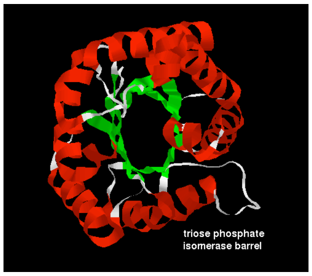
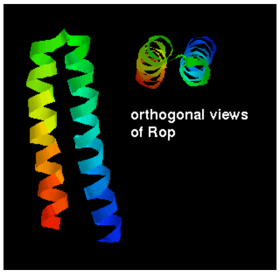
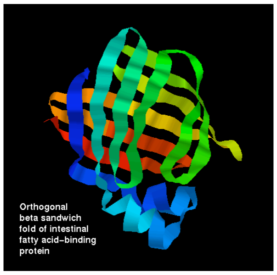
Protein Structure Assessment & Protein Interactions







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

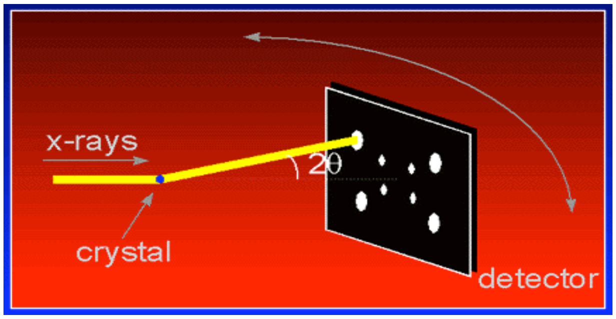
Much Ado About Structure

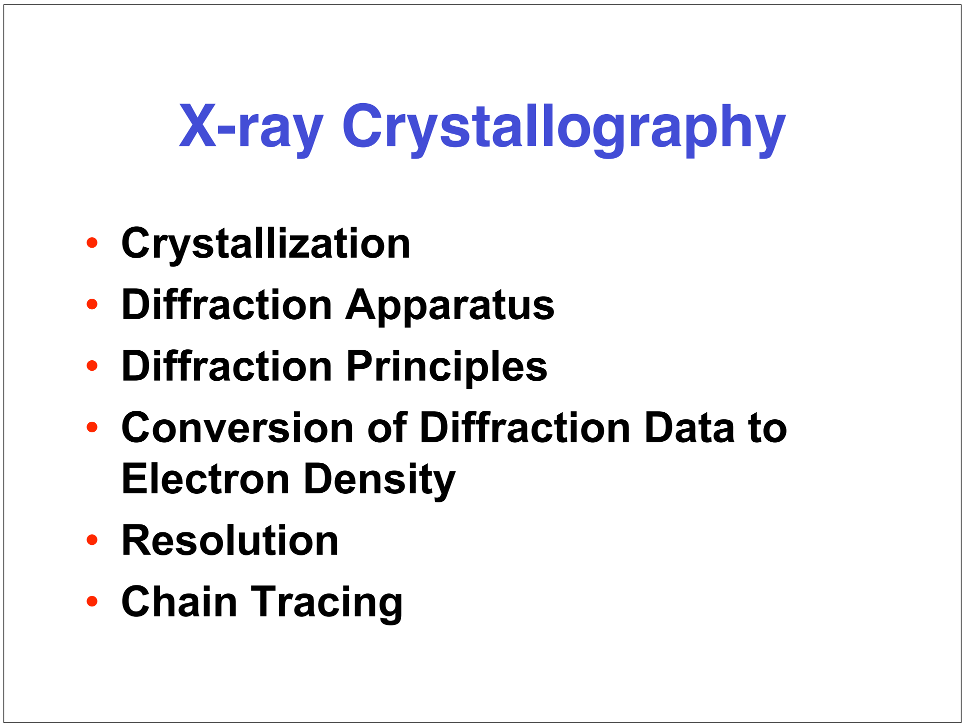
- **Structure ↔ Function**
- **Structure ↔ Mechanism**
- **Structure ↔ Origins/Evolution**
- **Structure-based Drug Design**
- **Solving the Protein Folding Problem**

Routes to 3D Structure

- X-ray Crystallography (the best)
- NMR Spectroscopy (close second)
- Cryoelectron microsocopy (distant 3rd)
- Homology Modelling (sometimes VG)
- Threading (sometimes VG)

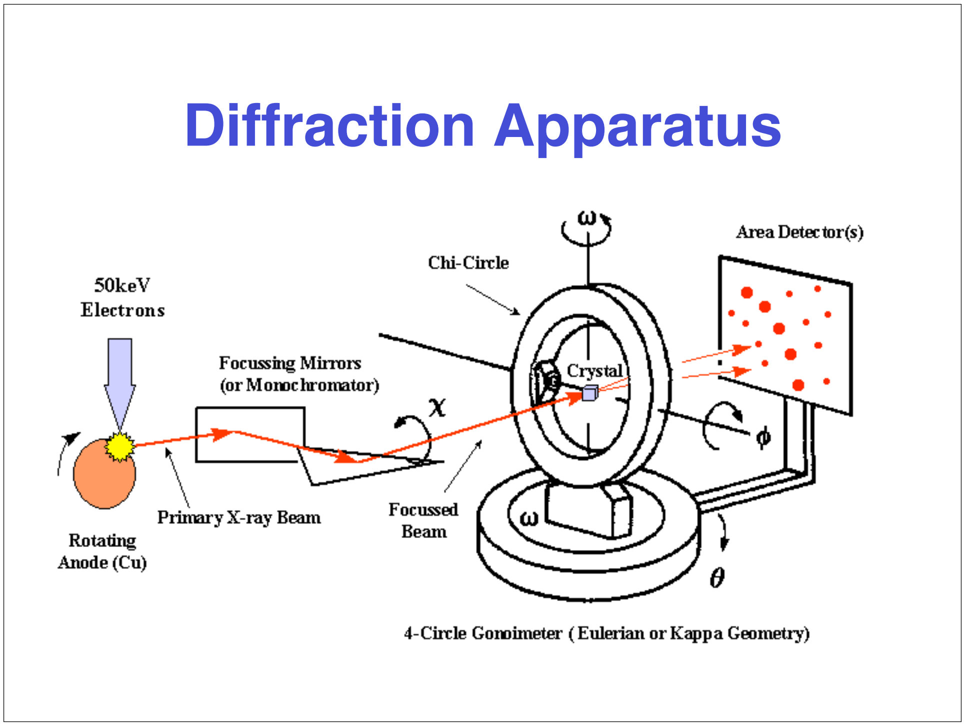
X-ray Crystallography





X-ray Crystallography

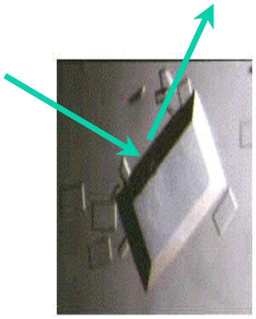
- Crystallization
- Diffraction Apparatus
- Diffraction Principles
- Conversion of Diffraction Data to Electron Density
- Resolution
- Chain Tracing

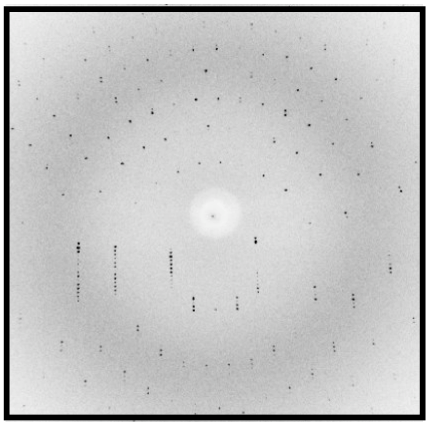


Diffraction Apparatus

4-Circle Gonoimeter (Eulerian or Kappa Geometry)

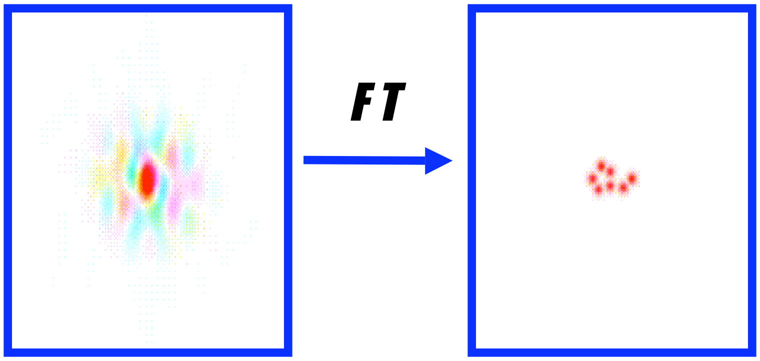
Protein Crystal Diffraction



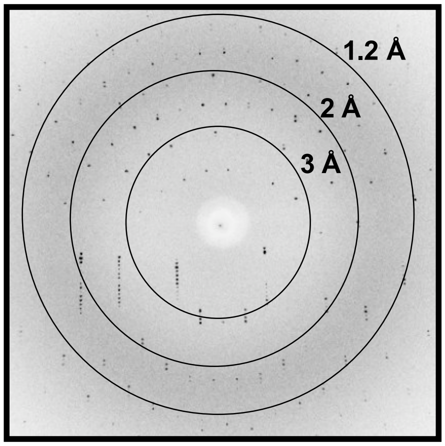


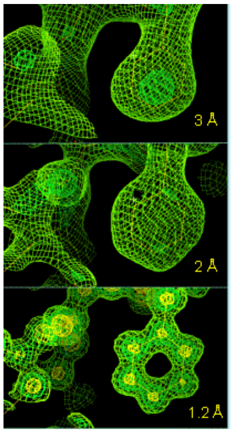
Diffraction Pattern

Converting Diffraction Data to Electron Density



Resolution



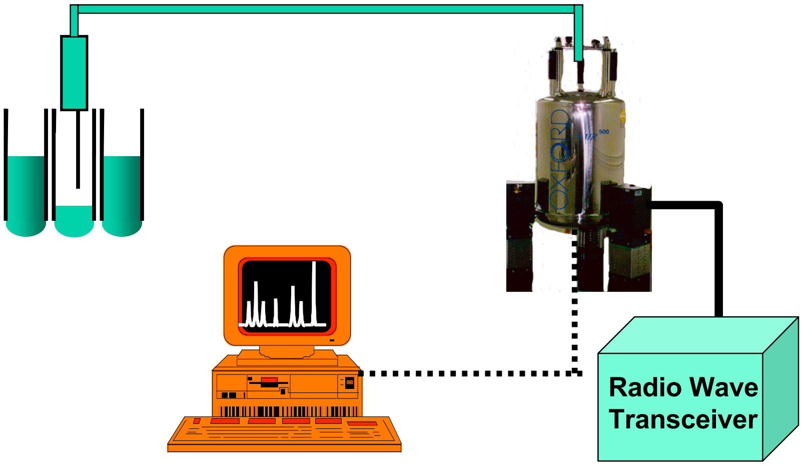


The Final Result

```
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
```

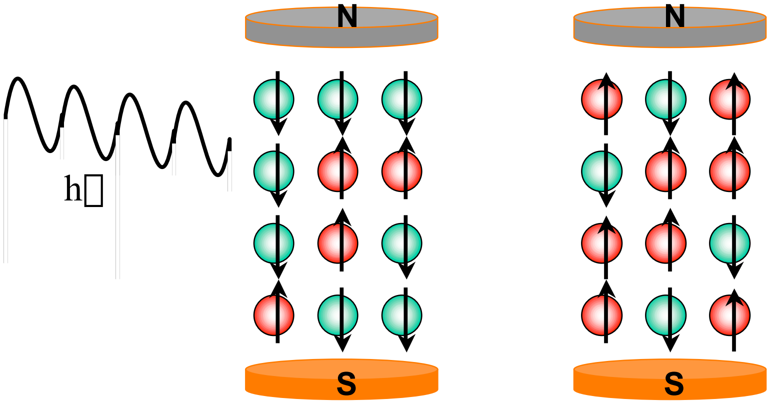
http://www-structure.llnl.gov/Xray/101index.html

NMR Spectroscopy



Radio Wave Transceiver

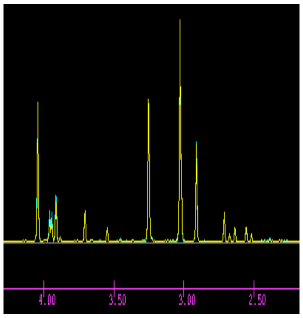
Principles of NMR



Low Energy High Energy

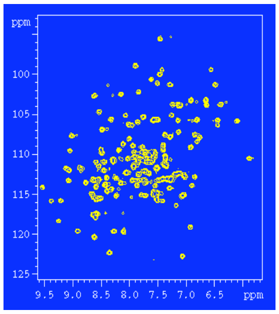
Multidimensional NMR

1D



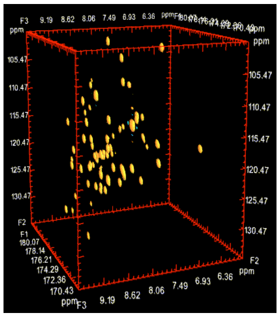
MW ~ 500

2D



MW ~ 10,000

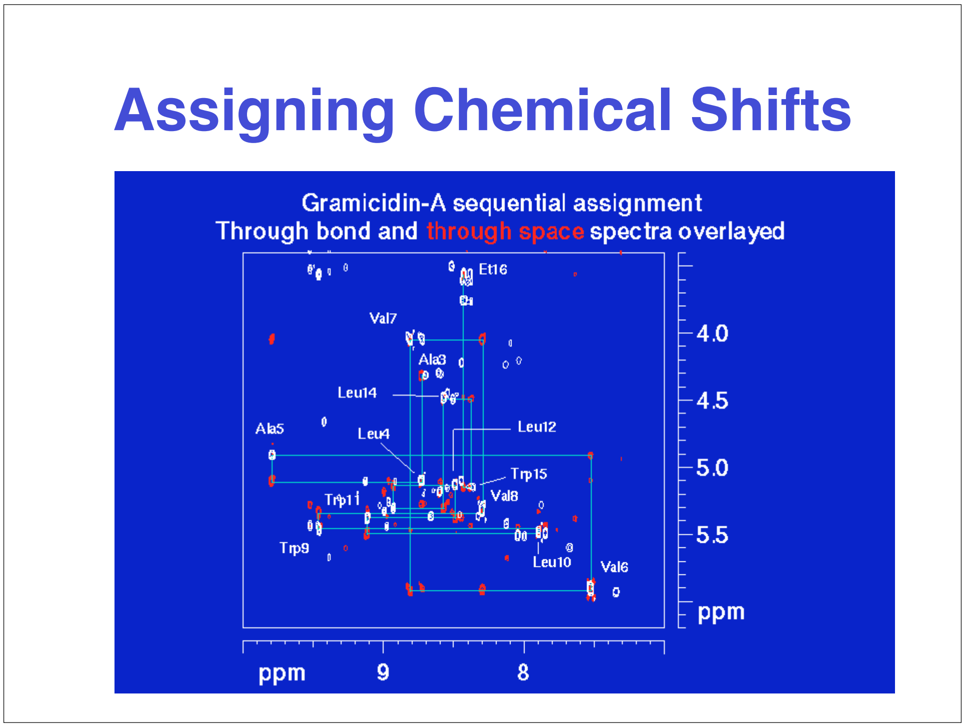
3D



MW ~ 30,000

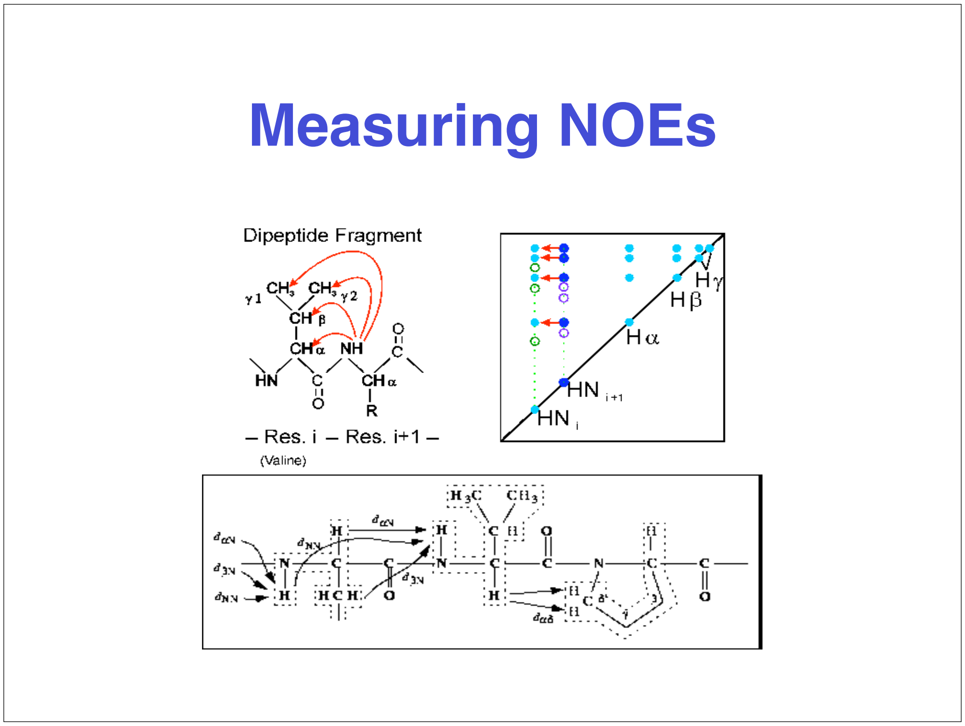
The NMR Process

- **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**

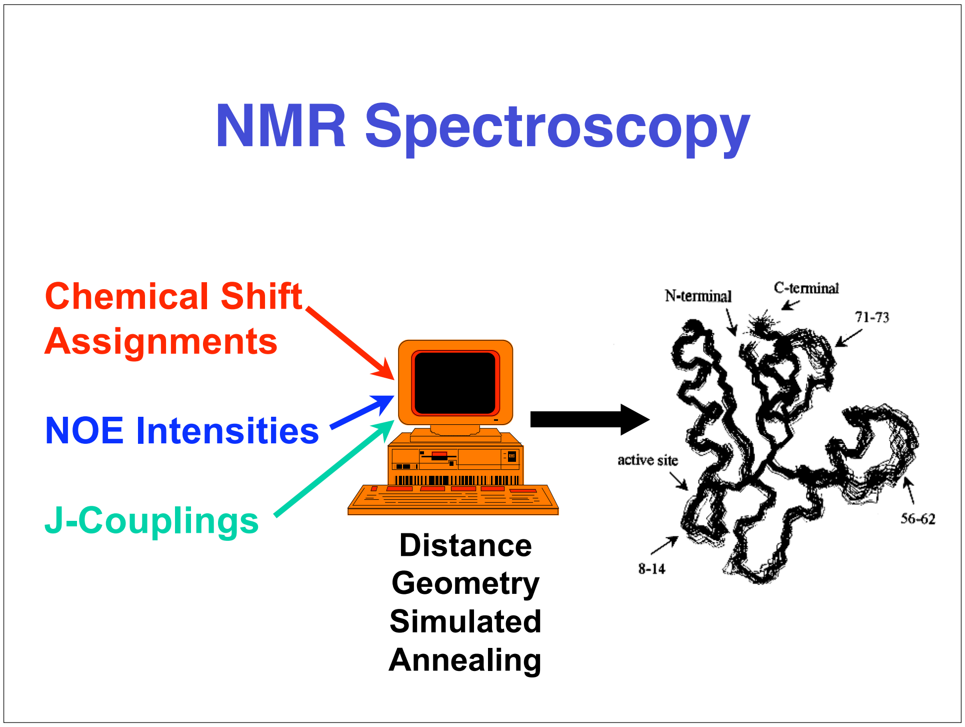


Assigning Chemical Shifts

Gramicidin-A sequential assignment
Through bond and through space spectra overlayed



