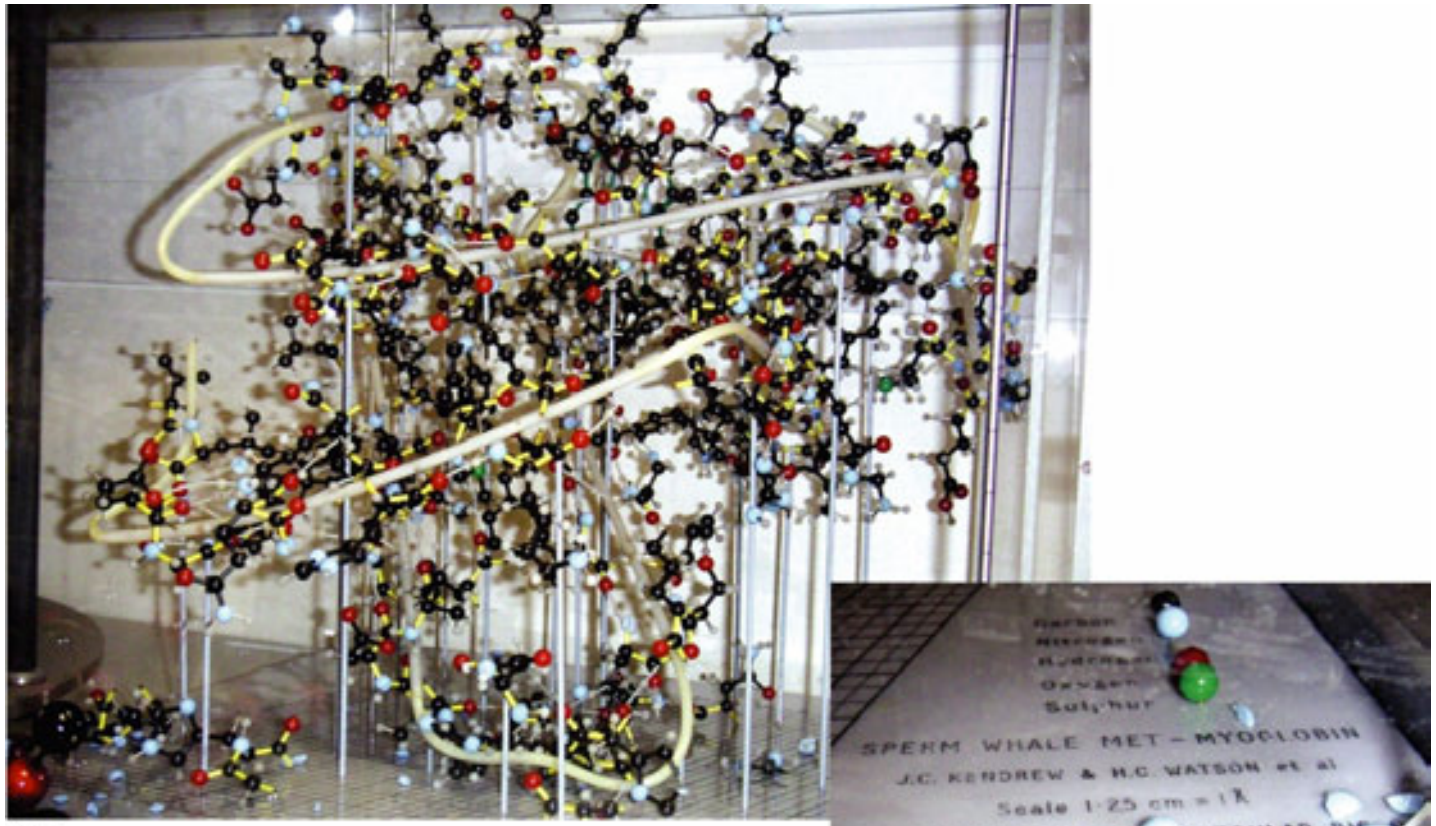
Protein modeling

Arne Elofsson



(A)

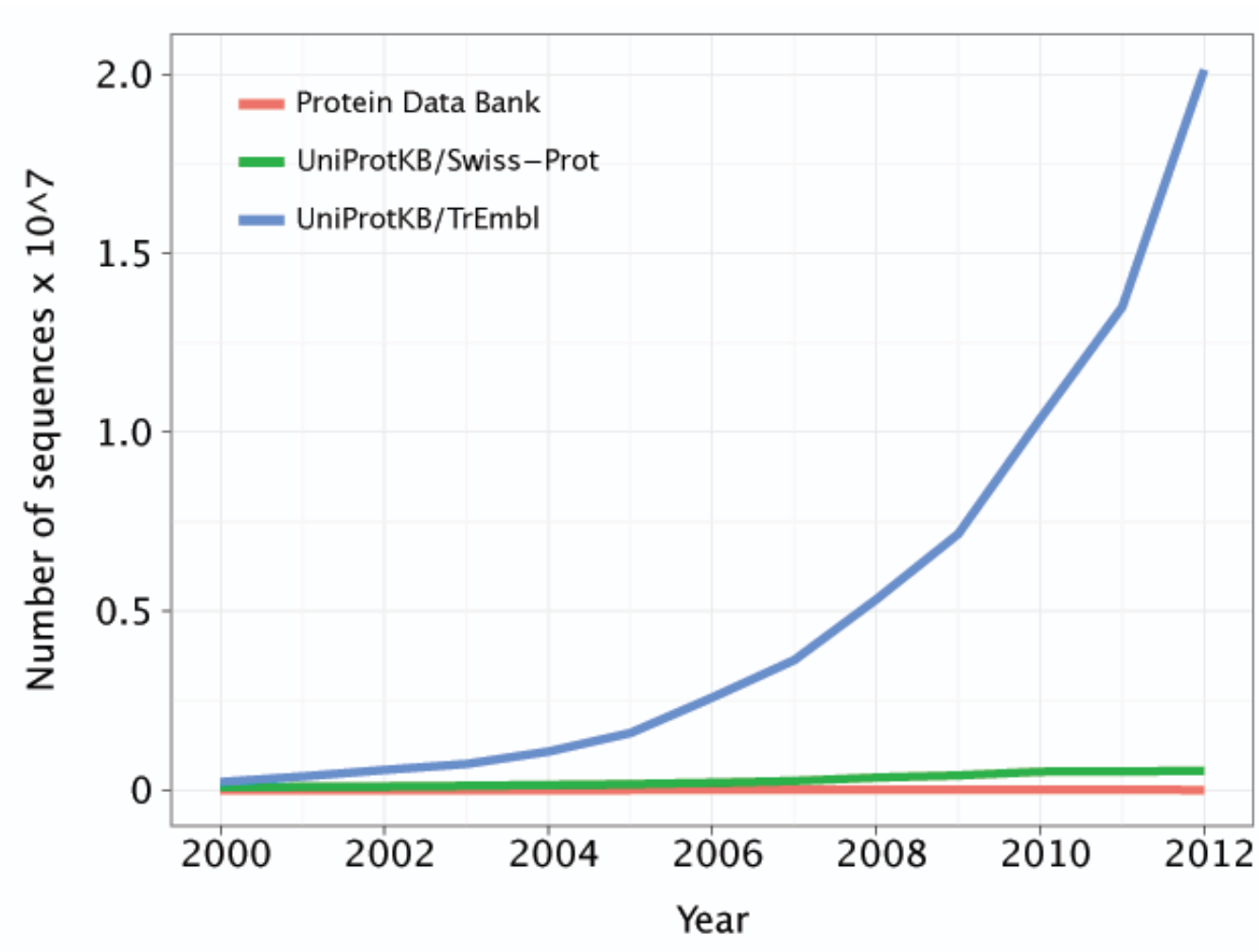

Sulphur
SPERM WHALE MET – MYOGLOBIN
J.C. KENDREW & H.C. WATSON et al
Scale 1·25 cm = 1Å
CAMBRIDGE, ENGLAND
A.A. BARKER M.A.
ENGLAND

(B)

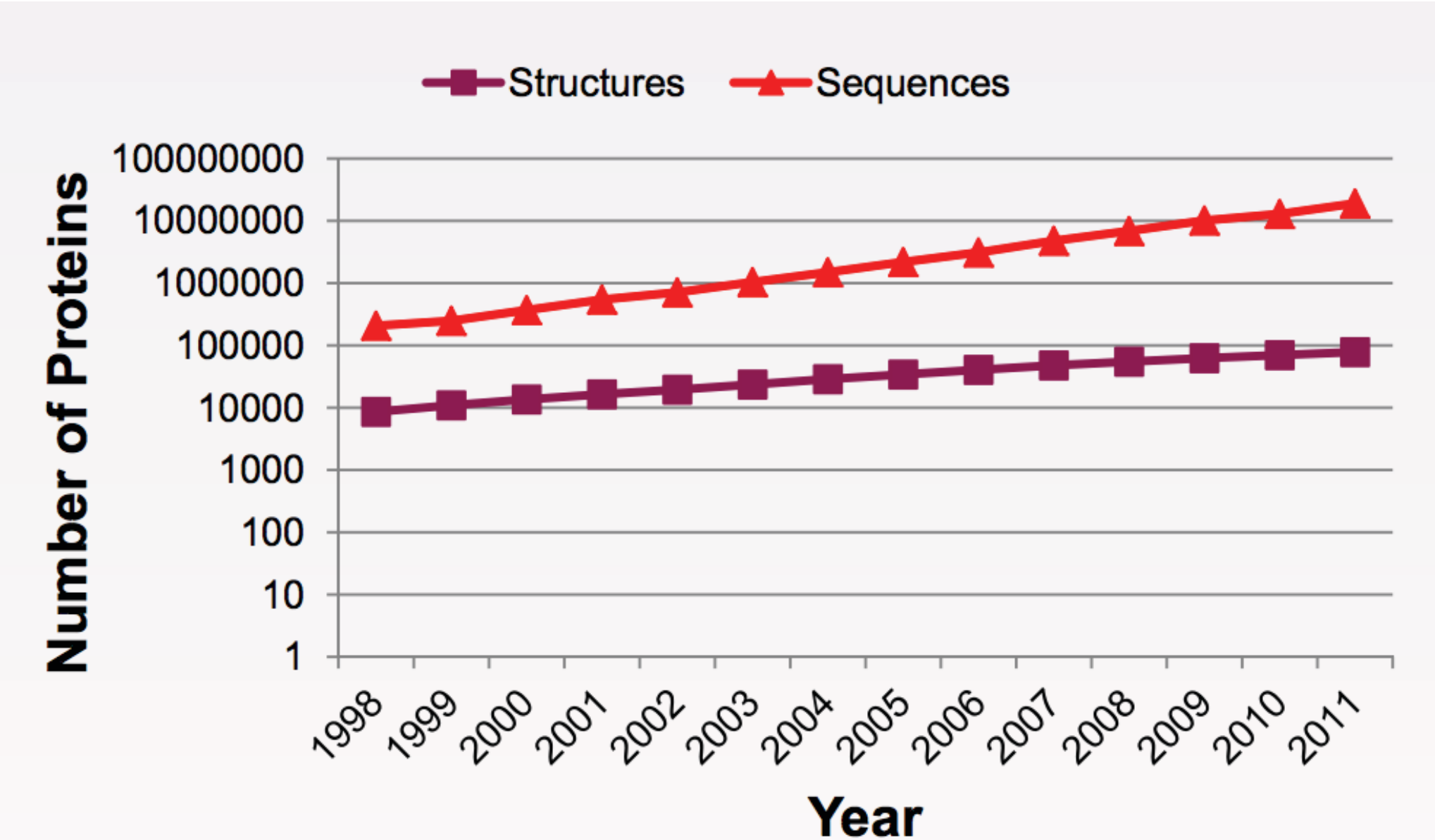
Why protein modeling?

- Experimental effort to determine protein structure is very large and costly
- The gap between the size of the protein sequence data and protein structure data is large and increasing
- Close to 50% of all new sequences can be homology modeled

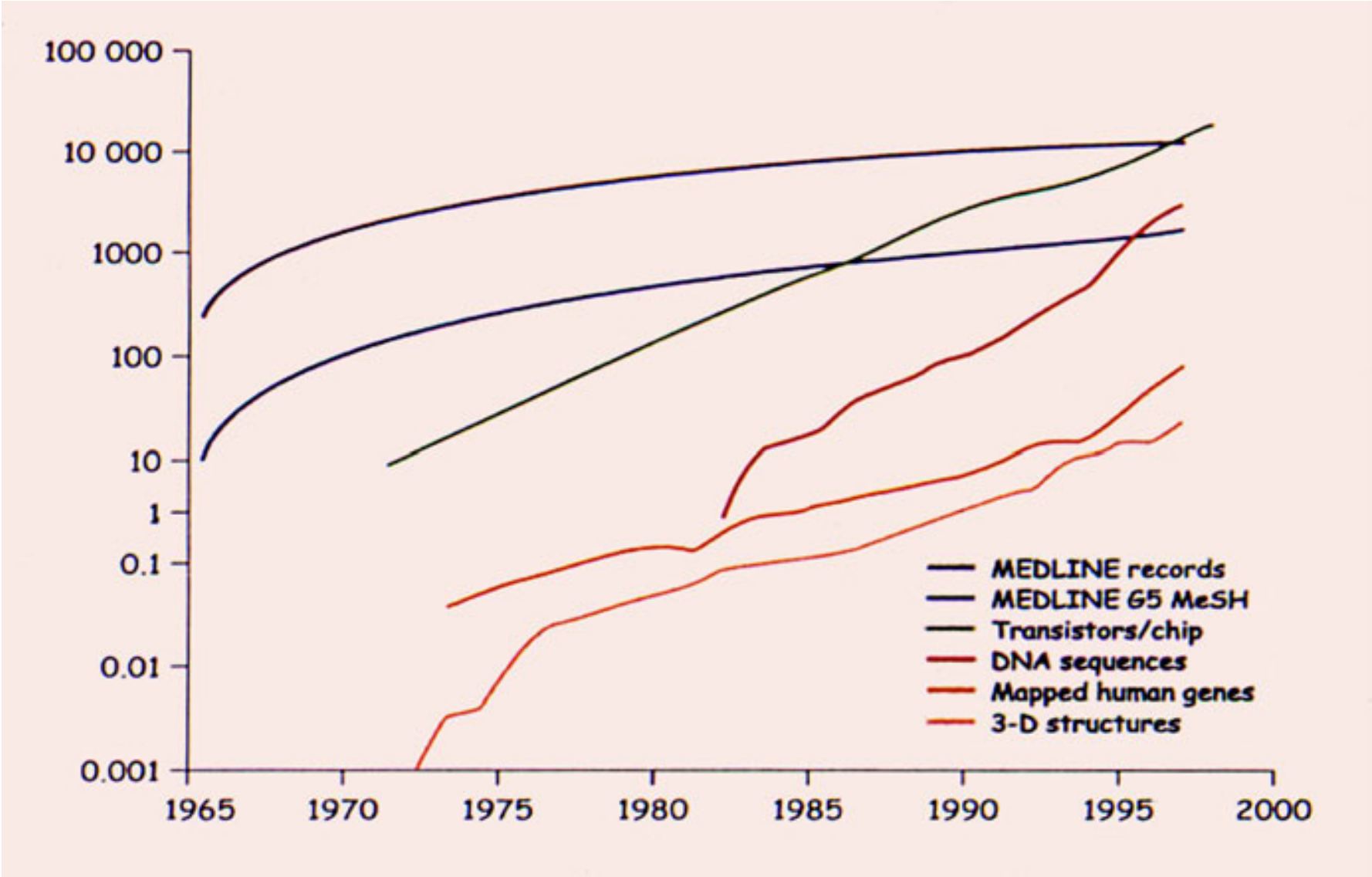
Number of protein sequences and structures is increasing.



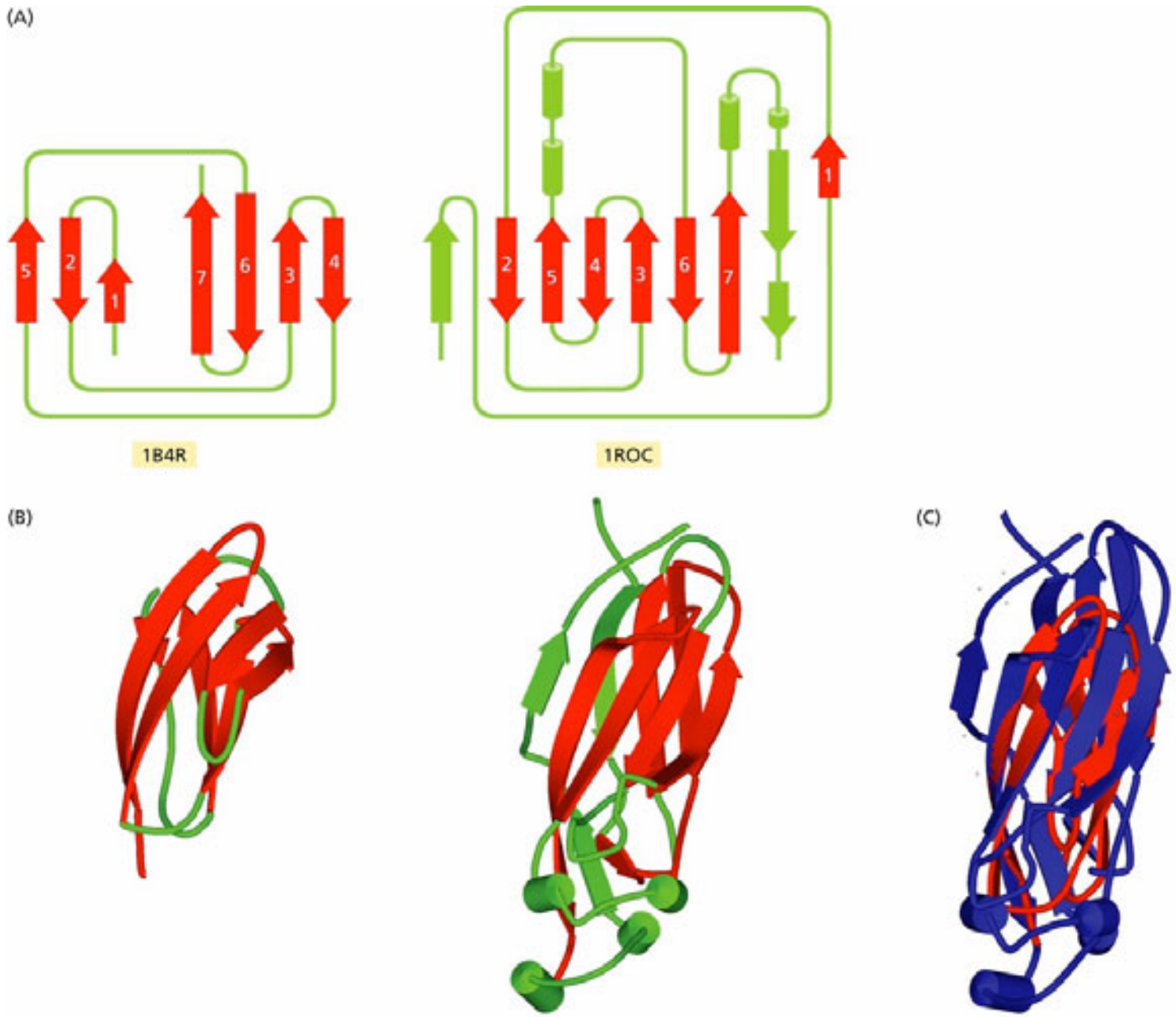
Exponential increase



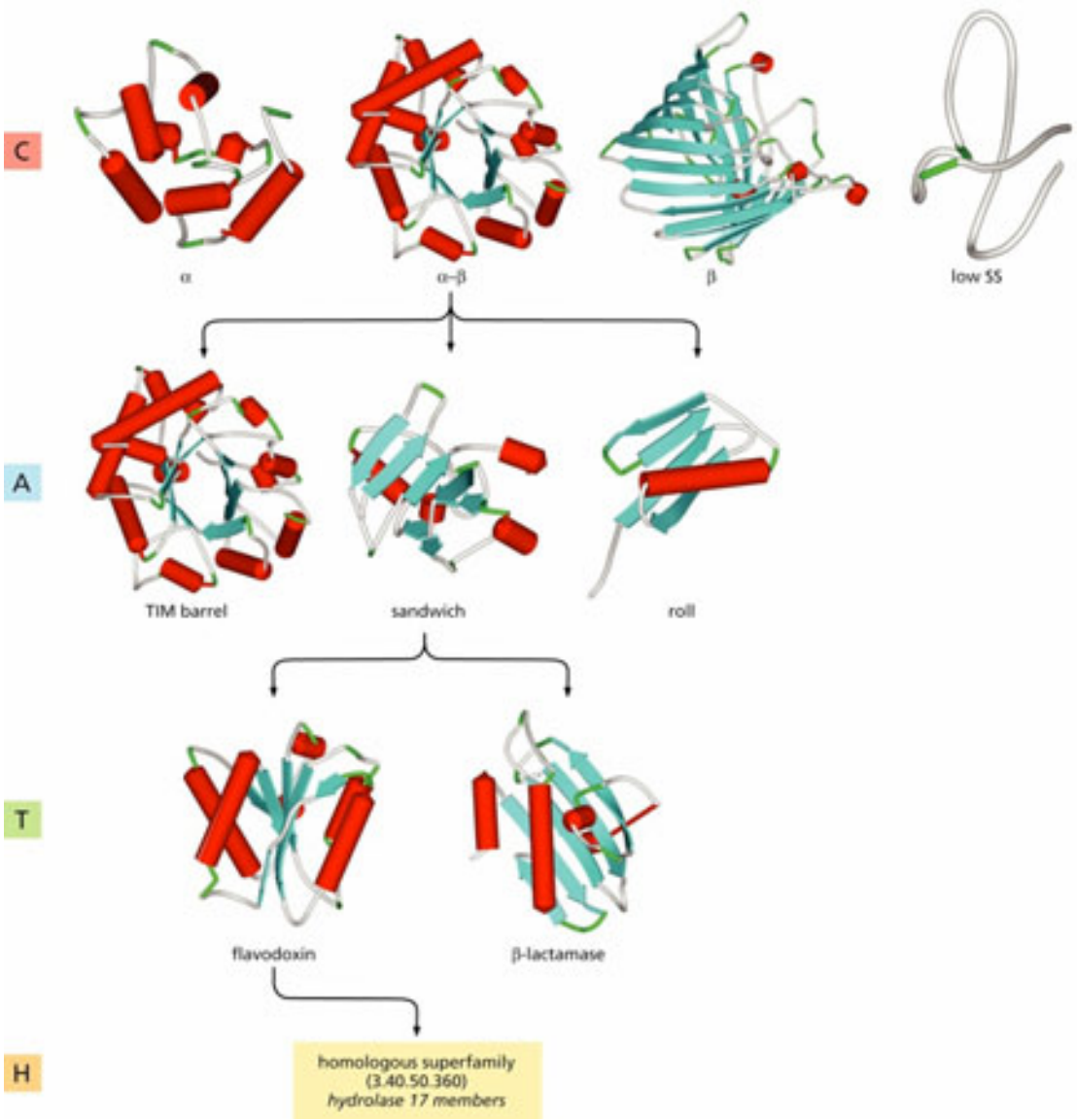
David Jones 2012


100 000
10 000
1000
100
10
1
0.1
0.01
0.001
1965
1970
1975
1980
1985
1990
1995
2000
MEDLINE records
MEDLINE G5 MeSH
Transistors/chip
DNA sequences
Mapped human genes
3-D structures

