

Comparative Modeling

Methods and Applications

Marc A. Martí-Renom, András Fiser & Andrej Šali

*Laboratories of Molecular Biophysics
Pels Family Center for Biochemistry and Structural Biology
The Rockefeller University*

Summary

- ✓ What is comparative modeling and why is it useful?
- ✓ Steps in CM (overview + some details)
- ✓ Accuracy of comparative models
- ✓ Target-Template alignment
- ✓ Loop modeling
- ✓ CM and Structural Genomics

Summary

- ✓ What is comparative modeling and why is it useful?
- ✓ Steps in CM (overview + some details)
- ✓ Accuracy of comparative models
- ✓ Target-Template alignment
- ✓ Loop modeling
- ✓ CM and Structural Genomics

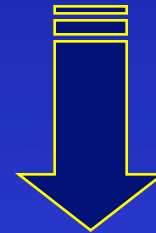
Why protein structure prediction?

	Y 2001	Y 2005
Sequences	700,000	millions
Structures	16,000	50,000

Why protein structure prediction?

	Y 2001
Sequences	700,000
Structures	15,000

Theory

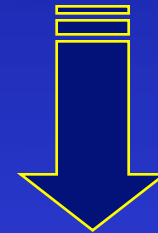


Experiment

Why protein structure prediction?

	Y 2001
Sequences	700,000
Structures	300,000

Theory



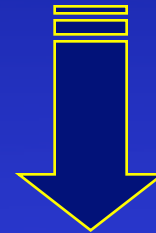
Experiment

<http://pipe.rockefeller.edu/modbase/>

Why protein structure prediction?

	Y 2001
Sequences	700,000
Structures	300,000

Theory



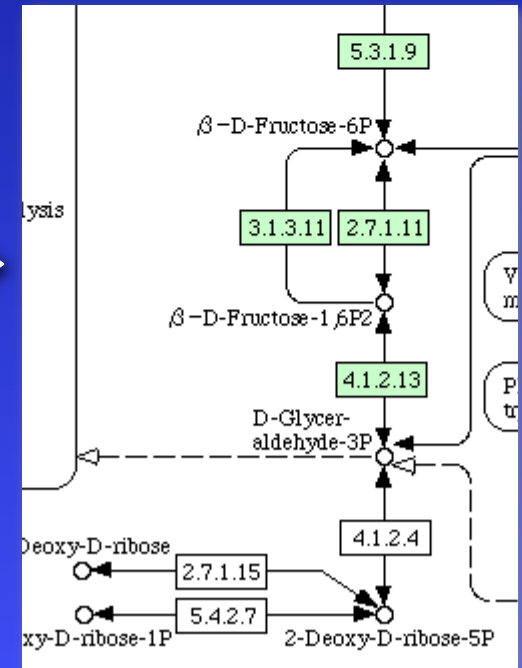
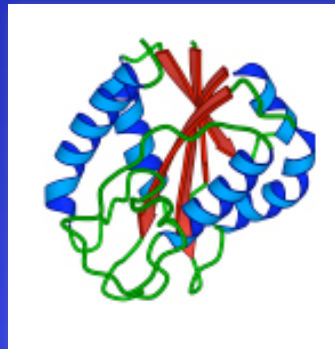
Experiment

<http://pipe.rockefeller.edu/modbase/>

Function *via* Structure

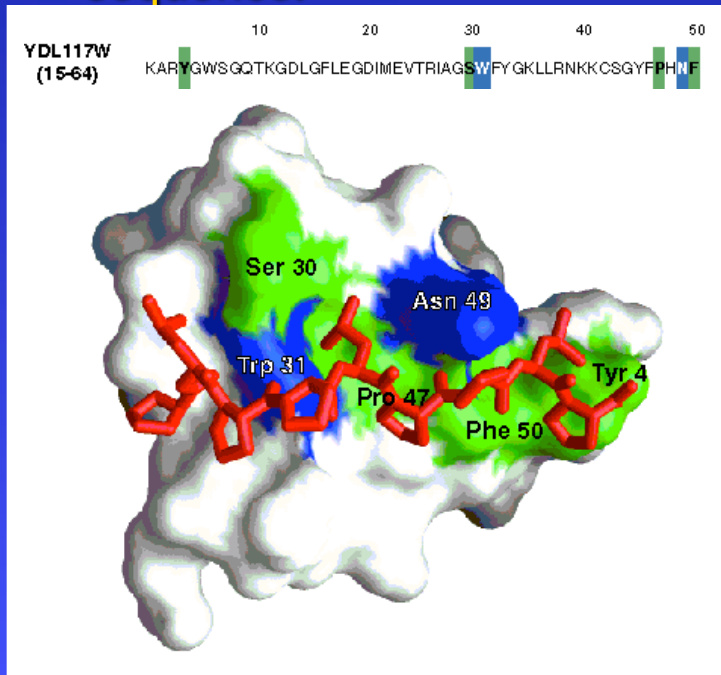
Sequence → Structure → Function

GFCHIKAYTRLIM...



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

Principles of Protein Structure

GFCHIKAYTRLIMVG...



Folding

Ab initio prediction

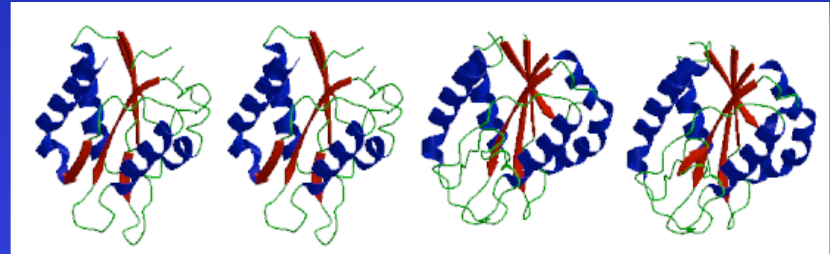
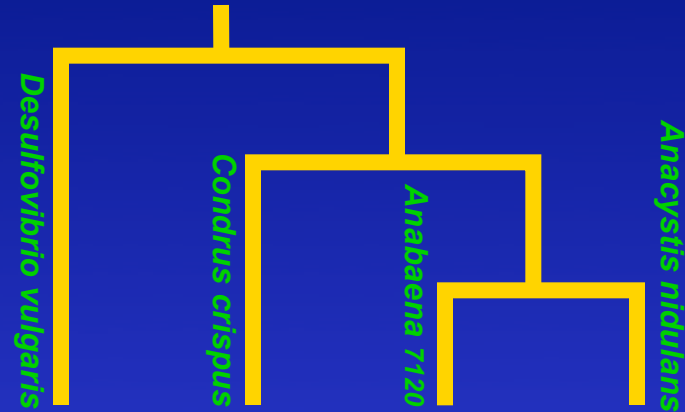
Principles of Protein Structure

GFCHIKAYTRLIMVG...



Folding

Ab initio prediction



Evolution

Threading
Comparative Modeling

Summary

- ✓ What is comparative modeling and why is it useful?
- ✓ Steps in CM (overview + some details)
- ✓ Accuracy of comparative models
- ✓ Target-Template alignment
- ✓ Loop modeling
- ✓ CM and Structural Genomics

Steps in Comparative Protein Structure Modeling

START

TARGET

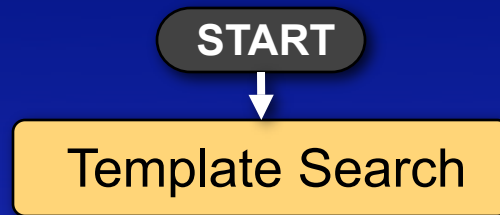
ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCLKIDD
VPERLIPERASFQWMNDK

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

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

M. A. Martí-Renom et al. *Ann. Rev. Biophys. Biomolec. Struct.*, 29, 291, 2000.

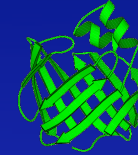
Steps in Comparative Protein Structure Modeling



TARGET

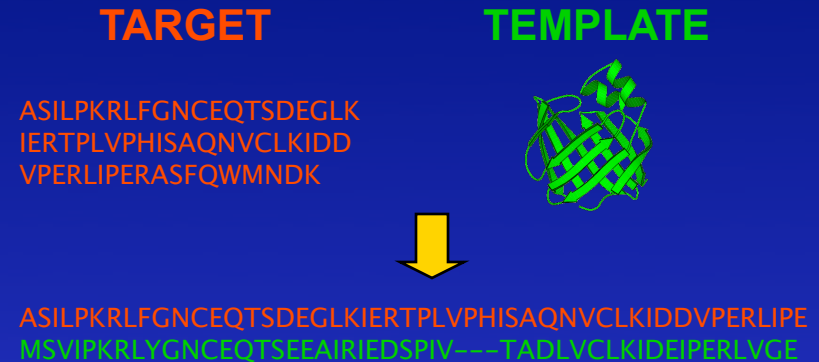
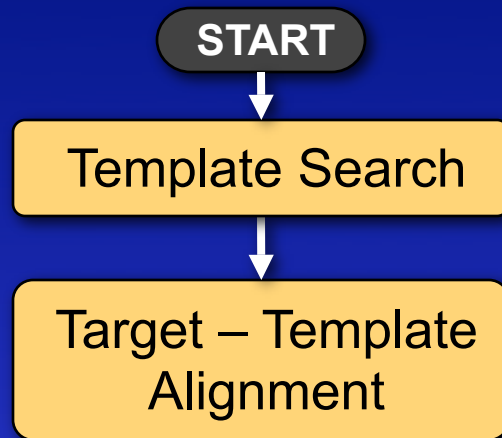
ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCLKIDD
VPERLIPERASFQWMNDK

TEMPLATE



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

Steps in Comparative Protein Structure Modeling

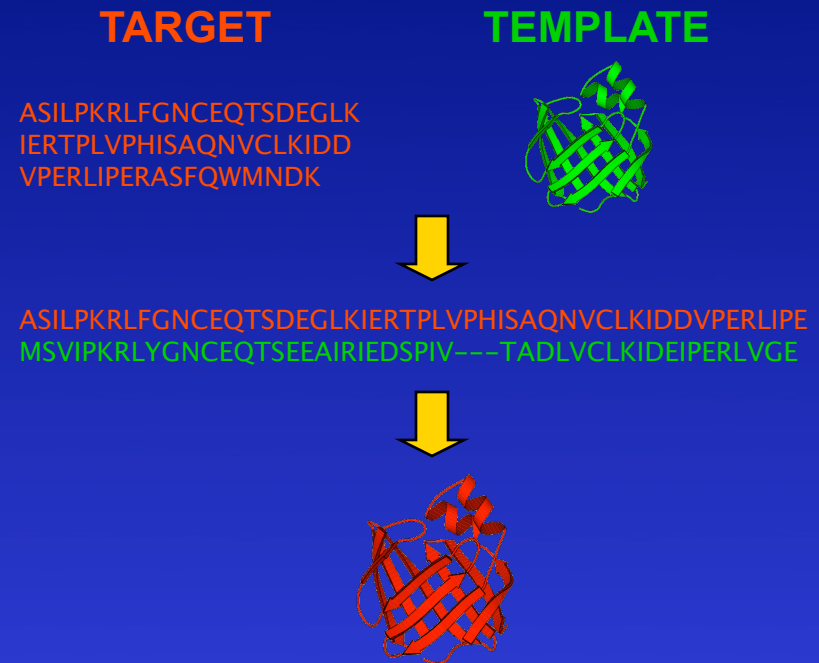
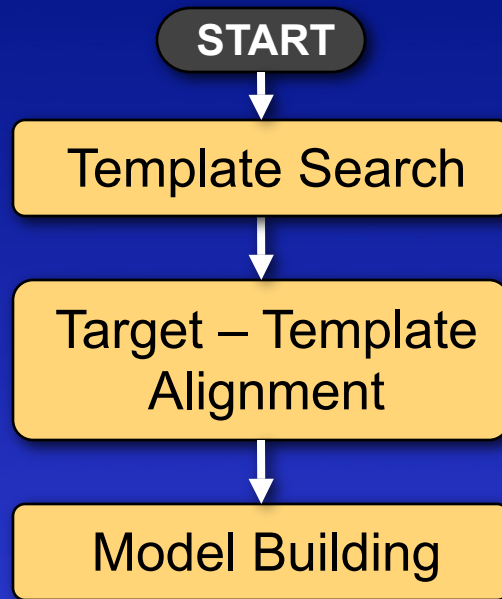


A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.

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

M. A. Martí-Renom et al. *Ann. Rev. Biophys. Biomolec. Struct.*, 29, 291, 2000.

Steps in Comparative Protein Structure Modeling

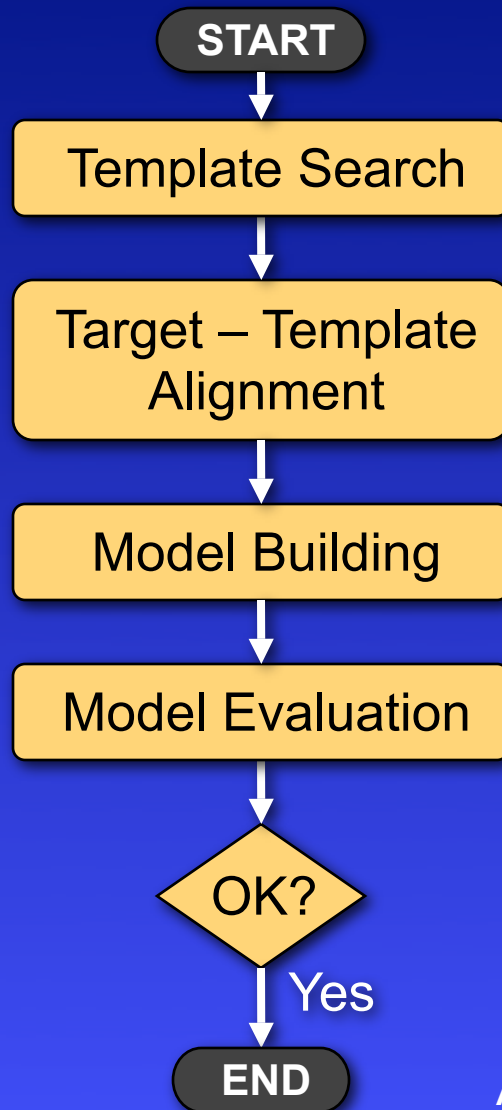


A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.

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

M. A. Martí-Renom et al. *Ann. Rev. Biophys. Biomolec. Struct.*, 29, 291, 2000.

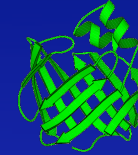
Steps in Comparative Protein Structure Modeling



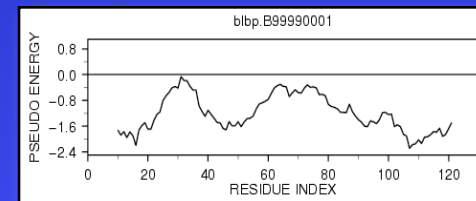
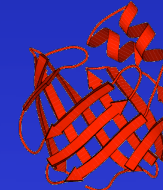
TARGET

ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCLKIDD
VPERLIPERASFQWMNDK

TEMPLATE



ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE

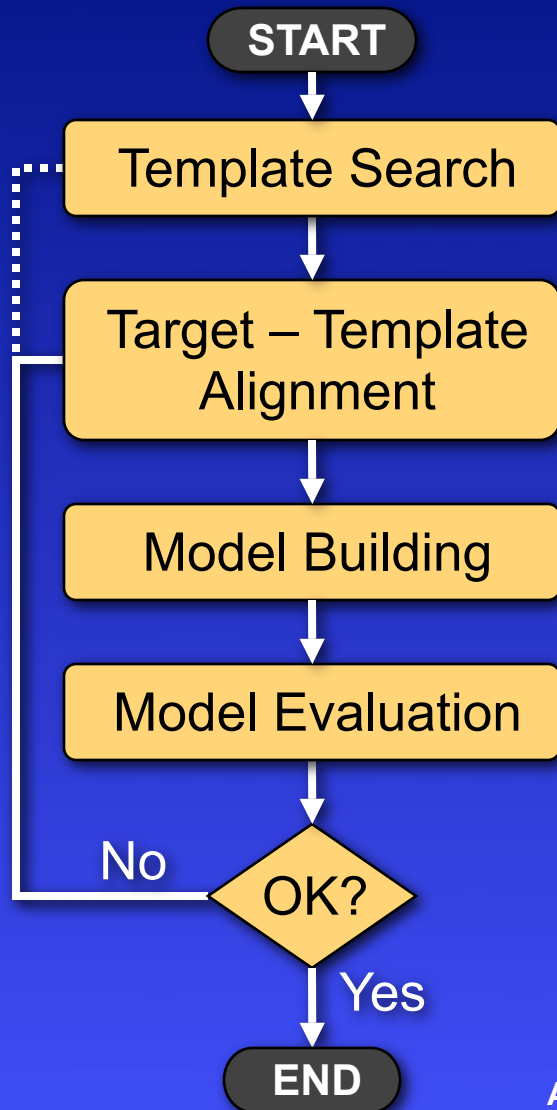


A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.

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

M. A. Martí-Renom et al. *Ann. Rev. Biophys. Biomolec. Struct.*, 29, 291, 2000.

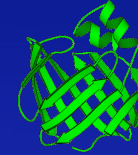
Steps in Comparative Protein Structure Modeling



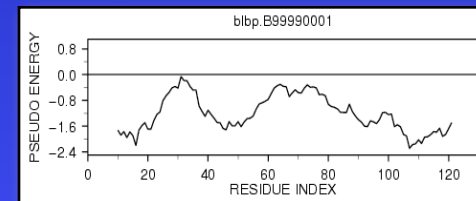
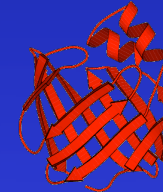
TARGET

ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCLKIDD
VPERLIPERASFQWMNDK

TEMPLATE



ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE

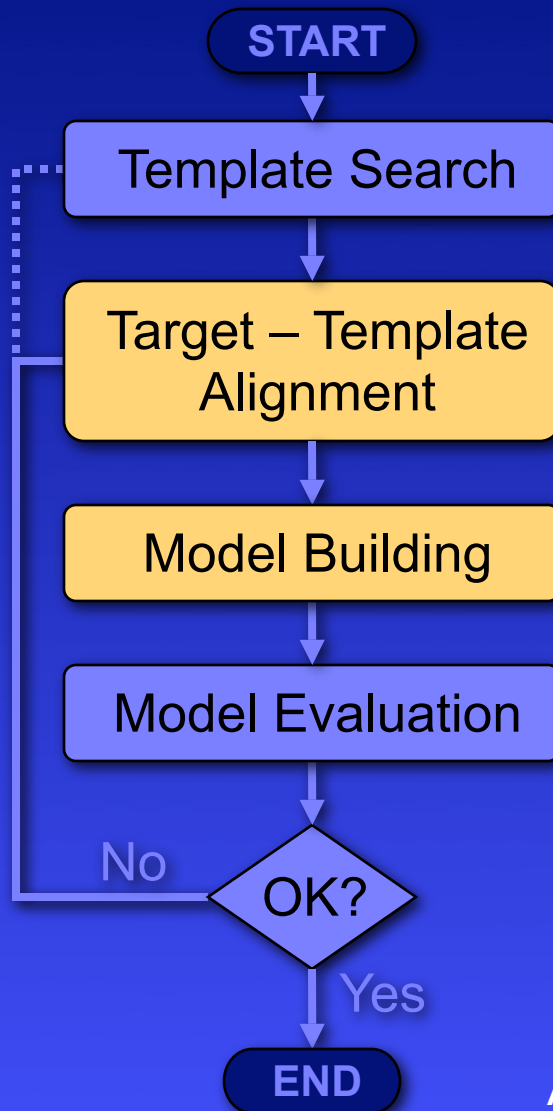


A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.

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

M. A. Martí-Renom et al. *Ann. Rev. Biophys. Biomolec. Struct.*, 29, 291, 2000.

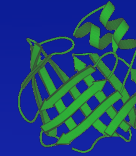
Steps in Comparative Protein Structure Modeling



TARGET

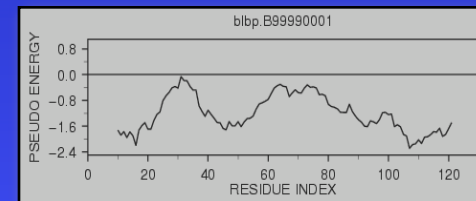
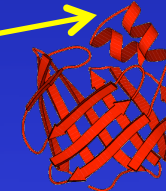
ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCLKIDD
VPERLIPERASFQWMNDK

TEMPLATE



ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE

Loop modeling



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

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

M. A. Martí-Renom et al. *Ann. Rev. Biophys. Biomolec. Struct.*, 29, 291, 2000.

Template Search Methods

✓ Sequence similarity searches

- ✓ BLAST [<http://www.ncbi.nlm.nih.gov/>]
- ✓ FastA program [<http://www.ebi.ac.uk/fasta33/>]

✓ Profile and iterative methods

- ✓ HMMs [<http://www.cse.ucsc.edu/research/compbio/HMM-apps/>]
- ✓ PSI-BLAST [<http://www.ncbi.nlm.nih.gov/>]

✓ Structure based threading

- ✓ THREADER [<http://bioinf.cs.ucl.ac.uk/>]
- ✓ PROFIT [<http://www.came.sbg.ac.at/>]

Target – Template Alignment Methods

- ✓ Dynamic Programming Pairwise Alignment
 - ✓ ALIGN [<http://guitar.rockefeller.edu/modeller/>]
- ✓ Multiple Alignments,
 - ✓ Psi-Blast [<http://www.ncbi.nlm.nih.gov/>]
 - ✓ HMM [<http://www.cse.ucsc.edu/research/compbio/HMM-apps/>]
 - ✓ ALIGN4D [<http://guitar.rockefeller.edu/modeller/>]
- ✓ Structure based approaches
 - ✓ Threading [<http://bioinf.cs.ucl.ac.uk/>]

Model Building Methods

- ✓ Rigid Body Assembly

- ✓ COMPOSER [<http://www-cryst.bioc.cam.ac.uk/>]

- ✓ Segment Matching

- ✓ SEGMOD

- ✓ Satisfaction of Spatial Restraints

- ✓ MODELLER [<http://guitar.rockefeller.edu/modeller/>]

Comparative Modeling by Satisfaction of Spatial Restraints (MODELLER)

3D GKITYFIERGFQGHCIYESDC-NLQP...
SEQ GKITYFIERG---RCYESDCPNLQP...

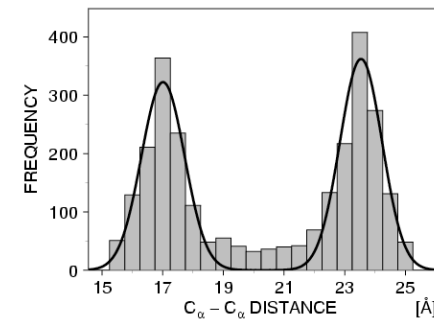
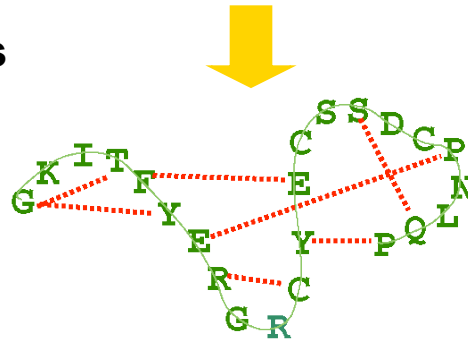
A. Šali & T. Blundell. *J. Mol. Biol.* **234**, 779, 1993.
J.P. Overington & A. Šali. *Prot. Sci.* **3**, 1582, 1994.
A. Fiser, R. Do & A. Šali, *Prot. Sci.*, in press.

<http://guitar.rockefeller.edu/>

Comparative Modeling by Satisfaction of Spatial Restraints (MODELLER)

3D GKITYERGFQGHCEYSDC-NLQP...
SEQ GKITYERG---RCYESDCPNLQP...

1. Extract spatial restraints



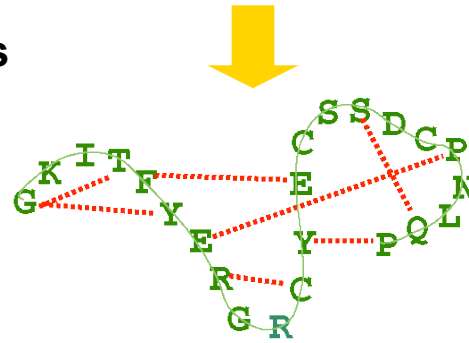
A. Šali & T. Blundell. *J. Mol. Biol.* **234**, 779, 1993.
J.P. Overington & A. Šali. *Prot. Sci.* **3**, 1582, 1994.
A. Fiser, R. Do & A. Šali, *Prot. Sci.*, in press.

