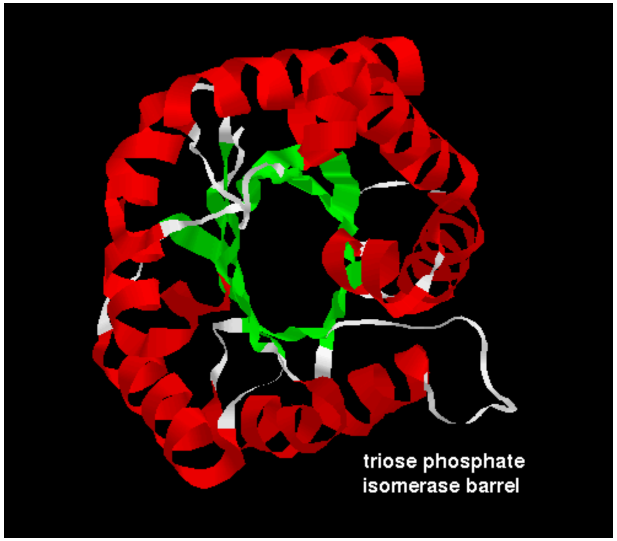
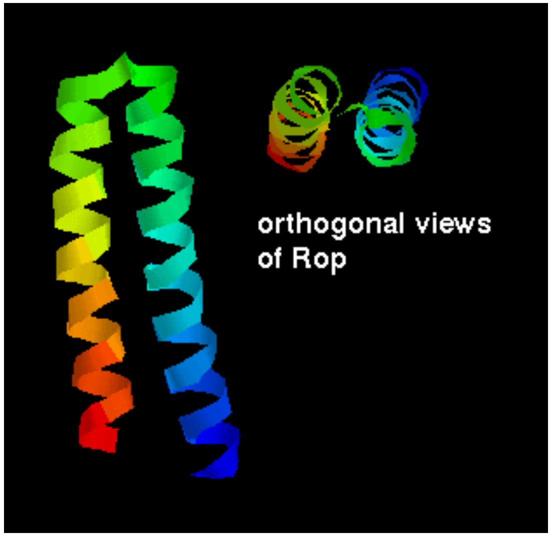
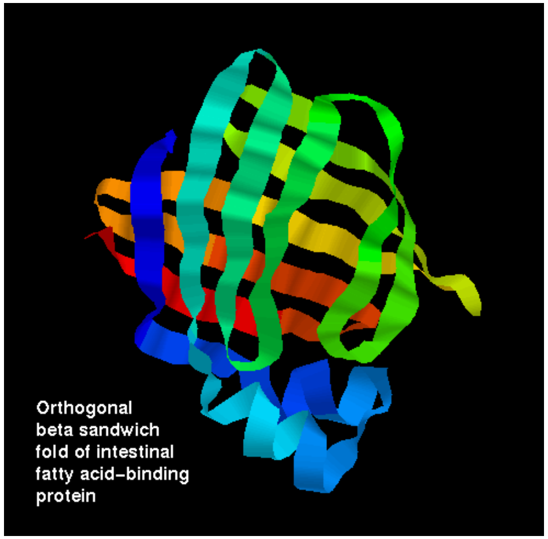
3D Structure
Prediction and Assessment







David Wishart

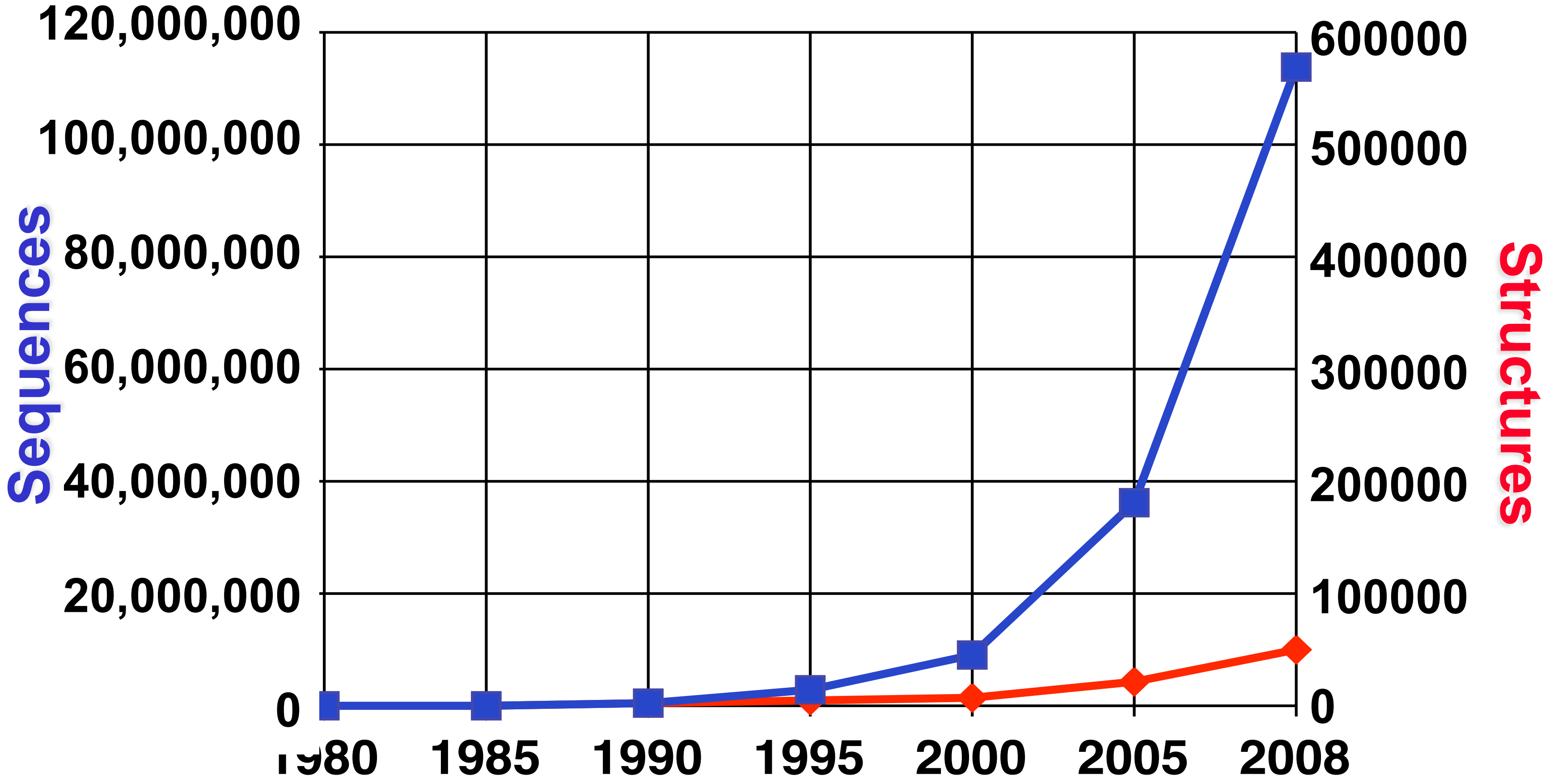
Athabasca 3-41

david.wishart@ualberta.ca

Outline & Objectives*

- Become familiar with the Protein Universe and the Protein Structure Initiative
- Learn principles of how to do homology (comparative) modelling of 3D protein structures
- Learn how to do homology modelling on the Web
- Learn how to assess 3D structures (modelled and experimental)

Structural Proteomics: The Motivation



Protein Structure Initiative*

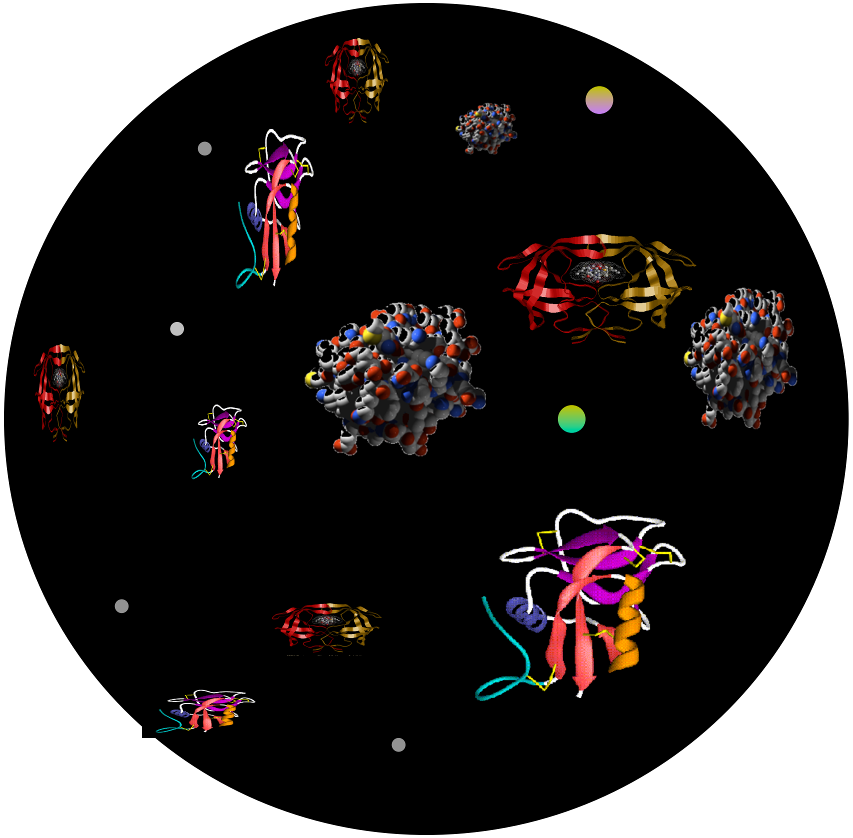
- Organize all known protein sequences into sequence families
- Select family representatives as targets
- Solve the 3D structures of these targets by X-ray or NMR
- Build models for the remaining proteins via comparative (homology) modeling

Protein Structure Initiative*

- Organize and recruit interested structural biologists and structure biology centres from around the world
- Coordinate target selection
- Develop new kinds of high throughput techniques
- Solve, solve, solve, solve….

The Protein Fold Universe

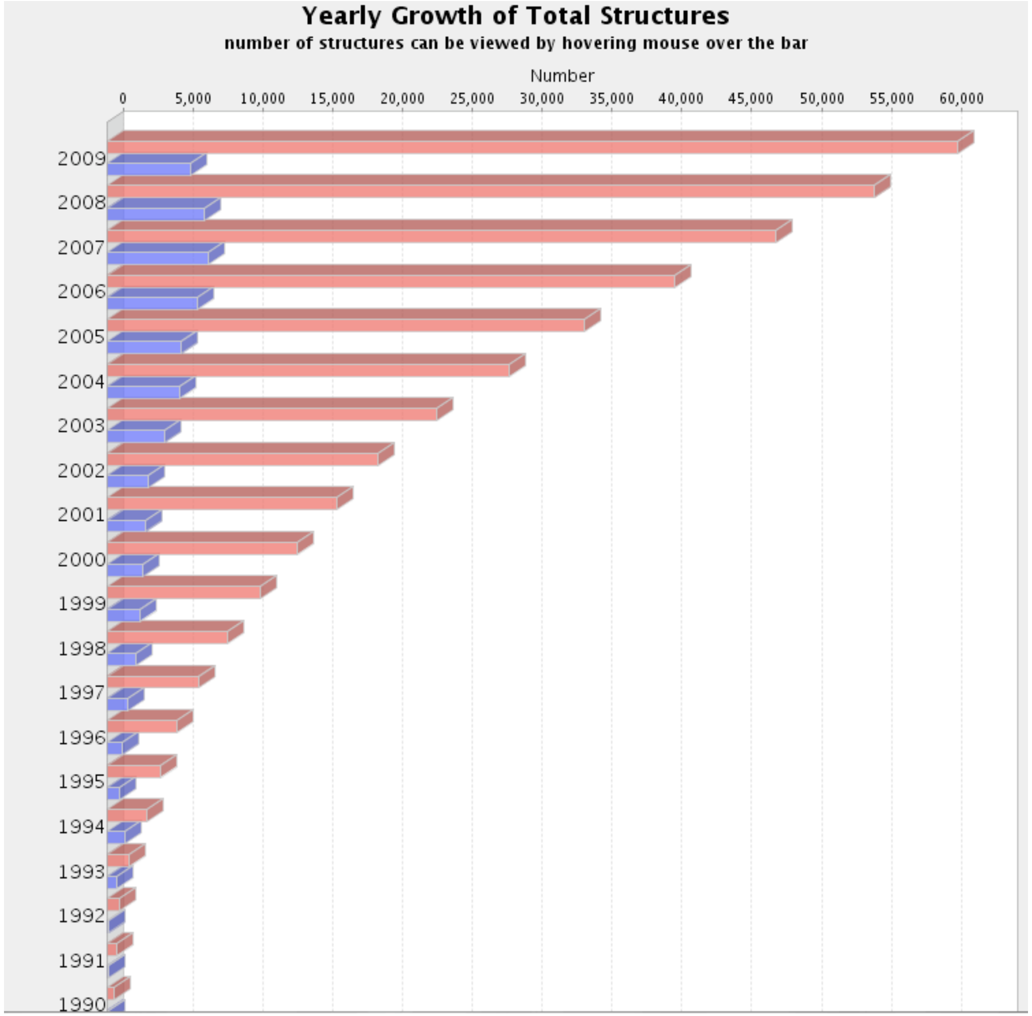
How
Big
Is
It???



500?
2000?
10000?
∞ ?

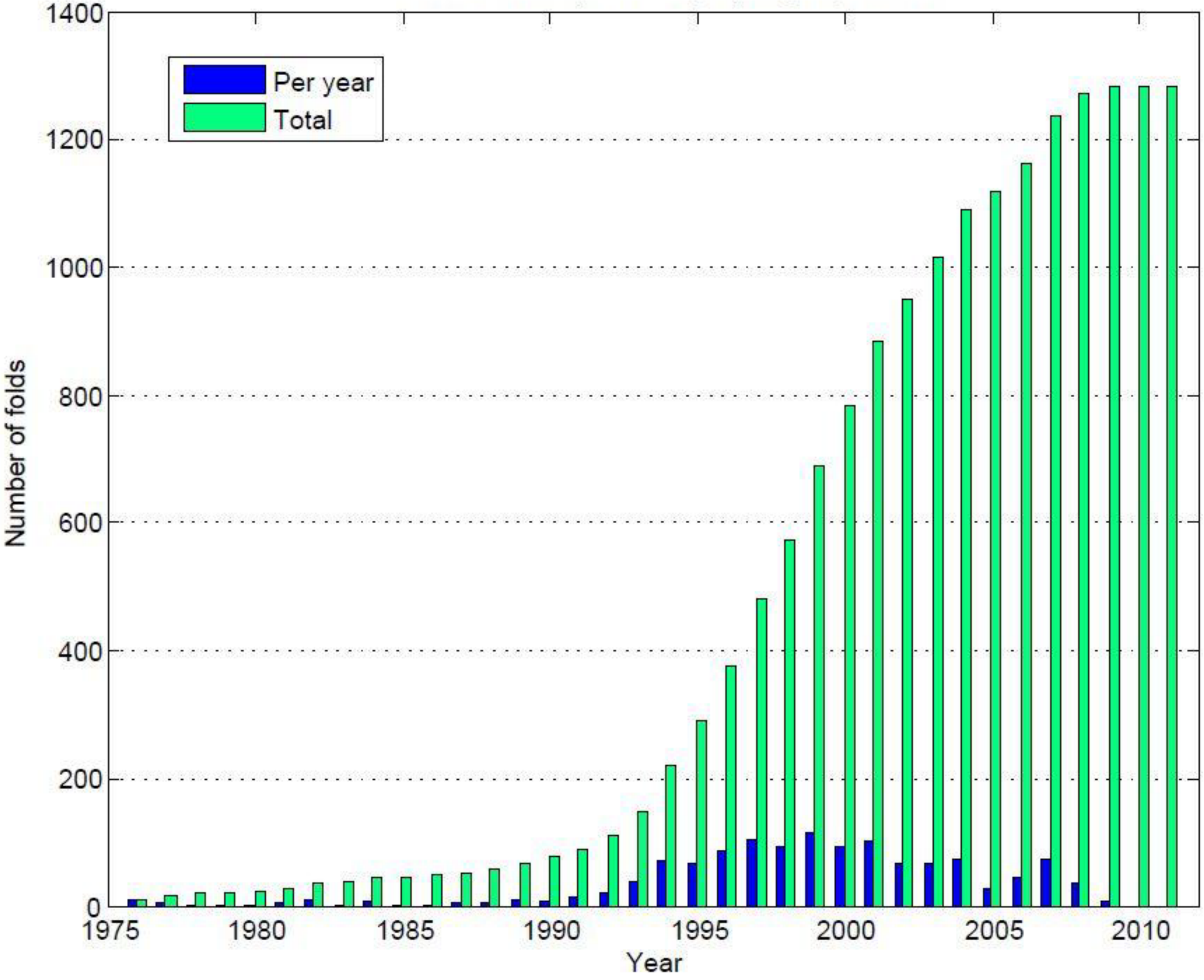
Human Genome Codes for ~21,000 Proteins

Structure Deposition Rate



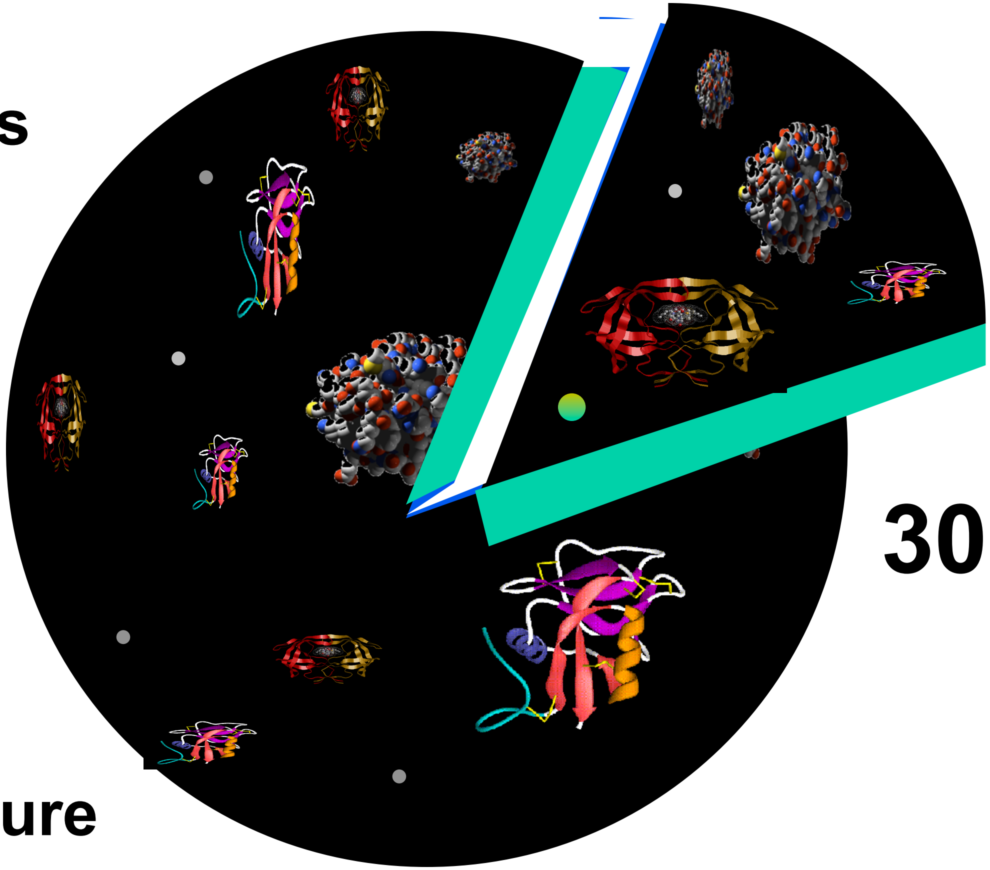
- Growth has been exponential for the past 10 years
- Approximately 8000 new structures being added each year

Number of New Folds in The PDB*



Protein Structure Initiative

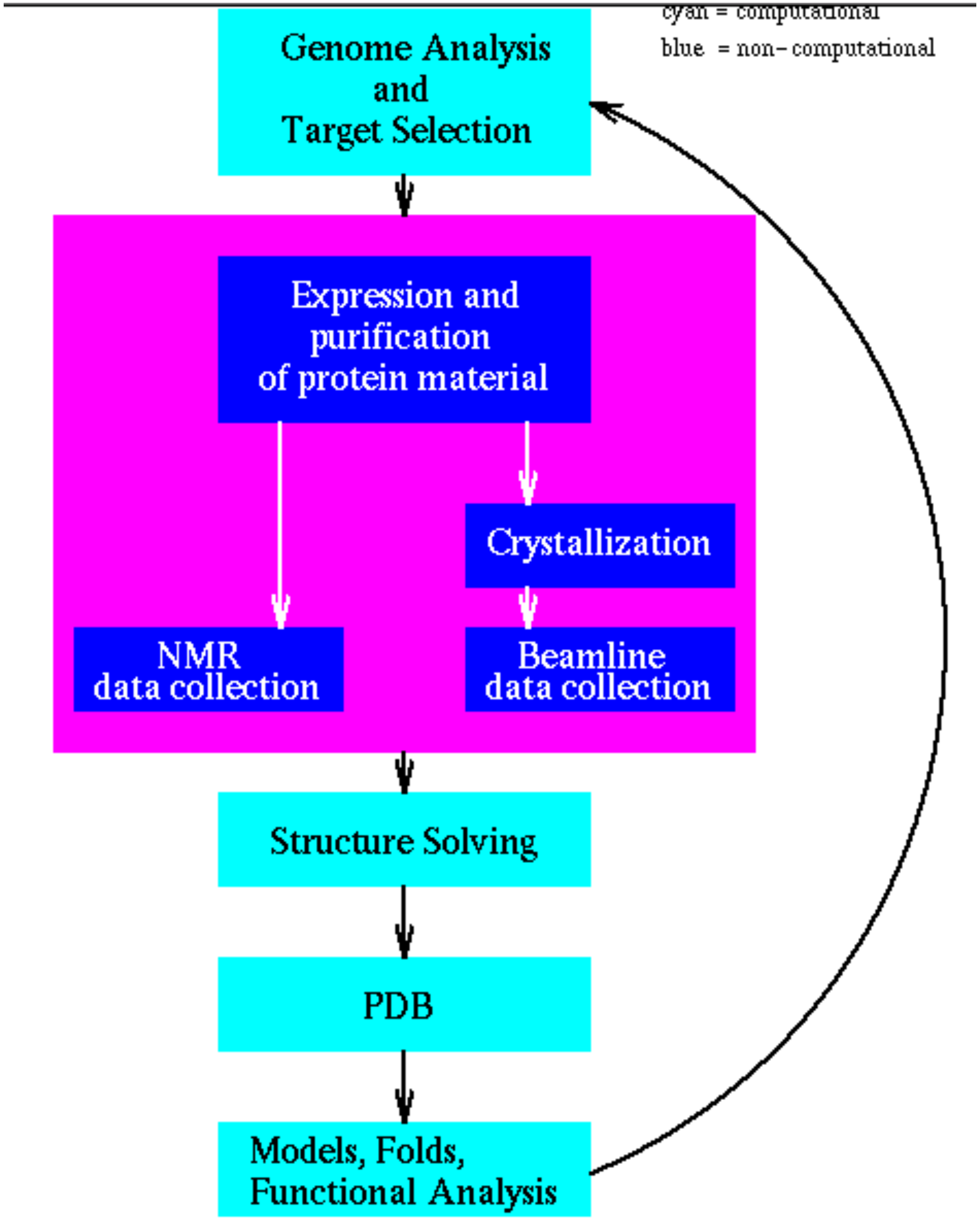
- •25,000 proteins
- •10,000 subset
- •30% ID or
- •30 seq
- •Solve by 2010
- •$20,000/Structure



