

BMI-206

Structure-Structure comparisons Sequence-Structure comparisons

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

February 3rd, 2005

How to use this lectures

- Ask!
- Outline
 - Basic introduction
 - Theory (representation-scoring-optimization)
 - Available programs
 - Application
- Assignment
 - *The POM152 sequence. Modeling exercise.*

Structure-Structure comparison

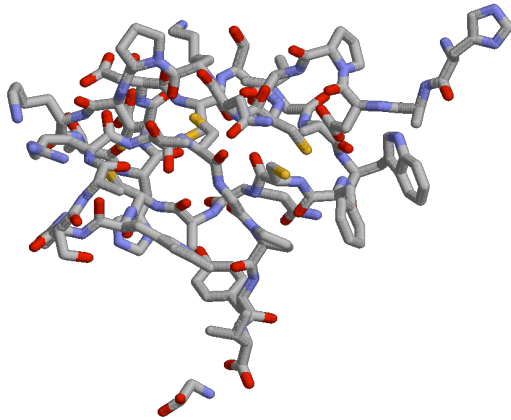
- Outline
 - Before we start...
 - Some theory
 - Coverage .vs. Accuracy
 - How can we compare structures...
 - SALIGN (properties comparison)
 - VAST (vector alignment)
 - CE (local heuristic comparison)
 - MAMMOTH (vector alignment)
 - How we classify the structural space...
 - SCOP (manual)
 - CATH (semi-automatic)
 - DBAli (fully automatic and comprehensive)

Structure-Structure alignments

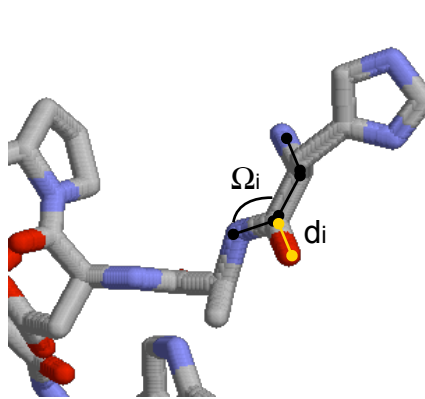
As any other bioinformatics problem...

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

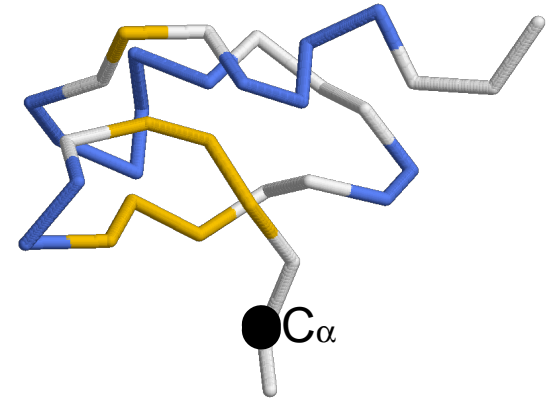
Representation Structures



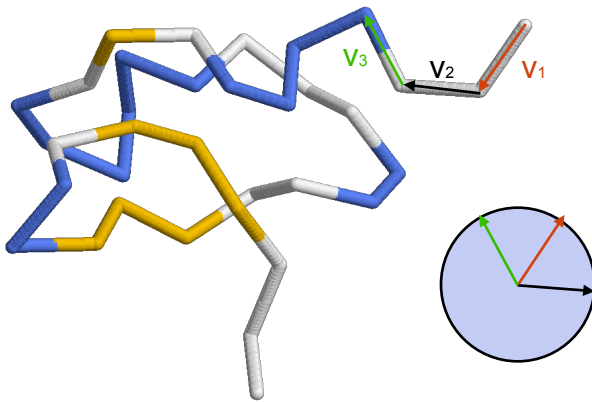
All atoms and coordinates



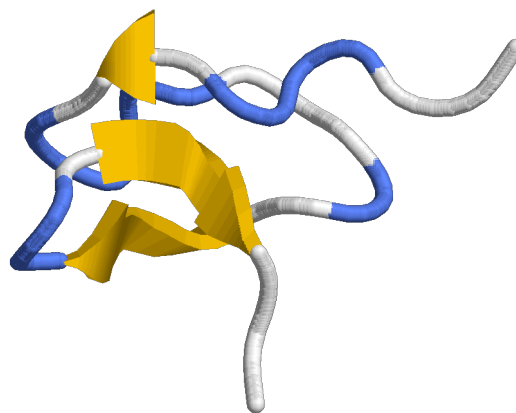
Dihedral space or distance space



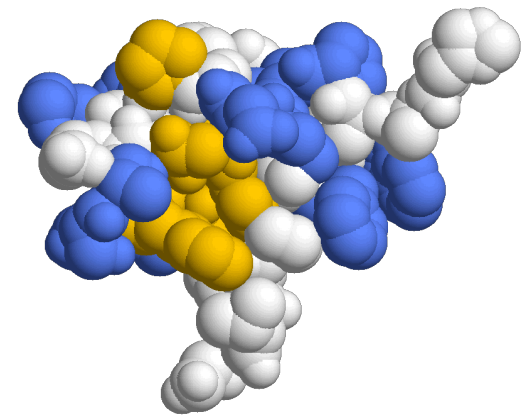
Reduced atom representation



Vector representation



Secondary Structure



Accessible surface (and others)

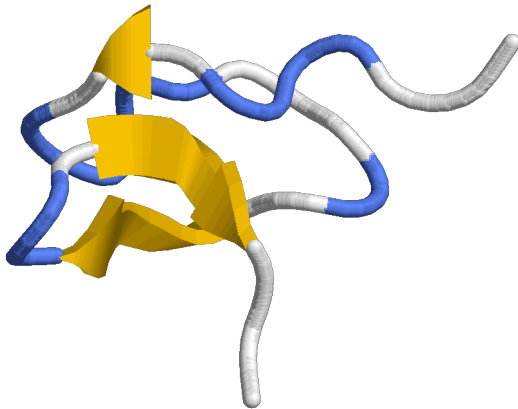
Scoring Raw scores

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-3	-1	-1	-1	-2	-2	-2
S	-1	4	1	-1	1	0	1	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-2	-3
T	-1	1	4	1	-1	1	0	1	0	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-3
P	-3	-1	1	7	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	4	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	6	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	6	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	6	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	5	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	5	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	5	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	5	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

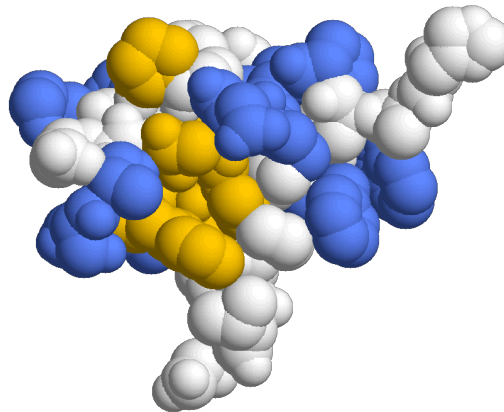
Aminoacid substitutions

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

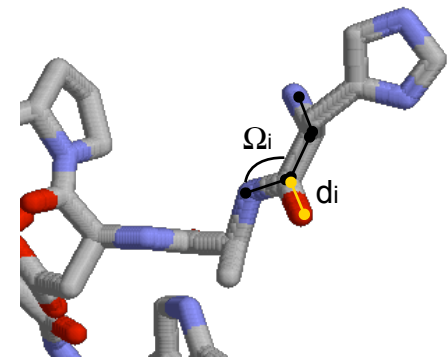
Root Mean Square Deviation



Secondary Structure (H,B,C)



Accessible surface (B,A [%])



Angles or distances

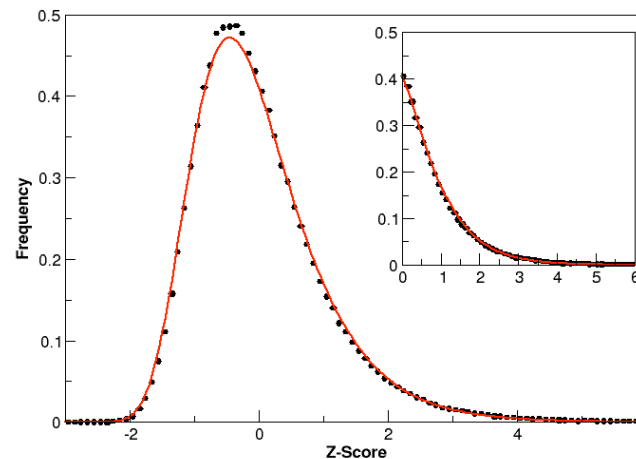
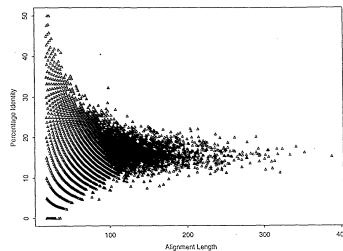
Scoring

Significance of an alignment (score)

remember Patsy's class

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

Empirical



Sometimes approximated by Z-score (normal distribution).

Analytic

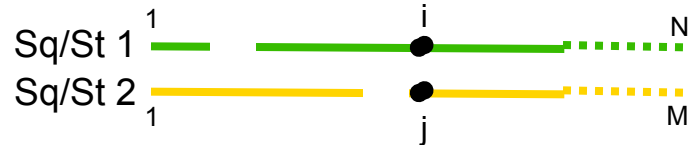
$$P(s) = e^{-\lambda (s-\mu)}$$

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

Optimizer

Global dynamic programming alignment

remember Patsy's class



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

$$D_{ij} = \min \begin{cases} D_{i,j-1} + \text{Score}(\ddot{A}, r_j) \\ D_{i-1,j-1} + \text{Score}(r_i, r_j) \\ D_{i-1,j} + \text{Score}(r_i, \ddot{A}) \\ 0 \end{cases}$$

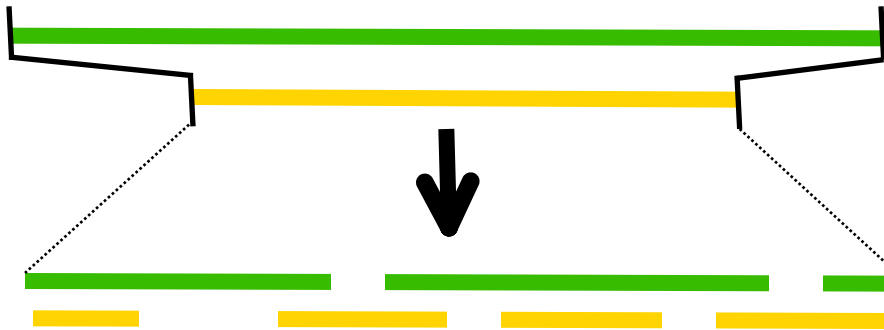
Best alignment score

Backtracking to get the best alignment

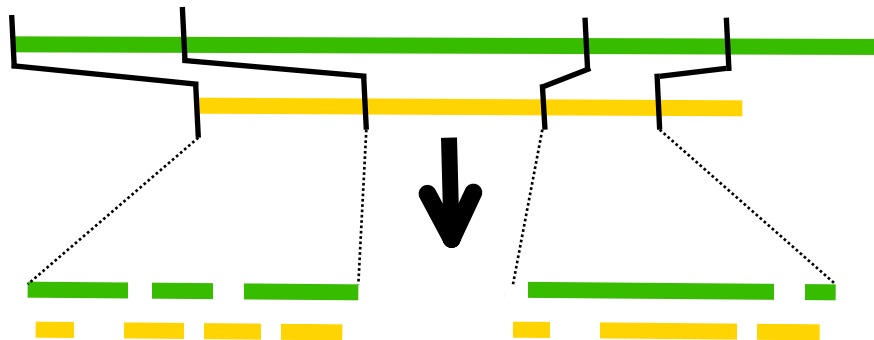
Optimizer

Global .vs. local alignment

remember Patsy's class



Global alignment



Local alignment

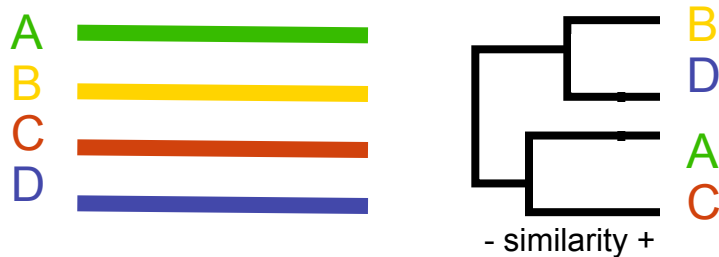
Optimizer

Multiple alignment

remember Patsy's class

Pairwise alignments

Example – 4 sequences A, B, C, D.



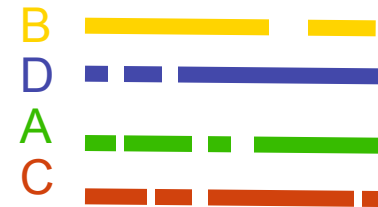
6 pairwise comparisons
then cluster analysis

Multiple alignments

Following the tree from step 1



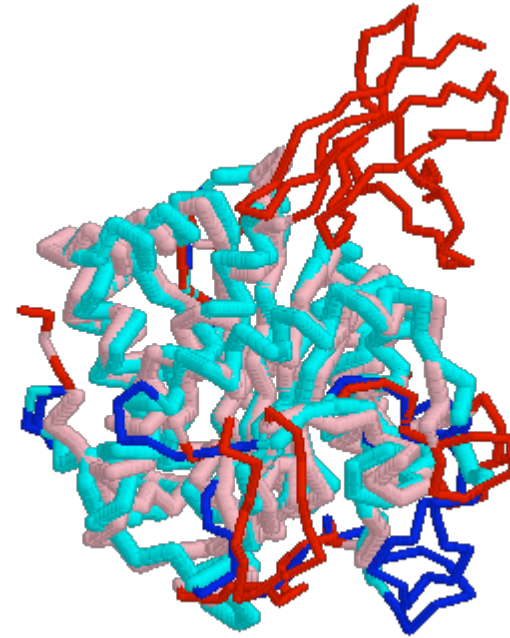
Align B-D with A-C



Coverage .vs. Accuracy



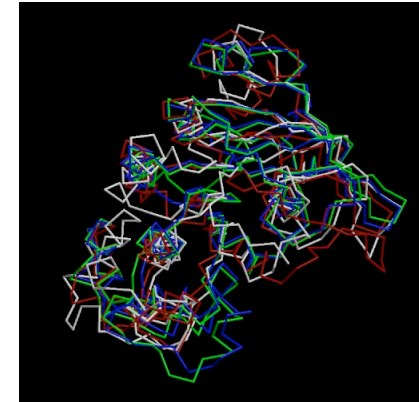
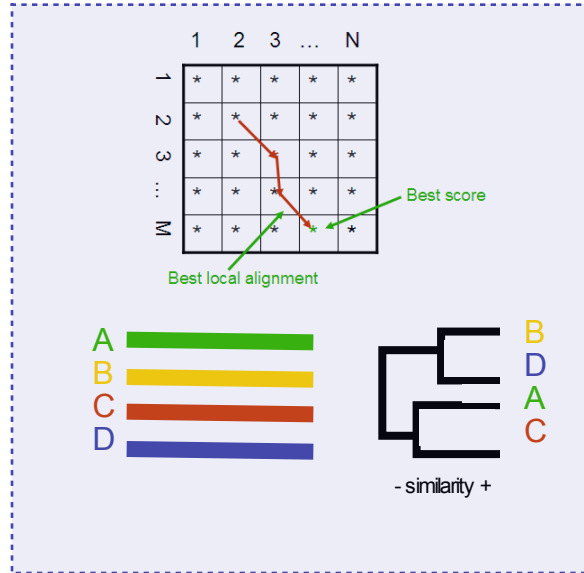
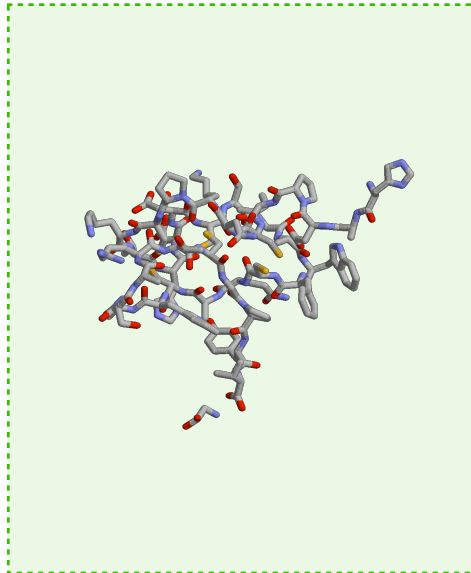
Coverage ~90% $C\alpha$



Coverage ~75% $C\alpha$

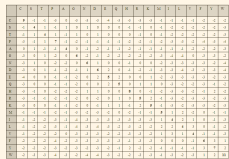
Same RMSD ~ 2.5Å

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

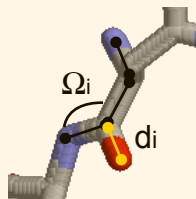


- ✓ Uses all available structural information
- ✓ Provides the optimal alignment

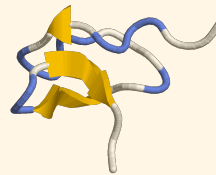
Computationally expensive



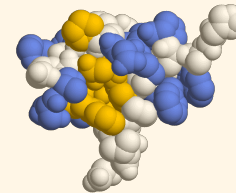
$R_{i,j}$



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



$S_{i,j}$



$B_{i,j}$

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

$I_{i,j}$

$$\text{Score}_{i,j} = w_1 * R_{i,j} + w_2 * D_{i(a),j(a)} + w_3 * S_{i,j} + w_4 * B_{i,j} + w_5 * I_{i,j} + w_6 * X_{i,j}$$

Structural alignment by properties conservation (SALIGN-MODELLER)

<http://alto.compbio.ucsf.edu/salign-cgi/index.cgi>



The screenshot shows a web browser window titled "SALIGN Server" with the URL <http://alto.compbio.ucsf.edu/salign-cgi/index.cgi>. The page has a blue background and contains the following text and form elements:

SALIGN Multiple Structure/Sequence Alignment Server

SALIGN is a general alignment module of the modeling program MODELLER

The alignments are computed using dynamic programming, making use of several features of the protein sequences and structures

Users can either upload their own sequences/structures to align or choose structures from the PDB

sequences can either be pasted or uploaded as FASTA or PIR format alignment files

Paste sequence to align

Multiple sequences can be pasted by iteratively clicking 'upload' after every pasted sequence

Specify file to upload (PIR, FASTA, PDB, .zip or .tar.gz)

Multiple files can be uploaded by iteratively clicking 'upload' after every file uploaded

Uploaded files:

No files uploaded

Enter 4 letter code(s) to choose PDB structures

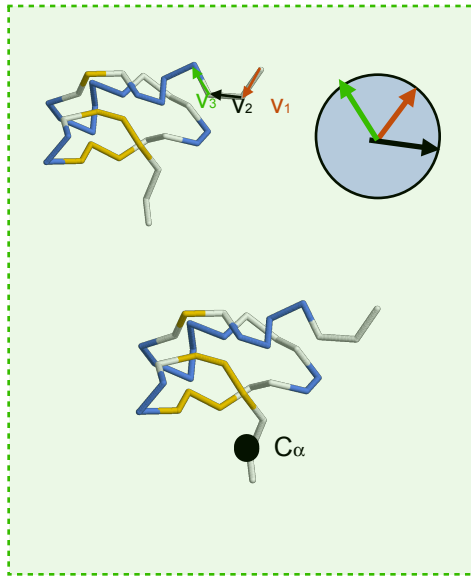
e-mail address, to receive results:

Reference:

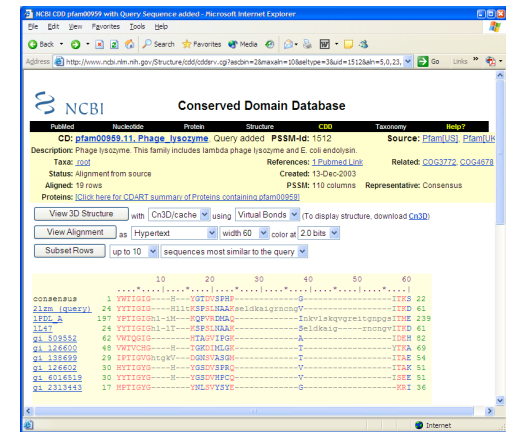
Madhusudhan, M.S., Marti-Renom, M.A., Eswar, N., Sali, A.

