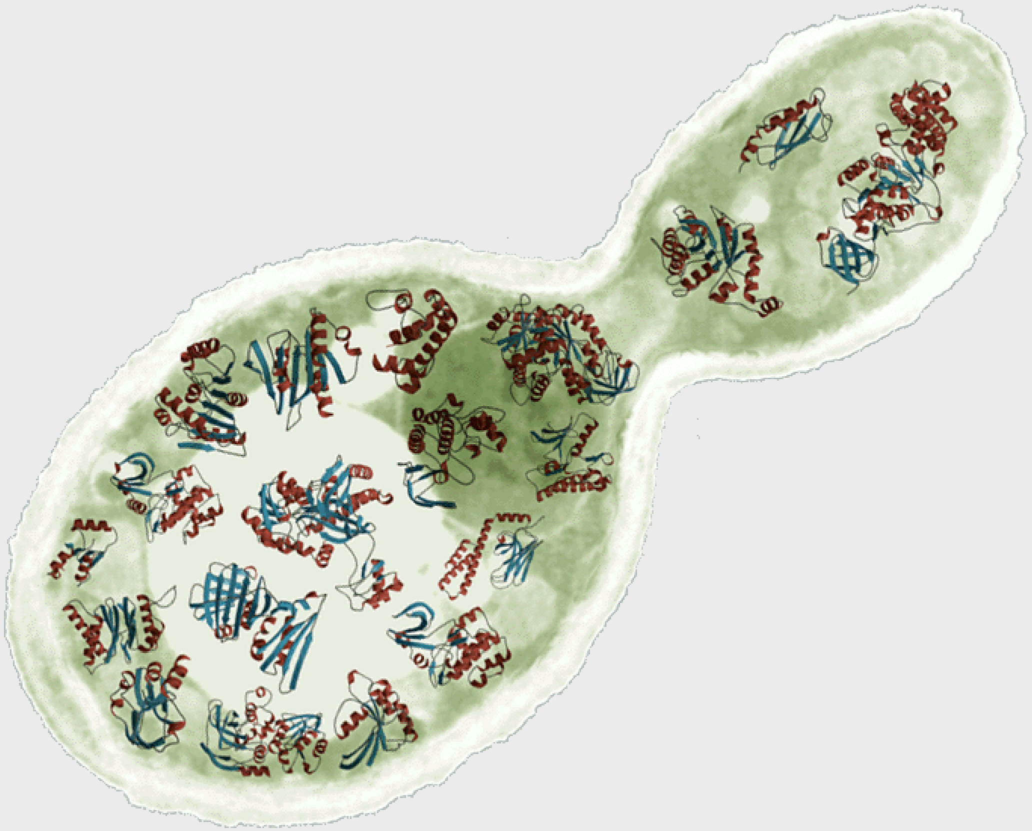
Master Bioinformatics for Health Sciences

Comparative Protein Structure Prediction



Marc A. Marti-Renom

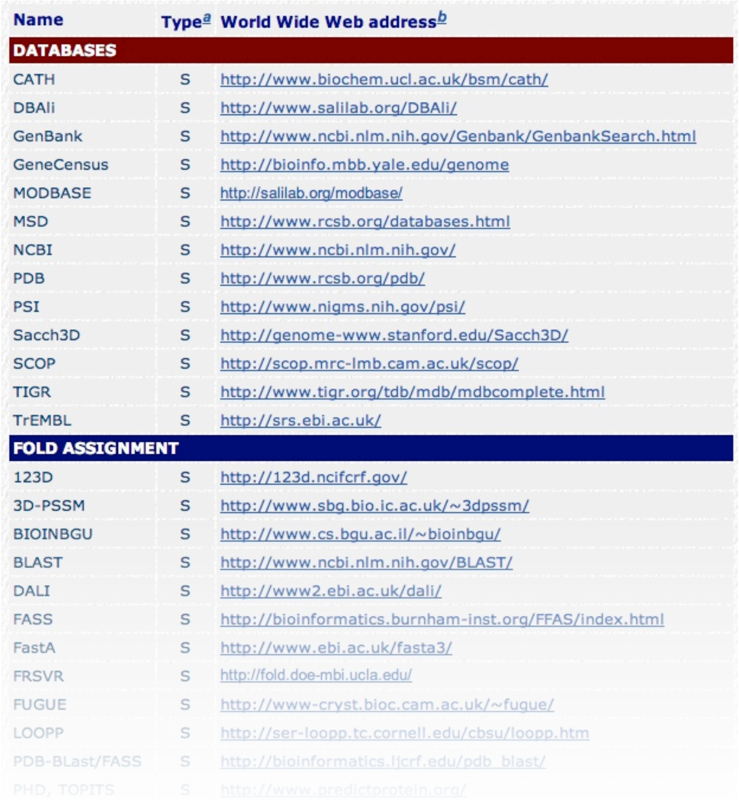
Adjunct Assistant Professor

http://salilab.org/~marcius

UCSF
University of California
San Francisco

Depts. of Biopharmaceutical Sciences and Pharmaceutical Chemistry
California Institute for Quantitative Biomedical Research
University of California at San Francisco

DISCLAIMER!



Name	Type[a]	World Wide Web address[b]
DATABASES		
CATH	S	http://www.biochem.ucl.ac.uk/bsm/cath/
DBAli	S	http://www.salilab.org/DBAli/
GenBank	S	http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html
GeneCensus	S	http://bioinfo.mbb.yale.edu/genome
MODBASE	S	http://salilab.org/modbase/
MSD	S	http://www.rcsb.org/databases.html
NCBI	S	http://www.ncbi.nlm.nih.gov/
PDB	S	http://www.rcsb.org/pdb/
PSI	S	http://www.nigms.nih.gov/psi/
Sacch3D	S	http://genome-www.stanford.edu/Sacch3D/
SCOP	S	http://scop.mrc-lmb.cam.ac.uk/scop/
TIGR	S	http://www.tigr.org/tdb/mdb/mdbcomplete.html
TrEMBL	S	http://srs.ebi.ac.uk/
FOLD ASSIGNMENT		
123D	S	http://123d.ncifcrf.gov/
3D-PSSM	S	http://www.sbg.bio.ic.ac.uk/~3dpssm/
BIOINBGU	S	http://www.cs.bgu.ac.il/~bioinbgu/
BLAST	S	http://www.ncbi.nlm.nih.gov/BLAST/
DALI	S	http://www2.ebi.ac.uk/dali/
FASS	S	http://bioinformatics.burnham-inst.org/FFAS/index.html
FastA	S	http://www.ebi.ac.uk/fasta3/
FRSVR	S	http://fold.doe-mbi.ucla.edu/
FUGUE	S	http://www-cryst.bioc.cam.ac.uk/~fugue/
LOOPP	S	http://ser-loopp.tc.cornell.edu/cbsu/loopp.htm
PDB-BLast/FASS	S	http://bioinformatics.ljcrf.edu/pdb_blast/
PHD, TOPITS	S	http://www.predictprotein.org/

http://salilab.org/bioinformatics_resources.shtml

Program

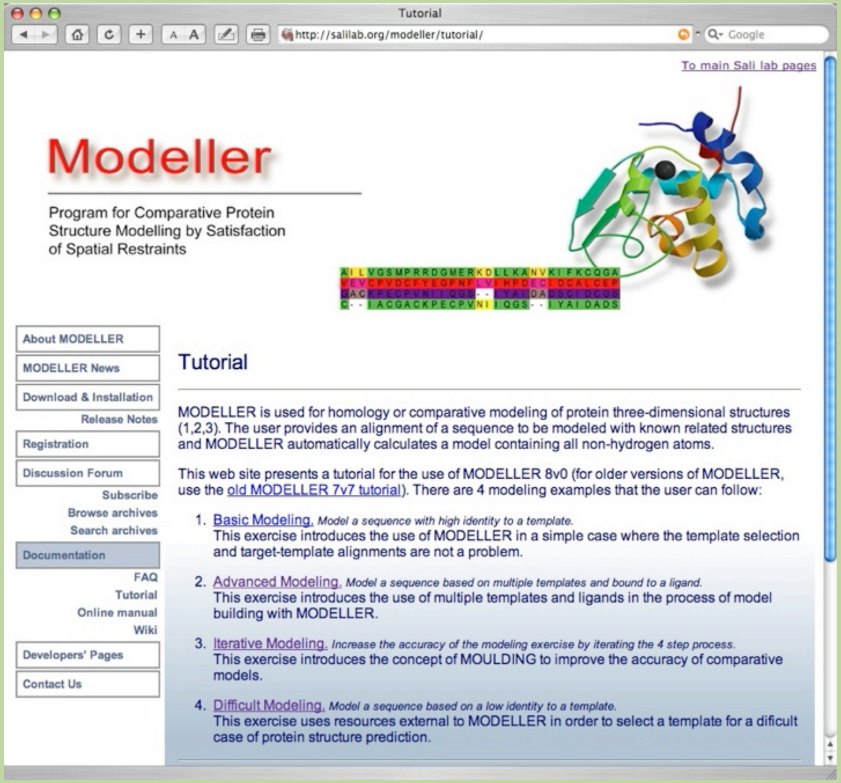
Intro to comparative protein structure prediction

Template Search*

Target – Template Alignment*

Model Building

Model Evaluation



Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints

Tutorial

MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1,2,3). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms.

This web site presents a tutorial for the use of MODELLER 8v0 (for older versions of MODELLER, use the old MODELLER 7v7 tutorial). There are 4 modeling examples that the user can follow:

1. Basic Modeling. *Model a sequence with high identity to a template.* This exercise introduces the use of MODELLER in a simple case where the template selection and target-template alignments are not a problem.
2. Advanced Modeling. *Model a sequence based on multiple templates and bound to a ligand.* This exercise introduces the use of multiple templates and ligands in the process of model building with MODELLER.
3. Iterative Modeling. *Increase the accuracy of the modeling exercise by iterating the 4 step process.* This exercise introduces the concept of MOULDING to improve the accuracy of comparative models.
4. Difficult Modeling. *Model a sequence based on a low identity to a template.* This exercise uses resources external to MODELLER in order to select a template for a dificult case of protein structure prediction.

http://www.salilab.org/modeller/tutotial/

Objective

TO LEARN HOW-TO MODEL A 3D-STRUCTURE FROM A SEQUENCE AND A KNOWN STRUCTURE

What are we going to do?

Ask!

- Each day…
  - Basic introduction
  - Theory (representation-scoring-optimization)
  - Available programs
  - Application

Nomenclature

Homology: Sharing a common ancestor, may have similar or dissimilar functions

Similarity: Score that quantifies the degree of relationship between two sequences.

Identity: Fraction of identical aminoacids between two aligned sequences (case of similarity).

Target: Sequence corresponding to the protein to be modeled.

Template: 3D structure/s to be used during protein structure prediction.

Model: Predicted 3D structure of the target sequence.

protein prediction .vs. protein determination

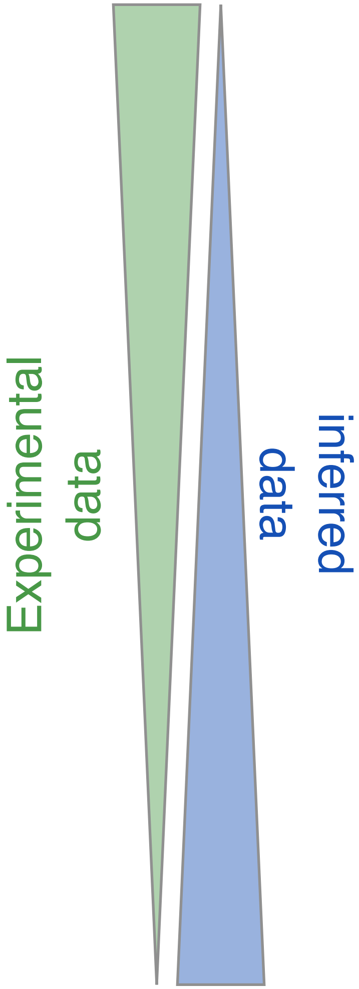
X-Ray

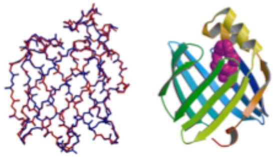
NMR

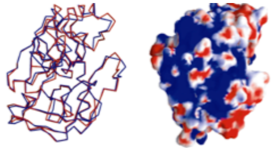
Comparative Modeling

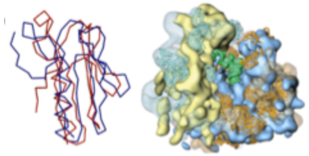
Threading

Ab-initio









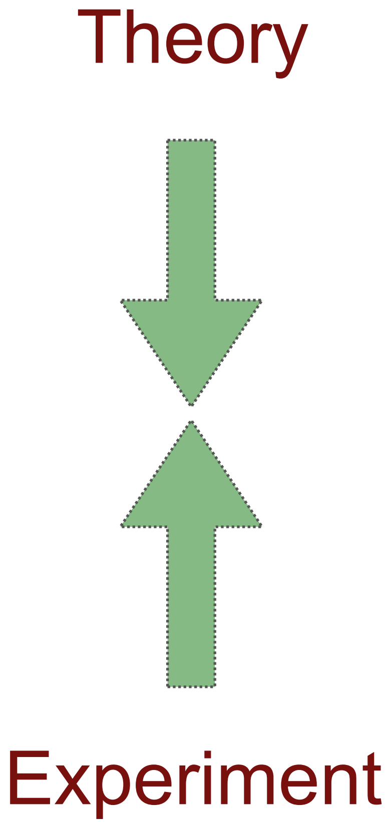
Why protein structure prediction?

	Y 2005	Y 2006
Sequences	1,700,000	millions
Structures	28,000	50,000

Why protein structure prediction?

	Y 2005
Sequences	1,700,000
Structures	**900,000**

http://salilab.org/modbase/

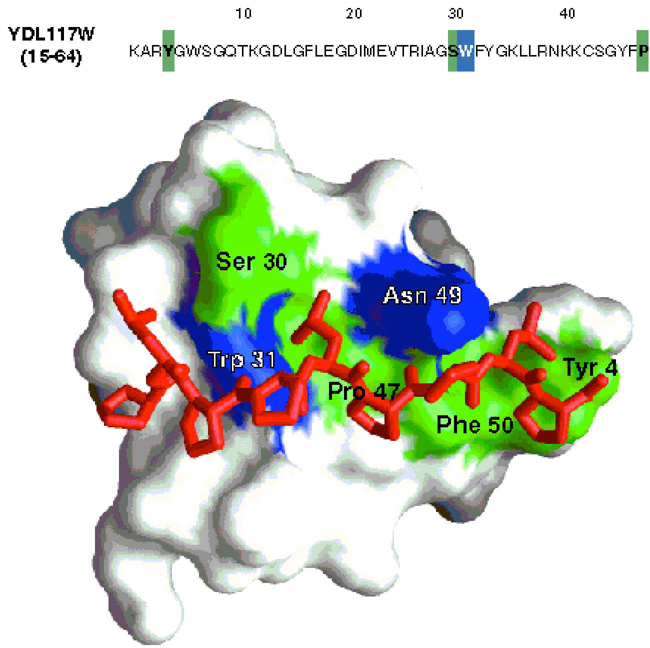


Theory

Experiment

Why is it useful to know the structure of a protein, not only its sequence?

- The biochemical function (activity) of a protein is defined by its interactions with other molecules.
- The biological function is in large part a consequence of these interactions.
- The 3D structure is more informative than sequence because interactions are determined by residues that are close in space but are frequently distant in sequence.

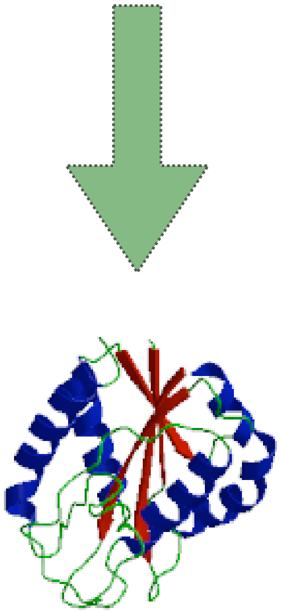


In addition, since evolution tends to conserve function and function depends more directly on structure than on sequence, **structure is more conserved in evolution than sequence**.

The net result is that **patterns in space are frequently more recognizable than patterns in sequence**.

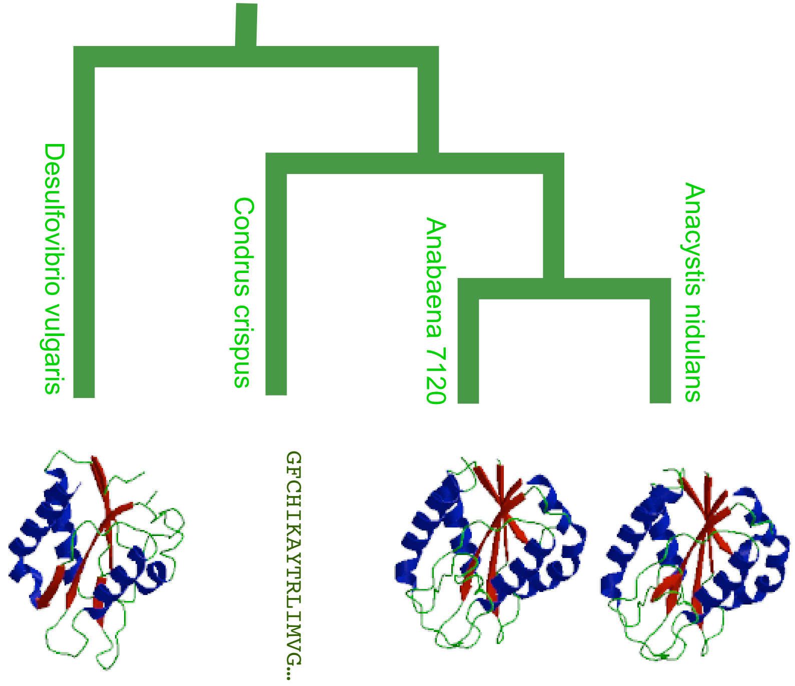
Principles of Protein Structure

GFCHIKAYTRLIMVG...



Folding

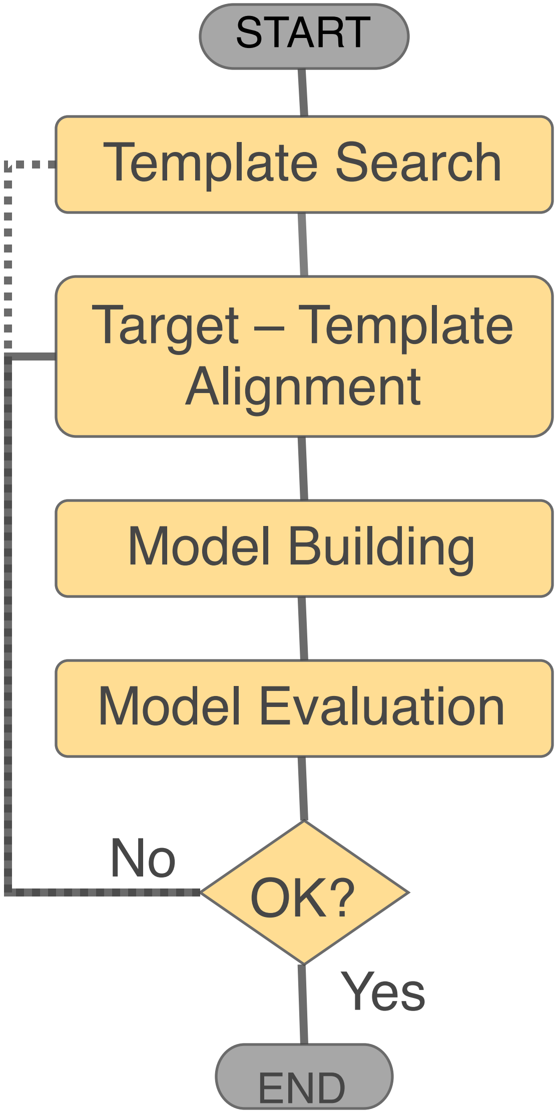
Ab initio prediction



Evolution

Threading
Comparative Modeling

Steps in Comparative Protein Structure Modeling

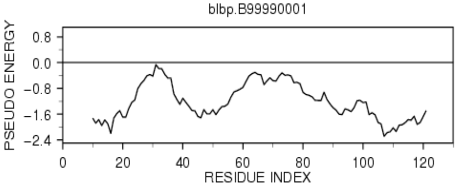


TARGET

ASILPKRLFGNCEQTSDEGL
KIERTPLVPHISAQNVCLKID
DVPERLIPERASFQWMNDK

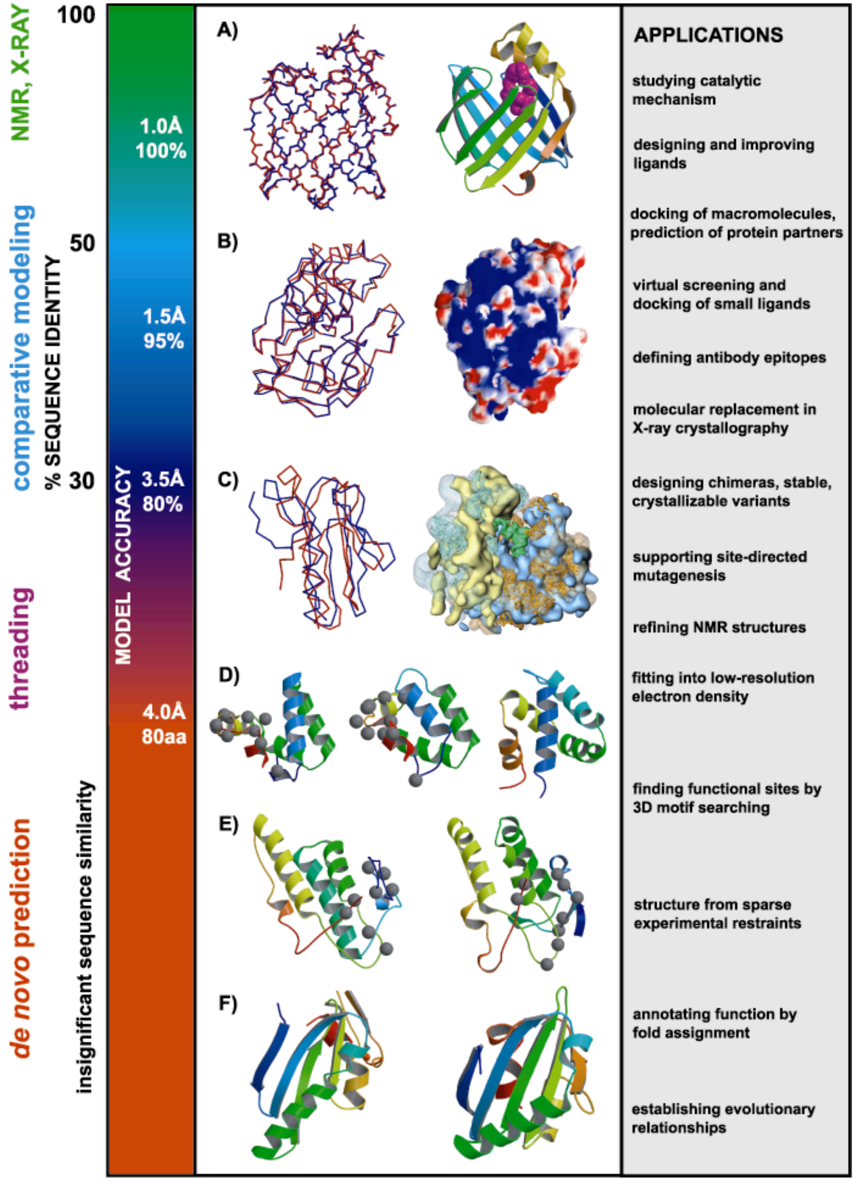
TEMPLATE

ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE



A. Šali, Curr. Opin. Biotech. 6, 437, 1995.
R. Sánchez & A. Šali, Curr. Opin. Str. Biol. 7, 206, 1997.
M. Marti et al. Ann. Rev. Biophys. Biomolec. Struct., 29, 291, 2000.

Utility of protein structure models, despite errors


NMR, X-RAY
comparative modeling
threading
de novo prediction
% SEQUENCE IDENTITY
100
50
30
insignificant sequence similarity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
4.0Å
80aa
A)
B)
C)
D)
E)
F)
APPLICATIONS
studying catalytic mechanism
designing and improving ligands
docking of macromolecules, prediction of protein partners
virtual screening and docking of small ligands
defining antibody epitopes
molecular replacement in X-ray crystallography
designing chimeras, stable, crystallizable variants
supporting site-directed mutagenesis
refining NMR structures
fitting into low-resolution electron density
finding functional sites by 3D motif searching
structure from sparse experimental restraints
annotating function by fold assignment
establishing evolutionary relationships

D. Baker & A. Sali. Science 294, 93, 2001.

General References

Protein Structure Prediction:

Marti-Renom el al. Annu. Rev. Biophys. Biomol. Struct. 29, 291-325, 2000.
Baker & Sali. Science 294, 93-96, 2001.

Comparative Modeling:

Marti-Renom el al. Annu. Rev. Biophys. Biomol. Struct. 29, 291-325, 2000.
Marti-Renom el al. Current Protocols in Protein Science 1, 2.9.1-2.9.22, 2002.

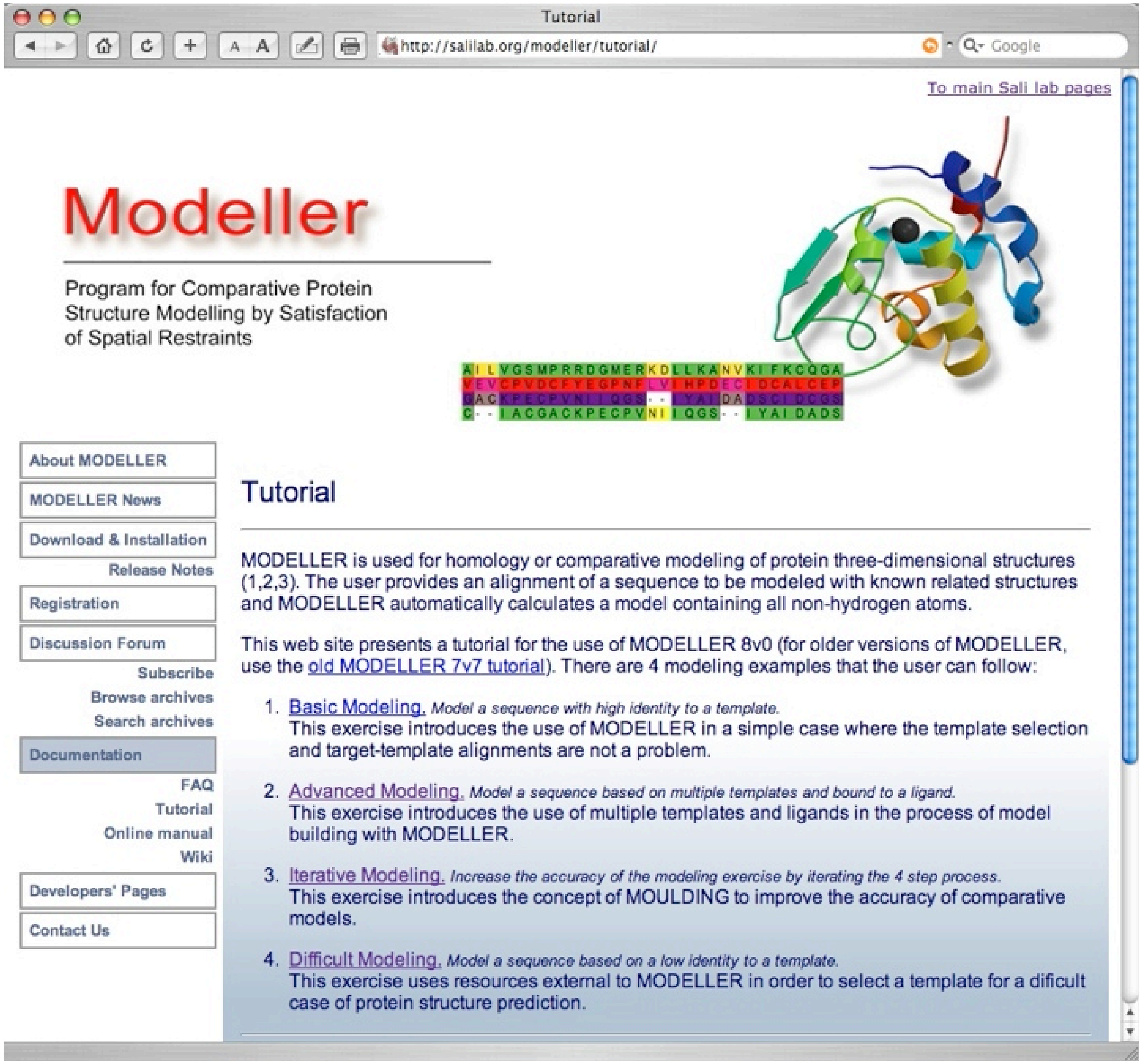
MODELLER:

Sali & Blundell. J. Mol. Biol. 234, 779-815, 1993.

Structural Genomics:

Sali. Nat. Struct. Biol. 5, 1029, 1998.
Burley et al. Nat. Genet. 23, 151, 1999.
Sali & Kuriyan. TIBS 22, M20, 1999.
Sanchez et al. Nat. Str. Biol. 7, 986, 2000.
Baker & Sali. Science 294, 93-96, 2001.
Vitkup et al. Nat. Struct. Biol. 8, 559, 2001.

http://www.salilab.org/modeller/tutorial/



To main Sali lab pages

Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints

Tutorial

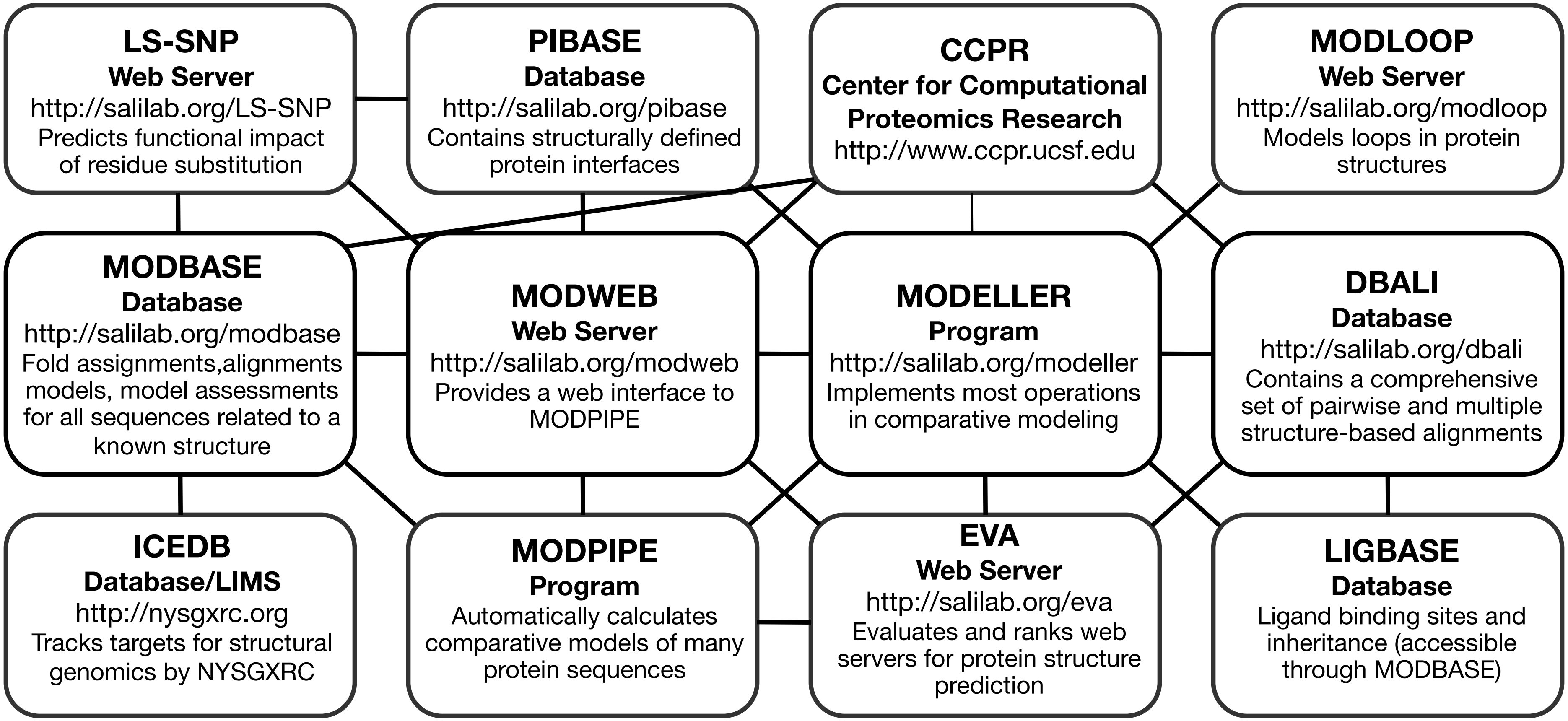
MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1,2,3). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms.

This web site presents a tutorial for the use of MODELLER 8v0 (for older versions of MODELLER, use the old MODELLER 7v7 tutorial). There are 4 modeling examples that the user can follow:

1. Basic Modeling. *Model a sequence with high identity to a template.* This exercise introduces the use of MODELLER in a simple case where the template selection and target-template alignments are not a problem.
2. Advanced Modeling. *Model a sequence based on multiple templates and bound to a ligand.* This exercise introduces the use of multiple templates and ligands in the process of model building with MODELLER.
3. Iterative Modeling. *Increase the accuracy of the modeling exercise by iterating the 4 step process.* This exercise introduces the concept of MOULDING to improve the accuracy of comparative models.
4. Difficult Modeling. *Model a sequence based on a low identity to a template.* This exercise uses resources external to MODELLER in order to select a template for a dificult case of protein structure prediction.

Programs, servers and databases

http://salilab.org



LS-SNP
Web Server
http://salilab.org/LS-SNP
Predicts functional impact of residue substitution

PIBASE
Database
http://salilab.org/pibase
Contains structurally defined protein interfaces

CCPR
Center for Computational Proteomics Research
http://www.ccpr.ucsf.edu

MODLOOP
Web Server
http://salilab.org/modloop
Models loops in protein structures

MODBASE
Database
http://salilab.org/modbase
Fold assignments,alignments models, model assessments for all sequences related to a known structure

MODWEB
Web Server
http://salilab.org/modweb
Provides a web interface to MODPIPE

MODELLER
Program
http://salilab.org/modeller
Implements most operations in comparative modeling

DBALI
Database
http://salilab.org/dbali
Contains a comprehensive set of pairwise and multiple structure-based alignments

ICEDB
Database/LIMS
http://nysgxrc.org
Tracks targets for structural genomics by NYSGXRC

MODPIPE
Program
Automatically calculates comparative models of many protein sequences

EVA
Web Server
http://salilab.org/eva
Evaluates and ranks web servers for protein structure prediction

LIGBASE
Database
Ligand binding sites and inheritance (accessible through MODBASE)

External Resources
PDB, Uniprot, GENBANK, NR, PIR, INTERPRO, Kinase Resource
UCSC Genome Browser, CHIMERA, Pfam, SCOP, CATH

Acknowledgments

COMPARATIVE MODELING
Andrej Sali
M. S. Madhusudhan
Narayanan Eswar
David Eramian
Ursula Pieper
Ben Webb
Min-Yi Shen
Mark Peterson
Ash Stuart
Andras Fiser (AECOM)
Roberto Sanchez (MSSM)
Bino John (Pitsburg U.)
Eric Feyfant (GI)

ASSEMBLIES
Frank Alber
Damien Devos
Maya Topf
Dmitry Korkin
Fred Davis
Mike Kim

FUNCTIONAL ANNOTATION
Andrea Rossi
Rachel Karchin
Libusha Kelly
Nebojša Mirkovic

STRUCTURAL GENOMICS
Stephen Burley (SGX)
John Kuriyan (UCB)
NY-SGXRC

BRCA1
Alvaro Monteiro (Cornell)

MODEL ASSESSMENT
Francisco Melo (CU, Chile)
Alejandro Panjkovich (CU, Chile)

NIH
NSF
Sinsheimer Foundation
A. P. Sloan Foundation
Burroughs-Wellcome Fund
Merck Genome Res. Inst.
Mathers Foundation
I.T. Hirschl Foundation
The Sandler Family Foundation
Human Frontiers Science Program
SUN
IBM
Intel
Structural Genomix

Master Bioinformatics for Health Sciences

Comparative Protein Structure Prediction

template selection & sequence-structure alignment*

Marc A. Marti-Renom

Adjunct Assistant Professor

http://salilab.org/~marcius

UCSF
University of California
San Francisco

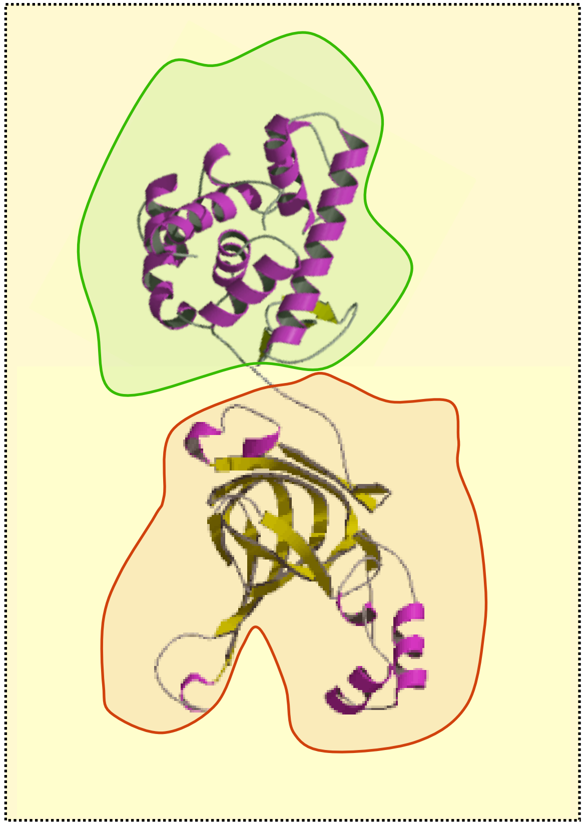
Depts. of Biopharmaceutical Sciences and Pharmaceutical Chemistry
California Institute for Quantitative Biomedical Research
University of California at San Francisco

Summary

- **Sequence and structure space (domains)**
  - Domains from sequence
  - Structure-Structure comparisons
  - How can we compare structures
  - How we classify the structural space
  - Aligning sequences and structures

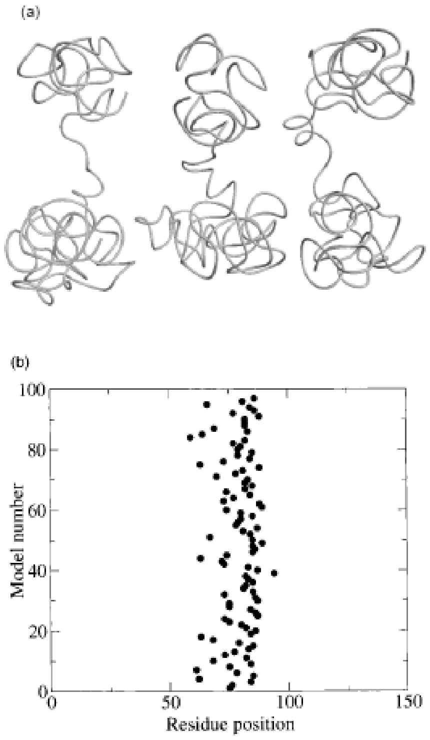
Domain boundaries from sequence

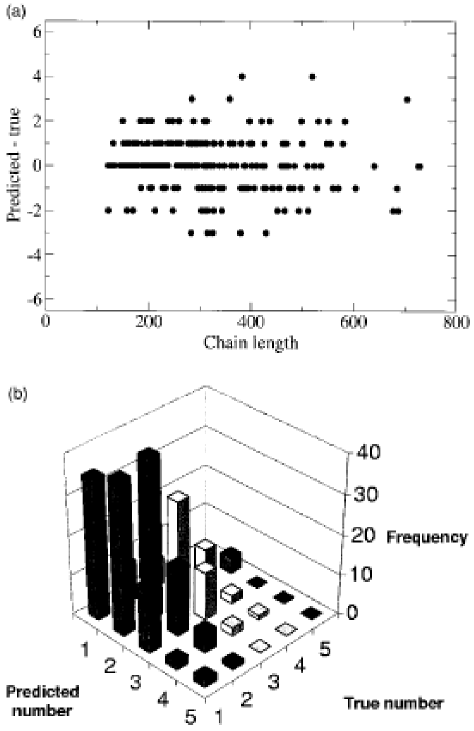
VERY DIFFICULT!!!!



```
MENFEIWVEKYRPRTLDEVVGQDEVIQRLKGYVERKNIPHLLFSGPPGTGKTATAIALARDLFGENWRDN
FIEMNASDERGIDVVRHKIKEFARTAPIGGAPFKIIFLDEADALTADAQAALRRTMEMYSKSCRFILSCN
YVSRIIEPIQSRCAVFRFKPVPKEAMKKRLLEICEKEGVKITEDGLEALIYISGGDFRKAINALQGAAAI
GEVVDADTIYQITATARPEEMTELIQTALKGNFMEARELLDRLMVEYGMSGEDIVAQLFREIISMPIKDS
LKVQLIDKLGEVDFRLTEGANERIQLDAYLAYLSTLAKK
```

Domain boundaries from sequence (SnapDragon)





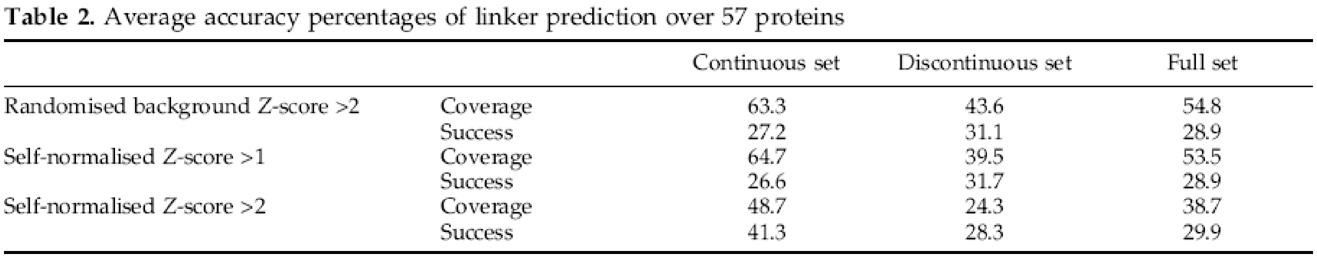
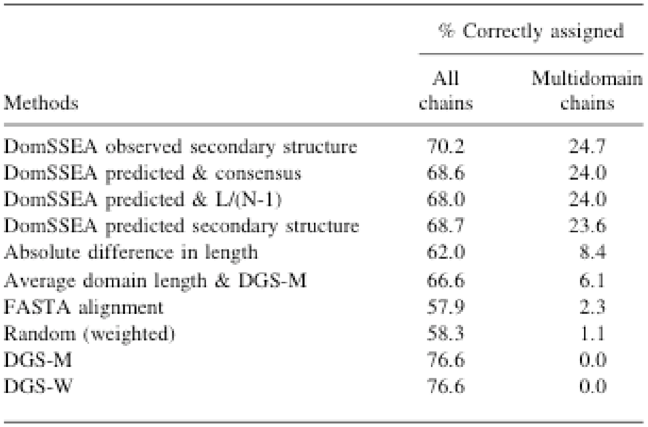


Table 2. Average accuracy percentages of linker prediction over 57 proteins

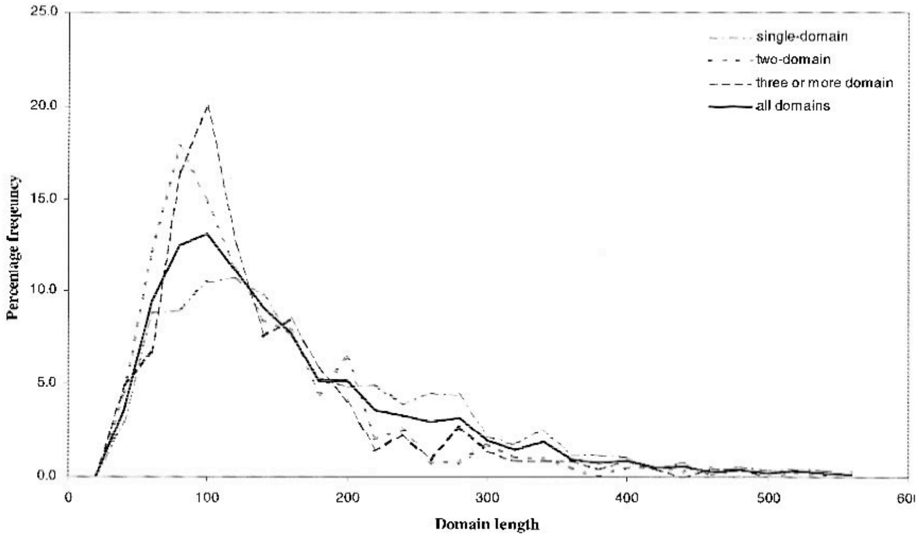
		Continuous set	Discontinuous set	Full set
Randomised background *Z*-score >2	Coverage	63.3	43.6	54.8
	Success	27.2	31.1	28.9
Self-normalised *Z*-score >1	Coverage	64.7	39.5	53.5
	Success	26.6	31.7	28.9
Self-normalised *Z*-score >2	Coverage	48.7	24.3	38.7
	Success	41.3	28.3	29.9

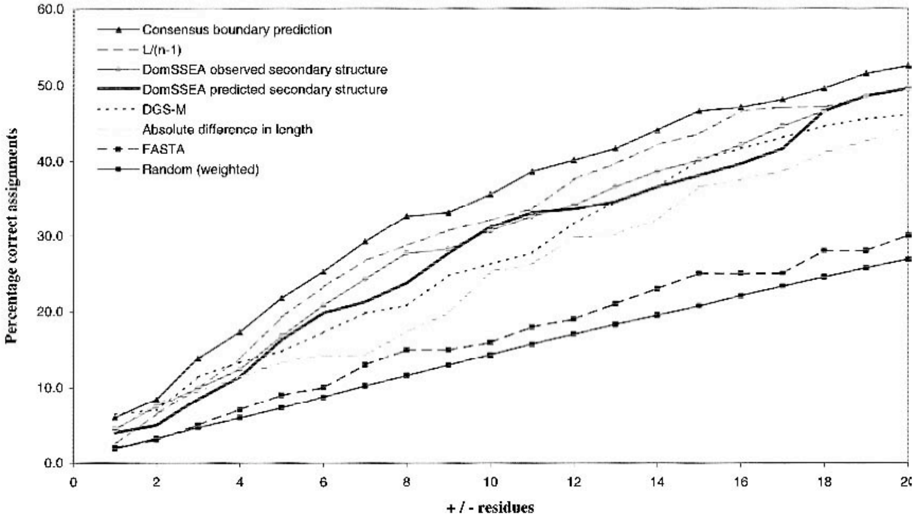
Georgea and Heringa (2002) J. Mol. Biol. ***316*** *pp839*

Domain boundaries from sequence and predicted SSE (DomSSEA)



Methods	% Correctly assigned	
	All chains	Multidomain chains
DomSSEA observed secondary structure	70.2	24.7
DomSSEA predicted & consensus	68.6	24.0
DomSSEA predicted & L/(N-1)	68.0	24.0
DomSSEA predicted secondary structure	68.7	23.6
Absolute difference in length	62.0	8.4
Average domain length & DGS-M	66.6	6.1
FASTA alignment	57.9	2.3
Random (weighted)	58.3	1.1
DGS-M	76.6	0.0
DGS-W	76.6	0.0



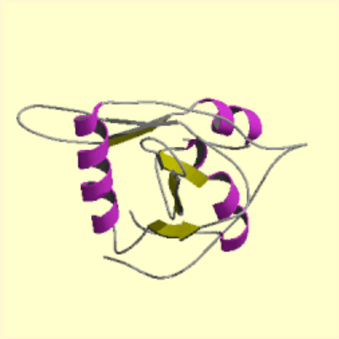


Dersden et al. (2003) Prot. Science **11** pp2014

Prediction of Secondary Structure (PSI-PRED)

```
>gi42541361
MDIRSVSSLRGLLCLPPSWPRR
```

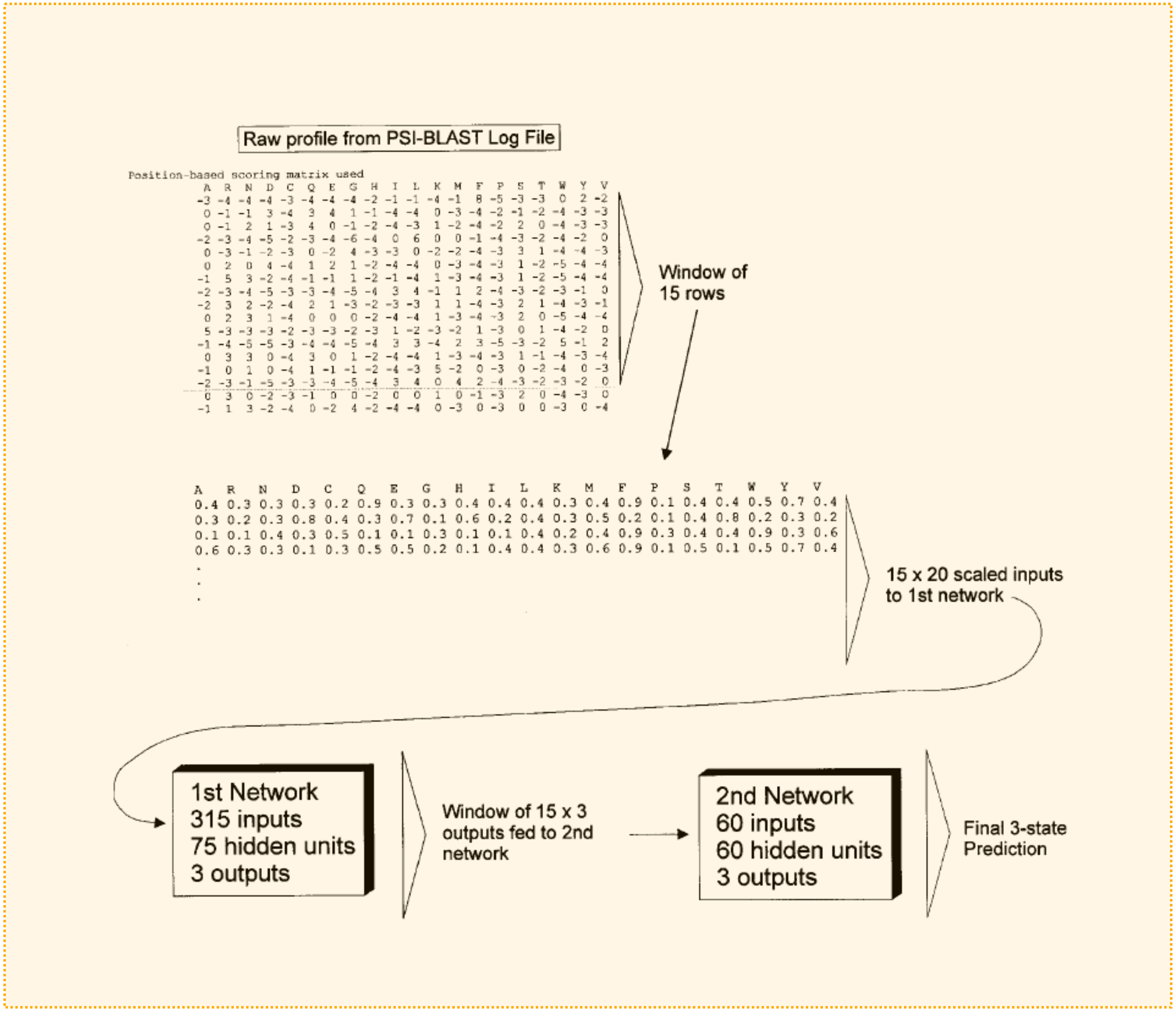
- Neural Network



✓ **Very simple idea**

✓ **Simple scoring**

Obscure optimizer



Raw profile from PSI-BLAST Log File

```
Position-based scoring matrix used
         A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
        -3 -4 -4 -4 -3 -4 -4 -4 -2 -1 -1 -4 -1  8 -5 -3 -3  0  2 -2
         0 -1 -1  3 -4  3  4  1 -1 -4 -4  0 -3 -4 -2 -1 -2 -4 -3 -3
         0 -1  2  1 -3  4  0 -1 -2 -4 -3  1 -2 -4 -2  2  0 -4 -3 -3
        -2 -3 -4 -5 -2 -3 -4 -6 -4  0  6  0  0 -1 -4 -3 -2 -4 -2  0
         0 -3 -1 -2 -3  0 -2  4 -3 -3  0 -2 -2 -4 -3  3  1 -4 -4 -3
         0  2  0  4 -4  1  2  1 -2 -4 -4  0 -3 -4 -3  1 -2 -5 -4 -4
        -1  5  3 -2 -4 -1 -1  1 -2 -1 -4  1 -3 -4 -3  1 -2 -5 -4 -4
        -2 -3 -4 -5 -3 -3 -4 -5 -4  3  4 -1  1  2 -4 -3 -2 -3 -1  0
        -2  3  2 -2 -4  2  1 -3 -2 -3 -3  1  1 -4 -3  2  1 -4 -3 -1
         0  2  3  1 -4  0  0  0 -2 -4 -4  1 -3 -4 -3  2  0 -5 -4 -4
         5 -3 -3 -3 -2 -3 -3 -2 -3  1 -2 -3 -2  1 -3  0  1 -4 -2  0
        -1 -4 -5 -5 -3 -4 -4 -5 -4  3  3 -4  2  3 -5 -3 -2  5 -1  2
         0  3  3  0 -4  3  0  1 -2 -4 -4  1 -3 -4 -3  1 -1 -4 -3 -4
        -1  0  1  0 -4  1 -1 -1 -2 -4 -3  5 -2  0 -3  0 -2 -4  0 -3
        -2 -3 -1 -5 -3 -3 -4 -5 -4  3  4  0  4  2 -4 -3 -2 -3 -2  0
      ------------------------------------------------------------
         0  3  0 -2 -3 -1  0  0 -2  0  0  1  0 -1 -3  2  0 -4 -3  0
        -1  1  3 -2 -4  0 -2  4 -2 -4 -4  0 -3  0 -3  0  0 -3  0 -4
```

Window of 15 rows

```
A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V
0.4 0.3 0.3 0.3 0.2 0.9 0.3 0.3 0.4 0.4 0.4 0.3 0.4 0.9 0.1 0.4 0.4 0.5 0.7 0.4
0.3 0.2 0.3 0.8 0.4 0.3 0.7 0.1 0.6 0.2 0.4 0.3 0.5 0.2 0.1 0.4 0.8 0.2 0.3 0.2
0.1 0.1 0.4 0.3 0.5 0.1 0.1 0.3 0.1 0.1 0.4 0.2 0.4 0.9 0.3 0.4 0.4 0.9 0.3 0.6
0.6 0.3 0.3 0.1 0.3 0.5 0.5 0.2 0.1 0.4 0.4 0.3 0.6 0.9 0.1 0.5 0.1 0.5 0.7 0.4
.
.
.
```

15 x 20 scaled inputs to 1st network

1st Network
315 inputs
75 hidden units
3 outputs

Window of 15 x 3 outputs fed to 2nd network

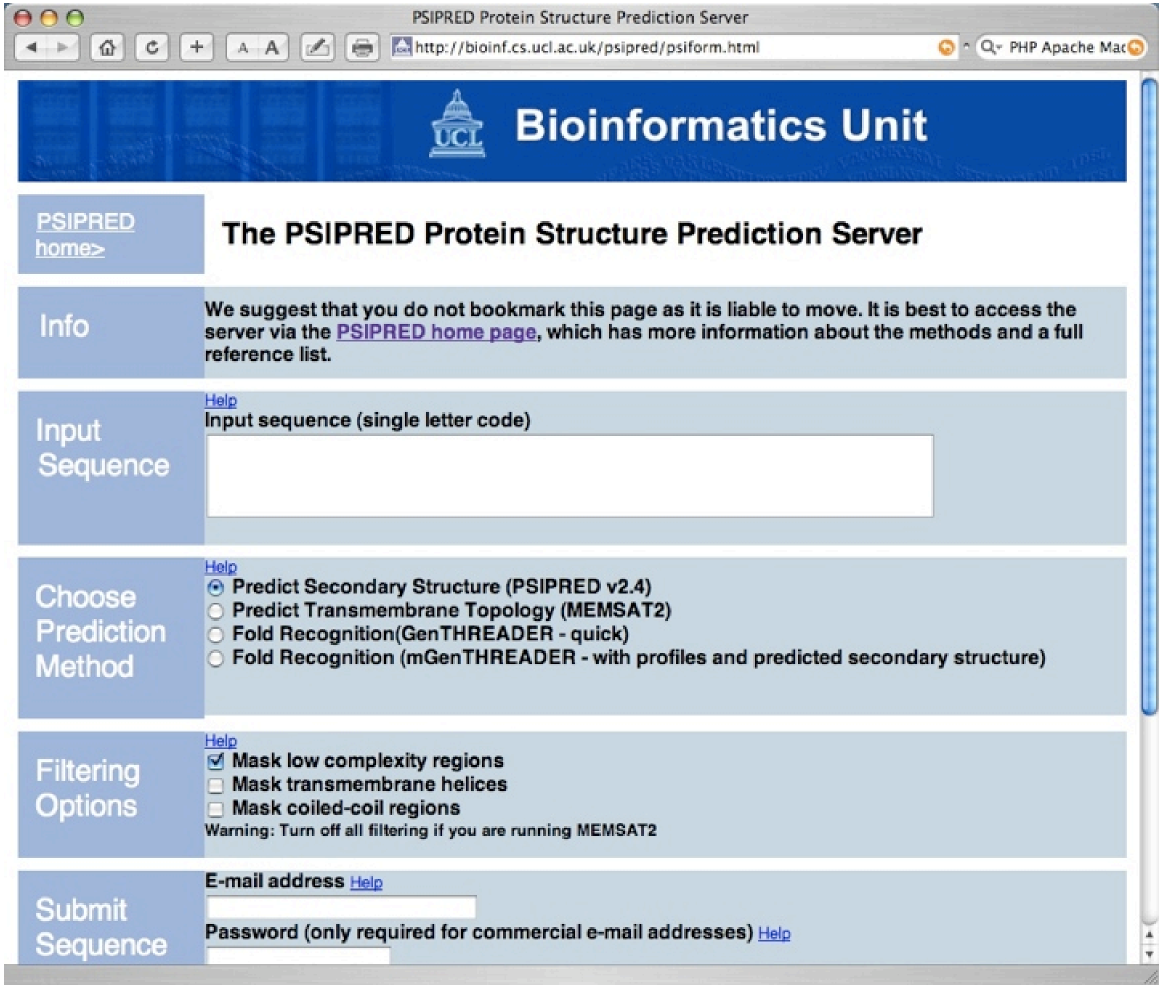
2nd Network
60 inputs
60 hidden units
3 outputs

Final 3-state Prediction

Jones DT. (1999) J. Mol. Biol. ***292*** *pp195*

Prediction of Secondary Structure (PSI-PRED)

http://bioinf.cs.ucl.ac.uk/psipred/



PSIPRED Protein Structure Prediction Server
http://bioinf.cs.ucl.ac.uk/psipred/psiform.html
PHP Apache Mac
Bioinformatics Unit
UCL

PSIPRED home>

The PSIPRED Protein Structure Prediction Server

Info

We suggest that you do not bookmark this page as it is liable to move. It is best to access the server via the PSIPRED home page, which has more information about the methods and a full reference list.