30 seq

cyan = computational

blue = non-computational



Genome Analysis and Target Selection

Expression and purification of protein material

Crystallization

NMR data collection

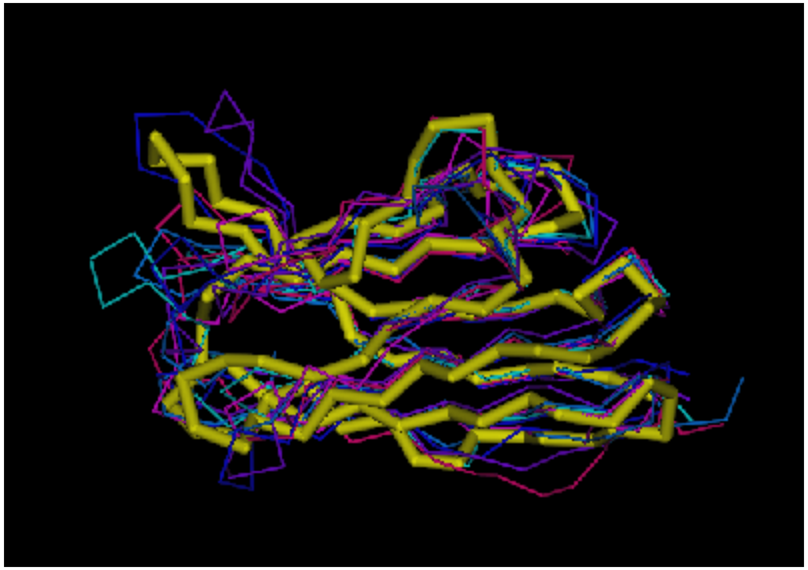
Beamline data collection

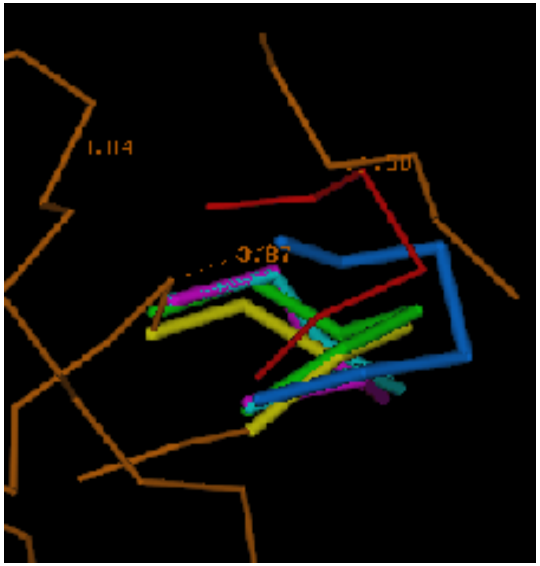
Structure Solving

PDB

Models, Folds, Functional Analysis

Comparative (Homology) Modelling





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

Homology Modelling*

- Based on the observation that “Similar sequences exhibit similar structures”
- Known structure is used as a template to model an unknown (but likely similar) structure with known sequence
- First applied in late 1970’s using early computer imaging methods (Tom Blundell)

Homology Modelling*

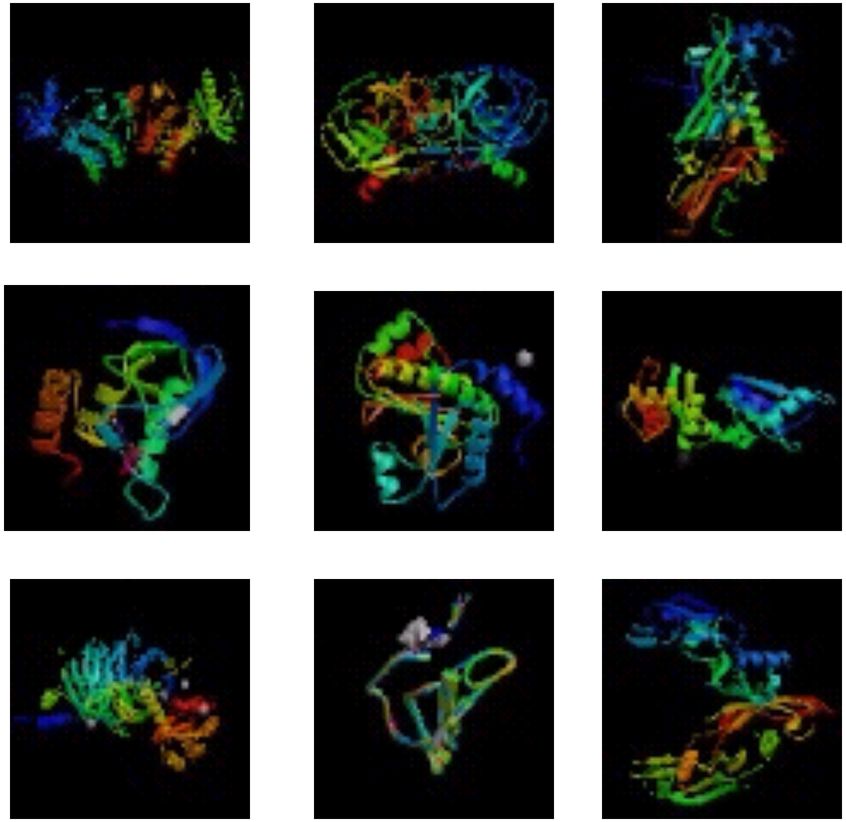
- Offers a method to “Predict” the 3D structure of proteins for which it is not possible to obtain X-ray or NMR data
- Can be used in understanding function, activity, specificity, etc.
- Of interest to drug companies wishing to do structure-aided drug design
- A keystone of Structural Proteomics

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

Step 1: ID Homologues in PDB

PRTEINSEQENCEPRTEINSEQUENC
EPRTEINSEQNCEQWERYTRASDFHG
TREWQIYPASDFGHKLMCNASQERWW
PRETWQLKHGFDSADAMNCVCNQWER
GFDHSDASFWERQWK



Query Sequence

PDB

Step 1: ID Homologues in PDB

PRTEINSEQENCEPRTEINSEQUENC
EPRTEINSEQNCEQWERYTRASDFHG
TREWQIYPASDFGHKLMCNASQERWW
PRETWQLKHGFDSADAMNCVCNQWER
GFDHSDASFWERQWK

Query Sequence

PRTEINSEQENCEPRTEINSEQUENC
EPRTEINSEQNCEQWERYTRASDFHG
TREWQIYPASDFGHKLMCNASQERWW
PRETWQLKHGFDSADAMNCVCNQWER
GFDHSDASFWERQWK

PRTEINSEQENCEPRTEINSEQUENC
EPRTEINSEQQWEWEWQWEWEQWEWEWQ
RYEYEWQWNCEQWERYTRASDFHG
TREWQIYPASDWERWEREWRFDSFG

Hit #1

PRTEINSEQENCEPRTEINSEQUENC
EPRTEINSEQNCEQWERYTRASDFHG
TREWQIYPASDFGHKLMCNASQERWW
PRETWQLKHGFDSADAMNCVCNQWER
GFDHSDASFWERQWK

PRTEINSEQENCEPRTEINSEQUENC
EPRTEINSEQQWEWEWQWEWEQWEWEWQ
RYEYEWQWNCEQWERYTRASDFHG
TR

PRTEINSEQENCEPRTEINSEQUENC
EPRTEINSEQNCEQWERYTRASDFHG
TREWQIYPASDFG

Hit #2

PRTEINSEQENCEPRTEINSEQUENC
EPRTEINSEQNCEQWERYTRASDFHG
TREWQIYPASDFGPRTEINSEQENCEPR
TEINSEQUENCEPRTEINSEQNCEQWER
YTRASDFHGTREWQIYPASDFG
TREWQIYPASDFGPRTEINSEQENCEPR
TEINSEQUENCEPRTEINSEQNCEQWER

YTRASDFHGTREWQ

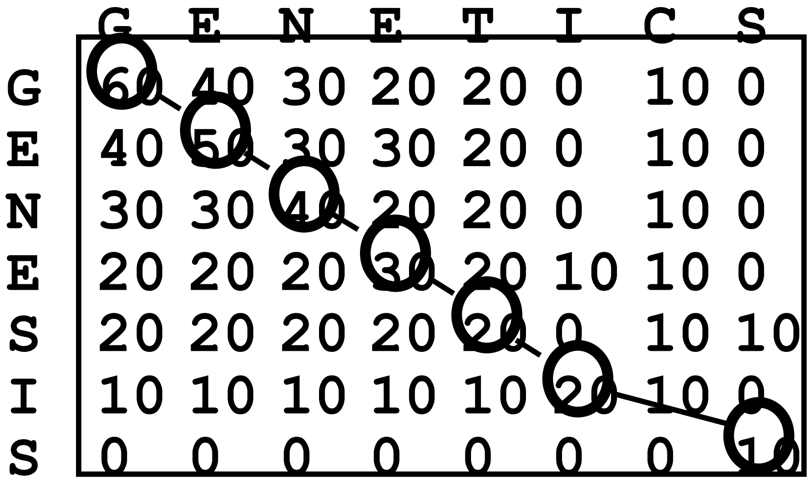
PRTEINSEQENCEPRTEINSEQUENC
EPRTEINSEQNCEQWERYTRASDFHG
TREWQIYPASDFG

PRTEINSEQENCEPRTEINSEQUENC
EPRTEINSEQNCEQWERYTRASDFHG
TREWQIYPASDFGPRTEINSEQENC

PDB

Step 2: Align Sequences

	G	E	N	E	T	I	C	S
G	10	0	0	0	0	0	0	0
E	0	10	0	10	0	0	0	0
N	0	0	10	0	0	0	0	0
E	0	0	0	10	0	0	0	0
S	0	0	0	0	0	0	0	10
I	0	0	0	0	0	10	0	0
S	0	0	0	0	0	0	0	10

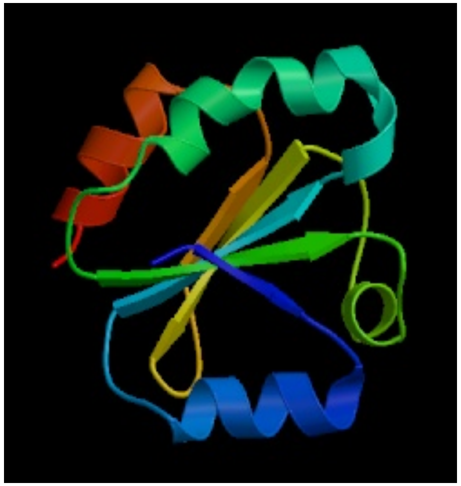


	G	E	N	E	T	I	C	S
G	60	40	30	20	20	0	10	0
E	40	50	30	30	20	0	10	0
N	30	30	40	20	20	0	10	0
E	20	20	20	30	20	10	10	0
S	20	20	20	20	20	0	10	10
I	10	10	10	10	10	20	10	0
S	0	0	0	0	0	0	0	10

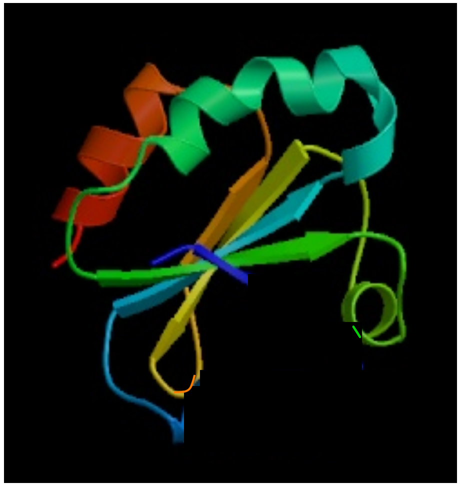
Dynamic Programming

Step 2: Align Sequences

```
Query   ACDEFGHIKLMNPQRST--FGHQWERT-----TYREWYEG
Hit #1  ASDEYAHLRILDPQRSTVAYAYE--KSFAPPGSFKWEYEA
Hit #2  MCDEYAHIRLMNPERSTVAGGHQWERT----GSFKEWYAA
```



Hit #1

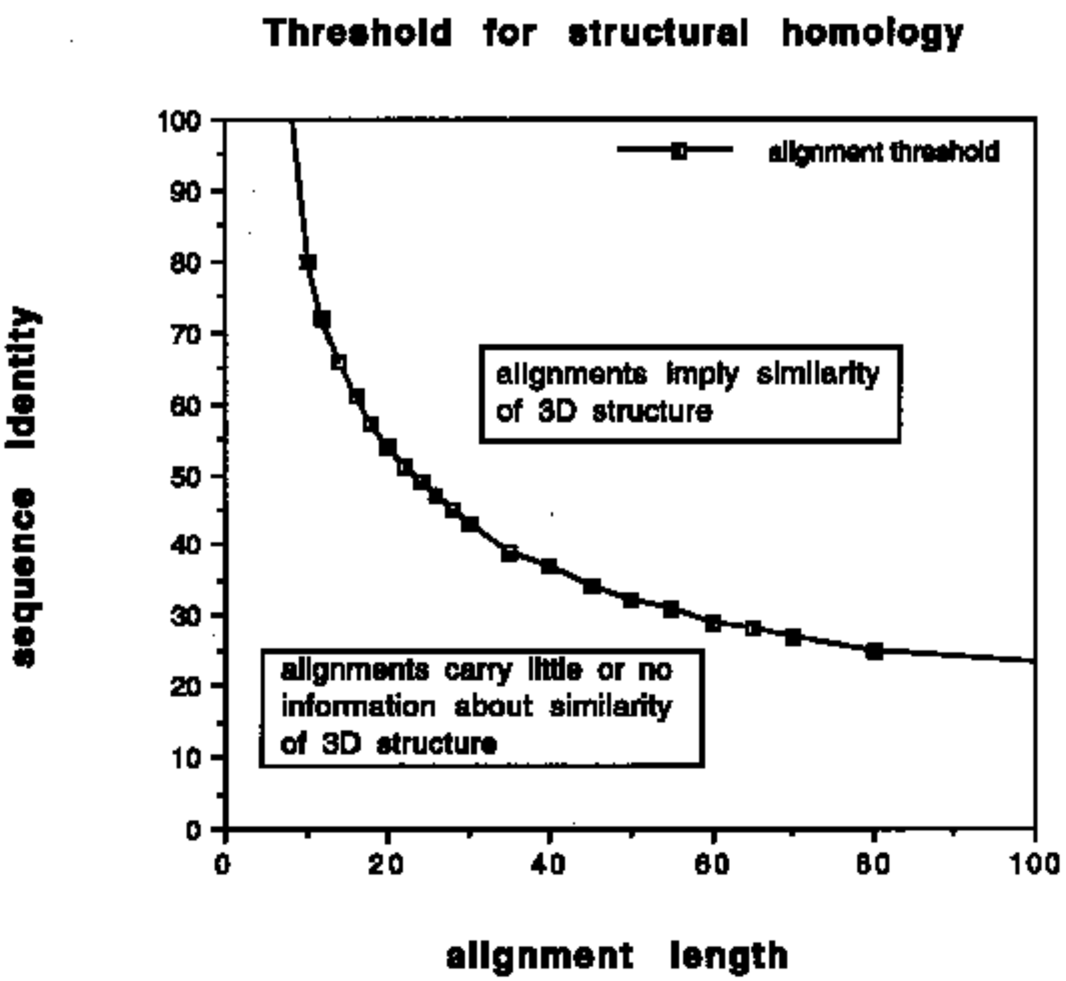


Hit #2

Alignment*

- Key step in Homology Modelling
- Global (Needleman-Wunsch) alignment is absolutely required
- Small error in alignment can lead to big error in structural model
- Multiple alignments are usually better than pairwise alignments

Alignment Thresholds*

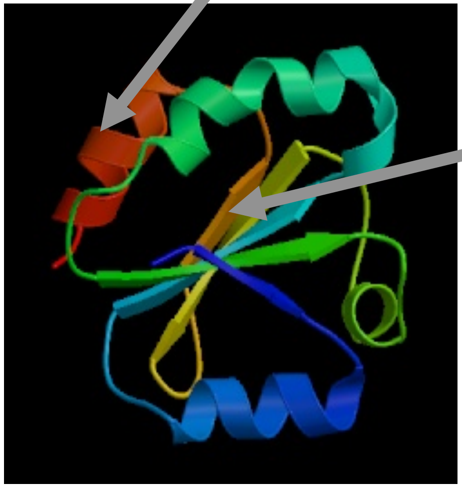


Step 3: Find SCR’s

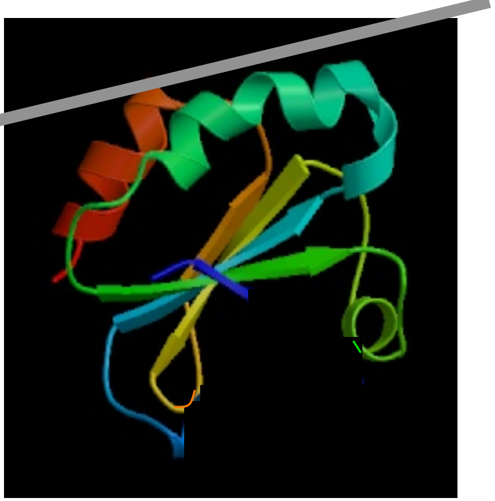
```
Query   ACDEFGHIKLMNPQRST--FGHQWERT-----TYREWYEG
Hit #1  ASDEYAHLRILDPQRSTVAYAYE--KSFAPPGSFKWEYEA
Hit #2  MCDEYAHIRLMNPERSTVAGGHQWERT----GSFKEWYAA
        HHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCBBBBBBBBB
```

SCR #1

SCR #2



Hit #1



Hit #2

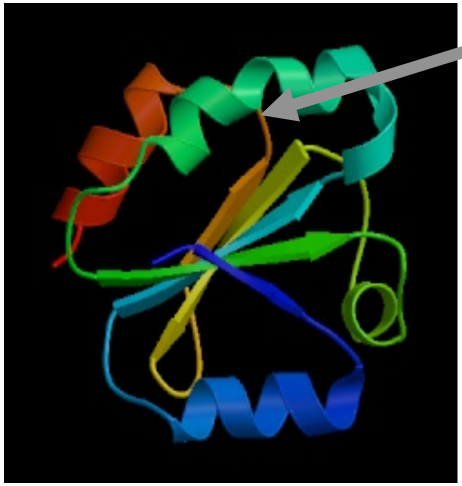
Structurally Conserved Regions (SCR’s)*

- Corresponds to the most stable structures or regions (usually interior) of protein
- Corresponds to sequence regions with lowest level of gapping, highest level of sequence conservation
- Usually corresponds to secondary structures

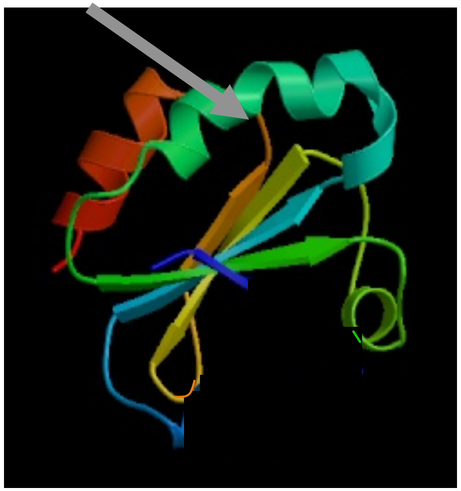
Step 4: Find SVR’s

```
Query   ACDEFGHIKLMNPQRST--FGHQWERT-----TYREWYEG
Hit #1  ASDEYAHLRILDPQRSTVAYAYE--KSFAPRGSFKWEYEA
Hit #2  MCDEYAHIRLMNPERSTVAGGHQWERT----GSFKEWYAA
        HHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCBBBBBBBBB
```

SVR (loop)



Hit #1



Hit #2

Structurally Variable Regions (SVR’s)*

- Corresponds to the least stable or most flexible regions (usually exterior) of protein
- Corresponds to sequence regions with highest level of gapping, lowest level of sequence conservation
- Usually corresponds to loops and turns

