

# Web-based Bioinformatics Applications in Proteomics

Chiquito Crasto

[ccrasto@genetics.uab.edu](mailto: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)

All Databases PubMed Nucleotide Protein Genome Structure OMIM PNC Journals Books

Search Structure for Go Clear

Limits Preview/Index History Clipboard Details

Molecular Modeling Database (MMDB) RESOURCES SEARCH METHODS HOW TO HELP NEWS FTP PUBLICATIONS DISCOVER

Hints on Finding 3D Macromolecular Structures

- This page is used for searching by text term (other search methods allow queries by protein sequence)
- Enter one or more search terms (e.g., chloride channel)
- Use search fields and other Advanced Search options (Limits, Preview/Index, and History) to refine a search
- Boolean operators AND, OR, NOT must be in upper case
- Use quotes to force a phrase search (e.g., "voltage gated")
- Use a wildcard (e.g., glyco\*[bd]) to search for a word stem
- Search results and structure record displays are described in the help document.

About the Database

Three dimensional structures provide a wealth of information on the biological function, on mechanisms linked to the function, and on the evolutionary history of and relationships between macromolecules. Most 3D-structure data are obtained from X-ray crystallography and NMR-spectroscopy.

The Molecular Modeling Database (MMDB), also known as "Entrez Structure", is a database of experimentally determined structures obtained from the RCSB Protein Data Bank (PDB). MMDB is developed by the Structure Group of the NCBI Computational Biology Branch. The data processing procedure at NCBI results in the addition of a number of useful features that facilitate computation on the data and link them to many other data types in the Entrez system. The help document and how-to pages provides examples of how the database can be used.

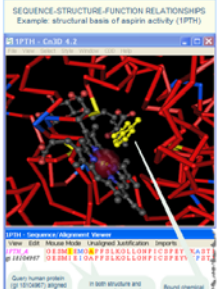
The structure database is considerably smaller than Entrez's Protein or Nucleotide databases, but a large fraction of all known protein sequences have homologs in this set, and one may often learn more about a protein by examining 3-D structures of its homologs. These are accessible as "Related Structures" in the "Links" menu of Entrez Protein sequence records (illustrated example). It is then possible to align the query protein to the structure-based sequence, as shown in the illustration on this page.

Additional resources can be used along with MMDB to interactively view the structures, find similar 3D structures, learn about the types of interactions and bound chemicals that have been found to exist among the similar 3D structures, and more.

RETRIEVE STRUCTURES THAT HAVE		
Protein Only	DNA Only	RNA Only
Protein + Chemical	DNA + Chemical	RNA + Chemical
Protein + DNA	Protein + RNA	DNA + RNA

The "How To" page provides tips for searching by gene/protein product, molecule type, and more.

SEQUENCE-STRUCTURE-FUNCTION RELATIONSHIPS  
Example: structural basis of aspirin activity (1PTM)



1PTM: Sequence by Structural View

View: Entrez Structure Original Refinement: 2007

1PTM: 4 33 58 73 88 103 118 133 148 163 178 193 208 223 238 253 268 283 298 313 328 343 358 373 388 403 418 433 448 463 478 493 508 523 538 553 568 583 598 613 628 643 658 673 688 703 718 733 748 763 778 793 808 823 838 853 868 883 898 913 928 943 958 973 988 1003

Quick Search options: (1) 1000000 proteins (2) 1000000 PDB structures (3) 1000000 related chemicals

## Structure of Actin—Genbank Structure View

NCBI

Structure Summary  
MMDB

Entrez Structure Protein CDD PubMed Taxonomy PubChem Help Cn3D

MMDB ID: 69126 PDB ID: 2ZWH Search PDB or MMDB ID

Description: Model For The F-Actin Structure.  
Deposition: Oda T, Iwasa M, Aihara T, Maeda Y, Nanta A, 2008/12/5  
Taxonomy: Oryctolagus cuniculus  
Related Structure: VAST

Structure View in Cn3D Structure View in RasMol

Tasks: Display Drawing: All Atoms

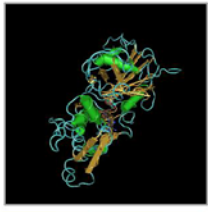

Download Cn3D View Cn3D Tutorial Visualization software

Molecular components in the MMDB structure are listed below and may include macromolecular chains, 3D domains, protein classifications (domain families), and ligands, as available. Mouse over each icon for more information on the component.

Protein  
3D Domains  
Domain Families  
Specific hits  
Superfamilies  
Multi-domain


Sequence (A)

ACTIN superfamily  
Actin

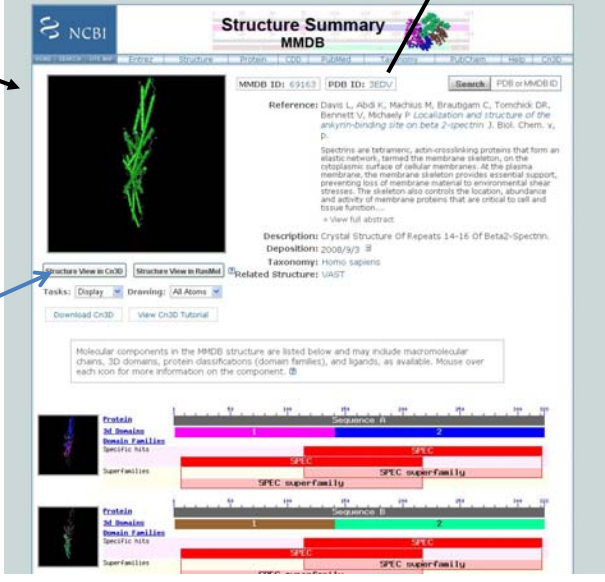
**Structure of Domains in Genbank**

List of domains related to or associated with Actin



Cn3D

Link to Protein Databank



**Genbank: Amino Acid Explorer**

NCBI

NCBI Structures

Amino Acid Explorer

PSSM Viewer

Course Main Page

Key to Symbols

Description of Displayed Data

Mutation Analyzer

Compare

A-Ala

to

C-Oys

using

text

Compare

Questions or comments

NCBI Structures
This tool was created as part of an NCBI course and is still under development.

Amino Acid Explorer

Biochemical Properties

View a table displaying various properties of all 20 amino acids.

Display Table

Structure and Chemistry

View structural views and detailed properties of a given amino acid.

Choose an amino acid: alanine

View Properties

Common Substitutions

Using data from the BLOSUM62 matrix, view a list of amino acids ranked by how often they substitute for a given amino acid.

Choose an amino acid: alanine

View Substitutions

Mutation Analyzer

Interactively discover what amino acid substitutions result from selected codon mutations.

Analyze Mutations

Amino Acids at Work

Using data from NCBI curated CD records, explore functional sites within proteins in which a given amino acid plays a pivotal part.

Choose an amino acid: alanine

View Functional Sites

Amino Acids as Ligands

Using Entrez Structure, retrieve 3D protein structures containing a given amino acid as a ligand.

Choose an amino acid: alanine

Retrieve Structures

## 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

Items 1 - 20 of 150

**Actin** [c00012]

ACTIN: Actin, An ubiquitous protein involved in the formation of filaments that are a major component of the cytoskeleton. Interaction with various proteins provides the basis of muscular contraction and many aspects of cell motility. Each actin protomer binds one molecule of ATP and either calcium or magnesium ions. Actin exists as a monomer in low salt concentrations, but filaments form rapidly at salt concentrations near, with the conserved hydrophobic ATP. Polymerization is regulated by co-actin capping proteins. The ATPase domain of actin shares evolutionary with ATPase domains of heme kinase and myo-20 proteins. [c00012] [20096]

**F-actin\_head, F-actin binding** [pfam01219]

F-actin\_head, F-actin binding: The F-actin binding domain forms a compact bundle of four antiparallel alpha-helices, which are arranged in a left-handed topology. Binding of F-actin to the F-actin binding domain may result in cytoplasmic retention and subcellular distribution of the protein, as well as possible inhibition of protein function. [pfam01219] [17466]

**ACTIN, ACTIN subfamily of ACT1 domain [superfamily]** [smart00468] [7597]

ACTIN, ACTIN subfamily of ACT1 domain [superfamily]

**F-actin\_cup\_A, F-actin capping protein, alpha subunit** [pfam01267] [10380]

F-actin\_cup\_A, F-actin capping protein, alpha subunit

**F-actin\_cup\_B, F-actin capping protein, beta subunit** [pfam01113] [10130]

F-actin\_cup\_B, F-actin capping protein, beta subunit

**Actin** [c00020] [109991]

Actin

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

NCBI OMSSA

Search Search Status Download FAQ Help

File name:  Browse...

