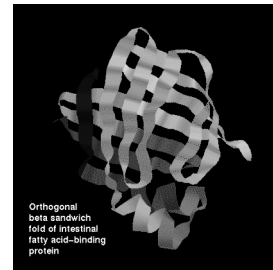
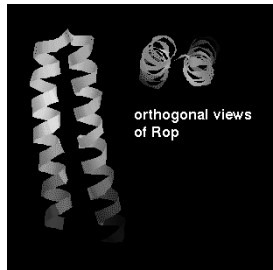


# Protein Structure Analysis & Protein-Protein Interactions



David Wishart

University of Alberta, Edmonton, Canada

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

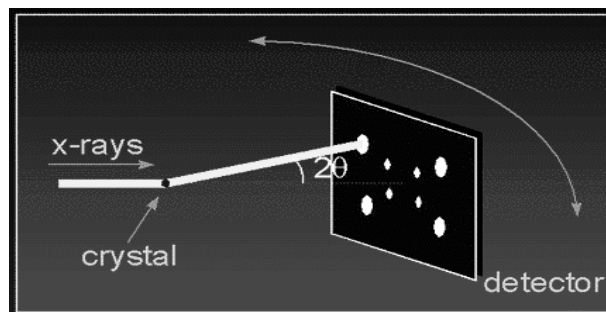
## Much Ado About Structure

- 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)

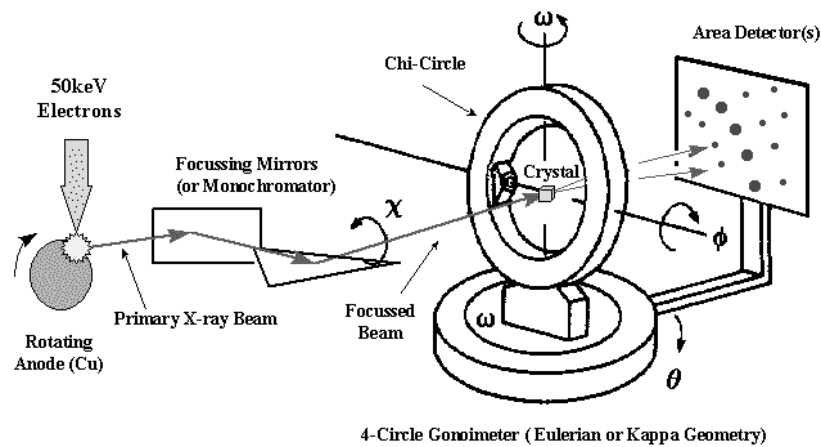
## X-ray Crystallography



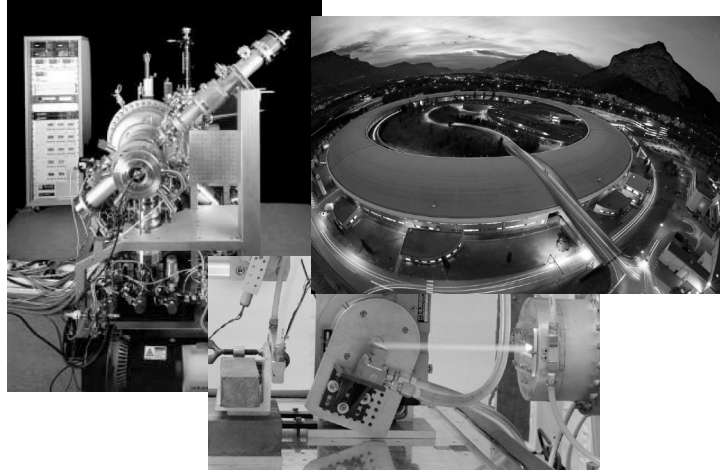
# X-ray Crystallography

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

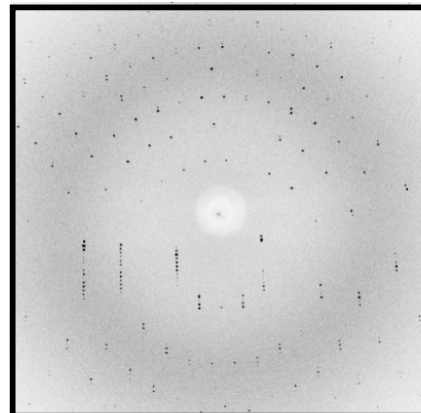
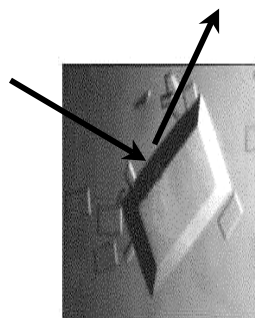
## Diffraction Apparatus



