

BMC WorkShop

Protein Structure Prediction

Introduction

Marc A. Marti-Renom & Damien Devos

Department of Biopharmaceutical Sciences, UCSF

Objective

**TO LEARN HOW-TO MODEL A
3D-STRUCTURE FROM SEQUENCE**

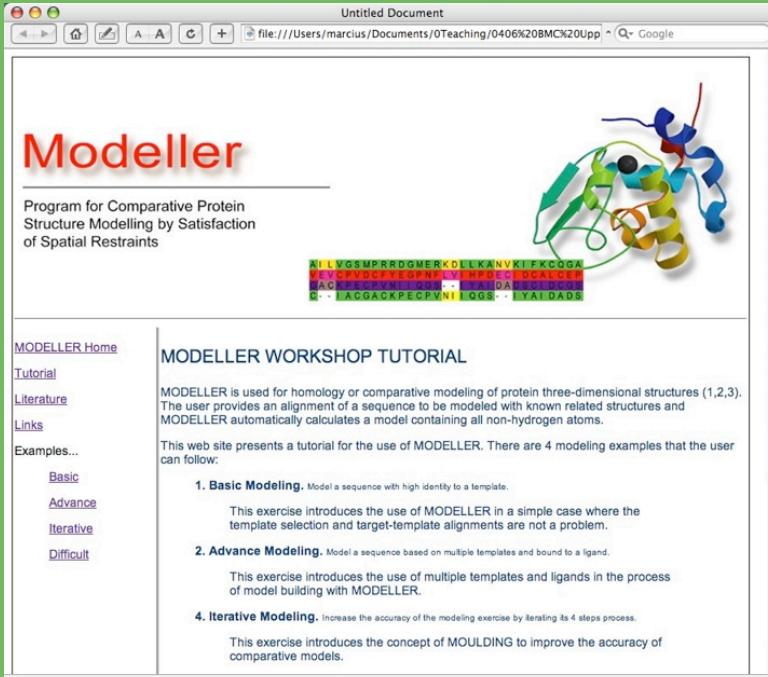
Program

Template Search

Target – Template Alignment

Model Building

Model Evaluation



The screenshot shows the MODELLER software interface. At the top, it says "Untitled Document" and "file:///Users/marcius/Documents/0Teaching/0406%20BMC%20Up...". The main title is "Modeller" in red. Below it is a subtitle: "Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints". To the right is a 3D ribbon model of a protein structure. Below the title is a sequence alignment grid:

N	I	L	V	G	S	M	P	R	R	D	O	M	E	R	K	D	L	K	A	N	V	K	I	F	K	C	G	A
E	V	F	P	D	C	T	E	G	P	N	V	H	P	D	E	C	D	A	C	E	T							
A	C	T	T	C	T	T	A	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
Q	A	G	A	C	K	P	E	C	P	V	N	I	T	Q	G	S	T	L	A	I	D	A	D	A	D	A	D	A

On the left, there's a sidebar with links: MODELLER Home, Tutorial, Literature, Links, Examples..., Basic, Advance, Iterative, and Difficult. On the right, under "MODELLER WORKSHOP TUTORIAL", it describes MODELLER's purpose and provides a 4-step tutorial:

- 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. Advance 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 its 4 steps process.
This exercise introduces the concept of MOULDING to improve the accuracy of comparative models.

<http://www.salilab.org/modeller/workshop/>

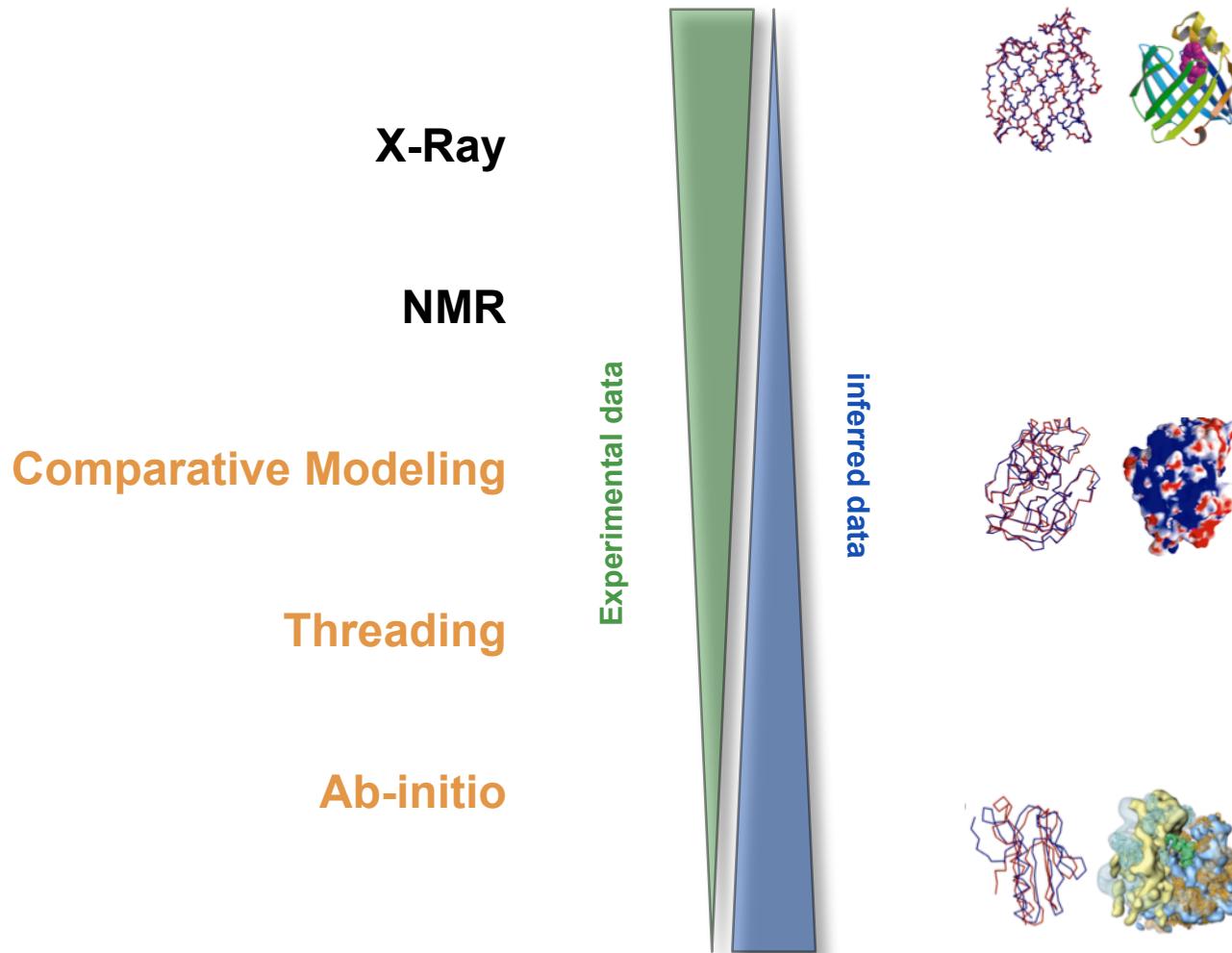
What are we going to do?

- 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.

protein prediction .vs. protein determination



Why protein structure prediction?

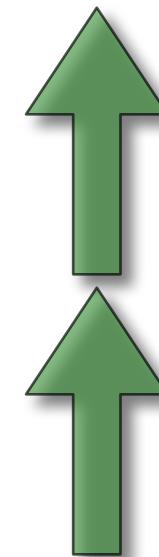
	Y 2004	Y 2006
Sequences	1,500,000	millions
Structures	28,000	50,000

Why protein structure prediction?

	Y 2004
Sequences	1,500,000
Structures	400,000

<http://salilab.org/modbase/>

Theory

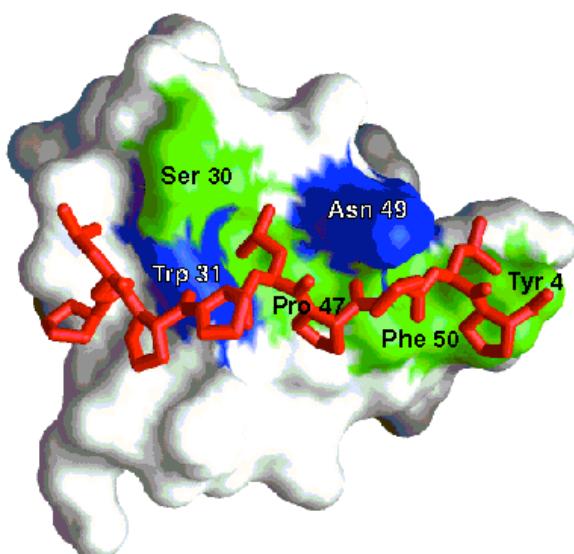


Experiment

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.

YDL117W
(15-64) KAR_YGWSGQTKGDLGFLEGDIMEVTRIAGS_WFYGKLLRNKKCSGYFPHN_F

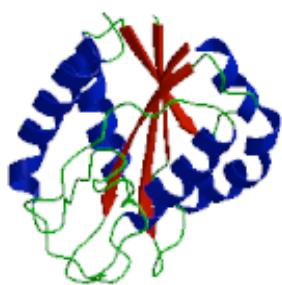


In addition, since evolution tends to conserve function and function depends more directly on structure than on sequence, **structure is more conserved in evolution than sequence**.

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

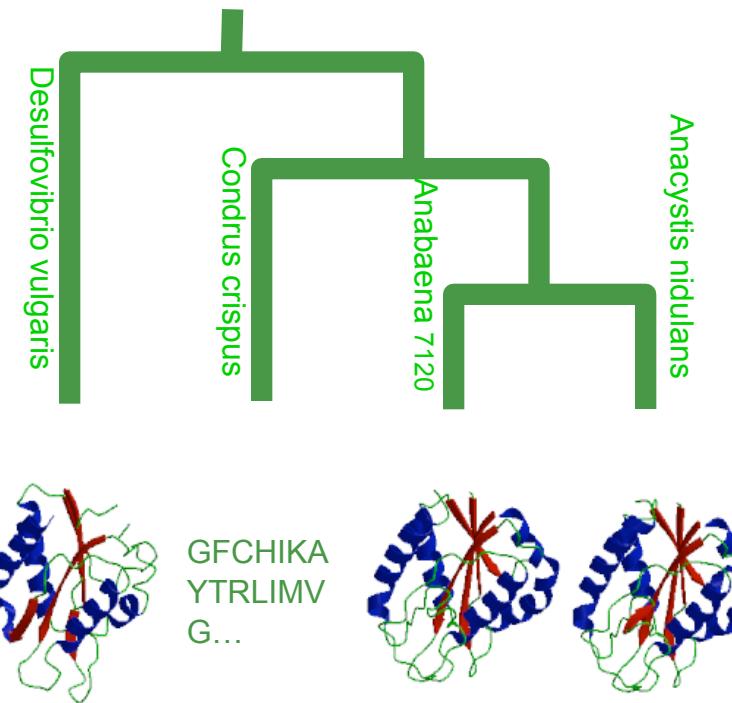
Principles of Protein Structure

GFCHIKAYTRLIMVG...



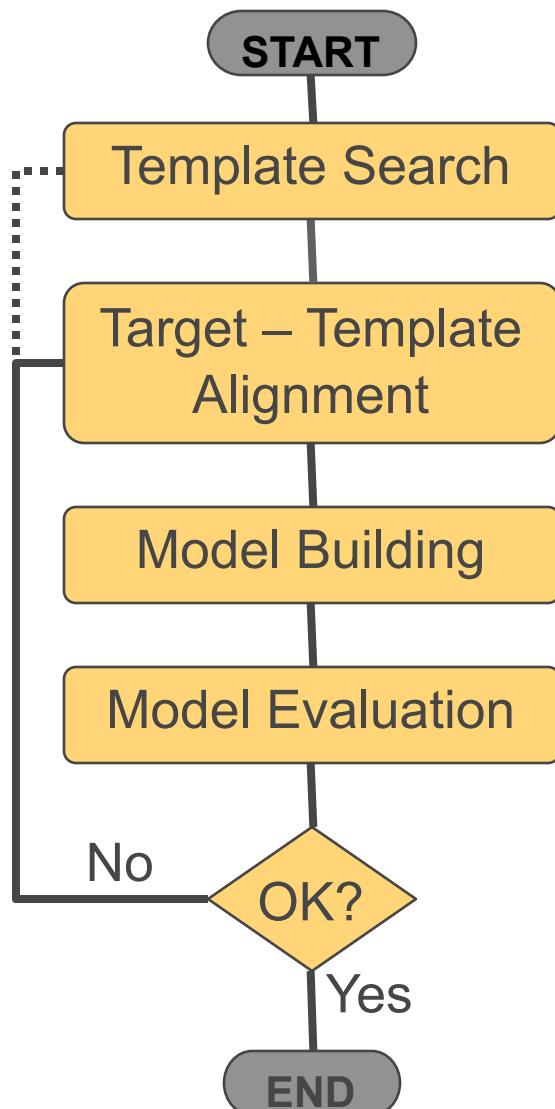
Folding

Ab initio prediction



Evolution
Threading
Comparative Modeling

Steps in Comparative Protein Structure Modeling



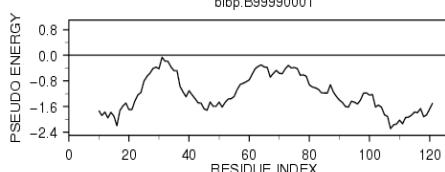
TARGET

ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVLKIDDV
PERLIPERASFQWMNDK

TEMPLATE



ASILPKRLFGNCEQTSDEGLK IERTPLVPHISAQNVLKIDDV PERLIPERASFQWMNDK
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE



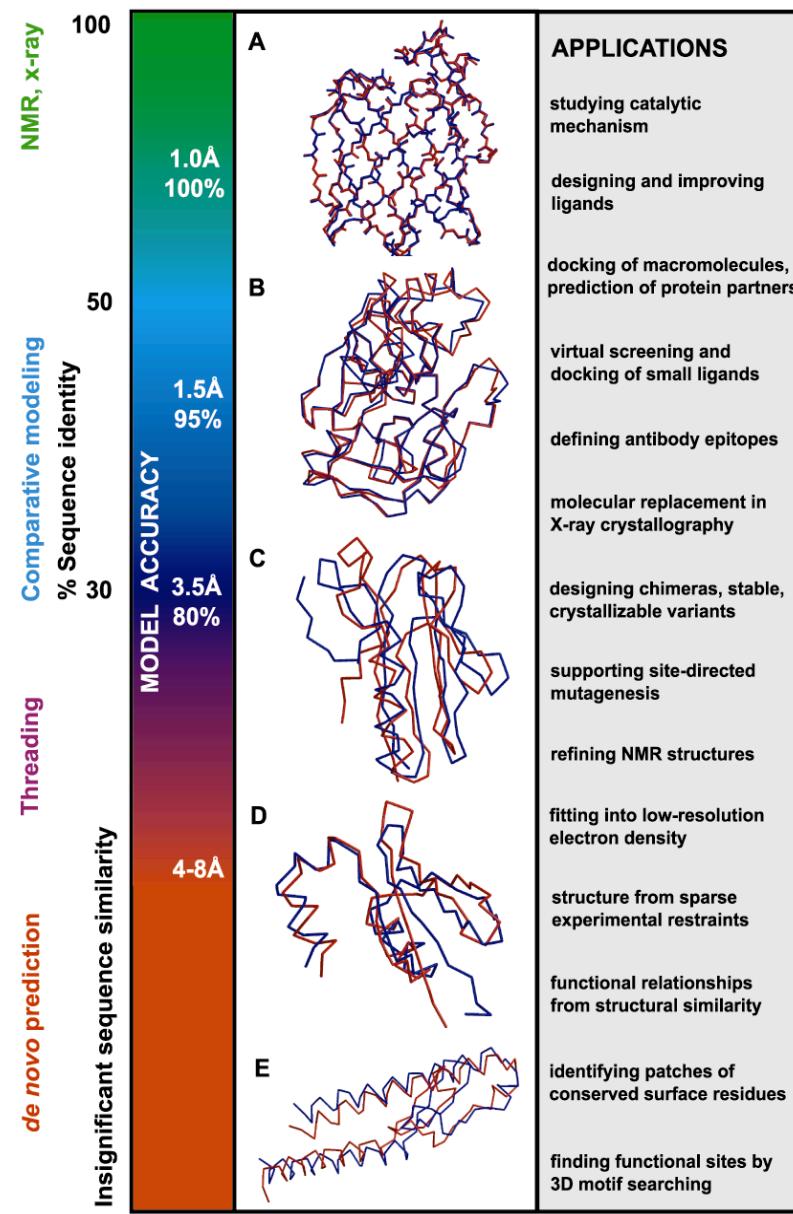
A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.

R. Sánchez & A. Šali, *Curr. Opin. Str. Biol.* 7, 206, 1997.

M. Marti et al. *Ann. Rev. Biophys. Biomol. Struct.*, 29, 291, 2000.

<http://salilab.org/>

Utility of protein structure models, despite errors



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

General References

Protein Structure Prediction:

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

Comparative Modeling:

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

<http://www.salilab.org/modeller/workshop/links/>

Untitled Document
file:///Users/marcius/Documents/0Teaching/0406%20BMC%20Upp ~ Google

Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints



A I L V G S M P R R D G M E R K D L L K A N V K I F K C O G A
V E V C P N D C P T S G P N D V I M P D E C D G A L G P
I A C T P C P Y N I Q G S - I Y A I D A D A D A D A D A
Q - I A C G A C K P E C P V N I Q G S - I Y A I D A D S

[MODELLER Home](#)
[Tutorial](#)
[Literature](#)
[Links](#)
[Examples...](#)
[Basic](#)
[Advance](#)
[Iterative](#)
[Difficult](#)

MODELLER WORKSHOP LINKS

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/
PS	S	http://www.structuralgenomics.org/
Sacch3	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/

Web site...

<http://www.salilab.org/modeller/workshop/>

The screenshot shows a web browser window titled "Untitled Document" with the URL "file:///Users/marcius/Documents/0Teaching/0406%20BMC%20Upp" in the address bar. The page content is as follows:

Modeller
Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints

MODELLER Home | Tutorial | Literature | Links | Examples... | MODELLER WORKSHOP TUTORIAL

MODELLER WORKSHOP 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. 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. Advance 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 its 4 steps process.
This exercise introduces the concept of MOULDING to improve the accuracy of comparative models.

Acknowledgments

Protein Structure Modeling

Andrej Sali

Bino John

Narayanan Eswar

Ursula Pieper

Roberto Sánchez (MSSM)

András Fiser (AECOM)

Francisco Melo (CU, Chile)

Azat Badretdinov (Accelrys)

M. S. Madhusudhan

Ash Stuart

Nebojša Mirkovic

Valentin Ilyin (NE)

Eric Feyfant (GI)

Min-Yi Shen

Ben Webb

Rachel Karchin

Mark Peterson

Brain Lipid Binding Protein

Liang Zhu (RU)

Nat Heintz (RU)

BRCA1

A. Monteiro (Cornel)

Fly p53

Shengkan Jin (RU)

Arnie Levine (RU)

<http://salilab.org>

1D to 3D for biologists

David Huassler (UCSC)

Jim Kent (UCSC)

Daryl Thomas (UCSC)

Mark (UCSC)

Rolf Apweiler (EBI)

Chimera

P. Babbitt

T. Ferrin

Ribosomes

J. Frank

Structural Genomics

Stephen Burley (SGX)

John Kuriyan (UCB)

NY-SGRC

Mast Cell Proteases

Rick Stevens (BWH)

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

Yeast NPC

Tari Suprapto (RU)

Julia Kipper (RU)

Wenzhu Zhang (RU)

Liesbeth Veenhoff (RU)

Sveta Dokudovskaya (RU)

J. Zhou (USC)

Mike Rout (RU)

Brian Chait (RU)

BMC WorkShop

Protein Structure Prediction

template selection

(sequence-structure alignment)

Marc A. Marti-Renom & Damien Devos

Department of Biopharmaceutical Sciences, UCSF

Summary

- Structural space! (and domains)
 - Structure-Structure comparisons
 - Some theory
 - Coverage .vs. Accuracy
 - How can we compare structures...
 - SALIGN (properties comparison)
 - VAST (vector alignment)
 - CE (local heuristic comparison)
 - MAMMOTH (vector alignment)
 - How we classify the structural space...
 - SCOP (manual)
 - CATH (semi-automatic)
 - DBAli (fully automatic and comprehensive)
 - ModDom application
 - What we know...
 - Sparseness in the protein structure and sequence spaces

Template Selection

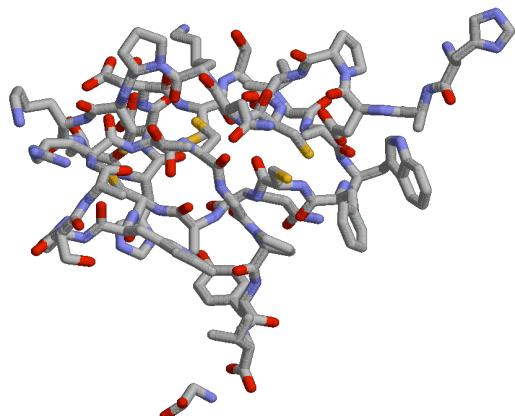
“Structural Space”

Structure-Structure alignments

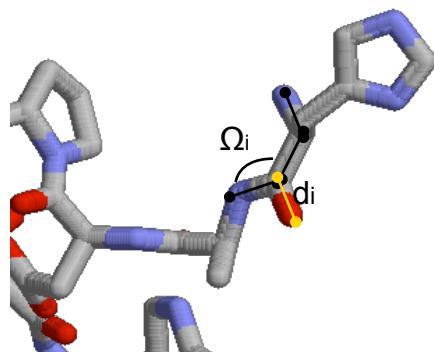
As any other bioinformatics problem...

- Representation
- Scoring
- Optimizer

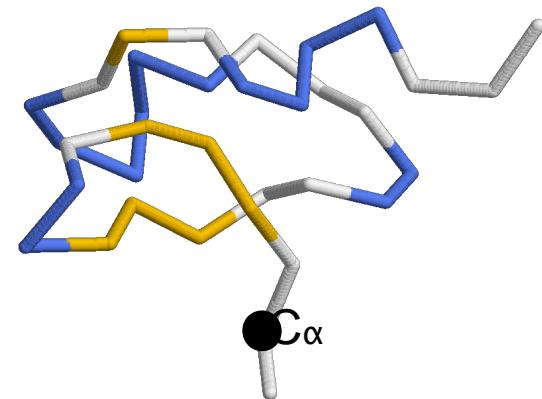
Representation Structures



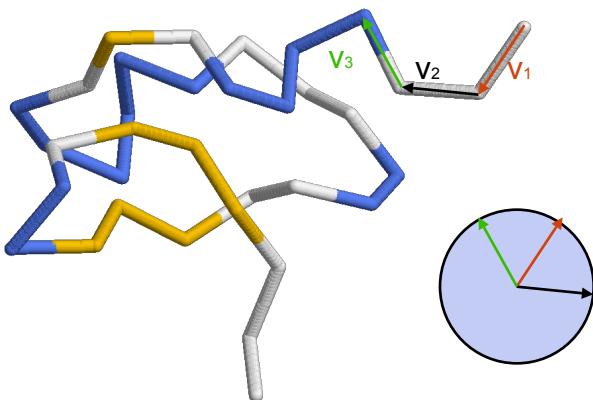
All atoms and coordinates



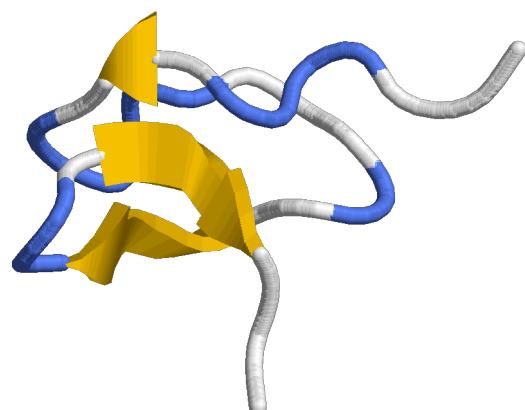
Dihedral space or distance space



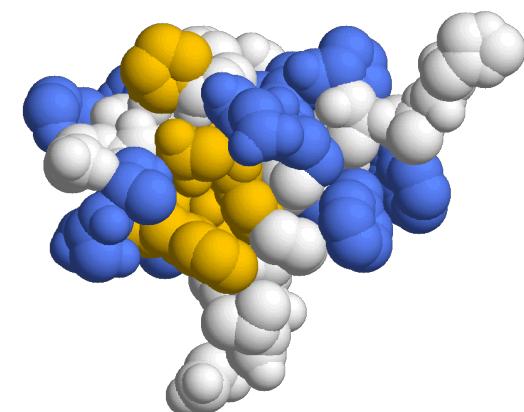
Reduced atom representation



Vector representation



Secondary Structure



Accessible surface (and others)

Scoring Raw scores

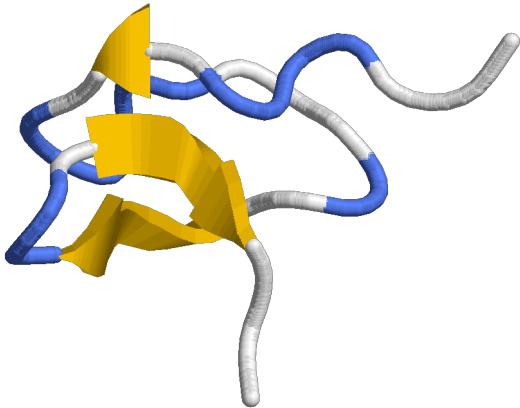
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9	-1	-1	-3	0	-3	-3	-4	-3	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-3	
T	-1	1	4	1	-1	1	0	1	0	0	-1	0	0	-1	-2	-2	-2	-2	-3	
P	-3	-1	1	7	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	
A	0	1	-1	-1	4	0	-1	-2	-1	-1	-2	-1	-1	-1	-1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

2/

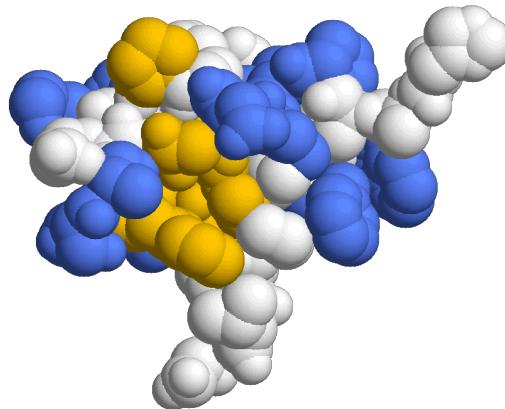
Aminoacid substitutions

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

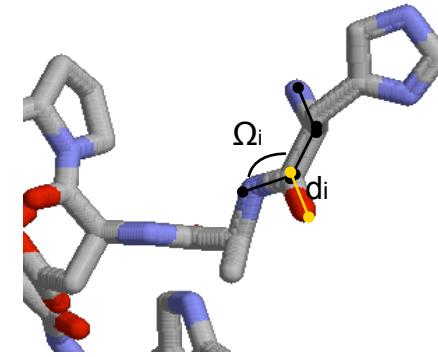
Root Mean Square Deviation



Secondary Structure (H,B,C)



Accessible surface (B,A [%])



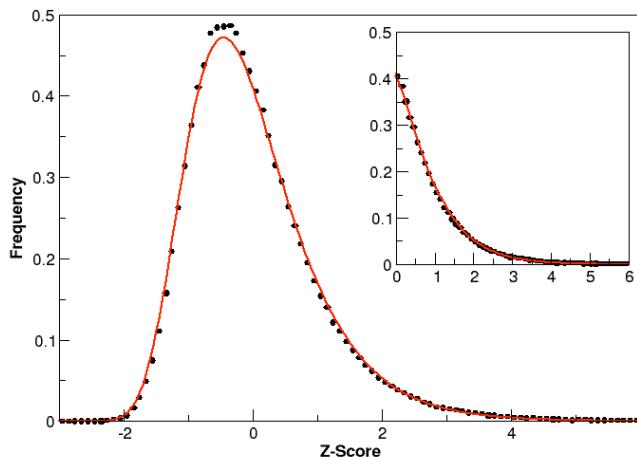
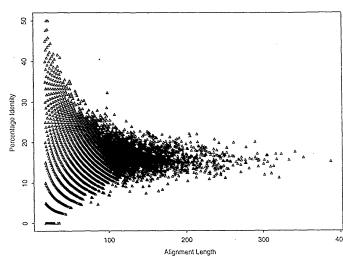
Angles or distances

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

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

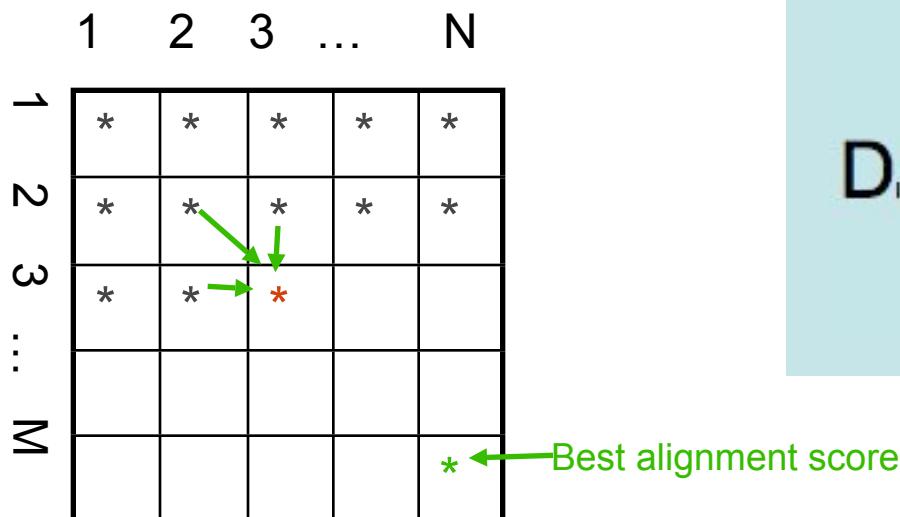
$$P(s \geq x) = 1 - \exp(e^{-\lambda (s-\mu)})$$

Karlin and Altschul, 1990 PNAS 87, pp2264

06/10/2004

Optimizer

Global dynamic programming alignment

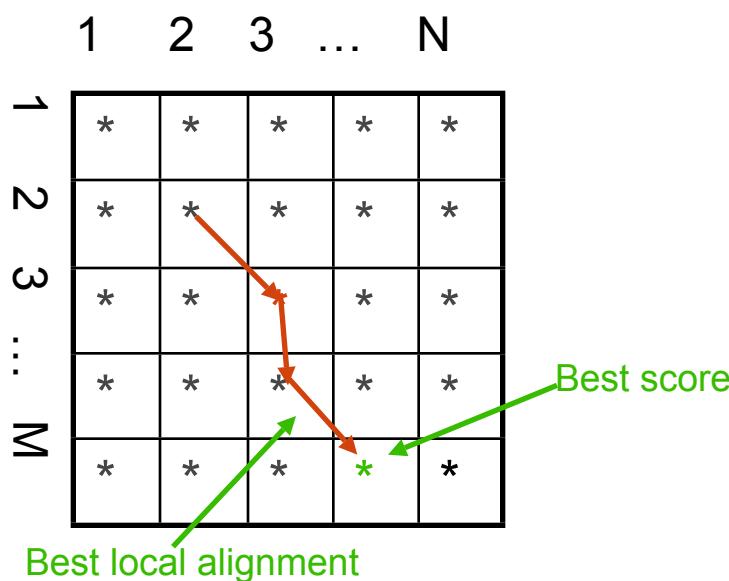


$$D_{i,j} = \min \begin{cases} D_{i,j-1} + \text{Score}_{(\Delta, r_j)} \\ D_{i-1,j-1} + \text{Score}_{(r_i, r_j)} \\ D_{i-1,j} + \text{Score}_{(r_i, \Delta)} \end{cases}$$

Backtracking to get the best alignment

Optimizer

Local dynamic programming alignment

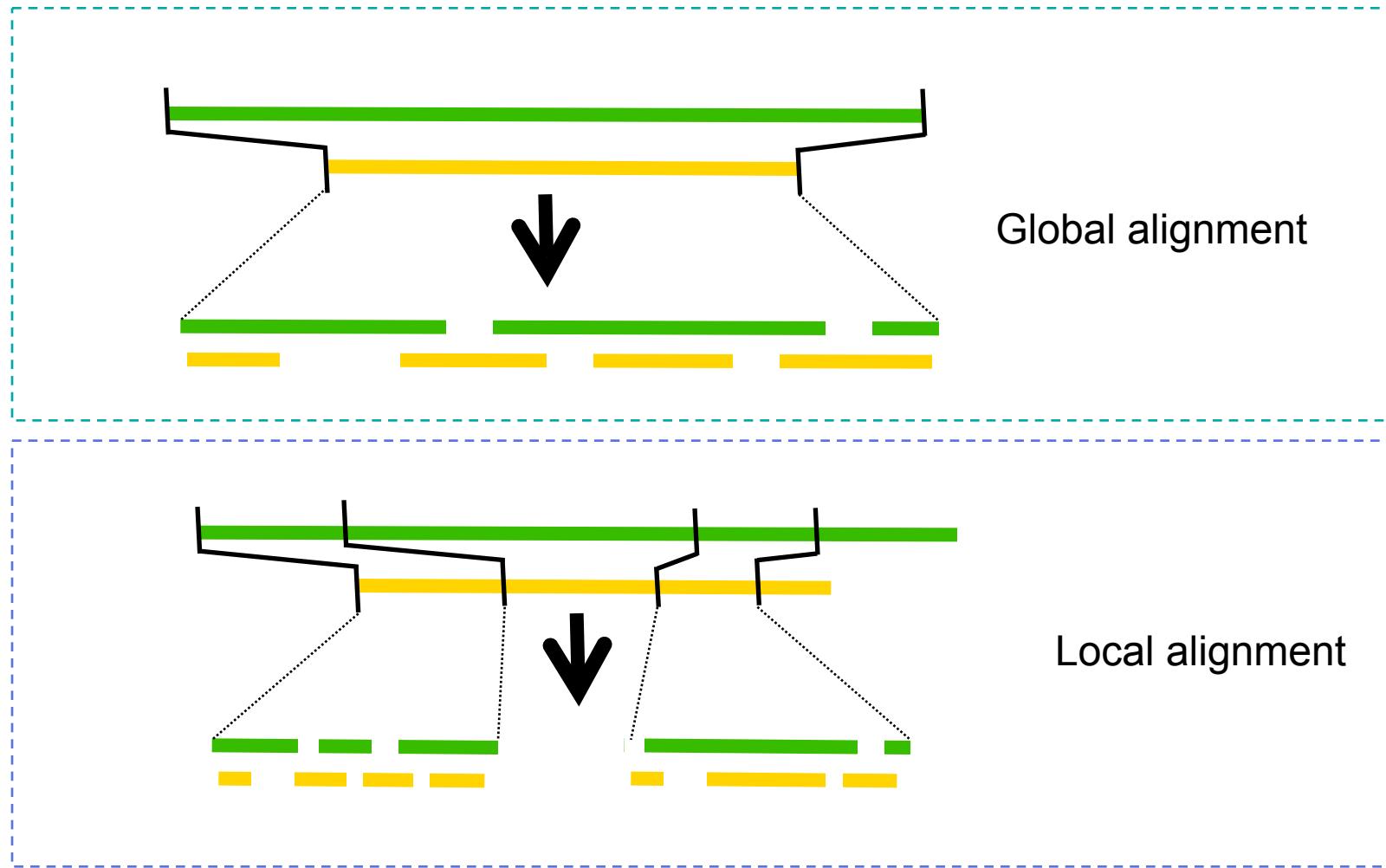


$$D_{i,j} = \min \begin{cases} D_{i,j-1} + \text{Score}_{(\Delta, rj)} \\ D_{i-1,j-1} + \text{Score}_{(ri, rj)} \\ D_{i-1,j} + \text{Score}_{(ri, \Delta)} \\ 0 \end{cases}$$

Backtracking to get the best alignment

Optimizer

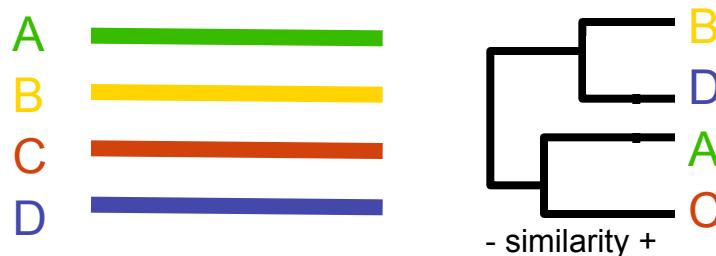
Global .vs. local alignment



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

B Align the most similar pair

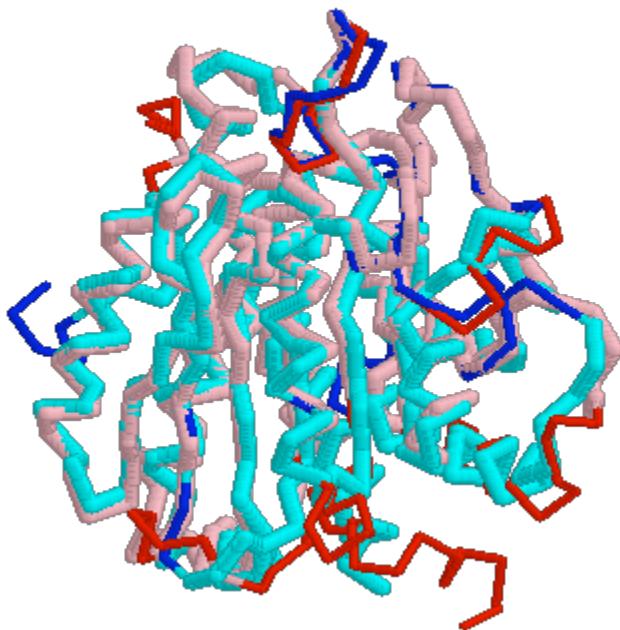
A Align next most similar pair

Align B-D with A-C

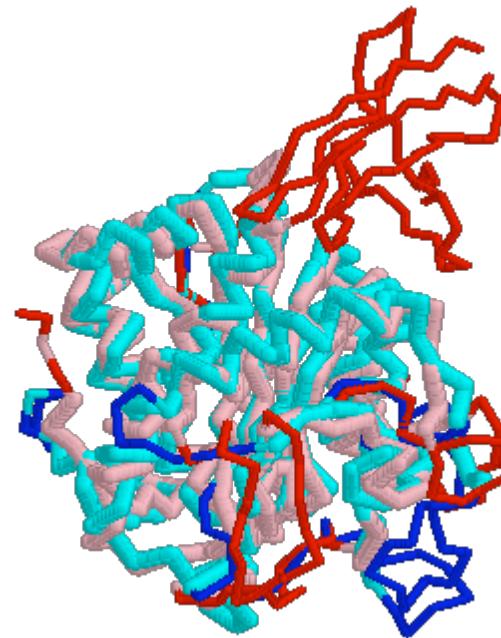
B New gap in A-C to optimize its alignment with B-D

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

Coverage .vs. Accuracy



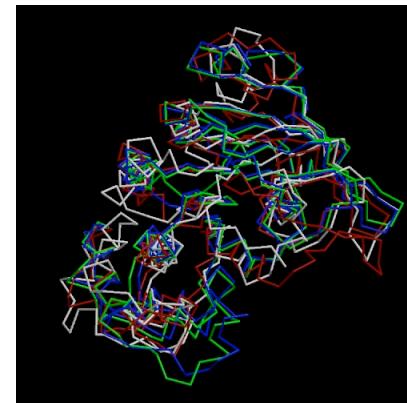
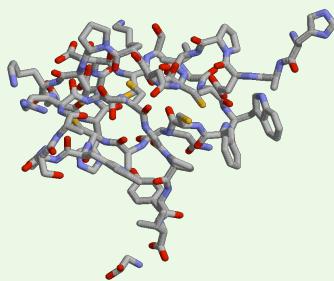
Coverage ~90% C α



Coverage ~75% C α

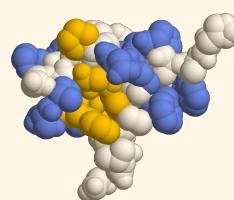
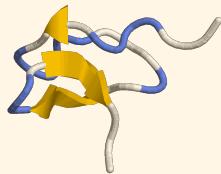
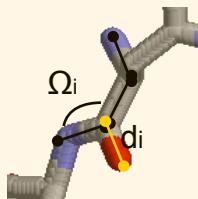
Same RMSD $\sim 2.5\text{\AA}$

Structural alignment by properties conservation (SALIGN-MODELLER)



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

Computationally expensive



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

R_{i,j}

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

$S_{i,j}$

B_{i,j}

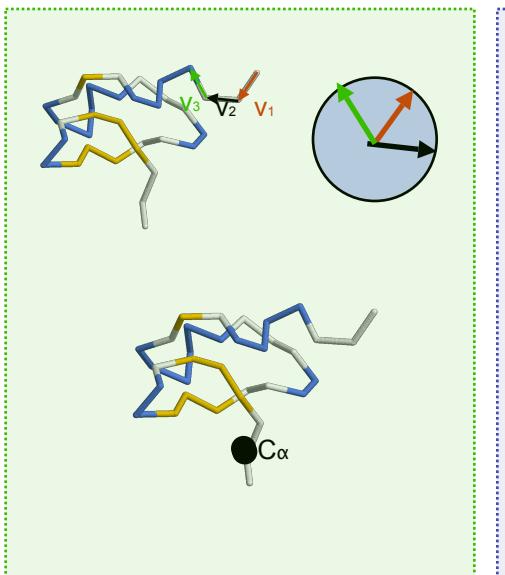
i,j

Structural alignment by properties conservation (SALIGN-MODELLER)

<http://www.salilab.org/dbali/>

The screenshot shows a Microsoft Internet Explorer window displaying the DBAli v2.0 tools page. The title bar reads "DBAli v2.0 tools page - Microsoft Internet Explorer". The address bar shows the URL http://salilab.org/DBAli/?page=tools&action=f_salign. The page header includes the "UCSF | Sali Lab | MAMMOTH" logo and the "DBAli v2.0" title. A banner at the top features a red ribbon-like 3D molecular structure. On the left, a sidebar menu lists "Home", "Search DBAli", "Tools", and "Help". A yellow box titled "DBAli ALERT!" contains the message: "09/21/2003 -- You are visiting the DBAli v2 pages. This pages contain the updated DBAli database. You can still visit the old DBAli database [here](#)". The main content area is titled "DBAli. Tools associated to the database." and lists several tools: "Cluster a list of chains", "Cluster from a chain", "Define domains from a chain", "Get a multiple structure alignment of a list of chains", "Database statistics", and "Download DBAli". Below this is a section titled "Get a multiple structure alignment of a list of chains." with a form. The form has a label "File with a list of chains:" followed by a file input field, a "Browse..." button, and a help icon. At the bottom of the form are two buttons: "SALIGN" and "Clear". The footer of the page includes links for "Reference", "Download", "Statistics", "Suggestions", and "Visitors: 1407 © 2003 - 2004 Marti-Renom". The status bar at the bottom of the browser window shows "Internet".

Vector Alignment Search Tool (VAST)



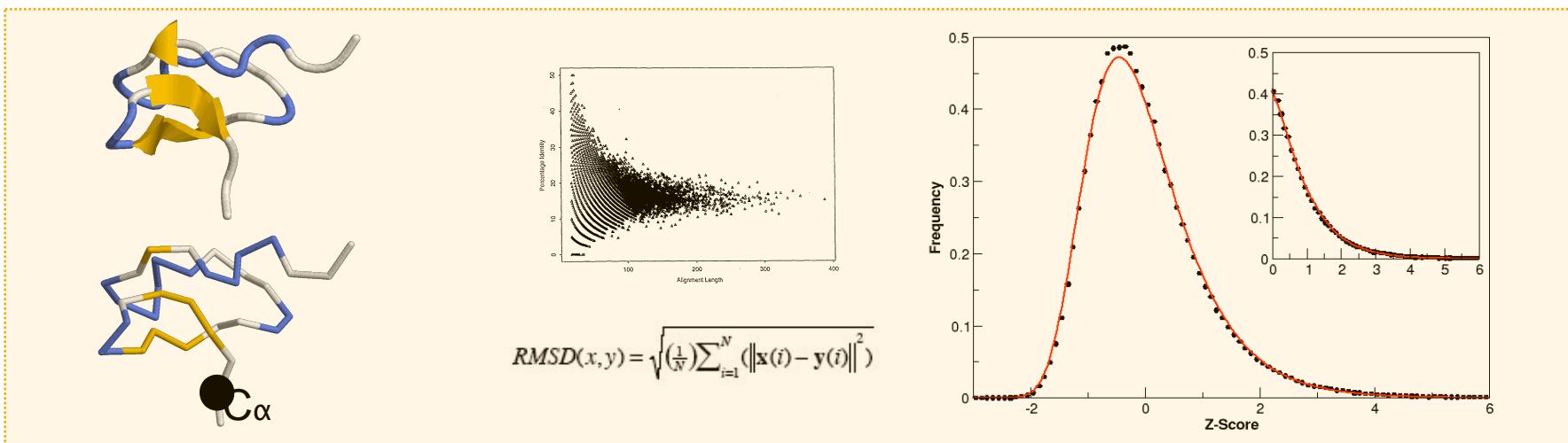
- Graph theory search of similar SSE
- Refining by Monte Carlo at all atom resolution

NCBI CDD phm00599 with Query Sequence added: Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Stop Search Favorites Media
Address http://www.ncbi.nlm.nih.gov/Structure/cdd/cdderv.cgi?wdir=2&maxid=20&eltype=2&uid=1512&eh=5,0,23
Description: Phage lysozyme. This family includes lambda phage lysozyme and E. coli endolysin.
Taxon: 111
Status: Alignment from source
Aligned: 19 rows
Proteins: Click here for CDART summary of Proteins containing phm00599
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******
21am (query)	V	V	T	I	G	I
1PDC_A	V	V	T	I	G	I
1L2A	V	V	T	I	G	I
g1_509532	V	V	T	I	G	I
g1_126600	V	V	T	I	G	I
g1_126629	V	V	T	I	G	I
g1_126602	V	V	T	I	G	I
g1_6014318	V	V	T	I	G	I
g1_2313443	V	V	T	I	G	I

✓ Good scoring system with significance

Reduces the protein representation



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 Go Links

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

VAST Help Vector Alignment Search Tool try:

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

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 neighbor 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).

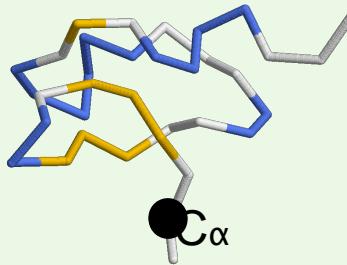
Structure Summary via PDB/MMDB Code [] Get

Install and test structure alignment viewers:
Get Cn3D v4.1 and [look at this example](#) to test!
[Read a bit more about VAST...](#)

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

Internet

Incremental combinatorial extension (CE)

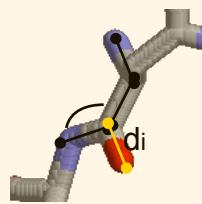


- Exhaustive combination of fragments
- Longest combination of AFPs
- Heuristic similar to PSI-BLAST



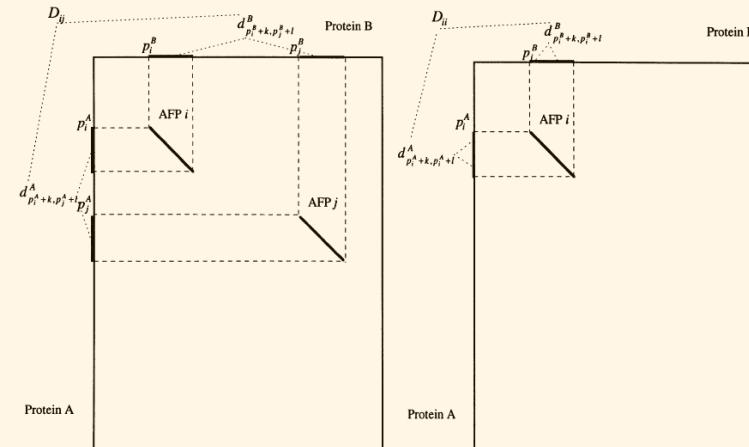
- ✓ FAST!
- ✓ Good quality of local alignments

Complicated scoring and heuristics



8 residues peptides

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



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 Links Go

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

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

Using the Combinatorial Extension (CE) Method

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

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

FIND

CALCULATE

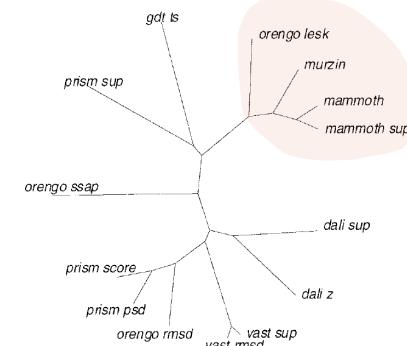
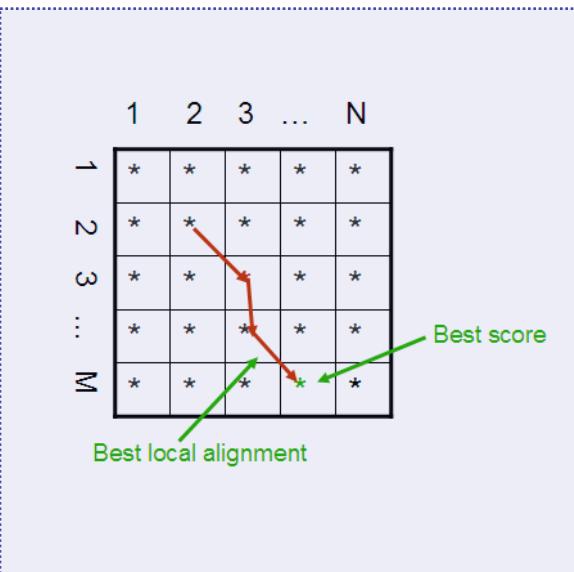
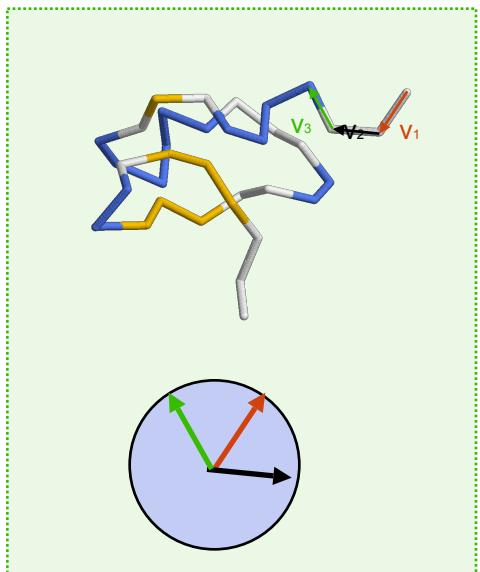
Find structural alignments by selecting from **ALL** or **REPRESENTATIVES** from the PDB.

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**.

Done Internet

Matching molecular models obtained from theory (MAMMOTH)

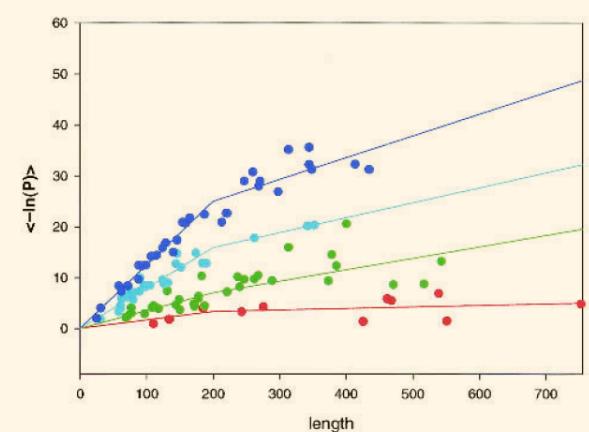
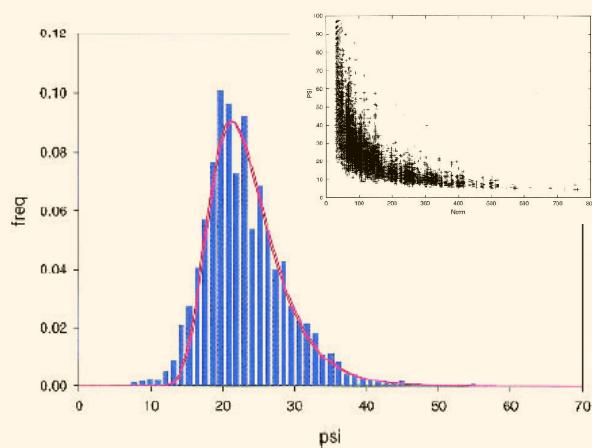


- ✓ VERY FAST!
- ✓ Good scoring system with significance

Reduces the protein representation

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

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



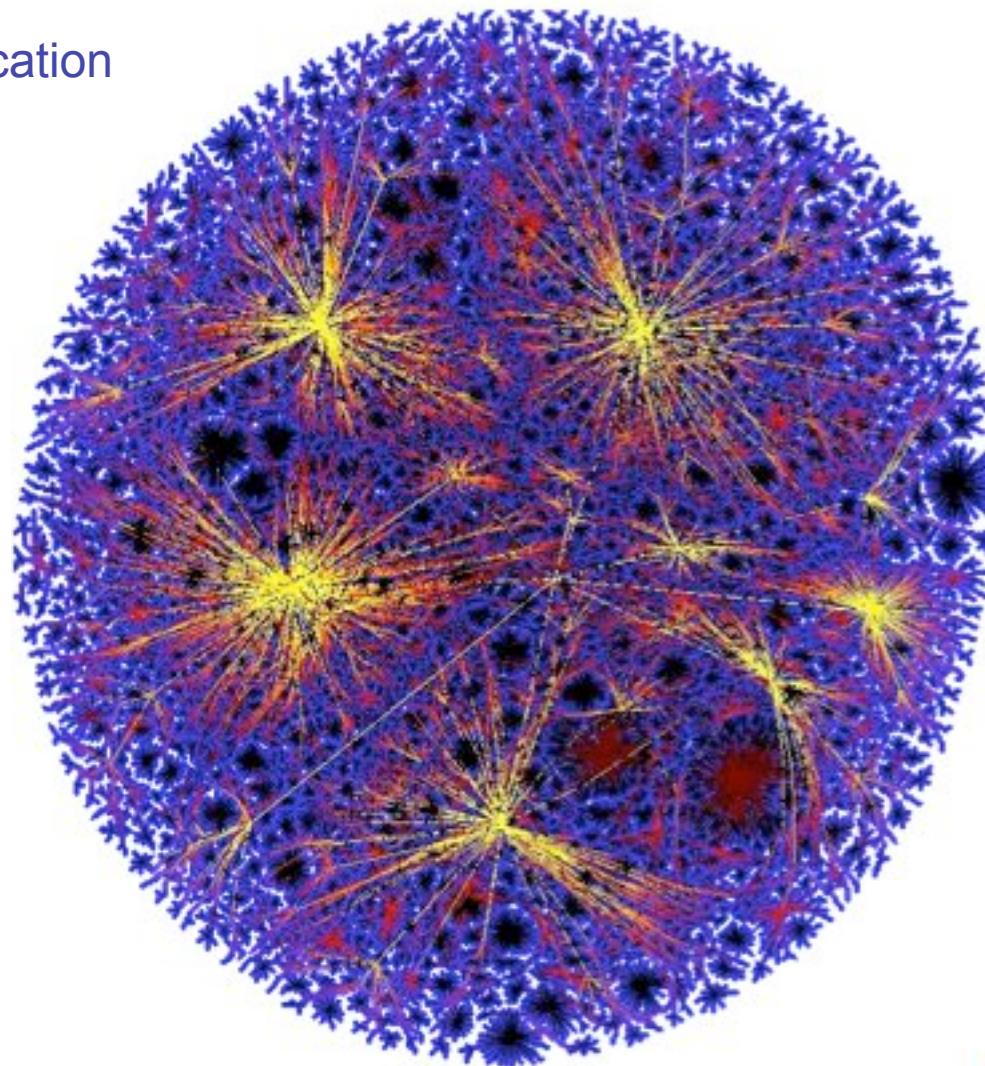
Matching molecular models obtained from theory (MAMMOTH)

<http://fulcrum.physbio.mssm.edu:8083/>

The screenshot shows the MAMMOTH web interface running in Microsoft Internet Explorer. The title bar reads "Protein Structure Alignment Server - Microsoft Internet Explorer". The address bar contains the URL "http://fulcrum.physbio.mssm.edu:8083/mammoth/". The main content area features a logo of a mammoth on the left and the text "MAMMOTH" and "MAatching Molecular Models Obtained from THeory" in the center. Below this, there are two input fields: one for "PREDICTION" coordinates (PDB format) and one for "EXPERIMENT" coordinates (PDB format), each with a "Browse..." button. The bottom status bar shows "Done" and "Internet".

Classification of the structural space

SCOP classification



SCOP 1.65 database

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

The screenshot shows the SCOP homepage with the following content:

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 browse 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 [link 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 [link 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

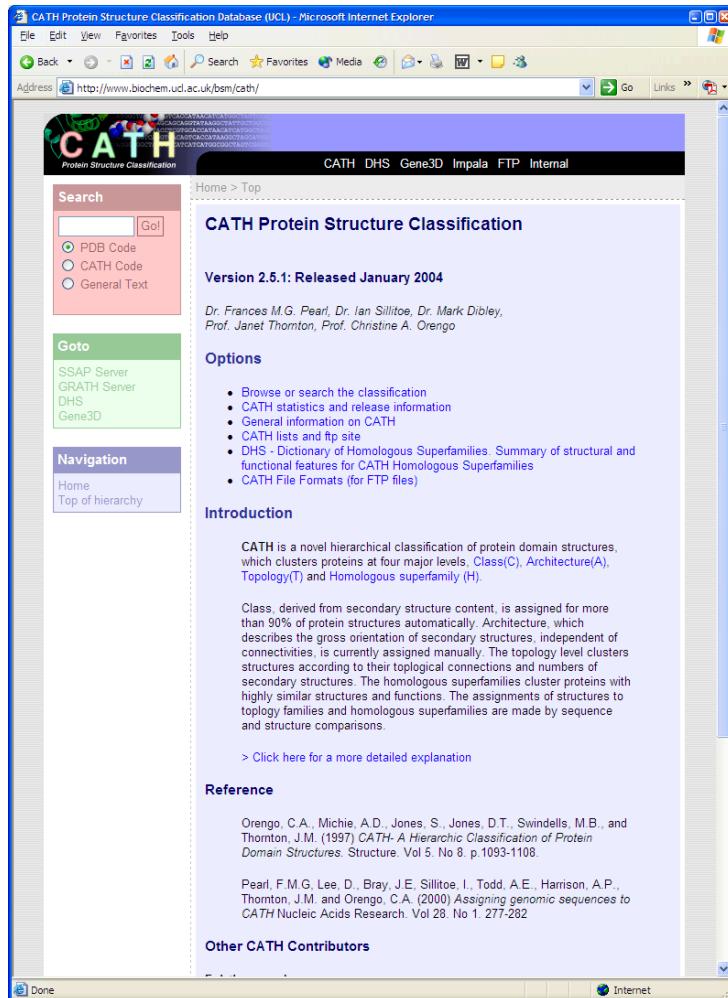
Class	Domain boundaries definition	Number of folds	Number of superfamilies	Number of families
All alpha proteins		179	299	480
All beta proteins		126	248	462
Alpha and beta proteins (a/b)		121	199	542
Alpha and beta proteins (a+b)		234	349	567
Multi-domain proteins		38	38	53
Membrane and cell surface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

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

06/10/2004

CATH 2.5.1 database

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

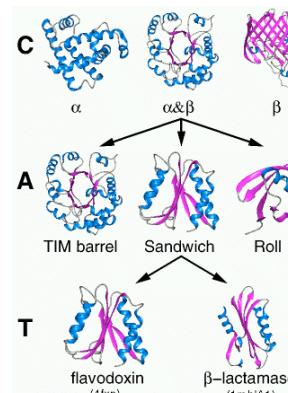


The screenshot shows the CATH 2.5.1 database homepage. The left sidebar includes a search bar with options for PDB Code, CATH Code, and General Text, and a Goto section with links to SSAP Server, GRATH Server, DHS, and Gene3D. The main content area features a heading 'CATH Protein Structure Classification' and 'Version 2.5.1: Released January 2004'. Below this are sections for 'Options' (listing browse/search, stats/info, general info, lists/ftp, and dictionary), 'Introduction' (describing the classification as hierarchical by Class, Architecture, Topology, and Homologous superfamily), and 'Reference' (citing Orengo et al., 1997 and Pearl et al., 2000). The bottom of the page includes a 'Done' button and an 'Internet' link.

