BMC WorkShop

Protein Structure Prediction

Introduction

Marc A. Marti-Renom & Damien Devos

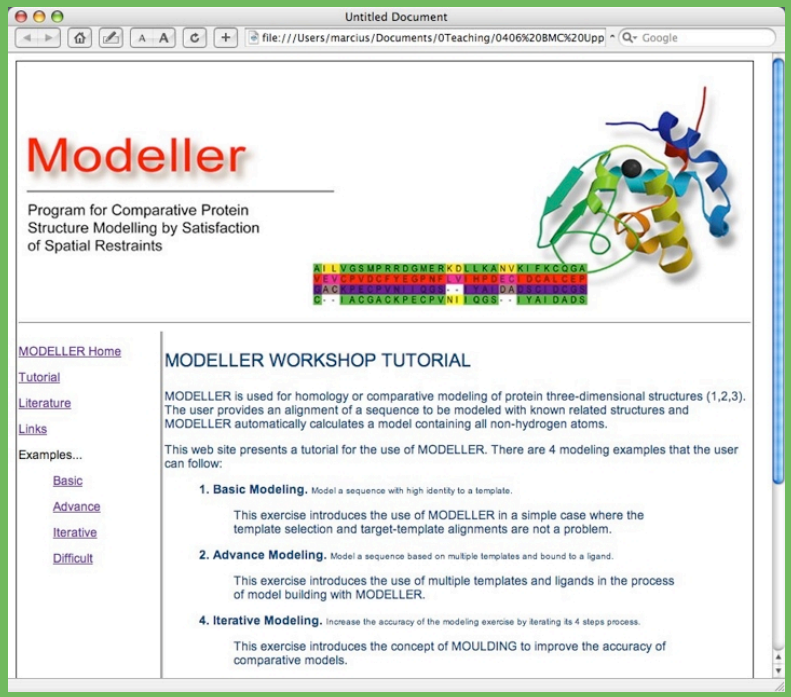
Department of Biopharmaceutical Sciences, UCSF

Objective

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

Program

- Template Search
- Target – Template Alignment
- Model Building
- Model Evaluation



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

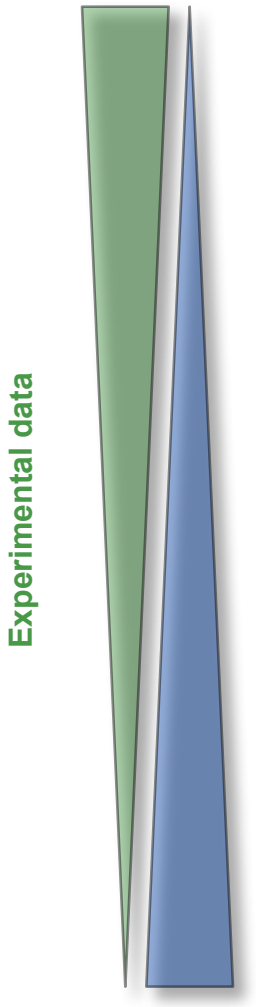
X-Ray

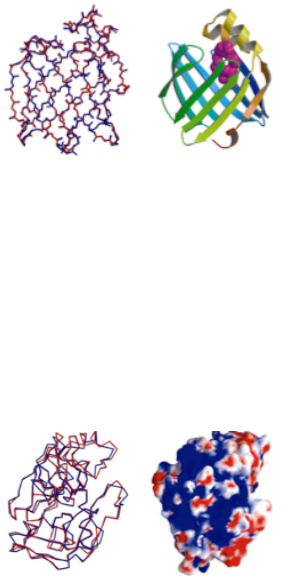
NMR

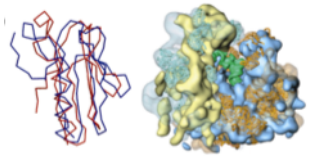
Comparative Modeling

Threading

Ab-initio







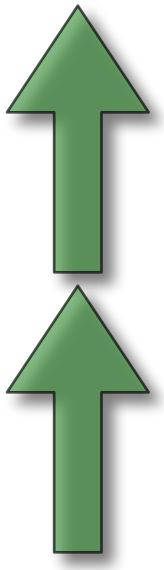
Why protein structure prediction?

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

Why protein structure prediction?

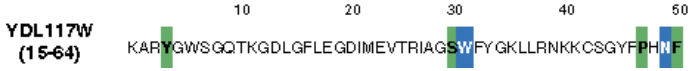
	Y 2004
Sequences	1,500,000
Structures	400,000

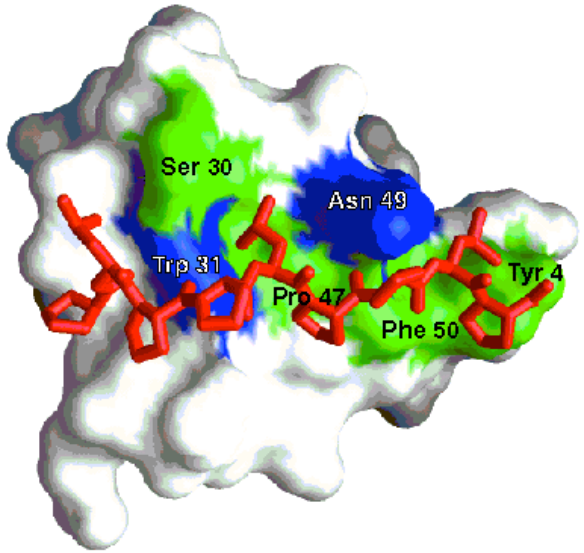
http://salilab.org/modbase/



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

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



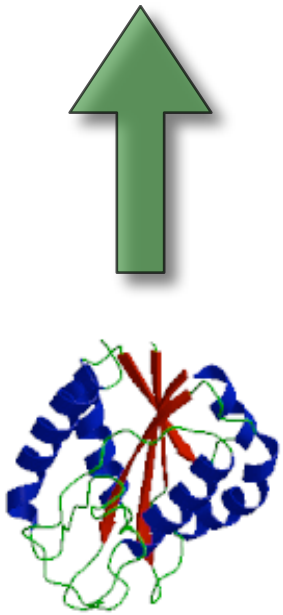


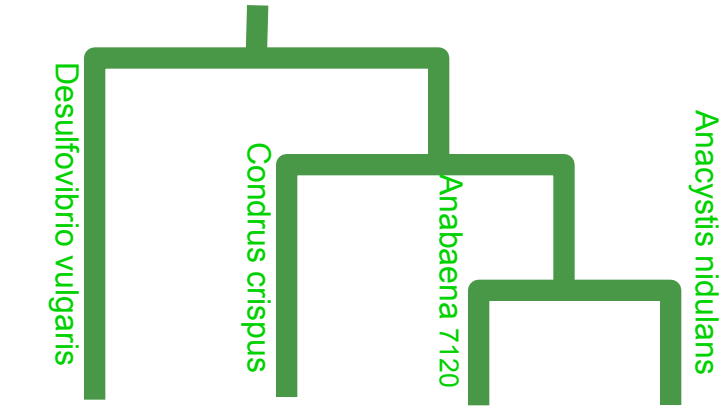
In addition, since evolution tends to conserve function and function depends more directly on structure than on sequence, structure is more conserved in evolution than sequence.

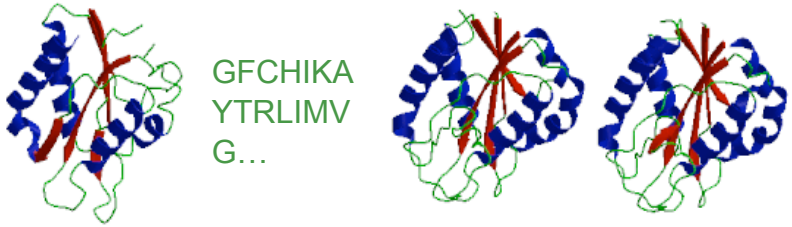
The net result is that patterns in space are frequently more recognizable than patterns in sequence.

Principles of Protein Structure

GFCHIKAYTRLIMVG…







Folding

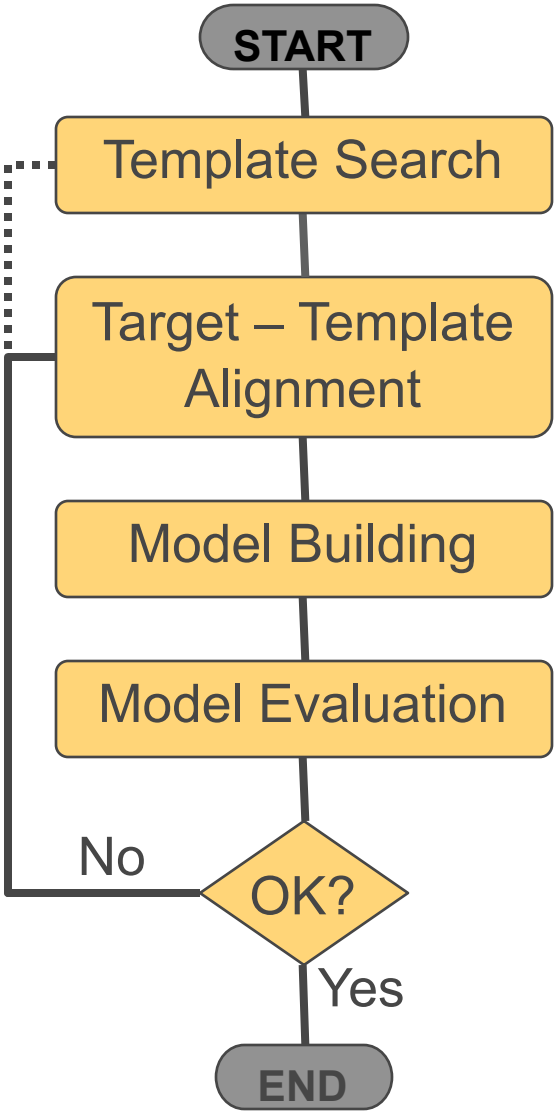
Ab initio prediction

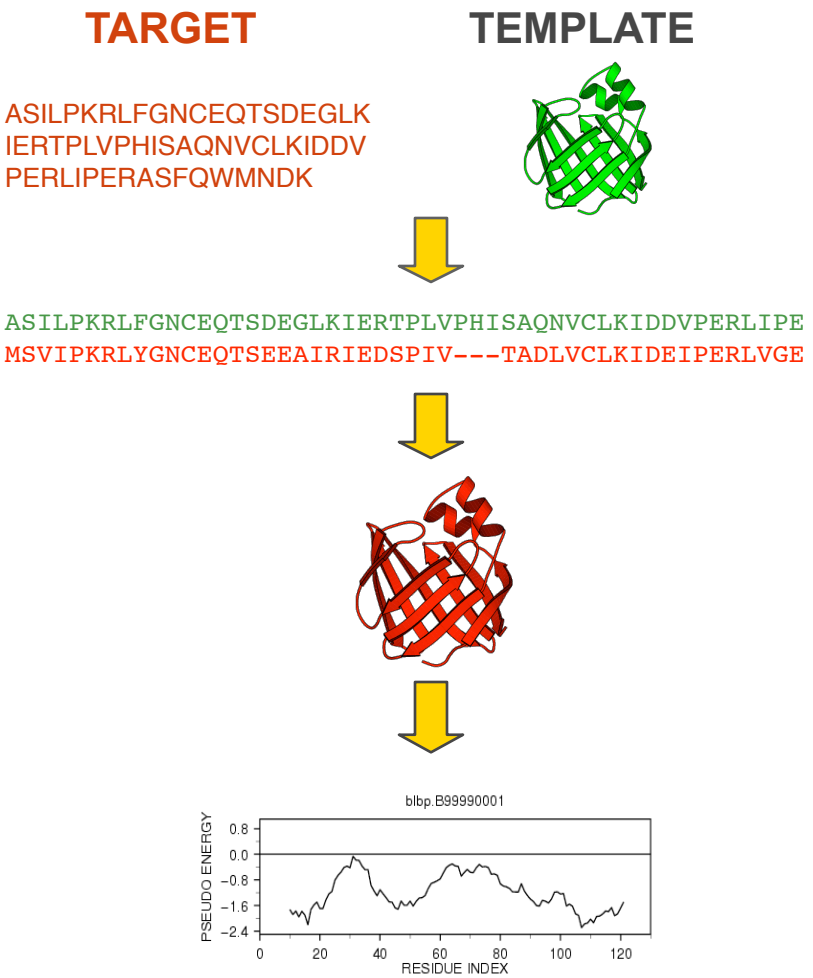
Evolution

Threading

Comparative Modeling

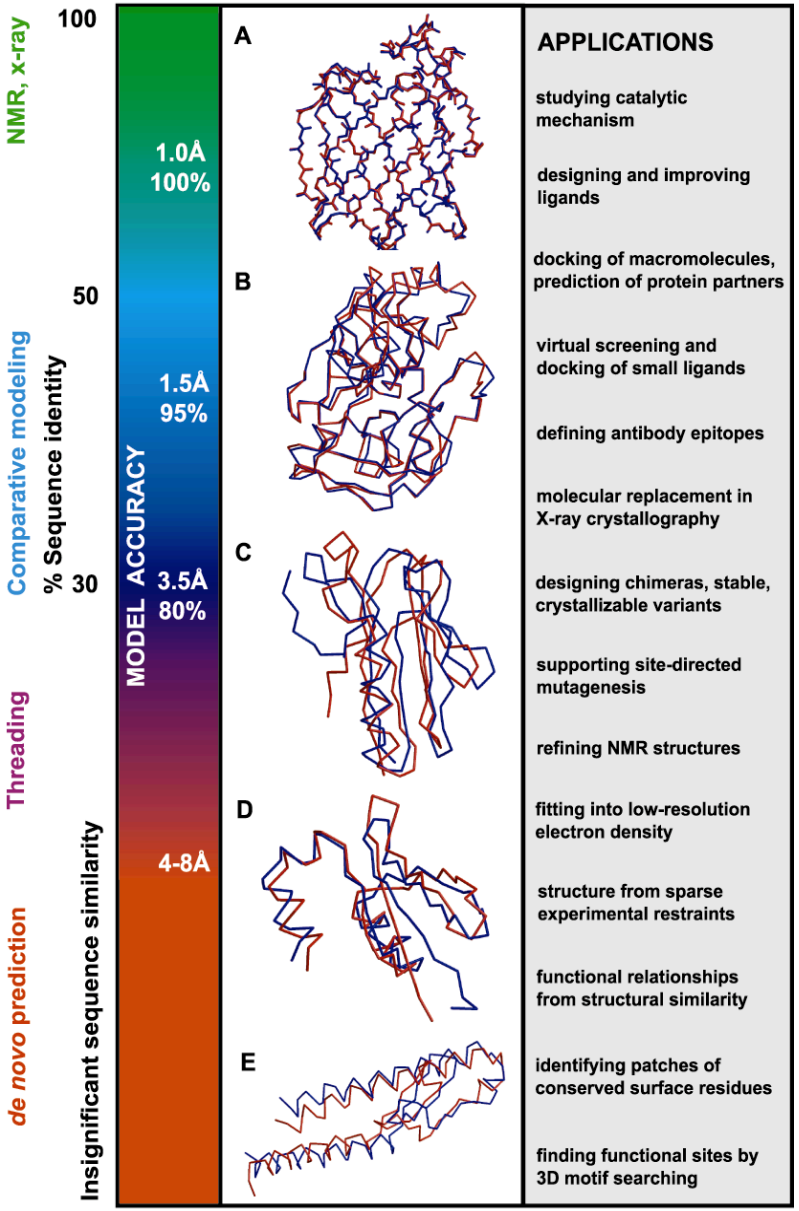
Steps in Comparative Protein Structure Modeling





A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.
R. Sánchez & A. Šali, *Curr. Opin. Str. Biol.* 7, 206, 1997.
M. Marti *et al. Ann. Rev. Biophys. Biomolec. Struct.*, 29, 291, 2000.
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 el al. Annu. Rev. Biophys. Biomol. Struct. 29, 291-325, 2000.

Baker & Sali. Science 294, 93-96, 2001.

Comparative Modeling:

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

Marti-Renom el al. Current Protocols in Protein Science 1, 2.9.1-2.9.22, 2002.

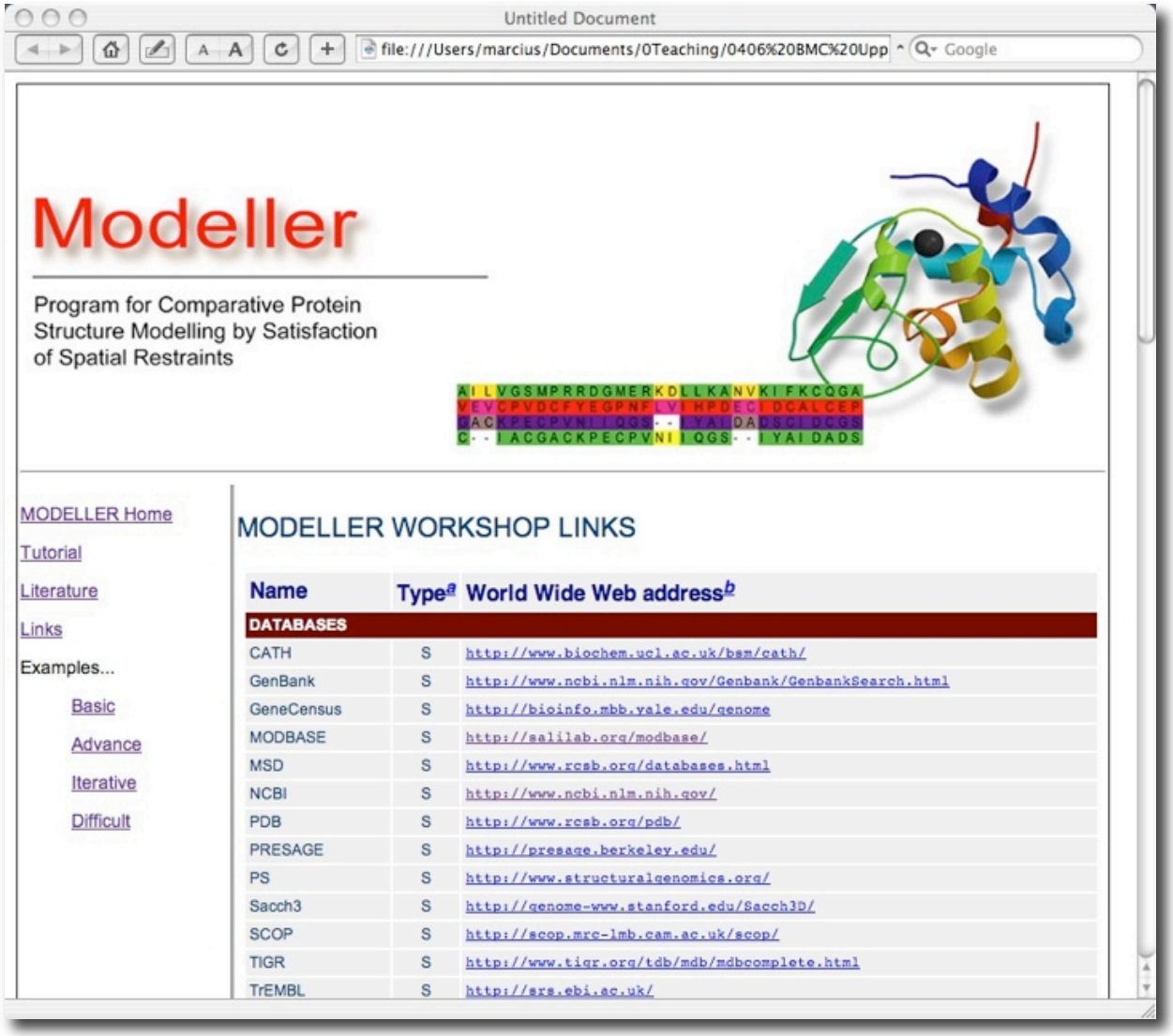
MODELLER:

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

Structural Genomics:

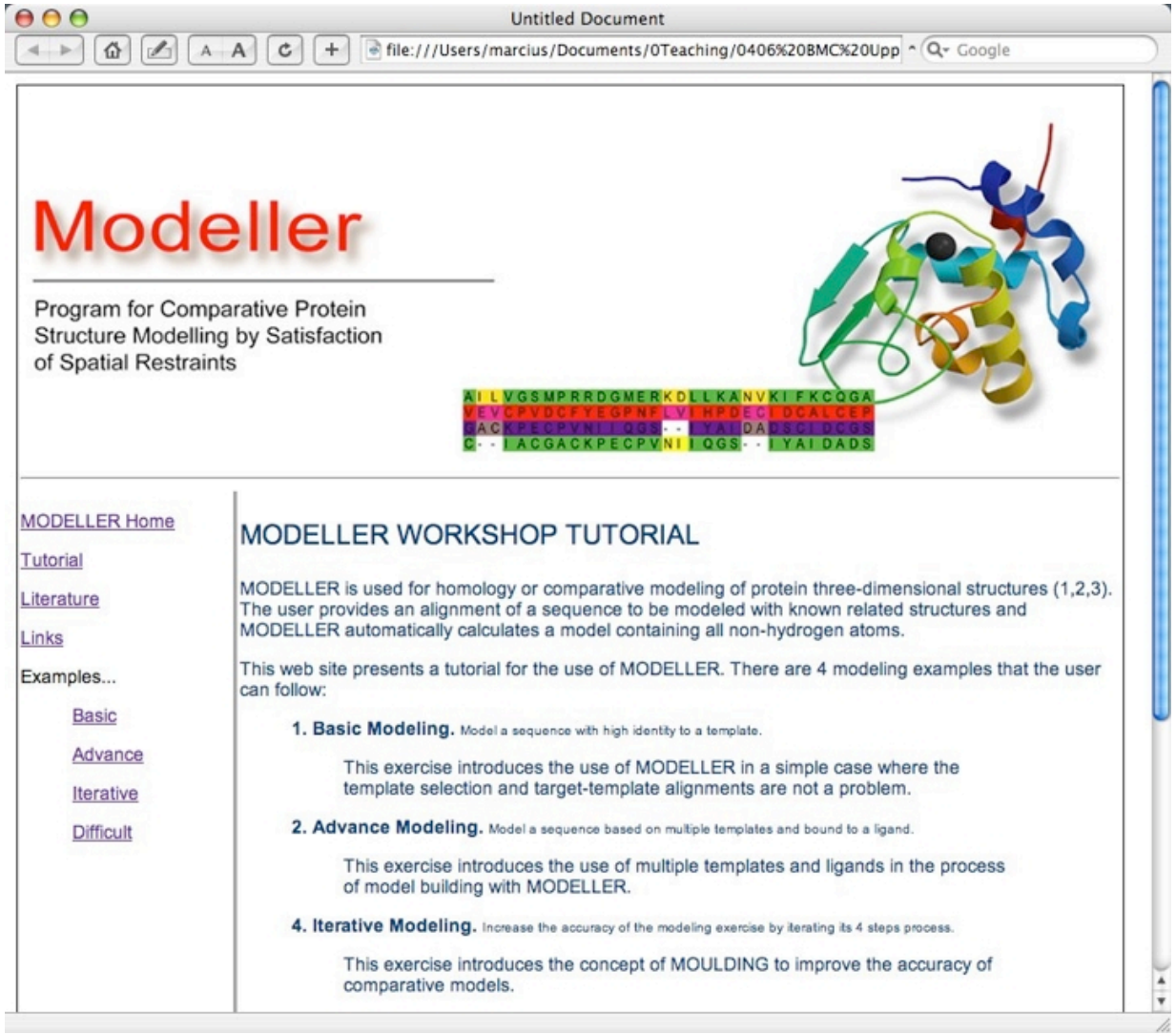
Sali. Nat. Struct. Biol. 5, 1029, 1998.
Burley et al. Nat. Genet. 23, 151, 1999.
Sali & Kuriyan. TIBS 22, M20, 1999.

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



Web site...

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



Acknowledgments

http://salilab.org

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

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)

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

Brain Lipid Binding Protein
Liang Zhu (RU)
Nat Heintz (RU)

Structural Genomics
Stephen Burley (SGX)
John Kuriyan (UCB)
NY-SGXRC

BRCA1
A. Monteiro (Cornel)

Mast Cell Proteases
Rick Stevens (BWH)

Fly p53
Shengkan Jin (RU)
Arnie Levine (RU)

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

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”

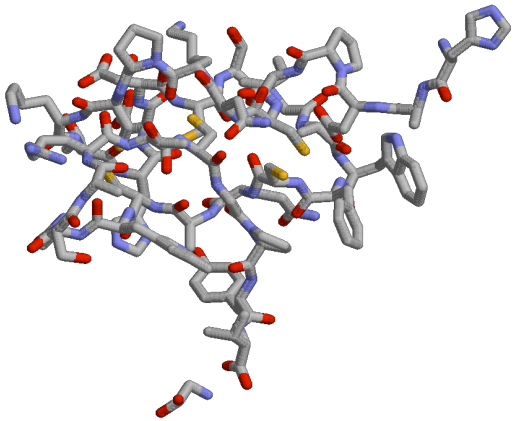
Structure-Structure alignments

As any other bioinformatics problem…

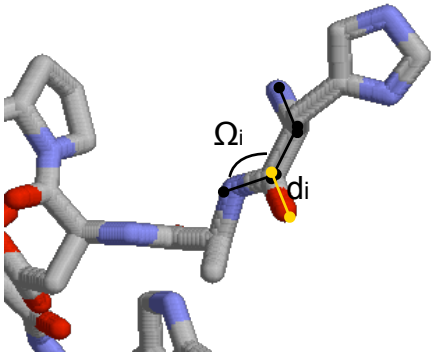
- **Representation**
- **Scoring**
- **Optimizer**

Representation

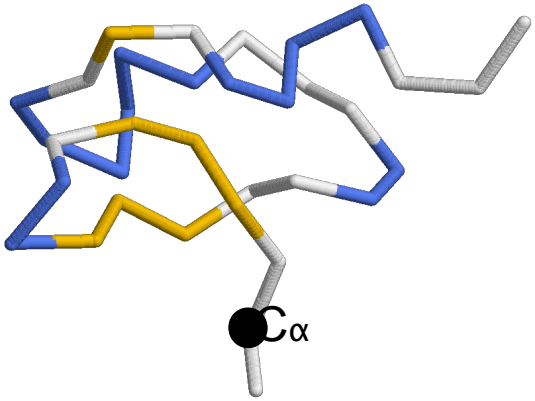
Structures



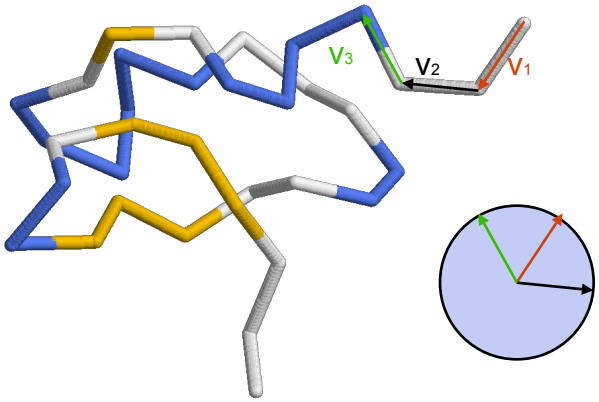
All atoms and coordinates



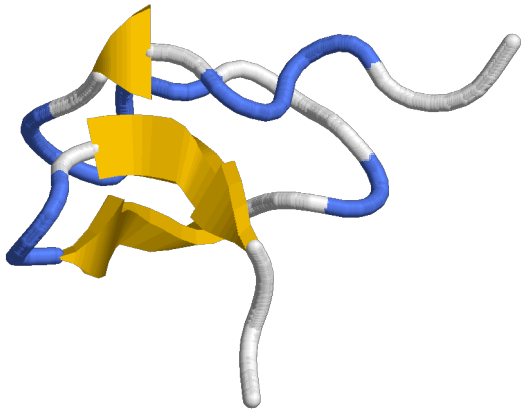
Dihedral space or distance space



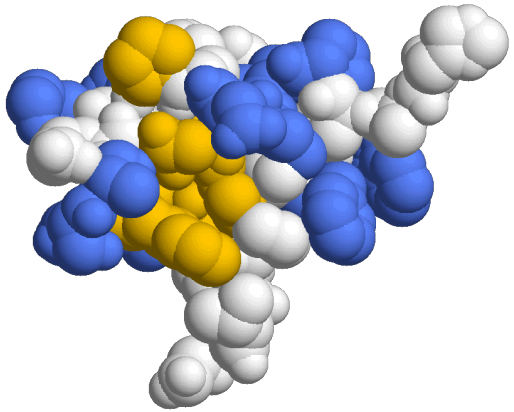
Reduced atom representation



Vector representation



Secondary Structure



Accessible surface (and others)

Scoring

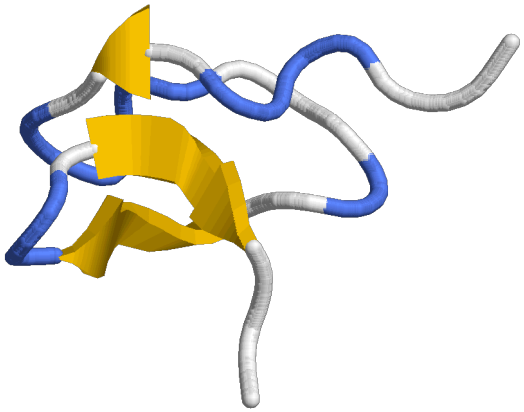
Raw scores

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

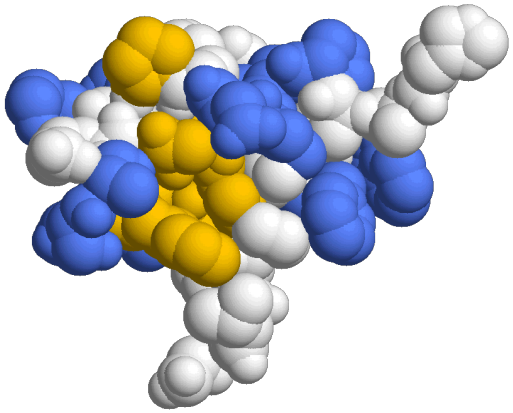
Aminoacid substitutions

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

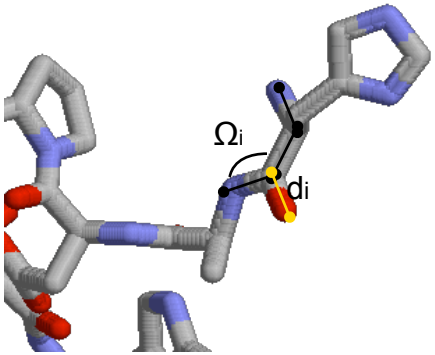
Root Mean Square Deviation



Secondary Structure (H,B,C)



Accessible surface (B,A [%])

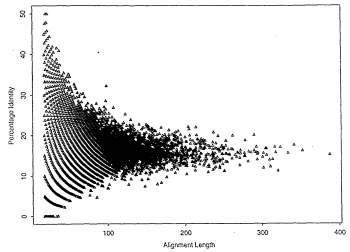


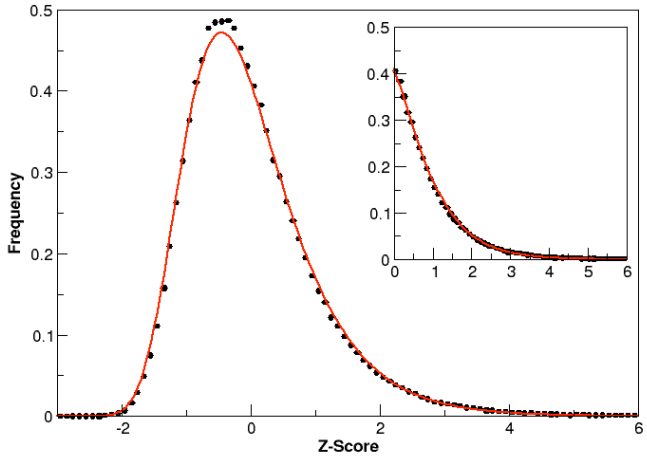
Angles or distances

Significance of an alignment (score)

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

Empirical





Sometimes approximated by Z-score (normal distribution).

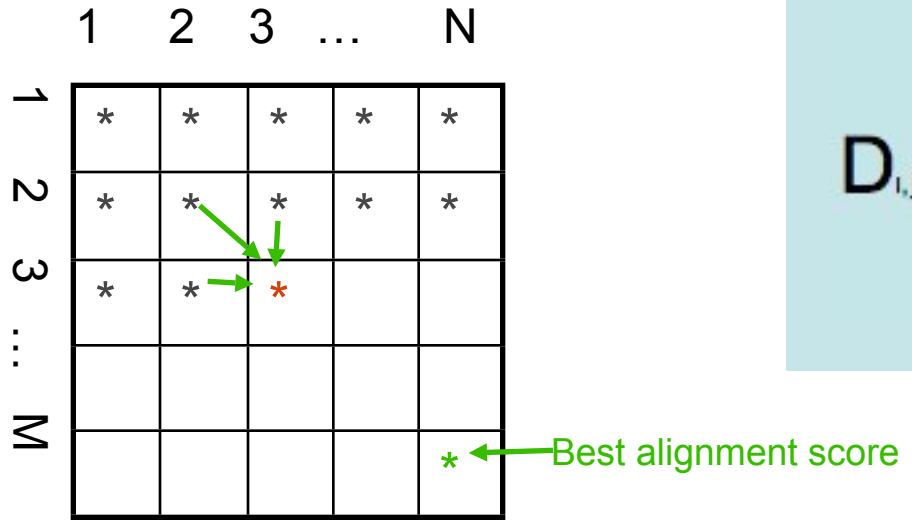
Analytic

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

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

Karlin and Altschul, 1990 *PNAS* **87**, pp2264

Global dynamic programming alignment

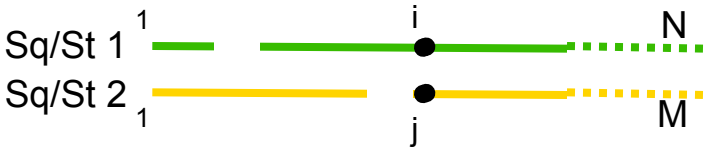


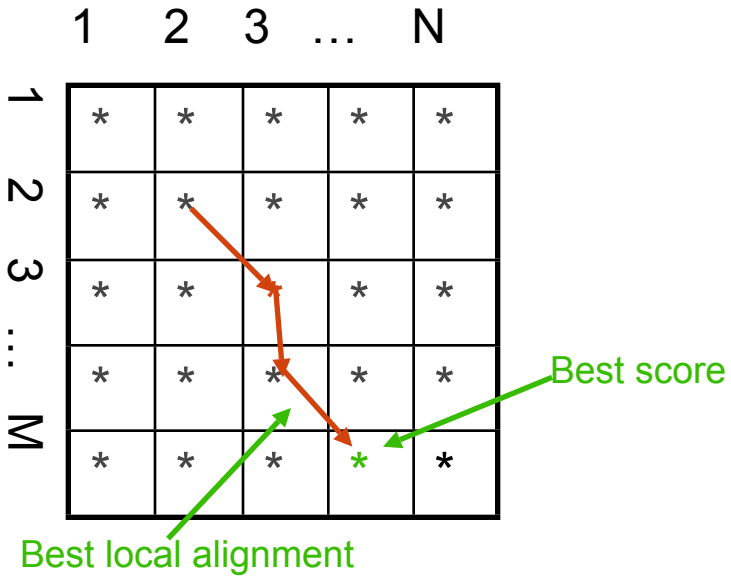
$$D_{i,j}=\min\begin{cases} D_{i,j-1}+\text{Score}_{(\Delta,rj)} \\ D_{i-1,j-1}+\text{Score}_{(ri,rj)} \\ D_{i-1,j}+\text{Score}_{(ri,\Delta)} \end{cases}$$

Backtracking to get the best alignment

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

Local dynamic programming alignment



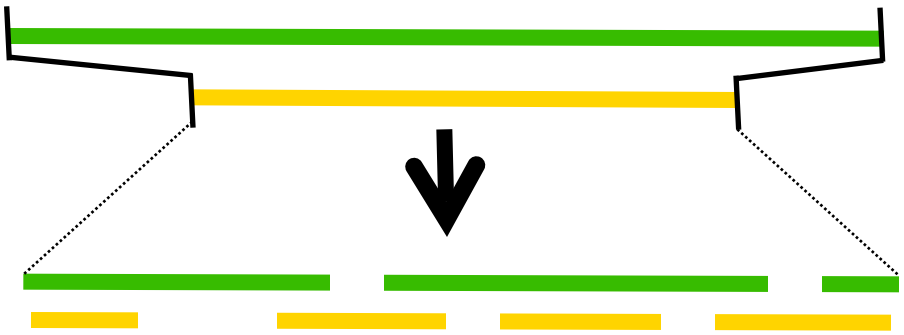


$$D_{i,j} = \min \begin{cases} D_{i,j-1} + Score_{(\Delta,rj)} \\ D_{i-1,j-1} + Score_{(ri,rj)} \\ D_{i-1,j} + Score_{(ri,\Delta)} \\ 0 \end{cases}$$

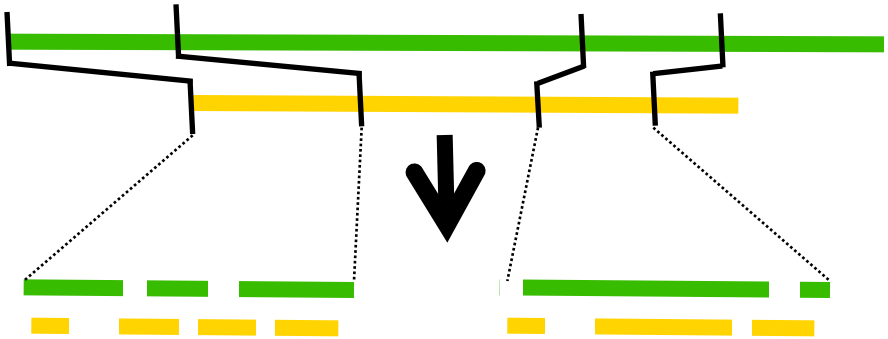
Backtracking to get the best alignment

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

Global *.vs.* local alignment



Global alignment



Local alignment

Multiple alignment

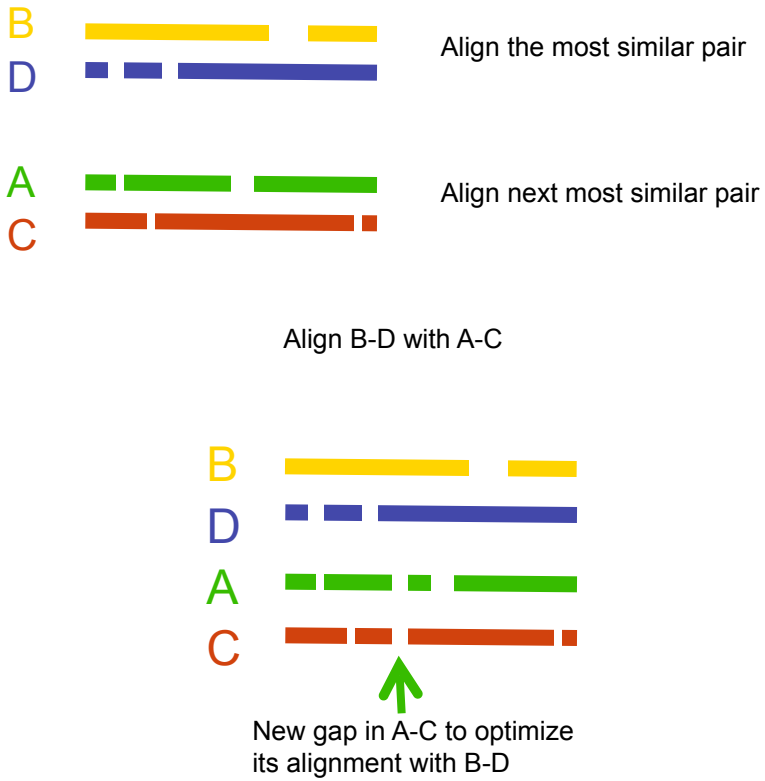
Pairwise alignments

Example – 4 sequences A, B, C, D.

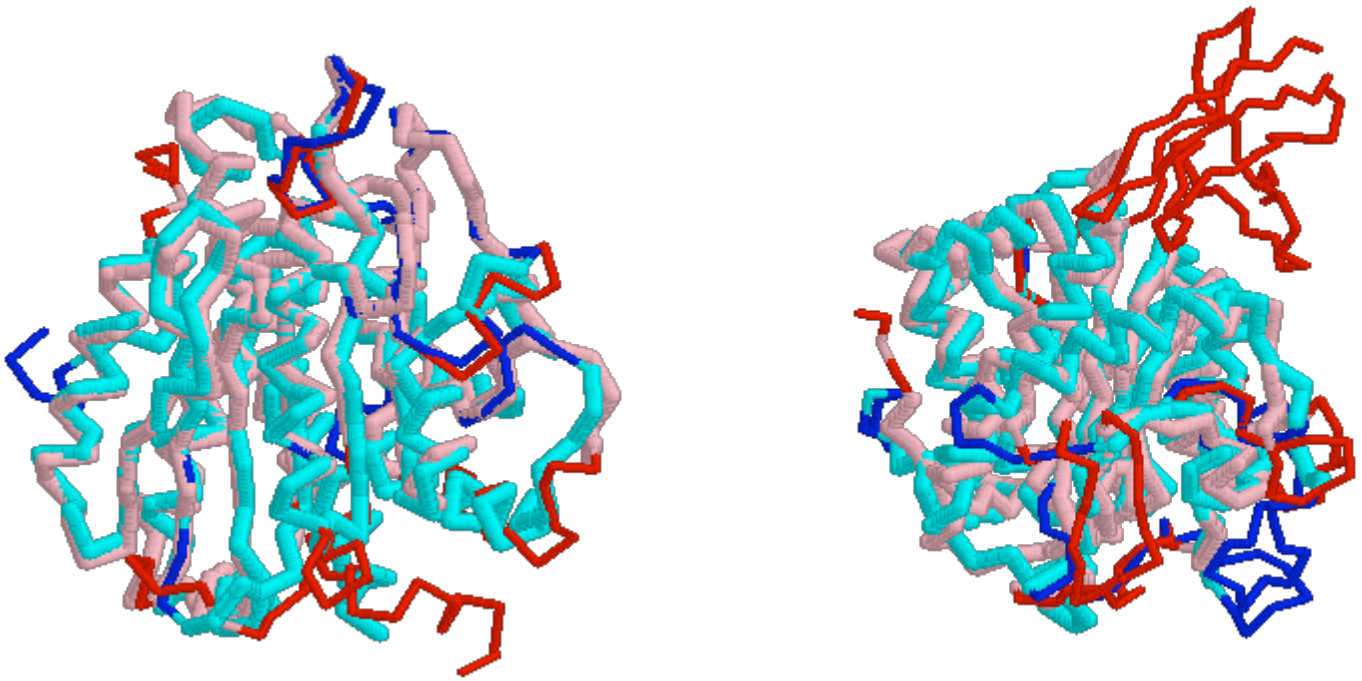
6 pairwise comparisons
then cluster analysis

Multiple alignments

Following the tree from step 1



Coverage .vs. Accuracy

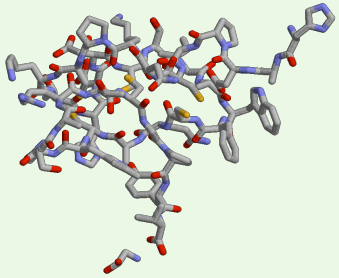


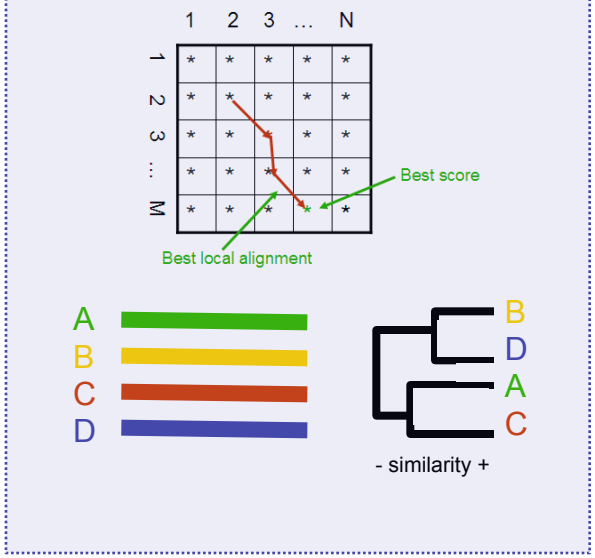
Same RMSD ~ 2.5Å

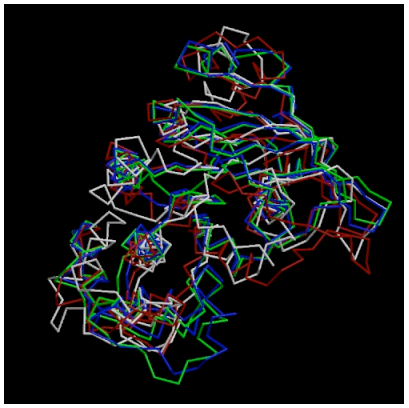
Coverage ~90% Cα

Coverage ~75% Cα

Structural alignment by properties conservation (SALIGN-MODELLER)

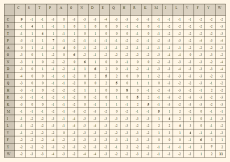


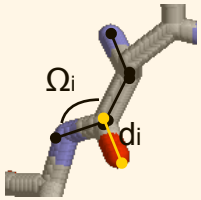


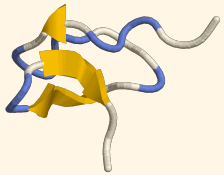


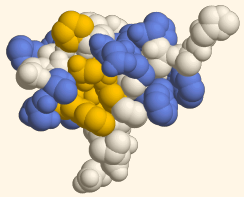
✓ Uses all available structural information
✓ Provides the optimal alignment

Computationally expensive









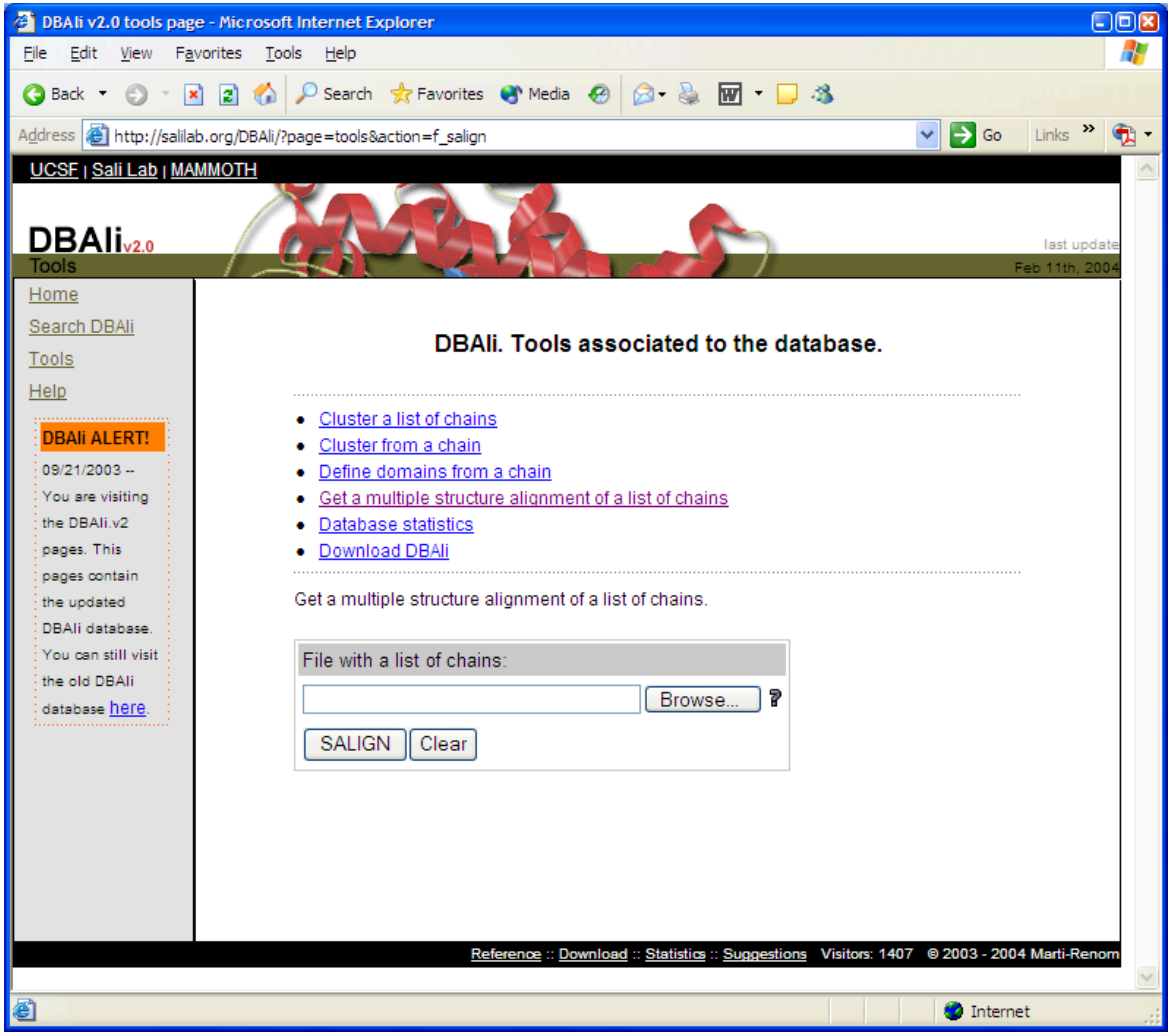
$$RMSD(x,y) = \sqrt{\left(\frac{1}{N}\right)\sum_{i=1}^{N}\left(\left\|\mathbf{x}(i)-\mathbf{y}(i)\right\|^{2}\right)}$$

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

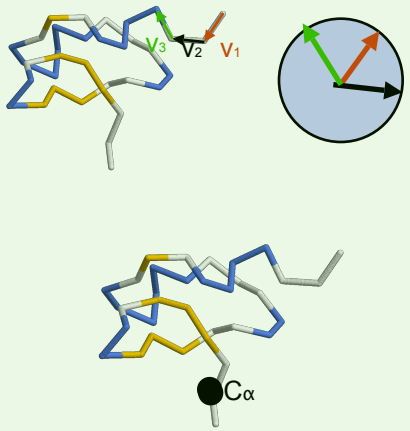
Madhusudhan et al. *in preparation*

Structural alignment by properties conservation (SALIGN-MODELLER)

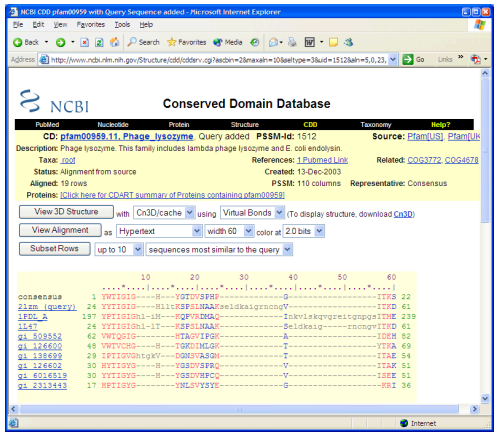
http://www.salilab.org/dbali/



Vector Alignment Search Tool (VAST)

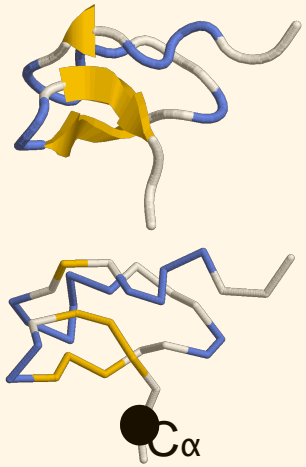


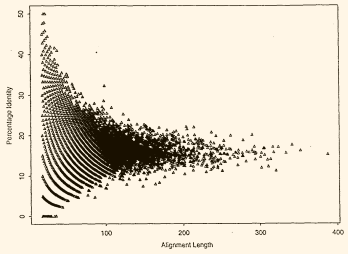
- Graph theory search of similar SSE
- Refining by Monte Carlo at all atom resolution



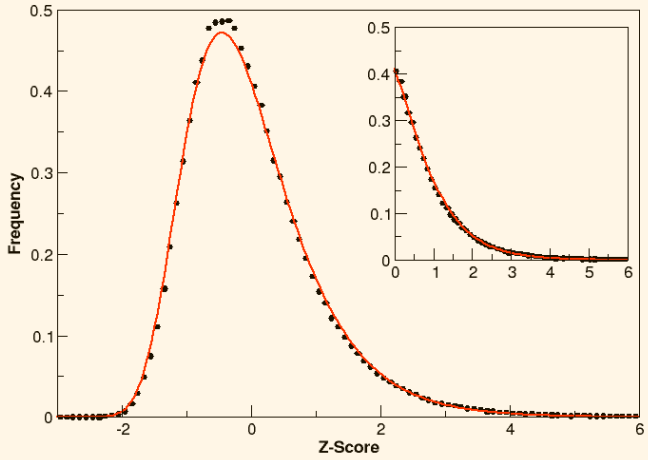
✓ **Good scoring system with significance**

Reduces the protein representation





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

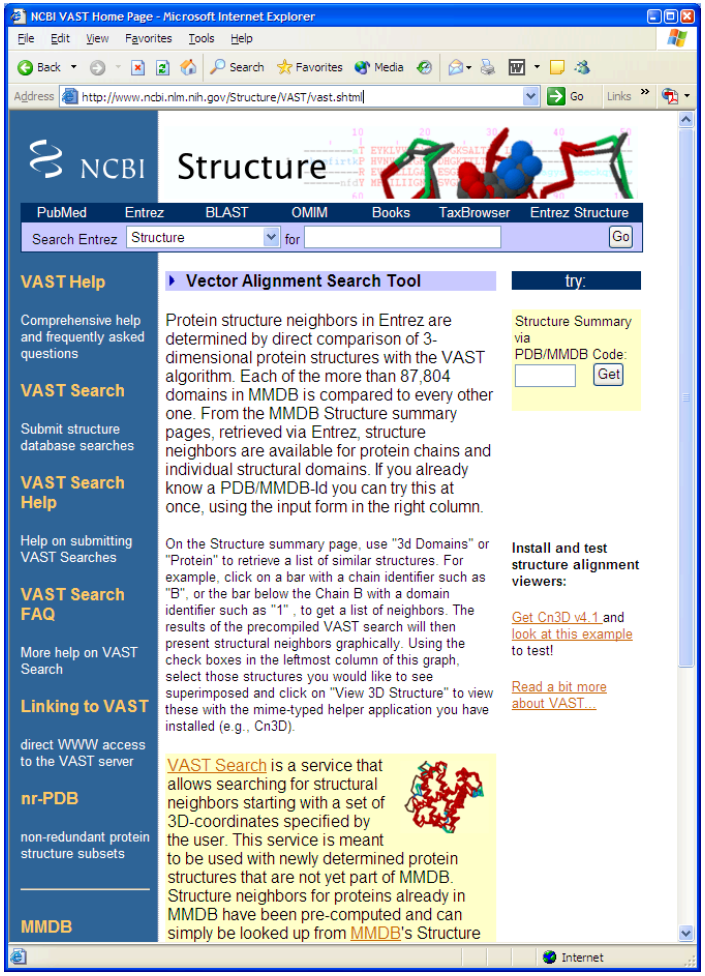


Gibrat JF et al. (1996) *Curr Opin Struct Biol* **3** pp377

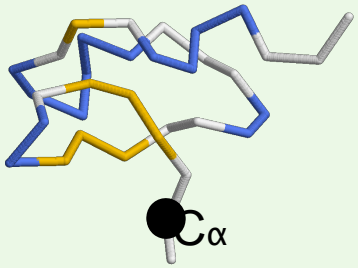
06/10/2004

Vector Alignment Search Tool (VAST)

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



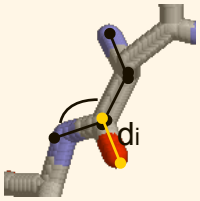
Incremental combinatorial extension (CE)



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

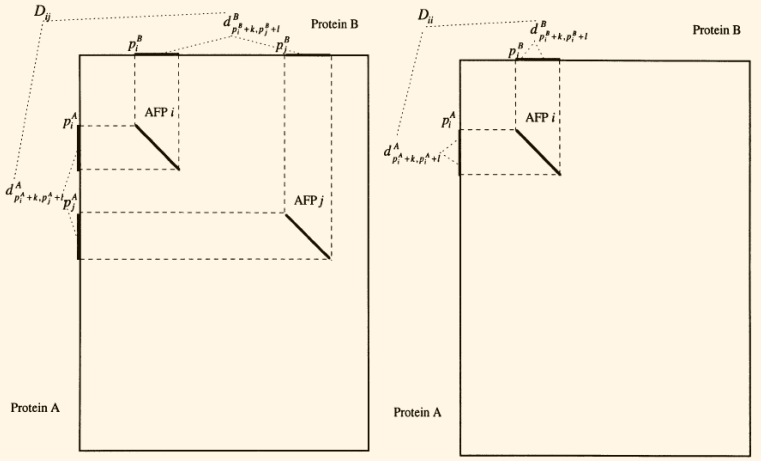
✓ **FAST!**
✓ **Good quality of local alignments**

Complicated scoring and heuristics



8 residues peptides

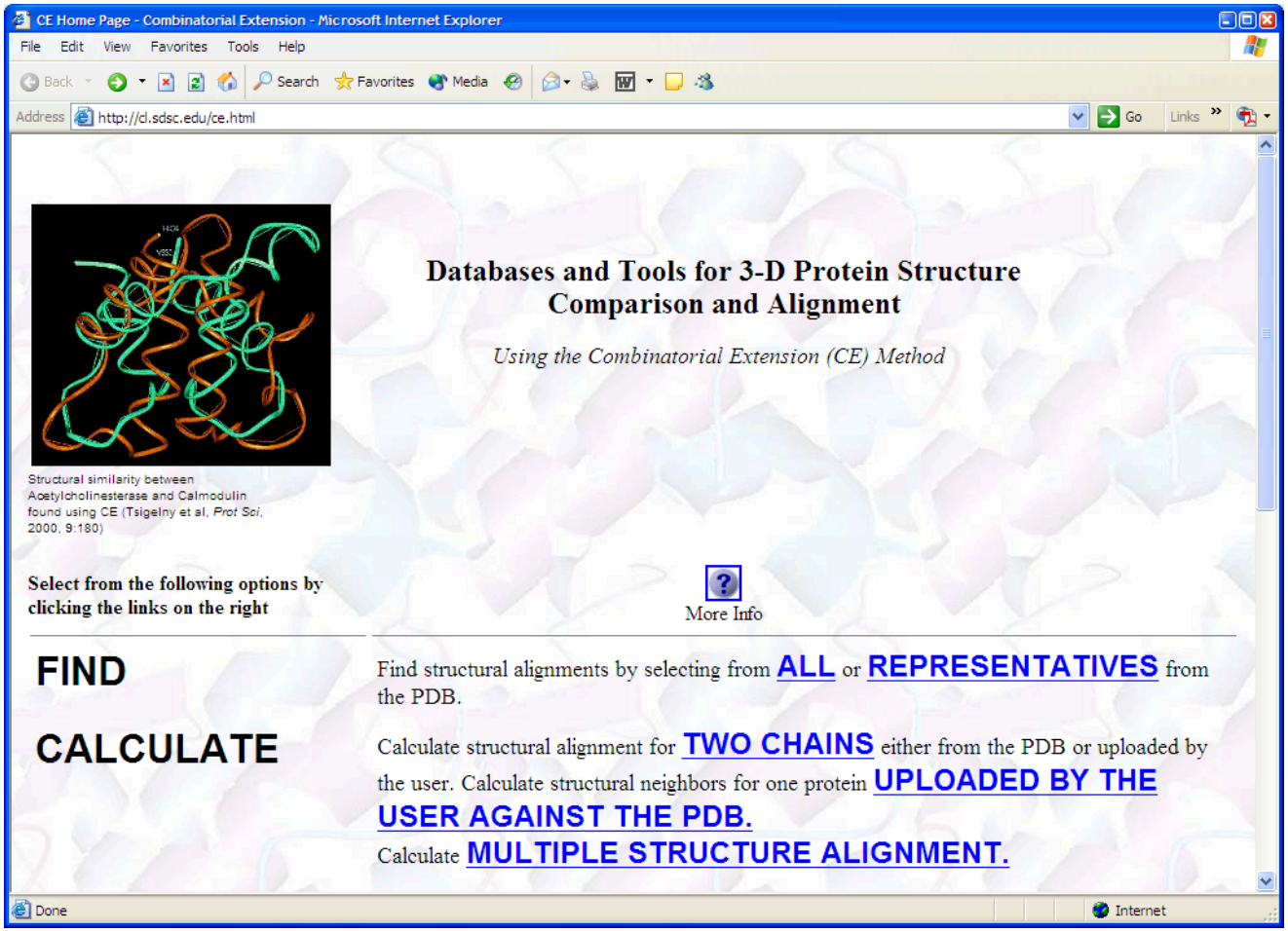
$$RMSD(x,y) = \sqrt{\left(\frac{1}{N}\right)\sum_{i=1}^{N}\left(\left\|\mathbf{x}(i)-\mathbf{y}(i)\right\|^{2}\right)}$$



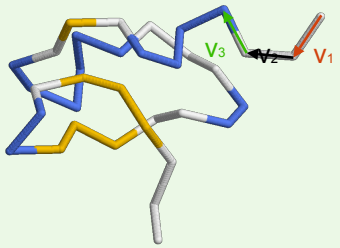
Shindyalov IN, amd Bourne PE. (1998) *Protein Eng.* **9** pp739

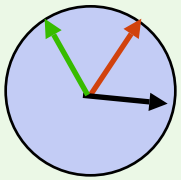
Incremental combinatorial extension (CE)

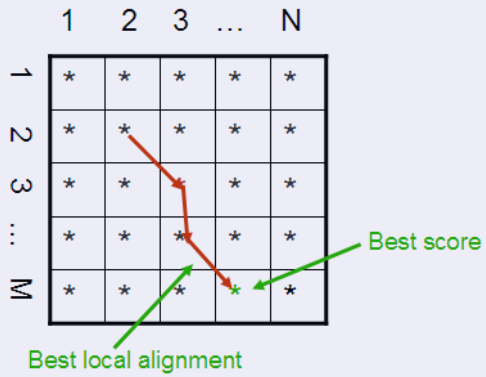
http://cl.sdsc.edu/ce.html

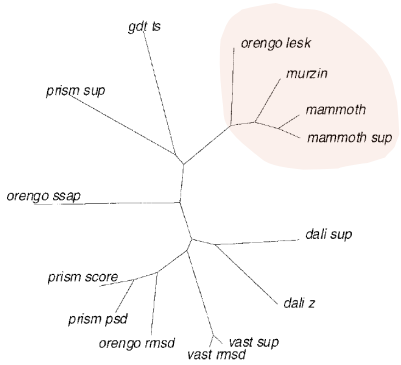


Matching molecular models obtained from theory (MAMMOTH)







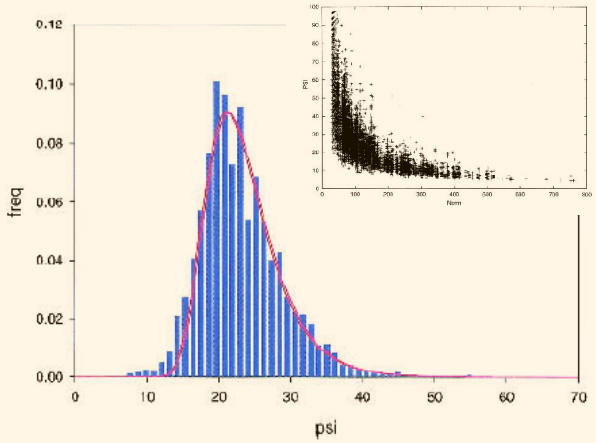


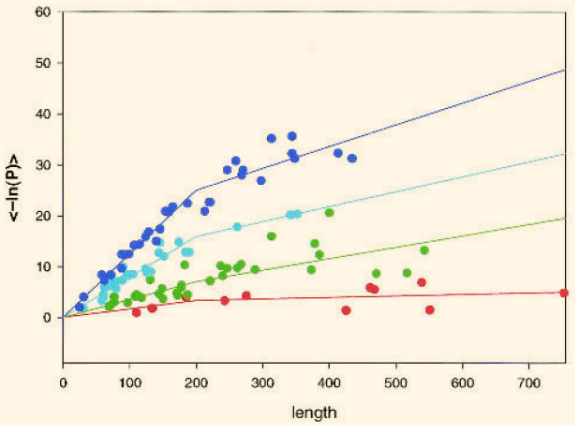
- ✓ **VERY FAST!**
- ✓ **Good scoring system with significance**

