Web-based Bioinformatics Applications in Proteomics

Chiquito Crasto

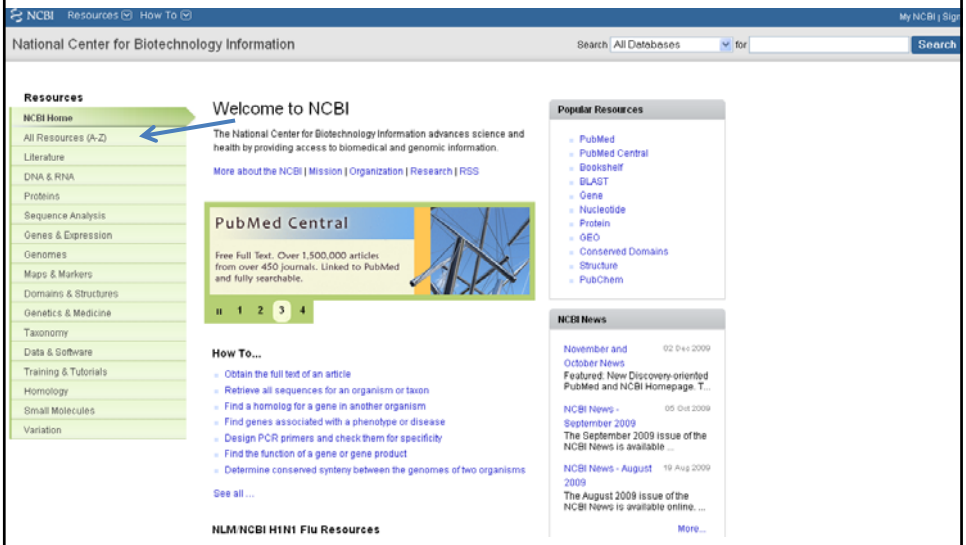
ccrasto@genetics.uab.edu

February 9, 2010

Genbank

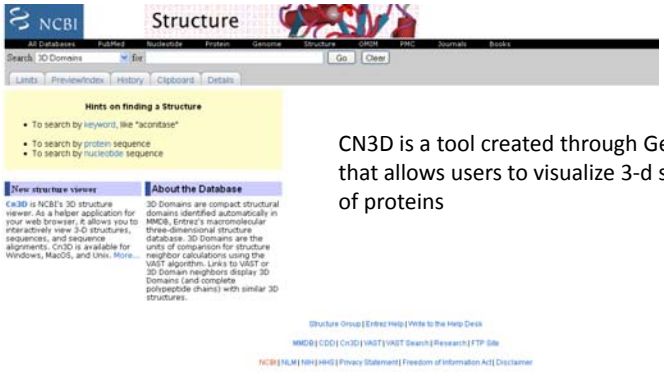
- Primary nucleic acid sequence database
- Maintained by NCBI
  - National Center for Biotechnology Information
  - http://www.ncbi.nlm.nih.gov
- As of August, 2009;
  - 106,533,156,756
    - 101,467,270,308 (Early 2009)
  - 148,165,117,763 (Whole Genome Shotgun Sequences)
    - 101,815,678 sequences (Early 2009)

Genbank …



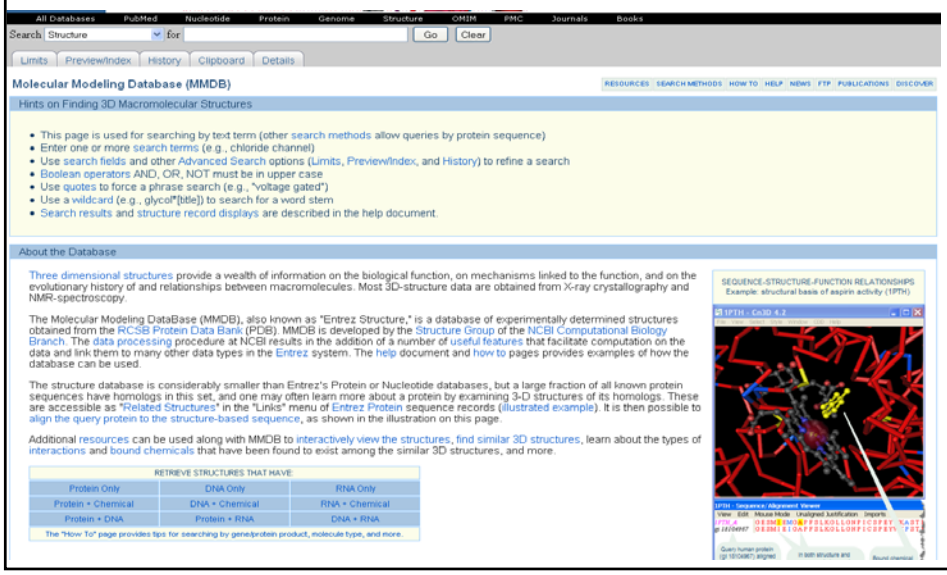
3D domain database

- 3d Domain Database

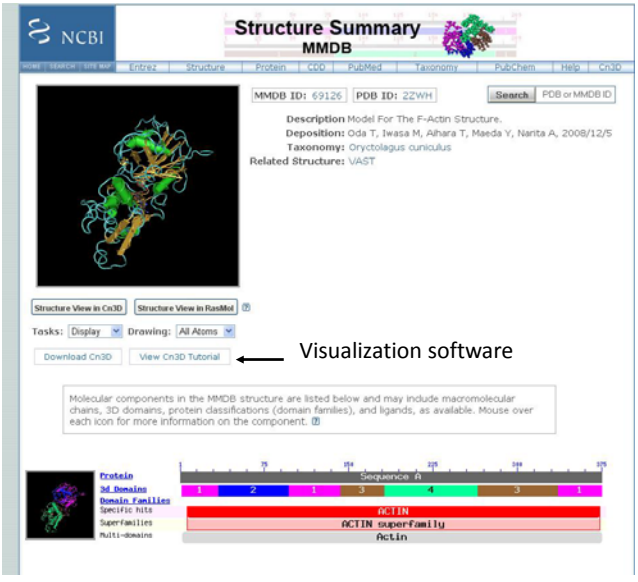


CN3D is a tool created through Genbank that allows users to visualize 3-d structures of proteins

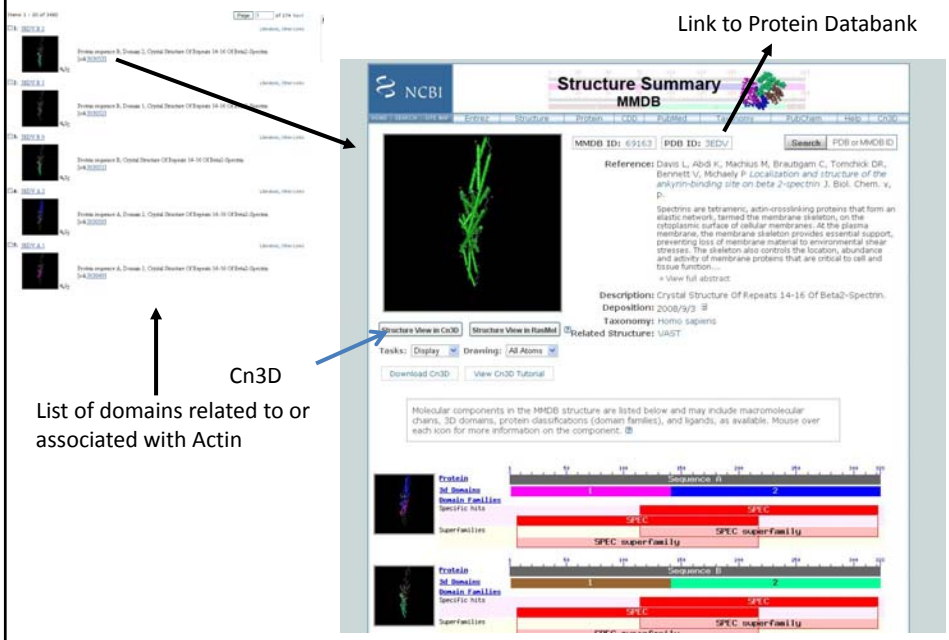
MMDB (Structures from PDB)



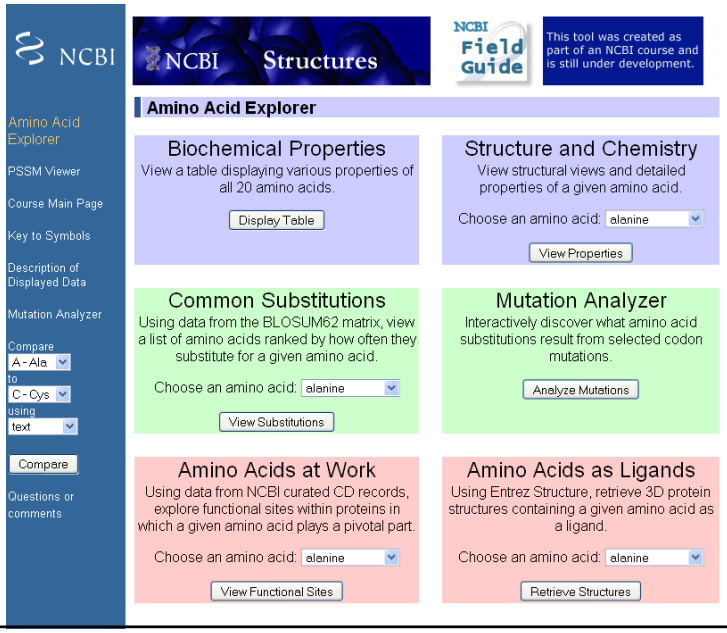
Structure of Actin—Genbank Structure View