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	2.5.1						
Date	28-01-2004						
	A	T	H	S	N	I	D
Mainly Alpha	5	227	428	948	1713	3946	10155
Mainly Beta	19	139	292	951	2344	5011	14259
Alpha Beta	12	368	648	2010	3631	8639	23025
Few Secondary Structures	1	86	91	114	225	378	952
Multi-domain chains	1	1053	1057	1071	2186	5801	12471
Preliminary single domain assignments	1	371	374	422	479	789	1663
Multi-domain domains	2	31	31	49	67	139	287
CATH-35 Sequence families	1	997	997	997	1108	2154	3431
Fragments from multi-chain domains	1	28	28	30	33	56	106

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

06/10/2004

DBAli_{v2.0} database

<http://salilab.org/DBAli/>

The screenshot shows the DBAli v2.0 home page in Microsoft Internet Explorer. The title bar reads "DBAli v2.0 home page - Microsoft Internet Explorer". The address bar shows the URL "http://salilab.org/DBAli/". The page content includes a banner for "UCSF | Sali Lab | MAMMOTH" and a red ribbon logo. The main text area says "DBAli. A Database of Pairwise Structure Alignments." by "Marc A. Marti-Renom and Andrej Sali" with help from "A. Ortiz's MAMMOTH program". A sidebar on the left has links for "Home", "Search DBAli", "Tools", "Help", and a "DBAli ALERT!" section. The alert section contains a message about visiting the DBAli v2 pages and provides a link to the old DBAli database. At the bottom, there are links for "Reference", "Download", "Statistics", "Suggestions", and "Visitors: 1150 © 2003 - 2004 Marti-Renom".

Uses MAMMOTH for superimposition

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

Does not provide (yet) a stable classification

Last updated:

February 11th, 2004 (18:49h)

Number of chains in database:

48,094

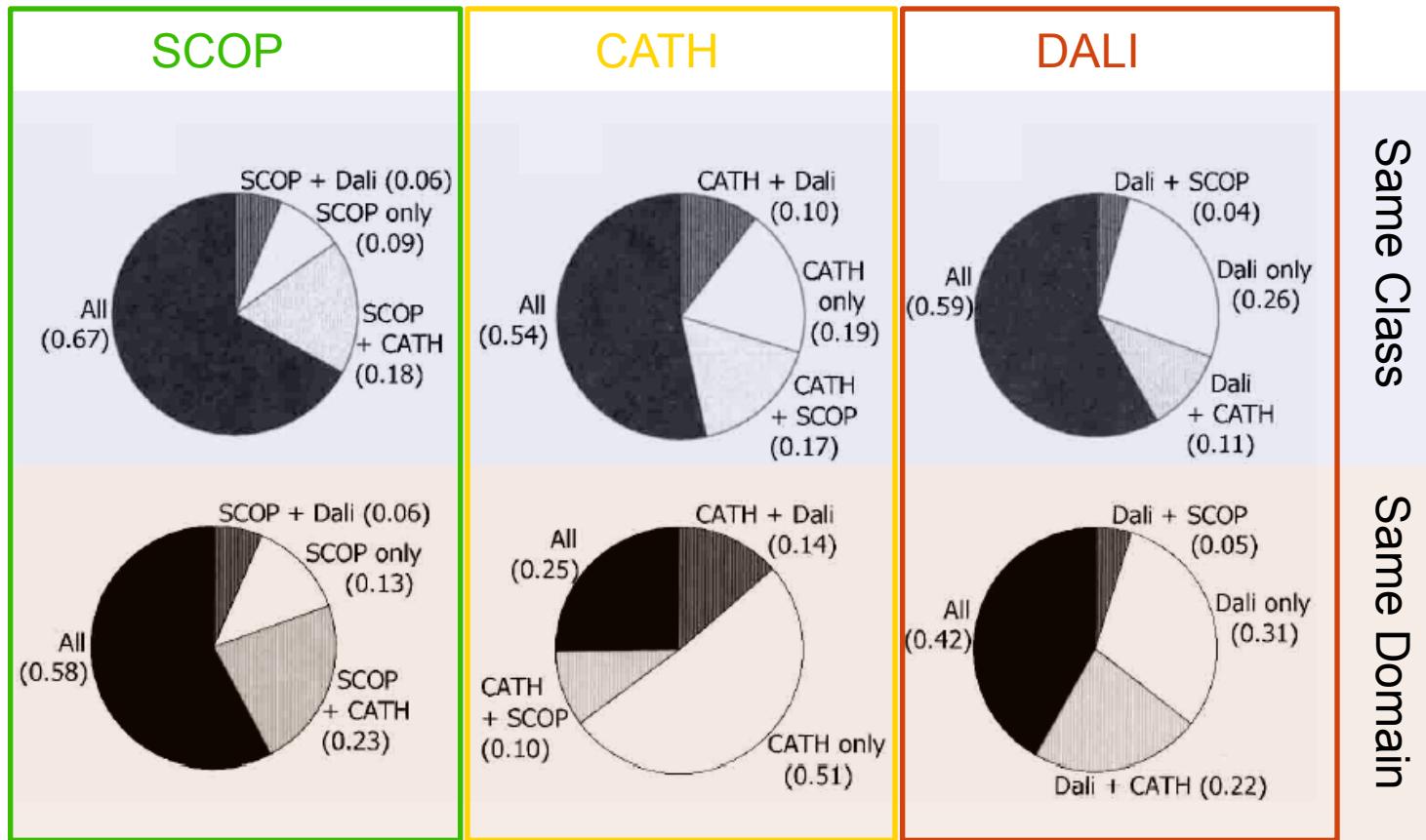
Number of structure-structure comparisons:

330,514,636

Classification of the structural space

Not an easy task!

Domain definition AND domain classification



Application (ModDom)

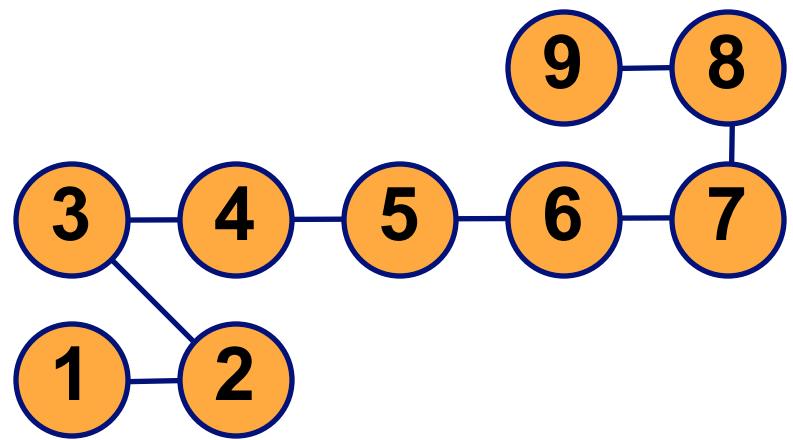
- Use of the DBAli data to define...
 - Protein Domains
 - Protein Fragments



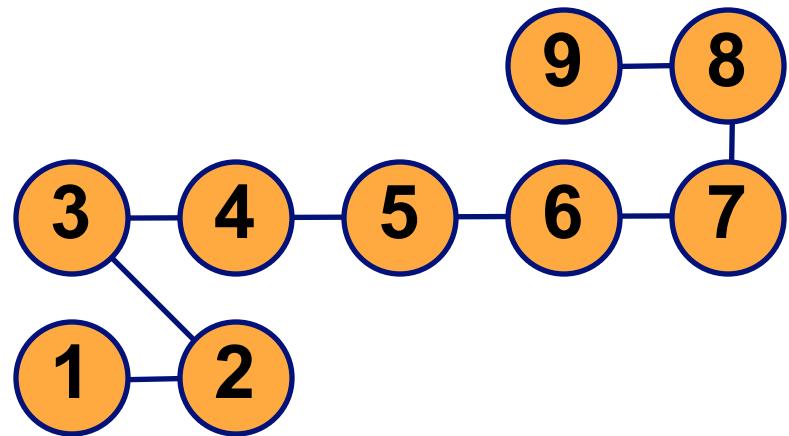
ModDom

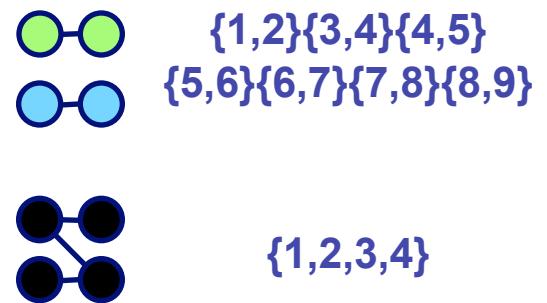
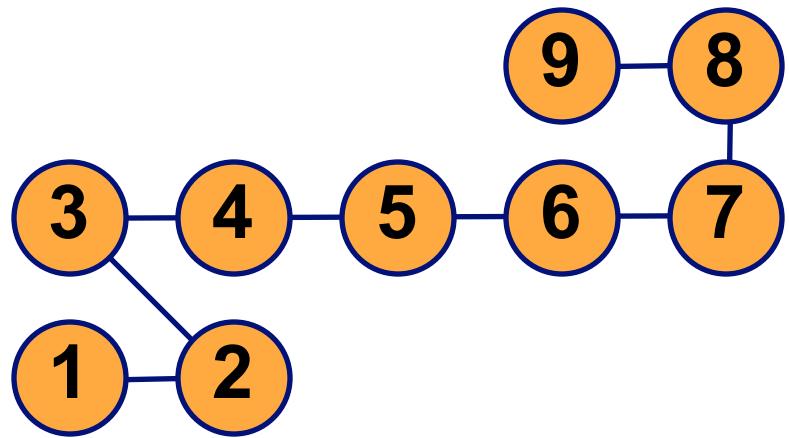
1phh (Oxydoreductase from *Pseudomonas fluorescens*)

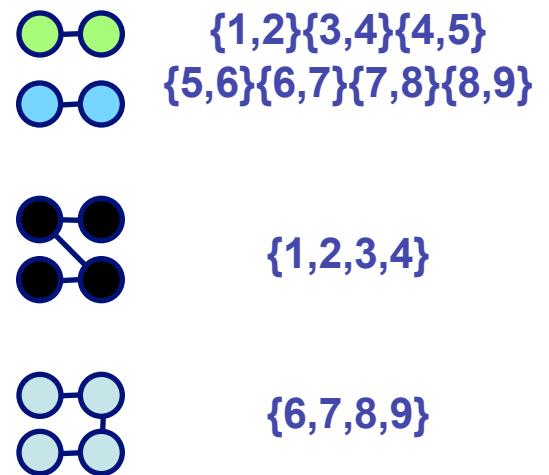
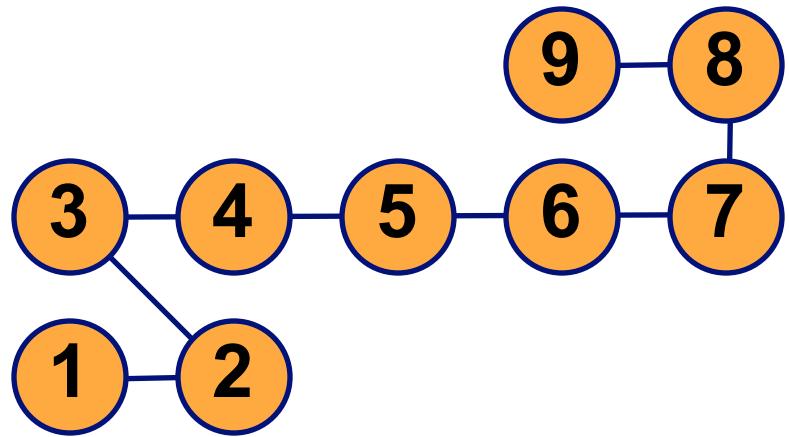


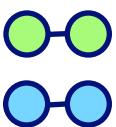
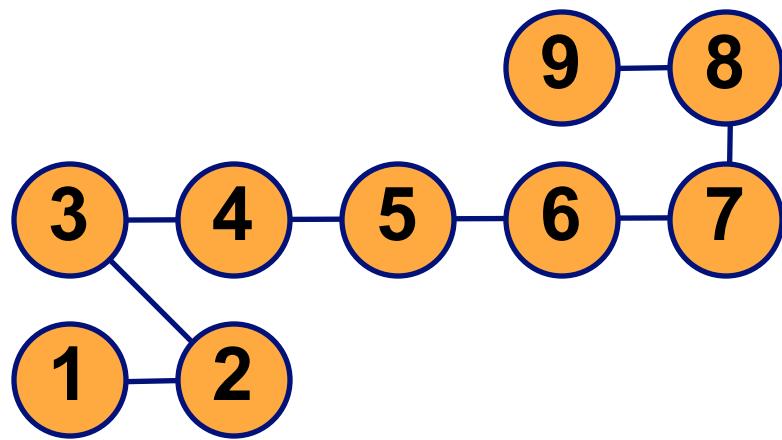


 $\{1,2\}\{3,4\}\{4,5\}$
 $\{5,6\}\{6,7\}\{7,8\}\{8,9\}$

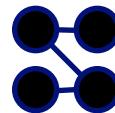




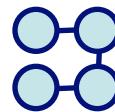




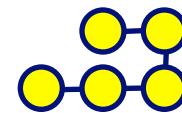
$\{1,2\}\{3,4\}\{4,5\}$
 $\{5,6\}\{6,7\}\{7,8\}\{8,9\}$



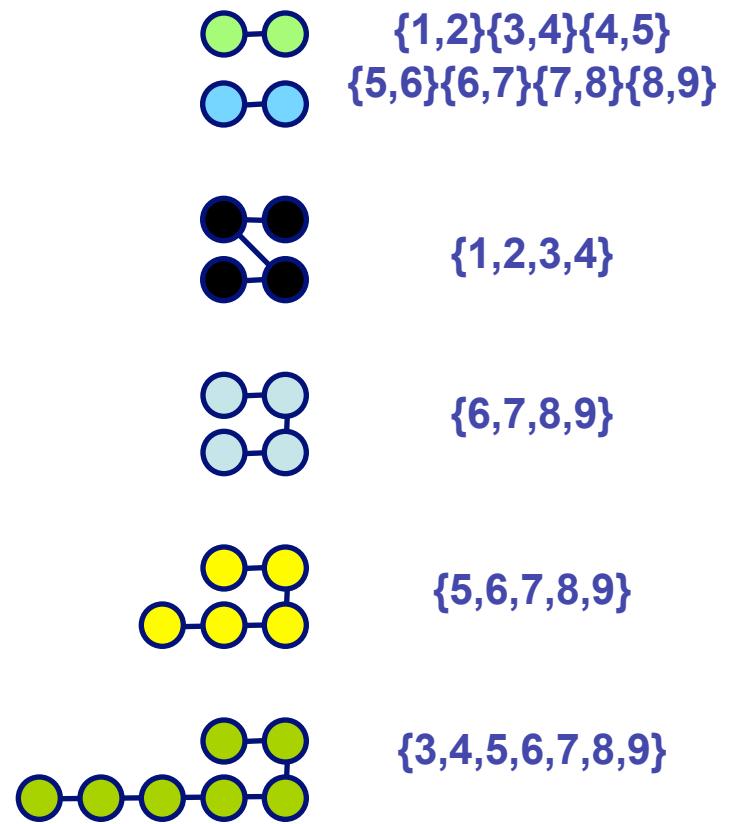
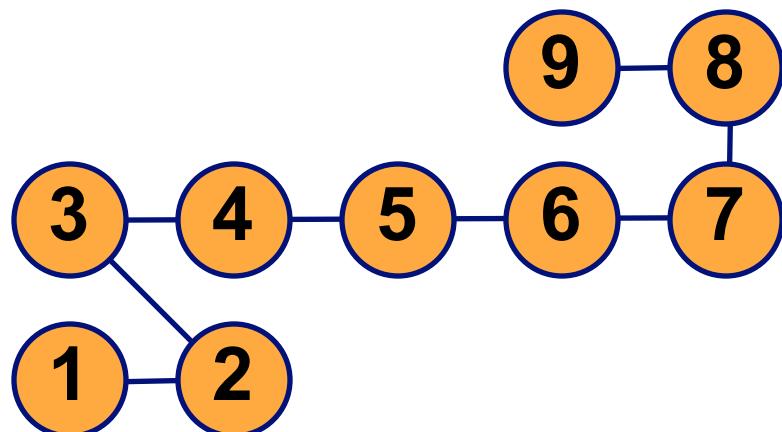
$\{1,2,3,4\}$

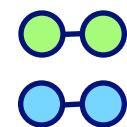
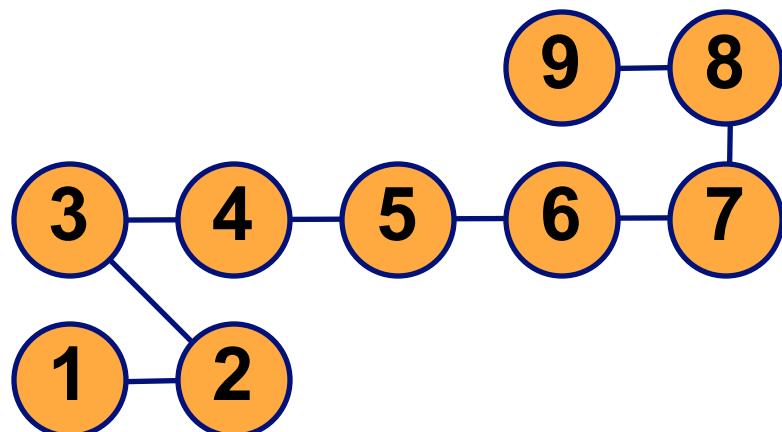


$\{6,7,8,9\}$



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

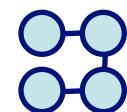




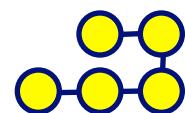
$\{1,2\}\{3,4\}\{4,5\}$
 $\{5,6\}\{6,7\}\{7,8\}\{8,9\}$



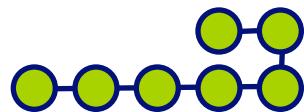
$\{1,2,3,4\}$



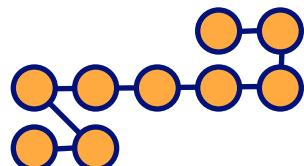
$\{6,7,8,9\}$



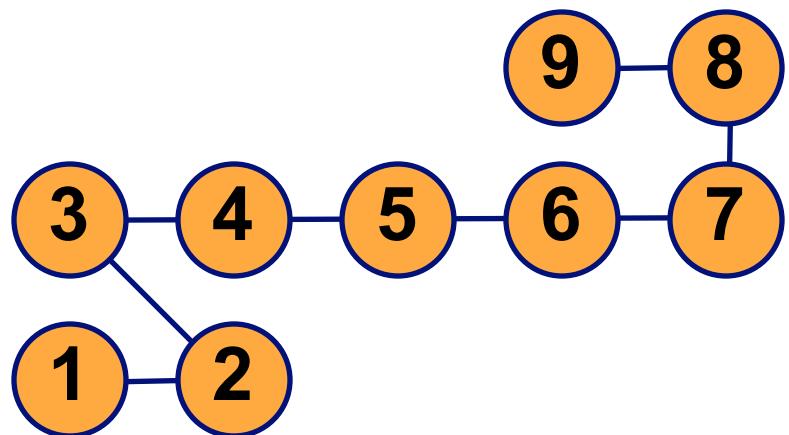
$\{5,6,7,8,9\}$



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



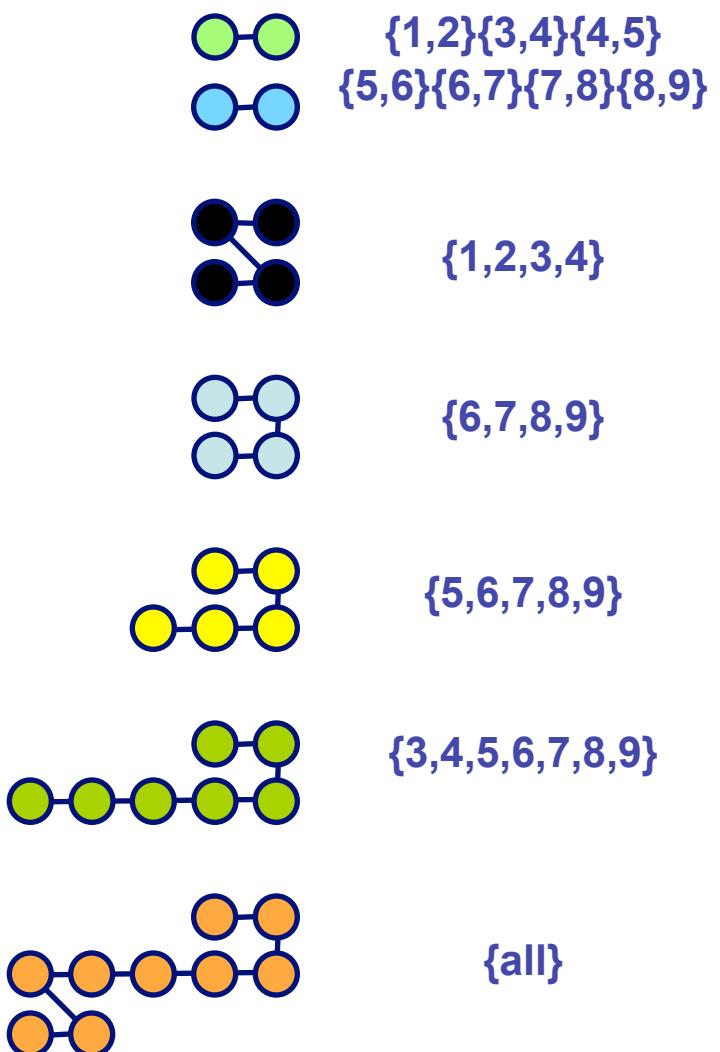
$\{\text{all}\}$

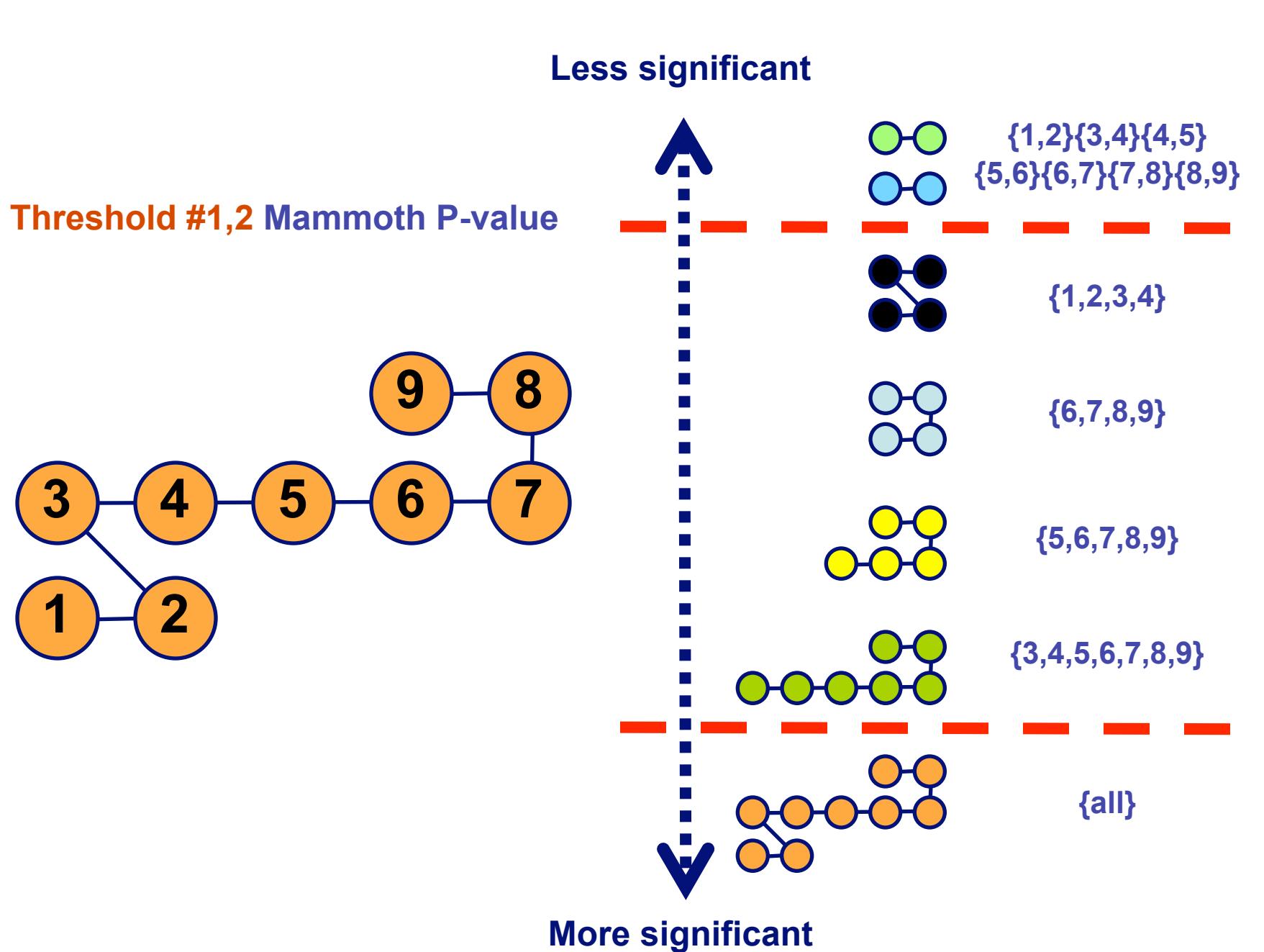


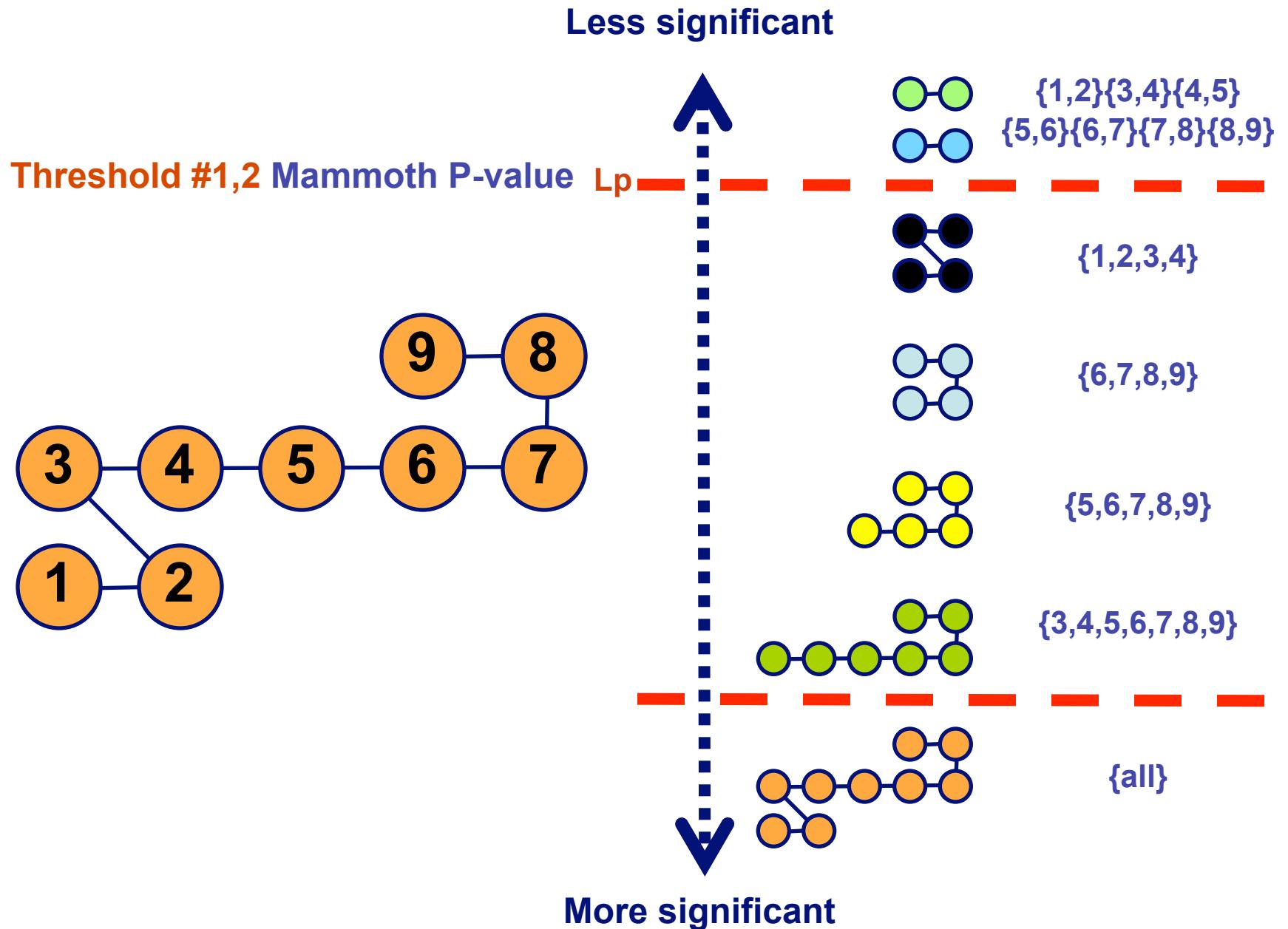
Less significant

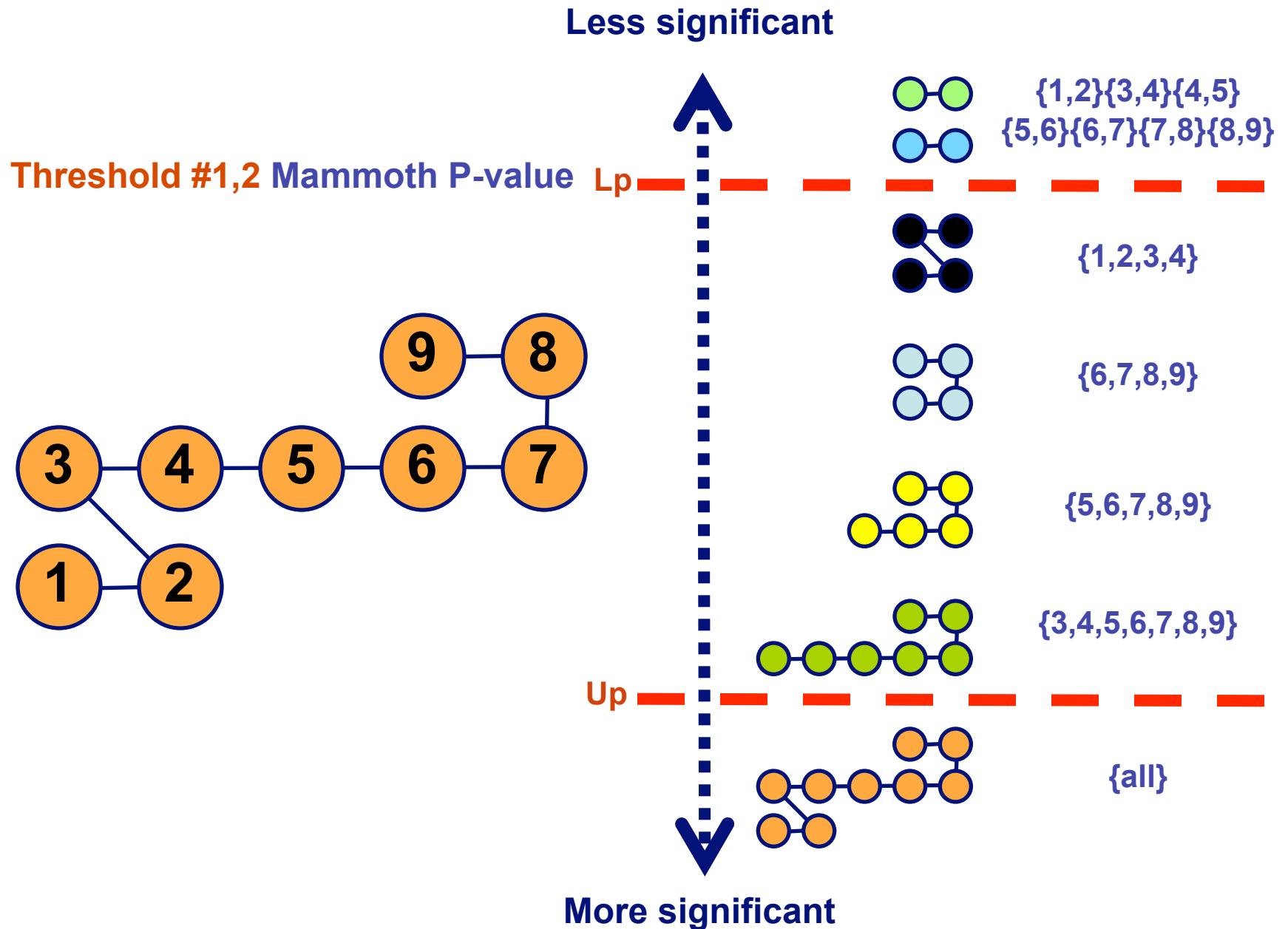


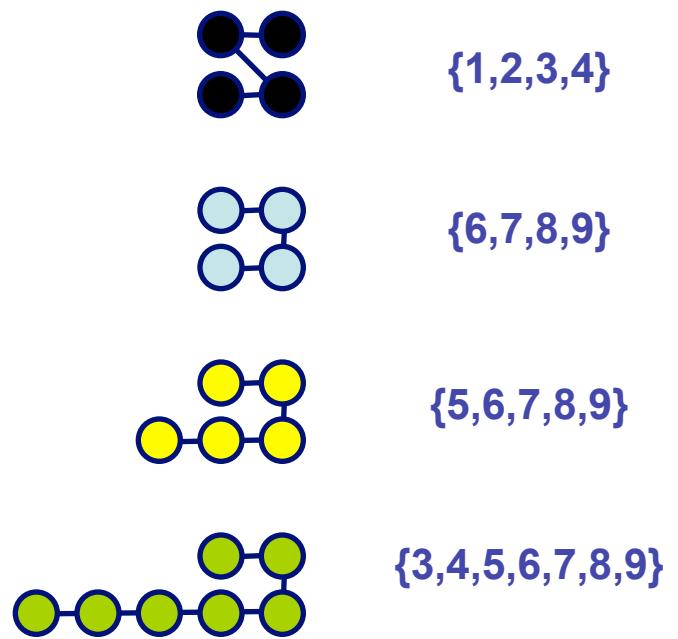
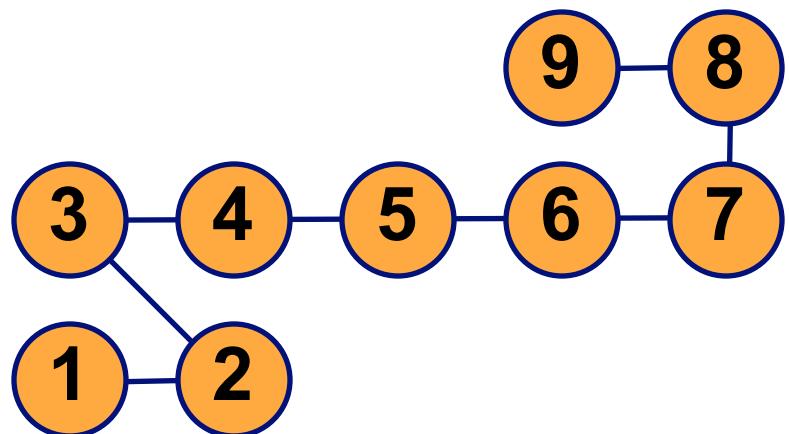
More significant

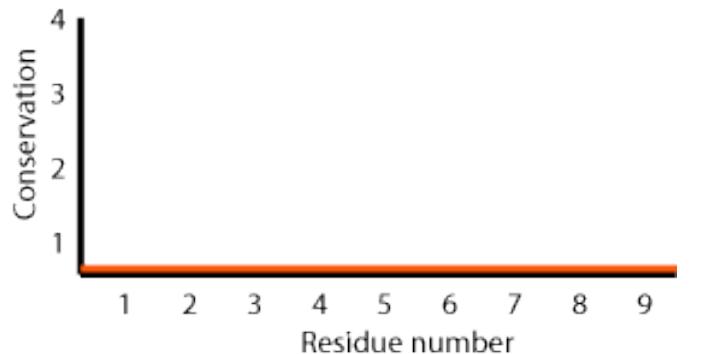
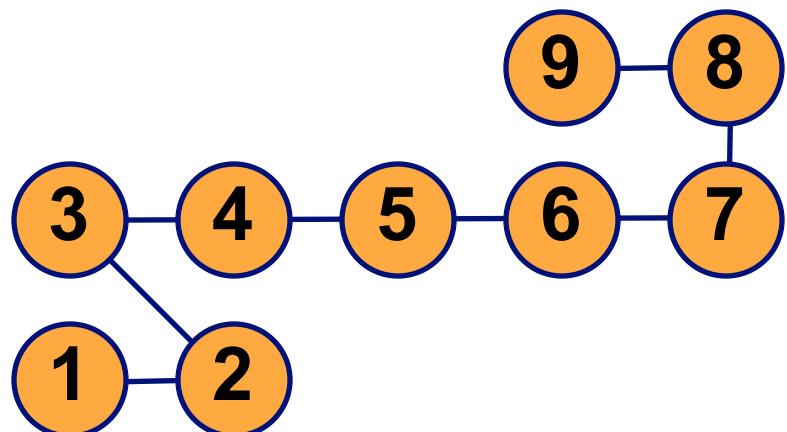
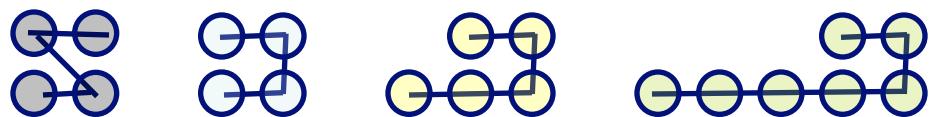




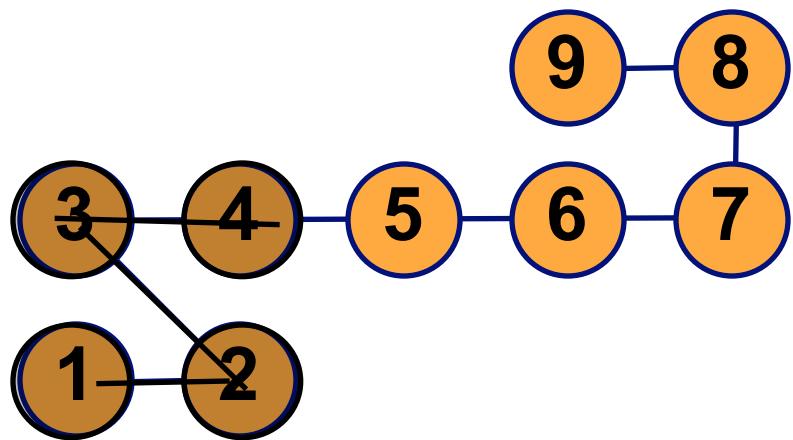
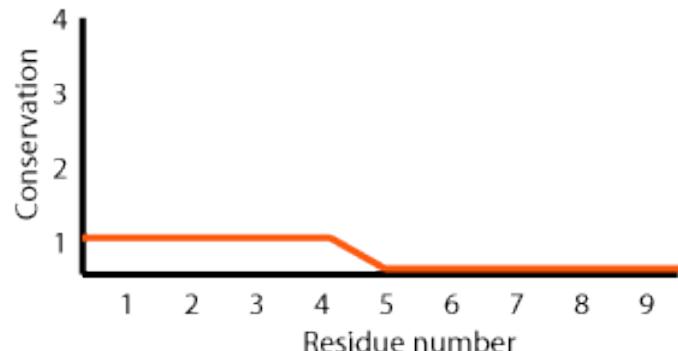
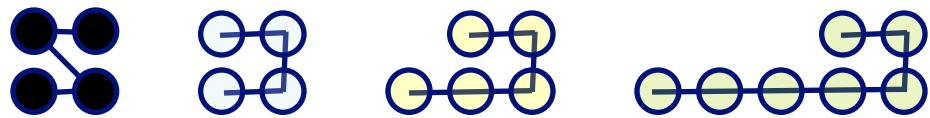




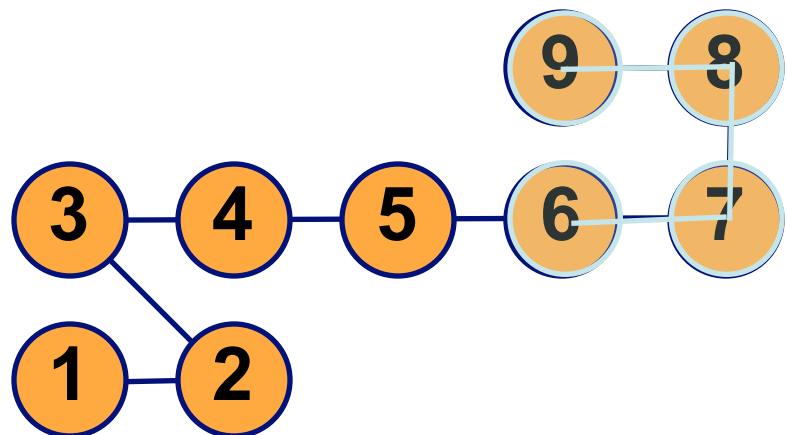
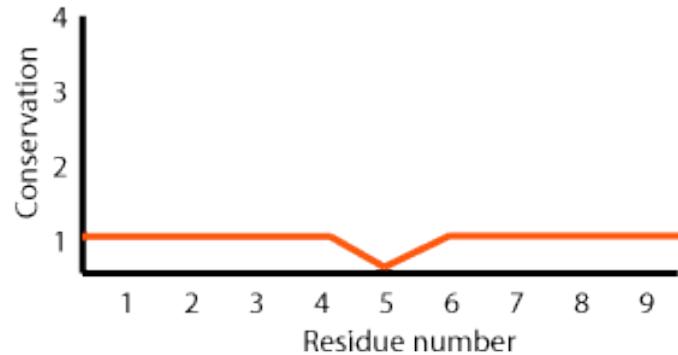
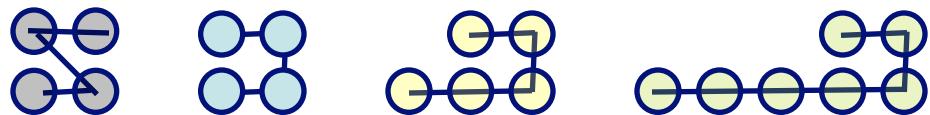




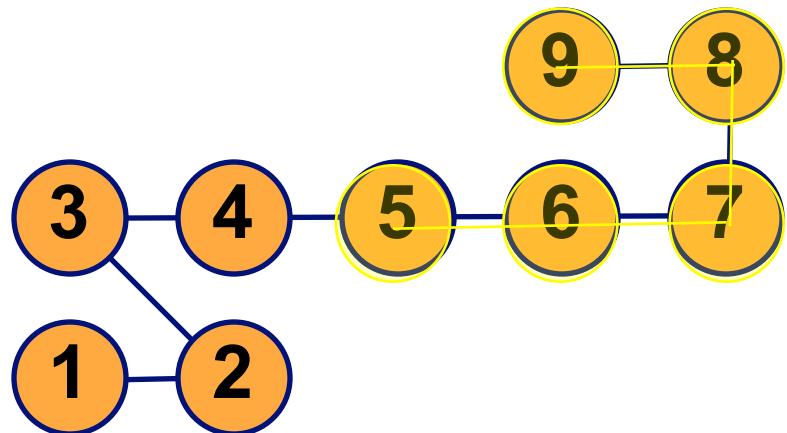
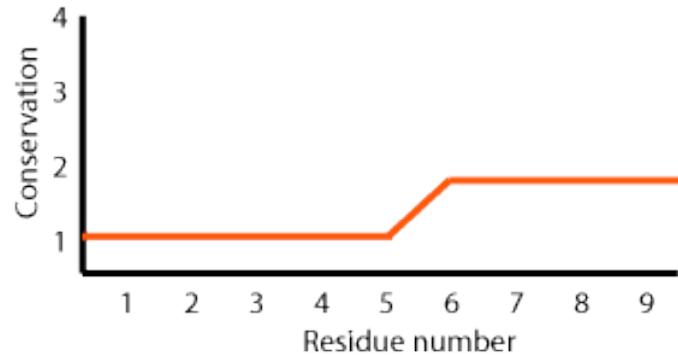
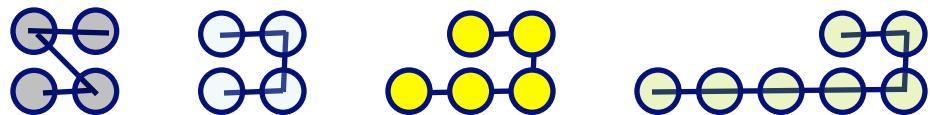
#	1	2	3	4	5	6	7	8	9
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0



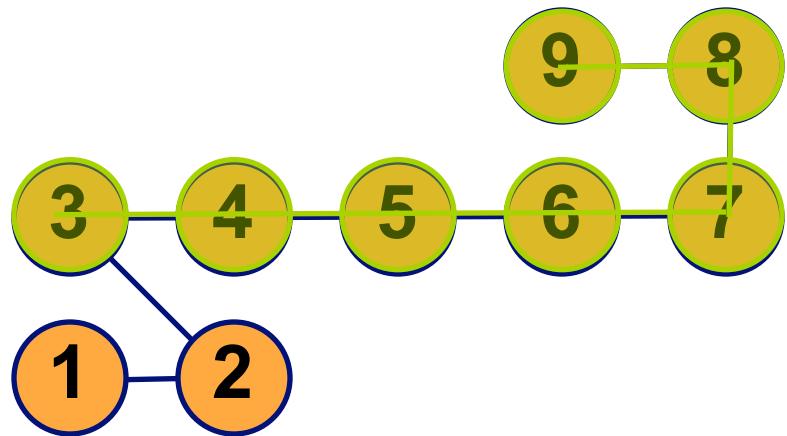
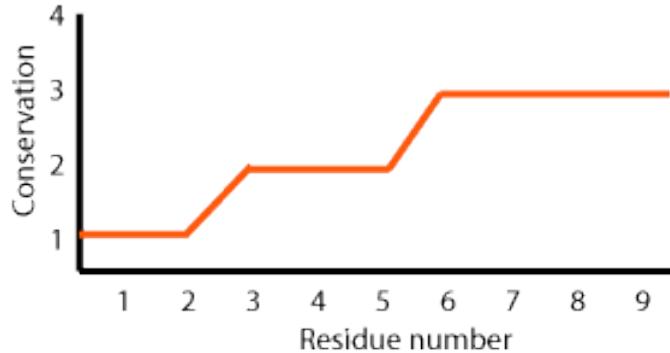
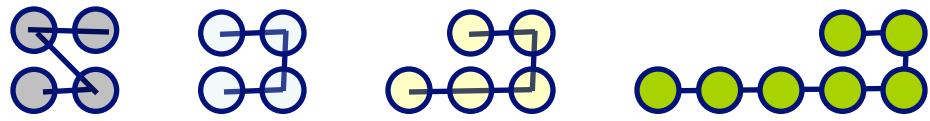
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0



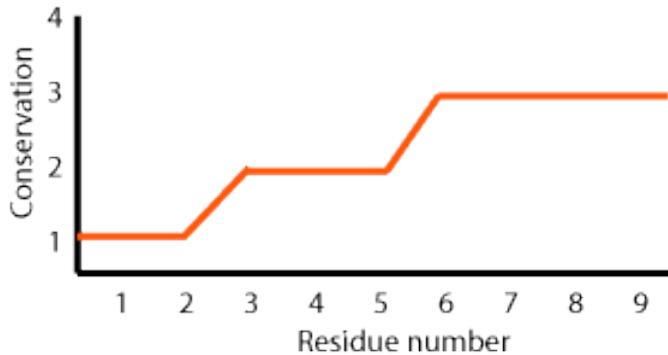
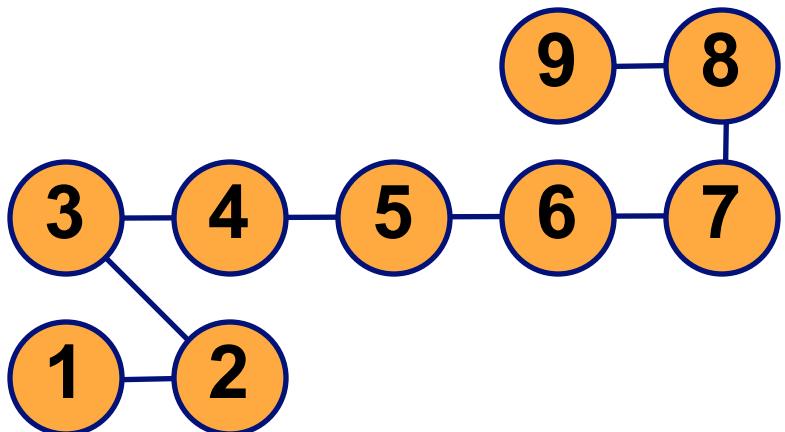
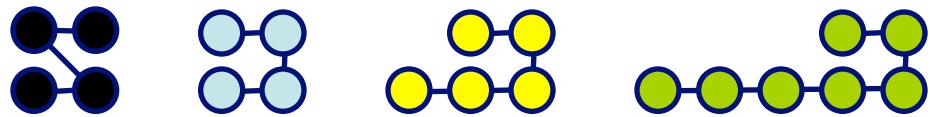
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	1	1	1
7	0	0	0	0	0	1	1	1	1
8	0	0	0	0	0	1	1	1	1
9	0	0	0	0	0	1	1	1	1



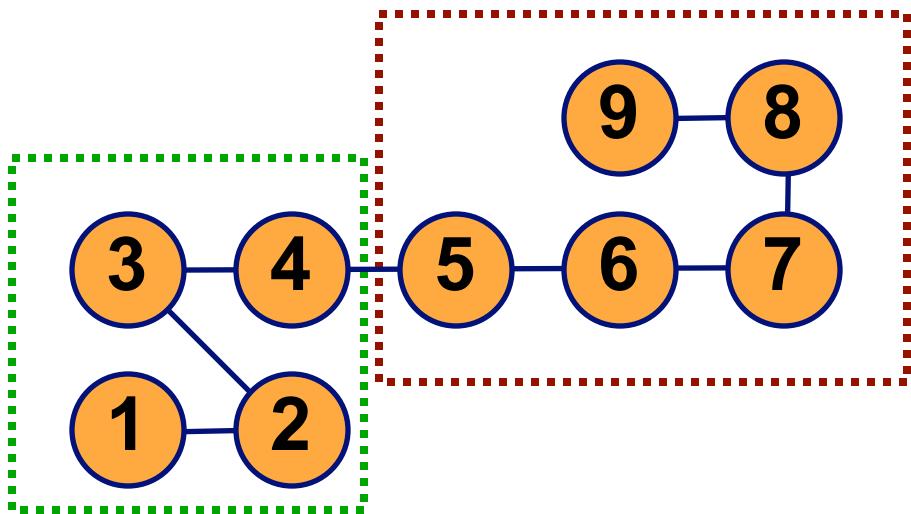
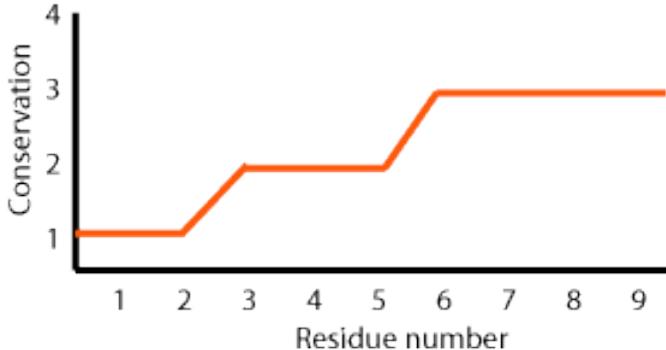
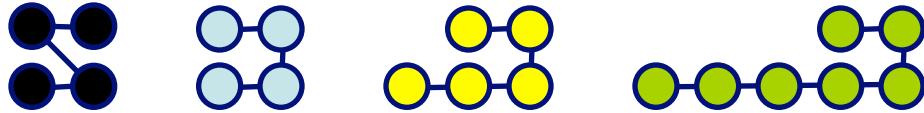
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	1	1	1	1	1
6	0	0	0	0	1	2	2	2	2
7	0	0	0	0	1	2	2	2	2
8	0	0	0	0	1	2	2	2	2
9	0	0	0	0	1	2	2	2	2



#	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 (-l)

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

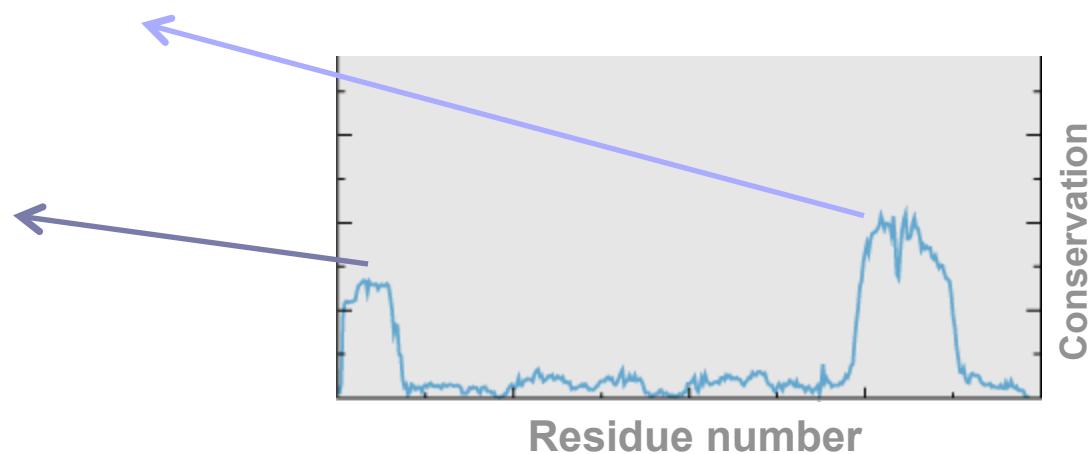
Thresholds #1,2 → MAMMOTH P-Value (Lp, Up)
High P-values → fewer partitions

Threshold #3 → Cluster Level (-l)
Low -l cluster value → fewer partitions

Applied to the ~45,000 chains in PDB (Dec 2003)

1phh	290-329	2.7Å	3.1
1hadB	72-111		

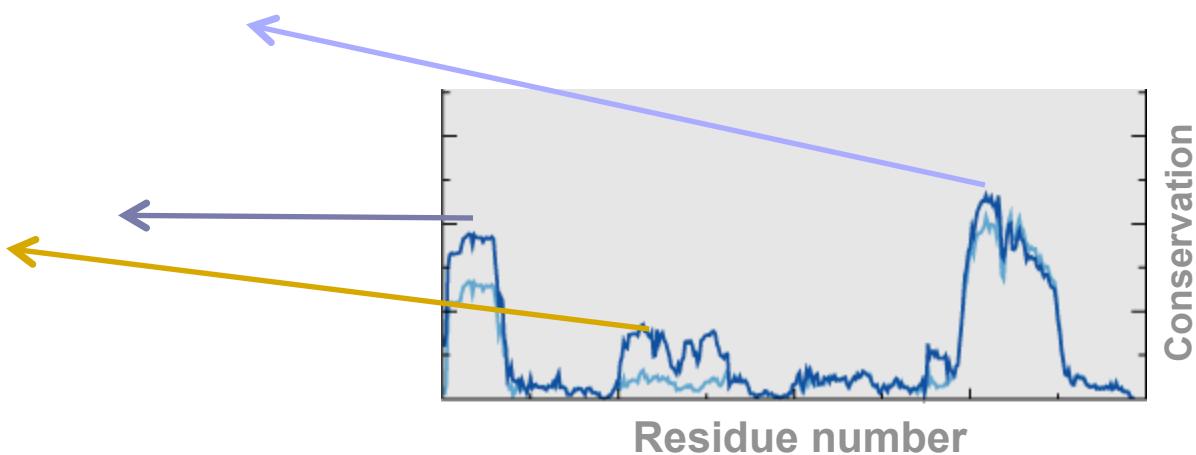
1phh	279-373	3.9Å	4.7
1bke	310-410		



1phh (Oxydoreductase from *Pseudomonas fluorescens*)

1phh	1-213	3.0 Å	8.1
1qjda	125-379		

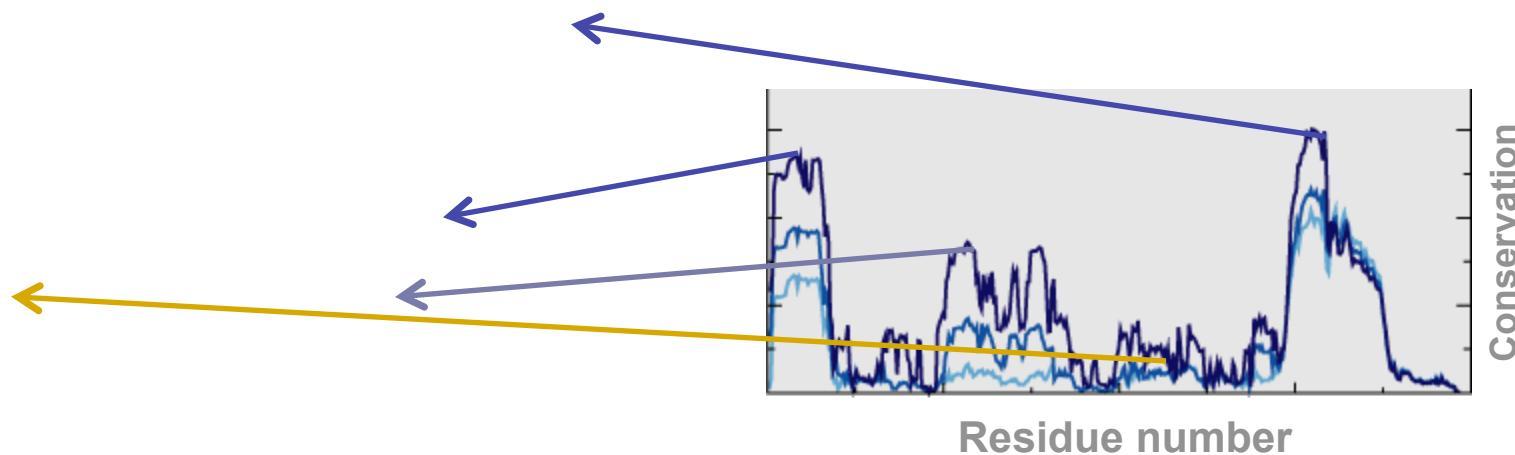
1phh	1-319	3.6 Å	9.8
1gerA	3-327		



1phh (Oxydoreductase from *Pseudomonas fluorescens*)

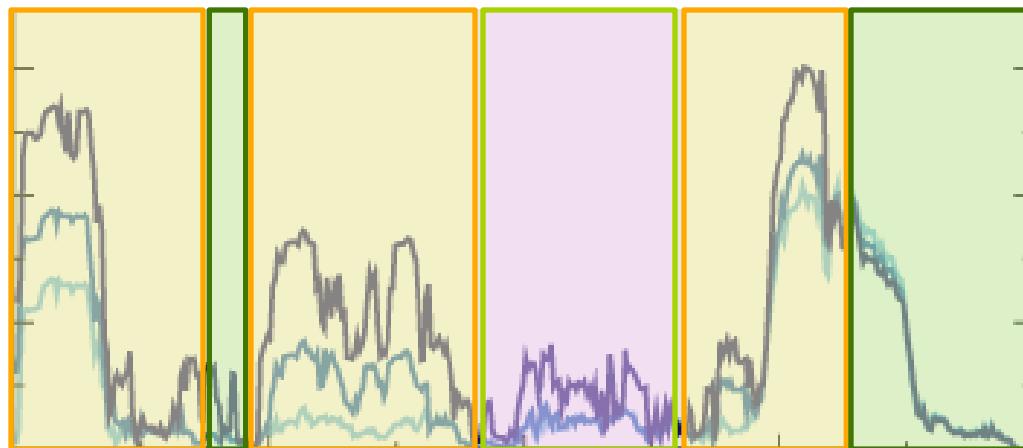
1phh	1-378	3.8Å	10.3
1feaC	2-464		

1phh	1-316	3.8Å	17.2
1l9dB	2-364		



1phh (Oxydoreductase from *Pseudomonas fluorescens*)

