

# BMI-206

## Structure-Structure comparisons Sequence-Structure comparisons

Marc A. Marti-Renom  
Assistant Adjunct Professor  
Department of Biopharmaceutical Sciences

February 17<sup>th</sup>, 2004

# How to use this lectures

- Ask!
- Each day...
  - Basic introduction
  - Theory (representation-scoring-optimization)
  - Available programs
  - Application
- Second day we discuss the assignment for the class
  - *The BMI206 genome. Structural and functional annotation.*

# Outline of the lectures

- Day 1.
  - Structure-Structure comparisons
  - Databases of protein structure classification
- Day 2.
  - Sequence-Structure comparison
  - Description of the assignment

# Day 1. Summary

- Structure-Structure comparisons
  - Before we start...
    - 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

# Day 2. Summary

- Sequence-Structure comparisons
  - Before we start...
    - Some theory...
    - Domain boundaries
  - Structural predictions from sequence...
    - PROF (SSE prediction)
    - SALIGN (gap penalties and substitution matrices)
    - mGenThreader (SSE prediction and alignment/potential scores)
    - Fugue (gap penalties and substitution matrices)
    - 3D-Jury (as a meta server example)
  - What does work?
    - EVA server (EVA-Threading)

# DAY 1

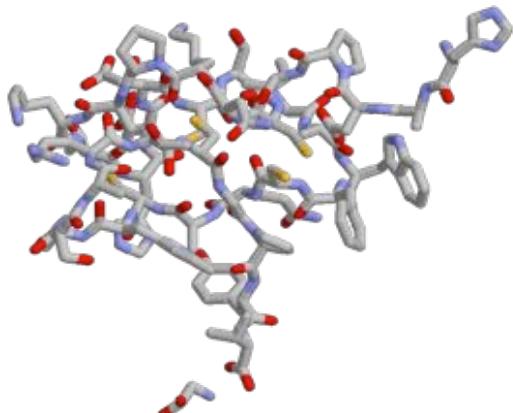
## “Structural Space”

# Structure-Structure alignments

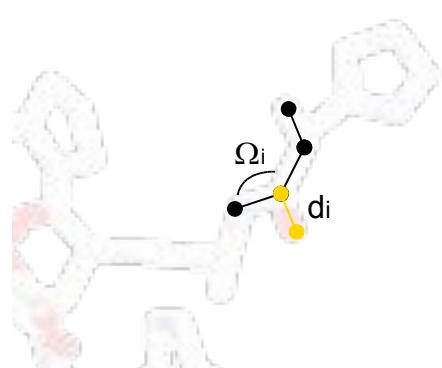
As any other bioinformatics problem...

- Representation
- Scoring
- Optimizer

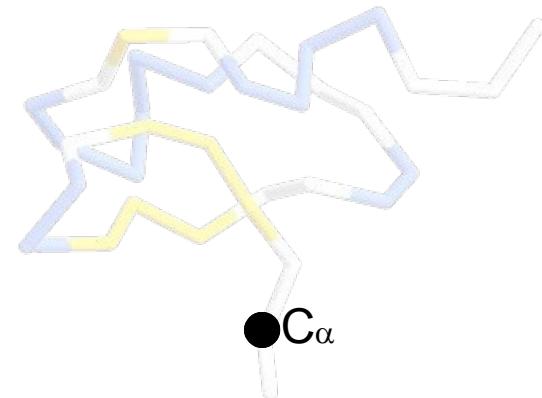
# Representation Structures



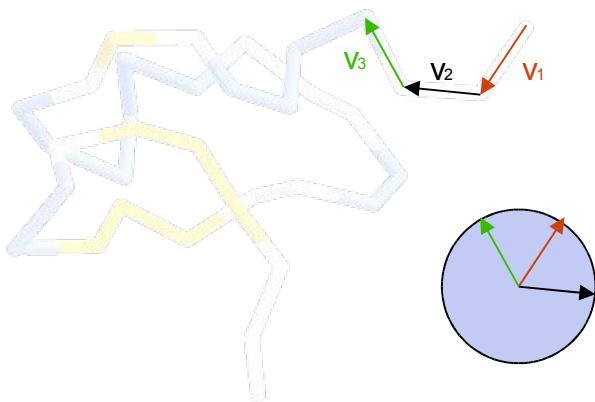
All atoms and coordinates



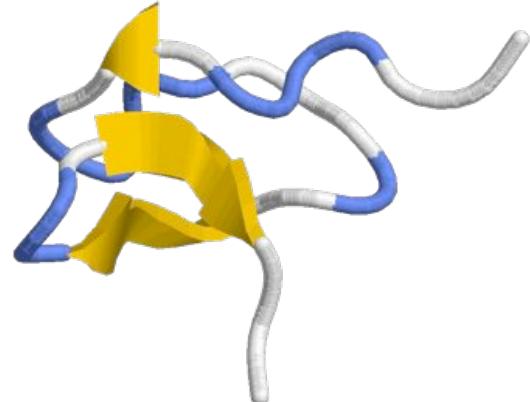
Dihedral space or distance space



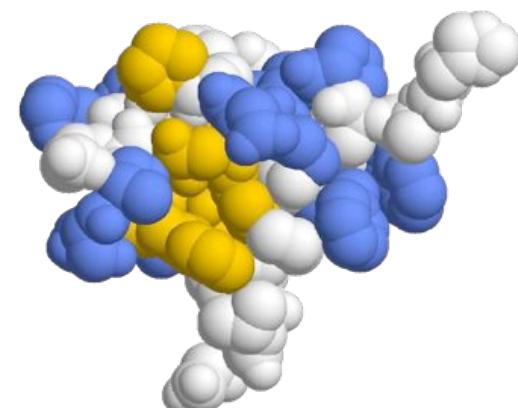
Reduced atom representation



Vector representation



Secondary Structure



Accessible surface (and others)

# Scoring

## Raw scores

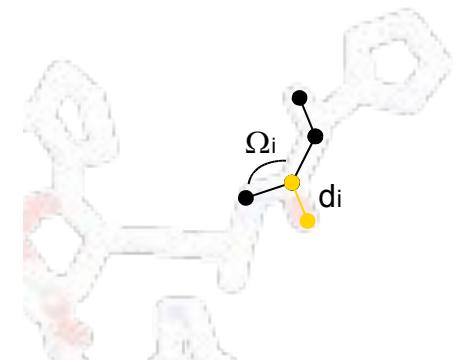
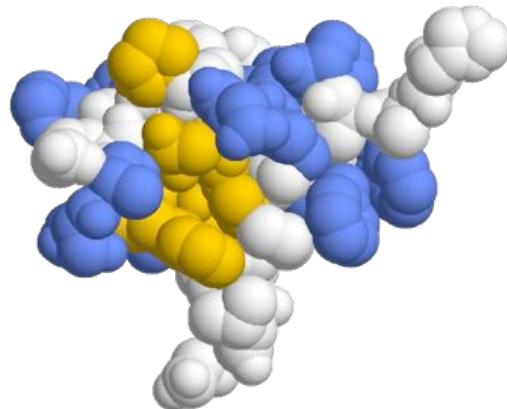
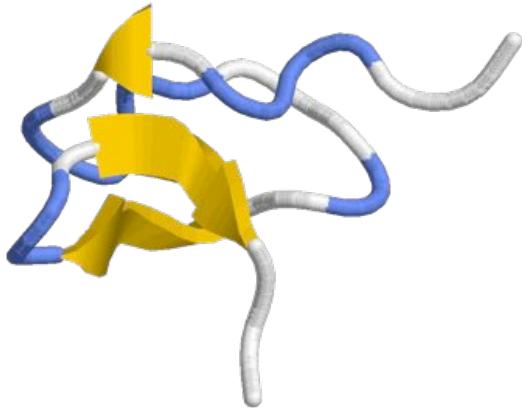
	C	S	T	F	A	G	M	D	E	Q	H	R	K	M	I	L	V	P	N	Y	W
C	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	-2	-2	
S	-1	4	-1	1	0	1	0	0	0	-1	0	0	2	-1	-1	-1	-1	-1	-1	-1	
T	-1	1	4	-1	0	1	0	0	0	-1	0	0	2	-1	-1	-1	-1	-1	-1	-1	
F	0	-1	1	4	-1	0	-1	-1	0	0	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	
A	0	1	-1	4	-1	0	-1	-1	0	0	-1	-1	-1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

21

Aminoacid substitutions

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

Root Mean Square Deviation

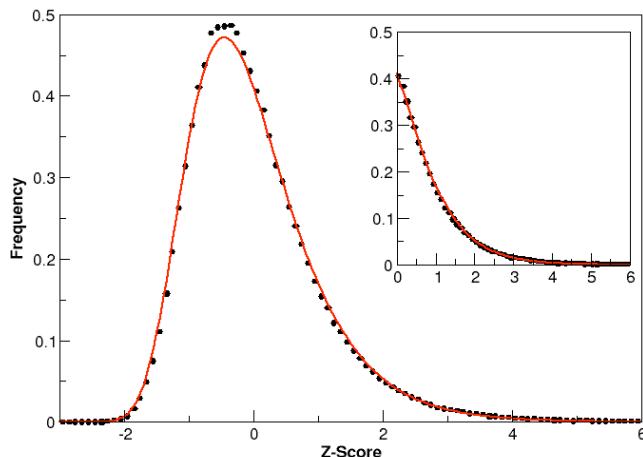
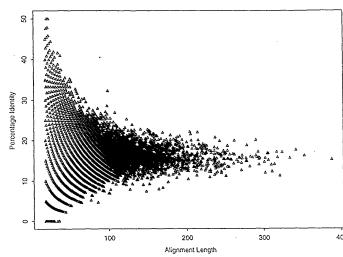


## Scoring

# Significance of an alignment (score)

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

Empirical



Sometimes approximated by Z-score (normal distribution).

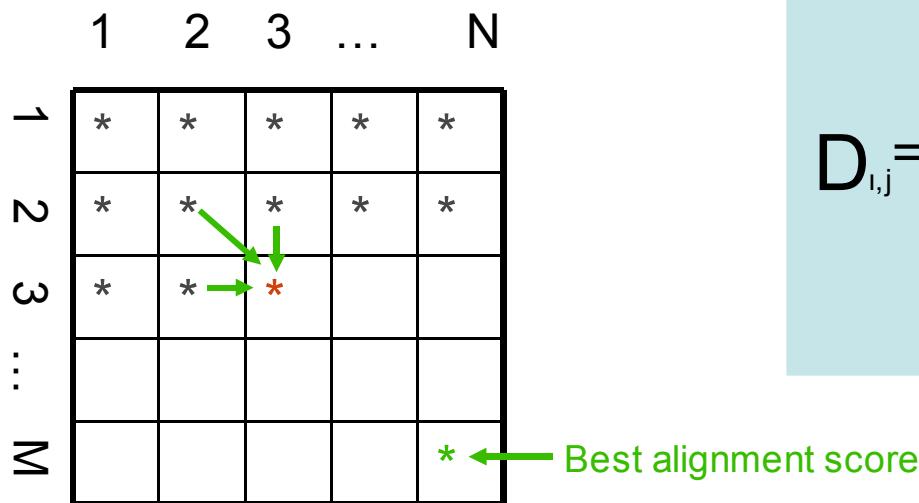
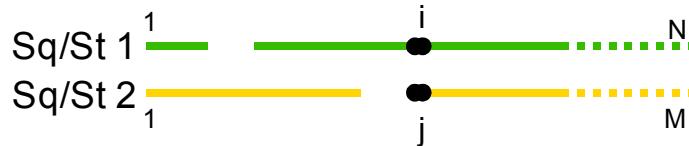
Analytic

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

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

Karlin and Altschul, 1990 PNAS 87, pp2264

# Global dynamic programming alignment

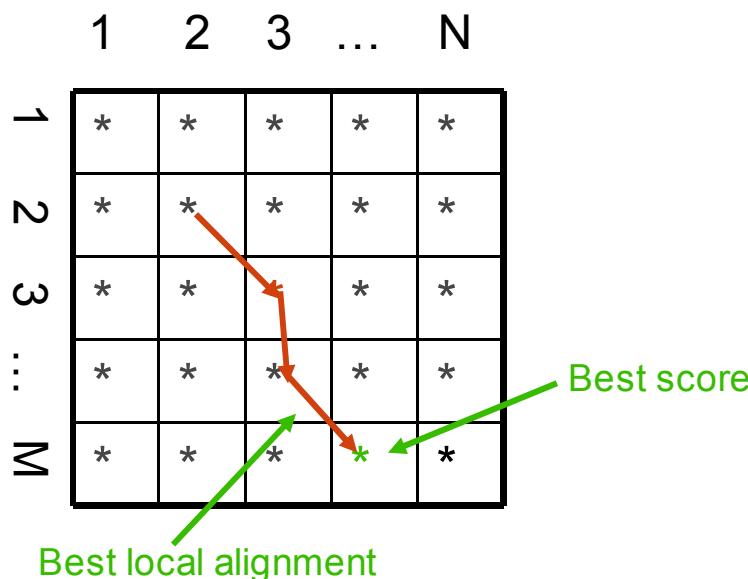
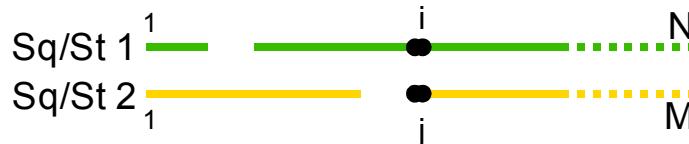


$$D_{i,j} = \min \begin{cases} D_{i,j-1} + \text{Score}_{(\Delta, rj)} \\ D_{i-1,j-1} + \text{Score}_{(ri, rj)} \\ D_{i-1,j} + \text{Score}_{(ri, \Delta)} \end{cases}$$

Backtracking to get the best alignment

## Optimizer

# Local dynamic programming alignment



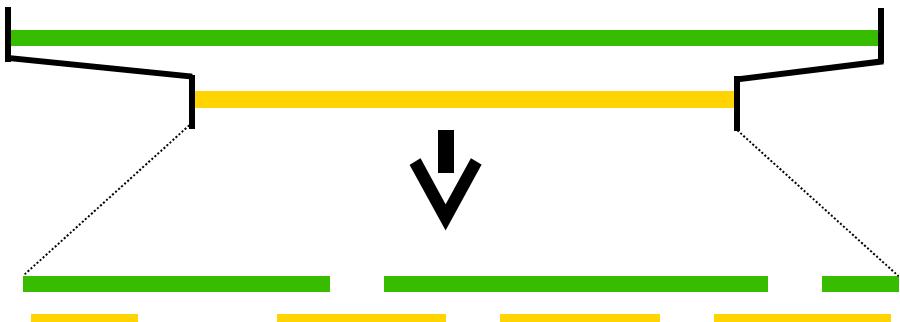
$$D_{i,j} = \min \left\{ \begin{array}{l} D_{i-1,j-1} + \text{Score}_{(\Delta, rj)} \\ D_{i-1,j-1} + \text{Score}_{(ri, rj)} \\ D_{i-1,j} + \text{Score}_{(ri, \Delta)} \\ 0 \end{array} \right.$$

Backtracking to get the best alignment

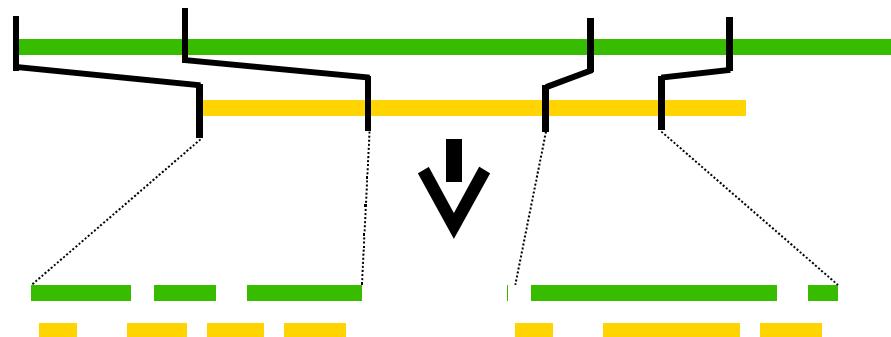
Smith and Waterman (1981) *J. Mol Biol.*, 147 pp195

Optimizer

# Global .vs. local alignment



Global alignment

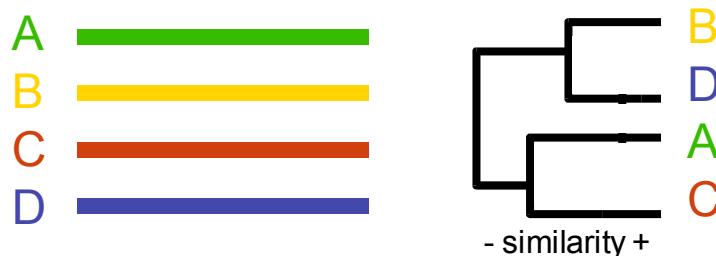


Local alignment

# Multiple alignment

## Pairwise alignments

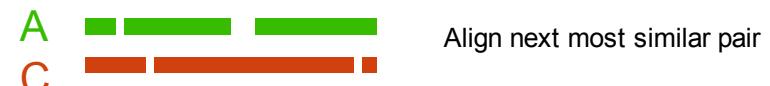
Example – 4 sequences A, B, C, D.



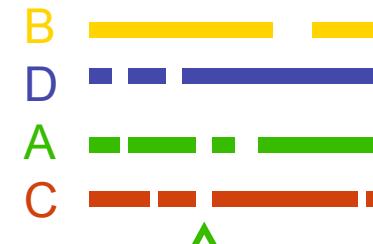
6 pairwise comparisons  
then cluster analysis

## Multiple alignments

Following the tree from step 1



Align B-D with A-C

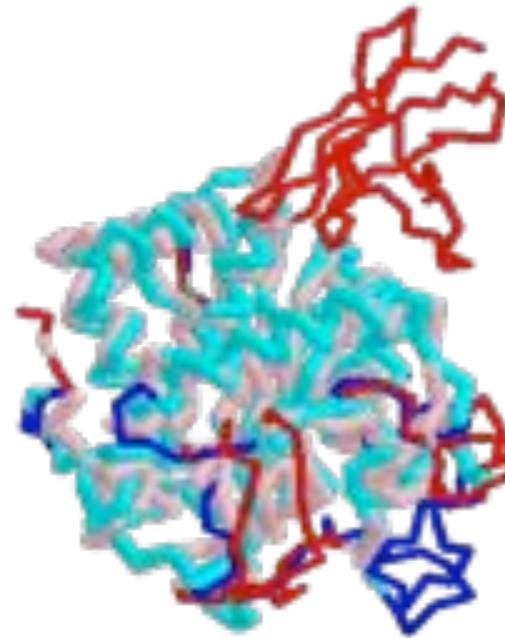


New gap in A-C to optimize  
its alignment with B-D

# Coverage .vs. Accuracy



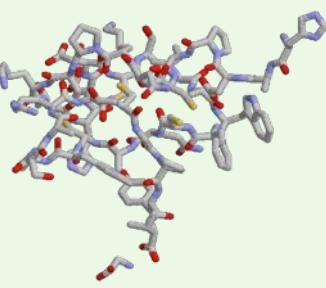
Coverage ~90% C $\alpha$



Coverage ~75% C $\alpha$

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

# Structural alignment by properties conservation (SALIGN-MODELLER)



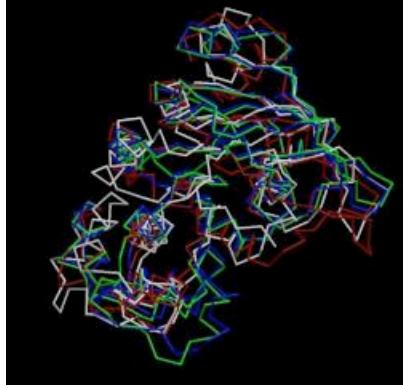


Best score  
Best local alignment

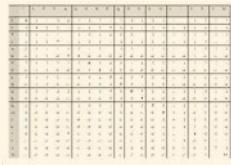
A B C D

B D A C

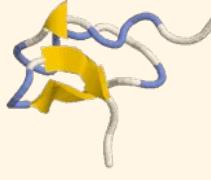
- similarity +

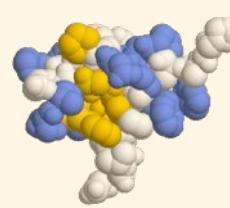


- ✓ Uses all available structural information
- ✓ Provides the optimal alignment
- ✗ Computationally expensive









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

$R_{i,j}$

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

$S_{i,j}$

$B_{i,j}$

$I_{i,j}$

# Structural alignment by properties conservation (SALIGN-MODELLER)

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

The screenshot shows a Microsoft Internet Explorer window displaying the DBAli v2.0 Tools page. The title bar reads "DBAli v2.0 Tools page - Microsoft Internet Explorer". The address bar shows the URL "http://salilab.org/DBAli/?page=tools&action=f\_salign". The main content area is titled "DBAli. Tools associated to the database." and lists several tools:

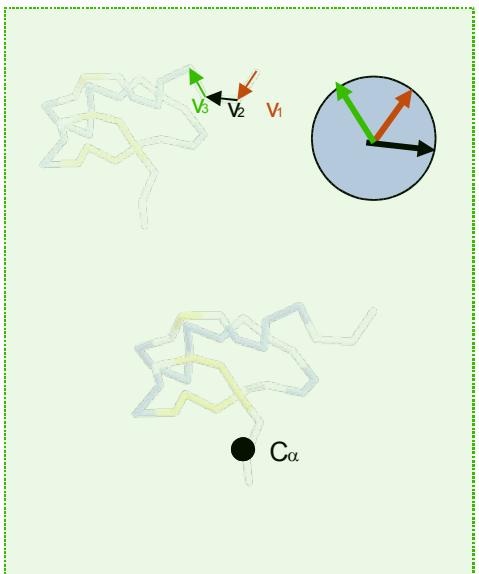
- Cluster a list of chains
- Cluster from a chain
- Define domains from a chain
- Get a multiple structure alignment of a list of chains
- Database statistics
- Download DBAli

Below this list, a sub-section titled "Get a multiple structure alignment of a list of chains." contains a form with the following fields:

File with a list of chains:  
  ?

At the bottom of the page, there is a navigation bar with links: Reference, Download, Statistics, Suggestions, Visitors: 1407, © 2003 - 2004 Marti-Renom.

# Vector Alignment Search Tool (VAST)

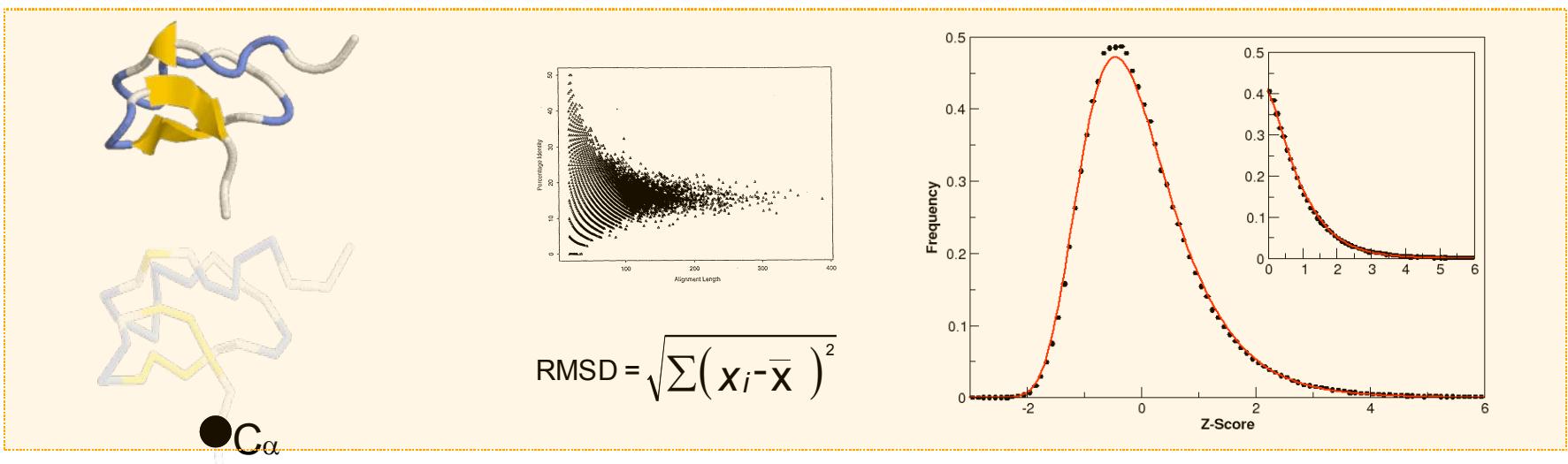


- Graph theory search of similar SSE
- Refining by Monte Carlo at all atom resolution

A screenshot of the NCBI Conserved Domain Database (CDD) interface. The query sequence is phm0089.11, which encodes Phage lysozyme. The results table shows various hits with their E-values and sequence alignments. The interface includes tabs for Published, ModelDB, Protein, Structure, CDD, Taxonomy, and Help.