Input Sequence

Help
Input sequence (single letter code)

Choose Prediction Method

Help
● Predict Secondary Structure (PSIPRED v2.4)
○ Predict Transmembrane Topology (MEMSAT2)
○ Fold Recognition(GenTHREADER - quick)
○ Fold Recognition (mGenTHREADER - with profiles and predicted secondary structure)

Filtering Options

Help
☑ Mask low complexity regions
□ Mask transmembrane helices
□ Mask coiled-coil regions
Warning: Turn off all filtering if you are running MEMSAT2

Submit Sequence

E-mail address Help

Password (only required for commercial e-mail addresses) Help

Template Selection

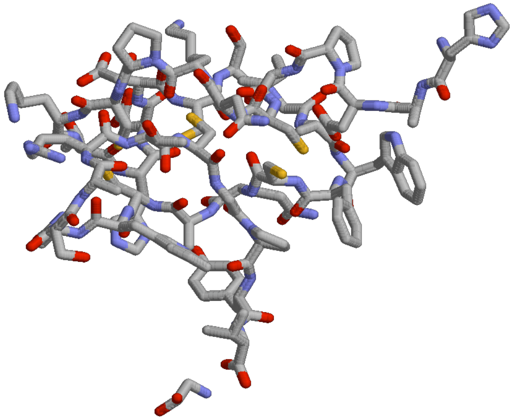
“Structural Space”

Structure-Structure alignments

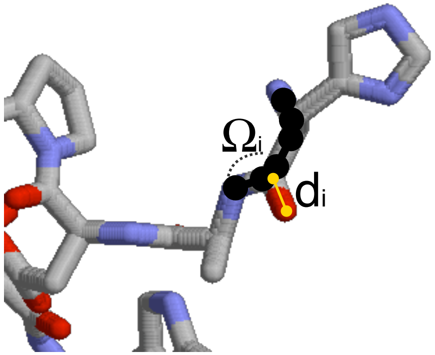
As any other bioinformatics problem…

- Representation
- Scoring
- Optimizer

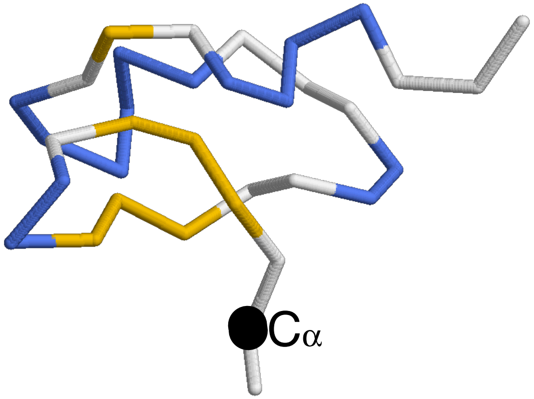
Structures



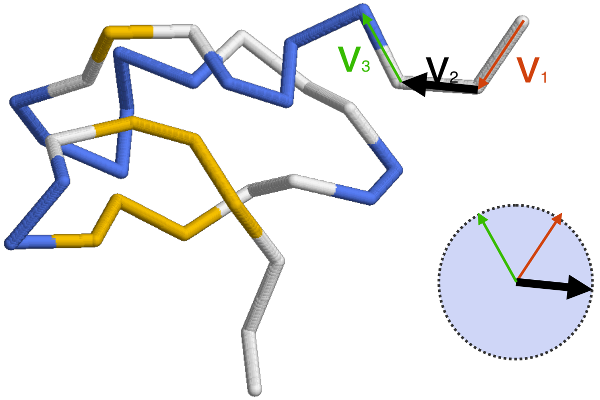
All atoms and coordinates



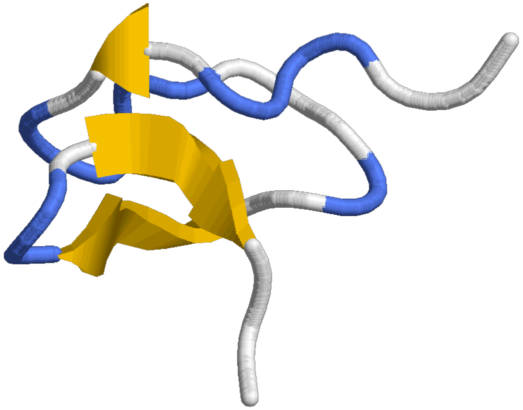
Dihedral space or distance space



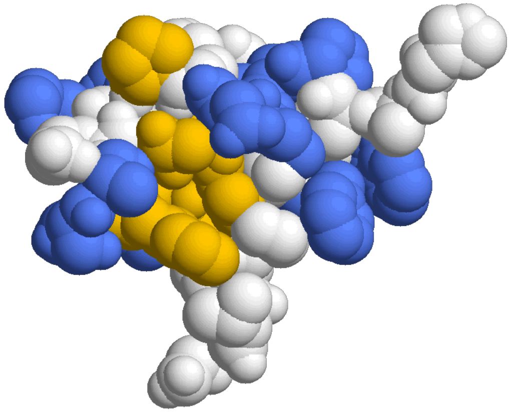
Reduced atom representation



Vector representation

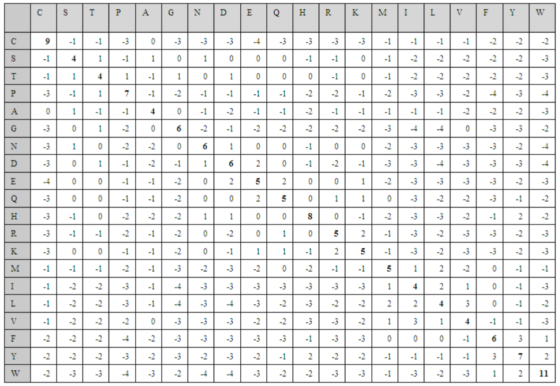


Secondary Structure



Accessible surface (and others)

Raw scores



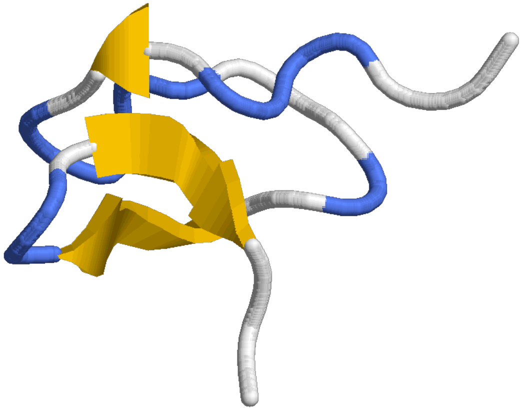
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	**9**	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	**4**	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	**4**	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	**7**	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	**4**	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	**6**	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	**6**	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	**6**	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	**5**	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	**5**	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	**8**	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	**5**	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	**7**	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	**11**

2/

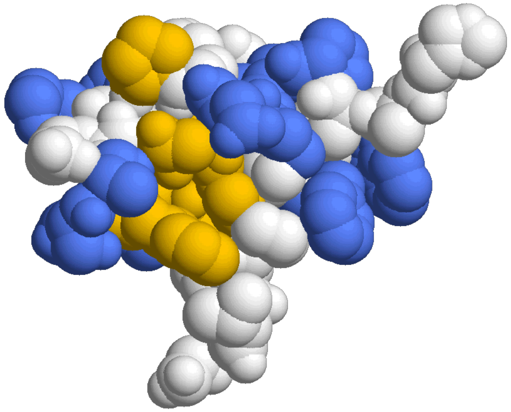
Aminoacid substitutions

$$RMSD(x,y) = \sqrt{\left(\tfrac{1}{N}\right)\sum_{i=1}^{N}\left(\left\|\mathbf{x}(i) - \mathbf{y}(i)\right\|^{2}\right)}$$

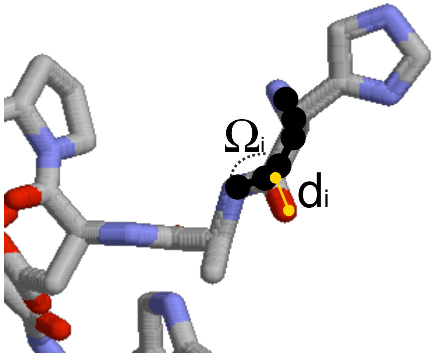
Root Mean Square Deviation



Secondary Structure (H,B,C)



Accessible surface (B,A [%])



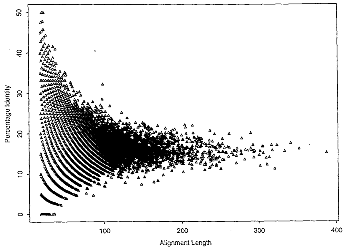
Angles or distances

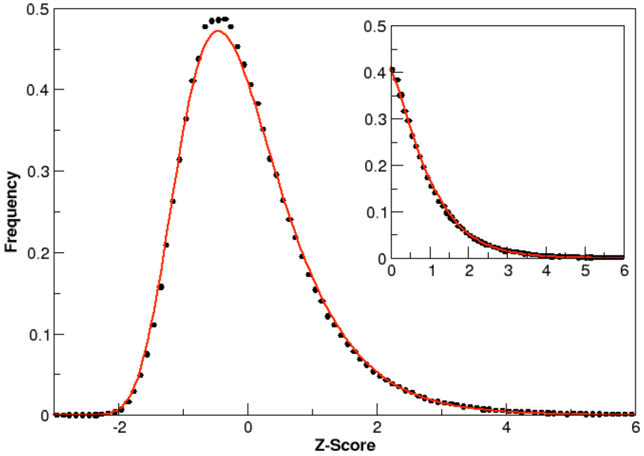
Scoring

Significance of an alignment (score)

Probability that the optimal alignment of two random sequences/structures of the same length and composition as the aligned sequences/structures have at least as good a score as the evaluated alignment.

Empirical





Sometimes approximated by Z-score (normal distribution).

Analytic

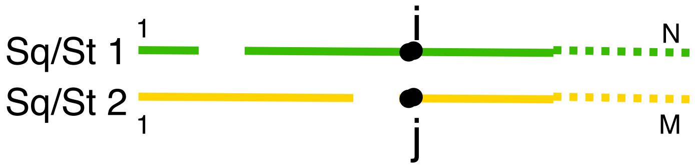
$$P(s) = e^{-\lambda\ (s-\mu)}$$

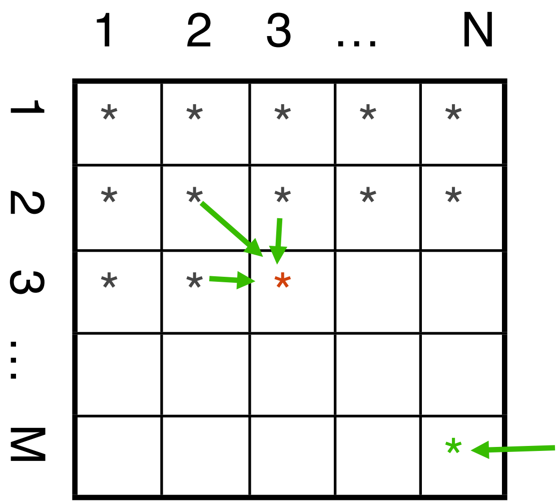
$$P(s \geq x) = 1-\exp\left(e^{-\lambda\ (s-\mu)}\right)$$

Karlin and Altschul, 1990 PNAS 87, pp2264

Optimizer

Global dynamic programming alignment





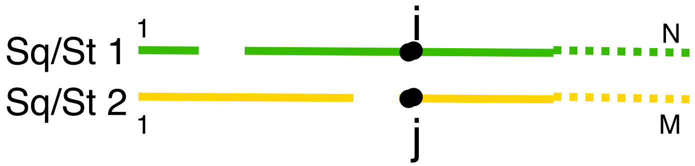
Best alignment score

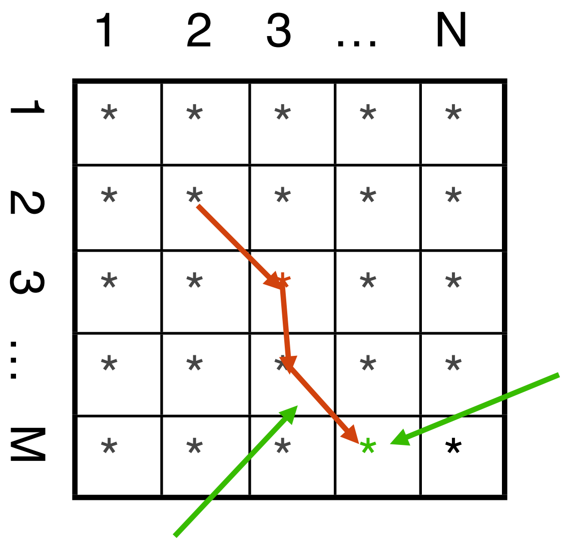
$$D_{i,j}=\min\begin{cases} D_{i,j-1}+\mathrm{Score}_{(\Delta,rj)} \\ D_{i-1,j-1}+\mathrm{Score}_{(ri,rj)} \\ D_{i-1,j}+\mathrm{Score}_{(ri,\Delta)} \end{cases}$$

Backtracking to get the best alignment

Needleman and Wunsch (1970) J. Mol Biol, 3 pp443

Local dynamic programming alignment





Best score

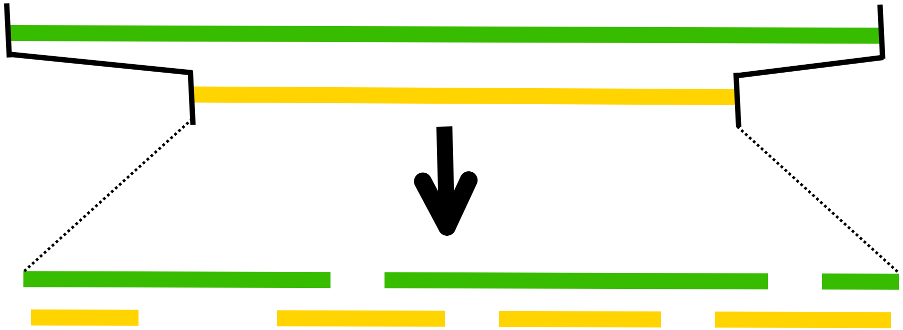
Best local alignment

$$D_{i,j}=\min\begin{cases} D_{i,j-1}+Score_{(\Delta,rj)} \\ D_{i-1,j-1}+Score_{(ri,rj)} \\ D_{i-1,j}+Score_{(ri,\Delta)} \\ \textcircled{0} \end{cases}$$

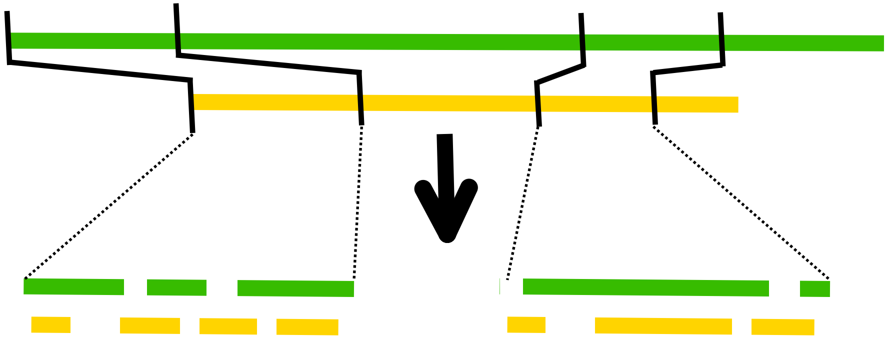
Backtracking to get the best alignment

Smith and Waterman (1981) J. Mol Biol, 147 pp195

Global .vs. local alignment



Global alignment

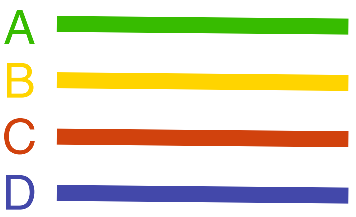


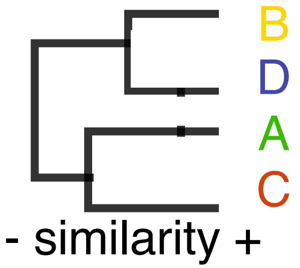
Local alignment

Multiple alignment

Pairwise alignments

Example – 4 sequences A, B, C, D.

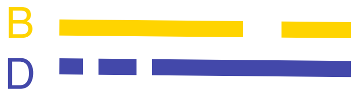




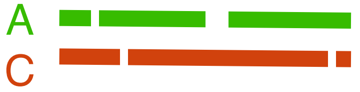
6 pairwise comparisons
then cluster analysis

Multiple alignments

Following the tree from step 1

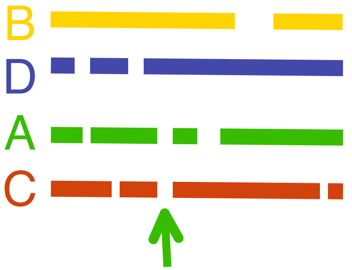


Align the most similar pair



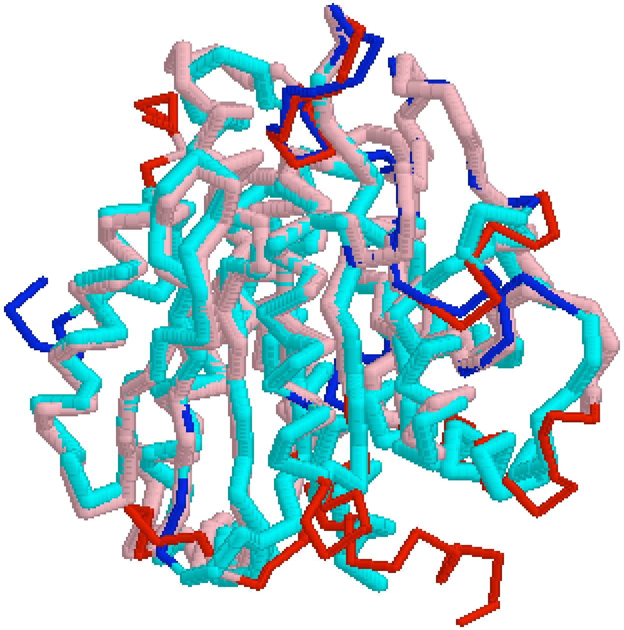
Align next most similar pair

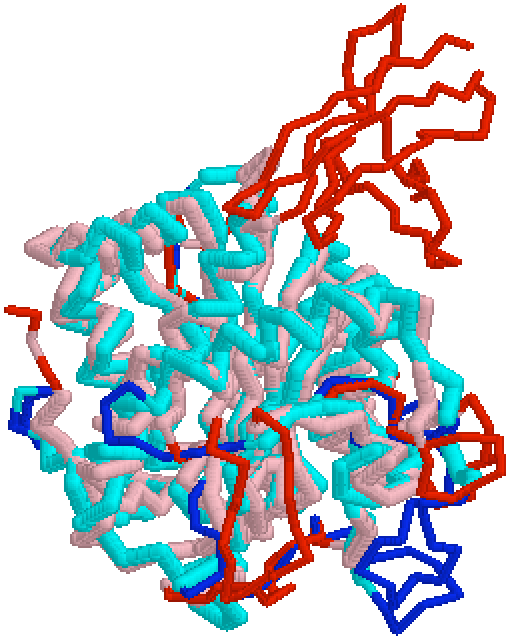
Align B-D with A-C



New gap in A-C to optimize
its alignment with B-D

Coverage .vs. Accuracy



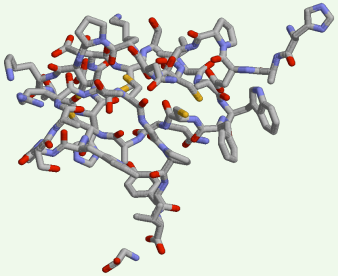


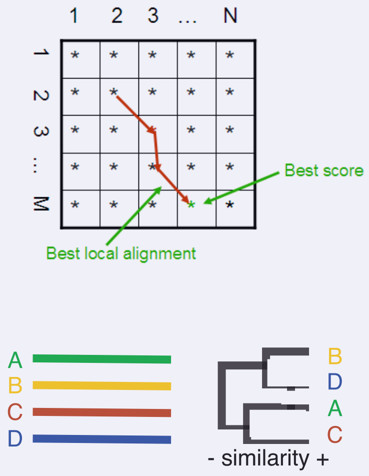
Same RMSD ~ 2.5Å

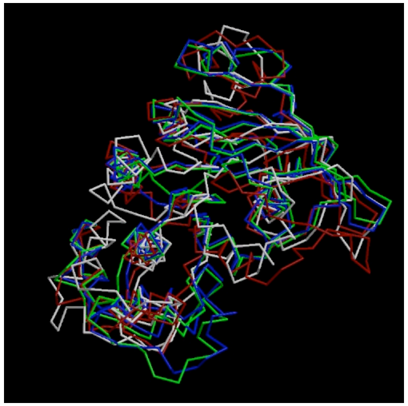
Coverage ~90% Cα

Coverage ~75% Cα

Structural alignment by properties conservation (SALIGN-MODELLER)

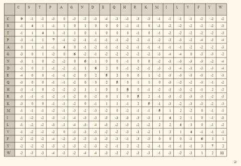


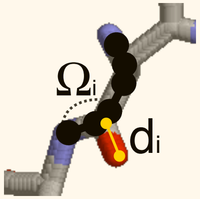


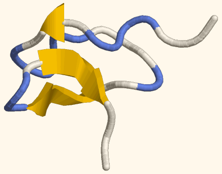


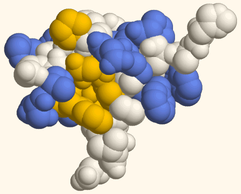
- ✓ Uses all available structural information
- ✓ Provides the optimal alignment

Computationally expensive









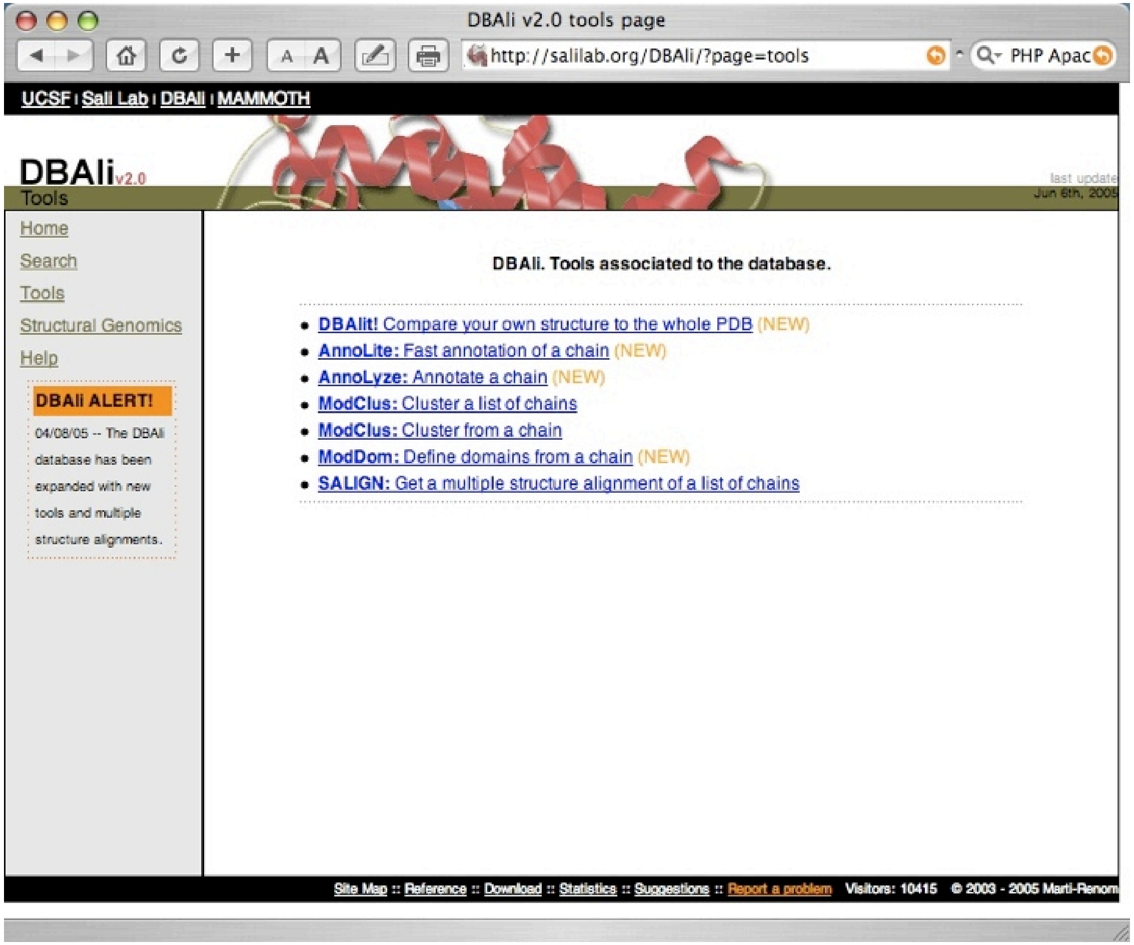
$$RMSD(x,y)=\sqrt{\left(\tfrac{1}{N}\right)\sum_{i=1}^{N}\left(\left\|\mathbf{x}(i)-\mathbf{y}(i)\right\|^{2}\right)}$$

$R_{i,j}$ $D_{,i(3),j(3)}$ $S_{i,j}$ $B_{i,j}$ $I_{i,j}$

Madhusudhan et al. in preparation

Structural alignment by properties conservation (SALIGN-MODELLER)

`http://salilab.org/DBAli`



DBAli ALERT!

04/08/05 -- The DBAli database has been expanded with new tools and multiple structure alignments.

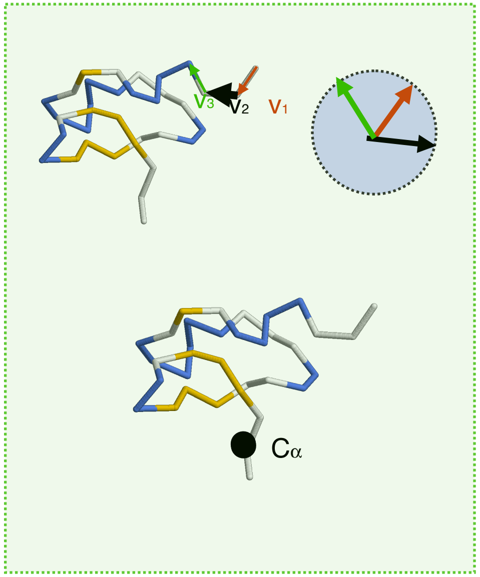
DBAli. Tools associated to the database.

- **DBAlit!** Compare your own structure to the whole PDB (NEW)
- **AnnoLite:** Fast annotation of a chain (NEW)
- **AnnoLyze:** Annotate a chain (NEW)
- **ModClus:** Cluster a list of chains
- **ModClus:** Cluster from a chain
- **ModDom:** Define domains from a chain (NEW)
- **SALIGN:** Get a multiple structure alignment of a list of chains

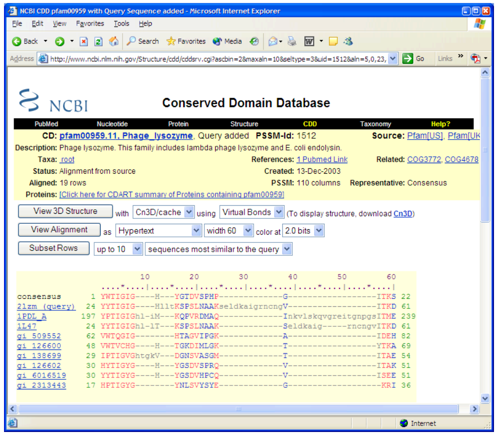
Site Map :: Reference :: Download :: Statistics :: Suggestions :: Report a problem Visitors: 10415 © 2003 - 2005 Marti-Renom

Madhusudhan, in preparation

Vector Alignment Search Tool (VAST)



Graph theory search
of similar SSE
Refining by Monte Carlo
at all atom resolution



Address: http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?ascbin=2&maxaln=10&seltype=3&uid=1512&aln=5,0,23,

NCBI Conserved Domain Database

CD: pfam00959.11, Phage_lysozyme, Query added **PSSM-Id:** 1512 **Source:** Pfam[US], Pfam[Uk

Description: Phage lysozyme. This family includes lambda phage lysozyme and E. coli endolysin.

Taxa: root **References:** 1 Pubmed Link **Related:** COG3772, COG4678

Status: Alignment from source **Created:** 13-Dec-2003

Aligned: 19 rows **PSSM:** 110 columns **Representative:** Consensus

Proteins: [Click here for CDART summary of Proteins containing pfam00959]

View 3D Structure with Cn3D/cache using Virtual Bonds (To display structure, download Cn3D)

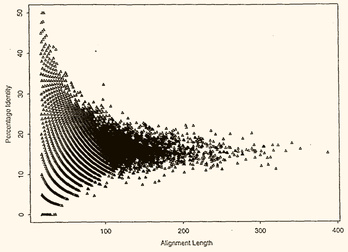
View Alignment as Hypertext width 60 color at 2.0 bits

Subset Rows up to 10 sequences most similar to the query

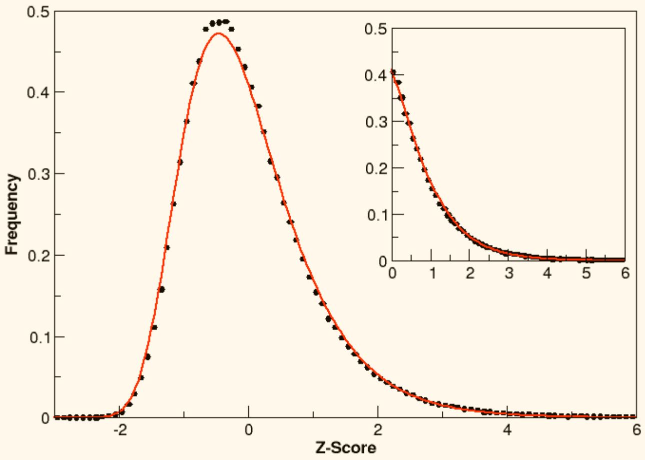
```
                         10        20        30        40        50        60
                 ....*....|....*....|....*....|....*....|....*....|....*....|
consensus      1 YWTIGIG----H---YGTDVSPHP-------------G------------------ITKS 22
2lzm (query)  24 YYTIGIG----HlltKSPSLNAAKseldkaigrncngV------------------ITKD 61
1PDL_A       197 YPTIGIGhl-iM---KQPVRDMAQ-------------InkvlskqvgreitgnpgsITME 239
1L47          24 YYTIGIGhl-lT---KSPSLNAAK-------------Seldkaig-----rncngvITKD 61
gi 509552     62 VWTQGIG--------HTAGVIPGK-------------A------------------IDEH 82
gi 126600     48 VWTVCHG----H---TGKDIMLGK-------------T------------------YTKA 69
gi 138699     29 IPTIGVGhtgkV---DGNSVASGM-------------T------------------ITAE 54
gi 126602     30 HYTIGYG----H---YGSDVSPRQ-------------V------------------ITAK 51
gi 6016519    30 YYTIGYG----H---YGSDVHPCQ-------------V------------------ISEE 51
gi 2313443    17 HPTIGYG--------YNLSVYSYE-------------G-------------------KRI 36
```

✓ Good scoring system with significance

Reduces the protein representation



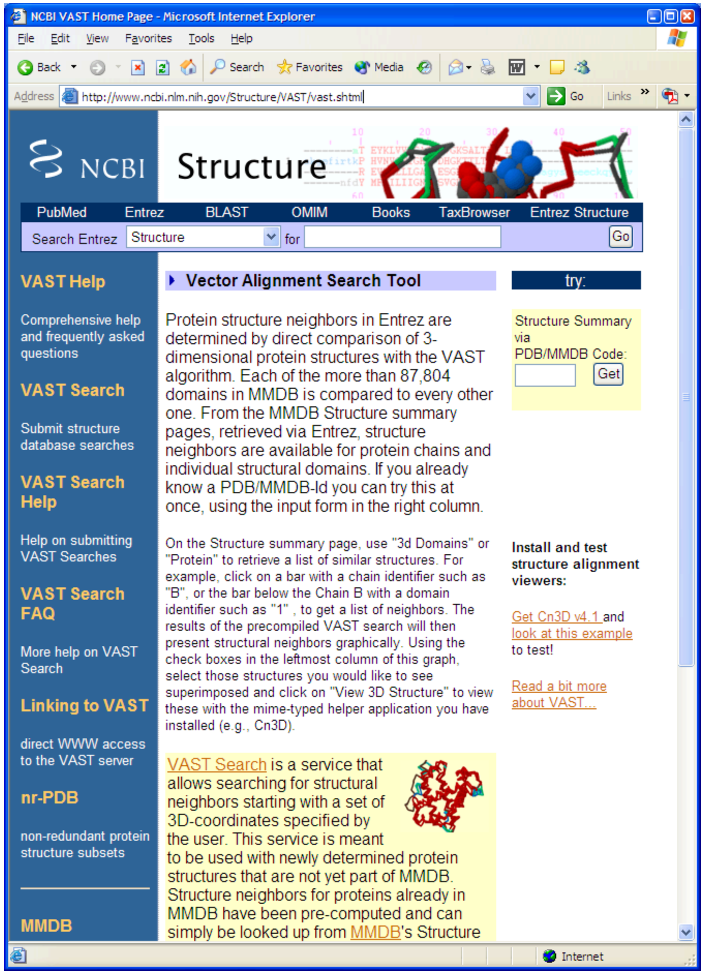
$$RMSD(x,y)=\sqrt{\left(\tfrac{1}{N}\right)\sum_{i=1}^{N}\left(\left\|\mathbf{x}(i)-\mathbf{y}(i)\right\|^{2}\right)}$$



Gibrat JF et al. (1996) Curr Opin Struct Biol 3 pp377

Vector Alignment Search Tool (VAST)

`http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml`


NCBI VAST Home Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml Go Links
NCBI Structure
PubMed Entrez BLAST OMIM Books TaxBrowser Entrez Structure
Search Entrez Structure for Go
try:
Get
Internet

VAST Help

Comprehensive help and frequently asked questions

VAST Search

Submit structure database searches

VAST Search Help

Help on submitting VAST Searches

VAST Search FAQ

More help on VAST Search

Linking to VAST

direct WWW access to the VAST server

nr-PDB

non-redundant protein structure subsets

MMDB

Vector Alignment Search Tool

Protein structure neighbors in Entrez are determined by direct comparison of 3-dimensional protein structures with the VAST algorithm. Each of the more than 87,804 domains in MMDB is compared to every other one. From the MMDB Structure summary pages, retrieved via Entrez, structure neighbors are available for protein chains and individual structural domains. If you already know a PDB/MMDB-Id you can try this at once, using the input form in the right column.

On the Structure summary page, use "3d Domains" or "Protein" to retrieve a list of similar structures. For example, click on a bar with a chain identifier such as "B", or the bar below the Chain B with a domain identifier such as "1" , to get a list of neighbors. The results of the precompiled VAST search will then present structural neighbors graphically. Using the check boxes in the leftmost column of this graph, select those structures you would like to see superimposed and click on "View 3D Structure" to view these with the mime-typed helper application you have installed (e.g., Cn3D).

VAST Search is a service that allows searching for structural neighbors starting with a set of 3D-coordinates specified by the user. This service is meant to be used with newly determined protein structures that are not yet part of MMDB. Structure neighbors for proteins already in MMDB have been pre-computed and can simply be looked up from MMDB's Structure

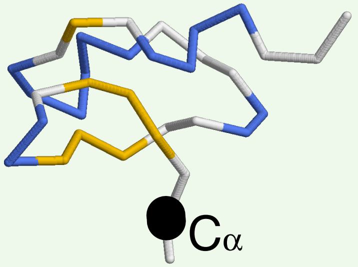
Structure Summary via PDB/MMDB Code:

Install and test structure alignment viewers:

Get Cn3D v4.1 and look at this example to test!

Read a bit more about VAST...

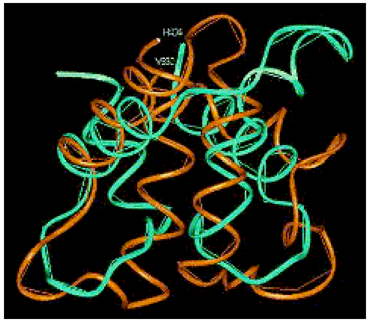
Incremental combinatorial extension (CE)



Exhaustive combination of fragments

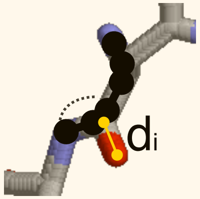
Longest combination of AFPs

Heuristic similar to PSI-BLAST



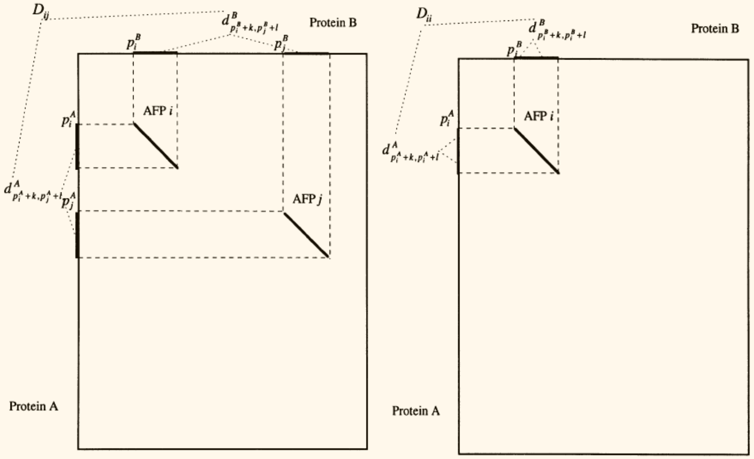
- ✓ FAST!
- ✓ Good quality of local alignments

Complicated scoring and heuristics



8 residues peptides

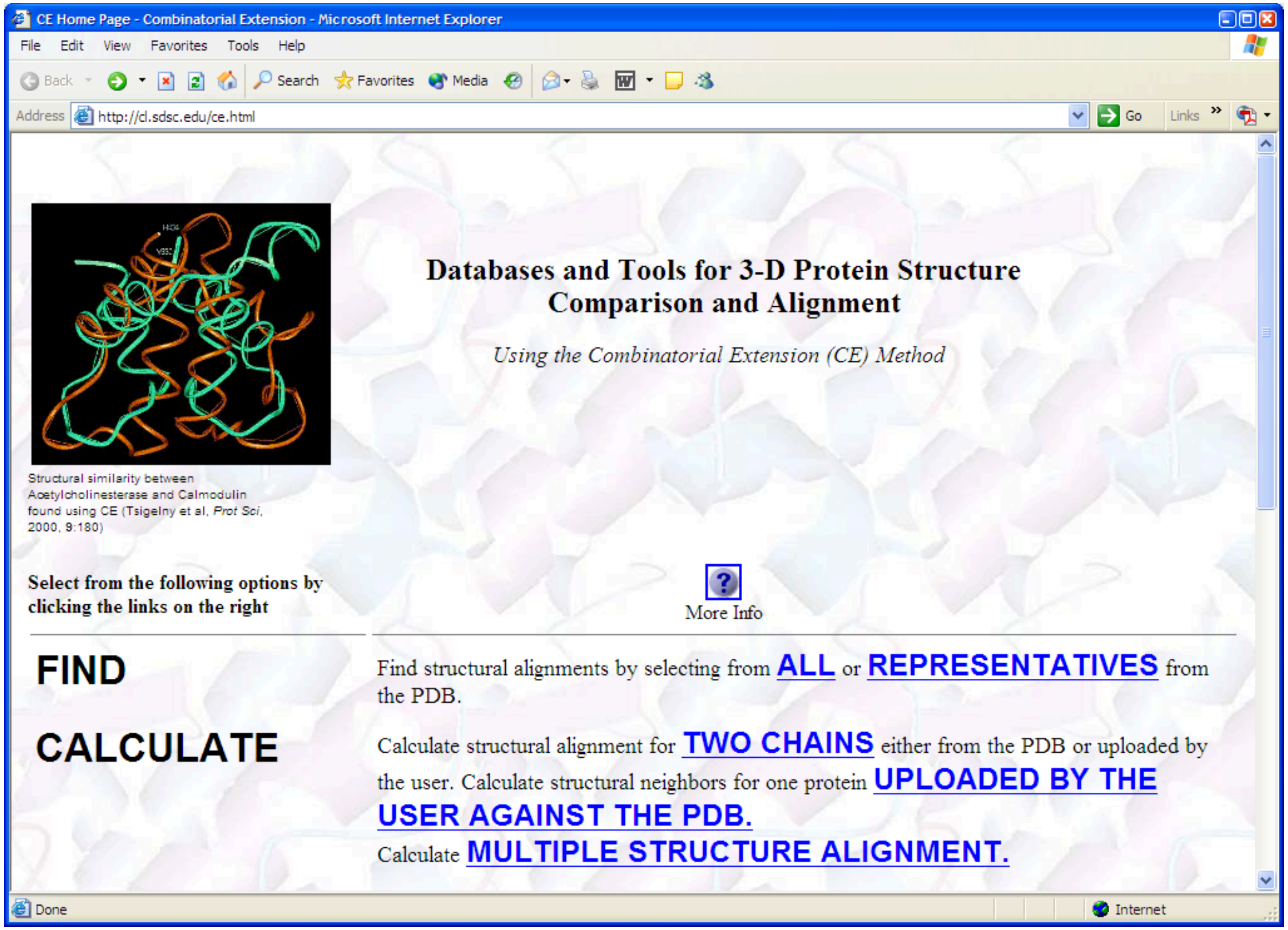
$$RMSD(x,y) = \sqrt{\left(\frac{1}{N}\right)\sum_{i=1}^{N}\left(\left\|\mathbf{x}(i) - \mathbf{y}(i)\right\|^{2}\right)}$$



Shindyalov IN, amd Bourne PE. (1998) Protein Eng. 9 pp739

Incremental combinatorial extension (CE)

http://cl.sdsc.edu/ce.html


CE Home Page - Combinatorial Extension - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://cl.sdsc.edu/ce.html
Go Links
Done
Internet

Structural similarity between Acetylcholinesterase and Calmodulin found using CE (Tsigelny et al, *Prot Sci*, 2000, 9:180)

Databases and Tools for 3-D Protein Structure Comparison and Alignment

Using the Combinatorial Extension (CE) Method

Select from the following options by clicking the links on the right

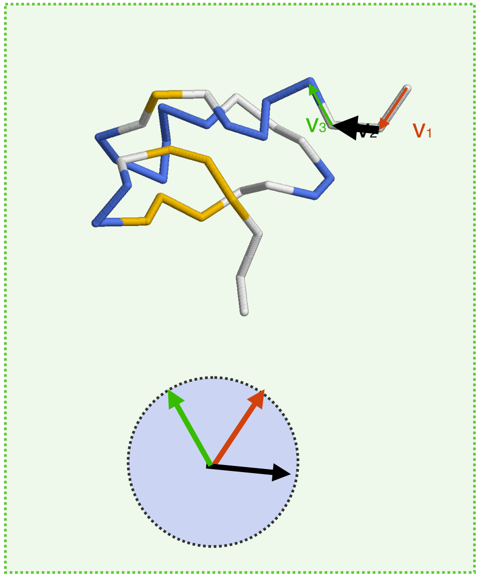
? More Info

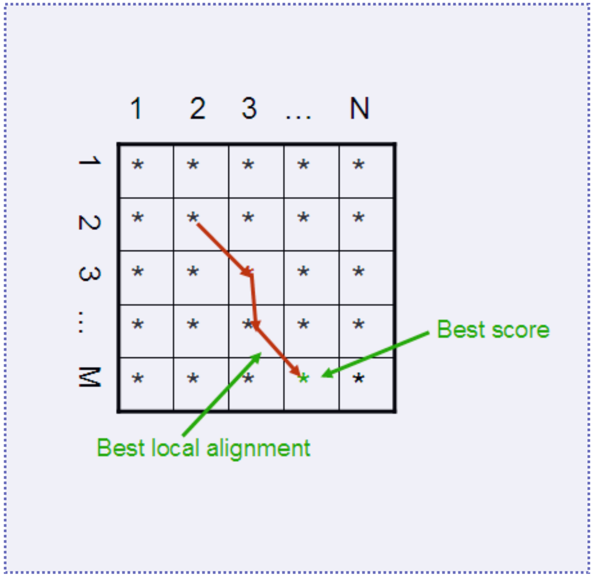
FIND Find structural alignments by selecting from ALL or REPRESENTATIVES from the PDB.

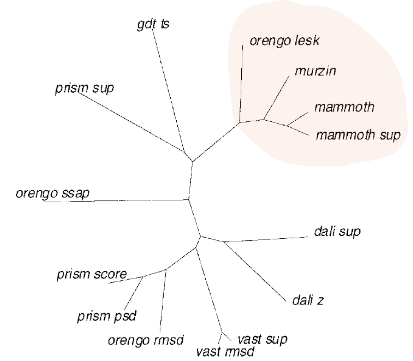
CALCULATE Calculate structural alignment for TWO CHAINS either from the PDB or uploaded by the user. Calculate structural neighbors for one protein UPLOADED BY THE USER AGAINST THE PDB.

Calculate MULTIPLE STRUCTURE ALIGNMENT.

Matching molecular models obtained from theory (MAMMOTH)





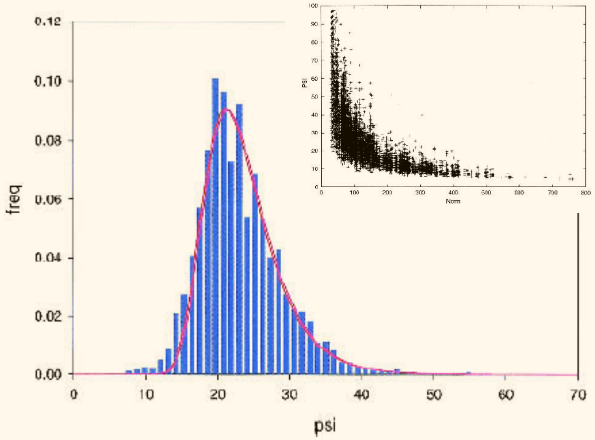


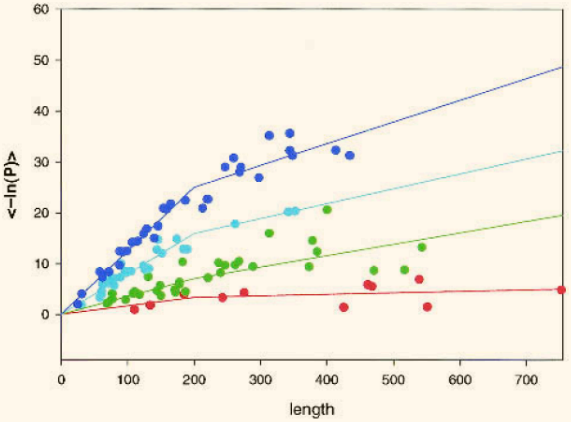
- ✓ VERY FAST!
- ✓ Good scoring system with significance

Reduces the protein representation

$$URMS^R = \sqrt{2.0 - \frac{2.84}{\sqrt{n}}}$$

$$S_{AB} = \frac{(URMS^R - URMS^{AB})D}{URMS^R}$$

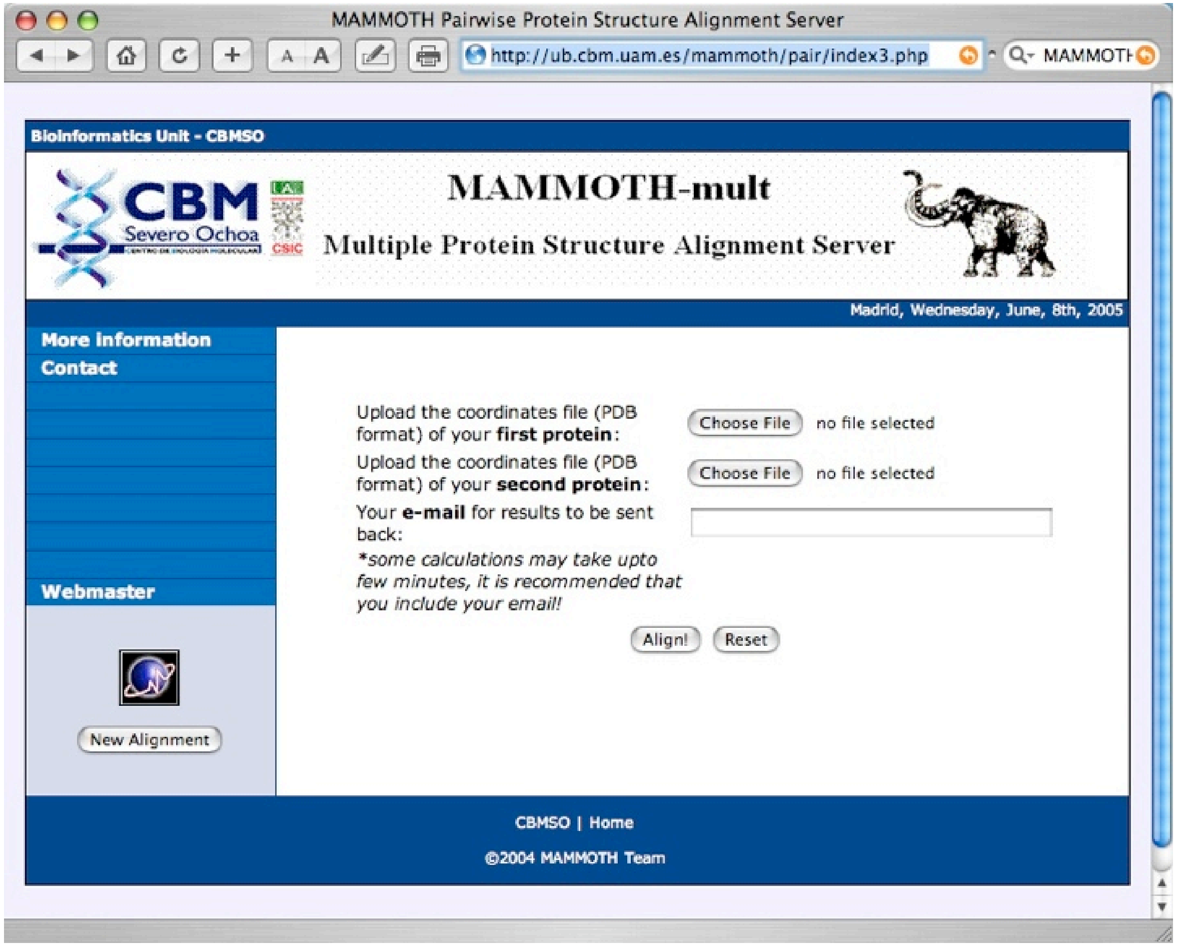




Ortiz AR, (2002) Protein Sci. 11 pp2606

Matching molecular models obtained from theory (MAMMOTH)

`http://ub.cbm.uam.es/mammoth/pair/index3.php`



MAMMOTH Pairwise Protein Structure Alignment Server

http://ub.cbm.uam.es/mammoth/pair/index3.php

Bioinformatics Unit - CBMSO

CBM Severo Ochoa

MAMMOTH-mult

Multiple Protein Structure Alignment Server

Madrid, Wednesday, June, 8th, 2005

Upload the coordinates file (PDB format) of your **first protein**: Choose File no file selected

Upload the coordinates file (PDB format) of your **second protein**: Choose File no file selected

Your **e-mail** for results to be sent back:

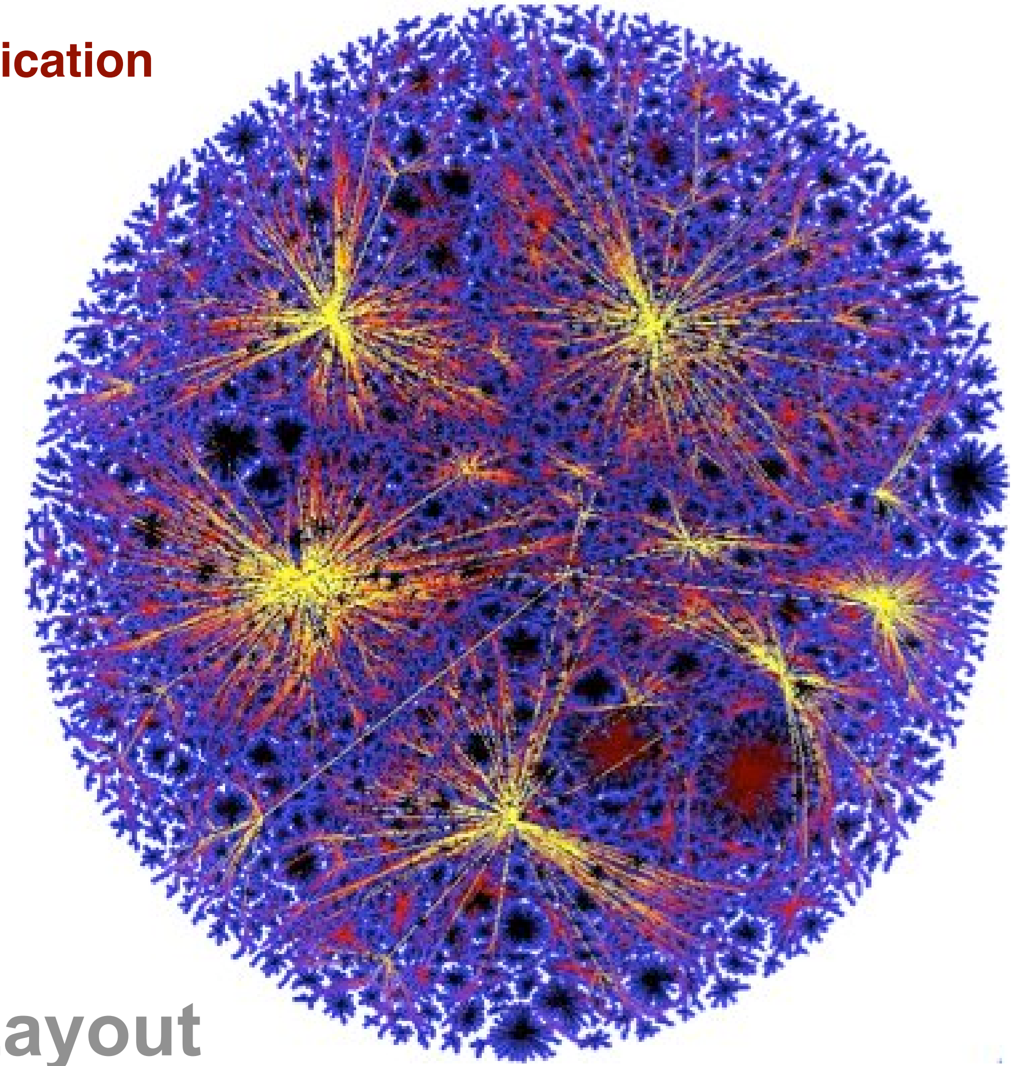
**some calculations may take upto few minutes, it is recommended that you include your email!*

CBMSO | Home

©2004 MAMMOTH Team

Classification of the structural space

SCOP classification



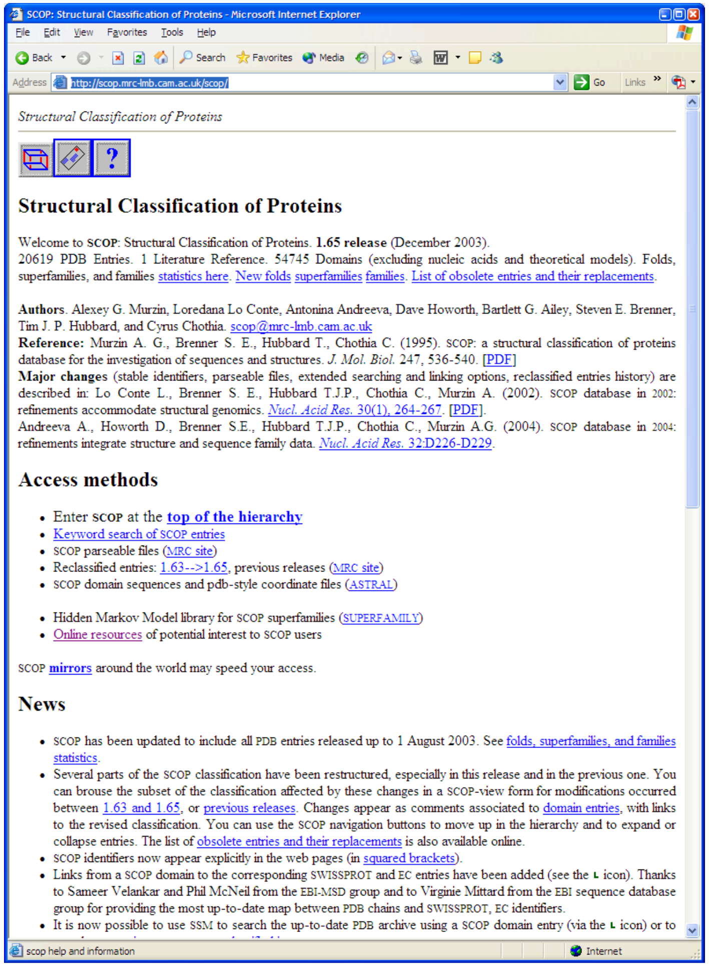
Large Graph Layout

Adai AT, Date SV, Wieland S, Marcotte EM. J Mol Biol. 2004 Jun 25;340(1):179-90

http://bioinformatics.icmb.utexas.edu/lgl/

SCOP$_{1.65}$ database

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


SCOP: Structural Classification of Proteins - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address
http://scop.mrc-lmb.cam.ac.uk/scop/
Go
Links
?
scop help and information
Internet

Structural Classification of Proteins

Structural Classification of Proteins

Welcome to SCOP: Structural Classification of Proteins. **1.65 release** (December 2003).
20619 PDB Entries. 1 Literature Reference. 54745 Domains (excluding nucleic acids and theoretical models). Folds, superfamilies, and families statistics here. New folds superfamilies families. List of obsolete entries and their replacements.

