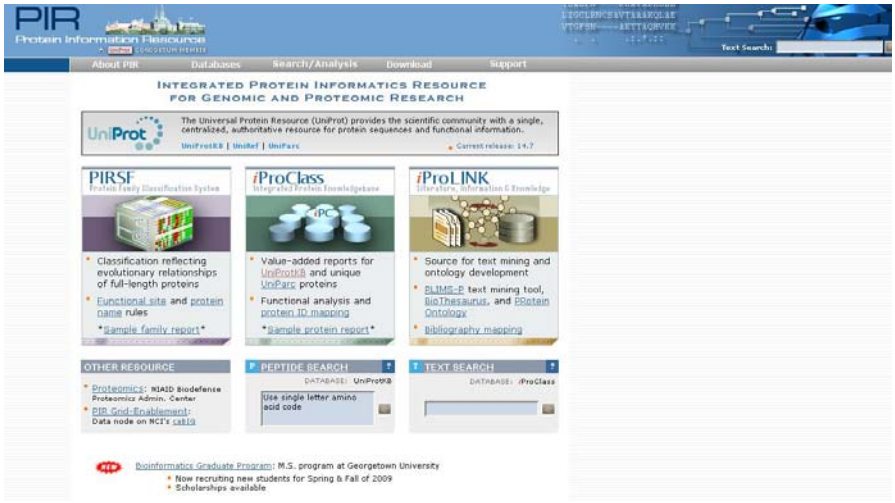
- http://ca.expasy.org/tools/dna.html
- http://bioportal.weizmann.ac.il/HORDE/search_horde.html
- http://bip.weizmann.ac.il/cgi-bin/HORDE/showGene.pl?key=symbol&value=OR1E3P
- TMHMM-- http://www.cbs.dtu.dk/services/TMHMM/

Problems during Protein Identification

- No sequence in database --- nothing to correlate with
- Problems with entries in database: human errors in entering information (typographical errors and curation); sequencing errors; errors during transcription
- Modifications in large proteins: degradation, oxidation of methionine, deamidation of N and Q, remember glycosylations, phosphorylations, and acetylations ….

  http://www.unimod.org/ lists the possible modifications that can occur

Sources for Protein Identification

- EBI, European Bioinformatics Institute, EBI Download site "The EBI is a centre for research and services in bioinformatics. The Institute manages databases of biological data including nucleic acid, protein sequences and macromolecular structures."
  - Expasy "The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE" Databases, Tools and Software Packages.
  - INFOBIOGEN, DBCAT, The Public Catalog of Databases
  - NCBI, The National Center for Biotechnology Information "provides an integrated approach to the use of gene and protein sequence information" Databases and Tools
  - PIR Protein Information Resource "An integrated public resource of protein informatics to support genomic and proteomic research and scientific discovery." Located at Georgetown University.
  - Plant Genome Database (PlantGDB) Resource for Plant Comparative Genomics
  -

Protein Information Resource--
http://pir.georgetown.edu/