Reduces the protein representation

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

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

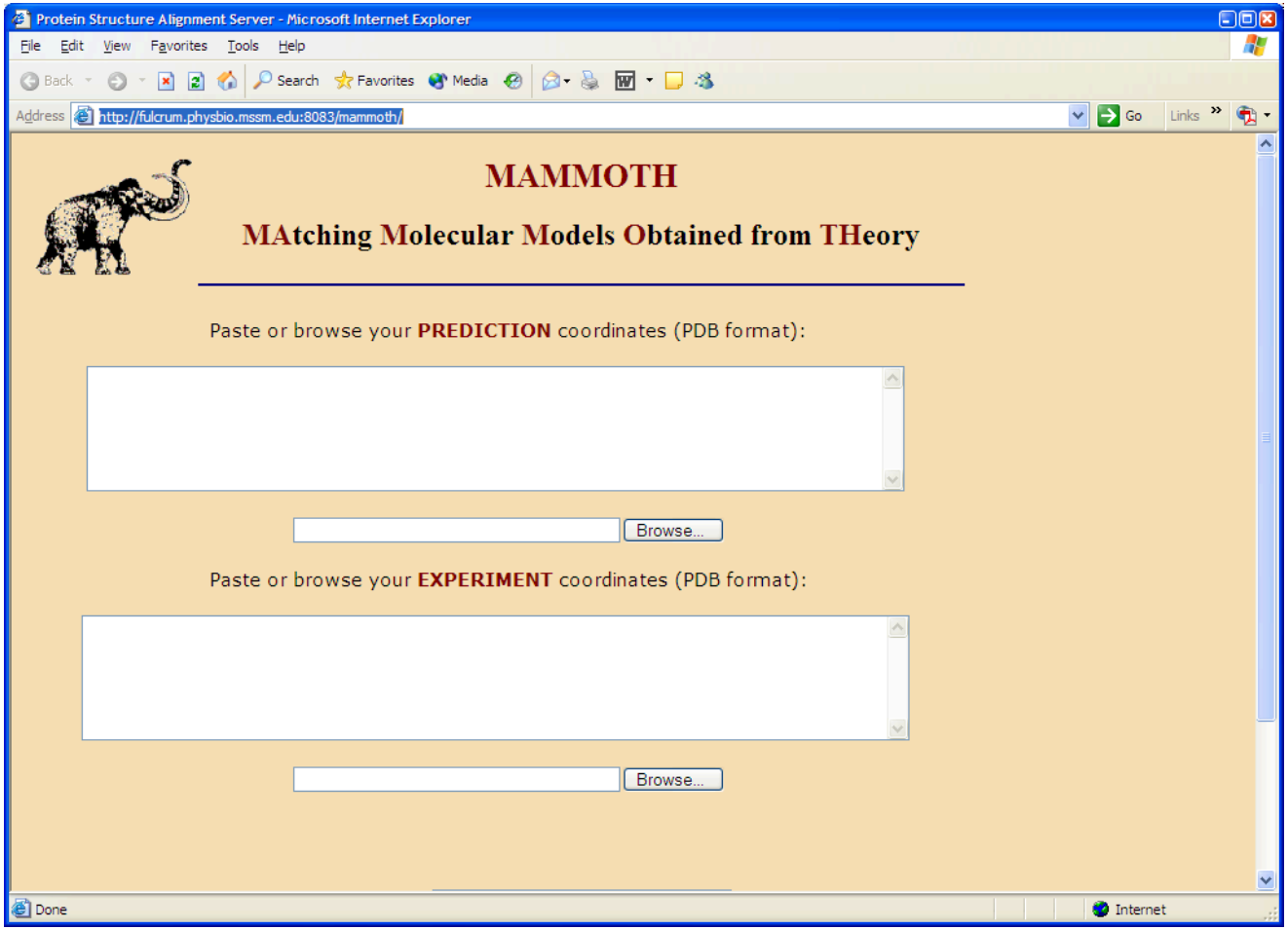




Ortiz AR, (2002) *Protein Sci.* **11** pp2606

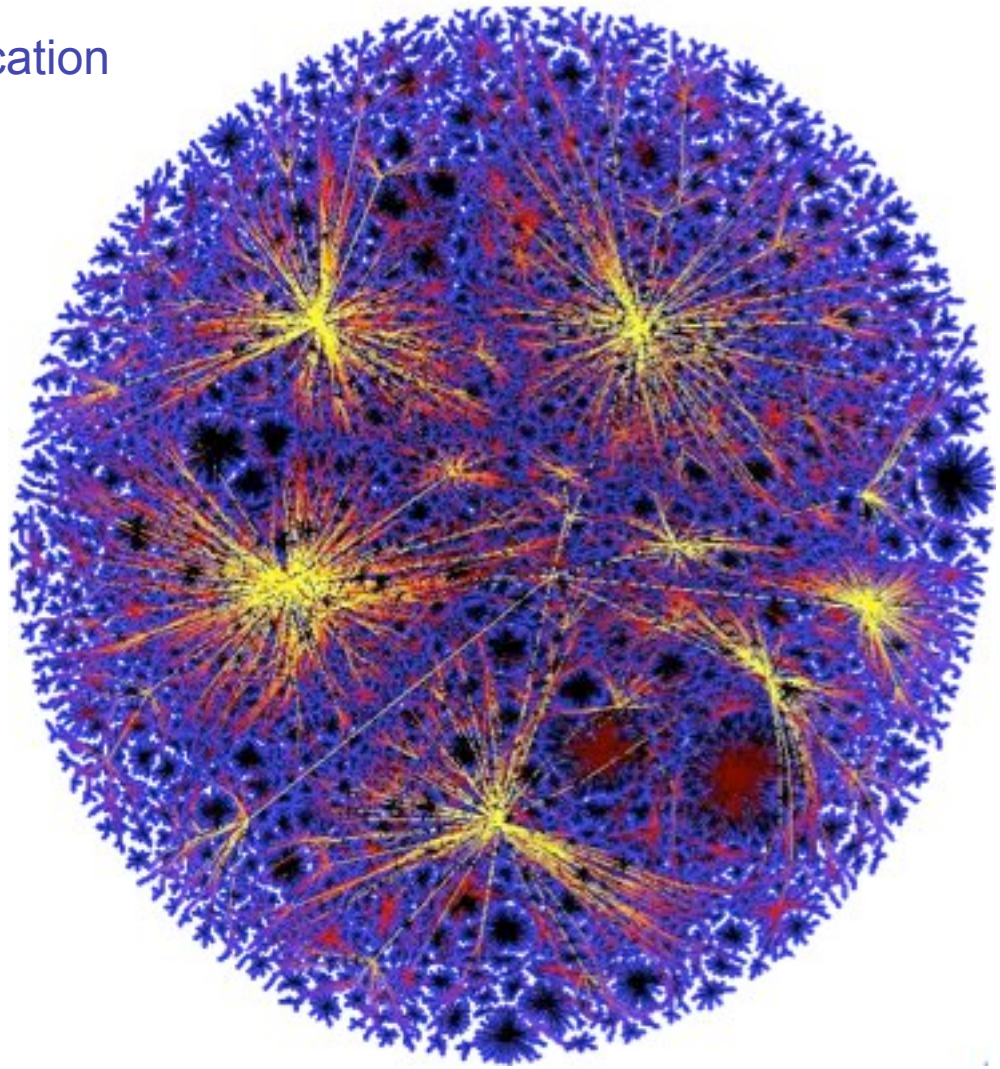
Matching molecular models obtained from theory (MAMMOTH)

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



Classification of the structural space

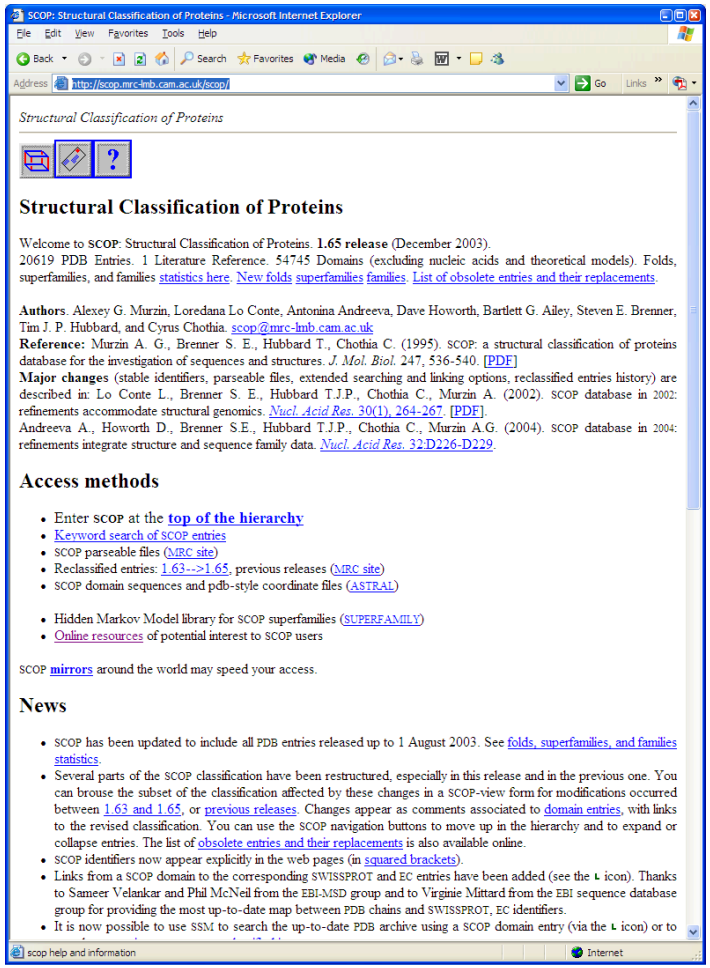
SCOP classification



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

SCOP$_{1.65}$ database

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



- ✓ Largely recognized as “standard of gold”
- ✓ Manually classification
- ✓ Clear classification of structures in:

CLASS
FOLD
SUPER-FAMILY
FAMILY

- ✓ Some large number of tools already available

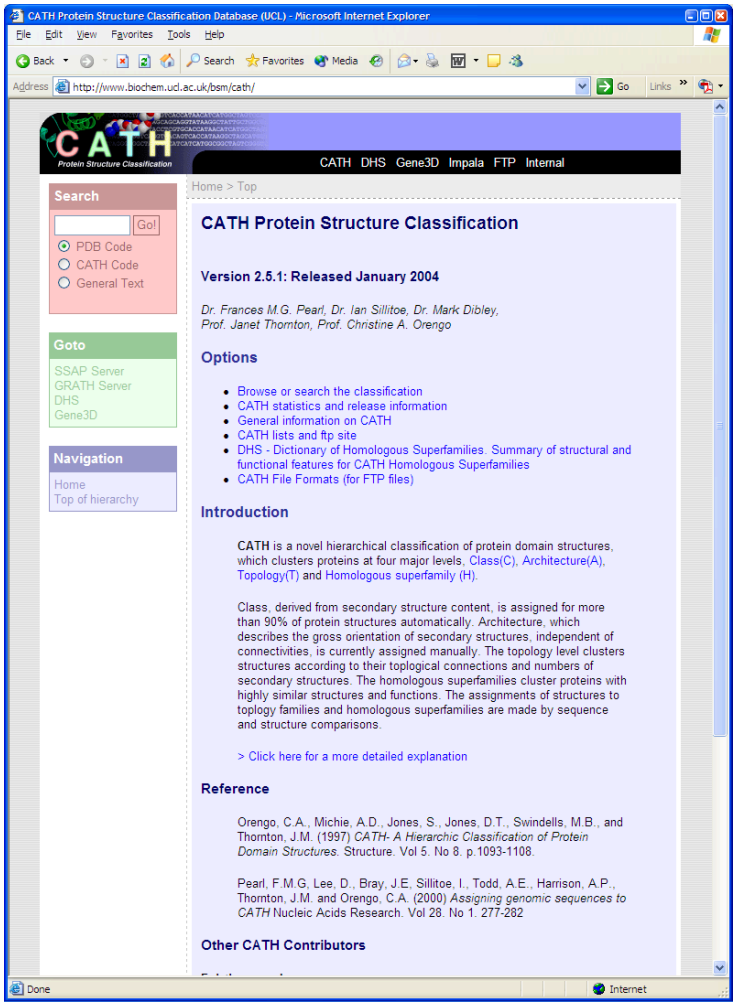
Manually classification
Not 100% up-to-date
Domain boundaries definition

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

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

CATH$_{2.5.1}$ database

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



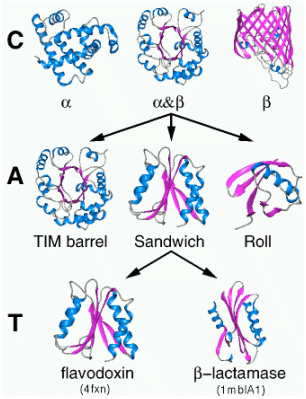
Uses FSSP for superimposition

- ✓ **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



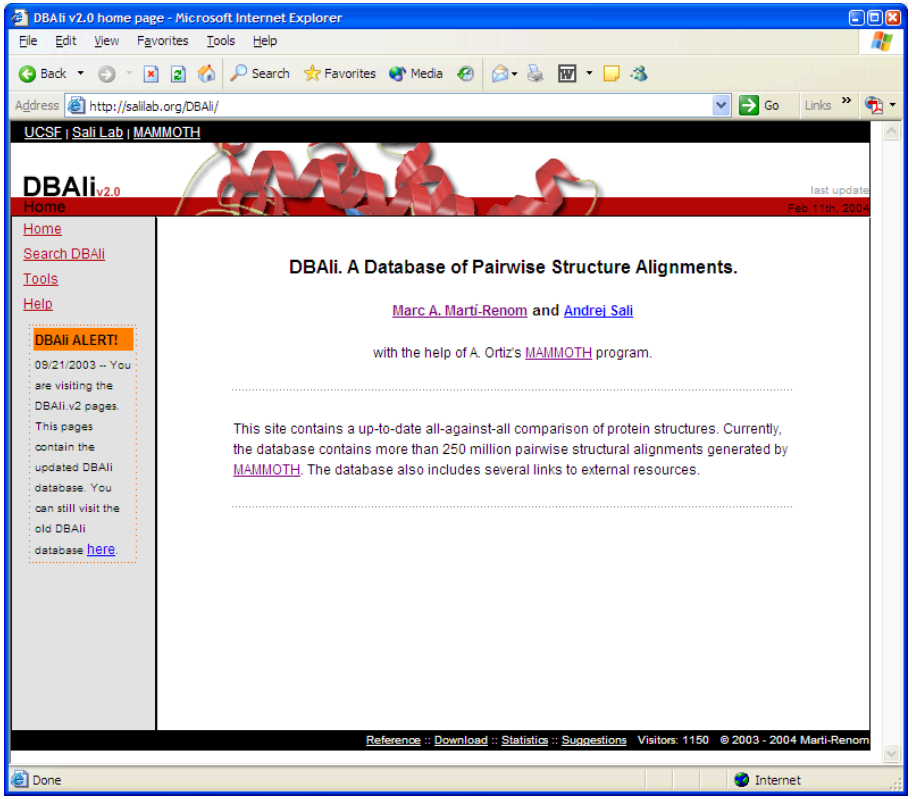
Version	2.5.1
Date	28-01-2004

	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 assigments	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.

DBAli$_{v2.0}$ database

http://salilab.org/DBAli/



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

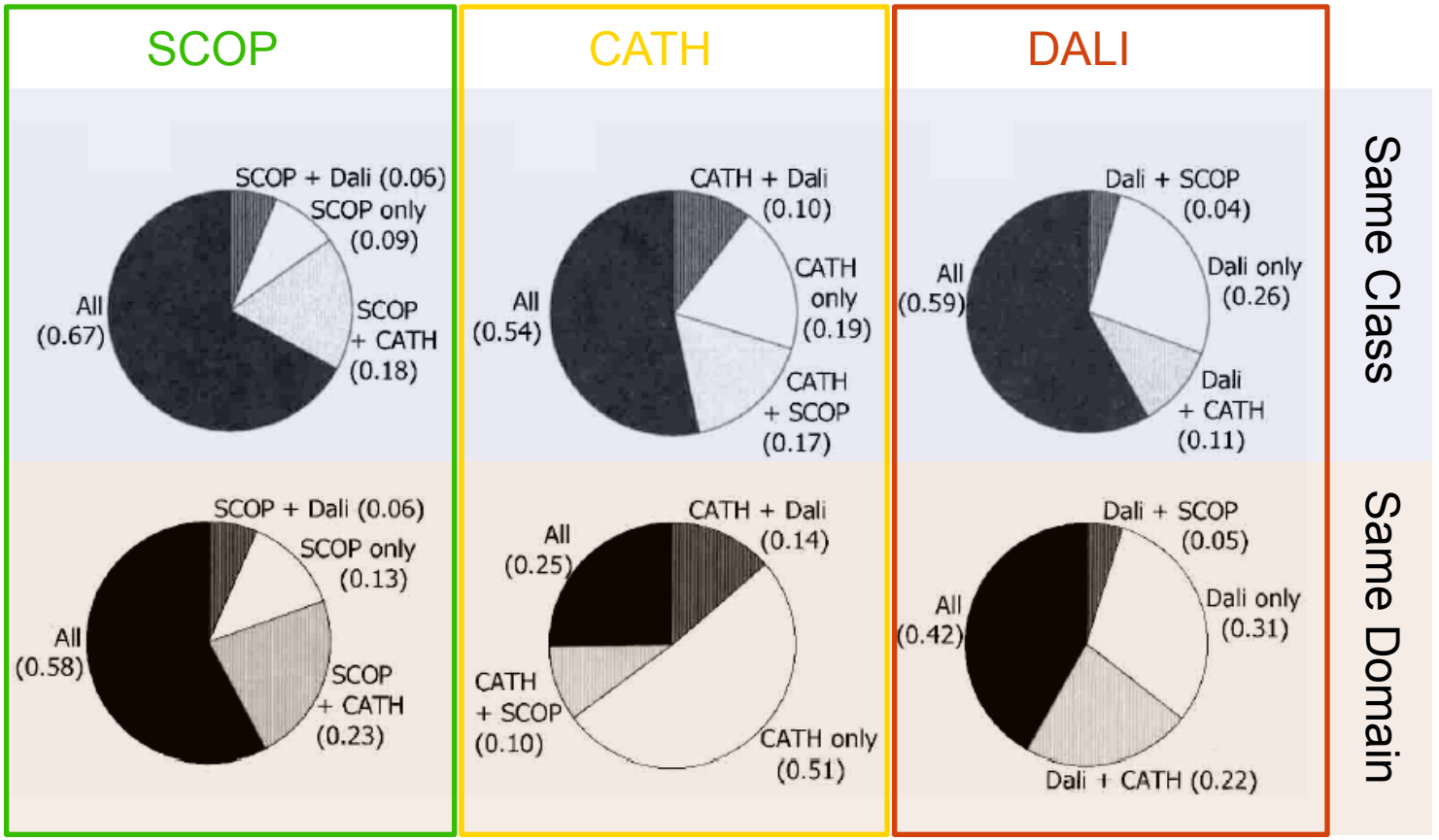
Last updated:	February 11th, 2004 (18:49h)
Number of chains in database:	48,094
Number of structure-structure comparisons:	330,514,636

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

Classification of the structural space

Not an easy task!

Domain definition AND domain classification

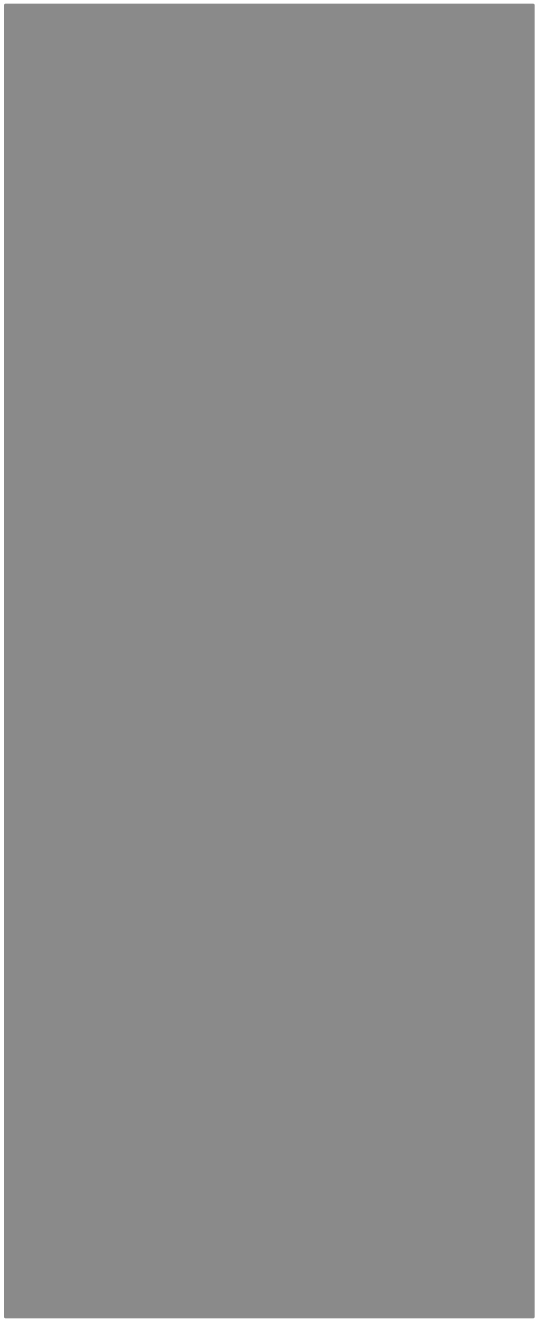


Day, et al. (2003) *Protein Sciences*, **12** pp2150

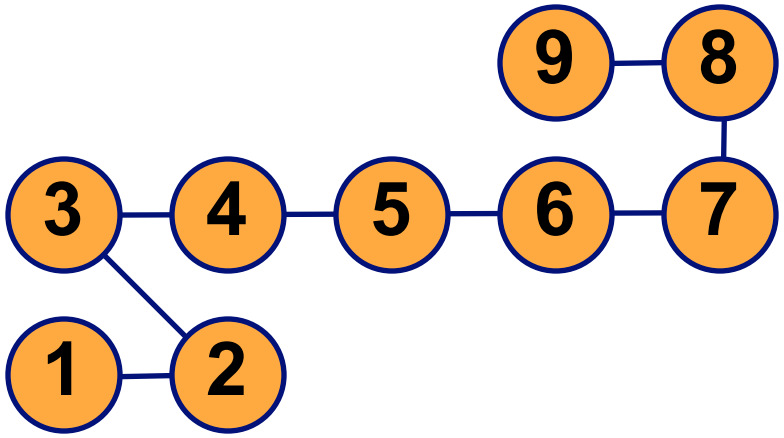
Application (ModDom)

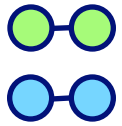
- Use of the DBAli data to define…
  - Protein Domains
  - Protein Fragments

ModDom

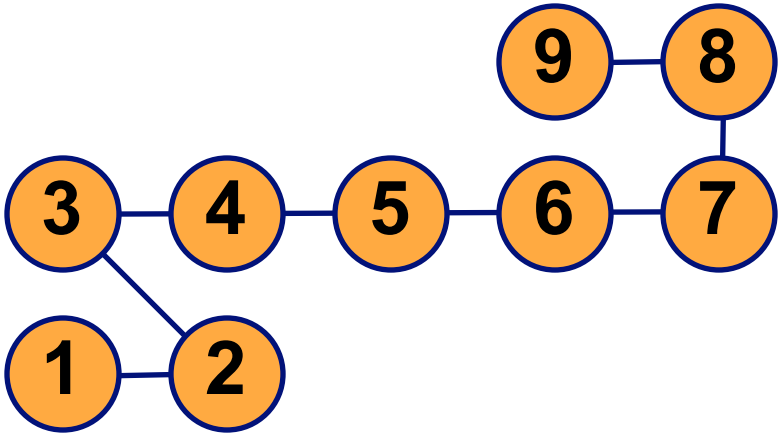


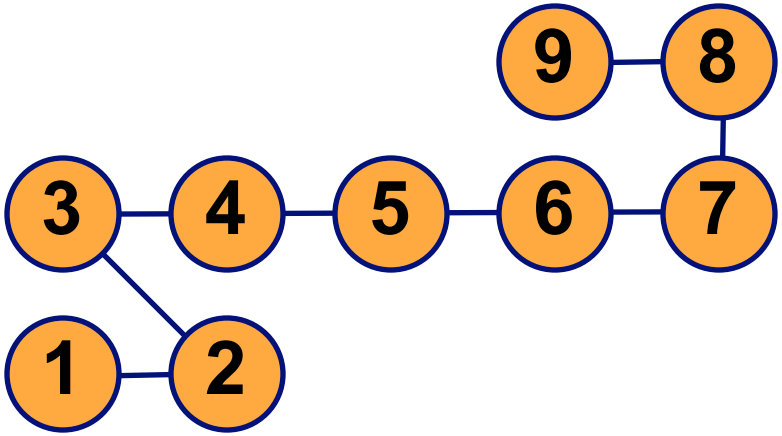
1phh (Oxydoreductase from *Pseudomonas fluorescens*)

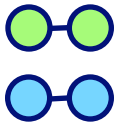

9
8
3
4
5
6
7
1
2



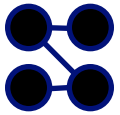
{1,2}{3,4}{4,5}
{5,6}{6,7}{7,8}{8,9}



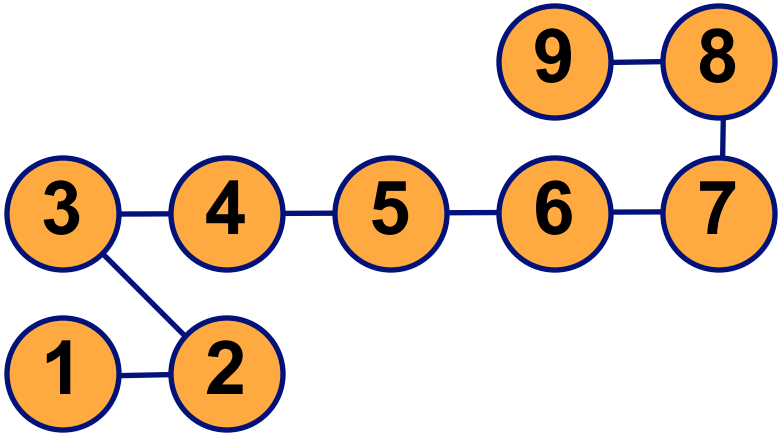


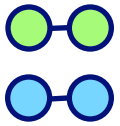


{1,2}{3,4}{4,5}
{5,6}{6,7}{7,8}{8,9}

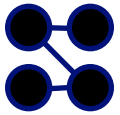


{1,2,3,4}

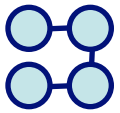




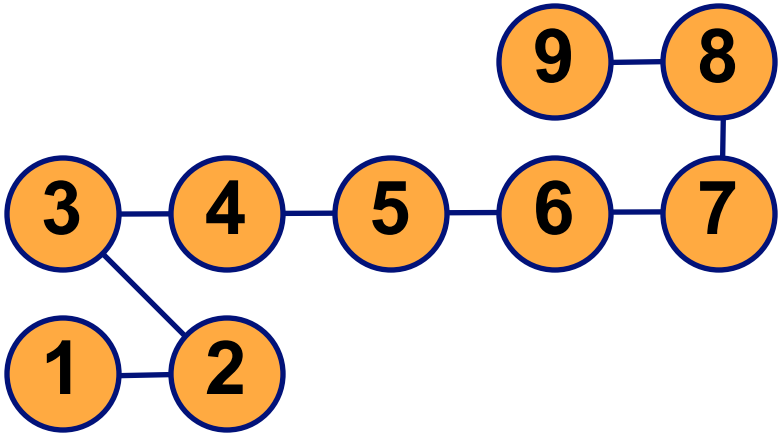
{1,2}{3,4}{4,5}
{5,6}{6,7}{7,8}{8,9}

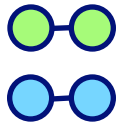


{1,2,3,4}

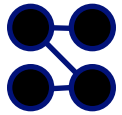


{6,7,8,9}

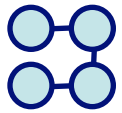




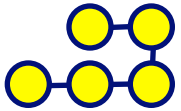
{1,2}{3,4}{4,5}
{5,6}{6,7}{7,8}{8,9}



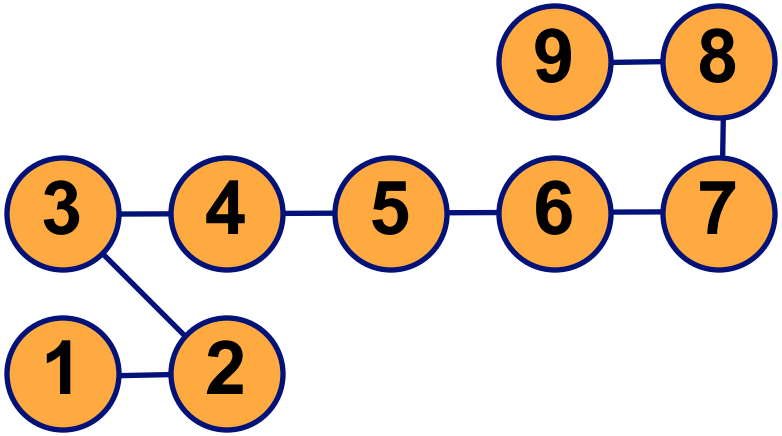
{1,2,3,4}

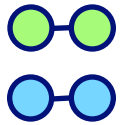


{6,7,8,9}

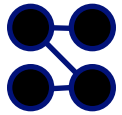


{5,6,7,8,9}

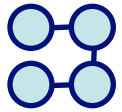




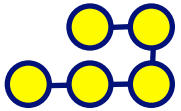
{1,2}{3,4}{4,5}
{5,6}{6,7}{7,8}{8,9}



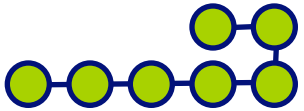
{1,2,3,4}



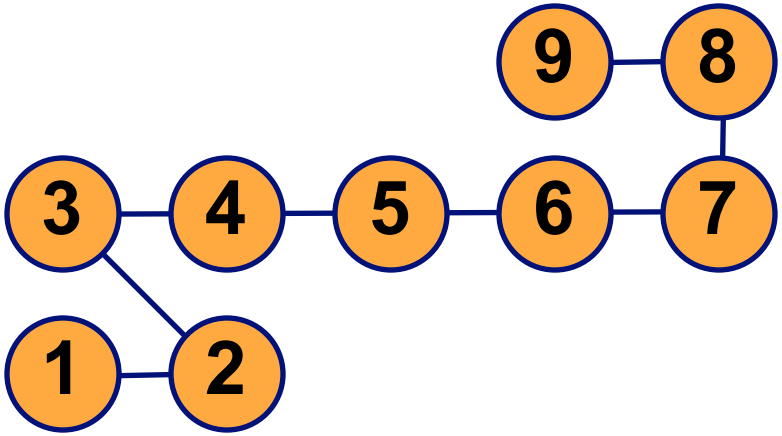
{6,7,8,9}

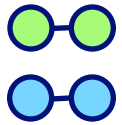


{5,6,7,8,9}

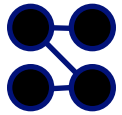


{3,4,5,6,7,8,9}

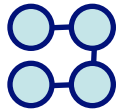




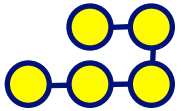
{1,2}{3,4}{4,5}
{5,6}{6,7}{7,8}{8,9}



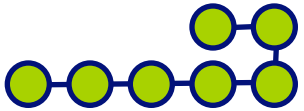
{1,2,3,4}



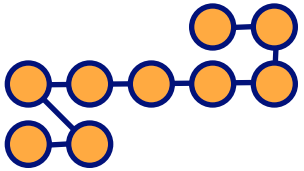
{6,7,8,9}



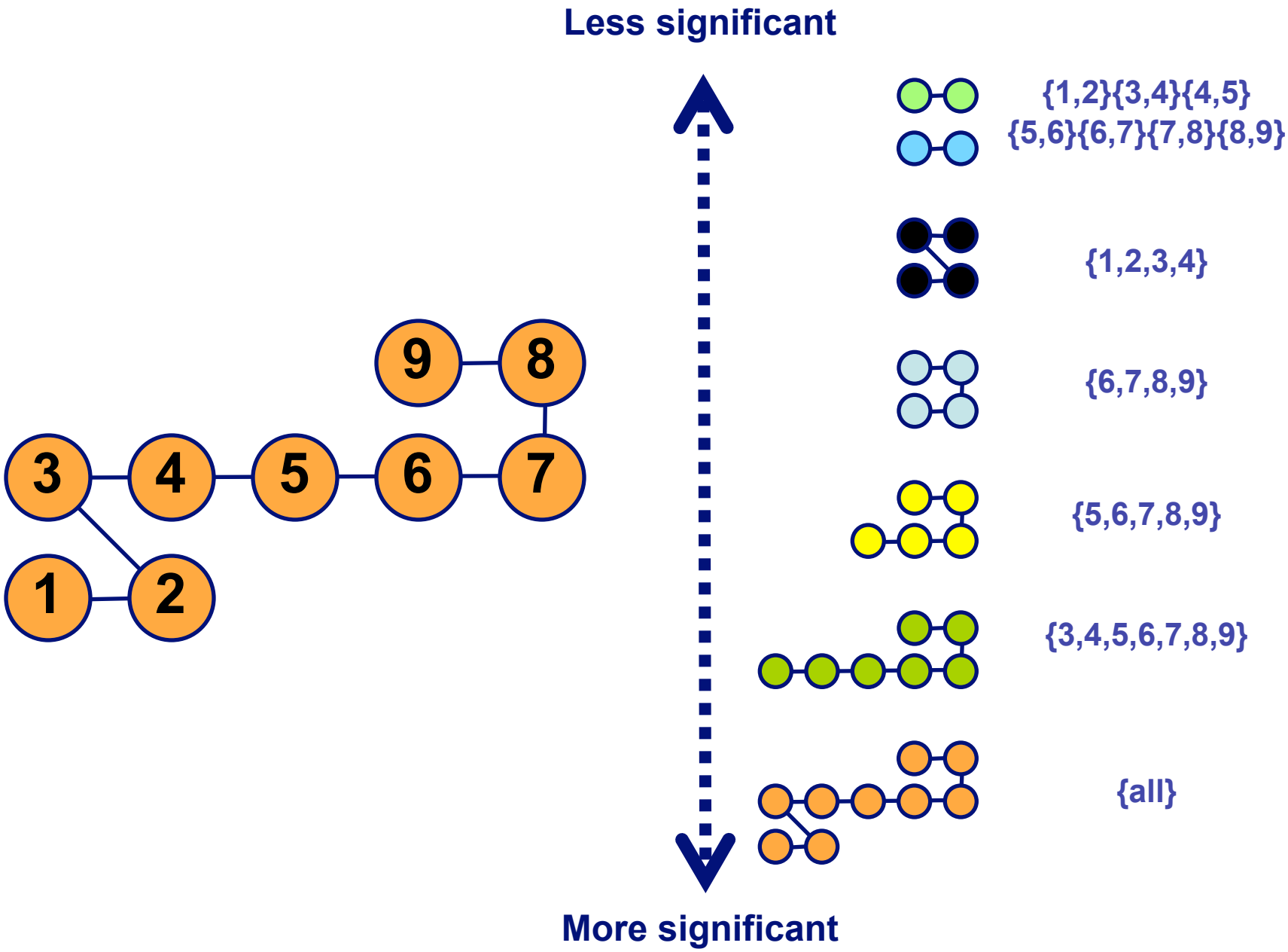
{5,6,7,8,9}

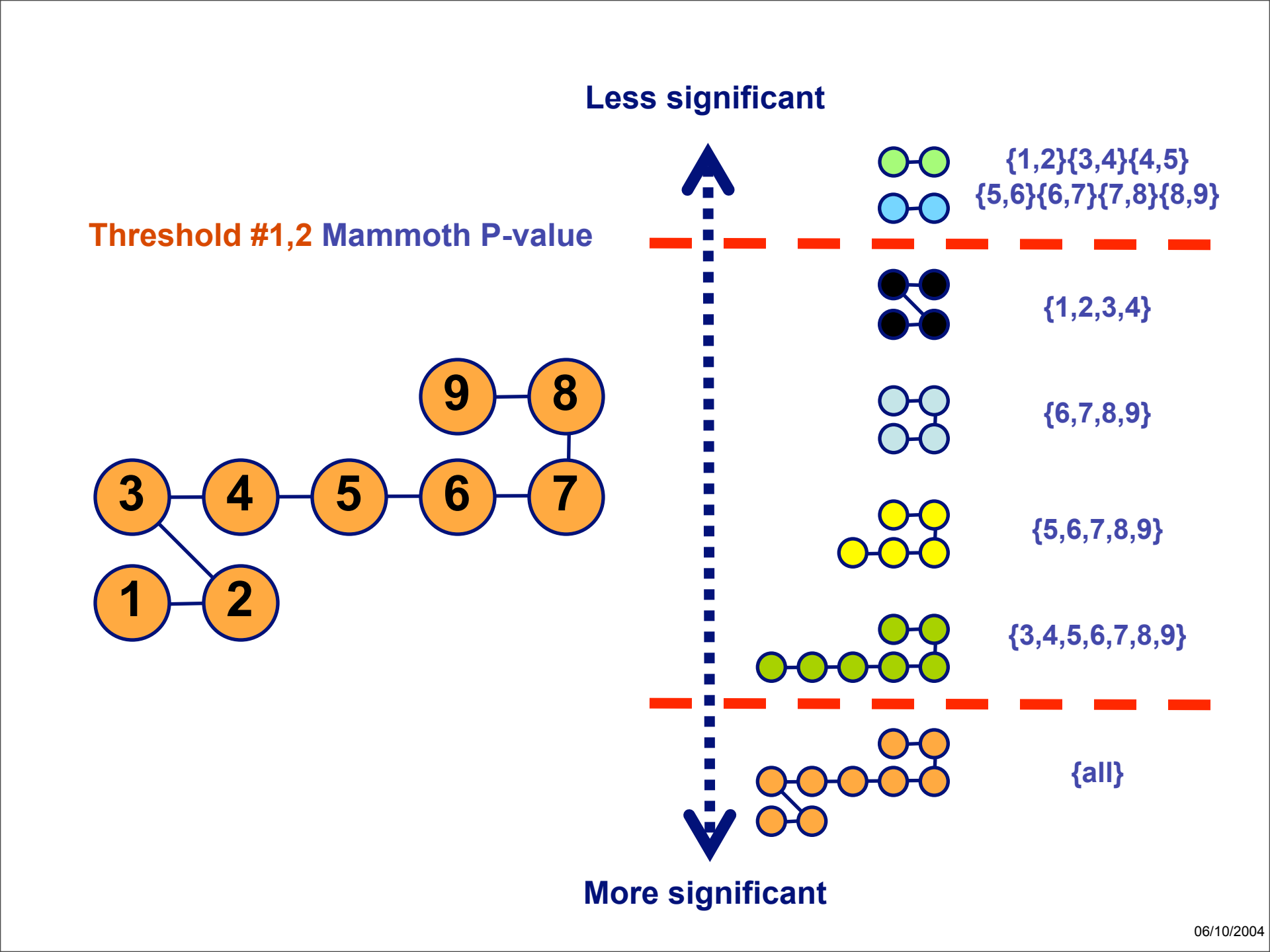


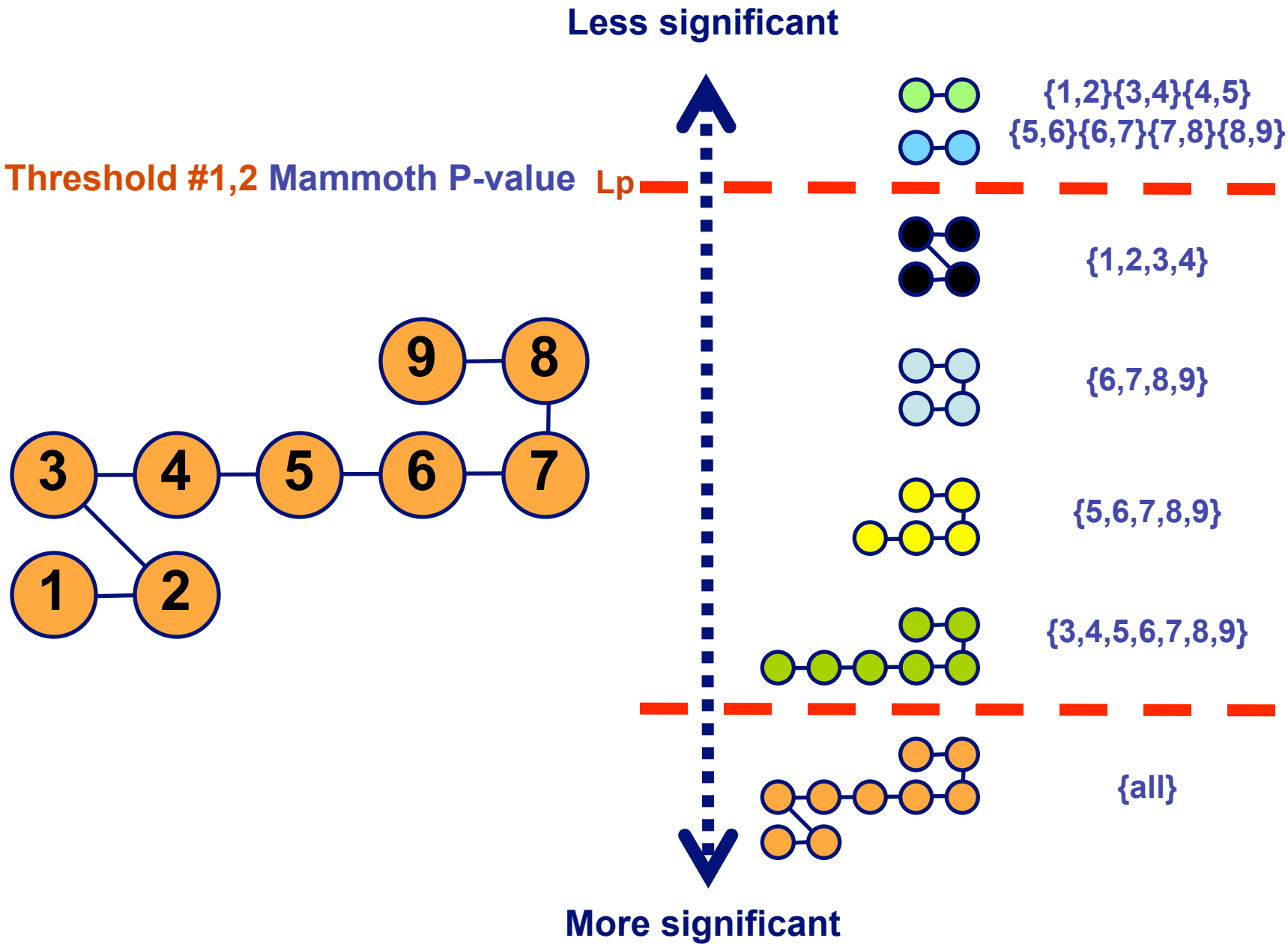
{3,4,5,6,7,8,9}

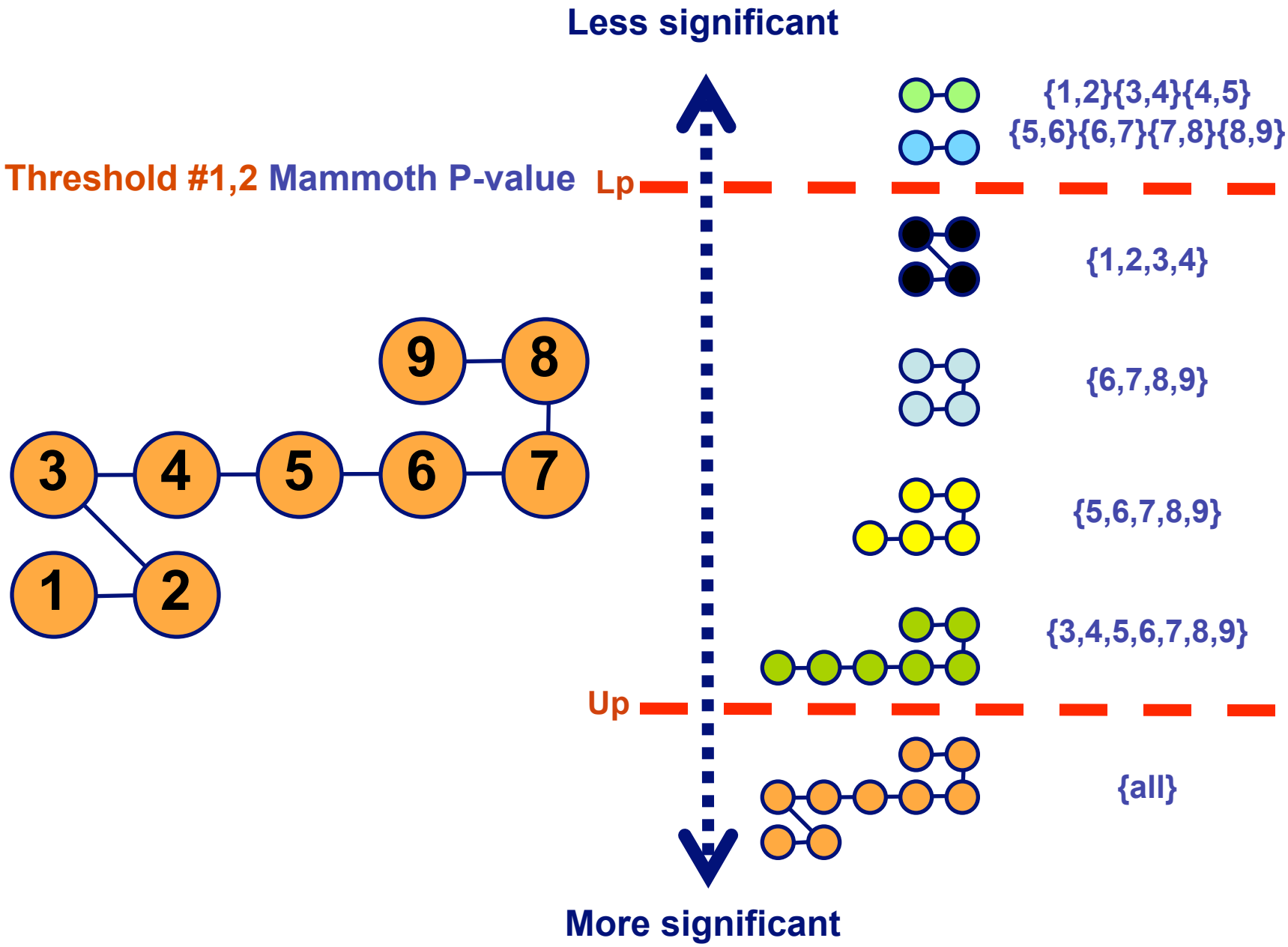


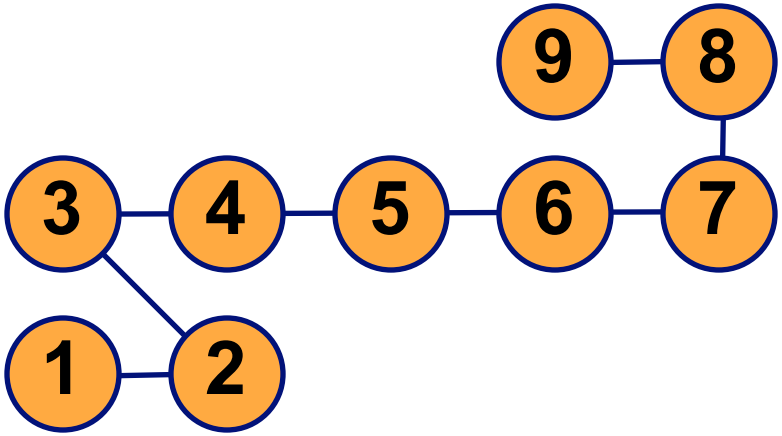
{all}


Less significant
{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}
{all}
9
8
3
4
5
6
7
1
2
More significant


Less significant
Threshold #1,2 Mammoth P-value
{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}
{all}
9
8
3
4
5
6
7
1
2
More significant

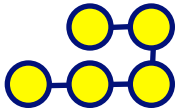

Less significant
Threshold #1,2 Mammoth P-value
Lp
{1,2}{3,4}{4,5}
{5,6}{6,7}{7,8}{8,9}
{1,2,3,4}
{6,7,8,9}
{5,6,7,8,9}
{3,4,5,6,7,8,9}
{all}
9
8
3
4
5
6
7
1
2
More significant


Less significant
Threshold #1,2 Mammoth P-value
Lp
{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}
Up
{all}
More significant
9
8
3
4
5
6
7
1
2

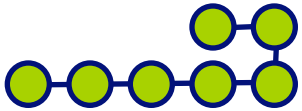


{1,2,3,4}

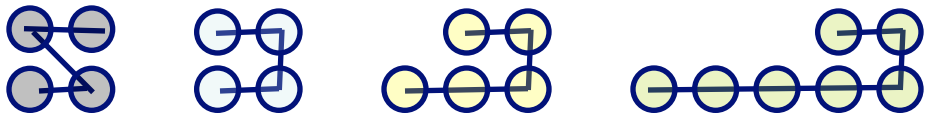
{6,7,8,9}

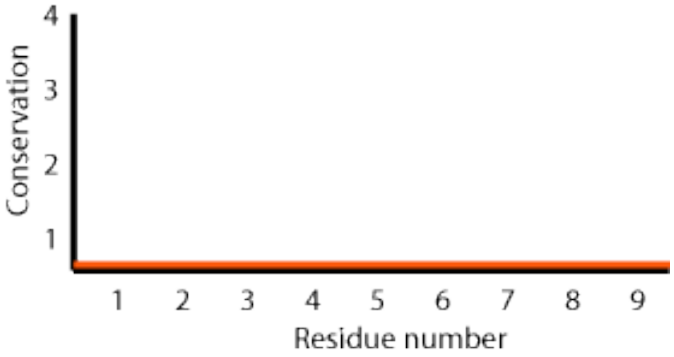


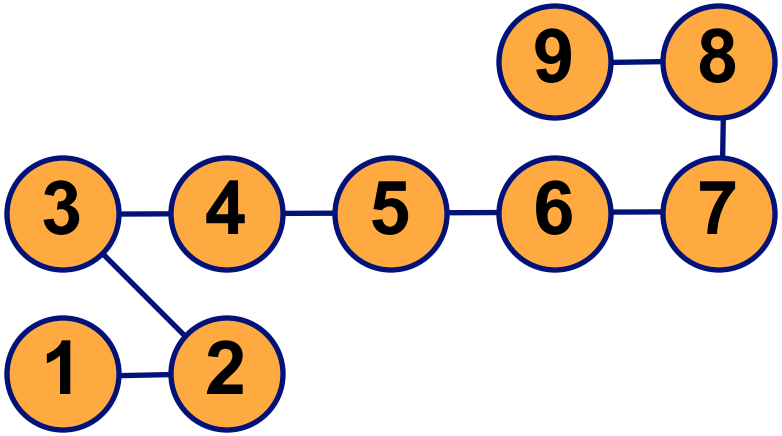
{5,6,7,8,9}



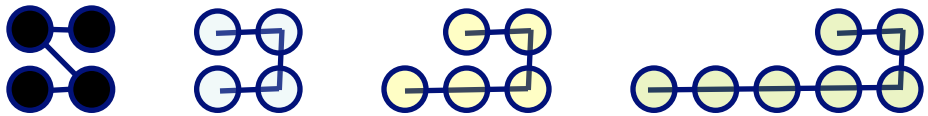
{3,4,5,6,7,8,9}

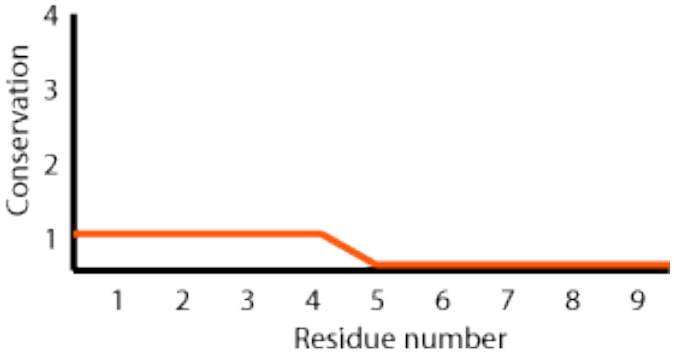


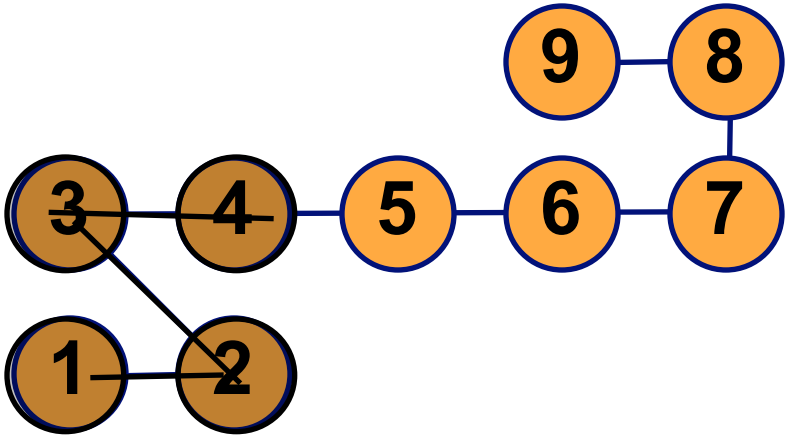




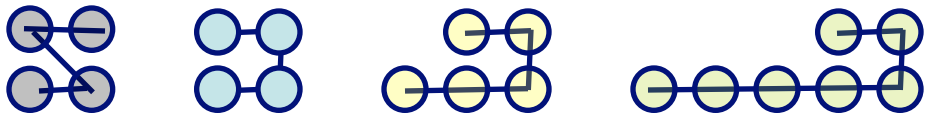
#	1	2	3	4	5	6	7	8	9
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0

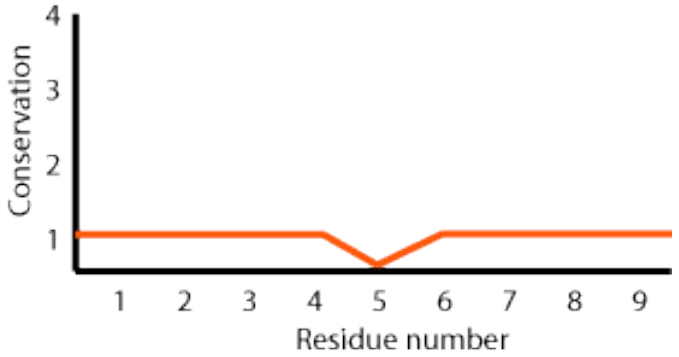


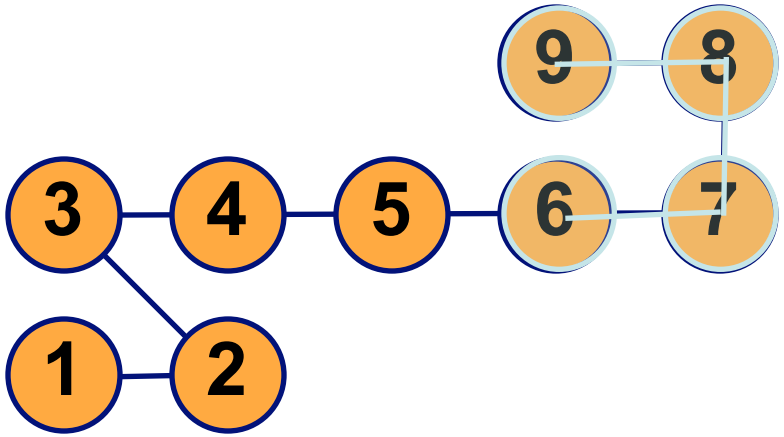




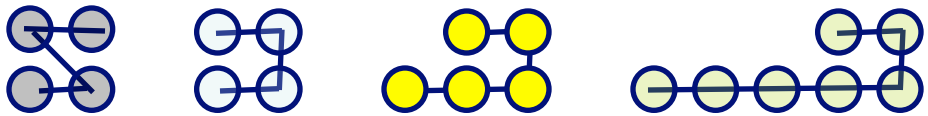
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0

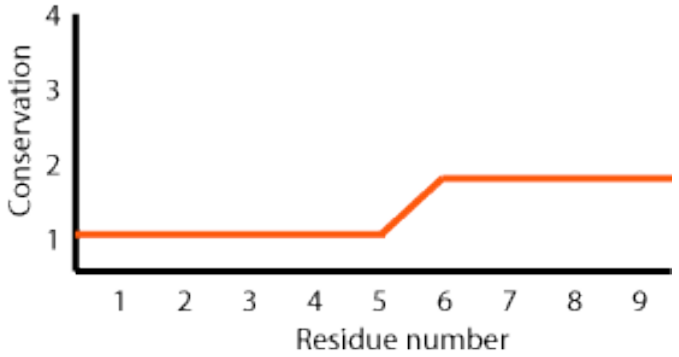


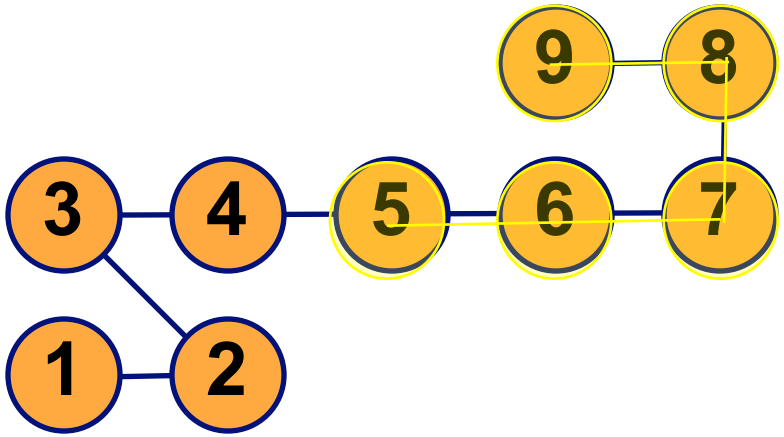




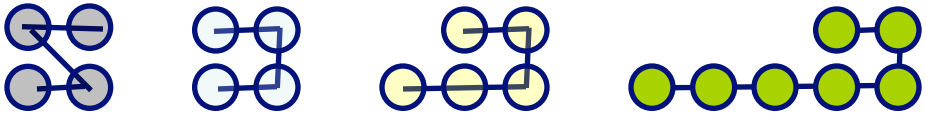
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	1	1	1
7	0	0	0	0	0	1	1	1	1
8	0	0	0	0	0	1	1	1	1
9	0	0	0	0	0	1	1	1	1

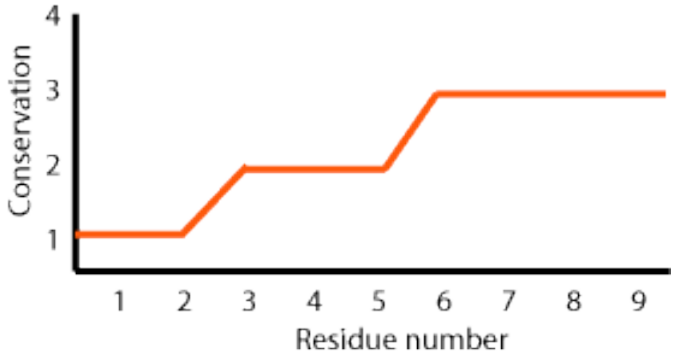


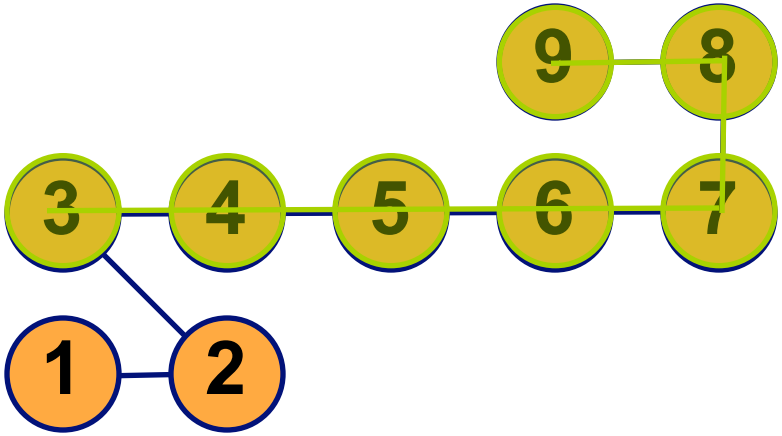




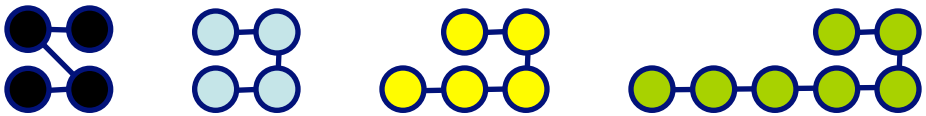
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	1	1	1	1	1
6	0	0	0	0	1	2	2	2	2
7	0	0	0	0	1	2	2	2	2
8	0	0	0	0	1	2	2	2	2
9	0	0	0	0	1	2	2	2	2

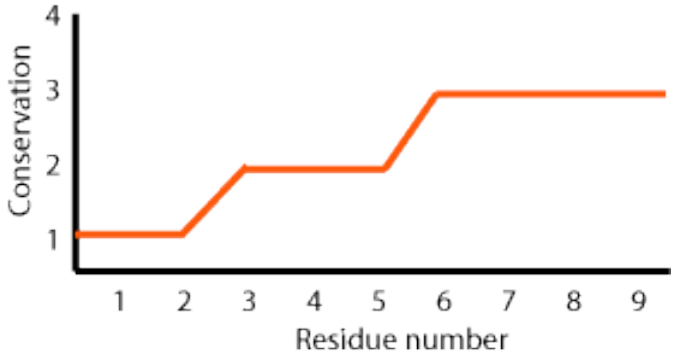


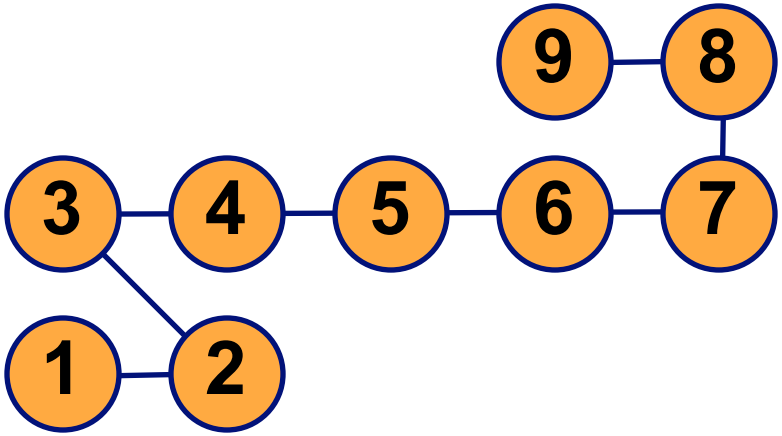




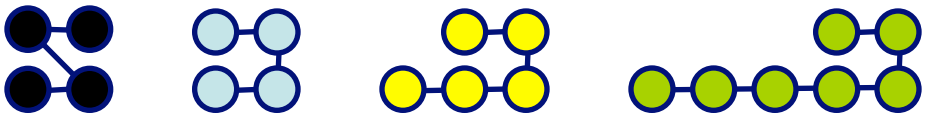
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	2	2	1	1	1	1	1
4	1	1	2	2	1	1	1	1	1
5	0	0	1	1	2	2	2	2	2
6	0	0	1	1	2	3	3	3	3
7	0	0	1	1	2	3	3	3	3
8	0	0	1	1	2	3	3	3	3
9	0	0	1	1	2	3	3	3	3

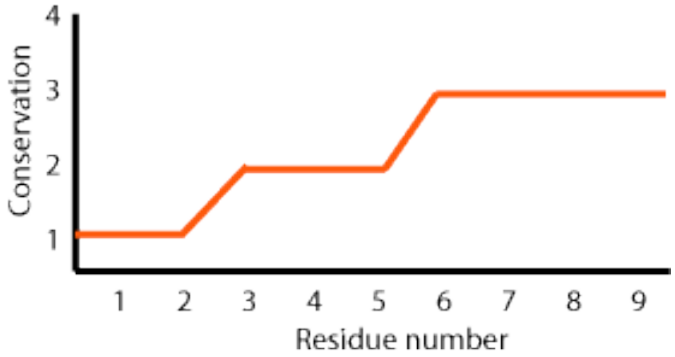


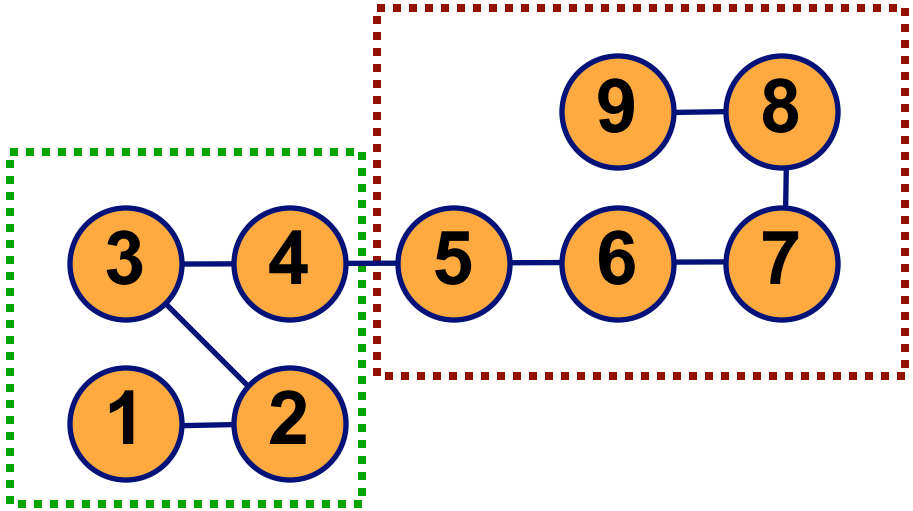




#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	2	2	1	1	1	1	1
4	1	1	2	2	1	1	1	1	1
5	0	0	1	1	2	2	2	2	2
6	0	0	1	1	2	3	3	3	3
7	0	0	1	1	2	3	3	3	3
8	0	0	1	1	2	3	3	3	3
9	0	0	1	1	2	3	3	3	3







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

Threshold #3 MCL Cluster level (-I)

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

Thresholds #1,2 → MAMMOTH P-Value (Lp, Up)

High P-values → fewer partitions

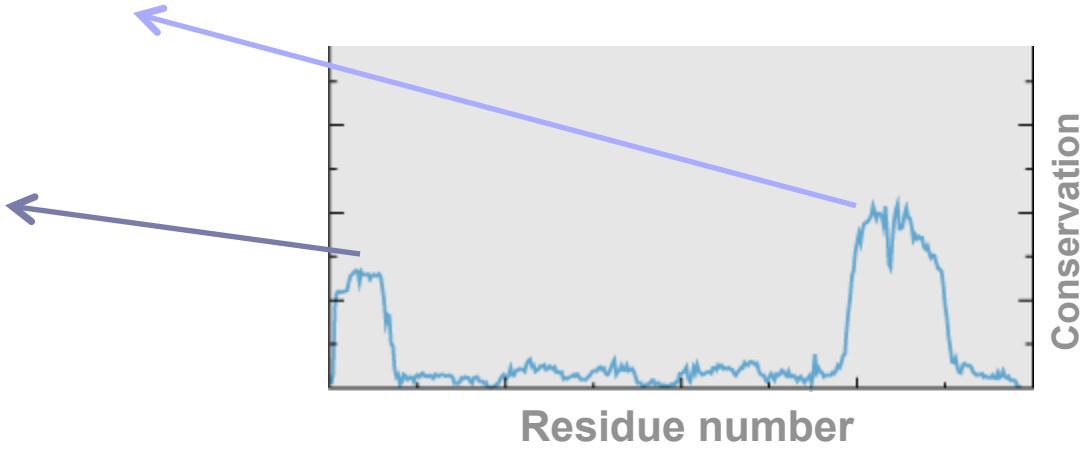
Threshold #3 → Cluster Level (-I)

Low –I cluster value → fewer partitions

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

1phh	290-329	**2.7Å**	**3.1**
1hadB	72-111		

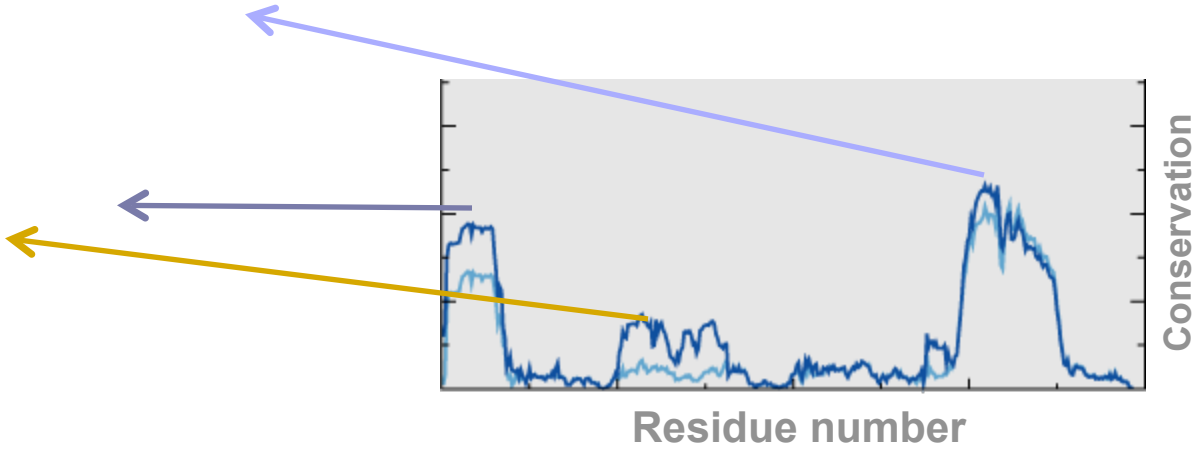
1phh	279-373	**3.9Å**	**4.7**
1bke	310-410		



1phh (Oxydoreductase from *Pseudomonas fluorescens*)

1phh	1-213	3.0Å	8.1
1qjdA	125-379		

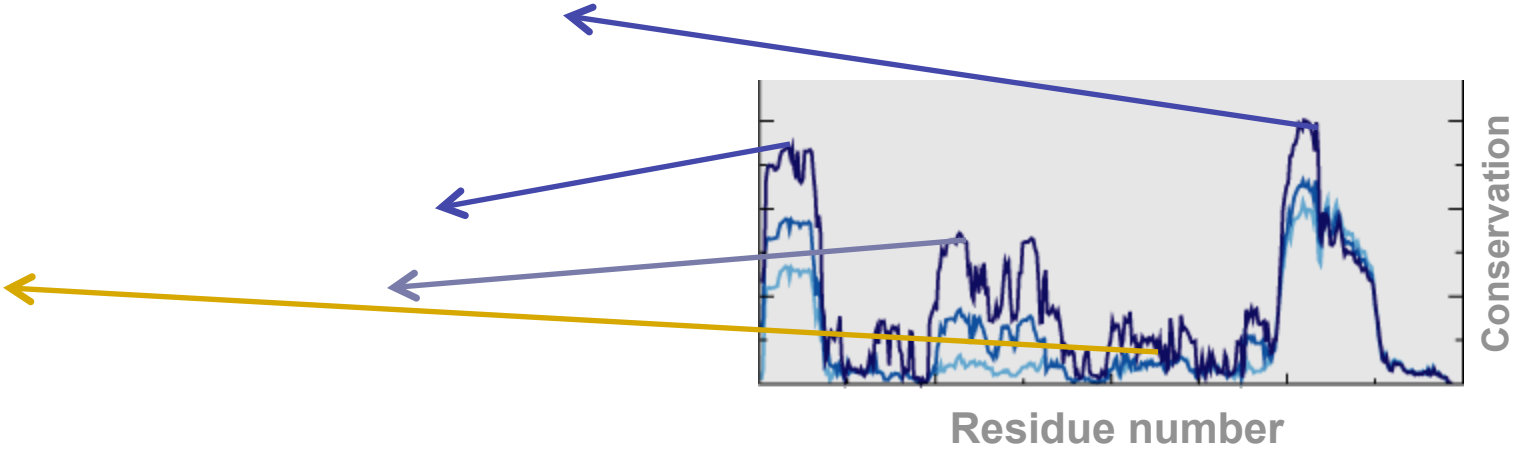
1phh	1-319	3.6Å	9.8
1gerA	3-327		



1phh (Oxydoreductase from *Pseudomonas fluorescens*)

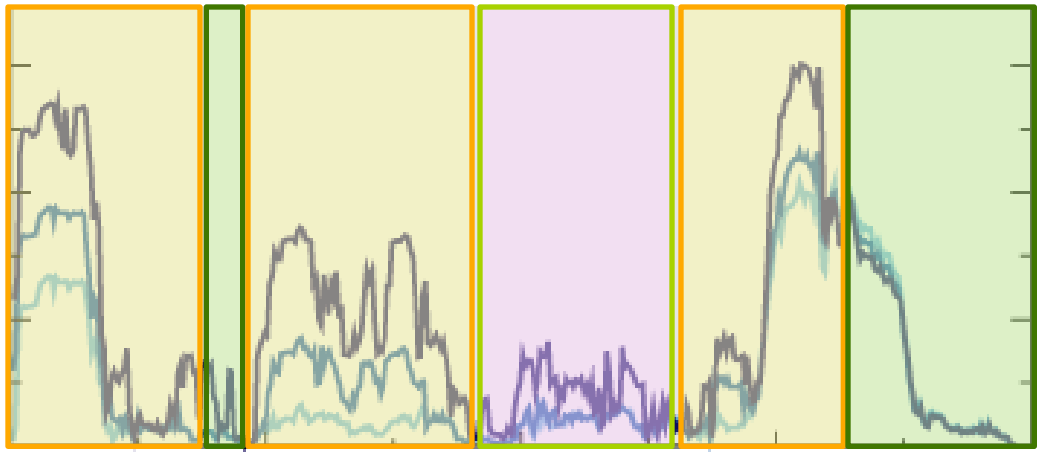
1phh	1-378	3.8Å	10.3
1feaC	2-464		

1phh	1-316	3.8Å	17.2
1l9dB	2-364		



1phh (Oxydoreductase from *Pseudomonas fluorescens*)

1phh (Oxydoreductase from *Pseudomonas fluorescens*)



Domain assignment from structure

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

<2Å && <30aa diff.