✓ Good scoring system with significance

✗ Reduces the protein representation

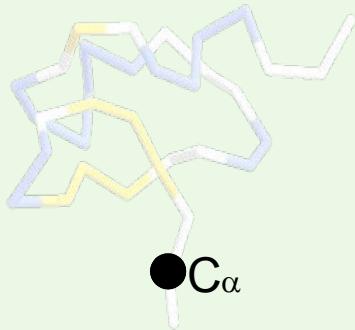


# Vector Alignment Search Tool (VAST)

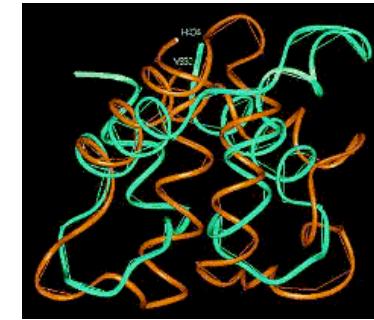
<http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml>



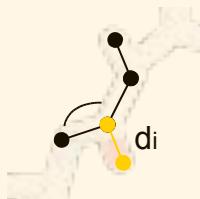
# Incremental combinatorial extension (CE)



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

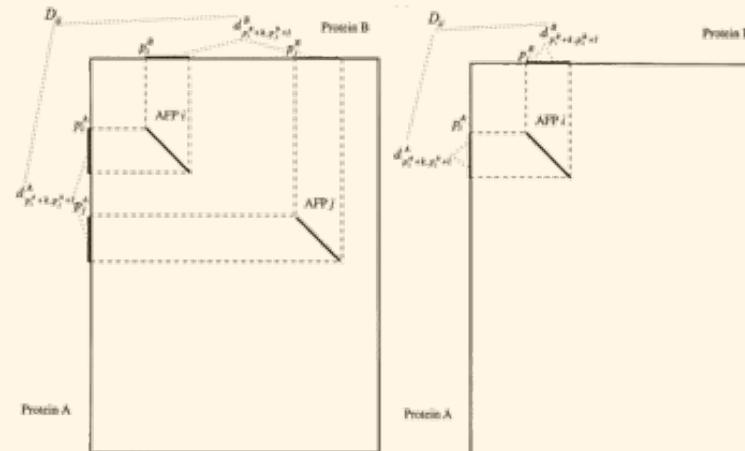


- ✓ FAST!
- ✓ Good quality of local alignments
- ✗ Complicated scoring and heuristics



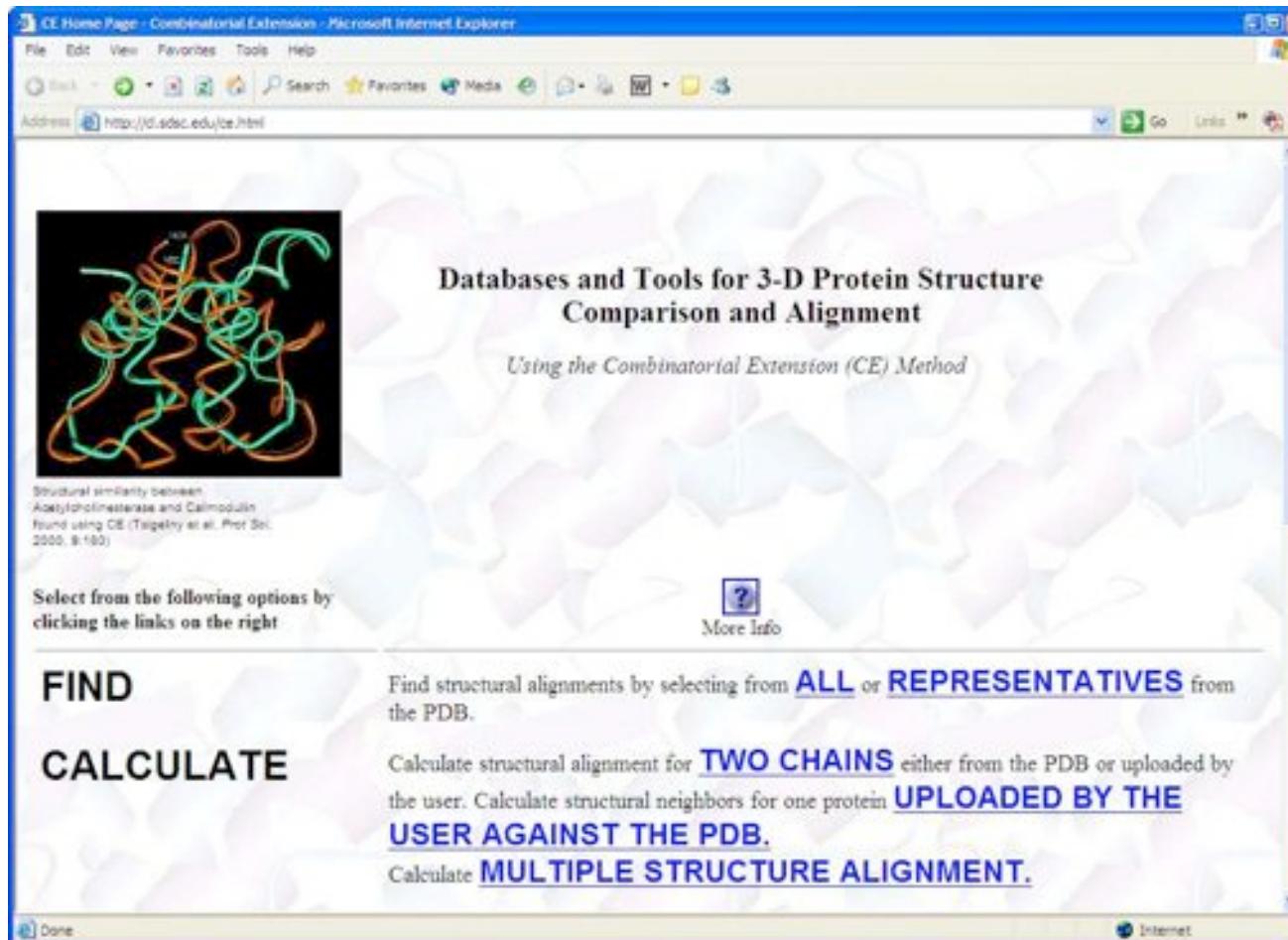
8 residues peptides

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

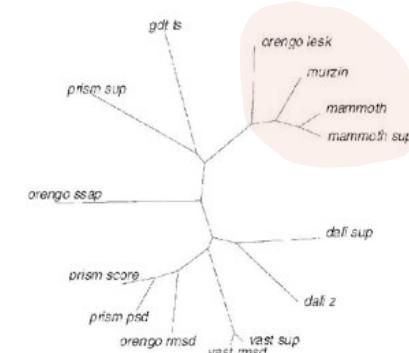
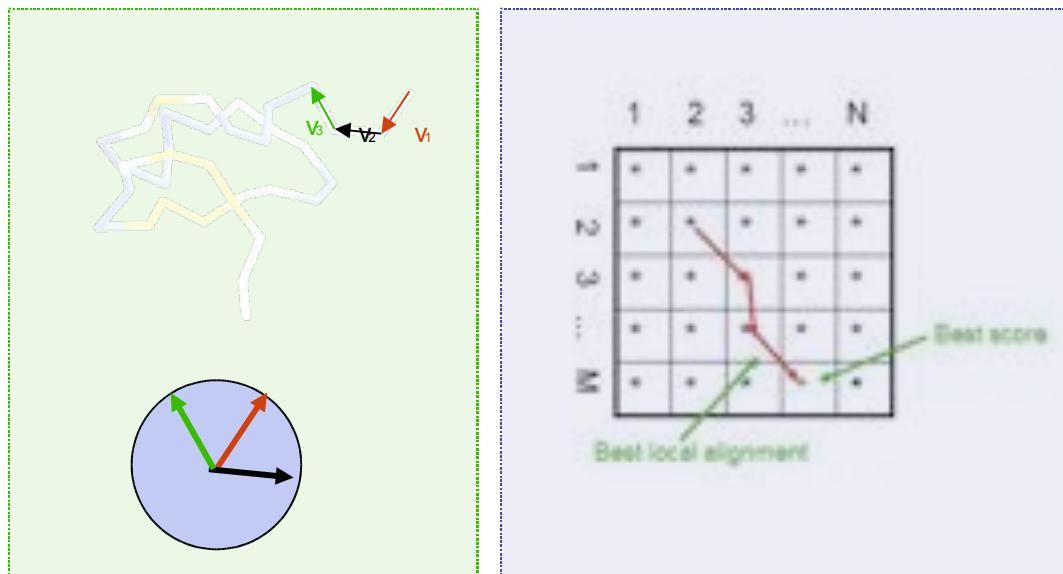


# Incremental combinatorial extension (CE)

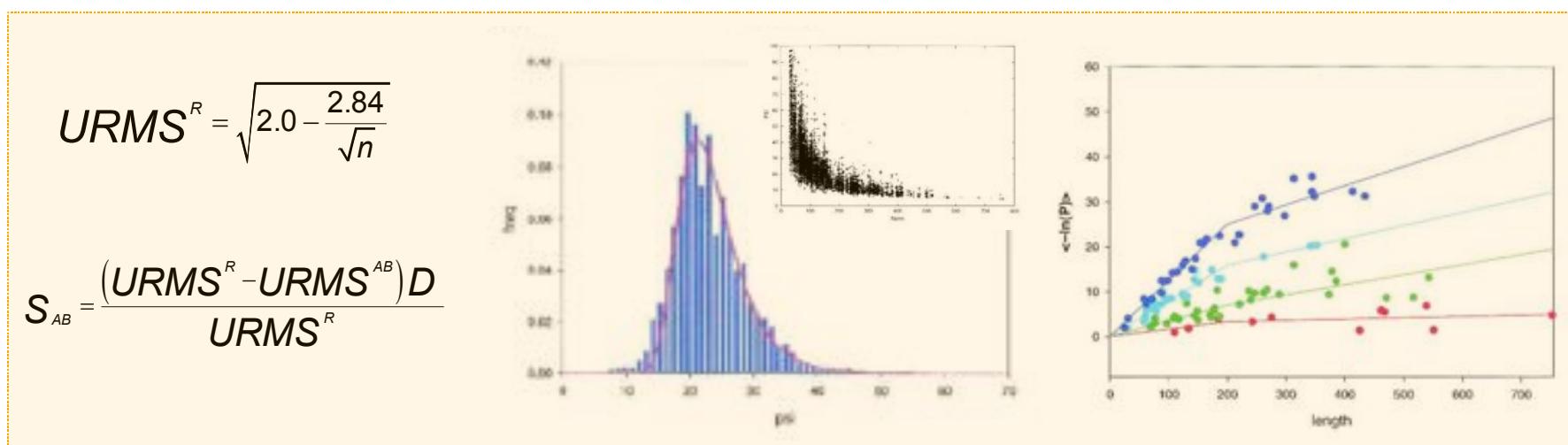
<http://cl.sdsc.edu/ce.html>



# Matching molecular models obtained from theory (MAMMOTH)

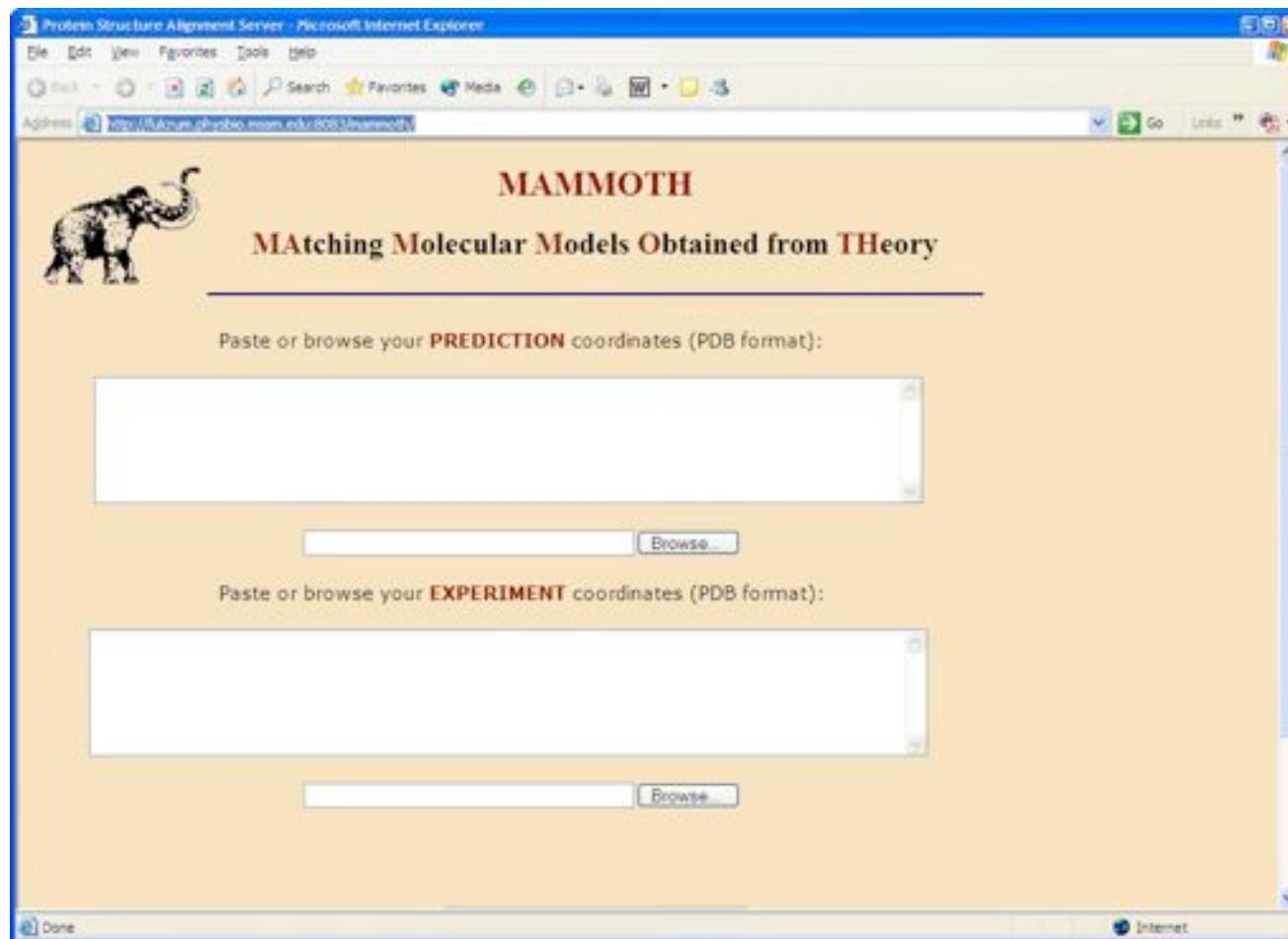


- ✓ VERY FAST!
- ✓ Good scoring system with significance
- ✗ Reduces the protein representation



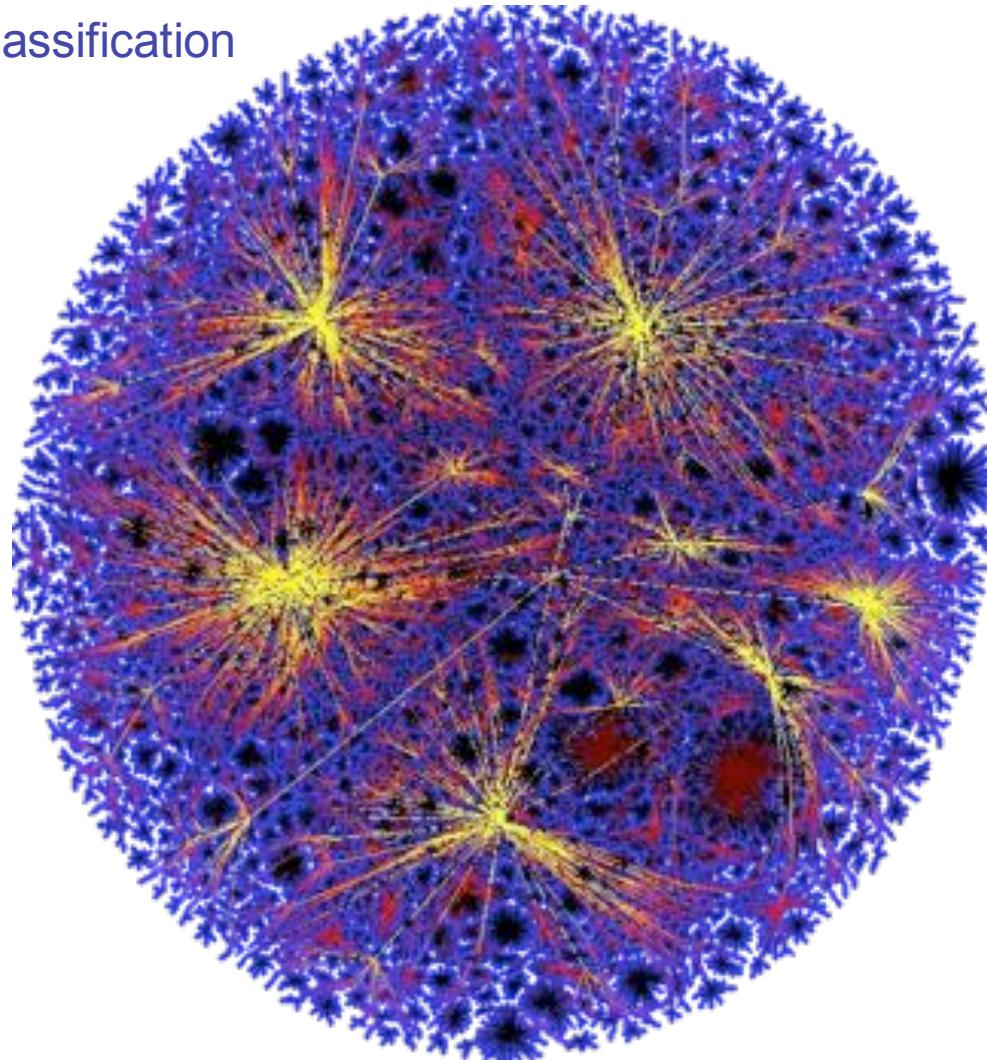
# Matching molecular models obtained from theory (MAMMOTH)

<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/>

The screenshot shows the SCOP 1.65 database homepage. At the top, there's a navigation bar with links like Home, Help, Back, Forward, Stop, Refresh, Favorites, Tools, and Help. The address bar shows the URL. Below the bar, the title "Structural Classification of Proteins" is displayed, along with three small icons: a magnifying glass, a pencil, and a question mark. The main content area has a heading "Structural Classification of Proteins". It includes a welcome message for the 1.65 release (December 2003), information about PDB entries (20619 total), literature references (54745), and domain counts (54745). It also lists authors (Murzin et al., 1995; Brenner et al., 2002; Andreeva et al., 2004), major changes (described in J. Mol. Biol. 247: 536-540), and a PDF version of the paper. A section on "Major changes" details structural refinements and genomic integration. Below this, there's a "Access methods" section with links to various file formats (top of hierarchy, keyword search, parseable files, previous releases, domain sequences, ASTRAL), a hidden Markov Model library, and online resources. A note about mirrors follows. The "News" section contains a list of updates from August 2003, mentioning the inclusion of all PDB entries, reclassification changes, and new identifiers. At the bottom, there's a "Help and Information" link.

- ✓ 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.

02/15/04

# CATH 2.5.1 database

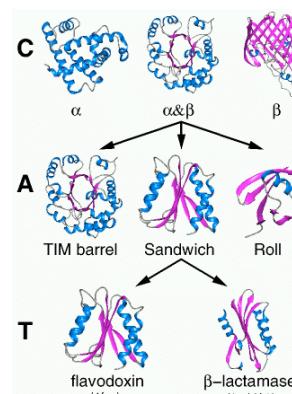
<http://www.biochem.ucl.ac.uk/bsm/cath/>

The screenshot shows the CATH 2.5.1 database homepage. It features a search bar with options for PDB Code, CATH Code, and General Text. A 'Goto' section includes links for SOAP Server, GRANTH Server, DHS, and Gene3D. The main content area displays the 'CATH Protein Structure Classification' and 'Version 2.5.1: Released January 2004'. It lists authors: Dr. Frances M.G. Pearl, Dr. Ian Sillitoe, Dr. Mark Dibley, Prof. Janet Thornton, and Prof. Christine A. Orengo. An 'Options' section provides links to browse or search the classification, CATH statistics, general information, CATH lists, and FTP site. The 'DHS - Dictionary of Homologous Superfamilies' is also mentioned. Below this is an 'Introduction' section with a detailed explanation of the CATH classification levels (Class, Architecture, Topology, and Homologous superfamily) and how they are assigned. A reference section cites work by Orengo et al. (1997) and Pearl et al. (2000). Other CATH contributors are listed at the bottom.

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

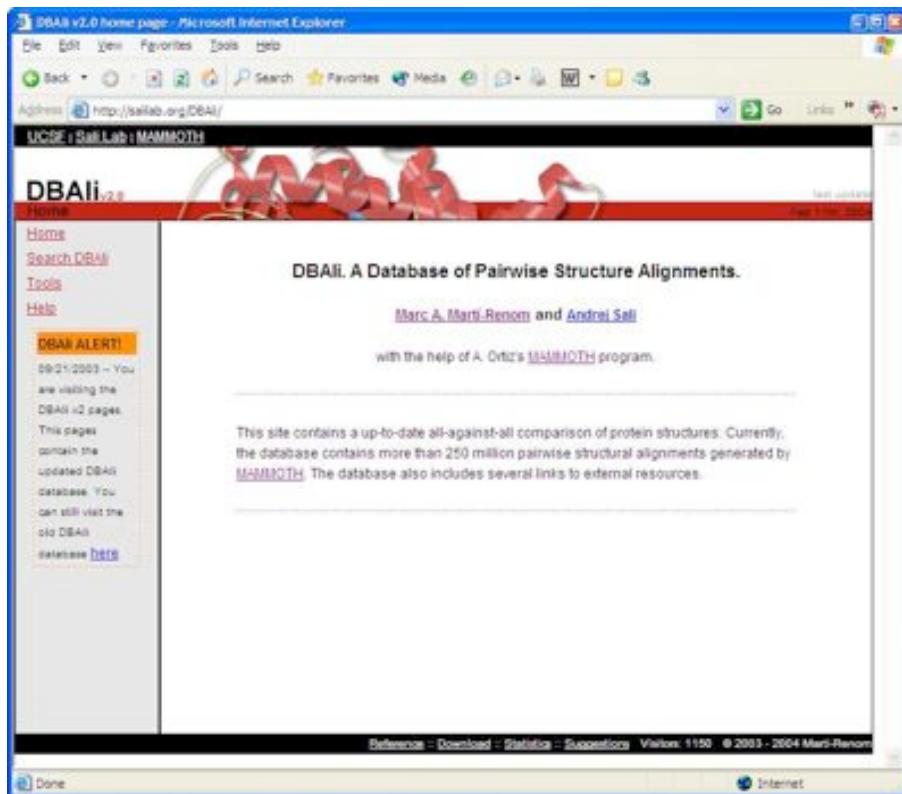


Version	2.5.1
Date	28-01-2004
	(1) (2) (3) (4) (5) (6) (7) (8) (9) (D)
Mainly Alpha	5 227 429 948 1713 3946 10155
Mainly Beta	19 139 292 951 2344 5011 14259
Alpha Beta	12 368 648 2010 3631 8639 23025
Few Secondary Structures	1 86 91 114 225 378 952
Multi-domain chains	1 1053 1057 1071 2186 5801 12471
Preliminary single domain assignments	1 371 374 422 479 785 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 66 106

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:

Number of chains in database:  
Number of structure-structure  
comparisons:

February 11th, 2004

(18:49h)

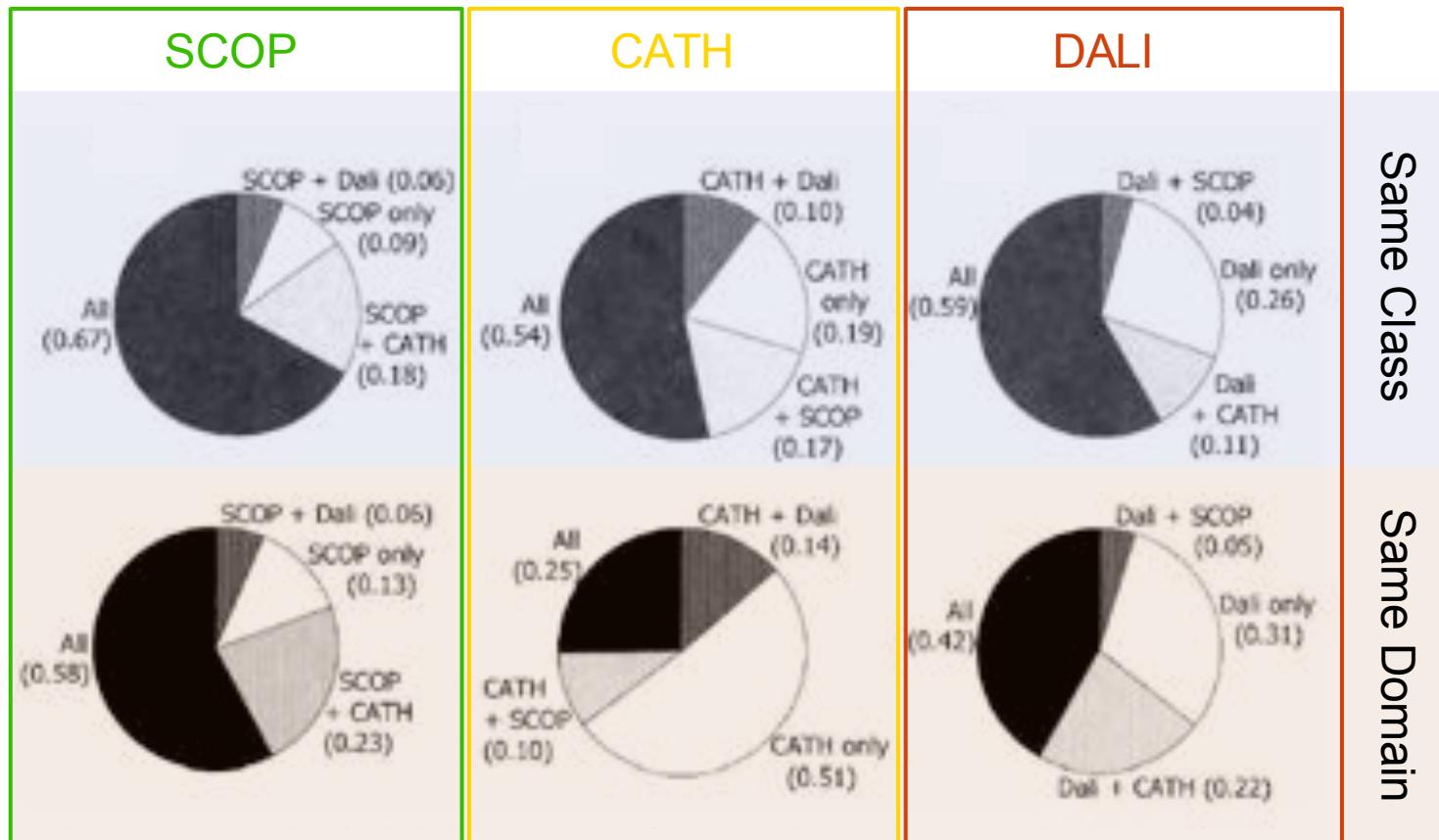
48,094

330,514,636

# Classification of the structural space

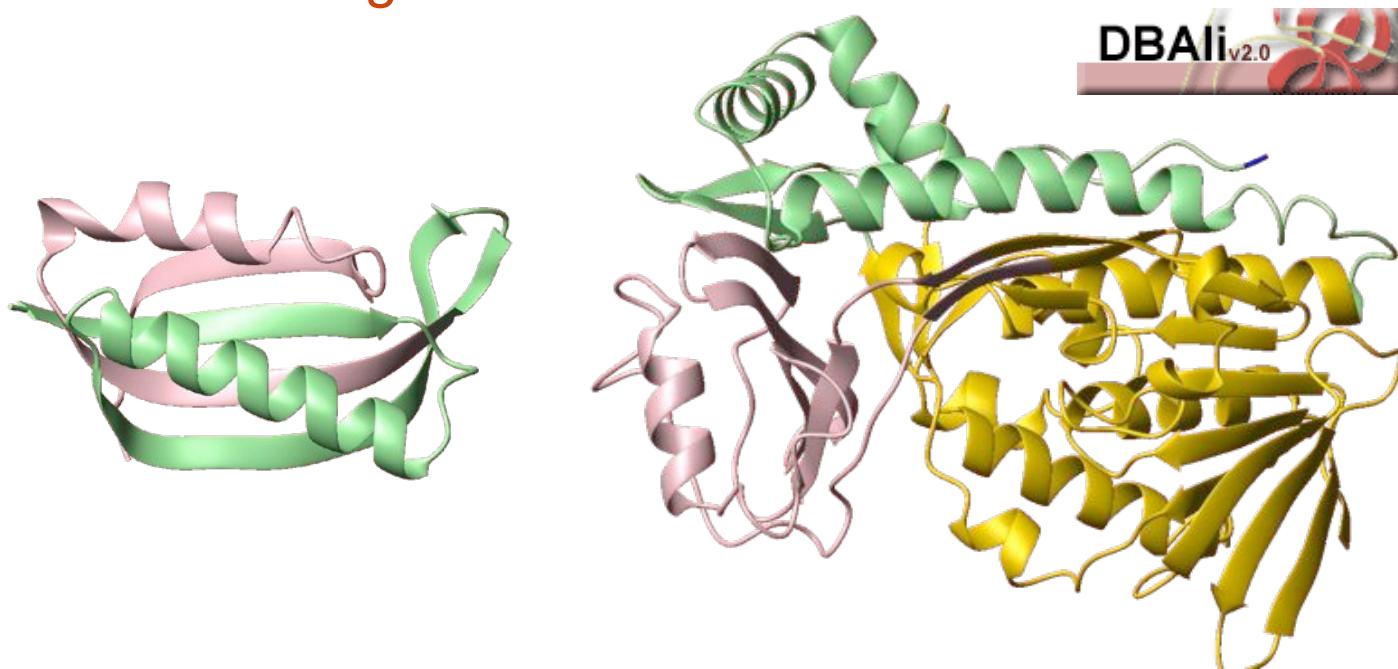
*Not an easy task!*

Domain definition AND domain classification

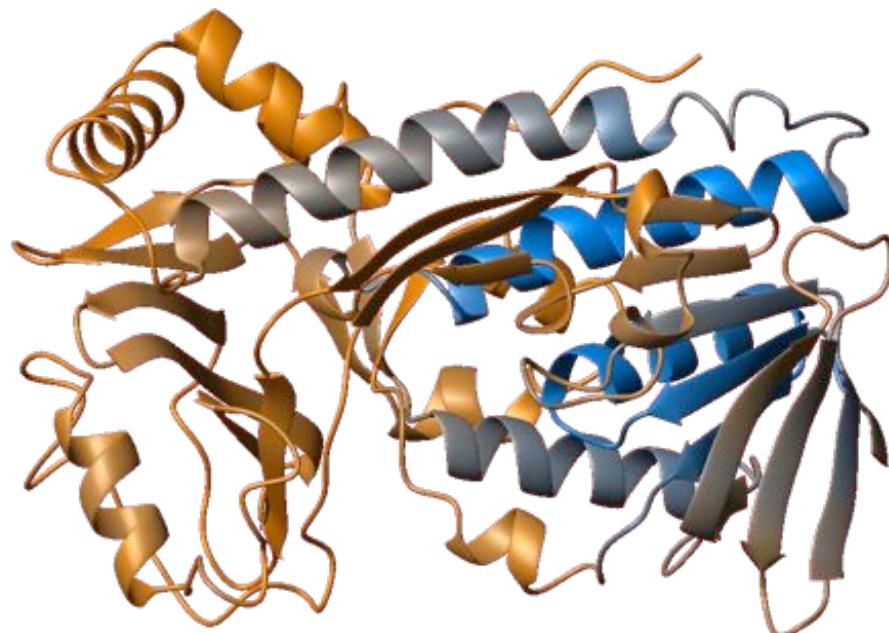


# Application (ModDom)

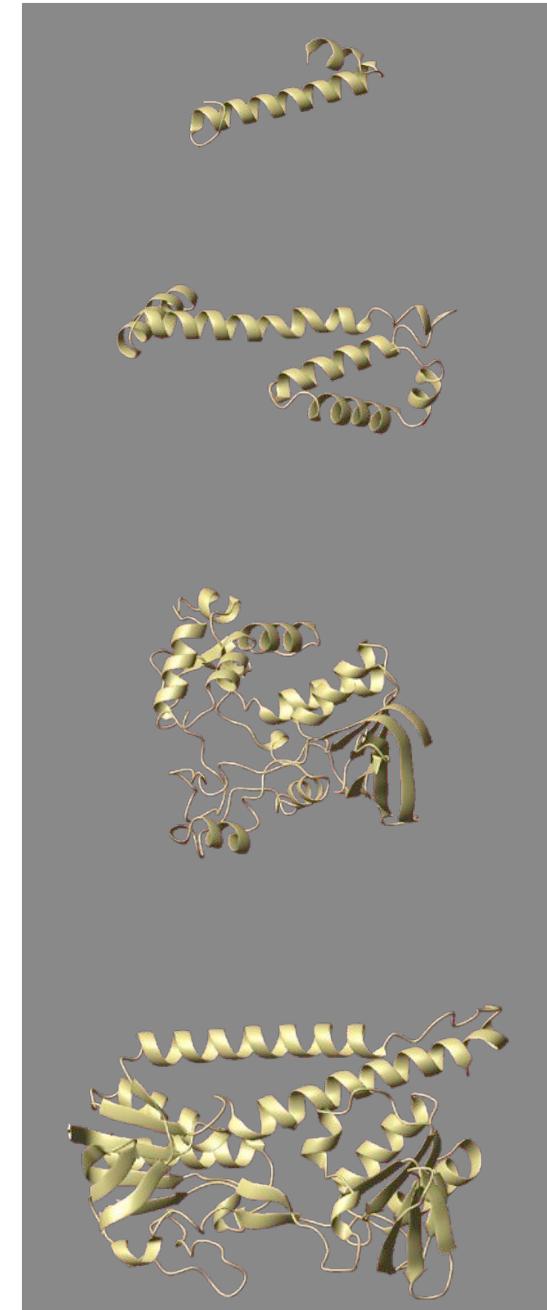
- Use of the DBAli data to define...
  - Protein Domains
  - Protein Fragments

