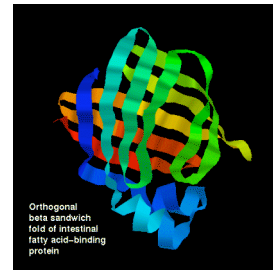
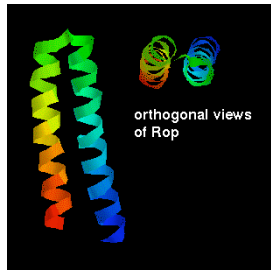
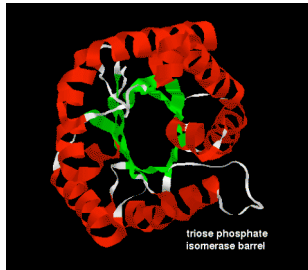


# Protein Structure Analysis & Protein-Protein Interactions



David Wishart

University of Alberta, Edmonton, Canada

[david.wishart@ualberta.ca](mailto:david.wishart@ualberta.ca)

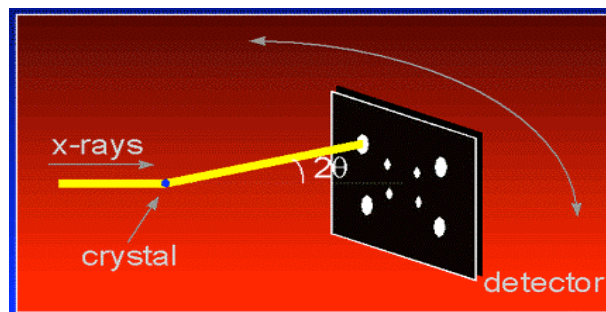
## Much Ado About Structure

- Structure ↔ Function
- Structure ↔ Mechanism
- Structure ↔ Origins/Evolution
- Structure-based Drug Design
- Solving the Protein Folding Problem

## Routes to 3D Structure

- X-ray Crystallography (the best)
- NMR Spectroscopy (close second)
- Cryoelectron microscopy (distant 3rd)
- Homology Modelling (sometimes VG)
- Threading (sometimes VG)
- Ab initio prediction (getting better)

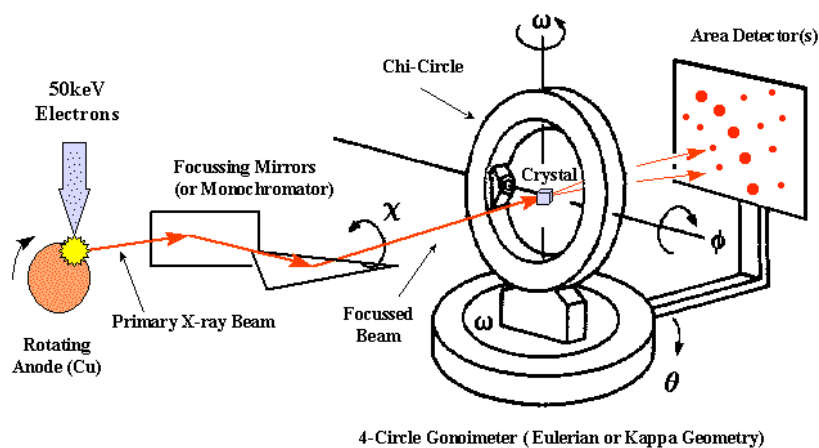
## X-ray Crystallography



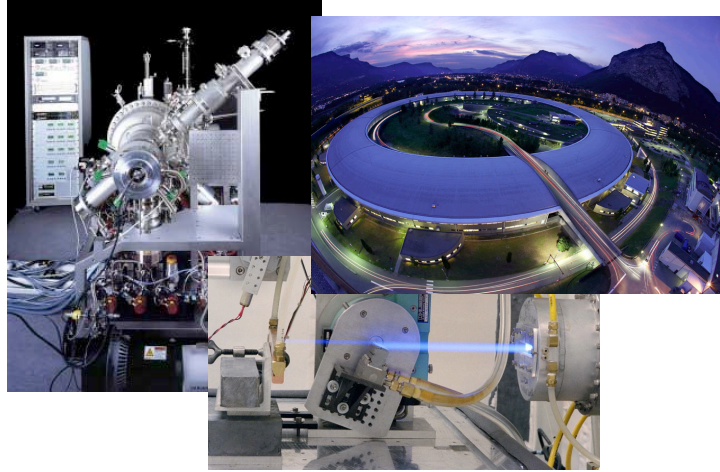
# X-ray Crystallography

- Crystallization
- Crystal Mounting (cryo-mounting)
- Diffraction and Data Collection
- Conversion of Diffraction Data to Electron Density (FT)
- Chain Tracing

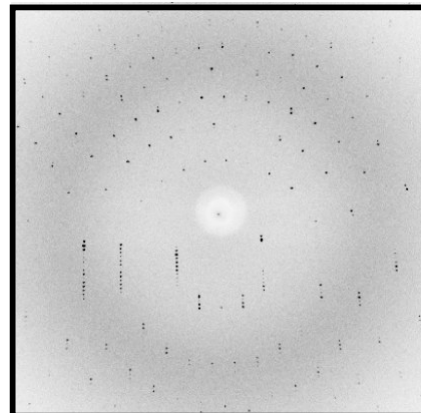
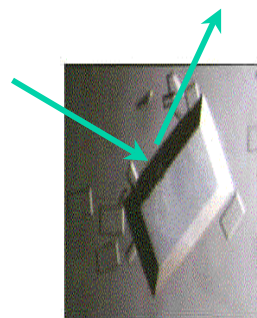
## Diffraction Apparatus



