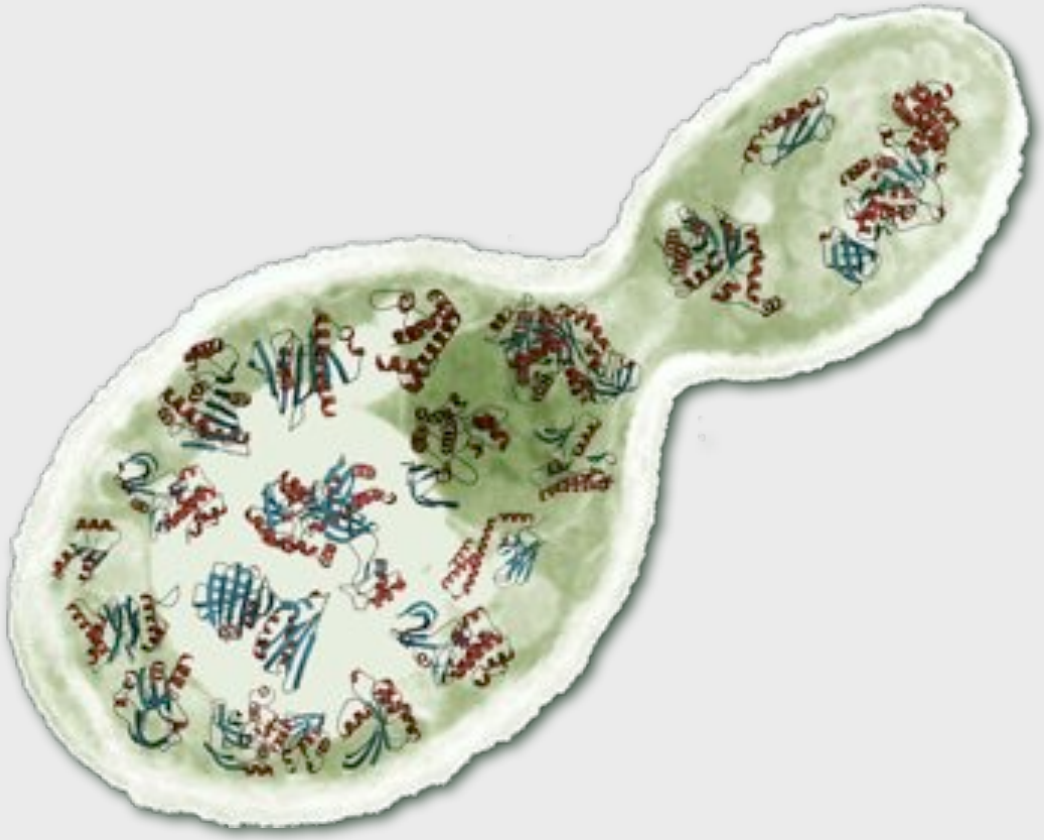
Comparative Protein Structure Prediction



Marc A. Marti-Renom
http://bioinfo.cipf.es/sgu/

Structural Genomics Unit
Bioinformatics Department
Prince Felipe Resarch Center (CIPF), Valencia, Spain

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

DISCLAIMER!

Name	Type[a]	World Wide Web address[b]
DATABASES		
CATH	S	http://www.biochem.ucl.ac.uk/bsm/cath/
DBAli	S	http://www.salilab.org/DBAli/
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/
PSI	S	http://www.nigms.nih.gov/psi/
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://srs.ebi.ac.uk/
FOLD ASSIGNMENT		
123D	S	http://123d.ncifcrf.gov/
3D-PSSM	S	http://www.sbg.bio.ic.ac.uk/~3dpssm/
BIOINBGU	S	http://www.cs.bgu.ac.il/~bioinbgu/
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://fold.doe-mbi.ucla.edu/
FUGUE	S	http://www-cryst.bioc.cam.ac.uk/~fugue/
LOOPP	S	http://ser-loopp.tc.cornell.edu/cbsu/loopp.htm
PDB-BLast/FASS	S	http://bioinformatics.ljcrf.edu/pdb_blast/
PHD, TOPITS	S	http://www.predictprotein.org/

http://salilab.org/bioinformatics_resources.shtml

Summary

- INTRO
- Structural Space
- Profile-Profile alignment
- MOULDER
- MODELLER example

Nomenclature

Homology: Sharing a common ancestor, may have similar or dissimilar functions

Similarity: Score that quantifies the degree of relationship between two sequences.

Identity: Fraction of identical aminoacids between two aligned sequences (case of similarity).

Target: Sequence corresponding to the protein to be modeled.

Template: 3D structure/s to be used during protein structure prediction.

Model: Predicted 3D structure of the target sequence.

protein prediction .vs. protein determination

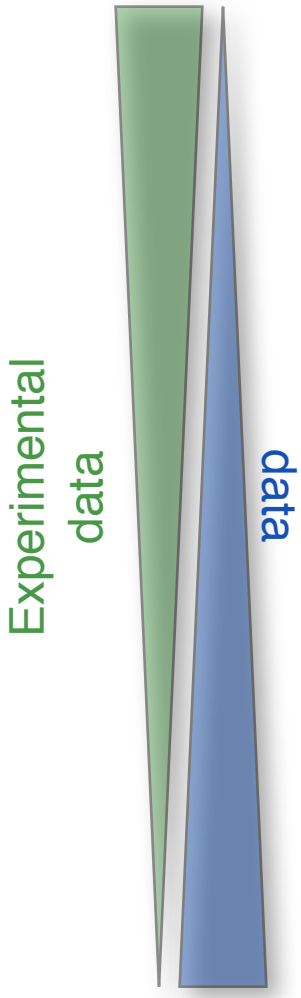
X-Ray

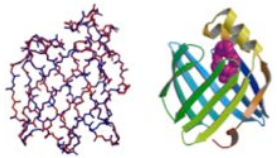
NMR

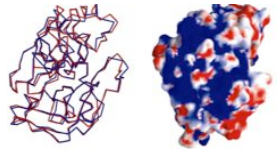
Comparative Modeling

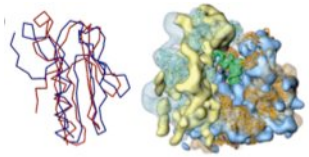
Threading

Ab-initio



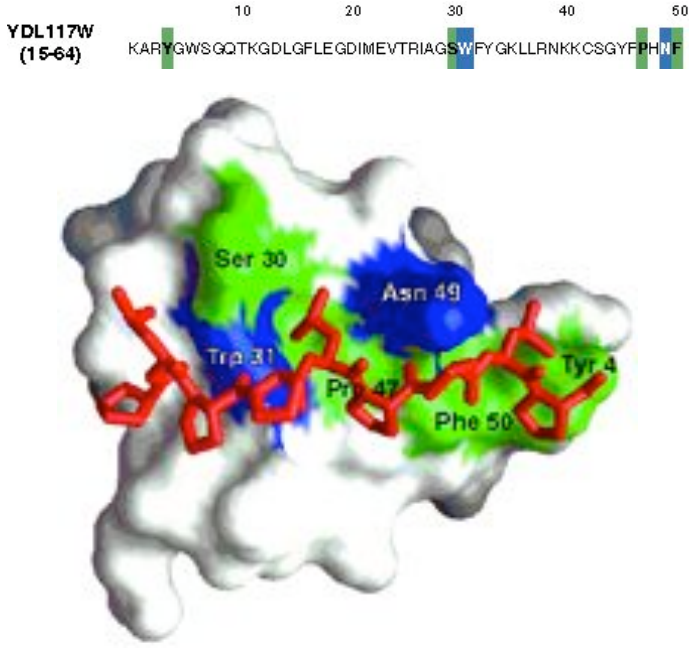






Why is it useful to know the structure of a protein, not only its sequence?

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

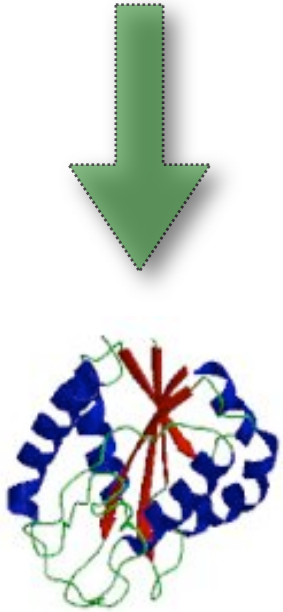


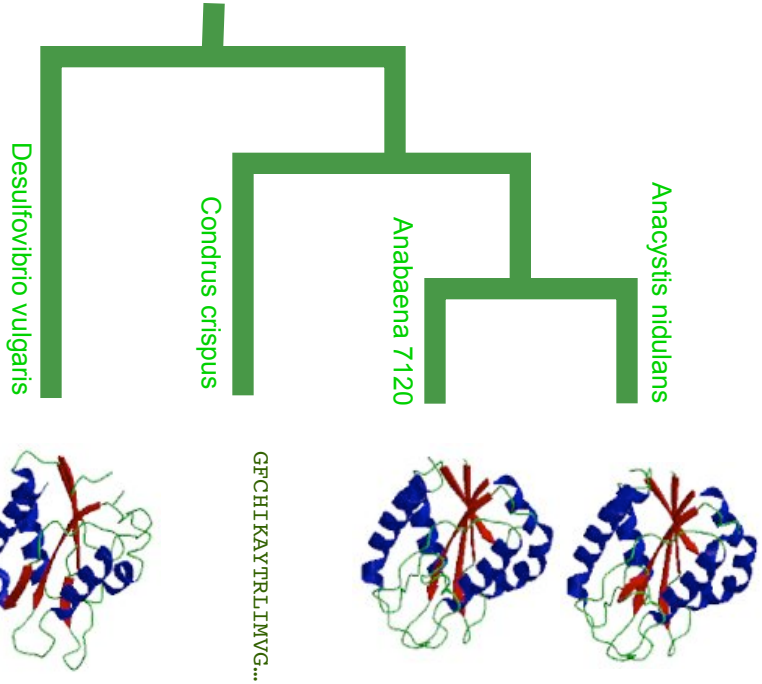
In addition, since evolution tends to conserve function and function depends more directly on structure than on sequence, **structure is more conserved in evolution than sequence**.

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

Principles of Protein Structure

GFCHIKAYTRLIMVG…





Folding

Ab initio prediction

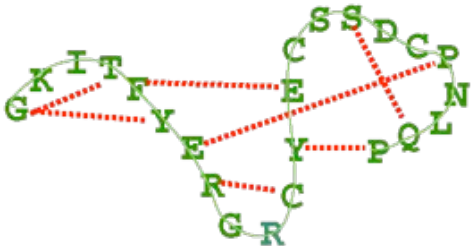
Evolution

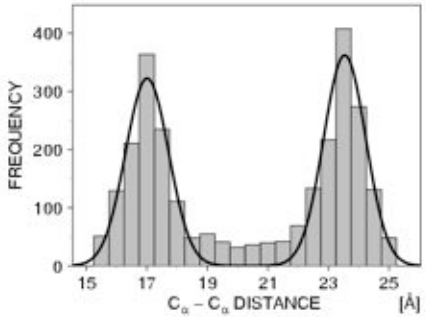
Threading
Comparative Modeling

Comparative Modeling by Satisfaction of Spatial Restraints (MODELLER)

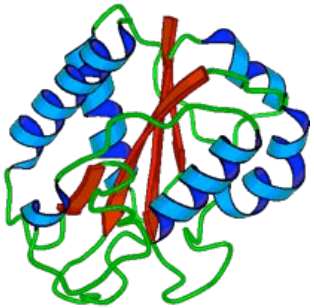
3D GKITFYERGFQGHCYESDC-NLQP…
SE GKITFYERG---RCYESDCPNLQP…