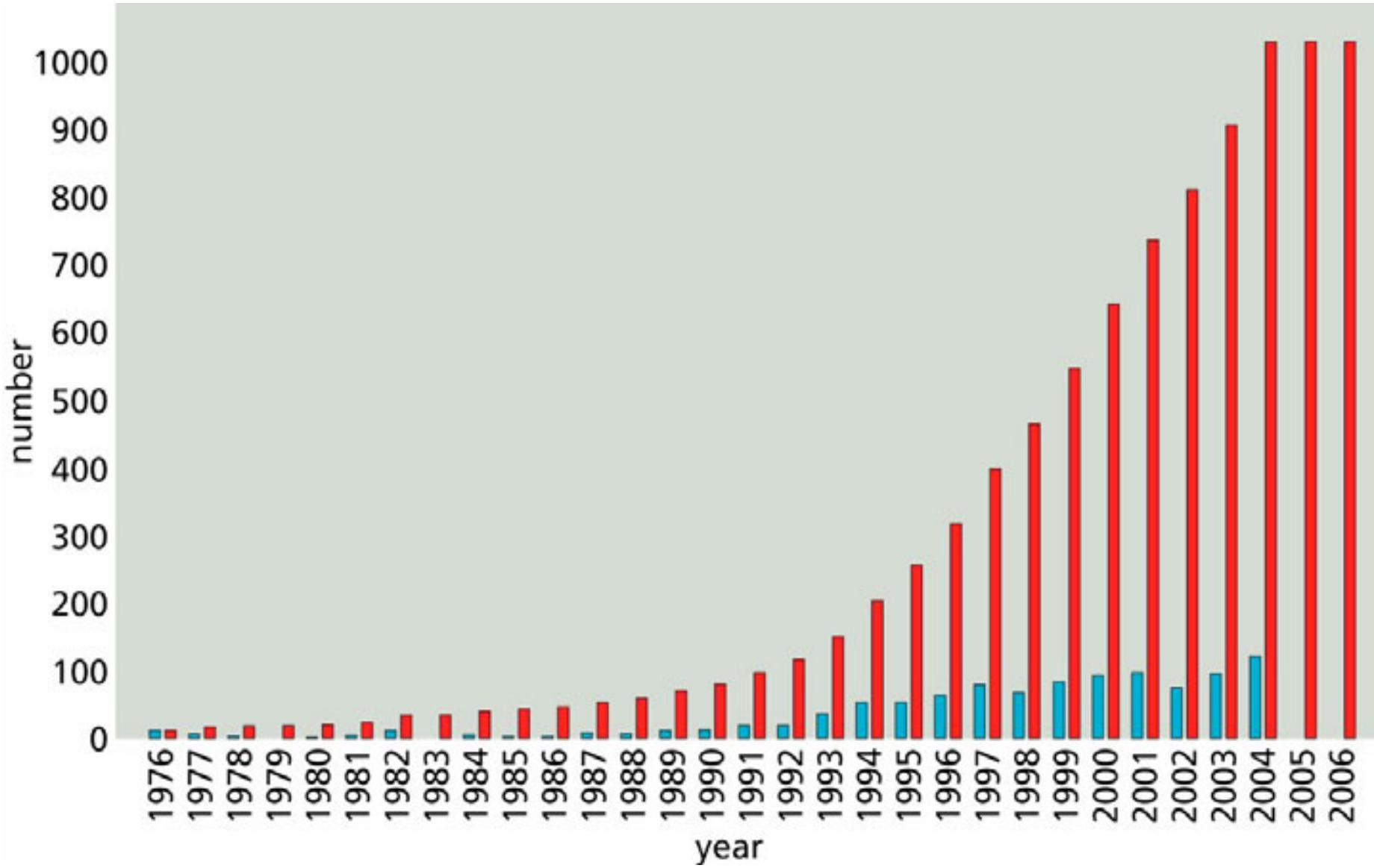
FOLDS

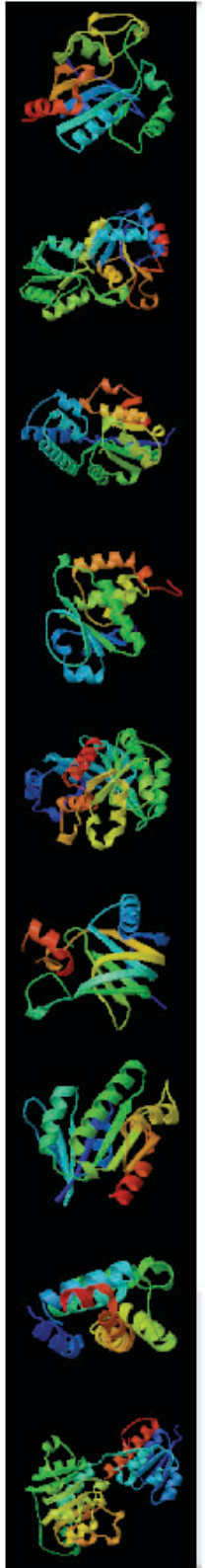


WHAT IS A FOLD SCOP/CATH



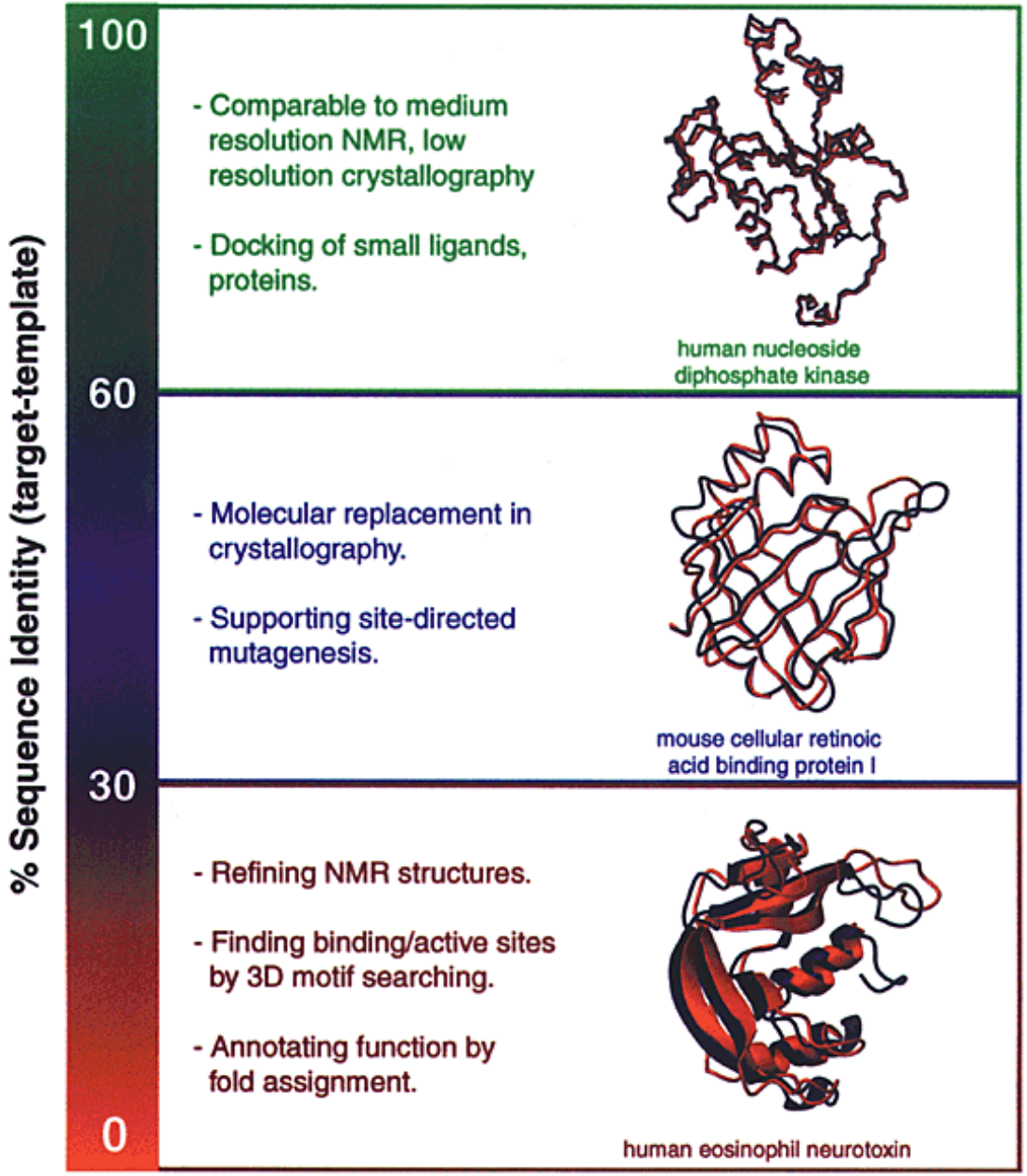
NUMBER OF FOLDS IS NOT INFINITE





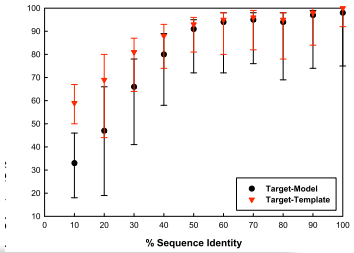
How Well Can We Model a structure?

Sali, A. & Kuriyan, J. *Trends Biochem. Sci.* **22**, M20–M24 (1999)



Model Accuracy

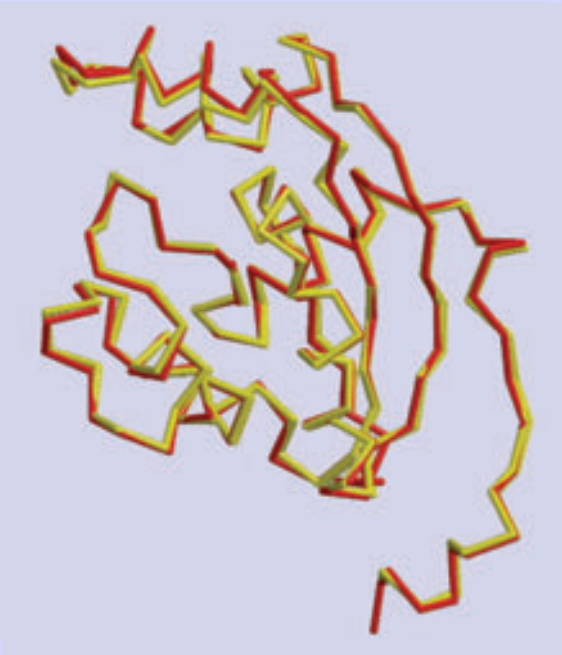
Marti-Renom *et al.* Annu.Rev.Biophys.Biomol.Struct. **29**, 291-325, 2000.



HIGH ACCURACY

NM23
Seq id 77%

Cα equiv 147/148
RMSD 0.41Å



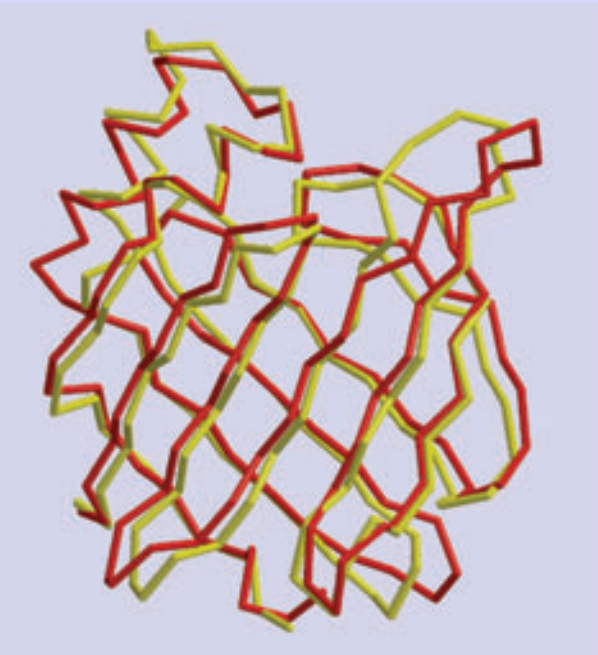
Sidechains
Core backbone
Loops

X-RAY / MODEL

MEDIUM ACCURACY

CRABP
Seq id 41%

Cα equiv 122/137
RMSD 1.34Å

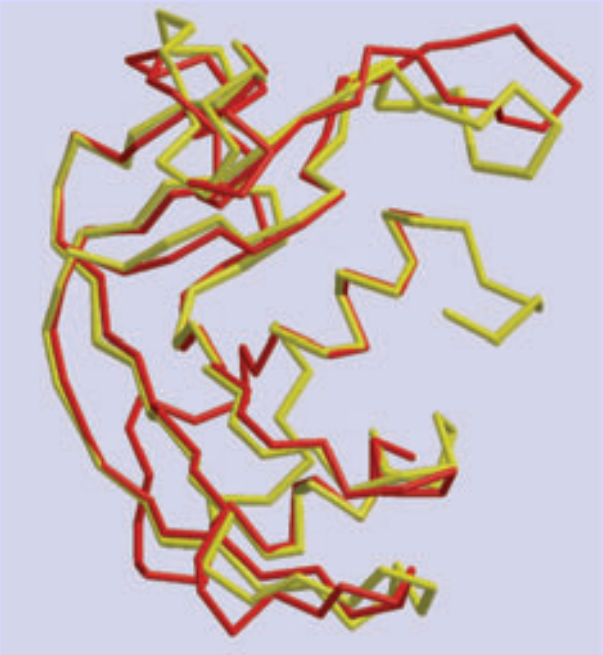


Sidechains
Core backbone
Loops
Alignment

LOW ACCURACY

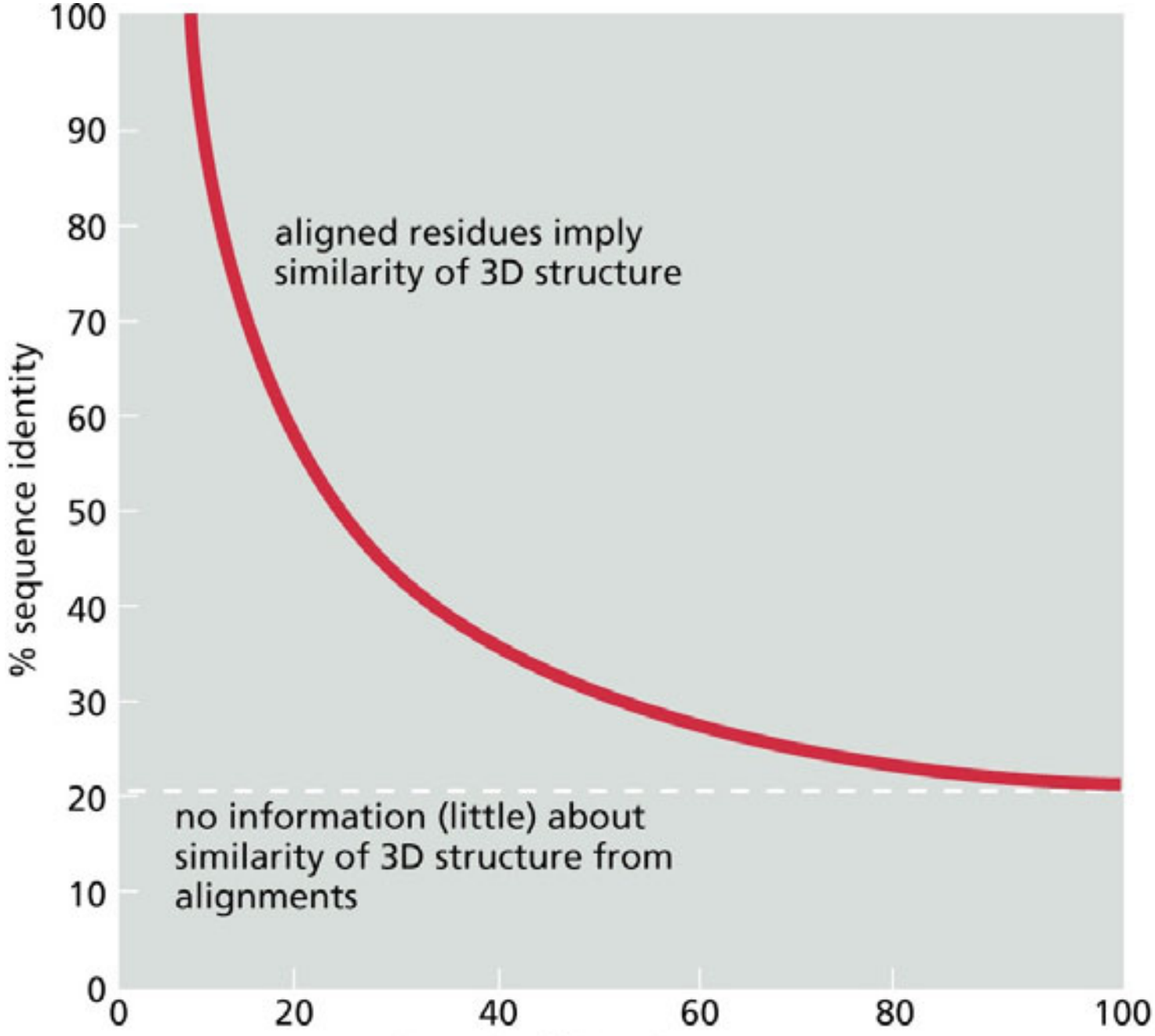
EDN
Seq id 33%

Cα equiv 90/134
RMSD 1.17Å



Sidechains
Core backbone
Loops
Alignment
Fold assignment

Twilight zone

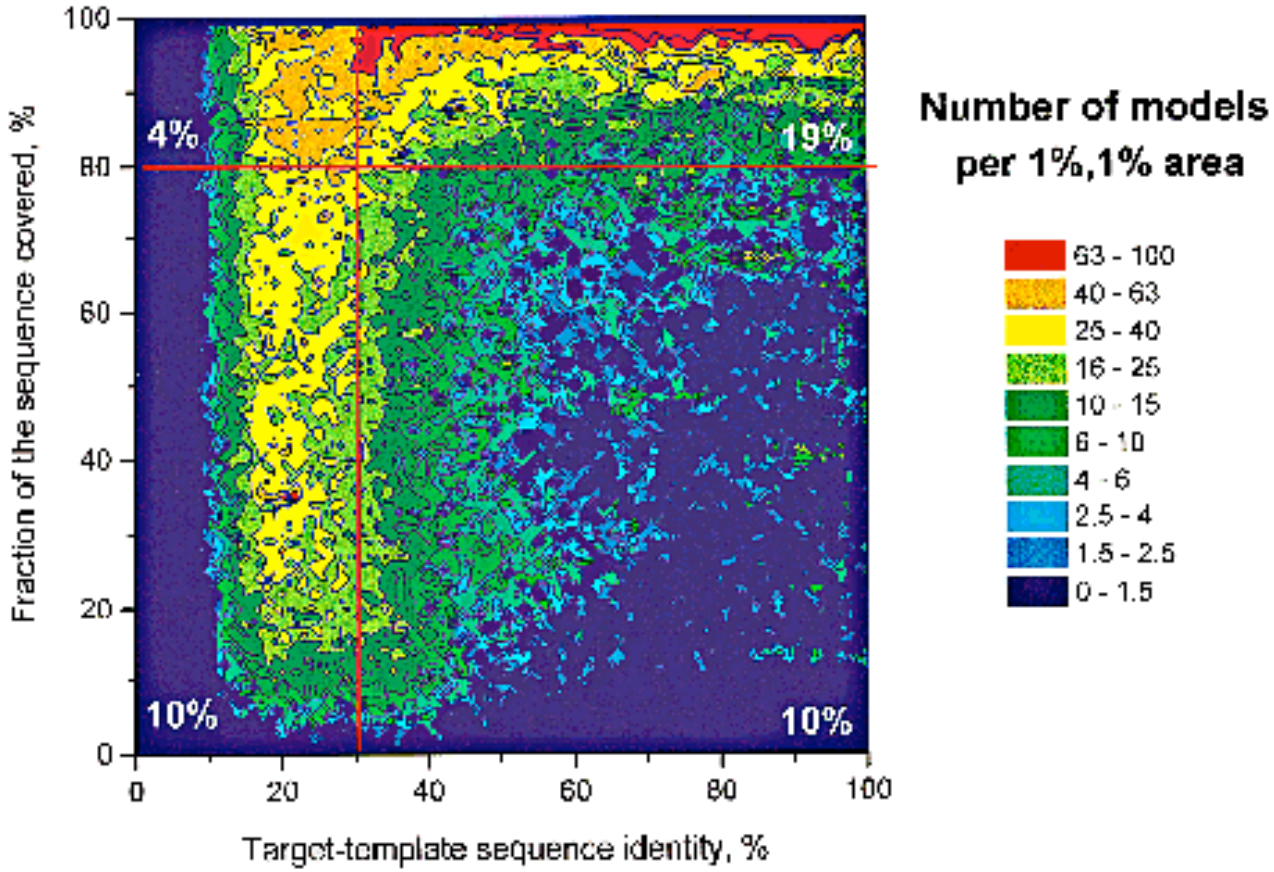


Structural coverage

high quality models

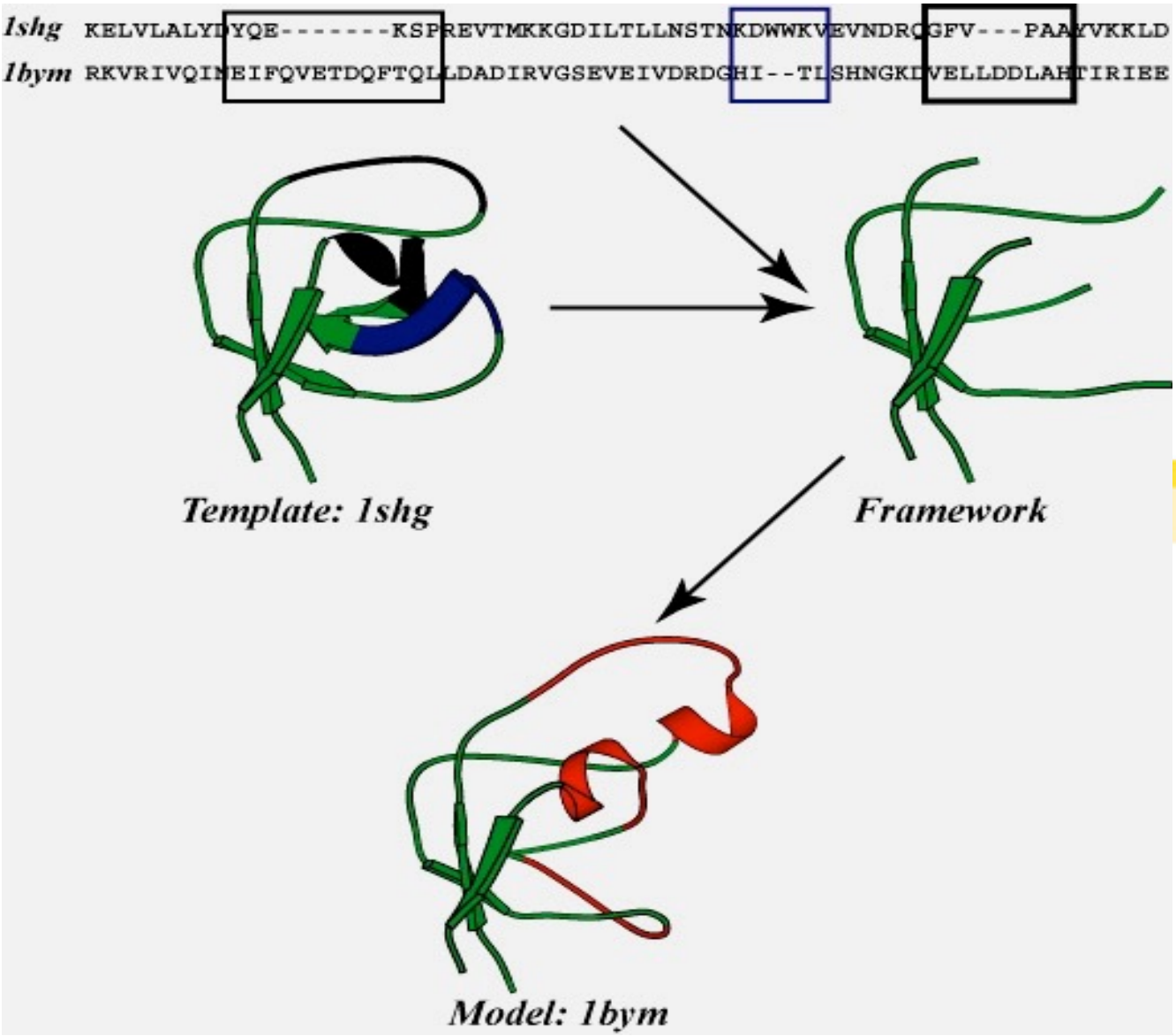
Complete models

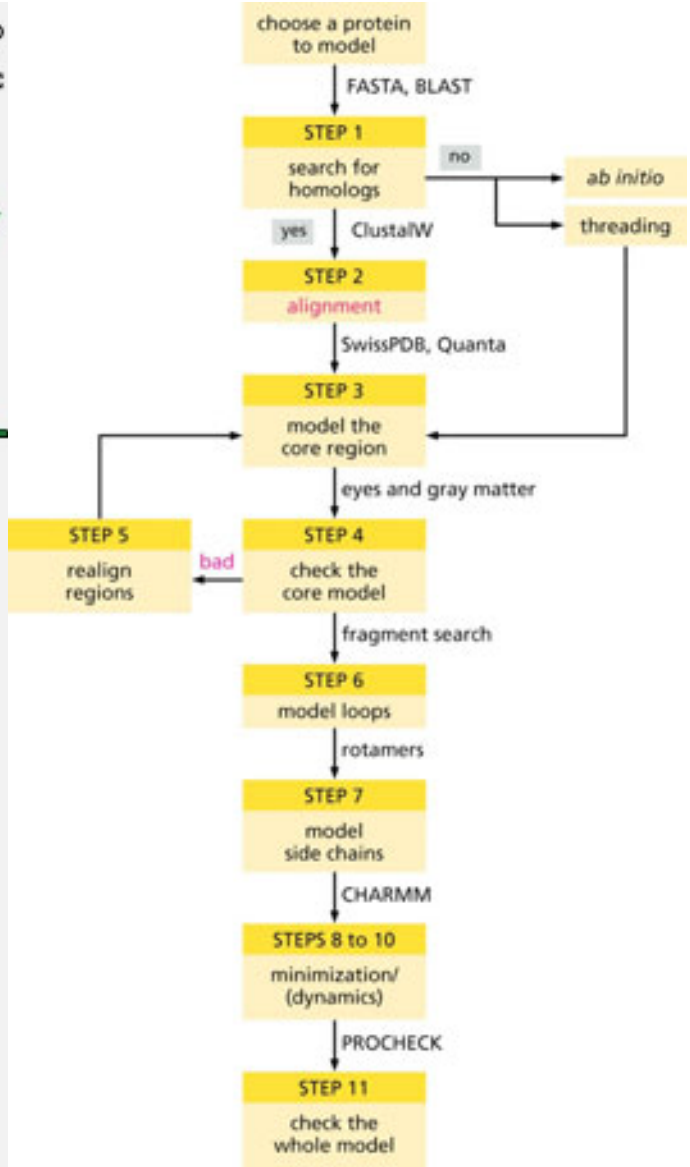
Total = 43 %



Vitkup et al. (2001)

