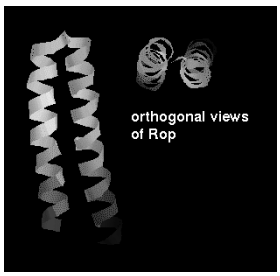
Protein Structure Analysis & Protein-Protein Interactions



David Wishart
University of Alberta, Edmonton, Canada
david.wishart@ualberta.ca

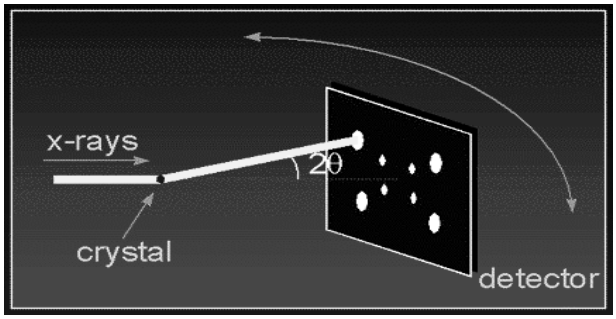
Much Ado About Structure

- **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 microsocopy (distant 3rd)
- Homology Modelling (sometimes VG)
- Threading (sometimes VG)
- Ab initio prediction (getting better)

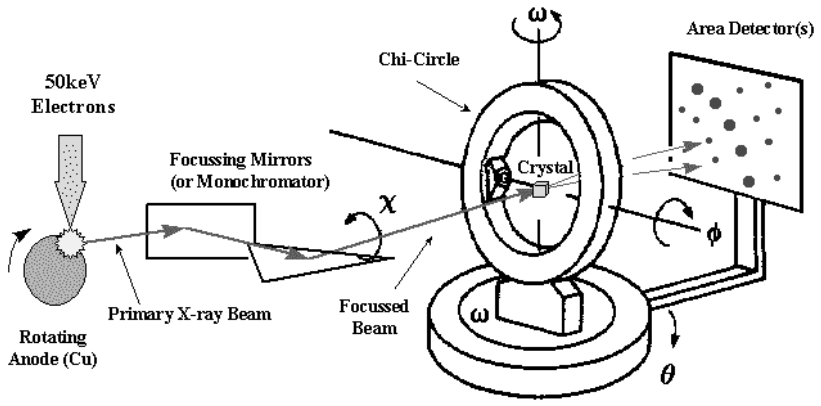
X-ray Crystallography



X-ray Crystallography

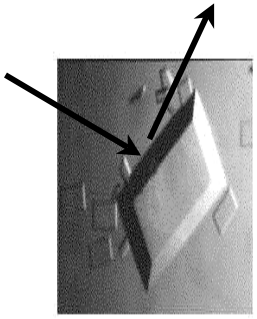
- Crystallization
- Diffraction Apparatus
- Diffraction Principles
- Conversion of Diffraction Data to Electron Density
- Resolution
- Chain Tracing

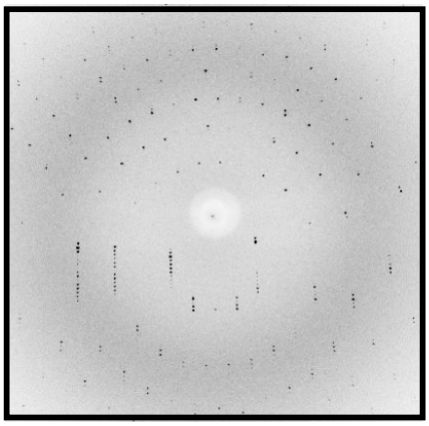
Diffraction Apparatus



4-Circle Goniometer (Eulerian or Kappa Geometry)