1phh (Oxydoreductase from *Pseudomonas fluorescens*)



Domain assignment from structure

**2163 chains from Islam *et al.* 1995 → 569 Non-redundant
 $<2\text{\AA}$ && $<30\text{aa}$ diff.**

**Divide randomly into two sets
Remove of incomplete or obsolete entries.**

FINAL:

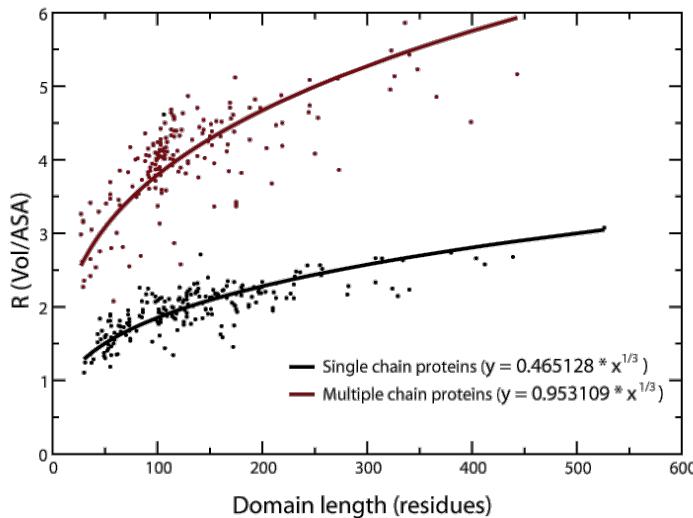
Training set → 242 chains

Testing set → 234 chains

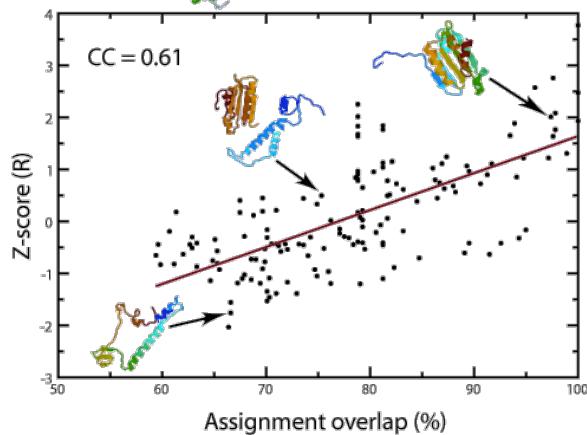
Thresholds #1,2 → MAMMOTH P-Value (Lp, Up)
High P-values → fewer partitions

Threshold #3 → Cluster Level (-l)
Low -l cluster value → fewer partitions

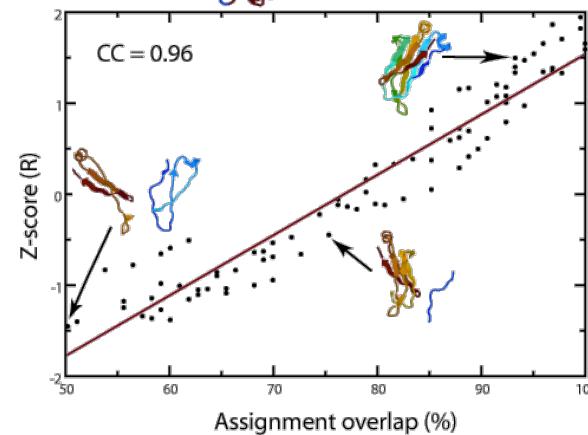
R = Volume/ASA



1lv1
3 domains protein



8fabA
2 domains protein



Domain → $\max(\langle \text{dist } f(R) \rangle)$

$\langle \text{dist to } f(R) \rangle$

-0.11

5-46



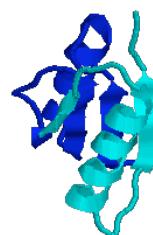
-0.10

47-84



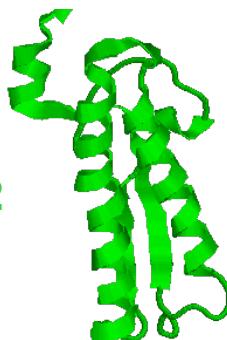
-0.08

1-84

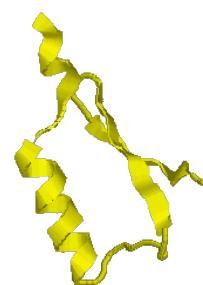


-0.09

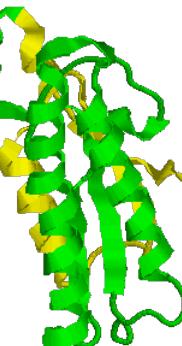
85-192



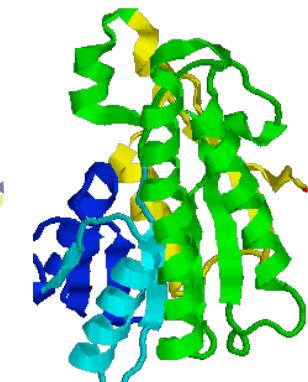
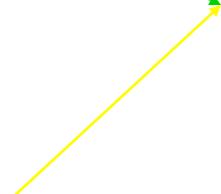
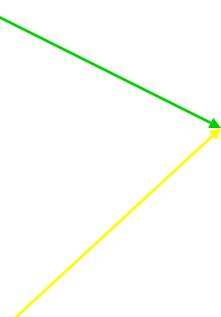
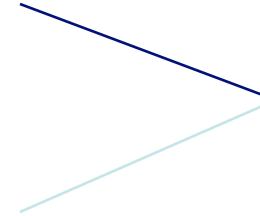
193-239



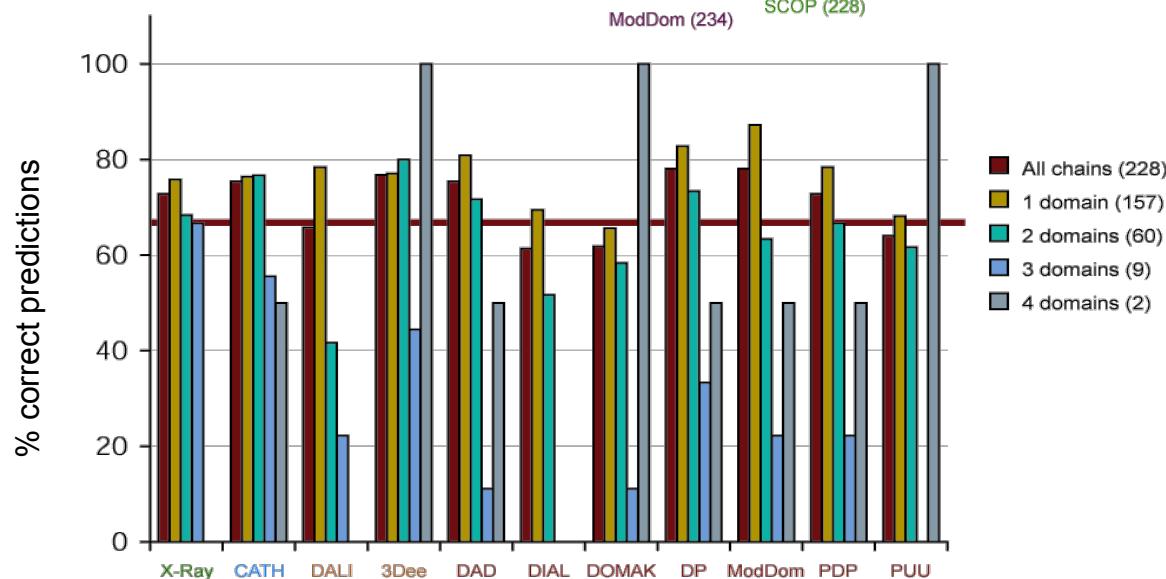
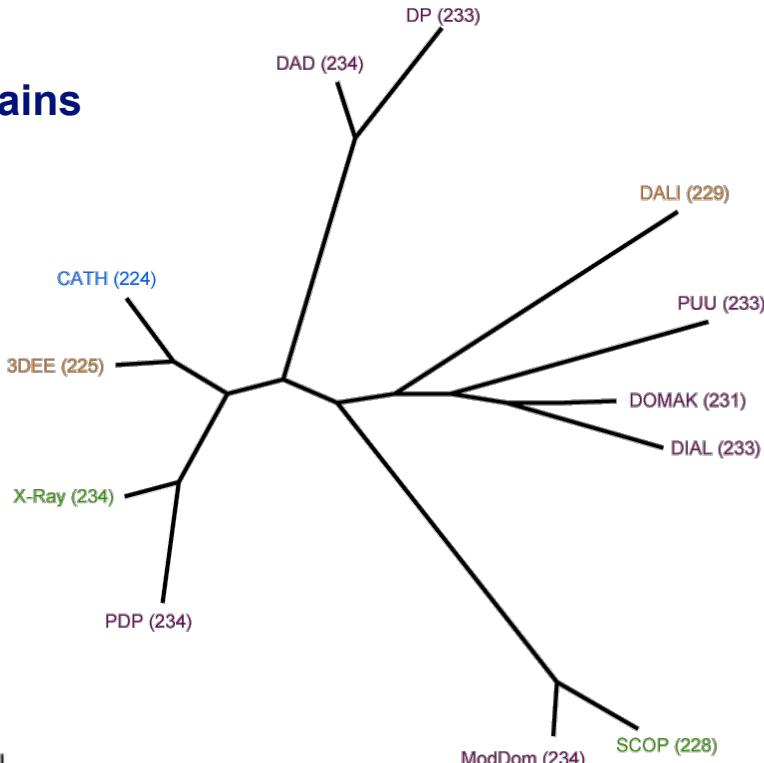
85-239



1dhr_ (dihydropteridine reductase)

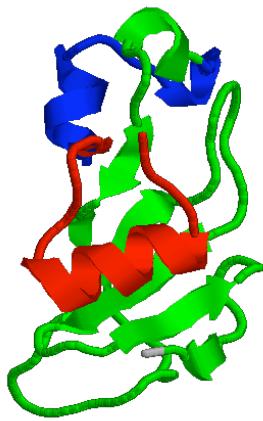


Non-redundant 234 chains

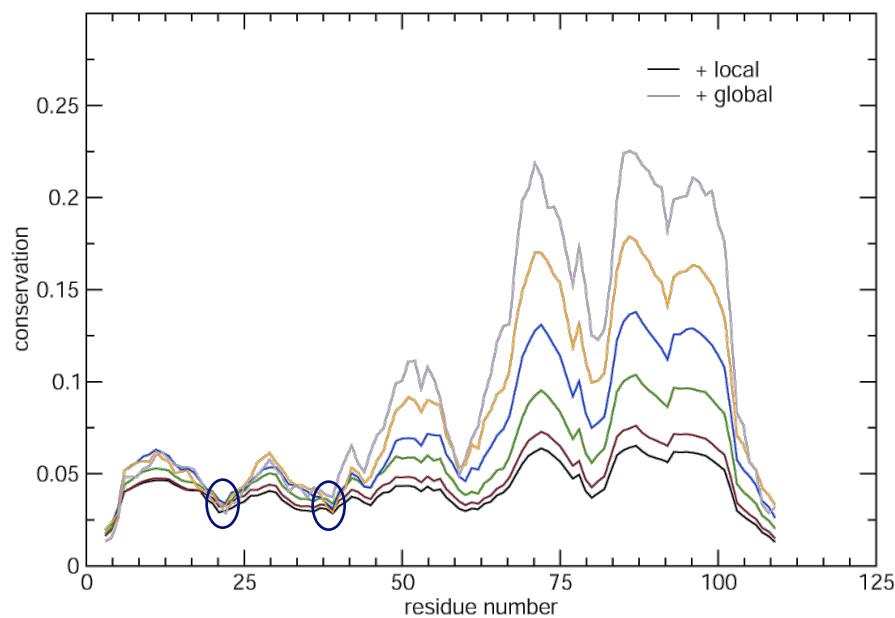


Fragments assignment from structure

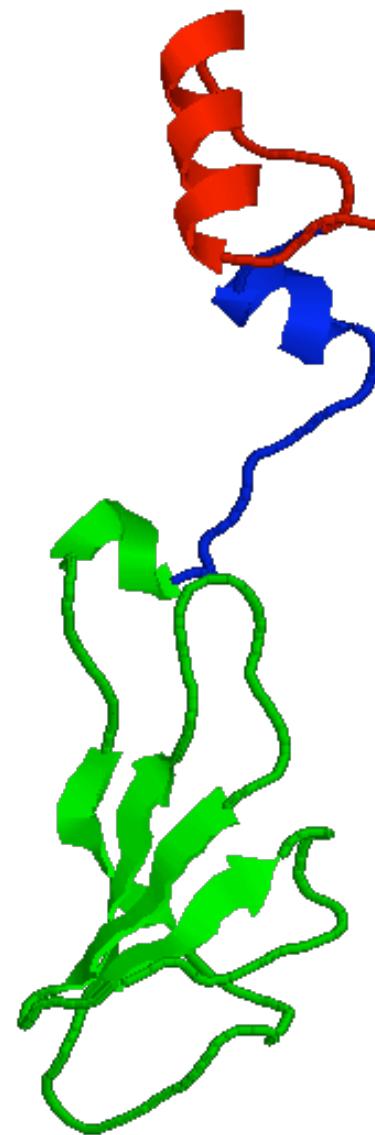
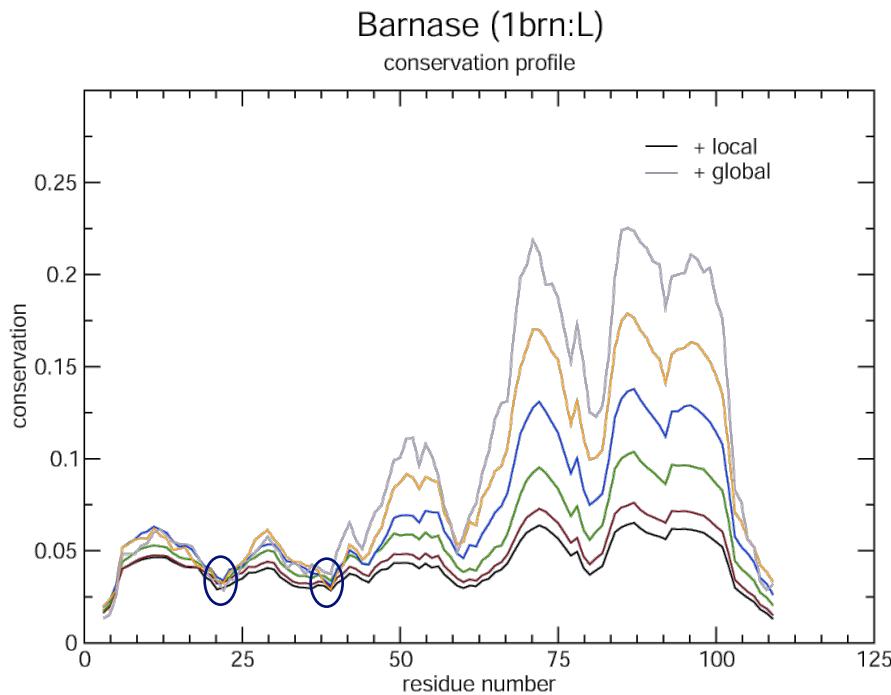
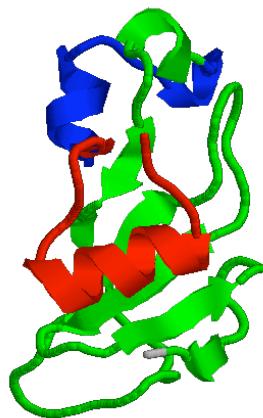
Barnase Domain-Swapping



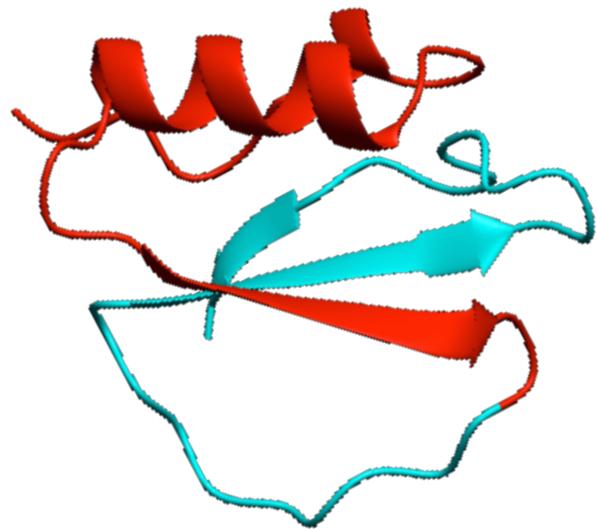
Barnase (1brn:L)
conservation profile



Barnase Domain-Swapping

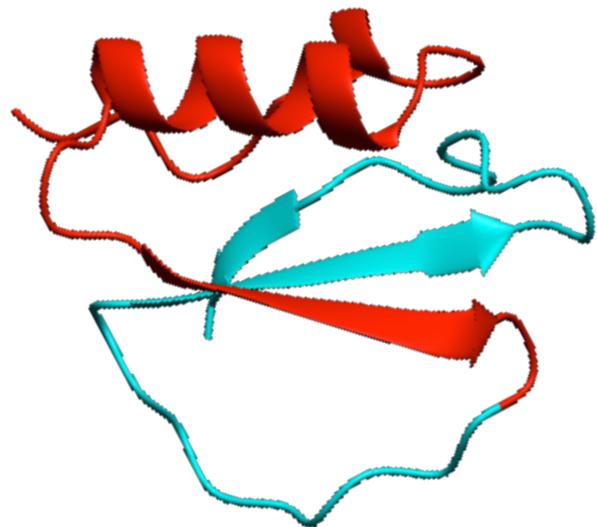


chymotrypsin inhibitor 2

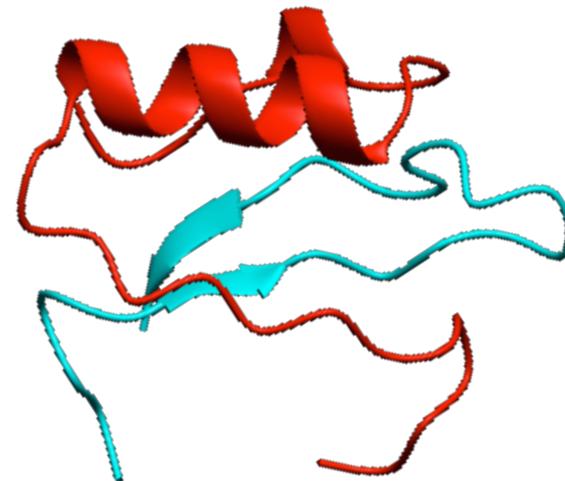


1-37 | 38-64

chymotrypsin inhibitor 2



1-37 | 38-64



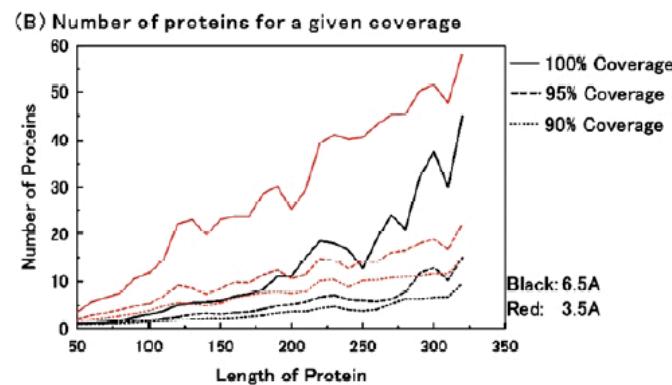
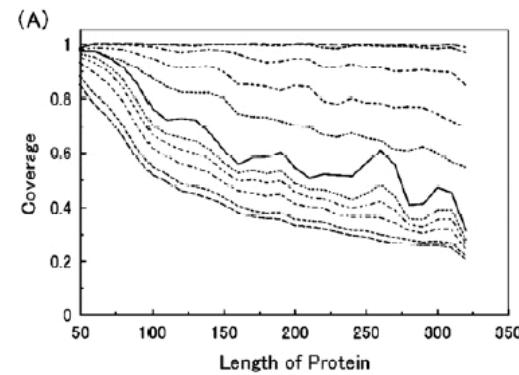
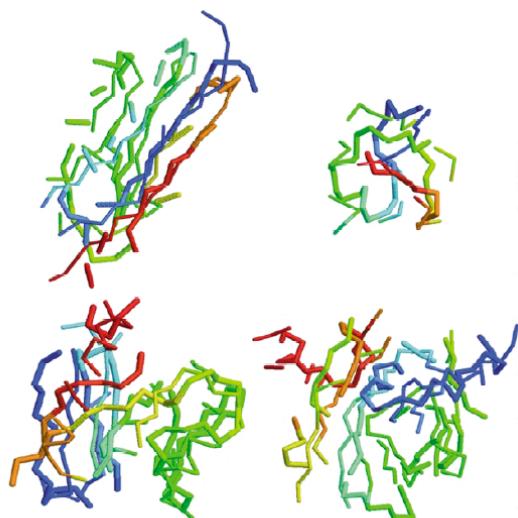
1-40 | 41-64

- Neira JL, Davis B, Ladurner AG, Buckle AM, Gay GP, Fersht AR. 1996. Towards the complete structural characterization of a protein folding pathway: the structures of the denatured, transition and native states for the association/folding of two complementary fragments of cleaved chymotrypsin inhibitor 2. Direct evidence for a nucleation-condensation mechanism. *Fold Des* 1:189-208.

- Ladurner AG, Itzhaki LS, de Prat GG, Fersht AR. 1997. Complementation of peptide fragments of the single domain protein chymotrypsin inhibitor 2. *J Mol Biol* 273:317-329.

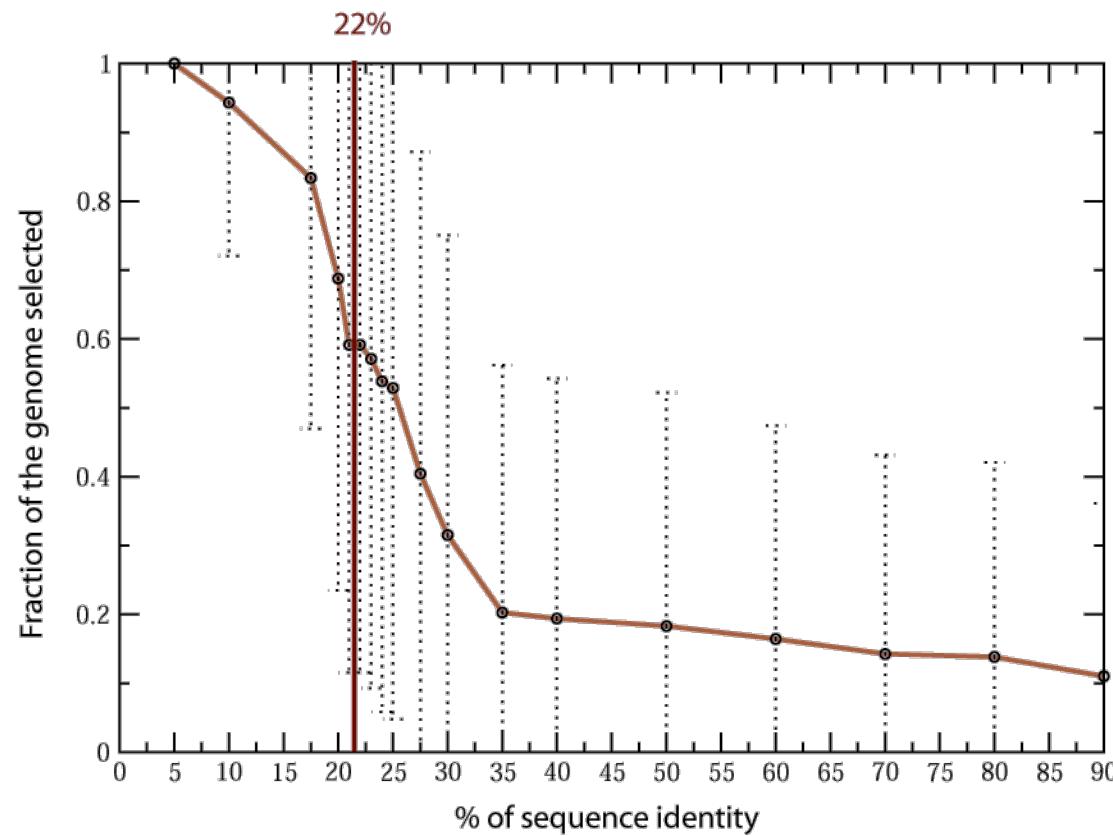
Sequence space .vs. Structure space

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



Data from DBAliv2.0 with maximum search space of 2000 chains

Sequence space .vs. Structure space



Template Search Methods

● Sequence similarity searches

- MODELLER <http://www.salilab.org/modeller/>
- BLAST <http://www.ncbi.nlm.nih.gov/BLAST/>
- FastA <http://www.ebi.ac.uk/fasta33/>

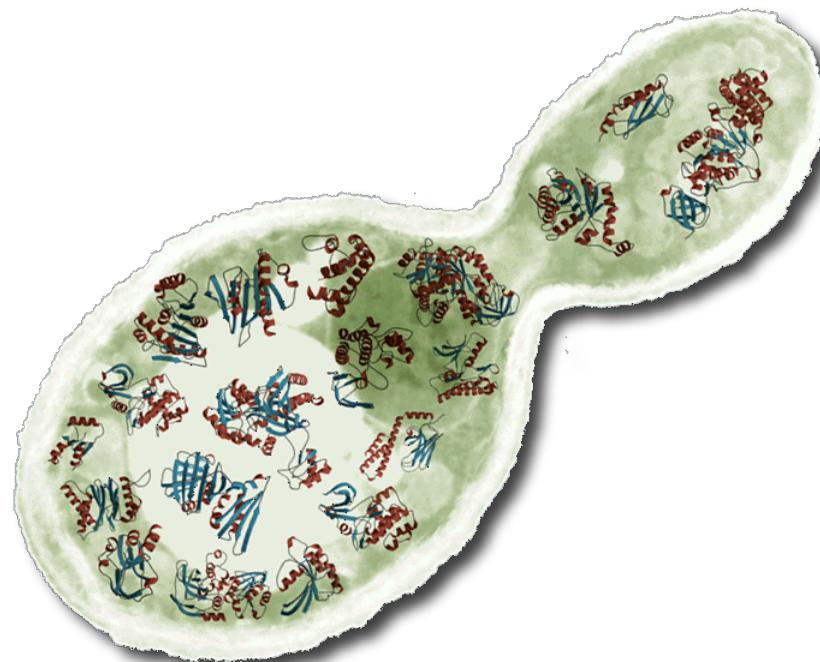
● Sequence profile and iterative methods

- HMMs <http://www.cse.ucsc.edu/research/compbio/HMM-apps/>
- PSI-BLAST <http://www.ncbi.nlm.nih.gov/BLAST/>

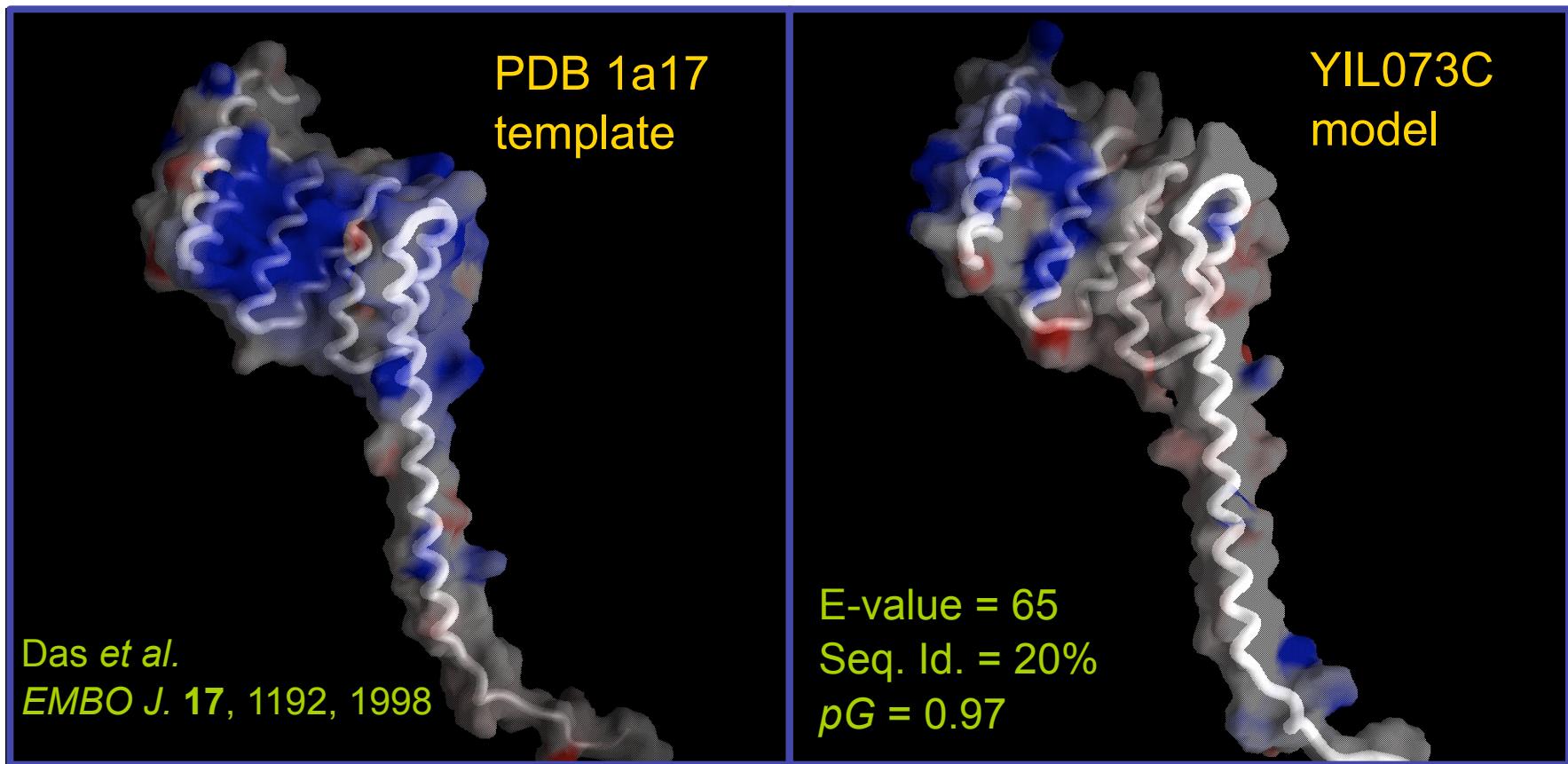
● Structure based threading

- mGenTHREADER <http://bioinf.cs.ucl.ac.uk/psipred/>
- PROFIT <http://www.came.sbg.ac.at/>

Fold assignment from sequence examples....



MODPIPE Model of Yeast Hypothetical Protein YIL073C (high e-value and good model score)

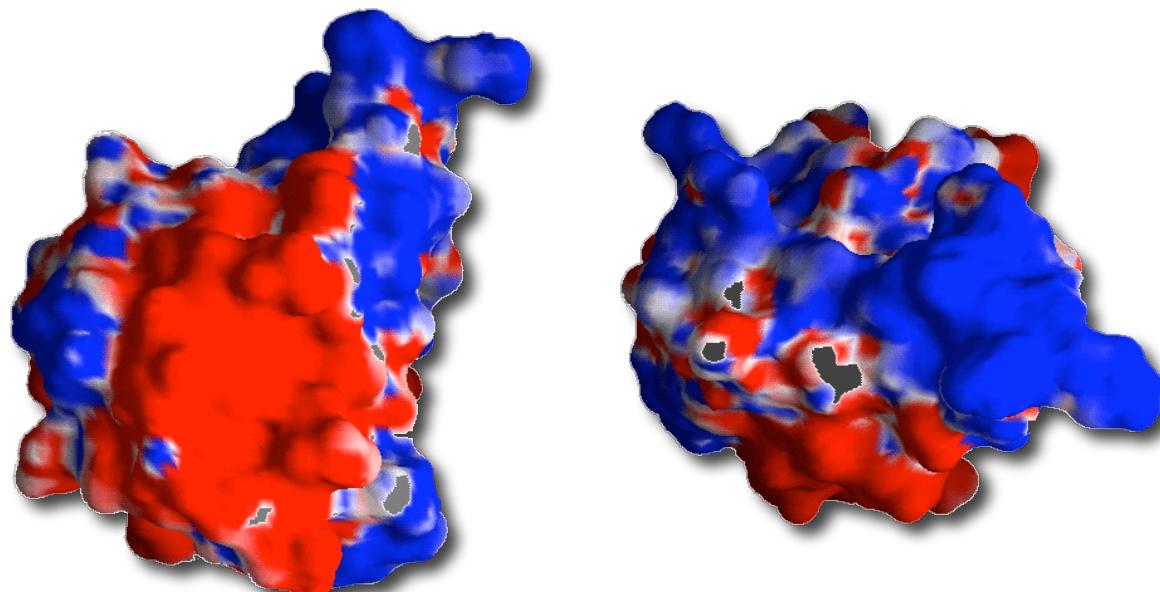
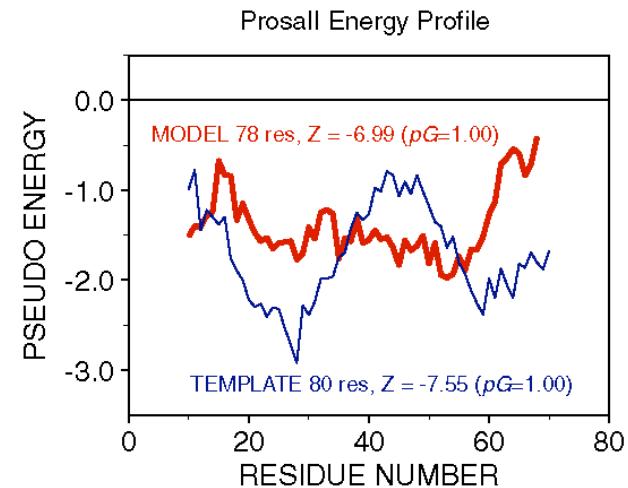


The tetratricopeptide repeat (TPR) is a degenerate 34 aa sequence identified in a variety of proteins, present in tandem arrays, mediates protein-protein interactions.

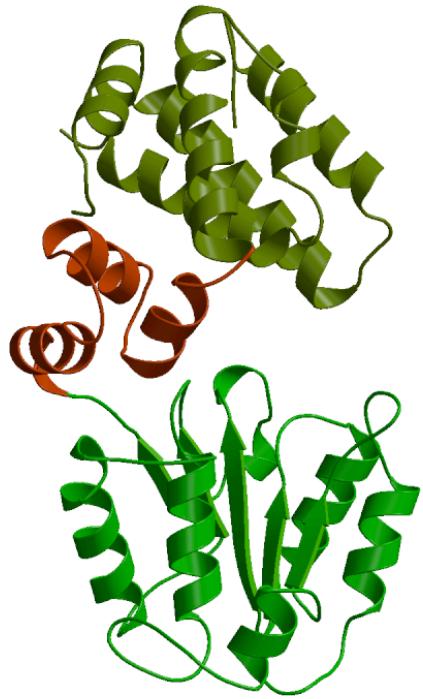
GRAP: Fold Assignment by PSI-BLAST + Model Evaluation

(significant e-value and good model score)

SEG FILTER	QUERY	MODEL SIZE	E-VALUE	pG
Y	Target	58	0.200	0.99
N	Target	64	0.029	1.00
Y	Template	NO HIT	NO HIT	NO HIT
N	Template	78	6×10^{-14}	1.00



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



Ec d' MRWYPWLRLPDEKLVASYQAGRGGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRG

RUVB LEEYVGQPQVRSQMEIFIFIKAAKLRGDALDHLLIFGPPGLGKTTLANIVANEMG-----

Ec d' CQLMQAGTHPDYYTLAPEKGKATLGVDREVTEKLNEAARLGGAKVVWVTDAALLTDAAANALLKTL

RUVB -----VNLRTT-----SGPVLEKAGDLAAMLTNLEPHDVLFIDEIHRLSPVVEEVLYPAM

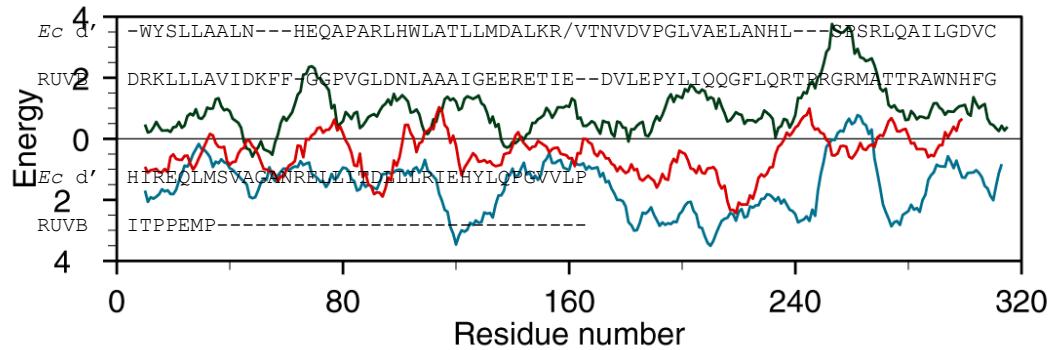
Ec d' -----EEPPAETWFFLATREPERL---LATLRSRCRLHYLAPPPEQYAVTWLSRE

Pdpd EDYQLDIMIGEGPAARSIKIDLPPFTLIGATTRAGSLTSPLRDRFGIVQRLEFY--QVPDLQYIVSRS

Ec d' VTM-----SQDALIAALRLSAGSPGAALALFQ-----GDNWQARETLCQALAYSVP SGD--

RUVB ARFMGLEMSDDGALEVARRARGTPRIANRLRRVRDFAEVKHDGTISADIAAQALDMNVDAEGFDYM

Energy profiles (Prossal by M. Sippl)



B. Guenther, R. Onrust, A. Šali, M. O'Donnell & J. Kuriyan. *Cell* **91**, 335, 1997.

Yamada, K., Kunishima, N., Mayanagi, K., Ohnishi, T., Nishino, T., Iwasaki, H., Shinagawa, H., Morikawa, K. Crystal Structure of the Holliday Junction Migration Motor Protein Ruvb from *Thermus Thermophilus* Hb8. *Proc.Nat.Acad.Sci.USA* **98**, 1442, 2001.

BMC WorkShop

Protein Structure Prediction

Sequence-Structure alignment (template selection)

Marc A. Marti-Renom & Damien Devos

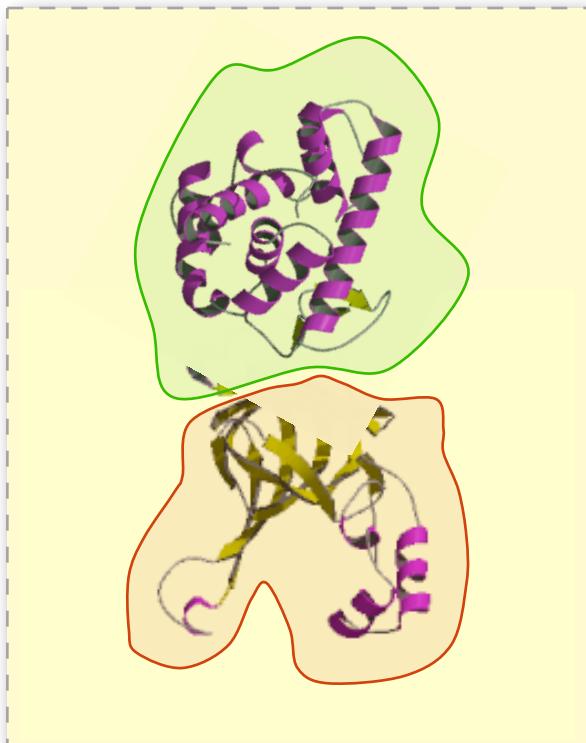
Department of Biopharmaceutical Sciences, UCSF

June 17th and 18th, 2004

Domains (?)

Domain boundaries from sequence

VERY DIFFICULT!!!!



MENFEIWVEKYRPRTLDEVVGQDEVIQRLKGYVERKNIPHLLFSGPPGTGKTATAIALARDLFGENWRDN
FIEMNASDERGIDVVRHKIKEFARTAPIGGAPFKIIFLDEADALTADAQAALRRTMEMYSKSCRFLSCN
YVSRIIEPIQSRCAVFRFKPVPKEAMKKRLLEICEKEGVKITEDGLEALIYISGGDFRKAINALQGAAAI
GEVVDADTIYQITATARPEEMTELIQTALKGNFMEARELLDRLMVEYGMSEDIVAQLFREIISMPIKDS
LKVQLIDKLGEVDTRLTEGANERIQLDAYLAYLSTLAKK

Domain boundaries from sequence (SnapDragon)

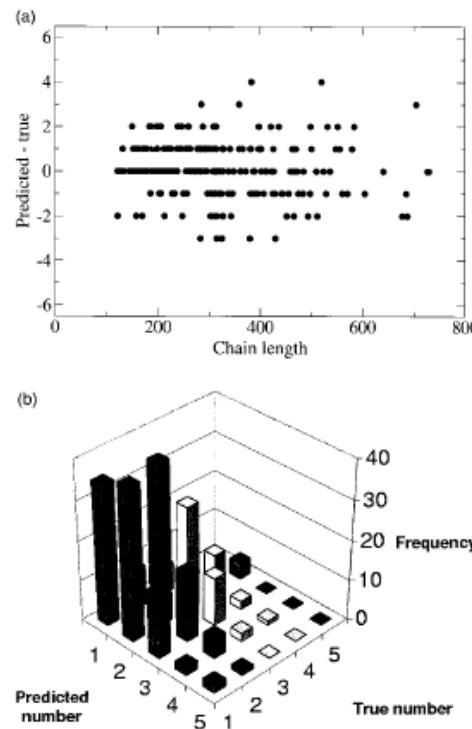
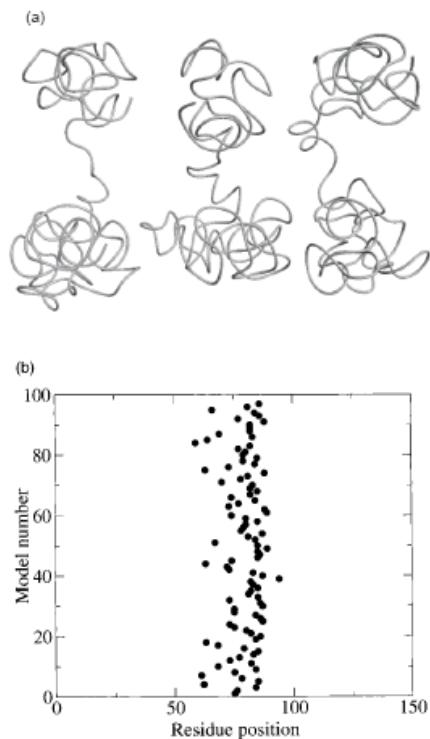
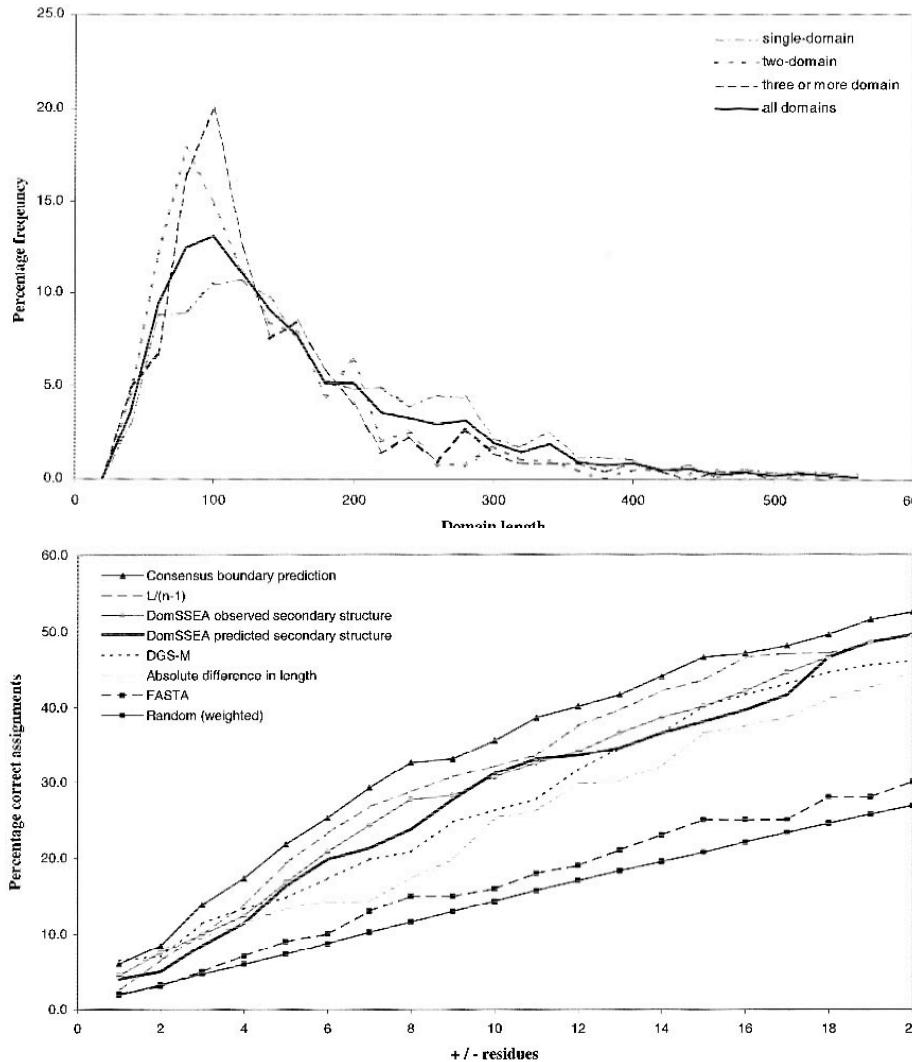


Table 2. Average accuracy percentages of linker prediction over 57 proteins

		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

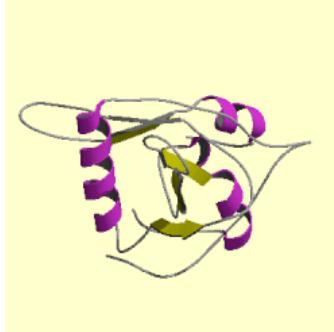
Domain boundaries from sequence (DomSSEA)



Prediction of Secondary Structure (PSI-PRED)

>gi42541361
MDIRSVSSLRGLLCLPPSWPRR

- Neural Network



- ✓ Very simple idea
- ✓ Simple scoring

Obscure optimizer

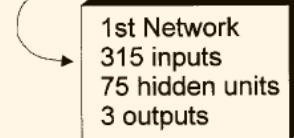
Position-based scoring matrix used																							
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V			
-2	-4	-4	-4	-3	-4	-4	-2	-1	-1	-4	-1	8	-5	-3	-3	0	2	-2					
0	-1	-1	3	-4	3	4	1	-1	-4	-4	0	-3	-4	-2	-1	-2	-4	-3	-3				
0	-1	2	1	-3	4	0	-1	-2	-4	-3	1	-2	-4	-2	2	0	-4	-3	-3				
-2	-3	-4	-5	-2	-3	-4	-6	-4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

Raw profile from PSI-BLAST Log File

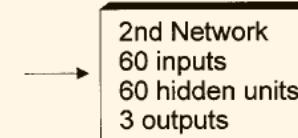
Window of 15 rows

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	0.4	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	0.1	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



Window of 15 x 3 outputs fed to 2nd network



Final 3-state Prediction

Prediction of Secondary Structure (PSI-PRED)

<http://bioinf.cs.ucl.ac.uk/psiform.html>

PSIPRED Protein Structure Prediction Server - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Stop Refresh Search Favorites Media Mail Links Stop

Address http://bioinf.cs.ucl.ac.uk/psiform.html Go Links Stop

Bioinformatics Unit

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 (MEMSAT)
 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 MEMSAT

Submit Sequence E-mail address [Help](#)
Password (only required for commercial e-mail addresses) [Help](#)
Short name for sequence [Help](#)
Predict Clear form

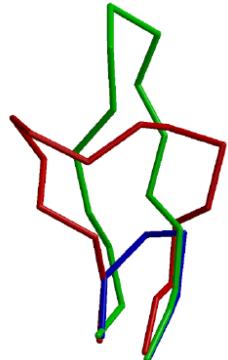
Internet

Why the alignment is so important?

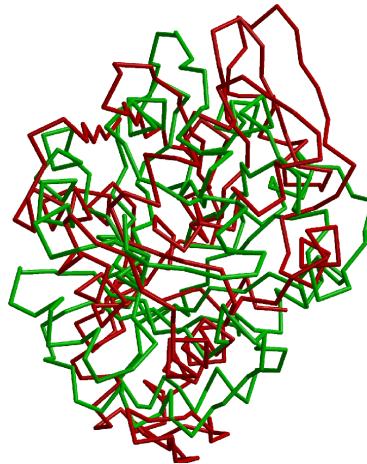
Typical errors in comparative models

MODEL
X-RAY
TEMPLATE

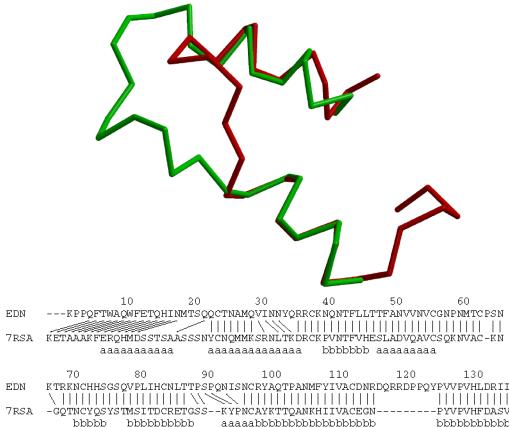
Region without a template



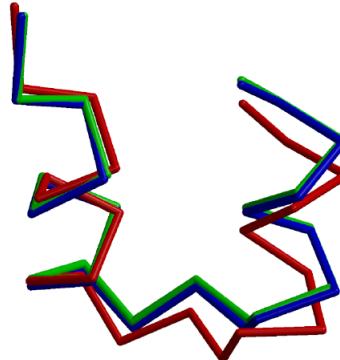
Incorrect template



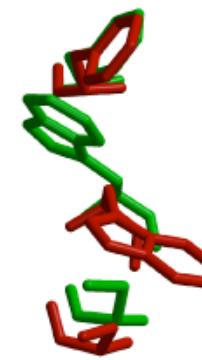
Misalignment



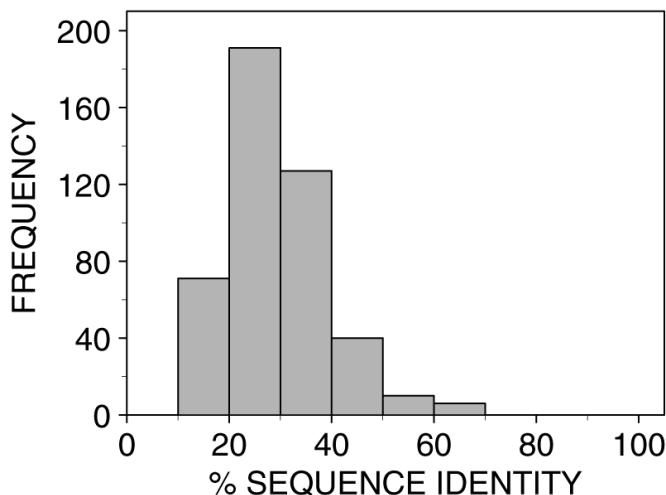
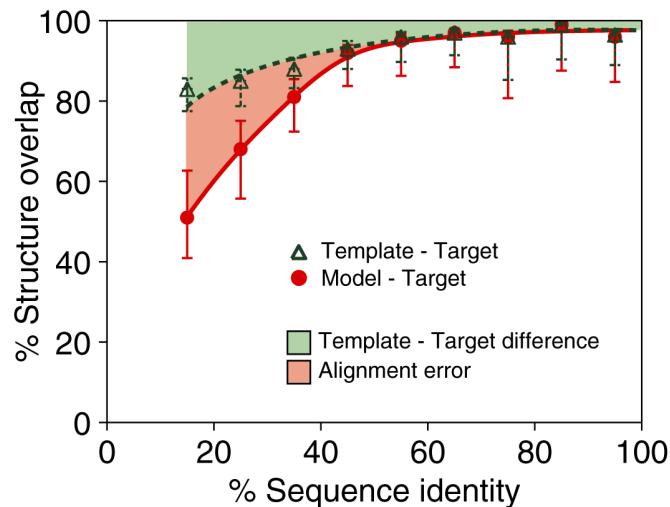
Distortion/shifts in aligned regions



Sidechain packing



Alignment errors are frequent and large



Minimizing errors in sequence-structure alignment

- Threading.
- Complex gap penalty functions.
- Multiple sequence profiles.
- Iterative process (model assessment)

Threading

General overview (Threading)

- Matches sequences to 3D structures
 - Requires a scoring function to assess the fit of a sequence to a given fold
 - Scoring functions derived from known structures and include atom contact and solvation terms evaluated in a pairwise fashion
 - May include secondary structure terms, multiple alignments...
- Threading servers available using several different approaches
 - Fold recognition server at Imperial College, UK
<http://www.sbg.bio.ic.ac.uk/~3dpssm/>
 - ProteinPredict server at EMBL
<http://www.embl-heidelberg.de/predictprotein/predictprotein.html>
 - Protein sequence-structure threading at NCBI
<http://www.ncbi.nlm.nih.gov/Structure/RESEARCH/threading.shtml>

Template comparison methods

- Uses 3D “templates” for searching structural databases
 - active site or binding site templates generated to reflect functionally important structural signatures
- Available software/servers
 - Template Search and Superposition (TESS), Thornton Group
<http://www.biochem.ucl.ac.uk/bsm/PROCAT/PROCAT.html>
Wallace AC; Borkakoti N; Thornton JM. (1997) Protein Science 6 pp2308
 - “Fuzzy Functional Forms”, Skolnick - commercial availability
Fetrow, JS and Skolnick, J (1998) J. Mo. Biol 281 pp949
 - Spatial Arrangements of Side-chain and Main-chain (SPASM), Kleywegt, Univ. of Uppsala
<http://portray.bmc.uu.se/cgi-bin/dennis/spasm.pl>
Kleywegt GJ (1999). J. Mol. Biol. 285 pp1887

Sequence-Structure alignments

As any other bioinformatics problem...

- Representation
- Scoring
- Optimizer

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
- Simple

Types:

- Contact potential
- Distance potentials
- Surface potentials

Approximations/Limitations in PMFs

Database size.

PMF versus Energy (additive/higher order terms).

Reference state.

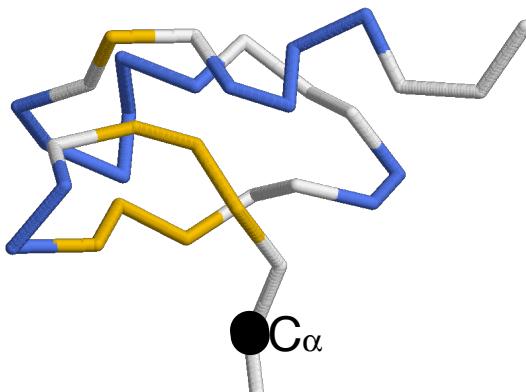
Physical origin.

Representation

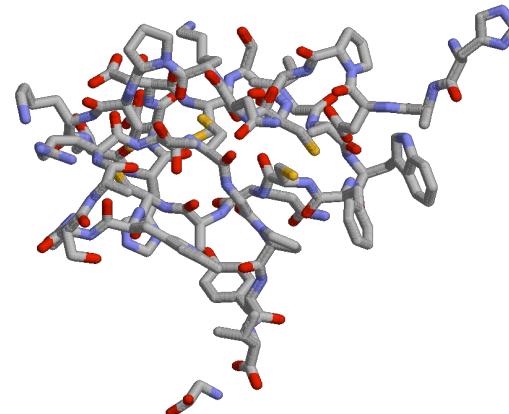
Sequence/Structures

>gi42541361
MDIRSVSSLRGLLCLPPSWPRR

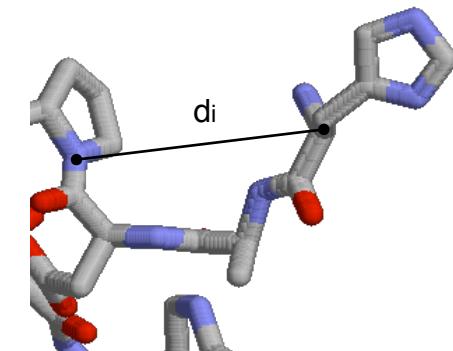
Primary sequence



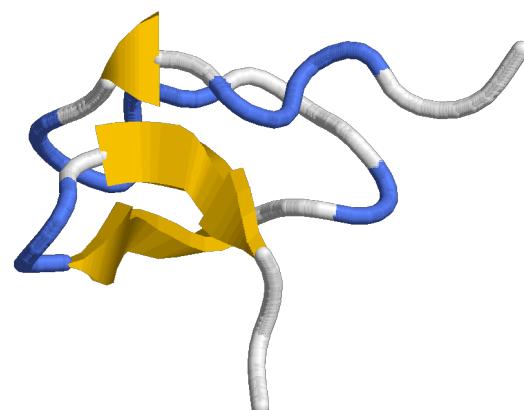
Reduced atoms representation



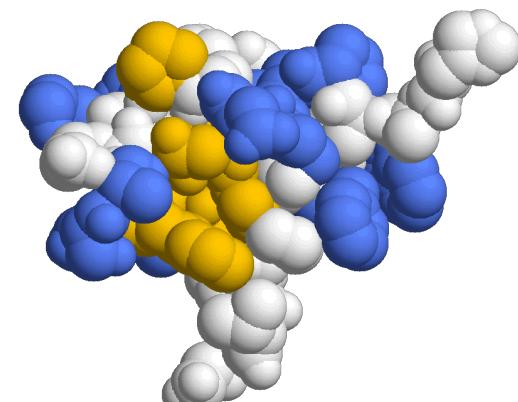
All atoms and coordinates



Distance space



Secondary Structure



Accessible surface

Scoring

Statistical Potential... inspiration

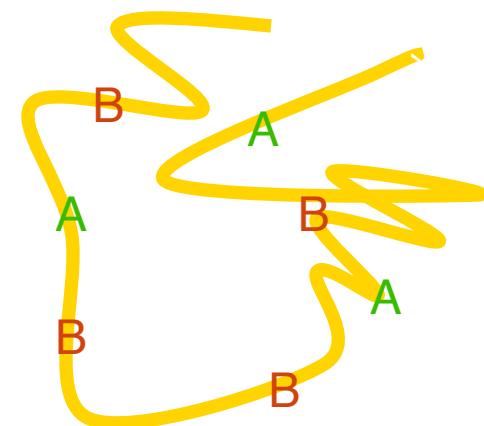
$$K = \frac{[AB]}{[A] \cdot [B]}$$

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

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.