Divide randomly into two sets

Remove of incomplete or obsolete entries.

FINAL:

Training set → 242 chains

Testing set → 234 chains

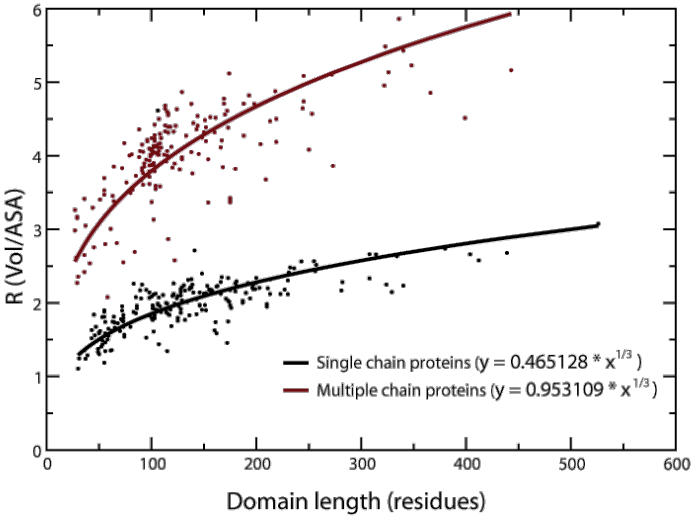
Thresholds #1,2 → MAMMOTH P-Value (Lp, Up)

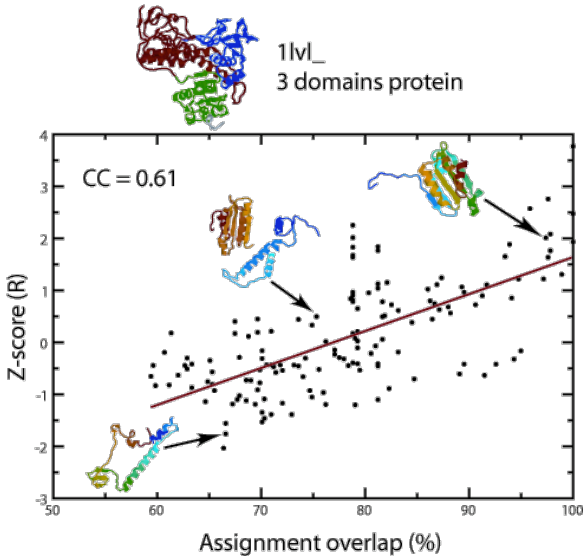
High P-values → fewer partitions

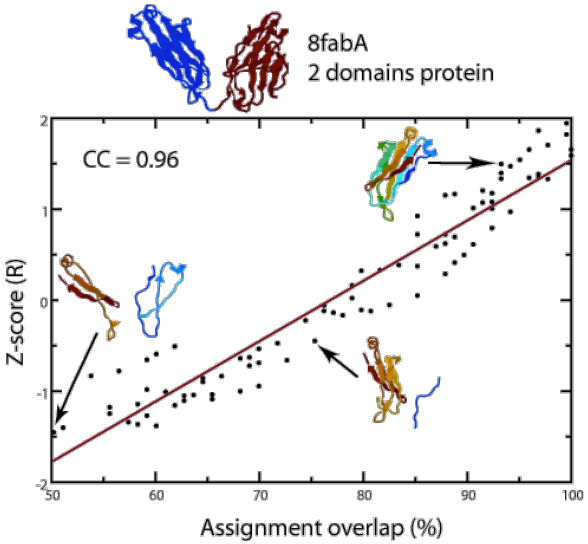
Threshold #3 → Cluster Level (-l)

Low –l cluster value → fewer partitions

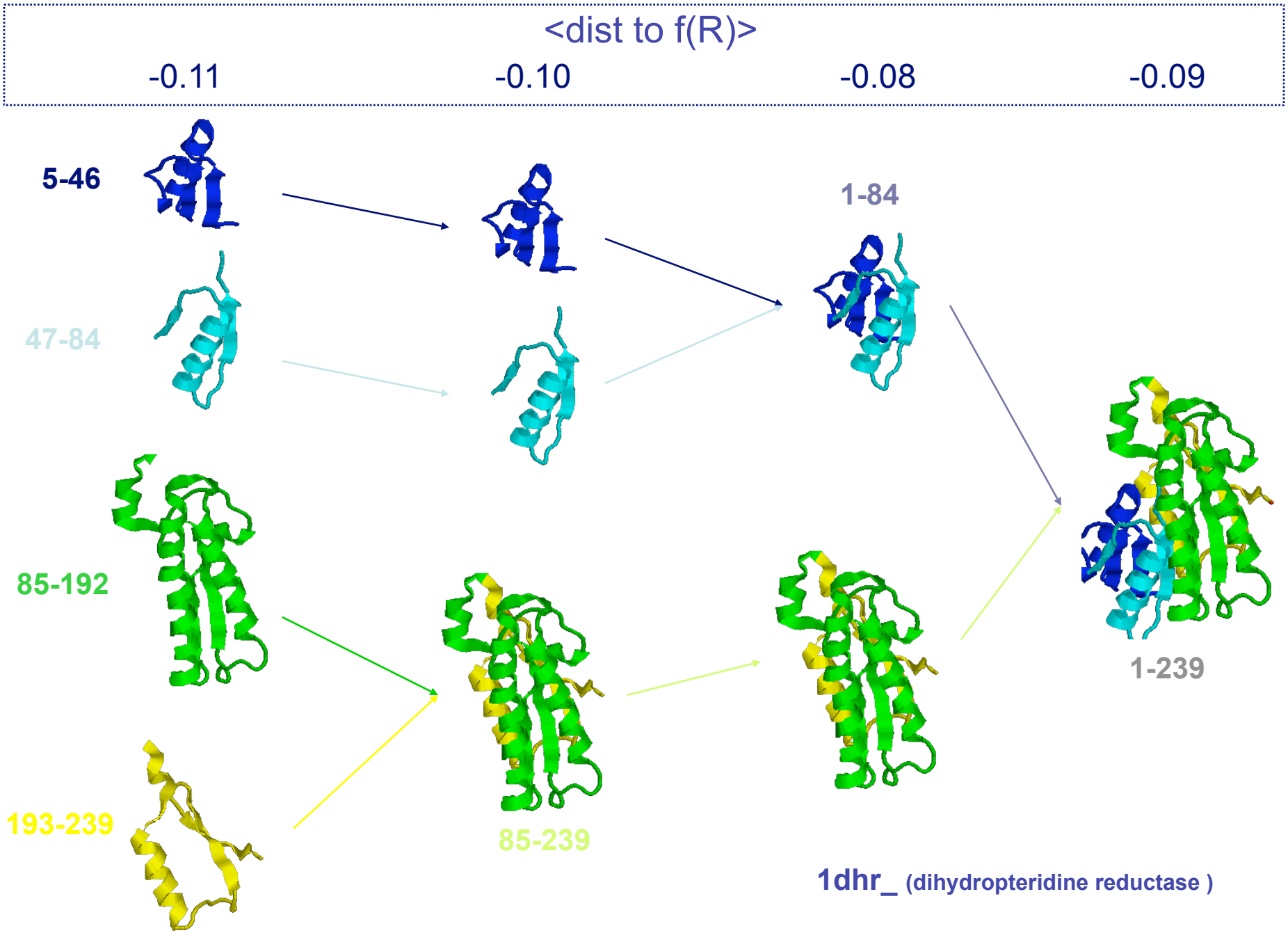
R = Volume/ASA



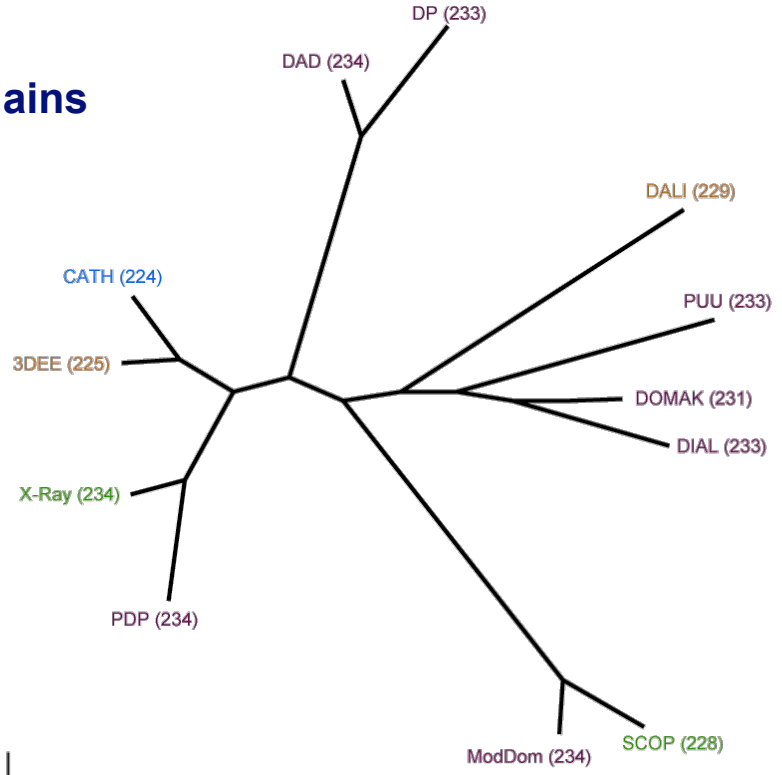


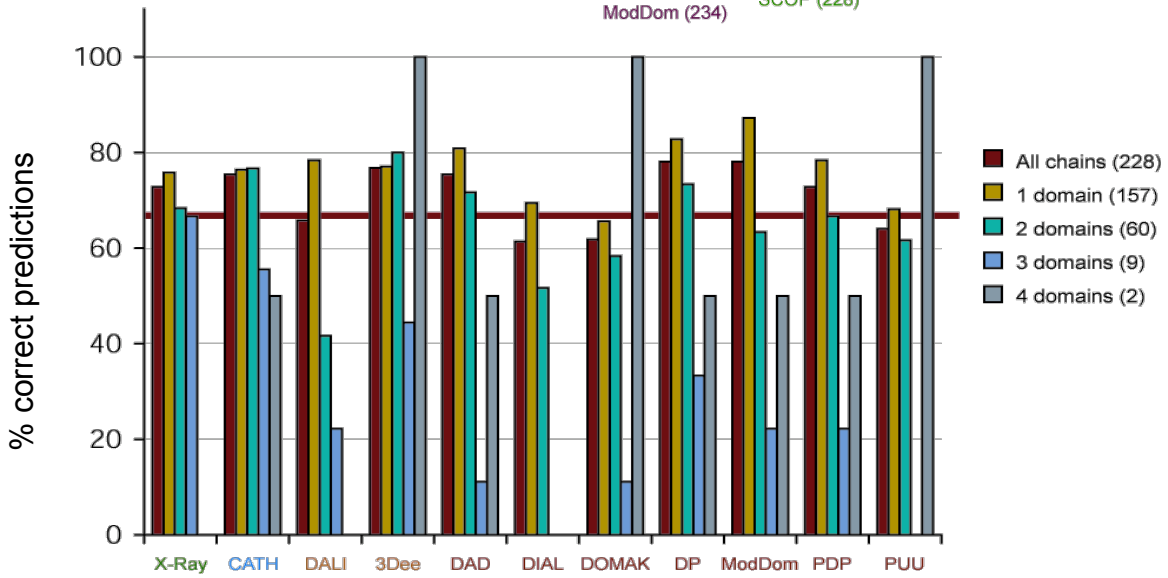


Domain → max(<dist f(R)>)


<dist to f(R)>
-0.11
-0.10
-0.08
-0.09
5-46
47-84
85-192
193-239
85-239
1-84
1-239
1dhr_ (dihydropteridine reductase)

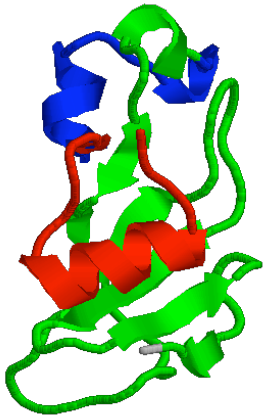
Non-redundant 234 chains

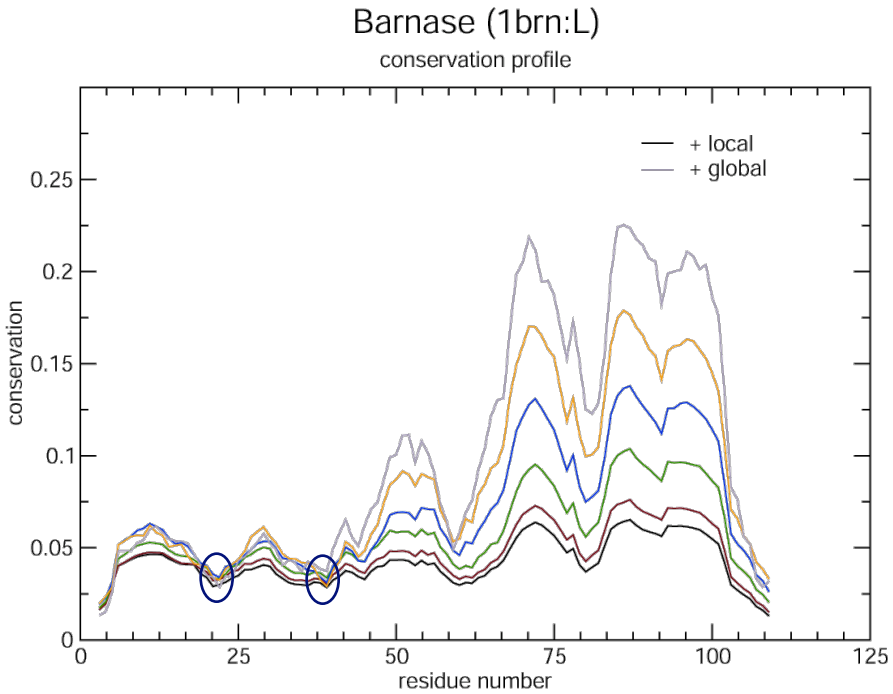




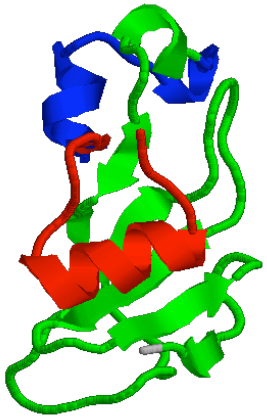
Fragments assignment from structure

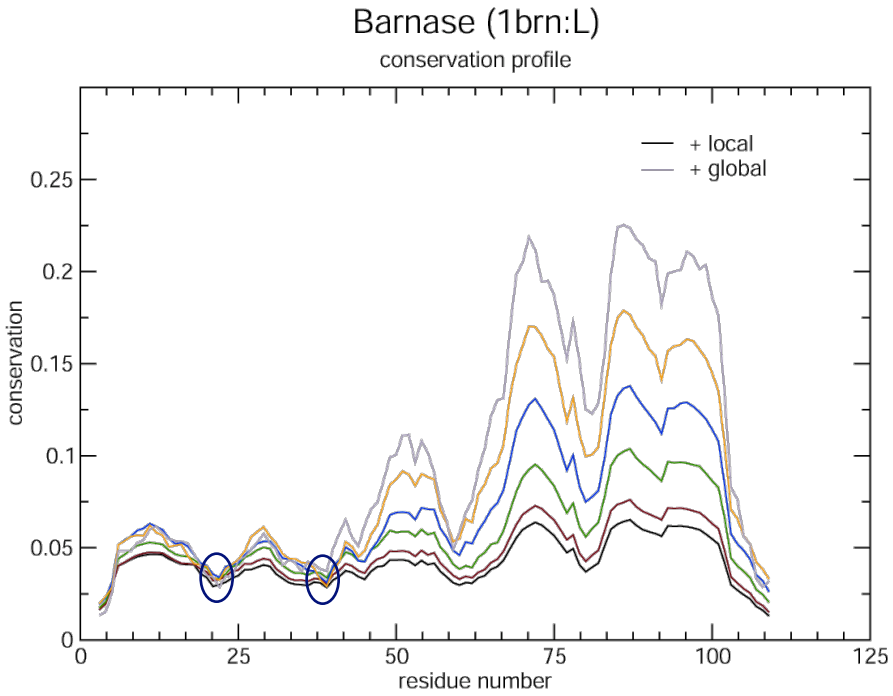
Barnase Domain-Swapping

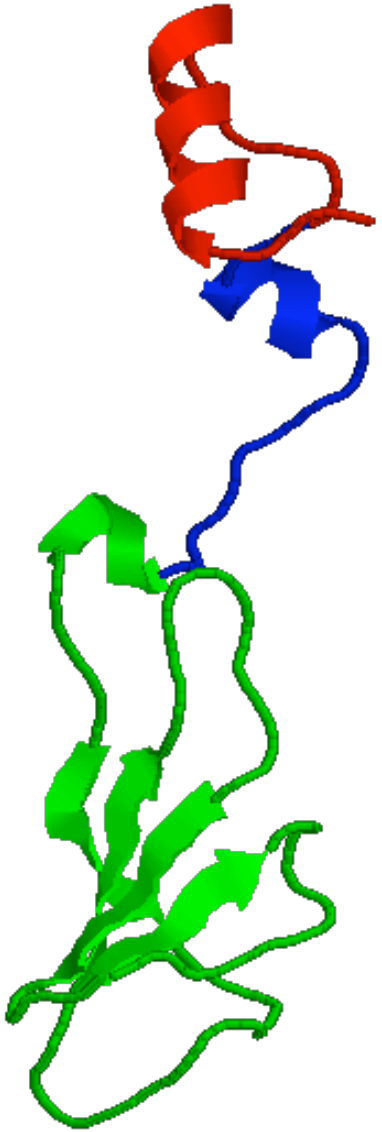




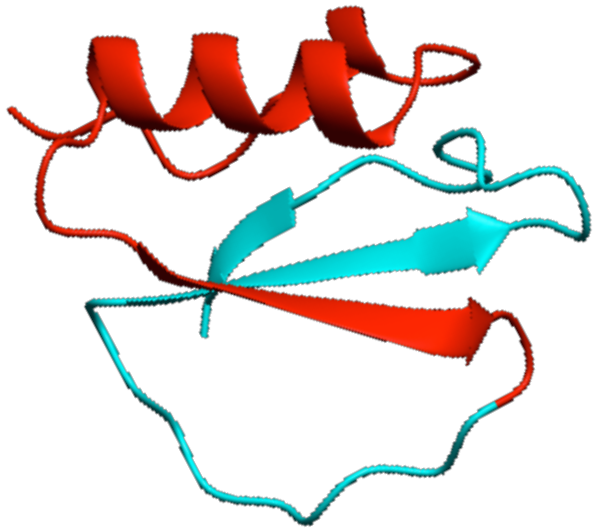
Barnase Domain-Swapping







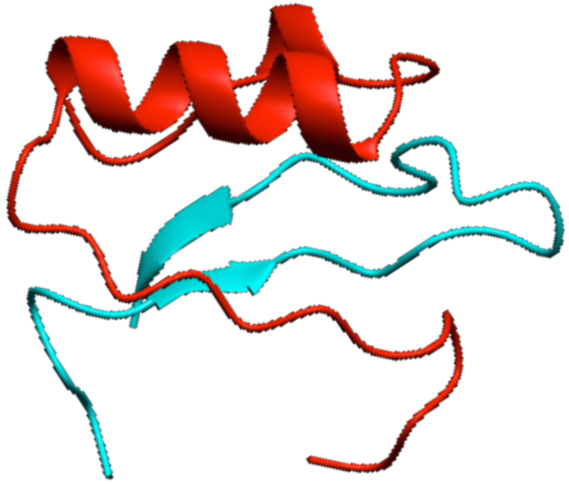
chymotrypsin inhibitor 2



1-37 | 38-64

chymotrypsin inhibitor 2

1-37 | 38-64



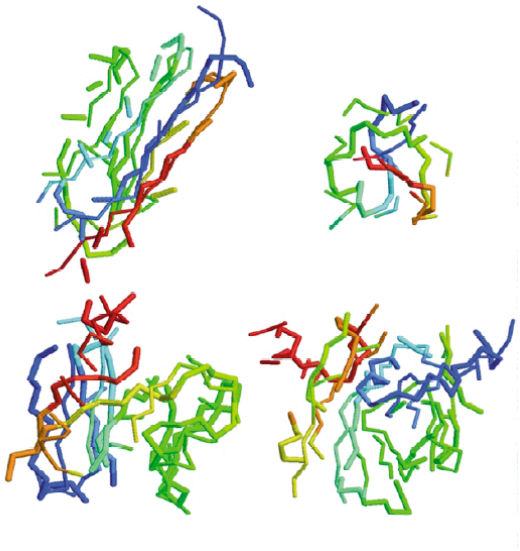
1-40 | 41-64

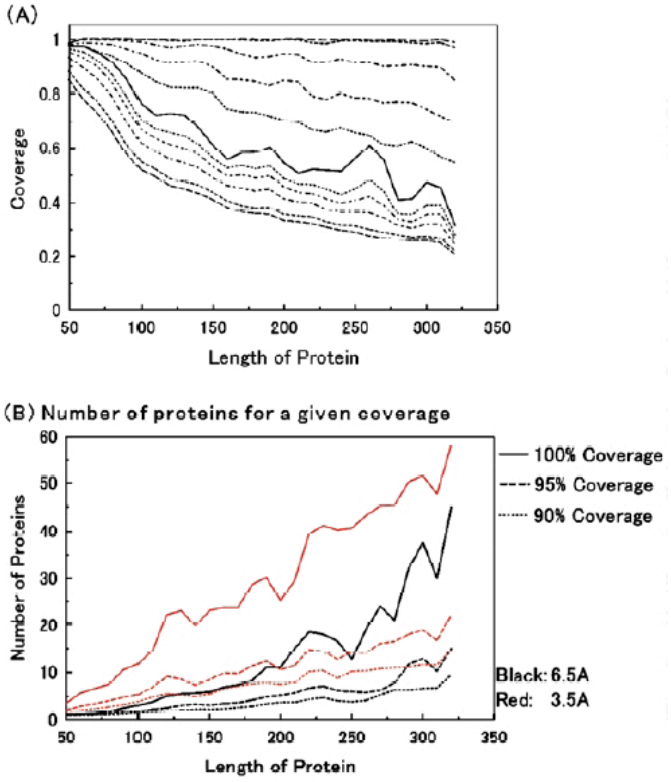
- Neira JL, Davis B, Ladurner AG, Buckle AM, Gay GP, Fersht AR. 1996. Towards the complete structural characterization of a protein folding pathway: the structures of the denatured, transition and native states for the association/folding of two complementary fragments of cleaved chymotrypsin inhibitor 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

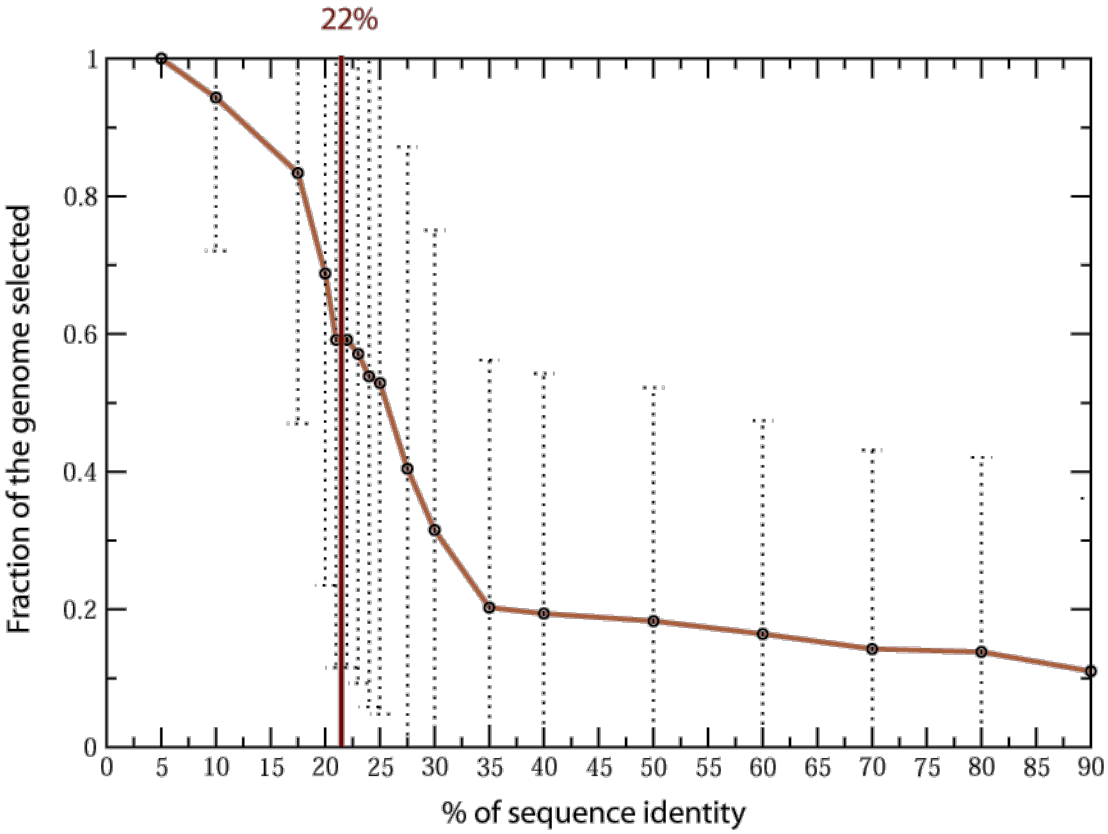
The PDB is a covering set of small protein structures.





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

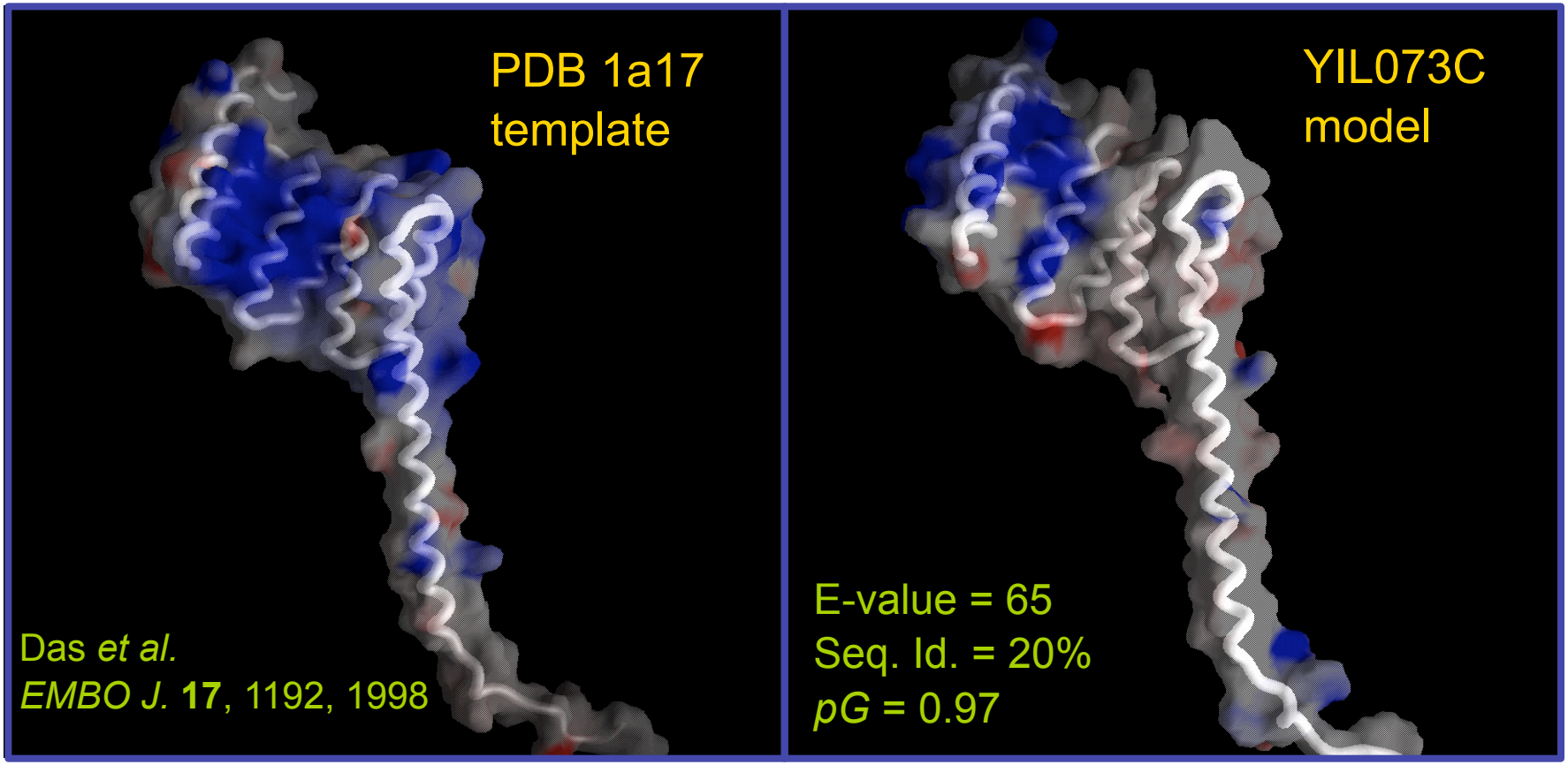
Sequence space .vs. Structure space

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)



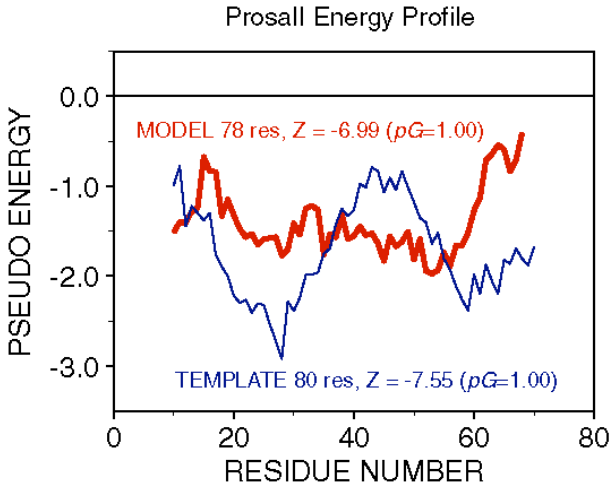
The tetratricopeptide repeat (TPR) is a degenerate 34 aa sequence identified in a variety of proteins, present in tandem arrays, mediates protein-protein interactions.

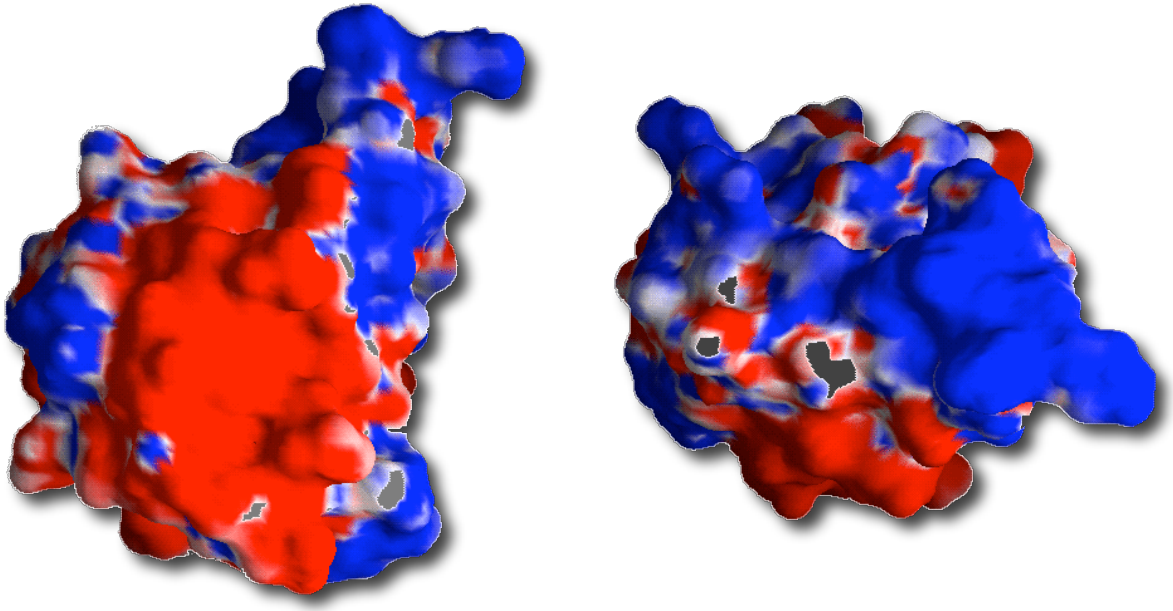
R. Sánchez, F. Melo, N. Mirkovic, A. Sali

GRAP: *Fold Assignment by PSI-BLAST + Model Evaluation*

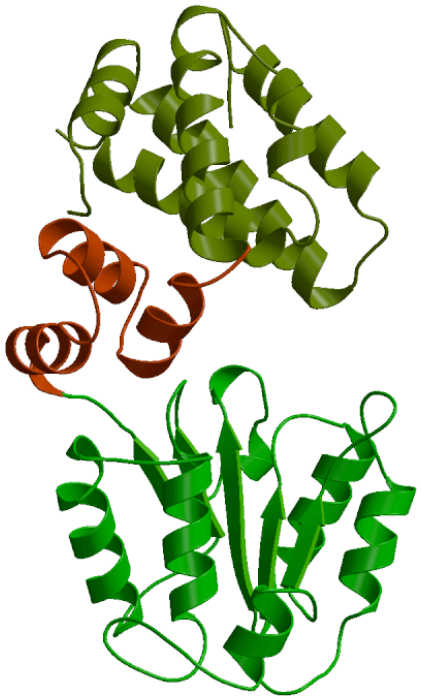
(significant e-value and good model score)

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	$6x10^{-14}$	1.00

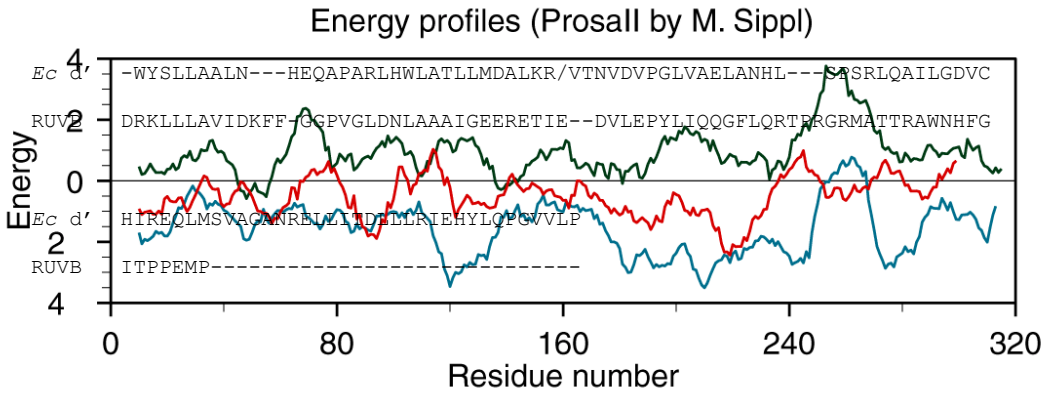




Does RuvB have the same fold as δ' of *E.coli* DNA polymerase III?



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

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

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

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



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

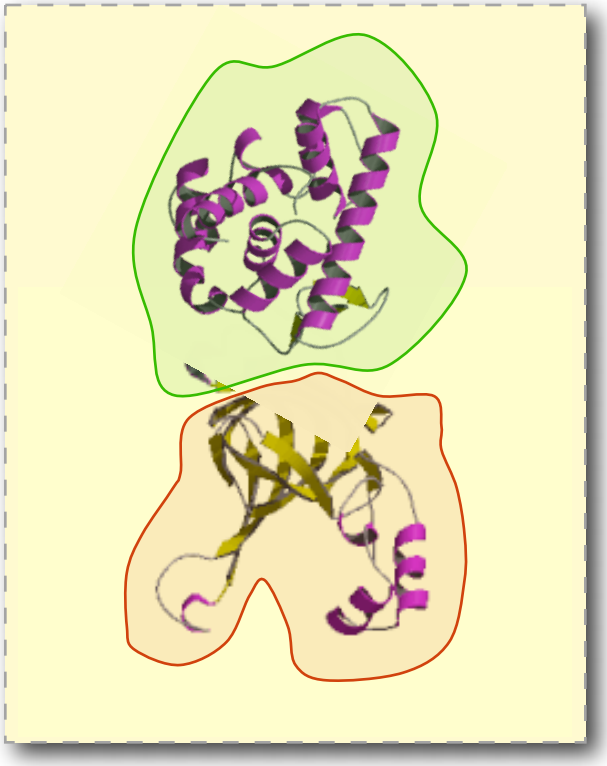
Department of Biopharmaceutical Sciences, UCSF

June 17th and 18th, 2004

Domains (?)

Domain boundaries from sequence

VERY DIFFICULT!!!!



MENFEIWVEKYRPRTLDEVVGQDEVIQRLKGYVERKNIPHLLFSGPPGTGKTATAIALARDLFGENWRDN
FIEMNASDERGIDVVRHKIKEFARTAPIGGAPFKIIFLDEADALTADAQAALRRTMEMYSKSCRFILSCN
YVSRIIEPIQSRCAVFRFKPVPKEAMKKRLLEICEKEGVKITEDGLEALIYISGGDFRKAINALQGAAAI
GEVVDADTIYQITATARPEEMTELIQTALKGNFMEARELLDRLMVEYGMSGEDIVAQLFREIISMPIKDS
LKVQLIDKLGEVDFRLTEGANERIQLDAYLAYLSTLAKK

Domain boundaries from sequence (SnapDragon)

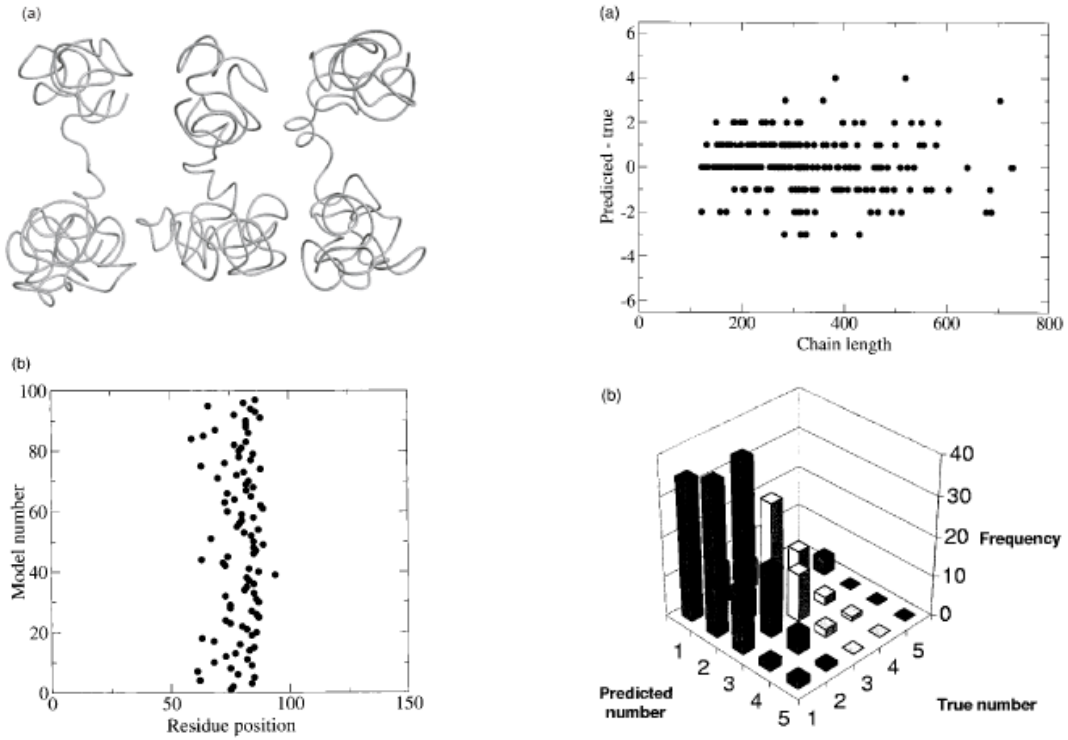
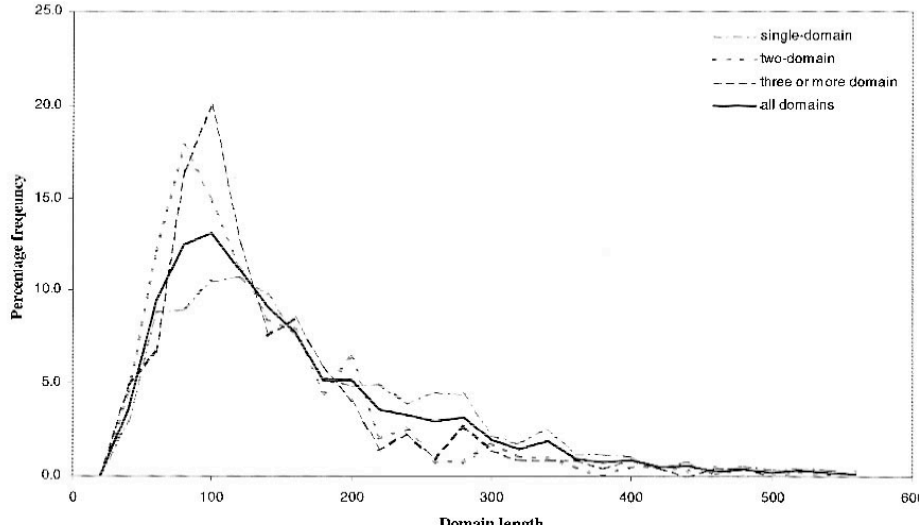


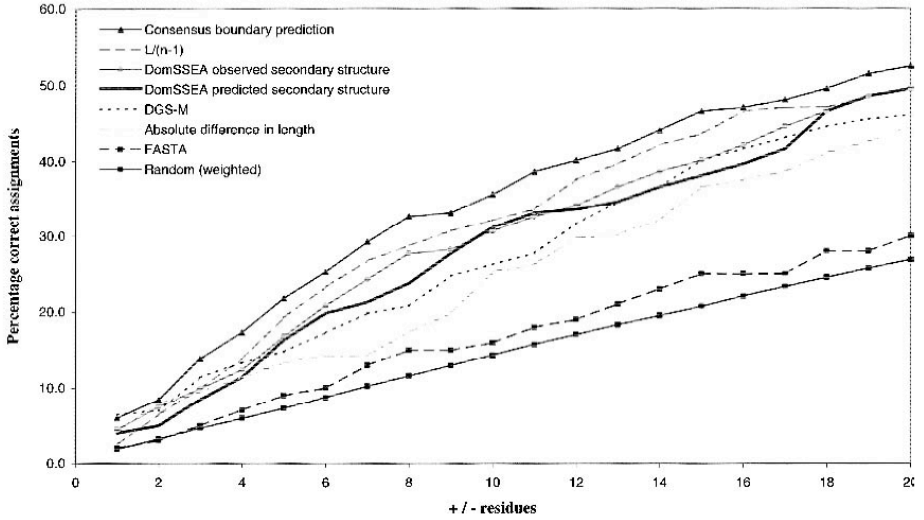
Table 2. Average accuracy percentages of linker prediction over 57 proteins

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

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

Domain boundaries from sequence (DomSSEA)



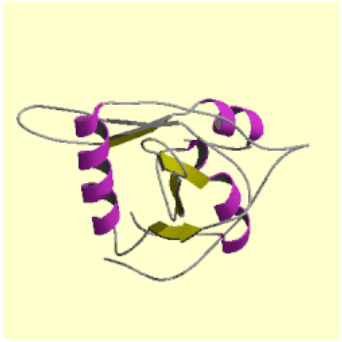


Dersden et al. (2003) Prot. Science 11 pp2014

Prediction of Secondary Structure (PSI-PRED)

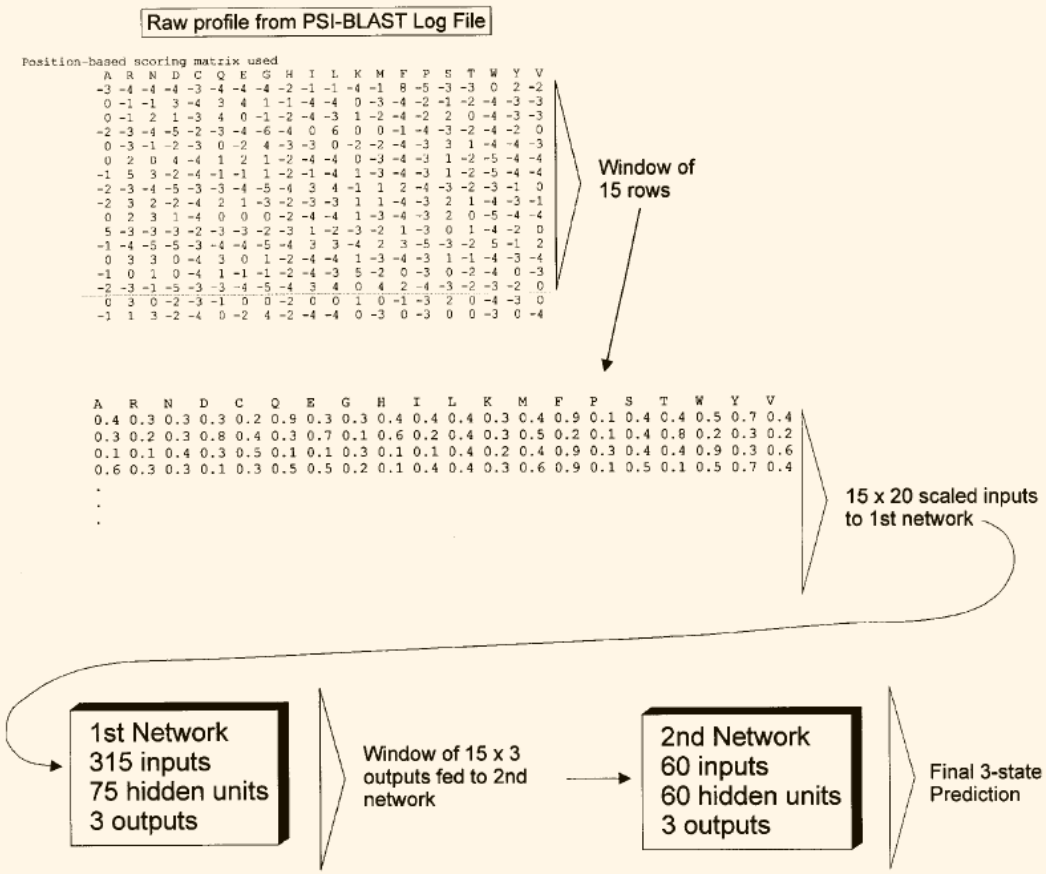
```
>gi42541361
MDIRSVSSLRGLLCLPPSWPRR
```

- Neural Network



- ✓ Very simple idea
- ✓ Simple scoring

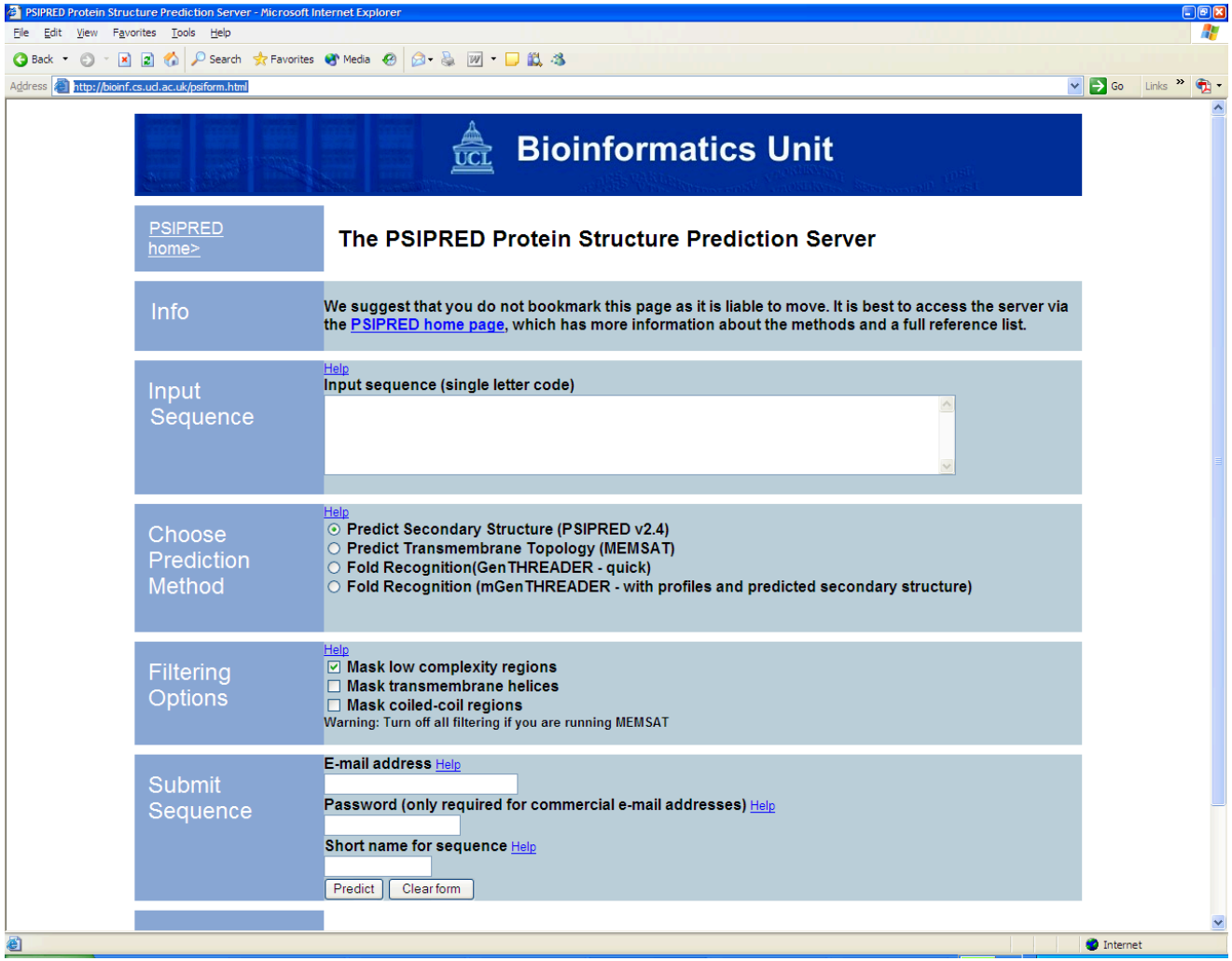
Obscure optimizer



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

Prediction of Secondary Structure (PSI-PRED)

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



Why the alignment is so important?

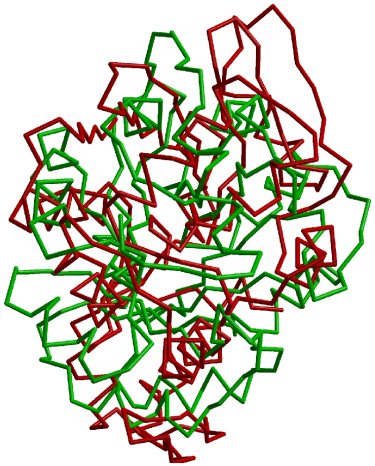
Typical errors in comparative models

MODEL

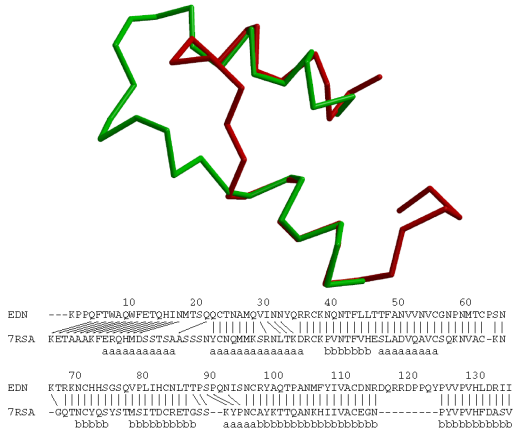
X-RAY

TEMPLATE

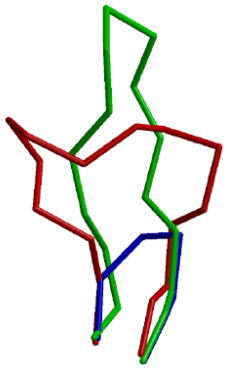
Incorrect template



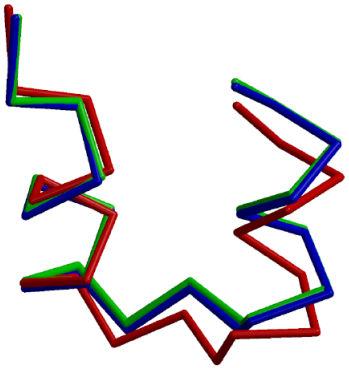
Misalignment



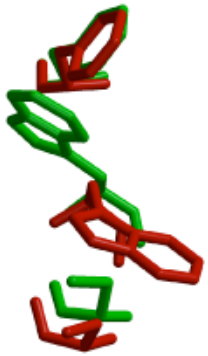
Region without a template



Distortion/shifts in aligned regions

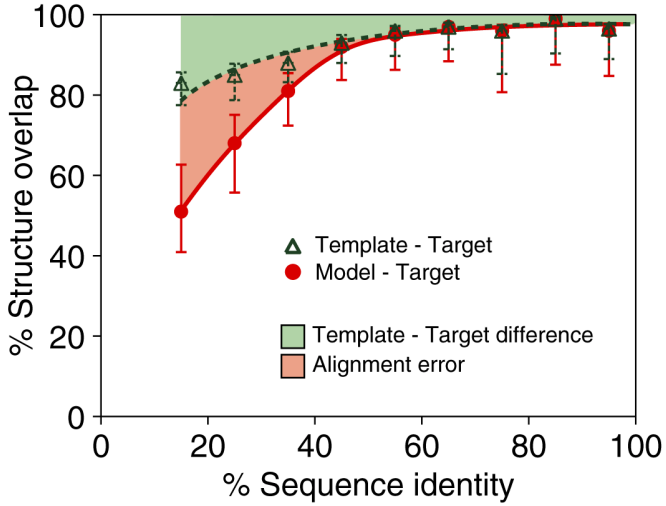


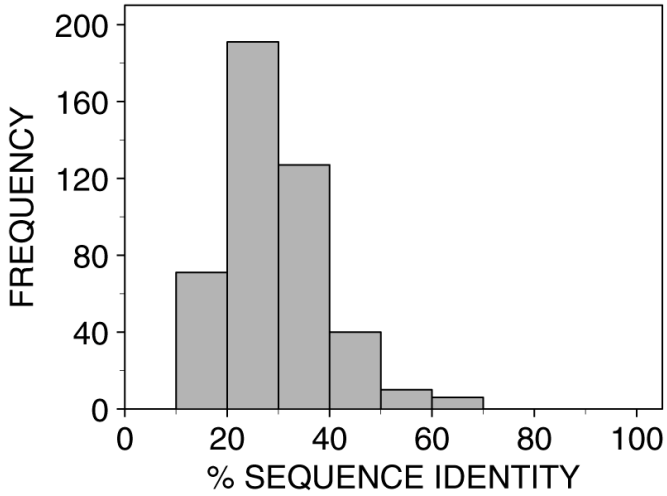
Sidechain packing



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

Alignment errors are frequent and large





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

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 asses the fit of a sequence to a given fold
  - Scoring functions deried 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.

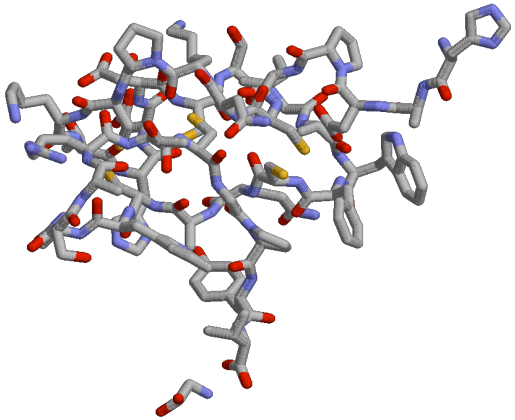
Finkelstein et al. (1995) Proteins 23, pp142

Representation

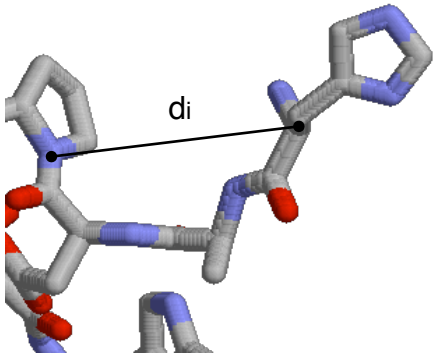
Sequence/Structures

```
>gi42541361
MDIRSVSSLRGLLCLPPSWPRR
```

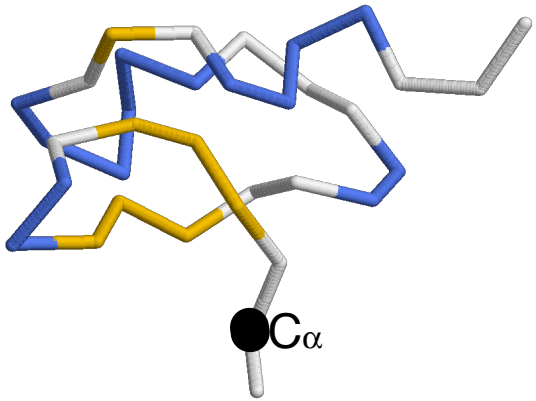
Primary sequence



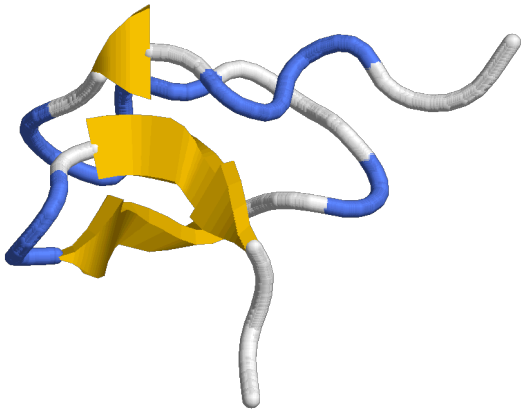
All atoms and coordinates



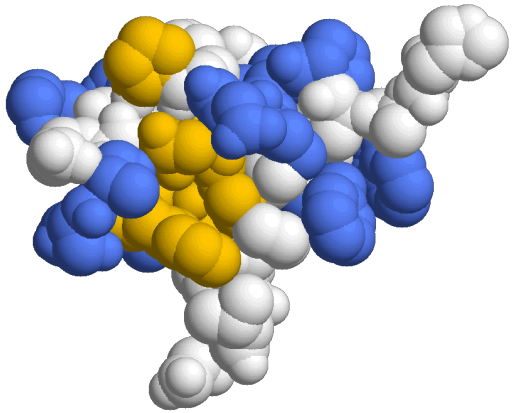
Distance space



Reduced atoms representation



Secondary Structure

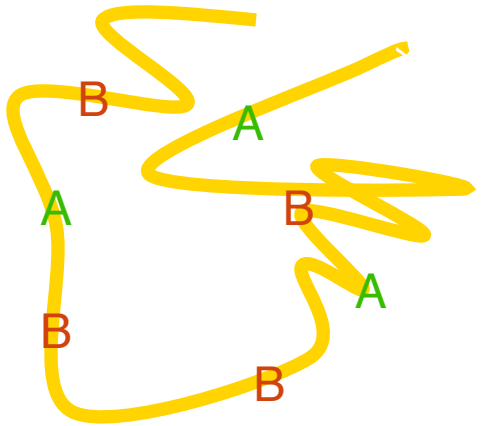


Accessible surface

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]:

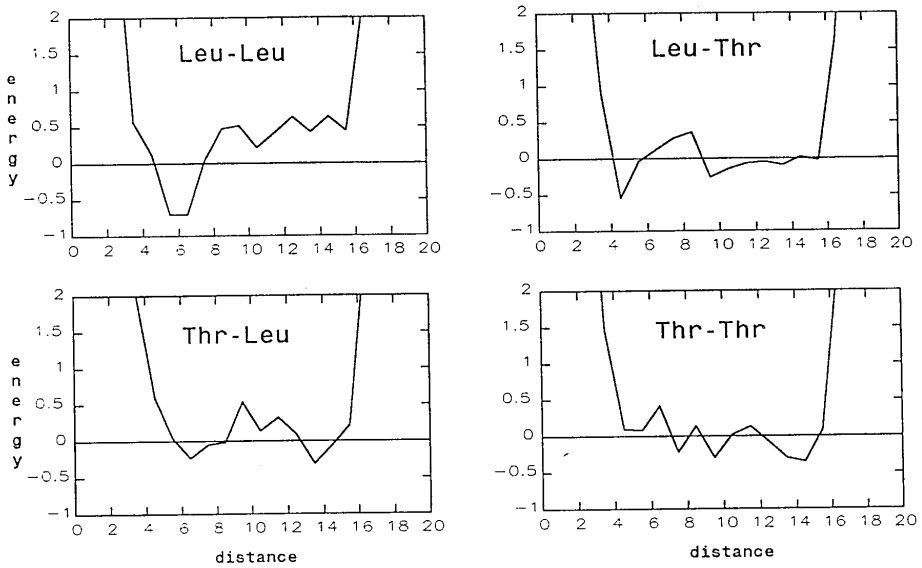
$$\Delta E = -kT \ln\left(\frac{N_1}{N_2}\right) \qquad (1)$$

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.