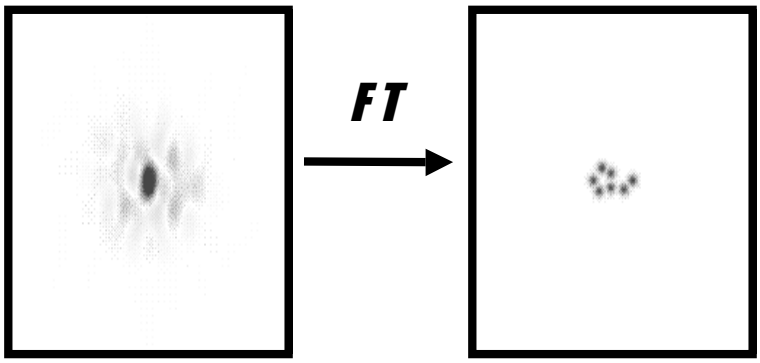
Protein Crystal Diffraction



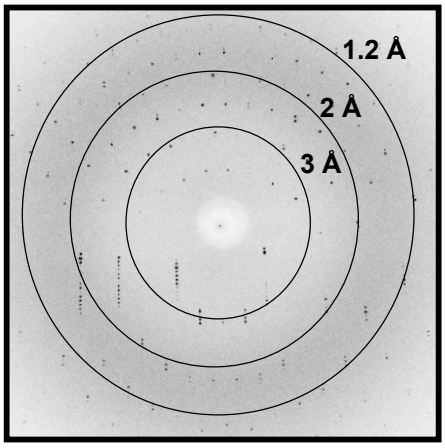


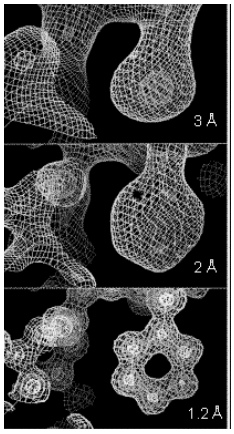
Diffraction Pattern

Converting Diffraction Data to Electron Density



Resolution



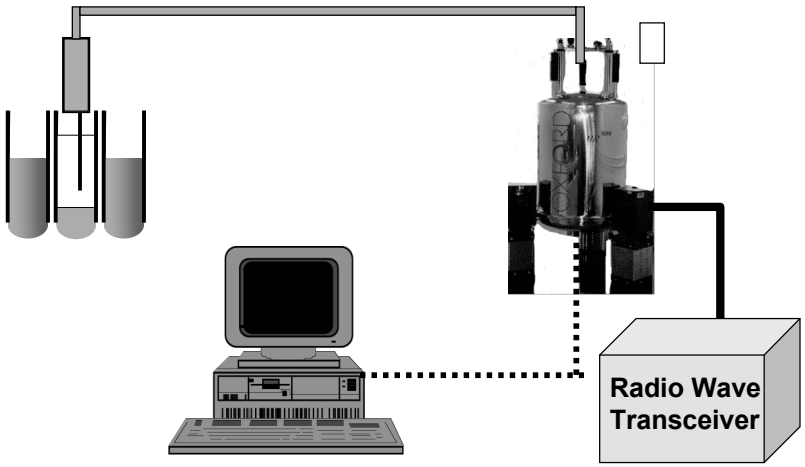


The Final Result

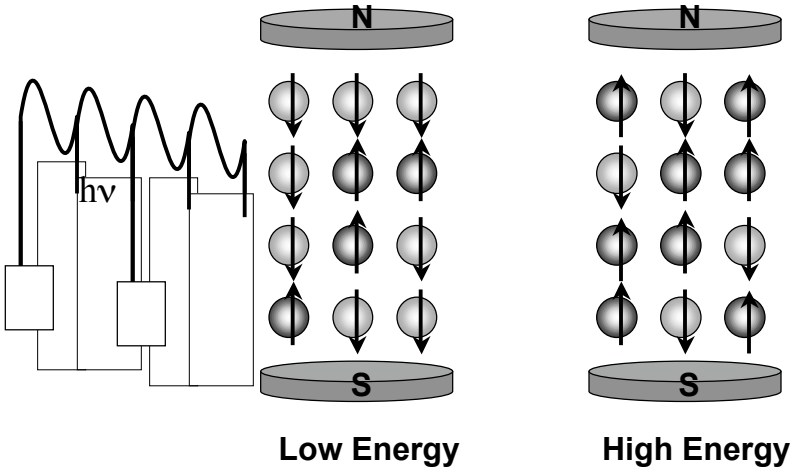
```
ORIGX2      0.000000  1.000000  0.000000        0.00000                  2TRX 147
ORIGX3      0.000000  0.000000  1.000000        0.00000                  2TRX 148
SCALE1      0.011173  0.000000  0.004858        0.00000                  2TRX 149
SCALE2      0.000000  0.019585  0.000000        0.00000                  2TRX 150
SCALE3      0.000000  0.000000  0.018039        0.00000                  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-structure.llnl.gov/Xray/101index.html

NMR Spectroscopy

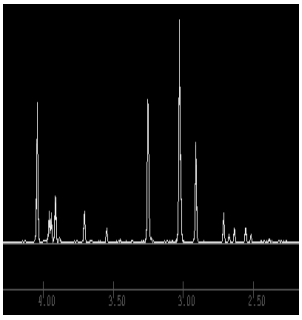


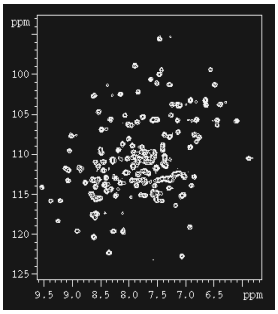
Principles of NMR

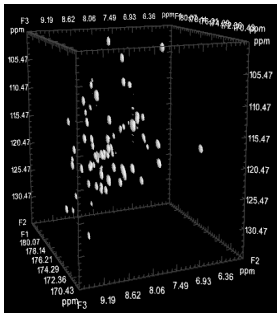


Multidimensional NMR

1D **2D** **3D**





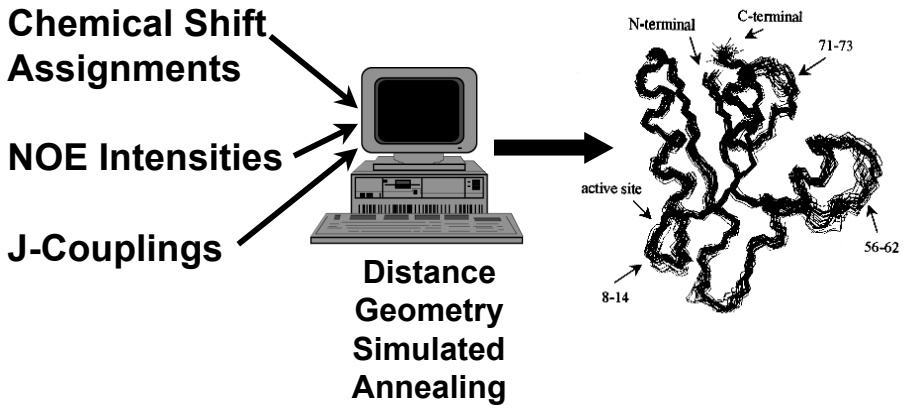


MW ~ 500 **MW ~ 10,000** **MW ~ 30,000**

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



The Final Result

```
ORIGX2      0.000000  1.000000  0.000000        0.00000                    2TRX 147
ORIGX3      0.000000  0.000000  1.000000        0.00000                    2TRX 148
SCALE1      0.011173  0.000000  0.004858        0.00000                    2TRX 149
SCALE2      0.000000  0.019585  0.000000        0.00000                    2TRX 150
SCALE3      0.000000  0.000000  0.018039        0.00000                    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
```

X-ray Versus NMR

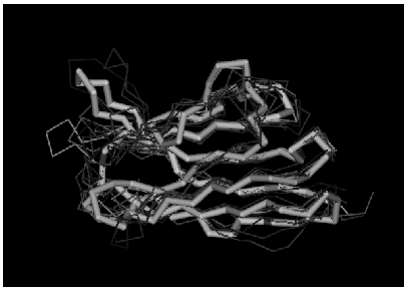
X-ray

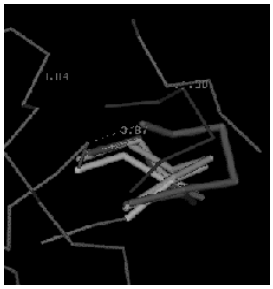
- Producing enough protein for trials
- Crystallization time and effort
- Crystal quality, stability and size control
- Finding isomorphous derivatives
- Chain tracing & checking

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

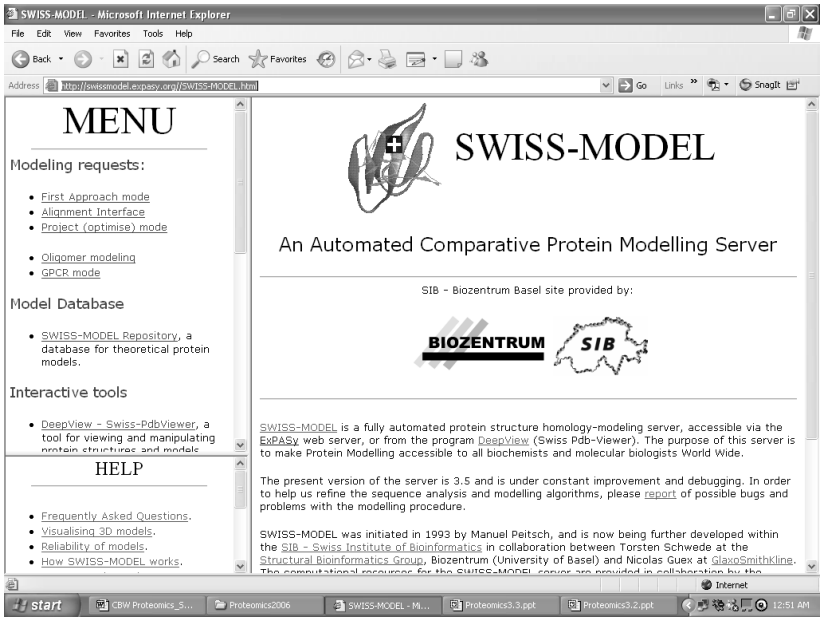
- Offers a method to "Predict" the 3D structure of proteins for which it is not possible to obtain X-ray or NMR data
- Can be used in understanding function, activity, specificity, etc.
- Of interest to drug companies wishing to do structure-aided drug design
- A keystone of Structural Proteomics

Homology Modelling

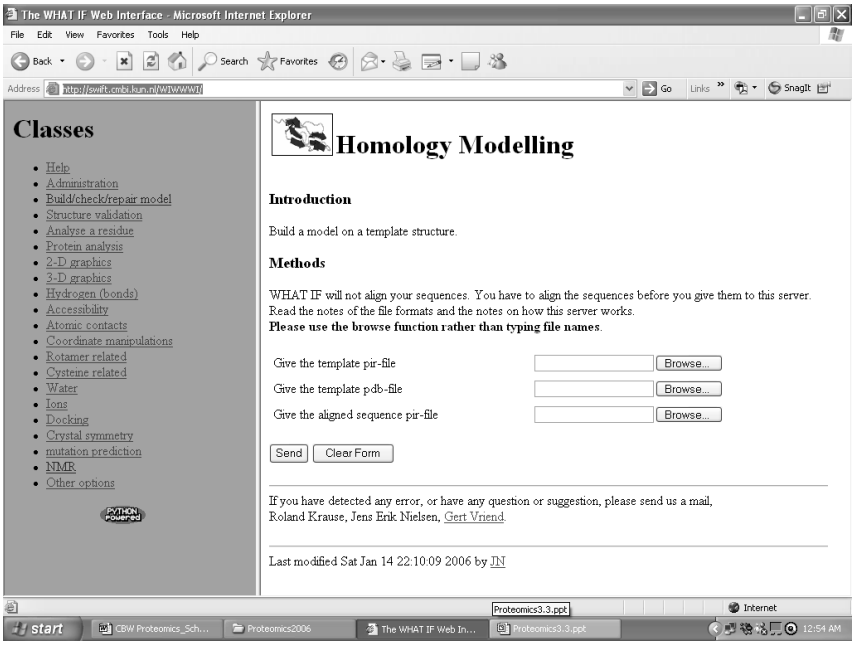
- Identify homologous sequences in PDB
- Align query sequence with homologues
- Find Structurally Conserved Regions (SCRs)
- Identify Structurally Variable Regions (SVRs)
- Generate coordinates for core region
- Generate coordinates for loops
- Add side chains (Check rotamer library)
- Refine structure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



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



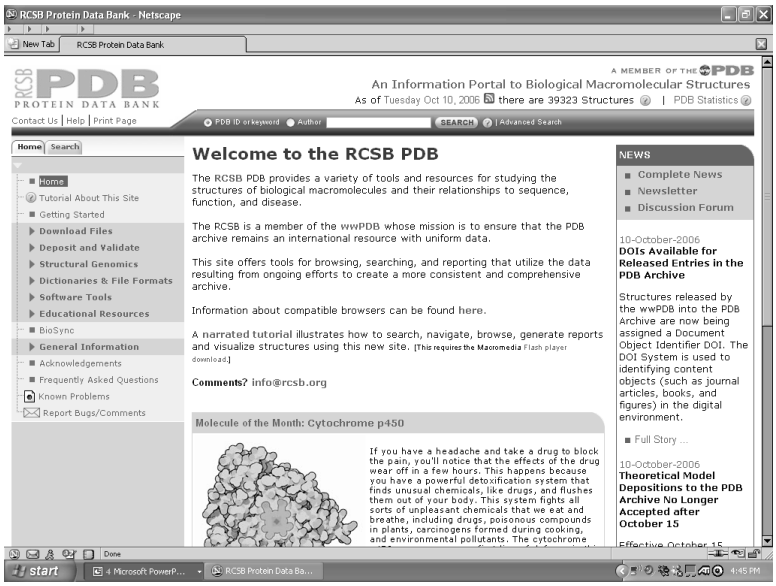
http://swift.cmbi.kun.nl/WIWWWI/

The Final Result

