

BMC WorkShop

Protein Structure Prediction

Sequence-Structure alignment (template selection)

Marc A. Marti-Renom & Damien Devos

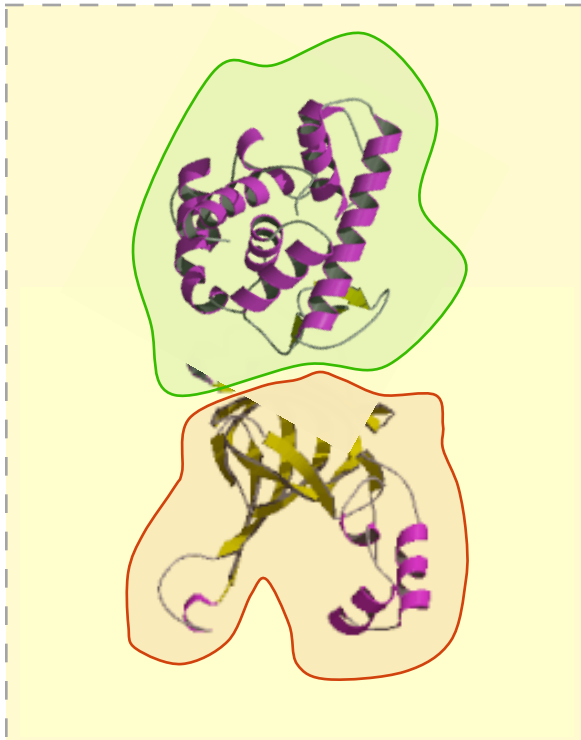
Department of Biopharmaceutical Sciences, UCSF

June 17th and 18th, 2004

Domains (?)

Domain boundaries from sequence

VERY DIFFICULT!!!!



```
MENFEIWVEKYRPRTLDEVVGQDEVIQRLKGYVERKNIPELLFSGPPGTGKTATAIALARDLFGENWRDN  
FIEMNASDERGIDVVRHKIKEFARTAPIGGAPFKIIFLDEADALTADAQAALRRRTMEMYSKSCRFILSCN  
YVSRIIEPIQSRCAVFRFKVPKPEAMKKRLEICEKEGVKITEDGLEALIYISGGDFRKAINALQAAAI  
GEVVDADTIYQITATARPEEMTELIQTALKGNFMEARELLDRLMVEYGMSEGEDIVAQLFREIISMPIKDS  
LKVQLIDKLGEVDFRLTEGANERIQLDAYLAYLSTLAKK
```

Domain boundaries from sequence (SnapDragon)

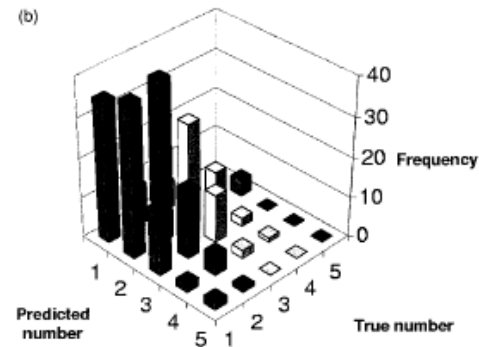
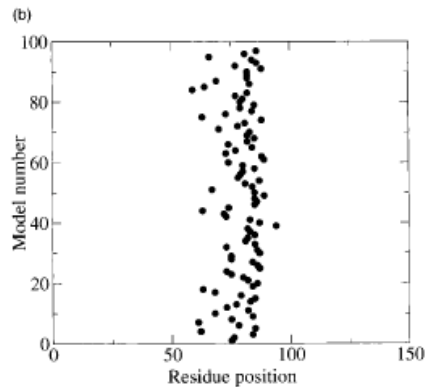
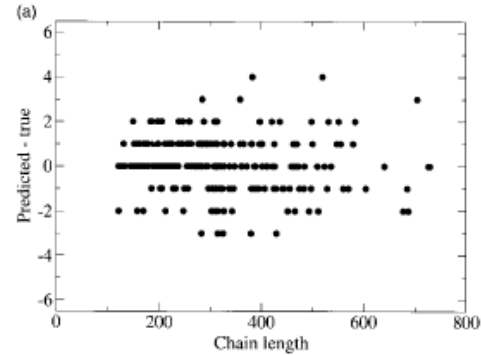
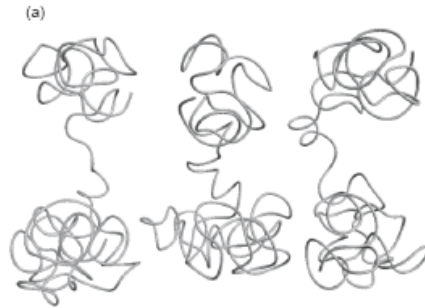
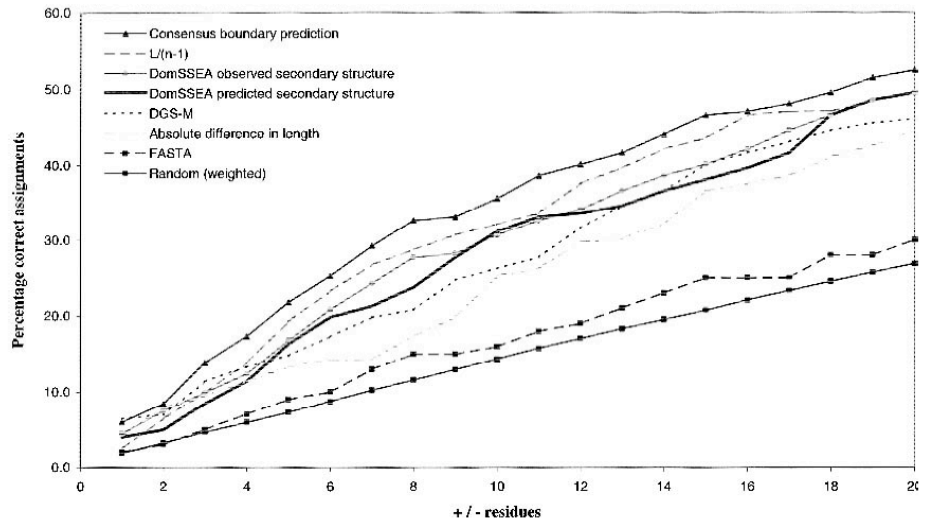
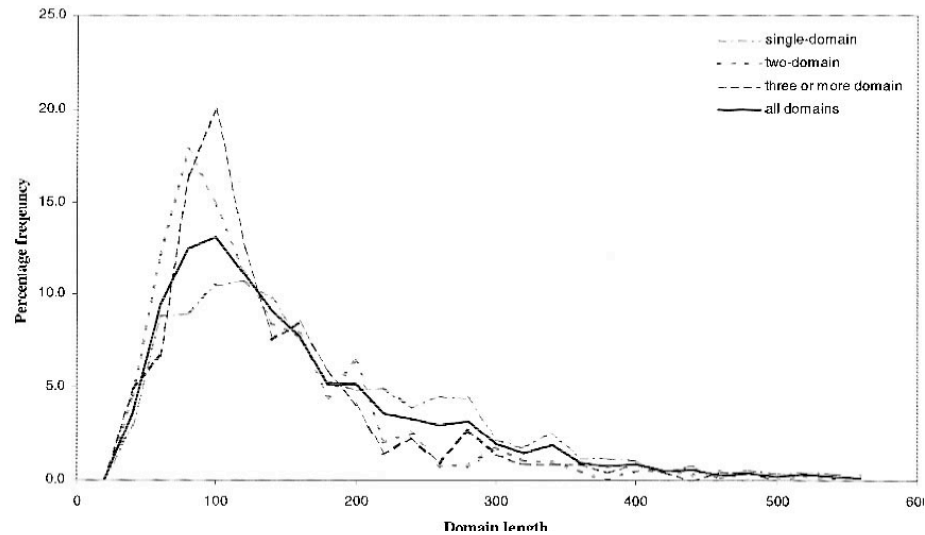


Table 2. Average accuracy percentages of linker prediction over 57 proteins

		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

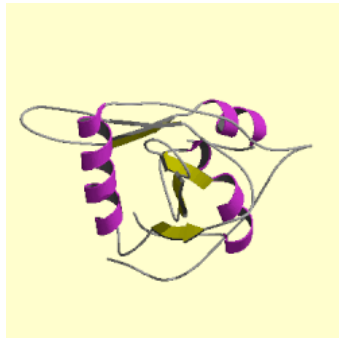
Domain boundaries from sequence (DomSSEA)



Prediction of Secondary Structure (PSI-PRED)

>gi42541361
MDIRSVSSLRGLLCLPPSWPRR

- Neural Network



- ✓ Very simple idea
- ✓ Simple scoring

Obscure optimizer

Raw profile from PSI-BLAST Log File

Position-based scoring matrix used

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

Window of 15 rows

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

15 x 20 scaled inputs to 1st network

1st Network
315 inputs
75 hidden units
3 outputs

Window of 15 x 3
outputs fed to 2nd
network

2nd Network
60 inputs
60 hidden units
3 outputs

Final 3-state
Prediction

Prediction of Secondary Structure (PSI-PRED)

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

The screenshot shows the PSIPRED Protein Structure Prediction Server web interface. The browser window title is "PSIPRED Protein Structure Prediction Server - Microsoft Internet Explorer". The address bar shows the URL "http://bioinf.cs.ucl.ac.uk/psiform.html". The page features a blue header with the UCL logo and the text "Bioinformatics Unit". Below the header, there are several sections:

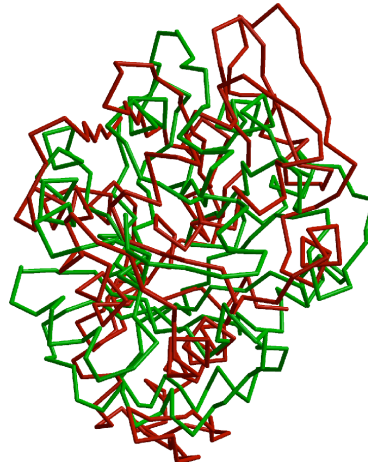
- PSIPRED home>**: A link to the home page.
- Info**: A message stating: "We suggest that you do not bookmark this page as it is liable to move. It is best to access the server via the [PSIPRED home page](#), which has more information about the methods and a full reference list."
- Input Sequence**: A section with a "Help" link and the text "Input sequence (single letter code)". Below this is a text input field.
- Choose Prediction Method**: A section with a "Help" link and four radio button options:
 - Predict Secondary Structure (PSIPRED v2.4)
 - Predict Transmembrane Topology (MEMSAT)
 - Fold Recognition(GenTHREADER - quick)
 - Fold Recognition (mGenTHREADER - with profiles and predicted secondary structure)
- Filtering Options**: A section with a "Help" link and three checkboxes:
 - Mask low complexity regions
 - Mask transmembrane helices
 - Mask coiled-coil regionsA warning message below reads: "Warning: Turn off all filtering if you are running MEMSAT".
- Submit Sequence**: A section with three text input fields and two buttons:
 - E-mail address** [Help](#)
 - Password (only required for commercial e-mail addresses)** [Help](#)
 - Short name for sequence** [Help](#)Buttons for "Predict" and "Clear form" are located at the bottom of this section.