Structure of Domains in Genbank

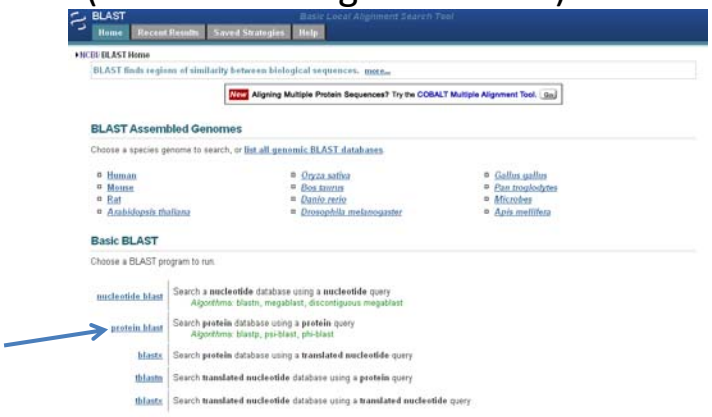


Genbank: Amino Acid Explorer



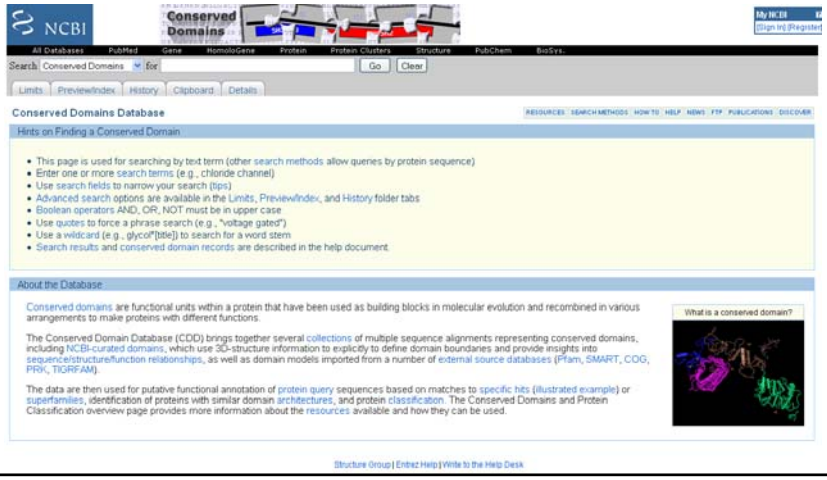
Additional tools and resources

- Batch Protein– Allows users to upload protein information in batches (saves time)
- BLAST (Basic Local Alignment Tool)

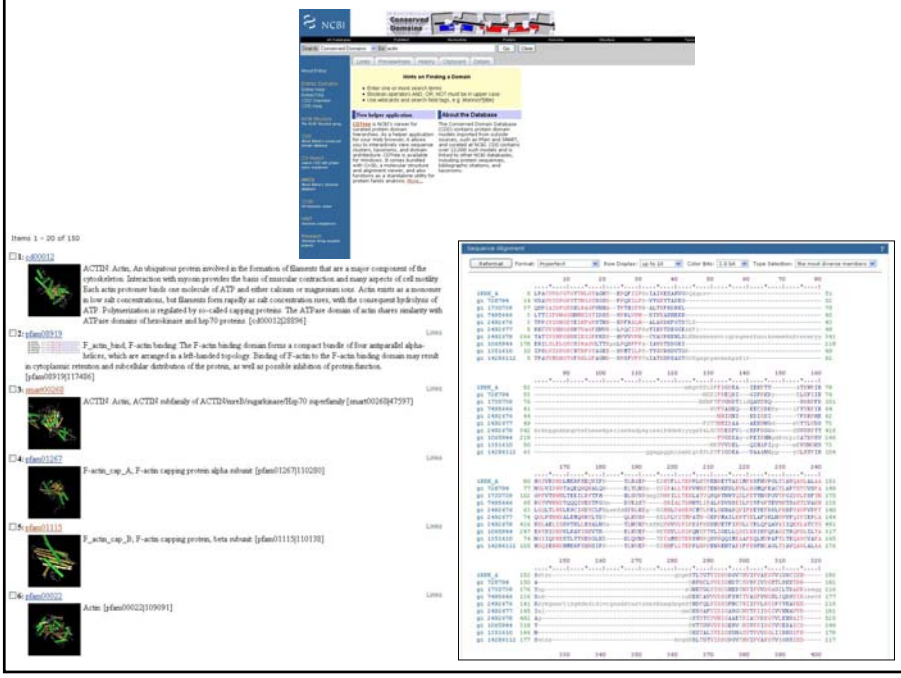


Conserved Domains

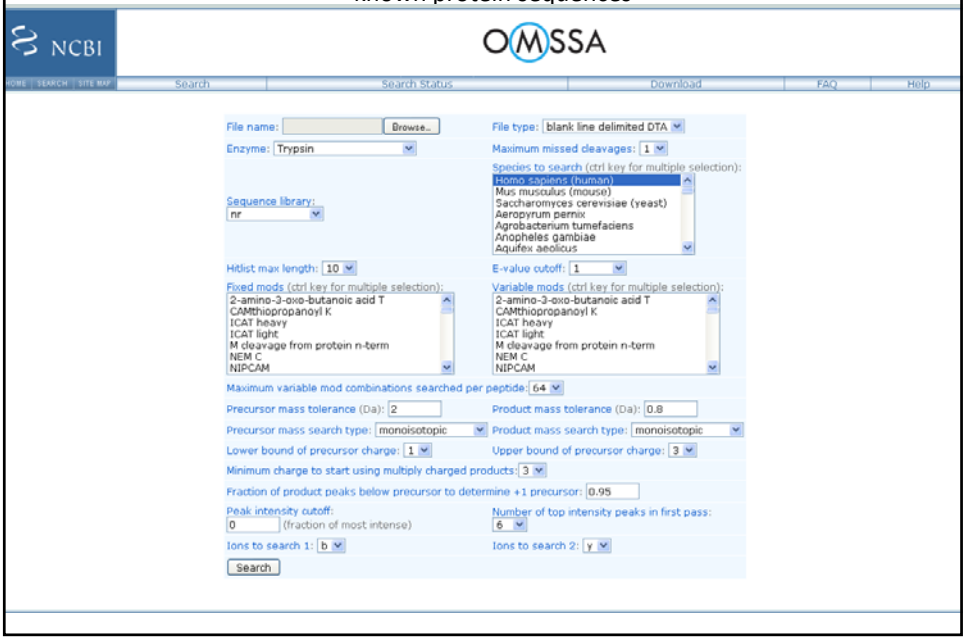
- CDART (Conserved Domain Architecture Retrieval Tool)
- CDD (Conserved Domain Database)



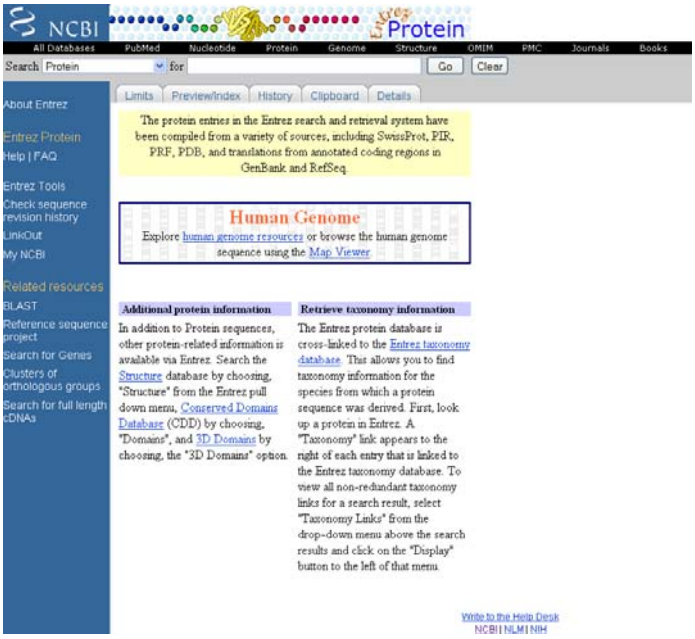
Conserved domain database (CDD) in Genbank



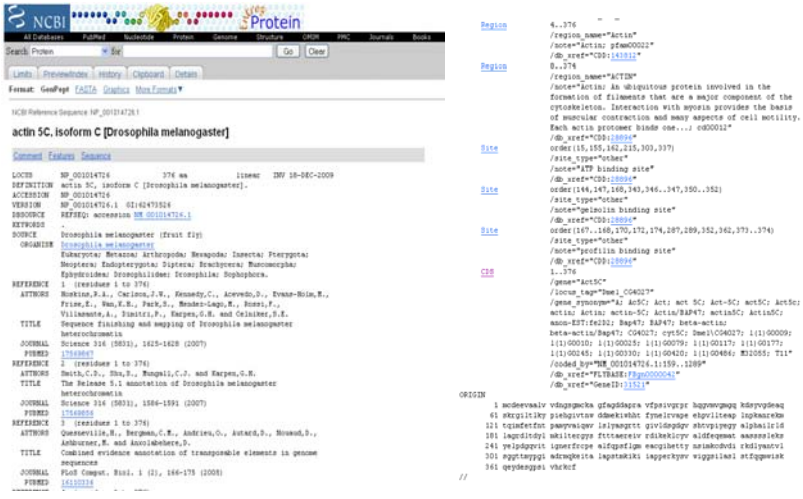
OMSSA—search engine that identifies ms/ms spectra by searching libraries of known protein sequences