1. Extract spatial restraints





2. Satisfy spatial restraints

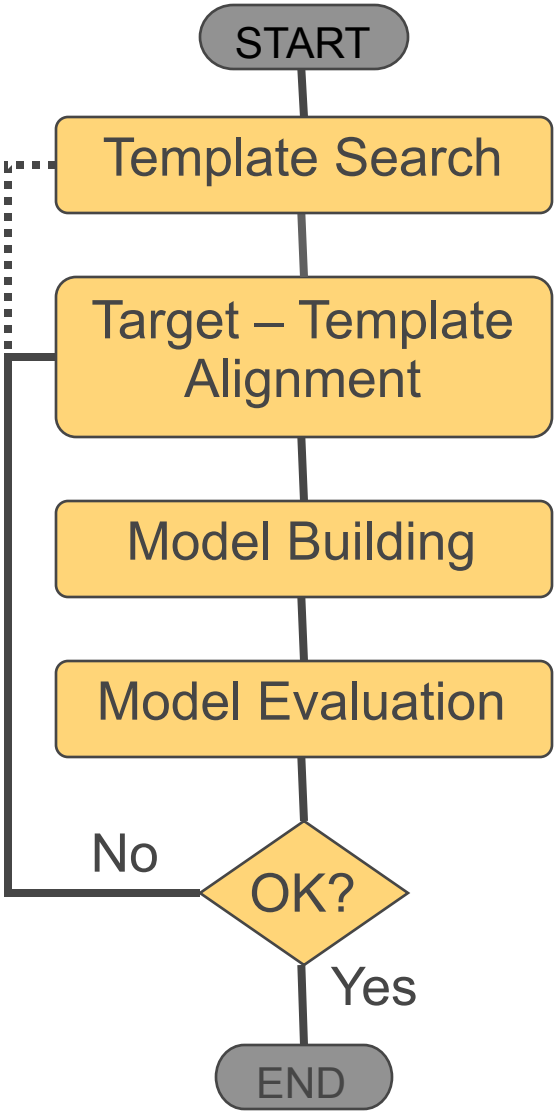


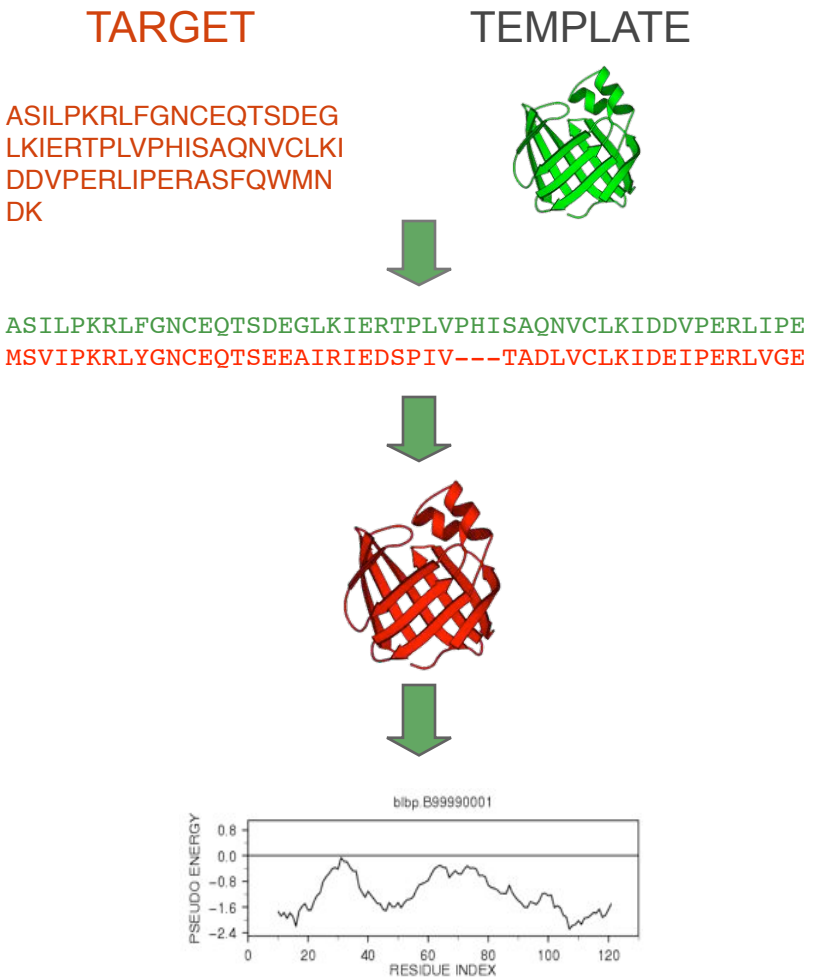
$$F(R) = \prod_i p_i(f_i/I)$$

http://www.salilab.org/modeller

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

Steps in Comparative Protein Structure Modeling



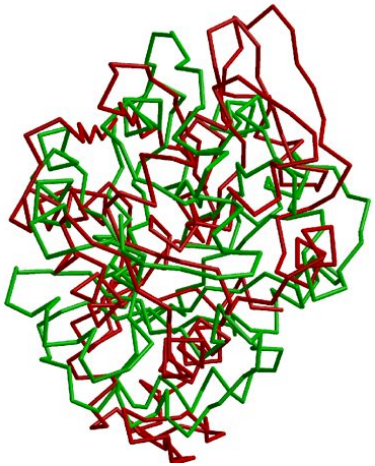


A. Šali, Curr. Opin. Biotech. 6, 437, 1995.
R. Sánchez & A. Šali, Curr. Opin. Str. Biol. 7, 206, 1997.
M. Marti et al. Ann. Rev. Biophys. Biomolec. Struct., 29, 291, 2000.

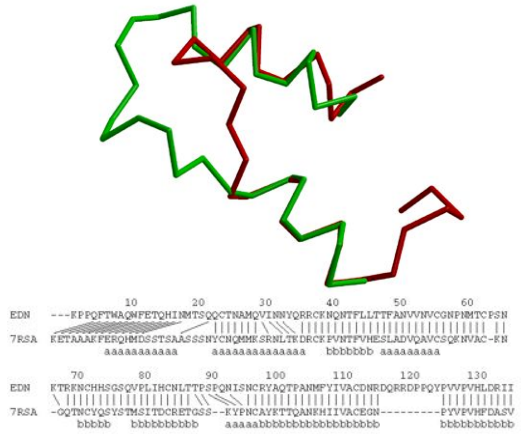
Typical errors in comparative models

MODEL
X-RAY
TEMPLATE

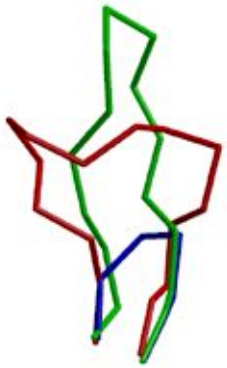
Incorrect template



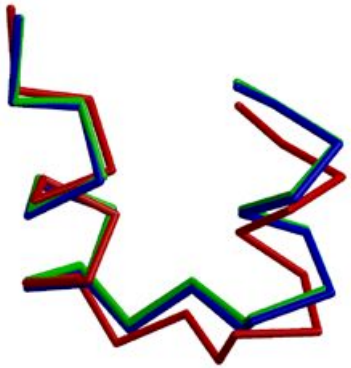
Misalignment



Region without a template



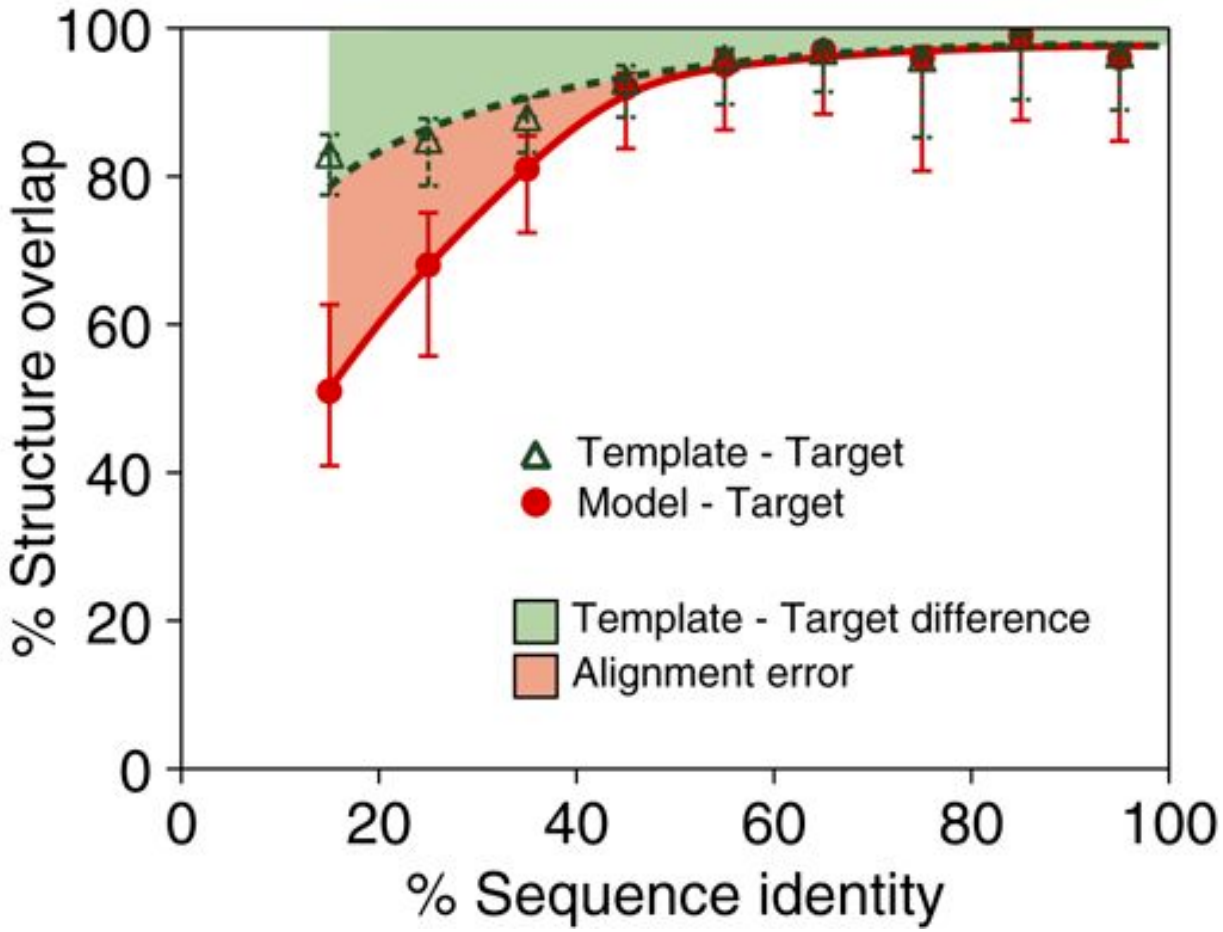
Distortion/shifts in aligned regions



Sidechain packing

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

Model Accuracy as a Function of Target-Template Sequence Identity

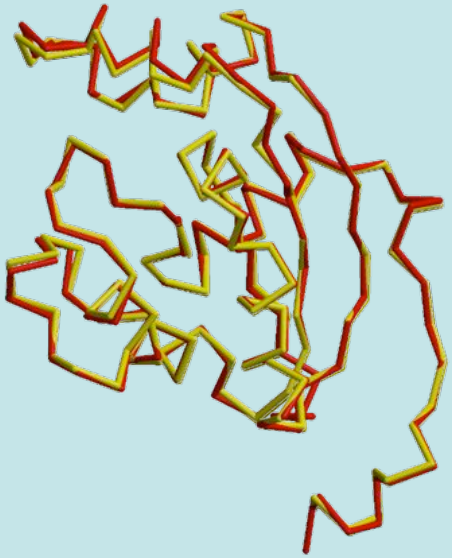


Sánchez, R., Šali, A. Proc Natl Acad Sci U S A. 95 pp13597-602. (1998).