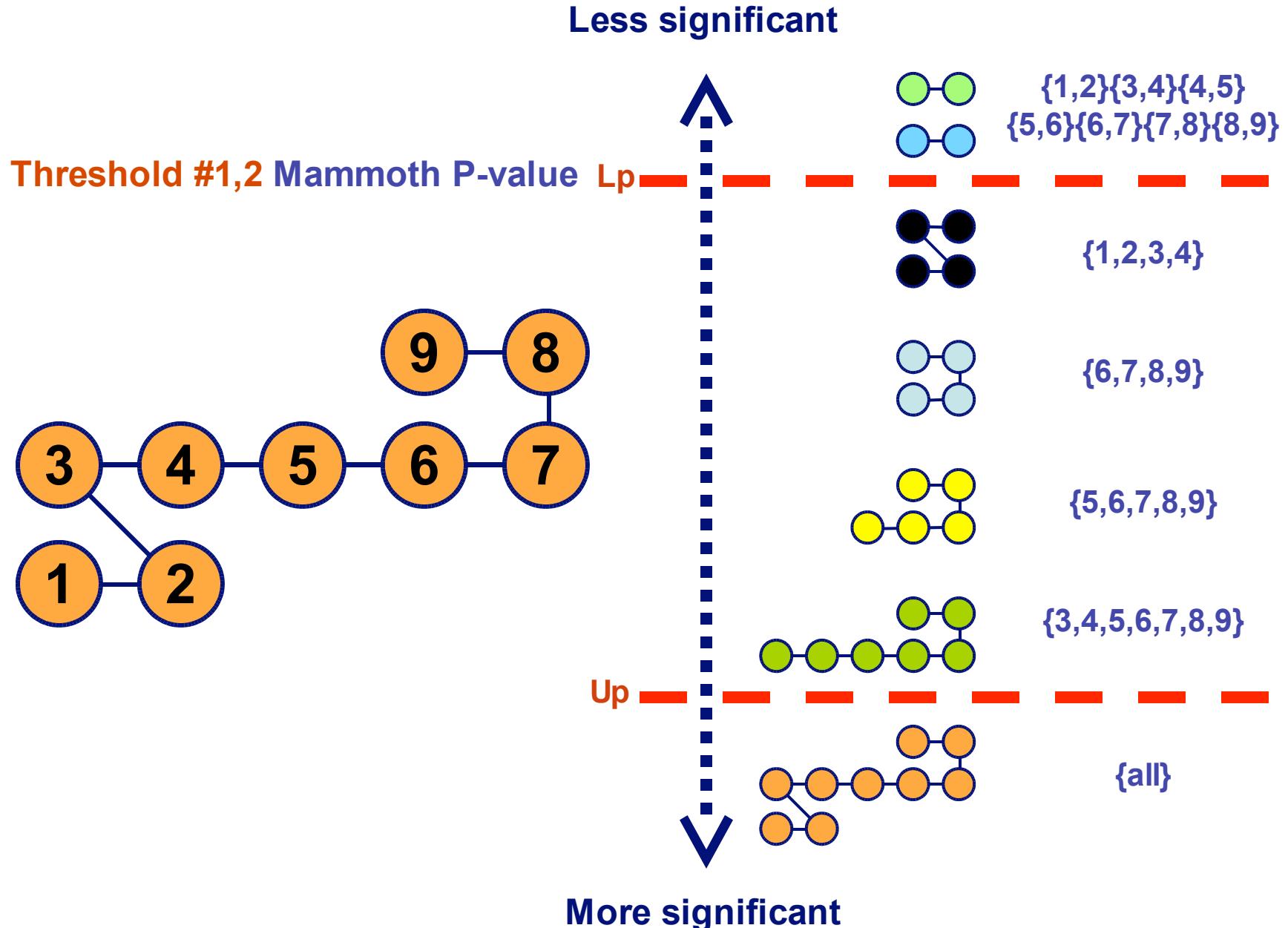


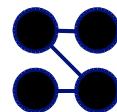
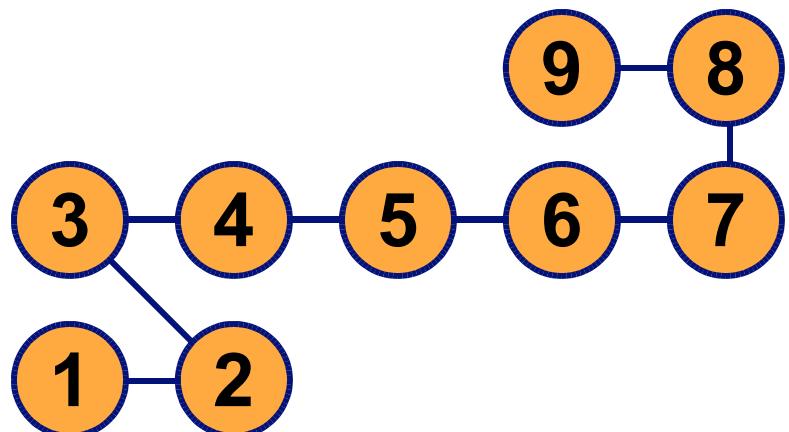
# ModDom



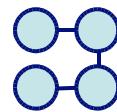
1phh (Oxydoreductase from *Pseudomonas fluorescens*)



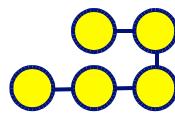




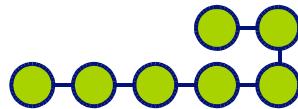
{1,2,3,4}



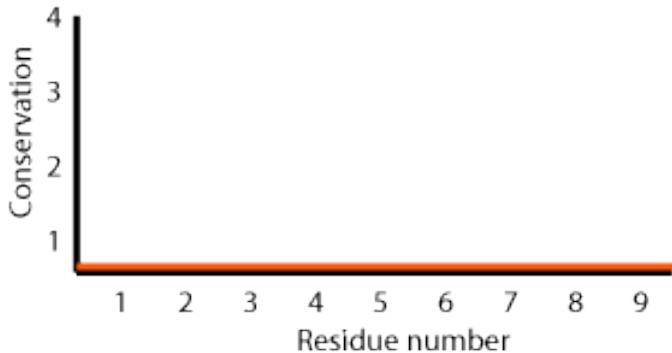
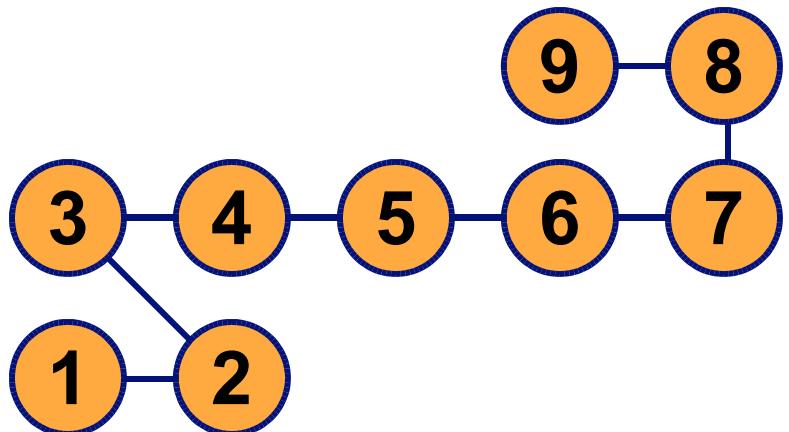
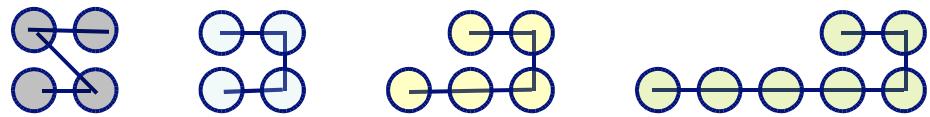
{6,7,8,9}



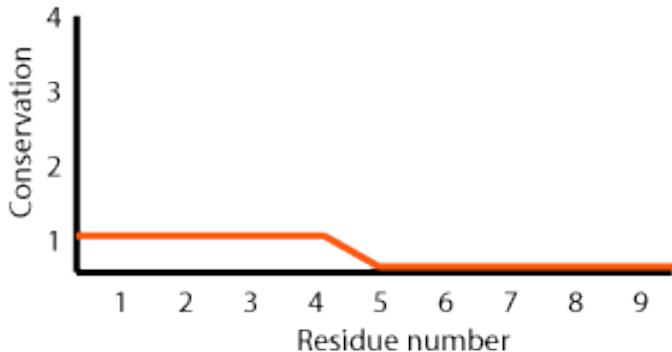
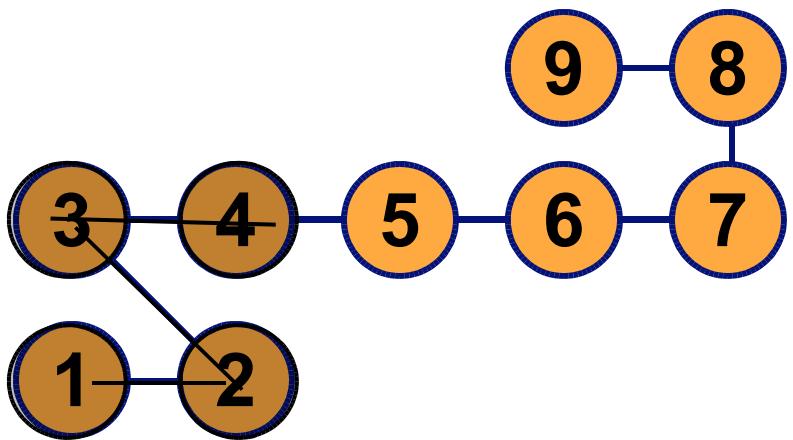
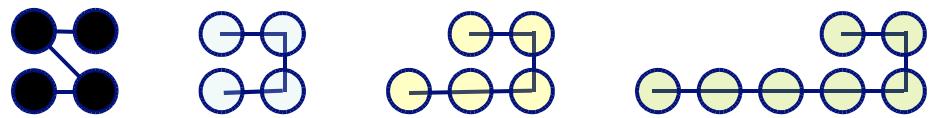
{5,6,7,8,9}



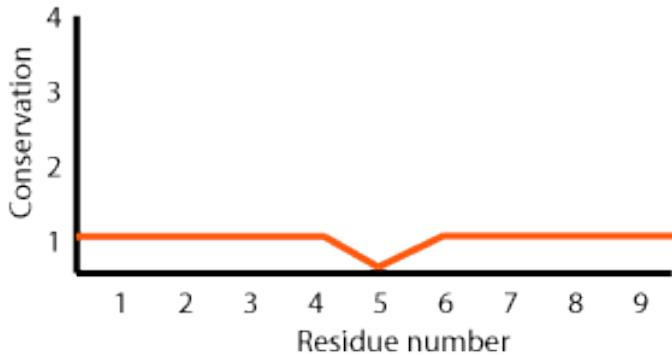
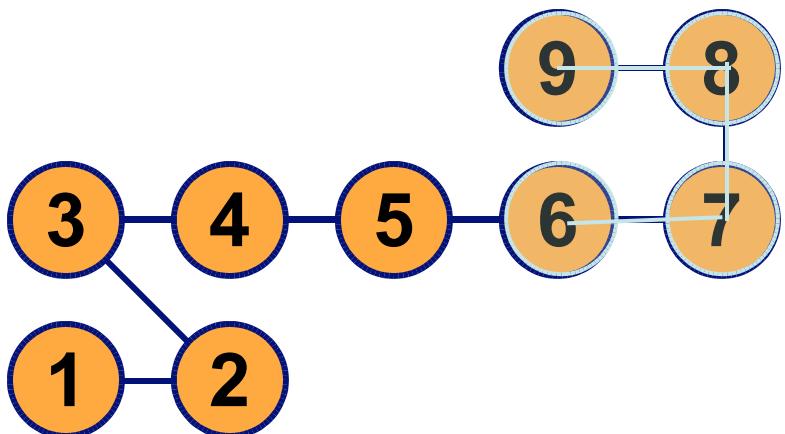
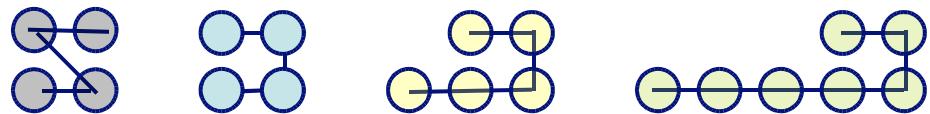
{3,4,5,6,7,8,9}



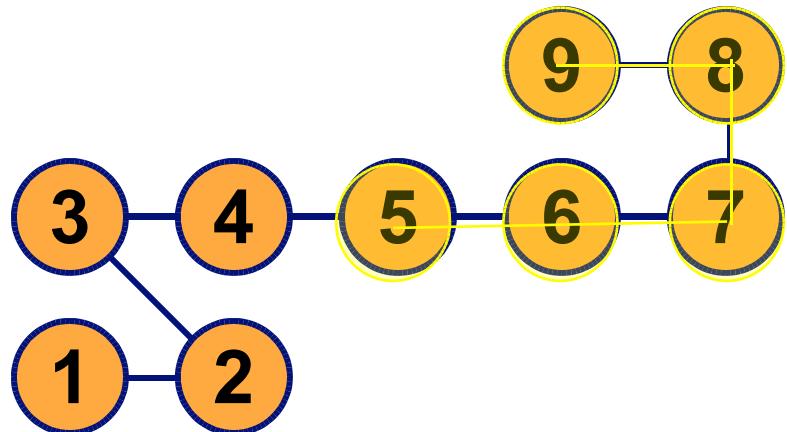
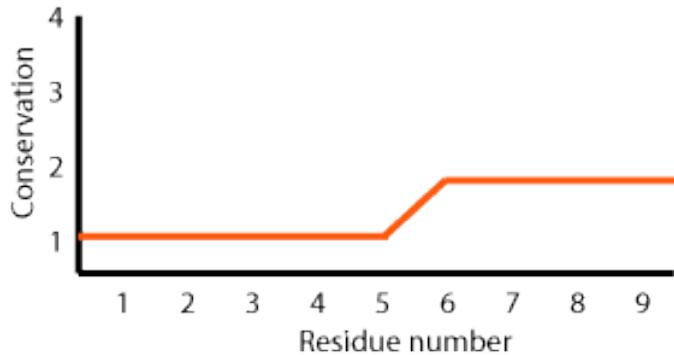
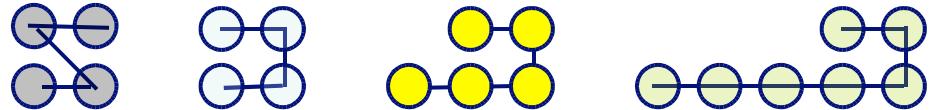
#	1	2	3	4	5	6	7	8	9
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0



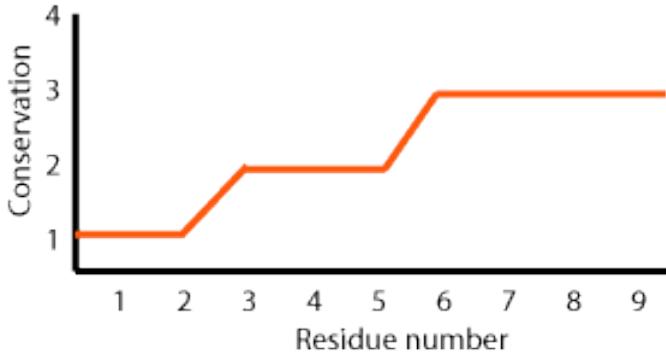
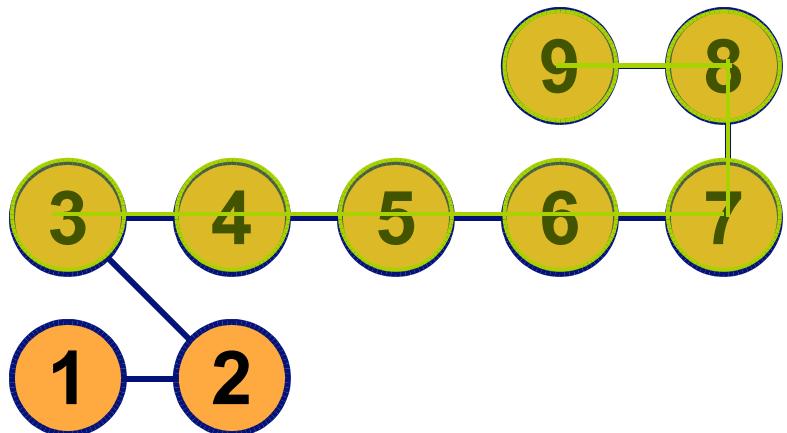
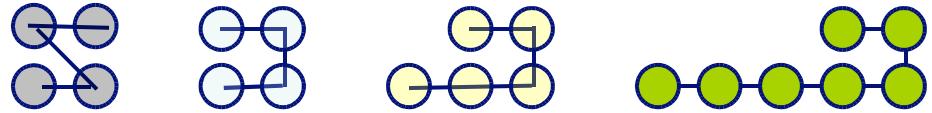
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0



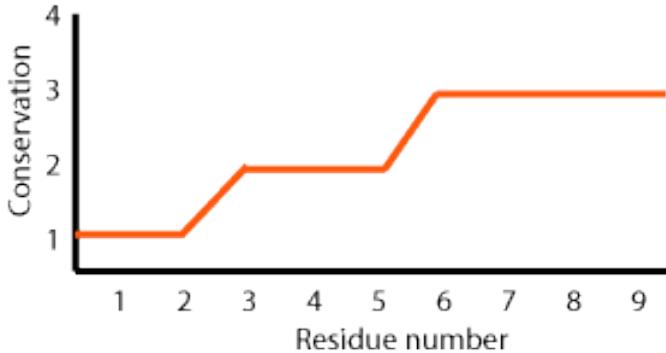
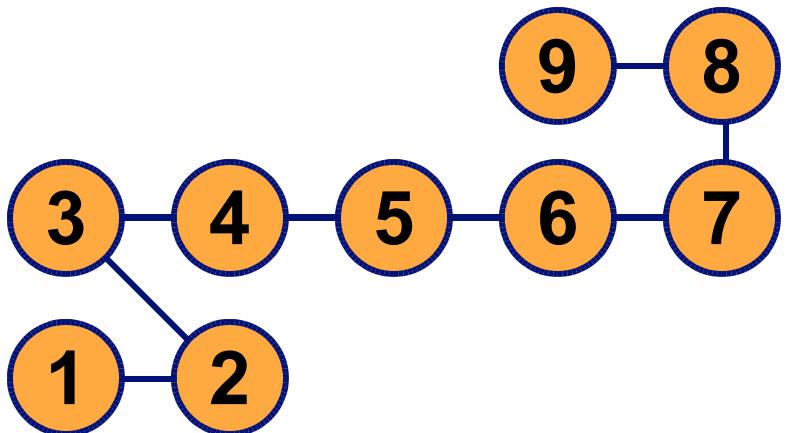
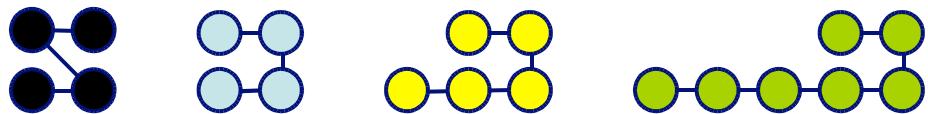
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	1	1	1
7	0	0	0	0	0	1	1	1	1
8	0	0	0	0	0	1	1	1	1
9	0	0	0	0	0	1	1	1	1



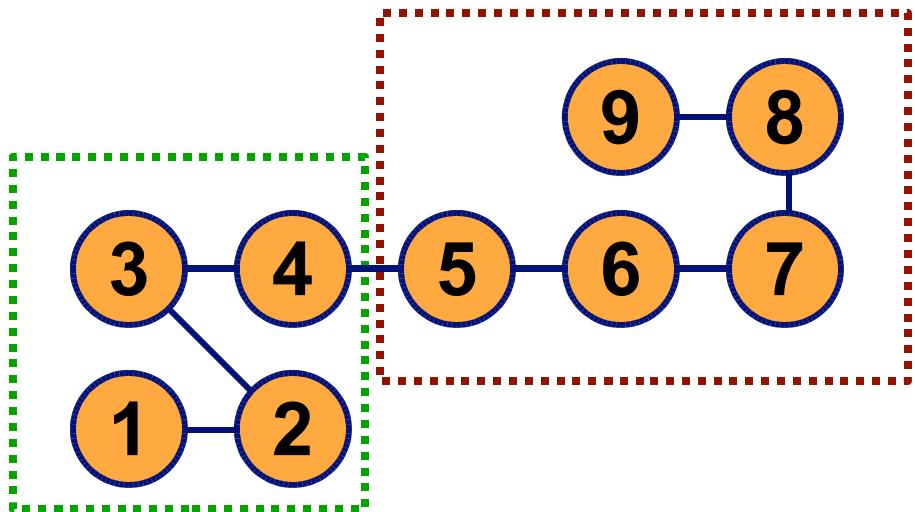
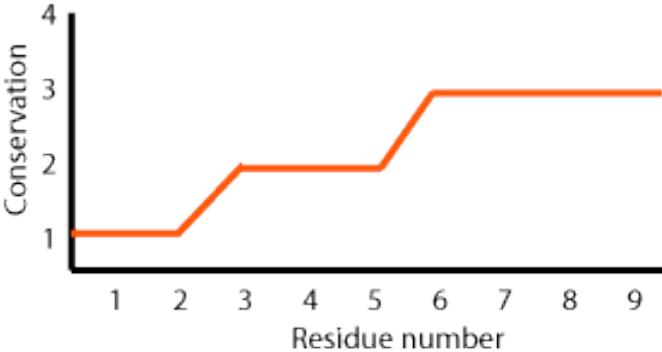
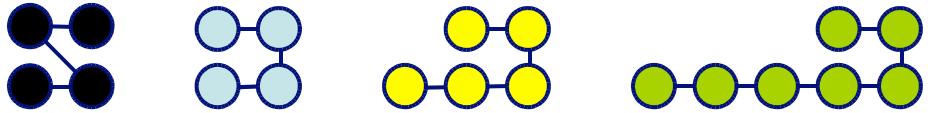
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	1	1	1	1	1
6	0	0	0	0	1	2	2	2	2
7	0	0	0	0	1	2	2	2	2
8	0	0	0	0	1	2	2	2	2
9	0	0	0	0	1	2	2	2	2



#	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

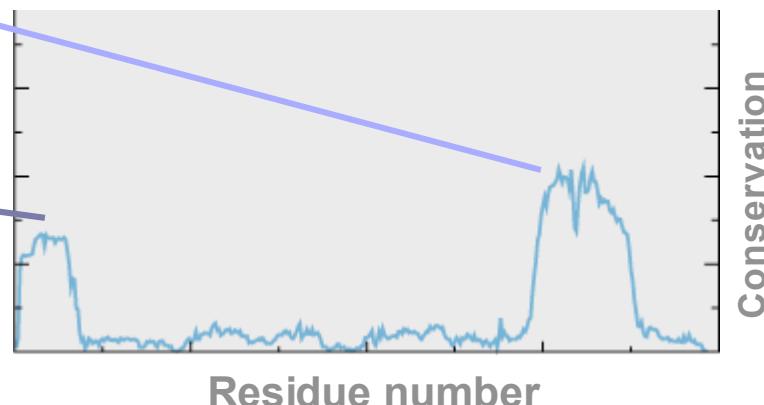
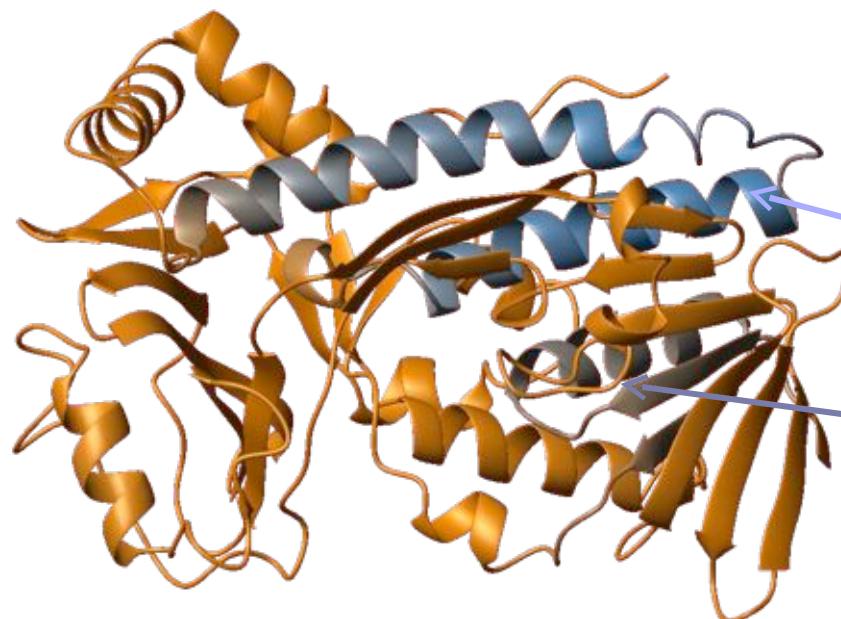
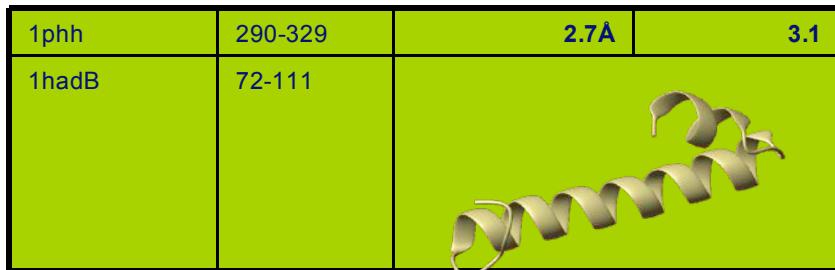
## Threshold #3 MCL Cluster level (-l)

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

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

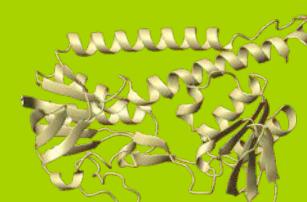
**Threshold #3 → Cluster Level (-l)**  
**Low -l cluster value → fewer partitions**