<http://guitar.rockefeller.edu/>

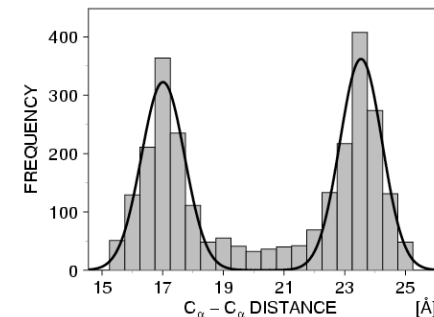
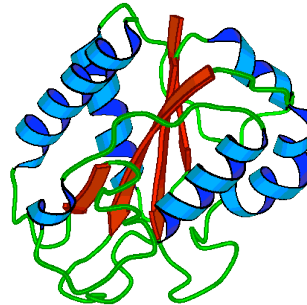
Comparative Modeling by Satisfaction of Spatial Restraints (MODELLER)

3D GKITFYERGFQGH CYESDC-NLQP...
SEQ GKITFYERG---RCYESDCPNLQP...

1. Extract spatial restraints



2. Satisfy spatial restraints



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

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.*, in press.

<http://guitar.rockefeller.edu/>

Model Evaluation methods

- ✓ Stereochemistry

- ✓ PROCHECK [<http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html>]

- ✓ Environment

- ✓ VERIFY3D [http://www.doe-mbi.ucla.edu/Services/Verify_3D/]

- ✓ Statistical potentials based methods

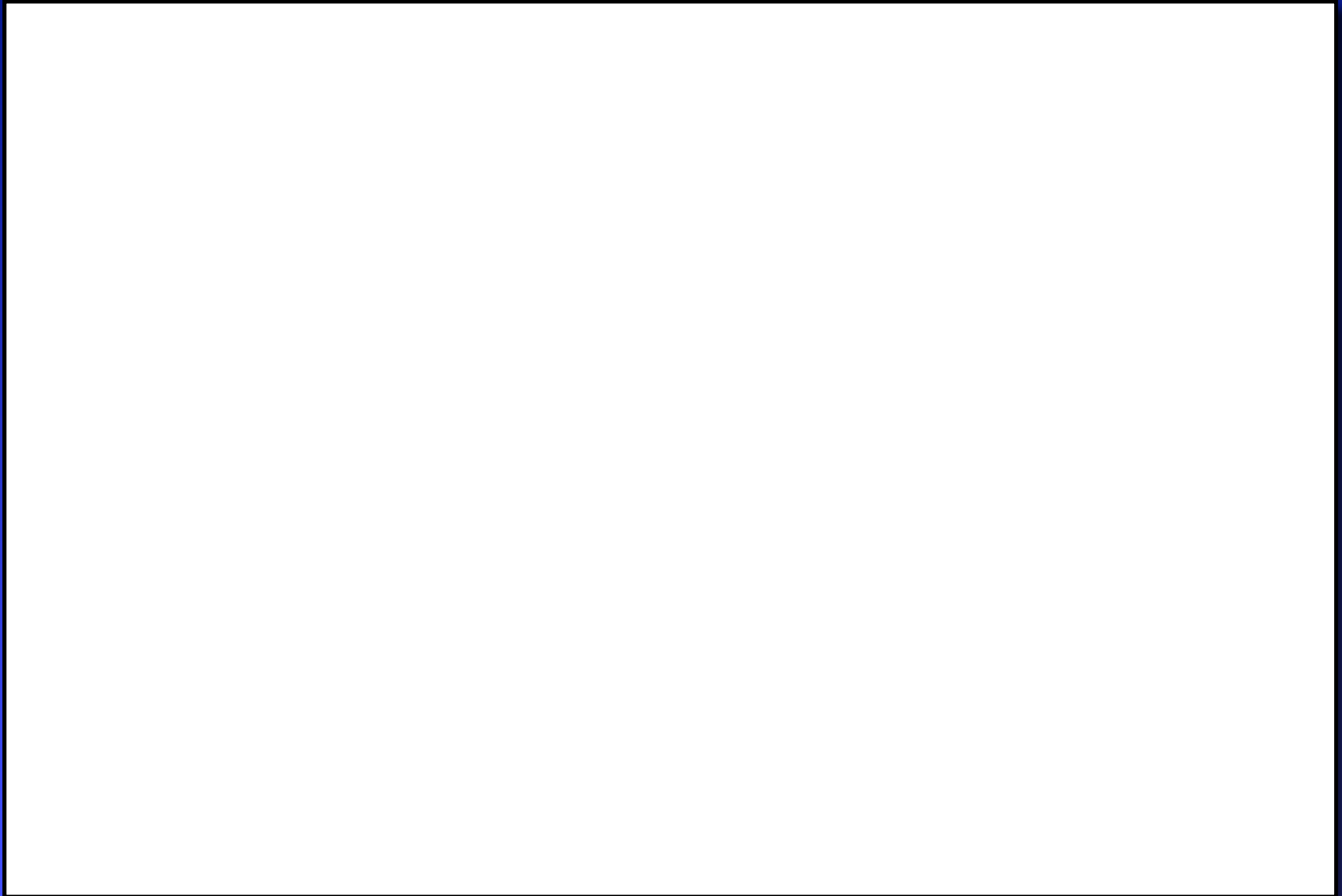
- ✓ PROSAIL [<http://www.came.sbg.ac.at/>]

http://guitar.rockefeller.edu/bioinformatics_resources.shtml

Summary

- ✓ What is comparative modeling and why is it useful?
- ✓ Steps in CM (overview + some details)
- ✓ **Accuracy of comparative models**
- ✓ Target-Template alignment
- ✓ Loop modeling
- ✓ CM and Structural Genomics

Typical Errors in Comparative Models



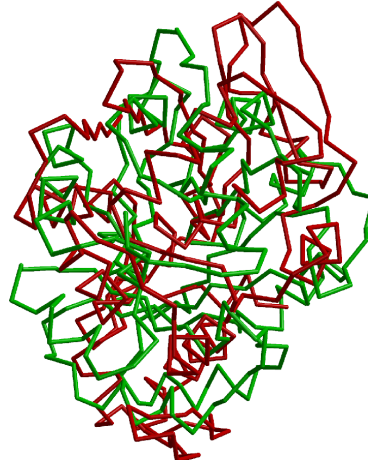
Typical Errors in Comparative Models

MODEL

X-RAY

TEMPLATE

Incorrect template



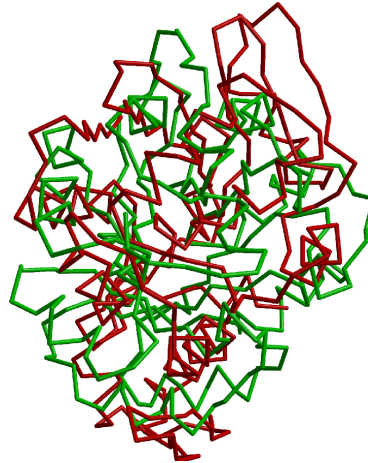
Typical Errors in Comparative Models

MODEL

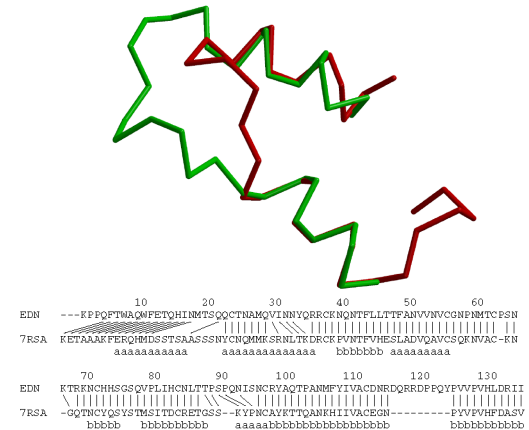
X-RAY

TEMPLATE

Incorrect template



Misalignment



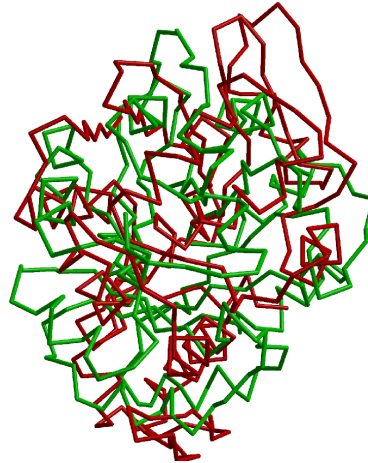
Typical Errors in Comparative Models

MODEL

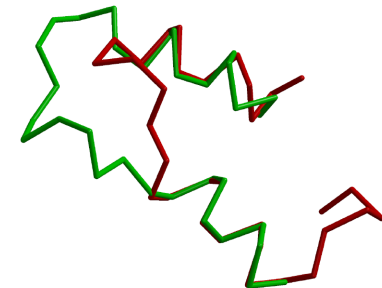
X-RAY

TEMPLATE

Incorrect template



Misalignment

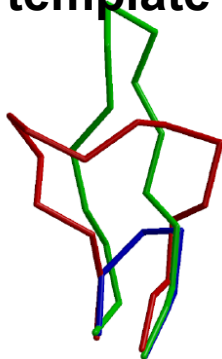


```
EDN  ---FFPQFTWAGMFETQHINMTSCQCTHAGVIRNNVQRCENQHTFLLTTPAVVUVCNFAHTCSN
7RSA KETAAAFERQHMDSSTAASSNNYCNQMMESRLTEDECKPVNTPVHESLADVQVCSQENFVAC-KN
      aaaaaaaaaa aaaaaaaaaa bbbbbbb aaaaaaaaaa

EDN  FTRENCHSGSQVPLIHCLNLTSPQNISNCRYAQTFANMFYIVACDNRDQREDPEQVFPVPHLDRII
7RSA -GQTNCYQSYSYMSITDCRETGSS--FYFNCAYETTDANEHIIVACEGN-----FYFVPHFDASV
      bbbb  bbbbbbbbbb aaaaabbbbbbbbbbbbbbbb bbbbbbbbbb
```

Region without a

template



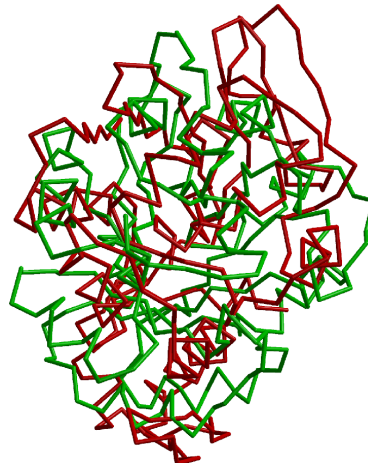
Typical Errors in Comparative Models

MODEL

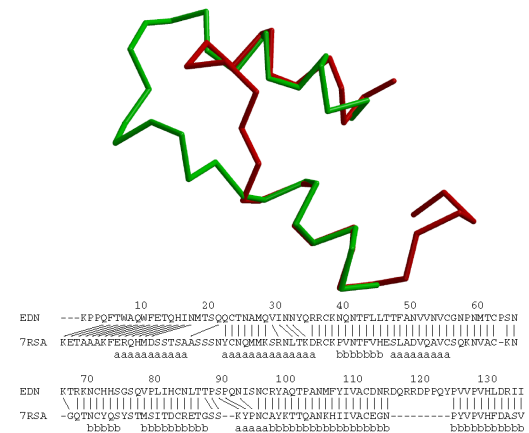
X-RAY

TEMPLATE

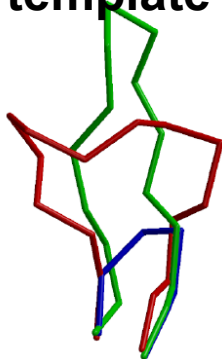
Incorrect template



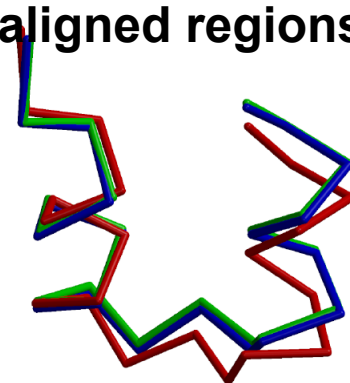
Misalignment



Region without a
template



Distortion in correctly
aligned regions



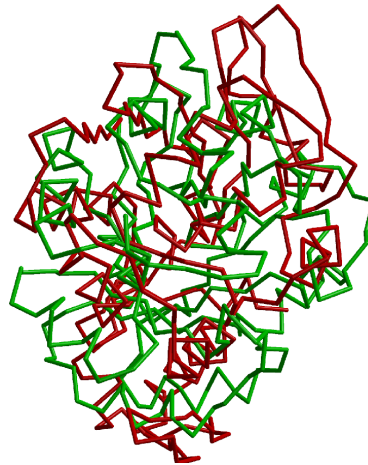
Typical Errors in Comparative Models

MODEL

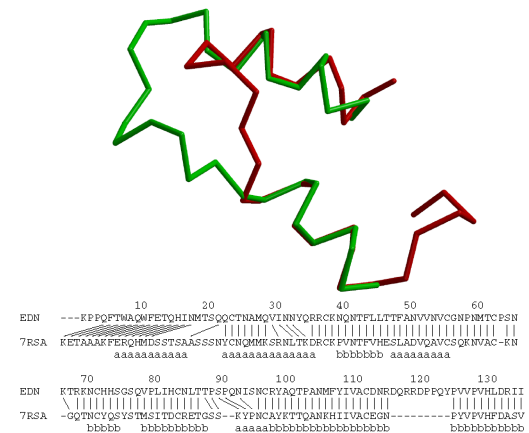
X-RAY

TEMPLATE

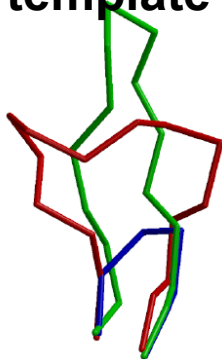
Incorrect template



Misalignment



Region without a template



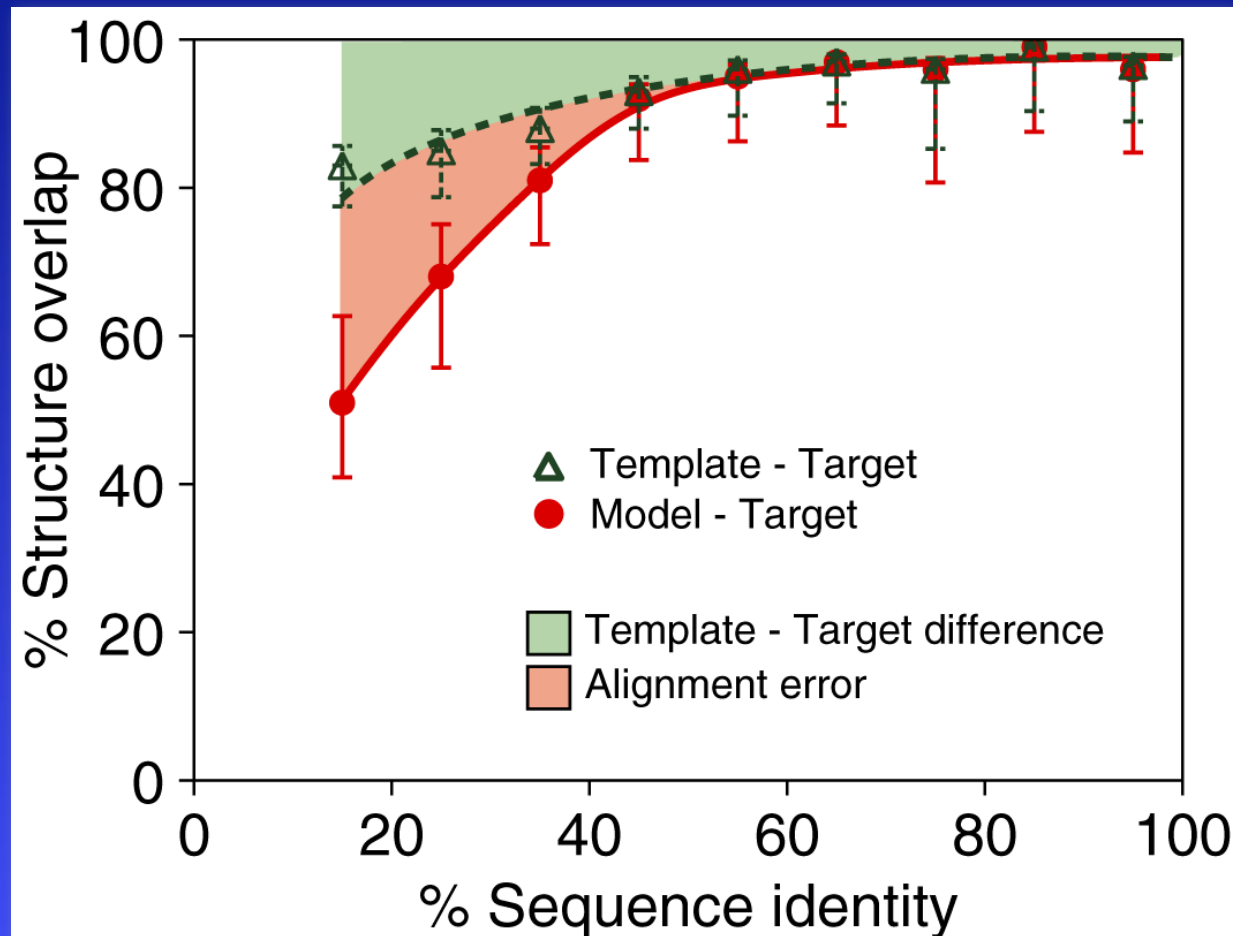
Distortion in correctly aligned regions



Sidechain packing



Model Accuracy as a Function of Target-Template Sequence Identity

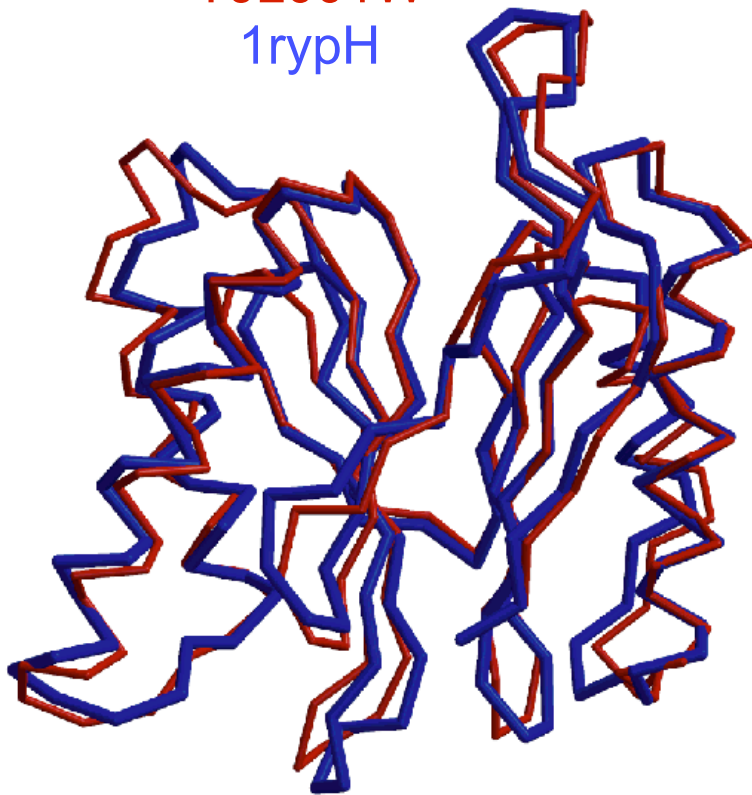


Some Models Can Be Surprisingly Accurate
(in Some Core or Active Site Regions)

Some Models Can Be Surprisingly Accurate (in Some Core or Active Site Regions)

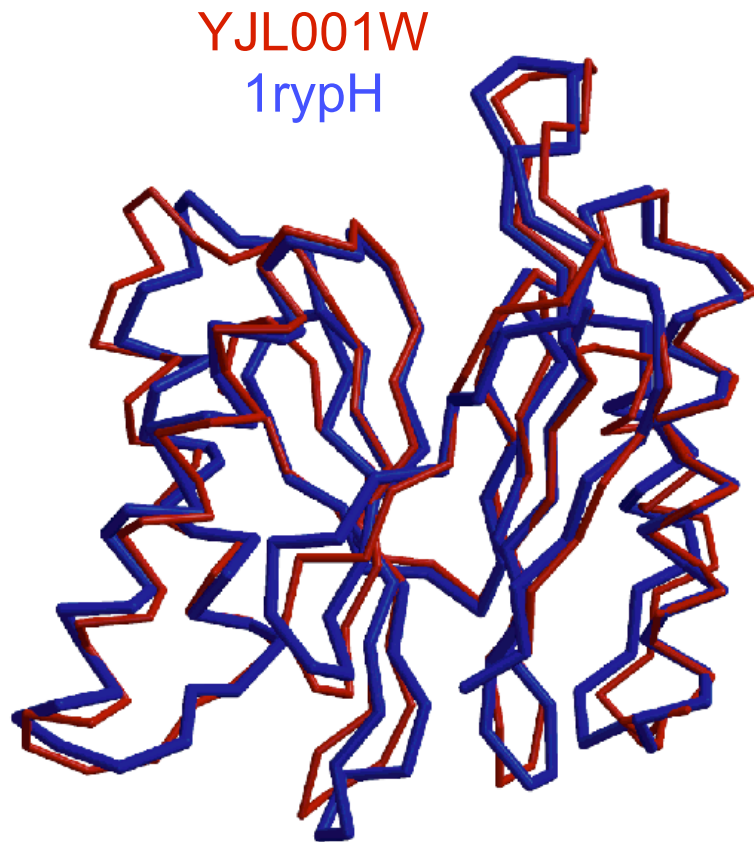
24% sequence identity

YJL001W
1rypH

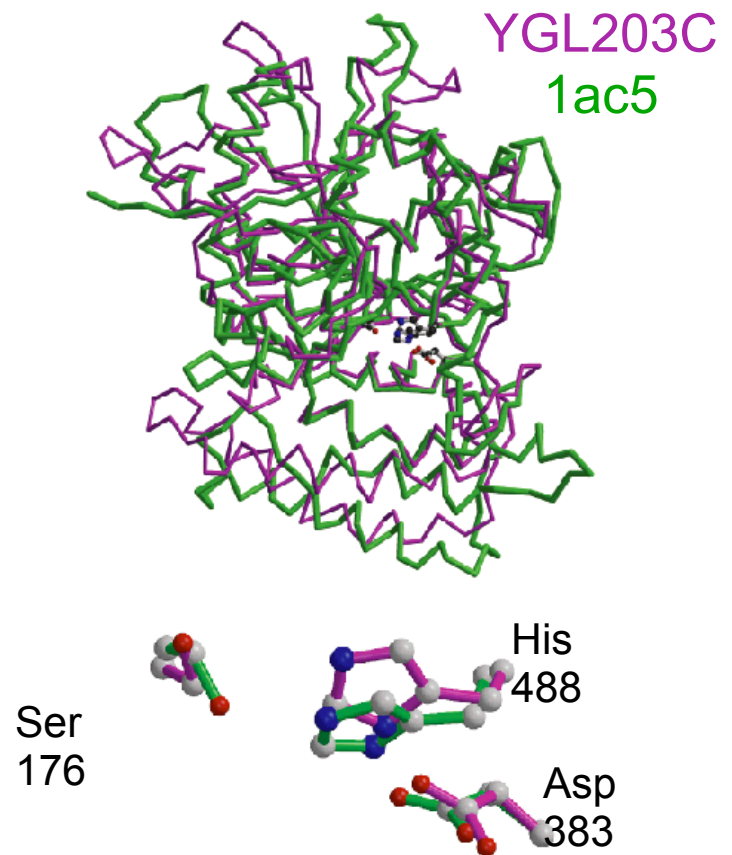


Some Models Can Be Surprisingly Accurate (in Some Core or Active Site Regions)

24% sequence identity



25% sequence identity



Do mast cell proteases bind proteoglycans? Where? When?