Model Accuracy

HIGH ACCURACY

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



Sidechains
Core backbone
Loops

X-RAY / MODEL

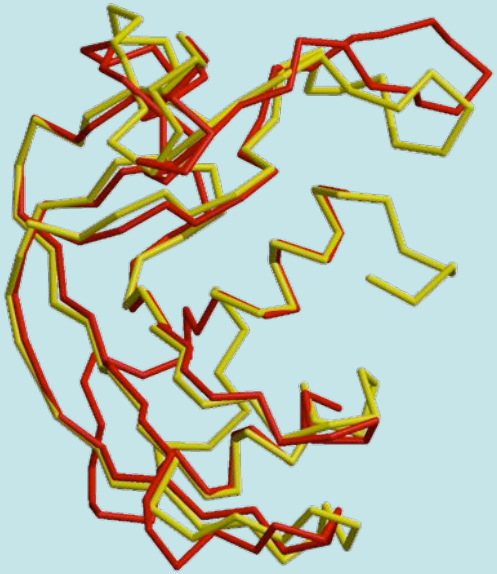
MEDIUM ACCURACY

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

Sidechains
Core backbone
Loops
Alignment

LOW ACCURACY

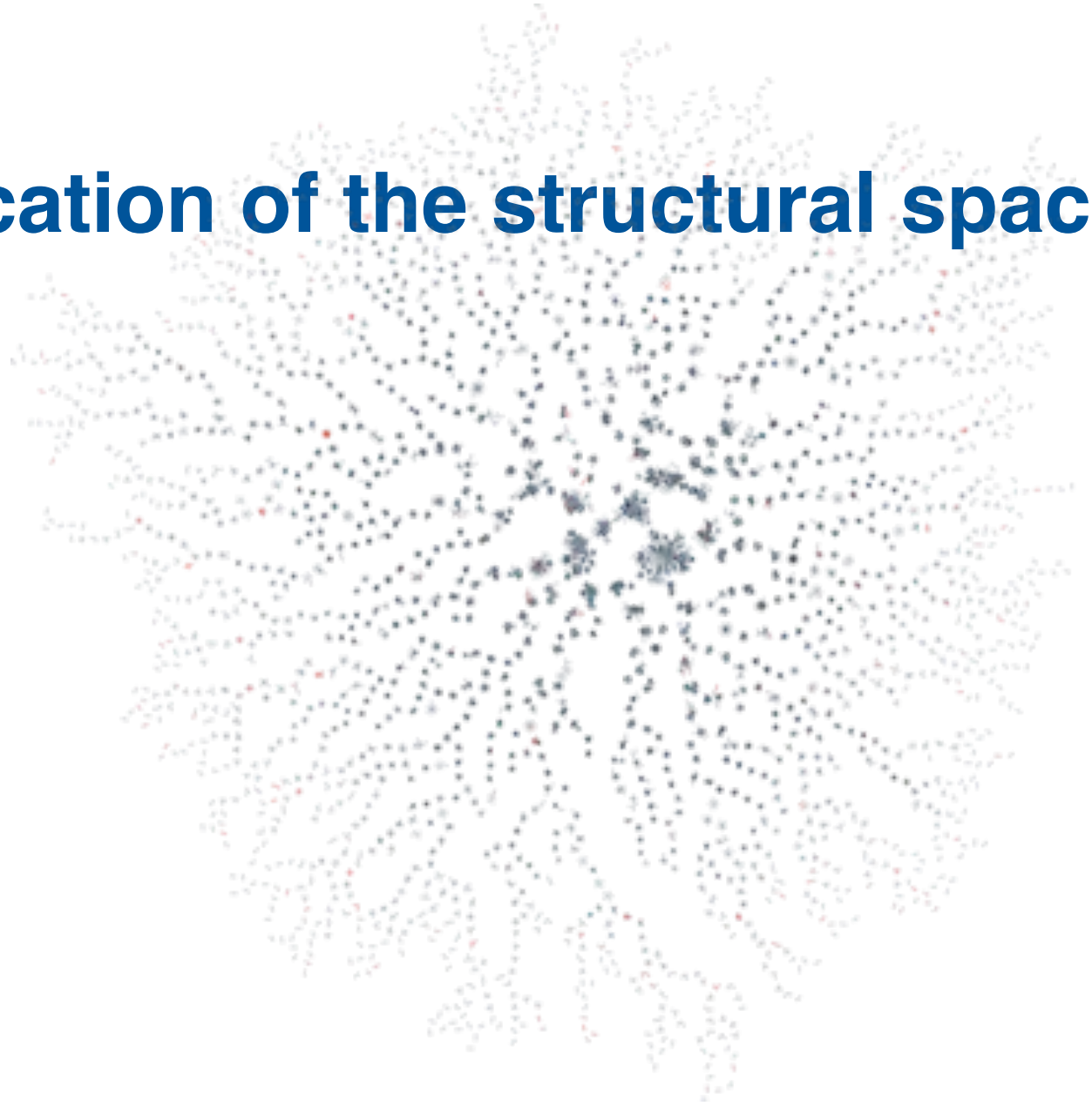
EDN
Seq id 33%
Cα equiv 90/134
RMSD 1.17Å



Sidechains
Core backbone
Loops
Alignment
Fold assignment

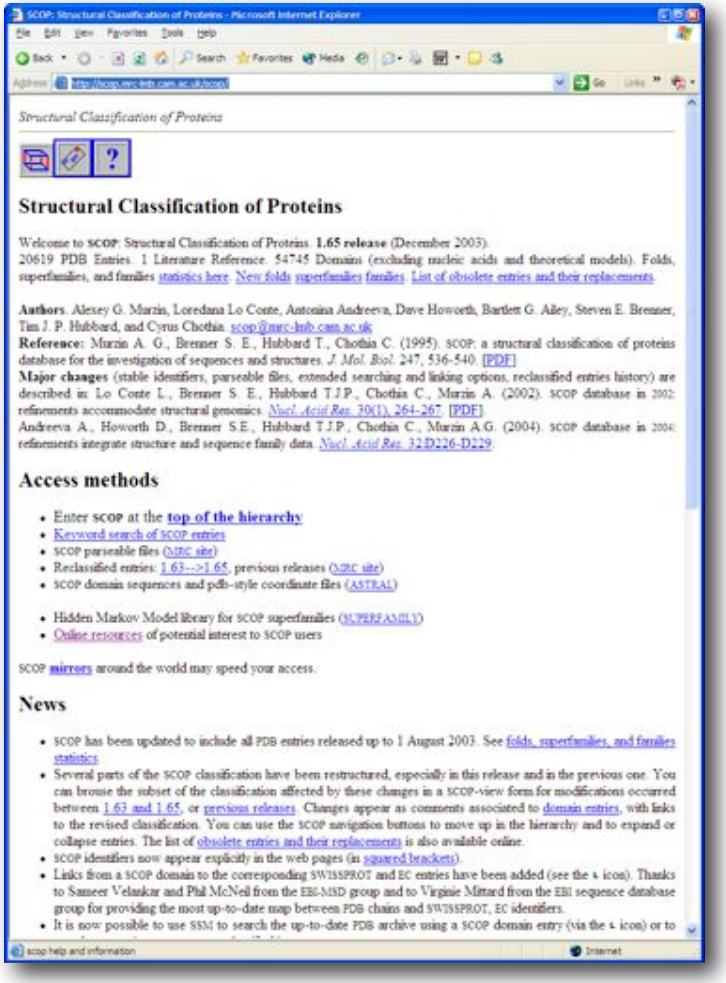
Marti-Renom et al. Annu.Rev.Biophys.Biomol.Struct. 29, 291-325, 2000.

Classification of the structural space



SCOP1.71 database

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



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

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

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	226	392	645
All beta proteins	149	300	594
Alpha and beta proteins (a/b)	134	221	661
Alpha and beta proteins (a+b)	286	424	753
Multi-domain proteins	48	48	64
Membrane and cell surface proteins	49	90	101
Small proteins	79	114	186
Total	971	1589	3004

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

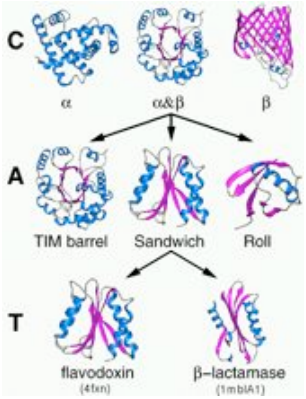
CATH3.1.0 database

http://www.cathdb.info

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

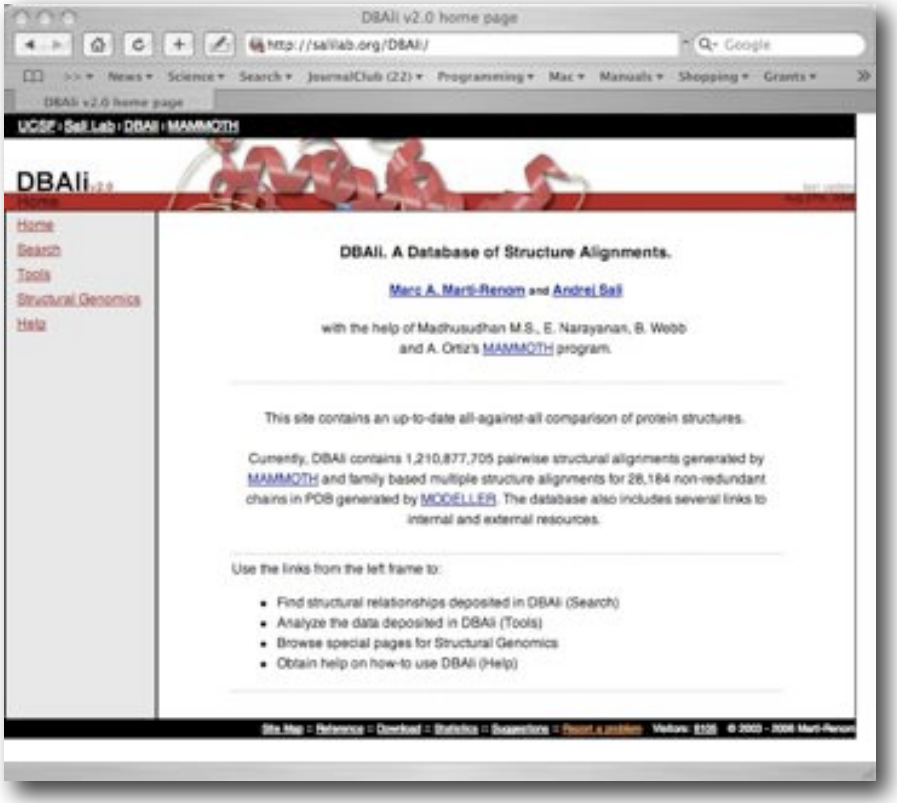


	A	T	H	S	O	L	I	D
Mainly Alpha	5	305	652	1850	2329	3001	5587	19729
Mainly Beta	20	191	415	1860	2531	3846	6503	25537
Alpha Beta	14	496	922	3922	5303	6659	12998	47193
Few Secondary Structures	1	92	102	162	200	275	403	1426
Total	**40**	**1084**	**2091**	**7794**	**10363**	**13781**	**25491**	**93885**

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

DBAli$_{v2.0}$ database

http://bioinfo.cipf.es/sgu/services/DBAli/
http://www.salilab.org/DBAli/



- ✓ Fully-automatic
- ✓ Data is kept up-to-date with PDB releases
- ✓ Tools for "on the fly" classification of families.
- ✓ Easy to navigate
- ✓ Provides tools for structure analysis

Does not provide a stable classification similar to that of CATH or SCOP

Pairwise structure alignments	
Last update:	February 15th, 2007
Number of chains:	88,276
Number of structure-structure comparisons:*	1,425,479,365
Multiple structure alignments	
Last update:	January 23rd, 2007
Number of representative chains:	30,900
Number of families:	11,615

Uses MAMMOTH for similarity detection

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

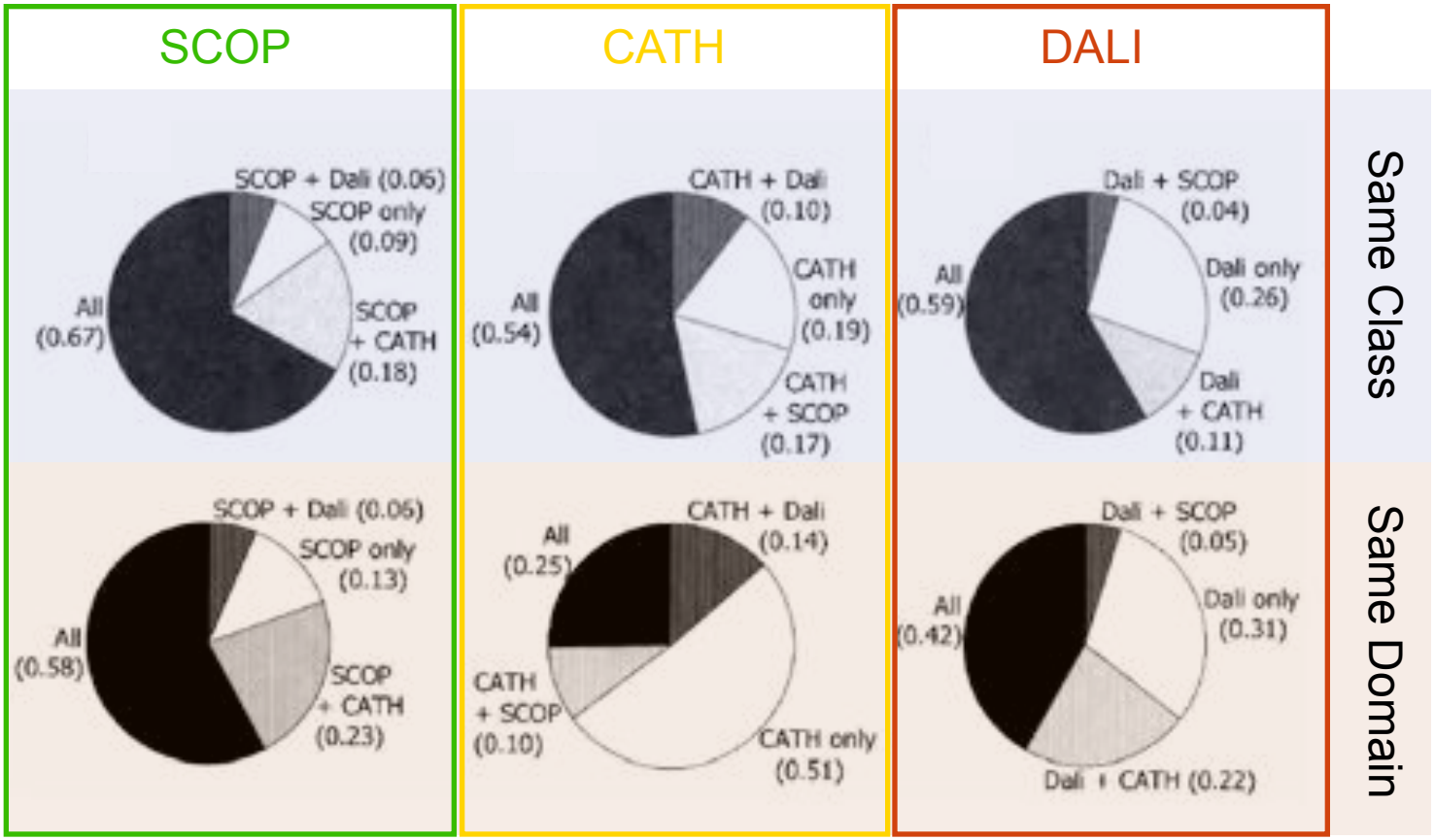
Ortiz AR, (2002) Protein Sci. 11 pp2606

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

Classification of the structural space

Not an easy task!