Protein—Genbank's Protein Search System



Genbank resource … Protein

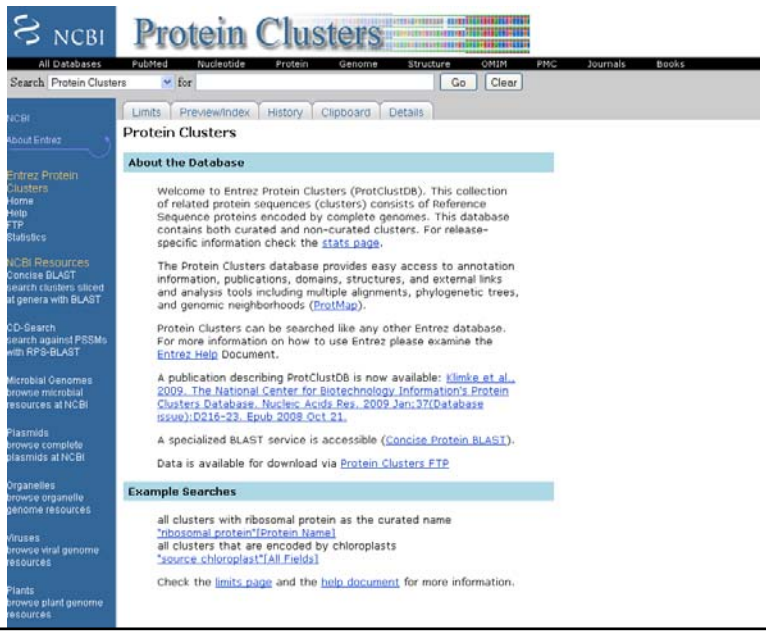


Protein …

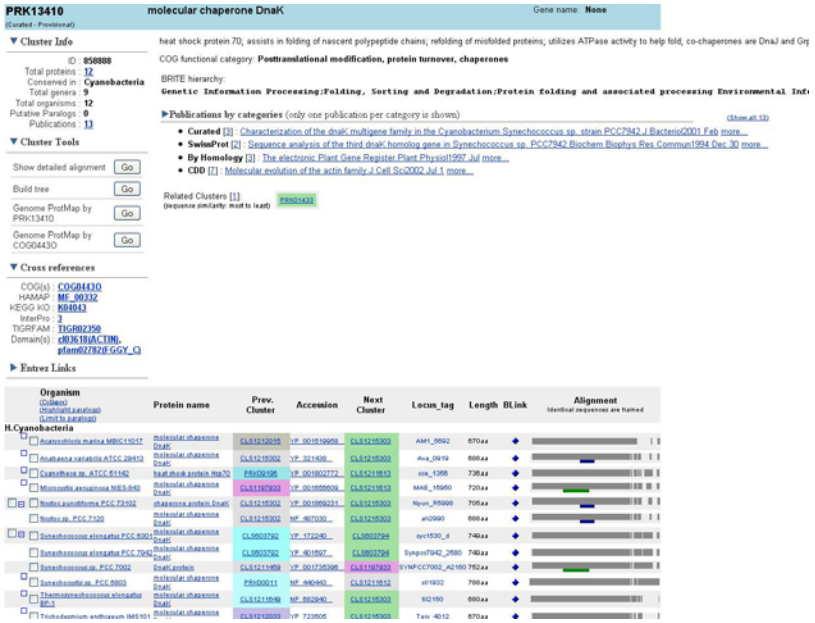
- The sequence can be visualized in different formats
  - FASTA—important to know because most software asks that you input information in the FASTA format
  - >gi|71031658|ref|XP_765471.1| actin [Theileria parva strain Muguga]

    MSDEETTALVVDNGSGNVKAGFAGDDAPRCVFPSIVGRPKNPALMVGMDEKDTYVGDEAQSKRGILTLKY
    PIEHGIVTNWEDMEKIWHHTFYNELRIAPEEHPVLLTEAPMNPKANREKMTTIMFETHNVPAMYVAIQAV
    LSLYSSGRTTGIVLDSGDGVTHTVPIYEGYALPHAIMRLDLAGRDLTEFMQKILVERGFSFTTTAEKEIV
    RDIKEKLCYIALDFDEEMTTSSSSSEVEKSYELPDGNIITVGNERFRCPEVLFQPTFIGMEAPGIHTTTY
    NSIVRCDVDIRKDLYANVVLSGGTTMFEGIGQRMTKELNALVPSTMKIKVVAPPERKYSVWIGGSILSSL
    STFQQMWITKEEFDESGPNIVHRKCF