```
ORIGX2      0.000000  1.000000  0.000000        0.00000                  2TRX 147
ORIGX3      0.000000  0.000000  1.000000        0.00000                  2TRX 148
SCALE1      0.011173  0.000000  0.004858        0.00000                  2TRX 149
SCALE2      0.000000  0.019585  0.000000        0.00000                  2TRX 150
SCALE3      0.000000  0.000000  0.018039        0.00000                  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

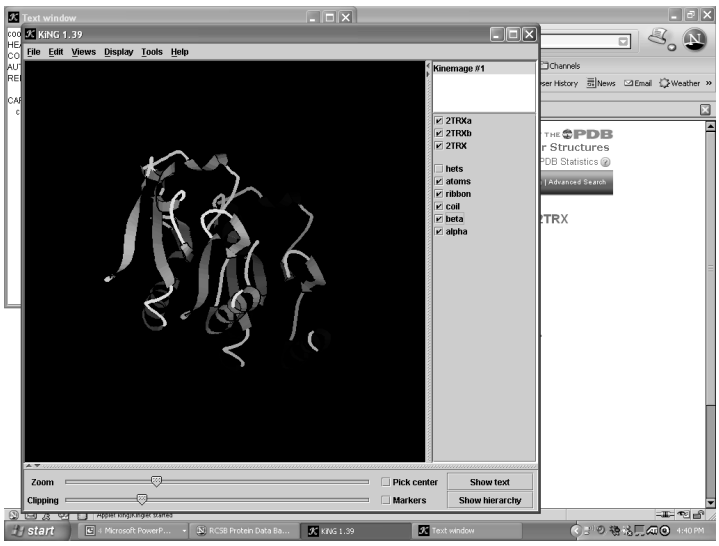
- PDB - Protein Data Bank
- Established in 1971 at Brookhaven National Lab (7 structures)
- Primary archive for macromolecular structures (proteins, nucleic acids, carbohydrates – now 40,000 structrs)
- Moved from BNL to RCSB (Research Collaboratory for Structural Bioinformatics) in 1998



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

Viewing 3D Structures

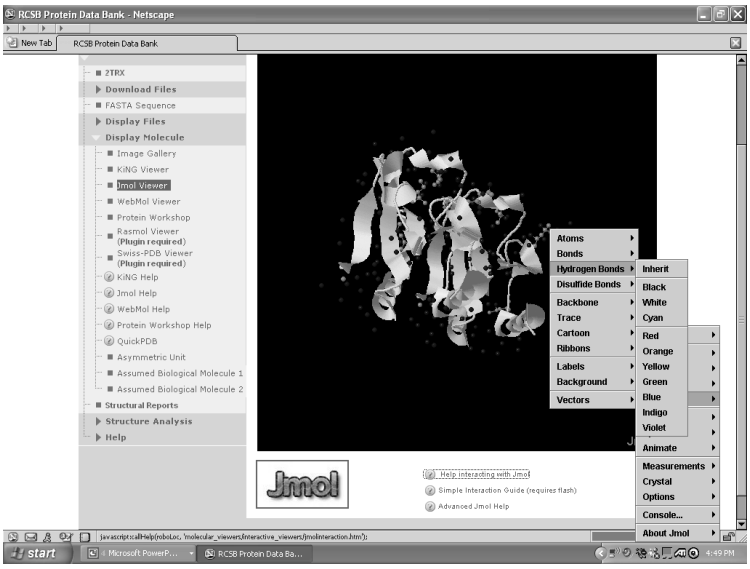
KiNG (Kinemage) 1.39



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

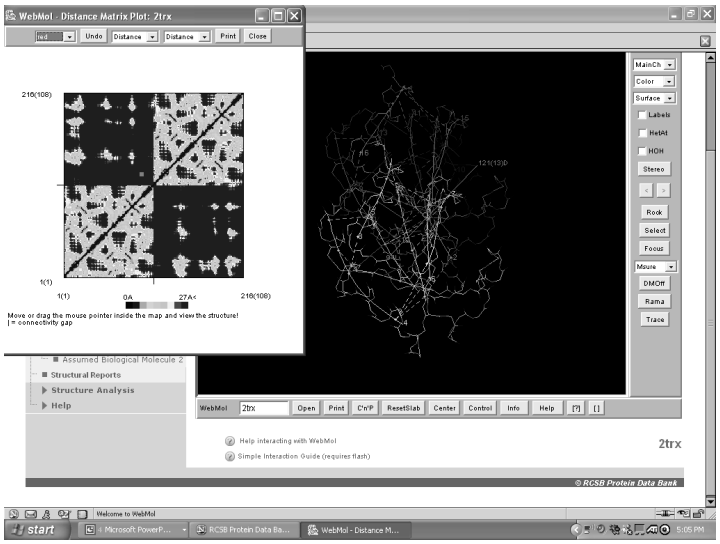
JMol Applet



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)

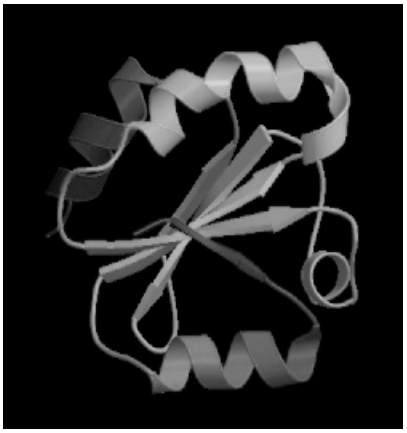
Structure Variability

X-ray to X-ray
Interleukin 1β
(41bi vs 2mlb)

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

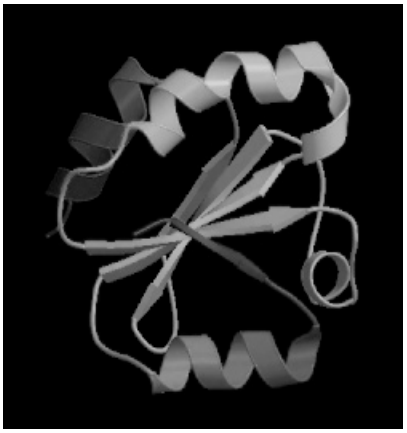
A Good Protein Structure..

