Protein Structure Modeling for Structural Genomics

Marc A. Marti-Renom

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

Summary

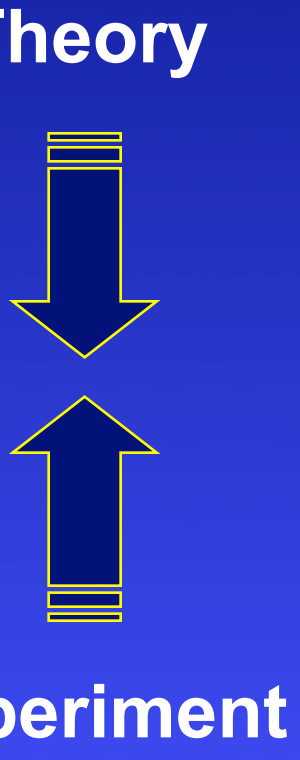
- ✓ Comparative Modeling
- ✓ Alignment problem
- ✓ Modeling genes
- ✓ Modeling genomes and structural genomics

Why protein structure prediction?

	Y 2002	Y 2005
Sequences	700,000	millions
Structures	17,000	50,000

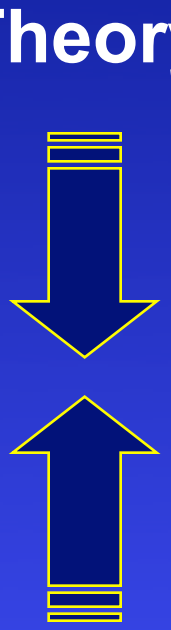
Why protein structure prediction?

	Y 2002
Sequences	700,000
Structures	17,000



Why protein structure prediction?

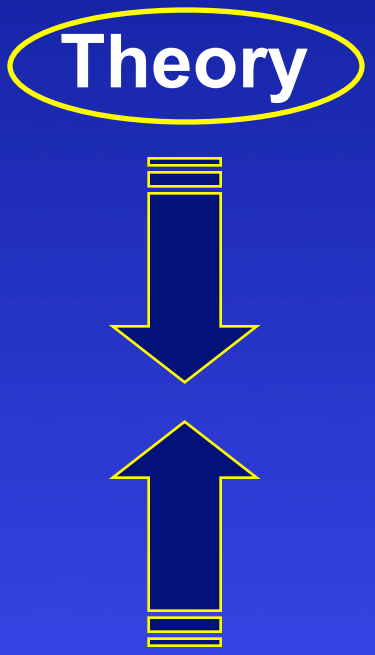
	Y 2002
Sequences	700,000
Structures	400,000



http://guitar.rockefeller.edu/modbase/

Why protein structure prediction?

	Y 2002
Sequences	700,000
Structures	400,000

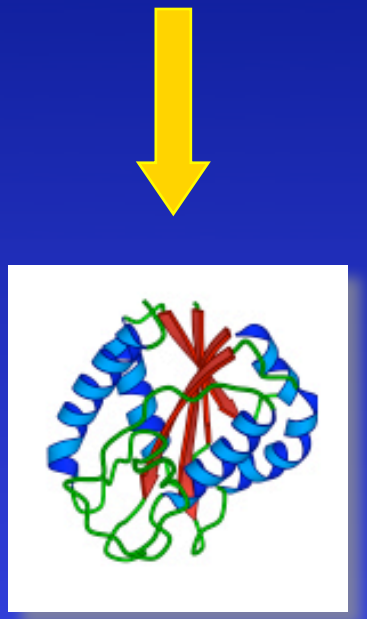


http://guitar.rockefeller.edu/modbase/

Principles of Protein Structure

Principles of Protein Structure

GFCHIKAYTRLIMVG…

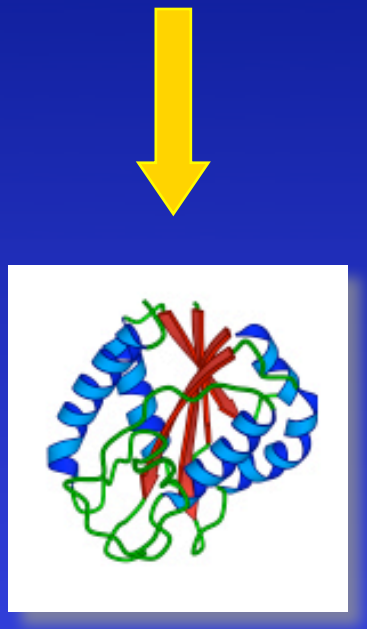


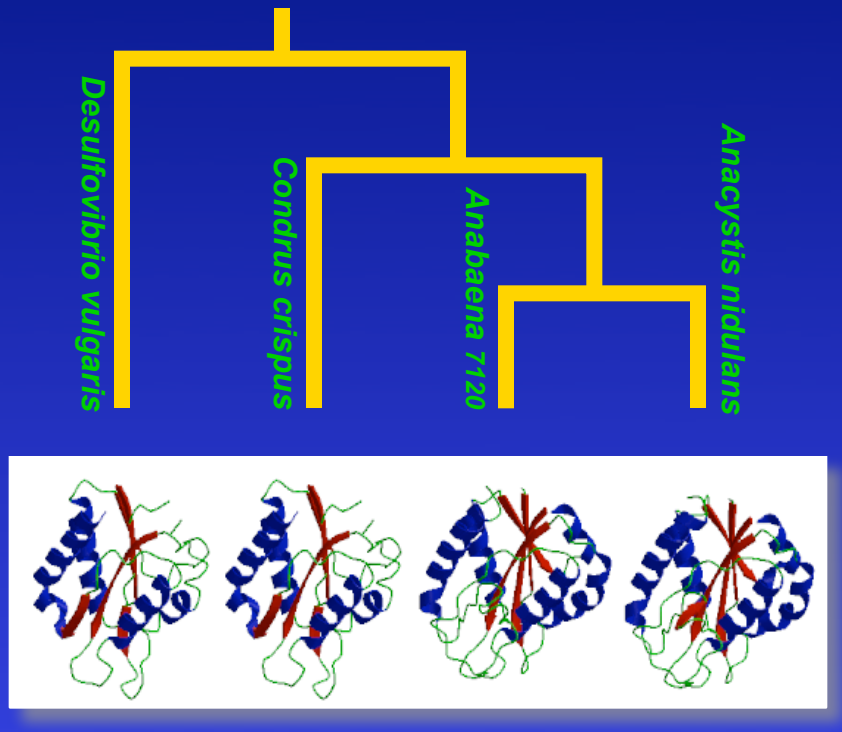
Folding

Ab initio prediction

Principles of Protein Structure

GFCHIKAYTRLIMVG…





Folding

Ab initio prediction

Evolution

Threading
Comparative Modeling

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

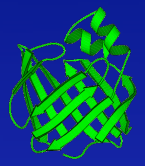
START

Template Search

TARGET

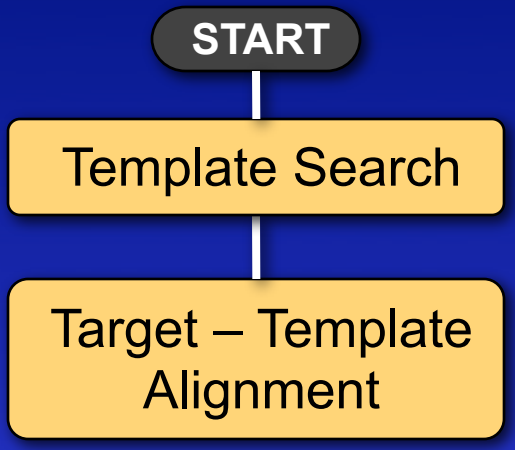
ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCLKIDD
VPERLIPERASFQWMNDK

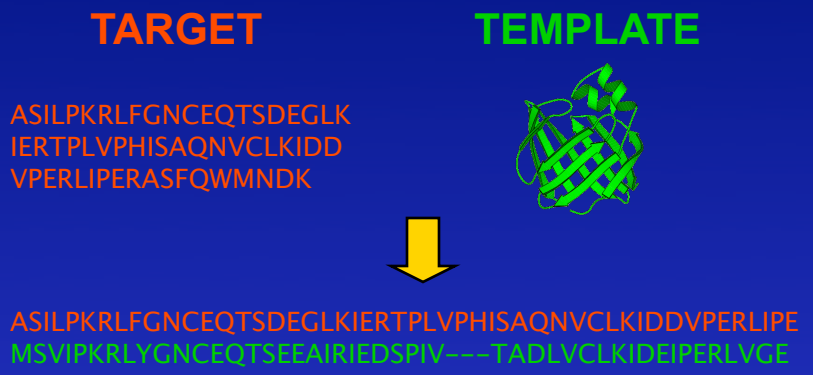
TEMPLATE



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.