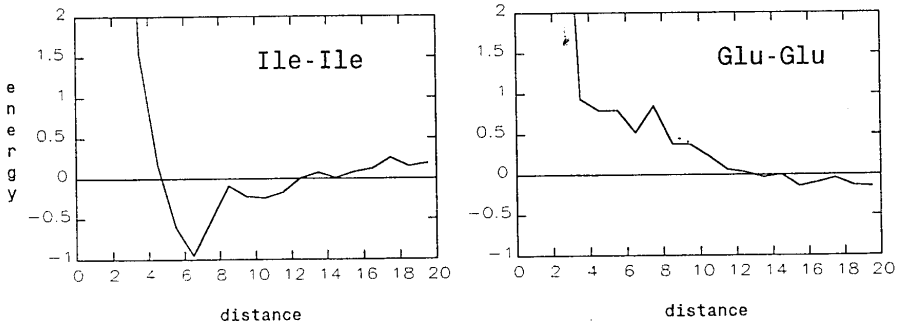
Tanaka and Sheraga (1975) PNAS, 72 pp3802
A. Godzik, (1996) Structure 15 pp363

Statistical Potential… Distance Potentials

Long range free energy



Short range free energy



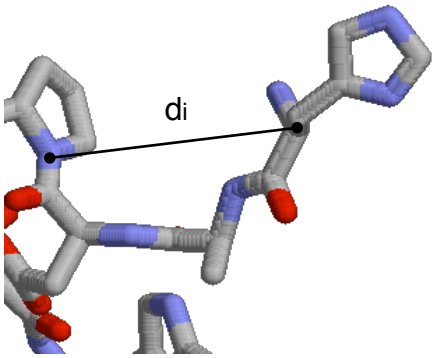
Sippl (1993). JCAM 7 pp473

Scoring

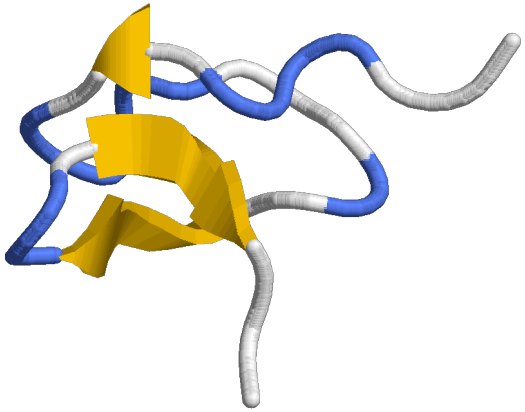
Raw scores of an alignment

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

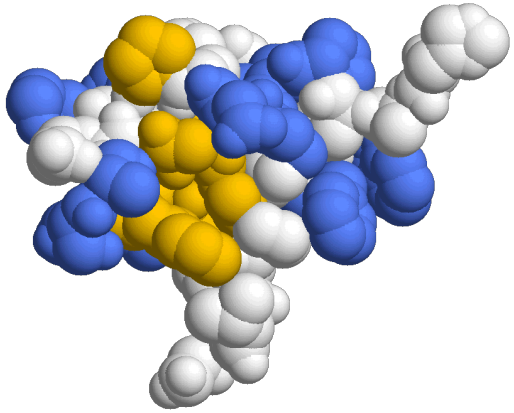
Aminoacid substitutions



Distance space



Secondary Structure (H,B,C)

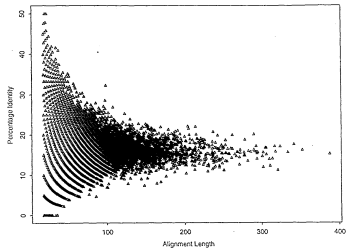


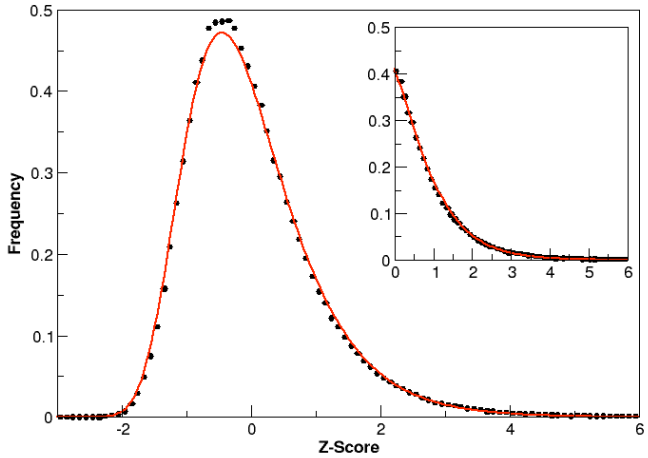
Accessible surface (B,A [%])

Significance of an alignment (score)

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

Empirical





Sometimes approximated by Z-score (normal distribution).

Analytic

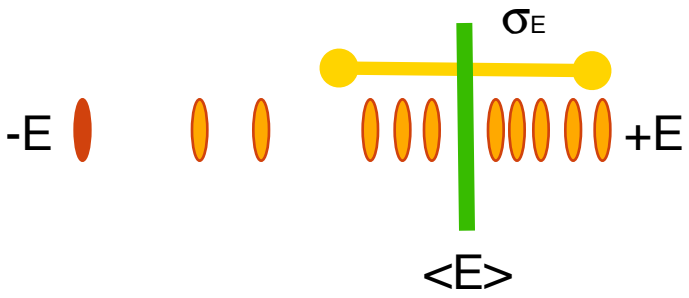
$$P(s) = e^{-\lambda (s-\mu)}$$

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

Karlin and Altschul, 1990 PNAS 87, pp2264

Significance of an alignment (score)

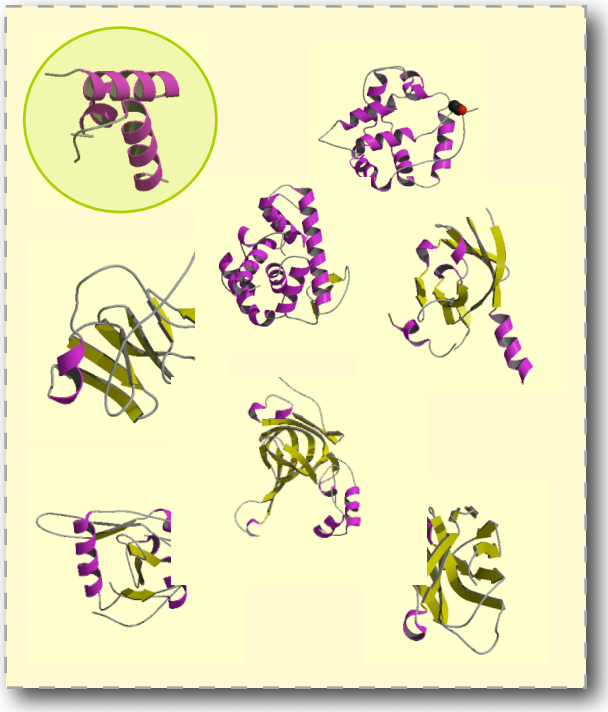
Energy Z-score the model with respect the energy of random models (or rest of decoys).



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

Significance of an alignment (background)

Structural space



Sequence space

MKLLIVLTCISLCSCICTVVQRCASNKPHVLEDPCKVQH

HLSVNQCVLLPQCCPKSCKICTHLISIEVVLTCRAVDKM

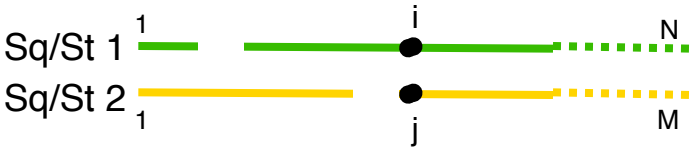
MHVNCVEQCSLQDCIKIAPRVLKTCILCVLKPCLTSVSH

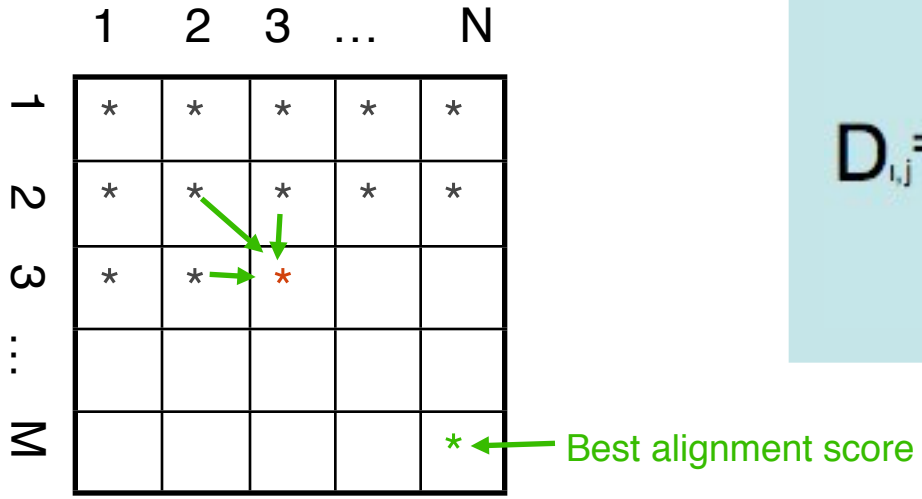
VHLVQPTSCCCKKNCICHVEIRSLDILTKSVQLACLVPM

⋮

MQCCRVQKICDLLAVELCKLHISTPSCKILCVVTSVPHN

Global dynamic programming alignment



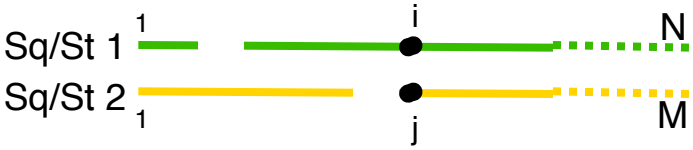


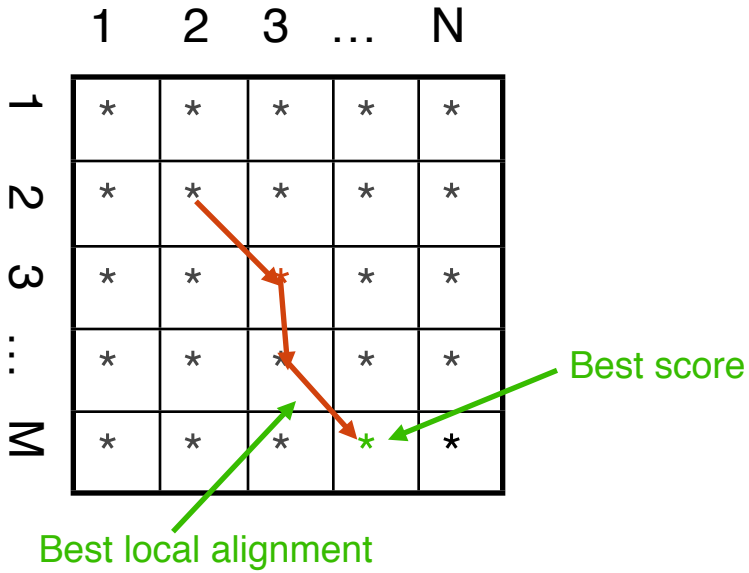
$$D_{i,j} = \min \begin{cases} D_{i,j-1} + \text{Score}_{(\Delta, rj)} \\ D_{i-1,j-1} + \text{Score}_{(ri, rj)} \\ D_{i-1,j} + \text{Score}_{(ri, \Delta)} \end{cases}$$

Backtracking to get the best alignment

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

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

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

Applications of PMFs

Model assessment.

Ab initio folding simulations.

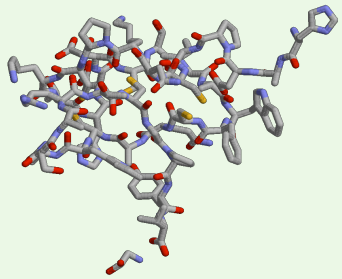
Sequence-structure matching (threading).

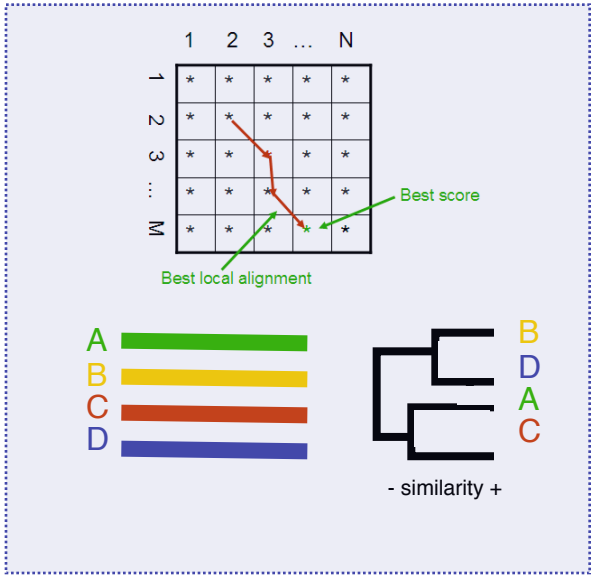
Comparative protein structure modeling (loops, sidechains, …).

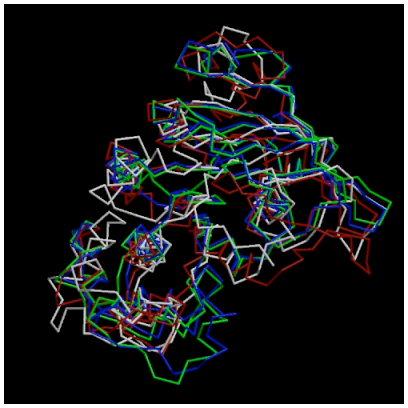
Secondary structure prediction, etc.

Finkelstein et al. (1995) Proteins 23, pp142

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

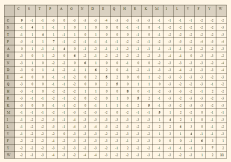


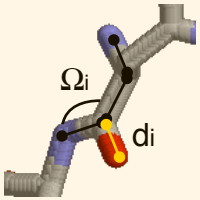


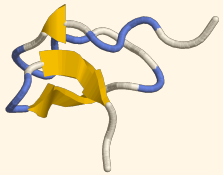


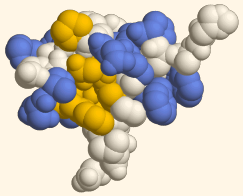
- ✓ Uses all available structural information
- ✓ Provides the optimal alignment

Computationally expensive









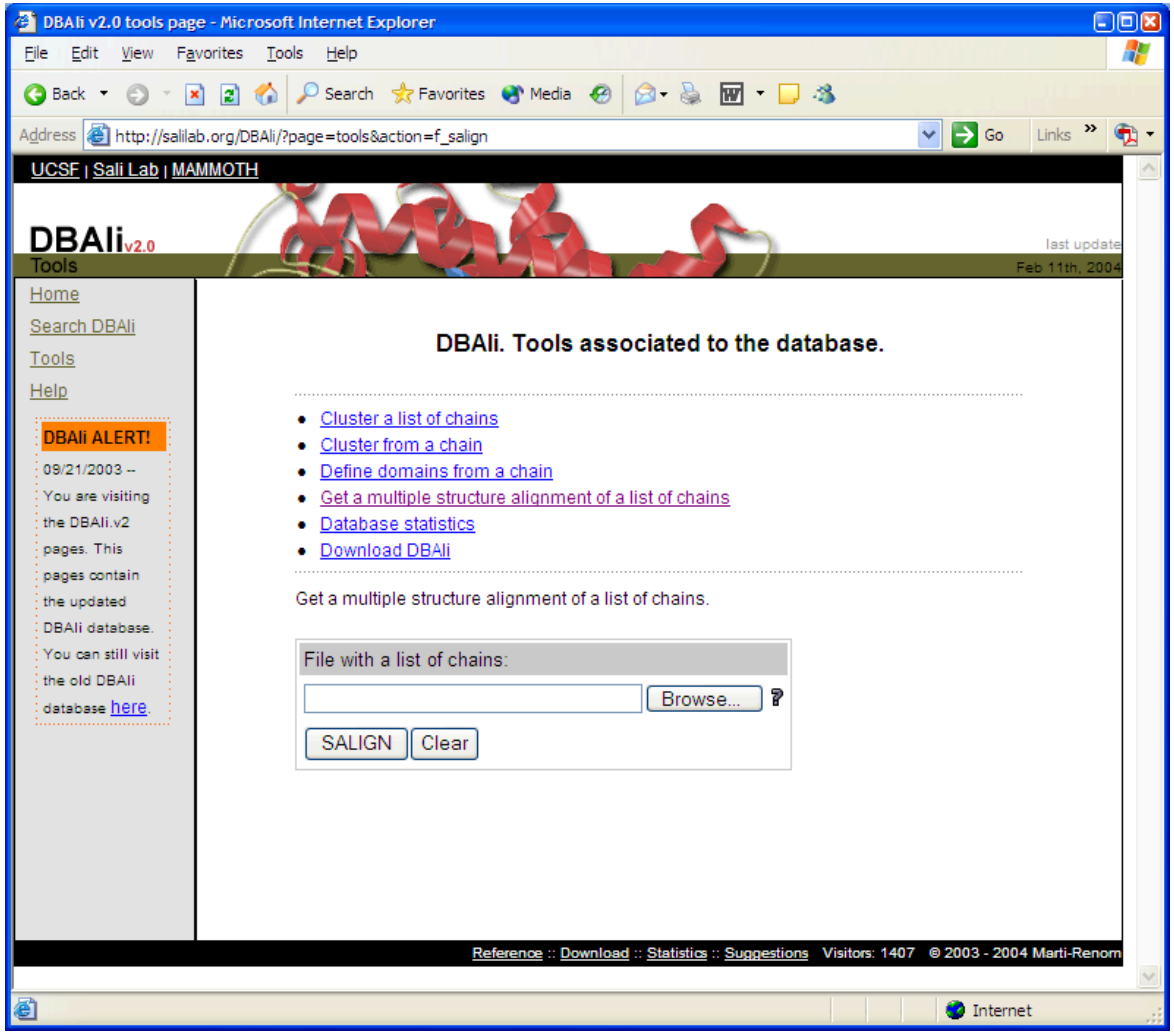
$$RMSD(x,y) = \sqrt{\left(\frac{1}{N}\right)\sum_{i=1}^{N}\left(\left\|\mathbf{x}(i)-\mathbf{y}(i)\right\|^{2}\right)}$$

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

Madhusudhan et al. in preparation

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

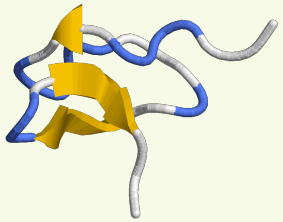
http://www.salilab.org/dbali/

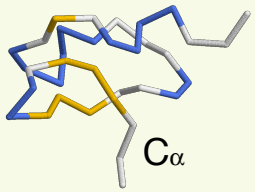


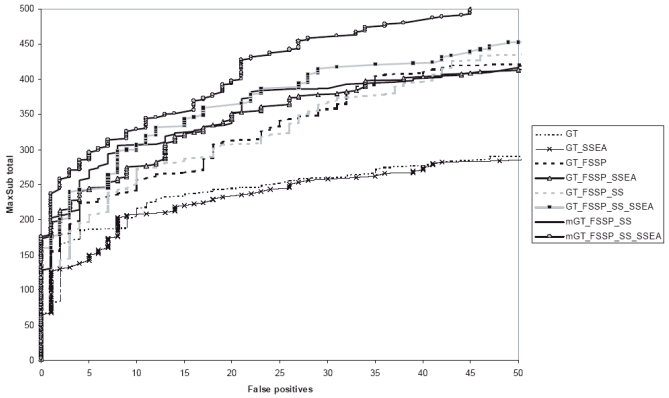
Threading (mGenThreader)

```
>gi42541361
MDIRSVSSLRGLLCLPPSWPRR
```

- Neural Network

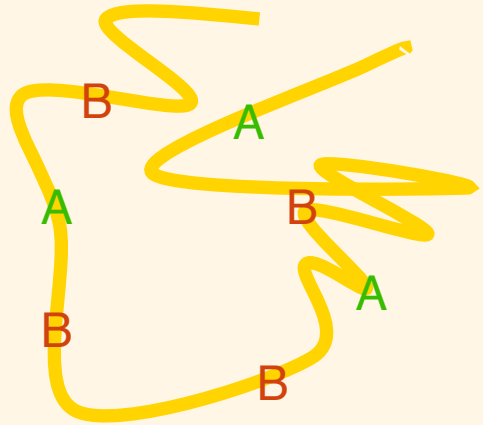


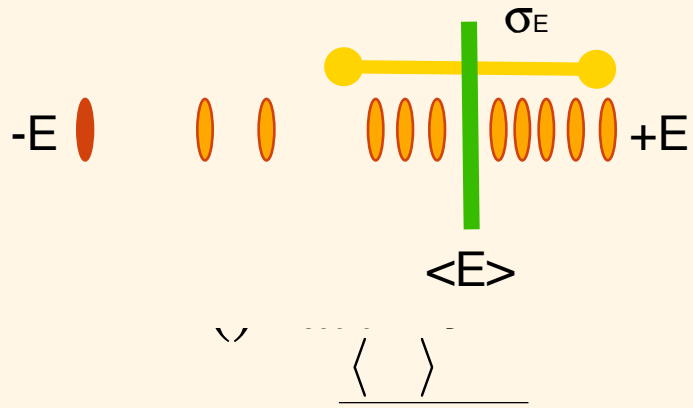




✓ Good row and significance scoring

Obscure optimizer

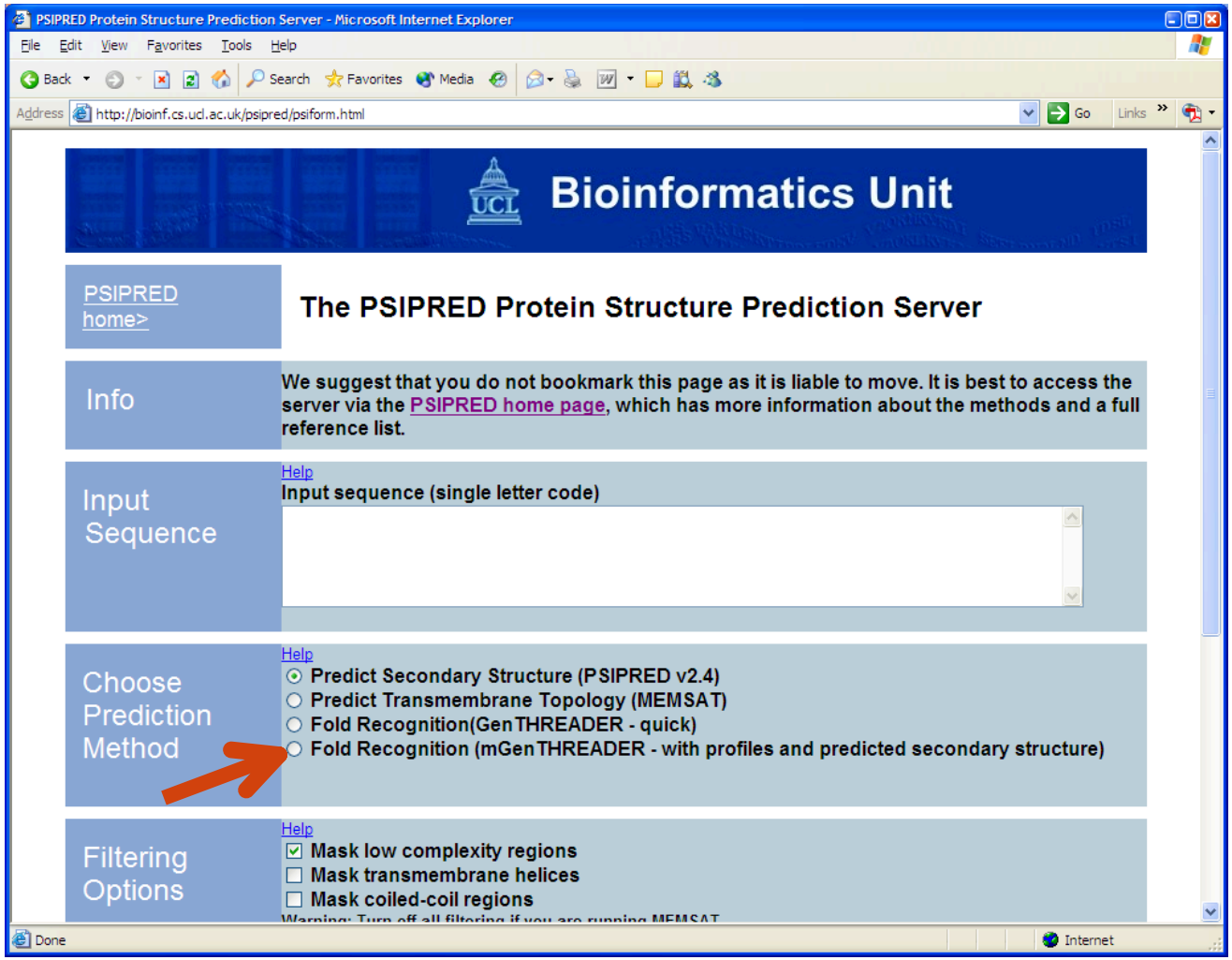




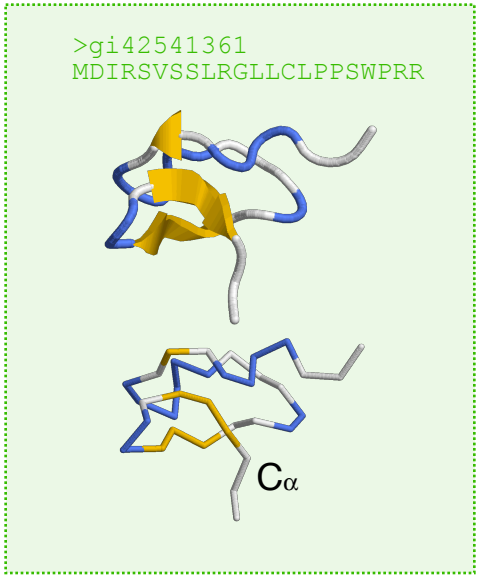
McGuffin LJ, Jones DT. (2003) Bioinformatics, 19, pp874

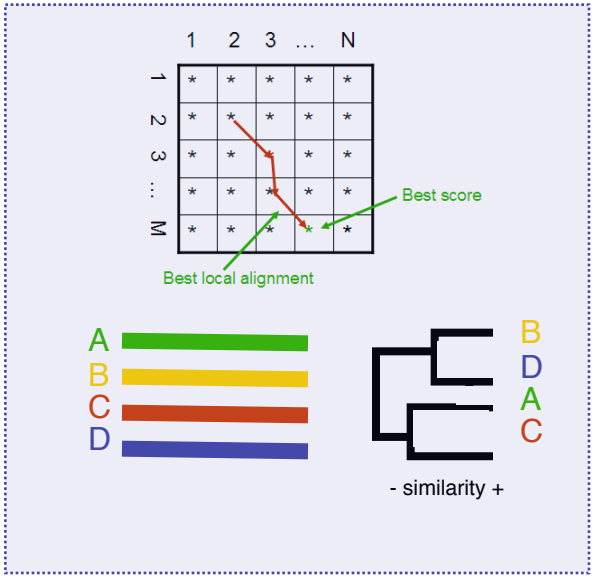
Threading (mGenThreader)

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



Remote homology detection (FUGUE)

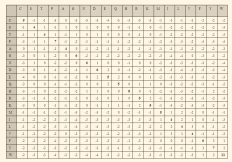




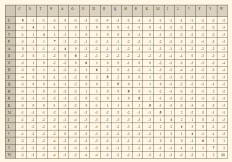
FUGUE

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

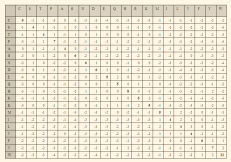
Does not use multiple sequence information



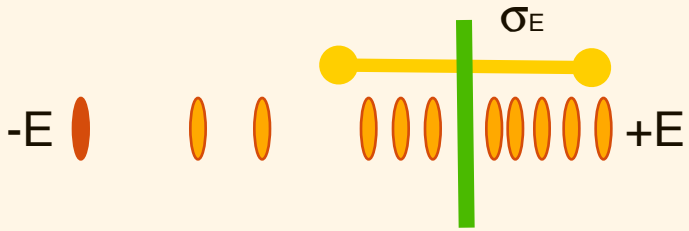
R_{ij}^{H}



R_{ij}^{B}



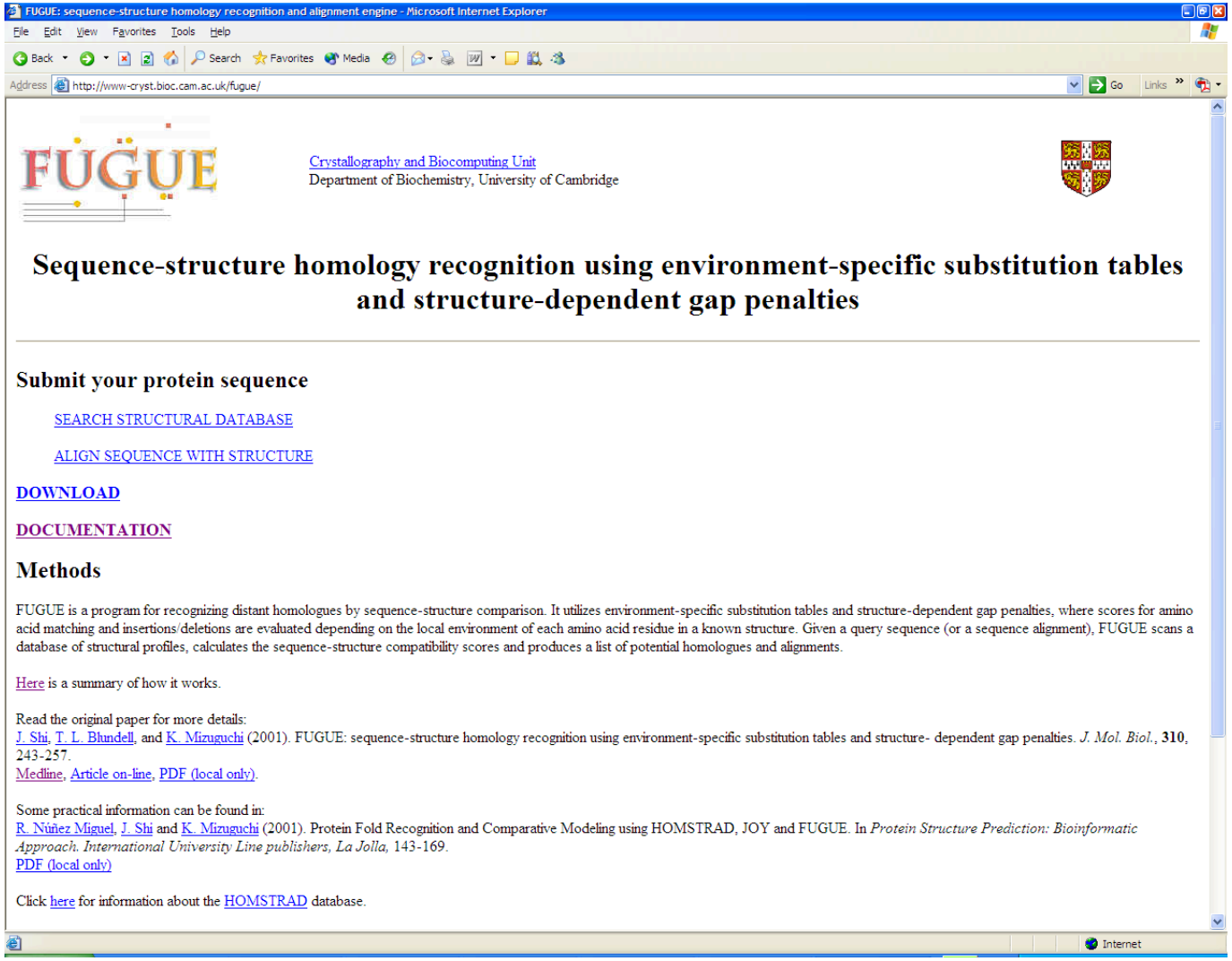
R_{ij}^{C}



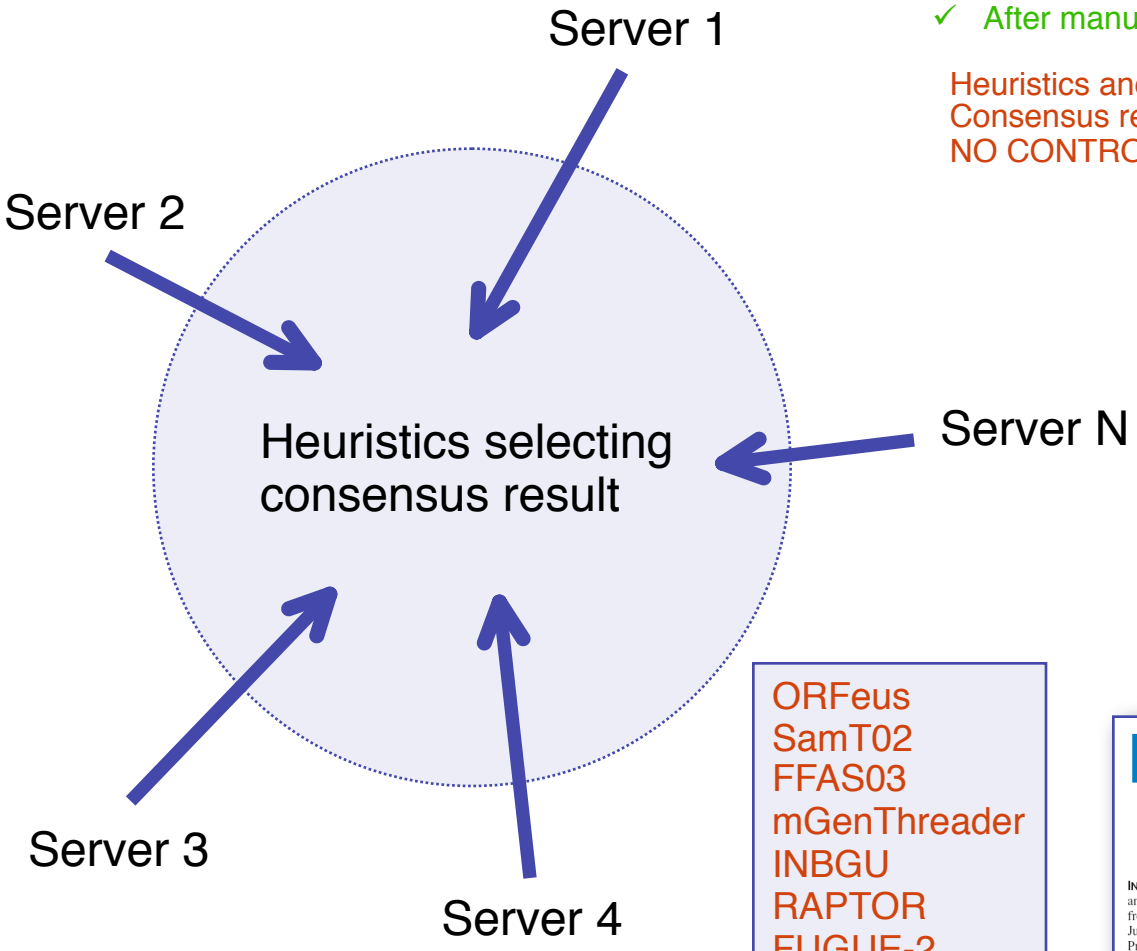
$$Zscore = \frac{(\langle E \rangle - E_m)}{\sigma_E}$$

Remote homology detection (FUGUE)

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



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!

ORFeus
SamT02
FFAS03
mGenThreader
INBGU
RAPTOR
FUGUE-2
3D-PSSM


Letter to the Editor

mRNA Cap-1 Methyltransferase in the SARS Genome

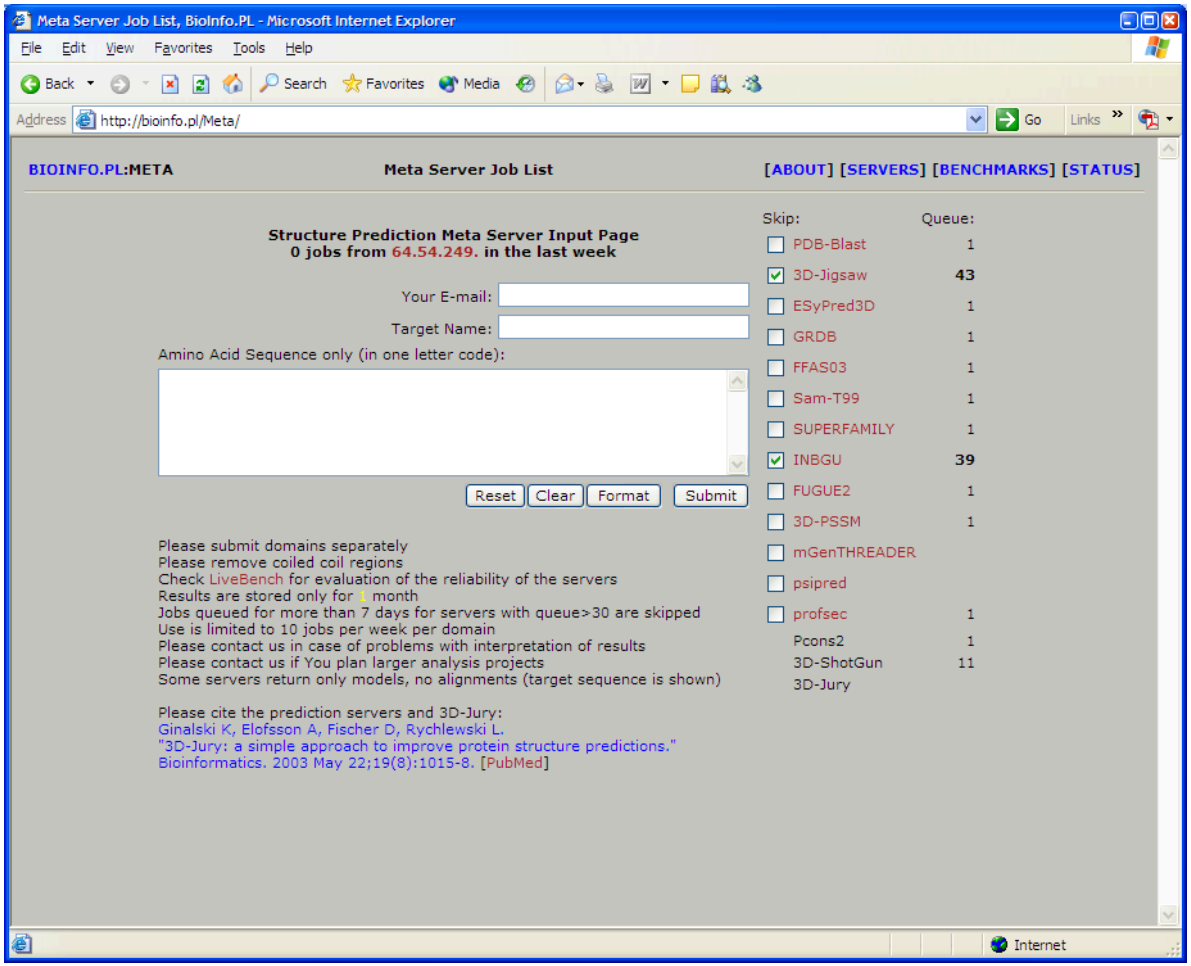
LETTERS

How Unique Is the Rice Transcriptome?

Ginalski K, et al. (2003) Bioinformatics 19 pp1015

Meta-Servers (3D-Jury)

http://bioinfo.pl/Meta/



Complex gap penalty functions

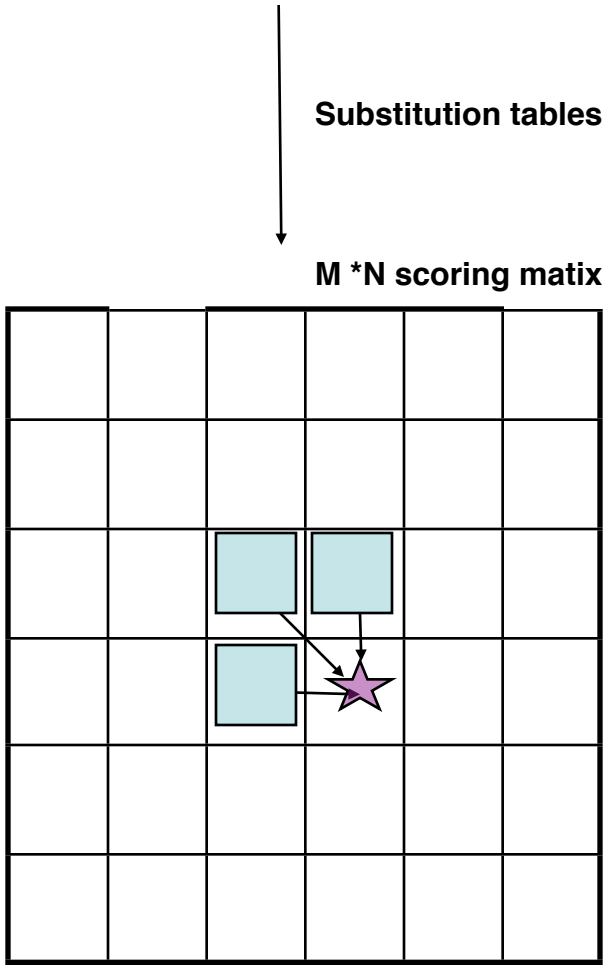
MODELLER SALIGN (ALIGN2D)

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

Regular dynamic programming (ALIGN)

Seq1-> **DEFGHLKSMV**

Seq2 -> **FGHISAVCSSMLPQ**



Creation of a gap penalized

Gap penalty = U + V

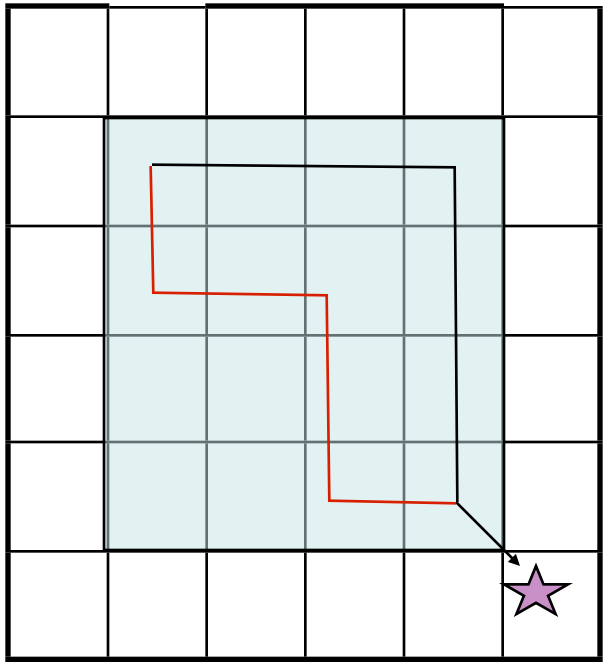
Align2D

Seq1-> **DEFGHLKSMV**

Seq2 -> **FGHISAVCSSMLPQ**

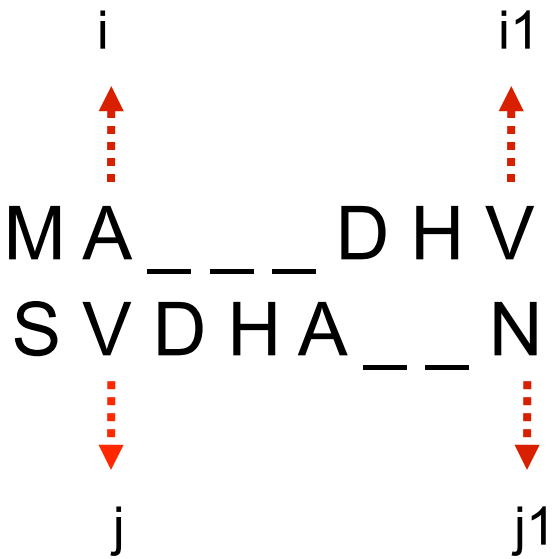
Substitution tables

M *N scoring matix



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

ALIGN2D



L1 = i – i1 –1

L2 = j – j1 -1

G = u*f(i1,i) + (L1+L2)*v – min(L1,L2) *t

opening

extension

diagonal

$$f(i1,i) = 1 + W_a H + W_b S + W_B B + W_{st} C + W_d (\max(0, d - d_0))^g$$

H {=1 if helix unbroken, 0 otherwise

S {=1 if strand unbroken, 0 otherwise

B {average burial

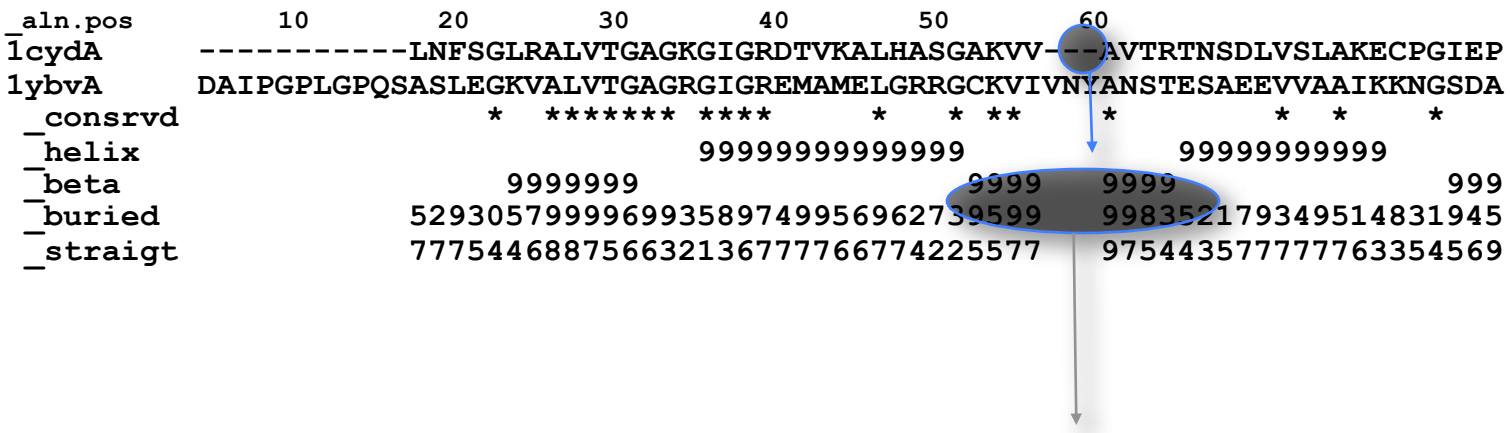
C { curvature, 1 if H or S, f(q) otherwise

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

d = gap spanning distance

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

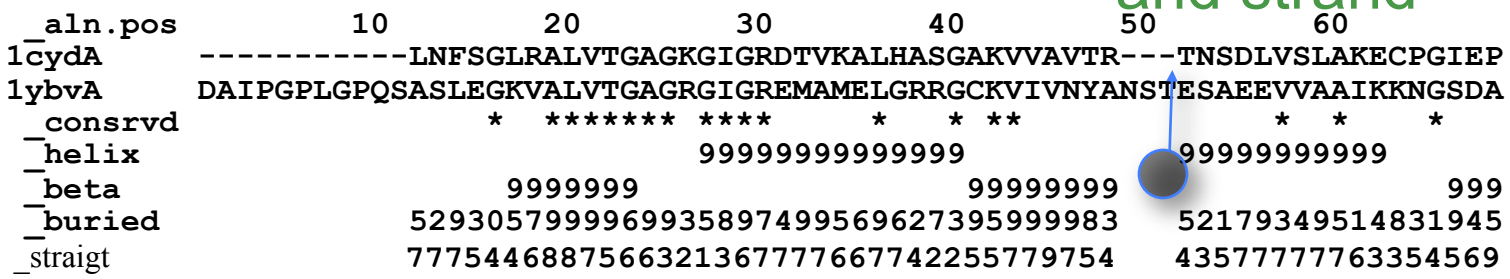
Align



Strand interrupted

Align2D

Gap placed between helix and strand



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

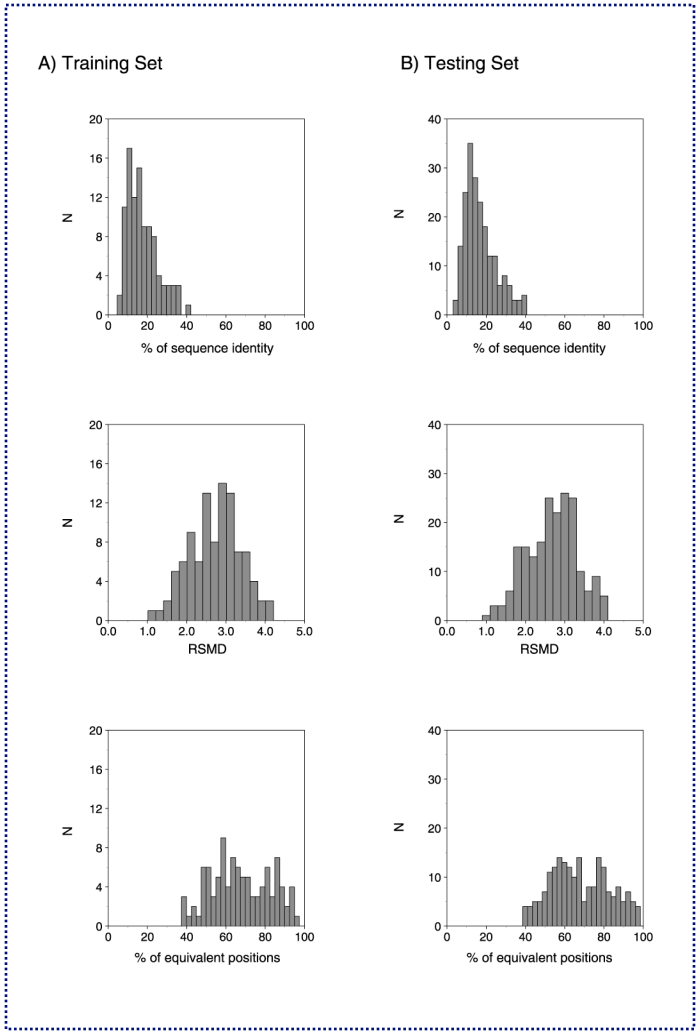
387

Filter: MAMMOTH alignments with
- > 50% EqPos

300

100 Training set

200 Testing set



Methods: Evaluated methods

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

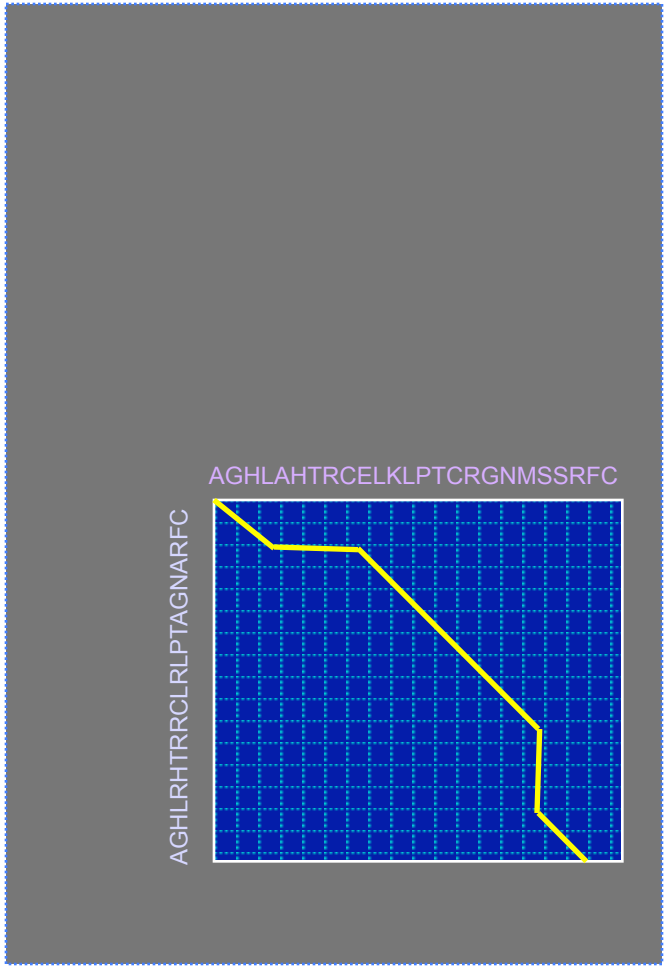
BLAST2SEQ: Local method

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

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

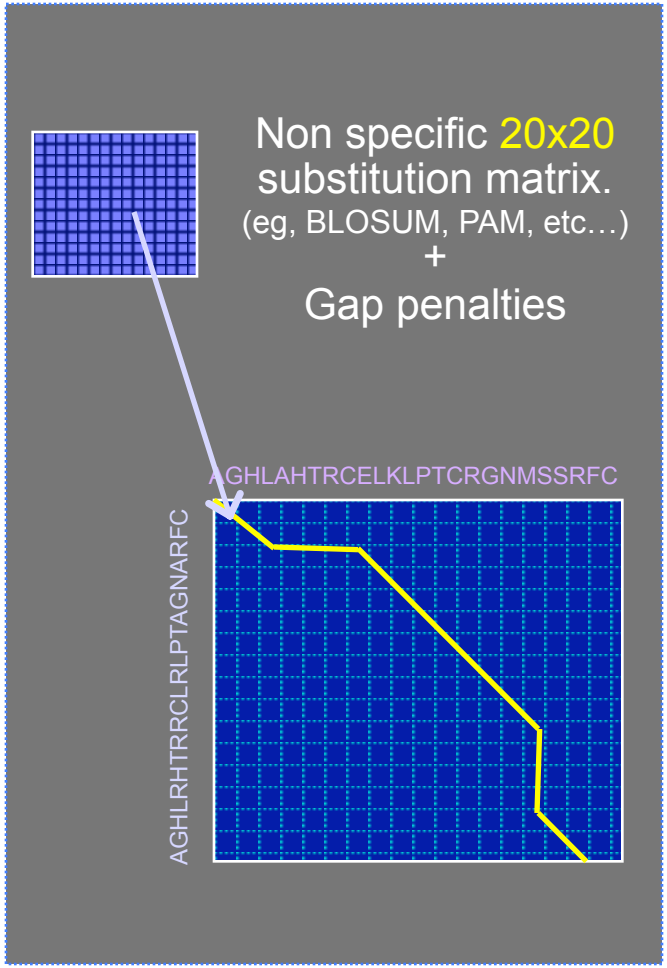
BLAST2SEQ: Local method

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

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

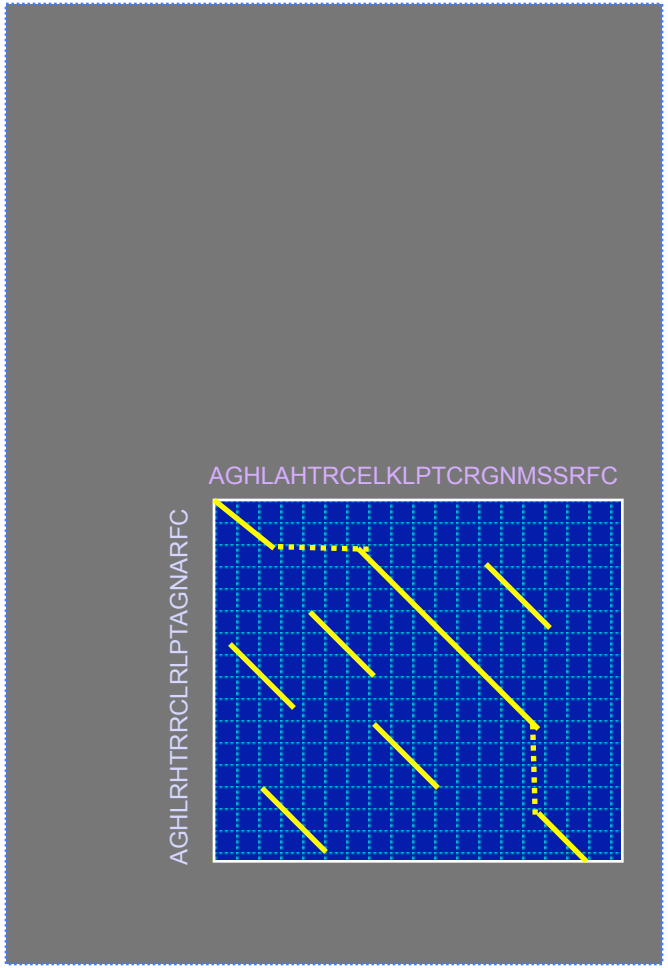
BLAST2SEQ: Local method

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

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

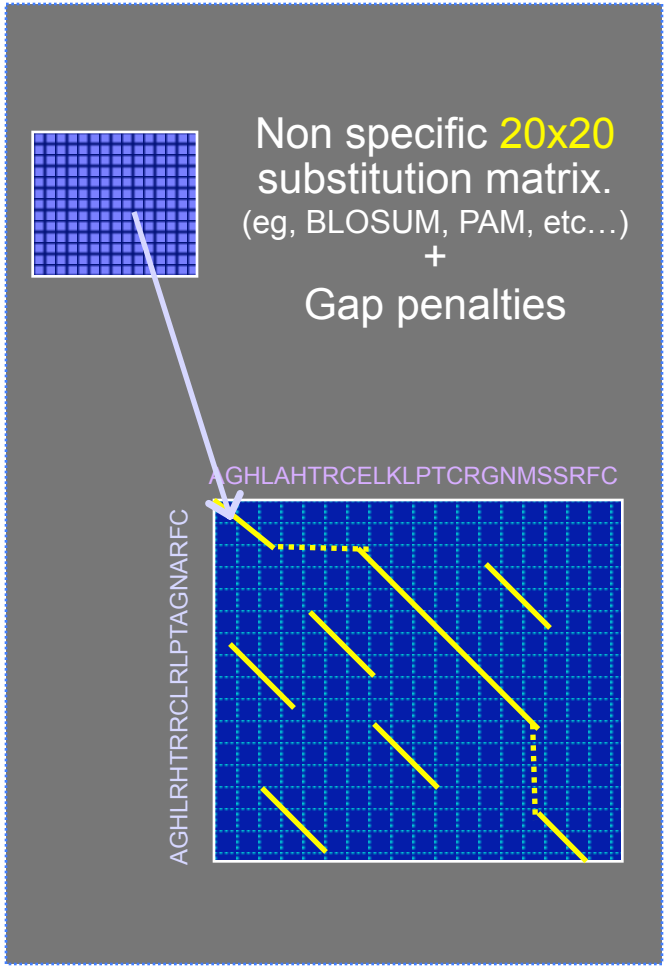
BLAST2SEQ: Local method

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

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

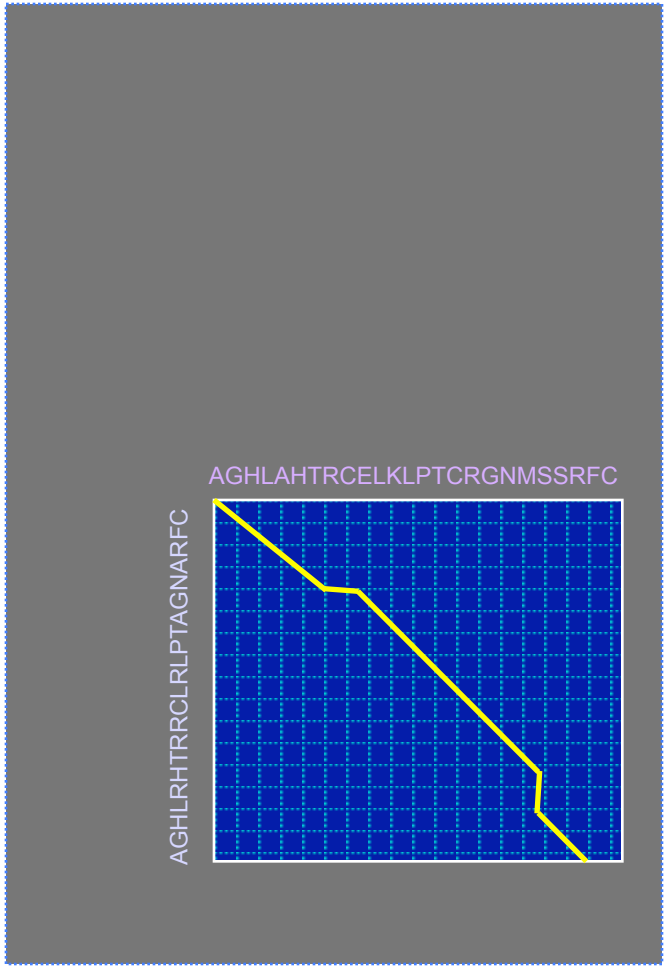
BLAST2SEQ: Local method

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

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

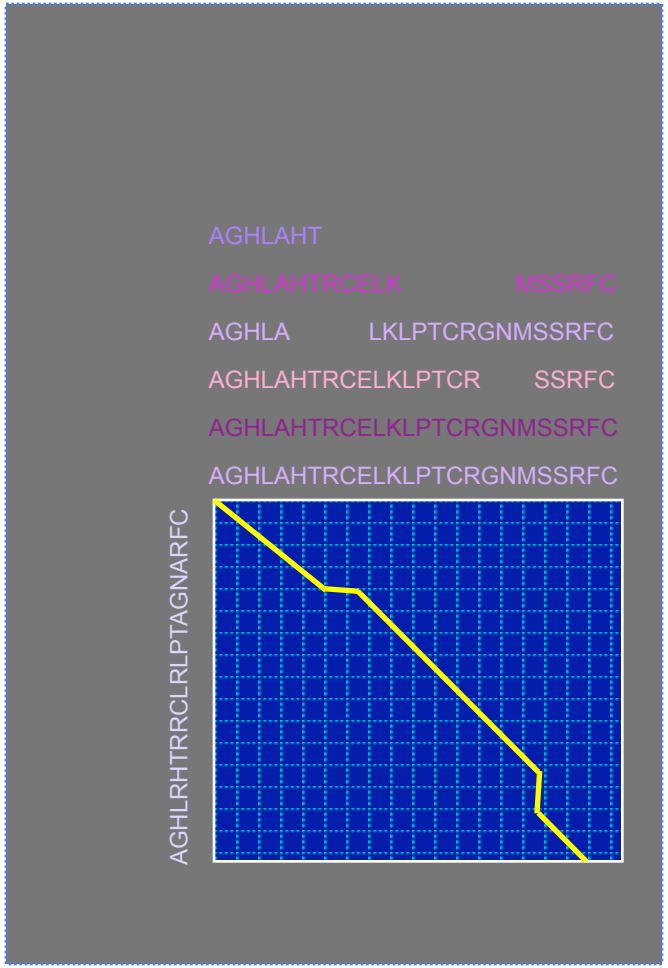
BLAST2SEQ: Local method

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

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

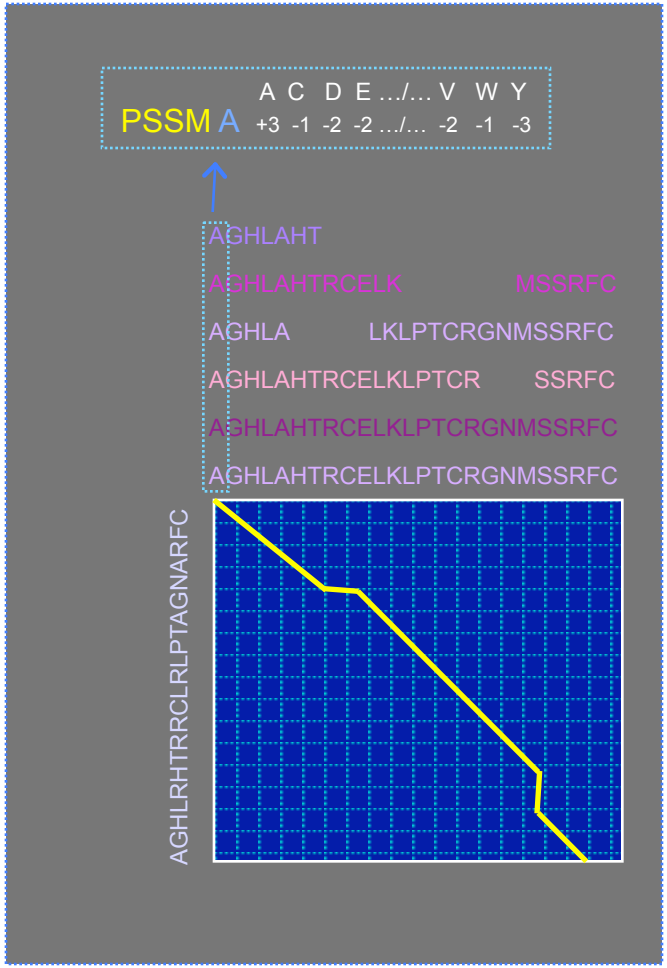
BLAST2SEQ: Local method

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

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

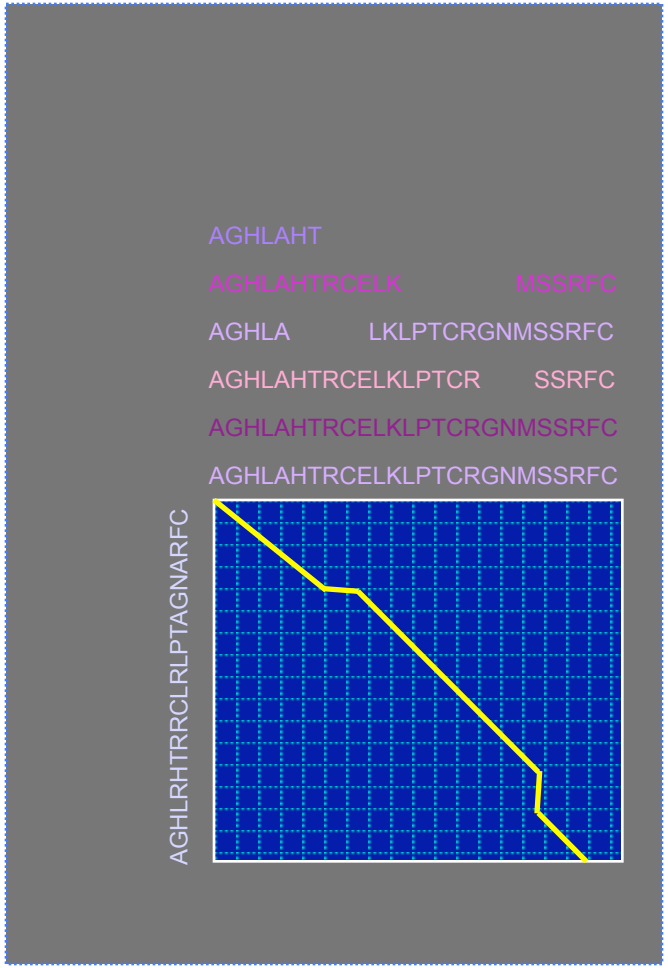
BLAST2SEQ: Local method

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

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

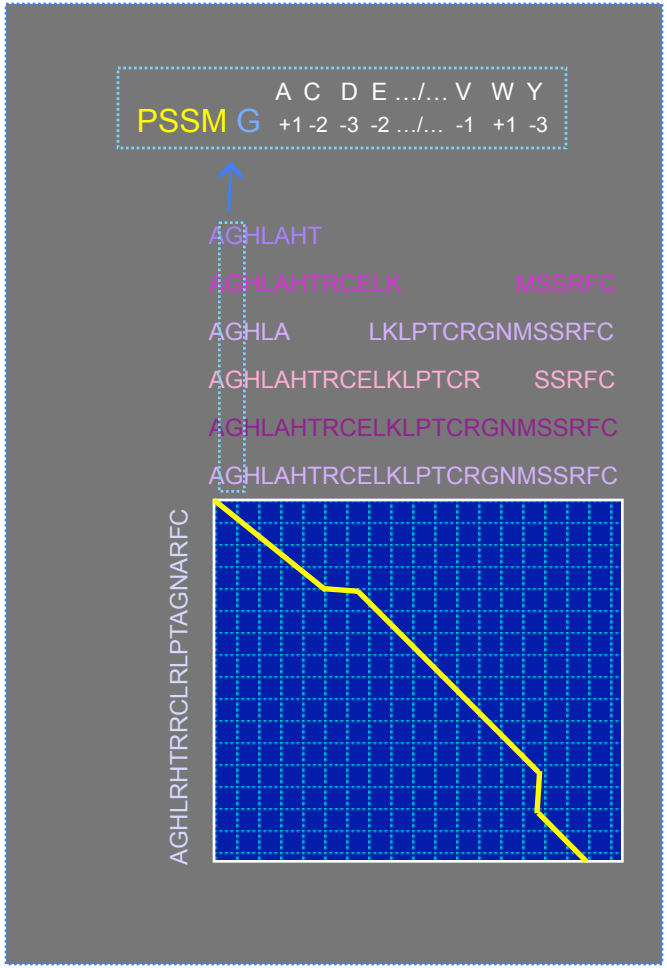
BLAST2SEQ: Local method

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.