Enzyme:

Sequence library:

Hitlist max length:

Fixed mods (ctrl key for multiple selection):

- 2-amino-3-oxo-butyric acid T
- CAMthio-propanoyl K
- ICAT heavy
- ICAT light
- M cleavage from protein n-term
- NEM C
- NIP-CAM

Maximum variable mod combinations searched per peptide:

Precursor mass tolerance (Da):

Precursor mass search type:

Lower bound of precursor charge:

Minimum charge to start using multiply charged products:

Fraction of product peaks below precursor to determine +1 precursor:

Peak intensity cutoff:  (fraction of most intense)

Ions to search 1:

Search

File type:

Maximum missed cleavages:

Species to search (ctrl key for multiple selection):

- Mus musculus (mouse)
- Saccharomyces cerevisiae (yeast)
- Aeropyrum pernix
- Agrobacterium tumefaciens
- Anopheles gambiae
- Aquifex aeolicus

E-value cutoff:

Variable mods (ctrl key for multiple selection):

- 2-amino-3-oxo-butyric acid T
- CAMthio-propanoyl K
- ICAT heavy
- ICAT light
- M cleavage from protein n-term
- NEM C
- NIP-CAM

Product mass tolerance (Da):

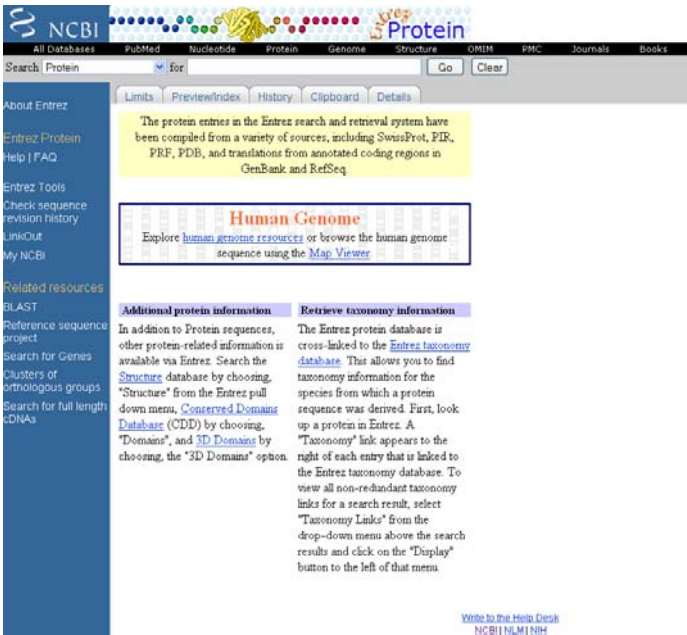
Product mass search type:

Upper bound of precursor charge:

Number of top intensity peaks in first pass:

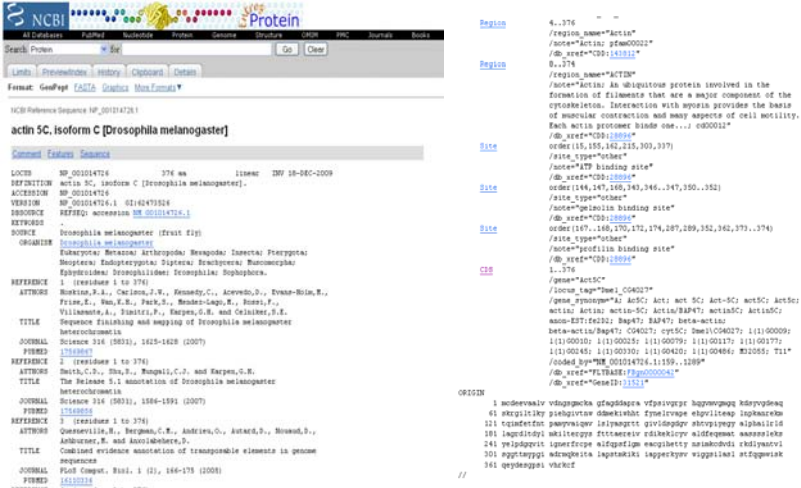
Ions to search 2:

## Protein—Genbank's Protein Search System



The screenshot displays the NCBI Protein search system interface. At the top, there are navigation tabs for 'All Databases', 'PubMed', 'Nucleotide', 'Protein', 'Genome', 'Structure', 'OMIM', 'PMC', 'Journals', and 'Books'. The search bar contains 'Protein' and 'for'. Below the search bar are buttons for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. A yellow box highlights that protein entries are compiled from various sources like SwissProt, PIR, PRF, PDB, and GenBank. A blue box titled 'Human Genome' encourages exploring human genome resources. The main content area has two columns: 'Additional protein information' and 'Retrieve taxonomy information', both providing detailed instructions on how to use the search system's features.