Measuring NOEs



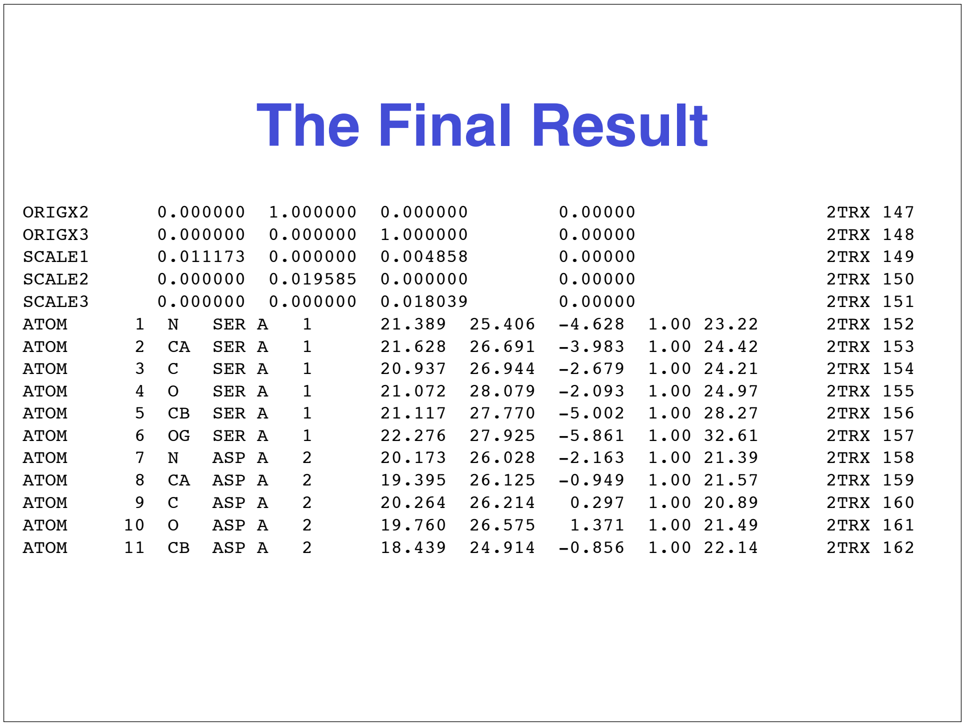
NMR Spectroscopy

Chemical Shift Assignments

NOE Intensities

J-Couplings

Distance Geometry Simulated Annealing



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

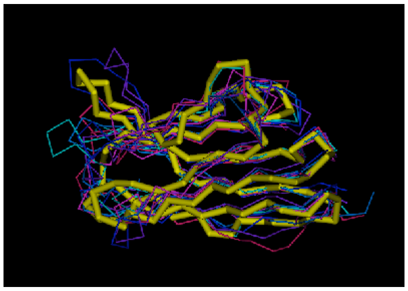
X-ray

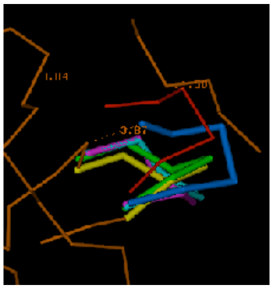
- Producing enough protein for trials
- Crystallization time and effort
- Crystal quality, stability and size control
- Finding isomorphous derivatives
- Chain tracing & checking

NMR

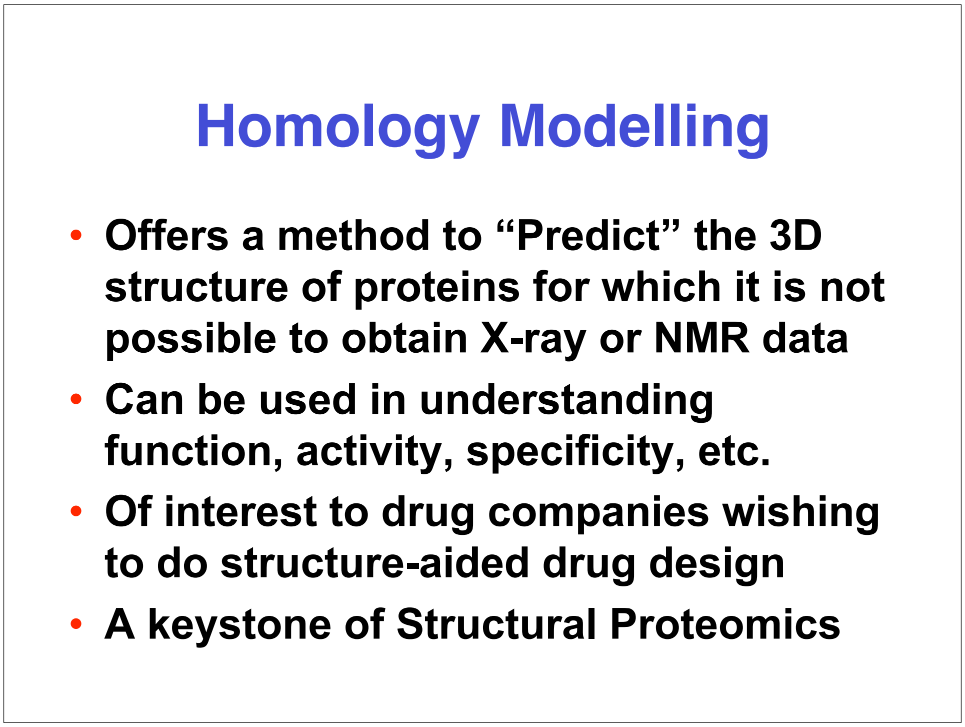
- Producing enough labeled protein for collection
- Sample “conditioning”
- Size of protein
- Assignment process is slow and error prone
- Measuring NOE’s is slow and error prone

Comparative (Homology) Modelling



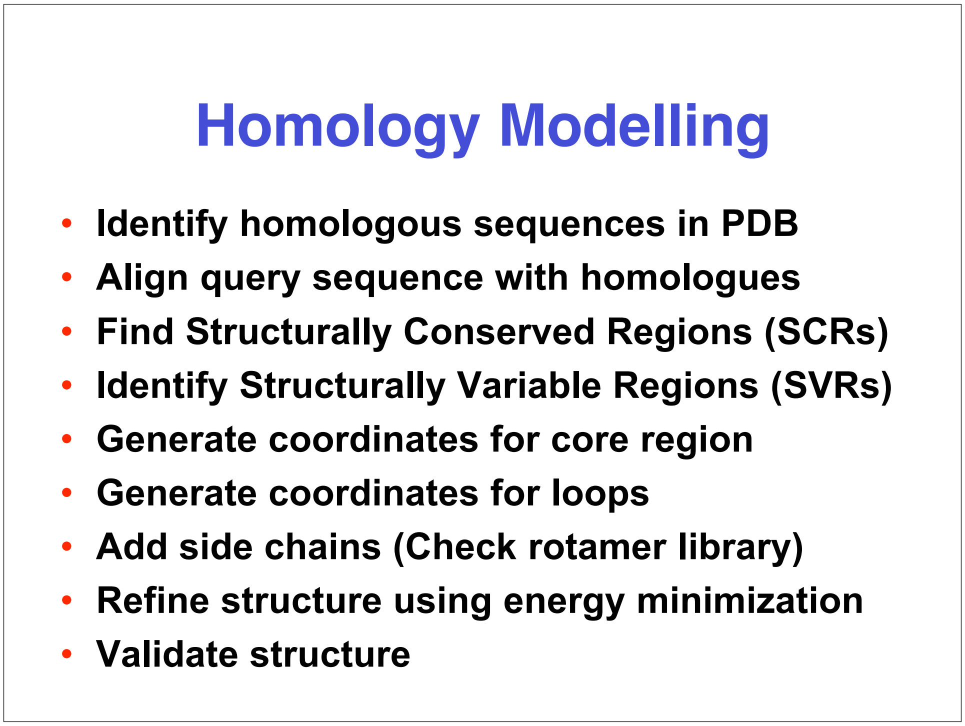


```
ACDEFGHIKLMNPQRST--FGHQWERT-----TYREWYEGHADS
ASDEYAHLRILDPQRSTVAYAYE--KSFAPPGSFKWEYEAHADS
MCDEYAHIRLMNPERSTVAGGHQWERT----GSFKEWYAAHADD
```



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

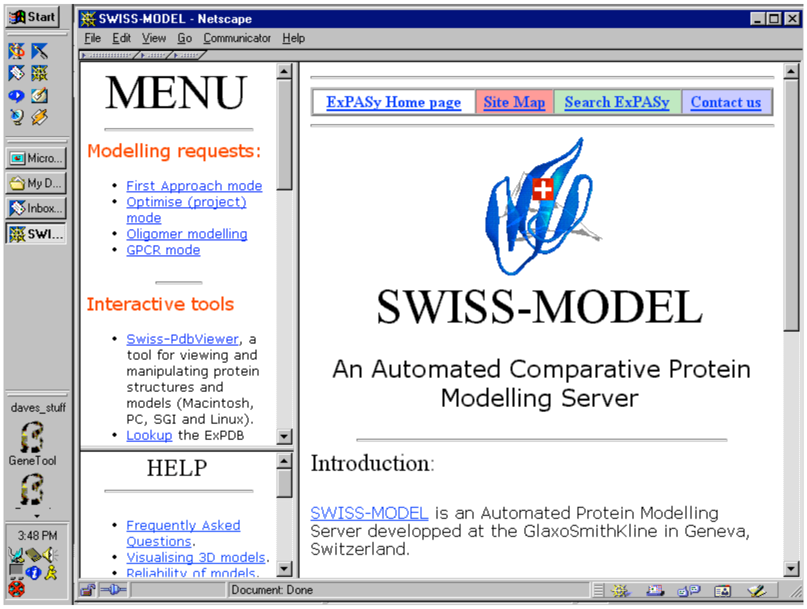


Homology Modelling

- 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!



MENU

Modelling requests:

- First Approach mode
- Optimise (project) mode
- Oligomer modelling
- GPCR mode

Interactive tools

- Swiss-PdbViewer, a tool for viewing and manipulating protein structures and models (Macintosh, PC, SGI and Linux).
- Lookup the ExPDB

HELP

- Frequently Asked Questions.
- Visualising 3D models.
- Reliability of models.

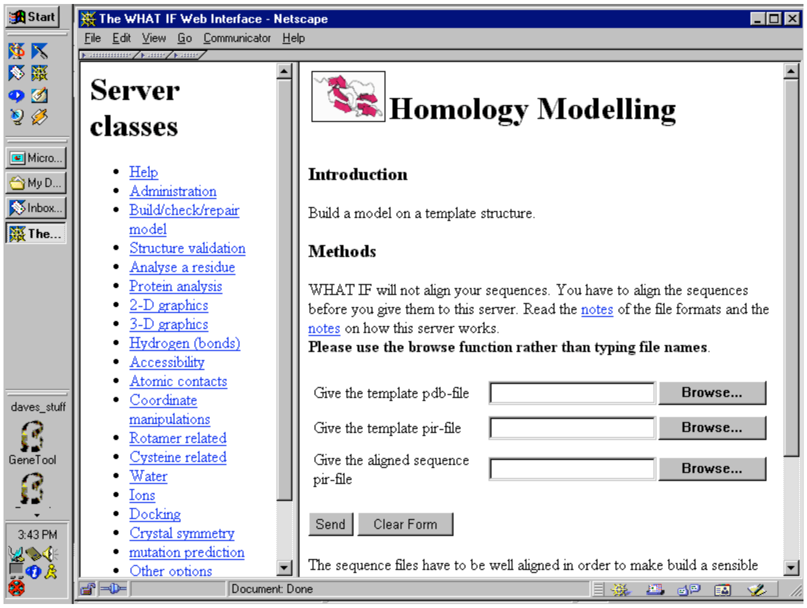
SWISS-MODEL

An Automated Comparative Protein Modelling Server

Introduction:

SWISS-MODEL is an Automated Protein Modelling Server developped at the GlaxoSmithKline in Geneva, Switzerland.

http://www.expasy.ch/swissmod/SWISS-MODEL.html



Server classes

- Help
- Administration
- Build/check/repair model
- Structure validation
- Analyse 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
- Other options

Homology Modelling

Introduction

Build a model on a template structure.

Methods

WHAT IF will not align your sequences. You have to align the sequences before you give them to this server. Read the notes of the file formats and the notes on how this server works.
Please use the browse function rather than typing file names.

Give the template pdb-file [ ] Browse...

Give the template pir-file [ ] Browse...

Give the aligned sequence pir-file [ ] Browse...

Send Clear Form

The sequence files have to be well aligned in order to make build a sensible

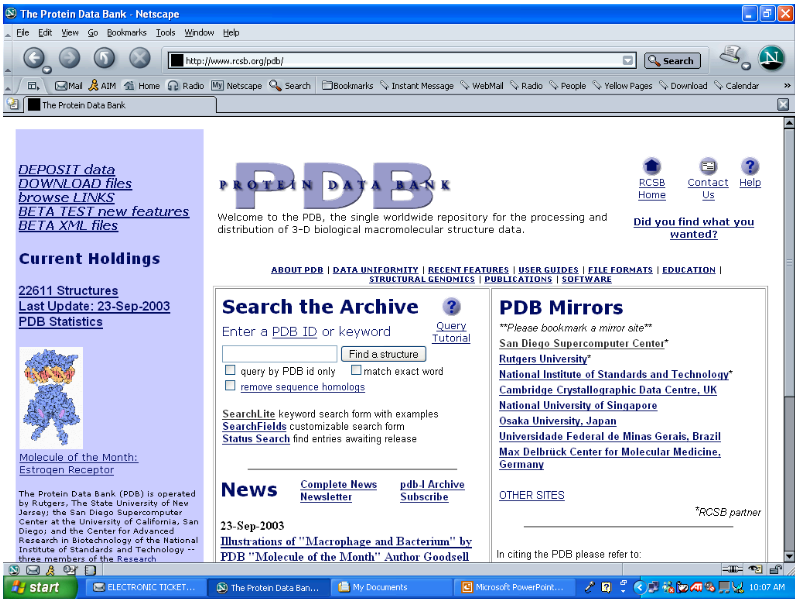
http://www.cmbi.kun.nl:1100/WIWWWI/

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)
- Moved from BNL to RCSB (Research Collaboratory for Structural Bioinformatics) in 1998



DEPOSIT data
DOWNLOAD files
browse LINKS
BETA TEST new features
BETA XML files

Current Holdings

22611 Structures
Last Update: 23-Sep-2003
PDB Statistics

Molecule of the Month: Estrogen Receptor

The Protein Data Bank (PDB) is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology of the National Institute of Standards and Technology -- three members of the Research

Welcome to the PDB, the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

Did you find what you wanted?