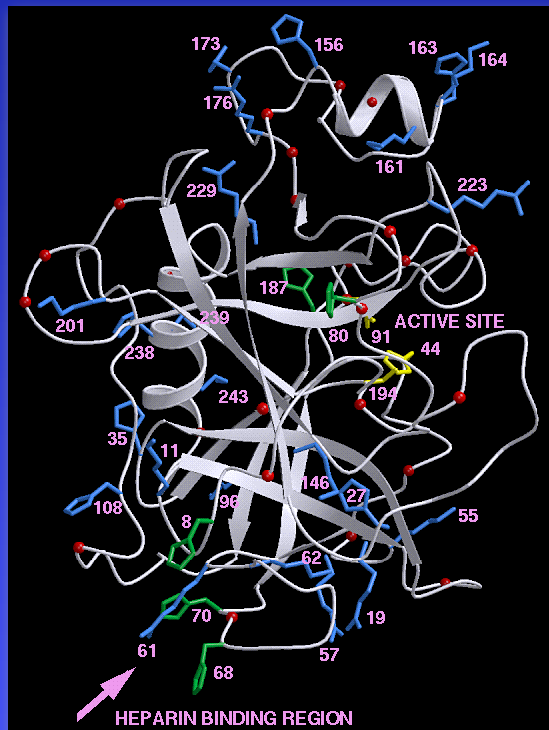
Predicting features of a model that are not present in the template

1. mMCPs bind negatively charged proteoglycans through electrostatic interactions?
2. Comparative models used to find clusters of positively charged surface residues.
3. Tested by site-directed mutagenesis.

Do mast cell proteases bind proteoglycans? Where? When?

Predicting features of a model that are not present in the template

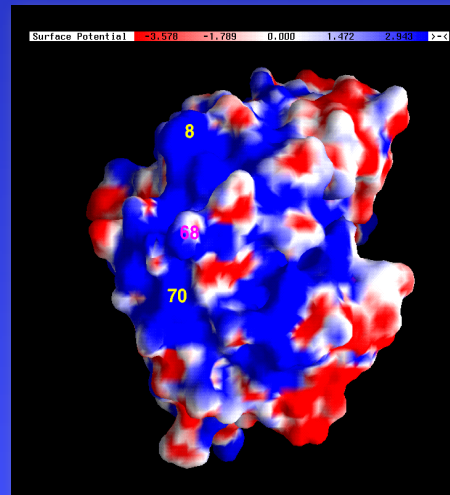
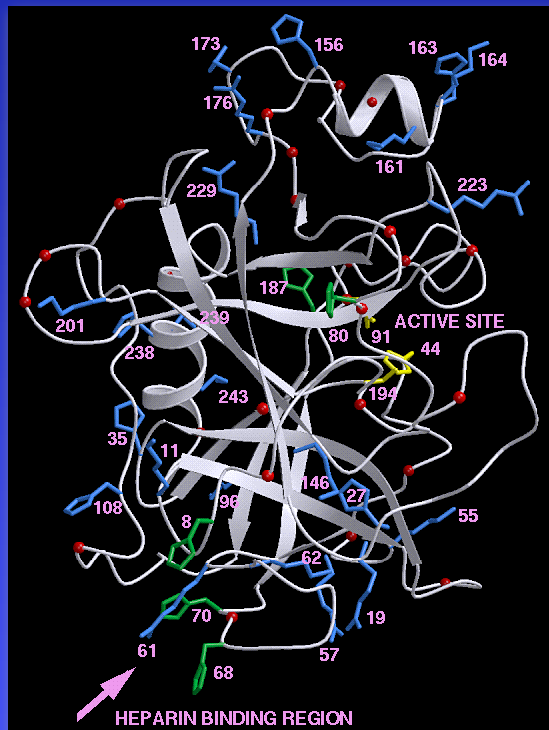
1. mMCPs bind negatively charged proteoglycans through electrostatic interactions?
2. Comparative models used to find clusters of positively charged surface residues.
3. Tested by site-directed mutagenesis.



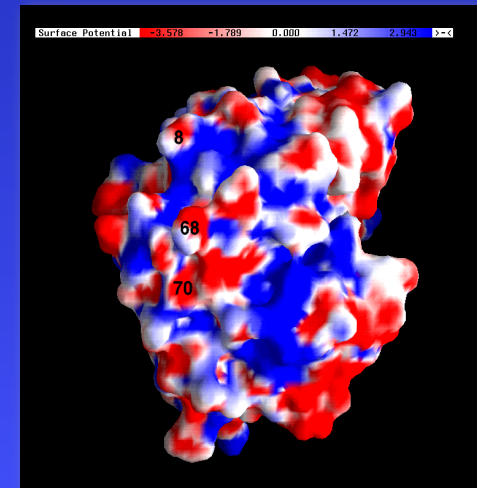
Do mast cell proteases bind proteoglycans? Where? When?

Predicting features of a model that are not present in the template

1. mMCPs bind negatively charged proteoglycans through electrostatic interactions?
2. Comparative models used to find clusters of positively charged surface residues.
3. Tested by site-directed mutagenesis.



Native mMCP-7 at pH=5 (His⁺)

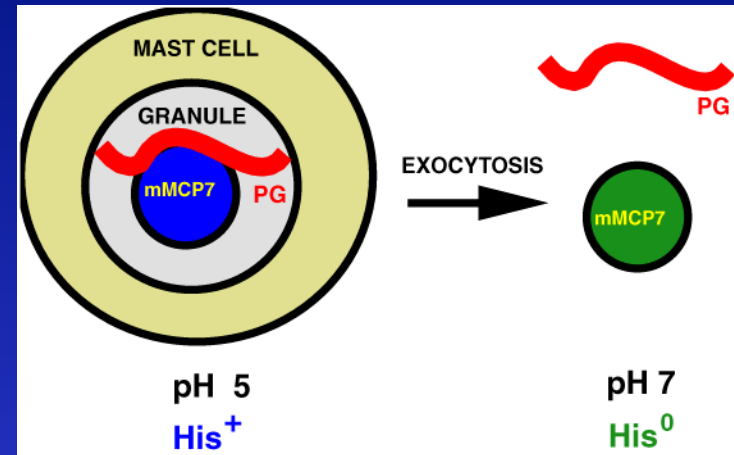


Native mMCP-7 at pH=7 (His⁰)

Do mast cell proteases bind proteoglycans? Where? When?

Predicting features of a model that are not present in the template

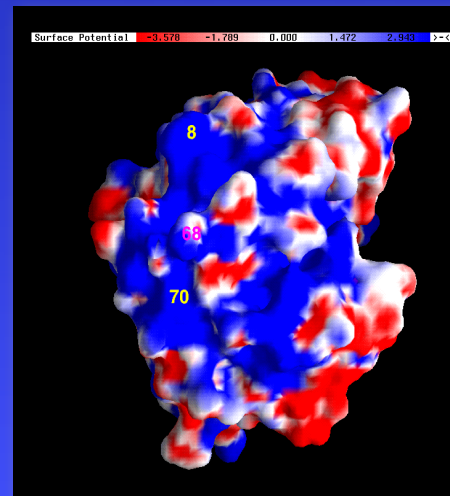
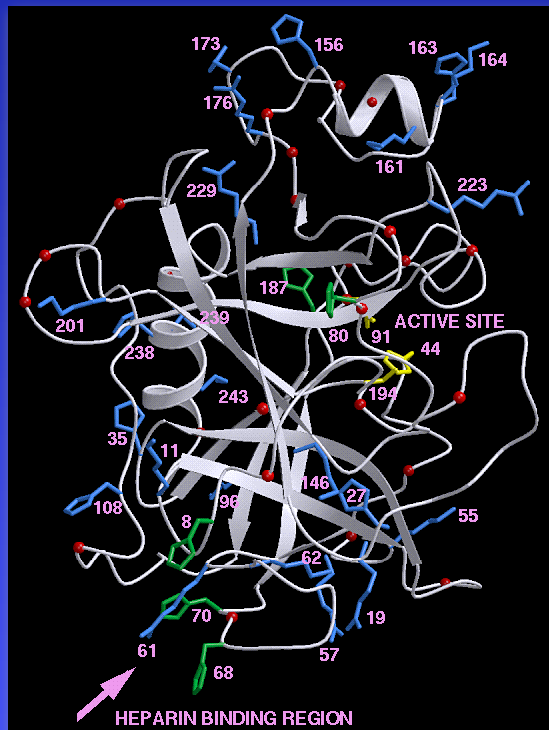
1. mMCPs bind negatively charged proteoglycans through electrostatic interactions?
2. Comparative models used to find clusters of positively charged surface residues.
3. Tested by site-directed mutagenesis.



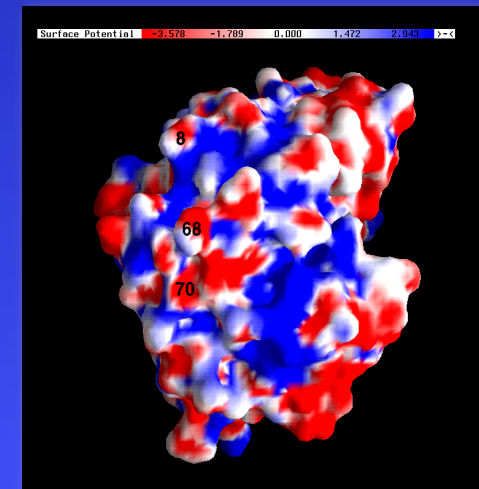
Huang et al. *J. Clin. Immunol.* **18**,169,1998.

Matsumoto et al. *J. Biol. Chem.* **270**,19524,1995.

Šali et al. *J. Biol. Chem.* **268**, 9023, 1993.



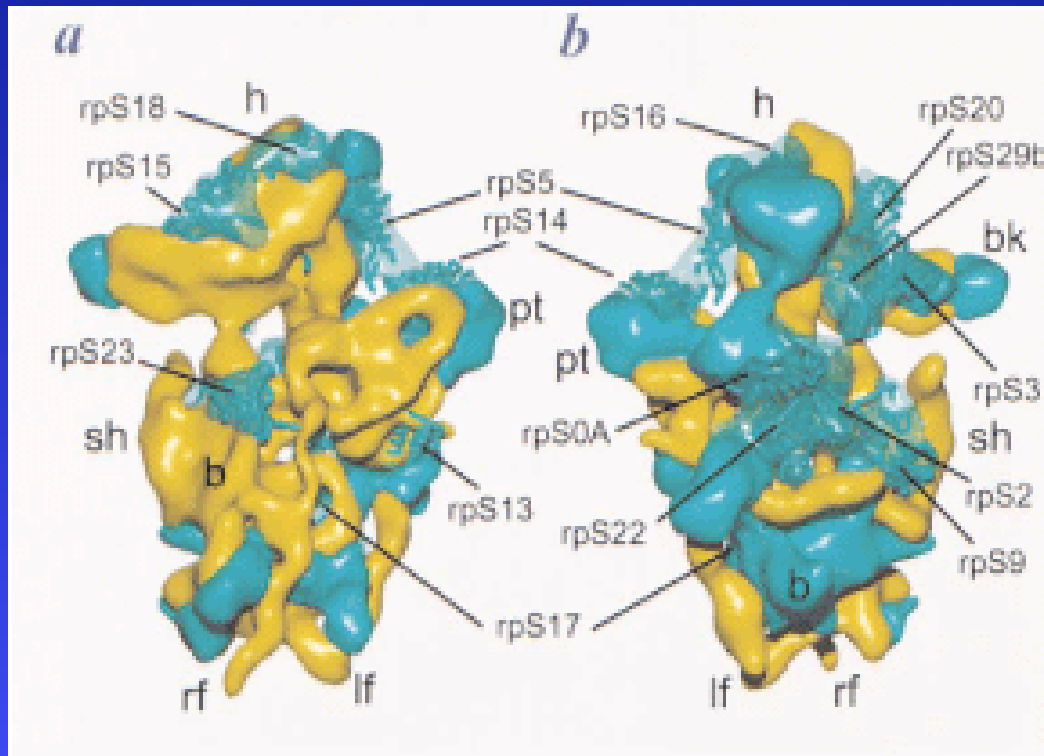
Native mMCP-7 at pH=5 (His⁺)



Native mMCP-7 at pH=7 (His⁰)

Some Models Can Be Used in Docking to Density Maps

(Yeast Ribosomal 40S subunit)

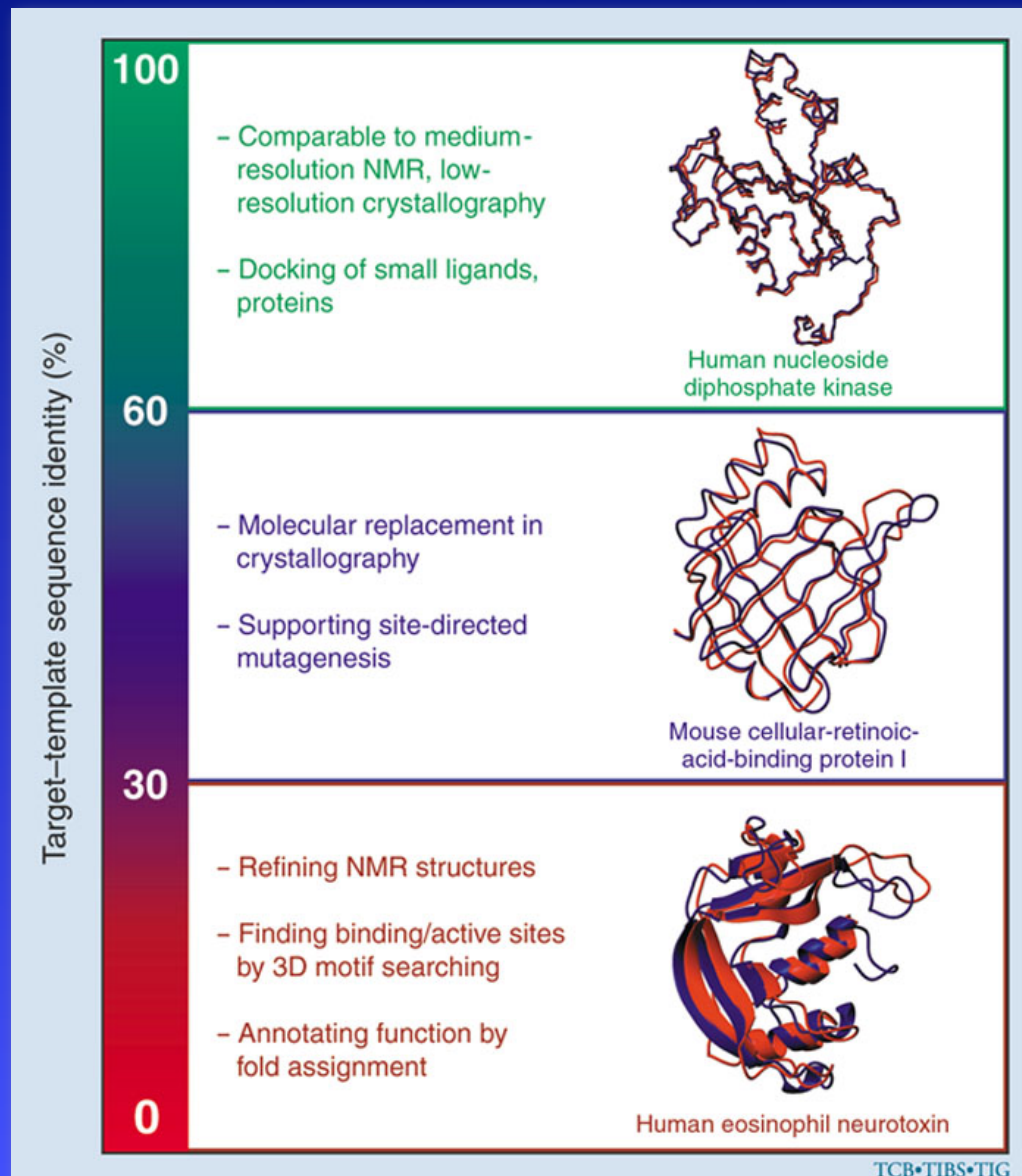


Docking of comparative models into the cryo-EM map.

Spahn *et al.* 2001 Cell **107**:373-386

Small 30S subunit from *Thermus thermophilus*
Large 50S subunit from *Haloarcula marismortui*

Applications of Comparative Models



Summary

- ✓ What is comparative modeling and why is it useful?
- ✓ Steps in CM (overview + some details)
- ✓ Accuracy of comparative models
- ✓ **Target-Template alignment**
- ✓ Loop modeling
- ✓ CM and Structural Genomics

Experiment (*in silico*)

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

Methods: Reference set

CE alignments with

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

387

Filter: MAMMOTH alignments with

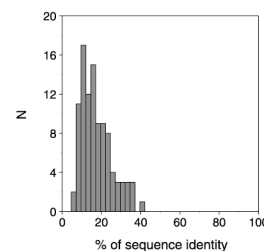
- > 50% EqPos

300

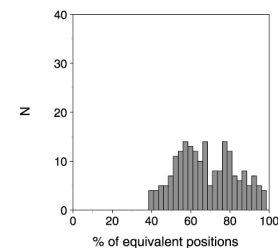
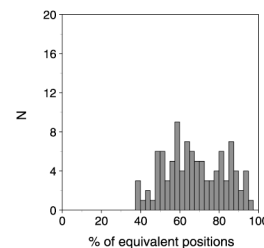
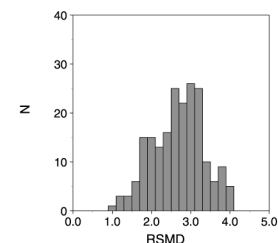
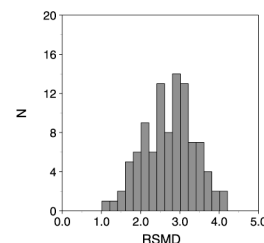
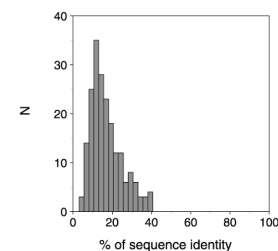
100 Training set

200 Testing set

A) Training Set



B) Testing Set



Methods: Evaluated methods

Sequence A: AGHLAHTRC~~EL~~KLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

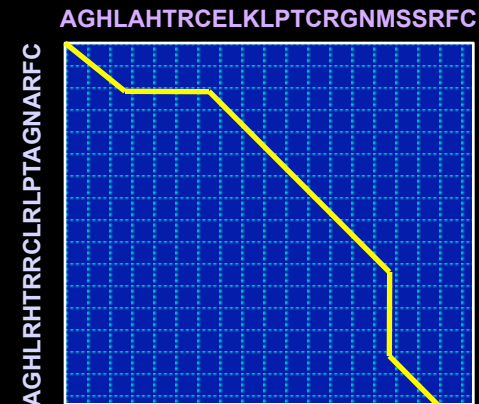
BLAST2SEQ: Local method

Prof.-Seq.

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

Prof.-Prof.

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



Methods: Evaluated methods

Sequence A: AGHLAHTRCCLKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

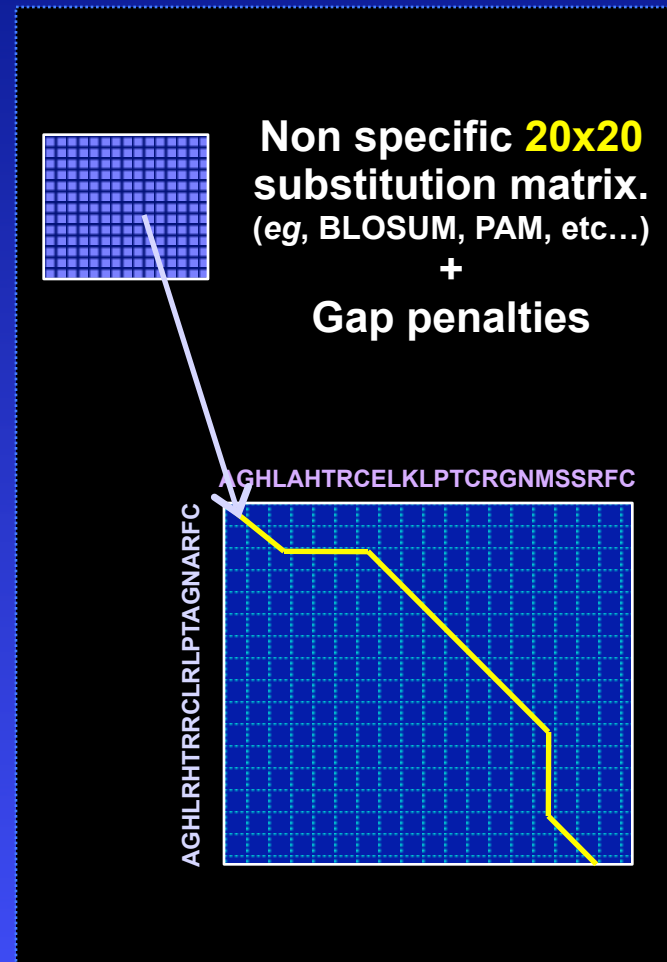
BLAST2SEQ: Local method

Prof.-Seq.

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

Prof.-Prof.

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



Methods: Evaluated methods

Sequence A: AGHLAHTRCCLKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

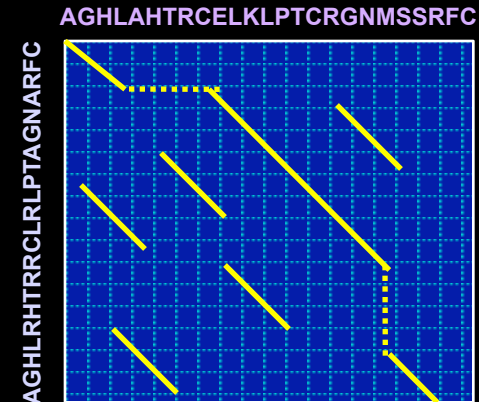
BLAST2SEQ: Local method

Prof.-Seq.

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

Prof.-Prof.

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



Methods: Evaluated methods

Sequence A: AGHLAHTRCCLKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

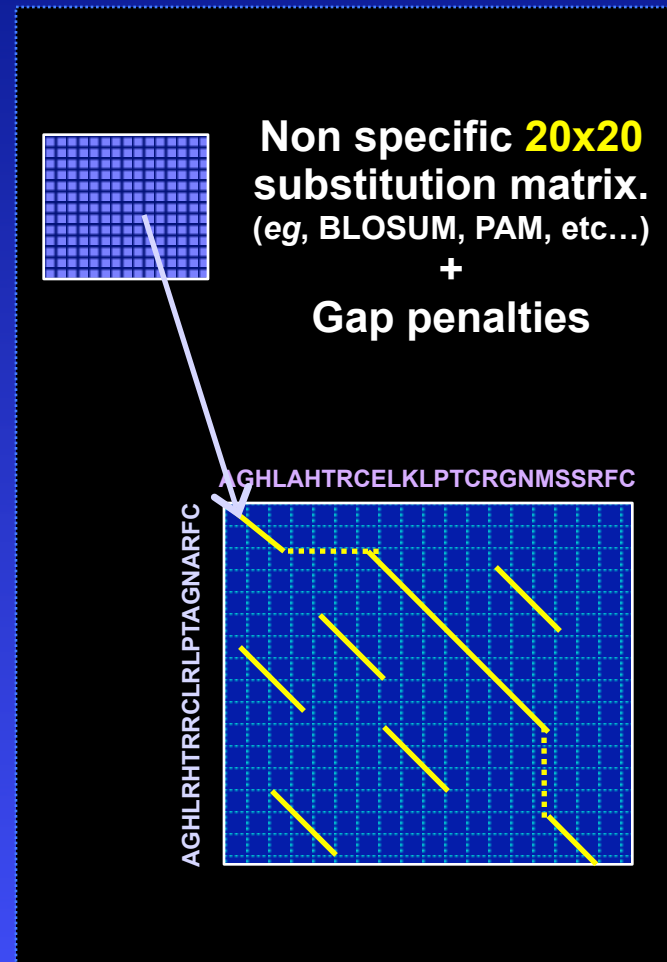
BLAST2SEQ: Local method

Prof.-Seq.

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

Prof.-Prof.

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



Methods: Evaluated methods

Sequence A: AGHLAHTRCCLKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

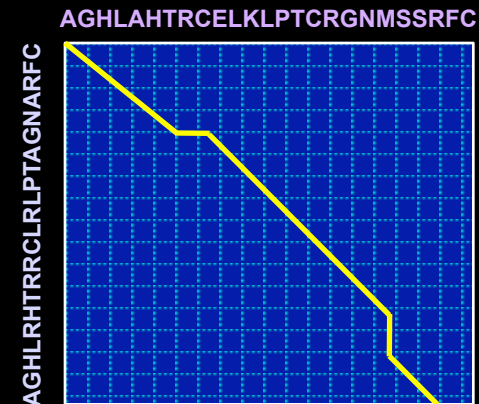
BLAST2SEQ: Local method

Prof.-Seq.

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

Prof.-Prof.

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



Methods: Evaluated methods

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

BLAST2SEQ: Local method

Prof.-Seq.