## Genbank resource ... Protein



This screenshot shows the detailed search results for the protein 'actin 5C, isoform C [Drosophila melanogaster]'. The top section includes the NCBI logo and search bar. The main content is divided into several sections: 'Summary', 'Features', 'Sequence', and 'References'. The 'Features' section lists key identifiers like LOCID (NP\_001014726), DEFINITION, and SOURCE. The 'Sequence' section shows the amino acid sequence of the protein, with various domains and motifs highlighted in color and labeled with 'Region' and 'Site'. The 'References' section provides a list of scientific publications related to this 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]
MSDEETALVVDNGSGNVKAGFAGDDAPRCVFPISVGRPKNPALMVGMDEKDTYVGDEAQSKRGILTLY
PIEHGIVTNWEDMEKIWHHTFYNELRIAPEEHPVLLTEAPMNPKANREKMTTTFETHNVPMYVAIQAV
LSLYSSGRTTGIVLDSGDGVHTVPIYEGYALPHAIMRLDLAGRDLEFMQKILVERGFSFTTAEKEIV
RDIKEKLCYIALDFDEEMTTSSSSSEVEKSYELPDGNIITVGNERFRCPVLFQPTFIGMEAPGIHTTTY
NSIVRCDVDIRKDYANVVLSSGTTMFEIGIGQRMTKELNALVPSTMKIKV VAPPERKYSVWIGGSILSSL
STFQQMWITKEEFDESGPNIVHRKCF
```

## Protein Clusters

NCBI Protein Clusters

All Databases PubMed Nucleotide Protein Genome Structure OMIM PRC Journals Books

Search Protein Clusters For  Go Clear

Limits Preview/Index History Clipboard Details

### Protein Clusters

#### About the Database

Welcome to Entrez Protein Clusters (ProtClustDB). This collection of related protein sequences (clusters) consists of reference Sequence proteins encoded by complete genomes. This database contains both curated and non-curated clusters. For release-specific information check the [stats page](#).

The Protein Clusters database provides easy access to annotation information, publications, domains, structures, and external links and analysis tools including multiple alignments, phylogenetic trees, and genomic neighborhoods ([ProtMap](#)).

Protein Clusters can be searched like any other Entrez database. For more information on how to use Entrez please examine the [Entrez Help](#) Document.

A publication describing ProtClustDB is now available: [Klimke et al., 2009. The National Center for Biotechnology Information's Protein Clusters Database. Nucleic Acids Res. 2009 Jan;37\(Database issue\):O216-23. Epub 2009 Oct 21.](#)

A specialized BLAST service is accessible ([Concise Protein BLAST](#)).

Data is available for download via [Protein Clusters FTP](#).

#### Example Searches

all clusters with ribosomal protein as the curated name  
`"ribosomal protein"[Protein Name]`  
 all clusters that are encoded by chloroplasts  
`"source:chloroplast"[All Fields]`

Check the [limits page](#) and the [help document](#) for more information.



## Clustering Proteins in terms of Sequence Similarities--Genbank

**PRK13410**
molecular chaperone DnaK
Gene name: None

(Curated - ProteinSet)

**Cluster Info**

Total proteins: 12  
Conserved in: **Cyanobacteria**  
Total genes: 9  
Total organisms: 12  
Putative Paralogs: 0  
Publications: 11

**Cluster Tools**

Show detailed alignment:

Build tree:

Genome ProtMap by PRK13410:

Genome ProtMap by COG04430:

**Cross references**

COG(s): [COG04430](#)  
HAMAP: [MF\\_00332](#)  
KEGG ID: [M08543](#)  
InterPro: [I:TCR02258](#)  
TIGRFAM: [TIGR02258](#)  
Domain(s): [c03378](#), [PF00121](#), [pfam02782](#), [GGV\\_Q](#)

**Entrez Links**

heat shock protein 70, assists in folding of nascent polypeptide chains, refolding of misfolded proteins, utilizes ATPase activity to help fold, co-chaperones are DnaJ and Grp

COG functional category: **Posttranslational modification, protein turnover, chaperones**

BRITe hierarchy: **Genetic Information Processing/Folding, Sorting and Degradation/Protein Folding and associated processing/Environmental Inf.**

**Publications by categories (only one publication per category is shown)**

- Curated [1]**: Characterization of the dnaK multigene family in the Cyanobacterium *Synechococcus* sp. strain PCC7942. *J Bacteriol* 2001 Feb more...
- SwissProt [2]**: Sequence analysis of the third dnaK homolog gene in *Synechococcus* sp. PCC7942. *Biochem Biophys Res Commun* 1994 Dec 30 more...
- By Homology [3]**: The electronic Plant Gene Register. *Plant Physiol* 1997 Jul more...
- CDI [7]**: Molecular evolution of the actin family. *J Cell Sci* 2002 Jul 1 more...

Related Clusters [1]: [PRK13410](#)  
(sequence similarity, must be exact)

Organism	Protein name	Prev. Cluster	Accession	Next Cluster	Locus_tag	Length	BLink	Alignment
<b>H.Cyanobacteria</b>								
<input type="checkbox"/> <i>Synechococcus</i> sp. strain PCC7942	DnaK	<a href="#">CL1312615</a>	<a href="#">YP_001519059</a>	<a href="#">CL1312630</a>	ANI_0562	870aa		
<input type="checkbox"/> <i>Synechococcus</i> sp. strain ATCC 29413	DnaK	<a href="#">CL1312632</a>	<a href="#">YP_3211438</a>	<a href="#">CL1312630</a>	Ans_0919	868aa		
<input type="checkbox"/> <i>Synechococcus</i> sp. ATCC 6116	heat shock protein 70	<a href="#">P5104196</a>	<a href="#">YP_001802772</a>	<a href="#">CL1312613</a>	hsp_1395	770aa		
<input type="checkbox"/> <i>Synechococcus</i> sp. strain NIES-84	DnaK	<a href="#">CL1312632</a>	<a href="#">YP_001899609</a>	<a href="#">CL1312613</a>	MAK_10560	720aa		
<input type="checkbox"/> <i>Synechococcus</i> sp. strain PCC 7232	DnaK	<a href="#">CL1312632</a>	<a href="#">YP_001899621</a>	<a href="#">CL1312630</a>	Nepu_05900	705aa		
<input type="checkbox"/> <i>Synechococcus</i> sp. PCC 7320	DnaK	<a href="#">CL1312632</a>	<a href="#">WP_007030</a>	<a href="#">CL1312630</a>	sp099	880aa		
<input type="checkbox"/> <i>Synechococcus</i> sp. strain PCC 6803	DnaK	<a href="#">CL0802762</a>	<a href="#">YP_174248</a>	<a href="#">CL0802769</a>	sp1500_f	740aa		
<input type="checkbox"/> <i>Synechococcus</i> sp. strain PCC 7945	DnaK	<a href="#">CL0802762</a>	<a href="#">YP_401567</a>	<a href="#">CL0802769</a>	Synp07942_2480	740aa		
<input type="checkbox"/> <i>Synechococcus</i> sp. PCC 2962	DnaK	<a href="#">CL1312689</a>	<a href="#">YP_001725596</a>	<a href="#">CL1312693</a>	Synp07902_2160	702aa		
<input type="checkbox"/> <i>Synechococcus</i> sp. PCC 6803	DnaK	<a href="#">P5509511</a>	<a href="#">WP_008465</a>	<a href="#">CL1312612</a>	sp1032	700aa		