Step 5: Generate Coordinates

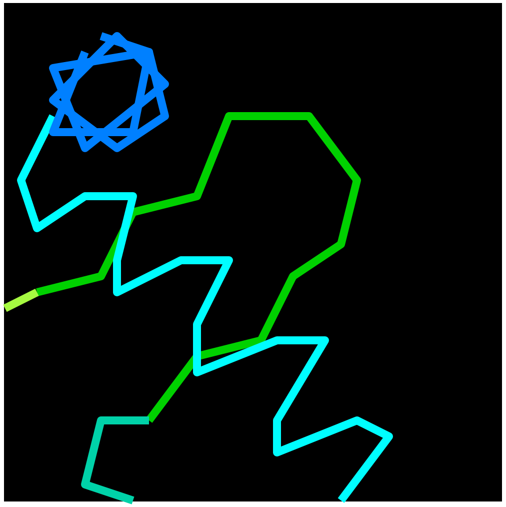
```
                           ALA
ATOM      1  N   ~~SER~~ A   1      21.389  25.406  -4.628  1.00 23.22      2TRX 152
ATOM      2  CA  SER A   1      21.628  26.691  -3.983  1.00 24.42      2TRX 153
ATOM      3  C   SER A   1      20.937  26.944  -2.679  1.00 24.21      2TRX 154
ATOM      4  O   SER 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      1  N   ALA A   1      21.389  25.406  -4.628  1.00 23.22      2TRX 152
ATOM      2  CA  ALA A   1      21.628  26.691  -3.983  1.00 24.42      2TRX 153
ATOM      3  C   ALA A   1      20.937  26.944  -2.679  1.00 24.21      2TRX 154
ATOM      4  O   ALA A   1      21.072  28.079  -2.093  1.00 24.97      2TRX 155
ATOM      5  CB  ALA 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   GLU A   2      20.173  26.028  -2.163  1.00 21.39      2TRX 158
ATOM      8  CA  GLU A   2      19.395  26.125  -0.949  1.00 21.57      2TRX 159
ATOM      9  C   GLU A   2      20.264  26.214   0.297  1.00 20.89      2TRX 160
ATOM     10  O   GLU A   2      19.760  26.575   1.371  1.00 21.49      2TRX 161
```

Step 5: Generate Core Coordinates*

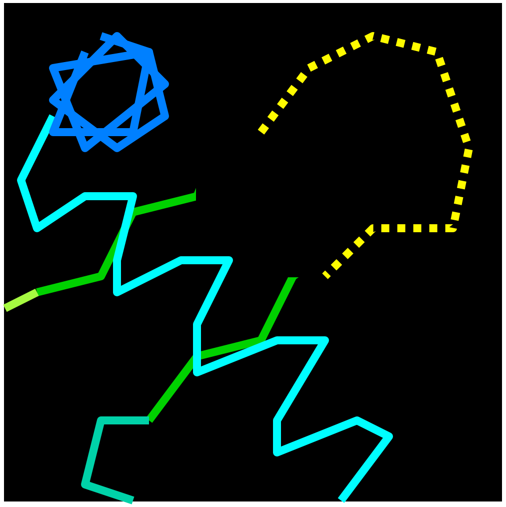
- For identical amino acids, transfer all atom coordinates (XYZ) to query protein
- For similar amino acids, transfer backbone coordinates & replace side chain atoms while respecting χ angles
- For different amino acids, transfer only the backbone coordinates (XYZ) to query sequence

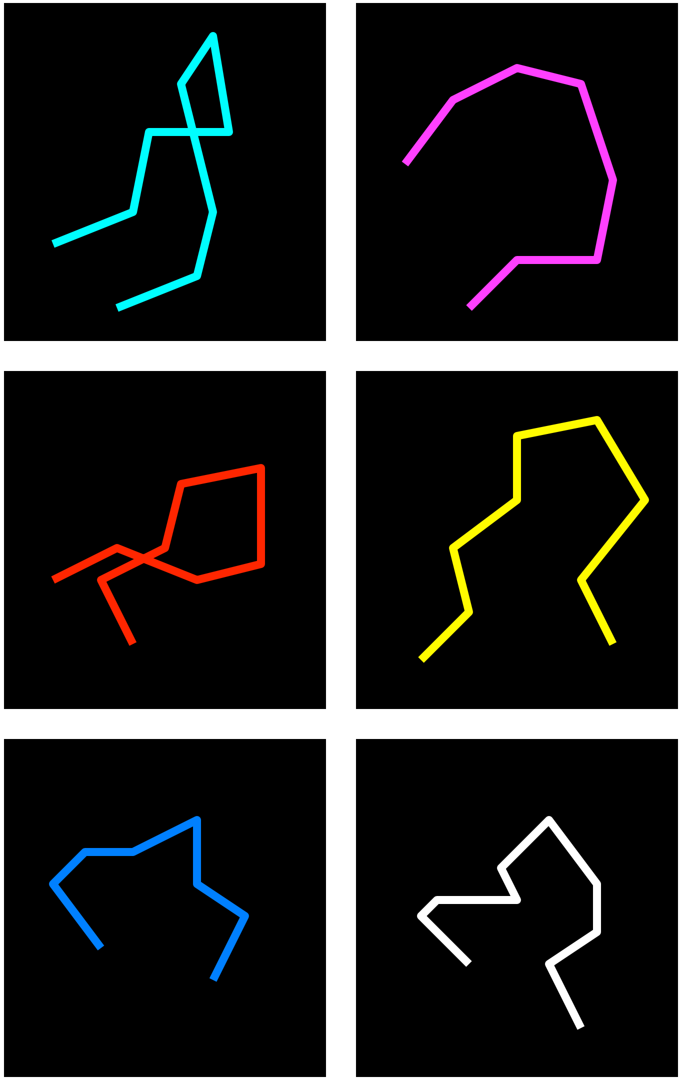
Step 6: Replace SVRs (loops)



Query FGHQWERT

Hit #1 YAYE--KS





Loop Library*

- Loops extracted from PDB using high resolution (<2 Å) X-ray structures
- Typically thousands of loops in DB
- Includes loop coordinates, sequence, # residues in loop, Ca-Ca distance, preceding 2° structure and following 2° structure (or their Ca coordinates)

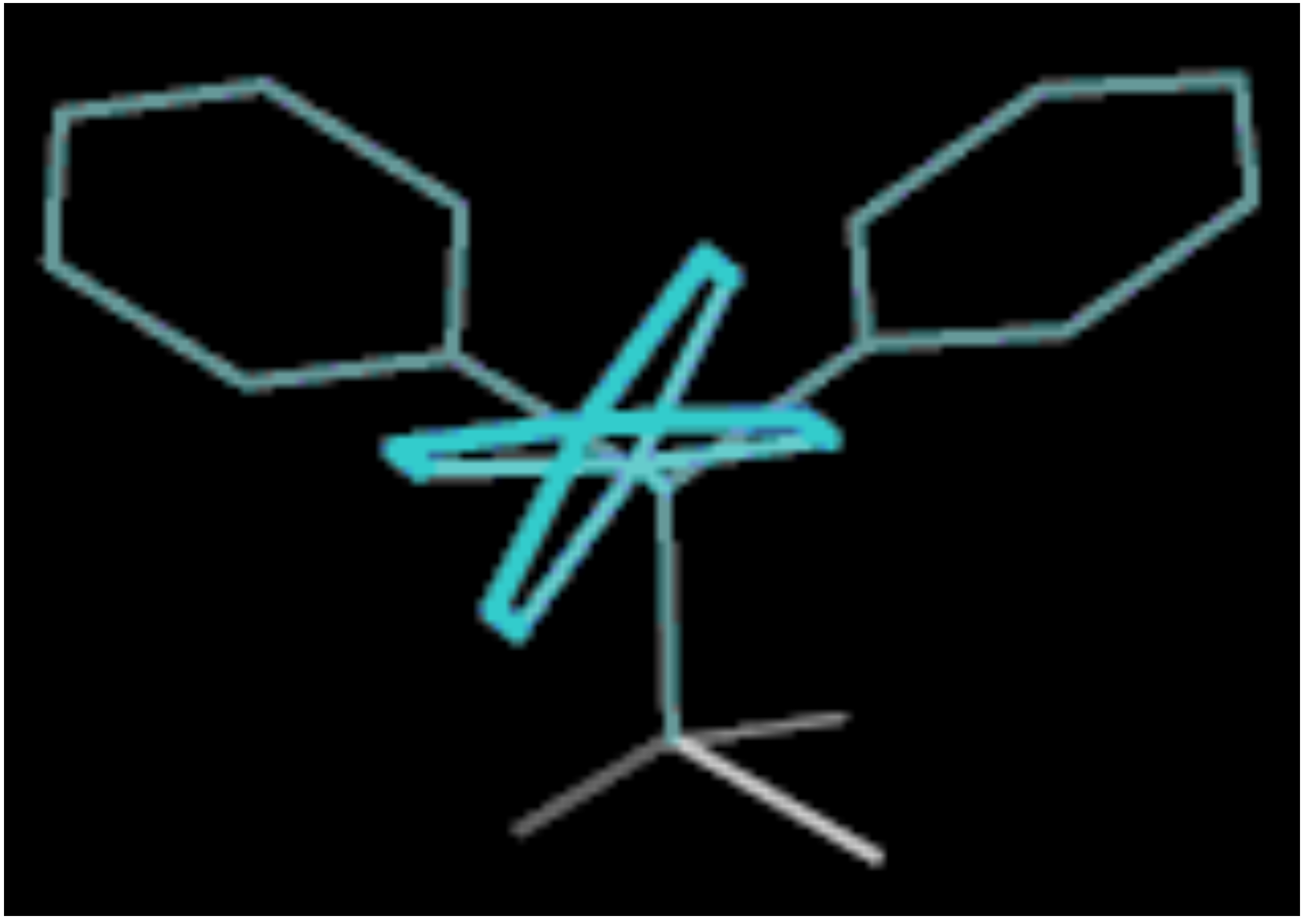
Step 6: Replace SVRs (loops)*

- Must match desired # residues
- Must match Ca-Ca distance (<0.5 Å)
- Must not bump into other parts of protein (no Ca-Ca distance <3.0 Å)
- Preceding and following Ca' s (3 residues) from loop should match well with corresponding Ca coordinates in template structure

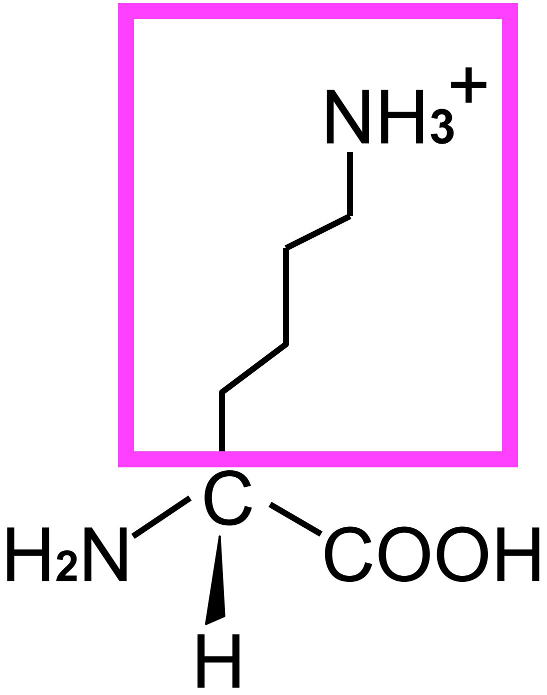
Step 6: Replace SVRs (loops)

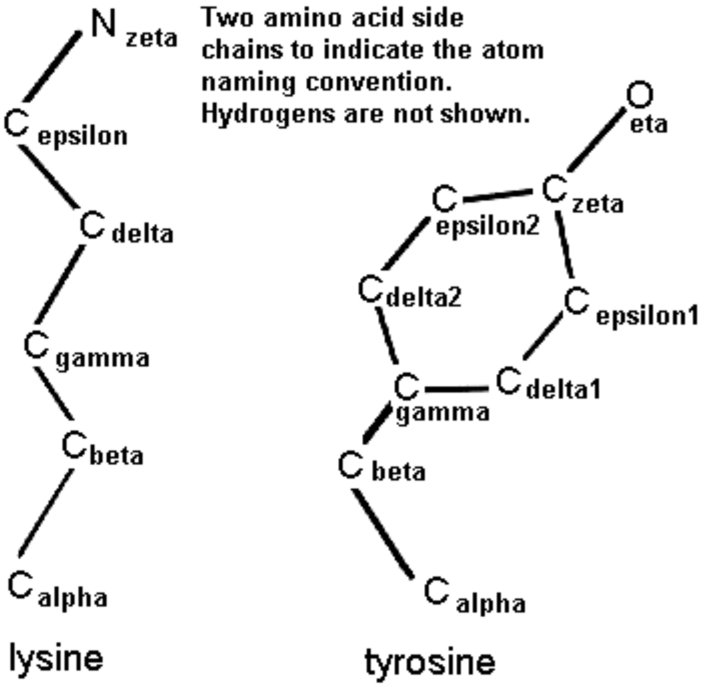
- Loop placement and positioning is done using superposition algorithm
- Loop fits are evaluated using RMSD calculations and standard “bump checking”
- If no “good” loop is found, some algorithms create loops using randomly generated ϕ/ψ angles

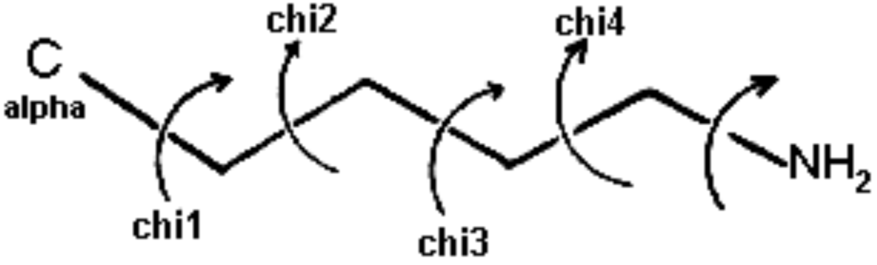
Step 7: Add Side Chains



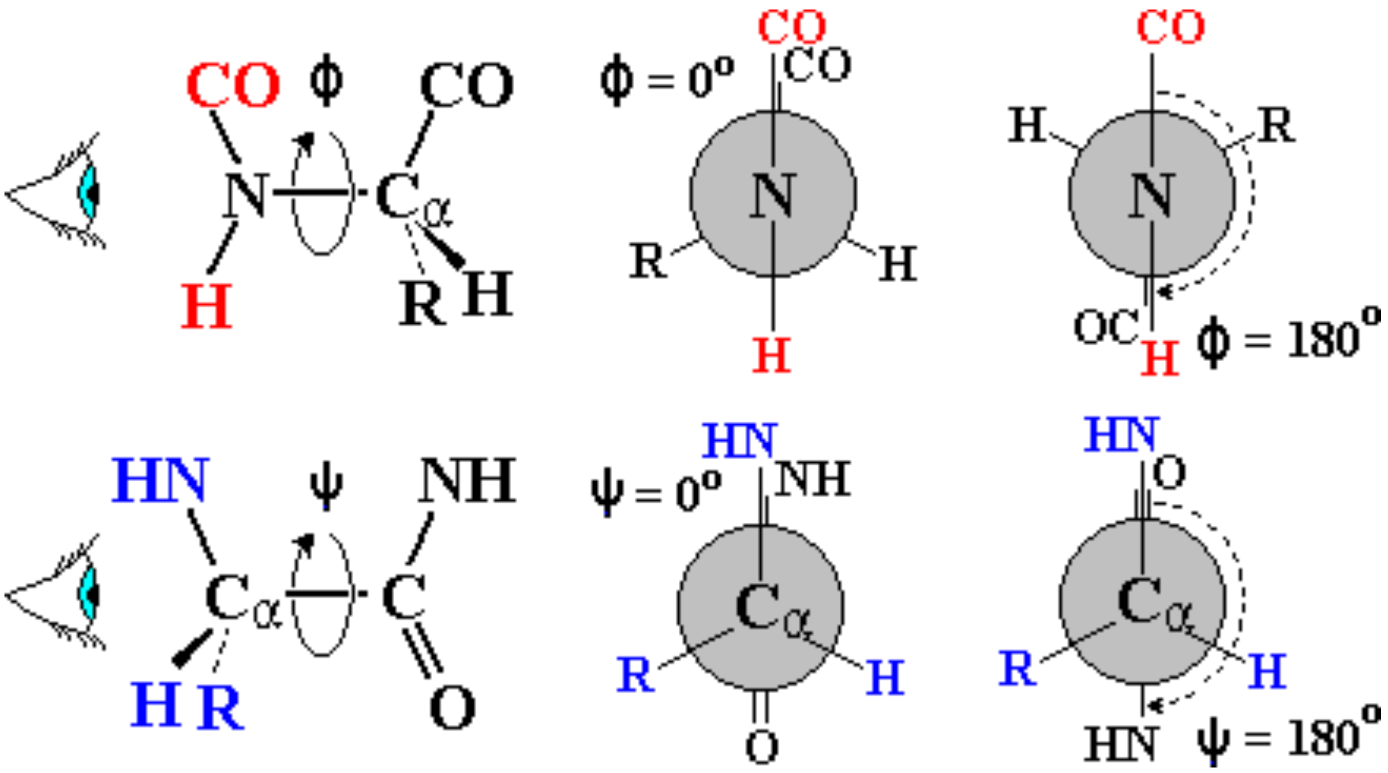
Amino Acid Side Chains*



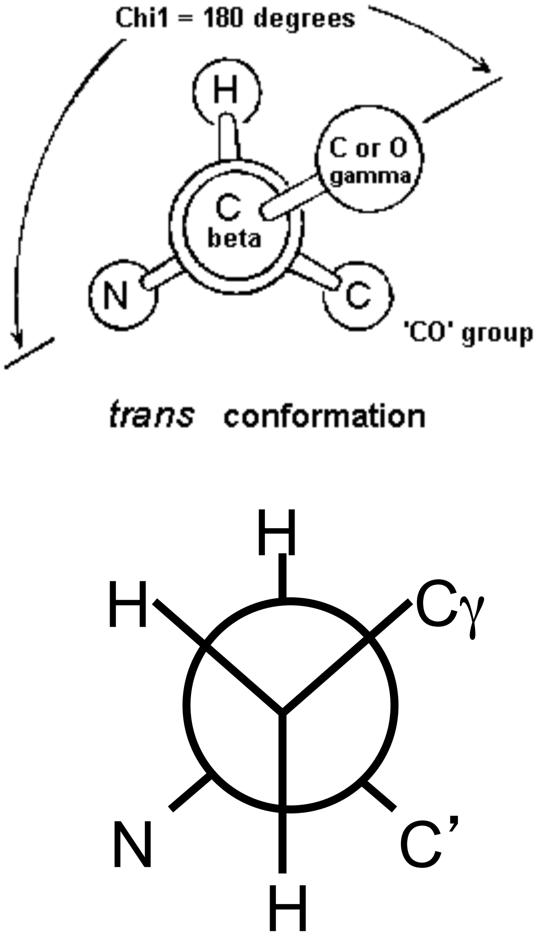

Two amino acid side chains to indicate the atom naming convention. Hydrogens are not shown.



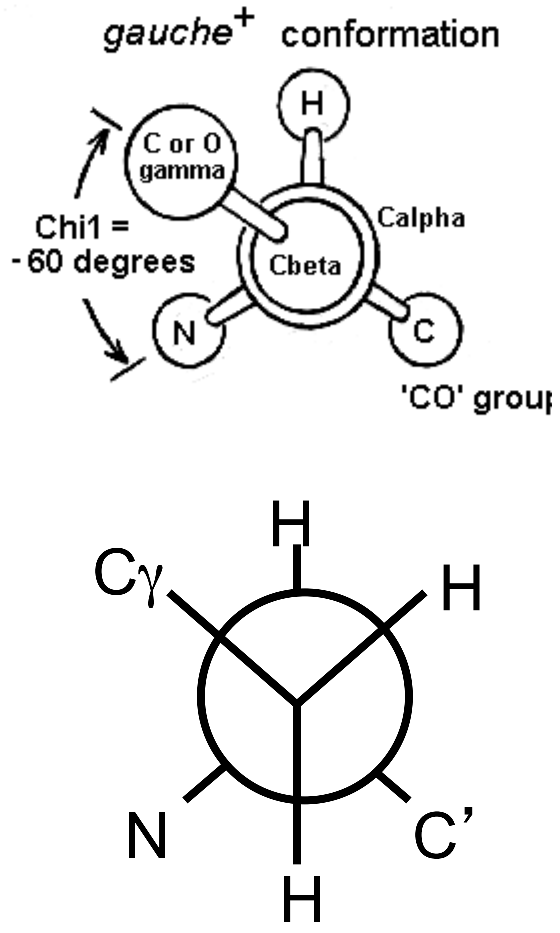
Newman Projections



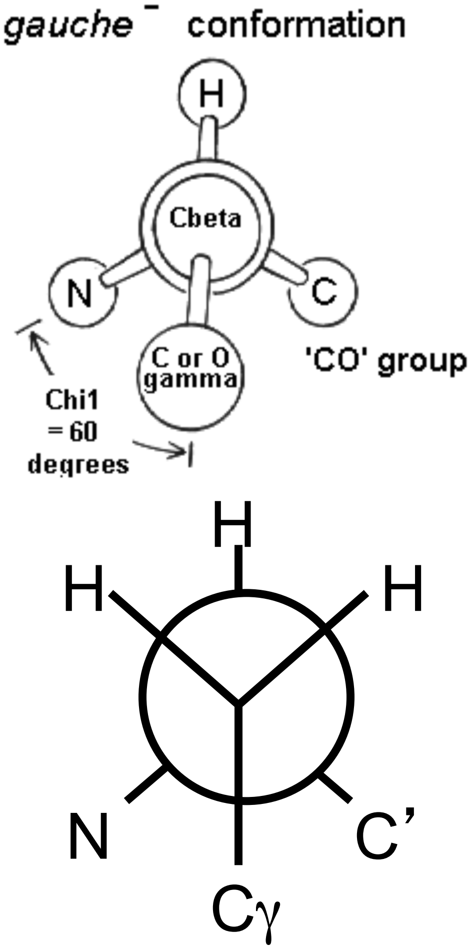
Newman Projections*



t



g+

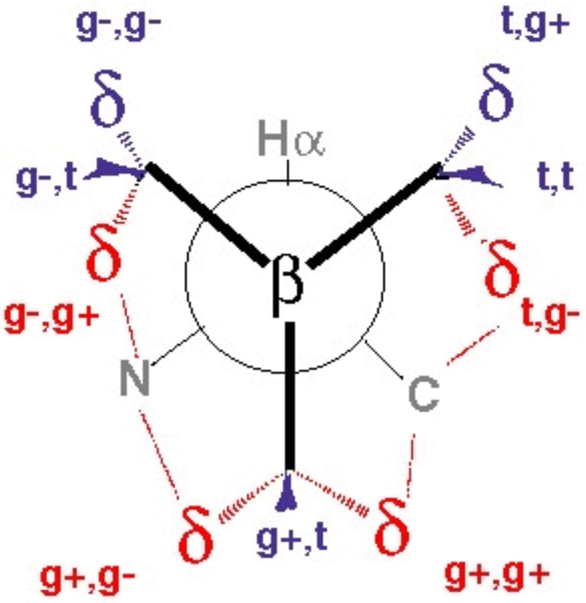


g-

Preferred Side Chain χ Angles*

Some combinations are BAD.