# Synchrotron Diffractometer

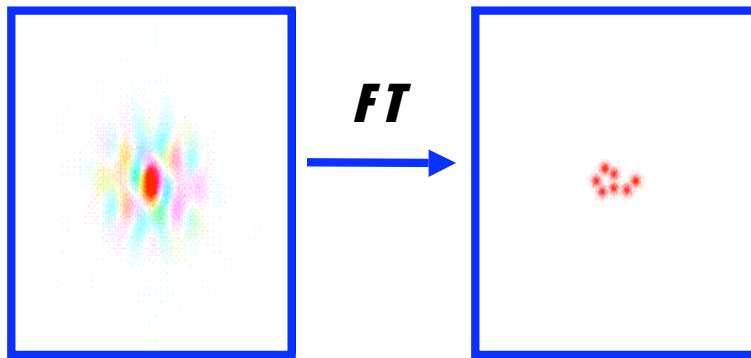


# Protein Crystal Diffraction

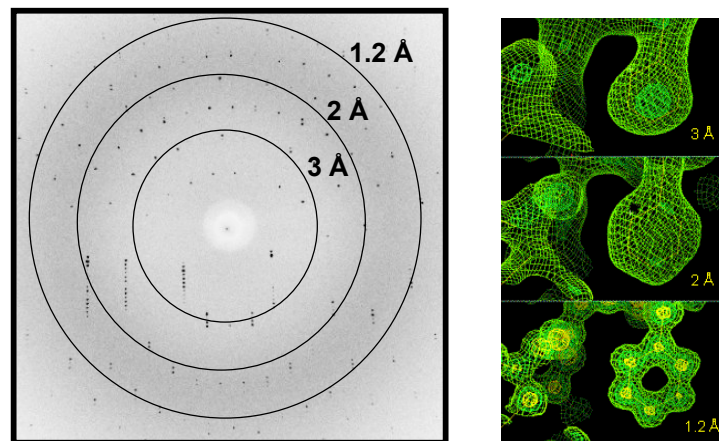


**Diffraction Pattern**

## Converting Diffraction Data to Electron Density



## Resolution

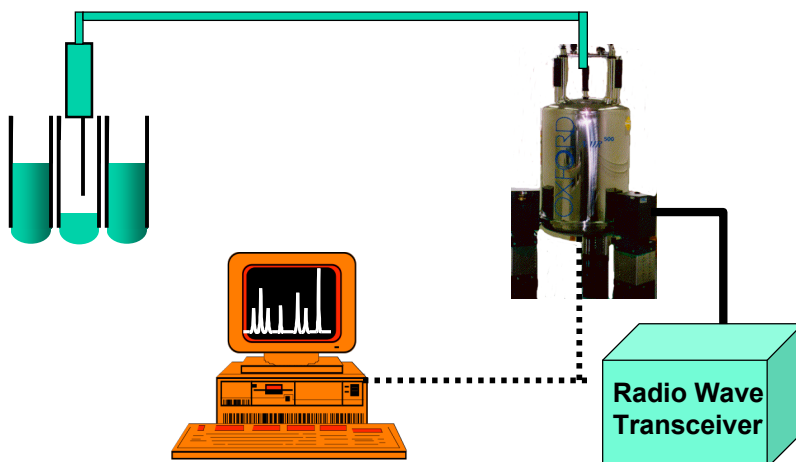


## The Final Result

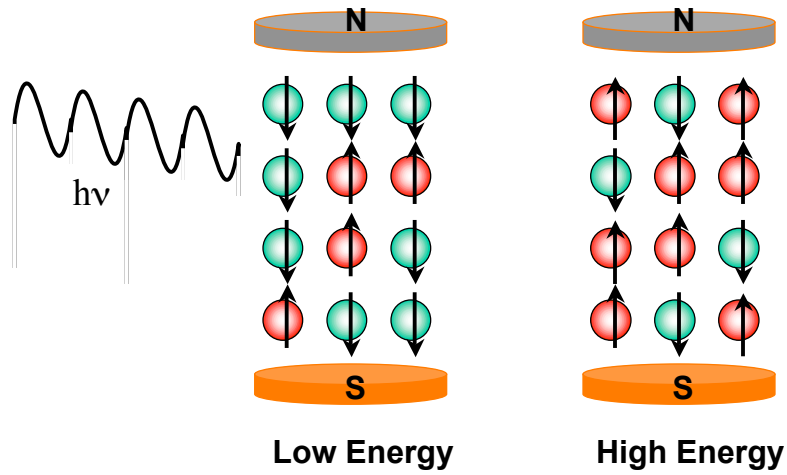
ORIGX2	0.000000	1.000000	0.000000	0.000000		2TRX	147					
ORIGX3	0.000000	0.000000	1.000000	0.000000		2TRX	148					
SCALE1	0.011173	0.000000	0.004858	0.000000		2TRX	149					
SCALE2	0.000000	0.019585	0.000000	0.000000		2TRX	150					
SCALE3	0.000000	0.000000	0.018039	0.000000		2TRX	151					
ATOM	1	N	SER	A	1	21.389	25.406	-4.628	1.00	23.22	2TRX	152
ATOM	2	CA	SER	A	1	21.628	26.691	-3.983	1.00	24.42	2TRX	153
ATOM	3	C	SER	A	1	20.937	26.944	-2.679	1.00	24.21	2TRX	154
ATOM	4	O	SER	A	1	21.072	28.079	-2.093	1.00	24.97	2TRX	155
ATOM	5	CB	SER	A	1	21.117	27.770	-5.002	1.00	28.27	2TRX	156
ATOM	6	OG	SER	A	1	22.276	27.925	-5.861	1.00	32.61	2TRX	157
ATOM	7	N	ASP	A	2	20.173	26.028	-2.163	1.00	21.39	2TRX	158
ATOM	8	CA	ASP	A	2	19.395	26.125	-0.949	1.00	21.57	2TRX	159
ATOM	9	C	ASP	A	2	20.264	26.214	0.297	1.00	20.89	2TRX	160
ATOM	10	O	ASP	A	2	19.760	26.575	1.371	1.00	21.49	2TRX	161
ATOM	11	CB	ASP	A	2	18.439	24.914	-0.856	1.00	22.14	2TRX	162

<http://www.ruppweb.org/xray/101index.html>

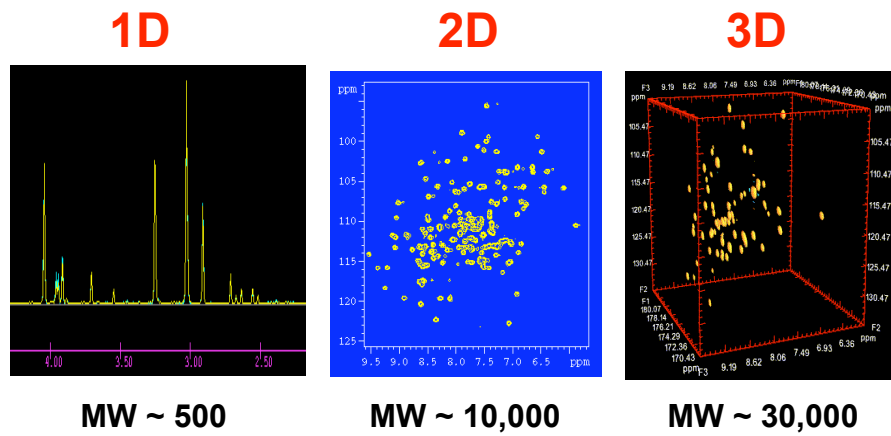
## NMR Spectroscopy



# Principles of NMR



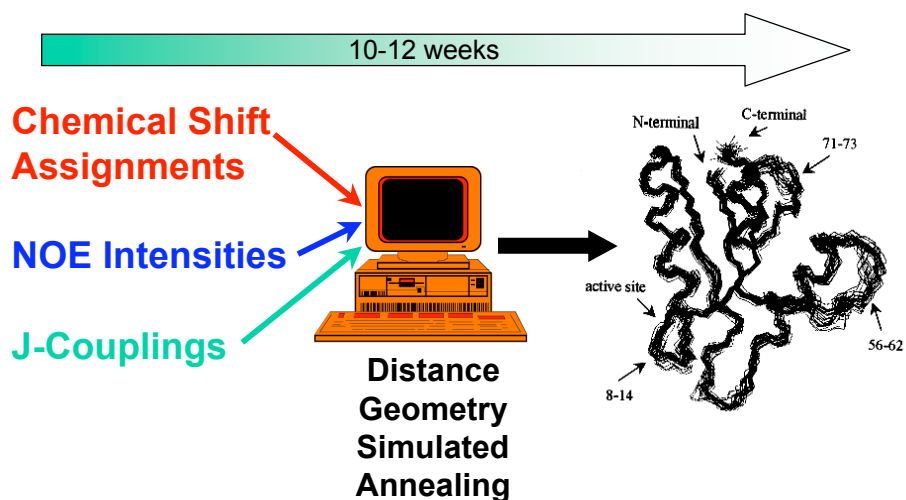
# Multidimensional NMR



## The NMR Process

- Obtain protein sequence
- Collect TOCSY & NOESY data
- Use chemical shift tables and known sequence to assign TOCSY spectrum
- Use TOCSY to assign NOESY spectrum
- Obtain inter and intra-residue distance information from NOESY data
- Feed data to computer to solve structure

## NMR Spectroscopy



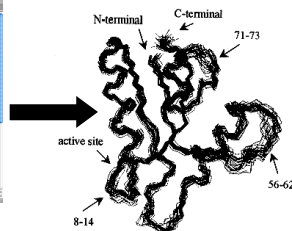
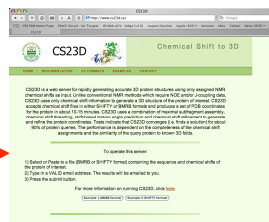


# NMR Spectroscopy

**New!**

10-12 minutes

**Chemical Shift Assignments**



**CS23D**

<http://www.cs23d.ca>

## The Final Result

ORIGX2	0.000000	1.000000	0.000000	0.000000	2TRX	147						
ORIGX3	0.000000	0.000000	1.000000	0.000000	2TRX	148						
SCALE1	0.011173	0.000000	0.004858	0.000000	2TRX	149						
SCALE2	0.000000	0.019585	0.000000	0.000000	2TRX	150						
SCALE3	0.000000	0.000000	0.018039	0.000000	2TRX	151						
ATOM	1	N	SER	A	1	21.389	25.406	-4.628	1.00	23.22	2TRX	152
ATOM	2	CA	SER	A	1	21.628	26.691	-3.983	1.00	24.42	2TRX	153
ATOM	3	C	SER	A	1	20.937	26.944	-2.679	1.00	24.21	2TRX	154
ATOM	4	O	SER	A	1	21.072	28.079	-2.093	1.00	24.97	2TRX	155
ATOM	5	CB	SER	A	1	21.117	27.770	-5.002	1.00	28.27	2TRX	156
ATOM	6	OG	SER	A	1	22.276	27.925	-5.861	1.00	32.61	2TRX	157
ATOM	7	N	ASP	A	2	20.173	26.028	-2.163	1.00	21.39	2TRX	158
ATOM	8	CA	ASP	A	2	19.395	26.125	-0.949	1.00	21.57	2TRX	159
ATOM	9	C	ASP	A	2	20.264	26.214	0.297	1.00	20.89	2TRX	160
ATOM	10	O	ASP	A	2	19.760	26.575	1.371	1.00	21.49	2TRX	161
ATOM	11	CB	ASP	A	2	18.439	24.914	-0.856	1.00	22.14	2TRX	162

<http://www.cryst.bbk.ac.uk/PPS2/projects/schirra/html/home.htm>

# X-ray Versus NMR

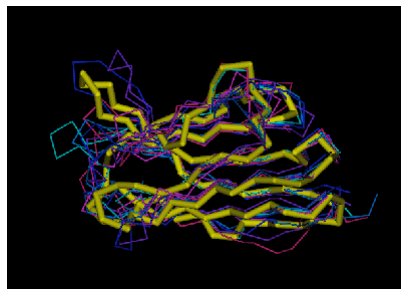
## X-ray

- Producing enough protein for trials
- Crystallization time and effort
- Crystal quality, stability and size control
- Finding isomorphous derivatives
- Chain tracing & checking

## NMR

- Producing enough labeled protein for collection
- Sample “conditioning”
- Size of protein
- Assignment process is slow and error prone
- Measuring NOE’s is slow and error prone

# Comparative (Homology) Modelling



```

ACDEFAGHBIKCLMDNEPFQGRHSITJ--FGKHLQMWNEORPTQ-----TYRRESWTYUEVGHWADXS
ASADEBYCADHLERFILGLDHPIQJRKSTLVMAYNAYOEP--KSQFRAPSPGTSUFKVWEWYXEAYHZADAAS
MCADEBYCADHLERLFLMGNHPEIRJSTKVLAGMGHNQOWPEQRRTS-----GSTFUKEVWWYXAAYHZADAAD
  
```

## Homology Modelling

- Offers a method to “Predict” the 3D structure of proteins for which it is not possible to obtain X-ray or NMR data
- Can be used in understanding function, activity, specificity, etc.
- Of interest to drug companies wishing to do structure-aided drug design
- A keystone of Structural Proteomics

## Homology Modelling

- Identify homologous sequences in PDB
- Align query sequence with homologues
- Find Structurally Conserved Regions (SCRs)
- Identify Structurally Variable Regions (SVRs)
- Generate coordinates for core region
- Generate coordinates for loops
- Add side chains (Check rotamer library)
- Refine structure using energy minimization
- Validate structure

# Modelling on the Web

- Prior to 1998 homology modelling could only be done with commercial software or command-line freeware
- The process was time-consuming and labor-intensive
- The past few years has seen an explosion in automated web-based homology modelling servers
- Now anyone can homology model!

The screenshot displays the Swiss-Model website interface. At the top, the title "Swiss-Model" is prominently displayed in blue. Below it, a navigation menu lists various services: "Modeling requests" (including First Approach, Alignment Interface, Project, Oligomer modelling, and gPCR), "Model Database" (SWISS-MODEL Repository), and "Interactive tools" (SWISS-MODEL Workspace, DeepView, Swiss-PdbViewer, Lookup ExPDB, Search, Examples, and ANOLEA). The main content area features the Swiss-Model logo with a "15 years" anniversary mark, the title "An Automated Comparative Protein Modelling Server", and a brief description of the server's purpose. It also includes a "History" section detailing the server's development since 1993 and an "Acknowledgements" section. A "HELP" section is visible at the bottom left, listing frequently asked questions and other resources. The URL <http://swissmodel.expasy.org//SWISS-MODEL.html> is shown at the bottom of the page.