<input type="checkbox"/> <i>Synechococcus</i> sp. strain PCC 6803	DnaK	<a href="#">CL1312689</a>	<a href="#">WP_008465</a>	<a href="#">CL1312630</a>	sp1150	800aa		
<input type="checkbox"/> <i>Thermotoga</i> sp. strain ATCC 49619	DnaK	<a href="#">CL0120009</a>	<a href="#">YP_729505</a>	<a href="#">CL0120003</a>	Tev_4012	870aa		

# ORF-Finder

## ORF Finder (Open Reading Frame Finder)

PubMed
Entrez
BLAST
OMIM

NCBI

Tools for data mining

GenBank sequence submission support and software

FTP site download data and software

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames in a sequence already in the database.

This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence is displayed and searched against the sequence database using the WWW BLAST server. The ORF Finder should be used for all sequence submissions. It is also packaged with the Sequin sequence submission software.

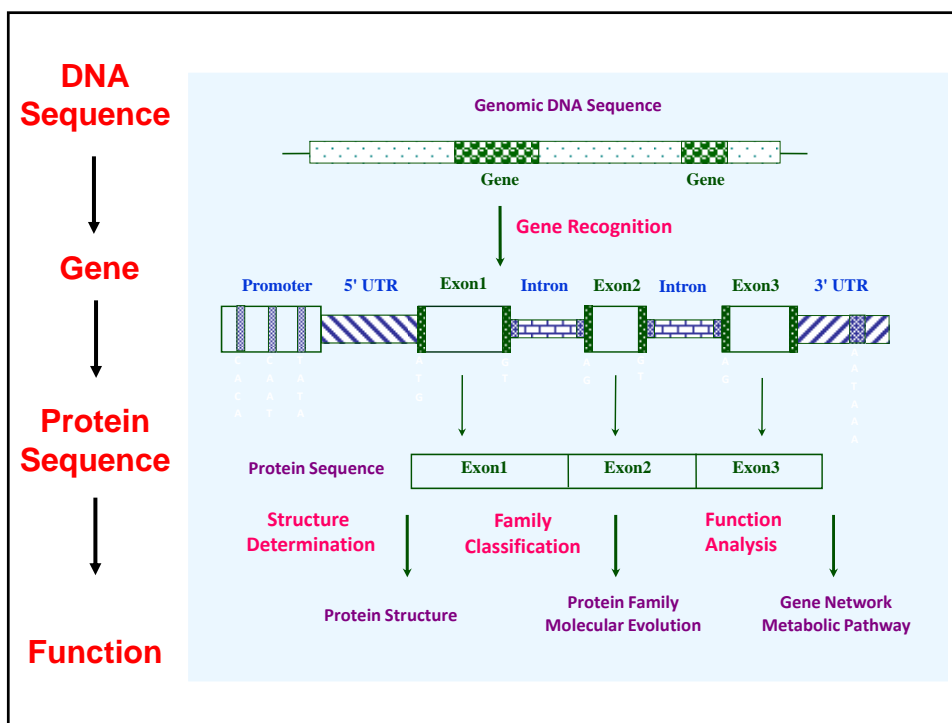
**Enter GI or ACCESSION**

**or sequence in FASTA format**

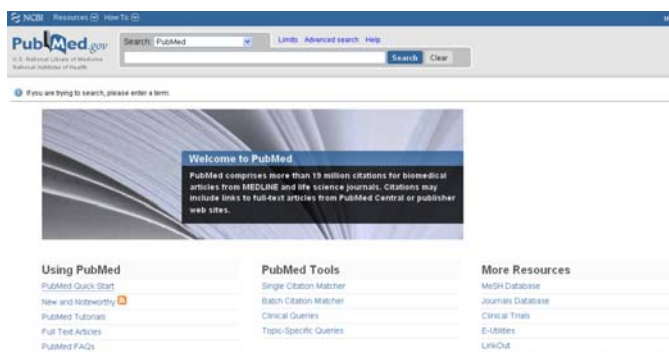
**FROM:**  **TO:**

Genetic codes:

9



## 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

The screenshot shows the PubMed website interface. At the top, there are logos for NCBI and PubMed, along with the text 'A service of the U.S. National Library of Medicine and the National Institutes of Health'. Below the logos, there are navigation tabs for 'All Databases', 'PubMed', 'Nucleotide', 'Protein', 'Genome', 'Structure', 'OMIM', 'PMC', 'Journals', and 'Books'. A search bar contains the text 'PubMed' and a 'Go' button. To the right of the search bar, there is a 'Link to Full Text if available' button. Below the search bar, there are options for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. The main content area displays the abstract for the article 'Dependence of alignment direction on magnitude of strain in esophageal smooth muscle cells.' by Ritchie AC, Wilava S, Ong WF, Zhong SP, Chan KS. The abstract text describes the response of cells in vitro to mechanical forces and the effect of cyclic stretch on cell alignment and proliferation. The abstract is followed by a 'Related Articles' section with several links to other articles. At the bottom of the page, there is a 'Write to the Help Desk' link and a 'Privacy Statement/ Freedom of Information Act/ Disclaimer' link.

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

The screenshot shows the Ensembl genome browser homepage. At the top, there is the Ensembl logo and a search bar. The search bar contains the text 'Human' and 'BRCA2'. Below the search bar, there is a 'Go' button. The main content area is divided into several sections. On the left, there is a 'Browse a Genome' section with a list of popular genomes including Human, Chimpanzee, Macaque, Mouse, Rat, Dog, Chicken, Zebrafish, Fruitfly, and C.elegans. On the right, there is a 'New to Ensembl?' section with a list of features and a 'What's New in Release 52' section with a list of updates. At the bottom, there is a 'Sanger' logo and a 'What's New in Release 52' section with a list of updates.

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

**Gene: BRCA2 (ENSG00000139618)**  
 Breast cancer type 2 susceptibility protein (Fanconi anemia group D1 protein) [Ensembl: UniProtKB/Swiss-Prot:P81987](#)

**Location** [Chromosome 13: 31,707,617-31,871,808](#) forward strand

**Transcripts** There is one transcript in this gene: [BRCA2-001 \(ENST00000360152\)](#), with protein product [ENSP00000369497](#).

**Gene summary** [/ho/p](#) Splice variants **w**

**Name** [BRCA2](#) (HGNC (curated))

**Synonyms** BRCC2, FACD, FAD, FAD1, FANCD, FANCD1 [To view all Ensembl genes linked to the name [click here](#)]

**CCDS** This gene is a member of the Human CCDS set: [CCDS3344](#)

**Gene type** Known protein coding


**Prediction Method** Gene containing both Ensembl genebuild transcripts and [Vega](#) manual curation, see [article](#).

**Transcripts**

Ensembl release 52 - Dec 2009 © [WTSI](#) | [EBI](#) Ensembl is available to [download for public use](#) - please see the [code licence](#) for details. This is a mirror site of Ensembl from [BOL.SZ](#). [Permanent link](#) - [View in archive site](#)

[About Ensembl](#) | [Contact Us](#) | [Help](#)

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

 **Swiss Institute of Bioinformatics** Search

**EXPASY Proteomics Server** [Databases](#) [Tools](#) [Services](#) [Mirrors](#) [About](#) [Contact](#)

You are here: EXPASY.CH

The EXPASY (Expert Protein Analysis System) 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 ([Disclaimer](#) / [References](#) / [Linking to EXPASY](#)).

**Databases**

[UniProtKB](#), [PROSITE](#), [HAMAP](#), [SwissVar](#), [ViralZone](#), [SWISS-MODEL](#) Repository, [SWISS-2DPAGE](#), [World-2DPAGE](#) Repository, [MAPE@eDB](#), [ENZYME](#), [GlycoSuiteDB](#), [UniPathway](#)  
[\[details\]](#) [\[full list\]](#)

**Education & services**


[Downloads](#), [Protein Spotlight](#), [Protéines à la «Une»](#), [e-proxemis](#), [Bioinformatics core facility for Proteomics](#)  
[\[full list\]](#)

**Tools & Software**

[Proteomics tools](#), [Blast](#), [ScanProsite](#), [Melanie](#), [MSight](#), [Make2D-DB](#), [SWISS-MODEL](#), [Swiss-PdbViewer](#)  
[\[full list\]](#)

**Documentation**

[What's New?](#), [E-mail alerts](#), [UniProtKB documentation](#), [How to link to EXPASY](#), [Advanced search](#)  
[\[full list\]](#)

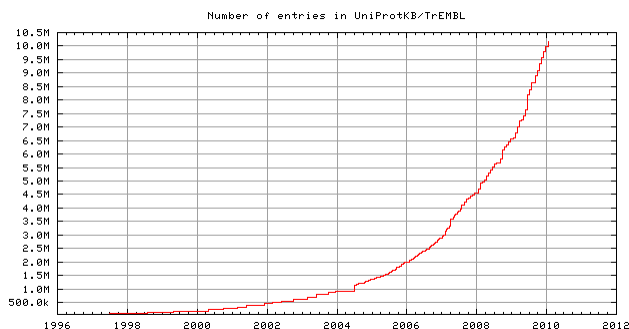
**Latest News** 

**Protein Spotlight** - Dec 21, 2009 [String of intrusion](#)  
 When I was little, I used to wear small cotton shirts that were knitted by my grandmother. So? Well, onto them she sewed tiny ruche buttons you could never get hold of and which mesmerized me because of the different colours that shone off them depending on how you oriented them in the light. [\[more\]](#)

**World-2DPAGE** - Oct 23, 2009  
 New data uploaded into the [World-2DPAGE](#) Repository. Currently, 113 maps for 16 species are available from the [World-2DPAGE Portal](#).  
[\[more news\]](#) [\[SIB news\]](#)

Last modified 06/Nov/2009 by CHH

# Uniprot (Swissprot and TrEMBL)



Contains (to date) more than ten million protein sequences