Domain definition AND domain classification



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

template search and template-target alignment

(build_profile & pp_scan)

Marti-Renom, et al. (2004) Prot. Sci. 13 pp1071
Narayanan, et al. in prepration

Preparation of Sequence Database

Generation of Alignment Scores

Construction of PSSM

- Position-Specific Scoring Matrix
- Data-dependent Pseudocounts
- Position-Based Sequence Weights

Assessment of Statistical Significance

Select Sequences Based on E-value

Create Multiple Alignment

1,803,406

LENGTH FILTER
(≤30aa / ≥3000aa)

1,774,668

SEG FILTER
(≥40aa / ≥40% of length)

1,460,796

SEQID FILTER

90%	799,201
80%	688,726
70%	609,238
60%	532,251

Preparation of Sequence Database

Generation of Alignment Scores

Construction of PSSM

- Position-Specific Scoring Matrix
- Data-dependent Pseudocounts
- Position-Based Sequence Weights

Assessment of Statistical Significance

Select Sequences Based on E-value

Create Multiple Alignment

		S	M	L	K	P
	0	0	0	0	0	0
T	0	S11	S12	S13	S14	S15
C	0	S21	S22	S23	S24	S25
I	0	S31	S32	S33	S34	**S35**
R	0	S41	S42	S43	S44	S45

Score-only Implementation of Smith-Waterman Dynamic Programing Algorithm

Miller & Myers, 1988

- Preparation of Sequence Database
- Generation of Alignment Scores
- Construction of PSSM
  - Position-Specific Scoring Matrix
    - Data-dependent Pseudocounts
    - Position-Based Sequence Weights
- Assessment of Statistical Significance
- Select Sequences Based on E-value
- Create Multiple Alignment

	V	L	S	E	G	...
A	450	400	500	400	450	
C	400	400	400	250	300	
D	300	250	450	550	440	
E	350	300	450	700	350	
F	400	450	350	300	300	
G	300	350	450	350	750	
H	300	300	450	450	300	
...						

a (row) i (column)

$$w_{ia} = \frac{1}{\lambda_u} \ln\left(\frac{p_{ia}}{P_a}\right)$$

where:

λ_u is a scaling factor

p_{ia} is the estimated probability of residue a to be found at position i

P_a is the background probability of residue a

Henikoff & Henikoff, 1994

- Preparation of Sequence Database
- Generation of Alignment Scores
- Construction of PSSM
  - Position-Specific Scoring Matrix
  - Data-dependent Pseudo-counts
    - Position-Based Sequence Weights
- Assessment of Statistical Significance
- Select Sequences Based on E-value
- Create Multiple Alignment

$$p_{ia} = \frac{\alpha_i}{\alpha_i + \beta} f_{ia} + \frac{\beta}{\alpha_i + \beta} \sum_{b=1}^{20} f_{ib} \frac{q_{ab}}{P_b}$$

where:

f_{ia}, f_{ib} are the observed weighted counts of residues a, b at position i

q_{ab} are the target frequencies implicit in the substitution matrix (BLOSUM62)

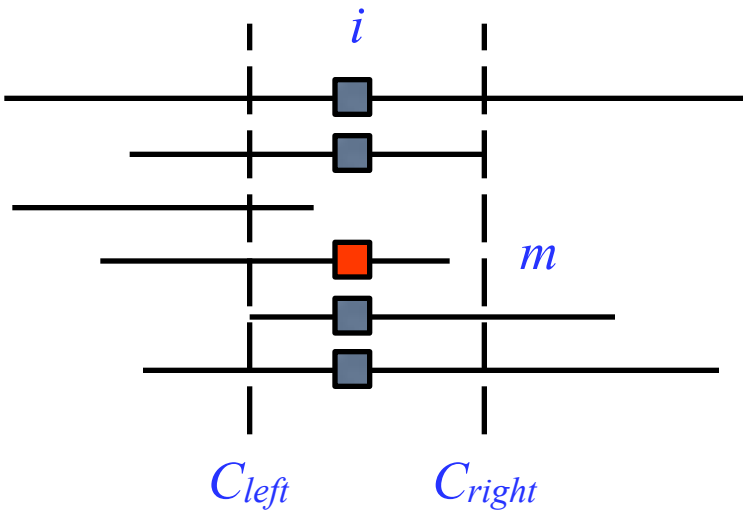
$$\alpha_i = N^i_{diff} - 1 \qquad \beta = 10$$

where:

N^i_{diff} is the number of different residues at i

Tatusov et.al., 1994; Altschul et.al., 1997

- Preparation of Sequence Database
- Generation of Alignment Scores
- Construction of PSSM
  - Position-Specific Scoring Matrix
  - Estimation of Target Frequencies
  - Position-Based Sequence Weights
- Assessment of Statistical Significance
- Select Sequences Based on E-value
- Create Multiple Alignment



$$W_m^i = \frac{1}{C_{right} - C_{left} + 1} \sum_{j=C_{left},C_{right}} \frac{1}{N_{diff}^j n_m^j}$$

where:

n^j_m is the number of times the residue in sequence m occurs in the column

Henikoff & Henikoff, 1994; Wang & Dunbrack, 2004

Preparation of Sequence Database

Generation of Alignment Scores

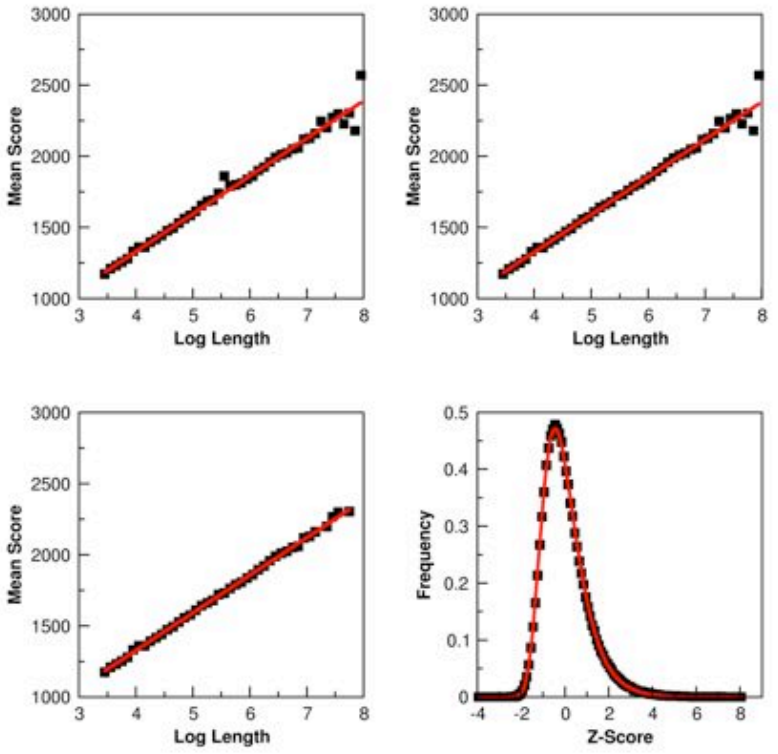
Construction of PSSM

- Position-Specific Scoring Matrix
- Estimation of Target Frequencies
- Position-Based Sequence Weights

Assessment of Statistical Significance

Select Sequences Based on E-value

Create Multiple Alignment



$$P(Z \geq z) = 1 - \exp\left(-e^{-z\pi/\sqrt{6}-\Gamma'(1)}\right)$$

$$E(Z) = P(Z)N$$

Pearson, 1998

Preparation of Sequence Database

Generation of Alignment Scores

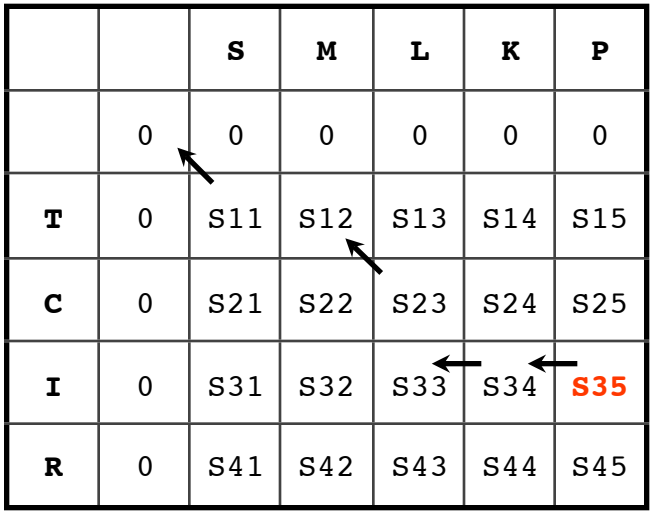
Construction of PSSM

- Position-Specific Scoring Matrix
- Estimation of Target Frequencies
- Position-Based Sequence Weights

Assessment of Statistical Significance

Re-align Significant Alignments

Create Multiple Alignment



Full Implementation of Smith-Waterman Dynamic Programing Algorithm

Gotoh, 1987

Preparation of Sequence Database

Generation of Alignment Scores

Construction of PSSM

- Position-Specific Scoring Matrix
- Estimation of Target Frequencies
- Position-Based Sequence Weights