Some are OK.

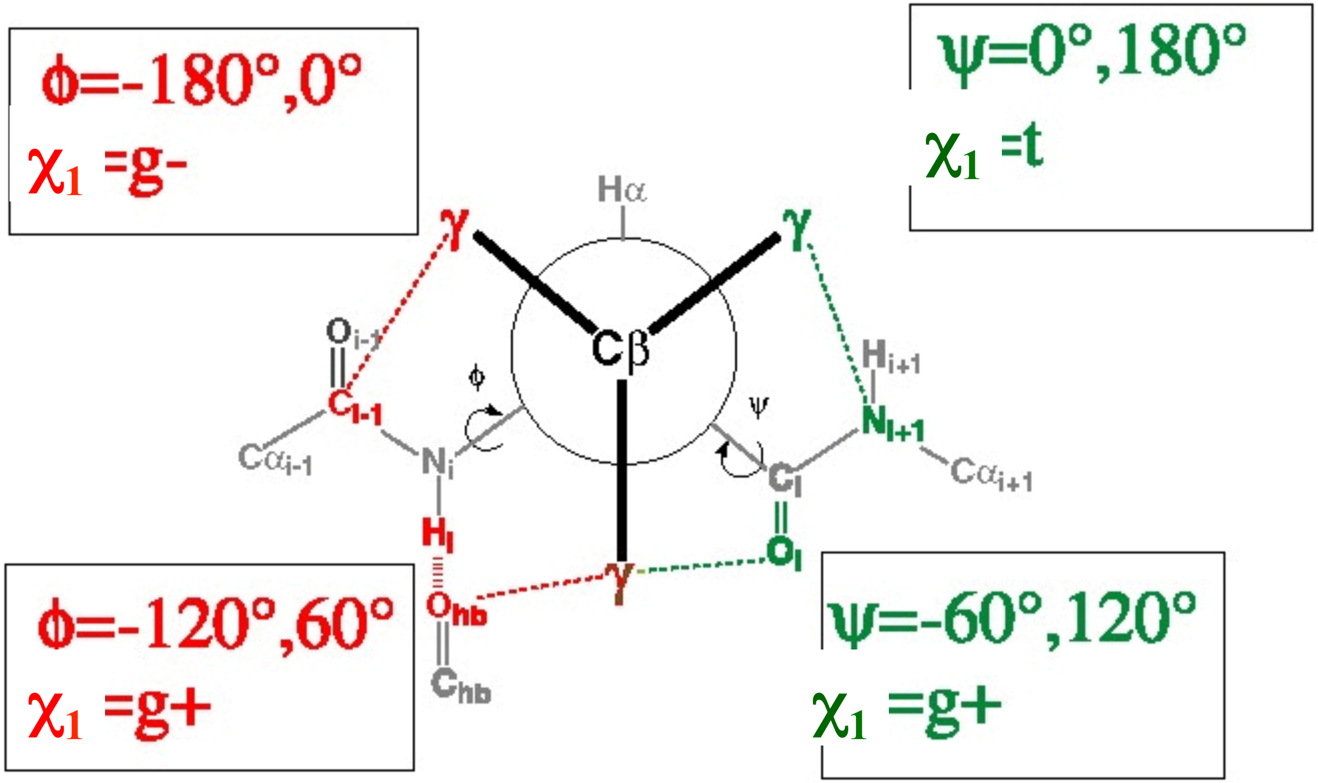


Relation Between χ and ϕ/ψ*

Some ϕ χ_1 combinations are **BAD.**

Some ψ χ_1 combinations are **BAD.**

The rest are **OK.**



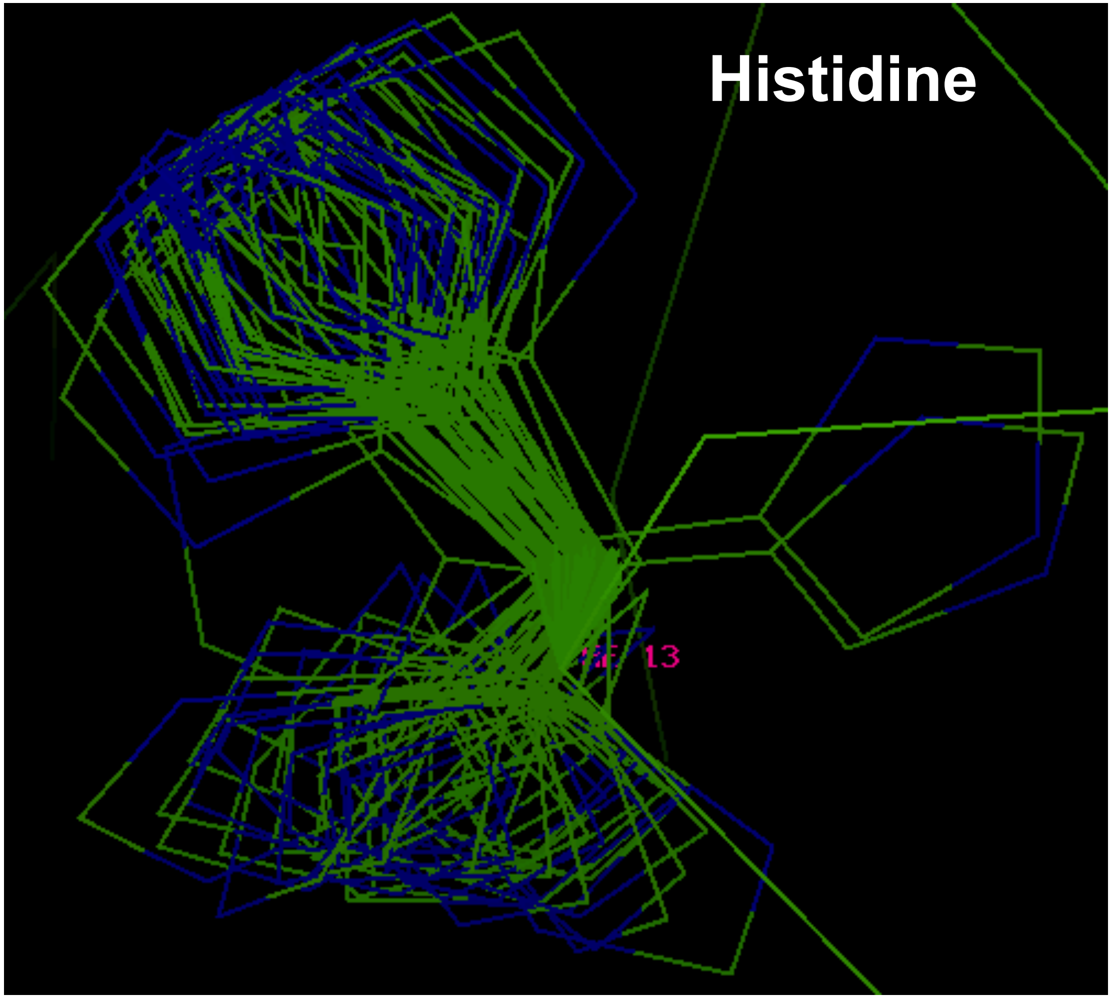
ϕ=-180°,0°
χ_1 =g-

ψ=0°,180°
χ_1 =t

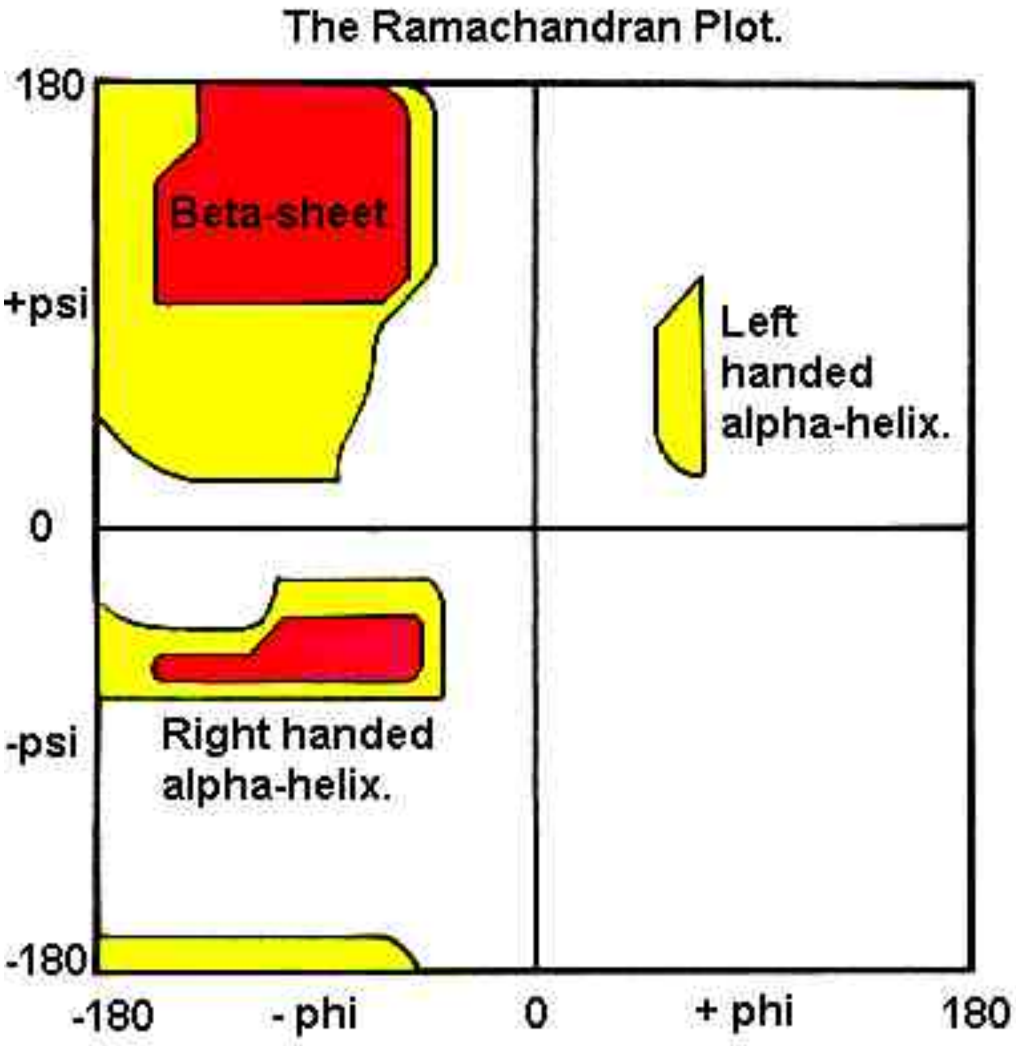
ϕ=-120°,60°
χ_1 =g+

ψ=-60°,120°
χ_1 =g+

Relation Between χ and ϕ/ψ

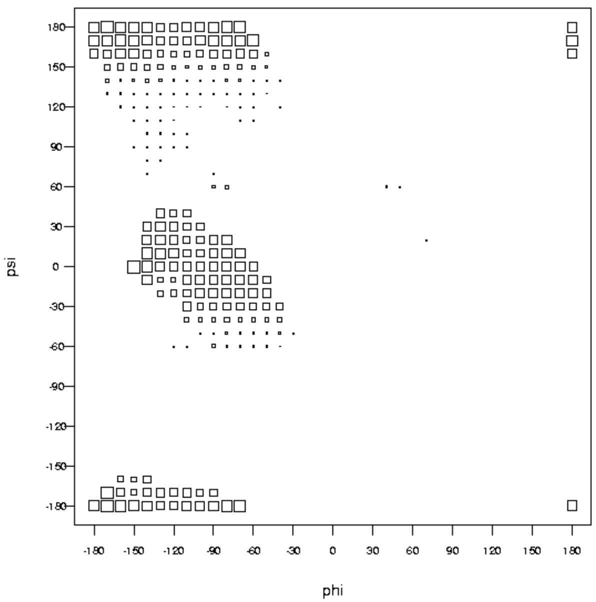


Relation Between χ and φ/ψ



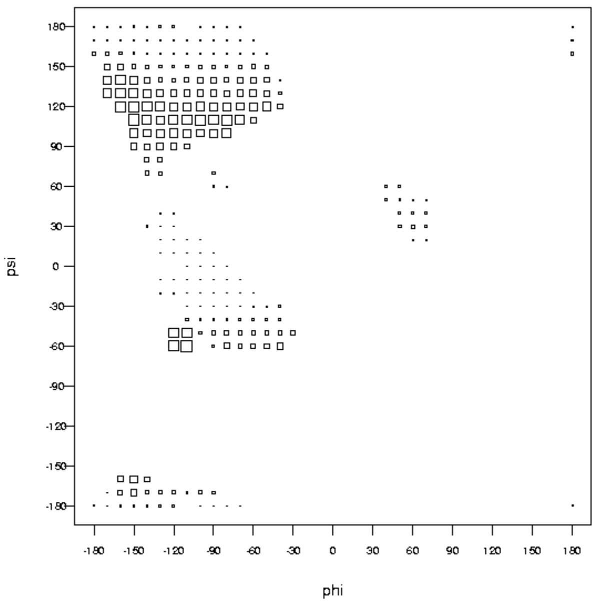
Relation Between χ and φ/ψ*

SER r1=g+



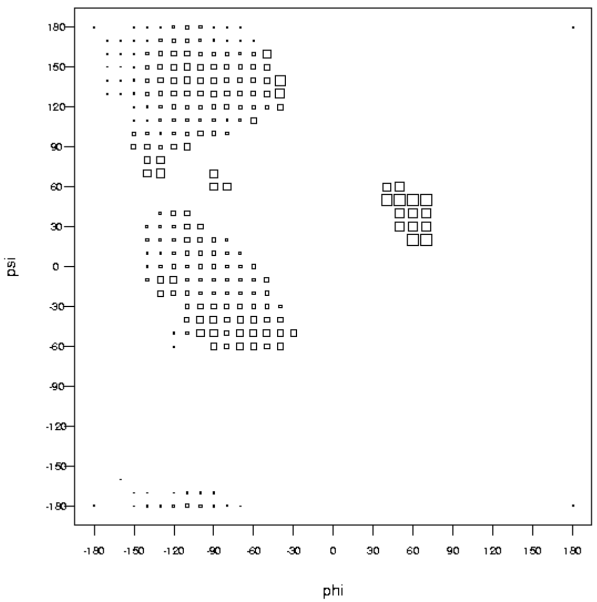
Box widths proportional to r1 populations in each phi,psi range

SER r1=t



Box widths proportional to r1 populations in each phi,psi range

SER r1=g-



Box widths proportional to r1 populations in each phi,psi range

g+

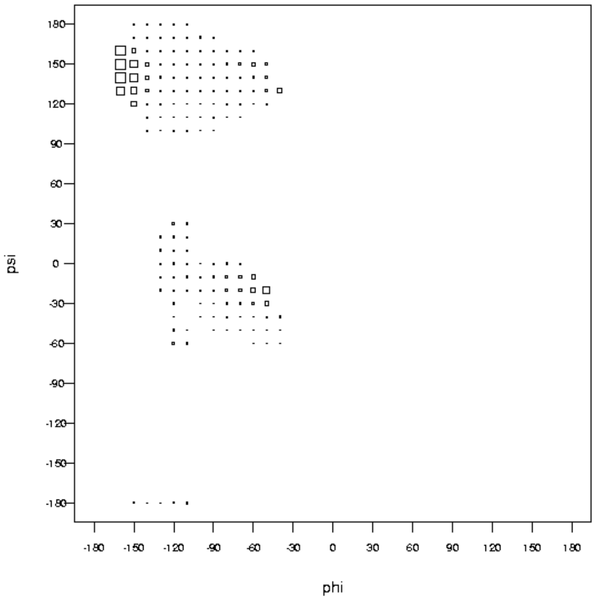
t

g-

Serine

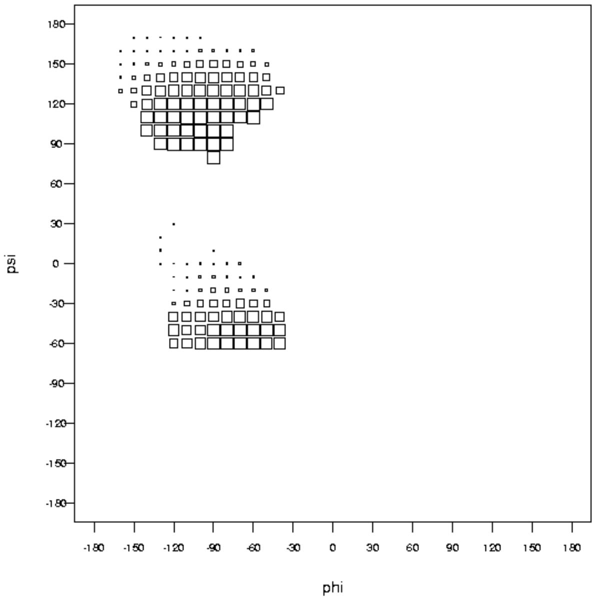
Relation Between χ and φ/ψ*

VAL r1=g+



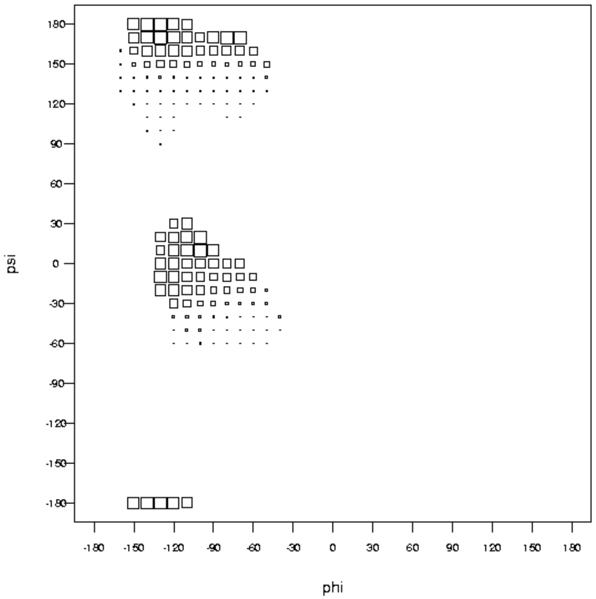
Box widths proportional to r1 populations in each phi,psi range

VAL r1=t



Box widths proportional to r1 populations in each phi,psi range

VAL r1=g-



Box widths proportional to r1 populations in each phi,psi range

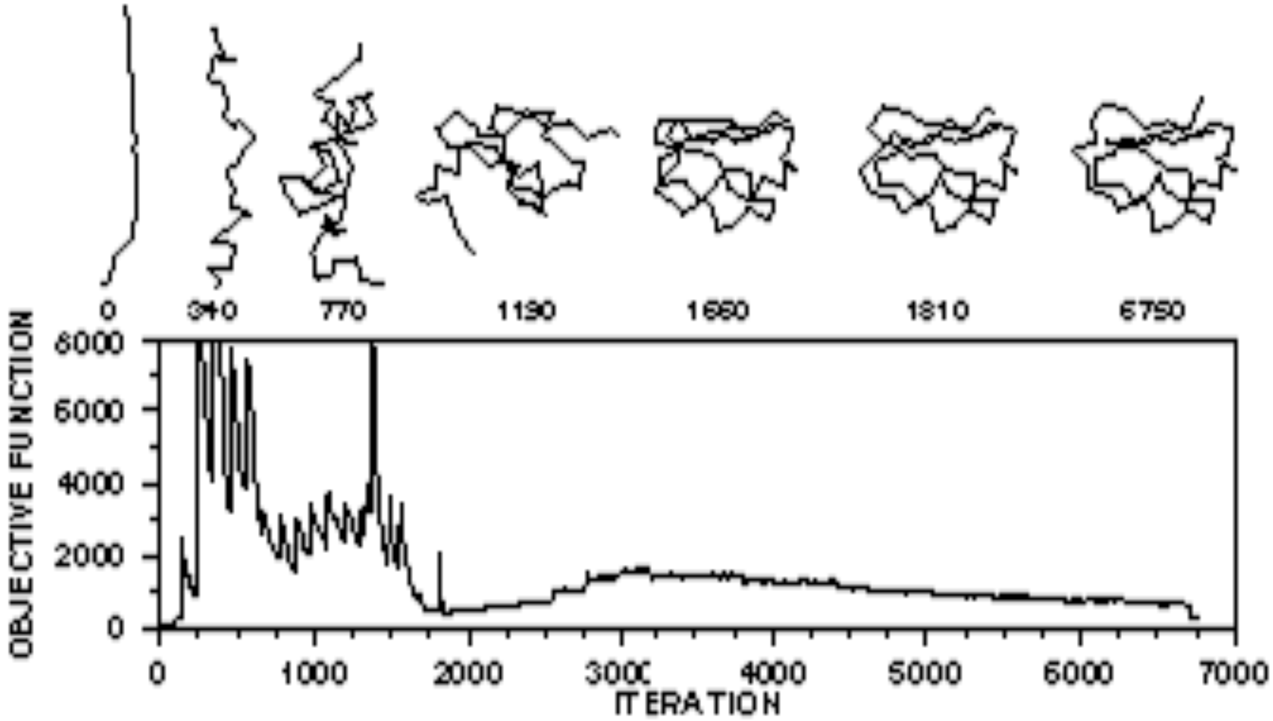
g+ **t** **g-**

Valine

Step 7: Add Side Chains*

- Done primarily for SVRs (not SCRs)
- Rotamer placement and positioning is done via a superposition algorithm using rotamers taken from a standardized library (Trial & Error)
- Rotamer fits are evaluated using simple “bump checking” methods

Step 8: Energy Minimization*



Energy Minimization*

- Efficient way of “polishing and shining” your protein model
- Removes atomic overlaps and unnatural strains in the structure
- Stabilizes or reinforces strong hydrogen bonds, breaks weak ones
- Brings protein to lowest energy in about 1-2 minutes CPU time

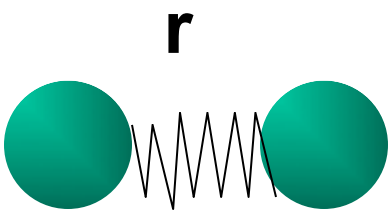
Energy Minimization (Theory)

- Treat Protein molecule as a set of balls (with mass) connected by rigid rods and springs
- Rods and springs have empirically determined force constants
- Allows one to treat atomic-scale motions in proteins as classical physics problems (OK approximation)