ABOUT PDB | DATA UNIFORMITY | RECENT FEATURES | USER GUIDES | FILE FORMATS | EDUCATION | STRUCTURAL GENOMICS | PUBLICATIONS | SOFTWARE

Search the Archive

Enter a PDB ID or keyword

Query Tutorial

□ query by PDB id only □ match exact word
□ remove sequence homologs

SearchLite keyword search form with examples
SearchFields customizable search form
Status Search find entries awaiting release

News Complete News Newsletter pdb-l Archive Subscribe

23-Sep-2003
Illustrations of "Macrophage and Bacterium" by PDB "Molecule of the Month" Author Goodsell

PDB Mirrors

Please bookmark a mirror site
San Diego Supercomputer Center*
Rutgers University*
National Institute of Standards and Technology*
Cambridge Crystallographic Data Centre, UK
National University of Singapore
Osaka University, Japan
Universidade Federal de Minas Gerais, Brazil
Max Delbrück Center for Molecular Medicine, Germany

OTHER SITES

*RCSB partner

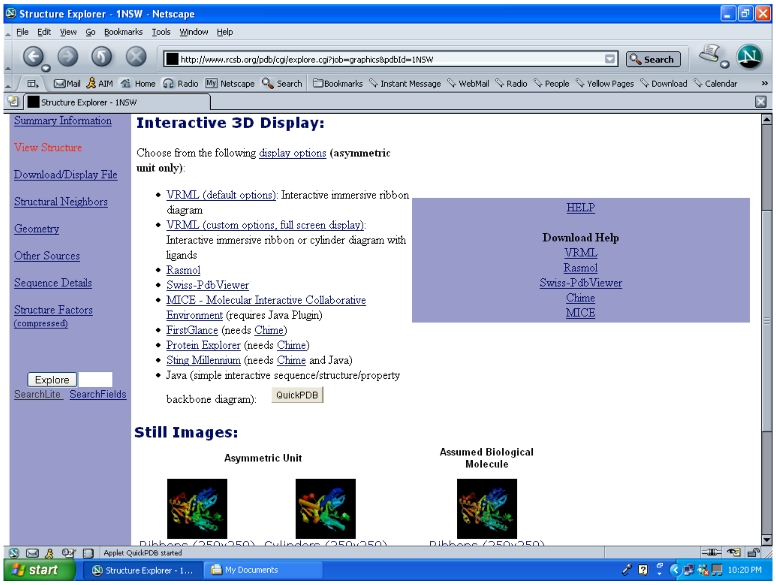
In citing the PDB please refer to:

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

The PDB

- Contains coordinate data (primarily) from X-ray, NMR and modelling
- Contains files in 2 formats
  - PDB format
  - mmCIF (macrmolecular Crystallographic Information File Format)
- Contains 22,000+ entries
- Currently growing exponentially

Viewing 3D Structures


Structure Explorer - 1NSW - Netscape
File Edit View Go Bookmarks Tools Window Help
Search
Mail AIM Home Radio Netscape Search Bookmarks Instant Message WebMail Radio People Yellow Pages Download Calendar
Structure Explorer - 1NSW
Summary Information
View Structure
Download/Display File
Structural Neighbors
Geometry
Other Sources
Sequence Details
Structure Factors (compressed)
Explore
SearchLite SearchFields
Applet QuickPDB started
start
Structure Explorer - 1...
My Documents
10:20 PM

http://www.rcsb.org/pdb/cgi/explore.cgi?job=graphics&pdbId=1NSW

Interactive 3D Display:

Choose from the following display options **(asymmetric unit only)**:

- VRML (default options): Interactive immersive ribbon diagram
- VRML (custom options, full screen display): Interactive immersive ribbon or cylinder diagram with ligands
- Rasmol
- Swiss-PdbViewer
- MICE - Molecular Interactive Collaborative Environment (requires Java Plugin)
- FirstGlance (needs Chime)
- Protein Explorer (needs Chime)
- Sting Millennium (needs Chime and Java)
- Java (simple interactive sequence/structure/property backbone diagram): QuickPDB

HELP

Download Help
VRML
Rasmol
Swiss-PdbViewer
Chime
MICE

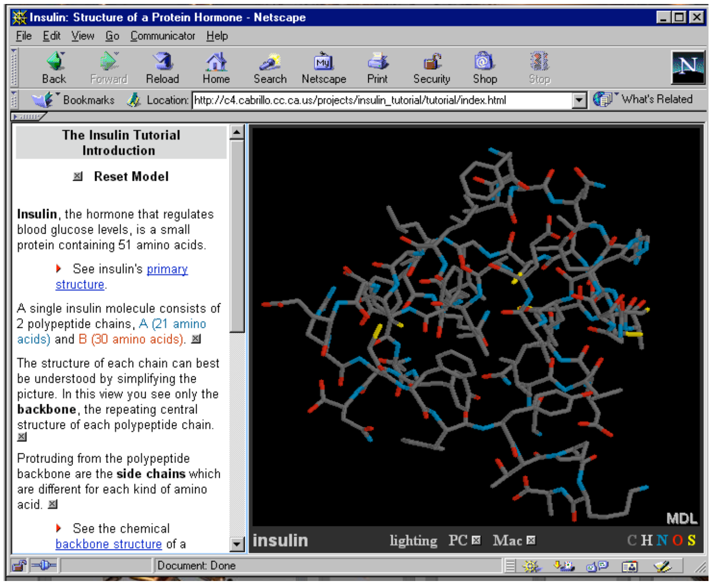
Still Images:

Asymmetric Unit

Assumed Biological Molecule

Ribbons (250x250) Cylinders (250x250) Ribbons (250x250)

Chime


Insulin: Structure of a Protein Hormone - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://c4.cabrillo.cc.ca.us/projects/insulin_tutorial/tutorial/index.html
What's Related
MDL
insulin
lighting PC ☒ Mac ☒
C H N O S
Document: Done

The Insulin Tutorial
Introduction

☒ **Reset Model**

Insulin, the hormone that regulates blood glucose levels, is a small protein containing 51 amino acids.

- See insulin's primary structure.

A single insulin molecule consists of 2 polypeptide chains, A (21 amino acids) and B (30 amino acids). ☒

The structure of each chain can best be understood by simplifying the picture. In this view you see only the **backbone**, the repeating central structure of each polypeptide chain. ☒

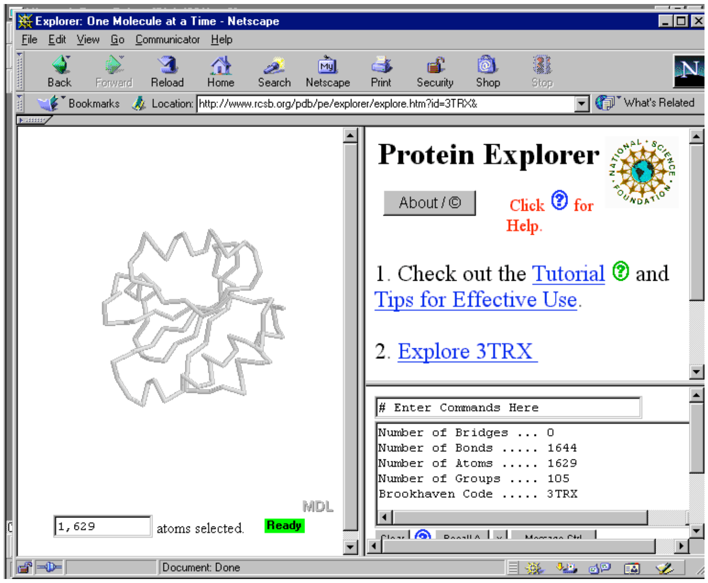
Protruding from the polypeptide backbone are the **side chains** which are different for each kind of amino acid. ☒

- See the chemical backbone structure of a

Chime

- http://www.mdlchime.com/chime/
- Very simple viewing program with limited manipulation capacity
- Uses Rasmol for its back end source
- View both large and small molecules
- Browser Plug-in (Like PDF reader)
- Compatible with Netscape 4.7X and higher as well as IE 5.5 and higher

Protein Explorer (Chime)


Explorer: One Molecule at a Time - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://www.rcsb.org/pdb/pe/explorer/explore.htm?id=3TRX&
What's Related
MDL
Document: Done

Protein Explorer

About / ©

Click ? for Help.

1. Check out the Tutorial and Tips for Effective Use.

2. Explore 3TRX

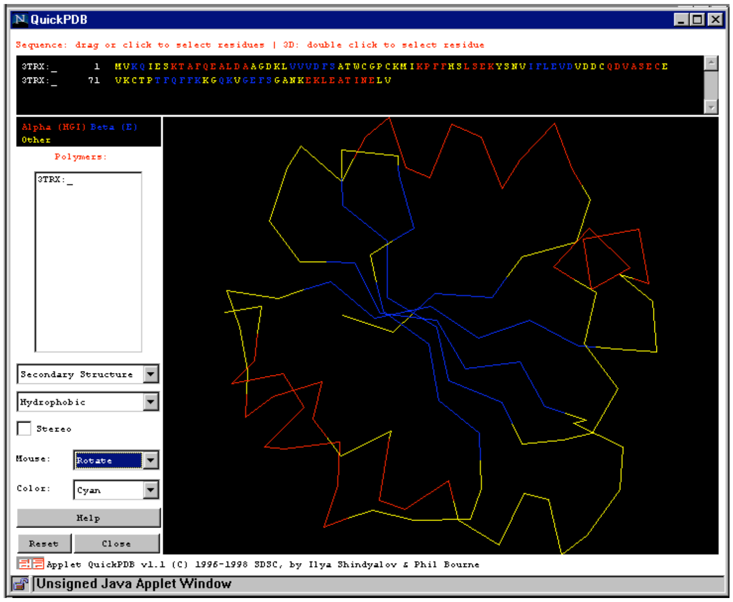
```
# Enter Commands Here
Number of Bridges ... 0
Number of Bonds ..... 1644
Number of Atoms ..... 1629
Number of Groups .... 105
Brookhaven Code ..... 3TRX
```

1,629 atoms selected. Ready

Protein Explorer

- http://www.umass.edu/microbio/chime/explorer/
- **Uses Chime & Rasmol for its back-end**
- **Very flexible, user friendly, well documented, offers morphing, sequence structure interface, comparisons, context-dependent help, smart zooming, off-line**
- **Browser Plug-in (Like PDF reader)**
- **Compatible with Netscape (Mac & Win)**

QuickPDB



QuickPDB

Sequence: drag or click to select residues | 3D: double click to select residue

```
3TRX:_      1  MVKQIESKTAFQEALDAAGDKLVVVDFSATWCGPCKMIKPFFHSLSEKYSNVIFLEVDVDDCQDVASECE
3TRX:_     71  VKCTPTFQFFKKGQKVGEFSGANKEKLEATINELV
```

Alpha (HGI) Beta (E) Other

Polymers:

3TRX:_

Secondary Structure

Hydrophobic

□ Stereo

Mouse: Rotate

Color: Cyan

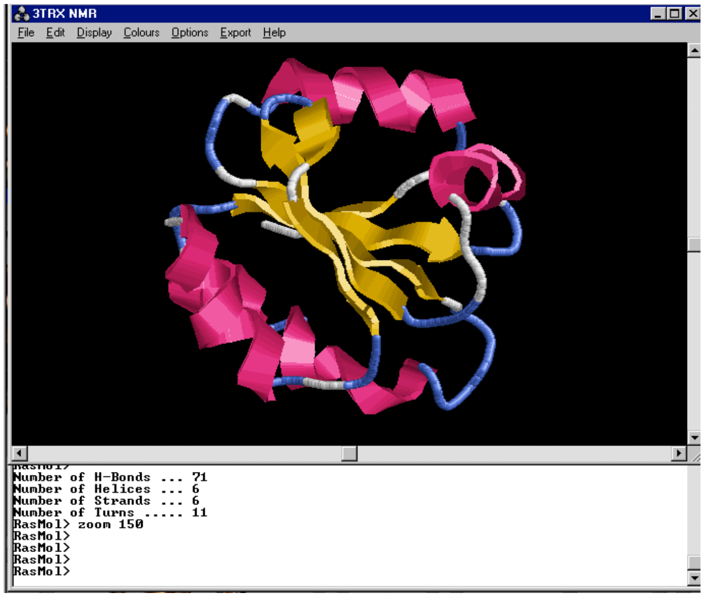
Applet QuickPDB v1.1 (C) 1996-1998 SDSC, by Ilya Shindyalov & Phil Bourne

Unsigned Java Applet Window

Quick PDB

- http://www.sdsc.edu/pb/Software.html
- Very simple viewing program with limited manipulation and very limited rendering capacity -- Very fast
- Java Applet (Source code available)
- Compatible with most browsers and computer platforms

Rasmol

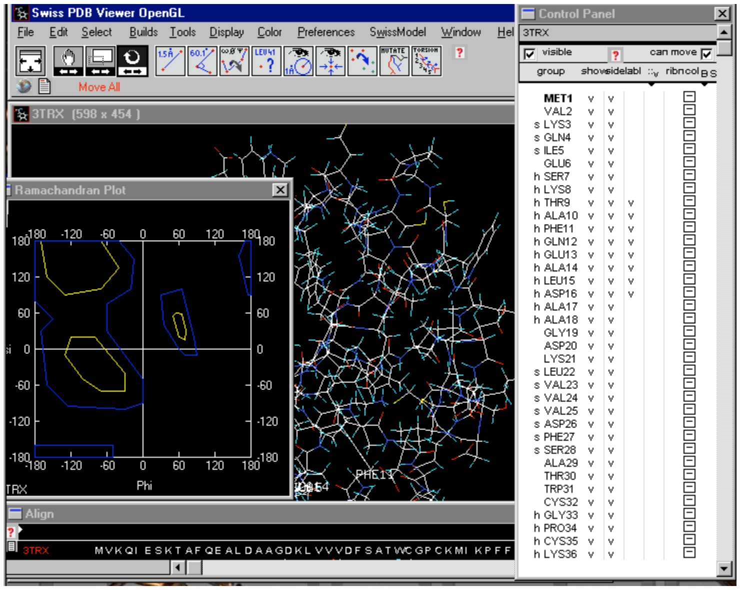


```
RasMol>
Number of H-Bonds ... 71
Number of Helices ... 6
Number of Strands ... 6
Number of Turns ..... 11
RasMol> zoom 150
RasMol>
RasMol>
RasMol>
RasMol>
```

Rasmol

- http://www.umass.edu/microbio/rasmol/
- Very simple viewing program with limited manipulation capacity, easy to use!
- “Grand-daddy” of all visual freeware
- Runs as installed “stand-alone” program
- Source code available
- Runs on Mac, Windows, Linux, SGI and most other UNIX platforms

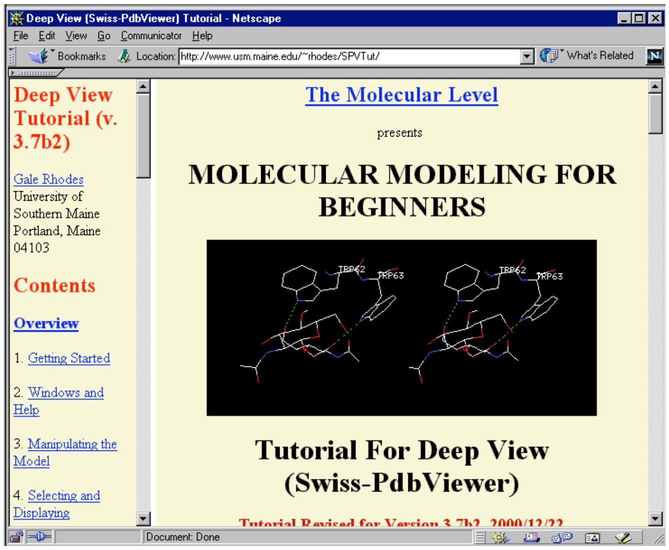
Swiss PDB Viewer



Swiss PDB Viewer

- http://www.expasy.ch/spdbv/
- Among most sophisticated molecular rendering, manipulation and modelling packages (commercial or freeware)
- Supports threading, hom. Modelling, energy minimization, seq/struc interface
- Stand-alone version only
- Compatible on Mac, Win, Linux, SGI

Swiss PDB Tutorial


Deep View (Swiss-PdbViewer) Tutorial - Netscape
File Edit View Go Communicator Help
Bookmarks
Location: http://www.usm.maine.edu/~rhodes/SPVTut/
What's Related
TRP62
TRP63
Document: Done

Deep View Tutorial (v. 3.7b2)

Gale Rhodes
University of
Southern Maine
Portland, Maine
04103

Contents

Overview

1. Getting Started

2. Windows and Help

3. Manipulating the Model

4. Selecting and Displaying

The Molecular Level

presents

MOLECULAR MODELING FOR BEGINNERS

Tutorial For Deep View (Swiss-PdbViewer)

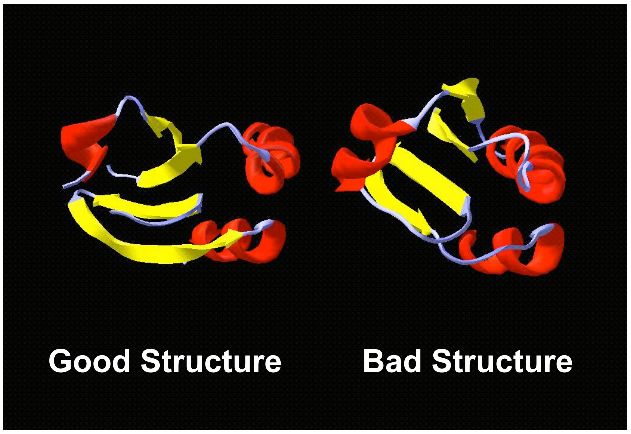
Tutorial Revised for Version 3.7b2, 2000/12/22

http://www.usm.maine.edu/~rhodes/SPVTut/

Summary

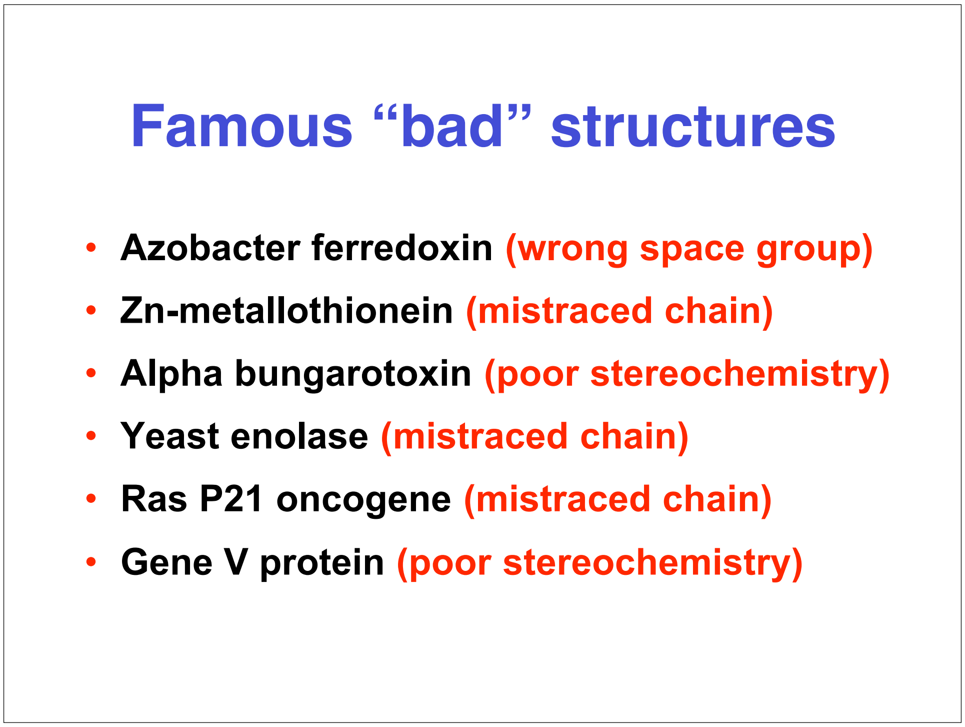
	Mac	Win	Unix	Rendr	SeqView	Super	E Min	Modeling
Rasmol	+	+	+	++	–	–	–	–
Chime	+	+	–	+	–	–	–	–
Prot. Expl.	+	+	–	++	+	+	–	–
Quick PDB	+	+	+	+	+	–	–	–
Biomer	+	+	+	++	–	+	+	+
SwP Viewer	+	+	+	+++	+	+	+	+
MolMol	–	+	+	+++	–	+	–	+