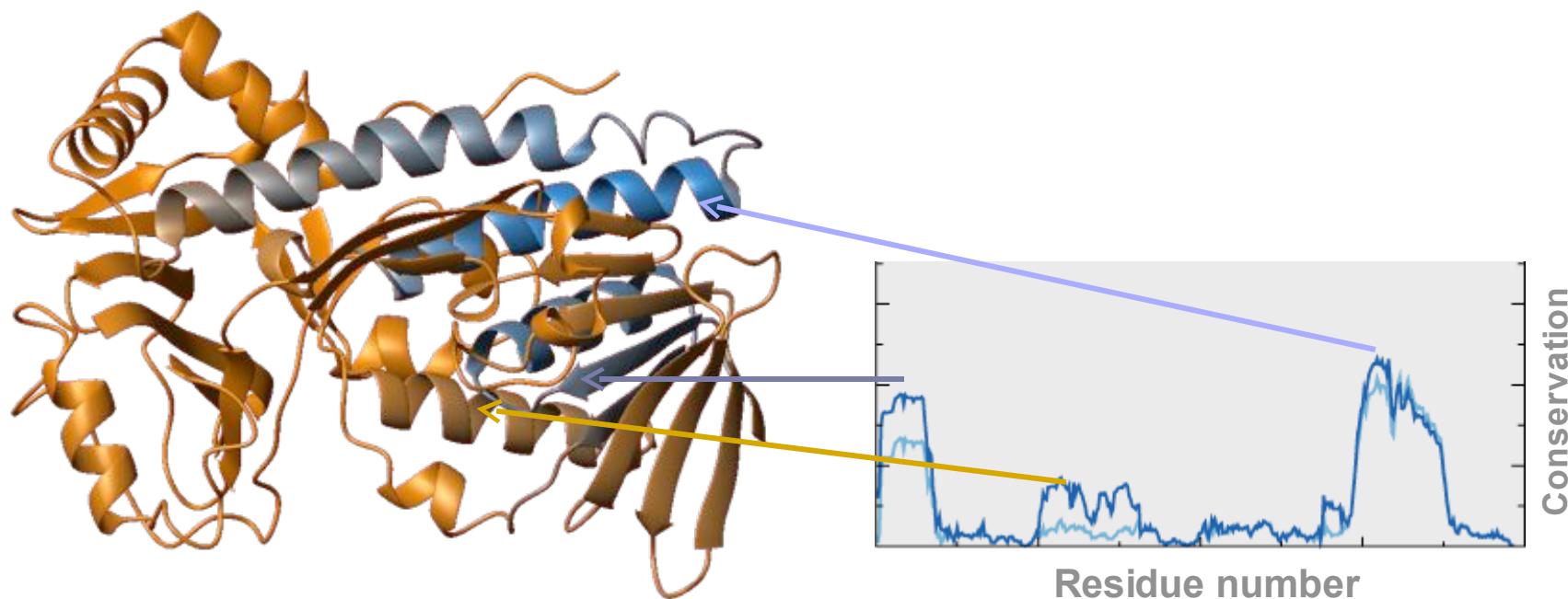
**Applied to the ~45,000 chains in PDB (Dec 2003)**



**1phh** (Oxydoreductase from *Pseudomonas fluorescens*)

1phh	1-213	3.0 Å	8.1
1qjda	125-379		

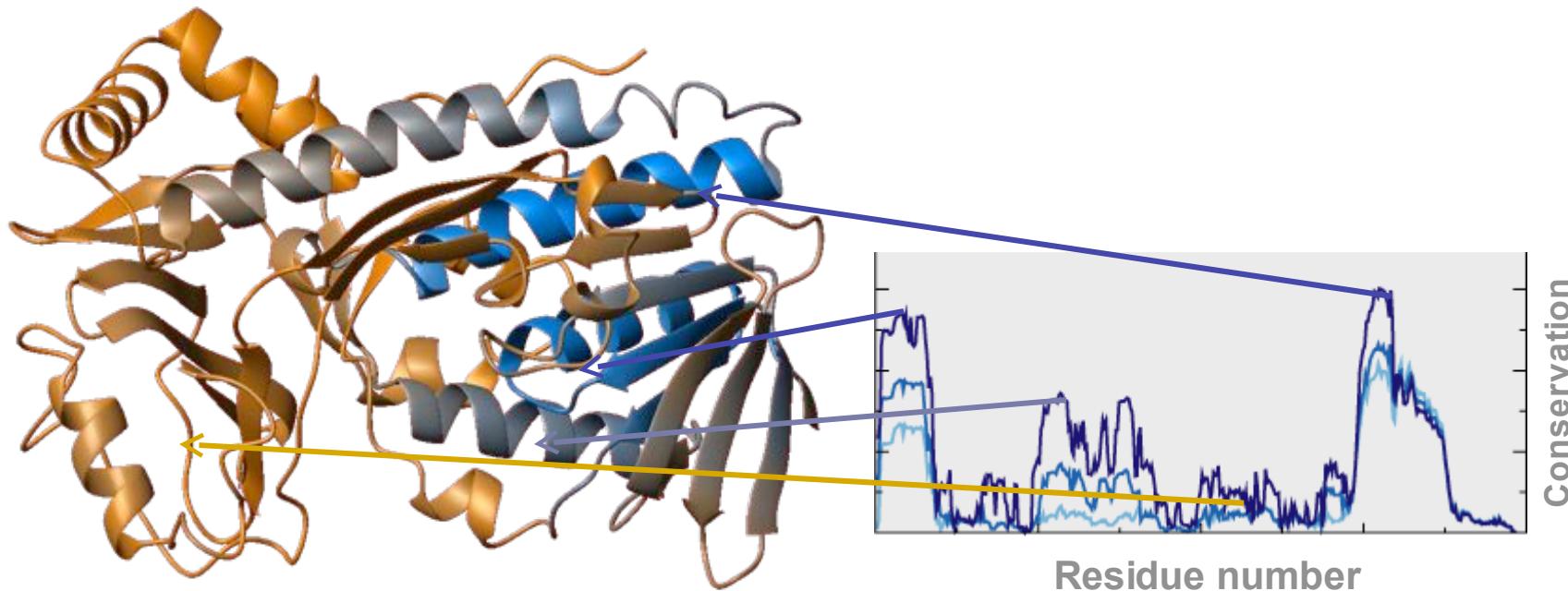
1phh	1-319	3.6 Å	9.8
1gerA	3-327		



## 1phh (Oxydoreductase from *Pseudomonas fluorescens*)

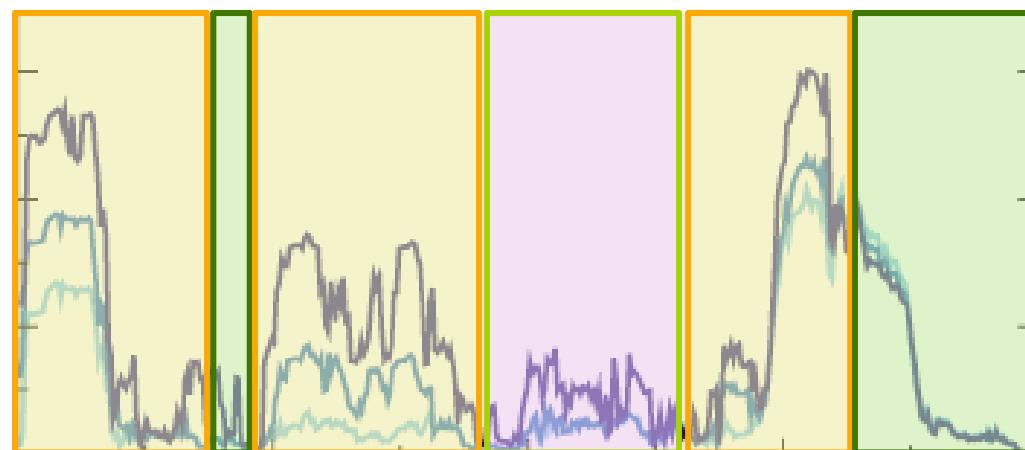
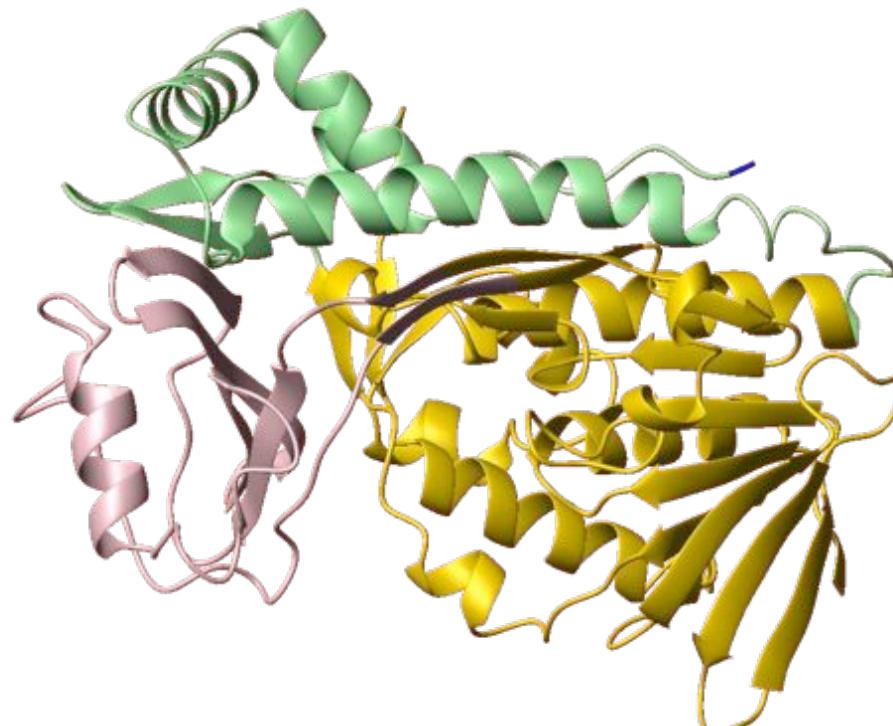
1phh	1-378	3.8Å	10.3
1feaC	2-464		

1phh	1-316	3.8Å	17.2
1I9dB	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:**

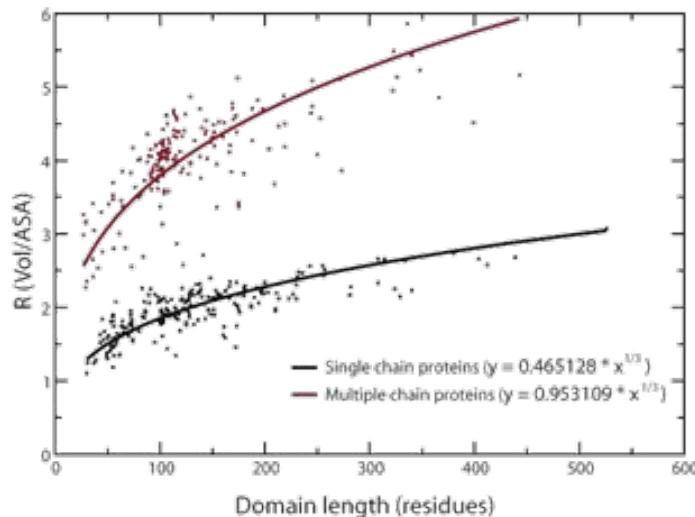
**Training set → 242 chains**

**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**

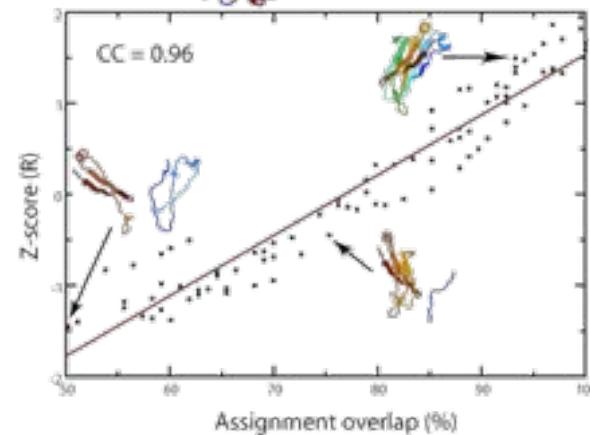
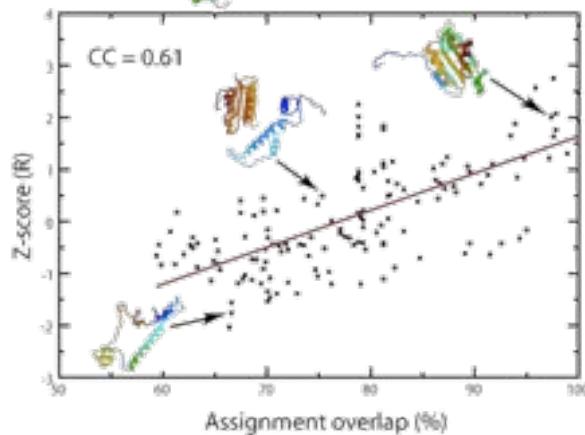
# $R = \text{Volume/ASA}$



1lvj  
3 domains protein



8fabA  
2 domains protein



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

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

-0.11

5-46

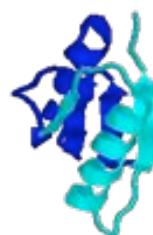


-0.10



-0.08

1-84



-0.09

47-84



85-192



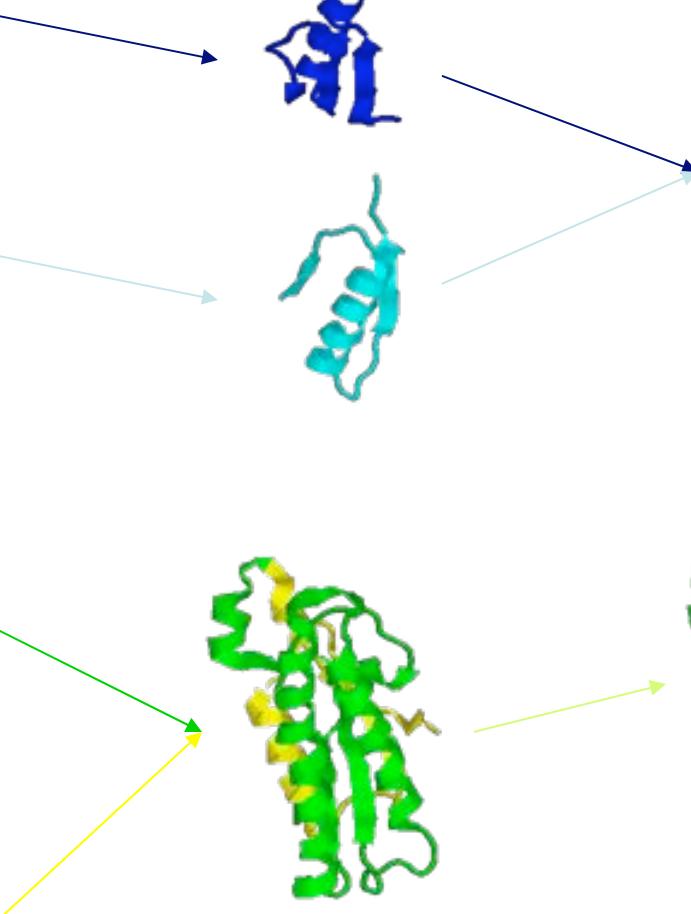
193-239



85-239

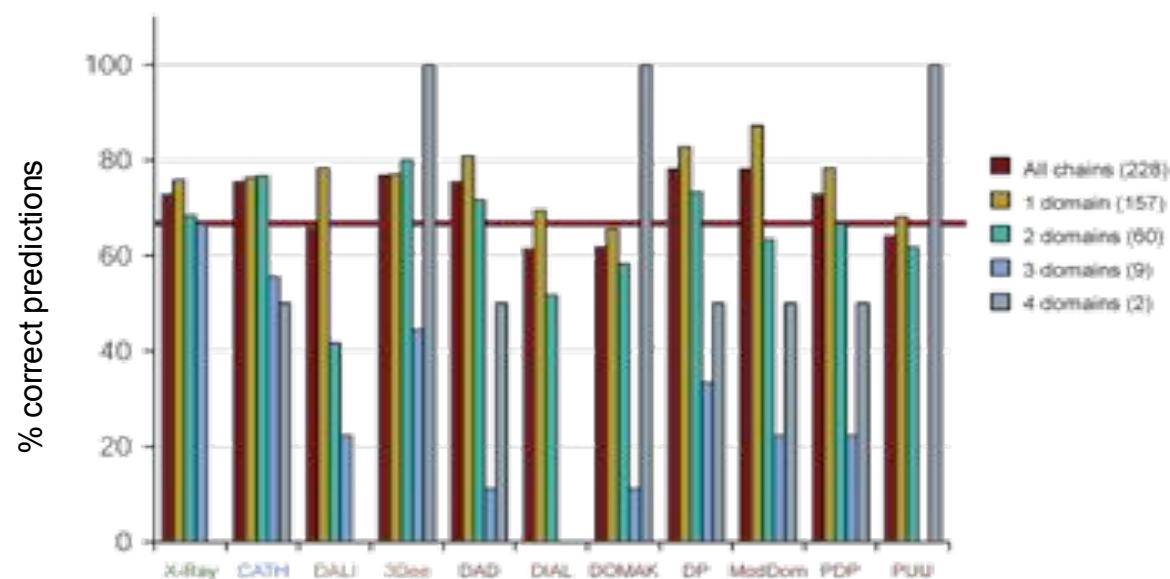
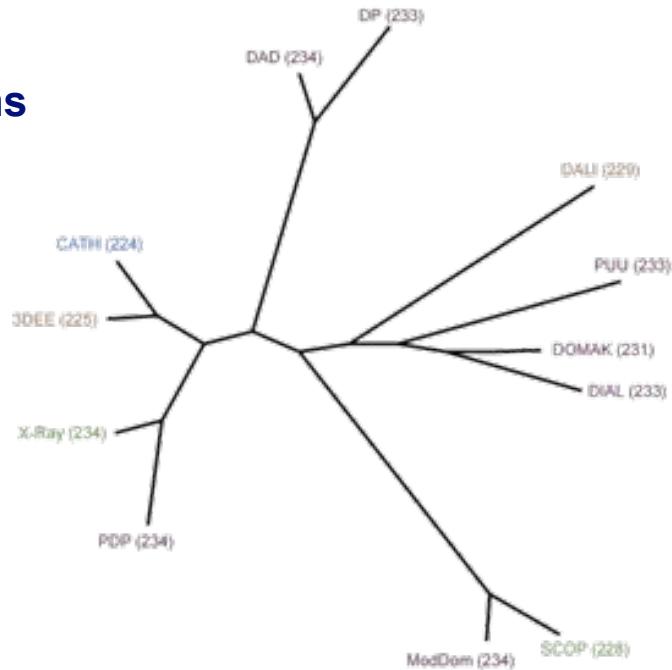


**1dhr\_ (dihydropteridine reductase )**



1-239

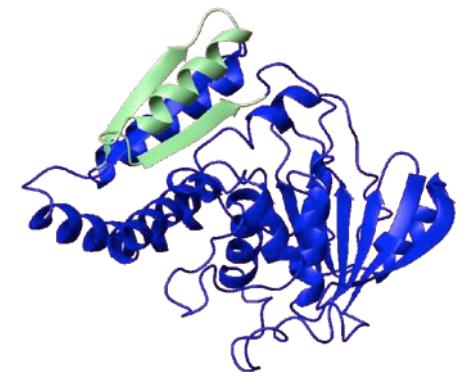
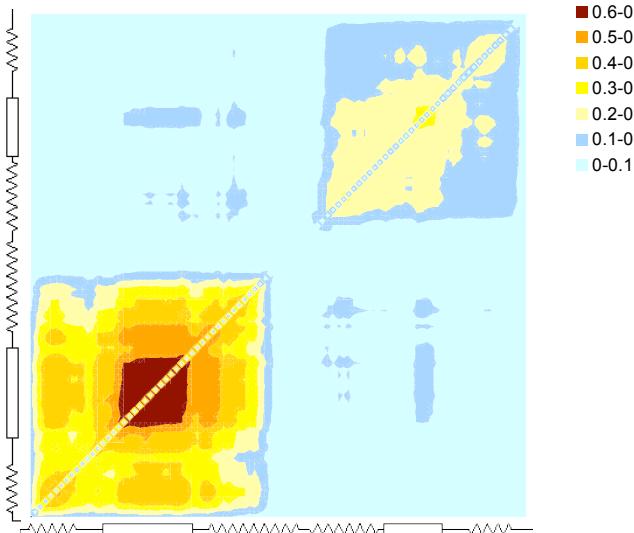
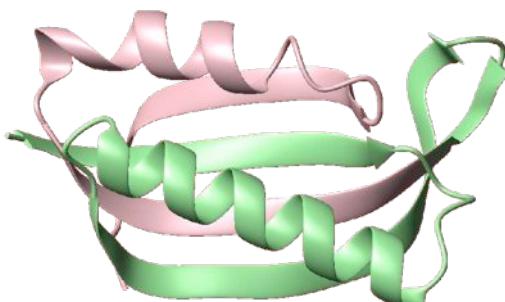
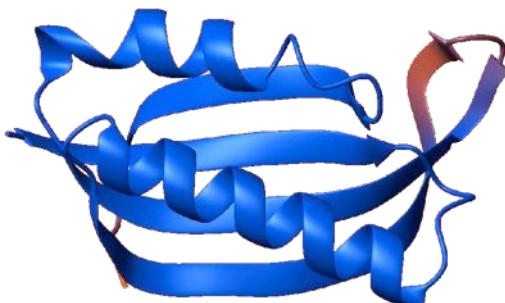
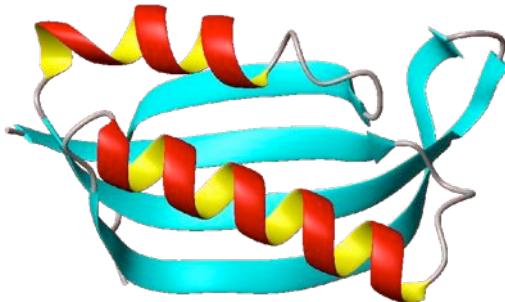
## Non-redundant 234 chains



# Fragments assignment from structure

Repetitions  
Swapping  
Complementarities

# Ribosomal protein S6 (1ris) $\alpha+\beta \rightarrow$ Ferrodoxin Like domain

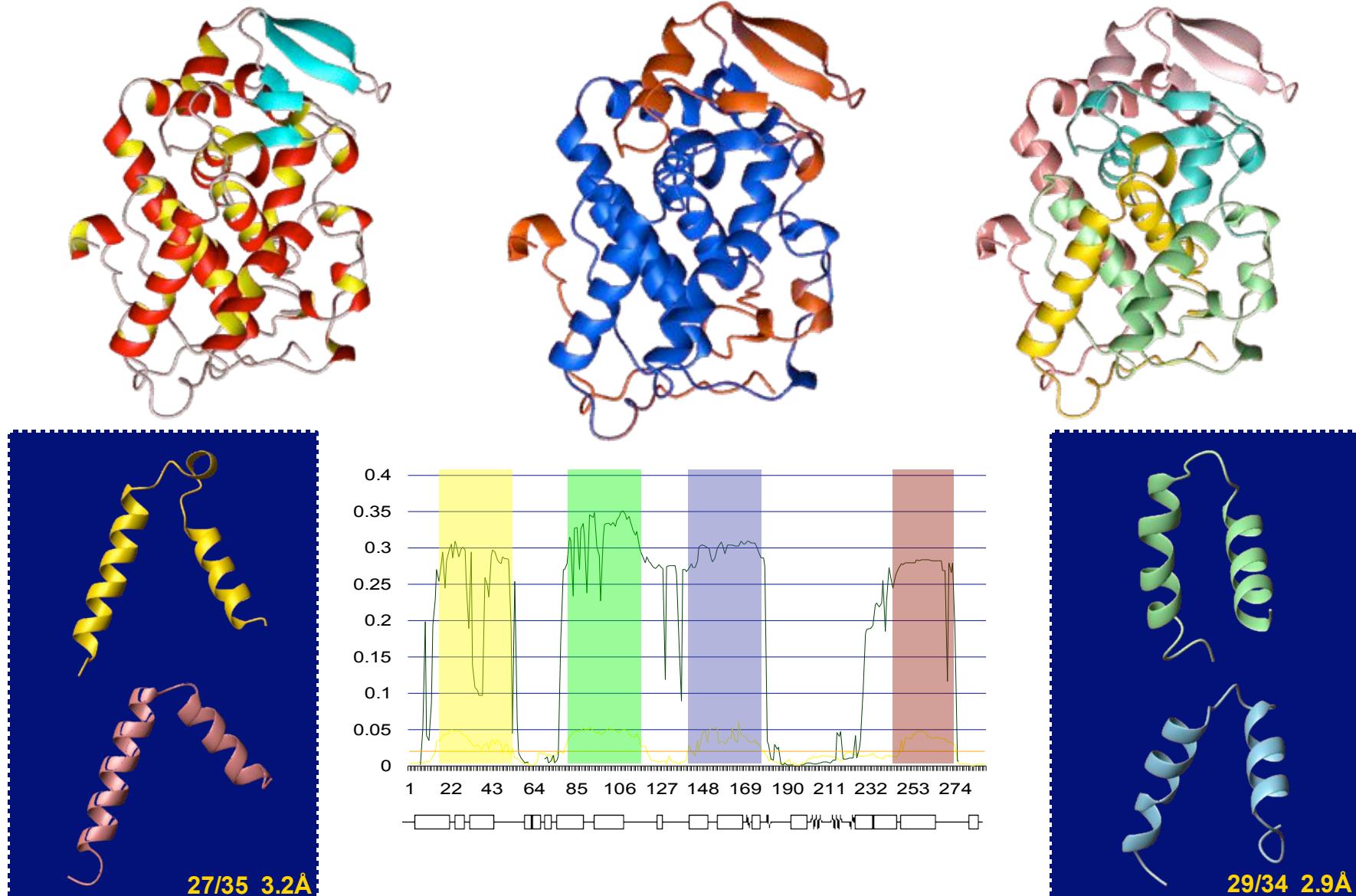


**1ee9A** 17.9% id. 2.3 $\text{\AA}$

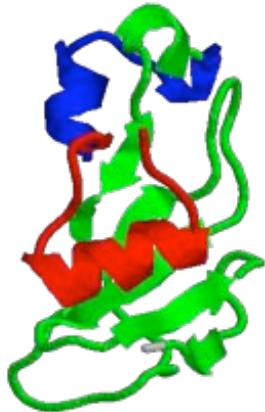


**6timB** 11.1% id. 2.6 $\text{\AA}$

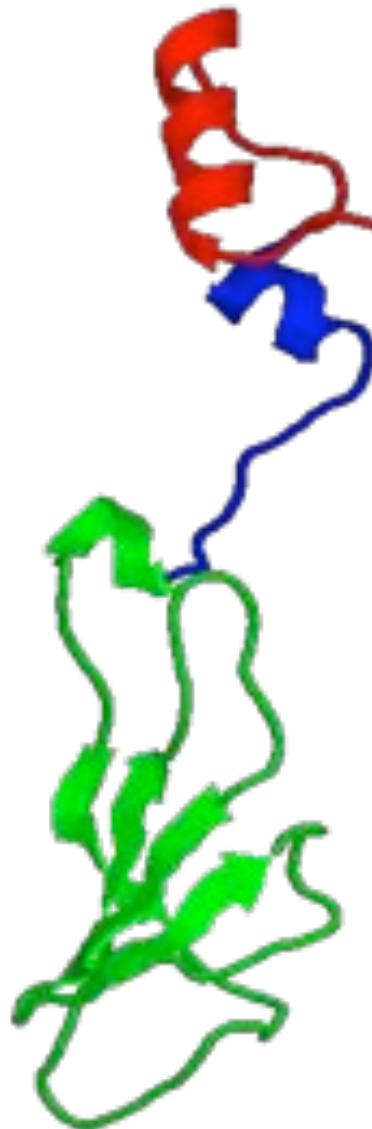
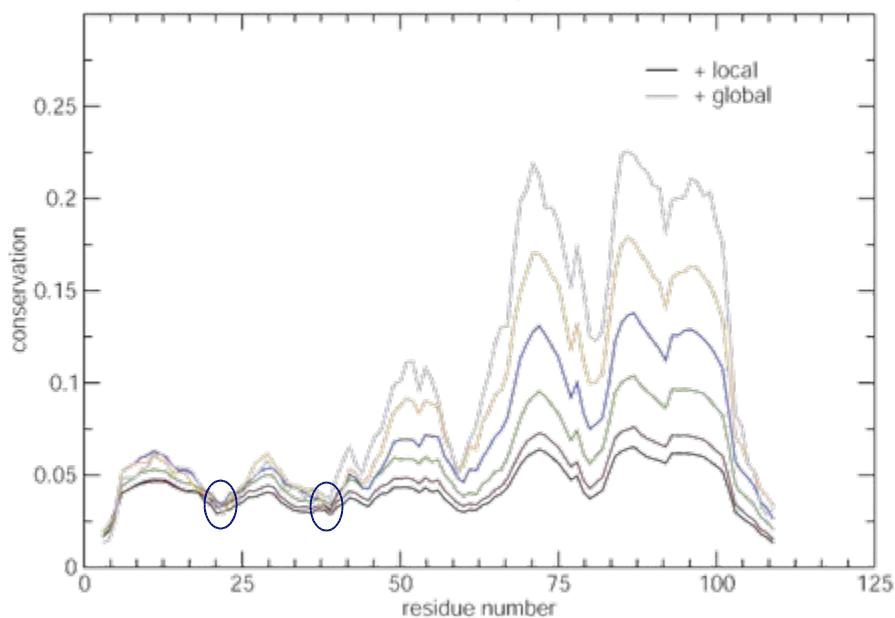
# Cytochrome C Peroxidase (2cyp) *all- $\alpha$* $\rightarrow$ CCP-like domain