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



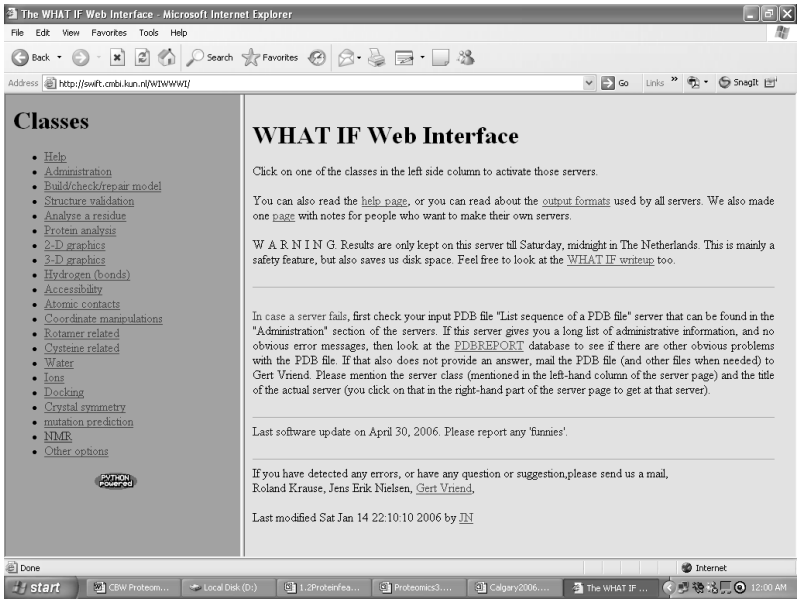
A Good Protein Structure..

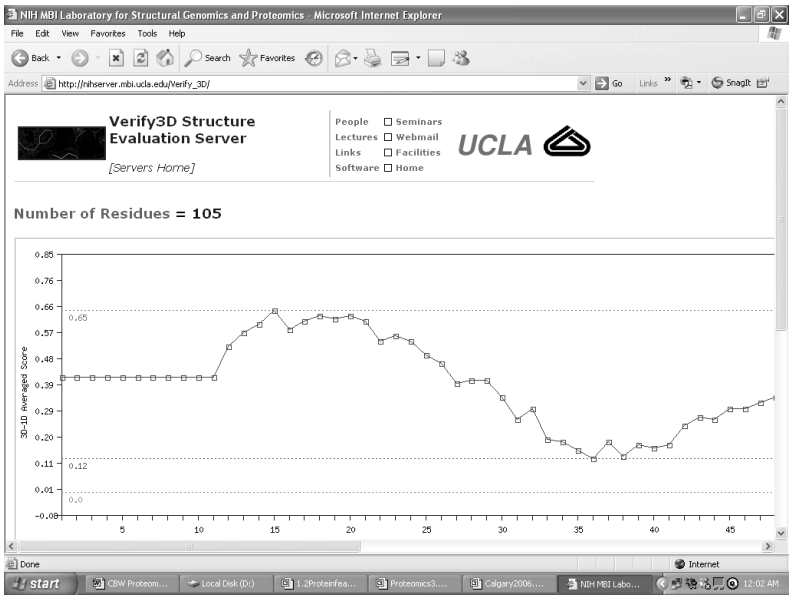
- Minimizes number of “bad” contacts
- Minimizes number of buried charges
- Minimizes radius of gyration
- Minimizes covalent and noncovalent (van der Waals and coulombic) energies



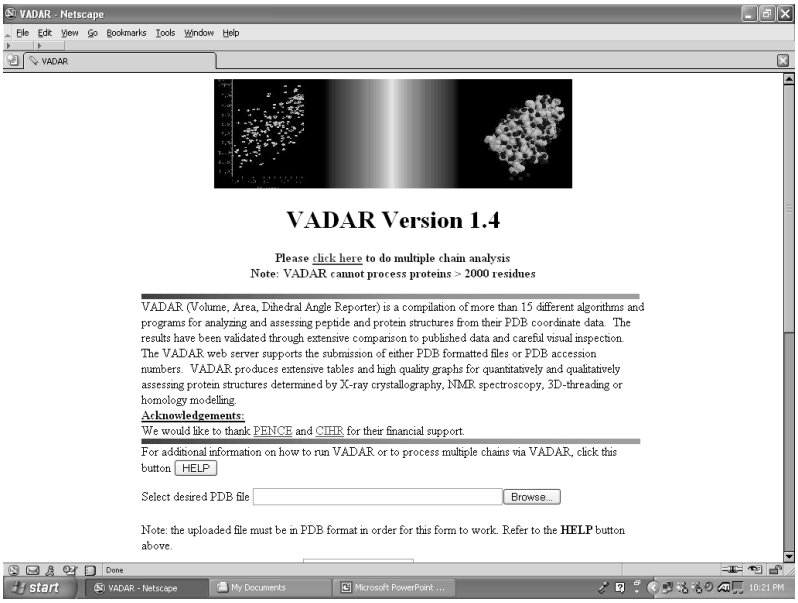
Structure Validation Servers

- **Whatlf Web Server -**
  http://swift.cmbi.kun.nl/WIWWWI/
- **Biotech Validation Suite -**
  http://biotech.ebi.ac.uk:8400/cgi-bin/sendquery
- **Verify3D -**
  http://www.doe-mbi.ucla.edu/Services/Verify_3D/
- **VADAR - http://redpoll.pharmacy.ualberta.ca**





High scores = good Low scores = bad

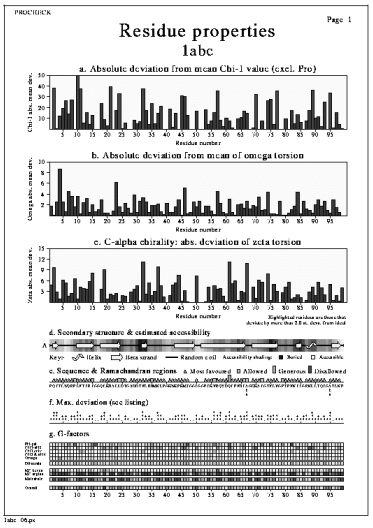


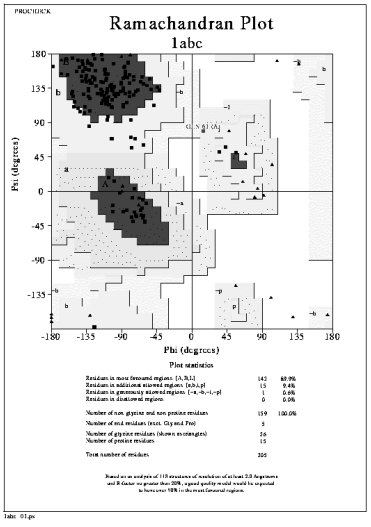
http://redpoll.pharmacy.ualberta.ca

Structure Validation Programs

- **PROCHECK -**
  http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html
- **PROSA II -**
  http://lore.came.sbg.ac.at/People/mo/Prosa/prosa.html
- **VADAR -**
  http://www.pence.ualberta.ca/ftp/vadar/
- **DSSP -**
  http://www.embl-heidelberg.de/dssp/