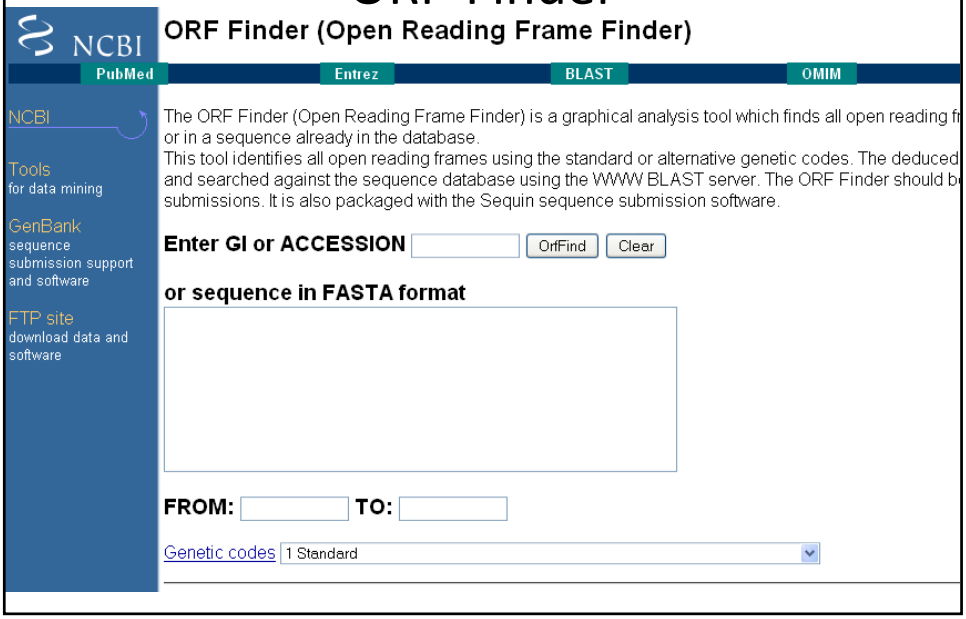
Protein Clusters

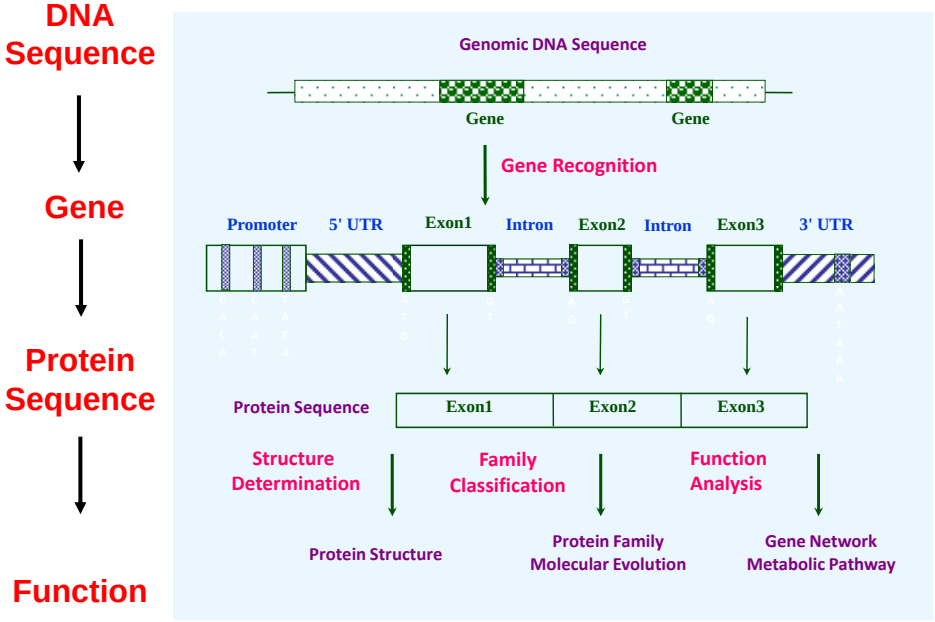


Clustering Proteins in terms of Sequence Similarities--Genbank

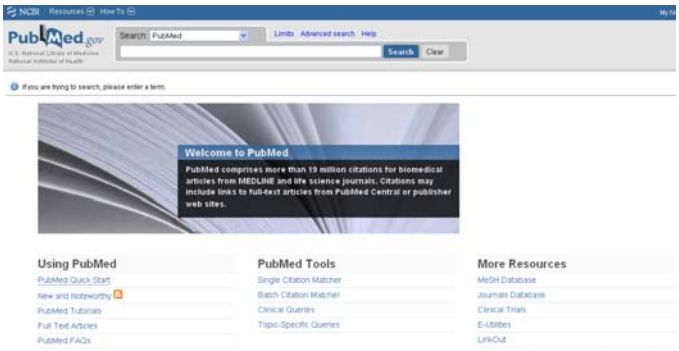


ORF-Finder



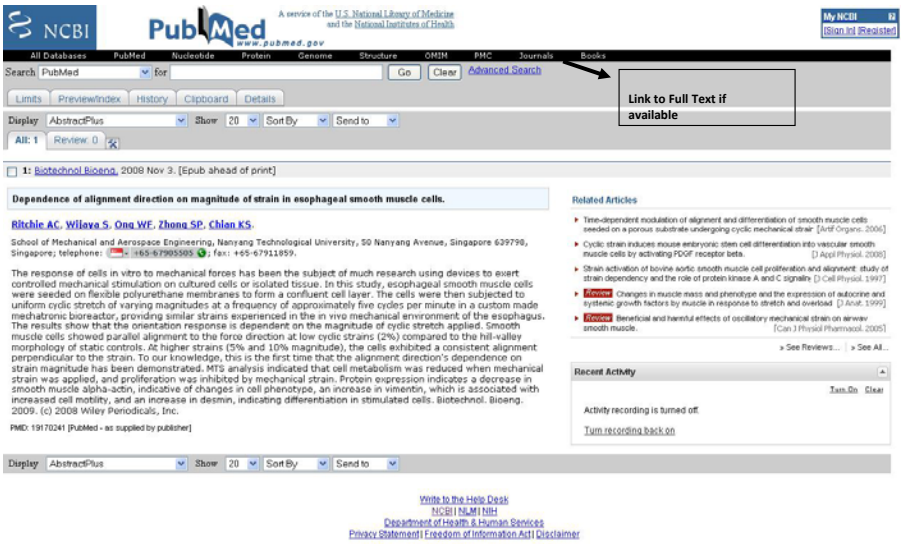


Pubmed—repository of biomedical abstracts

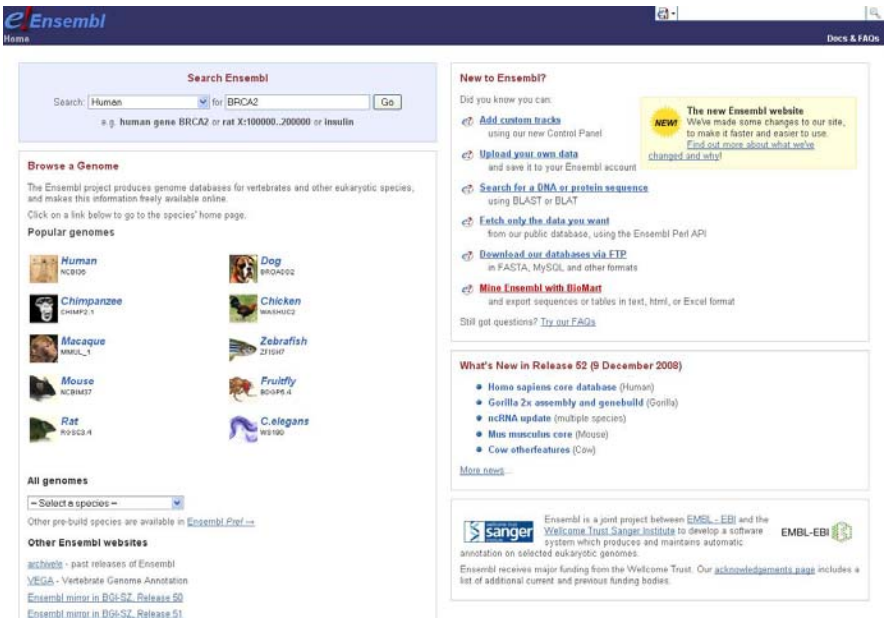


Information in Pubmed is available in several formats.
Abstracts can be downloaded 500 at a time
Abstracts can be specified in terms of date of publication, author lists, etc
If subscriptions are available, a user can access the full text of articles
NCBI has made several utility tools available to automatically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

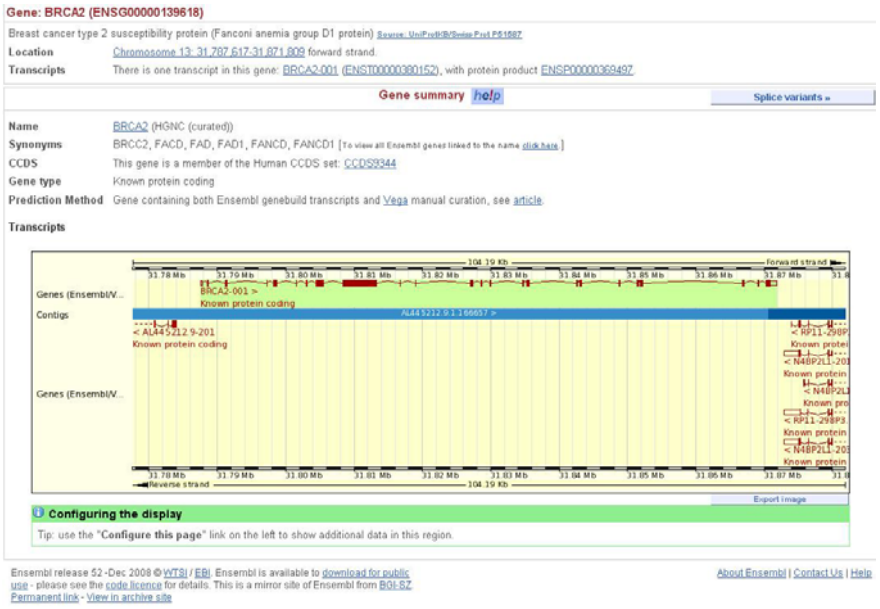
A single Abstract in Pubmed



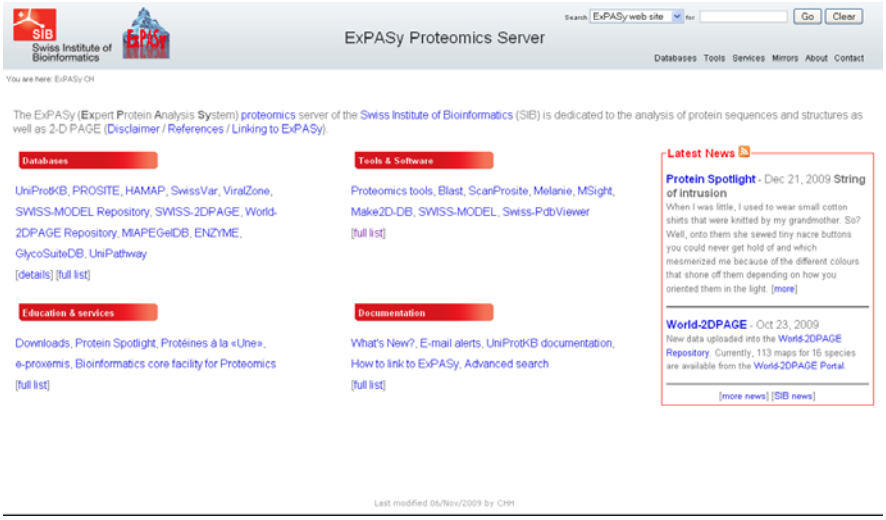
ENSEMBL—European version of Genbank—now focused exclusively on genome wide applications



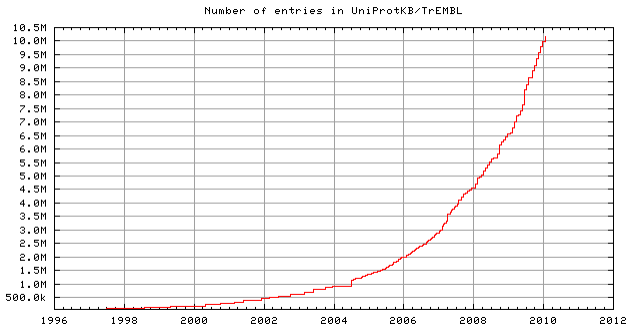
Sample Ensembl Result—Chromosomal location and other features for downloading information



EXPASY Proteomics Server (SwissProt) http://www.expasy.ch/



Uniprot (Swissprot and TREMBL)

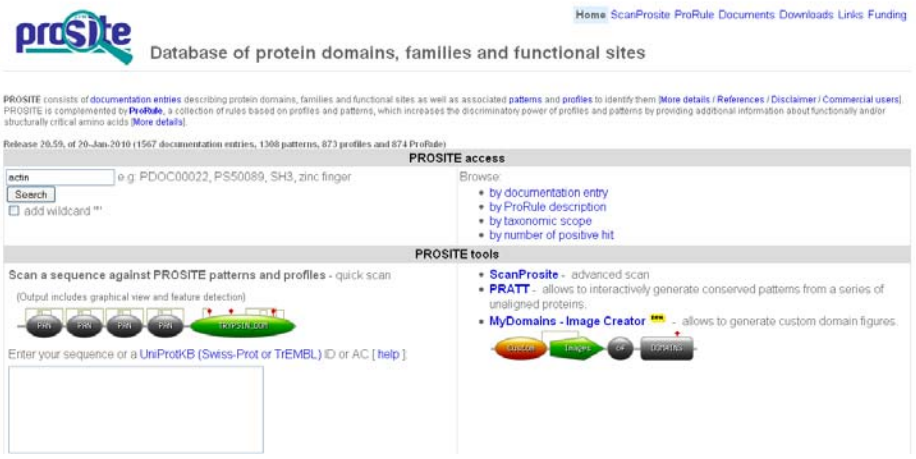


Contains (to date) more than ten million protein sequences

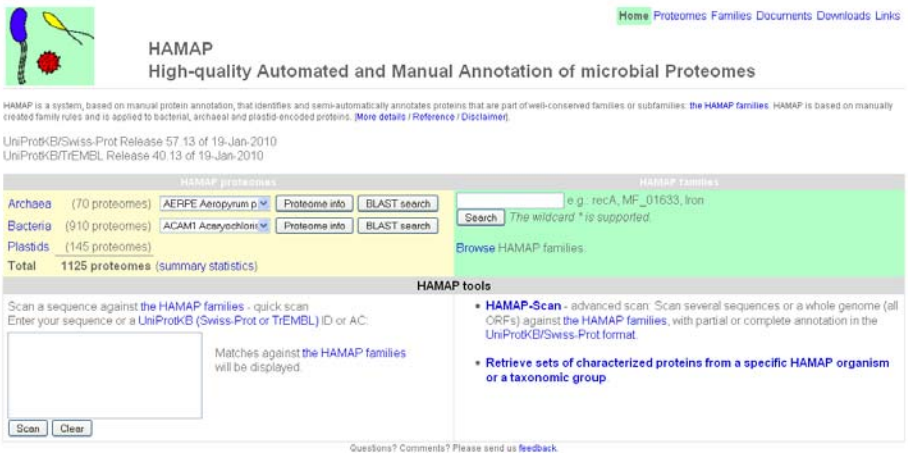
Courtesy: http://www.ebi.ac.uk/uniprot/TrEMBLstats/