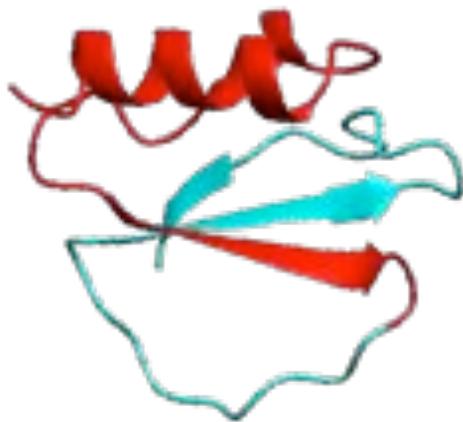
# Barnase Domain-Swapping



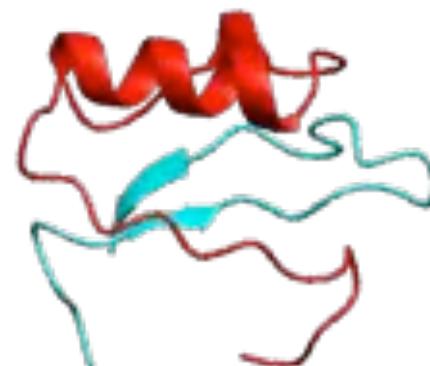
Barnase (1brn:L)  
conservation profile



# chymotrypsin inhibitor 2



1-37 | 38-64



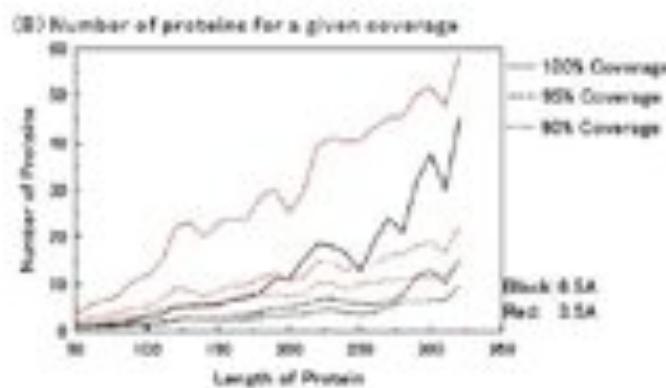
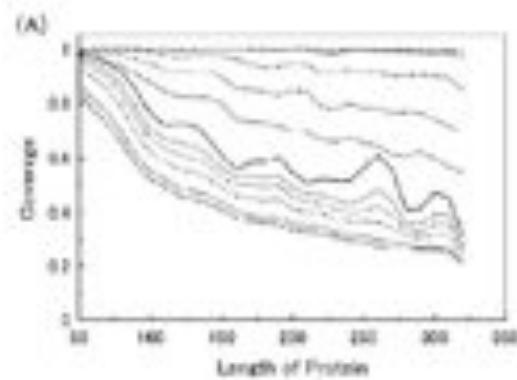
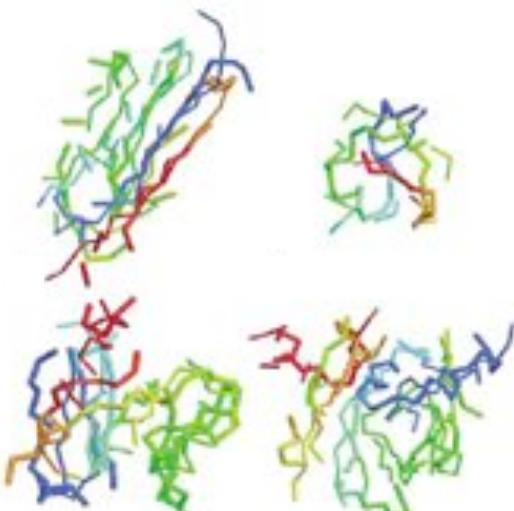
1-40 | 41-64

- 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.

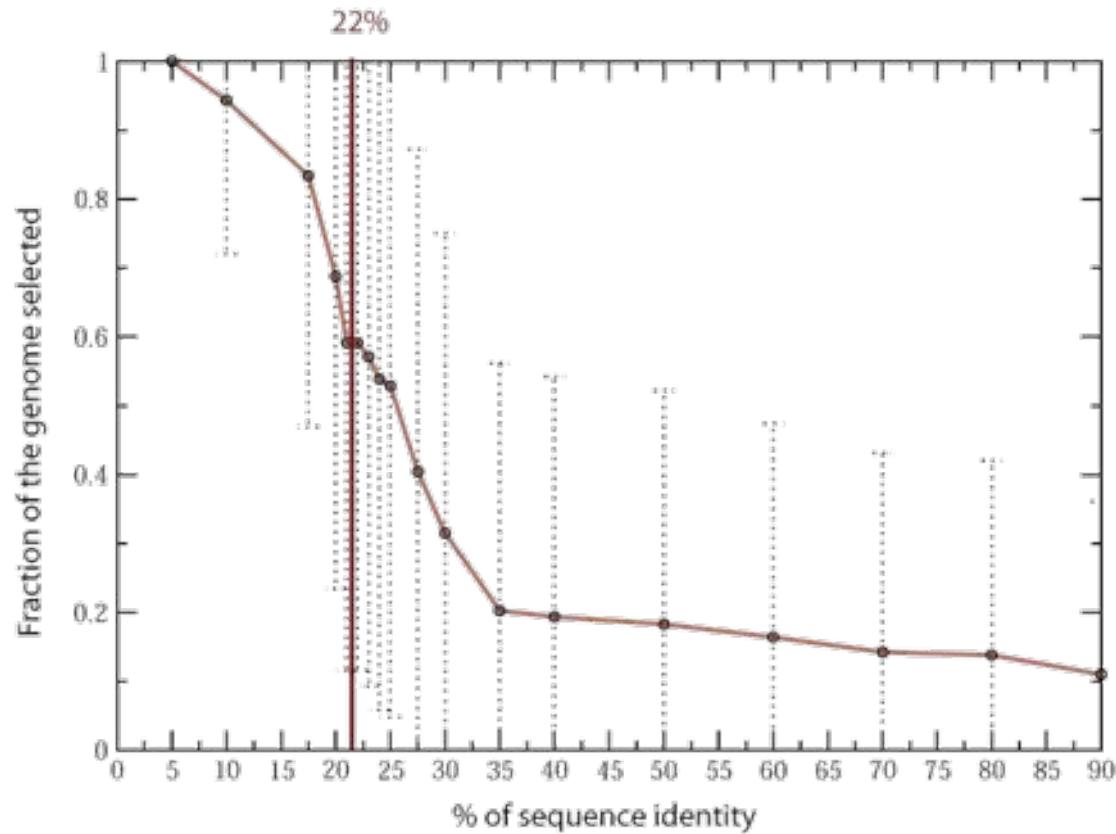
- Ladurner AG, Itzhaki LS, de Prat GG, Fersht AR. 1997. Complementation of peptide fragments of the single domain protein chymotrypsin inhibitor 2. *J Mol Biol* 273:317-329.

# Sequence space .vs. Structure space

The PDB is a covering set of small protein structures.



# Sequence space .vs. Structure space



Data from DBAli v2.0 with maximum search space of 2000 chains

# Sequence space .vs. Structure space



# BMI-206

## Structure-Structure comparisons Sequence-Structure comparisons

Marc A. Marti-Renom  
Assistant Adjunct Professor  
Department of Biopharmaceutical Sciences

February 19<sup>th</sup>, 2004

# How to use this lectures

- Ask me!
- Each day goes like...
  - Basic introduction
  - Theory (representation-scoring-optimization)
  - Available programs
  - Their results
  - Examples
- Second day we discuss the assignment for the class
  - *The BMI206 genome. Structural and functional annotation.*

# Outline of the lectures

- Day 1.
  - Structure-Structure comparisons
  - Some look at databases of protein structure classification
- Day 2.
  - Sequence-Structure comparison
  - Description of the class assignment

# Day 2. Summary

- Sequence-Structure comparisons
  - Before we start...
    - Some theory...
    - Domain boundaries
  - Structural predictions from sequence...
    - PSI-PRED (SSE prediction)
    - SALIGN (gap penalties and substitution matrices)
    - mGenThreader (SSE prediction and alignment/potential scores)
    - Fugue (gap penalties and substitution matrices)
    - 3D-Jury (as a meta server example)
  - What works...
    - EVA server (EVA-Threading)

# DAY 2

## “Fold assignment”

# Sequence space .vs. Structure space



# General overview (Threading)

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

# Template comparison methods

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

# Sequence-Structure alignments

**As any other bioinformatics problem...**

- Representation
- Scoring
- Optimizer

# Empirical energy functions (PMF)

- Idea: **energy leads to structure, thus it should be possible to infer energy from many known structures**
- To be used in: model refinement and assessment
- Properties needed:
  - Deep minimum at correct state (native)
  - Smooth
  - Simple
- Types:
  - Contact potential
  - Distance potentials
  - Surface potentials

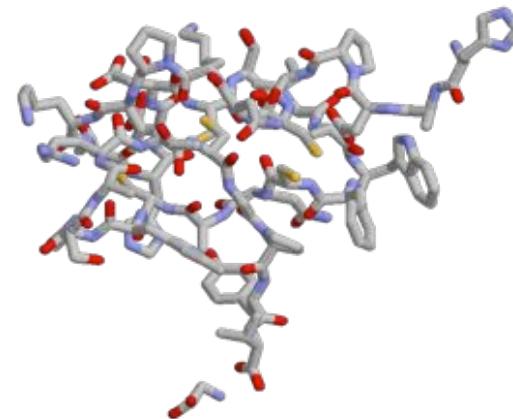
# Approximations/Limitations in PMFs

- Database size.
- PMF versus Energy (additive/higher order terms).
- Reference state.
- Physical origin.

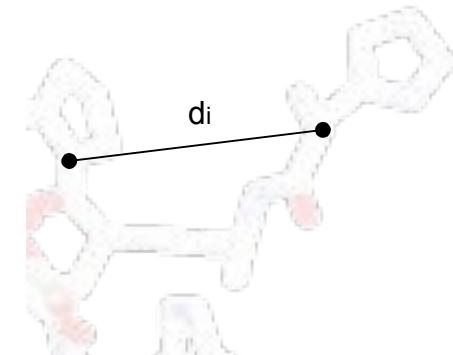
# Representation Sequence/Structures

>gi42541361  
MDIRSVSSLRGLLCLPPSWPRR

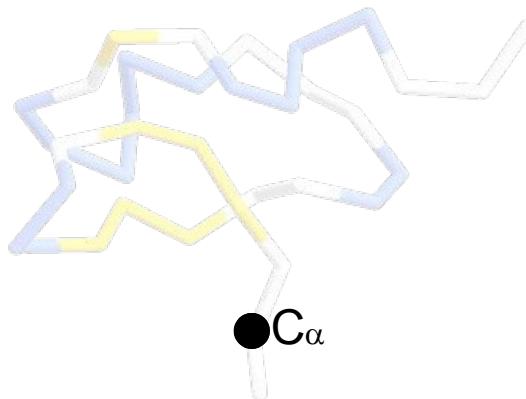
Primary sequence



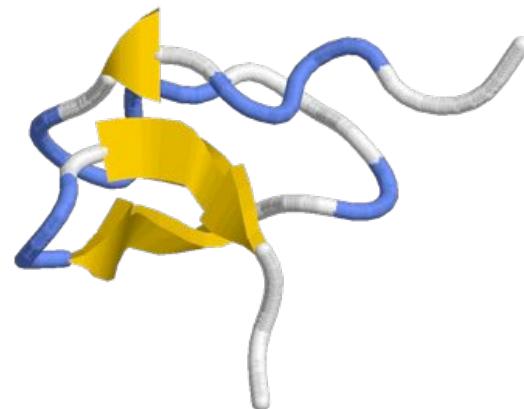
All atoms and coordinates



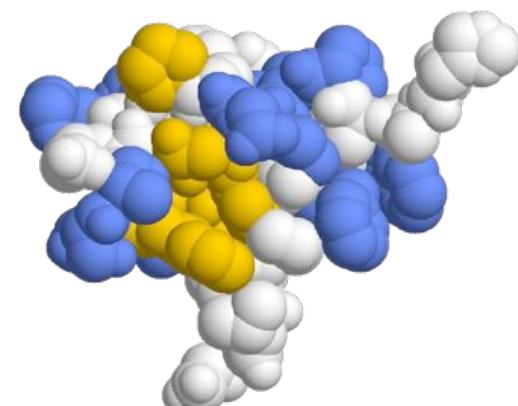
Distance space



Reduced atoms representation



Secondary Structure



Accessible surface

# Scoring Statistical Potential... inspiration

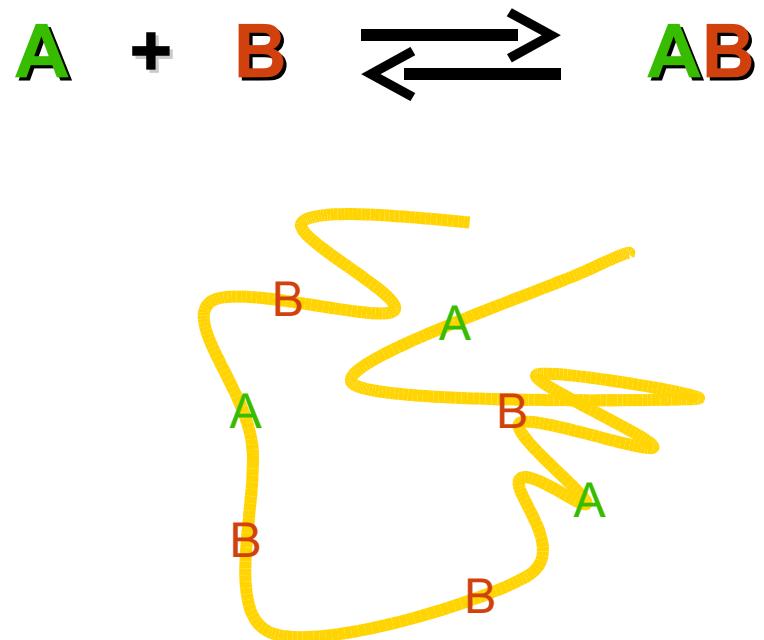
$$K = \frac{[AB]}{[A] \cdot [B]}$$

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

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

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

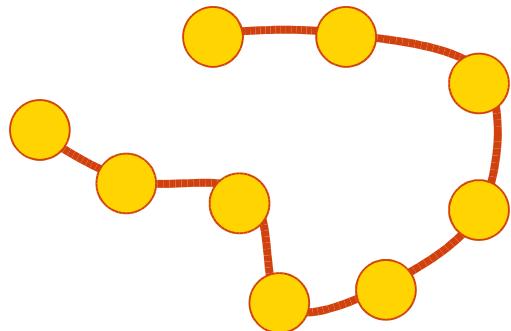
in which T is the absolute temperature and k is the Boltzmann's constant. As we are interested in an interaction energy between two amino acid side chains, it would seem natural to define  $N_1$  as the number of interactions between these two residues types in a group of real protein structures, a number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



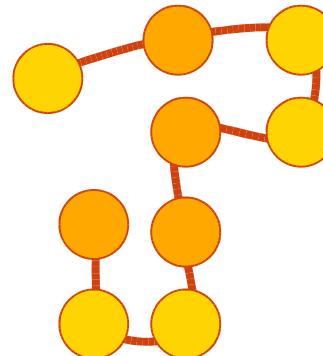
Tanaka and Sheraga (1975) PNAS, 72 pp3802  
A. Godzik, (1996) Structure 15 pp363

Scoring

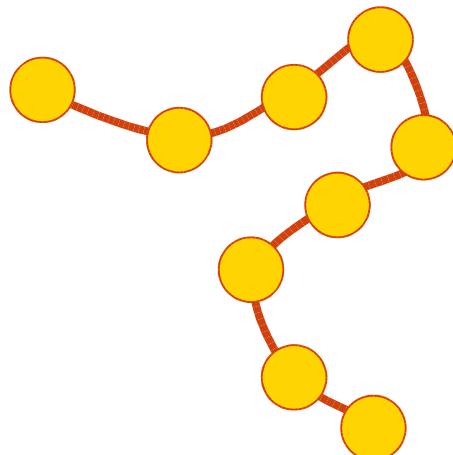
# Statistical Potential... interaction types



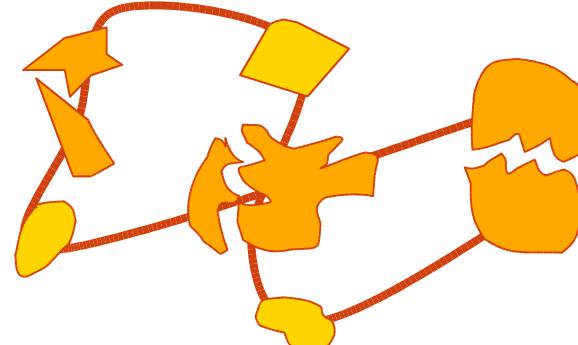
Neutral interactions



Hydrophobic interactions



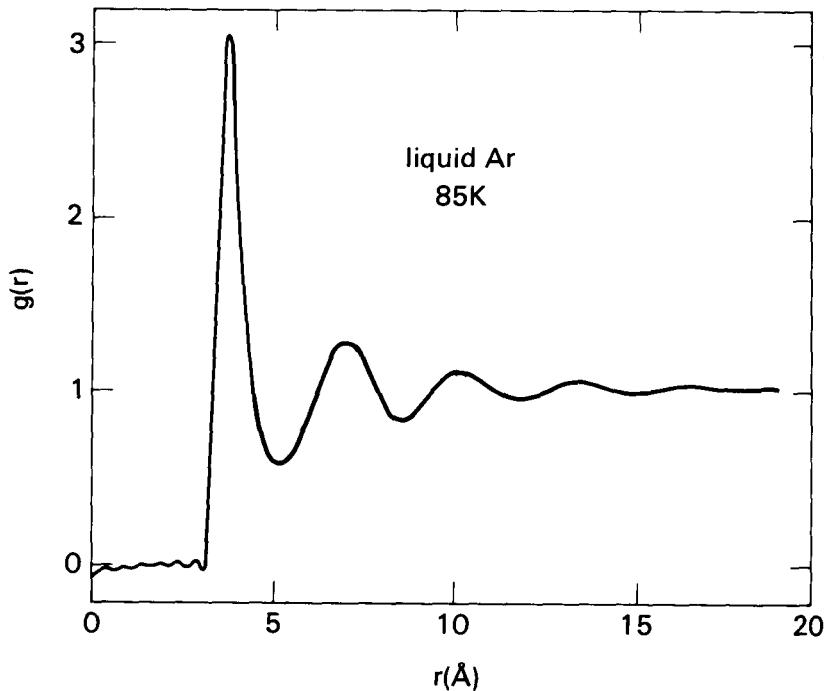
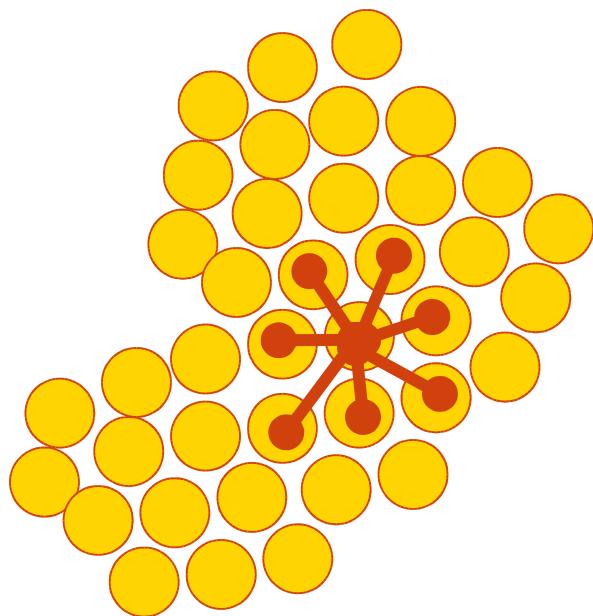
Compact interactions



Specific interactions

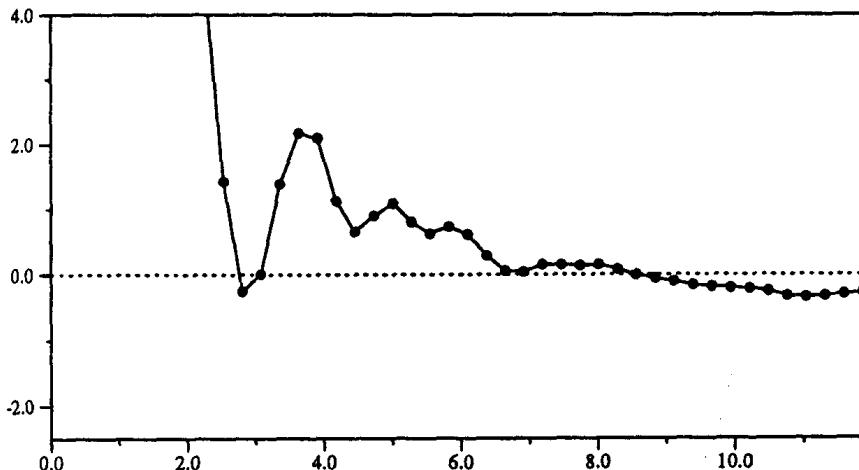
## Scoring

# Statistical Potential... reference state

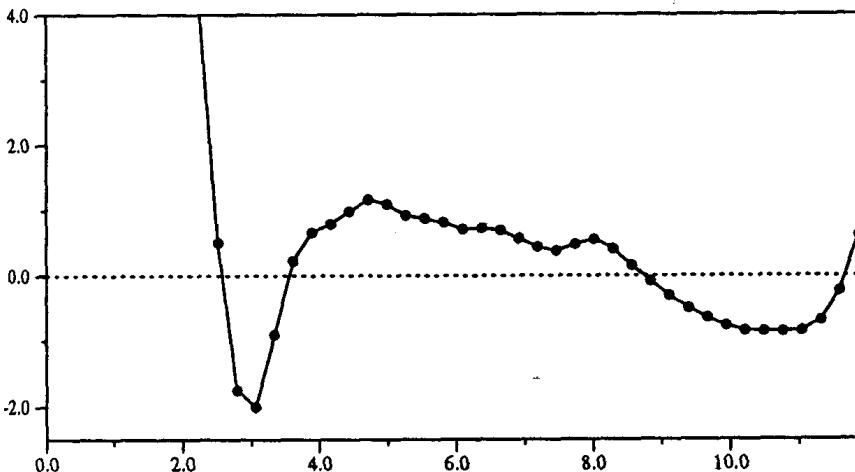


# Scoring Statistical Potential... Hydrogen Bonds

Long range free energy



Short range free energy



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

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

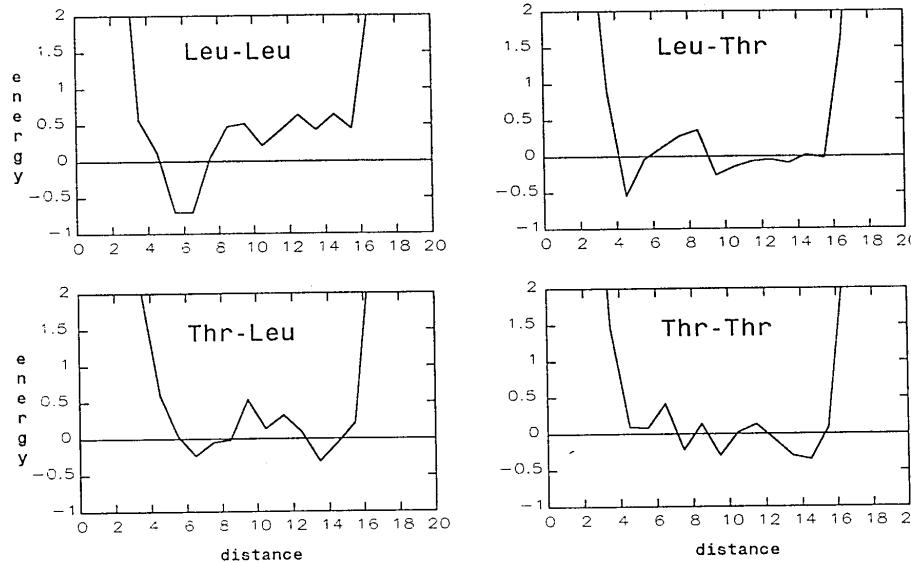
$$g_{NO}(r) = \frac{\rho_{NO}(r)}{\rho}$$

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

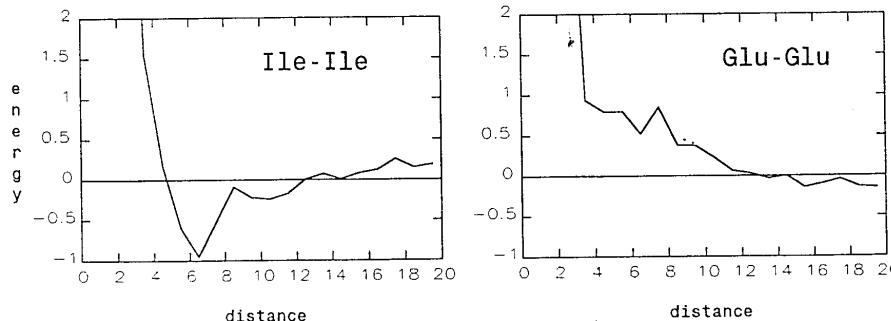
## Scoring

# Statistical Potential... Distance Potentials

Long range free energy



Short range free energy

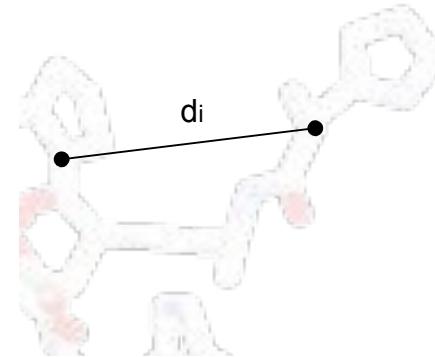


## Scoring

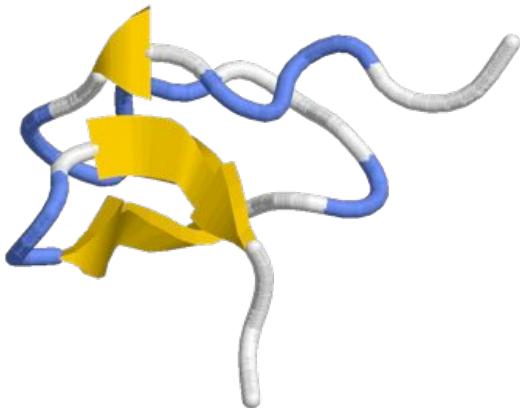
# Raw scores of an alignment

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

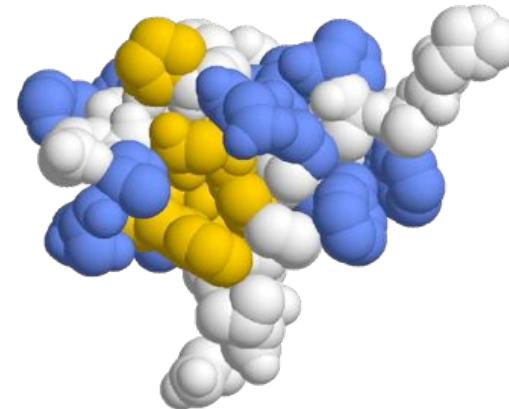
Aminoacid substitutions



Distance space



Secondary Structure (H,B,C)



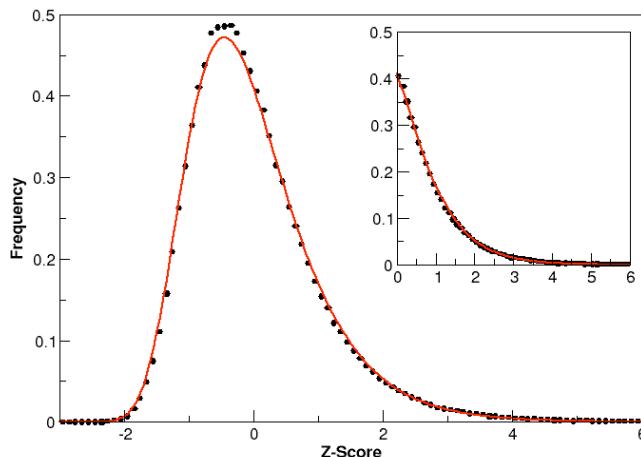
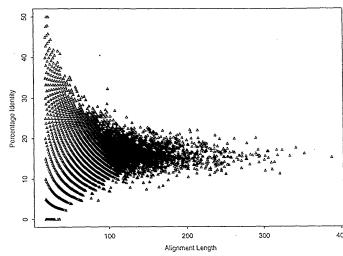
Accessible surface (B,A [%])

## Scoring

# Significance of an alignment (score)

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

Empirical



Sometimes approximated by Z-score (normal distribution).