The principle of homology modeling





THE STEPS OF MODELING

1 Detect template

2 Get alignment

3 Exchange side chains

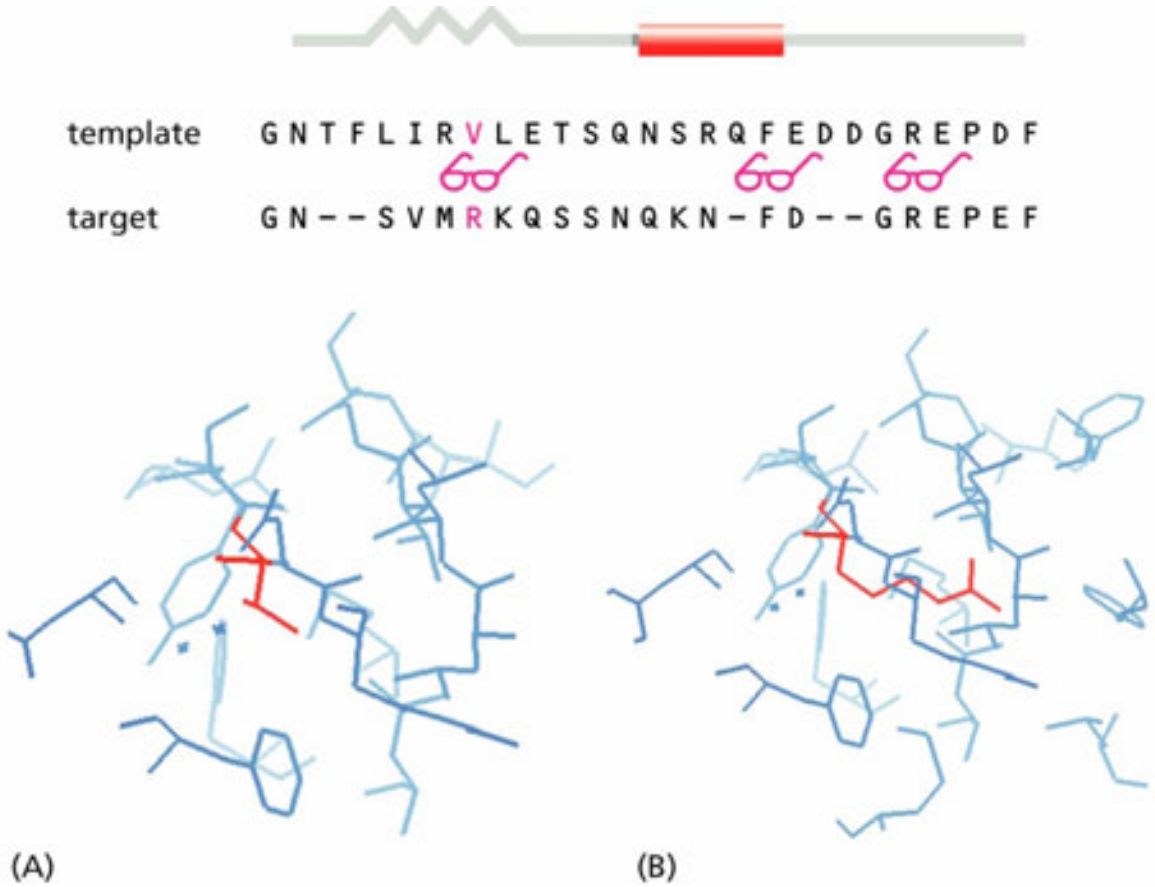
4 Insertions/deletions -Loop

5 Refine model

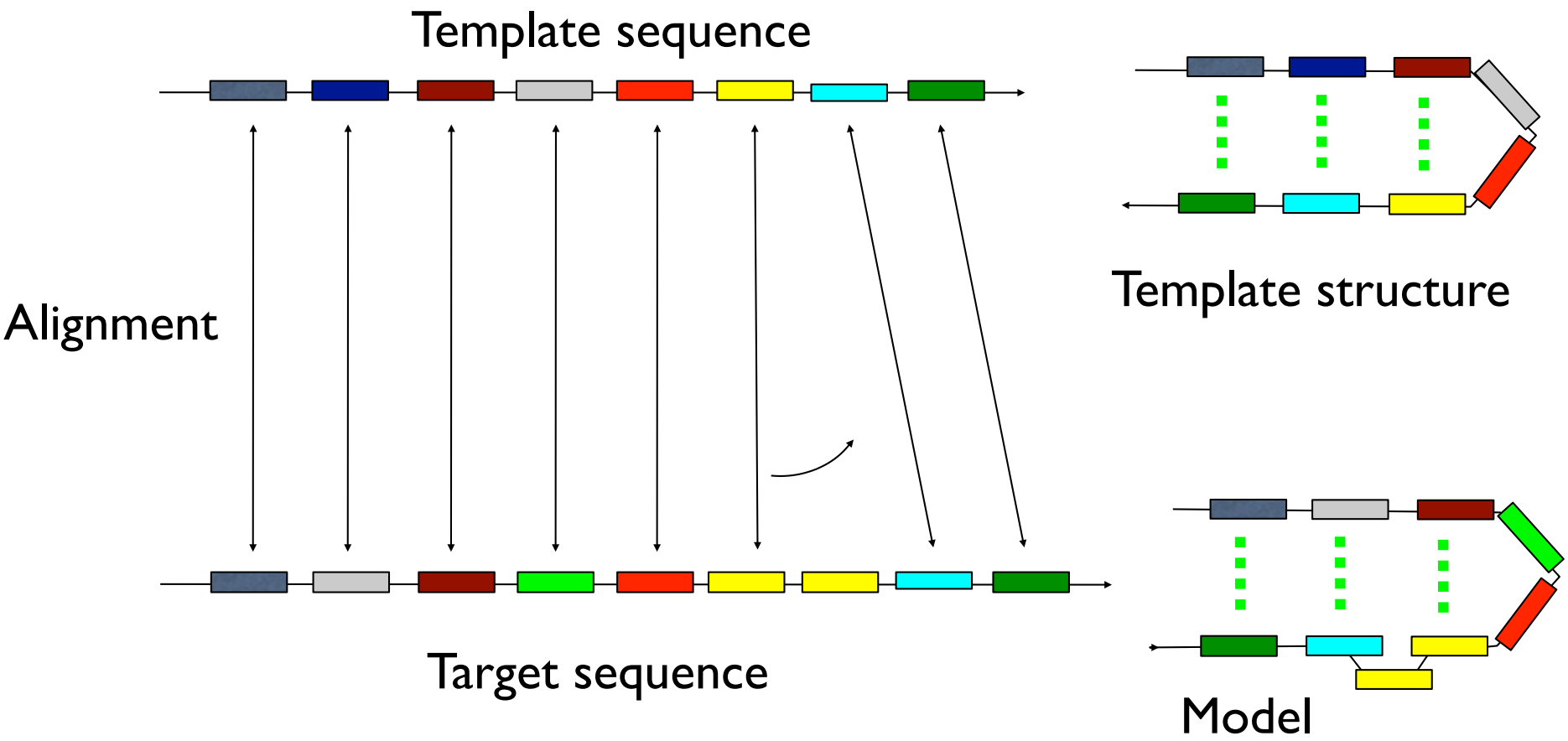
6 Evaluate model

7 Iterate

No modeling method can correct incorrect alignments

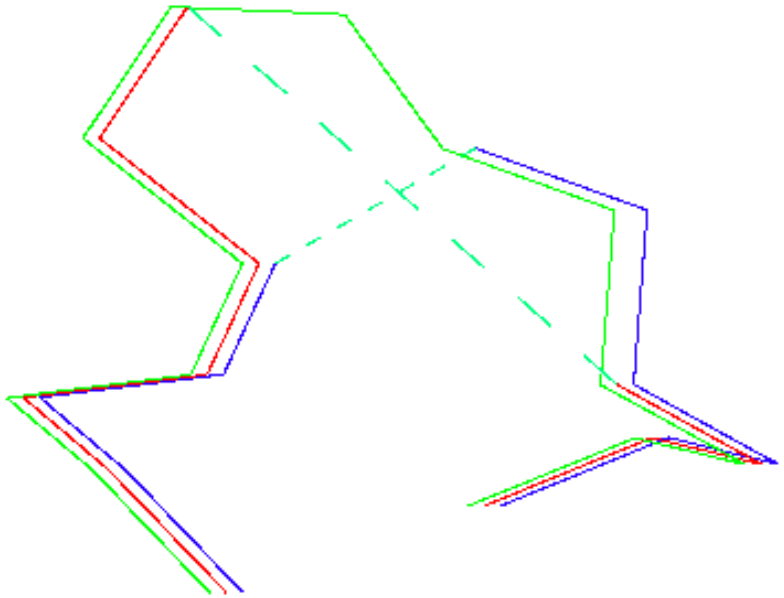


The crucial importance of the alignment



Improving the Alignment

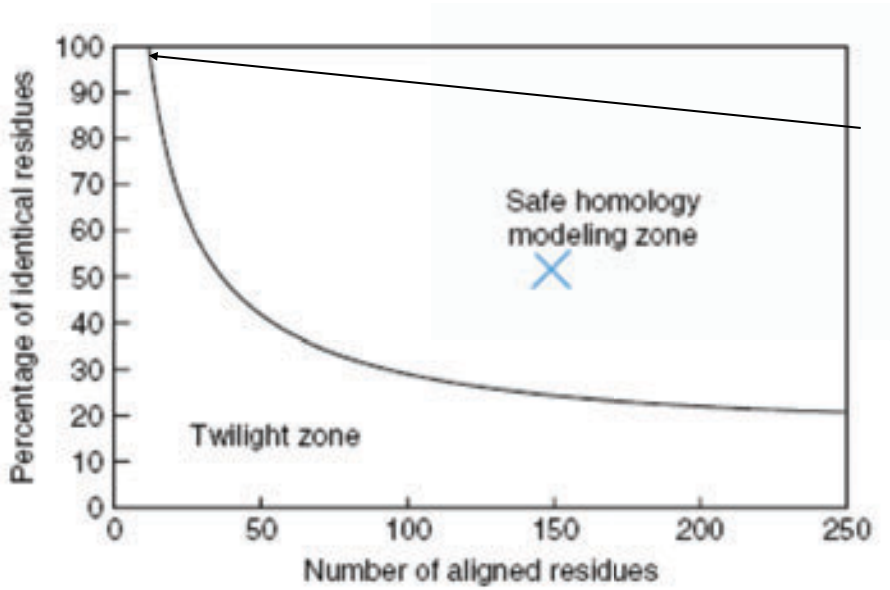
1	2	3	4	5	6	7	8	9	10	11	12	13	14
PHE	ASP	ILE	CYS	ARG	LEU	PRO	GLY	SER	ALA	GLU	ALA	VAL	CYS
PHE	ASN	VAL	CYS	ARG	THR	PRO	---	---	---	GLU	ALA	ILE	CYS
PHE	ASN	VAL	CYS	ARG	---	---	---	THR	PRO	GLU	ALA	ILE	CYS

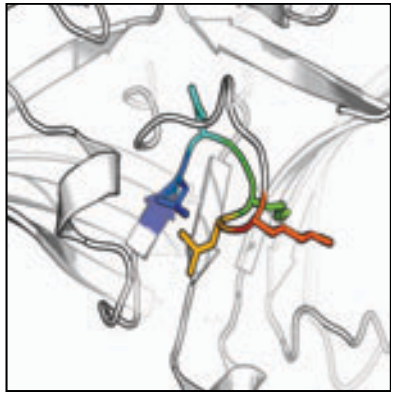


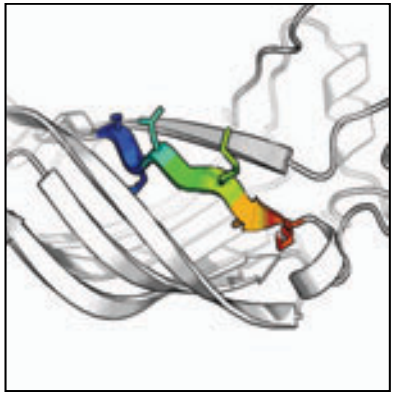
From "Professional Gambling" by Gert Vriend
http://www.cmbi.kun.nl/gv/articles/text/gambling.html

Template Quality

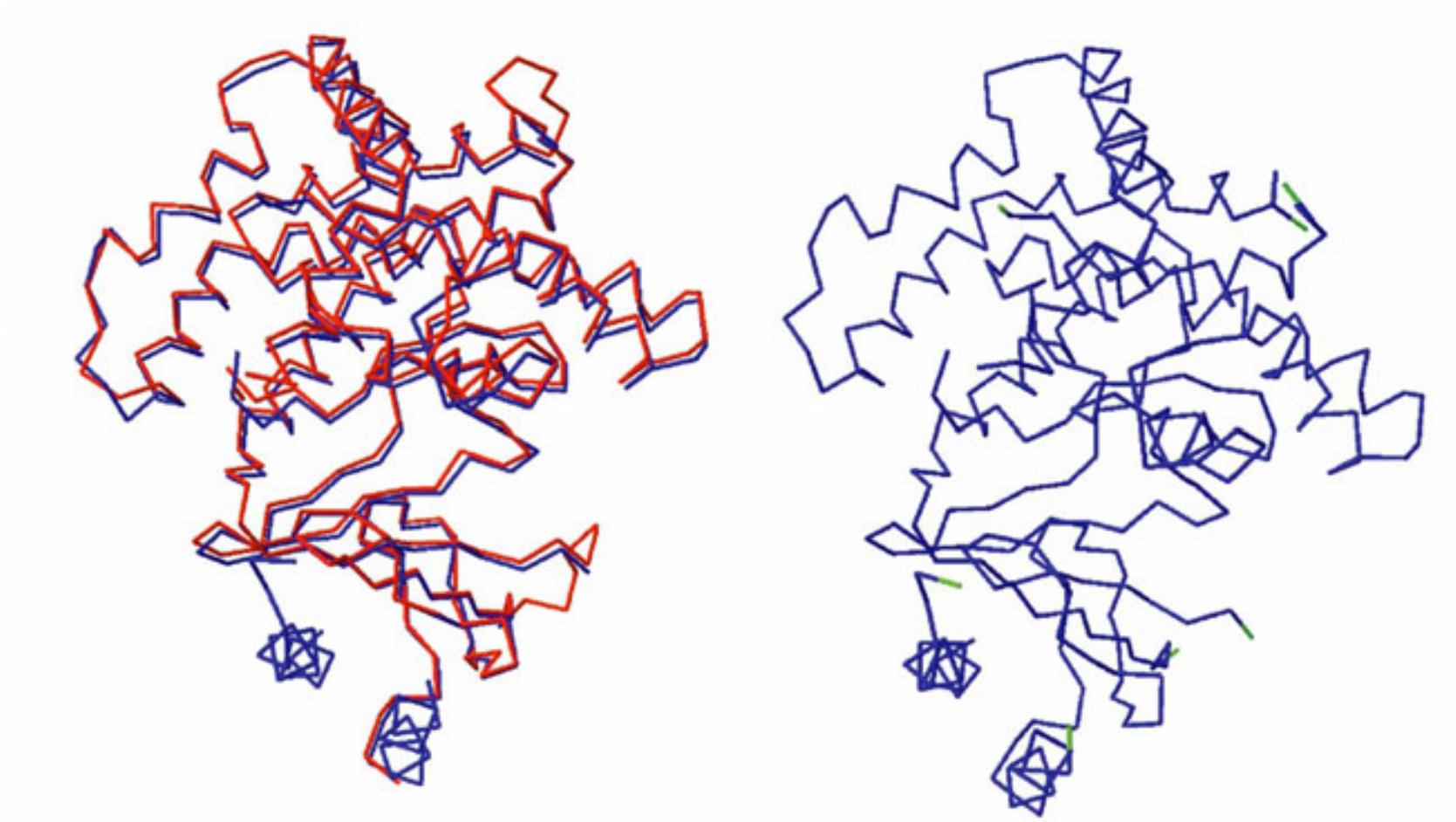
- Selecting the best template is crucial!
- The best template may not be the one with the highest % id (best p-value…)
  - Template 1: 93% id, 3.5 Å resolution ☹
  - Template 2: 90% id, 1.5 Å resolution ☺





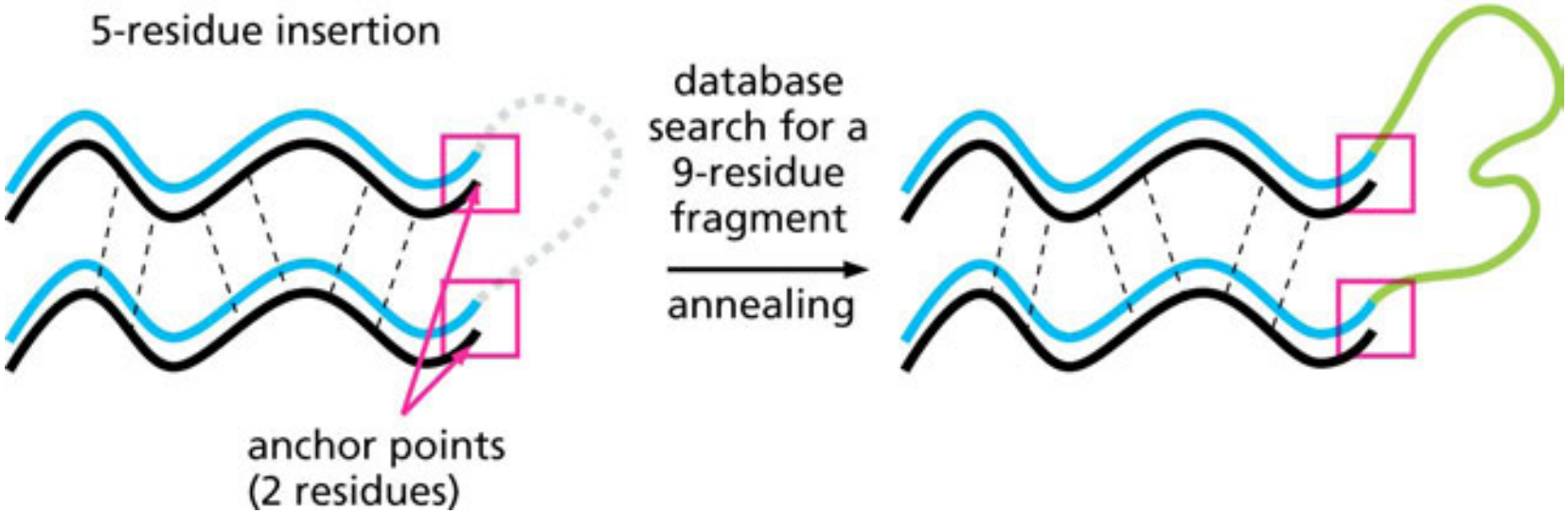


Conserved cores

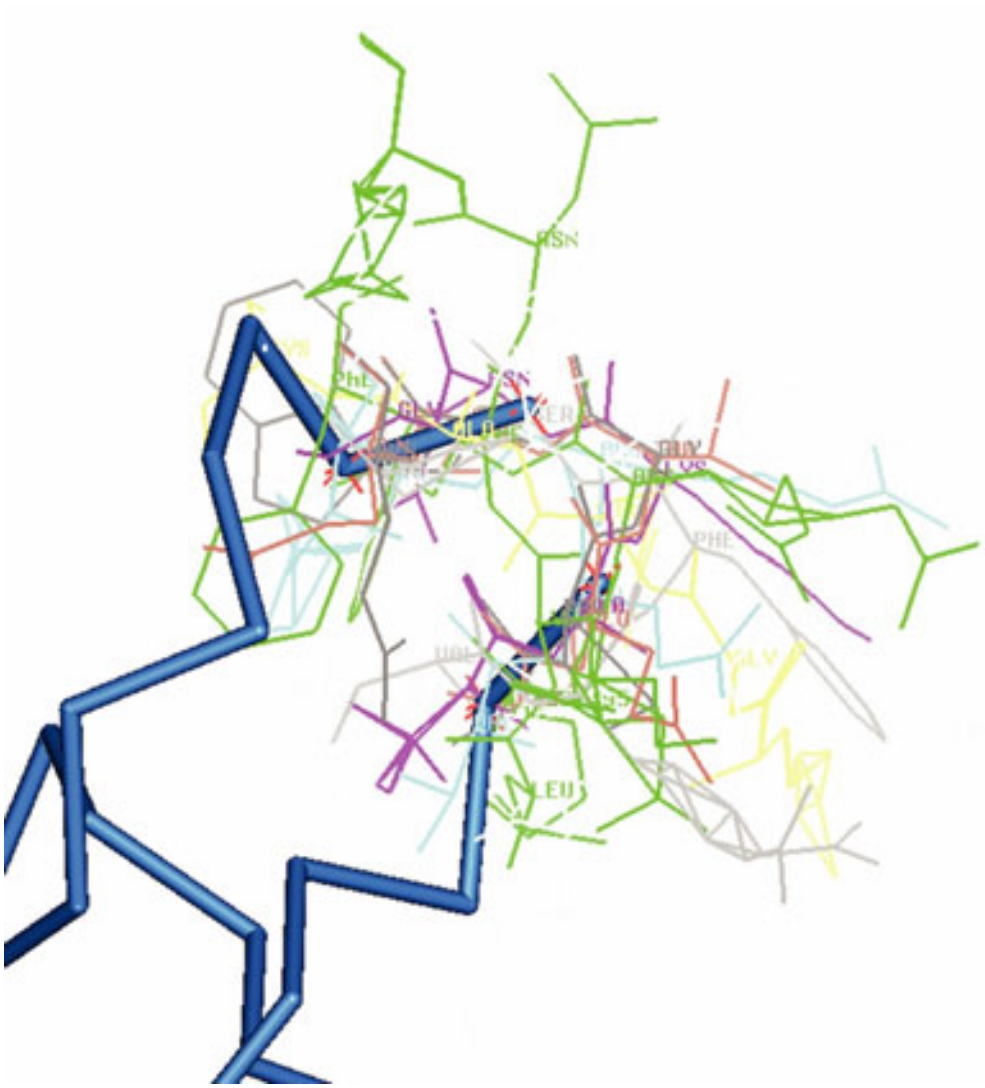


Fragments can be used to search for loops

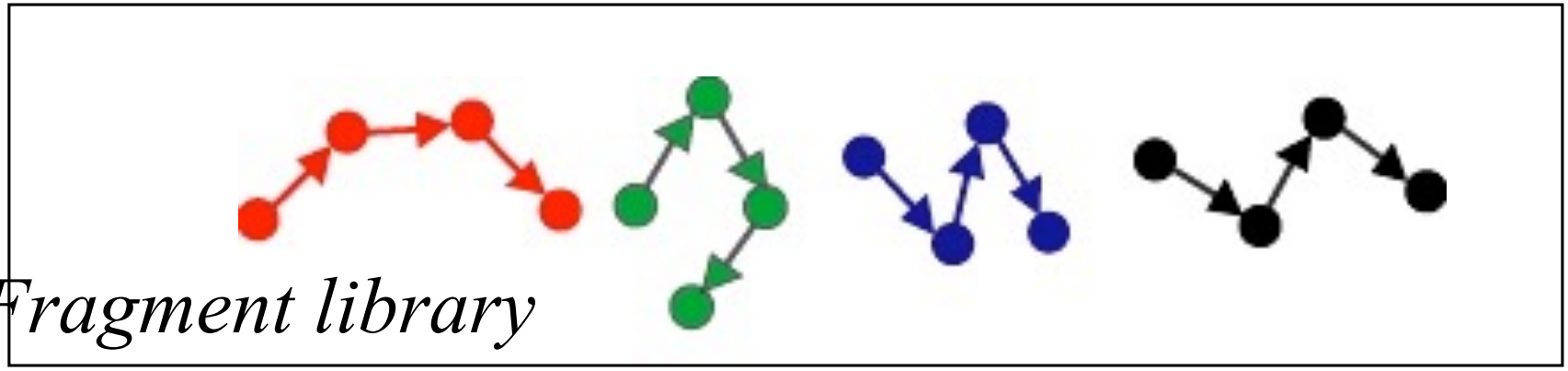
```
Target:      VLVATY      HDFVLI ...
Template:    VLIISYFGNSGREFVIL ...
```