Why the alignment is so important?

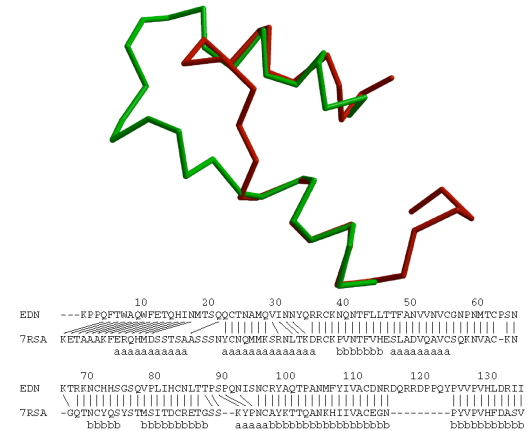
Typical errors in comparative models

MODEL
X-RAY
TEMPLATE

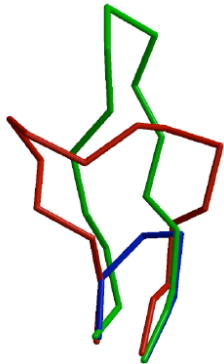
Incorrect template



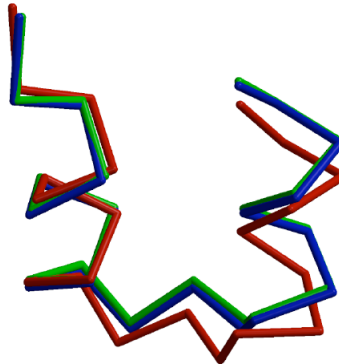
Misalignment



Region without a template



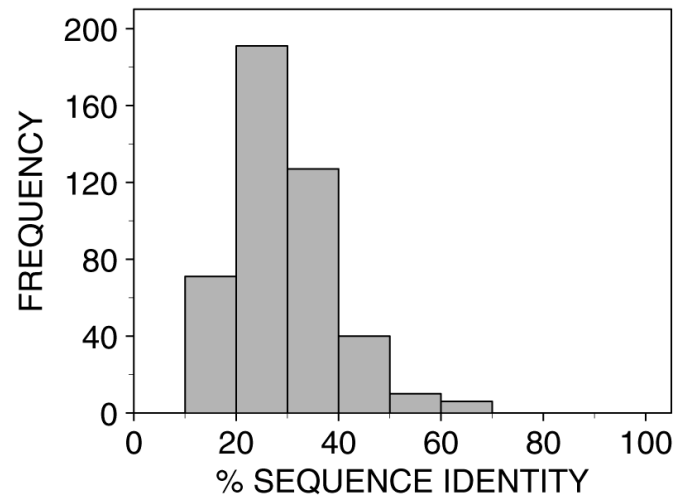
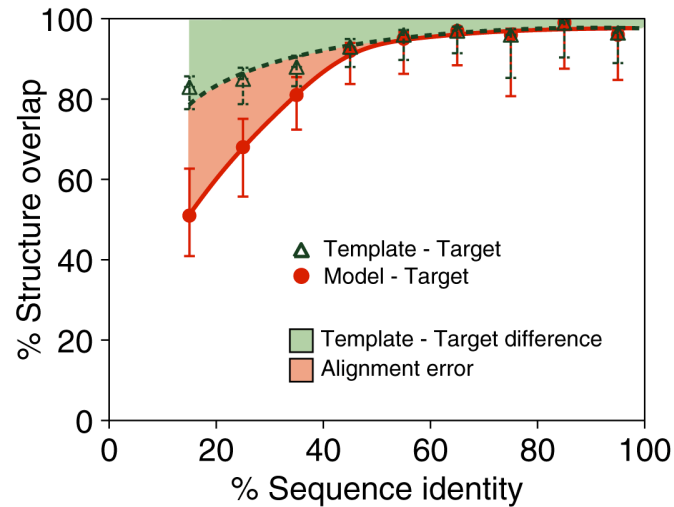
Distortion/shifts in aligned regions



Sidechain packing



Alignment errors are frequent and large



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

Minimizing errors in sequence-structure alignment

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

Threading

General overview (Threading)

- Matches sequences to 3D structures
 - Requires a scoring function to 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.

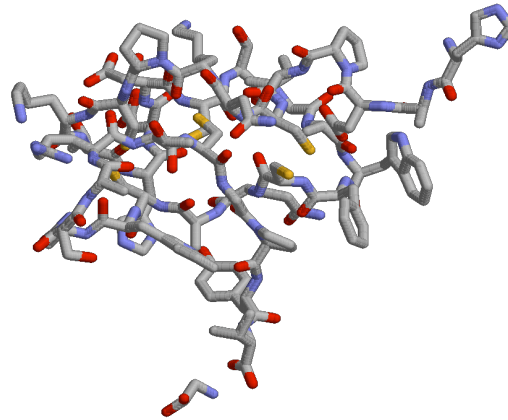
PMF versus Energy (additive/higher order terms).

Reference state.

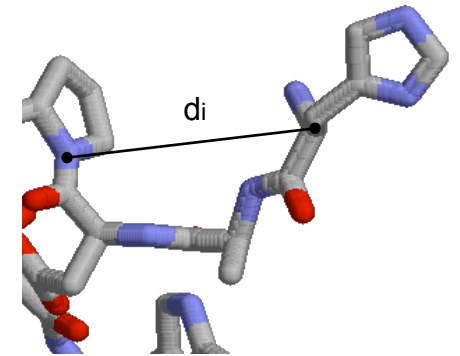
Physical origin.

Representation Sequence/Structures

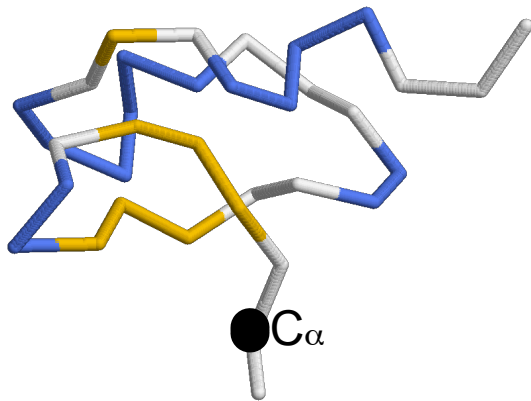
```
>gi42541361  
MDIRSVSSLRGLLCLPPSWPRR
```



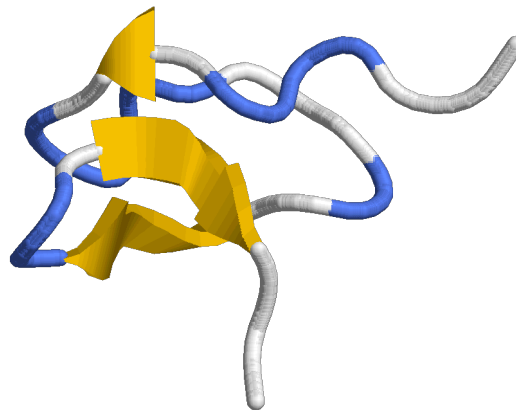
All atoms and coordinates



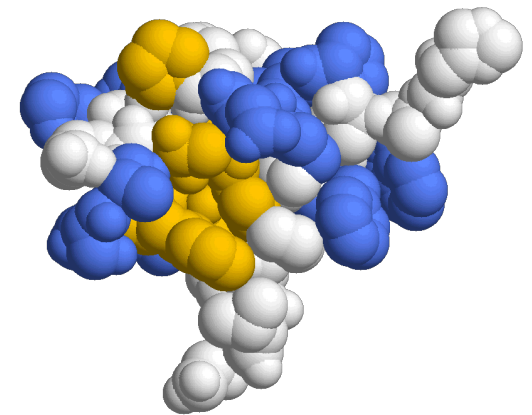
Distance space



Reduced atoms representation



Secondary Structure



Accessible surface

Scoring Statistical Potential... inspiration

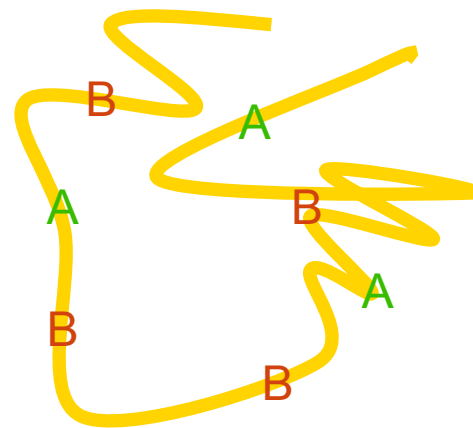
$$K = \frac{[AB]}{[A] \cdot [B]}$$

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

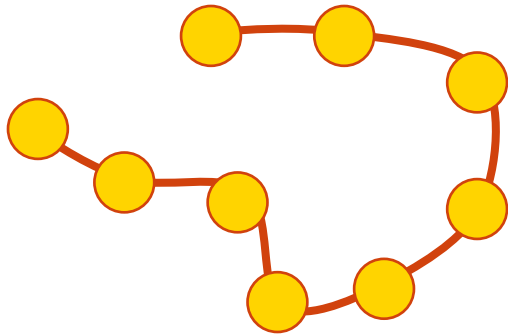
From statistical physics, we know that energy difference between two states (ΔE) and the ratio of their occupancies ($N_1:N_2$) are related [9]:

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

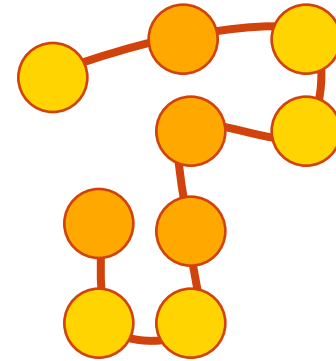
in which T is the absolute temperature and k is the Boltzmann's constant. As we are interested in an interaction energy between two amino acid side chains, it would seem natural to define N_1 as the number of interactions between these two residues types in a group of real protein structures, a number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



