BMI-206

Structure-Structure comparisons
Sequence-Structure comparisons

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

February 17th, 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”

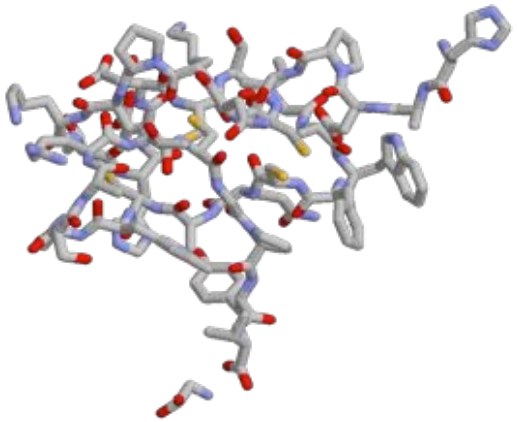
Structure-Structure alignments

As any other bioinformatics problem…

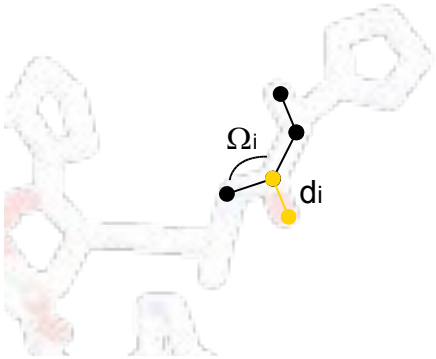
- **Representation**
- **Scoring**
- **Optimizer**

Representation

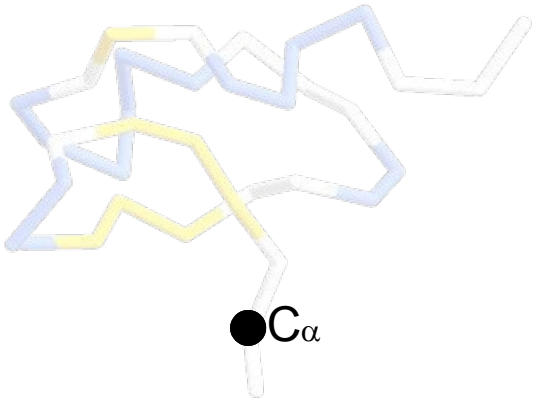
Structures



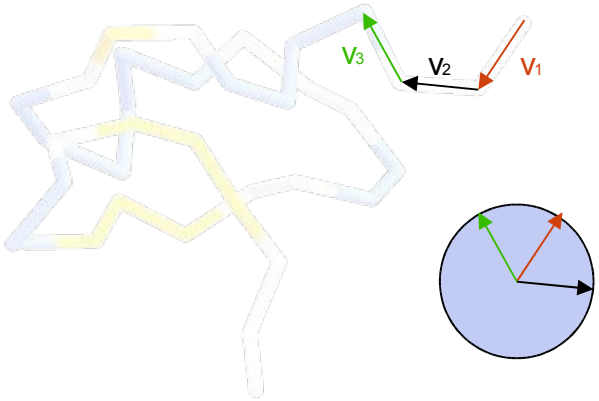
All atoms and coordinates



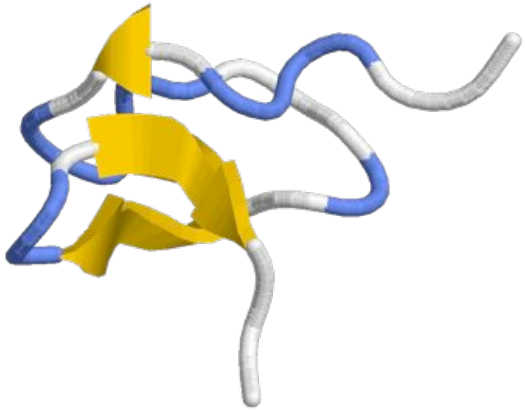
Dihedral space or distance space



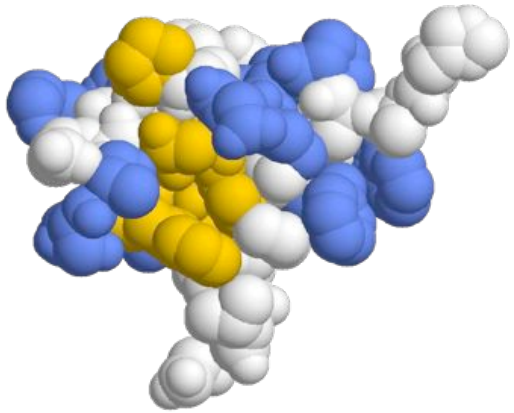
Reduced atom representation



Vector representation



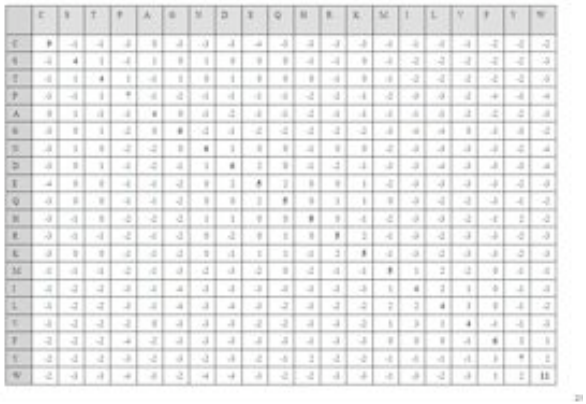
Secondary Structure



Accessible surface (and others)

Scoring

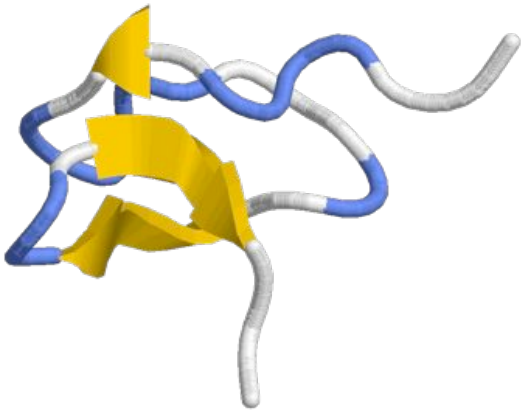
Raw scores



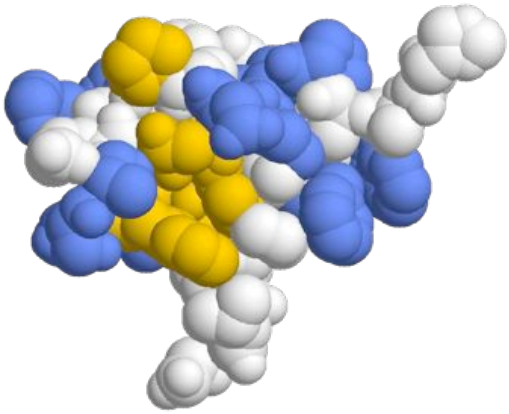
Aminoacid substitutions

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

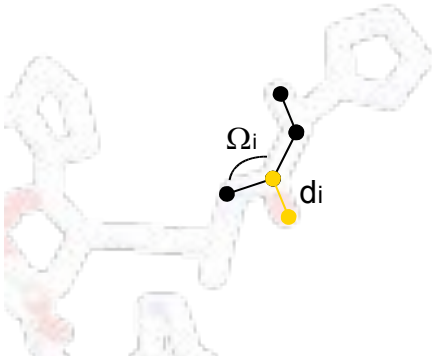
Root Mean Square Deviation



Secondary Structure (H,B,C)



Accessible surface (B,A [%])

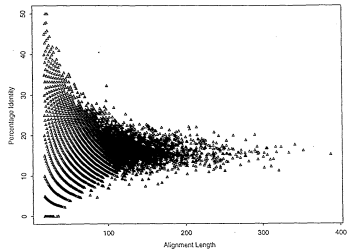


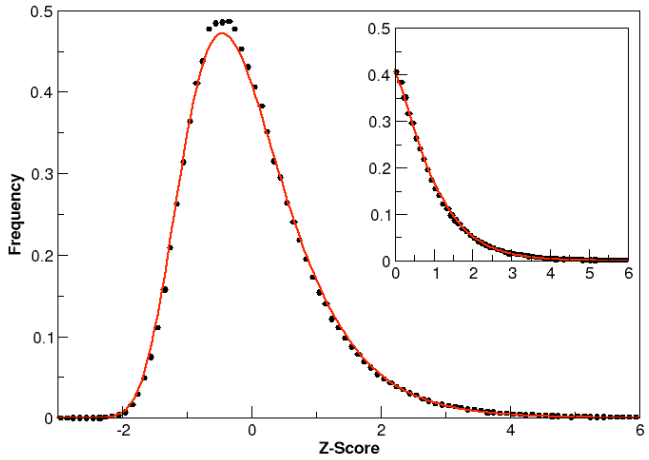
Angles or distances

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).

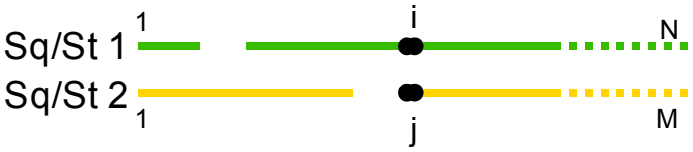
Analytic

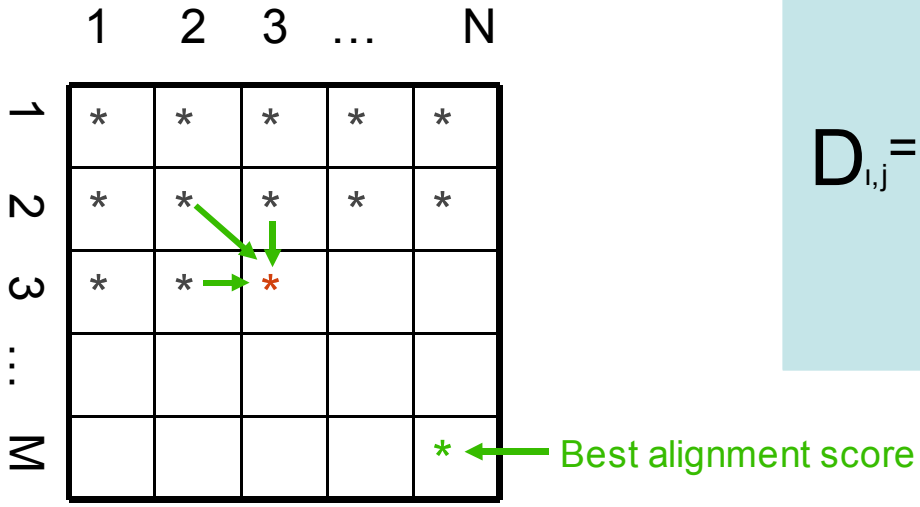
$$P(s) = e^{-\lambda (s-\mu)}$$

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

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

Global dynamic programming alignment



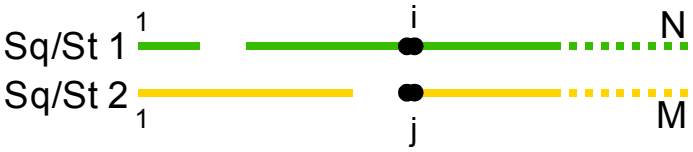


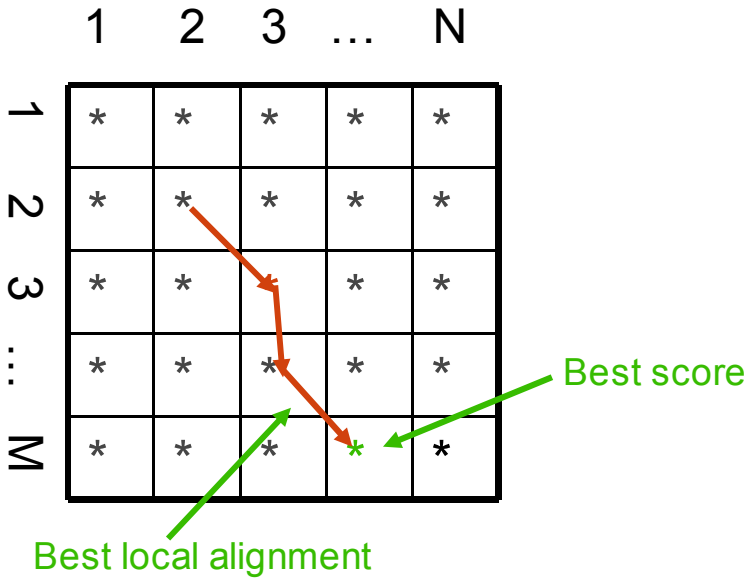
$$D_{i,j} = \min \begin{cases} D_{i,j-1} + Score_{(\Delta, rj)} \\ D_{i-1,j-1} + Score_{(ri, rj)} \\ D_{i-1,j} + Score_{(ri, \Delta)} \end{cases}$$

Backtracking to get the best alignment

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

Local dynamic programming alignment





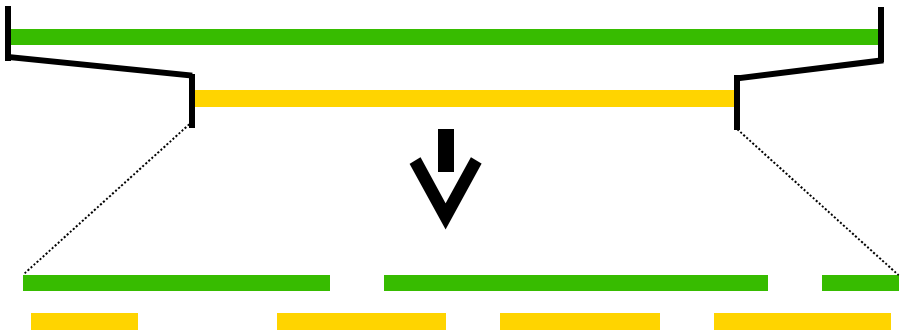
$$D_{i,j} = \min \begin{cases} D_{i,j-1} + Score_{(\Delta, rj)} \\ D_{i-1,j-1} + Score_{(ri, rj)} \\ D_{i-1,j} + Score_{(ri, \Delta)} \\ 0 \end{cases}$$

Backtracking to get the best alignment

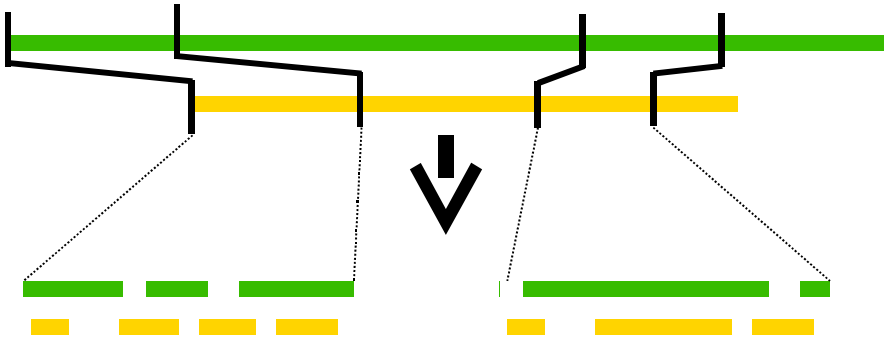
Smith and Waterman (1981) *J. Mol Biol*, **147** pp195

Optimizer

Global *.vs.* local alignment



Global alignment



Local alignment

Multiple alignment

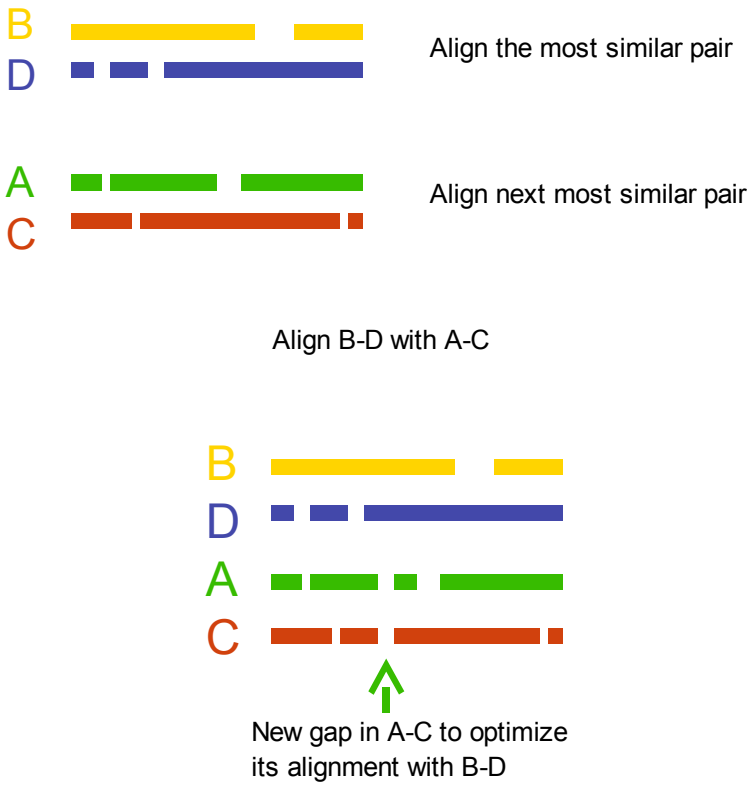
Pairwise alignments

Example – 4 sequences A, B, C, D.

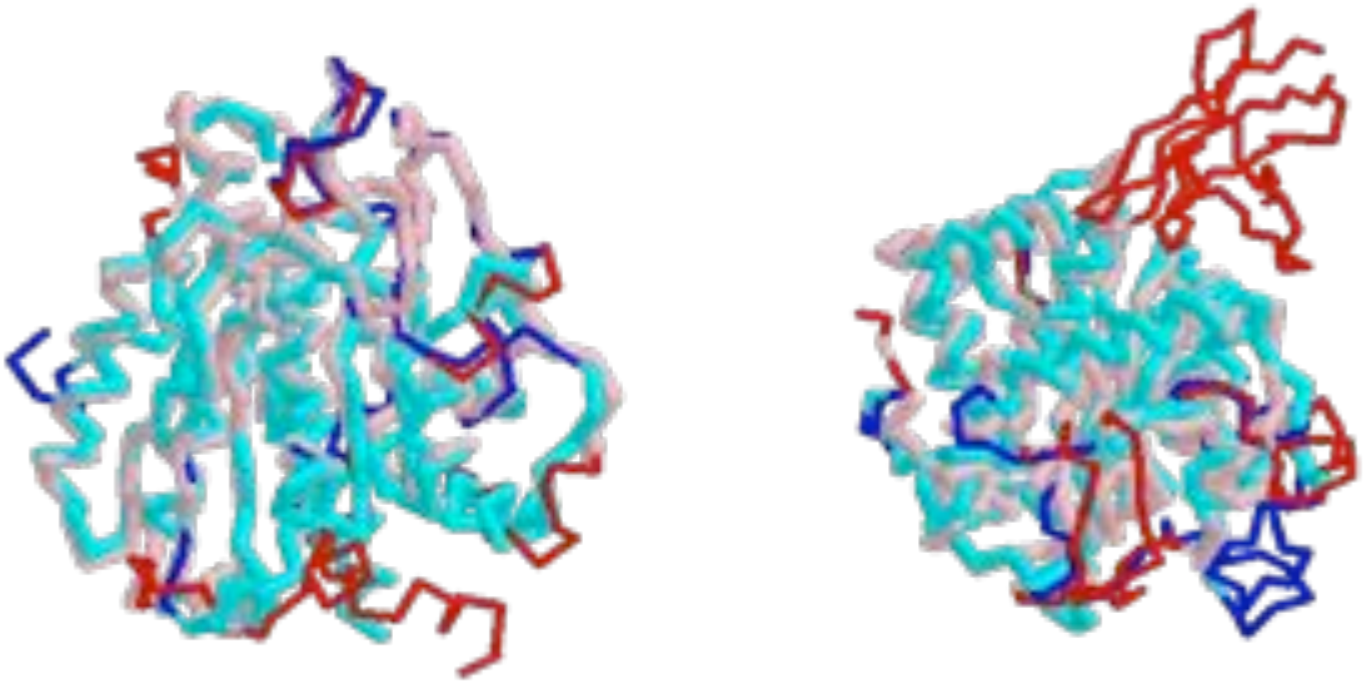
6 pairwise comparisons
then cluster analysis

Multiple alignments

Following the tree from step 1



Coverage .vs. Accuracy

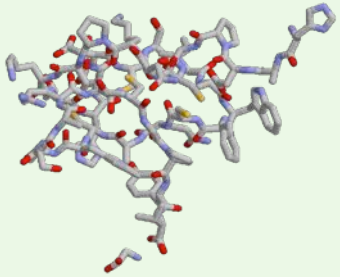


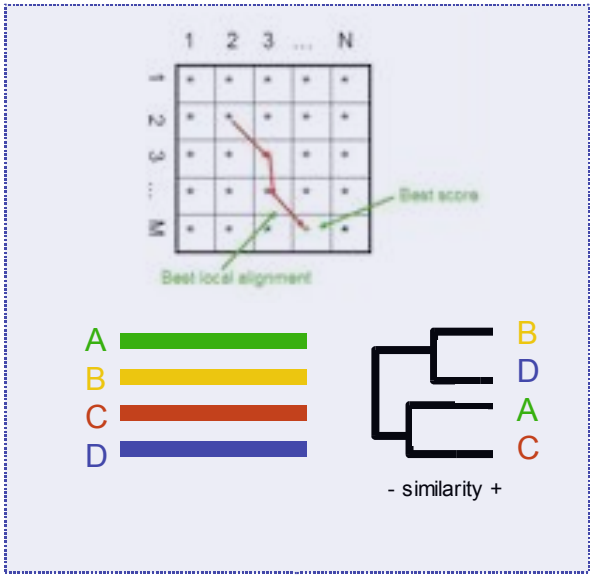
Same RMSD ~ 2.5Å

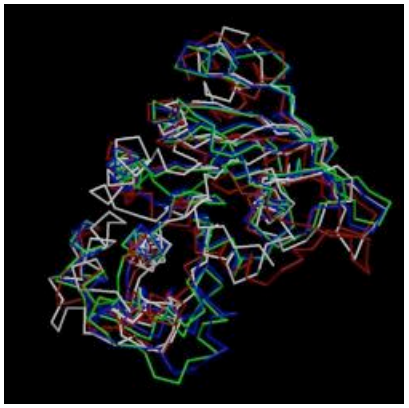
Coverage ~90% Cα

Coverage ~75% Cα

Structural alignment by properties conservation (SALIGN-MODELLER)

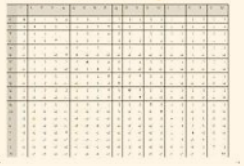




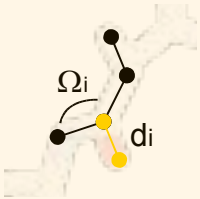


✓ Uses all available structural information
✓ Provides the optimal alignment

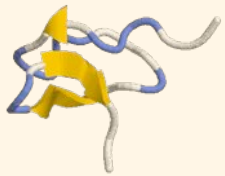
✗ Computationally expensive



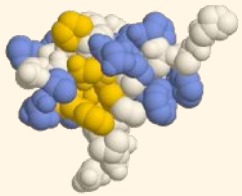
$R_{i,j}$



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



$S_{i,j}$



$B_{i,j}$

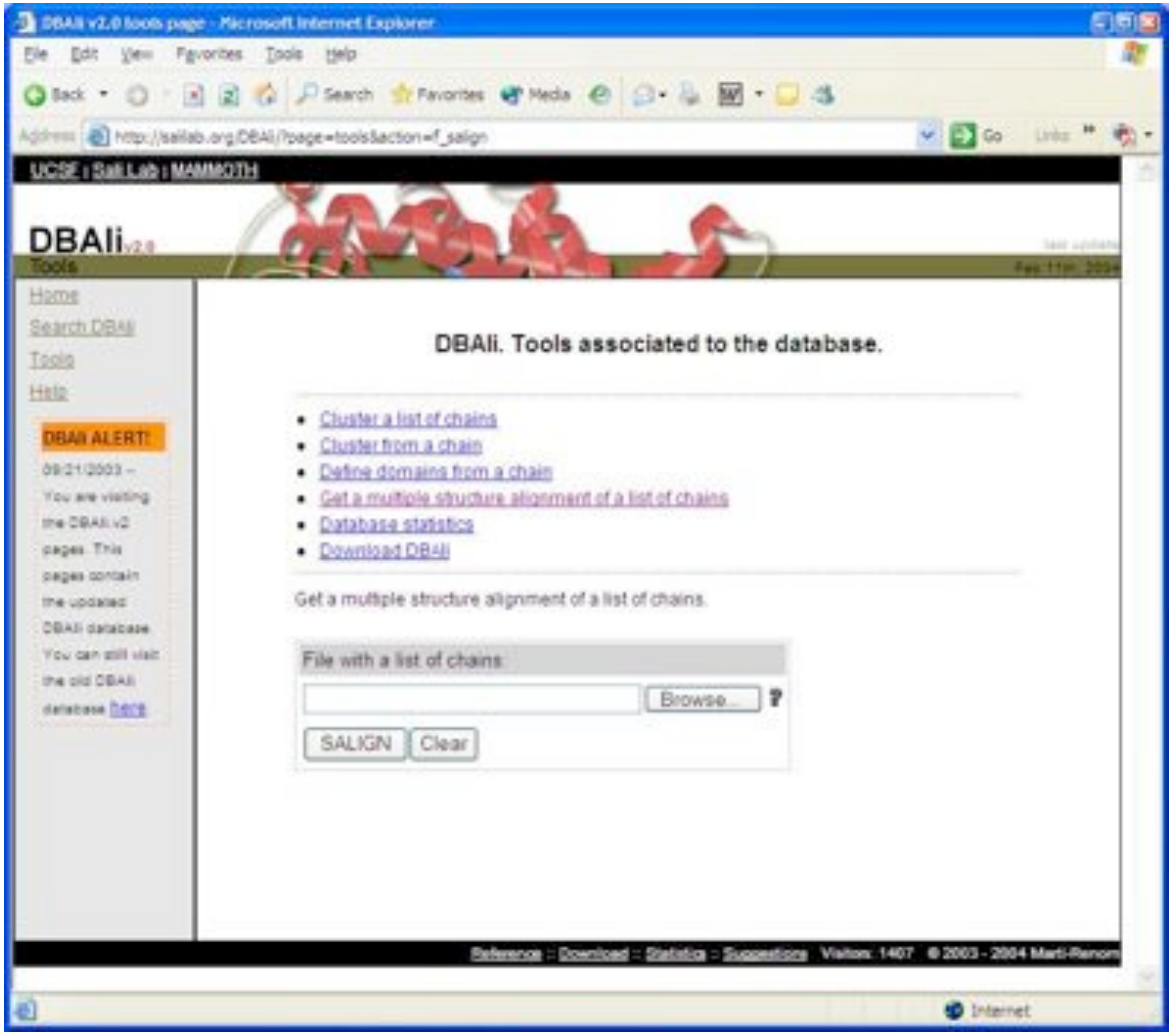
$$RMSD = \sqrt{\sum\left(x_i - \bar{x}\right)^2}$$

$I_{i,j}$

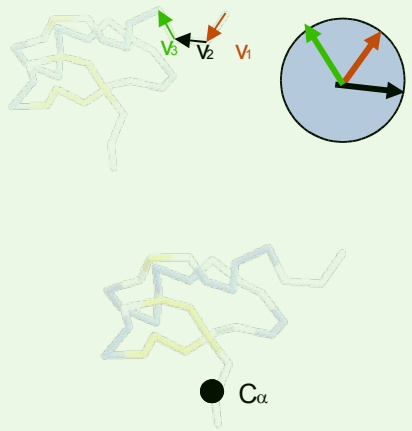
Madhusudhan et al. *in preparation*

Structural alignment by properties conservation (SALIGN-MODELLER)

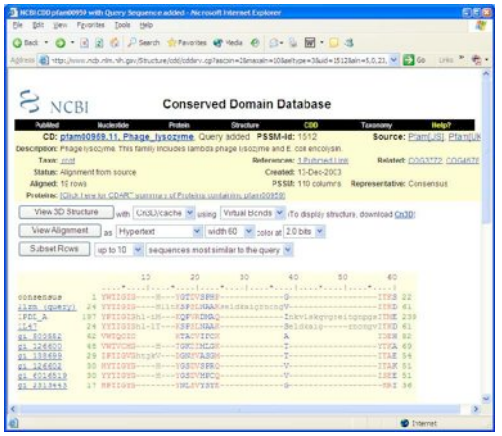
http://www.salilab.org/dbali/



Vector Alignment Search Tool (VAST)

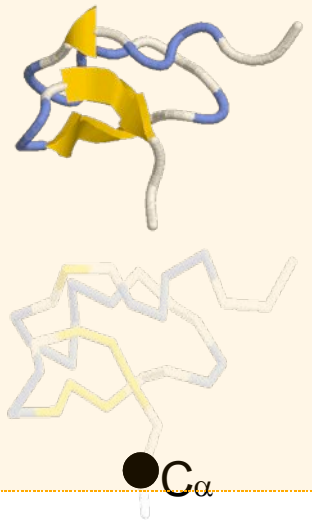


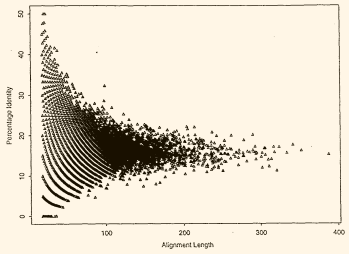
- Graph theory search of similar SSE
- Refining by Monte Carlo at all atom resolution