Assessment of Statistical Significance

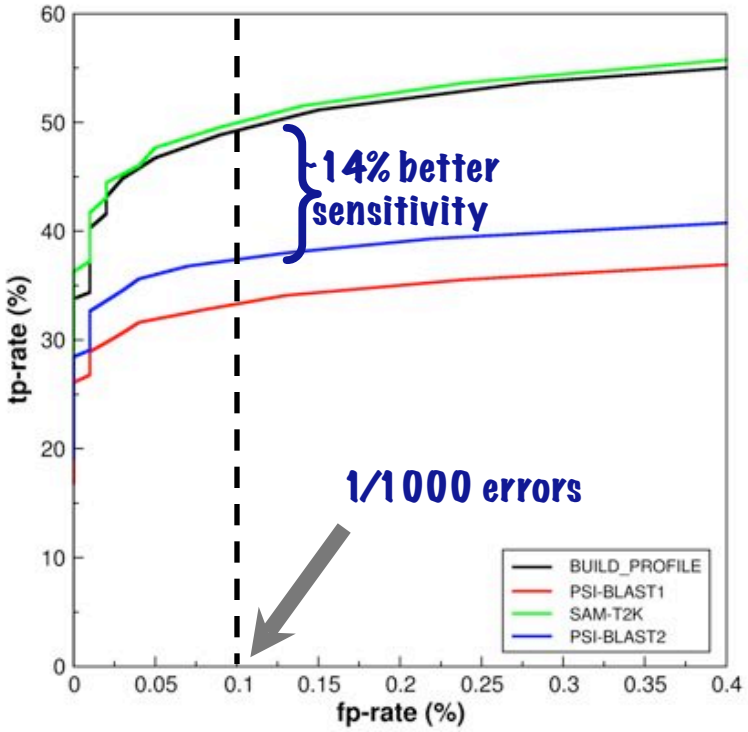
Re-align Significant Alignments

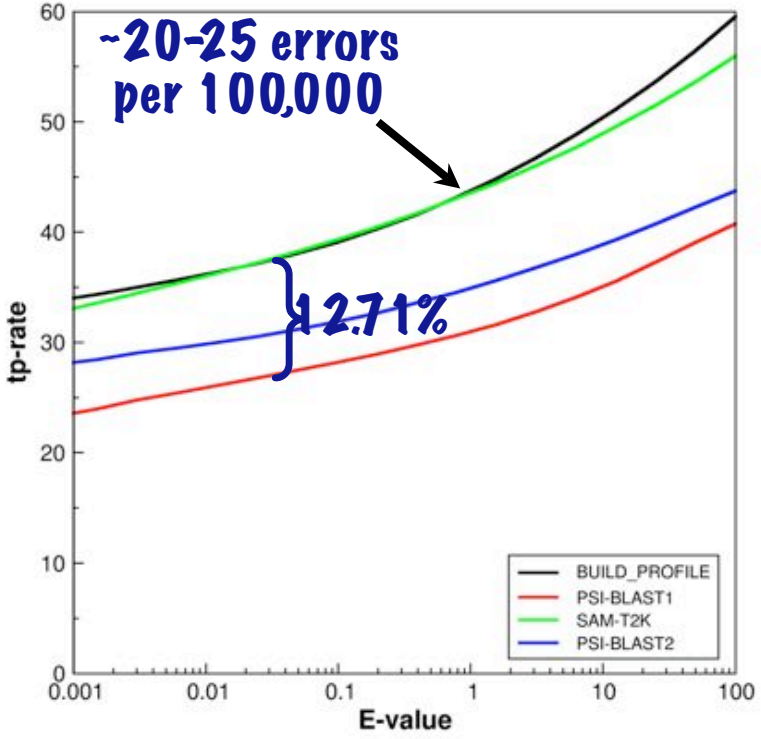
Create Multiple Alignment

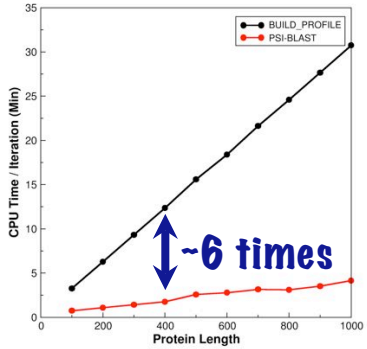
```
VLSEGEWQLVIWMQLC
-LSEGEWQLVTFLNLC
TLAEGEYQLI--LNLC
T--IAADGEYNLVALC
```

Iterate or 

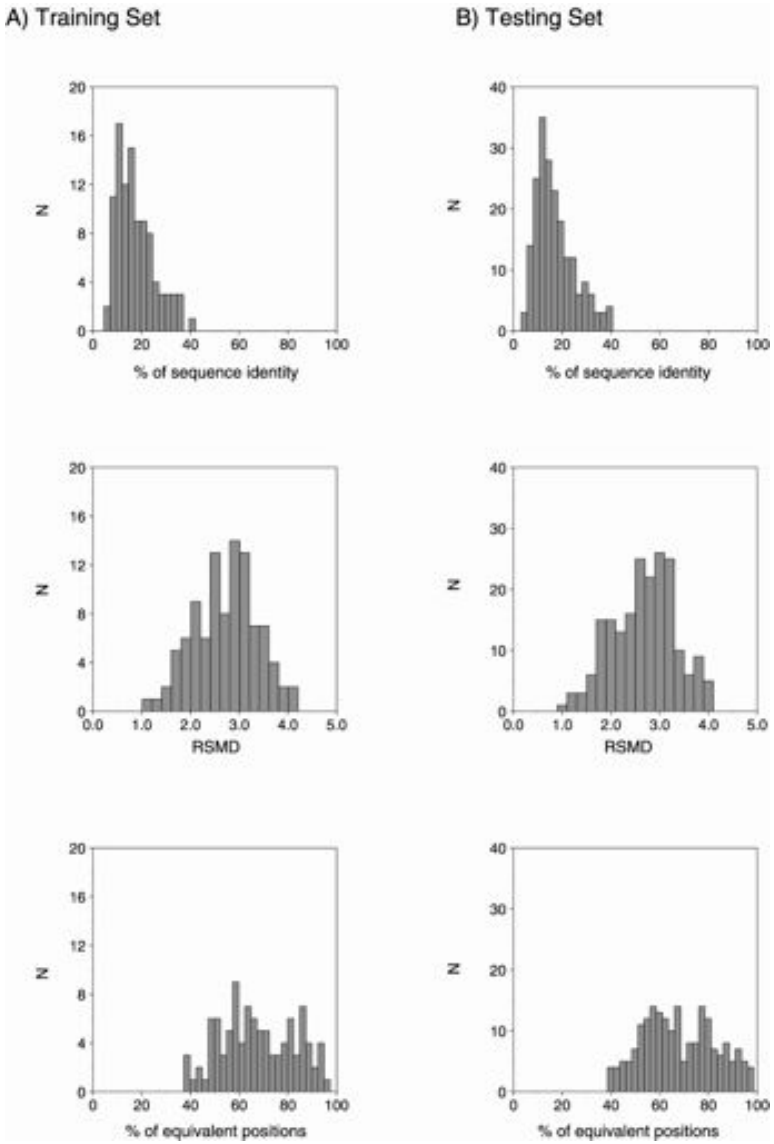
Only 26 (out of 6600) profiles showed corruption







PP_SCAN or profile-profile alignments



Seq.-Seq.

ALIGN: DP pairwise method

BLAST2SEQ: Local heuristic method

Seq.-Str.

SEA: Local structure prediction method

Prof.-Seq.

SAM: HMM method

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

LOBSTER: HHM + Phylogeny Method

Prof.-Prof.

CLUSTALW: DP multiple sequence method.

COMPASS: DP profile-profile method

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

PP_SCAN protocols

Profile generation

- PSI-Blast (PBP)
- Henikoff & Henikoff (HH)
- Henikoff & Henikoff + Similarity (HS)
- Henikoff & Henikoff substitution matrix (MAT)

Profile comparison

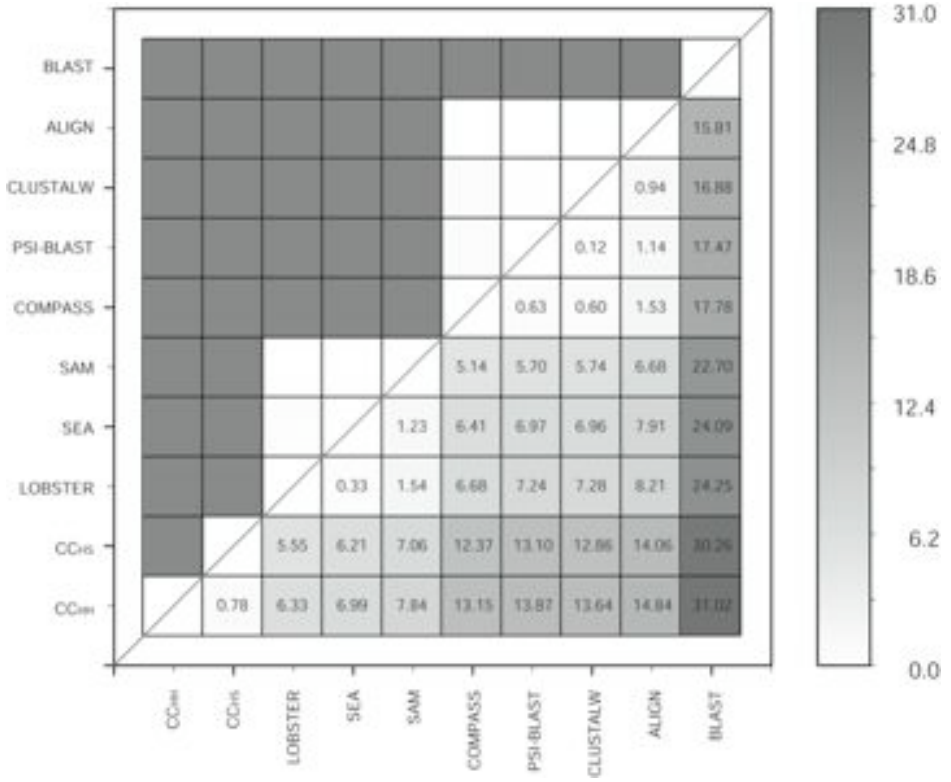
- Correlation coefficient (CC)
- Euclidean distance (ED)
- Dot product (DP)
- Jensen-Shannon distance (JS)
- Average value (Ave)

PP_SCAN protocols accuracy

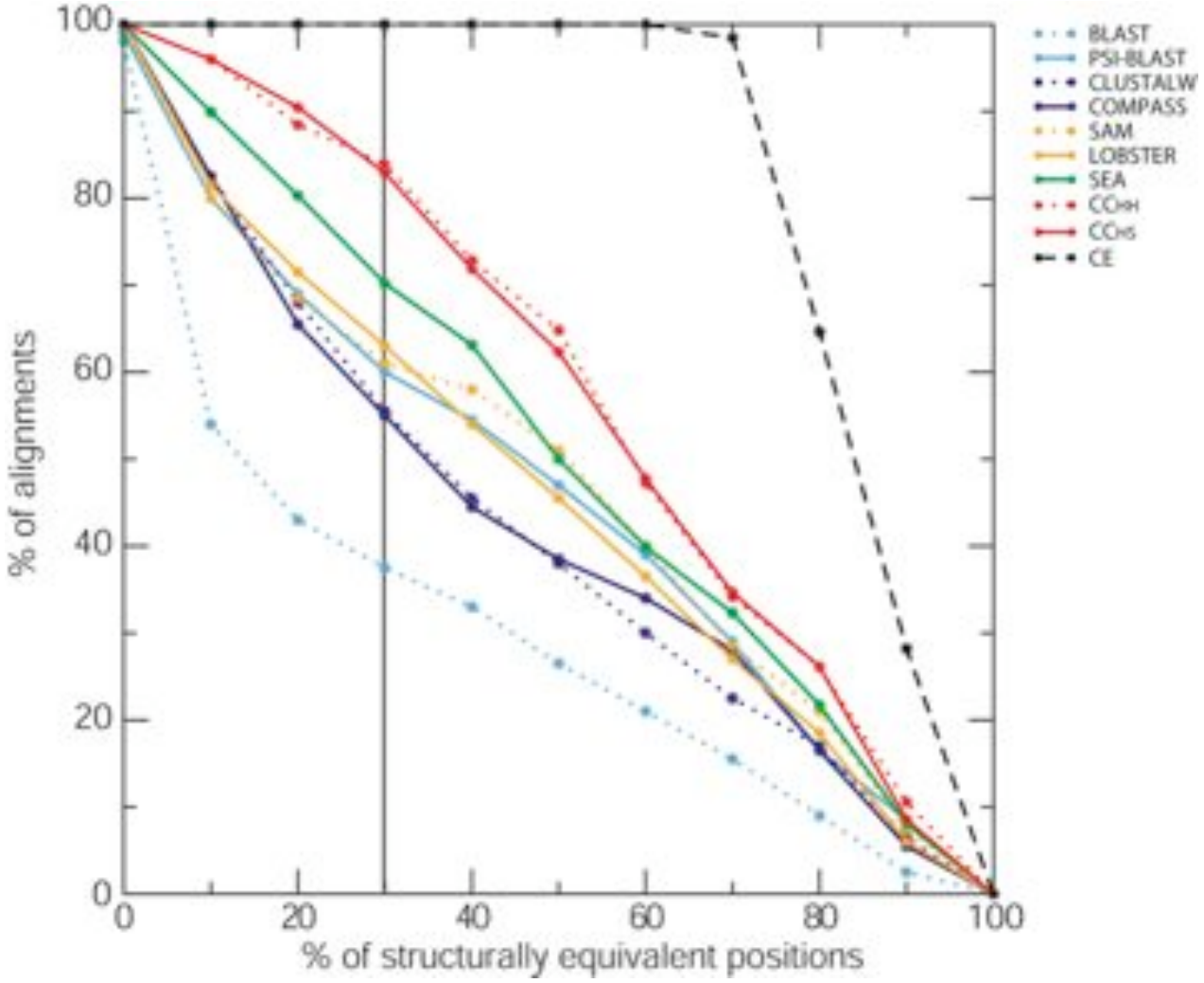
SALIGN protocol	CE overlap [%]	Shift score
CC$_{PBP}$	55 ± 23	0.61 ± 0.24
CC$_{HH}$	**56 ± 23**	**0.61 ± 0.24**
CC$_{HS}$	**56 ± 24**	**0.62 ± 0.23**
CC$_{MAT}$	51 ± 25	0.55 ± 0.27
ED$_{PBP}$	54 ± 24	0.60 ± 0.25
ED$_{HH}$	54 ± 24	0.59 ± 0.26
ED$_{HS}$	55 ± 24	0.59 ± 0.26
DP$_{PBP}$	55 ± 23	0.61 ± 0.24
DP$_{HH}$	56 ± 23	0.60 ± 0.25
DP$_{HS}$	55 ± 24	0.61 ± 0.24
JS$_{HH}$	53 ± 24	0.60 ± 0.24
JS$_{HS}$	54 ± 24	0.60 ± 0.24
Ave$_{MAT}$	49 ± 26	0.52 ± 0.29
TOP	62 ± 20	0.67 ± 0.20