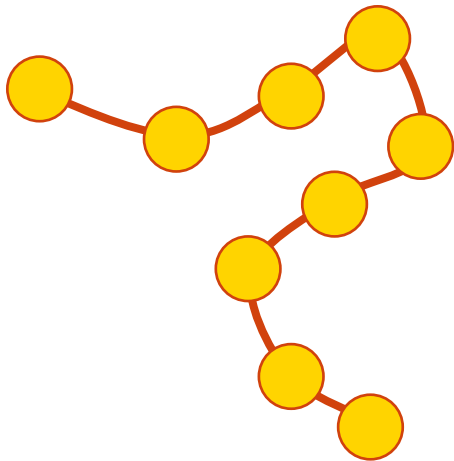
Scoring Statistical Potential... interaction types



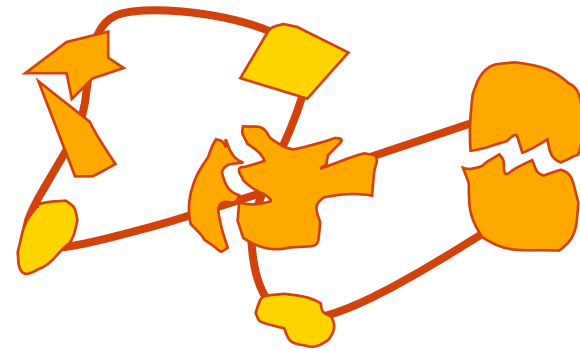
Neutral interactions



Hydrophobic interactions

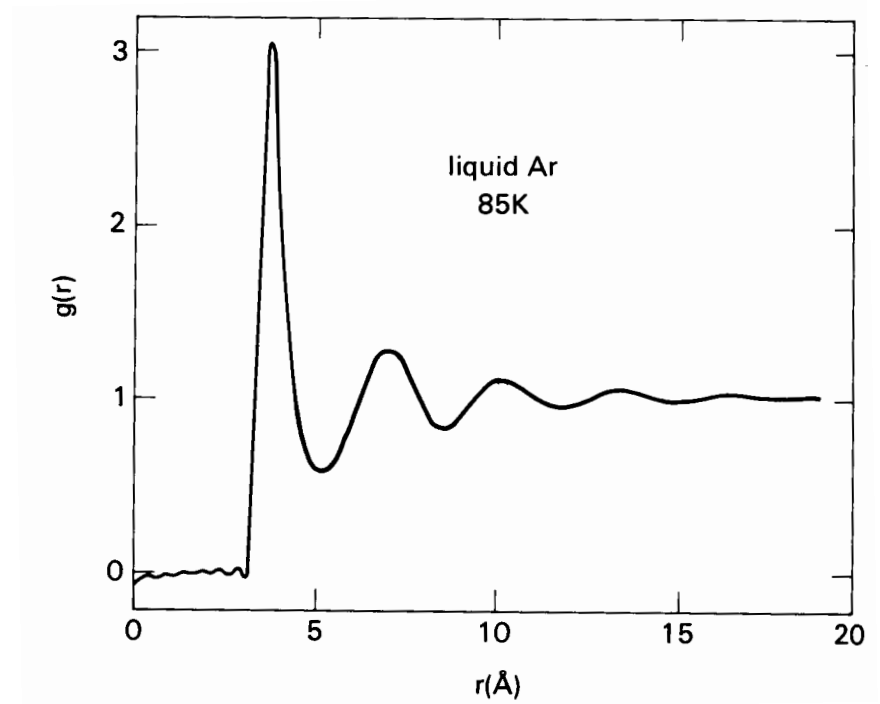
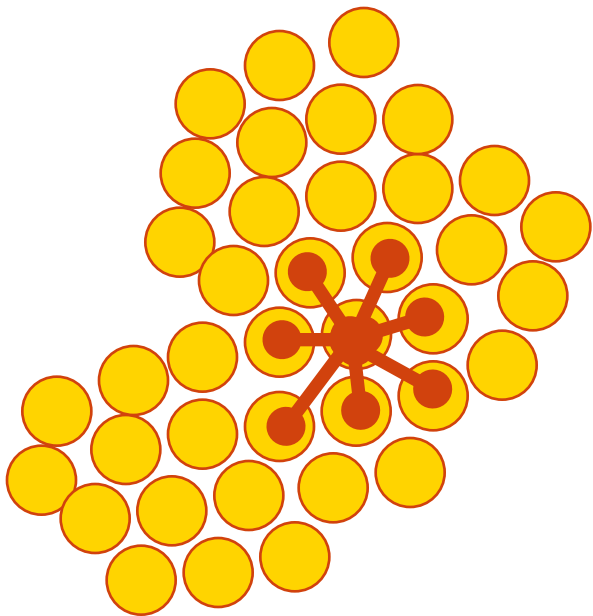


Compact interactions



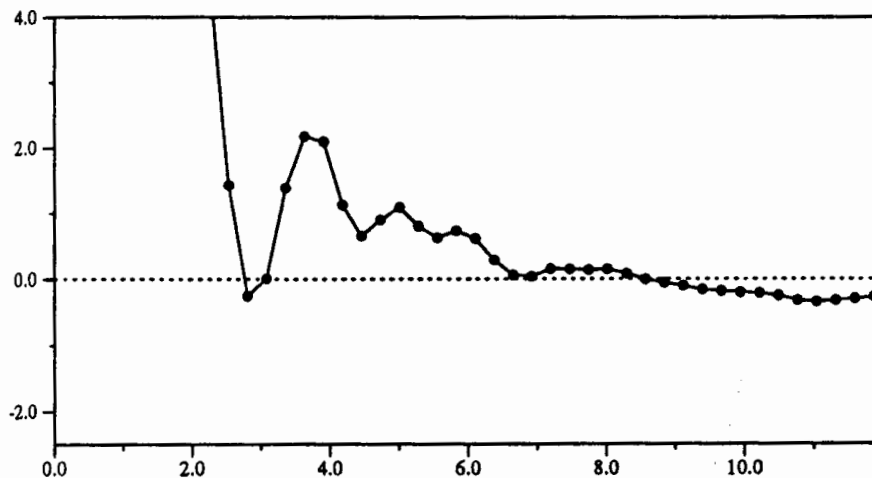
Specific interactions

Scoring Statistical Potential... reference state

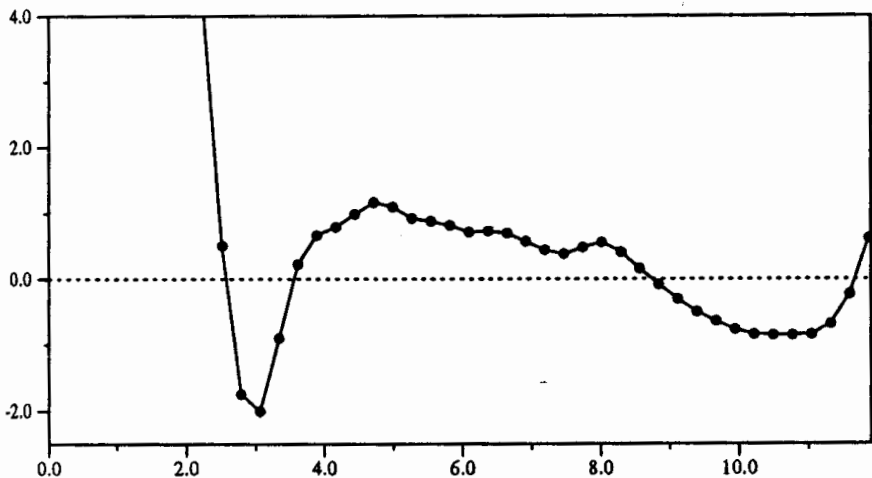


Scoring Statistical Potential... Hydrogen Bonds

Long range free energy



Short range free energy



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

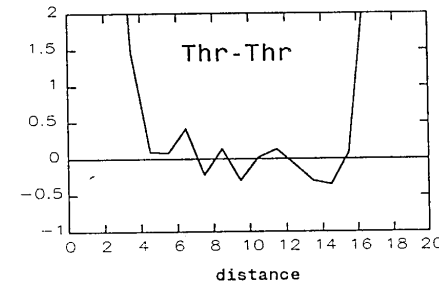
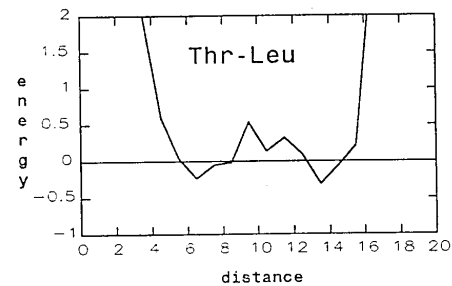
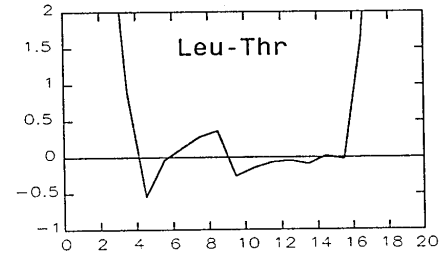
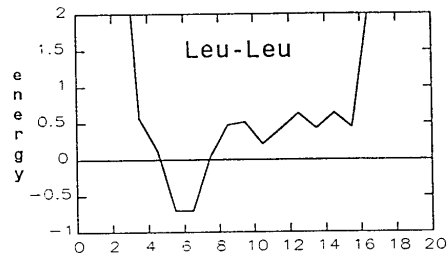
$$\rho_{NO}(r) = \sum_j \delta(r - r_j)$$

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

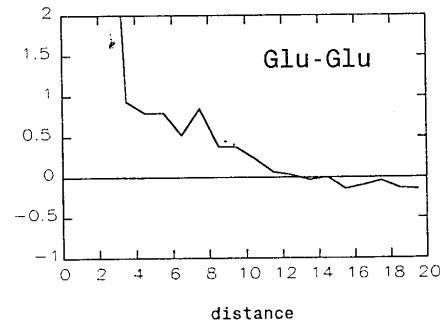
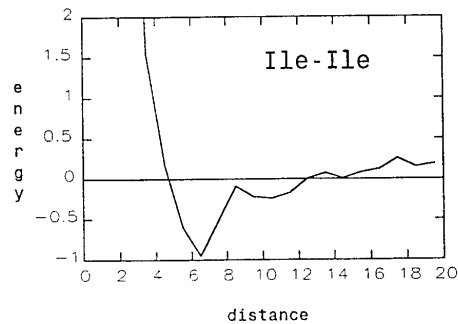
$$w_{NO}(r) = -kT \ln(g_{NO}(r))$$