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

Prof.-Prof.

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

AGHLAHT

AGHLAHTRCELK

MSSRFC

AGHLA

LKLPTCRGNMSSRFC

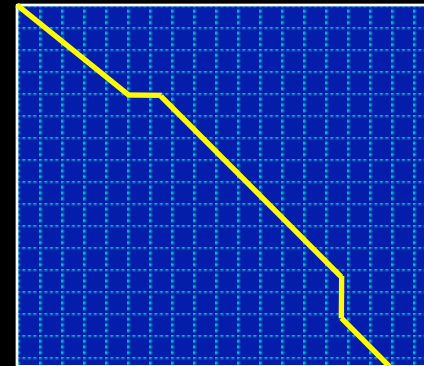
AGHLAHTRCELKLPTCR

SSRFC

AGHLAHTRCELKLPTCRGNMSSRFC

AGHLAHTRCELKLPTCRGNMSSRFC

AGHLRHTRRCLRLPTAGNARFC



Methods: Evaluated methods

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

BLAST2SEQ: Local method

Prof.-Seq.

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

Prof.-Prof.

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

	A	C	D	E	...	V	W	Y
PSSM A	+3	-1	-2	-2	...	-2	-1	-3



AGHLAHT

AGHLAHTRCELK

MSSRFC

AGHLA

LKLPTCRGNMSSRFC

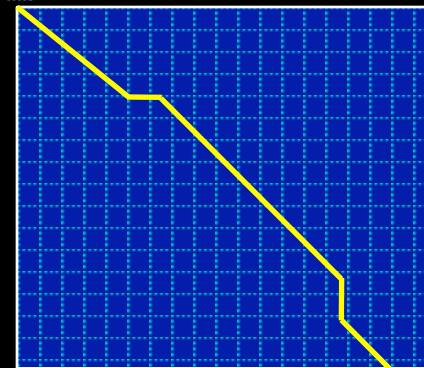
AGHLAHTRCELKLPTCR

SSRFC

AGHLAHTRCELKLPTCRGNMSSRFC

AGHLAHTRCELKLPTCRGNMSSRFC

AGHLRHTRRCLRLPTAGNARFC



Methods: Evaluated methods

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

BLAST2SEQ: Local method

Prof.-Seq.

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

Prof.-Prof.

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

AGHLAHT

AGHLAHTRCELK

MSSRFC

AGHLA

LKLPTCRGNMSSRFC

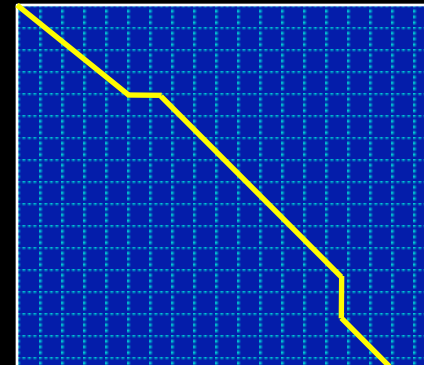
AGHLAHTRCELKLPTCR

SSRFC

AGHLAHTRCELKLPTCRGNMSSRFC

AGHLAHTRCELKLPTCRGNMSSRFC

AGHLRHTRRCLRLPTAGNARFC



Methods: Evaluated methods

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

BLAST2SEQ: Local method

Prof.-Seq.

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

Prof.-Prof.

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

	A	C	D	E	...	V	W	Y
PSSM G	+1	-2	-3	-2	...	-1	+1	-3



AGHLAHT

AGHLAHTRCELK

MSSRFC

AGHLA

LKLPTCRGNMSSRFC

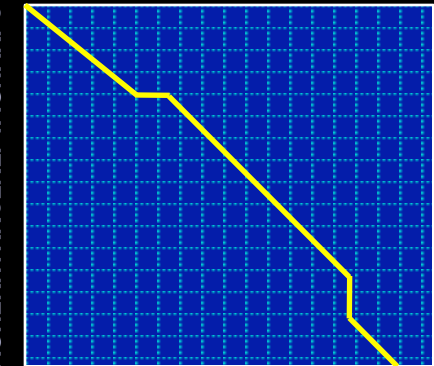
AGHLAHTRCELKLPTCR

SSRFC

AGHLAHTRCELKLPTCRGNMSSRFC

AGHLAHTRCELKLPTCRGNMSSRFC

AGHLRHTRRCLRLPTAGNARFC



Methods. Evaluated methods.

Sequence A: AGHLAHTRC~~EL~~KLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

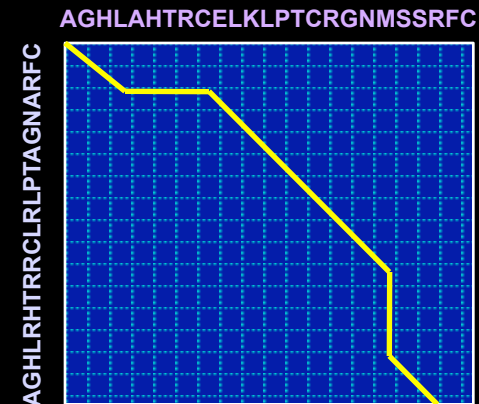
BLAST2SEQ: Local method

Prof.-Seq.

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

Prof.-Prof.

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



Methods. Evaluated methods.

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

BLAST2SEQ: Local method

Prof.-Seq.

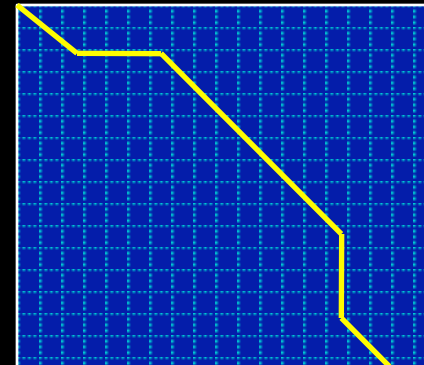
PSI-BLAST: Local search method that uses multiple sequence information for one of the sequences.

Prof.-Prof.

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

AGHLAHTRCELK MSSRFC
AGHLAHT
AGHLA LKLPTCRGNMSSRFC
AGHLAHTRCELKLPTCR SSRFC
AGHLAHTRCELKLPTCRGNMSSRFC
AGHLAHTRCELKLPTCRGNMSSRFC

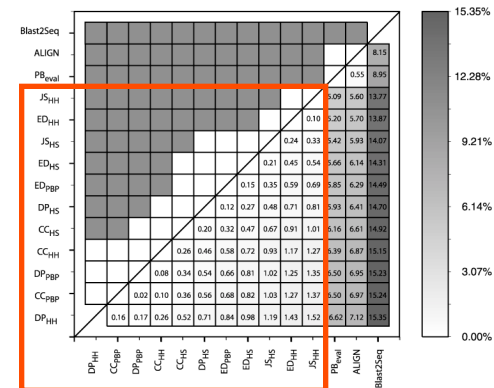
RRCLRLPTAGNARFC
AGHLRHTR AGNARFC
AGHLR RRCLRLPTAGNARFC
AGHLRHTRRCLRLPTAGNARFC



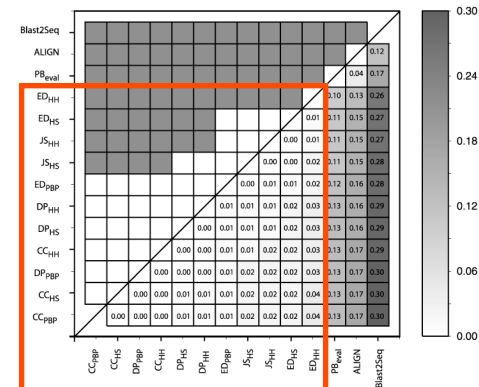
Results: Comparison of alignment dependent measures

<i>ALIGN4D</i> protocol	% of Correct SeqA	% of Correct SeqB	Shift Score
CC_{PBP}	55.34 [8.00 - 100.00]	55.49 [7.00 - 100.00]	0.61 [0.08 - 1.00]
CC_{HH}	54.96 [8.00 - 100.00]	55.30 [7.00 - 100.00]	0.61 [-0.07 - 1.00]
CC_{HS}	54.48 [6.00 - 100.00]	54.80 [7.00 - 100.00]	0.61 [0.04 - 1.00]
ED_{PBP}	54.22 [6.00 - 99.00]	54.17 [7.00 - 99.00]	0.60 [-0.07 - 0.99]
ED_{HH}	52.90 [8.00 - 100.00]	53.01 [7.00 - 100.00]	0.58 [-0.07 - 1.00]
ED_{HS}	53.70 [9.00 - 100.00]	53.89 [7.00 - 100.00]	0.59 [-0.07 - 1.00]
DP_{PBP}	55.02 [7.00 - 100.00]	55.47 [7.00 - 100.00]	0.61 [0.00 - 1.00]
DP_{HH}	55.50 [7.00 - 100.00]	55.81 [9.00 - 100.00]	0.61 [-0.06 - 1.00]
DP_{HS}	54.07 [6.00 - 100.00]	54.41 [7.00 - 100.00]	0.61 [0.01 - 1.00]
JS_{HH}	52.56 [6.00 - 100.00]	52.82 [7.00 - 100.00]	0.59 [0.03 - 1.00]
JS_{HS}	53.24 [6.00 - 100.00]	53.48 [7.00 - 100.00]	0.60 [-0.01 - 1.00]
ALIGN	41.55 [6.00 - 94.00]	41.84 [5.00 - 94.00]	0.44 [-0.07 - 0.96]
BLAST2SEQ	26.09 [0.00 - 92.00]	26.07 [0.00 - 93.00]	0.32 [-0.08 - 0.95]
PB (e-val)	42.95 [0.00 - 96.00]	43.11 [0.00 - 95.00]	0.48 [-0.12 - 0.98]

A) % of correctly aligned positions.



B) Shift Score.



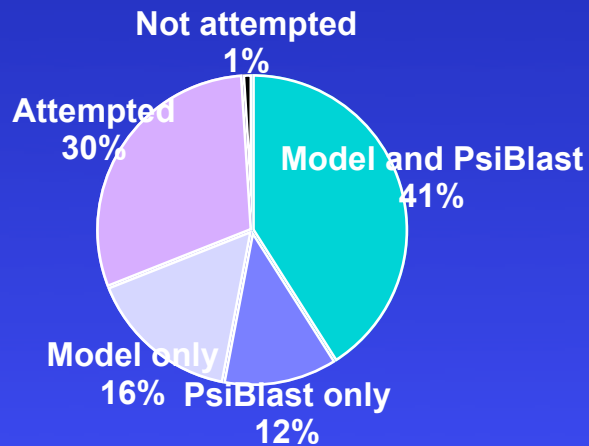
Results: Comparison of success rates

Method	% of alignments at 1Å	% of alignments at 2Å	% of alignments at 3Å	% of alignments at average
CE	20.50	82.50	100.00	82.50
ALIGN	8.50	23.00	35.00	21.00
BLAST2SEQ	8.00	21.50	30.00	20.00
PB (e-val)	8.00	31.00	45.50	29.50
CC _{PBP}	11.50	37.00	55.50	35.50
DP _{PBP}	11.00	37.50	53.50	35.50

Results. Turn over.

Mycoplasma genitalium MODPIPE Models

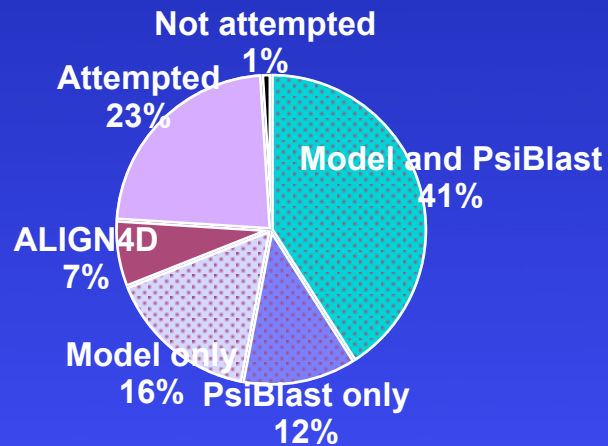
Number of ORFs	479
Average ORF length	364



Results. Turn over.

Mycoplasma genitalium MODPIPE Models

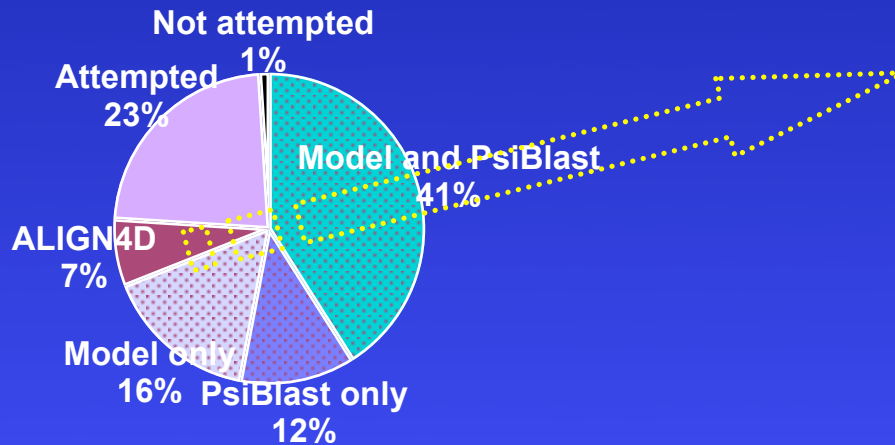
Number of ORFs	479
Average ORF length	364



Results. Turn over.

Mycoplasma genitalium MODPIPE Models

Number of ORFs	479
Average ORF length	364



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

~ **50,000** models
for TrEMBL-SP
“genome”.

Examples: T0092 model

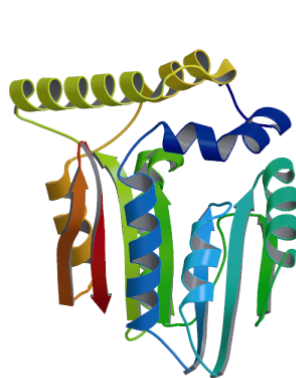
- Target T0092 at CASP4:

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

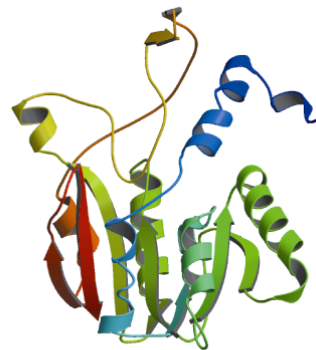
Method	RMSD Å	% of EqPos
ALIGN4D CC_{PBP}	5.9	67.84
PSI-BLAST	4.9	31.72
Best predictions at CASP4	6.0	65.20

Data from CASP4, Asilomar, CA, December 2000.

B) Target T0092



X-Ray structure



ALIGN4D (CC_{PBP}) model

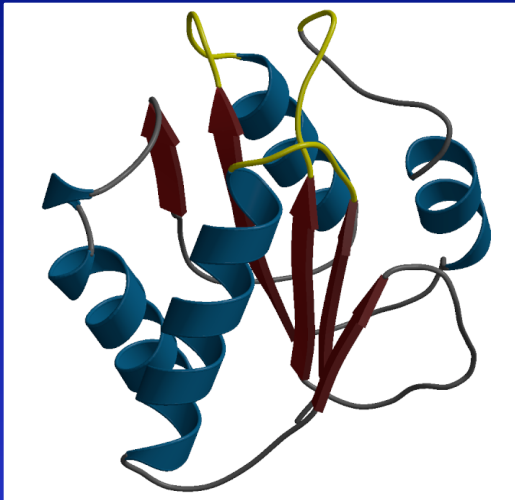


Psi-Blast model

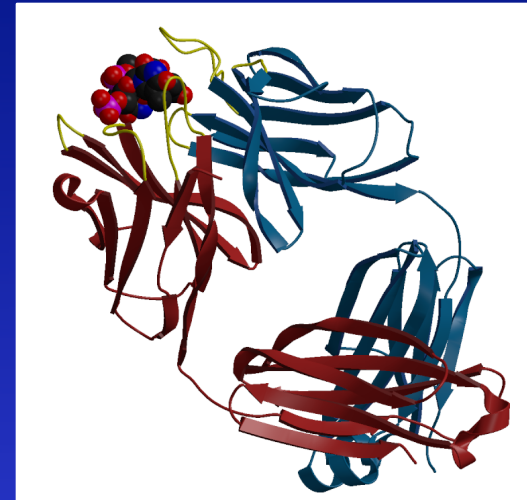
Summary

- ✓ What is comparative modeling and why is it useful?
- ✓ Steps in CM (overview + some details)
- ✓ Accuracy of comparative models
- ✓ Target-Template alignment
- ✓ **Loop modeling**
- ✓ CM and Structural Genomics

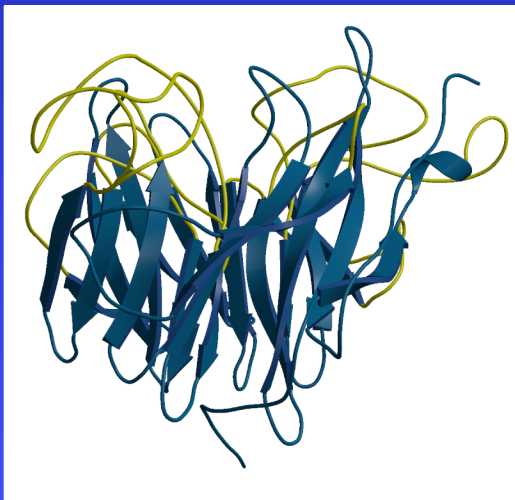
Loop Modeling in Protein Structures



$\alpha+\beta$ barrel: flavodoxin



IG fold: immunoglobulin

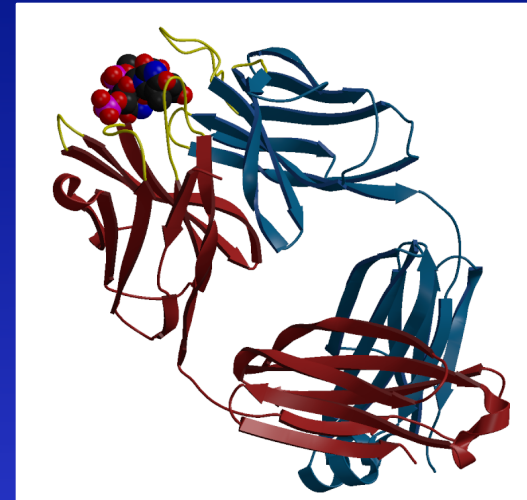


antiparallel β -barrel

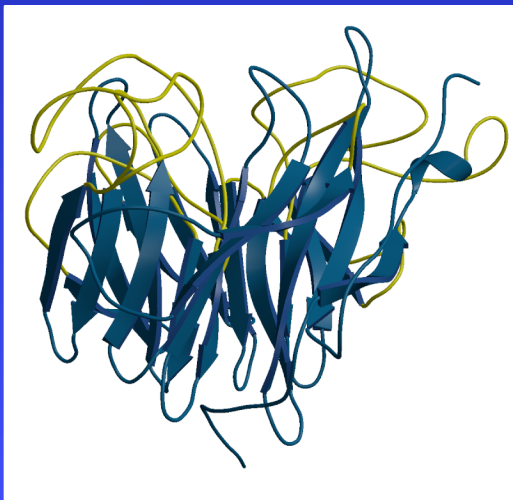
Loop Modeling in Protein Structures



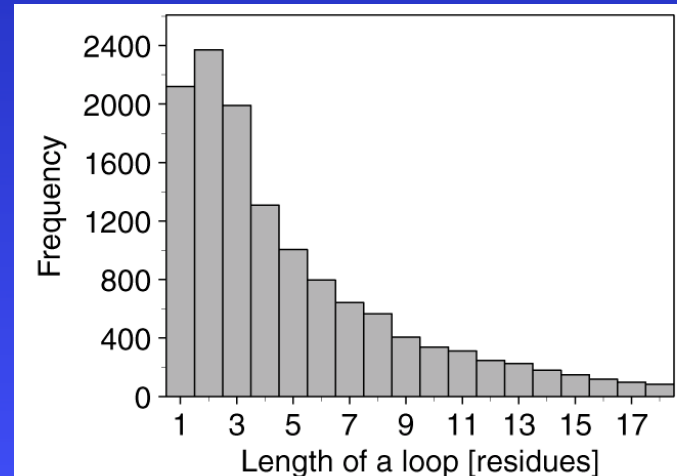
$\alpha+\beta$ barrel: flavodoxin



IG fold: immunoglobulin



antiparallel β -barrel



Loop modeling strategies

Loop modeling strategies

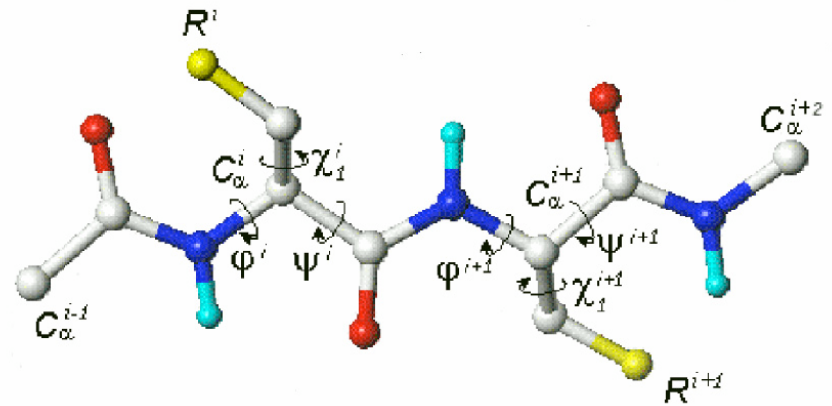
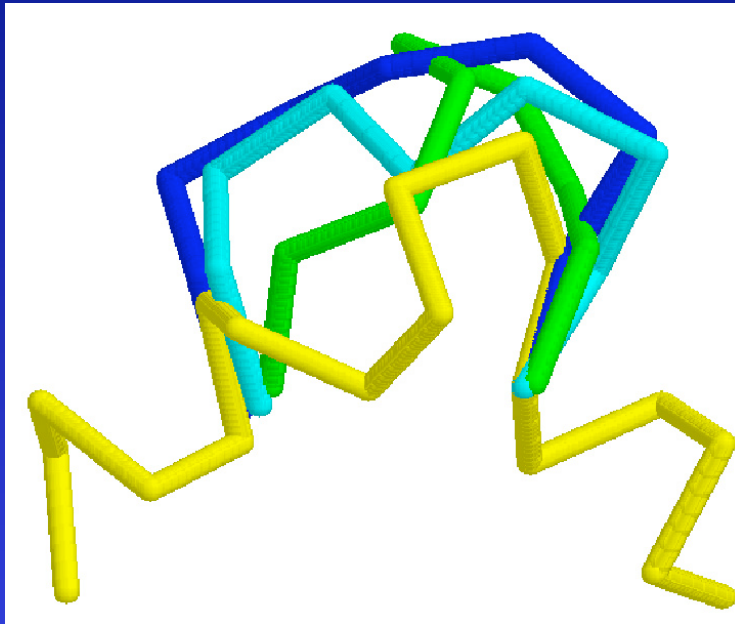
Database search