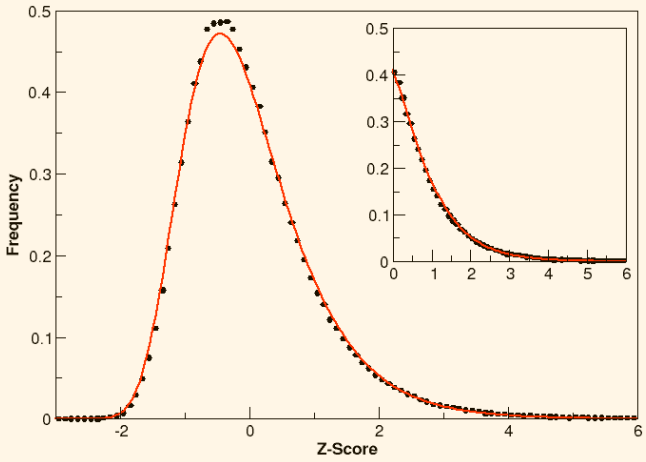
✓ **Good scoring system with significance**

✘ **Reduces the protein representation**





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

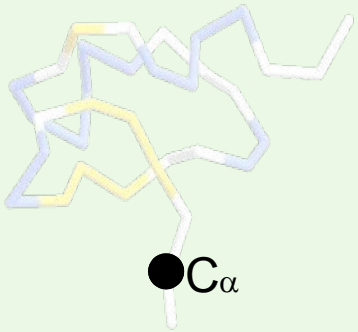


Gibrat JF et al. (1996) *Curr Opin Struct Biol* **3** pp377

Vector Alignment Search Tool (VAST)

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

Incremental combinatorial extension (CE)



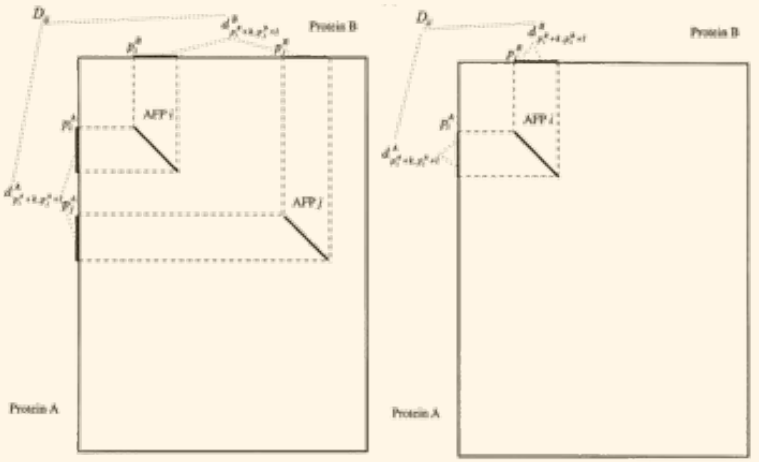
- Exhaustive combination of fragments
- Longest combination of AFPs
- Heuristic similar to PSI-BLAST

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

✗ **Complicated scoring and heuristics**

8 residues peptides

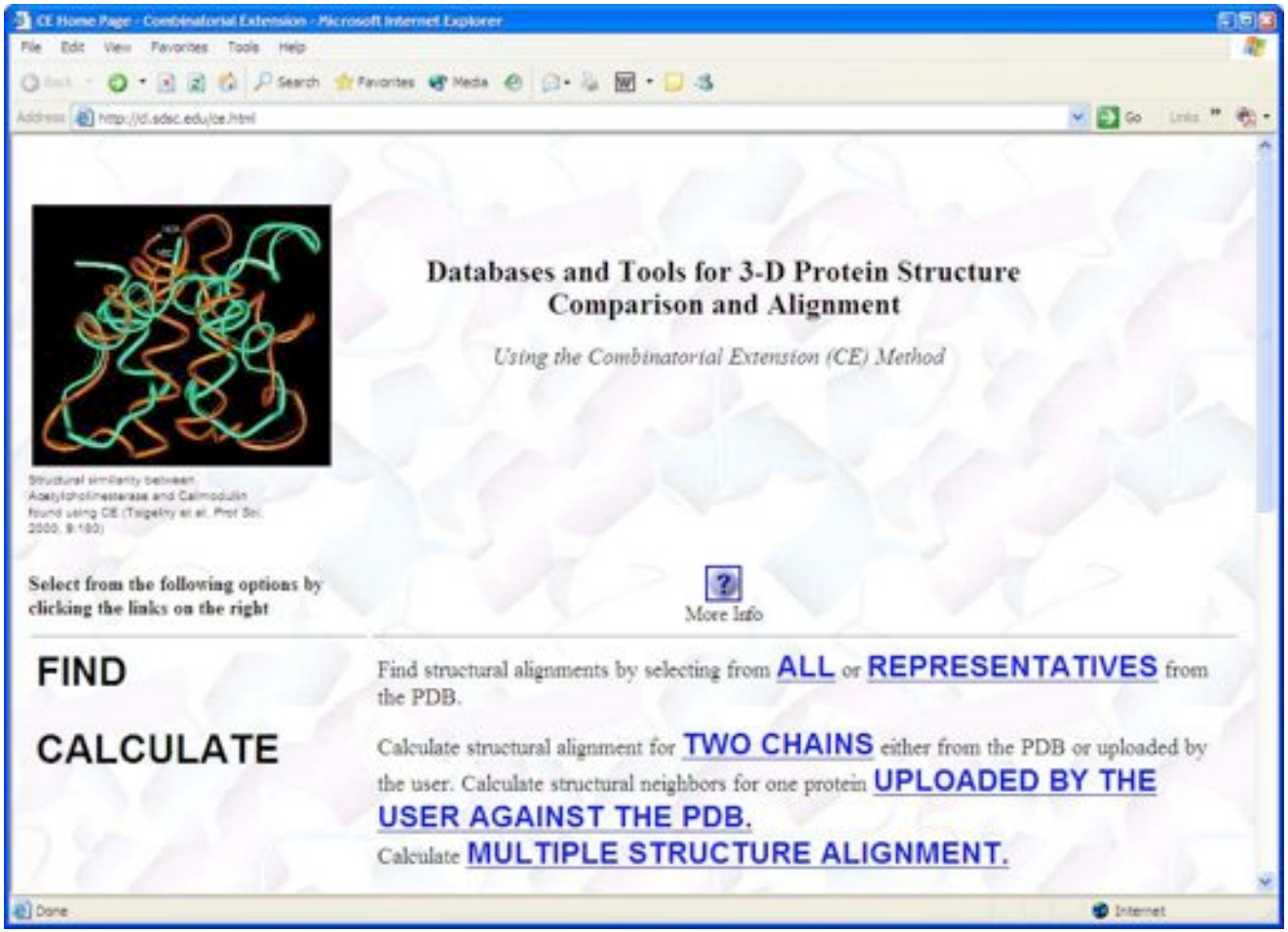
$$RMSD = \sqrt{\sum\left(x_i - \bar{x}\right)^2}$$



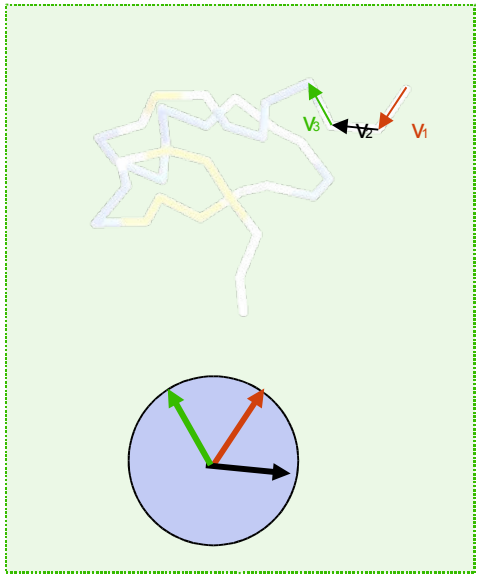
Shindyalov IN, amd Bourne PE. (1998) *Protein Eng.* **9** pp739

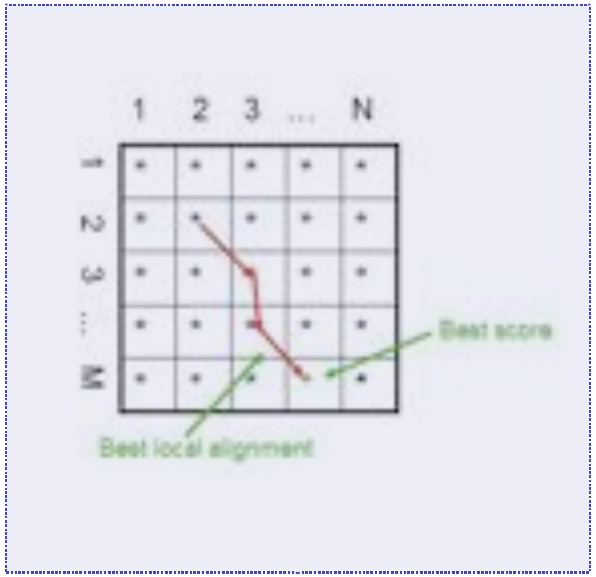
Incremental combinatorial extension (CE)

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



Matching molecular models obtained from theory (MAMMOTH)



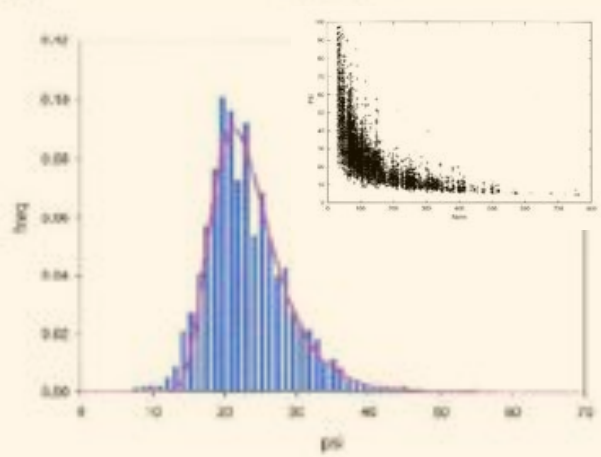


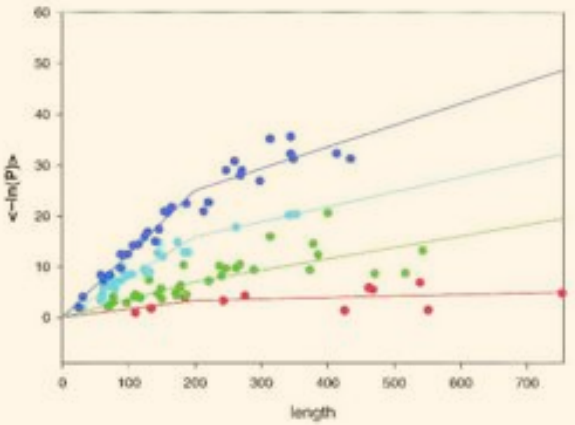
✓ **VERY FAST!**
✓ **Good scoring system with significance**

✘ **Reduces the protein representation**

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

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

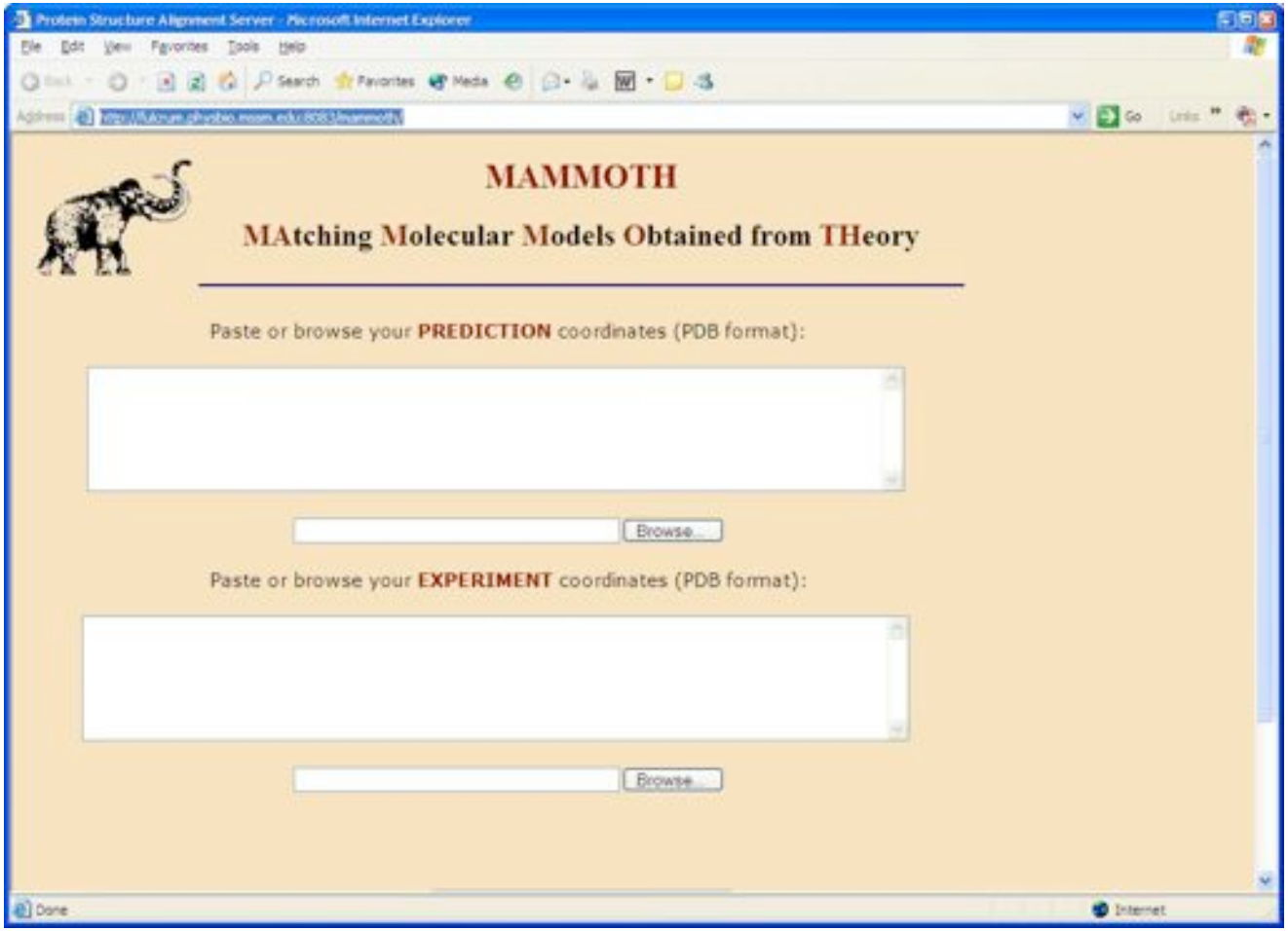




Ortiz AR, (2002) *Protein Sci.* **11** pp2606

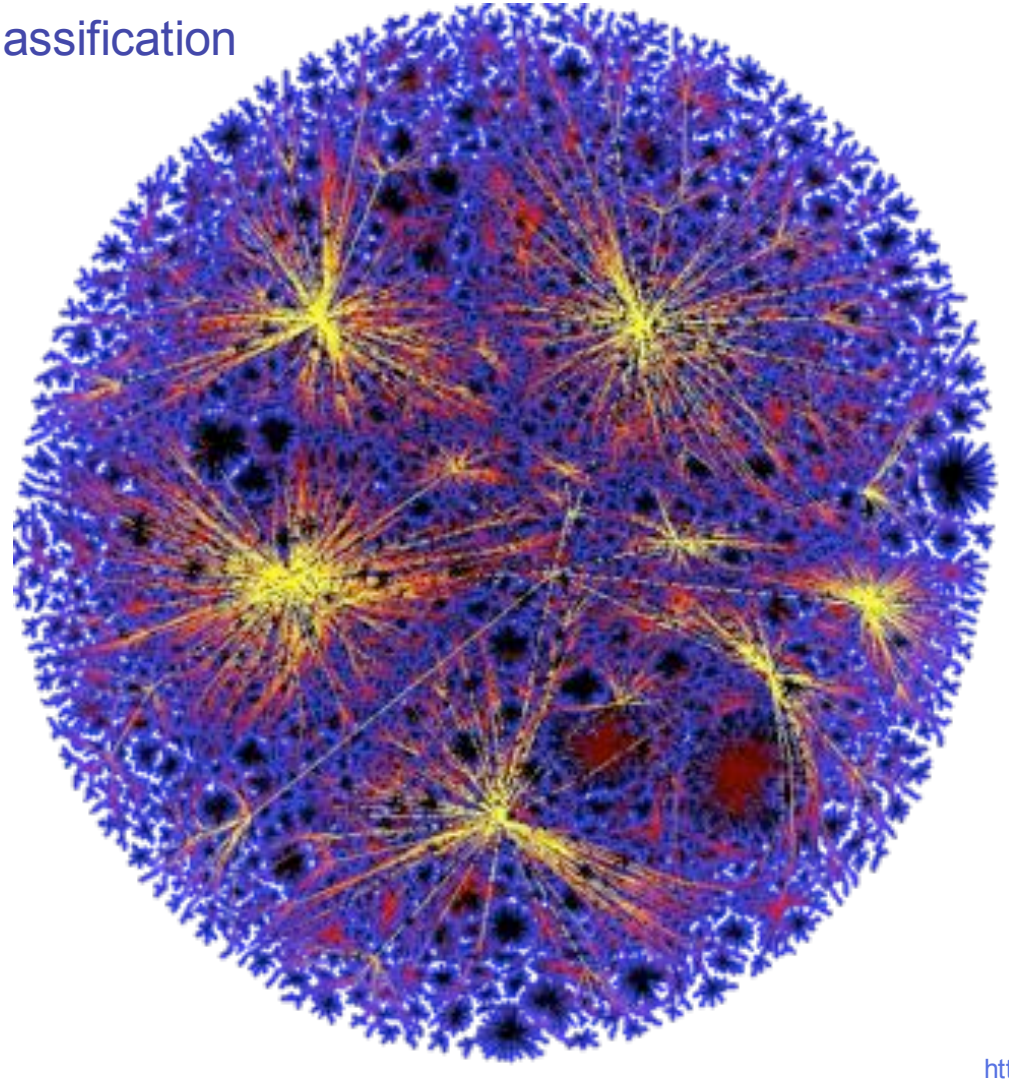
Matching molecular models obtained from theory (MAMMOTH)

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



Classification of the structural space

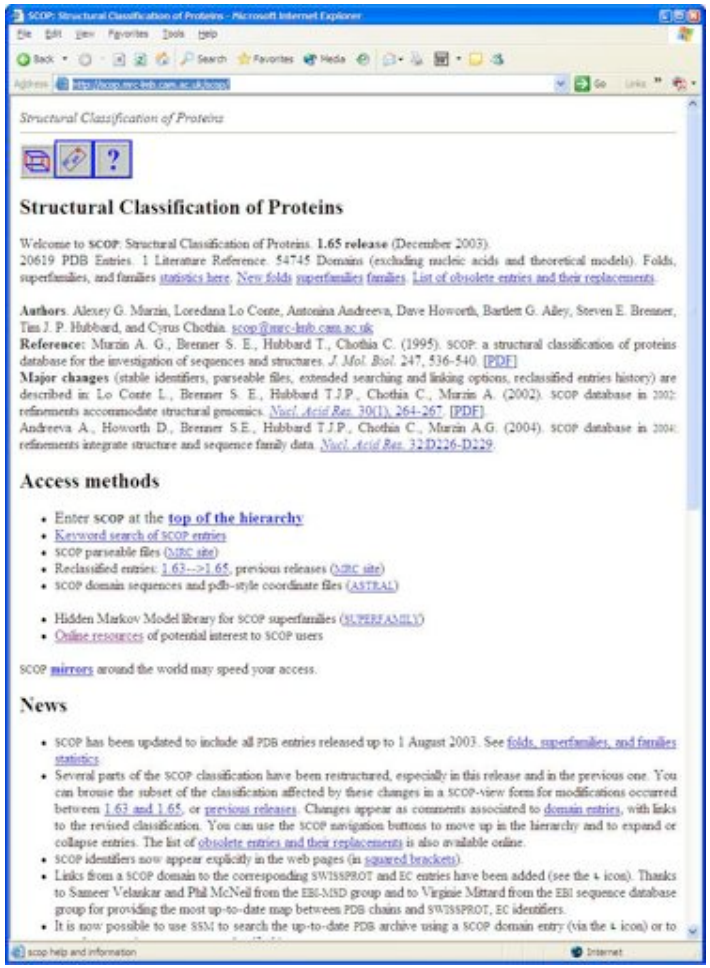
SCOP classification



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.

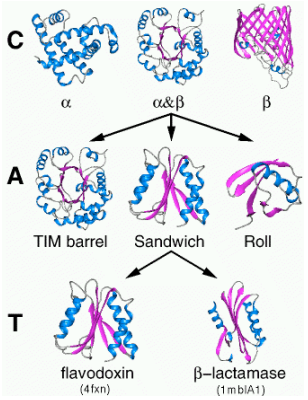
CATH2.5.1 database

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**

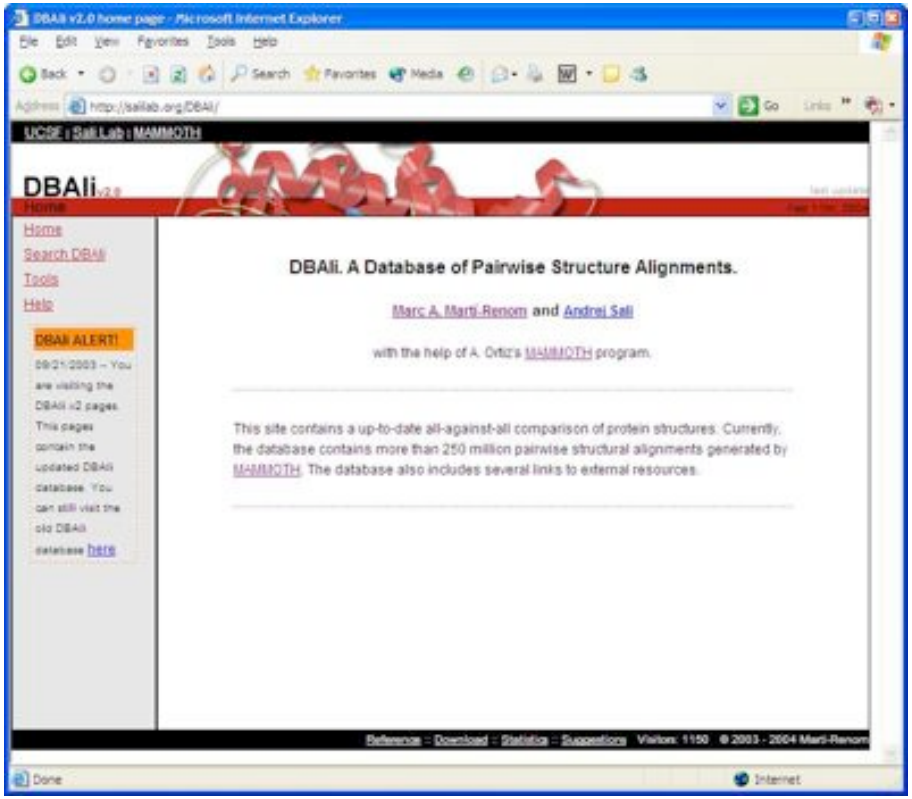


Version	2.5.1						
Date	28-01-2004						
							D
Mainly Alpha	5	227	428	948	1713	3946	10155
Mainly Beta	19	139	292	951	2344	5011	14259
Alpha Beta	12	368	648	2010	3631	8639	23025
Few Secondary Structures	1	86	91	114	225	378	952
Multi-domain chains	1	1053	1057	1071	2186	5801	12471
Preliminary single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

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

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**

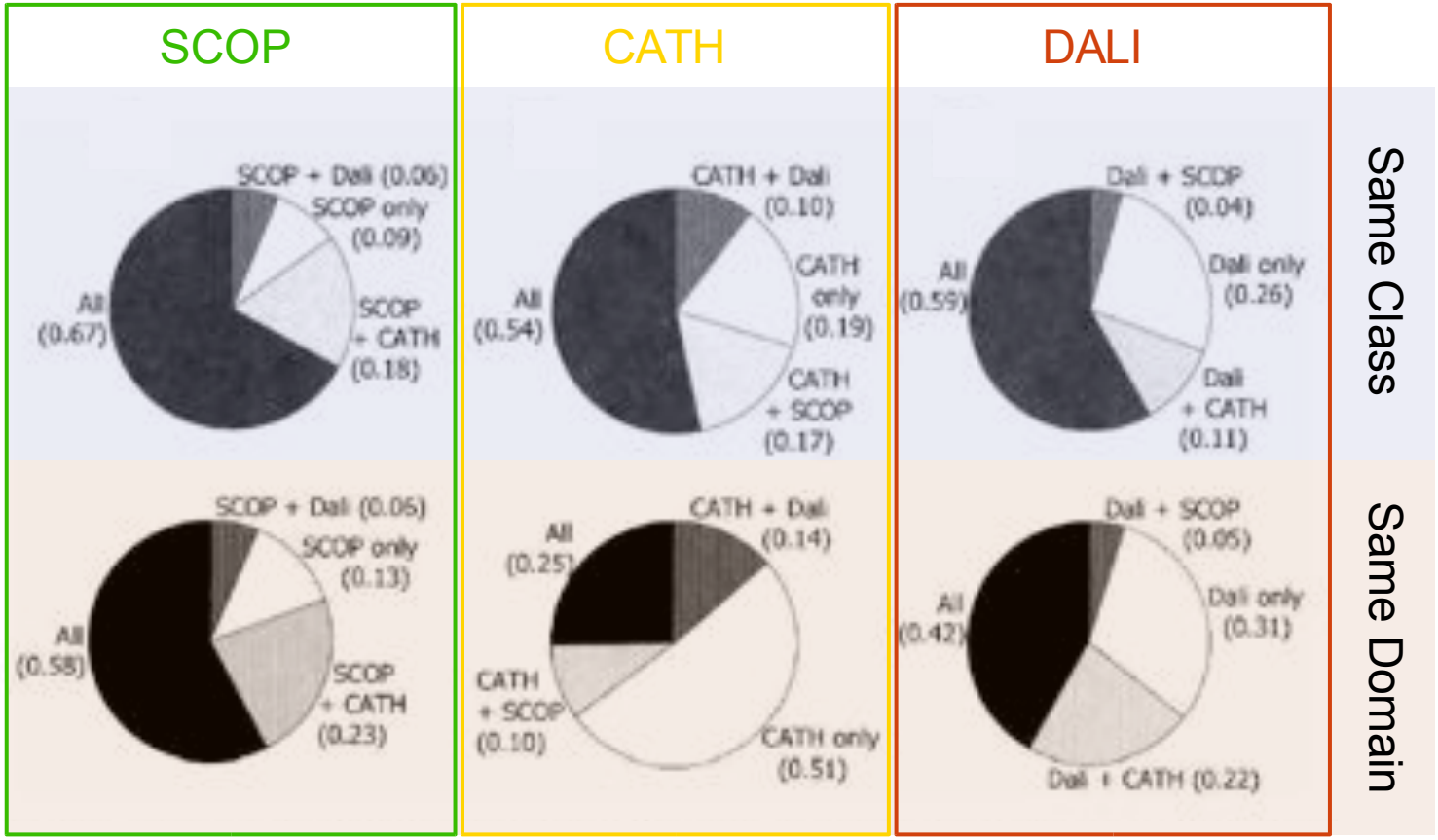
Last updated:	February 11th, 2004 (18:49h)
Number of chains in database:	48,094
Number of structure-structure comparisons:	330,514,636

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

Classification of the structural space

Not an easy task!