# Synchrotron Diffractometer

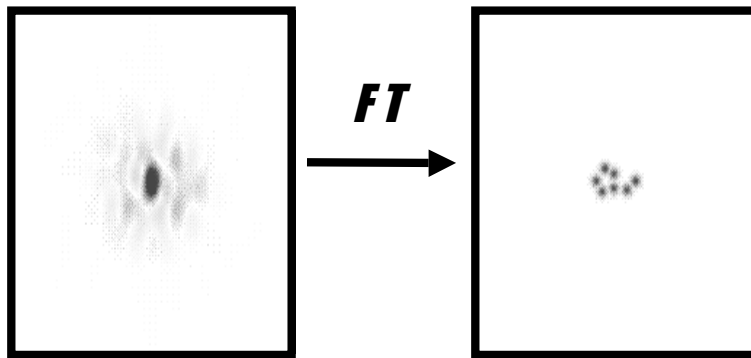


# Protein Crystal Diffraction

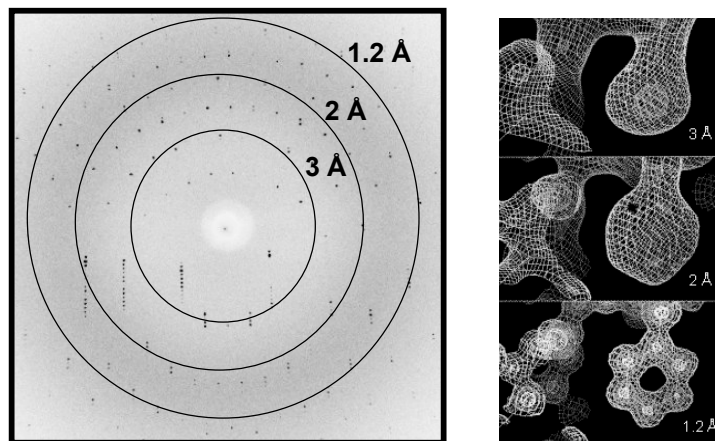


**Diffraction Pattern**

## Converting Diffraction Data to Electron Density



## Resolution

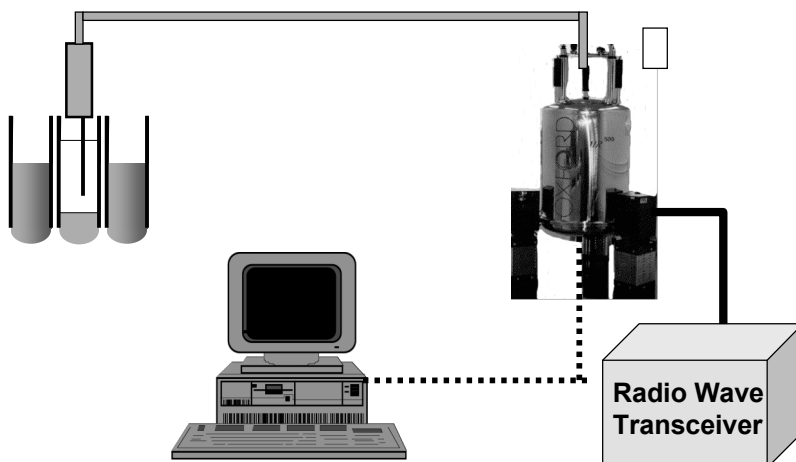


## The Final Result

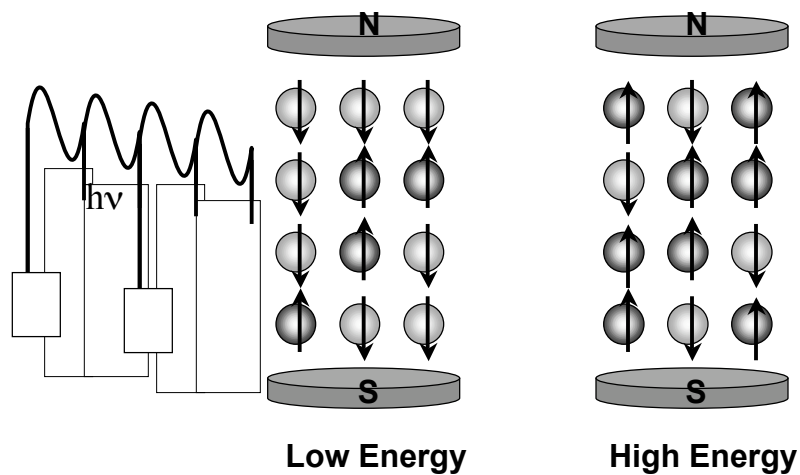
ORIGX2	0.000000	1.000000	0.000000	0.000000		2TRX	147					
ORIGX3	0.000000	0.000000	1.000000	0.000000		2TRX	148					
SCALE1	0.011173	0.000000	0.004858	0.000000		2TRX	149					
SCALE2	0.000000	0.019585	0.000000	0.000000		2TRX	150					
SCALE3	0.000000	0.000000	0.018039	0.000000		2TRX	151					
ATOM	1	N	SER	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>

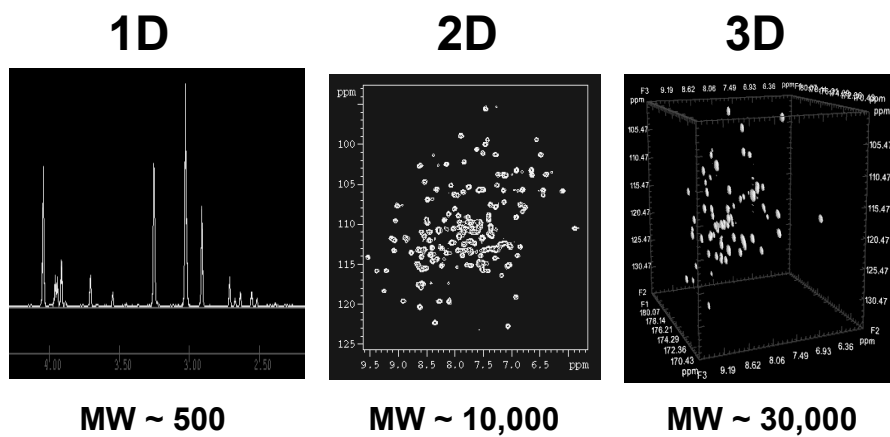
## NMR Spectroscopy



# Principles of NMR



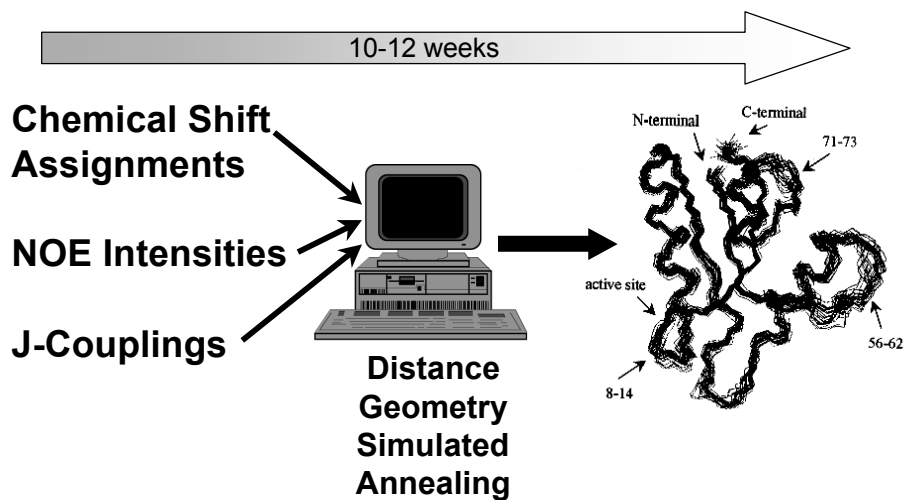
# Multidimensional NMR



## 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



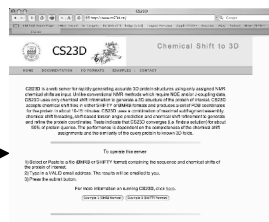


# NMR Spectroscopy

**New!**

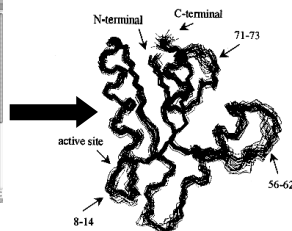
10-12 minutes

**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

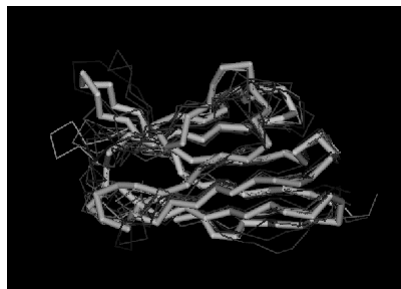
## X-ray

- 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



```
ACDEFGHIKLMNPQRST--FGHQWERT-----TYREWYEGHADS  
ASDEYAHLRILDPQRSTVAYAYE--KSFAPPGSFKWEYEAHADS  
MCDEYAHIRLMNPERSTVAGGHQWERT-----GSFKEWYAAHADD
```

## **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!

**Swiss-Model**

MENU

Modeling requests:

- First Approach mode
- Alignment Interface
- Project (optimise) mode
- Oligomer modelling
- GPCR mode

Model Database

- SWISS-MODEL Repository, a database for theoretical protein models.

Interactive tools

- SWISS-MODEL Workspace, an interactive working environment for protein structure modelling and assessment.
- DeepView - Swiss-PdbViewer, a tool for viewing and manipulating protein structures and models.
- Lookup ExPDB template codes accessible to SWISS-MODEL.
- Search the SWISS-MODEL Template Library.
- Examples using SWISS-MODEL and the Swiss-PdbViewer.
- ANOLEA Protein structure quality check (atomic non-local environment assessment)

HELP

- Frequently Asked Questions.
- Visualising 3D models.
- Reliability of models.
- How SWISS-MODEL works.
- How ProModel works.
- Modelling of oligomeric proteins.
- Model Confidence factors.
- About model quality.