# 3D-Jigsaw

Warning: You must provide a valid E-mail address to retrieve the results of your query.

Your name

Your E-Mail Address

Your E-Mail Address (verification)

Protein identifier  Automatic  Interactive! Split your sequence into domains, choose the modelling templates and edit the alignments

**3D-JIGSAW**

Protein amino acid sequence in one letter code

Please Note: If you need to submit a large number of jobs to this server, please [contact us](#) first.

(NEW) You can now try the latest version The computing time is significantly longer but the results should be even better!

[Home](#) [Submission](#) [Help](#) [Cite Us](#) [Links](#) [Contact Us](#) [Disclaimer](#) CANCER RESEARCH UK

<http://bmm.cancerresearchuk.org/~3djigsaw/>

## The Final Result

ORIGX2	0.000000	1.000000	0.000000	0.000000	2TRX	147						
ORIGX3	0.000000	0.000000	1.000000	0.000000	2TRX	148						
SCALE1	0.011173	0.000000	0.004858	0.000000	2TRX	149						
SCALE2	0.000000	0.019585	0.000000	0.000000	2TRX	150						
SCALE3	0.000000	0.000000	0.018039	0.000000	2TRX	151						
ATOM	1	N	SER	A	1	21.389	25.406	-4.628	1.00	23.22	2TRX	152
ATOM	2	CA	SER	A	1	21.628	26.691	-3.983	1.00	24.42	2TRX	153
ATOM	3	C	SER	A	1	20.937	26.944	-2.679	1.00	24.21	2TRX	154
ATOM	4	O	SER	A	1	21.072	28.079	-2.093	1.00	24.97	2TRX	155
ATOM	5	CB	SER	A	1	21.117	27.770	-5.002	1.00	28.27	2TRX	156
ATOM	6	OG	SER	A	1	22.276	27.925	-5.861	1.00	32.61	2TRX	157
ATOM	7	N	ASP	A	2	20.173	26.028	-2.163	1.00	21.39	2TRX	158
ATOM	8	CA	ASP	A	2	19.395	26.125	-0.949	1.00	21.57	2TRX	159
ATOM	9	C	ASP	A	2	20.264	26.214	0.297	1.00	20.89	2TRX	160
ATOM	10	O	ASP	A	2	19.760	26.575	1.371	1.00	21.49	2TRX	161
ATOM	11	CB	ASP	A	2	18.439	24.914	-0.856	1.00	22.14	2TRX	162

# The PDB

- PDB - Protein Data Bank
- Established in 1971 at Brookhaven National Lab (7 structures)
- Primary archive for macromolecular structures (proteins, nucleic acids, carbohydrates – now 50,000 structures)
- Moved from BNL to RCSB (Research Collaboratory for Structural Bioinformatics) in 1998

# The PDB

The screenshot shows the RCSB Protein Data Bank homepage. At the top, it says "RCSB Protein Data Bank" and "A MEMBER OF THE PDB". Below that, it says "An Information Portal to Biological Macromolecular Structures" and "As of Tuesday Apr 01, 2008 there are 49974 Structures". The page has a navigation menu on the left with links like Home, Getting Started, Download Files, Deposit and Validate, Structural Genomics, Dictionaries & File Formats, Software Tools, General Education, Site Tutorials, BioSync, General Information, Acknowledgments, Frequently Asked Questions, and Report Bugs/Comments. The main content area has a yellow banner with the text "Are you missing data updates? The PDB archive has moved to ftp://ftp.wwpdb.org. For more information click here." Below this is a "Welcome to the RCSB PDB" section with a paragraph about the RCSB PDB's mission and a "Molecule of the Month: Adrenergic Receptors" section featuring a 3D model of the receptor. On the right, there is a "News" section with links to Complete News, Newsletter, Discussion Forum, and Job Listings, and a "Molecule of the Month" section with a 3D model of the adrenergic receptor and text about the 100th installment.

<http://www.rcsb.org/pdb/>

# Viewing 3D Structures

The screenshot shows the RCSB PDB Structure Explorer interface for entry 2TRX. The main content area displays the following information:

- Title:** CRYSTAL STRUCTURE OF THIOREDOXIN FROM ESCHERICHIA COLI AT 1.68 ANGSTROMS RESOLUTION
- Authors:** Katti, S.K., Lemaster, D.M., Eklund, H.
- Primary Citation:** Katti, S.K., Lemaster, D.M., Eklund, H. Crystal structure of thioredoxin from Escherichia coli at 1.68 Å resolution. *Mol Biol* 21(2):267-74, 1990. [Abstract]
- History:** Deposition 1990-03-19 Release 1991-10-15
- Experimental Method:** Type X-RAY DIFFRACTION Data NA
- Parameters:**

Resolution(Å)	R-Value	R-Free	Space Group
1.68	0.165 (obs.)	n/a	C 2 (C 1 2 1)
- Unit Cell:**

Length (Å)	a	b	c
89.50	51.06	60.45	

Angles (°)	alpha	beta	gamma
90.00	113.50	90.00	
- Molecular Description:** Polymer: 1 Molecule: THIOREDOXIN Chains: A,B
- Functional Class:** Electron Transport

The 'Images and Visualization' section on the right shows a 3D ribbon model of the protein. Below it, the 'Display Options' menu is highlighted with a black circle, listing the following options:

- KING
- Jmol
- WebMol
- Protein Workshop
- QuickPDB
- All Images

# KiNG (Kinemage) 1.39

The screenshot shows the KiNG (Kinemage) 1.39 software interface. The main window displays a 3D ribbon model of the protein structure. The interface includes a menu bar (File, Edit, Views, Display, Tools, Help) and a toolbar with various controls. On the right side, there is a 'Channels' panel and a 'Kinemage #1' panel with the following options:

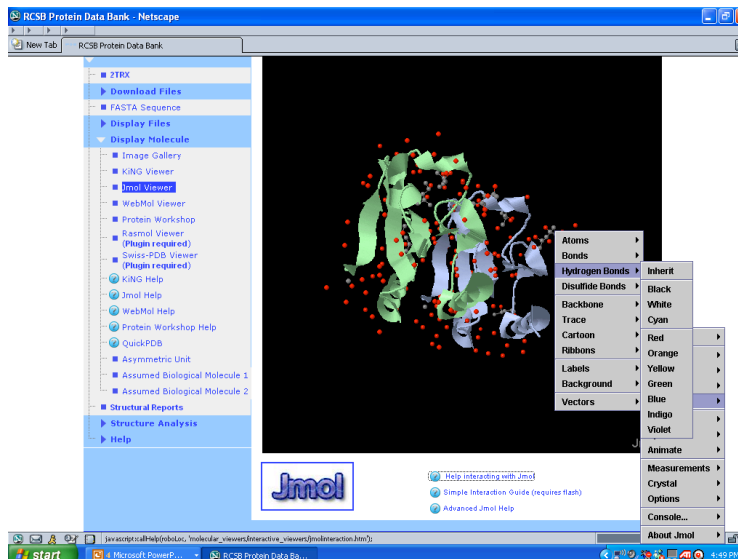
- 2TRXa
- 2TRXb
- 2TRX
- beta
- atoms
- ribbon
- coil
- beta
- alpha

At the bottom of the window, there are controls for 'Zoom' (a slider), 'Clipping', 'Pick center', 'Markers', 'Show text', and 'Show hierarchy'. The status bar at the bottom shows the system tray with the time 4:40 PM.

## KiNG (Kinemage)

- Both a (signed) Java Applet and a downloadable application
- Application is compatible with most Operating systems
- Compatible with most Java (1.3+) enabled browsers including:
  - Internet Explorer (Win32)
  - Mozilla/Firefox (Win32, OSX, \*nix)
  - Safari (Mac OS X) and Opera 7.5.4

## JMol Applet

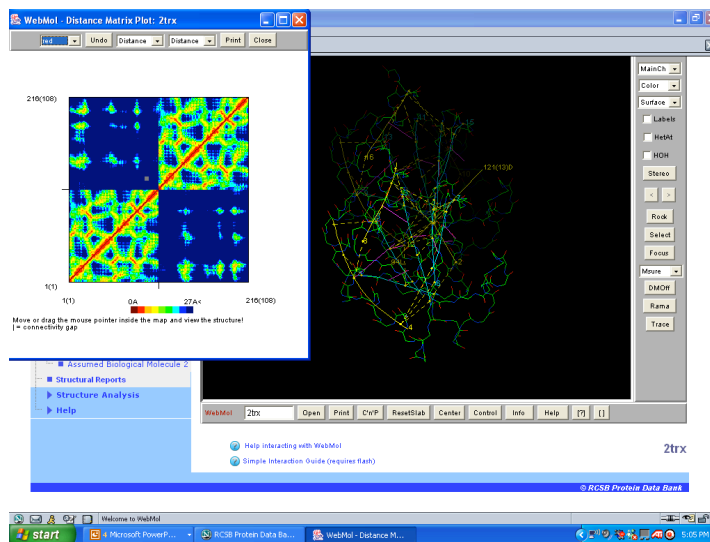




# JMol

- **Java-based program**
- **Open source applet and application**
  - Compatible with Linux, MacOS, Windows
- **Menus access by clicking on Jmol icon on lower right corner of applet**
- **Supports all major web browsers**
  - Internet Explorer (Win32)
  - Mozilla/Firefox (Win32, OSX, \*nix)
  - Safari (Mac OS X) and Opera 7.5.4

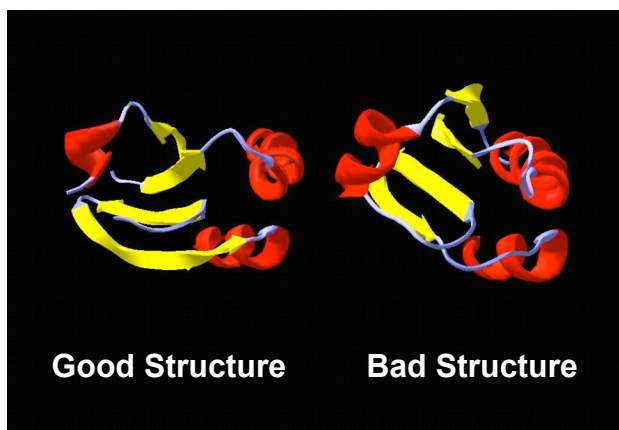
# WebMol



## WebMol

- Both a Java Applet and a downloadable application
- Offers many tools including distance, angle, dihedral angle measurements, detection of steric conflicts, interactive Ramachandran plot, diff. distance plot
- Compatible with most Java (1.3+) enabled browsers including:
  - Internet Explorer 6.0 on Windows XP
  - Safari on Mac OS 10.3.3
  - Mozilla 1.6 on Linux (Redhat 8.0)

## Analyzing and Assessing 3D Structures



## Why Assess Structure?

- A structure can (and often does) have mistakes
- A poor structure will lead to poor models of mechanism or relationship
- Unusual parts of a structure may indicate something important (or an error)

## Famous “bad” structures

- Azobacter ferredoxin (wrong space group)
- Zn-metallothionein (mistraced chain)
- Alpha bungarotoxin (poor stereochemistry)
- Yeast enolase (mistraced chain)
- Ras P21 oncogene (mistraced chain)
- Gene V protein (poor stereochemistry)

## How to Assess Structure?

- Assess experimental fit (look at R factor {X-ray} or rmsd {NMR})
- Assess correctness of overall fold (look at disposition of hydrophobes, location of charged residues)
- Assess structure quality (packing, stereochemistry, bad contacts, etc.)