UniProt

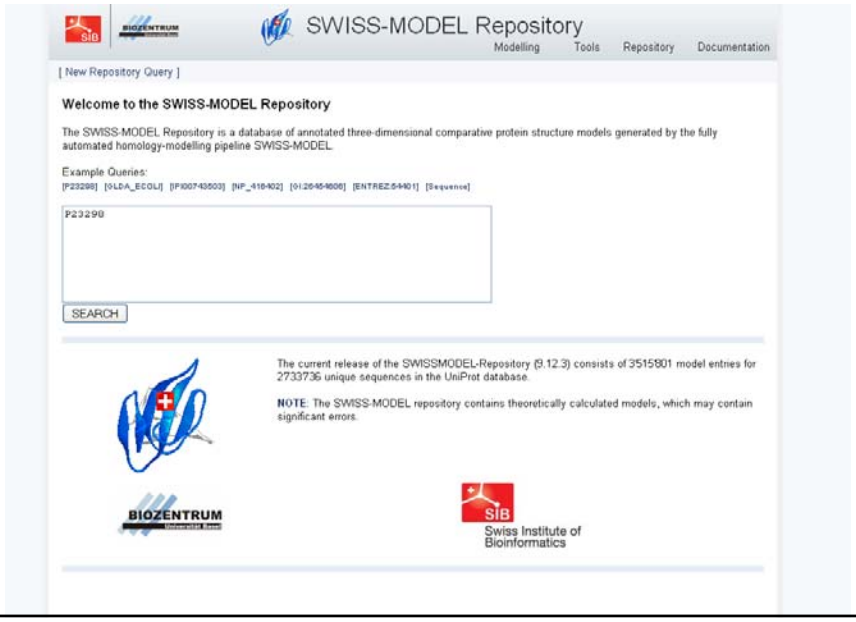
PROSITE—families, patterns, profiles and functional sites



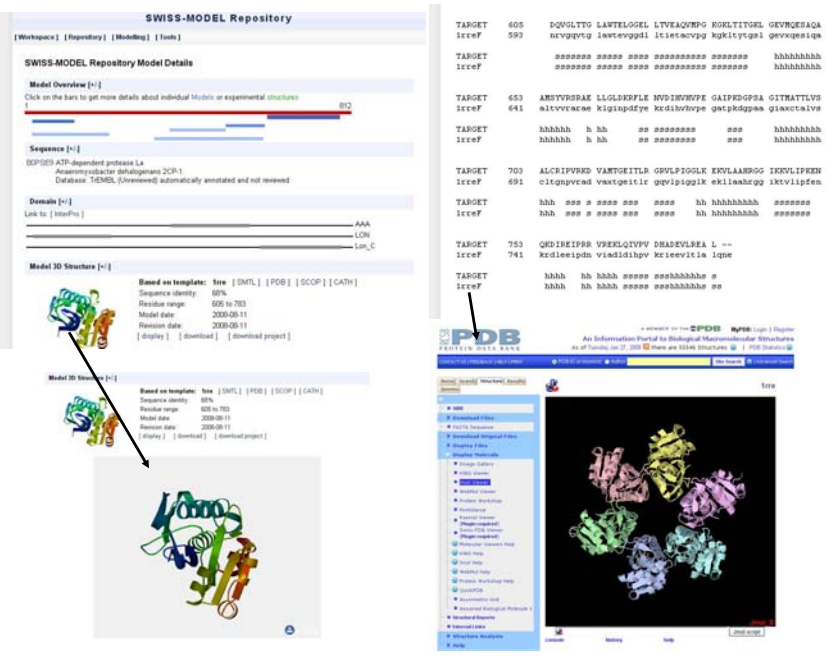
HAMAP--High-quality Automated and Manual Annotation of microbial Proteomes



Swiss Model Repository

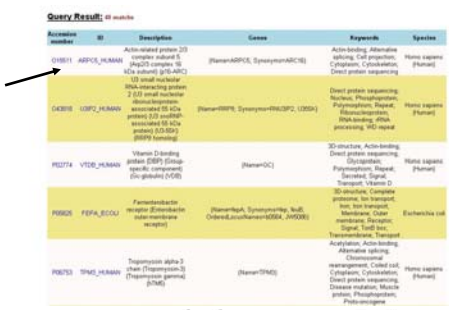


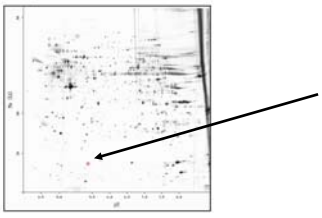
Swiss PDB Model Viewer



Swiss 2D-PAGE

Swiss 2DPAGE --Actin





LYMPHOCYTE_HUMAN

MAP LOCATIONS:

- SPOT 2D-001YF9: pI=5.44; Mw=18173 [identification data]

MAPPING (identification):
MALDI tandem mass spectrometry [1].

World 2D Page portal allows users all over the world to upload their gels to EXPASY MIAPEGelDB, allows users to document their electrophoresis experiments in such a way that it can be published, stored and retrieved, when necessary

EXPASY .. Other resources

- ENZYME—allows users to search for enzymes based on nomenclature, function, etc.
- Proteomics Tools (http://www.expasy.ch/tools/#proteome) -- several hundred resources that perform various functions in identifying, characterizing, translating sequences, processing MS data, prediction, etc.

EXPASY –MS tools

- Popitam - Identification and characterization tool for peptides with unexpected modifications (e.g. post-translational modifications or mutations) by tandem mass spectrometry
- Phenyx - Protein and peptide identification/characterization from MS/MS data from GeneBio, Switzerland
- Mascot - Sequence query and MS/MS ion search from Matrix Science Ltd., London
- OMSSA - MS/MS peptide spectra identification by searching libraries of known protein sequences
- PepFrag - Search known protein sequences with peptide fragment mass information from Rockefeller and NY Universities
- ProteinProspector - UCSF tools for fragment-ion masses data (MS-Tag, MS-Seq, MS-Product, etc.)
- xQuest - Search machine to identify cross-linked peptides from complex samples and large protein sequence databases

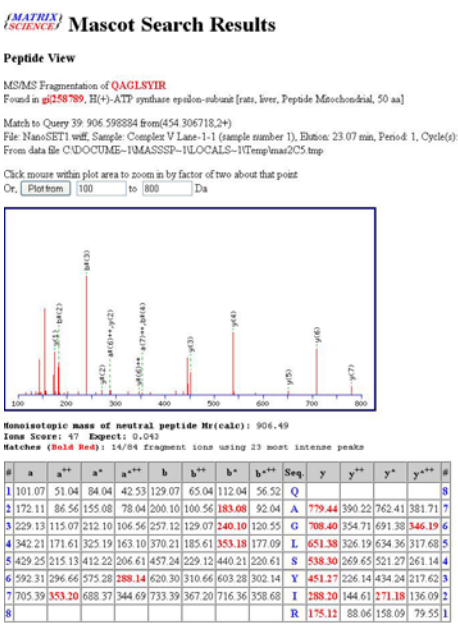
EXPASY—Visualization tools for MS data

- HCD/CID spectra merger - a tool to merge the peptide sequence-ion m/z range from CID spectra and the reporter-ion m/z range from HCD spectra into the appropriate single file, to be further used in identification and quantification search engines
- MALDIPepQuant - Quantify MALDI peptides (SILAC) from Phenyx output
- MSight - Mass Spectrometry Imager
- pIcarver - Visualize theoretical distributions of peptide pI on a given pH range and generate fractions with similar peptide frequencies

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: 0.043
Matches (Bold Red): 14/84 fragment ions using 23 most intense peaks