15 years SWISS-MODEL

An Automated Comparative Protein Modelling Server

SWISS-MODEL is a fully automated protein structure homology-modelling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make Protein Modelling accessible to all biochemists and molecular biologists World Wide.

The present version of the server is 3.5 and is under constant improvement and debugging. In order to help us refine the sequence analysis and modelling algorithms, please report possible bugs and problems with the modelling procedure.

SWISS-MODEL is provided by:

BIOZENTRUM SIB

History

SWISS-MODEL was initiated in 1993 by Manuel Peitsch, and further developed at Glaxo Wellcome Experimental Research in Geneva and the SIB Swiss Institute of Bioinformatics by Manuel Peitsch, Nicolas Guex and Torsten Schwede. Since 2001, SWISS-MODEL is being developed by Torsten Schwede's Structural Bioinformatics Group at the SIB & Biozentrum (University of Basel). The SWISS-MODEL Repository, a relational database of annotated three-dimensional comparative protein structure models, was established in 2004. In 2005, SWISS-MODEL service was extended by SWISS-MODEL Workspace, a web-based work bench for protein structure modelling and assessment. Computational resources for the SWISS-MODEL server are provided in collaboration by the Biozentrum (University Basel), the Swiss Institute of Bioinformatics and the Advanced Biomedical Computing Center (NCI Frederick, USA).

Acknowledgements

SWISS-MODEL would not have been possible without a lot of help and support. We are particularly thankful to Nicolas Guex for his many crucial contributions to the development efforts of Swiss-Model and specifically DeepView and to Gale Rhodes of the University of Southern Maine for coordinating the active DeepView user community. We also thank Alexander Diemand, Konstantin Arnold, Jürgen Kiss and Lorenzo Bordoli for their many contributions to the development and operations for the modeling platform. Furthermore, we

<http://swissmodel.expasy.org//SWISS-MODEL.html>

# 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 "An Information Portal to Biological Macromolecular Structures". Below this, there is a navigation bar with "CONTACT US | HELP | PRINT PAGE" and search options for "PDB ID or keyword" and "Author". The main content area features a "Welcome to the RCSB PDB" section with a message about data updates and a "Molecule of the Month: Adrenergic Receptors" feature. A sidebar on the left contains a "Home" menu with links to "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". A "Quick Tips" box is also present in the sidebar. The right sidebar includes "News" with links to "Complete News", "Newsletter", "Discussion Forum", and "Job Listings", along with a "Molecule of the Month" section.

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

# Viewing 3D Structures

**RCSB PDB Structure Explorer - Netscape**

**PDB**  
PROTEIN DATA BANK

An Information Portal to Biological Macromolecular Structures  
As of Tuesday Oct 10, 2006 there are 39323 Structures | PDB Statistics

Home | Search | Structure | Queries | Structure Summary | Biology & Chemistry | Materials & Methods | Sequence Details | Geometry

2TRX

**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* 2002, 26:167-169, 1992 [Abstract]

**History**  
Deposition 1990-03-19 Release 1991-10-15

**Experimental Method**  
Type X-RAY DIFFRACTION Data NA

Resolution(Å)	R-Value	R-Free	Space Group
1.68	0.165 (obs)	n/a	C 2 (C 1 2 1)

**Parameters**

Length (Å)	a	b	c
89.50	51.06	60.45	

Angles (°)	alpha	beta	gamma
90.00	113.50	90.00	

**Unit Cell**

**Molecular Description**  
Asymmetric Unit  
Polymer 1 Molecule THIOREDOXIN Chains: A,B

**Functional Class**  
Electron Transport

**Display Options**  
KING  
Jmol  
WebMol  
Protein Workshop  
QuickPDB  
All Images

# KiNG (Kinemage) 1.39

**KiNG 1.39**

File Edit Views Display Tools Help

Channels

Kinemage #1

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

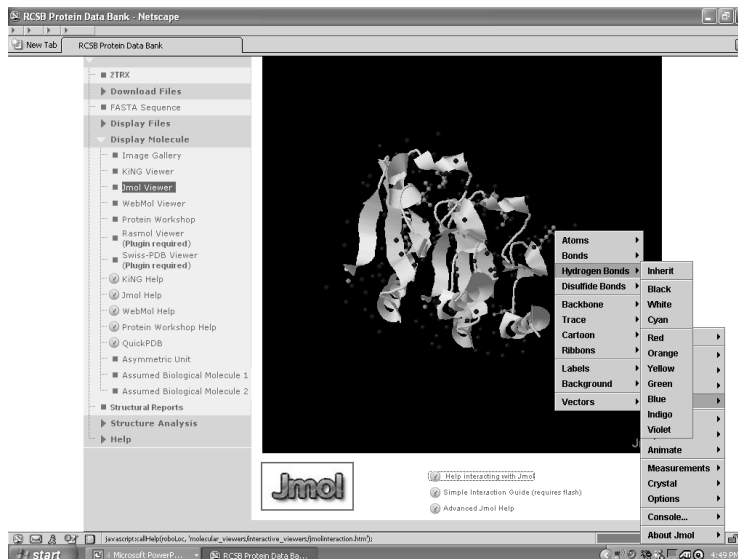
Zoom: [Slider]

Clipping:  Pick center  Show text  
 Markers  Show hierarchy

## 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

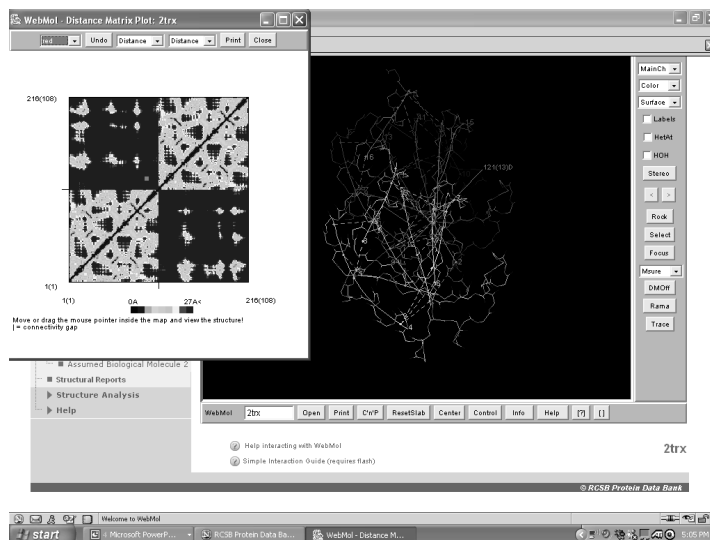




# JMol

- **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

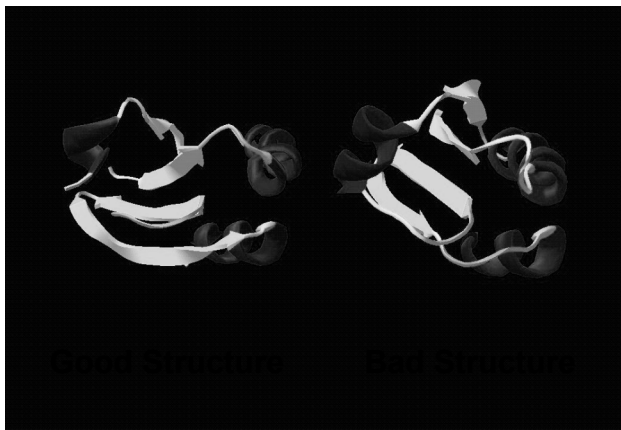
# WebMol



## 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

### NMR structure

- |                              |                            |
|------------------------------|----------------------------|
| • R = 0.59 random chain      | • rmsd = 4 Å random        |
| • R = 0.45 initial structure | • rmsd = 2 Å initial fit   |
| • R = 0.35 getting there     | • rmsd = 1.5 Å OK          |
| • R = 0.25 typical protein   | • rmsd = 0.8 Å typical     |
| • R = 0.15 best case         | • rmsd = 0.4 Å best case   |
| • R = 0.05 small molecule    | • 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