Procheck

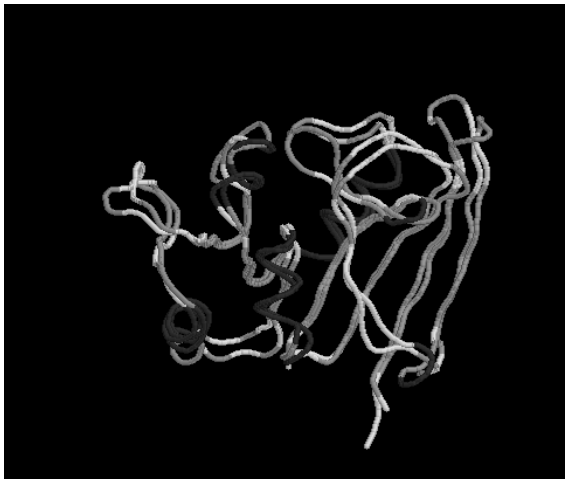




Comparing 3D Structures

Qualitative vs. Quantitative

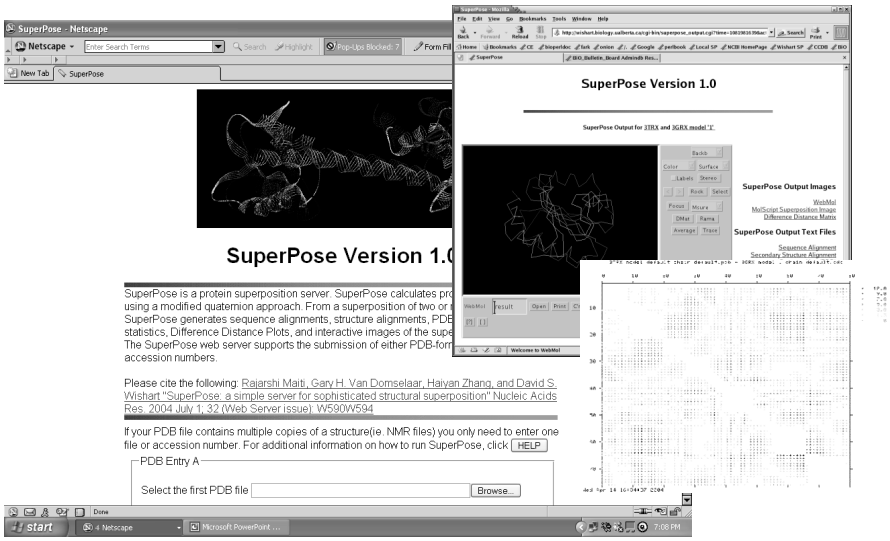
Rigid Body Superposition



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

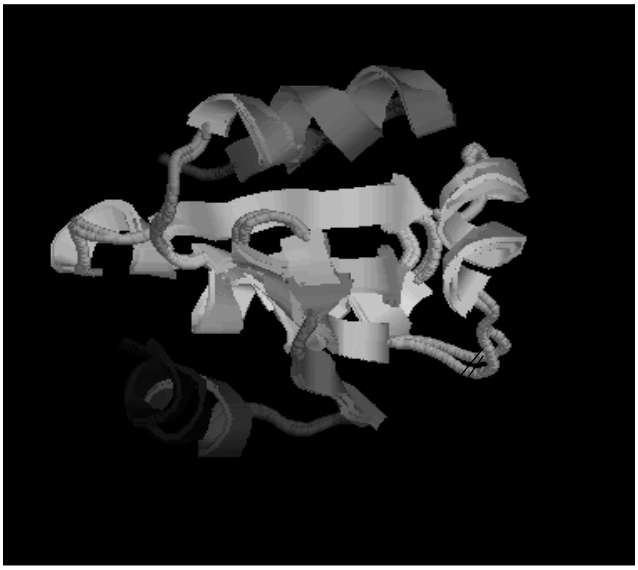


http://wishart.biology.ualberta.ca/SuperPose/

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

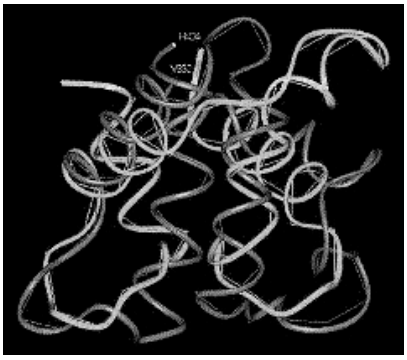
- Method to quantify structural similarity - same as standard deviation
- Requires 2 superimposed structures (designated here as "a" & "b")
- N = number of atoms being compared

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

RMSD

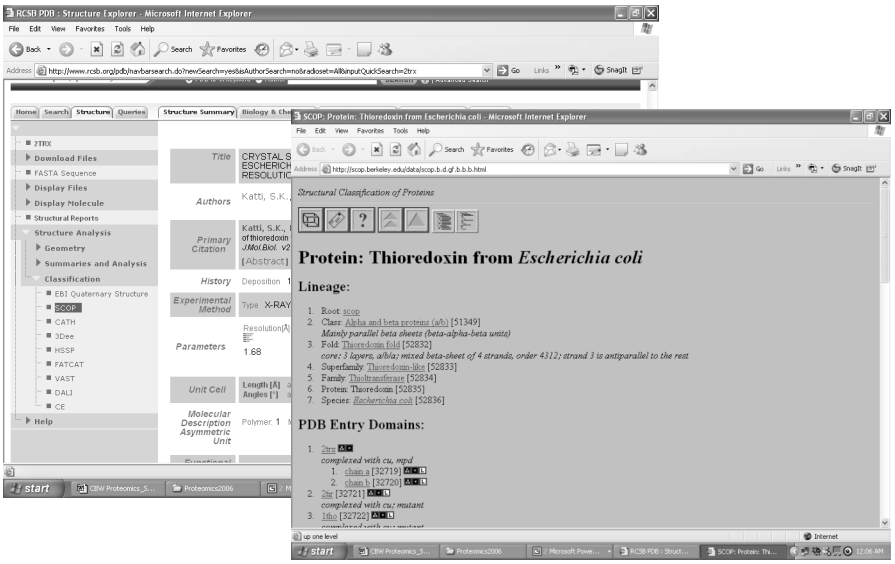
- 0.0-0.5 Å → Essentially Identical
- <1.5 Å → Very good fit
- < 5.0 Å → Moderately good fit
- 5.0-7.0 Å → Structurally related
- > 7.0 Å → Dubious relationship
- > 12.0 Å → Completely unrelated

Detecting Unusual Relationships

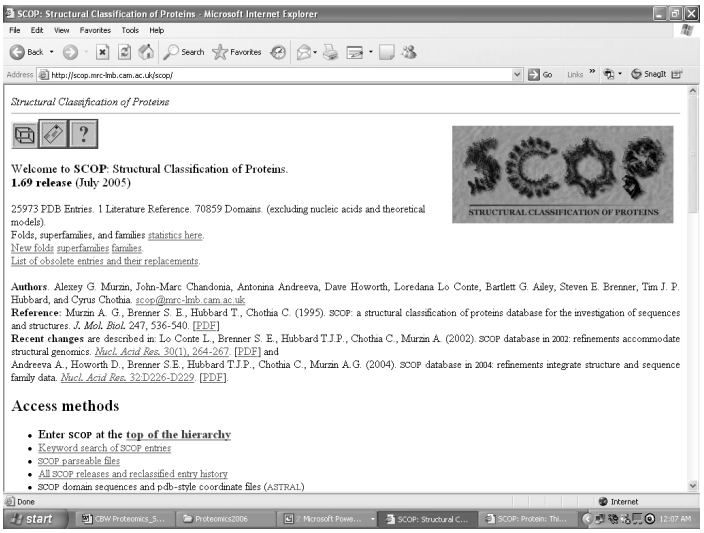


Similarity between Calmodulin and Acetylcholinesterase

Classifying Protein Folds



SCOP Database

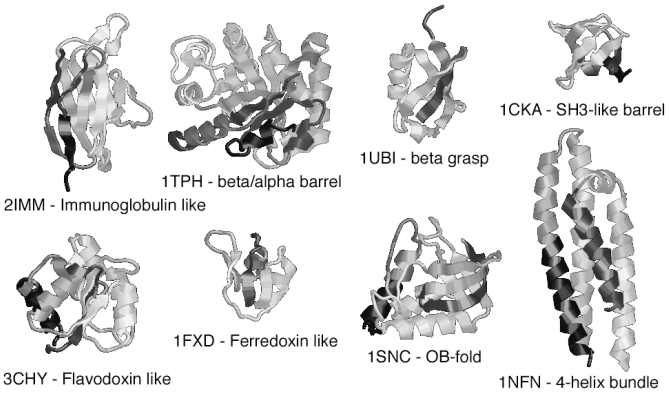


http://scop.mrc-lmb.cam.ac.uk/scop

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 clusers of proteins with seq ID > 30% with v. similar struct. & function