Analytic

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

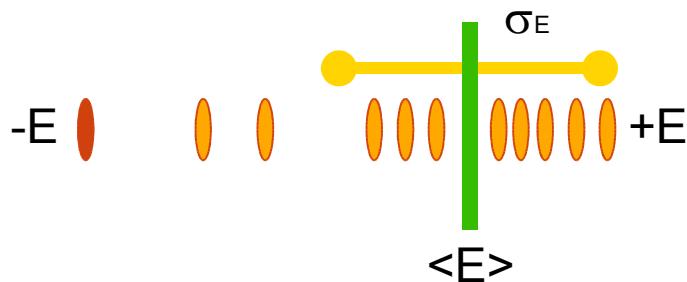
$$P(s \geq x) = 1 - \exp(e^{-\lambda (s - \mu)})$$

Karlin and Altschul, 1990 PNAS 87, pp2264

## Scoring

# Significance of an alignment (score)

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

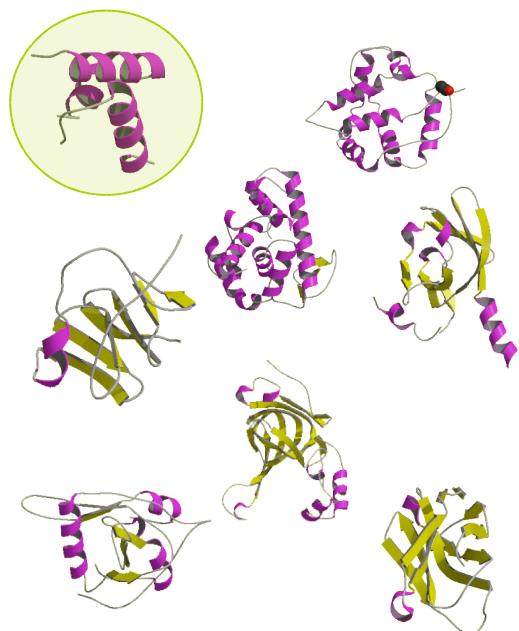


$$Zscore = \frac{(\langle E \rangle - E_m)}{\sigma_E}$$

## Scoring

# Significance of an alignment (background)

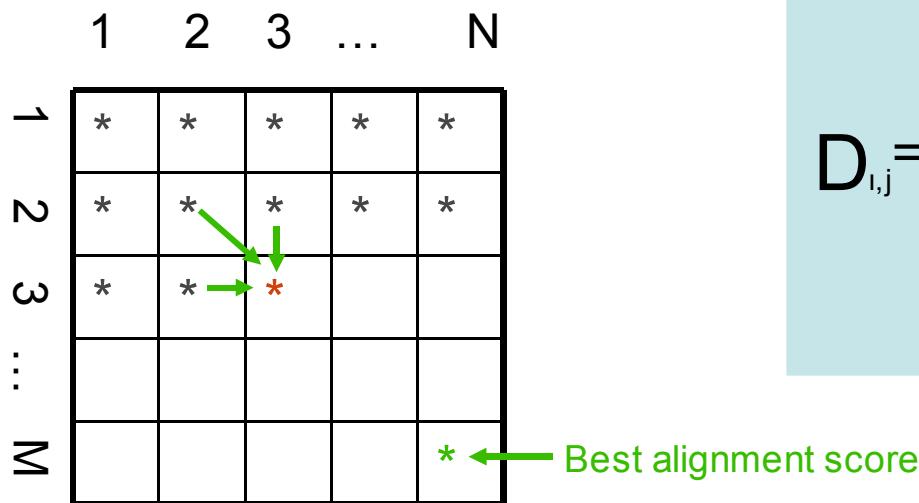
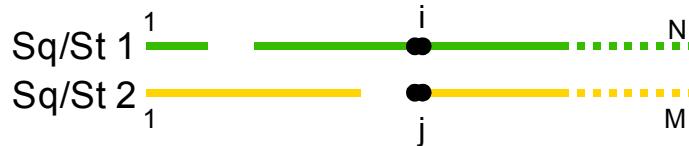
Structural space



Sequence space

MKLLIVLTCISLCSCICTVVQRCASNKPHVLEDPCKVQH  
HLSVNQCVLLPQCCPKSCKICTHLISIEVVLTCAVDKM  
MHVNCVEQCSLQDCIKIAPRVLKTCLCVLKPCLTSVSH  
VHLVQPTSCCCKKNICICHVEIRSLDILTKSVQLACLVPM  
⋮  
MQCCRVQKICDLLAVELCKLHISTPSCKILCVVTSVPHN

# Global dynamic programming alignment

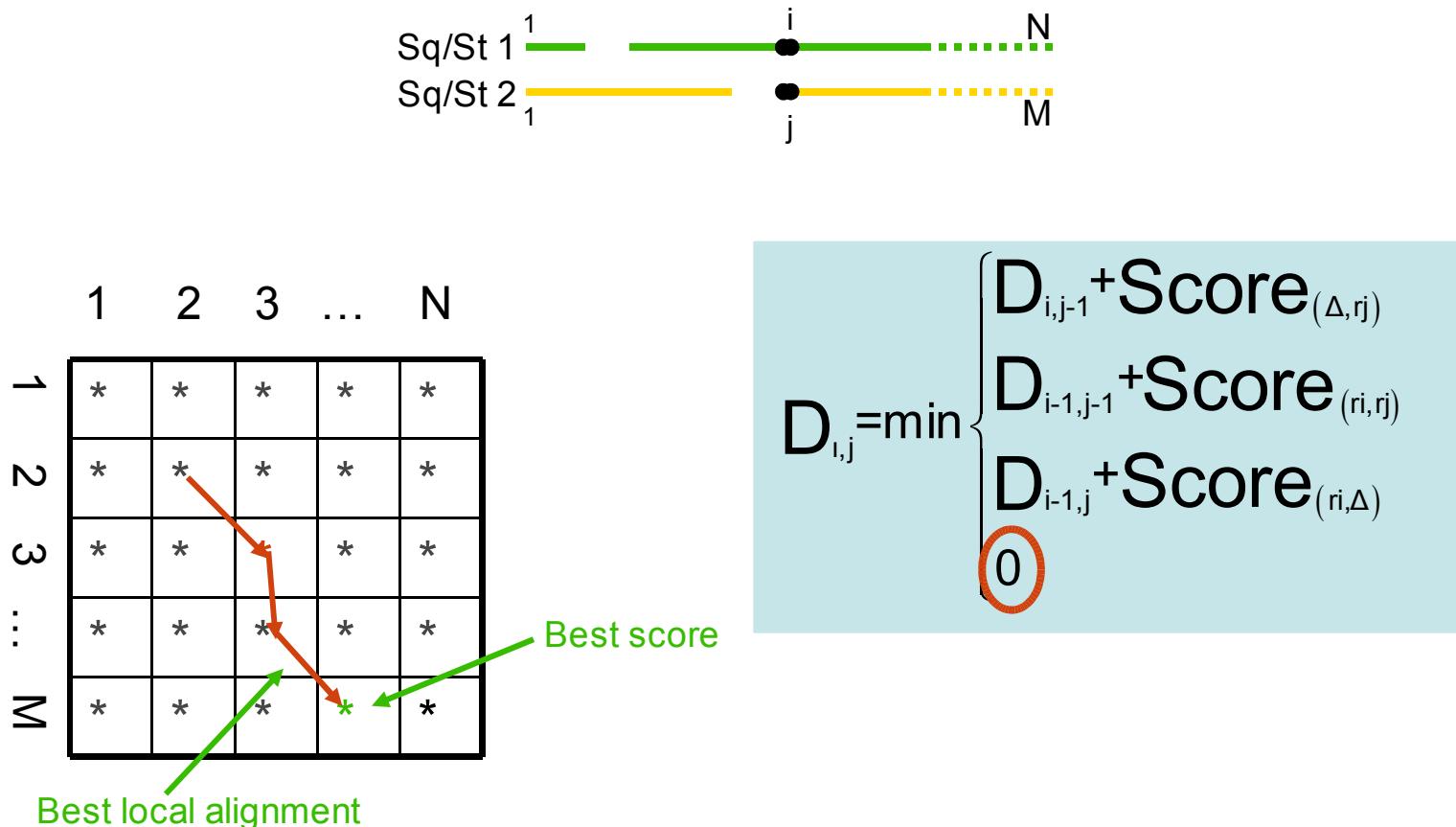


$$D_{i,j} = \min \begin{cases} D_{i,j-1} + \text{Score}_{(\Delta, rj)} \\ D_{i-1,j-1} + \text{Score}_{(ri, rj)} \\ D_{i-1,j} + \text{Score}_{(ri, \Delta)} \end{cases}$$

Backtracking to get the best alignment

## Optimizer

# Local dynamic programming alignment



Backtracking to get the best alignment

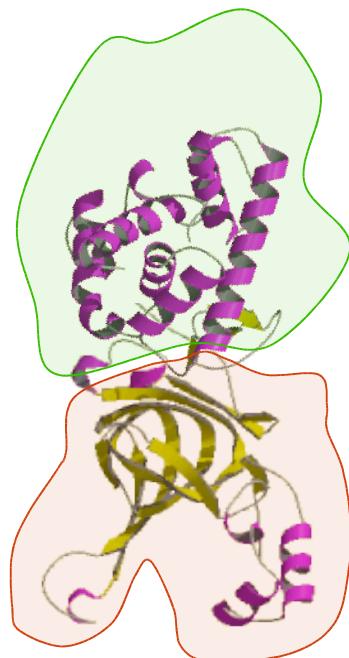
Smith and Waterman (1981) *J. Mol Biol.*, 147 pp195

# Applications of PMFs

- Model assessment.
- *Ab initio* folding simulations.
- Sequence-structure matching (threading).
- Comparative protein structure modeling (loops, sidechains, ...).
- Secondary structure prediction, etc.

# Domain boundaries from sequence

**VERY DIFFICULT!!!!**



MENFEIWVEKYRPTLDEVVGQDEVIQRLKGYVERKNIPHLLFSGPPGTGKTATAIALARDLFGENWRDN  
FIEMNASDERGIDVVRHKIKEFARTAPIGGAPFKIIFLDEADALTADAQAALRRTMEEMYSKSCRFILSCN  
YVSRIIEPIQSRCAVFRFKPVPKEAMKKRLLEICEKEGVKITEDGLEALIYISGGDFRKAINALQGAAAI  
GEVVDADTIYQITATARPEEMTELIQTALKGNFMEARELLDRLMVEYGMMSGDIVAQLFREIISMPIKDS  
LKVLQOLIDKLGEVDFRLTEGANERIQLDAYLAYLSTLAKK

# Domain boundaries from sequence (SnapDragon)

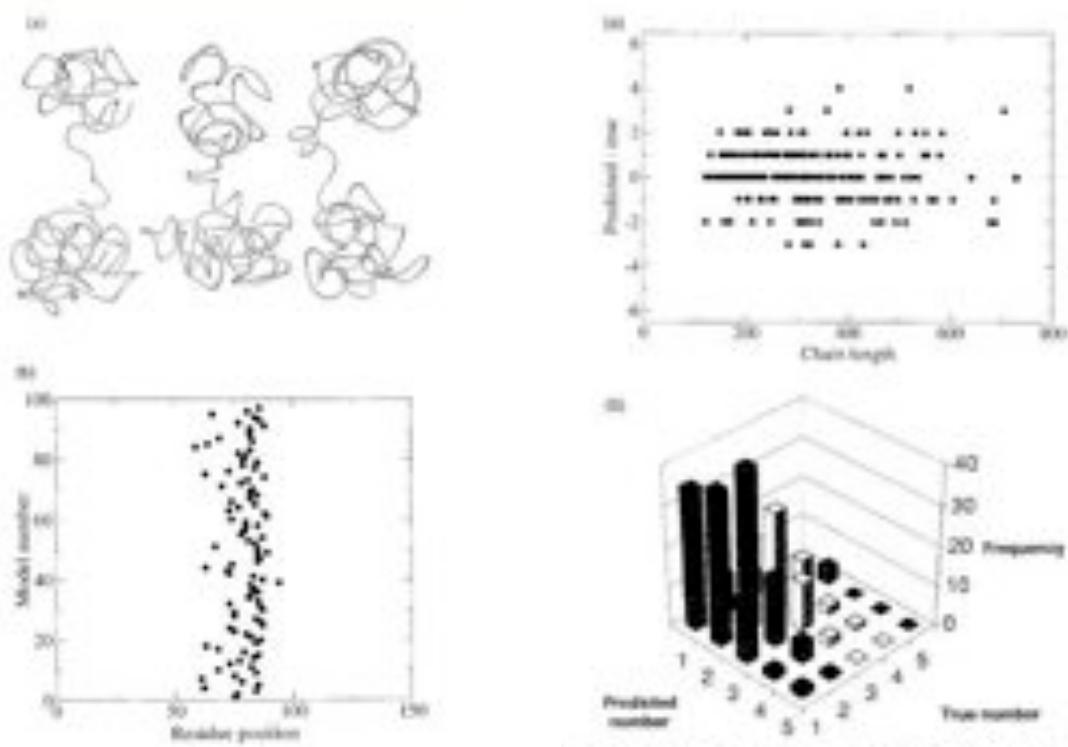
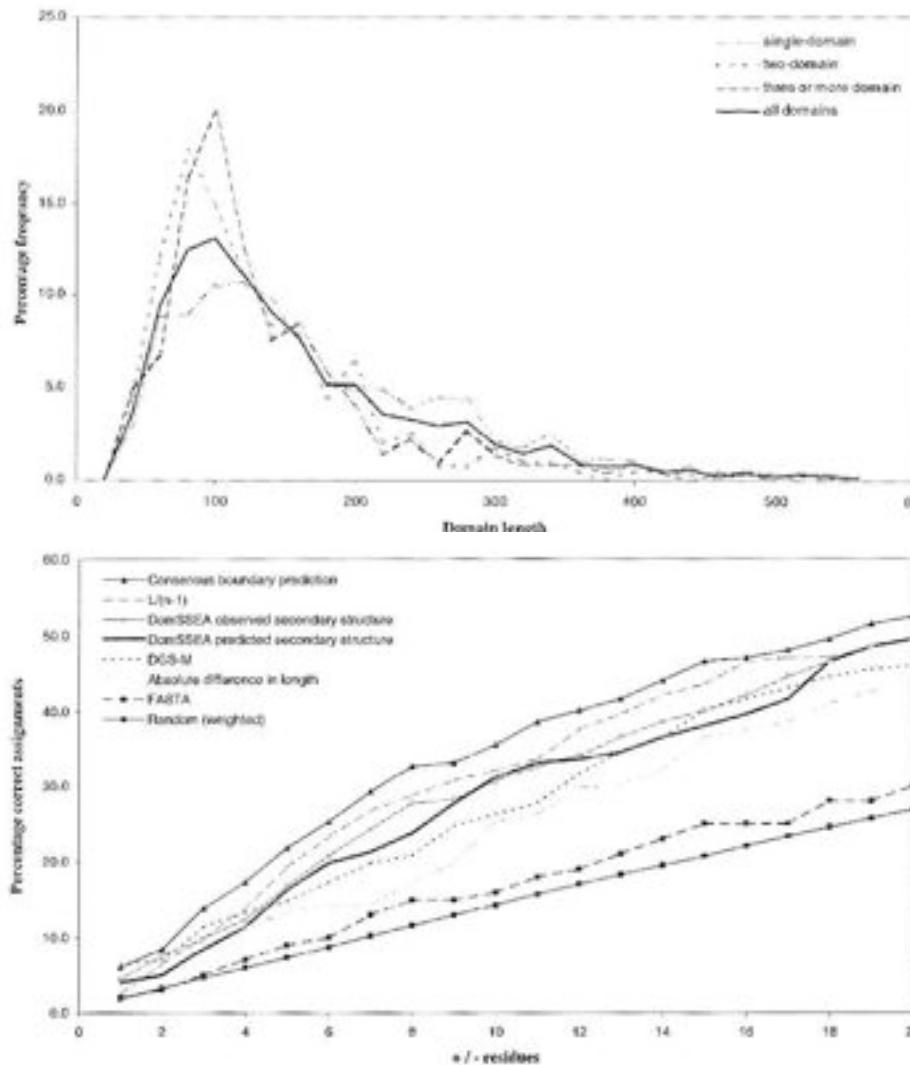


Table 2. Average accuracy percentages of linker prediction over 57 proteins

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

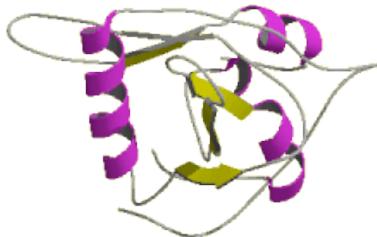
# Domain boundaries from sequence (DomSSEA)



# Prediction of Secondary Structure (PSI-PRED)

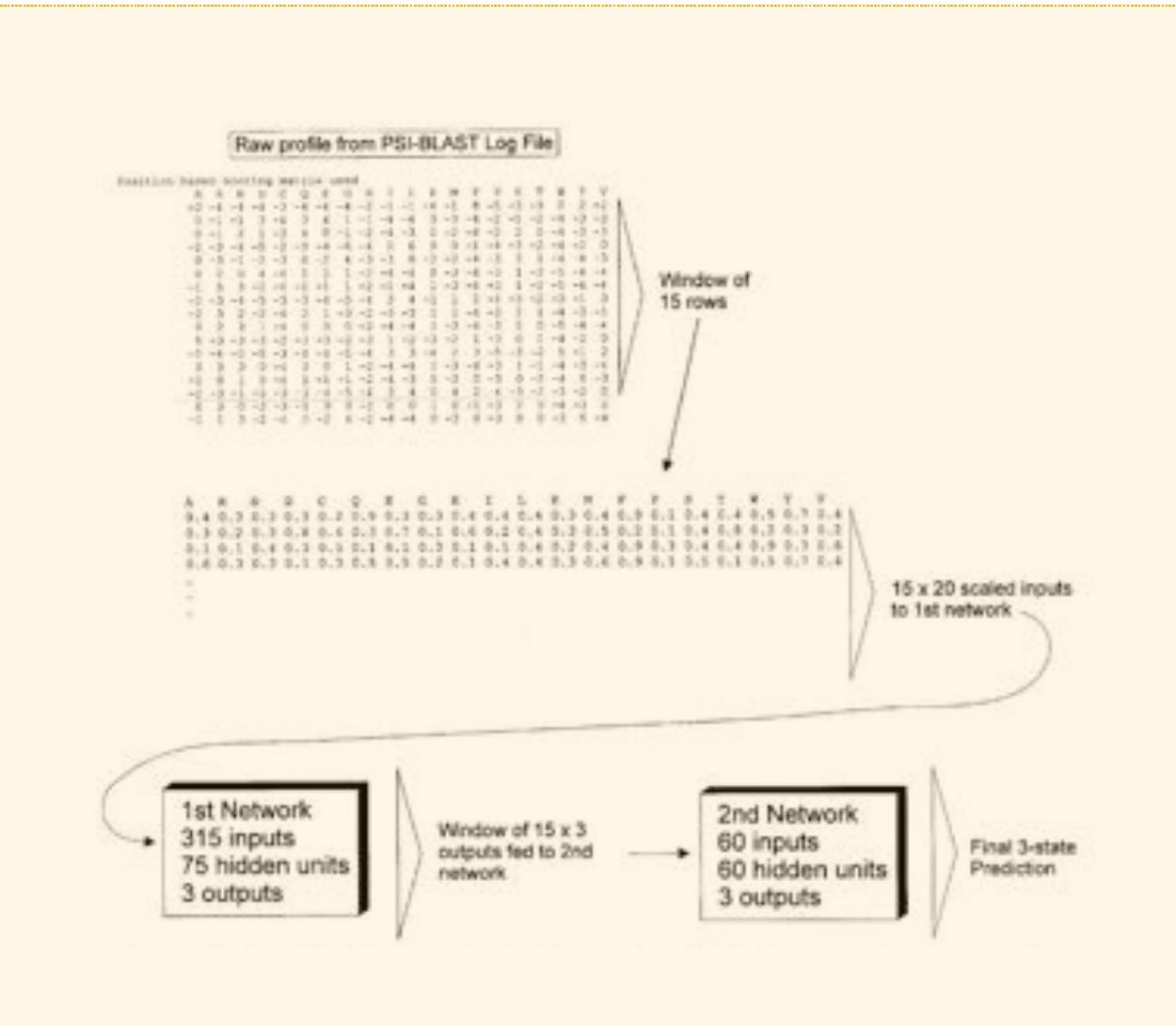
>gi42541361  
MDIRSVSSLRGLLLCLPPSWPRR

- Neural Network



- ✓ Very simple idea
- ✓ Simple scoring

- ✗ Obscure optimizer



# Prediction of Secondary Structure (PSI-PRED)

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

Screenshot of the PSIPRED Protein Structure Prediction Server interface in Internet Explorer.

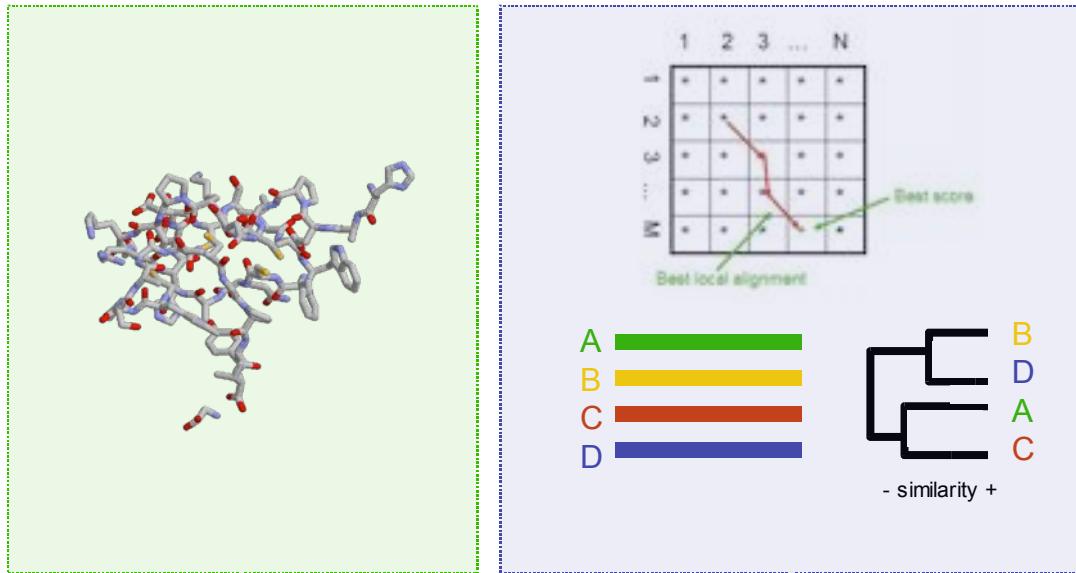
The page title is "The PSIPRED Protein Structure Prediction Server".

Key sections include:

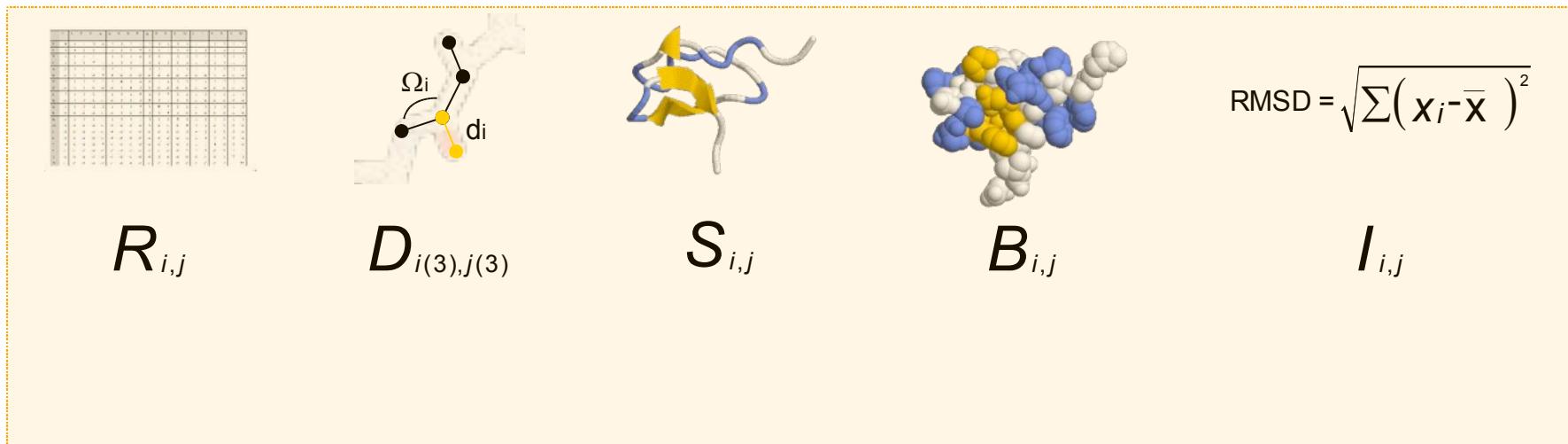
- PSIPRED home**: A link to the main PSIPRED website.
- Info**: A note suggesting not to bookmark the page due to its potential movement and directing users to the PSIPRED home page for more information.
- Input Sequence**: A text input field for entering the protein sequence in single-letter code.
- Choose Prediction Method**: A section with four radio button options:
  - Predict Secondary Structure (PSIPRED v2.4) (selected)
  - Predict Transmembrane Topology (MEMSAT)
  - Fold Recognition(GenTHREADER - quick)
  - Fold Recognition (mGenTHREADER - with profiles and predicted secondary structure)
- Filtering Options**: A section with three checkboxes:
  - Mask low complexity regions (checked)
  - Mask transmembrane helices (unchecked)
  - Mask coiled-coil regions (unchecked)A warning message states: "Warning: Turn off all filtering if you are running MEMSAT."
- Submit Sequence**: Fields for E-mail address, Password (only required for commercial e-mail addresses), and Short name for sequence.

At the bottom are "Predict" and "Clearform" buttons.