Assessing 3D Structures



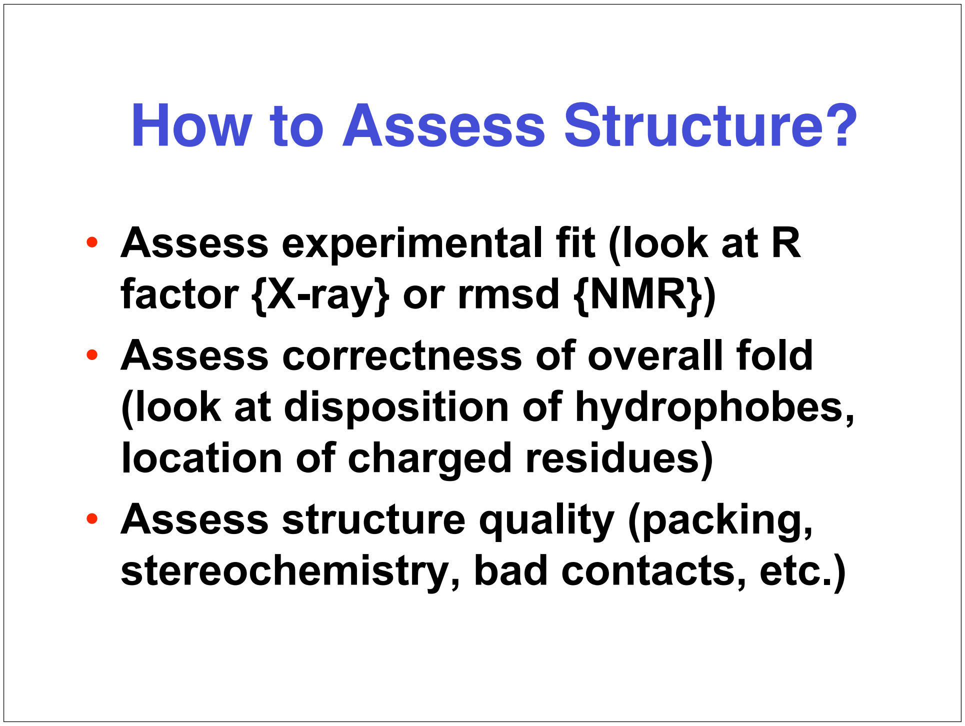
Why Assess Structure?

- 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)



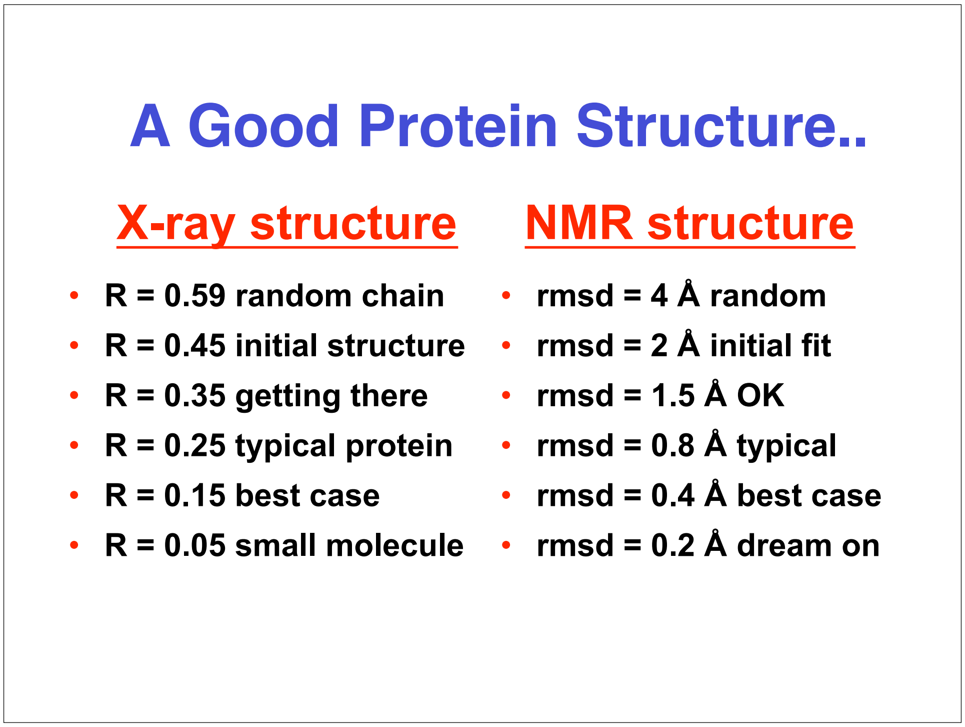
Famous “bad” structures

- 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)



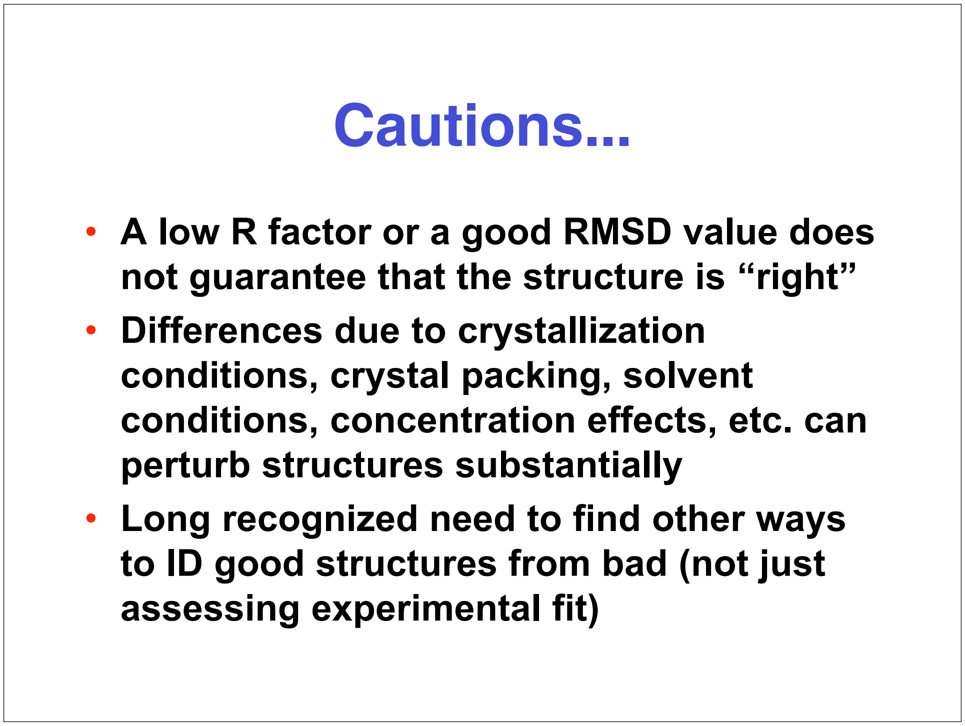
How to Assess Structure?

- 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.)



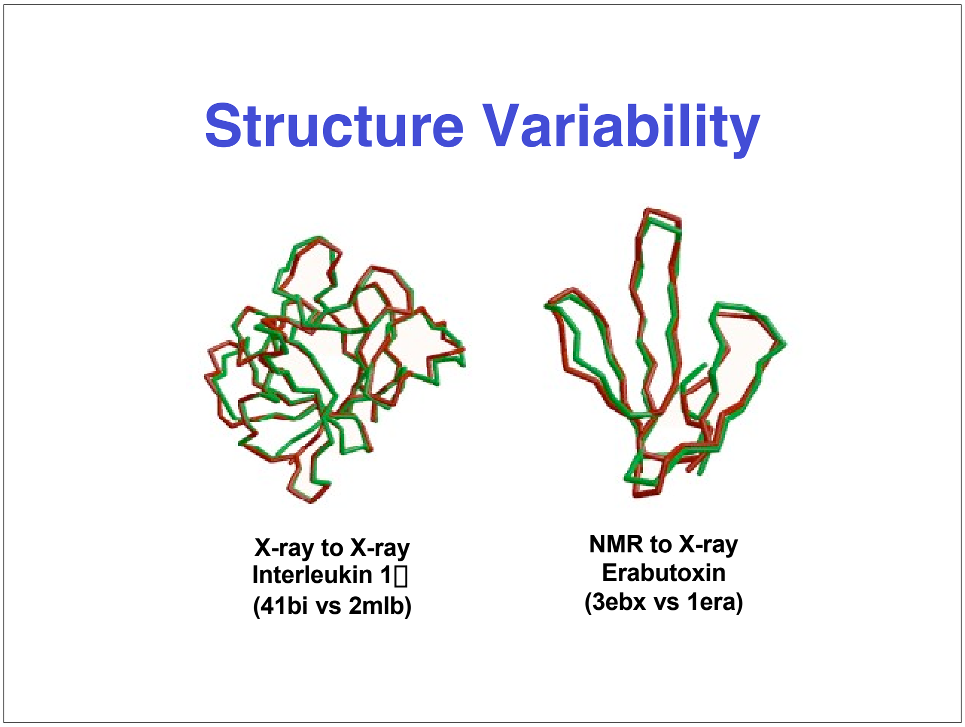
A Good Protein Structure..

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...

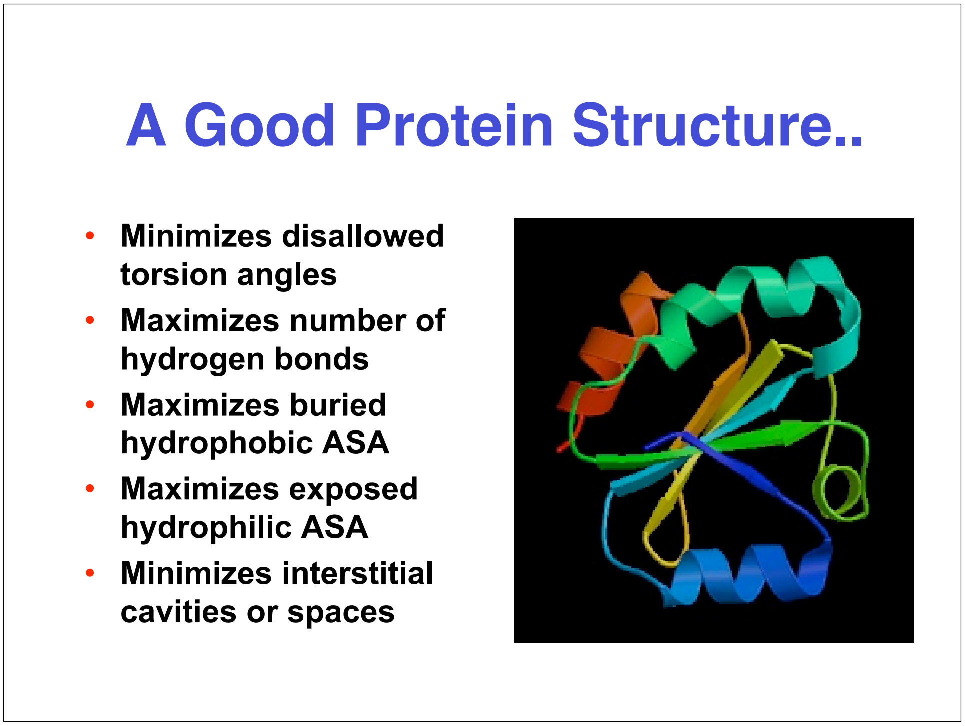
- A low R factor or a good RMSD value does not guarantee that the structure is “right”
- Differences due to crystallization conditions, crystal packing, solvent conditions, concentration effects, etc. can perturb structures substantially
- Long recognized need to find other ways to ID good structures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



Structure Variability

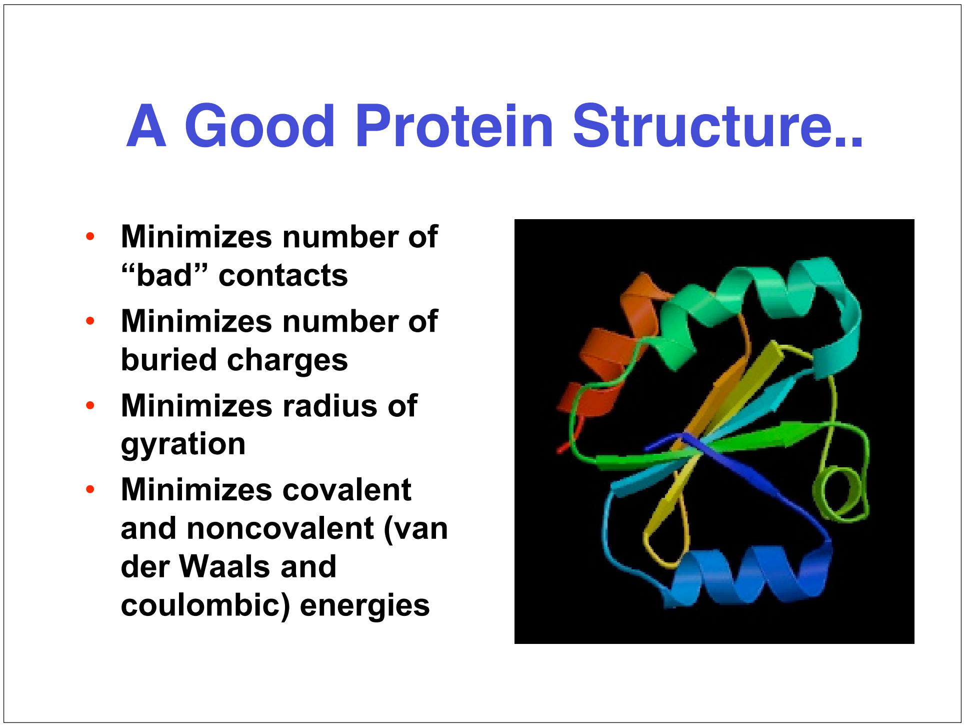
X-ray to X-ray
Interleukin 1□
(41bi vs 2mlb)

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



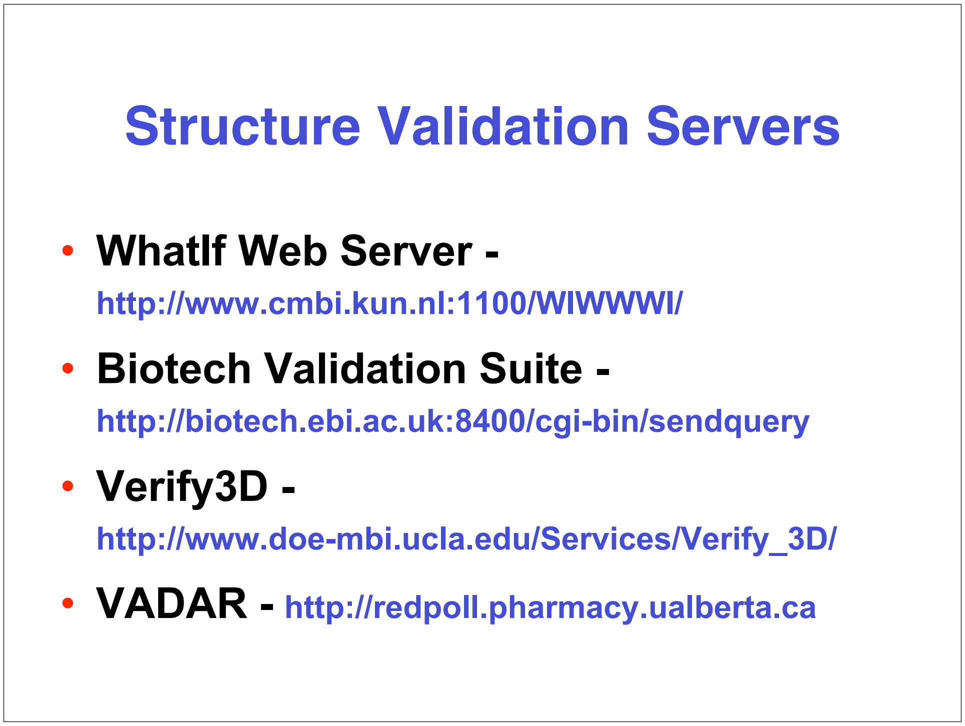
A Good Protein Structure..