SALIGN - a multiple structure/sequence alignment tool, under preparation

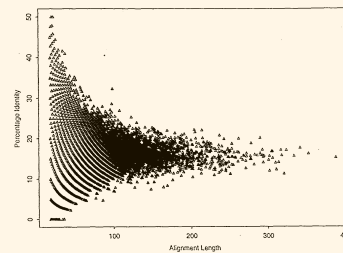
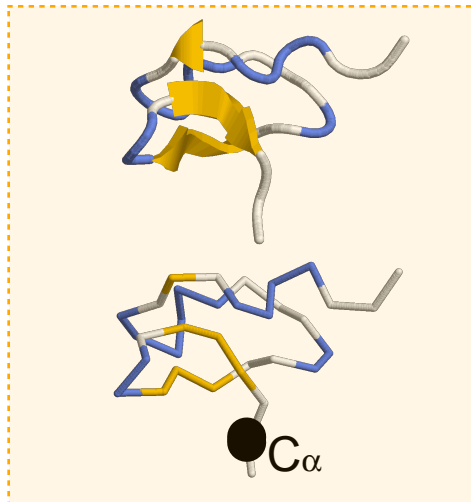
Vector Alignment Search Tool (VAST)



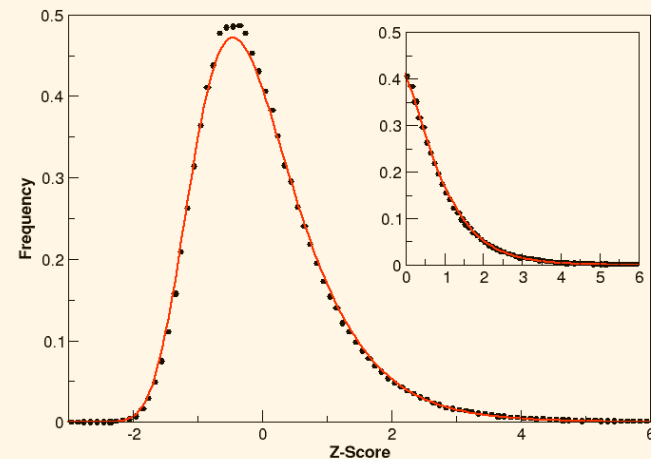
- Graph theory search of similar SSE
- Refining by Monte Carlo at all atom resolution



✓ Good scoring system with significance
Reduces the protein representation

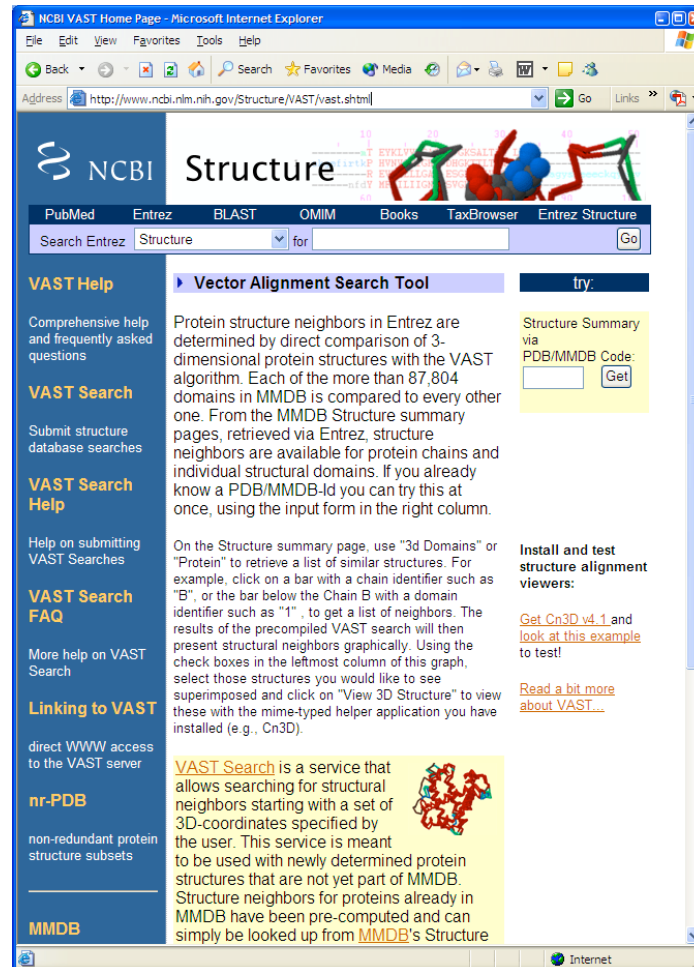


$$\text{RMSD} = \sqrt{\sum (x_i - \bar{x})^2}$$



Vector Alignment Search Tool (VAST)

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



NCBI VAST Home Page - Microsoft Internet Explorer

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

NCBI Structure

PubMed Entrez BLAST OMIM Books TaxBrowser Entrez Structure

Search Entrez Structure for Go

VAST Help

Comprehensive help and frequently asked questions

VAST Search

Submit structure database searches

VAST Search Help

Help on submitting VAST Searches

VAST Search FAQ

More help on VAST Search

Linking to VAST

direct WWW access to the VAST server

nr-PDB

non-redundant protein structure subsets

MMDB

Vector Alignment Search Tool

try:

Protein structure neighbors in Entrez are determined by direct comparison of 3-dimensional protein structures with the VAST algorithm. Each of the more than 87,804 domains in MMDB is compared to every other one. From the MMDB Structure summary pages, retrieved via Entrez, structure neighbors are available for protein chains and individual structural domains. If you already know a PDB/MMDB-Id you can try this at once, using the input form in the right column.

Structure Summary via PDB/MMDB Code: Get

On the Structure summary page, use "3d Domains" or "Protein" to retrieve a list of similar structures. For example, click on a bar with a chain identifier such as "B", or the bar below the Chain B with a domain identifier such as "1", to get a list of neighbors. The results of the precompiled VAST search will then present structural neighbors graphically. Using the check boxes in the leftmost column of this graph, select those structures you would like to see superimposed and click on "View 3D Structure" to view these with the mime-typed helper application you have installed (e.g., Cn3D).

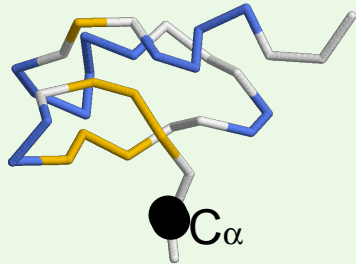
Install and test structure alignment viewers:

[Get Cn3D v4.1 and look at this example](#) to test!

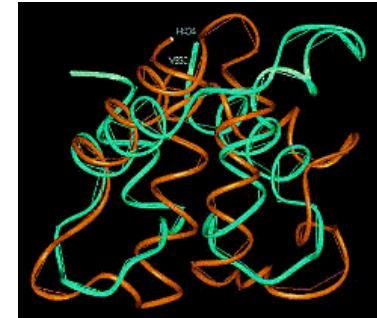
[Read a bit more about VAST...](#)

VAST Search is a service that allows searching for structural neighbors starting with a set of 3D-coordinates specified by the user. This service is meant to be used with newly determined protein structures that are not yet part of MMDB. Structure neighbors for proteins already in MMDB have been pre-computed and can simply be looked up from MMDB's Structure

Incremental combinatorial extension (CE)

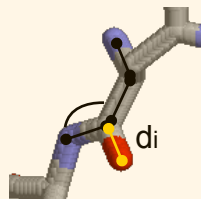


- Exhaustive combination of fragments
- Longest combination of AFPs
- Heuristic similar to PSI-BLAST



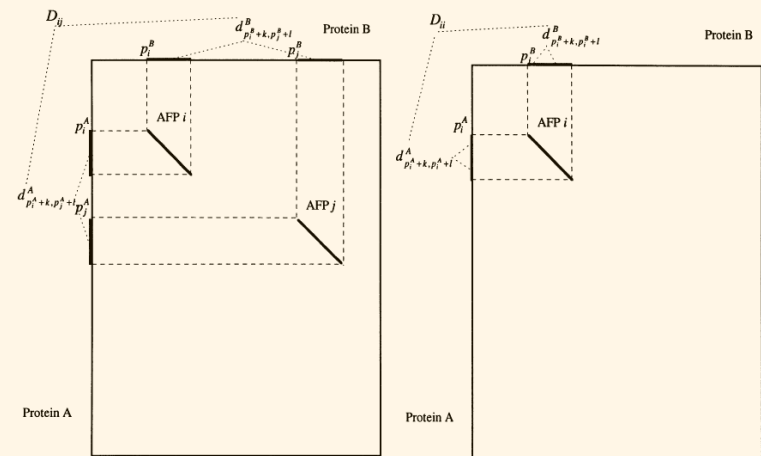
- ✓ **FAST!**
- ✓ **Good quality of local alignments**

Complicated scoring and heuristics



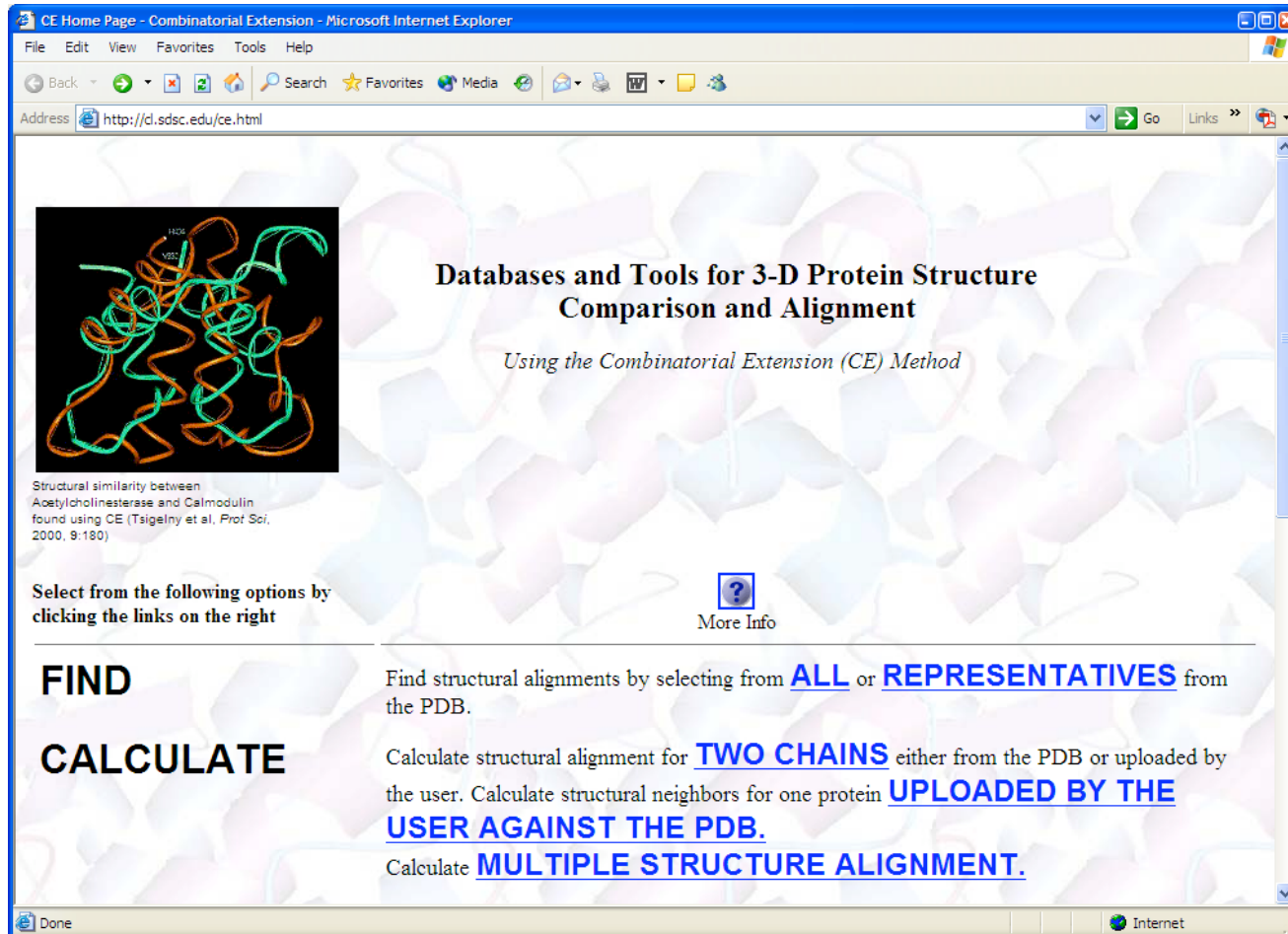
8 residues peptides

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



Incremental combinatorial extension (CE)

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



CE Home Page - Combinatorial Extension - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Search Favorites Media

Address <http://cl.sdsc.edu/ce.html> Go Links



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

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

Using the Combinatorial Extension (CE) Method

[More Info](#)

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

FIND

Find structural alignments by selecting from [ALL](#) or [REPRESENTATIVES](#) from the PDB.

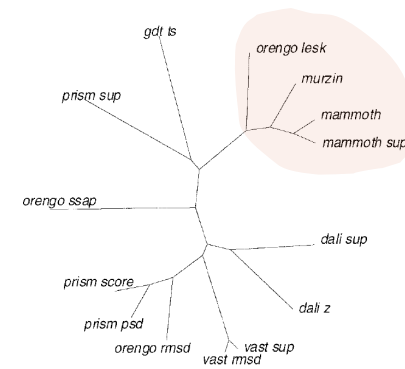
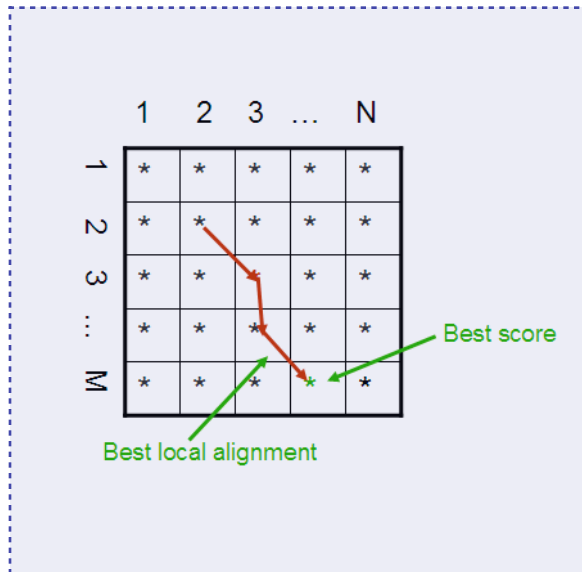
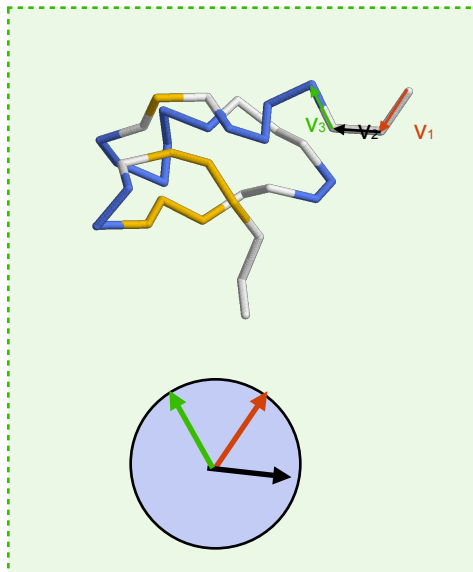
CALCULATE

Calculate structural alignment for [TWO CHAINS](#) either from the PDB or uploaded by the user. Calculate structural neighbors for one protein [UPLOADED BY THE USER AGAINST THE PDB](#).

Calculate [MULTIPLE STRUCTURE ALIGNMENT](#).

Done Internet

Matching molecular models obtained from theory (MAMMOTH)

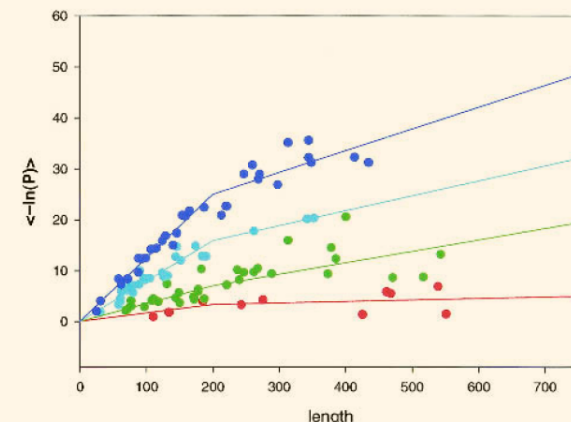
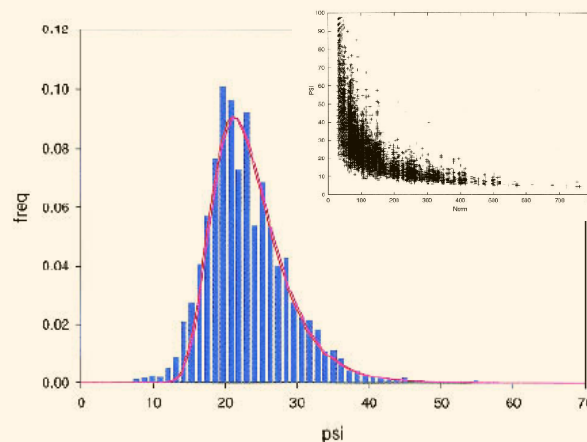


- ✓ VERY FAST!
- ✓ Good scoring system with significance

Reduces the protein representation

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

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



Matching molecular models obtained from theory (MAMMOTH)

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

Protein Structure Alignment Server - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media

Address <http://fulcrum.physbio.mssm.edu:8083/mammoth/> Go Links



MAMMOTH

Matching Molecular Models Obtained from Theory

Paste or browse your **PREDICTION** coordinates (PDB format):

Browse...

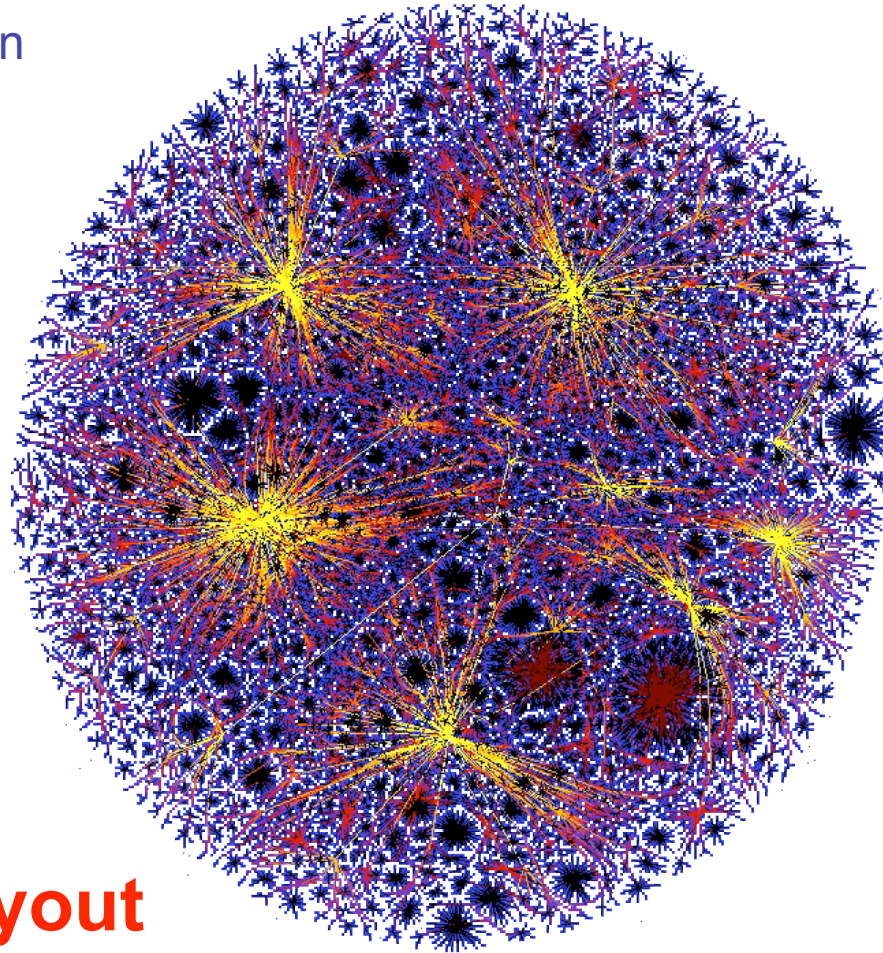
Paste or browse your **EXPERIMENT** coordinates (PDB format):

Browse...

Done Internet

Classification of the structural space

SCOP classification



Large Graph Layout

Alex Adai

Adai AT, Date SV, Wieland S, Marcotte EM. J Mol Biol. 2004 Jun 25;340(1):179-90

SCOP_{1.65} database

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

Structural Classification of Proteins

Welcome to SCOP: Structural Classification of Proteins. **1.65 release** (December 2003).
 20619 PDB Entries. 1 Literature Reference. 54745 Domains (excluding nucleic acids and theoretical models). Folds, superfamilies, and families [statistics here](#). [New folds](#) [superfamilies](#) [families](#). [List of obsolete entries and their replacements](#).

Authors: Alexey G. Murzin, Loredana Lo Conte, Antonina Andreeva, Dave Howorth, Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia. scop@mrc-lmb.cam.ac.uk

Reference: Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540. [\[PDF\]](#)

Major changes (stable identifiers, parseable files, extended searching and linking options, reclassified entries history) are described in: Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.* 30(1), 264-267. [\[PDF\]](#)

Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res.* 32:D226-D229.

Access methods

- Enter SCOP at the [top of the hierarchy](#)
- [Keyword search of SCOP entries](#)
- SCOP parseable files ([MRC site](#))
- Reclassified entries: [1.63-->1.65](#), previous releases ([MRC site](#))
- SCOP domain sequences and pdb-style coordinate files ([ASTRAL](#))
- Hidden Markov Model library for SCOP superfamilies ([SUPERFAMILY](#))
- [Online resources](#) of potential interest to SCOP users

SCOP [mirrors](#) around the world may speed your access.

News

- SCOP has been updated to include all PDB entries released up to 1 August 2003. See [folds](#), [superfamilies](#), and [families statistics](#).
- Several parts of the SCOP classification have been restructured, especially in this release and in the previous one. You can browse the subset of the classification affected by these changes in a SCOP-view form for modifications occurred between [1.63](#) and [1.65](#), or [previous releases](#). Changes appear as comments associated to [domain entries](#), with links to the revised classification. You can use the SCOP navigation buttons to move up in the hierarchy and to expand or collapse entries. The list of [obsolete entries and their replacements](#) is also available online.
- SCOP identifiers now appear explicitly in the web pages (in [squared brackets](#)).
- Links from a SCOP domain to the corresponding SWISSPROT and EC entries have been added (see the [u](#) icon). Thanks to Sameer Velankar and Phil McNeil from the EBI-MSD group and to Virginie Mittard from the EBI sequence database group for providing the most up-to-date map between PDB chains and SWISSPROT, EC identifiers.
- It is now possible to use SSM to search the up-to-date PDB archive using a SCOP domain entry (via the [t](#) icon) or to

- ✓ Largely recognized as “standard of gold”
- ✓ Manually classification
- ✓ Clear classification of structures in:
 CLASS
 FOLD
 SUPER-FAMILY
 FAMILY
- ✓ Some large number of tools already available

Manually classification
Not 100% up-to-date
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

CATH_{2.5.1} database

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

Uses FSSP for superimposition

- ✓ Recognized as “standard of gold”
- ✓ Semi-automatic classification
- ✓ Clear classification of structures in:
 - CLASS
 - ARCHITECTURE
 - TOPOLOGY
 - HOMOLOGOUS SUPERFAMILIES
- ✓ Some large number of tools already available
- ✓ Easy to navigate

Semi-automatic classification
Domain boundaries definition

CATH Protein Structure Classification

Version 2.5.1: Released January 2004

Dr. Frances M.G. Pearl, Dr. Ian Sillicoe, Dr. Mark Dibley, Prof. Janet Thornton, Prof. Christine A. Orengo

Options

- Browse or search the classification
- CATH statistics and release information
- General information on CATH
- CATH lists and ftp site
- DHS - Dictionary of Homologous Superfamilies. Summary of structural and functional features for CATH Homologous Superfamilies
- CATH File Formats (for FTP files)

Introduction

CATH is a novel hierarchical classification of protein domain structures, which clusters proteins at four major levels, Class(C), Architecture(A), Topology(T) and Homologous superfamily (H).