Scoring Statistical Potential... Distance Potentials

Long range free energy



Short range free energy

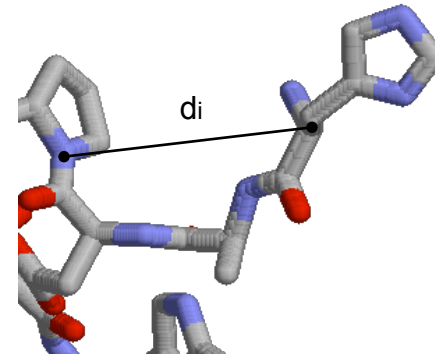


Scoring

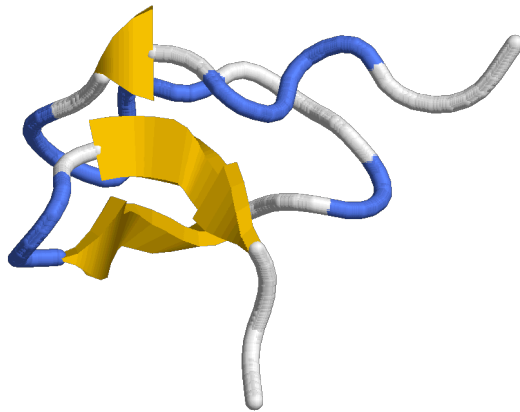
Raw scores of an alignment

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

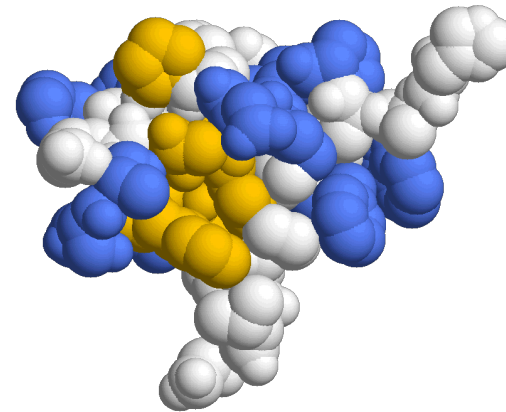
Aminoacid substitutions



Distance space



Secondary Structure (H,B,C)



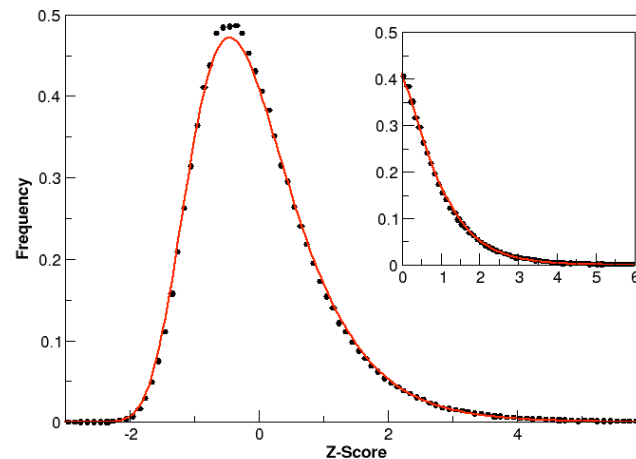
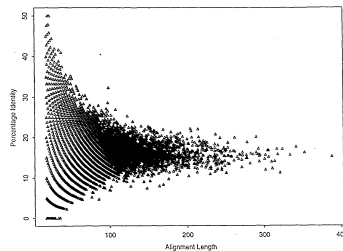
Accessible surface (B,A [%])

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

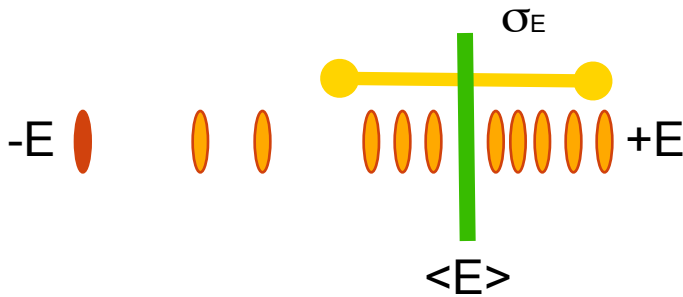
$$P(s) = e^{-\lambda (s-\mu)}$$

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

Scoring

Significance of an alignment (score)

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

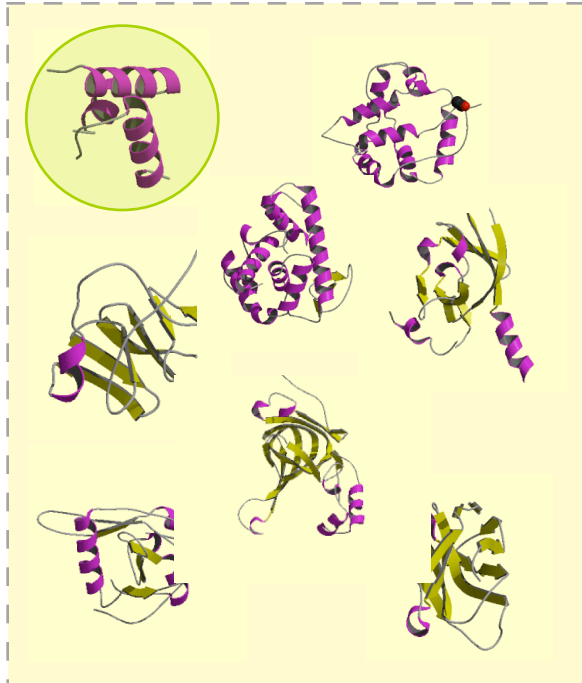


$$Zscore = \frac{(\langle E \rangle - E_m)}{\sigma_E}$$

Scoring

Significance of an alignment (background)

Structural space

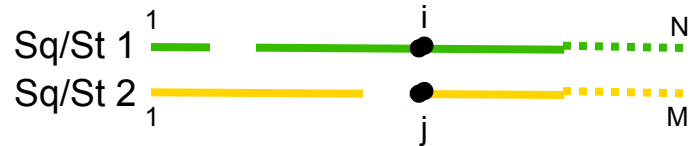


Sequence space

```
MKLLIVLTCISLCSICTVVQRCASNKPHVLEDPCVKVQH
HLSVNQCVLLPQCCPKSCKICTHLISIEVVLTCRAVDKM
MHVNCVEQCSLQDCIKIAPRVLKTCILCVLKPCLTSVSH
VHLVQPTSCCCKKNCICHVEIRSLDILTKSVQLACLVPM
      ■
      ■
      ■
MQCCRVOQKICDLLAVELCKLHISTPCKILCVVTSVPHN
```

Optimizer

Global dynamic programming alignment



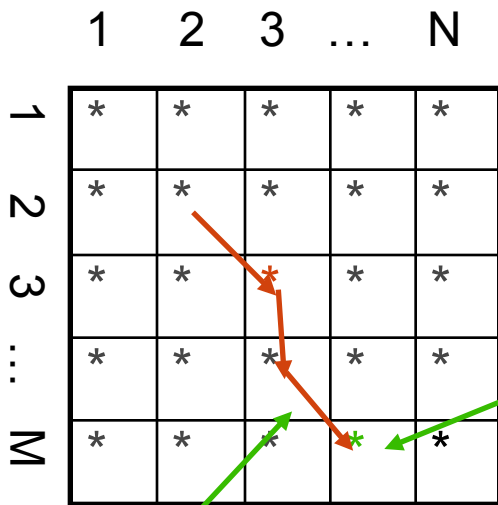
	1	2	3	...	N
1	*	*	*	*	*
2	*	*	*	*	*
3	*	*	*		
...					
M					*

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

Best alignment score

Backtracking to get the best alignment

Local dynamic programming alignment



Best local alignment

Best score

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

Backtracking to get the best alignment

Applications of PMFs

Model assessment.

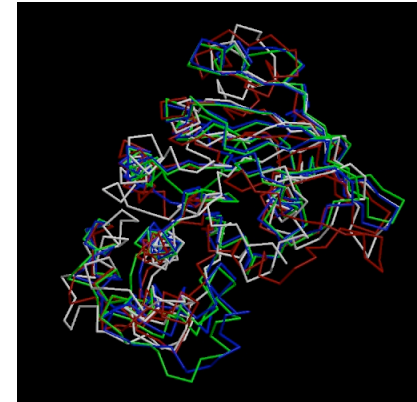
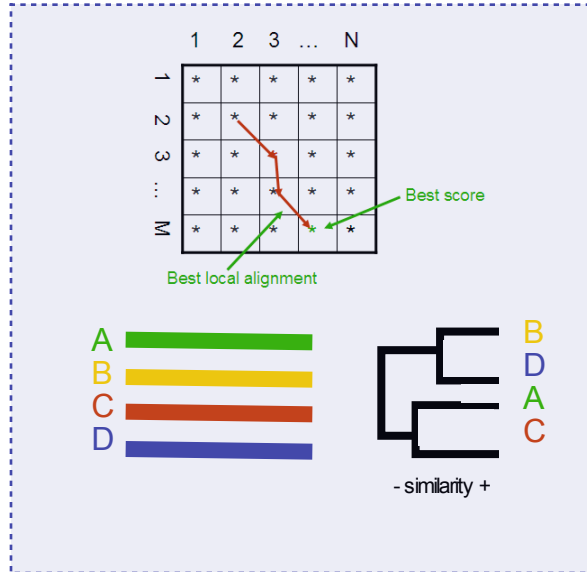
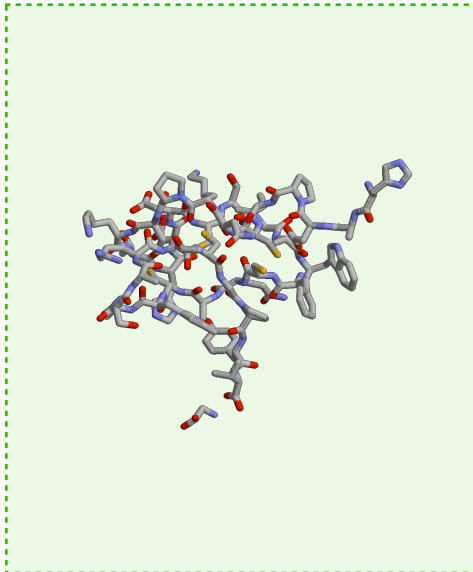
Ab initio folding simulations.