Conformational search

Loop modeling strategies

Database search

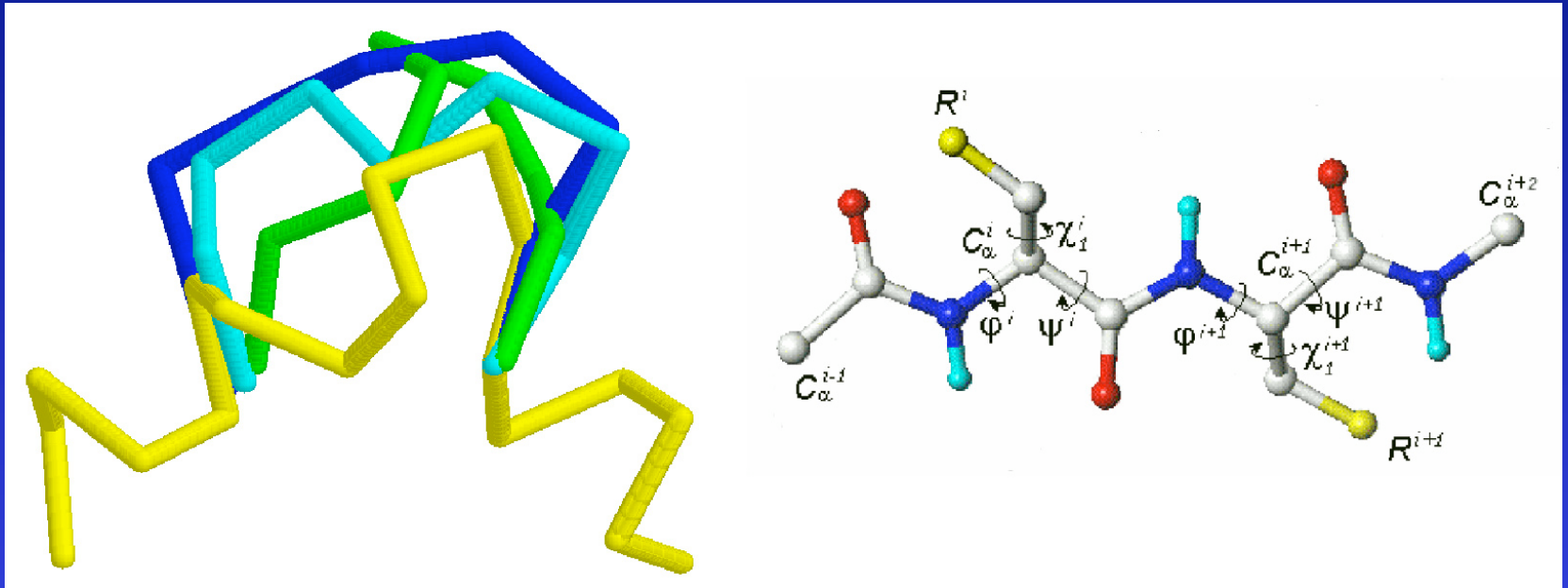
Conformational search



Loop modeling strategies

Database search

Conformational search

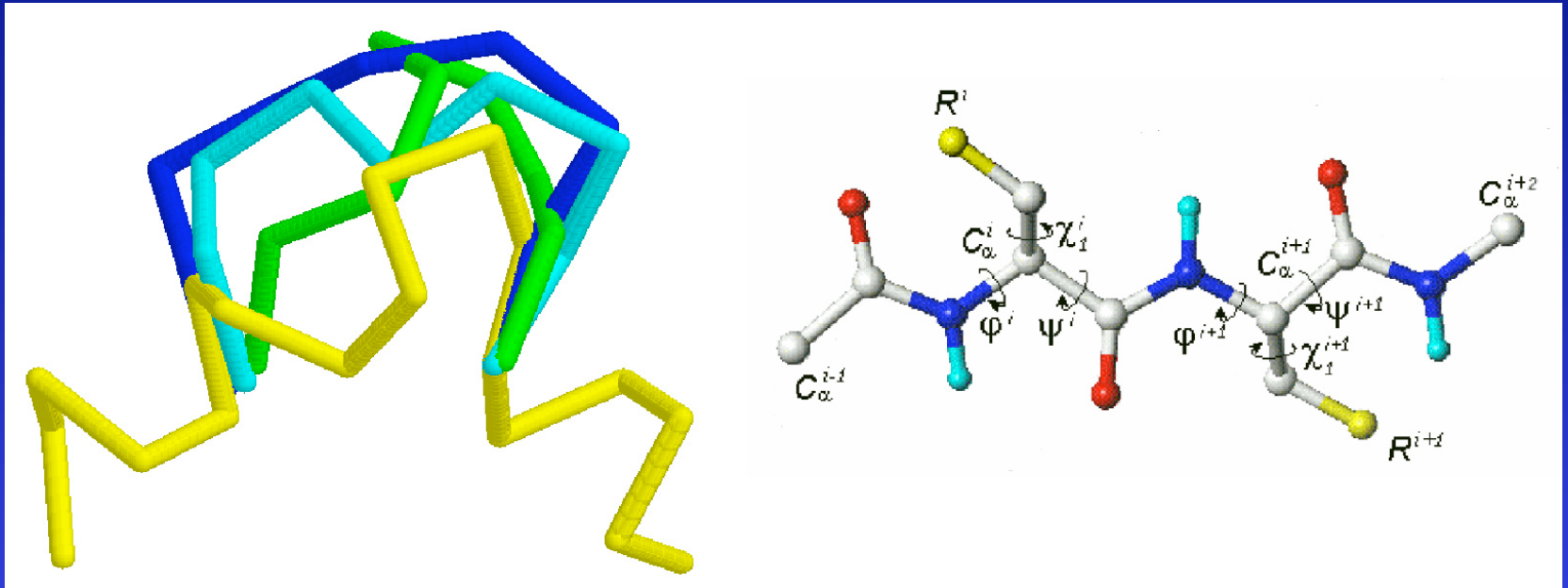


- database is complete only up to 6-8 residues

Loop modeling strategies

Database search

Conformational search

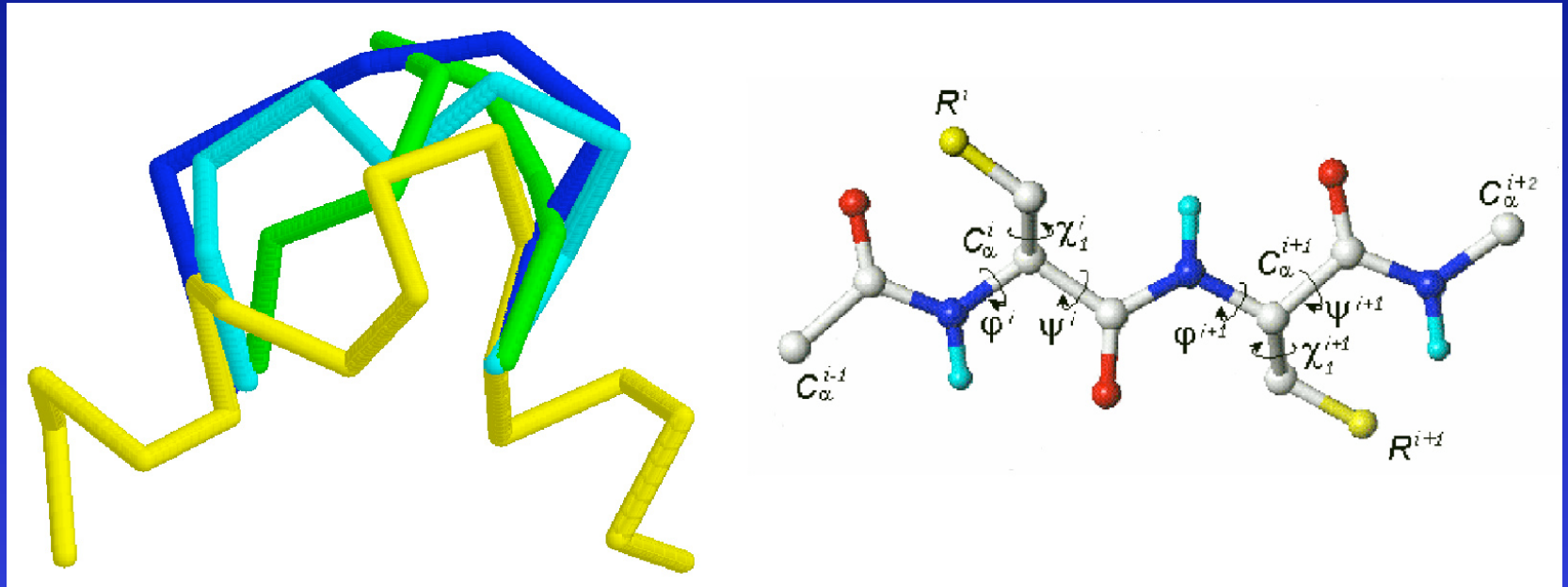


- database is complete only up to 6-8 residues
- even in DB search, the different conformations must be ranked

Loop modeling strategies

Database search

Conformational search

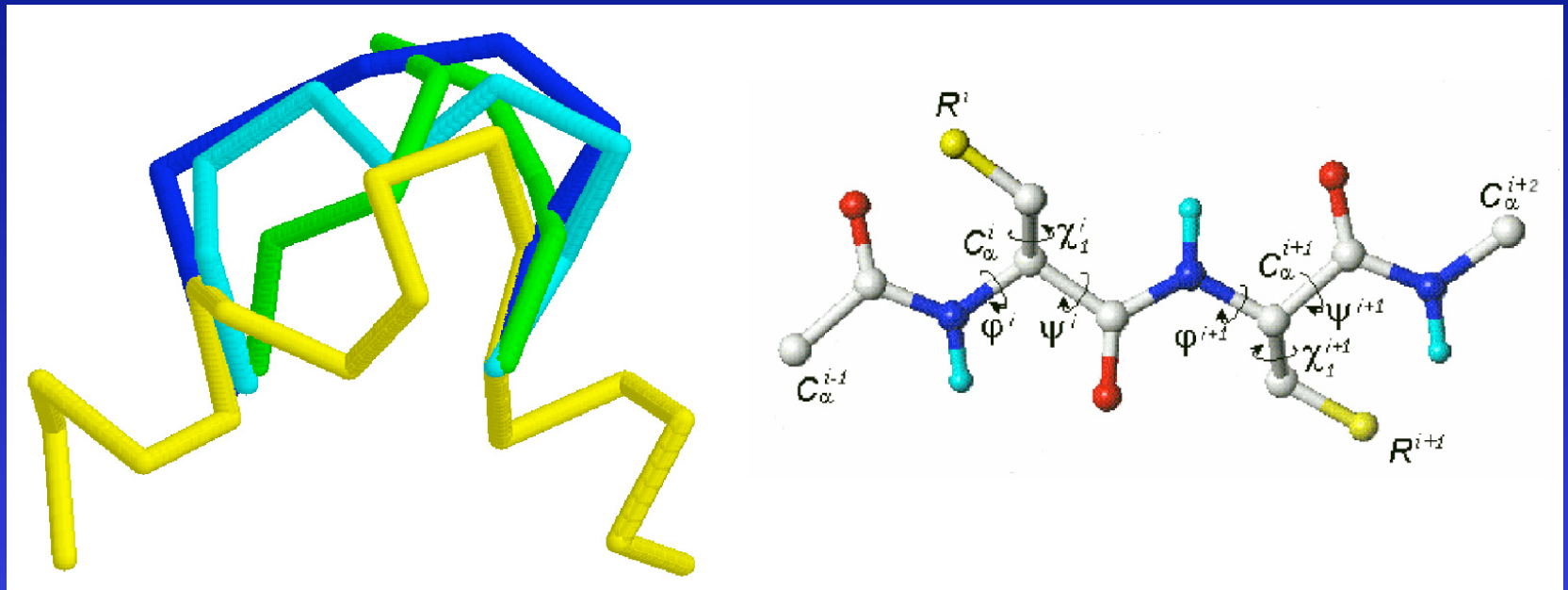


- database is complete only up to 6-8 residues
- even in DB search, the different conformations must be ranked
- loops longer than 4 residues need extensive optimization

Loop modeling strategies

Database search

Conformational search

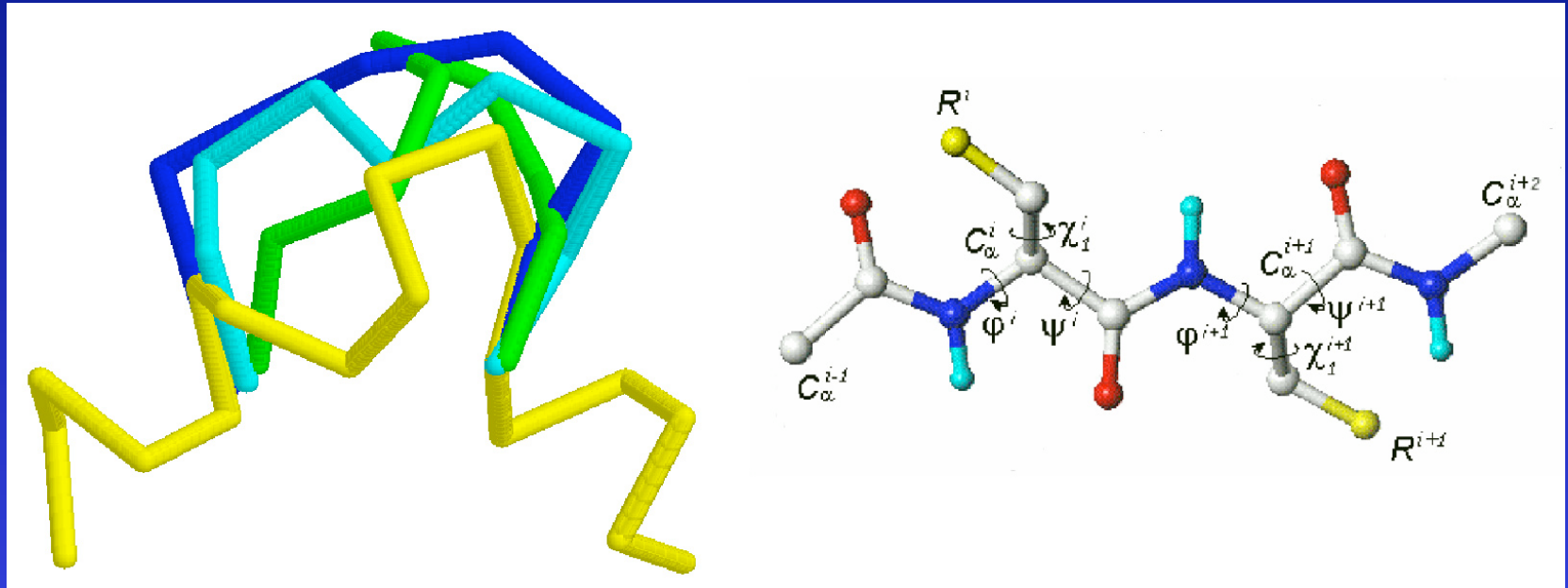


- database is complete only up to 6-8 residues
- even in DB search, the different conformations must be ranked
- loops longer than 4 residues need extensive optimization
- DB method is efficient for specific families (eg. Canonical loops in Ig's,

Loop modeling strategies

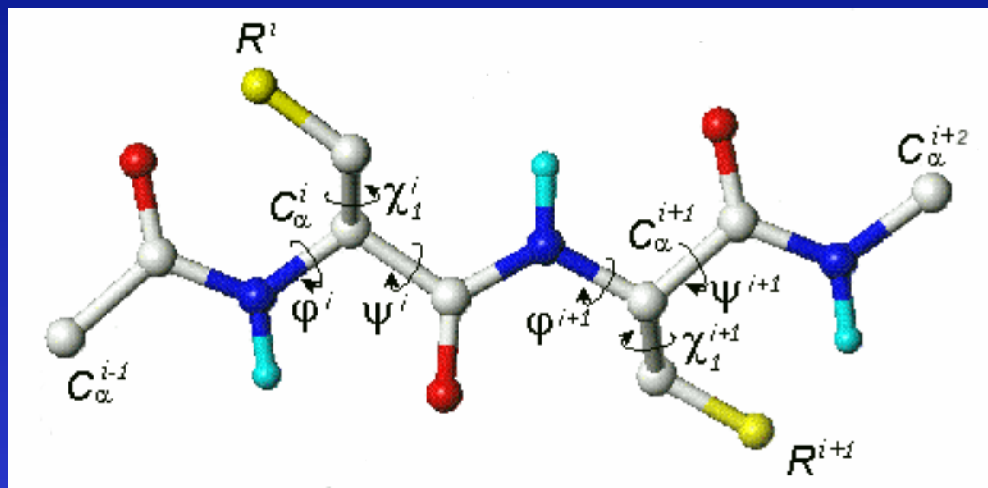
Database search

Conformational search

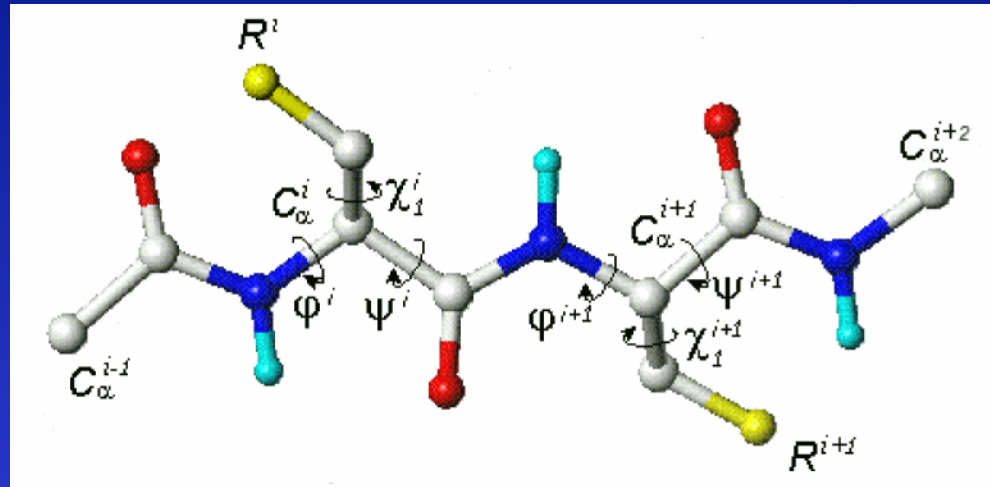


- database is complete only up to 6-8 residues
- even in DB search, the different conformations must be ranked
- loops longer than 4 residues need extensive optimization
- DB method is efficient for specific families (eg. Canonical loops in Ig's, β -hairpins etc)

Loop Modeling by Conformational Search

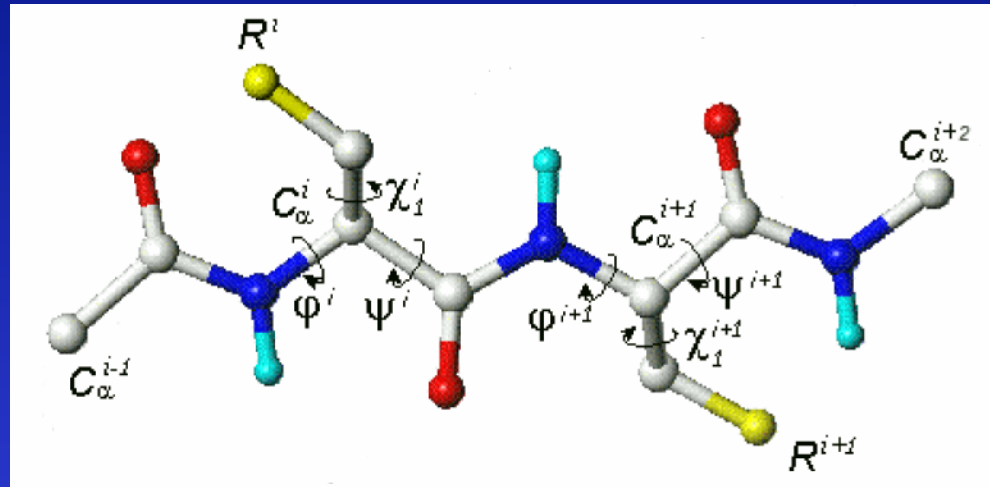


Loop Modeling by Conformational Search



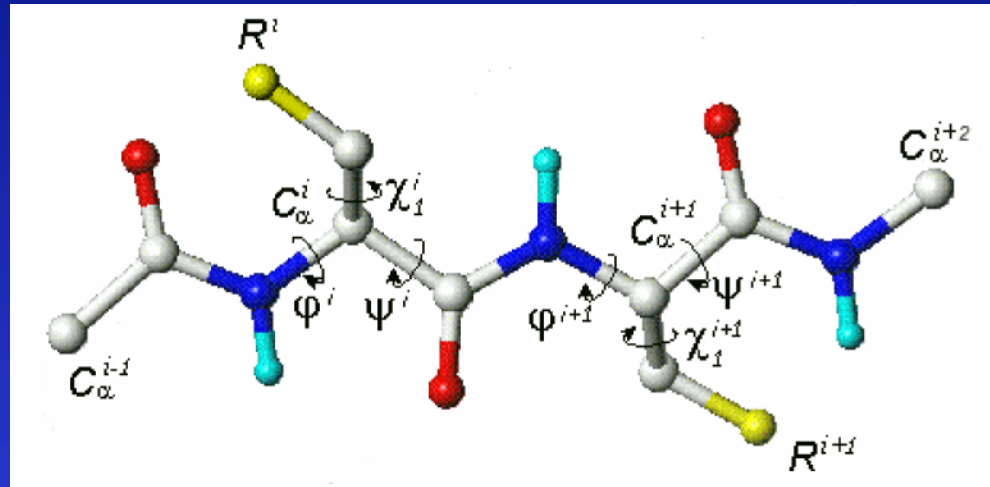
1. Protein representation.

Loop Modeling by Conformational Search



1. Protein representation.
2. Energy (scoring) function.

Loop Modeling by Conformational Search



1. Protein representation.
2. Energy (scoring) function.
3. Optimization algorithm.

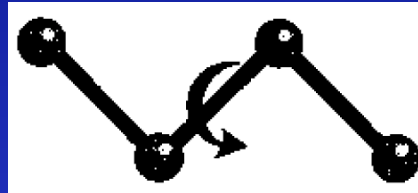
Energy Function for Loop Modeling

The energy function is a sum of many terms:

Energy Function for Loop Modeling

The energy function is a sum of many terms:

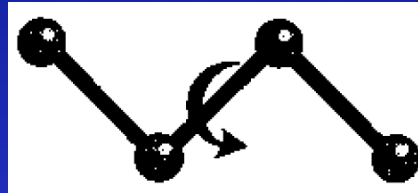
1) Statistical preferences for dihedral angles:



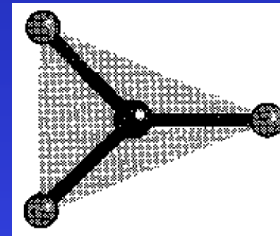
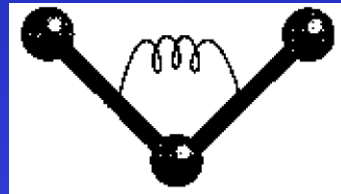
Energy Function for Loop Modeling

The energy function is a sum of many terms:

1) Statistical preferences for dihedral angles:



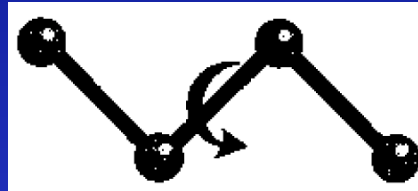
2) Restraints from the CHARMM-22 force field:



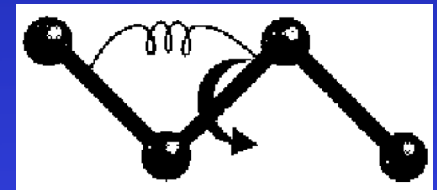
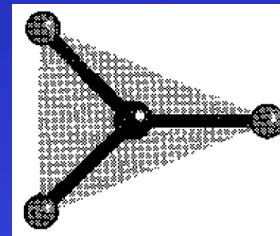
Energy Function for Loop Modeling

The energy function is a sum of many terms:

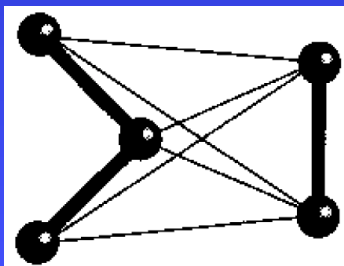
1) Statistical preferences for dihedral angles:



2) Restraints from the CHARMM-22 force field:

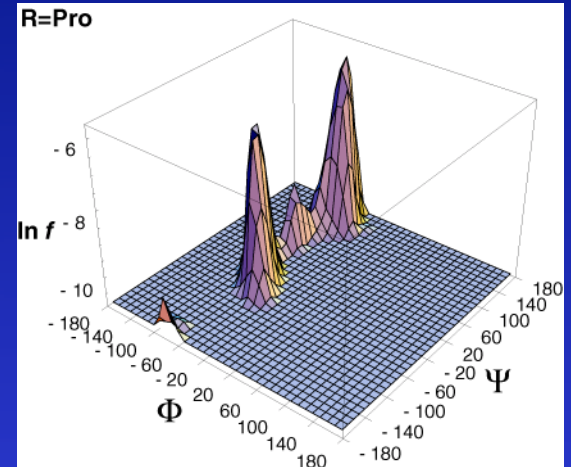
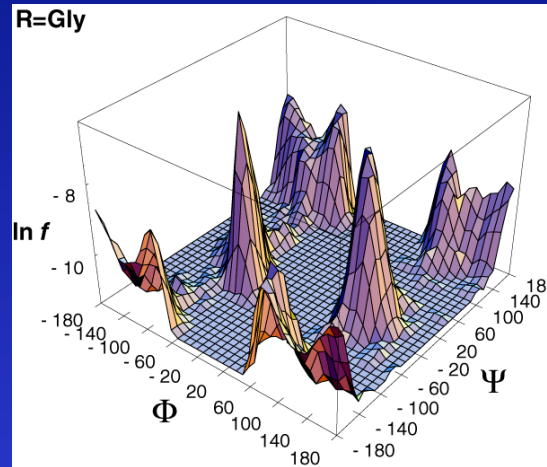
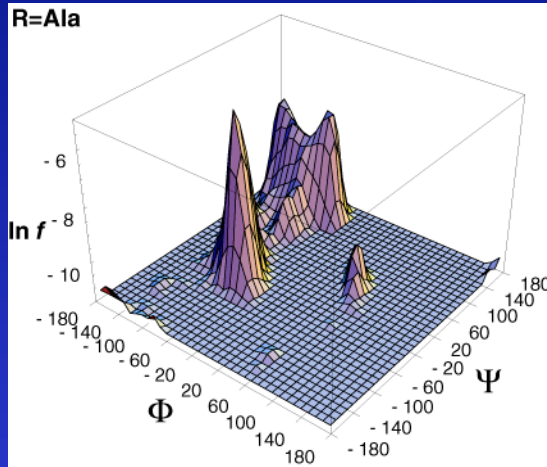