Class, derived from secondary structure content, is assigned for more than 90% of protein structures automatically. Architecture, which describes the gross orientation of secondary structures, independent of connectivities, is currently assigned manually. The topology level clusters structures according to their topological connections and numbers of secondary structures. The homologous superfamilies cluster proteins with highly similar structures and functions. The assignments of structures to topology families and homologous superfamilies are made by sequence and structure comparisons.

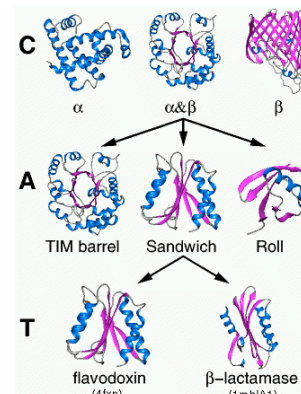
[Click here for a more detailed explanation](#)

Reference

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. (1997) CATH - A Hierarchic Classification of Protein Domain Structures. *Structure*. Vol 5. No 8. p.1093-1108.

Pearl, F.M.G., Lee, D., Bray, J.E., Sillicoe, I., Todd, A.E., Harrison, A.P., Thornton, J.M. and Orengo, C.A. (2000) Assigning genomic sequences to CATH *Nucleic Acids Research*. Vol 28. No 1. 277-282

Other CATH Contributors

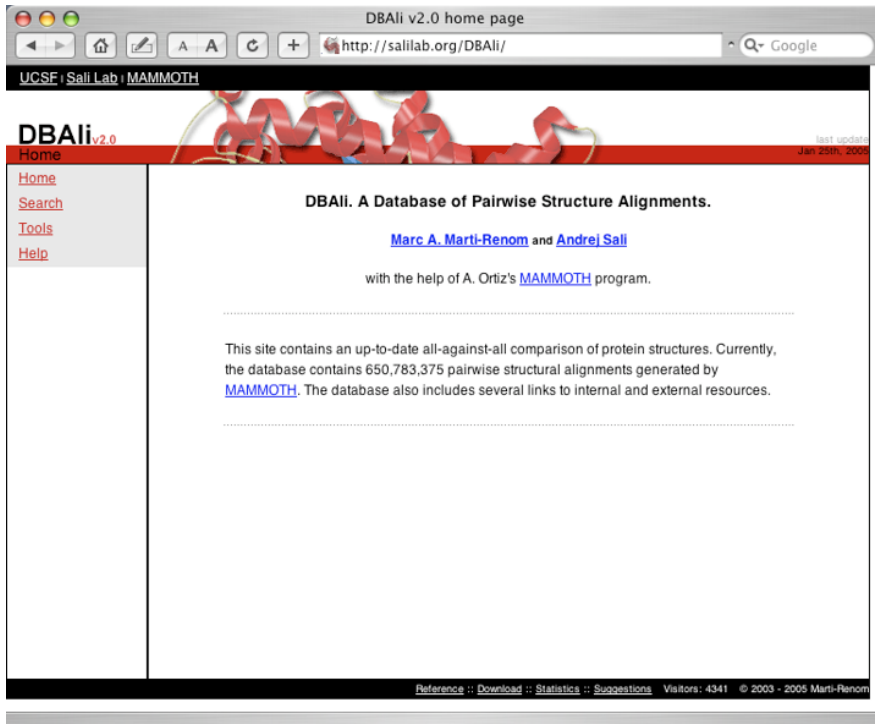


Version	2.5.1						
Date	28-01-2004						
	A	T	H	S	N	I	D
Mainly Alpha	5	227	428	948	1713	3946	10155
Mainly Beta	19	139	292	951	2344	5011	14259
Alpha Beta	12	368	648	2010	3631	8639	23025
Few Secondary Structures	1	86	91	114	225	378	952
Multi-domain chains	1	1053	1057	1071	2186	5801	12471
Preliminary single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

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

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:

January 25th, 2005

Number of chains in database:

60,656

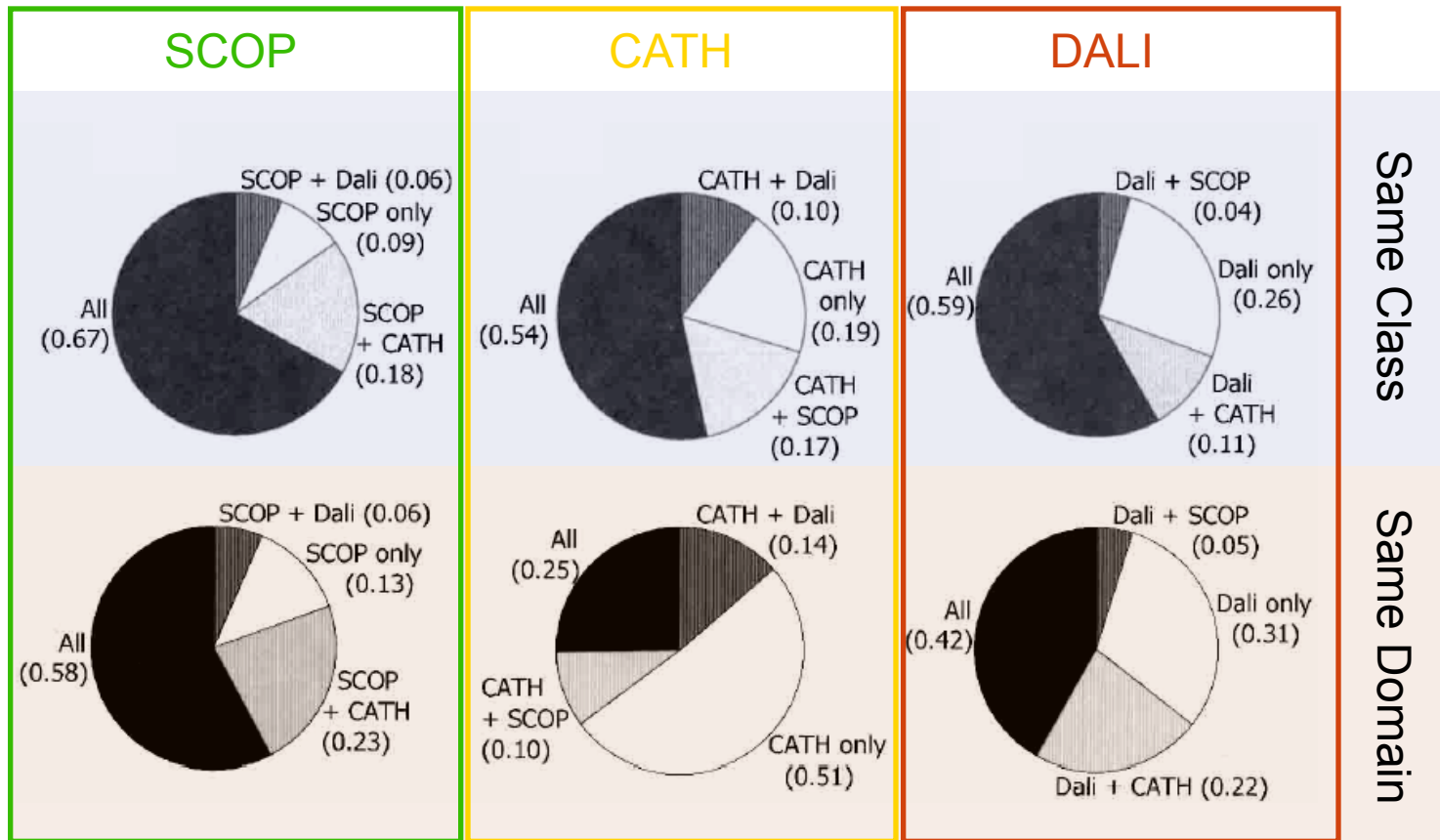
Number of structure-structure comparisons:

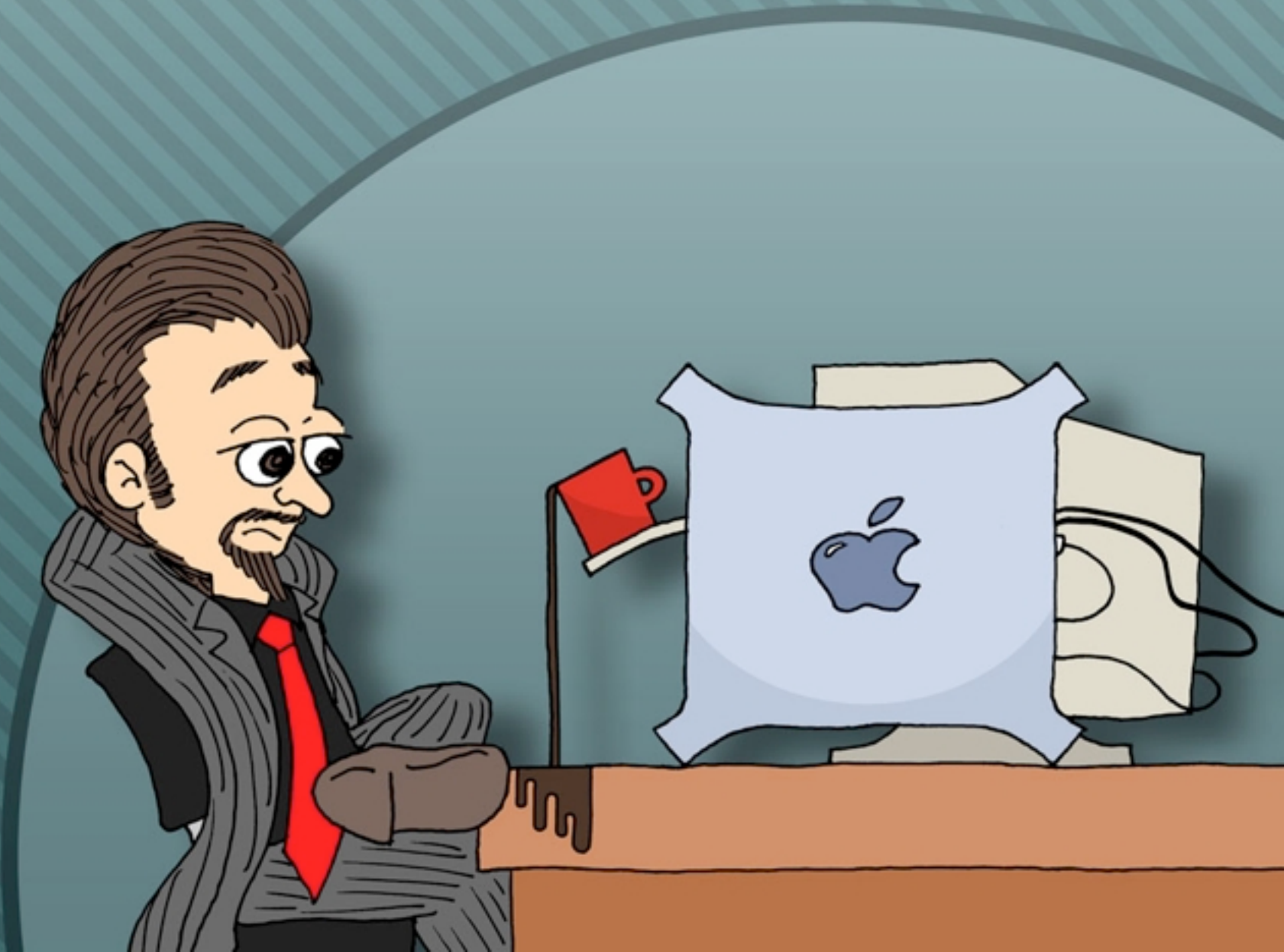
650,783,375

Classification of the structural space

Not an easy task!

Domain definition AND domain classification





Sequence-Structure comparison

- Outline
 - Before we start...
 - Some theory...
 - Domain boundaries
 - Structural predictions from sequence...
 - 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)

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/>
 - PredictProtein 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

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 (energy landscape)

- Simple (CPU calculation)

Types:

- Contact potential

- Distance potentials

- Surface potentials

Approximations/Limitations in PMFs

Database size.

PMF versus Energy (additive/higher order terms).

Reference state.

Physical origin.

Sequence-Structure alignments

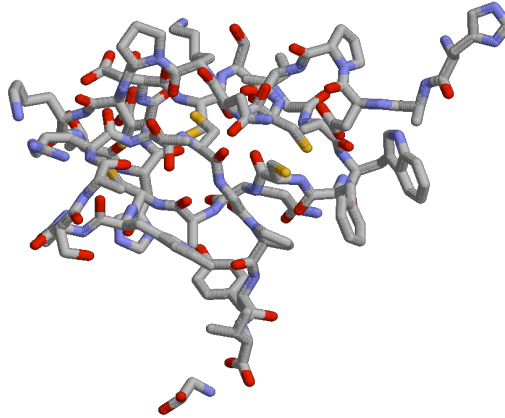
As any other bioinformatics problem...

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

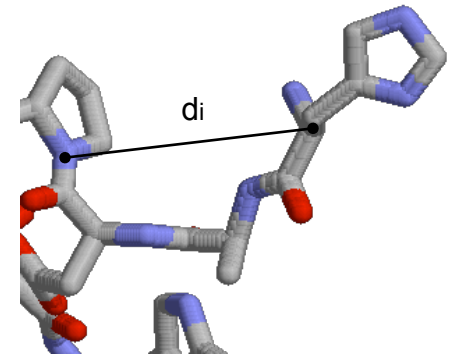
Representation Sequence/Structures

```
>gi42541361  
MDIRSVSSLRGLLCLPPSWPRR
```

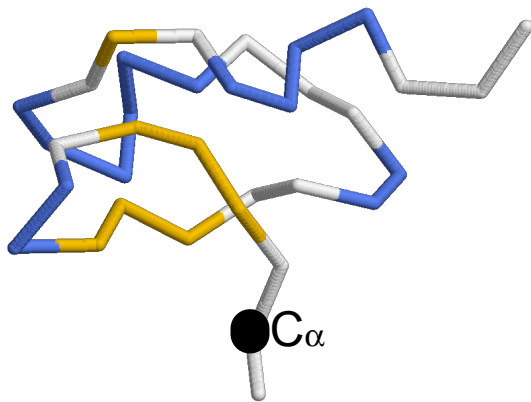
Primary sequence



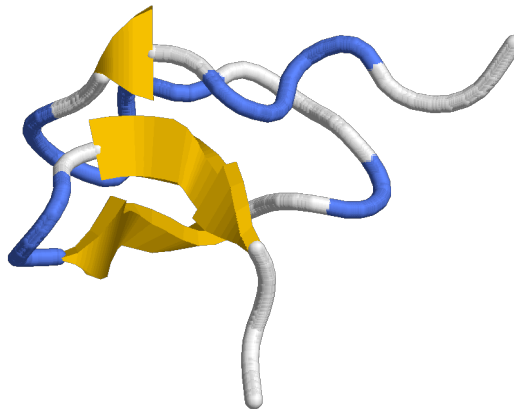
All atoms and coordinates



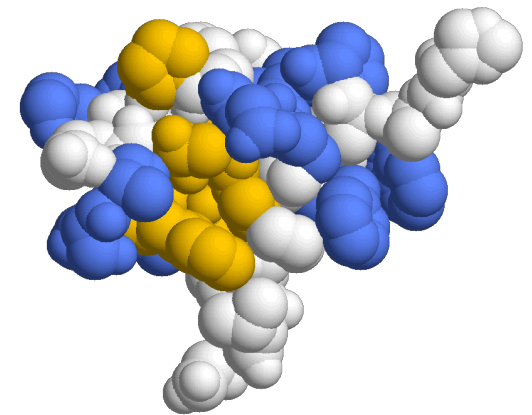
Distance space



Reduced atoms representation



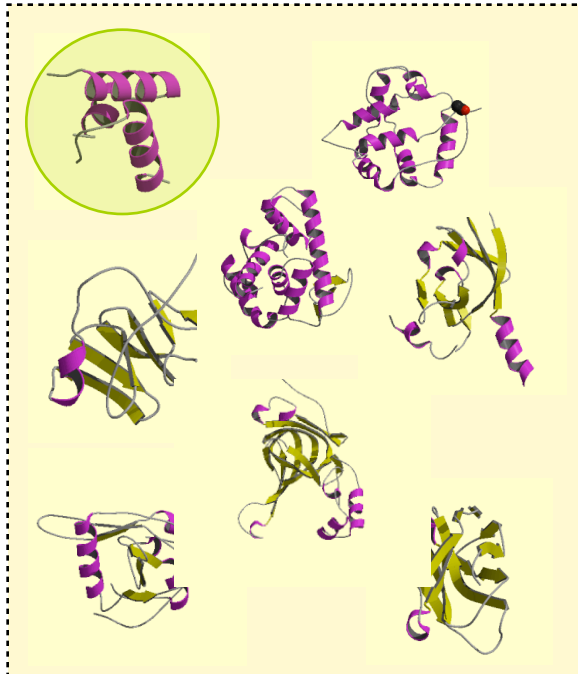
Secondary Structure



Accessible surface

Scoring Statistical Potentials (background)

Structural space



Sequence space

MKLLIVLTCISLCSCICTVVQRCASNKPHVLEDPCVKVQH
HLSVNQCVLLPQCCPKSCKICTHLISIEVVLT CRAVDKM
MHVNCVEQCSLQDCIKIAPRVLKTCILCVLKPCLT SVSH
VHLVQPTSCCCKKNCICHVEIRSLDILT KSVQLACL VPM
⋮
MQCCRVQKICDLLAVELCKLHISTPCKILCVVTSVPHN

Scoring Statistical Potential (inspiration)

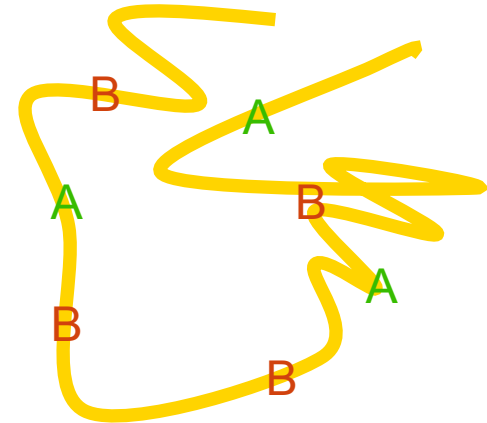
$$K = \frac{[AB]}{[A] \cdot [B]}$$

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

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

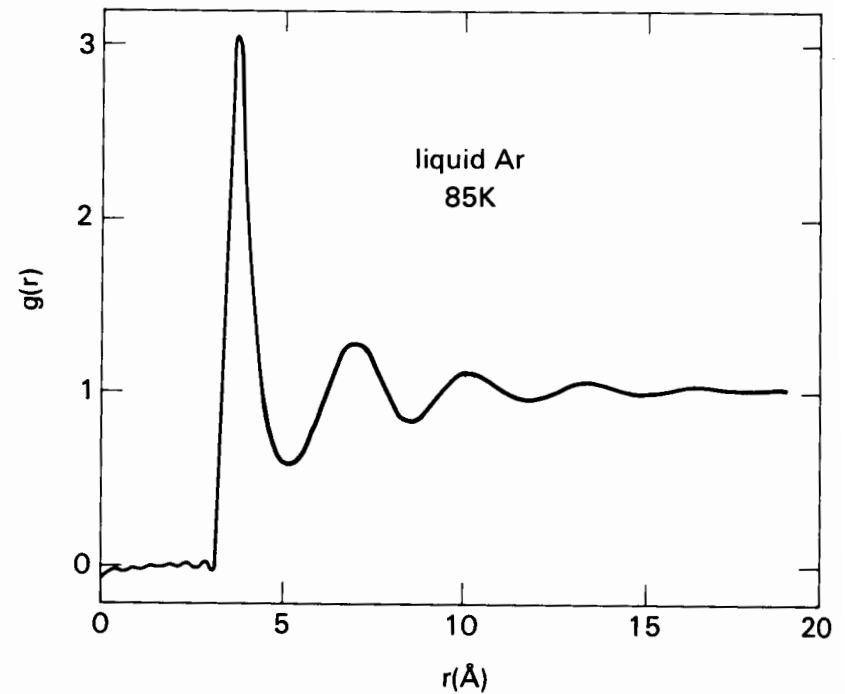
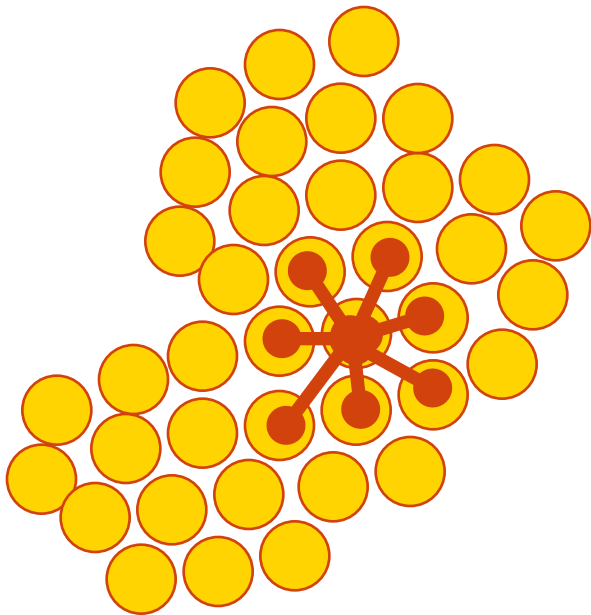
$$\Delta E = -kT \ln \left(\frac{N_1}{N_2} \right) \quad (1)$$