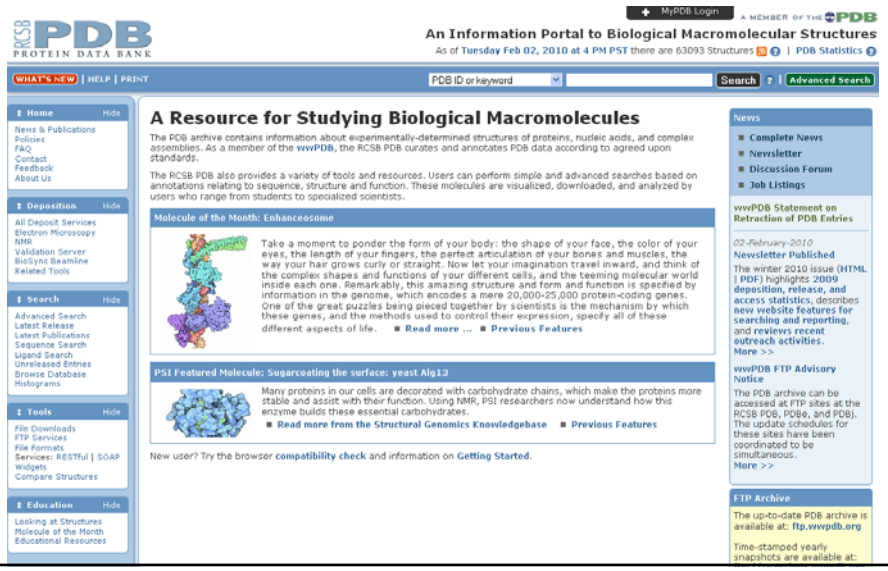
#	a	a^{++}	a*	a*$^{++}$	b	b^{++}	b*	b*$^{++}$	Seq.	y	y^{++}	y*	y*$^{++}$	#
1	101.07	51.04	84.04	42.53	129.07	65.04	112.04	56.52	Q					8
2	172.11	86.56	155.08	78.04	200.10	100.56	**183.08**	92.04	A	**779.44**	390.22	762.41	381.71	7
3	229.13	115.07	212.10	106.56	257.12	129.07	**240.10**	120.55	G	**708.40**	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

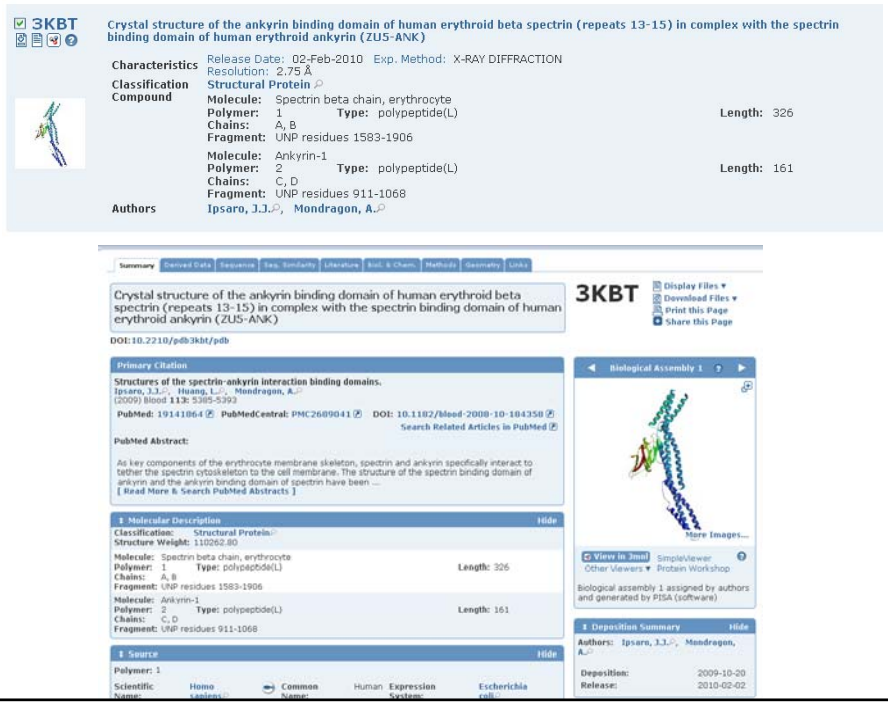
Problems during Protein Identification from Mass Spec Data

- 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

Protein Data Bank—repository of experimentally and computationall obntained structures of proteins, protein-DNA and DNA (http://www.rcsb.org/pdb/home/home.do)

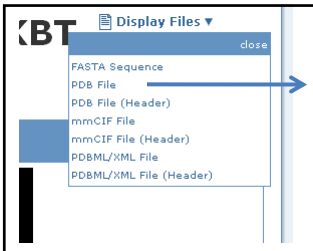




Tip: right-mouse click on Jmol to get access to additional Jmol functionality.

This view is modifiable via the web—requires Java

Other very useful visualization and more, tools: Rasmol, ArgusLab, Pymol and Chimera

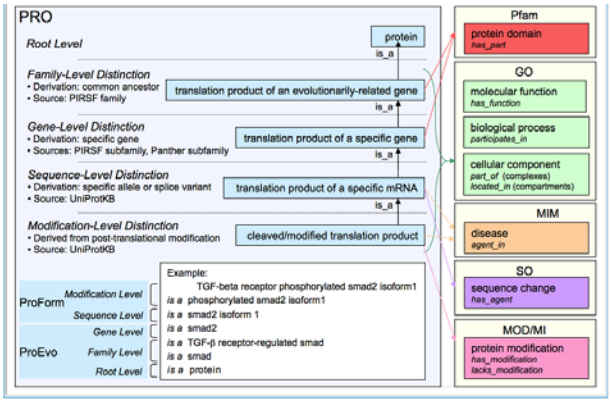



```
ATOM      1  N   GLU A1585      30.618 -42.928  39.826  1.00 41.08           N
ATOM      2  CA  GLU A1585      31.123 -41.839  38.894  1.00 41.42           C
ATOM      3  C   GLU A1585      30.409 -41.714  37.524  1.00 40.78           C
ATOM      4  O   GLU A1585      30.134 -40.597  37.079  1.00 40.82           O
ATOM      5  CB  GLU A1585      32.655 -41.841  38.751  1.00 41.48           C
ATOM      6  CG  GLU A1585      33.385 -41.150  39.936  1.00 44.28           C
ATOM      7  CD  GLU A1585      33.424 -41.988  41.250  1.00 46.48           C
ATOM      8  OE1 GLU A1585      34.095 -43.046  41.282  1.00 46.53           O
ATOM      9  OE2 GLU A1585      32.810 -41.570  42.263  1.00 47.68           O
ATOM     10  N   ALA A1586      30.143 -42.826  36.840  1.00 39.99           N
ATOM     11  CA  ALA A1586      29.307 -42.757  35.641  1.00 39.58           C
ATOM     12  C   ALA A1586      27.876 -42.482  36.144  1.00 39.76           C
ATOM     13  O   ALA A1586      27.135 -41.647  35.582  1.00 39.71           O
ATOM     14  CB  ALA A1586      29.364 -44.027  34.857  1.00 39.22           C
ATOM     15  N   GLN A1587      27.515 -43.182  37.227  1.00 39.61           N
ATOM     16  CA  GLN A1587      26.236 -43.011  37.944  1.00 39.37           C
ATOM     17  C   GLN A1587      26.050 -41.522  38.260  1.00 39.07           C
ATOM     18  O   GLN A1587      24.940 -40.993  38.118  1.00 39.54           O
ATOM     19  CB  GLN A1587      26.193 -43.858  39.236  1.00 39.25           C
ATOM     20  CG  GLN A1587      27.320 -43.533  40.252  1.00 40.14           C
ATOM     21  CD  GLN A1587      27.905 -44.762  40.970  1.00 40.47           C
ATOM     22  OE1 GLN A1587      29.041 -45.188  40.696  1.00 39.31           O
ATOM     23  NE2 GLN A1587      27.135 -45.321  41.896  1.00 39.23           N
ATOM     24  N   GLN A1588      27.137 -40.859  38.674  1.00 38.16           N
ATOM     25  CA  GLN A1588      27.125 -39.432  38.989  1.00 37.55           C
ATOM     26  C   GLN A1588      26.968 -38.577  37.725  1.00 37.67           C
ATOM     27  O   GLN A1588      26.334 -37.489  37.757  1.00 37.34           O
ATOM     28  CB  GLN A1588      28.376 -39.029  39.773  1.00 37.29           C
ATOM     29  CG  GLN A1588      28.381 -37.591  40.319  1.00 36.68           C
ATOM     30  CD  GLN A1588      27.139 -37.221  41.130  1.00 36.39           C
ATOM     31  OE1 GLN A1588      26.691 -37.974  41.982  1.00 36.65           O
ATOM     32  NE2 GLN A1588      26.593 -36.044  40.868  1.00 36.30           N
ATOM     33  N   TYR A1589      27.565 -39.046  36.623  1.00 37.07           N
ATOM     34  CA  TYR A1589      27.435 -38.336  35.367  1.00 36.28           C
ATOM     35  C   TYR A1589      25.932 -38.269  34.974  1.00 36.05           C
ATOM     36  O   TYR A1589      25.420 -37.204  34.667  1.00 36.04           O
ATOM     37  CB  TYR A1589      28.320 -38.941  34.259  1.00 36.23           C
ATOM     38  CG  TYR A1589      28.058 -38.308  32.916  1.00 35.18           C
```