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



- ✓ Uses all available structural information
- ✓ Provides the optimal alignment
- ✗ Computationally expensive



# Structural alignment by properties conservation (SALIGN-MODELLER)

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

The screenshot shows a Microsoft Internet Explorer window displaying the DBAli v2.0 Tools page. The title bar reads "DBAli v2.0 Tools page - Microsoft Internet Explorer". The address bar shows the URL "http://salilab.org/DBAli/?page=tools&action=f\_salign". The main content area is titled "DBAli. Tools associated to the database." and lists several tools:

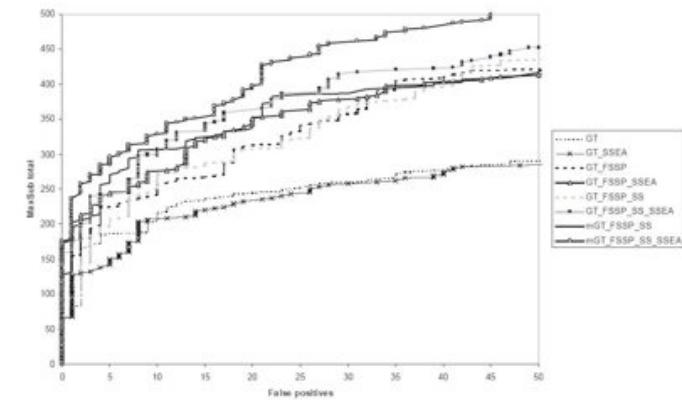
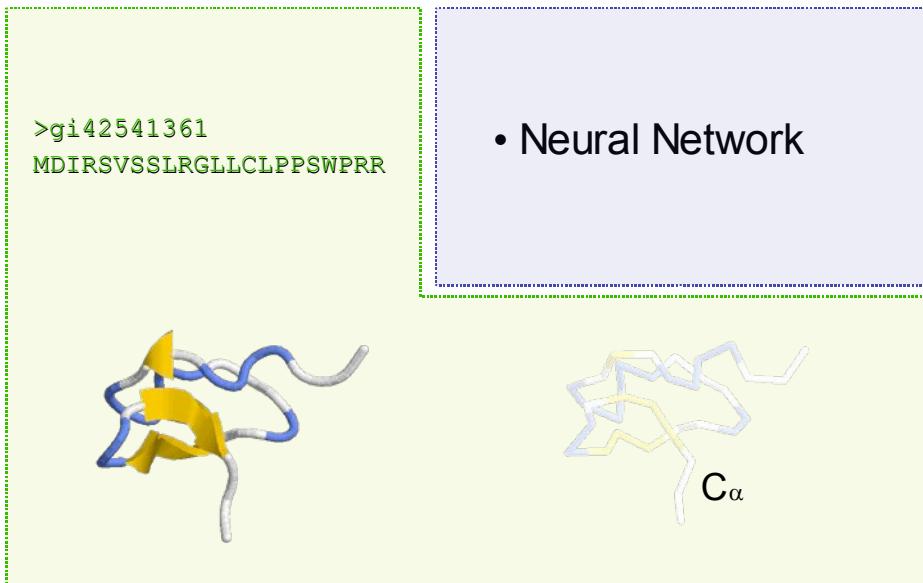
- Cluster a list of chains
- Cluster from a chain
- Define domains from a chain
- Get a multiple structure alignment of a list of chains
- Database statistics
- Download DBAli

Below this list, a sub-section titled "Get a multiple structure alignment of a list of chains." contains a form with the following fields:

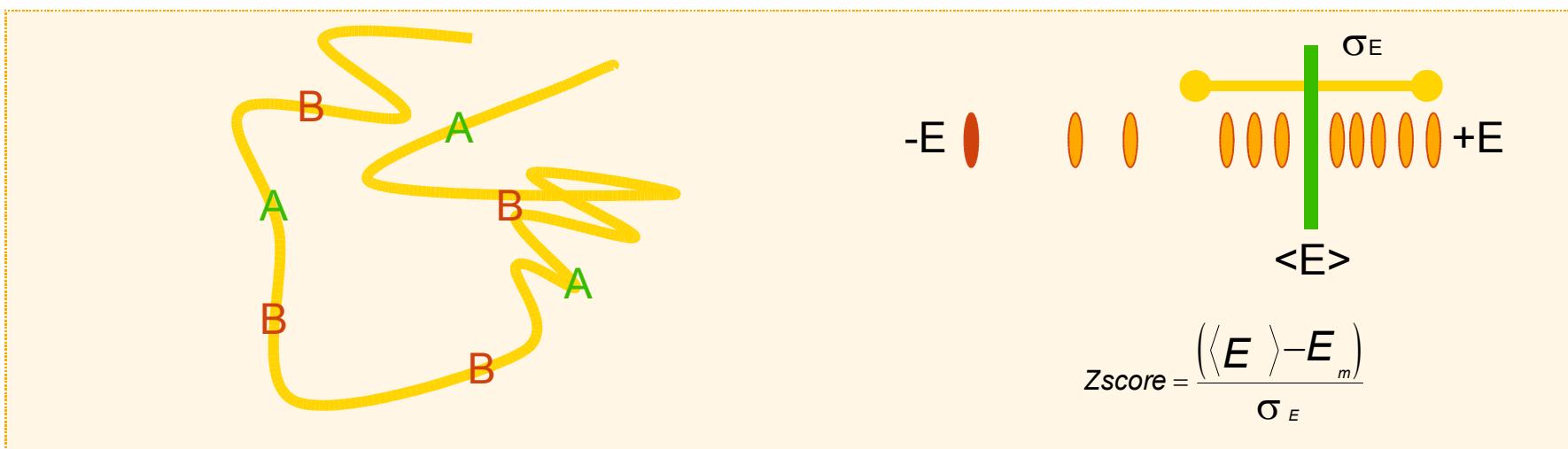
File with a list of chains:  
  ?

At the bottom of the page, there is a navigation bar with links: Reference, Download, Statistics, Suggestions, Visitors: 1407, © 2003 - 2004 Marti-Renom.

# Threading (mGenThreader)



- ✓ Good row and significance scoring
- ✗ Obscure optimizer

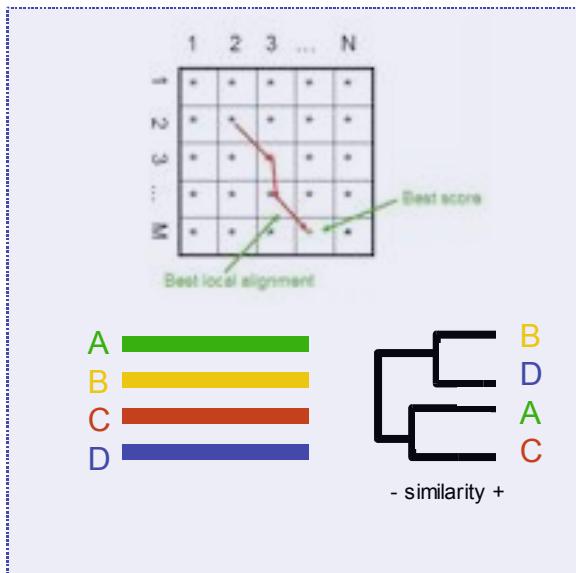
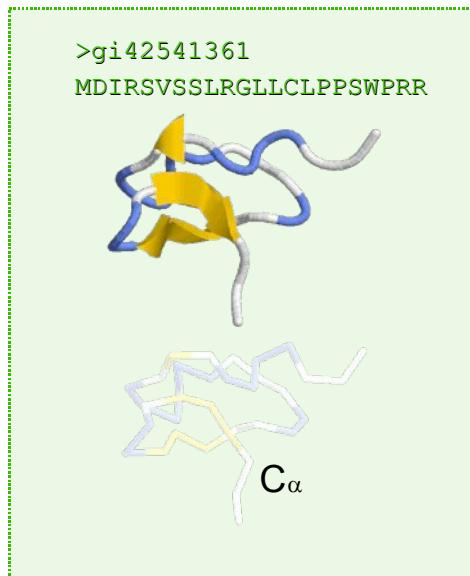


# Threading (mGenThreader)

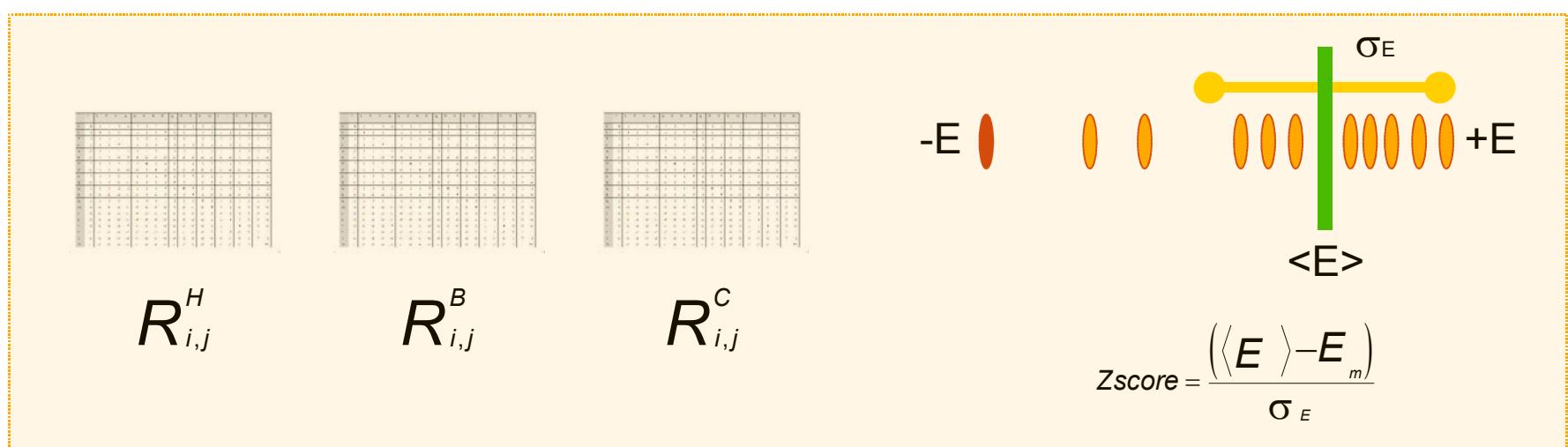
<http://bioinf.cs.ucl.ac.uk/psiform.html>

The screenshot shows the PSIPRED Protein Structure Prediction Server interface. At the top, there's a blue header bar with the UCL logo and the text "Bioinformatics Unit". Below this, the main content area has a left sidebar with categories: "PSIPRED home", "Info", "Input Sequence", "Choose Prediction Method", and "Filtering Options". The "Choose Prediction Method" section contains a list of four options with radio buttons. The fourth option, "Fold Recognition (mGenTHREADER - with profiles and predicted secondary structure)", is highlighted with a red arrow pointing to it. The "Input Sequence" section has a text input field labeled "Input sequence (single letter code)". The "Filtering Options" section has three checkboxes: "Mask low complexity regions" (checked), "Mask transmembrane helices" (unchecked), and "Mask coiled-coil regions" (unchecked). The bottom of the window shows standard Internet Explorer navigation buttons like Back, Forward, Stop, and Refresh.

# Remote homology detection (FUGUE)



- ✓ Uses most of the structural information
- ✓ Easy to access either locally and on the web
- ✓ Good row and significance scoring
- ✗ Does not uses multiple sequence information

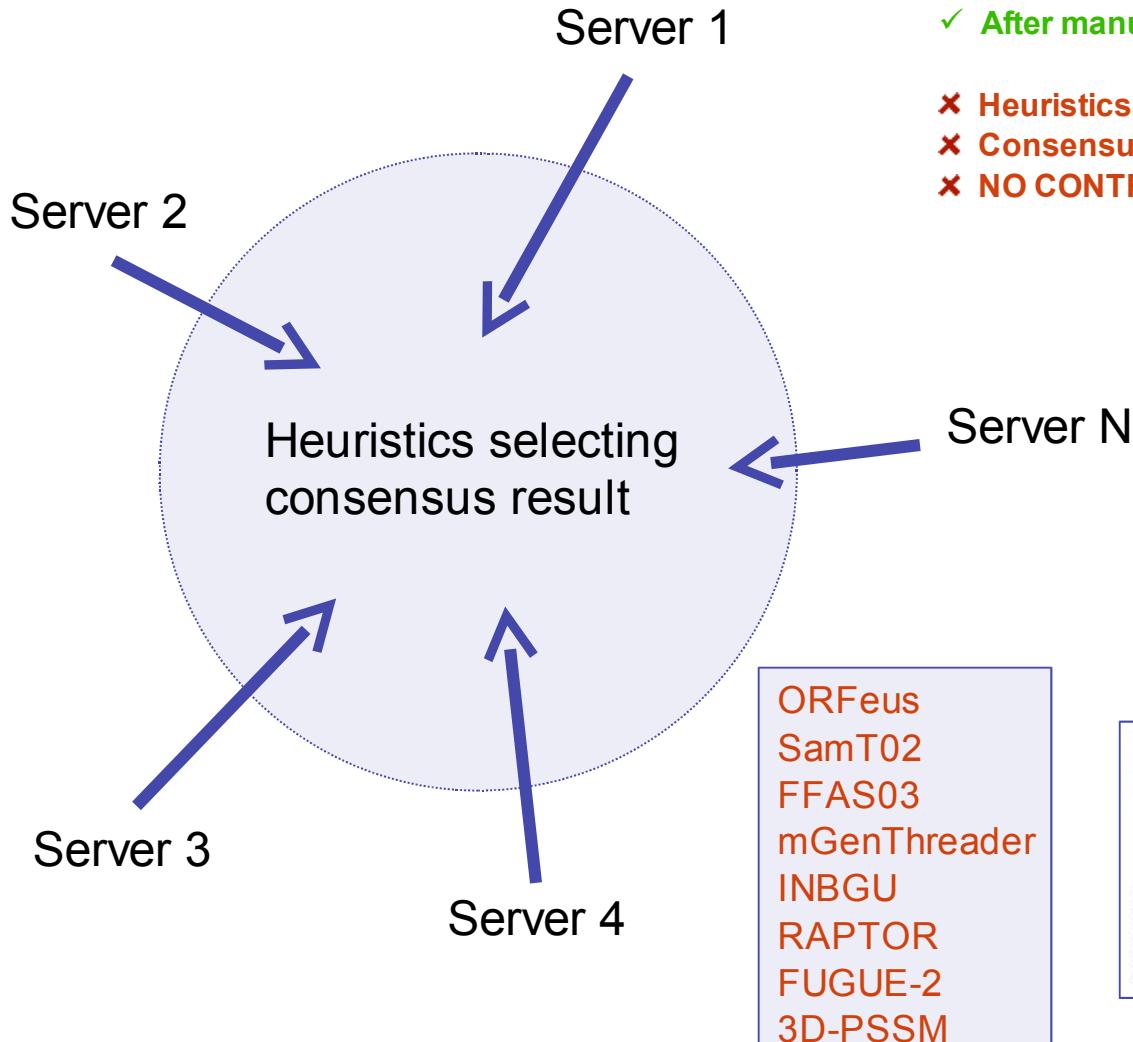


# Remote homology detection (FUGUE)

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

The screenshot shows a Microsoft Internet Explorer window displaying the FUGUE website. The title bar reads "FUGUE: sequence-structure homology recognition and alignment engine - Microsoft Internet Explorer". The address bar shows the URL "http://www-cryst.bioc.cam.ac.uk/fugue/". The page content includes the FUGUE logo, the text "Sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties", and several navigation links: "SEARCH STRUCTURAL DATABASE", "ALIGN SEQUENCE WITH STRUCTURE", "DOWNLOAD", and "DOCUMENTATION". Below these are sections for "Methods" and "Some practical information can be found in:". There is also a link to "Click here for information about the HOMSTRAD database". The University of Cambridge crest is visible in the top right corner.

# 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!

**Letter to the Editor**

**mRNA Cap-1 Methytransferase in the SAD1 Gene**

The 3D-jury option has predicted the methyltransferase-like domain of the SAD1 transcript. Based on the conservation of a characteristic feature of mRNAs, the mRNA cap-1 methytransferase domain has been assigned to this protein, which has potential implications for antibiotic discovery.

The silent outbreak of the severe acute respiratory syndrome (SARS) epidemic has led to thousands of potentially infected patients and hundreds of deaths. The SARS virus is a member of the coronaviridae, a group of viruses causing major medical and economic concerns. Recently, the SARS coronavirus identified as the pathogen responsible for the disease has been found to contain a unique sequence element [1].

We have applied the 3D-jury meta-predictor (Ginalski et al., 2003) to predict the structure and function of the SARS virus by the analysis of its genome. Three domain recognition methods utilize the global network of independent structure prediction servers. Detection of patterns of structure similarity between the predicted structures and known structures resulted from a set of structure predictions. Each method made a significant impact on the test validation assessment of protein structure prediction (CASP-5 and CASP-6) and was also used to validate the structure predicted by the servers of SAD1. One of the most interesting predictions was the identification of the methyltransferase domain of the SAD1 transcript. The predicted amino acid sequence of the SAD1 transcript sequence (100% identity with the SAD1 transcript) was submitted to the Swiss-Prot database [2] and was found to be identical to the one described by Ginalski et al. (2003). This entry confirmed the presence of the SAD1 transcript in rice and its high structural conservation.

**LETTERS**

**How Unique Is the Rice Transcriptome?**

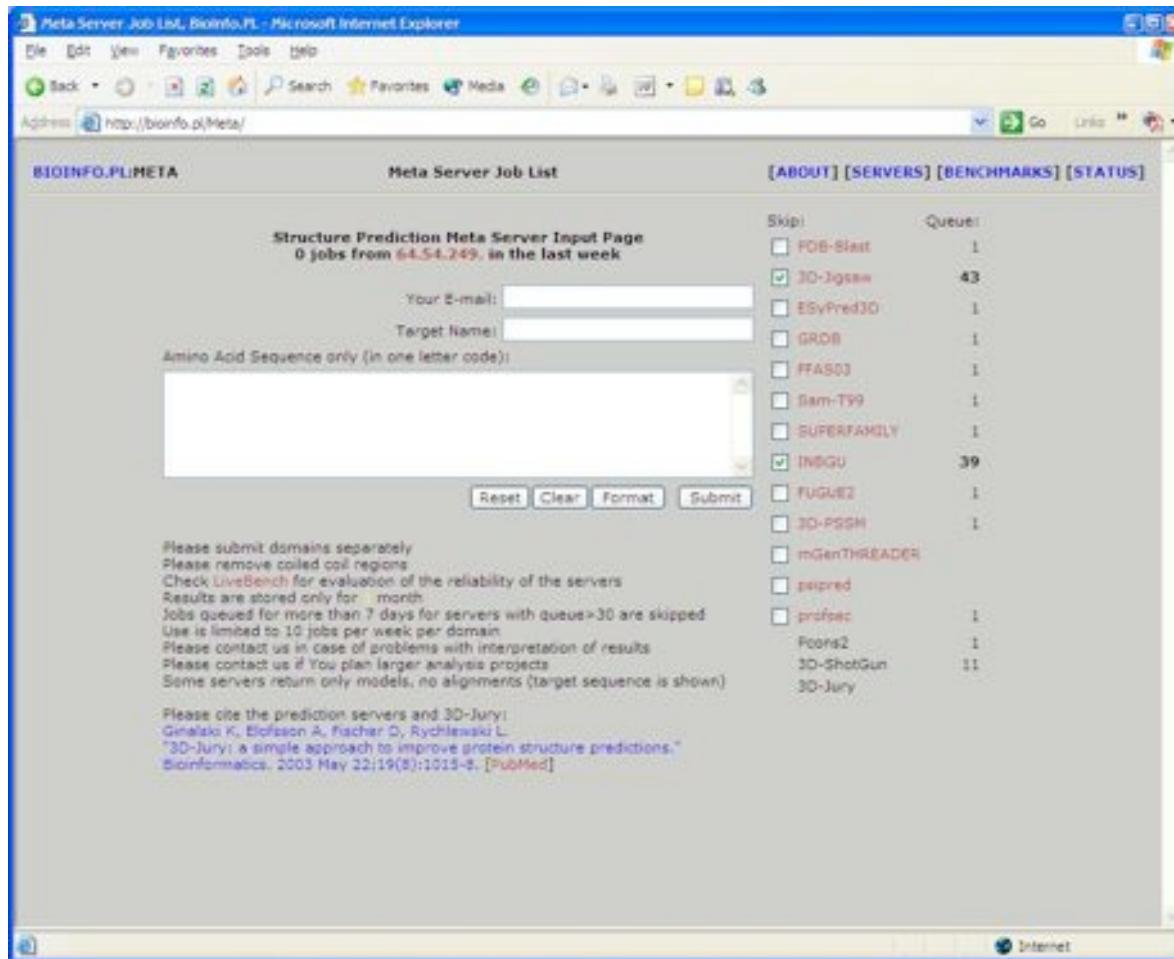
**IN THE REPORT "COLLECTION, MAPPING, AND ANNOTATION OF OVER 28,000 cDNA CLONES FROM JAPANESE RICE" (S. Kikuchi et al., 18 July, p. 376), THE RICE FULL-LENGTH cDNA PROJECT TEAM PROVIDES A DETAILED DESCRIPTION OF THE RICE TRANSCRIPTOME. THE AUTHORS CLAIM THAT 36% OF THE TESTED RICE TRANSCRIPTS ARE NOT**

**Figure 1. Structure of the methyltransferase domain of the SAD1 transcript.**

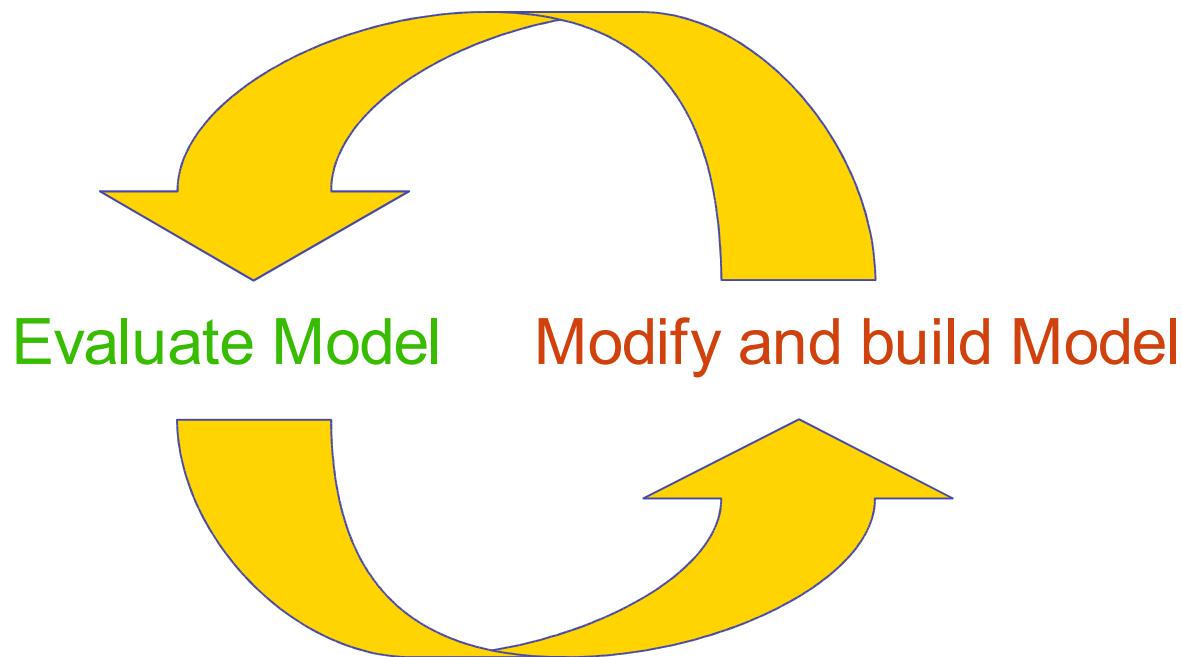
This figure is based on the nucleotide sequence of the SAD1 transcript. The structure was predicted by the 3D-jury meta-predictor. The predicted structure is composed of three domains: a N-terminal domain, a middle domain, and a C-terminal domain. The N-terminal domain is a methyltransferase domain, while the middle and C-terminal domains are less conserved. The predicted structure is compared with the known structure of the SAD1 transcript (PDB ID: 1SAD). The predicted structure is shown in stick representation, while the known structure is shown in ribbon representation. The predicted structure is highly similar to the known structure, with a root mean square deviation of approximately 0.3 Å.

# Meta-Servers (3D-Jury)

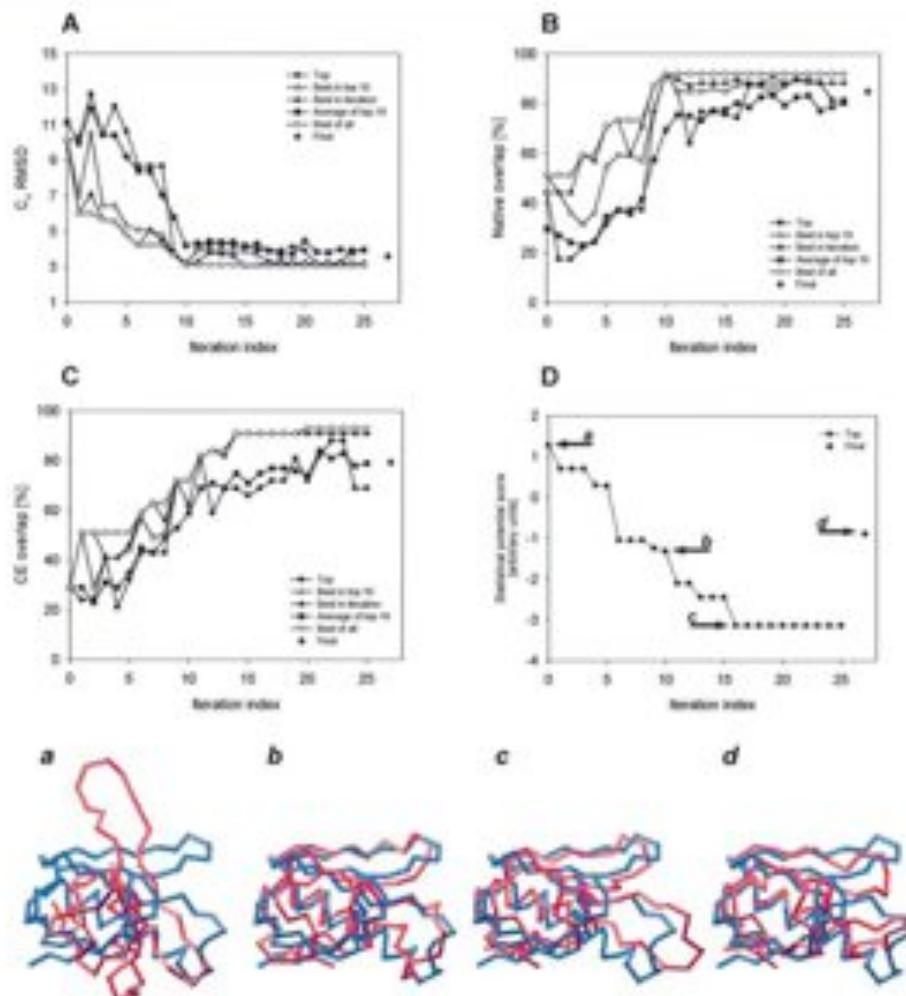
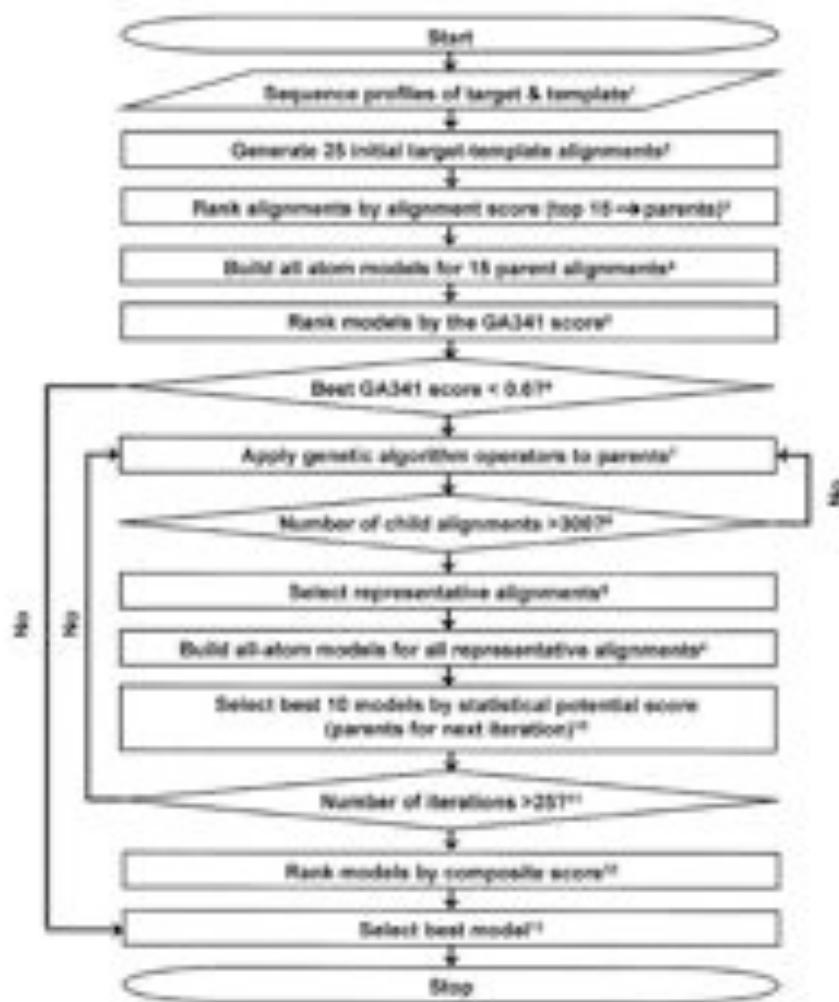
<http://bioinfo.pl/Meta/>



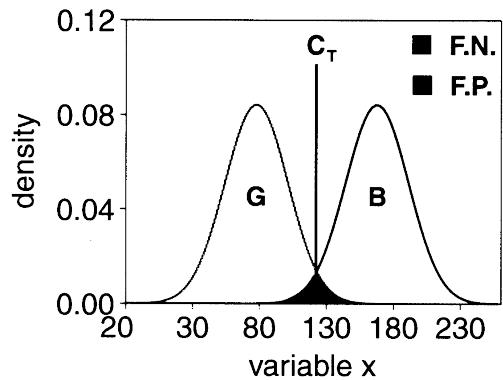
# Iterative process... better models(?)