SCOP Structural Classification



The eight most frequent SCOP superfolds

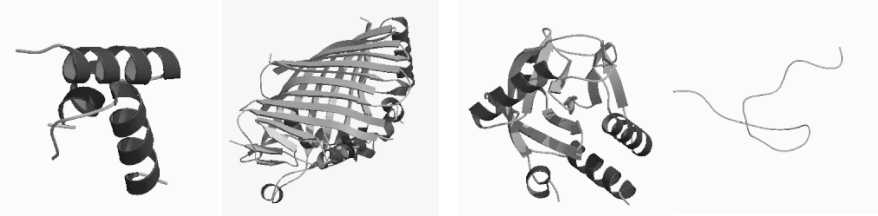
The CATH Database

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

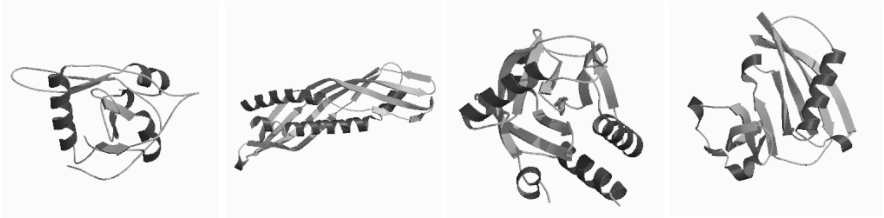
Class 2: Mainly Beta

Class 3: Mixed Alpha/Beta

Class 4: Few Secondary Structures

Secondary structure content (automatic)

CATH - Architecture



Roll | Super Roll | Barrel | 2-Layer Sandwich

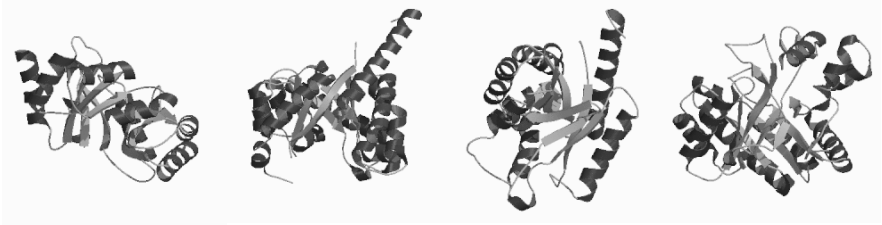
Orientation of secondary structures (manual)

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

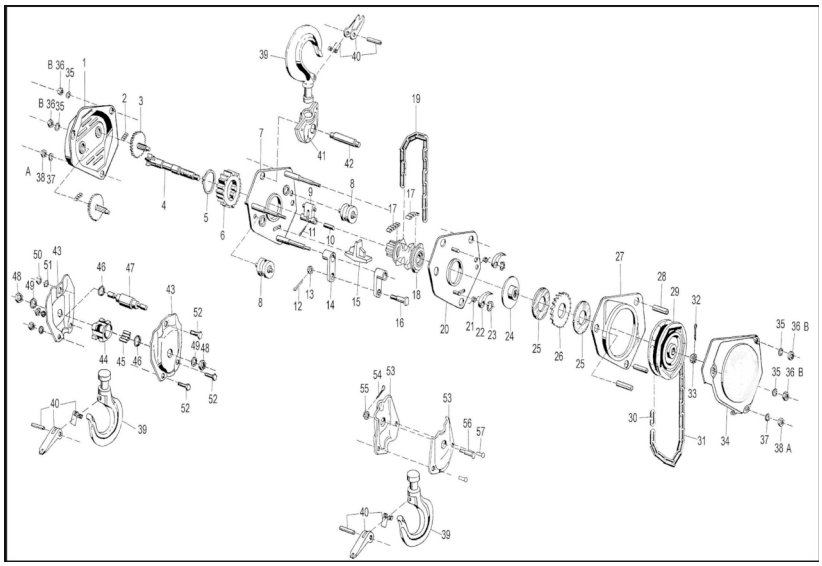
Superfamily clusters of similar structures & functions

Other Servers/Databases

- **Dali** - http://www.ebi.ac.uk/dali/
- **VAST** - www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml
- **CE** - http://cl.sdsc.edu/ce.html
- **FSSP** - http://www.ebi.ac.uk/dali/fssp/fssp.html
- **PDBsum** - www.biochem.ucl.ac.uk/bsm/pdbsum/