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



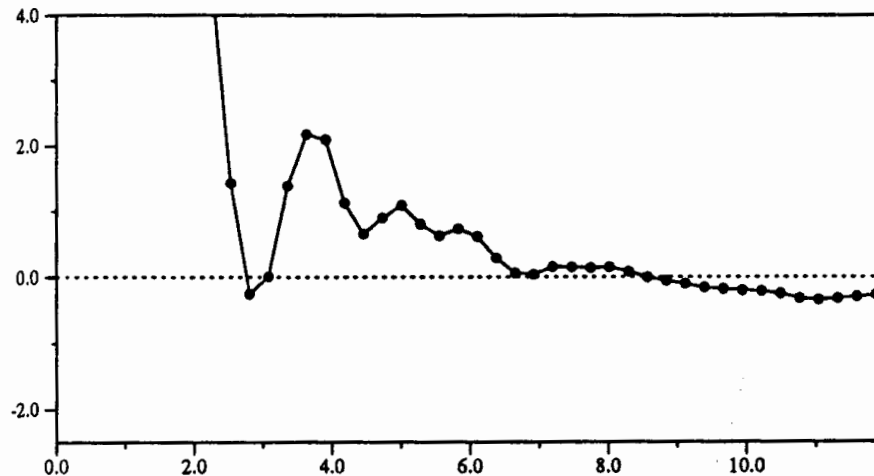
Tanaka and Sheraga (1975) *PNAS*, **72** pp3802
Sippl, (1990) *J.Mo.Biol.* **213** pp859
Godzik, (1996) *Structure* **15** pp363

Scoring Statistical Potential (reference state)



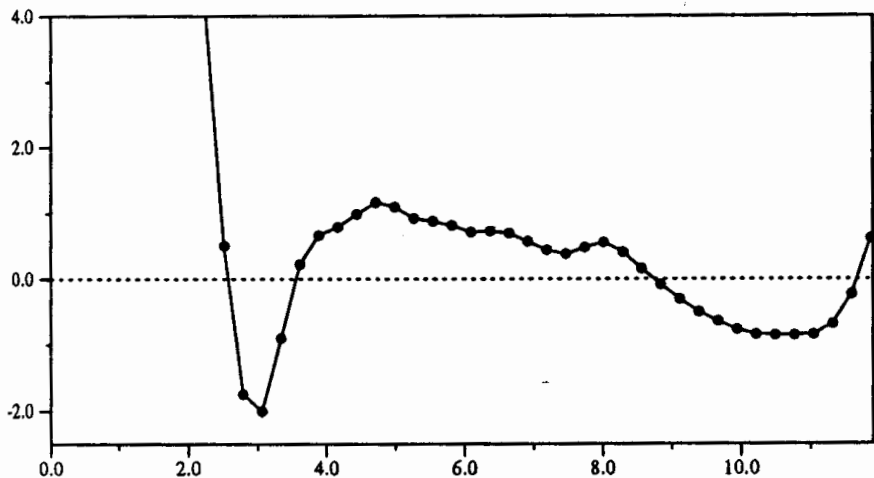
Scoring Statistical Potential... Hydrogen Bonds

Long range free energy



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

Short range free energy



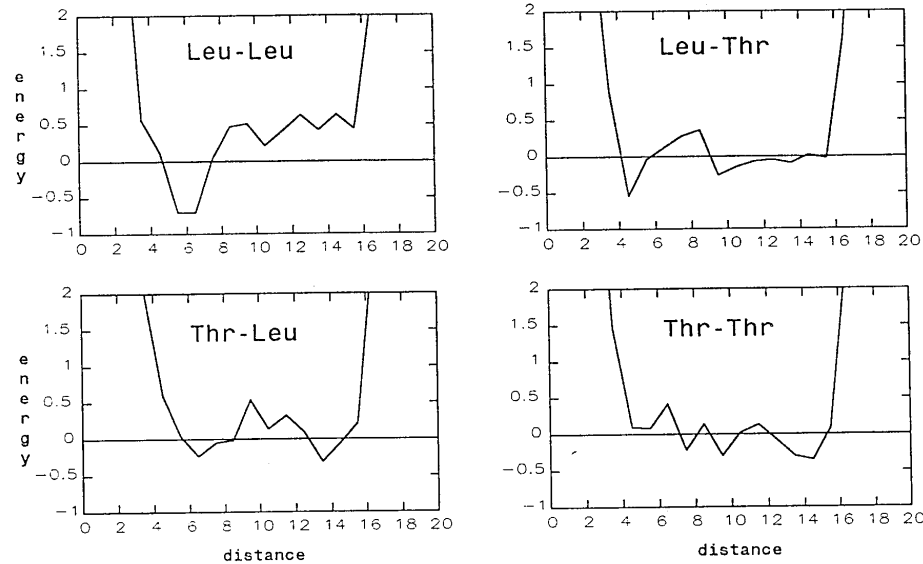
$$\rho_{NO}(r) = \sum_{ij} \delta(r - r_{ij})$$

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

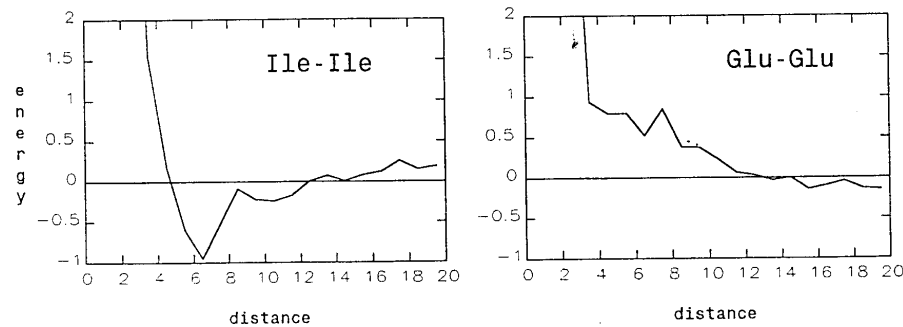
$$w_{NO}(r) = -kT \ln(g_{NO}(r))$$

Scoring Statistical Potential... Distance Potentials

Long range free energy



Short range free energy

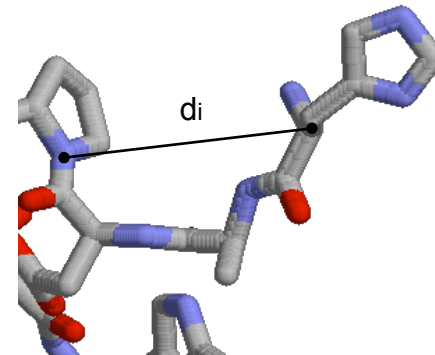


Scoring

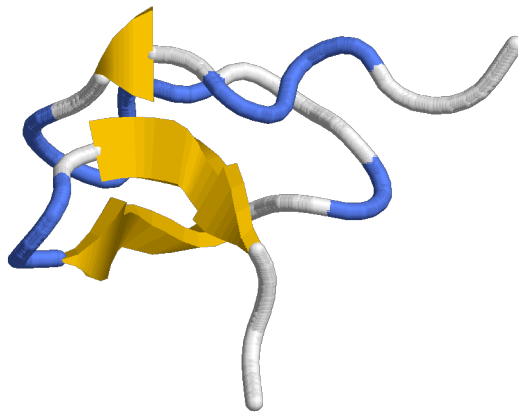
Raw scores of an alignment

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

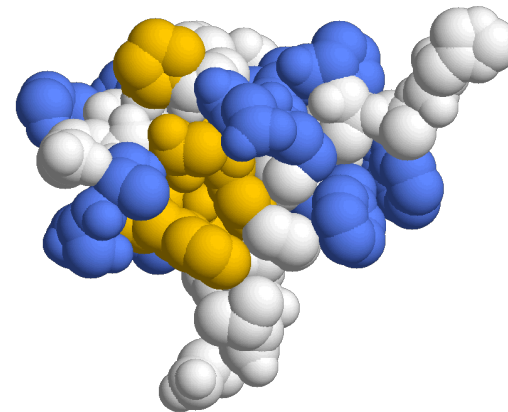
Aminoacid substitutions



Distance space



Secondary Structure (H,B,C)



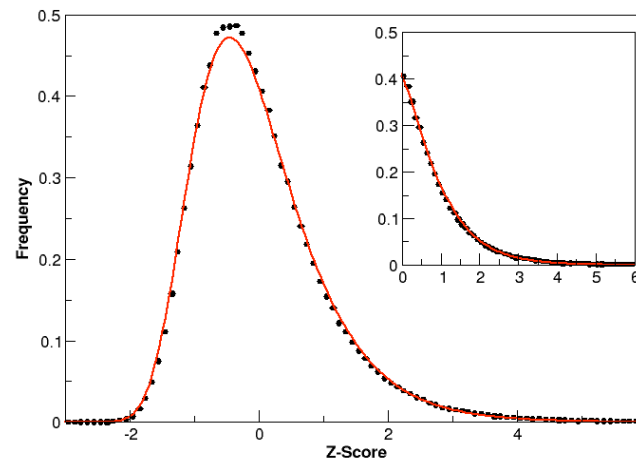
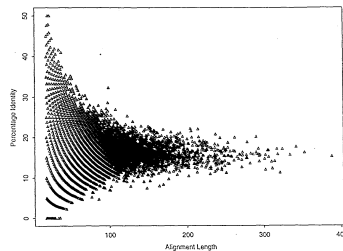
Accessible surface (B,A [%])

Scoring

Significance of an alignment (score)

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

Empirical



Sometimes approximated by Z-score (normal distribution).

Analytic

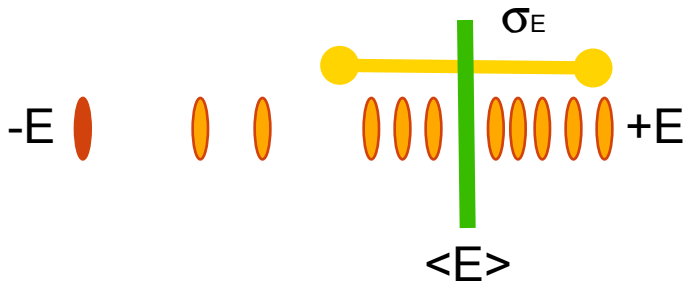
$$P(s) = e^{-\lambda (s-\mu)}$$

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

Scoring

Significance of an alignment (score)

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

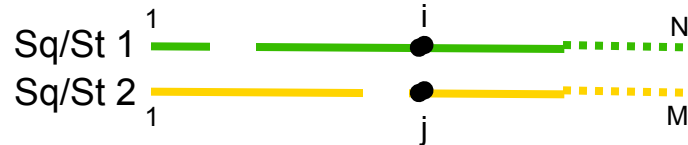


$$Zscore = \frac{(\langle E \rangle - E_m)}{\sigma_E}$$

Optimizer

Global dynamic programming alignment

remember Patsy's class



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

$$D_{ij} = \min \begin{cases} D_{i,j-1} + \text{Score}(\ddot{A}, r_j) \\ D_{i-1,j-1} + \text{Score}(r_i, r_j) \\ D_{i-1,j} + \text{Score}(r_i, \ddot{A}) \\ 0 \end{cases}$$

Best alignment score

Backtracking to get the best alignment

Applications of PMFs

Model assessment.

Ab initio folding simulations.

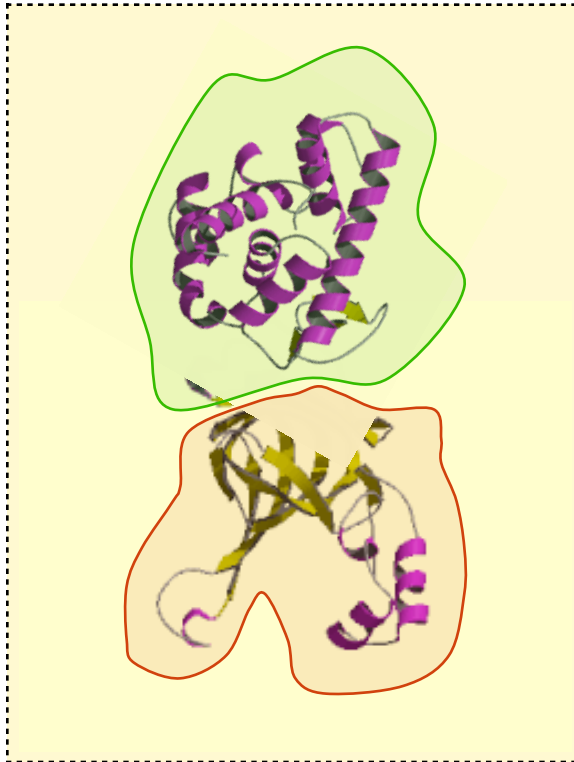
Sequence-structure matching (threading).

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

Secondary structure prediction, *etc.*

Domain boundaries from sequence

VERY DIFFICULT!!!!



MENFEIWVEKYRPRTLDEVVGQDEVIQRLKGYVERKNIPHLLFSGPPGTGKTATAIALARDLFGENWRDN
FIEMNASDERGIDVVRHKEFARTAPIGGAPFKIIFLDEADALTADAQAALRRTMEMYSKSCRFILSCN
YVSRIIEPIQSRCAVFRFKPVPKEAMKKRLLEICEKEGVKITEDGLEALIYISGGDFRKAINALQAAAI
GEVVDADTIYQITATARPEEMTELIQTALKGNFMEARELLDRLMVEYGMSEDIVAQLFREIISMPIKDS
LKVQLIDKLGEVDFRLTEGANERIQLDAYLAYLSTLAKK

Domain boundaries from sequence (SnapDragon)

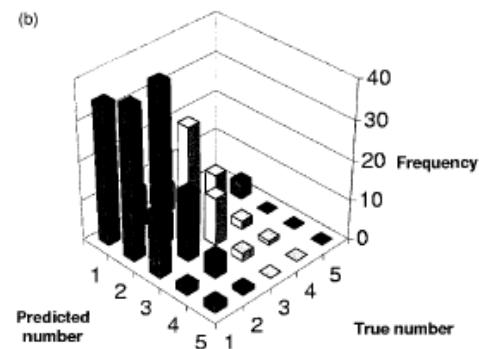
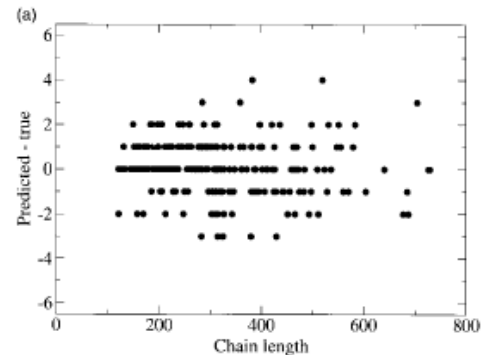
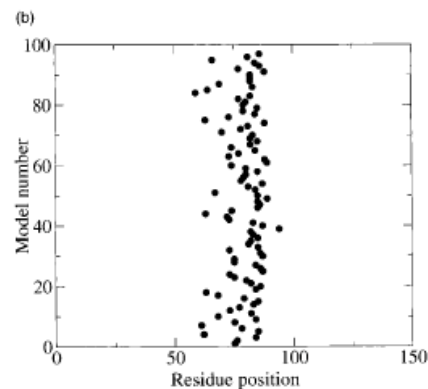
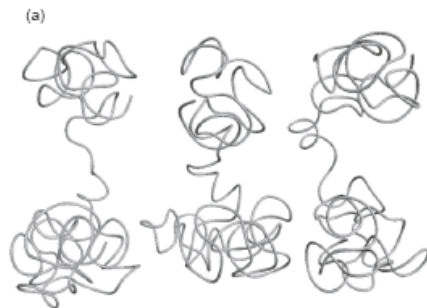
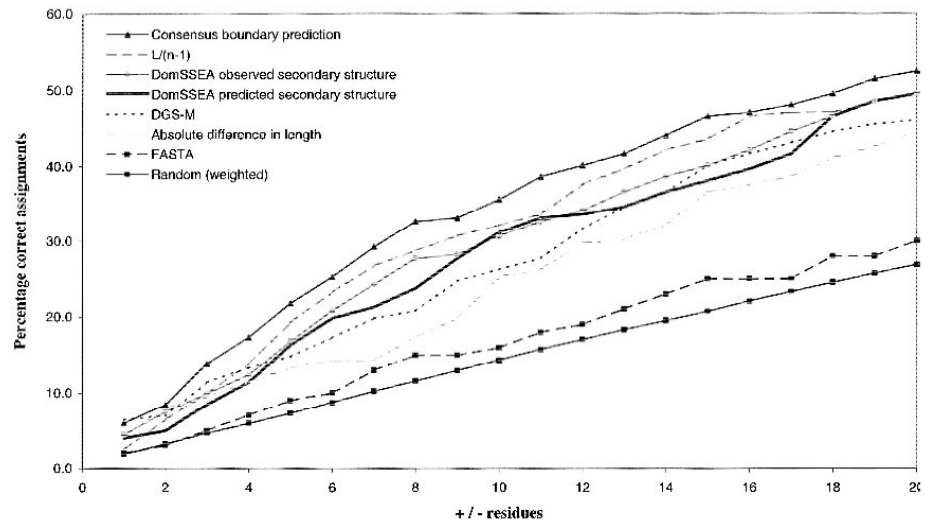
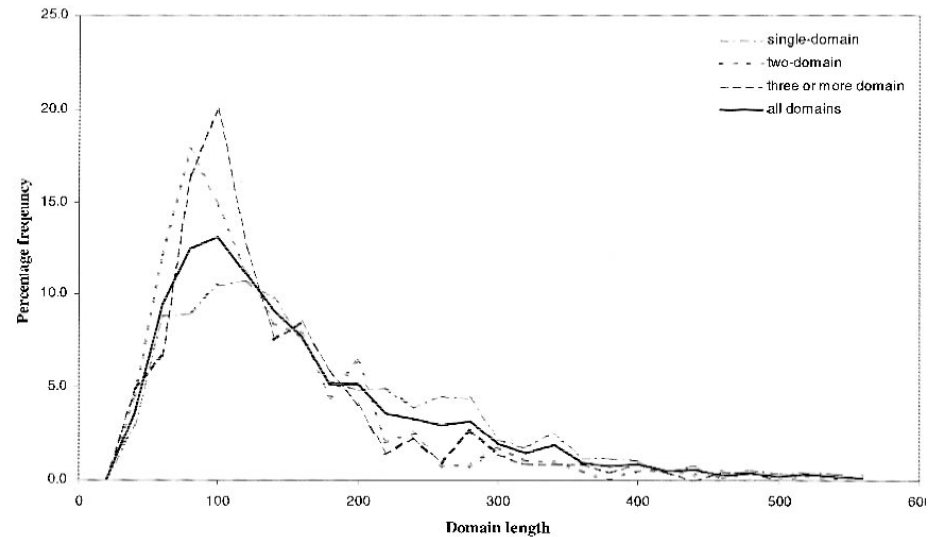


Table 2. Average accuracy percentages of linker prediction over 57 proteins

		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 and predicted SSE (DomSSEA)

Methods	% Correctly assigned	
	All chains	Multidomain chains
DomSSEA observed secondary structure	70.2	24.7
DomSSEA predicted & consensus	68.6	24.0
DomSSEA predicted & L/(N-1)	68.0	24.0
DomSSEA predicted secondary structure	68.7	23.6
Absolute difference in length	62.0	8.4
Average domain length & DGS-M	66.6	6.1
FASTA alignment	57.9	2.3
Random (weighted)	58.3	1.1
DGS-M	76.6	0.0
DGS-W	76.6	0.0



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	-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	
-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	-3	
0	2	0	4	-4	1	2	1	-2	-4	0	-3	-4	-3	1	-2	-5	-4	-4	
-1	5	3	-2	-4	-1	-1	1	-2	-1	-4	1	-3	-4	-3	1	-2	-5	-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	-4	-3	2	1	-4	-3	-1	
0	2	3	1	-4	0	0	0	-2	-4	1	-3	-4	-3	2	0	-5	-4	-4	
5	-3	-3	-3	-2	-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	
0	3	3	0	4	3	0	1	-2	-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	
-2	-3	-1	-5	-3	-3	-4	-5	-4	3	4	0	4	2	-4	-3	-2	-3	-2	
0	3	0	-2	-3	-1	0	0	-2	0	0	1	0	-1	-3	2	0	-4	-3	
-1	1	3	-2	-4	0	-2	4	-2	-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 "http://bioinf.cs.ucl.ac.uk/psiform.html". The page has a blue header with the UCL logo and "Bioinformatics Unit". The main content area is divided into sections: "PSIPRED home>", "Info", "Input Sequence", "Choose Prediction Method", "Filtering Options", and "Submit Sequence".