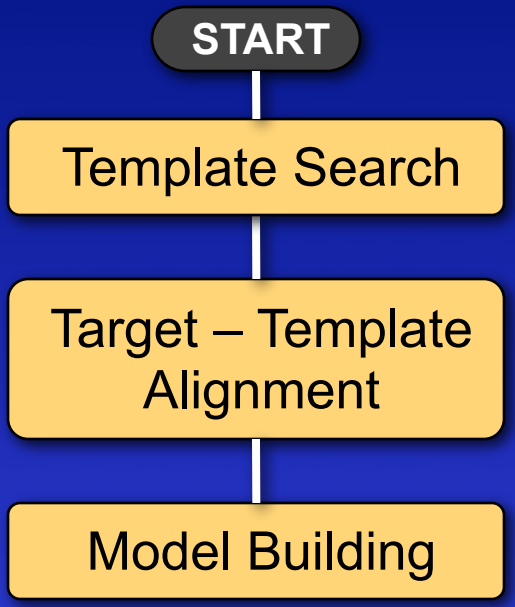
Steps in Comparative Protein Structure Modeling

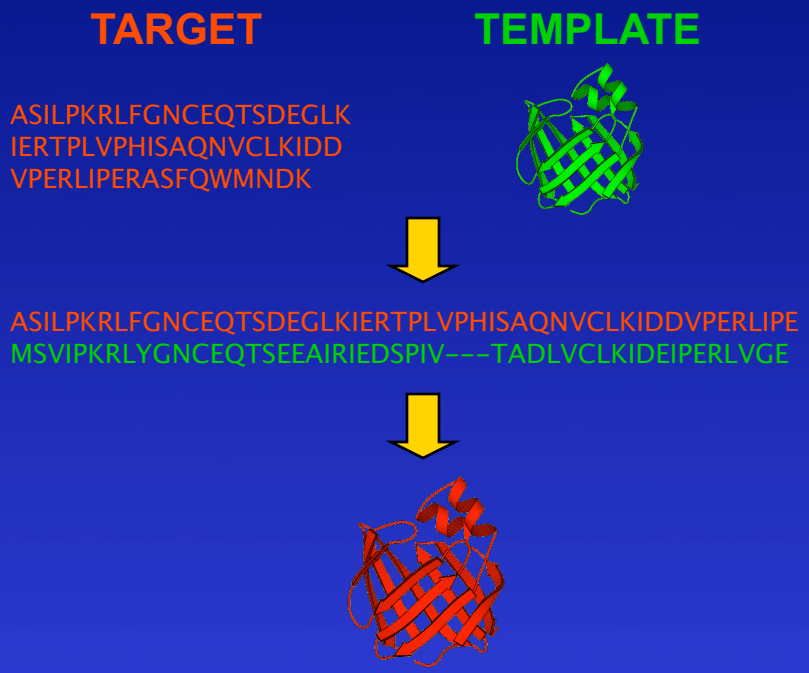




A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.
R. Sánchez & A. Šali, *Curr. Opin. Str. Biol.* 7, 206, 1997.
M. A. Martí-Renom *et al. Ann. Rev. Biophys. Biomolec. Struct.*, 29, 291, 2000.

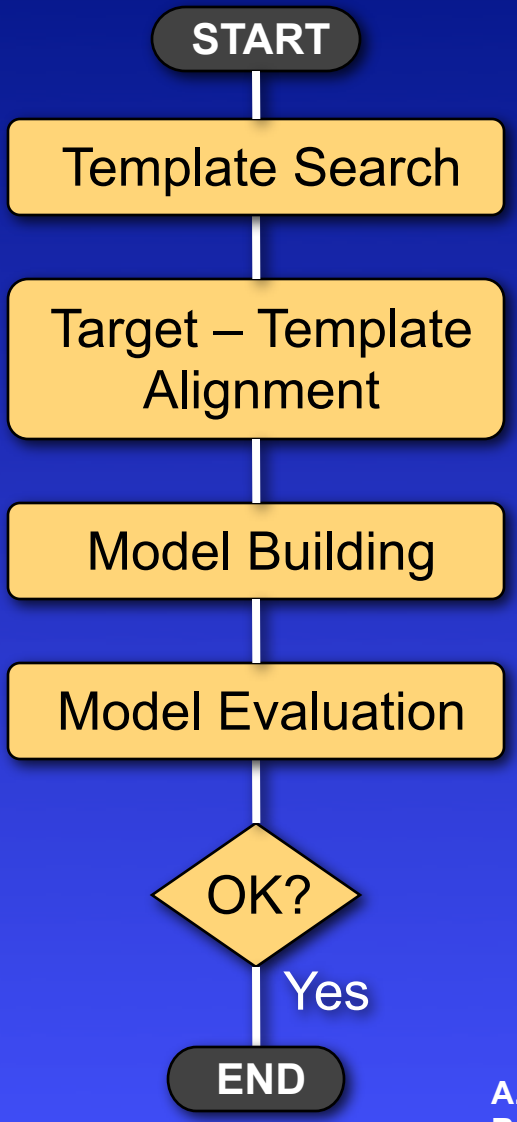
Steps in Comparative Protein Structure Modeling

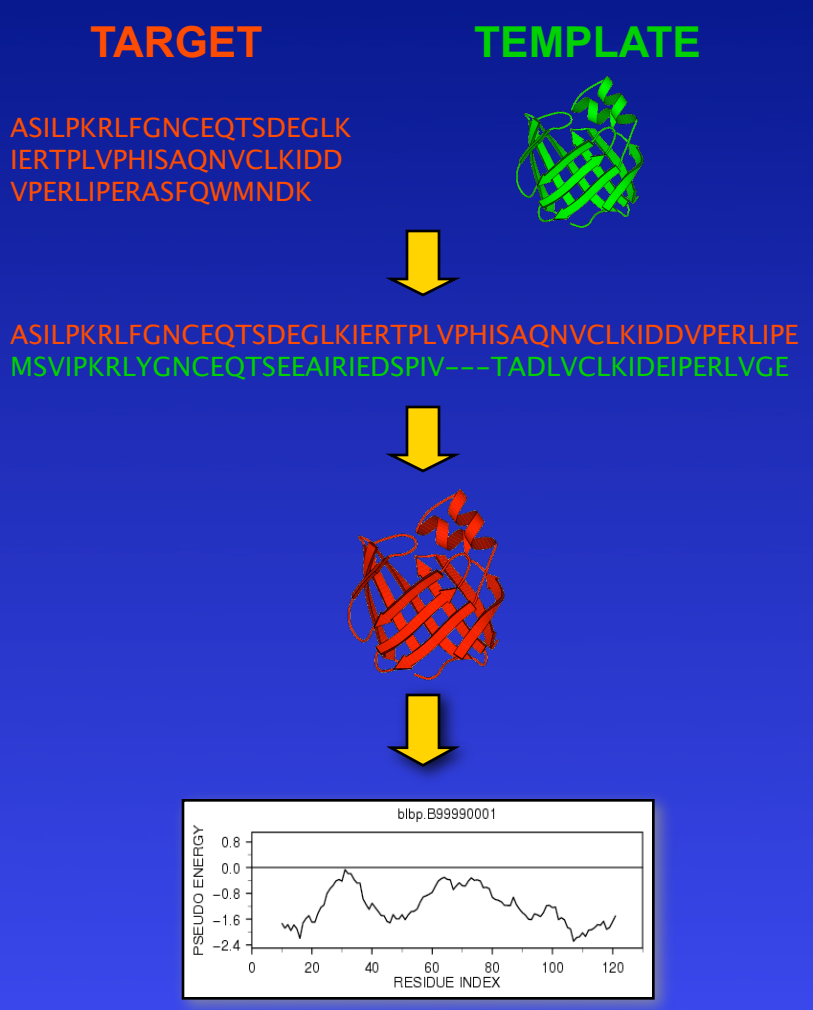




A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.
R. Sánchez & A. Šali, *Curr. Opin. Str. Biol.* 7, 206, 1997.
M. A. Martí-Renom *et al. Ann. Rev. Biophys. Biomolec. Struct.*, 29, 291, 2000.