3) Statistical potential for non-bonded contacts:

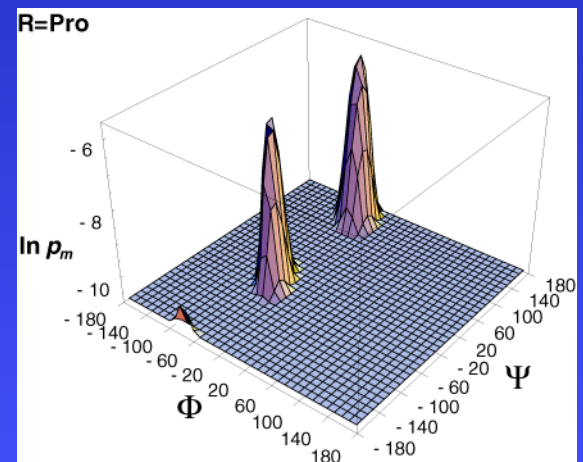
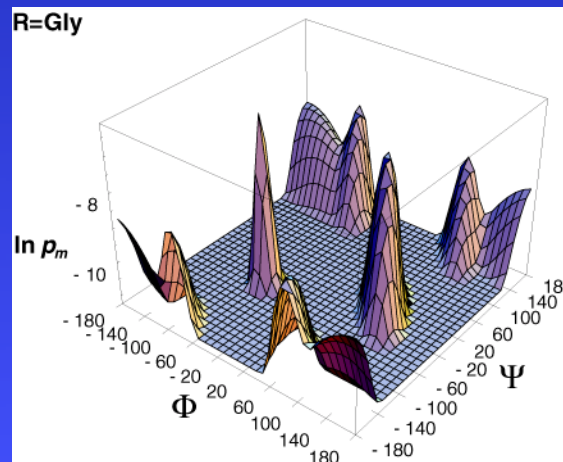
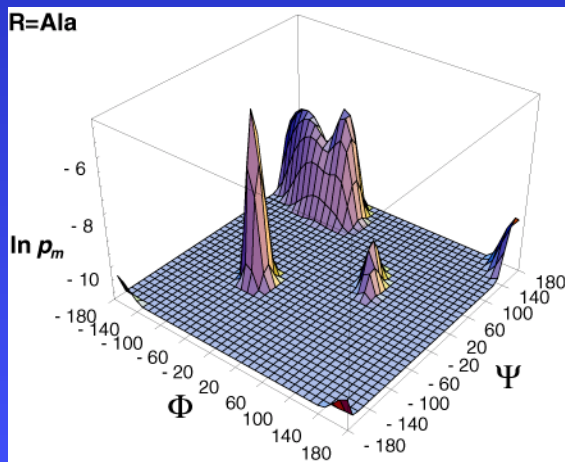
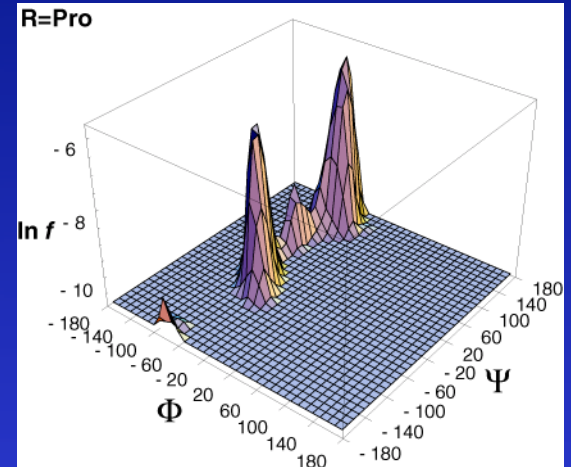
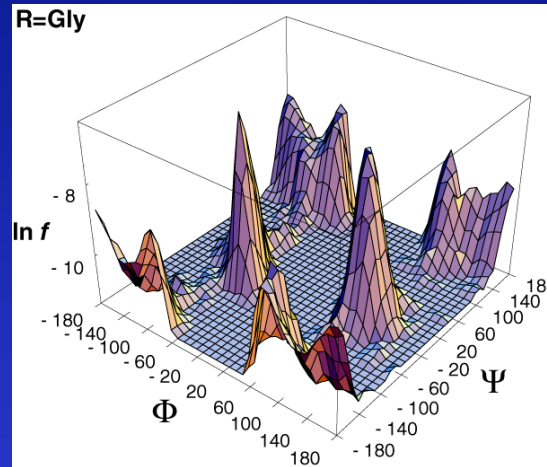
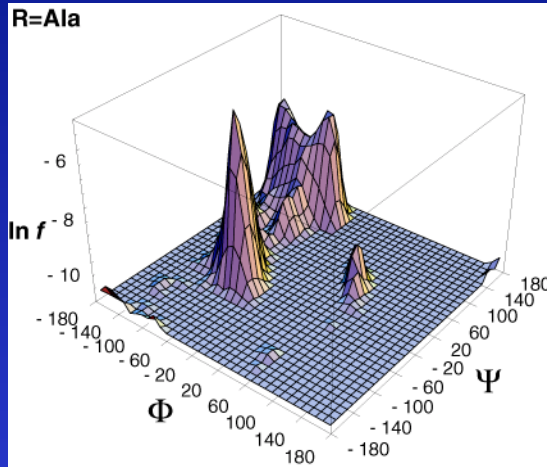


Mainchain Terms for Loop Modeling

Mainchain Terms for Loop Modeling

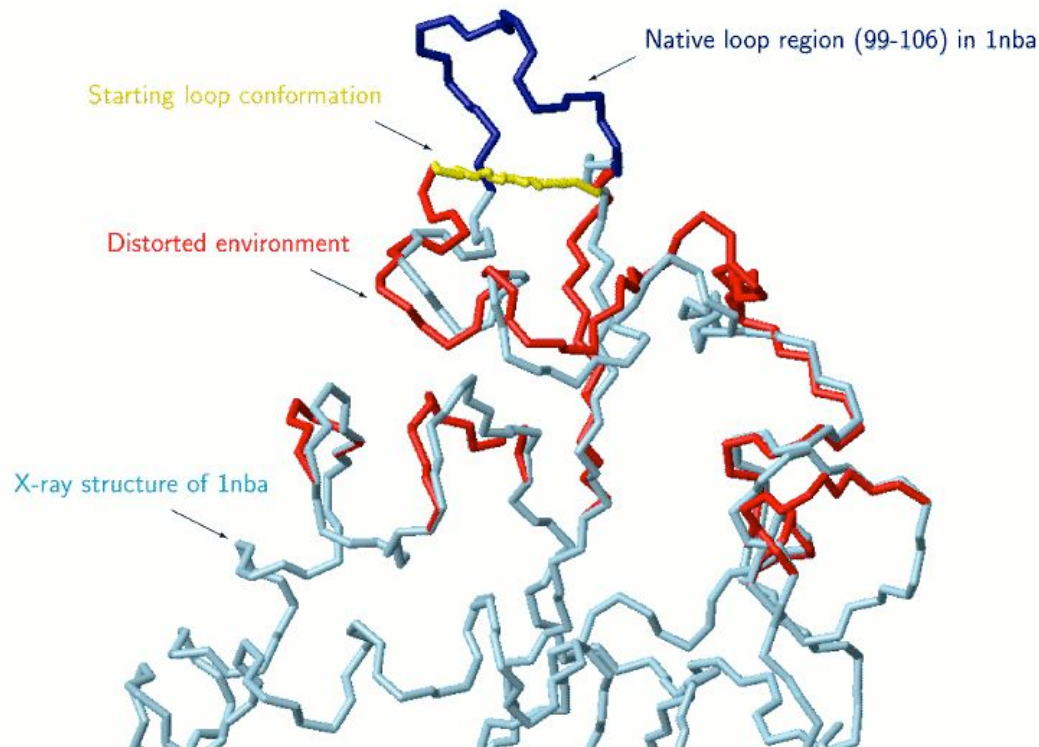


Mainchain Terms for Loop Modeling



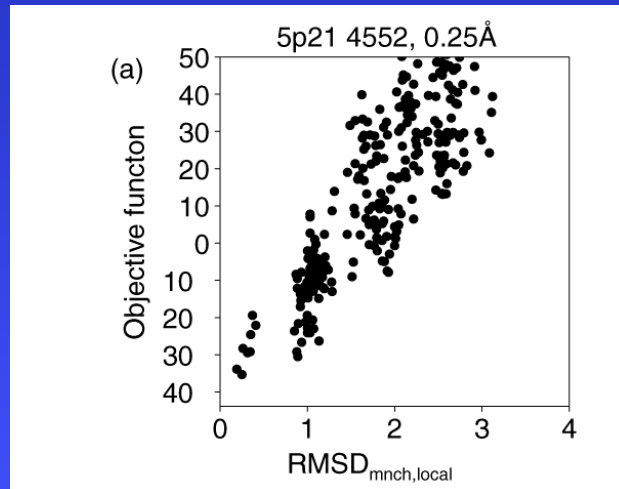
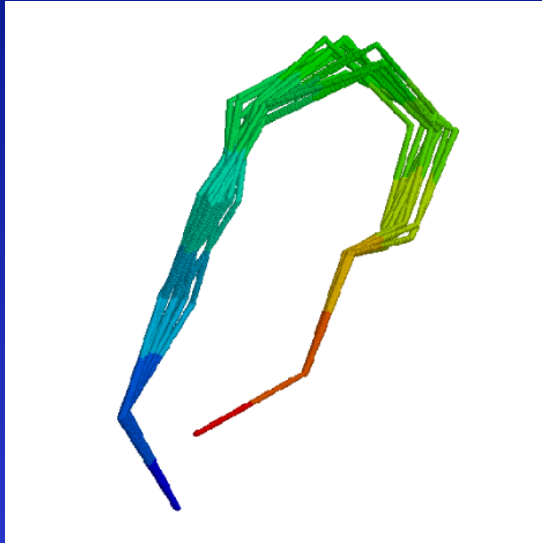
Optimization of Objective Function

- Test set: 40 randomly selected loops of known structures, for each length from 1 to 14 residues.
- Starting conformation: Loop atoms were spaced evenly on a line spanning the two anchor regions, then randomized by ± 5 Å.
- To simulate real comparative modeling situations, performance of the loop modeling problem was determined by predicting loops in only approximately correct environment.

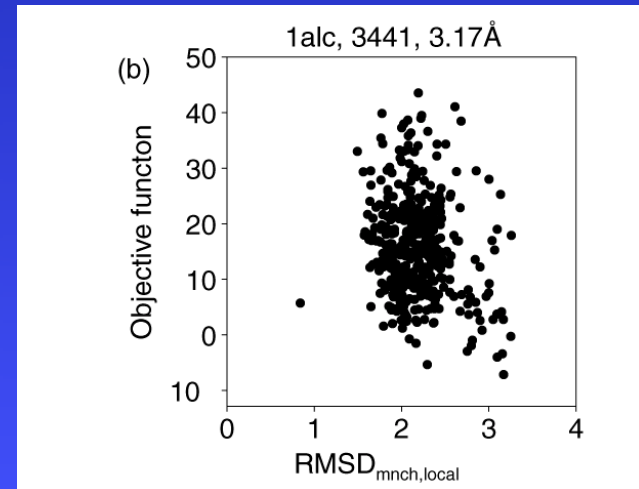
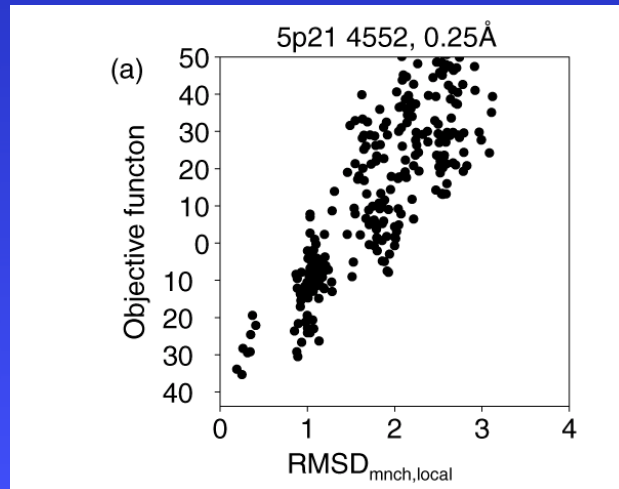
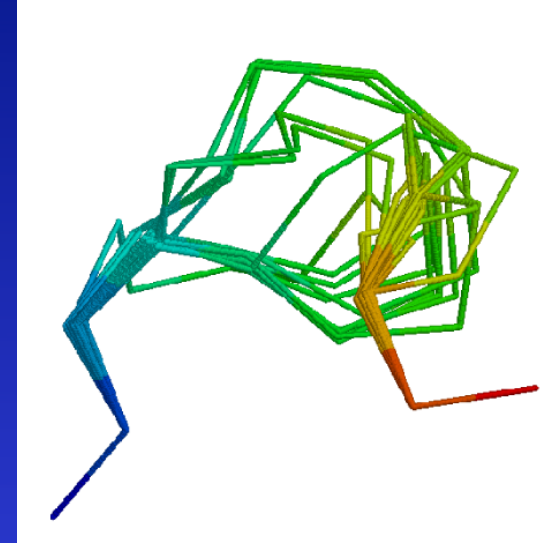
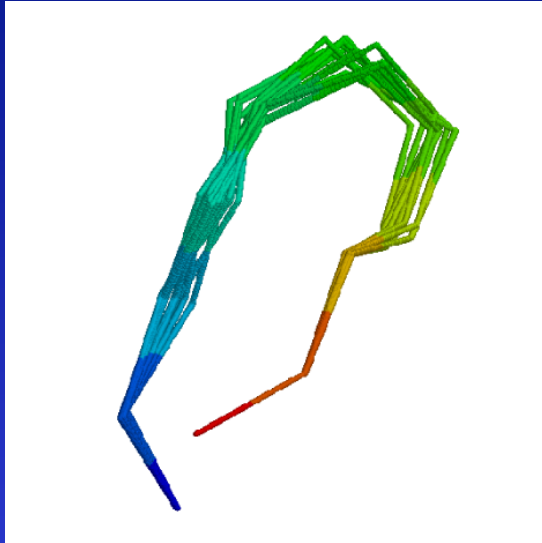


Calculating an Ensemble of Loop Models

Calculating an Ensemble of Loop Models

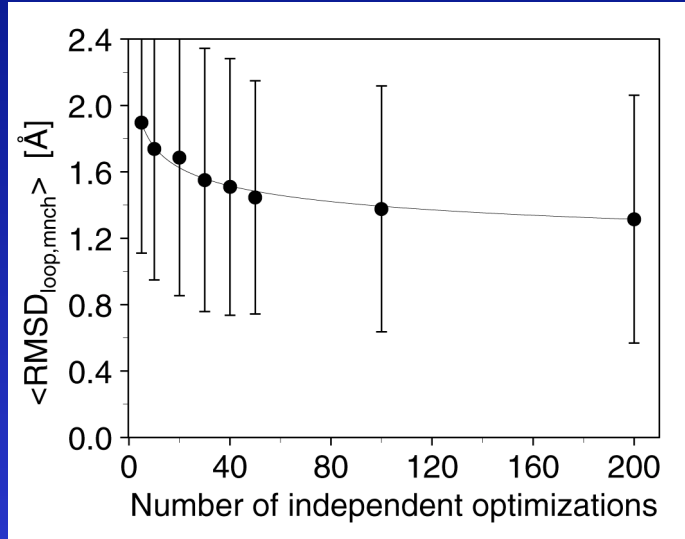


Calculating an Ensemble of Loop Models

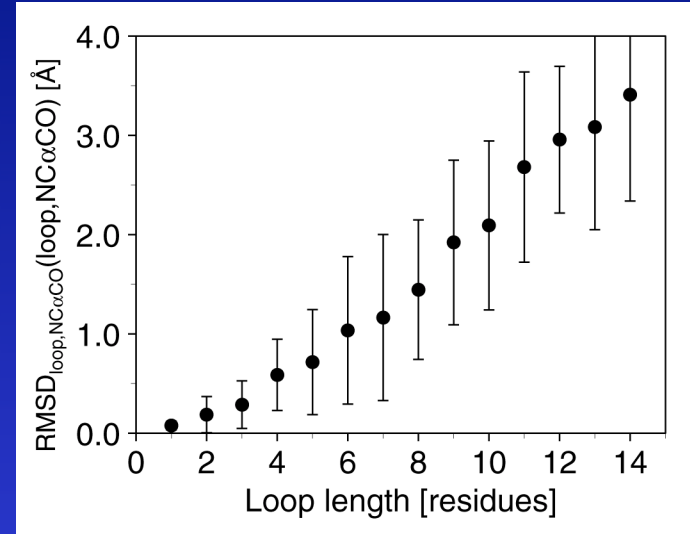
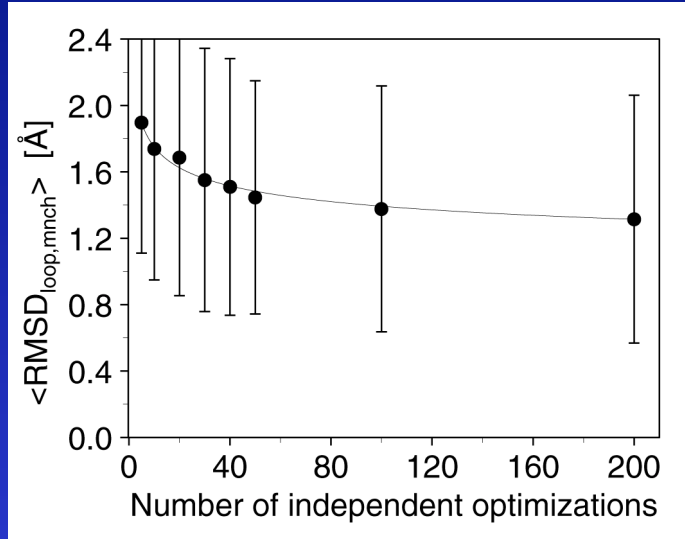


Accuracy of loop models

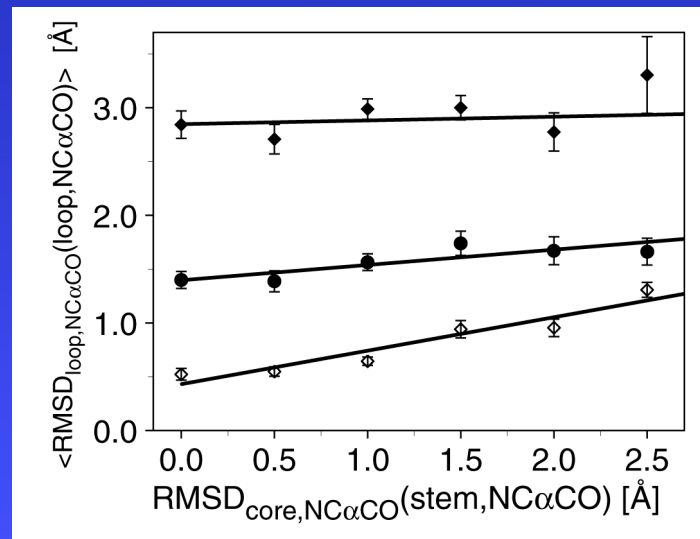
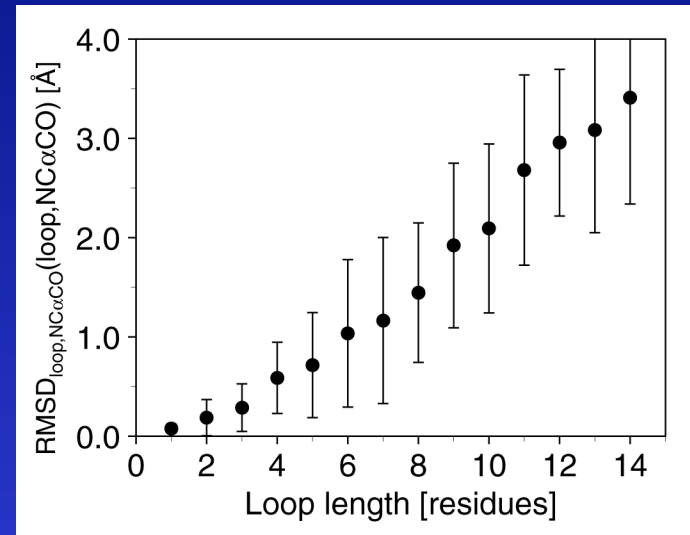
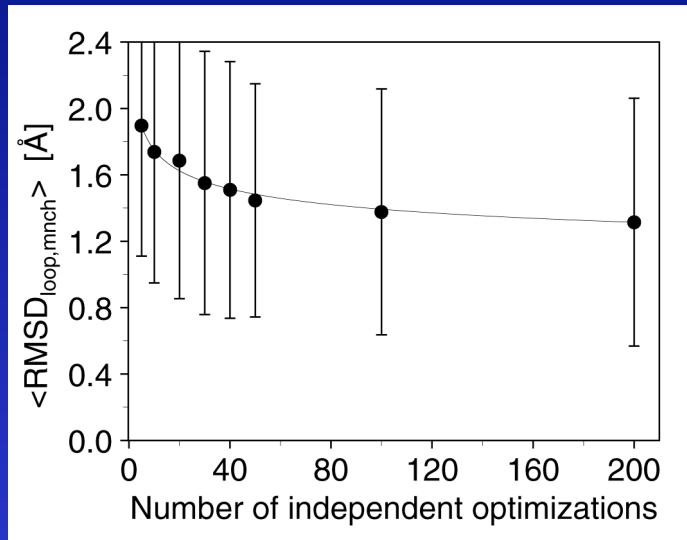
Accuracy of loop models



Accuracy of loop models



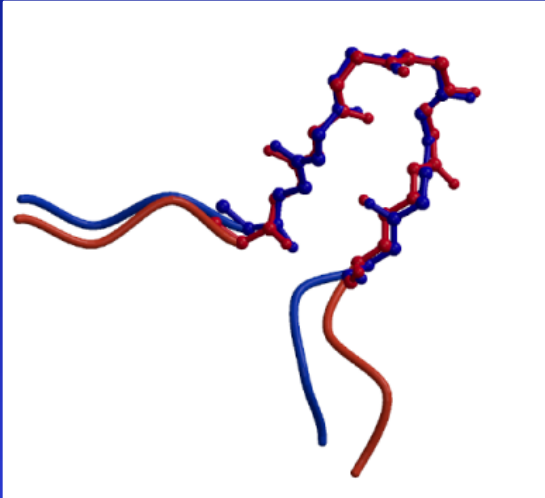
Accuracy of loop models



Accuracy of Loop Modeling

A. Fiser, R. Do & A. Šali, *Prot. Sci.*, **9**, 1537, 2000

Accuracy of Loop Modeling

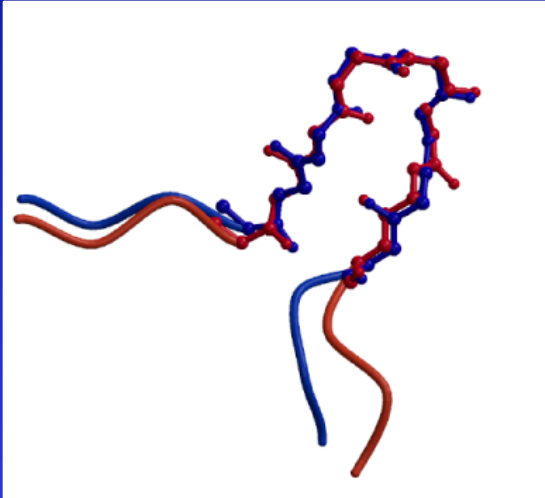


HIGH ACCURACY ($<1\text{\AA}$)