Sequence-structure matching (threading).

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

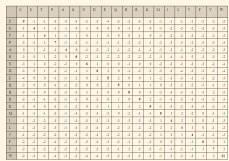
Secondary structure prediction, *etc.*

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

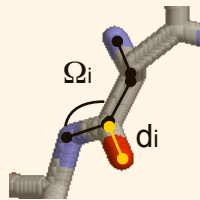


- ✓ Uses all available structural information
- ✓ Provides the optimal alignment

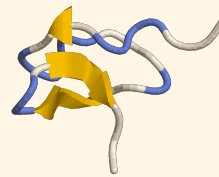
Computationally expensive



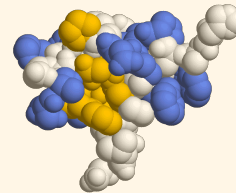
$R_{i,j}$



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



$S_{i,j}$



$B_{i,j}$

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

$I_{i,j}$

Structural alignment by properties conservation (SALIGN-MODELLER)

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

DBAli v2.0 tools page - Microsoft Internet Explorer

Address: http://salilab.org/DBAli/?page=tools&action=f_salign

UCSE | Salilab | MAMMOTH

DBAli v2.0

Tools last update Feb 11th, 2004

[Home](#)
[Search DBAli](#)
[Tools](#)
[Help](#)

DBAli ALERT!
09/21/2003 --
You are visiting the DBAli.v2 pages. This pages contain the updated DBAli database. You can still visit the old DBAli database [here](#).

DBAli. Tools associated to the database.

- [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](#)

Get a multiple structure alignment of a list of chains.

File with a list of chains:

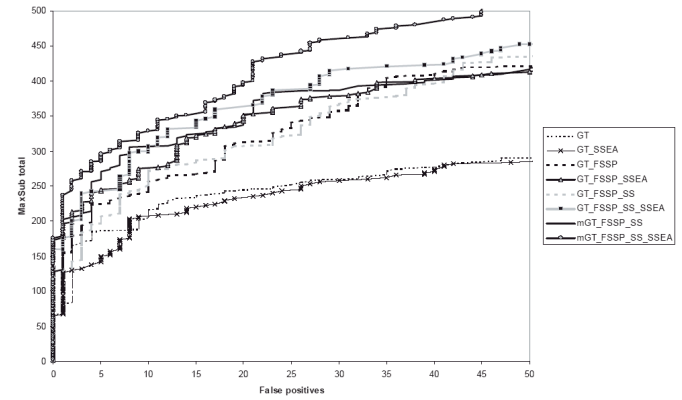
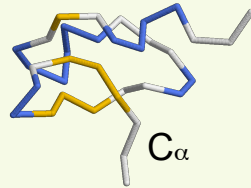
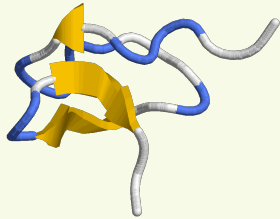
 ?

Reference :: [Download](#) :: [Statistics](#) :: [Suggestions](#) Visitors: 1407 © 2003 - 2004 Marti-Renom

Threading (mGenThreader)

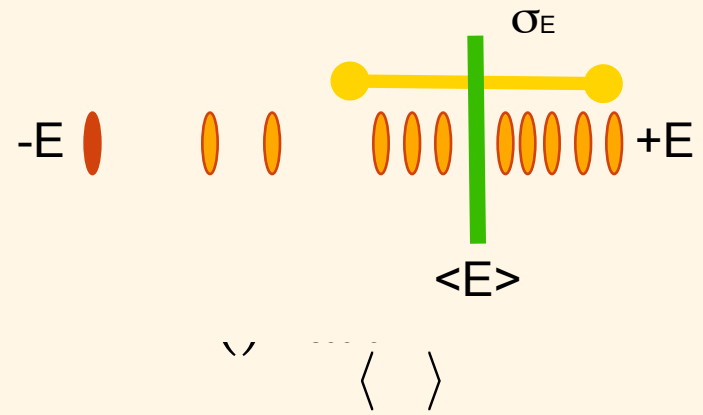
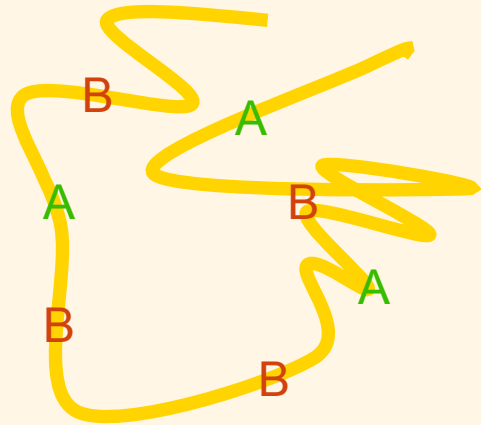
>gi42541361
MDIRSVSSLRGLLCLPPSWPRR

- Neural Network



✓ Good row and significance scoring

Obscure optimizer



Threading (mGenThreader)

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

PSIPRED Protein Structure Prediction Server - Microsoft Internet Explorer

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

Bioinformatics Unit

PSIPRED home>

The PSIPRED Protein Structure Prediction Server

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

Input Sequence [Help](#)
Input sequence (single letter code)

Choose Prediction Method [Help](#)

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

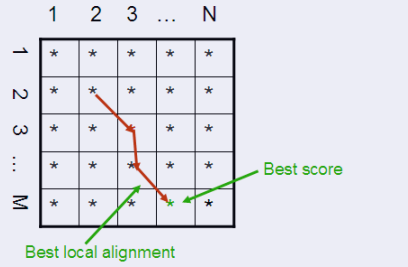
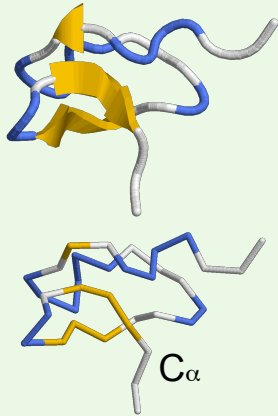
Filtering Options [Help](#)

- Mask low complexity regions
- Mask transmembrane helices
- Mask coiled-coil regions

Warning: Turn off all filtering if you are running MEMSAT

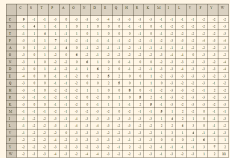
Remote homology detection (FUGUE)

>gi42541361
MDIRSVSSLRGLLCLPPSWPRR

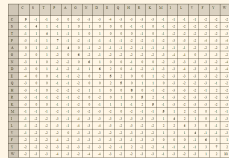


- ✓ Uses most of the structural information
- ✓ Easy to access either locally and on the web
- ✓ Good row and significance scoring

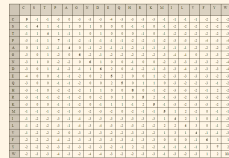
Does not uses multiple sequence information



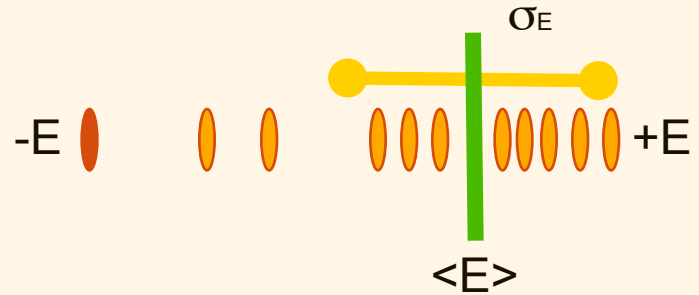
R_{ij}^H



R_{ij}^B



R_{ij}^C



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

Remote homology detection (FUGUE)

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

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

File Edit View Favorites Tools Help

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

FUGUE Crystallography and Biocomputing Unit
Department of Biochemistry, University of Cambridge

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

Submit your protein sequence

[SEARCH STRUCTURAL DATABASE](#)

[ALIGN SEQUENCE WITH STRUCTURE](#)

[DOWNLOAD](#)

[DOCUMENTATION](#)

Methods

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

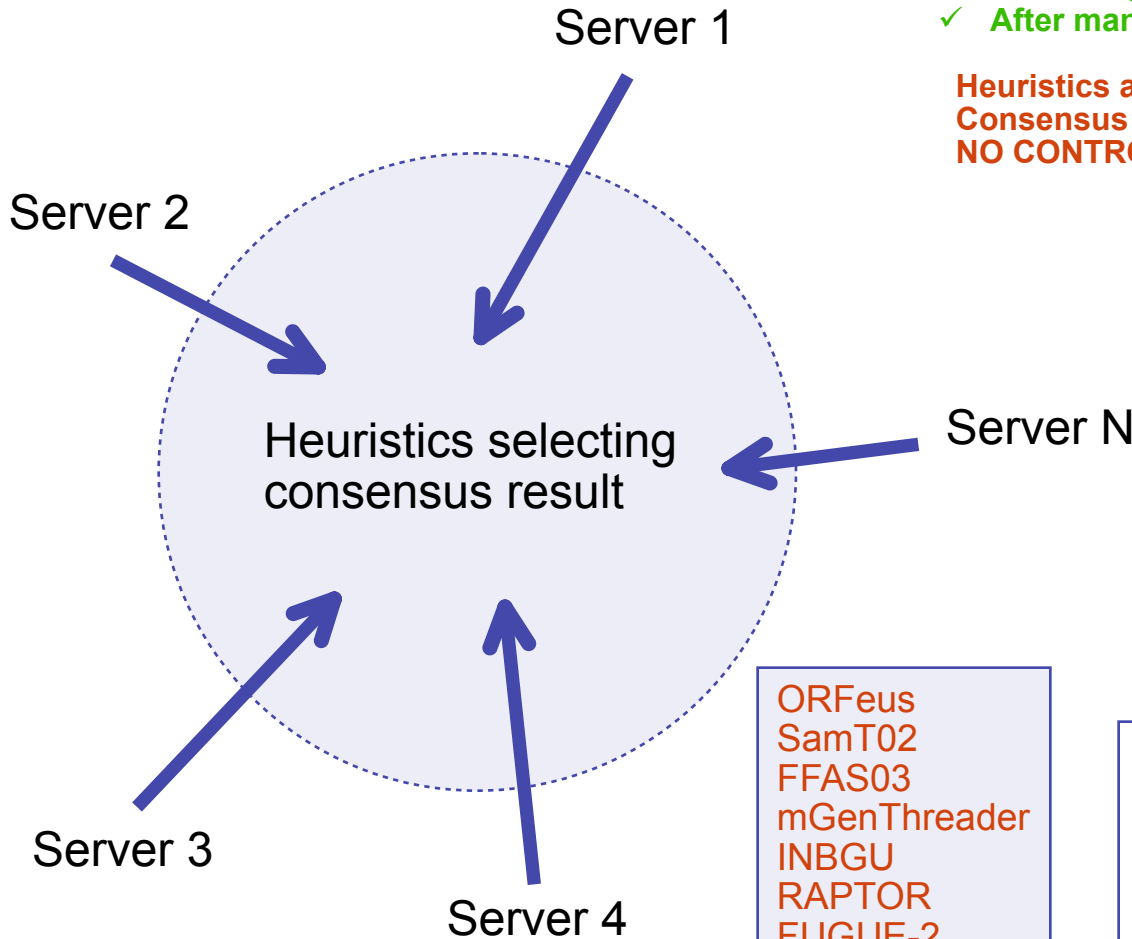
[Here](#) is a summary of how it works.