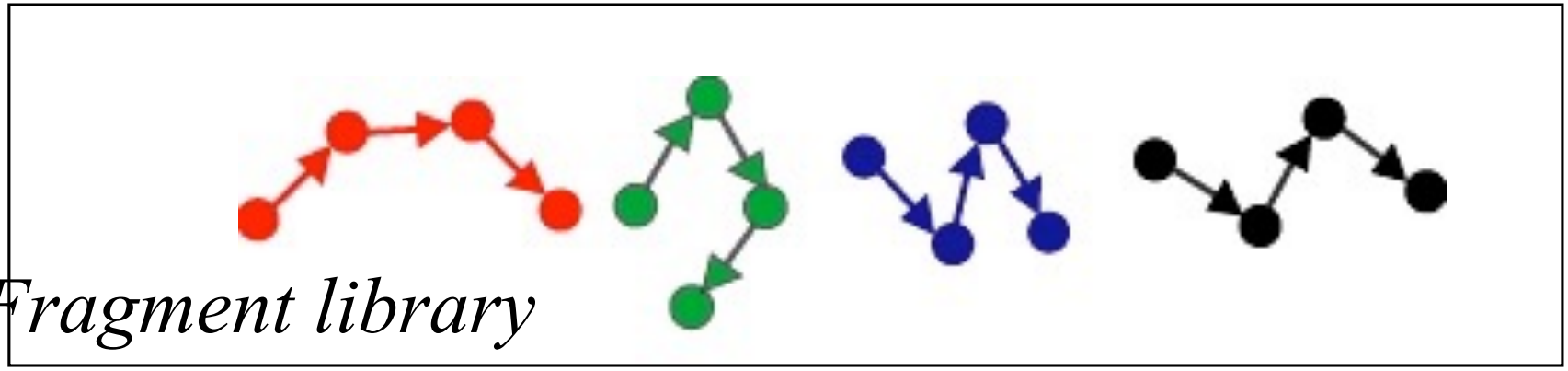
Example of loop fragments

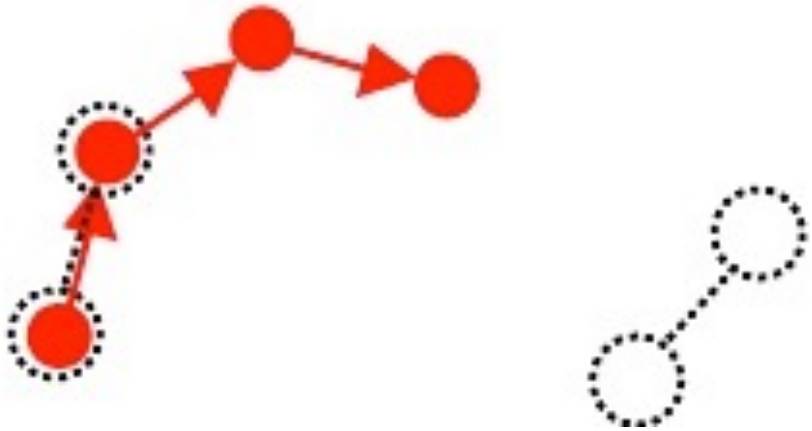


Generating Loops

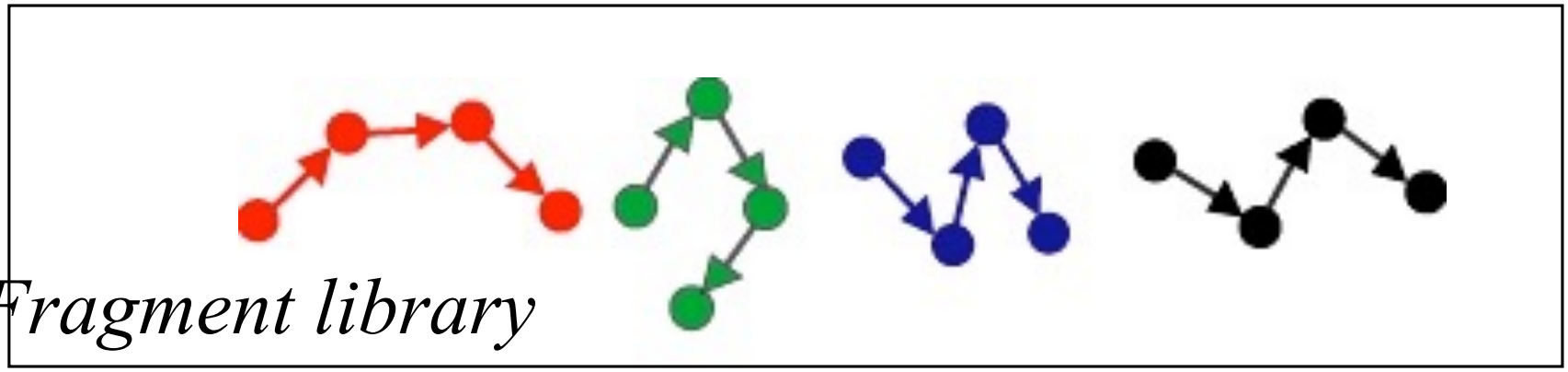


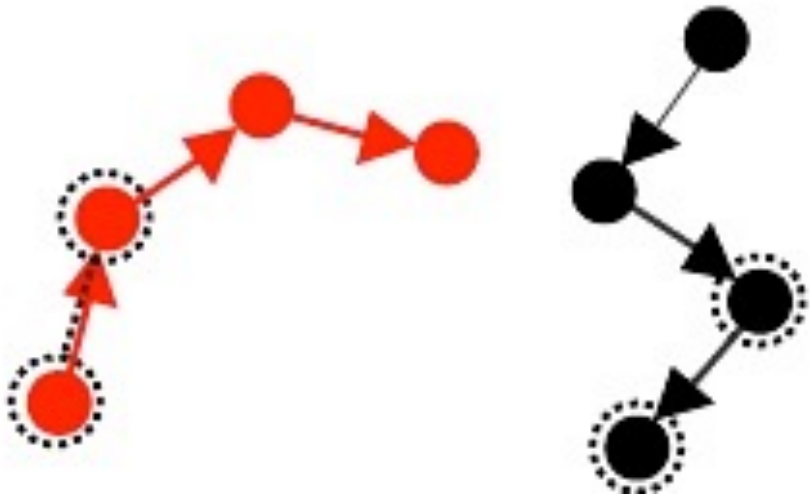
Generating Loops



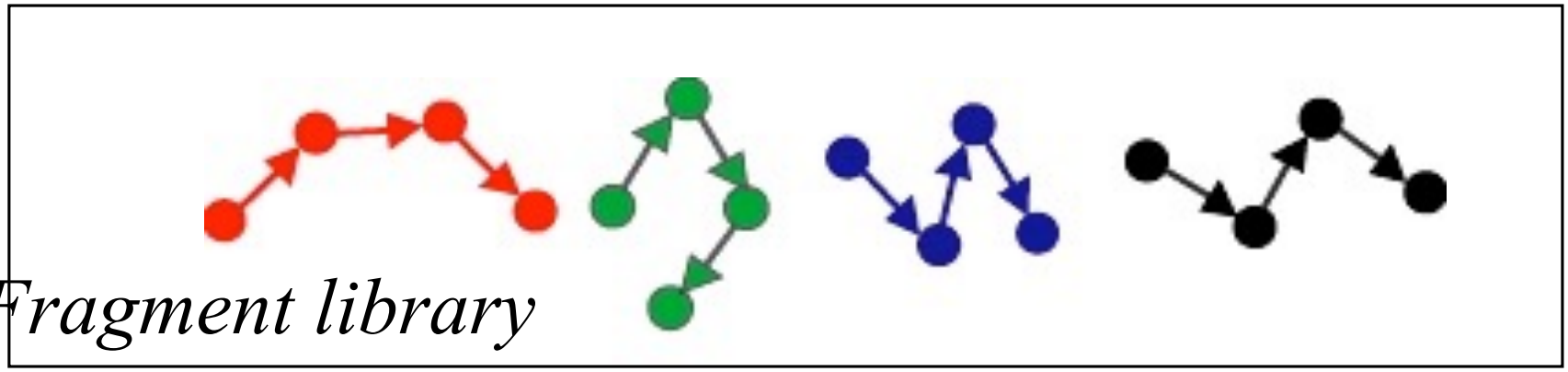


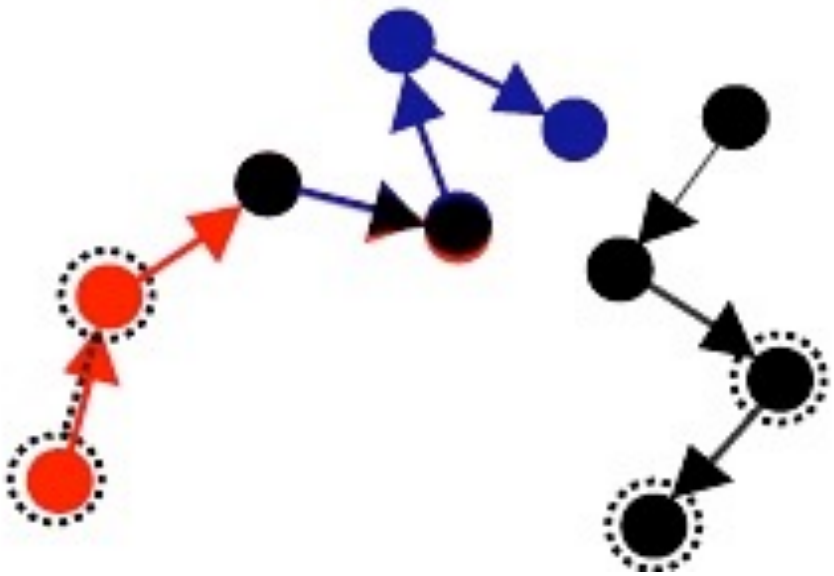
Generating Loops



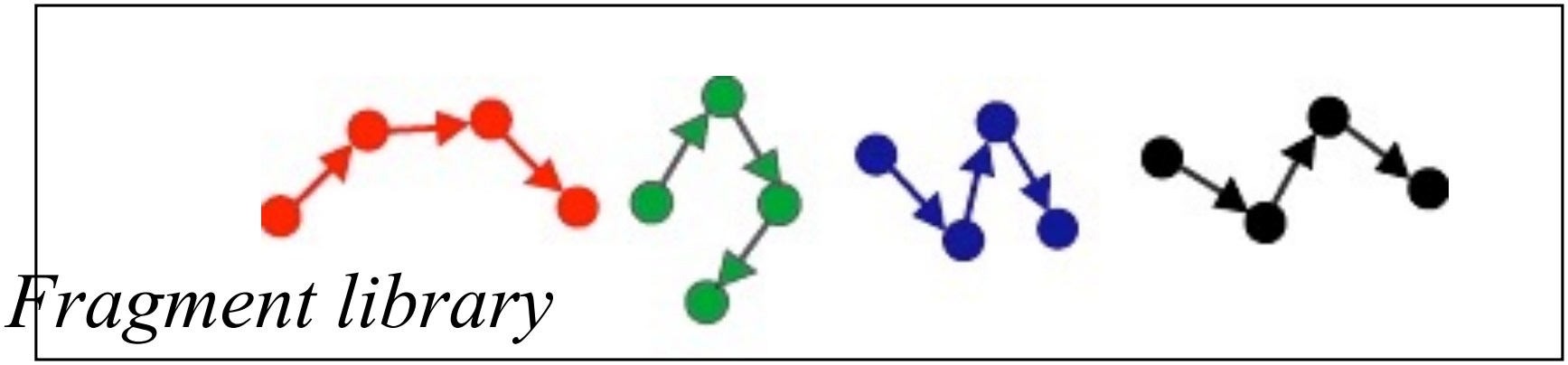


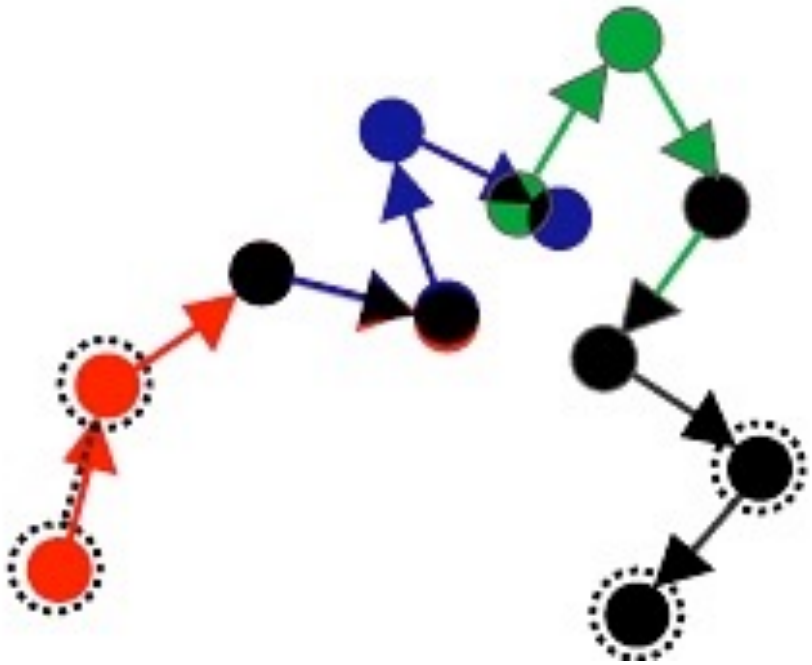
Generating Loops



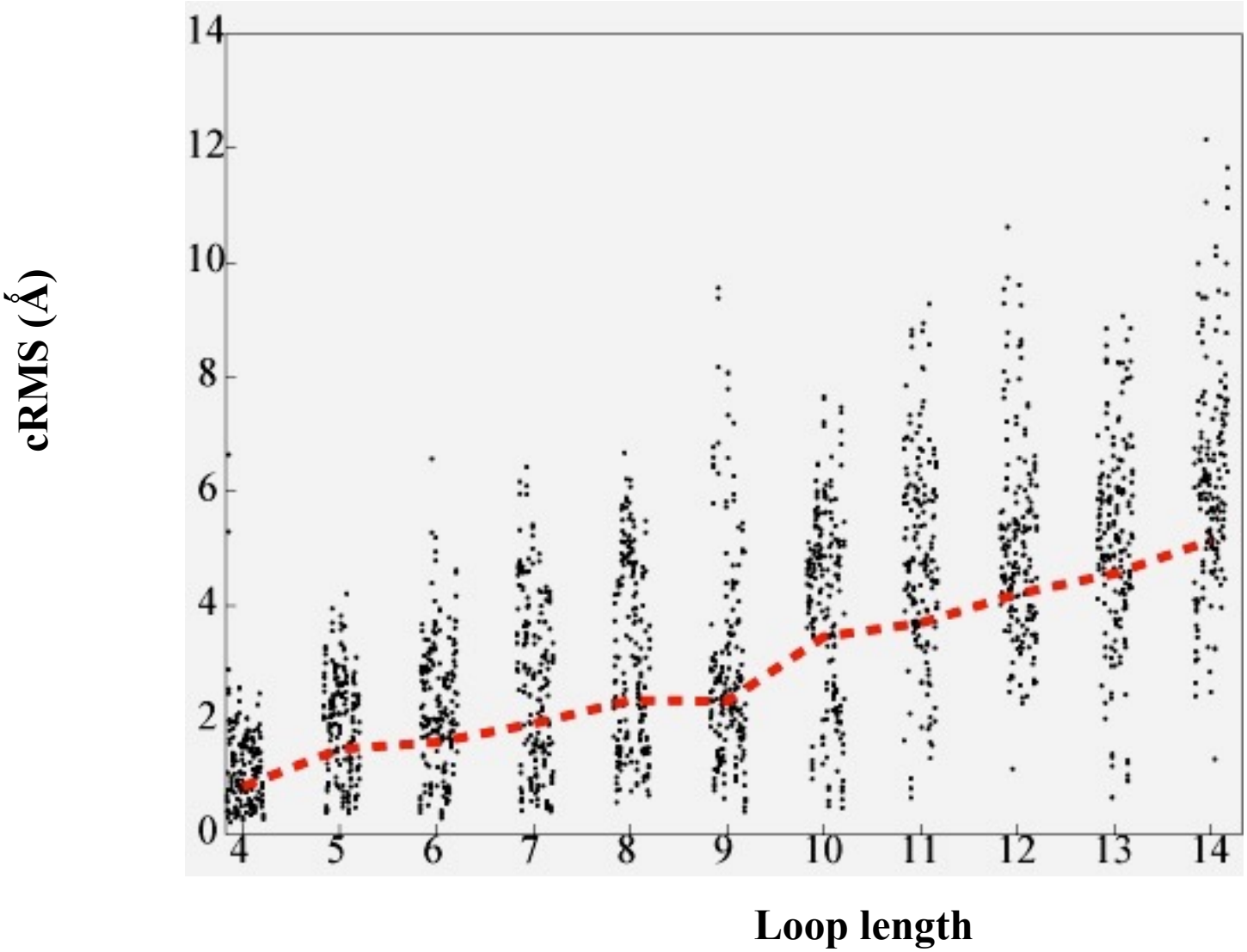


Generating Loops



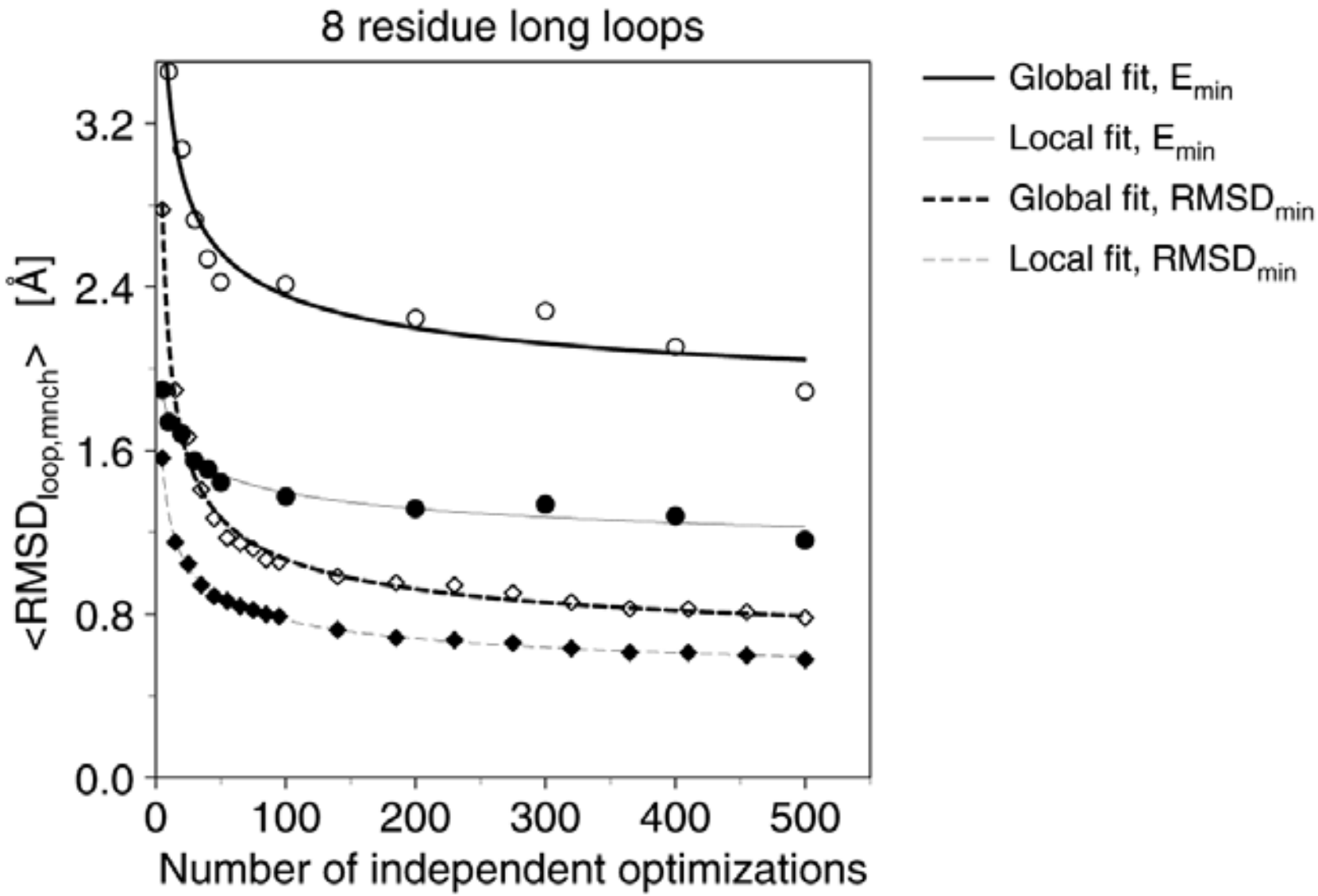


Medium loops: A database approach

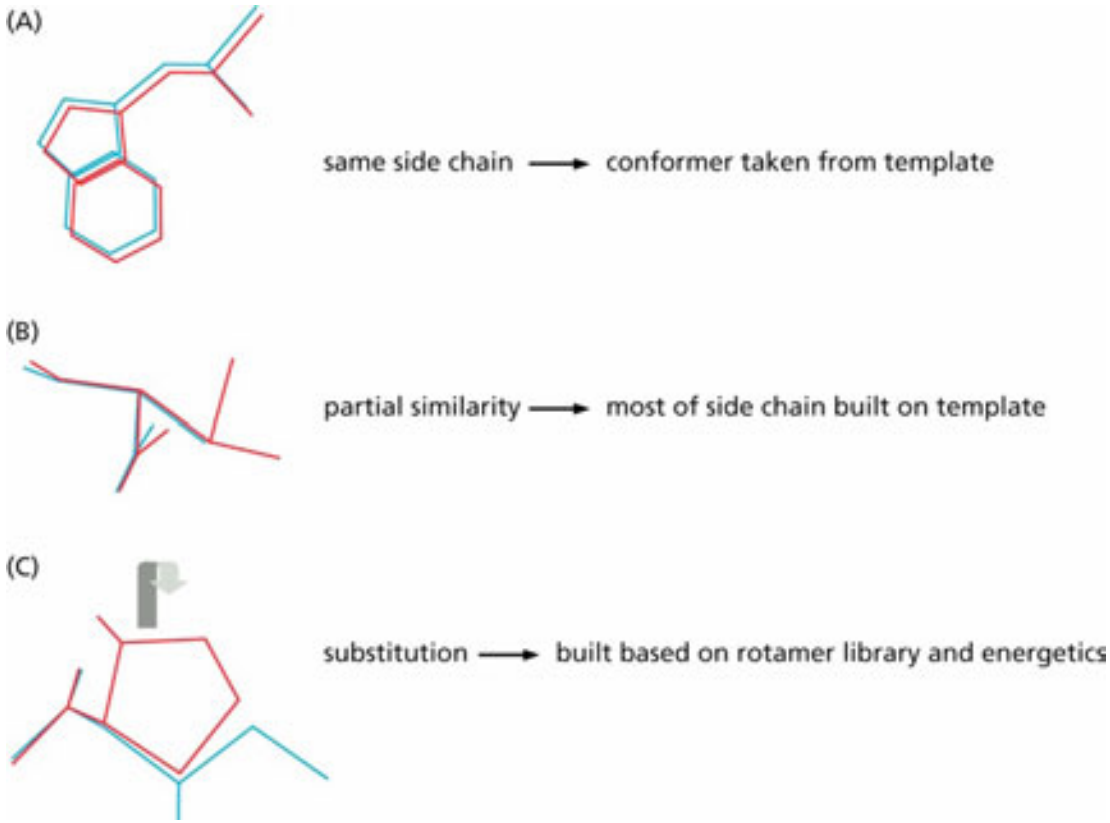


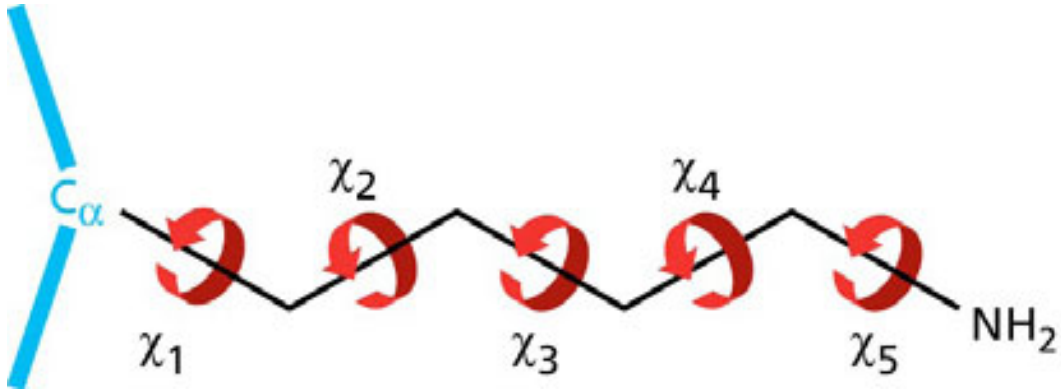
Method breaks down for loops larger than 9