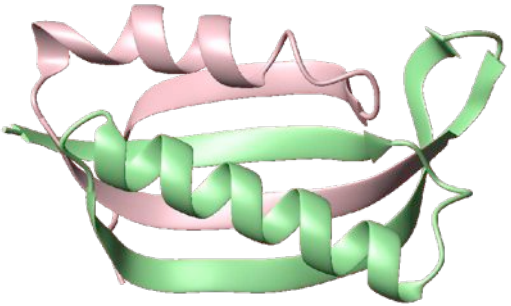
Domain definition AND domain classification

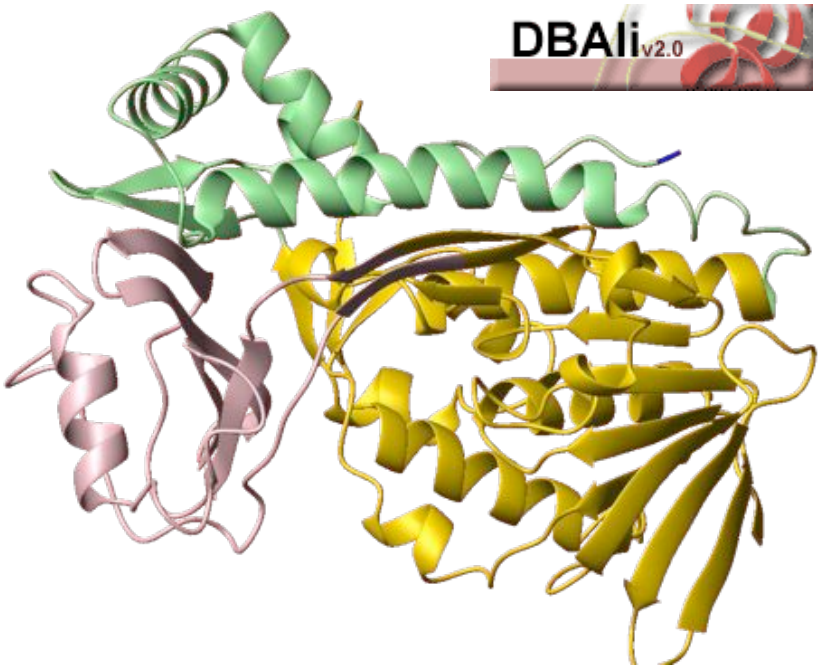


Day, et al. (2003) *Protein Sciences*, **12** pp2150

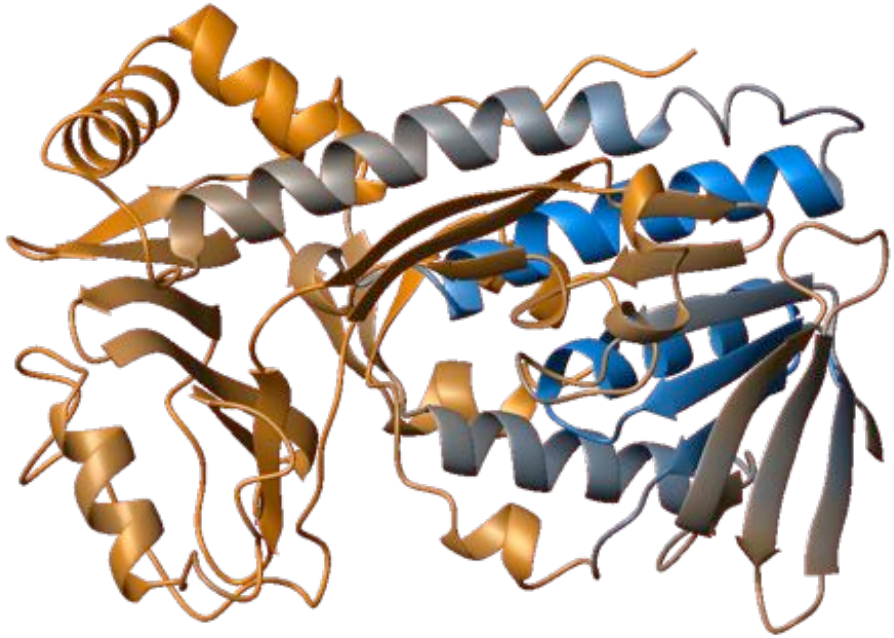
Application (ModDom)

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

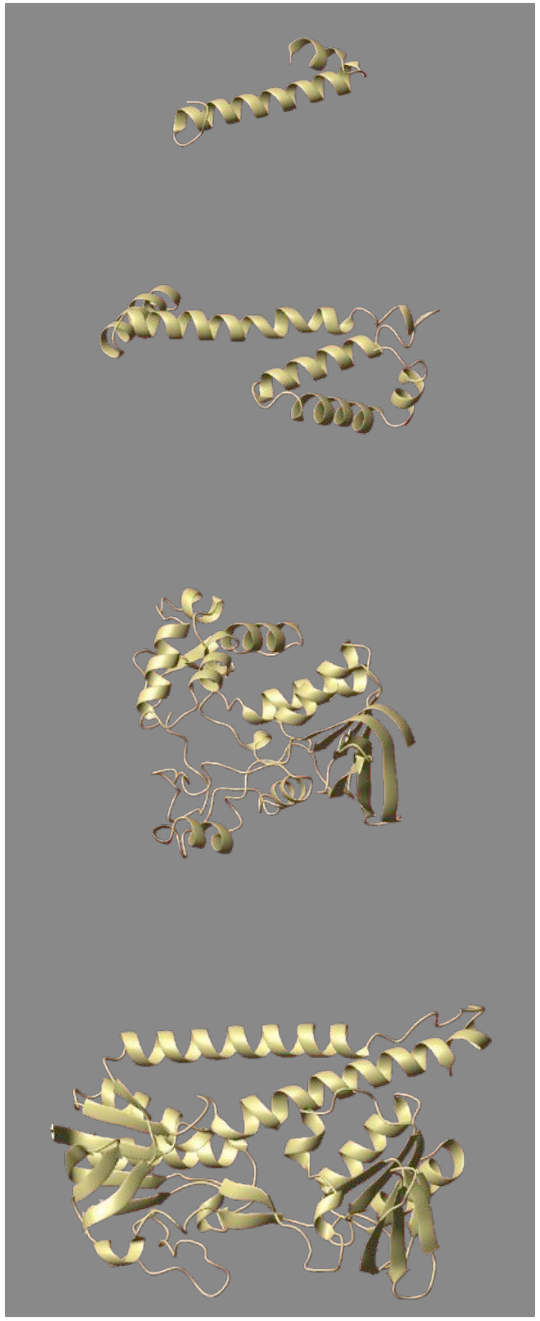


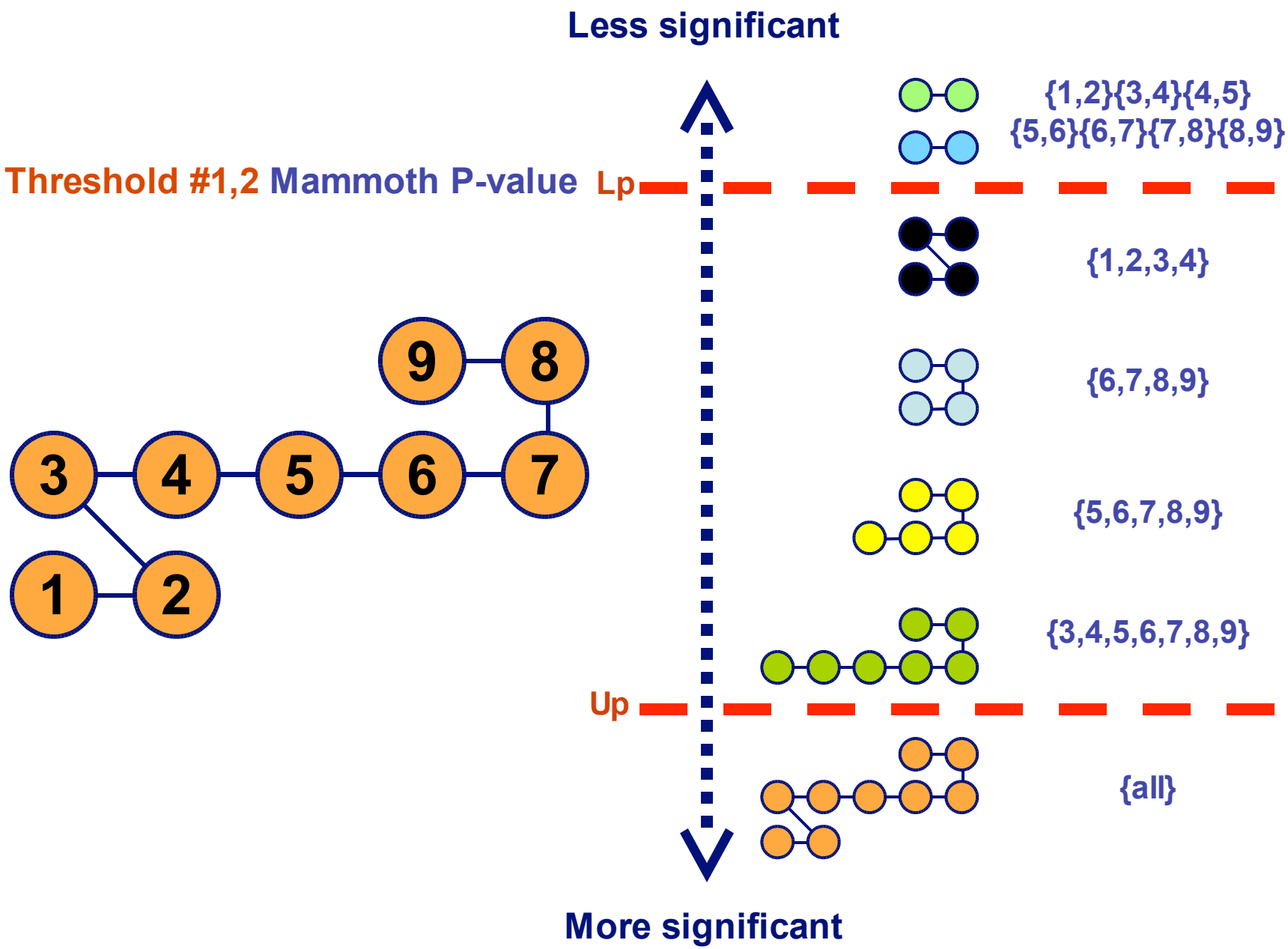


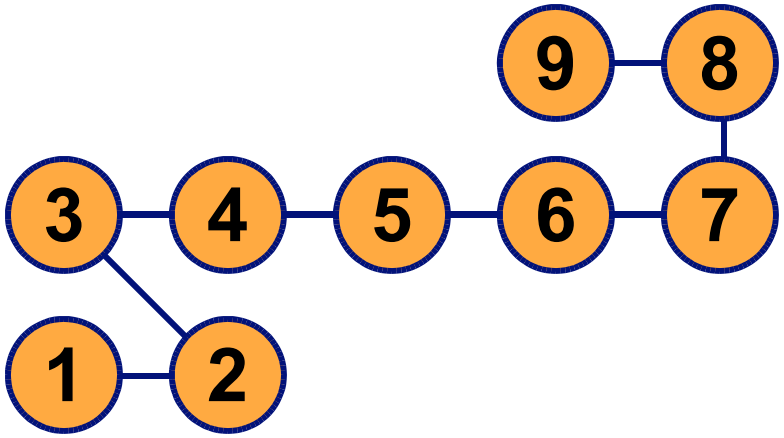
ModDom



1phh (Oxydoreductase from *Pseudomonas fluorescens*)

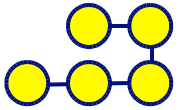



Less significant
Threshold #1,2 Mammoth P-value
Lp
{1,2}{3,4}{4,5}
{5,6}{6,7}{7,8}{8,9}
{1,2,3,4}
{6,7,8,9}
{5,6,7,8,9}
{3,4,5,6,7,8,9}
Up
{all}
More significant
9
8
3
4
5
6
7
1
2

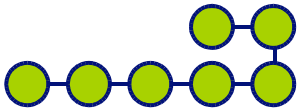


{1,2,3,4}

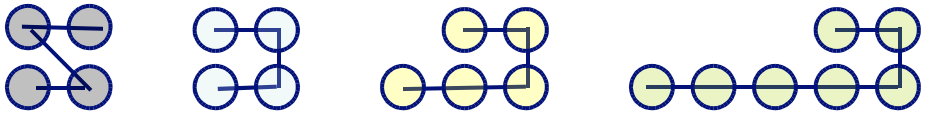
{6,7,8,9}

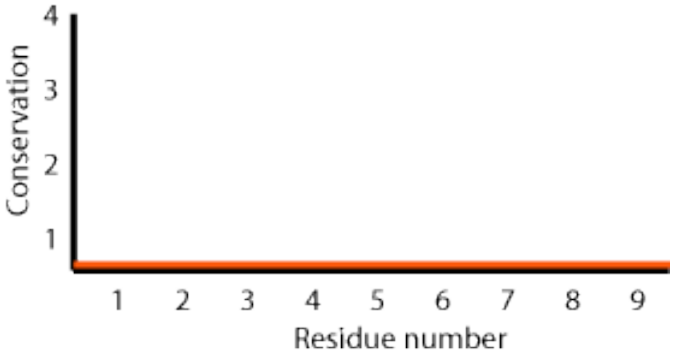


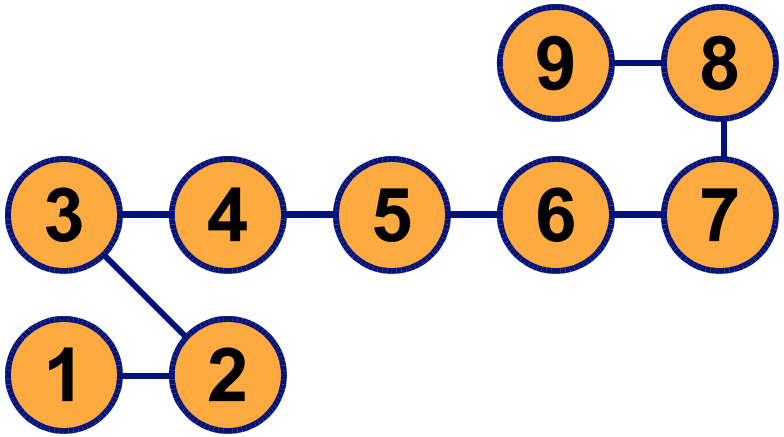
{5,6,7,8,9}



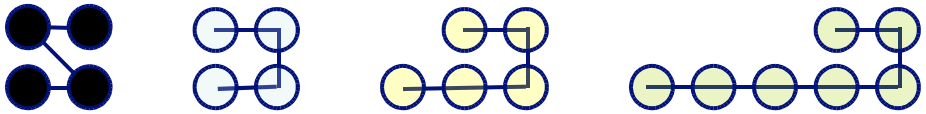
{3,4,5,6,7,8,9}

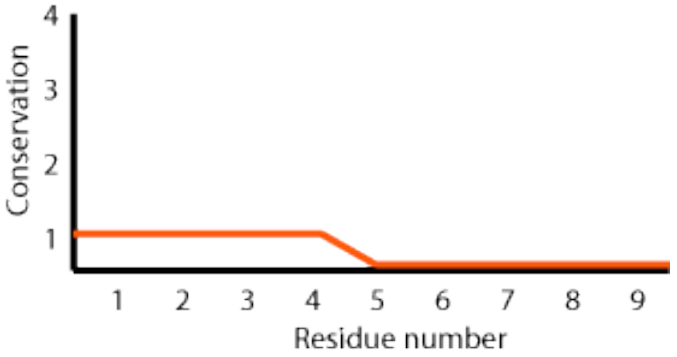


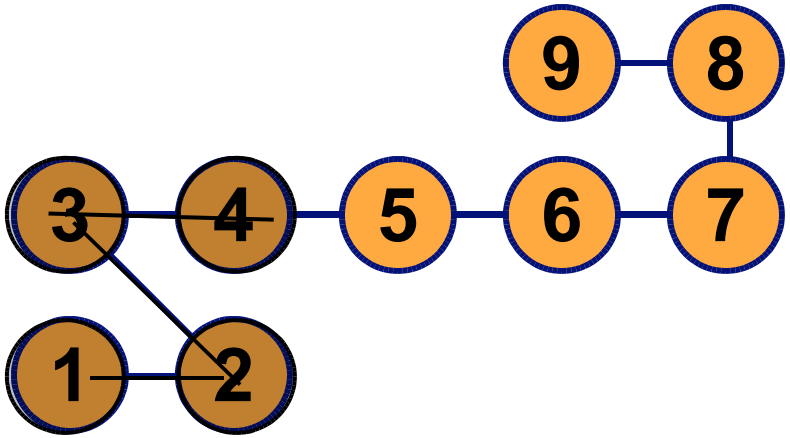




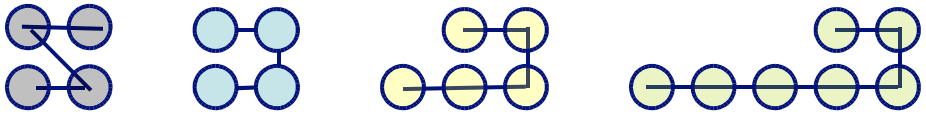
#	1	2	3	4	5	6	7	8	9
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0

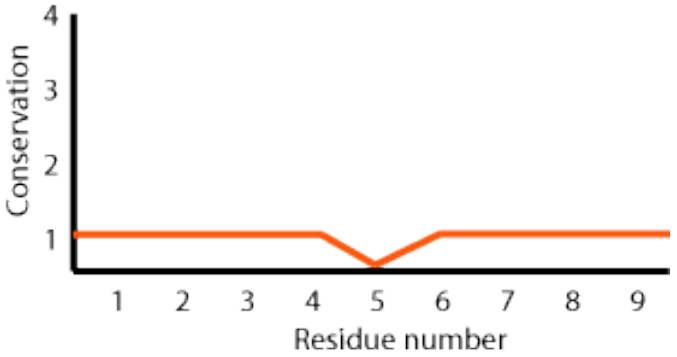


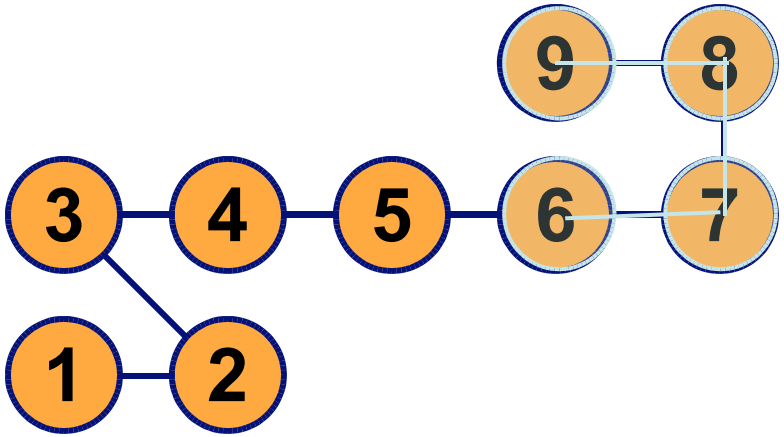




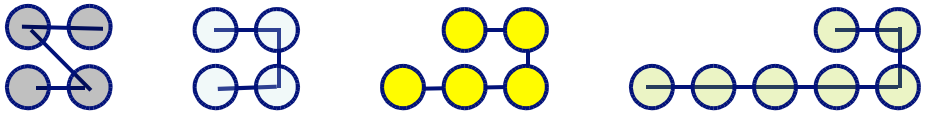
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0

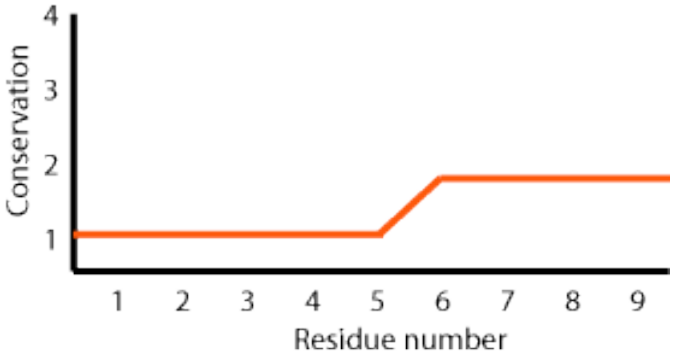


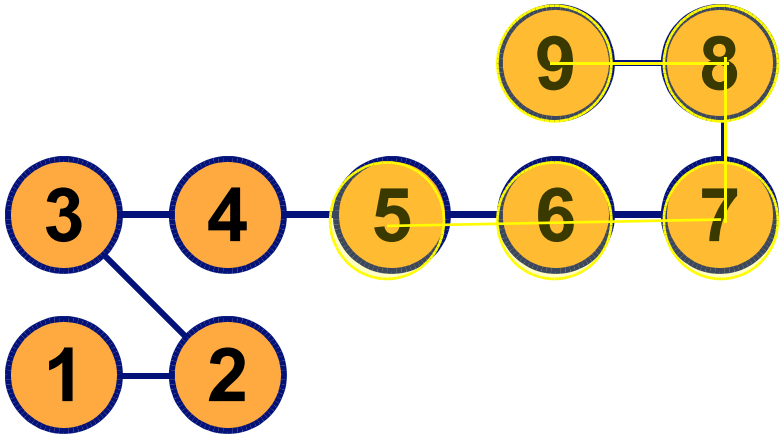




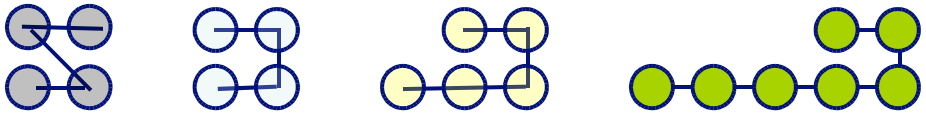
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	1	1	1
7	0	0	0	0	0	1	1	1	1
8	0	0	0	0	0	1	1	1	1
9	0	0	0	0	0	1	1	1	1

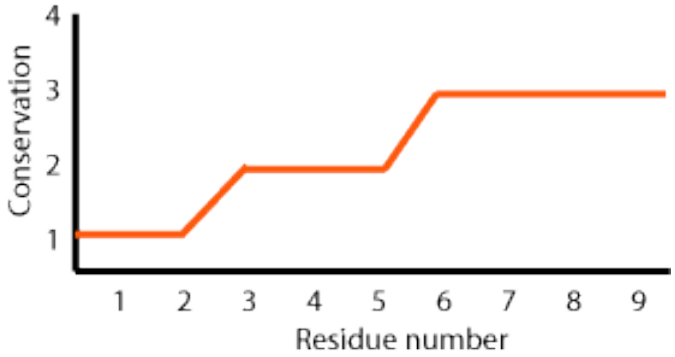


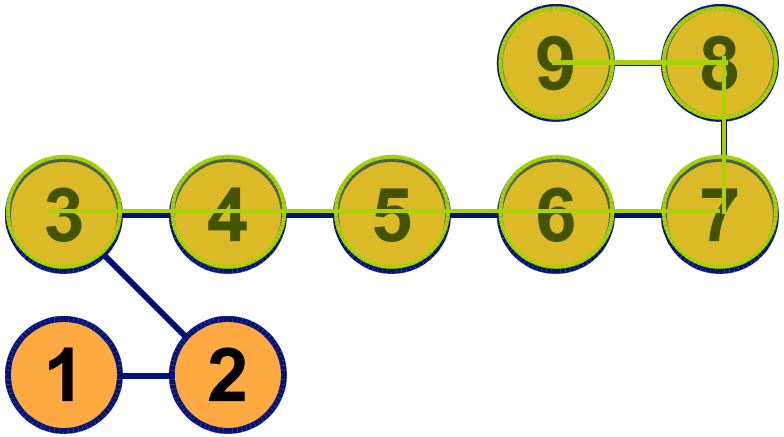




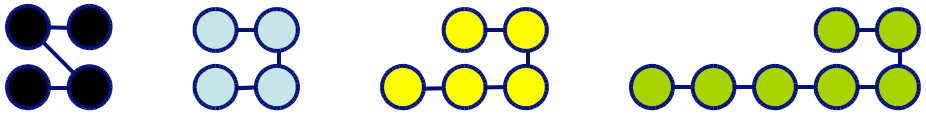
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	0	0	0	0	1	1	1	1	1
6	0	0	0	0	1	2	2	2	2
7	0	0	0	0	1	2	2	2	2
8	0	0	0	0	1	2	2	2	2
9	0	0	0	0	1	2	2	2	2

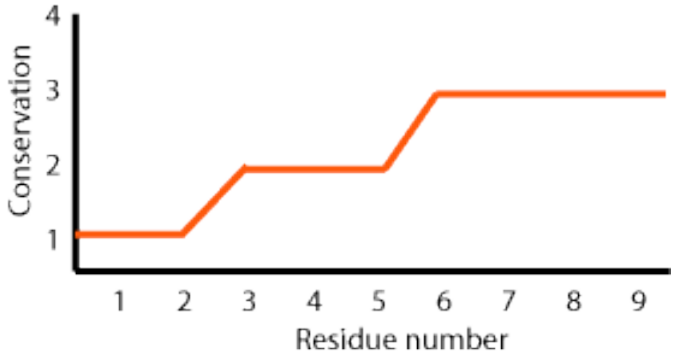


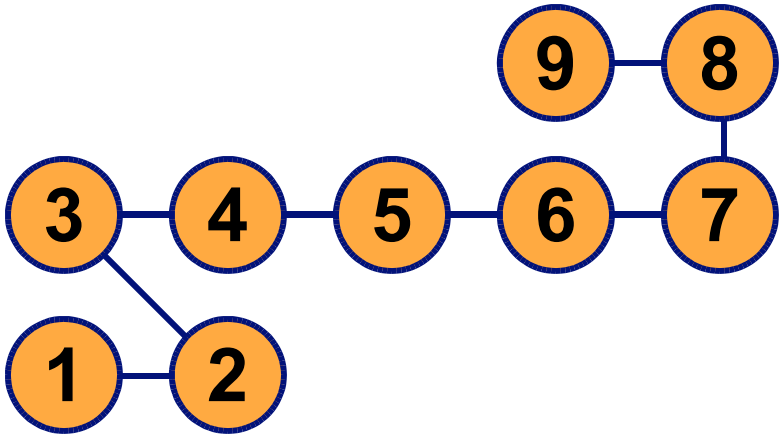




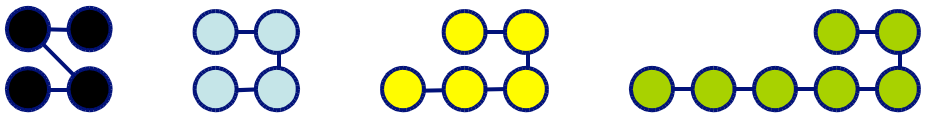
#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	2	2	1	1	1	1	1
4	1	1	2	2	1	1	1	1	1
5	0	0	1	1	2	2	2	2	2
6	0	0	1	1	2	3	3	3	3
7	0	0	1	1	2	3	3	3	3
8	0	0	1	1	2	3	3	3	3
9	0	0	1	1	2	3	3	3	3

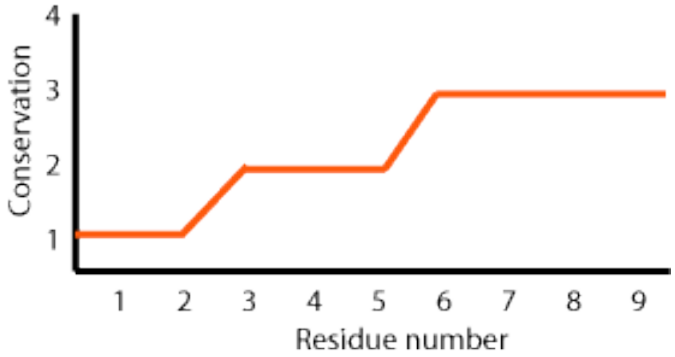


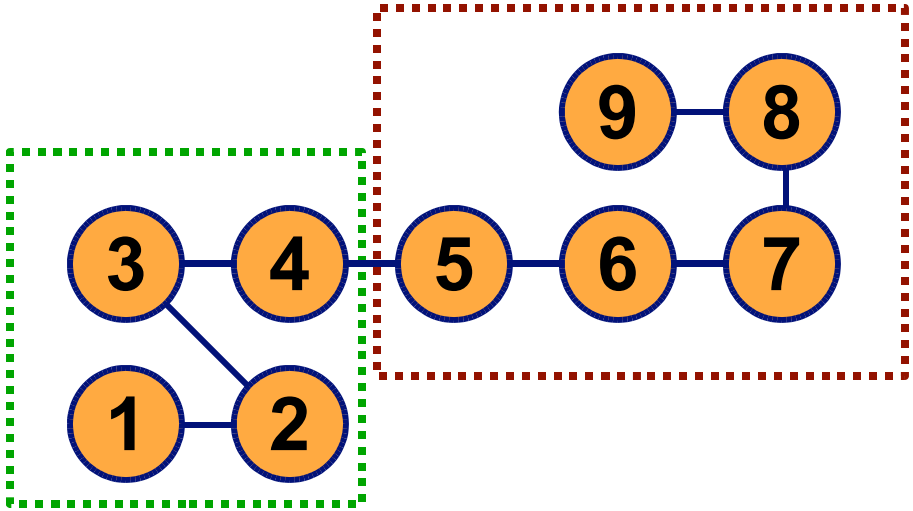




#	1	2	3	4	5	6	7	8	9
1	1	1	1	1	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0
3	1	1	2	2	1	1	1	1	1
4	1	1	2	2	1	1	1	1	1
5	0	0	1	1	2	2	2	2	2
6	0	0	1	1	2	3	3	3	3
7	0	0	1	1	2	3	3	3	3
8	0	0	1	1	2	3	3	3	3
9	0	0	1	1	2	3	3	3	3







#	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/)

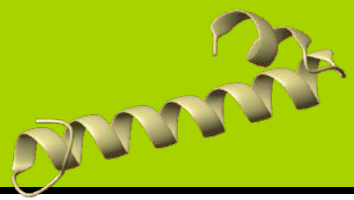
Thresholds #1,2 → MAMMOTH P-Value (Lp, Up)