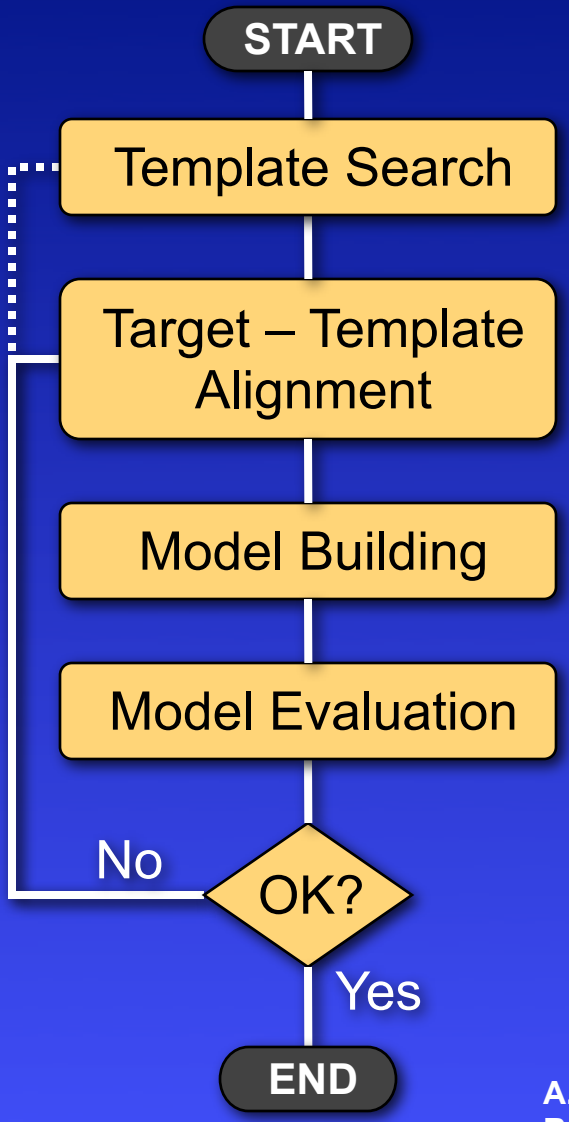
Steps in Comparative Protein Structure Modeling

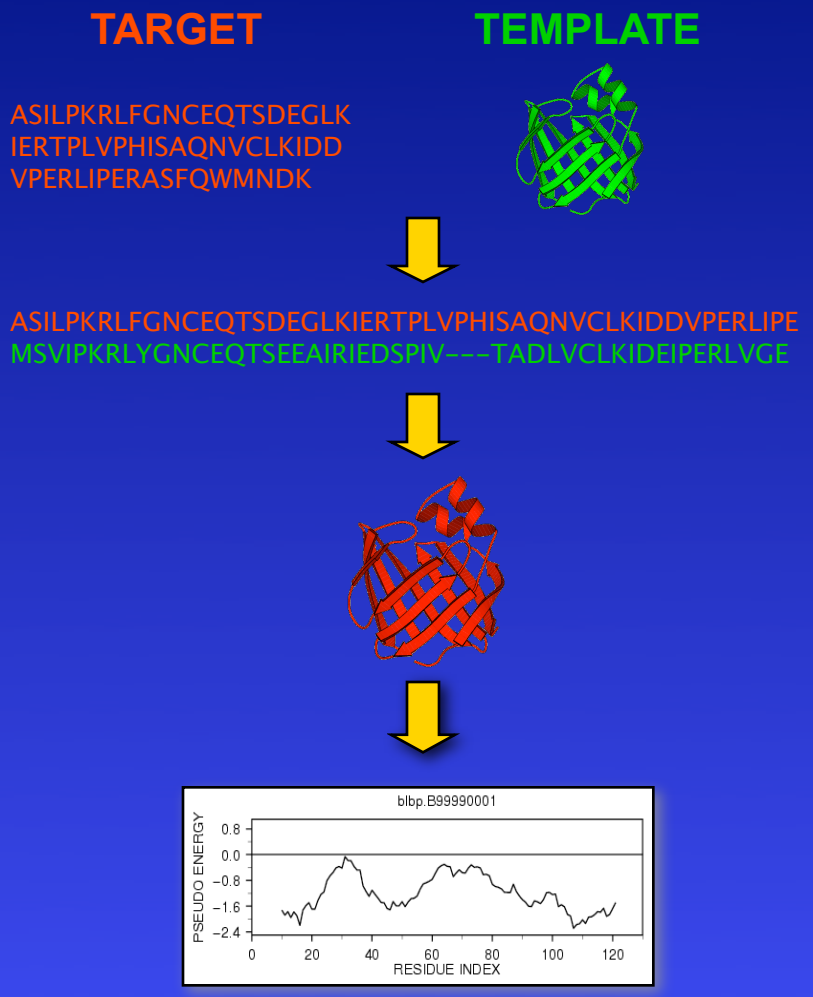




A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.
R. Sánchez & A. Šali, *Curr. Opin. Str. Biol.* 7, 206, 1997.
M. A. Martí-Renom *et al. Ann. Rev. Biophys. Biomolec. Struct.*, 29, 291, 2000.

Steps in Comparative Protein Structure Modeling





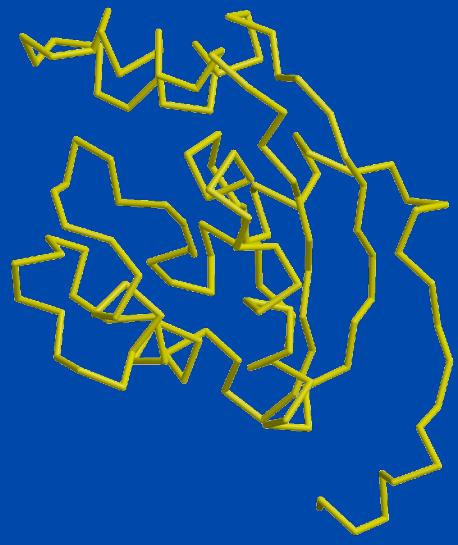
A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.
R. Sánchez & A. Šali, *Curr. Opin. Str. Biol.* 7, 206, 1997.
M. A. Martí-Renom *et al. Ann. Rev. Biophys. Biomolec. Struct.*, 29, 291, 2000.