PP_SCAN accuracy

Method	CE overlap	Shift score
CE	100 ± 0	1.00 ± 0.00
BLAST	26 ± 29	0.32 ± 0.33
PSI-BLAST	43 ± 31	0.48 ± 0.35
SAM	48 ± 26	0.50 ± 0.34
LOBSTER	50 ± 27	0.51 ± 0.32
SEA	49 ± 27	0.53 ± 0.29
ALIGN	42 ± 25	0.44 ± 0.28
CLUSTALW	43 ± 27	0.44 ± 0.31
COMPASS	43 ± 32	0.49 ± 0.35
CC_{HH}	**56 ± 23**	**0.61 ± 0.24**
CC_{HS}	**56 ± 24**	**0.62 ± 0.24**
TOP	62 ± 20	0.67 ± 0.20



PP_SCAN success



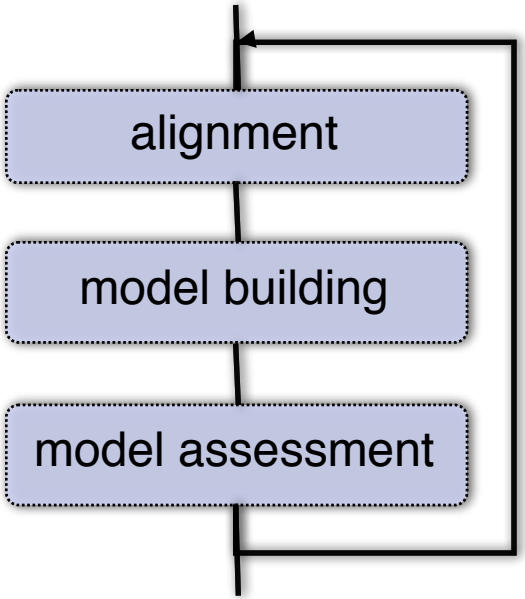
Alignment accuracy (CE overlap)

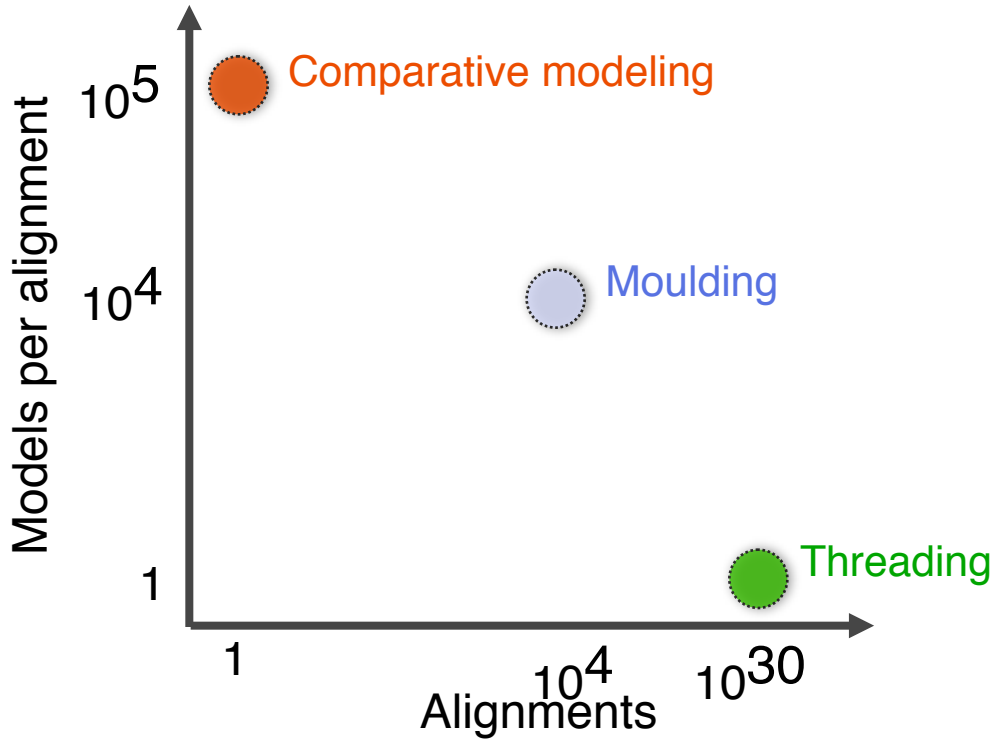
200 pairwise DBAli alignments

PSI-BLAST (sequence-profile alignment)	43%
SEA (local structure alignment)	49%
PP_SCAN (profile-profile alignment)	56%

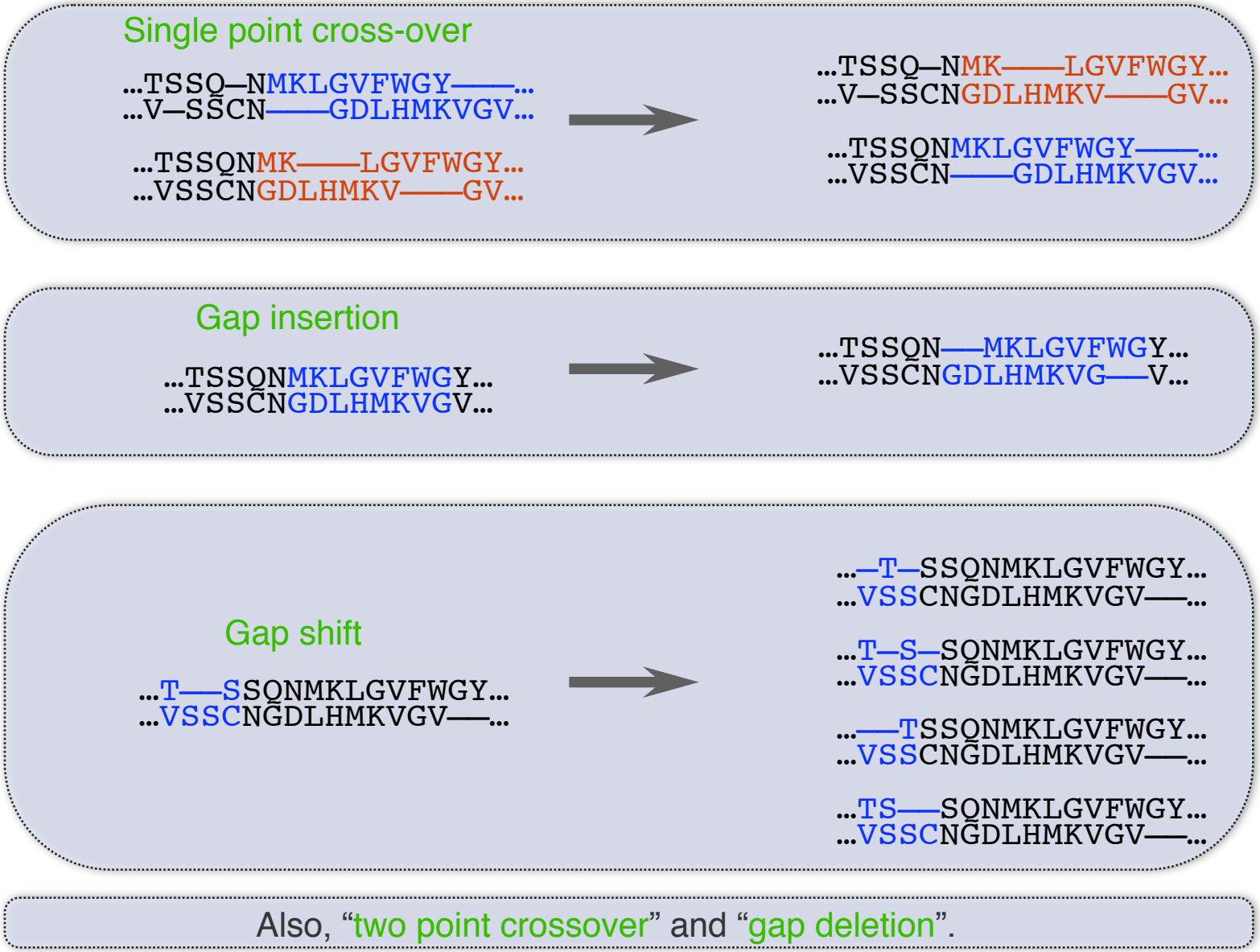
John, Sali (2003). NAR pp31 3982

Moulding: iterative alignment, model building, model assessment





Genetic algorithm operators



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).

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

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

μ … average score of all models

σ … standard deviation of the scores

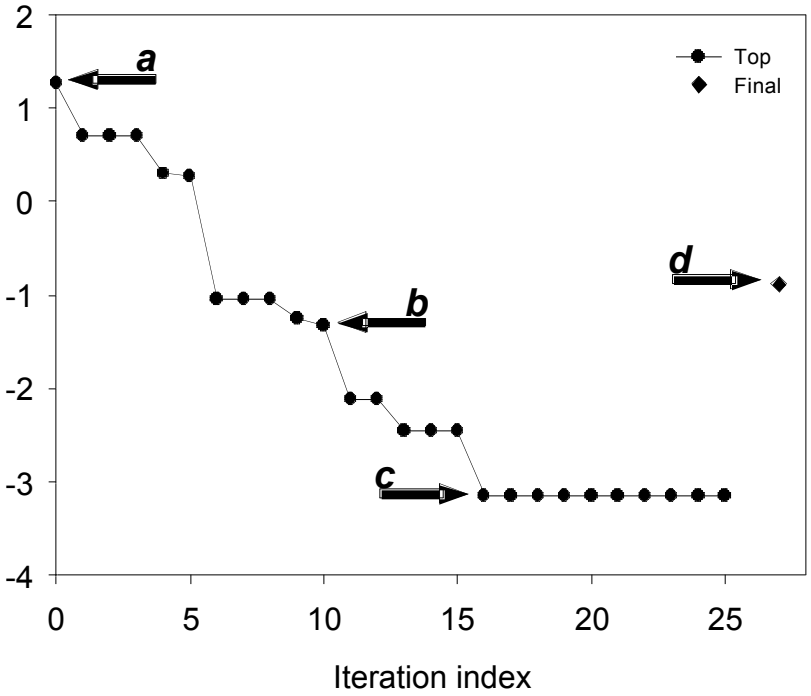
Benchmark with the “very difficult” test set

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

Application to a difficult modeling case

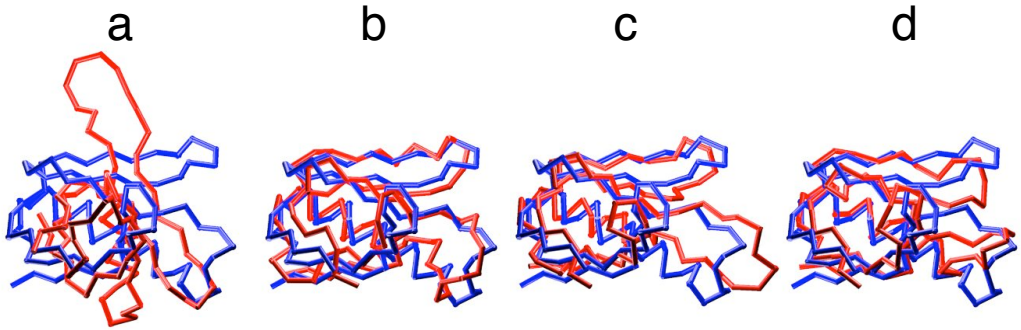
1BOV-1LTS



Sequence identity 4.4%

Initial model Cα RMSD 10.1Å

Final model Cα RMSD 3.6Å



Comparative Protein Structure Prediction
MODELLER tutorial

```
$>mod9v1 model.py
```

Marc A. Marti-Renom
http://bioinfo.cipf.es/sgu/

Structural Genomics Unit
Bioinformatics Department
Prince Felipe Resarch Center (CIPF), Valencia, Spain

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Obtaining MODELLER and related information

- MODELLER (9v1) web page
- http://www.salilab.org/modeller/
  - Download Software (Linux/Windows/Mac/Solaris)
  - HTML Manual
  - **Join Mailing List**