PSIPRED home>

The PSIPRED Protein Structure Prediction Server

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

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

Choose Prediction Method
[Help](#)
☐ Predict Secondary Structure (PSIPRED v2.4)
☐ Predict Transmembrane Topology (MEMSAT)
☐ Fold Recognition(GenTHREADER - quick)
☐ Fold Recognition (mGenTHREADER - with profiles and predicted secondary structure)

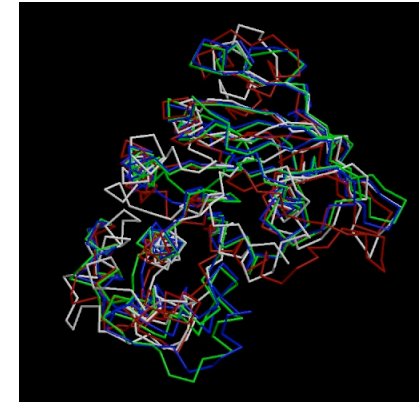
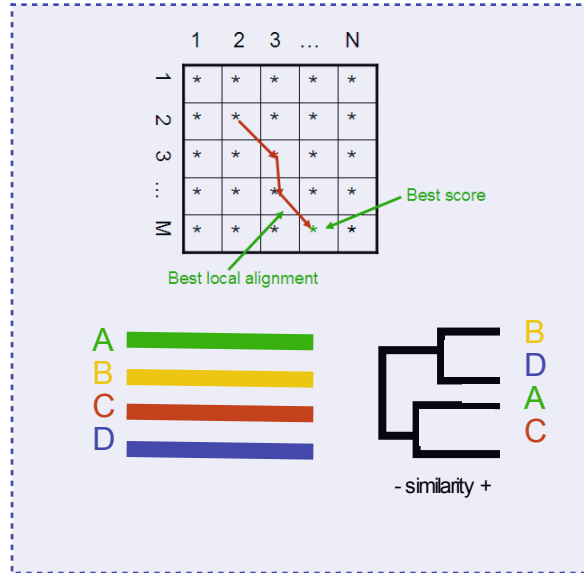
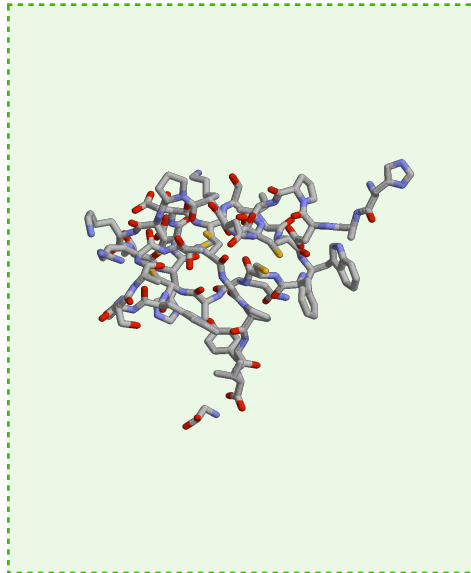
Filtering Options
[Help](#)
☒ Mask low complexity regions
☐ Mask transmembrane helices
☐ Mask coiled-coil regions
Warning: Turn off all filtering if you are running MEMSAT

Submit Sequence
E-mail address [Help](#)

Password (only required for commercial e-mail addresses) [Help](#)

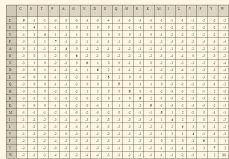
Short name for sequence [Help](#)

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

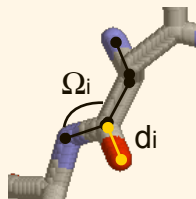


- ✓ Uses all available structural information
- ✓ Provides the optimal alignment

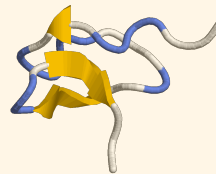
Computationally expensive



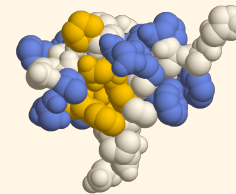
$R_{i,j}$



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



$S_{i,j}$



$B_{i,j}$

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

$I_{i,j}$

$$\text{Score}_{i,j} = w_1 * R_{i,j} + w_2 * D_{i(a),j(a)} + w_3 * S_{i,j} + w_4 * B_{i,j} + w_5 * I_{i,j} + w_6 * X_{i,j}$$

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

<http://alto.compbio.ucsf.edu/salign-cgi/index.cgi>



The screenshot shows a web browser window titled "SALIGN Server" with the address bar displaying "http://alto.compbio.ucsf.edu/salign-cgi/index.cgi". The page content is as follows:

SALIGN Multiple Structure/Sequence Alignment Server

SALIGN is a general alignment module of the modeling program MODELLER
The alignments are computed using dynamic programming, making use of several features of the protein sequences and structures

Users can either upload their own sequences/structures to align or choose structures from the PDB
sequences can either be pasted or uploaded as FASTA or PIR format alignment files

Paste sequence to align
Multiple sequences can be pasted by iteratively clicking 'upload' after every pasted sequence

[Text input field]

Specify file to upload (PIR, FASTA, PDB, .zip or .tar.gz)
Multiple files can be uploaded by iteratively clicking 'upload' after every file uploaded

[localized string not found] [localized string not found]

[Upload]

Uploaded files:
No files uploaded

Enter 4 letter code(s) to choose PDB structures

[Text input field]

e-mail address, to receive results:
[Text input field]

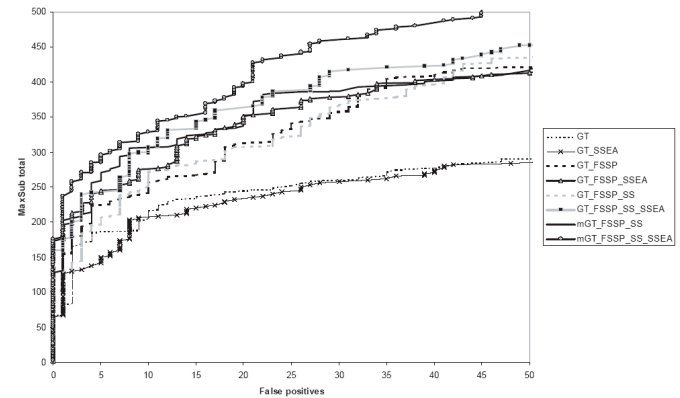
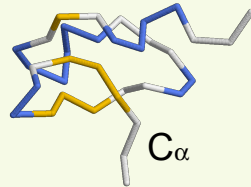
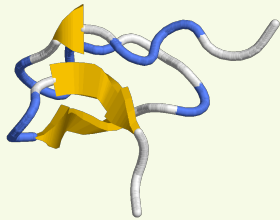
[Submit] [localized string not found]

Reference:
Madhusudhan, M.S., Marti-Renom, M.A., Eswar, N., Sali, A.
SALIGN - a multiple structure/sequence alignment tool, under preparation

Threading (mGenThreader)

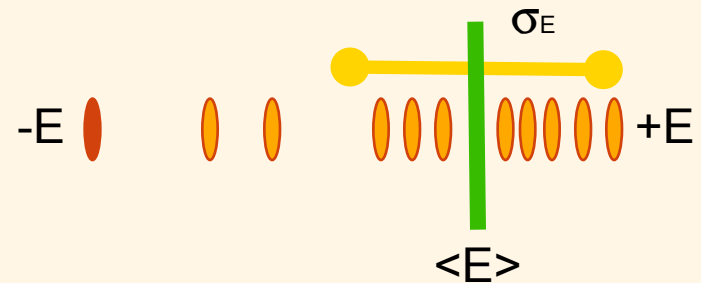
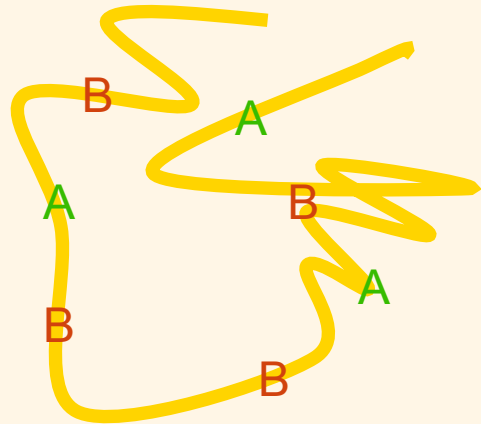
>gi42541361
MDIRSVSSLRGLLCLPPSWPRR

- Neural Network



✓ Good row and significance scoring

Obscure optimizer



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

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> Go Links

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)

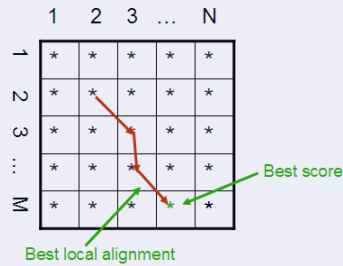
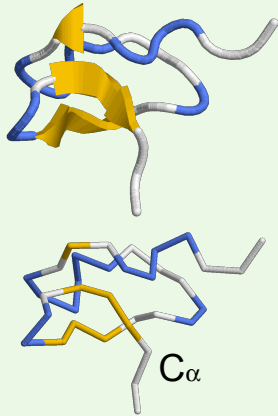
Choose Prediction Method
[Help](#)
☒ Predict Secondary Structure (PSIPRED v2.4)
☐ Predict Transmembrane Topology (MEMSAT)
☐ Fold Recognition(GenTHREADER - quick)
☐ Fold Recognition (mGenTHREADER - with profiles and predicted secondary structure)

Filtering Options
[Help](#)
☒ Mask low complexity regions
☐ Mask transmembrane helices
☐ Mask coiled-coil regions
Warning: Turn off all filtering if you are running MEMSAT

Done Internet

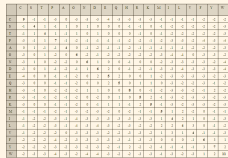
Remote homology detection (FUGUE)

>gi42541361
MDIRSVSSLRGLLCLPPSWPRR

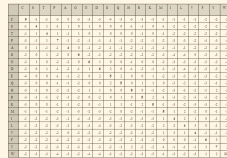


- ✓ Uses most of the structural information
- ✓ Easy to access either locally and on the web
- ✓ Good row and significance scoring

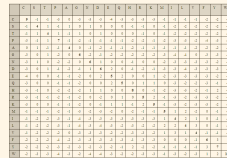
Does not uses multiple sequence information



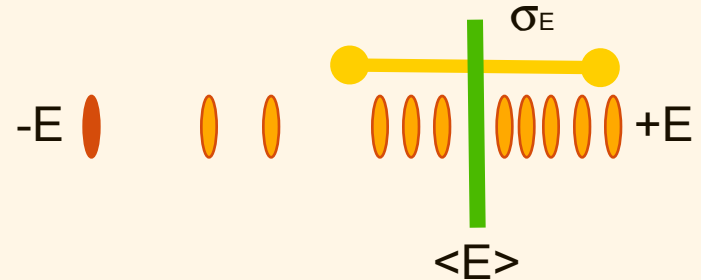
$$R_{i,j}^H$$



$$R_{i,j}^B$$



$$R_{i,j}^C$$



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

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

Back Forward Stop Search Favorites Media Print Mail News RSS Feeds

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

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.

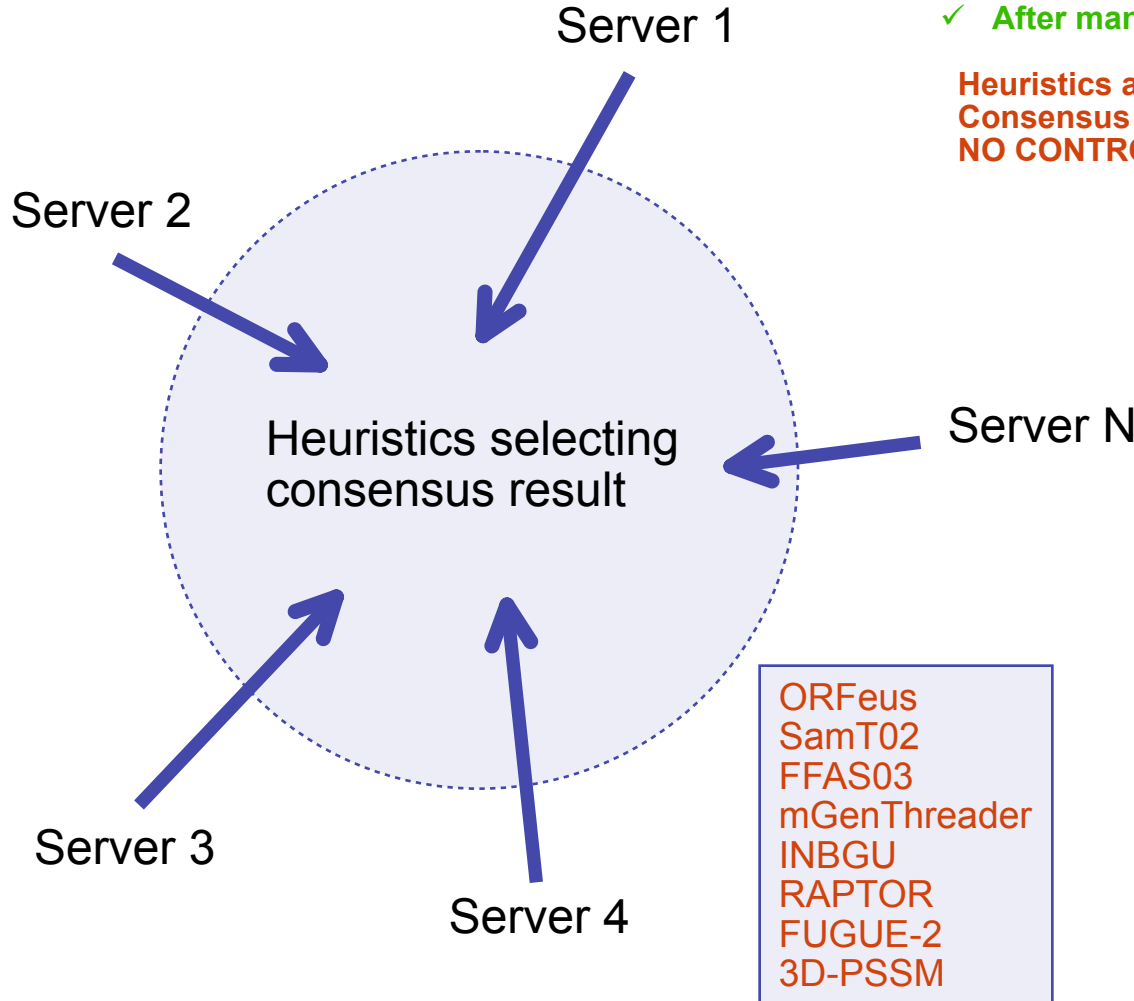
Read the original paper for more details:
[J. Shi, T. L. Blundell, and K. Mizuguchi](#) (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.*, **310**, 243-257.
[Medline](#), [Article on-line](#), [PDF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

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.

Internet

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!

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 (Mam et al., 2002; Rota et al., 2002).

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 predictor 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 interesting findings obtained during the SARS genome annotation process is a surprisingly reliable (3D jury score >100) assignment of the methyltransferase fold to the nsP13 (GI20133975) domain located in the C-terminal part of the almost 7000 amino acid large pp1a viral polyprotein (Figure 1). Standard sequence comparison tools such as PSI-BLAST or RPS-BLAST applied using the conserved domain database (Marchler-Bauer et al., 2003) failed to assign any function to this domain. The domain belongs to the ancient family of AdoMet-dependent ribose 2'-O-methyltransferases, which has been adapted by numerous viruses before the three domains of life evolved from the last universal common ancestor (LUCA) (Ficker et al., 2002). The enzymatic role of the protein was confirmed by the presence of the conserved tetrad of residues K-D-K-E essential for mRNA cap-1 (m⁷GpppN) formation.

The mRNA cap methyltransferase is found indispensable for efficient replication of many viruses (Bach et al., 1995; Woycikowski et al., 1995; Vlot et al., 2002) and represents an active area for drug development. Nevertheless, direct inhibitors of the nsP13 enzyme may fail to suppress viral replication, as the cap-1 formation seems to be less critical than the preceding cap-0 (m⁷GpppN) formation (Luthe et al., 2002; Wu and Guarino, 2002). The existence of the cap-1-forming enzyme in the genome would suggest that the virus also requires the AdoMet-dependent cap-0 methyltransferase. Both functions can be inhibited by carboxylic analogs of adenosine, such as Neplanodin A or 3-desazaneplanodin A, which interfere with the AdoMet-AdoHcy 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 Grotthuss, Lucjan S. Wysocki, and Leszek Rykaczewski*
 Bionotbank Institute
 Ulanowoskiego 24A
 60-744 Poznań
 Poland

*Correspondence: leszek@bionotbank.pl

Figure 1. 3D Model of the nsP13 Domain of the SARS Coronavirus pp1a Polyprotein

This model is based on the reassigned (Ginalski and Rykaczewski, 2001) cap-1 methyltransferase of the reovirus 12 protein (4 app [PDB: 1vot]). While other templates (that) or (that) obtained marginally higher 3D jury scores, the selected template had the lowest number of insertions and deletions. Disruption of the conserved tetrad of residues (K-D-K-E) essential for cap-1 methylation and the docked AdoMet 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/>

Meta Server Job List, BioInfo.PL - Microsoft Internet Explorer

Address <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):

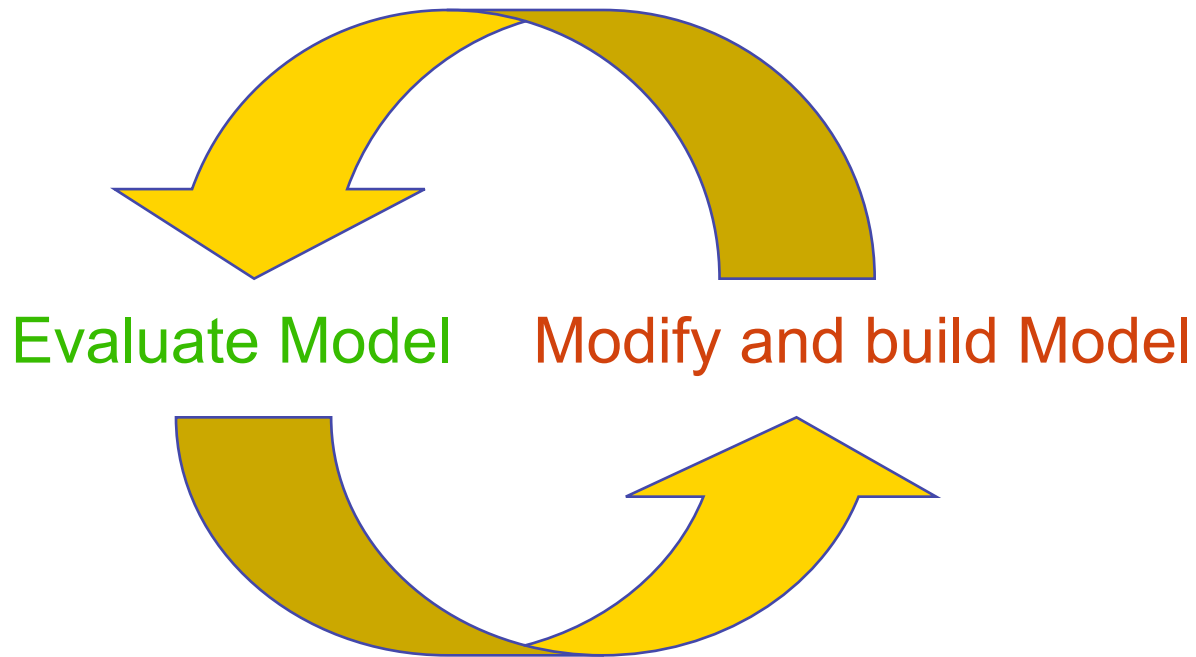
Please submit domains separately
Please remove coiled coil regions
Check [LiveBench](#) for evaluation of the reliability of the servers
Results are stored only for 1 month
Jobs queued for more than 7 days for servers with queue>30 are skipped
Use is limited to 10 jobs per week per domain
Please contact 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. \[PubMed\]](#)

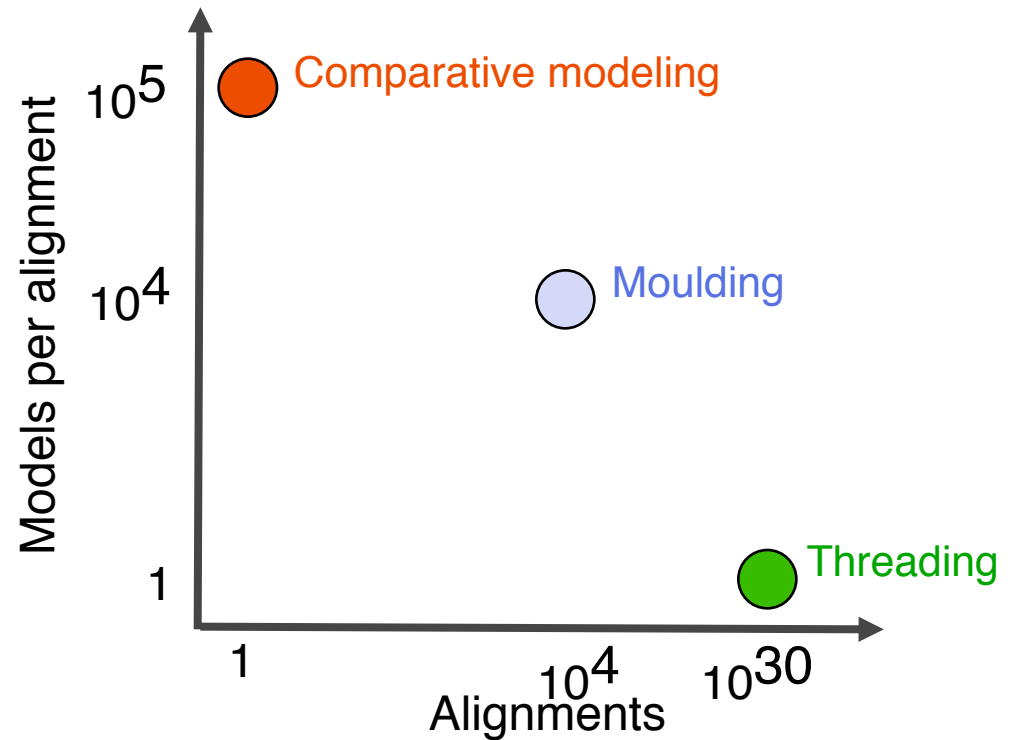
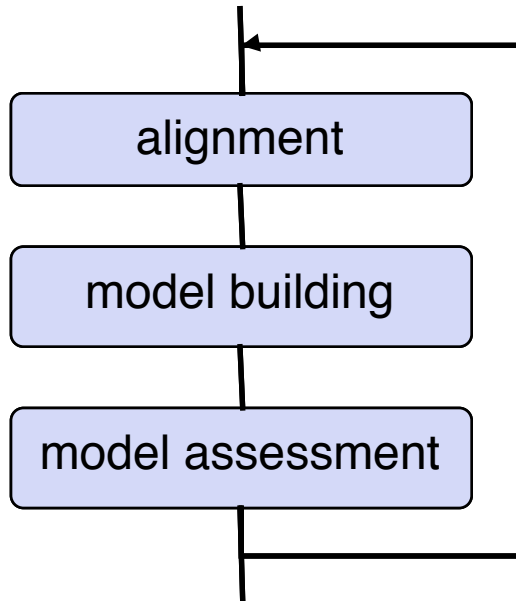
Skip: Queue:

☐ PDB-Blast	1
☒ 3D-Jigsaw	43
☐ ESyPred3D	1
☐ GRDB	1
☐ FFAS03	1
☐ Sam-T99	1
☐ SUPERFAMILY	1
☒ INBGU	39
☐ FUGUE2	1
☐ 3D-PSSM	1
☐ mGenTHREADER	
☐ psipred	
☐ profsec	1
Pcons2	1
3D-ShotGun	11
3D-Jury	

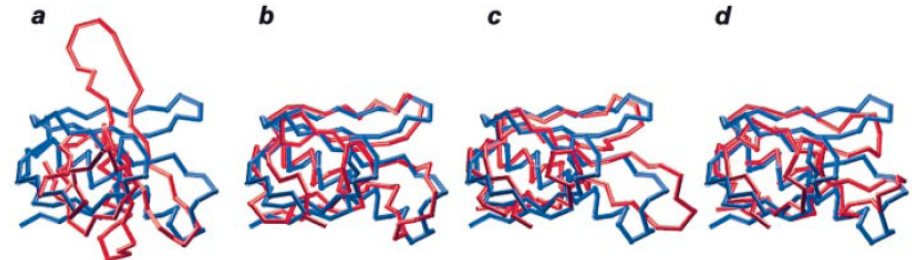
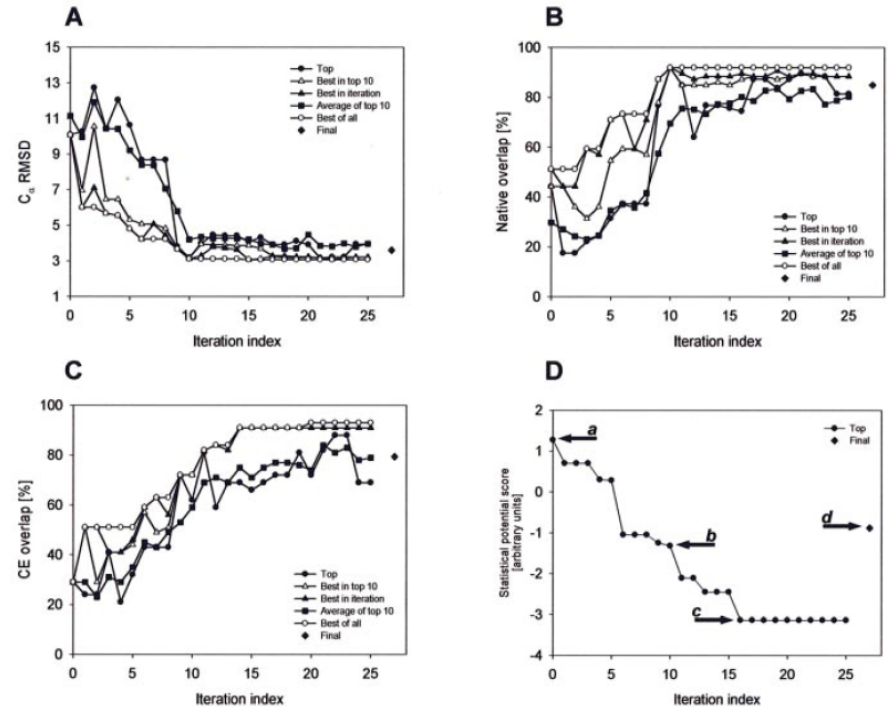
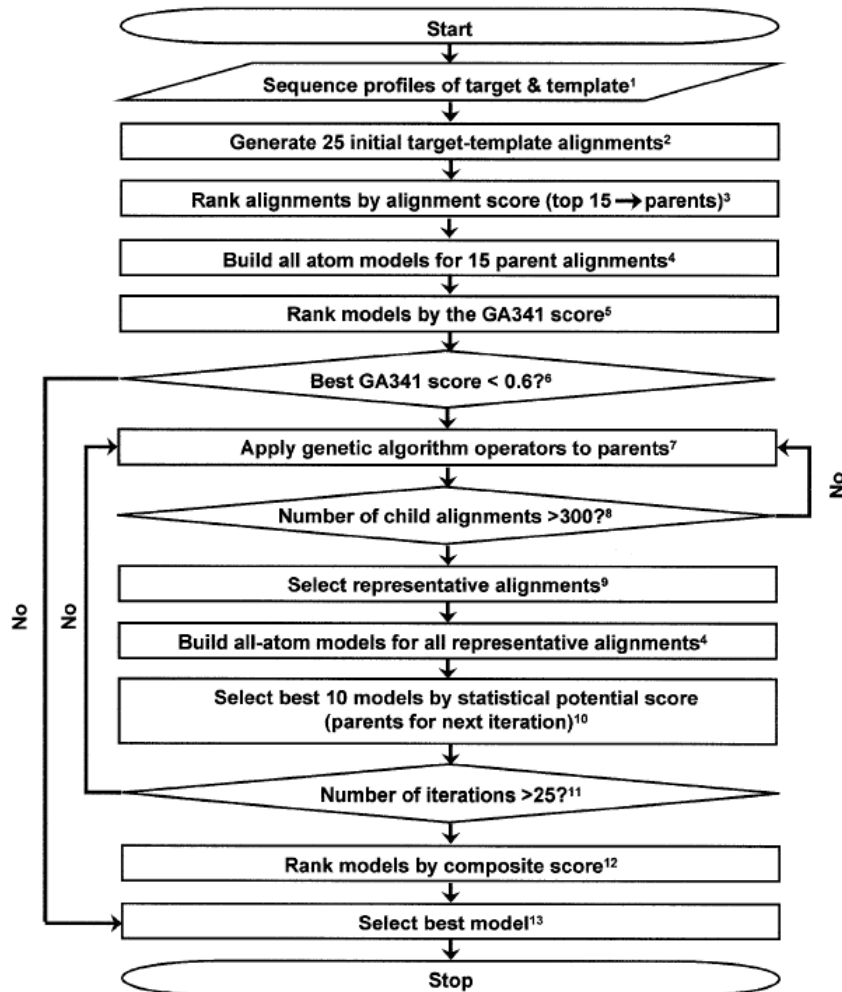
Iterative process... better models(?)



Moulding: iterative alignment, model building, model assessment



Iterative process... MOULDER



Genetic algorithm operators

Single point cross-over

...TSSQ—NMKLGVFWGY—...
...V—SSCN—GDLHMKVGV...
...TSSQNMK—LGVFWGY...
...VSSCNGDLHMKV—GV...



...TSSQ—NMK—LGVFWGY...
...V—SSCNGDLHMKV—GV...
...TSSQNMKLGVFWGY—...
...VSSCN—GDLHMKVGV...

Gap insertion

...TSSQNMKLGVFWGY...
...VSSCNGDLHMKVGV...



...TSSQN—MKLGVFWGY...
...VSSCNGDLHMKVG—V...

Gap shift

...T—SSQNMKLGVFWGY...
...VSSCNGDLHMKVGV—...



...—T—SSQNMKLGVFWGY...
...VSSCNGDLHMKVGV—...
...T—S—SSQNMKLGVFWGY...
...VSSCNGDLHMKVGV—...
...—TSSQNMKLGVFWGY...
...VSSCNGDLHMKVGV—...
...TS—SSQNMKLGVFWGY...
...VSSCNGDLHMKVGV—...

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

Composite model assessment score

Weighted linear combination of several scores:

- Pair (P_p) and surface (P_s) statistical potentials;
- Structural compactness (S_c);
- Harmonic average distance score (H_a);
- Alignment score (A_s).

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

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

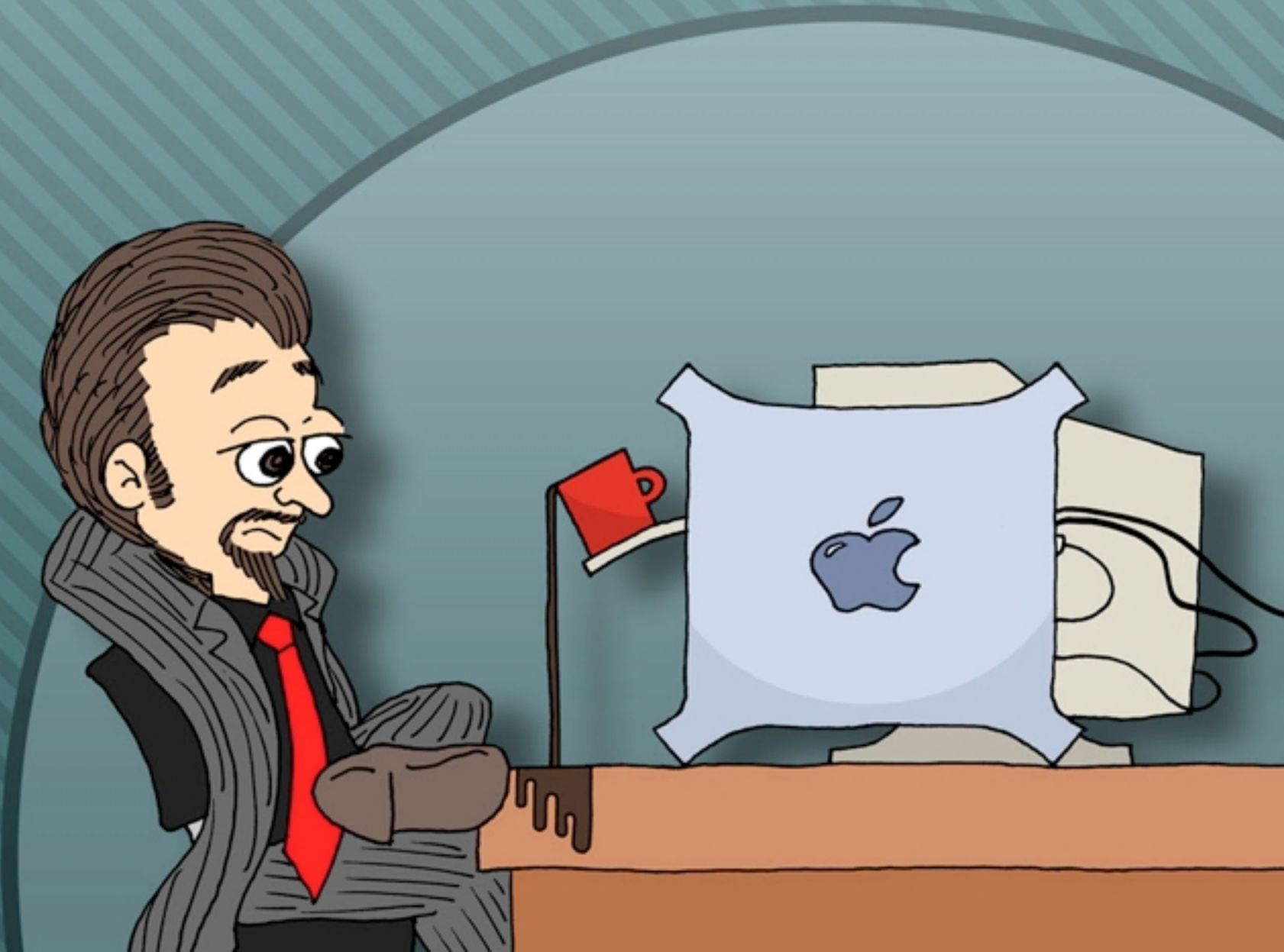
μ ... average score of all models

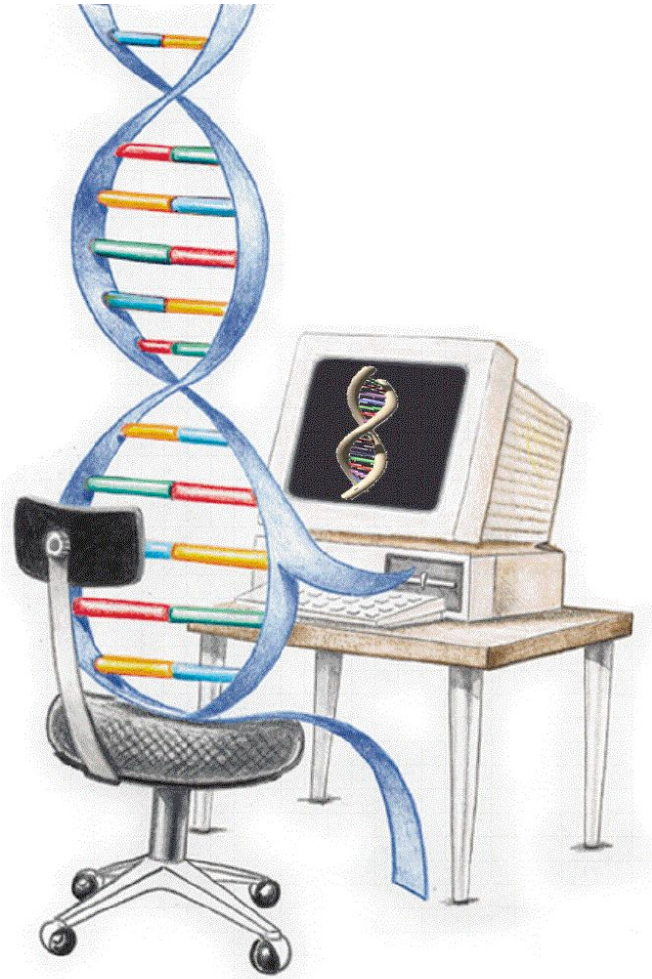
σ ... standard deviation of the scores

Benchmark with the “very difficult” test set

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

Target -template	Sequence identity [%]	Coverage [% aa]	Initial prediction		Final prediction		Best prediction	
			C α RMSD [Å]	CE overlap [%]	C α RMSD [Å]	CE overlap [%]	C α RMSD [Å]	CE overlap [%]
1ATR-1ATN	13.8	94.3	19.2	20.2	18.8	20.2	17.1	24.6
1BOV-1LTS	4.4	83.5	10.1	29.4	3.6	79.4	3.1	92.6
1CAU-1CAU	18.8	96.7	11.7	15.6	10.0	27.4	7.6	47.4
1COL-1CPC	11.2	81.4	8.6	44.0	5.6	58.6	4.8	59.3
1LFB-1HOM	17.6	75.0	1.2	100.0	1.2	100.0	1.1	100.0
1NSB-2SIM	10.1	89.2	13.2	20.2	13.2	20.1	12.3	26.8
1RNH-1HRH	26.6	91.2	13.0	21.2	4.8	35.4	3.5	57.5
1YCC-2MTA	14.5	55.1	3.4	72.4	5.3	58.4	3.1	75.0
2AYH-1SAC	8.8	78.4	5.8	33.8	5.5	48.0	4.8	64.9
2CCY-1BBH	21.3	97.0	4.1	52.4	3.1	73.0	2.6	77.0
2PLV-1BBT	20.2	91.4	7.3	58.9	7.3	58.9	6.2	60.7
2POR-2OMF	13.2	97.3	18.3	11.3	11.4	14.7	10.5	25.9
2RHE-1CID	21.2	61.6	9.2	33.7	7.5	51.1	4.4	71.1
2RHE-3HLA	2.4	96.0	8.1	16.5	7.6	9.4	6.7	43.5
3ADK-1GKY	19.5	100.0	13.8	26.6	11.5	37.7	7.7	48.1
3HHR-1TEN	18.4	98.9	7.3	60.9	6.0	66.7	4.9	79.3
4FGF-81IB	14.1	98.6	11.3	24.0	9.3	30.6	5.4	41.2
6XIA-3RUB	8.7	44.1	10.5	14.5	10.1	11.0	9.0	34.3
9RNT-2SAR	13.1	88.5	5.8	41.7	5.1	51.2	4.8	69.0
AVERAGE	14.2	85.2	9.6	36.7	7.7	44.8	6.3	57.8





**some biology?
please...**

Common Evolutionary Origin of Coated Vesicles and Nuclear Pore Complexes

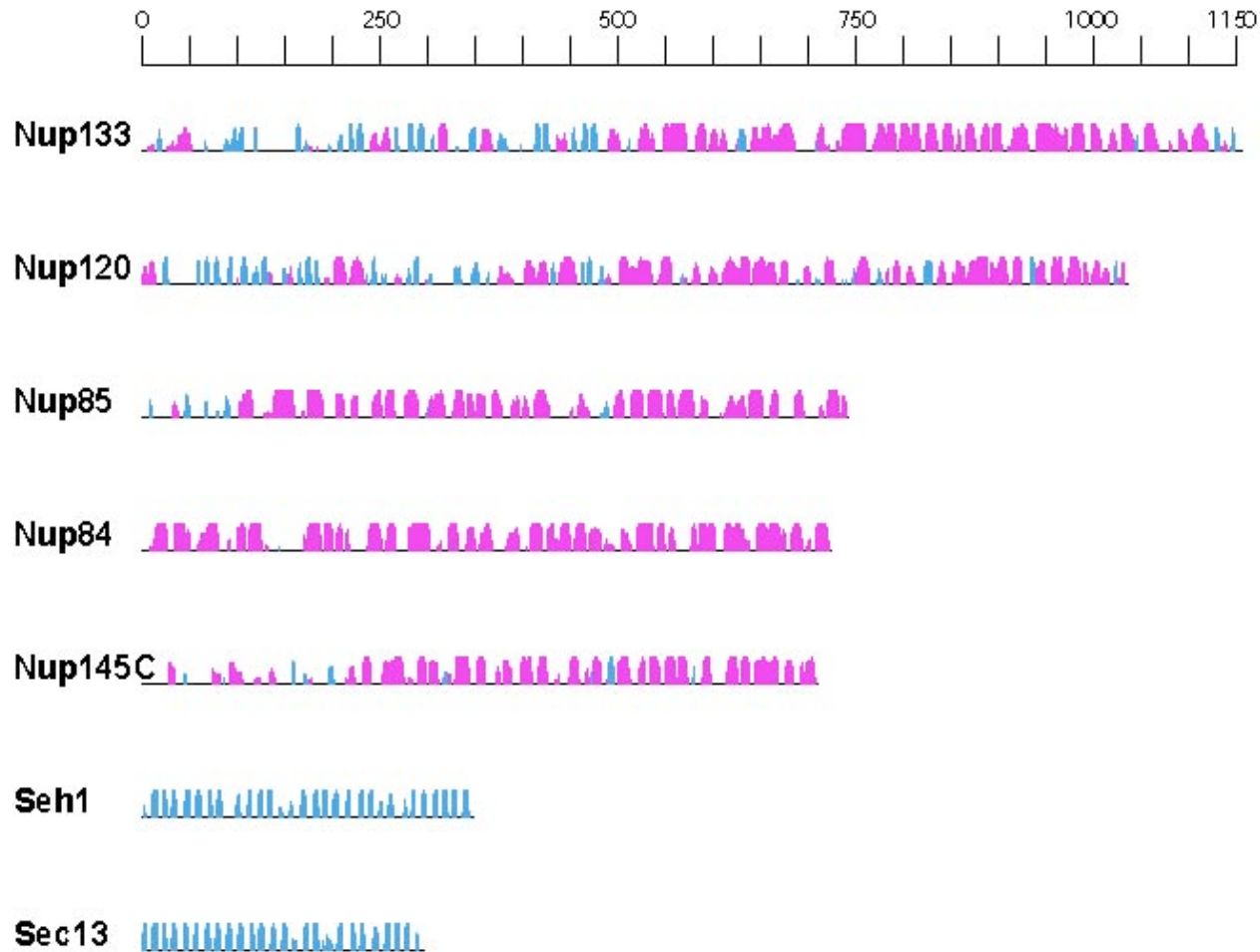
mGenThreader + SALIGN + MOULDER

D. Devos, S. Dokudovskaya, F. Alber, R. Williams, B.T. Chait, A. Sali, M.P. Rout.

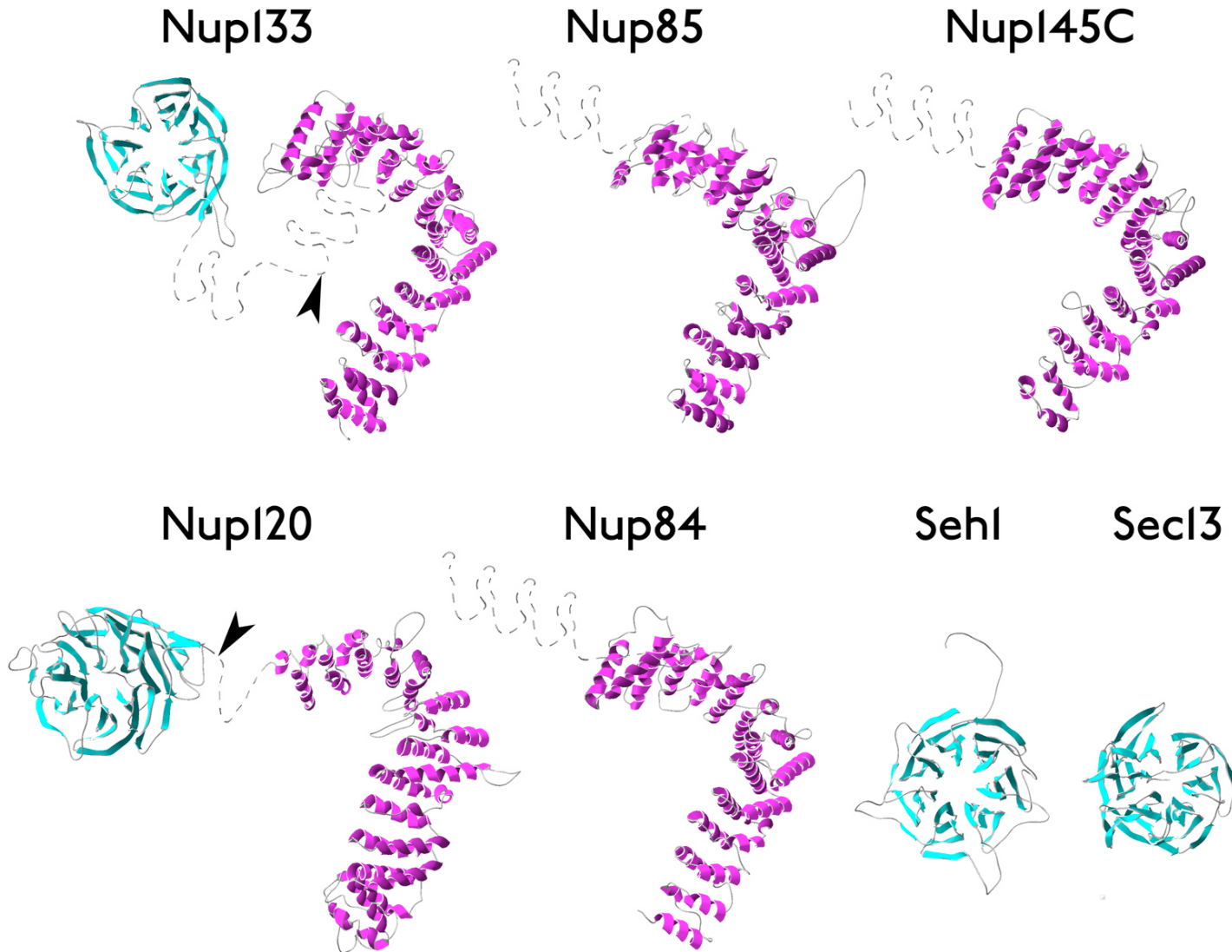
Components of Coated Vesicles and Nuclear Pore Complexes Share a Common Molecular Architecture.