Model Accuracy

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

HIGH ACCURACY

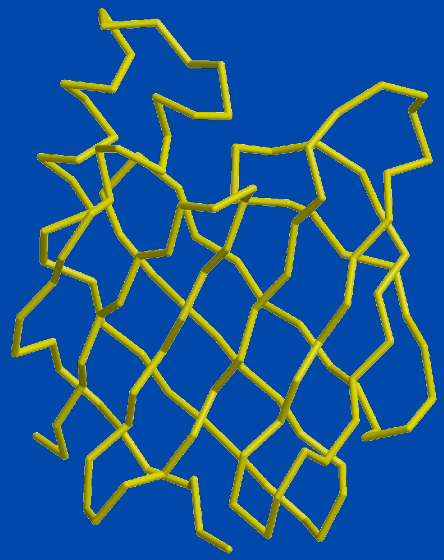
NM23
Seq id 77%



X-RAY

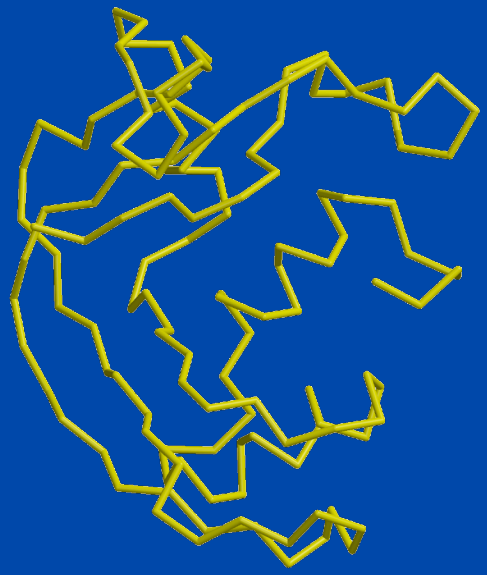
MEDIUM ACCURACY

CRABP
Seq id 41%



LOW ACCURACY

EDN
Seq id 33%

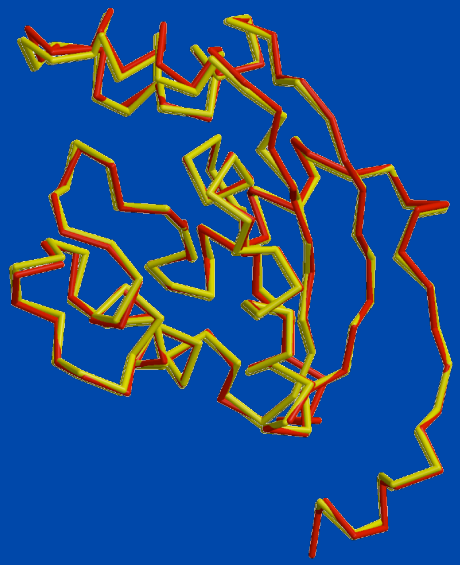


Model Accuracy

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

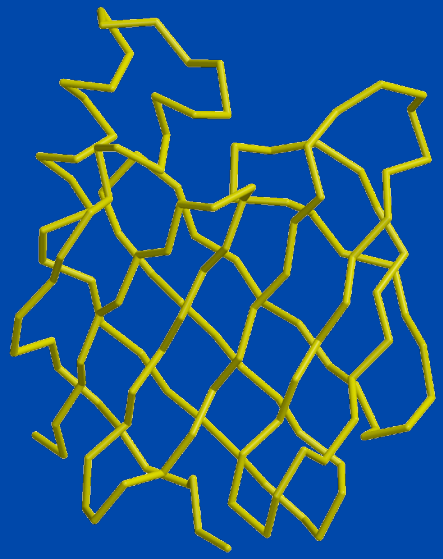


Sidechains
Core backbone
Loops

X-RAY / MODEL

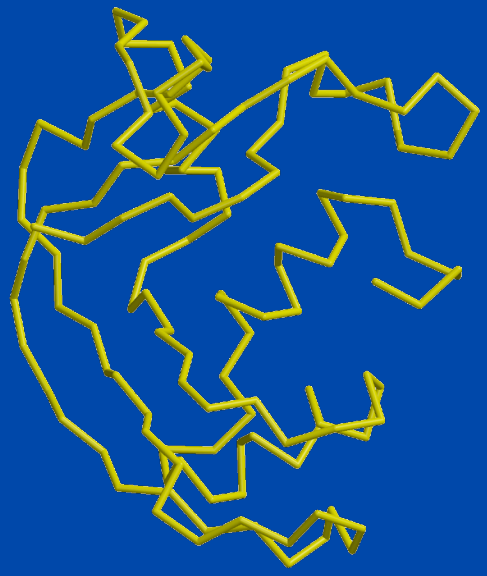
MEDIUM ACCURACY

CRABP
Seq id 41%



LOW ACCURACY

EDN
Seq id 33%

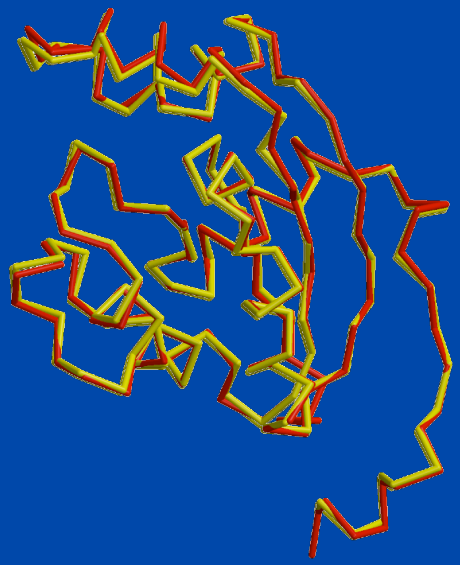


Model Accuracy

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

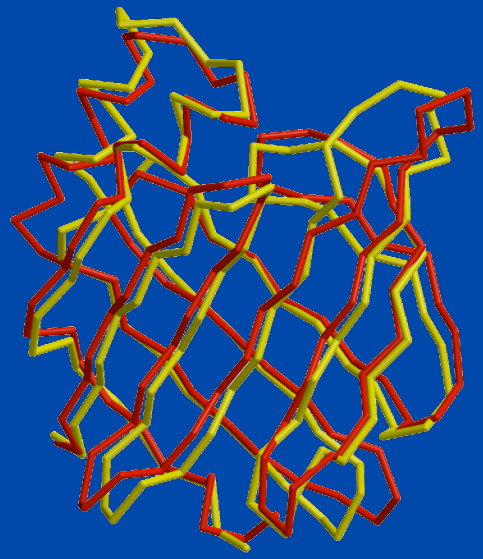


Sidechains
Core backbone
Loops

X-RAY / MODEL

MEDIUM ACCURACY

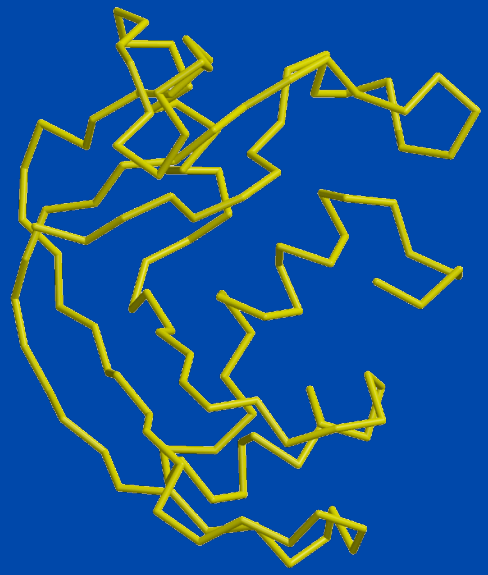
CRABP
Seq id 41%
Cα equiv 122/137
RMSD 1.34Å



Sidechains
Core backbone
Loops
Alignment

LOW ACCURACY

EDN
Seq id 33%

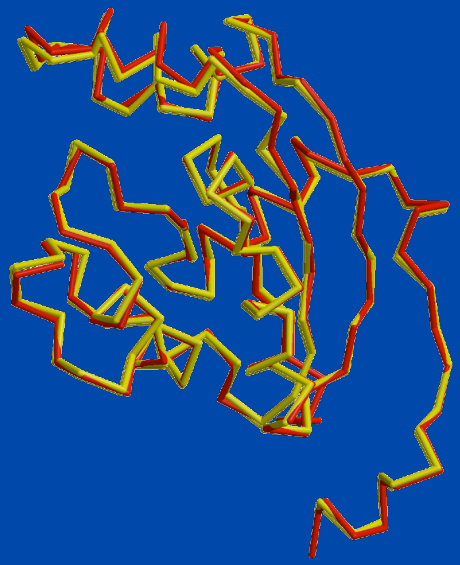


Model Accuracy

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

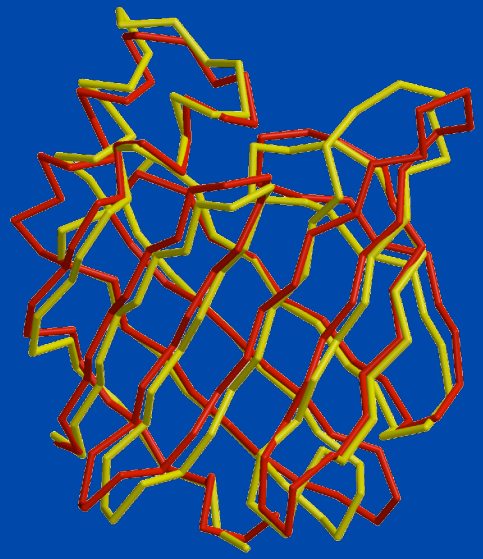


Sidechains
Core backbone
Loops

X-RAY / MODEL

MEDIUM ACCURACY

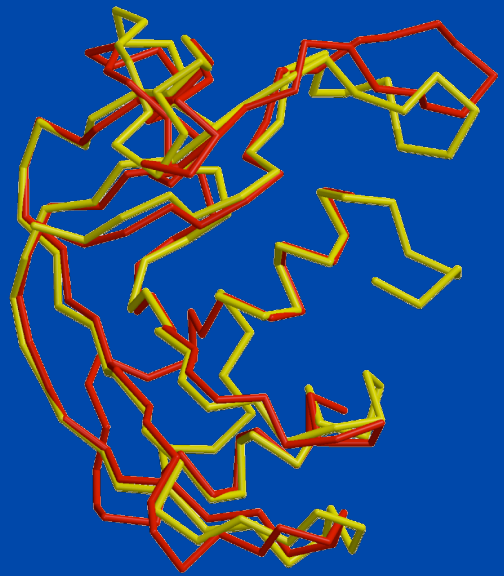
CRABP
Seq id 41%
Cα equiv 122/137
RMSD 1.34Å