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

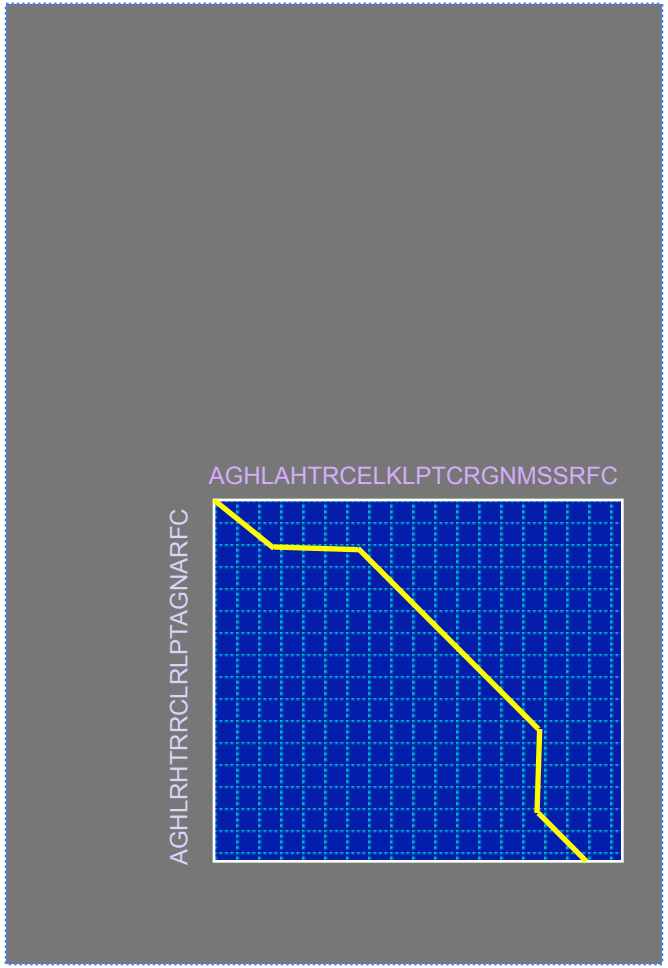
BLAST2SEQ: Local method

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.

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

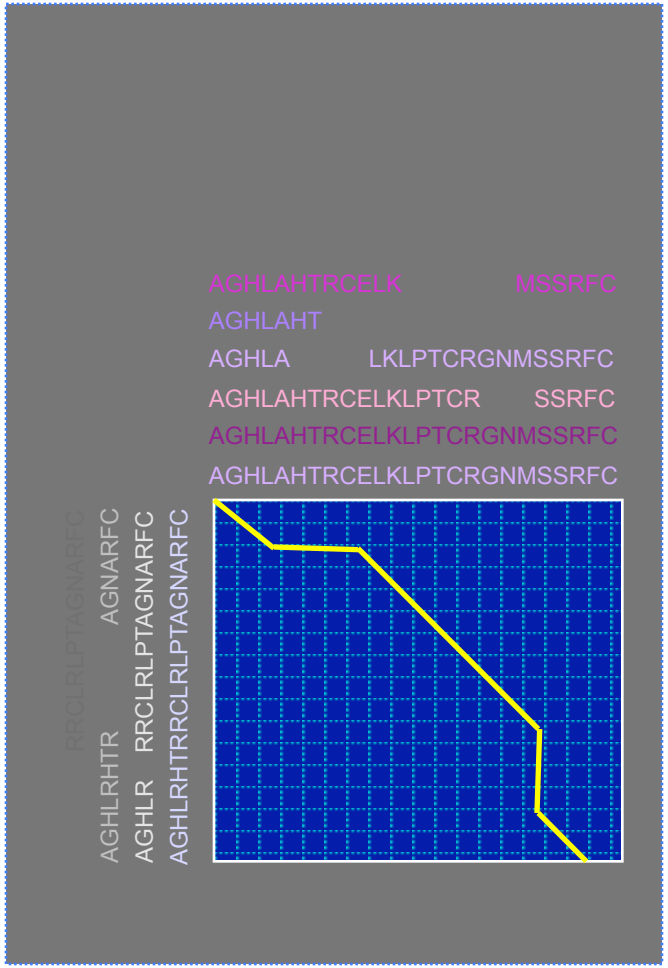
BLAST2SEQ: Local method

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. SALIGN.

ALIGN4D protocol	Profile	Comparison	Open	Extension
CC_{PBP}	PSI-BLAST	Correlation Coefficient	-300	0
CC_{HH}	Henikoff-Henikoff	Correlation Coefficient	-300	0
CC_{HS}	H-H + similarity weight	Correlation Coefficient	-150	0
ED_{PBP}	PSI-BLAST	Euclidian Distance	-450	-30
ED_{HH}	Henikoff-Henikoff	Euclidian Distance	-550	0
ED_{HS}	H-H + similarity weight	Euclidian Distance	-450	-10
DP_{PBP}	PSI-BLAST	Dot Product	-250	-30
DP_{HH}	Henikoff-Henikoff	Dot Product	-550	0
DP_{HS}	H-H + similarity weight	Dot Product	-100	-30
JS_{HH}	Henikoff-Henikoff	Jansen-Shannon Distance	-150	0
JS_{HS}	H-H + similarity weight	Jansen-Shannon Distance	-250	0

Methods: Coverage and accuracy

High coverage

Low accuracy

High accuracy

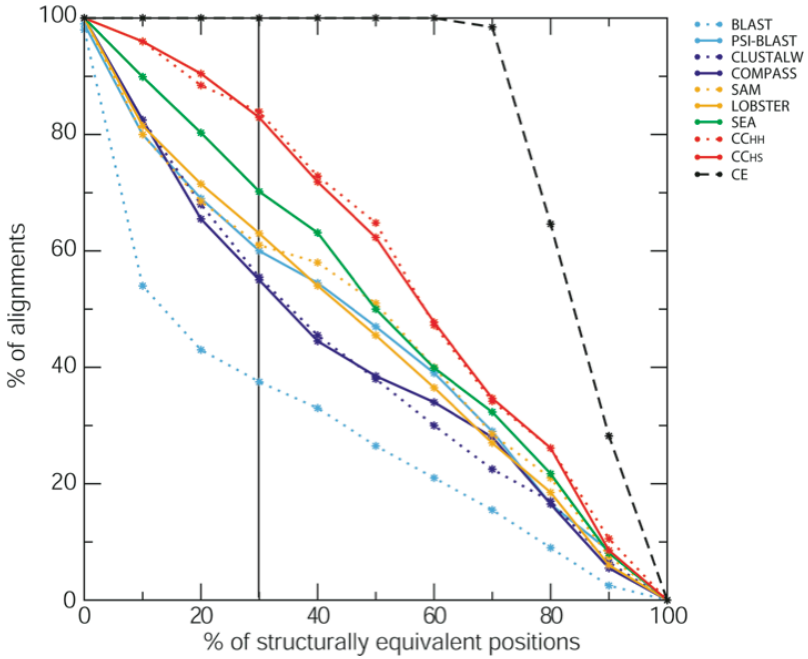
Low coverage

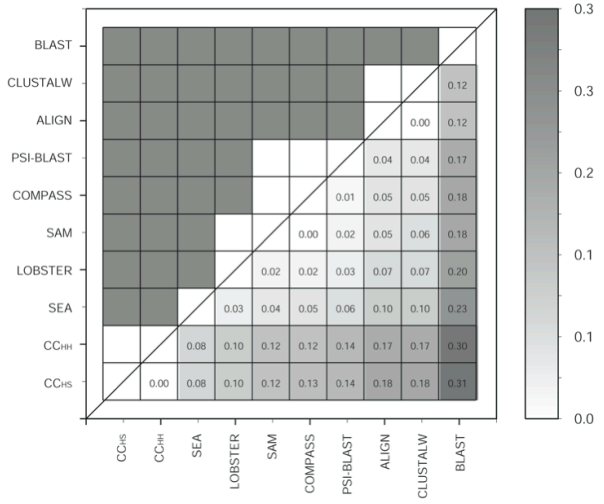
High accuracy

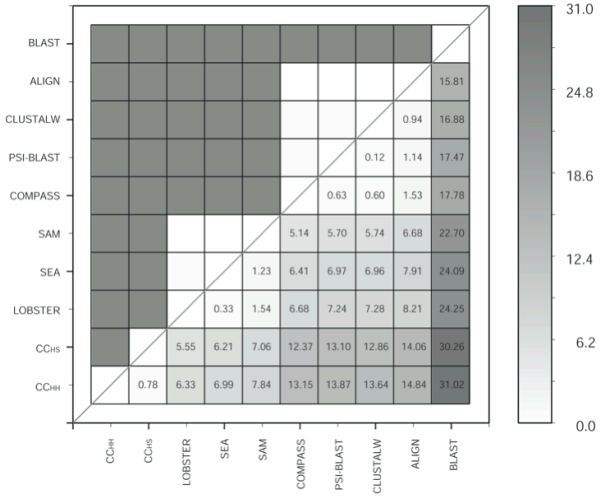
Low accuracy

Results: Comparison of alignment dependent measures

Method	CE overlap [%]	Shift score	RMSD [A]	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



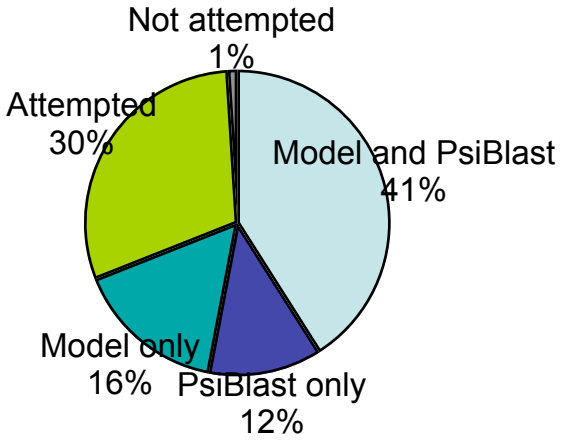




Results. Turn over.

Mycoplasma genitalium MODPIPE Models

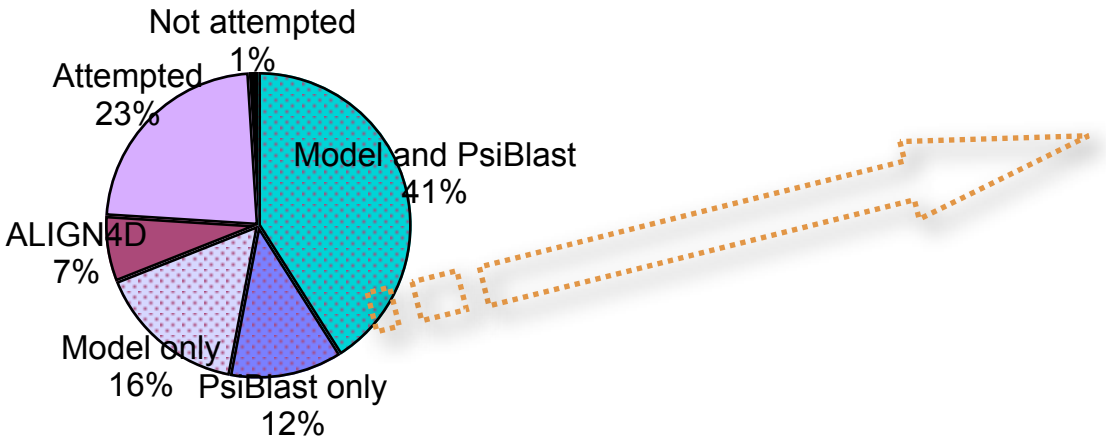
Number of ORFs	479
Average ORF length	364



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”.

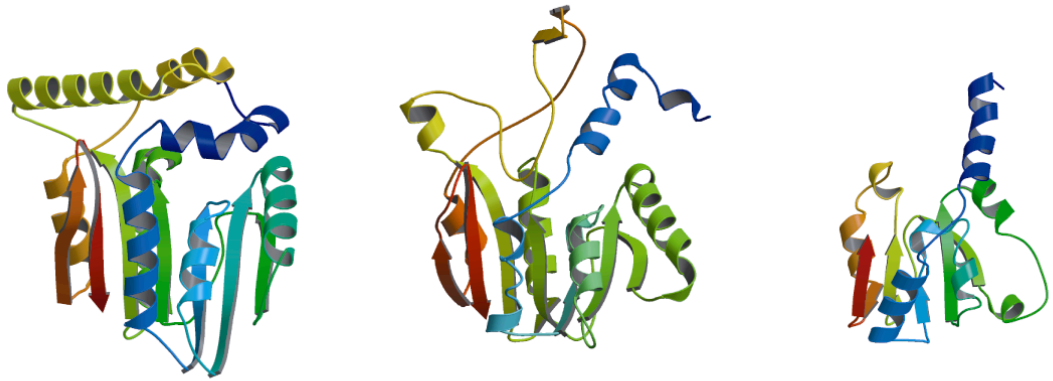
Examples: T0092 model

- 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



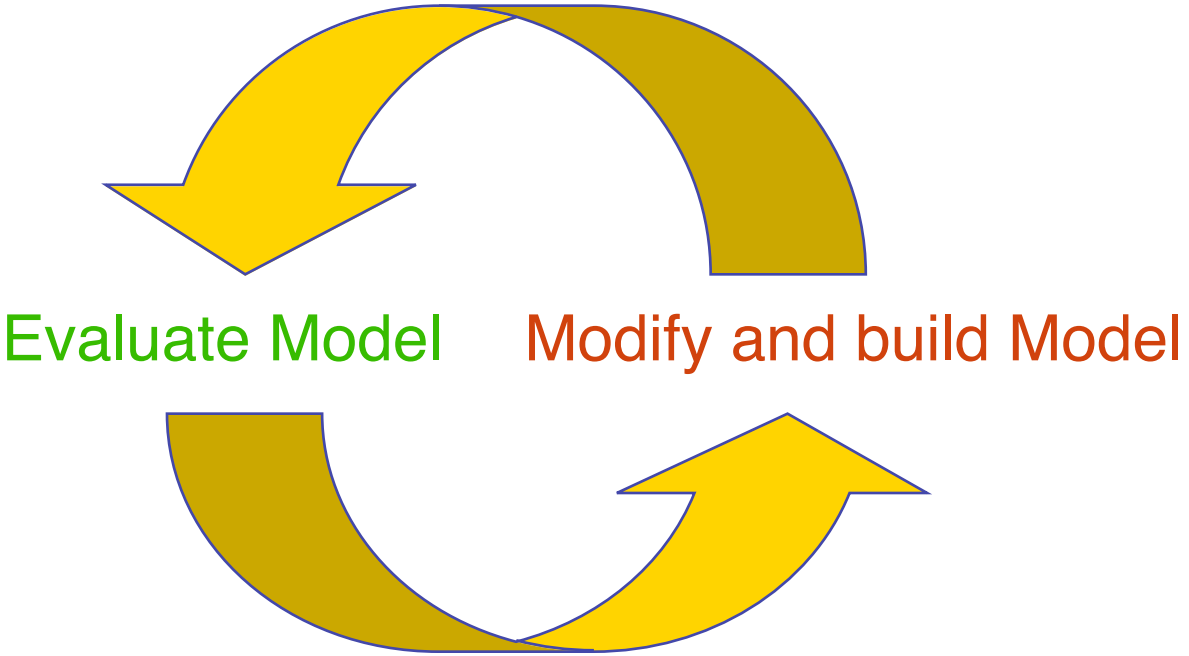
X-Ray structure | ALIGN4D (CC_{PBP}) model | Psi-Blast model

Iterative process

MOULDER

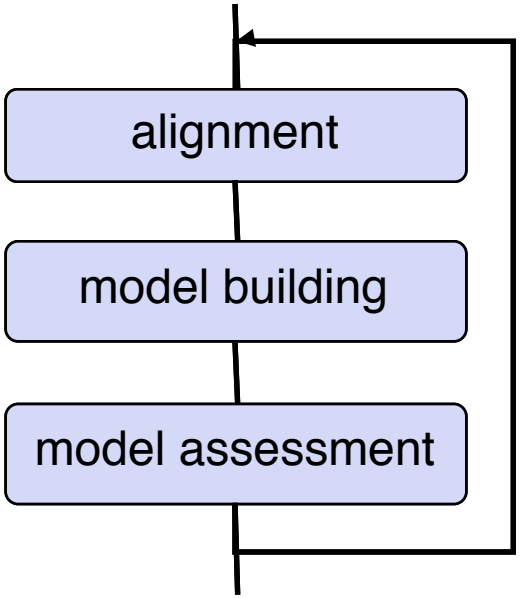
John, B. and Sali, A. (2003) Nucleic Acids Research. **31**:1982-1992

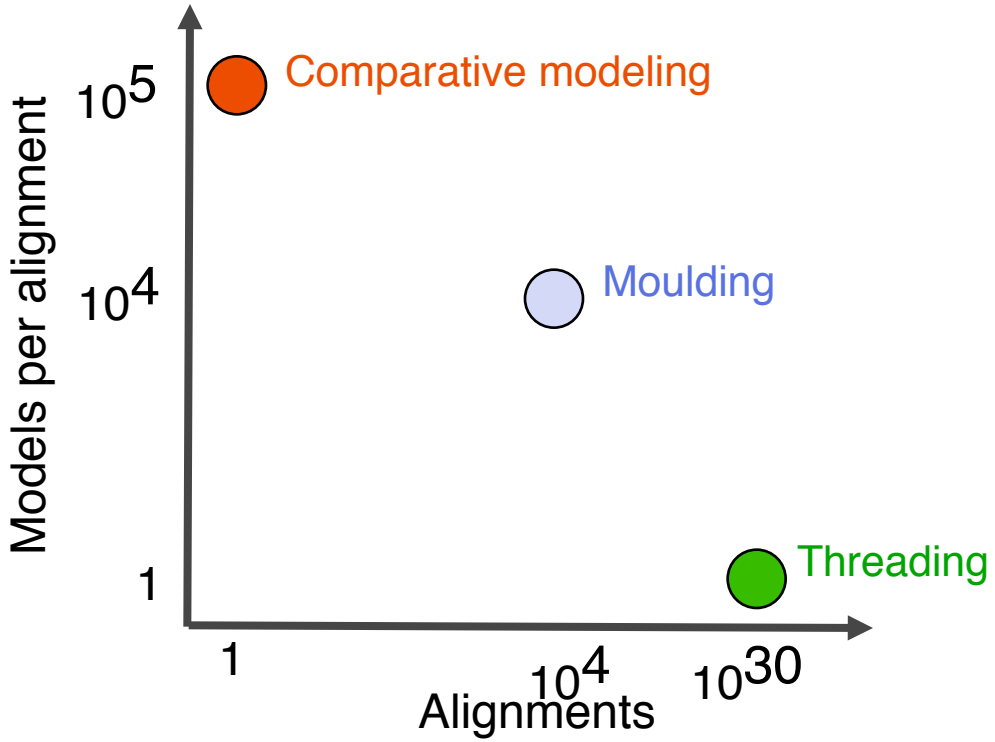
Iterative process… better models(?)



Moulding: iterative alignment, model building, model assessment

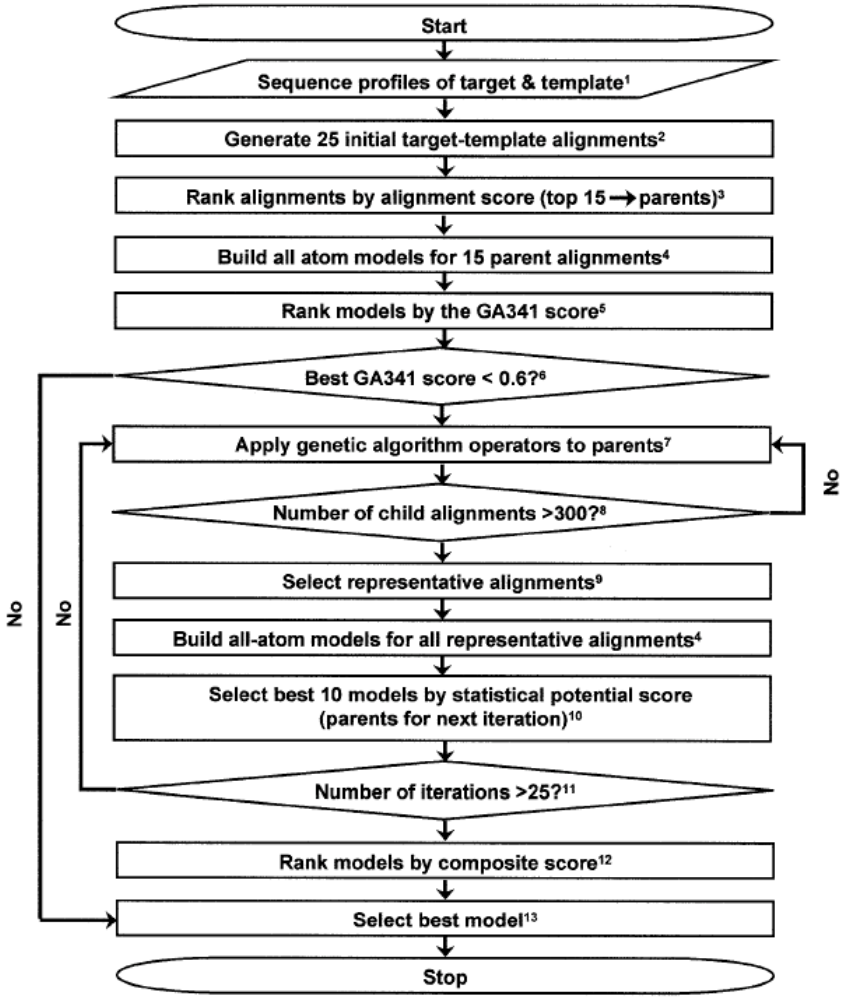
B. John, A. Sali. Nucl. Acids Res., 31, 1982-1992, 2003.

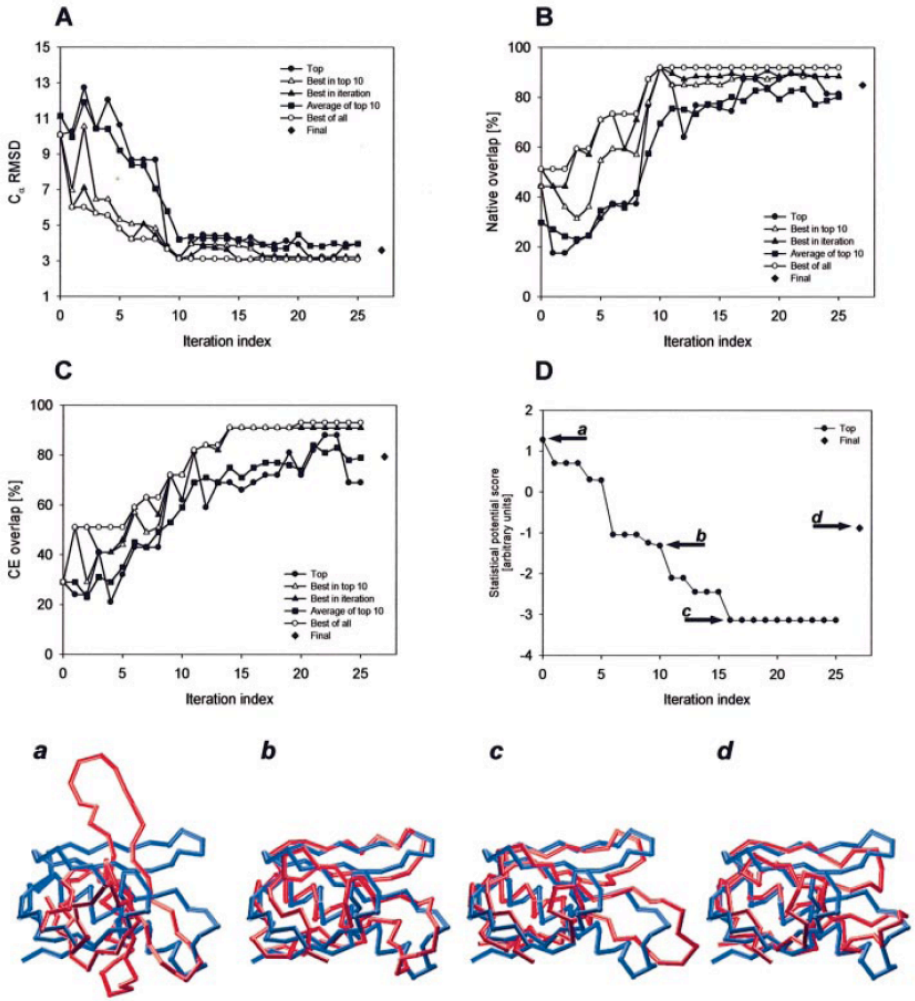




Iterative process… MOULDER

more in model evaluation





John & Sali (2003). NAR 31 pp3982

BMC WorkShop

Protein Structure Prediction

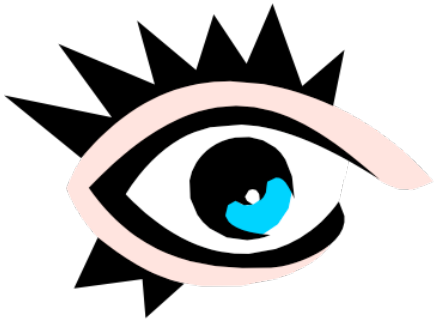
model building

(model assessment)

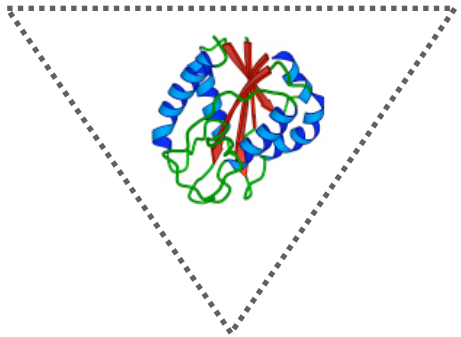
Marc A. Marti-Renom & Damien Devos

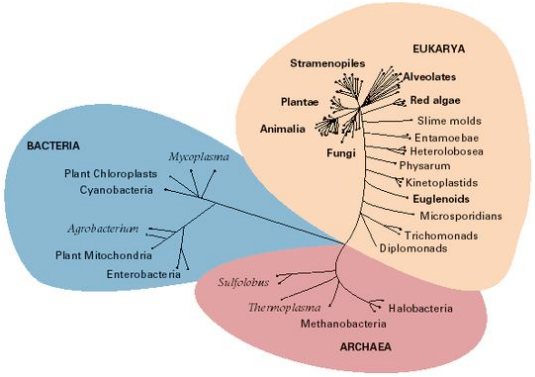
Department of Biopharmaceutical Sciences, UCSF

Information about a protein can come from three distinct sources

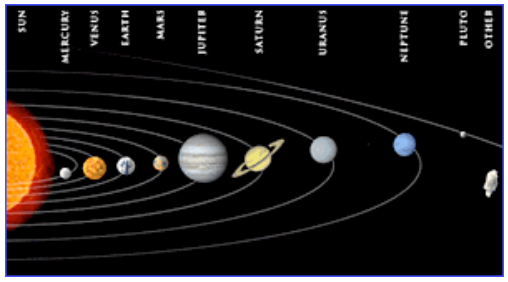


Experimental observations





Statistical rules



Laws of physics

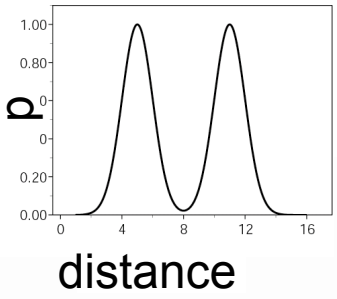
Modeling by optimization

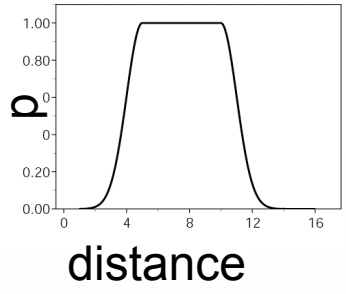
There is nothing but points and restraints on them.

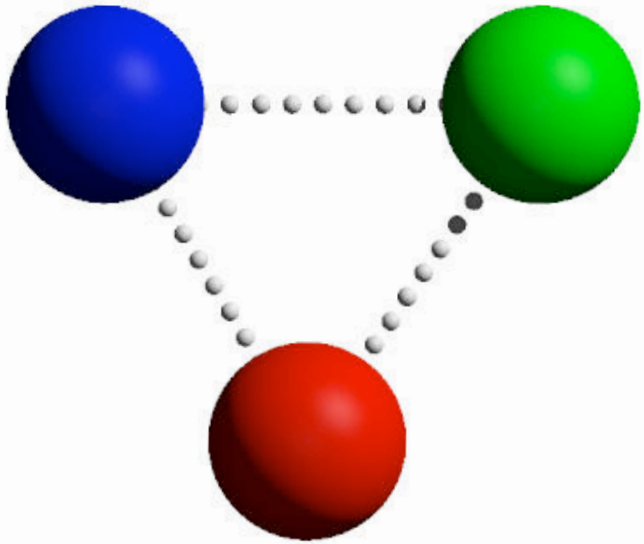
P(r/I) feature

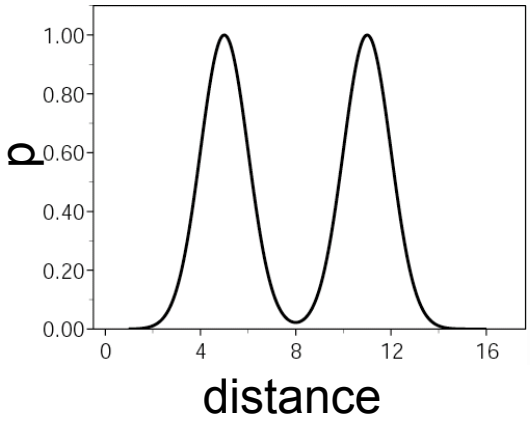
↓

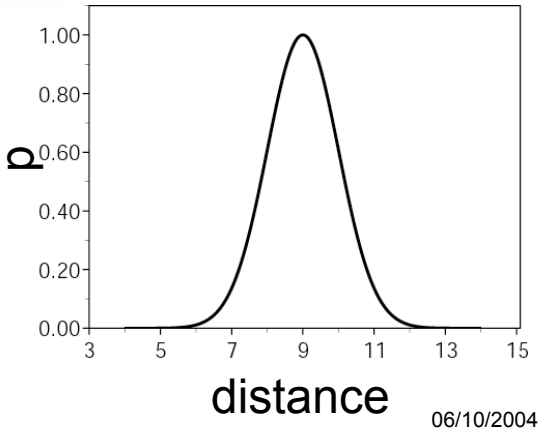
P(R/I) molecule











Classes of methods for comparative protein structure modeling

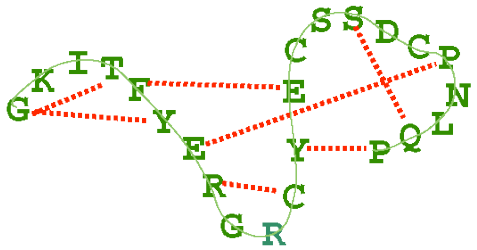
- Model building by assembly of rigid bodies: core, loops, sidechains.
- Model building by segment matching.
- Model building by satisfaction of spatial restraints.

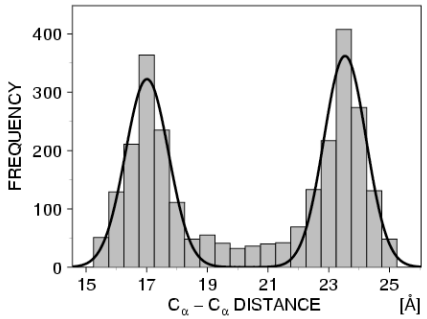
Marti-Renom *et al.* Annu.Rev.Biophys.Biomol.Struct. **29**, 291-325, 2000.

Comparative modeling by satisfaction of spatial restraints
MODELLER

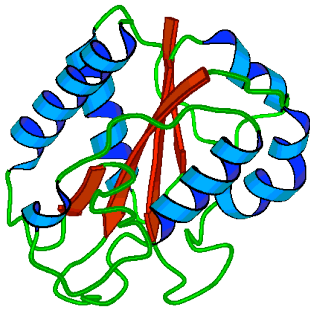
3D GKITFYERGFQGHCYESDC-NLQP…
SEQ GKITFYERG---RCYESDCPNLQP…

1. Extract spatial restraints





2. Satisfy spatial restraints



$$F(\mathbf{R}) = \prod_{i} p_i\,(f_i / I)$$

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

http://salilab.org/

Restraints

p(d)

p(d/d')

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

p(d/d',d'',…)

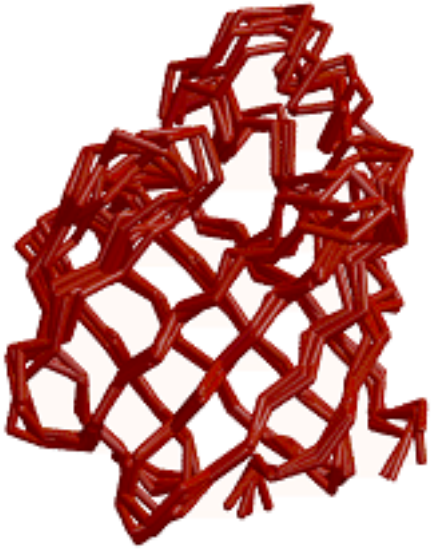
p(S/R,S',R',t,s)

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

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

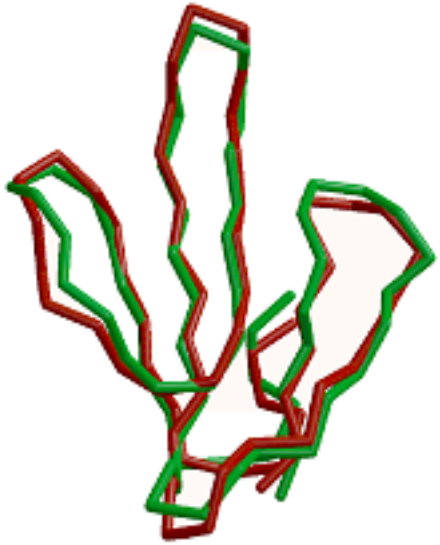
Accuracy and applicability of comparative models

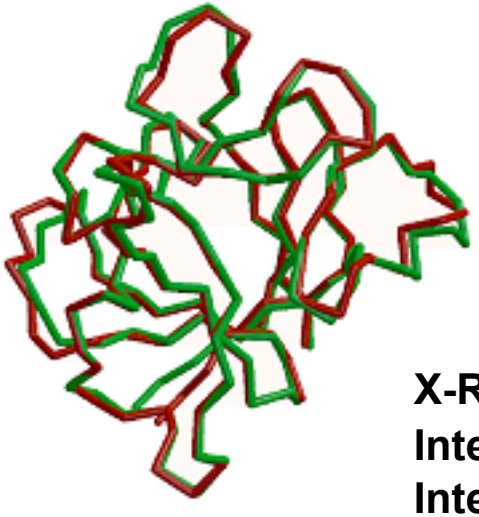
“Biological” significance of modeling errors



NMR
Ileal lipid-binding protein
1eal

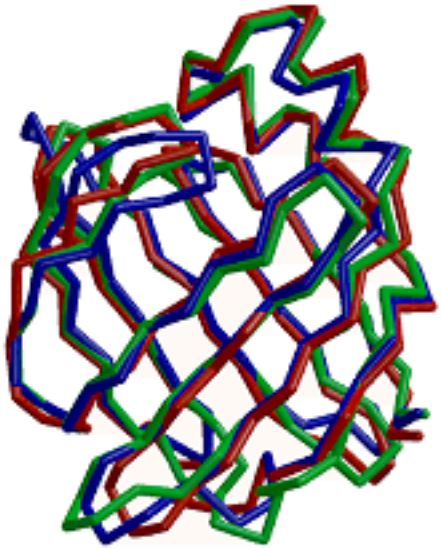
NMR – X-RAY
Erabutoxin 3ebx
Erabutoxin 1era





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

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



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/)

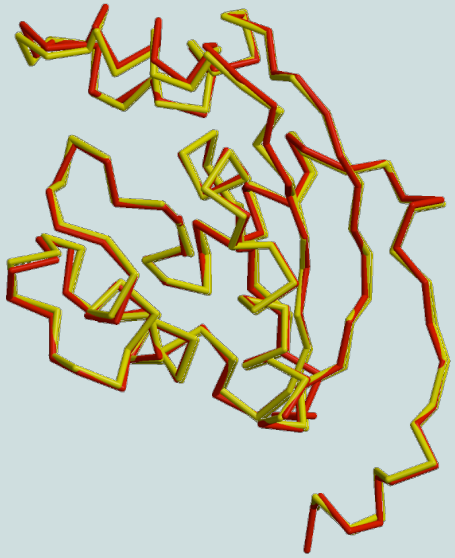
LiveBench (http://bioinfo.pl/)

Model Accuracy

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

HIGH ACCURACY

NM23
Seq id 77%
Cα equiv 147/148
RMSD 0.41Å

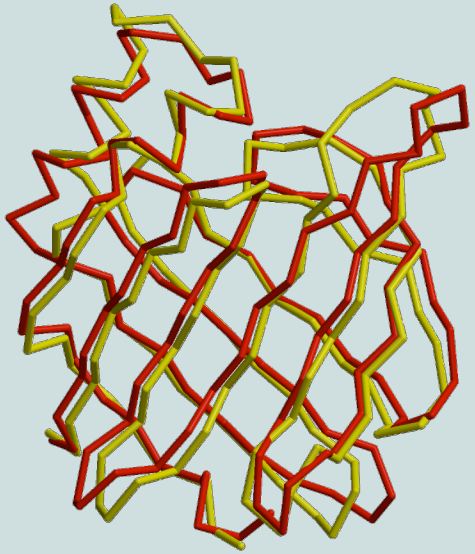


Sidechains
Core backbone
Loops

X-RAY / MODEL

MEDIUM ACCURACY

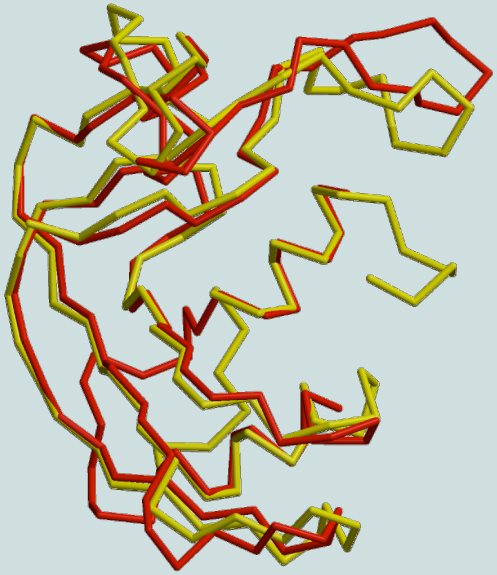
CRABP
Seq id 41%
Cα equiv 122/137
RMSD 1.34Å



Sidechains
Core backbone
Loops
Alignment

LOW ACCURACY

EDN
Seq id 33%
Cα equiv 90/134
RMSD 1.17Å



Sidechains
Core backbone
Loops
Alignment
Fold assignment

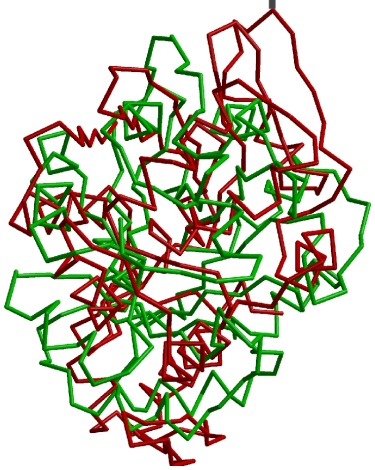
Typical errors in comparative models

MODEL

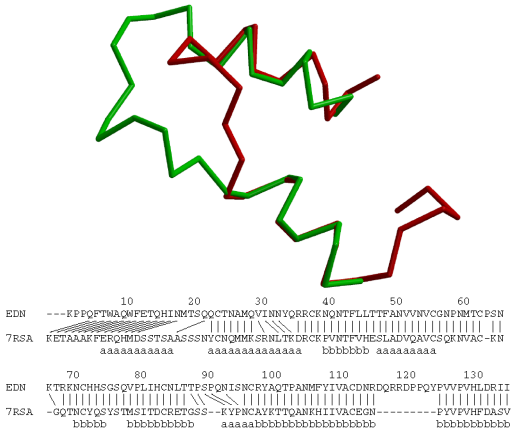
X-RAY

TEMPLATE

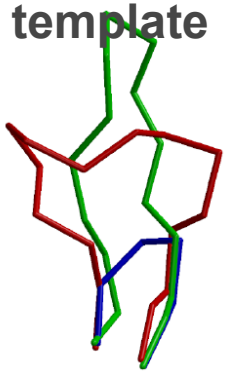
Incorrect template



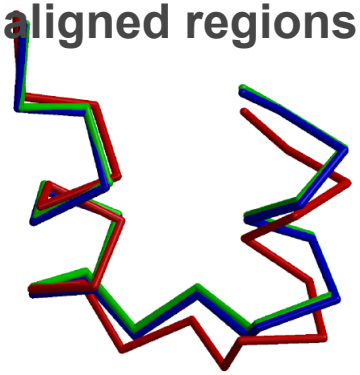
Misalignment



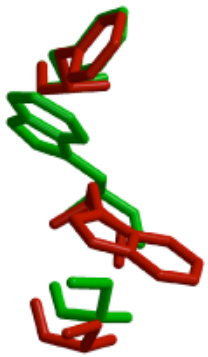
Region without a template



Distortion/shifts in aligned regions



Sidechain packing



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

Typical errors in comparative models

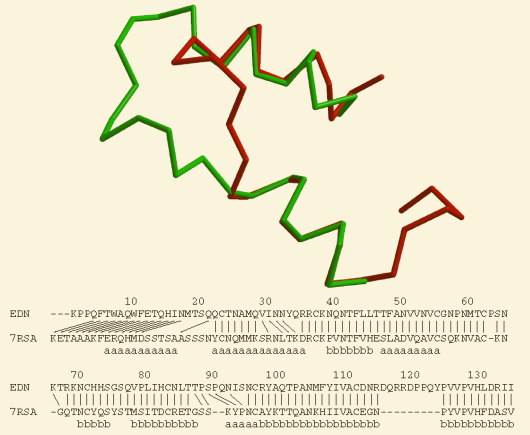
MODEL

X-RAY

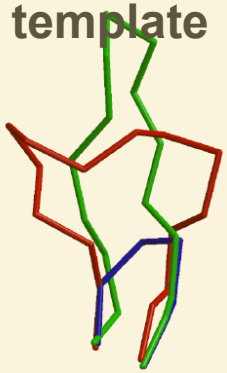
TEMPLATE

Incorrect template

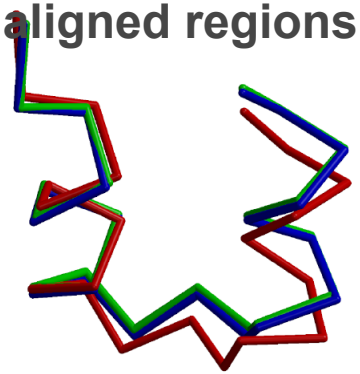
Misalignment



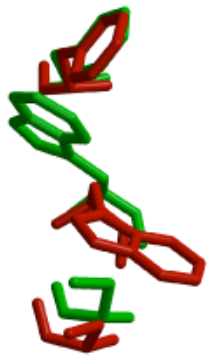
Region without a template



Distortion/shifts in aligned regions

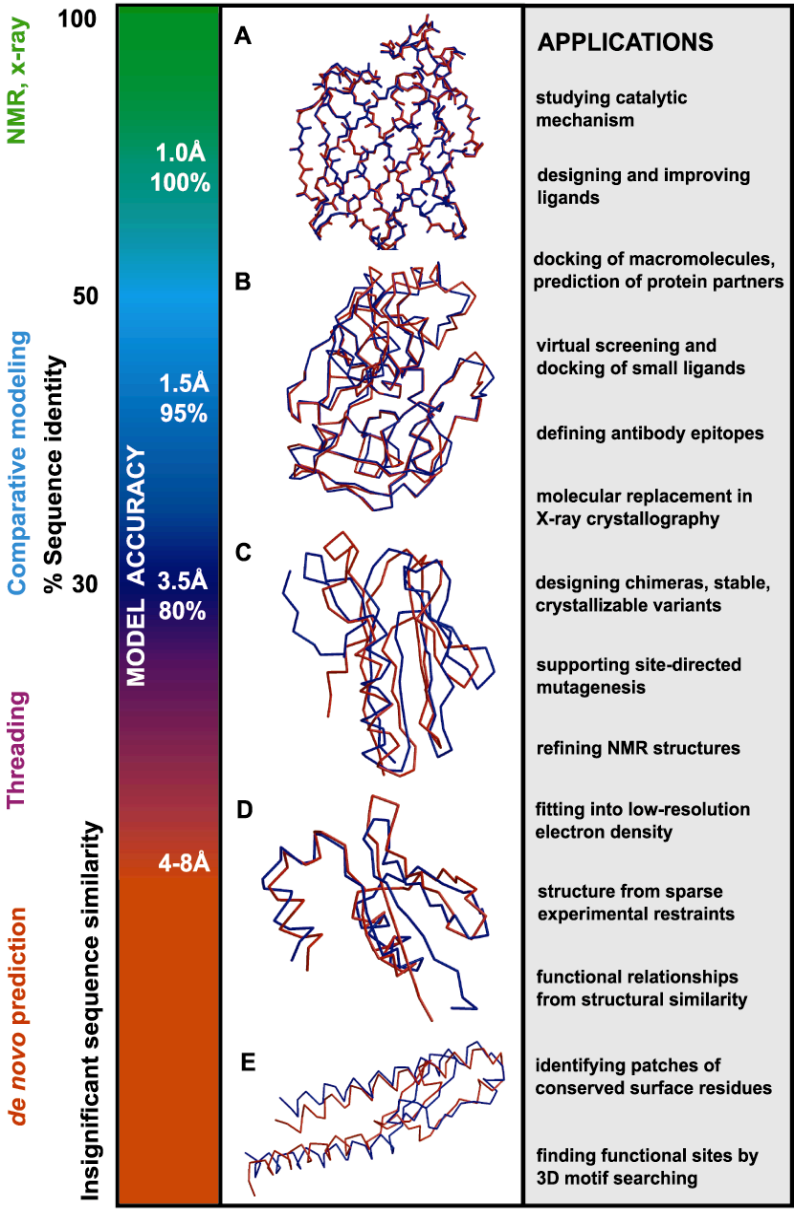


Sidechain packing



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

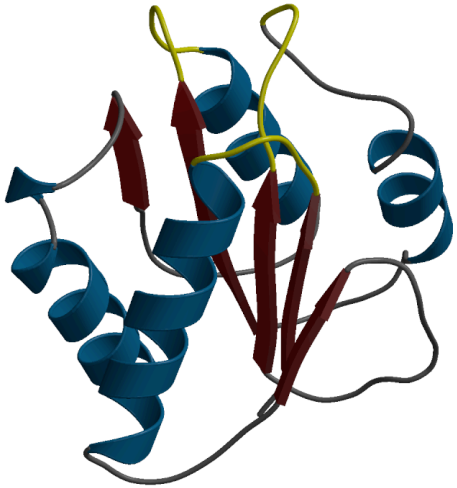
Utility of protein structure models, despite errors



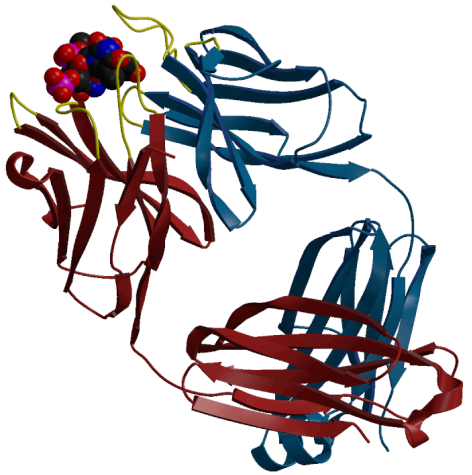
D. Baker & A. Sali.
Science **294,** 93, 2001.

Modeling of loops in protein structures (modeling of insertions)

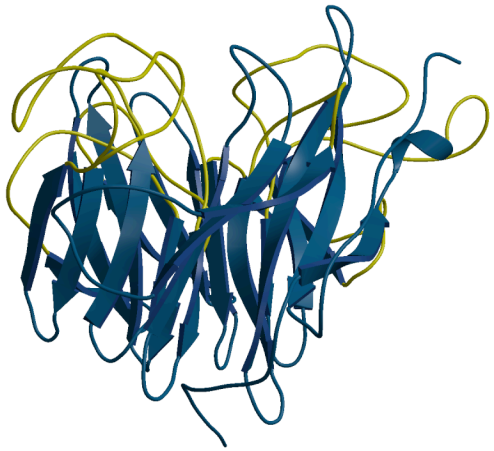
Loop Modeling in Protein Structures



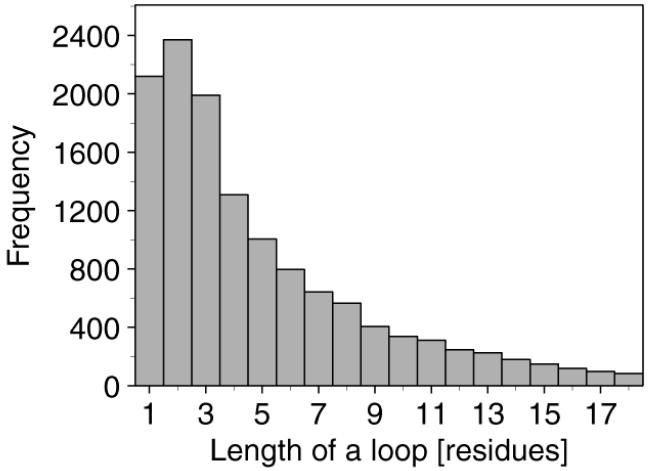
α+β barrel: flavodoxin



IG fold: immunoglobulin



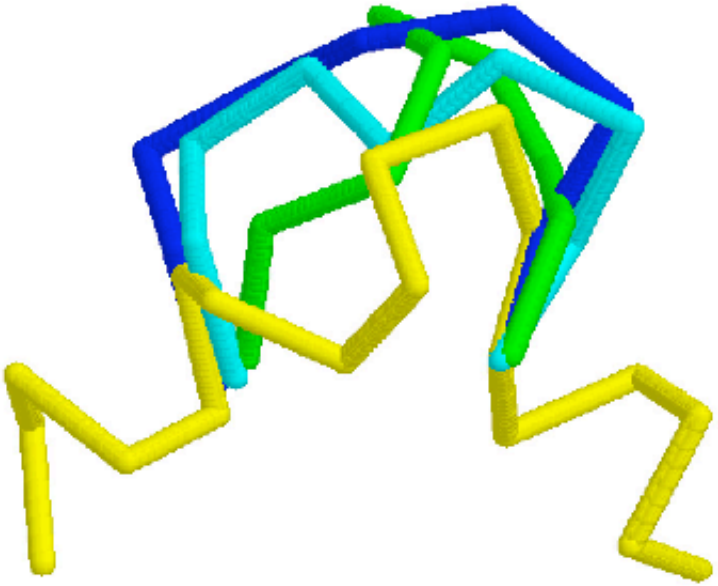
antiparallel β-barrel

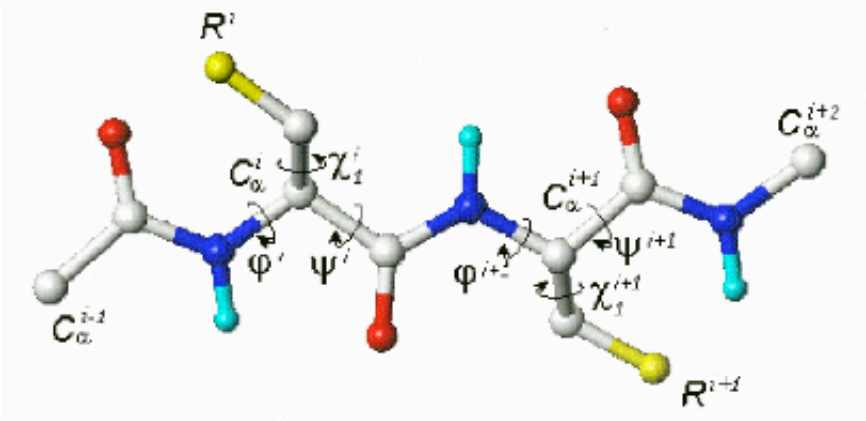


A. Fiser, R. Do & A. Šali, *Prot. Sci.* **9**, 1753, 2000

Loop modeling strategies

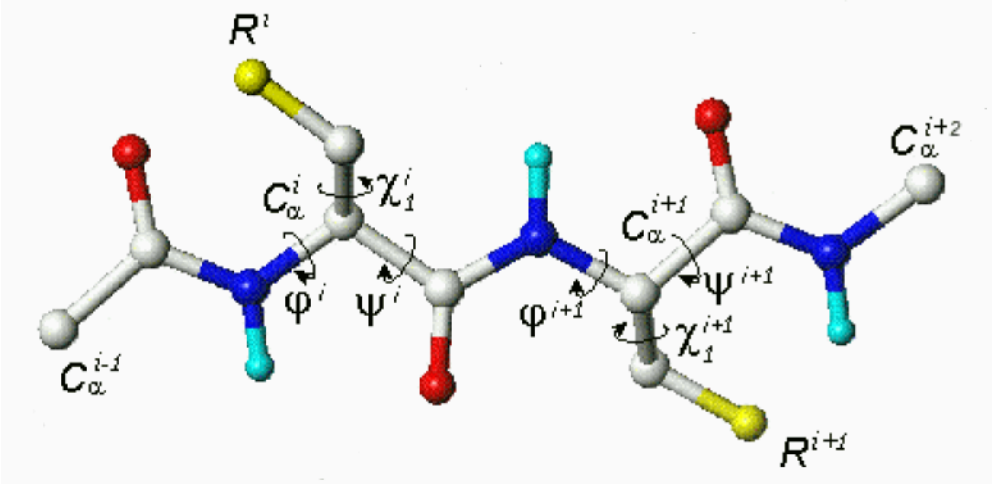
Database search Conformational search





- 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).

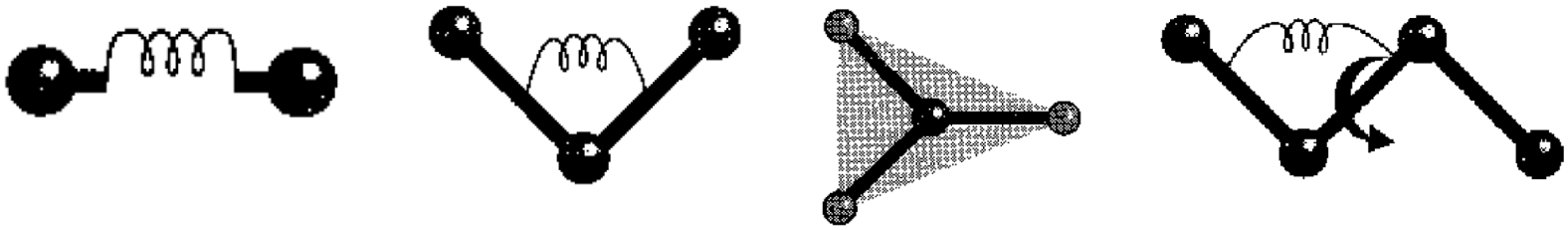
3. Non-bonded contacts.