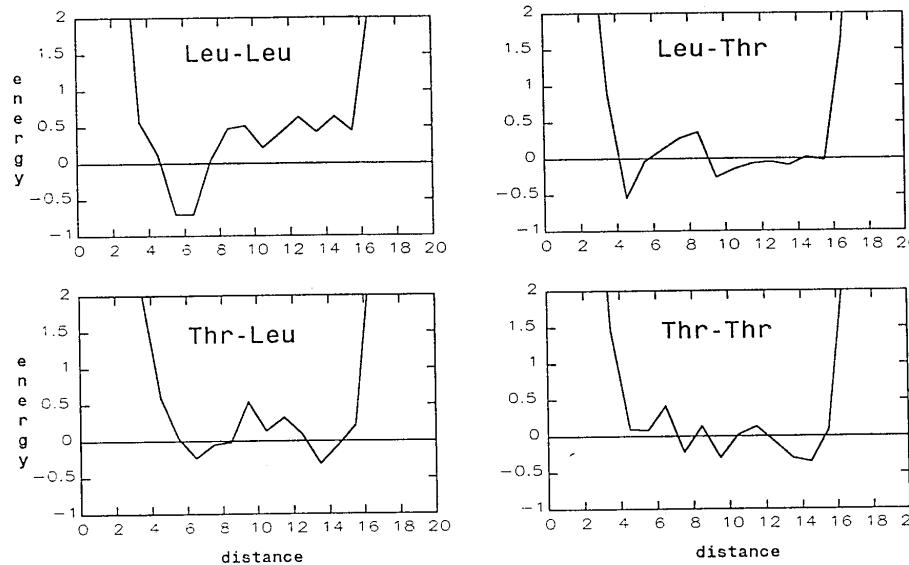


Tanaka and Sheraga (1975) PNAS, 72 pp3802
A. Godzik, (1996) Structure 15 pp363

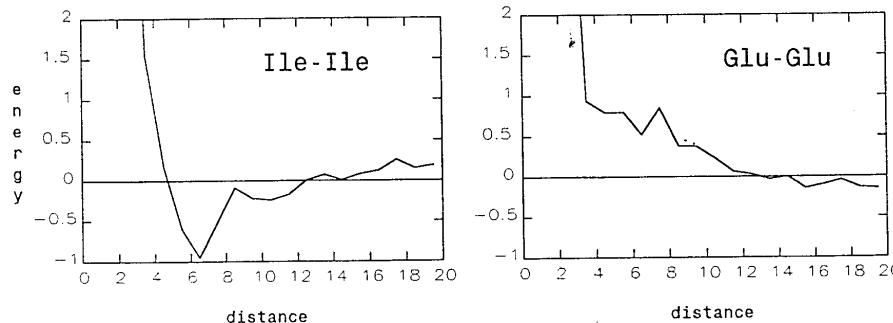
Scoring

Statistical Potential... Distance Potentials

Long range free energy



Short range free energy



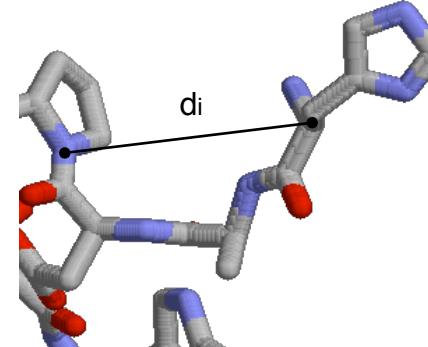
Scoring

Raw scores of an alignment

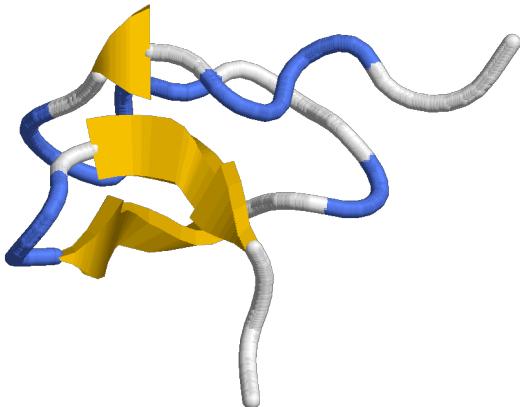
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2	
S	-1	4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-3		
T	-1	1	4	1	-1	1	0	1	0	0	-1	-1	0	-1	-2	-2	-2	-2	-3		
P	-3	-1	1	7	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4	
A	0	1	-1	-1	4	0	-1	-2	-1	-1	-2	-1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

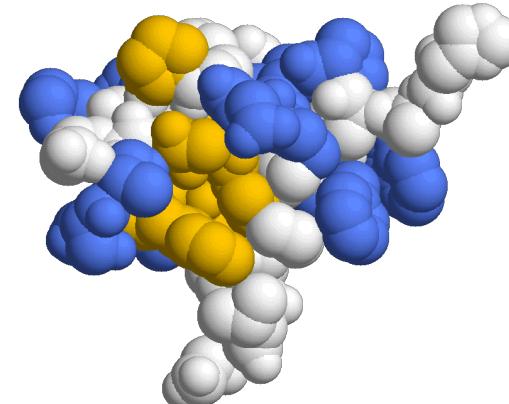
Aminoacid substitutions



Distance space



Secondary Structure (H,B,C)



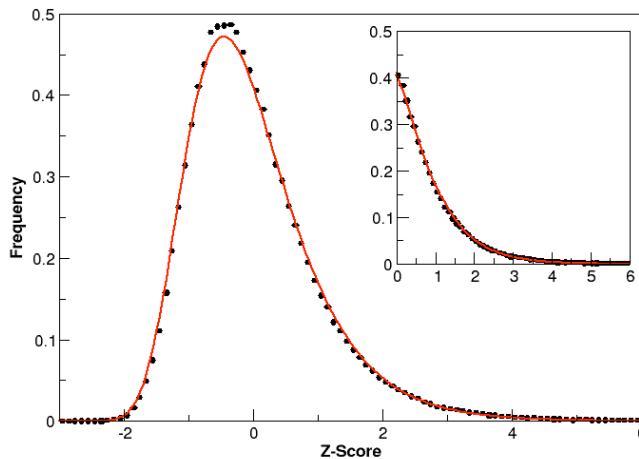
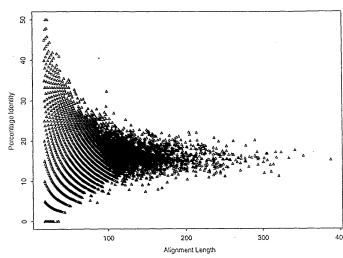
Accessible surface (B,A [%])

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

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

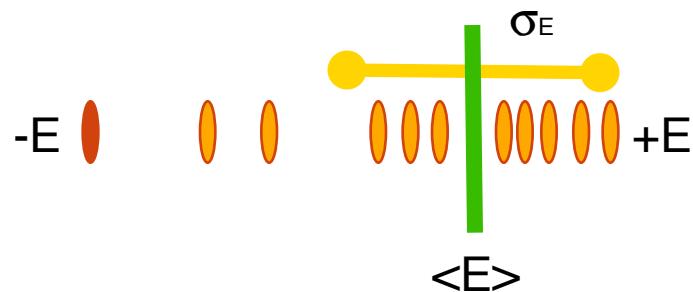
$$P(s \geq x) = 1 - \exp(e^{-\lambda (s-\mu)})$$

Karlin and Altschul, 1990 PNAS 87, pp2264

Scoring

Significance of an alignment (score)

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

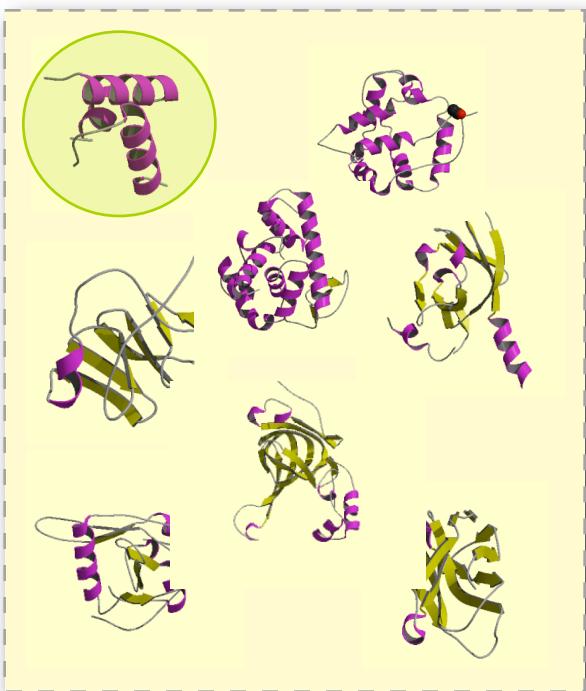


$$Zscore = \frac{(\langle E \rangle - E_m)}{\sigma_E}$$

Scoring

Significance of an alignment (background)

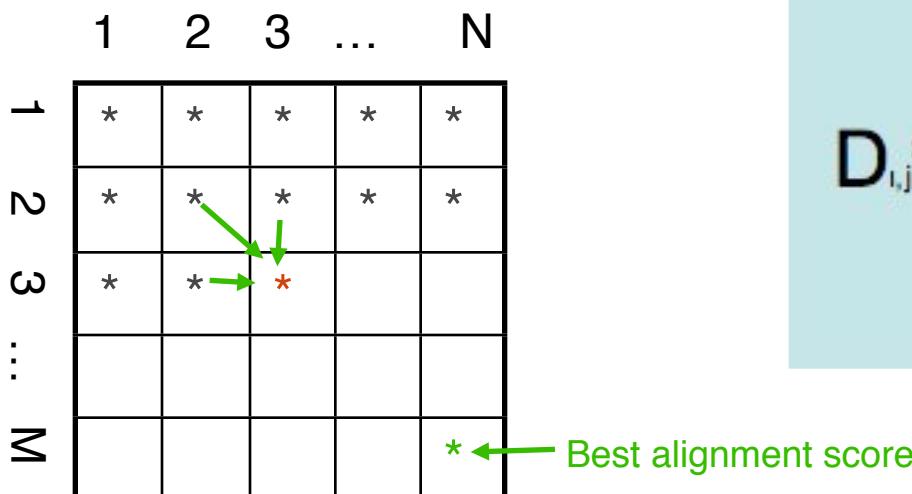
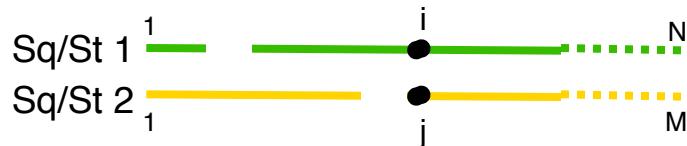
Structural space



Sequence space

MKLLIVLTCISLCSCICTVVQRCASNKPHVLEDPCKVQH
HLSVNQCVLLPQCCPKSCKICTHLISIEVVLTCRAVDKM
MHVNCVEQCSLQDCIKIAPRVLKTCILCVLKPCLTSH
VHLVQPTSCCCKNCICHVEIRSLDILTKSVQLACLVPM
⋮
MQCCRVQKICDLLAVALCKLHISTPSCKILCVVTSPHN

Global dynamic programming alignment

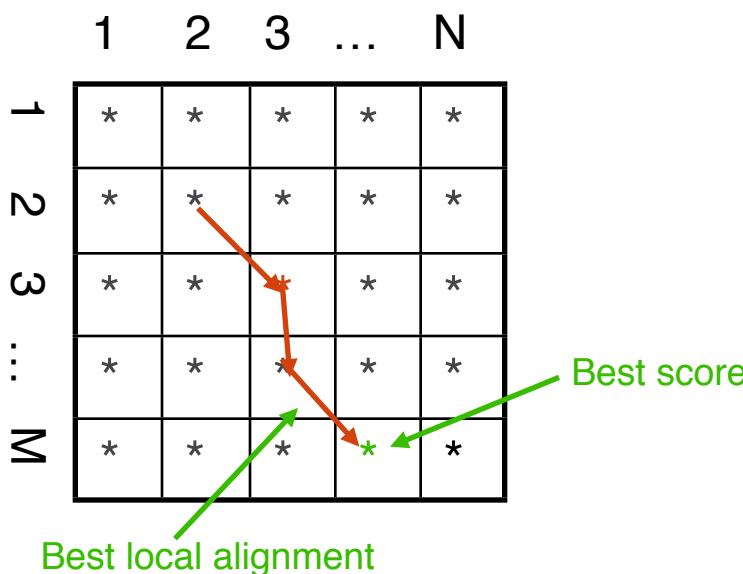
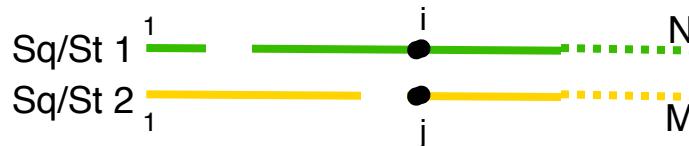


$$D_{i,j} = \min \begin{cases} D_{i,j-1} + \text{Score}_{(\Delta, rj)} \\ D_{i-1,j-1} + \text{Score}_{(ri, rj)} \\ D_{i-1,j} + \text{Score}_{(ri, \Delta)} \end{cases}$$

Backtracking to get the best alignment

Optimizer

Local dynamic programming alignment



$$D_{i,j} = \min \left\{ \begin{array}{l} D_{i,j-1} + \text{Score}_{(\Delta, rj)} \\ D_{i-1,j-1} + \text{Score}_{(ri, rj)} \\ D_{i-1,j} + \text{Score}_{(ri, \Delta)} \\ 0 \end{array} \right\}$$

Backtracking to get the best alignment

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

Applications of PMFs

Model assessment.

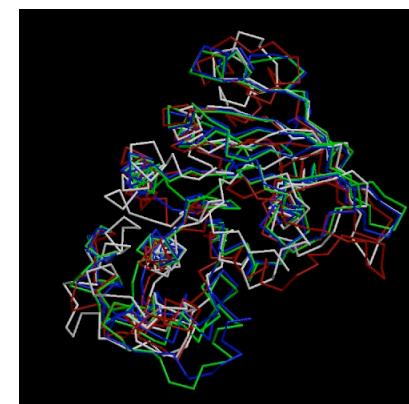
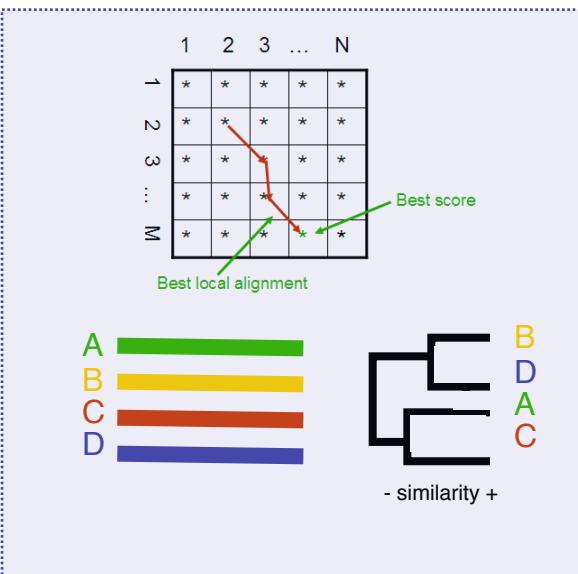
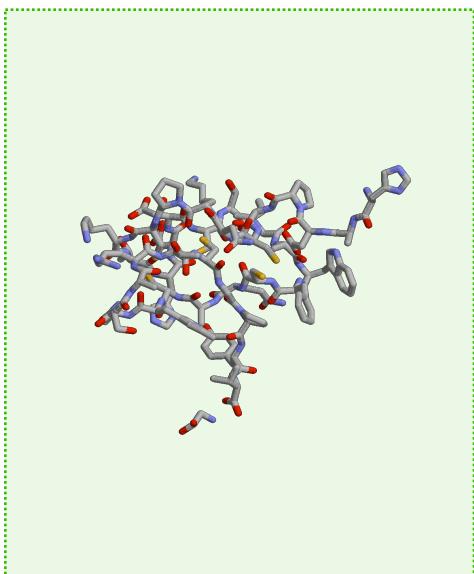
Ab initio folding simulations.

Sequence-structure matching (threading).

Comparative protein structure modeling (loops, sidechains, ...).

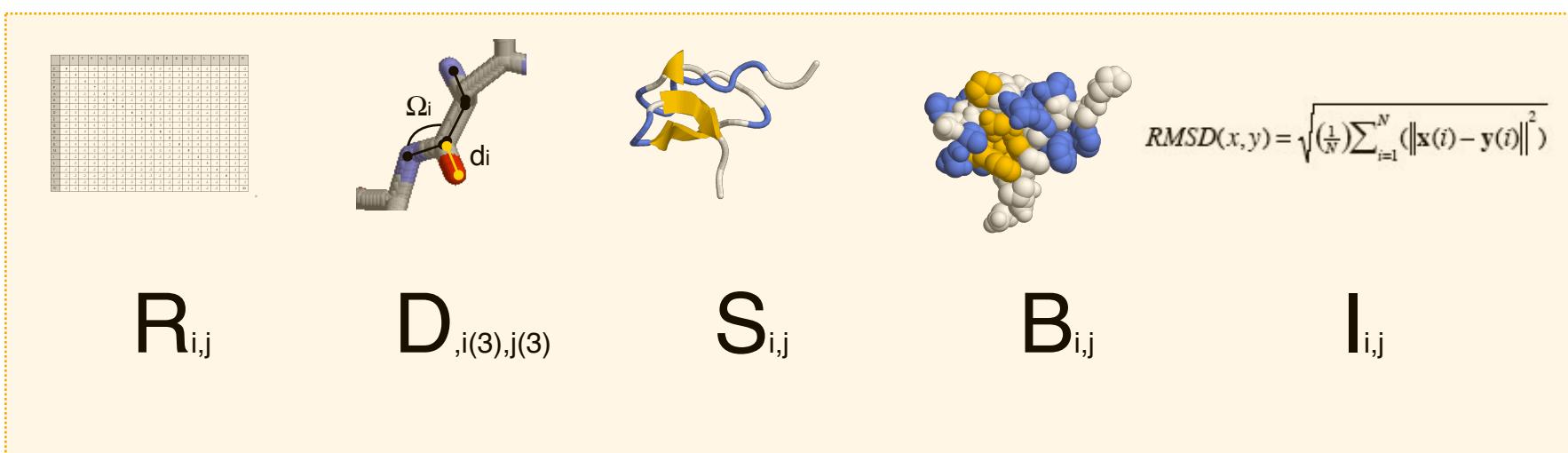
Secondary structure prediction, etc.

Sequence-Structural alignment by properties conservation (SALIGN-MODELLER)



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

Computationally expensive

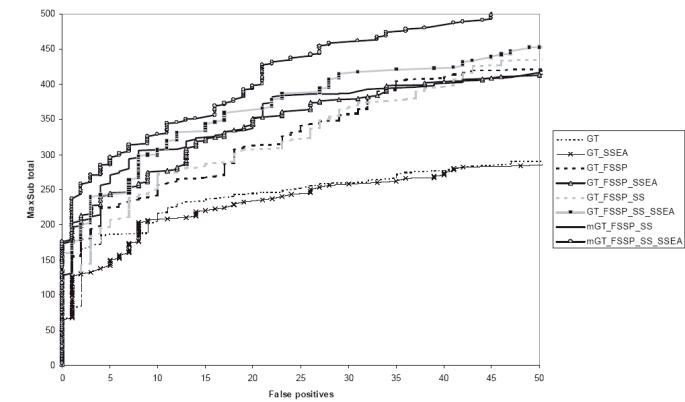
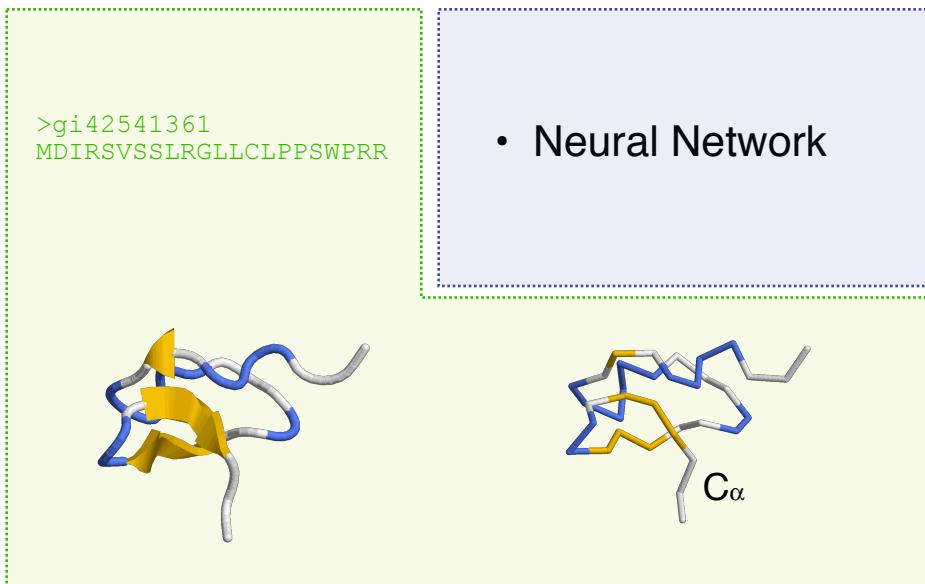


Sequence-Structural alignment by properties conservation (SALIGN-MODELLER)

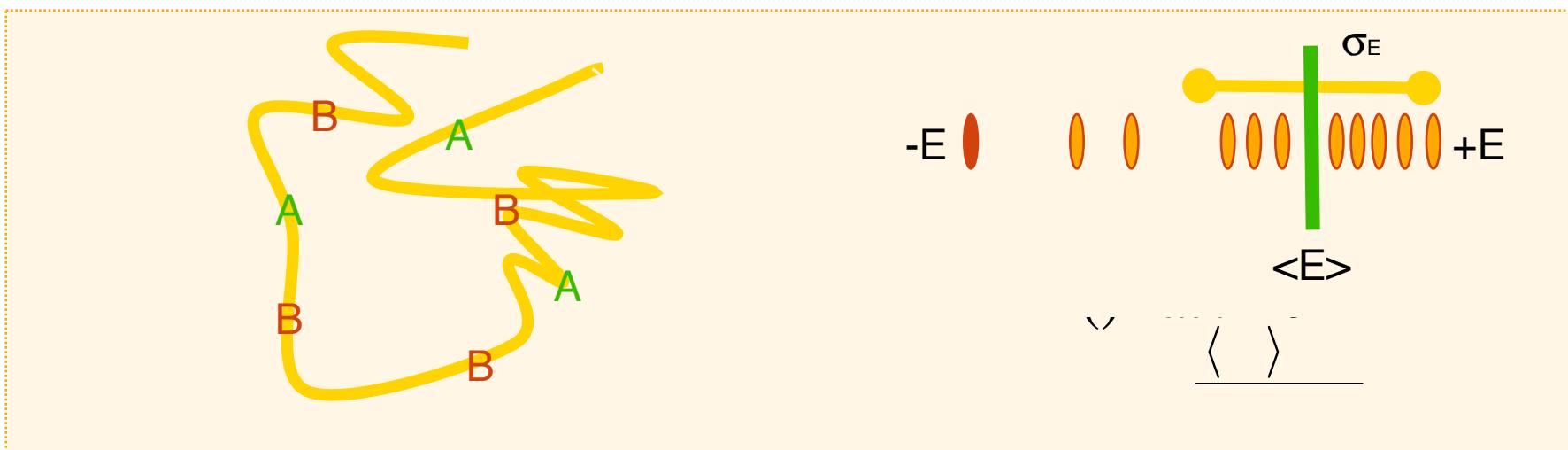
<http://www.salilab.org/dbali/>

The screenshot shows a Microsoft Internet Explorer window displaying the DBAli v2.0 tools page. The title bar reads "DBAli v2.0 tools page - Microsoft Internet Explorer". The address bar shows the URL "http://salilab.org/DBAli/?page=tools&action=f_salign". The page header includes the "UCSF | Sali Lab | MAMMOTH" logo and the DBAli v2.0 logo. A banner at the top features a red ribbon-like protein structure. On the left, a sidebar menu lists "Home", "Search DBAli", "Tools", and "Help". A yellow box titled "DBAli ALERT!" contains a message about visiting the updated DBAli database. The main content area is titled "DBAli. Tools associated to the database." and lists several tools: "Cluster a list of chains", "Cluster from a chain", "Define domains from a chain", "Get a multiple structure alignment of a list of chains", "Database statistics", and "Download DBAli". Below this is a section titled "Get a multiple structure alignment of a list of chains." with a form. The form has a label "File with a list of chains:" followed by a file input field, a "Browse..." button, and a help icon. At the bottom of the form are "SALIGN" and "Clear" buttons. The footer contains links for "Reference", "Download", "Statistics", "Suggestions", and "Visitors: 1407 © 2003 - 2004 Marti-Renom". The status bar at the bottom right shows "Internet".

Threading (mGenThreader)



✓ Good row and significance scoring
Obscure optimizer

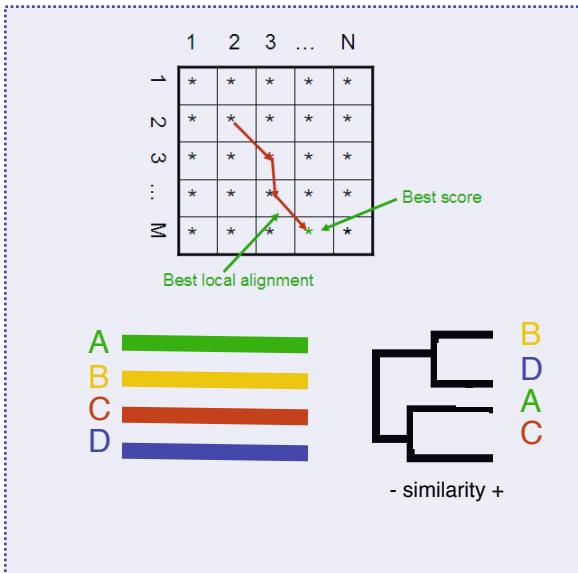
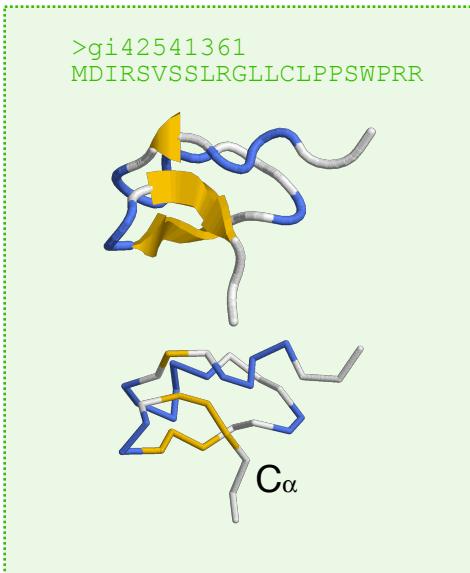


Threading (mGenThreader)

<http://bioinf.cs.ucl.ac.uk/psiform.html>

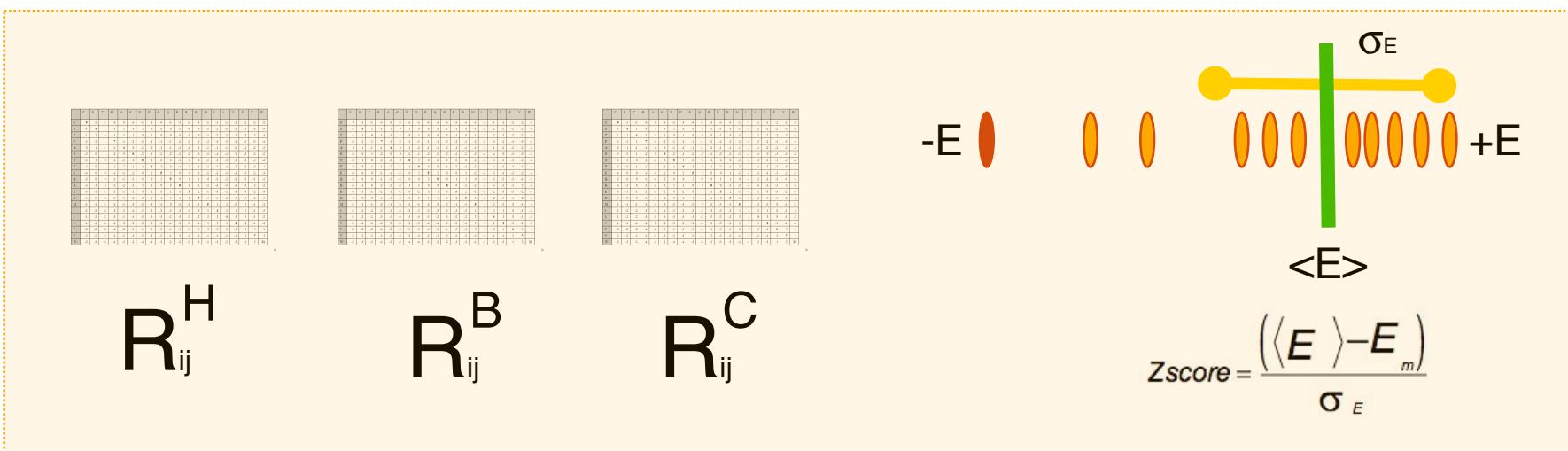
The screenshot shows the PSIPRED Protein Structure Prediction Server interface. At the top, there's a banner for the Bioinformatics Unit at UCL. The main menu includes 'PSIPRED home', 'Info', 'Input Sequence', 'Choose Prediction Method', and 'Filtering Options'. A red arrow points to the 'Choose Prediction Method' section, which contains four radio button options: 'Predict Secondary Structure (PSIPRED v2.4)', 'Predict Transmembrane Topology (MEMSAT)', 'Fold Recognition(GenTHREADER - quick)', and 'Fold Recognition (mGenTHREADER - with profiles and predicted secondary structure)'. Below this, the 'Filtering Options' section has three checkboxes: 'Mask low complexity regions' (checked), 'Mask transmembrane helices' (unchecked), and 'Mask coiled-coil regions' (unchecked). A warning message at the bottom says 'Warning: Turn off all filtering if you are running MEMSAT'.

Remote homology detection (FUGUE)



- ✓ Uses most of the structural information
- ✓ Easy to access either locally and on the web
- ✓ Good row and significance scoring

Does not use multiple sequence information



Remote homology detection (FUGUE)

<http://www-cryst.bioc.cam.ac.uk/fugue/>

The screenshot shows a Microsoft Internet Explorer window displaying the FUGUE homepage. The title bar reads "FUGUE: sequence-structure homology recognition and alignment engine - Microsoft Internet Explorer". The address bar shows the URL "http://www-cryst.bioc.cam.ac.uk/fugue/". The page content includes the FUGUE logo, the text "Crystallography and Biocomputing Unit Department of Biochemistry, University of Cambridge", and the University of Cambridge crest. Below this, a main heading states "Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties". A horizontal line separates this from a "Submit your protein sequence" section. This section contains links for "SEARCH STRUCTURAL DATABASE", "ALIGN SEQUENCE WITH STRUCTURE", "DOWNLOAD", and "DOCUMENTATION". Another horizontal line separates this from a "Methods" section. The "Methods" section contains text about the FUGUE program's purpose and how it works, along with a link to the original paper and practical information. At the bottom, there is a link to the HOMSTRAD database.

FUGUE: sequence-structure homology recognition and alignment engine - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Mail Links

Address http://www-cryst.bioc.cam.ac.uk/fugue/ Go Links

FUGUE

Crystallography and Biocomputing Unit
Department of Biochemistry, University of Cambridge

University of Cambridge crest

Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties

Submit your protein sequence

[SEARCH STRUCTURAL DATABASE](#)

[ALIGN SEQUENCE WITH STRUCTURE](#)

[DOWNLOAD](#)

[DOCUMENTATION](#)

Methods

FUGUE is a program for recognizing distant homologues by sequence-structure comparison. It utilizes environment-specific substitution tables and structure-dependent gap penalties, where scores for amino acid matching and insertions/deletions are evaluated depending on the local environment of each amino acid residue in a known structure. Given a query sequence (or a sequence alignment), FUGUE scans a database of structural profiles, calculates the sequence-structure compatibility scores and produces a list of potential homologues and alignments.

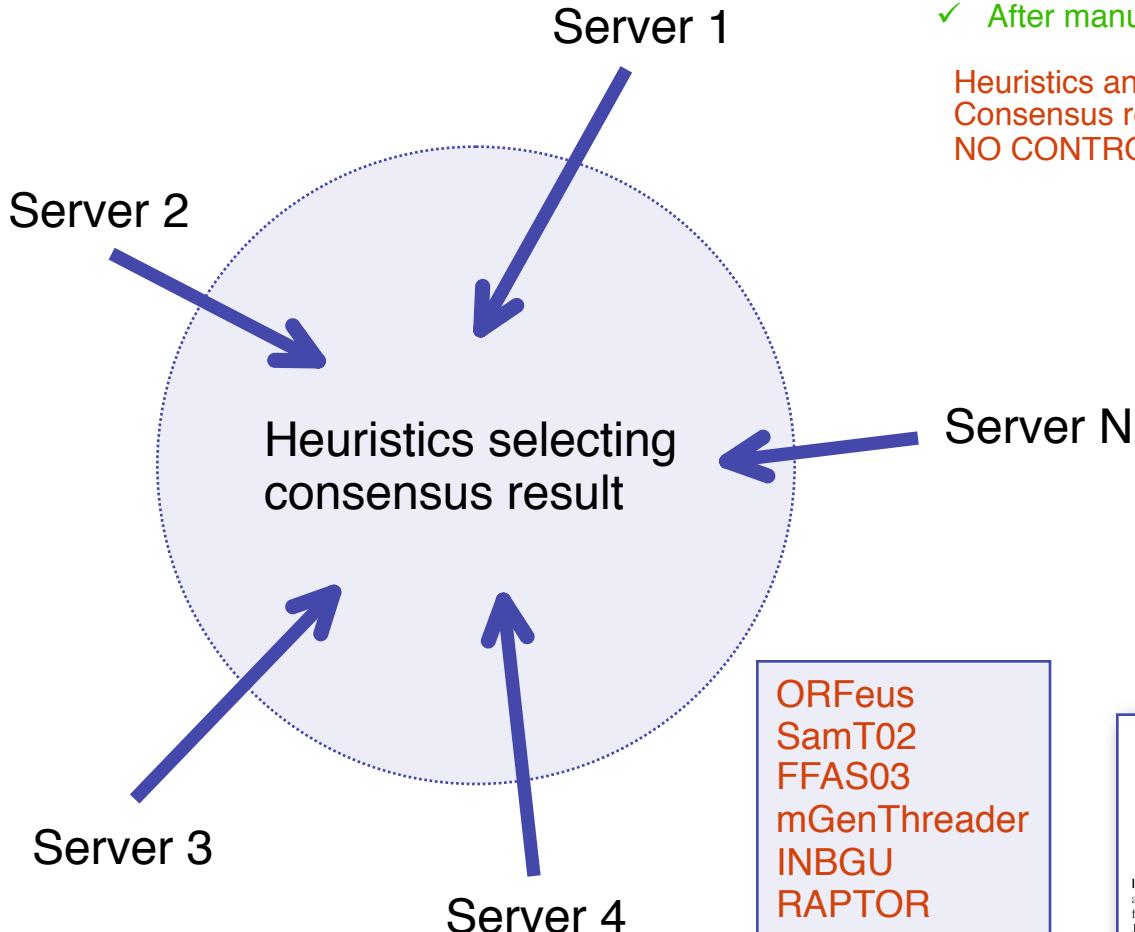
[Here](#) is a summary of how it works.

Read the original paper for more details:
[J. Shi, T. L. Blundell, and K. Mizuguchi \(2001\). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure- dependent gap penalties. *J. Mol. Biol.*, 310, 243-257.](#)
[Medline](#), [Article on-line](#), [PDF \(local only\)](#).

Some practical information can be found in:
[R. Núñez Miguel, J. Shi and K. Mizuguchi \(2001\). Protein Fold Recognition and Comparative Modeling using HOMSTRAD, JOY and FUGUE. In *Protein Structure Prediction: Bioinformatic Approach*. International University Line publishers, La Jolla, 143-169.](#)
[PDF \(local only\)](#)

Click [here](#) for information about the [HOMSTRAD](#) database.

Meta-Servers (3D-Jury)



- ✓ Collecting several results
- ✓ After manual analysis... good results

Heuristics and complicated scoring
Consensus results
NO CONTROL OF DATA GENERATION or SERVERS!

Cell, Vol. 115, 701–702, June 13, 2003, Copyright ©2003 by Cell Press

Letter to the Editor

mRNA Cap-1 Methyltransferase in the SARS Genome

The 3D jury system has predicted the methyltransferase fold for the nsP13 protein of the SARS coronavirus. Based on the conservation of a characteristic tetrad of residues, the mRNA cap-1 methyltransferase function has been assigned to this protein, which has potential implications for antiviral therapy.

Marcin von Drathen, Lucjan S. Wyrwicz,
Bikini Beach Institute
Uranowskiego 24A
60-744 Poznań
Poland

*Correspondence: leczak@biomed.pl

suggest that the virus also requires the AdoMet-dependent cap-0 methyltransferase. Both functions can be inhibited by carbocyclic analogs of adenosine, such as Haplosporin C, which is currently being evaluated together with the AdoMet-AdoHcy metabolism of the host cell (De Clercq, 1998; Bray et al., 2002). These compounds could complement other therapeutic strategies aimed at blocking viral enzymes, such as the RNA-dependent RNA polymerase, the protease, or the helicase encoded by the SARS virus.

Marcin von Drathen, Lucjan S. Wyrwicz,
Bikini Beach Institute
Uranowskiego 24A
60-744 Poznań
Poland

LETTERS

How Unique Is the Rice Transcriptome?

IN THE REPORT "COLLECTION, MAPPING, AND annotation of over 28,000 cDNA clones from japonica rice" (S. Kikuchi et al., 18 July, p. 376), the Rice Full-Length cDNA Project Team provides a detailed description of the rice transcriptome. The authors claim that 36% of the tested rice transcripts are not

practically reliable (3D jury score > 0.5) and contain a domain located in the first 7000 amino acid large (aa 1). A standard sequence logo-based or RPS-BLAST-like domain detection algorithm was used to assign any function to domains to the ancient family 2'-O-methyltransferases. Domains were derived from the last universal ruler (Liu et al., 2003). The enzymatic activity is determined by the presence of the K-C-X-E essential motif.

Figure 1. 3D Model of the nsP13 Domain of the SARS Coronavirus nsP13B Polyprotein. This model is based on the redesigned Bielski and Rybczynski template (Bielski et al., 1998; Rybczynski et al., 2000). While other templates (rat or 149) obtained marginally higher 3D jury scores, the selected template had the lowest error rate. The template is shown in blue, the conserved tetrad of residues (K-D-A-E) essential for cap-1 methylation and the docked AdoMet/AdoHcy are shown. Four blocks of aligned motifs containing the conserved, function-specific residues are shown in upper right corner.

Ginalski K, et al. (2003) Bioinformatics 19 pp1015
05/26/200

Meta-Servers (3D-Jury)

http://bioinfo.pl/Meta/

BIOINFO.PL: META

Meta Server Job List

[ABOUT] [SERVERS] [BENCHMARKS] [STATUS]

Structure Prediction Meta Server Input Page
0 jobs from 64.54.249. in the last week

Your E-mail:

Target Name:

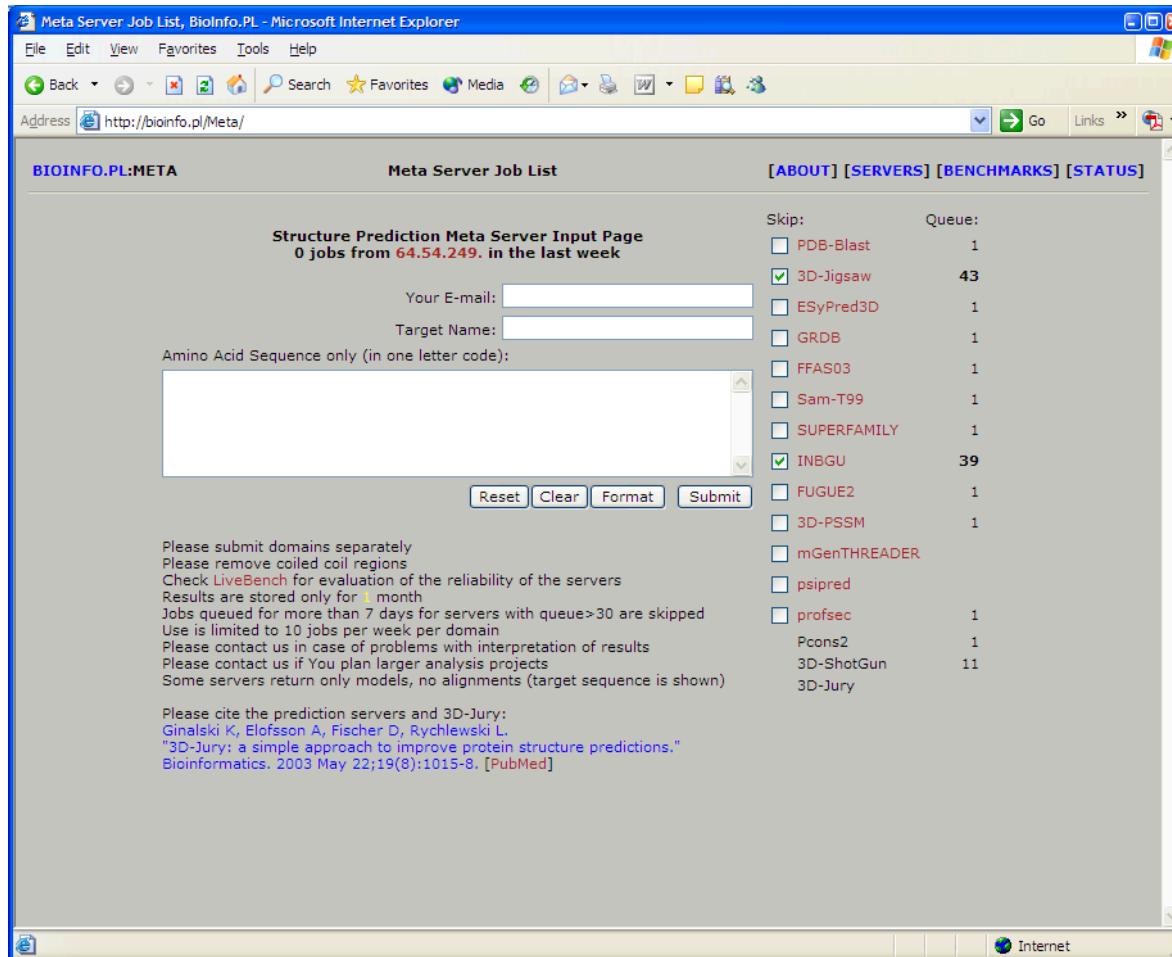
Amino Acid Sequence only (in one letter code):

Skip: Queue:

<input type="checkbox"/> PDB-Blast	1
<input checked="" type="checkbox"/> 3D-Jigsaw	43
<input type="checkbox"/> ESyPred3D	1
<input type="checkbox"/> GRDB	1
<input type="checkbox"/> FFAS03	1
<input type="checkbox"/> Sam-T99	1
<input type="checkbox"/> SUPERFAMILY	1
<input checked="" type="checkbox"/> INBGU	39
<input type="checkbox"/> FUGUE2	1
<input type="checkbox"/> 3D-PSSM	1
<input type="checkbox"/> mGenTHREADER	
<input type="checkbox"/> psipred	
<input type="checkbox"/> profsec	1
Pcons2	1
3D-ShotGun	11
3D-Jury	

Please submit domains separately
Please remove coiled coil regions
Check [LiveBench](#) for evaluation of the reliability of the servers
Results are stored only for 1 month
Jobs queued for more than 7 days for servers with queue>30 are skipped
Use is limited to 10 jobs per week per domain
Please contact us in case of problems with interpretation of results
Please contact us if You plan larger analysis projects
Some servers return only models, no alignments (target sequence is shown)

Please cite the prediction servers and 3D-Jury:
Ginalski K, Elofsson A, Fischer D, Rychlewski L.
"3D-Jury: a simple approach to improve protein structure predictions."
Bioinformatics. 2003 May 22;19(8):1015-8. [PubMed]



Complex gap penalty functions

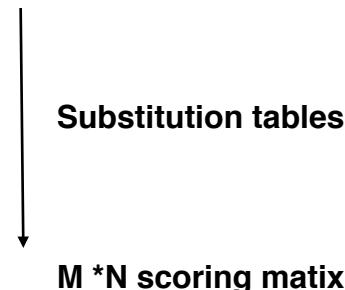
MODELLER SALIGN (ALIGN2D)

Madusudhan M.S. *et al.* in preparation

Regular dynamic programming (ALIGN)

Seq1-> DEFGHLKSMV

Seq2 -> FGHISAVCSSMLPQ



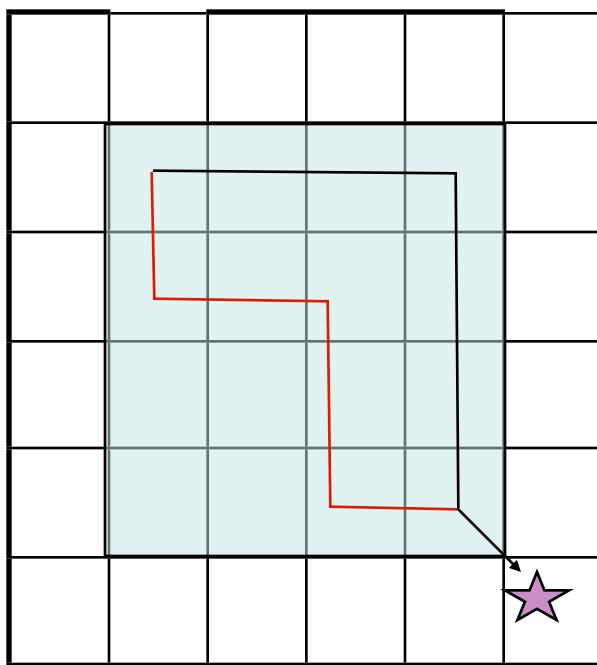
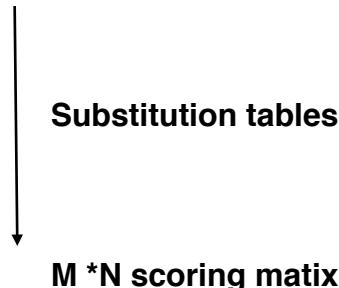
Creation of a gap penalized

Gap penalty = U + V

Align2D

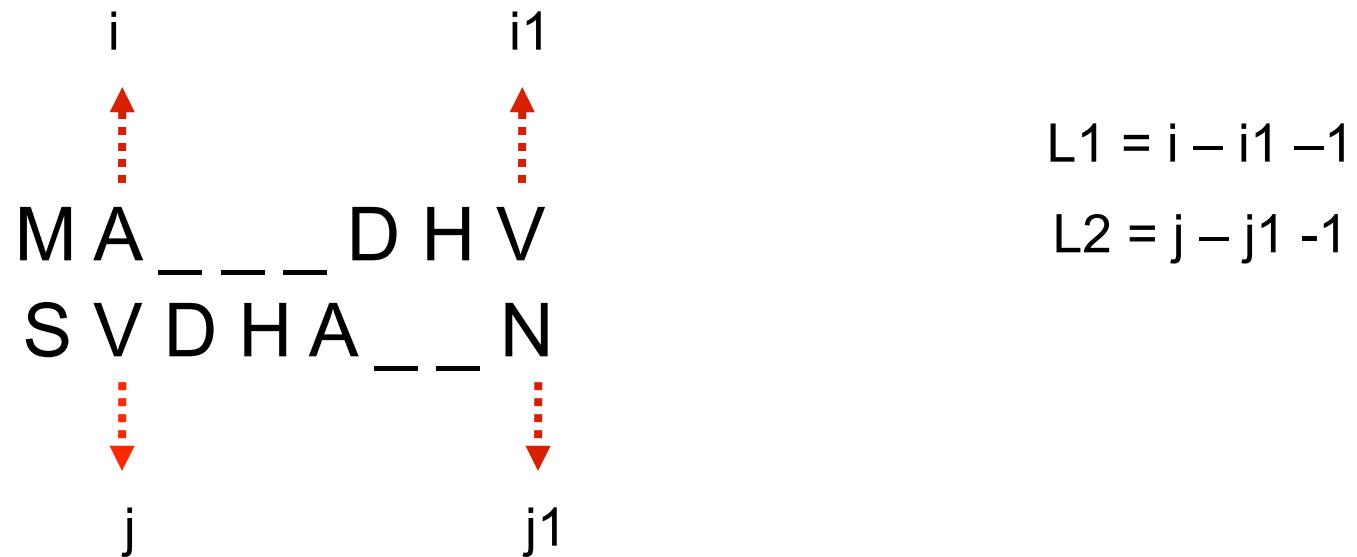
Seq1-> DEFGHLKSMV

Seq2 -> FGHISAVCSSMLPQ



$$G = u*f(i_1, i) + (L_1 + L_2)*v - \min(L_1, L_2) * t$$

ALIGN2D



$$G = u^*f(i1, i) + (L1+L2)^*v - \min(L1, L2) * t$$

opening extension diagonal

$$f(i_1, i) = 1 + W_a H + W_b S + W_B B + W_{st} C + W_d \max(0, d - d_0) g$$

H {=1 if helix unbroken, 0 otherwise}

S {=1 if strand unbroken, 0 otherwise}

B {average burial}

C { curvature, 1 if H or S, f(q) otherwise}

$$f(q) = 1 - \min(180, \max(0, q)) / 180$$

d = gap spanning distance

All averages are over residues i to i1 and over all template structures

Align

aln.pos	10	20	30	40	50	60	
1cydA	-----	LNFSGRLALVTGAGKGIGRDTVKALHASGAKVV			--AVTRTNSDLVSLAKECPGIEP		
1ybvA	DAIPGPLGPQSASLEGKVALVTGAGRGIGREMAMELGRRGCKVIVNYANSTESAEVVAAIKKNNGSDA	*	*****	****	*	*	
_consrvd					*	*	
_helix					999999999999999		
_beta					9999	9999	999
_buried					529305799996993589749956962739599	998352179349514831945	
_straight					77754468875663213677766774225577	97544357777763354569	



Strand interrupted

Align2D

Gap placed between helix
and strand

aln.pos	10	20	30	40	50	60	
1cydA	-----	LNFSGRLALVTGAGKGIGRDTVKALHASGAKVV	A	VTR	--TNSDLVSLAKECPGIEP		
1ybvA	DAIPGPLGPQSASLEGKVALVTGAGRGIGREMAMELGRRGCKVIVNYANSTESAEVVAAIKKNNGSDA	*	*****	****	*	*	
_consrvd					*	*	
_helix					999999999999999		
_beta					99999999		999
_buried					529305799996993589749956962739599983	52179349514831945	
_straight					777544688756632136777667742255779754	4357777763354569	

Profile-Profile alignments MODELLER SALIGN ('PROFILE')

Marti-Renom *et al.* (2004) Protein Science. 13:1071

Experiment (in silico)

- Benchmarking the best alignment methods.
- New alignment method.
- Projected gains.

Methods: Reference set

CE alignments with

- < 40% sequence identity
- > 100 EqPos
- > 50% EqPos
- > 90% coverage for one chain

387

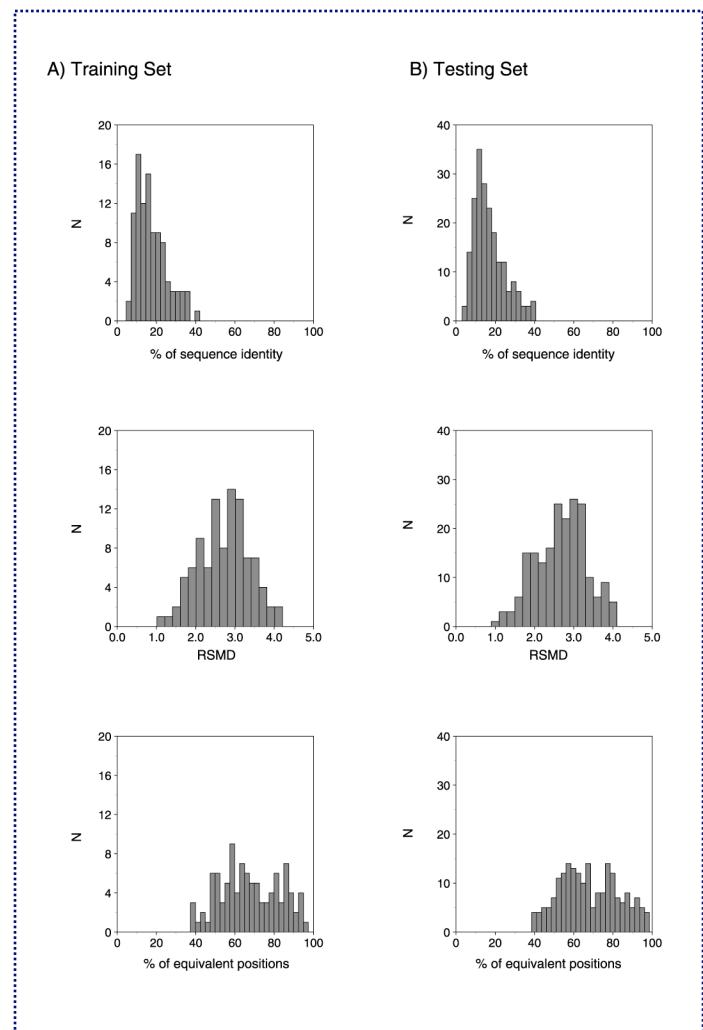
Filter: MAMMOTH alignments with

- > 50% EqPos

300

100 Training set

200 Testing set



Methods: Evaluated methods

Prof.-Prof. Prof.-Seq. Seq.-Seq.

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

ALIGN: DP pairwise method

BLAST2SEQ: Local method

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

ALIGN4D: DP pairwise method that uses multiple sequence information for both sequences.



Methods: Evaluated methods

Seq.-Seq.

ALIGN: DP pairwise method

BLAST2SEQ: Local method

Prof.-Seq.

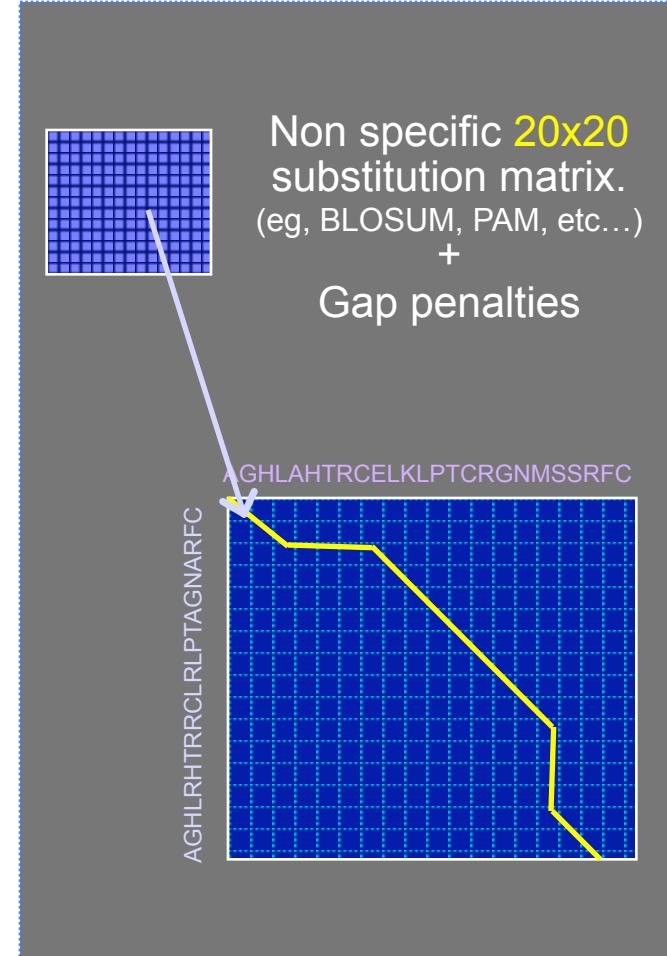
PSI-BLAST: Local search method that uses multiple sequence information for one of the sequences.

Prof.-Prof.

ALIGN4D: DP pairwise method that uses multiple sequence information for both sequences.

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRRCLRLPTAGNARFC



Methods: Evaluated methods

Prof.-Prof. Prof.-Seq. Seq.-Seq.

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

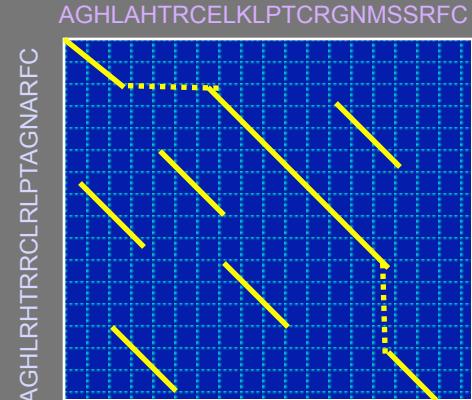
Sequence B: AGHLRHTRRCLRLPPTAGNARFC

ALIGN: DP pairwise method

BLAST2SEQ: Local method

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

ALIGN4D: DP pairwise method that uses multiple sequence information for both sequences.



Methods: Evaluated methods

Seq.-Seq.

ALIGN: DP pairwise method

BLAST2SEQ: Local method

Prof.-Seq.

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

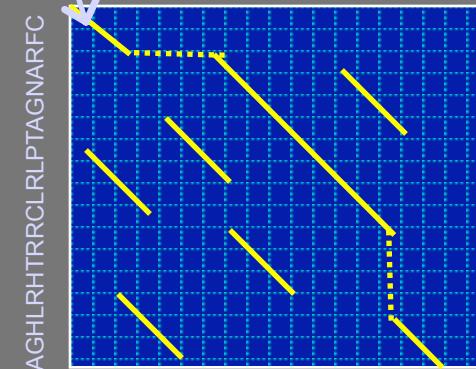
Prof.-Prof.

ALIGN4D: DP pairwise method that uses multiple sequence information for both sequences.

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Non specific **20x20** substitution matrix.
(eg, BLOSUM, PAM, etc...) +
Gap penalties



Methods: Evaluated methods

Prof.-Prof. Prof.-Seq. Seq.-Seq.

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

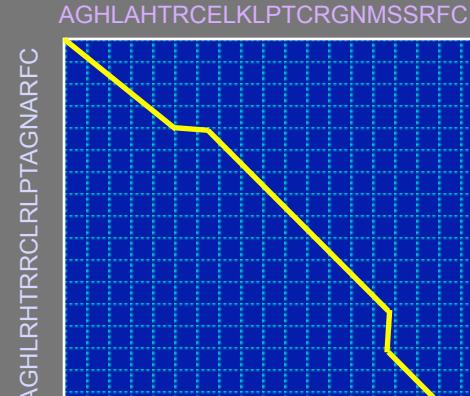
Sequence B: AGHLRHTRRCLRLPTAGNARFC

ALIGN: DP pairwise method

BLAST2SEQ: Local method

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

ALIGN4D: DP pairwise method that uses multiple sequence information for both sequences.



Methods: Evaluated methods

Prof.-Prof. Prof.-Seq. Seq.-Seq.

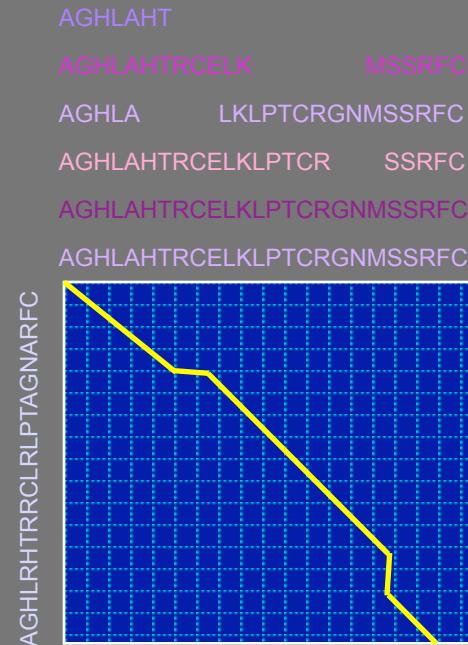
Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC
Sequence B: AGHLRHTRRRCLRLPTAGNARFC

ALIGN: DP pairwise method

BLAST2SEQ: Local method

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

ALIGN4D: DP pairwise method that uses multiple sequence information for both sequences.



Methods: Evaluated methods

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

BLAST2SEQ: Local method

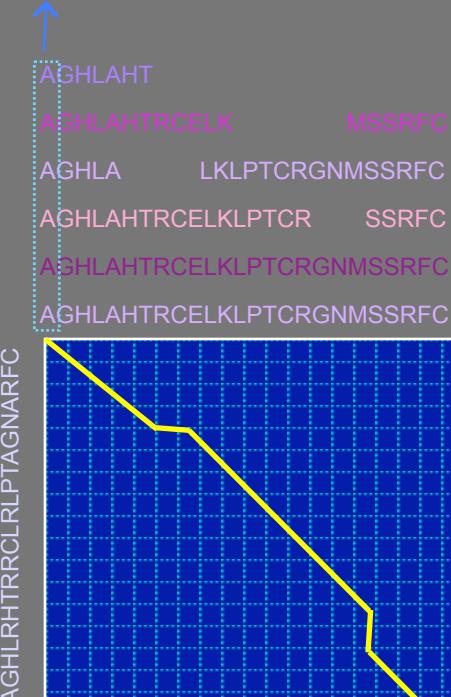
Prof.-Seq.

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

Prof.-Prof.

ALIGN4D: DP pairwise method that uses multiple sequence information for both sequences.

PSSMA A C D E .../... V W Y
+3 -1 -2 -2 .../... -2 -1 -3



Methods: Evaluated methods

Prof.-Prof. Prof.-Seq. Seq.-Seq.