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



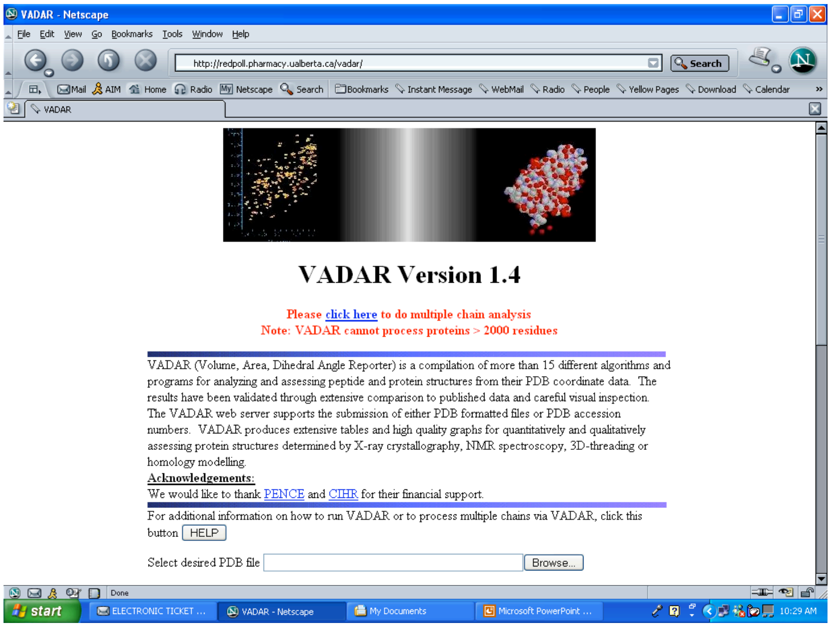
A Good Protein Structure..

- 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://www.cmbi.kun.nl:1100/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


VADAR - Netscape
File Edit View Go Bookmarks Tools Window Help
http://redpoll.pharmacy.ualberta.ca/vadar/
Search
Mail AIM Home Radio Netscape Search Bookmarks Instant Message WebMail Radio People Yellow Pages Download Calendar
VADAR
Done
start
ELECTRONIC TICKET ...
VADAR - Netscape
My Documents
Microsoft PowerPoint ...
10:29 AM

VADAR Version 1.4

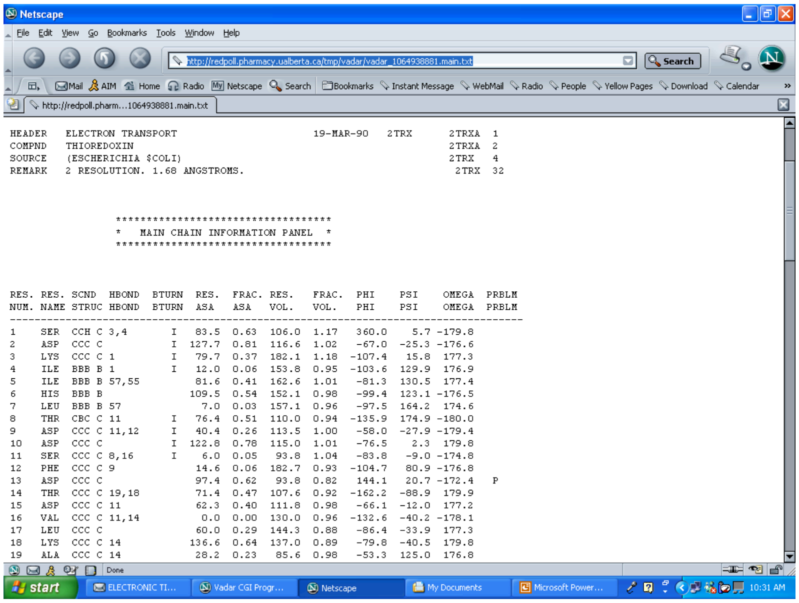
Please click here to do multiple chain analysis
Note: VADAR cannot process proteins > 2000 residues

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.

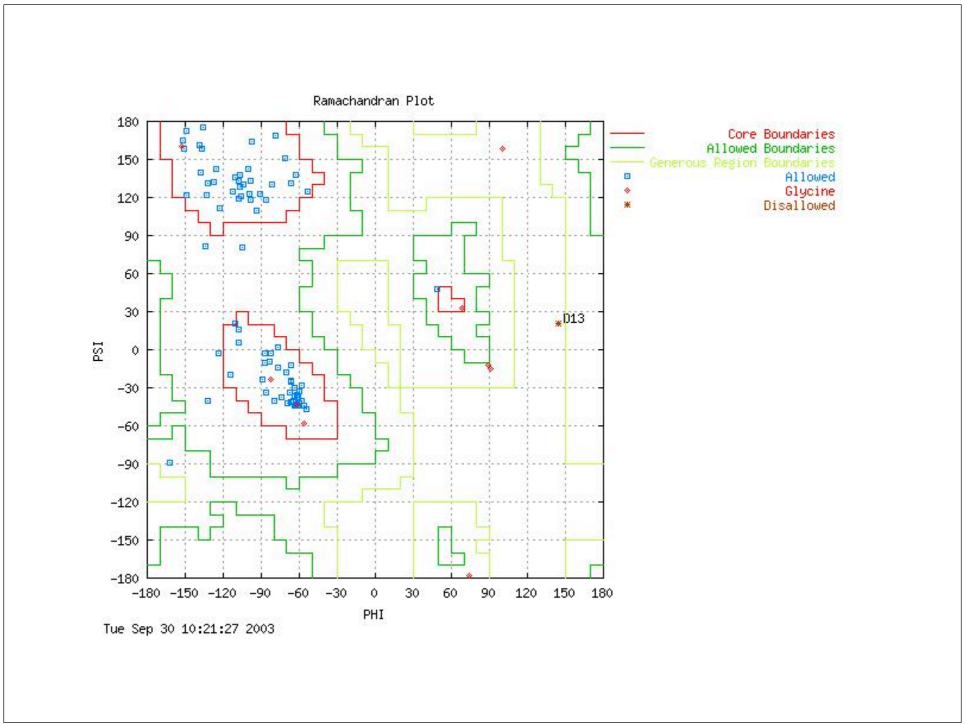
Acknowledgements:
We would like to thank PENCE and CIHR for their financial support.

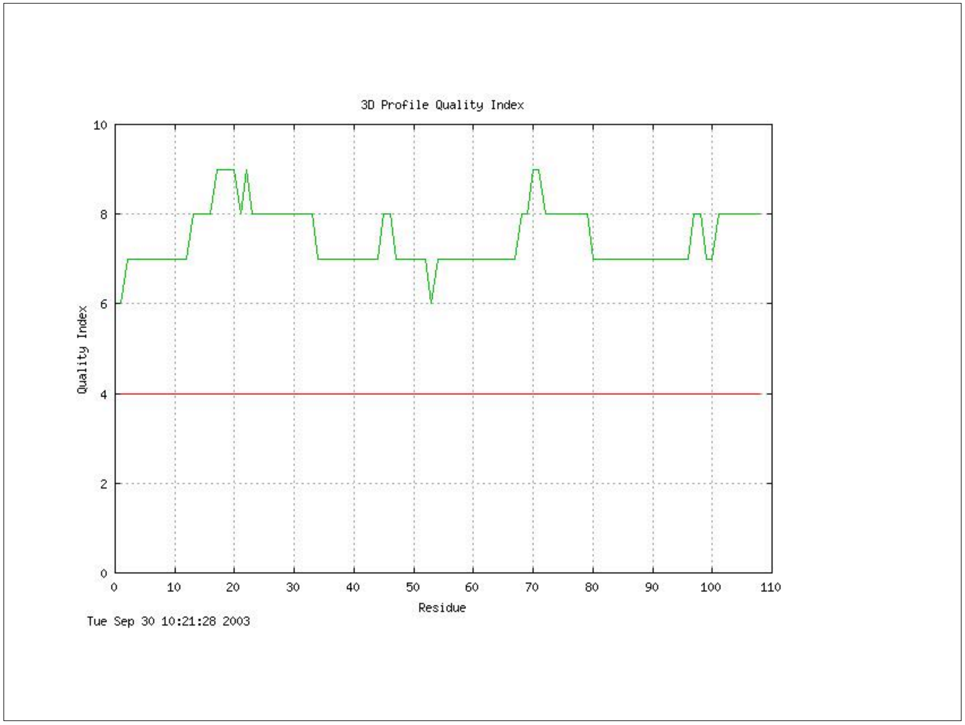
For additional information on how to run VADAR or to process multiple chains via VADAR, click this button HELP

Select desired PDB file Browse...


Netscape
File Edit View Go Bookmarks Tools Window Help
http://redpoll.pharmacy.ualberta.ca/tmp/vadar/vadar_1064938881.main.txt
Search
Mail AIM Home Radio Netscape Search Bookmarks Instant Message WebMail Radio People Yellow Pages Download Calendar
http://redpoll.pharm...1064938881.main.txt
Done
start
ELECTRONIC TI...
Vadar CGI Progr...
Netscape
My Documents
Microsoft Power...
10:31 AM

```
HEADER   ELECTRON TRANSPORT                       19-MAR-90   2TRX       2TRXA  1
COMPND   THIOREDOXIN                                                     2TRXA  2
SOURCE   (ESCHERICHIA $COLI)                                             2TRX   4
REMARK   2 RESOLUTION. 1.68 ANGSTROMS.                                    2TRX  32



                 **********************************
                 *   MAIN CHAIN INFORMATION PANEL  *
                 **********************************



RES. RES. SCND  HBOND  BTURN  RES.  FRAC. RES.   FRAC.  PHI    PSI    OMEGA  PRBLM
NUM. NAME STRUC HBOND  BTURN  ASA   ASA   VOL.   VOL.   PHI    PSI    OMEGA  PRBLM
-----------------------------------------------------------------------------------
1    SER  CCH C 3,4        I   83.5  0.63  106.0  1.17   360.0    5.7 -179.8
2    ASP  CCC C            I  127.7  0.81  116.6  1.02   -67.0  -25.3 -176.6
3    LYS  CCC C 1          I   79.7  0.37  182.1  1.18  -107.4   15.8  177.3
4    ILE  BBB B 1          I   12.0  0.06  153.8  0.95  -103.6  129.9  176.9
5    ILE  BBB B 57,55          81.6  0.41  162.6  1.01   -81.3  130.5  177.4
6    HIS  BBB B               109.5  0.54  152.1  0.98   -99.4  123.1 -176.5
7    LEU  BBB B 57              7.0  0.03  157.1  0.96   -97.5  164.2  174.6
8    THR  CBC C 11         I   76.4  0.51  110.0  0.94  -135.9  174.9 -180.0
9    ASP  CCC C 11,12      I   40.4  0.26  113.5  1.00   -58.0  -27.9 -179.4
10   ASP  CCC C            I  122.8  0.78  115.0  1.01   -76.5    2.3  179.8
11   SER  CCC C 8,16       I    6.0  0.05   93.8  1.04   -83.8   -9.0 -174.8
12   PHE  CCC C 9              14.6  0.06  182.7  0.93  -104.7   80.9 -176.8
13   ASP  CCC C                97.4  0.62   93.8  0.82   144.1   20.7 -172.4   P
14   THR  CCC C 19,18          71.4  0.47  107.6  0.92  -162.2  -88.9  179.9
15   ASP  CCC C 11             62.3  0.40  111.8  0.98   -66.1  -12.0  177.2
16   VAL  CCC C 11,14           0.0  0.00  130.0  0.96  -132.6  -40.2 -178.1
17   LEU  CCC C                60.0  0.29  144.3  0.88   -86.4  -33.9  177.3
18   LYS  CCC C 14            136.6  0.64  137.0  0.89   -79.8  -40.5  179.8
19   ALA  CCC C 14             28.2  0.23   85.6  0.98   -53.3  125.0  176.8
```

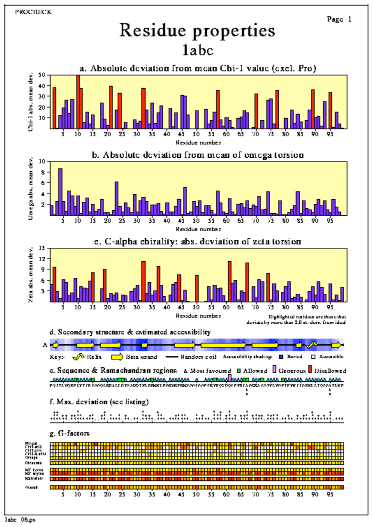

Ramachandran Plot
PSI
PHI
180
150
120
90
60
30
0
−30
−60
−90
−120
−150
−180
−180 −150 −120 −90 −60 −30 0 30 60 90 120 150 180
D13
Core Boundaries
Allowed Boundaries
Generous Region Boundaries
Allowed
Glycine
Disallowed
Tue Sep 30 10:21:27 2003

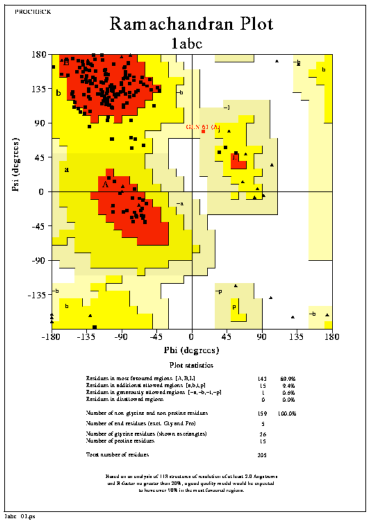

3D Profile Quality Index
Quality Index
Residue
10
8
6
4
2
0
0 10 20 30 40 50 60 70 80 90 100 110
Tue Sep 30 10:21:28 2003

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

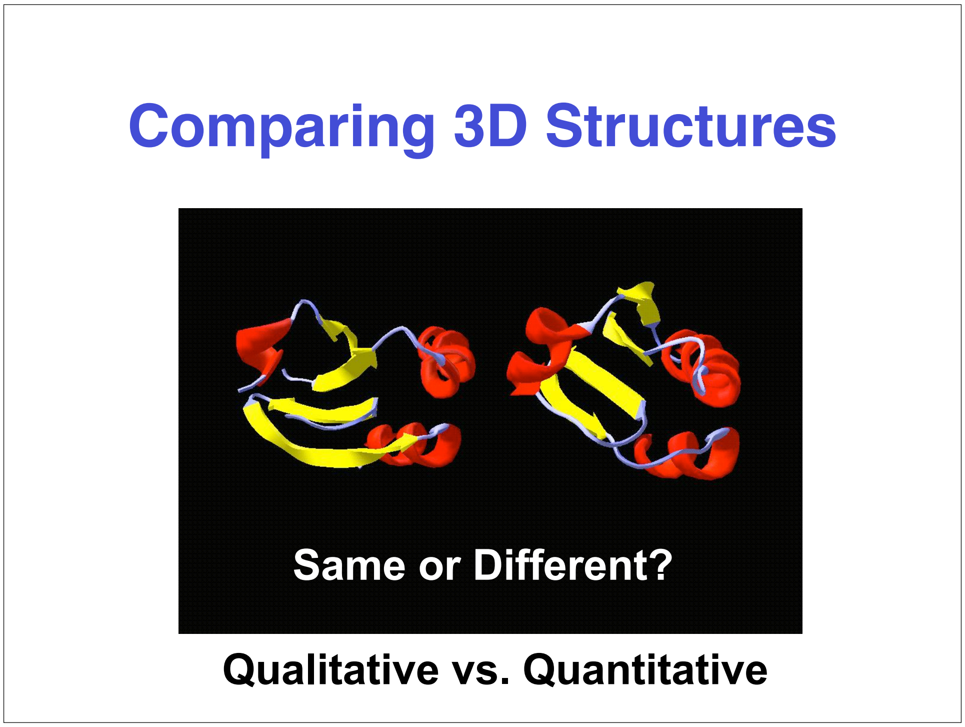




Plot statistics

Residues in most favoured regions [A,B,L]	143	89.9%
Residues in additional allowed regions [a,b,l,p]	15	9.4%
Residues in generously allowed regions [~a,~b,~l,~p]	1	0.6%
Residues in disallowed regions	0	0.0%
Number of non-glycine and non-proline residues	159	100.0%
Number of end residues (excl. Gly and Pro)	5	
Number of glycine residues (shown as triangles)	26	
Number of proline residues	15	
Total number of residues	205	

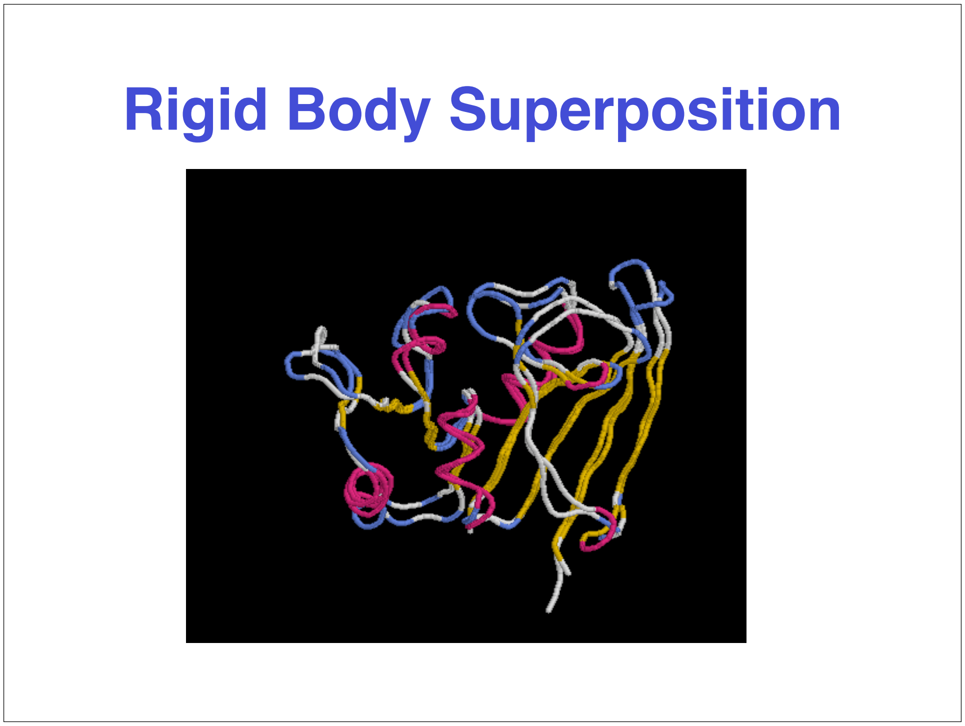
Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.