Standard Energy Function*

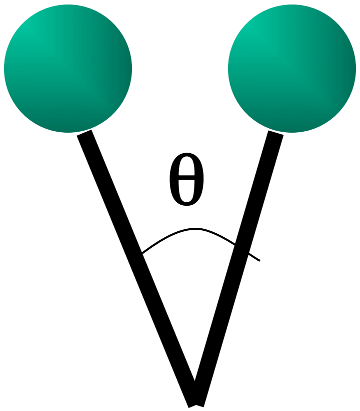
$E =$	$K_r(r_i - r_j)^2 +$	Bond length
	$K_\theta(\theta_i - \theta_j)^2 +$	Bond bending
	$K_\phi(1-\cos(n\phi_j))^2 +$	Bond torsion
	$q_i q_j/4\pi\varepsilon r_{ij} +$	Coulomb
	$A_{ij}/r^6 - B_{ij}/r^{12} +$	van der Waals
	$C_{ij}/r^{10} - D_{ij}/r^{12}$	H-bond

Energy Terms*



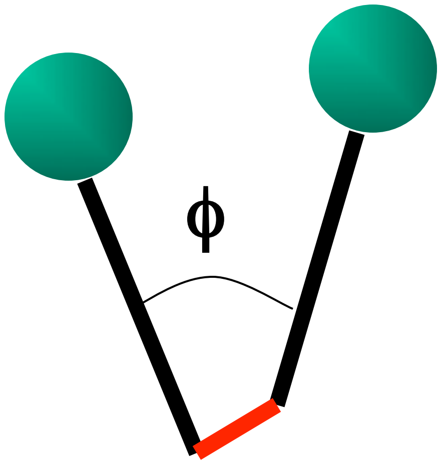
$K_r(r_i - r_j)^2$

Stretching



$K_\theta(\theta_i - \theta_j)^2$

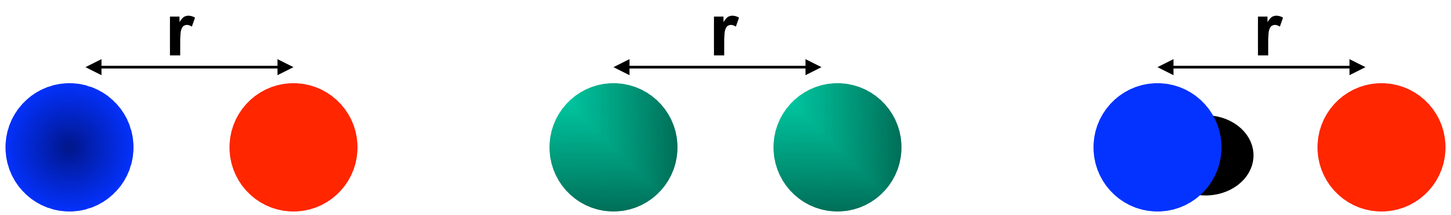
Bending



$K_\phi(1-\cos(n\phi_j))^2$

Torsional

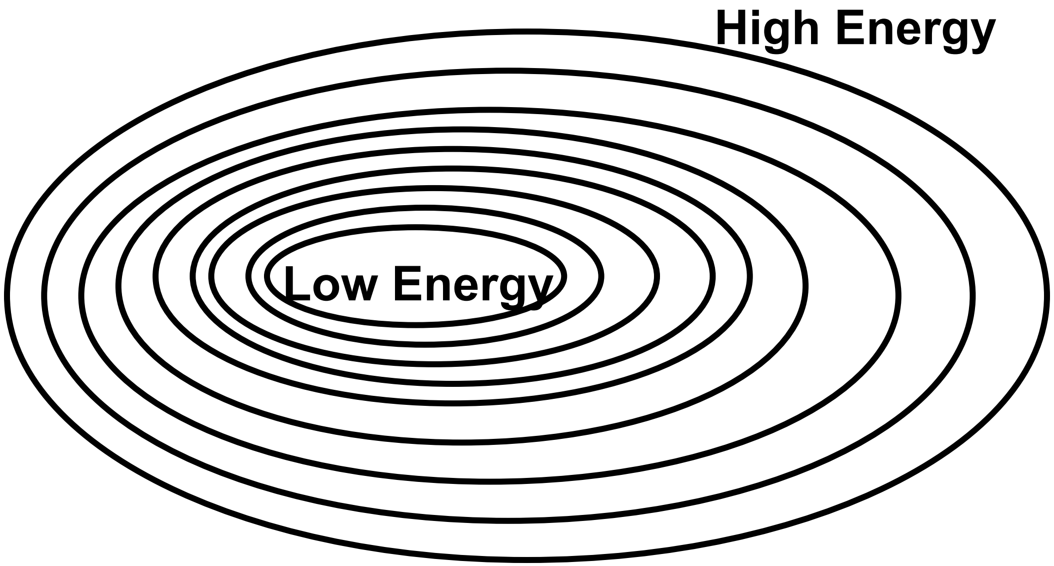
Energy Terms*



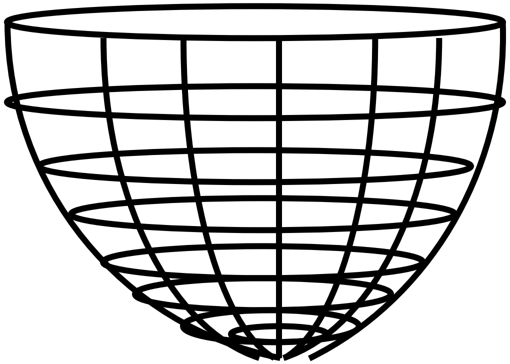
$q_iq_j/4\pi\varepsilon r_{ij}$ $A_{ij}/r^6 - B_{ij}/r^{12}$ $C_{ij}/r^{10} - D_{ij}/r^{12}$

Coulomb van der Waals H-bond

An Energy Surface



Overhead View



Side View

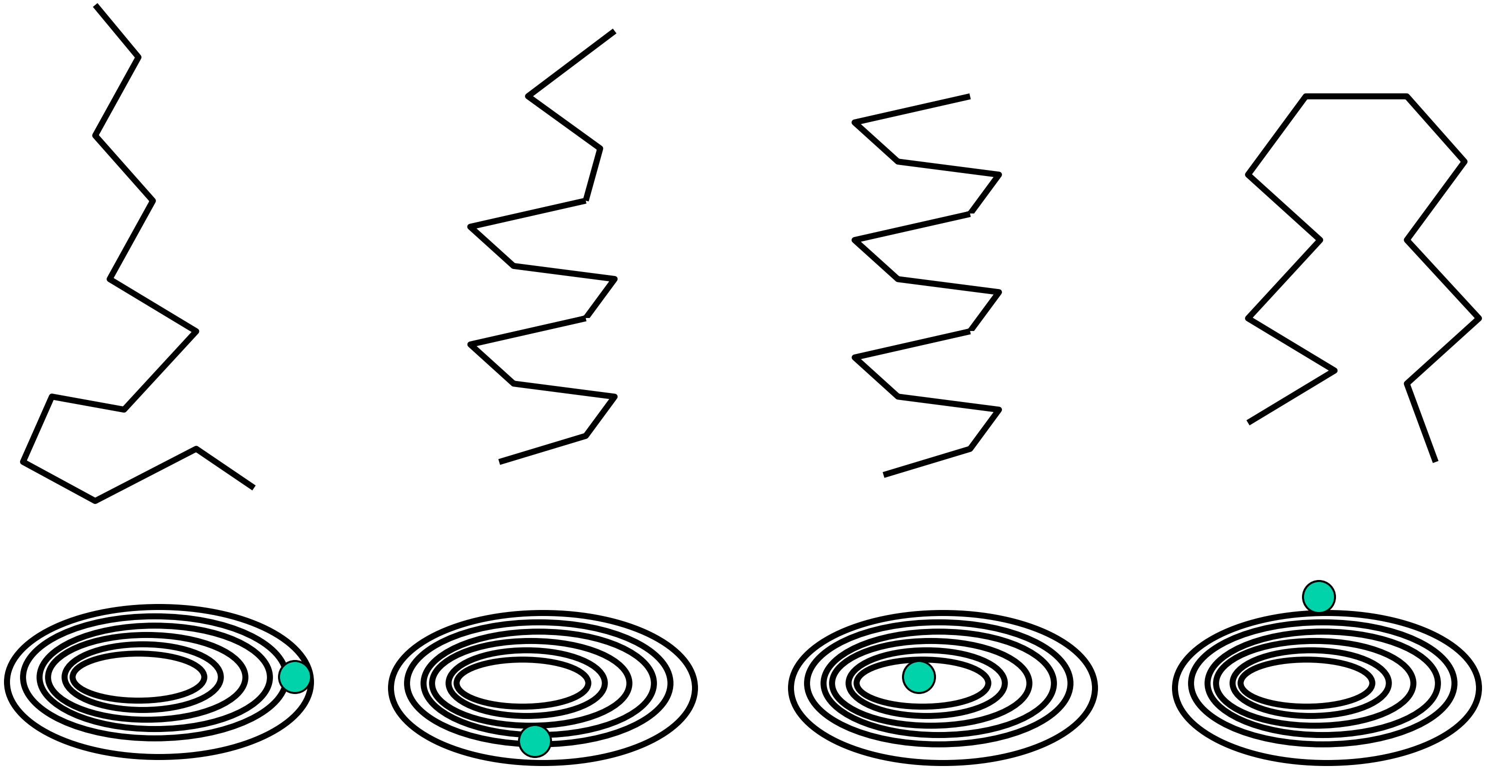
Minimization Methods*

- Energy surfaces for proteins are complex hyperdimensional spaces
- Biggest problem is overcoming local minimum problem
- Simple methods (slow) to complex methods (fast)
  - Monte Carlo Method
  - Steepest Descent
  - Conjugate Gradient

Monte Carlo Algorithm

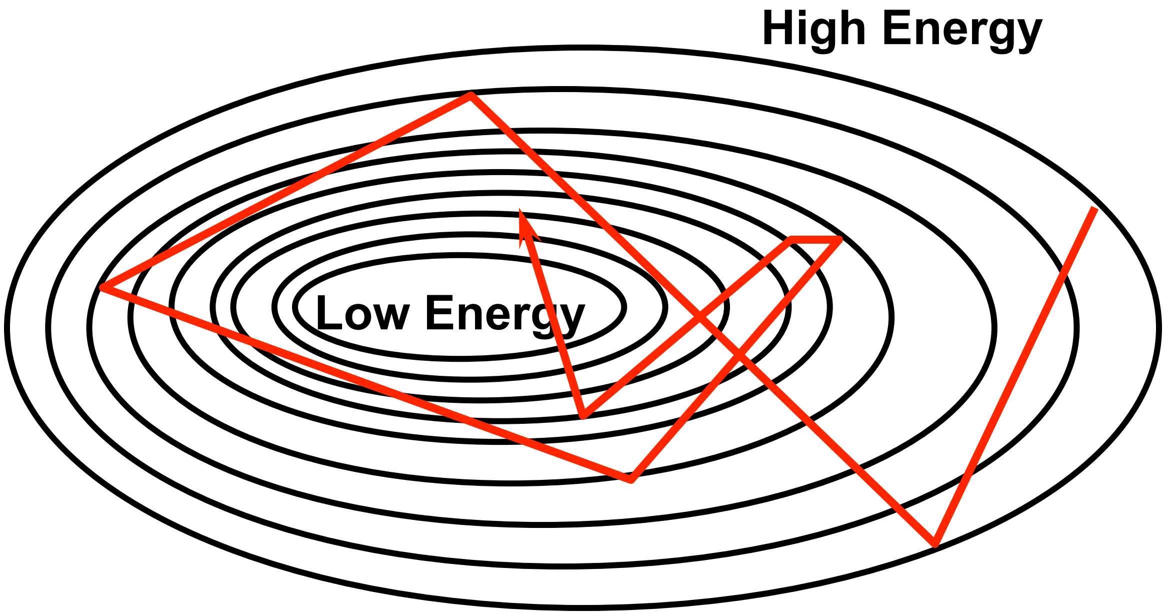
- Generate a conformation or alignment (a state)
- Calculate that state's energy or "score"
- If that state's energy is less than the previous state accept that state and go back to step 1
- If that state's energy is greater than the previous state accept it if a randomly chosen number is < $e^{-E/kT}$ where E is the state energy otherwise reject it
- Go back to step 1 and repeat until done

Conformational Sampling



Mid-energy lower energy lowest energy highest energy

Monte Carlo Minimization

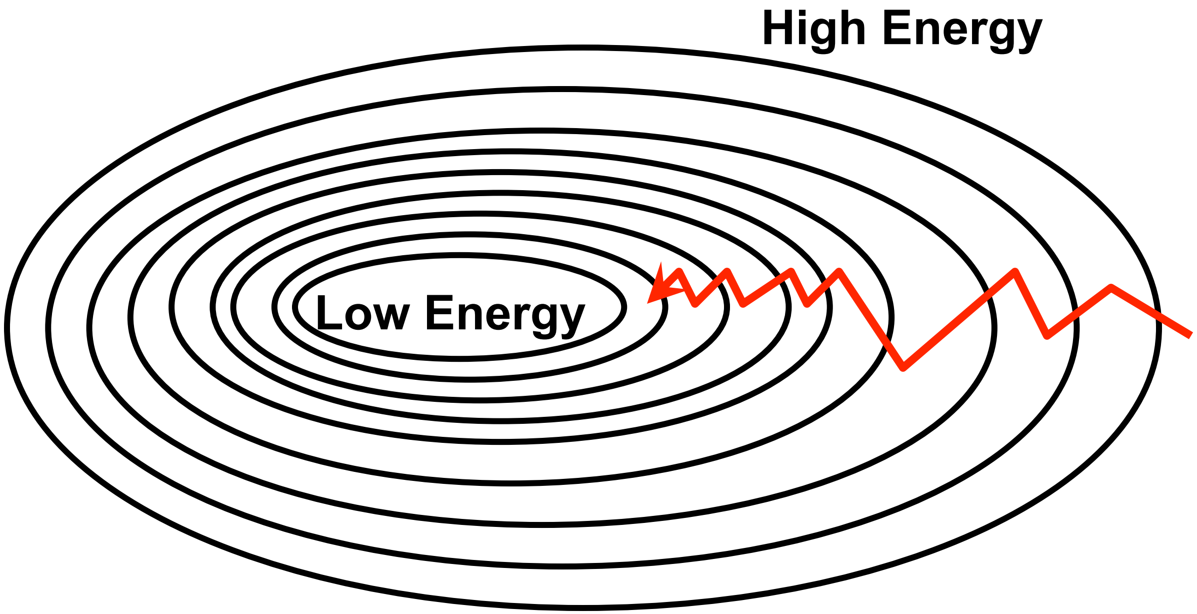


Performs a progressive or directed random search

Steepest Descent & Conjugate Gradients

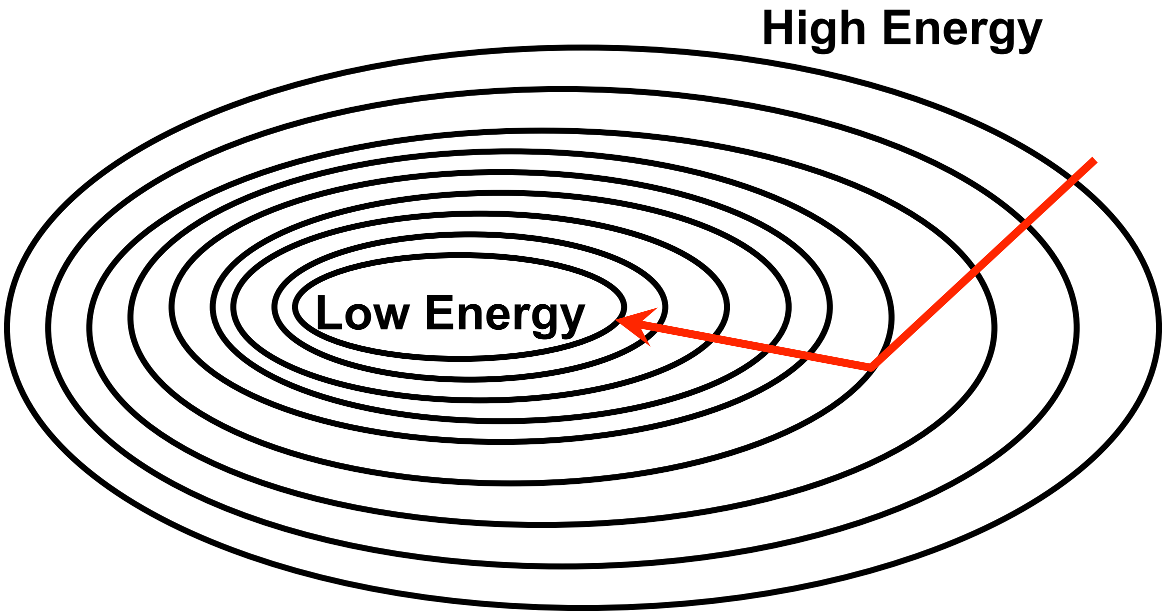
- Frequently used for energy minimization of large (and small) molecules
- Ideal for calculating minima for complex (I.e. non-linear) surfaces or functions
- Both use derivatives to calculate the slope and direction of the optimization path
- Both require that the scoring or energy function be differentiable (smooth)

Steepest Descent Minimization



Makes small locally steep moves down gradient

Conjugate Gradient Minimization

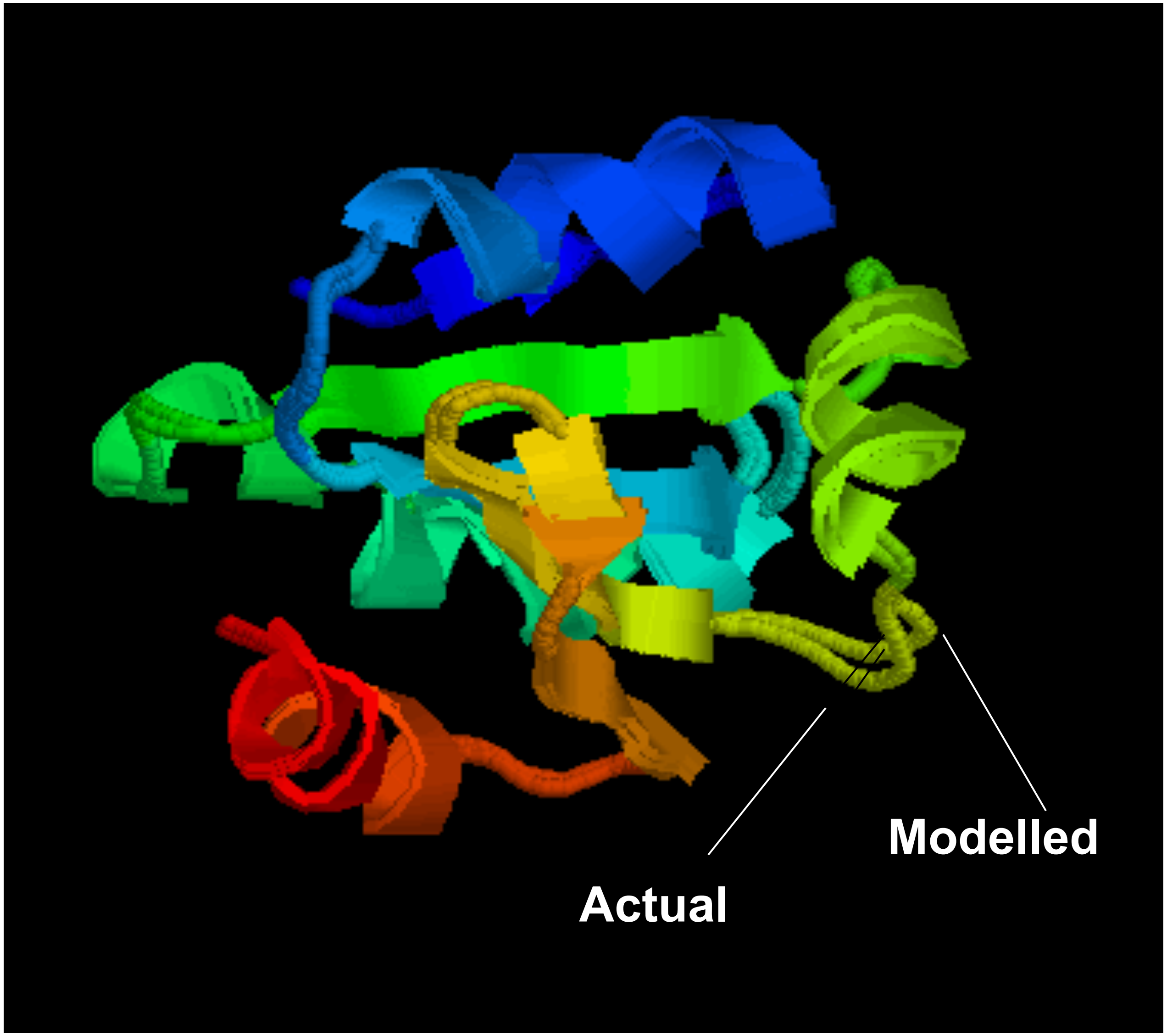


Includes information about the prior history of path

Energy Minimization*

- Very complex programs that have taken years to develop and refine
- Several freeware options to choose
  - XPLOR (Axel Brunger, Yale)
  - GROMACS (Gronnigen, The Netherlands)
  - AMBER (Peter Kollman, UCSF)
  - CHARMM (Martin Karplus, Harvard)
  - TINKER (Jay Ponder, Wash U))

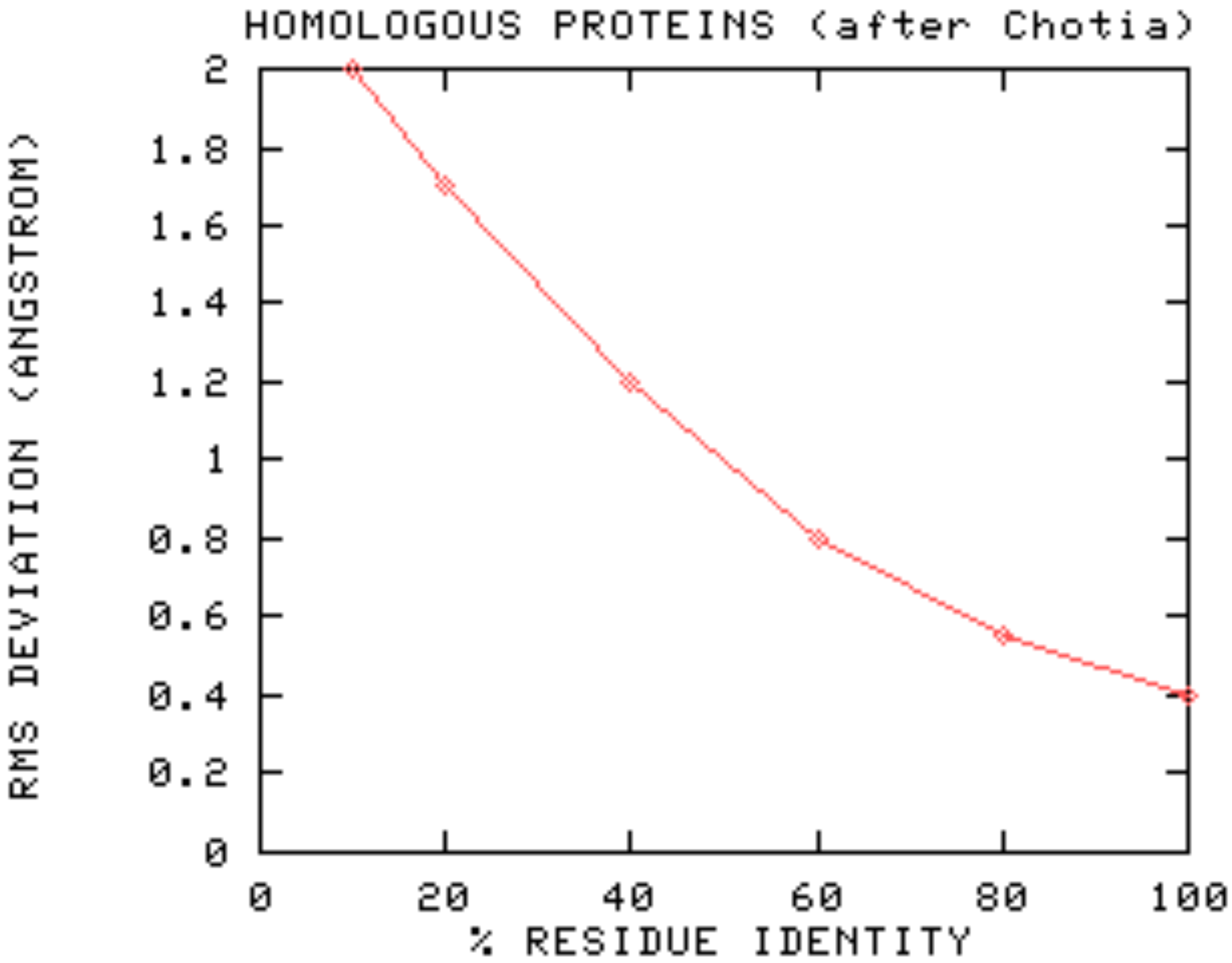
The Final Result



Summary*

- 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

How Good are Homology Models?