## A Good Protein Structure..

### X-ray structure

- R = 0.59 random chain
- R = 0.45 initial structure
- R = 0.35 getting there
- R = 0.25 typical protein
- R = 0.15 best case
- R = 0.05 small molecule

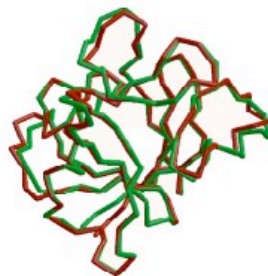
### NMR structure

- rmsd = 4 Å random
- rmsd = 2 Å initial fit
- rmsd = 1.5 Å OK
- rmsd = 0.8 Å typical
- rmsd = 0.4 Å best case
- rmsd = 0.2 Å dream on...

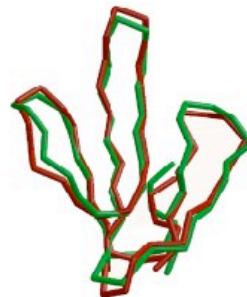
## Cautions...

- A low R factor or a good RMSD value does not guarantee that the structure is “right”
- Differences due to crystallization conditions, crystal packing, solvent conditions, concentration effects, etc. can perturb structures substantially
- Long recognized need to find other ways to ID good structures from bad (not just assessing experimental fit)

## Structure Variability



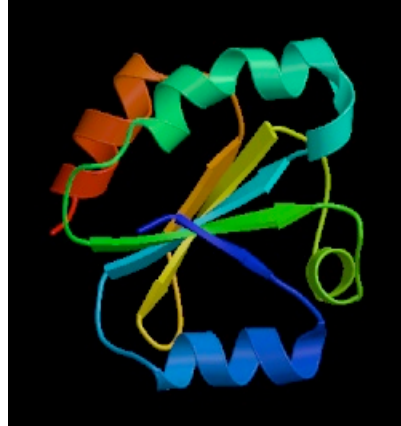
X-ray to X-ray  
Interleukin 1 $\beta$   
(41bi vs 2mlb)



NMR to X-ray  
Erabutoxin  
(3ebx vs 1era)

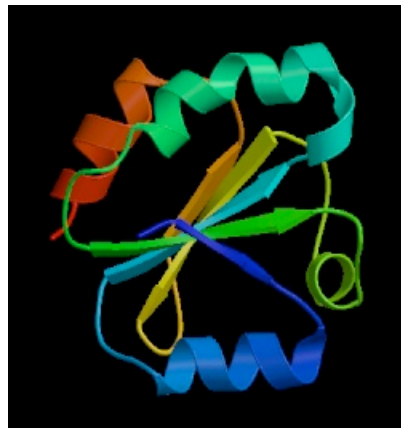
## A Good Protein Structure..

- **Minimizes disallowed torsion angles**
- **Maximizes number of hydrogen bonds**
- **Maximizes buried hydrophobic ASA**
- **Maximizes exposed hydrophilic ASA**
- **Minimizes interstitial cavities or spaces**



## A Good Protein Structure..

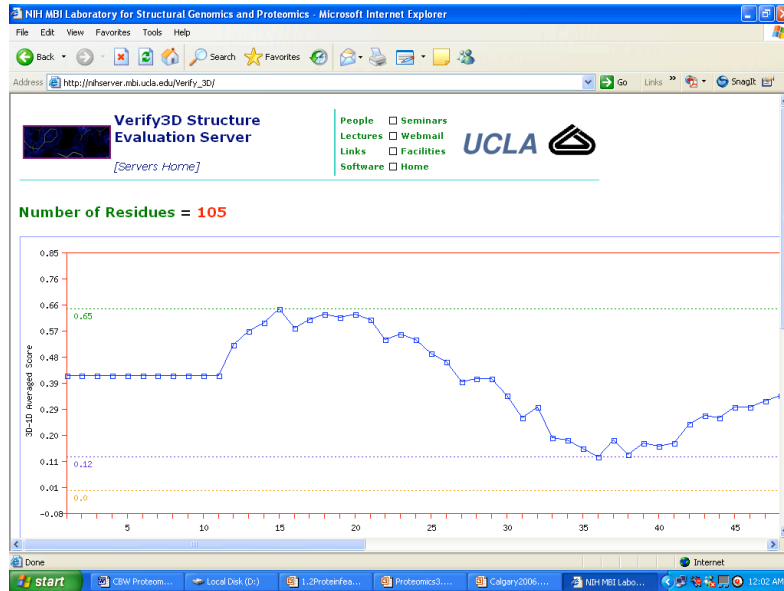
- **Minimizes number of “bad” contacts**
- **Minimizes number of buried charges**
- **Minimizes radius of gyration**
- **Minimizes covalent and noncovalent (van der Waals and coulombic) energies**



# Structure Validation Servers

- **WhatIf Web Server -**  
<http://swift.cmbi.ru.nl/servers/html/index.html>
- **Biotech Validation Suite -**  
<http://biotech.ebi.ac.uk:8400/cgi-bin/sendquery>
- **ProSA-web -**  
<https://prosa.services.came.sbg.ac.at/prosa.php>
- **Verify3D -**  
[http://nihserver.mbi.ucla.edu/Verify\\_3D/](http://nihserver.mbi.ucla.edu/Verify_3D/)
- **VADAR -**  
<http://redpoll.pharmacy.ualberta.ca/vadar/>

The screenshot shows a web browser window titled "The WHAT IF Web Interface - Microsoft Internet Explorer". The address bar displays "http://swift.cmbi.kun.nl/WWW/WTI/". The page content is divided into two main sections. On the left, a pink sidebar titled "Classes" lists various services: Help, Administration, Build/check/repair model, Structure validation, Analyze a residue, Protein analysis, 2-D graphics, 3-D graphics, Hydrogen (bonds), Accessibility, Atomic contacts, Coordinate manipulations, Rotamer related, Cysteine related, Water, Ions, Docking, Crystal symmetry, mutation prediction, NMR, and Other options. On the right, the main content area has a yellow background and is titled "WHAT IF Web Interface". It contains the following text: "Click on one of the classes in the left side column to activate those servers. You can also read the [help page](#), or you can read about the [output formats](#) used by all servers. We also made one [page](#) with notes for people who want to make their own servers. W A R N I N G Results are only kept on this server till Saturday, midnight in The Netherlands. This is mainly a safety feature, but also saves us disk space. Feel free to look at the [WHAT IF writeup](#) too. In case a server fails, first check your input PDB file "List sequence of a PDB file" server that can be found in the "Administration" section of the servers. If this server gives you a long list of administrative information, and no obvious error messages, then look at the [PDBREPORT](#) database to see if there are other obvious problems with the PDB file. If that also does not provide an answer, mail the PDB file (and other files when needed) to Gert Vriend. Please mention the server class (mentioned in the left-hand column of the server page) and the title of the actual server (you click on that in the right-hand part of the server page to get at that server). Last software update on April 30, 2006. Please report any 'buzzes'. If you have detected any errors, or have any question or suggestion, please send us a mail, Roland Krause, Jens Erik Nielsen, [Gert Vriend](#). Last modified Sat Jan 14 22:10:10 2006 by [TN](#)". The browser's taskbar at the bottom shows several open windows, including "start", "CBW Proteom...", "Local Disk (D:)", "1:3Proteinfes...", "Proteomics3...", "Calgary2006...", and "The WHAT IF ...". The system clock shows "12:00 AM".



High scores = good Low scores = bad

# VADAR

VADAR

<http://redpoll.pharmacy.ualberta.ca/vadar/>

VADAR

**VADAR Version 1.5**

Please click [here](#) to do multiple chain analysis  
**Note: VADAR cannot process proteins < 15 residues or > 2000 residues**

VADAR (Volume, Area, Dihedral Angle Reporter) is a compilation of more than 15 different algorithms and programs for analyzing and assessing peptide and protein structures from their PDB coordinate data. The results have been validated through extensive comparison to published data and careful visual inspection. The VADAR web server supports the submission of either PDB formatted files or PDB accession numbers. VADAR produces extensive tables and high quality graphs for quantitatively and qualitatively assessing protein structures determined by X-ray crystallography, NMR spectroscopy, 3D-threading or homology modelling.

Please cite the following: Leigh Willard, Anuj Ranjan, Haiyan Zhang, Hassan Monzavi, Robert F. Hoyko, Brian D. Sykes, and David S. Wishart "VADAR: a web server for quantitative evaluation of protein structure quality" *Nucleic Acids Res.* 2003 July 1; 31 (13): 3316-3319

For additional information on how to run VADAR or to process multiple chains via VADAR, click this button [HELP](#)

Select desired PDB file  no file selected

Note: the uploaded file must be in PDB format in order for this form to work. Refer to the [HELP](#) button above.