Authors. Alexey G. Murzin, Loredana Lo Conte, Antonina Andreeva, Dave Howorth, Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia. scop@mrc-lmb.cam.ac.uk
Reference: Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540. [PDF]
Major changes (stable identifiers, parseable files, extended searching and linking options, reclassified entries history) are described in: Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.* 30(1), 264-267. [PDF].
Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res.* 32:D226-D229.

Access methods

- Enter SCOP at the **top of the hierarchy**
- Keyword search of SCOP entries
- SCOP parseable files (MRC site)
- Reclassified entries: 1.63-->1.65, previous releases (MRC site)
- SCOP domain sequences and pdb-style coordinate files (ASTRAL)

- Hidden Markov Model library for SCOP superfamilies (SUPERFAMILY)
- Online resources of potential interest to SCOP users

SCOP **mirrors** around the world may speed your access.

News

- SCOP has been updated to include all PDB entries released up to 1 August 2003. See folds, superfamilies, and families statistics.
- Several parts of the SCOP classification have been restructured, especially in this release and in the previous one. You can brouse the subset of the classification affected by these changes in a SCOP-view form for modifications occurred between 1.63 and 1.65, or previous releases. Changes appear as comments associated to domain entries, with links to the revised classification. You can use the SCOP navigation buttons to move up in the hierarchy and to expand or collapse entries. The list of obsolete entries and their replacements is also available online.
- SCOP identifiers now appear explicitly in the web pages (in squared brackets).
- Links from a SCOP domain to the corresponding SWISSPROT and EC entries have been added (see the L icon). Thanks to Sameer Velankar and Phil McNeil from the EBI-MSD group and to Virginie Mittard from the EBI sequence database group for providing the most up-to-date map between PDB chains and SWISSPROT, EC identifiers.
- It is now possible to use SSM to search the up-to-date PDB archive using a SCOP domain entry (via the L icon) or to

✓ **Largely recognized as "standard of gold"**
✓ **Manually classification**
✓ **Clear classification of structures in:**
CLASS
FOLD
SUPER-FAMILY
FAMILY
✓ **Some large number of tools already available**

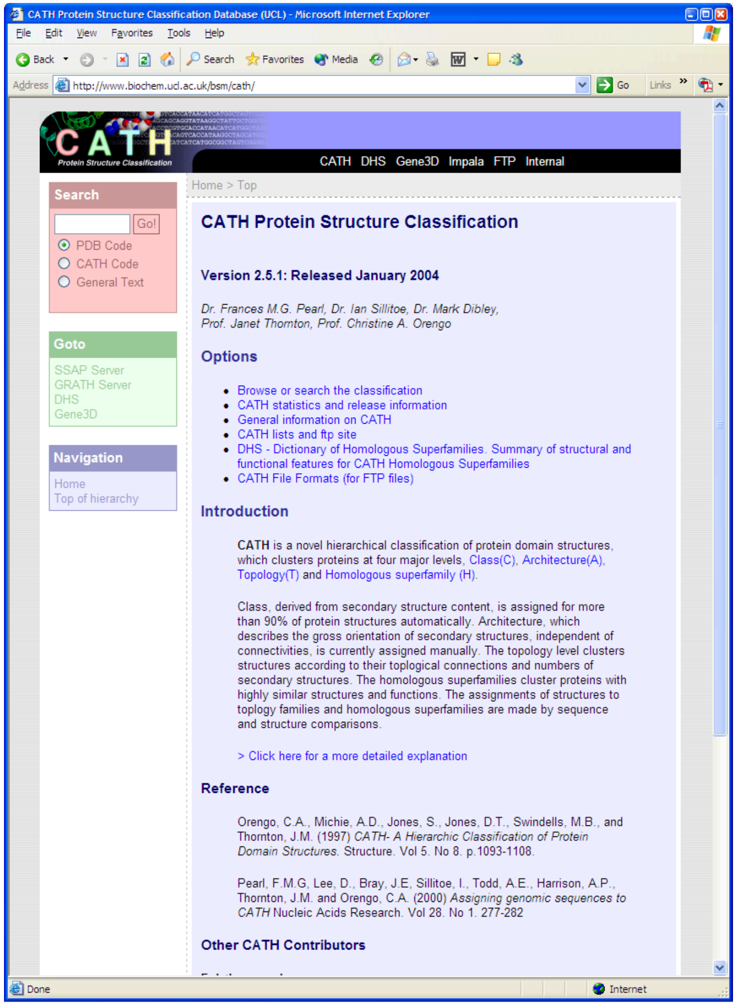
Manually classification
Not 100% up-to-date
Domain boundaries definition

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	179	299	480
All beta proteins	126	248	462
Alpha and beta proteins (a/b)	121	199	542
Alpha and beta proteins (a+b)	234	349	567
Multi-domain proteins	38	38	53
Membrane and cell surface proteins	36	66	73
Small proteins	66	95	150
Total	800	1294	2327

*Murzin A. G.,el at. (1995). J. Mol. Biol. **247**, 536-540.*

CATH$_{2.6.0}$ database

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


CATH Protein Structure Classification Database (UCL) - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.biochem.ucl.ac.uk/bsm/cath/
Go Links
CATH
Protein Structure Classification
CATH DHS Gene3D Impala FTP Internal
Done
Internet

Search

[ ] Go!

● PDB Code
○ CATH Code
○ General Text

Goto

SSAP Server
GRATH Server
DHS
Gene3D

Navigation

Home
Top of hierarchy

Home > Top

CATH Protein Structure Classification

Version 2.5.1: Released January 2004

Dr. Frances M.G. Pearl, Dr. Ian Sillitoe, Dr. Mark Dibley,
Prof. Janet Thornton, Prof. Christine A. Orengo

Options

- Browse or search the classification
- CATH statistics and release information
- General information on CATH
- CATH lists and ftp site
- DHS - Dictionary of Homologous Superfamilies. Summary of structural and functional features for CATH Homologous Superfamilies
- CATH File Formats (for FTP files)

Introduction

CATH is a 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.

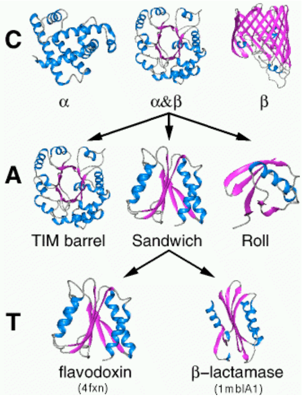
Pearl, F.M.G, Lee, D., Bray, J.E, Sillitoe, I., Todd, A.E., Harrison, A.P., Thornton, J.M. and Orengo, C.A. (2000) *Assigning genomic sequences to CATH* Nucleic Acids Research. Vol 28. No 1. 277-282

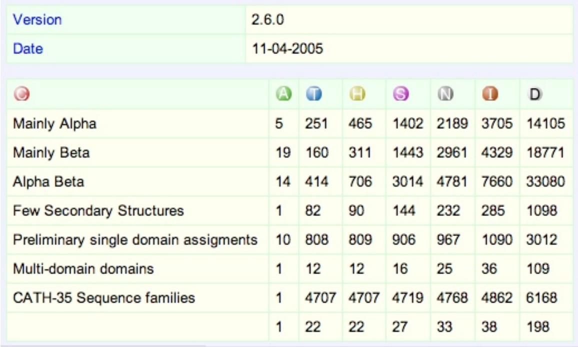
Other CATH Contributors

Uses FSSP for superimposition

✓ Recognized as “standard of gold”
✓ Semi-automatic classification
✓ Clear classification of structures in:
CLASS
ARCHITECTURE
TOPOLOGY
HOMOLOGOUS SUPERFAMILIES
✓ Some large number of tools already available
✓ Easy to navigate

Semi-automatic classification
Domain boundaries definition





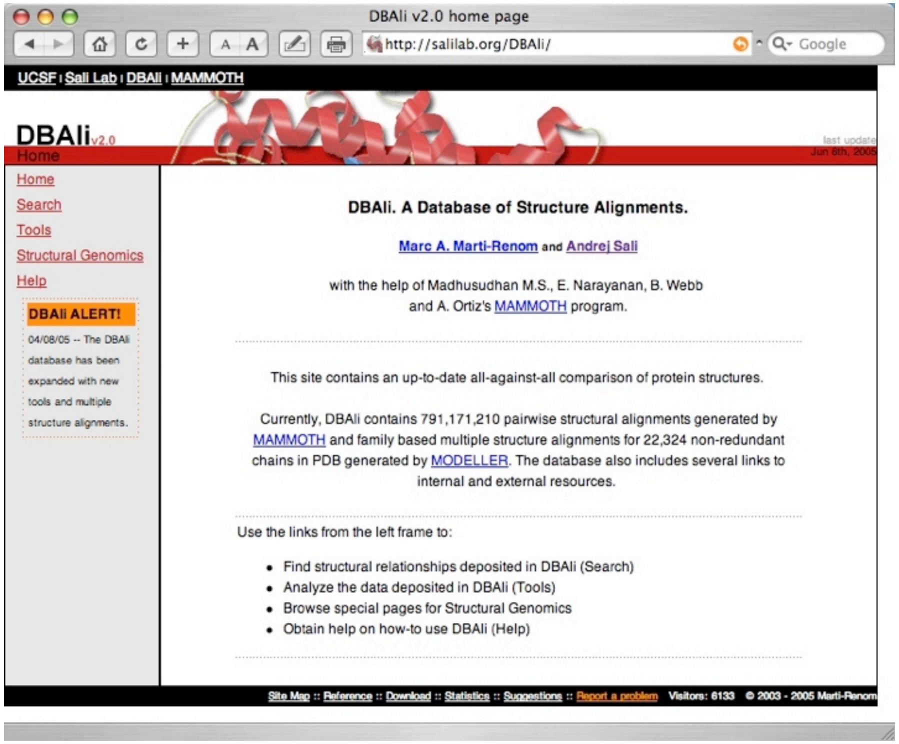
Version	2.6.0
Date	11-04-2005

C	A	T	H	S	N	I	D
Mainly Alpha	5	251	465	1402	2189	3705	14105
Mainly Beta	19	160	311	1443	2961	4329	18771
Alpha Beta	14	414	706	3014	4781	7660	33080
Few Secondary Structures	1	82	90	144	232	285	1098
Preliminary single domain assigments	10	808	809	906	967	1090	3012
Multi-domain domains	1	12	12	16	25	36	109
CATH-35 Sequence families	1	4707	4707	4719	4768	4862	6168
	1	22	22	27	33	38	198

Orengo, C.A., et al. (1997) Structure. ***5****. 1093-1108.*

DBAli$_{v2.0}$ database

http://salilab.org/DBAli/


DBAli v2.0 home page
http://salilab.org/DBAli/
Google
UCSF | Sali Lab | DBAli | MAMMOTH
DBAli v2.0
Home
last update
Jun 6th, 2005
Home
Search
Tools
Structural Genomics
Help
Site Map :: Reference :: Download :: Statistics :: Suggestions :: Report a problem Visitors: 6133 © 2003 - 2005 Marti-Renom

DBAli ALERT!

04/08/05 -- The DBAli database has been expanded with new tools and multiple structure alignments.

DBAli. A Database of Structure Alignments.

Marc A. Marti-Renom and Andrej Sali

with the help of Madhusudhan M.S., E. Narayanan, B. Webb
and A. Ortiz's MAMMOTH program.

This site contains an up-to-date all-against-all comparison of protein structures.

Currently, DBAli contains 791,171,210 pairwise structural alignments generated by MAMMOTH and family based multiple structure alignments for 22,324 non-redundant chains in PDB generated by MODELLER. The database also includes several links to internal and external resources.

Use the links from the left frame to:

- Find structural relationships deposited in DBAli (Search)
- Analyze the data deposited in DBAli (Tools)
- Browse special pages for Structural Genomics
- Obtain help on how-to use DBAli (Help)

Uses MAMMOTH for superimposition

- ✓ Fully-automatic
- ✓ Data is kept up-to-date with PDB releases
- ✓ Tools for “on the fly” classification of families
- ✓ Up-to-date multiple structure alignments
- ✓ Easy to navigate
- ✓ Provides some tools for structure comparison

Does not provide (yet) a stable classification

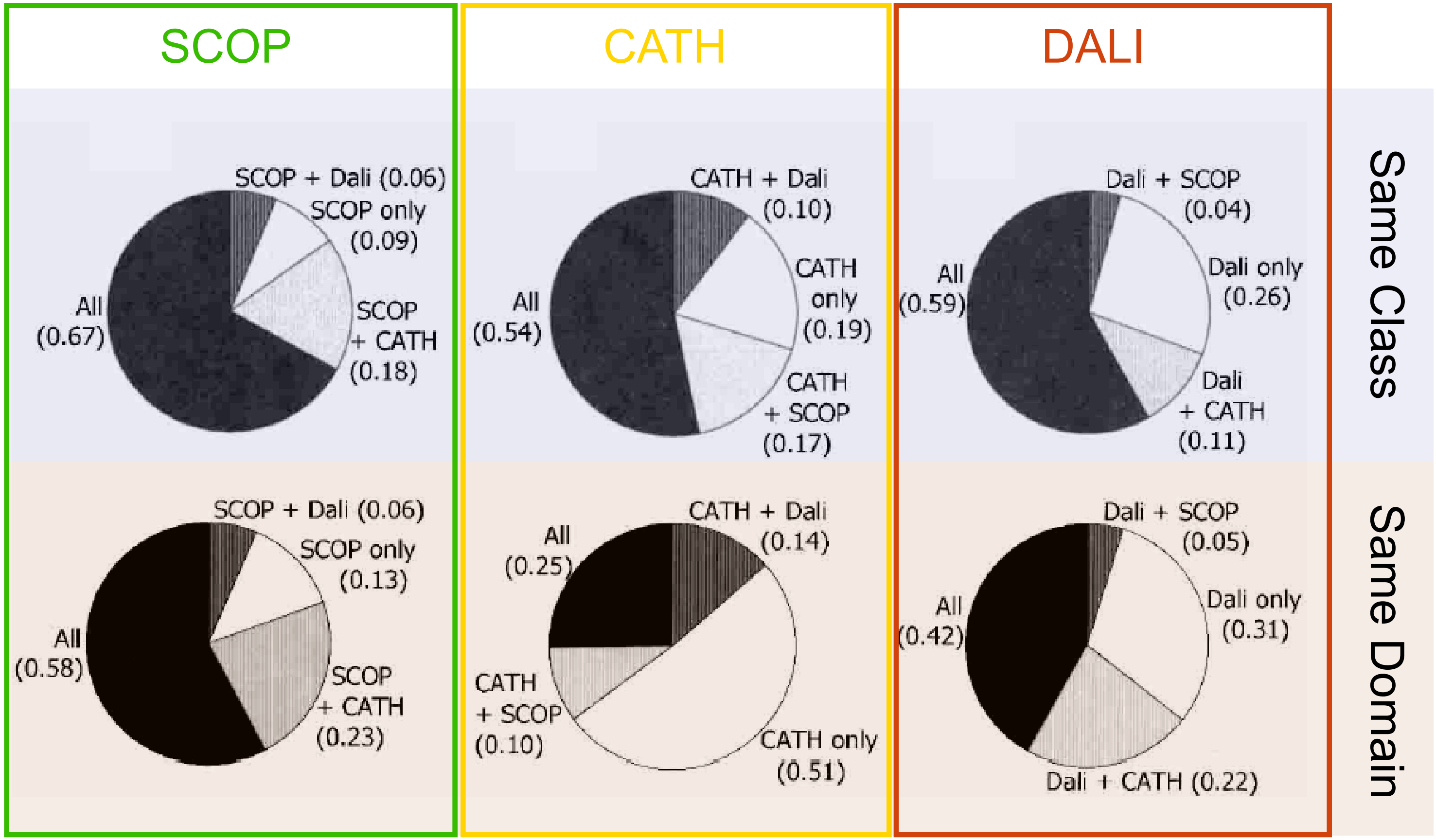
	Pairwise structure alignments
Last update:	June 6th, 2005
Number of chains:	65,286
Number of structure-structure comparisons:*	791,171,210
	Multiple structure alignments
Last update:	May 14th, 2005
Number of representative chains:	22,324
Number of families:	8,737

Marti-Renom et al. 2001. Bioinformatics. ***17****, 746*

Classification of the structural space

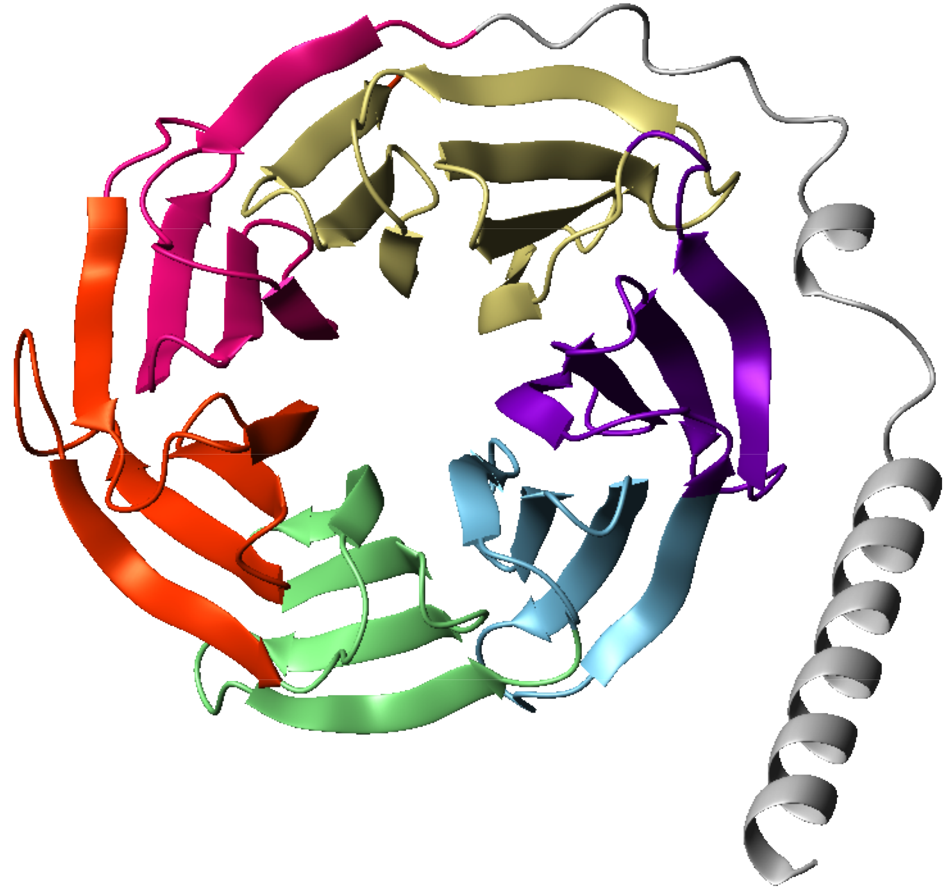
Not an easy task!

Domain definition AND domain classification

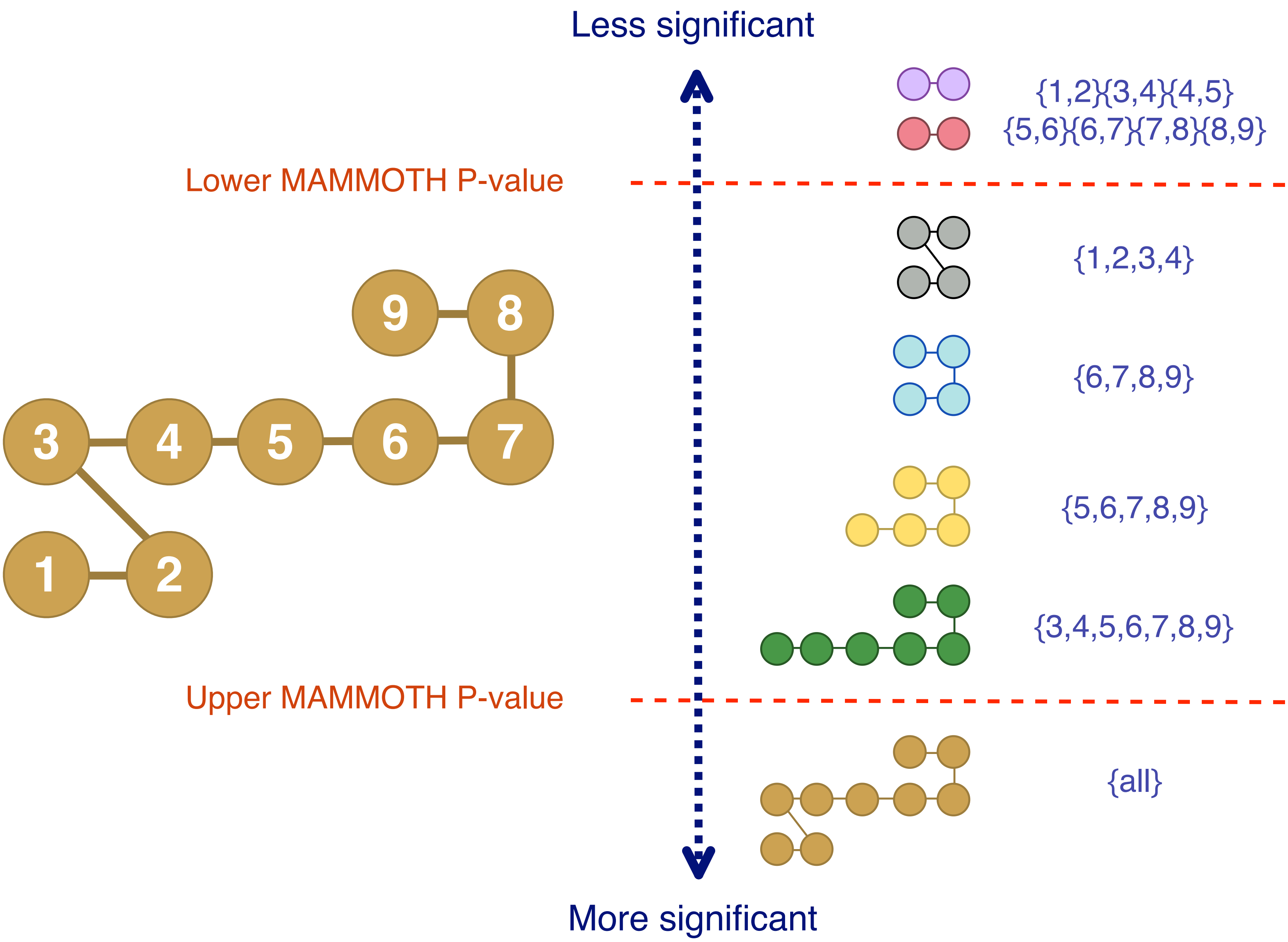


*Day, et al. (2003) Protein Sciences, **12** pp2150*

Application (ModDom)



Assigning domains from structure



Less significant

{1,2}{3,4}{4,5}

{5,6}{6,7}{7,8}{8,9}

Lower MAMMOTH P-value

{1,2,3,4}

{6,7,8,9}

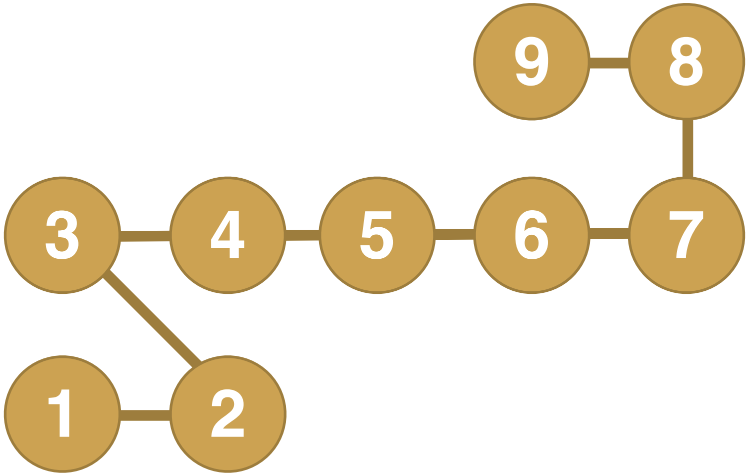
{5,6,7,8,9}

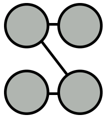
{3,4,5,6,7,8,9}

Upper MAMMOTH P-value

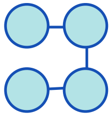
{all}

More significant

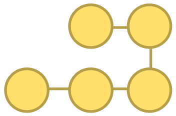




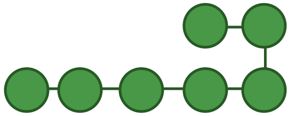
{1,2,3,4}



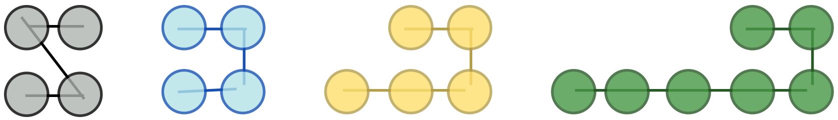
{6,7,8,9}

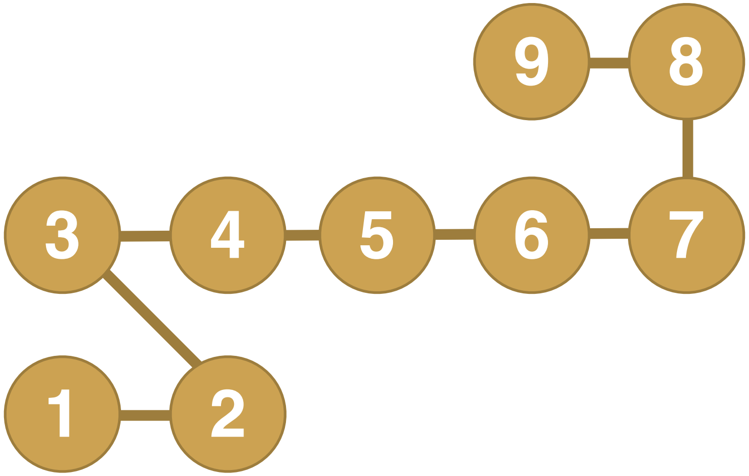


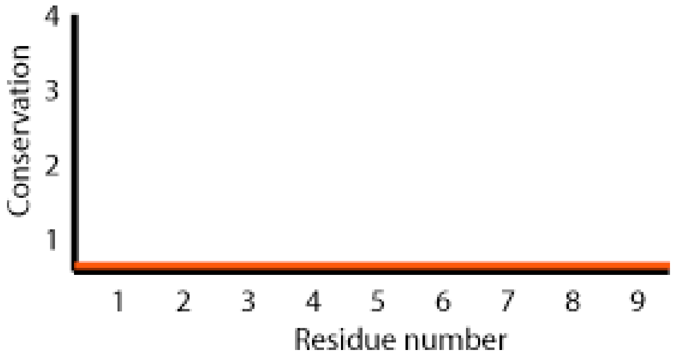
{5,6,7,8,9}



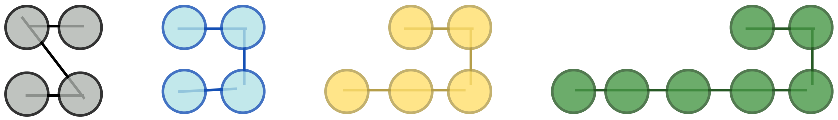
{3,4,5,6,7,8,9}

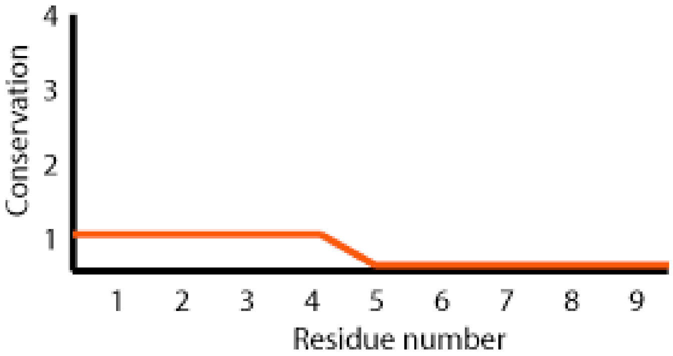


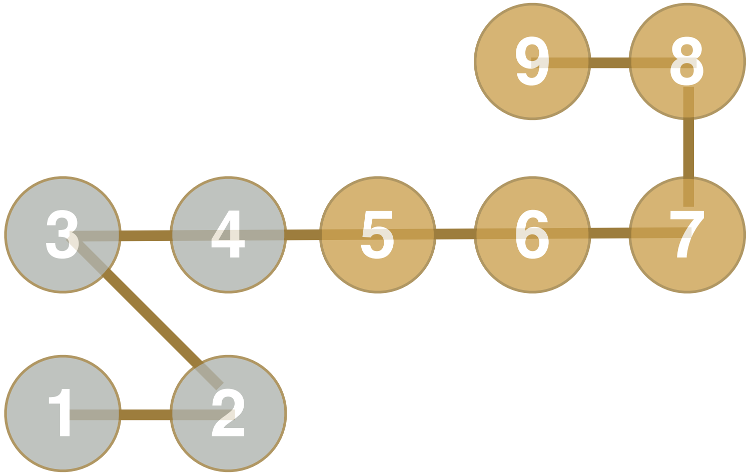




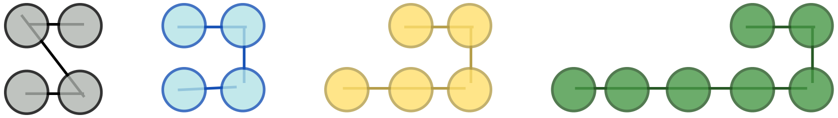
#	1	2	3	4	5	6	7	8	9
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0

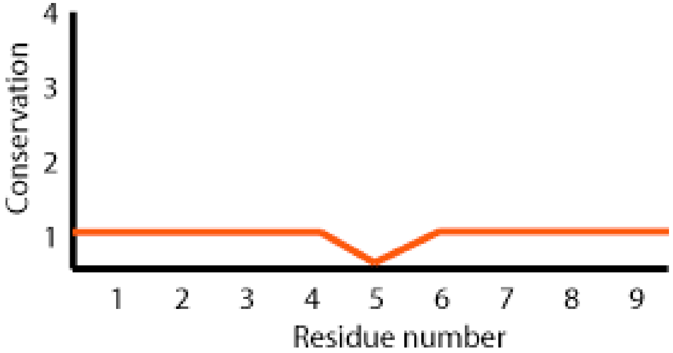


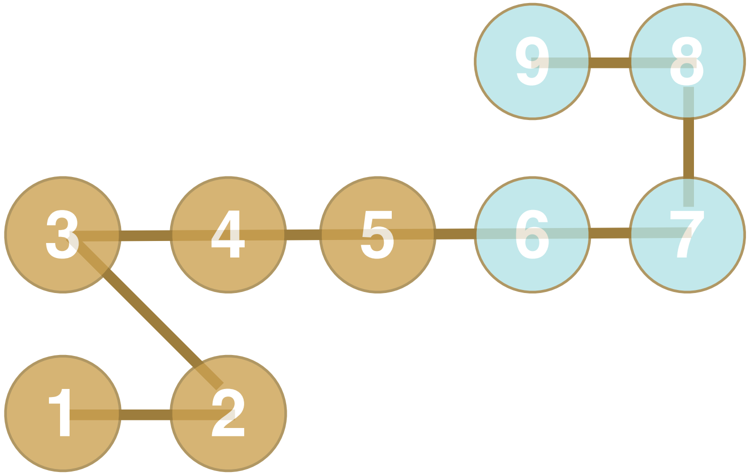




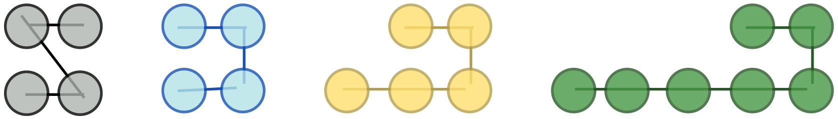
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0

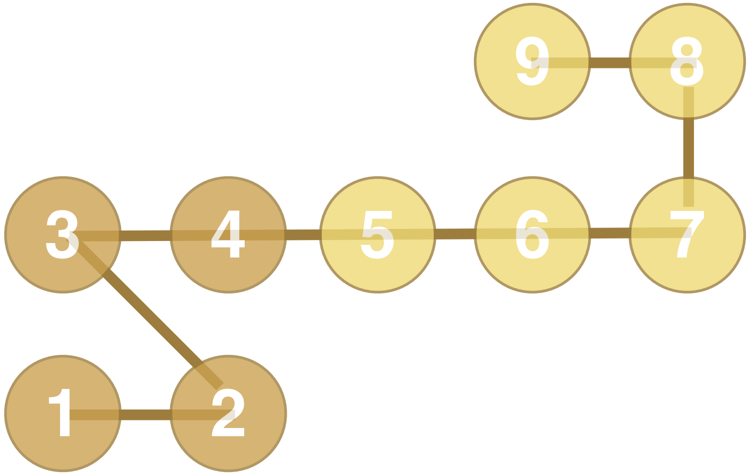


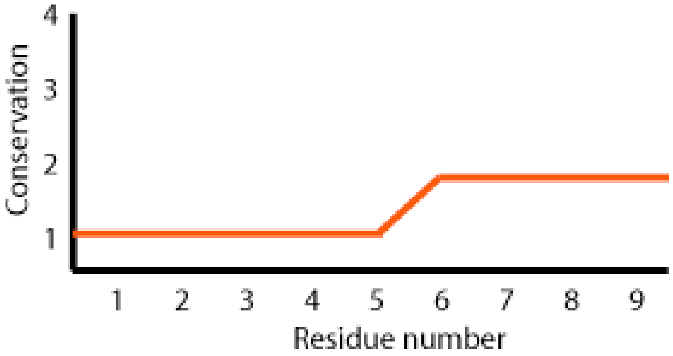




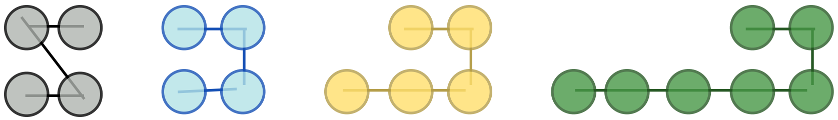
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	1	1	1
7	0	0	0	0	0	1	1	1	1
8	0	0	0	0	0	1	1	1	1
9	0	0	0	0	0	1	1	1	1

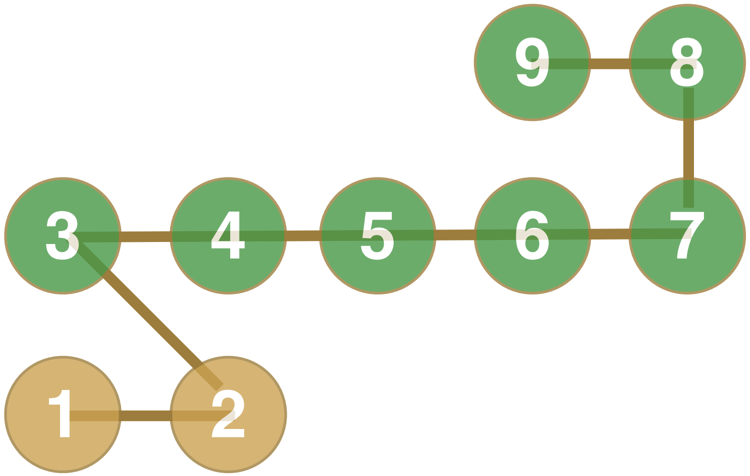


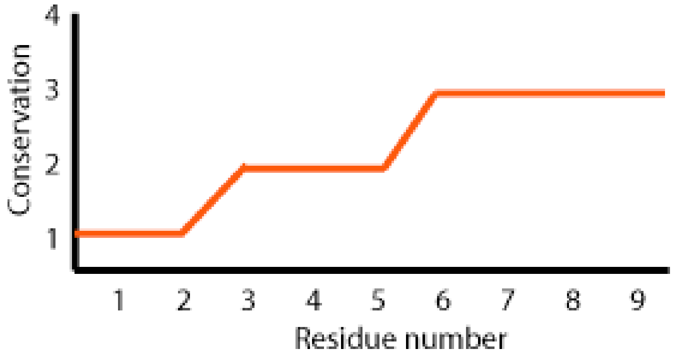




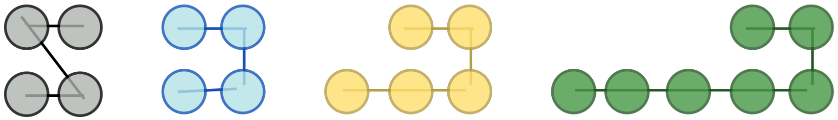
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	1	1	1	1	1
6	0	0	0	0	1	2	2	2	2
7	0	0	0	0	1	2	2	2	2
8	0	0	0	0	1	2	2	2	2
9	0	0	0	0	1	2	2	2	2

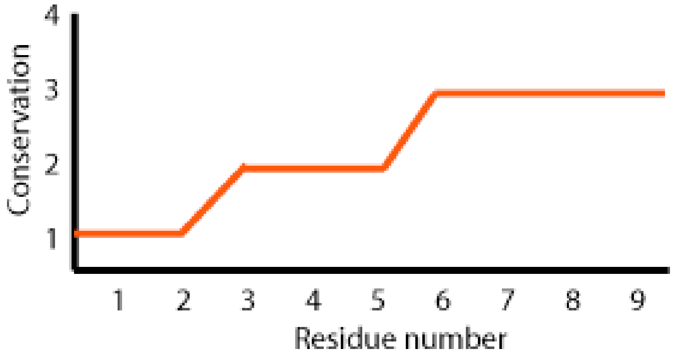


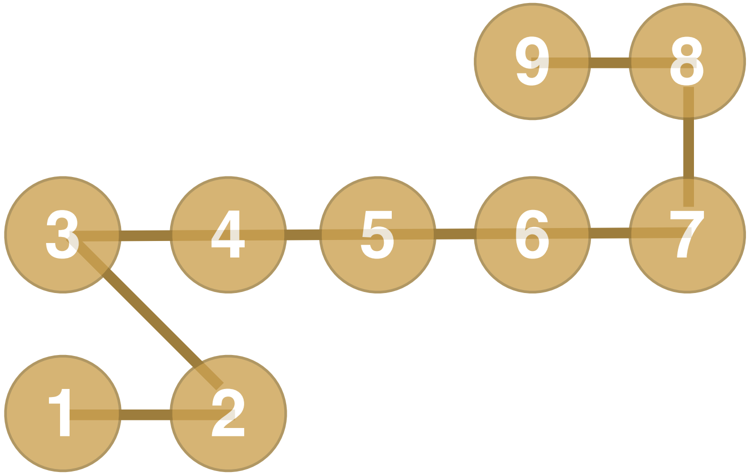




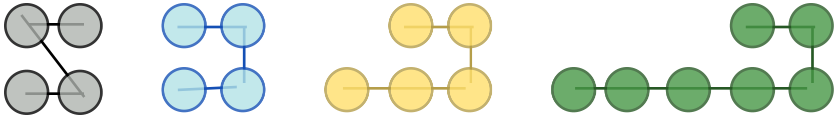
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	2	2	1	1	1	1	1
4	1	1	2	2	1	1	1	1	1
5	0	0	1	1	2	2	2	2	2
6	0	0	1	1	2	3	3	3	3
7	0	0	1	1	2	3	3	3	3
8	0	0	1	1	2	3	3	3	3
9	0	0	1	1	2	3	3	3	3

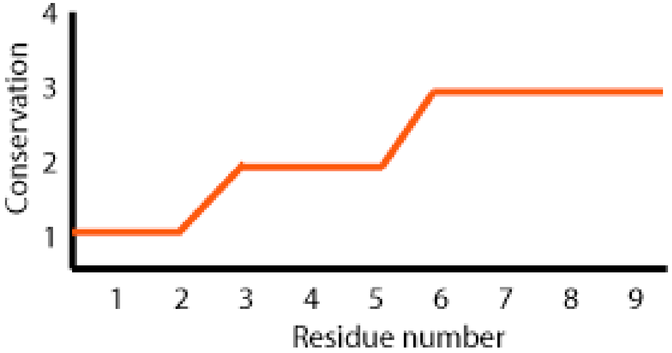




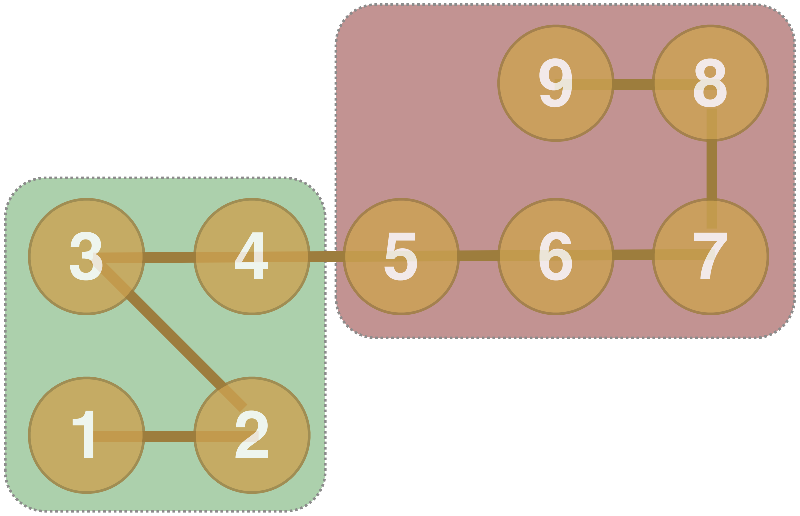


#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	2	2	1	1	1	1	1
4	1	1	2	2	1	1	1	1	1
5	0	0	1	1	2	2	2	2	2
6	0	0	1	1	2	3	3	3	3
7	0	0	1	1	2	3	3	3	3
8	0	0	1	1	2	3	3	3	3
9	0	0	1	1	2	3	3	3	3



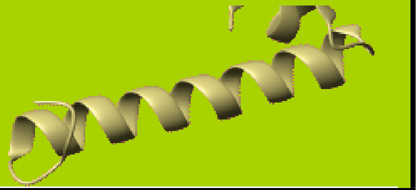


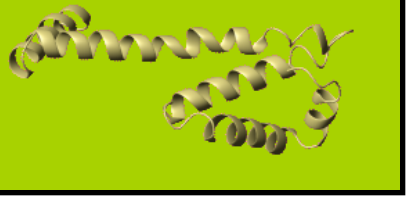
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	2	2	1	1	1	1	1
4	1	1	2	2	1	1	1	1	1
5	0	0	1	1	2	2	2	2	2
6	0	0	1	1	2	3	3	3	3
7	0	0	1	1	2	3	3	3	3
8	0	0	1	1	2	3	3	3	3
9	0	0	1	1	2	3	3	3	3

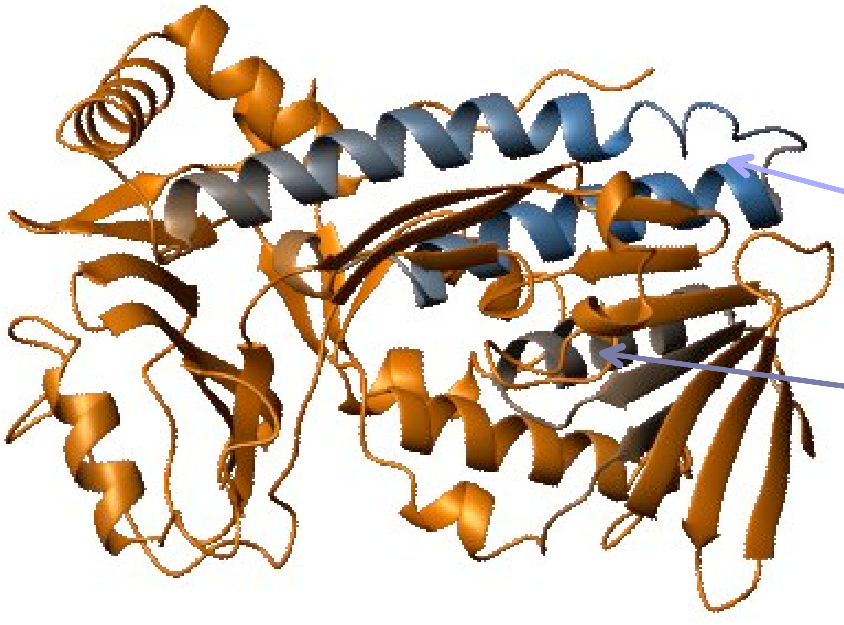


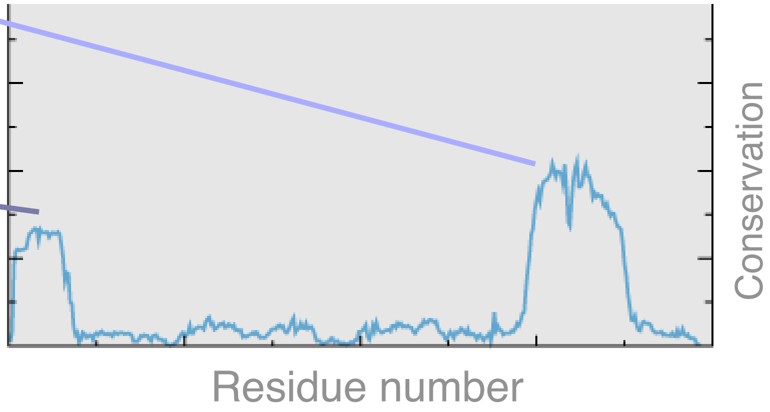
Threshold #3 MCL Cluster level (-I)

Stijn van Dongen (`http://micans.org/mcl/`)

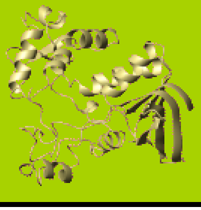
1phh	290-329	2.7Å	3.1
1hadB	72-111		

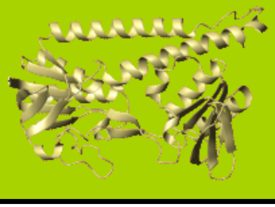
1phh	279-373	3.9Å	4.7
1bke	310-410		

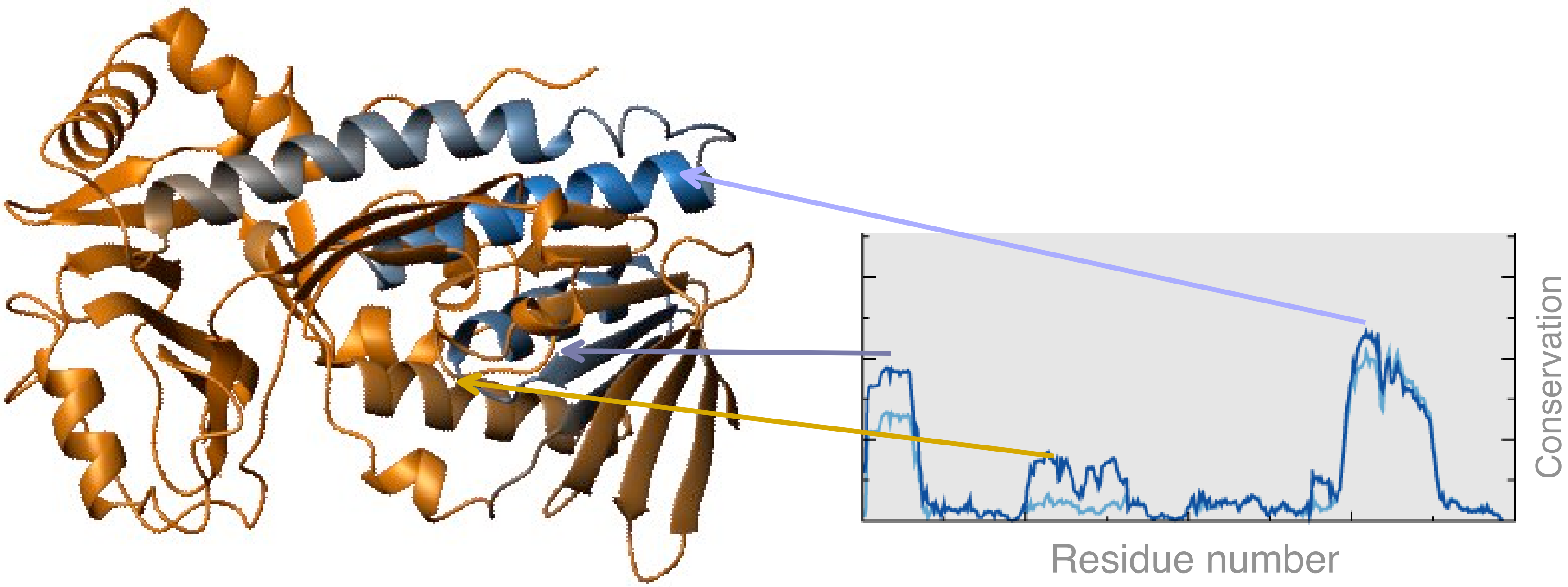




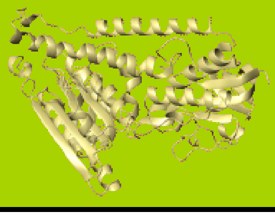
1phh (Oxydoreductase from Pseudomonas fluorescens)

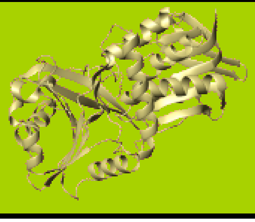
1phh	1-213	3.0Å	8.1
1qjdA	125-379		

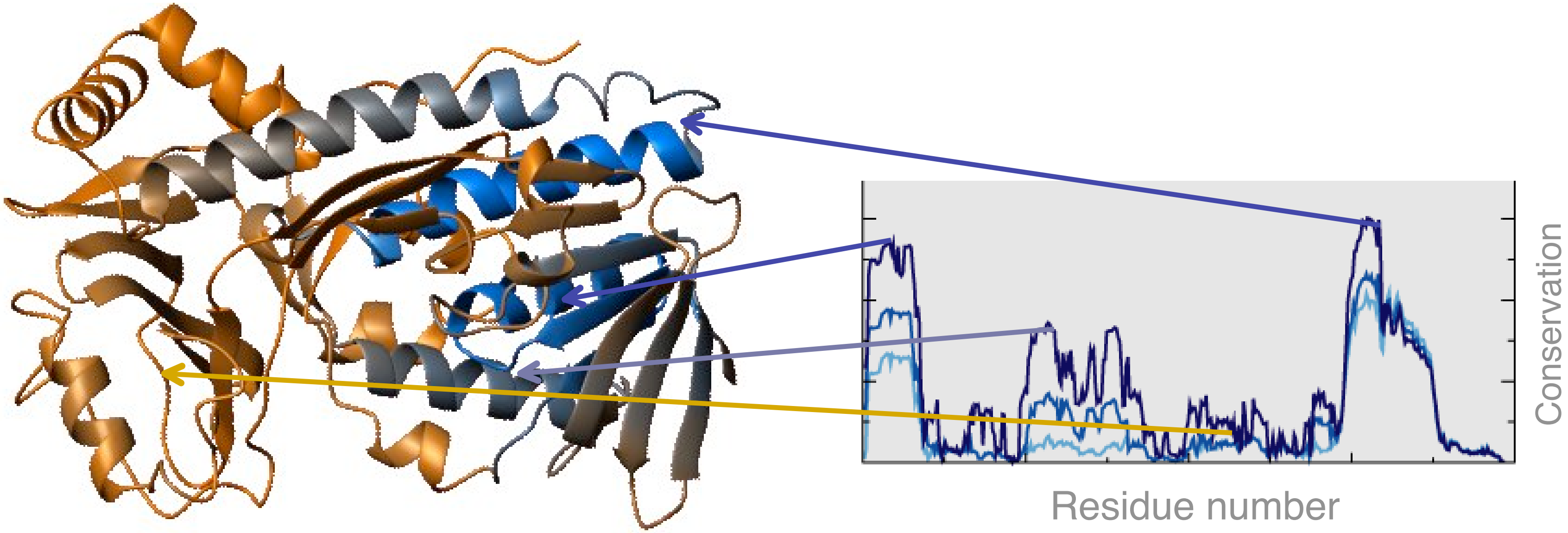
1phh	1-319	3.6Å	9.8
1gerA	3-327		



1phh (Oxydoreductase from Pseudomonas fluorescens)

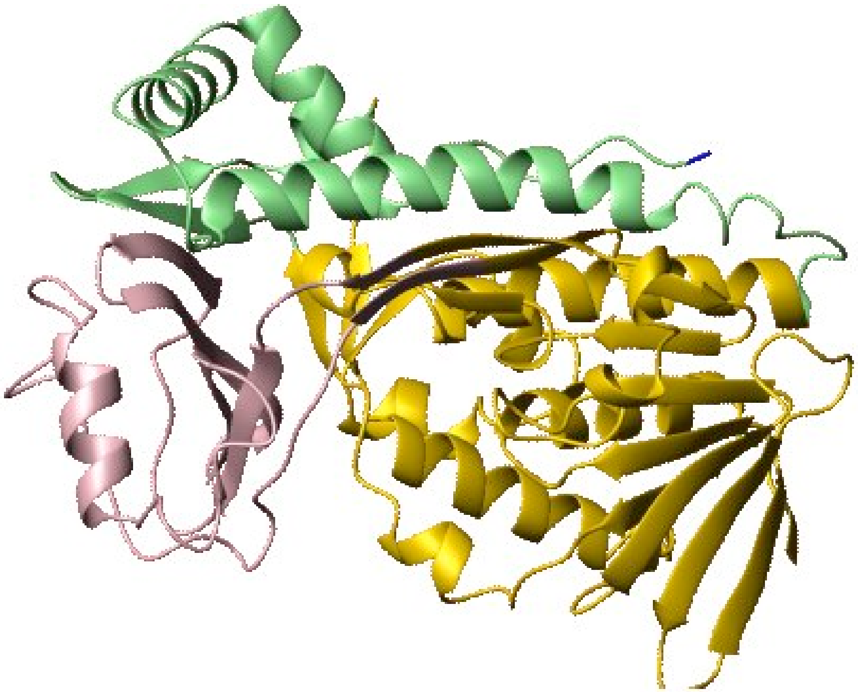
1phh	1-378	3.8Å	10.3
1feaC	2-464		

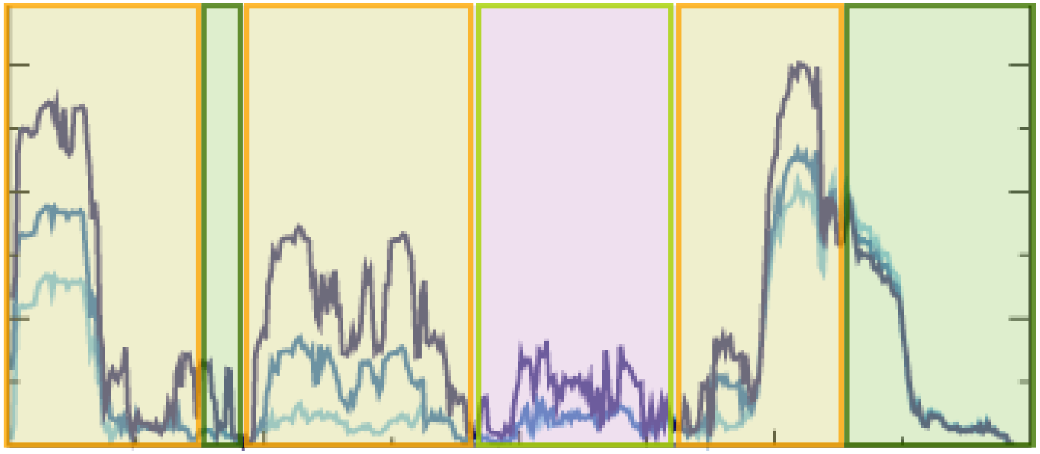
1phh	1-316	3.8Å	17.2
1l9dB	2-364		



1phh (Oxydoreductase from Pseudomonas fluorescens)

1phh (Oxydoreductase from Pseudomonas fluorescens)





Benchmark dataset

2163 chains from Islam et al. 1995 → 569 Non-redundant

<2Å && <30aa diff.