PLOS Biology **2(12)**:e380, 2004

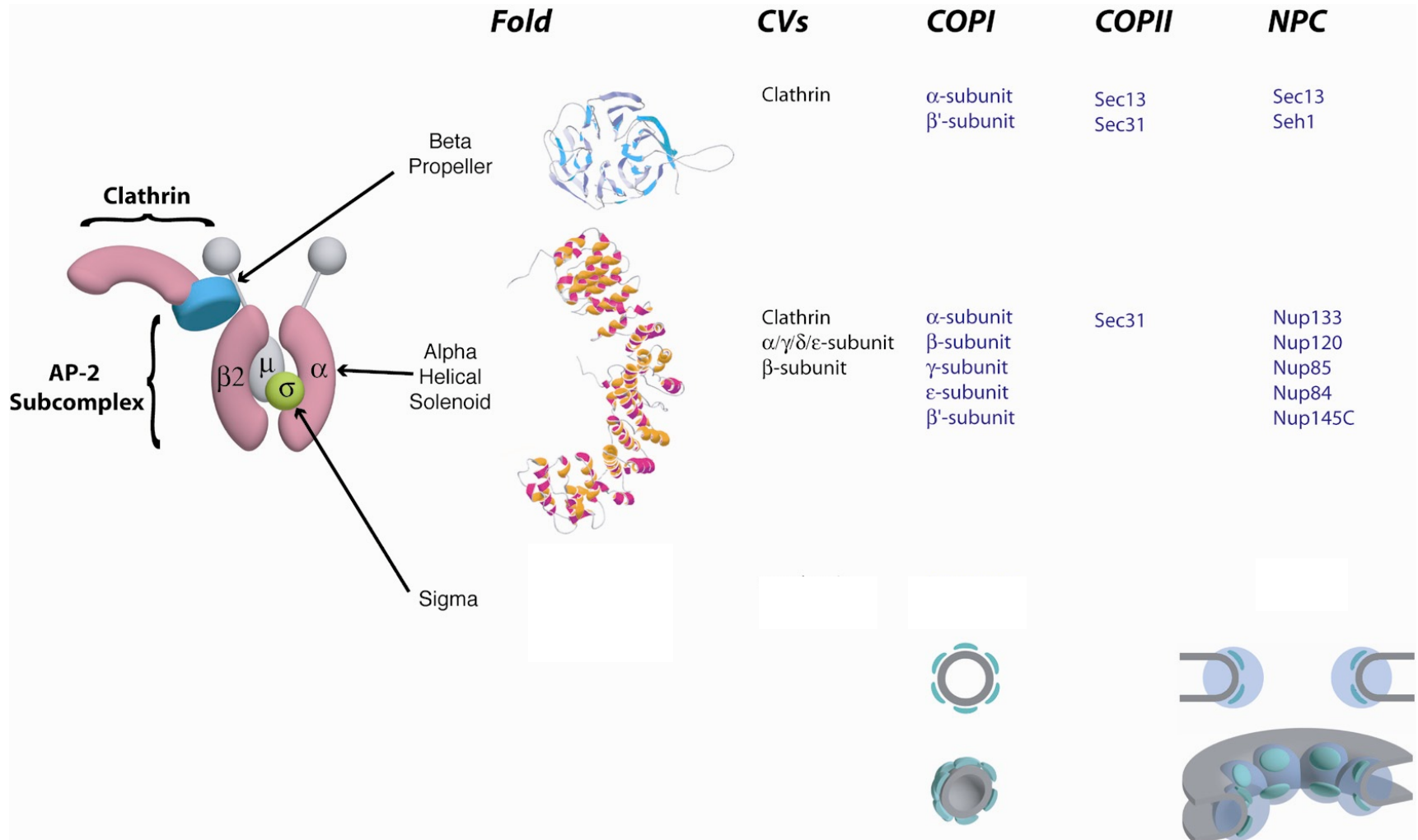
yNup84 complex proteins



All Nucleoporins in the Nup84 Complex are Predicted to Contain β -Propeller and/or α -Solenoid Folds



NPC and Coated Vesicles Share the β -Propeller and α -Solenoid Folds and Associate with Membranes

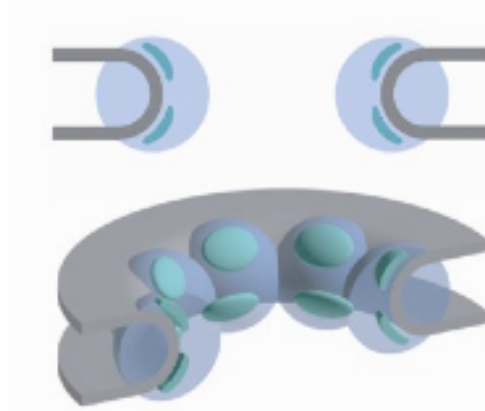


NPC and Coated Vesicles Both Associate with Membranes

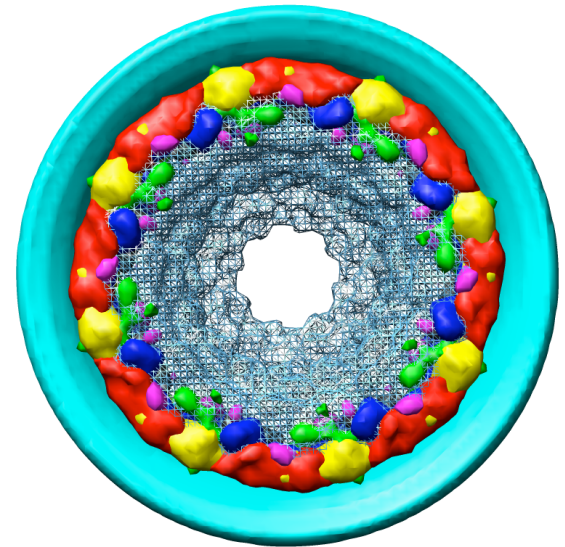
Coated Vesicle



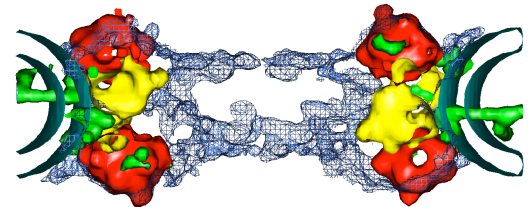
NPC model



top view



side view

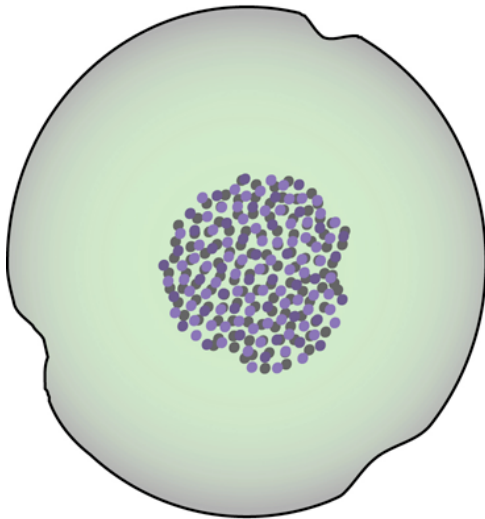


Nup 84 complex

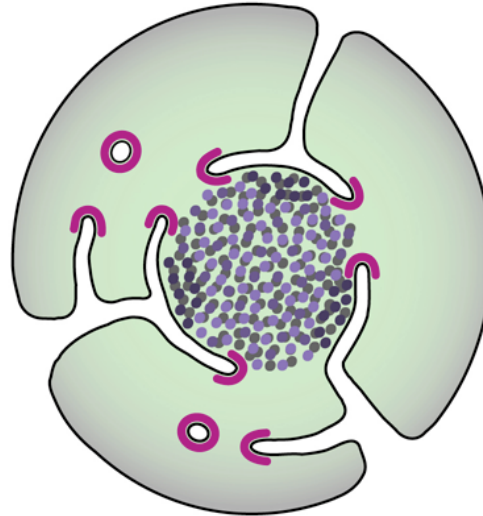
A Common Evolutionary Origin for Nuclear Pore Complexes and Coated Vesicles?

The proto-coatomer hypothesis

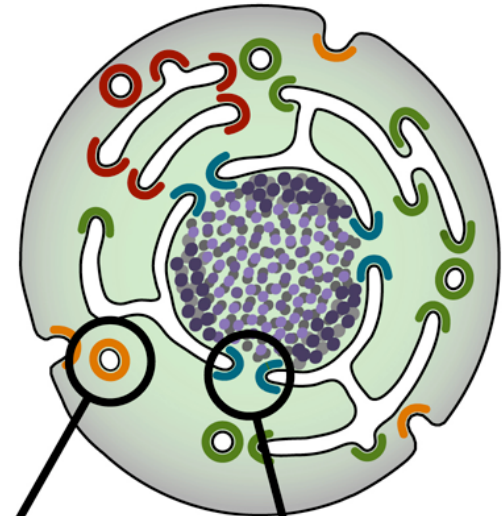
Prokaryote



Early Eukaryote

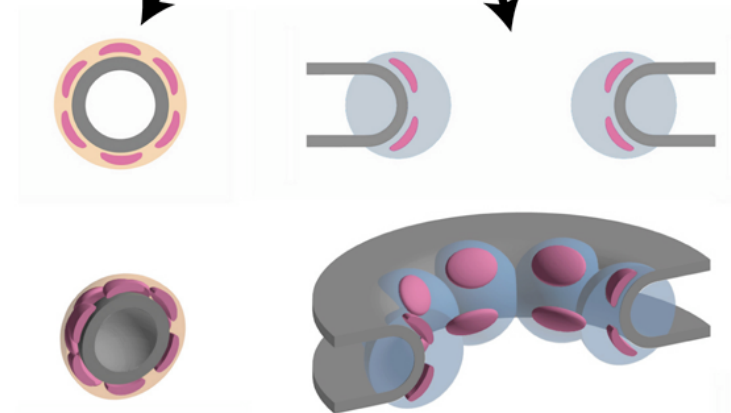


Modern Eukaryote



A simple coating module containing minimal copies of the two conserved folds evolved in proto-eukaryotes to bend membranes.

The progenitor of the NPC arose from a membrane-coating module that wrapped extensions of an early ER around the cell's chromatin.



Course assignment

The POM152 nucleoporin protein

Course assignment

The POM152 nucleoporin protein

Introduction

The POM152 protein functions as a component of the nuclear pore complex (NPC). NPC components, collectively referred to as nucleoporins (NUPs), can play the role of both NPC structural components and of docking or interaction partners for transiently associated nuclear transport factors. POM152 is important for the de novo assembly of NPCs.

The nuclear pore complex (NPC) constitutes the exclusive means of nucleocytoplasmic transport. NPCs allow the passive diffusion of ions and small molecules and the active, nuclear transport receptor-mediated bidirectional transport of macromolecules such as proteins, RNAs, ribonucleoparticles (RNPs), and ribosomal subunits across the nuclear envelope. The 55-60 MDa NPC is composed of at least 31 different subunits. Due to its 8-fold rotational symmetry, all subunits are present with 8 copies or multiples thereof. POM152 is known to interact with NUP188.

Assignment

1. Predict the domain boundaries for the POM152 sequence
2. Search for a suitable template/s for the POM152 domains
3. Align the sequences of POM152 domains against the template/s sequences
4. Build a 3D-models of the POM152 domains
5. Evaluate the models
6. Indicate possible applications of the models

GRADING: The entire assignment is worth 20 points.

The screenshot displays the NCBI Protein database entry for the POM152 protein (P39685). The top navigation bar includes links for Nucleotide, Protein, Genome, Structure, PMC, Taxonomy, and other resources. The search interface shows the protein name and accession number. The main content area provides a detailed summary of the protein, including its locus, definition, accession number, version, and database source. It also lists keywords, source organism, and references. The protein is identified as Nucleoporin POM152 (Nuclear pore protein POM152) (Pore membrane protein POM152) (P150).

LOCUS	P39685	1337 aa	linear	PLN 25-JAN-2005
DEFINITION	Nucleoporin POM152 (Nuclear pore protein POM152) (Pore membrane protein POM152) (P150).			
ACCESSION	P39685			
VERSION	P39685 GI:730249			
DBSOURCE	swissprot: locus P152_YEAST, accession P39685; class: standard.			

created: Feb 1, 1995.
sequence updated: Feb 1, 1995.
annotation updated: Jan 25, 2005.

xrefs: gi: 473153, gi: 473154, gi: 728663, gi: 728668, gi: 626784
xrefs (non-sequence databases): IntAct:P39685, GermOnline:42798, SCDS000004736, G00005739, G00005643, G00005198, G00006609, G00006466, G00006607, G00006999, G00006611, G00006610, G00006407, G00006408, G00006608, G00006409

KEYWORDS
Direct protein sequencing; Glycoprotein; mRNA transport; Nuclear pore complex; Nuclear protein; Phosphorylation; Protein transport; Repeat; Translocation; Transmembrane; Transport.

SOURCE
ORGANISM
Saccharomyces cerevisiae (baker's yeast)
Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.

REFERENCE
AUTHORS
Wozniak,R.W., Blobel,G. and Rout,M.P.
TITLE
POM152 is an integral protein of the pore membrane domain of the yeast nuclear envelope
J. Cell Biol. 125 (1), 31-42 (1994)
JOURNAL
PUBMED
8138573
REMARK
NUCLEOTIDE SEQUENCE, PARTIAL PROTEIN SEQUENCE, GLYCOSYLATION, AND REPEATS.

REFERENCE
AUTHORS
Bowman,S., Churcher,C., Badcock,K., Brown,D., Chillingworth,T., Connor,R., Dedman,K., Devlin,K., Gentles,S., Hamlin,N., Hunt,S., Jagels,K., Lye,G., Moule,S., Odell,C., Pearson,D., Rajandream,M., Rice,P., Skelton,J., Walsh,S., Whitehead,S. and Barrett,I.B.
TITLE
The nucleotide sequence of Saccharomyces cerevisiae chromosome XIII
Nature 387 (6632 SUPPL), 90-93 (1997)
JOURNAL
PUBMED
9169872
REMARK
NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
STRAIN=S288c / AB972

REFERENCE
AUTHORS
Nehrbass,U., Rout,M.P., Maguire,S., Blobel,G. and Wozniak,R.W.
TITLE
The yeast nucleoporin Nup188p interacts genetically and physically with the core structures of the nuclear pore complex
J. Cell Biol. 133 (6), 1153-1162 (1996)
JOURNAL
PUBMED
8682855

<http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=protein&val=730249>

BMI 206

MODELLER TUTORIAL

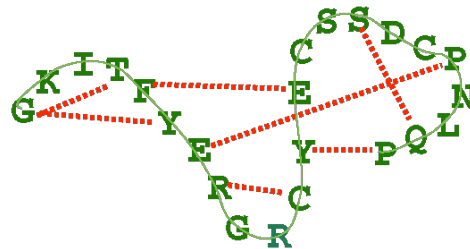
<http://www.salilab.org/modeller/tutorial/>

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

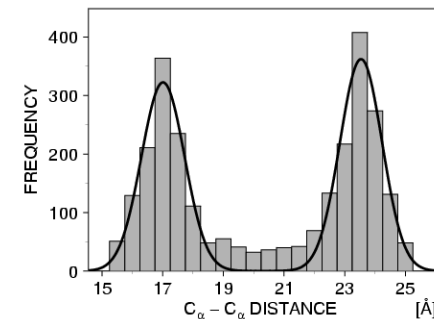
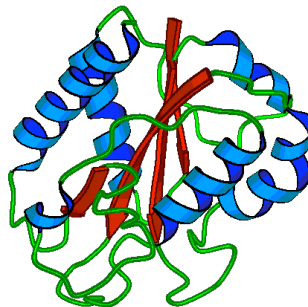
Comparative Modeling by Satisfaction of Spatial Restraints (MODELLER)

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

1. Extract spatial restraints

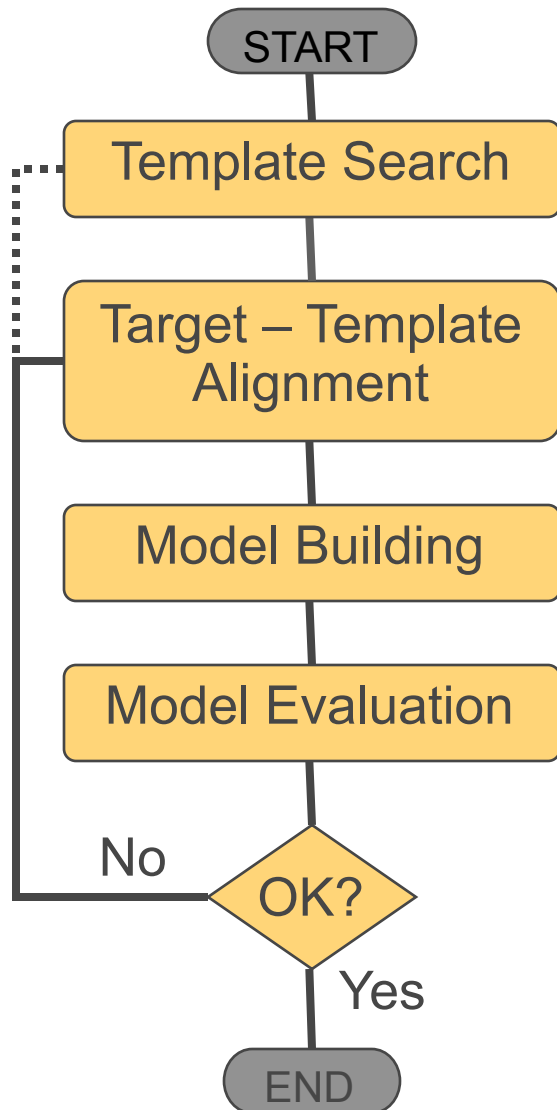


2. Satisfy spatial restraints



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

Steps in Comparative Protein Structure Modeling



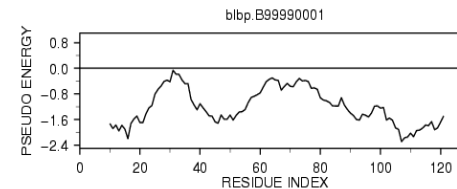
TARGET

ASILPKRLFGNCEQTSDEG
LKIERTPLVPHISAQNVCLKI
DDVPERLIPERASFQWMN
DK

TEMPLATE



ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCETSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE



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

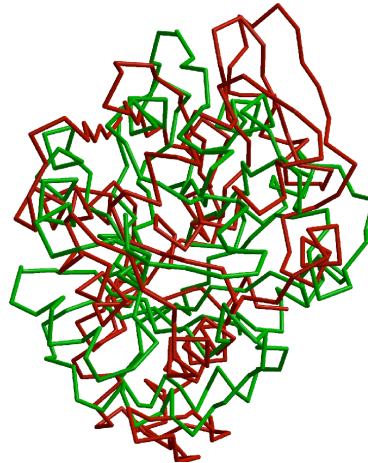
R. Sánchez & A. Šali, *Curr. Opin. Str. Biol.* 7, 206, 1997.