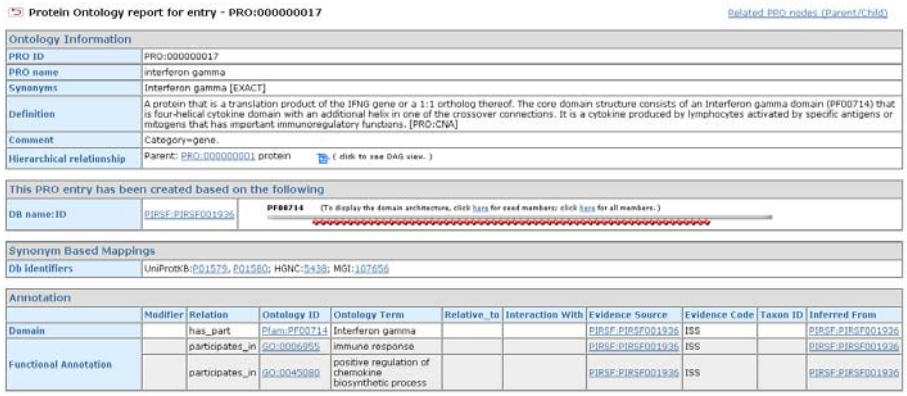
PIR-Protein Information Resource
http://pir.georgetown.edu/

PIR—PRO (Protein Ontology)

- Identifies hierarchies and relationships realted to proteins supplied by the user



Protein Ontology, example



Protein Ontology report for entry - PRO:000000017

Related PRO nodes (Parent/Child)

Ontology Information

PRO ID	PRO:000000017
PRO name	interferon gamma
Synonyms	Interferon gamma [EXACT]
Definition	A protein that is a translation product of the IFNG gene or a 1:1 ortholog thereof. The core domain structure consists of an Interferon gamma domain (PF00714) that is four-helical cytokine domain with an additional helix in one of the crossover connections. It is a cytokine produced by lymphocytes activated by specific antigens or mitogens that has important immunoregulatory functions. [PRO:CNA]
Comment	Category=gene.
Hierarchical relationship	Parent: PRO:000000001 protein (click to see DAG view.)

This PRO entry has been created based on the following

DB name:ID	PIRSF:PIRSF001936

PF00714 (To display the domain architecture, click here for seed members; click here for all members.)

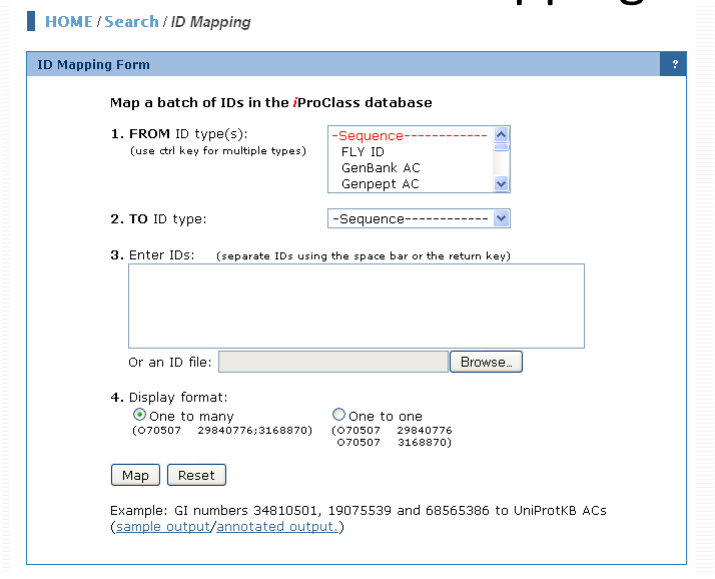
Synonym Based Mappings

Db identifiers	UniProtKB:P01579, P01580; HGNC:5438; MGI:107656

Annotation

	Modifier	Relation	Ontology ID	Ontology Term	Relative_to	Interaction With	Evidence Source	Evidence Code	Taxon ID	Inferred From
Domain		has_part	Pfam:PF00714	Interferon gamma			PIRSF:PIRSF001936	ISS		PIRSF:PIRSF001936
Functional Annotation		participates_in	GO:0006955	immune response			PIRSF:PIRSF001936	ISS		PIRSF:PIRSF001936
		participates_in	GO:0045080	positive regulation of chemokine biosynthetic process			PIRSF:PIRSF001936	ISS		PIRSF:PIRSF001936

PIR—ProClass– ID Mapping



HOME / Search / *ID Mapping*

ID Mapping Form

Map a batch of IDs in the *i*ProClass database

1. **FROM** ID type(s):
(use ctrl key for multiple types)
-Sequence----------
FLY ID
GenBank AC
Genpept AC

2. **TO** ID type:
-Sequence----------

3. Enter IDs: (separate IDs using the space bar or the return key)

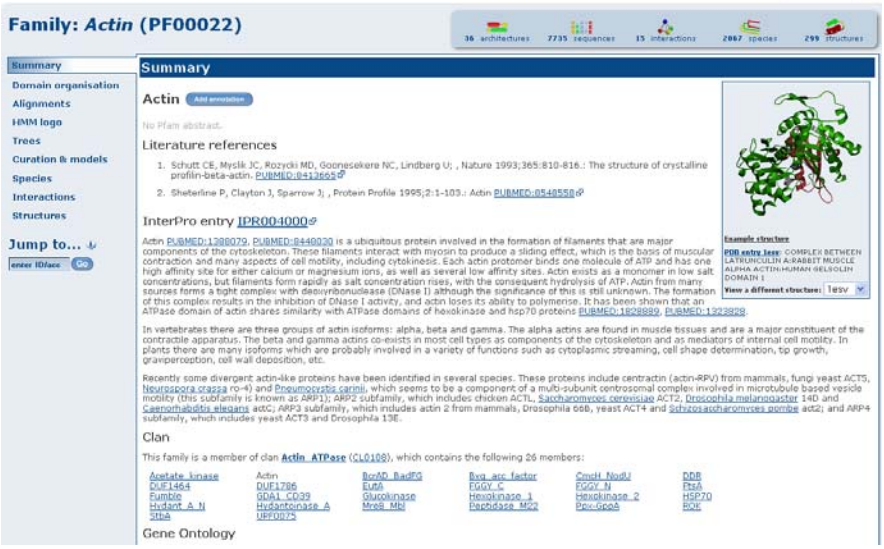
Or an ID file: Browse...

4. Display format:
One to many (O70507 29840776;3168870)
One to one (O70507 29840776
O70507 3168870)

Map Reset

Example: GI numbers 34810501, 19075539 and 68565386 to UniProtKB ACs
(sample output/annotated output.)

Pfam-Protein Families (http://pfam.sanger.ac.uk/)



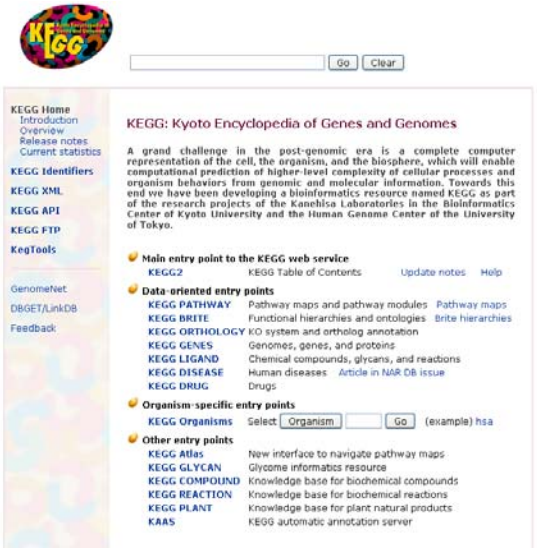
SCOP—Structural Characterization of Proteins (http://scop.mrc-lmb.cam.ac.uk/scop/)

- Following input of a sequence whose structure is unknown, SCOP will identify regions in the test protein which have sequence similarities with regions in proteins that have a structure.
- SCOP will identify the PDB entry and the structure of the queried region

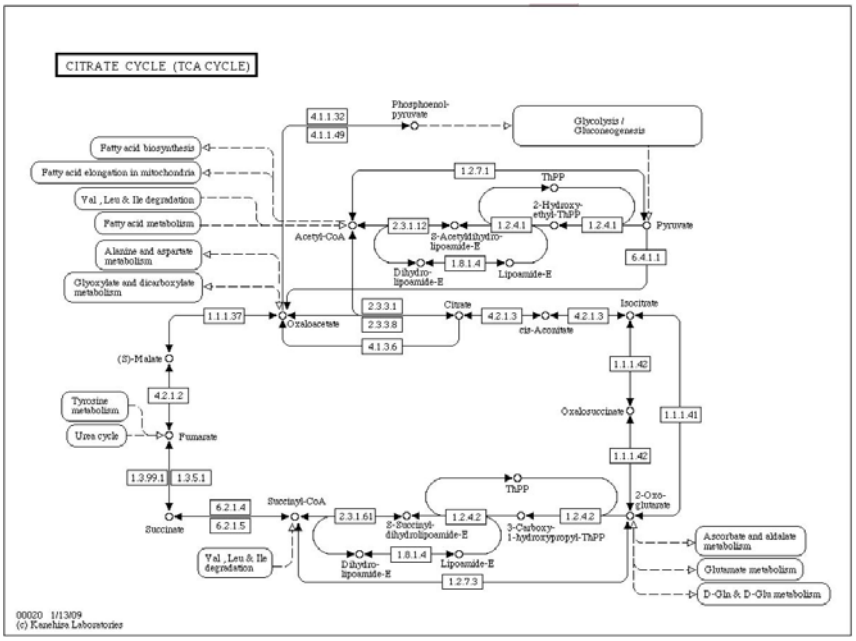
Biological pathways

- Critical because of the role that proteins play in it. Pathways are responsible for myriad functions. Each step in a pathway is catalyzed by an enzyme (protein)