High P-values → fewer partitions

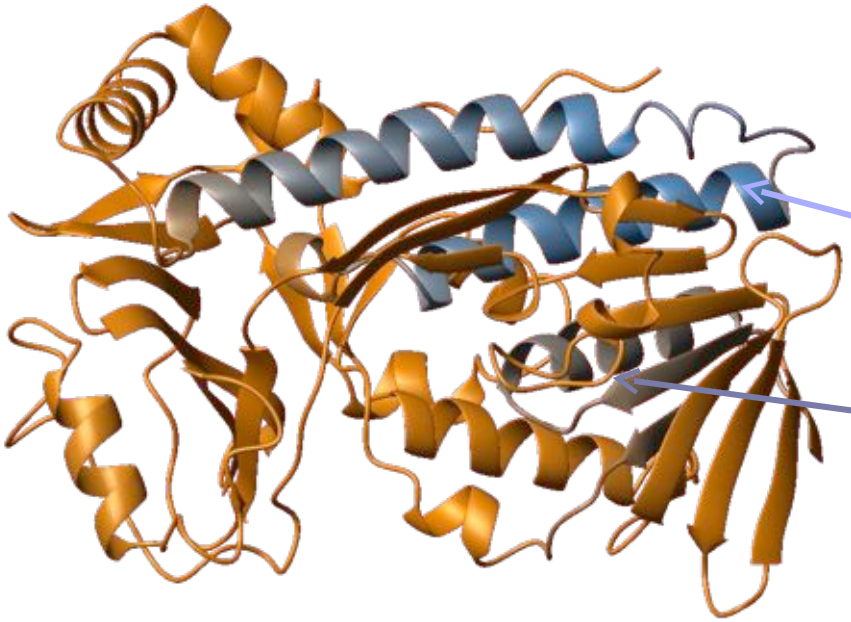
Threshold #3 → Cluster Level (-l)

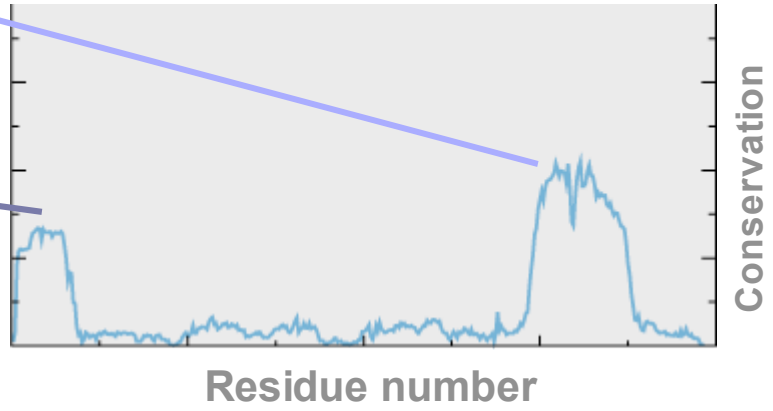
Low –l cluster value → fewer partitions

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

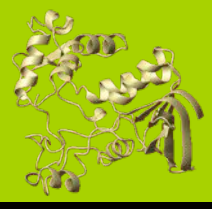
1phh	290-329	2.7Å	3.1
1hadB	72-111		

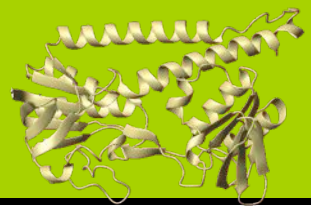
1phh	279-373	3.9Å	4.7
1bke	310-410		

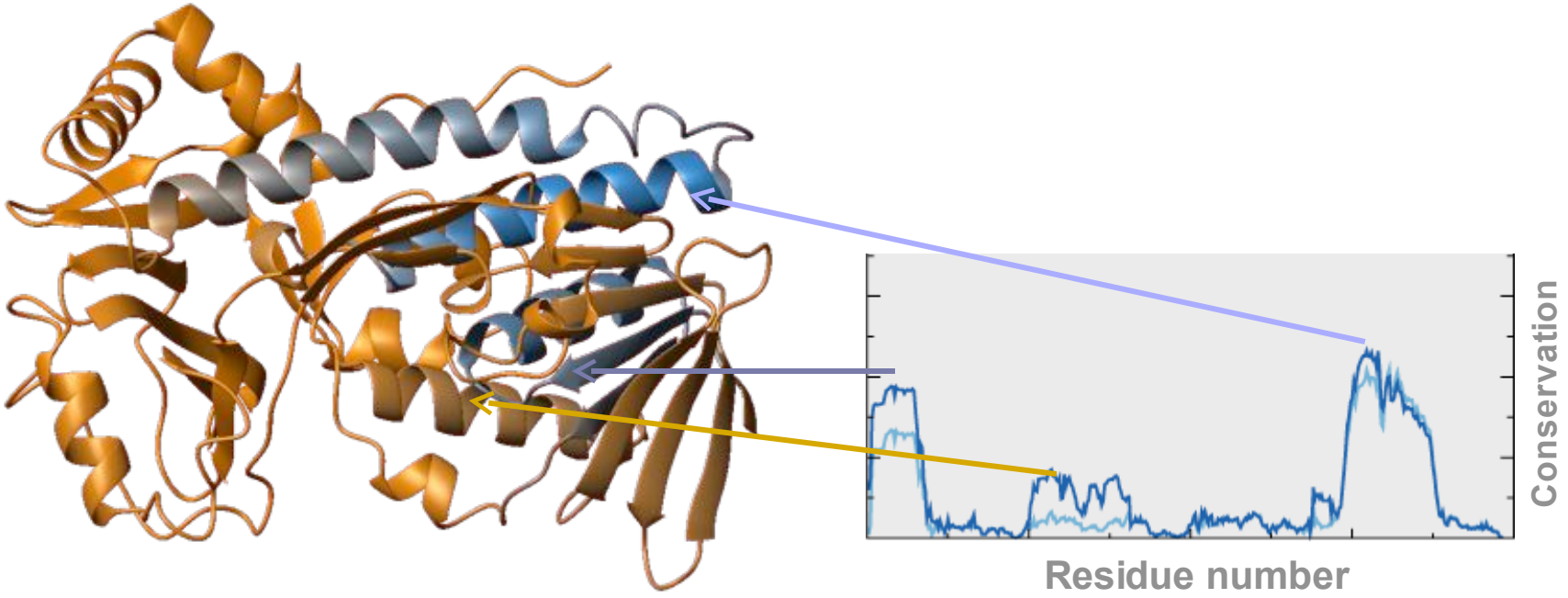




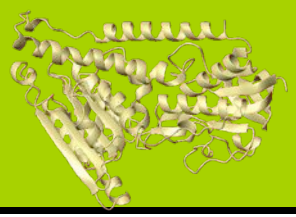
1phh (Oxydoreductase from *Pseudomonas fluorescens*)

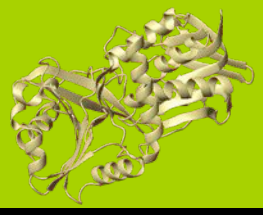
1phh	1-213	3.0Å	8.1
1qjdA	125-379		

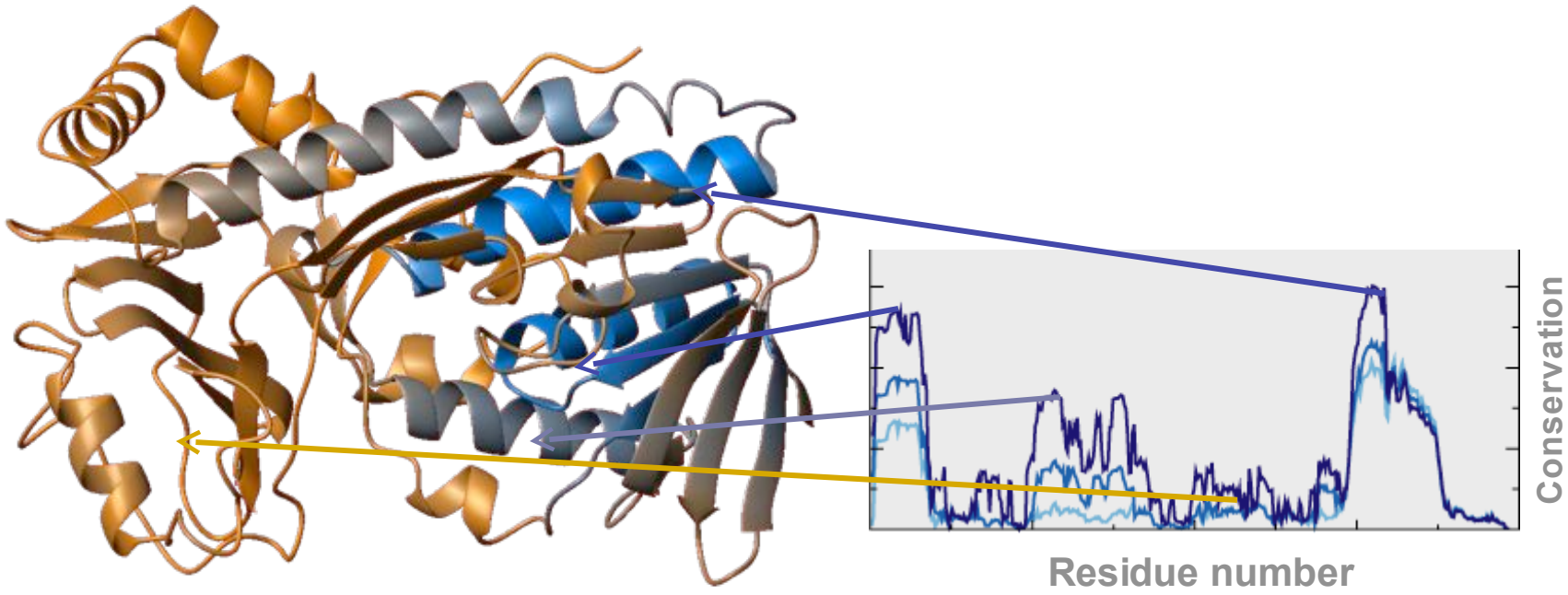
1phh	1-319	3.6Å	9.8
1gerA	3-327		



1phh (Oxydoreductase from *Pseudomonas fluorescens*)

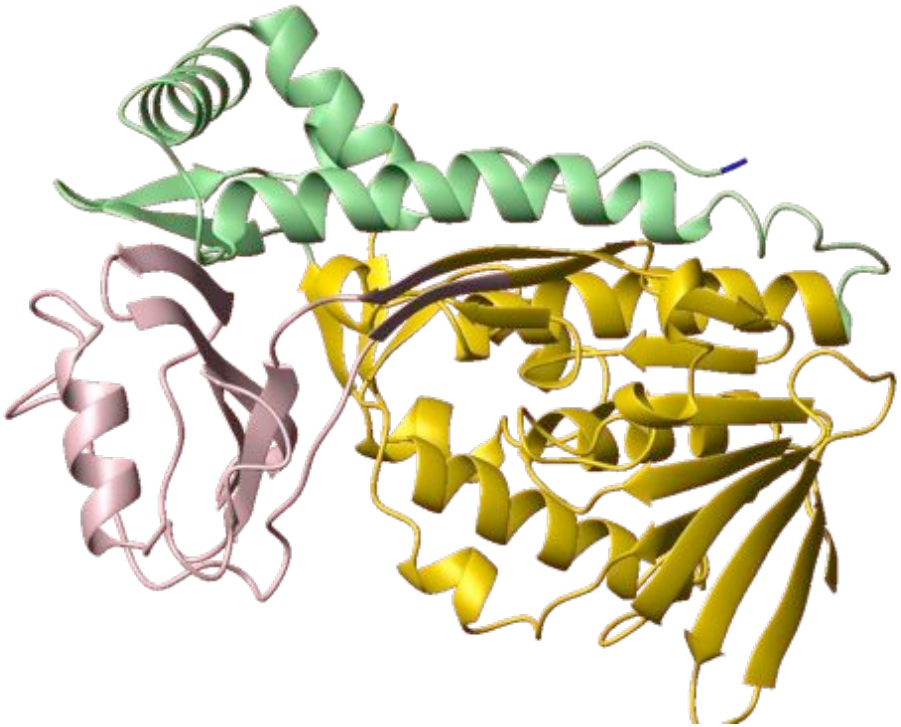
1phh	1-378	3.8Å	10.3
1feaC	2-464		

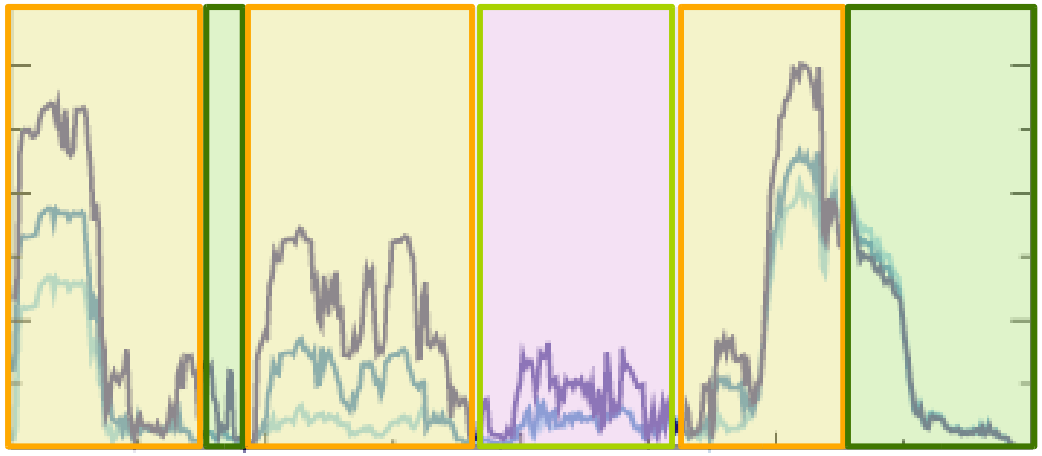
1phh	1-316	3.8Å	17.2
1l9dB	2-364		



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

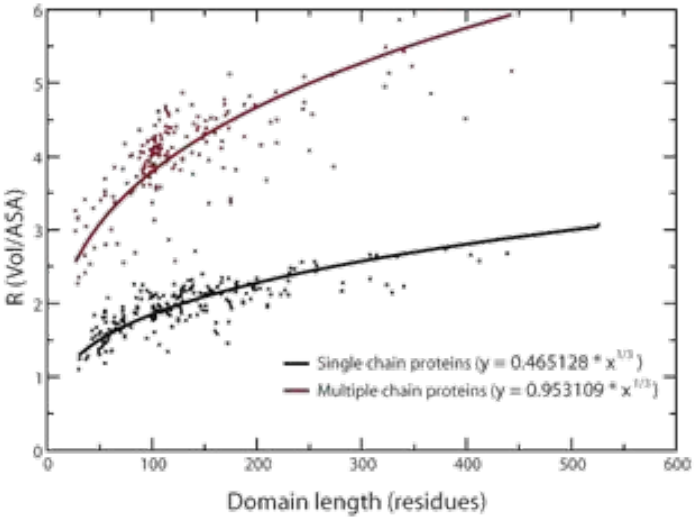
Thresholds #1,2 → MAMMOTH P-Value (Lp, Up)

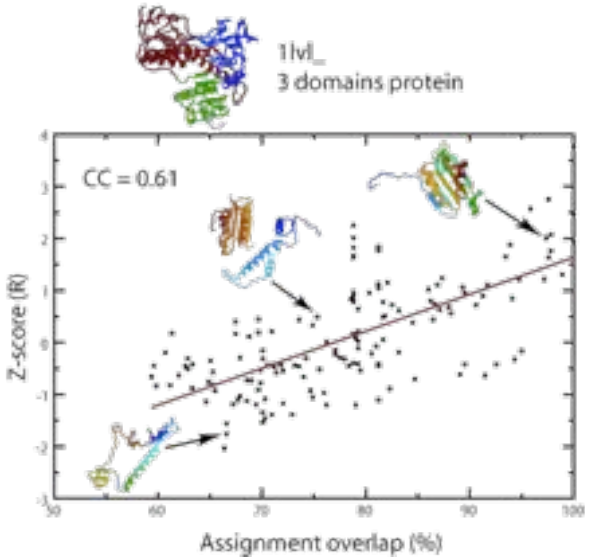
High P-values → fewer partitions

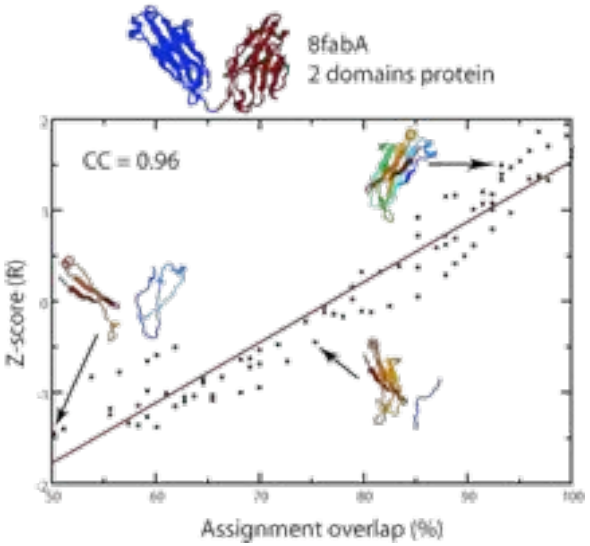
Threshold #3 → Cluster Level (-l)

Low –l cluster value → fewer partitions

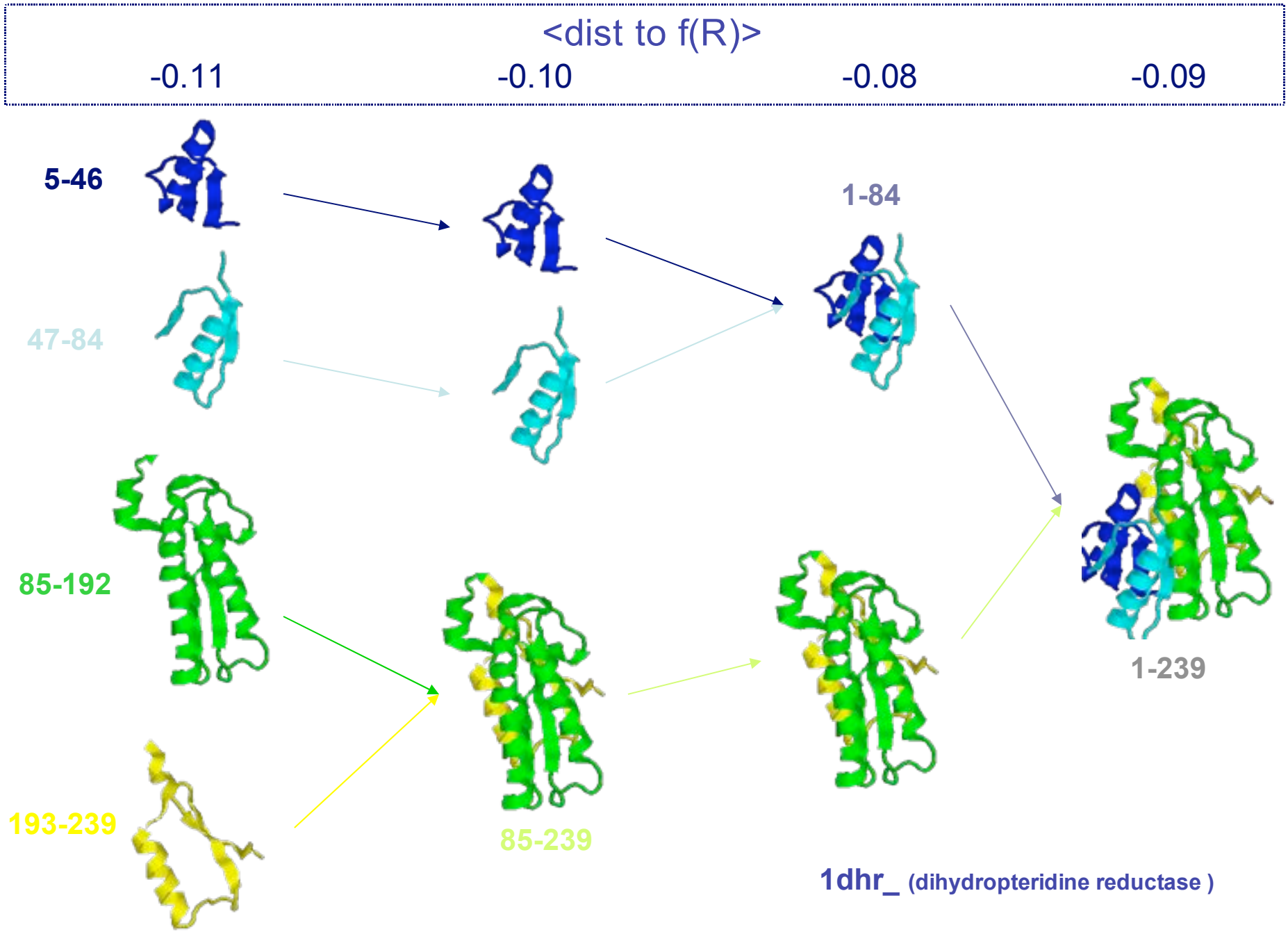
R = Volume/ASA



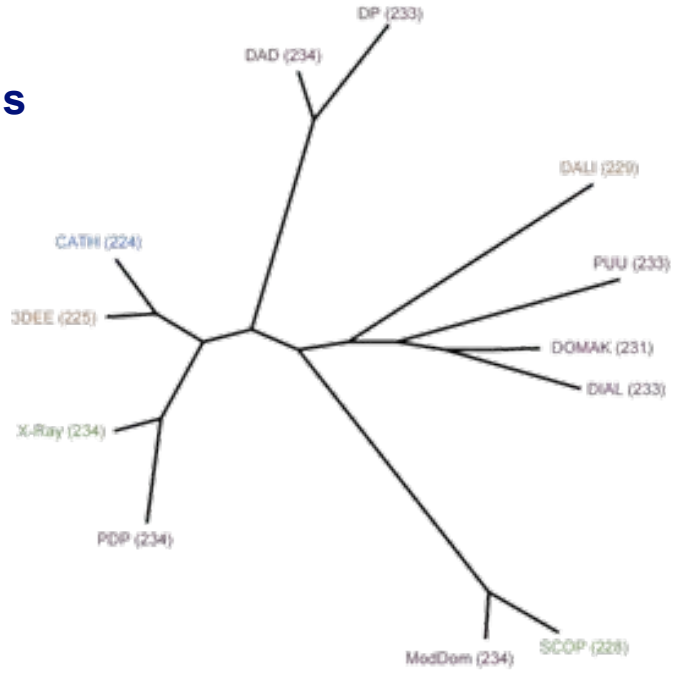


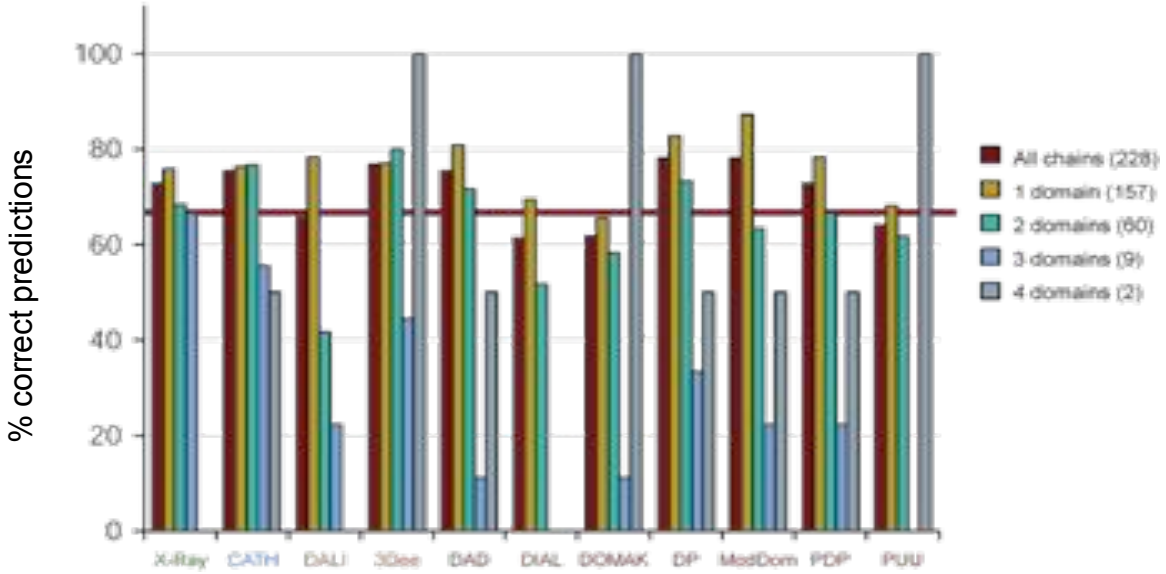


Domain → max(<dist f(R)>)


<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

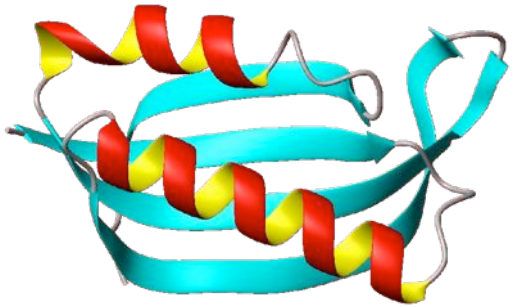


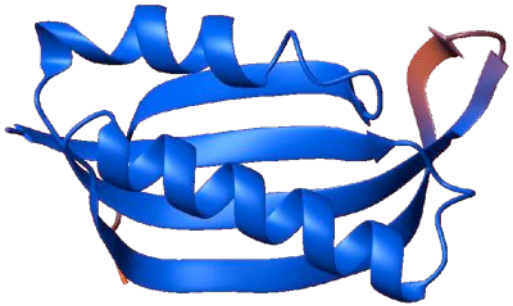


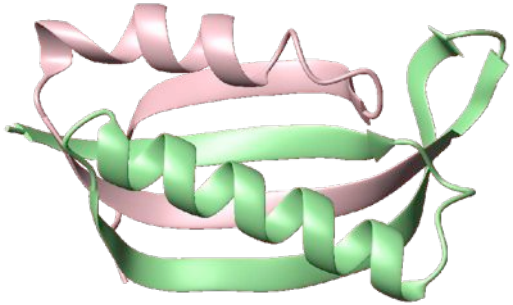
Fragments assignment from structure

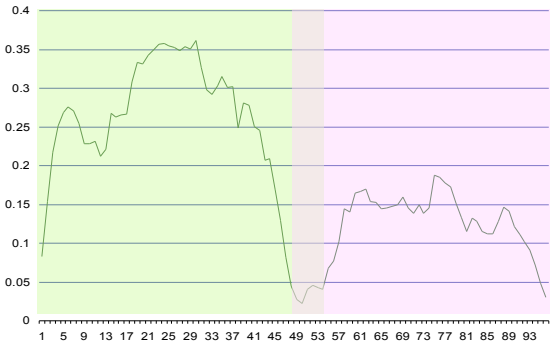
Repetitions
Swapping
Complementarities

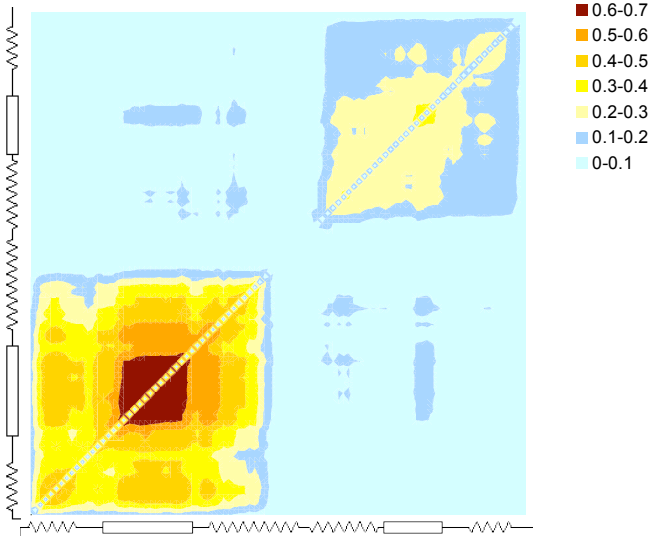
Ribosomal protein S6 (1ris) $\alpha+\beta$ → *Ferrodoxin Like domain*