50% (30%) of 8-residue loops

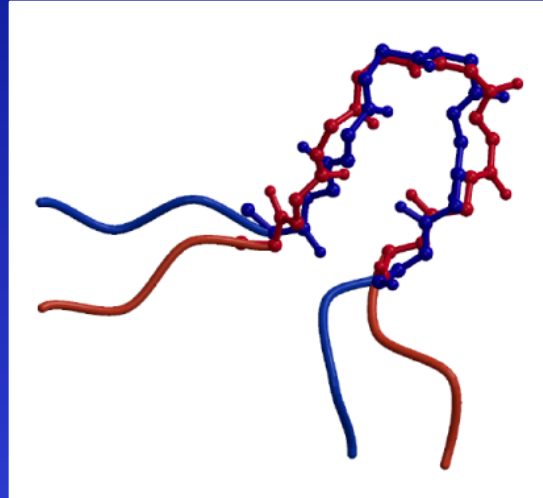
A. Fiser, R. Do & A. Šali, *Prot. Sci.*, **9**, 1537, 2000

Accuracy of Loop Modeling



HIGH ACCURACY ($<1\text{\AA}$)

50% (30%) of 8-residue loops

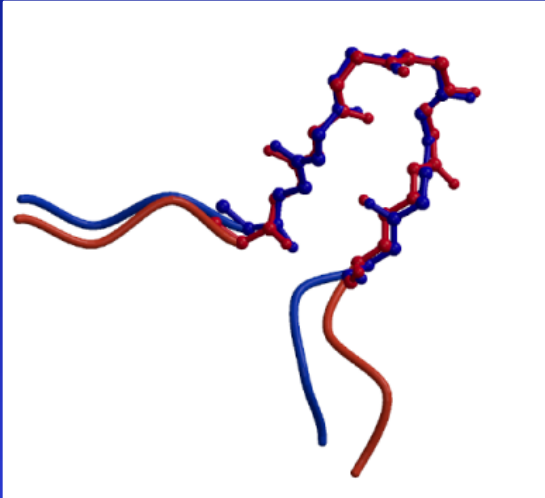


MEDIUM ACCURACY ($<2\text{\AA}$)

40% (48%) of 8-residue loops

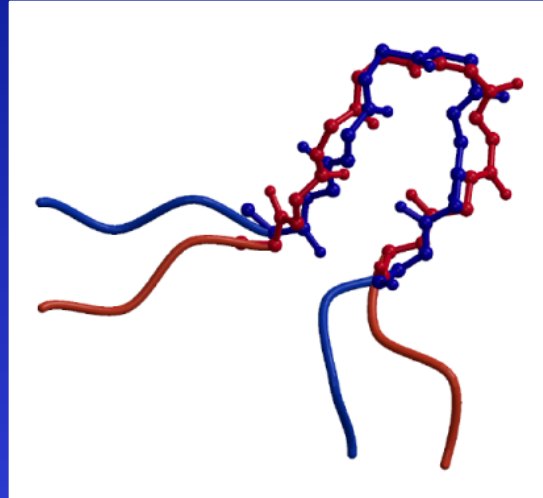
A. Fiser, R. Do & A. Šali, *Prot. Sci.*, **9**, 1537, 2000

Accuracy of Loop Modeling



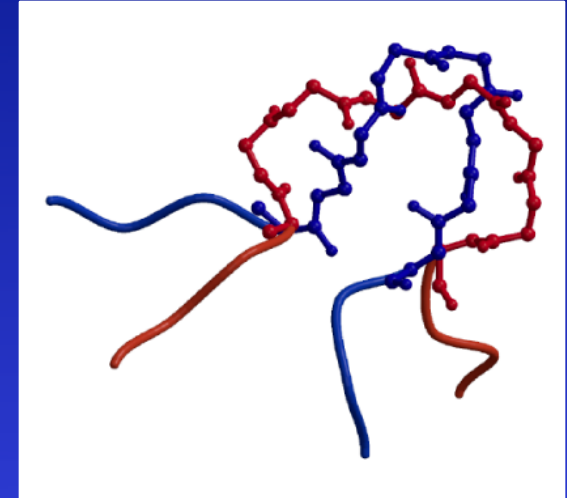
HIGH ACCURACY ($<1\text{\AA}$)

50% (30%) of 8-residue loops



MEDIUM ACCURACY ($<2\text{\AA}$)

40% (48%) of 8-residue loops

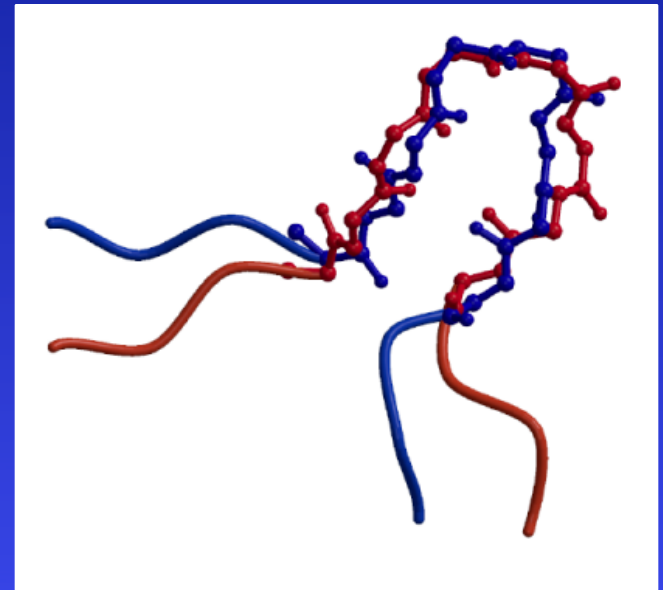
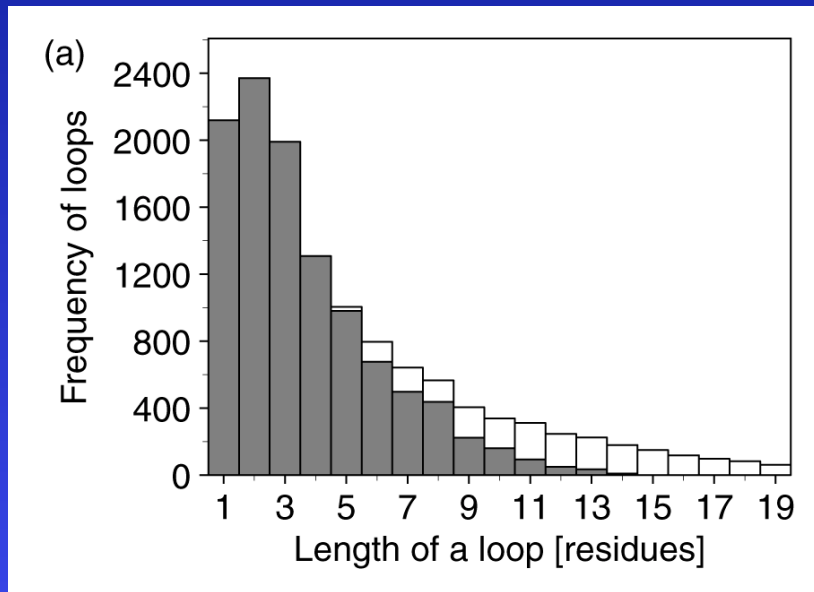


LOW ACCURACY ($>2\text{\AA}$)

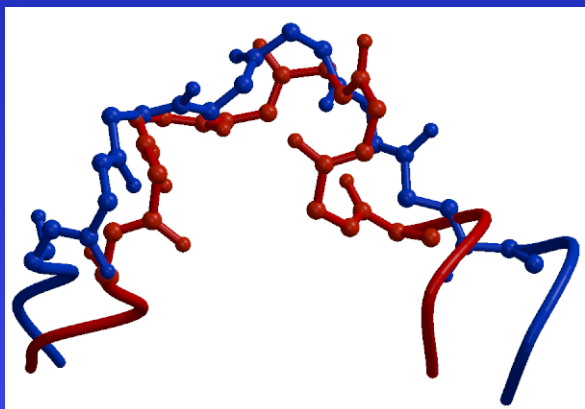
10% (22%) of 8-residue loops

A. Fiser, R. Do & A. Šali, *Prot. Sci.*, **9**, 1537, 2000

Fraction of Loops Modeled With at Least Medium Accuracy



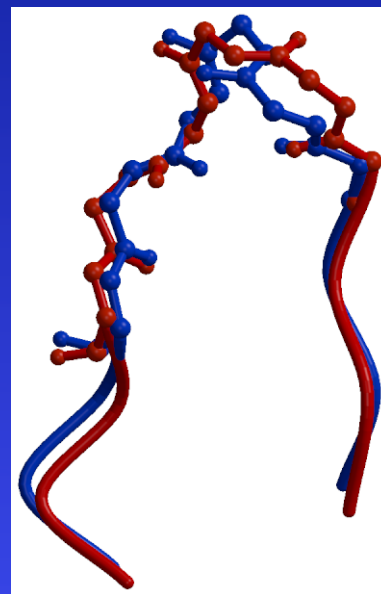
Problems in Practical Loop Modeling



T0076: 46-53

$\text{RMSD}_{\text{mnch}}$ loop = 1.37 Å

$\text{RMSD}_{\text{mnch}}$ anchors = 1.52 Å



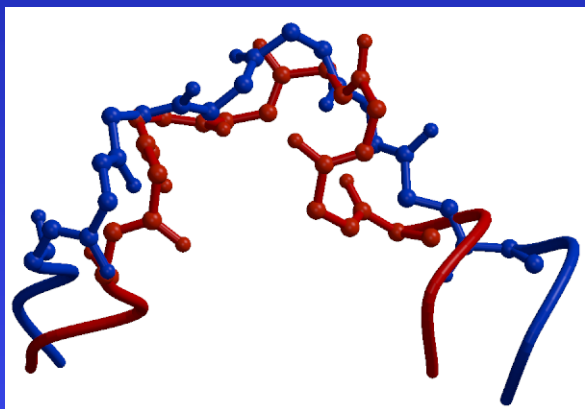
T0058: 80-85

$\text{RMSD}_{\text{mnch}}$ loop = 1.09 Å

$\text{RMSD}_{\text{mnch}}$ anchors = 0.29 Å

Problems in Practical Loop Modeling

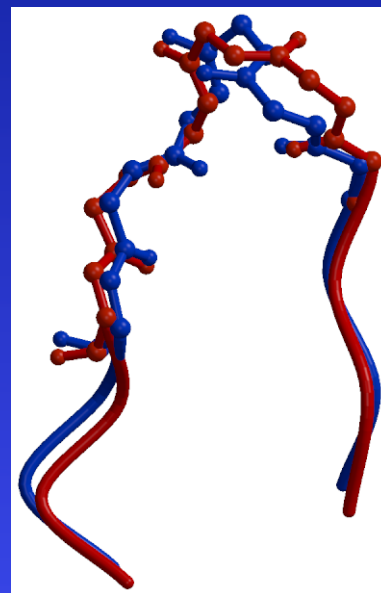
1. Decide which regions to model as loops.



T0076: 46-53

$\text{RMSD}_{\text{mnch}}$ loop = 1.37 Å

$\text{RMSD}_{\text{mnch}}$ anchors = 1.52 Å



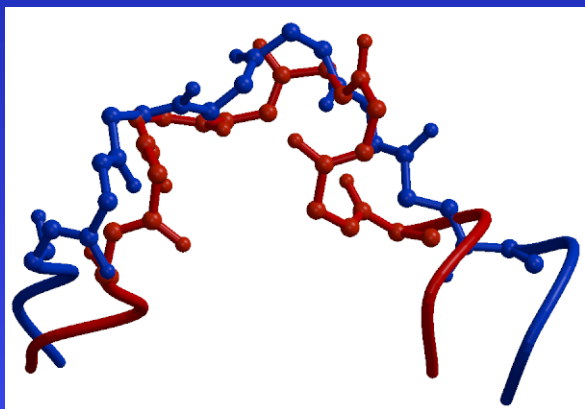
T0058: 80-85

$\text{RMSD}_{\text{mnch}}$ loop = 1.09 Å

$\text{RMSD}_{\text{mnch}}$ anchors = 0.29 Å

Problems in Practical Loop Modeling

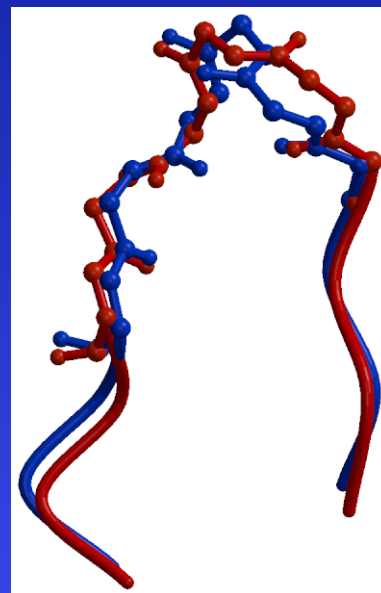
1. Decide which regions to model as loops.
2. Correct alignment of anchor regions & environment.



T0076: 46-53

$\text{RMSD}_{\text{mnch}}$ loop = 1.37 Å

$\text{RMSD}_{\text{mnch}}$ anchors = 1.52 Å



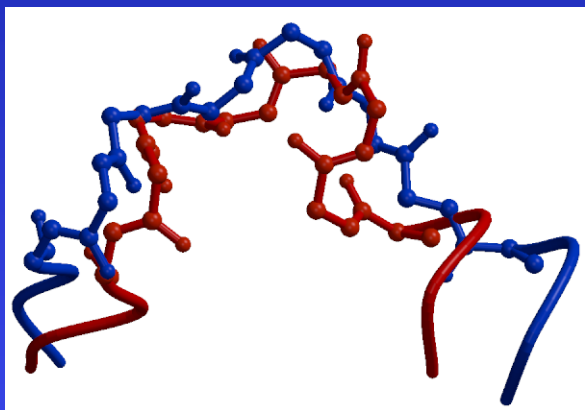
T0058: 80-85

$\text{RMSD}_{\text{mnch}}$ loop = 1.09 Å

$\text{RMSD}_{\text{mnch}}$ anchors = 0.29 Å

Problems in Practical Loop Modeling

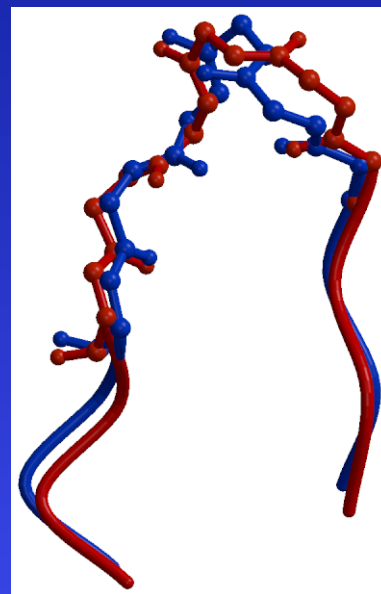
1. Decide which regions to model as loops.
2. Correct alignment of anchor regions & environment.
3. Modeling of a loop.



T0076: 46-53

RMSD_{mnch} loop = 1.37 Å

RMSD_{mnch} anchors = 1.52 Å



T0058: 80-85

RMSD_{mnch} loop = 1.09 Å

RMSD_{mnch} anchors = 0.29 Å

Summary

- ✓ What is comparative modeling and why is it useful?
- ✓ Steps in CM (overview + some details)
- ✓ Accuracy of comparative models
- ✓ Loop modeling
- ✓ **CM and Structural Genomics**

Structural Genomics

Šali. *Nat. Struct. Biol.* **5**, 1029, 1998.
Šali & Kuriyan. *TIBS* **22**, M20, 1999.

Burley *et al.* *Nat. Genet.* **23**, 151, 1999.
Sanchez *et al.* *Nat. Str. Biol.* **7**, 986, 2000

Structural Genomics

✓ **Definition:**

The aim of structural genomics is to put every protein sequence within a modeling distance of a known protein structure.

Šali. *Nat. Struct. Biol.* **5**, 1029, 1998.
Šali & Kuriyan. *TIBS* **22**, M20, 1999.

Burley *et al.* *Nat. Genet.* **23**, 151, 1999.
Sanchez *et al.* *Nat. Str. Biol.* **7**, 986, 2000

Structural Genomics

✓ **Definition:**

The aim of structural genomics is to put every protein sequence within a modeling distance of a known protein structure.

Šali. *Nat. Struct. Biol.* **5**, 1029, 1998.
Šali & Kuriyan. *TIBS* **22**, M20, 1999.

Burley *et al.* *Nat. Genet.* **23**, 151, 1999.
Sanchez *et al.* *Nat. Str. Biol.* **7**, 986, 2000

Structural Genomics

✓ **Definition:**

The aim of structural genomics is to put every protein sequence within a modeling distance of a known protein structure.

✓ **Size of the problem:**

- ✓ There are a few thousand domain fold families.
- ✓ There are ~20,000 sequence families (30% sequence id).

Structural Genomics

✓ **Definition:**

The aim of structural genomics is to put every protein sequence within a modeling distance of a known protein structure.

✓ **Size of the problem:**

- ✓ There are a few thousand domain fold families.
- ✓ There are ~20,000 sequence families (30% sequence id).

✓ **Solution:**

- ✓ Determine many protein structures.
- ✓ Increase modeling distance.

Šali. *Nat. Struct. Biol.* **5**, 1029, 1998.
Šali & Kuriyan. *TIBS* **22**, M20, 1999.

Burley *et al.* *Nat. Genet.* **23**, 151, 1999.
Sanchez *et al.* *Nat. Str. Biol.* **7**, 986, 2000

How can Comparative Modeling be used in Structural Genomics?

- Target Selection

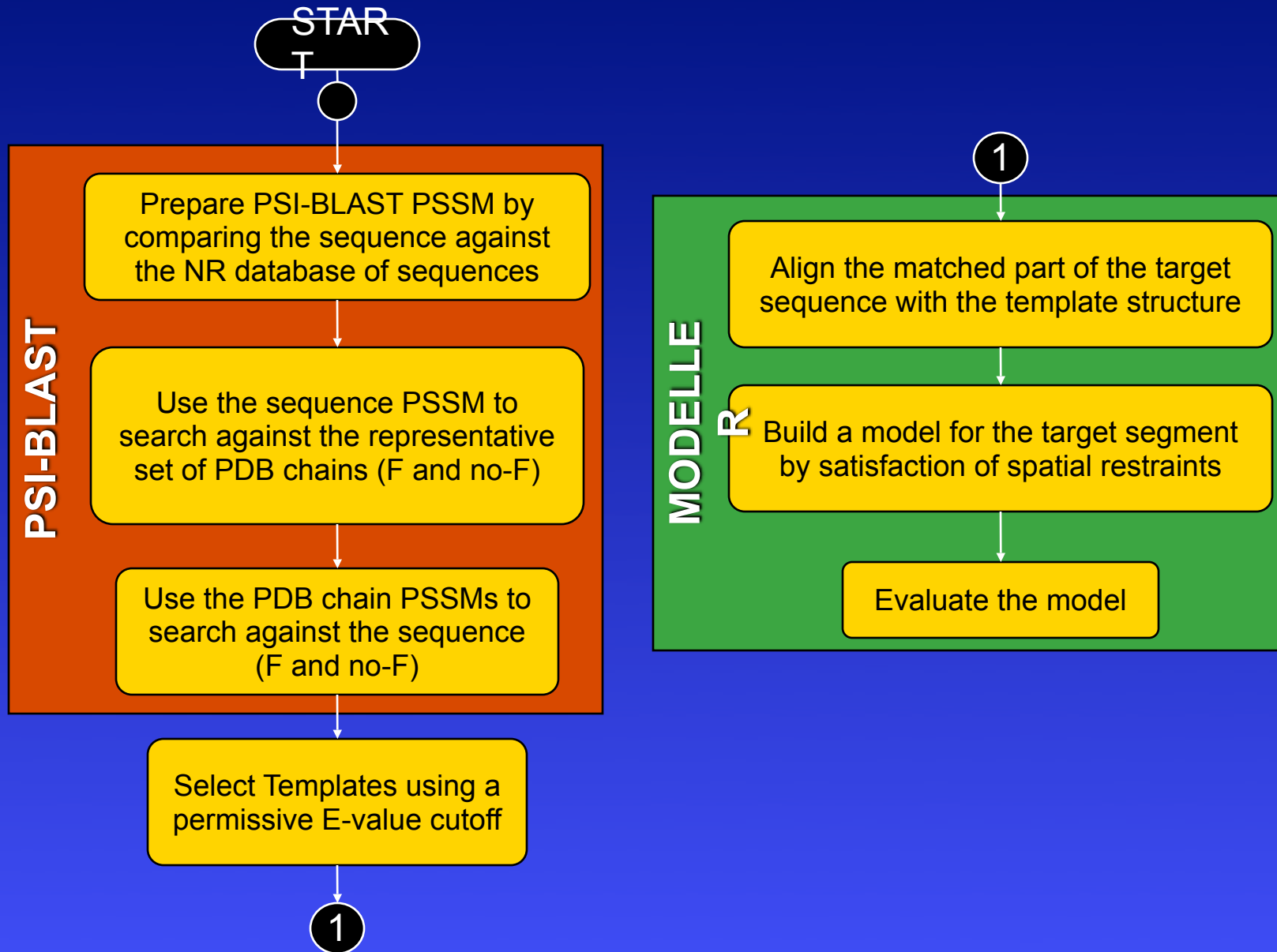
How many structures need to be solved?
Which structures should we solve first?