Energy Function for Loop Modeling

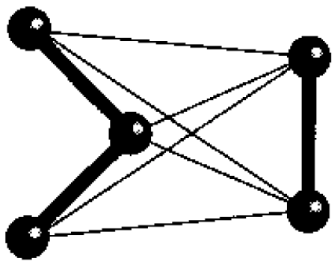
The energy function is a sum of many terms:

1) Statistical preferences for dihedral angles:

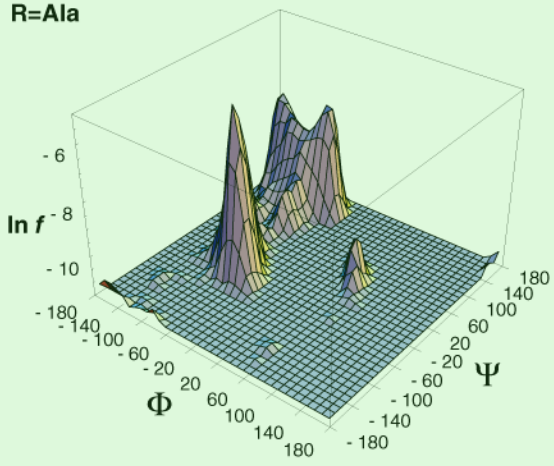
2) Restraints from the CHARMM-22 force field:

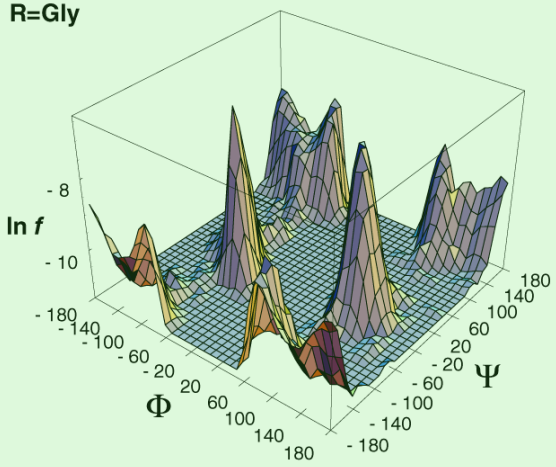


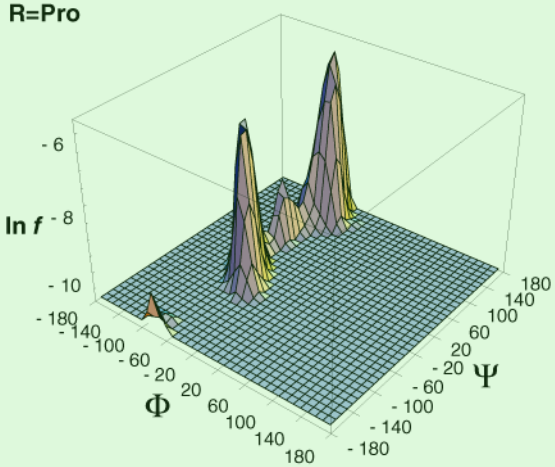
3) Statistical potential for non-bonded contacts:

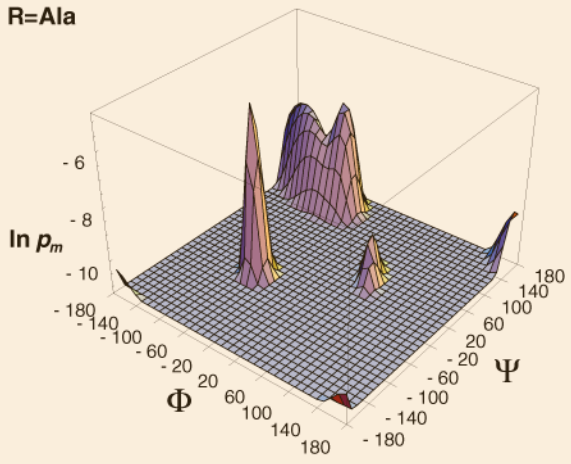


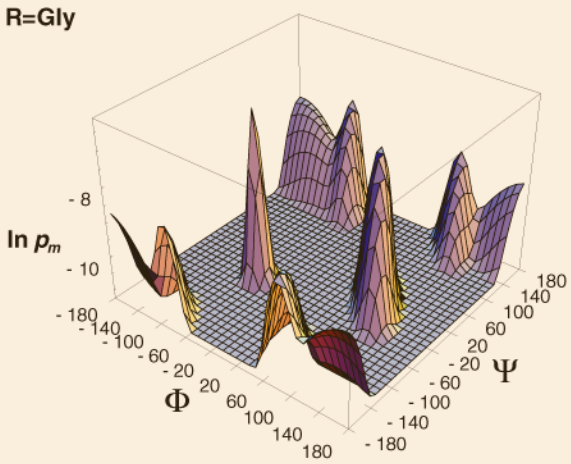
Mainchain Terms for Loop Modeling

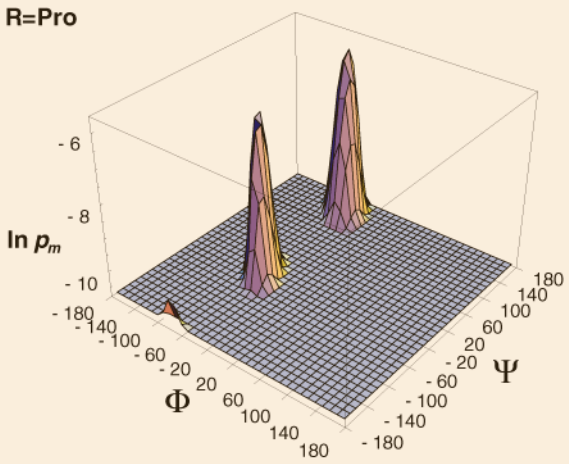






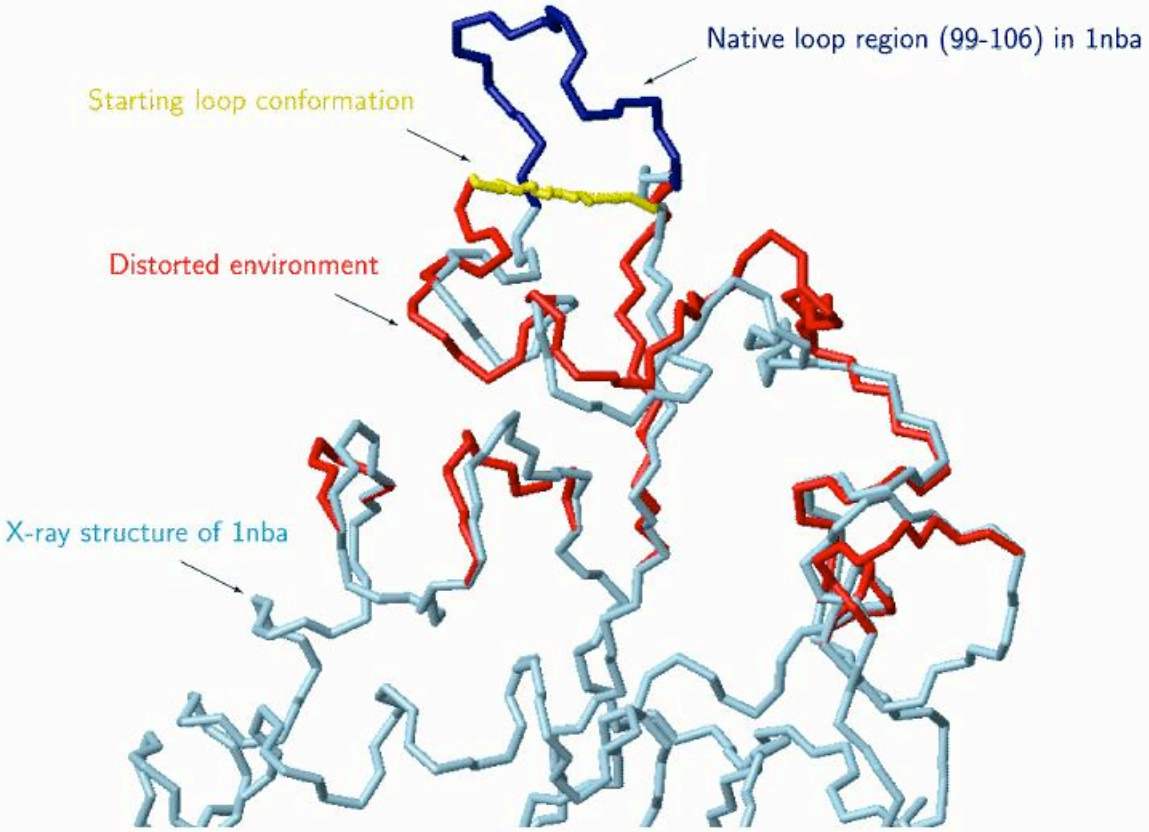




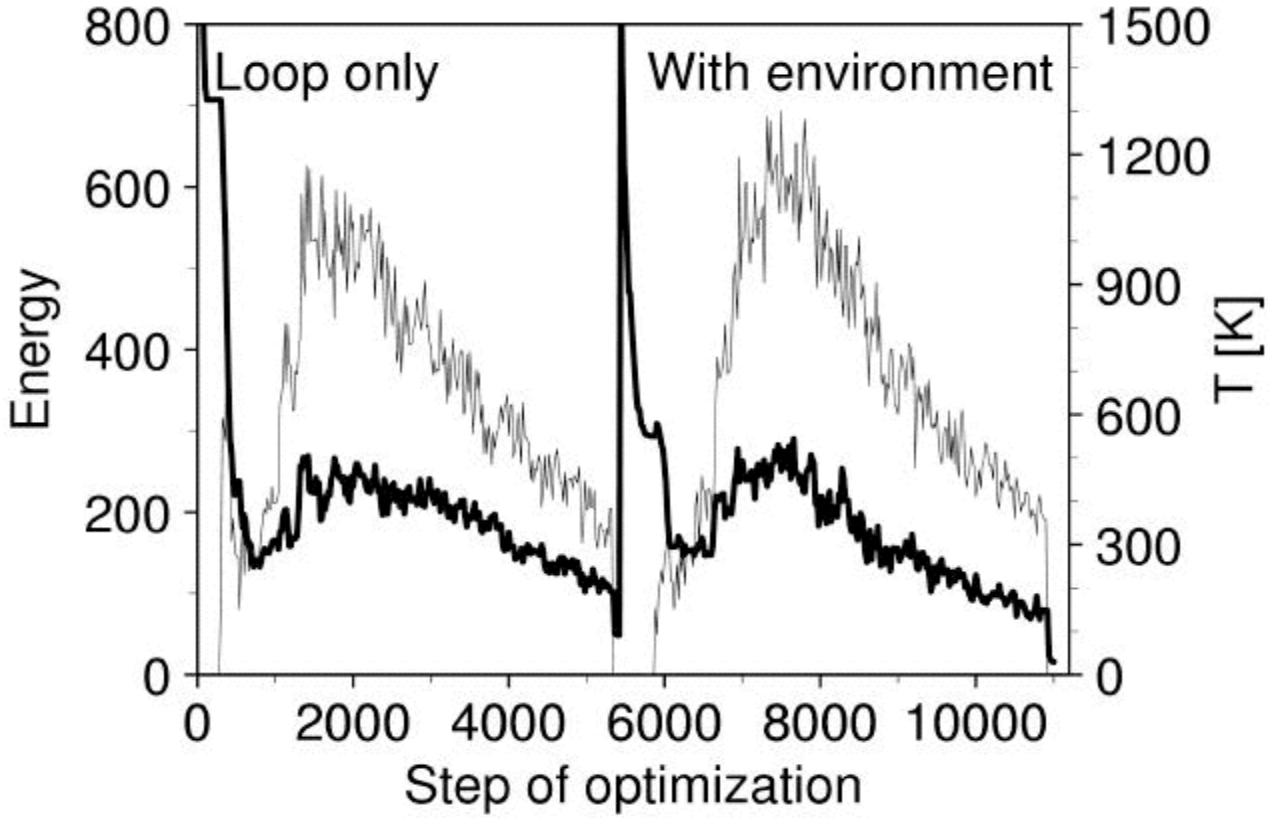


Optimization of Objective Function

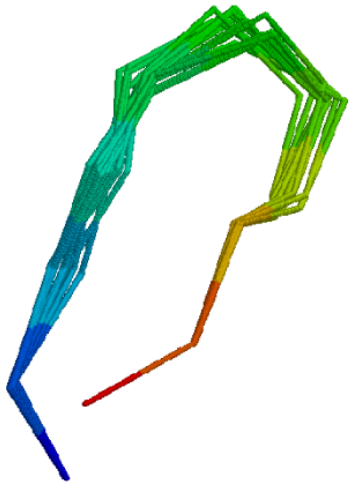
- Test set: 40 randomly selected loops of known structures, for each length from 1 to 14 residues.
- Starting conformation: Loop atoms were spaced evenly on a line spanning the two anchor regions, then randomized by $\pm$ 5 Å.
- To simulate real comparative modeling situations, performance of the loop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

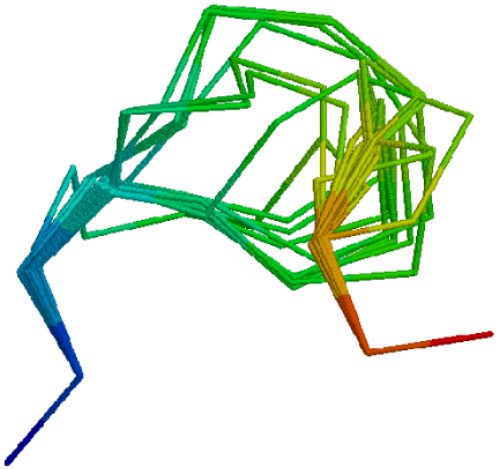


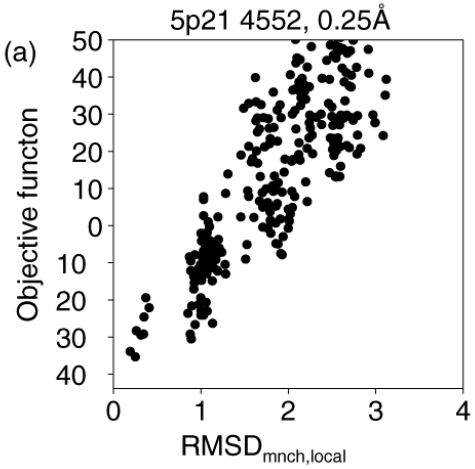
Optimization of Objective Function

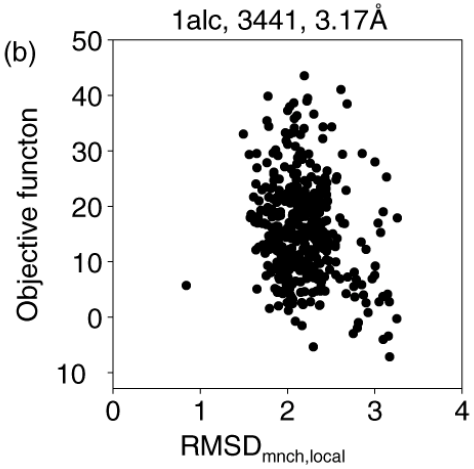


Calculating an Ensemble of Loop Models

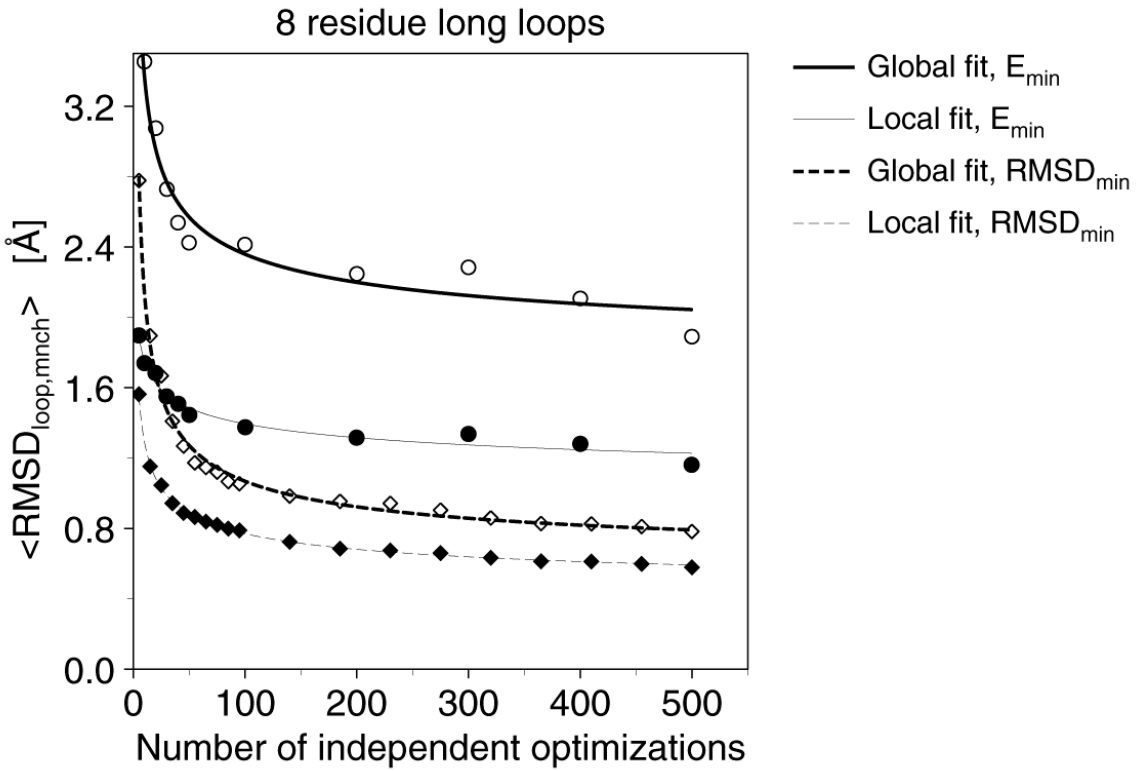




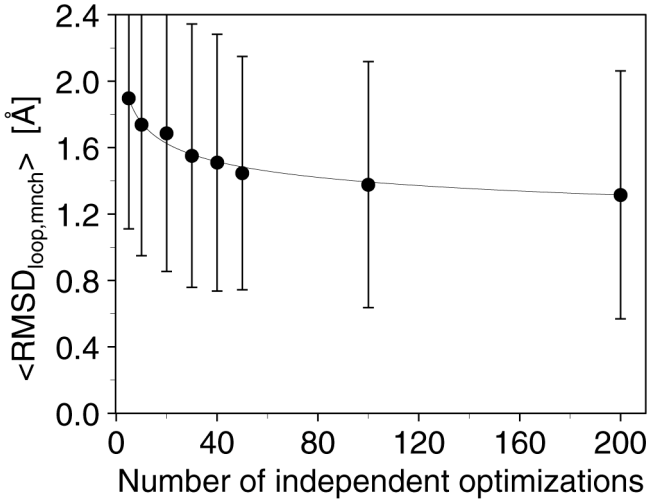


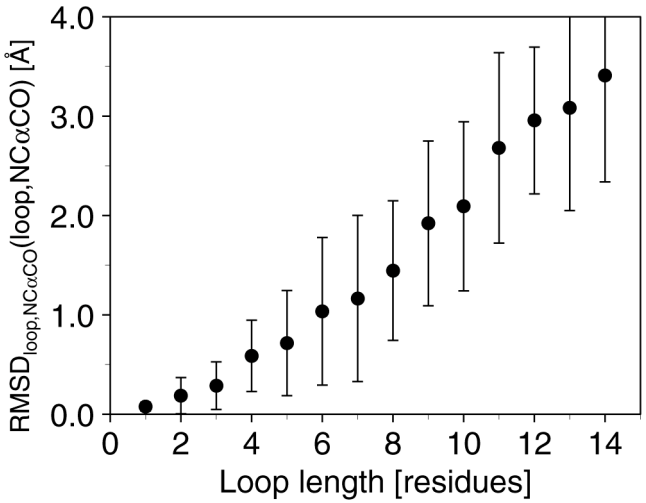


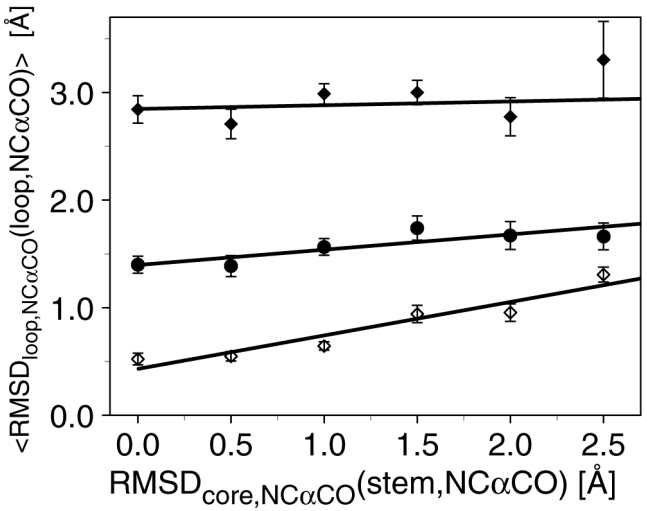
Accuracy of loop models as a function of amount of optimization



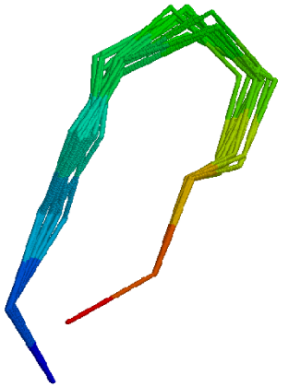
Accuracy of loop models

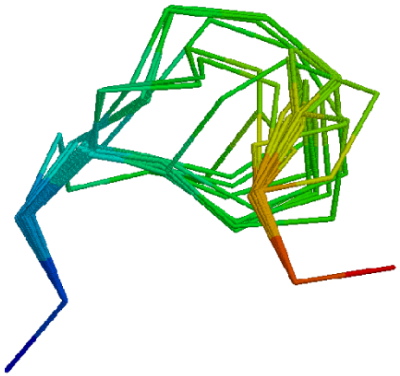


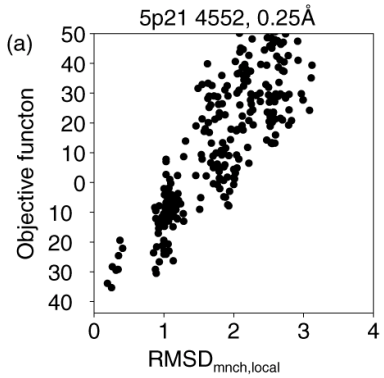


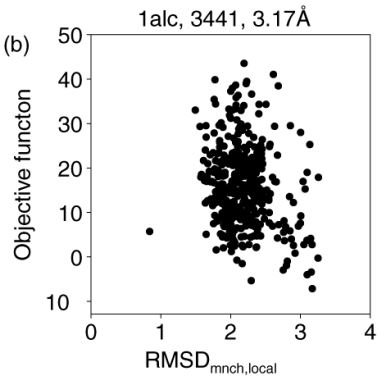


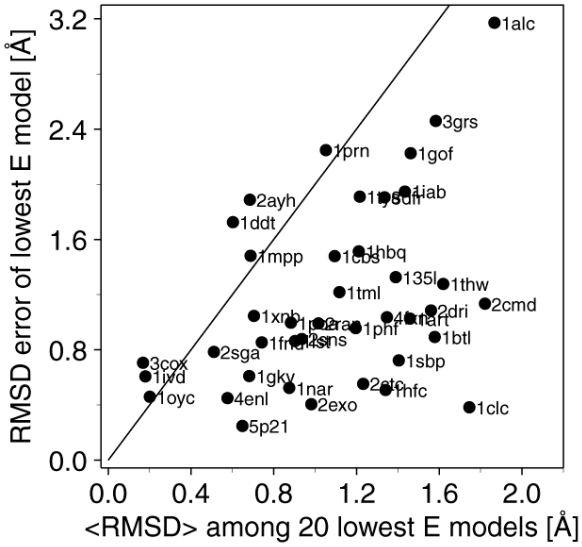
Assessing Accuracy of Loop Models



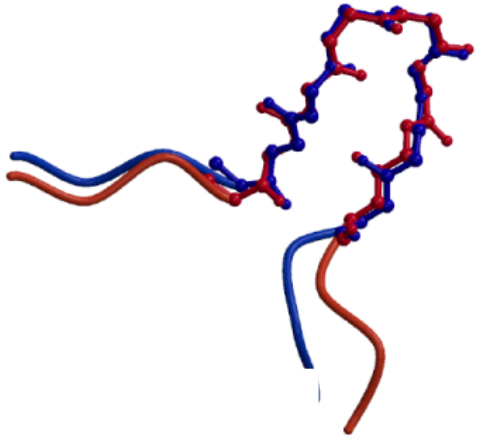








Accuracy of Loop Modeling



RMSD=0.6Å

HIGH ACCURACY (<1Å)

50% (30%) of

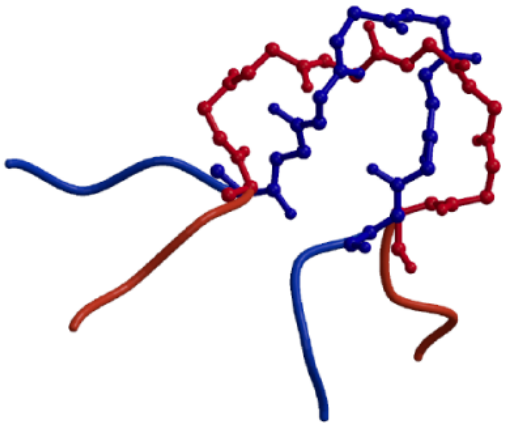
8-residue loops

RMSD=1.1Å

MEDIUM ACCURACY (<2Å)

40% (48%) of

8-residue loops



RMSD=2.8Å

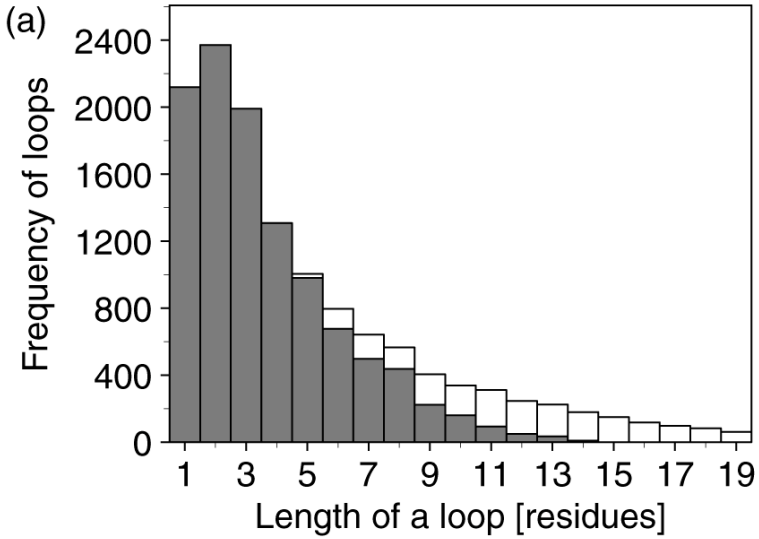
LOW ACCURACY (>2Å)

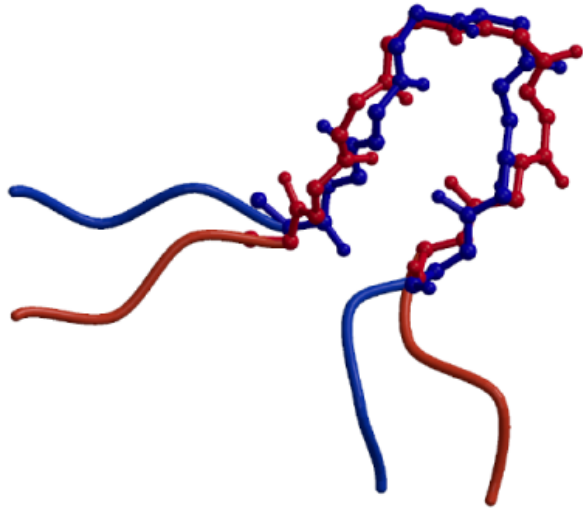
10% (22%) of

8-residue loops

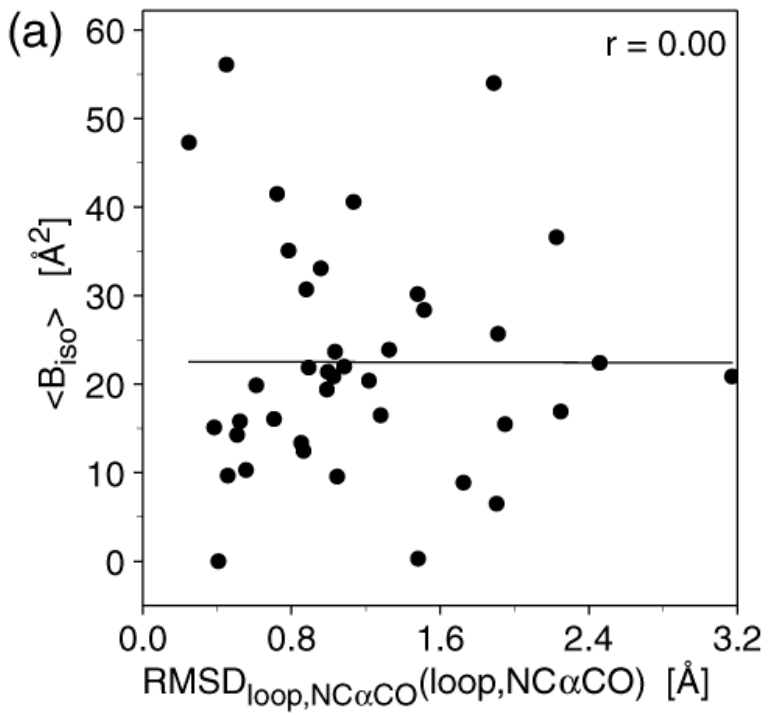
A. Fiser, R. Do & A. Šali, Prot. Sci. 9, 1753, 2000.

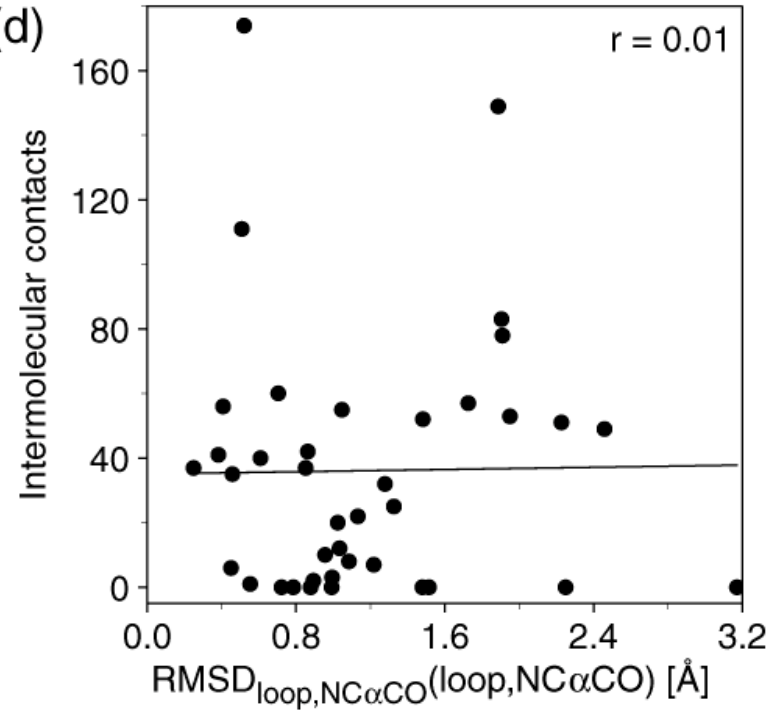
Fraction of Loops Modeled With at Least Medium Accuracy



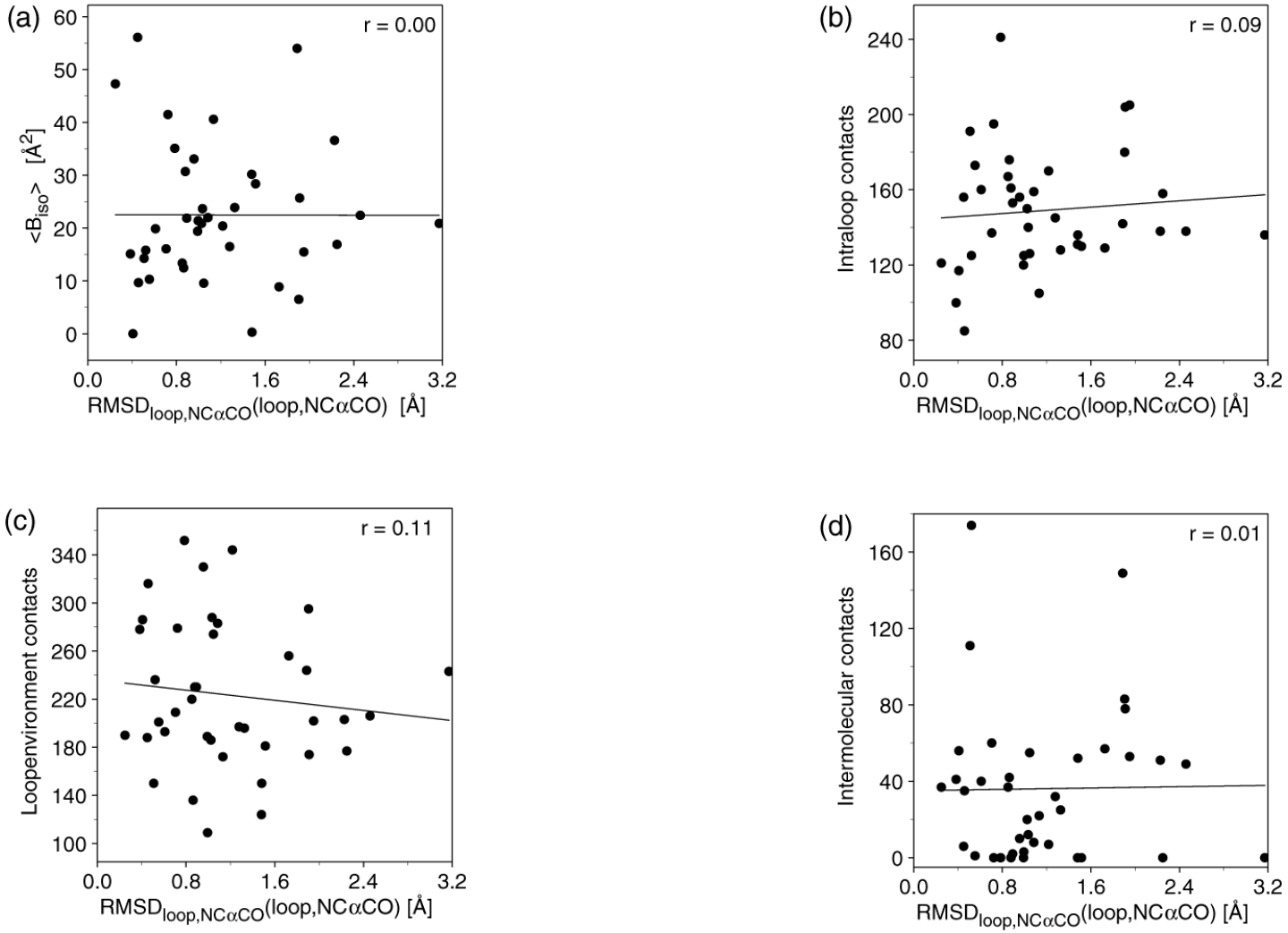


Accuracy of Loop Modeling as a Function of Loop Properties

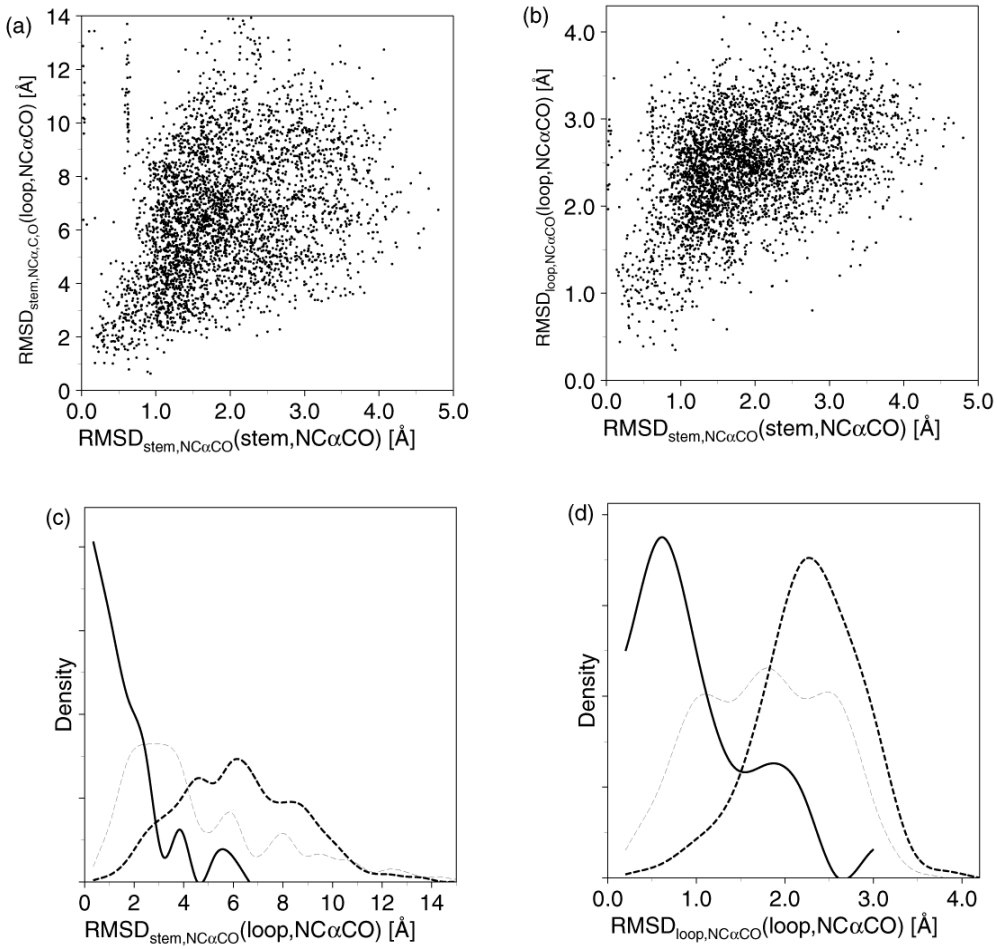




Accuracy of Loop Modeling as a Function of Loop Properties

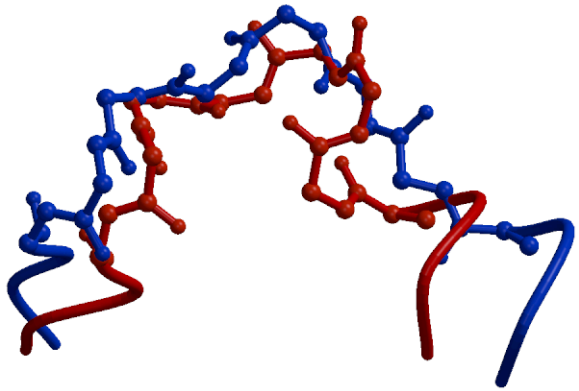


Comparison of the Loop Modeling Errors With Reference RMSD Distributions



Problems in Practical Loop Modeling

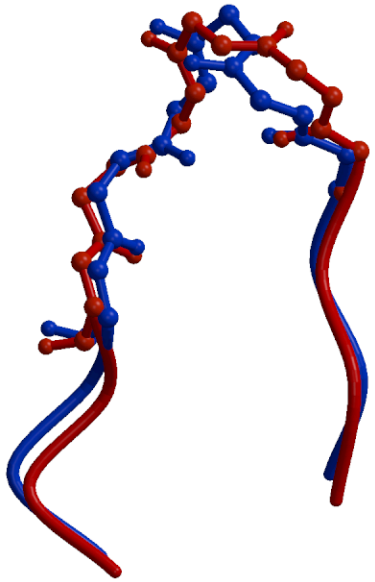
1. Decide which regions to model as loops.
2. Correct alignment of anchor regions & environment.
3. Modeling of a loop.



T0058: 80-85

$RMSD_{mnch}$ loop = 1.09 Å

$RMSD_{mnch}$ anchors = 0.29 Å



T0076: 46-53

$RMSD_{mnch}$ loop = 1.37 Å

$RMSD_{mnch}$ 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[i], Andrej Sali[i], Mark D. Yandell[¶], Nikola P. Pavletich[§], Michael W. Young[3], and Arnold J. Levine*,**

PNAS 97, 7301, 2000.

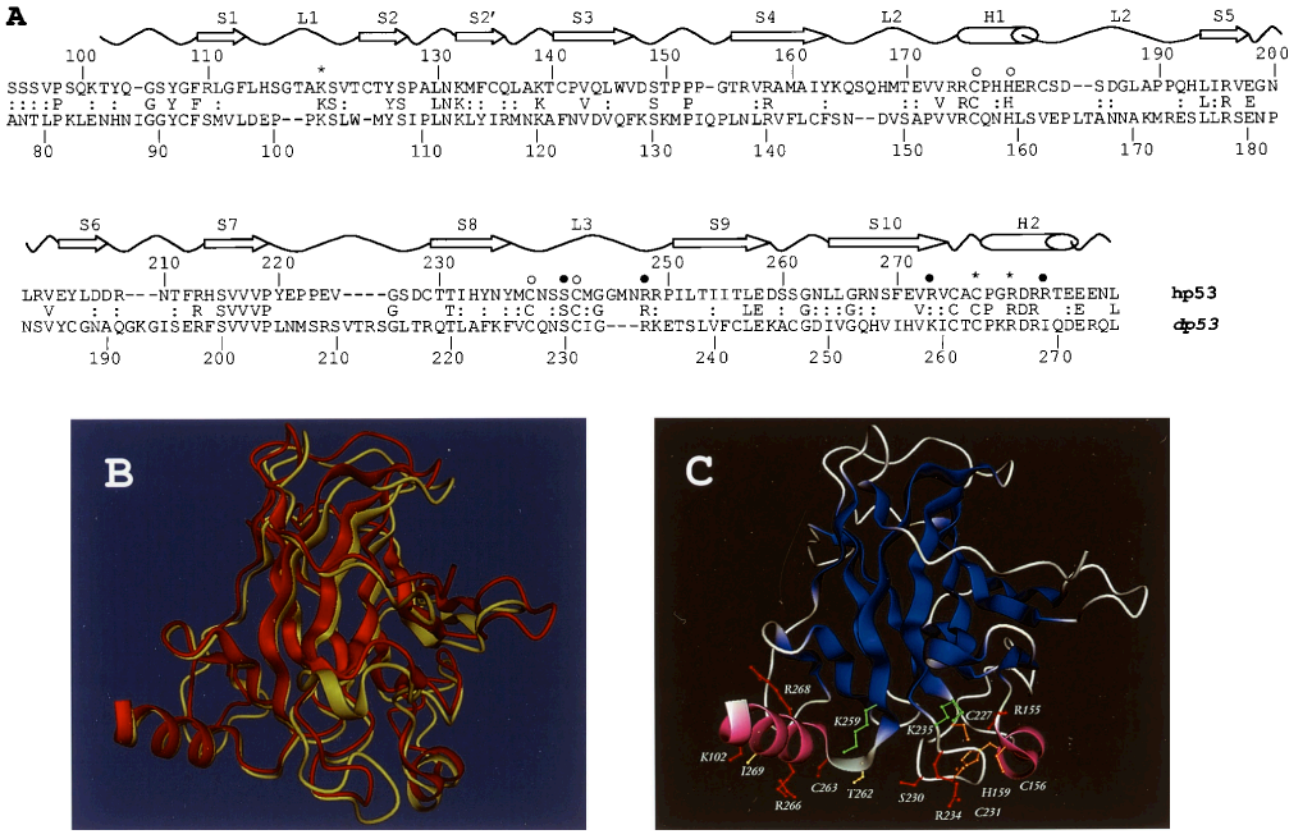


Fig. 1. (*A*) Sequence and structure comparison of the *Drosophila* and human p53 DNA-binding domains. Sequence alignment of the dp53 and hp53 DNA-binding domains as produced by PSI-BLAST. The secondary structure elements of hp53 are shown above (S, *b*-strand; L, loop; H, *a*-helix). Residues involved 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:yywww.biozentrum.unibas.chy ; x-rayydino).

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.**

S. Jin, N. Mirkovic, A. Sali, A. Levine *et al*. PNAS **97**, 7301, 2000.

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

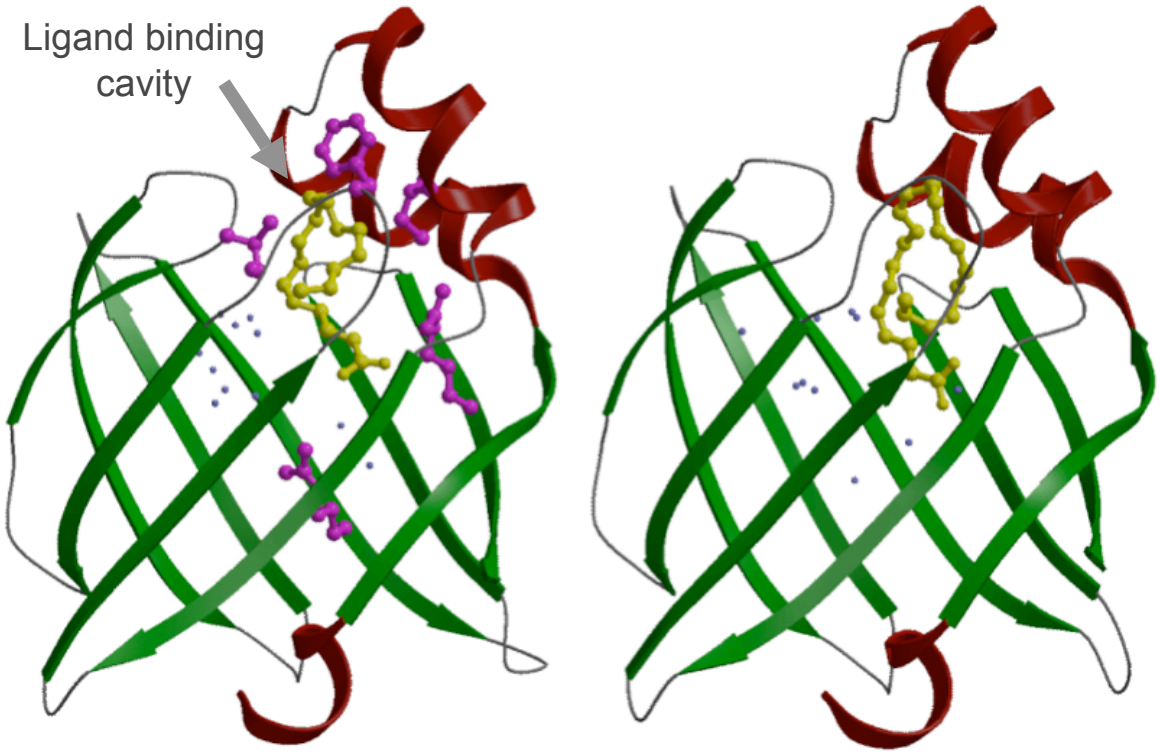
Predicting features of a model that are not present in the template

BLBP/oleic acid

Cavity is **not** filled

BLBP/docosahexaenoic acid

Cavity **is** filled



1. BLBP binds fatty acids.

2. Build a 3D model.

3. Find the fatty acid that fits most snuggly into the ligand binding cavity.

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

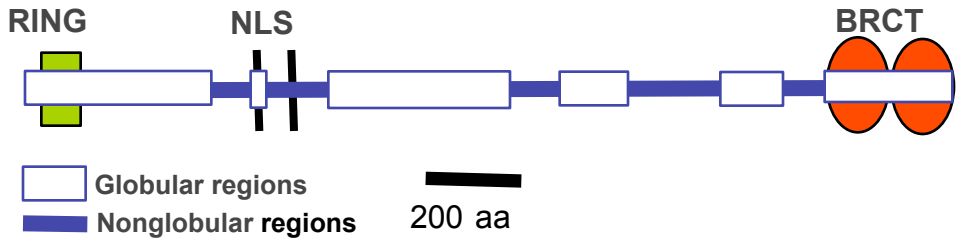
Structural analysis of missense mutations in human BRCA1 BRCT domains

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

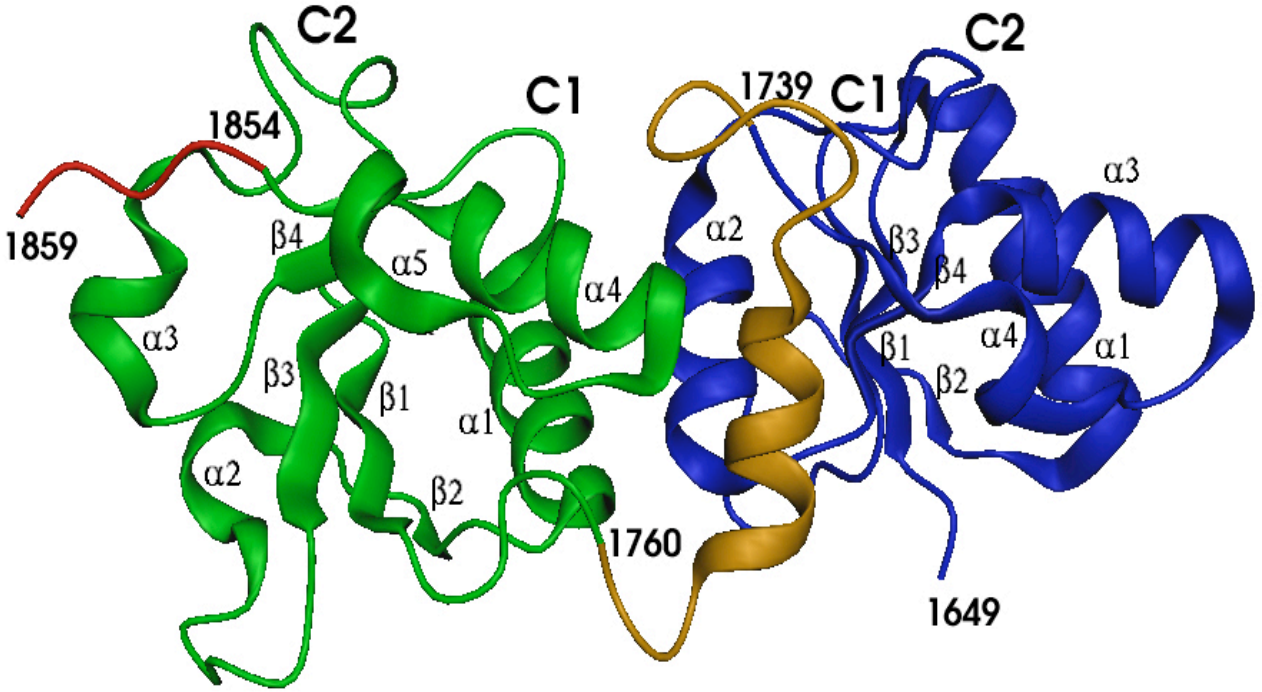
Cancer Research (June 2004). 64:3790-97

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



BRCA1 BRCT repeats, 1jnx



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

CONFIDENTIAL

MYRIAD

BRACAnalysis™

Comprehensive BRCA1-BRCA2 Gene Sequence Analysis Result

Niecee Singer, MS Strang Cancer Prevention Center 428 E 72nd St New York, NY 10021	**SPECIMEN** Specimen Type: Blood Draw Date: n/a Accession Date: Oct 27, 2000 Report Date: Nov 17, 2000	**PATIENT** Name: Date of Birth: Feb 02, 1953 Patient ID: Gender: Female Accession #: 00019998 Requisition #: 56694
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

MYRIAD

BRACAnalysis™

Comprehensive BRCA1-BRCA2 Gene Sequence Analysis Result

Niecee 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 #: 56694
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.

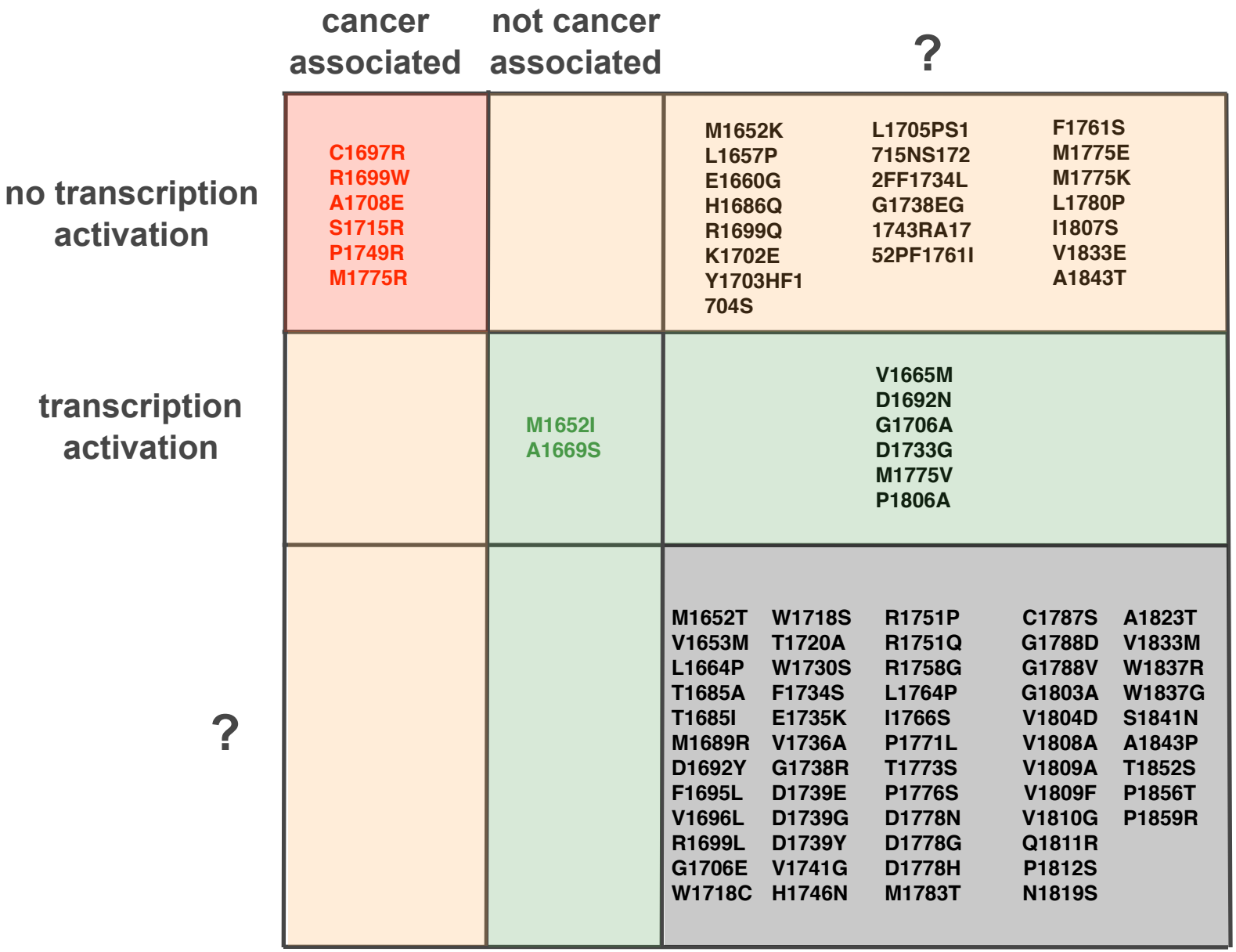
Authorized Signature:

Brian E. Ward, Ph.D.
Laboratory Director

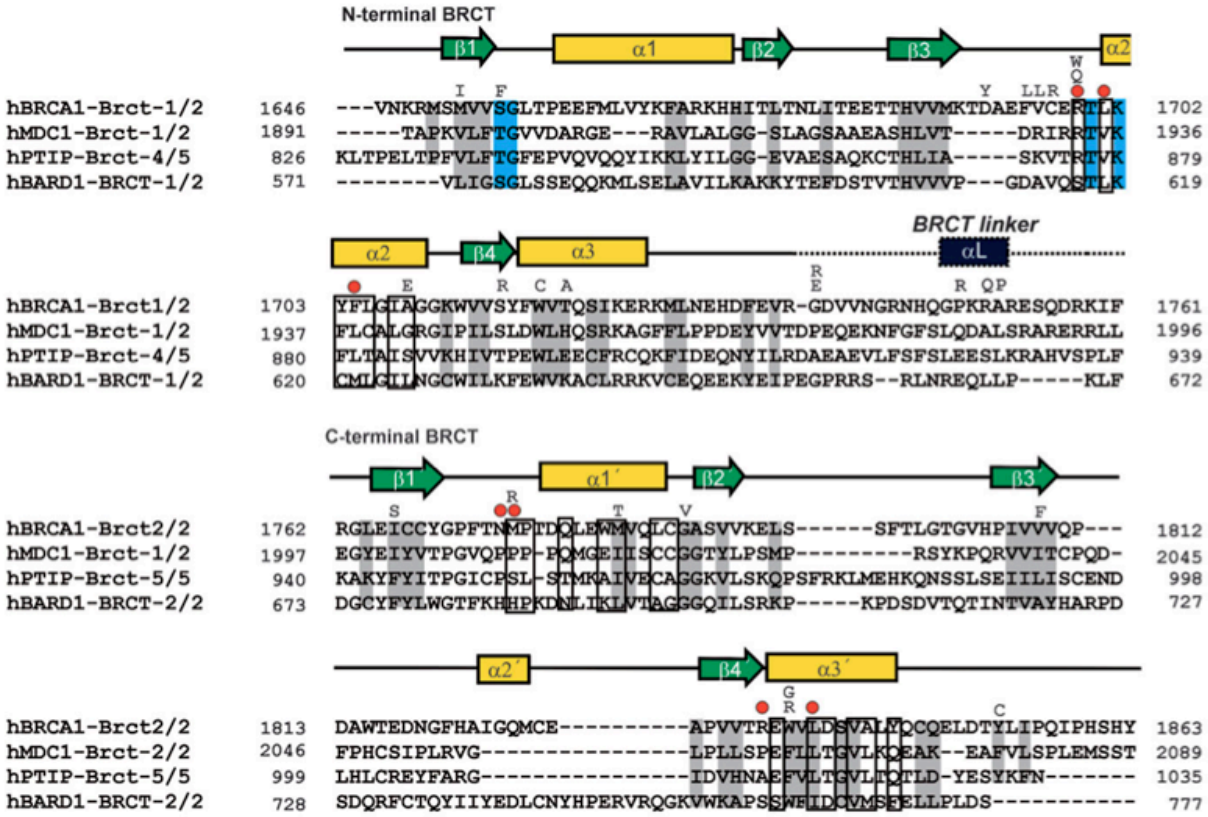
Thomas S. Frank, M.D.
Medical Director

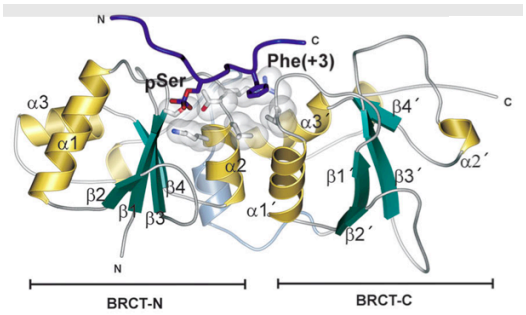
These test results should only be used in conjunction with the patient's clinical history and any previous analysis of appropriate family members. It is strongly recommended that these results be communicated to the patient in a setting that includes appropriate counseling. The accompanying Technical 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



Putative binding site on BRCA1

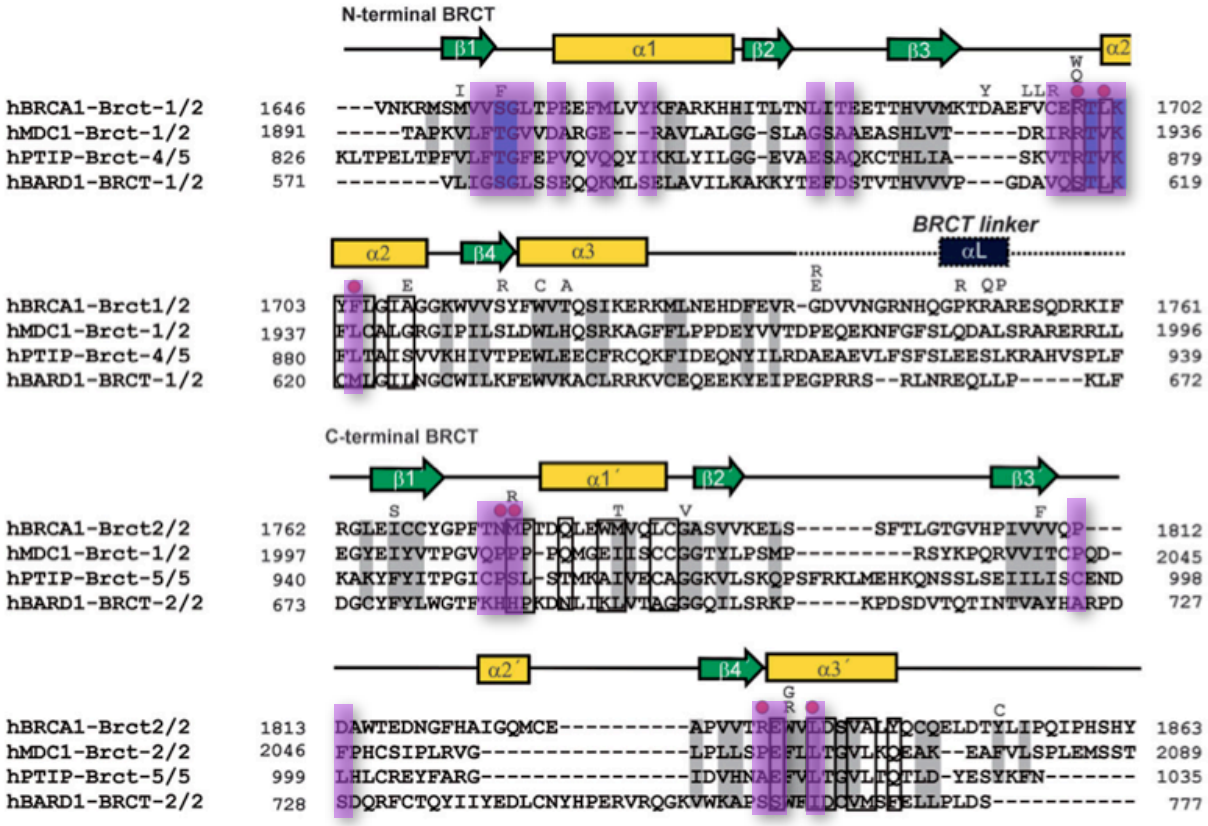


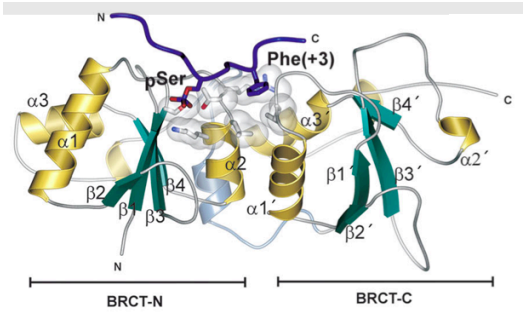


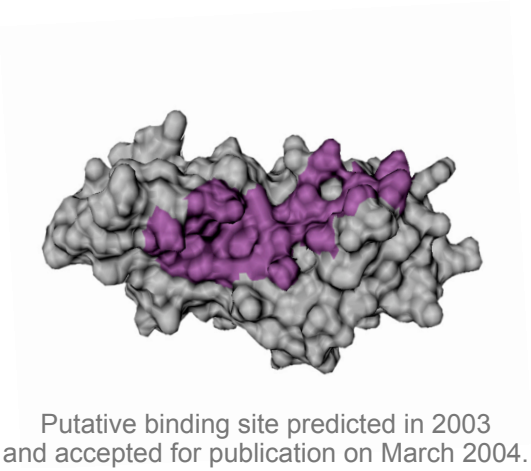
Williams *et al.* 2004 Nature Structure Biology. **June 2004 11**:519

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

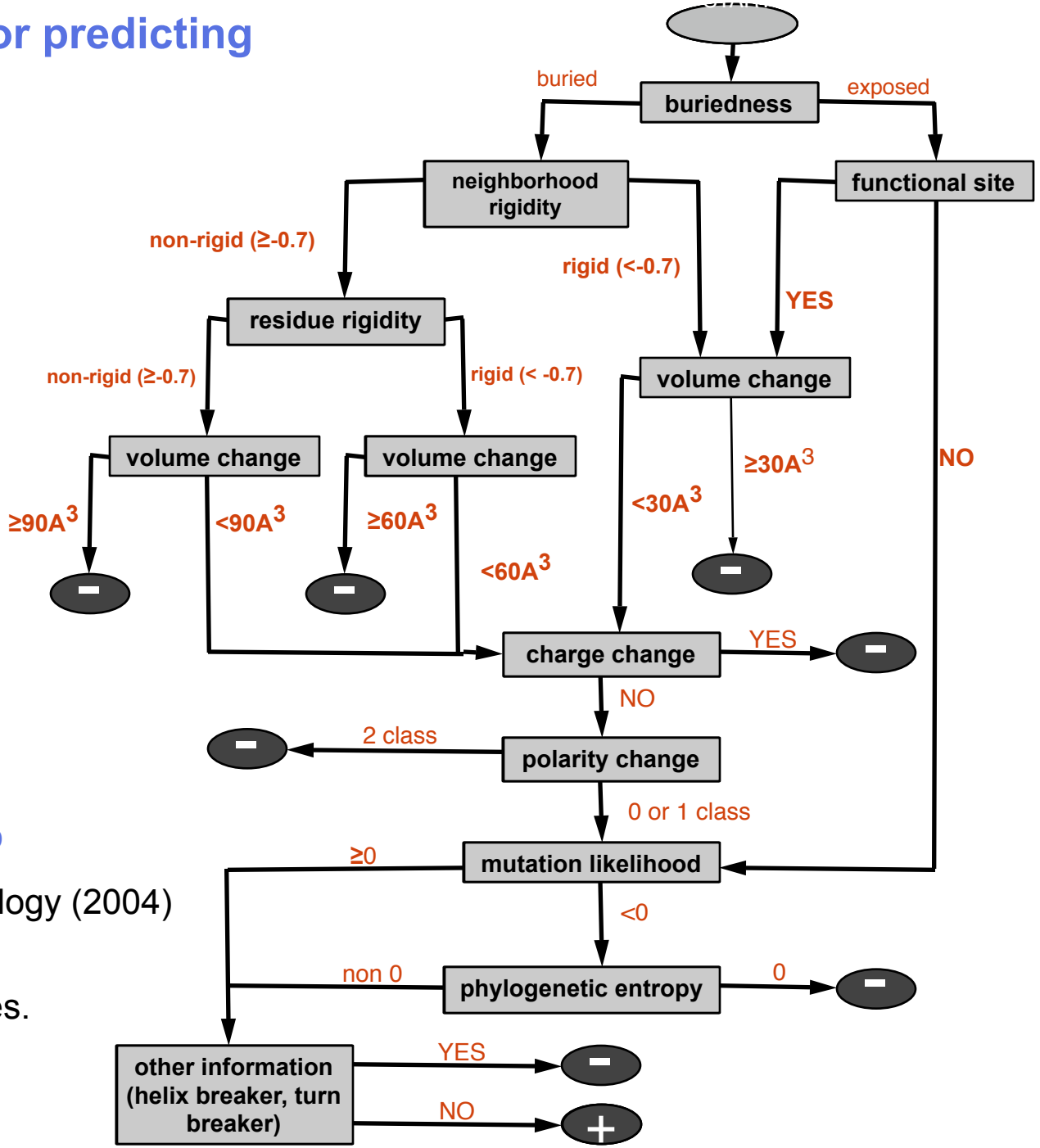
Mirkovic *et al.* 2004 Cancer Research. **June 2004 64**:3790

"Decision" tree for predicting

functional impact

of genetic

variants



http://salilab.org/snpweb

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

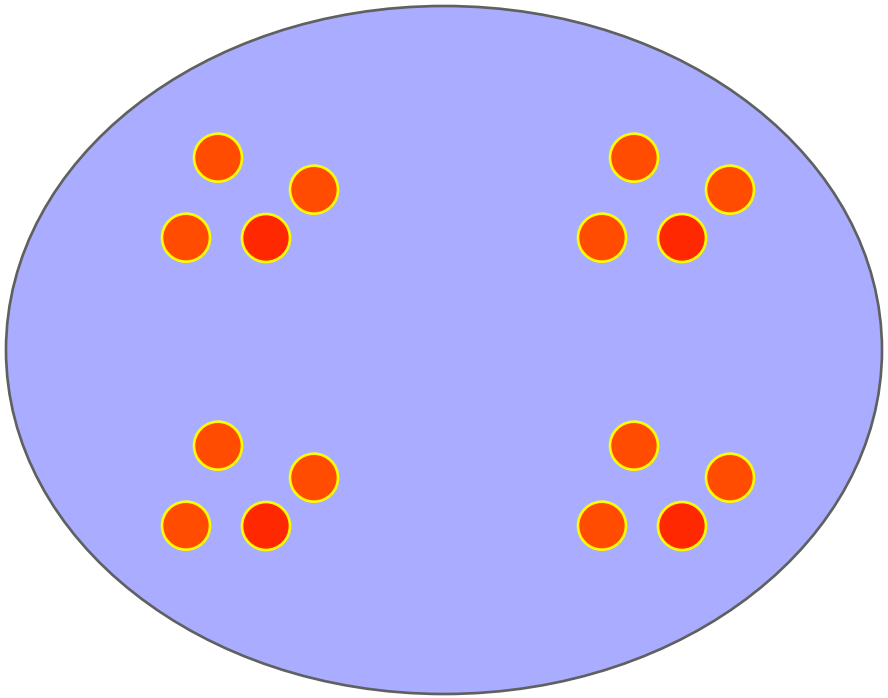
Eswar *et al*. Nucl.Acids Res. **31**, 3375, 2003.

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.

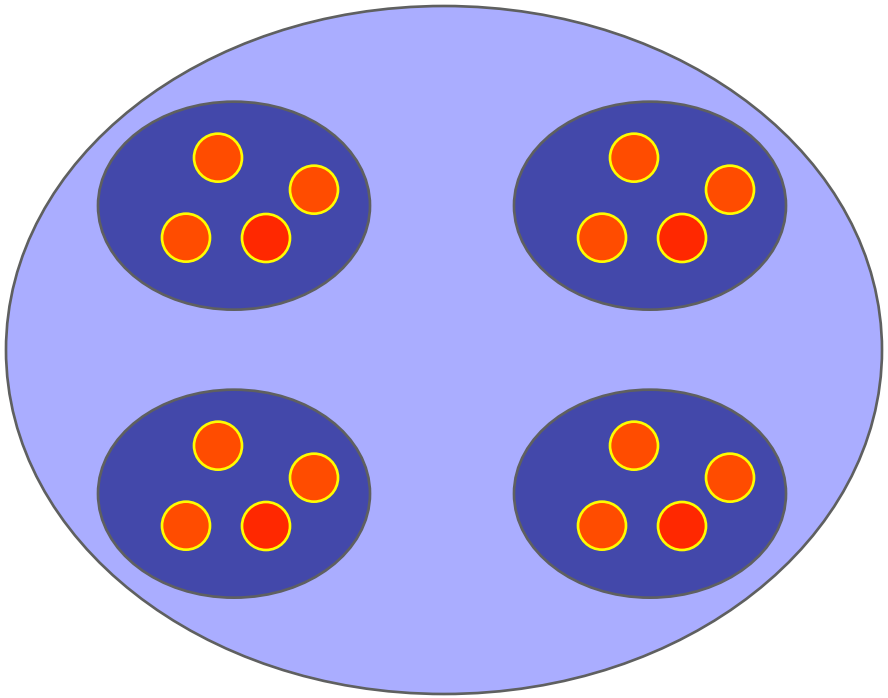
Any one of the members of a family is fine.

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.

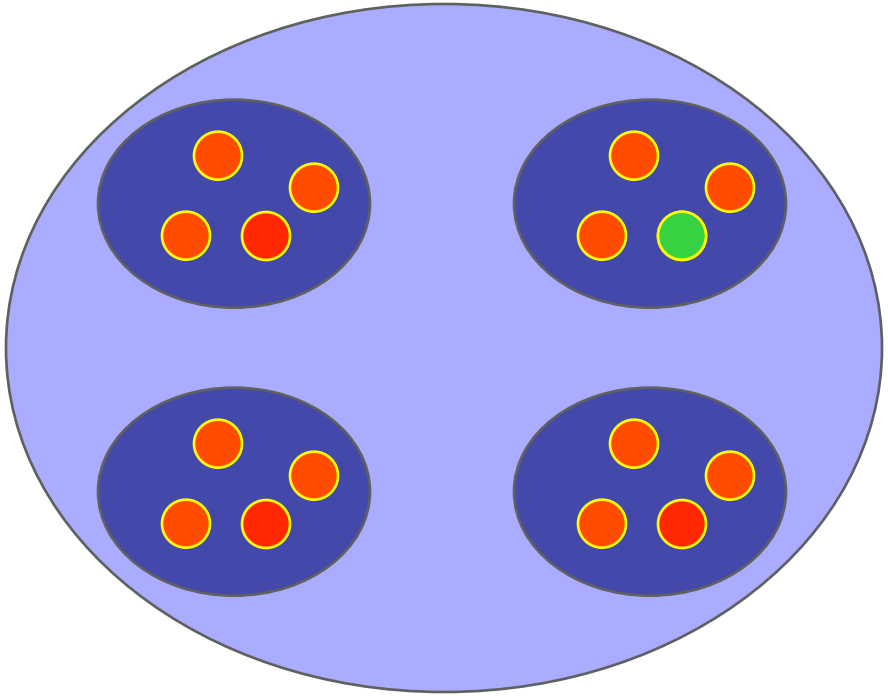
Any one of the members of a family is fine.