OR Enter PDB accession number   
 (Please specify the chain e.g. 2TRXB (2TRX chain B). If not specified, the first chain will be processed. e.g. 2TRX)

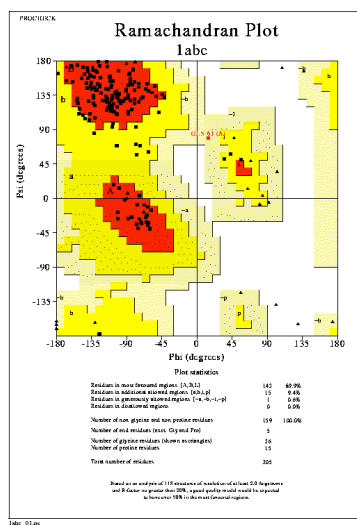
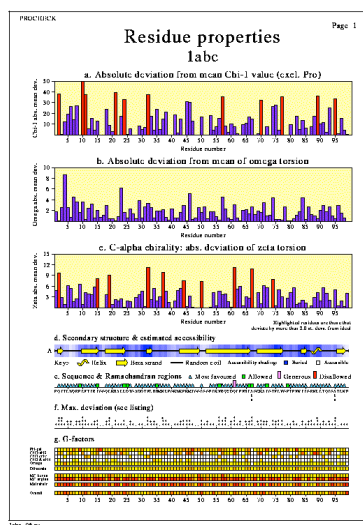
<http://redpoll.pharmacy.ualberta.ca/vadar>



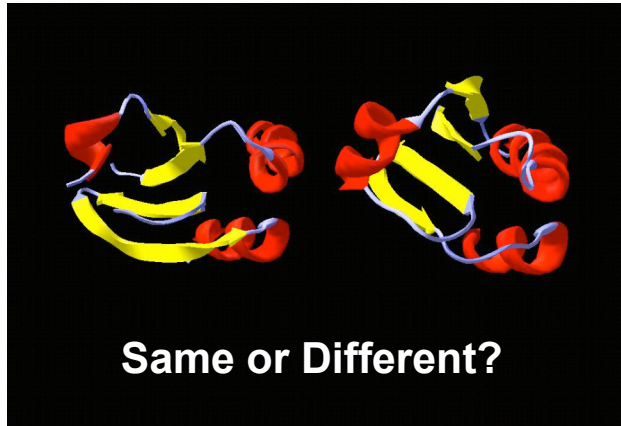
# Structure Validation Programs

- **PROCHECK** -  
<http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html>
- **PROSA II** -  
<https://prosa.services.came.sbg.ac.at/download/download.php>
- **VADAR** -  
<http://www.pence.ualberta.ca/ftp/vadar/>
- **DSSP** -  
<http://swift.cmbi.ru.nl/gv/dssp/index.html>

## Procheck



## Comparing 3D Structures



Qualitative vs. Quantitative

## Rigid Body Superposition



# Superposition

- Objective is to match or overlay 2 or more similar objects
- Requires use of translation and rotation operators (matrices/vectors)
- Least squares or conjugate gradient minimization (McLachlan/Kabsch)
- Lagrangian multipliers
- Quaternion-based methods (**fastest**)

## SuperPose Web Server

SuperPose Version 1.0

SuperPose is a protein superposition server. SuperPose calculates protein superpositions using a modified quaternion approach. From a superposition of two or more protein structures, SuperPose generates sequence alignments, structure alignments, RMS statistics, Difference Distance Plots, and interactive images of the superposition. The SuperPose web server supports the submission of either PDB-formatted accession numbers.

Please cite the following: Rajarshi Maiti, Gary H. Van Domselaar, Haiyan Zhang, and David S. Wishart "SuperPose: a simple server for sophisticated structural superposition" *Nucleic Acids Res.* 2004 July 1; 32 (Web Server issue): W590-W594

If your PDB file contains multiple copies of a structure (i.e. NMR files) you only need to enter one file or accession number. For additional information on how to run SuperPose, click [HELP](#).

PDB Entry A:

Select the first PDB file:

SuperPose Output Images

WebMail  
McLachlan Superposition Image  
Difference Distance Matrix  
SuperPose Output Text Files

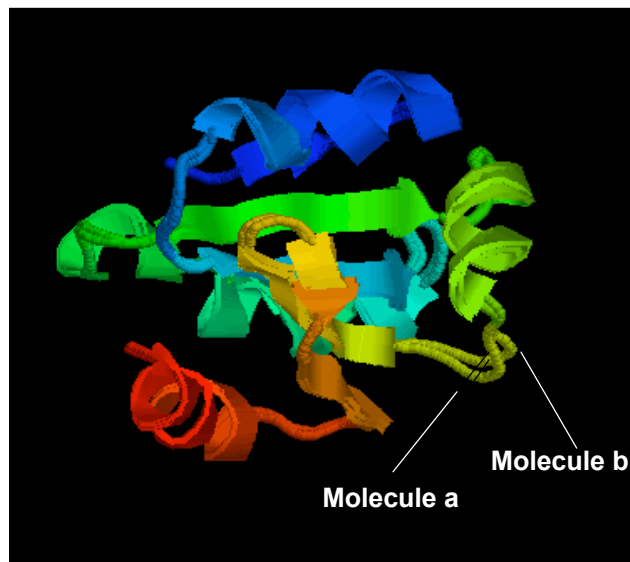
Sequence Alignment  
SuperPose Structure Alignment

<http://wishart.biology.ualberta.ca/SuperPose/>

## Superposition - Applications

- Ideal for comparing or overlaying two or more protein structures
- Allows identification of structural homologues (CATH and SCOP)
- Allows loops to be inserted or replaced from loop libraries (comparative modelling)
- Allows side chains to be replaced or inserted with relative ease

## Measuring Superpositions



## RMSD - Root Mean Square Deviation

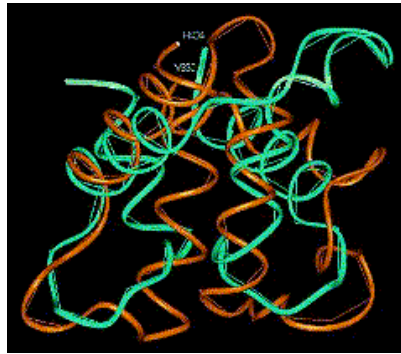
- Method to quantify structural similarity - same as standard deviation
- Requires 2 superimposed structures (designated here as “a” & “b”)
- N = number of atoms being compared

$$\text{RMSD} = \frac{\sqrt{\sum_i (x_{ai} - x_{bi})^2 + (y_{ai} - y_{bi})^2 + (z_{ai} - z_{bi})^2}}{\sqrt{N}}$$

## RMSD

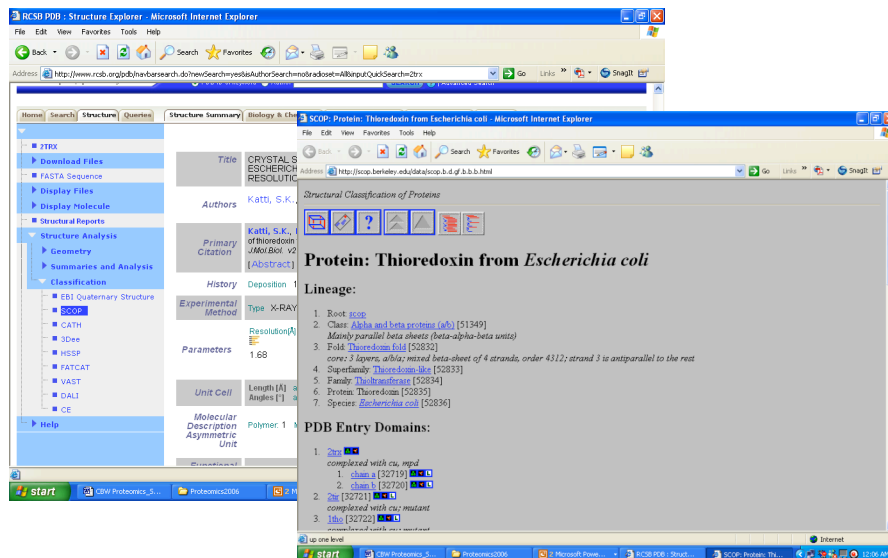
- 0.0-0.5 Å → Essentially Identical
- <1.5 Å → Very good fit
- < 5.0 Å → Moderately good fit
- 5.0-7.0 Å → Structurally related
- > 7.0 Å → Dubious relationship
- > 12.0 Å → Completely unrelated

# Detecting Unusual Relationships



Similarity between Calmodulin and Acetylcholinesterase

# Classifying Protein Folds



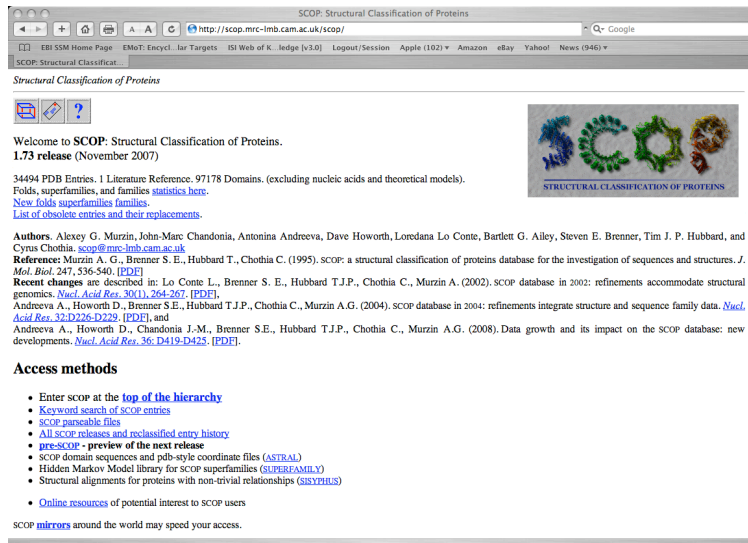
The screenshot shows the RCSB PDB Structure Explorer interface. The main content area displays the structural classification of Thioredoxin from *Escherichia coli*. The title is "Protein: Thioredoxin from *Escherichia coli*". The author is Katti, S.K., and the primary citation is "Katti, S.K., of thioredoxin J.Mol.Biol. v2 [ABSTRACT]". The experimental method is X-RAY, and the resolution is 1.68 Å. The unit cell is described as Polymer 1. The lineage is as follows:

- 1. Root: *scop*
- 2. Class: *Alpha and beta proteins (Ab)* [51349]
- 3. Fold: *Thioredoxin-like* [52833]
- 4. Superfamily: *Thioredoxin-like* [52833]
- 5. Family: *Thioredoxin* [52834]
- 6. Protein: *Thioredoxin* [52835]
- 7. Species: *Escherichia coli* [52836]

The PDB Entry Domains are listed as follows:

- 1. *1ttr* [12719] [MOL]
- 2. *1ttr* [12720] [MOL]
- 3. *1ttr* [12721] [MOL]
- 4. *1ttr* [12722] [MOL]

# SCOP Database



The screenshot shows the SCOP Database homepage in a web browser. The browser's address bar displays the URL <http://scop.mrc-lmb.cam.ac.uk/scop/>. The page title is "SCOP: Structural Classification of Proteins". The main content area includes a welcome message, a 1.73 release date (November 2007), and statistics: 34494 PDB Entries, 1 Literature Reference, and 97178 Domains. It also lists authors, references, recent changes, and access methods. A small graphic with the letters SCOP is visible on the right side of the page.