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

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

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

Meta-Servers (3D-Jury)



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

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

Cell, Vol. 113, 791-792, June 13, 2003, Copyright ©2003 by Cell Press

Letter to the Editor

mRNA Cap-1 Methyltransferase in the SARS Genome

The 3D jury system has predicted the methyltransferase fold for the nsp13 protein of the SARS coronavirus. Based on the conservation of a characteristic tetrad of residues, the mRNA cap-1 methyltransferase function has been assigned to this protein, which has potential implications for antiviral therapy.

The latest outbreak of the severe acute respiratory syndrome (SARS) epidemic has led to thousands of potentially lethally infected patients and hundreds of deaths. These numbers are likely to rise, and the spreading disease is already causing major medical and economical concerns. Meanwhile, the SARS coronavirus identified as the pathogen responsible for the disaster has been isolated, and its genome sequenced (Marrs et al., 2002; Rota et al., 2003).

We have applied the 3D jury meta predictor (Ginalski et al., 2002) to annotate the structure and function of proteins encoded by the viral positive-strand ssRNA. Novel fold recognition methods utilize the global network of independent structure prediction servers. Detection of patterns of structural similarity between diverse models is used to consistently select the correct fold from a set of borderline predictions. Such methods made a dramatic impact on the last critical assessment of protein structure prediction (CASP-5 experiment) conducted in the summer of 2002. One of the most striking successes was the prediction of the correct fold of the nsp13 protein of the SARS coronavirus. This protein is a methyltransferase fold domain located in the nt 7200 amino acid large (re 1). Standard sequence RFI-BLAST or RFI-BLAST domain database search to assign any function to a protein is not possible for the SARS coronavirus, because numerous viruses before were not in the last universal set of sequences (S. Kikuchi et al., 2003). The identity confirmed by the presence of the conserved K-C-K-E essential residues is found indispensable for viruses (Bach et al., 1995; Kikuchi et al., 2002) and represents a novel fold. Nevertheless, dicyclic may fail to suppress formation seems to be less ap-0 (m)pppN formation (Guarnita, 2003). The enzyme in the genome would suggest that the virus also requires the AdoMet-dependent cap-0 methyltransferase. Both functions can be inhibited by carbocyclic analogs of adenosine, such as Neplanodin A or 3-deazaadenosine A, which interfere with the AdoMet-AdoMet metabolism of the host cell (De Clercq, 1998; Bray et al., 2002). Those compounds could complement other therapeutic strategies aimed at blocking enzymatic functions such as the RNA-dependent RNA polymerase, the protease, or the helicase encoded by the SARS virus.

Marcin von Grothuss, Lucjan S. Wywiץ, and Leszek Rychlewski*
Bioinformatics Institute
Limanowskiego 24A
60-744 Poznan
Poland

*Correspondence: leszek@bioinformatics.pl

SARS VNTTLCQYLN
1E36 LVSTVAVLQKAY
SARS NQMLLILIMD
1E36 QSTVPTVYVQ
SARS STATLSTREEM
1E36 SARFVVLNRPVTS
SARS STATLSTREEM
1E24 YLFFVAVVQV

Glu202
Lys46
Asp30
AdoMet

LETTERS

How Unique Is the Rice Transcriptome?

IN THE REPORT "COLLECTION, MAPPING, AND ANNOTATION OF OVER 28,000 cDNA CLONES FROM *JAPONICA* RICE" (S. Kikuchi et al., 18 July, p. 376), the Rice Full-Length cDNA Project Team provides a detailed description of the rice transcriptome. The authors claim that 36% of the tested rice transcripts are not

Figure 1. 3D Model of the nsp13 Domain of the SARS Coronavirus p13a Polyprotein. This model is based on the assignment (Ginalski and Rychlewski, 2003) of the methyltransferase fold of the nsp13 protein (P. Bouchard et al., 2003). While other templates (Pao et al., 2003) related marginally higher 30% sequence, the selected template had the lowest number of residues and disulfide bonds. Side chains of the conserved tetrad of residues (K-C-K-E) essential for cap-0 methylation and the docked AdoMet cofactor are shown. Four blocks of aligned motifs containing the conserved, function-specific residues are shown in upper right corner.

Meta-Servers (3D-Jury)

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

BIOINFO.PL:META **Meta Server Job List** [ABOUT] [SERVERS] [BENCHMARKS] [STATUS]

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

Your E-mail:
Target Name:
Amino Acid Sequence only (in one letter code):

Reset Clear Format Submit

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

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

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

Complex gap penalty functions

MODELLER SALIGN (ALIGN2D)

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

Profile-Profile alignments

MODELLER SALIGN ('PROFILE')

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

Experiment (in silico)

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

Methods: Reference set

CE alignments with

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

387

Filter: MAMMOTH alignments with

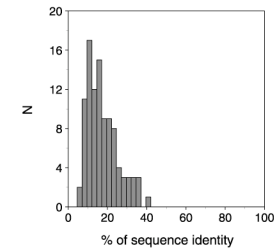
- > 50% EqPos

300

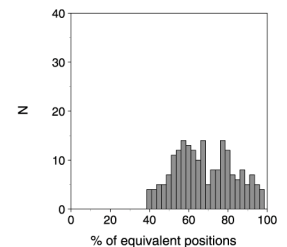
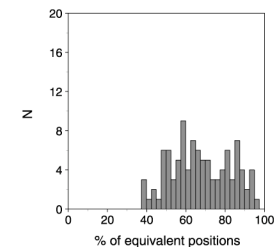
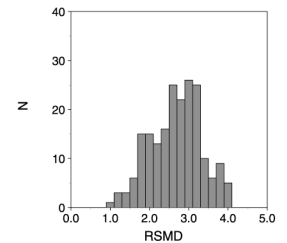
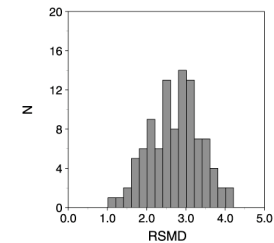
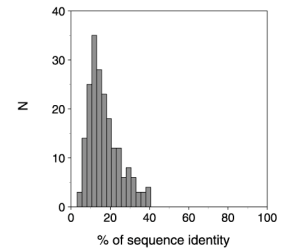
100 Training set

200 Testing set

A) Training Set



B) Testing Set



Methods: Evaluated methods

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

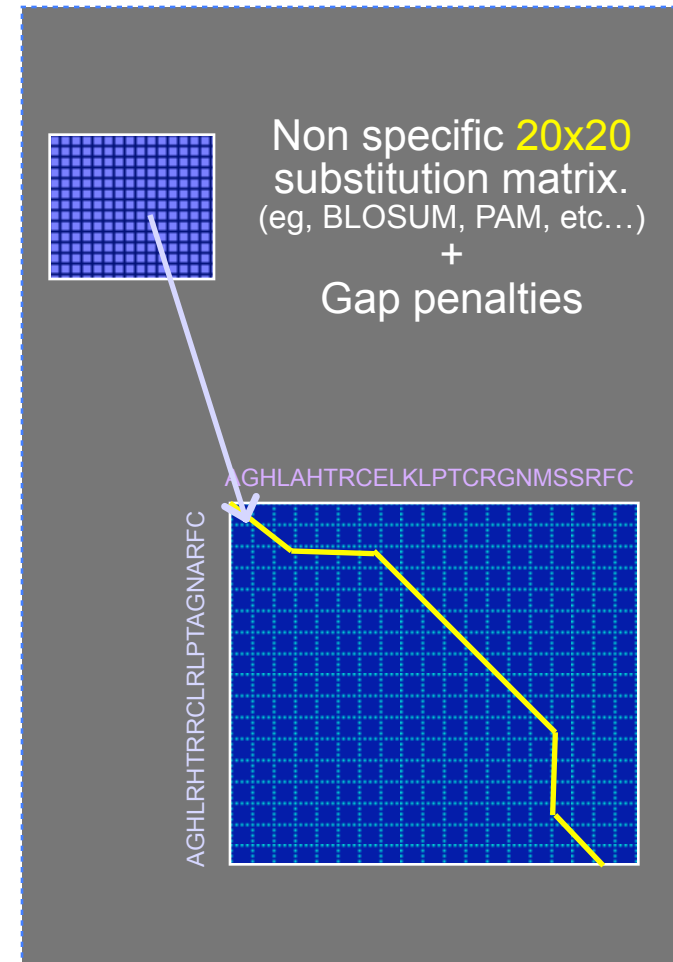
BLAST2SEQ: Local method

Prof.-Seq.

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

Prof.-Prof.