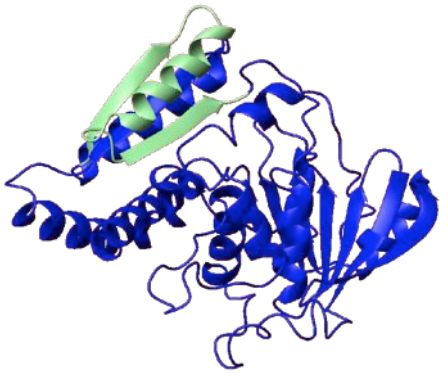








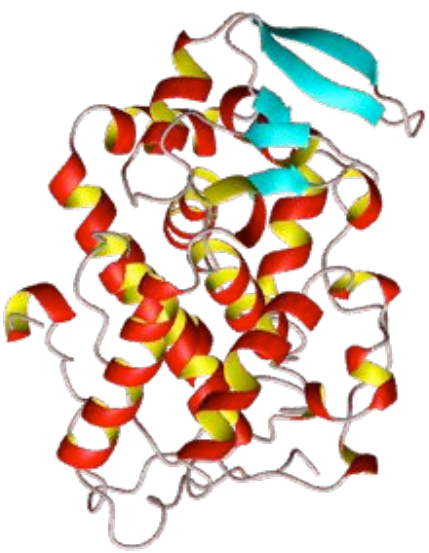


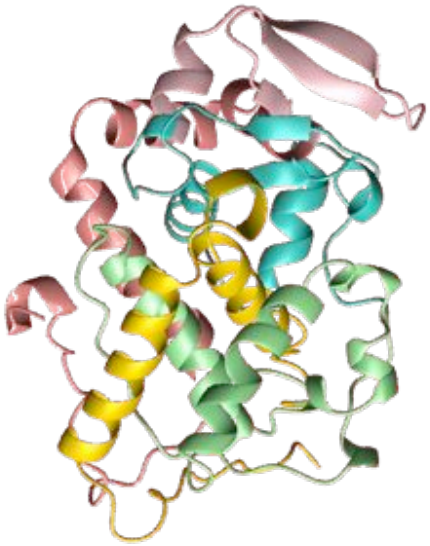


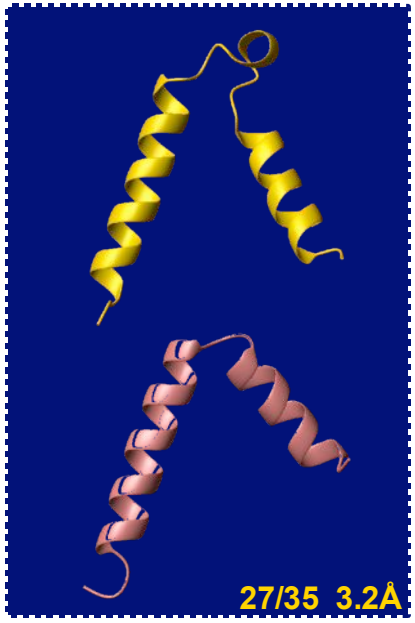
1ee9A 17.9% id. 2.3Å

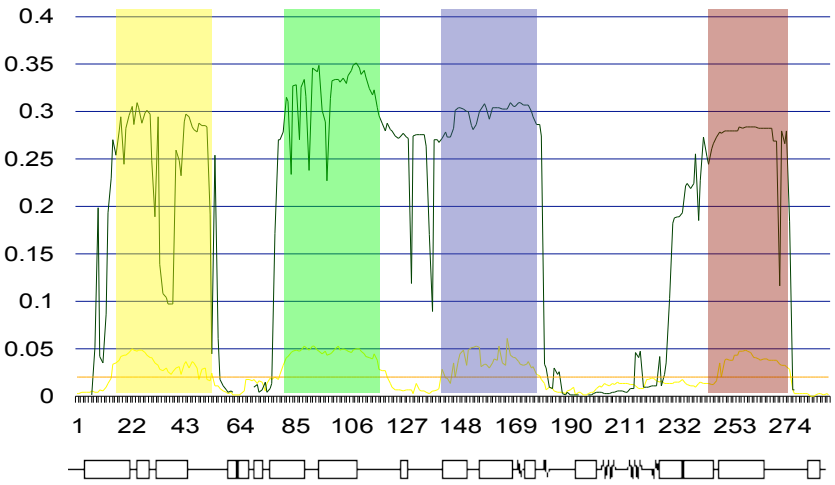
6timB 11.1% id. 2.6Å

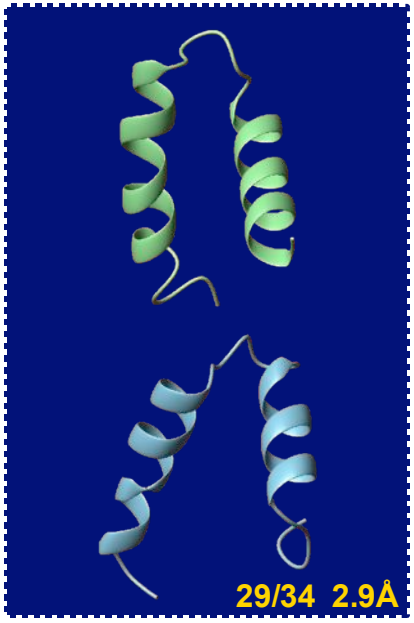
Cytochrome C Peroxidase (2cyp) *all-α → CCP-like domain*



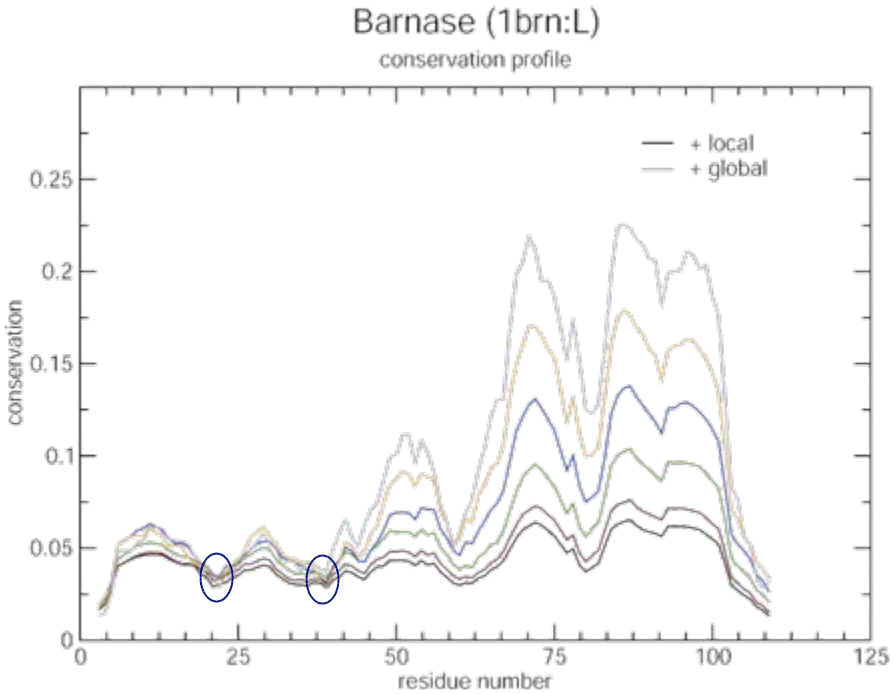






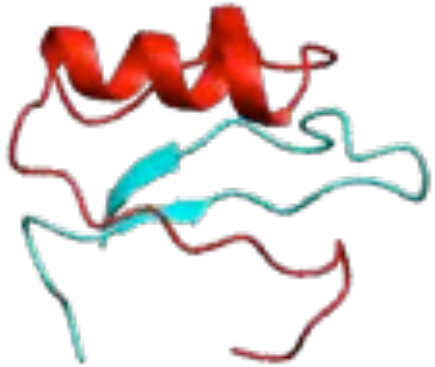


Barnase Domain-Swapping



chymotrypsin inhibitor 2

1-37 | 38-64



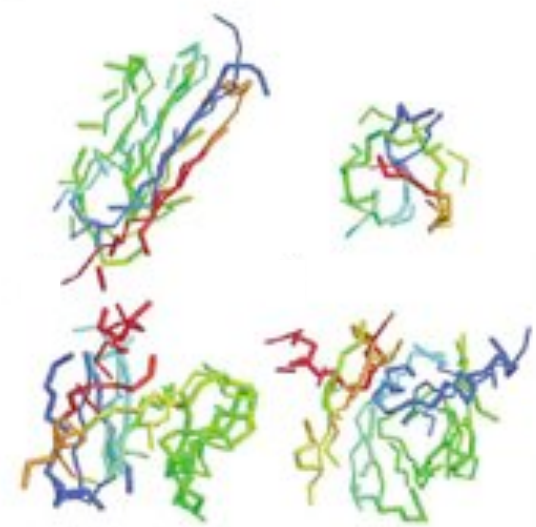
1-40 | 41-64

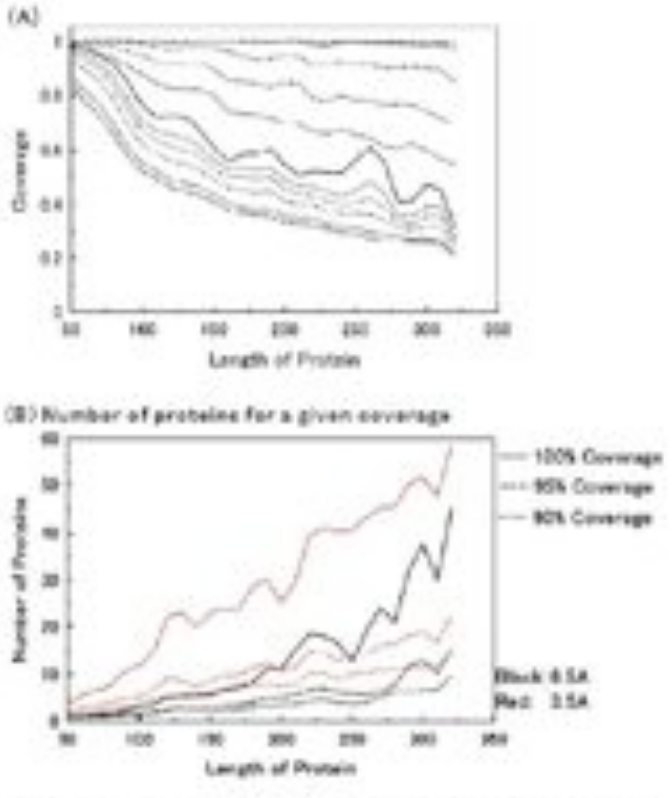
- Neira JL, Davis B, Ladurner AG, Buckle AM, Gay GP, Fersht AR. 1996. Towards the complete structural characterization of a protein folding pathway: the structures of the denatured, transition and native states for the association/folding of two complementary fragments of cleaved chymotrypsin inhibitor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

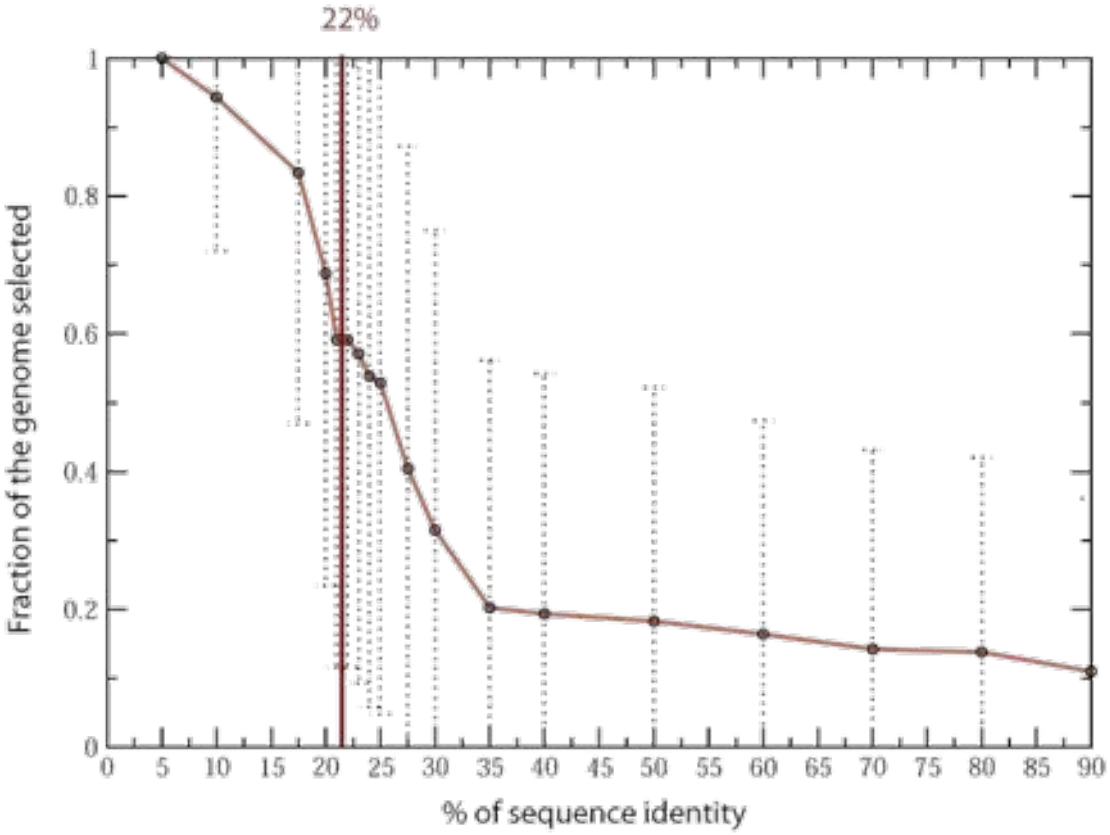
The PDB is a covering set of small protein structures.





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

BMI-206

Structure-Structure comparisons
Sequence-Structure comparisons

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

February 19th, 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 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.

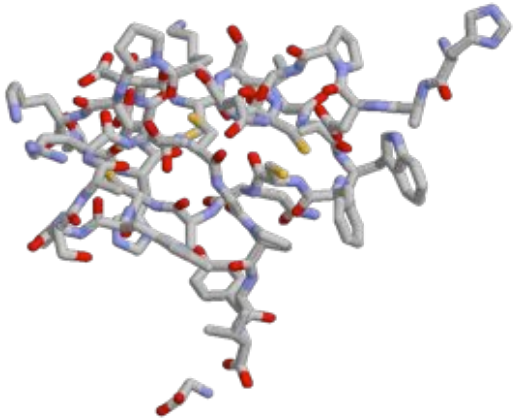
Finkelstein et al. (1995) *Proteins* **23**, pp142

Representation

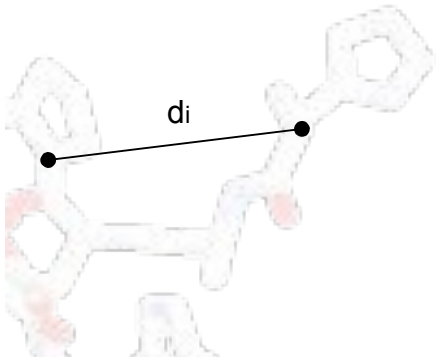
Sequence/Structures

```
>gi42541361
MDIRSVSSLRGLLCLPPSWPRR
```

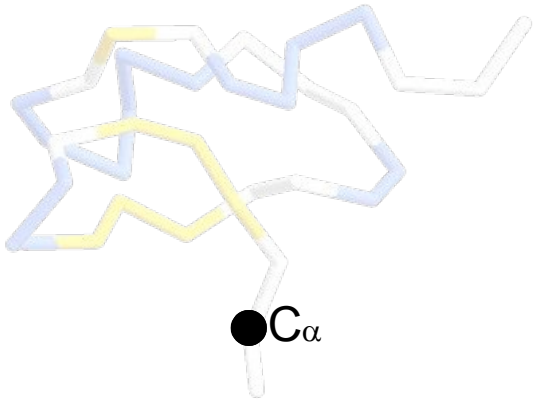
Primary sequence



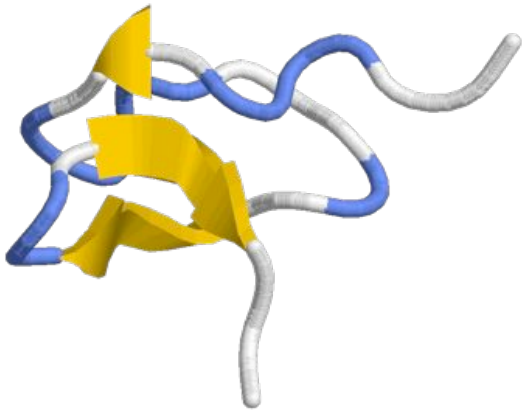
All atoms and coordinates



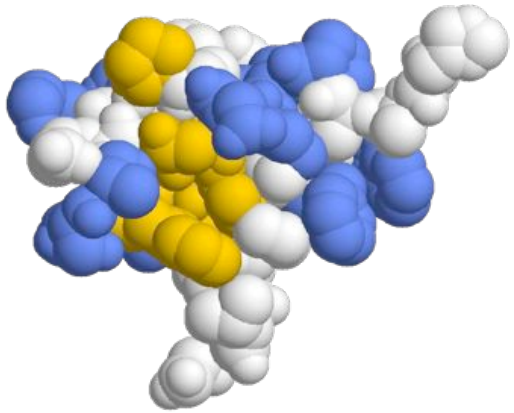
Distance space



Reduced atoms representation



Secondary Structure

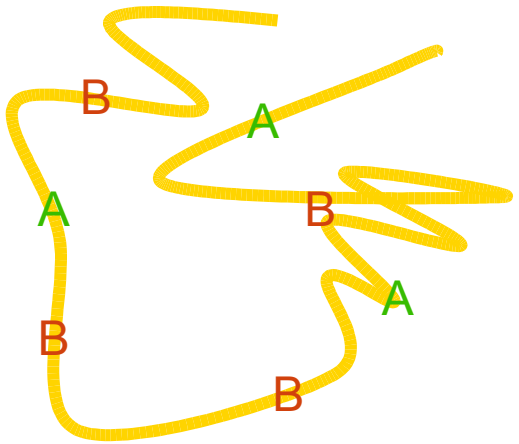


Accessible surface

Statistical Potential… inspiration

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

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



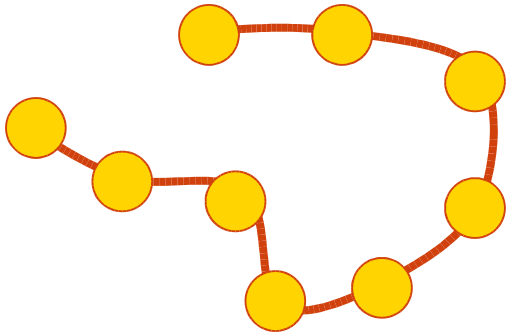
From statistical physics, we know that energy difference between two states (ΔE) and the ratio of their occupancies (N_1:N_2) are related [9]:

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

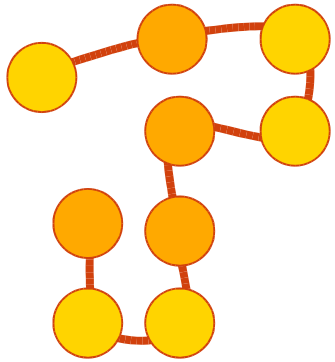
in which T is the absolute temperature and k is the Boltzmann's constant. As we are interested in an interaction energy between two amino acid side chains, it would seem natural to define N_1 as the number of interactions between these two residues types in a group of real protein structures, a number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

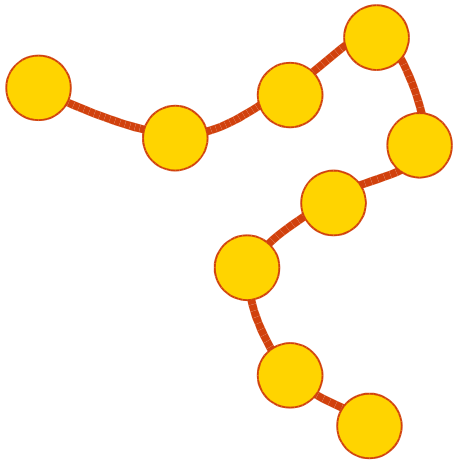
Statistical Potential… interaction types



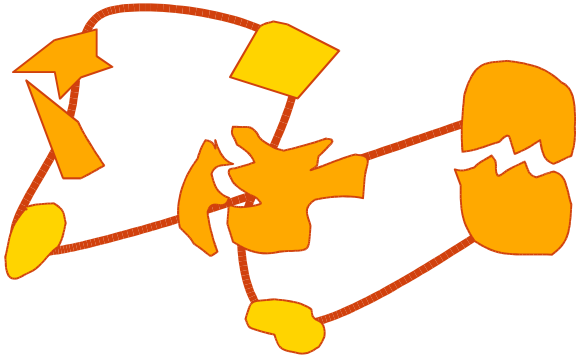
Neutral interactions



Hydrophobic interactions



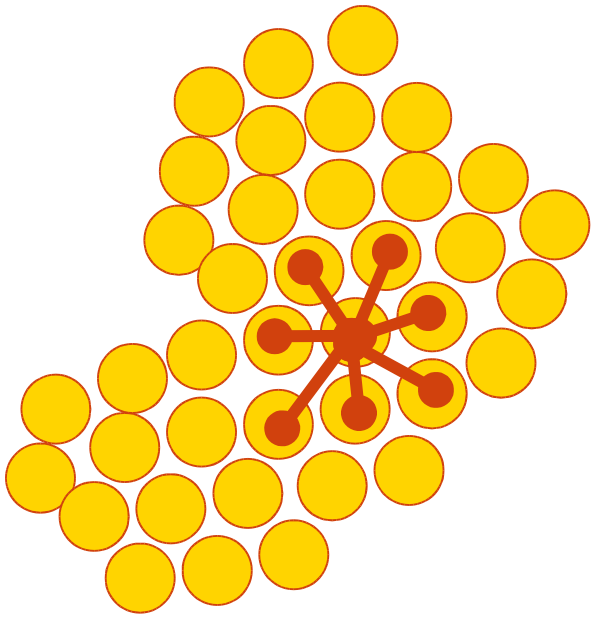
Compact interactions

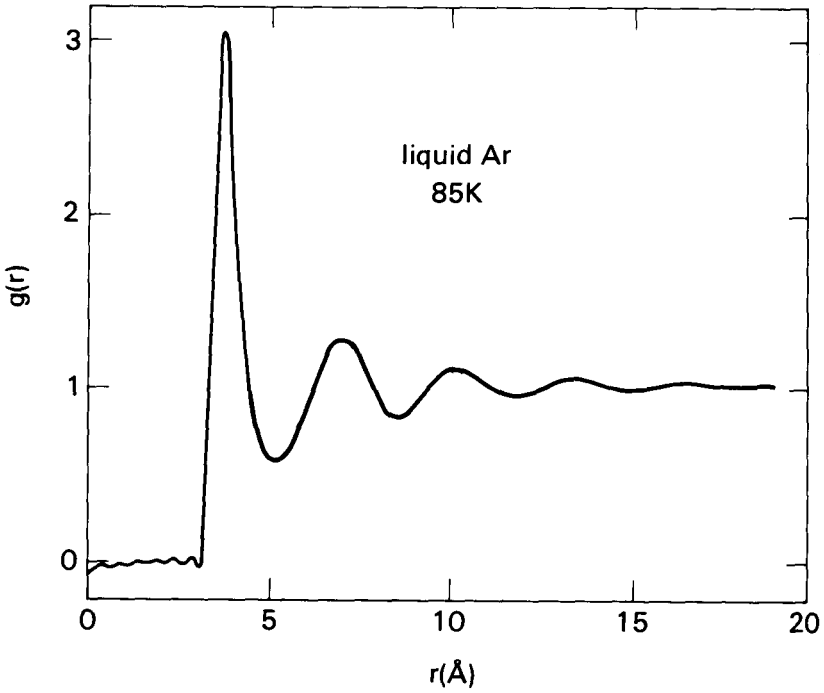


Specific interactions

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

Statistical Potential… reference state

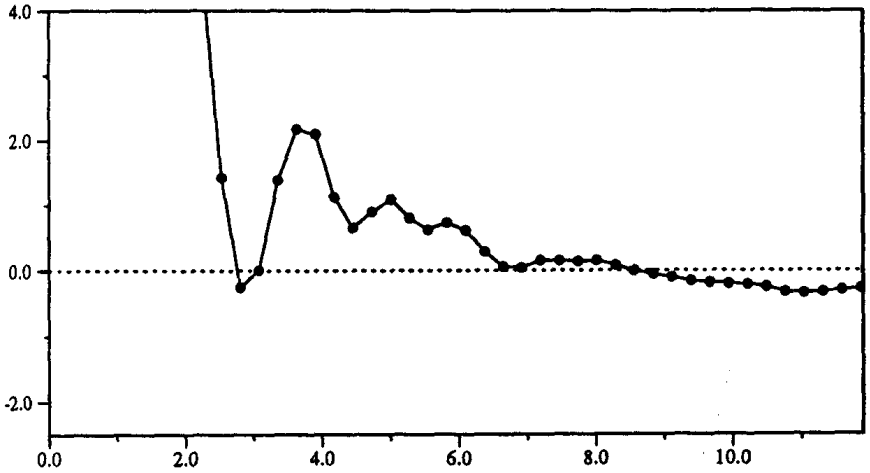




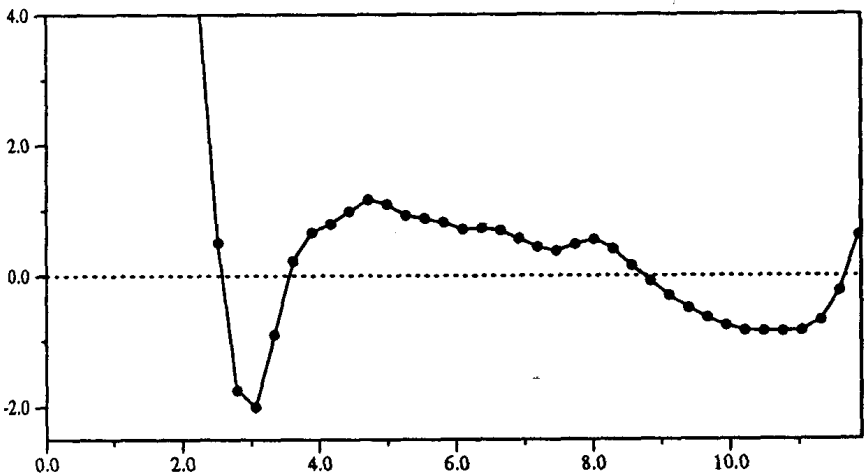
Theory of simple liquids 2nd edition JP Hansen and IR McDonald, Academic Press.