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

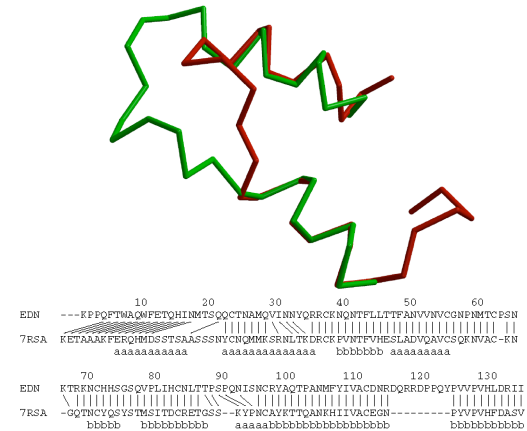
Typical errors in comparative models

MODEL
X-RAY
TEMPLATE

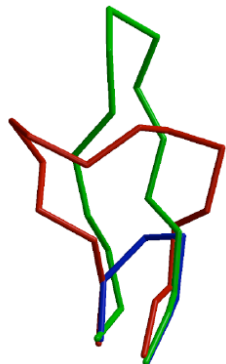
Incorrect template



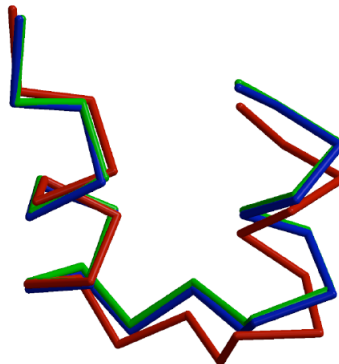
Misalignment



Region without a
template



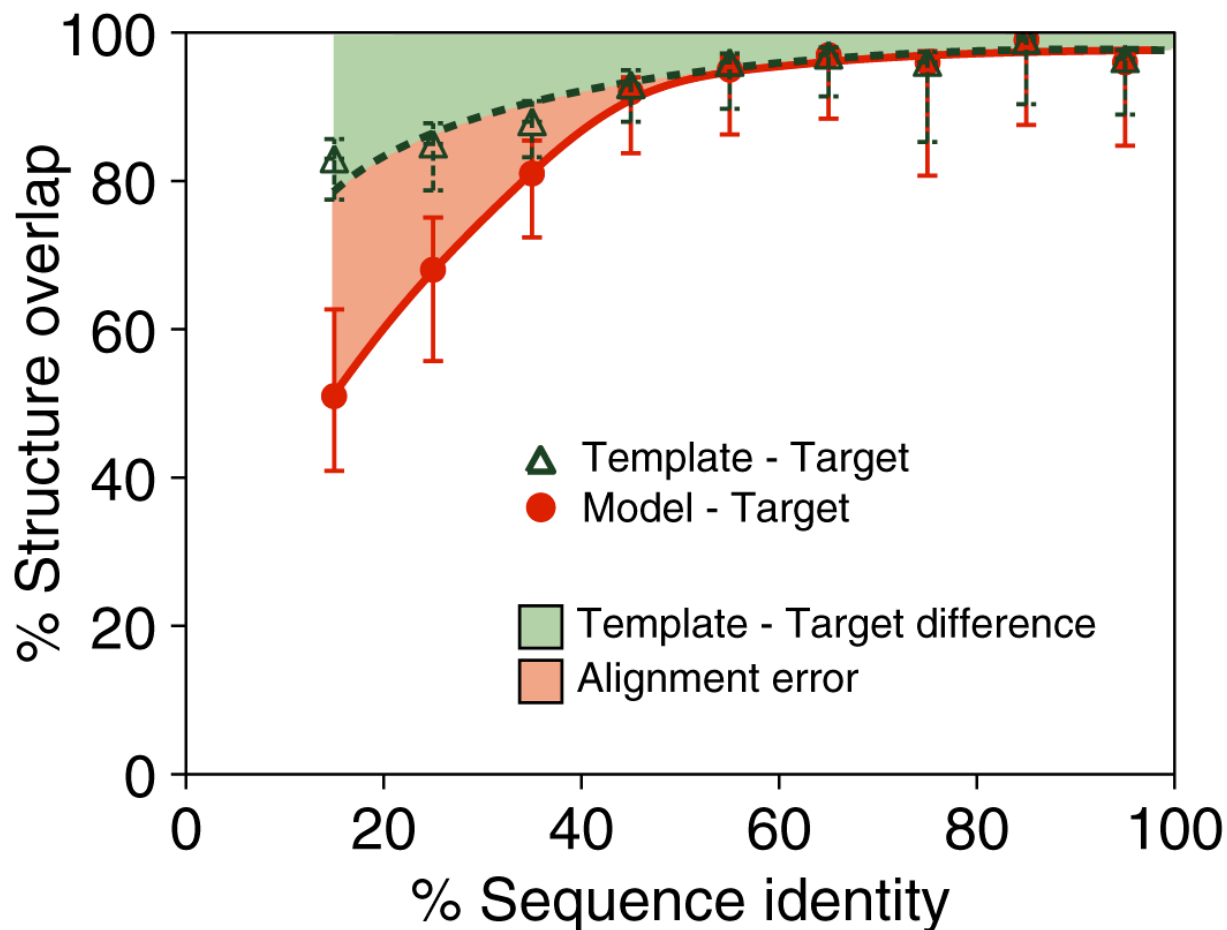
Distortion/shifts in
aligned regions



Sidechain packing



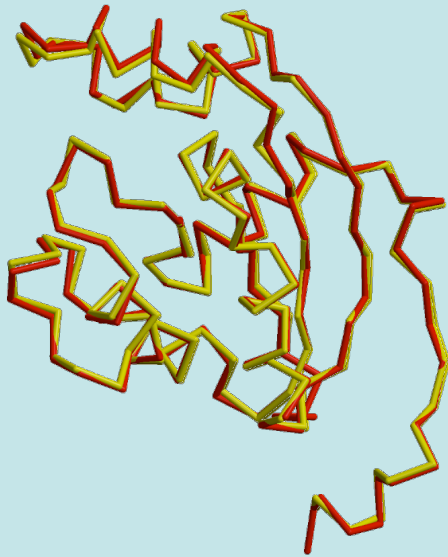
Model Accuracy as a Function of Target-Template Sequence Identity



Model Accuracy

HIGH ACCURACY

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

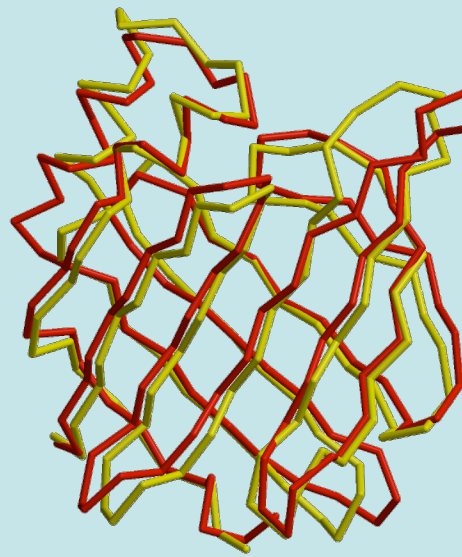


Sidechains
Core backbone
Loops

X-RAY / MODEL

MEDIUM ACCURACY

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



Sidechains
Core backbone
Loops
Alignment

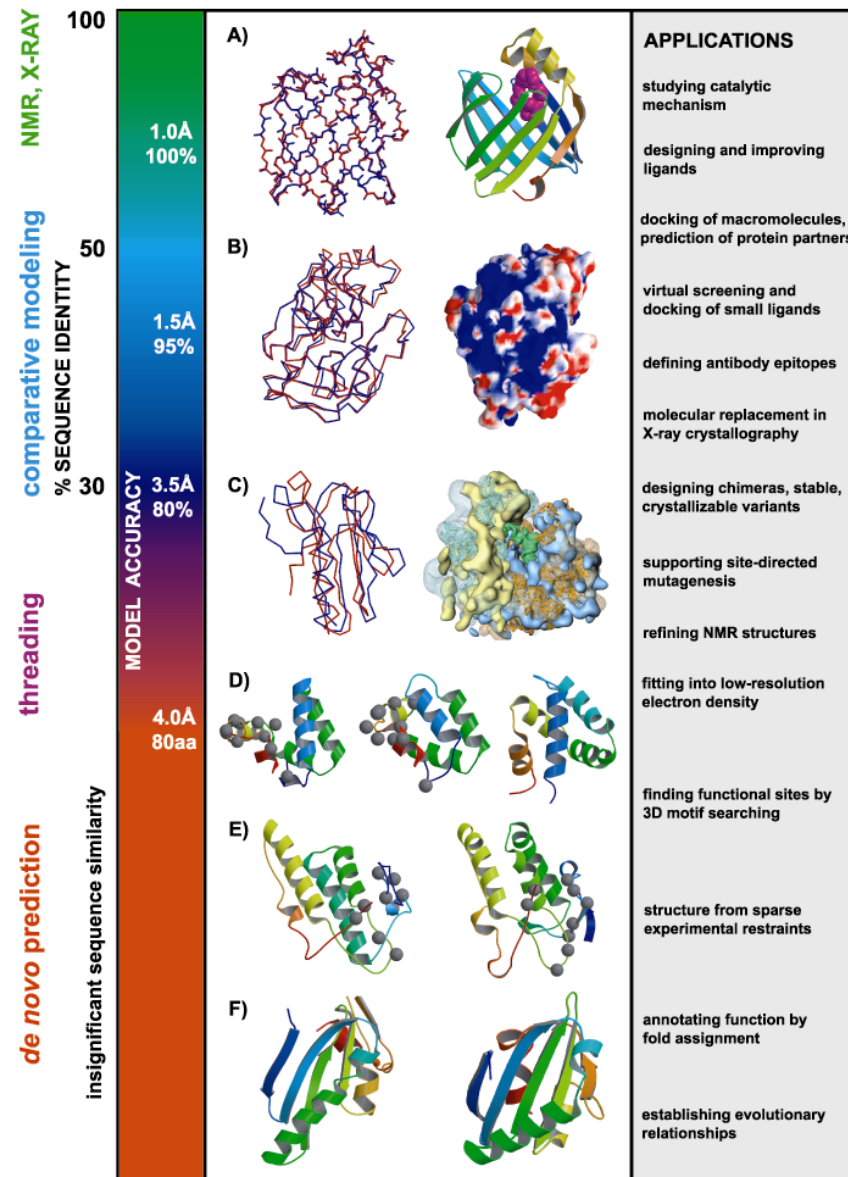
LOW ACCURACY

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



Sidechains
Core backbone
Loops
Alignment
Fold assignment

Applications of Protein Structure Models



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

Obtaining MODELLER and related information

- MODELLER (7v7) web page
- <http://www.salilab.org/modeller/>
 - Download Software (Linux/Windows/Mac/Solaris)
 - HTML Manual
 - Join Mailing List



Using MODELLER

- No GUI! 😞
- Controlled by command file (script) 😞😞
- Script is written in TOP language 😞😞😞
- TOP language is simple 😊😊😊😊

Using MODELLER

- **INPUT:**
 - Target Sequence (FASTA/PIR format)
 - Template Structure (PDB format)
 - TOP command file
- **OUTPUT:**
 - Target-Template Alignment
 - Model in PDB format
 - Other data

Modeling of BLBP

Input

- ✓ Target: Brain lipid-binding protein (BLBP)
- ✓ BLBP sequence in PIR (MODELLER) format:

```
>P1;blbp
sequence:blbp:::::::::
VDAFCATWKLTDSONFDEYMKALGVGFATRQVGNVTKPTVIIISQEGGKVIVIRTQCTFKNTEINFQLGEEFEETSI
DDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

- PSI-BLAST template search: Template: PDB file 1HMS:_

Modeling of BLBP

STEP 1: Align **blbp** and **1hms** sequences
TOP script for target-template alignment

```
READ_MODEL FILE = '1hms.pdb'  
SEQUENCE_TO_ALI ALIGN_CODES = '1hms'  
READ_ALIGNMENT FILE = 'blbp.seq', ALIGN_CODES = 'blbp', ADD_SEQUENCE = on  
ALIGN  
WRITE_ALIGNMENT FILE='blbp-1hms.ali', ALIGNMENT_FORMAT = 'PIR'  
WRITE_ALIGNMENT FILE='blbp-1hms.pap', ALIGNMENT_FORMAT = 'PAP'
```

Run by typing `mod align.top` directory where you have the TOP file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align **blbp** and **1hms** sequences
TOP script for target-template alignment

```
READ_MODEL FILE = '1hms.pdb'
SEQUENCE_TO_ALI ALIGN_CODES = '1hms'
READ_ALIGNMENT FILE = 'blbp.seq', ALIGN_CODES = 'blbp', ADD_SEQUENCE = on
ALIGN
WRITE_ALIGNMENT FILE='blbp-1hms.ali', ALIGNMENT_FORMAT = 'PIR'
WRITE_ALIGNMENT FILE='blbp-1hms.pap', ALIGNMENT_FORMAT = 'PAP'
```

Run by typing `mod align.top` directory where you have the TOP file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align **blbp** and **1hms** sequences
TOP script for target-template alignment

```
READ_MODEL FILE = '1hms.pdb'
SEQUENCE_TO_ALI ALIGN_CODES = '1hms'
READ_ALIGNMENT FILE = 'blbp.seq', ALIGN_CODES = 'blbp', ADD_SEQUENCE = on
ALIGN
WRITE_ALIGNMENT FILE 'blbp-1hms.ali', ALIGNMENT_FORMAT = 'PIR'
WRITE_ALIGNMENT FILE 'blbp-1hms.pap', ALIGNMENT_FORMAT = 'PAP'
```

Run by typing `mod align.top` directory where you have the TOP file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align **blbp** and **1hms** sequences
TOP script for target-template alignment

```
READ_MODEL FILE = '1hms.pdb'  
SEQUENCE_TO_ALI ALIGN_CODES = '1hms'  
READ_ALIGNMENT FILE = 'blbp.seq', ALIGN_CODES = 'blbp', ADD_SEQUENCE = on  
ALIGN  
WRITE_ALIGNMENT FILE='blbp-1hms.ali', ALIGNMENT_FORMAT = 'PIR'  
WRITE_ALIGNMENT FILE='blbp-1hms.pap', ALIGNMENT_FORMAT = 'PAP'
```

Run by typing `mod align.top` directory where you have the TOP file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align **blbp** and **1hms** sequences

Output

```
>P1;1hms
```

```
structureX:1hms: 1 : : 131 : :undefined:undefined:-1.00:-1.00
```

```
VDAFLGTWKLVD SKNFDDYMKSLGVGFATRQVASMTKPTTIEKNGDILTLKTHSTFKNTEISFKLGVEFDETTA  
DDRKVKSIIVTL DGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTRYEKE*
```

```
>P1;blbp
```

```
sequence:blbp: : : : : : 0.00: 0.00
```

```
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVIRTQCTFKNTEINFQLGEEFEETSI  
DDRNCKSVVRLDGD KLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

Modeling of BLBP

STEP 1: Align **blbp** and **1hms** sequences

Output

```
>P1;1hms
```

```
structureX:1hms: 1 : : 131 : :undefined:undefined:-1.00:-1.00
```

```
VDAFLGTWKLVD SKNFDDYMKSLGVGFATRQVASMTKPTTIEKNGDILTLKTHSTFKNTEISFKLGVEFDETTA  
DDRKVKSIIVTL DGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTRYEKE*
```

```
>P1;blbp
```

```
sequence:blbp: : : : : : 0.00: 0.00
```

```
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVIRTQCTFKNTEINFQLGEEFEETSI  
DDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

Modeling of BLBP

STEP 1: Align **blbp** and **1hms** sequences

Output

```
_aln.pos      10      20      30      40      50      60
1hms          VDAFLGTWKLVD SKNFDDYMKSLGVGFATRQVASMTKPTTIEKNGDILTLKTHSTFKNT
blbp          VDAFCATWKLTD SQNFDEYMKALGVGFATRQVG NVTKPTV IISQEGGKV VIRTQCTFKNT
_consrvd      ****  ****  **  ***  ***  *****  ****  **  *      *  *****

_aln.pos      70      80      90     100     110     120
1hms          EISFKLGVEFDETTADDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHG
blbp          EINFQLGEEFEETSIDDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFG
_consrvd      ** *  **  **  **  ***  **  *  ***  **  *  *****  **  ***  ***  *

_aln.pos      130
1hms          TAVCTR TYEKE
blbp          DIVAVR CYEKA
_consrvd      *  *  ***
```

Modeling of BLBP

STEP 2: Model the **blbp** structure using the alignment from step 1.

TOP script for model building

```
INCLUDE  
SET ALNFILE = 'blbp-1hms.ali'  
SET KNOWN = '1hms'  
SET SEQUENCE = 'blbp'  
SET STARTING_MODEL = 1  
SET ENDING_MODEL = 1  
CALL ROUTINE = 'model'
```

Run by typing `mod model.top.`

Check file `model.log`

Modeling of BLBP

STEP 2: Model the **blbp** structure using the alignment from step 1.

TOP script for model building

```
INCLUDE  
SET ALNFILE = 'blbp-1hms.ali'  
SET KNOWN = '1hms'  
SET SEQUENCE = 'blbp'  
SET STARTING_MODEL = 1  
SET ENDING_MODEL = 1  
CALL ROUTINE = 'model'
```

Run by typing `mod model.top.`

Check file `model.log`

Modeling of BLBP

STEP 2: Model the **blbp** structure using the alignment from step 1.

TOP script for model building

```
INCLUDE  
SET ALNFILE = 'blbp-1hms.ali'  
SET KNOWNs = '1hms'  
SET SEQUENCE = 'blbp'  
SET STARTING_MODEL = 1  
SET ENDING_MODEL = 1  
CALL ROUTINE = 'model'
```

Run by typing `mod model.top`

Check file `model.log`

Modeling of BLBP

STEP 2: Model the **blbp** structure using the alignment from step 1.

Output coordinates file

Model file → `blbp.B99990001`

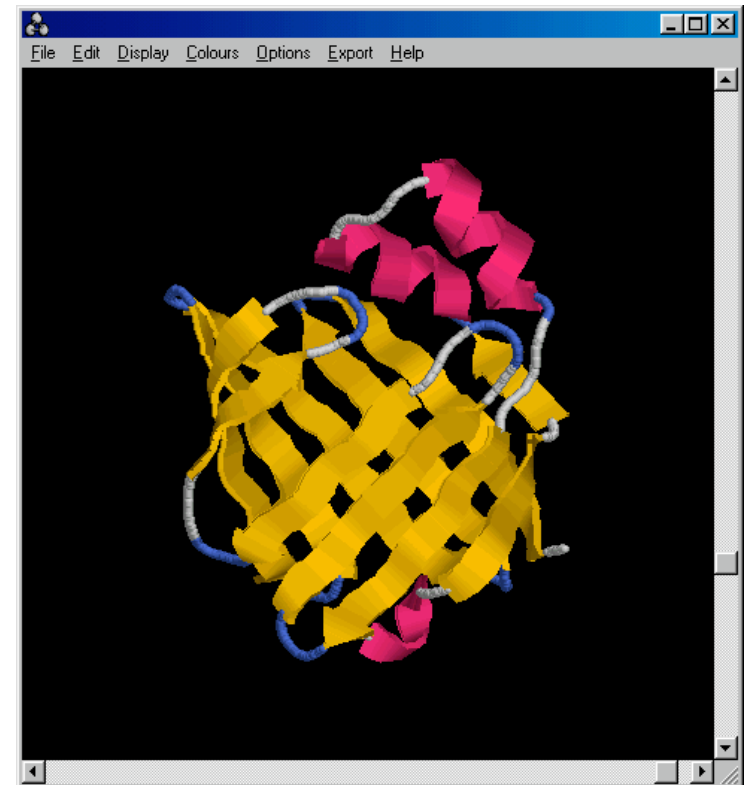
PDB file

Can be viewed with Chimera

<http://www.cgl.ucsf.edu/chimera/>

Rasmol

<http://www.bernstein-plus-sons.com/software/rasmol/>



http://www.salilab.org/bioinformatics_resources.shtml

Andrej Sali Lab

University of California, San Francisco
Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and
California Institute for Quantitative Biomedical Research

Bioinformatics Resources

Programs and World Wide Web servers useful in comparative modeling

Name	Type ^a	World Wide Web address ^b
DATABASES		
CATH	S	http://www.biochem.ucl.ac.uk/bsm/cath/
GenBank	S	http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html
GeneCensus	S	http://bioinfo.mbb.yale.edu/genome
ModBASE	S	http://salilab.org/modbase/
MSD	S	http://www.rcsb.org/databases.html
NCBI	S	http://www.ncbi.nlm.nih.gov/
PDB	S	http://www.rcsb.org/pdb/
PRESAGE	S	http://presage.berkeley.edu/
PSI	S	http://www.structuralgenomics.org/
Sacch3D	S	http://genome-www.stanford.edu/Sacch3D/
SCOP	S	http://scop.mrc-lmb.cam.ac.uk/scop/
TIGR	S	http://www.tigr.org/tdb/mdb/mdbcomplete.html
TrEMBL	S	http://ars.ebi.ac.uk/
FOLD ASSIGNMENT		
123D	S	http://www.lecb.ncifcrf.gov/~nicka/123D.html
3D-PSSM	S	http://www.bmm.icnet.uk/~3dpsm/html/ffrecog.html
BIOINBGU	S	http://www.cs.bgu.ac.il/~biolnbg/
BLAST	S	http://www.ncbi.nlm.nih.gov/BLAST/
DALI	S	http://www2.ebi.ac.uk/dali/
FASS	S	http://bioinformatics.burnham-inst.org/FFAS/index.html
FastA	S	http://www.ebi.ac.uk/fasta3/
FRSVR	S	http://www.doe-mbi.ucla.edu/~frsvr/preds/MG/MG.html
FUGUE	S	http://www-cryst.bioc.cam.ac.uk/~fugue
Koonin Group	S	ftp://ncbi.nlm.nih.gov/pub/koonin/FOLD3/index.html
LOOPP	S	http://www.tc.cornell.edu/reports/NIH/resource/CompBiologyTools/loopp/
PDB-BLAST/FASS	S	http://bioinformatics.ljcrf.edu/pdb_blast/
PHD, TOPITS	S	http://www.embl-heidelberg.de/predictprotein/predictprotein.html
PROFIT	P	http://www.came.sbg.ac.at
SAM-T99/T98	S	http://www.cse.ucsc.edu/research/compbio/WMH-apps/
THREADER	S	http://insulin.brunel.ac.uk/threader/threader.html
ToPLign/123D	S	http://caxtan.gmd.de/Genome/

Go to "http://salilab.org/our_resources.shtml"

References

Protein Structure Prediction:

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

Comparative Modeling:

Marti-Renom et al. Annu. Rev. Biophys. Biomol. Struct. 29, 291-325, 2000.
Marti-Renom et al. Current Protocols in Protein Science 1, 2.9.1-2.9.22, 2002.

MODELLER:

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

Structural Genomics:

Sali. Nat. Struct. Biol. 5, 1029, 1998.
Burley et al. Nat. Genet. 23, 151, 1999.
Sali & Kuriyan. TIBS 22, M20, 1999.
Sanchez et al. Nat. Str. Biol. 7, 986, 2000.
Baker & Sali. Science 294, 93-96, 2001.
Vitkup et al. Nat. Struct. Biol. 8, 559, 2001.