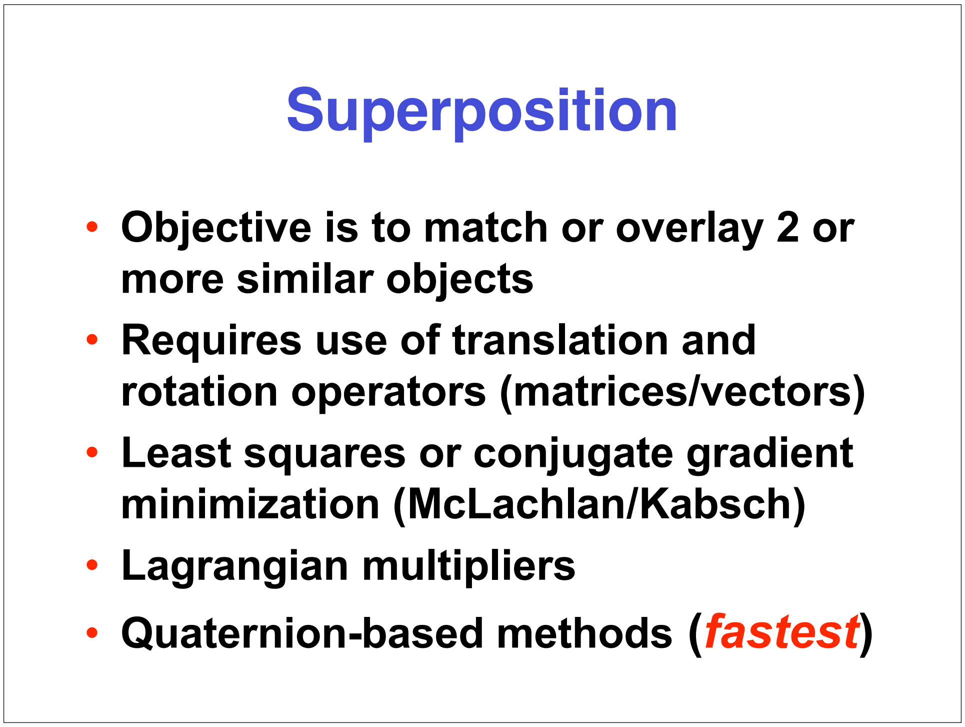
Comparing 3D Structures

Same or Different?

Qualitative vs. Quantitative

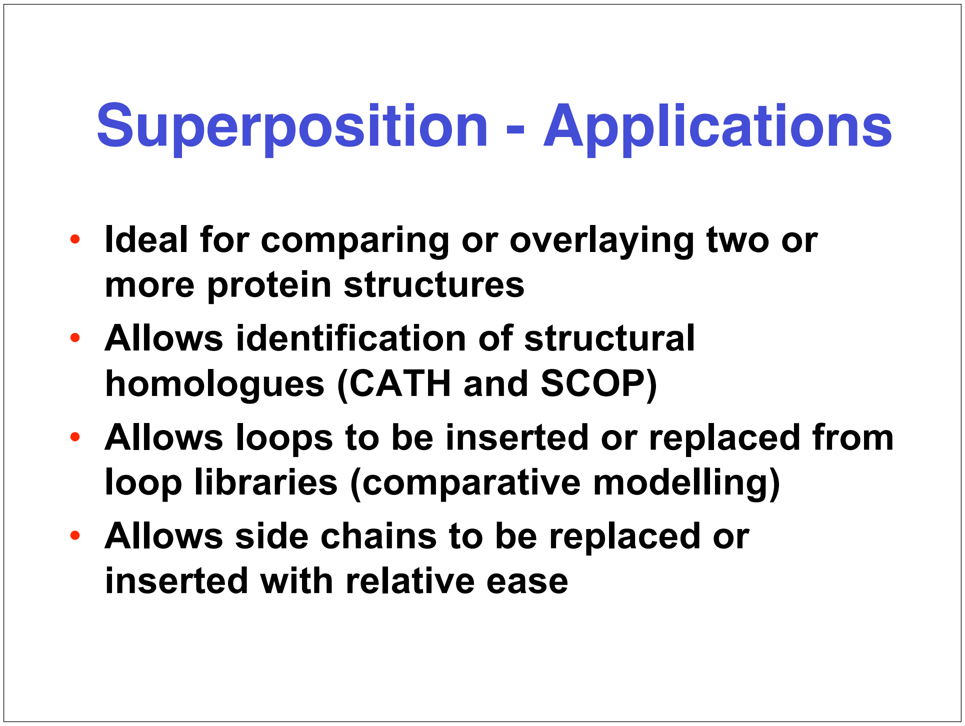


Rigid Body Superposition



Superposition

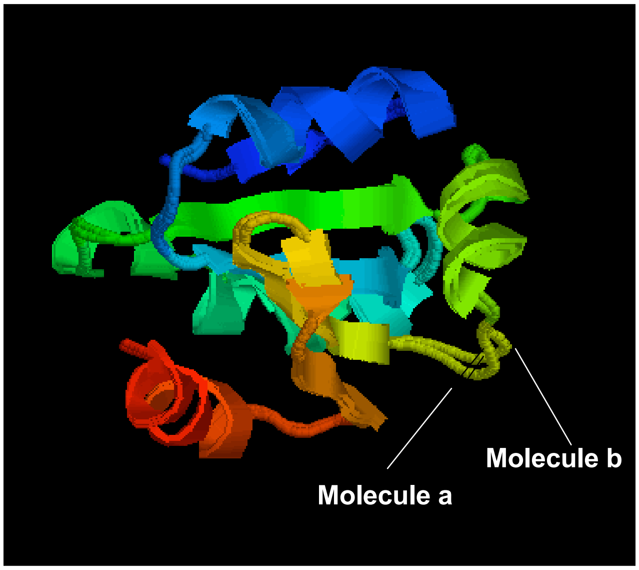
- Objective is to match or overlay 2 or more similar objects
- Requires use of translation and rotation operators (matrices/vectors)
- Least squares or conjugate gradient minimization (McLachlan/Kabsch)
- Lagrangian multipliers
- Quaternion-based methods (*fastest*)



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

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

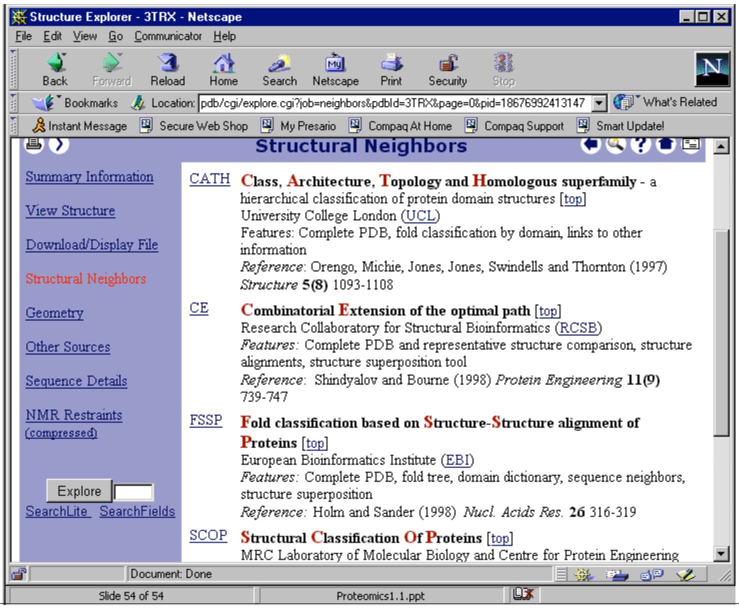
RMSD

- 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


Structure Explorer - 3TRX - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location: pdb/cgi/explore.cgi?job=neighbors&pdbId=3TRX&page=0&pid=18676992413147 What's Related
Instant Message Secure Web Shop My Presario Compaq At Home Compaq Support Smart Update!
Summary Information
View Structure
Download/Display File
Structural Neighbors
Geometry
Other Sources
Sequence Details
NMR Restraints (compressed)
Explore
SearchLite SearchFields
Document: Done
Slide 54 of 54
Proteomics1.1.ppt

Structural Neighbors

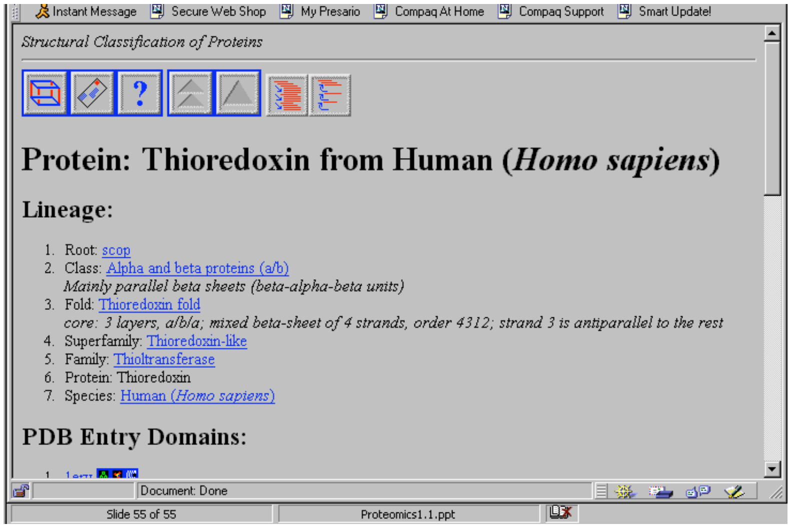
CATH **Class, Architecture, Topology and Homologous superfamily** - a hierarchical classification of protein domain structures [top]
University College London (UCL)
Features: Complete PDB, fold classification by domain, links to other information
Reference: Orengo, Michie, Jones, Jones, Swindells and Thornton (1997) *Structure* **5(8)** 1093-1108

CE **Combinatorial Extension of the optimal path** [top]
Research Collaboratory for Structural Bioinformatics (RCSB)
Features: Complete PDB and representative structure comparison, structure alignments, structure superposition tool
Reference: Shindyalov and Bourne (1998) *Protein Engineering* **11(9)** 739-747

FSSP **Fold classification based on Structure-Structure alignment of Proteins** [top]
European Bioinformatics Institute (EBI)
Features: Complete PDB, fold tree, domain dictionary, sequence neighbors, structure superposition
Reference: Holm and Sander (1998) *Nucl. Acids Res.* **26** 316-319

SCOP **Structural Classification Of Proteins** [top]
MRC Laboratory of Molecular Biology and Centre for Protein Engineering

SCOP Database


Instant Message Secure Web Shop My Presario Compaq At Home Compaq Support Smart Update!
Document: Done
Slide 55 of 55
Proteomics1.1.ppt

Structural Classification of Proteins

Protein: Thioredoxin from Human (*Homo sapiens*)

Lineage:

1. Root: scop
2. Class: Alpha and beta proteins (a/b)
   Mainly parallel beta sheets (beta-alpha-beta units)
3. Fold: Thioredoxin fold
   core: 3 layers, a/b/a; mixed beta-sheet of 4 strands, order 4312; strand 3 is antiparallel to the rest
4. Superfamily: Thioredoxin-like
5. Family: Thioltransferase
6. Protein: Thioredoxin
7. Species: Human *(Homo sapiens)*

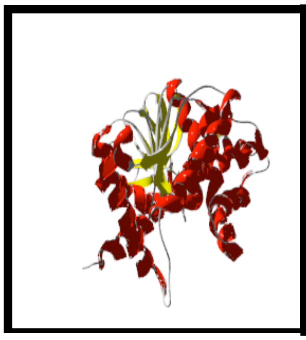
PDB Entry Domains:

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

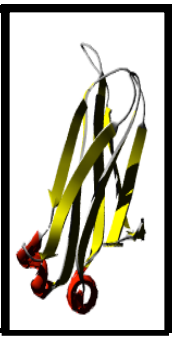
SCOP

- 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

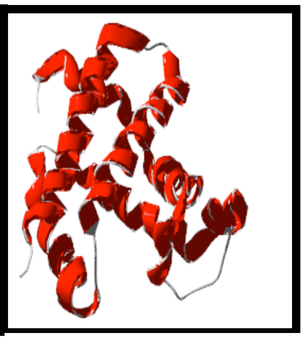
Different Folding Classes



Lactate
Dehydrogenase:
Mixed □ / □

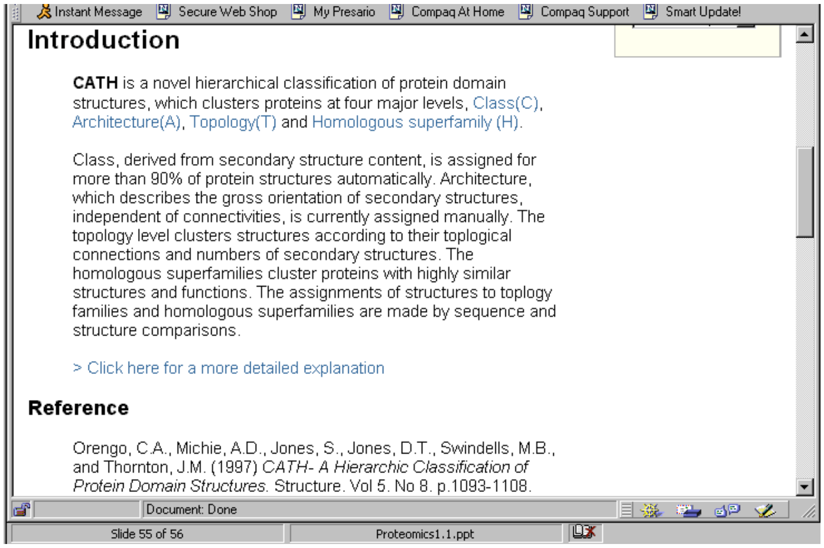


Immunoglobulin
Fold: □



Hemoglobin B
Chain: □

CATH Database



Introduction

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

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 according to their toplogical connections and numbers of secondary structures. The homologous superfamilies cluster proteins with highly similar structures and functions. The assignments of structures to toplogy families and homologous superfamilies are made by sequence and structure comparisons.

> Click here for a more detailed explanation

Reference

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. (1997) *CATH- A Hierarchic Classification of Protein Domain Structures*. Structure. Vol 5. No 8. p.1093-1108.

http://www.biochem.ucl.ac.uk/bsm/cath/

CATH

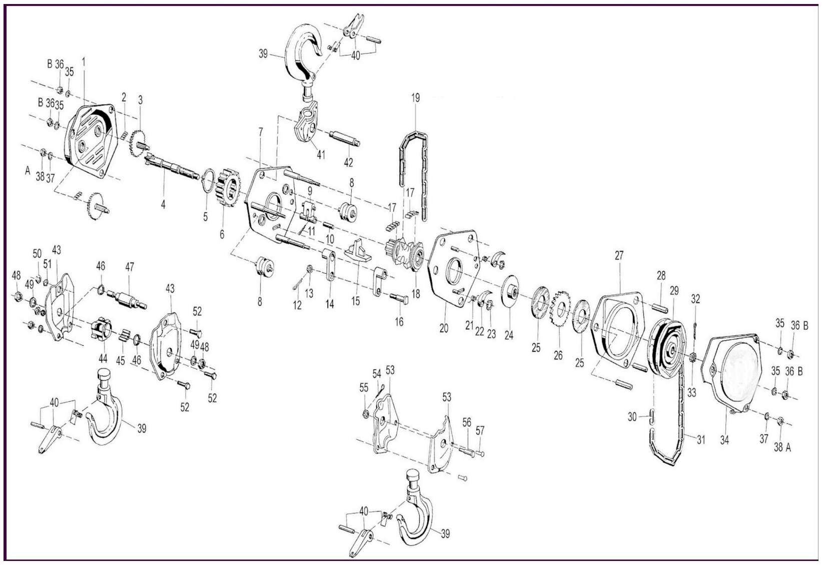
- **Class [C] derived from secondary structure content (automatic)**
- **Architecture (A) derived from orientation of 2° structures (manual)**
- **Topology (T) derived from topological connection and # 2° structures**
- **Homologous Superfamily (H) clusters of similar structures & functions**