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



Methods: Evaluated methods

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

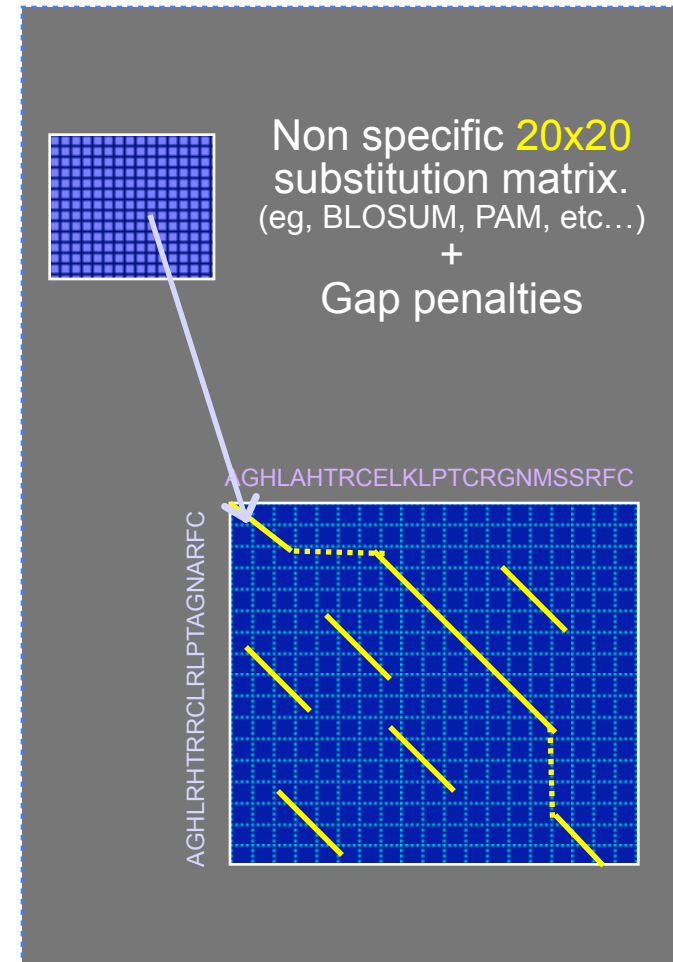
BLAST2SEQ: Local method

Prof.-Seq.

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

Prof.-Prof.

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



Methods: Evaluated methods

Sequence A: AGHLAHTRCELKLPCTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

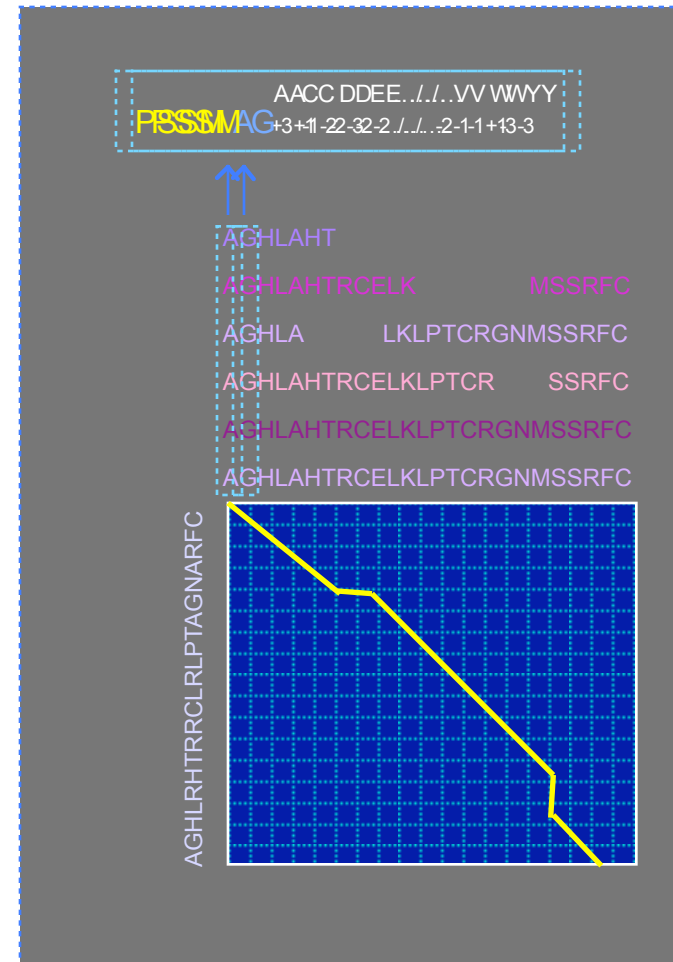
BLAST2SEQ: Local method

Prof.-Seq.

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

Prof.-Prof.

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



Methods. Evaluated methods.

Sequence A: AGHLAHTRCELKLPCTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

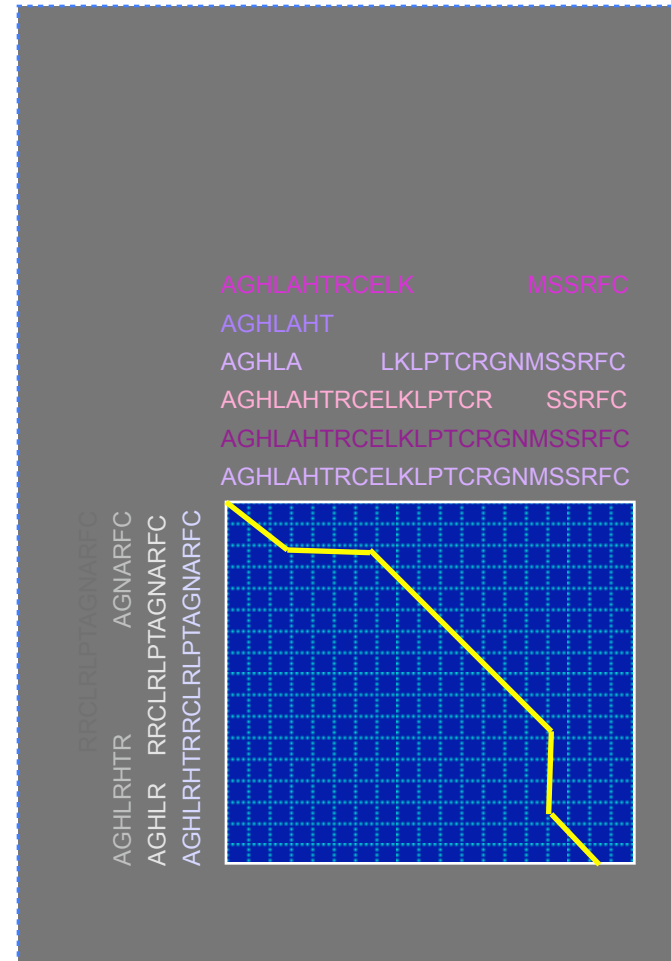
BLAST2SEQ: Local method

Prof.-Seq.

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

Prof.-Prof.

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



Methods. SALIGN.

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

Methods: Coverage and accuracy

High coverage



Low accuracy

High accuracy

Low coverage

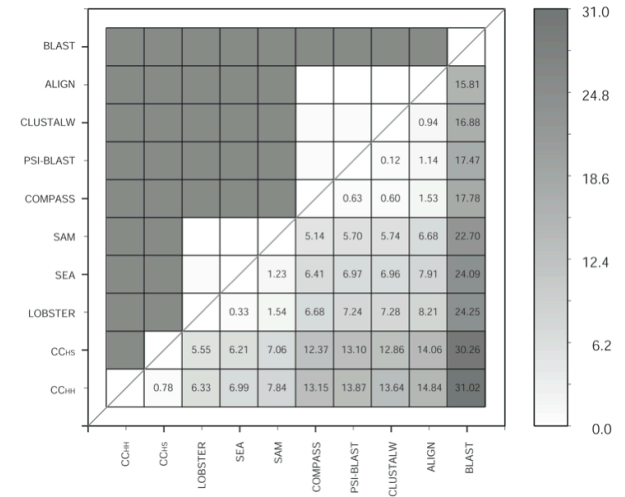
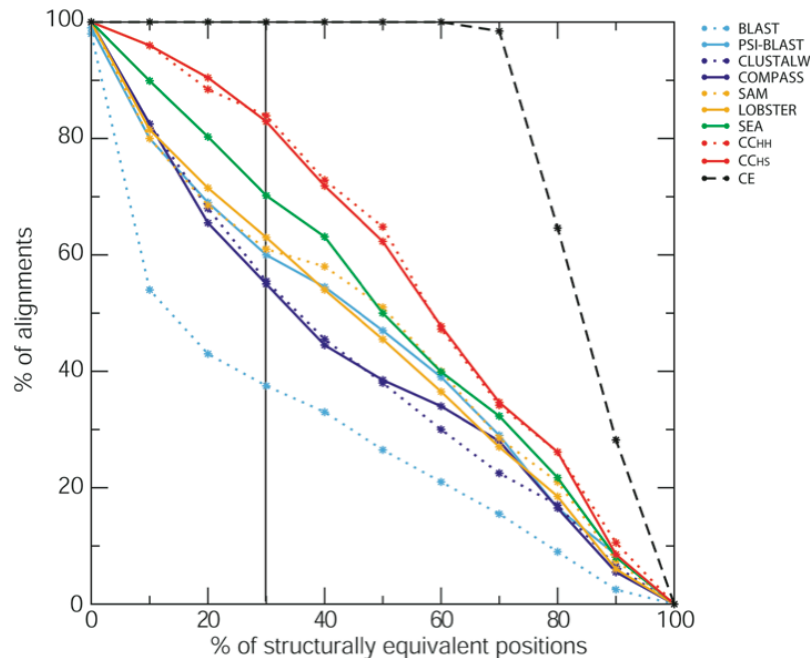
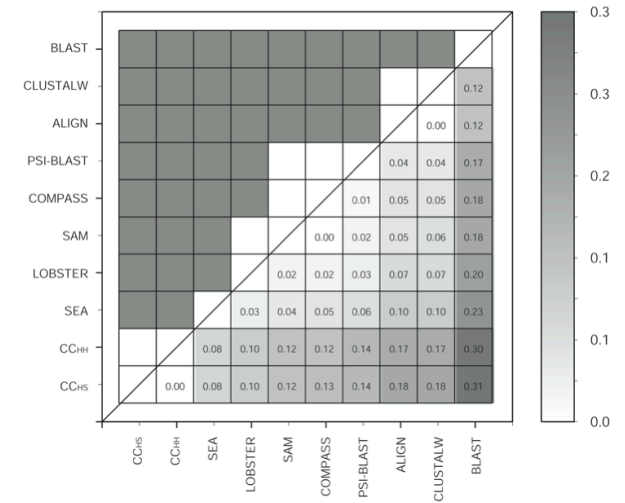


High accuracy

Low accuracy

Results: Comparison of alignment dependent measures

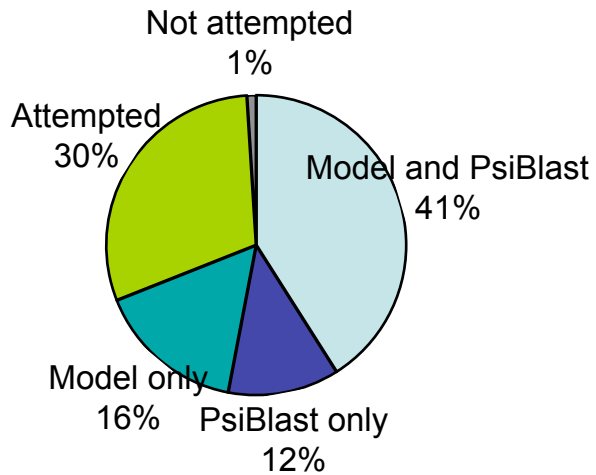
Method	CE overlap [%]	Shift score	RMSD [Å]	Structure overlap [%]
CE	100 ± 0	1.00 ± 0.00	2.7 ± 0.6	59.8 ± 12.9
BLAST	26 ± 29	0.32 ± 0.33	5.6 ± 3.7	20.6 ± 23.7
PSI-BLAST	43 ± 31	0.48 ± 0.35	6.5 ± 3.9	30.3 ± 24.9
SAM	48 ± 26	0.50 ± 0.34	9.2 ± 4.7	28.9 ± 24.8
LOBSTER	50 ± 27	0.51 ± 0.32	9.1 ± 4.9	31.1 ±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



Results. Turn over.