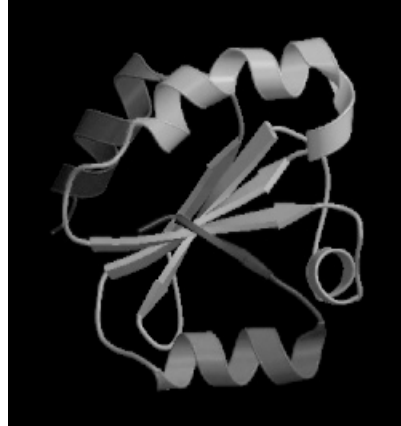
X-ray to X-ray  
Interleukin 1 $\beta$   
(41bi vs 2mlb)



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

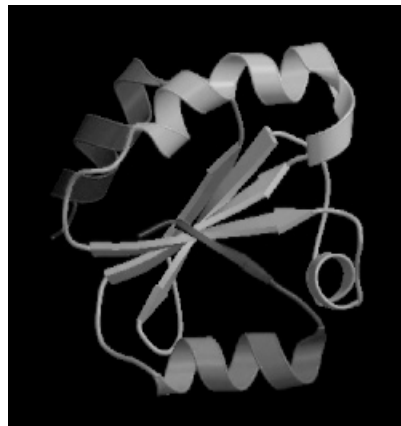
## A Good Protein Structure..

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



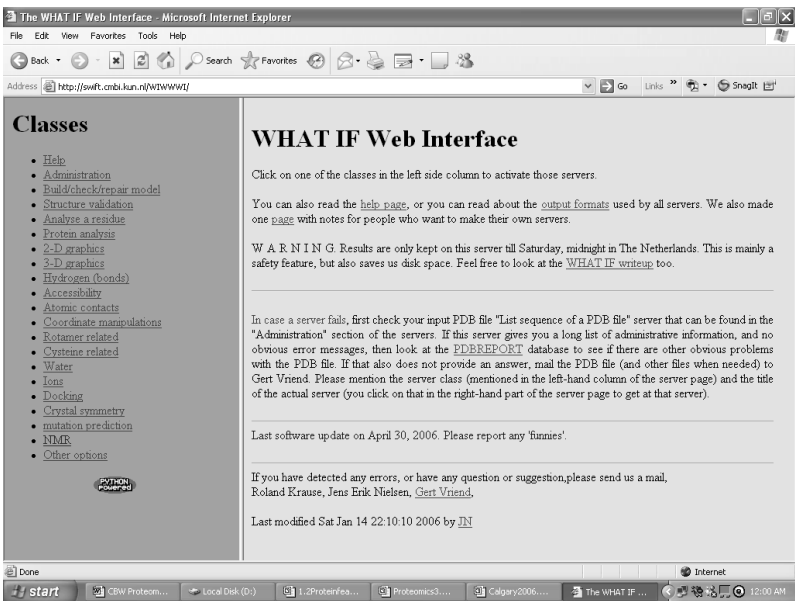
## A Good Protein Structure..

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

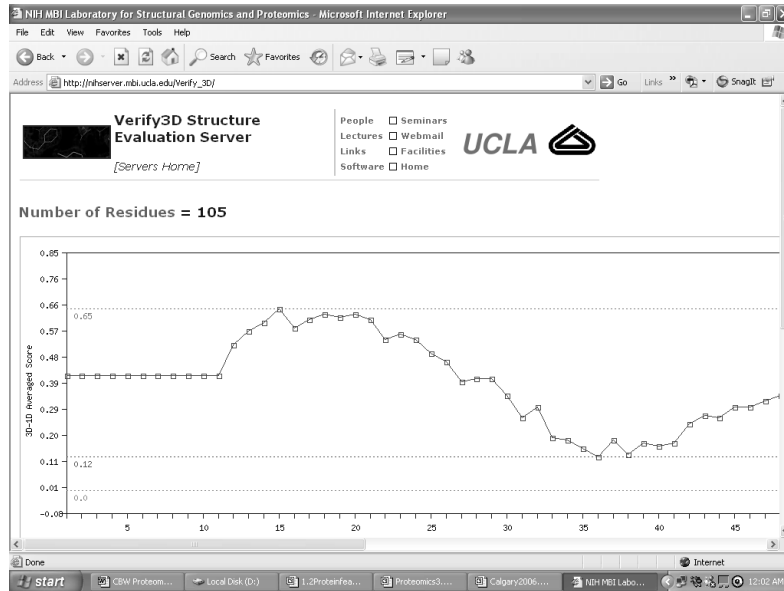


# Structure Validation Servers

- **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.ru.nl/WWW/". The page content is divided into two main sections. On the left, under the heading "Classes", there is a list of links: 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, under the heading "WHAT IF Web Interface", there is a paragraph of 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." Below this is a "WARNING" section: "WARNING 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." Further down, it says: "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)." Below that is a note: "Last software update on April 30, 2006. Please report any 'funnies'." At the bottom, it says: "If you have detected any errors, or have any question or suggestion, please send us a mail, Roland Krause, Jens Erik Nielsen, Gert Vriend." and "Last modified Sat Jan 14 22:10:10 2006 by IN". The browser's taskbar at the bottom shows several open windows: "start", "CBM Proteom...", "Local Disk (D:)", "1:Proteinfes...", "Proteomics3...", "Calgary2006...", "The WHAT IF...", and "Internet". The system clock shows "12:00 AM".