Welcome to **SCOP: Structural Classification of Proteins**.  
**1.73 release** (November 2007)

34494 PDB Entries, 1 Literature Reference, 97178 Domains. (excluding nucleic acids and theoretical models).  
Folds, superfamilies, and families [statistics here](#).  
[New folds superfamilies families](#).  
[List of obsolete entries and their replacements](#).

**Authors:** Alexey G. Murzin, John-Marc Chandonia, Antonina Andreeva, Dave Howorth, Loredana Lo Conte, Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia. [scop@mrc-lmb.cam.ac.uk](mailto:scop@mrc-lmb.cam.ac.uk)

**References:** Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540. [PDF]

**Recent changes** are described in: Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.* 30(1), 264-267. [PDF].  
Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res.* 32:D226-D229. [PDF], and  
Andreeva A., Howorth D., Chandonia J.-M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucl. Acid Res.* 36: D419-D425. [PDF].

**Access methods**

- Enter scop at the [top of the hierarchy](#)
- [Keyword search of scop entries](#)
- [SCOP parseable files](#)
- [All SCOP releases and reclassified entry history](#)
- [pre-SCOP - preview of the next release](#)
- [SCOP domain sequences and pdb-style coordinate files \(ASTRAL\)](#)
- [Hidden Markov Model library for SCOP superfamilies \(SUPERFAMILY\)](#)
- [Structural alignments for proteins with non-trivial relationships \(SISSYFUS\)](#)

• [Online resources](#) of potential interest to SCOP users

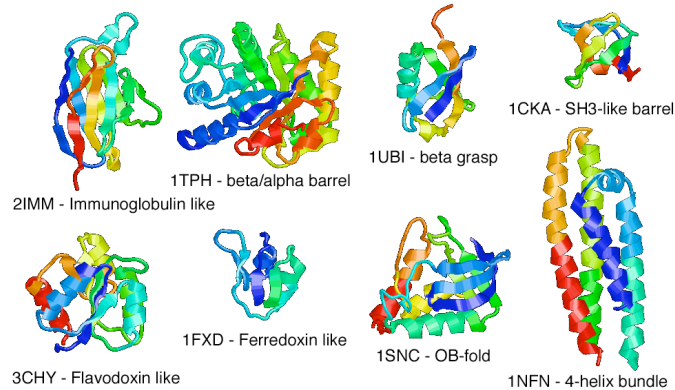
SCOP [mirrors](#) around the world may speed your access.

<http://scop.mrc-lmb.cam.ac.uk/scop>

## SCOP

- **Class** folding class derived from secondary structure content
- **Fold** derived from topological connection, orientation, arrangement and # 2° structures
- **Superfamily** clusters of low sequence ID but related structures & functions
- **Family** clusters of proteins with seq ID > 30% with v. similar struct. & function

# SCOP Structural Classification



The eight most frequent SCOP superfolds

# The CATH Database

CATH Protein Structure Classification Database (UCL)

http://www.cathdb.info/latest/index.html

EBI SSM Home Page EMO: Encycl. Lar Targets ISI Web of K. ledge [v3.0] Logout/Session Apple (102) Amazon eBay Yahoo! News (946)

CATH Protein Structure Cla

**CATH**  
Protein Structure Classification

Home > Top

**Search**

**Go to**

SSAP Server  
CATHEDRAL Server  
DHS  
Gene3D

**Navigation**

Home  
Top of Hierarchy

**CATH Protein Structure Classification**

Version 3.1.0: Released Jan 2007

**CATH Group**

Dr. Alison Cuff, Dr. Ian Silfver, Dr. Mark Dobley, Mr. Tony Lewis, Mr. Oliver Redfern, Dr. Frances M.G. Peart

**Contributors to the CATH Version 3.1.0 Release**

Ms. Sarah Abdou, Mr. Tim Dainman, Mr. Benoit Desnary, Dr. Lesley Greene, Dr. David Lee, Dr. Jon Lees, Dr. Russell L. Mancini, Mr. Adam Field, Mr. Stathis Sotiros, Dr. Corn Yeates, Prof. Janet Thornton, Prof. Christine A. Orengo

**Links**

- Browse or search the classification
- CATH statistics and release information
- General information on CATH
- CATH lists and FTP site
- [NEW]** Raw data files for CATH (including CATH Domain PDB files)
  - Full HMM Library (right-click link and select "Save as...")
  - Concatenated file of 7784 models representing all sequence families in CATH v3.1.0 (zipped HMM32.2 format: 63MB)
- [NEW]** Cross-links between superfamilies in CATH
- DHS - Dictionary of Homologous Superfamilies. Summary of structural and functional features for CATH Homologous Superfamilies
- CATH File Formats (for FTP files)

**Introduction**

**CATH** is a hierarchical classification of protein domain structures, which clusters proteins at four major levels, **Class(C)**, **Architecture(A)**, **Topology(T)** and **Homologous superfamily (S)**.

Class, derived from secondary structure content, is assigned for more than 90% of protein structures automatically. Architecture, which describes the gross orientation of secondary structures, independent of connectivities, is currently assigned manually. The topology level clusters structures into fold groups according to their topological connections and numbers of secondary structures. The homologous superfamilies cluster proteins with highly similar structures and functions. The assignments of structures to fold groups and homologous superfamilies are made by sequence and

**CATH v3.1.0**  
Release statistics

	v3.0.0	v3.1.0	New
<b>Domains</b>	69191	93895	7734
<b>Chains</b>	57741	63453	5712
<b>PDBs</b>	27522	30028	2506

**Technical notes**  
This release has incorporated a great deal of external development including:

- Development of backend ProteoSC, databases
- Development of the central code library
- New web interface for domain clipping (DomClip)
- Improved public pages to show very brief information
- Added numerous maintenance scripts and regression tests

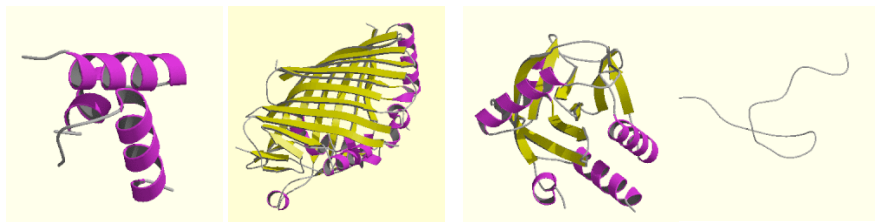
http://www.cathdb.info/latest/index.html



# CATH

- **Class [C]** derived from secondary structure content (automatic)
- **Architecture (A)** derived from orientation of 2° structures (manual)
- **Topology (T)** derived from topological connection and # 2° structures
- **Homologous Superfamily (H)** clusters of similar structures & functions

## CATH - Class



Class 1:  
Mainly Alpha

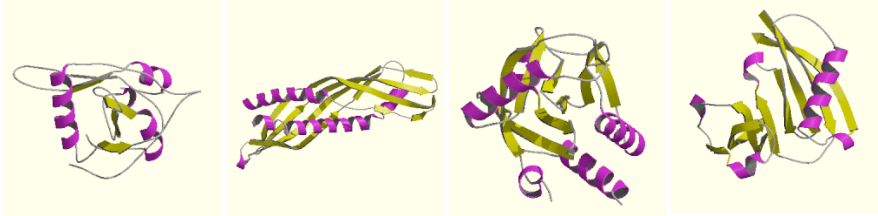
Class 2:  
Mainly Beta

Class 3:  
Mixed  
Alpha/Beta

Class 4:  
Few Secondary  
Structures

**Secondary structure content (automatic)**

## CATH - Architecture



Roll

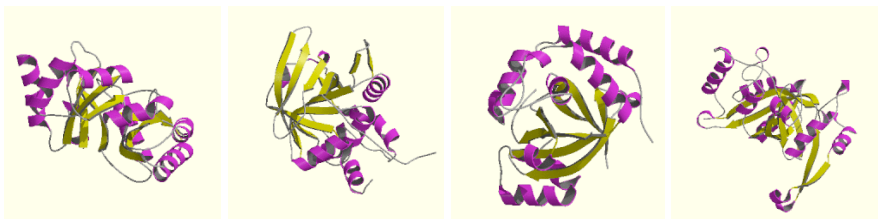
Super Roll

Barrel

2-Layer  
Sandwich

Orientation of secondary structures (manual)