Divide randomly into two sets

Remove of incomplete or obsolete entries.

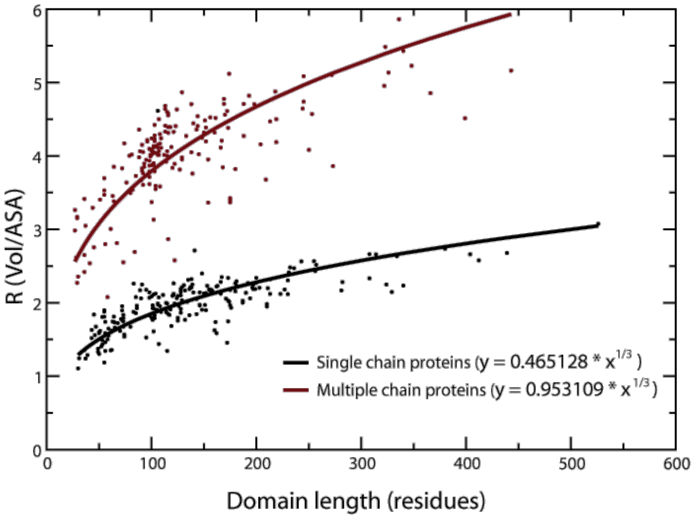
FINAL:

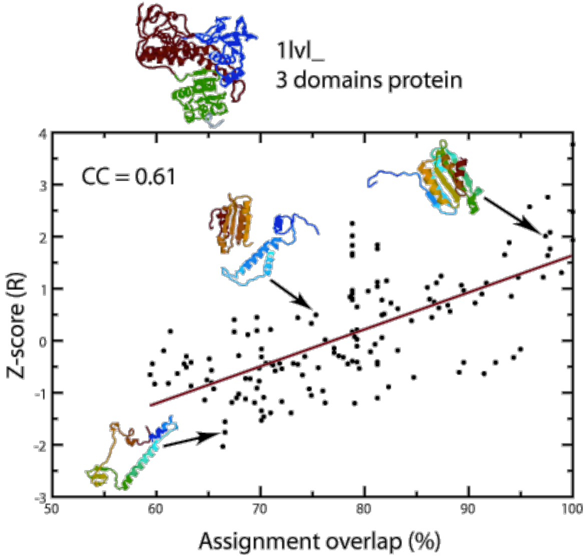
Training set → 242 chains

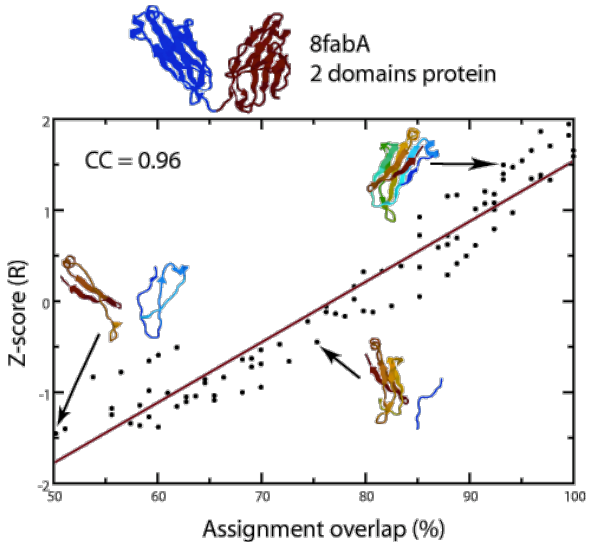
Testing set → 234 chains

Scoring function

R = Volume/ASA



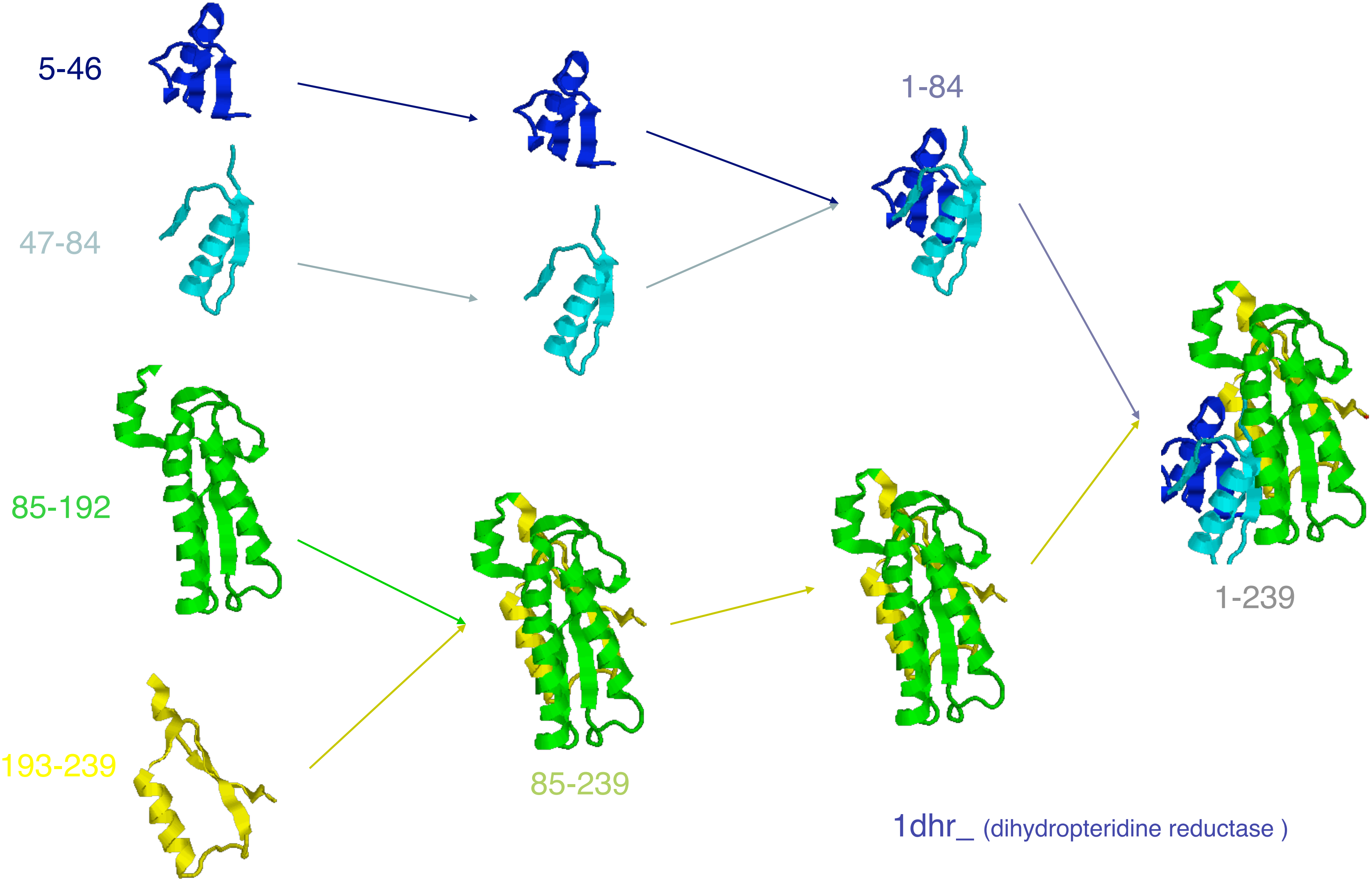




Domain → max(<dist f(R)>)

<dist to f(R)>

-0.11 -0.10 -0.08 -0.09



5-46

47-84

85-192

193-239

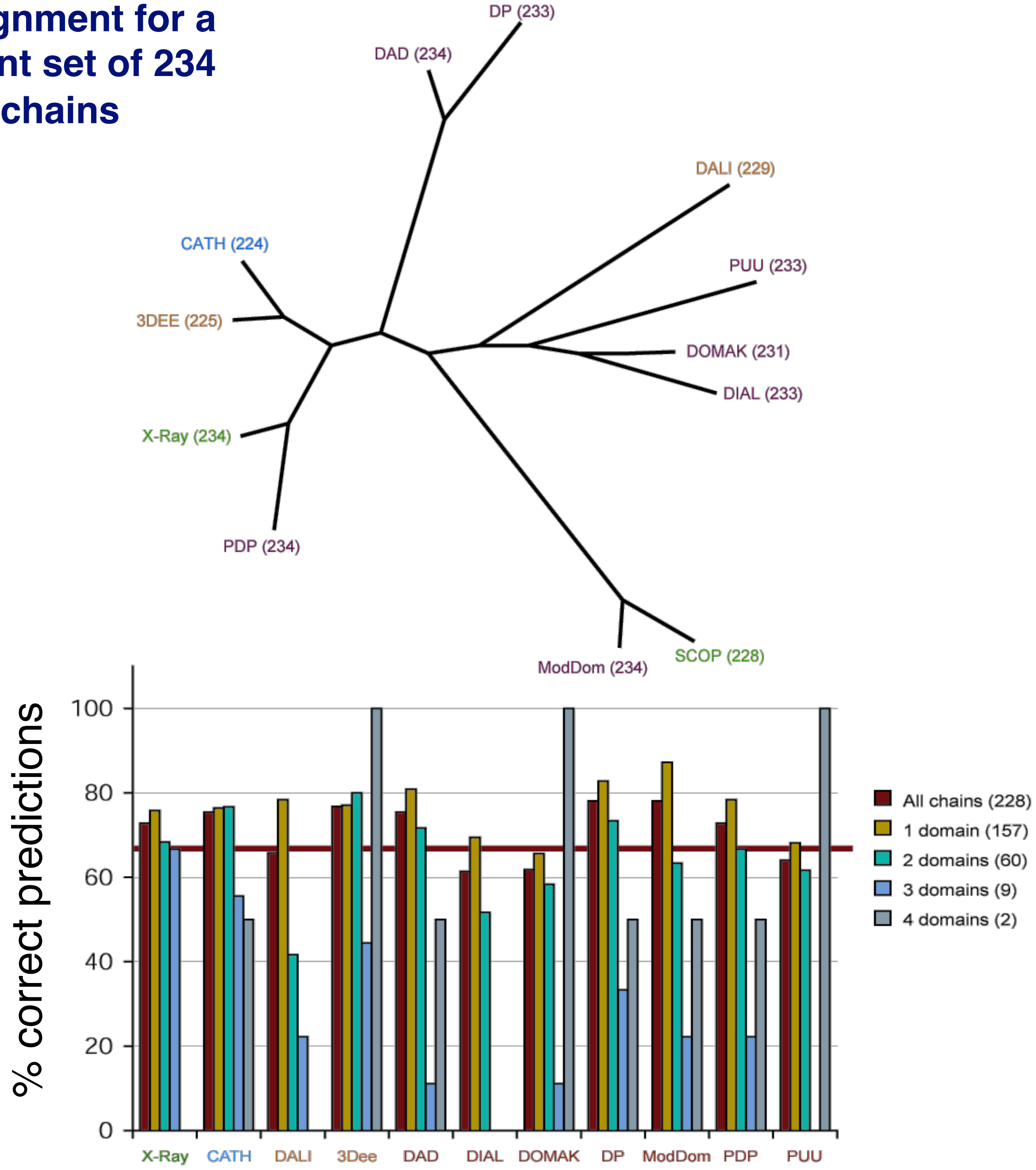
85-239

1-84

1-239

1dhr_ (dihydropteridine reductase)

Domain assignment for a non-redundant set of 234 protein chains

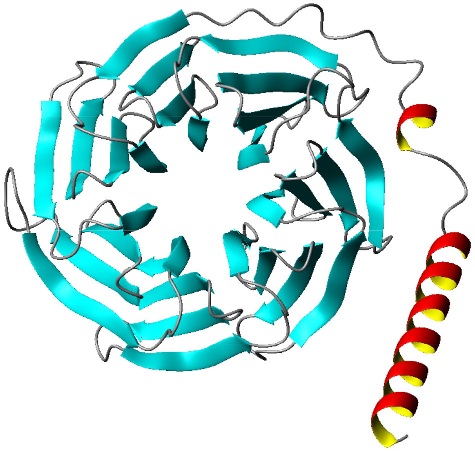


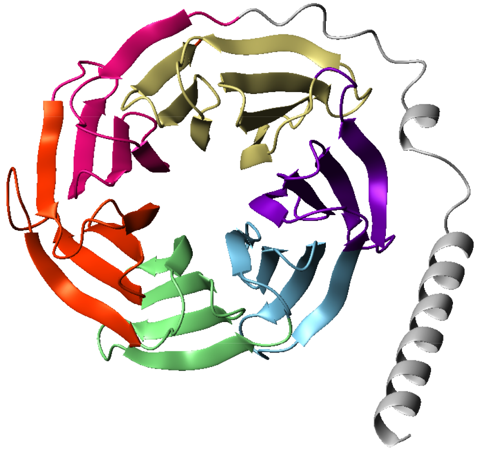
What are domains?

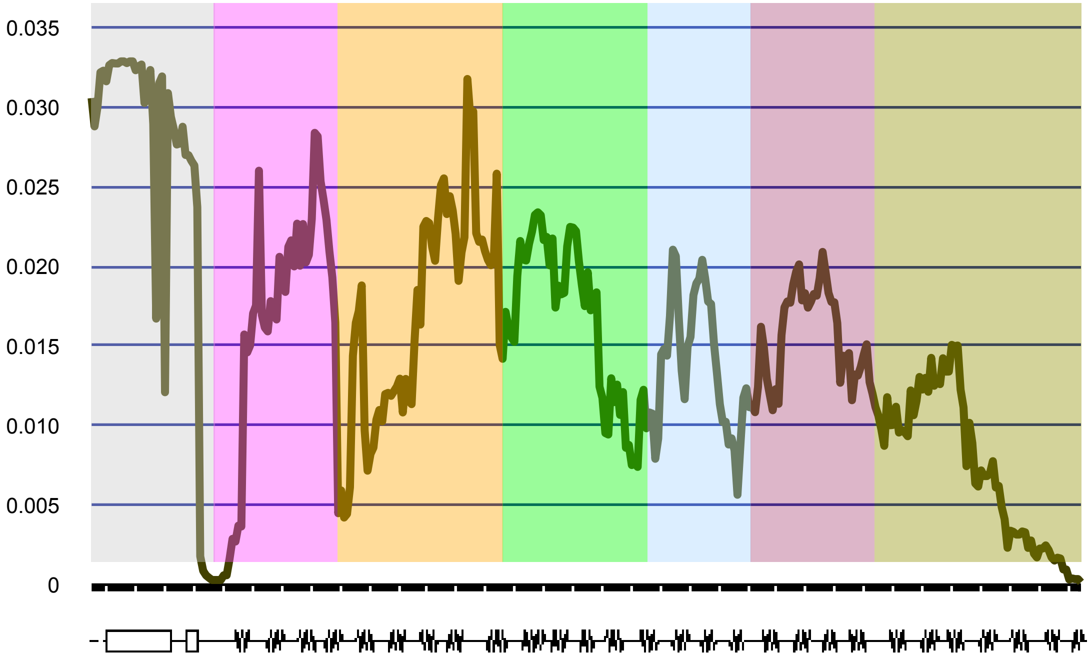
↕

Structural recurrent fragments

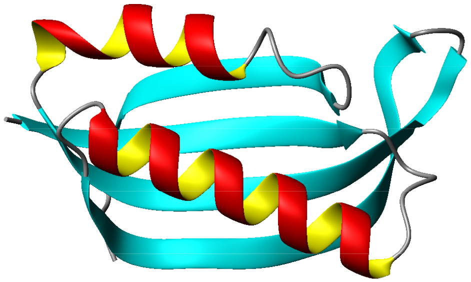
G-protein (1gotB) *all-β* → *7 bladed beta propeller domain*

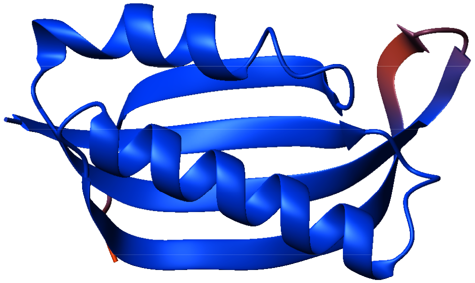


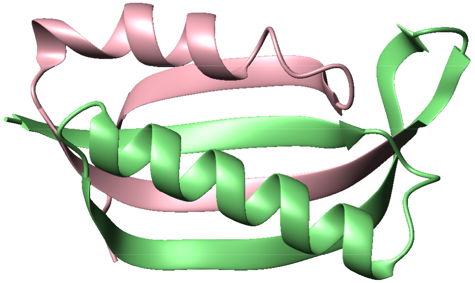


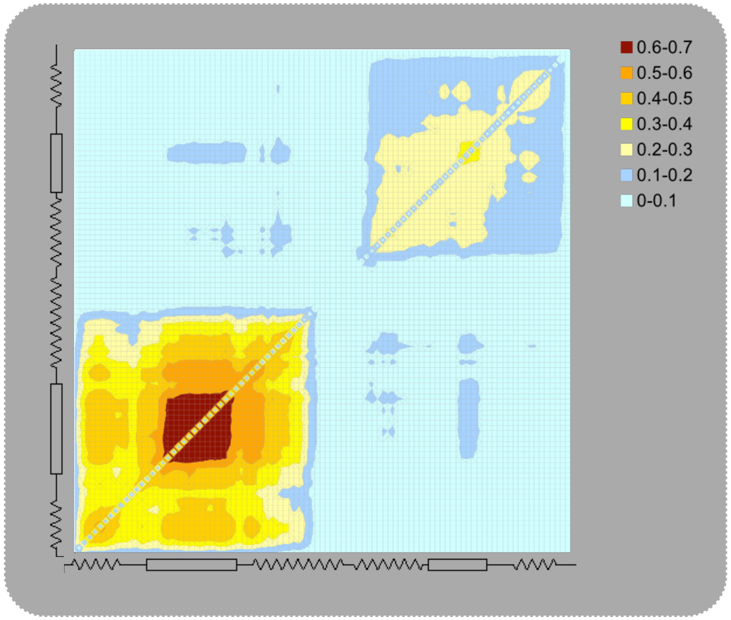


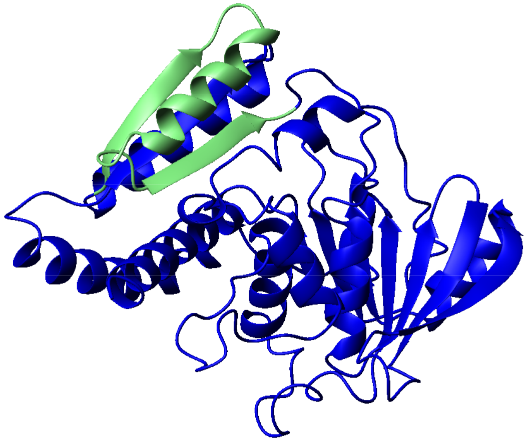
Ribosomal protein S6 (1ris) α+β → *Ferrodoxin Like domain*



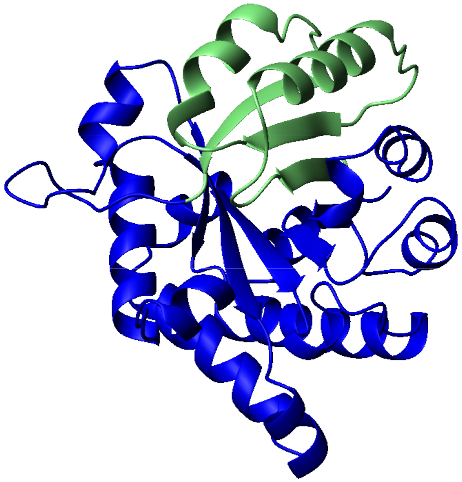






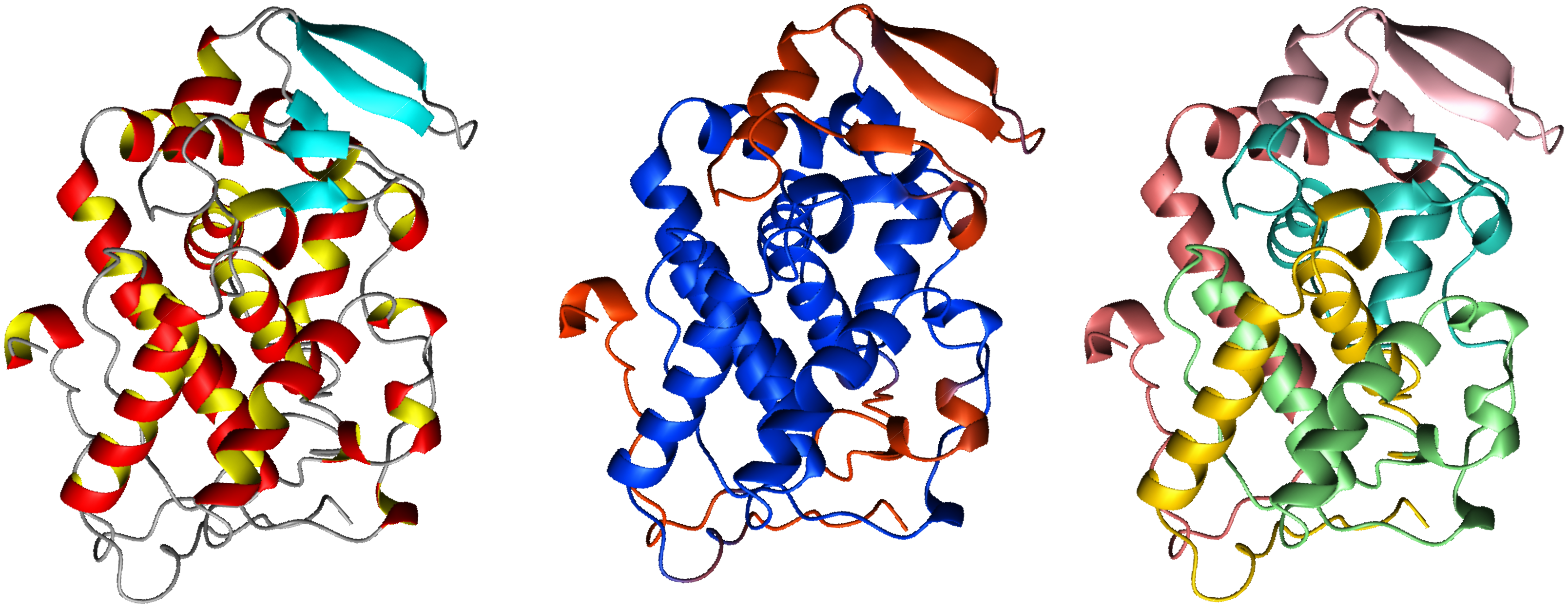


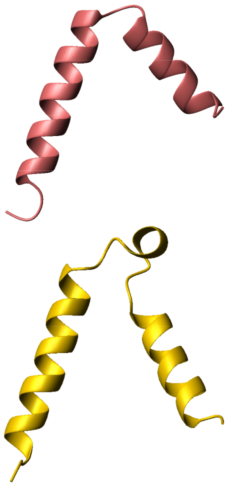
1ee9A 17.9% id. 2.3Å



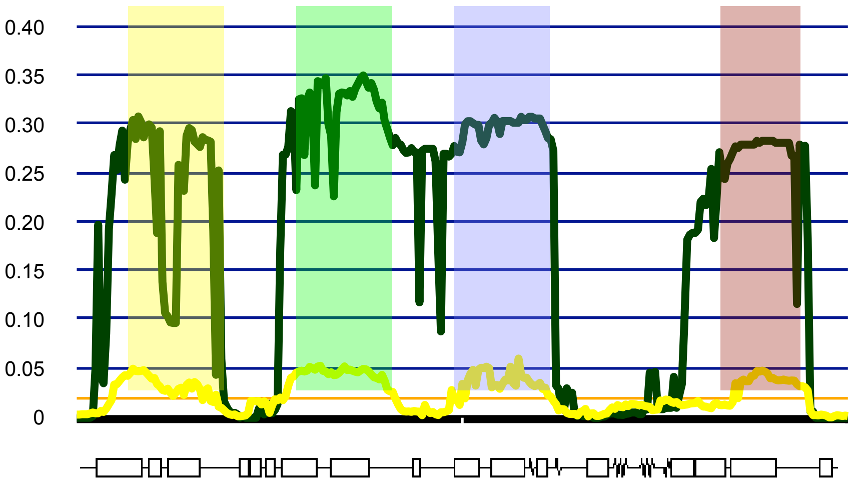
6timB 11.1% id. 2.6Å

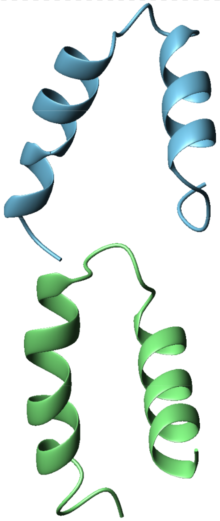
Cytochrome C Peroxidase (2cyp) *all-α → CCP-like domain*





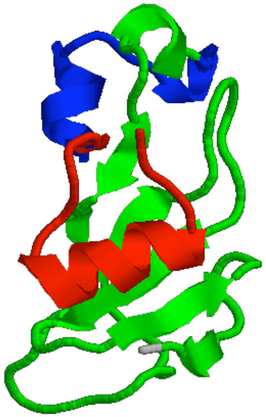
27/35 3.2Å

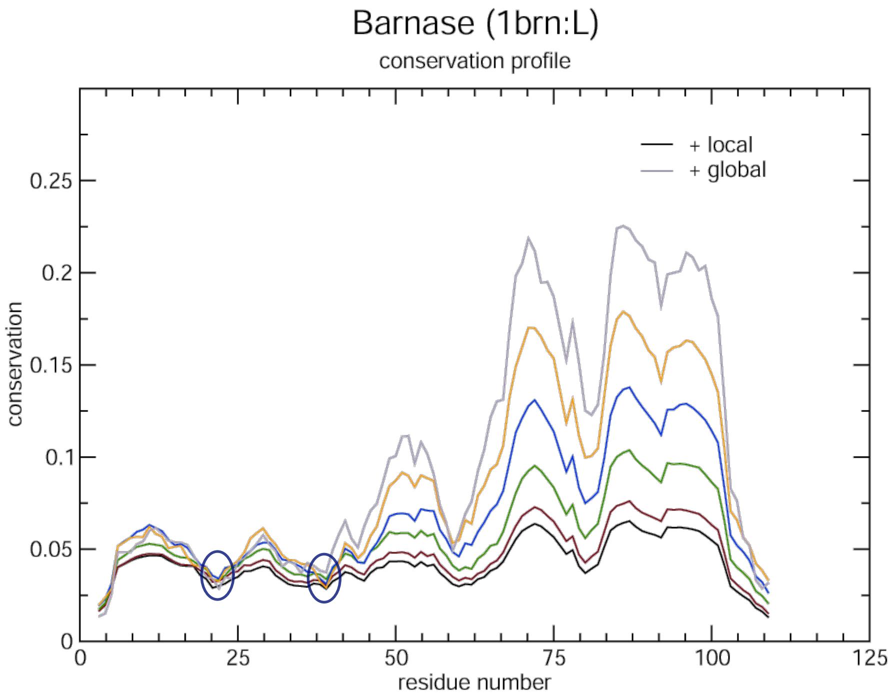


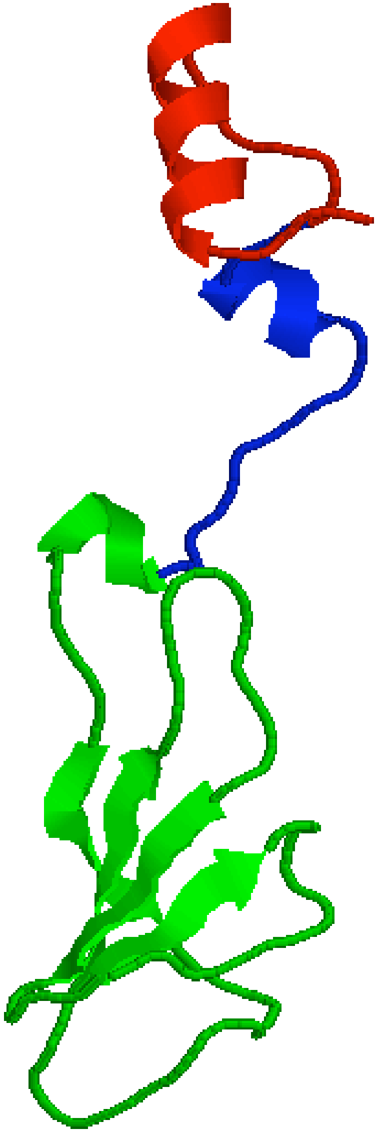


29/34 2.9Å

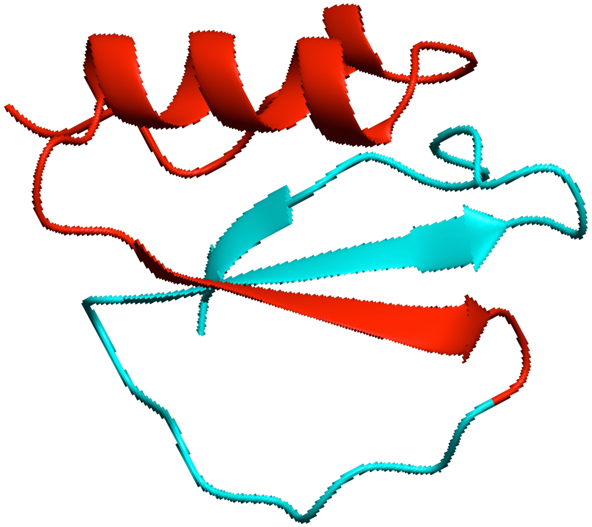
Barnase Domain-Swapping



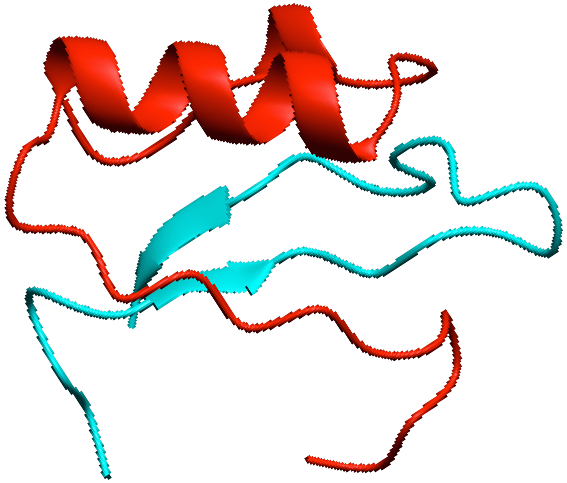




chymotrypsin inhibitor 2



1-37 | 38-64

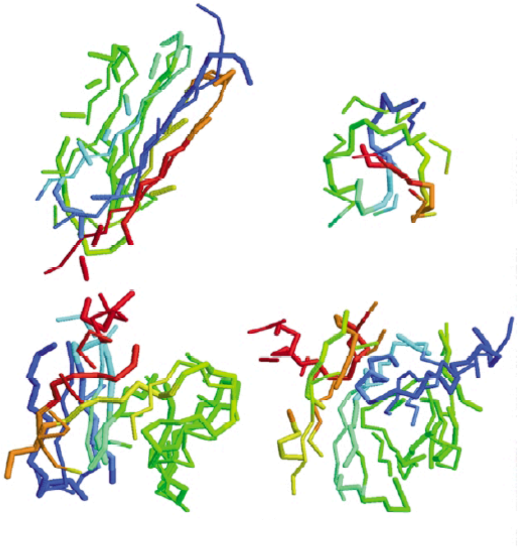


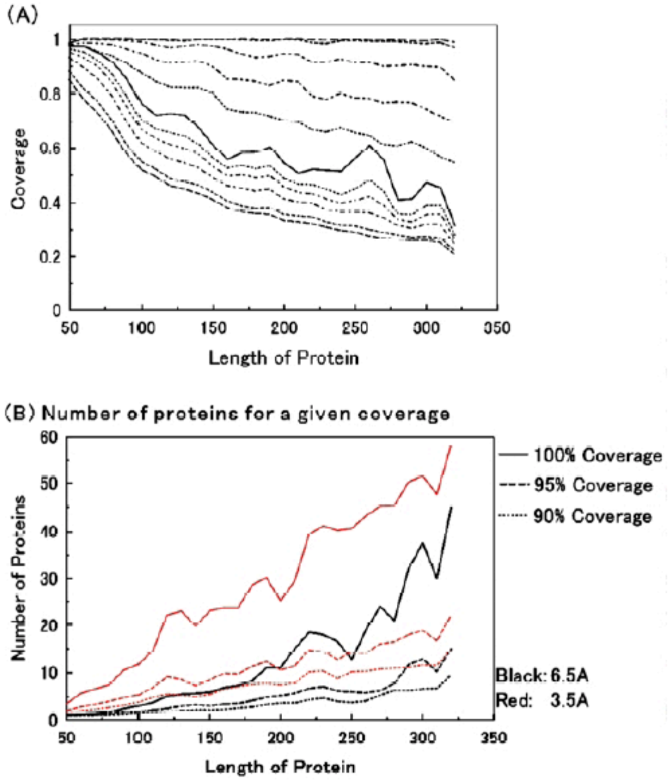
1-40 | 41-64

Neira JL, Davis B, Ladurner AG, Buckle AM, Gay GP, Fersht AR. 1996. Fold Des 1:189-208.
Ladurner AG, Itzhaki LS, de Prat GG, Fersht AR. 1997. J Mol Biol 273:317-329.

Sequence space .vs. Structure space

The PDB is a covering set of small protein structures.

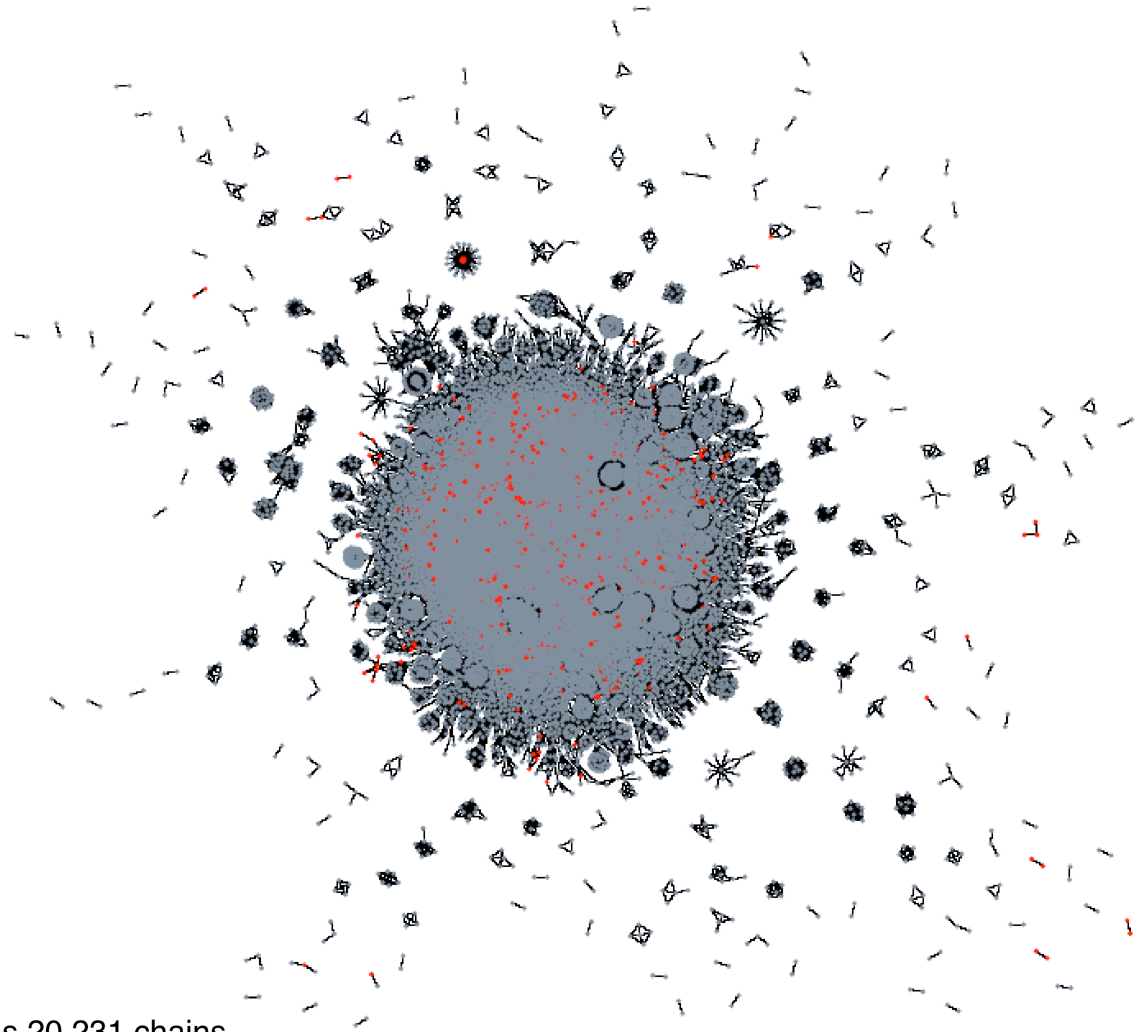




Kihara D, Skolnick J. (2003) *J Mol Biol*. **334** pp793

Sequence space .vs. Structure space

Structure map @ >20% sequence identity



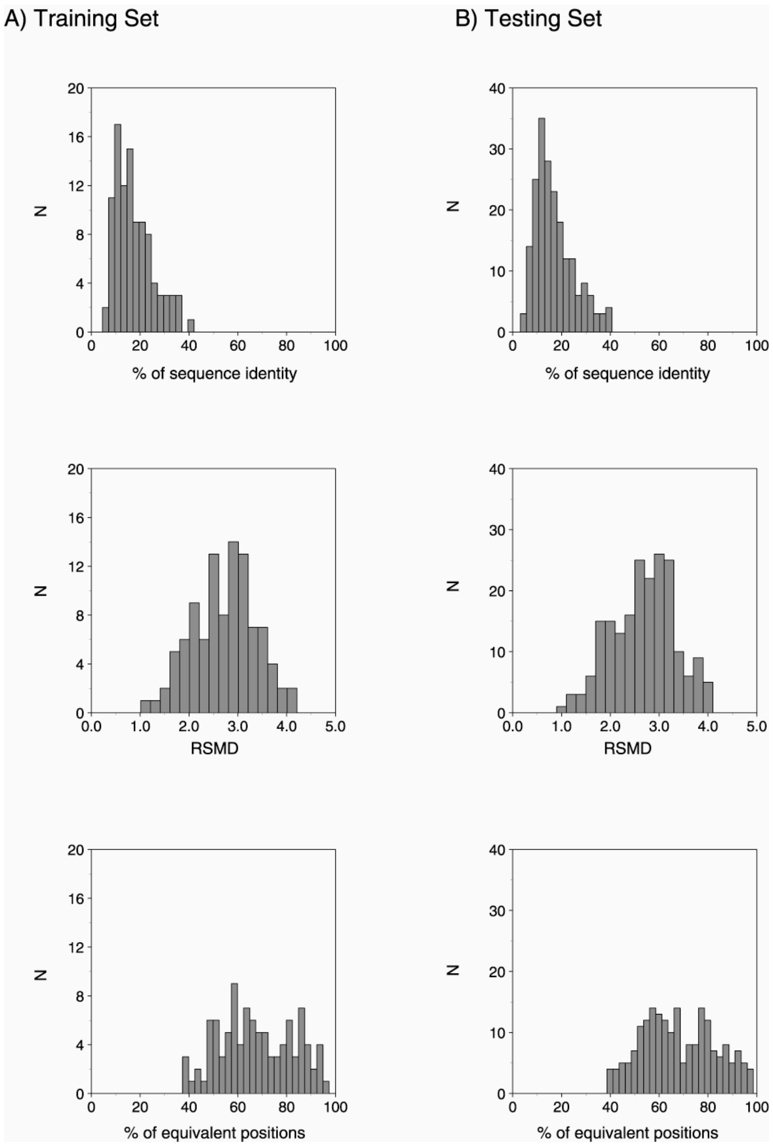
~22,000 nodes
186 clusters
Larger cluster contains 20,231 chains

Sequence space .vs. Structure space

SALIGN

aligning profiles (PP_SCAN in MODELLER)

Marti-Renom, et al. (2004) Prot. Sci. 13 pp1071
Narayanan, et al. in prepration



Seq.-Seq.

ALIGN: DP pairwise method

BLAST2SEQ: Local heuristic method

Seq.-Str.

SEA: Local structure prediction method

Prof.-Seq.

SAM: HMM method

PSI-BLAST: Local search method that uses multiple sequence information for one of the sequences.

LOBSTER: HHM + Phylogeny Method

Prof.-Prof.

CLUSTALW: DP multiple sequence method.

COMPASS: DP profile-profile method

SALIGN: DP pairwise method that uses multiple sequence information for both sequences.

SALIGN protocols

Profile generation

- PSI-Blast (PBP)
- Henikoff & Henikoff (HH)
- Henikoff & Henikoff + Similarity (HS)
- Henikoff & Henikoff substitution matrix (MAT)

Profile comparison

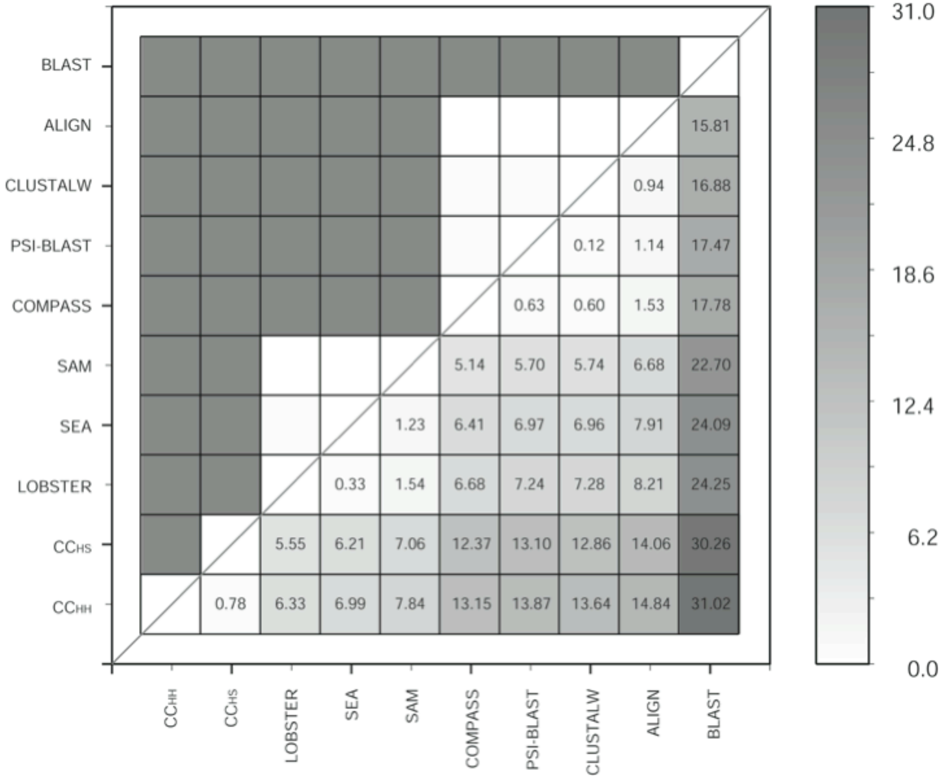
- Correlation coefficient (CC)
- Euclidean distance (ED)
- Dot product (DP)
- Jensen-Shannon distance (JS)
- Average value (Ave)

SALIGN protocols accuracy

SALIGN protocol	CE overlap [%]	Shift score
CCPBP	55 ± 23	0.61 ± 0.24
CCHH	**56 ± 23**	**0.61 ± 0.24**
CCHS	**56 ± 24**	**0.62 ± 0.23**
CCMAT	51 ± 25	0.55 ± 0.27
EDPBP	54 ± 24	0.60 ± 0.25
EDHH	54 ± 24	0.59 ± 0.26
EDHS	55 ± 24	0.59 ± 0.26
DPPBP	55 ± 23	0.61 ± 0.24
DPHH	56 ± 23	0.60 ± 0.25
DPHS	55 ± 24	0.61 ± 0.24
JSHH	53 ± 24	0.60 ± 0.24
JSHS	54 ± 24	0.60 ± 0.24
AveMAT	49 ± 26	0.52 ± 0.29
TOP	62 ± 20	0.67 ± 0.20

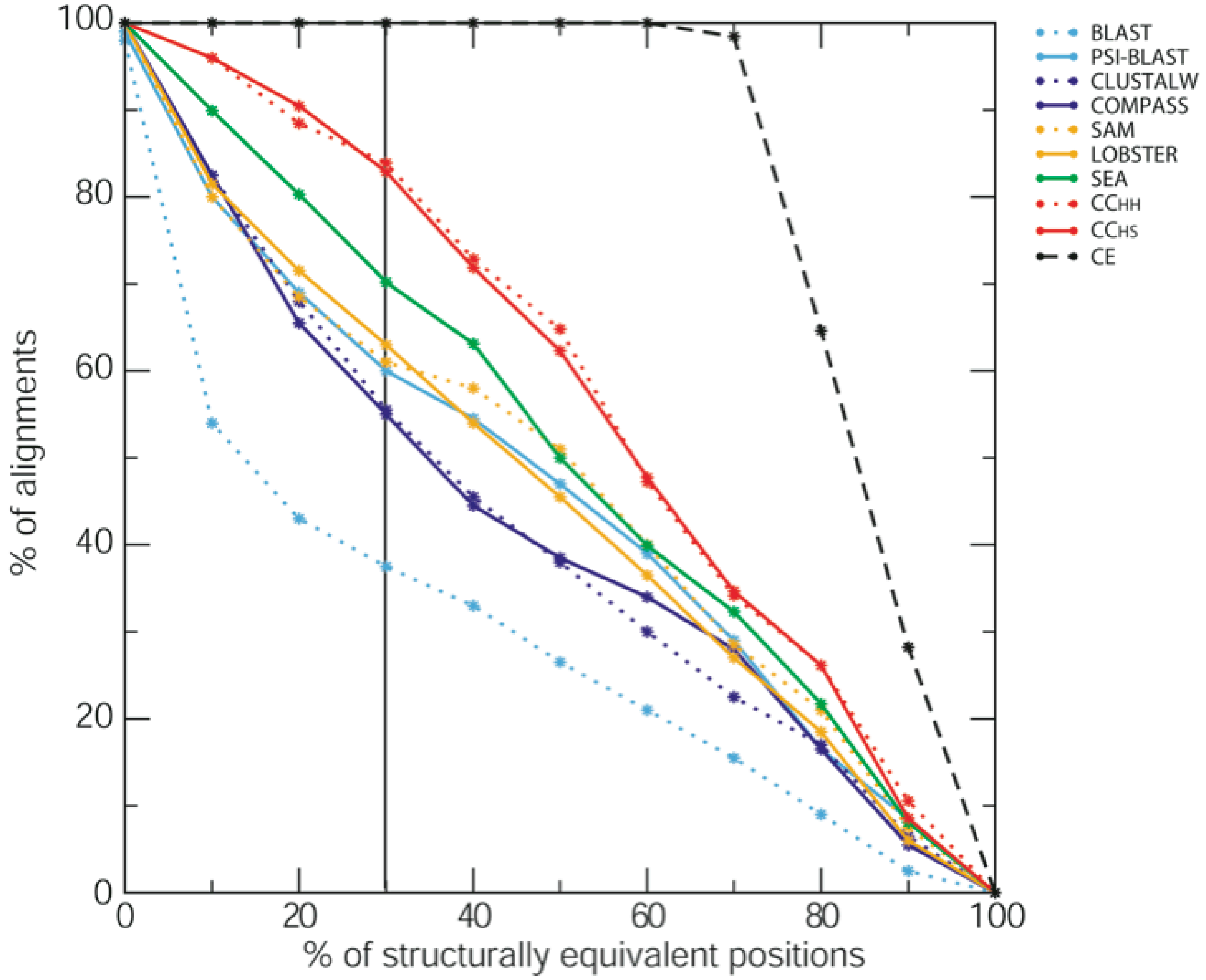
SALIGN accuracy

Method	CE overlap	Shift score
CE	100 ± 0	1.00 ± 0.00
BLAST	26 ± 29	0.32 ± 0.33
PSI-BLAST	43 ± 31	0.48 ± 0.35
SAM	48 ± 26	0.50 ± 0.34
LOBSTER	50 ± 27	0.51 ± 0.32
SEA	49 ± 27	0.53 ± 0.29
ALIGN	42 ± 25	0.44 ± 0.28
CLUSTALW	43 ± 27	0.44 ± 0.31
COMPASS	43 ± 32	0.49 ± 0.35
CCHH	56 ± 23	0.61 ± 0.24
CCHS	56 ± 24	0.62 ± 0.24
TOP	62 ± 20	0.67 ± 0.20



	CCHH	CCHS	LOBSTER	SEA	SAM	COMPASS	PSI-BLAST	CLUSTALW	ALIGN	BLAST
BLAST										
ALIGN										15.81
CLUSTALW									0.94	16.88
PSI-BLAST								0.12	1.14	17.47
COMPASS							0.63	0.60	1.53	17.78
SAM						5.14	5.70	5.74	6.68	22.70
SEA					1.23	6.41	6.97	6.96	7.91	24.09
LOBSTER				0.33	1.54	6.68	7.24	7.28	8.21	24.25
CCHS			5.55	6.21	7.06	12.37	13.10	12.86	14.06	30.26
CCHH		0.78	6.33	6.99	7.84	13.15	13.87	13.64	14.84	31.02

SALIGN success



Alignment accuracy (CE overlap)

200 pairwise DBAli alignments

PSI-BLAST (sequence-profile alignment)	43%
SEA (local structure alignment)	49%
SALIGN (profile-profile alignment)	56%

Program

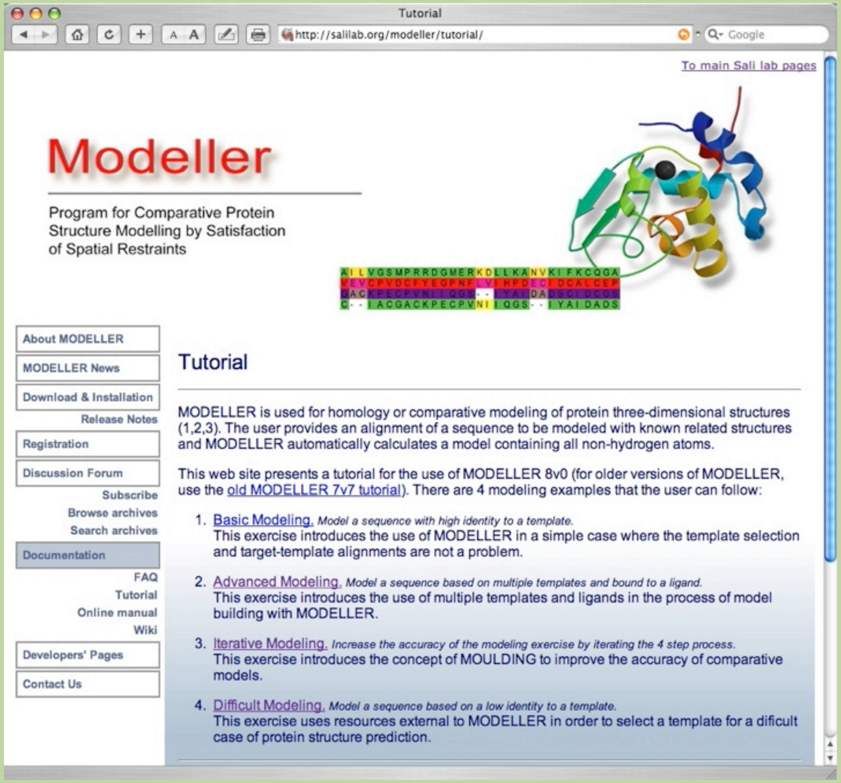
Intro to comparative protein structure prediction

Template Search*

Target – Template Alignment*

Model Building

Model Evaluation



Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints

Tutorial

MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1,2,3). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms.