High scores = good Low scores = bad

# 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. Boyko, 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)

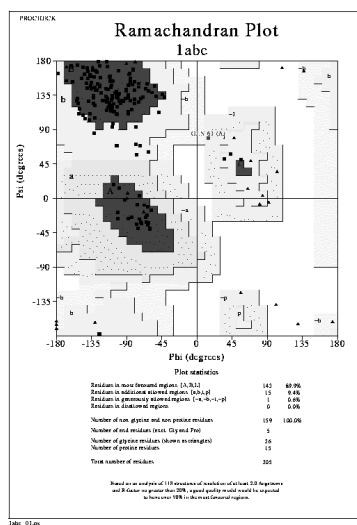
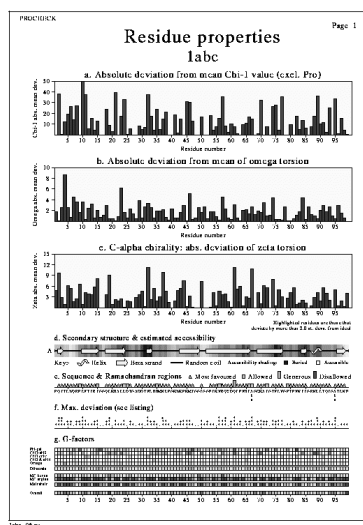
<http://redpoll.pharmacy.ualberta.ca/vadar>



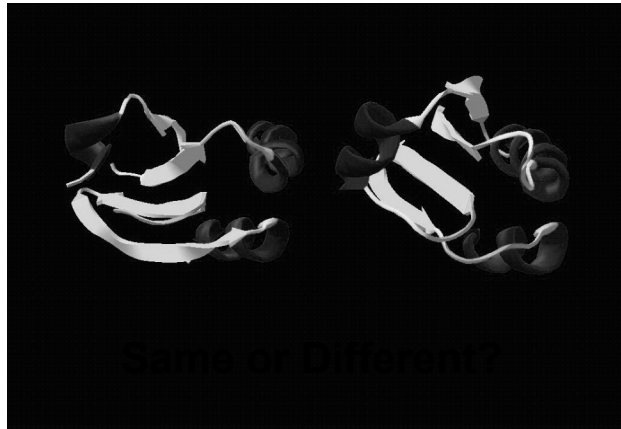
# Structure Validation Programs

- **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

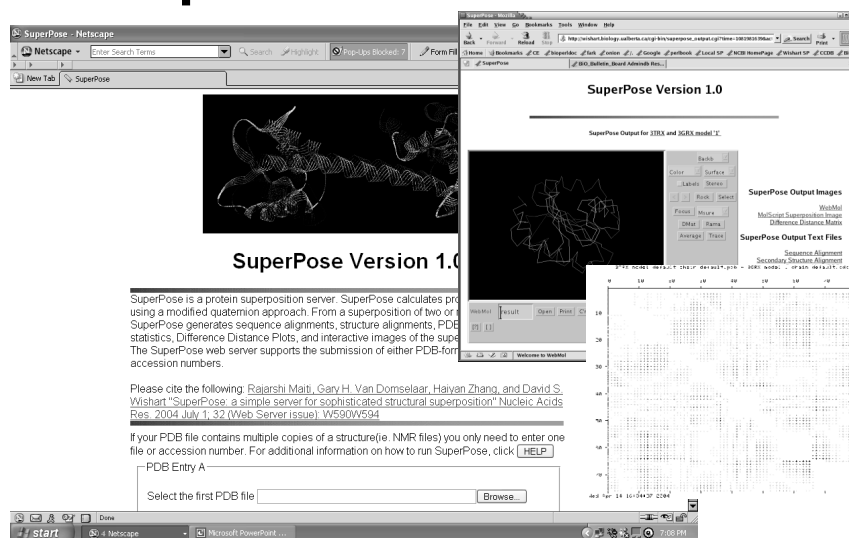
## Rigid Body Superposition



# 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



The screenshot displays the SuperPose web server interface. At the top, there's a navigation bar with "SuperPose" and "Netscape" tabs. The main content area features a 3D protein structure on the left, a "SuperPose Version 1.0" title, and a "SuperPose Output for 3DXX and 3DXX model 1" section. Below this, there's a "SuperPose Output Images" section with a "WebMail" link and a "SuperPose Output Text Files" section with a "Sequence Alignment" link. A form for submitting PDB files is visible at the bottom, with a "Browse" button and a "PDB Entry A" label. The interface also includes a "Please cite the following" section with a citation to Rajarshi Maiti, Garv H. Van Domselaar, Haiyan Zhang, and David S. Wishart's 2004 paper on SuperPose.

<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**

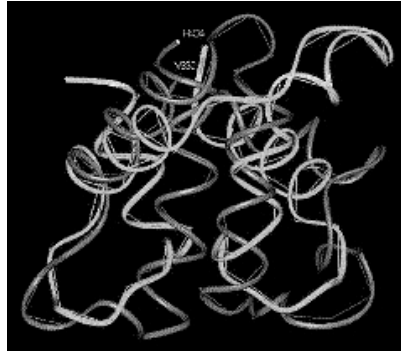
- **Method to quantify structural similarity - same as standard deviation**
- **Requires 2 superimposed structures (designated here as “a” & “b”)**
- **N = number of atoms being compared**

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

## **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

**Structure Explorer - Microsoft Internet Explorer**

Address: <http://www.rcsb.org/pdb/home/home.do?method=author&search=mod&mol=All&input=Q&idSearch=23r>

**Structure Summary** | Biology & Chemistry

**SCOP: Thioredoxin from Escherichia coli**

Address: <http://scop.bemley.ed.ac.uk/scop.b.d.f/b.b.html>

**Structural Classification of Proteins**

**Protein: Thioredoxin from *Escherichia coli***

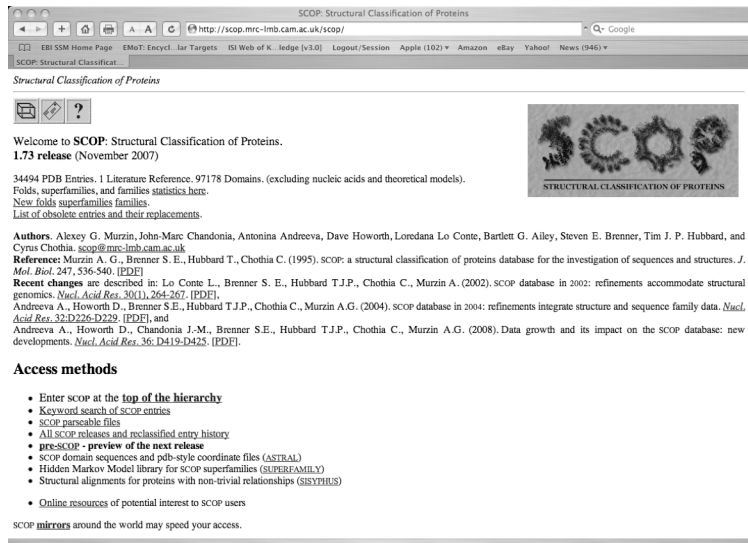
**Lineage:**

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

**PDB Entry Domains:**

1. *23r* [52835] complexed with *cu*, *mpd*
  1. *chain a* [32719] [52835]
  2. *chain b* [32720] [52835]
2. *23r* [32721] [52835]
3. *23r* [32720] [52835] complexed with *cu*, *mutant*