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

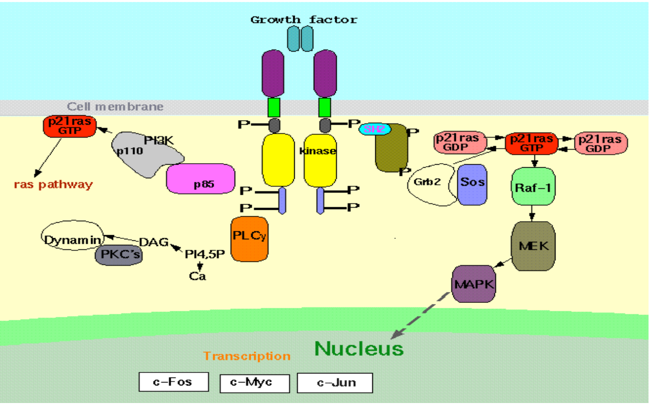
The Protein Parts List



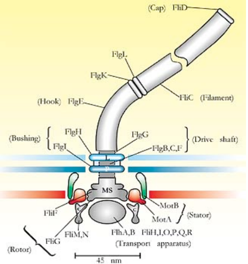
The Parts List

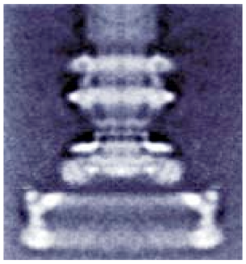
- Sequencing gives “serial number”
- Sequence alignment gives a name
- Microarrays give # of parts
- X-ray and NMR give a picture
- However, having a collection of parts and names doesn’t tell you how to put something together or how things connect -- *this is biology*

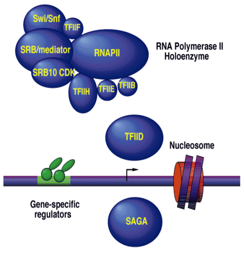
Remember: *Proteins Interact*

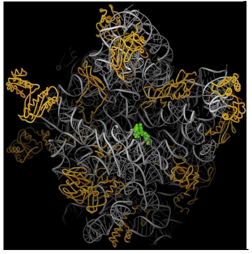


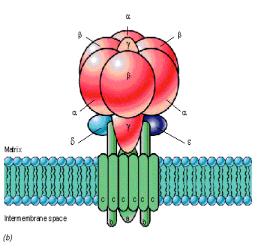
Proteins *Assemble*

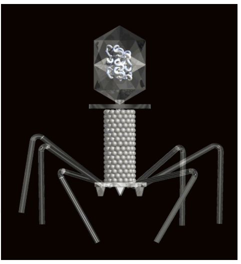


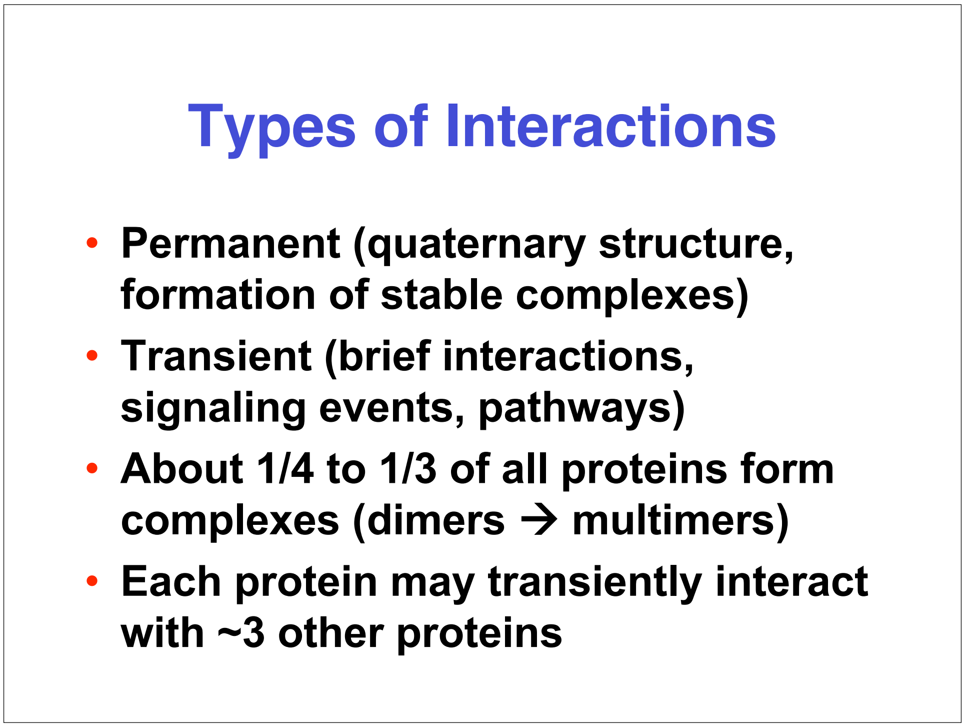










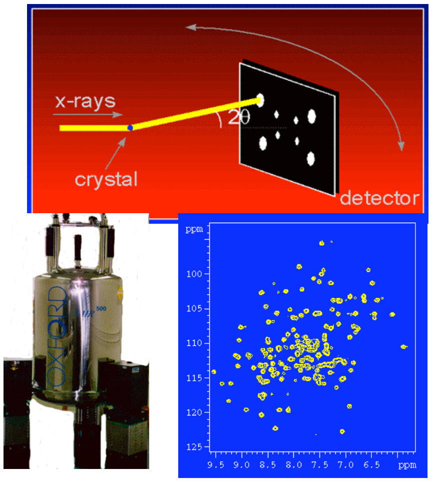


Types of Interactions

- 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

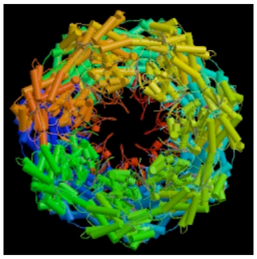
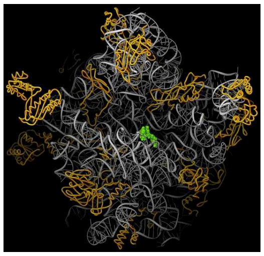
Protein Interaction Tools and Techniques - Experimental Methods

3D Structure Determination

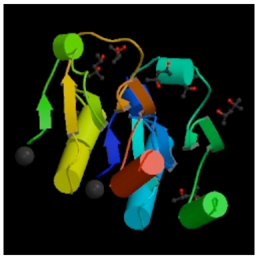
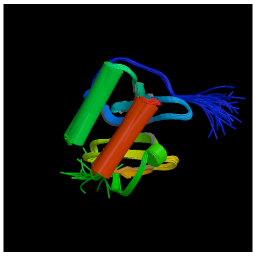


- X-ray crystallography
  - grow crystal
  - collect diffract. data
  - calculate e- density
  - trace chain
- NMR spectroscopy
  - label protein
  - collect NMR spectra
  - assign spectra & NOEs
  - calculate structure using distance geom.

Quaternary Structure

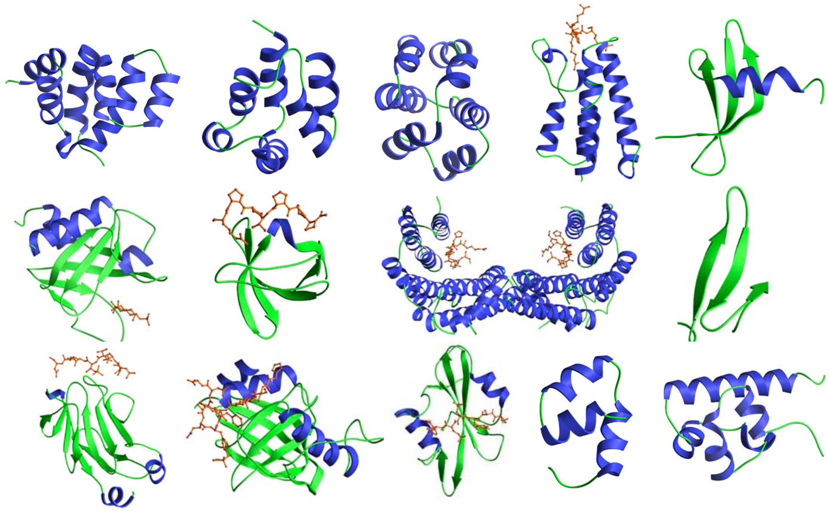



Some interactions are real

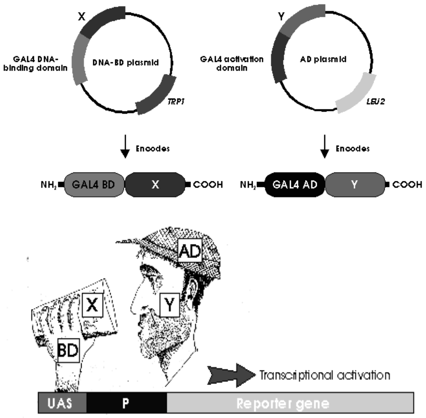
Others are not

Protein Interaction Domains



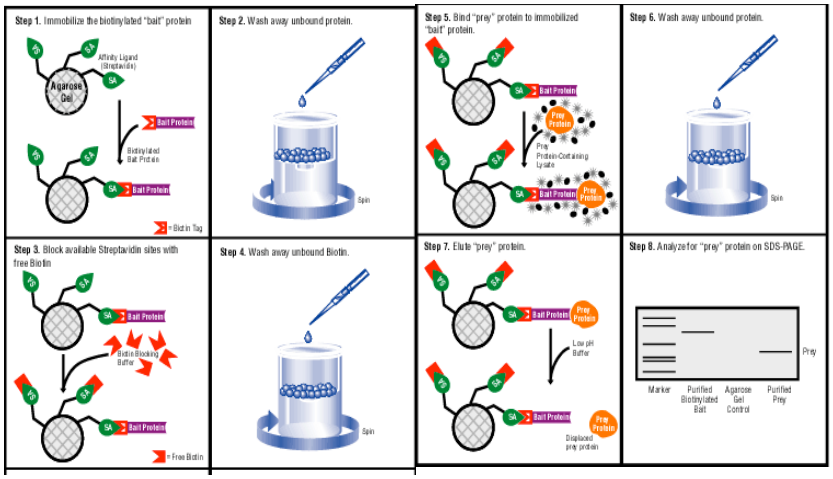
http://www.mshri.on.ca/pawson/domains.html

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



Step 1. Immobilize the biotinylated "bait" protein

Step 2. Wash away unbound protein.

Step 5. Bind "prey" protein to immobilized "bait" protein.

Step 6. Wash away unbound protein.

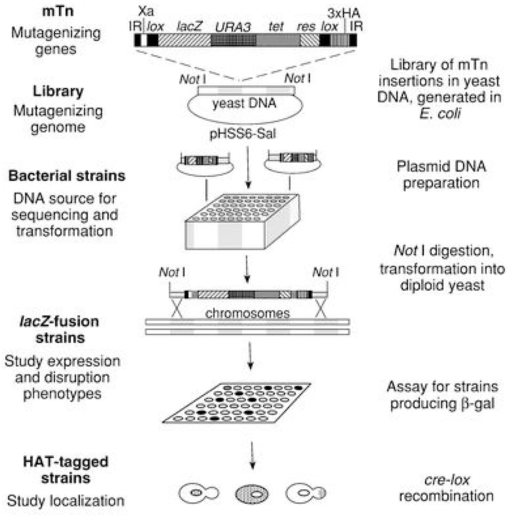
Step 3. Block available Streptavidin sites with free Biotin

Step 4. Wash away unbound Biotin.

Step 7. Elute "prey" protein.

Step 8. Analyze for "prey" protein on SDS-PAGE.

Transposon Tagging



mTn
Mutagenizing genes

Library
Mutagenizing genome

Bacterial strains
DNA source for sequencing and transformation

***lacZ*-fusion strains**
Study expression and disruption phenotypes

HAT-tagged strains
Study localization

Library of mTn insertions in yeast DNA, generated in *E. coli*

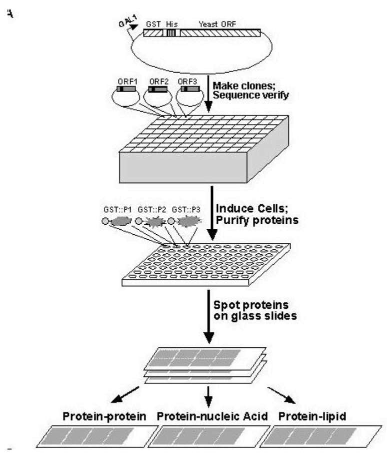
Plasmid DNA preparation

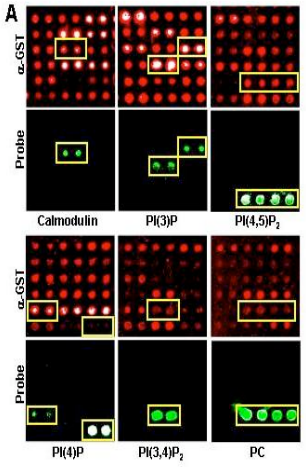
Not I digestion, transformation into diploid yeast

Assay for strains producing β-gal

cre-lox recombination

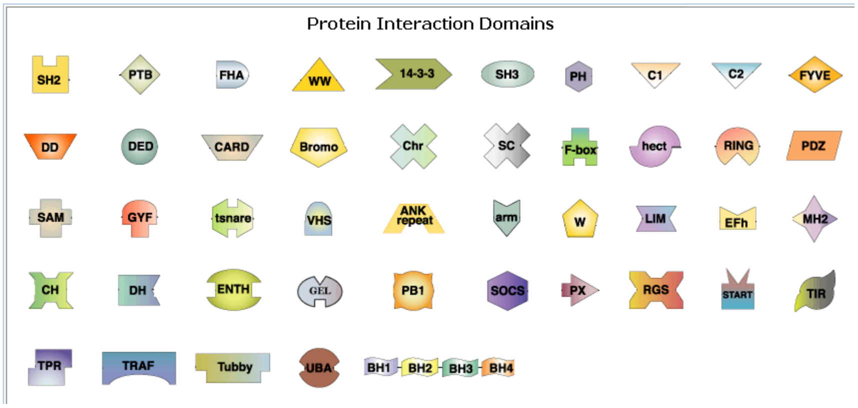
Protein Arrays





Protein Interaction Tools and Techniques - Computational Methods

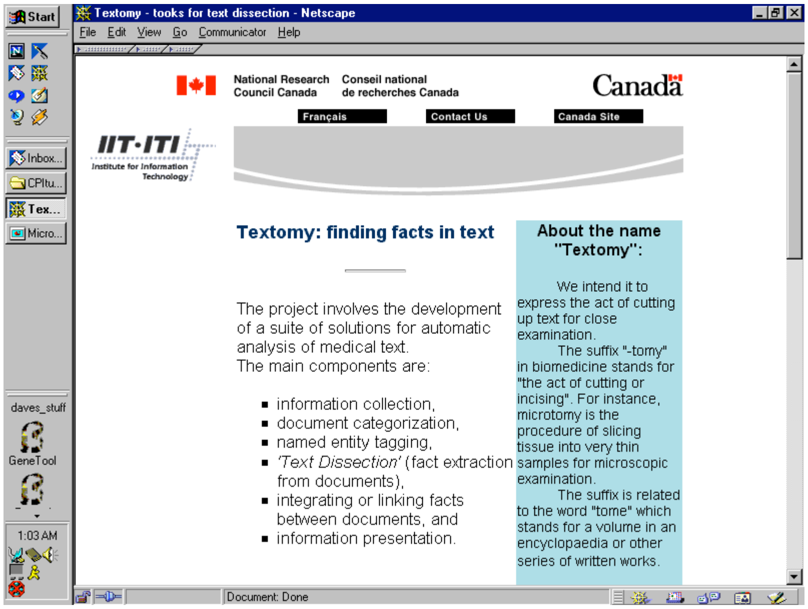
Sequence Searching Against Known Domains



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

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
- Sometimes called “Textomy”


Start
Textomy - tooks for text dissection - Netscape
File Edit View Go Communicator Help
Inbox...
CPltu...
Tex...
Micro...
daves_stuff
GeneTool
1:03 AM
National Research Council Canada
Conseil national de recherches Canada
Canadä
Français
Contact Us
Canada Site
IIT-ITI
Institute for Information Technology
Document: Done

Textomy: finding facts in text

The project involves the development of a suite of solutions for automatic analysis of medical text.
The main components are:

- information collection,
- document categorization,
- named entity tagging,
- *'Text Dissection'* (fact extraction from documents),
- integrating or linking facts between documents, and
- information presentation.

About the name "Textomy":

We intend it to express the act of cutting up text for close examination.

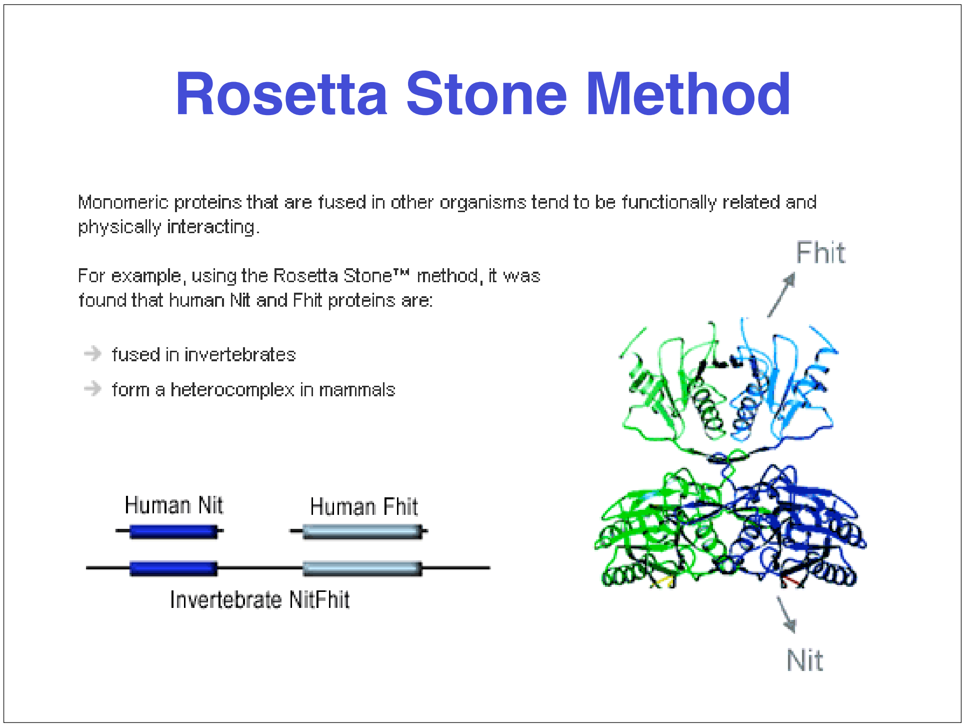
The suffix "-tomy" in biomedicine stands for "the act of cutting or incising". For instance, microtomy is the procedure of slicing tissue into very thin samples for microscopic examination.