This web site presents a tutorial for the use of MODELLER 8v0 (for older versions of MODELLER, use the old MODELLER 7v7 tutorial). There are 4 modeling examples that the user can follow:

1. Basic Modeling. *Model a sequence with high identity to a template.* This exercise introduces the use of MODELLER in a simple case where the template selection and target-template alignments are not a problem.
2. Advanced Modeling. *Model a sequence based on multiple templates and bound to a ligand.* This exercise introduces the use of multiple templates and ligands in the process of model building with MODELLER.
3. Iterative Modeling. *Increase the accuracy of the modeling exercise by iterating the 4 step process.* This exercise introduces the concept of MOULDING to improve the accuracy of comparative models.
4. Difficult Modeling. *Model a sequence based on a low identity to a template.* This exercise uses resources external to MODELLER in order to select a template for a dificult case of protein structure prediction.

http://www.salilab.org/modeller/tutotial/

Master Bioinformatics for Health Sciences

Protein Structure Prediction

model building & model assessment

Marc A. Marti-Renom

Adjunct Assistant Professor

http://salilab.org/~marcius

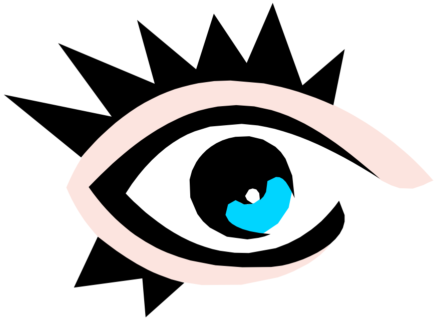
UCSF
University of California
San Francisco

Depts. of Biopharmaceutical Sciences and Pharmaceutical Chemistry
California Institute for Quantitative Biomedical Research
University of California at San Francisco

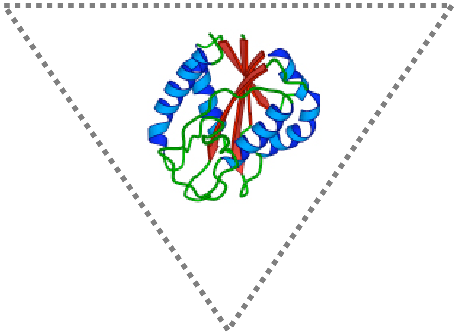
Summary

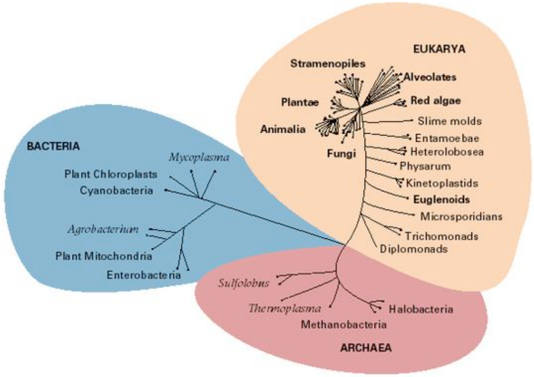
- **Model building with MODELLER**
  - Points and restraints
  - Model accuracy
  - Modeling loops
  - Evaluating models
  - MOULDER
  - Modeing genes (examples)
  - Modeling genomes (large-scale modeling)

Information about a protein can come from three distinct sources

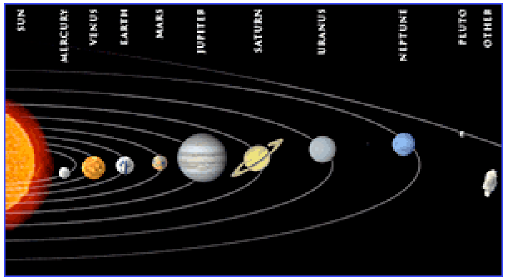


Experimental observations





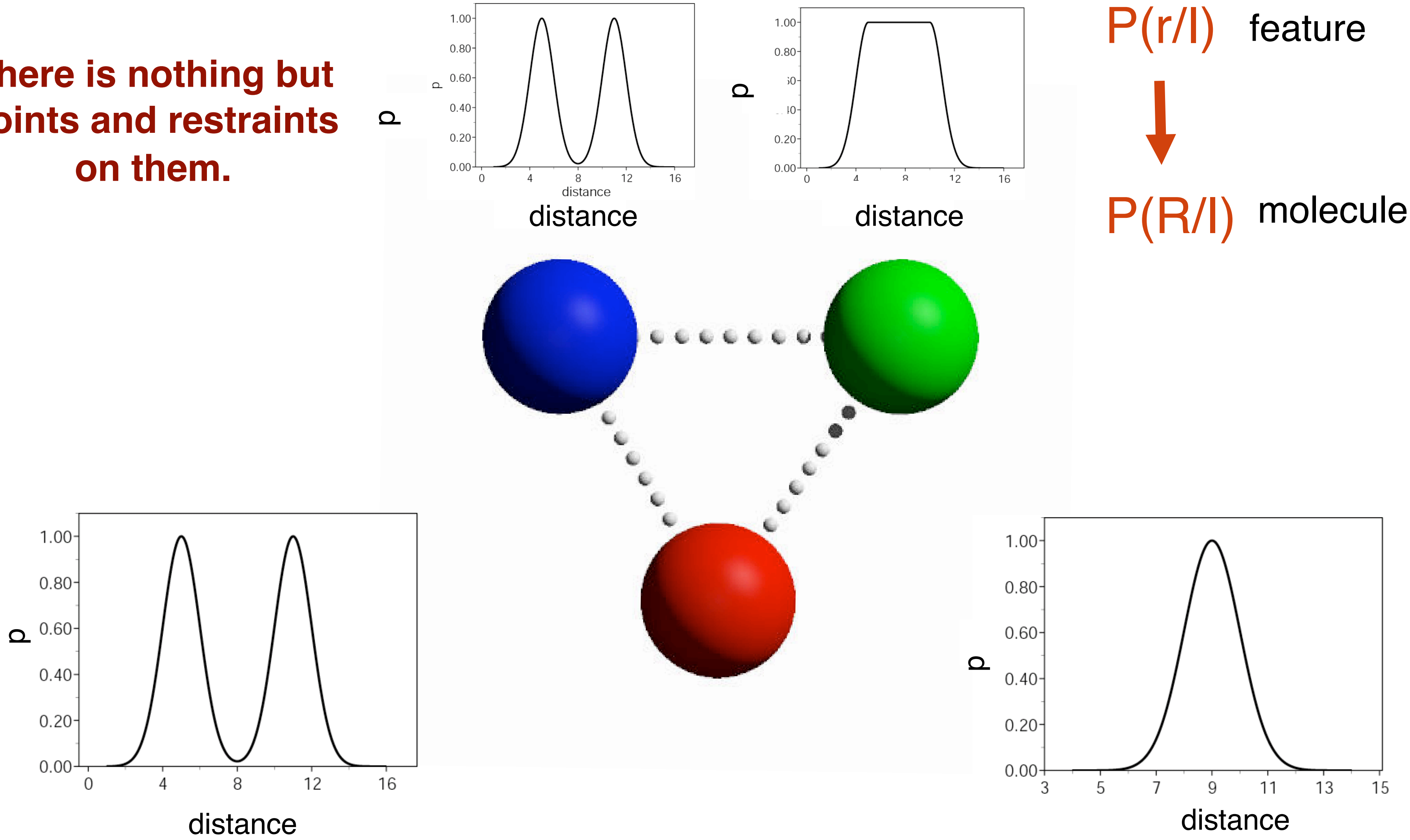
Statistical rules



Laws of physics

Modeling by optimization

There is nothing but points and restraints on them.



P(r/I) feature

↓

P(R/I) molecule

Classes of methods for comparative protein structure modeling

- Model building by assembly of rigid bodies

  core, loops, sidechains.

- Model building by segment matching.

- Model building by satisfaction of spatial restraints.

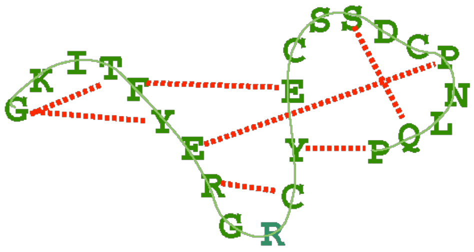
Marti-Renom et al. Annu.Rev.Biophys.Biomol.Struct. 29, 291-325, 2000.

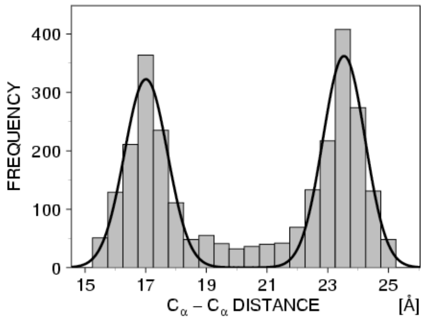
Comparative modeling by satisfaction of spatial restraints

MODELLER

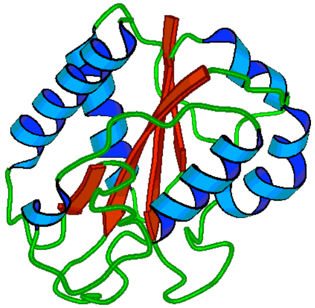
```
3D  GKITFYERGFQGHCYESDC-NLQP…
SEQ GKITFYERG---RCYESDCPNLQP…
```

1. Extract spatial restraints



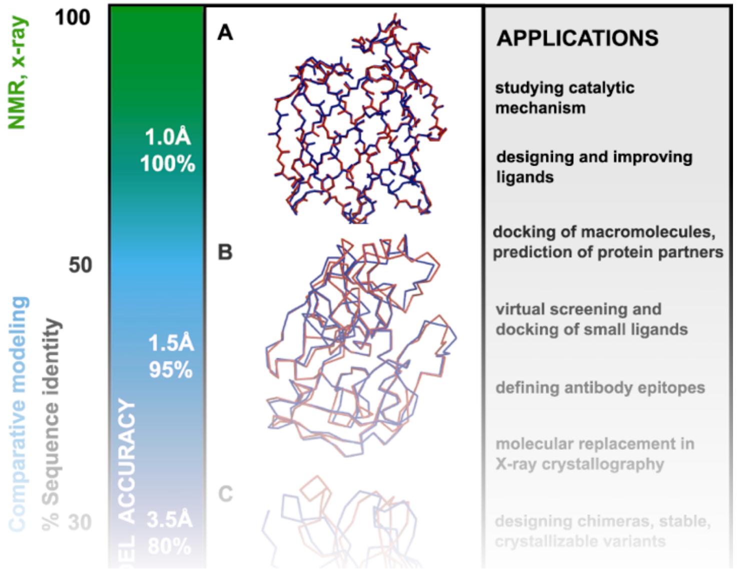


2. Satisfy spatial restraints



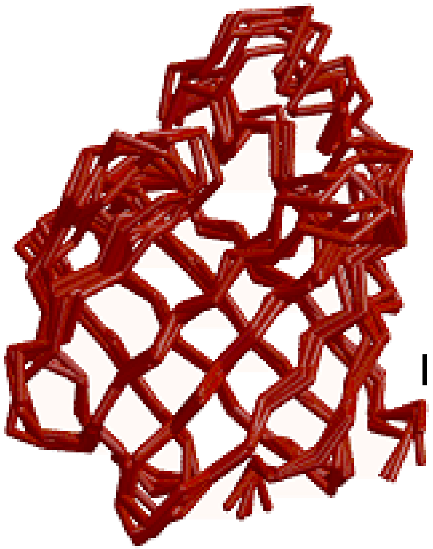
$$F(R) = \prod_i p_i \, (f_i / I)$$

A. Šali & T. Blundell. J. Mol. Biol. 234, 779, 1993.
J.P. Overington & A. Šali. Prot. Sci. 3, 1582, 1994.
A. Fiser, R. Do & A. Šali, Prot. Sci., 9, 1753, 2000.


NMR, x-ray
Comparative modeling
% Sequence identity
100
50
30
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
A
B
C
APPLICATIONS
studying catalytic mechanism
designing and improving ligands
docking of macromolecules, prediction of protein partners
virtual screening and docking of small ligands
defining antibody epitopes
molecular replacement in X-ray crystallography
designing chimeras, stable, crystallizable variants

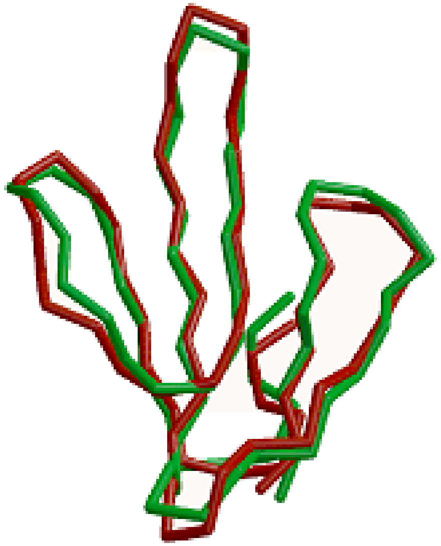
Accuracy and applicability of comparative models

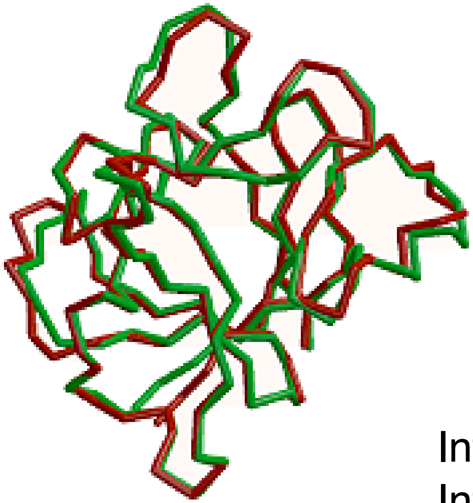
“Biological” significance of modeling errors



NMR
Ileal lipid-binding protein
1eal

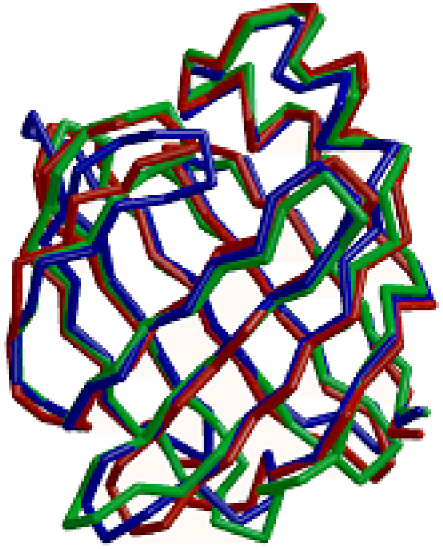
NMR – X-RAY
Erabutoxin 3ebx
Erabutoxin 1era





X-RAY
Interleukin 1β 41bi (2.9Å)
Interleukin 1β 2mib (2.8Å)

CRABPII 1opbB
FABP 1ftpA
ALBP 1lib
40% seq. id.

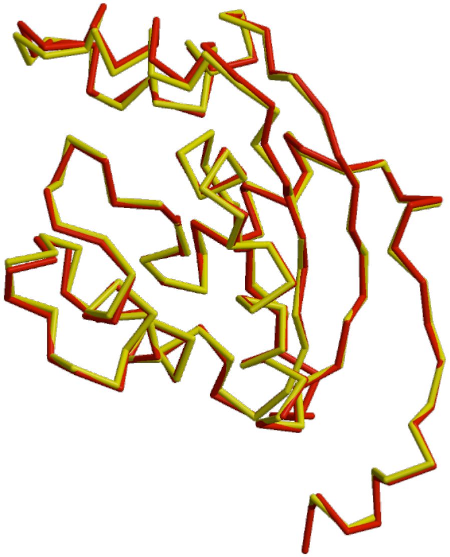


Model Accuracy

HIGH ACCURACY

NM23   Seq id 77%

Cα equiv 147/148
RMSD 0.41Å

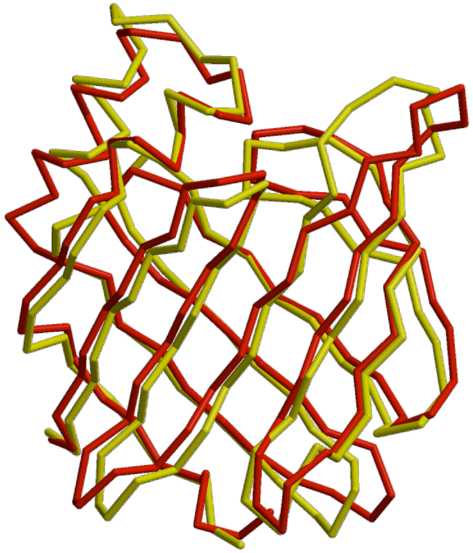


Sidechains
Core backbone
Loops

MEDIUM ACCURACY

CRABP   Seq id 41%

Cα equiv 122/137
RMSD 1.34Å

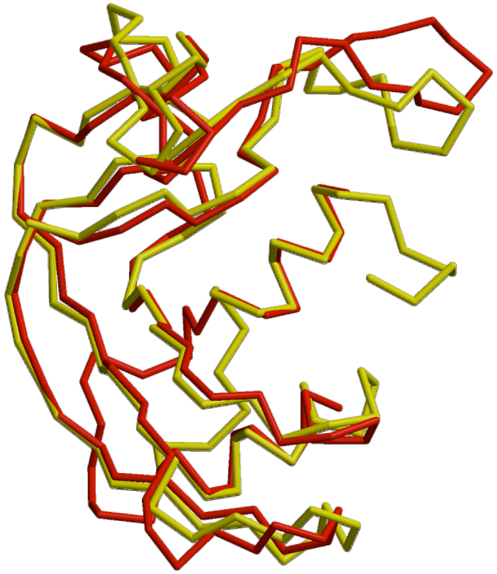


Sidechains
Core backbone
Loops
Alignment

LOW ACCURACY

EDN Seq id 33%

Cα equiv 90/134
RMSD 1.17Å



Sidechains
Core backbone
Loops
Alignment
Fold assignment

X-RAY / MODEL

Marti-Renom et al. Annu.Rev.Biophys.Biomol.Struct. 29, 291-325, 2000.

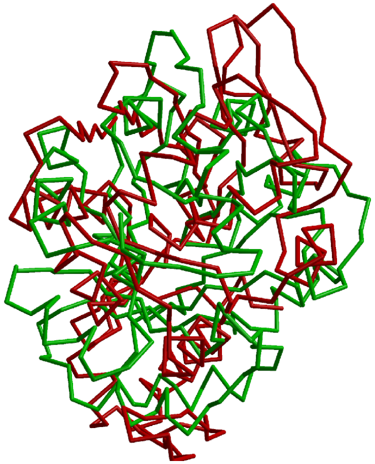
Typical errors in comparative models

MODEL

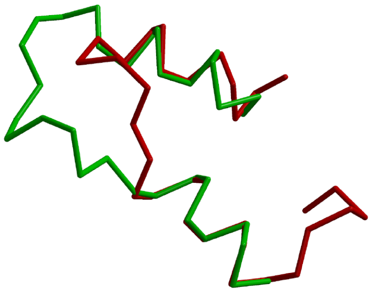
X-RAY

TEMPLATE

Incorrect template

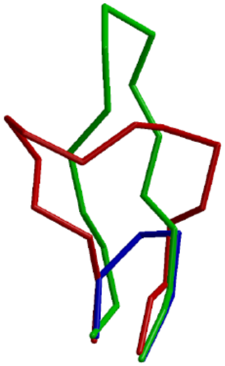


Misalignment

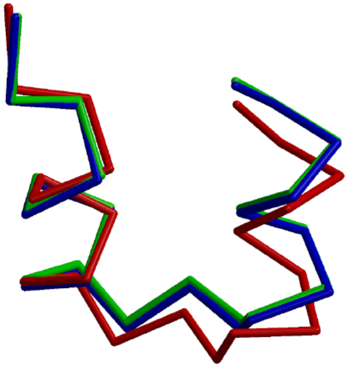


```
                 10        20        30        40        50        60
EDN   ---KPPQFTWAQWFETQHINMTSQQCTNAMQVINNYQRRCKNQNTFLLTTFANVVNVCGNPNMTCPSN
      /////////////////   /   |||||||\\\\  ||||||||||||||||||||||||||||  ||
7RSA  KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHESLADVQAVCSQKNVAC-KN
              aaaaaaaaaaa     aaaaaaaaaaaaaa   bbbbbbb aaaaaaaaa

         70        80        90       100       110       120       130
EDN   KTRKNCHHSGSQVPLIHCNLTTPSPQNISNCRYAQTPANMFYIVACDNRDQRRDPPQYPVVPVHLDRII
      \ ||||||||||||||||||| \\ \\ \ ||||||||||||||||||||         |||||||||||
7RSA  -GQTNCYQSYSTMSITDCRETGSS--KYPNCAYKTTQANKHIIVACEGN---------PYVPVHFDASV
          bbbbb   bbbbbbbbbb    aaaaabbbbbbbbbbbbbbbbbb         bbbbbbbbbbb
```

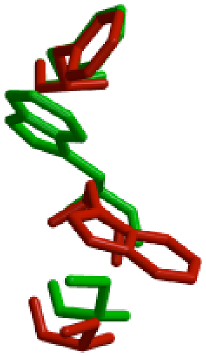
Region without a template



Distortion/shifts in aligned regions

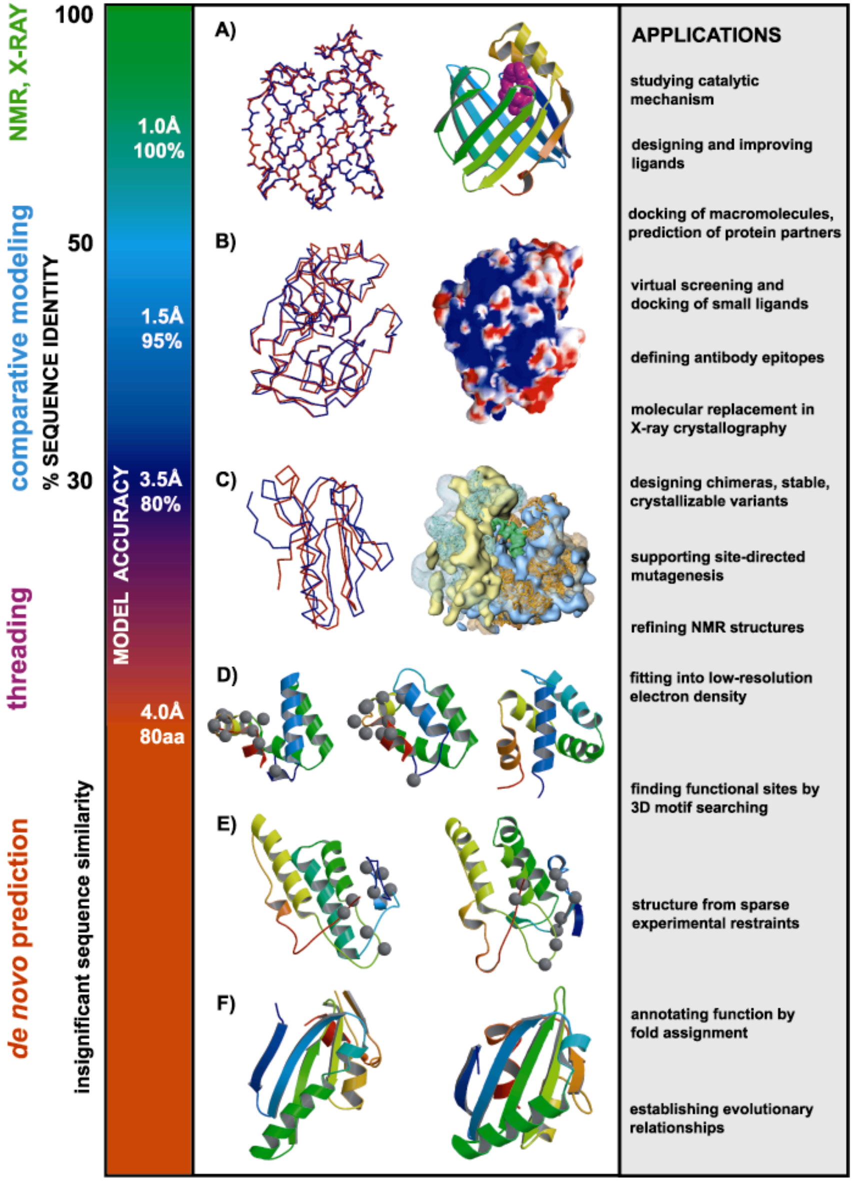


Sidechain packing



Marti-Renom et al. Annu.Rev.Biophys.Biomol.Struct. 29, 291-325, 2000.

Utility of protein structure models, despite errors



APPLICATIONS

studying catalytic mechanism

designing and improving ligands

docking of macromolecules, prediction of protein partners

virtual screening and docking of small ligands

defining antibody epitopes

molecular replacement in X-ray crystallography

designing chimeras, stable, crystallizable variants

supporting site-directed mutagenesis

refining NMR structures

fitting into low-resolution electron density

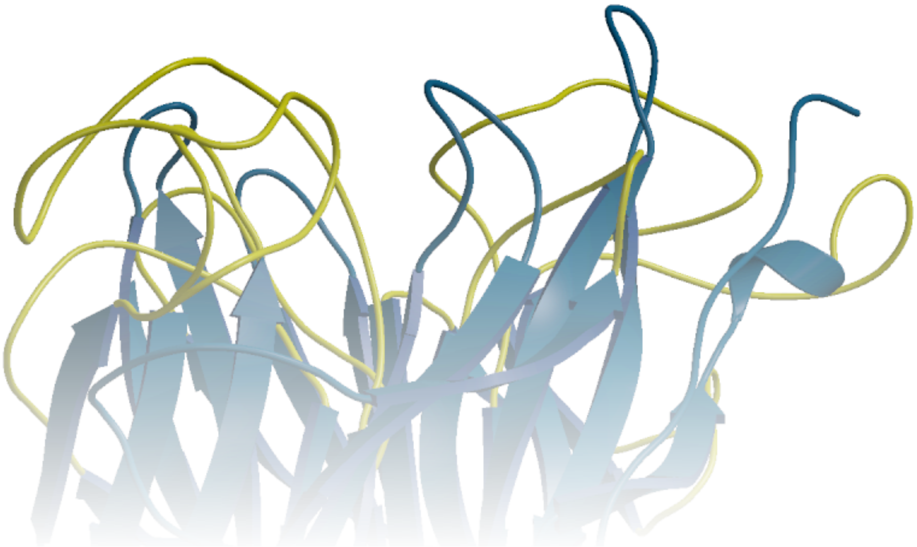
finding functional sites by 3D motif searching

structure from sparse experimental restraints

annotating function by fold assignment

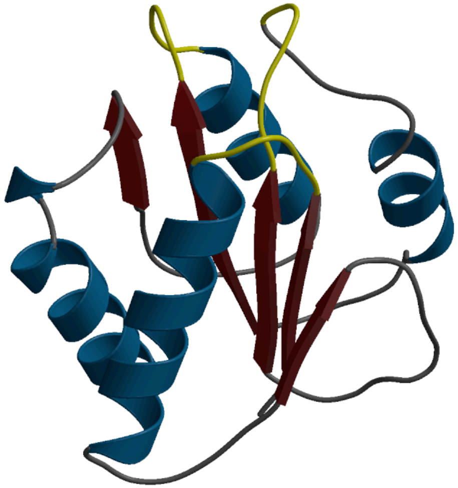
establishing evolutionary relationships

D. Baker & A. Sali. Science 294, 93, 2001.

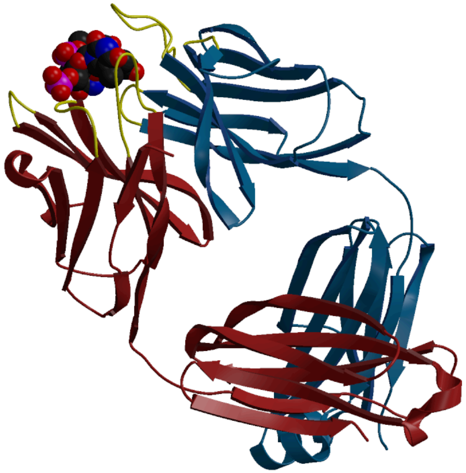


Modeling of loops in protein structures

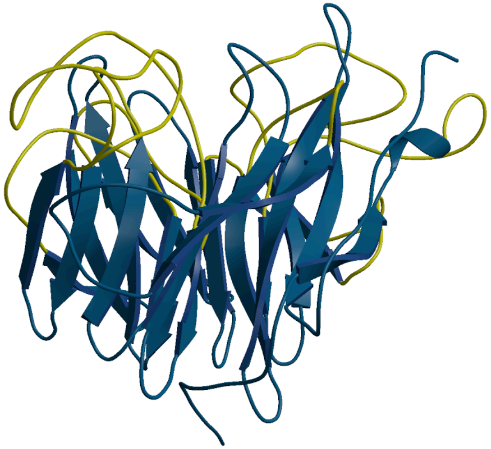
Loop Modeling in Protein Structures



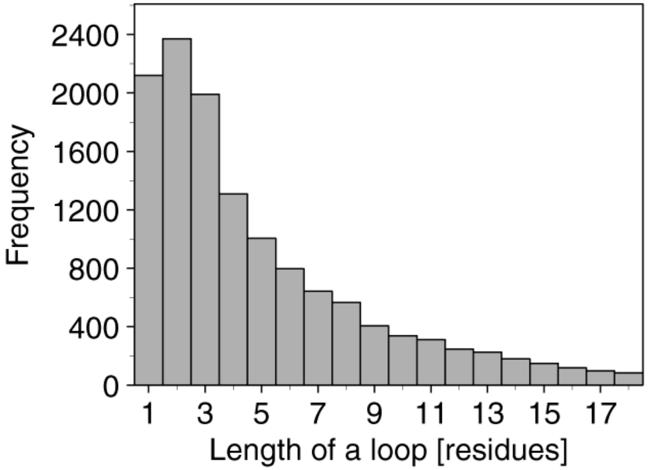
α+β barrel: flavodoxin



IG fold: immunoglobulin

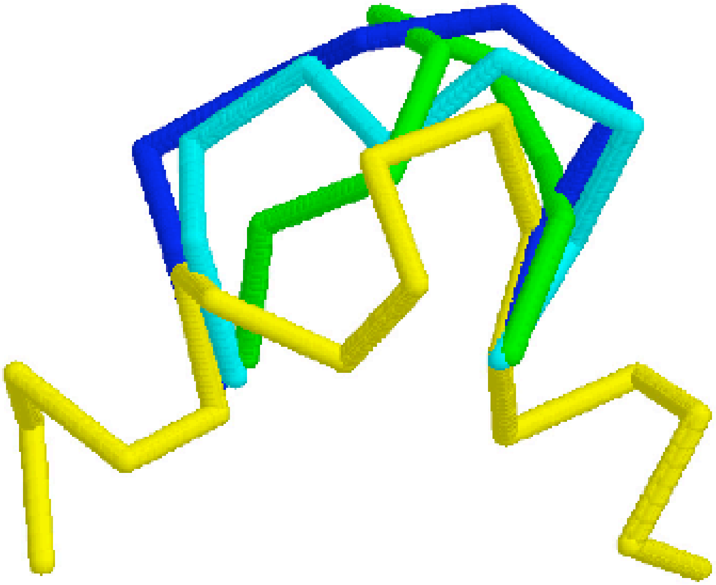


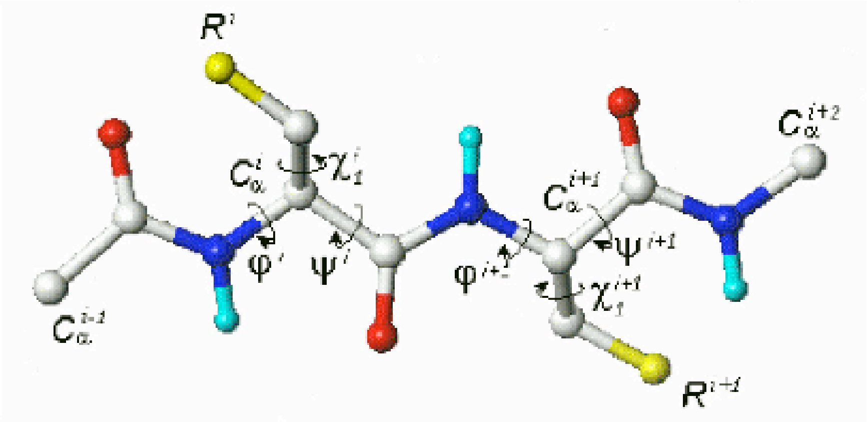
antiparallel β-barrel



A. Fiser, R. Do & A. Šali, Prot. Sci. 9, 1753, 2000.

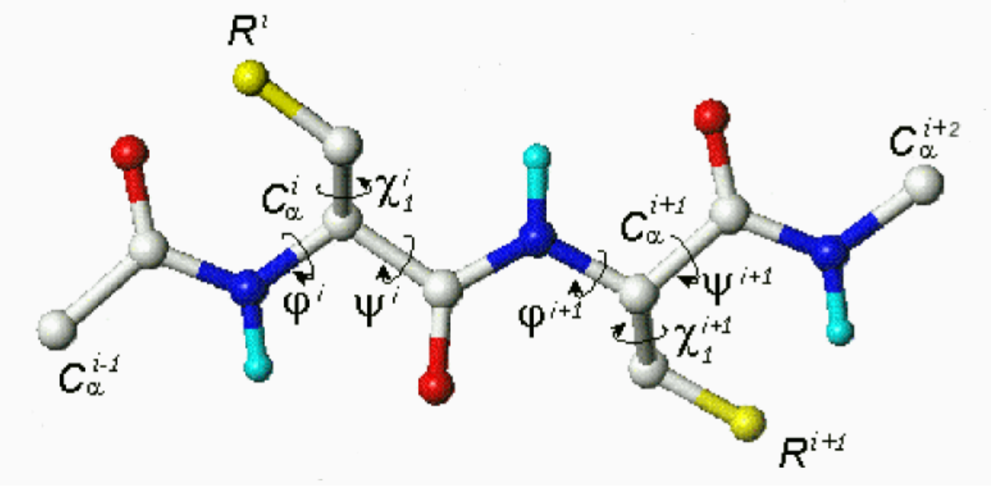
Loop modeling strategies





- database is complete only up to 4-6 residues
- even in DB search, the different conformations must be ranked
- loops longer than 4 residues need extensive optimization
- DB method is efficient for specific families (eg, canonical loops in Ig’s, β–hairpins)

Loop Modeling by Conformational Search



1. Protein representation.

2. Energy (scoring) function.

3. Optimization algorithm.

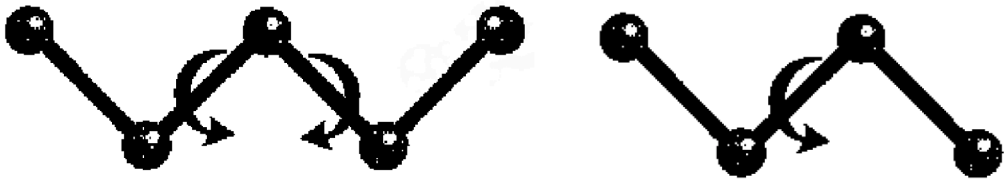
Energy Function for Loop Modeling

The energy function is a sum of many terms:

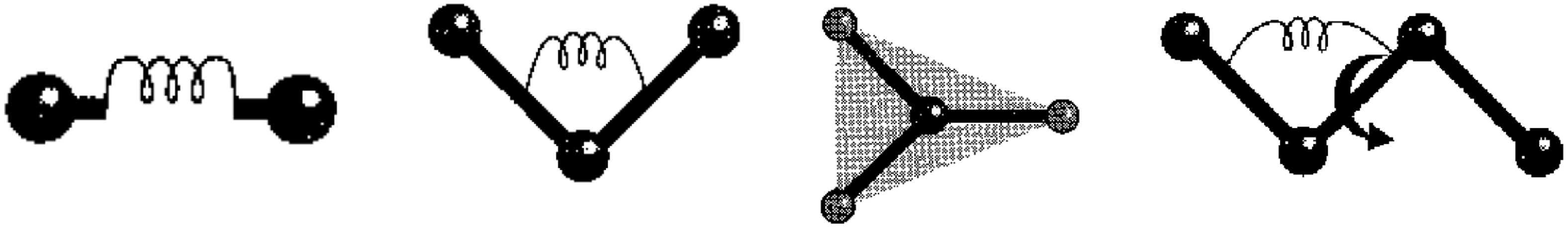
1. Stereochemistry (CHARMM).
2. Mainchain conformation (Φ， Ψ).
3. Non-bonded contacts.

Energy Function for Loop Modeling

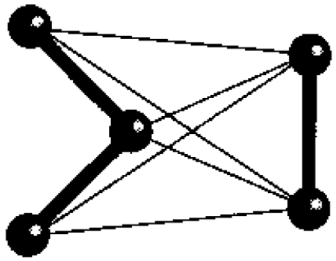
1) Statistical preferences for dihedral angles:



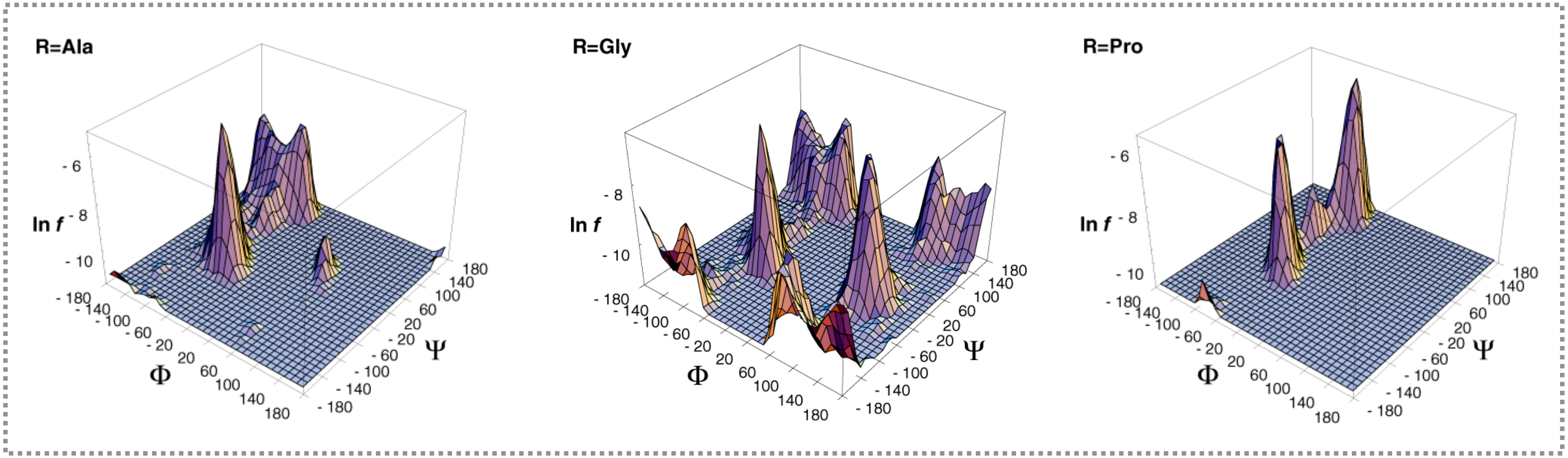
2) Restraints from the CHARMM-22 force field:

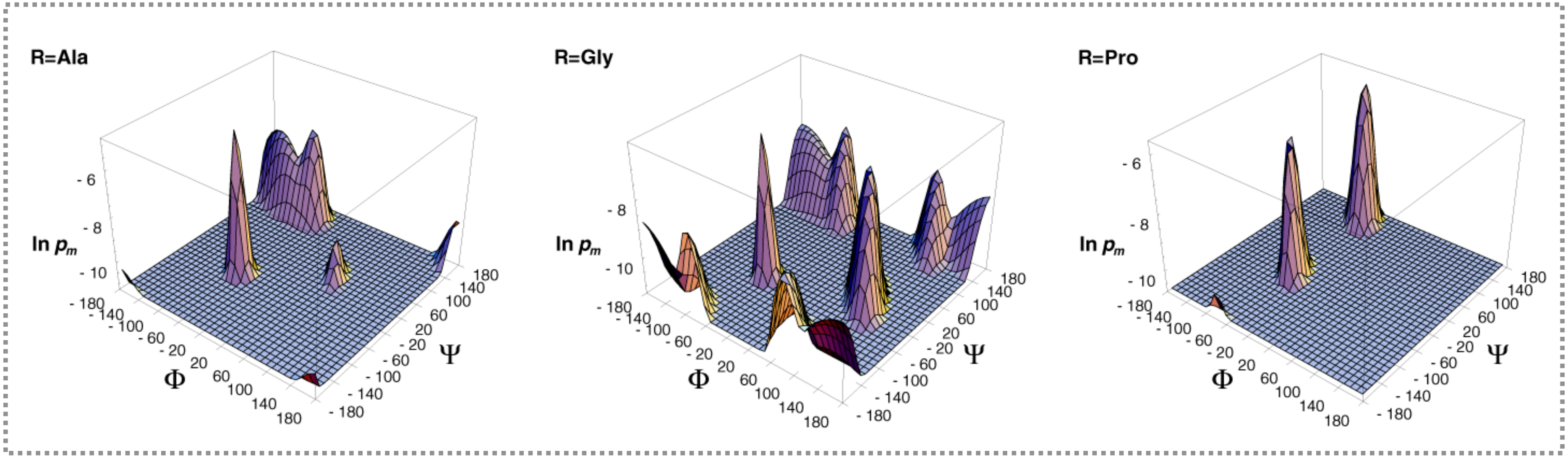


3) Statistical potential for non-bonded contacts:

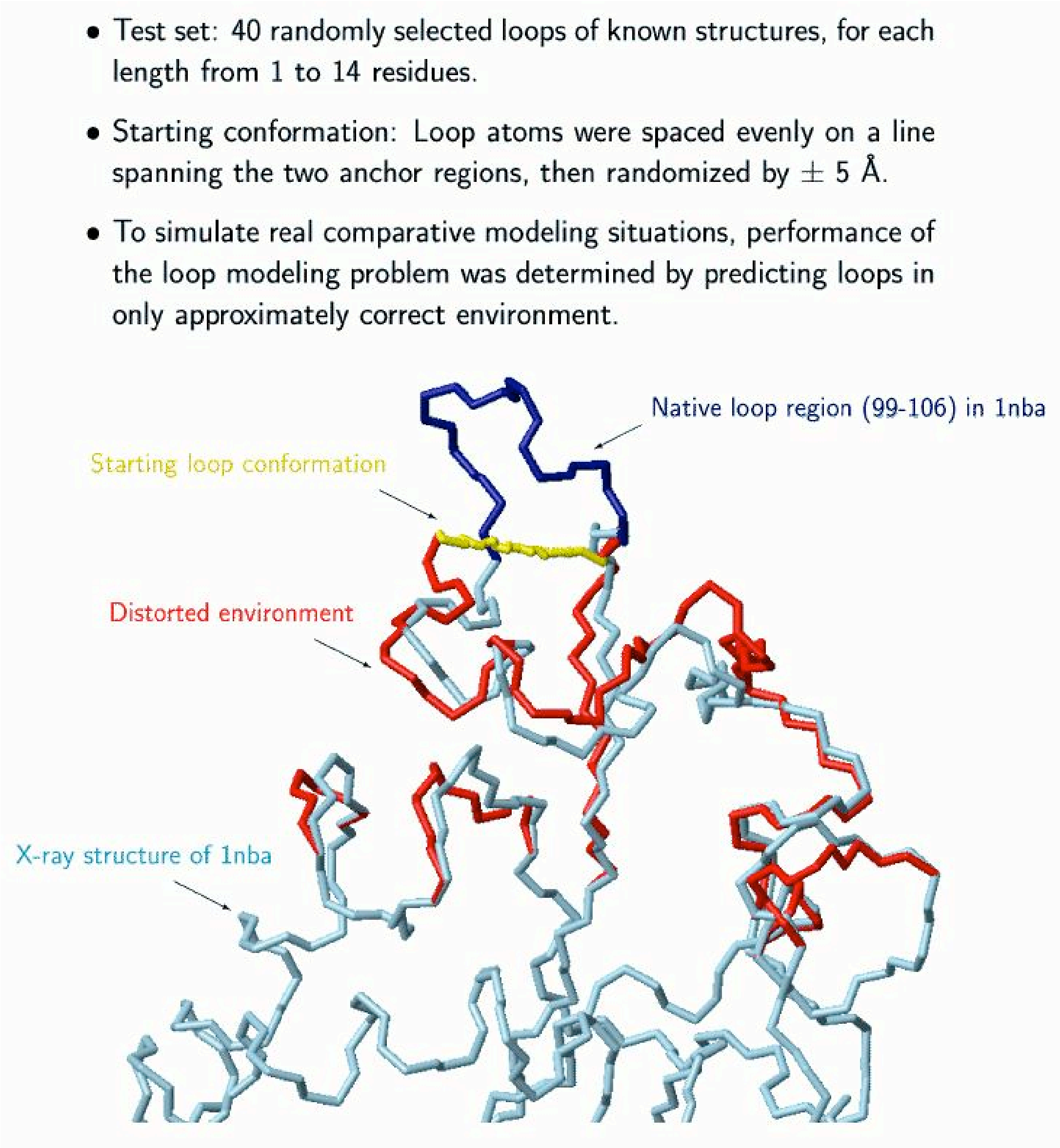


Mainchain Terms for Loop Modeling





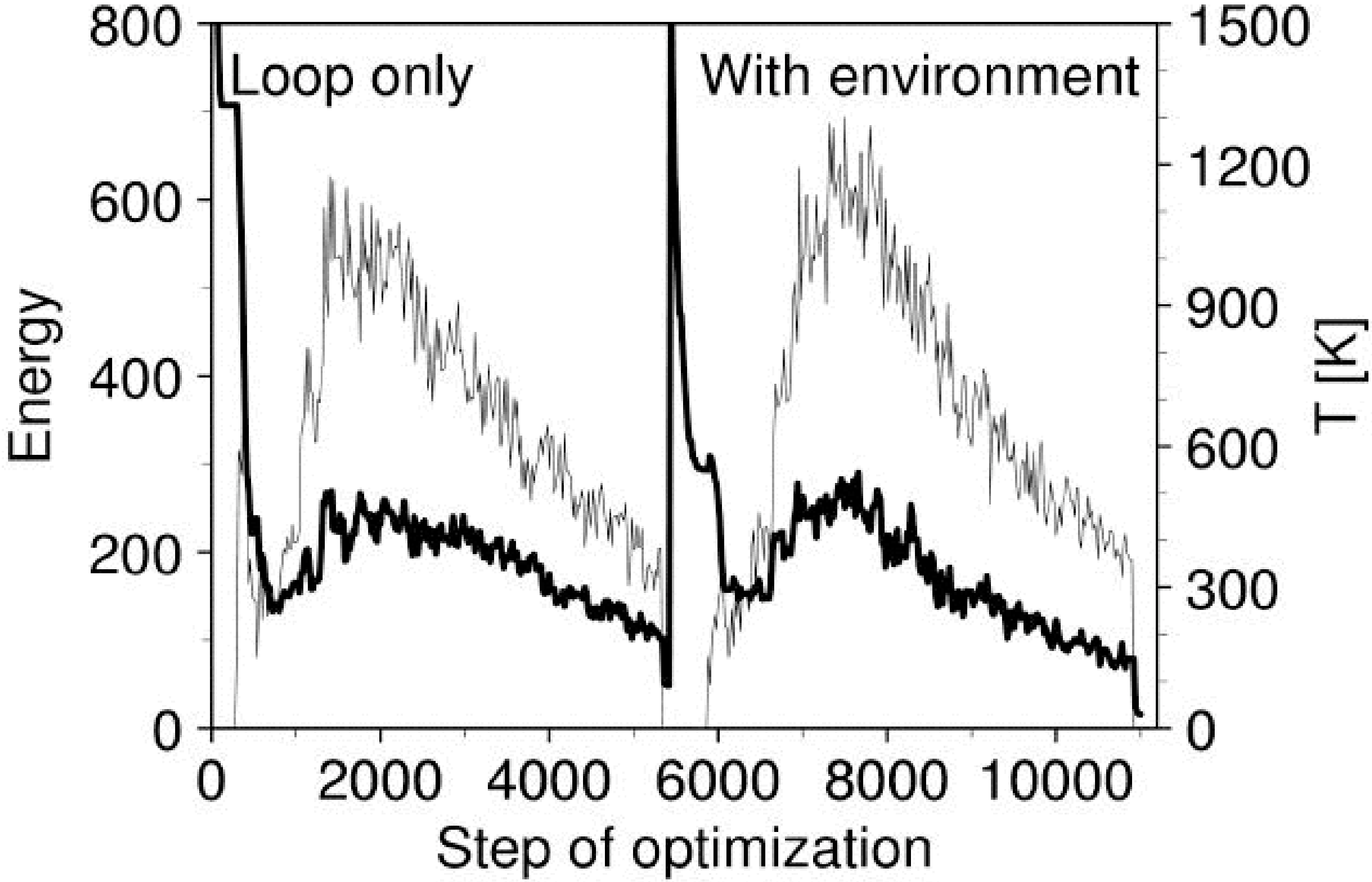
Optimization of Objective Function



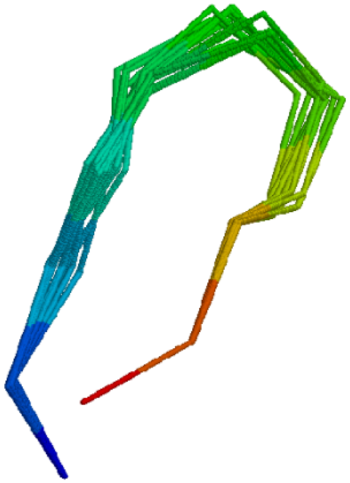
- Test set: 40 randomly selected loops of known structures, for each length from 1 to 14 residues.
- Starting conformation: Loop atoms were spaced evenly on a line spanning the two anchor regions, then randomized by ± 5 Å.
- To simulate real comparative modeling situations, performance of the loop modeling problem was determined by predicting loops in only approximately correct environment.

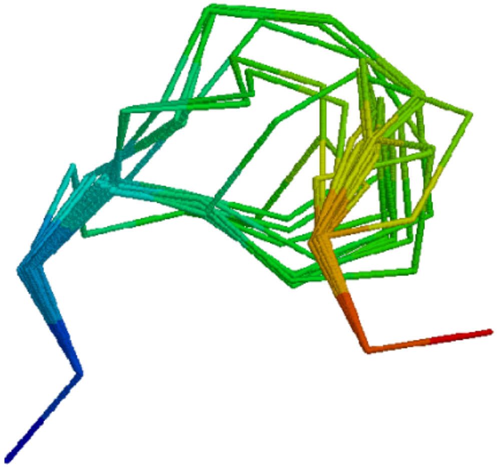
Native loop region (99-106) in 1nba
Starting loop conformation
Distorted environment
X-ray structure of 1nba

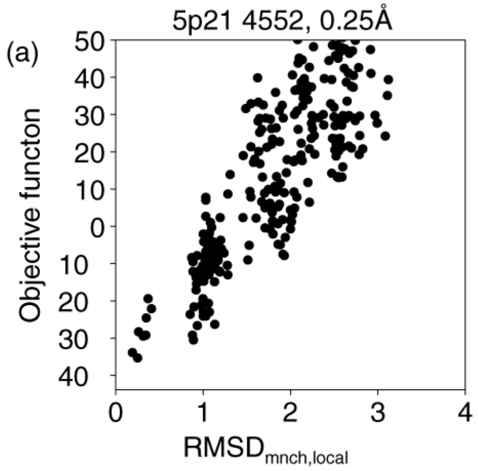
Optimization of Objective Function

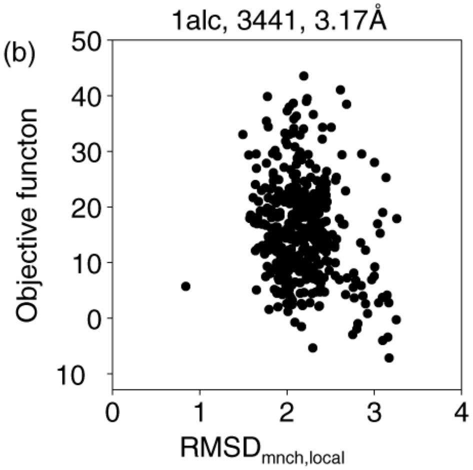


Calculating an Ensemble of Loop Models

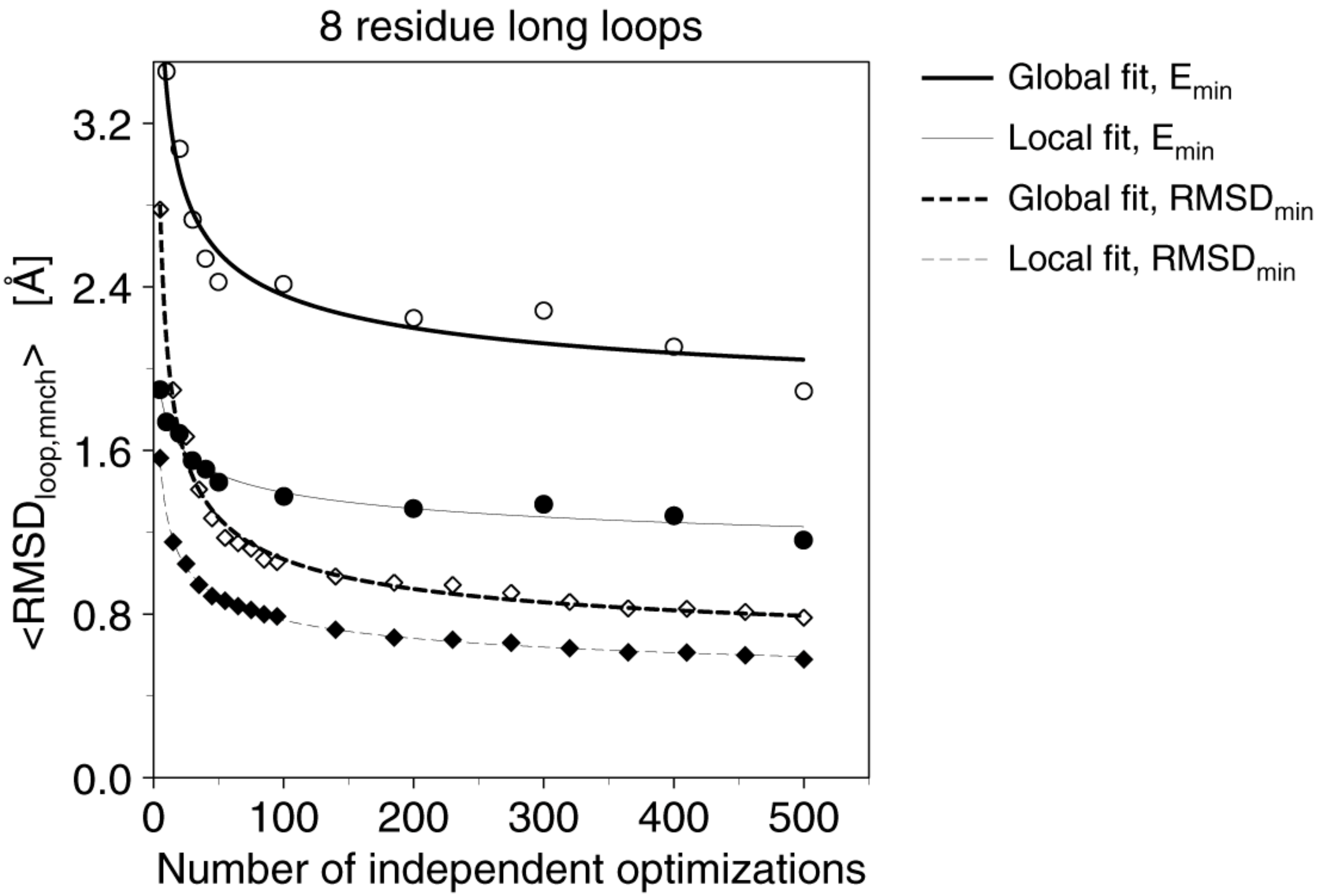




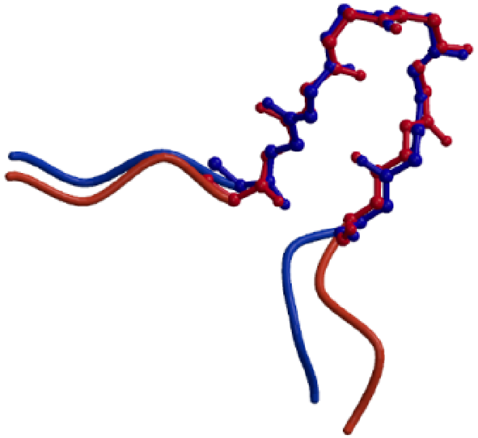




Accuracy of loop models as a function of amount of optimization



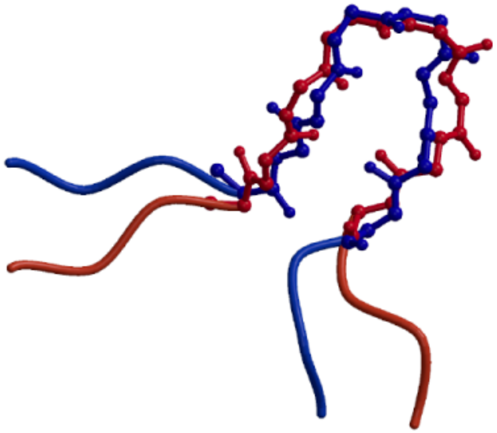
Accuracy of Loop Modeling



RMSD=0.6Å

HIGH ACCURACY (<1Å)

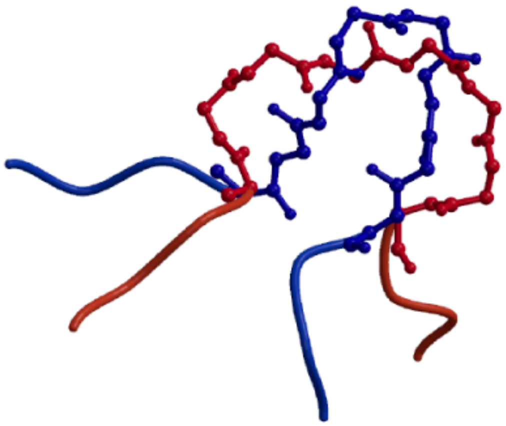
50% (30%) of
8-residue loops



RMSD=1.1Å

MEDIUM ACCURACY (<2Å)

40% (48%) of
8-residue loops



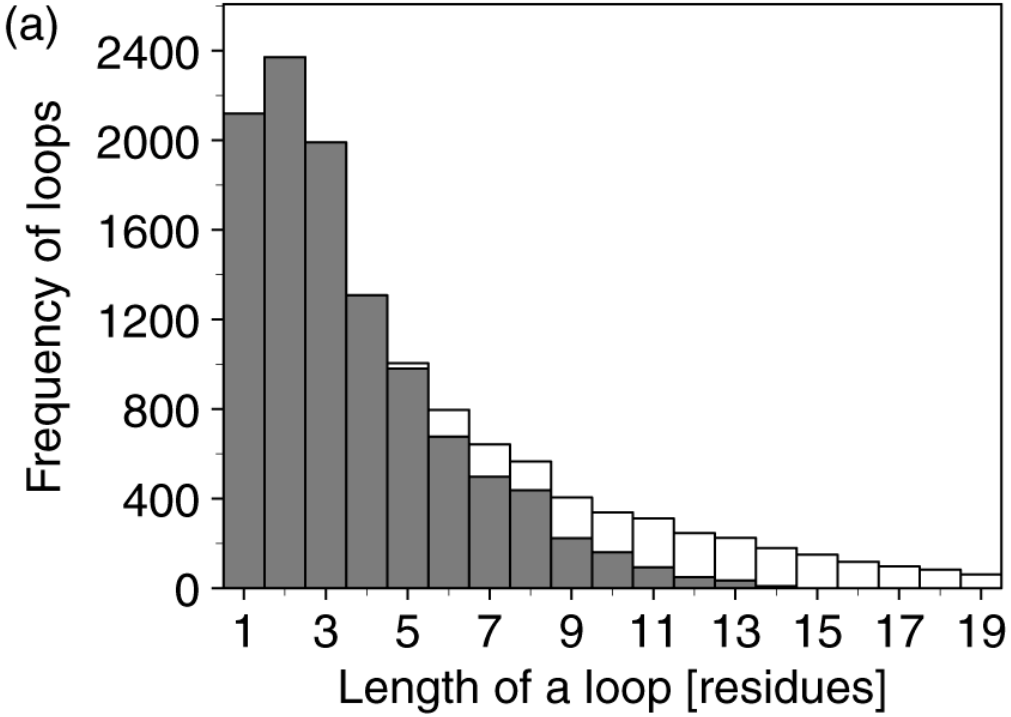
RMSD=2.8Å

LOW ACCURACY (>2Å)

10% (22%) of
8-residue loops

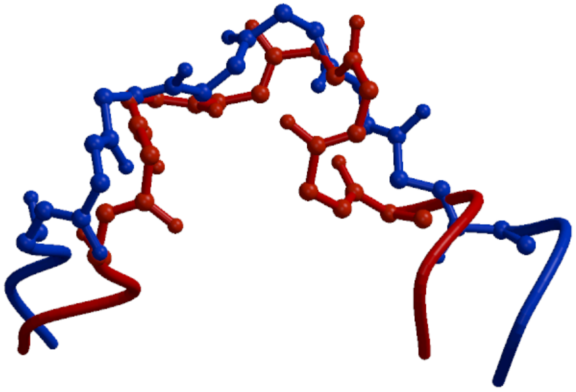
A. Fiser, R. Do & A. Šali, Prot. Sci. 9, 1753, 2000.