Using MODELLER

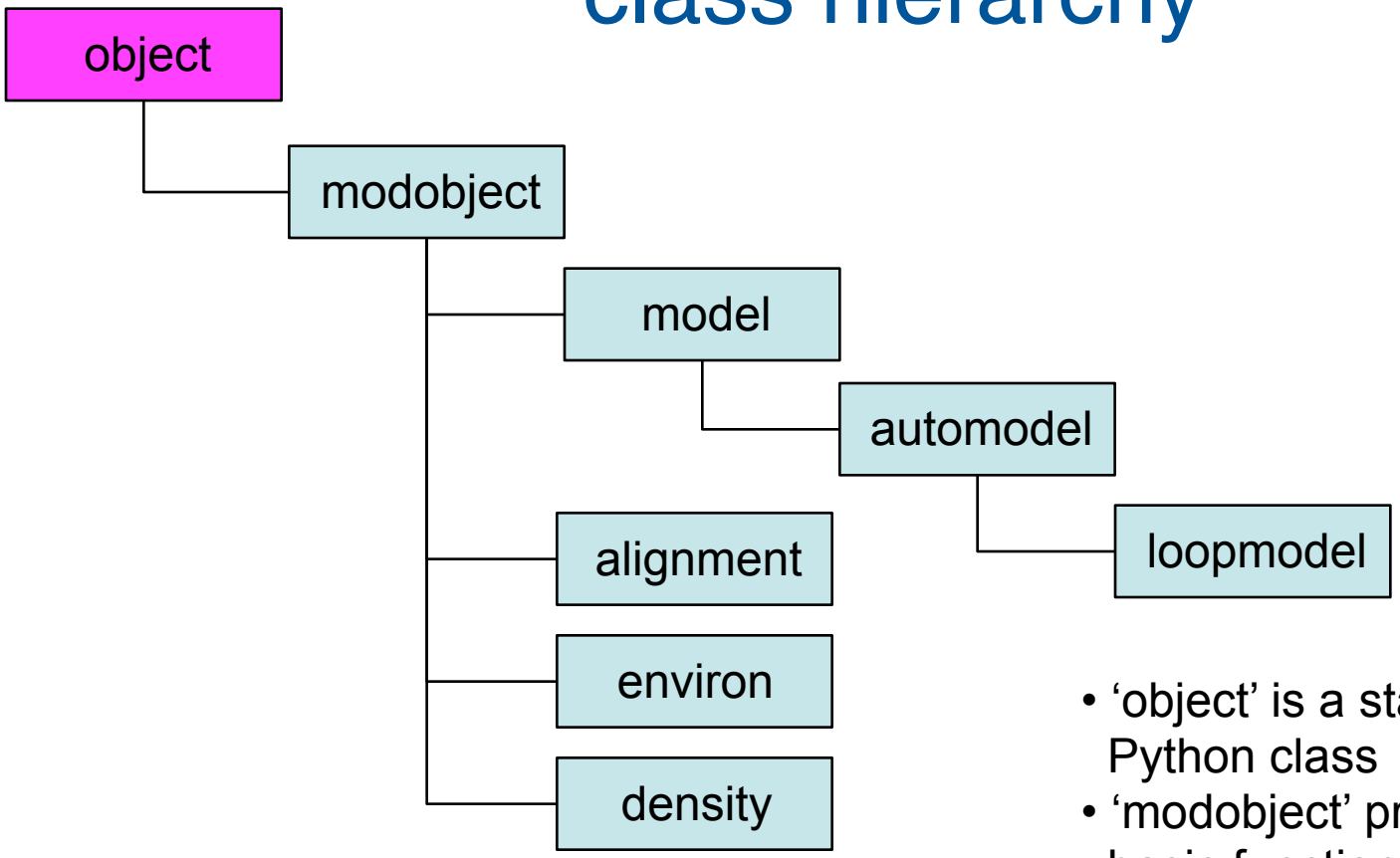
- No GUI! ☹
- Controlled by command file ☹☹
- Script is written in PYTHON language ☺
- You may know Python language is simple ☺☺

MODELLER 9v1
Python interface

- Modeller Python interface uses classes, e.g.:
  - 'alignment' *holds and manipulates aligned sequences*
  - 'model' *holds and manipulates protein models*
  - 'environ' *keeps the configuration of the environment*
  - 'profile' *holds and manipulates sequence profiles*
  - 'sequence_db' *is for sequence databases*
- These behave just like ordinary Python classes, but Modeller Fortran code is linked to them
- The Modeller data is automatically freed when the Python object is deleted (explicitly or implicitly)

MODELLER 8

class hierarchy



- ‘object’ is a standard Python class
- ‘modobject’ provides basic functions for most Modeller classes
- Not all classes are shown in this diagram

Using MODELLER

- INPUT:
  - Target Sequence (FASTA/PIR format)
  - Template Structure (PDB format)
  - Python 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:::::::
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEETSID
DRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Python script for target-template alignment

```
# Example for: alignment.align()

# This will read two sequences, align them, and write the alignment
# to a file:

log.verbose()
env = environ()

aln = alignment(env)
mdl = model(env, file='1hms')
aln.append_model(mdl, align_codes='1hms')
aln.append(file='blbp.seq', align_codes=('blbp'))

# The as1.sim.mat similarity matrix is used by default:
aln.align(gap_penalties_1d=(-600, -400))
aln.write(file='blbp-1hms.ali', alignment_format='PIR')
aln.write(file='blbp-1hms.pap', alignment_format='PAP')
```

Run by typing `mod9v1 align.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Python script for target-template alignment

```
# Example for: alignment.align()

# This will read two sequences, align them, and write the alignment
# to a file:

log.verbose()
env = environ()

aln = alignment(env)
mdl = model(env, file='1hms')
aln.append_model(mdl, align_codes='1hms')
aln.append(file='blbp.seq', align_codes=('blbp'))

# The as1.sim.mat similarity matrix is used by default:
aln.align(gap_penalties_1d=(-600, -400))
aln.write(file='blbp-1hms.ali', alignment_format='PIR')
aln.write(file='blbp-1hms.pap', alignment_format='PAP')
```

Run by typing `mod9v1 align.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Python script for target-template alignment

```
# Example for: alignment.align()

# This will read two sequences, align them, and write the alignment
# to a file:

log.verbose()
env = environ()

aln = alignment(env)
mdl = model(env, file='1hms')
aln.append_model(mdl, align_codes='1hms')
aln.append(file='blbp.seq', align_codes=('blbp'))

# The as1.sim.mat similarity matrix is used by default:
aln.align(gap_penalties_1d=(-600, -400))
aln.write(file='blbp-1hms.ali', alignment_format='PIR')
aln.write(file='blbp-1hms.pap', alignment_format='PAP')
```

Run by typing `mod9v1 align.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Python script for target-template alignment

```
# Example for: alignment.align()

# This will read two sequences, align them, and write the alignment
# to a file:

log.verbose()
env = environ()

aln = alignment(env)
mdl = model(env, file='1hms')
aln.append_model(mdl, align_codes='1hms')
aln.append(file='blbp.seq', align_codes=('blbp'))

# The as1.sim.mat similarity matrix is used by default:
aln.align(gap_penalties_1d=(-600, -400))
aln.write(file='blbp-1hms.ali', alignment_format='PIR')
aln.write(file='blbp-1hms.pap', alignment_format='PAP')
```

Run by typing `mod9v1 align.py` in the directory where you have the python 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
VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGVEFDETTA
DDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTYEKE*
>P1;blbp
sequence:blbp:      : :      : : : : 0.00: 0.00
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEETSI
DDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Output

```
>P1;1hms
structureX:1hms:    1 : : 131 : :undefined:undefined:-1.00:-1.00
VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGVEFDETTA
DDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTYEKE*
>P1;blbp
sequence:blbp:      : :      : : : : 0.00: 0.00
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEETSI
DDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Output

```
 _aln.pos          10         20         30         40         50         60
1hms       VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGV
blbp       VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGE
 _consrvd  ****   **** ** *** *** **********    **** **    *       *   ******* * **


 _aln.p    70         80         90         100        110        120        130
1hms       EFDETTADDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTYEKE
blbp       EFEETSIDDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA
 _consrvd  ** **   ***   ** * *** ** * ***** **    **   ***    *** *   *   * ***
```

Modeling of BLBP

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

Python script for model building

```
# Homology modelling by the automodel class
from modeller.automodel import *    # Load the automodel class
log.verbose()                       # request verbose output
env = environ()                     # create a new MODELLER environment

# directories for input atom files
env.io.atom_files_directory = './:../atom_files'

a = automodel(env,
              alnfile  = 'blbp-1hms.ali',     # alignment filename
              knowns   = '1hms',              # codes of the templates
              sequence = 'blbp')              # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  = 1                 # index of the last model
                                    # (determines how many models to calculate)
a.make()                            # do the actual homology modelling
```

Run by typing `mod9v1 model.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

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

Python script for model building

```
# Homology modelling by the automodel class
from modeller.automodel import *    # Load the automodel class
log.verbose()                       # request verbose output
env = environ()                     # create a new MODELLER environment

# directories for input atom files
env.io.atom_files_directory = './:../atom_files'

a = automodel(env,
              alnfile  = 'blbp-1hms.ali',     # alignment filename
              knowns   = '1hms',              # codes of the templates
              sequence = 'blbp')              # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  = 1                 # index of the last model
                                    # (determines how many models to calculate)
a.make()                            # do the actual homology modelling
```

Run by typing `mod9v1 model.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

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

Python script for model building

```
# Homology modelling by the automodel class
from modeller.automodel import *    # Load the automodel class
log.verbose()                       # request verbose output
env = environ()                     # create a new MODELLER environment