Sidechains
Core backbone
Loops
Alignment

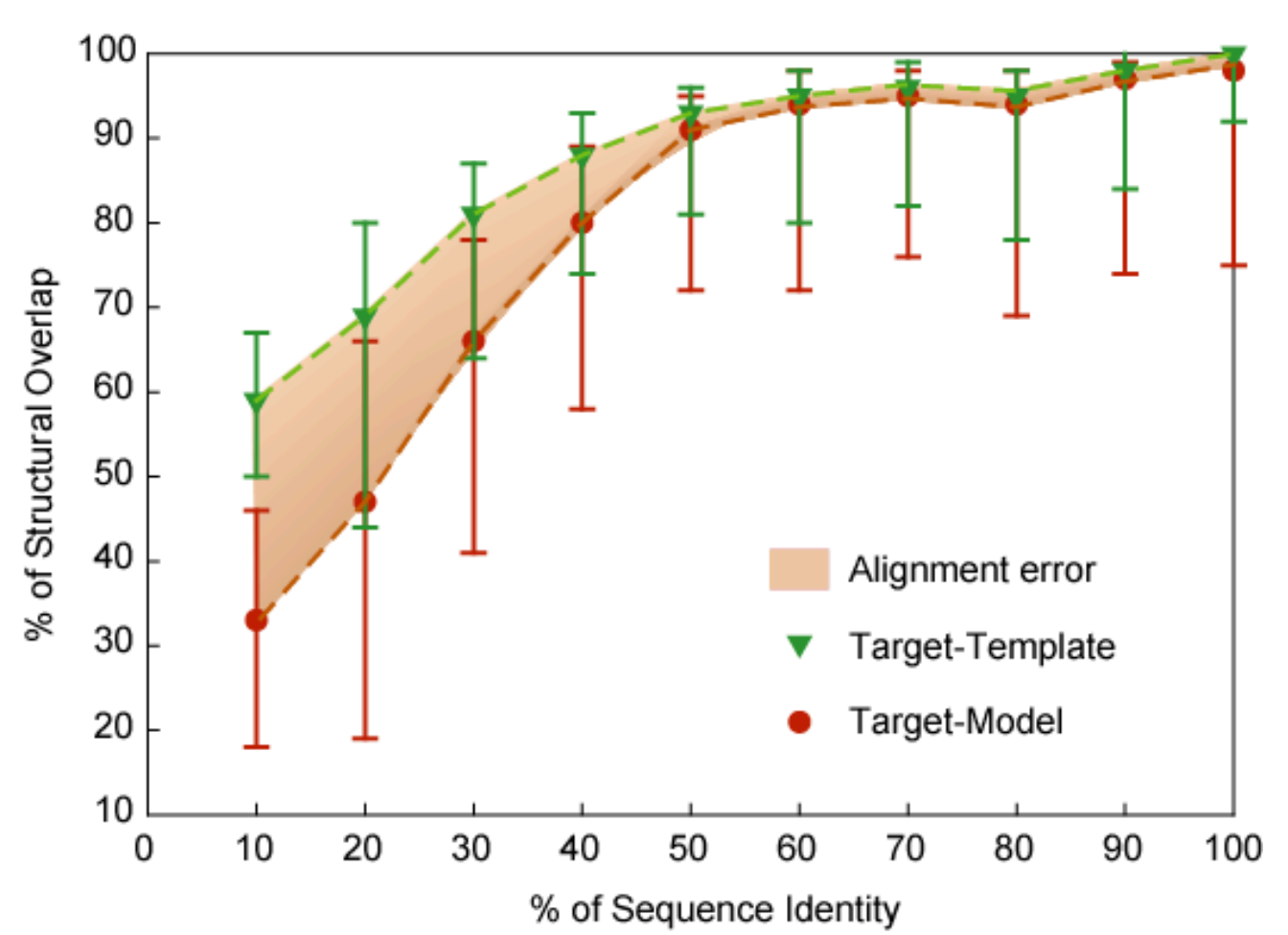
LOW ACCURACY

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



Sidechains
Core backbone
Loops
Alignment
Fold assignment

Model Accuracy as a Function of Target-Template Sequence Identity



Alignment problem: Methods

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

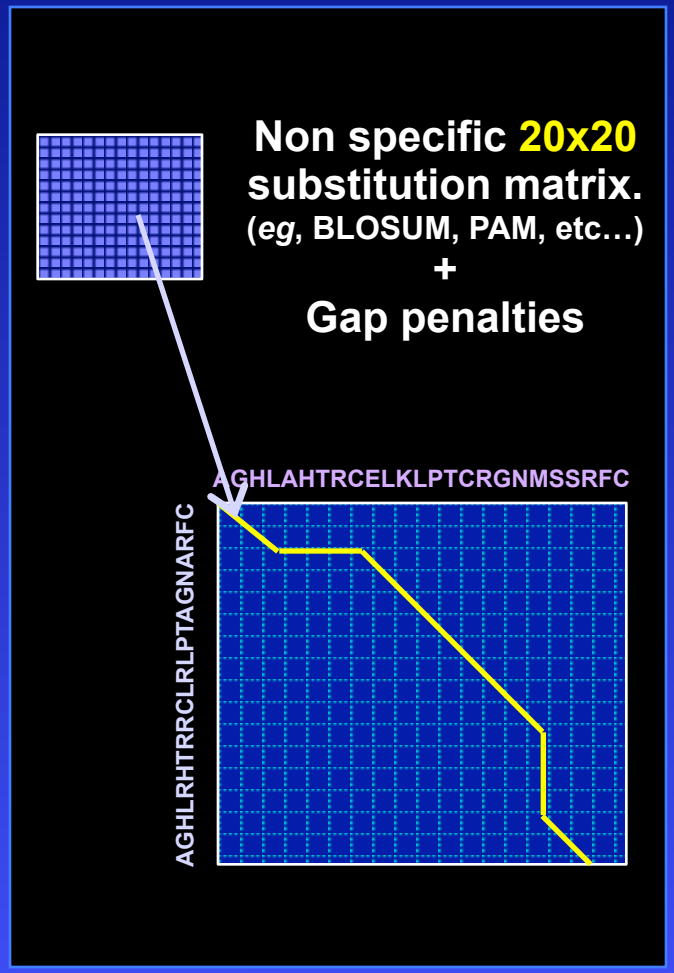
BLAST2SEQ: Local method

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.



Alignment problem: Methods

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

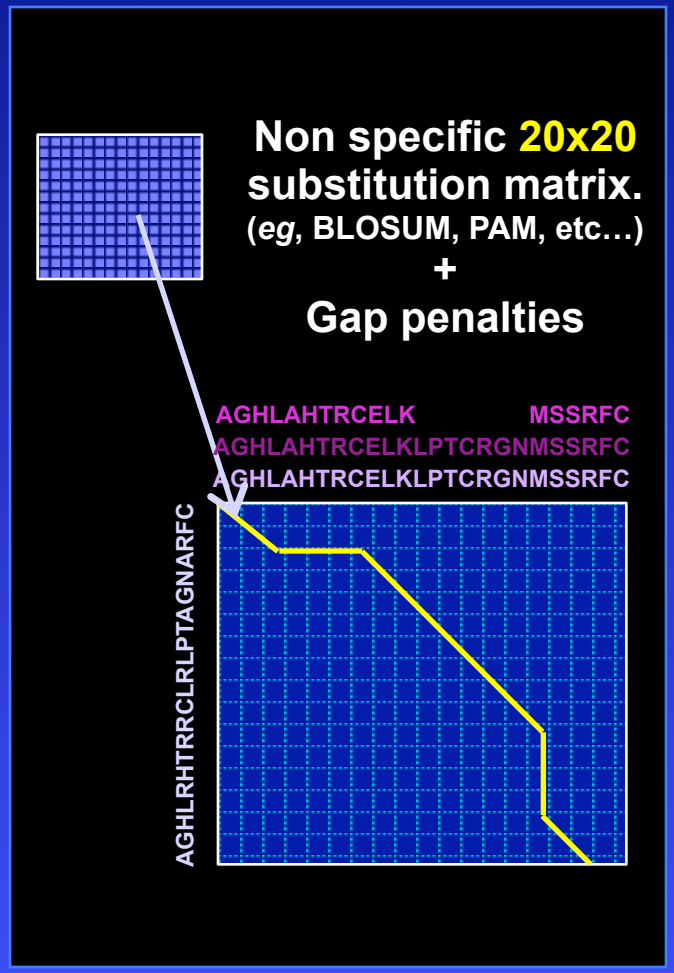
BLAST2SEQ: Local method

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.



Alignment problem: Methods

Sequence A: AGHLAHTRCELKLPTCRGNMSSRFC

Sequence B: AGHLRHTRRCLRLPTAGNARFC

Seq.-Seq.

ALIGN: DP pairwise method

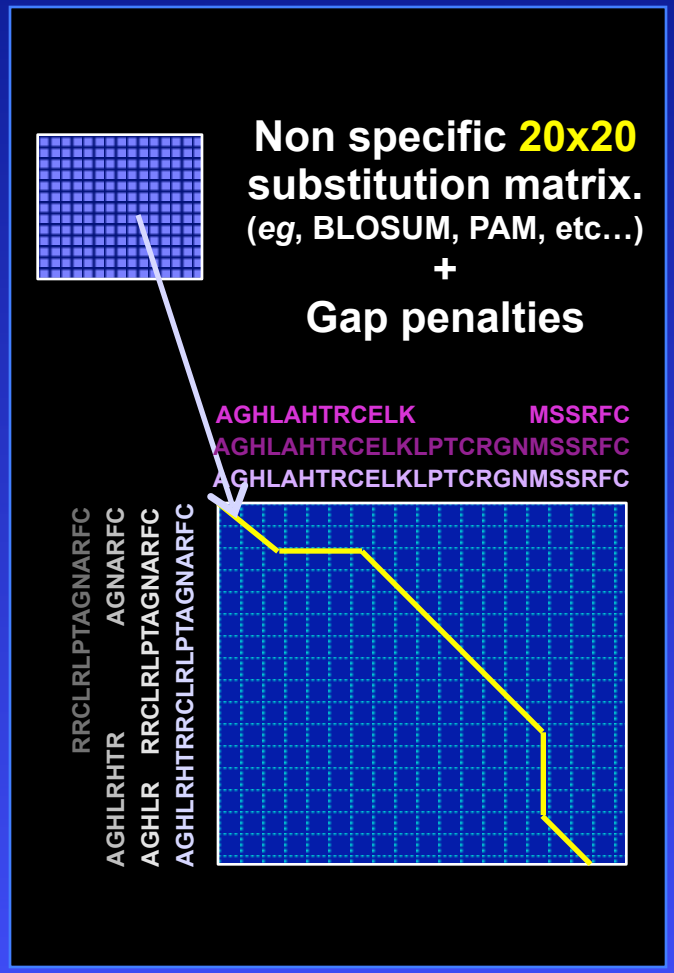
BLAST2SEQ: Local method

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.