Fraction of Loops Modeled With at Least Medium Accuracy

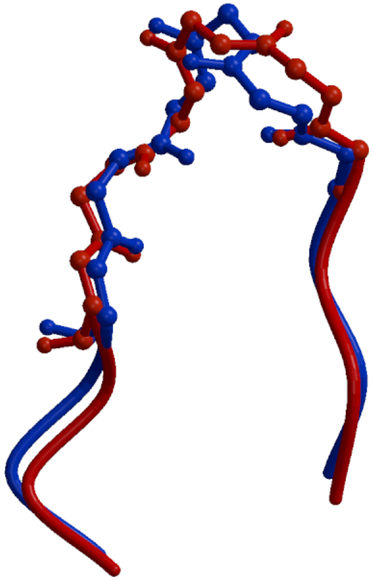


Problems in Practical Loop Modeling

1. Decide which regions to model as loops.
2. Correct alignment of anchor regions & environment.
3. Modeling of a loop.

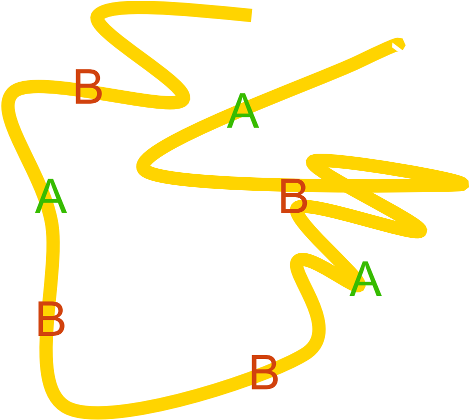


T0058: 80-85
$\text{RMSD}_{\text{mnch}}$ loop = 1.09 Å
$\text{RMSD}_{\text{mnch}}$ anchors = 0.29 Å



T0076: 46-53
$\text{RMSD}_{\text{mnch}}$ loop = 1.37 Å
$\text{RMSD}_{\text{mnch}}$ anchors = 1.52 Å

Potentials of Mean Force (PMF)



Empirical energy functions (PMF)

Idea: **energy leads to structure, thus it should be possible to infer energy from many known structures**

To be used in: **model refinement and assessment**

Properties needed:

- Deep minimum at correct state (native)
- Smooth (energy landscape)
- Simple (CPU calculation)

Types:

- Contact potential
- Distance potentials
- Surface potentials

Approximations/Limitations in PMFs

Database size.

PMF versus Energy (additive/higher order terms).

Reference state.

Physical origin.

*Finkelstein et al. (1995) Proteins **23**, pp142*

Potentials of Mean Force

As any other bioinformatics problem…

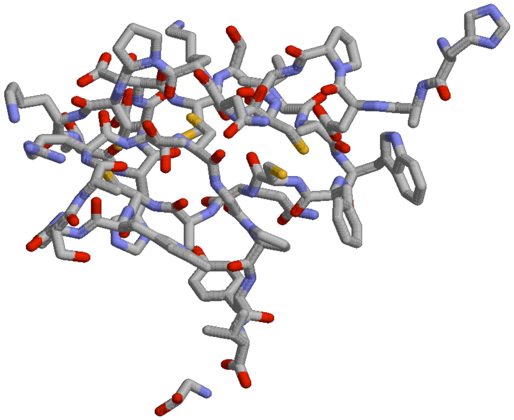
- Representation
- Scoring
- Optimizer

Representation

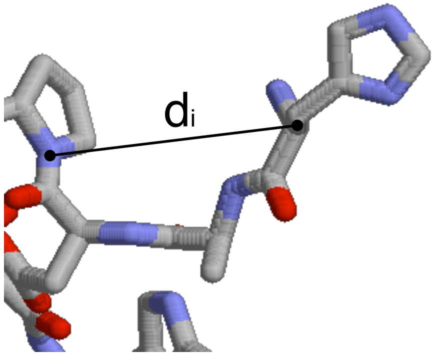
Sequence/Structures

>gi42541361
MDIRSVSSLRGLLCLPPSWPRR

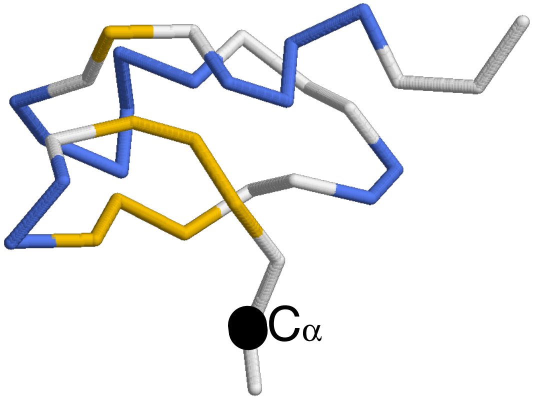
Primary sequence



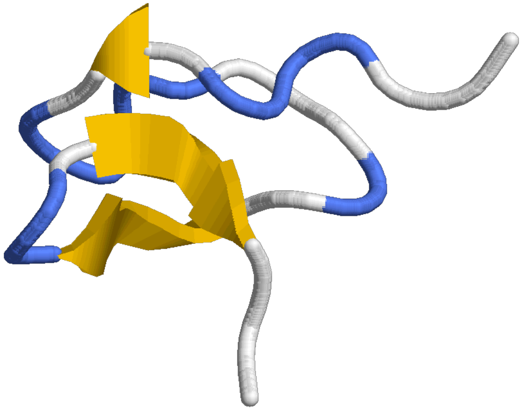
All atoms and coordinates



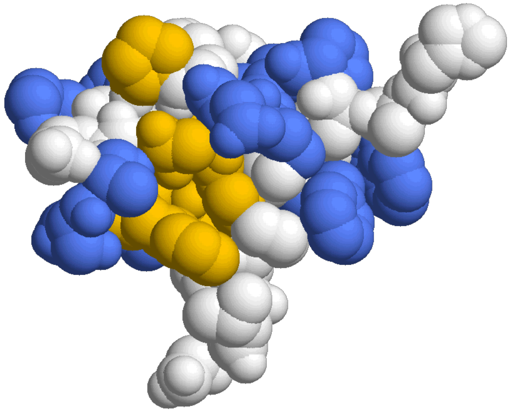
Distance space



Reduced atoms representation



Secondary Structure

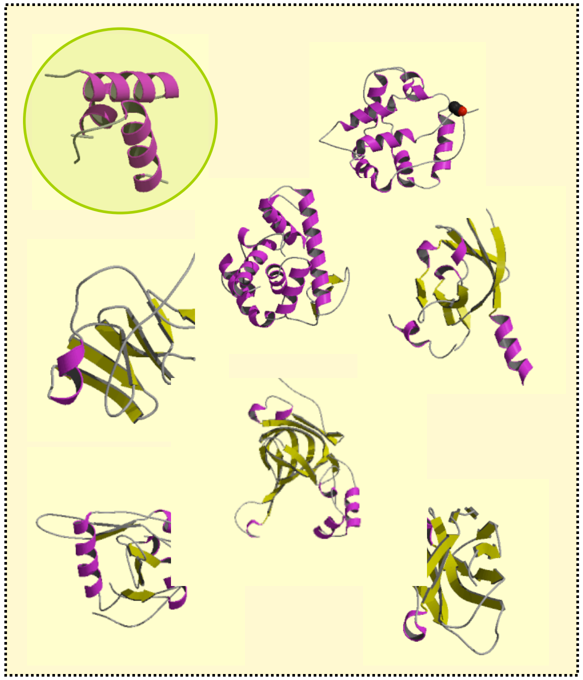


Accessible surface

Scoring

Statistical Potentials (background)

Structural space



Sequence space

MKLLIVLTCISLCSCICTVVQRCASNKPHVLEDPCKVQH

HLSVNQCVLLPQCCPKSCKICTHLISIEVVLTCRAVDKM

MHVNCVEQCSLQDCIKIAPRVLKTCILCVLKPCLTSVSH

VHLVQPTSCCCKKNCICHVEIRSLDILTKSVQLACLVPM

⋮

MQCCRVQKICDLLAVELCKLHISTPSCKILCVVTSVPHN

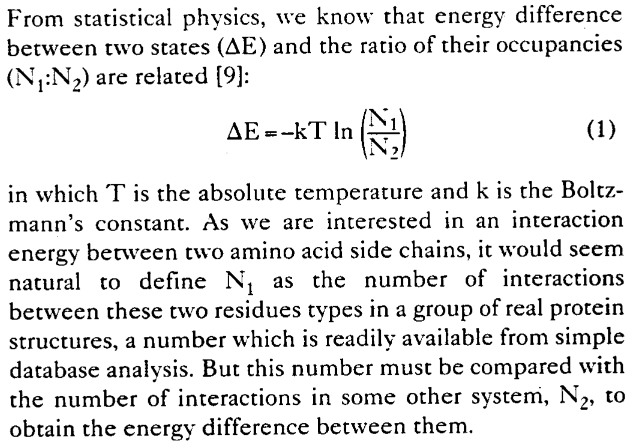
Scoring

Statistical Potential (inspiration)

$$K = \frac{[AB]}{[A]\cdot[B]}$$

$$\Delta G = -RT\ln(K) = -RT\ln\frac{[AB]}{[A]\cdot[B]}$$

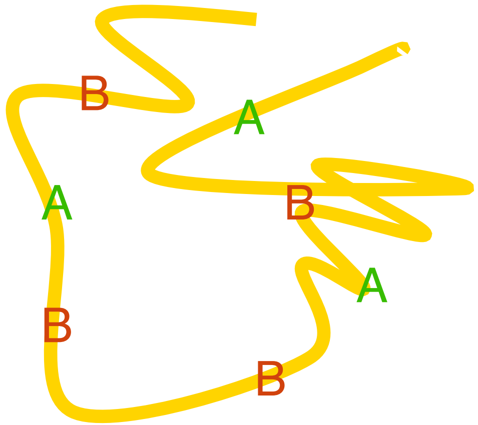
$$A + B \rightleftharpoons AB$$



From statistical physics, we know that energy difference between two states (ΔE) and the ratio of their occupancies (N_1:N_2) are related [9]:

$$\Delta E = -kT \ln\left(\frac{N_1}{N_2}\right) \quad (1)$$

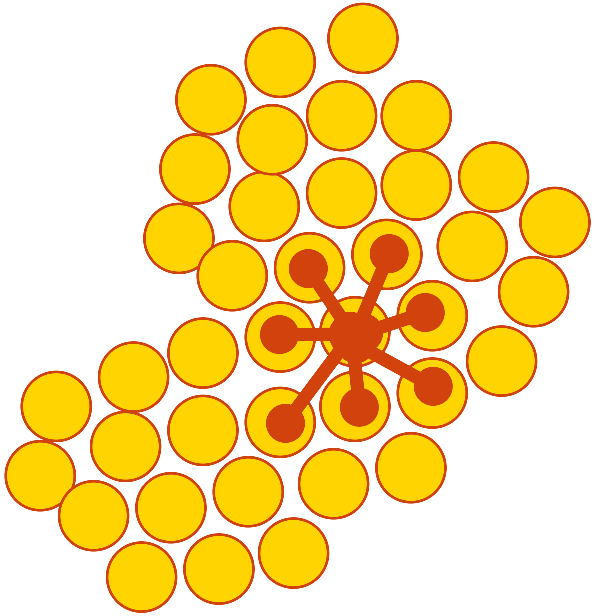
in which T is the absolute temperature and k is the Boltzmann's constant. As we are interested in an interaction energy between two amino acid side chains, it would seem natural to define N_1 as the number of interactions between these two residues types in a group of real protein structures, a number which is readily available from simple database analysis. But this number must be compared with the number of interactions in some other system, N_2, to obtain the energy difference between them.

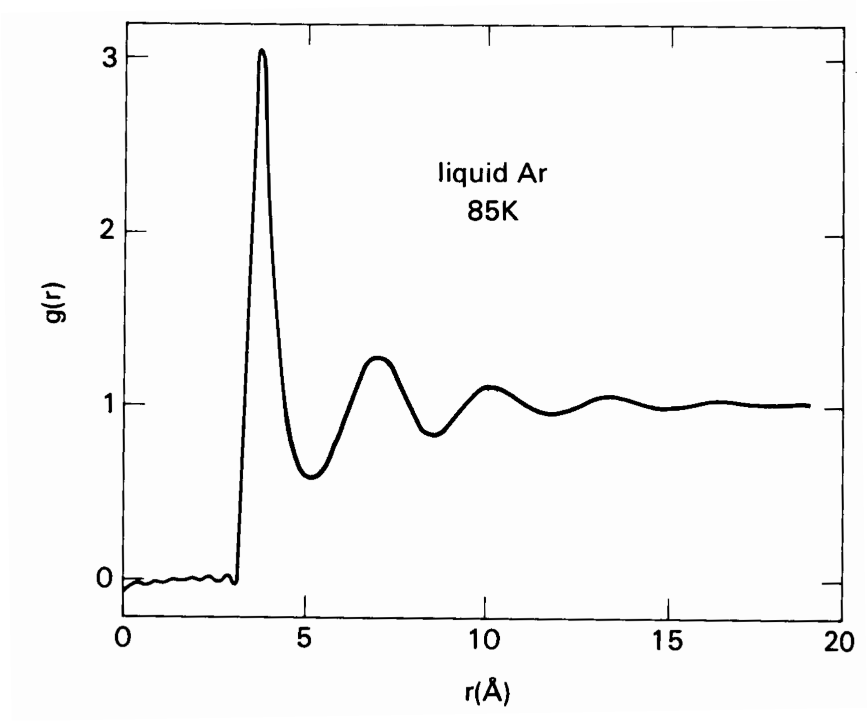


*Tanaka and Sheraga (1975) PNAS, **72** pp3802*
*Sippl, (1990) J.Mo.Biol. **213** pp859*
*Godzik, (1996) Structure **15** pp363*

Scoring

Statistical Potential (reference state)



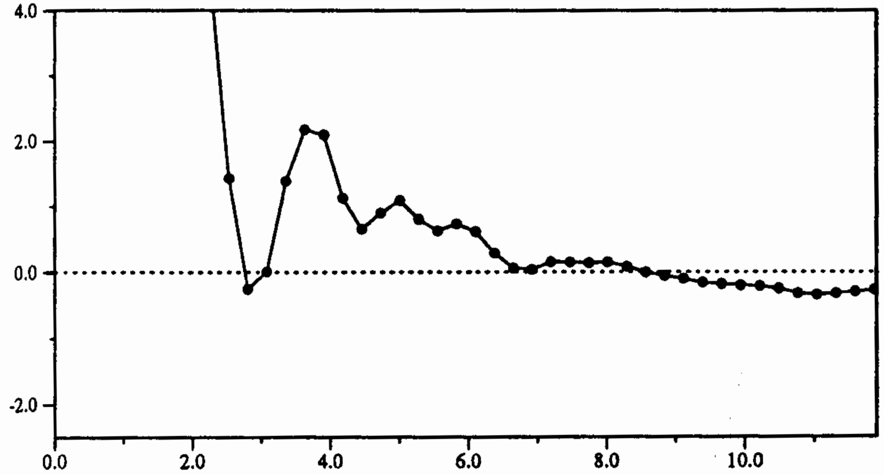


Theory of simple liquids 2nd edition JP Hansen and IR McDonald, Academic Press.

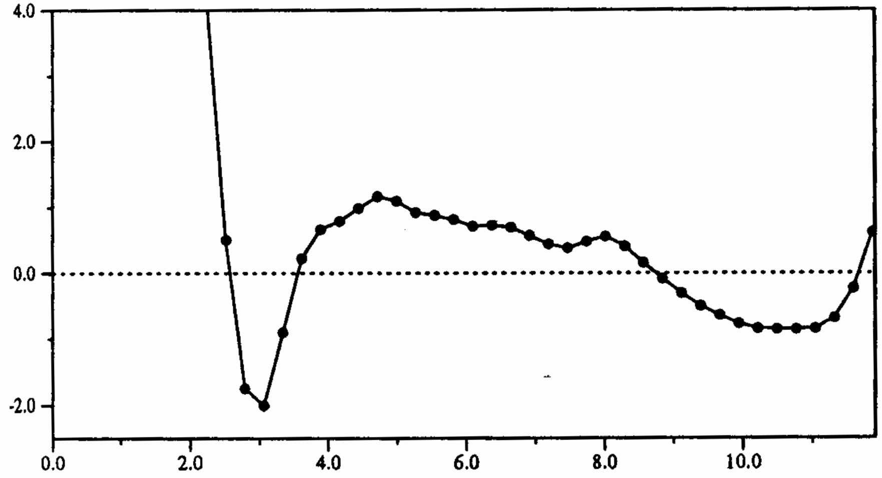
Scoring

Statistical Potential… Hydrogen Bonds

Long range free energy



Short range free energy



Free energy of the protein backbone hydrogen bond N · · · O compiled from a database of 289 X-ray structures

$$\rho_{NO}(r) = \sum_{ij} \delta(r - r_{ij})$$

$$g_{NO}(r) = \frac{\rho_{NO}(r)}{\rho^2}$$

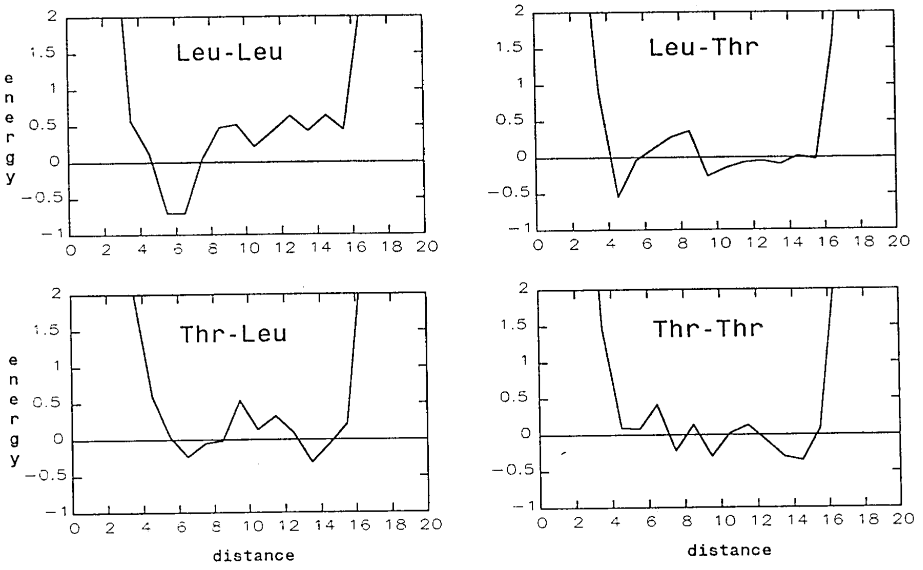
$$w_{NO}(r) = -kT \ln\left(g_{NO}(r)\right)$$

Sippl (1996). JMB ***260*** *pp644*

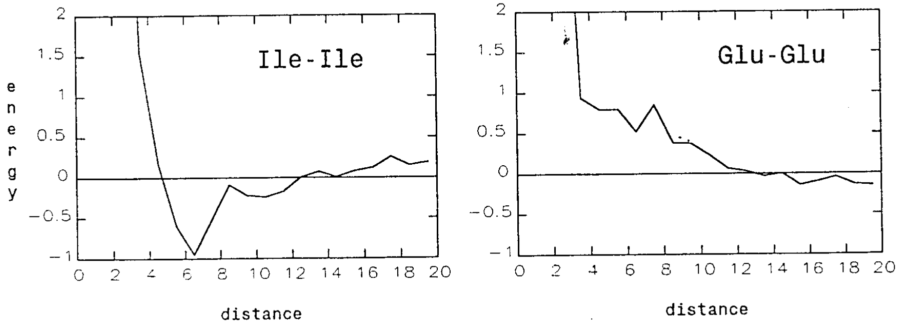
Scoring

Statistical Potential… Distance Potentials

Long range free energy



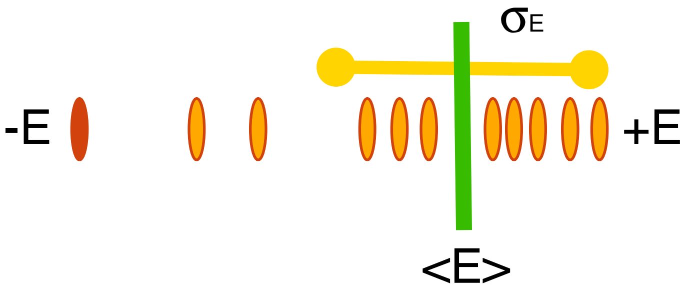
Short range free energy



*Sippl (1993). JCAM **7** pp473*

Significance of an alignment (score)

Energy Z-score the model with respect the energy of random models (or rest of decoys).



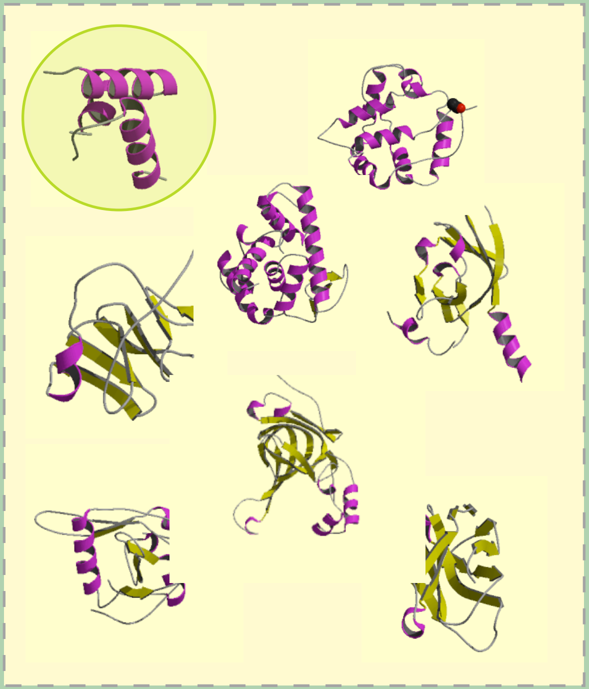
$$Zscore = \frac{\left(\langle E \rangle - E_m\right)}{\sigma_E}$$

Prosall

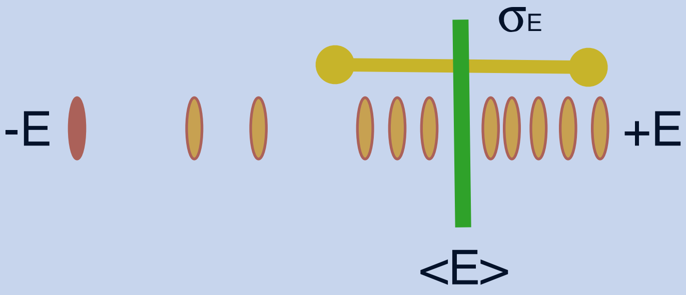
http://www.came.sbg.ac.at

Deriving

Structural space



Scoring



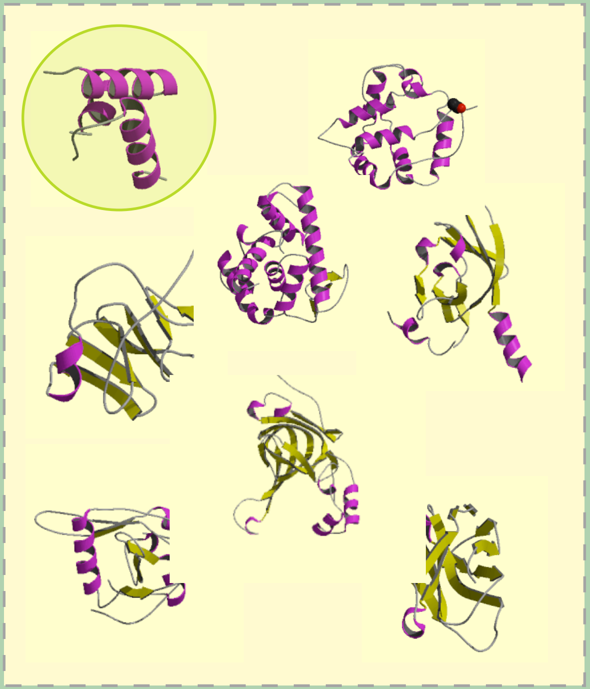
$$Zscore = \frac{\left(\langle E \rangle - E_m\right)}{\sigma_E}$$

ANOLEA

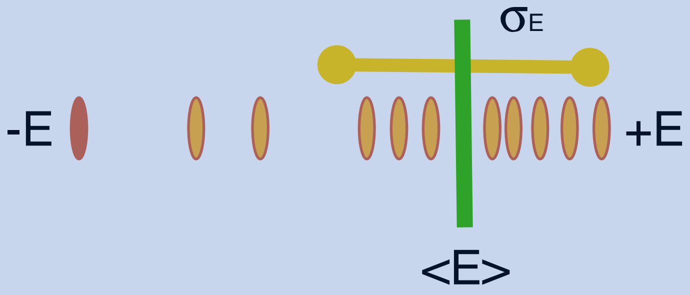
http://protein.bio.puc.cl/cardex/servers/anolea/

Deriving

Structural space



Scoring



$$Zscore = \frac{(\langle E \rangle - E_m)}{\sigma_E}$$

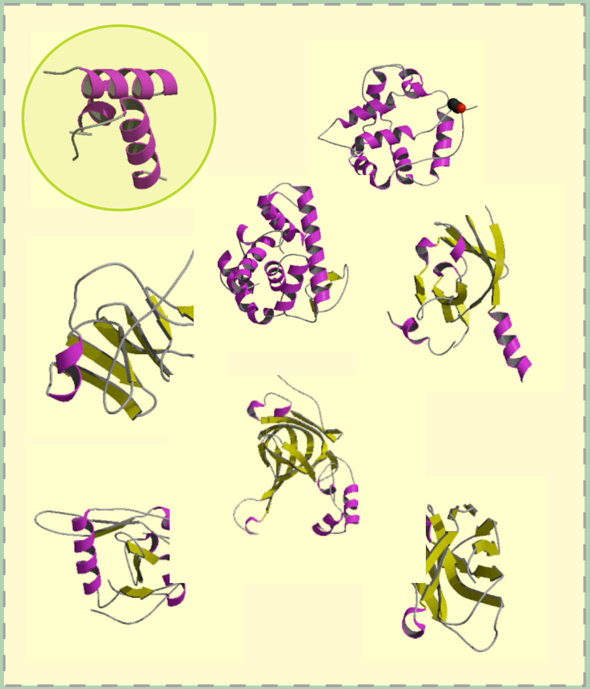
all atom potential

Verify3D

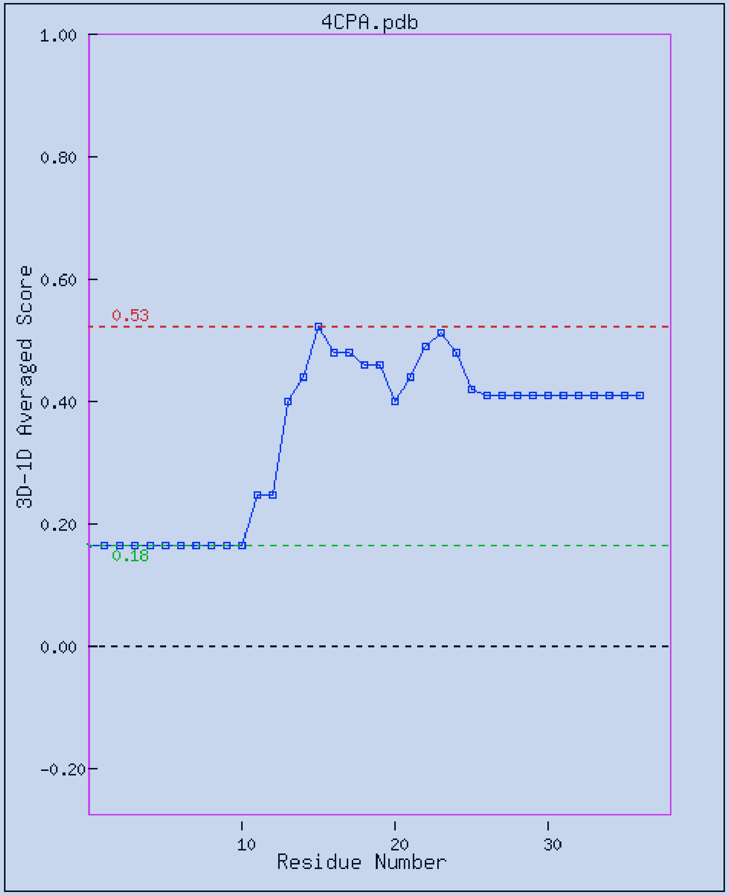
http://shannon.mbi.ucla.edu/DOE/Services/Verify_3D/

Deriving

Structural space



Scoring

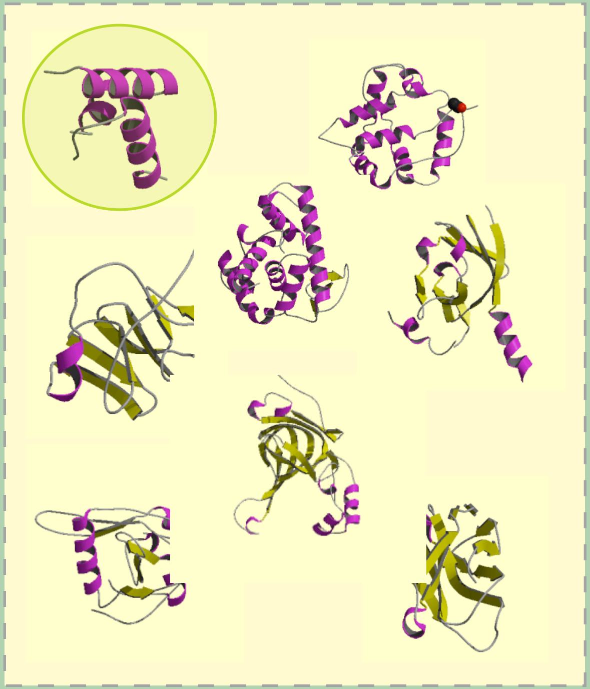


DFIRE

http://phyyz4.med.buffalo.edu/hzhou/dmonomer.html

Deriving

Structural space



Scoring

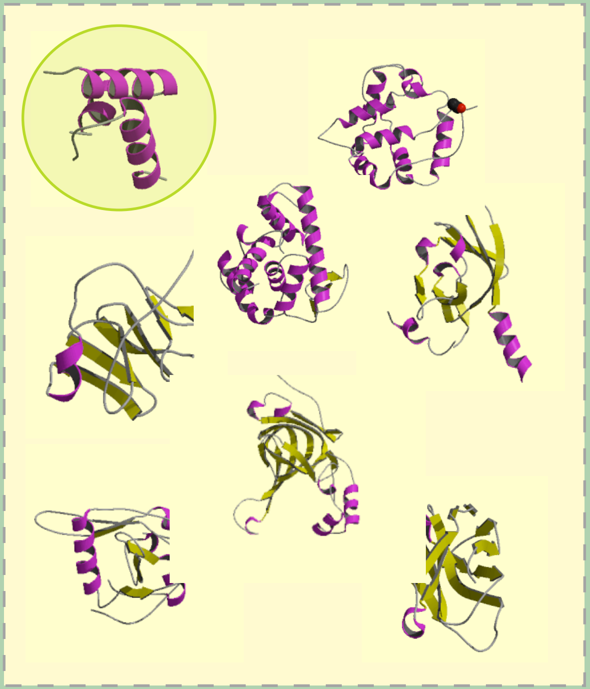
Pseudo-Energy with respect a ideal gas-phase reference state

DOPE (MODELLER)

http://www.salilab.org/modeller/

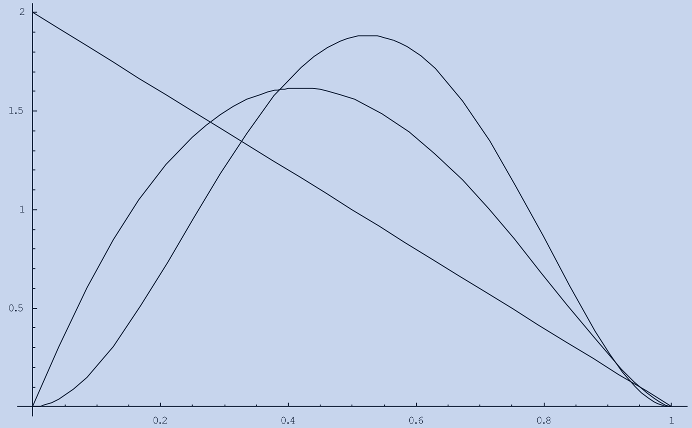
Deriving

Structural space



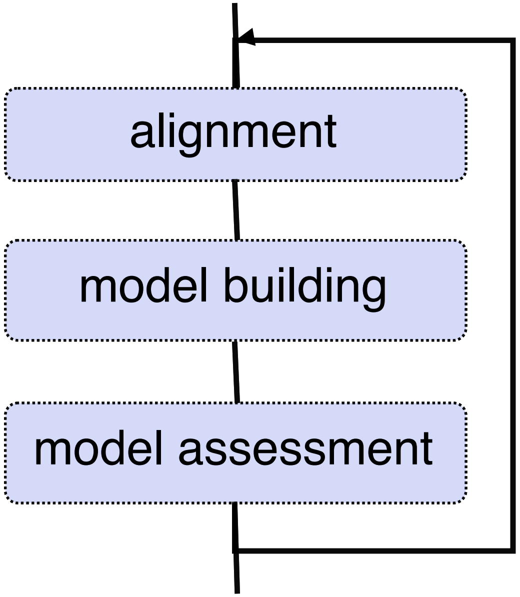
Scoring

Pseudo-Energy with respect a ideal spherical protein as a reference state



MOULDER

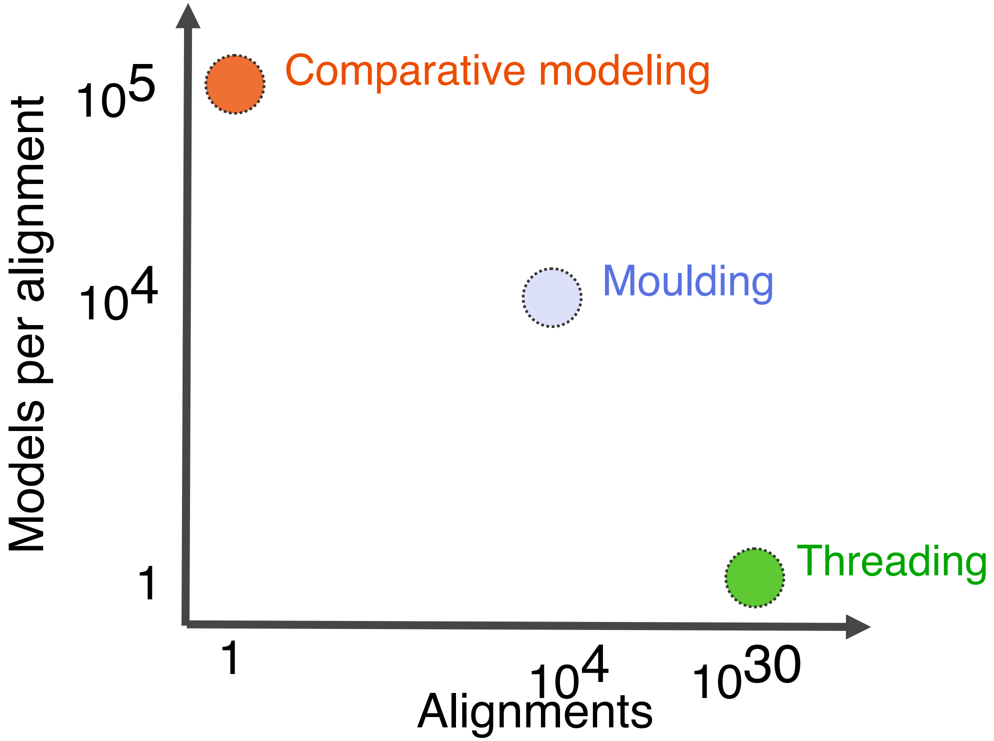
Moulding: iterative alignment, model building, model assessment



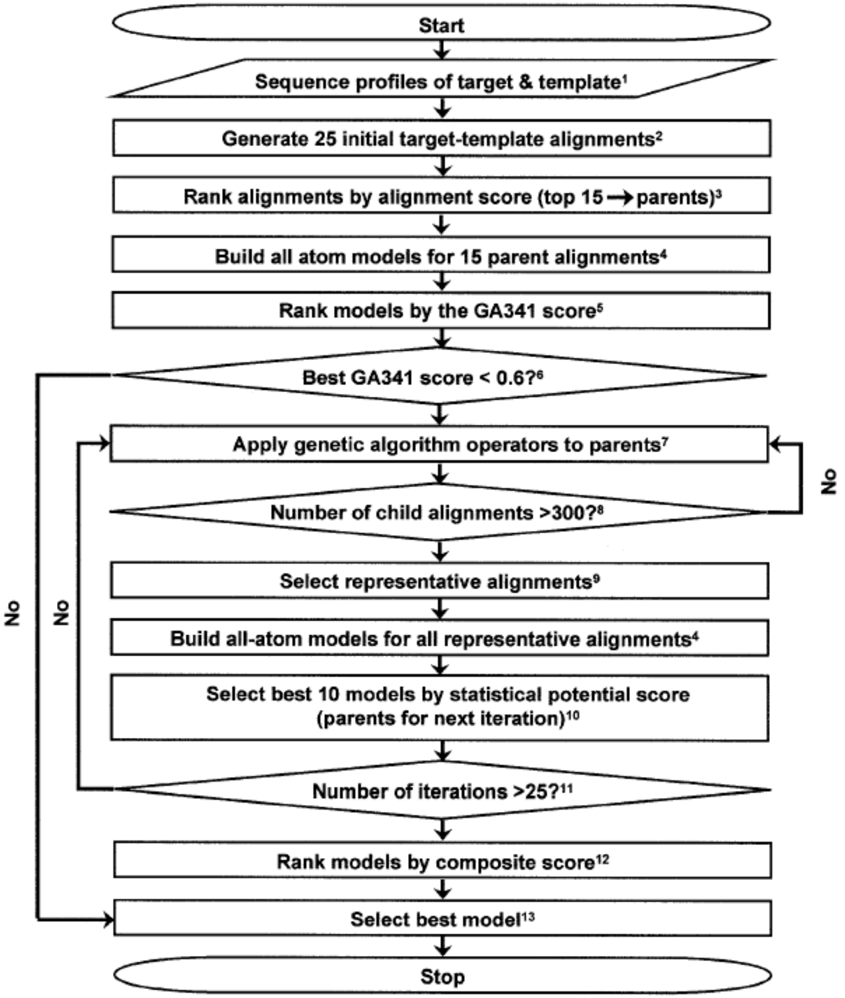
alignment

model building

model assessment



Iterative process… MOULDER



Start

Sequence profiles of target & template[1]

Generate 25 initial target-template alignments[2]

Rank alignments by alignment score (top 15 → parents)[3]

Build all atom models for 15 parent alignments[4]

Rank models by the GA341 score[5]

Best GA341 score < 0.6?[6]

Apply genetic algorithm operators to parents[7]

Number of child alignments >300?[8]

Select representative alignments[9]

Build all-atom models for all representative alignments[4]

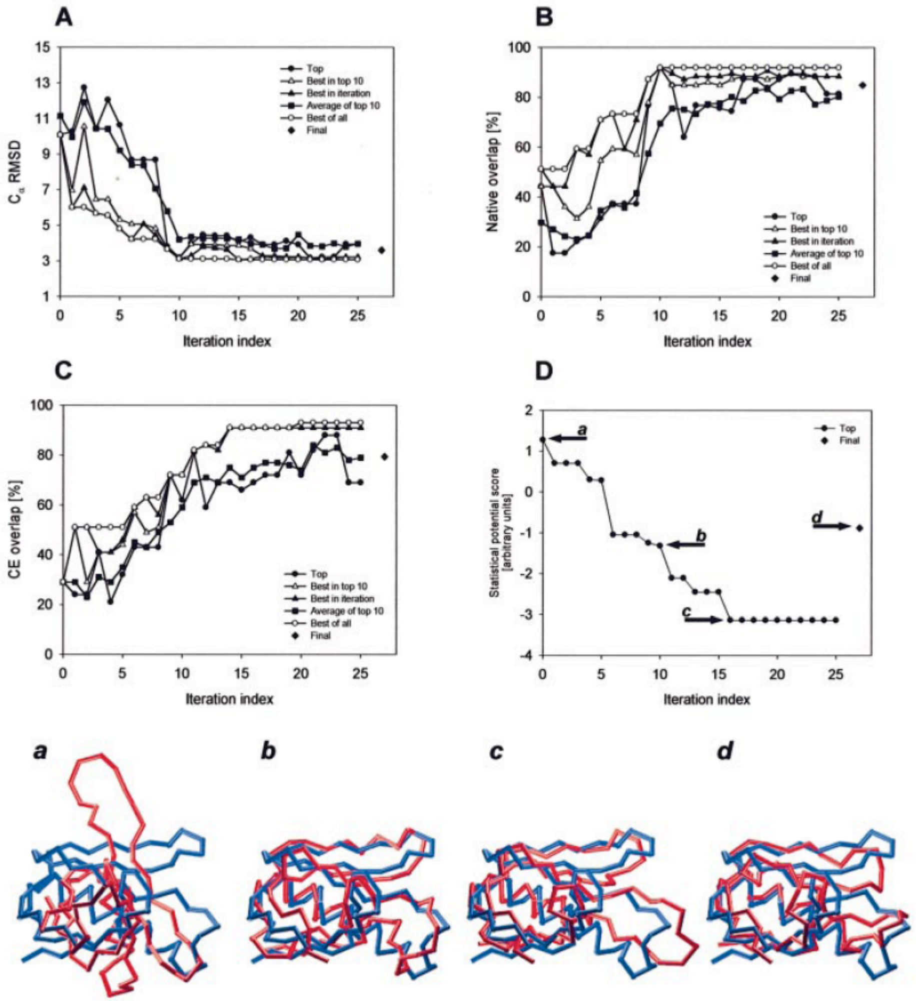
Select best 10 models by statistical potential score (parents for next iteration)[10]

Number of iterations >25?[11]

Rank models by composite score[12]

Select best model[13]

Stop



*John & Sali (2003). NAR **31** pp3982*

Genetic algorithm operators

Single point cross-over

```
...TSSQ-NMKLGVFWGY---...
...V-SSCN---GDLHMKVGV...

...TSSQNMK---LGVFWGY...
...VSSCNGDLHMKV---GV...
```

→

```
...TSSQ-NMK---LGVFWGY...
...V-SSCNGDLHMKV---GV...

...TSSQNMKLGVFWGY---...
...VSSCN---GDLHMKVGV...
```

Gap insertion

```
...TSSQNMKLGVFWGY...
...VSSCNGDLHMKVGV...
```

→

```
...TSSQN--MKLGVFWGY...
...VSSCNGDLHMKVG--V...
```

Gap shift

```
...T--SSQNMKLGVFWGY...
...VSSCNGDLHMKVGV--...
```

→

```
...-T-SSQNMKLGVFWGY...
...VSSCNGDLHMKVGV--...

...T-S-SQNMKLGVFWGY...
...VSSCNGDLHMKVGV--...

...--TSSQNMKLGVFWGY...
...VSSCNGDLHMKVGV--...

...TS--SQNMKLGVFWGY...
...VSSCNGDLHMKVGV--...
```

Also, “two point crossover” and “gap deletion”.

Composite model assessment score

Weighted linear combination of several scores:

- Pair (P_p) and surface (P_s) statistical potentials;
- Structural compactness (S_c);
- Harmonic average distance score (H_a);
- Alignment score (A_s).

$$Z = 0.17\ Z(P_P) + 0.02\ Z(P_S) + 0.10\ Z(S_C) + 0.26\ Z(H_a) + 0.45\ (A_S)$$

$Z(\text{score}) = (\text{score}- \mu)/\sigma$

μ … average score of all models

σ … standard deviation of the scores

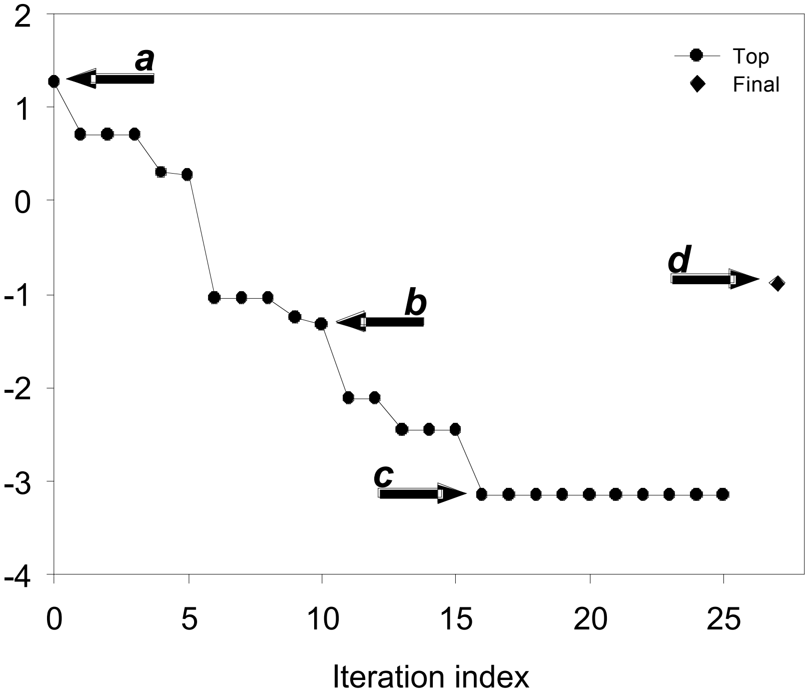
Benchmark with the “very difficult” test set

D. Fischer threading test set of 68 structural pairs (a subset of 19)

Target -template	Sequence identity [%]	Coverage [% aa]	Initial prediction		Final prediction		Best prediction	
			Cα RMSD [Å]	CE overlap [%]	Cα RMSD [Å]	CE overlap [%]	Cα RMSD [Å]	CE overlap [%]
1ATR-1ATN	13.8	94.3	19.2	20.2	18.8	20.2	17.1	24.6
1BOV-1LTS	4.4	83.5	10.1	29.4	3.6	79.4	3.1	92.6
1CAU-1CAU	18.8	96.7	11.7	15.6	10.0	27.4	7.6	47.4
1COL-1CPC	11.2	81.4	8.6	44.0	5.6	58.6	4.8	59.3
1LFB-1HOM	17.6	75.0	1.2	100.0	1.2	100.0	1.1	100.0
1NSB-2SIM	10.1	89.2	13.2	20.2	13.2	20.1	12.3	26.8
1RNH-1HRH	26.6	91.2	13.0	21.2	4.8	35.4	3.5	57.5
1YCC-2MTA	14.5	55.1	3.4	72.4	5.3	58.4	3.1	75.0
2AYH-1SAC	8.8	78.4	5.8	33.8	5.5	48.0	4.8	64.9
2CCY-1BBH	21.3	97.0	4.1	52.4	3.1	73.0	2.6	77.0
2PLV-1BBT	20.2	91.4	7.3	58.9	7.3	58.9	6.2	60.7
2POR-2OMF	13.2	97.3	18.3	11.3	11.4	14.7	10.5	25.9
2RHE-1CID	21.2	61.6	9.2	33.7	7.5	51.1	4.4	71.1
2RHE-3HLA	2.4	96.0	8.1	16.5	7.6	9.4	6.7	43.5
3ADK-1GKY	19.5	100.0	13.8	26.6	11.5	37.7	7.7	48.1
3HHR-1TEN	18.4	98.9	7.3	60.9	6.0	66.7	4.9	79.3
4FGF-81IB	14.1	98.6	11.3	24.0	9.3	30.6	5.4	41.2
6XIA-3RUB	8.7	44.1	10.5	14.5	10.1	11.0	9.0	34.3
9RNT-2SAR	13.1	88.5	5.8	41.7	5.1	51.2	4.8	69.0
AVERAGE	**14.2**	**85.2**	**9.6**	**36.7**	**7.7**	**44.8**	**6.3**	**57.8**

Application to a difficult modeling case

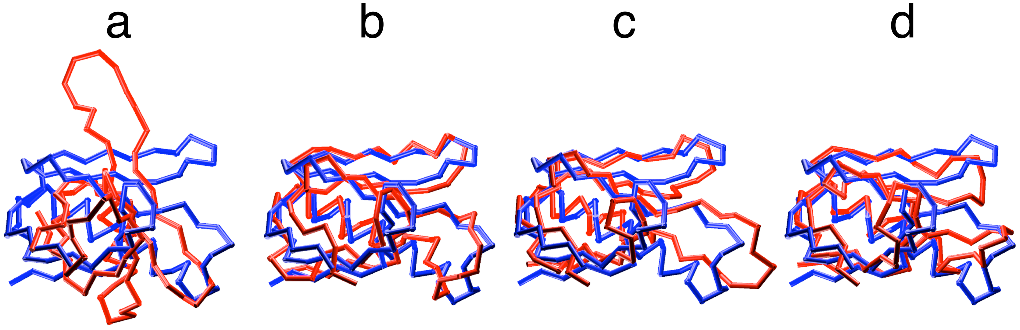
1BOV-1LTS

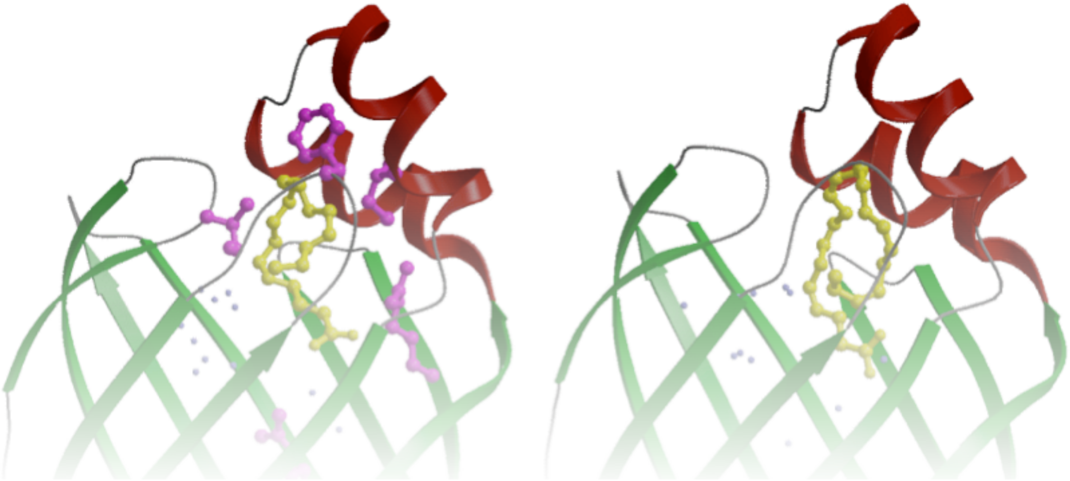


Sequence identity 4.4%

Initial model Cα RMSD 10.1Å

Final model Cα RMSD 3.6Å





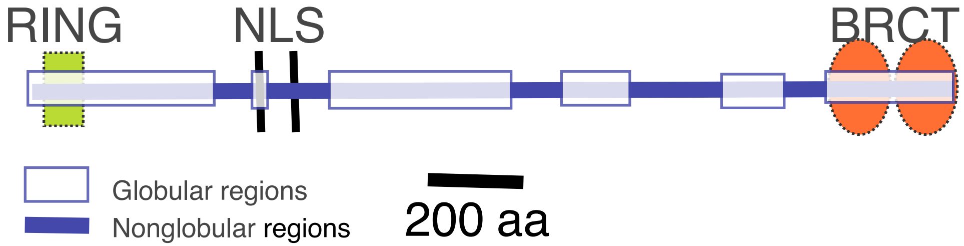
Modeling genes

Structural analysis of missense mutations in human BRCA1 BRCT domains

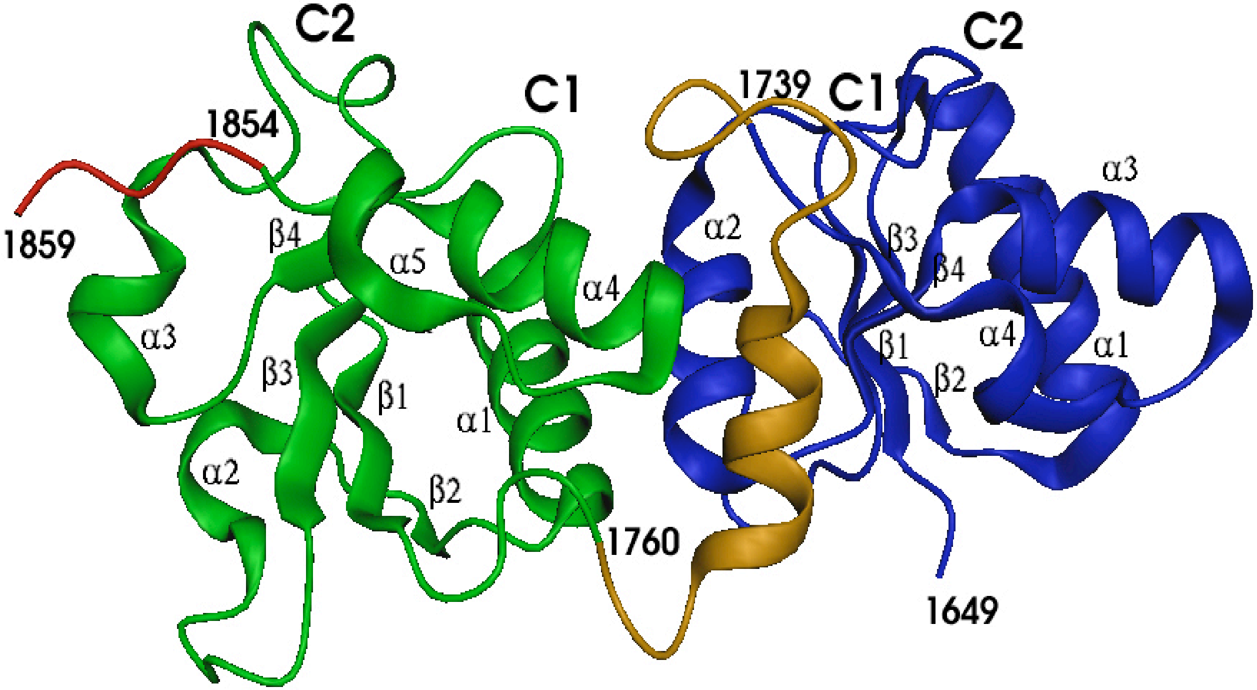
Cannot measure the functional impact of every possible SNP at all positions in each protein! Thus, prediction based on general principles of protein structure is needed.