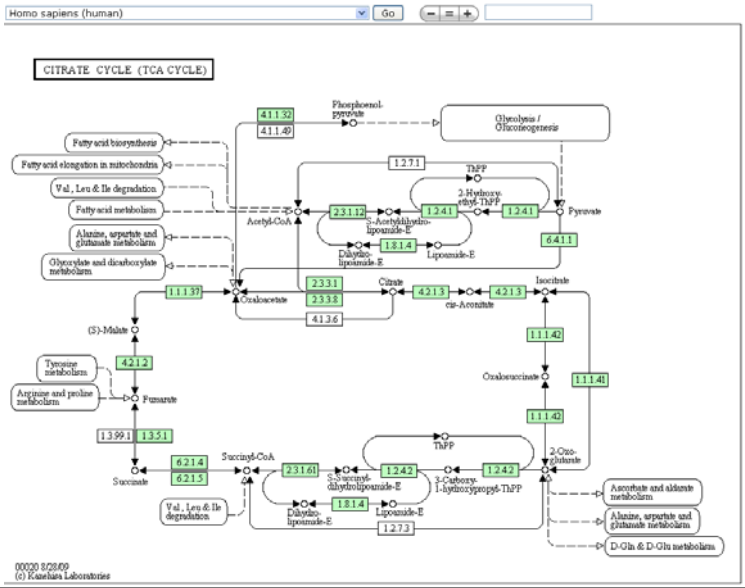
KEGG (**Kyoto Encyclopedia of Genes and Genomes) (http://www.genome.jp/kegg/)**



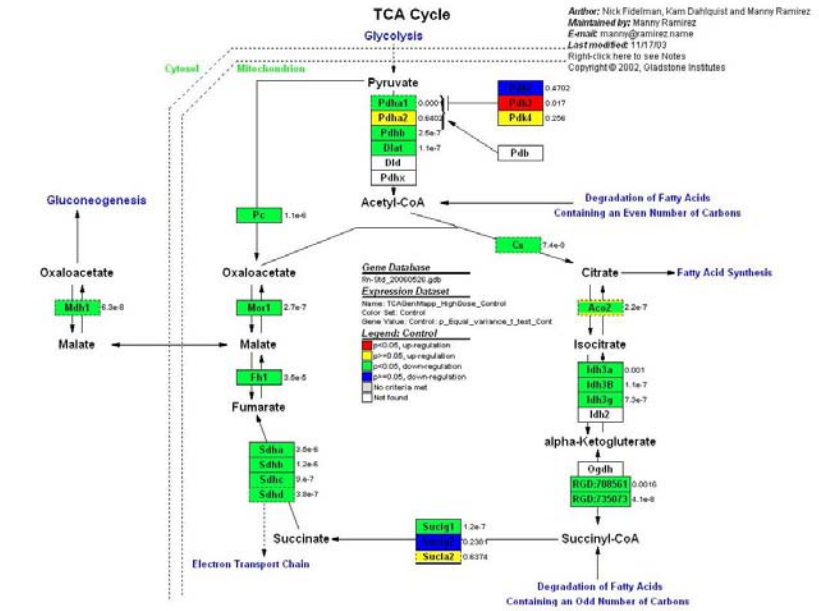
KEGG Pathway



KEGG—TCA Cycle in Humans



Center for Gene Nutrition Interactions in Cancer --UAB



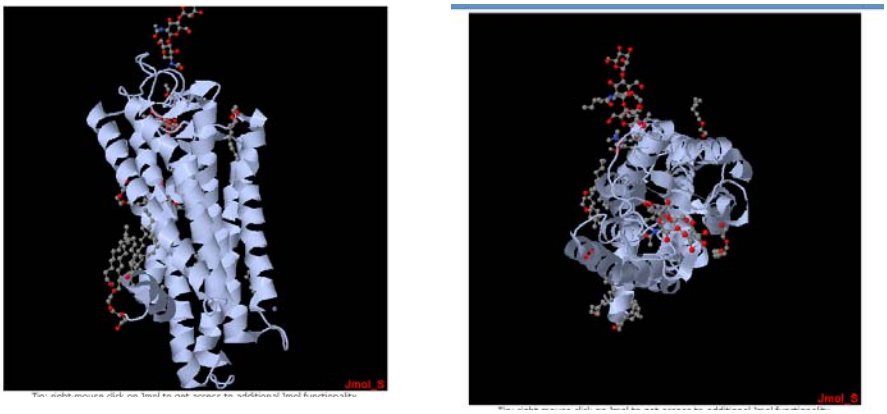
Information Exchange!!

Bioinformatics and Computational Biology as a Drive or Discovery

- Olfactory receptors are G-protein coupled receptors embedded in the mucus membrane lining the nostrils.
- They interact with a G-protein following activation by an odorant molecule and catalyze a signal transduction cascade; the signal gets processed in the olfactory processing region of the brain—resulting in the perception of smell
  - Important consequences for neurobiological disorders

A GPCR structure

PDB entry 3C9L



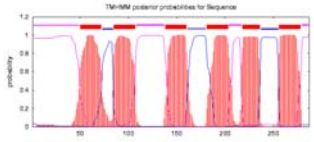
Structurally, a GPCR and (presumably) an OR contains seven helical regions connected by six intra-helical loops, and an N- and C-terminus in the extended (loop like) conformation. The N-terminus is extra-cellular; the C-terminus is intra-cellular

OR17-210

- **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.**

- **Sequence Features for OR17-210**
- **This protein sequence for olfactory receptor OR17-210 appears as a pseudogene in the HORDE1 database** (http://bip.weizmann.ac.il/cgi-bin/HORDE/showGene.pl?key=symbol&value=OR1E3P):
  - ATGATGAAGAAGAACCAAACCATGATCTCAGAGTTCCTGCTCCTGGGCCTTCCATCCAACCTGAGCAGCAGAATCTGTTCTATGCCTTGTTCTTGGCCGTGTATCTTACC
    ACCCTCCTGGGGAACCTCCTCGTCATTGTCCTCATTCGACTGGACTCCCACCTCCACATGCCTATGTATTTGTGTCTCAGCAACTTGTCCTTCTCTGACCTCTGCTTTTCCT
    CGGTCACAATGCCCAAATTGCTGCAGAACATGCAGAGCCAAAACCCATCCATCCCCTTTGCGGACTGCCTGGCTCAGATGTACTTTCATCTGTTTTATGGAGTTCTGGA
    GAGCTTCCTCCTTGTGGTCATGGCTTATCACTGCTATGTGGCTATTTGCTTTCCTCTGCACTACACCACTATCATGAGCCCCAAGTGTTGCCTTGGTCTGCTGACACTCTC
    CTGGCTGTTGACCACTGCCCATGCCACGTTGCACACCTTGCTTATGGCCAGGCTGTCCTTTTGTGCTGAGAATGTGATTCCTCACTTTTTCTGTGATACATCTACCTTGTT
    GAAGCTGGCCTGCTCCAACACGCAAGTCAATGGGTGGGTGATGTTTTTCATGGGCGGGCTCATCCTTGTCATCCCATTCCTACTCCTCATCATGTCCTGTGCAAGAATC
    GTCTCCACCATCCTCAGGGTCCCTTCCACTGGGGGCATCCAGAAGGCTTTCTCCACCTGTGGCCCCCACCTCTCTGTGGTGTCTCTCTTCTATGGGACAATTATTGGTCTC
    TACTTGTGCCCATTGACGAATCATAACACTGTGAAGGACACTGTCATGGCTGTGATGTACACTGGGGTGACCCACATGCTGAACCCCTTCATCTACAGCCTGAGGAAC
    AGAGACATGAGGGGGAACCCTGGGCAGAGTCTTCAGCACAAAGAAAATTTTTTTGTCTTTAAAATAGTAATAGTTGGCATTTTACCGTTATTGAAT
- **Intuitively Translated as**:
  - MMKKNQTMISEFLLLGL/PIQPEQQNLFYALFLAVYLTTLLGNLLVIVLIRLDSHLHMPMYLCLSNLSFSDLCFSSVTMPKLLQNMQSQNPSIPFADCLAQMYFHLFYGVLESFL
    LVVMAYHCYVAICFPLHYTTIMSPKCCLGLLTLSWLLTTAHATLHTLLMARLSFCAENVIPHFFCDTSTLLKLACSNTQVNGWVMFFMGGLILVIPFLLLIMSCARIVSTILRVPS
    TGGIQKAFSTCGPHLSVVSLFYGTIIGLYLCPLTNHNTVKDTVMAVMYTGVTHMLNPFIYSLRNRDMRGNPGQSLQHKENFFVFKIVIVGILPLLN
- **A two nucleotide frame shift however results in a functional protein with the following sequence** :
  - MPMYLCLSNLSFSDLCFSSVTMPKLLQNMQSQNPSIPFADCLAQMYFHLFYGVLESFLLVVMAYHCYVAICFPLHYTTIMSPKCCLGLLTLSWLLTTAHATLHTLLMARLSFCA
    ENVIPHFFCDTSTLLKLACSNTQVNGWVMFFMGGLILVIPFLLLIMSCARIVSTILRVPSTGGIQKAFSTCGPHLSVVSLFYGTIIGLYLCPLTNHNTVKDTVMAVMYTGVTHML
    NPFIYSLRNRDMRGNPGQSLQHKENFFVFKIVIVGILPLLN
- The TRANSLATE tool in EXPASY will translate a nucleotide sequence into the protein sequence. It will also do so following a one and two-nucleotide frame shift

- 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. (TMHMM --http://www.cbs.dtu.dk/services/TMHMM/)



- What is typically helix 3 in ORs is he ich 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