# SCOP Database



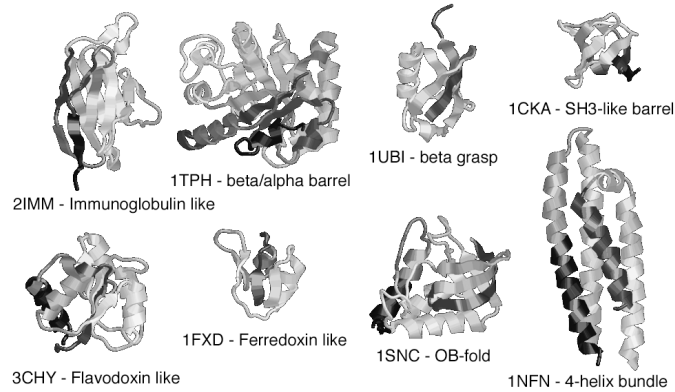
The screenshot shows the SCOP Database homepage in a web browser. The browser's address bar displays the URL <http://scop.mrc-lmb.cam.ac.uk/scop/>. The page title is "SCOP: Structural Classification of Proteins". Below the title, there is a navigation menu with icons for home, search, and help. The main content area features a welcome message: "Welcome to SCOP: Structural Classification of Proteins. 1.73 release (November 2007)". It provides statistics: "34494 PDB Entries, 1 Literature Reference, 97178 Domains (excluding nucleic acids and theoretical models). Folds, superfamilies, and families statistics here." There are links for "New folds superfamilies families" and "List of obsolete entries and their replacements". The authors listed are 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. The page also includes a list of references and a section titled "Access methods" with several bullet points: "Enter SCOP at the top of the hierarchy", "Keyword search of SCOP entries", "SCOP parsable 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 (SISYPHUS)", and "Online resources of potential interest to SCOP users". At the bottom, it mentions "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. Jedge [v3.0] Logout/Session Apple (102) Amazon eBay Yahoo! News (946)

CATH Protein Structure Cla

**CATH**  
Protein Structure Classification

Home > Top

**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 Addou, Mr. Tim Dallman, Mr. Benoit Dessailly, Dr. Lesley Greene, Dr. David Lee, Dr. Jon Lees, Dr. Russel L. Marsden, Mr. Adam Reid, Mr. Stathis Soltes, Dr. Corin 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 (gziped HMMER2.0 format: 63MB)
- [NEW]** CrossLinks 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
Domains	86191	93867	7734
Chains	57741	63453	5712
PDBs	27522	30028	2506

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

- Development of backend PostgreSQL database
- Development of the central code library
- New web interface for domain grouping (DoinChop)
- Improved public pages to show very latest information
- Added numerous maintenance scripts and regression tests

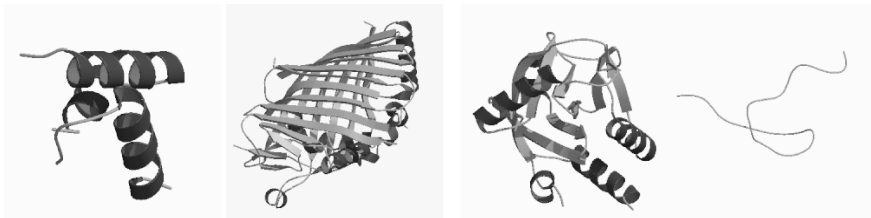
http://www.cathdb.info/latest/index.html



# 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

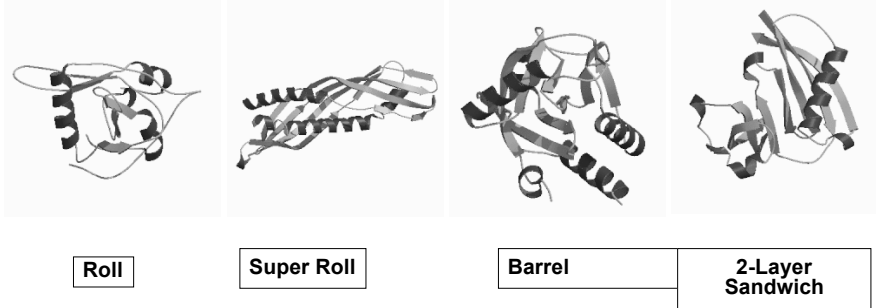
**Class 2:**  
Mainly Beta

**Class 3:**  
Mixed  
Alpha/Beta

**Class 4:**  
Few Secondary  
Structures

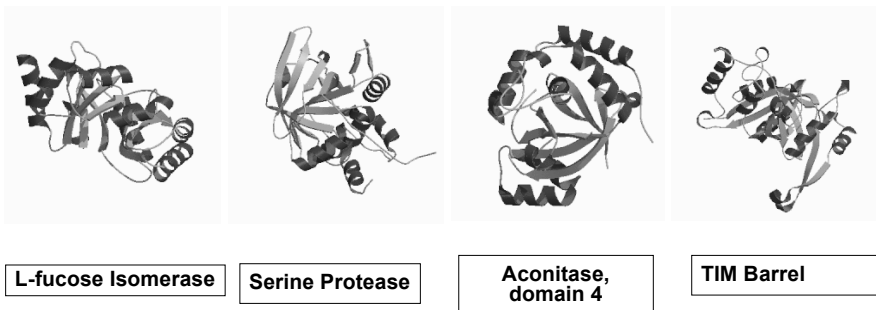
**Secondary structure content (automatic)**

## CATH - Architecture



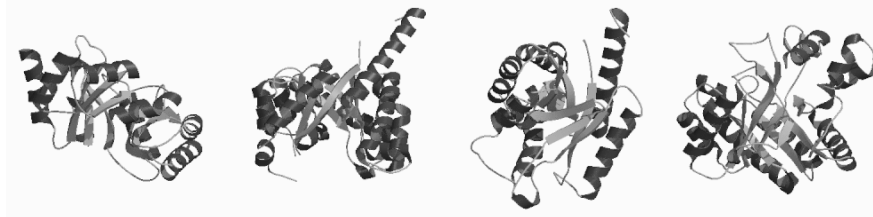
Orientation of secondary structures (manual)