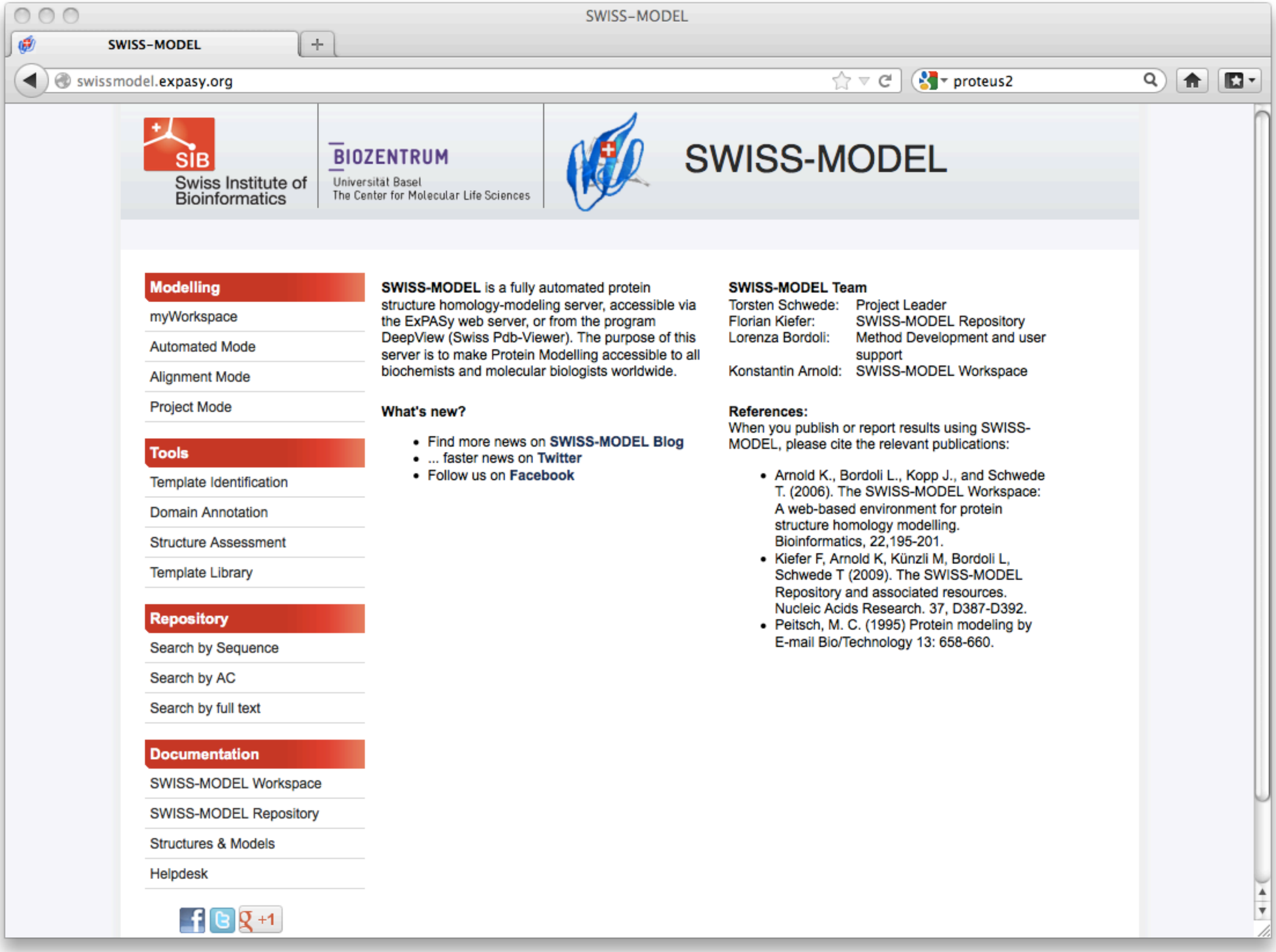
Outline

- The Protein Universe and the Protein Structure Initiative
- Homology (Comparative) Modelling of 3D Protein Structures
- Homology Modelling on the Web
- Assessing 3D Structures (modelled and experimental)

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*



Modelling
- myWorkspace
- Automated Mode
- Alignment Mode
- Project Mode

Tools
- Template Identification
- Domain Annotation
- Structure Assessment
- Template Library

Repository
- Search by Sequence
- Search by AC
- Search by full text

Documentation
- SWISS-MODEL Workspace
- SWISS-MODEL Repository
- Structures & Models
- Helpdesk

SWISS-MODEL is a fully automated protein structure homology-modeling 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 worldwide.

What's new?

- Find more news on **SWISS-MODEL Blog**
- ... faster news on **Twitter**
- Follow us on **Facebook**

SWISS-MODEL Team

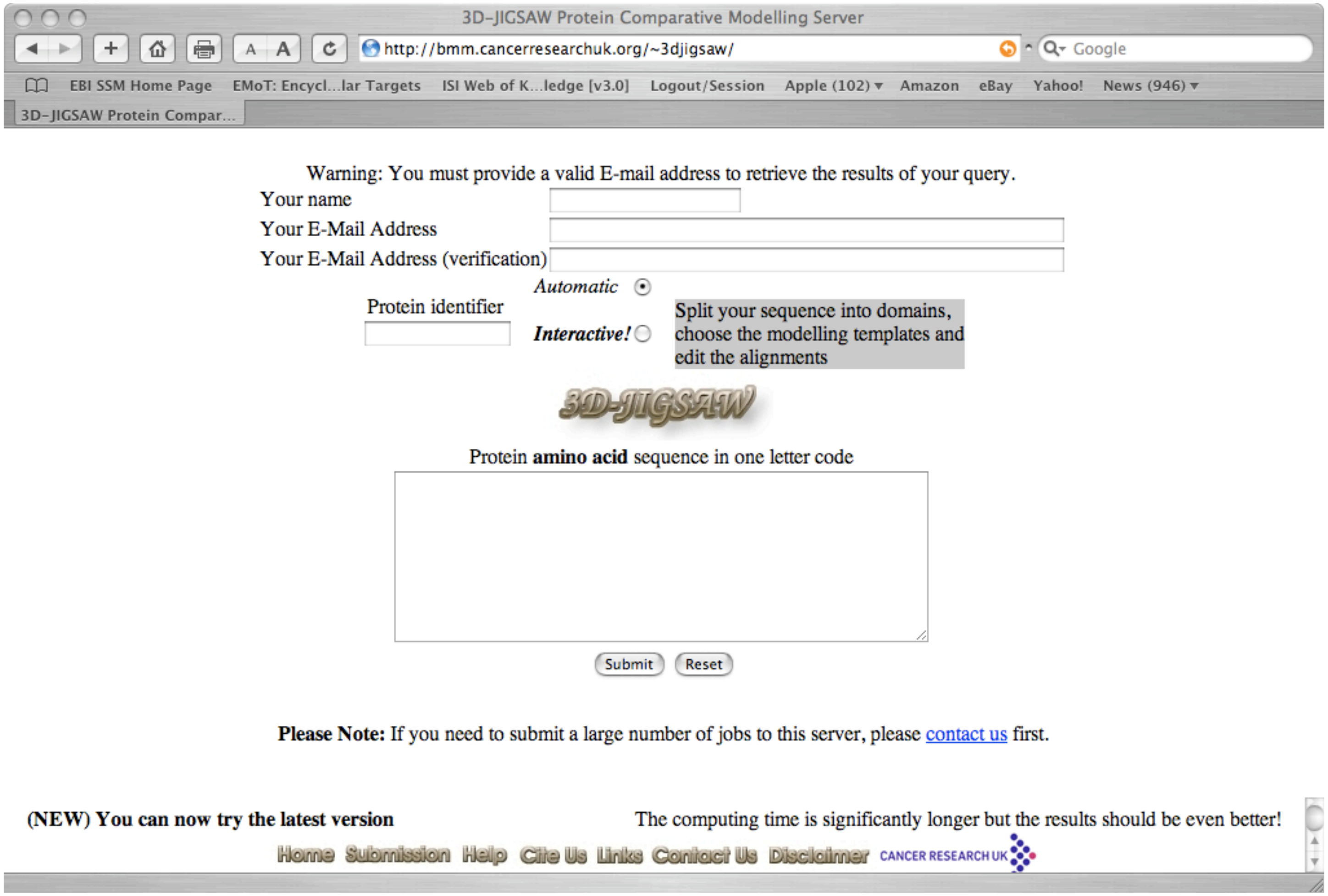
Torsten Schwede:	Project Leader
Florian Kiefer:	SWISS-MODEL Repository
Lorenza Bordoli:	Method Development and user support
Konstantin Arnold:	SWISS-MODEL Workspace

References:
When you publish or report results using SWISS-MODEL, please cite the relevant publications:

- Arnold K., Bordoli L., Kopp J., and Schwede T. (2006). The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. Bioinformatics, 22,195-201.
- Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T (2009). The SWISS-MODEL Repository and associated resources. Nucleic Acids Research. 37, D387-D392.
- Peitsch, M. C. (1995) Protein modeling by E-mail Bio/Technology 13: 658-660.

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

3D-Jigsaw



3D-JIGSAW Protein Comparative Modelling Server
http://bmm.cancerresearchuk.org/~3djigsaw/
Google
EBI SSM Home Page
EMoT: Encycl...lar Targets
ISI Web of K...ledge [v3.0]
Logout/Session
Apple (102)
Amazon
eBay
Yahoo!
News (946)
3D-JIGSAW Protein Compar...

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

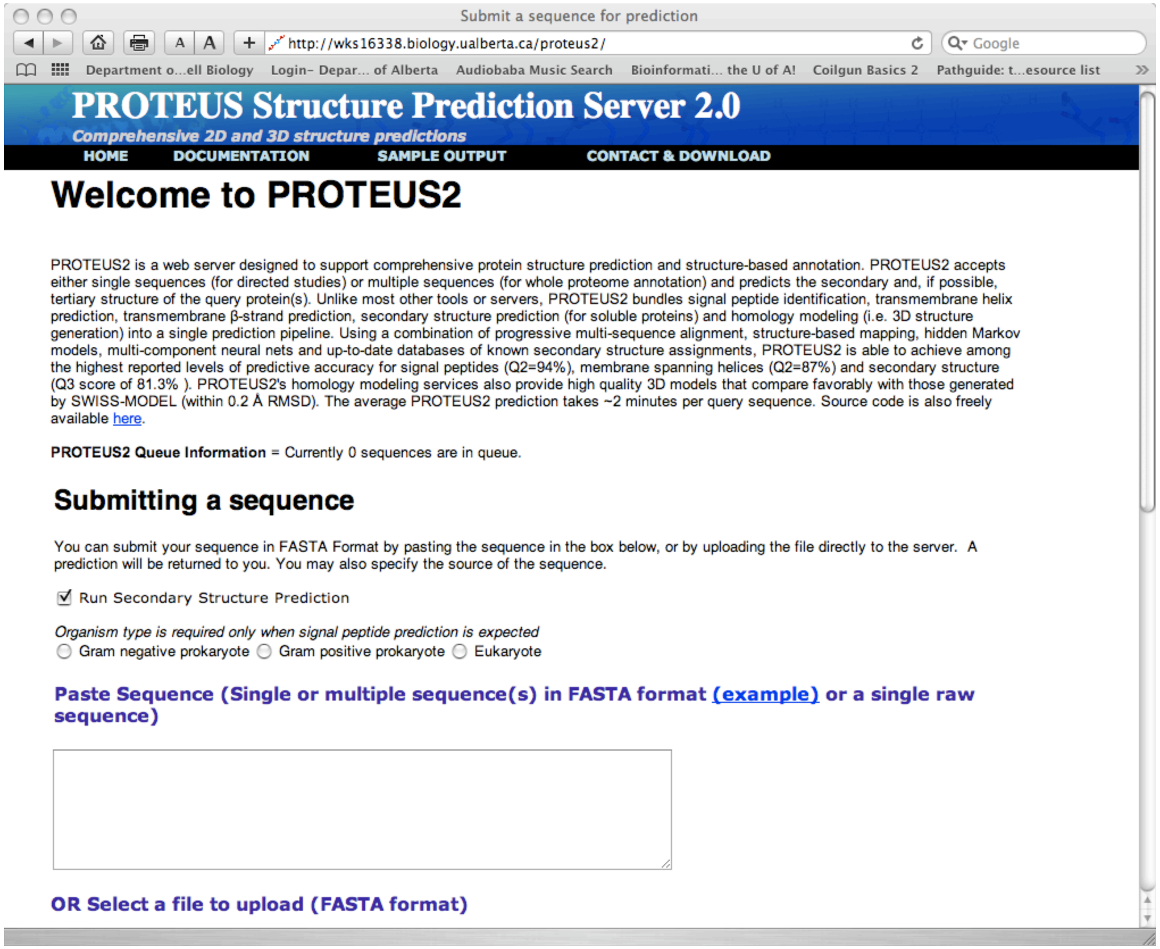
Protein **amino acid** sequence in one letter code

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!

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

Proteus2*


Submit a sequence for prediction
http://wks16338.biology.ualberta.ca/proteus2/
Google
Department o...ell Biology
Login– Depar... of Alberta
Audiobaba Music Search
Bioinformati... the U of A!
Coilgun Basics 2
Pathguide: t...esource list
PROTEUS Structure Prediction Server 2.0
Comprehensive 2D and 3D structure predictions
HOME
DOCUMENTATION
SAMPLE OUTPUT
CONTACT & DOWNLOAD

Welcome to PROTEUS2

PROTEUS2 is a web server designed to support comprehensive protein structure prediction and structure-based annotation. PROTEUS2 accepts either single sequences (for directed studies) or multiple sequences (for whole proteome annotation) and predicts the secondary and, if possible, tertiary structure of the query protein(s). Unlike most other tools or servers, PROTEUS2 bundles signal peptide identification, transmembrane helix prediction, transmembrane β-strand prediction, secondary structure prediction (for soluble proteins) and homology modeling (i.e. 3D structure generation) into a single prediction pipeline. Using a combination of progressive multi-sequence alignment, structure-based mapping, hidden Markov models, multi-component neural nets and up-to-date databases of known secondary structure assignments, PROTEUS2 is able to achieve among the highest reported levels of predictive accuracy for signal peptides (Q2=94%), membrane spanning helices (Q2=87%) and secondary structure (Q3 score of 81.3%). PROTEUS2's homology modeling services also provide high quality 3D models that compare favorably with those generated by SWISS-MODEL (within 0.2 Å RMSD). The average PROTEUS2 prediction takes ~2 minutes per query sequence. Source code is also freely available here.

PROTEUS2 Queue Information = Currently 0 sequences are in queue.

Submitting a sequence

You can submit your sequence in FASTA Format by pasting the sequence in the box below, or by uploading the file directly to the server. A prediction will be returned to you. You may also specify the source of the sequence.

☑ Run Secondary Structure Prediction

Organism type is required only when signal peptide prediction is expected
○ Gram negative prokaryote ○ Gram positive prokaryote ○ Eukaryote

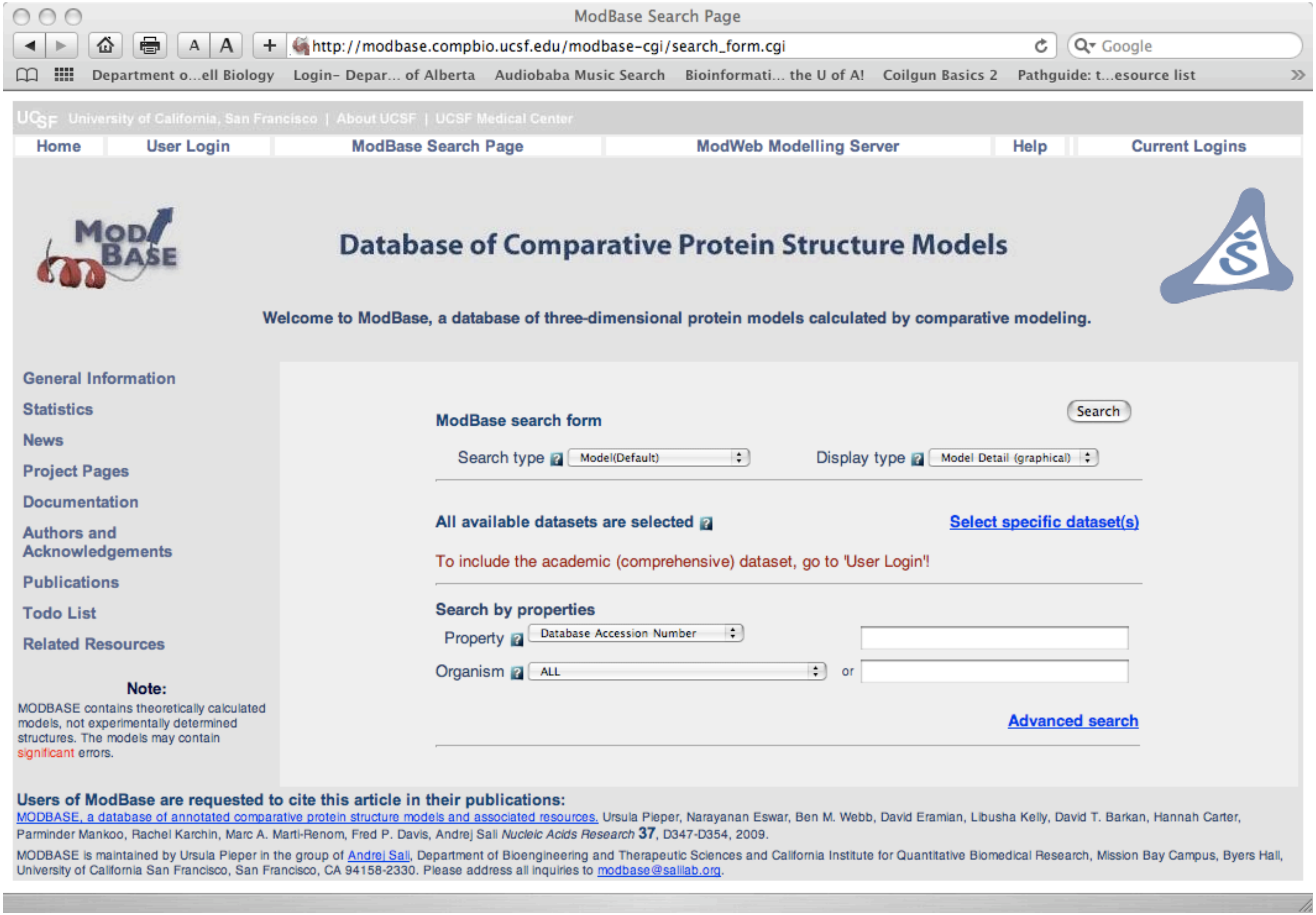
Paste Sequence (Single or multiple sequence(s) in FASTA format (example) or a single raw sequence)

OR Select a file to upload (FASTA format)

http://www.proteus2.ca/proteus2/

Modelled Protein Databases

- Databases containing 3D structural models of 100,000’s of proteins and protein domains
- Idea is to generate a 3D equivalent of GenBank (saves on everyone having to model everytime they want to look at a structure)
- Helps in Proteomics Target Selection


ModBase Search Page
http://modbase.compbio.ucsf.edu/modbase-cgi/search_form.cgi
Google
Department o...ell Biology
Login- Depar... of Alberta
Audiobaba Music Search
Bioinformati... the U of A!
Coilgun Basics 2
Pathguide: t...esource list
UCSF University of California, San Francisco | About UCSF | UCSF Medical Center
Home
User Login
ModBase Search Page
ModWeb Modelling Server
Help
Current Logins
MODBASE
General Information
Statistics
News
Project Pages
Documentation
Authors and Acknowledgements
Publications
Todo List
Related Resources
Search
Select specific dataset(s)
Advanced search

Database of Comparative Protein Structure Models

Welcome to ModBase, a database of three-dimensional protein models calculated by comparative modeling.