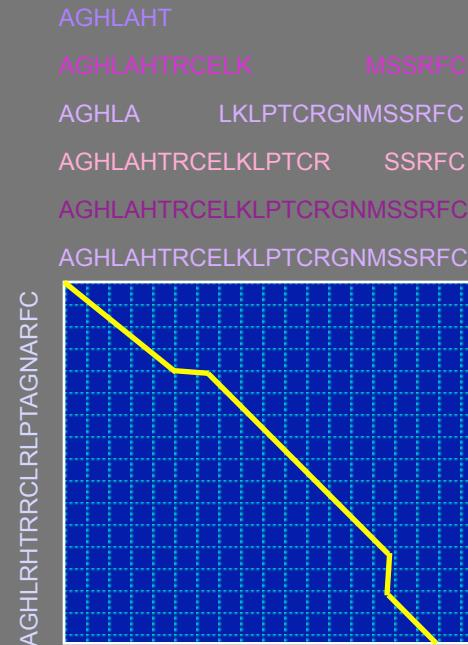
Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC
Sequence B: AGHLRHTRRRCLRLPTAGNARFC

ALIGN: DP pairwise method

BLAST2SEQ: Local method

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

ALIGN4D: DP pairwise method that uses multiple sequence information for both sequences.



Methods: Evaluated methods

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

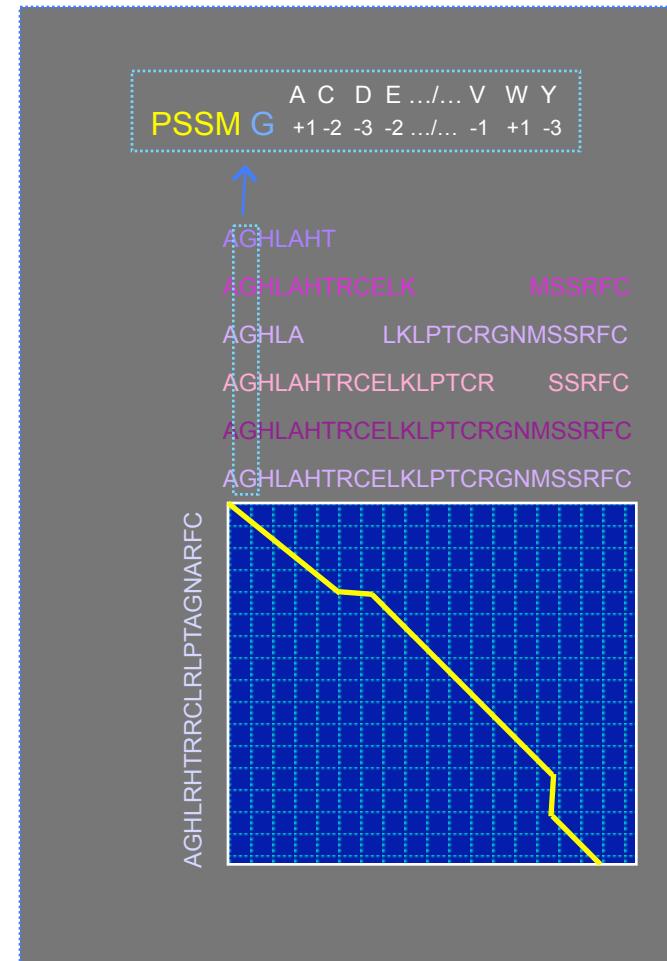
BLAST2SEQ: Local method

Prof.-Seq.

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

Prof.-Prof.

ALIGN4D: DP pairwise method that uses multiple sequence information for both sequences.



Methods. Evaluated methods.

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

BLAST2SEQ: Local method

Prof.-Seq.

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

Prof.-Prof.

ALIGN4D: DP pairwise method that uses multiple sequence information for both sequences.



Methods. Evaluated methods.

Prof.-Prof. Prof.-Seq. Seq.-Seq.

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC
Sequence B: AGHLRHTRRCLRLPTAGNARFC

ALIGN: DP pairwise method

BLAST2SEQ: Local method

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

ALIGN4D: DP pairwise method that uses multiple sequence information for both sequences.



Methods. SALIGN.

ALIGN4D protocol	Profile	Comparison	Open	Extension
CCPBP	PSI-BLAST	Correlation Coefficient	-300	0
CC _{HH}	Henikoff-Henikoff	Correlation Coefficient	-300	0
CC _{HS}	H-H + similarity weight	Correlation Coefficient	-150	0
<hr/>				
EDPBP	PSI-BLAST	Euclidian Distance	-450	-30
ED _{HH}	Henikoff-Henikoff	Euclidian Distance	-550	0
ED _{HS}	H-H + similarity weight	Euclidian Distance	-450	-10
<hr/>				
DPPBP	PSI-BLAST	Dot Product	-250	-30
DP _{HH}	Henikoff-Henikoff	Dot Product	-550	0
DP _{HS}	H-H + similarity weight	Dot Product	-100	-30
<hr/>				
JSHH	Henikoff-Henikoff	Jansen-Shannon Distance	-150	0
JSHS	H-H + similarity weight	Jansen-Shannon Distance	-250	0

Methods: Coverage and accuracy

High coverage



Low accuracy

High accuracy

Low coverage

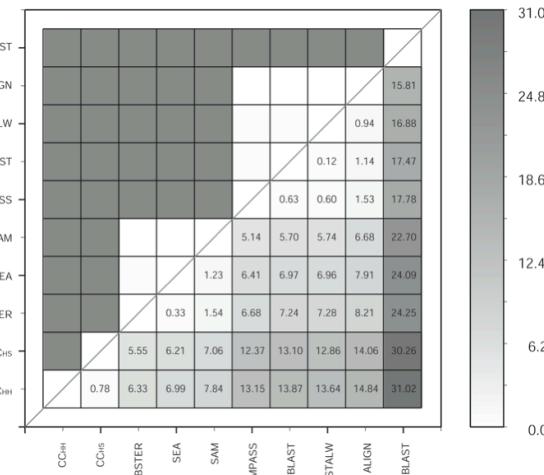
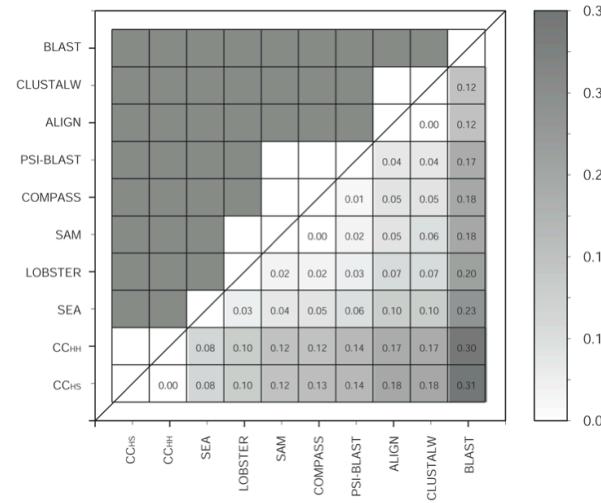
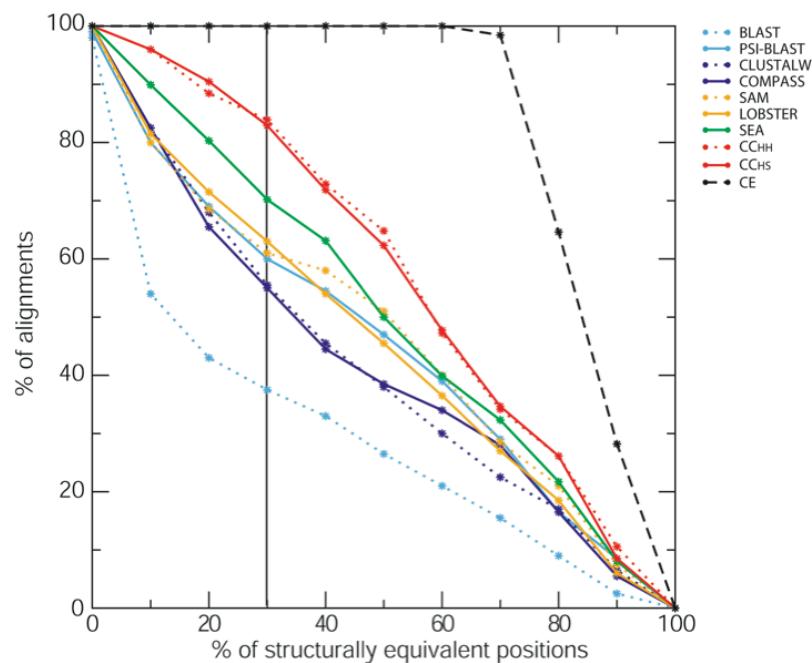


High accuracy

Low accuracy

Results: Comparison of alignment dependent measures

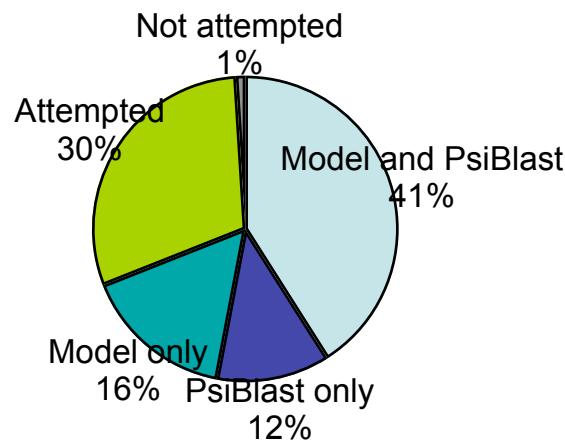
Method	CE overlap [%]	Shift score	RMSD [Å]	Structure overlap [%]
CE	100 ± 0	1.00 ± 0.00	2.7 ± 0.6	59.8 ± 12.9
BLAST	26 ± 29	0.32 ± 0.33	5.6 ± 3.7	20.6 ± 23.7
PSI-BLAST	43 ± 31	0.48 ± 0.35	6.5 ± 3.9	30.3 ± 24.9
SAM	48 ± 26	0.50 ± 0.34	9.2 ± 4.7	28.9 ± 24.8
LOBSTER	50 ± 27	0.51 ± 0.32	9.1 ± 4.9	31.1 ± 25.2
SEA	49 ± 27	0.53 ± 0.29	8.4 ± 4.4	33.4 ± 24.3
ALIGN	42 ± 25	0.44 ± 0.28	10.6 ± 5.0	25.7 ± 24.1
CLUSTALW	43 ± 27	0.44 ± 0.31	10.2 ± 4.9	26.4 ± 24.3
COMPASS	43 ± 32	0.49 ± 0.35	4.8 ± 3.2	32.3 ± 24.7
CC_{HH}	56 ± 23	0.61 ± 0.24	7.8 ± 4.2	36.7 ± 22.9
CC_{HS}	56 ± 24	0.62 ± 0.24	7.8 ± 4.2	36.5 ± 23.2



Results. Turn over.

Mycoplasma genitalium MODPIPE Models

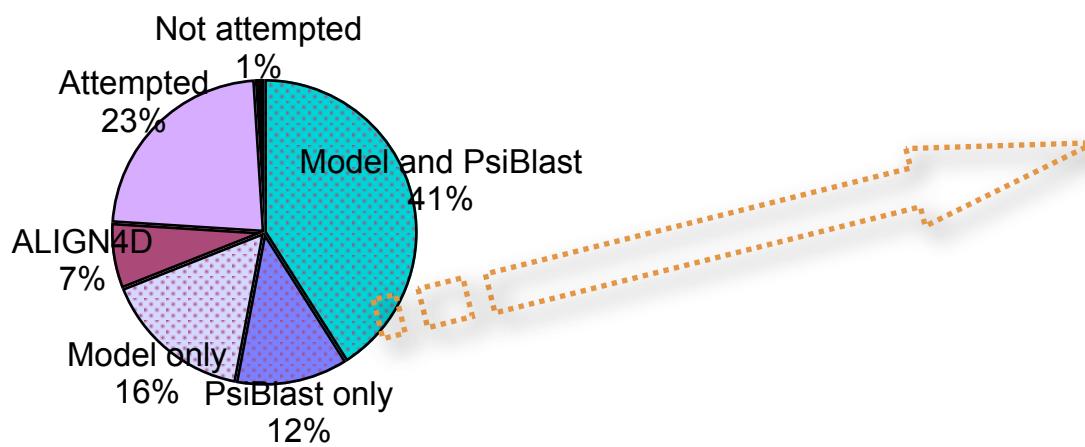
Number of ORFs	479
Average ORF length	364



Results. Turn over.

Mycoplasma genitalium MODPIPE Models

Number of ORFs	479
Average ORF length	364



~ 34 extra
accurate models
for M. g. genome.

~ 100,000 models
for TrEMBL-SP
“genome”.

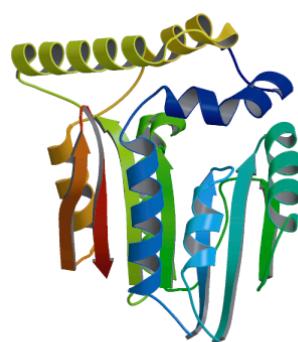
Examples: T0092 model

- Target T0092 at CASP4:
- Hypothetical protein HI0319
- Haemophilus influenzae
- Parent: 1d2cA (Methyltransferase)
- ALIGN4D alignment at 8.4% seq id.

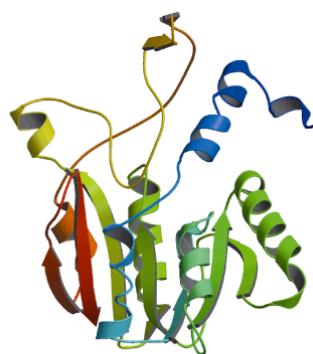
Method	RMSD Å	% of EqPos
ALIGN4D CC _{PBP}	5.9	67.84
PSI-BLAST	4.9	31.72
Best predictions at CASP4	6.0	65.20

Data from CASP4, Asilomar, CA, December 2000.

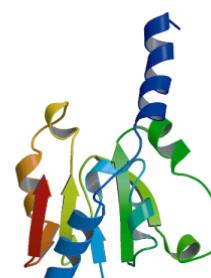
B) Target T0092



X-Ray structure



ALIGN4D (CC_{PBP}) model



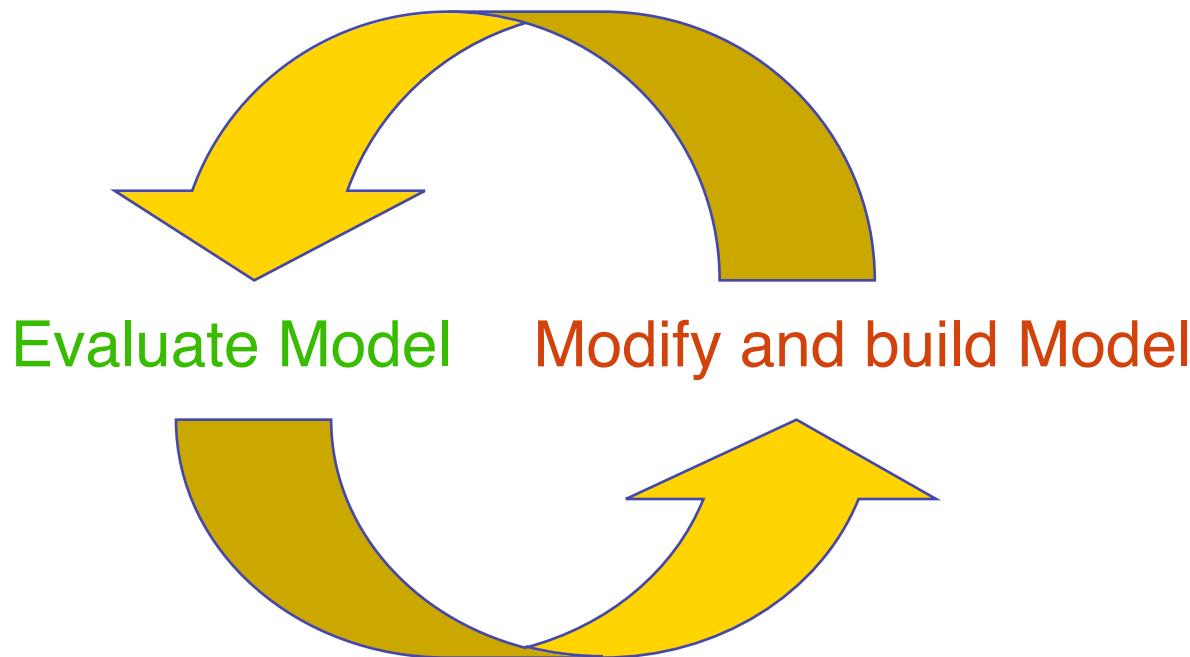
Psi-Blast model

Iterative process

MOULDER

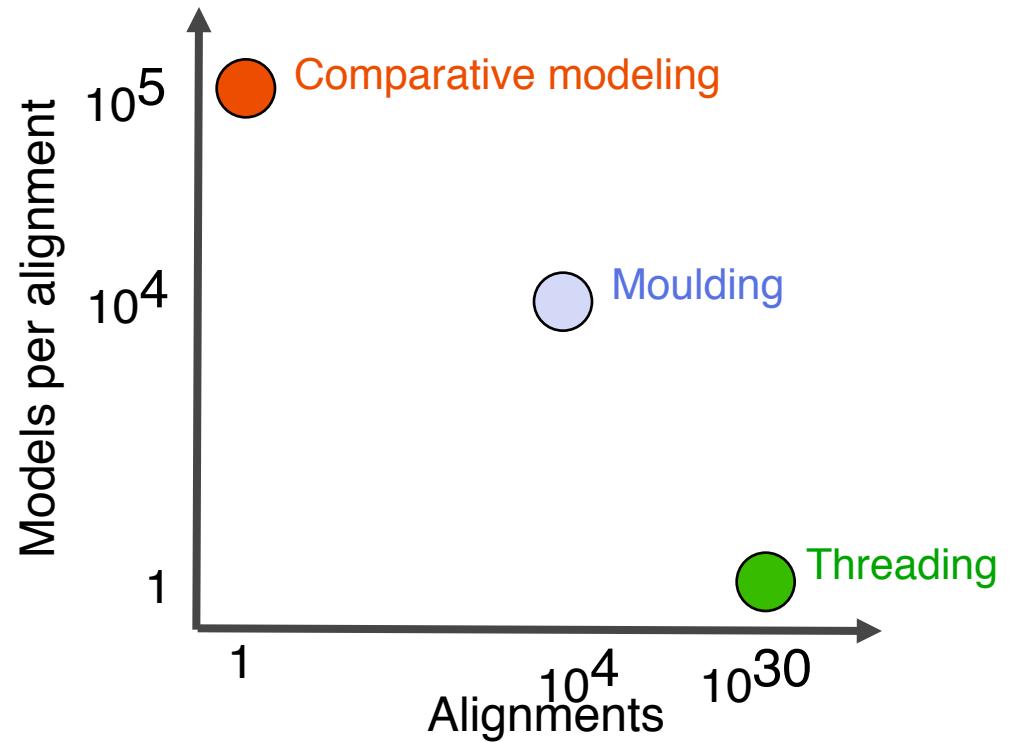
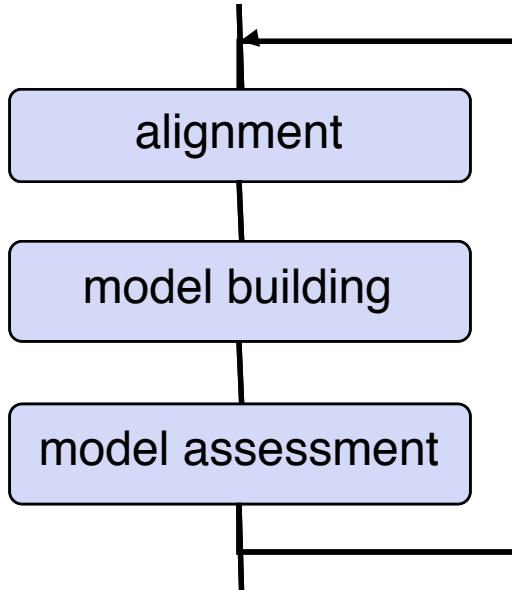
John, B. and Sali, A. (2003) Nucleic Acids Research. **31**:1982-1992

Iterative process... better models(?)



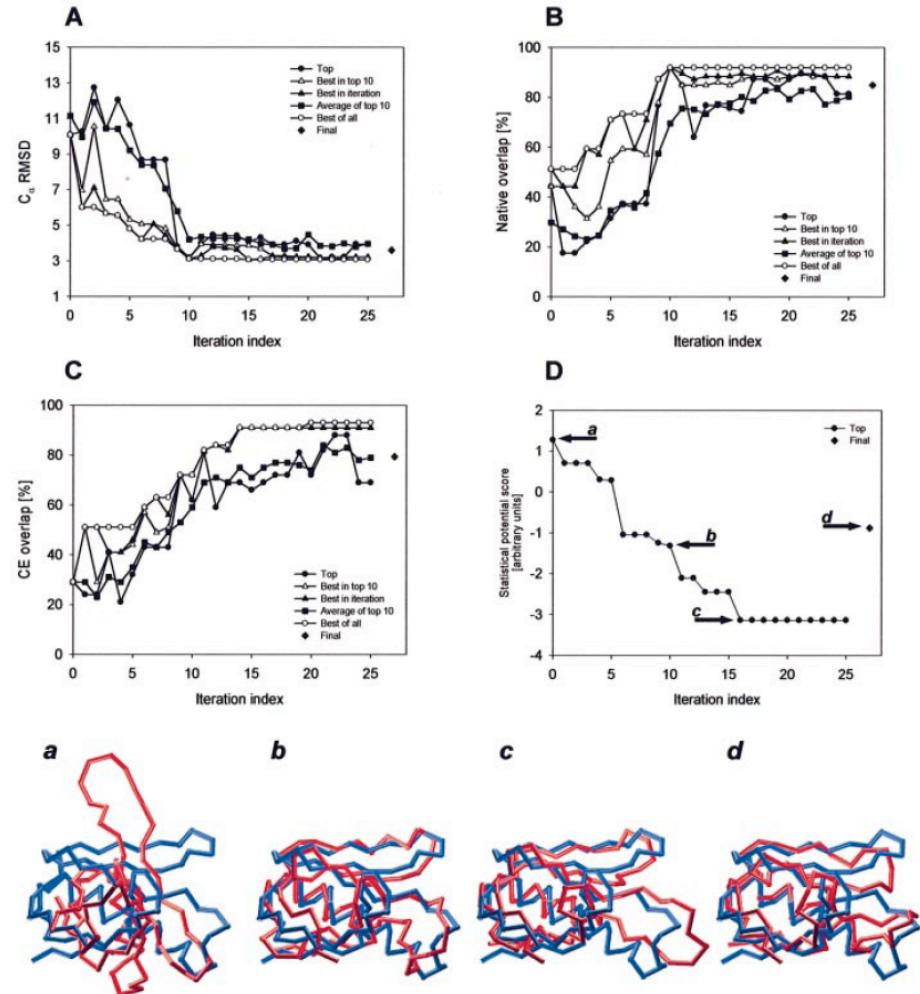
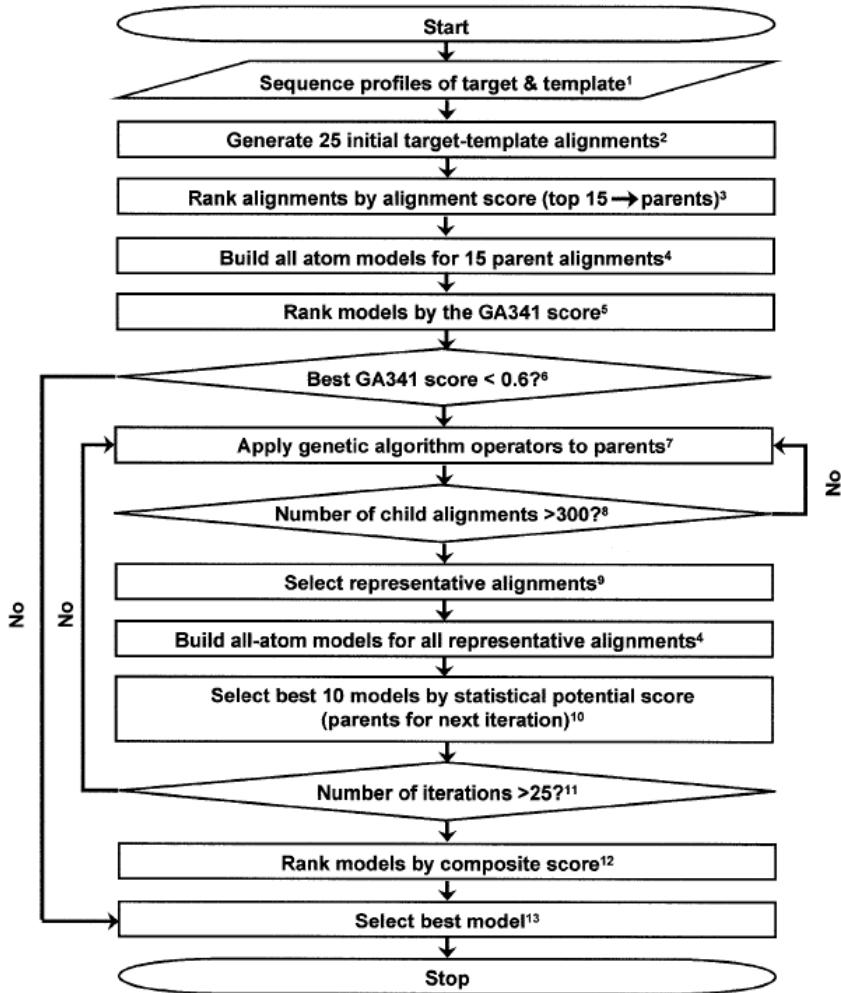
Moulding: iterative alignment, model building, model assessment

B. John, A. Sali. Nucl. Acids Res., 31, 1982-1992, 2003.



Iterative process... MOULDER

more in model evaluation



BMC WorkShop

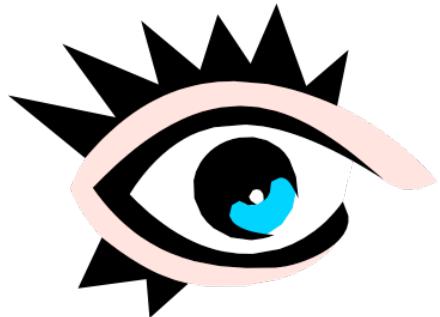
Protein Structure Prediction

model building
(model assessment)

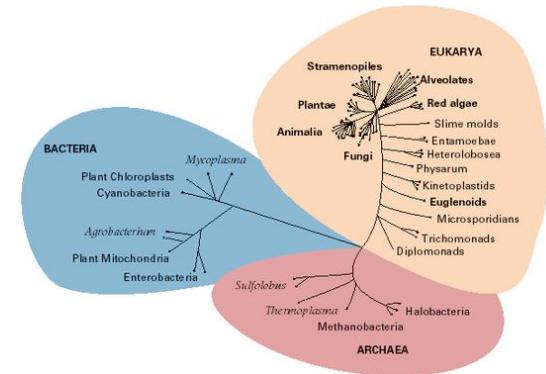
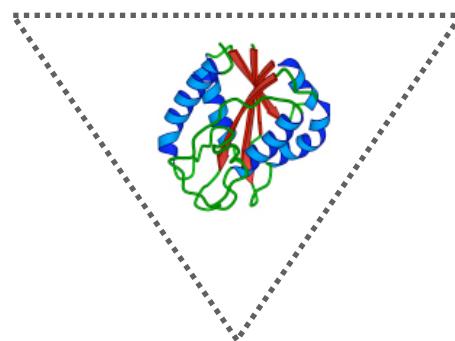
Marc A. Marti-Renom & Damien Devos

Department of Biopharmaceutical Sciences, UCSF

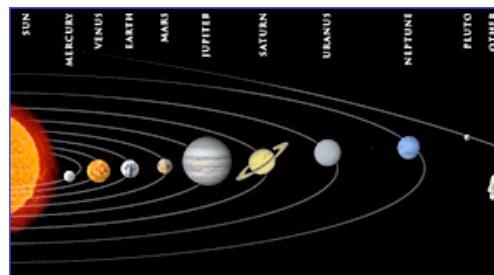
Information about a protein can come from three distinct sources



Experimental
observations



Statistical rules

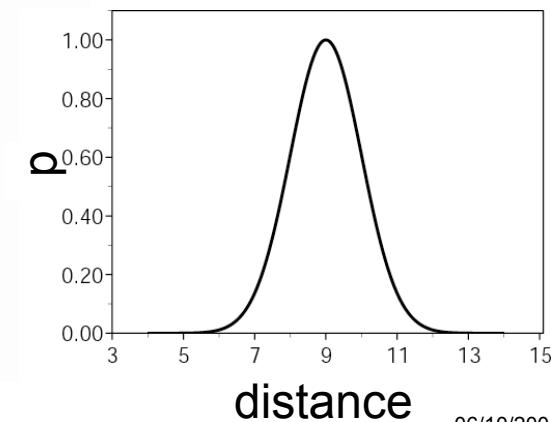
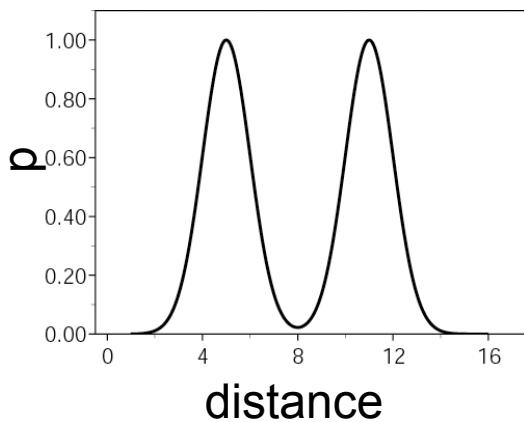
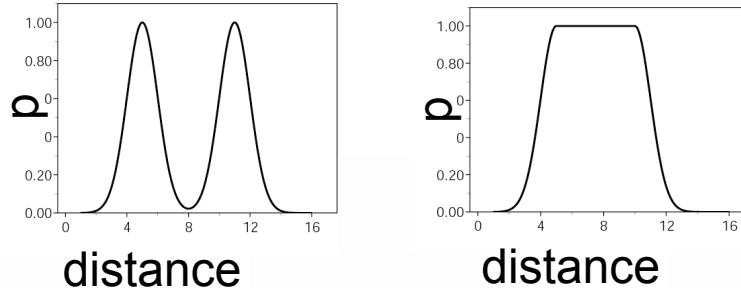


Laws of physics

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

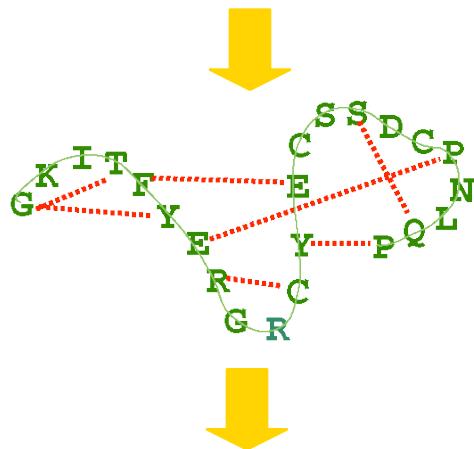
- Model building by assembly of rigid bodies: core, loops, sidechains.
- Model building by segment matching.
- Model building by satisfaction of spatial restraints.

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

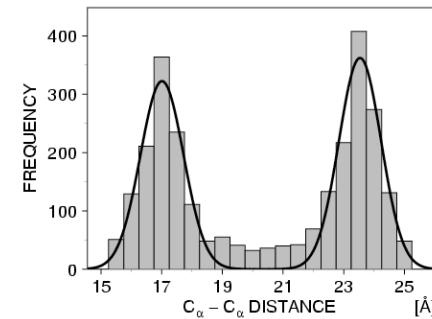
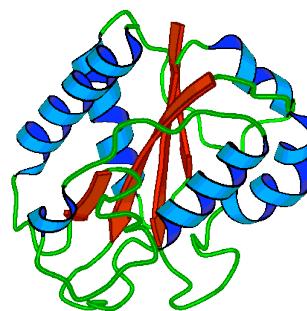
Comparative modeling by satisfaction of spatial restraints MODELLER

3D GKITFYERGFQGHCYESDC-NLQP...
SEQ GKITFYERG---RCYESDCPNLQP...

1. Extract spatial restraints



2. Satisfy spatial restraints



$$F(R) = \prod_i p_i (f_i / l)$$

- 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.

<http://salilab.org/>

Restraints

$p(d)$

$p(d/d')$

$p(d/d', a, g, s, i)$

$p(d/d', d'', \dots)$

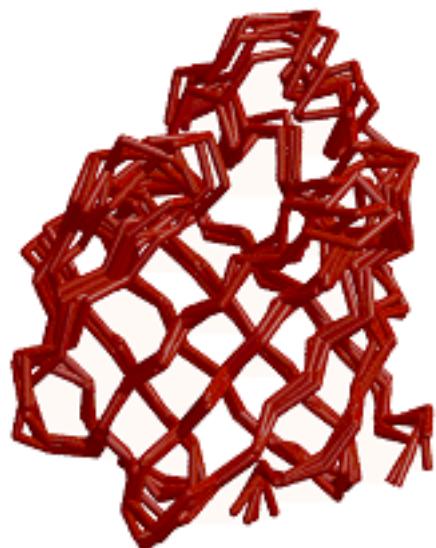
$p(S/R, S', R', t, s)$

$p(M/R, M', R, s)$

A. Šali & T. Blundell. *J. Mol. Biol.* **234**, 779, 1993.

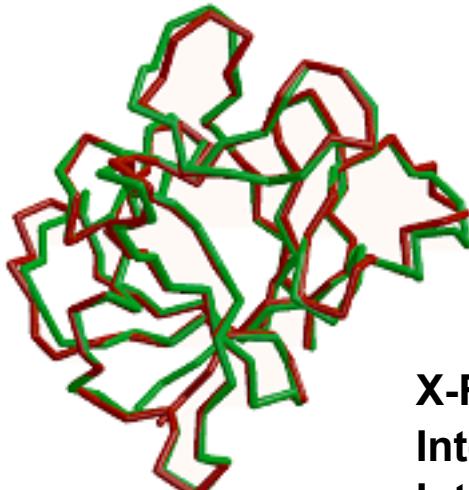
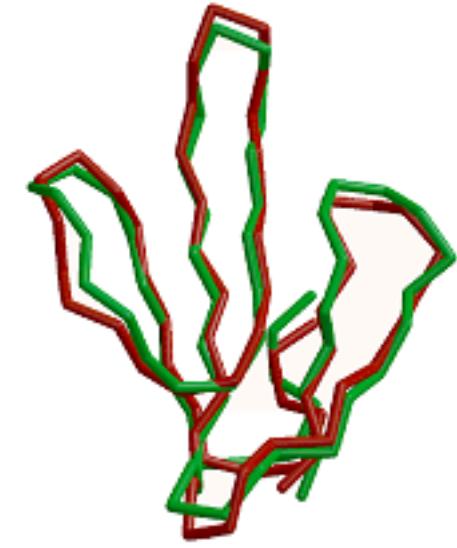
Accuracy and applicability of comparative models

“Biological” significance of modeling errors



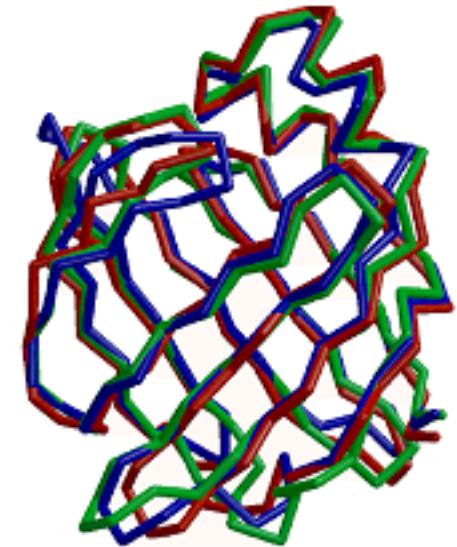
NMR
Ileal lipid-binding protein
1eal

NMR – X-RAY
Erbabutoxin 3ebx
Erbabutoxin 1era



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

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



Assessing errors is important

Manual:

Critical Assessment of Techniques for Protein Structure Prediction
(CASP) (<http://predictioncenter.llnl.gov/>)

Automated:

CAFASP

EVA (<http://salilab.org/~eva/>)

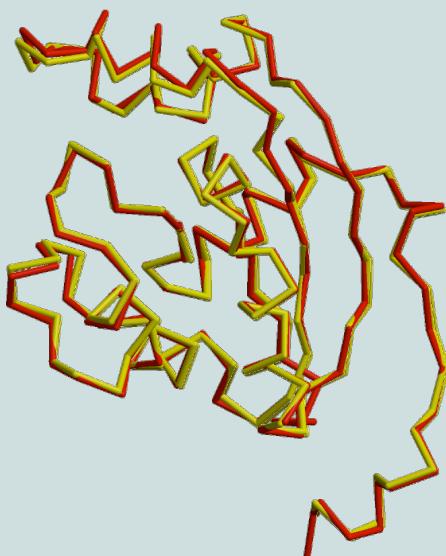
LiveBench (<http://bioinfo.pl/>)

Model Accuracy

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

HIGH ACCURACY

NM23
Seq id 77%
 $C\alpha$ equiv 147/148
RMSD 0.41 Å

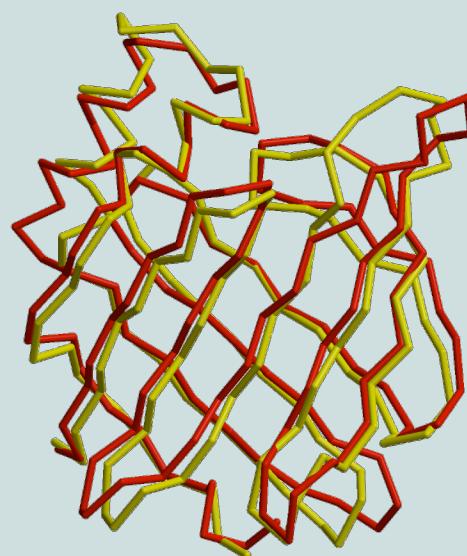


Sidechains
Core backbone
Loops

X-RAY / MODEL

MEDIUM ACCURACY

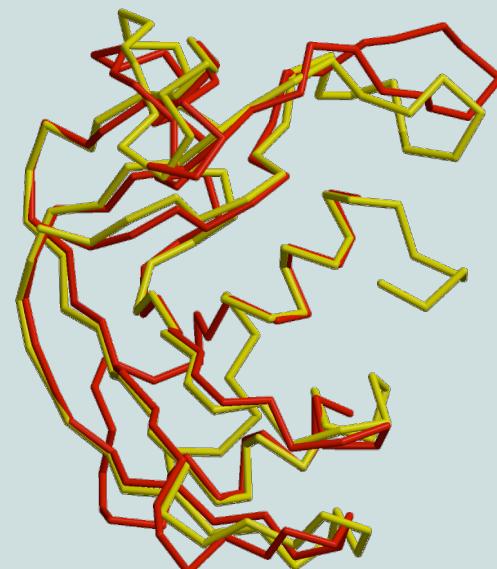
CRABP
Seq id 41%
 $C\alpha$ equiv 122/137
RMSD 1.34 Å



Sidechains
Core backbone
Loops
Alignment

LOW ACCURACY

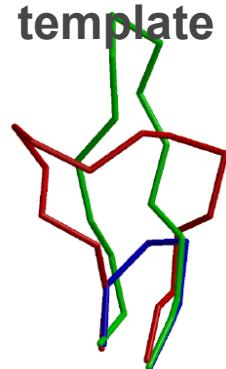
EDN
Seq id 33%
 $C\alpha$ equiv 90/134
RMSD 1.17 Å



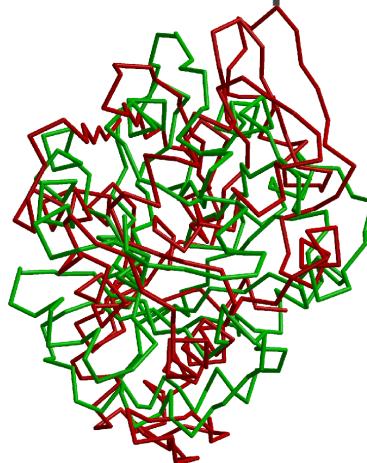
Sidechains
Core backbone
Loops
Alignment
Fold assignment

MODEL X-RAY TEMPLATE

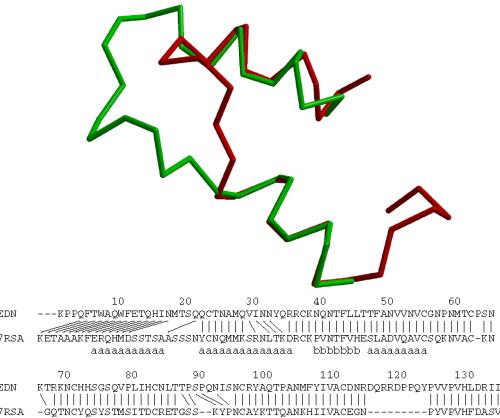
Region without a



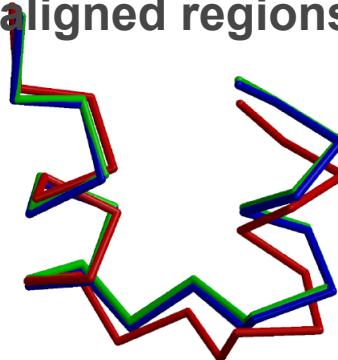
Incorrect template



Misalignment



Distortion/shifts in aligned regions



Sidechain packing



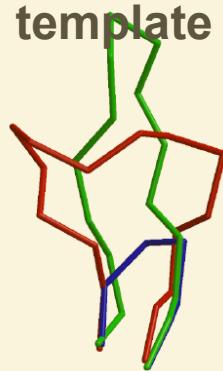
MODEL

X-RAY

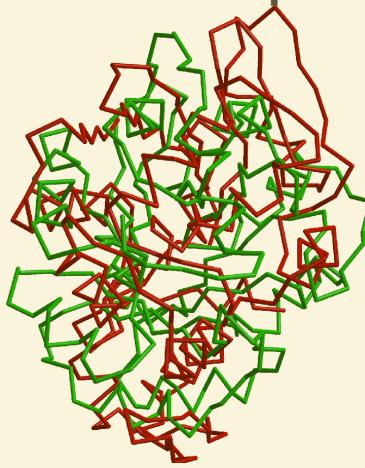
TEMPLATE

Typical errors in comparative models

Region without a



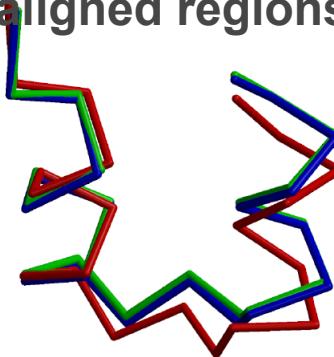
Incorrect template



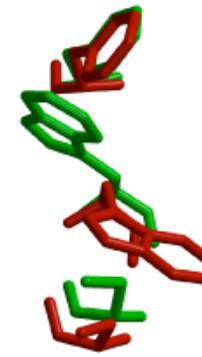
Misalignment



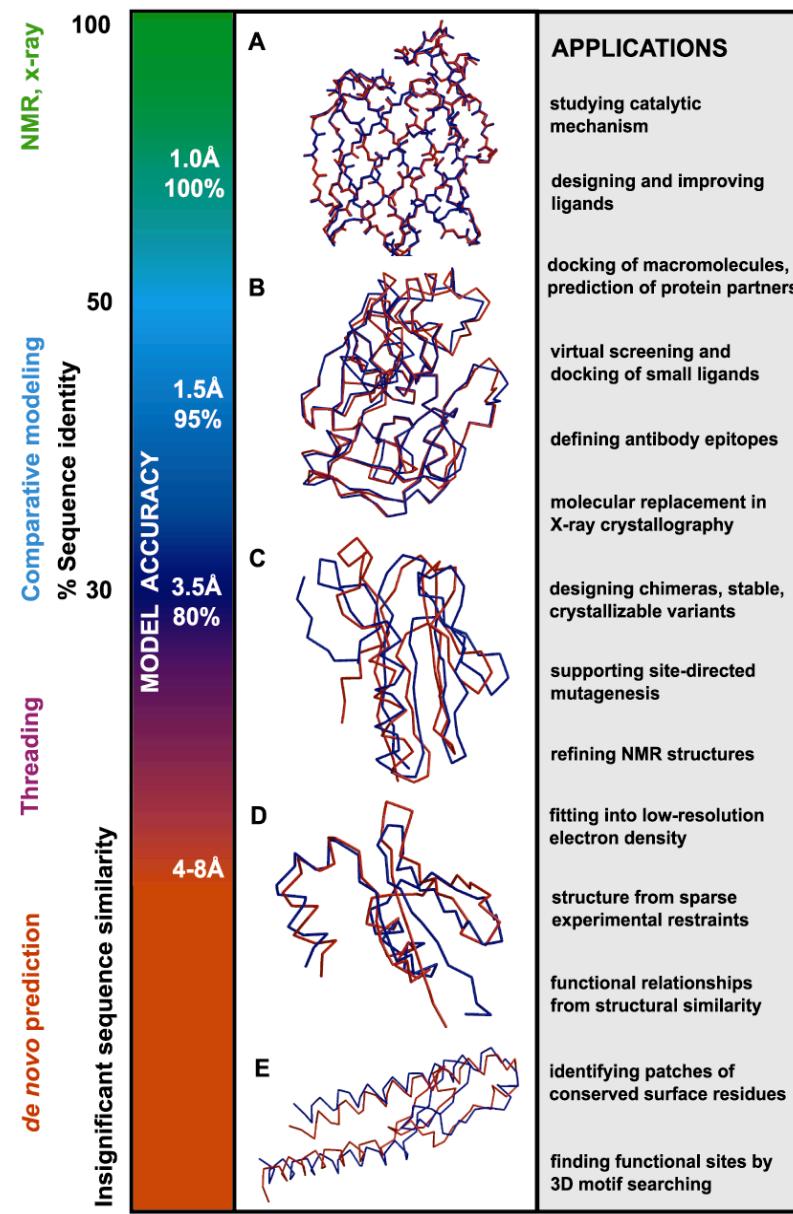
Distortion/shifts in aligned regions



Sidechain packing



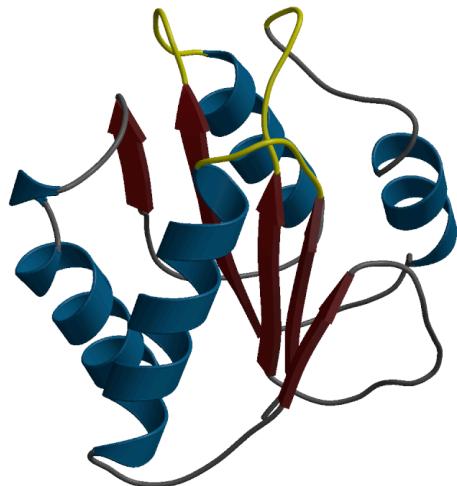
Utility of protein structure models, despite errors



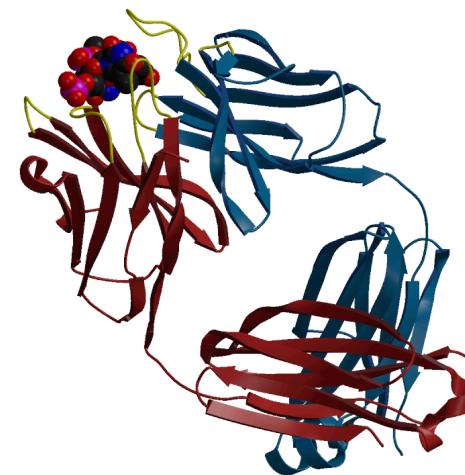
D. Baker & A. Sali.
Science 294, 93, 2001.

Modeling of loops in protein structures (modeling of insertions)

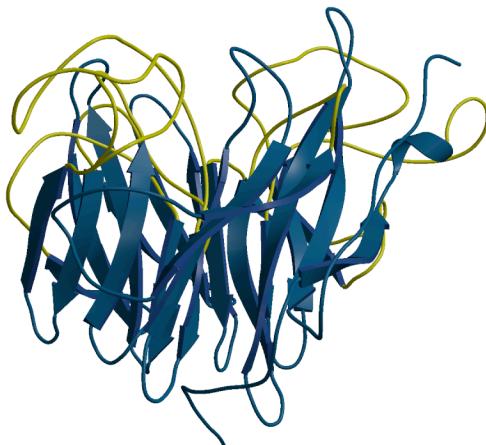
Loop Modeling in Protein Structures



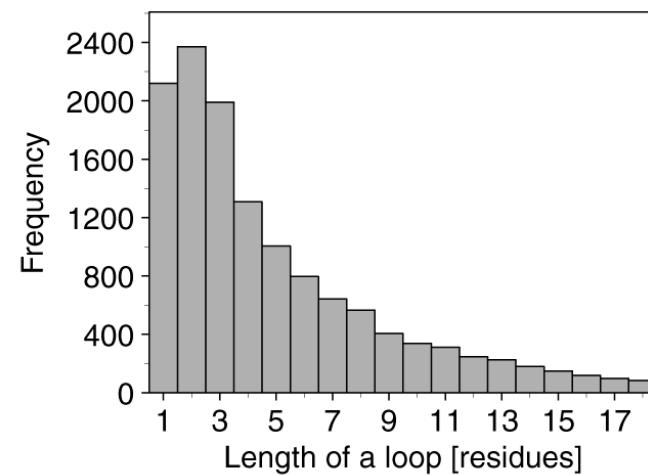
$\alpha+\beta$ barrel: flavodoxin



Ig fold: immunoglobulin



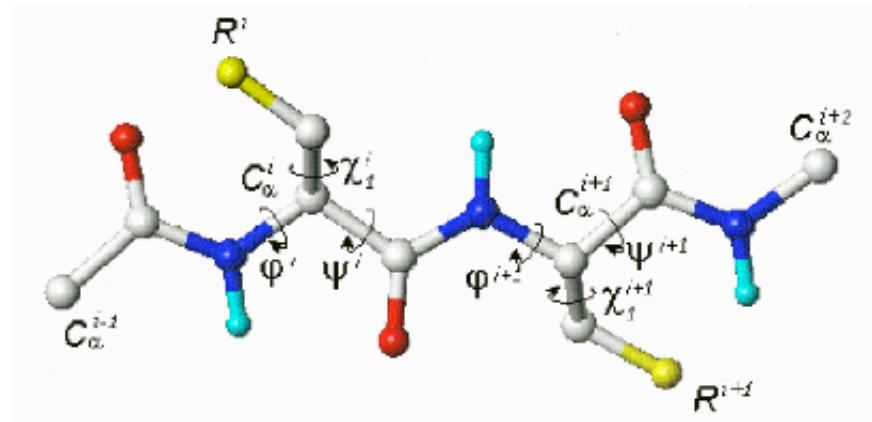
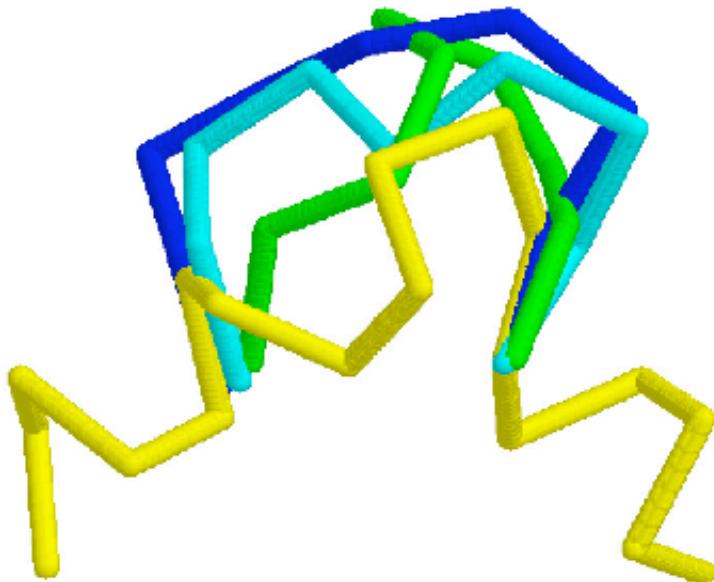
antiparallel β -barrel



Loop modeling strategies

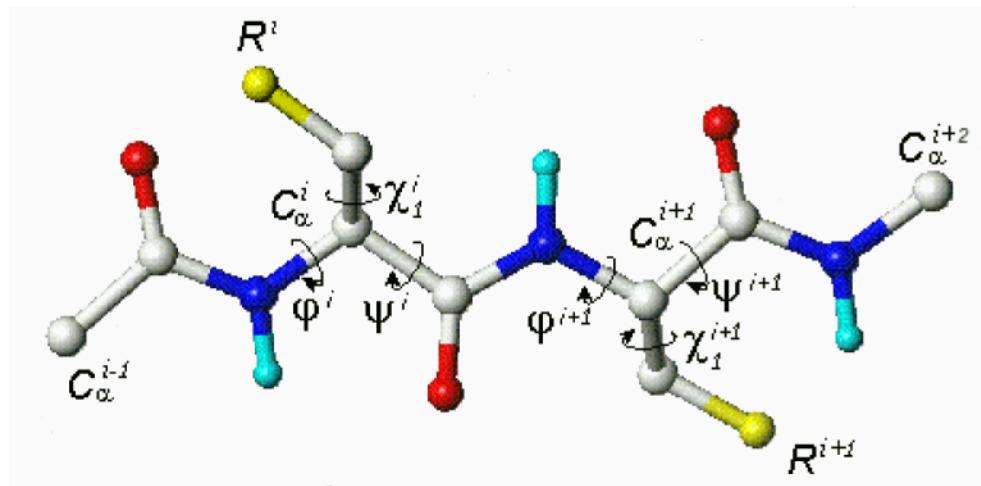
Database search

Conformational search



- 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.