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

# UniProt

The screenshot displays the UniProt website interface. At the top, there are navigation tabs: Search, Blast, Align, Refseq, and ID Mapping. Below these is a search bar with the text 'actin' entered. The search bar also includes a dropdown menu for 'Protein Knowledgebase (UniProtKB)' and buttons for 'Search', 'Clear', and 'Fields »'.

The main content area is divided into several sections:

- WELCOME**: A brief message about the mission of UniProt.
- What we provide**: A table listing various services:
 

Service	Description
UniProtKB	Protein knowledgebase, consists of two sections: <ul style="list-style-type: none"> <li>★ Swiss-Prot, which is manually annotated and reviewed</li> <li>★ TrEMBL, which is automatically annotated and is not reviewed.</li> </ul> Includes <a href="#">Complete Proteome Sets</a> .
UniRef	Sequence clusters, used to speed up similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, taxonomy, keywords and more.
- NEWS**: A section titled 'UniProt release 15.13 - Jan 19, 2010' with sub-sections for 'XMRV complete proteome in UniProtKB/Swiss-Prot', 'Cross-references to eggNOG', 'Change to cross-references to HAMAP and HOGENOM', 'Statistics for UniProtKB', 'Swiss-Prot - TrEMBL', 'Forthcoming changes', and 'News archives'.
- SITE TOUR**: A section with a thumbnail image and the text 'Learn how to make best use of the tools and data on this site.'
- PROTEIN SPOTLIGHT**: A section at the bottom of the page.

The UniProt logo is visible at the bottom center of the page.

## PROSITE—families, patterns, profiles and functional sites

prosite Database of protein domains, families and functional sites

Home ScanProsite ProRule Documents Downloads Links Funding

PROSITE consists of **documentation entries** describing protein domains, families and functional sites as well as associated **patterns** and **profiles** to identify them ([More details](#) / [References](#) / [Disclaimer](#) / [Commercial users](#)). PROSITE is complemented by **ProRule**, a collection of rules based on profiles and patterns, which increased the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids ([More details](#)).

Release 20.59, of 20-Jan-2010 (1567 documentation entries, 1366 patterns, 673 profiles and 674 ProRule)

**PROSITE access**

actin e.g. PDOC00022, PS50089, SH3, zinc finger

Search

add wildcard ""

Browse:

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hit

**PROSITE tools**

Scan a sequence against PROSITE patterns and profiles - quick scan  
(Output includes graphical view and feature detection)

Enter your sequence or a UniProtKB (Swiss-Prot or TrEMBL) ID or AC [ [help](#) ]

ScanProsite - advanced scan

PRATT - allows to interactively generate conserved patterns from a series of unaligned proteins.

MyDomains - Image Creator - allows to generate custom domain figures.

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

HAMAP High-quality Automated and Manual Annotation of microbial Proteomes

Home Proteomes Families Documents Downloads Links

HAMAP is a system, based on manual protein annotation, that identifies and semi-automatically annotates proteins that are part of well-conserved families or subfamilies: the **HAMAP families**. HAMAP is based on manually created family rules and is applied to bacterial, archaeal and plastid-encoded proteins. ([More details](#) / [Reference](#) / [Disclaimer](#))

UniProtKB/Swiss-Prot Release 57.13 of 19-Jan-2010  
UniProtKB/TrEMBL Release 40.13 of 19-Jan-2010

**HAMAP proteomes**

Archaea (70 proteomes) [AERPE Aeropyrum p.](#) [Proteome info](#) [BLAST search](#)

Bacteria (910 proteomes) [ACAM1 Acaryochloris](#) [Proteome info](#) [BLAST search](#)

Plastids (145 proteomes)

Total 1125 proteomes ([summary statistics](#))

**HAMAP families**

recA, MF\_01633, Iron

Search *The wildcard \* is supported.*

[Browse HAMAP families](#)

**HAMAP tools**

Scan a sequence against the HAMAP families - quick scan  
Enter your sequence or a UniProtKB (Swiss-Prot or TrEMBL) ID or AC.

Matches against the HAMAP families will be displayed.

Scan Clear

**HAMAP-Scan** - advanced scan Scan several sequences or a whole genome (all ORFs) against the HAMAP families, with partial or complete annotation in the UniProtKB/Swiss-Prot format.

**Retrieve sets of characterized proteins from a specific HAMAP organism or a taxonomic group**

[Questions? Comments? Please send us feedback.](#)

# Swiss Model Repository

**SWISS-MODEL Repository**  
Modelling Tools Repository Documentation

[ New Repository Query ]

**Welcome to the SWISS-MODEL Repository**

The SWISS-MODEL Repository is a database of annotated three-dimensional comparative protein structure models generated by the fully automated homology-modelling pipeline SWISS-MODEL.

Example Queries:  
[P23290] [OLGA\_ECOLI] [PF00743503] [NP\_418402] [0126454608] [ENTREZ54401] [Sequence]

P23290

SEARCH

The current release of the SWISS-MODEL-Repository (3.12.3) consists of 3515801 model entries for 2733736 unique sequences in the UniProt database.

**NOTE:** The SWISS-MODEL repository contains theoretically calculated models, which may contain significant errors.

BIOENTRUM  
SIB  
Swiss Institute of Bioinformatics

## Swiss PDB Model Viewer

**SWISS-MODEL Repository**  
[Workspaces] [Repository] [Modelling] [Tools]

**SWISS-MODEL Repository Model Details**

**Model Overview** [-]  
Click on the bars to get more details about individual Models or experimental structures

**Sequence** [-]  
ECOP23 AT3-dependent endonuclease La  
Aeromonas hydrophila subsp. hydrophila DSM 2011  
Database: TrEMBL (Unreviewed) automatically annotated and not reviewed

**Domain** [-]  
Link to: InterPro

AAA  
LON  
LON\_C

**Model 3D Structure** [-]  
Based on template: 1t9a [DML] [PDB] [SCOP] [CATH]  
Sequence identity: 88%  
Residue range: 625 to 783  
Model date: 2008-08-11  
Revision date: 2008-08-11  
[display] [download] [download project]

**Model 3D Structure** [-]  
Based on template: 1t9a [DML] [PDB] [SCOP] [CATH]  
Sequence identity: 88%  
Residue range: 625 to 783  
Model date: 2008-08-11  