## CATH - Topology



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

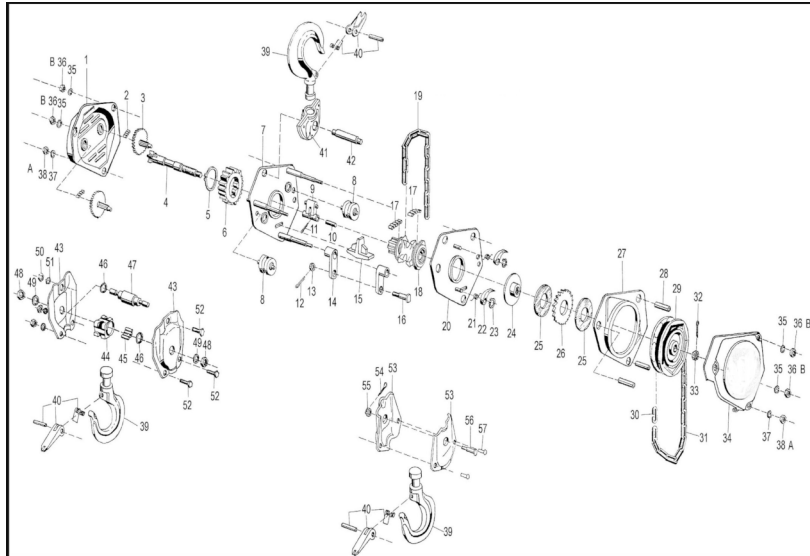
## Other Servers/Databases

- **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



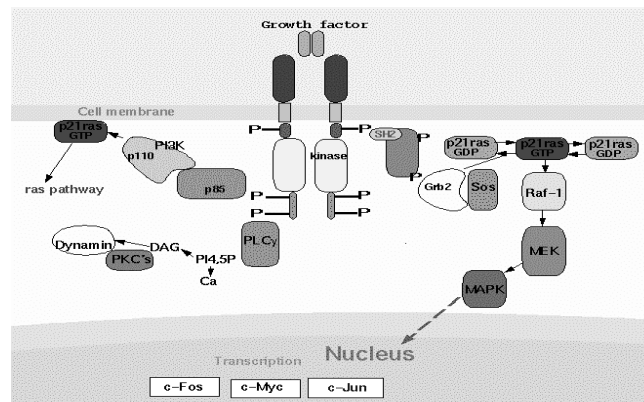
# The Protein Parts List



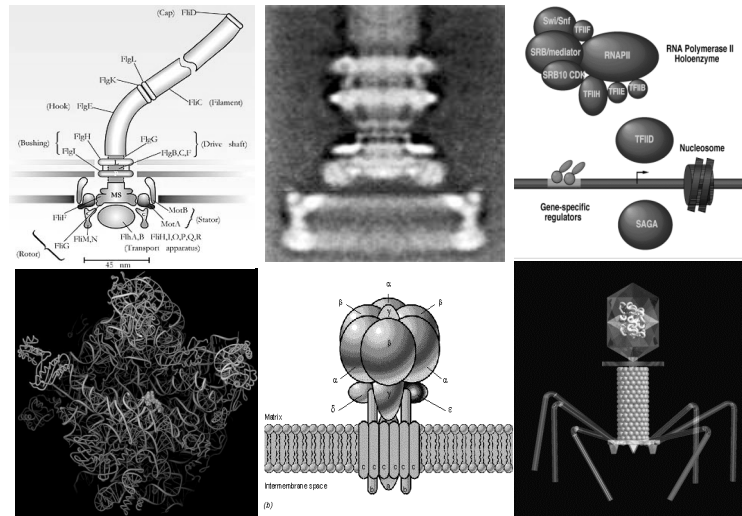
## The Parts List

- 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

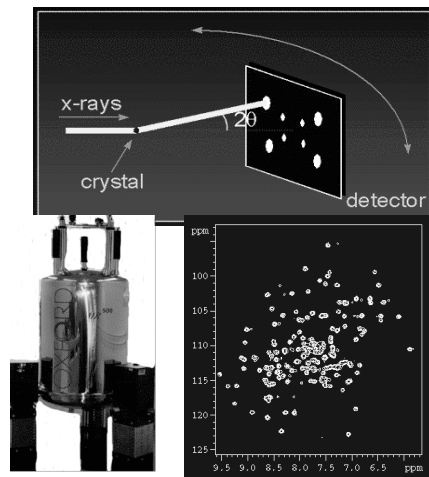


## Types of Interactions

- **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**

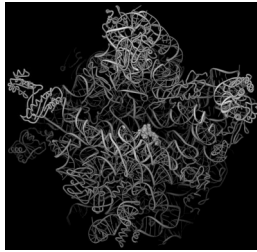
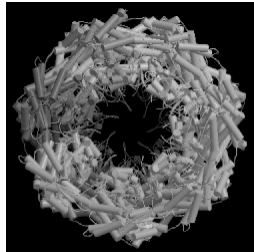
# Protein Interaction Tools and Techniques - Experimental Methods

## 3D Structure Determination

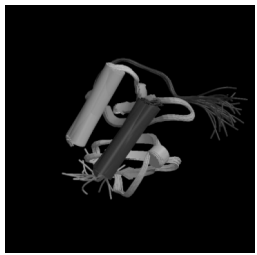


- **X-ray crystallography**
  - grow crystal
  - collect diffract. data
  - calculate e- density
  - trace chain
- **NMR spectroscopy**
  - label protein
  - collect NMR spectra
  - assign spectra & NOEs
  - calculate structure using distance geom.

## Quaternary Structure

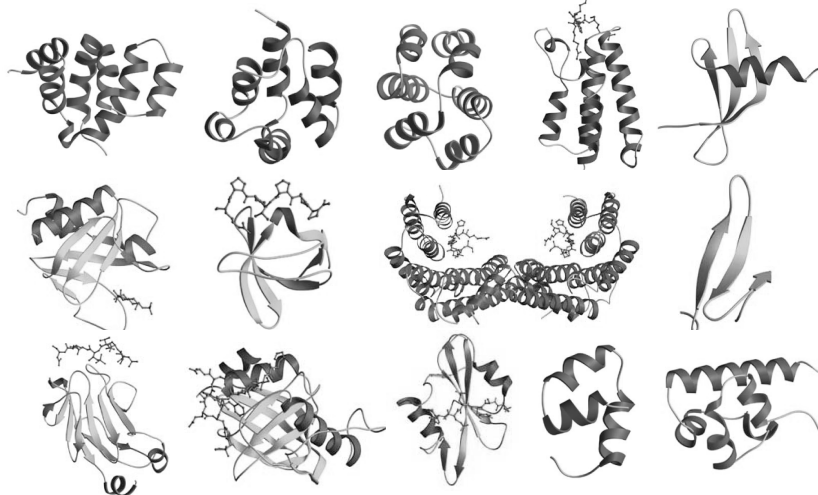


Some interactions  
are real



Others are not

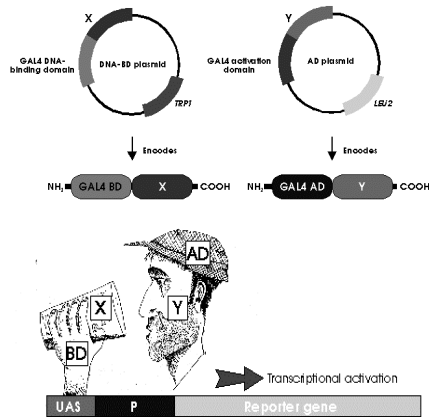
## Protein Interaction Domains



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

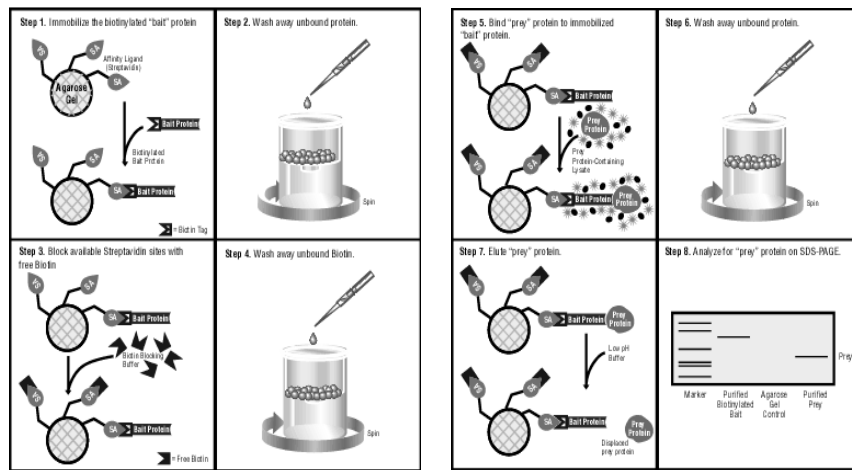


# Yeast Two-Hybrid Analysis

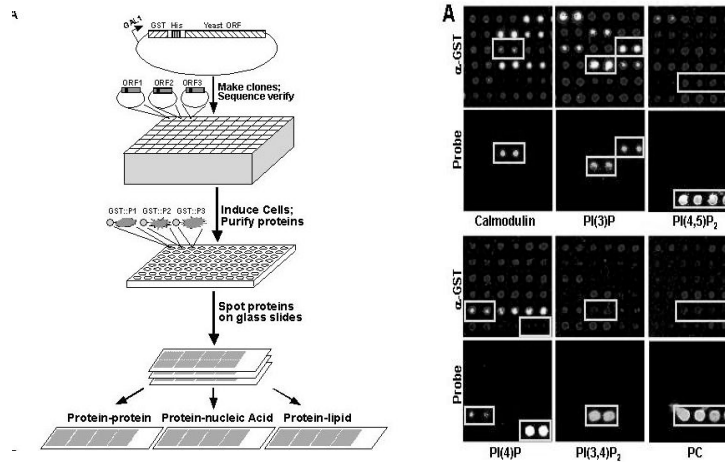


- Yeast two-hybrid experiments yield information on protein protein interactions
- GAL4 Binding Domain
- GAL4 Activation Domain
- X and Y are two proteins of interest
- If X & Y interact then reporter gene is expressed