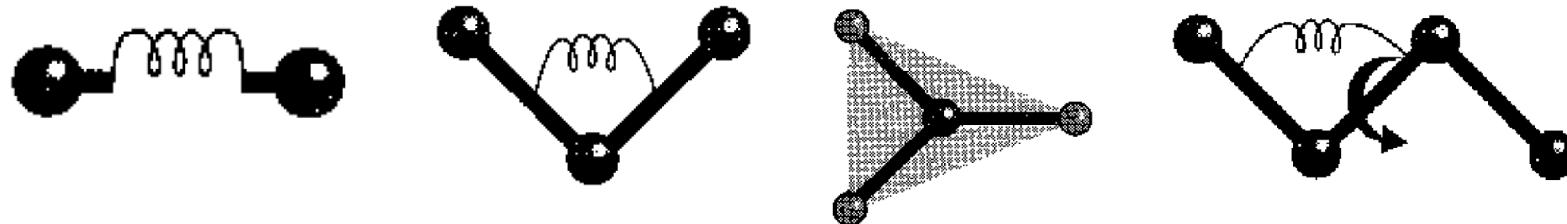
Energy Function for Loop Modeling

The energy function is a sum of many terms:

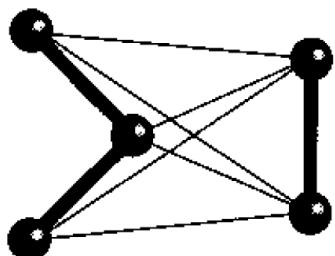
1) Statistical preferences for dihedral angles:



2) Restraints from the CHARMM-22 force field:

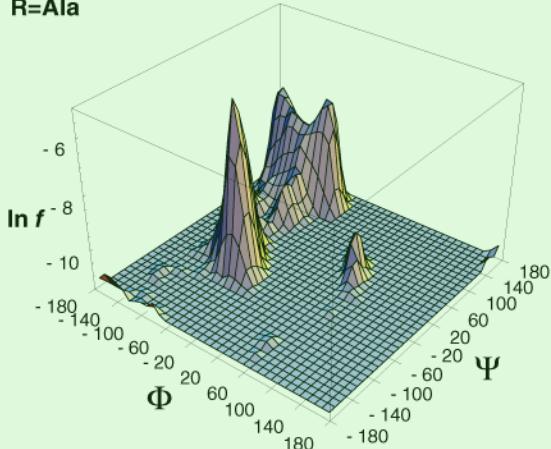


3) Statistical potential for non-bonded contacts:

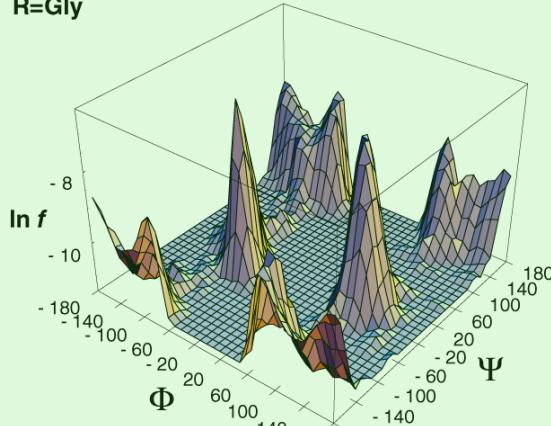


Mainchain Terms for Loop Modeling

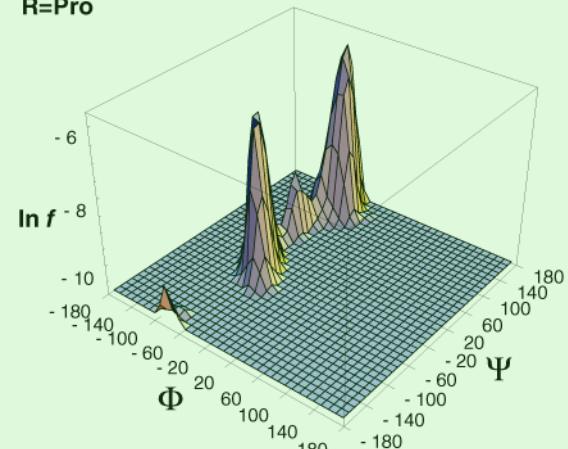
R=Ala



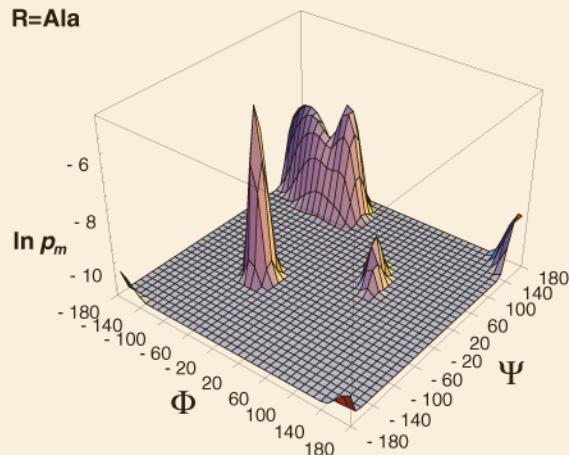
R=Gly



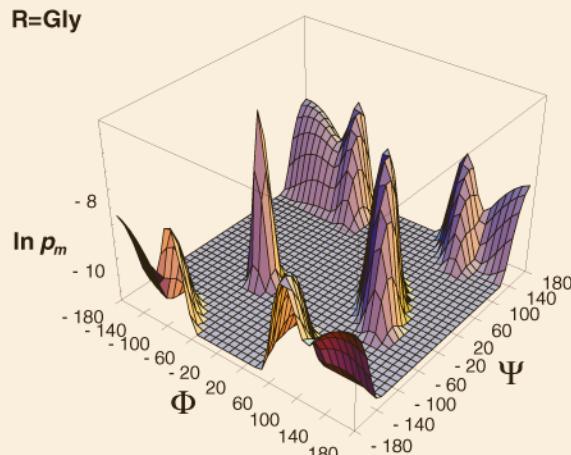
R=Pro



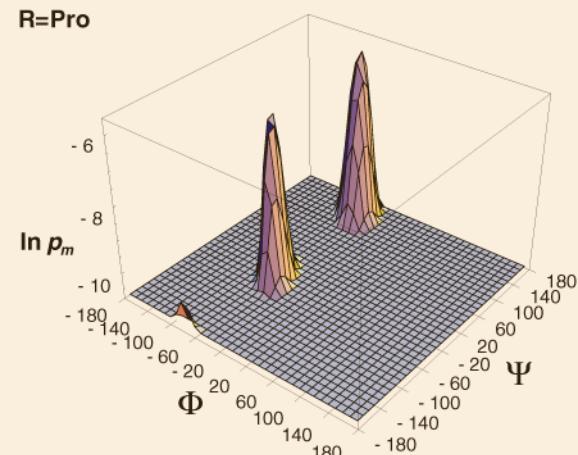
R=Ala



R=Gly

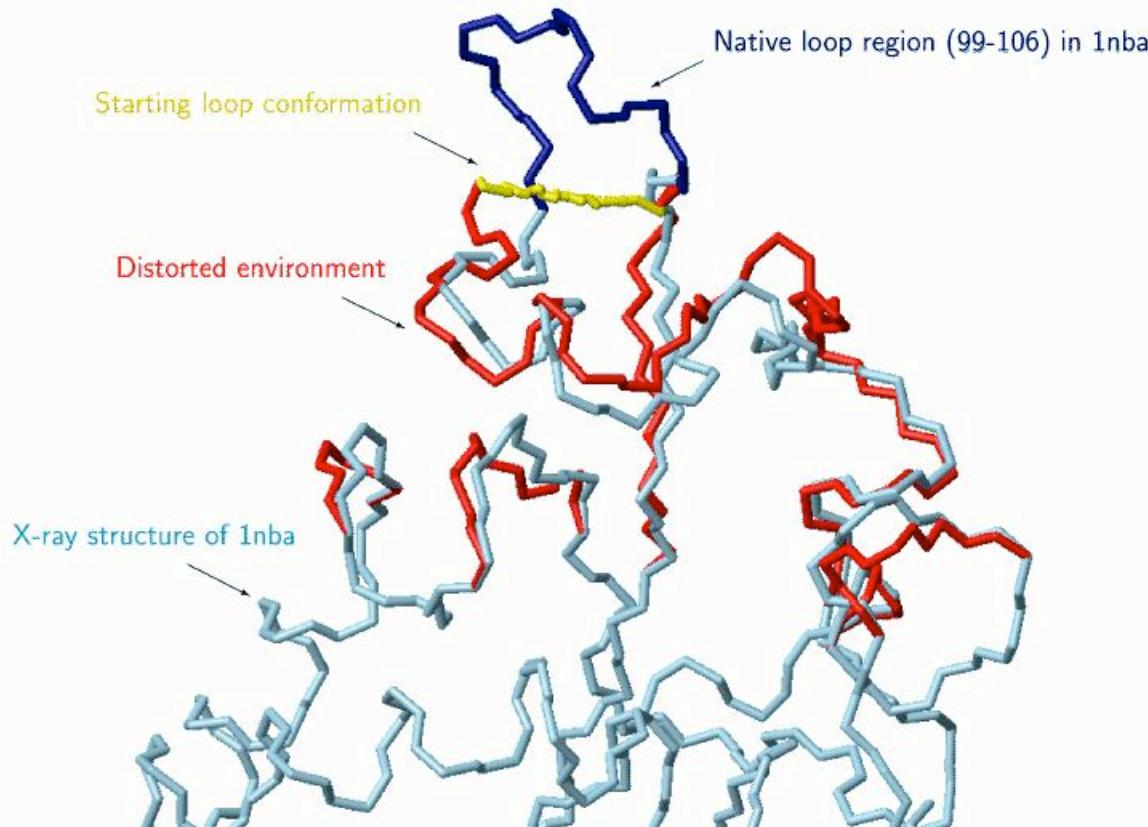


R=Pro

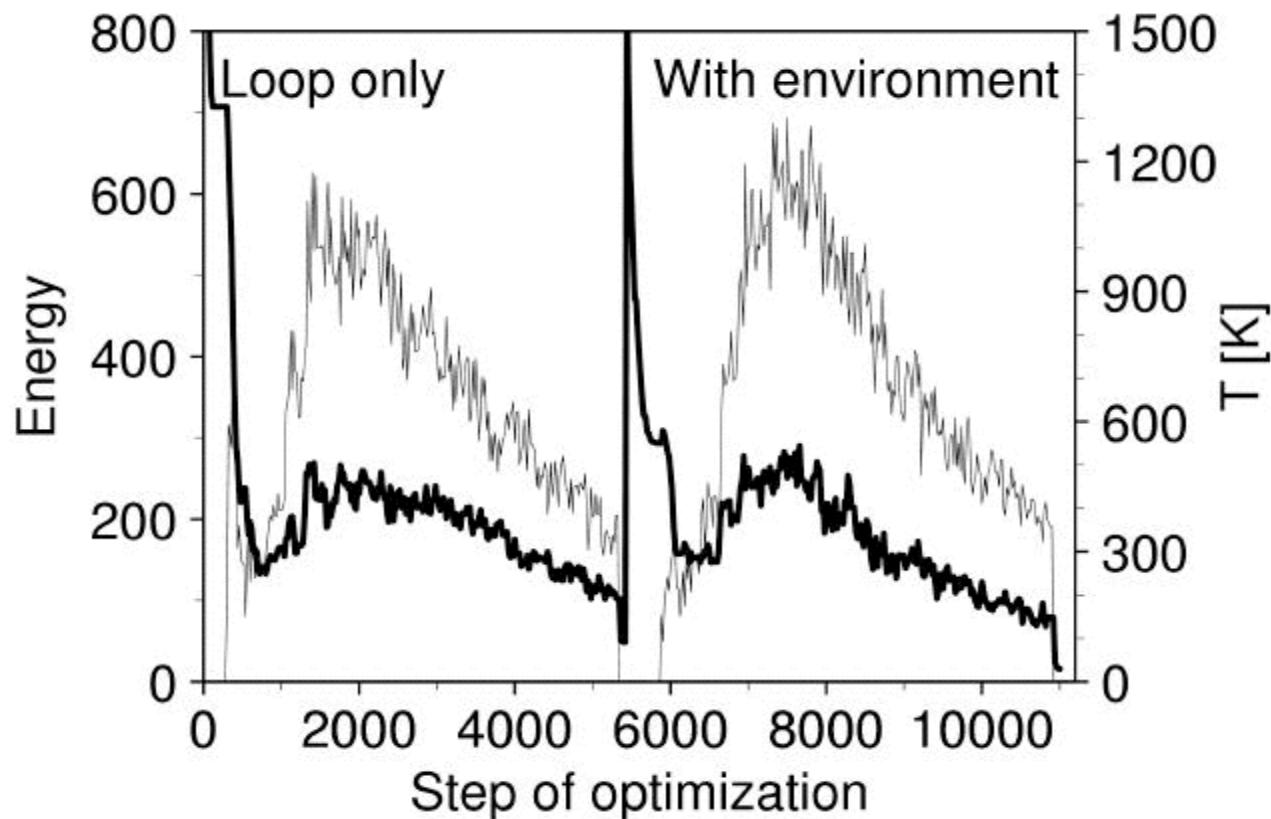


Optimization of Objective Function

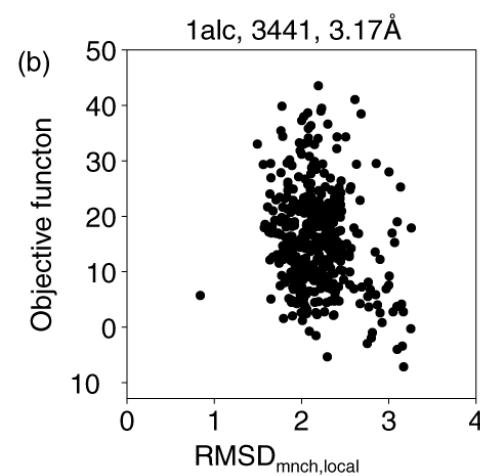
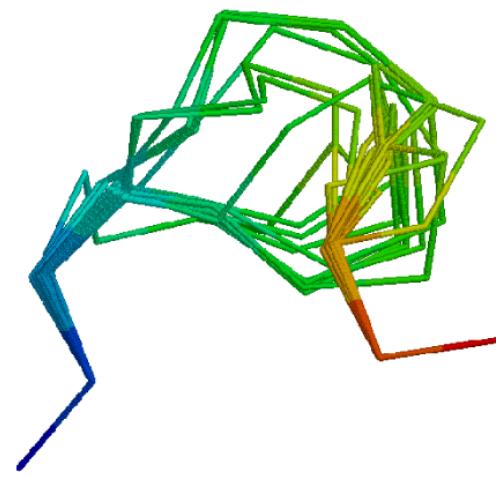
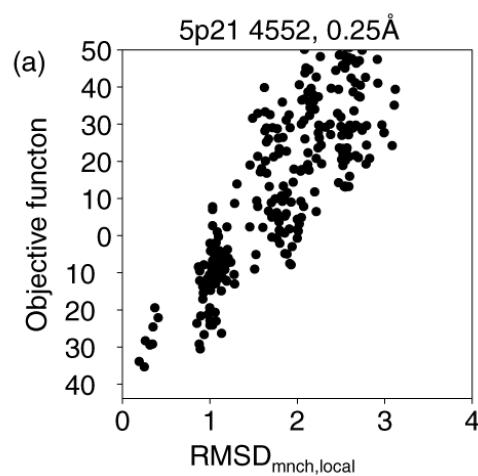
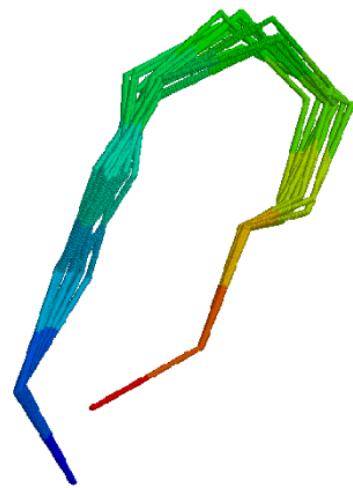
- Test set: 40 randomly selected loops of known structures, for each length from 1 to 14 residues.
- Starting conformation: Loop atoms were spaced evenly on a line spanning the two anchor regions, then randomized by $\pm 5 \text{ \AA}$.
- To simulate real comparative modeling situations, performance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



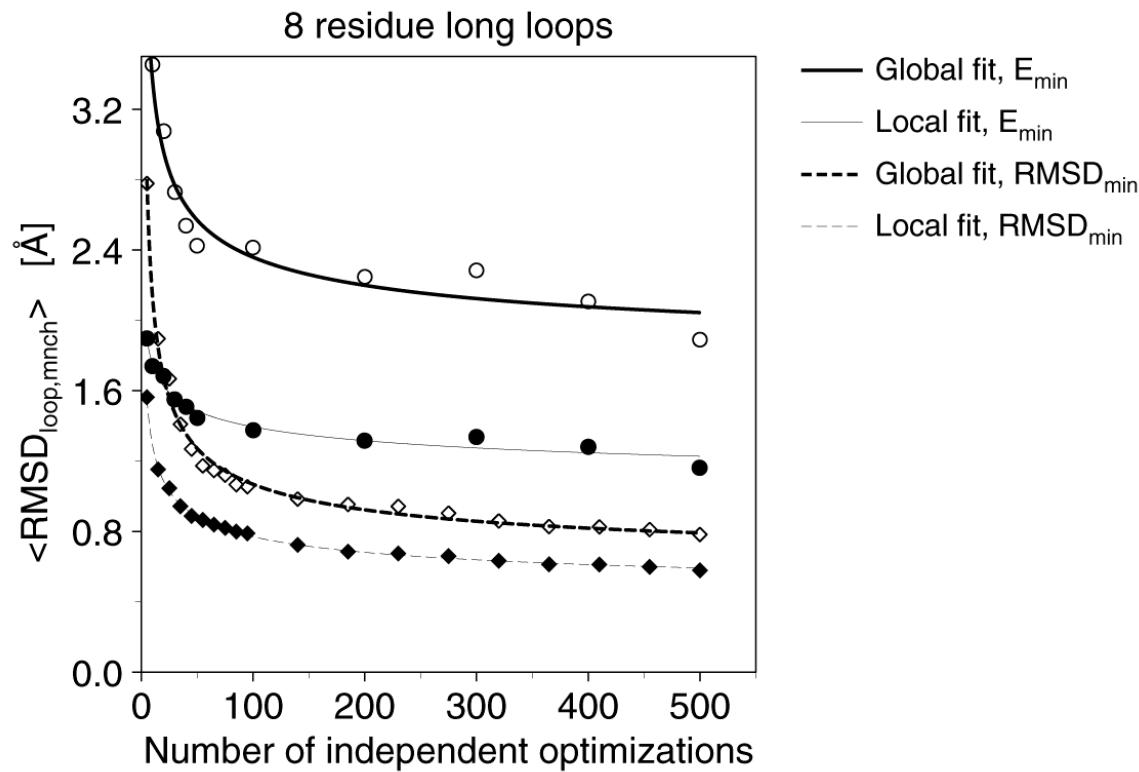
Optimization of Objective Function



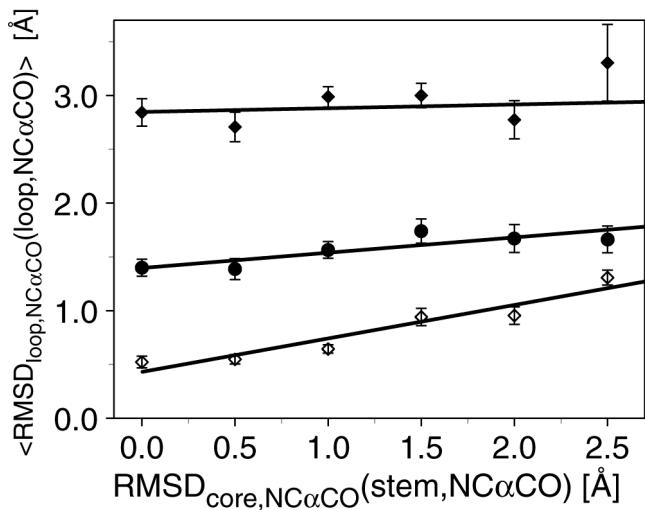
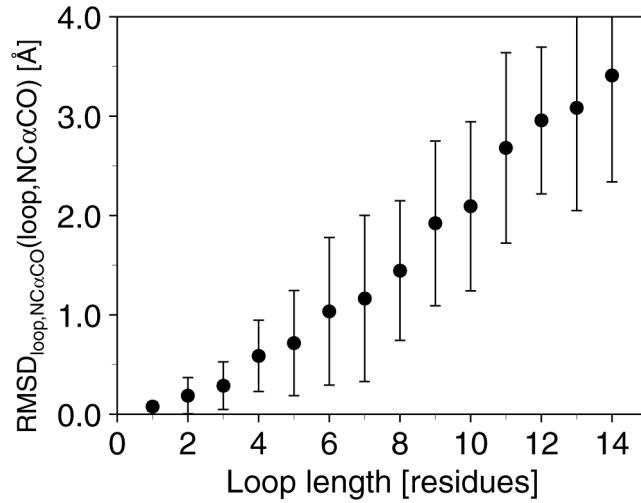
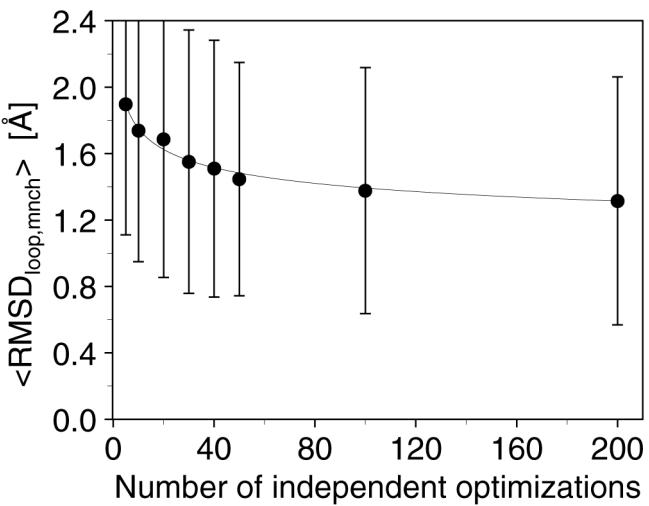
Calculating an Ensemble of Loop Models



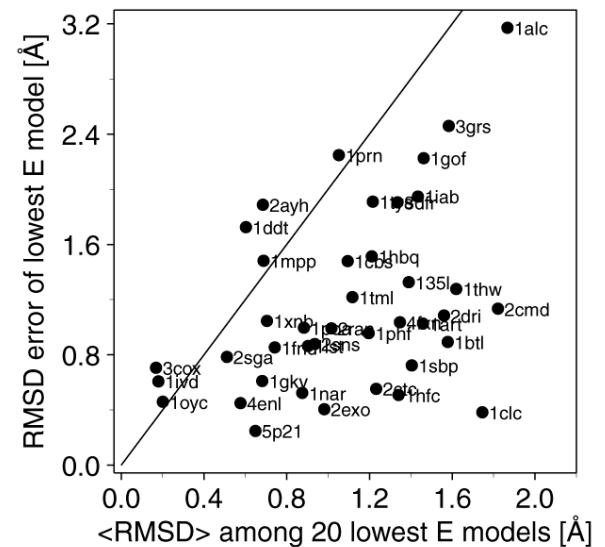
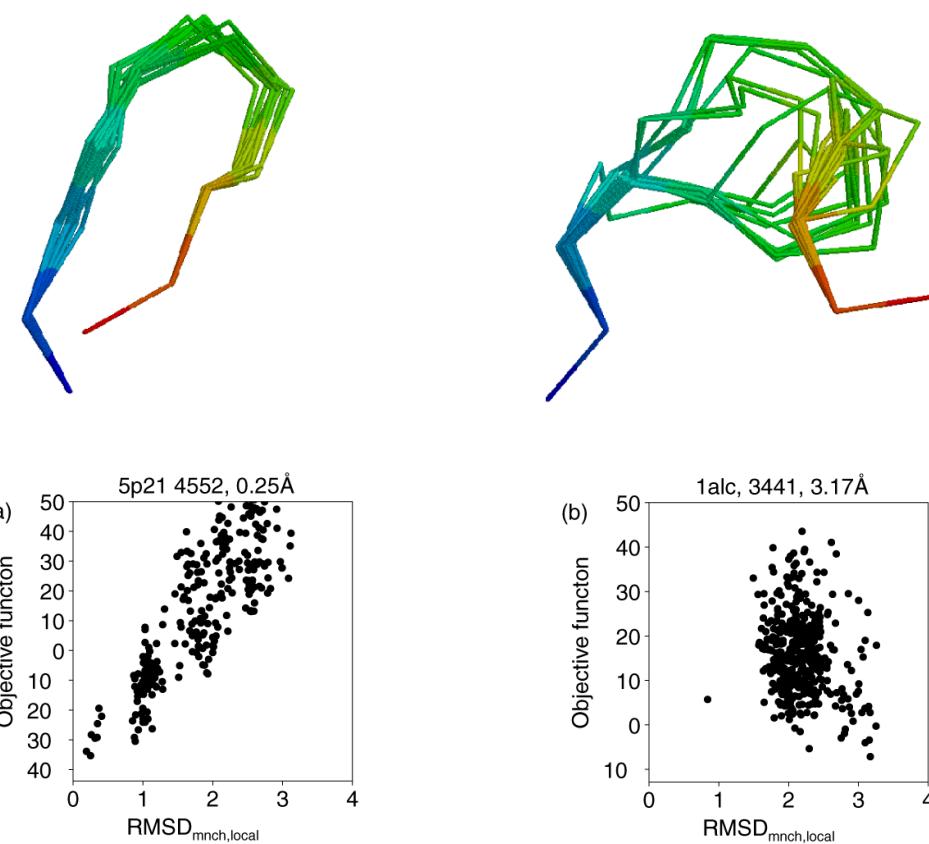
Accuracy of loop models as a function of amount of optimization



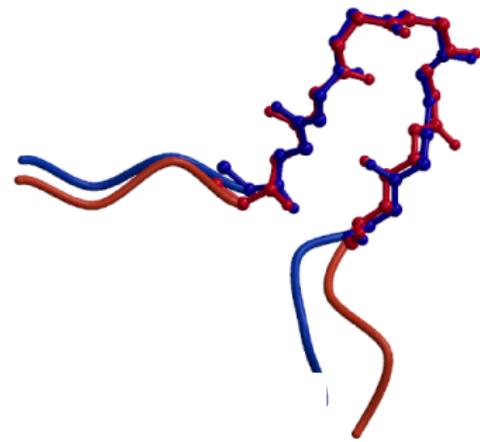
Accuracy of loop models



Assessing Accuracy of Loop Models

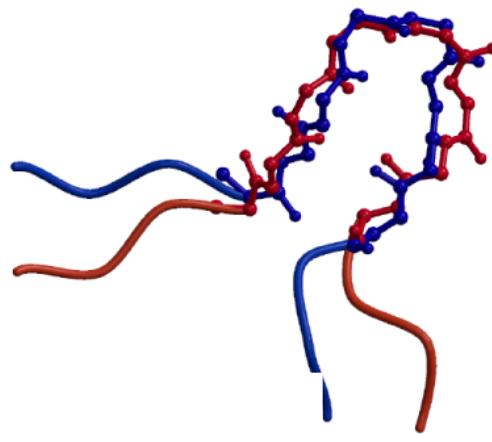


Accuracy of Loop Modeling



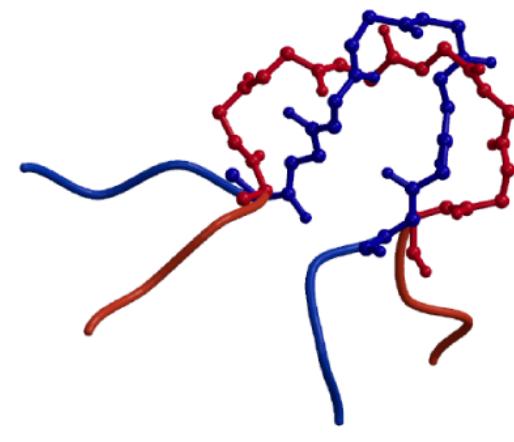
RMSD=0.6Å

HIGH ACCURACY (<1Å)



RMSD=1.1Å

MEDIUM ACCURACY (<2Å)



RMSD=2.8Å

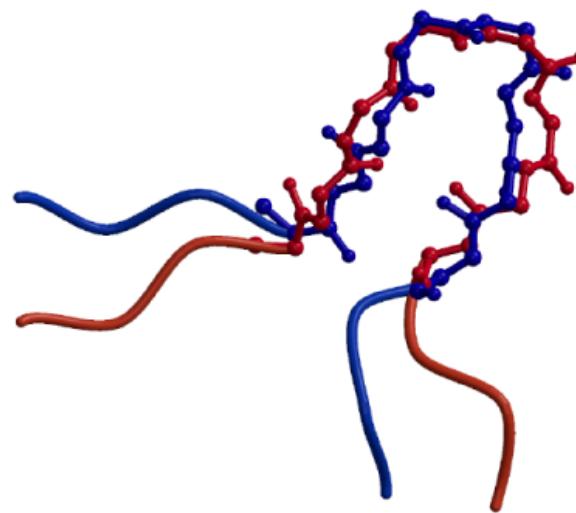
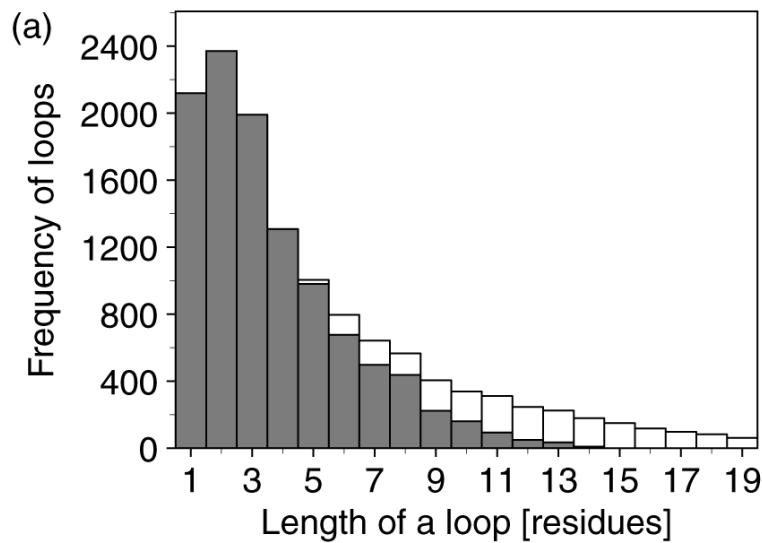
LOW ACCURACY (>2Å)

50% (30%) of
8-residue loops

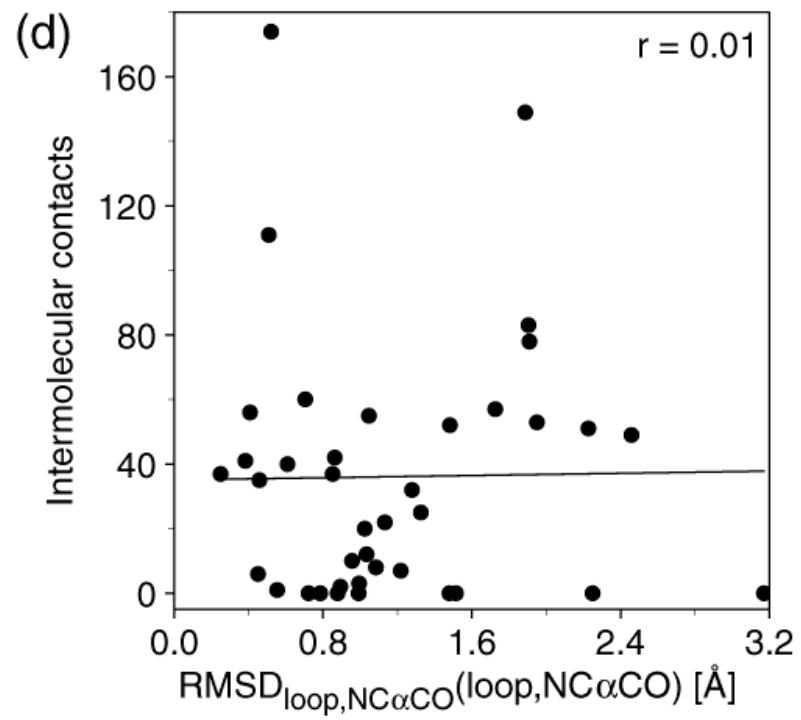
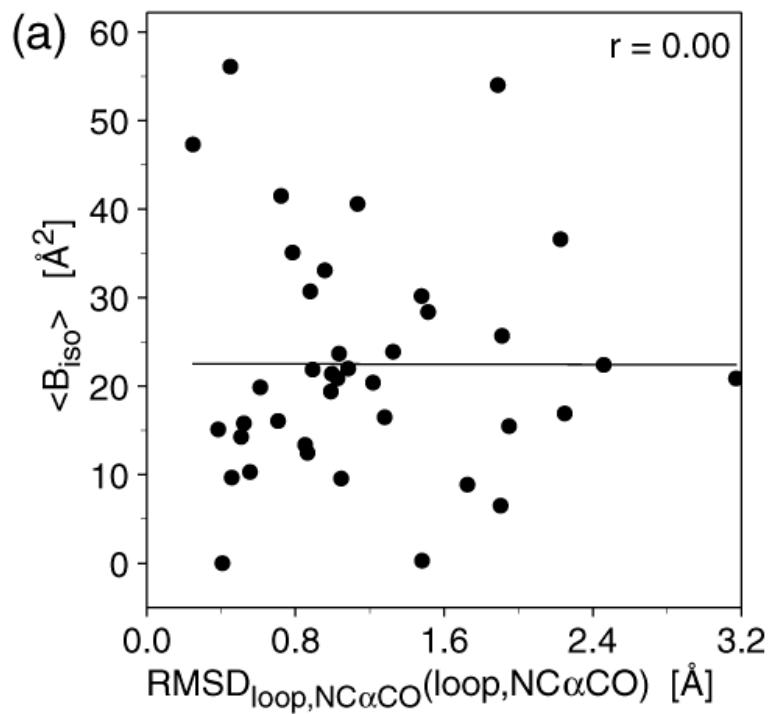
40% (48%) of
8-residue loops

10% (22%) of
8-residue loops

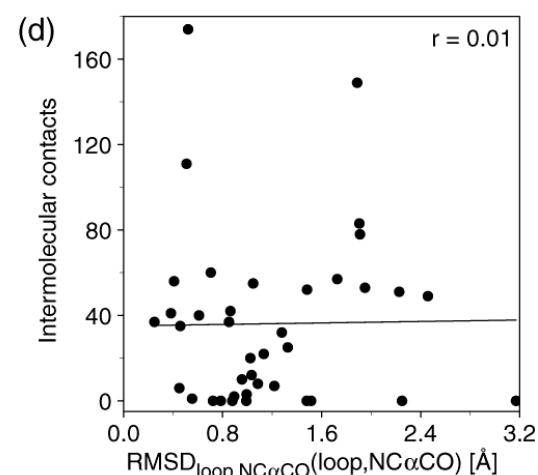
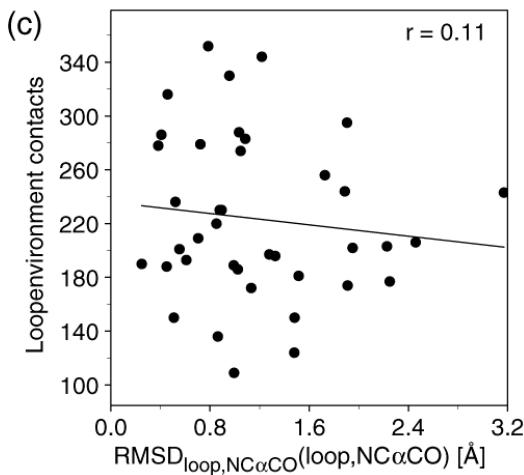
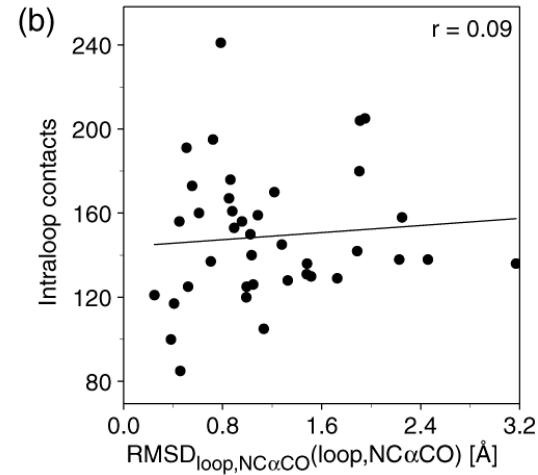
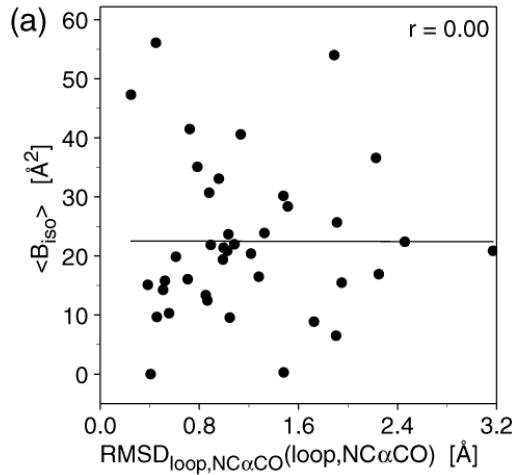
Fraction of Loops Modeled With at Least Medium Accuracy



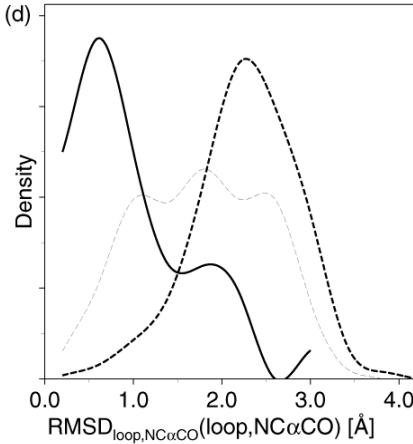
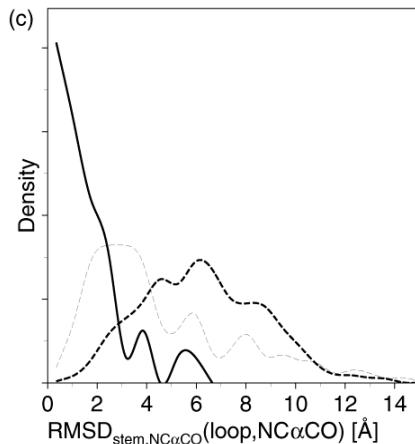
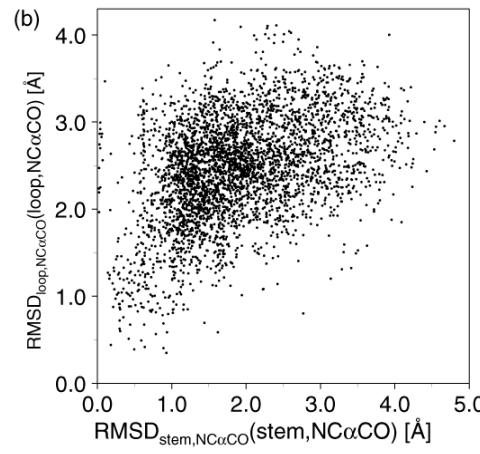
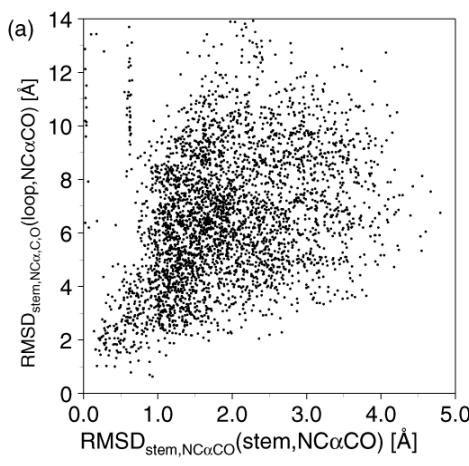
Accuracy of Loop Modeling as a Function of Loop Properties



Accuracy of Loop Modeling as a Function of Loop Properties

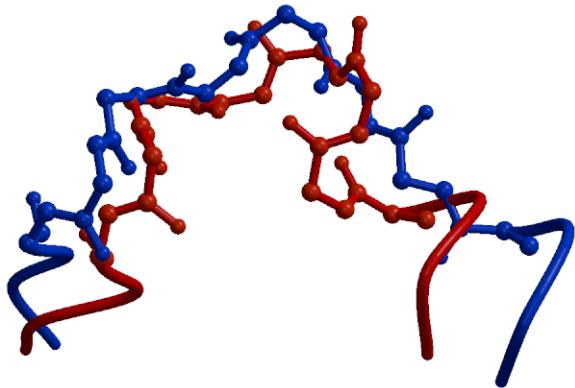


Comparison of the Loop Modeling Errors With Reference RMSD Distributions



Problems in Practical Loop Modeling

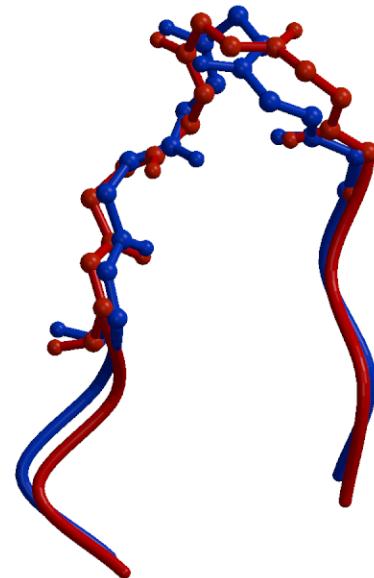
1. Decide which regions to model as loops.
2. Correct alignment of anchor regions & environment.
3. Modeling of a loop.



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

Modeling genes

Identification and characterization of a p53 homologue in *Drosophila melanogaster*

Shengkan Jin^{*2}, Sebastian Martinek^{2,3}, Woo S. Joo[§], Jennifer R. Wortman[¶], Nebojsa Mirkovic[¶], Andrej Sali[¶], Mark D. Yandell[¶], Nikola P. Pavletich[§], Michael W. Young³, and Arnold J. Levine^{*,**}

PNAS 97, 7301, 2000.

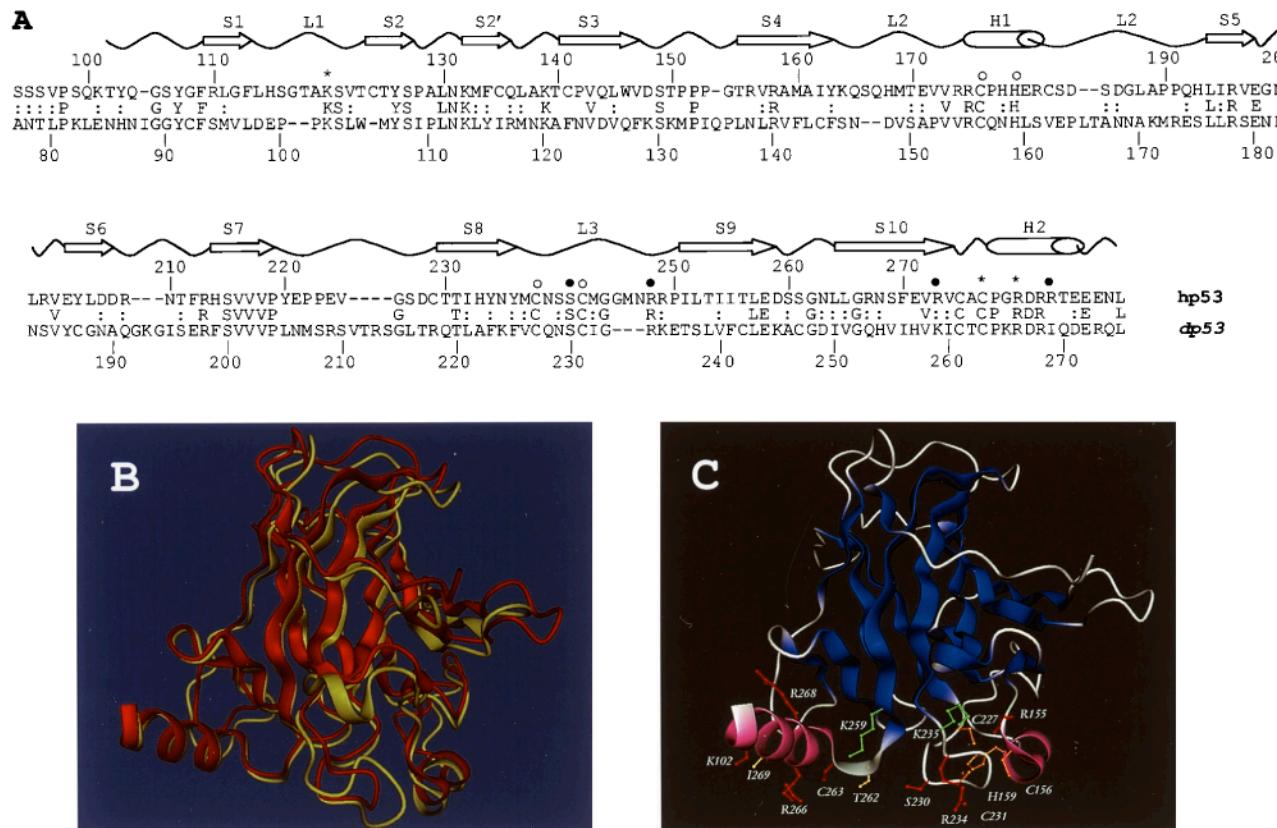


Fig. 1. (A) Sequence and structure comparison of the *Drosophila* and human p53 DNA-binding domains. Sequence alignment of the dp53 and hp53 DNA-binding domains as produced by PSI-BLAST. The secondary structure elements of hp53 are shown above (S, *b*-strand; L, loop; H, *a*-helix). Residues involved in DNA binding (*, contacting bases; F, contacting phosphate backbone) and zinc binding (E) also are indicated (6). (B) Superimposition of the crystal structure of hp53 (yellow cartoon) DNA-binding domain and the model of the dp53 domain (red cartoon) predicted by program MODELLER. (C) Protein structure model of the dp53 DNA-binding domain. Color scheme: red, residues preserved between the human and *Drosophila* sequences; green, conservative substitutions; orange, preserved Zn-coordinating residues; and yellow, nonconservative substitutions. *B* and *C* were rendered by program DINO (<http://ywww.biozentrum.unibas.ch/>; x-raydino).

Fly has p53

- purified dp53 DNA binding domain binds to the hp53 consensus binding site;
- a mutant dp53 exerted the same dominant negative effect on transactivation as its human counterpart;
- ectopic expression of dp53 in *Dm* eye disc caused cell death;
- dp53 expression pattern in the course of *Dm* lifespan is similar to that of hp53.

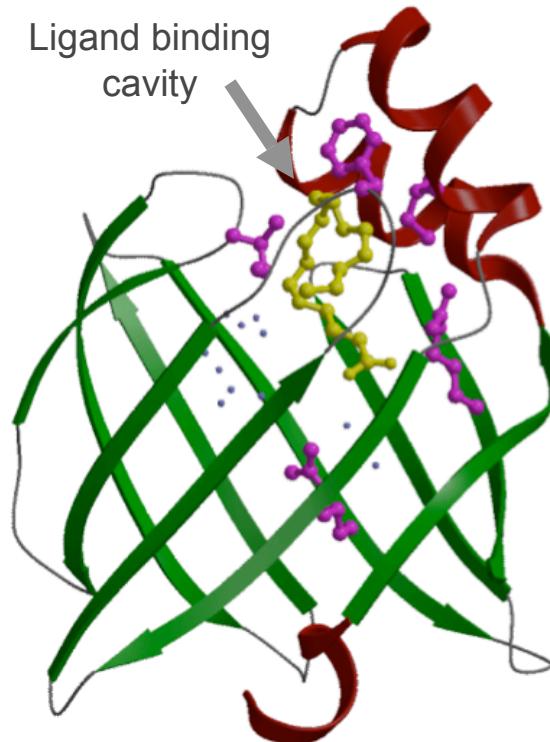
Dm may provide a convenient and simpler model genetic system in which to study p53.

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

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

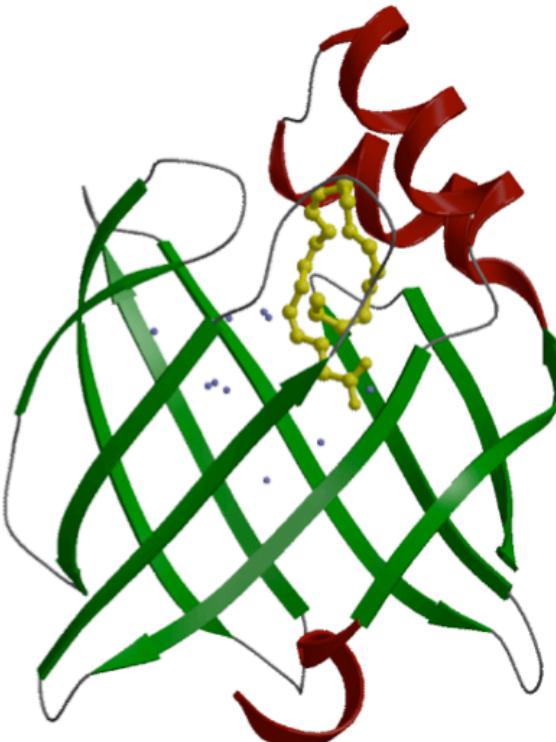
BLBP/oleic acid

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.

Structural analysis of missense mutations in human BRCA1 BRCT domains

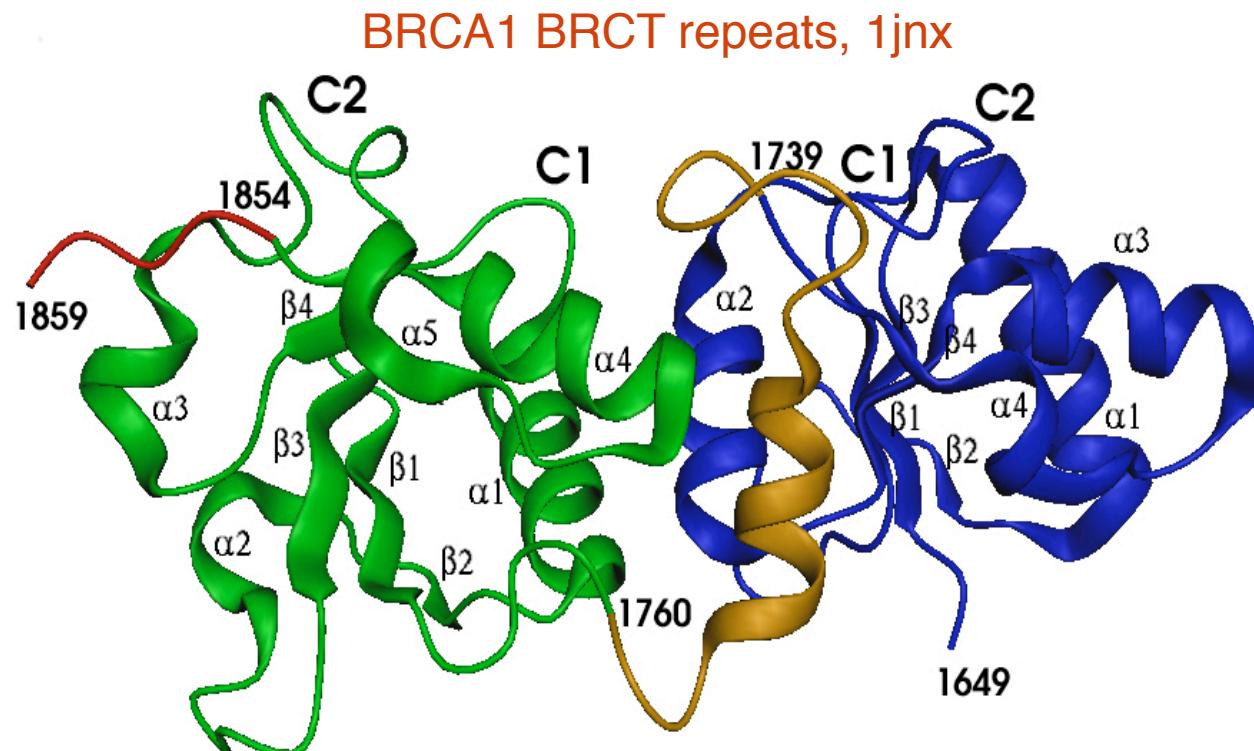
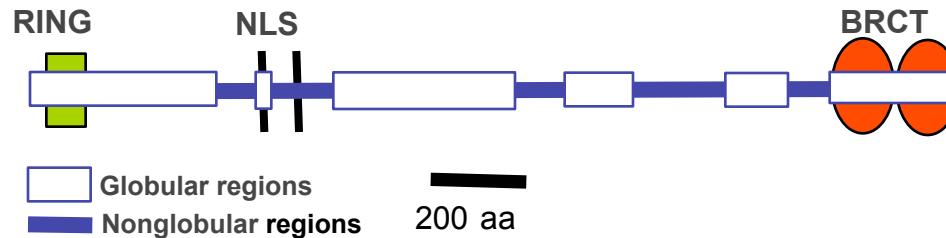
Nebojsa Mirkovic, Marc A. Marti-Renom, Barbara L. Weber,
Andrej Sali and Alvaro N.A. Monteiro

Cancer Research (June 2004). 64:3790-97

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.



Human BRCA1 and its two BRCT domains



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

CONFIDENTIAL



BRACAnalysis™

Comprehensive BRCA1-BRCA2 Gene Sequence Analysis Result

Nieco Singer, MS Strang Cancer Prevention Center 428 E 72nd St New York, NY 10021	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 #: 56594
--------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------

Physician: Fred Gilbert, MD

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.

CONFIDENTIAL



BRACAnalysis™
Comprehensive BRCA1-BRCA2 Gene Sequence Analysis Result

Nicole Singer, MS Strang Cancer Prevention Center 428 E 72nd St New York, NY 10021	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 #: 56594
Physician: Fred Gilbert, MD		

Test Result

Gene Analyzed	Specific Genetic Variant
BRCA2	H2116R
BRCA1	None Detected

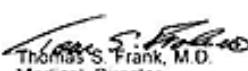
Interpretation

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

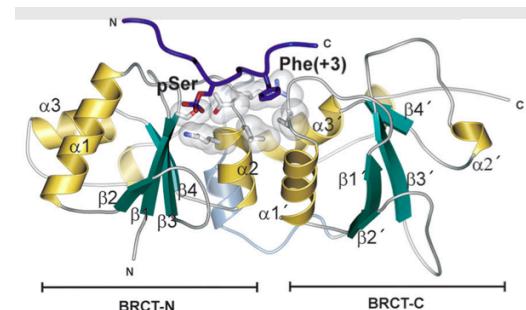
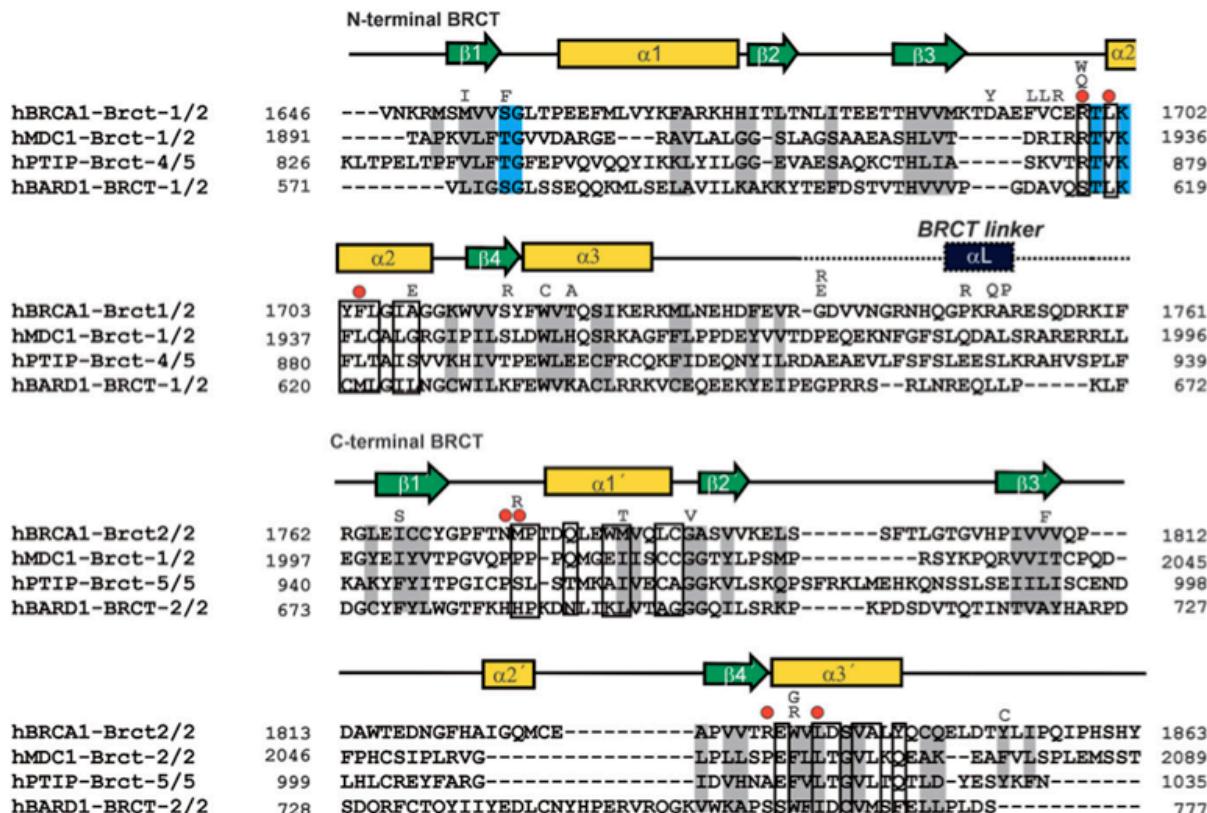

Thomas S. Frank, M.D.
Medical Director

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 Y1703HF1 704S	L1705PS1 715NS172 2FF1734L G1738EG 1743RA17 52PF1761I	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

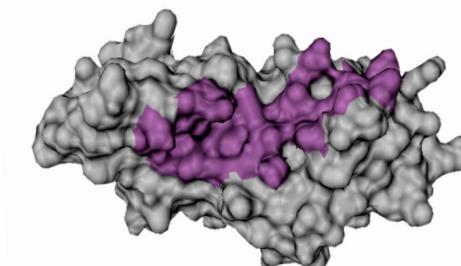
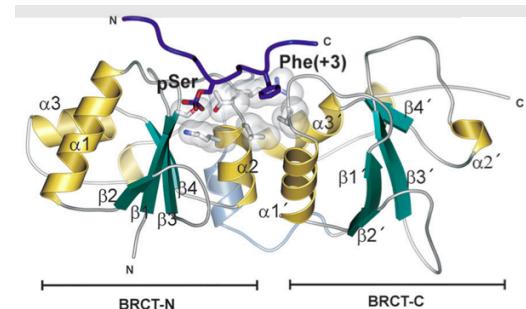
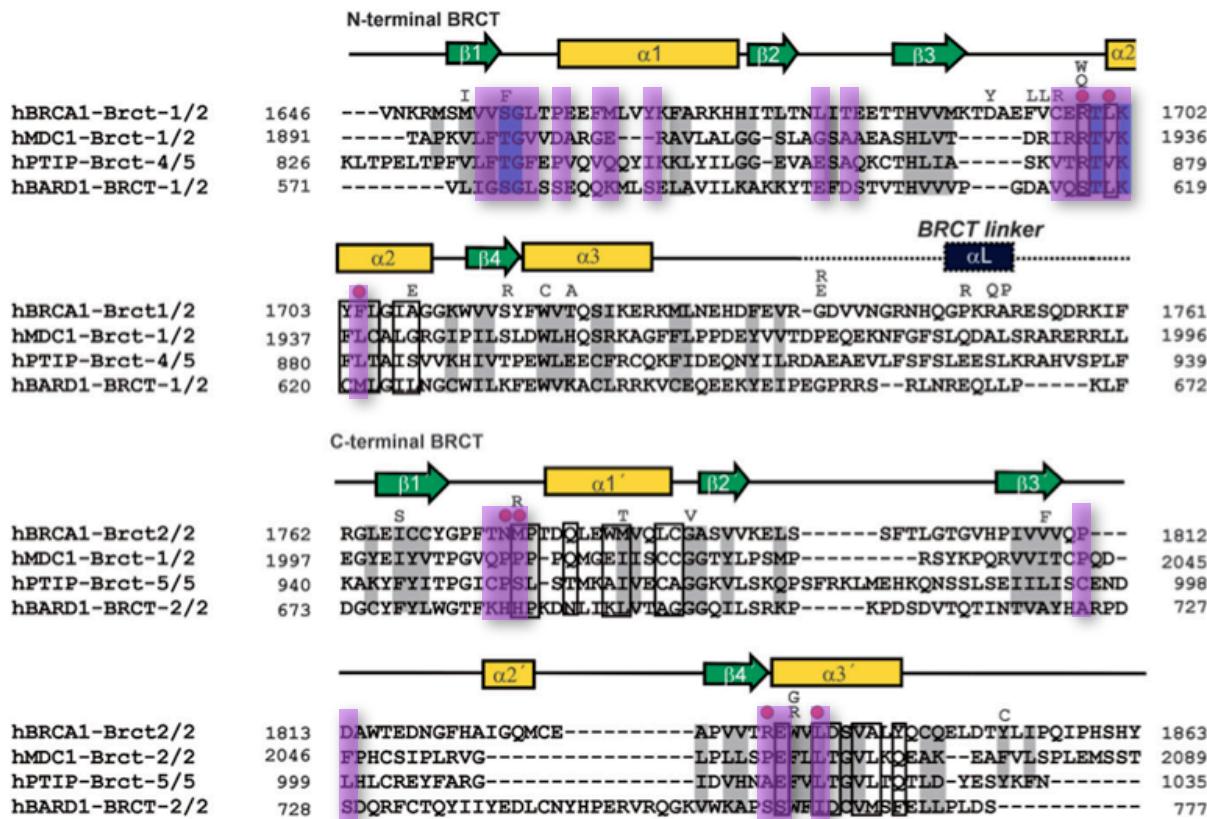
Putative binding site on BRCA1



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

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

Putative binding site on BRCA1



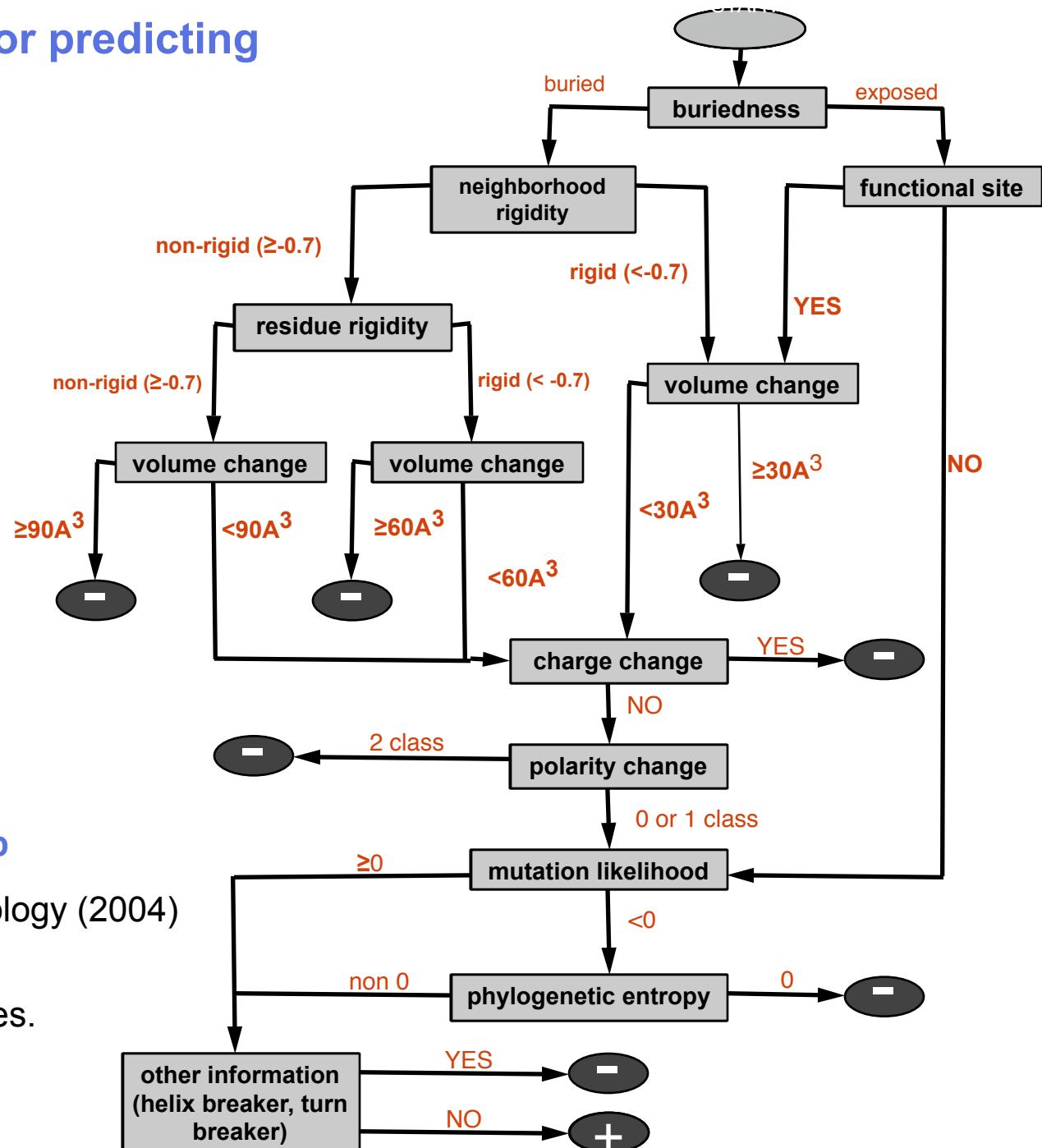
Putative binding site predicted in 2003
and accepted for publication on March 2004.