Mycoplasma genitalium MODPIPE Models

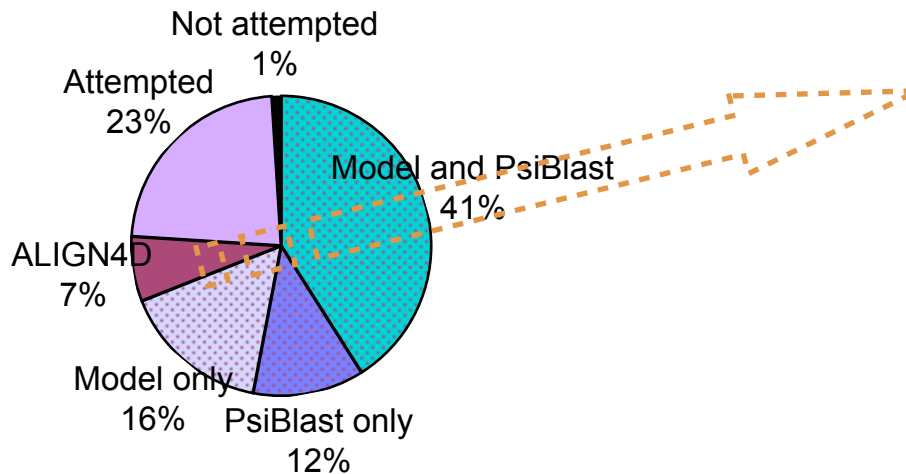
Number of ORFs	479
Average ORF length	364



Results. Turn over.

Mycoplasma genitalium MODPIPE Models

Number of ORFs	479
Average ORF length	364



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

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

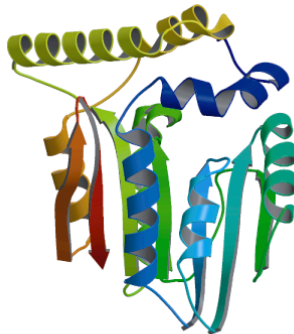
Examples: T0092 model

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

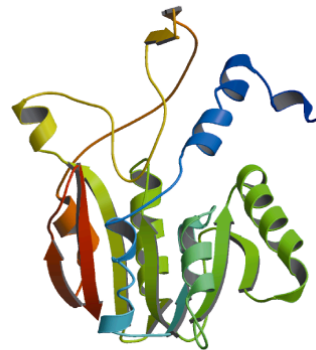
Method	RMSD Å	% of EqPos
ALIGN4D CC _{PP}	5.9	67.84
PSI-BLAST	4.9	31.72
Best predictions at CASP4	6.0	65.20

Data from CASP4, Asilomar, CA, December 2000.

B) Target T0092



X-Ray structure



ALIGN4D (CC_{PP}) model



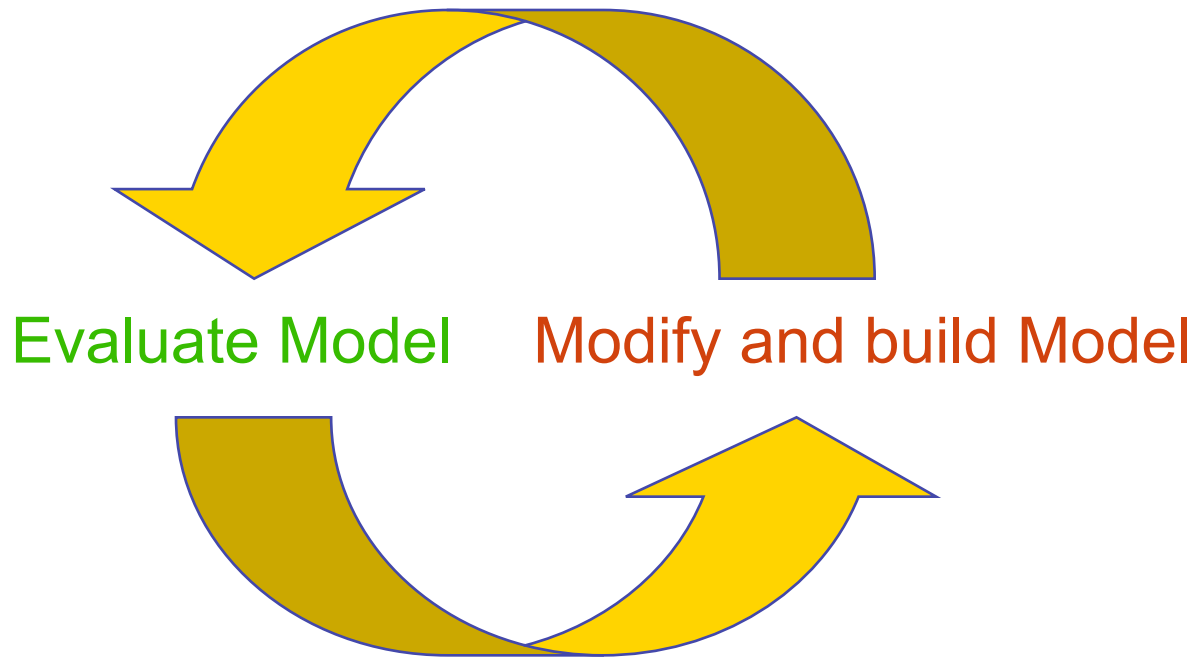
Psi-Blast model

Iterative process

MOULDER

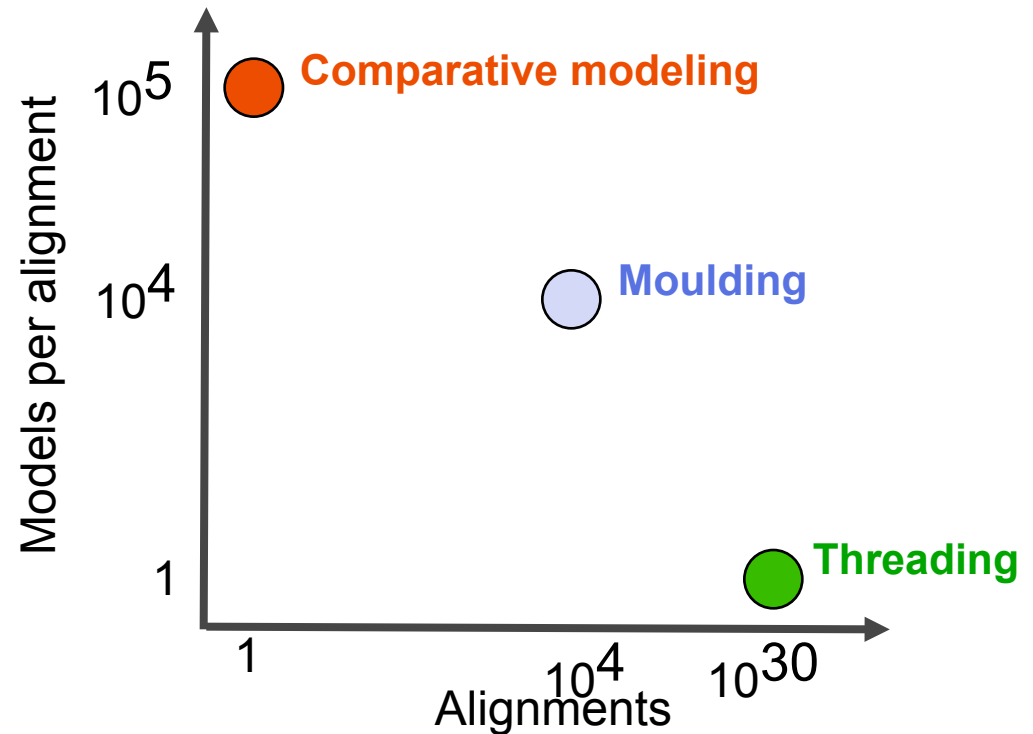
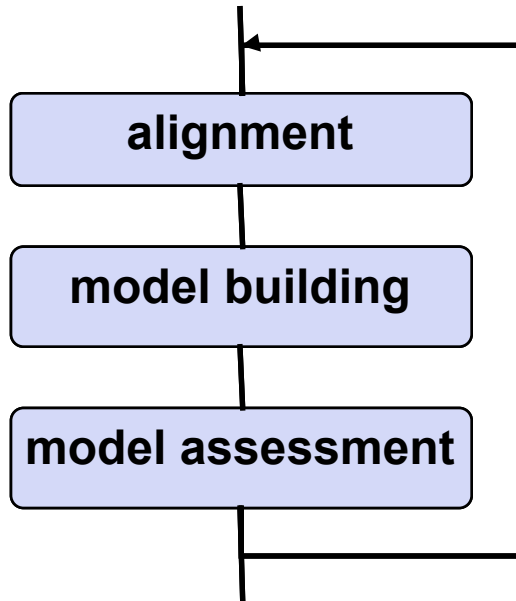
John, B. and Sali, A. (2003) Nucleic Acids Research. **31**:1982-1992

Iterative process... better models(?)



Moulding: iterative alignment, model building, model **assessment**

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



Iterative process... MOULDER

more in model evaluation