# 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)

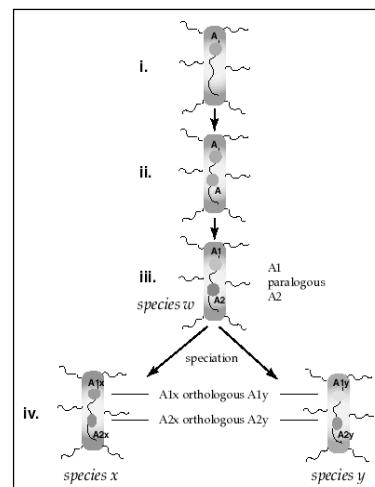
- **Assessed reliability using known interacting Pfam domains, Gene Ontology annotations and sequence homology**
- **56% of HT data for yeast are reliable**
- **27% of HT data for C. elegans are reliable**
- **18% of HT data for D. melanogaster are reliable**
- **68% of HT data for H. sapiens are reliable**

# Protein Interaction Tools and Techniques - Computational Methods

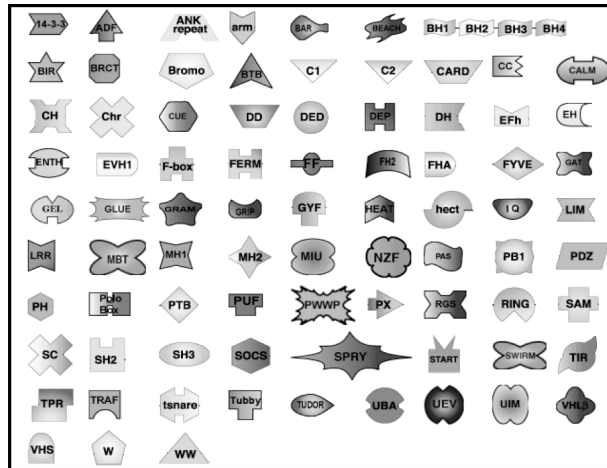
## Interologs, Homologs, Paralogs...

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

YM2



# Sequence Searching Against Known Domains



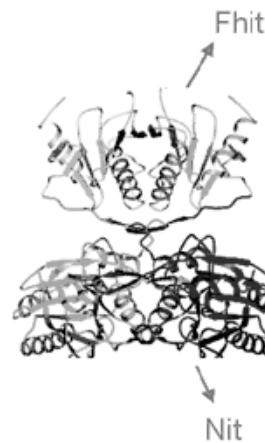
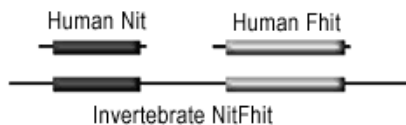
<http://pawsonlab.mshri.on.ca/>

## 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)

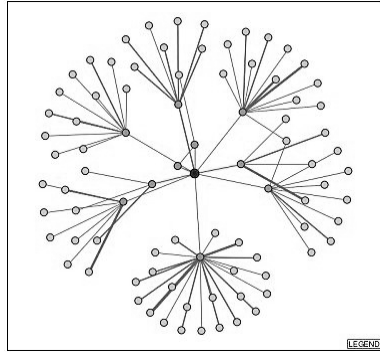
The screenshot displays the iHOP web application interface. The main content area shows a search result for 'iHop' with a detailed text description. The text describes the relevant sequences of the gene, its function in E. coli, and the effects of various mutations. The interface includes a search bar, navigation buttons, and a sidebar with links to 'PHYSIOLOGY' and 'INTERACTIONS'. The browser's address bar shows the URL: <http://www.ihop-net.org/UniPub/iHOP/>

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

# Visualizing Interactions

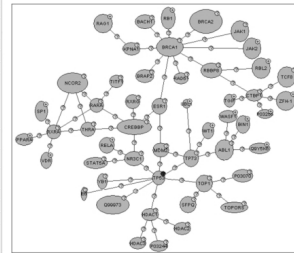
MINT *a Molecular INteractions database*

#754  
CELLULAR TUMOR ANTIGEN P53



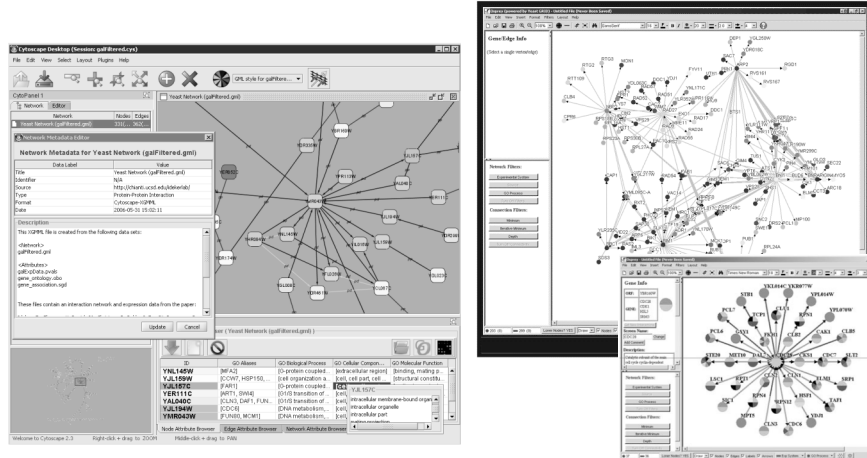
DIP

MINT View of 1  
P53  
CELLULAR TUMOR ANTIGEN P53



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. The top navigation bar includes 'FEATURES', 'PATHWAYS', 'OUTER-SERVICES', 'GENES', and 'PRODUCTS'. The main content area is titled 'PATHWAYS > All Pathways' and contains a list of pathway titles such as 'Acetylation and Deacetylation of Histone in the Nucleus', 'Activation of cdk by GMP-dependent protein kinase', and 'Adhesion and Diapedesis of Granulocytes'. An inset window displays a detailed signaling pathway diagram. This diagram illustrates the interaction between extracellular signals (like LPS, TNF, and IL-1) and intracellular components (including receptors like TLRs, MyD88, IRAK, and signaling molecules like MEK1, ERK1, and NF-κB) leading to nuclear events such as transcription factor activation and gene expression. The diagram is labeled with 'Extracellular' and 'Intracellular' regions and includes various protein names and their interactions.

<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