# directories for input atom files
env.io.atom_files_directory = './:../atom_files'

a = automodel(env,
              alnfile  = 'blbp-1hms.ali',     # alignment filename
              knowns   = '1hms',              # codes of the templates
              sequence = 'blbp')              # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  = 1                 # index of the last model
                                    # (determines how many models to calculate)
a.make()                            # do the actual homology modelling
```

Run by typing `mod9v1 model.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

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

Python script for model building

PDB file

Can be viewed with Chimera

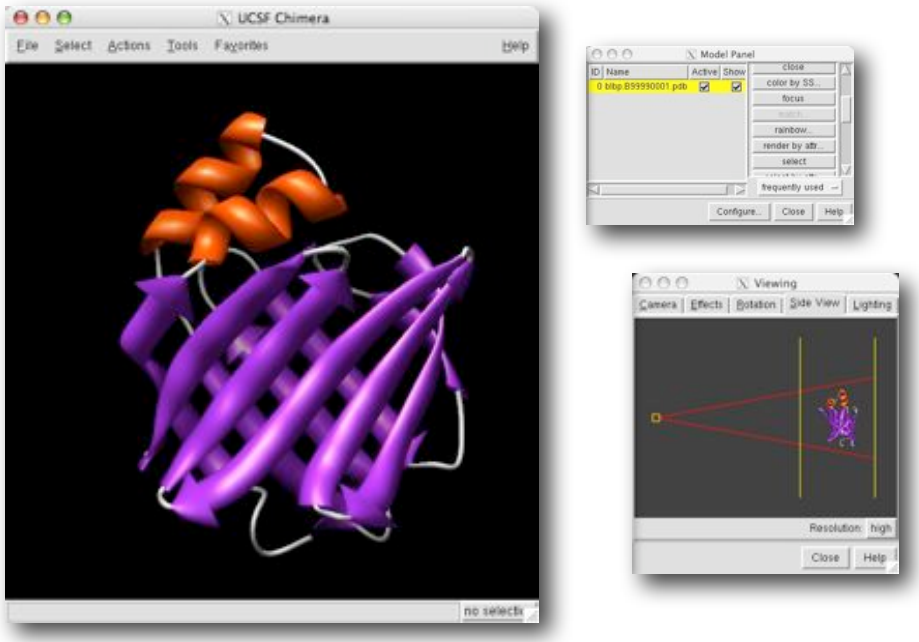
http://www.cgl.ucsf.edu/chimera/

Rasmol

http://www.openrasmol.org

PyMol

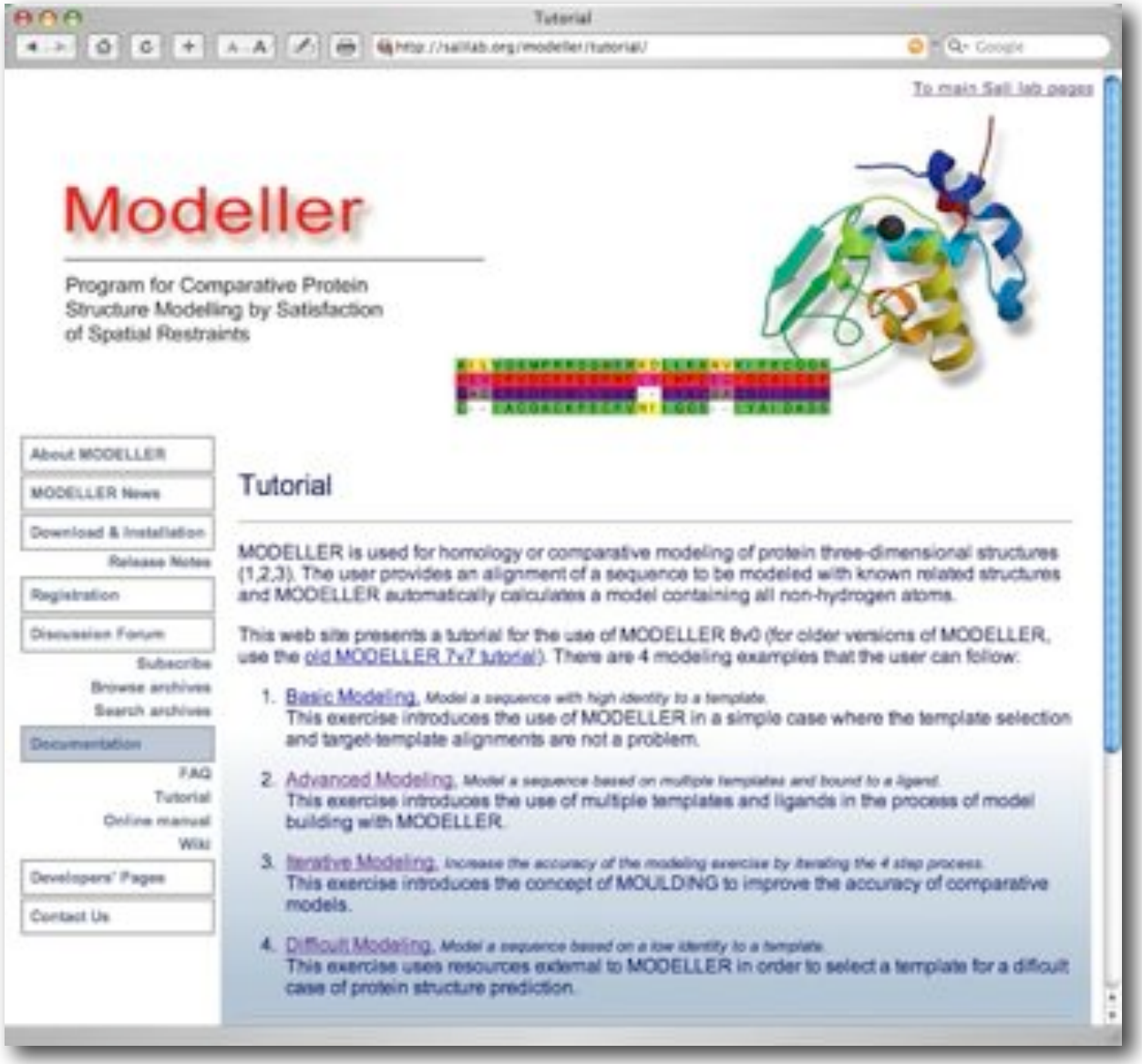
http://pymol.sourceforge.net/



Model file →

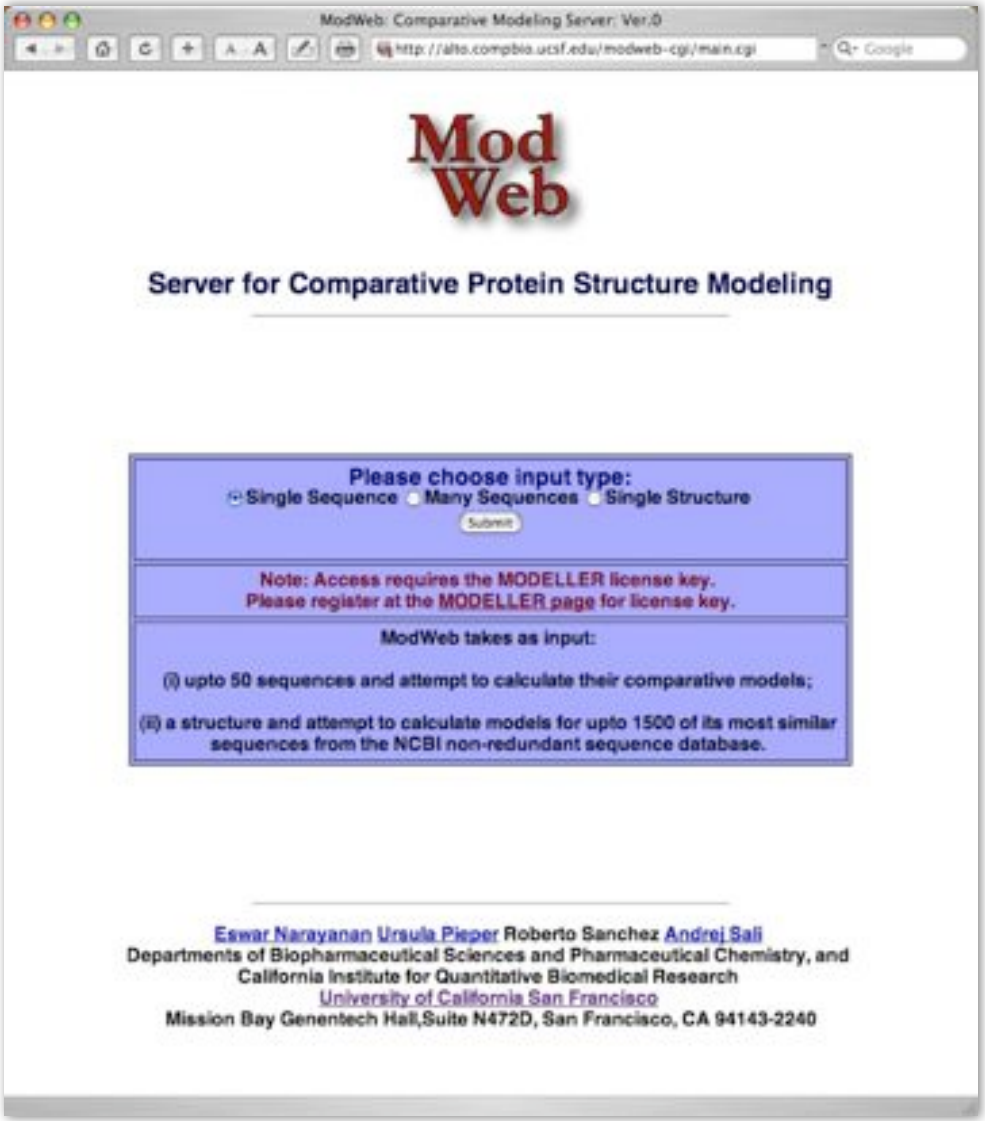
`blbp.B99990001.pdb`

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



MODWEB

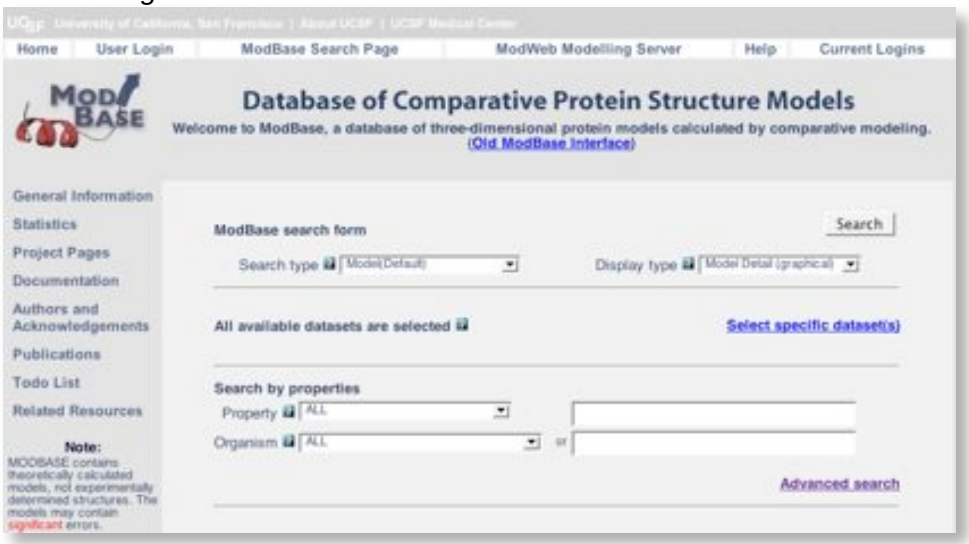
http://salilab.org/modweb



MODBASE

http://salilab.org/modbase

Search Page



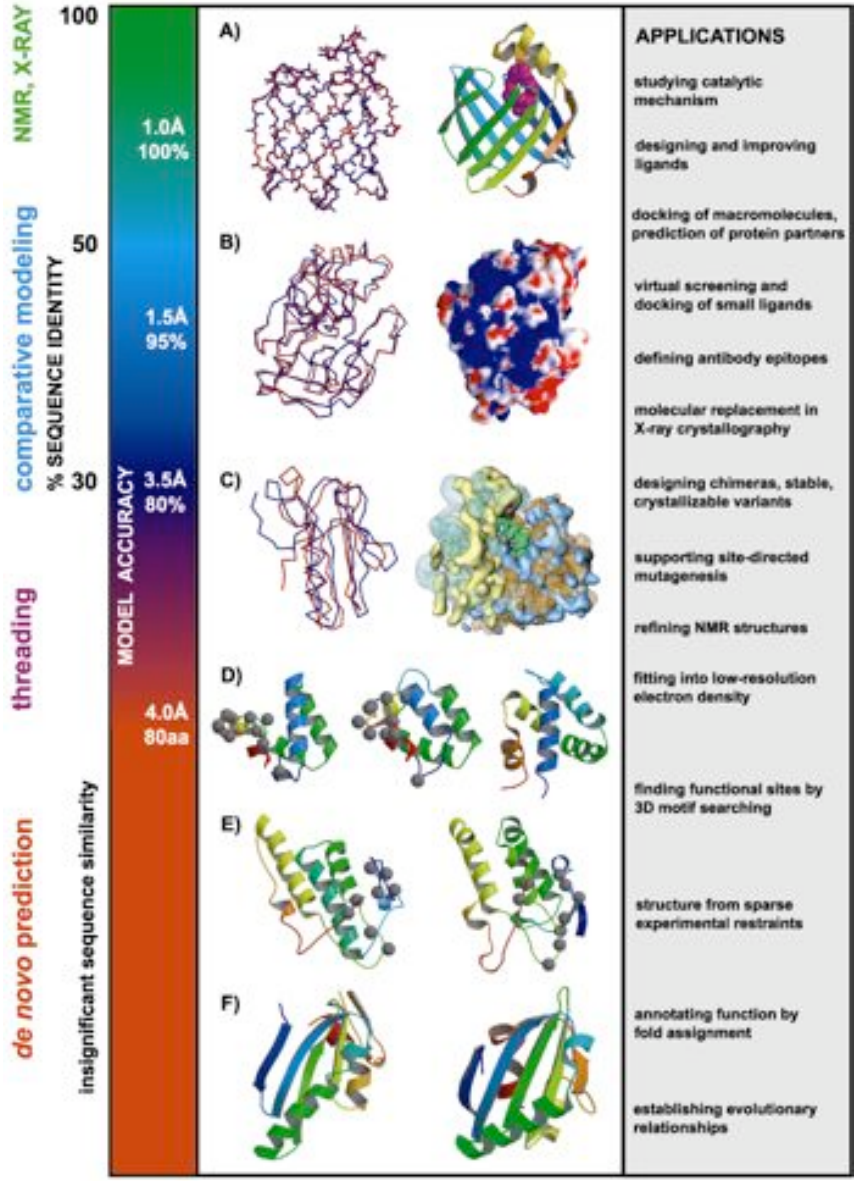
Model Details

Sequence Overview

Model Overview

Pieper et al. (2004) Nucleic Acids Research 32, D217-D222

Utility of protein structure models, despite errors



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

Acknowledgments

COMPARATIVE MODELING
Andrej Sali
M. S. Madhusudhan
Narayanan Eswar
Min-Yi Shen
Ursula Pieper
Ben Webb
Maya Topf

MODEL ASSESSMENT
David Eramian
Min-Yi Shen
Damien Devos

FUNCTIONAL ANNOTATION
Andrea Rossi
Fred Davis

FUNDING
Prince Felipe Research Center
Marie Curie Reintegration Grant
STREP EU Grant

MODEL ASSESSMENT
Francisco Melo (CU)
Alejandro Panjkovich (CU)

STRUCTURAL GENOMICS
Stephen Burley (SGX)
John Kuriyan (UCB)
NY-SGXRC

MAMMOTH
Angel R. Ortiz

FUNCTIONAL ANNOTATION
Fatima Al-Shahrour
Joaquin Dopazo

BIOLOGY
Jeff Friedman (RU)
James Hudsped (RU)
Partho Ghosh (UCSD)
Alvaro Monteiro (Cornell U)
Stephen Krilis (St.George H)

Tropical Disease Initiative
Stephen Maurer (UC Berkeley)
Arti Rai (Duke U)
Andrej Sali (UCSF)
Ginger Taylor (TSL)

CCPR Functional Proteomics
Patsy Babbitt (UCSF)
Fred Cohen (UCSF)
Ken Dill (UCSF)
Tom Ferrin (UCSF)
John Irwin (UCSF)
Matt Jacobson (UCSF)
Tack Kuntz (UCSF)
Andrej Sali (UCSF)
Brian Shoichet (UCSF)
Chris Voigt (UCSF)

EVA
Burkhard Rost (Columbia U)
Alfonso Valencia (CNB/UAM)

CAMP
Xavier Aviles (UAB)
Hans-Peter Nester (SANOFI)
Ernst Meinjohanns (ARPIDA)
Boris Turk (IJS)
Markus Gruetter (UE)
Matthias Wilmanns (EMBL)
Wolfram Bode (MPG)

http://bioinfo.cipf.es/sgu/