Revision date: 2008-08-11  
[display] [download] [download project]

**Sequence Alignment**

TARGET	605	DQVLTG	LAVTEGEL	LTVKQVPG	KKLTITGEL	GEVRESAQA
Irref	593	srvgvqv	lwtvrggll	lctctwvpg	ngklytgsi	qevrqsniq
TARGET		ssssss	ssss	ssss	ssssssss	ssssss
Irref		ssssss	ssss	ssss	ssssssss	ssssss
TARGET	653	AMTVSRK	LGLSEFLE	WVSHVDFE	GAIFEDGFA	GITRATLVE
Irref	641	altrvras	kigispdfe	krdihtvpe	gafpdsqas	qiaxtatve
TARGET		hhhhh	h hh	ss	ssssss	ss
Irref		hhhhh	h hh	ss	ssssss	ss
TARGET	703	ALCRIPVKE	VMTGKTLR	GVLPDGLK	KVLAARDGG	IKQVLIPEN
Irref	691	cltgsprcd	vxtgqtlr	gvvipgik	ekilwrrgy	iktvlipen
TARGET		hhh	ss	ss	ss	hh hhhhhhh
Irref		hhh	ss	ss	ss	hh hhhhhhh
TARGET	753	QSDREIFD	VREKQIVV	DHAEVLSA	L --	
Irref	741	krdielids	viadiilpiv	krseevitia	lqpe	
TARGET		hhh	hh	hhhh	ssss	sshhhhhs
Irref		hhh	hh	hhhh	ssss	sshhhhhs

**PDB**  
Protein Data Bank  
An Information Portal for Biological Macromolecular Structures  
As of Tuesday, Jan 12, 2010 there are 53246 Structures @ 1 PDB Station

Home | About | Help | Contact Us

Home | About | Help | Contact Us

Home | About | Help | Contact Us





## 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
- [MALDI PepQuant](#) - Quantify MALDI peptides (SILAC) from [Phenyx](#) output
- [MSight](#) - Mass Spectrometry Imager
- [plcarver](#) - 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

**WELCOME**

This site features **MASCOT**, a powerful search engine that uses mass spectrometry data to identify proteins from primary sequence databases. To assist you, the help text for Mascot forms a substantial knowledge base concerning protein identification by MS.

If this is your first visit, please check for browser compatibility and read the [initial page](#). If you include results from Mascot in a publication, please cite either [www.matrixscience.com](#) or [Electrophoresis](#), 20(18):3524-5 (1999) (abstract).

We value your feedback and suggestions for new features. If you find any problems, errors, oversights, or just get unexpected results then please let us know.

For information on licensing Mascot for in-house use, please refer to our [Products](#) and [Support](#) pages. For recent news, check [What's New](#).

Matrix Science develops and markets software products which integrate mass spectrometry into bioinformatics. Our interests extend to all aspects of mass spectrometry in the life sciences. Please contact us to discuss:

- Developing new applications
- Consultancy in mass spectrometry and bioinformatics
- Systems analysis and integration

**Collaborations**

Mascot, Mascot-database code from Matrix, developed by Daniel Højrup and David Perkins when working at the former Imperial Cancer Research Fund, and licensed from its technology transfer subsidiary, Cancer Research Technology.

LabVantage Solutions and Matrix Science are working together to develop data management and data mining solutions for proteomics.

We are grateful to the Swiss Institute of Bioinformatics for permission to make Swiss-Prot available on this web site for searching with Mascot.

## MASCOT Search Results

### *(MATRIX SCIENCE)* Mascot Search Results

#### Peptide View

#### MS/MS Fragmentation of QAGLSYIR

Found in [gi2258789](#), H(+)-ATP synthase epsilon-subunit [rat, liver, Peptide Mitochondrial, 50 aa]

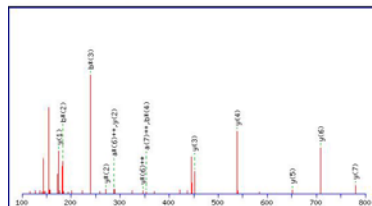
Match to Query 39: 906.598884 from(454 306718,24)

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:\DOCUMENTS-1\MASSTP-1\LOCALS-1\Temp\msa2c5.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 H(calc): 906.49

Ions Score: 47 Expect: 0.043

Hatches (Hold Red): 14/84 Fragment ions using 23 most intense peaks

#	a	a <sup>+</sup>	a <sup>+</sup>	a <sup>++</sup>	b	b <sup>+</sup>	b <sup>++</sup>	Seq	y	y <sup>+</sup>	y <sup>+</sup>	y <sup>++</sup>	#	
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 computationally obtained structures of proteins, protein-DNA and DNA  
(<http://www.rcsb.org/pdb/home/home.do>)

**PDB**  
PROTEIN DATA BANK

An Information Portal to Biological Macromolecular Structures  
As of Tuesday Feb 02, 2010 at 4 PM PST there are 63093 Structures | PDB Statistics

WHAT'S NEW | HELP | PRINT

PDB ID or keyword Search ? | Advanced Search

### 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 [wwwPDB](http://www.pdb.org), 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.

#### Molecule of the Month: Enhanceosome

Take a moment to ponder the form of your body: the shape of your face, the color of your eyes, the length of your fingers, the perfect articulation of your bones and muscles, the way your hair grows curly or straight. Now let your imagination travel inward, and think of the complex shapes and functions of your different cells, and the teeming molecular world inside each one. Remarkably, this amazing structure and form and function is specified by information in the genome, which encodes a mere 20,000-25,000 protein-coding genes. One of the great puzzles being pieced together by scientists is the mechanism by which these genes, and the methods used to control their expression, specify all of these different aspects of life. [Read more ...](#) [Previous Features](#)

#### PSI Featured Molecule: Sugarcoating the surface: yeast Aig13

Many proteins in our cells are decorated with carbohydrate chains, which make the proteins more stable and assist with their function. Using NMR, PSI researchers now understand how this enzyme builds these essential carbohydrates. [Read more from the Structural Genomics Knowledgebase](#) [Previous Features](#)

New user? Try the [browser compatibility check](#) and information on [Getting Started](#).

**News**

- Complete News
- Newsletter
- Discussion Forum
- Job Listings

**wwwPDB Statement on Retraction of PDB Entries**

02-February-2010  
Newsletter Published  
The winter 2010 issue (HTML | PDF) highlights 2009 deposition, release, and access statistics, describes new website features for searching and reporting, and reviews recent outreach activities. [More >>](#)

**wwwPDB FTP Advisory Notice**

The PDB archive can be accessed at FTP sites at the RCSB PDB, PDBe, and PDBj. The update schedules for these sites have been coordinated to be simultaneous. [More >>](#)