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

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

“Decision” tree for predicting

functional impact
of genetic
variants



<http://salilab.org/snpweb>

Mirkovic et al., Cancer Biology (2004)
64:3790-97

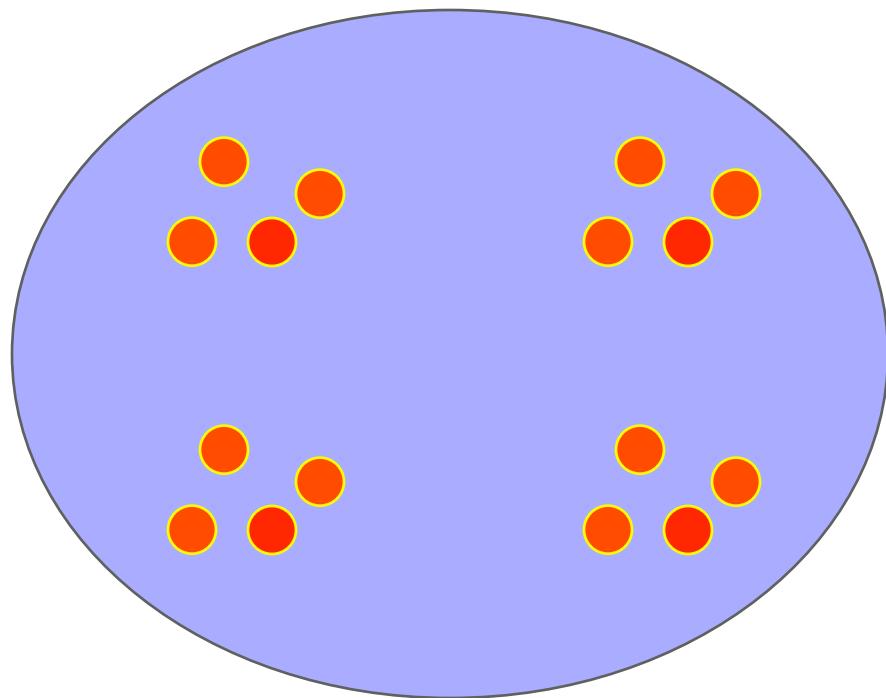
Eswar et al. Nucl.Acids Res.
31, 3375, 2003.

Modeling genomes

Structural Genomics

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.

Characterize most protein sequences based on related known structures.



The number of “families” is much smaller than the number of proteins.

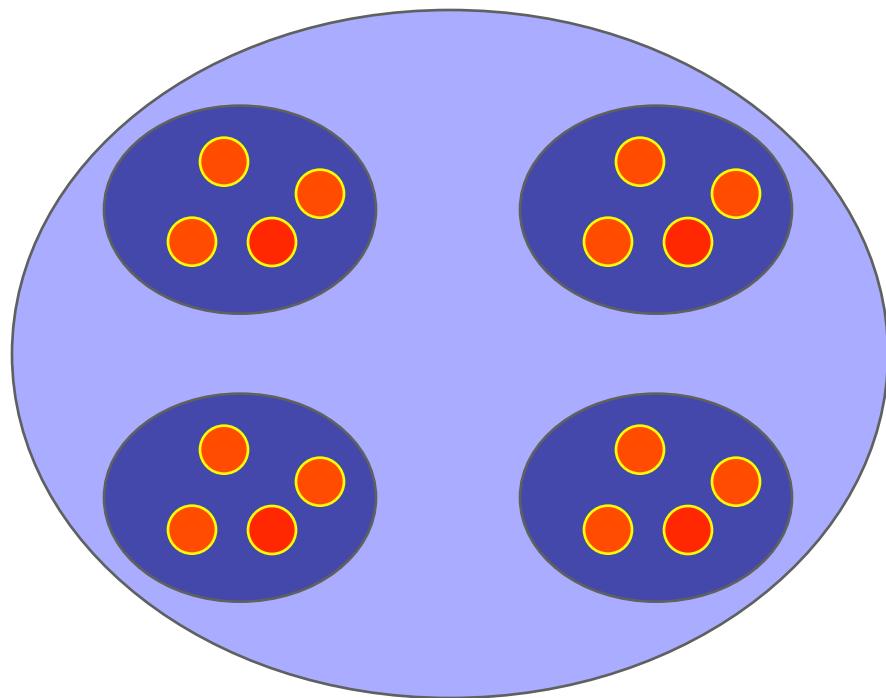
Any one of the members of a family is fine.

There are ~16,000 30% seq id families (90%)
(Vitkup et al. *Nat. Struct. Biol.* **8**, 559, 2001).

Structural Genomics

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.

Characterize most protein sequences based on related known structures.



The number of “families” is much smaller than the number of proteins.

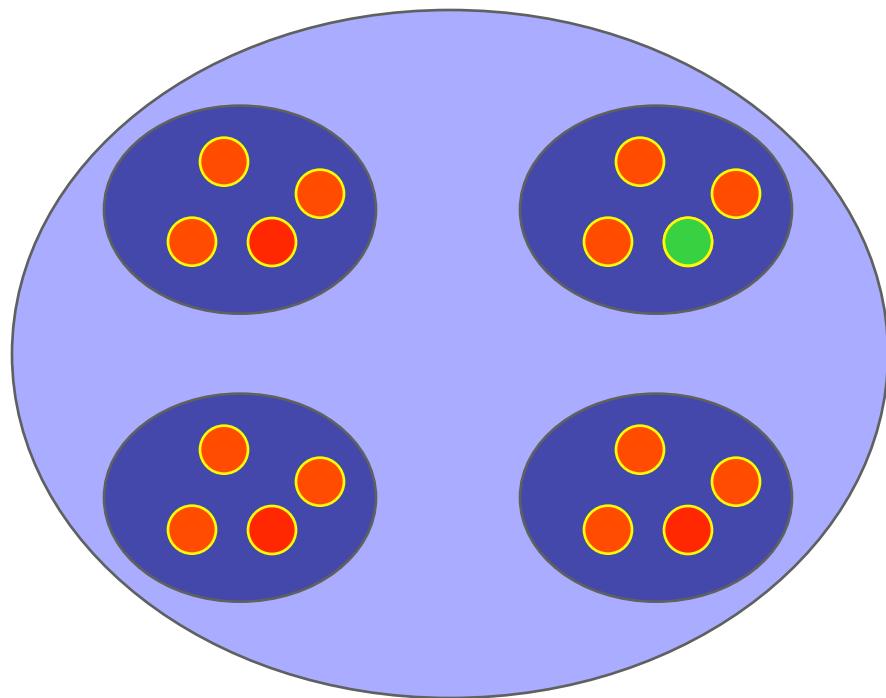
Any one of the members of a family is fine.

There are ~16,000 30% seq id families (90%)
(Vitkup et al. *Nat. Struct. Biol.* **8**, 559, 2001).

Structural Genomics

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.

Characterize most protein **sequences** based on related known structures.



The number of “families” is much smaller than the number of proteins.

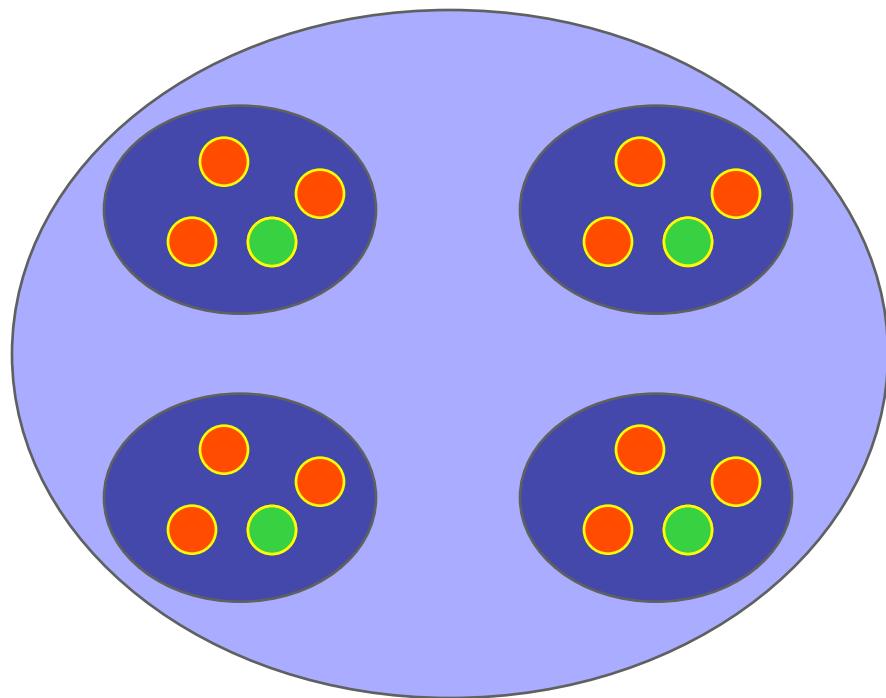
Any one of the members of a family is fine.

There are ~16,000 30% seq id families (90%)
(Vitkup et al. *Nat. Struct. Biol.* **8**, 559, 2001).

Structural Genomics

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.

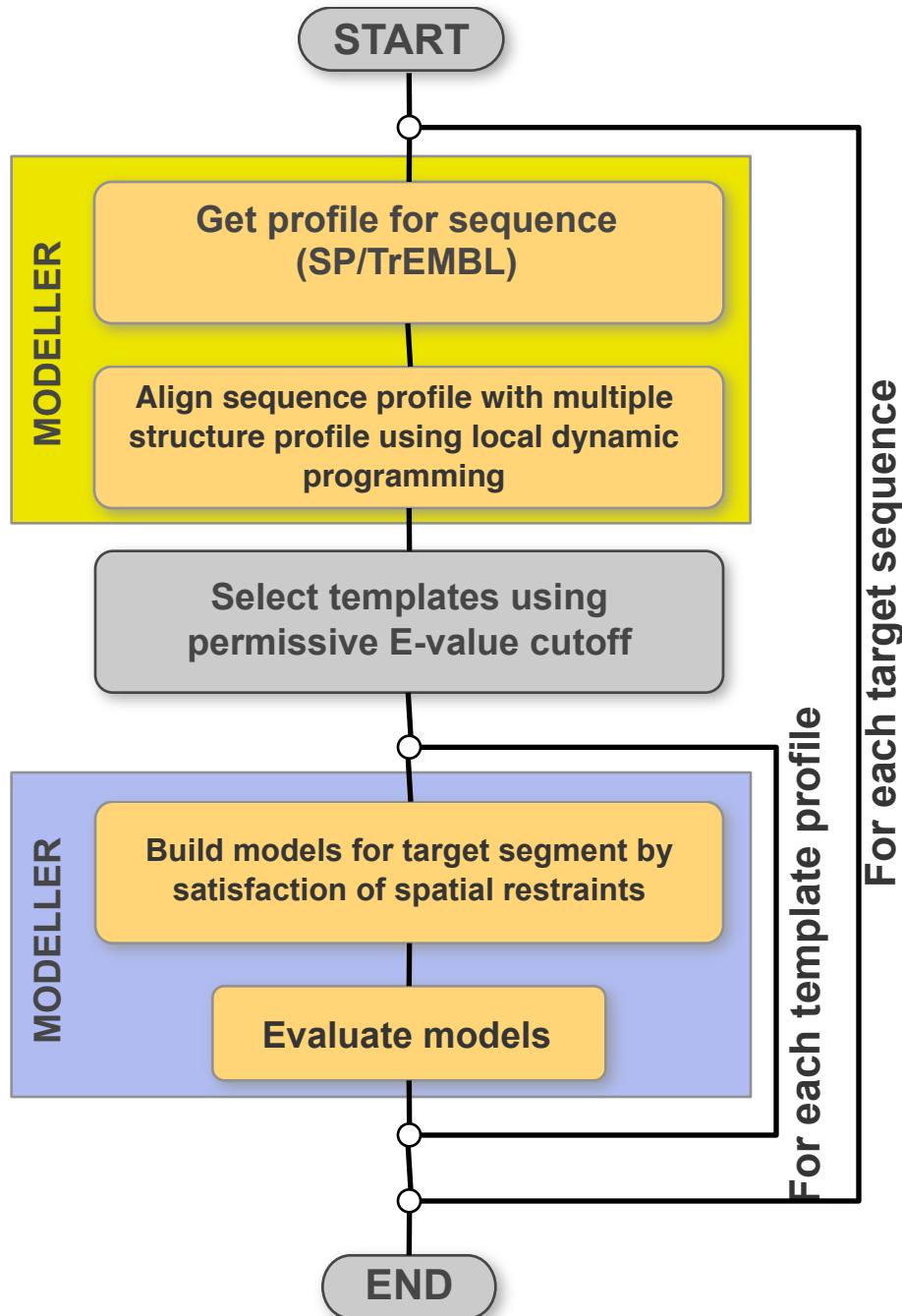
Characterize most protein **sequences** based on related known structures.



The number of “families” is much smaller than the number of proteins.

Any one of the members of a family is fine.

There are ~16,000 30% seq id families (90%)
(Vitkup et al. *Nat. Struct. Biol.* **8**, 559, 2001).



MODPIPE: Automated Large-Scale Comparative Modeling

R. Sánchez & A. Šali, *Proc. Natl. Acad. Sci. USA* 95, 13597, 1998.

Eswar *et al.* *Nucl. Acids Res.* 31, 3375–3380, 2003.

Pieper *et al.*, *Nucl. Acids Res.* 32, 2004.

N. Eswar, M. Marti-Renom, M.S. Madhusudhan, B. John, A. Fiser, R. Sánchez, F. Melo, N. Mirkovic, B. Webb, M.-Y. Shen, A. Šali.

Synergy of crystallography and comparative modeling in structural genomics

Pieper et al., *Nucl. Acids Res.* 32, 2004.

http://salilab.org/modbase/models_nsgxrc.html

NYSGXRC X-ray Structure			MODBASE Models			
PDB Code	Database Accession Number	Annotation	Total Sequences	Fold & Model	Fold	Model
1b54	P38197	Hypothetical UPF0001 protein YBL036C	151	132	2	17
1f89	P49954	Hypothetical 32.5 kDa protein YLR351C	553	488	55	10
1njr	Q04299	Hypothetical 32.1 kDa protein in ADH3-RCA1 intergenic region	4	1	0	3
1nkq	P53889	Hypothetical 28.8 kDa protein in PSD1-SKO1 intergenic region	379	207	172	0
1jzt	P40165	Hypothetical 27.5 kDa protein in SPX19-GCR2 intergenic region	1058	39	1006	13
1jr7	P76621	Hypothetical protein ygaT	11	10	0	1
		YF63_METJA hypothetical protein				05/10/2004

Comparative modeling of the TrEMBL database

Unique sequences processed: 1,182,126

Sequences with fold assignments or models: 659,495 (56%)

70% of models based on <30% sequence identity to template.

On average, only a domain per protein is modeled
(an “average” protein has 2.5 domains of 175 aa).



Database of Comparative Protein Structure Models

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

About MODBASE

ModBase moved to UCSF.
You have to login (academic or user logins)
in order to reset your cookies.
Sorry for the inconvenience!

Some datasets are not yet fully restored. We are working on it.

General Information

Glossary

Authors and acknowledgements

SUMMARY Search Criteria
Keyword Table

23 matches were found using the specified search criteria. Click on the links in the table header to resort your output.

TARGET		MODEL						
Model/Fold Reliability	Sequence based View	Select Sequence Database Links	Database Description	Organism	Protein Size	Modeled Segments	Size (%)	Reliability
●	●	● TR Q8ZQZL	L-arabinose binding periplasmic protein precursor Dataset: SPTR-2003 2EAM PDBID: 1M0BL	Escherichia coli O8	346	42-346	305	98
●	●	● TR Q8ZQZL	L-arabinose binding periplasmic protein precursor Dataset: SPTR-2003 2EAM PDBID: 1M0BL	Vibrio parahaemolyticus	333	34-333	300	76.00

MODBASE Contents

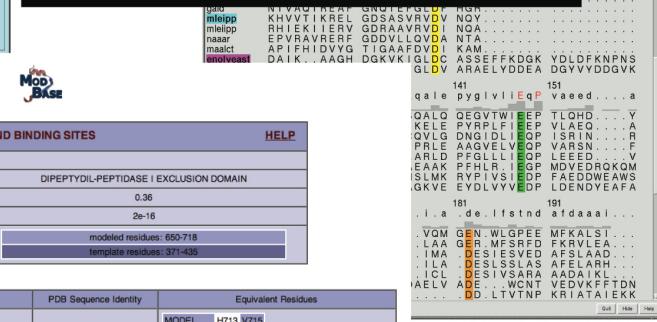
1,261,962 Reliable Models or PSI-BLAST
Fold Assignments for domains in 659,316
proteins. Last Update on 09/14/03. ModBASE



Database of Comparative Protein Structure Models

User: ModWeb0-1070773001,NYSGRC,Academic User
[Change User](#)

[SEARCH for Models](#) [SEARCH for Sequences](#)



TARGET		MODEL SUMMARY			
Sequence based View	Select Sequence Database Links	Database Description	Organism	Protein Size	Modeled Segments-Schemas
●	● TR Q8TY9S	SERA ANTIGEN/PAPAIN-LIKE PROTEASE WITH ACTIVE SER	Plasmodium falciparum	997	
		Dataset: PFAM PRODOM			

TARGET		MODEL SUMMARY			
Sequence based View	Select Sequence Database Links	Database Description	Organism	Protein Size	Modeled Segments-Schemas
●	● TR Q8TY9S	P putative sugar ABC transporter sugar-binding protein Dataset: SPTR-2003 2EAM PDBID: 1M0BL	Streptomyces avermitilis	370	43-362 320 15.00 78-31 1.00 8400
					3-301

TARGET		MODEL SUMMARY			
Sequence based View	Select Sequence Database Links	Database Description	Organism	Protein Size	Modeled Segments-Schemas
●	● TR Q8TY9S	SERA ANTIGEN/PAPAIN-LIKE PROTEASE WITH ACTIVE SER	Plasmodium falciparum	997	
		Dataset: PFAM PRODOM			

(406 Residues)

149 - 406

258 97.00 2e-7;

1qmc_B (1-258) CELL DIVISION PROTEIN KINASE 2: 1-7 - CATH 1.10.472.10.3.1.2 (57%)

Subset: SP/TR-2001

Model/Reliability

Seq Id

Size (%)

E-value

Model Data

Fold/Model Reliability

Template

1k3b_C

CL

500

CHLORIDE ION

Template PDB

1H73 V715

Template ID

I429 V431

L-arabinose-binding periplasmic protein precursor

Dataset: SPTR-2003 2EAM PDBID: 1M0BL

Organism: Escherichia coli O8

Protein Size: 346

Modeled Segments: 42-346

Size (%): 98

E-value: 1.0E-03

Reliability: 98

Model/Fold Reliability: 98

Sequence based View: 98

Select Sequence Database Links: 98

Database Description: 98

Organism: 98

Protein Size: 98

Modeled Segments-Schemas: 98

Size (%): 98

E-value: 98

Model Data: 98

Fold/Model Reliability: 98

Template: 98

1k3b_C: 98

CL: 98

500: 98

CHLORIDE ION: 98

Template PDB: 98

1H73: 98

V715: 98

I429: 98

V431: 98

L-arabinose-binding periplasmic protein precursor: 98

Dataset: SPTR-2003 2EAM PDBID: 1M0BL: 98

Organism: Escherichia coli O8: 98

Protein Size: 346: 98

Modeled Segments-Schemas: 98: 98

Size (%): 98: 98

E-value: 98: 98

Model Data: 98: 98

Fold/Model Reliability: 98: 98

Template: 98: 98

1k3b_C: 98: 98

CL: 98: 98

500: 98: 98

CHLORIDE ION: 98: 98

Template PDB: 98: 98

1H73: 98: 98

V715: 98: 98

I429: 98: 98

V431: 98: 98

L-arabinose-binding periplasmic protein precursor: 98: 98

Dataset: SPTR-2003 2EAM PDBID: 1M0BL: 98: 98

Organism: Escherichia coli O8: 98: 98

Protein Size: 346: 98: 98

Modeled Segments-Schemas: 98: 98: 98

Size (%): 98: 98: 98

E-value: 98: 98: 98

Model Data: 98: 98: 98

Fold/Model Reliability: 98: 98: 98

Template: 98: 98: 98

1k3b_C: 98: 98: 98

CL: 98: 98: 98

500: 98: 98: 98

CHLORIDE ION: 98: 98: 98

Template PDB: 98: 98: 98

1H73: 98: 98: 98

V715: 98: 98: 98

I429: 98: 98: 98

V431: 98: 98: 98

L-arabinose-binding periplasmic protein precursor: 98: 98: 98

Dataset: SPTR-2003 2EAM PDBID: 1M0BL: 98: 98: 98

Organism: Escherichia coli O8: 98: 98: 98

Protein Size: 346: 98: 98: 98

Modeled Segments-Schemas: 98: 98: 98: 98

Size (%): 98: 98: 98: 98

E-value: 98: 98: 98: 98

Model Data: 98: 98: 98: 98

Fold/Model Reliability: 98: 98: 98: 98

Template: 98: 98: 98: 98

1k3b_C: 98: 98: 98: 98

CL: 98: 98: 98: 98

500: 98: 98: 98: 98

CHLORIDE ION: 98: 98: 98: 98

Template PDB: 98: 98: 98: 98

1H73: 98: 98: 98: 98

V715: 98: 98: 98: 98

I429: 98: 98: 98: 98

V431: 98: 98: 98: 98

L-arabinose-binding periplasmic protein precursor: 98: 98: 98: 98

Dataset: SPTR-2003 2EAM PDBID: 1M0BL: 98: 98: 98: 98

Organism: Escherichia coli O8: 98: 98: 98: 98

Protein Size: 346: 98: 98: 98: 98

Modeled Segments-Schemas: 98: 98: 98: 98: 98

Size (%): 98: 98: 98: 98: 98

E-value: 98: 98: 98: 98: 98

Model Data: 98: 98: 98: 98: 98

Fold/Model Reliability: 98: 98: 98: 98: 98

Template: 98: 98: 98: 98: 98

1k3b_C: 98: 98: 98: 98: 98

CL: 98: 98: 98: 98: 98

500: 98: 98: 98: 98: 98

CHLORIDE ION: 98: 98: 98: 98: 98

Template PDB: 98: 98: 98: 98: 98

1H73: 98: 98: 98: 98: 98

V715: 98: 98: 98: 98: 98

I429: 98: 98: 98: 98: 98

V431: 98: 98: 98: 98: 98

L-arabinose-binding periplasmic protein precursor: 98: 98: 98: 98: 98

Dataset: SPTR-2003 2EAM PDBID: 1M0BL: 98: 98: 98: 98: 98

Organism: Escherichia coli O8: 98: 98: 98: 98: 98

Protein Size: 346: 98: 98: 98: 98: 98

Modeled Segments-Schemas: 98: 98: 98: 98: 98: 98

Size (%): 98: 98: 98: 98: 98: 98

E-value: 98: 98: 98: 98: 98: 98

Model Data: 98: 98: 98: 98: 98: 98

Fold/Model Reliability: 98: 98: 98: 98: 98: 98

Template: 98: 98: 98: 98: 98: 98

1k3b_C: 98: 98: 98: 98: 98: 98

CL: 98: 98: 98: 98: 98: 98

500: 98: 98: 98: 98: 98: 98

CHLORIDE ION: 98: 98: 98: 98: 98: 98

Template PDB: 98: 98: 98: 98: 98: 98

1H73: 98: 98: 98: 98: 98: 98

V715: 98: 98: 98: 98: 98: 98

I429: 98: 98: 98: 98: 98: 98

V431: 98: 98: 98: 98: 98: 98

L-arabinose-binding periplasmic protein precursor: 98: 98: 98: 98: 98: 98

Dataset: SPTR-2003 2EAM PDBID: 1M0BL: 98: 98: 98: 98: 98: 98

Organism: Escherichia coli O8: 98: 98: 98: 98: 98: 98

Protein Size: 346: 98: 98: 98: 98: 98: 98

Modeled Segments-Schemas: 98: 98: 98: 98: 98: 98: 98

Size (%): 98: 98: 98: 98: 98: 98: 98

E-value: 98: 98: 98: 98: 98: 98: 98

Model Data: 98: 98: 98: 98: 98: 98: 98

Fold/Model Reliability: 98: 98: 98: 98: 98: 98: 98

Template: 98: 98: 98: 98: 98: 98: 98

1k3b_C: 98: 98: 98: 98: 98: 98: 98

CL: 98: 98: 98: 98: 98: 98: 98

500: 98: 98: 98: 98: 98: 98: 98

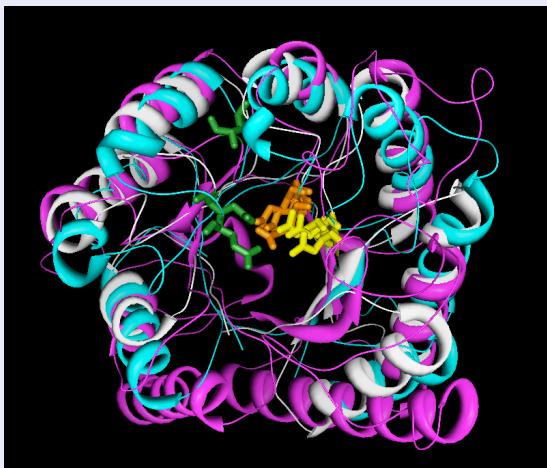
CHLORIDE ION: 98: 98: 98: 98: 98: 98: 98

Template PDB: 98: 98: 98: 98: 98: 98:

Major bidirectional resources involving ModBase

UCSF CHIMERA

an Extensible Molecular Modeling System



Conservation
DOI: 10.1101/003108
APE_BOVIN MWWVAVYVVA ALLLGGRRGQI
APE_CAVPO MKVLWAALLVV TLLAGGRADV
C60940 Y13652 MHSIVVFFAL AAVLTGQARS

51	R FWD Y L RWVQ	61	TLS d QVQEEEL
71	I s a QVTQELT	81	aLM e t TMKEV
91	KAYK a ELE e Q		

Conserve DOI: 10.1101/003108
APE_BOVIN R FWD Y L RWVQ TLSDQVQEEEL
APE_PIG R FWD Y L RWVQ TLSDQVQEEEL
APE_RAT R FWD Y L RWVQ TLSDQVQEEEL
APE_RAT R FWD Y L RWVQ TLSDQVQEEEL
APE_PAPAN R FWD Y L RWVQ TLSEQVQEEEL
APE_PAPAN R FWD Y L RWVQ TLSEQVQEEEL
APE_CHUMAN R FWD Y L RWVQ TLSDQVQEEEL
APE_RABIT R FWD Y L RWVQ TLSDQVQEEEL
APE_CAVPO R FWD Y L RWVQ TLSDQVQEEEL
C60940 R FWD Y L RWVQ TLSDQVQEEEL
Y13652 R FWD Y VVSELN TQIDGMVQNI

Quit Hide Help

<http://www.cgl.ucsf.edu/chimera/>
Daniel Greenblatt, Conrad C. Huang,
Thomas E. Ferrin

ExPASy Home page Site Map Search ExPASy Contact us PROSITE Proteomics tools
Hosted by NCSC US Mirror sites: Bolivia Canada China Korea Switzerland Taiwan
Search Swiss-Prot/TreEMBL for P2Y2_BOVIN Go Clear

Swiss-Prot
Protein knowledgebase
TreEMBL
Computer-annotated supplement to Swiss-Prot

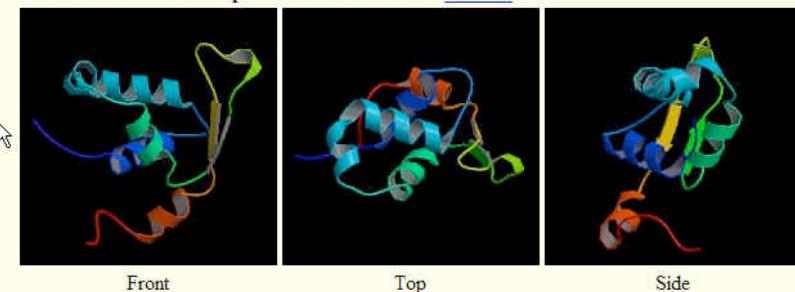
PS30262, G_PROTEIN_RECEP_F1_2; 1.
[Family / Alignment / Tree]
O18951
O18951
O18951
O18951
O18951
ModBase
O18951
SWISS-2DPAGE
Get region on 2D PAGE.

UCSC Human Gene Family Browser

Home genome Human assembly July 2003 search U39840 Go!
sort by Expression (GNF) configure filter (now off) display 50 output sequence text

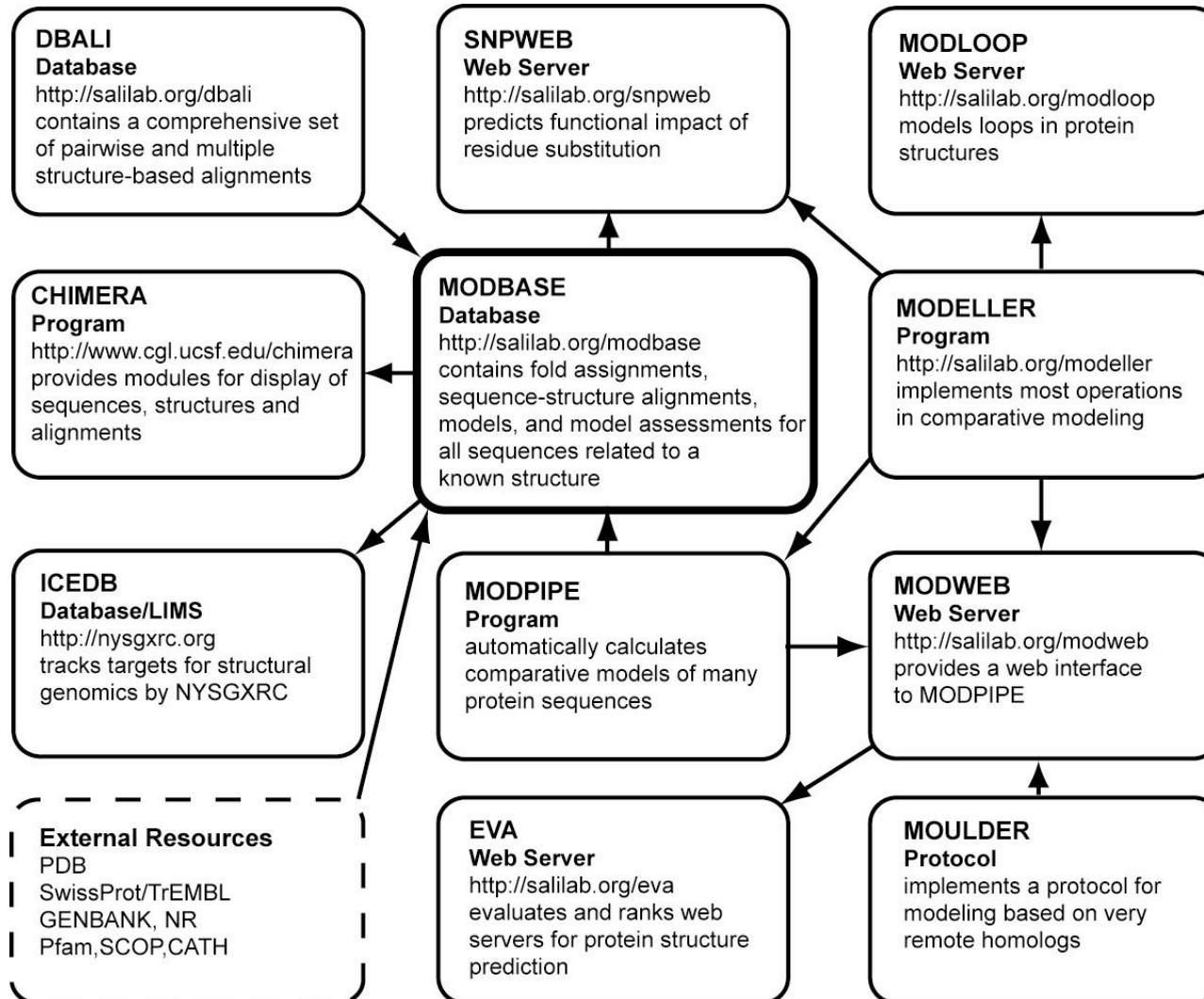
#	Name	cerebellum	wholeblood	thymus	pancreas	heart	trachea	kidney	liver	lung	tissue	E-Value	Genome	Position	Description
1	FOXA1	n/a										0	chr14	36,051,755	forkhead box A1

ModBase Predicted Comparative 3D Structure on P55317



MODBASE and associated resources

<http://salilab.org/>



Concluding remarks

- At present, useful 3D models can be obtained for domains in ~ 50% of the proteins (20% of domains).
- Completeness in structural coverage (structural genomics).
- Assembly of domains into higher order complexes.

BMC WorkShop

Protein Structure Prediction

model assessment (model building)

Marc A. Marti-Renom & Damien Devos

Department of Biopharmaceutical Sciences, UCSF

model assessment
.vs.
model evaluation

MODEL ASSESSMENT

prediction of the accuracy of methods and models

Model Assessment Methods

- Is the fold correct?
 - How correct is the overall structure?
 - What regions are modeled incorrectly?
 - What is the best model in the set of alternative models?
-
- Does the model satisfy the restraints used to calculate it?
 - What regions of the fold are variable?
 - Stereochemistry test (PROCHECK)
 - Residue environment test (Verify3D)
 - Statistical potential tests (PROSAIL, DFIRE, ANOLEA)
 - Other statistical tests, including tests with multiple criteria (GA341).

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
- Simple

Types:

- Contact potential
- Distance potentials
- Surface potentials

Approximations/Limitations in PMFs

Database size.

PMF versus Energy (additive/higher order terms).

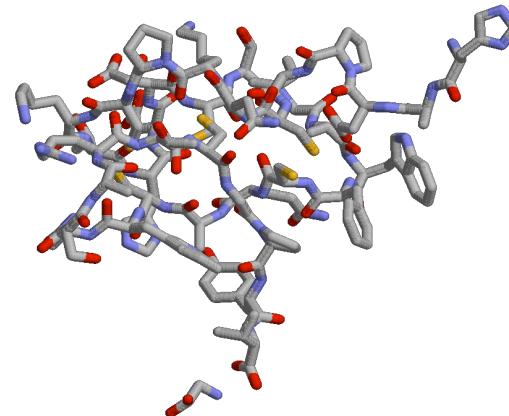
Reference state.

Physical origin.

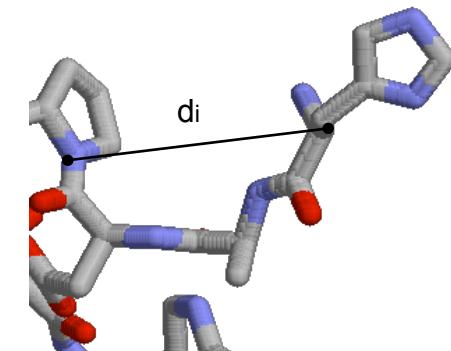
Representation Sequence/Structures

>gi42541361
MDIRSVSSLRGLLCLPPSWPRR

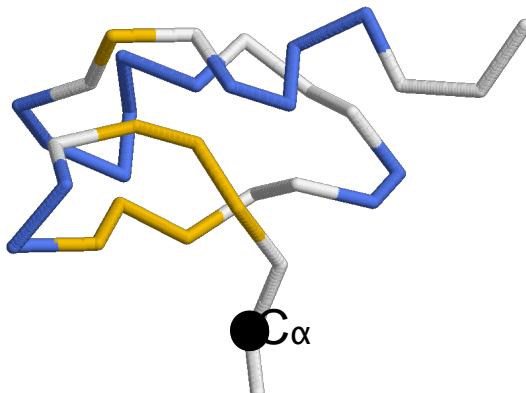
Primary sequence



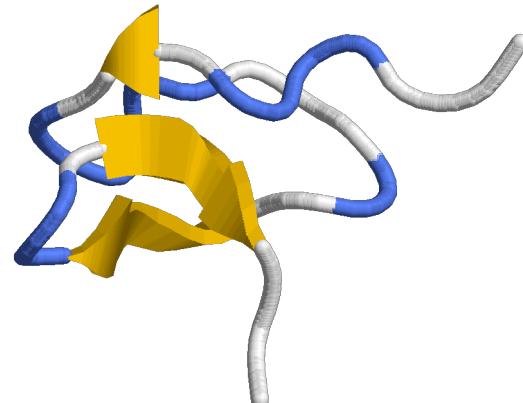
All atoms and coordinates



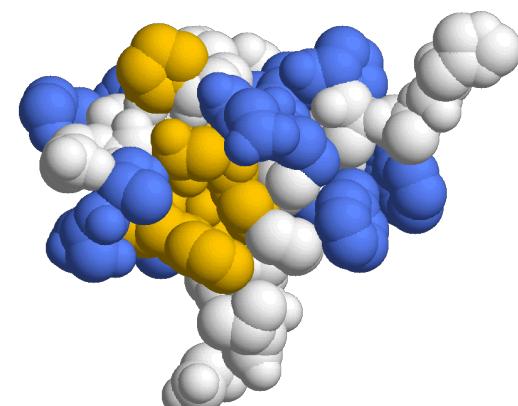
Distance space



Reduced atoms representation



Secondary Structure



Accessible surface

Scoring

Statistical Potential... inspiration

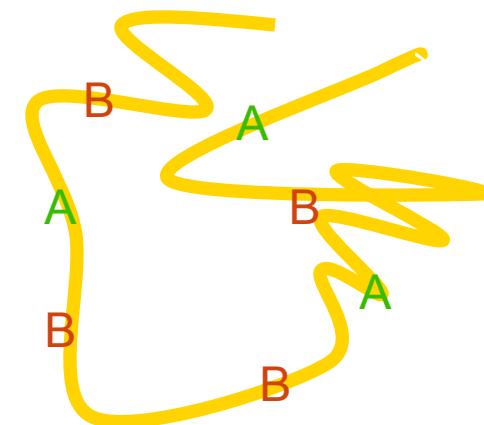
$$K = \frac{[AB]}{[A] \cdot [B]}$$

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

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]:

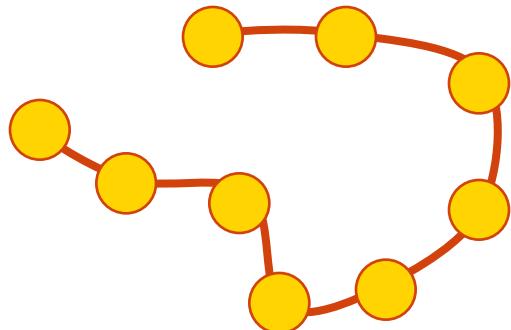
$$\Delta E = -kT \ln \left(\frac{N_1}{N_2} \right) \quad (1)$$

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.

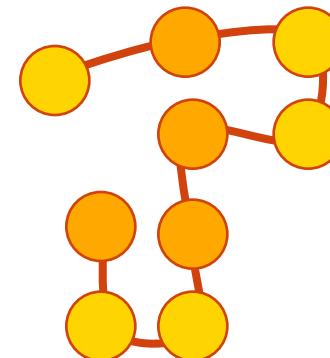


Scoring

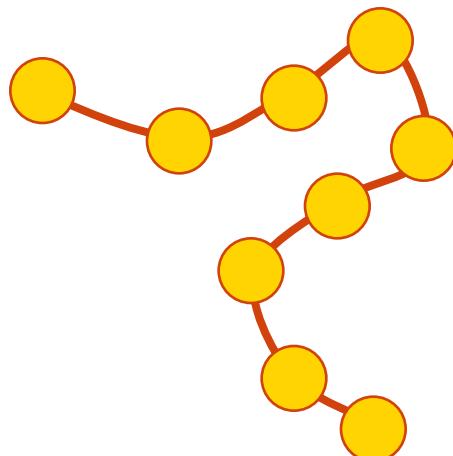
Statistical Potential... interaction types



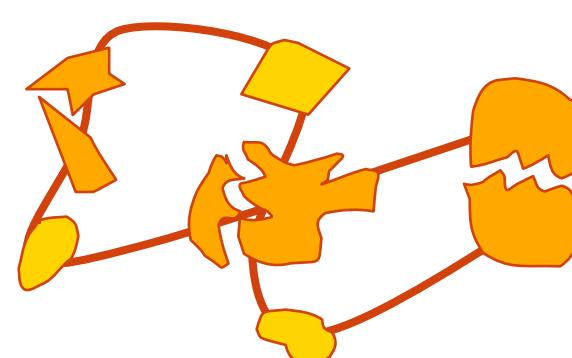
Neutral interactions



Hydrophobic interactions



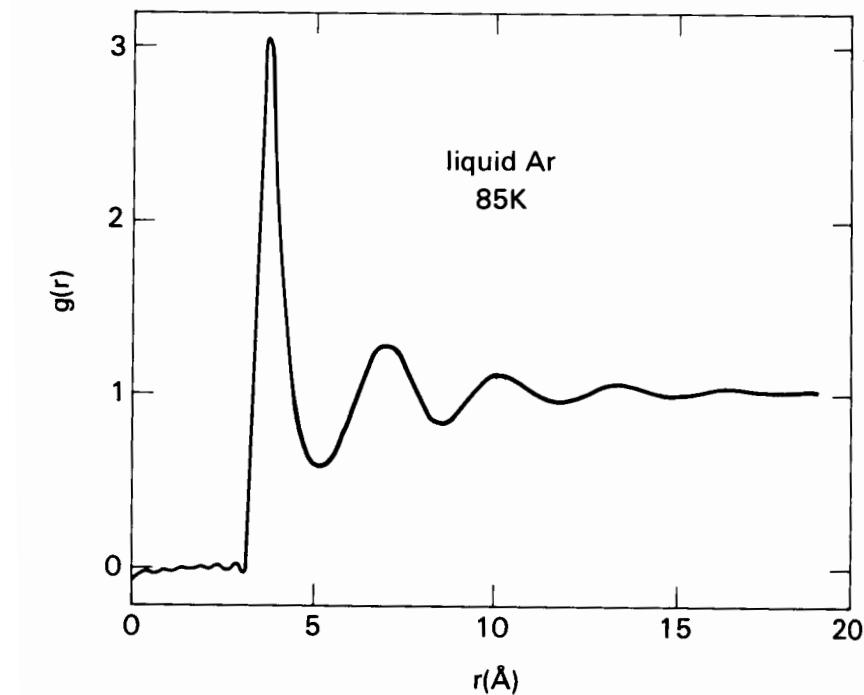
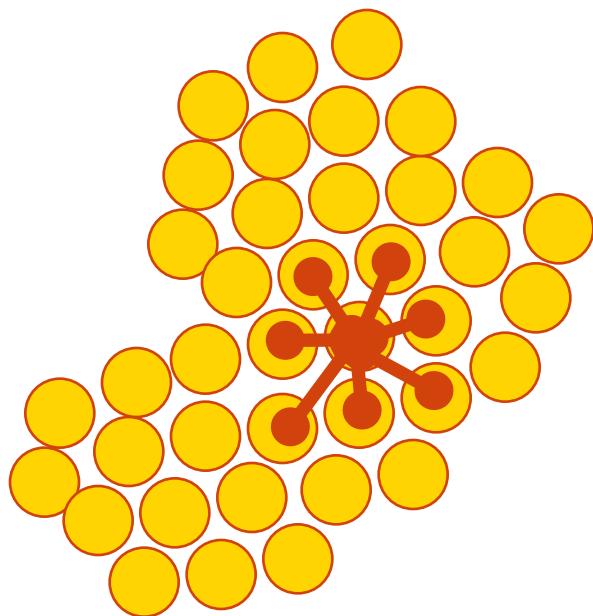
Compact interactions



Specific interactions

Scoring

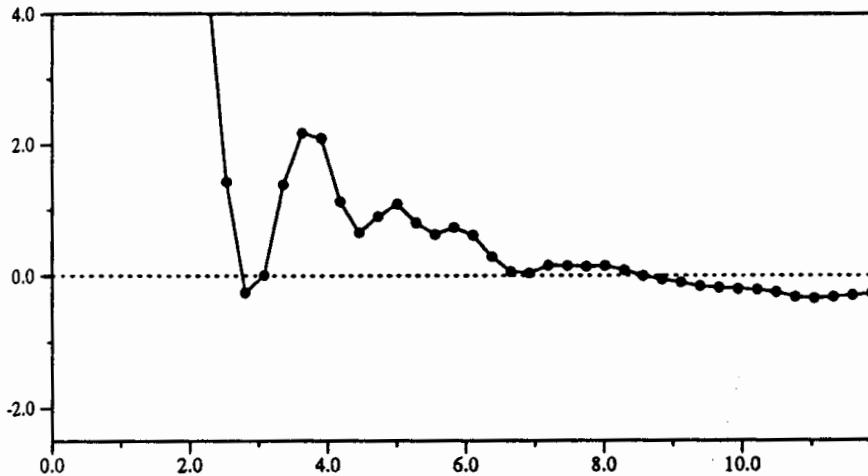
Statistical Potential... reference state



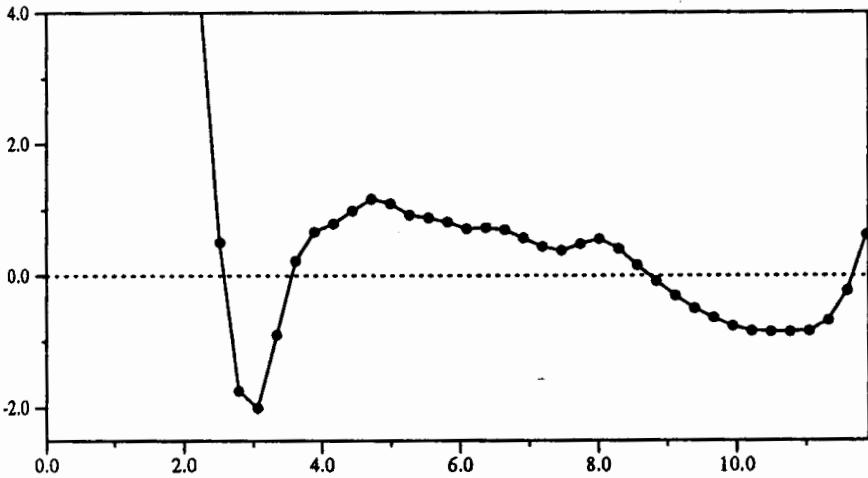
Scoring

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

$$\rho_{NO}(r) = \sum_j \delta(\mathbf{r} - \mathbf{r}_j)$$

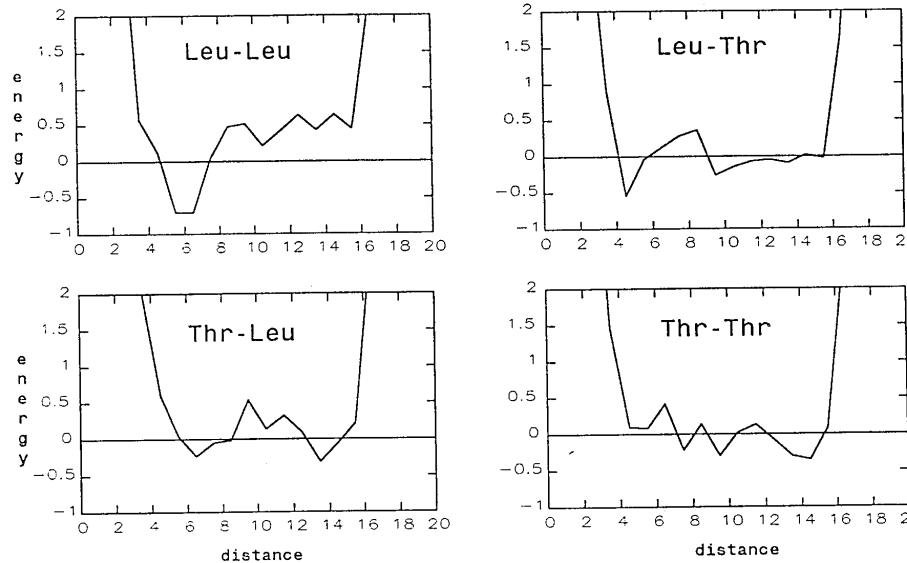
$$g_{NO}(r) = \frac{\rho_{NO}(r)}{\rho}$$

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

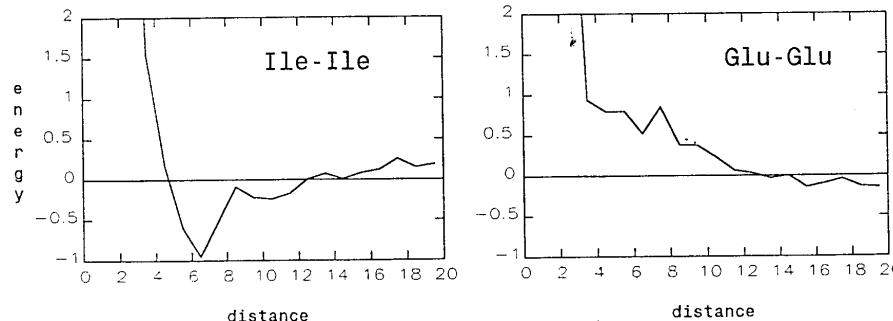
Scoring

Statistical Potential... Distance Potentials

Long range free energy



Short range free energy



Sippl (1993). JCAM 7 pp473

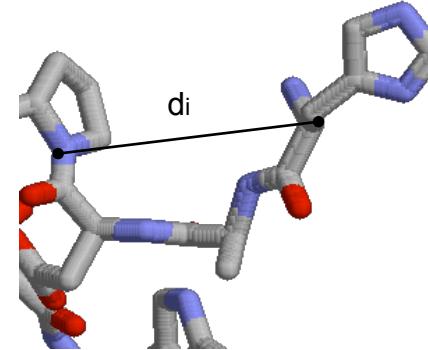
Scoring

Raw scores of an alignment

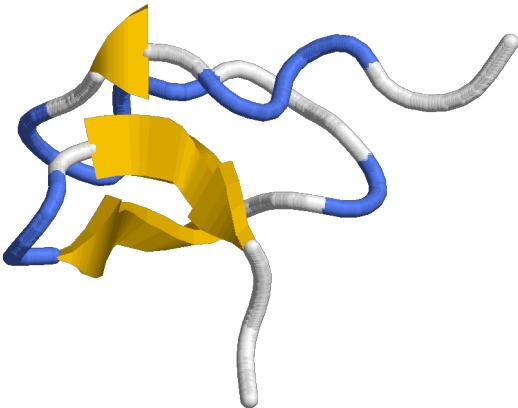
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	4	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-3	
T	-1	1	4	1	-1	1	0	1	0	0	-1	-1	-1	-1	-2	-2	-2	-2	-3	
P	-3	-1	1	7	-1	-2	-1	-1	-1	-1	-2	-2	-1	-1	-2	-3	-2	-4	-3	-4
A	0	1	-1	-1	4	0	-1	-2	-1	-1	-2	-1	-1	-1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

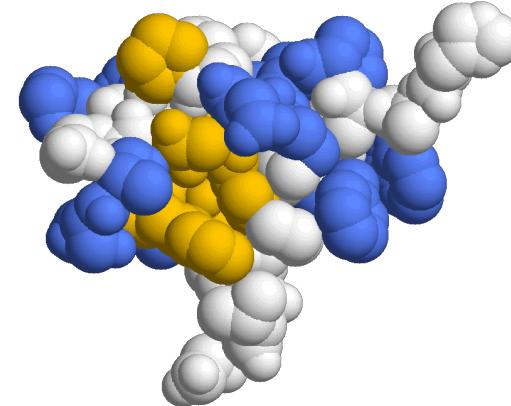
Aminoacid substitutions



Distance space



Secondary Structure (H,B,C)



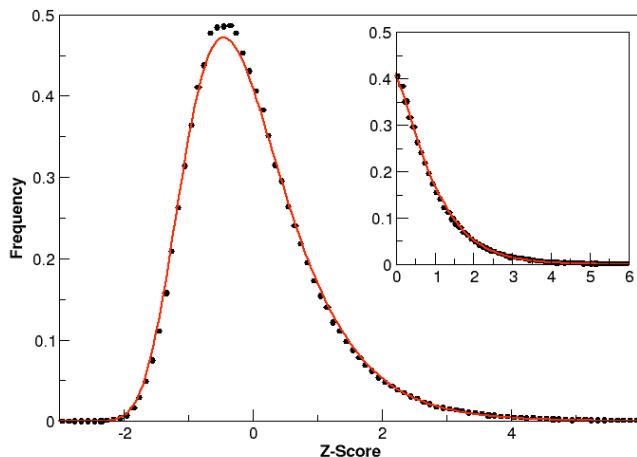
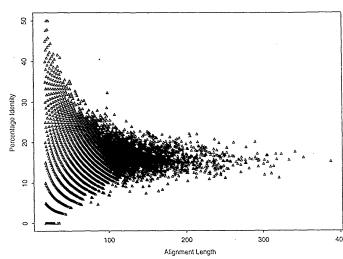
Accessible surface (B,A [%])

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

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

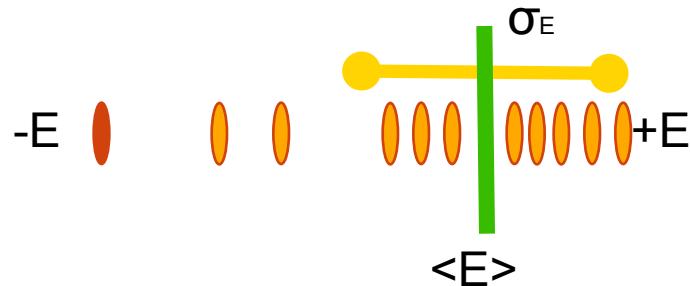
$$P(s \geq x) = 1 - \exp(-e^{-\lambda (s-\mu)})$$

Karlin and Altschul, 1990 PNAS 87, pp2264

Scoring

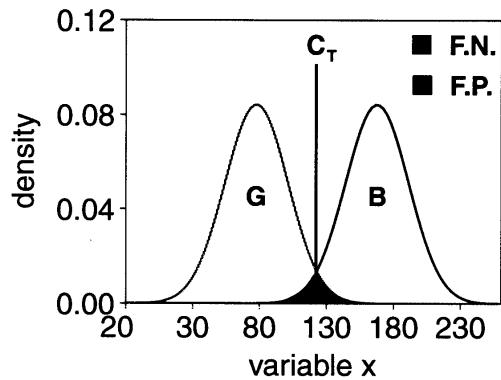
Significance of an alignment (score)

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

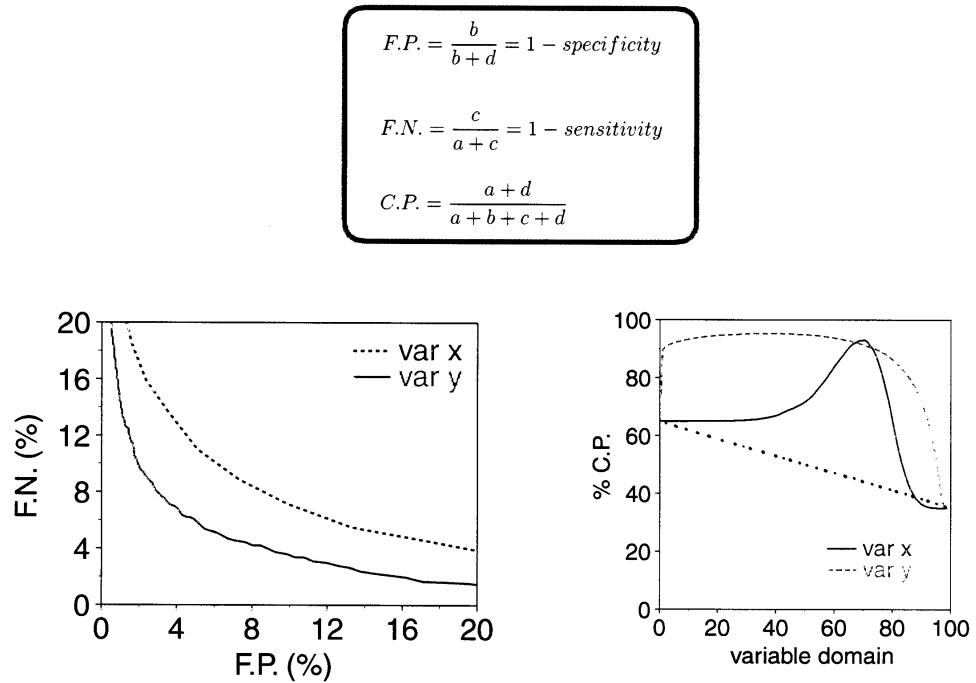


$$Zscore = \frac{(\langle E \rangle - E_m)}{\sigma_E}$$

Evaluating the assessment



	<i>is GOOD</i>	<i>is BAD</i>
<i>predicted as GOOD</i>	a	b
<i>predicted as BAD</i>	c	d



Evaluating the assessment

1

3900 GOOD MODELS

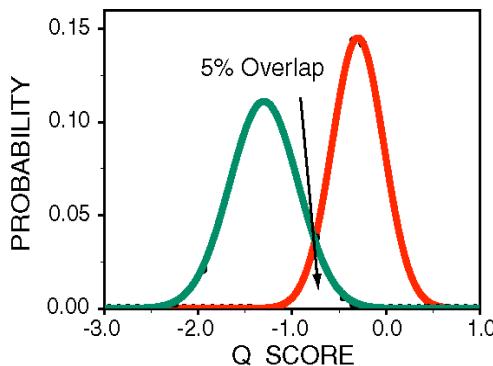
Models based on correct templates and approximately correct alignments

6000 BAD MODELS

Models based on incorrect templates or mostly incorrect alignments

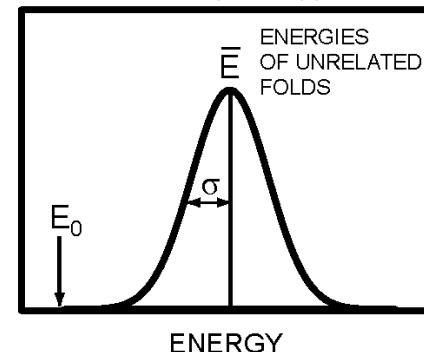
3

$p(Q\text{-SCORE}/\text{GOOD})$
 $p(Q\text{-SCORE}/\text{BAD})$



2

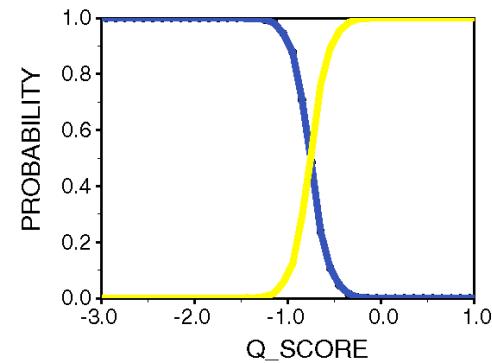
Prosall by M. Sippl



4

$$pG = \frac{p(Q\text{-SCORE}/\text{GOOD})}{p(Q\text{-SCORE}/\text{GOOD}) + p(Q\text{-SCORE}/\text{BAD})}$$

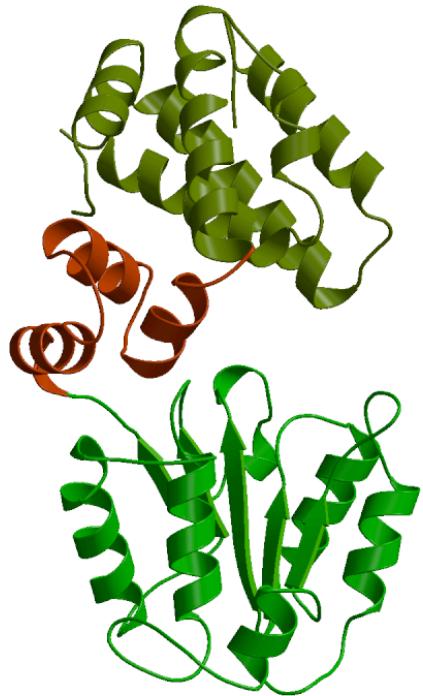
$$p(\text{BAD}/Q\text{-SCORE}) = 1 -$$



R. Sánchez & A. Šali, (1998) *Proc. Natl. Acad. Sci. USA* **95**, pp13597

Applications of methods for model assessment

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



Ec d' MRWYPWLRLPDEKLVASYQAGRGGHHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRG

RUVB LEEYVGQPQVRSQMEIFIFIKAAKLRGDALDHLLIFGPPGLGKTTLANIVANEMG-----

Ec d' CQLMQAGTHPDYYTLAPEKGKATLGVDREVTEKLNEARLGGAKVVWVTDAALLTDAAANALLKTL

RUVB -----VNLRTT-----SGPVLEKAGDLAAMLTNLEPHDVLFIDEIHRLSPVVEEVLYPAM

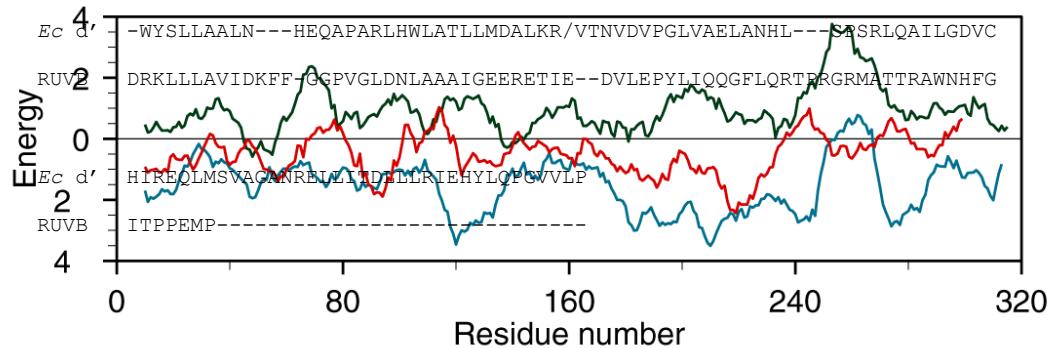
Ec d' -----EEPPAETWFFLATREPERL---LATLRSRCRLHYLAPPPEQYAVTWLSRE

PpdP EDYQLDIMIGEGPAARSIKIDLPPFTLIGATTRAGSLTSPLRDRFGIVQRLEFY--QVPDLQYIVSRS

Ec d' VTM-----SQDALLAALRLSAGSPGAALALFQ-----GDNWQARETLCQALAYSVP SGD--

RUVB ARFMGLEMSDDGALEVARRARGTPRIANRLRRVRDFAEVKHDGTISADIAAQALDMNVDAEGFDYM

Energy profiles (Prossal by M. Sippl)

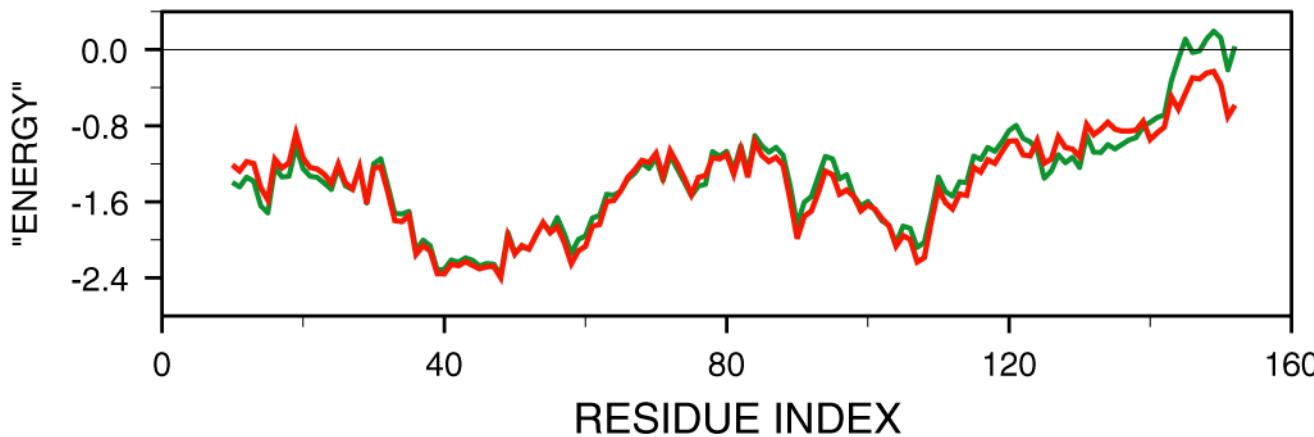


B. Guenther, R. Onrust, A. Šali, M. O'Donnell & J. Kuriyan. *Cell* **91**, 335, 1997.

Yamada, K., Kunishima, N., Mayanagi, K., Ohnishi, T., Nishino, T., Iwasaki, H., Shinagawa, H., Morikawa, K. Crystal Structure of the Holliday Junction Migration Motor Protein Ruvb from *Thermus Thermophilus* Hb8. *Proc.Nat.Acad.Sci.USA* **98**, 1442, 2001.