Alignment problem

Results: Comparison of alignment dependent measures

Method	% of Correct SeqA	% of Correct SeqB	Shift Score
ALIGN	41.55	41.84	0.44
BLAST2Seq	26.09	26.07	0.32
PB (e-val)	42.95	43.11	0.48
ALIGN4D	55.34	55.49	0.61

Alignment problem

Results: Comparison of success rates

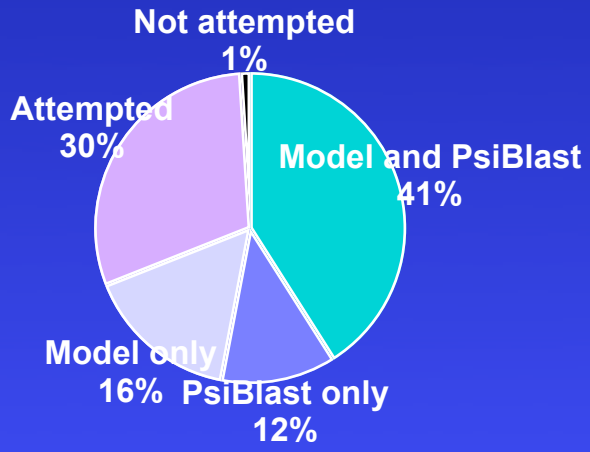
Method	% of alignments at 1Å	% of alignments at 2Å	% of alignments at 3Å	% of alignments at average
CE	20.50	82.50	100.00	82.50
ALIGN	8.50	23.00	35.00	21.00
BLAST2SEQ	8.00	21.50	30.00	20.00
PB (e-val)	8.00	31.00	45.50	29.50
ALIGN4D	11.50	37.00	55.50	35.50

Alignment problem

Results. Turn over.

Mycoplasma genitalium MODPIPE Models

Number of ORFs	479
Average ORF length	364

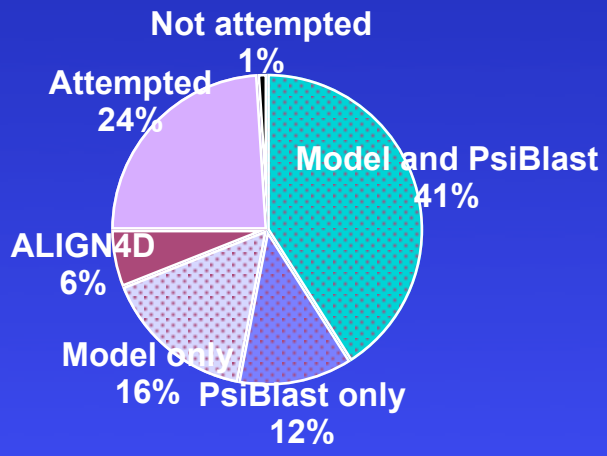


Alignment problem

Results. Turn over.

Mycoplasma genitalium MODPIPE Models

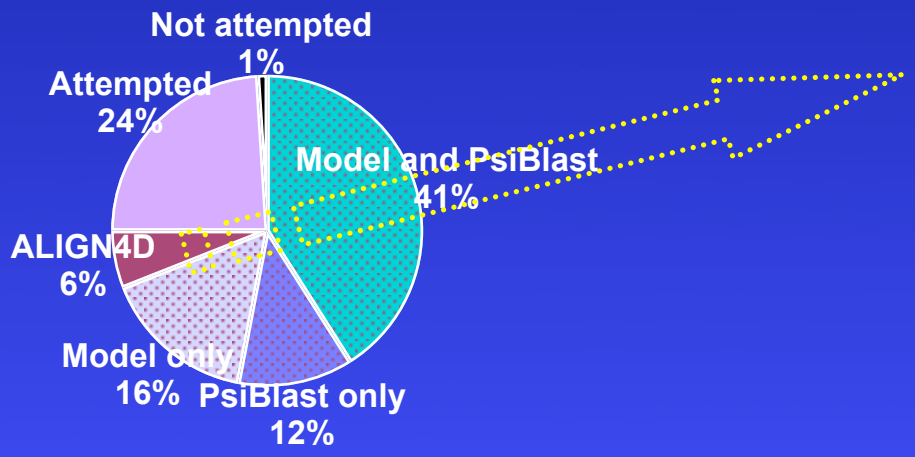
Number of ORFs	479
Average ORF length	364



Alignment problem
Results. Turn over.

Mycoplasma genitalium MODPIPE Models

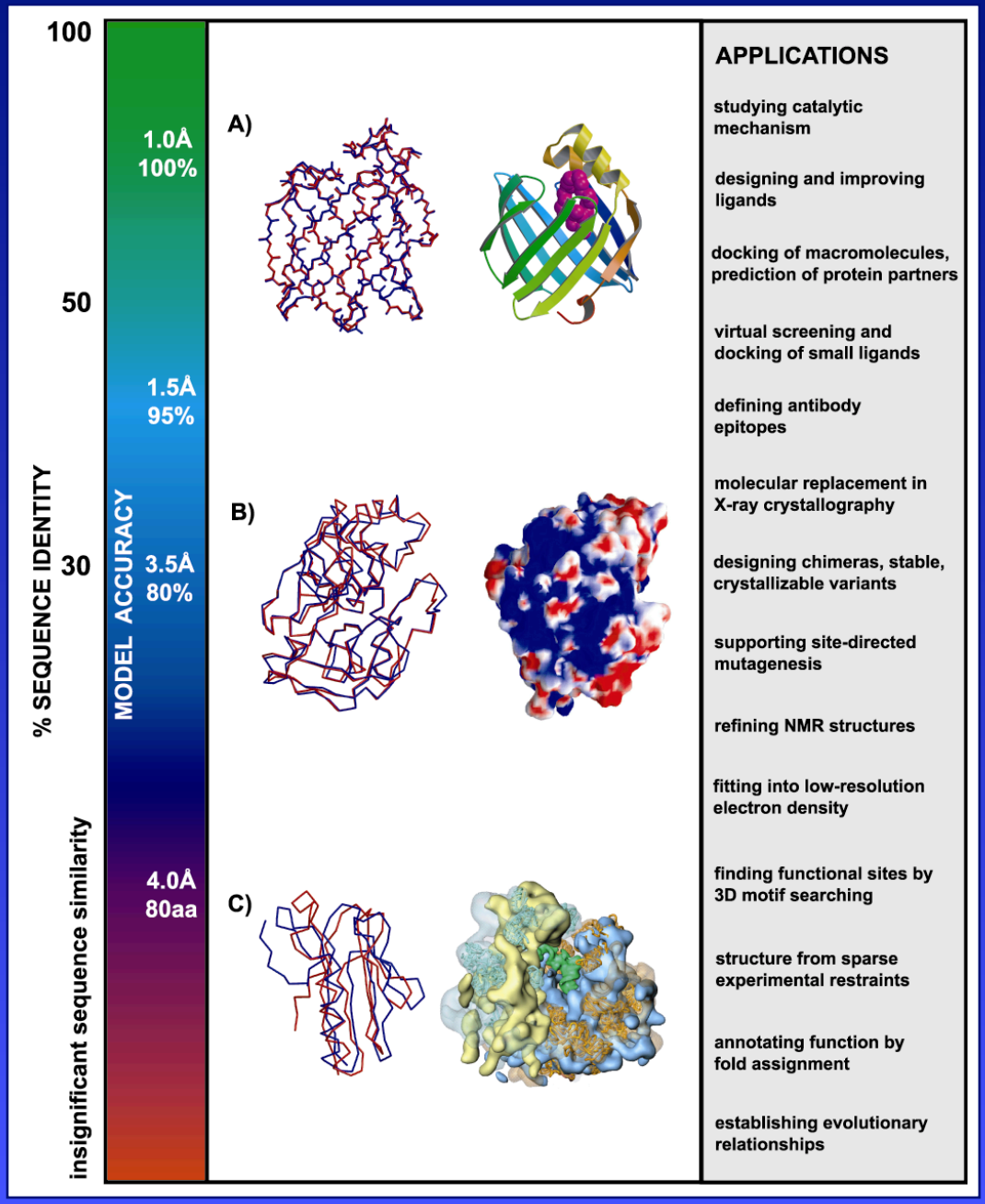
Number of ORFs	479
Average ORF length	364



~ 30 extra accurate models for *M. g.* genome.