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

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

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

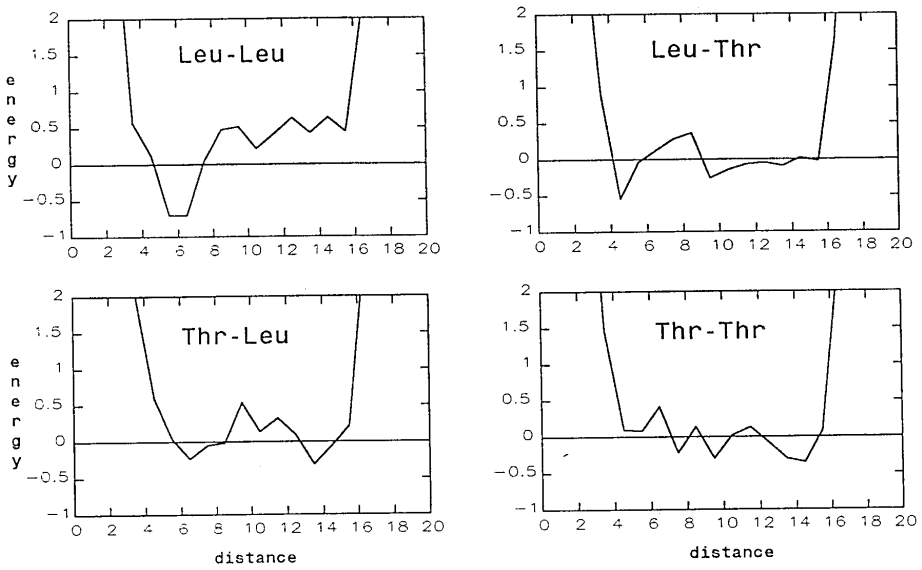
$$w_{NO}(r) = -kT \ln\left(g_{NO}(r)\right)$$

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

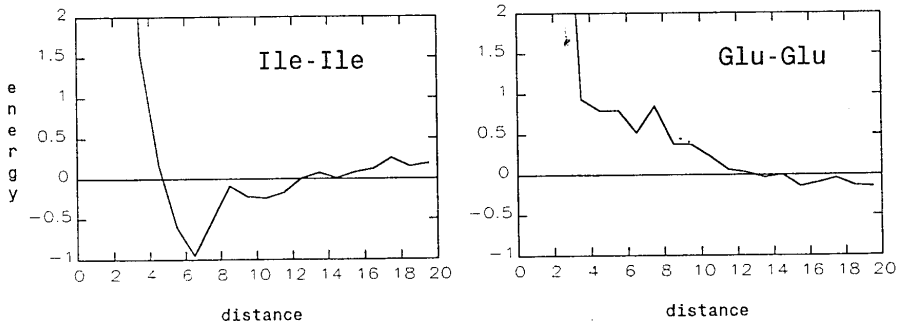
Scoring

Statistical Potential… Distance Potentials

Long range free energy



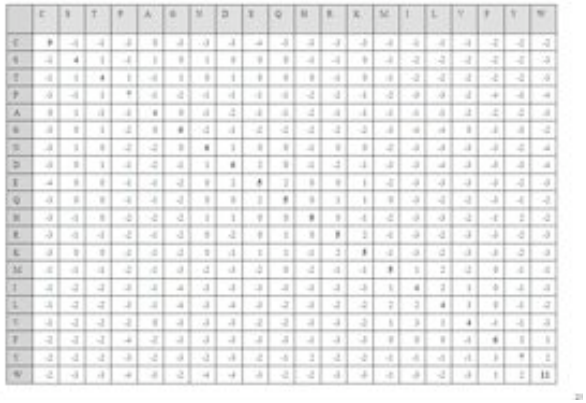
Short range free energy



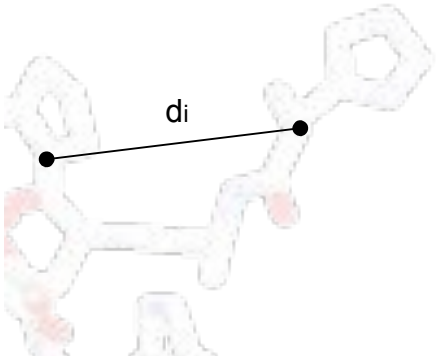
Sippl (1993). *JCAM* **7** pp473

Scoring

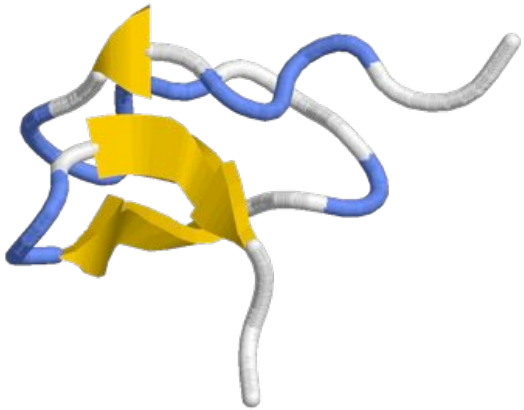
Raw scores of an alignment



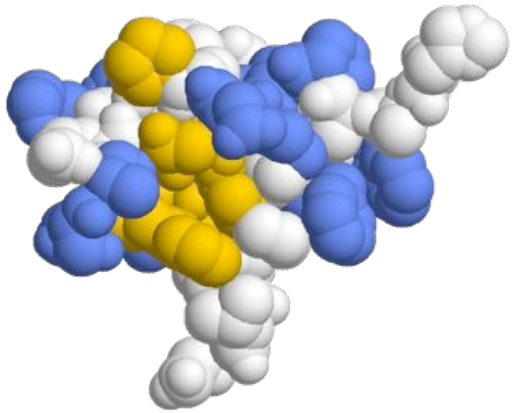
Aminoacid substitutions



Distance space



Secondary Structure (H,B,C)

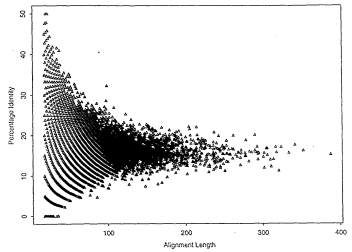


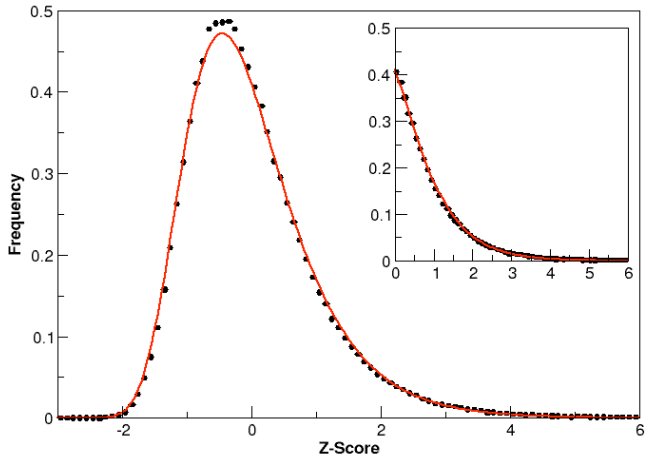
Accessible surface (B,A [%])

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

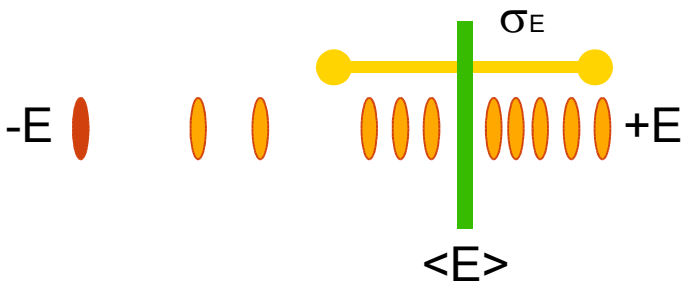
$$P(s) = e^{-\lambda (s-\mu)}$$

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

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

Significance of an alignment (score)

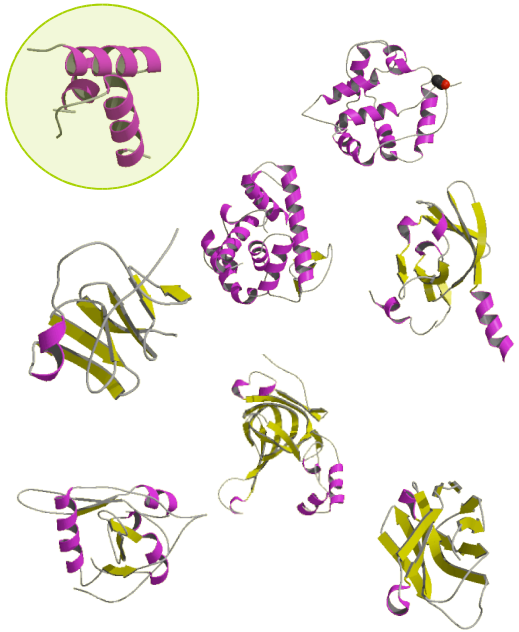
Energy Z-score the model with respect the energy of random models (or rest of decoys).



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

Significance of an alignment (background)

Structural space



Sequence space

MKLLIVLTCISLCSCICTVVQRCASNKPHVLEDPCKVQH

HLSVNQCVLLPQCCPKSCKICTHLISIEVVLTCRAVDKM

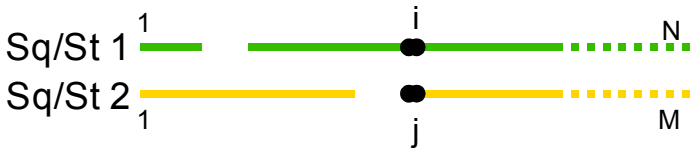
MHVNCVEQCSLQDCIKIAPRVLKTCILCVLKPCLTSVSH

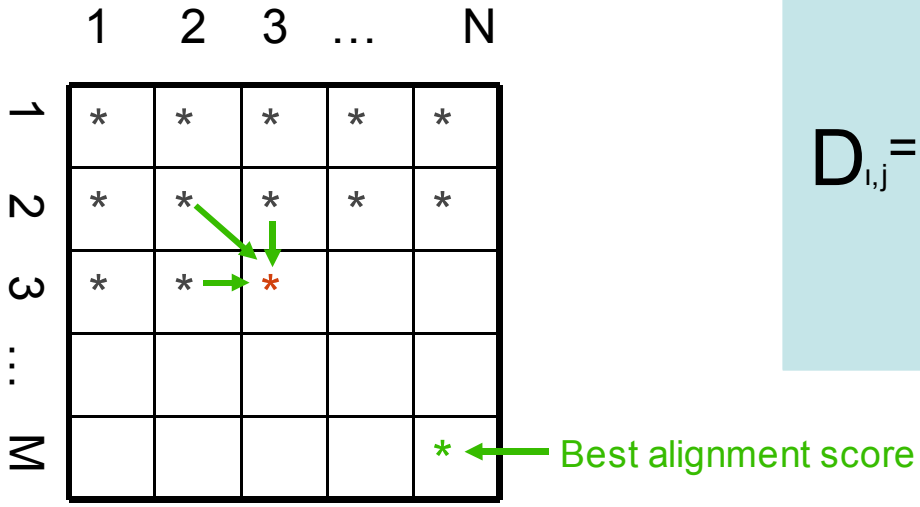
VHLVQPTSCCCKKNCICHVEIRSLDILTKSVQLACLVPM

⋮

MQCCRVQKICDLLAVELCKLHISTPSCKILCVVTSVPHN

Global dynamic programming alignment



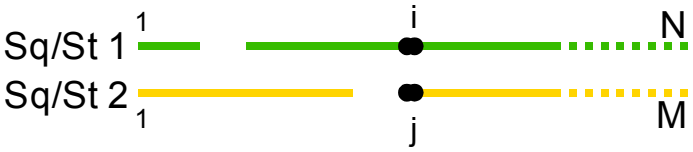


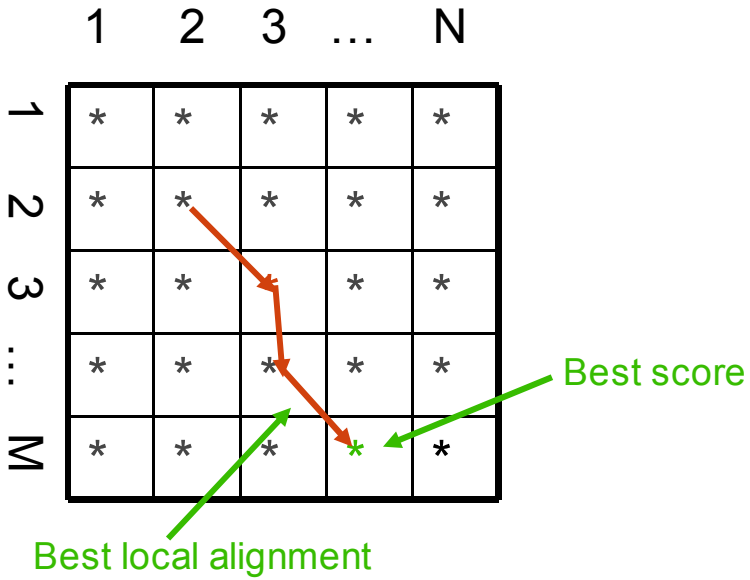
$$D_{i,j} = \min \begin{cases} D_{i,j-1} + Score_{(\Delta, rj)} \\ D_{i-1,j-1} + Score_{(ri, rj)} \\ D_{i-1,j} + Score_{(ri, \Delta)} \end{cases}$$

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

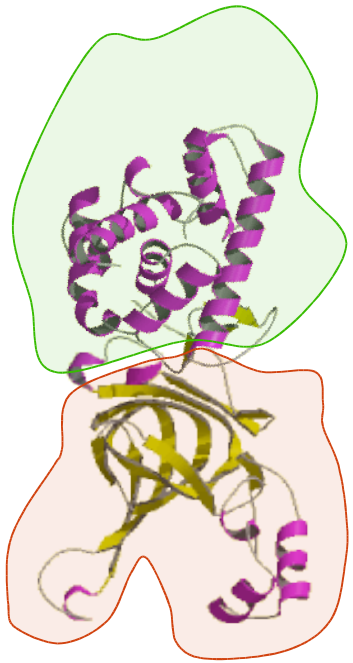
Applications of PMFs

- 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

Domain boundaries from sequence

VERY DIFFICULT!!!!



MENFEIWVEKYRPRTLDEVVGQDEVIQRLKGYVERKNIPHLLFSGPPGTGKTATAIALARDLFGENWRDN
FIEMNASDERGIDVVRHKIKEFARTAPIGGAPFKIIFLDEADALTADAQAALRRTMEMYSKSCRFILSCN
YVSRIIEPIQSRCAVFRFKPVPKEAMKKRLLEICEKEGVKITEDGLEALIYISGGDFRKAINALQGAAAI
GEVVDADTIYQITATARPEEMTELIQTALKGNFMEARELLDRLMVEYGMSGEDIVAQLFREIISMPIKDS
LKVQLIDKLGEVDFRLTEGANERIQLDAYLAYLSTLAKK

Domain boundaries from sequence (SnapDragon)

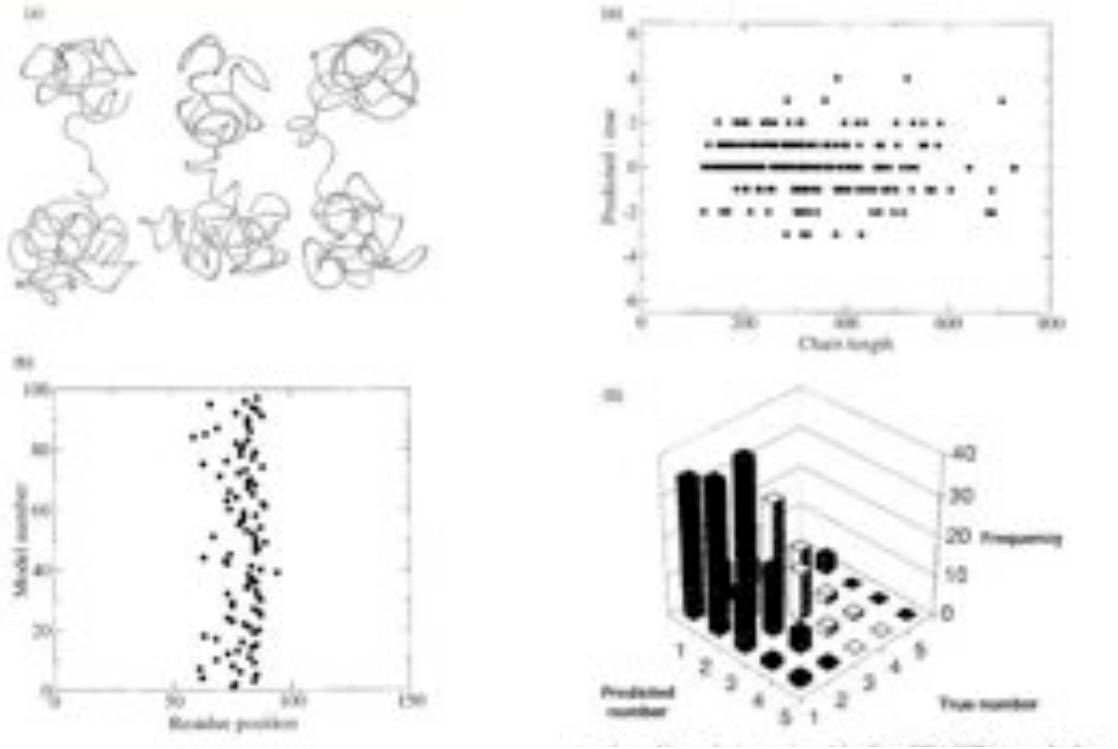
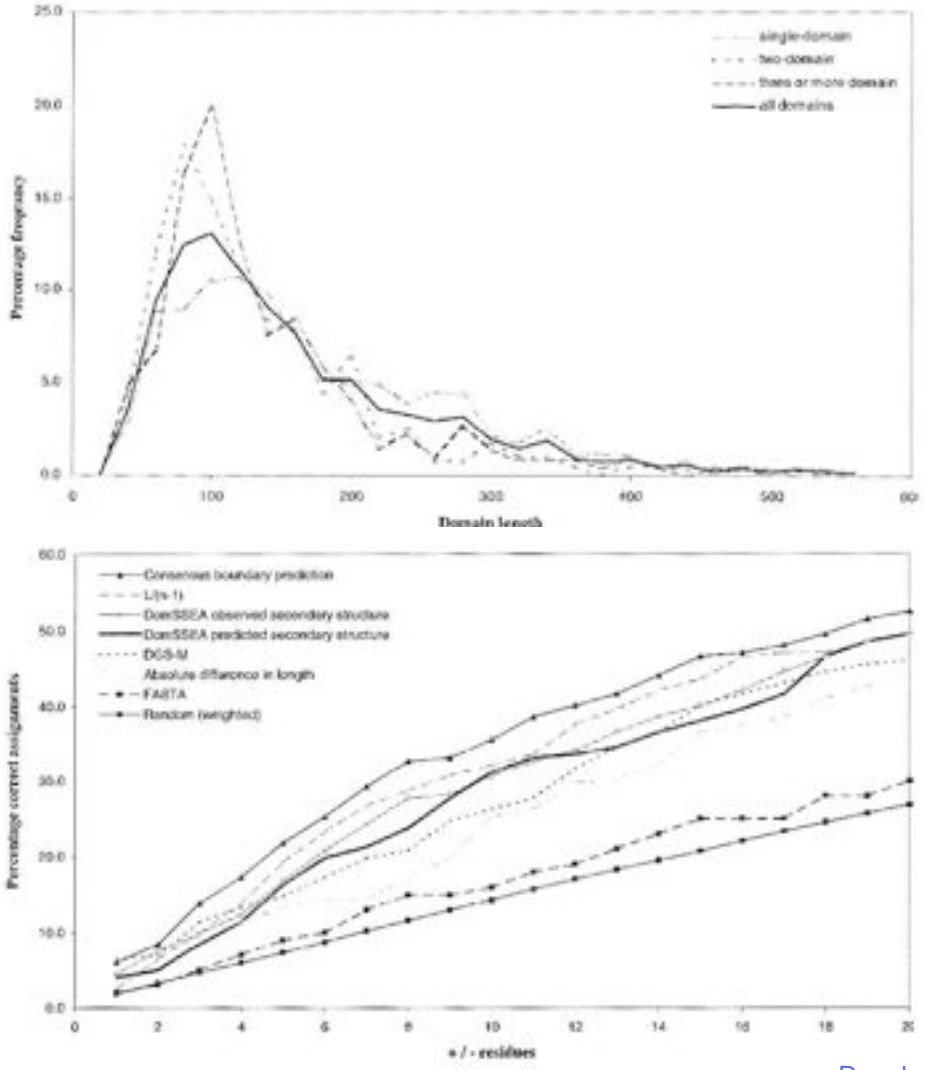


Table 2. Average accuracy percentages of linker prediction over 57 proteins

		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) *Proty. Science* **11** pp2014

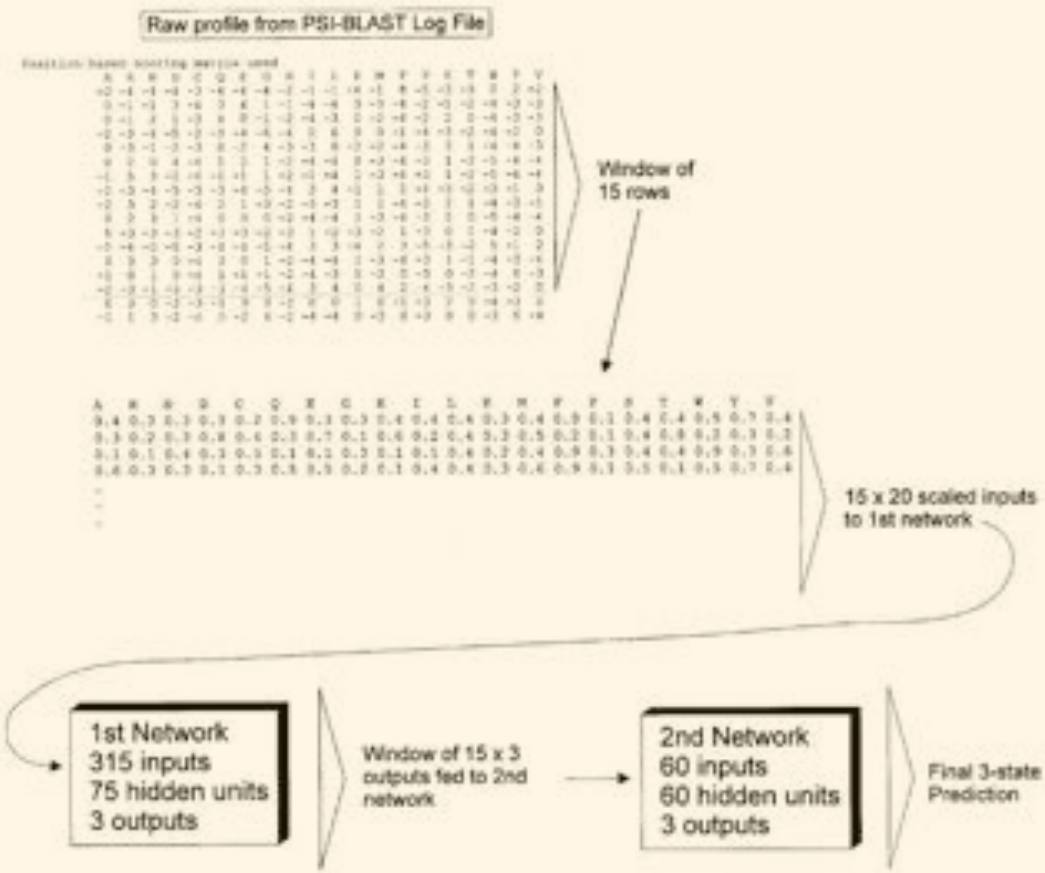
Prediction of Secondary Structure (PSI-PRED)

```
>gi42541361
MDIRSVSSLRGLLCLPPSWPRR
```

- Neural Network

✓ **Very simple idea**
✓ **Simple scoring**

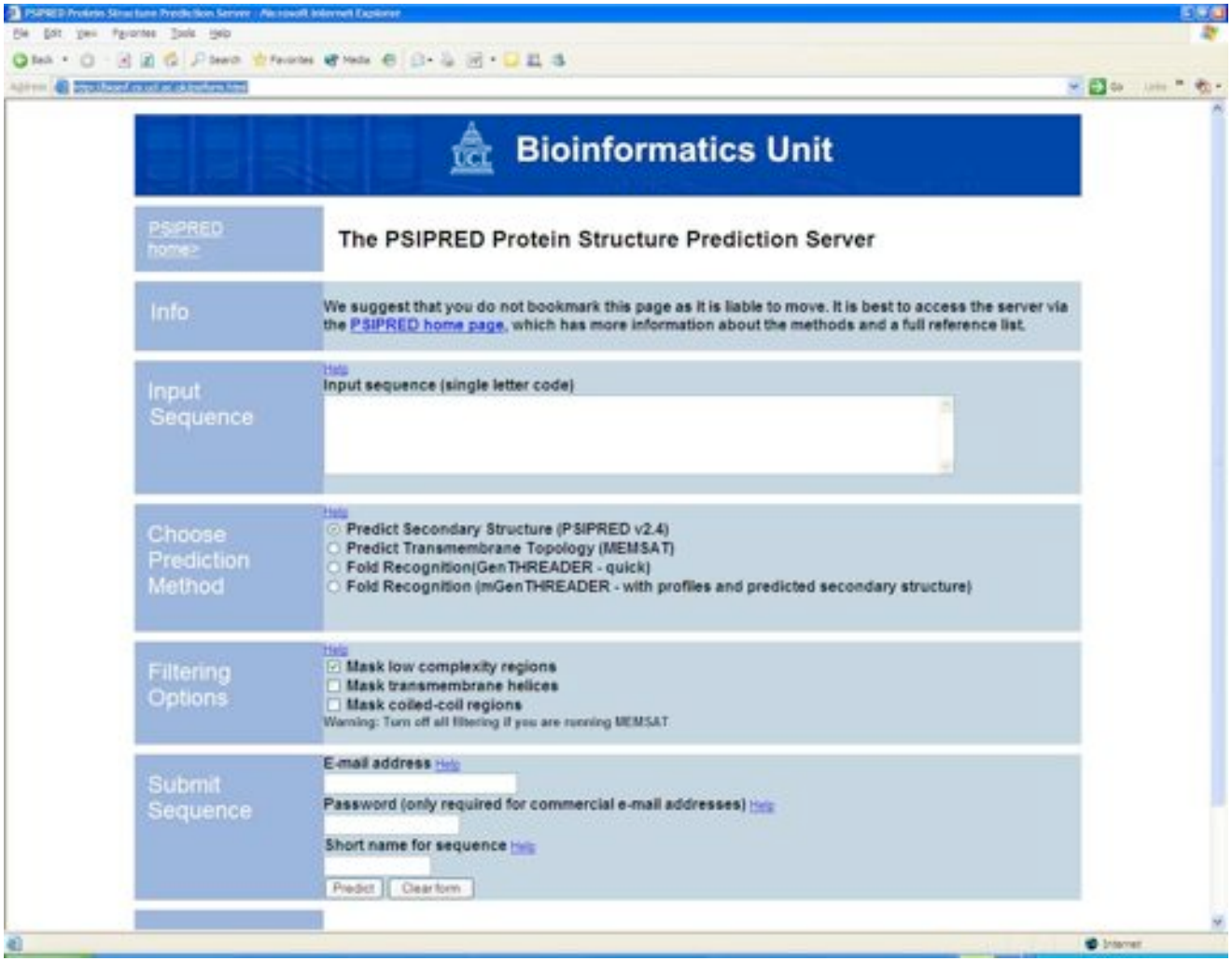
✘ **Obscure optimizer**



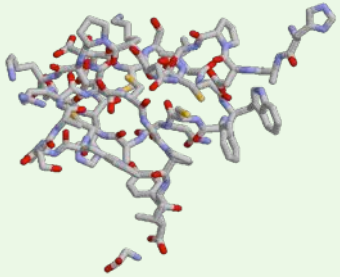
Jones DT. (1999) *J. Mol. Biol.* **292** pp195

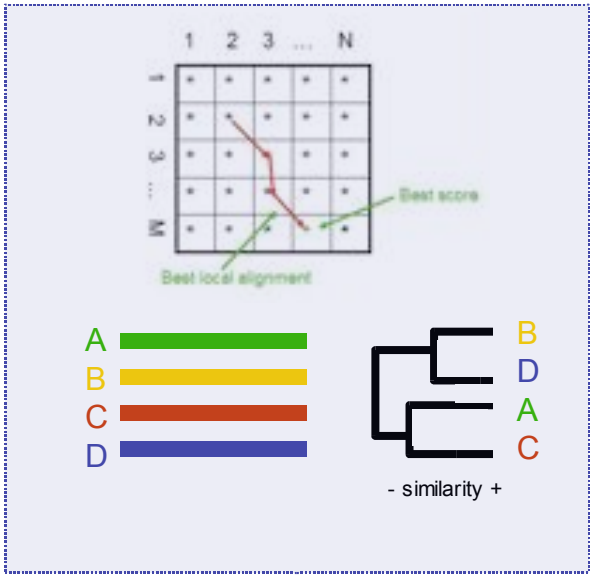
Prediction of Secondary Structure (PSI-PRED)

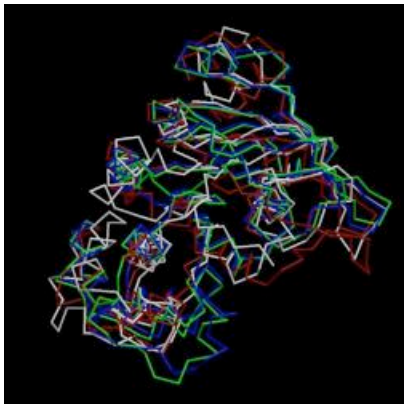
http://bioinf.cs.ucl.ac.uk/psiform.html



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

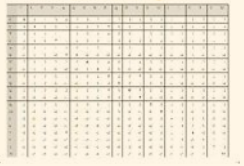




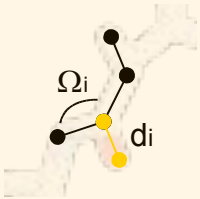


✓ Uses all available structural information
✓ Provides the optimal alignment

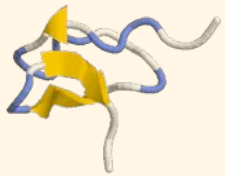
✗ Computationally expensive



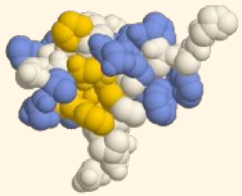
$R_{i,j}$



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



$S_{i,j}$



$B_{i,j}$

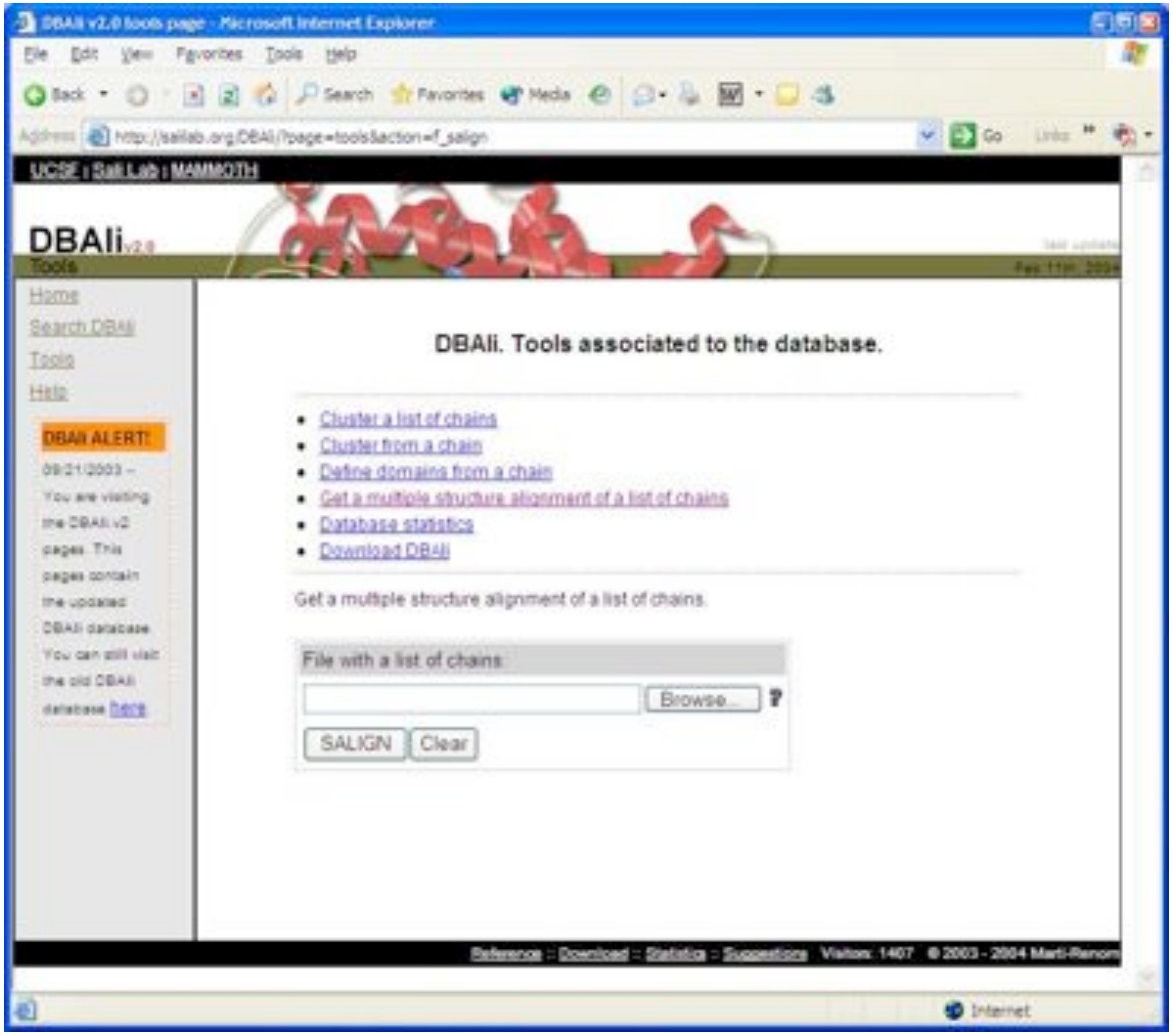
$$RMSD = \sqrt{\sum\left(x_i - \bar{x}\right)^2}$$