Accuracy of loop models as a function of amount of optimization



Rules for manual methods

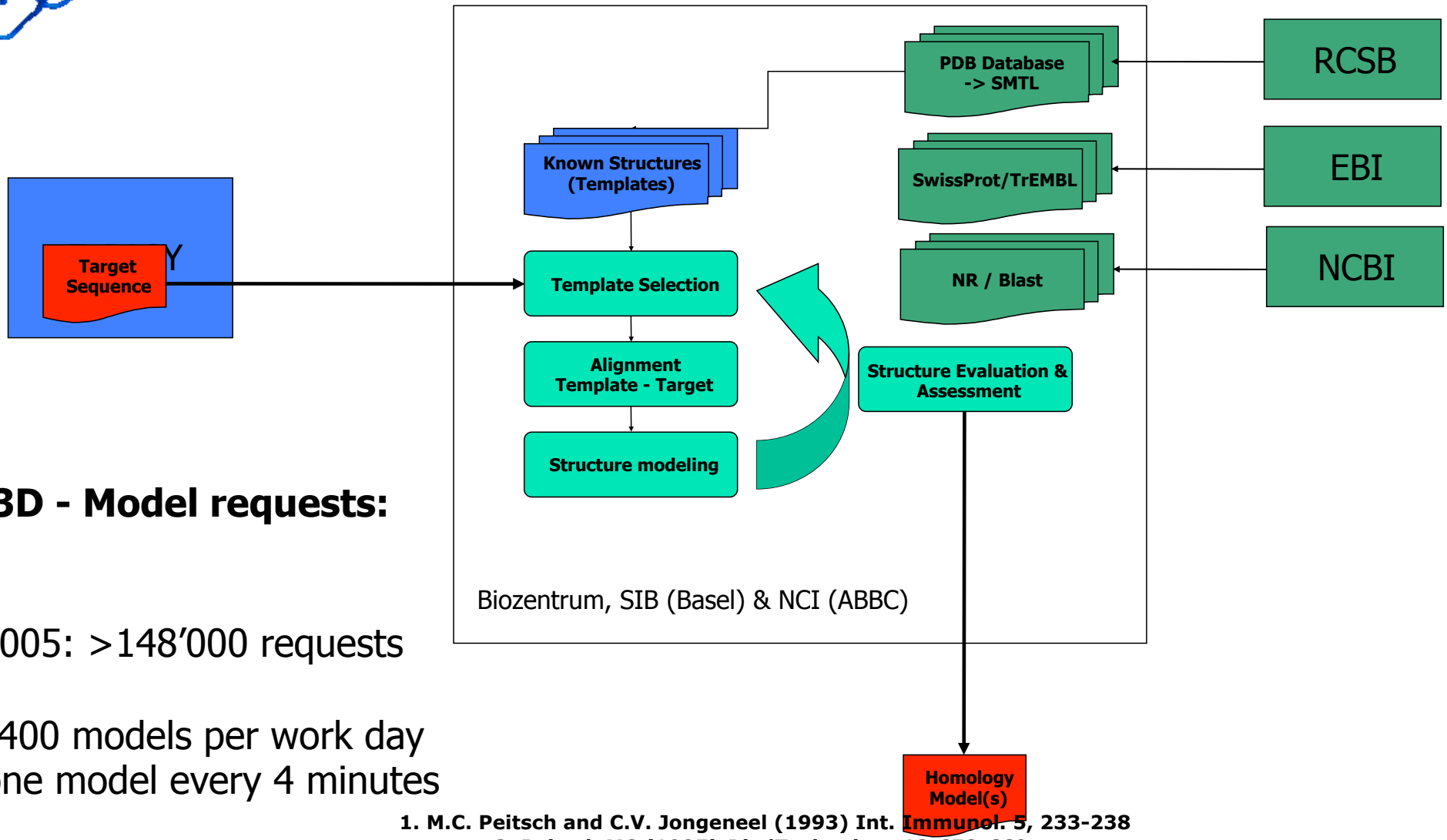




SWISS-MODELLER

- Copy backbone
- Copy sidechains (if they the same)
- Build loops
- Build sidechains

How to automate protein modeling?

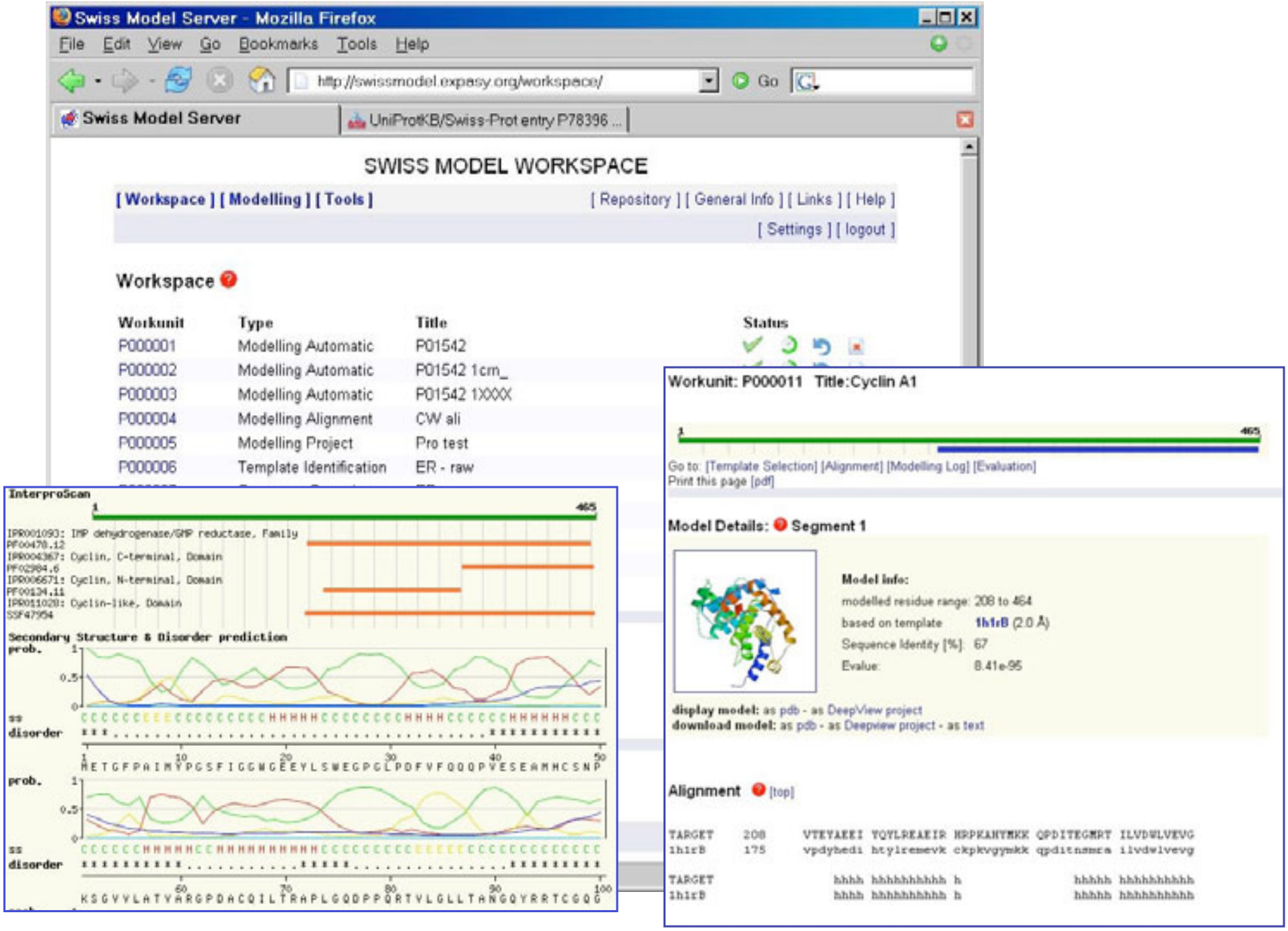


3D - Model requests:

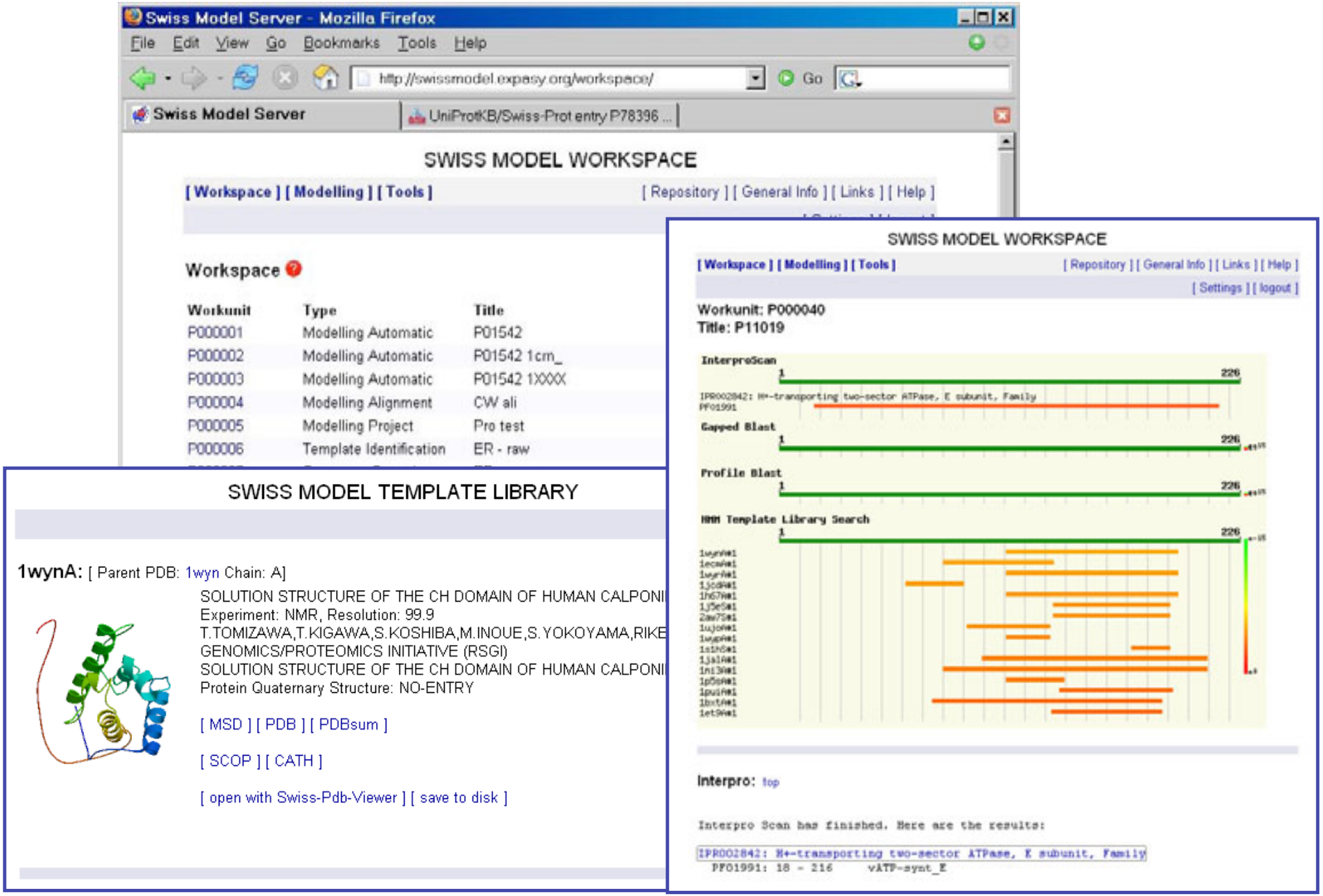
2005: >148'000 requests

~ 400 models per work day
~ one model every 4 minutes

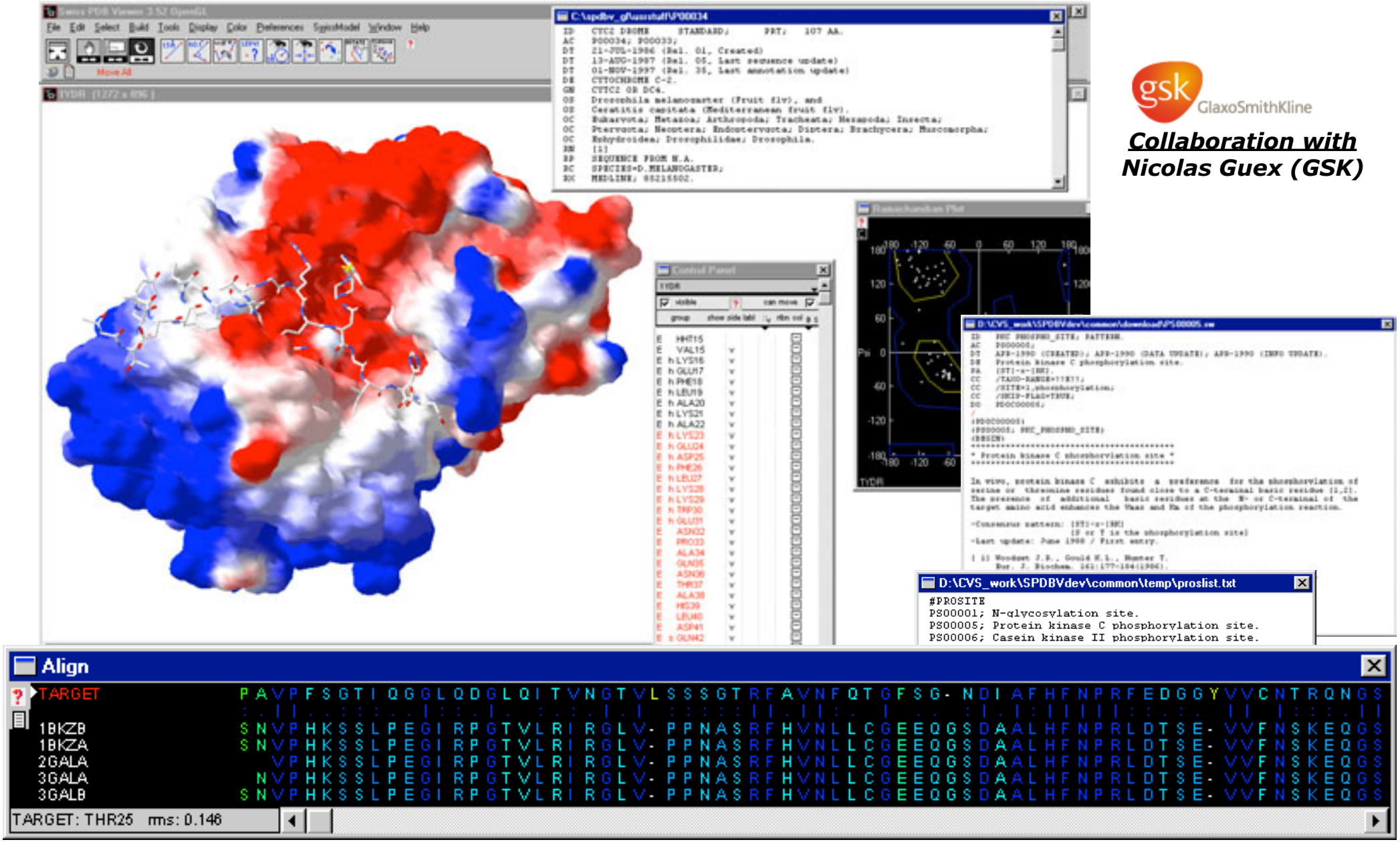
1. M.C. Peitsch and C.V. Jongeneel (1993) Int. Immunol. 5, 233-238
2. Peitsch MC (1995) *Bio/Technology* 13:658-660.
3. Guex N and Peitsch MC (1997) *Electrophoresis* 18:2714-2723.
4. Schwede T, Kopp J, Guex N, Peitsch MC (2003) Nucleic Acids Research 31, 3381-3385.



Arnold K, Bordoli L, Kopp J, Schwede T (2006) *Bioinformatics* **22**,195-201.



Arnold K, Bordoli L, Kopp J, Schwede T (2006) *Bioinformatics* **22**,195-201.

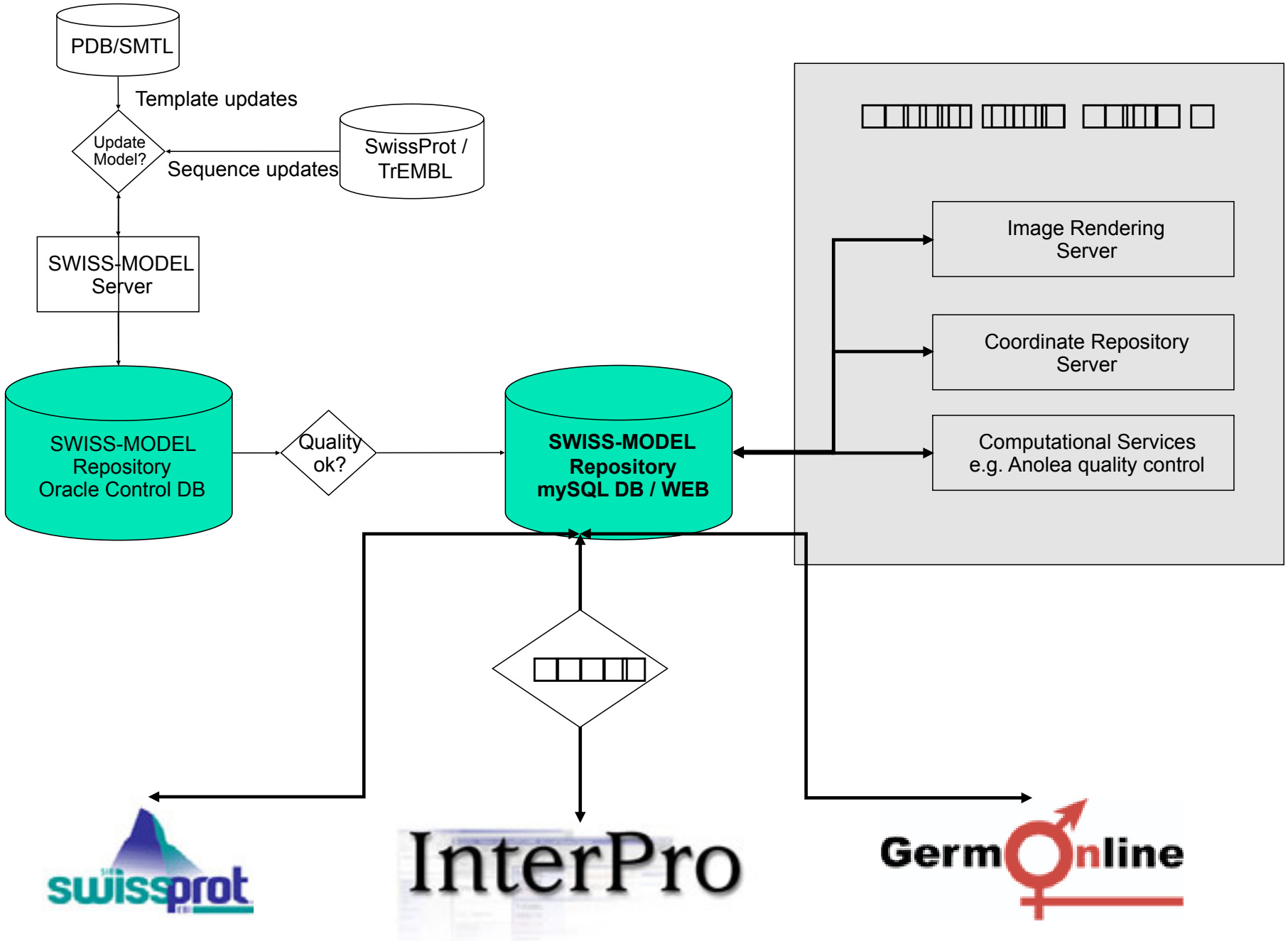


GlaxoSmithKline

Collaboration with Nicolas Guex (GSK)

[http://www.expasy.org/spdbv/]

Guex N and Peitsch MC (1997) *Electrophoresis* 18:2714-2723.



Kopp J, and Schwede T (2004). *Nucleic Acids Research* **32** , D230-D234.

SWISS-MODEL REPOSITORY

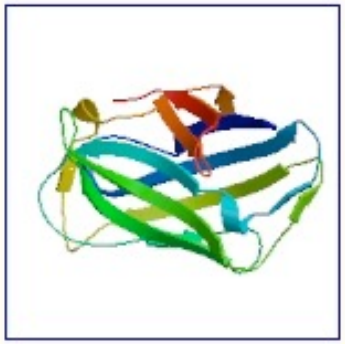
Home | Advanced Search | >>Swiss-Model | HELP Swiss-Prot/TrEMBL AC: search

Model Navigator

5 Models for Cellulosomal scaffoldin precursor from Acetivibrio cellulolyticus;

click on target sequence or model bars

Model Info



model name:	**Q9RPL0_C00002**
residue range:	**973** to **1121** of sequence Q9RPL0
based on template:	**1nbcA.pdb** (X-RAY; 1.80 Å)
	>>PDB >>SCOP >>CATH
sequence identity:	**55.9%** between target and template
alignment e-value:	**1.2E-36**

display model: in pdb format | as DeepView project
download model: in pdb format | as DeepView project | as text

Show Alignment

Anolea/Gromos

Kopp J, and Schwede T (2006). *Nucleic Acids Research* **34** , D315-D318.