The suffix is related to the word "tome" which stands for a volume in an encyclopaedia or other series of written works.

http://textomy.iit.nrc.ca/

Pre-BIND

- *Donaldson et al. BMC Bioinformatics 2003 4:11*
- **Used Support Vector Machine (SVM) to scan literature for protein interactions**
- **Precision, accuracy and recall of 92% for correctly classifying PI abstracts**
- **Estimated to capture 60% of all abstracted protein interactions for a given organism**

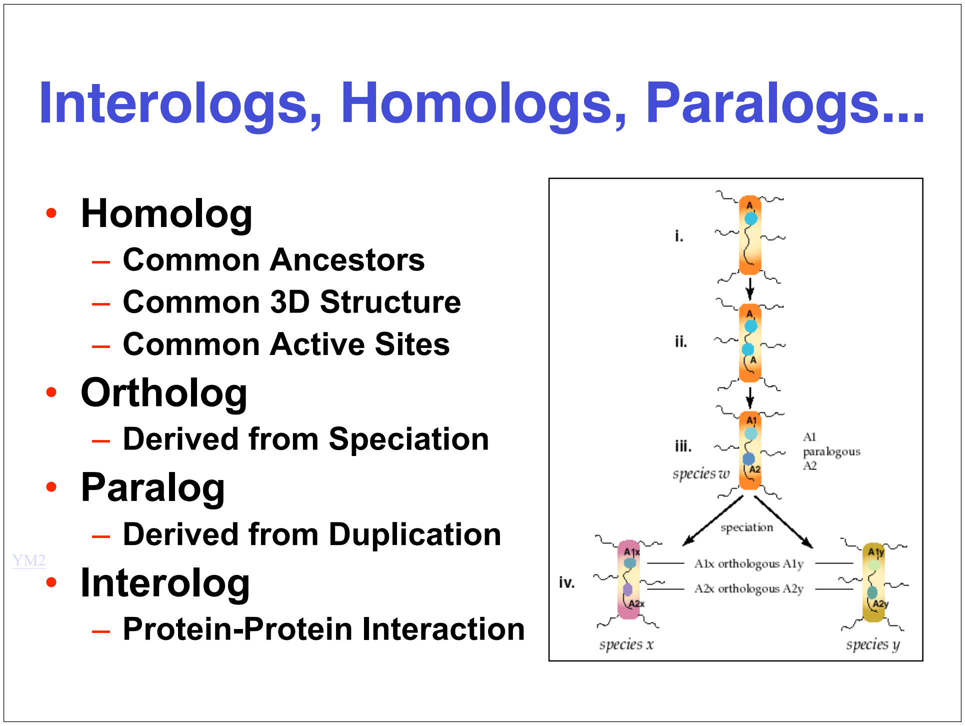


Rosetta Stone Method

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



Interologs, Homologs, Paralogs...

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

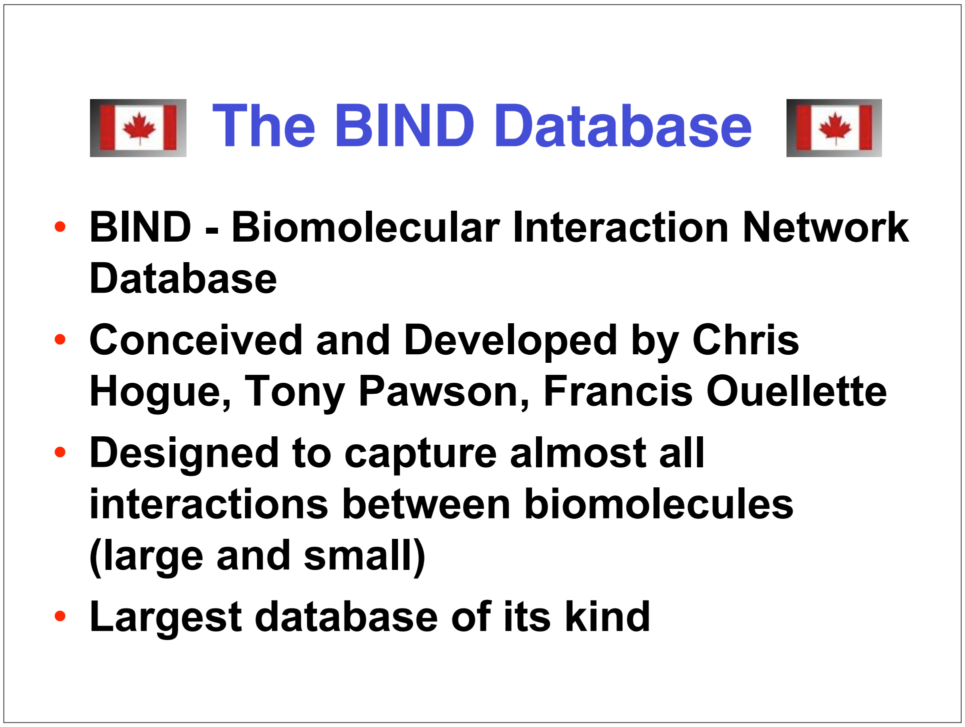
YM2

A Flood of Data

- High throughput techniques are leading to more and more data on protein interactions
- This is where bioinformatics can play a key role
- Some suggest that this is the “future” for bioinformatics

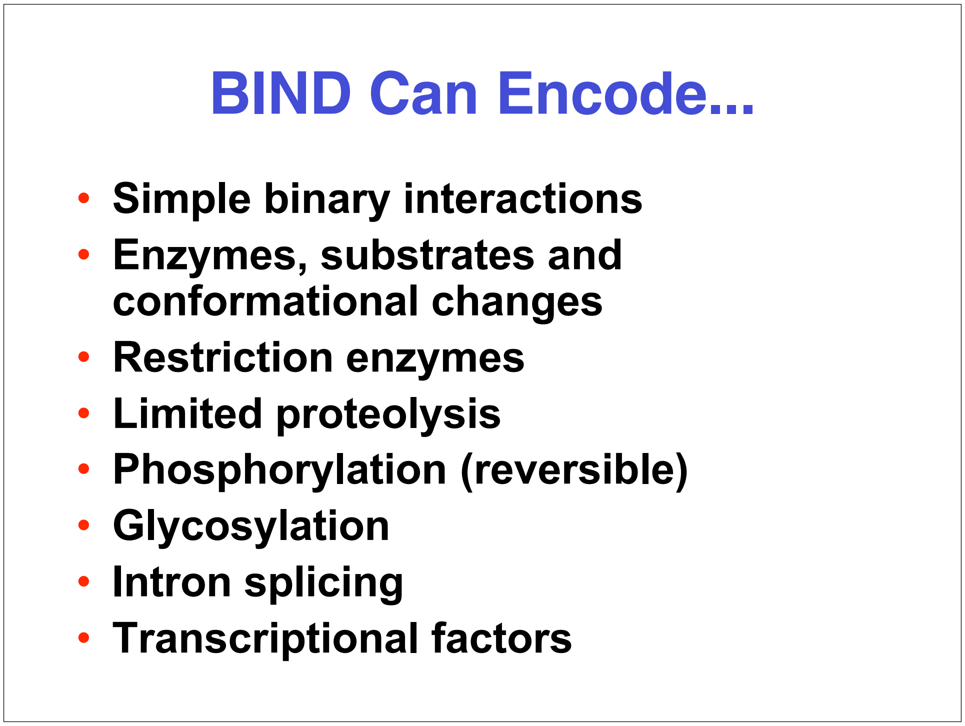
Interaction Databases

- BIND
  – http://www.bind.ca/
- DIP
  – http://dip.doe-mbi.ucla.edu/
- PIM
  – http://www.hybrigenics.fr/
- PathCalling
  – http://portal.curagen.com/extpc/com.curagen.portal.servlet.Yeast



The BIND Database

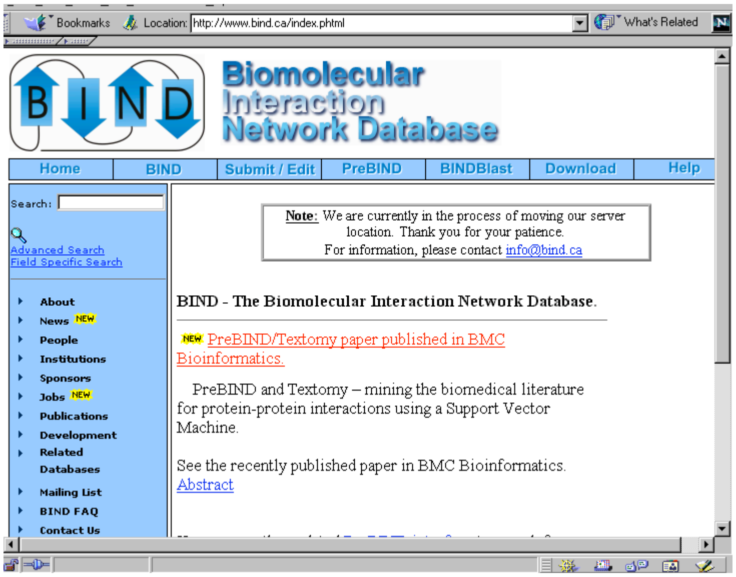
- BIND - Biomolecular Interaction Network Database
- Conceived and Developed by Chris Hogue, Tony Pawson, Francis Ouellette
- Designed to capture almost all interactions between biomolecules (large and small)
- Largest database of its kind



BIND Can Encode...

- Simple binary interactions
- Enzymes, substrates and conformational changes
- Restriction enzymes
- Limited proteolysis
- Phosphorylation (reversible)
- Glycosylation
- Intron splicing
- Transcriptional factors

BIND



Note: We are currently in the process of moving our server location. Thank you for your patience.
For information, please contact info@bind.ca

BIND - The Biomolecular Interaction Network Database.

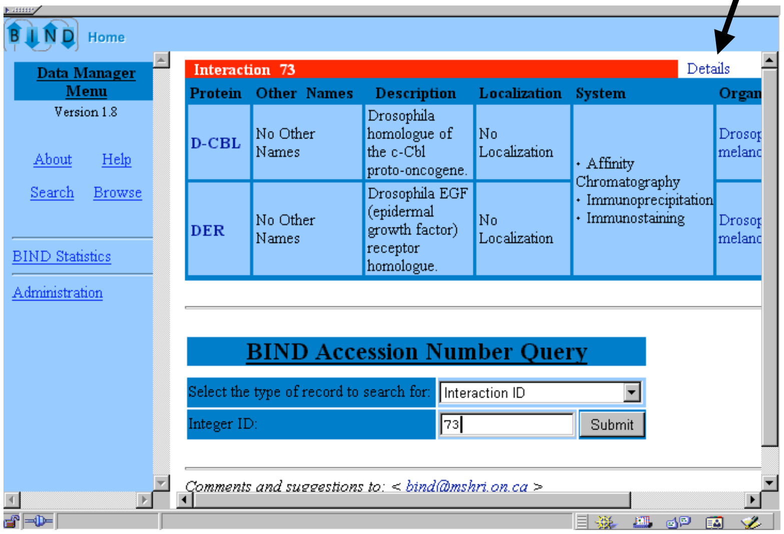
NEW PreBIND/Textomy paper published in BMC Bioinformatics.

PreBIND and Textomy – mining the biomedical literature for protein-protein interactions using a Support Vector Machine.

See the recently published paper in BMC Bioinformatics.
Abstract

BIND Query Result

click



Interaction 73

Protein	Other Names	Description	Localization	System	Organ
D-CBL	No Other Names	Drosophila homologue of the c-Cbl proto-oncogene.	No Localization	• Affinity Chromatography • Immunoprecipitation • Immunostaining	Drosop melanc
DER	No Other Names	Drosophila EGF (epidermal growth factor) receptor homologue.	No Localization		Drosop melanc

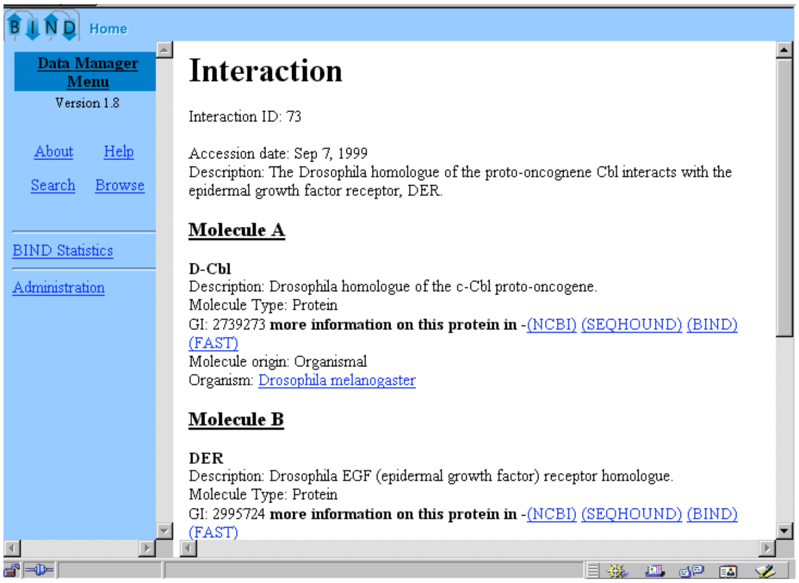
BIND Accession Number Query

Select the type of record to search for:	Interaction ID
Integer ID:	73

Submit

Comments and suggestions to: < bind@mshri.on.ca >

BIND Details



Interaction

Interaction ID: 73

Accession date: Sep 7, 1999
Description: The Drosophila homologue of the proto-oncognene Cbl interacts with the epidermal growth factor receptor, DER.

Molecule A

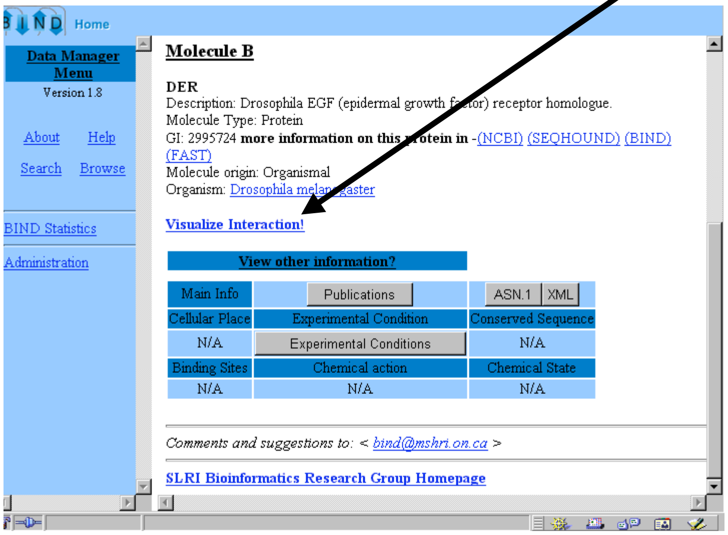
D-Cbl
Description: Drosophila homologue of the c-Cbl proto-oncogene.
Molecule Type: Protein
GI: 2739273 **more information on this protein in** -(NCBI) (SEQHOUND) (BIND) (FAST)
Molecule origin: Organismal
Organism: Drosophila melanogaster

Molecule B

DER
Description: Drosophila EGF (epidermal growth factor) receptor homologue.
Molecule Type: Protein
GI: 2995724 **more information on this protein in** -(NCBI) (SEQHOUND) (BIND) (FAST)

BIND Details

click



Molecule B

DER
Description: Drosophila EGF (epidermal growth factor) receptor homologue.
Molecule Type: Protein
GI: 2995724 **more information on this protein in** -(NCBI) (SEQHOUND) (BIND) (FAST)
Molecule origin: Organismal
Organism: Drosophila melanogaster

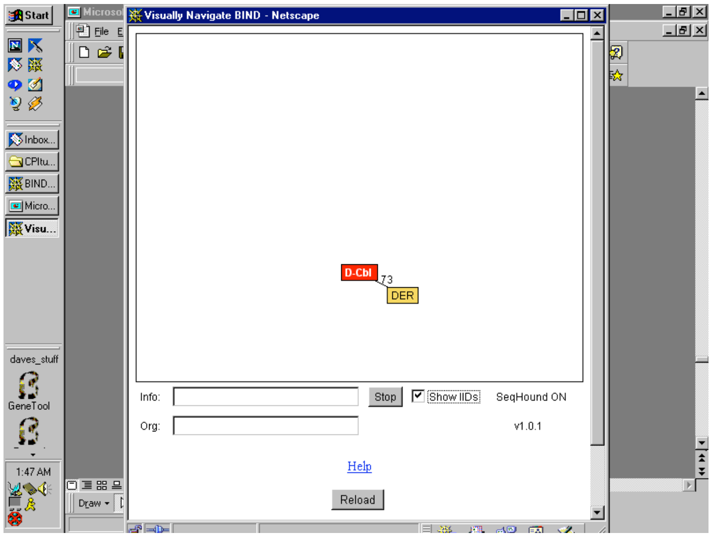
Visualize Interaction!

View other information?		
Main Info	Publications	ASN.1 XML
Cellular Place	Experimental Condition	Conserved Sequence
N/A	Experimental Conditions	N/A
Binding Sites	Chemical action	Chemical State
N/A	N/A	N/A

Comments and suggestions to: < *bind@mshri.on.ca* >

SLRI Bioinformatics Research Group Homepage

BIND Details



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