$I_{i,j}$

Madhusudhan et al. *in preparation*

Structural alignment by properties conservation (SALIGN-MODELLER)

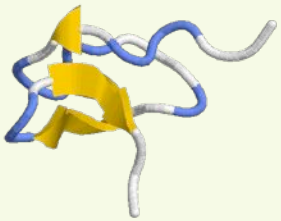
http://www.salilab.org/dbali/

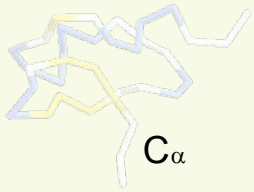


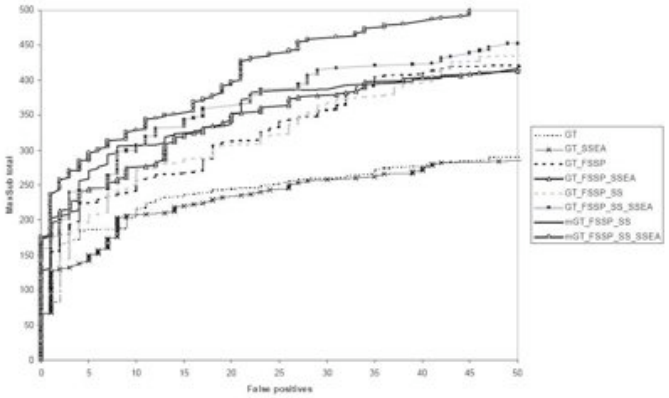
Threading (mGenThreader)

>gi42541361
MDIRSVSSLRGLLCLPPSWPRR

- Neural Network

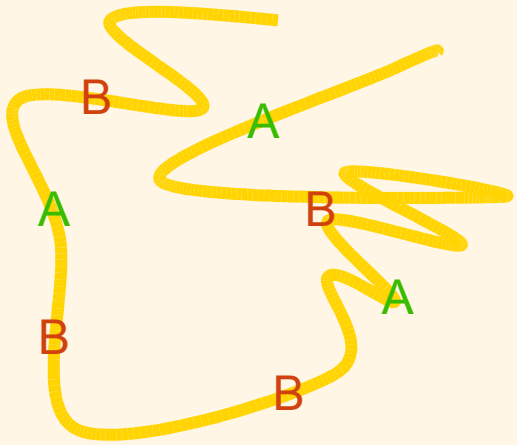


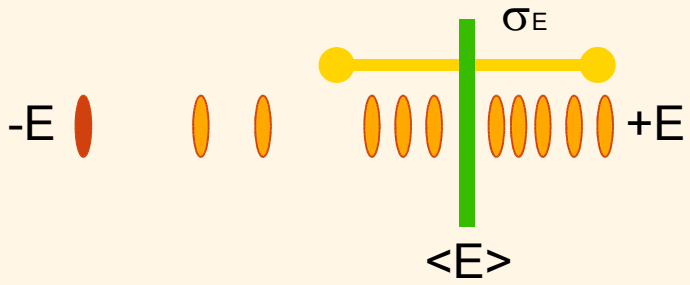




✓ Good row and significance scoring

✘ Obscure optimizer



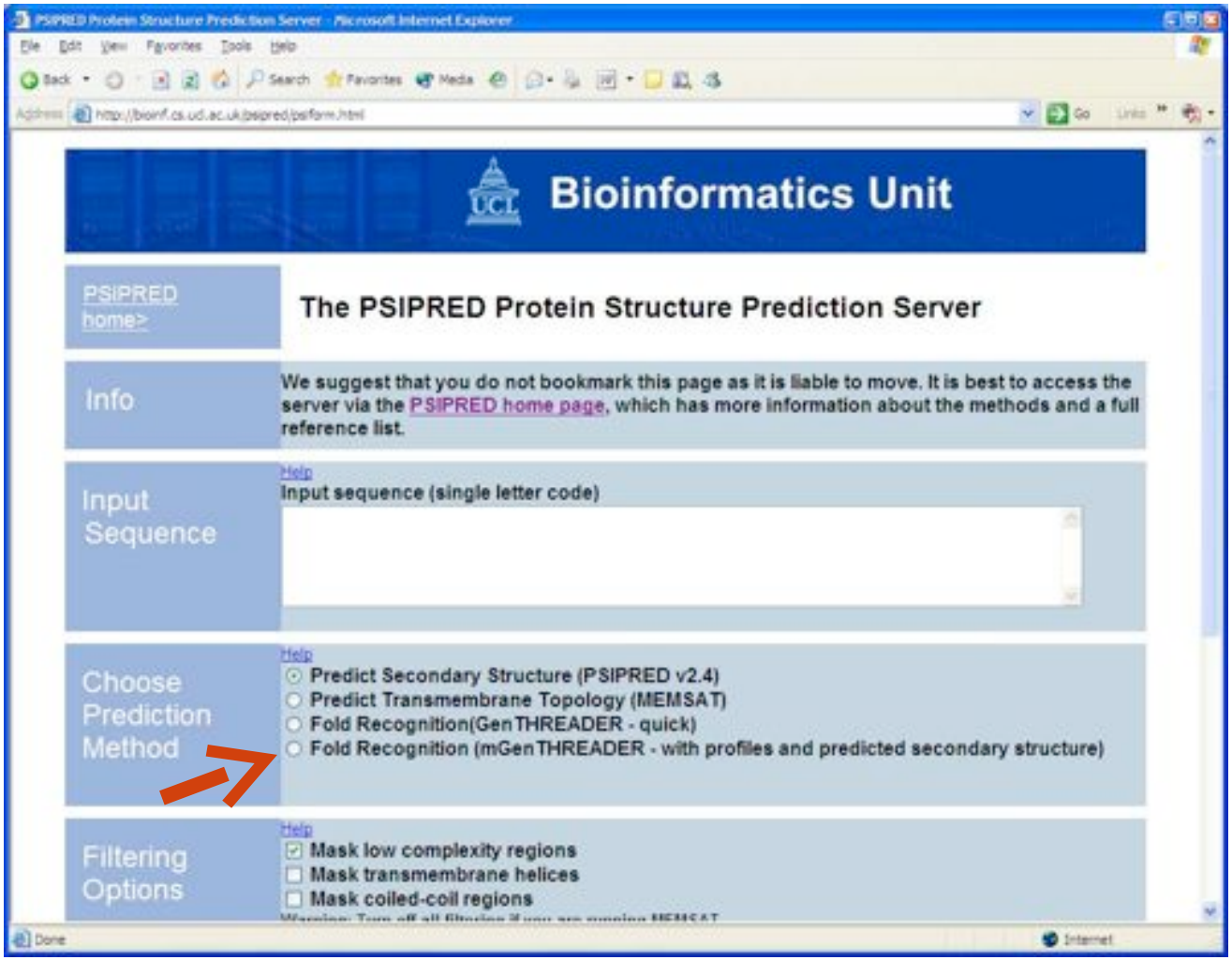


$$Zscore = \frac{\left(\langle E \rangle - E_m\right)}{\sigma_E}$$

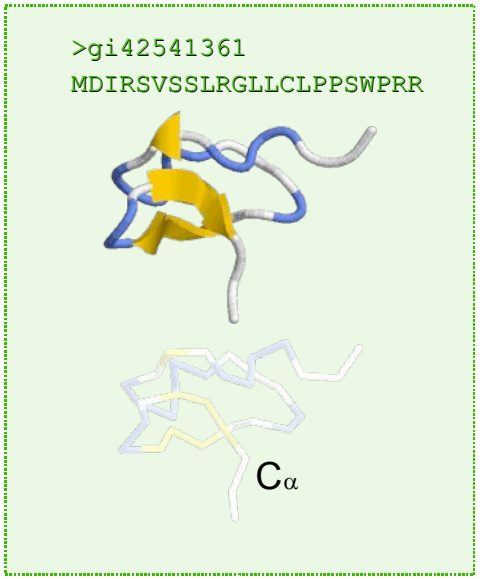
McGuffin LJ, Jones DT. (2003) *Bioinformatics*, **19**, pp874

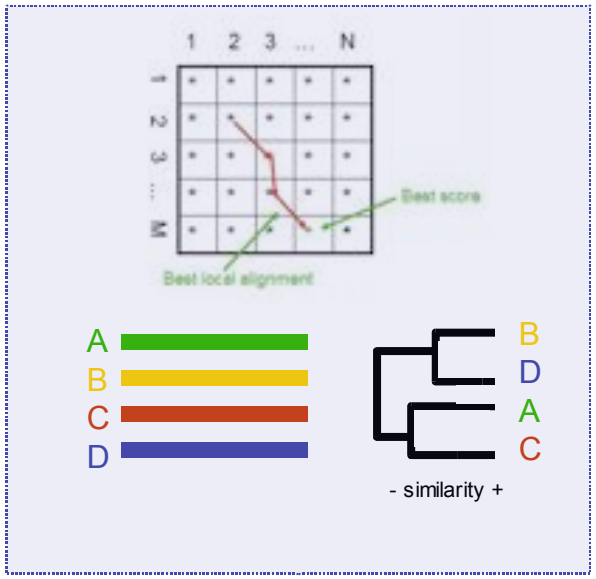
Threading (mGenThreader)

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



Remote homology detection (FUGUE)

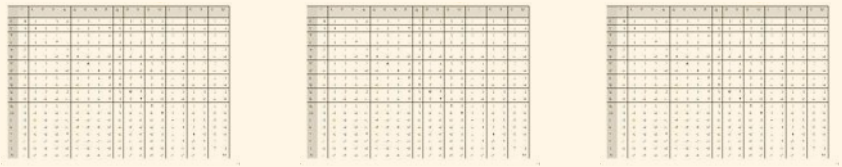




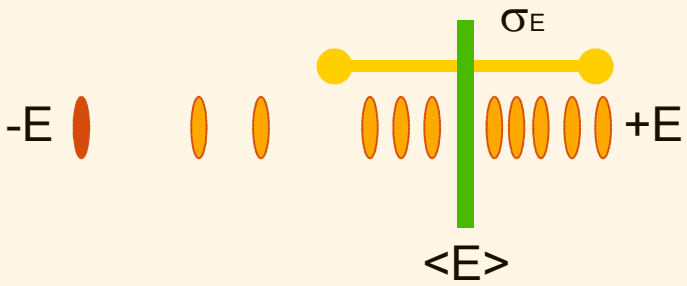
FUGUE

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

- ✘ Does not uses multiple sequence information



$R^{H}_{i,j}$ $R^{B}_{i,j}$ $R^{C}_{i,j}$

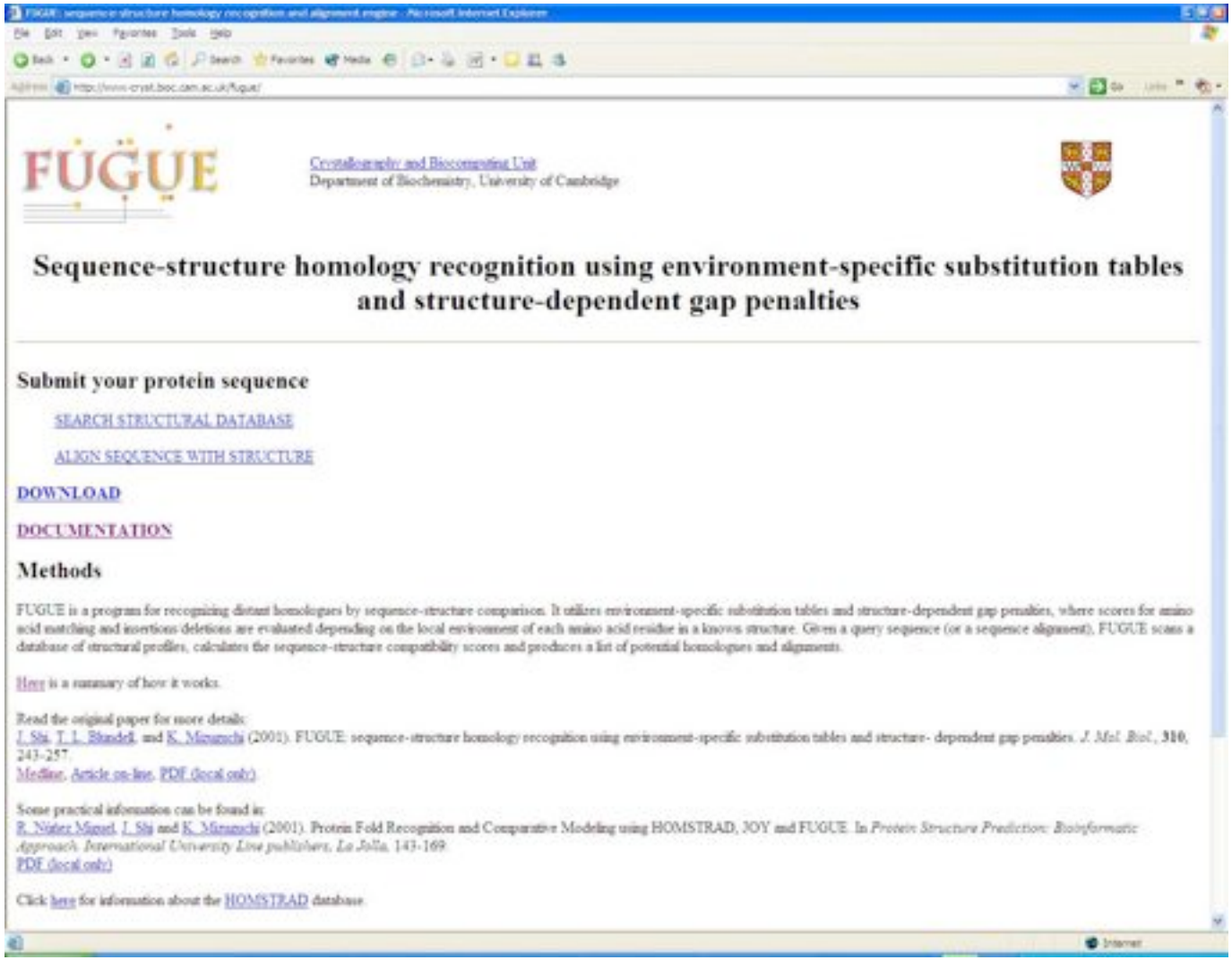


$$Zscore = \frac{\left(\langle E \rangle - E_m\right)}{\sigma_E}$$

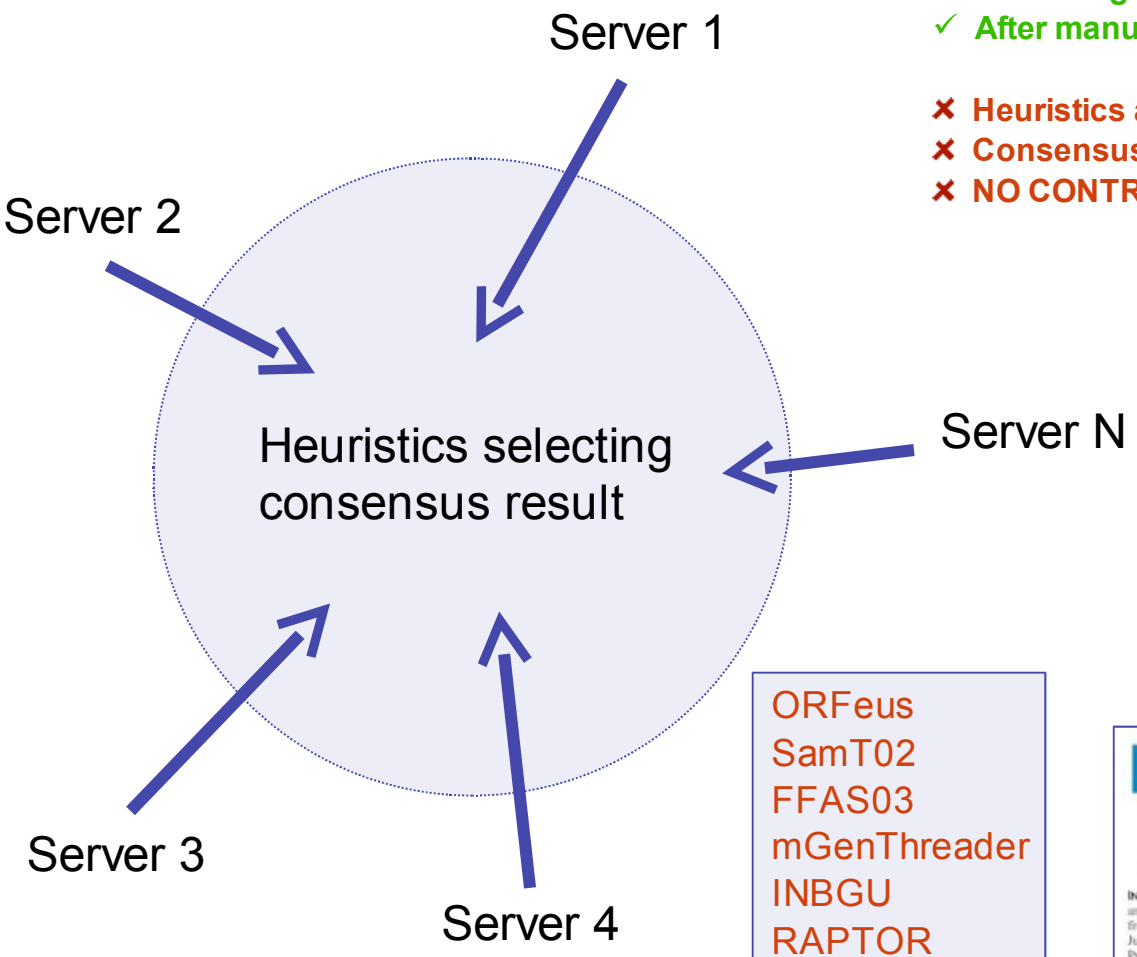
Shi et al. (2001) *J. Mol. Biol* **310** pp241

Remote homology detection (FUGUE)

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



Meta-Servers (3D-Jury)



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

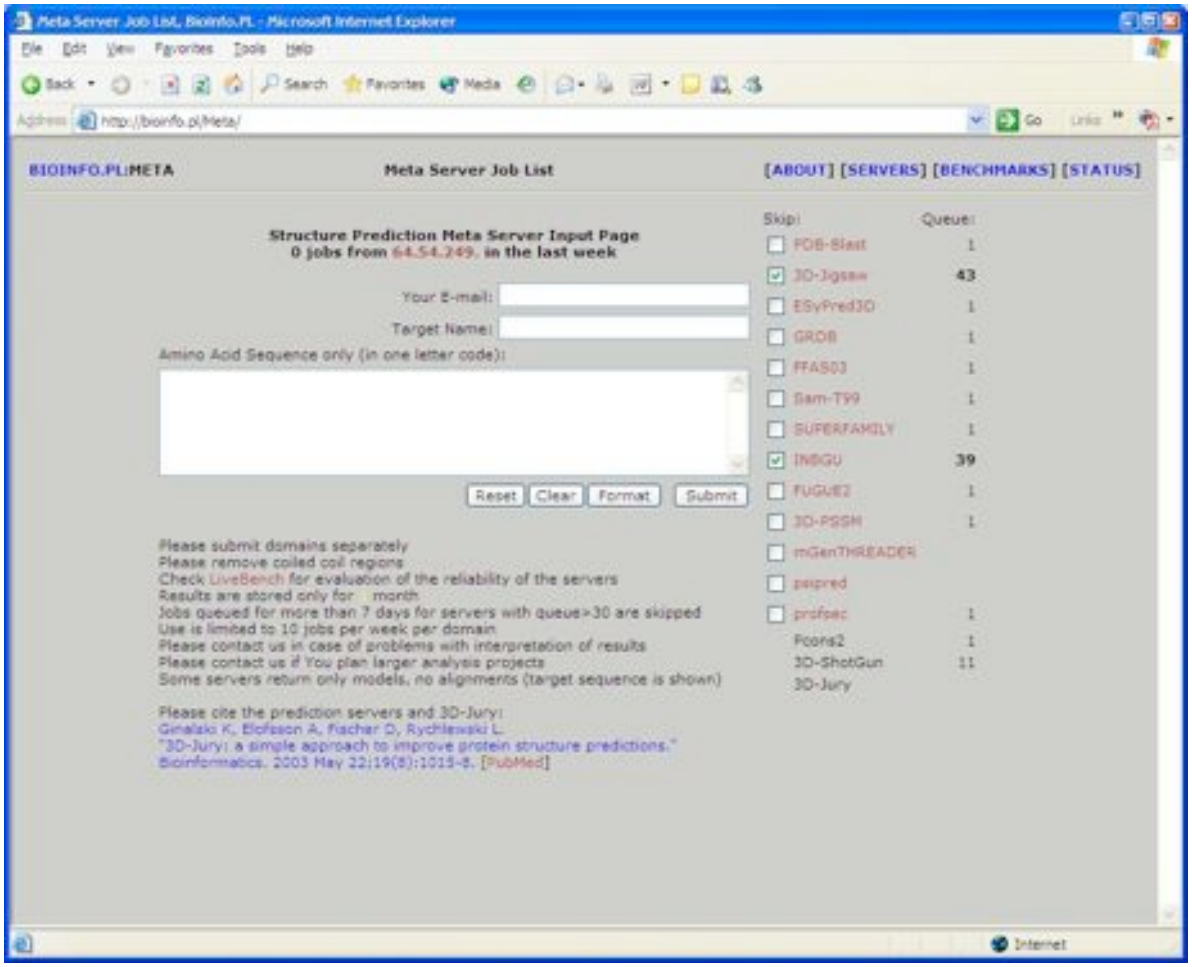
- ✘ Heuristics and complicated scoring
- ✘ Consensus results
- ✘ NO CONTROL OF DATA GENERATION or SERVERS!

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

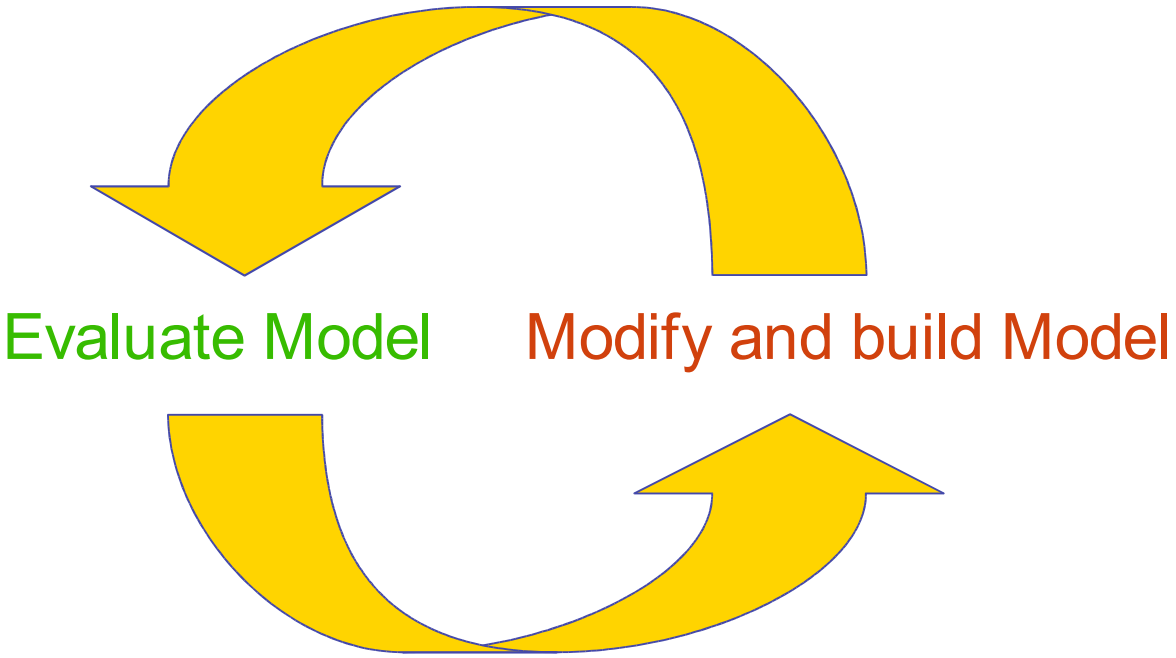
Ginalski K, et al. (2003) *Bioinformatics* **19** pp1015

Meta-Servers (3D-Jury)

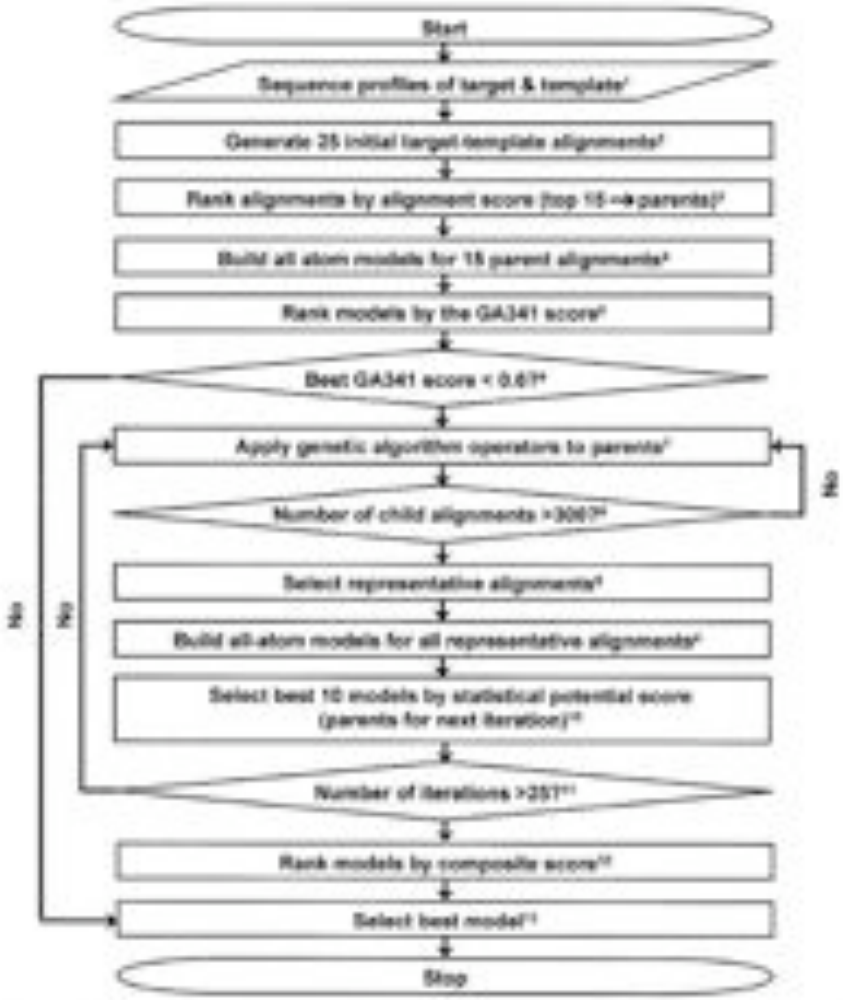
http://bioinfo.pl/Meta/

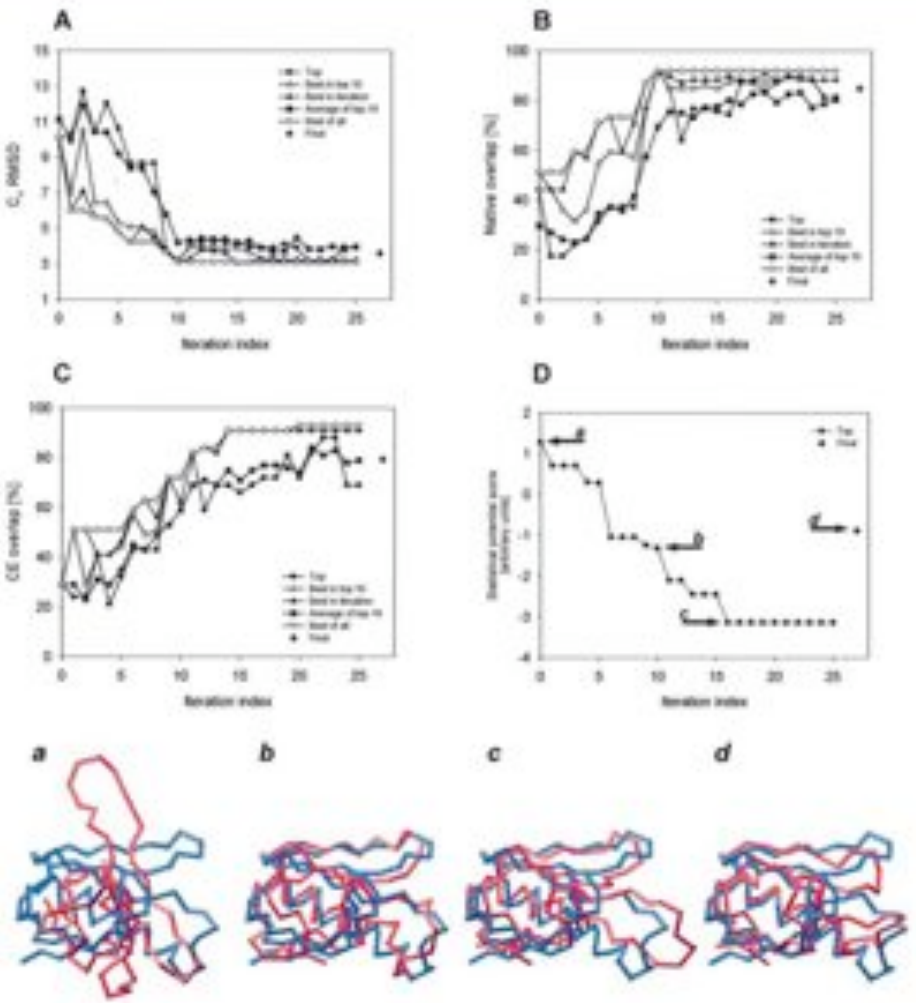


Iterative process… better models(?)