**FTP Archive**

The up-to-date PDB archive is available at: <ftp://www.pdb.org>  
Time-stamped yearly snapshots are available at:

**3KBT**

Crystal structure of the ankyrin binding domain of human erythroid beta spectrin (repeats 13-15) in complex with the spectrin binding domain of human erythroid ankyrin (ZUS-ANK)

**Characteristics** Release Date: 02-Feb-2010 Exp. Method: X-RAY DIFFRACTION  
Resolution: 2.75 Å  
Structural Protein

**Classification**

**Compound**

Molecule: Spectrin beta chain, erythrocyte Length: 326  
Polymer: 1 Type: polypeptide(L)  
Chains: A, B  
Fragment: UNP residues 1583-1906

Molecule: Ankyrin-1 Length: 161  
Polymer: 2 Type: polypeptide(L)  
Chains: C, D  
Fragment: UNP residues 911-1068

**Authors** Ipsaro, J.J., Mondragon, A.

Summary | Raw Data | Sequence | Top. Contacts | Literature | Biol. & Chem. | Methods | Geometry | Links

Crystal structure of the ankyrin binding domain of human erythroid beta spectrin (repeats 13-15) in complex with the spectrin binding domain of human erythroid ankyrin (ZUS-ANK)

DOI: 10.2210/pdb3kbt/pdb

**Primary Citation**

Structures of the spectrin-ankyrin interaction binding domains.  
Ipsaro, J.J., Huang, L., Mondragon, A.  
(2009) *Blood* 113: 5385-5392  
PubMed: 19141864 | PubMedCentral: PMC269041 | DOI: 10.1102/Meed-2009-10-184358  
[Search Related Articles in PubMed](#)

**PubMed Abstract:**

As key components of the erythrocyte membrane skeleton, spectrin and ankyrin specifically interact to tether the spectrin cytoskeleton to the cell membrane. The structure of the spectrin binding domain of ankyrin and the ankyrin binding domain of spectrin have been ...  
[Read More & Search PubMed Abstracts](#)

**Molecular Description**

Classification: Structural Proteins  
Structure Weight: 110262.80

Molecule: Spectrin beta chain, erythrocyte Length: 326  
Polymer: 1 Type: polypeptide(L)  
Chains: A, B  
Fragment: UNP residues 1583-1906

Molecule: Ankyrin-1 Length: 161  
Polymer: 2 Type: polypeptide(L)  
Chains: C, D  
Fragment: UNP residues 911-1068

**Source**

Polymer: 1  
Scientific Name: Homo sapiens  
Common Name: Human  
Expression System: Escherichia coli


**Biological Assembly 1**

[View in 3mol](#) [SimpleViewer](#)  
[Other Viewers](#) [Protein Workshop](#)

Biological assembly 1 assigned by authors and generated by PISA (software)

**Deposit/Release Summaries**

Authors: Ipsaro, J.J., Mondragon, A.  
Deposition: 2009-10-20  
Release: 2010-02-02



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

This view is modifiable via the web—requires Java

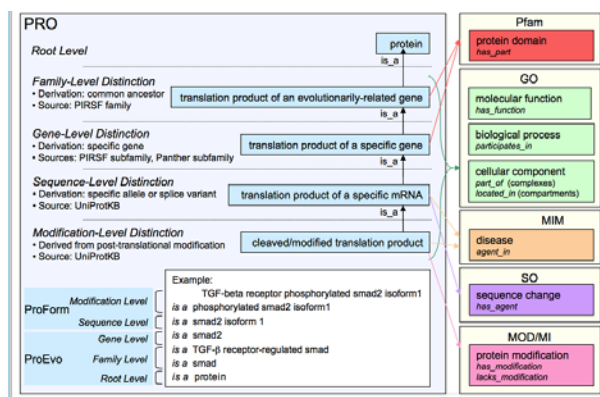
Other very useful visualization and more, tools: Rasmol, ArgusLab, Pymol and Chimera

ATOM	ID	Element	Residue	X	Y	Z	Occupancy	B-factor	Chain
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	32.910	-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 related to proteins supplied by the user



## Protein Ontology, example

Protein Ontology report for entry - PRO:00000017 [Related PRO nodes \(Parent/Child\)](#)

Ontology Information	
PRO ID	PRO:00000017
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:00000001 protein <a href="#">( click to see DAG view. )</a>

This PRO entry has been created based on the following

DB name:ID	PIR#	PF#	(To display the domain architecture, click <a href="#">here</a> for read members; click <a href="#">here</a> for all members.)
PIRSE:PIRSEF001936	PIRSEF001936	PF00714	

Synonym Based Mappings

Db Identifiers UniProtKB:P01579; P01580; HGNC:5438; MGI:107656

Annotation	Modifier	Relation	Ontology ID	Ontology Terms	Relative_to	Interaction With	Evidence Source	Evidence Code	Taxon ID	Inferred From
Domain		has_part	PIR:PF00714	Interferon gamma			PIRSE:PIRSEF001936	ISS		PIRSE:PIRSEF001936
Functional Annotation		participates_in	GO:0006953	immune response			PIRSE:PIRSEF001936	ISS		PIRSE:PIRSEF001936
		participates_in	GO:0045080	positive regulation of chemokine biosynthetic process			PIRSE:PIRSEF001936	ISS		PIRSE:PIRSEF001936

## PIR—ProClass— ID Mapping

[HOME](#) / [Search](#) / [ID Mapping](#)

**ID Mapping Form**

Map a batch of IDs in the *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:

4. Display format:

One to many (070507 29840776;3168870)

One to one (070507 29840776  
070507 3168870)

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

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

**Family: Actin (PF00022)**

36 architectures 7735 sequences 15 interactions 2067 species 299 structures

**Summary**

**Actin** [Add annotation](#)

No Pfam abstract.

**Literature references**

- Schutt CE, Myosik JC, Rozyczki MD, Goonesekere NC, Lindberg U, . Nature 1993;365:810-816.: The structure of crystalline profilin-beta-actin. [PUBMED:2413692](#)
- Shetlerline P, Clayton J, Sparrow J, . Protein Profile 1995;2:1-103.: Actin. [PUBMED:8548558](#)

**InterPro entry** [IPR004000](#)

Actin [PUBMED:1288079](#) [PUBMED:8449070](#) is a ubiquitous protein involved in the formation of filaments that are major components of the cytoskeleton. These filaments interact with myosin to produce a sliding effect, which is the basis of muscular contraction and many aspects of cell motility, including cytokinesis. Each actin protomer binds one molecule of ATP and has one high affinity site for either calcium or magnesium ions, as well as several low affinity sites. Actin exists as a monomer in low salt concentrations, but filaments form rapidly as salt concentration rises, with the consequent hydrolysis of ATP. Actin from many sources forms a tight complex with desferrioxalase (DNase I) although the significance of this is still unknown. The formation of this complex results in the inhibition of DNase I activity, and actin loses its ability to polymerise. It has been shown that an ATPase domain of actin shares similarity with ATPase domains of hexokinase and hsp70 proteins [PUBMED:1628289](#) [PUBMED:1232228](#).

In vertebrates there are three groups of actin isoforms: alpha, beta and gamma. The alpha actins are found in muscle tissues and are a major constituent of the contractile apparatus. The beta and gamma actins co-exists in most cell types as components of the cytoskeleton and as mediators of internal cell motility. In plants there are many isoforms which are probably involved in a variety of functions such as cytoplasmic streaming, cell shape determination, tip growth, graviperception, cell wall deposition, etc.