Mirkovic et al. (2004) Cancer Research 64 pp3790

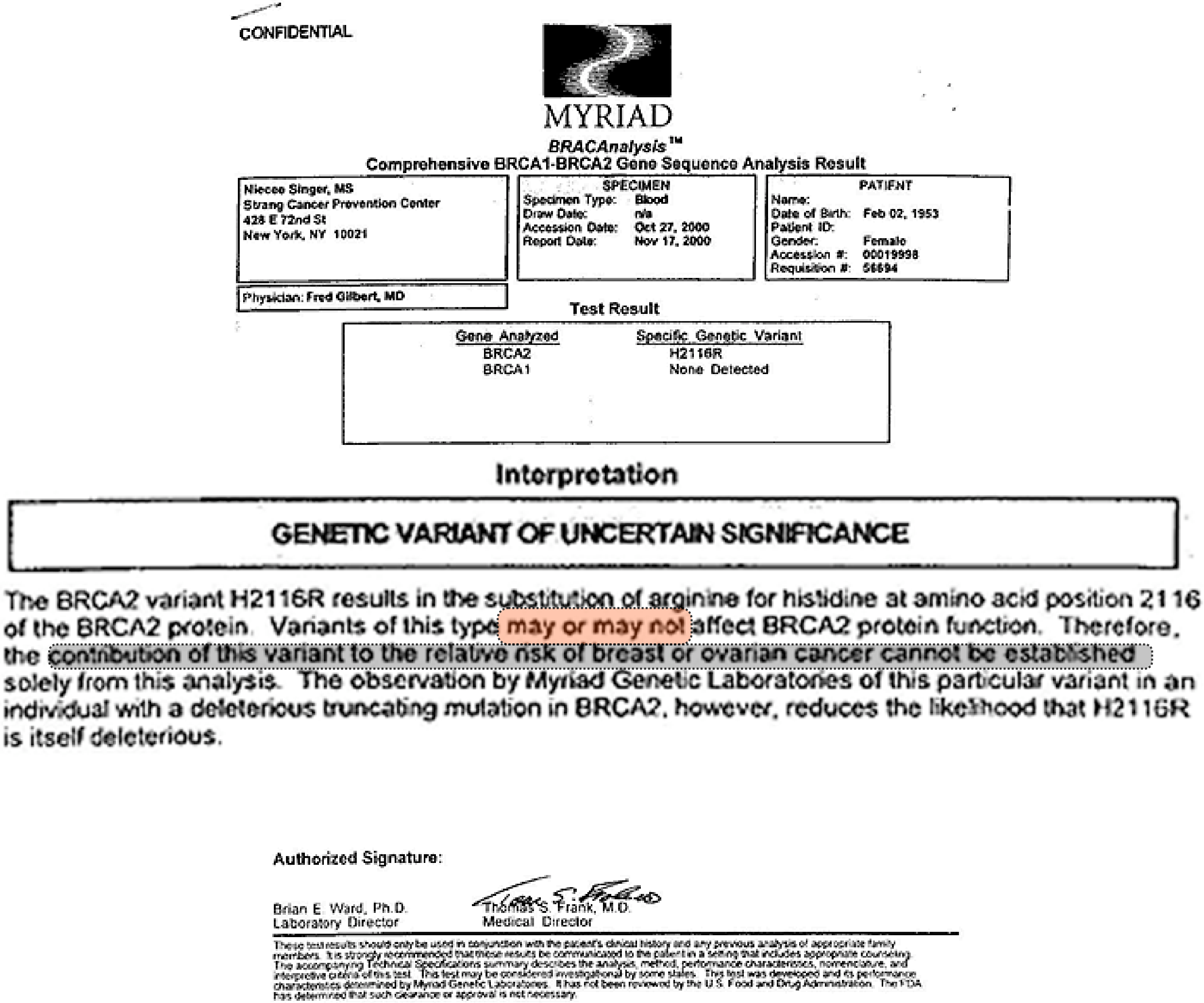
Human BRCA1 and its two BRCT domains



BRCA1 BRCT repeats, 1jnx



Williams, Green, Glover. Nat.Struct.Biol. 8, 838, 2001



CONFIDENTIAL

MYRIAD

BRACAnalysis™

Comprehensive BRCA1-BRCA2 Gene Sequence Analysis Result

Niecee Singer, MS
Strang Cancer Prevention Center
428 E 72nd St
New York, NY 10021

Physician: Fred Gilbert, MD

SPECIMEN	
Specimen Type:	Blood
Draw Date:	n/a
Accession Date:	Oct 27, 2000
Report Date:	Nov 17, 2000

PATIENT	
Name:	
Date of Birth:	Feb 02, 1953
Patient ID:	
Gender:	Female
Accession #:	00019998
Requisition #:	56694

Test Result

Gene Analyzed	Specific Genetic Variant
BRCA2	H2116R
BRCA1	None Detected

Interpretation

GENETIC VARIANT OF UNCERTAIN SIGNIFICANCE

The BRCA2 variant H2116R results in the substitution of arginine for histidine at amino acid position 2116 of the BRCA2 protein. Variants of this type may or may not affect BRCA2 protein function. Therefore, the contribution of this variant to the relative risk of breast or ovarian cancer cannot be established solely from this analysis. The observation by Myriad Genetic Laboratories of this particular variant in an individual with a deleterious truncating mutation in BRCA2, however, reduces the likelihood that H2116R is itself deleterious.

Authorized Signature:

Brian E. Ward, Ph.D.
Laboratory Director

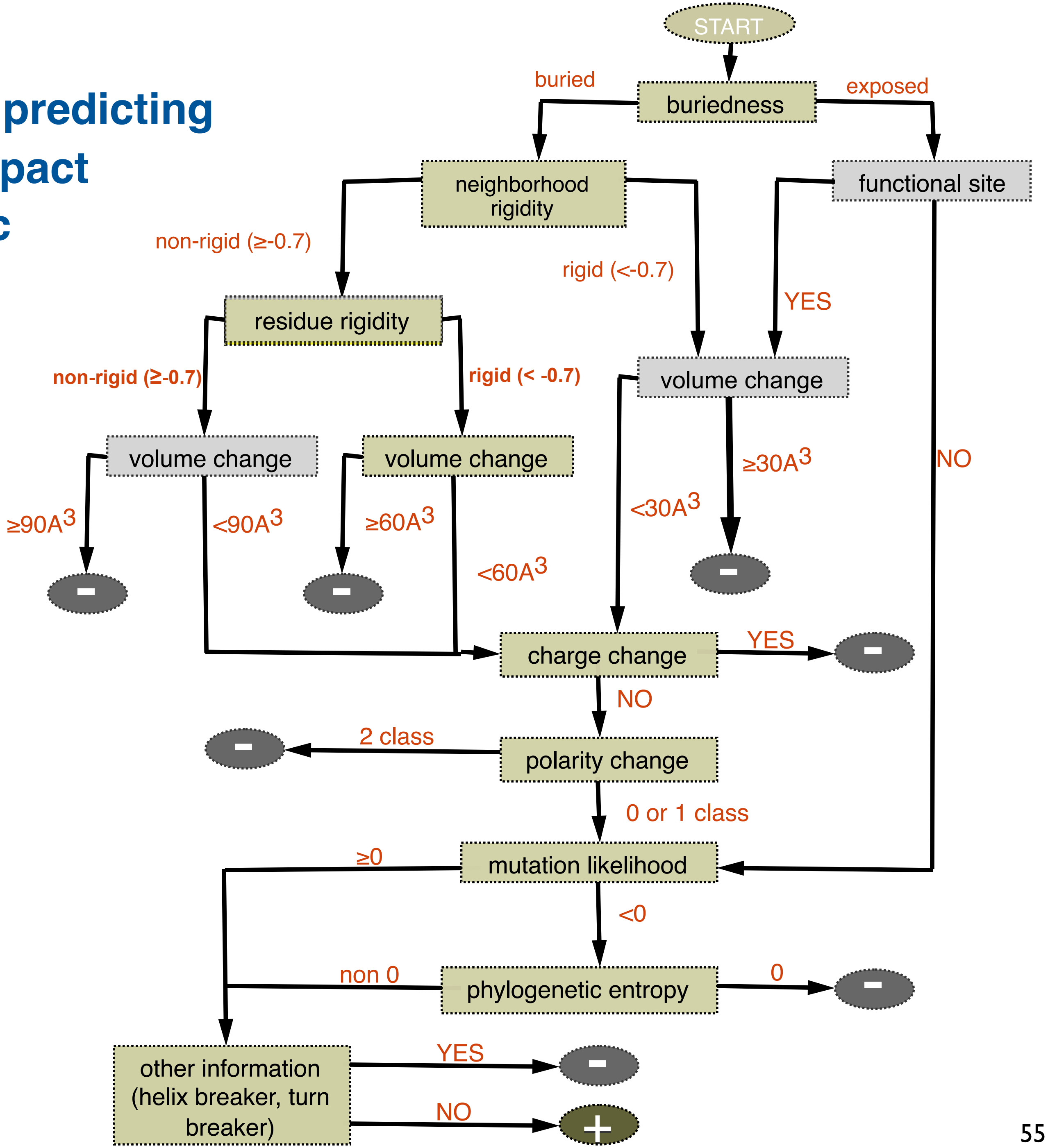
Thomas S. Frank, M.D.
Medical Director

These test results should only be used in conjunction with the patient's clinical history and any previous analysis of appropriate family members. It is strongly recommended that these results be communicated to the patient in a setting that includes appropriate counseling. The accompanying Technical Specifications summary describes the analysis, method, performance characteristics, nomenclature, and interpretive criteria of this test. This test may be considered investigational by some states. This test was developed and its performance characteristics determined by Myriad Genetic Laboratories. It has not been reviewed by the U.S. Food and Drug Administration. The FDA has determined that such clearance or approval is not necessary.

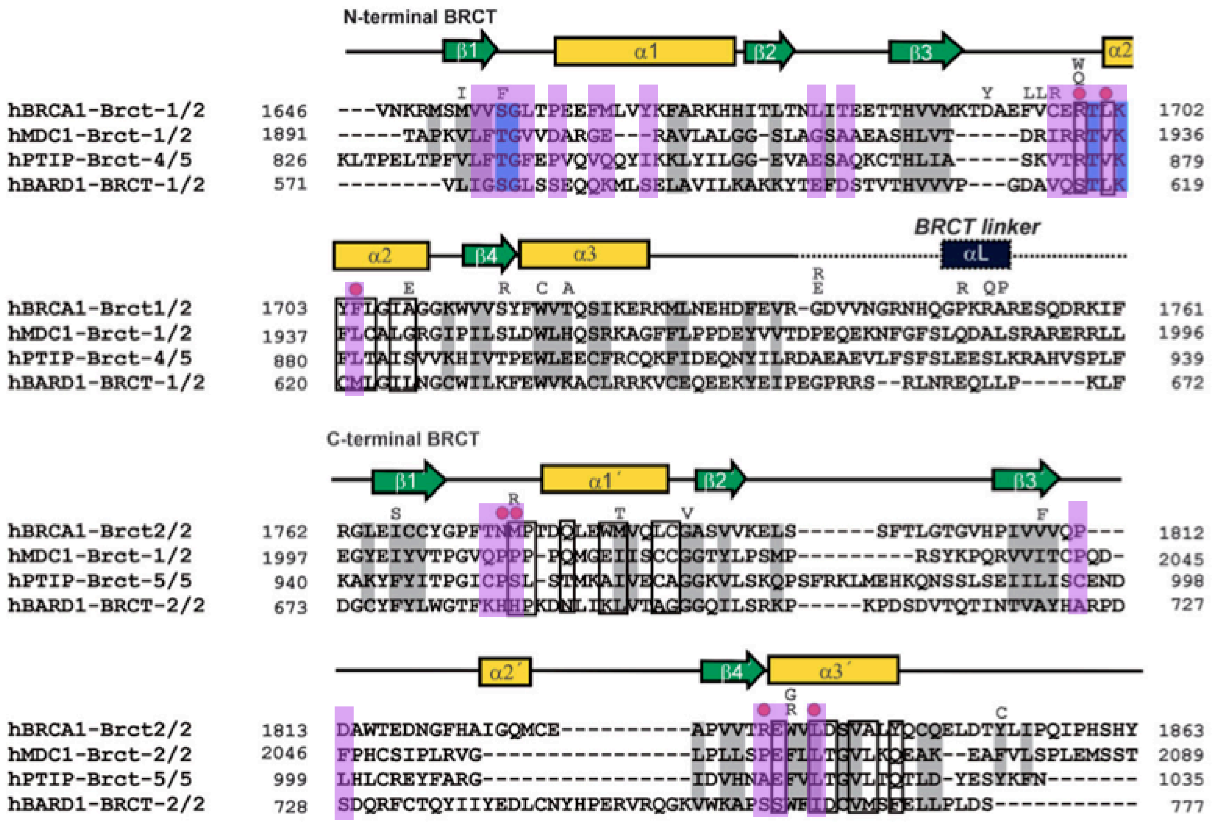
Missense mutations in BRCT domains by function

	cancer associated	not cancer associated	?
no transcription activation	C1697R R1699W A1708E S1715R P1749R M1775R		M1652K L1657P E1660G H1686Q R1699Q K1702E Y1703HF 1704S L1705PS 1715NS1 722FF17 34LG173 8EG174 3RA175 2PF1761 I F1761S M1775E M1775K L1780P I1807S V1833E A1843T
transcription activation		M1652I A1669S	V1665M D1692N G1706A D1733G M1775V P1806A
?			M1652T V1653M L1664P T1685A T1685I M1689R D1692Y F1695L V1696L R1699L G1706E W1718C W1718S T1720A W1730S F1734S E1735K V1736A G1738R D1739E D1739G D1739Y V1741G H1746N R1751P R1751Q R1758G L1764P I1766S P1771L T1773S P1776S D1778N D1778G D1778H M1783T C1787S G1788D G1788V G1803A V1804D V1808A V1809A V1809F V1810G Q1811R P1812S N1819S A1823T V1833M W1837R W1837G S1841N A1843P T1852S P1856T P1859R

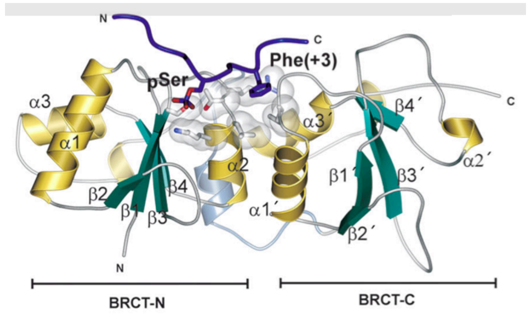
“Decision” tree for predicting functional impact of genetic variants

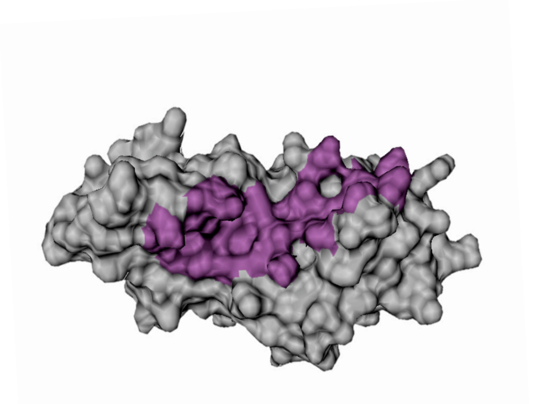


Putative binding site on BRCA1



N-terminal BRCT
β1
α1
β2
β3
α2
I F Y LLR W Q
hBRCA1-Brct-1/2 1646 ---VNKRMSMVVSGLTPEEFMLVYKFARKHHITLTNLITEETTHVVMKTDAEFVCERTLK 1702
hMDC1-Brct-1/2 1891 -----TAPKVLFTGVVDARGE---RAVLALGG-SLAGSAAEASHLVT-----DRIRRTVK 1936
hPTIP-Brct-4/5 826 KLTPELTPFVLFTGFEPVQVQQYIKKLYILGG-EVAESAQKCTHLIA-----SKVTRTVK 879
hBARD1-BRCT-1/2 571 --------VLIGSGLSSEQQKMLSELAVILKAKKYTEFDSTVTHVVVP---GDAVQSTLK 619
BRCT linker
α2
β4
α3
αL
E R C A R E R QP
hBRCA1-Brct1/2 1703 YFLGIAGGKWVVSYFWVTQSIKERKMLNEHDFEVR-GDVVNGRNHQGPKRARESQDRKIF 1761
hMDC1-Brct-1/2 1937 FLCALGRGIPILSLDWLHQSRKAGFFLPPDEYVVTDPEQEKNFGFSLQDALSRARERRLL 1996
hPTIP-Brct-4/5 880 FLTAISVVKHIVTPEWLEECFRCQKFIDEQNYILRDAEAEVLFSFSLEESLKRAHVSPLF 939
hBARD1-BRCT-1/2 620 CMIGILNGCWILKFEWVKACLRRKVCEQEEKYEIPEGPRRS--RLNREQLLP-----KLF 672
C-terminal BRCT
β1
α1´
β2
β3
S R T V F
hBRCA1-Brct2/2 1762 RGLEICCYGPFTNMPTDQLEWMVQLCGASVVKELS------SFTLGTGVHPIVVVQP--- 1812
hMDC1-Brct-1/2 1997 EGYEIYVTPGVQPPP-PQMGEIISCCGGTYLPSMP---------RSYKPQRVVITCPQD- 2045
hPTIP-Brct-5/5 940 KAKYFYITPGICPSL-STMKAIVECAGGKVLSKQPSFRKLMEHKQNSSLSEIILISCEND 998
hBARD1-BRCT-2/2 673 DGCYFYLWGTFKHHPKDNLIKLVTAGGGQILSRKP-----KPDSDVTQTINTVAYHARPD 727
α2´
β4
α3´
G R C
hBRCA1-Brct2/2 1813 DAWTEDNGFHAIGQMCE----------APVVTREWVLDSVALYQCQELDTYLIPQIPHSHY 1863
hMDC1-Brct-2/2 2046 FPHCSIPLRVG----------------LPLLSPEFILTGVLKQEAK--EAFVLSPLEMSST 2089
hPTIP-Brct-5/5 999 LHLCREYFARG----------------IDVHNAEFVLTGVLTQTLD-YESYKFN------- 1035
hBARD1-BRCT-2/2 728 SDQRFCTQYIIYEDLCNYHPERVRQGKVWKAPSSWFIDCVMSFELLPLDS----------- 777





Putative binding site predicted in 2003 and accepted for publication on March 2004.

Williams *et al.* 2004 Nature Structure Biology. **June 2004 11**:519

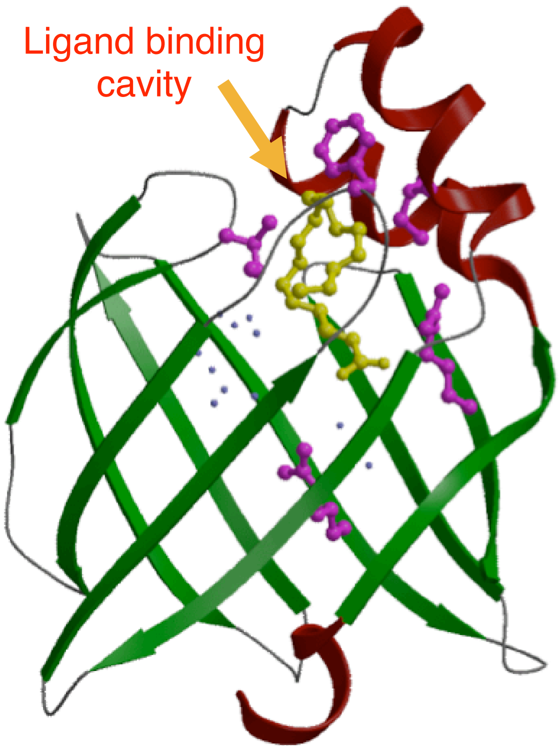
Mirkovic *et al.* 2004 Cancer Research. **June 2004 64**:3790

What is the physiological ligand of Brain Lipid-Binding Protein?

Predicting features of a model that are not present in the template

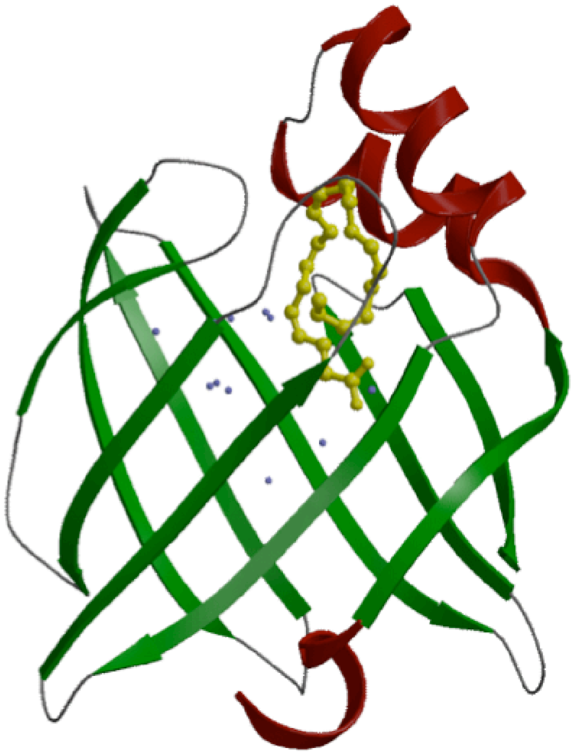
BLBP/oleic acid

Cavity is not filled



BLBP/docosahexaenoic acid

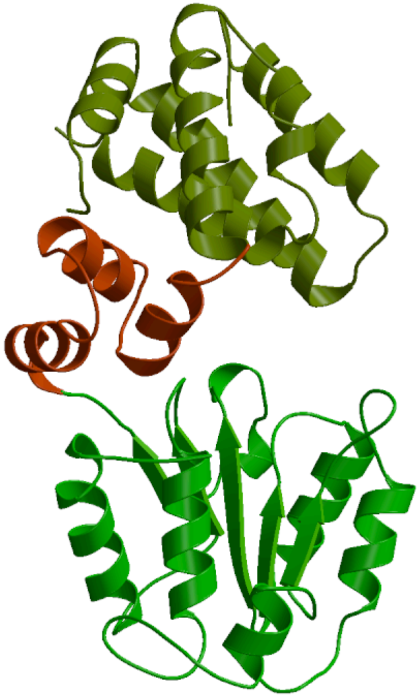
Cavity is filled



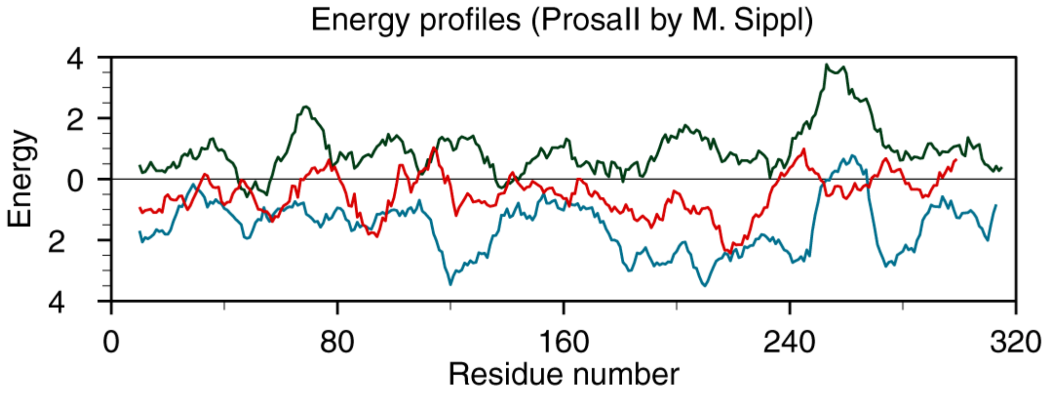
1. BLBP binds fatty acids.
2. Build a 3D model.
3. Find the fatty acid that fits most snuggly into the ligand binding cavity.

L. Xu, R. Sánchez, A. Šali, N. Heintz, J. Biol. Chem. 271, 24711, 1996.

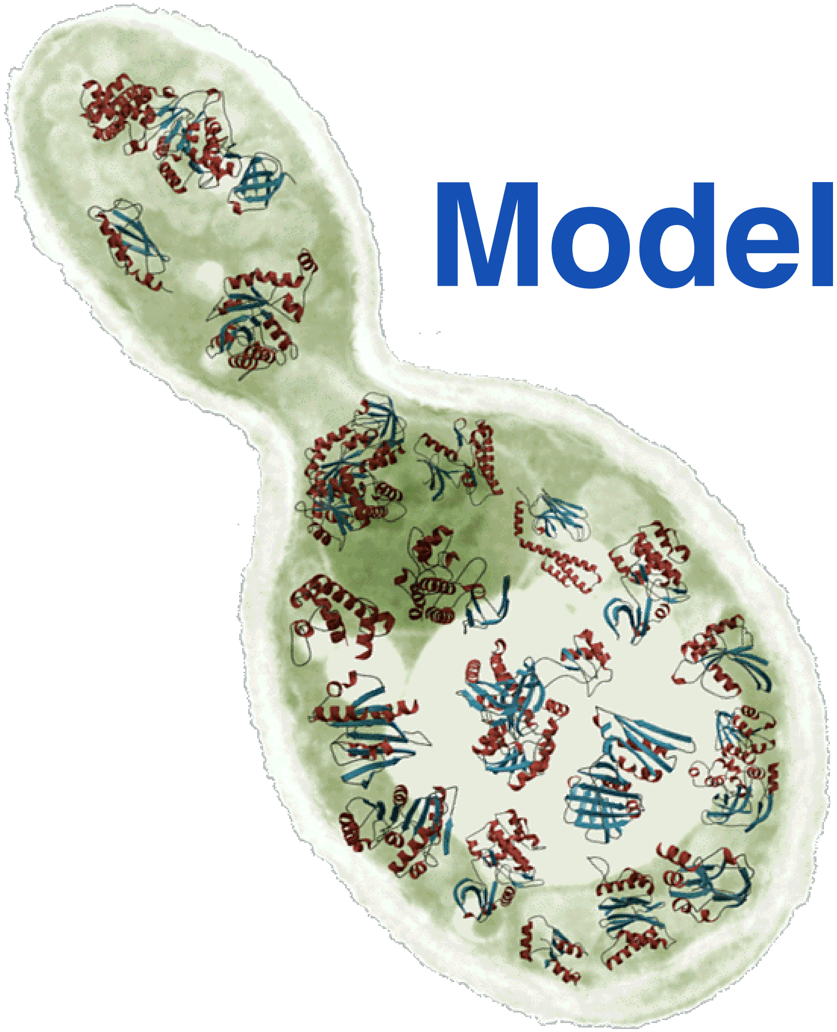
Does RuvB have the same fold as δ' of E.coli DNA polymerase III?



```
Ec d'   MRWYPWLRPDFEKLVASYQAGRG----HHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRG
RUVB    LEEYVGQPQVRSQMEIFIKAAKLRGDALDHLLIFGPPGLGKTTLANIVANEMG---------------

Ec d'   CQLMQAGTHPDYYTLAPEKGKATLGVDAVREVTEKLNEAARLGGAKVVWVTDAALLTDAAANALLKTL
RUVB    -----------VNLRTT-------SGPVLEKAGDLAAMLTNLEPHDVLFIDEIHRLSPVVEEVLYPAM

Ec d'   ------------------EEPPAETWFFLATREPERL---LATLRSRCRLHYLAPPPEQYAVTWLSRE
Ppdp    EDYQLDIMIGEGPAARSIKIDLPPFTLIGATTRAGSLTSPLRDRFGIVQRLEFY--QVPDLQYIVSRS

Ec d'   VTM-----SQDALLAALRLSAGSPGAALALFQ------------GDNWQARETLCQALAYSVPSGD--
RUVB    ARFMGLEMSDDGALEVARRARGTPRIANRLLRRVRDFAEVKHDGTISADIAAQALDMLNVDAEGFDYM

Ec d'   -WYSLLAALN---HEQAPARLHWLATLLMDALKR/VTNVDVPGLVAELANHL---SPSRLQAILGDVC
RUVB    DRKLLLAVIDKFF-GGPVGLDNLAAAIGEERETIE--DVLEPYLIQQGFLQRTPRGRMATTRAWNHFG

                Ec d'   HIREQLMSVAGANRELLITDLLLRIEHYLQPGVVLP
                RUVB    ITPPEMP-----------------------------
```



B. Guenther, et al. Cell 91, 335, 1997.
Yamada, K., et al. Proc.Nat.Acad.Sci.USA 98,1442, 2001.

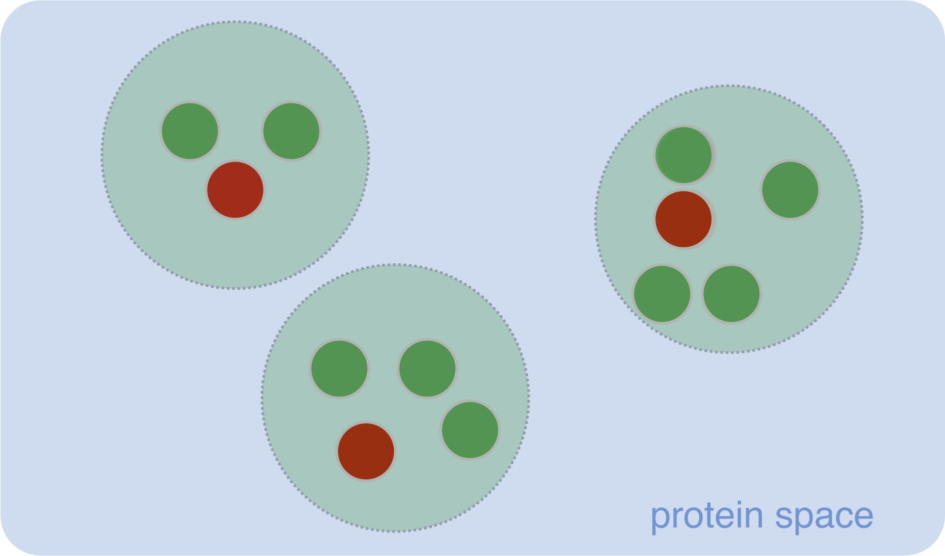


Modeling genomes

Structural Genomics

Characterize most protein sequences based on related known structures

1. The number of "**families**" is much **smaller** than the number of proteins.
2. **Any one** of the members of a family is **fine**.

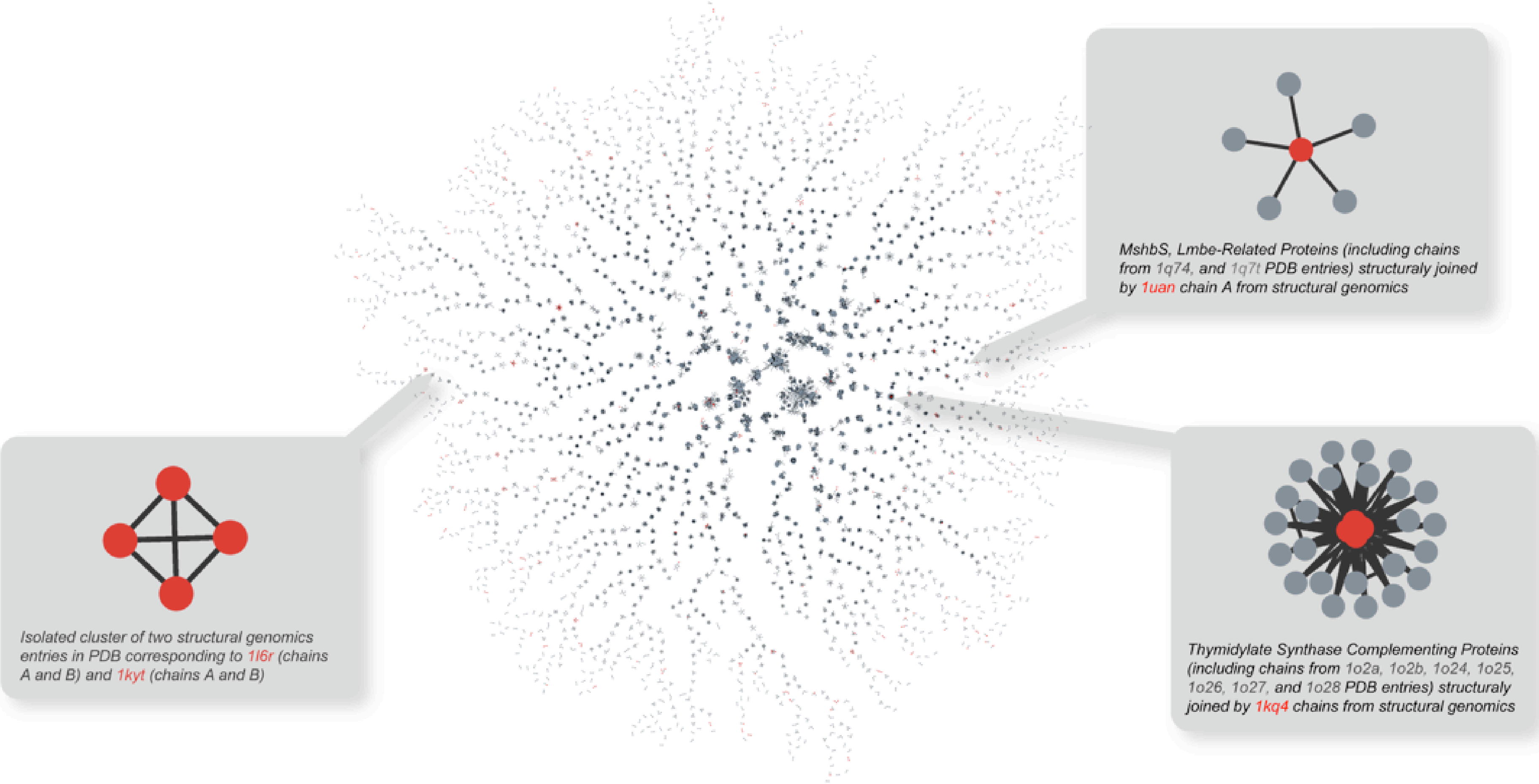


There are **~16,000** families (90%)
@ 30% sequence identity cutoff

*Sali. Nat. Struct. Biol. **5**, 1029, 1998.*
*Sali et al. Nat. Struct. Biol., **7**, 986, 2000.*
*Sali. Nat. Struct. Biol. **7**, 484, 2001.*
*Baker & Sali. Science **294,** 93, 2001**.***
*Vitkup et al. Nat. Struct. Biol. **8**, 559, 2001*

Structure Space & Structural Genomics

Structure map @ >30% sequence identity



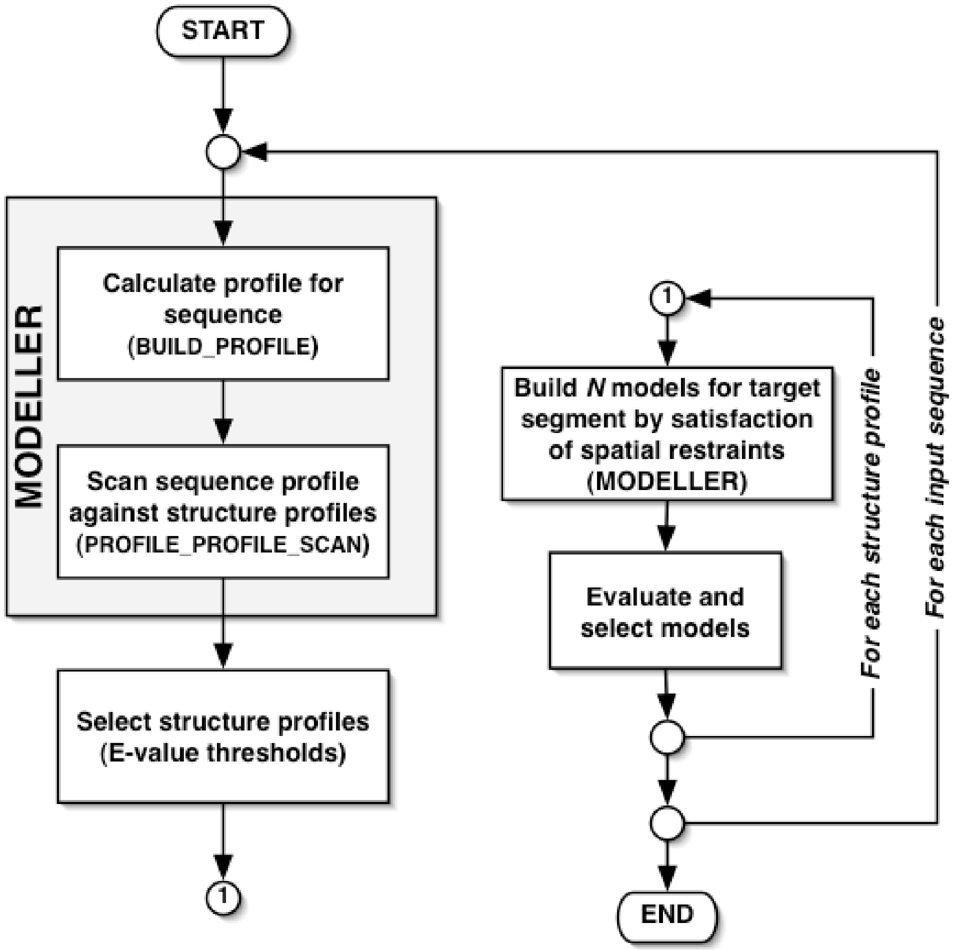
MshbS, Lmbe-Related Proteins (including chains from 1q74, and 1q7t PDB entries) structuraly joined by 1uan chain A from structural genomics

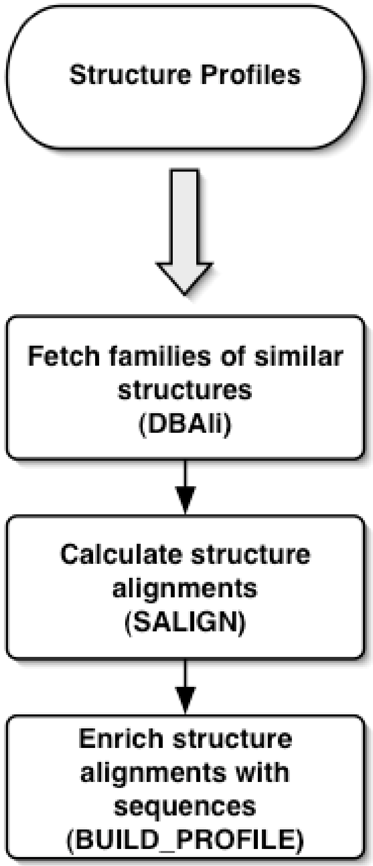
Isolated cluster of two structural genomics entries in PDB corresponding to 1l6r (chains A and B) and 1kyt (chains A and B)

Thymidylate Synthase Complementing Proteins (including chains from 1o2a, 1o2b, 1o24, 1o25, 1o26, 1o27, and 1o28 PDB entries) structuraly joined by 1kq4 chains from structural genomics

MODPIPE2.0

Large-Scale Protein Structure Modeling


START
MODELLER
Calculate profile for sequence (BUILD_PROFILE)
Scan sequence profile against structure profiles (PROFILE_PROFILE_SCAN)
Select structure profiles (E-value thresholds)
1
1
Build *N* models for target segment by satisfaction of spatial restraints (MODELLER)
Evaluate and select models
For each structure profile
For each input sequence
END


Structure Profiles
Fetch families of similar structures (DBAli)
Calculate structure alignments (SALIGN)
Enrich structure alignments with sequences (BUILD_PROFILE)

Eswar et.al., (2003) Nucl.Acids.Res. 31(13)

ModBase Statistics

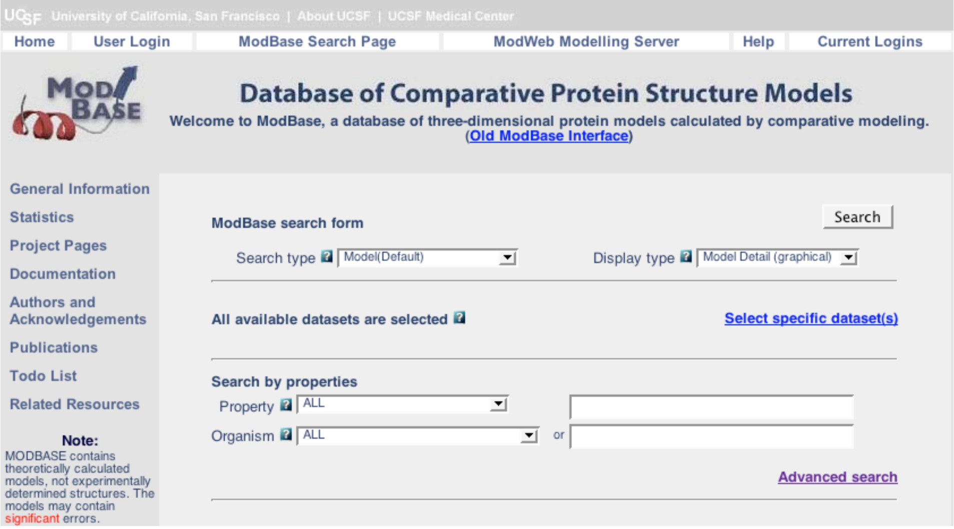
Large-scale modeling of the TrEMBL-SWISSPROT databases

Sequences (total)	1,679,742
Sequences (modeled)	964,442
Models	2,947,461

MODBASE

http://salilab.org/modbase

Search Page



UCSF University of California, San Francisco | About UCSF | UCSF Medical Center

Home | User Login | ModBase Search Page | ModWeb Modelling Server | Help | Current Logins

Database of Comparative Protein Structure Models

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

- General Information
- Statistics
- Project Pages
- Documentation
- Authors and Acknowledgements
- Publications
- Todo List
- Related Resources

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

ModBase search form — Search

Search type: Model(Default) — Display type: Model Detail (graphical)

All available datasets are selected — Select specific dataset(s)

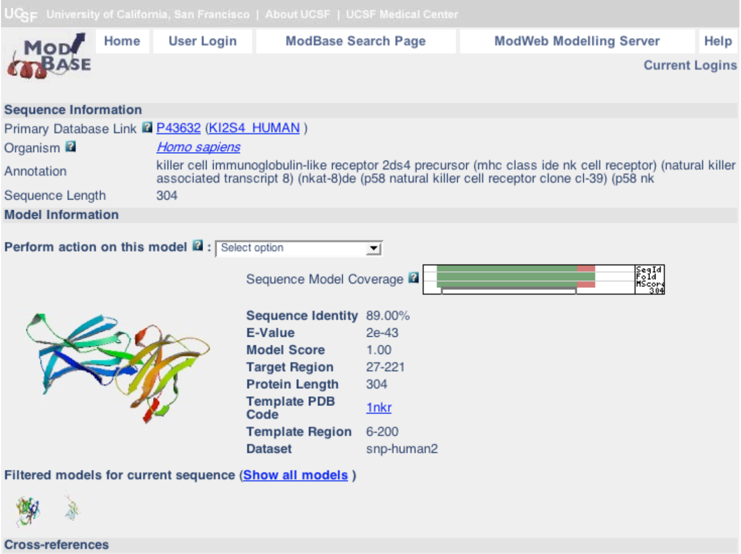
Search by properties

Property: ALL

Organism: ALL or

Advanced search

Model Details



UCSF University of California, San Francisco | About UCSF | UCSF Medical Center

Home | User Login | ModBase Search Page | ModWeb Modelling Server | Help

Current Logins

Sequence Information

Primary Database Link	P43632 (KI2S4_HUMAN)
Organism	*Homo sapiens*
Annotation	killer cell immunoglobulin-like receptor 2ds4 precursor (mhc class ide nk cell receptor) (natural killer associated transcript 8) (nkat-8)de (p58 natural killer cell receptor clone cl-39) (p58 nk
Sequence Length	304

Model Information

Perform action on this model : Select option

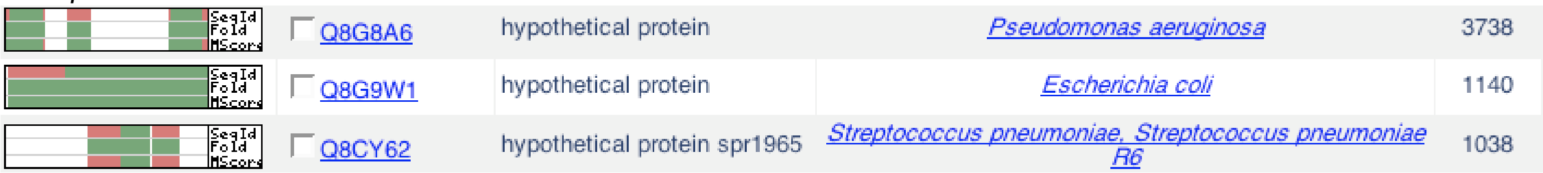
Sequence Model Coverage

Sequence Identity	89.00%
E-Value	2e-43
Model Score	1.00
Target Region	27-221
Protein Length	304
Template PDB Code	1nkr
Template Region	6-200
Dataset	snp-human2

Filtered models for current sequence (Show all models)

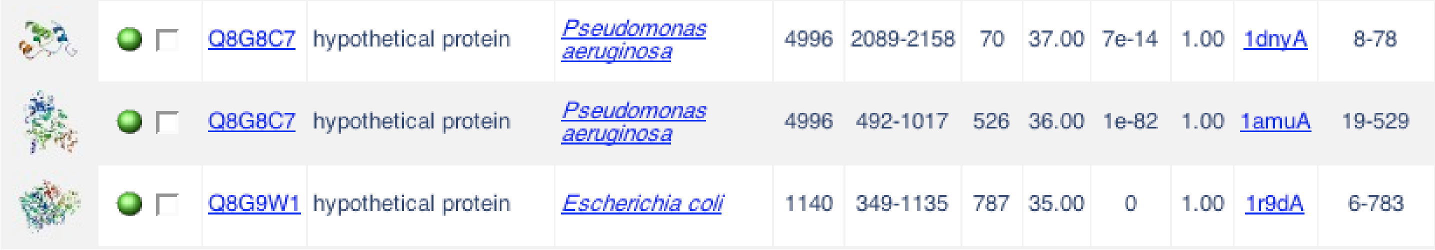
Cross-references

Sequence Overview



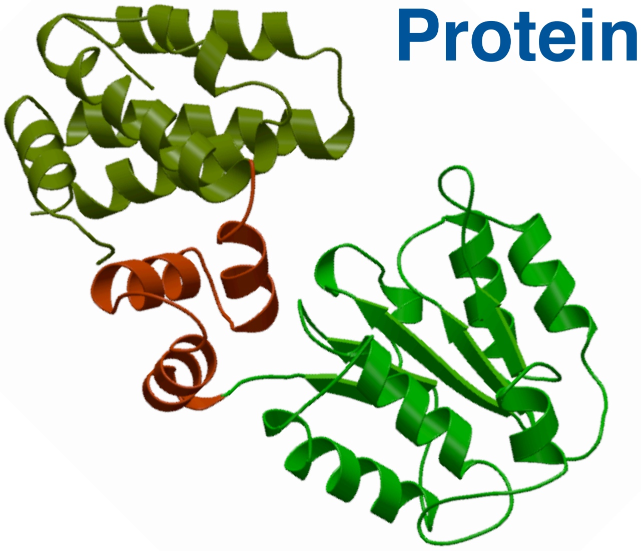
SeqId Fold MScore	☐	Q8G8A6	hypothetical protein	*Pseudomonas aeruginosa*	3738
SeqId Fold MScore	☐	Q8G9W1	hypothetical protein	*Escherichia coli*	1140
SeqId Fold MScore	☐	Q8CY62	hypothetical protein spr1965	*Streptococcus pneumoniae, Streptococcus pneumoniae R6*	1038

Model Overview



	●	☐	Q8G8C7	hypothetical protein	*Pseudomonas aeruginosa*	4996	2089-2158	70	37.00	7e-14	1.00	1dnyA	8-78
	●	☐	Q8G8C7	hypothetical protein	*Pseudomonas aeruginosa*	4996	492-1017	526	36.00	1e-82	1.00	1amuA	19-529
	●	☐	Q8G9W1	hypothetical protein	*Escherichia coli*	1140	349-1135	787	35.00	0	1.00	1r9dA	6-783

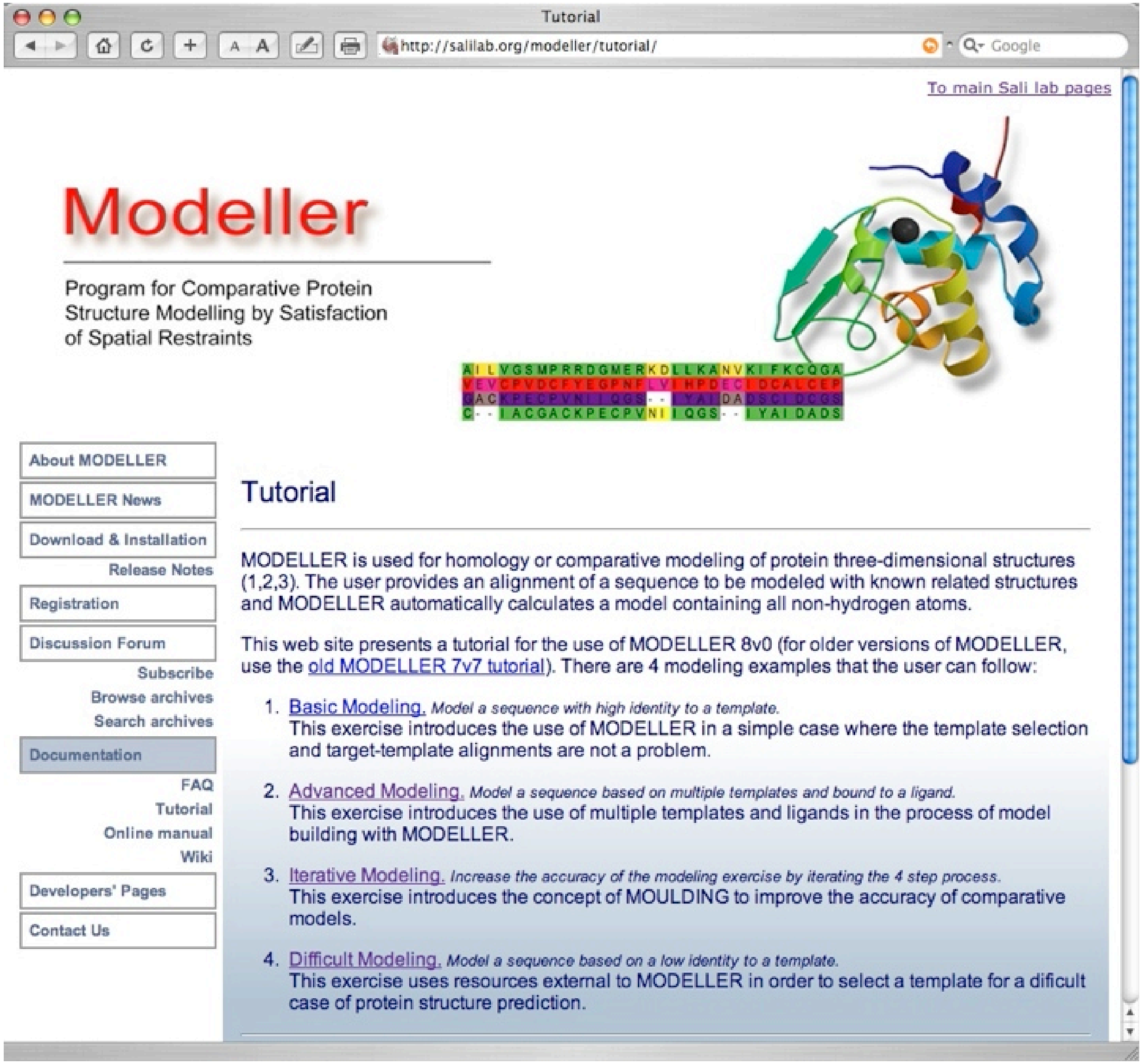
Pieper et al. (2004) Nucleic Acids Research 32, D217-D222



Protein Structure Prediction

SUMMARY

http://www.salilab.org/modeller/tutorial/



Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints

Tutorial

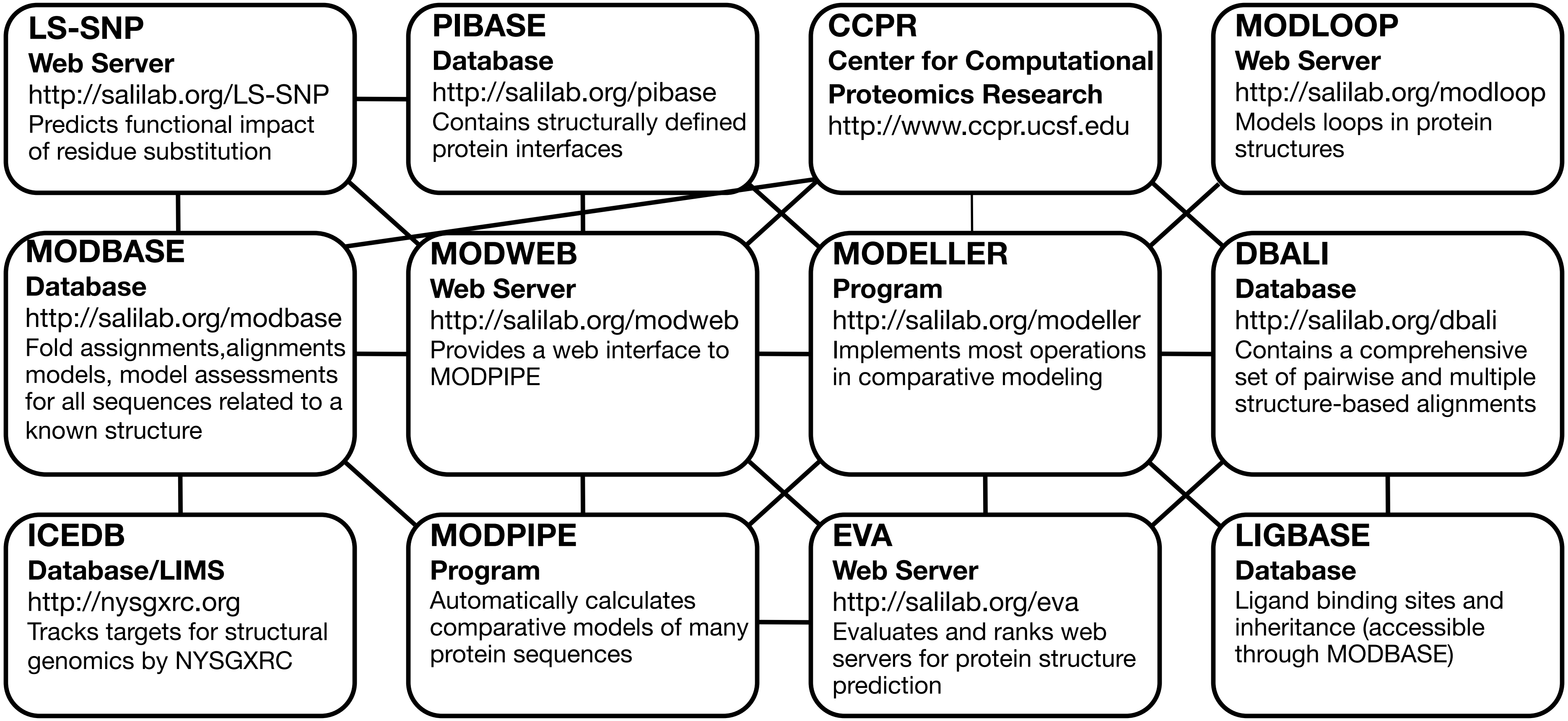
MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1,2,3). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms.

This web site presents a tutorial for the use of MODELLER 8v0 (for older versions of MODELLER, use the old MODELLER 7v7 tutorial). There are 4 modeling examples that the user can follow:

1. Basic Modeling. *Model a sequence with high identity to a template.*
   This exercise introduces the use of MODELLER in a simple case where the template selection and target-template alignments are not a problem.
2. Advanced Modeling. *Model a sequence based on multiple templates and bound to a ligand.*
   This exercise introduces the use of multiple templates and ligands in the process of model building with MODELLER.
3. Iterative Modeling. *Increase the accuracy of the modeling exercise by iterating the 4 step process.*
   This exercise introduces the concept of MOULDING to improve the accuracy of comparative models.
4. Difficult Modeling. *Model a sequence based on a low identity to a template.*
   This exercise uses resources external to MODELLER in order to select a template for a dificult case of protein structure prediction.