## CATH - Topology



L-fucose Isomerase

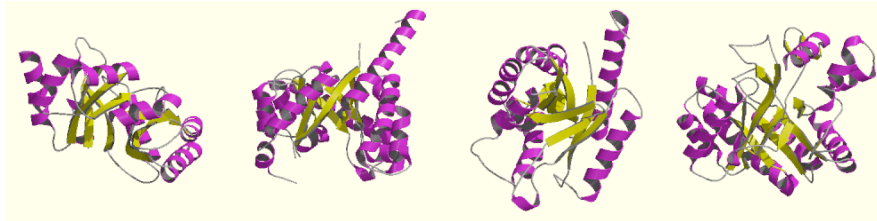
Serine Protease

Aconitase,  
domain 4

TIM Barrel

Topological connection and number of secondary structures

## CATH - Homology



Alanine racemase

Dihydropteroate (DHP) synthetase

FMN dependent fluorescent proteins

7-stranded glycosidases

**Superfamily clusters of similar structures & functions**

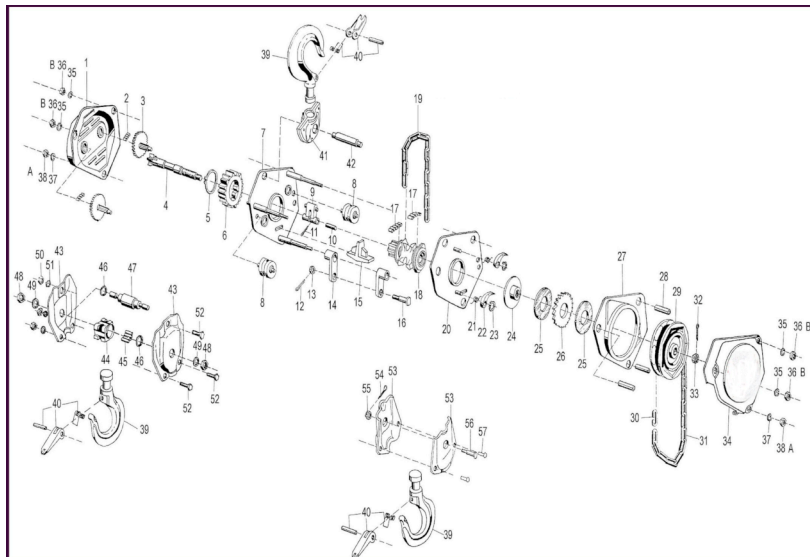
## Other Servers/Databases

- **Dali** - [http://ekhidna.biocenter.helsinki.fi/dali\\_server/](http://ekhidna.biocenter.helsinki.fi/dali_server/)
- **VAST** - [www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml](http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml)
- **CE** - <http://cl.sdsc.edu/ce.html>
- **SSM** - <http://www.ebi.ac.uk/msd-srv/ssm/>
- **PDBsum** - <http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/>

# Protein Interactions



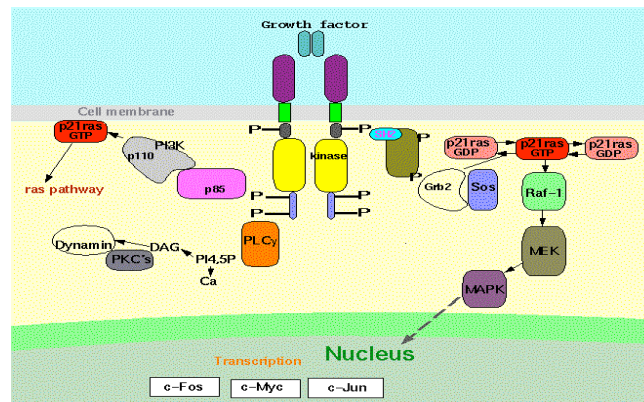
# The Protein Parts List



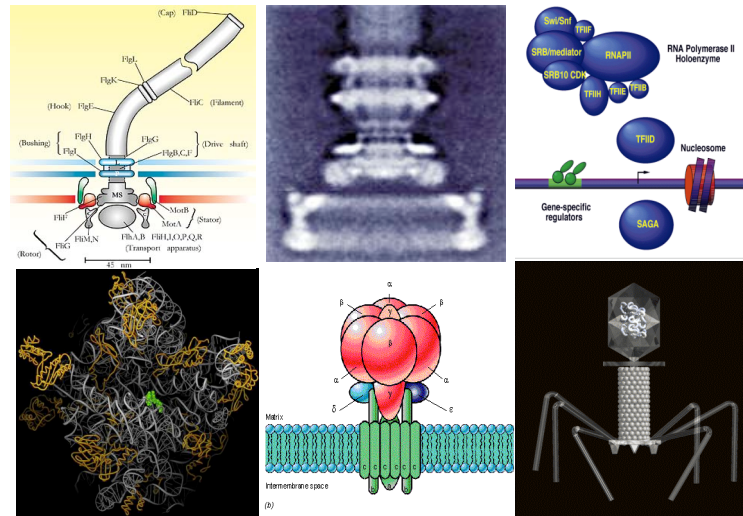
## The Parts List

- Sequencing gives “serial number”
- Sequence alignment gives a name
- Microarrays give # of parts
- X-ray and NMR give a picture
- However, having a collection of parts and names doesn't tell you how to put something together or how things connect -- *this is biology*

## Remember: *Proteins Interact*



## Proteins Assemble

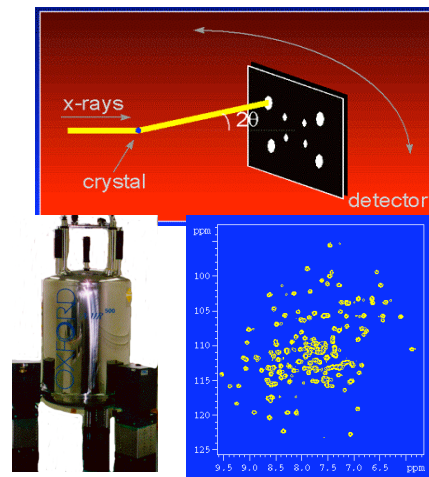


## Types of Interactions

- **Permanent (quaternary structure, formation of stable complexes)**
- **Transient (brief interactions, signaling events, pathways)**
- **About 1/4 to 1/3 of all proteins form complexes (dimers → multimers)**
- **Each protein may transiently interact with ~3 other proteins**

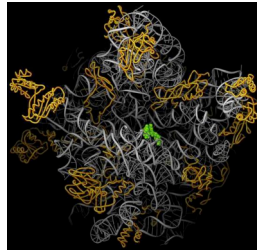
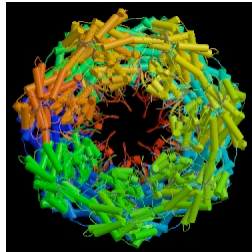
# Protein Interaction Tools and Techniques - Experimental Methods

## 3D Structure Determination

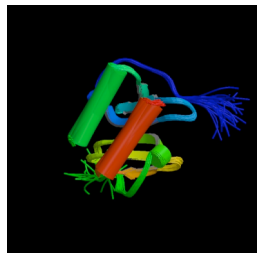
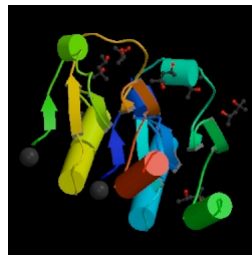


- **X-ray crystallography**
  - grow crystal
  - collect diffract. data
  - calculate e- density
  - trace chain
- **NMR spectroscopy**
  - label protein
  - collect NMR spectra
  - assign spectra & NOEs
  - calculate structure using distance geom.