ModBase search form

Search type: Model(Default)

Display type: Model Detail (graphical)

All available datasets are selected

To include the academic (comprehensive) dataset, go to 'User Login'!

Search by properties

Property: Database Accession Number

Organism: ALL or

Note:

MODBASE contains theoretically calculated models, not experimentally determined structures. The models may contain significant errors.

Users of ModBase are requested to cite this article in their publications:

MODBASE, a database of annotated comparative protein structure models and associated resources. Ursula Pieper, Narayanan Eswar, Ben M. Webb, David Eramian, Libusha Kelly, David T. Barkan, Hannah Carter, Parminder Mankoo, Rachel Karchin, Marc A. Marti-Renom, Fred P. Davis, Andrej Sali *Nucleic Acids Research* **37**, D347-D354, 2009.

MODBASE is maintained by Ursula Pieper in the group of Andrej Sali, Department of Bioengineering and Therapeutic Sciences and California Institute for Quantitative Biomedical Research, Mission Bay Campus, Byers Hall, University of California San Francisco, San Francisco, CA 94158-2330. Please address all inquiries to modbase@salilab.org.

Outline

- The Protein Universe and the Protein Structure Initiative
- Homology (Comparative) Modelling of 3D Protein Structures
- Homology Modelling on the Web
- Assessing 3D Structures (modelled and experimental)

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 or rmsd)
- Assess correctness of overall fold (look at disposition of hydrophobes)
- Assess structure quality (packing, stereochemistry, bad contacts, etc.)

A Good Protein Structure..*

X-ray structure

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

NMR structure

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

A Good Protein Structure..*

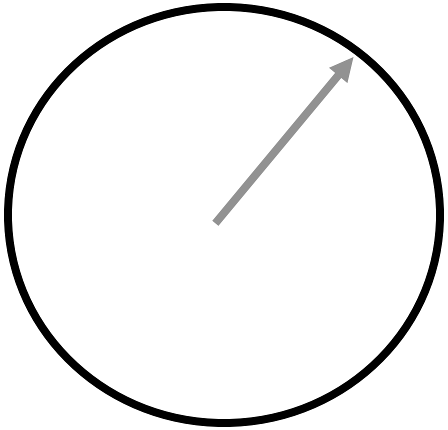
- Minimizes disallowed torsion angles
- Maximizes number of hydrogen bonds
- Maximizes buried hydrophobic ASA
- Maximizes exposed hydrophilic ASA
- Minimizes interstitial cavities or spaces

A Good Protein Structure..*

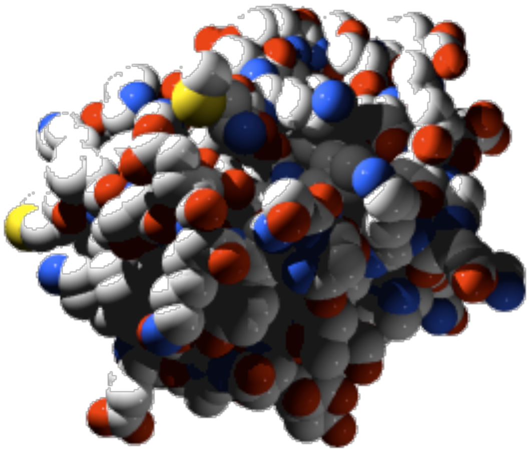
- Minimizes number of “bad” contacts
- Minimizes number of buried charges
- Minimizes radius of gyration
- Minimizes covalent and noncovalent (van der Waals and coulombic) energies

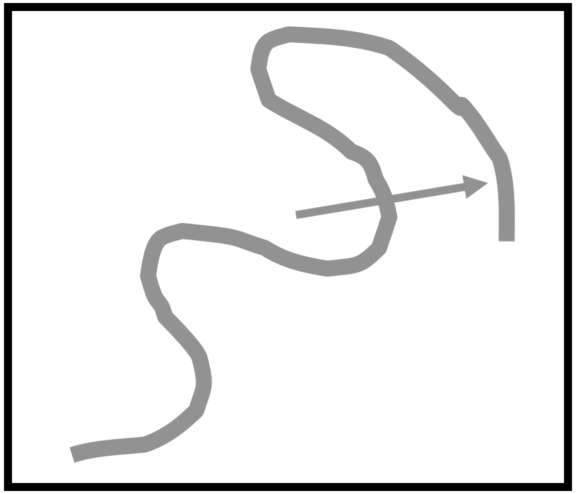
Radius & Radius of Gyration

- $\mathrm{RAD} = 3.95 \times \mathrm{NUMRES}^{0.6} + 7.25$ (Folded)
- $\mathrm{RADG} = 0.41 \times (110 \times \mathrm{NUMRES})^{0.5}$ (Unfolded)



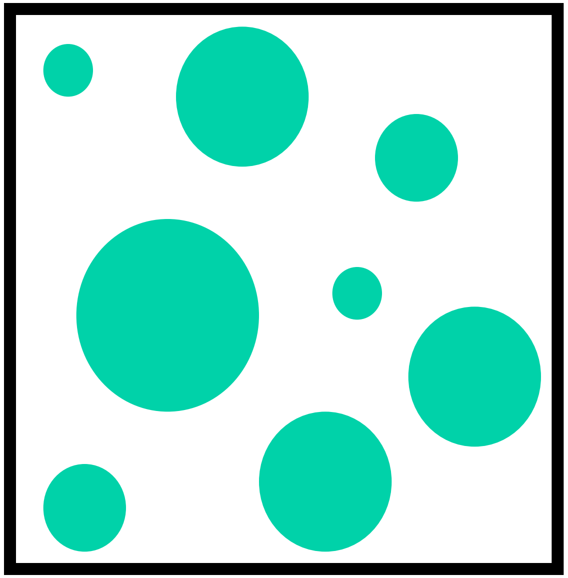
Radius



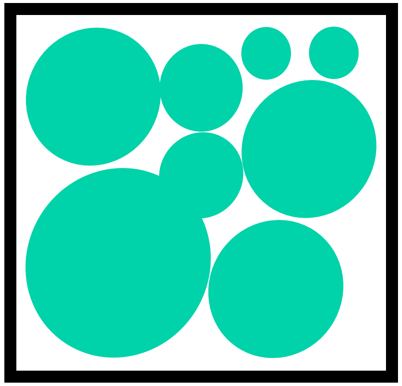


Radius of Gyration

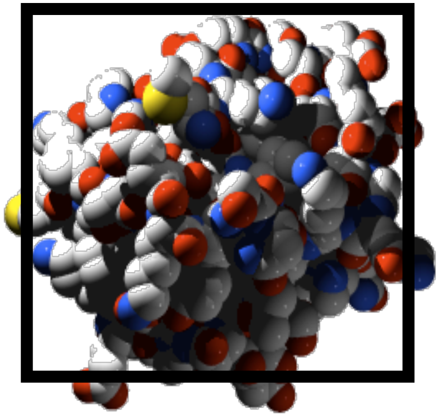
Packing Volume



Loose Packing



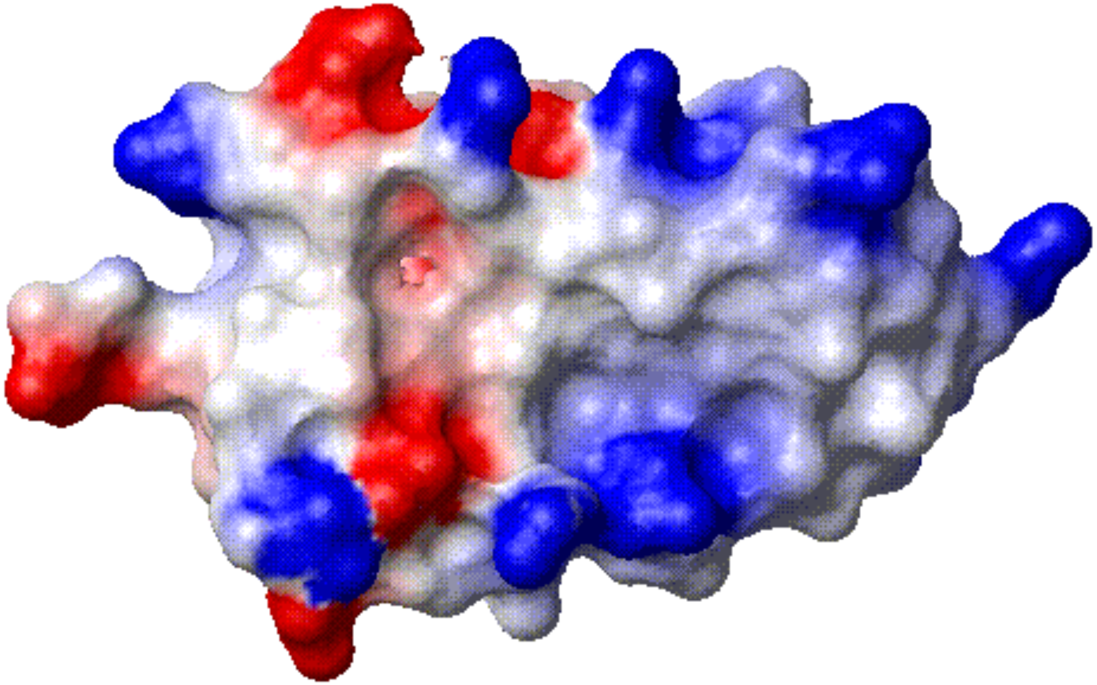
Dense Packing



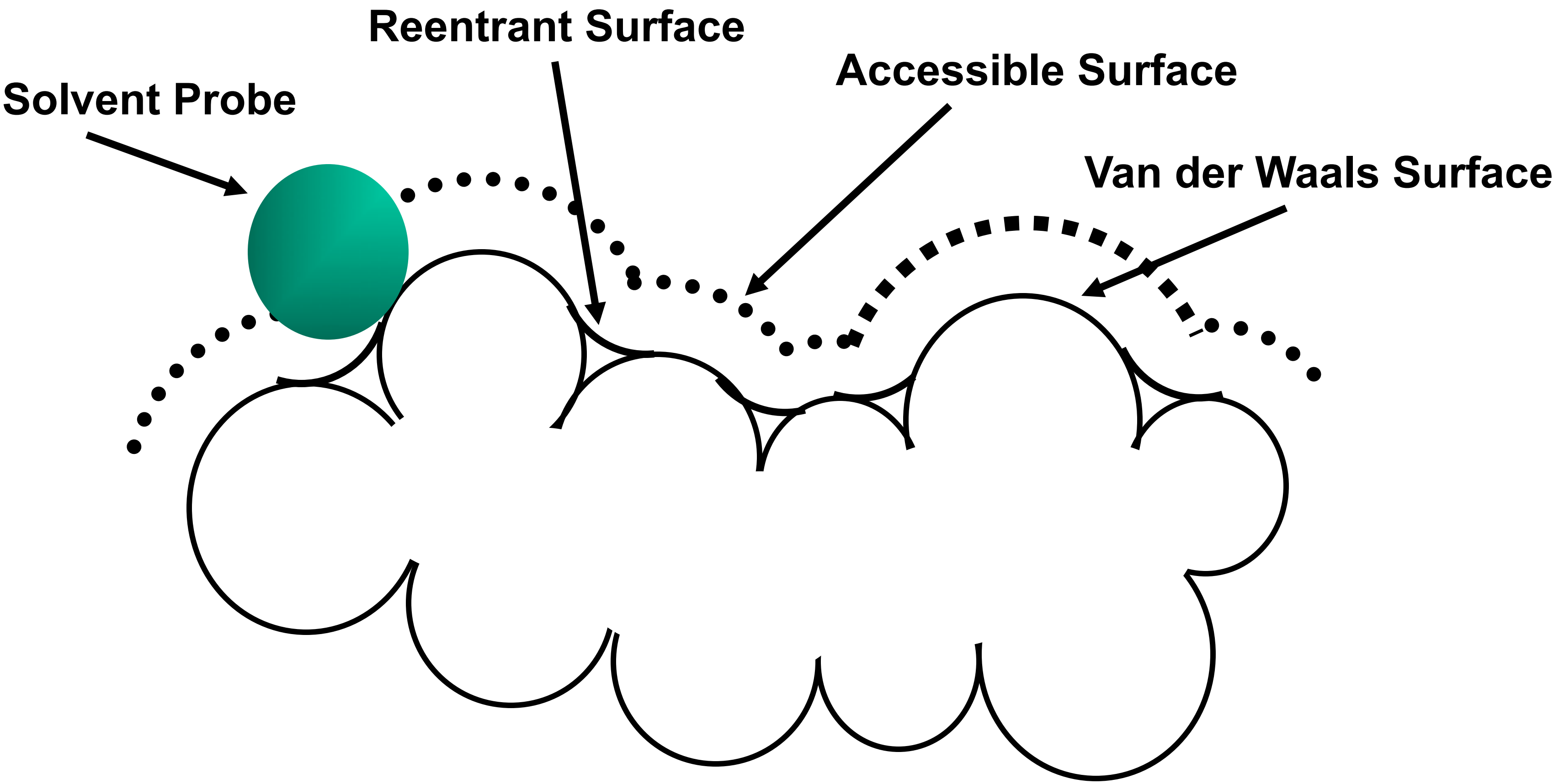
Protein

Proteins are Densely Packed

Accessible Surface Area



Accessible Surface Area*

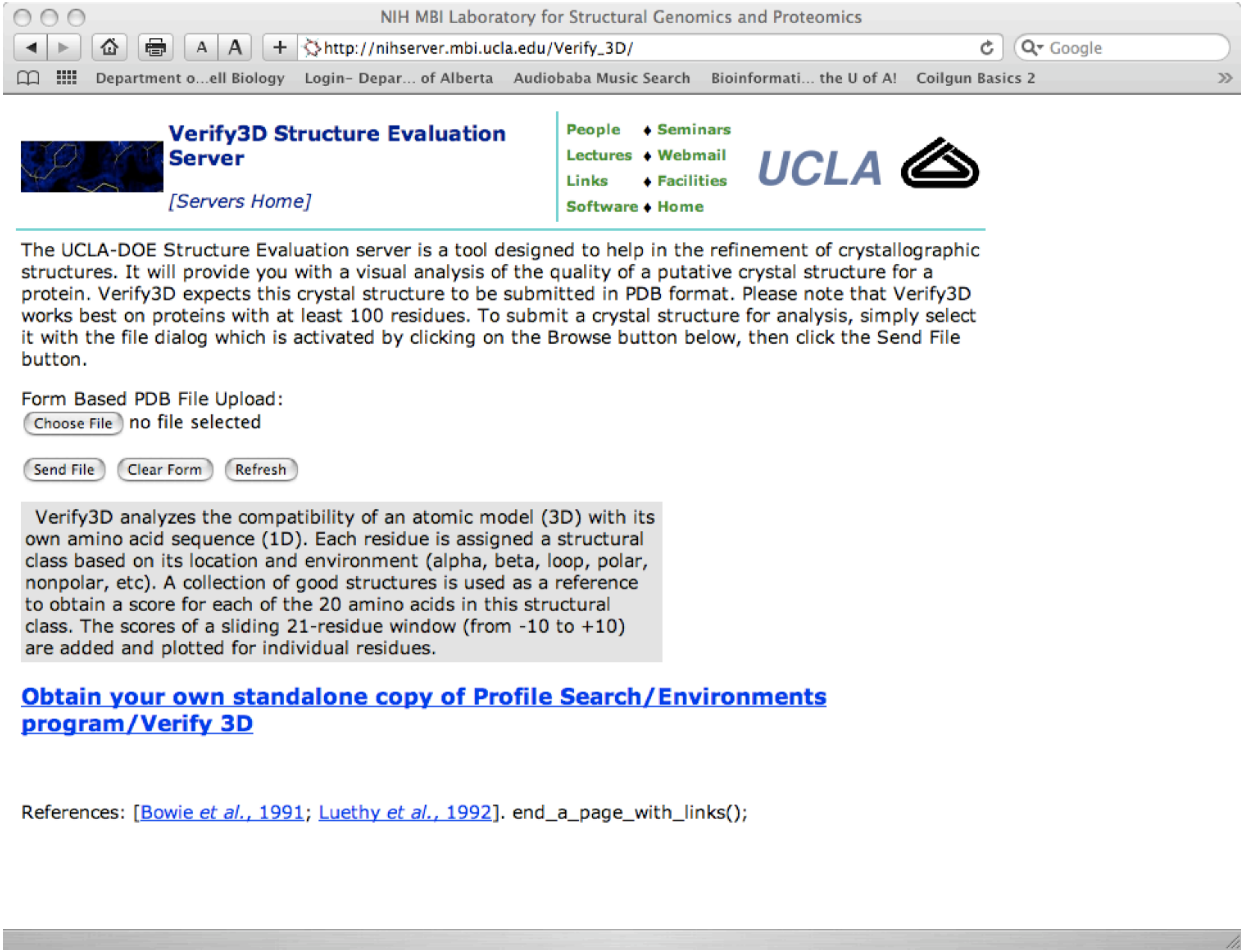


Accessible Surface Area*

- Solvation free energy is related to ASA
  - ♦ $\Delta G = \Sigma \Delta\sigma_i A_i$
- Proteins typically have 60% of their ASA comprised of polar atoms or residues
- Proteins typically have 40% of their ASA comprised of nonpolar atoms or residues
- ΔASA (obs - exp.) reveals shape/roughness

Structure Validation Servers

- **WhatIf Web Server -** http://swift.cmbi.ru.nl/servers/html/index.html
- **Protein Structure Validation Suite -** http://psvs-1_3.nesg.org/
- **Verify3D -** http://nihserver.mbi.ucla.edu/Verify_3D/
- **Molprobity -** http://molprobity.biochem.duke.edu/
- **PROSESS -** http://www.prosess.ca/
- **VADAR -** http://vadar.wishartlab.com/



NIH MBI Laboratory for Structural Genomics and Proteomics
http://nihserver.mbi.ucla.edu/Verify_3D/
Google
Department o...ell Biology
Login- Depar... of Alberta
Audiobaba Music Search
Bioinformati... the U of A!
Coilgun Basics 2

Verify3D Structure Evaluation Server

[Servers Home]

People ♦ Seminars
Lectures ♦ Webmail
Links ♦ Facilities
Software ♦ Home

UCLA

The UCLA-DOE Structure Evaluation server is a tool designed to help in the refinement of crystallographic structures. It will provide you with a visual analysis of the quality of a putative crystal structure for a protein. Verify3D expects this crystal structure to be submitted in PDB format. Please note that Verify3D works best on proteins with at least 100 residues. To submit a crystal structure for analysis, simply select it with the file dialog which is activated by clicking on the Browse button below, then click the Send File button.

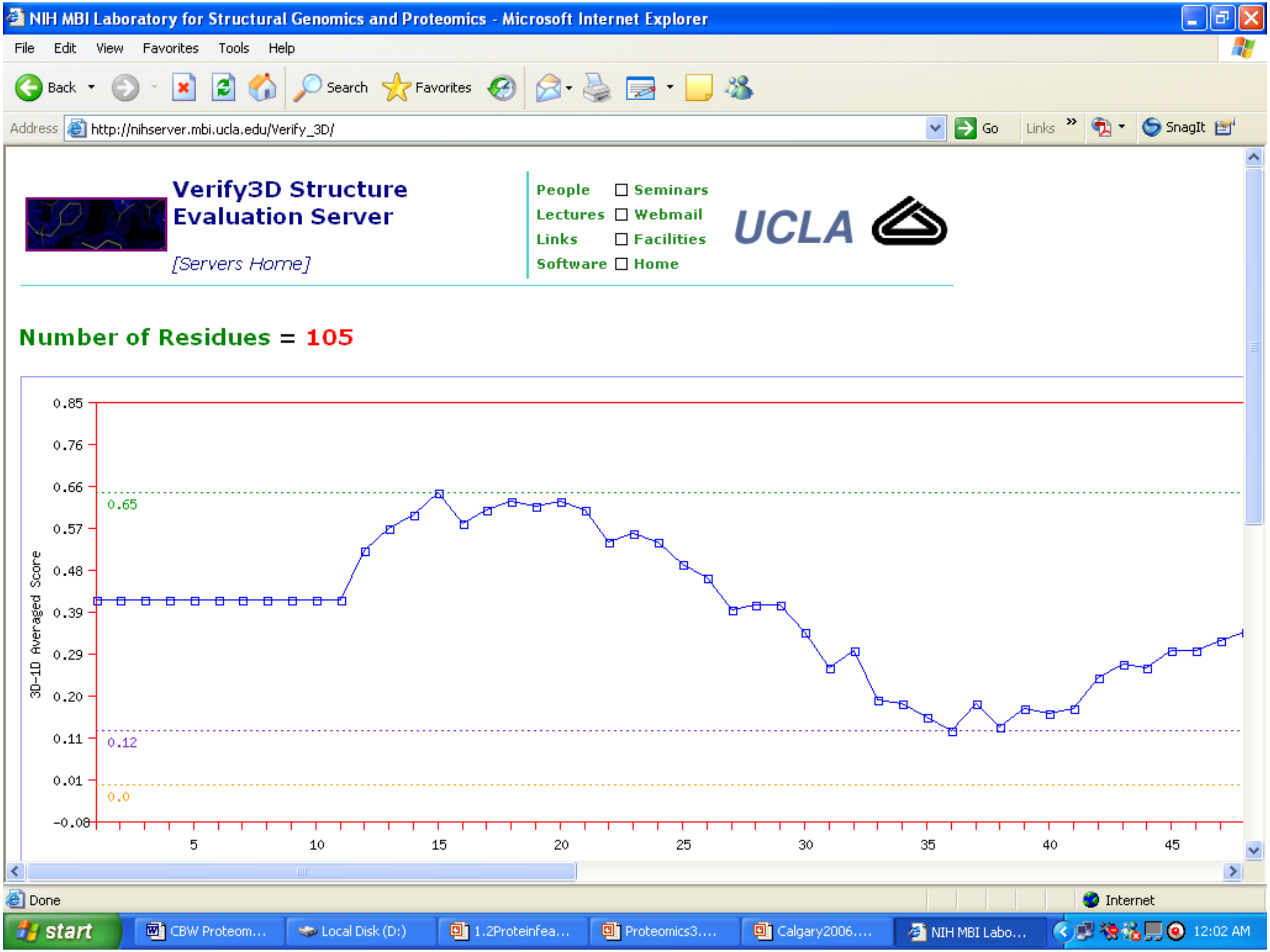
Form Based PDB File Upload:
Choose File no file selected

Send File Clear Form Refresh

Verify3D analyzes the compatibility of an atomic model (3D) with its own amino acid sequence (1D). Each residue is assigned a structural class based on its location and environment (alpha, beta, loop, polar, nonpolar, etc). A collection of good structures is used as a reference to obtain a score for each of the 20 amino acids in this structural class. The scores of a sliding 21-residue window (from -10 to +10) are added and plotted for individual residues.