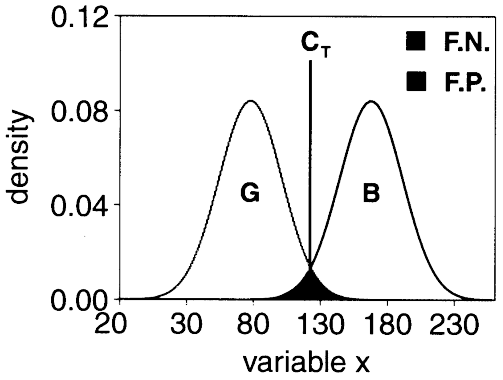
Iterative process… MOULDER





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

Iterative process… Evaluation

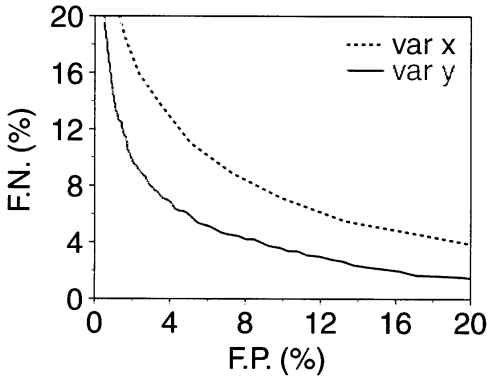


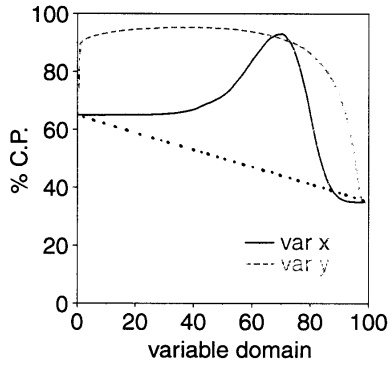
	is GOOD	*is BAD*
predicted as GOOD	a	b
predicted as BAD	c	d

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

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

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





Iterative process… Evaluation

1

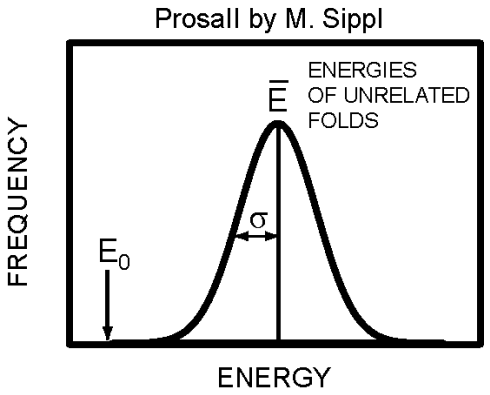
3900 GOOD MODELS

Models based on correct templates and approximately correct alignments

6000 BAD MODELS

Models based on incorrect templates or mostly incorrect alignments

2

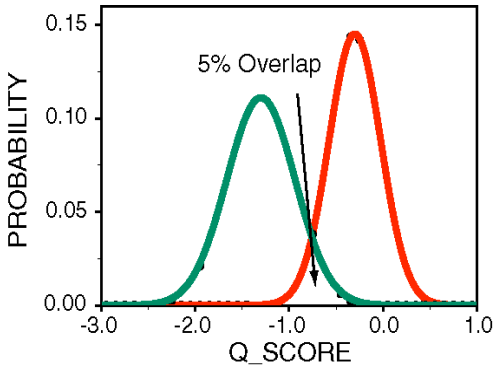


$$Z-SCORE = \frac{E_0 - \bar{E}}{\sigma}$$

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

3

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

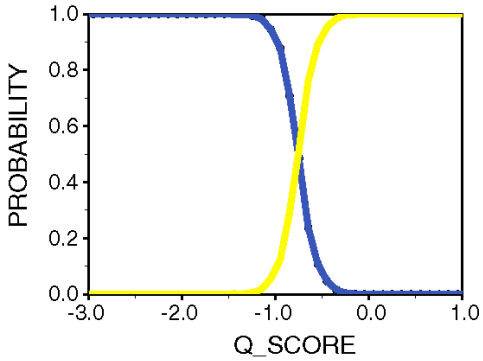


4

pG

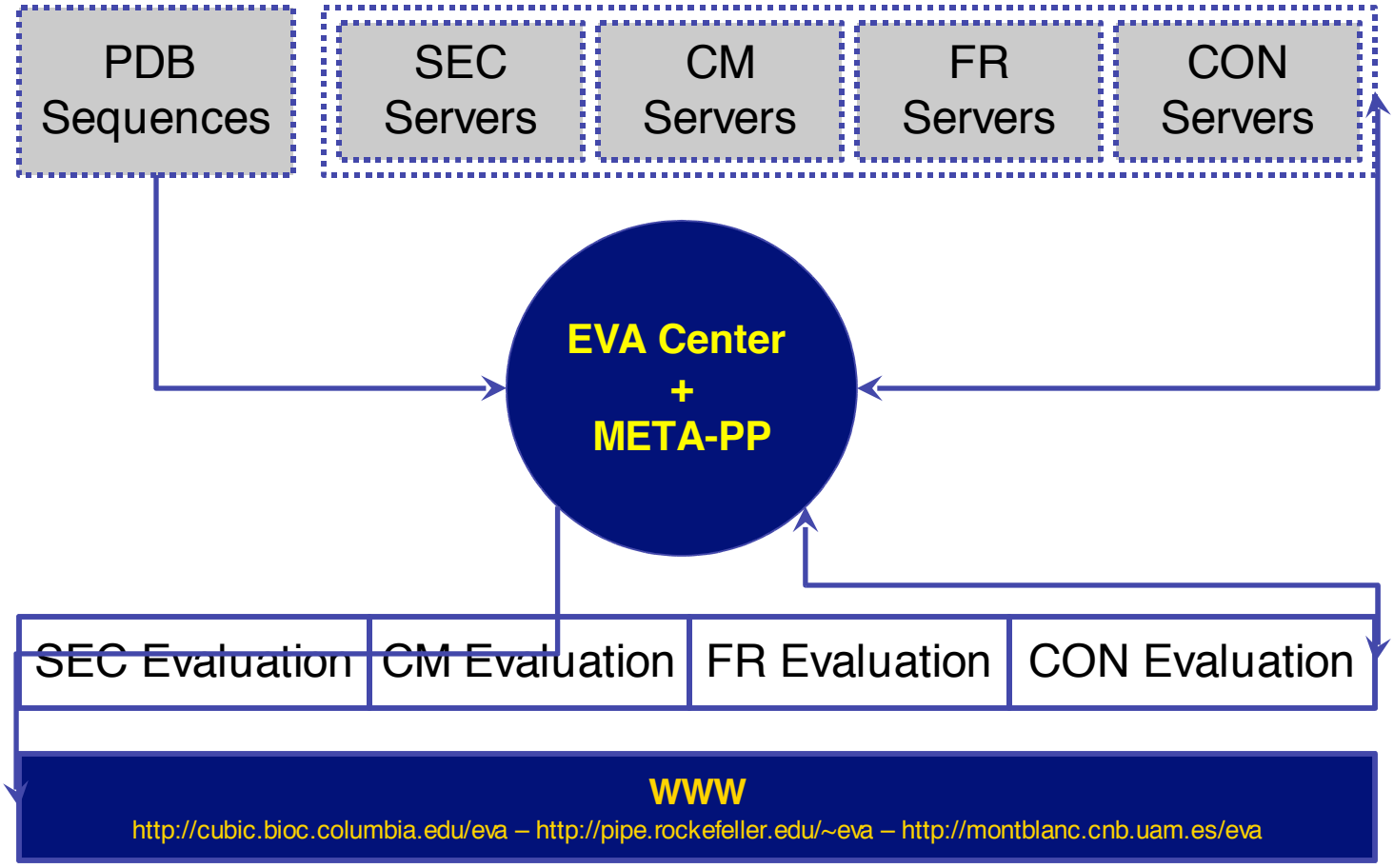
$$pG = \frac{p(\text{Q-SCORE/GOOD})}{p(\text{Q-SCORE/GOOD}) + p(\text{Q-SCORE/BAD})}$$

p(BAD/Q-SCORE) = 1 -



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

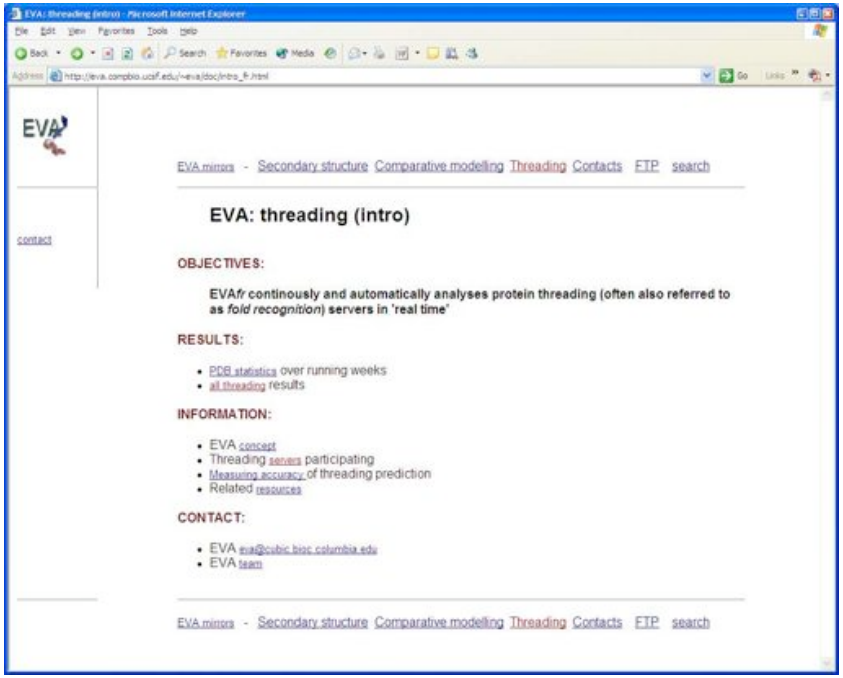
Eva Server

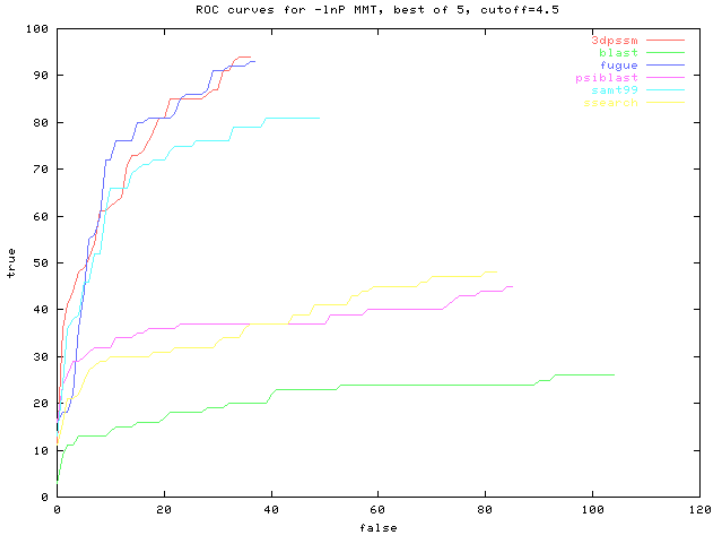


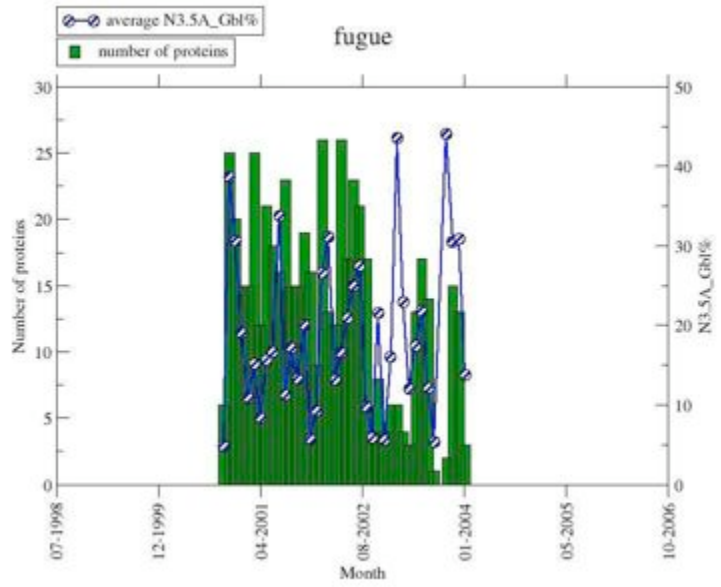
http://cubic.bioc.columbia.edu/eva - http://pipe.rockefeller.edu/~eva - http://montblanc.cnb.uam.es/eva

Eva-Threading Server

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



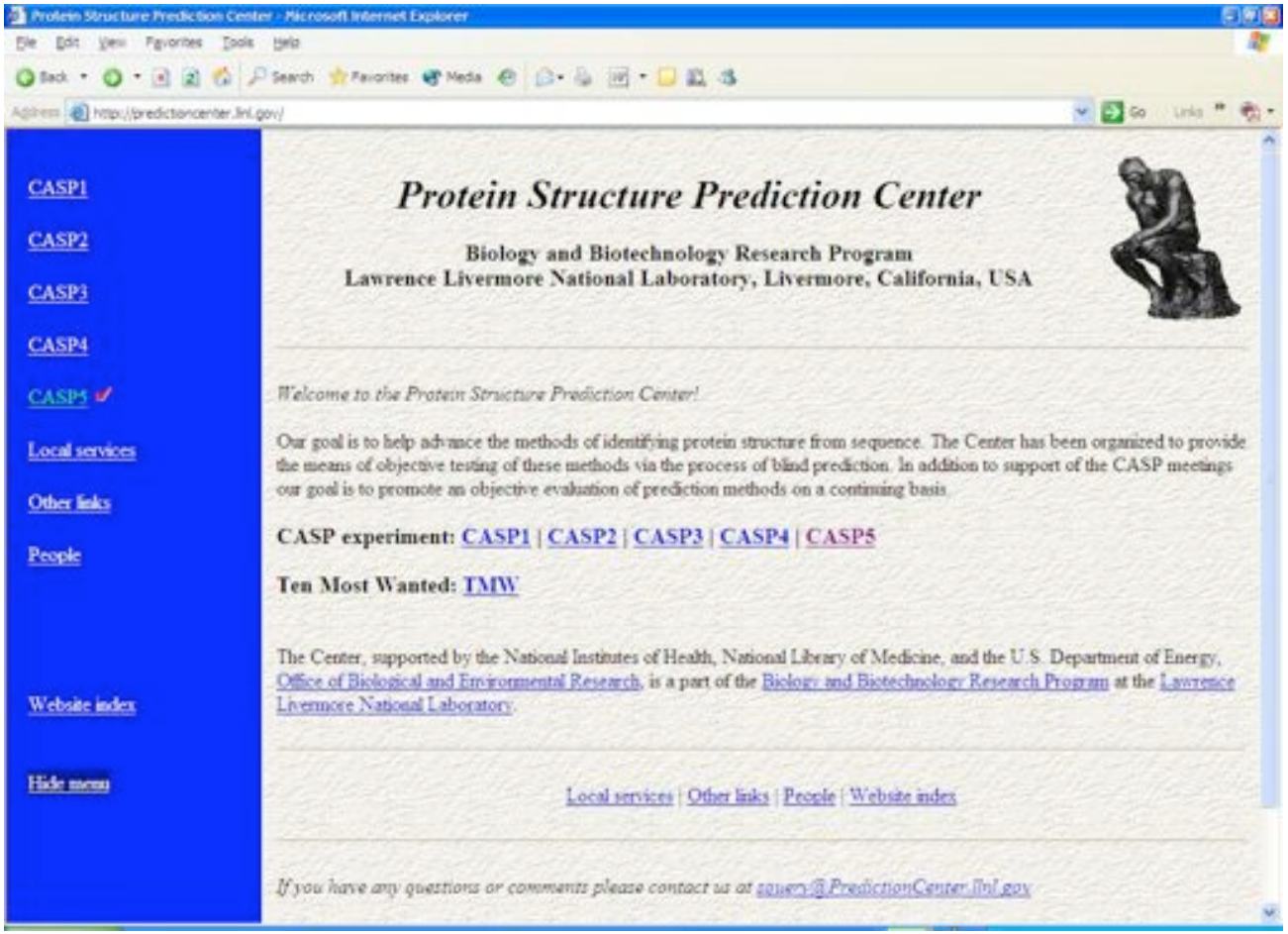




Other evaluation benchmarks

CASP **http://predictioncenter.llnl.gov**

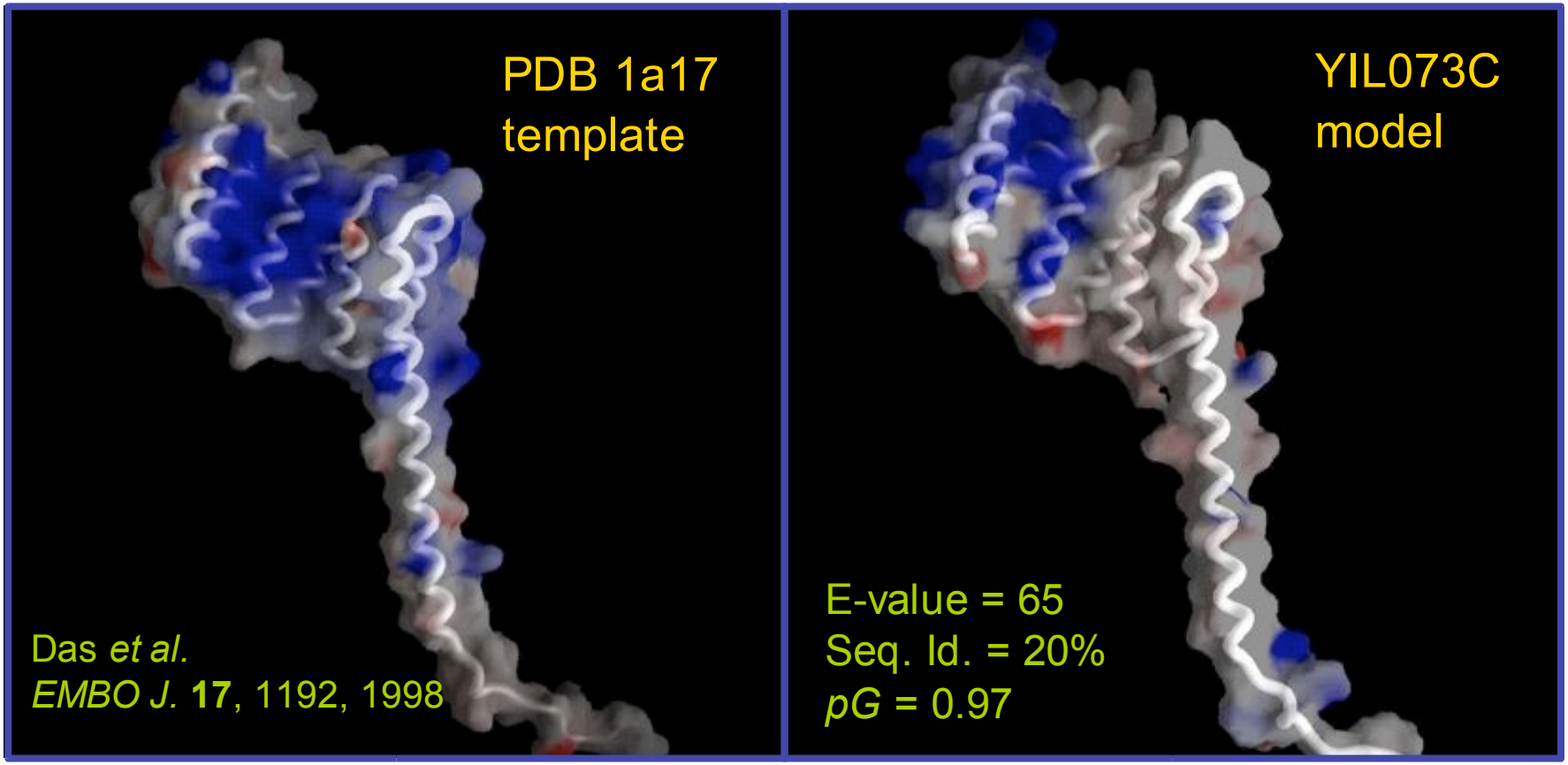
LiveBench **http://bioinfo.pl/LiveBench/**



Fold assignment from sequence examples….

MODPIPE Model of Yeast Hypothetical Protein YIL073C

(high e-value and good model score)



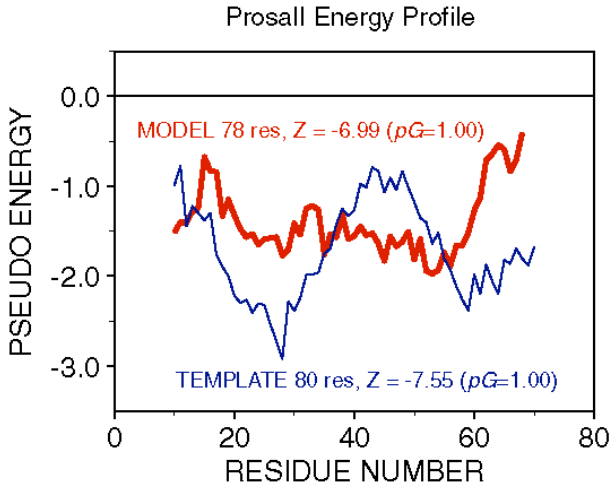
The tetratricopeptide repeat (TPR) is a degenerate 34 aa sequence identified in a variety of proteins, present in tandem arrays, mediates protein-protein interactions.

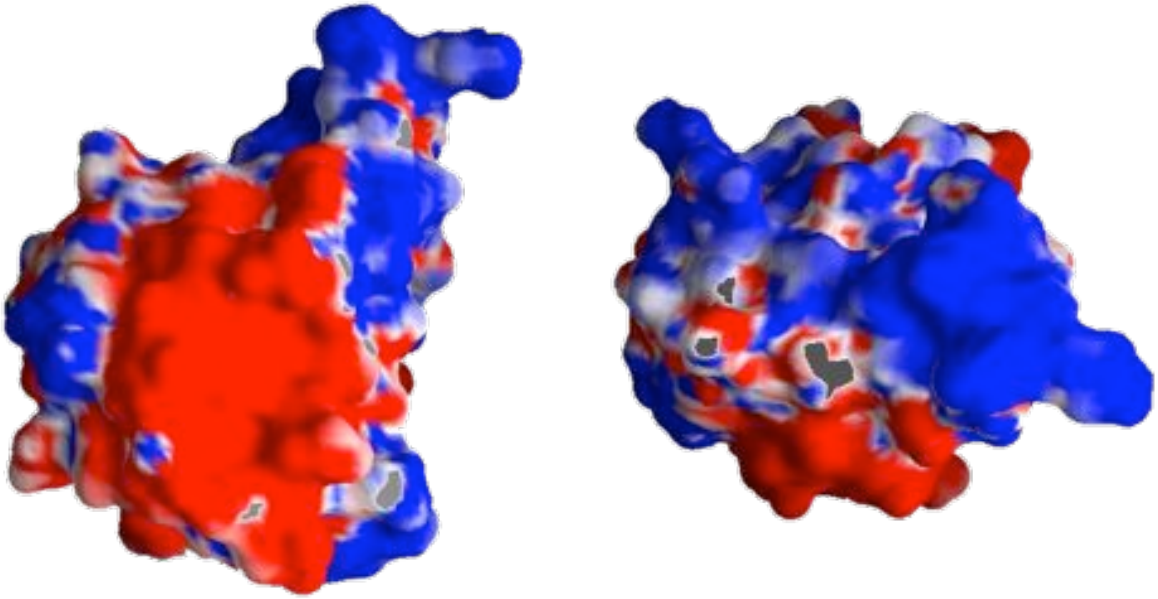
R. Sánchez, F. Melo, N. Mirkovic, A. Sali

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

(significant e-value and good model score)

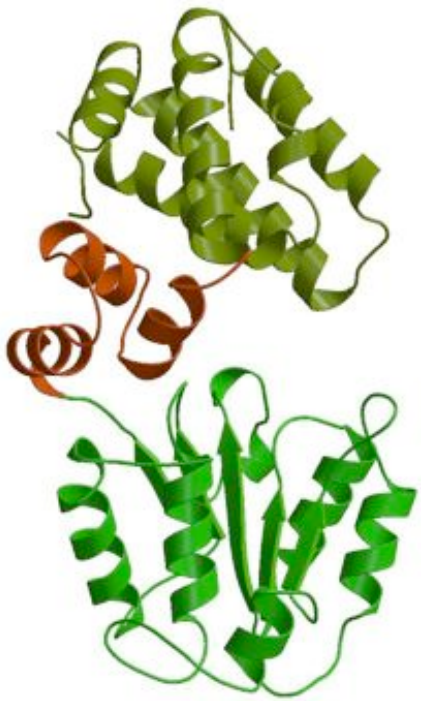
SEG FILTER	QUERY	MODEL SIZE	E-VALUE	*pG*
Y	Target	58	0.200	0.99
N	Target	64	0.029	1.00
Y	Template	NO HIT	NO HIT	NO HIT
N	Template	78	$6x10^{-14}$	1.00





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

(iterative manual process… Model building ← → Model assessment)

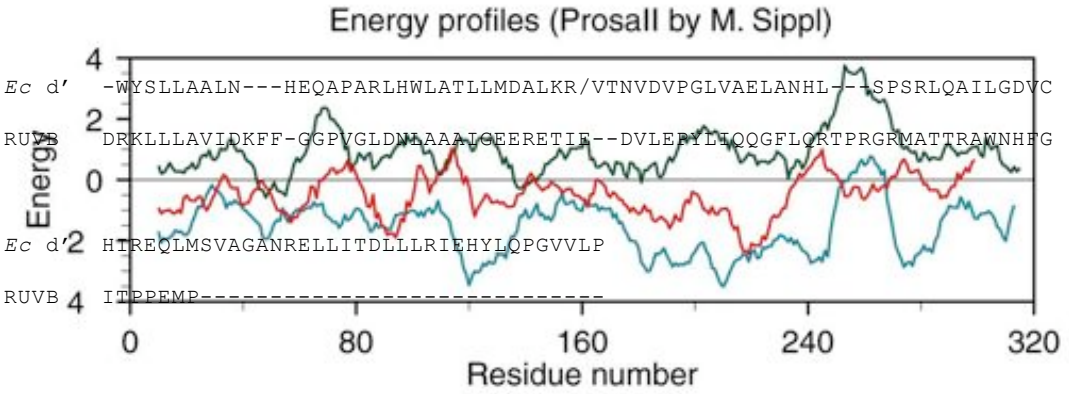


```
Ec d'   MRWYPWLRPDFEKLVASYQAGRG----HHALLIQALPGMGDDALIYALSRYLLCQQPQGHKSCGHCRG
RUVB    LEEYVGQPQVRSQMEIFIKAAKLRGDALDHLLIFGPPGLGKTTLANIVANEMG---------------

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-GGPVGLDNLAAAIGEEERETIR--DVLEPYLIQQGFLQRTPRGRMATTRAWNHFG
Ec d'   HTREQLMSVAGANRELLITDLLLRIEHYLQPGVVLP
RUVB    ITPPEMP-----------------------------
```

B. Guenther *et al.* (1997) *Cell* **91** pp335

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.