Protein Interactions

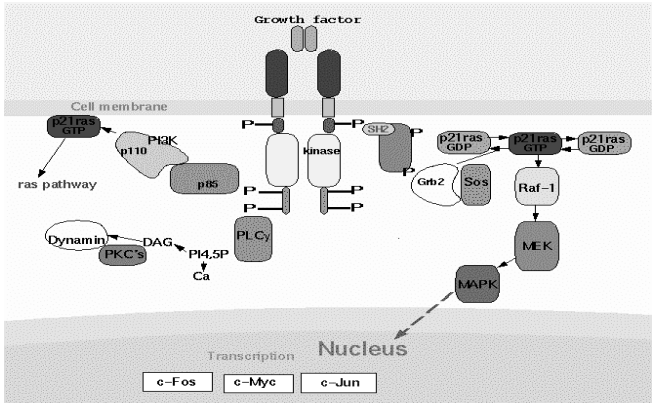
The Protein Parts List



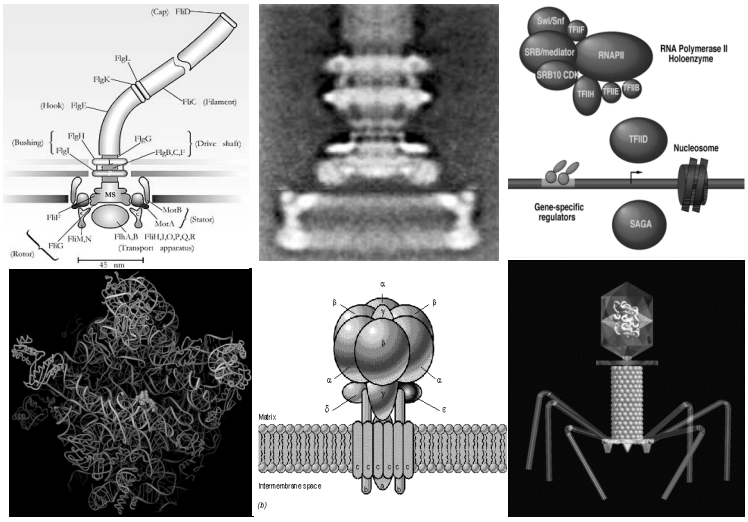
The Parts List

- 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*

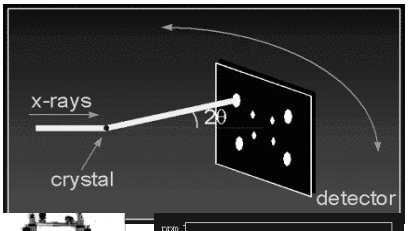


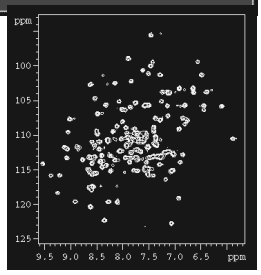
Types of Interactions

- **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

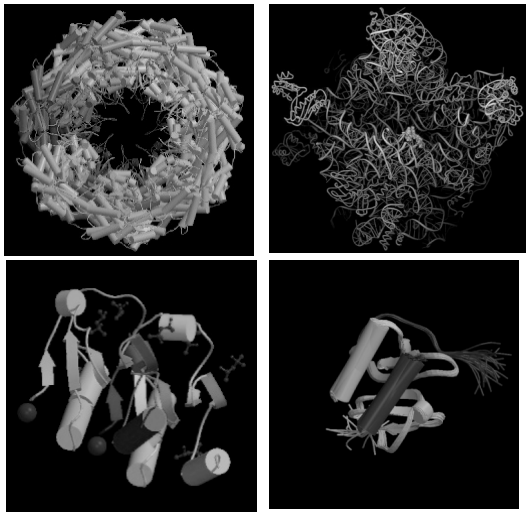
3D Structure Determination





- **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.**

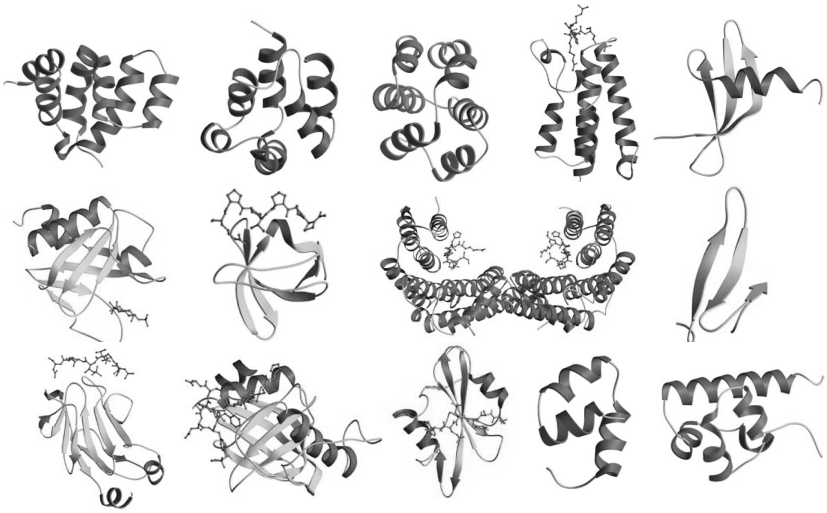
Quaternary Structure



Some interactions are real

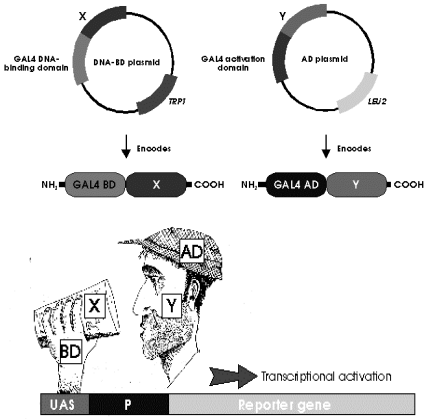
Others are not

Protein Interaction Domains



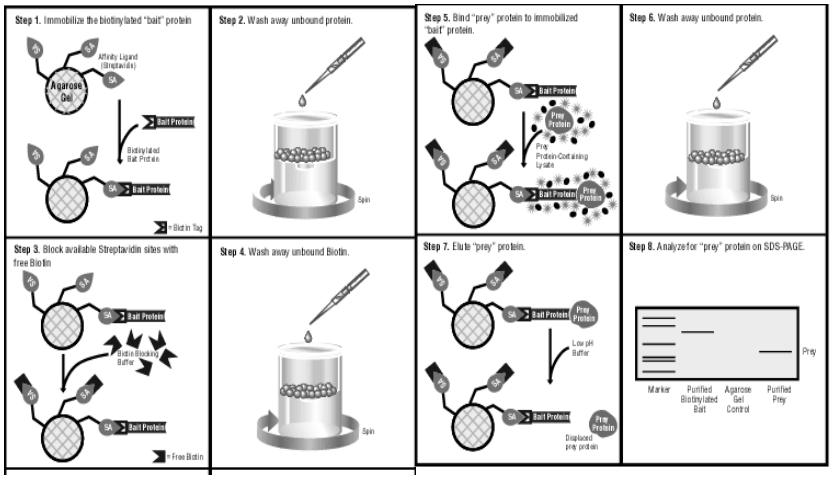
http://www.mshri.on.ca/pawson/domains.html

Yeast Two-Hybrid Analysis

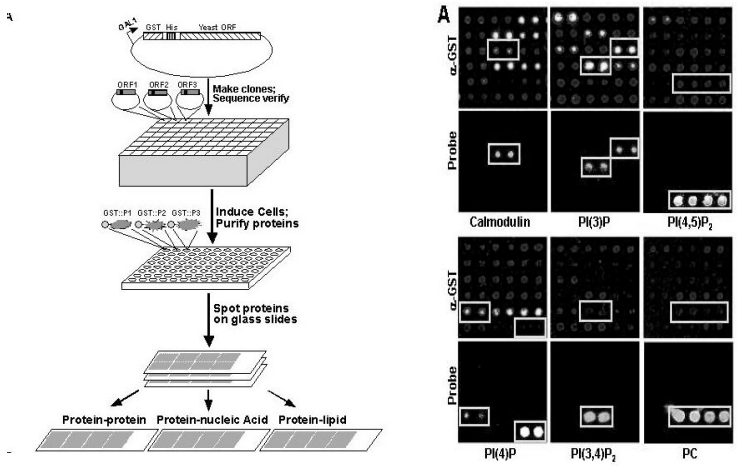


- Yeast two-hybrid experiments yield information on protein protein interactions
- GAL4 Binding Domain
- GAL4 Activation Domain
- X and Y are two proteins of interest
- If X & Y interact then reporter gene is expressed

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

- **BIND**
  - **http://www.bind.ca/**
- **DIP**
  - **http://dip.doe-mbi.ucla.edu/**
- **MINT**
  - **http://160.80.34.4/mint/**
- **IntAct**
  - **http://www.ebi.ac.uk/intact/index.jsp**

More Protein Interaction Databases
http://www.hgmp.mrc.ac.uk/GenomeWeb/prot-interaction.html

Reliability of HT Interaction Data (Patil & Nakamura, BMC Bioinf. 6:100, 2005)

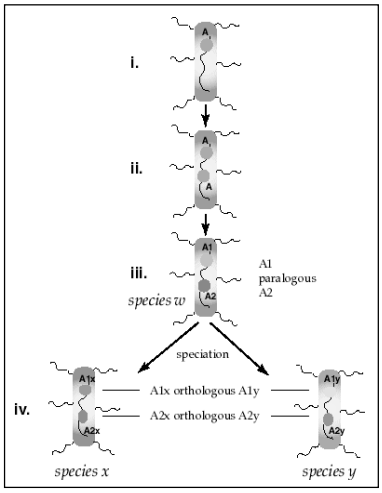
- **Assessed reliability using known interacting Pfam domains, Gene Ontology annotations and sequence homology**
- **56% of HT data for yeast are reliable**
- **27% of HT data for C. elegans are reliable**
- **18% of HT data for D. melanogaster are reliable**
- **68% of HT data for H. sapiens are reliable**

Protein Interaction Tools and Techniques - Computational Methods

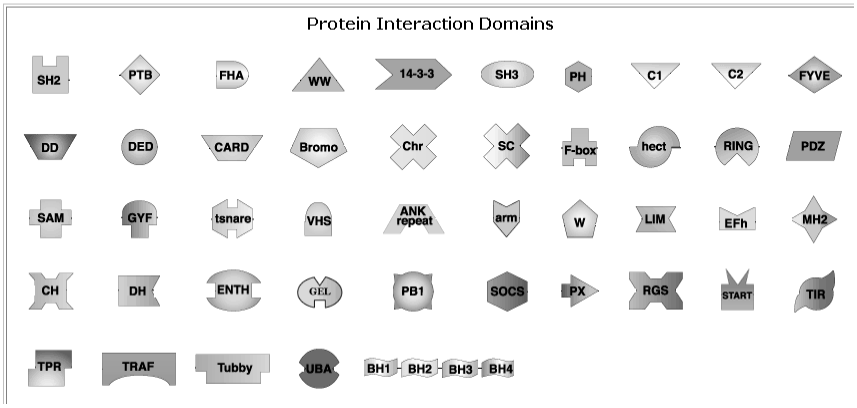
Interologs, Homologs, Paralogs...