- Target Amplification

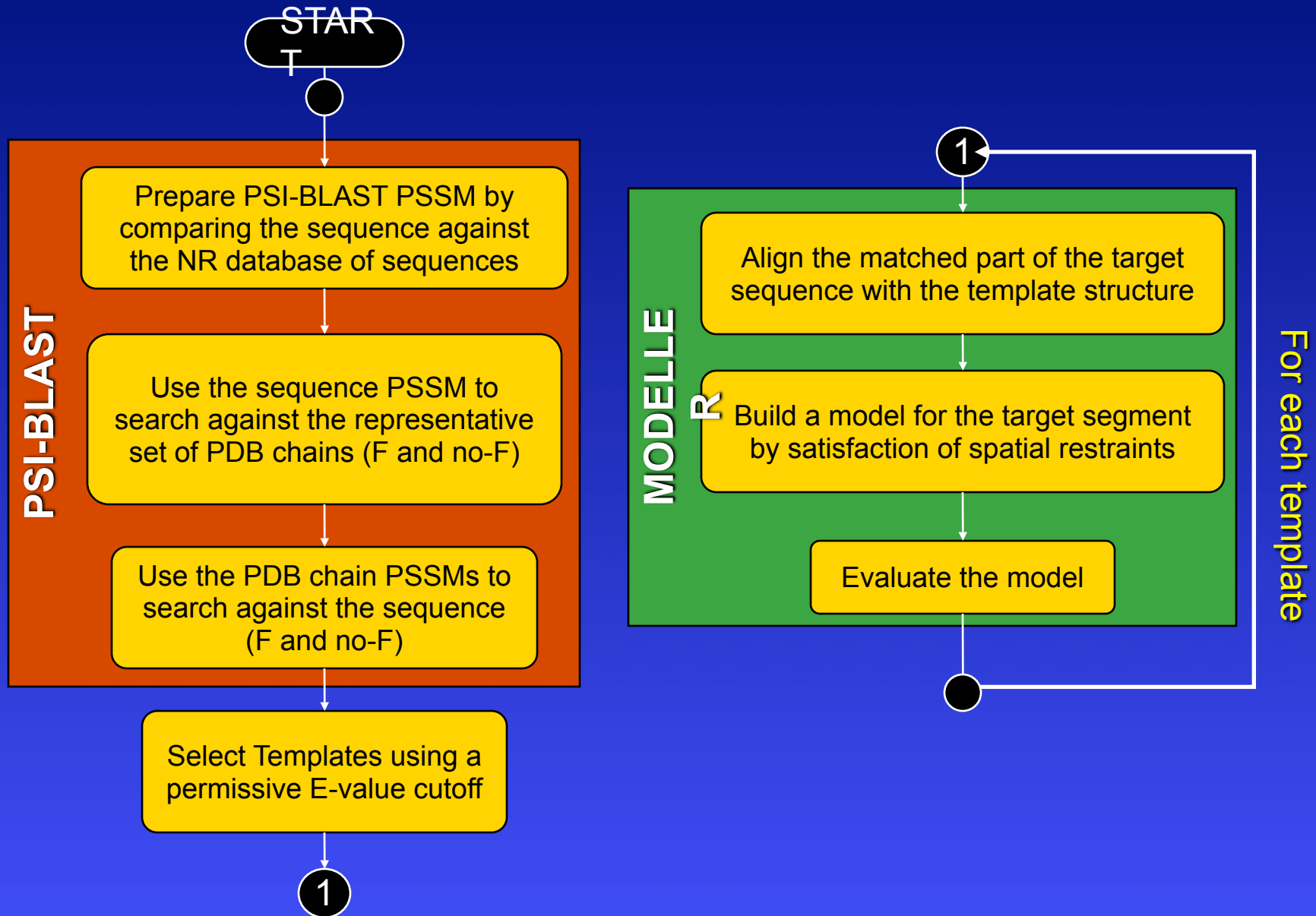
How much of the sequence space is covered by:

- a new structure
- all structures

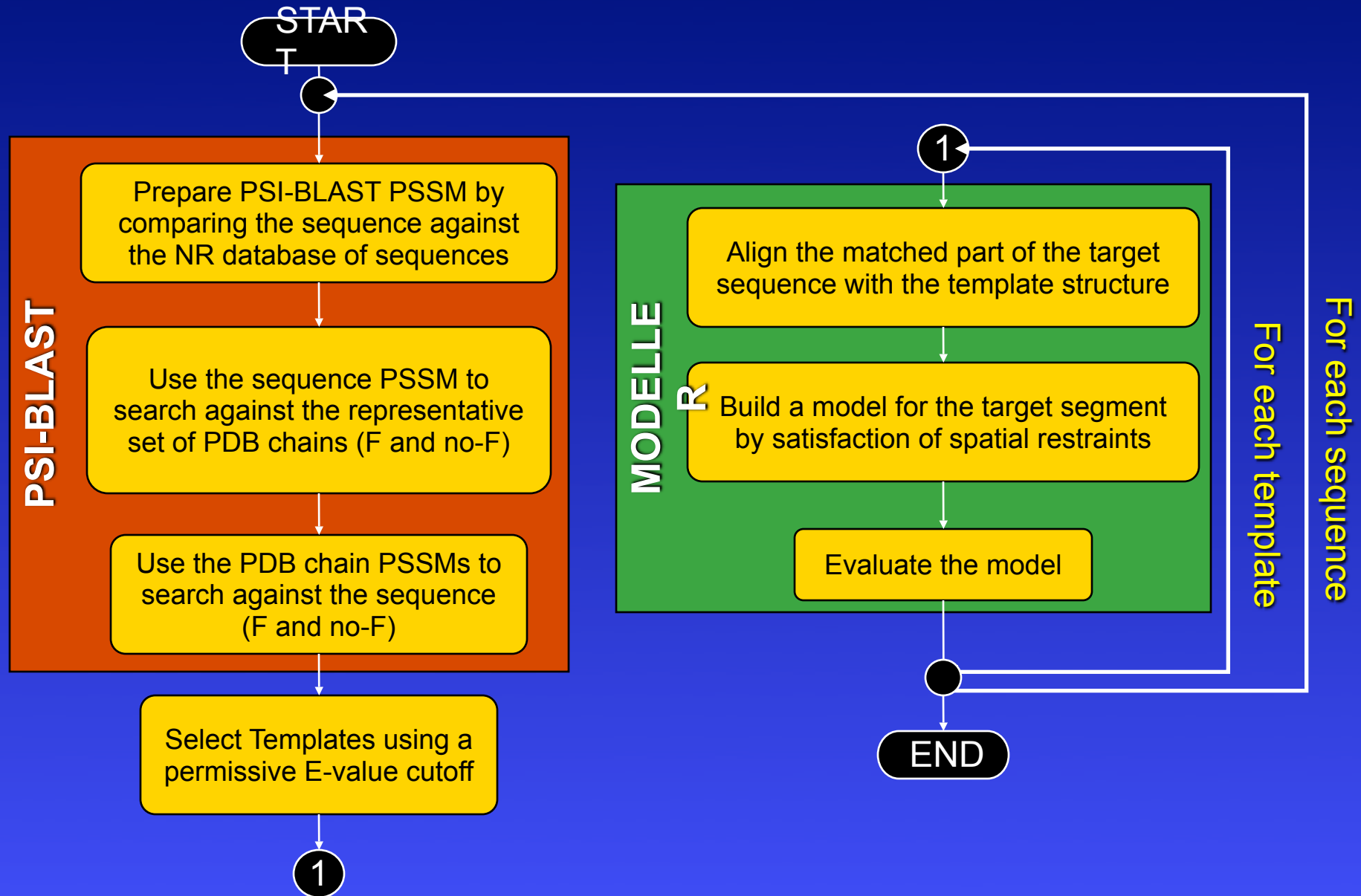
MODPIPE: Large-Scale Comparative Protein Structure Modeling



MODPIPE: Large-Scale Comparative Protein Structure Modeling



MODPIPE: Large-Scale Comparative Protein Structure Modeling



Comparative modeling of the TrEMBL database

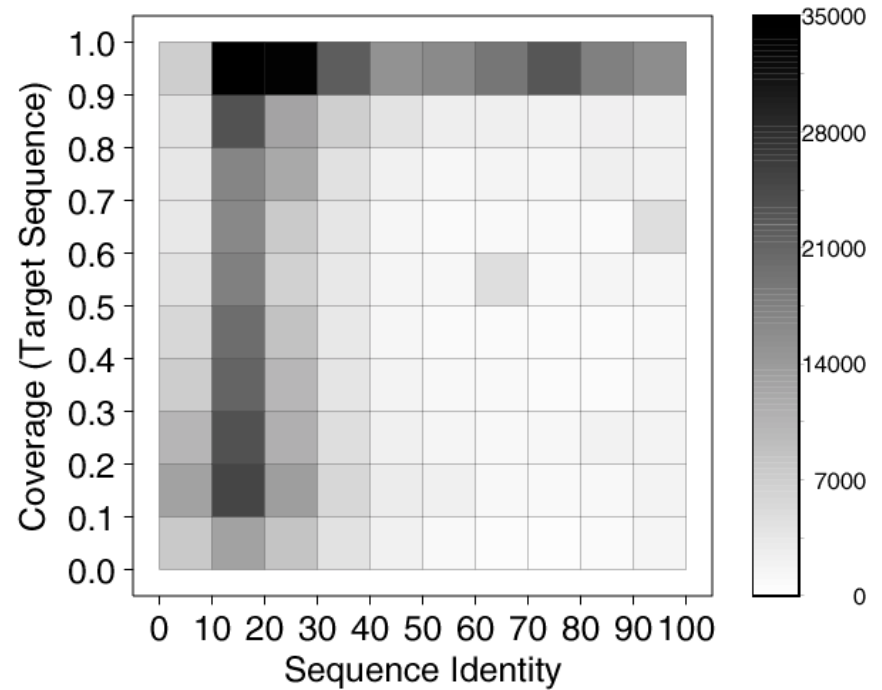
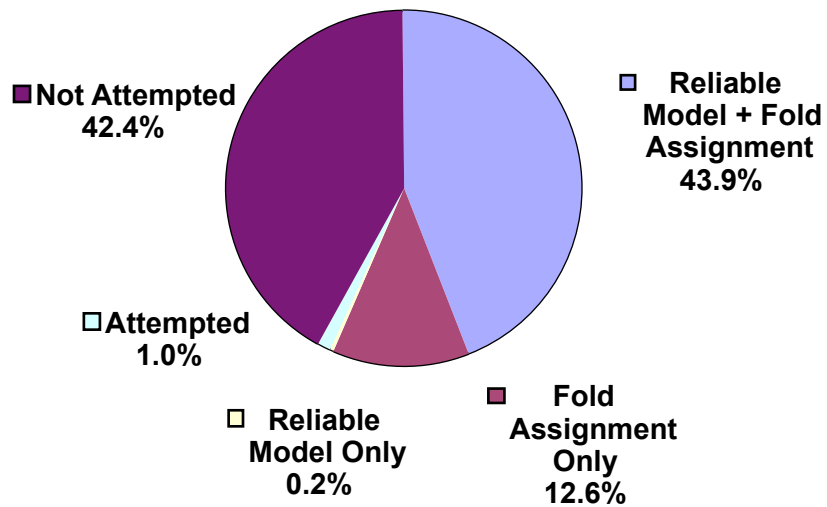
Fold Assignments

- ✓ Reliable fold assignments: **620,370**
- ✓ Sequences with reliable folds: **304,517 (56%)**
- ✓ Average length of queries: **514 amino acids**
- ✓ Average length of folds: **226 amino acids**

Comparative Models

- ✓ Reliable models: **392,397**
- ✓ Sequences with reliable models: **237,143 (44%)**
- ✓ Structures used as templates: **5523 (90%)**

Modeling Coverage Of The Sequence Space



Fold assignment: PSI-BLAST E-value $\leq 1e^{-4}$
Reliable Model: Model Score ≥ 0.7

Organism Statistics

Top 10 organism by number of models

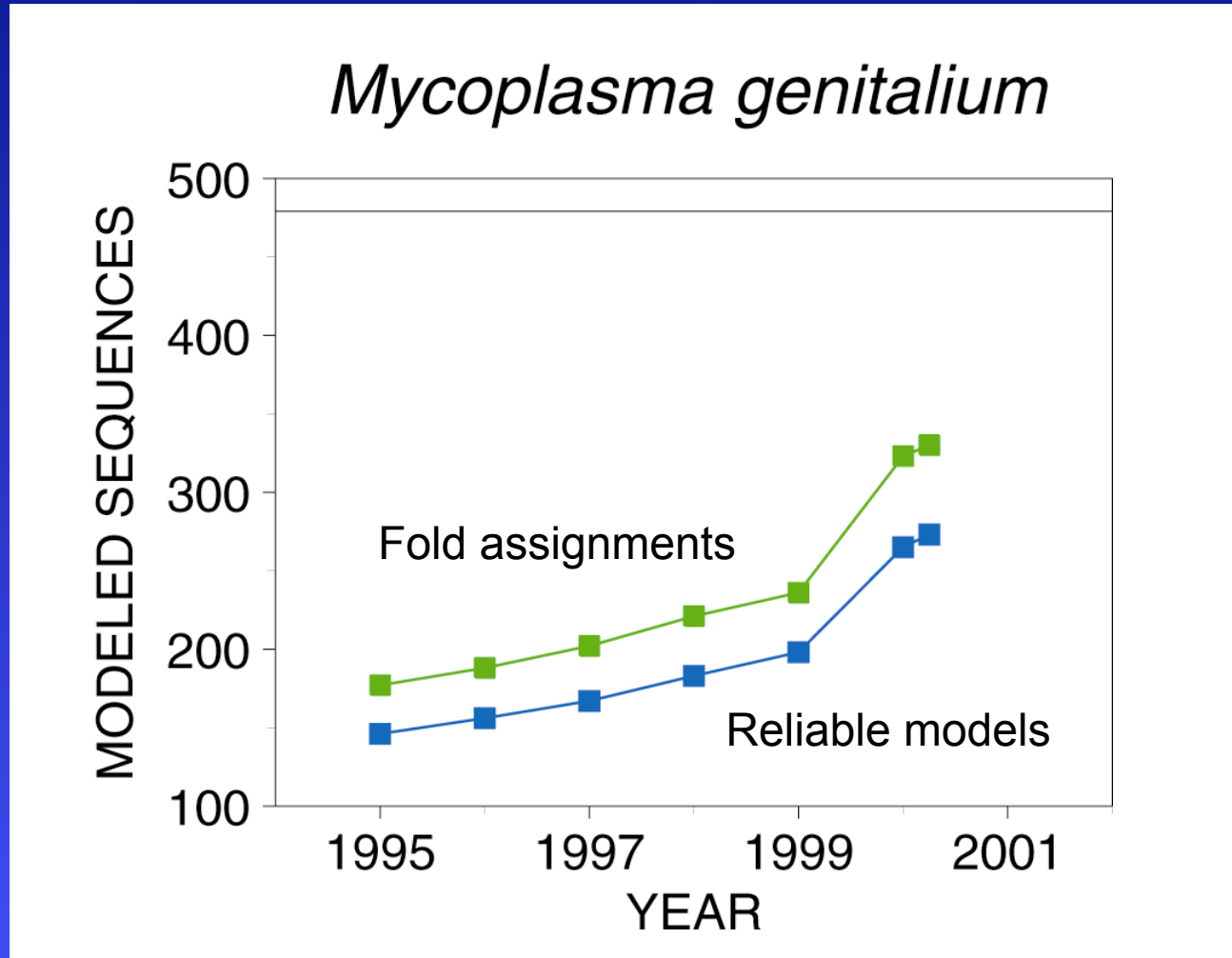
Organism	# sequences	# models	models/ seq#	# CATH folds
<i>Homo sapiens</i>	13,785	37,638	2.73	315
HIV type 1	25,654	33,180	1.29	12
<i>D. melanogaster</i>	8,248	25,314	3.06	299
<i>C. elegans</i>	7,260	20,095	2.76	289
<i>A. thaliana</i>	8,852	18,695	2.11	294
<i>Mus musculus</i>	6,232	17,248	2.76	271
<i>R. norvegicus</i>	3,586	9,299	2.59	246
<i>S. cerevisiae</i>	2,580	5,749	2.22	237
<i>S. Pombe</i>	2,315	4,497	1.94	221
<i>E. coli</i>	2,862	4,333	1.51	259

Organism Statistics

Top 10 organism by number of models

Organism	Avg. seq. length	Avg. model length	Avg. Sequence coverage	“Organism” coverage
<i>Homo sapiens</i>	517	191	0.55	0.36
HIV type 1	165	124	0.84	0.75
<i>D. melanogaster</i>	634	209	0.47	0.32
<i>C. elegans</i>	563	209	0.50	0.37
<i>A. thaliana</i>	480	218	0.55	0.45
<i>Mus musculus</i>	510	191	0.53	0.37
<i>R. norvegicus</i>	511	207	0.57	0.40
<i>S. cerevisiae</i>	590	255	0.55	0.43
<i>S. Pombe</i>	527	247	0.58	0.46
<i>E. coli</i>	367	248	0.75	0.67

Factors affecting coverage: PDB growth



MODBASE

Netscape: ModBase database of Comparative Protein Structure Models

File Edit View Go Communicator

Back Forward Reload Home Search Netscape Print Security Stop

Bookmarks Location: <http://pipe.rockefeller.edu/modbase/>

WebMail Contact People Yellow Pages Download Find Sites Channels

MODBASE

Database of Comparative Protein Structure Models

Welcome to MODBASE, a database of three-dimensional protein models calculated by **comparative modeling**.

About MODBASE

[General Information](#)

[Glossary](#)

[Authors and acknowledgements](#)

[Publications](#)

[Related resources](#)

Users of ModBase are requested to cite this article in their publications:
 ModBase: a database of annotated comparative protein structure models,
 R. Sánchez, U. Pieper, N. Melnik, F. J. D. Abad, E. Vittinger & A. Sá
 Nat. Acad. Sci. 98, 220-225, 2001.

Preliminary Use of ModBase for commercial purposes is not allowed.

ModBase is maintained by Uriah Pieper and Roberto Sánchez in the group of [Agustín Sáiz](#), Laboratories of Molecular Biophysics, Pab. Pinsky Center for Biochemistry and Structural Biology, The Rockefeller University, 1230 York Ave, New York, NY 10021. Please address all inquiries to modbase@pi.rockefeller.edu.

[The Rockefeller University](#)

MODBASE Contents

22015 **Reliable models** for domains in 2194 proteins.
 39096 **Reliable fold assignments** for domains in 2730 proteins.
 Last Update: Feb-10-2000.
[ModBase statistics and sequence sources.](#)

Search

Enter a SwissProt Identifier or Accession number, a GenBank GI, a PDB code or a Keyword:

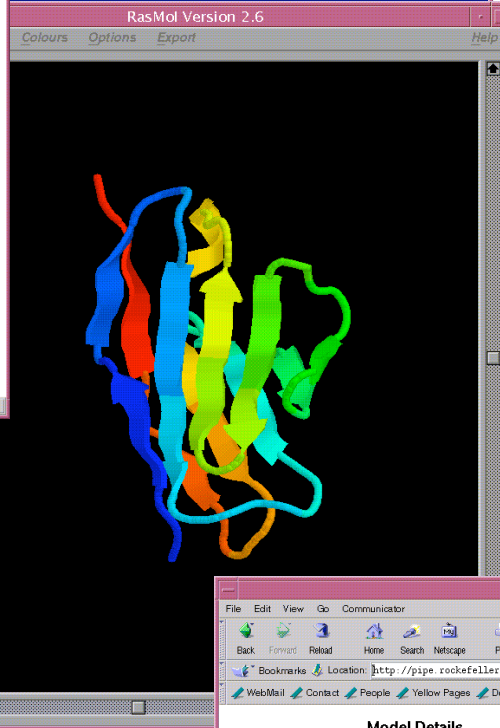
[Search for models by keywords and model properties.](#)
[Search for models by sequence similarity.](#)

Notes

MODBASE contains theoretically calculated models, which may contain **significant errors**, not experimentally determined structures.

MODBASE is undergoing a major transition, consequently only a subset of the models is currently available at this location.

[Link to the original MODBASE](#) for some additional models.



Netscape: ModBase: Model Search Form

File Edit View Go Communicator

Back Forward Reload Home Search Netscape Print Security Stop

Bookmarks Location: http://pipe.rockefeller.edu/modbase/cgi/search_form.cgi

WebMail Contact People Yellow Pages Download Find Sites Channels

Database of Comparative Protein Structure Models

KEYWORD & PROPERTIES SEARCH
 Go to Sequence Similarity Search

HELP

Search all

Category

Maximum number of hits

Sort output by

Model Size (residues) lower limit upper limit

Model Score

Sequence identity %

E-value

Netscape: ModBase: Search Results

File Edit View Go Communicator

Back Forward Reload Home Search Netscape Print Security Stop

Bookmarks Location: http://pipe.rockefeller.edu/modbase/cgi/model_search.cgi

WebMail Contact People Yellow Pages Download Find Sites Channels

Keyword Search

Search Criteria	Keywords	Values
Organism	ALL	Minimum Average Maximum
Model size range (residues)	-	63 223 424
Model score range	0.7-1.0	0.91 0.99 1.00
Sequence identity range (%)	-	16 30 77
E-value range	-	1.0e-06 5.0e-14 5.0e-13

More than 10 matches were found. Click on the links in the table header to resort your output.

TARGET SEQUENCE				MODEL DATA				TEMPLATE STRUCTURE						
Link to the model	Link to all reliable models	Selected Sequence Database Links	SwissProt/TrEMBL/TrEMBL-New Description	Organism	Protein Size	Modelled Segment	Size	Model Score	Seq. Id (%)	E-value	PDB code	Segment	Description	
Link	Link	Link	• SR 026514 • GI 4550394 • SR P33071 • GI 1137024	CYTOCHROME P450 71B7 (EC 1.14.~.~) PF319.13 protein	A. thaliana	498	121-443	323	1.00	22	5e-13	1ro0	95-356	CYTOCHROME P450
Link	Link	Link	• SR P33071 • GI 1137024	CALMODULIN-RELATED PROTEIN 5, TOUCH-INDUCED	A. thaliana	324	91-255	165	1.00	59	1e-64	1o3a	1-148	CALMODULIN 105A.3
Link	Link	Link	• GI 1492706	Calmodulin	A. thaliana	148	5-148	144	1.00	24	7e-50	2ncp	1-157	SARCOPLASMIC CALCIUM-BINDING PROTEIN 25CP.3
Link	Link	Link	• SR 028902 • GI 3803661	RAC-LIKE GTP BINDING PROTEIN ARAC1	A. thaliana	197	4-177	174	1.00	36	2e-23	5p2l	1-163	2C-M-RAS 2P21 PROTEIN (AMINO ACIDS 1-166)
Link	Link	Link	• SR 025199 • GI 166441	rac-related small GTP-binding protein	A. thaliana	439	346-408	63	1.00	38	4e-19	1aa3	268-330	REGA C-TERMINAL DOMAIN
Link	Link	Link	• SR P33133 • GI 1762437	DNA REPAIR PROTEIN RECA 1 PRECURSOR	A. thaliana	148	2-147	146	1.00	77	6e-73	2o0c	3-148	UBIQUITIN-CONJUGATING ENZYME SUCE 1
Link	Link	Link	• SR 042392	UBIQUITIN-CONJUGATING ENZYME E2-17 KD 10 (EC 6.3.2.19) UBIQUITIN-PROTEIN LIGASE 10 UBIQUITIN CARRIER PROTEIN 10	A. thaliana	22	22-22	22	1.00	100	1e-100	1a01	22-22	22-DALKYLYLCINE DECARBOXYLASE (PYRUVATE)

Netscape: ModBase : Model Details

File Edit View Go Communicator

Back Forward Reload Home Search Netscape Print Security Stop

Bookmarks Location: http://pipe.rockefeller.edu/modbase/cgi/model_page.cgi?orf=category=A&thaliana&seq=cb135fc80a069816aded8063dc0d1

WebMail Contact People Yellow Pages Download Find Sites Channels

Model Details

Categories:

- Organism: A. thaliana

Models in the Model Score Range: 0.7-1.0

Database Synonyms for this Sequence

Database	Accession	Description
SwissProt	Q06514	CYTOCHROME P450 71B7 (EC 1.14.~.~)
SwissProt	Q05794	CYTOCHROME P450 71B1 (EC 1.14.~.~)
GenBank	4950394	-
GenBank	1523796	-
GenBank	3184132	-
TrEMBL-New	AA031064	PF319.13 protein

MODEL DATA				LINKS			
Size	Model Score	Seq. Id (%)	E-value	Model	Template	Sequence	Structure
323	1.00	22	5e-13	R	C	A	I

(498 residues)

121 443

323 1.00 22 5e-13 R C A I

[1ro0](#) (95-356) CYTOCHROME P450
 CATH: no information available

30 498

460 1.00 22 6e-36 R C A I

[1aa3](#) (6-455) CYTOCHROME P450 HEME DOMAIN
 CATH: no information available

30 465

436 1.00 23 5e-52 R C A I

[1aa3](#) (6-426) CYTOCHROME P450 HEME DOMAIN
 CATH: no information available

121 443

323 1.00 10 1e-12 R C A I

[1a01](#) (111-381) CYTOCHROME P450-TERP 1CPT.3
 CATH: [3.90.60.10.1.1 (32%)] [1.10.63.0.10.1.1 (75%)]

<http://pipe.rockefeller.edu/modbase/>

Conclusions

Conclusions

- ✓ Comparative models help to understand protein's function:
 - ✓ Detecting remote structural (functional?) relationships.
 - ✓ Revealing features that are not present in the templates.
 - ✓ Revealing features that are not recognizable from the sequence.

Conclusions

- ✓ Comparative models help to understand protein's function:
 - ✓ Detecting remote structural (functional?) relationships.
 - ✓ Revealing features that are not present in the templates.
 - ✓ Revealing features that are not recognizable from the sequence.

Conclusions

- ✓ Comparative models help to understand protein's function:
 - ✓ Detecting remote structural (functional?) relationships.
 - ✓ Revealing features that are not present in the templates.
 - ✓ Revealing features that are not recognizable from the sequence.
- ✓ Currently, useful 3D models can be obtained for domains in approximately 50% of the proteins (30% of domains), because of the improved **methods** and because of the many **known protein structures** and **sequences**.

Conclusions

- ✓ Comparative models help to understand protein's function:
 - ✓ Detecting remote structural (functional?) relationships.
 - ✓ Revealing features that are not present in the templates.
 - ✓ Revealing features that are not recognizable from the sequence.
- ✓ Currently, useful 3D models can be obtained for domains in approximately 50% of the proteins (30% of domains), because of the improved **methods** and because of the many **known protein structures** and **sequences**.

Conclusions

- ✓ Comparative models help to understand protein's function:
 - ✓ Detecting remote structural (functional?) relationships.
 - ✓ Revealing features that are not present in the templates.
 - ✓ Revealing features that are not recognizable from the sequence.
- ✓ Currently, useful 3D models can be obtained for domains in approximately 50% of the proteins (30% of domains), because of the improved **methods** and because of the many **known protein structures** and **sequences**.
- ✓ We will be able to calculate useful models for most globular domains soon after the completion of the genome projects, because of **structural genomics**.

Acknowledgments



Burroughs Wellcome Fund

Andrej Šali

Frank Alber

Fred Davis

Narayanan Eswar

András Fiser

Valentin Ilyin

Bozidar Jerković

Bino John

M. S. Madhusudhan

Linda McMahan

Nebojša Mirković

Ursula Pieper

Andrea Rossi

<http://guitar.rockefeller.edu>