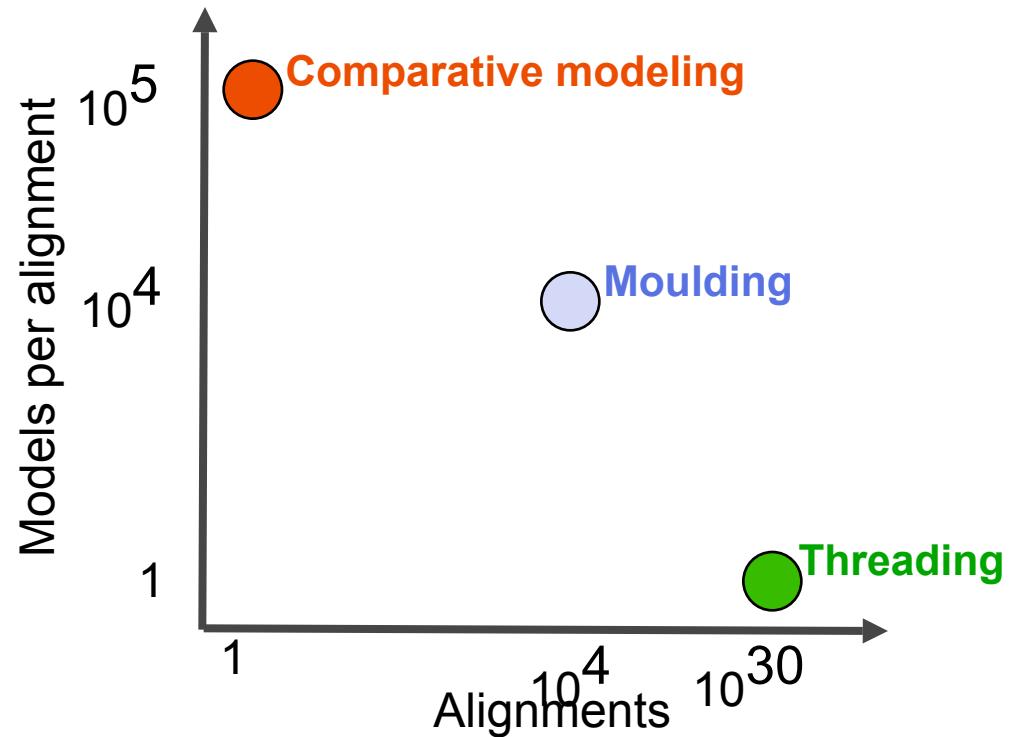
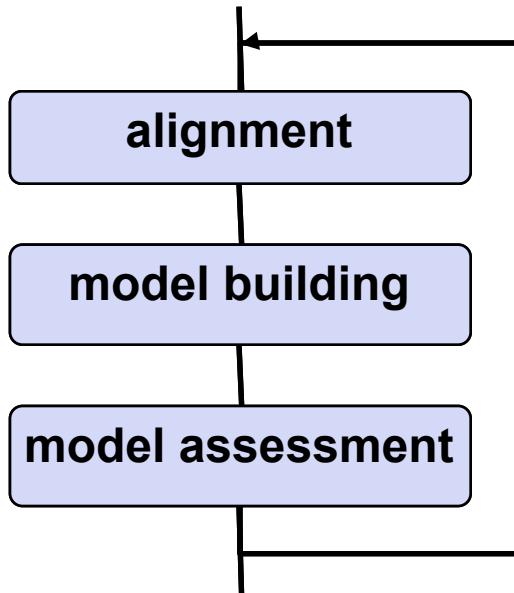
Model Evaluation: Alignment Errors

3dfr TKVSSRTVEDT---NPALTHTYEVWQKKA
4dfrB ESVFSEFHDADAQNS--HSYCFKILERR
DFR₁ model1 ELDAETDHEG-----FTLQEWFVRASSR
DFR₁ model2 ELDAETDHEG-----FTLQEWFVRASSR

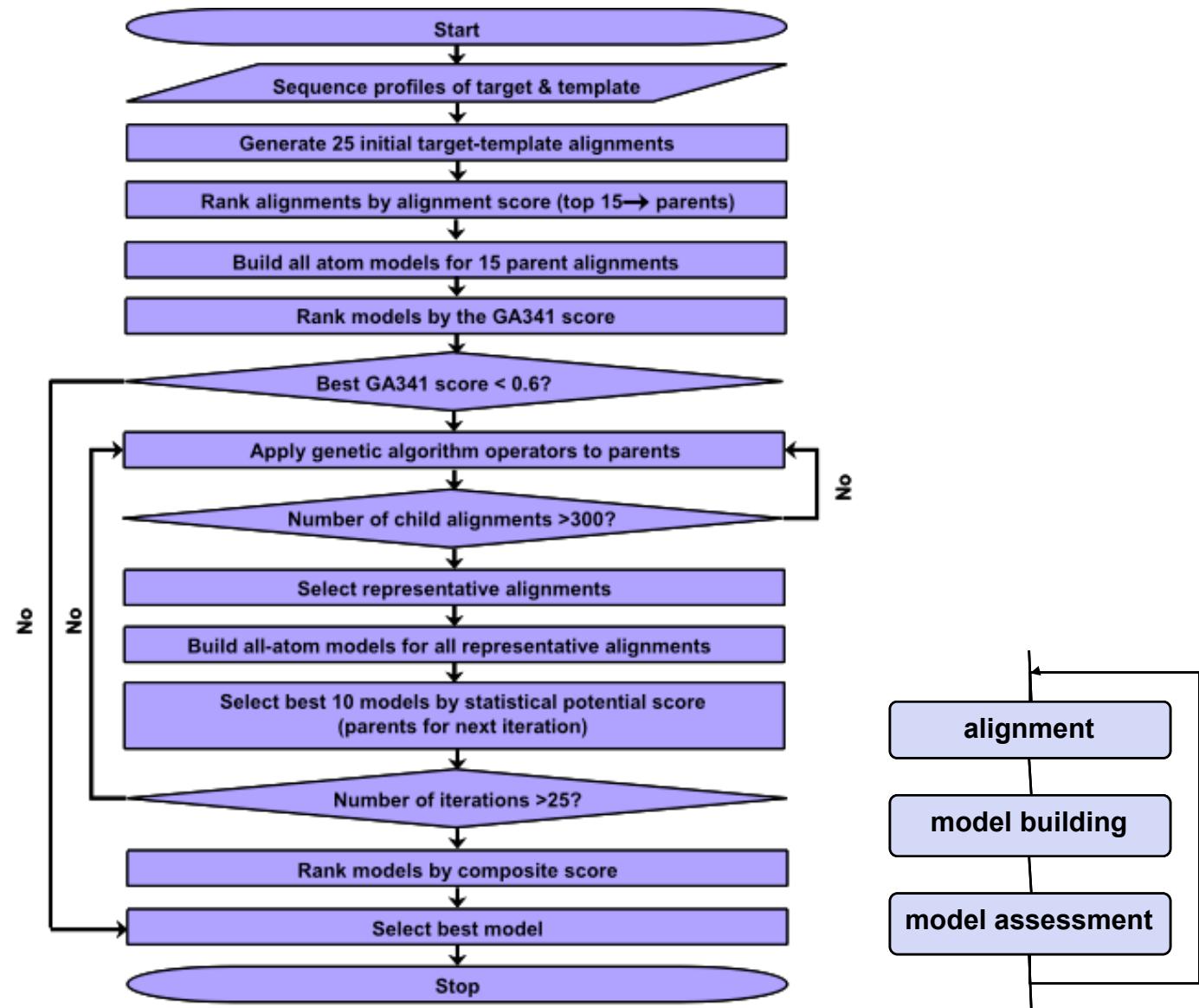


Moulding: iterative alignment, model building, model assessment

B. John, A. Sali. *Nucl. Acids Res.*, **31**, 1982-1992, 2003.



Moulding by a *Genetic Algorithm* approach



Genetic algorithm operators

Single point cross-over

...TSSQ—**NMKLGVFWGY**—...

...V—SSCN—**GDLHMKVG**V...



...TSSQ—**N**MK—**LGVFWGY**...

...V—SSCN**GDLHMKV**—GV...

...TSSQ**N**MK**LGVFWGY**—...

...VSSCN—**GDLHMKVG**V...

Gap insertion

...TSSQ**N**MK**LGVFWG**Y...

...VSSCN**GDLHMKVG**V...



...TSSQ**N**MK**LGVFWGY**...

...VSSCN**GDLHMKVG**—V...

Gap shift

...**T**—S**SQNMKLGVFWGY**...

...VSSC**NGDLHMKVG**V—...



...—**T**—SQNMKLGVFWGY...

...VSSC**NG**DLHMKVG—...

...**T**—S**QNMKLGVFWGY**...

...VSSC**NGDLHMKVG**V—...

...—**T**SSQ**N**MKLGVFWGY...

...VSSC**NG**DLHMKVG—...

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

...**TS**—SQNMKLGVFWGY...

...VSSC**NG**DLHMKVG—...

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).

$$Z = 0.17 Z(P_p) + 0.02 Z(P_s) + 0.10 Z(S_c) + 0.26 Z(H_a) + 0.45 (A_s)$$

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

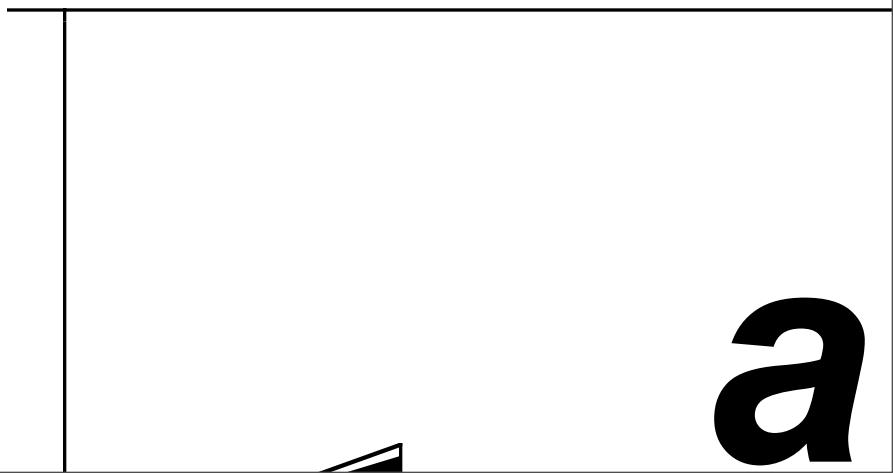
μ ... average score of all models

σ ... standard deviation of the scores

Application to a difficult modeling case

1BOV-1LTS

2



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	
			C α RMSD [Å]	CE overlap [%]	C α RMSD [Å]	CE overlap [%]	C α 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

Alignment accuracy (CE overlap)

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

PSI-BLAST (sequence-profile alignment)	25%
SAM (Hidden Markov Models)	36%
MOULDER (iterative sequence-structure alignment)	45%

Programs

Scoring
PROSAII

Deriving

Scoring

Scoring
PROSAII

Scoring ANOLEA

Deriving

Scoring

Scoring
ANOLEA

Scoring
VERIFY 3D

Deriving

Scoring

Scoring
VERIFY 3D

Scoring DFIRE

Deriving

Scoring

Scoring **DFIRE**

MODEL EVALUATION

a posterior assessment of methods and models

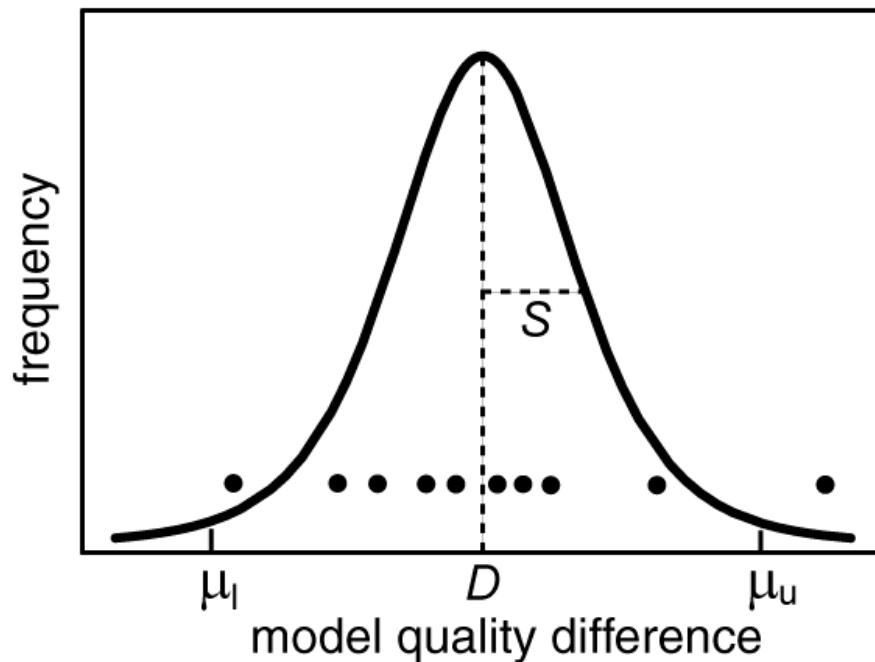
Reliability of assessment of protein structure prediction at CASP

M.A. Marti-Renom, M.S. Madhusudhan, A. Fiser, B. Rost, A. Sali

There were 14 target sequences in the comparative modeling category at CASP4.

Is this number sufficient for reliable ranking of the modeling methods?

Statistical significance of comparing two modeling methods



Compare methods based on common models only.

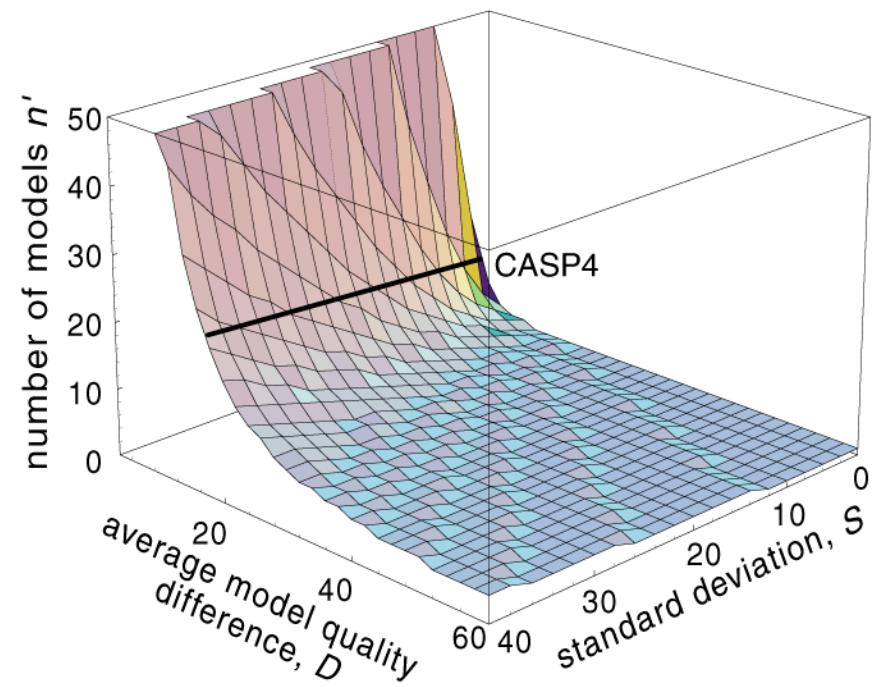
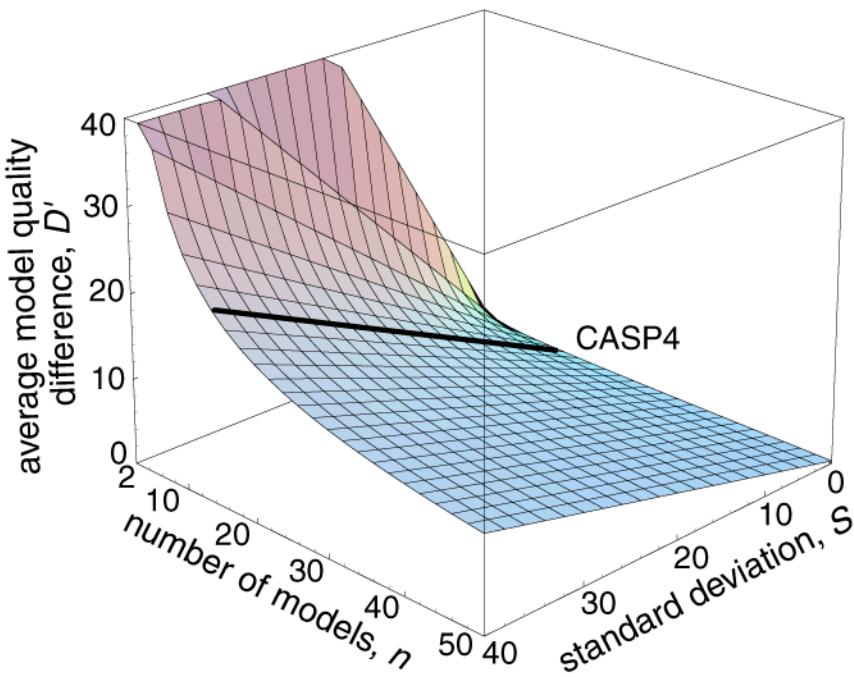
Model Quality Criterion

Quality = Coverage and Accuracy.

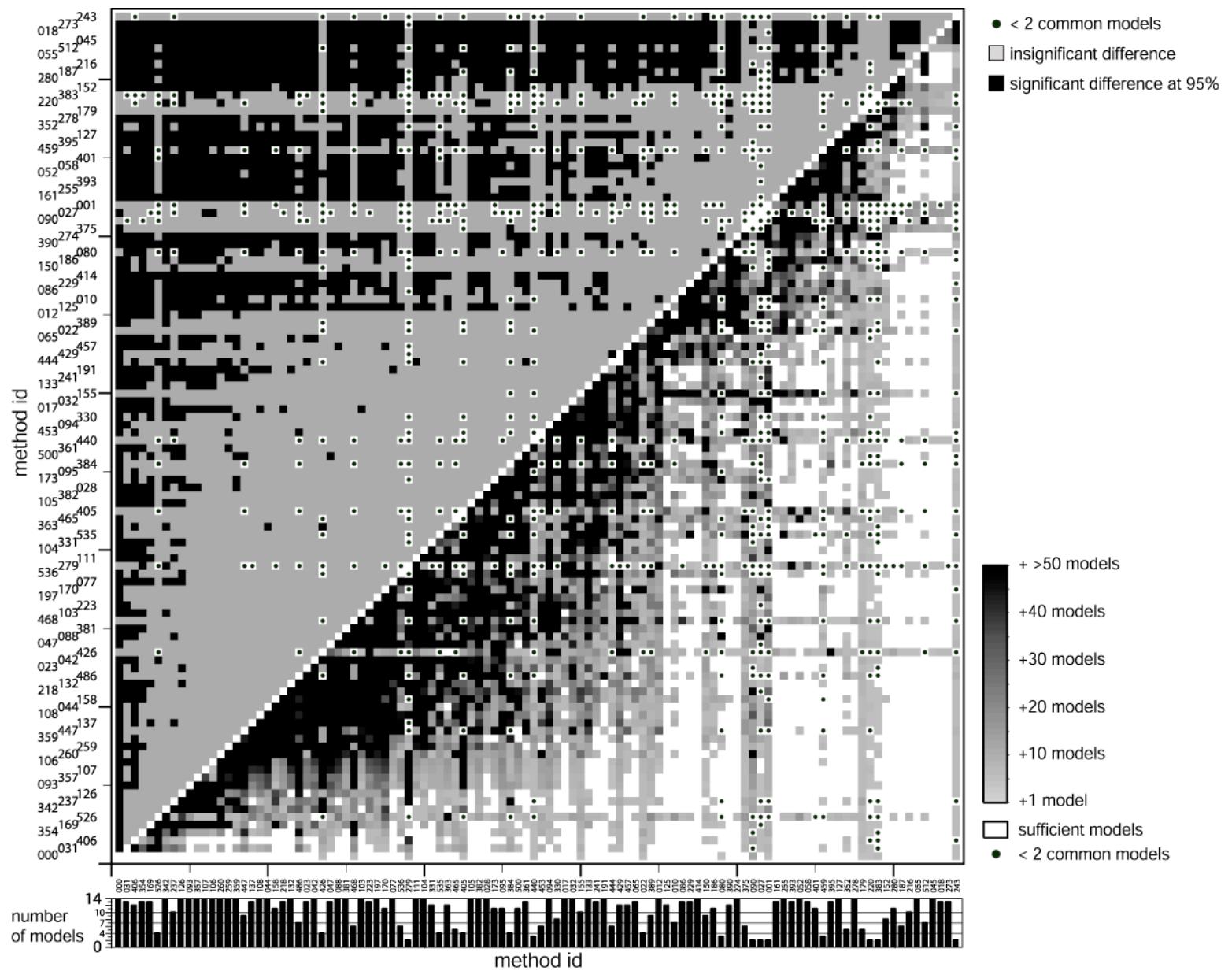
Quality = Average Coverage at 1, 2, 3 Å cuttofs.

From the CASP web site.

Statistical significance of comparing two modeling methods



Comparison of performances of comparative modeling methods at CASP4



Conclusions (CM at CASP4)

Conclusions (CM at CASP4)

- Not enough targets at CASP4 to discriminate between top ~8 modeling methods;

Conclusions (CM at CASP4)

- Not enough targets at CASP4 to discriminate between top ~8 modeling methods;
- Hundreds of target sequences needed (difficulty of models; fold assignment, alignment, loops, backbone distortions, sidechains);

Conclusions (CM at CASP4)

- Not enough targets at CASP4 to discriminate between top ~8 modeling methods;
- Hundreds of target sequences needed (difficulty of models; fold assignment, alignment, loops, backbone distortions, sidechains);
- Need automated modeling and automated assessment, such as EVA (Rost, Sali, Valencia, Eyrich, Marti-Renom, Przybylski, Pazos, Madhusudhan, Fiser) and LiveBench (Rychlewski, Fischer, Elofsson, Bujnicki).



Why?



Why?

Continuous....



Why?

Continuous....

Automatic....



Why?

Continuous....

Automatic....

Large scale...



Why?

Continuous....

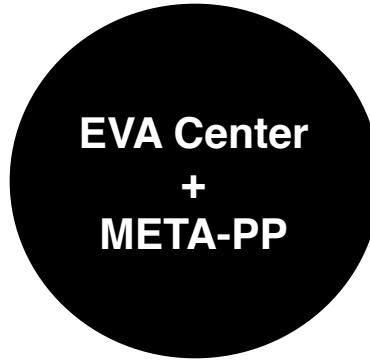
Automatic....

Large scale...

Accessible...



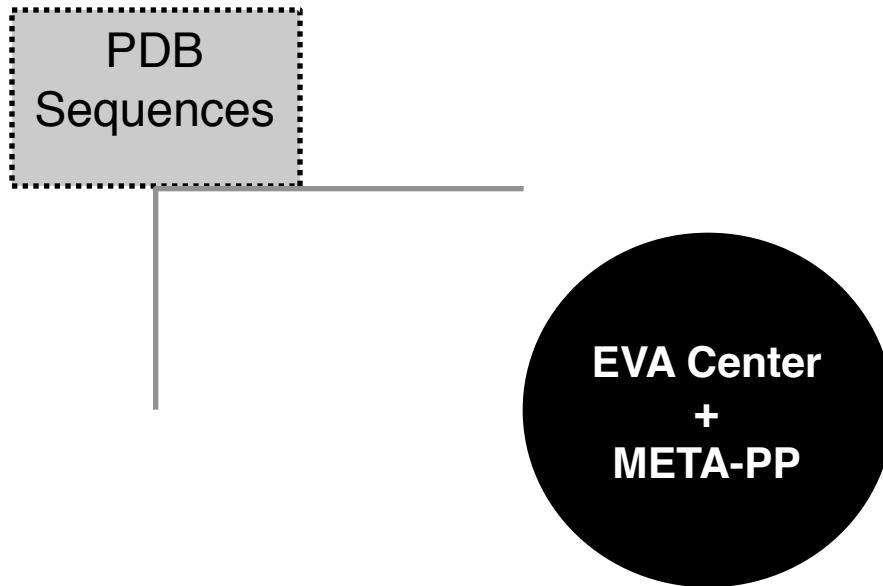
What?



EVA Center
+
META-PP

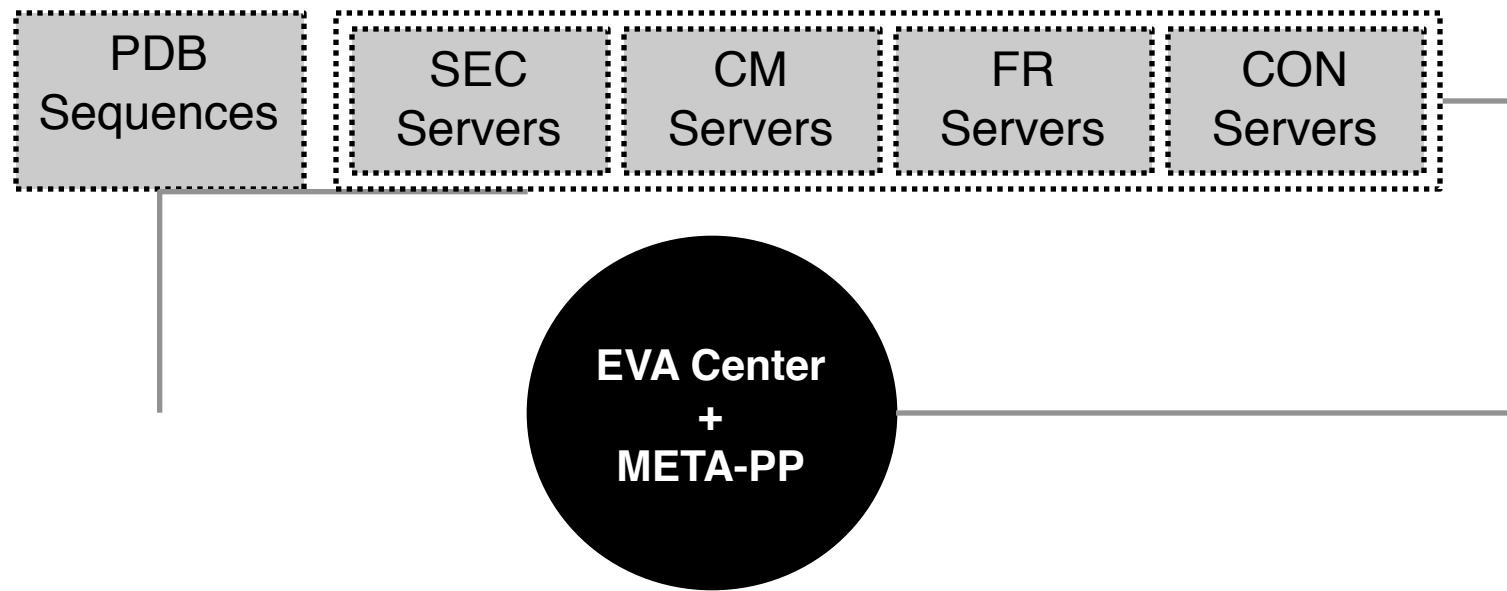


What?



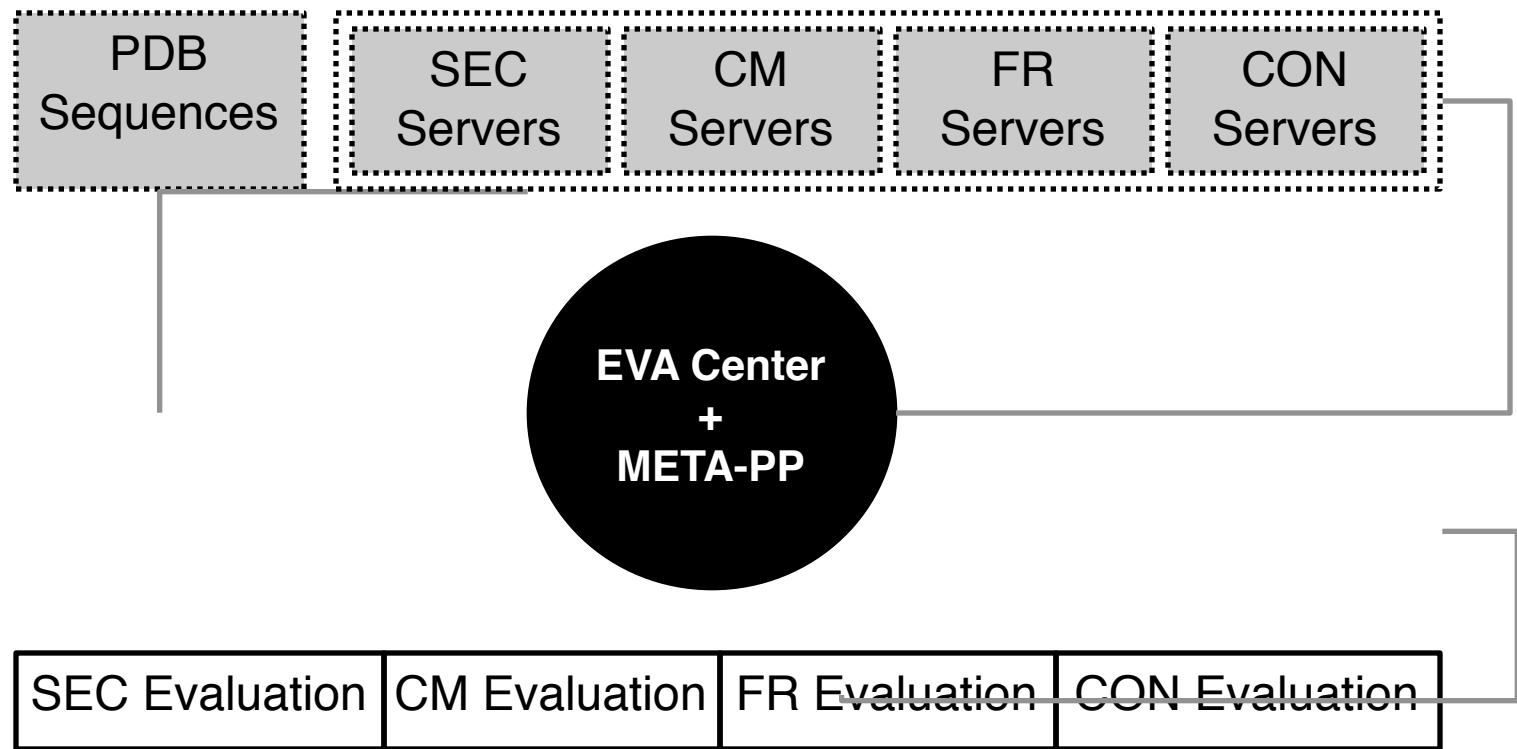


What?



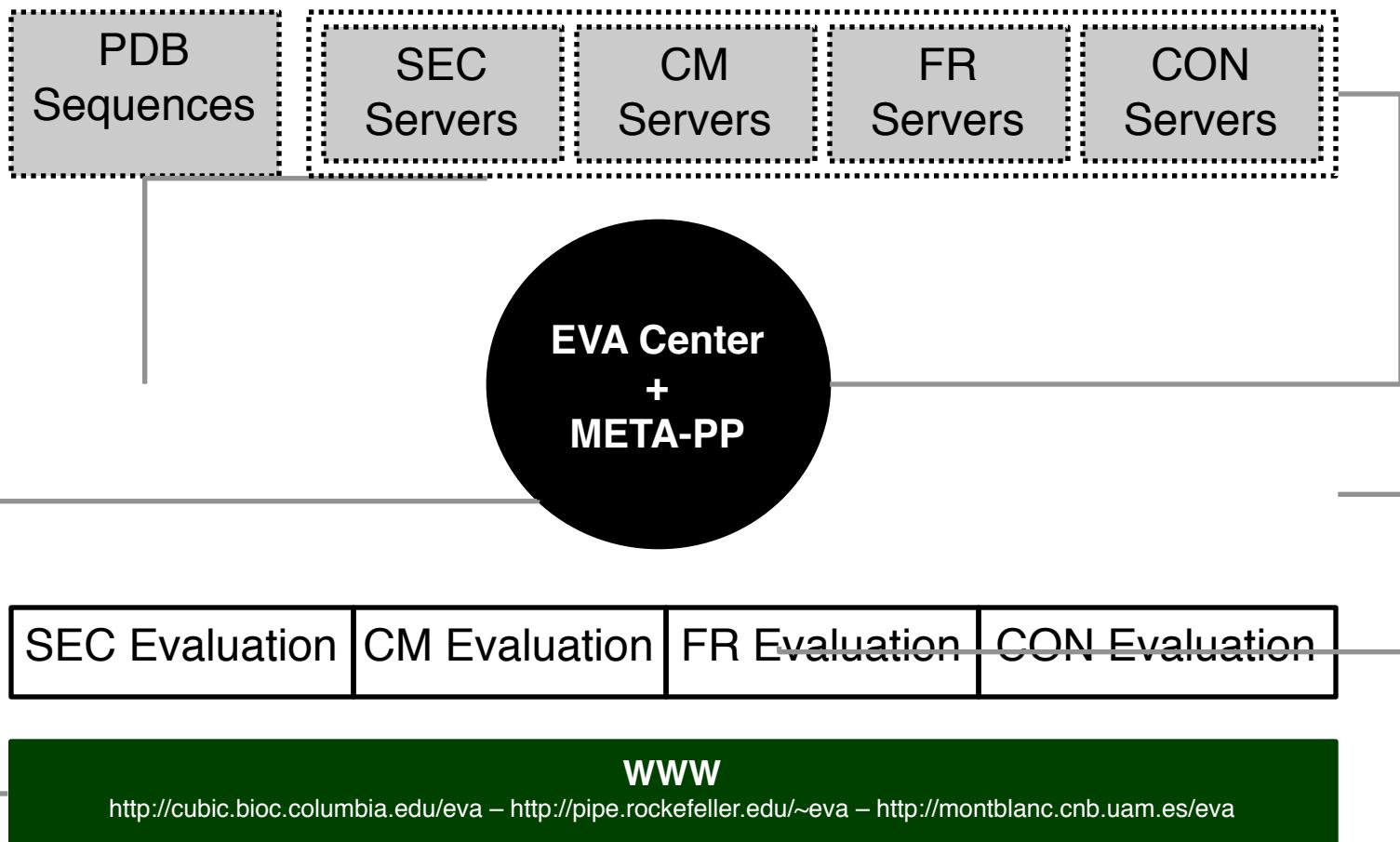


What?



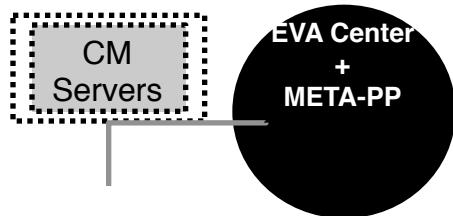


What?



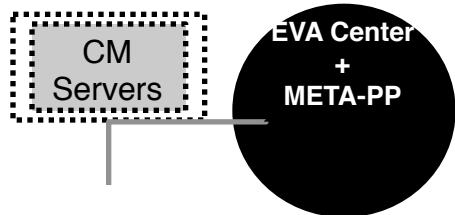


How?



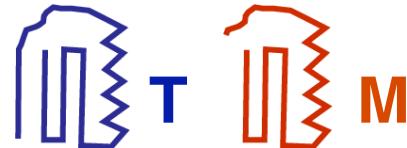


How?



Eva-CM

Fold accuracy...



3D alignment by CE



Calculate RMSD, equivalent positions under 3.5Å etc...

Alignment accuracy...



1 to 1 alignment



Calculate RMSD, equivalent positions under 3.5Å etc...

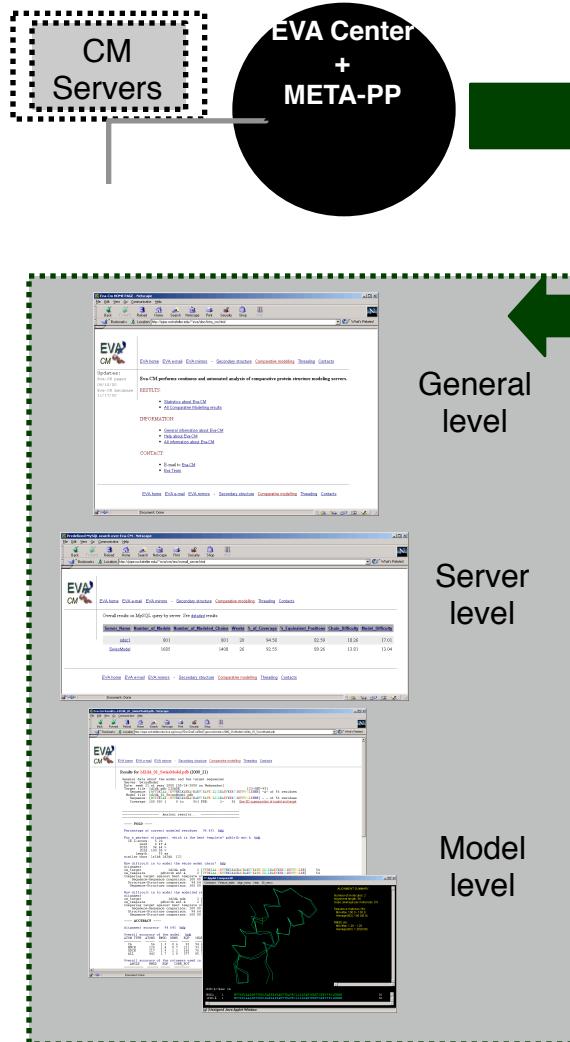
Structure quality...

MODELLER and PROCHECK

MODELLER and PROCHECK programs will calculate several structural properties of the model.

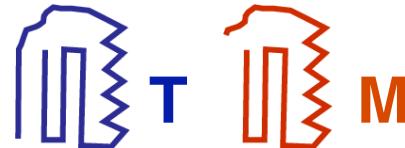


How?



Eva-CM

Fold accuracy...



aaaAAAAAAAAA
AAAAaaaAAAAAA

3D alignment by CE

Calculate RMSD, equivalent positions under 3.5Å etc...

Alignment accuracy...



---aaaAAAAAAAAA
AAAAaaaAAAAAA---

1 to 1 alignment

Calculate RMSD, equivalent positions under 3.5Å etc...

Structure quality...

MODELLER and PROCHECK

MODELLER and PROCHECK programs will calculate several structural properties of the model.



Where?

<http://cubic.bioc.columbia.edu/eva/> (NYC, USA)

<http://www.salilab.org/eva/> (UCSF, USA)

<http://montblanc.cnb.uam.es/eva/> (Madrid, Spain)



Main pages @ Eva-CM

Eva-CM HOME PAGE - Netscape

Eva-CM performs continuous and automated analysis of comparative protein structure modeling servers.

RESULTS:

- Statistics about Eva-CM
- All Comparative Modelling results

INFORMATION:

- General information about Eva-CM
- Help about Eva-CM
- All information about Eva-CM

CONTACT:

- E-mail to Eva-CM
- Eva Team

Preddefined MySQL search over Eva-CM - Netscape

Statistics about Eva-CM server/database.

Number of Weeks Running	Number of Public Servers	Number of Modeled Chains	Number of Models
26	2	1531	2486

Weekly statistics about Eva-CM server/database (only last weeks shown).

Week	Number of Models	Number of Modeled Chains	Number of Public Servers
2000_46	83	36	2
2000_45	180	94	2
2000_44	100	52	2
2000_43	233	121	2
2000_42	66	59	1
2000_41	135	58	2
2000_40	174	83	2
2000_39	165	88	2
2000_38	178	89	2
2000_37	156	80	2
2000_36	110	60	2
2000_35	58	32	2

Eva-CM RESULTS PAGE - Netscape

Results page for Eva-CM.

STATIC PAGES (local pages)

- Overall and detailed results by server
- Overall and detailed results by server and week
- Overall and detailed results for common subset of chains modelled by ALL evaluated servers

DYNAMIC PAGES (only @ pipe.rocketfeller.edu server in NYC)

- Limited and full search of results

PRIVATE AREA (only @ pipe.rocketfeller.edu server in NYC)

- Enter Server Id Enter Password Enter private area

This section is intended to offer privacy to servers that want to test their results using Eva-CM, but do not want to make them public. Private results are not exhaustive as public results.



Server results pages @ Eva-CM

Screenshot of the "Overall results" page for Eva-CM. The page title is "Predefined MySQL search over Eva-CM - Netscape". It shows a table of results for two servers: sdsc1 and SwissModel. An arrow points from this screenshot down to the "Detailed results" screenshot below.

Server_Name	Number_of_Models	Number_of_Modeled_Chains	Weeks	%_of_Coverage	%_Equivalent_Positions	Chain_Difficulty	Model_Difficulty
sdsc1	801	801	20	94.58	82.59	18.26	17.01
SwissModel	1685	1408	26	92.55	89.26	13.81	13.04

Screenshot of the "Detailed results" page for Eva-CM. The page title is "Predefined MySQL search over Eva-CM - Netscape". It shows a table of results for the same two servers as the previous screenshot, but includes additional columns for Ca_global_rmsd, Ca_global_%eq_pos, Ca_core_rmsd, and Ca_core_%eq_pos. An arrow points from the "Overall results" screenshot above down to this detailed view.

Server_Name	Number_of_Models	Number_of_Modeled_Chains	Weeks	%_of_Coverage	%_Equivalent_Positions	Chain_Difficulty	Model_Difficulty	Ca_global_rmsd	Ca_global_%eq_pos	Ca_core_rmsd	Ca_core_%eq_pos
sdsc1	801	801	20	94.58	82.59	18.26	17.01	3.32	81.44	2.96	1.49
SwissModel	1685	1408	26	92.55	89.26	13.81	13.04	1.90	89.38	1.49	1.49



Query and model results pages @ Eva-CM

The figure illustrates the workflow from a search query to a detailed model result.

Left Panel: Eva-CM SEARCH PAGE - Netscape

This panel shows the search interface for Eva-CM. It includes fields for Match PDB identifier, Match keyword on protein definition (PDB HEADER), Match given methods, Match % of coverage, Match % of equivalent positions, Match global C-alpha RMSD, Match difficulty, Match predictions between given dates, Order output by (Week Number), and Show maximum entries (20). A large grey arrow points from this page to the top right panel.

Top Right Panel: Results for Eva-CM Query using MySQL - Netscape

This panel displays a table of search results. The columns include Model Name, Week Number, Server Name, % of Coverage, % of Equivalent Positions, % correct aligned residues, C_a global rmsd, Chain Difficulty, and Model Difficulty. The results show various models (1t0A_01_SwissModel.pdb, 1t0A_01_InsuModel.pdb, etc.) with their respective statistics. A smaller grey arrow points from this table to the bottom right panel.

Bottom Right Panel: Eva-CM Results (1t0A_01_SwissModel.pdb) - Netscape

This panel provides detailed results for the SwissModel model (2000_21). It includes sections for General data about the model and the target sequences, Sequence alignment (with a green ribbon diagram showing alpha-helices), Analysis results (including FOLD, Percentage of correct modeled residues, and Structure-Structure comparison), and Accuracy (Alignment accuracy, Overall accuracy of the model, and Overall accuracy of the rotamers used in Angle). A final Alignment Summary table at the bottom lists various statistics like Min/Max, Average, and Standard Deviation for different metrics.



Ranking @ Eva-CM

Predefined MySQL search over Eva - Microsoft Internet Explorer

Analysis of Fold accuracy:

Ranking for Fold accuracy:

1. sdsc1	2. SwissModel	3. cphmodels
----------	---------------	--------------

t-Student statistical analysis of the comparisons:

	sdsc1	SwissModel	cphmodels
1. sdsc1		2.13 ± 16.15 [1524]	1.23 ± 10.19 [274]
2. SwissModel	-2.13 ± 16.15 [1524]		1.18 ± 15.63 [∞]
3. cphmodels	-1.23 ± 10.19 [274]	-1.18 ± 15.63 [∞]	

Analysis of Alignment accuracy:

Ranking for Alignment accuracy:

1. SwissModel	2. sdsc1	3. cphmodels
---------------	----------	--------------

t-Student statistical analysis of the comparisons:

	SwissModel	sdsc1	cphmodels
1. SwissModel		2.49 ± 19.65 [1524]	3.81 ± 24.75 [257]
2. sdsc1	-2.49 ± 19.65 [1524]		-0.22 ± 17.63 [∞]
3. cphmodels	-3.81 ± 24.75 [257]	0.22 ± 17.63 [∞]	



Acknowledgments:

- **CUBIC group @ Columbia U.**

Volker Eyrich
Dariusz Przybylski
Burkhard Rost

- **Sali Lab @ UCSF**

Marc A. Marti-Renom
Andrej Šali

- **CNB @ U. Atonónoma de Madrid**

Florencio Pazos
Alfonso Valencia

- **All groups developing servers evaluated by EVA.**

Special thanks to Ilya Shindyalov and Phil Bourne for providing CE and Compare3D software.



Other evaluation benchmarks

CASP <http://predictioncenter.llnl.gov>

LiveBench <http://bioinfo.pl/LiveBench/>

Protein Structure Prediction Center - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Mail Links

Address http://predictioncenter.llnl.gov/ Go Links

[CASP1](#)
[CASP2](#)
[CASP3](#)
[CASP4](#)
[CASP5 ✓](#)
[Local services](#)
[Other links](#)
[People](#)
[Website index](#)
[Hide menu](#)

Protein Structure Prediction Center

Biology and Biotechnology Research Program
Lawrence Livermore National Laboratory, Livermore, California, USA



Welcome to the Protein Structure Prediction Center!

Our goal is to help advance the methods of identifying protein structure from sequence. The Center has been organized to provide the means of objective testing of these methods via the process of blind prediction. In addition to support of the CASP meetings our goal is to promote an objective evaluation of prediction methods on a continuing basis.

CASP experiment: [CASP1](#) | [CASP2](#) | [CASP3](#) | [CASP4](#) | [CASP5](#)

Ten Most Wanted: [TMW](#)

The Center, supported by the National Institutes of Health, National Library of Medicine, and the U.S. Department of Energy, [Office of Biological and Environmental Research](#), is a part of the [Biology and Biotechnology Research Program](#) at the [Lawrence Livermore National Laboratory](#).

[Local services](#) | [Other links](#) | [People](#) | [Website index](#)

If you have any questions or comments please contact us at squery@PredictionCenter.llnl.gov

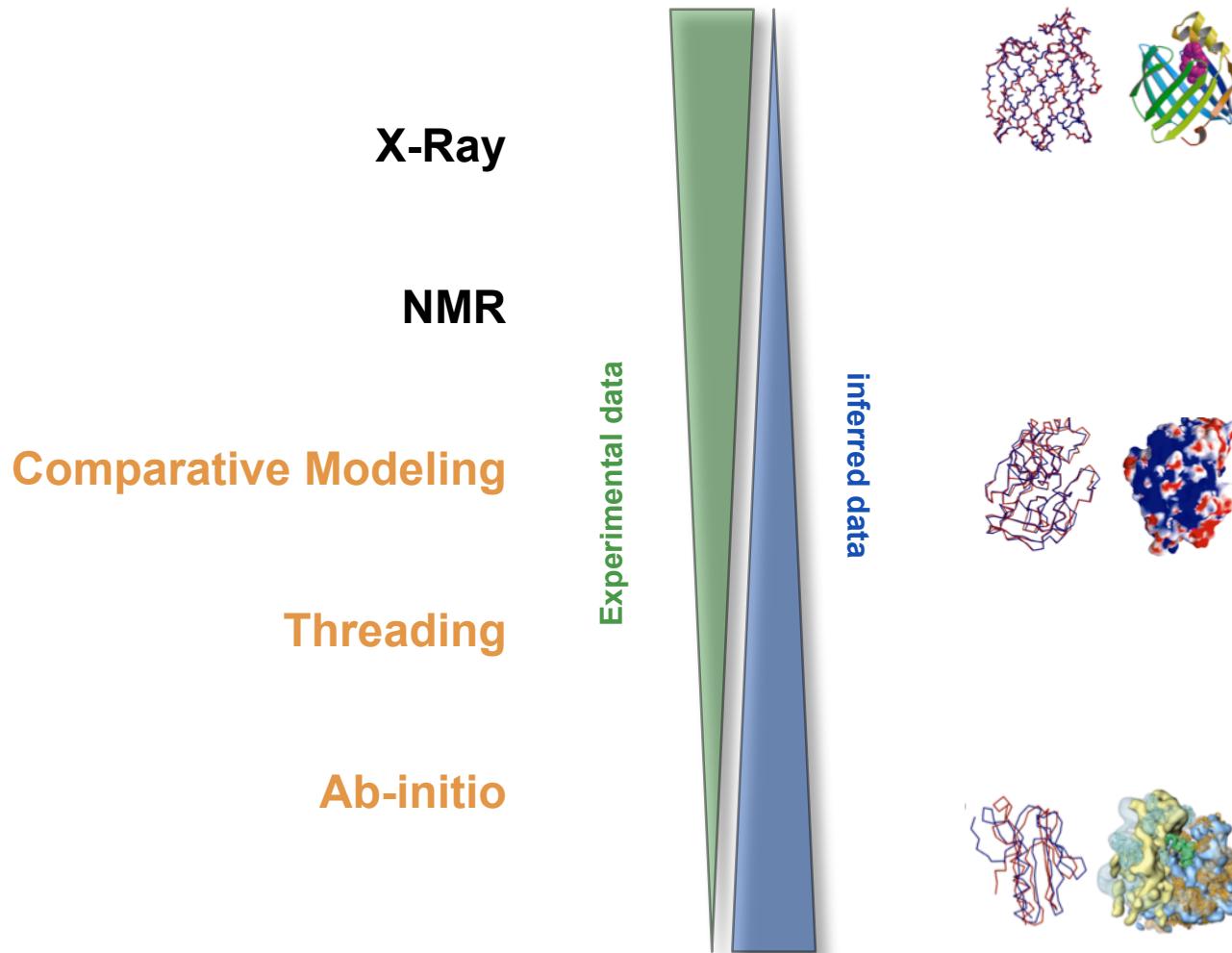
BMC WorkShop

Protein Structure Prediction Summary

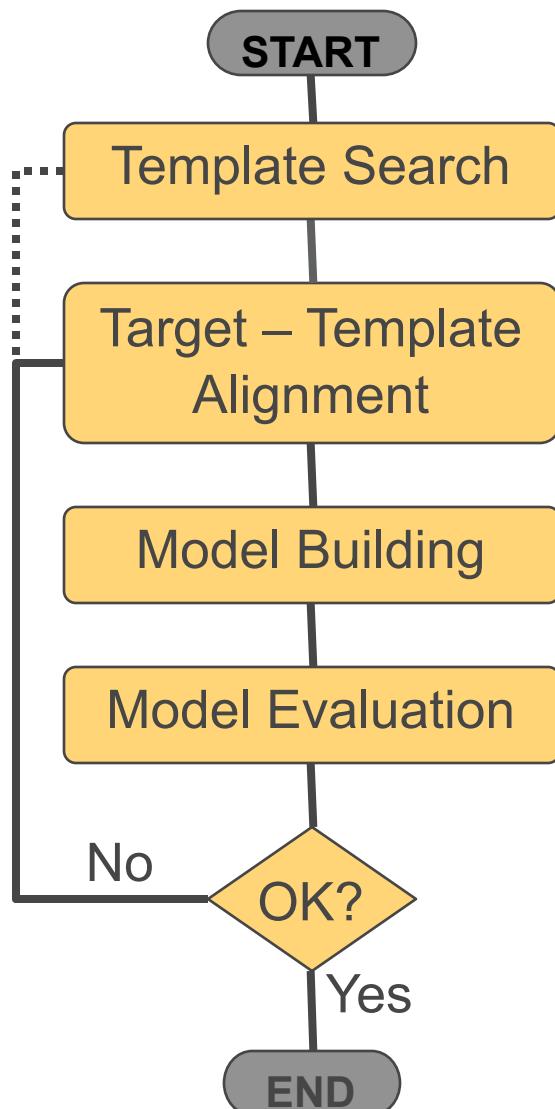
Marc A. Marti-Renom & Damien Devos

Department of Biopharmaceutical Sciences, UCSF

protein prediction .vs. protein determination



Steps in Comparative Protein Structure Modeling



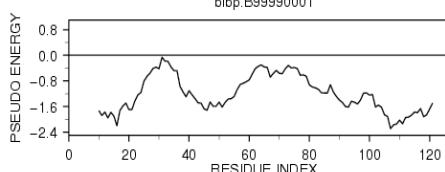
TARGET

ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVLKIDDV
PERLIPERASFQWMNDK

TEMPLATE



ASILPKRLFGNCEQTSDEGLK IERTPLVPHISAQNVLKIDDV PERLIPERASFQWMNDK
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE



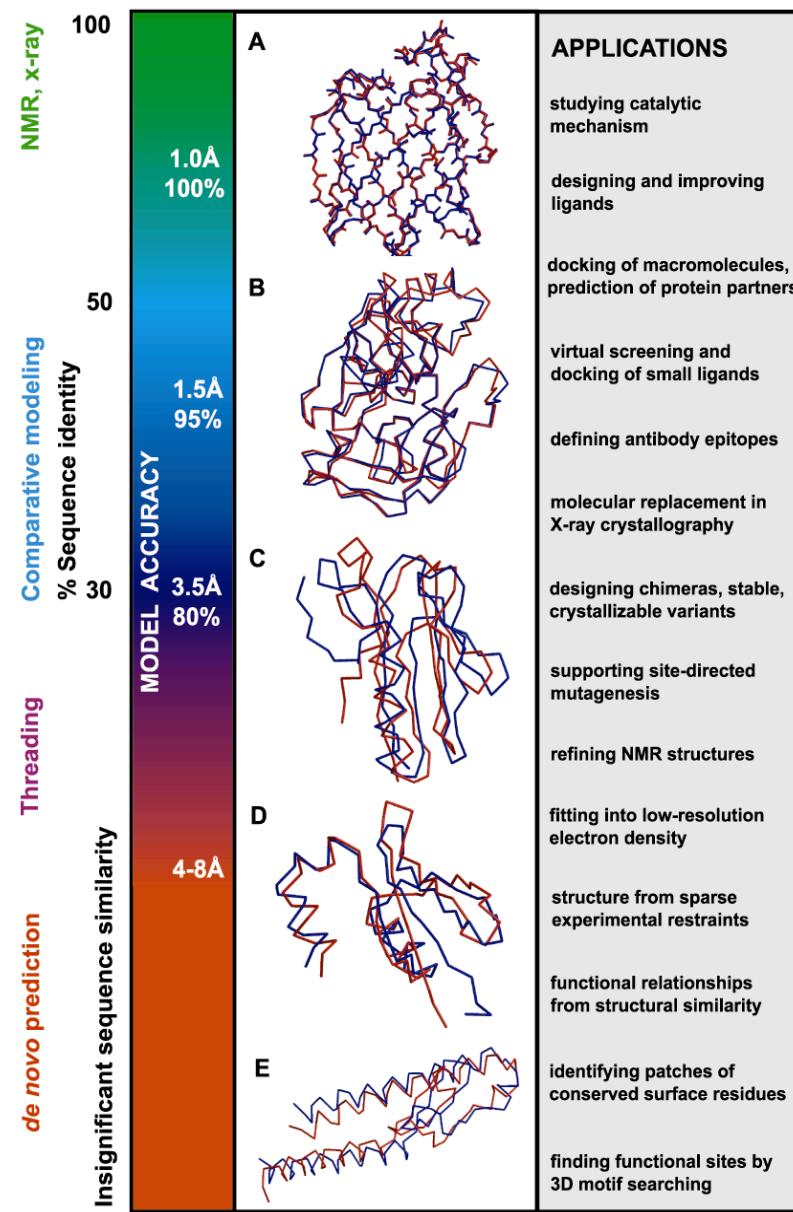
A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.

R. Sánchez & A. Šali, *Curr. Opin. Str. Biol.* 7, 206, 1997.

M. Marti et al. *Ann. Rev. Biophys. Biomol. Struct.*, 29, 291, 2000.

<http://salilab.org/>

Utility of protein structure models, despite errors



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

Acknowledgments

Protein Structure Modeling

Andrej Sali

Bino John

Narayanan Eswar

Ursula Pieper

Roberto Sánchez (MSSM)

András Fiser (AECOM)

Francisco Melo (CU, Chile)

Azat Badretdinov (Accelrys)

M. S. Madhusudhan

Ash Stuart

Nebojša Mirkovic

Valentin Ilyin (NE)

Eric Feyfant (GI)

Min-Yi Shen

Ben Webb

Rachel Karchin

Mark Peterson

Brain Lipid Binding Protein

Liang Zhu (RU)

Nat Heintz (RU)

BRCA1

A. Monteiro (Cornel)

Fly p53

Shengkan Jin (RU)

Arnie Levine (RU)

<http://salilab.org>

1D to 3D for biologists

David Huassler (UCSC)

Jim Kent (UCSC)

Daryl Thomas (UCSC)

Mark (UCSC)

Rolf Apweiler (EBI)

Chimera

P. Babbitt

T. Ferrin

Ribosomes

J. Frank

Structural Genomics

Stephen Burley (SGX)

John Kuriyan (UCB)

NY-SGRC

Mast Cell Proteases

Rick Stevens (BWH)

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

Yeast NPC

Tari Suprapto (RU)

Julia Kipper (RU)

Wenzhu Zhang (RU)

Liesbeth Veenhoff (RU)

Sveta Dokudovskaya (RU)

J. Zhou (USC)

Mike Rout (RU)

Brian Chait (RU)