~ 40,000 models for TrEMBL-SP "genome".

Applications of Comparative Models



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

A. Šali & J. Kuriyan.
TIBS **22**, M20, 1999.

Do mast cell proteases bind proteoglycans? Where? When?

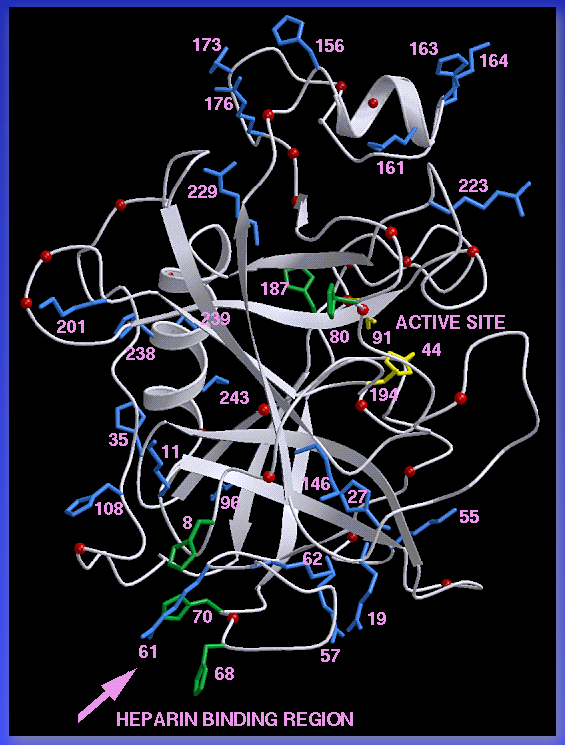
Predicting features of a model that are not present in the template

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

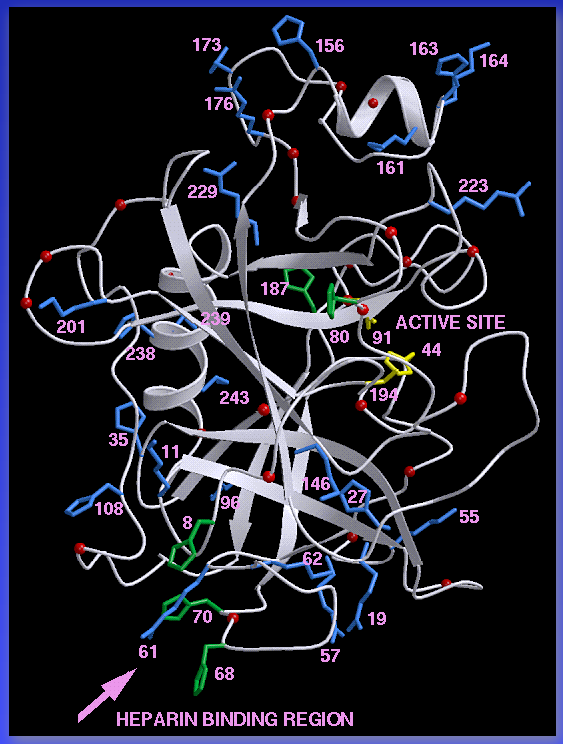
1. mMCPs bind negatively charged proteoglycans through electrostatic interactions?
2. Comparative models used to find clusters of positively charged surface residues.
3. Tested by site-directed mutagenesis.

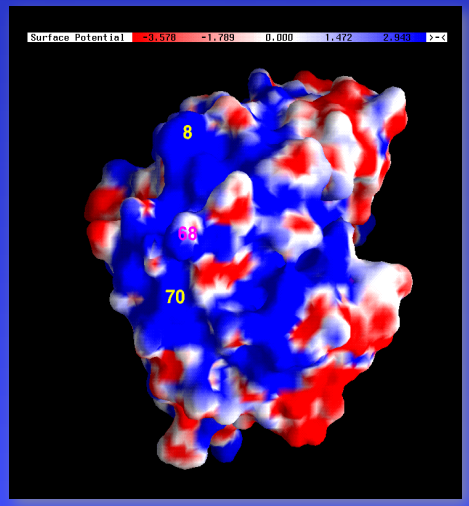


Do mast cell proteases bind proteoglycans? Where? When?

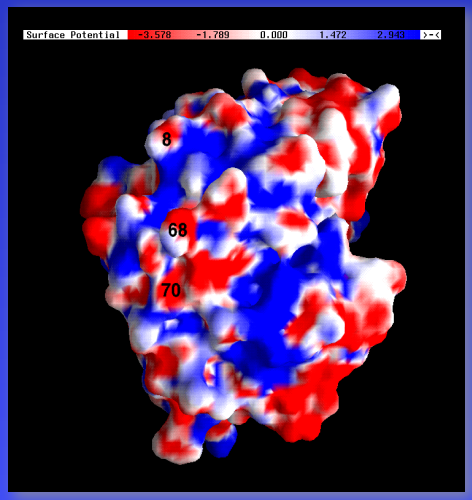
Predicting features of a model that are not present in the template

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 *p*H=5 (His^{+})

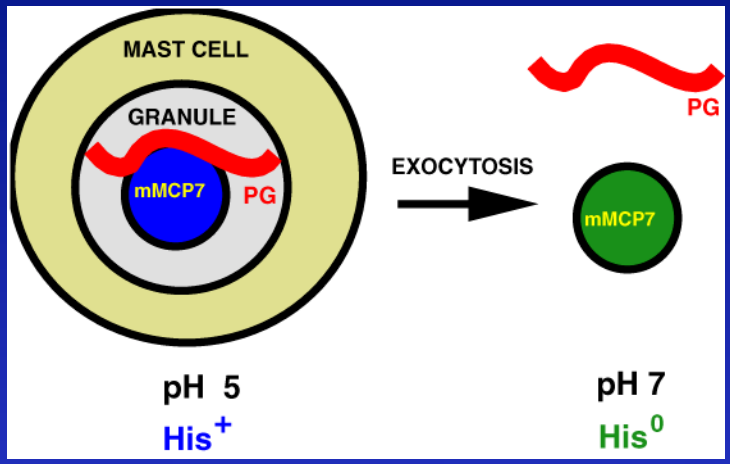


Native mMCP-7 at *p*H=7 (His^{0})

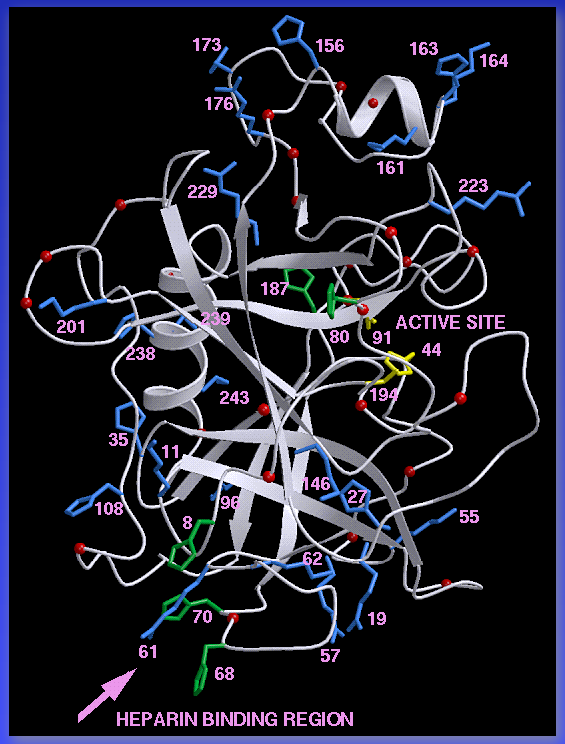
Do mast cell proteases bind proteoglycans? Where? When?

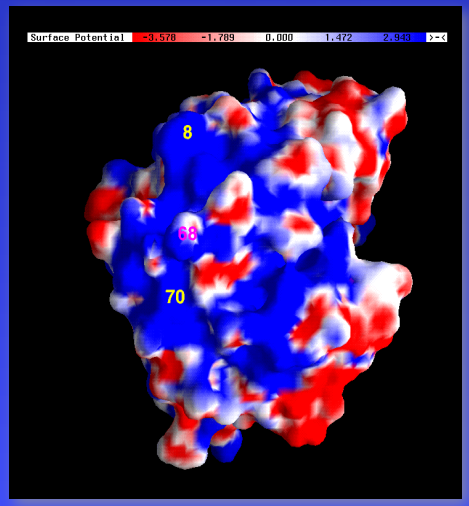
Predicting features of a model that are not present in the template

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.