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.

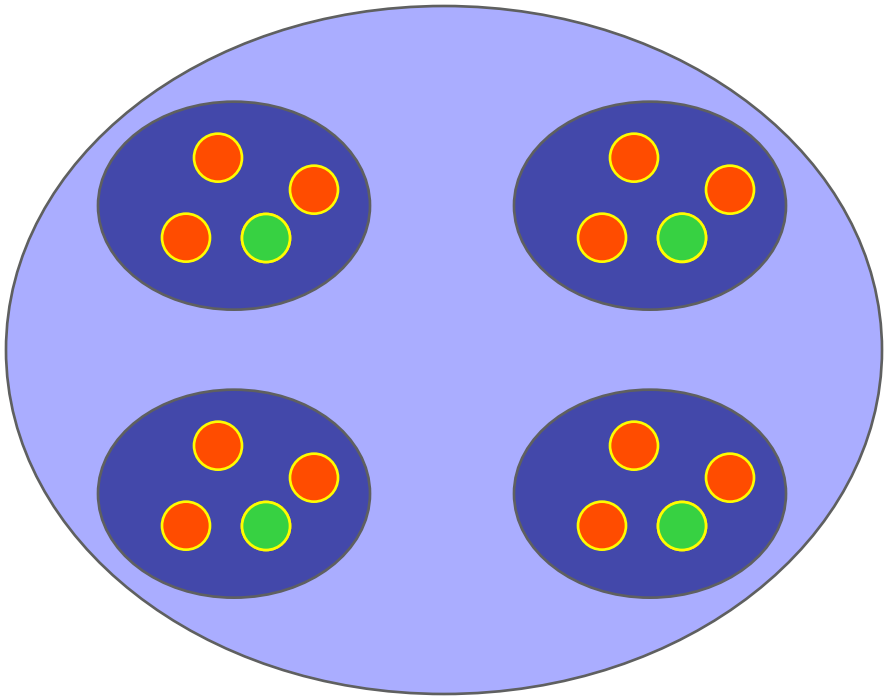
Any one of the members of a family is fine.

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.

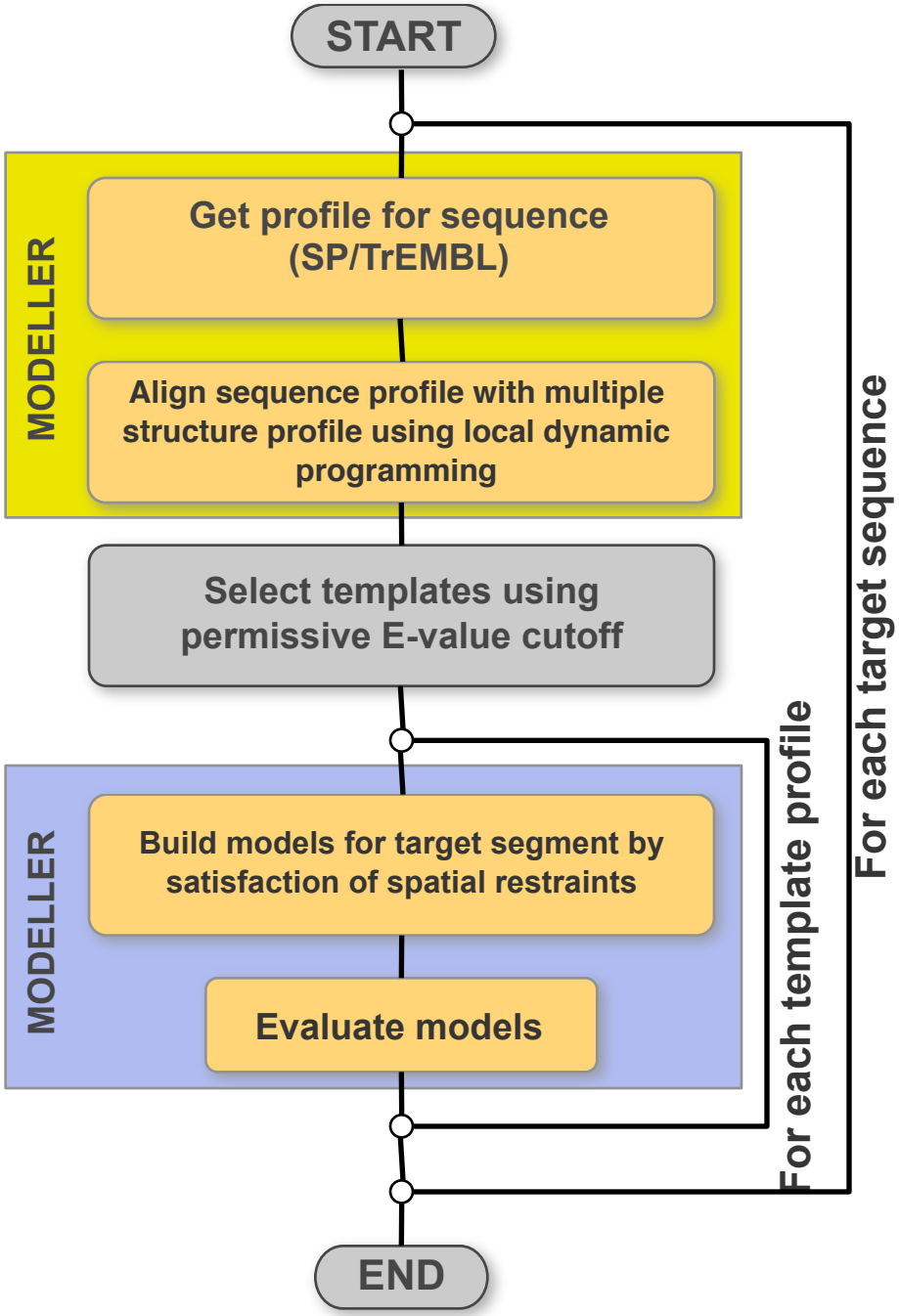
Characterize most protein sequences based on related known structures.



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_nysgxrc.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				

Comparative modeling of the TrEMBL database

Unique sequences processed: 1,182,126

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

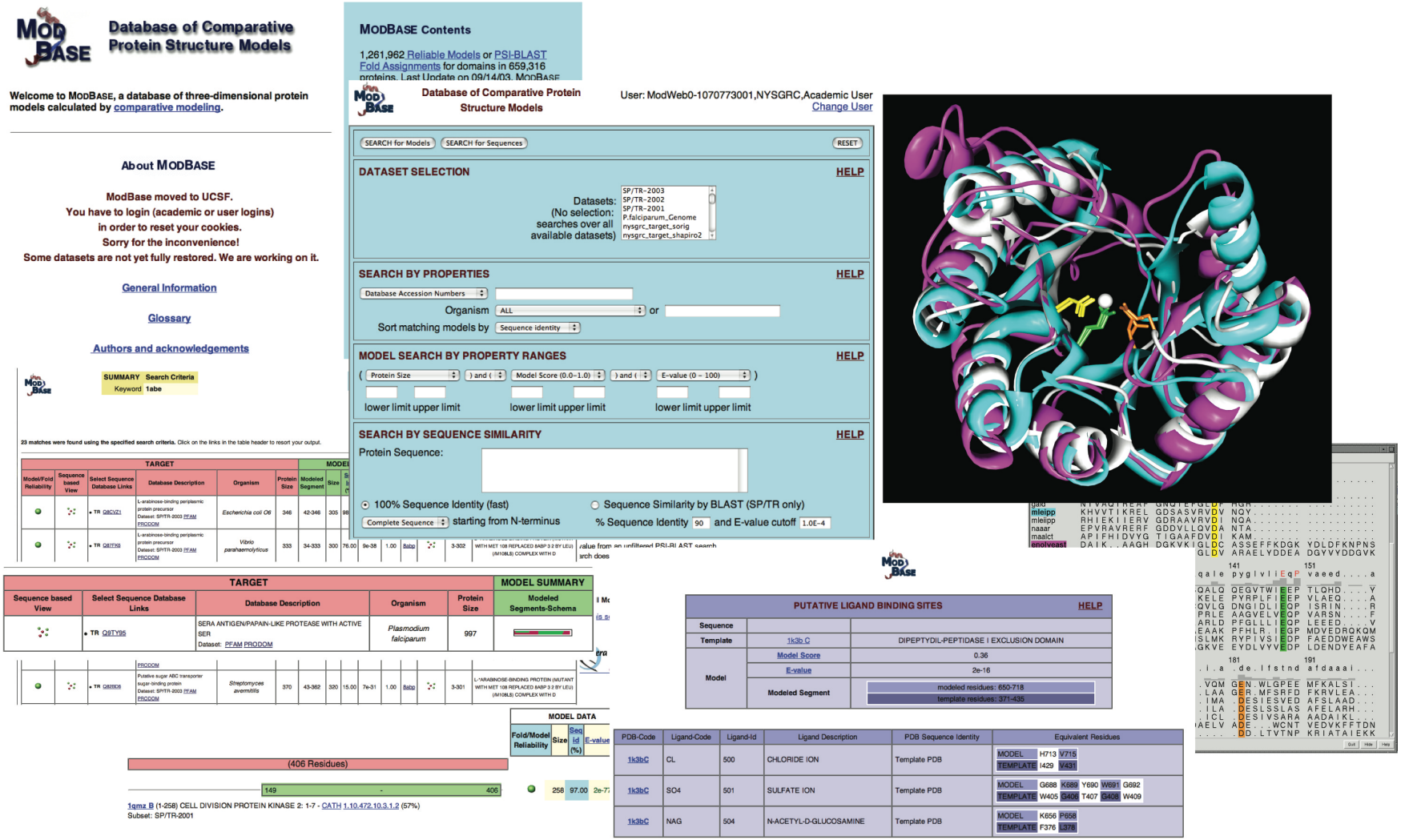
70% of models based on <30% sequence identity to template.

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

9/15/03 ~4 weeks on 660 Pentium CPUs

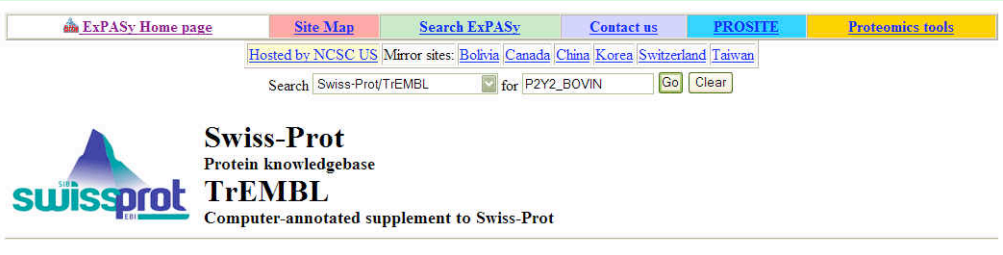
http://salilab.org/modbase

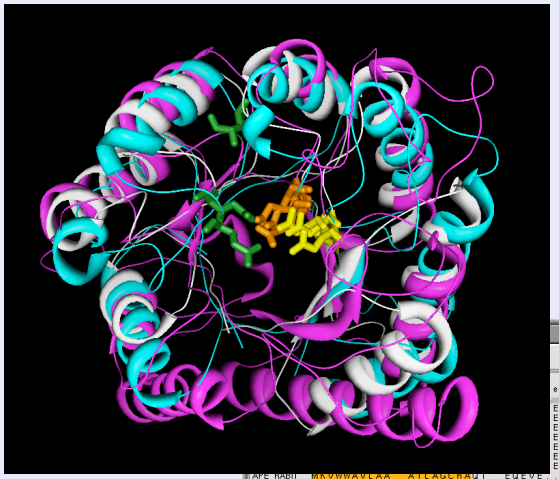
Pieper *et al.*, Nucl. Acids Res. 2004.

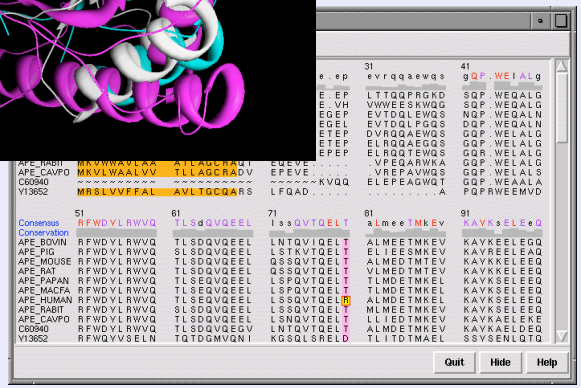


05/10/2004

Major bidirectional resources involving ModBase

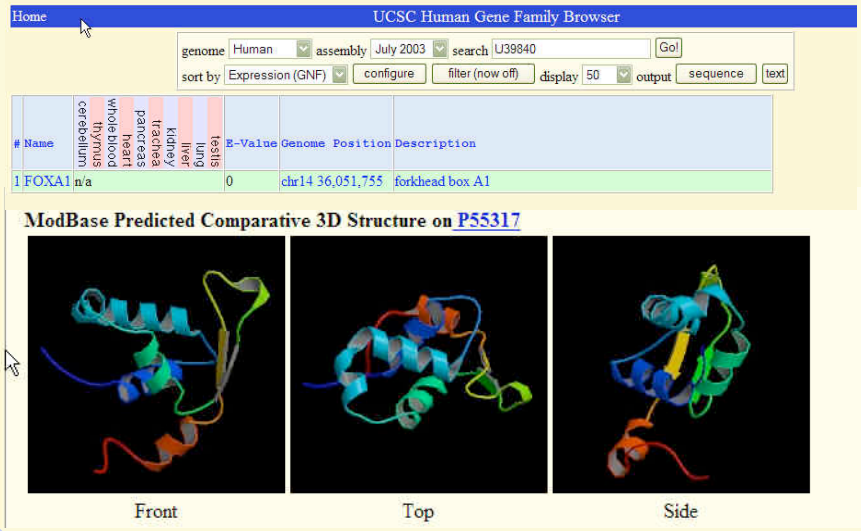




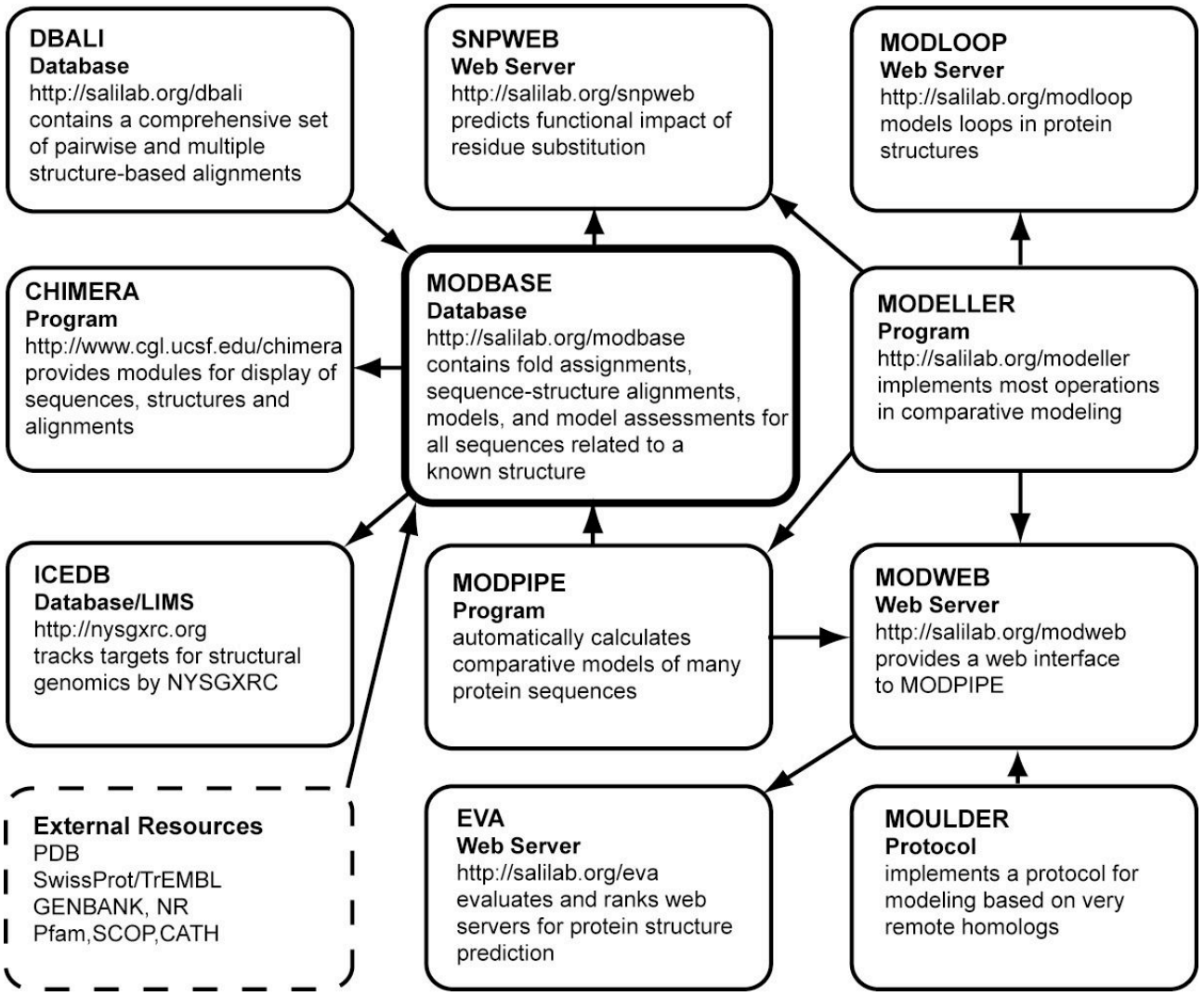


http://www.cgl.ucsf.edu/chimera/
Daniel Greenblatt, Conrad C. Huang,
Thomas E. Ferrin

UCSC Human Gene Family Browser



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 (PROSAII, 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.

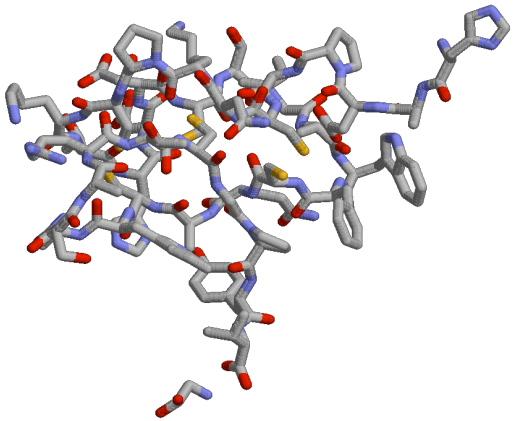
Finkelstein et al. (1995) *Proteins* **23**, pp142

Representation

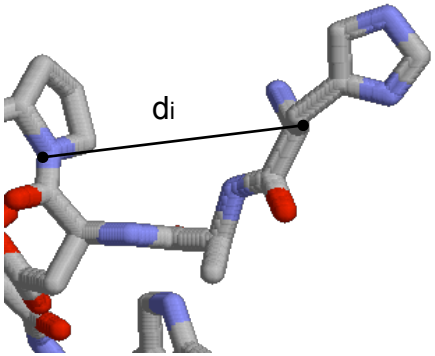
Sequence/Structures

```
>gi42541361
MDIRSVSSLRGLLCLPPSWPRR
```

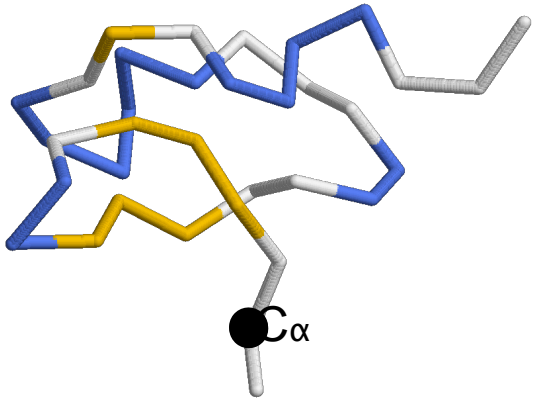
Primary sequence



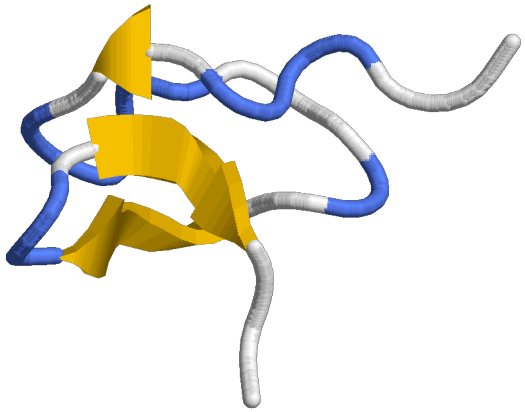
All atoms and coordinates



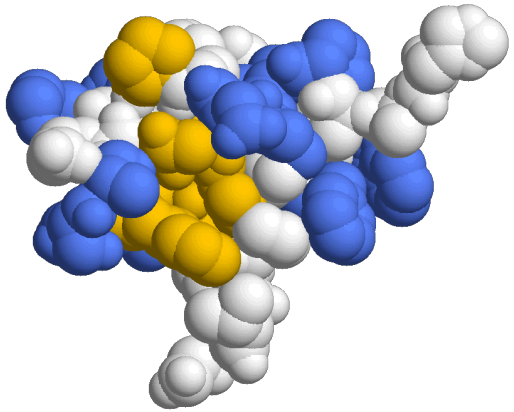
Distance space



Reduced atoms representation



Secondary Structure

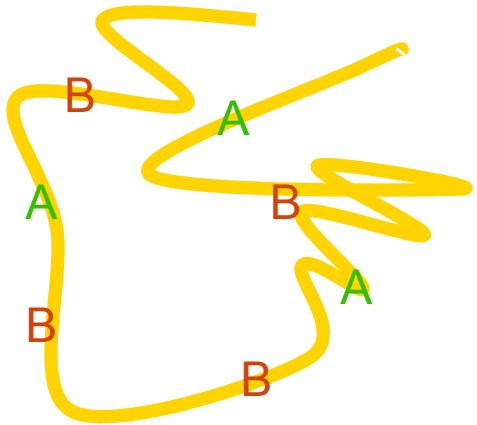


Accessible surface

Statistical Potential… inspiration

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

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



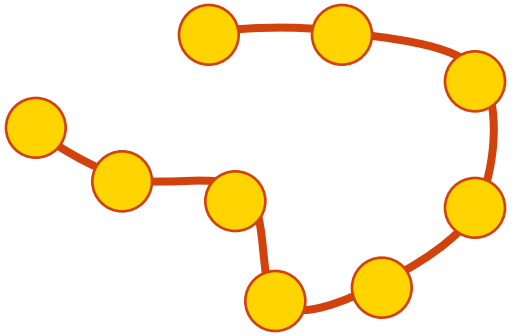
From statistical physics, we know that energy difference between two states (ΔE) and the ratio of their occupancies (N_1:N_2) are related [9]:

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

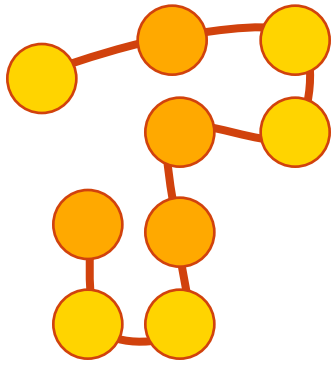
in which T is the absolute temperature and k is the Boltzmann's constant. As we are interested in an interaction energy between two amino acid side chains, it would seem natural to define N_1 as the number of interactions between these two residues types in a group of real protein structures, a number 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

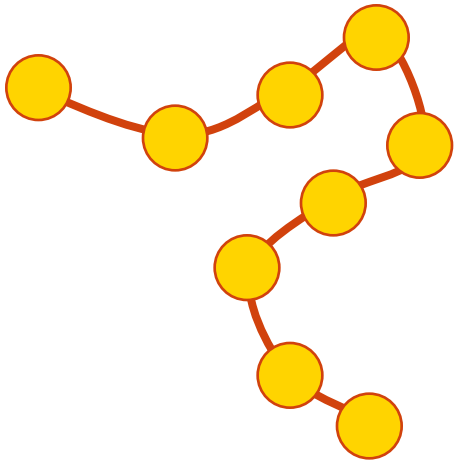
Statistical Potential… interaction types



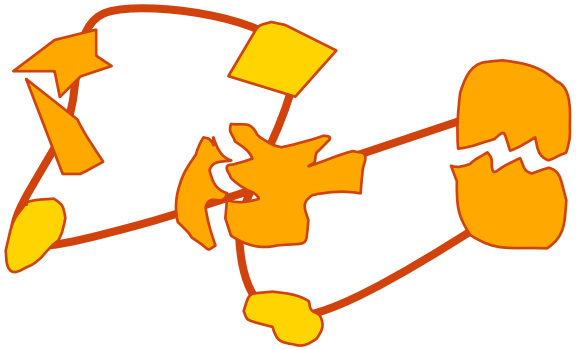
Neutral interactions



Hydrophobic interactions



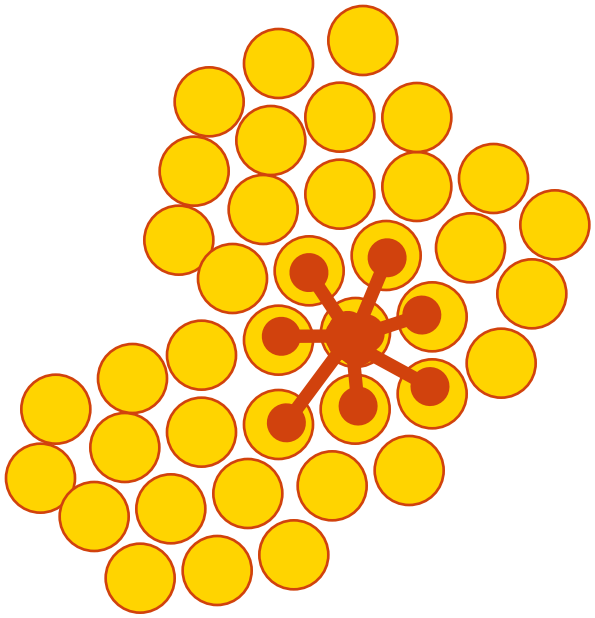
Compact interactions

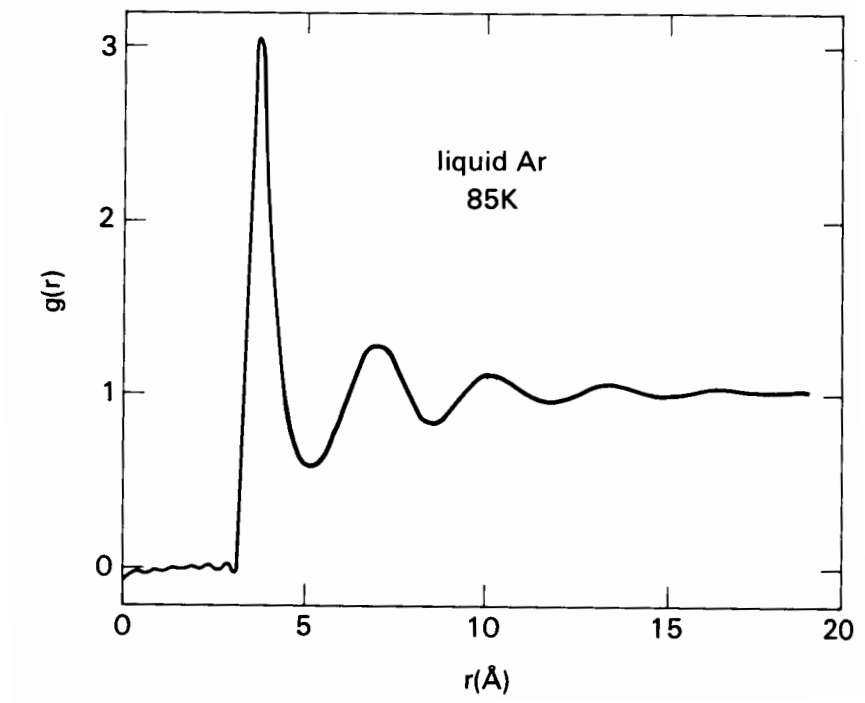


Specific interactions

A. Godzik, (1996) *Structure* **15** pp363

Statistical Potential… reference state

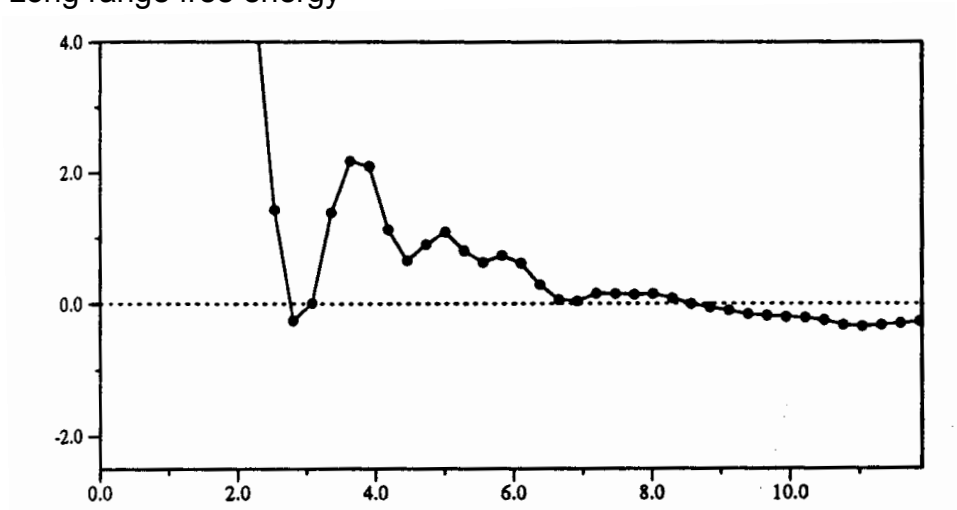




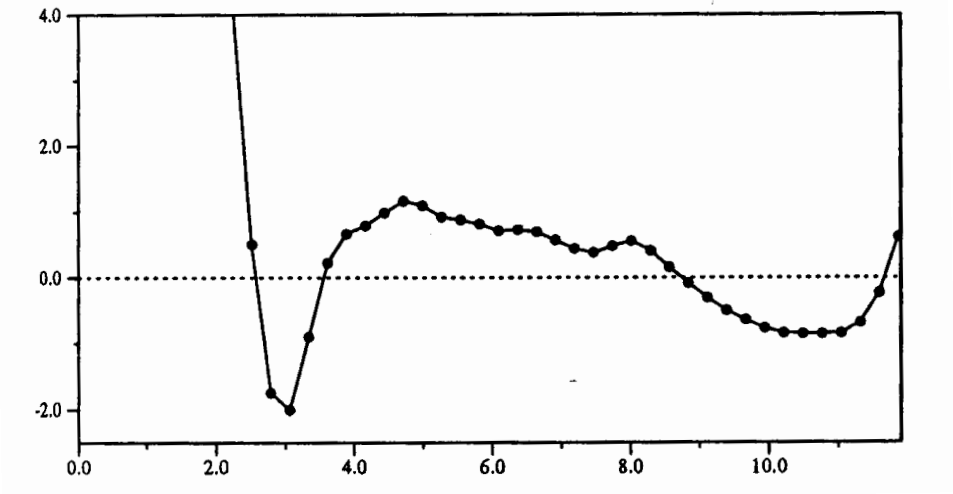
Theory of simple liquids 2nd edition JP Hansen and IR McDonald, Academic Press.

Statistical Potential… Hydrogen Bonds

Long range free energy



Short range free energy



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

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

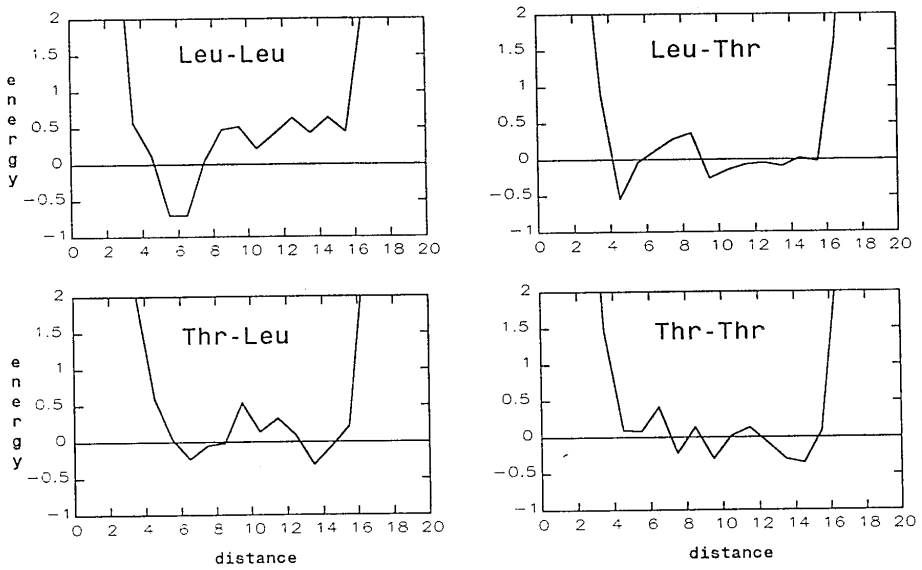
$$g_{NO}(r) = \frac{\rho_{NO}(r)}{\rho^2}$$

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

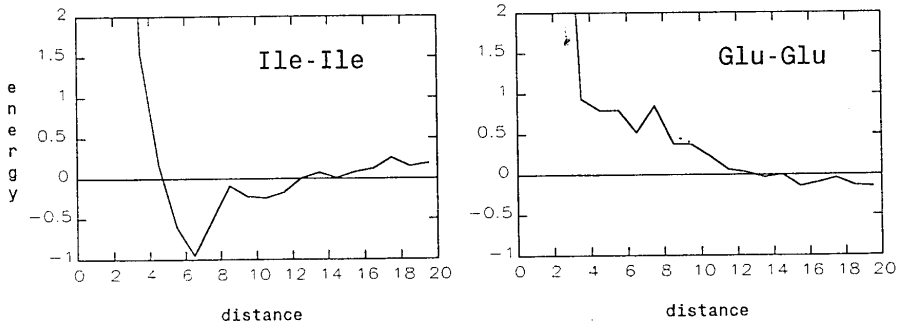
Sippl (1996). *JMB* **260** pp644

Statistical Potential… Distance Potentials

Long range free energy



Short range free energy



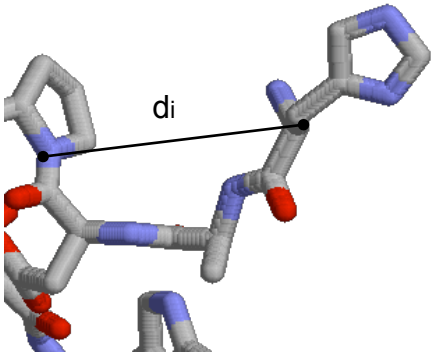
Sippl (1993). *JCAM* **7** pp473

Scoring

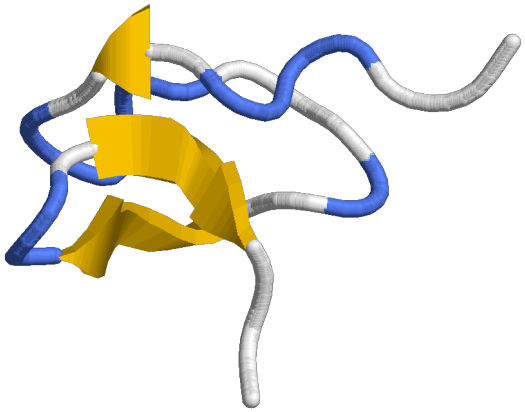
Raw scores of an alignment

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

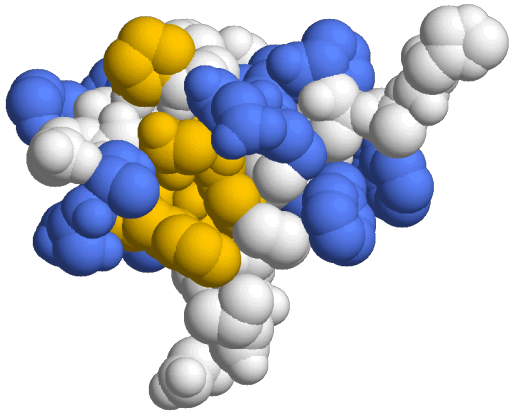
Aminoacid substitutions



Distance space



Secondary Structure (H,B,C)

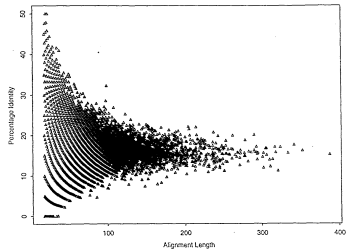


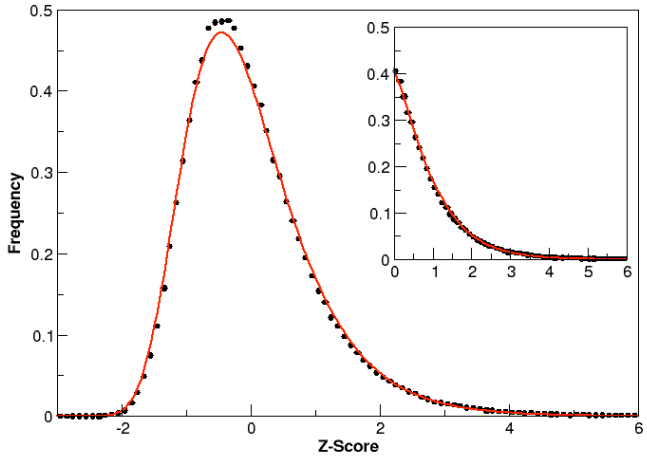
Accessible surface (B,A [%])

Significance of an alignment (score)

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

Empirical





Sometimes approximated by Z-score (normal distribution).

Analytic

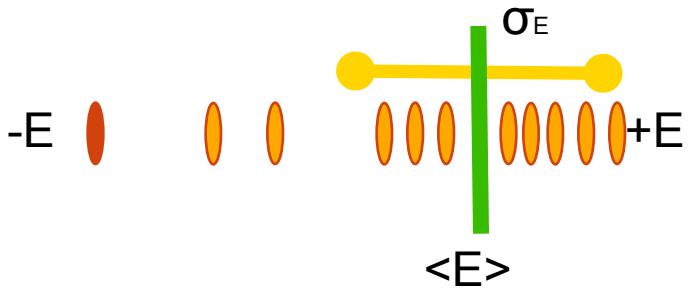
$$P(s) = e^{-\lambda (s-\mu)}$$

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

Karlin and Altschul, 1990 *PNAS* **87**, pp2264

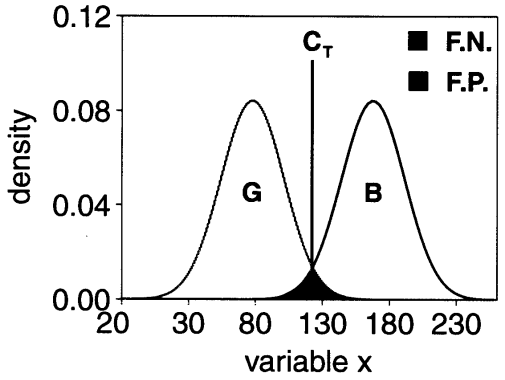
Significance of an alignment (score)

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



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

Evaluating the assessment

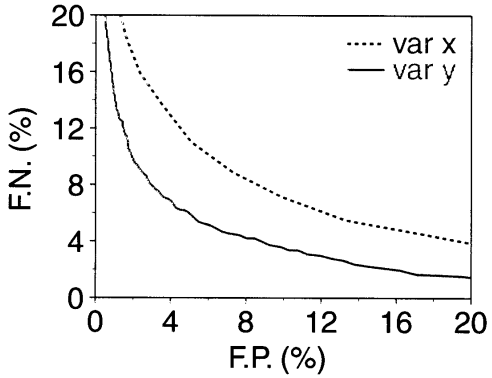


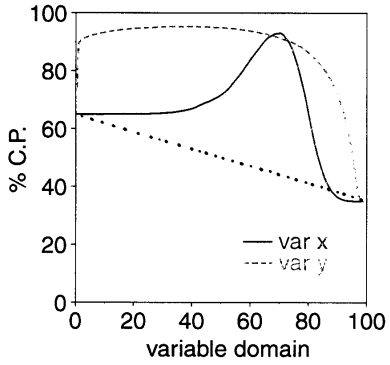
	is GOOD	*is BAD*
predicted as GOOD	a	b
predicted as BAD	c	d

$$F.P. = \frac{b}{b+d} = 1 - specificity$$

$$F.N. = \frac{c}{a+c} = 1 - sensitivity$$

$$C.P. = \frac{a+d}{a+b+c+d}$$





Evaluating the assessment

1

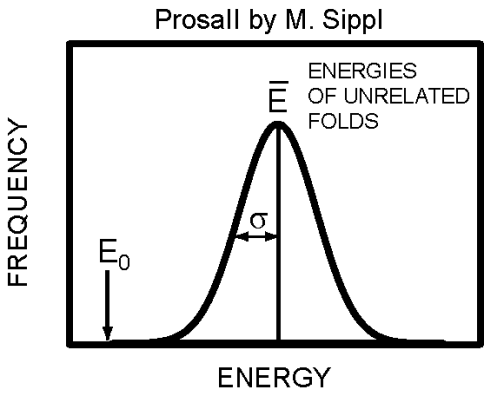
3900 GOOD MODELS

Models based on correct templates and approximately correct alignments

6000 BAD MODELS

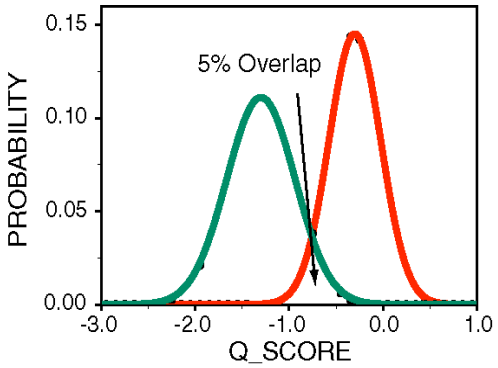
Models based on incorrect templates or mostly incorrect alignments

2



3

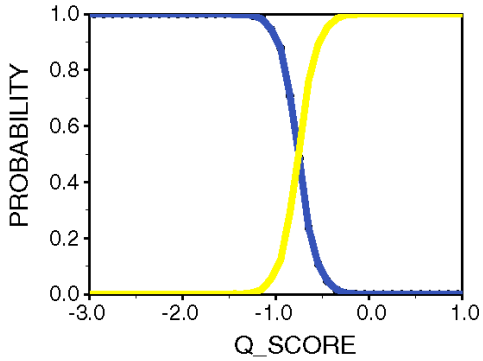
p(Q-SCORE/GOOD)
p(Q-SCORE/BAD)



4

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

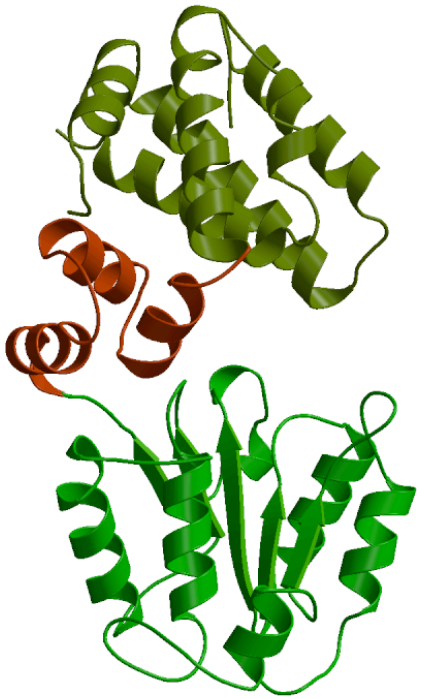
p(BAD/Q-SCORE) = 1 -



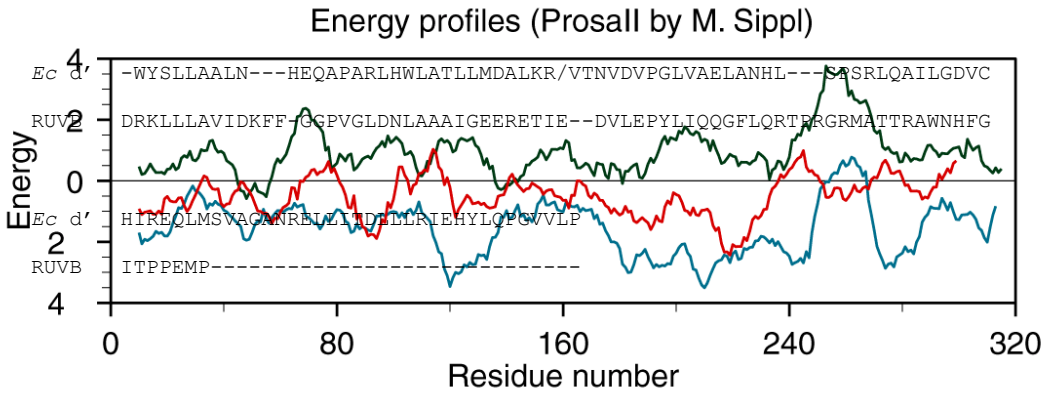
R. Sánchez & A. Šali, (1998) *Proc. Natl. Acad. Sci. USA* **95**, pp13597

Applications of methods for model assessment

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



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

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

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

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



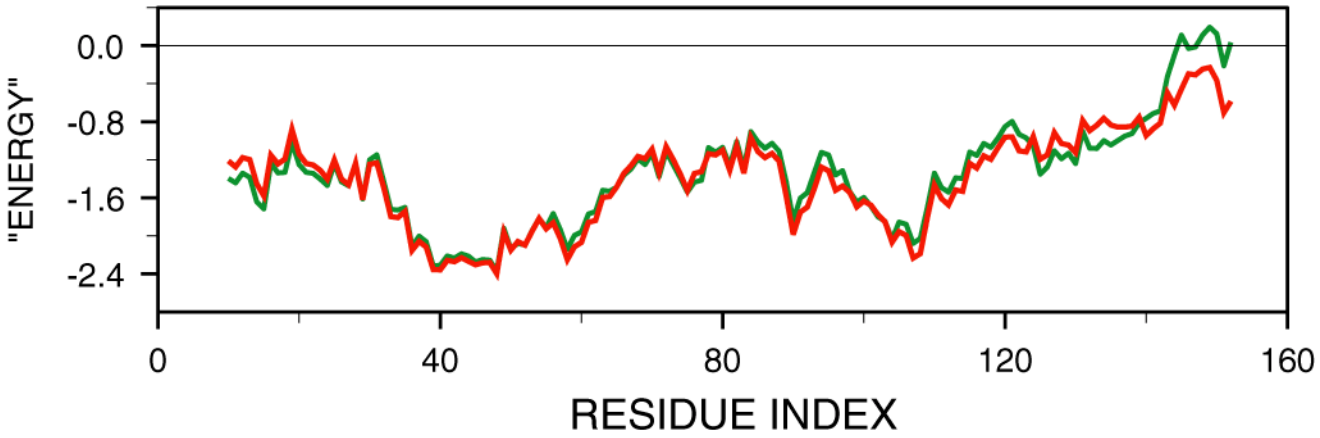

```
Ec d'   -WYSLLAALN---HEQAPARLHWLATLLMDALKR/VTNVDVPGLVAELANHL---SPSRLQAILGDVC
RUVB    DRKLLLAVIDKFF-GGPVGLDNLAAAIGEERETIE--DVLEPYLIQQGFLQRTPRGRMATTRAWNHFG

Ec d'   HIREQLMSVAQANREMVITDRLLRIEHYLQPGWVLP
RUVB    ITPPEMP-----------------------------
```


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

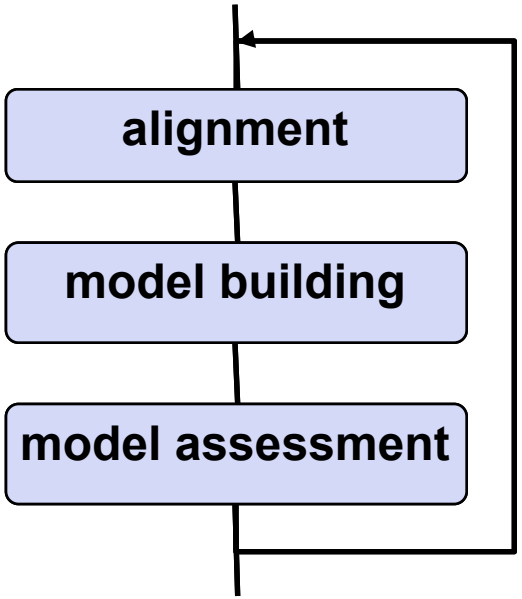
```
          3dfr  TKVSSRTVEDT---NPALTHTYEVWQKKA
         4dfrB  ESVFSEFHDADAQNS--HSYCFKILERR
DFR_1 model1    ELDAETDHEG---------FTLQEWVRSASSR
DFR_1 model2    ELDAETDHEG-------FTLQEWVRSASSR
```

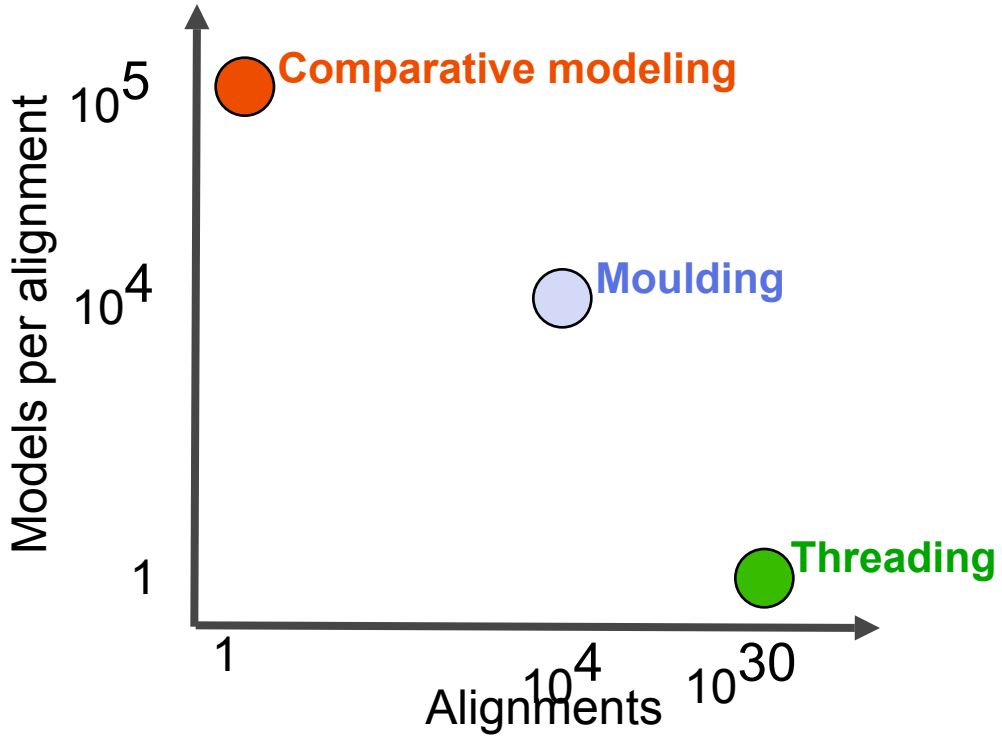


R. Sánchez & A. Šali, Proteins, Suppl. 1, 50-58, 1997

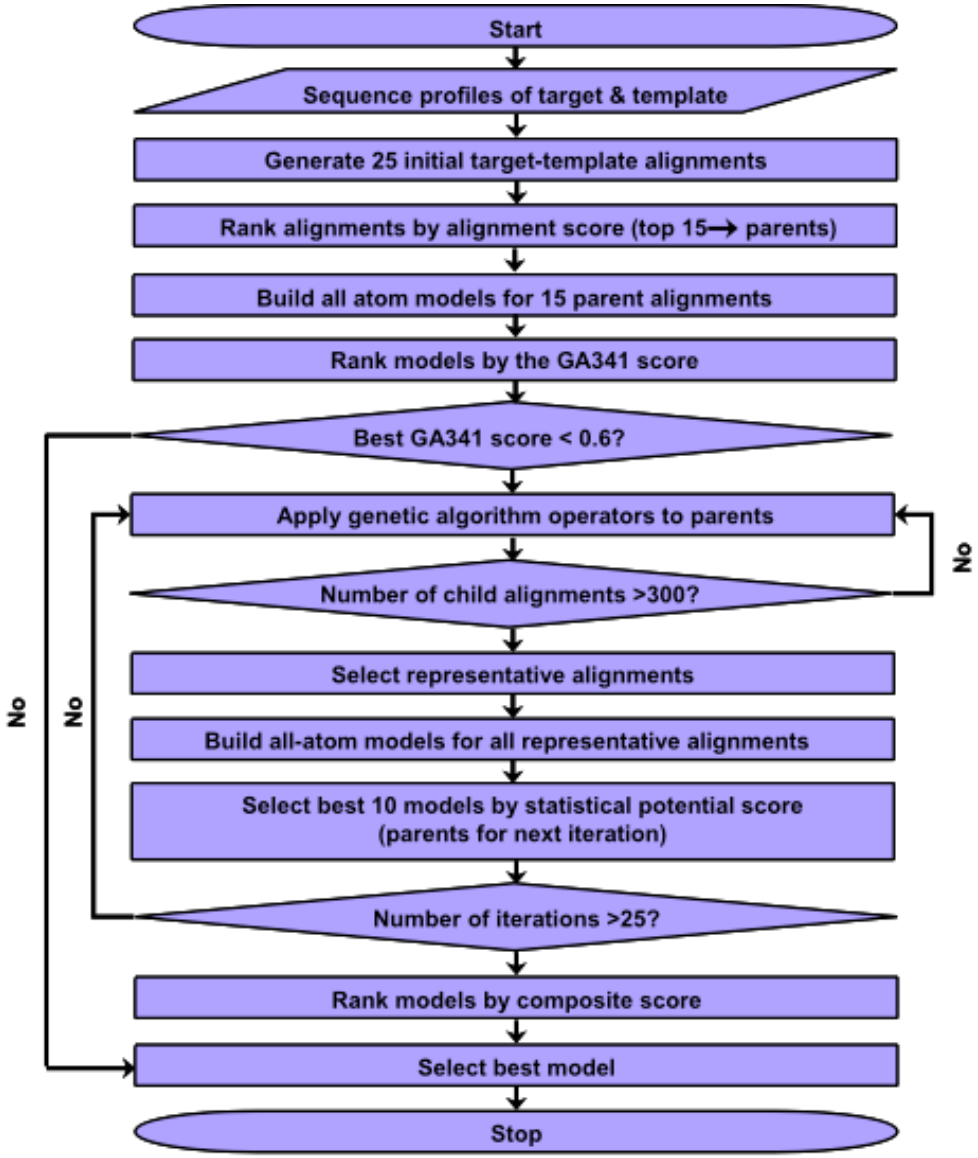
Moulding: iterative alignment, model building, model assessment

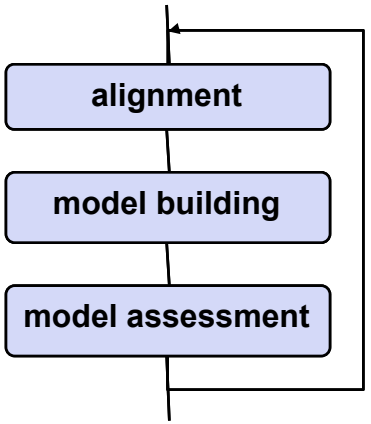
B. John, A. Sali. *Nucl. Acids Res.,* **31**, 1982-1992, 2003.



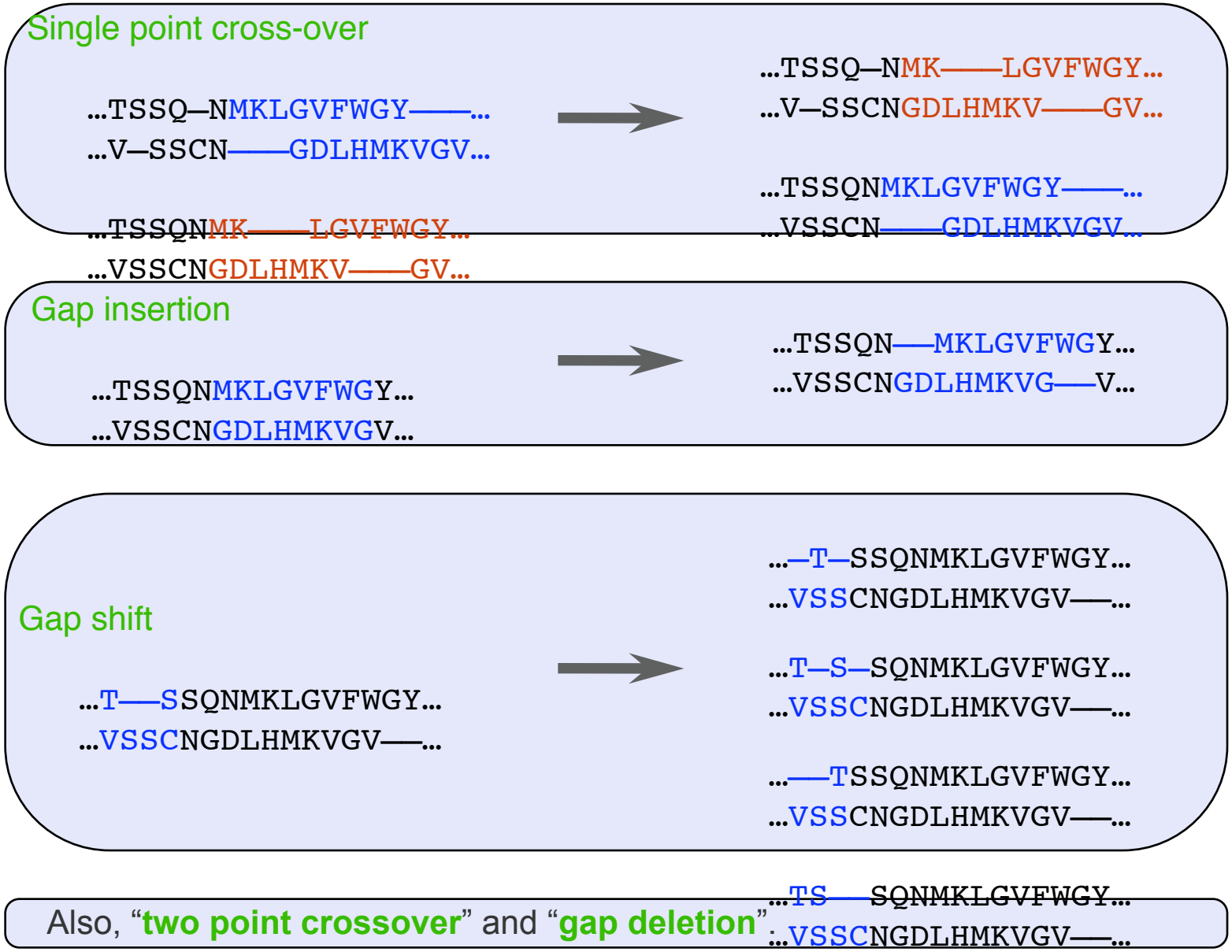


Moulding by a *Genetic Algorithm* approach





Genetic algorithm operators



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).