SWISS-MODEL Repository

Home | Advanced Search | >>Swiss-Mod

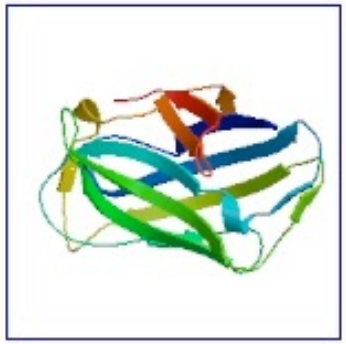
Model

5 Models for Cellulosomal scaffo

Q9RPL0 models

click on target s

Model Info



model name:

residue range

sequence ide

alignment e-v

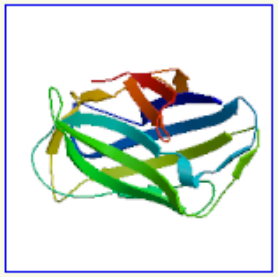
display model: in pdb format | as DeepView

download model: in pdb format | as DeepV

Show Alignment

Anolea/Gromos

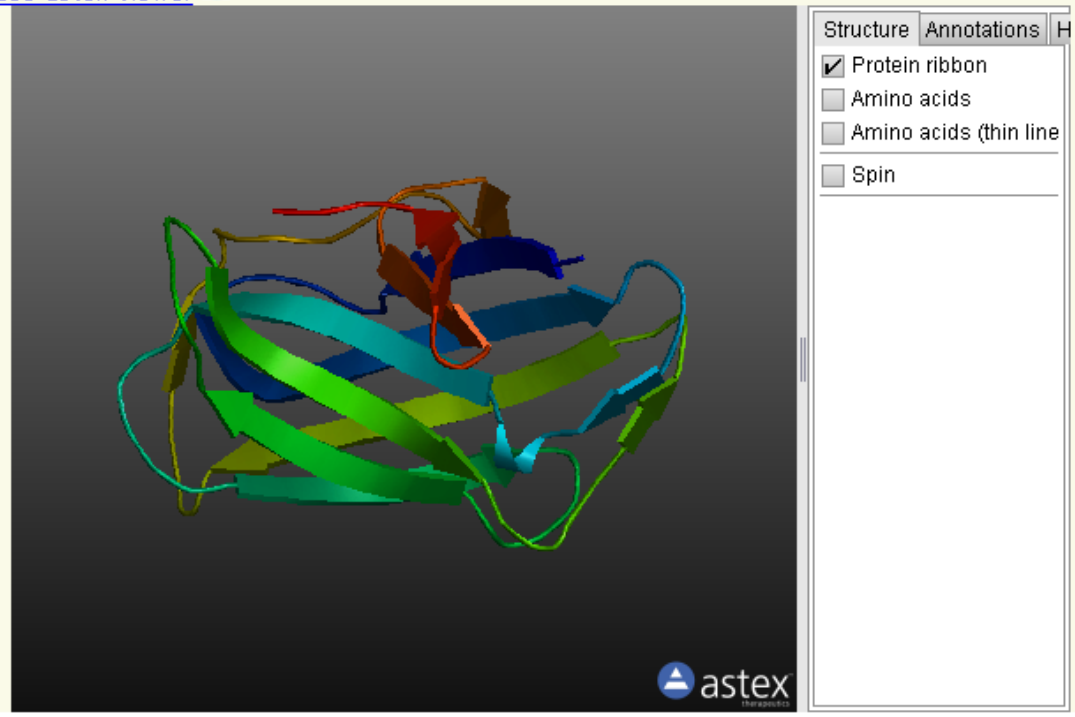
Model Info



interpro domain info:	Bacterial type 3a, cellulose-binding
residue range:	**973** to **1121** of sequence Q9RPL0
based on template:	**1nbcA.pdb** (X-RAY; 1.80 Å) >>PDB >>SCOP >>CATH
sequence identity:	**55.9%** between target and template
alignment e-value:	**2.2E-35**

download model: in pdb format | as DeepView project | as text

close astex viewer



Show Alignment

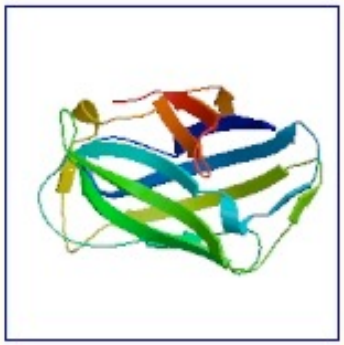
Show Anolea/Gromos

Show Modeling Log

Show Template Selection Log

Kopp J, and Schwede T (2006). *Nucleic Acids Research* **34** , D315-D318.

Model Info



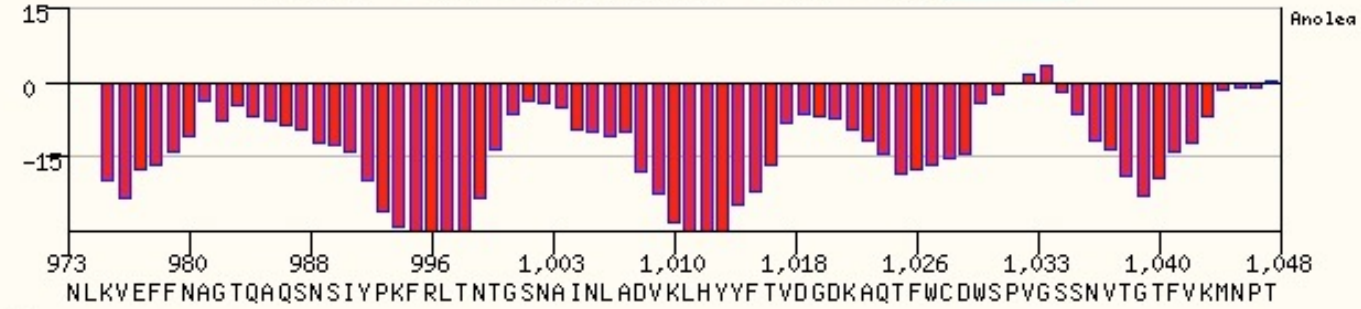
model name:	**Q9RPL0_C00002**
residue range:	**973** to **1121** of sequence Q9RPL0
based on template:	**1nbcA.pdb** (X-RAY; 1.80 Å)
	>>PDB >>SCOP >>CATH
sequence identity:	**55.9%** between target and template
alignment e-value:	**1.2E-36**

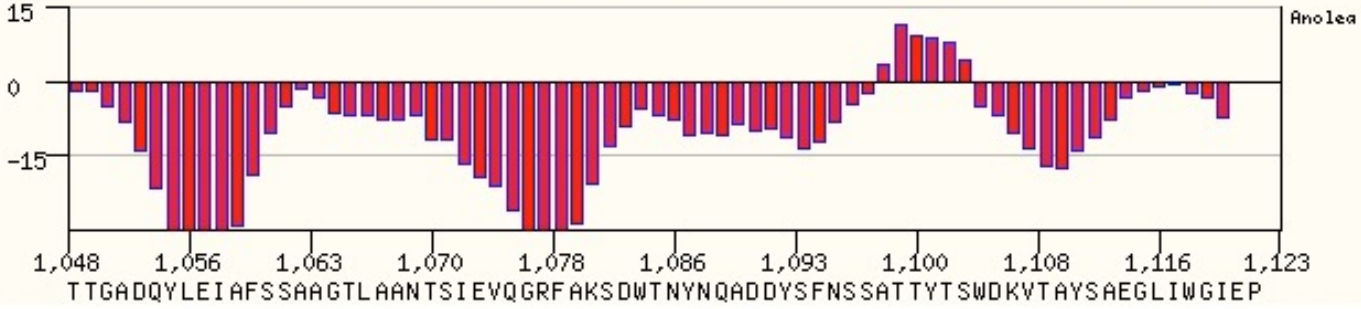
display model: in pdb format | as DeepView project
download model: in pdb format | as DeepView project | as text

Show Alignment

Anolea/Gromos

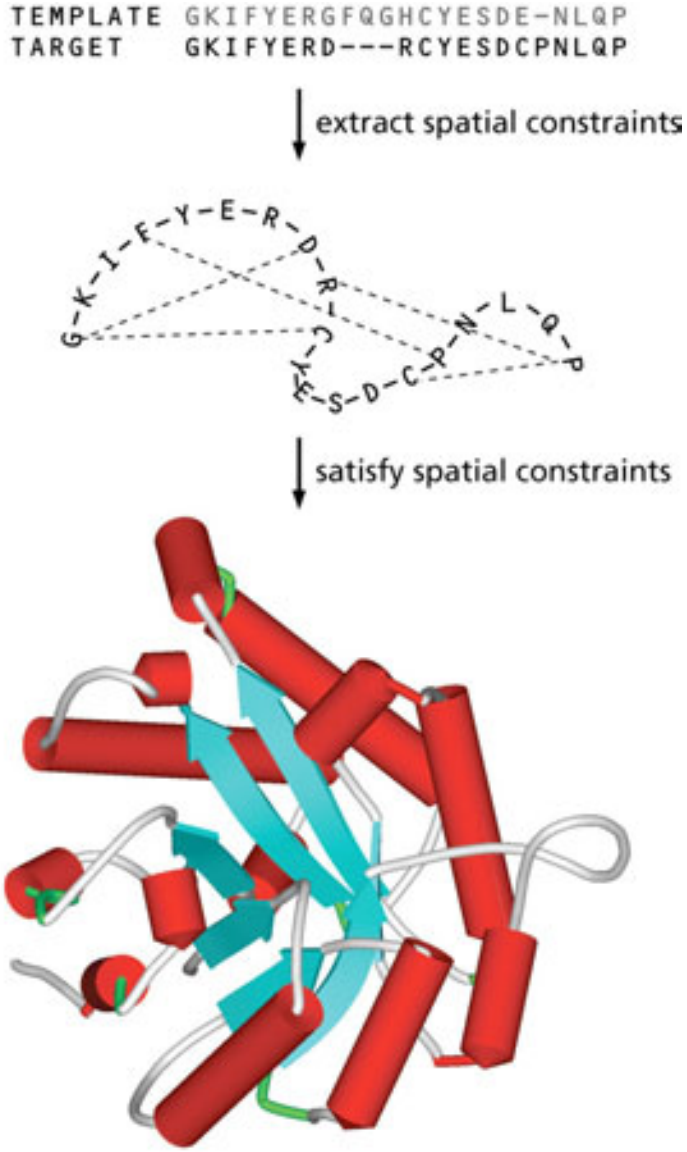
Anolea: on off **Gromos:** on off redraw





Kopp J, and Schwede T (2006). *Nucleic Acids Research* **34** , D315-D318.

Modeler use constraints to model

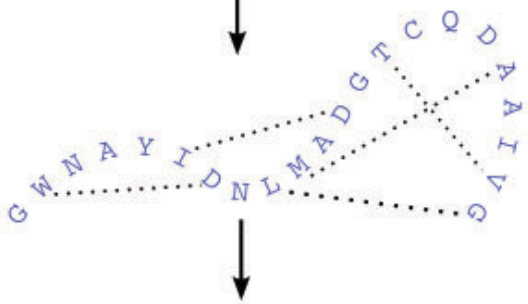


Comparative modeling by satisfaction of spatial restraints - MODELLER

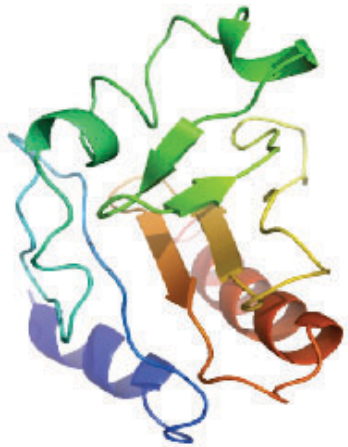
1. Align sequence with structures

Template structure(s) SWQTYVDTNLVGTGAVTQA--AI
Target sequence -GWNAYIDNLMADGTCQDAAIVG

2. Extract spatial restraints



3. Satisfy spatial restraints



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.

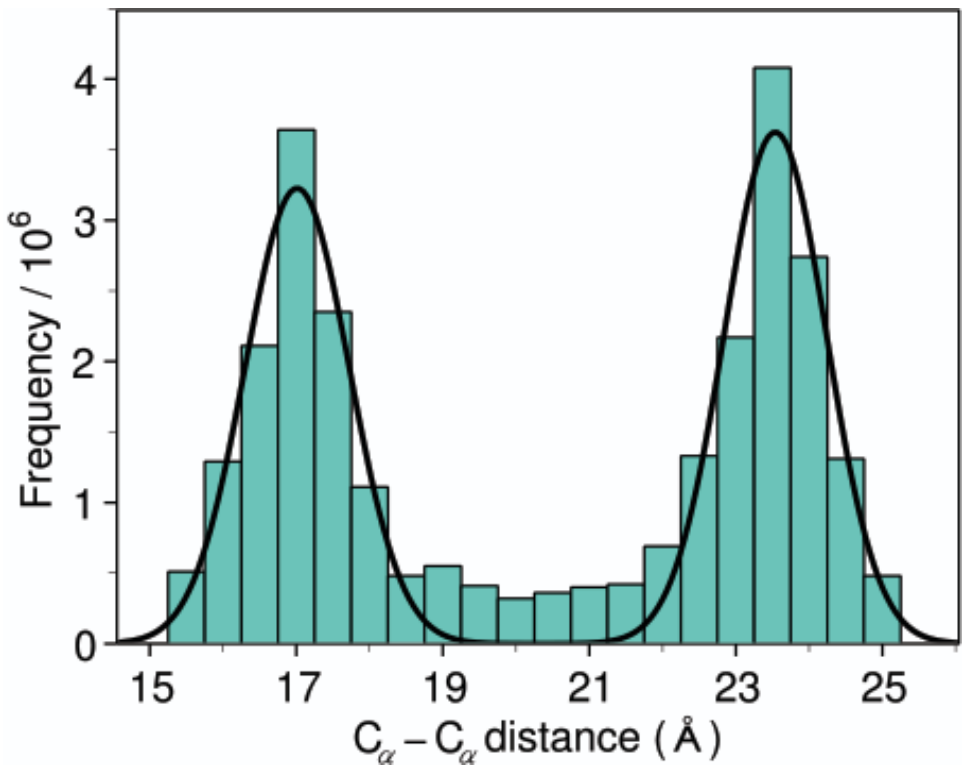
1. Align sequence with structures

- First, must determine the template structures
  - Simplistically, try to align the target sequence against every known structure's sequence
  - In practice, this is too slow, so heuristics are used (e.g. BLAST)
  - Profile or HMM searches are generally more sensitive in difficult cases (e.g. Modeller's profile.build method, or PSI-BLAST)
  - Could also use threading or other web servers
- Alignment to templates generally uses global dynamic programming
  - Sequence-sequence: relies purely on a matrix of observed residue-residue mutation probabilities ('align')
  - Sequence-structure: gap insertion is penalized within secondary structure (helices etc.) ('align2d')
  - Other features and/or user-defined ('salign') or use an external program

2. Extract spatial restraints

Spatial restraints incorporate homology information, statistical preferences, and physical knowledge

- Template Cα- Cα internal distances
- Backbone dihedrals (φ/ψ)
- Sidechain dihedrals given residue type of both target and template
- Force field stereochemistry (bond, angle, dihedral)
- Statistical potentials
- Other experimental constraints
- etc.



3. Satisfy spatial restraints

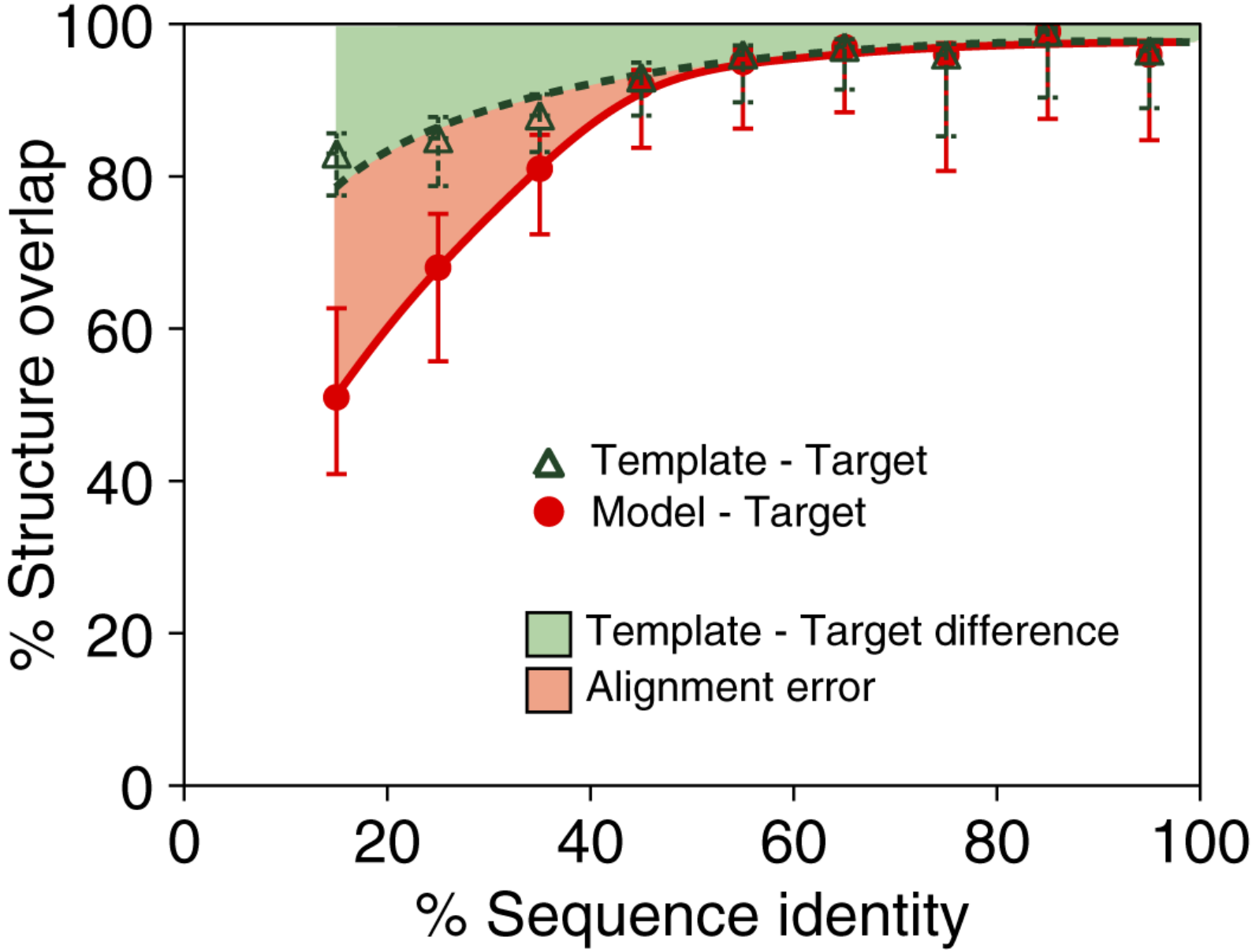
All information is combined into a single objective function

- Restraints and statistics are converted to an “energy” by taking the negative log
- Force field (CHARMM 22) simply added in

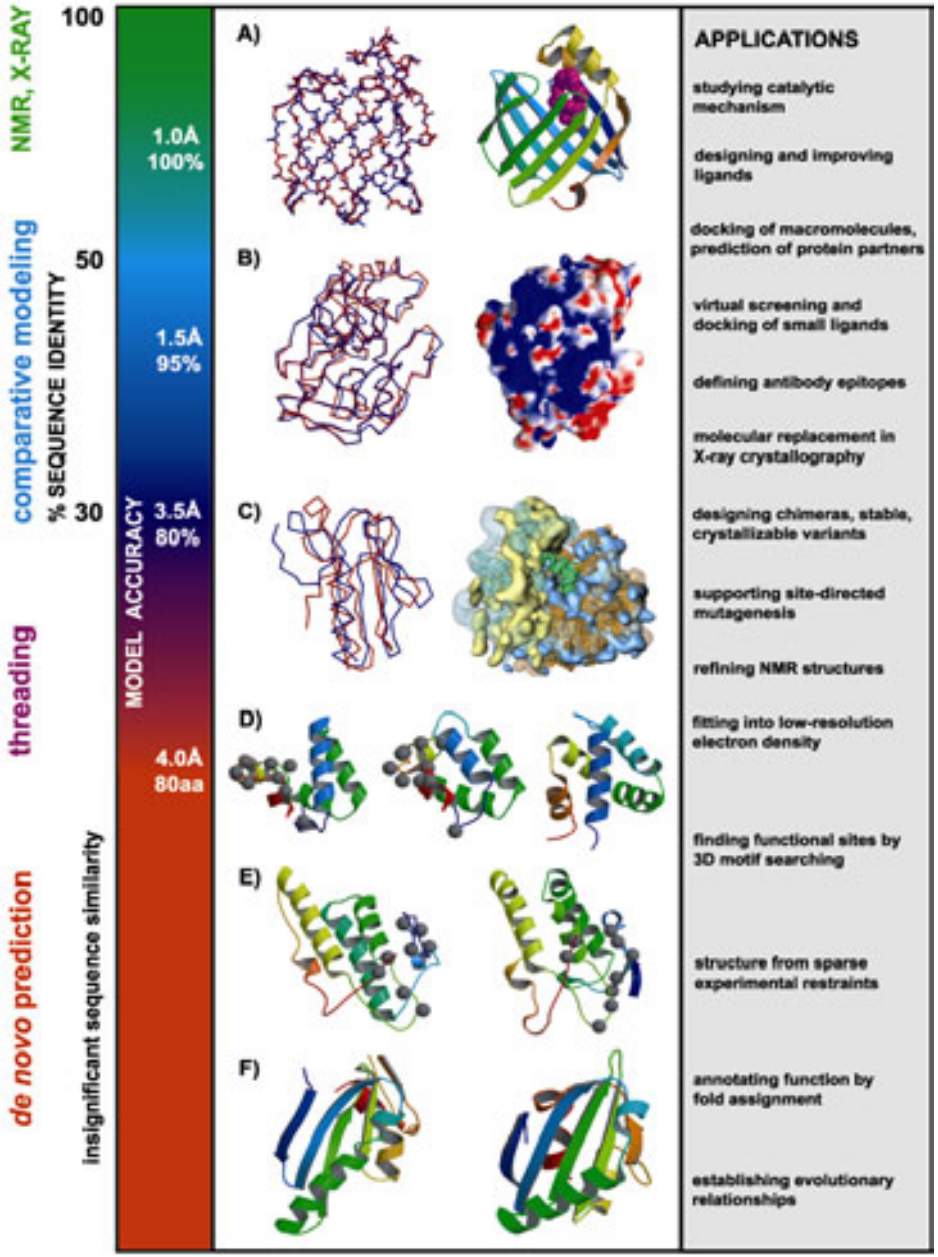
Function is optimized by conjugate gradients and simulated annealing molecular dynamics, starting from the target sequence threaded onto template structure(s)

Multiple models are generally recommended; ‘best’ model or cluster or models chosen by simply taking the lowest objective function score, or using a model assessment method such as Modeller’s own DOPE or GA341, fit to EM density, or external programs such as PROSA or DFIRE

Model Accuracy as a Function of Target-Template Sequence Identity



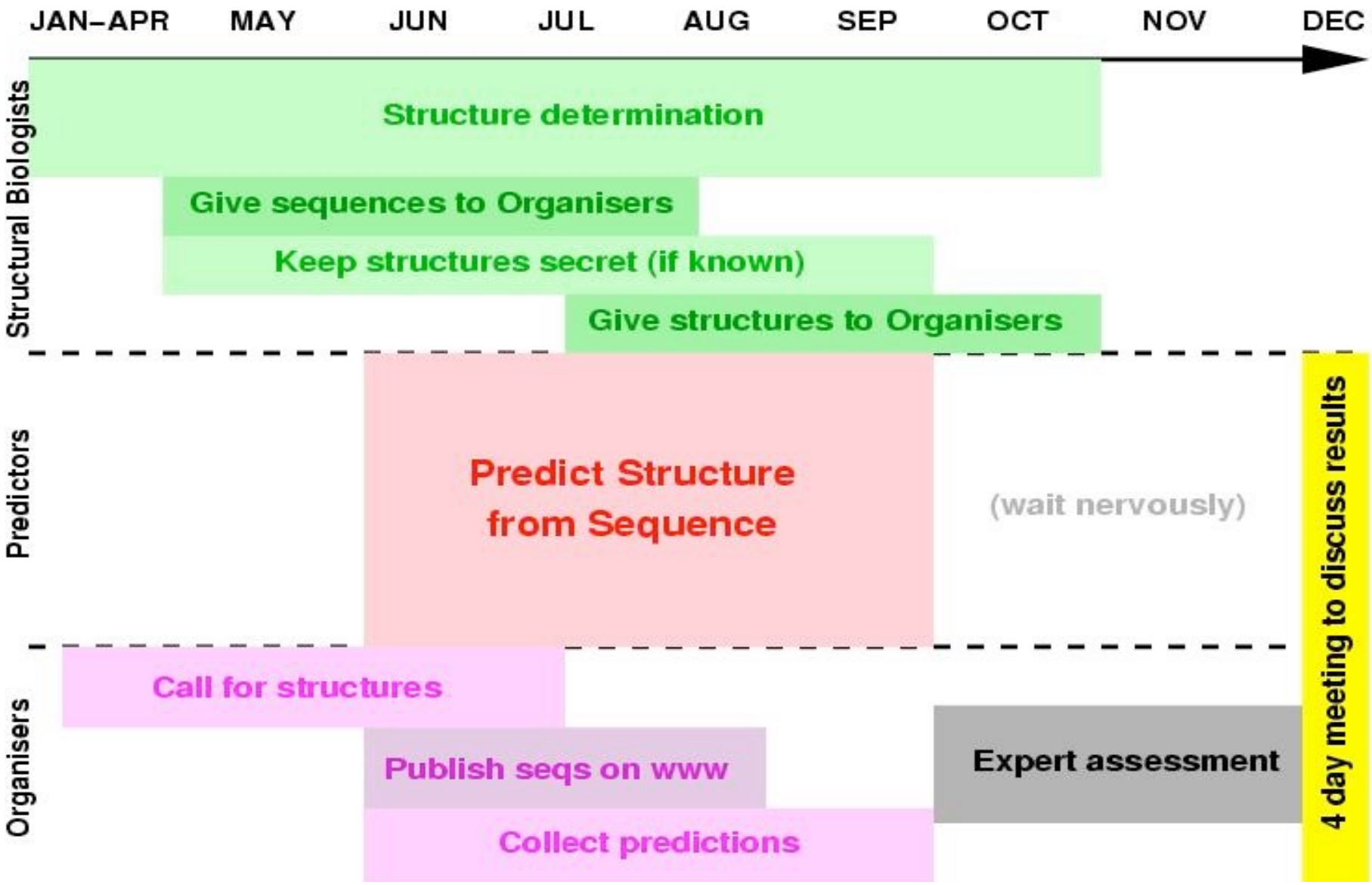
Applications of models



The CASP experiment

- *CASP= Critical Assessment of Structure Prediction*
- *Started in 1994, based on an idea from John Moult (Moult, Pederson, Judson, Fidelis, Proteins, 23:2-5 (1995))*
- *First run in 1994; now runs regularly every second year (CASP7 was held last December)*

CASP


JAN–APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
Structural Biologists
Structure determination
Give sequences to Organisers
Keep structures secret (if known)
Give structures to Organisers
Predictors
Predict Structure from Sequence
(wait nervously)
Organisers
Call for structures
Publish seqs on www
Collect predictions
Expert assessment
4 day meeting to discuss results

The CASP experiment: how it works

1) Sequences of target proteins are made available to CASP participants in June-July of a CASP year

- the structure of the target protein is know, but not yet released in the PDB, or even accessible

2) CASP participants have between 2 weeks and 2 months over the summer of a CASP year to generate up to 5 models for each of the target they are interested in.