## Quaternary Structure

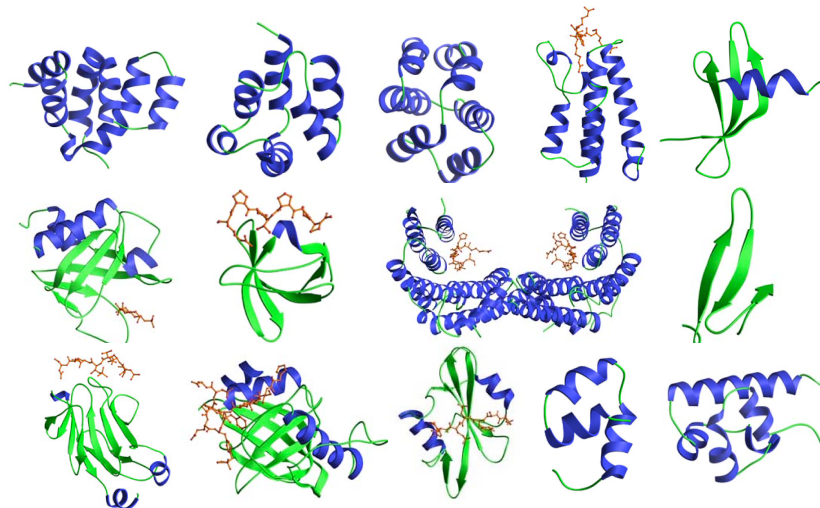


Some interactions  
are real



Others are not

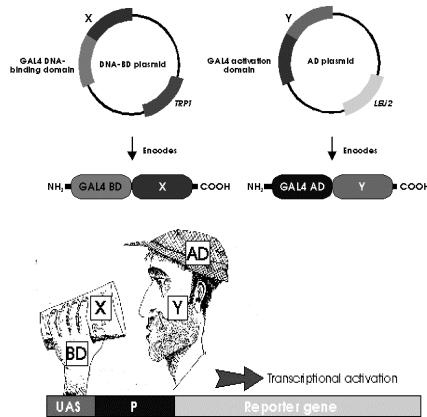
## Protein Interaction Domains



<http://pawsonlab.mshri.on.ca/> 82 domains

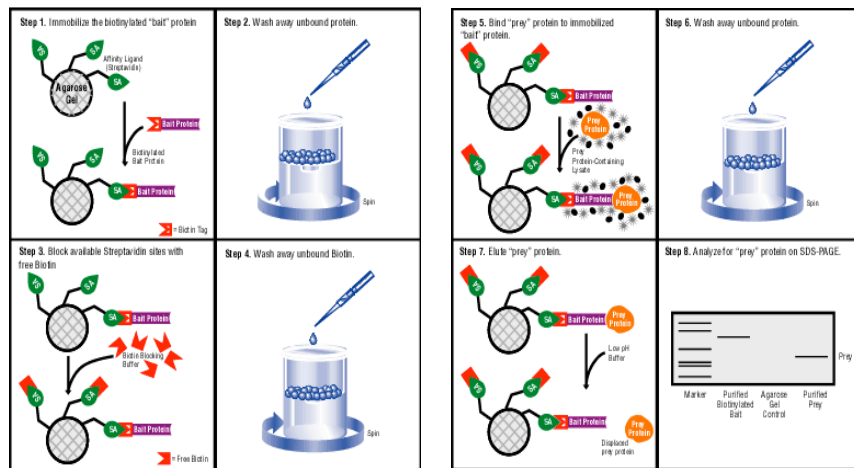


# Yeast Two-Hybrid Analysis

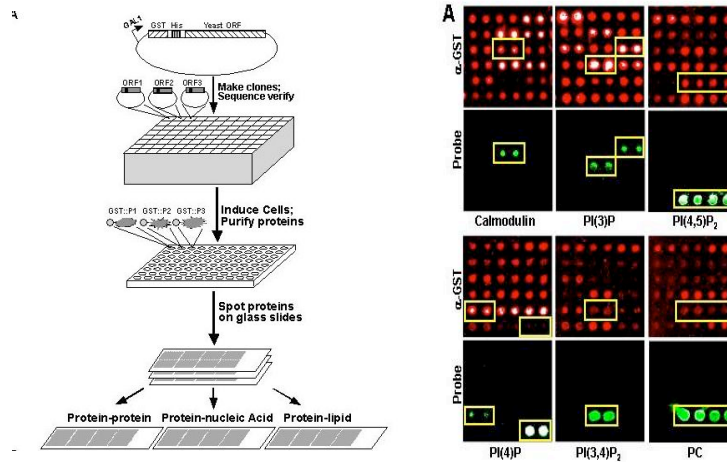


- Yeast two-hybrid experiments yield information on protein protein interactions
- GAL4 Binding Domain
- GAL4 Activation Domain
- X and Y are two proteins of interest
- If X & Y interact then reporter gene is expressed

# Affinity Pull-down



# Protein Arrays



## A Flood of Data

- High throughput techniques are leading to more and more data on protein interactions
- Very high level of false positives – need tools to sort and rationalize
- This is where bioinformatics can play a key role
- Some suggest that this is the “future” for bioinformatics

## Interaction Databases

- **BioGRID**
  - <http://www.thebiogrid.org/>
- **DIP**
  - <http://dip.doe-mbi.ucla.edu/>
- **MINT**
  - <http://160.80.34.4/mint/Welcome.do>
- **IntAct**
  - <http://www.ebi.ac.uk/intact/site/index.jsf>



*More Protein Interaction Databases are listed at  
<http://proteome.wayne.edu/PIDBL.html>*

## Reliability of HT Interaction Data

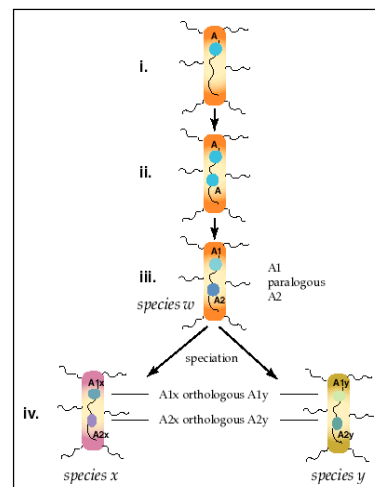
(Patil & Nakamura, BMC Bioinf. 6:100, 2005)

- **Assessed reliability using known interacting Pfam domains, Gene Ontology annotations and sequence homology**
- **56% of HT data for yeast are reliable**
- **27% of HT data for C. elegans are reliable**
- **18% of HT data for D. melanogaster are reliable**
- **68% of HT data for H. sapiens are reliable**

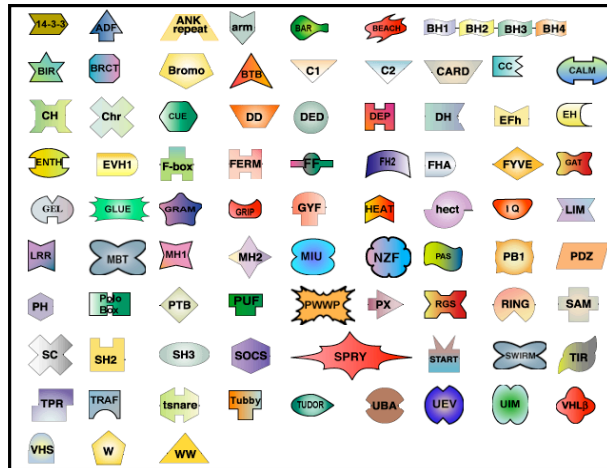
# Protein Interaction Tools and Techniques - Computational Methods

## Interologs, Homologs, Paralogs...

- **Homolog**
  - Common Ancestors
  - Common 3D Structure
  - Common Active Sites
- **Ortholog**
  - Derived from Speciation
- **Paralog**
  - Derived from Duplication
- **Interolog**
  - Protein-Protein Interaction



# Sequence Searching Against Known Domains



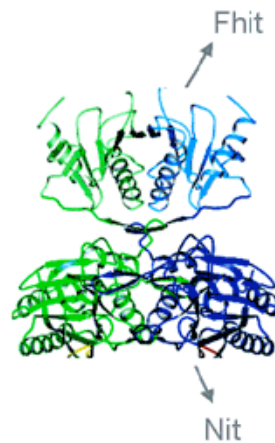
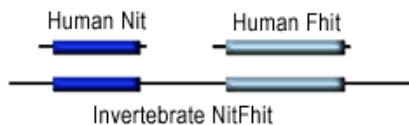
<http://pawsonlab.mshri.on.ca/>

# Rosetta Stone Method

Monomeric proteins that are fused in other organisms tend to be functionally related and physically interacting.

For example, using the Rosetta Stone™ method, it was found that human Nit and Fhit proteins are:

- fused in invertebrates
- form a heterocomplex in mammals



# Text Mining


- Searching Medline or Pubmed for words or word combinations
- “X binds to Y”; “X interacts with Y”; “X associates with Y” etc. etc.
- Requires a list of known gene names or protein names for a given organism (a protein/gene thesaurus)

# iHOP (Information hyperlinked over proteins)

The screenshot displays the iHOP web application. On the left, there is a search bar and navigation options. The central area features a network diagram with nodes and edges, representing interactions between proteins and genes. On the right, a list of search results is shown, each with a title and a brief description. The interface is designed to facilitate the discovery of relationships between different biological entities.

<http://www.ihop-net.org/UniPub/iHOP/>

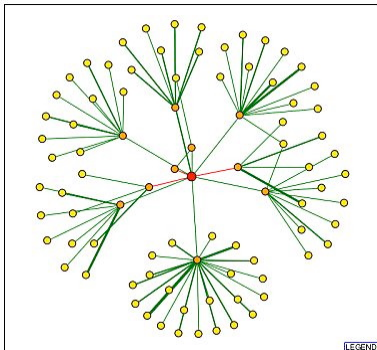
# Visualizing Interactions



#754  
**CELLULAR TUMOR ANTIGEN P53**

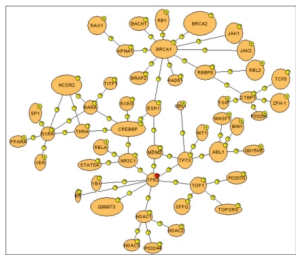
Binary Interactions - Pathways - Complexes

ID: 754  
AC: [R05037 \(5\)](#)



**DIP**

MINT View of 1  
P53  
CELLULAR TUMOR ANTIGEN P53



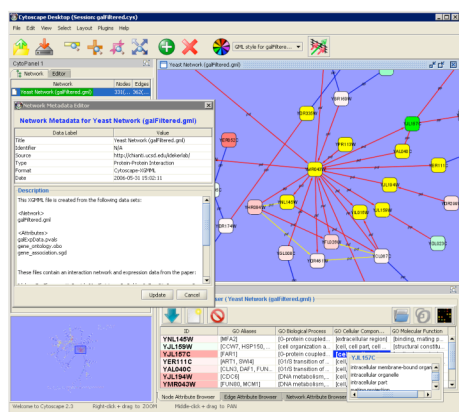
rbs arrest or apoptosis depending  
 rate are involved in tumor  
 but acts to negatively regulate cell  
 cycle of the arrested genes or in  
 cell to be maintained either by  
 1 of bcl-2 expression.

cell in cell-1 cell line  
 1 CA63036 AAC12971

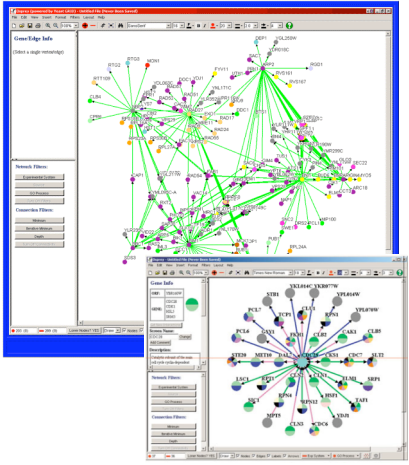
[USAL](#) [Y208](#) [L448](#) [Y180](#) [Y181X](#)  
 or 3Nucleic protein  
 Disease mutation: Polyspermy

**MINT**

# Visualizing Interactions



**Cytoscape** ([www.cytoscape.org](http://www.cytoscape.org))



**Osprey** <http://biodata.mshri.on.ca/osprey/servlet/Index>

# Pathway Visualization with BioCarta

The screenshot shows the BioCarta website interface. On the left, a list of pathways is displayed under the heading "ALL PATHWAYS". On the right, a detailed signaling pathway diagram is shown, illustrating the interaction of various proteins and molecules in a cell. The diagram is color-coded and includes labels for various components such as receptors, kinases, and transcription factors.

<http://www.biocarta.com/genes/allpathways.asp>

## Summary

- **First application of bioinformatics was probably in protein structure (the PDB)**
- **Structural biology continues to be a rich source for bioinformatics innovation and bioinformaticians**
- **Next “big” step in bioinformatics is to go from the “parts list” to figuring out how to put it all together**