- **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



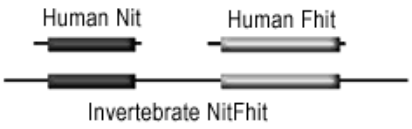
http://www.mshri.on.ca/pawson/domains.html

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:

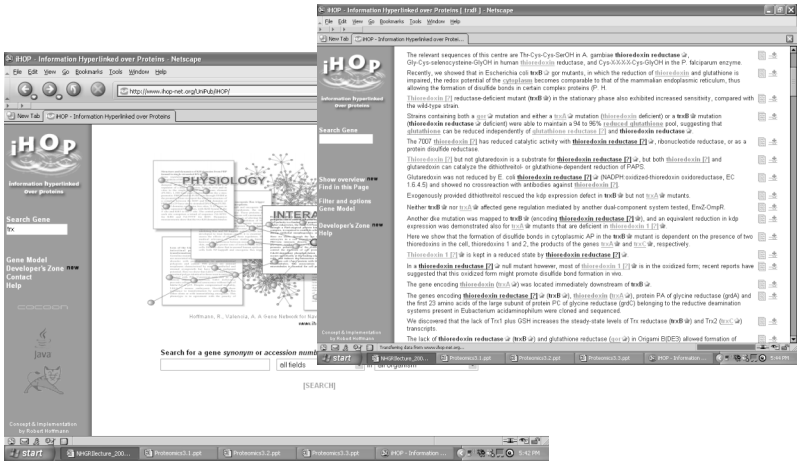
- fused in invertebrates
- form a heterocomplex in mammals



Text Mining

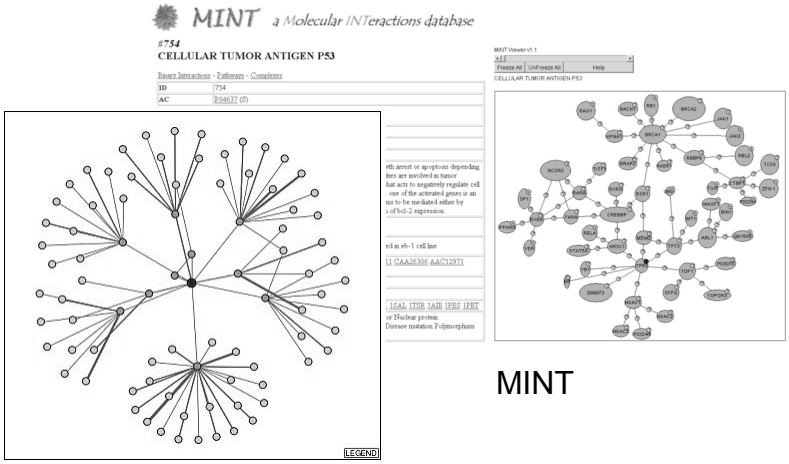
- Searching Medline or Pubmed for words or word combinations
- “X binds to Y”; “X interacts with Y”; “X associates with Y” etc. etc.
- Requires a list of known gene names or protein names for a given organism (a protein/gene thesaurus)

iHOP (Information hyperlinked over proteins)



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

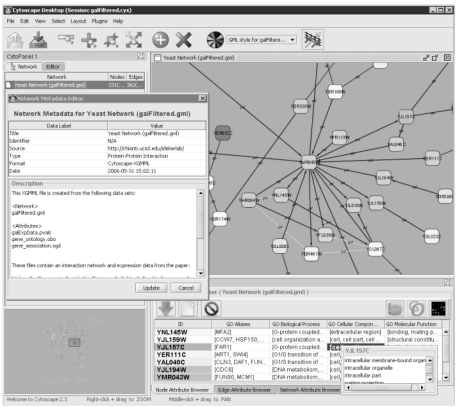
Visualizing Interactions

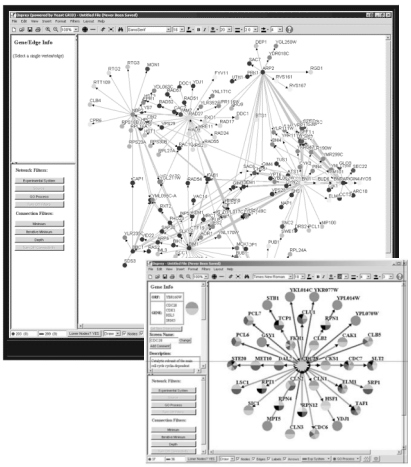


MINT

DIP

Visualizing Interactions

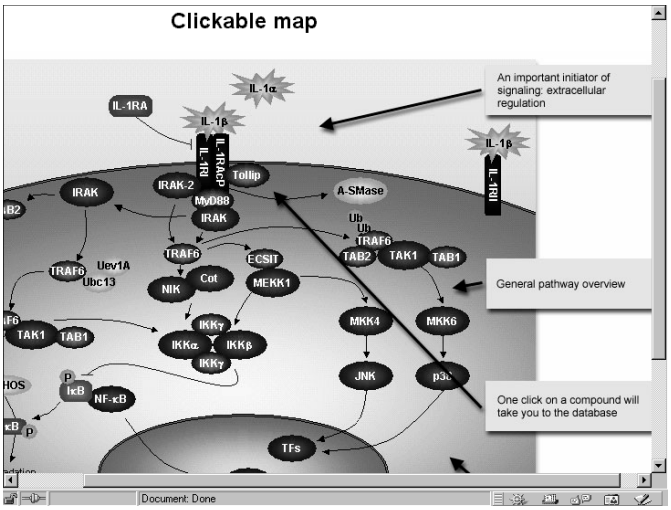




Cytoscape (www.cytoscape.org)

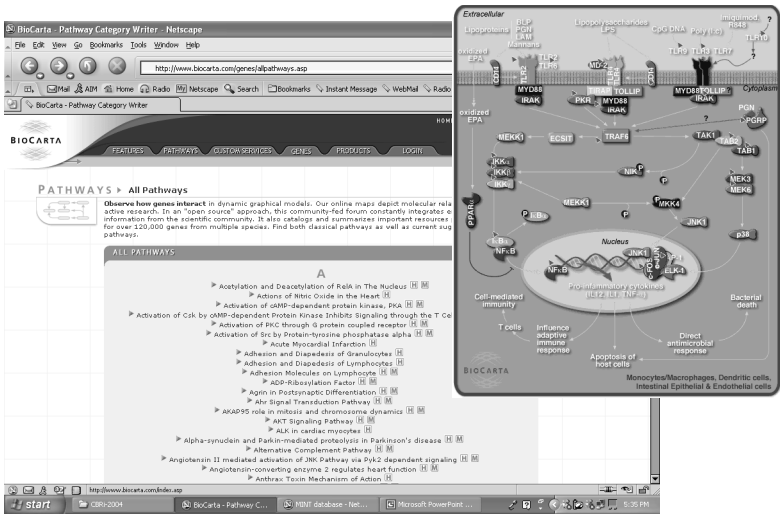
Osprey http://biodata.mshri.on.ca/osprey/servlet/Index

Pathway Visualization with TRANSPATH



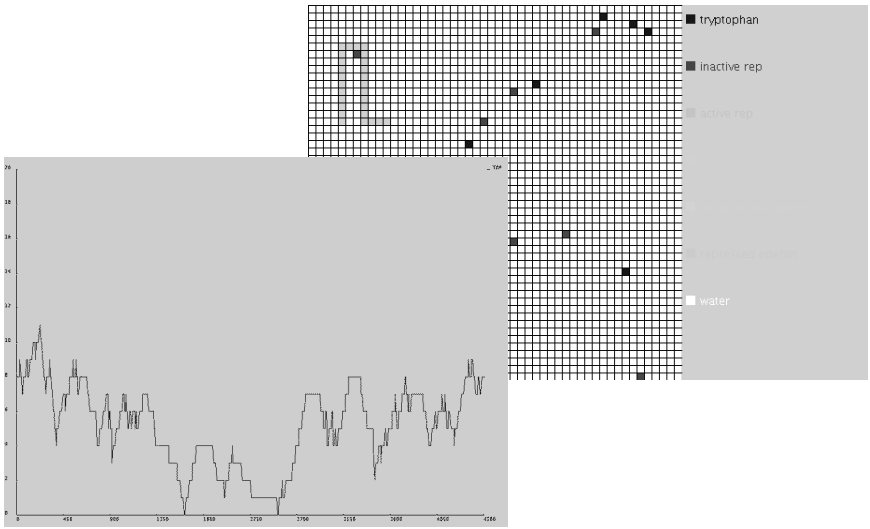
http://www.biobase.de/pages/products/transpath.html

Pathway Visualization with BioCarta



www.biocarta.com

Dynamic Simulations using SimCell



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