Obtain your own standalone copy of Profile Search/Environments program/Verify 3D

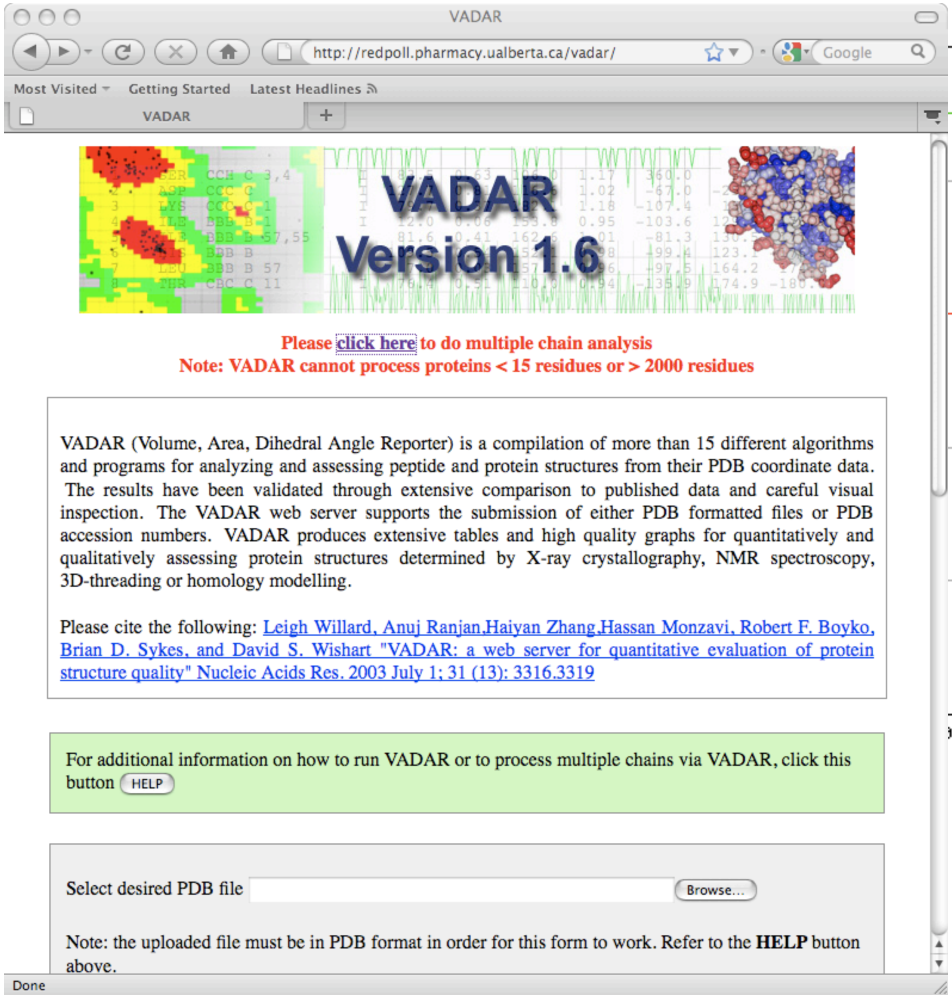
References: [Bowie *et al.*, 1991; Luethy *et al.*, 1992]. end_a_page_with_links();



Number of Residues = 105

High scores = good   Low scores = bad

VADAR*



Please click here to do multiple chain analysis
Note: VADAR cannot process proteins < 15 residues or > 2000 residues

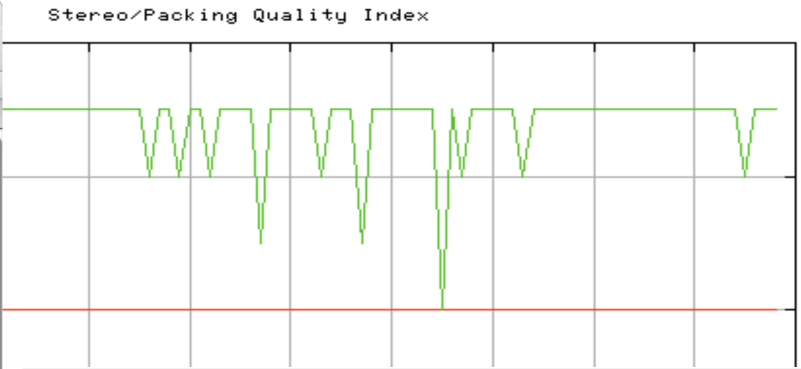
VADAR (Volume, Area, Dihedral Angle Reporter) is a compilation of more than 15 different algorithms and programs for analyzing and assessing peptide and protein structures from their PDB coordinate data. The results have been validated through extensive comparison to published data and careful visual 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.

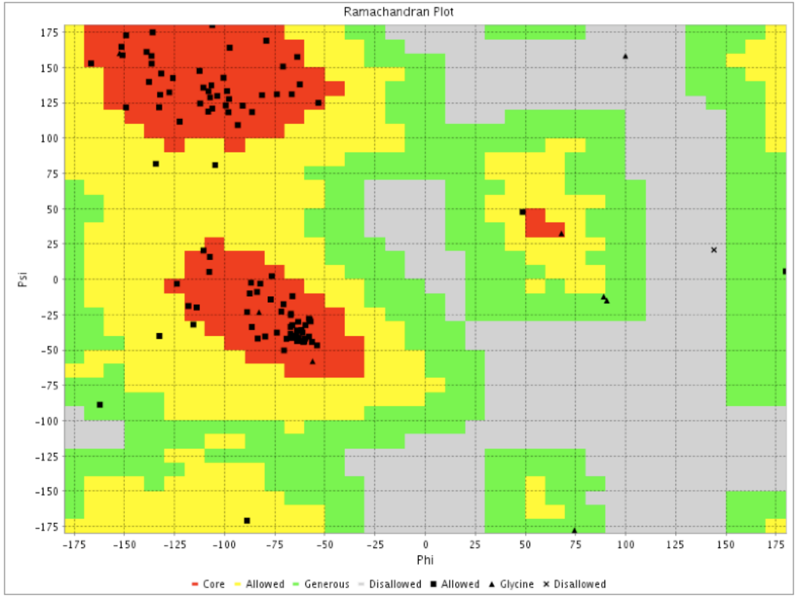
Please cite the following: Leigh Willard, Anuj Ranjan,Haiyan Zhang,Hassan Monzavi, Robert F. Boyko, Brian D. Sykes, and David S. Wishart "VADAR: a web server for quantitative evaluation of protein structure quality" Nucleic Acids Res. 2003 July 1; 31 (13): 3316.3319

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

Select desired PDB file Browse...

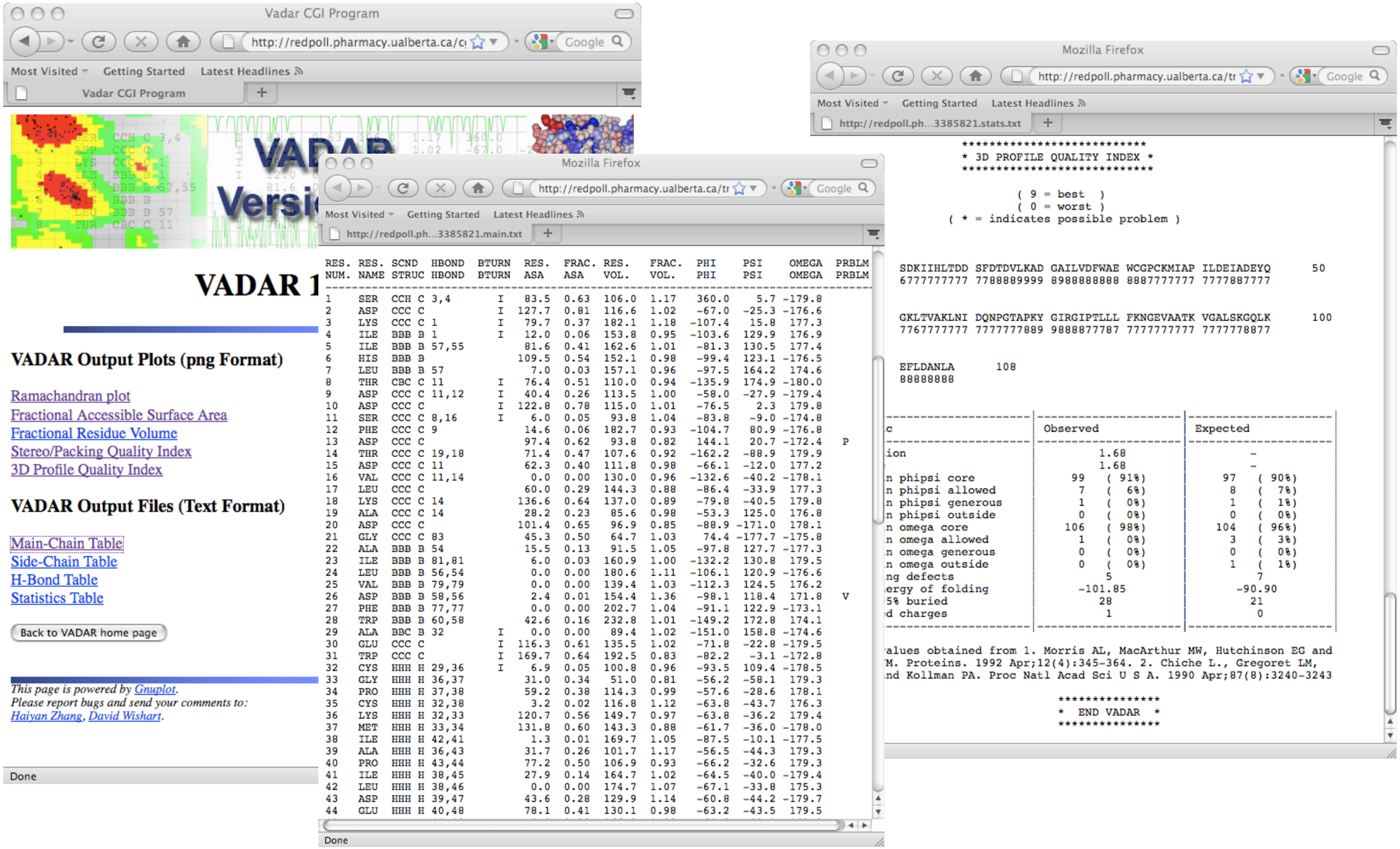
Note: the uploaded file must be in PDB format in order for this form to work. Refer to the HELP button above.





http://vadar.wishartlab.com/

VADAR


Vadar CGI Program
http://redpoll.pharmacy.ualberta.ca/c
Google
Most Visited
Getting Started
Latest Headlines
Vadar CGI Program
VADAR
Versi
Mozilla Firefox
http://redpoll.pharmacy.ualberta.ca/tr
http://redpoll.ph...3385821.main.txt
http://redpoll.ph...3385821.stats.txt
Done

VADAR 1

VADAR Output Plots (png Format)

- Ramachandran plot
- Fractional Accessible Surface Area
- Fractional Residue Volume
- Stereo/Packing Quality Index
- 3D Profile Quality Index

VADAR Output Files (Text Format)

- Main-Chain Table
- Side-Chain Table
- H-Bond Table
- Statistics Table

Back to VADAR home page

This page is powered by Gnuplot.
Please report bugs and send your comments to:
Haiyan Zhang, David Wishart.

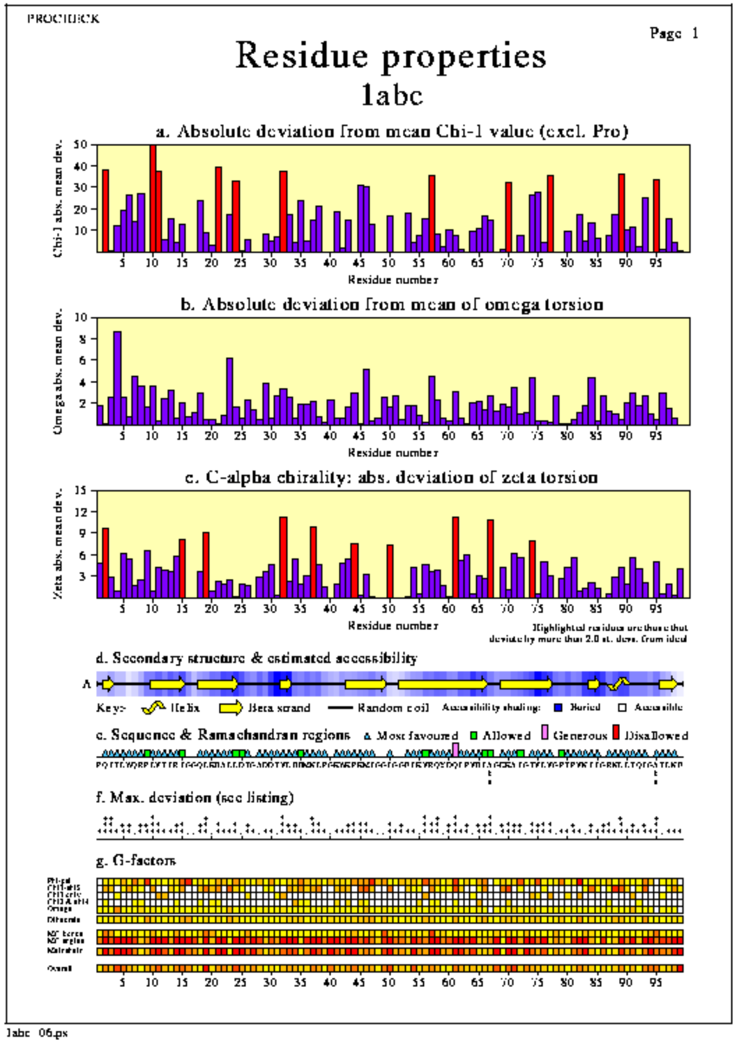
```
RES. RES. SCND  HBOND  BTURN  RES.  FRAC. RES.   FRAC.  PHI    PSI    OMEGA  PRBLM
NUM. NAME STRUC HBOND  BTURN  ASA   ASA   VOL.   VOL.   PHI    PSI    OMEGA  PRBLM
-----------------------------------------------------------------------------------
1    SER  CCH C 3,4      I   83.5  0.63  106.0  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
20   ASP  CCC C             101.4  0.65   96.9  0.85   -88.9 -171.0  178.1
21   GLY  CCC C 83           45.3  0.50   64.7  1.03    74.4 -177.7 -175.8
22   ALA  BBB B 54           15.5  0.13   91.5  1.05   -97.8  127.7 -177.3
23   ILE  BBB B 81,81         6.0  0.03  160.9  1.00  -132.2  130.8  179.5
24   LEU  BBB B 56,54         0.0  0.00  180.6  1.11  -106.1  120.9 -176.6
25   VAL  BBB B 79,79         0.0  0.00  139.4  1.03  -112.3  124.5  176.2
26   ASP  BBB B 58,56         2.4  0.01  154.4  1.36   -98.1  118.4  171.8   V
27   PHE  BBB B 77,77         0.0  0.00  202.7  1.04   -91.1  122.9 -173.1
28   TRP  BBB B 60,58        42.6  0.16  232.8  1.01  -149.2  172.8  174.1
29   ALA  BBC B 32       I    0.0  0.00   89.4  1.02  -151.0  158.8 -174.6
30   GLU  CCC C          I  116.3  0.61  135.5  1.02   -71.8  -22.8 -179.5
31   TRP  CCC C          I  169.7  0.64  192.5  0.83   -82.2   -3.1 -172.8
32   CYS  HHH H 29,36    I    6.9  0.05  100.8  0.96   -93.5  109.4 -178.5
33   GLY  HHH H 36,37        31.0  0.34   51.0  0.81   -56.2  -58.1  179.3
34   PRO  HHH H 37,38        59.2  0.38  114.3  0.99   -57.6  -28.6  178.1
35   CYS  HHH H 32,38         3.2  0.02  116.8  1.12   -63.8  -43.7  176.3
36   LYS  HHH H 32,33       120.7  0.56  149.7  0.97   -63.8  -36.2  179.4
37   MET  HHH H 33,34       131.8  0.60  143.3  0.88   -61.7  -36.0 -178.0
38   ILE  HHH H 42,41         1.3  0.01  169.7  1.05   -87.5  -10.1 -177.5
39   ALA  HHH H 36,43        31.7  0.26  101.7  1.17   -56.5  -44.3  179.3
40   PRO  HHH H 43,44        77.2  0.50  106.9  0.93   -66.2  -32.6  179.3
41   ILE  HHH H 38,45        27.9  0.14  164.7  1.02   -64.5  -40.0 -179.4
42   LEU  HHH H 38,46         0.0  0.00  174.7  1.07   -67.1  -33.8  175.3
43   ASP  HHH H 39,47        43.6  0.28  129.9  1.14   -60.8  -44.2 -179.7
44   GLU  HHH H 40,48        78.1  0.41  130.1  0.98   -63.2  -43.5  179.5
```

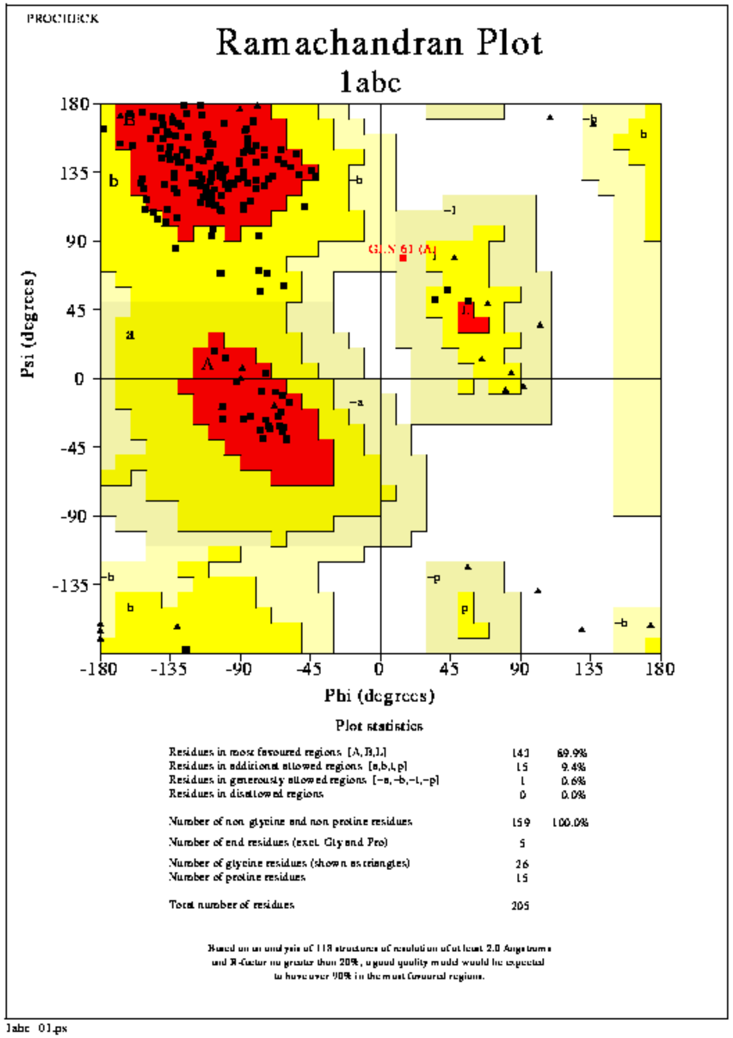
```
           ***************************
           * 3D PROFILE QUALITY INDEX *
           ***************************

                 ( 9 = best  )
                 ( 0 = worst )
          ( * = indicates possible problem )


 SDKIIHLTDD SFDTDVLKAD GAILVDFWAE WCGPCKMIAP ILDEIADEYQ      50
 6777777777 7788889999 8988888888 8887777777 7777887777


 GKLTVAKLNI DQNPGTAPKY GIRGIPTLLL FKNGEVAATK VGALSKGQLK      100
 7767777777 7777777889 9888877787 7777777777 7777778877


 EFLDANLA      108
 88888888


---------------------|--------------------|--------------------|
c                    | Observed           | Expected           |
---------------------|--------------------|--------------------|
ion                  |       1.68         |        -           |
                     |       1.68         |        -           |
n phipsi core        |    99   ( 91%)     |    97   ( 90%)     |
n phipsi allowed     |     7   (  6%)     |     8   (  7%)     |
n phipsi generous    |     1   (  0%)     |     1   (  1%)     |
n phipsi outside     |     0   (  0%)     |     0   (  0%)     |
n omega core         |   106   ( 98%)     |   104   ( 96%)     |
n omega allowed      |     1   (  0%)     |     3   (  3%)     |
n omega generous     |     0   (  0%)     |     0   (  0%)     |
n omega outside      |     0   (  0%)     |     1   (  1%)     |
ng defects           |       5            |       7            |
ergy of folding      |   -101.85          |    -90.90          |
5% buried            |      28            |      21            |
d charges            |       1            |       0            |
---------------------|--------------------|--------------------|

alues obtained from 1. Morris AL, MacArthur MW, Hutchinson EG and
M. Proteins. 1992 Apr;12(4):345-364. 2. Chiche L., Gregoret LM,
nd Kollman PA. Proc Natl Acad Sci U S A. 1990 Apr;87(8):3240-3243

                     ***************
                     *  END VADAR  *
                     ***************
```

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
- **WhatCheck** - http://swift.cmbi.kun.nl/gv/whatcheck/
- **PDB Validation Suite** - http://sw-tools.pdb.org/apps/VAL/index.html
- **DSSP** - http://swift.cmbi.kun.nl/gv/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.

Summary

- Homology modeling is the most accurate method known for predicting 3D protein structures
- Recent advances have made homology modeling trivial to do over the web
- There are many different ways of evaluating and validating the quality of 3D structure models
- *Homework: spend 15-20 minutes visiting the websites mentioned today*

How To Do Your Assignment

- Follow the instructions carefully
- Each of the programs or websites you need to use has been mentioned in the last 3 lectures, if you’ re smart you may only need to use 3 (local) tools
- This assignment will take 4-5 hours to complete and should be 6-8 pages long