3) Model structures are assessed against experimental structure

4) CASP participants meet in December to discuss results

CASP Statistics

Experiment	# of Targets	# of predictors	# of 3D models
CASP1	33	35	100
CASP2	42	72	947
CASP3	43	61	1256
CASP4	43	111	5150
CASP5	67	175	22909
CASP6	87	166	28965

CASP

Three categories at CASP

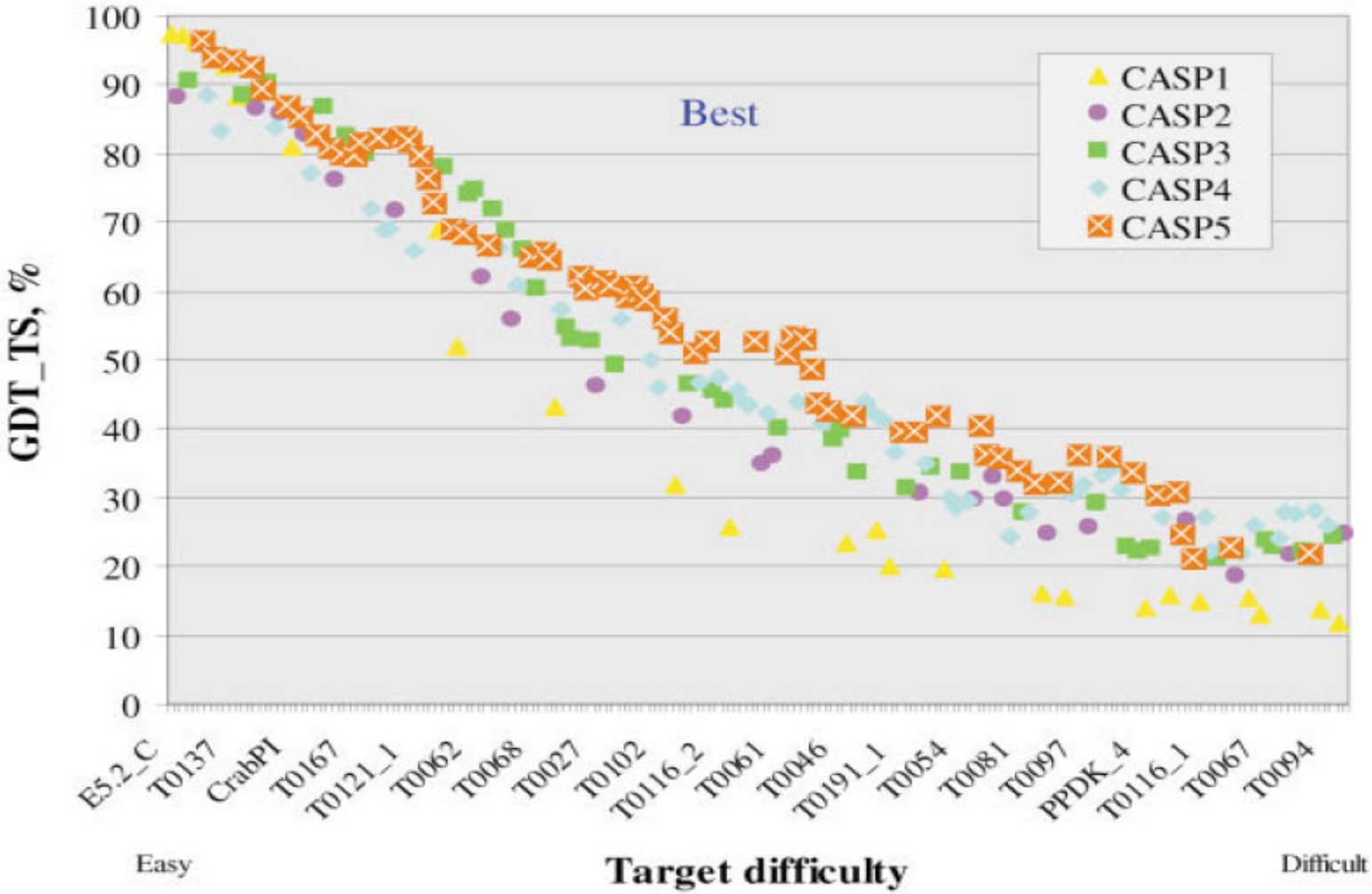
- Homology (or comparative) modeling
- Fold recognition
- Ab initio/new folds prediction

CASP dynamics:

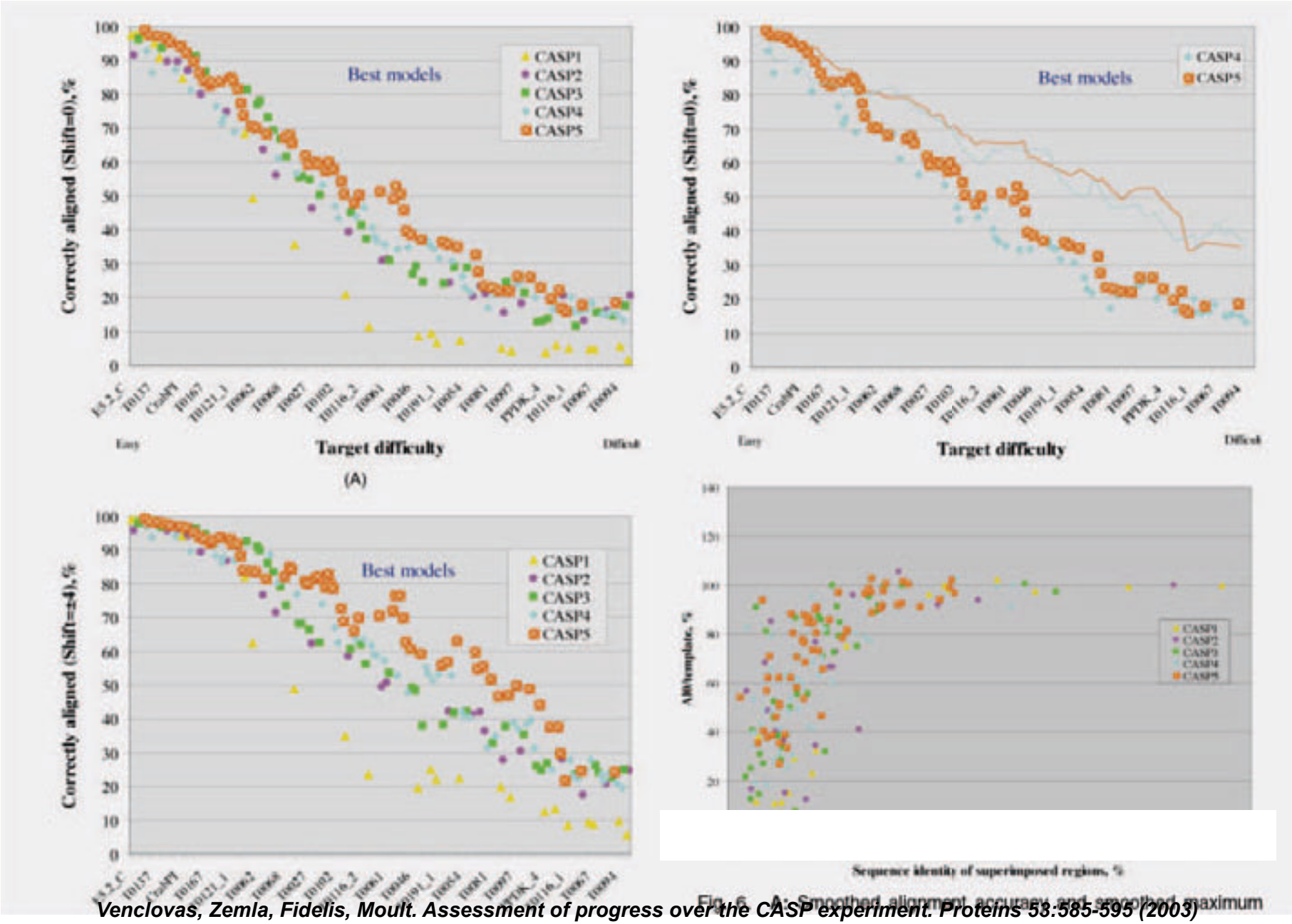
- Real deadlines; pressure: positive, or negative?
- Competition?
- Influence on science ?

Venclovas, Zemla, Fidelis, Moult. Assessment of progress over the CASP experiments. Proteins, 53:585-595 (2003)

CASP Progress

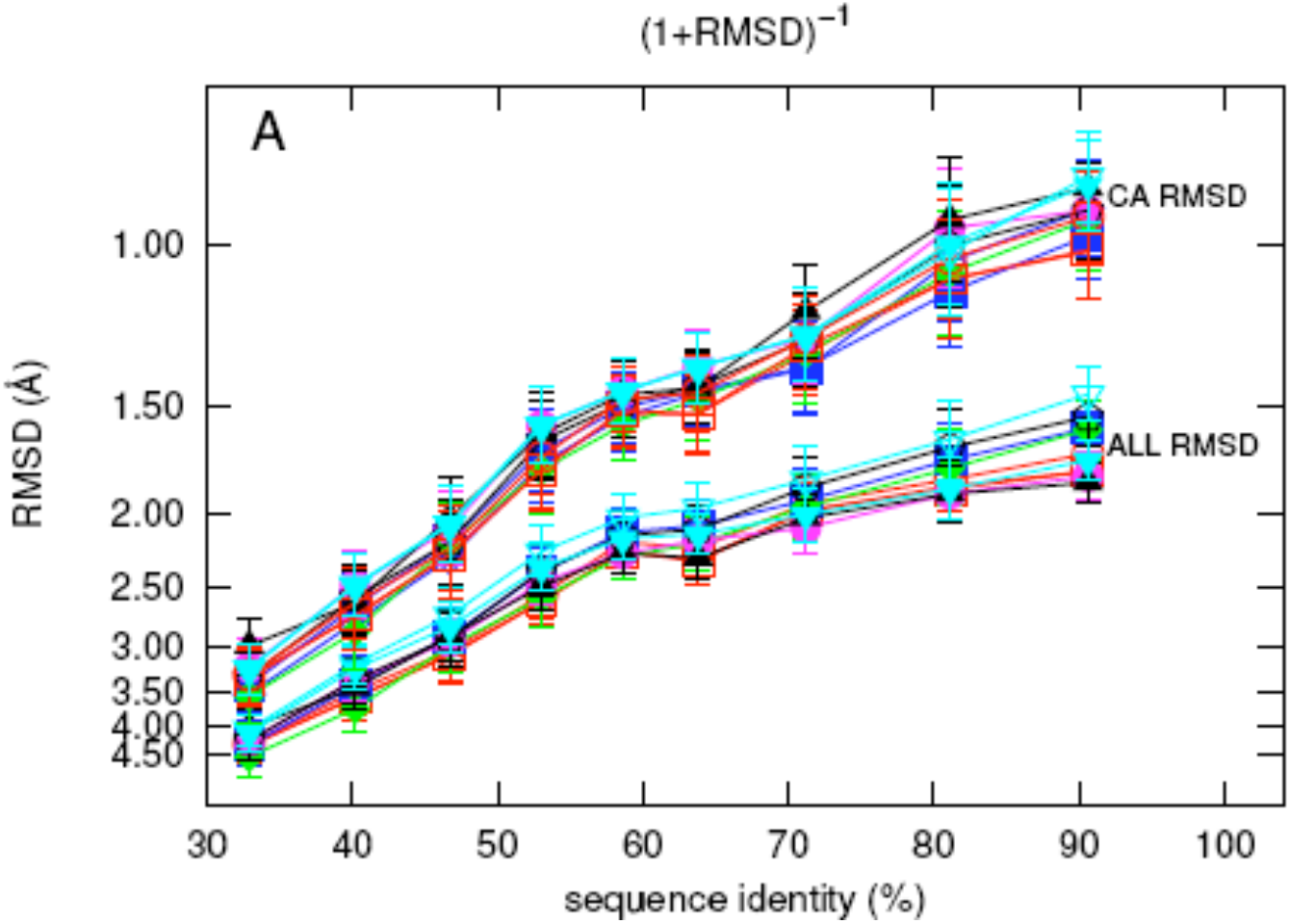


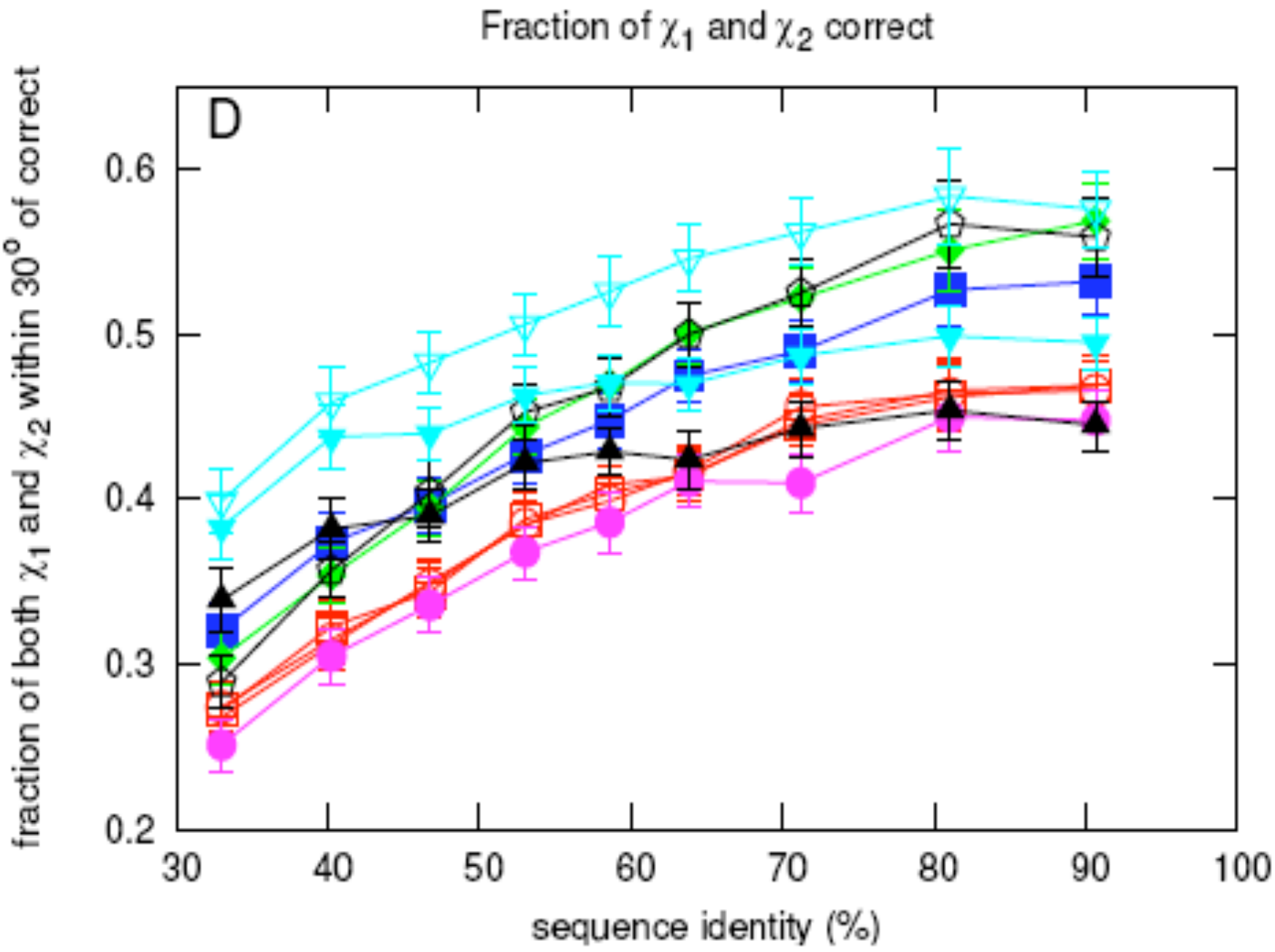
CASP: quality of alignment

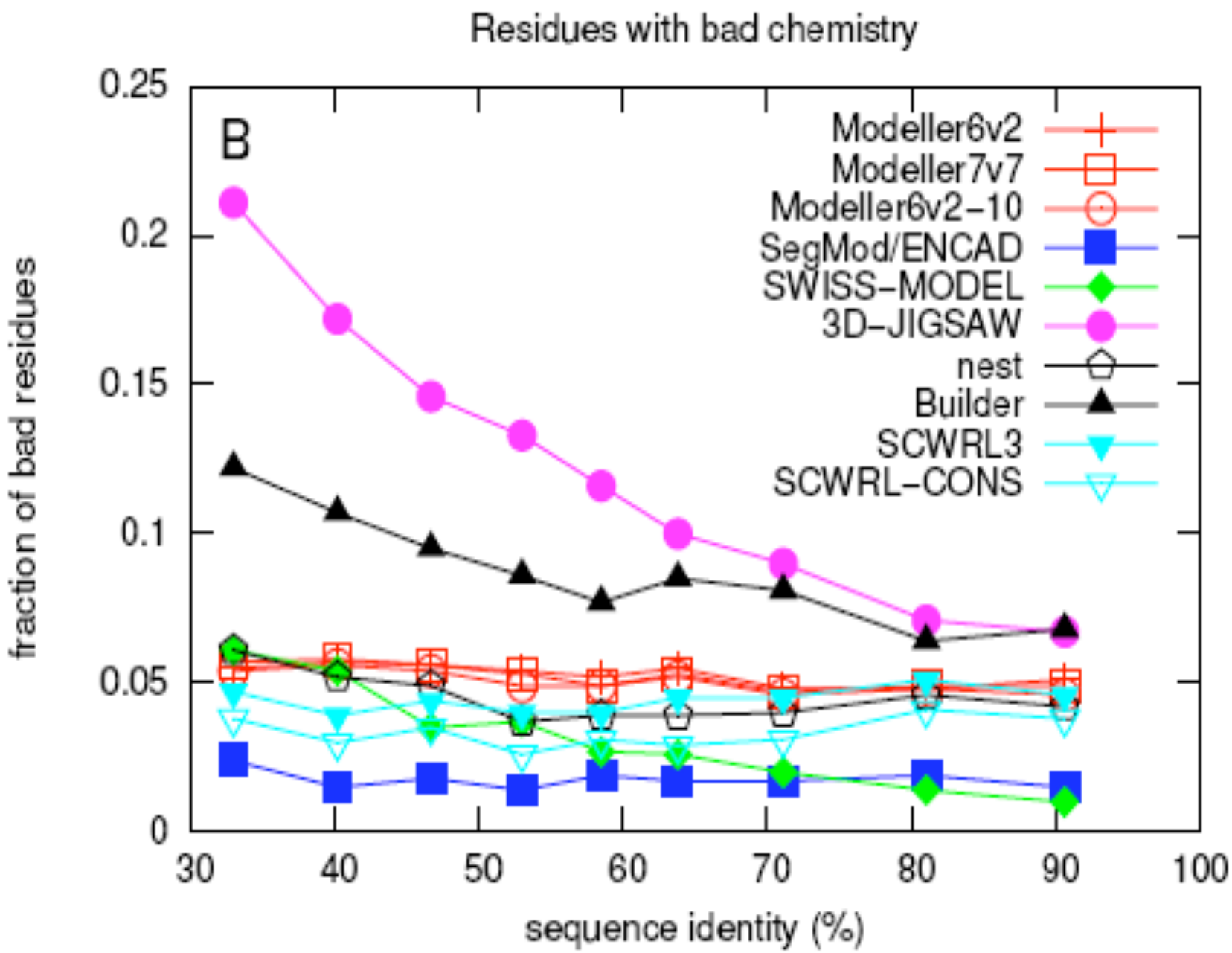


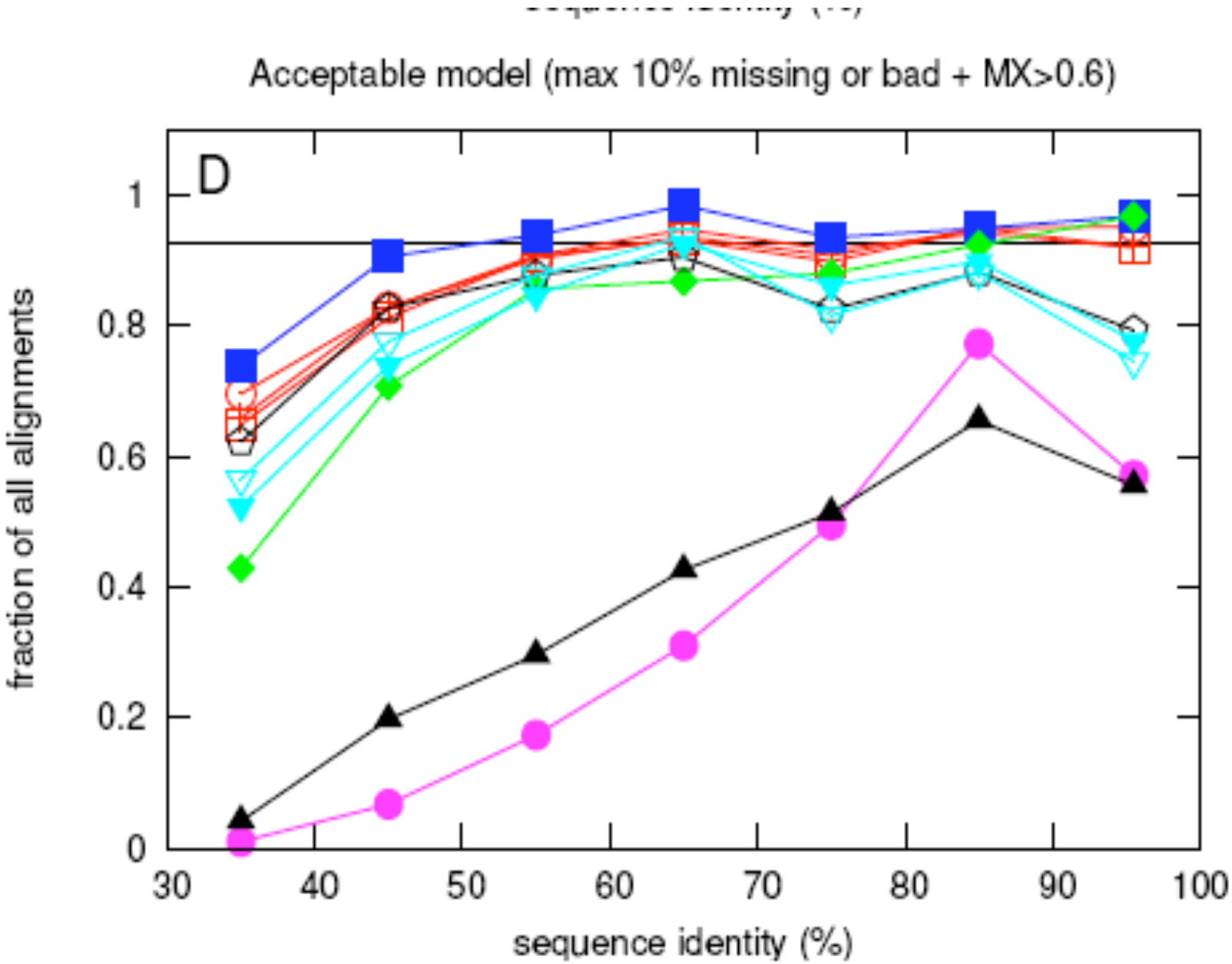
Venclovas, Zemla, Fidelis, Moult. Assessment of progress over the CASP experiment. Proteins 53:585-595 (2003)