# Iterative process... MOULDER



# Iterative process... Evaluation

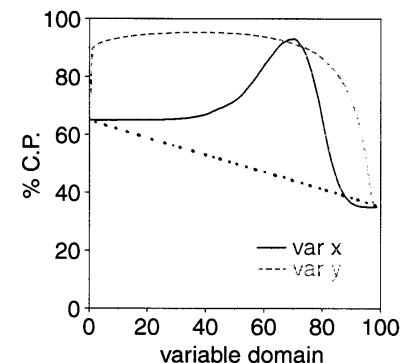
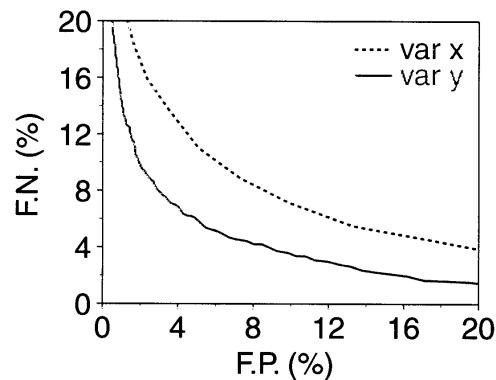


	<i>is GOOD</i>	<i>is BAD</i>
<i>predicted as GOOD</i>	a	b
<i>predicted as BAD</i>	c	d

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

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

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



# Iterative process... Evaluation

1

3900 GOOD MODELS

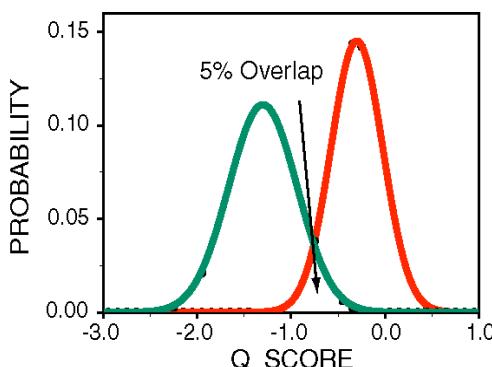
Models based on correct templates and approximately correct alignments

6000 BAD MODELS

Models based on incorrect templates or mostly incorrect alignments

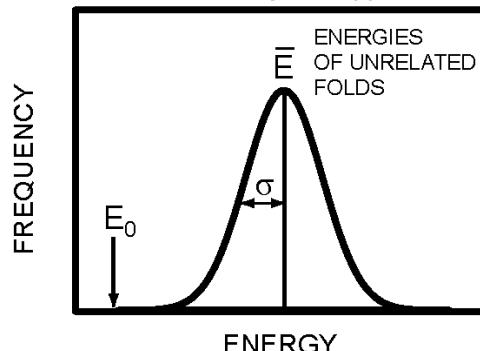
3

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



2

Prosall by M. Sippl



$$Z\text{-SCORE} = \frac{E_0 - \bar{E}}{\sigma}$$

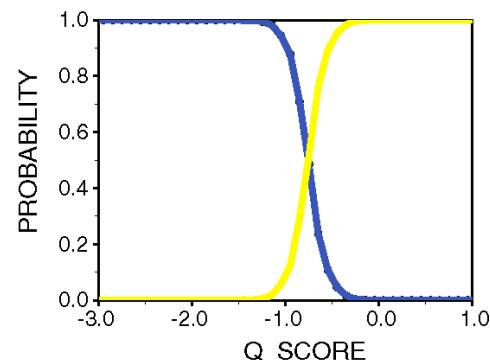
$$Q\text{-SCORE} = \frac{Z\text{-SCORE}}{\ln(\text{SEQ.LENGTH})}$$

4

$pG$

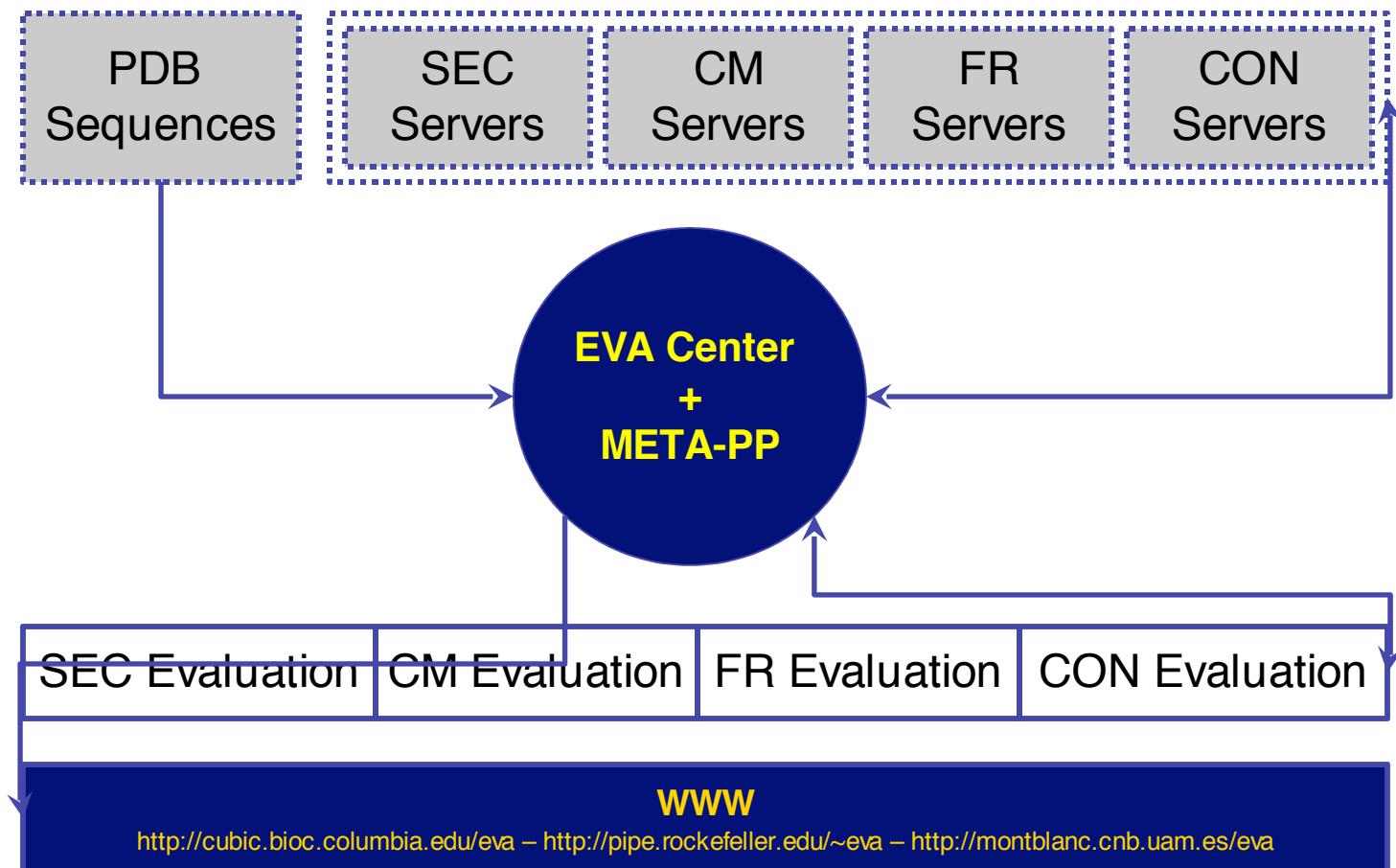
$$pG = \frac{p(Q\text{-SCORE}/\text{GOOD})}{p(Q\text{-SCORE}/\text{GOOD}) + p(Q\text{-SCORE}/\text{BAD})}$$

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



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

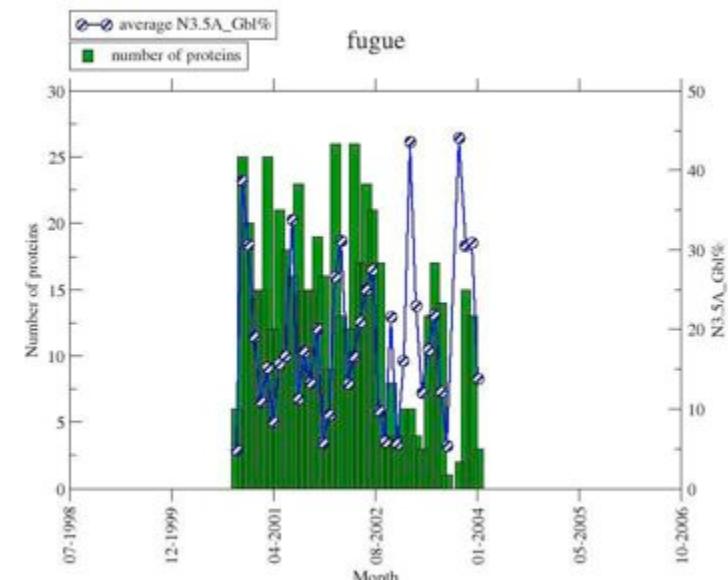
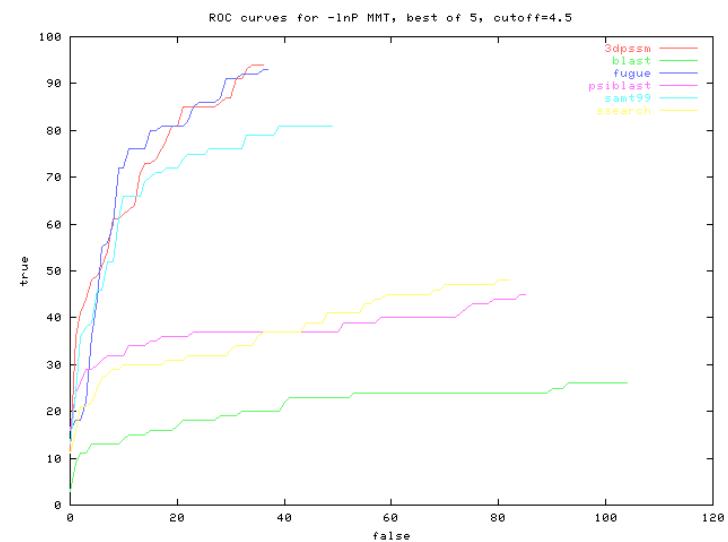
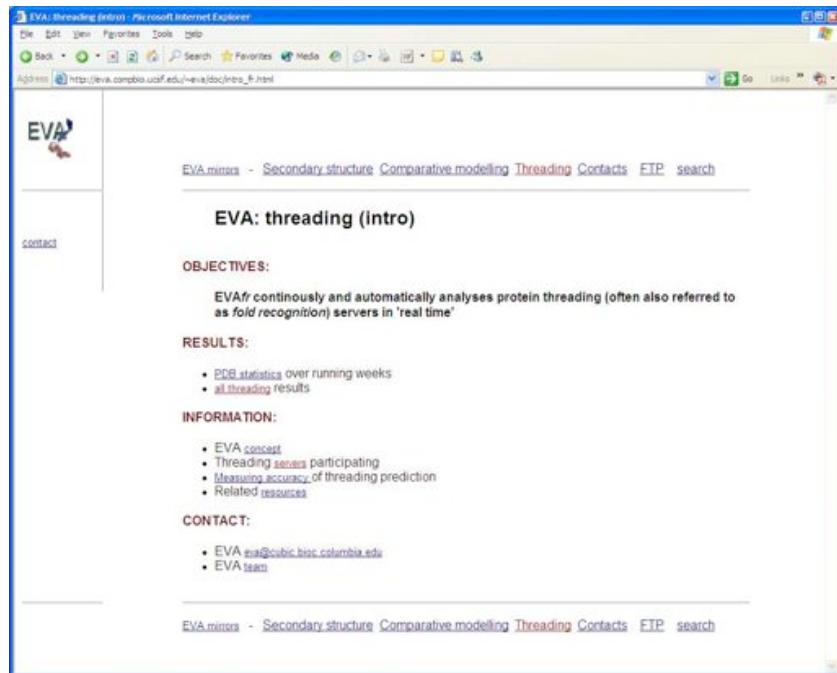
# Eva Server



<http://cubic.bioc.columbia.edu/eva> – <http://pipe.rockefeller.edu/~eva> – <http://montblanc.cnb.uam.es/eva>

# Eva-Threading Server

<http://eva.compbio.ucsf.edu/~eva/>



# Other evaluation benchmarks

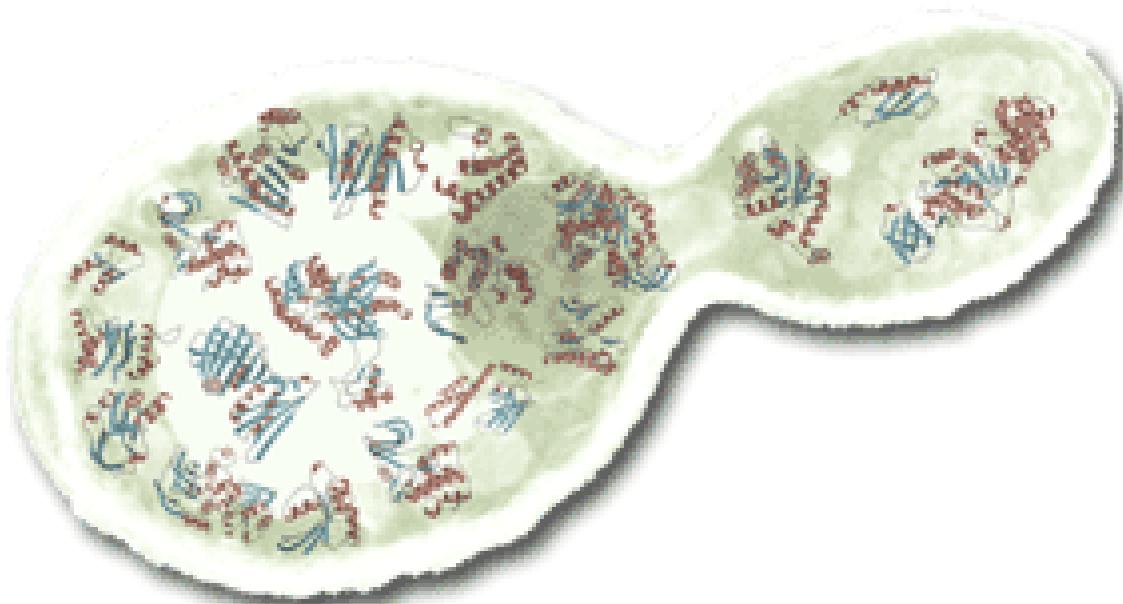
CASP <http://predictioncenter.llnl.gov>

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

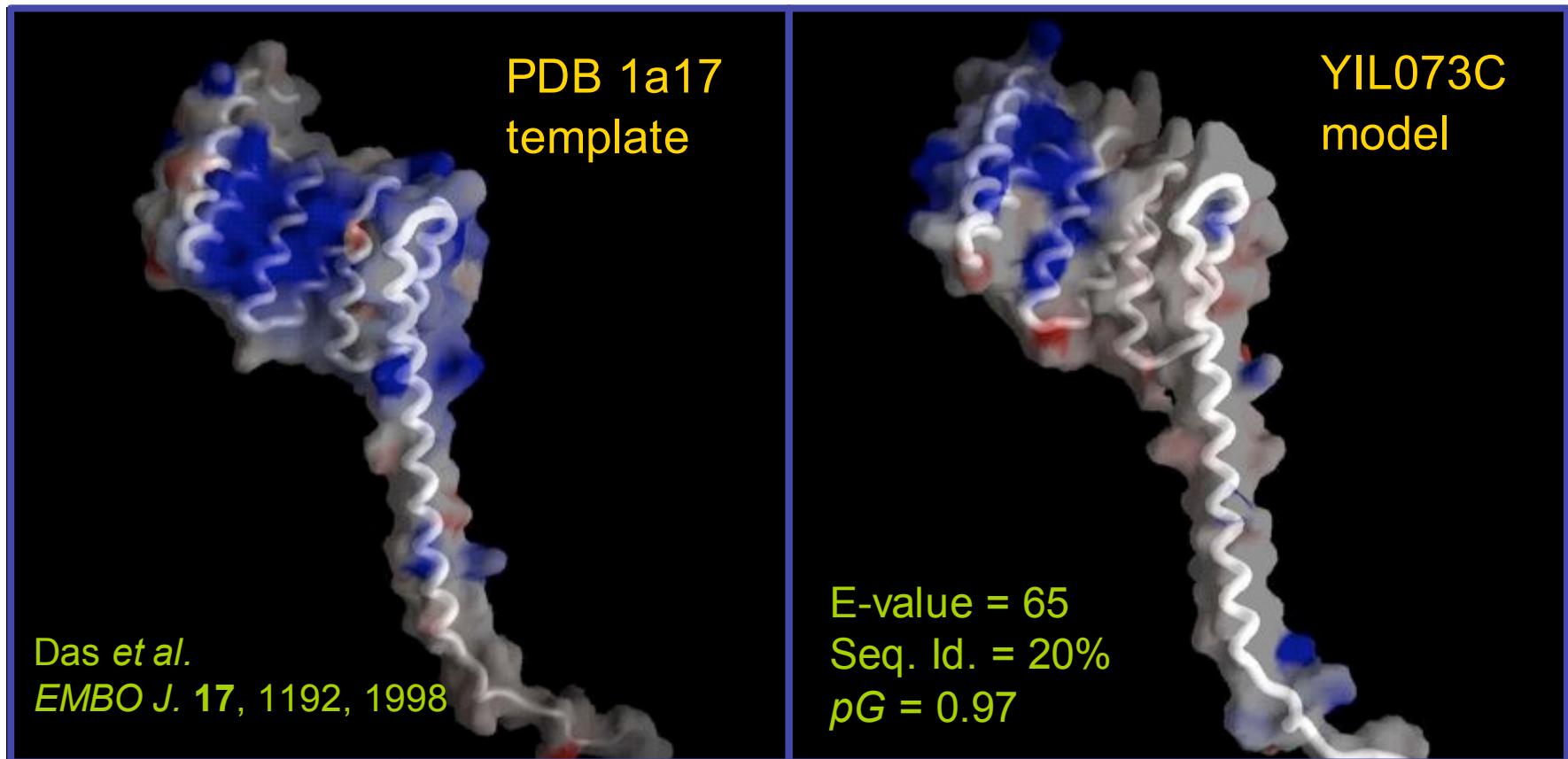
The screenshot shows a Microsoft Internet Explorer window displaying the "Protein Structure Prediction Center" website. The address bar shows the URL <http://predictioncenter.llnl.gov>. The page content includes:

- A sidebar on the left with links: CASP1, CASP2, CASP3, CASP4, CASP5 (with a red checkmark), Local services, Other links, People, Website index, and Hide menu.
- The main header "Protein Structure Prediction Center" with a subtitle "Biology and Biotechnology Research Program" and "Lawrence Livermore National Laboratory, Livermore, California, USA". To the right is a small statue of a seated figure.
- A welcome message: "Welcome to the Protein Structure Prediction Center!" followed by a paragraph about the center's goal to advance protein structure prediction methods.
- A link to the "CASP experiment": [CASP1](#) | [CASP2](#) | [CASP3](#) | [CASP4](#) | [CASP5](#).
- A section titled "Ten Most Wanted: TMW".
- Information at the bottom about the center's funding and affiliation: "The Center, supported by the National Institutes of Health, National Library of Medicine, and the U.S. Department of Energy, [Office of Biological and Environmental Research](#), is a part of the [Biology and Biotechnology Research Program](#) at the [Lawrence Livermore National Laboratory](#)".
- A footer navigation bar with links: Local services, Other links, People, and Website index.
- A footer note: "If you have any questions or comments please contact us at [squery@PredictionCenter.llnl.gov](mailto:squery@PredictionCenter.llnl.gov)".

# Fold assignment from sequence examples....



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

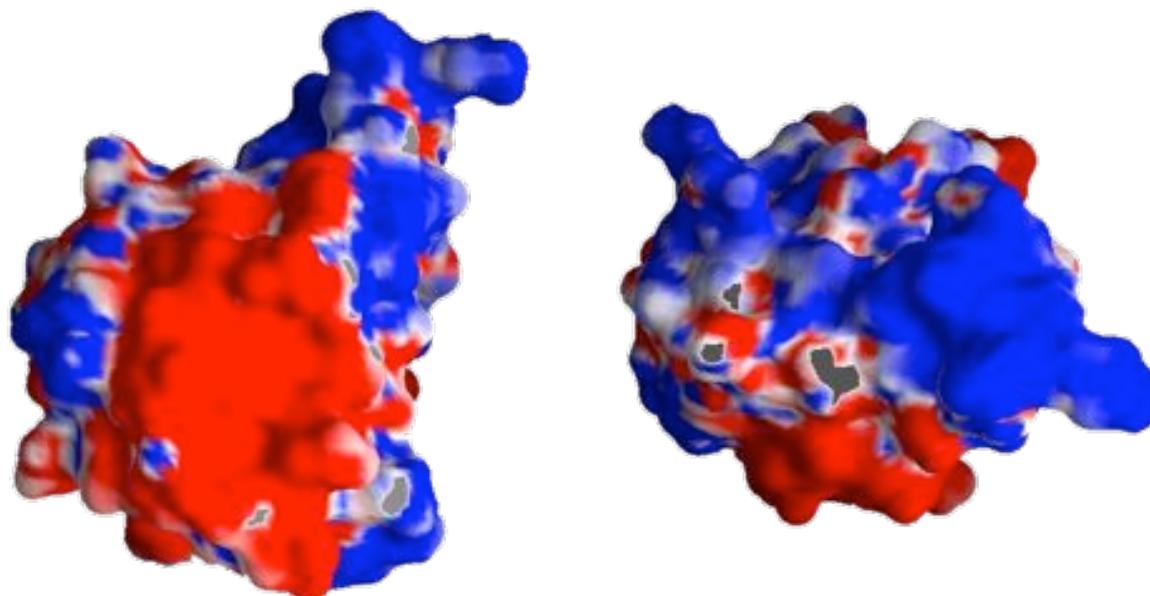
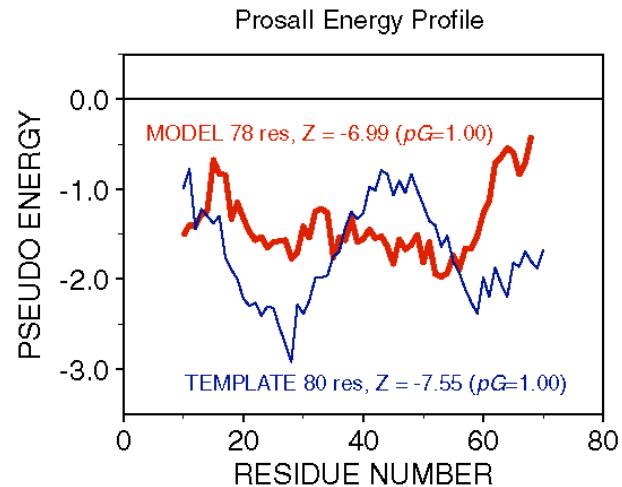


The tetratricopeptide repeat (TPR) is a degenerate 34 aa sequence identified in a variety of proteins, present in tandem arrays, mediates protein-protein interactions.

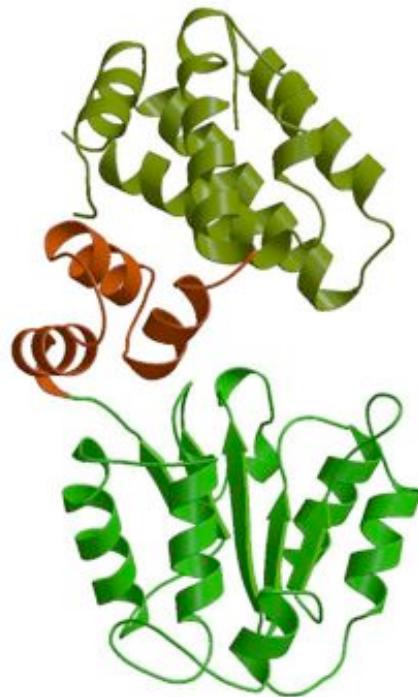
# GRAP: Fold Assignment by PSI-BLAST + Model Evaluation

(significant e-value and good model score)

SEG FILTER	QUERY	MODEL SIZE	E-VALUE	pG
Y	Target	58	0.200	0.99
N	Target	64	0.029	1.00
Y	Template	NO HIT	NO HIT	NO HIT
N	Template	78	6x10 <sup>-14</sup>	1.00



# Does RuvB have the same fold as $\delta'$ of *E.coli* DNA polymerase III? (iterative manual process... Model building $\leftarrow \rightarrow$ Model assessment )



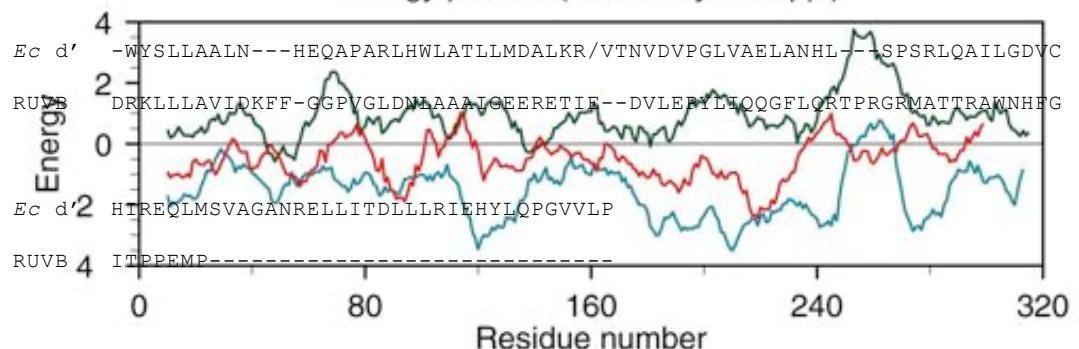
*Ec d'* MRWYPWLRLPDEKLVASYQAGRGG----HHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRG  
RUVB LEEYVGQPQVRSQMEIFIFIKAALKRGDALDHLLIFGPPGLGKTTLANIVANEMG-----

*Ec d'* CQLMQAGTHPDYYTLAPEKGKATLGVDAVREVTEKLNEAARLGGAKVVWVTDAALLTDAANALLKTL  
RUVB -----VNLRTT-----SGPVLEKAGDLAAMLTNLEPHDVLFIDEIHRLSPVVEEVLYPAM

*Ec d'* -----EEPPAETWFFLATREPERL---LATLRSRCRLHYLAPPPEQYAVTWLSRE  
PpdP EDYQLDIMIGEGPAARSIKIDLPPFTLIGATTRAGSLTSPLRDRFGIVQRLEFY--QVPDLQYIVSRS

*Ec d'* VTM-----SQDALLAALRLSAGSPGAALALFQ-----GDNWQARETLCQALAYSVPSGD--  
RUVB ARFMGLEMSDDGALEVARRARGTPRIANRLRRVRDFAEVKHDGTISADIAAQALDMLNVDAE GF DYM

Energy profiles (Prosali by M. Sippl)



# Course assignment

## The BMI206 Genome

Or your own  
sequence

### Introduction

A new highly contagious virus (known as BMI206) has been isolated and its genome sequenced. It is predicted that its genome, of about 30,000bp, codifies for 14 genes which may translate into 12 mature proteins. Your project will be to annotate and assign any structural and functional information to the proteins of the BMI206 genome. All proteins sequences codified in the BMI206 genome can be found in the [BMI206.fasta](#) file in FASTA format.

### Assignment

Write a report (2 to 4 pages) of the methods you have used, the results they produced and the conclusions you reached. The following questions can guide you while you do your research. However, there is no specific set of answers you are required to get as the goal is for you to address each question as well as you can. There are several "correct" answers to some of these questions.

- Actually, this is a genome from an existing virus. Which one?
- How many proteins in the BMI206 genome are membrane proteins?
- Is there a single known structure in the PDB that can be clearly associated to any of the proteins in the BMI206 genome?
- For how many of the sequences can you provide a fold assignment?
- Can you locate a putative binding site on the predicted folds?
- Do the findings you made make sense given that the genome is from a virus? For example, does any of your assignments contain a viral envelop proteins?, do you find polyprotein processing machinery?, etc...

**GRADING: The entire assignment is worth 20 points.**