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

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

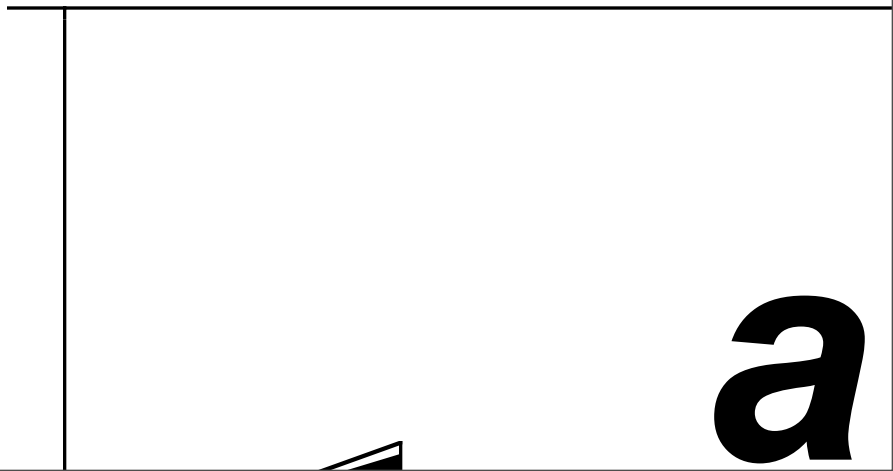
μ … average score of all models

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	
			Cα RMSD [Å]	CE overlap [%]	Cα RMSD [Å]	CE overlap [%]	Cα RMSD [Å]	CE overlap [%]
1ATR-1ATN	13.8	94.3	19.2	20.2	18.8	20.2	17.1	24.6
1BOV-1LTS	4.4	83.5	10.1	29.4	3.6	79.4	3.1	92.6
1CAU-1CAU	18.8	96.7	11.7	15.6	10.0	27.4	7.6	47.4
1COL-1CPC	11.2	81.4	8.6	44.0	5.6	58.6	4.8	59.3
1LFB-1HOM	17.6	75.0	1.2	100.0	1.2	100.0	1.1	100.0
1NSB-2SIM	10.1	89.2	13.2	20.2	13.2	20.1	12.3	26.8
1RNH-1HRH	26.6	91.2	13.0	21.2	4.8	35.4	3.5	57.5
1YCC-2MTA	14.5	55.1	3.4	72.4	5.3	58.4	3.1	75.0
2AYH-1SAC	8.8	78.4	5.8	33.8	5.5	48.0	4.8	64.9
2CCY-1BBH	21.3	97.0	4.1	52.4	3.1	73.0	2.6	77.0
2PLV-1BBT	20.2	91.4	7.3	58.9	7.3	58.9	6.2	60.7
2POR-2OMF	13.2	97.3	18.3	11.3	11.4	14.7	10.5	25.9
2RHE-1CID	21.2	61.6	9.2	33.7	7.5	51.1	4.4	71.1
2RHE-3HLA	2.4	96.0	8.1	16.5	7.6	9.4	6.7	43.5
3ADK-1GKY	19.5	100.0	13.8	26.6	11.5	37.7	7.7	48.1
3HHR-1TEN	18.4	98.9	7.3	60.9	6.0	66.7	4.9	79.3

Alignment accuracy (CE overlap)

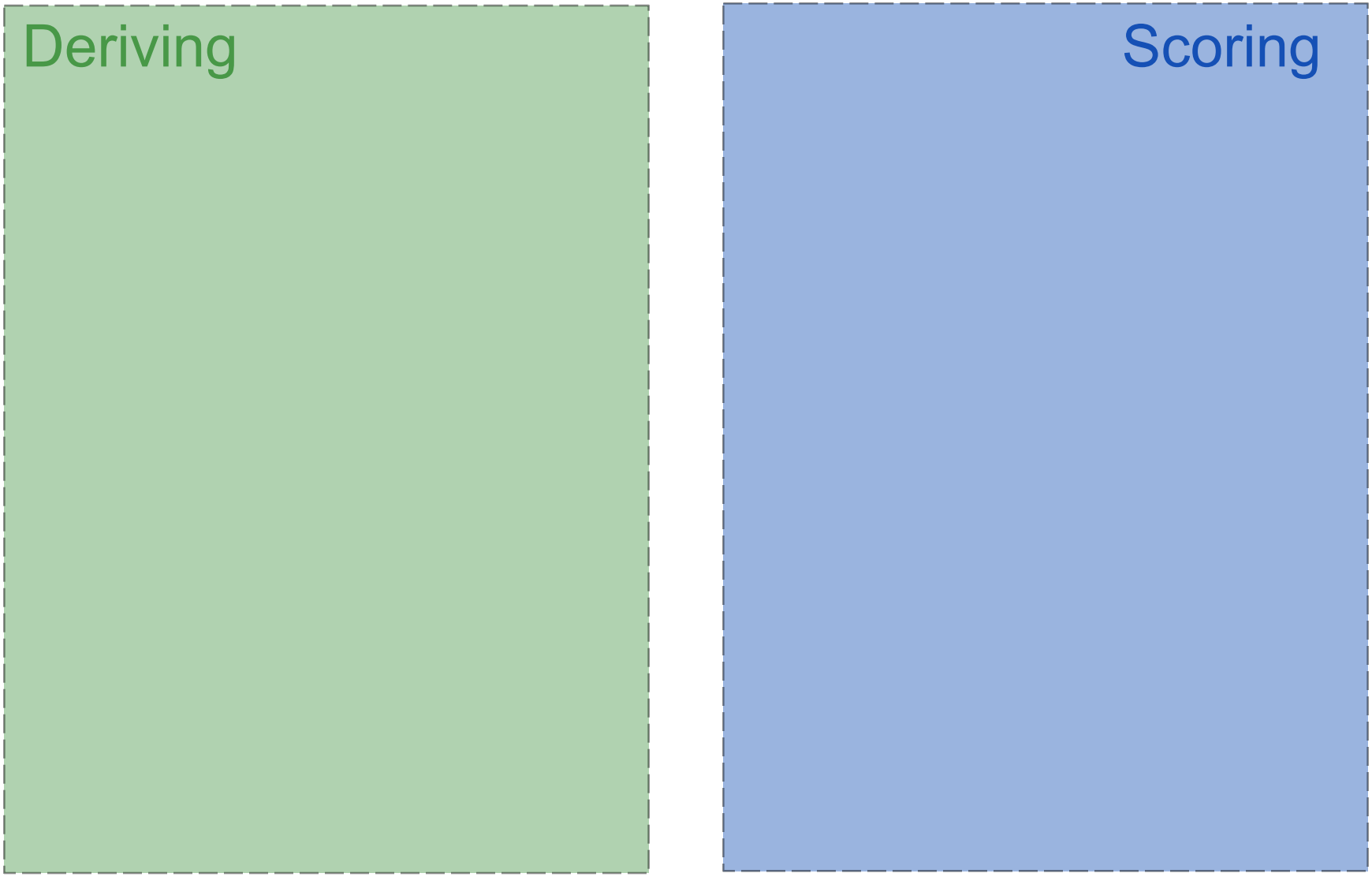
D. Fischer threading test set of 68 structural pairs (a subset of 19):

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

Programs

Scoring

PROSAII

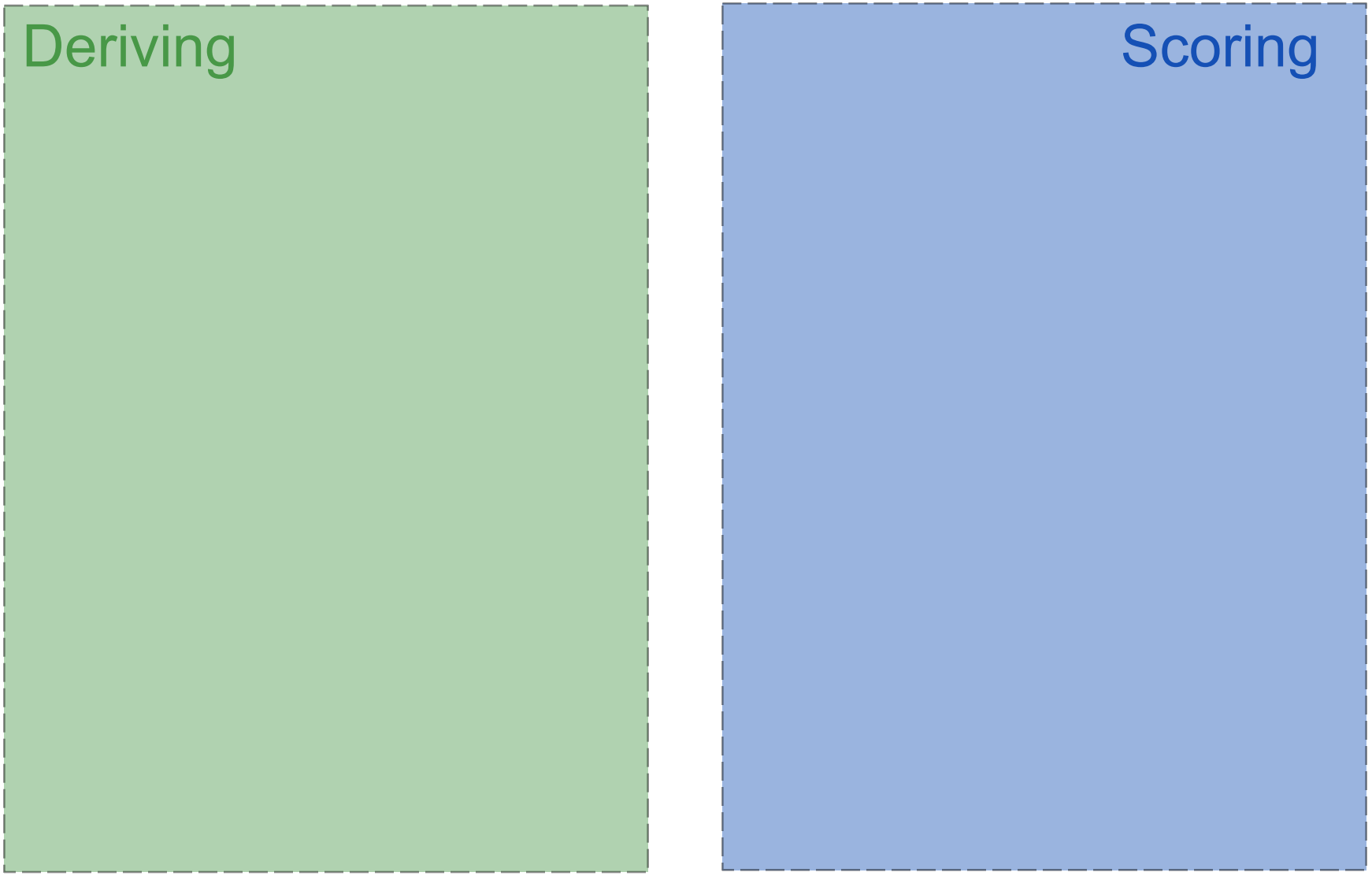


Scoring

PROSAII

Scoring

ANOLEA

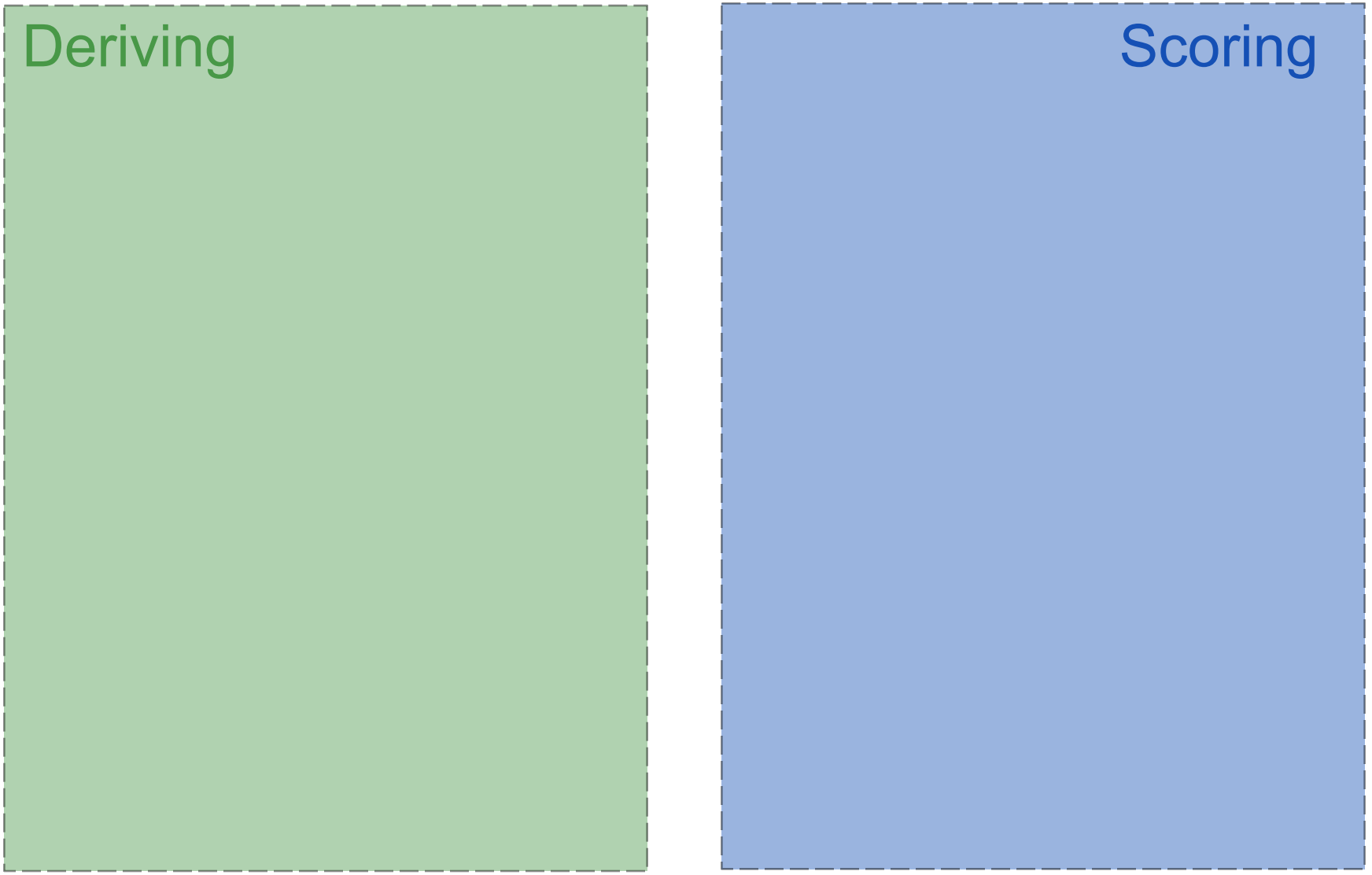


Scoring

ANOLEA

Scoring

VERIFY 3D

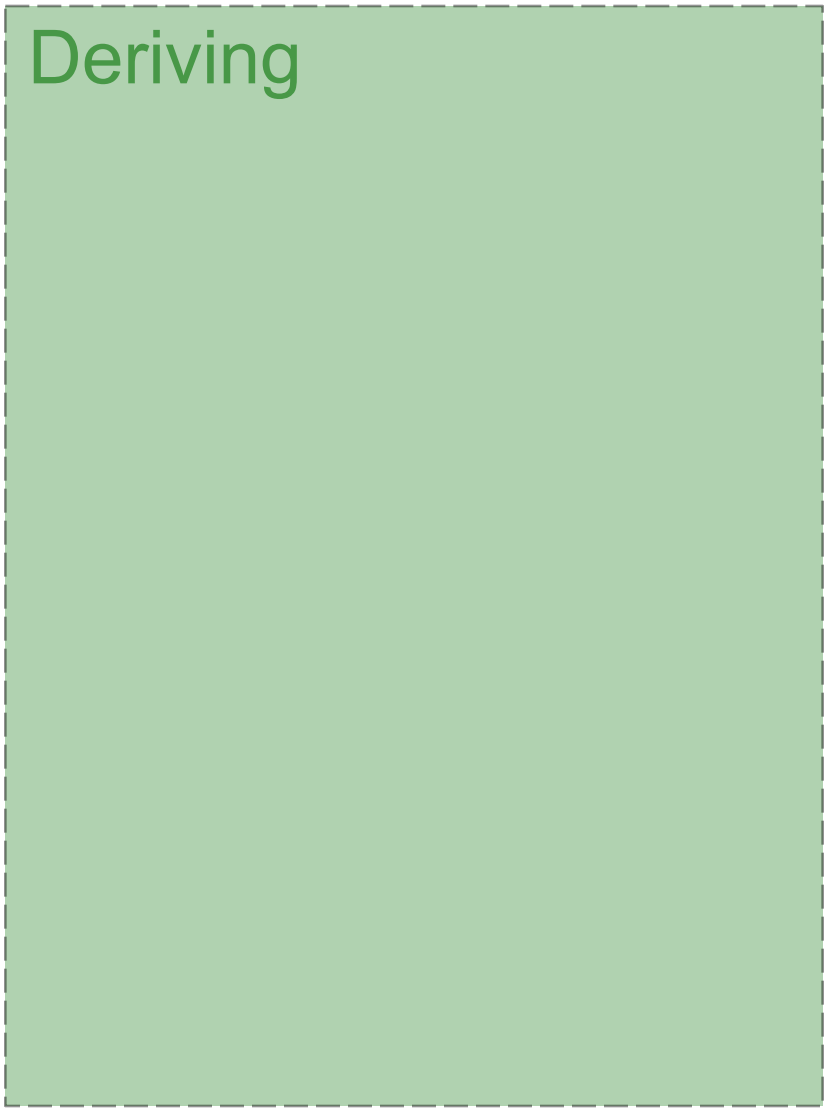


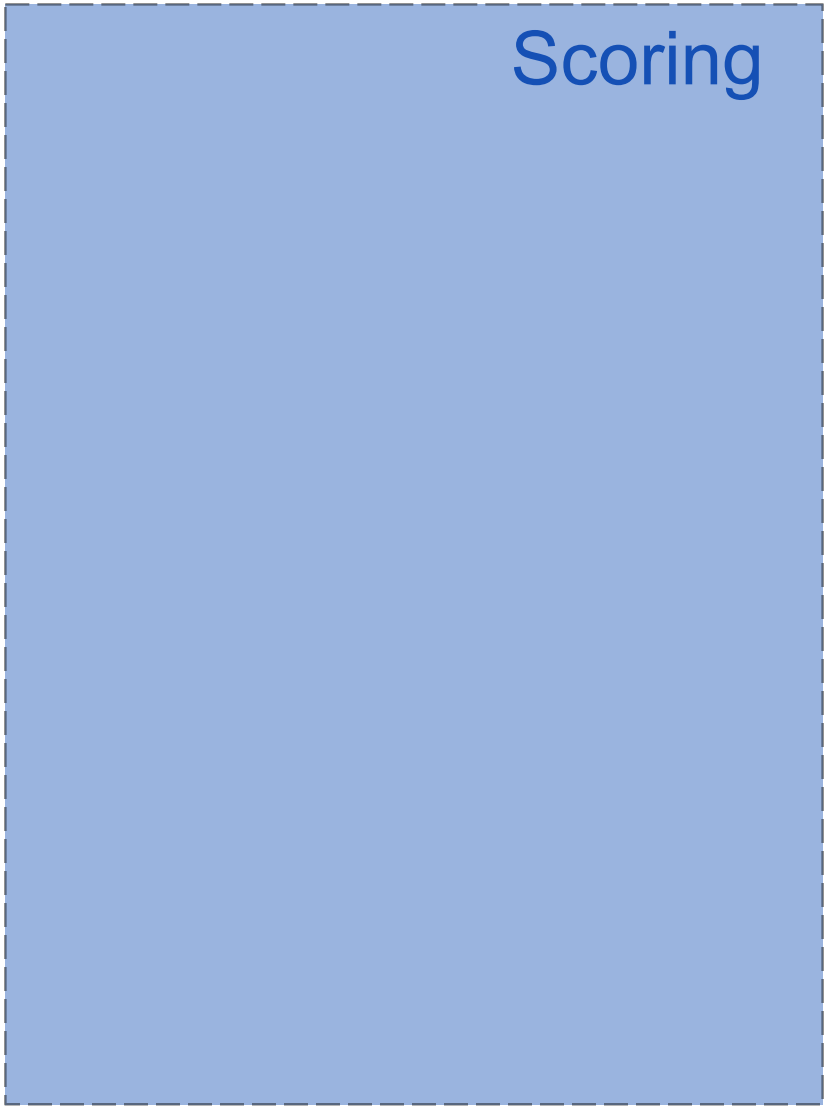
Scoring

VERIFY 3D

Scoring

DFIRE





Scoring

DFIRE

MODEL EVALUATION

a posterior assessment of methods and models

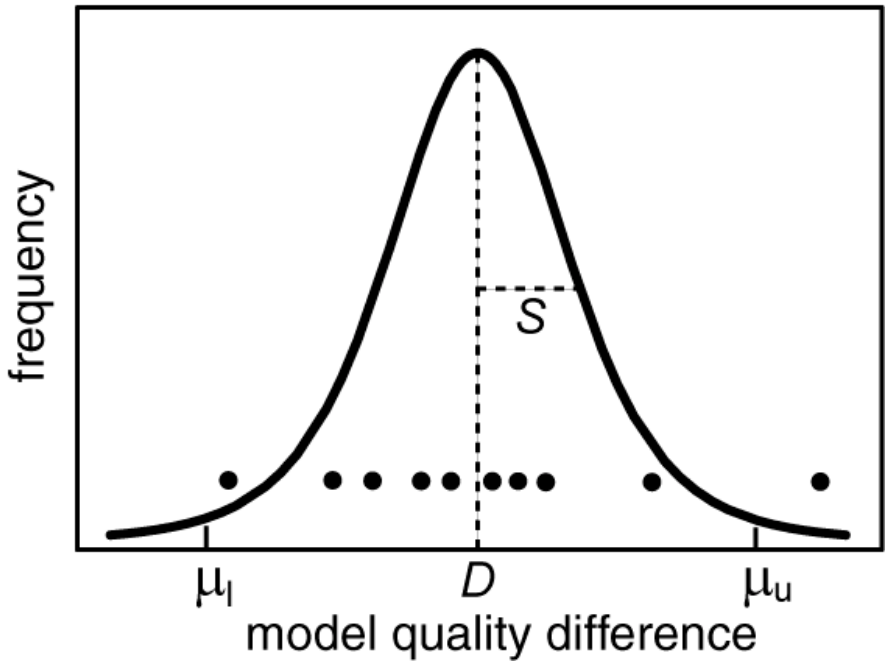
Reliability of assessment of protein structure prediction at CASP

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.

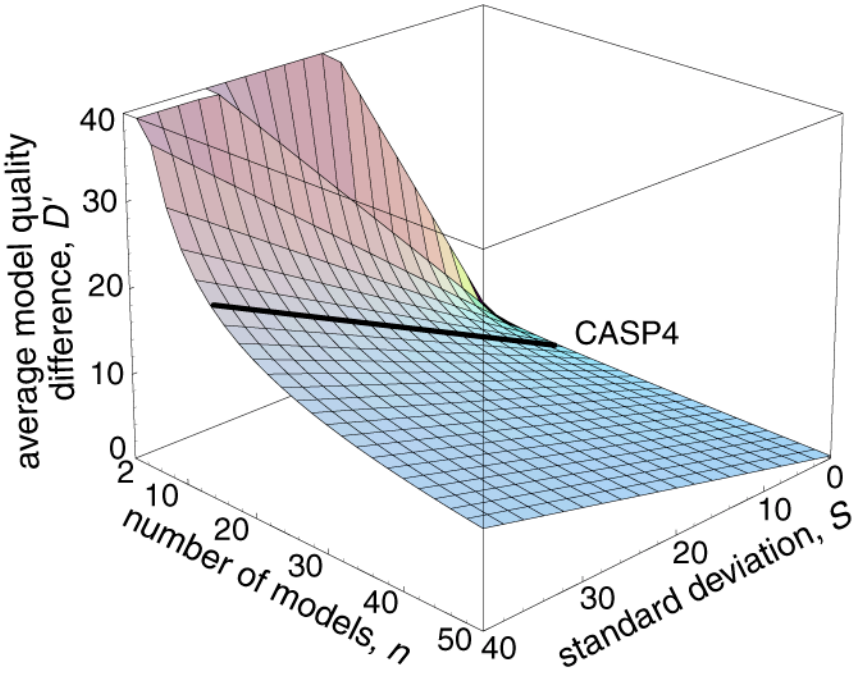
Model Quality Criterion

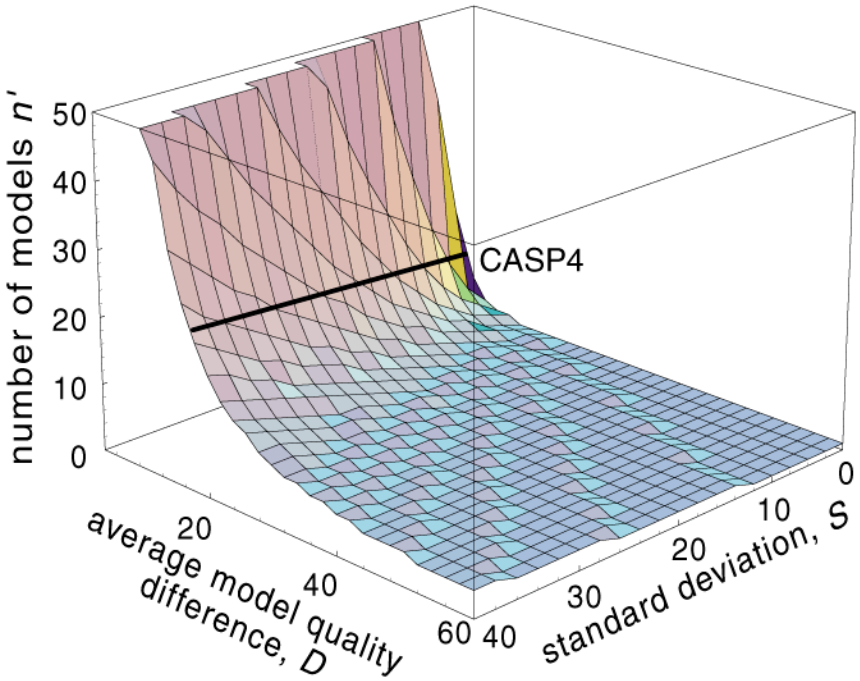
Quality = Coverage and Accuracy.

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

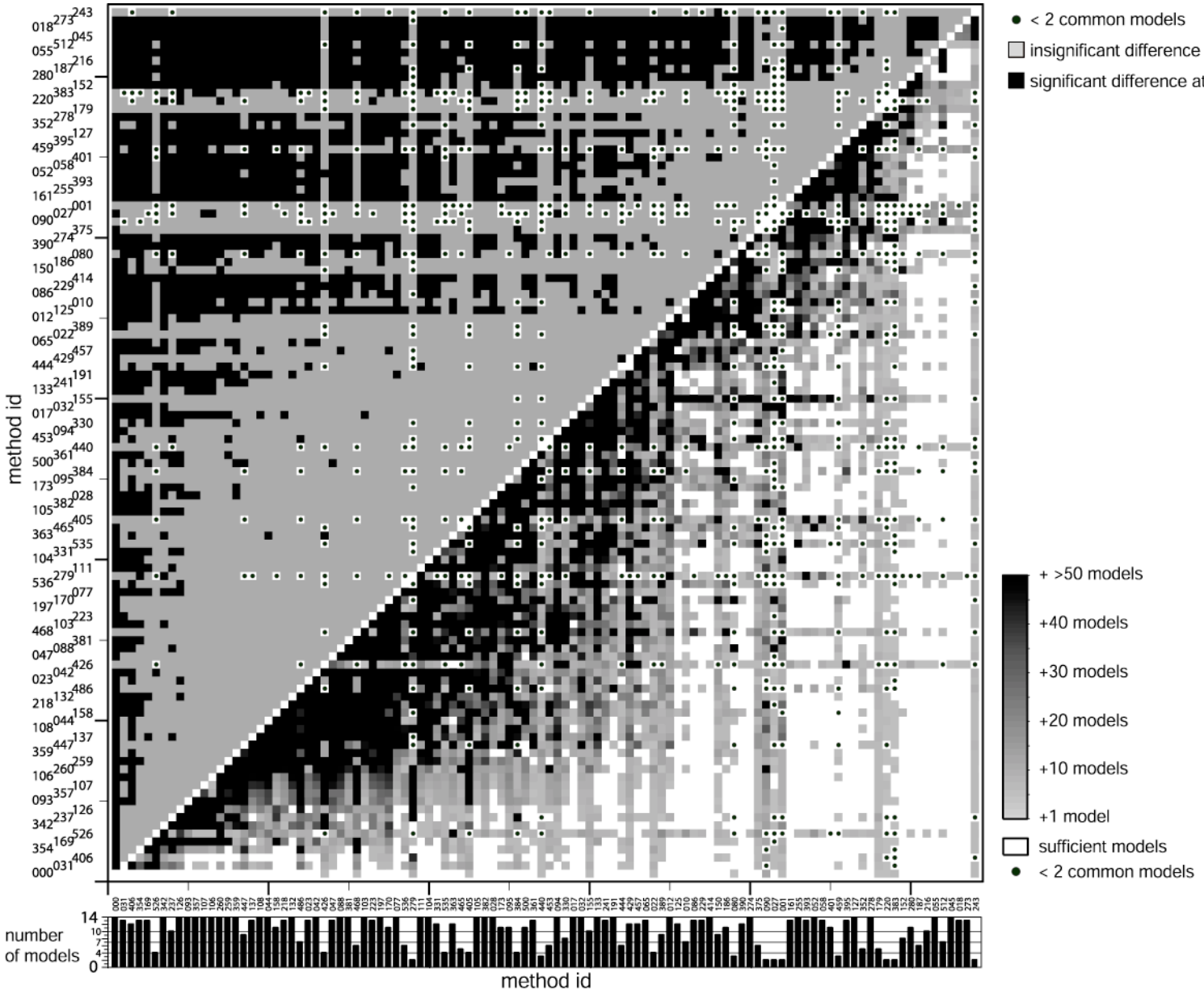
From the CASP web site.

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?

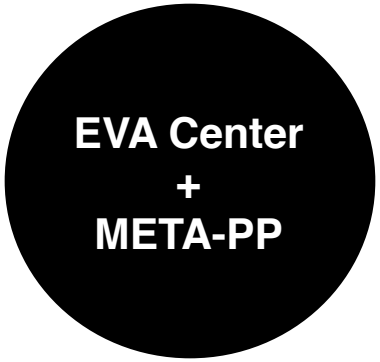
Continuous….

Automatic….

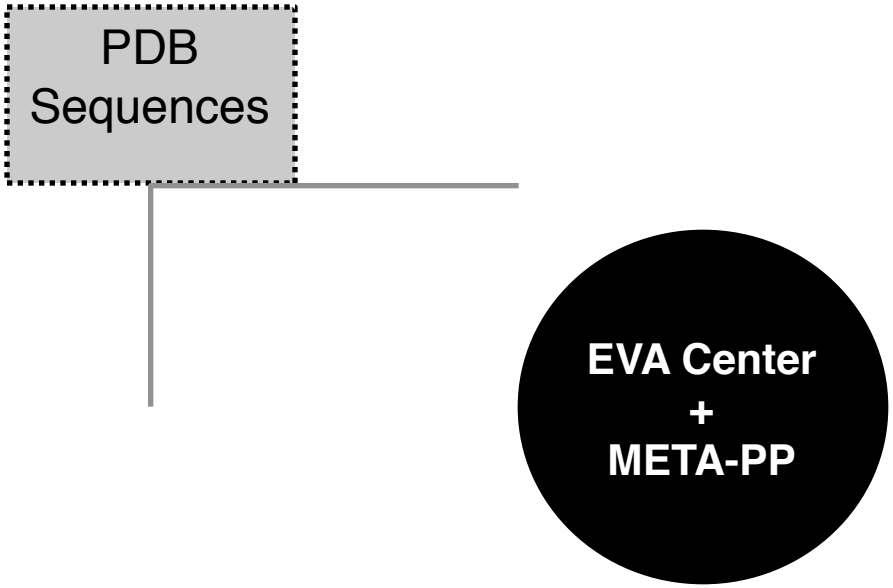
Large scale…

Accessible…

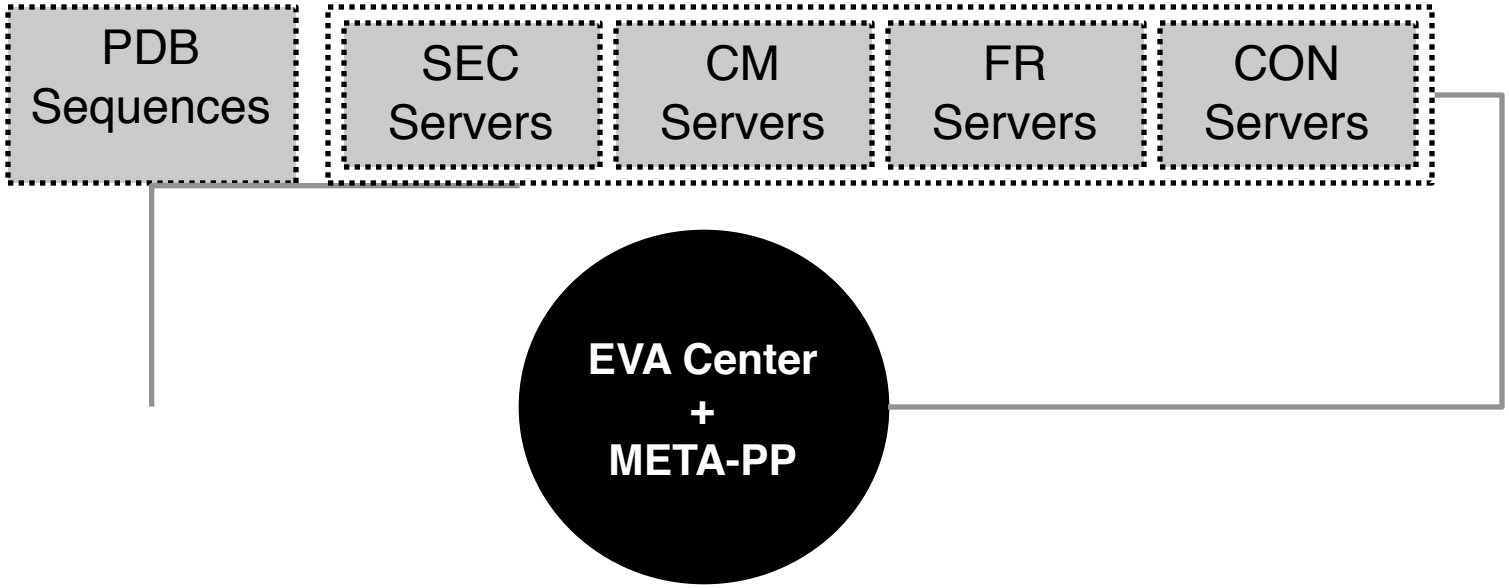
What?



What?

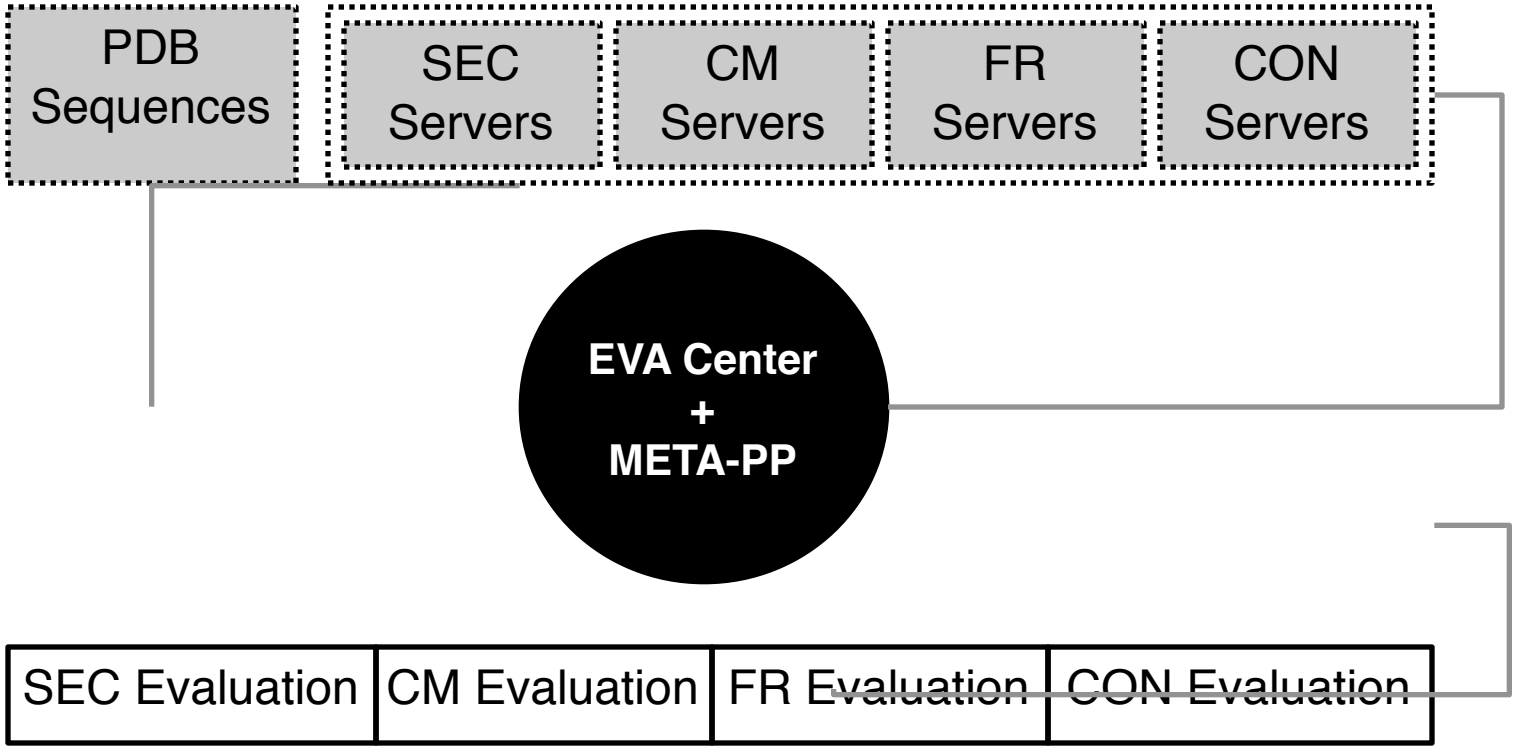


What?



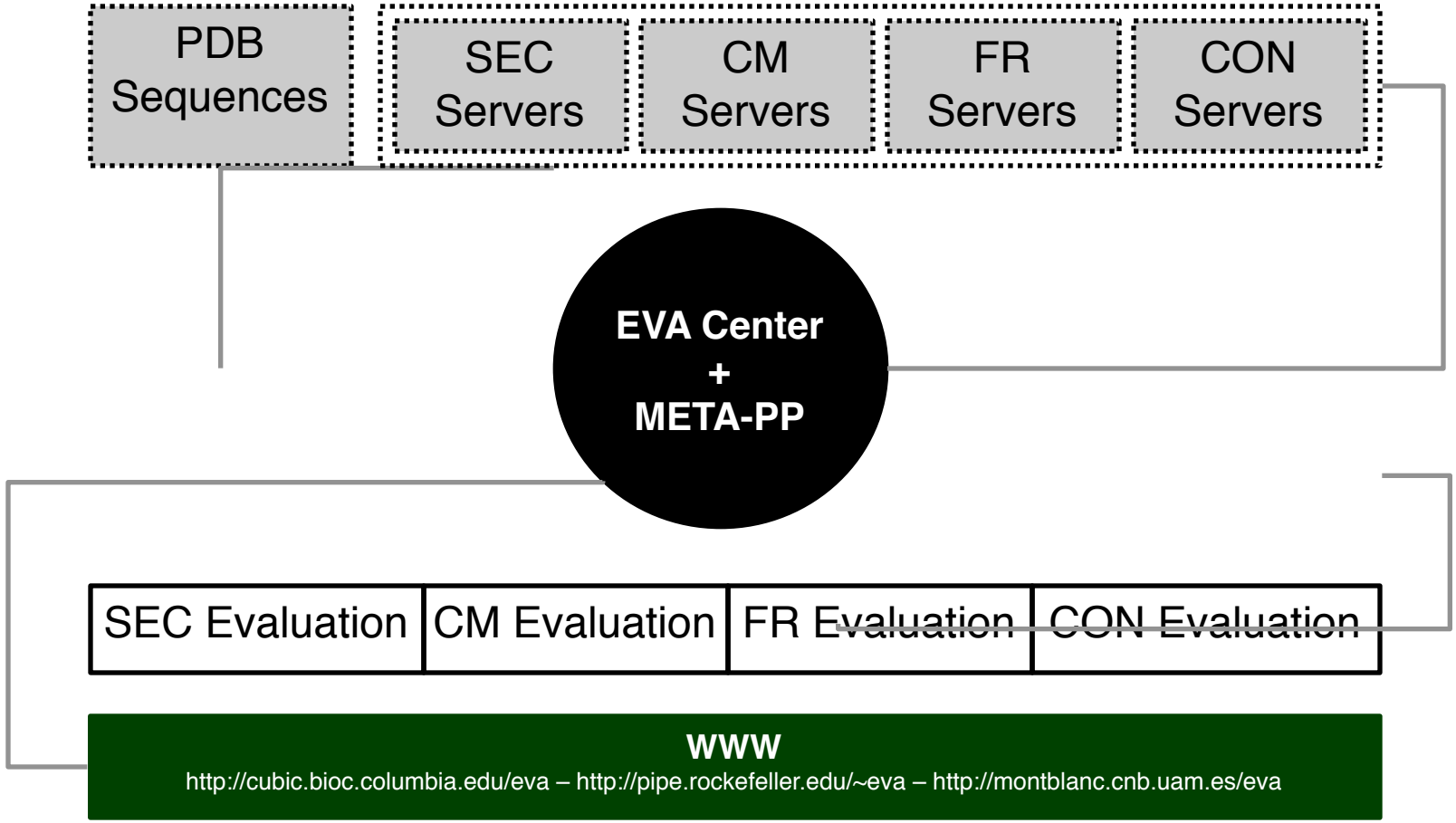
http://cubic.bioc.columbia.edu/eva – http://www.salilab.org/eva – http://montblanc.cnb.uam.es/eva

What?



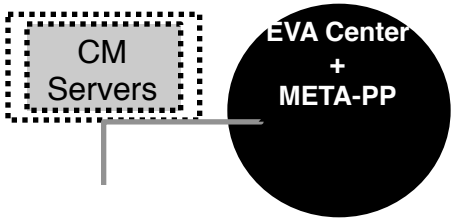
http://cubic.bioc.columbia.edu/eva – http://www.salilab.org/eva – http://montblanc.cnb.uam.es/eva

What?



http://cubic.bioc.columbia.edu/eva – http://www.salilab.org/eva – http://montblanc.cnb.uam.es/eva

How?



How?

CM Servers

EVA Center + META-PP

Eva-CM

T M

Fold accuracy…

aaaAAAAAAAAAA
AAAAaaaAAAAAA

3D alignment by CE

Calculate RMSD, equivalent positions under 3.5Å etc…

Alignment accuracy…

----aaaAAAAAAAAAA
AAAAaaaAAAAAA----

1 to 1 alignment

Calculate RMSD, equivalent positions under 3.5Å etc…

Structure quality…

MODELLER and PROCHECK

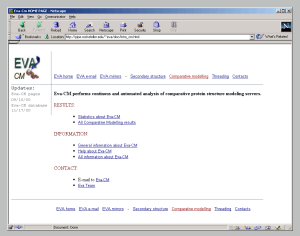
MODELLER and PROCHECK programs will calculate several structural properties of the model.

http://cubic.bioc.columbia.edu/eva – http://www.salilab.org/eva – http://montblanc.cnb.uam.es/eva

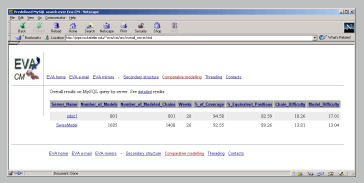
How?

CM Servers

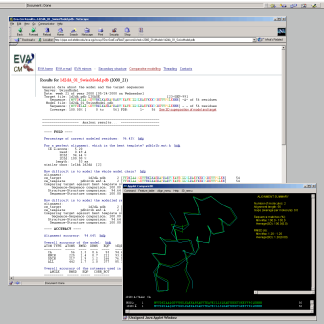
EVA Center + META-PP



General level



Server level



Model level

Eva-CM

Fold accuracy…

aaaAAAAAAAAAA
AAAAaaaAAAAAA

3D alignment by CE

Calculate RMSD, equivalent positions under 3.5Å etc…

Alignment accuracy…

----aaaAAAAAAAAAA
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.

http://cubic.bioc.columbia.edu/eva – http://www.salilab.org/eva – http://montblanc.cnb.uam.es/eva

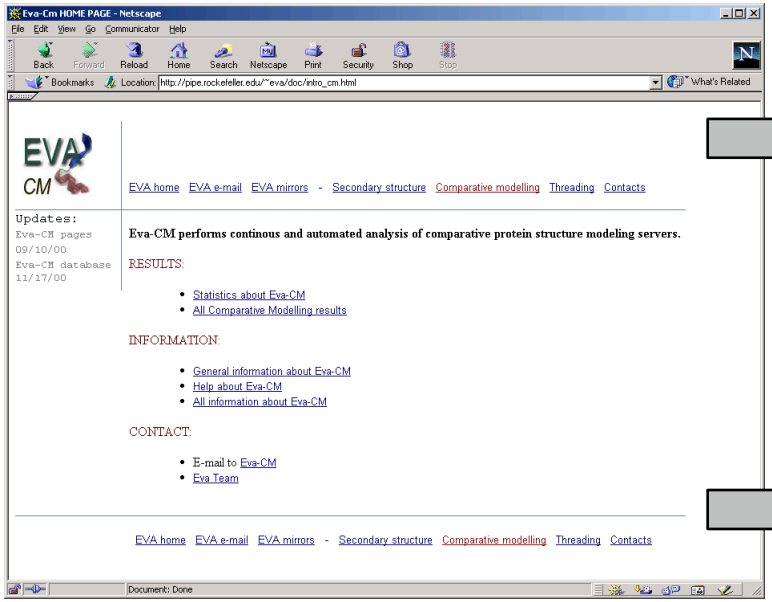
Where?

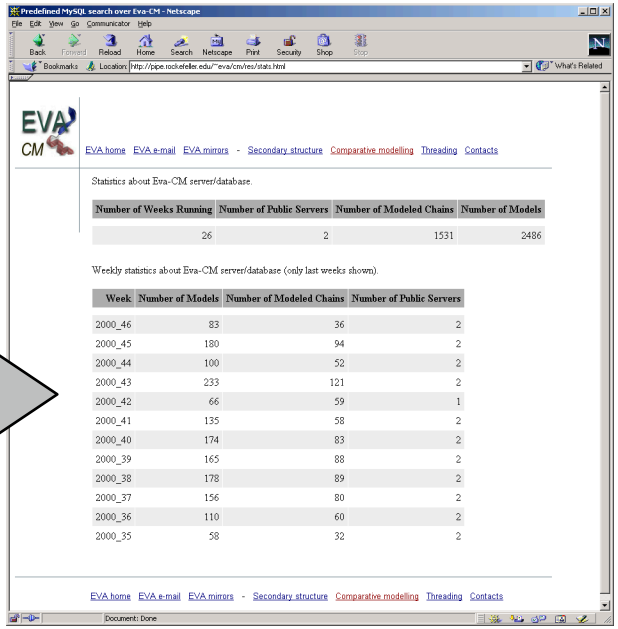
http://cubic.bioc.columbia.edu/eva/ (NYC, USA)

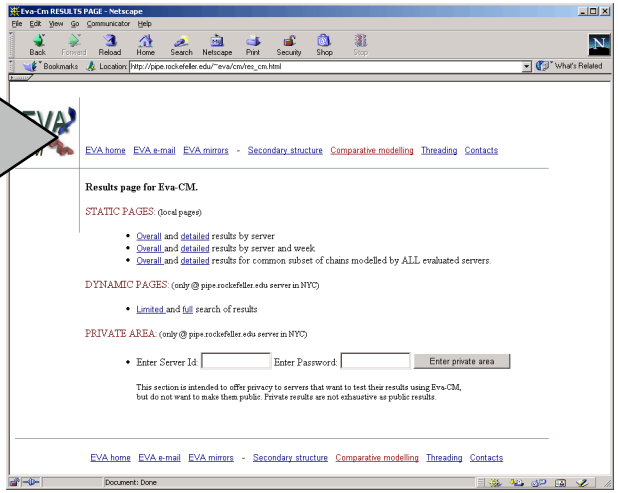
http://www.salilab.org/eva/ (UCSF, USA)

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

Main pages @ Eva-CM

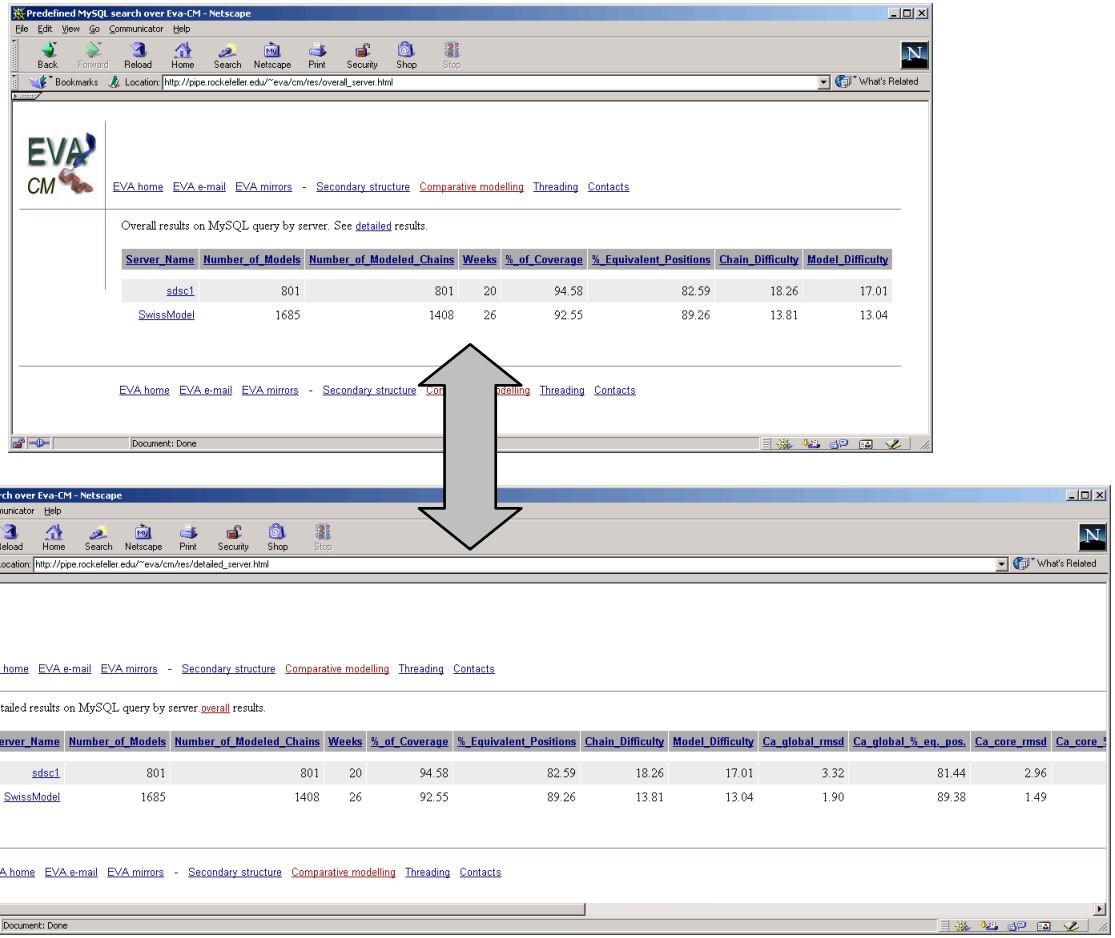






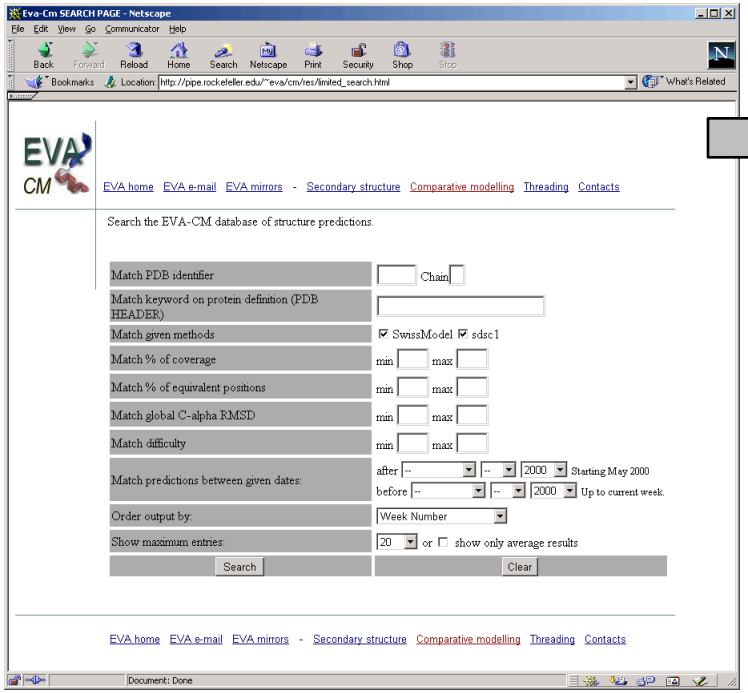
http://cubic.bioc.columbia.edu/eva – http://www.salilab.org/eva – http://montblanc.cnb.uam.es/eva

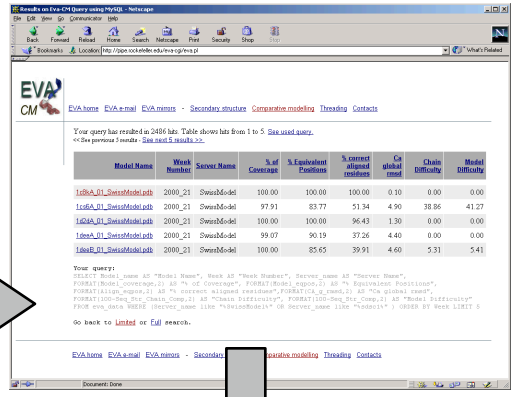
Server results pages @ Eva-CM

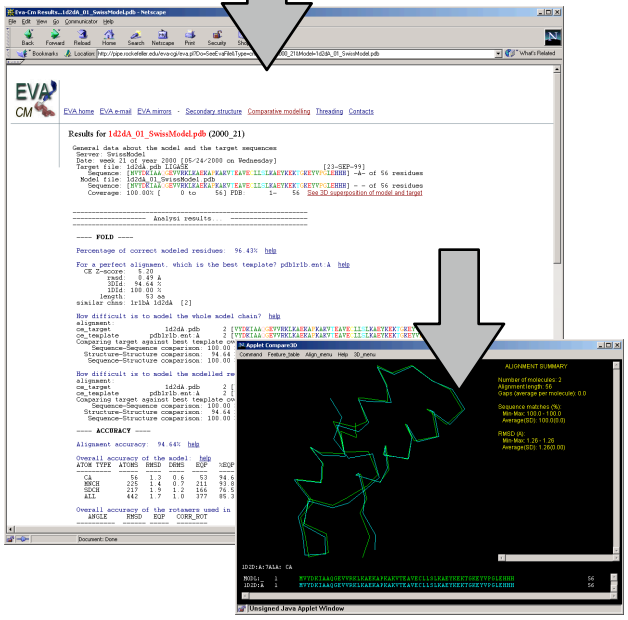


http://cubic.bioc.columbia.edu/eva – http://www.salilab.org/eva – http://montblanc.cnb.uam.es/eva

Query and model results pages @ Eva-CM

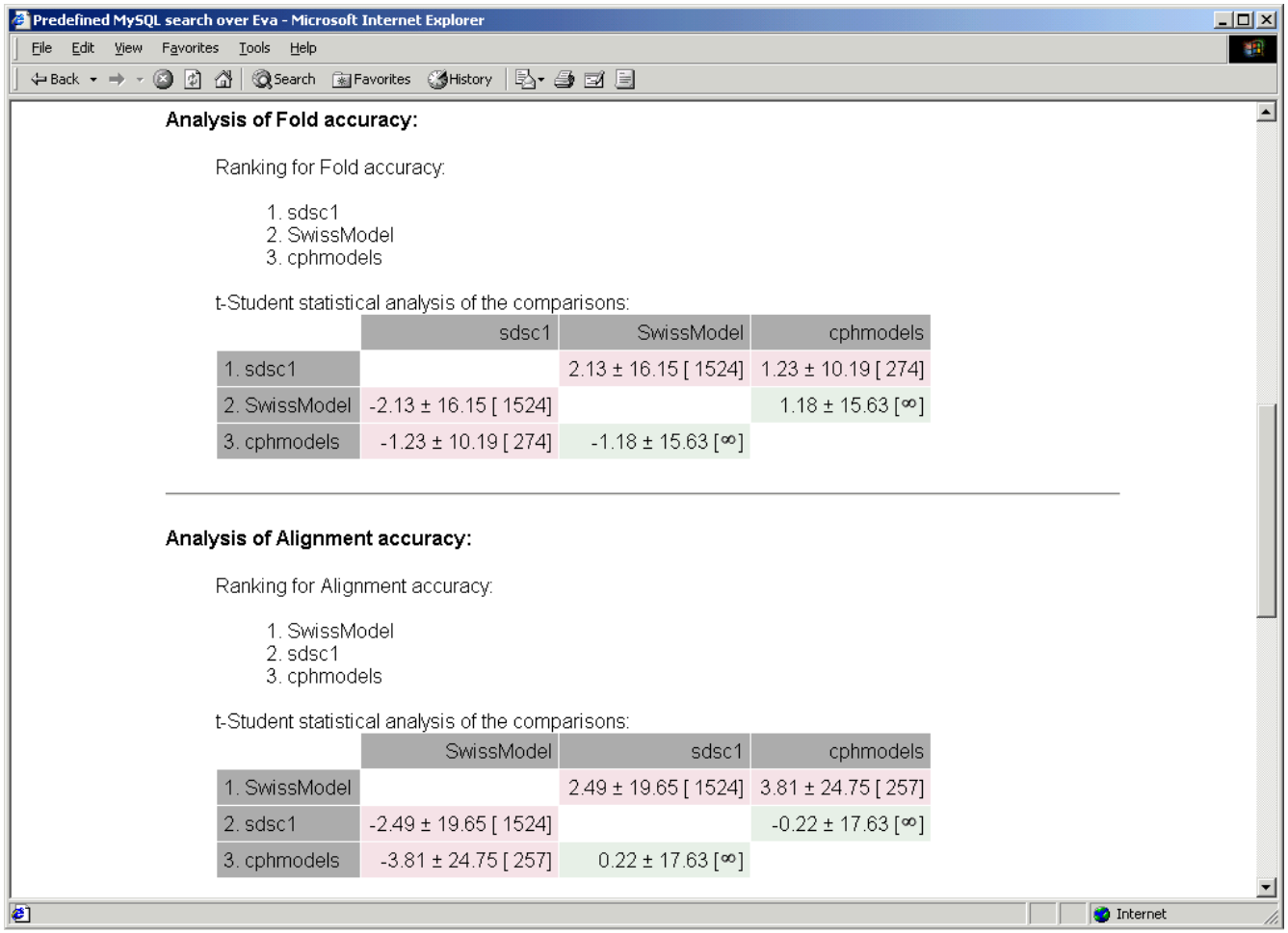






http://cubic.bioc.columbia.edu/eva – http://www.salilab.org/eva – http://montblanc.cnb.uam.es/eva

Ranking @ Eva-CM



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

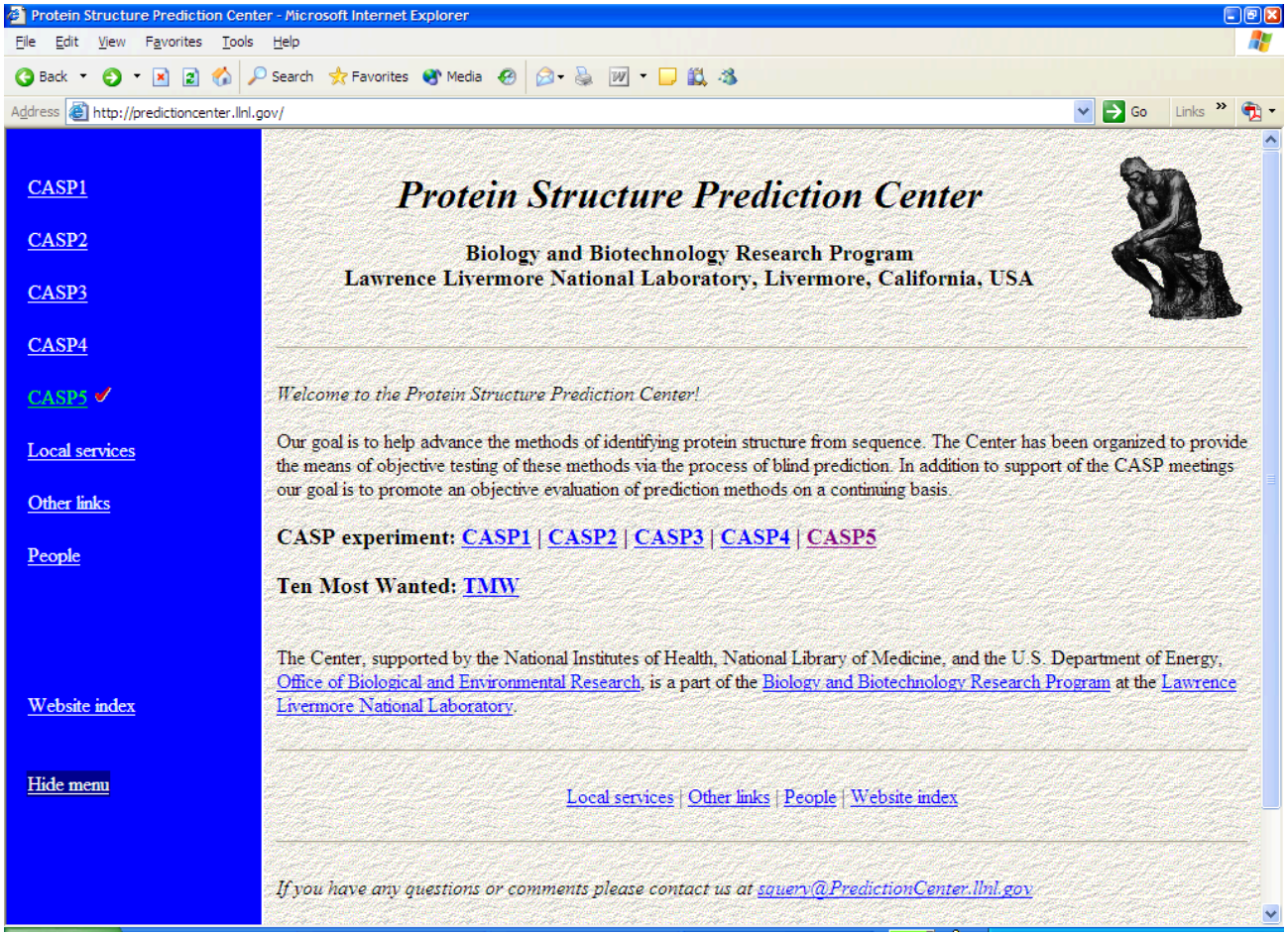
- **All groups developing servers evaluated by EVA.**

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/**



http://cubic.bioc.columbia.edu/eva – http://www.salilab.org/eva – http://montblanc.cnb.uam.es/eva

BMC WorkShop

Protein Structure Prediction

Summary

Marc A. Marti-Renom & Damien Devos

Department of Biopharmaceutical Sciences, UCSF

protein prediction .vs. protein determination

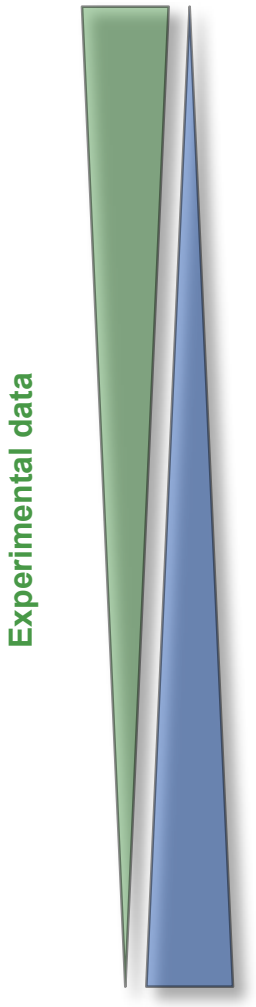
X-Ray

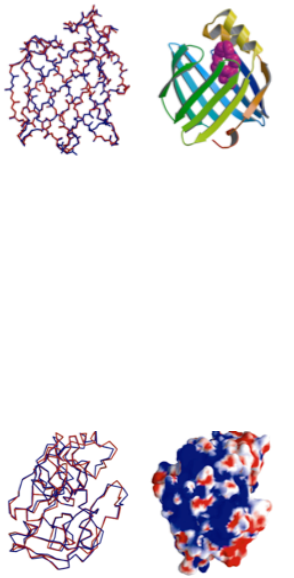
NMR

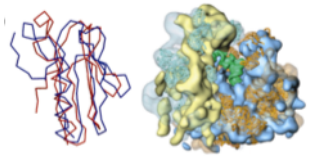
Comparative Modeling

Threading

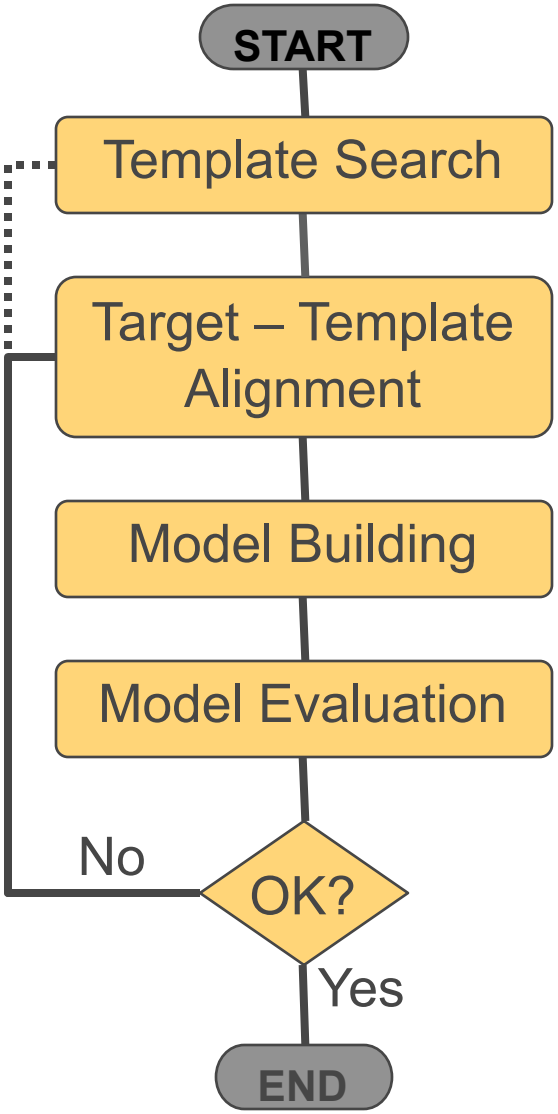
Ab-initio

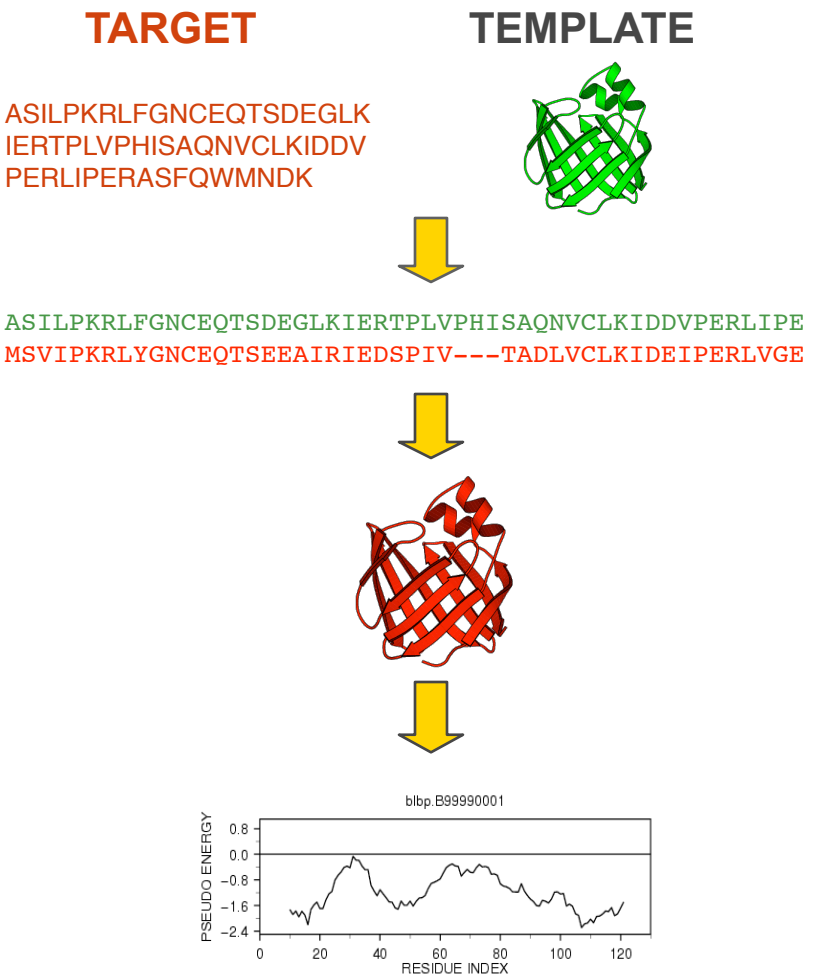






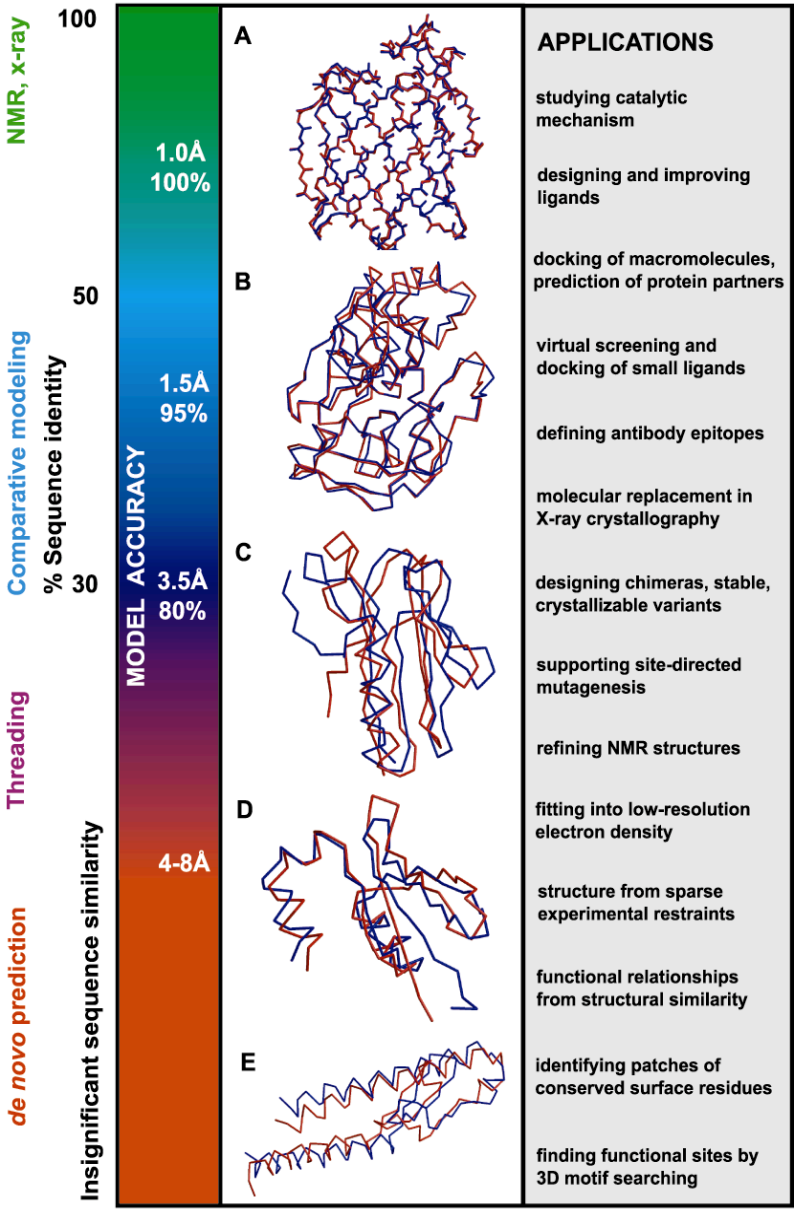
Steps in Comparative Protein Structure Modeling





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

Utility of protein structure models, despite errors



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

Acknowledgments

http://salilab.org

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

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)

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

Brain Lipid Binding Protein
Liang Zhu (RU)
Nat Heintz (RU)

Structural Genomics
Stephen Burley (SGX)
John Kuriyan (UCB)
NY-SGXRC

BRCA1
A. Monteiro (Cornel)

Mast Cell Proteases
Rick Stevens (BWH)

Fly p53
Shengkan Jin (RU)
Arnie Levine (RU)

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