HM benchmark: RMSD

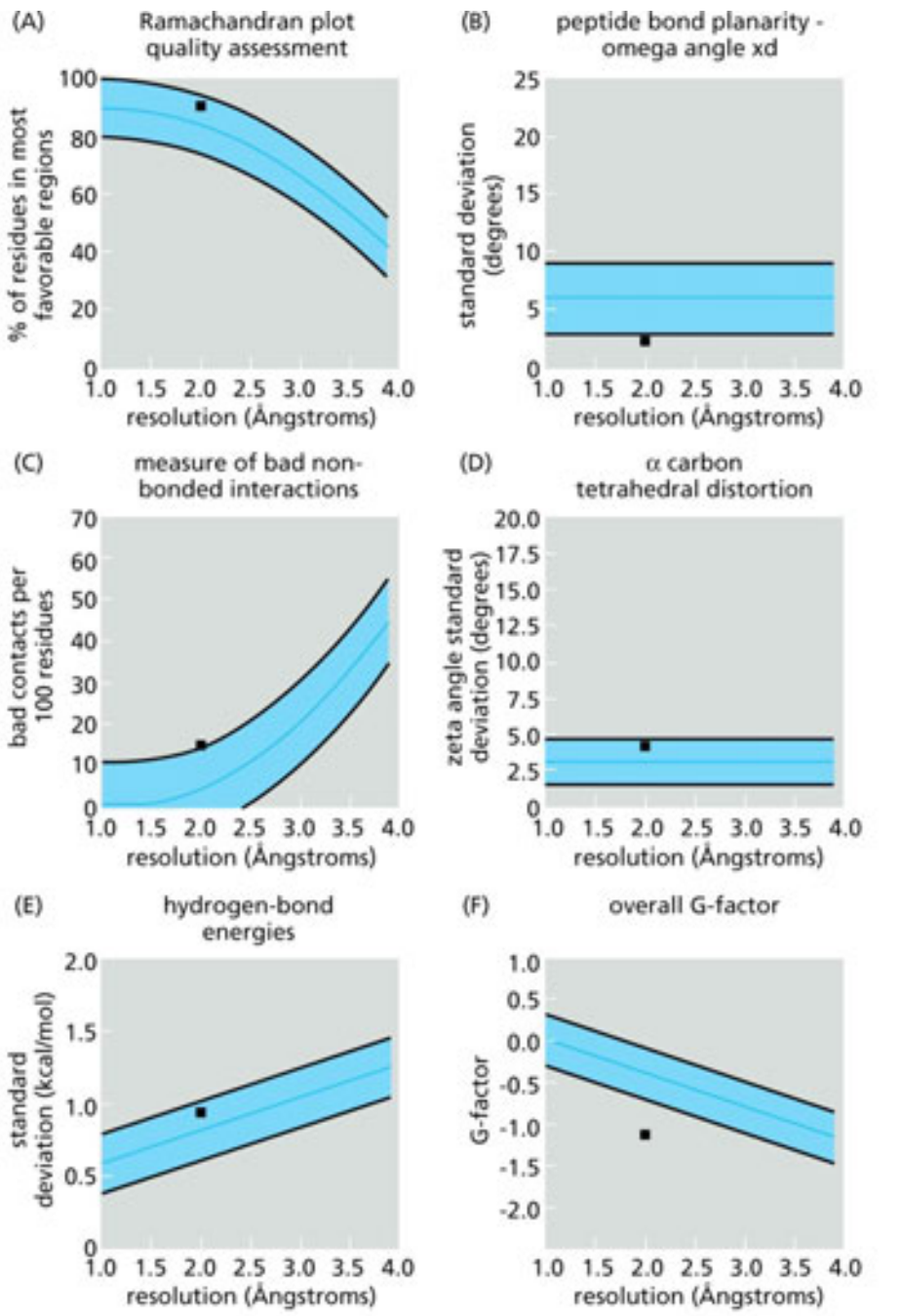



Fraction of χ_1 and χ_2 correct
D
fraction of both χ_1 and χ_2 within 30° of correct
0.2
0.3
0.4
0.5
0.6
30
40
50
60
70
80
90
100
sequence identity (%)


Residues with bad chemistry
B
Modeller6v2
Modeller7v7
Modeller6v2-10
SegMod/ENCAD
SWISS-MODEL
3D-JIGSAW
nest
Builder
SCWRL3
SCWRL-CONS
fraction of bad residues
0.25
0.2
0.15
0.1
0.05
0
30
40
50
60
70
80
90
100
sequence identity (%)


Acceptable model (max 10% missing or bad + MX>0.6)
D
fraction of all alignments
sequence identity (%)
0
0.2
0.4
0.6
0.8
1
30
40
50
60
70
80
90
100

Checks of the structure



PLOT STATISTICS

Typical errors in comparative modeling

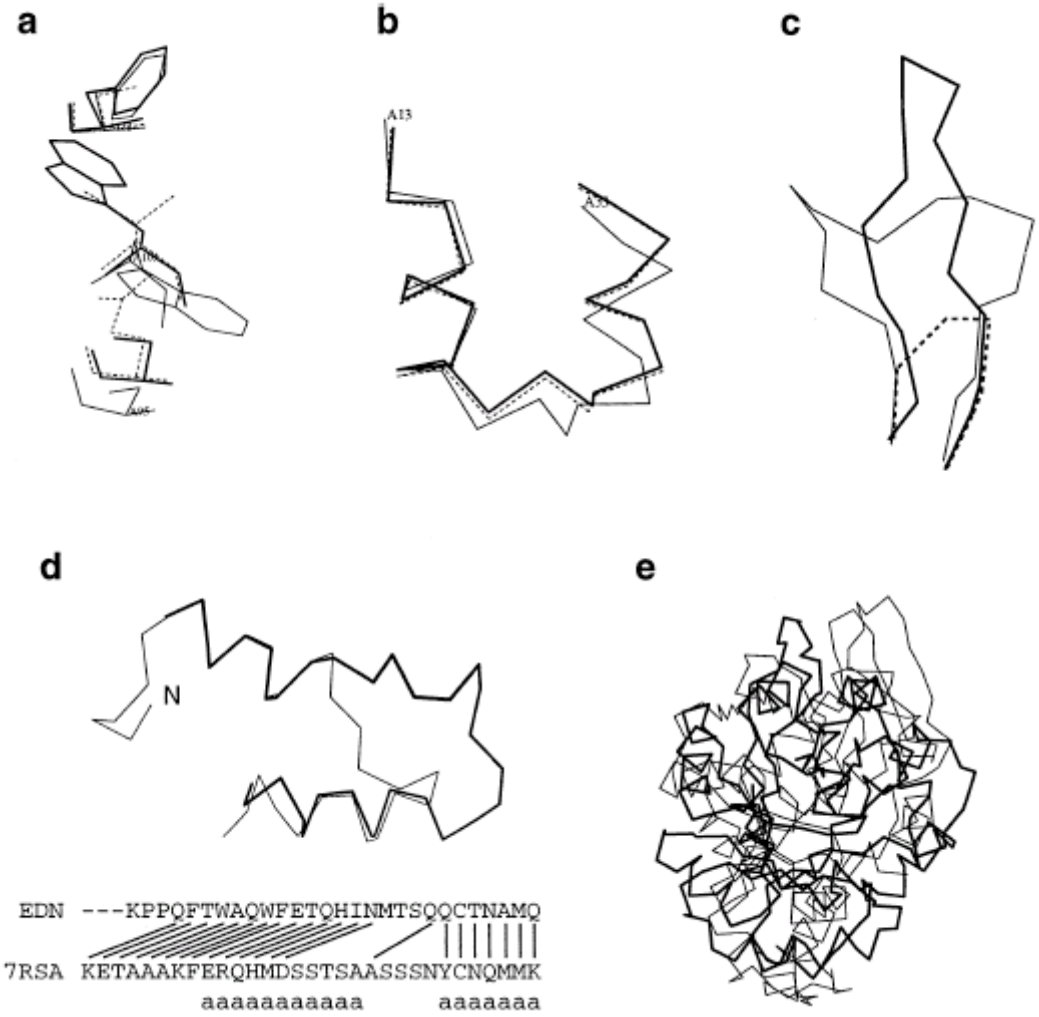
a) Side chain packing

b) Distortions and shifts

c) Loops

d) Misalignments

e) Incorrect template



True structure

Template

Model

Marti-Renom et al. (2000)

ProQ

(Wallner & Elofsson, 2003)

- Neural networks that predicts quality of proteins
- Trained on LiveBench models
- Predicts "quality" with
  - Cc=0.76
  - Z-score 2.7
  - Z-native 5.1

- Uses full atom models with these parameters:
  - Atom-Atom contacts (13 types)
  - Residue- Residue contacts (6 types)
  - Surface area (4 categories)
  - SecStr-Q3 compared to psipred
  - difference in $C\alpha$ between model template
  - fatness of model
  - fraction modeled

Development of ProQ

- Trained on 11108 LB2 models
- Quality measured with MaxSub and LGscore
- 1 894 correct models
- 8 270 incorrect models
- All atom models built by MODELLER
- MaxSub and LGscore predicted
- Testing different input parameters

Input parameters

- Atom-Atom and Residue-Residue contacts
- Fraction of contacts of a particular type
- Similar to Errat, (Colovos and Yeates, 1993)
- Different binning of atoms and contacts

Contact parameters

Training data	Predicting LGscore Correlation/Z-score	Predicting MaxSub Correlation/Z-score
Atom-3 contacts	0.43/0.9	0.33/0.9
Atom-13 contacts	0.52/1.2	0.42/1.1
Residue-6 contacts	0.49/1.2	0.37/1.0
Residue-20 contacts	0.40/0.9	0.28/0.7
Atom-13 + Residue-6	0.58/1.5	0.47/1.4

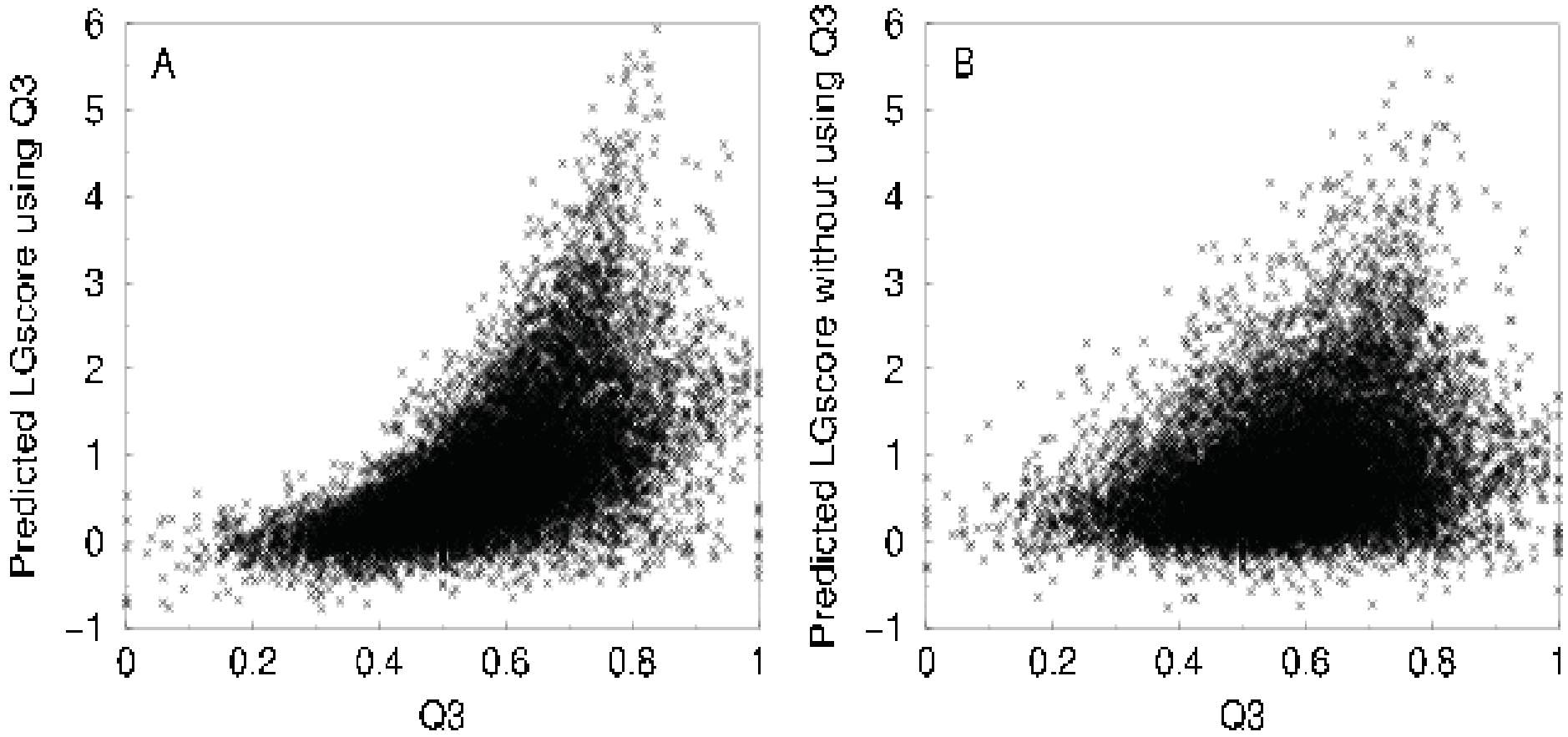
Surface parameters

Training data	Predicting LGscore Correlation/Z-score	Predicting MaxSub Correlation/Z-score
Surface accessibility less than 25%	0.53/1.3	0.40/1.3
Surface accessibility 25%-50%	0.28/0.4	0.07/0.2
Surface accessibility all	0.55/1.4	0.49/1.4

ProQ parameters

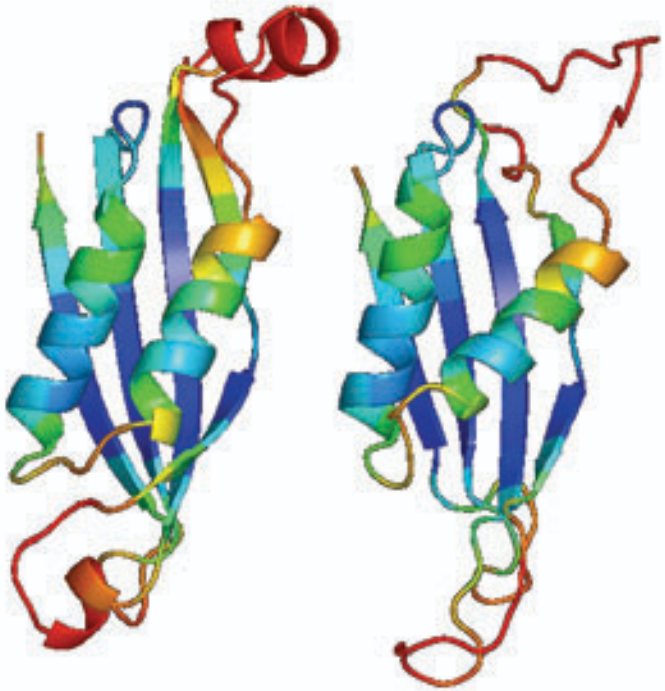
Input parameters	LGscore Cc/Z-score	MaxSub Cc/Z-score
Atom-13 + Residue-6	0.53/1.4	0.41/1.2
Atom-13 + Resi due-6 + Surface all	0.63/1.9	0.51/1.6
Atom-13 + Resi due-6 + Surface all + Q3	0.71/2.4	0.60/2.2
Atom-13 + Resi due-6 + Surface all + Q3 + Cα	0.75/2.6	0.61/2.3
Atom-13 + Resi due-6 + Surface all + Q3 + C α + fatness	0.75/2.7	0.64/2.4
Atom-13 + Resi due-6 + Surface all + Q3 + C α + fatness + frac	0.76/2.7	0.72/2.7

Use of secondary structure in ProQ

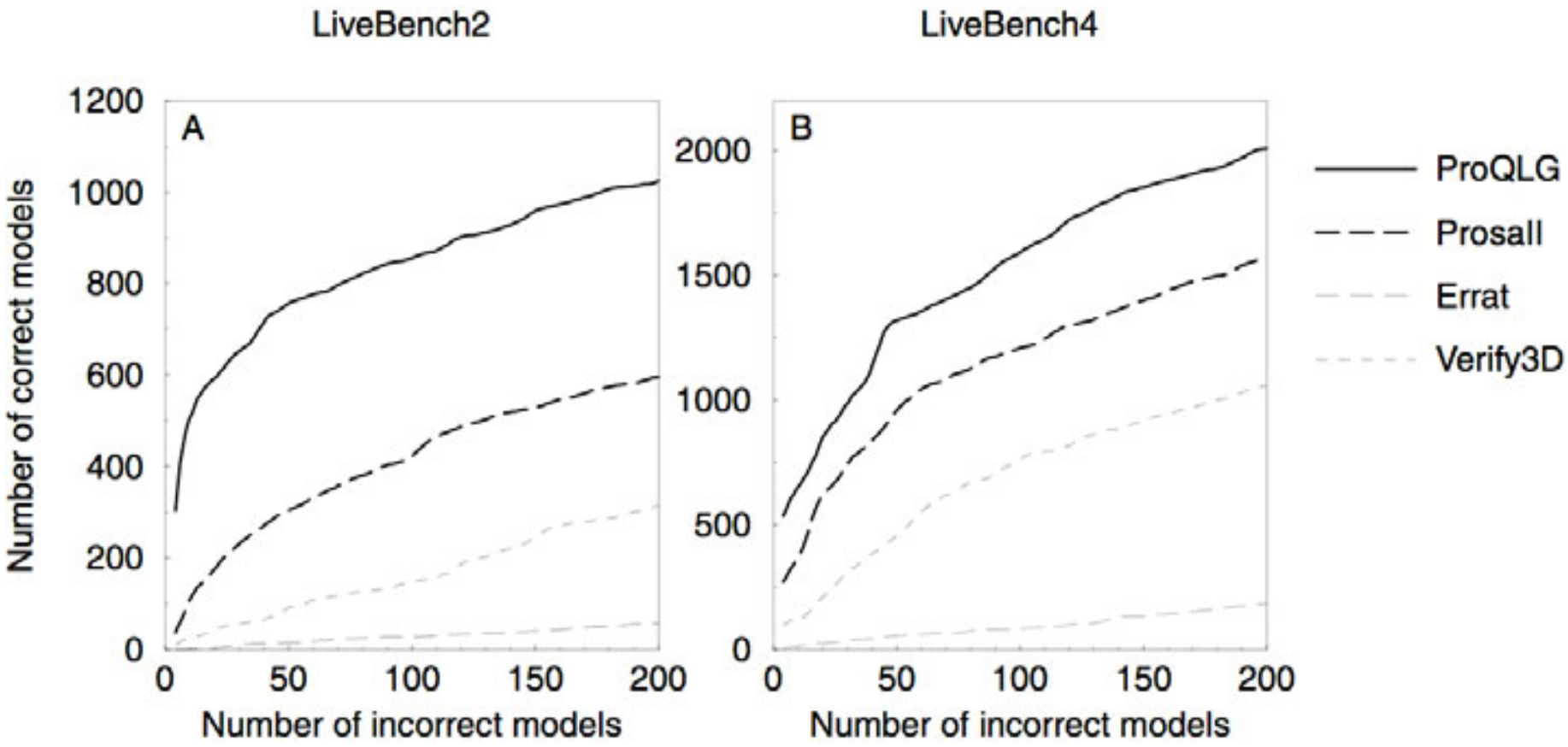


ProQ (Wallner & Elofsson, 2001)

- Neural networks
- Predicts quality of models
- Full atom models
- Many parameters
  - Atom-Atom contacts
  - Residue- Residue contacts
  - Surface area
  - Secondary structure info
- Predicts "quality" with
  - R = 0.76
  - Finding correct models
- Best in CAFASP

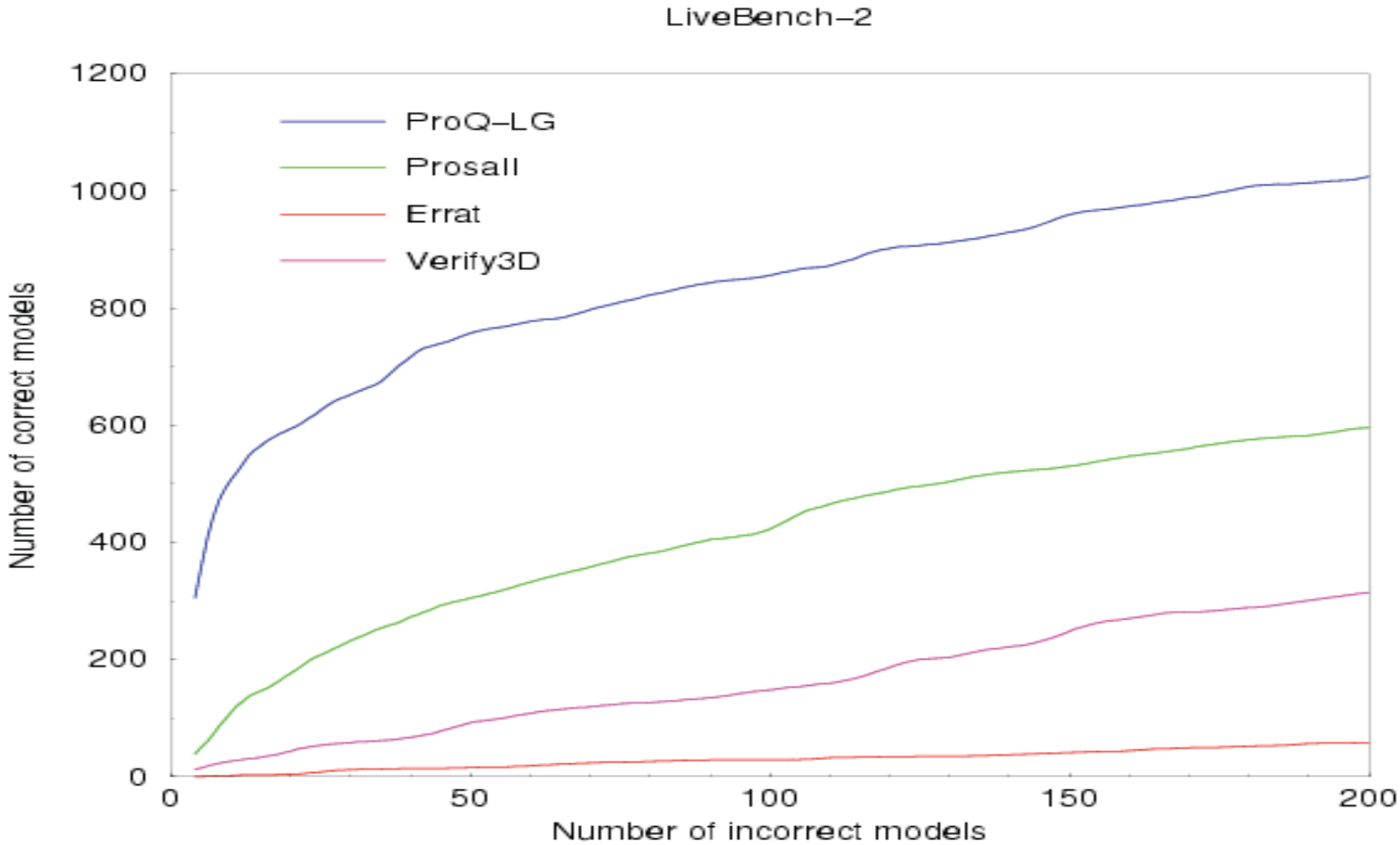


ProQ compared to other



(Wallner and Elofsson, *Protein Sci.* 2001)

ProQ specificity



ProQ better at finding correct models

$$Z \equiv \frac{\langle score_{correct} \rangle - \langle score_{incorrect} \rangle}{\sigma_{incorrect}}$$

$$Z_{nat} \equiv \frac{1}{n} \sum_{i=1}^{n} \frac{score^{i}_{native} - \langle score^{i}_{all} \rangle}{\sigma^{i}_{all}}$$

Method	LB-2 Z/Z_{nat}	LB-4 Z/Z_{nat}	4state_reduced Z/Z_{nat}	LMDS Z/Z_{nat}	Lattice_ssfit Z/Z_{nat}	Structal Z/Z_{nat}
ProQ-LG	2.7/5.2	2.7/5.1	2.3/4.4	–0.4/3.9	—/11.7	—/2.4
ProQ-MX	2.6/5.0	2.8/4.6	2.0/3.5	0.0/1.8	—/11.6	—/1.6
Errat	0.3/5.0	0.3/5.7	1.7/2.5	0.2/3.1	—/5.1	—/3.6
Prosa II	1.2/3.5	1.6/3.5	2.0/2.7	0.4/2.5	—/5.6	—/1.7
Verify3D	1.0/2.8	1.1/2.8	1.0/2.6	0.8/1.4	—/4.5	—/1.4

The lattice_ssfit and structal methods contained too few correct models to calculate Z-score.

(Wallner and Elofsson, *Protein Sci.* 2001)