A suite of programs, servers and databases for comparative protein structure modeling

http://salilab.org



LS-SNP
Web Server
http://salilab.org/LS-SNP
Predicts functional impact of residue substitution

PIBASE
Database
http://salilab.org/pibase
Contains structurally defined protein interfaces

CCPR
Center for Computational Proteomics Research
http://www.ccpr.ucsf.edu

MODLOOP
Web Server
http://salilab.org/modloop
Models loops in protein structures

MODBASE
Database
http://salilab.org/modbase
Fold assignments,alignments models, model assessments for all sequences related to a known structure

MODWEB
Web Server
http://salilab.org/modweb
Provides a web interface to MODPIPE

MODELLER
Program
http://salilab.org/modeller
Implements most operations in comparative modeling

DBALI
Database
http://salilab.org/dbali
Contains a comprehensive set of pairwise and multiple structure-based alignments

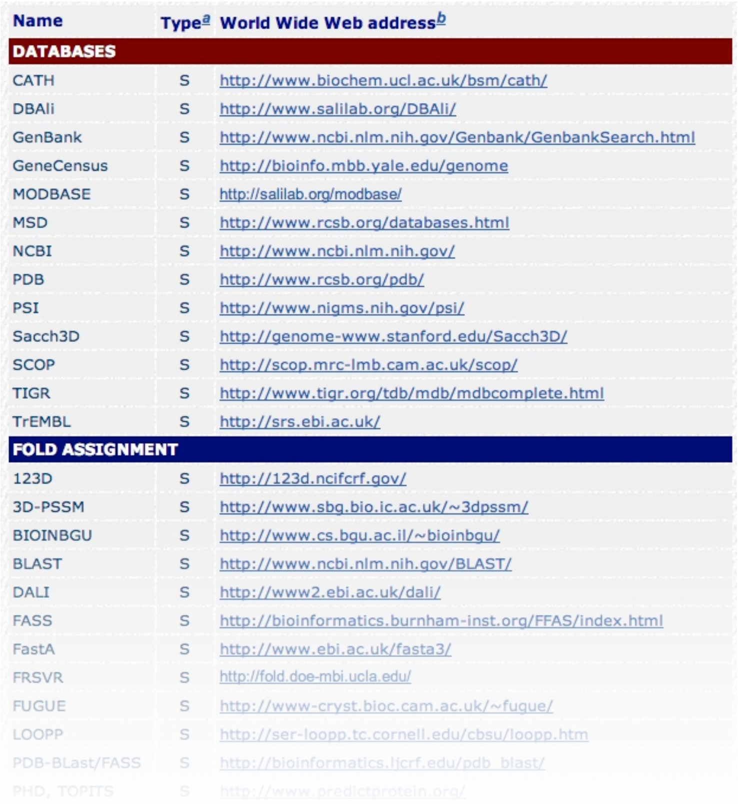
ICEDB
Database/LIMS
http://nysgxrc.org
Tracks targets for structural genomics by NYSGXRC

MODPIPE
Program
Automatically calculates comparative models of many protein sequences

EVA
Web Server
http://salilab.org/eva
Evaluates and ranks web servers for protein structure prediction

LIGBASE
Database
Ligand binding sites and inheritance (accessible through MODBASE)

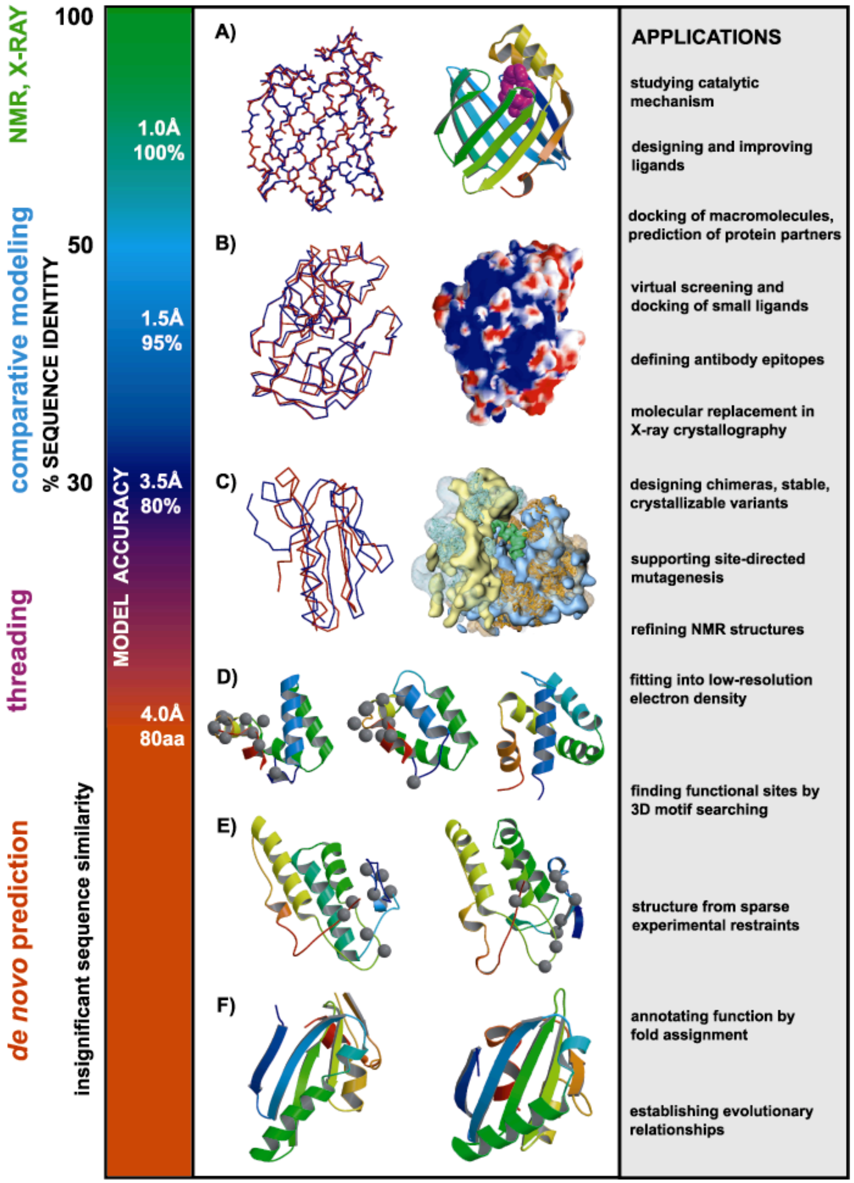
External Resources
PDB, Uniprot, GENBANK, NR, PIR, INTERPRO, Kinase Resource
UCSC Genome Browser, CHIMERA, Pfam, SCOP, CATH



Name	Type[a]	World Wide Web address[b]
DATABASES		
CATH	S	http://www.biochem.ucl.ac.uk/bsm/cath/
DBAli	S	http://www.salilab.org/DBAli/
GenBank	S	http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html
GeneCensus	S	http://bioinfo.mbb.yale.edu/genome
MODBASE	S	http://salilab.org/modbase/
MSD	S	http://www.rcsb.org/databases.html
NCBI	S	http://www.ncbi.nlm.nih.gov/
PDB	S	http://www.rcsb.org/pdb/
PSI	S	http://www.nigms.nih.gov/psi/
Sacch3D	S	http://genome-www.stanford.edu/Sacch3D/
SCOP	S	http://scop.mrc-lmb.cam.ac.uk/scop/
TIGR	S	http://www.tigr.org/tdb/mdb/mdbcomplete.html
TrEMBL	S	http://srs.ebi.ac.uk/
FOLD ASSIGNMENT		
123D	S	http://123d.ncifcrf.gov/
3D-PSSM	S	http://www.sbg.bio.ic.ac.uk/~3dpssm/
BIOINBGU	S	http://www.cs.bgu.ac.il/~bioinbgu/
BLAST	S	http://www.ncbi.nlm.nih.gov/BLAST/
DALI	S	http://www2.ebi.ac.uk/dali/
FASS	S	http://bioinformatics.burnham-inst.org/FFAS/index.html
FastA	S	http://www.ebi.ac.uk/fasta3/
FRSVR	S	http://fold.doe-mbi.ucla.edu/
FUGUE	S	http://www-cryst.bioc.cam.ac.uk/~fugue/
LOOPP	S	http://ser-loopp.tc.cornell.edu/cbsu/loopp.htm
PDB-BLast/FASS	S	http://bioinformatics.ljcrf.edu/pdb_blast/
PHD, TOPITS	S	http://www.predictprotein.org/

http://salilab.org/bioinformatics_resources.shtml

TAKE HOME SLIDE!!



APPLICATIONS

- studying catalytic mechanism
- designing and improving ligands
- docking of macromolecules, prediction of protein partners
- virtual screening and docking of small ligands
- defining antibody epitopes
- molecular replacement in X-ray crystallography
- designing chimeras, stable, crystallizable variants
- supporting site-directed mutagenesis
- refining NMR structures
- fitting into low-resolution electron density
- finding functional sites by 3D motif searching
- structure from sparse experimental restraints
- annotating function by fold assignment
- establishing evolutionary relationships

Happy Modeling!

Master Bioinformatics for Health Sciences

MODELLER tutorial

```
$>mod8v1 model.py
```

Marc A. Marti-Renom

Adjunct Assistant Professor

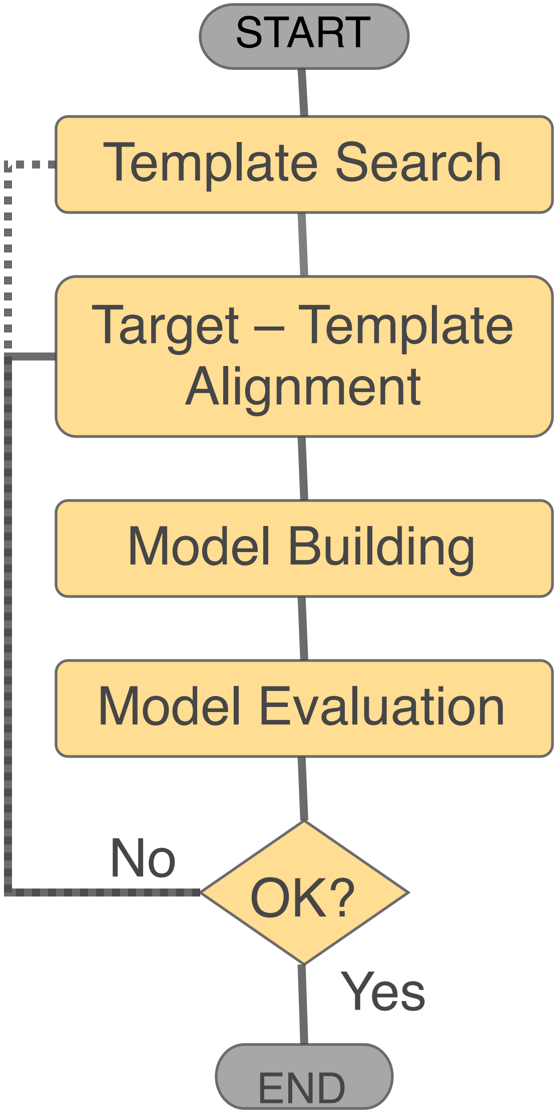
http://salilab.org/~marcius

University of California
San Francisco

Depts. of Biopharmaceutical Sciences and Pharmaceutical Chemistry
California Institute for Quantitative Biomedical Research
University of California at San Francisco

Steps in Comparative Protein Structure Modeling

Reminder!



START

Template Search

Target – Template Alignment

Model Building

Model Evaluation

No

OK?

Yes

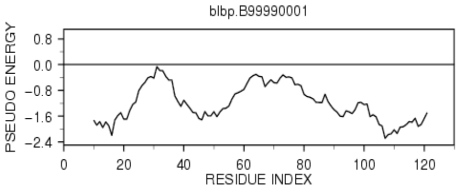
END

TARGET

TEMPLATE

ASILPKRLFGNCEQTSDEGL
KIERTPLVPHISAQNVCLKID
DVPERLIPERASFQWMNDK

```
ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE
```



A. Šali, Curr. Opin. Biotech. 6, 437, 1995.
R. Sánchez & A. Šali, Curr. Opin. Str. Biol. 7, 206, 1997.
M. Marti et al. Ann. Rev. Biophys. Biomolec. Struct., 29, 291, 2000.

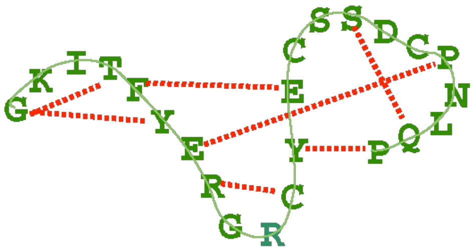
Comparative modeling by satisfaction of spatial restraints

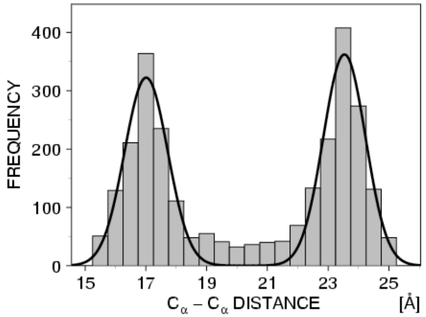
MODELLER

Reminder!

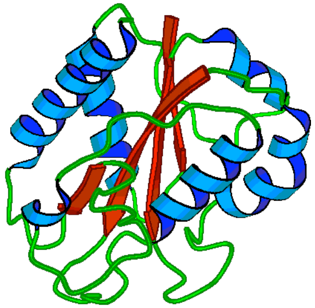
```
3D  GKITFYERGFQGHCYESDC-NLQP…
SEQ GKITFYERG---RCYESDCPNLQP…
```

1. Extract spatial restraints





2. Satisfy spatial restraints



$$F(R) = \prod_i p_i\,(f_i\,/I)$$

A. Šali & T. Blundell. J. Mol. Biol. 234, 779, 1993.
J.P. Overington & A. Šali. Prot. Sci. 3, 1582, 1994.
A. Fiser, R. Do & A. Šali, Prot. Sci., 9, 1753, 2000.

Typical errors in comparative models

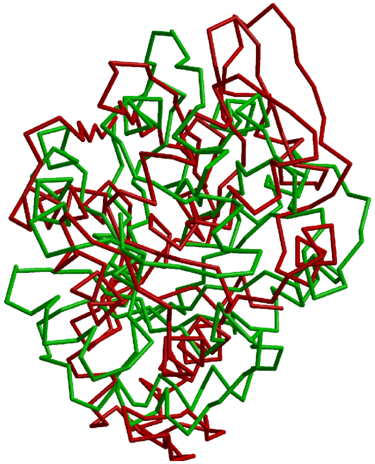
Reminder!

MODEL

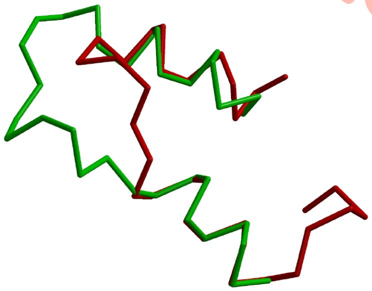
X-RAY

TEMPLATE

Incorrect template

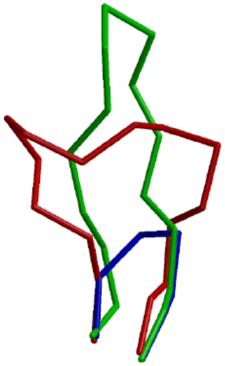


Misalignment

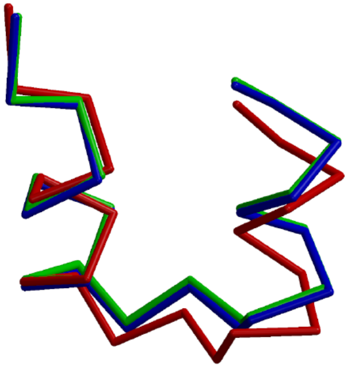


```
                 10        20        30        40        50        60
EDN   ---KPPQFTWAQWFETQHINMTSQQCTNAMQVINNYQRRCKNQNTFLLTTFANVVNVCGNPNMTCPSN
       /////////////////  /   ||||||| \\\\ |||||||||||||||||||||||||||| ||
7RSA  KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHESLADVQAVCSQKNVAC-KN
              aaaaaaaaaaa     aaaaaaaaaaaaaa   bbbbbbb aaaaaaaaa

         70        80        90       100       110       120       130
EDN   KTRKNCHHSGSQVPLIHCNLTTPSPQNISNCRYAQTPANMFYIVACDNRDQRRDPPQYPVVPVHLDRII
      \ ||||||||||||||||||| \\ \\ \ ||||||||||||||||||||         |||||||||||
7RSA  -GQTNCYQSYSTMSITDCRETGSS--KYPNCAYKTTQANKHIIVACEGN---------PYVPVHFDASV
          bbbbb   bbbbbbbbbb    aaaaabbbbbbbbbbbbbbbbbb         bbbbbbbbbbb
```

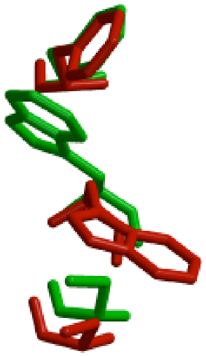
Region without a template



Distortion/shifts in aligned regions



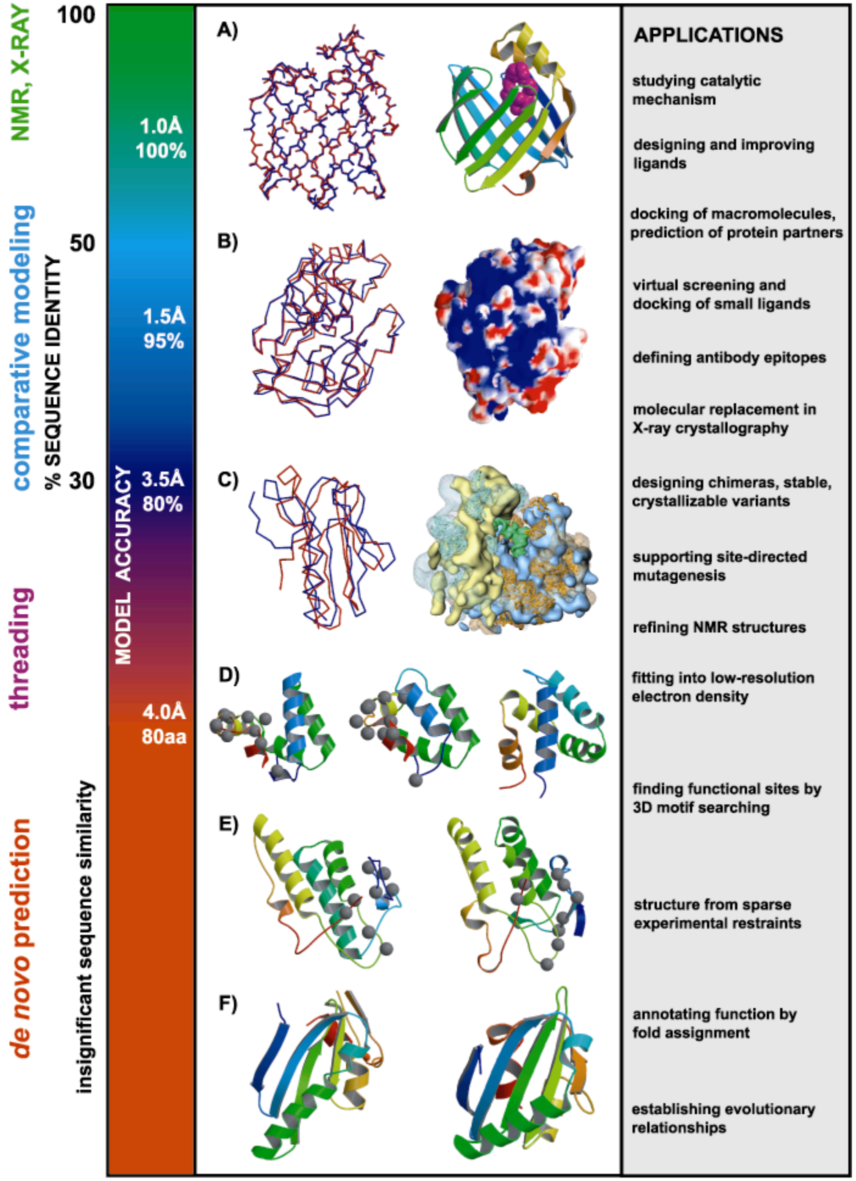
Sidechain packing



Marti-Renom et al. Annu.Rev.Biophys.Biomol.Struct. 29, 291-325, 2000.

Utility of protein structure models, despite errors

Reminder!


NMR, X-RAY
comparative modeling
threading
de novo prediction
100
50
30
% SEQUENCE IDENTITY
insignificant sequence similarity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
4.0Å
80aa
A)
B)
C)
D)
E)
F)
APPLICATIONS
studying catalytic mechanism
designing and improving ligands
docking of macromolecules, prediction of protein partners
virtual screening and docking of small ligands
defining antibody epitopes
molecular replacement in X-ray crystallography
designing chimeras, stable, crystallizable variants
supporting site-directed mutagenesis
refining NMR structures
fitting into low-resolution electron density
finding functional sites by 3D motif searching
structure from sparse experimental restraints
annotating function by fold assignment
establishing evolutionary relationships

D. Baker & A. Sali. Science 294, 93, 2001.

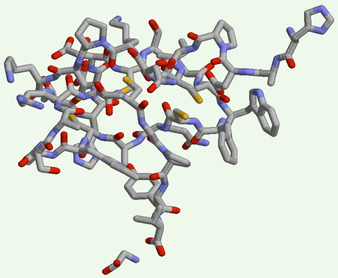
Modeller
Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints
AILVGSMPRRDGMERKDLLKANVKIFKCQGA
VEVCPVDCFYEGPNFLVIHPDECIDCALCEP
GACKPECPVNIIQGS--IYAIDADSCIDCGS
C--IACGACKPECPVNIIQGS--IYAIDADS

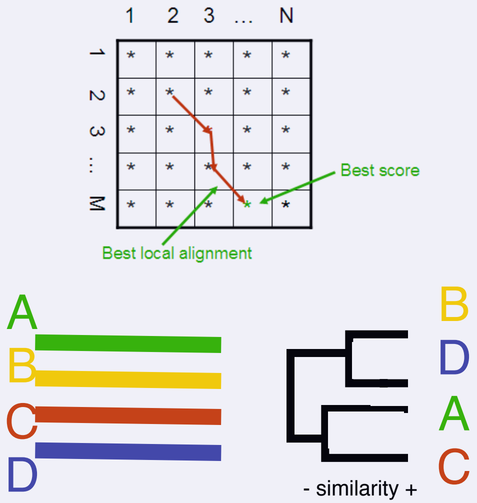
New features in MODELLER 8

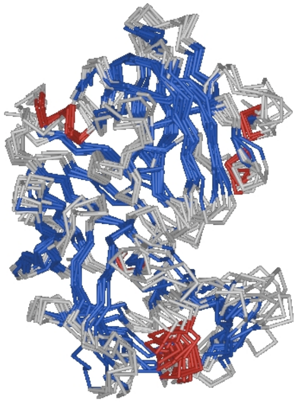
Overview

- SALIGN: versatile alignment module
- MODPIPE2.0: Large-scale protein structure modeling
- profile.build(): Iterative database searching & profile construction
- profile.scan(): fold-assignment & profile-profile alignments
- New objective functions
- Mod-EM: Fitting a model into an EM map
- Additional features and bugs fixed

SALIGN: Versatile alignment command in MODELLER

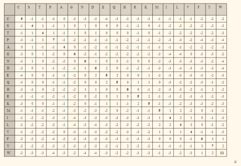




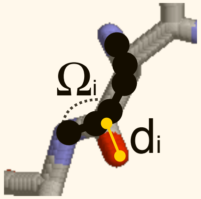


- ✓ Uses all available structural information
- ✓ Provides the optimal alignment

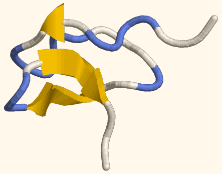
- Computationally expensive



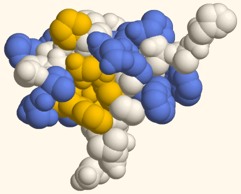
$R_{i,j}$



$D_{i(3),j(3)}$



$S_{i,j}$



$B_{i,j}$

$$\text{RMSD} = \sqrt{\sum(x_i - \bar{x})^2}$$

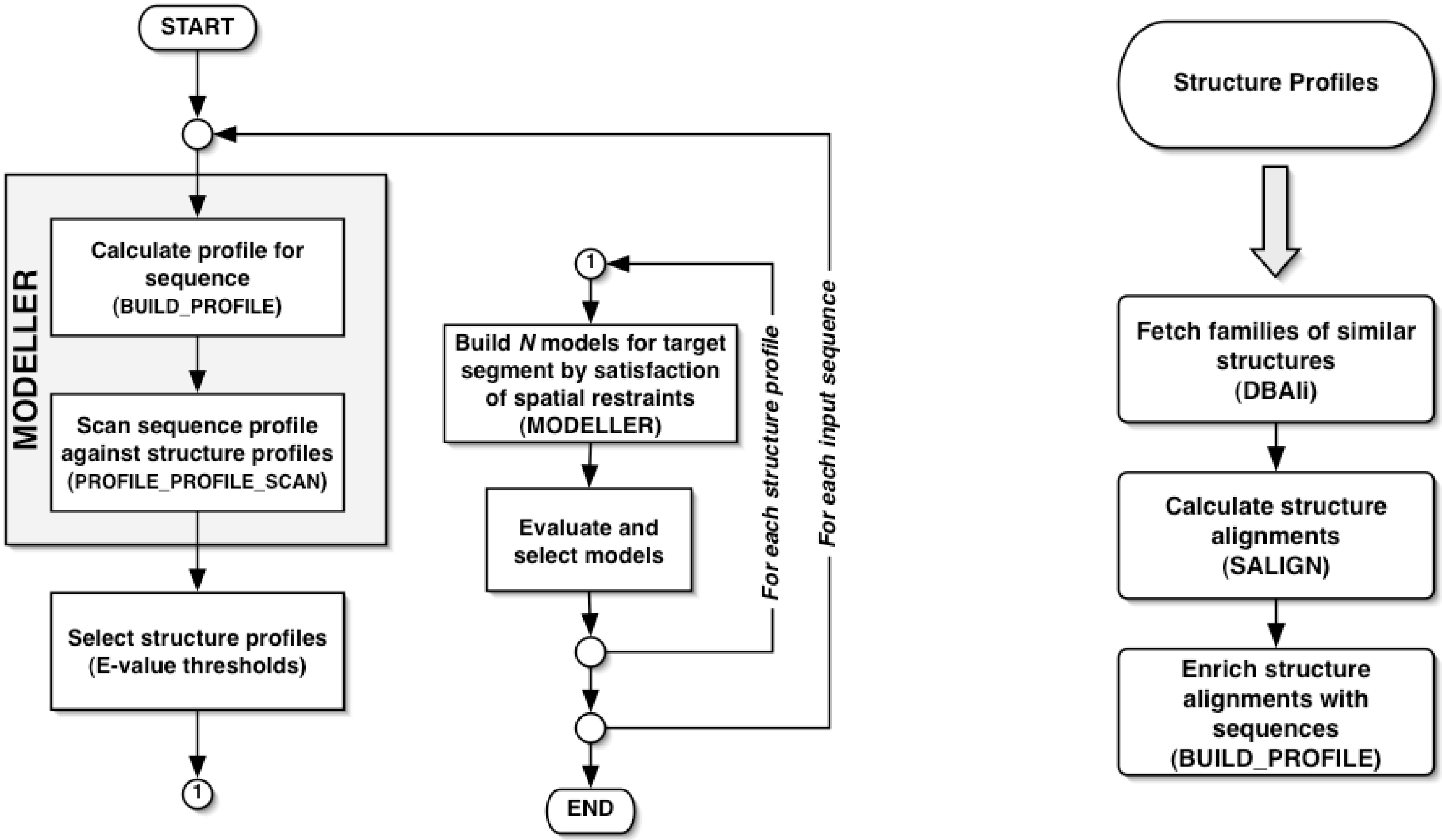
$I_{i,j}$

$$Score_{i,j} = w_1 * R_{i,j} + w_2 * D_{i(a),j(a)} + w_3 * S_{i,j} + w_4 * B_{i,j} + w_5 * I_{i,j} + w_6 * X_{i,j}$$

Madhusudhan et al. in preparation

MODPIPE2.0

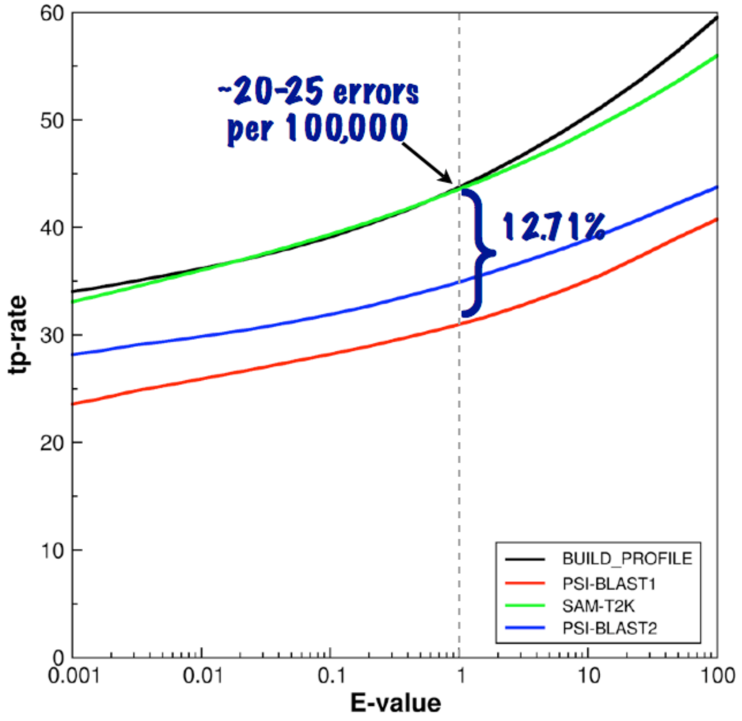
Large-Scale Protein Structure Modeling

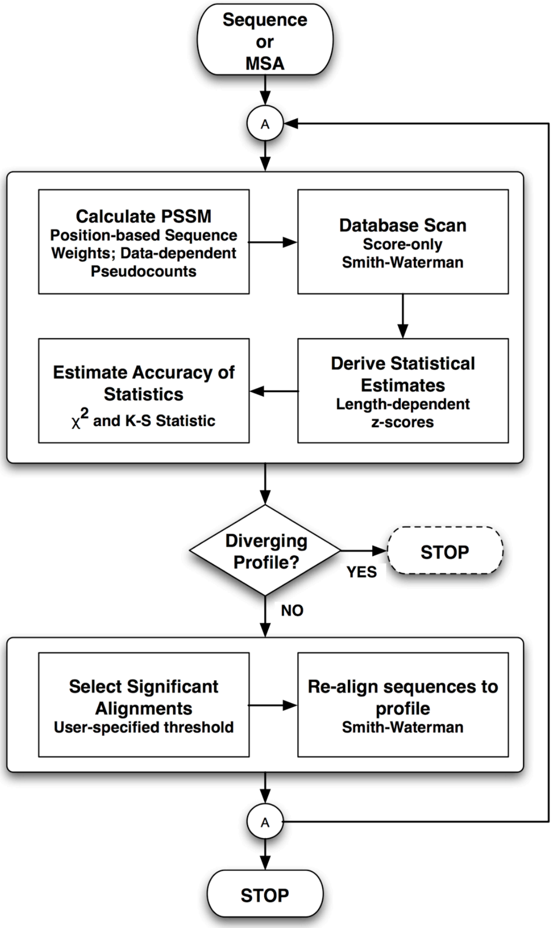

START
MODELLER
Calculate profile for sequence (BUILD_PROFILE)
Scan sequence profile against structure profiles (PROFILE_PROFILE_SCAN)
Select structure profiles (E-value thresholds)
1
1
Build *N* models for target segment by satisfaction of spatial restraints (MODELLER)
Evaluate and select models
For each structure profile
For each input sequence
END
Structure Profiles
Fetch families of similar structures (DBAli)
Calculate structure alignments (SALIGN)
Enrich structure alignments with sequences (BUILD_PROFILE)

Eswar et.al., (2003) Nucl.Acids.Res. 31(13)

profile.build(): Iterative database searching & profile construction

- Rigorous Smith-Waterman local alignments
- Statistical significance using z-scores
- Null model for statistics from actual sequences
- Built-in mechanism to detect profile divergence



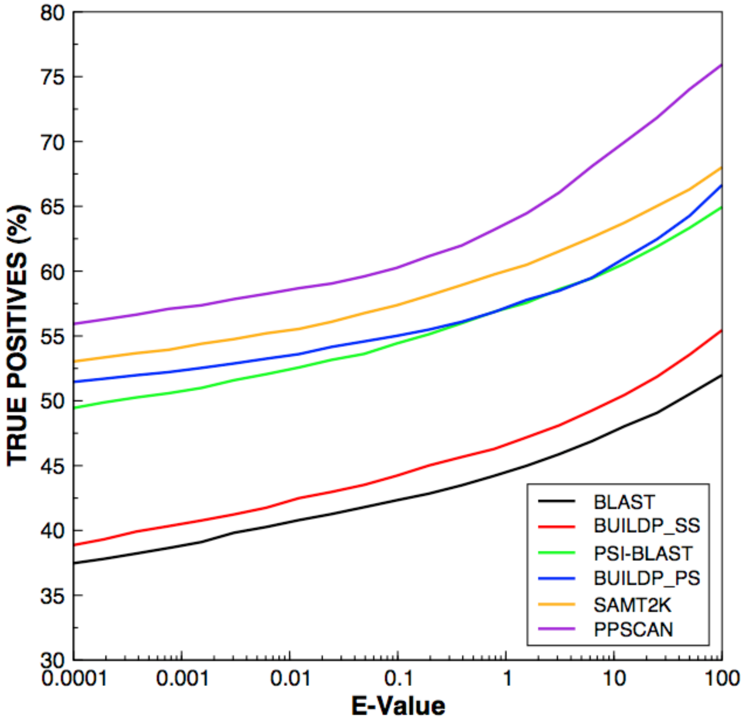


Sequence or MSA
A
Calculate PSSM
Position-based Sequence Weights; Data-dependent Pseudocounts
Database Scan
Score-only Smith-Waterman
Derive Statistical Estimates
Length-dependent z-scores
Estimate Accuracy of Statistics
χ^2 and K-S Statistic
Diverging Profile?
YES
STOP
NO
Select Significant Alignments
User-specified threshold
Re-align sequences to profile
Smith-Waterman
A
STOP

Eswar et.al., in preparation

profile.scan(): Large-scale fold-assignment & profile-profile alignments

- Correlation Coefficient as the scoring scheme
- Smith-Waterman local alignments
- Statistical significance using z-scores
- Null model for statistics from actual sequences



Method	CE overlap	Shift score
CE	100 ± 0	1.00 ± 0.00
BLAST	26 ± 29	0.32 ± 0.33
PSI-BLAST	43 ± 31	0.48 ± 0.35
SAM	48 ± 26	0.50 ± 0.34
LOBSTER	50 ± 27	0.51 ± 0.32
SEA	49 ± 27	0.53 ± 0.29
ALIGN	42 ± 25	0.44 ± 0.28
CLUSTALW	43 ± 27	0.44 ± 0.31
COMPASS	43 ± 32	0.49 ± 0.35
CCHH	**56 ± 23**	**0.61 ± 0.24**
CCHS	**56 ± 24**	**0.62 ± 0.24**
TOP	62 ± 20	0.67 ± 0.20

Eswar et.al., in preparation
Marti-Renom et al. (2004) Protein Science 13 pp1071

New objective functions in Modeller

- Discrete Optimized Protein Energy (DOPE)
    - A highly accurate distance dependent statistical potential
    - The best performer among 29 tested scoring functions
    - Has been released with Modeller 8v0
    - Applications to loop modeling and model assessment
- Solvation model in Modeller
    - The GB/SA solvation model is implemented in Modeller
    - A newer scheme for point charges that concurs with CHARMM22
    - Will be included in Modeller 8v1
- Other modifications
    - Minor modifications to the atomic van der Waals radii (radii.lib)

Mod-EM: Fitting a model into an EM map

Allows protein models to be docked into lower-resolution electron microscopy maps for scoring

Representation:

map: density in voxels, $\rho^{EM}(r)$
protein: based on atomic mass, $\rho^{probe}(r)$ - can be represented by different functions: Gaussian, uniform sphere model, hybrid Gaussian/sphere model
filters: cutoff filters, sqr filter, Laplacian

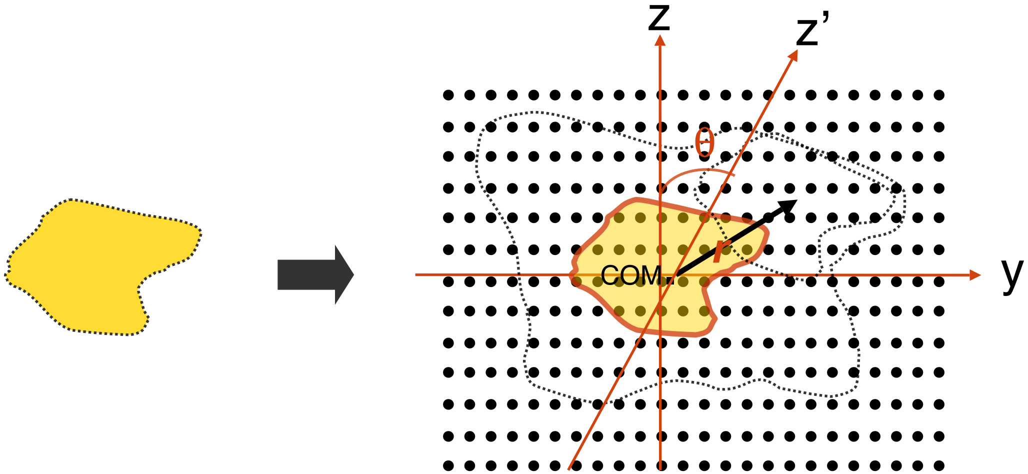
Scoring function:

Cross-correlation function (C) between the map density and the probe ‘density’:

$$C = \frac{\sum_{i=1}^{M} \rho_i^{EM} \left(\sum_{j=1}^{N} \rho_{i,j}^{probe} \right)}{\sqrt{\sum_{i=1}^{M} \left(\rho_i^{EM}\right)^2 \sum_{i=1}^{M} \left(\sum_{j=1}^{N} \rho_{i,j}^{probe} \right)^2}}$$

N - number of atoms in the probe

M - number of cryoEM grid points
covered by the probe density



Optimization methods (density.grid_search()):

Exhaustive rigid rotation and translation
Monte Carlo optimization

Topf et al., J. Struct. Biol. ***149****, 191-203 (2005)*

Additional features

- **New methods for model assessment:**
  - DOPE
  - GA341
  - ModEM scoring
- **Improved user interface:**
  - Python interface allows for more powerful and flexible scripting, and easier integration with other applications
  - Legacy TOP scripts are still parsed

Obtaining MODELLER and related information

- MODELLER (8v0) web page
- `http://www.salilab.org/modeller/`
    - Download Software (Linux/Windows/Mac/Solaris)
    - HTML Manual
    - **Join Mailing List**

Using MODELLER

- No GUI! ☹
- Controlled by command file ☹☹
- Script is written in PYTHON language ☺
- You may know Python language is simple ☺☺

Using MODELLER

- INPUT:
  - Target Sequence (FASTA/PIR format)
  - Template Structure (PDB format)
  - TOP command file
- OUTPUT:
  - Target-Template Alignment
  - Model in PDB format
  - Other data

Modeling of BLBP
Input

- Target: Brain lipid-binding protein (BLBP)
- BLBP sequence in PIR (MODELLER) format:

```
>P1;blbp
sequence:blbp::::::::
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEETSID
DRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Python script for target-template alignment

```
# Example for: alignment.align()

# This will read two sequences, align them, and write the alignment
# to a file:

log.verbose()
env = environ()

aln = alignment(env)
mdl = model(env, file='1hms')
aln.append_model(mdl, align_codes='1hms')
aln.append(file='blbp.seq', align_codes=('blbp'))

# The as1.sim.mat similarity matrix is used by default:
aln.align(gap_penalties_1d=(-600, -400))
aln.write(file='blbp-1hms.ali', alignment_format='PIR')
aln.write(file='blbp-1hms.pap', alignment_format='PAP')
```

Run by typing `mod8v0 align.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Python script for target-template alignment

```
# Example for: alignment.align()

# This will read two sequences, align them, and write the alignment
# to a file:

log.verbose()
env = environ()

aln = alignment(env)
mdl = model(env, file='1hms')
aln.append_model(mdl, align_codes='1hms')
aln.append(file='blbp.seq', align_codes=('blbp'))

# The as1.sim.mat similarity matrix is used by default:
aln.align(gap_penalties_1d=(-600, -400))
aln.write(file='blbp-1hms.ali', alignment_format='PIR')
aln.write(file='blbp-1hms.pap', alignment_format='PAP')
```

Run by typing `mod8v0 align.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Python script for target-template alignment

```
# Example for: alignment.align()

# This will read two sequences, align them, and write the alignment
# to a file:

log.verbose()
env = environ()

aln = alignment(env)
mdl = model(env, file='1hms')
aln.append_model(mdl, align_codes='1hms')
aln.append(file='blbp.seq', align_codes=('blbp'))

# The as1.sim.mat similarity matrix is used by default:
aln.align(gap_penalties_1d=(-600, -400))
aln.write(file='blbp-1hms.ali', alignment_format='PIR')
aln.write(file='blbp-1hms.pap', alignment_format='PAP')
```

Run by typing `mod8v0 align.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Python script for target-template alignment

```
# Example for: alignment.align()

# This will read two sequences, align them, and write the alignment
# to a file:

log.verbose()
env = environ()

aln = alignment(env)
mdl = model(env, file='1hms')
aln.append_model(mdl, align_codes='1hms')
aln.append(file='blbp.seq', align_codes=('blbp'))

# The as1.sim.mat similarity matrix is used by default:
aln.align(gap_penalties_1d=(-600, -400))
aln.write(file='blbp-1hms.ali', alignment_format='PIR')
aln.write(file='blbp-1hms.pap', alignment_format='PAP')
```

Run by typing `mod8v0 align.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Output

```
>P1;1hms
structureX:1hms:   1 : : 131 : :undefined:undefined:-1.00:-1.00
VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGVEFDETTA
DDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTYEKE*
>P1;blbp
sequence:blbp:     : :     : : : : 0.00: 0.00
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEETSI
DDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

Modeling of BLBP
STEP 1: Align blbp and 1hms sequences
Output

```
>P1;1hms
structureX:1hms:   1 : : 131 : :undefined:undefined:-1.00:-1.00
VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGVEFDETTA
DDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTYEKE*
>P1;blbp
sequence:blbp:     : :     : : : : 0.00: 0.00
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEETSI
DDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Output

```
 _aln.pos          10        20        30        40        50        60
1hms      VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGV
blbp      VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGE
 _consrvd ****  **** ** *** *** **********   **** **   *      *  ******* * **


 _aln.p   70        80        90       100       110       120       130
1hms      EFDETTADDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTYEKE
blbp      EFEETSIDDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA
 _consrvd ** **  ***  ** * *** ** * ***** **   **  ***   *** *  *  * ***
```

Modeling of BLBP

STEP 2: Model the blbp structure using the alignment from step 1.

Python script for model building

```
# Homology modelling by the automodel class
from modeller.automodel import *    # Load the automodel class
log.verbose()                       # request verbose output
env = environ()                     # create a new MODELLER environment

# directories for input atom files
env.io.atom_files_directory = './:../atom_files'

a = automodel(env,
              alnfile  = 'blbp-1hms.ali',     # alignment filename
              knowns   = '1hms',              # codes of the templates
              sequence = 'blbp')              # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  = 1                 # index of the last model
                                    # (determines how many models to calculate)
a.make()                            # do the actual homology modelling
```

Run by typing `mod8v0 model.py` in the directory where you have the python file. MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 2: Model the blbp structure using the alignment from step 1.

Python script for model building

```
# Homology modelling by the automodel class
from modeller.automodel import *    # Load the automodel class
log.verbose()                       # request verbose output
env = environ()                     # create a new MODELLER environment

# directories for input atom files
env.io.atom_files_directory = './:../atom_files'

a = automodel(env,
              alnfile  = 'blbp-1hms.ali',     # alignment filename
              knowns   = '1hms',              # codes of the templates
              sequence = 'blbp')              # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  = 1                 # index of the last model
                                    # (determines how many models to calculate)
a.make()                            # do the actual homology modelling
```

Run by typing `mod8v0 model.py` in the directory where you have the python file. MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 2: Model the blbp structure using the alignment from step 1.

Python script for model building

```
# Homology modelling by the automodel class
from modeller.automodel import *    # Load the automodel class
log.verbose()                       # request verbose output
env = environ()                     # create a new MODELLER environment

# directories for input atom files
env.io.atom_files_directory = './:../atom_files'

a = automodel(env,
              alnfile  = 'blbp-1hms.ali',     # alignment filename
              knowns   = '1hms',              # codes of the templates
              sequence = 'blbp')              # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  = 1                 # index of the last model
                                    # (determines how many models to calculate)
a.make()                            # do the actual homology modelling
```

Run by typing `mod8v0 model.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 2: Model the blbp structure using the alignment from step 1.

Python script for model building

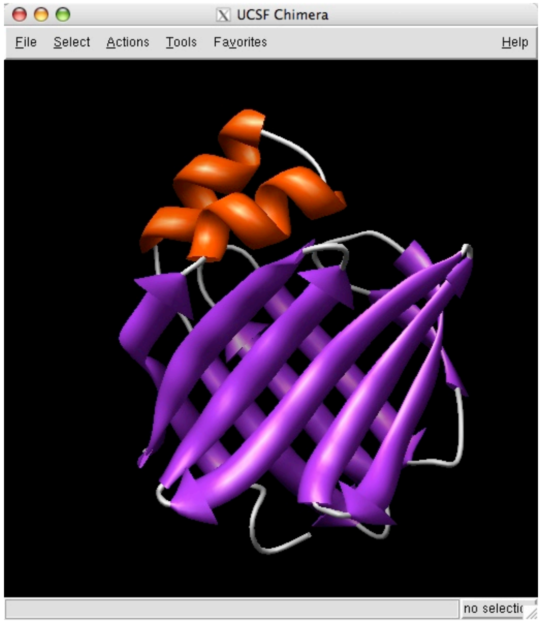
PDB file

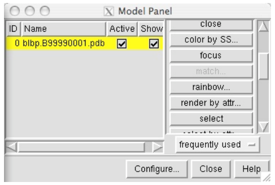
Can be viewed with Chimera

http://www.cgl.ucsf.edu/chimera/

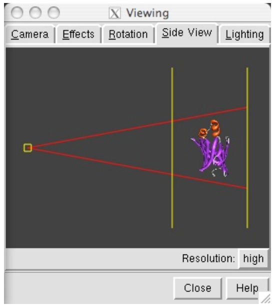
Rasmol

http://www.openrasmol.org





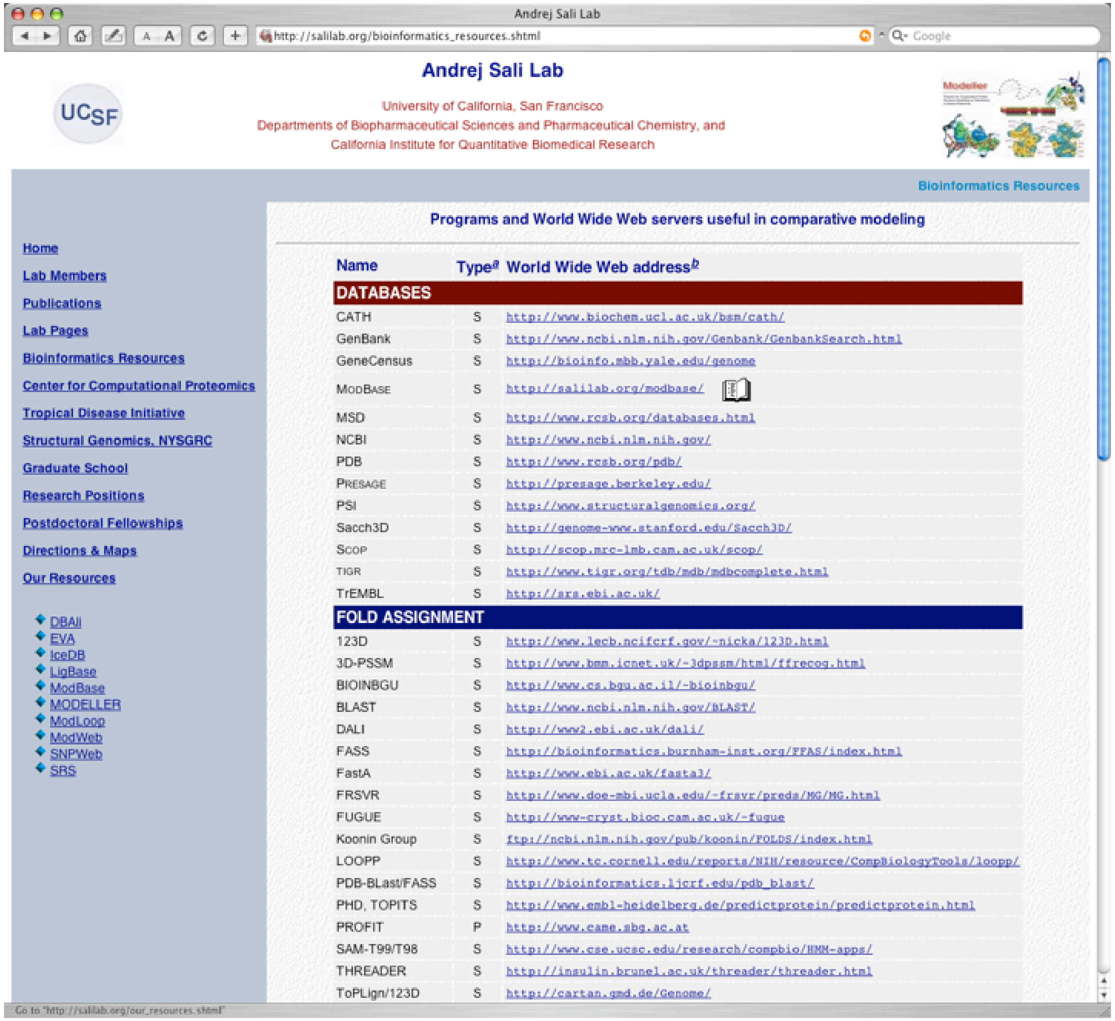
ID	Name	Active	Show
0	blbp.B99990001.pdb	☑	☑



Resolution: high

Model file → `blbp.B99990001`

http://www.salilab.org/bioinformatics_resources.shtml



Andrej Sali Lab

http://salilab.org/bioinformatics_resources.shtml

Andrej Sali Lab

University of California, San Francisco
Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and
California Institute for Quantitative Biomedical Research

Bioinformatics Resources

- Home
- Lab Members
- Publications
- Lab Pages
- Bioinformatics Resources
- Center for Computational Proteomics
- Tropical Disease Initiative
- Structural Genomics, NYSGRC
- Graduate School
- Research Positions
- Postdoctoral Fellowships
- Directions & Maps
- Our Resources

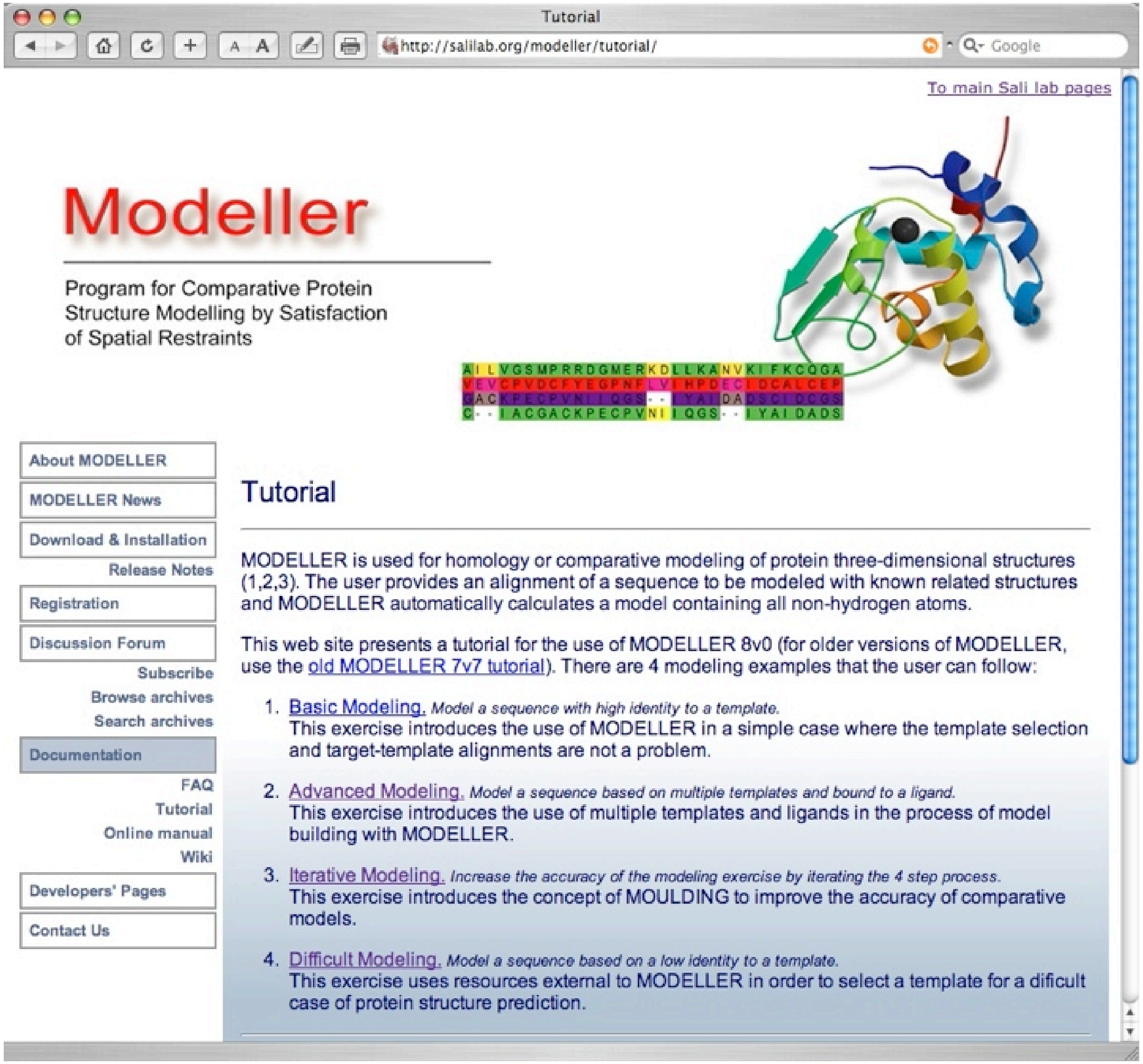
- DBAli
- EVA
- IceDB
- LigBase
- ModBase
- MODELLER
- ModLoop
- ModWeb
- SNPWeb
- SRS

Programs and World Wide Web servers useful in comparative modeling

Name	Type[a]	World Wide Web address[b]
DATABASES		
CATH	S	http://www.biochem.ucl.ac.uk/bsm/cath/
GenBank	S	http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html
GeneCensus	S	http://bioinfo.mbb.yale.edu/genome
MODBASE	S	http://salilab.org/modbase/ 📖
MSD	S	http://www.rcsb.org/databases.html
NCBI	S	http://www.ncbi.nlm.nih.gov/
PDB	S	http://www.rcsb.org/pdb/
PRESAGE	S	http://presage.berkeley.edu/
PSI	S	http://www.structuralgenomics.org/
Sacch3D	S	http://genome-www.stanford.edu/Sacch3D/
SCOP	S	http://scop.mrc-lmb.cam.ac.uk/scop/
TIGR	S	http://www.tigr.org/tdb/mdb/mdbcomplete.html
TrEMBL	S	http://srs.ebi.ac.uk/
FOLD ASSIGNMENT		
123D	S	http://www.lecb.ncifcrf.gov/~nicka/123D.html
3D-PSSM	S	http://www.bmm.icnet.uk/~3dpssm/html/ffrecog.html
BIOINBGU	S	http://www.cs.bgu.ac.il/~bioinbgu/
BLAST	S	http://www.ncbi.nlm.nih.gov/BLAST/
DALI	S	http://www2.ebi.ac.uk/dali/
FASS	S	http://bioinformatics.burnham-inst.org/FFAS/index.html
FastA	S	http://www.ebi.ac.uk/fasta3/
FRSVR	S	http://www.doe-mbi.ucla.edu/~frsvr/preds/MG/MG.html
FUGUE	S	http://www-cryst.bioc.cam.ac.uk/~fugue
Koonin Group	S	ftp://ncbi.nlm.nih.gov/pub/koonin/FOLDS/index.html
LOOPP	S	http://www.tc.cornell.edu/reports/NIH/resource/CompBiologyTools/loopp/
PDB-BLast/FASS	S	http://bioinformatics.ljcrf.edu/pdb_blast/
PHD, TOPITS	S	http://www.embl-heidelberg.de/predictprotein/predictprotein.html
PROFIT	P	http://www.came.sbg.ac.at
SAM-T99/T98	S	http://www.cse.ucsc.edu/research/compbio/HMM-apps/
THREADER	S	http://insulin.brunel.ac.uk/threader/threader.html
ToPLign/123D	S	http://cartan.gmd.de/Genome/

Go to "http://salilab.org/our_resources.shtml"

http://www.salilab.org/modeller/tutorial/



Tutorial
http://salilab.org/modeller/tutorial/
Google
AILVGSMPRRDGMERKDLLKANVKIFKCQGA
VEVCPVDCFYEGPNFLVIHPDECIDCALCEP
GACKPECPVNIIQGS..IYAIDADSCIDCGS
C..IACGACKPECPVNIIQGS..IYAIDADS

To main Sali lab pages

Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints

- About MODELLER
- MODELLER News
- Download & Installation
  - Release Notes
- Registration
- Discussion Forum
  - Subscribe
  - Browse archives
  - Search archives
- Documentation
  - FAQ
  - Tutorial
  - Online manual
  - Wiki
- Developers' Pages
- Contact Us

Tutorial

MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1,2,3). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms.

This web site presents a tutorial for the use of MODELLER 8v0 (for older versions of MODELLER, use the old MODELLER 7v7 tutorial). There are 4 modeling examples that the user can follow:

1. Basic Modeling. *Model a sequence with high identity to a template.*
   This exercise introduces the use of MODELLER in a simple case where the template selection and target-template alignments are not a problem.
2. Advanced Modeling. *Model a sequence based on multiple templates and bound to a ligand.*
   This exercise introduces the use of multiple templates and ligands in the process of model building with MODELLER.
3. Iterative Modeling. *Increase the accuracy of the modeling exercise by iterating the 4 step process.*
   This exercise introduces the concept of MOULDING to improve the accuracy of comparative models.
4. Difficult Modeling. *Model a sequence based on a low identity to a template.*
   This exercise uses resources external to MODELLER in order to select a template for a dificult case of protein structure prediction.