Recently some divergent actin-like proteins have been identified in several species. These proteins include cetractin (actin-RPV) from mammals, fungi yeast ACT5, *Neurospora crassa* (nc-4) and *Drosophila* *carina*, which seems to be a component of a multi-subunit centrosomal complex involved in microtubule based vesicle motility (this subfamily is known as ARP1) ARP2 subfamily, which includes chicken ACT1, *Saccharomyces cerevisiae* ACT2, *Drosophila melanogaster* 140, and *Caenorhabditis elegans* atcC, ARP3 subfamily, which includes actin 2 from mammals, *Drosophila* o6b, yeast ACT4 and *Schistosoma mansoni* *smact2*; and ARP4 subfamily, which includes yeast ACT3 and *Drosophila* 13E.

**Clan**

This family is a member of clan **Actin\_ATPase** (CL0108), which contains the following 26 members:

<a href="#">Acetate_kinase</a>	<a href="#">Actin</a>	<a href="#">BovAD_BadPQ</a>	<a href="#">Bov_arc_factor</a>	<a href="#">CmH1_NspB</a>	<a href="#">DDB</a>
<a href="#">DUF1464</a>	<a href="#">DUF1786</a>	<a href="#">EUG</a>	<a href="#">EUGV_C</a>	<a href="#">EUGV_N</a>	<a href="#">Eps</a>
<a href="#">Gimble</a>	<a href="#">GimA1_C039</a>	<a href="#">Glucokinase</a>	<a href="#">hexokinase_1</a>	<a href="#">hexokinase_2</a>	<a href="#">HSP70</a>
<a href="#">Hudant A.1</a>	<a href="#">Hudantouase_A</a>	<a href="#">MreB_Mki</a>	<a href="#">DeoDase_MKc</a>	<a href="#">Dev-Gp65</a>	<a href="#">RDE</a>
<a href="#">HhA</a>	<a href="#">UFP0075</a>				

**Gene Ontology**

**Example structure**

[PDB entry 1aen](#): COMPLEX BETWEEN LATRUNCULIN A-RABBIT MUSCLE ALPHA-ACTIN-HUMAN MUSCULIN DOMAIN I

[View a different structure.](#) [1aen](#)

## 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: Kyoto Encyclopedia of Genes and Genomes**

A grand challenge in the post-genomic era is a complete computer representation of the cell, the organism, and the biosphere, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and molecular information. Towards this end we have been developing a bioinformatics resource named KEGG as part of the research projects of the Kanehisa Laboratories in the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo.

**Main entry point to the KEGG web service**  
 KEGG2

**Data-oriented entry points**

<b>KEGG PATHWAY</b>	Pathway maps and pathway modules	<a href="#">Pathway maps</a>
<b>KEGG BRITTE</b>	Functional hierarchies and ontologies	<a href="#">Britte hierarchies</a>
<b>KEGG ORTHOLOGY</b>	KO system and ortholog annotation	
<b>KEGG GENES</b>	Genomes, genes, and proteins	
<b>KEGG LIGAND</b>	Chemical compounds, glycans, and reactions	
<b>KEGG DISEASE</b>	Human diseases	<a href="#">Article in NAR DB issue</a>
<b>KEGG DRUG</b>	Drugs	

**Organism-specific entry points**

**KEGG Organisms** Select:   (example) hsa

**Other entry points**

<b>KEGG Atlas</b>	New interface to navigate pathway maps
<b>KEGG GLYCAN</b>	Glycome informatics resource
<b>KEGG COMPOUND</b>	Knowledge base for biochemical compounds
<b>KEGG REACTION</b>	Knowledge base for biochemical reactions
<b>KEGG PLANT</b>	Knowledge base for plant natural products
<b>KAAS</b>	KEGG automatic annotation server

**KEGG Home**  
[Introduction](#)  
[Overview](#)  
[Release notes](#)  
[Current statistics](#)

**KEGG Identifiers**

**KEGG XML**

**KEGG API**

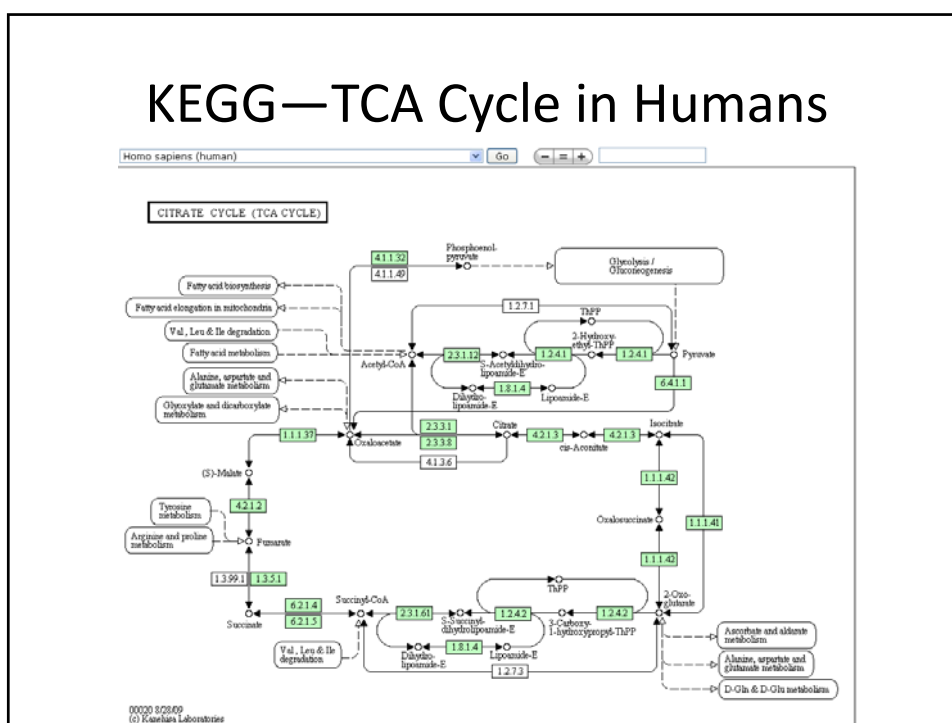
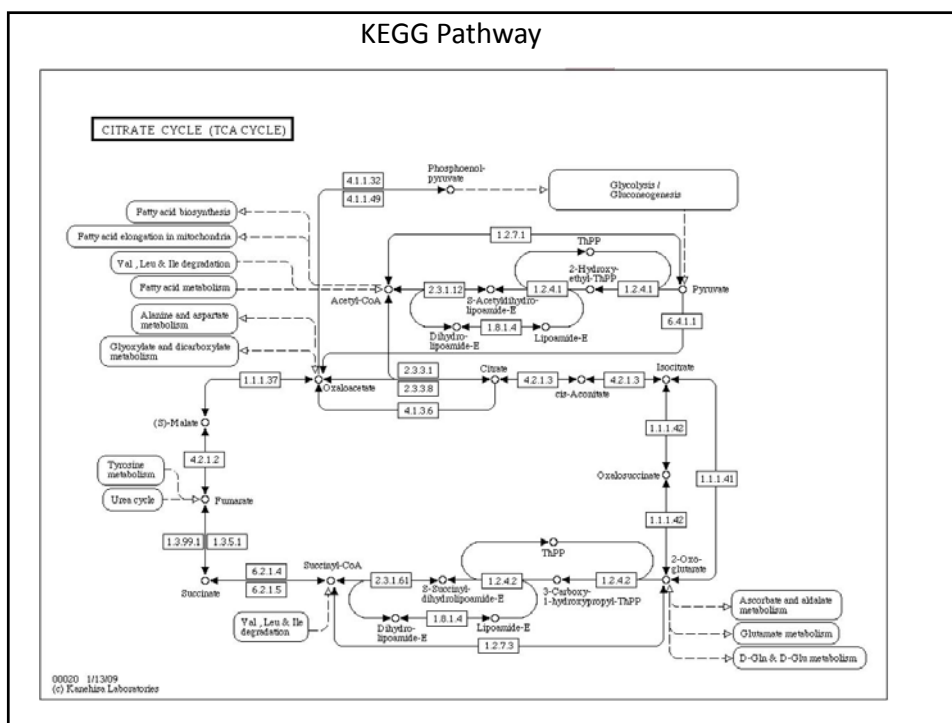
**KEGG FTP**

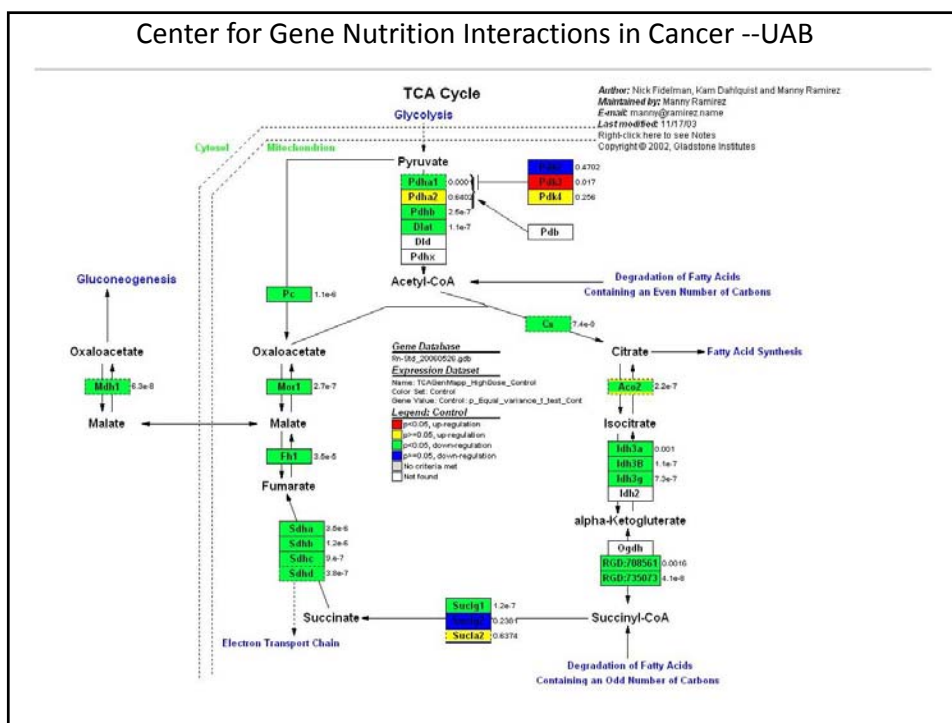
**KegTools**

**GenomeNet**

**DBGETA/LinkDB**

**Feedback**





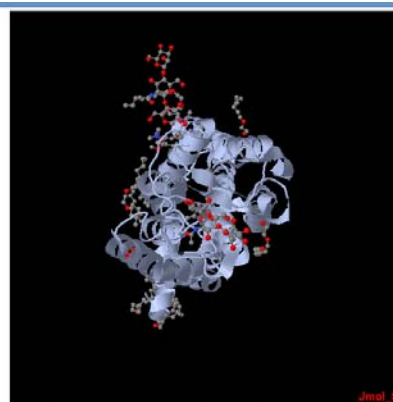
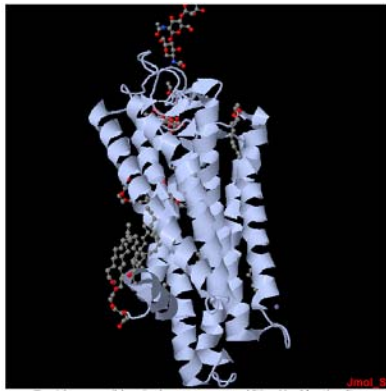
Information Exchange!!

## 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



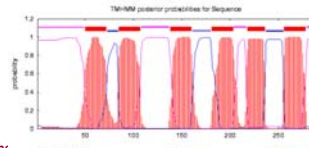
Structurally, a GPCR and (presumably) an OR contains seven helical regions connected by six intra-helical loops, and an N- and C-terminus in the extended (loop like) conformation. The N-terminus is extra-cellular; the C-terminus is intra-cellular

## 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://bjp.weizmann.ac.il/cgi-bin/HORDE/showGene.pl?key=symbol&value=OR1E3P>):**
  - ATGATGAAGAAGAACCAACATGATCTCAGAGTTCTGCTCTGGGCCTTCCATCCAACCTGAGCAGCAGAATCTGTTCTATGCCTTGTCTTGGCCGTGATCTTACCACCTCTGGGAACTCTCTGCTATTGTCTCATTGACTGGACTCCACCTCCACATGCCTATGTAATTTGTCTCAGCAACTGTCTCTGACCTCTGCTTTTCTCGGTCAAAATGCCCCAATTCTGTGAGAACATGGAGAGCCAAACCCATCCATCCCTTTGGGGAGCTGCTGGCTCAGATGTACTTTCATCTGTTTATGGAATTTCTGGAAGCTTCTCTTGGTCTATGGCTTATCACTGCTATGGCTATTGCTTCTCTGCACACACCCTATCATGAGCCCAAGTGTGCTTGGCTGTGACACTCTCTGGCTGTGACCACTGCCATGCCAGTTGCCACACTTGTCTATGGCCAGGCTGTCTTTTGTCTGAGAAATGTGATTCCTCACTTTTCTGTGATACATCACTGTTGAAGCTGGCTCTCCAAACCCAGTCAATGGGTGGGTGATTTTTTATGGGGGGGCTCATCTTGTATCCATCTCACTCTCTCATGTCCTGTGCAAGAACTGTCTCCACCATCTCAGGGTCCCTTCCACTGGGGGATCCAGAAGCTTCTCCACCTGTGGCCCACTCTCTGTGTCTCTCTCTATGGGACAATATTGGTCTTACTTGTGCCATTGACGAATCAAACTGTGAAGGACACTGCATGGCTGTATGACACTGGGGTGACCCACATGCTGAACCCCTCATCTACAGCTGAGGAACAGAGACATGAGGGGAAACCTGGGAGAGTCTCAGCACAAAGAAAATTTTTGTCTTAAAAAGTAAATAGTTGGCATTACCGTATTGAAT
- **Intuitively Translated as:**
  - MMKKNQTMISEFLLGL/PIQPEQNLFYALFLAVYLLTLLGNLLVILRLDHLHMPMYLCLSNLSFSDLCFSSVTMPKLLQNMOSQNPSPFADCLQAMYFHLFYGVLESFLVVMAYHCYVAICFPLHYTTIMSPKCCLGLLTSWLLTAHATLHLLMARLSFCAENWIPHFFCDSTLLKACSNQVNGWVWFFMGGILVIFLLIMSCARIVSTILRVPSFGGIQAFSTCGPHLSVSLFYGTIGLYLCLPLTNHNTVKTVMAMVYGVTHMLNPFYSLRNRMRGNPQSLQHKENFEYFKVIVGILPLLN
- **A two nucleotide frame shift however results in a functional protein with the following sequence :**
  - MPMYLCNLSFSDLCFSSVTMPKLLQNMOSQNPSPFADCLQAMYFHLFYGVLESFLVVMAYHCYVAICFPLHYTTIMSPKCCLGLLTSWLLTAHATLHLLMARLSFCAENWIPHFFCDSTLLKACSNQVNGWVWFFMGGILVIFLLIMSCARIVSTILRVPSFGGIQAFSTCGPHLSVSLFYGTIGLYLCLPLTNHNTVKTVMAMVYGVTHMLNPFYSLRNRMRGNPQSLQHKENFEYFKVIVGILPLLN
- 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 the intracellular loop 2) of TM3. The directionality of this TM1 is extracellular to intracellular. This correctly positions the DRY region of the TM intracellularly—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