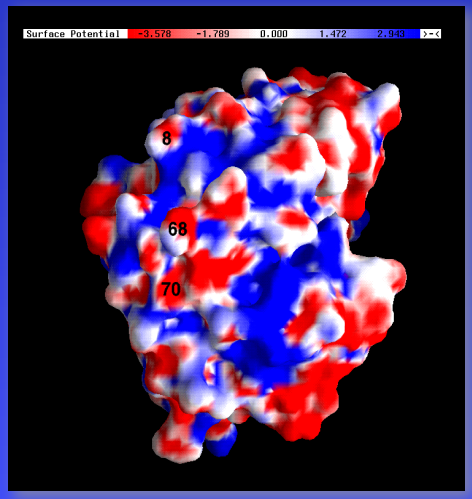


Huang *et al*. *J. Clin. Immunol*. **18**,169,1998.
Matsumoto *et al*. *J.Biol.Chem*. **270**,19524,1995.
Šali *et al*. *J. Biol. Chem*. **268**, 9023, 1993.





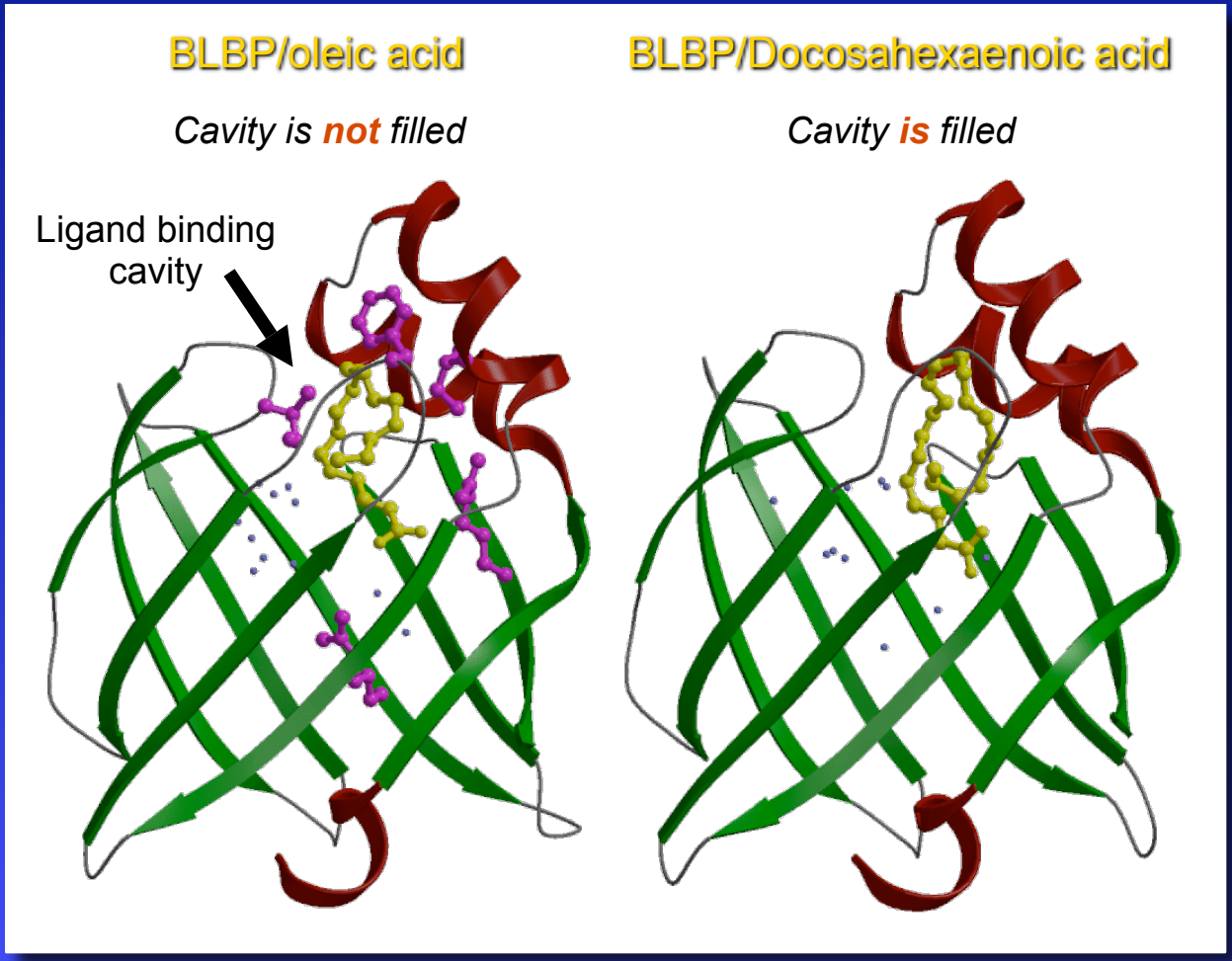
Native mMCP-7 at *p*H=5 (His^{+})



Native mMCP-7 at *p*H=7 (His^{0})

What is the physiological ligand of Brain Lipid-Binding Protein?

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



1. BLBP binds fatty acids.
2. Build a 3D model.
3. Find the fatty acid that fits most snuggly into the ligand binding cavity.

L. Xu, R. Sánchez, A. Šali, N. Heintz, *J. Biol. Chem.* **271**, 24711, 1996.

Structural Genomics

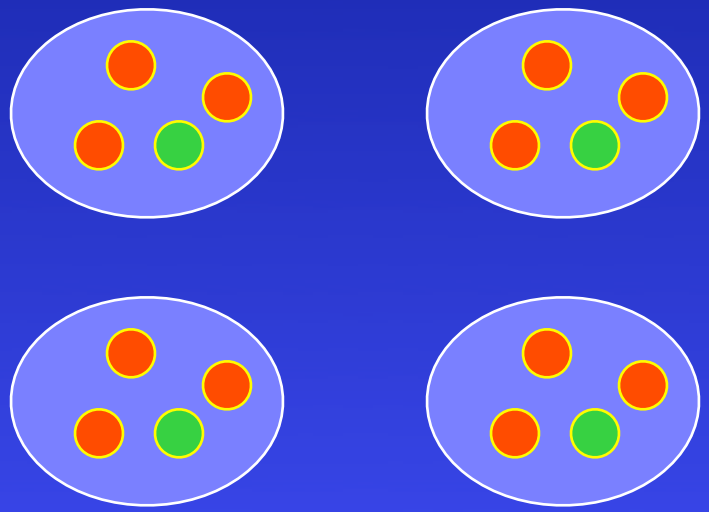
Sali. *Nat. Struct. Biol*. **5**, 1029, 1998.
Sali *et al. Nat. Struct. Biol*., **7**, 986, 2000.
Sali. *Nat. Struct. Biol.* **7**, 484, 2001.

Characterize most protein sequences based on related known structures.

Structural Genomics

Sali. *Nat. Struct. Biol.* **5**, 1029, 1998.
Sali *et al. Nat. Struct. Biol.*, **7**, 986, 2000.
Sali. *Nat. Struct. Biol.* **7**, 484, 2001.

Characterize most protein sequences based on related known structures.



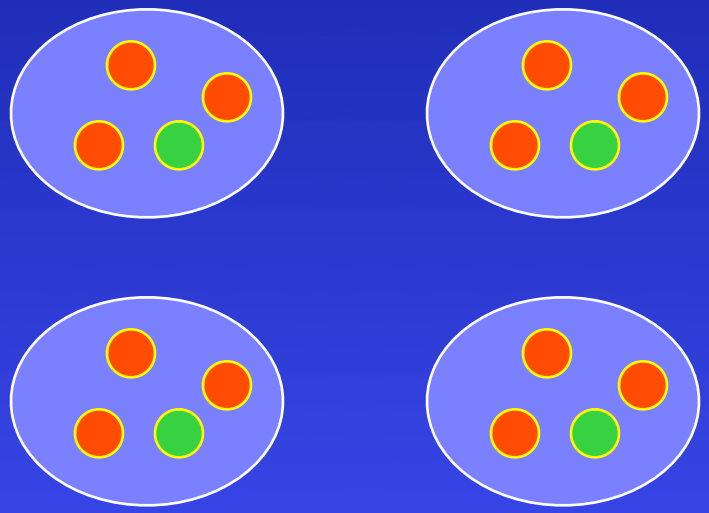
The number of “families” is much smaller than the number of proteins.

Any one of the members of a family is fine.

Structural Genomics

Sali. *Nat. Struct. Biol.* **5**, 1029, 1998.
Sali *et al. Nat. Struct. Biol.*, **7**, 986, 2000.
Sali. *Nat. Struct. Biol.* **7**, 484, 2001.

Characterize most protein sequences based on related known structures.

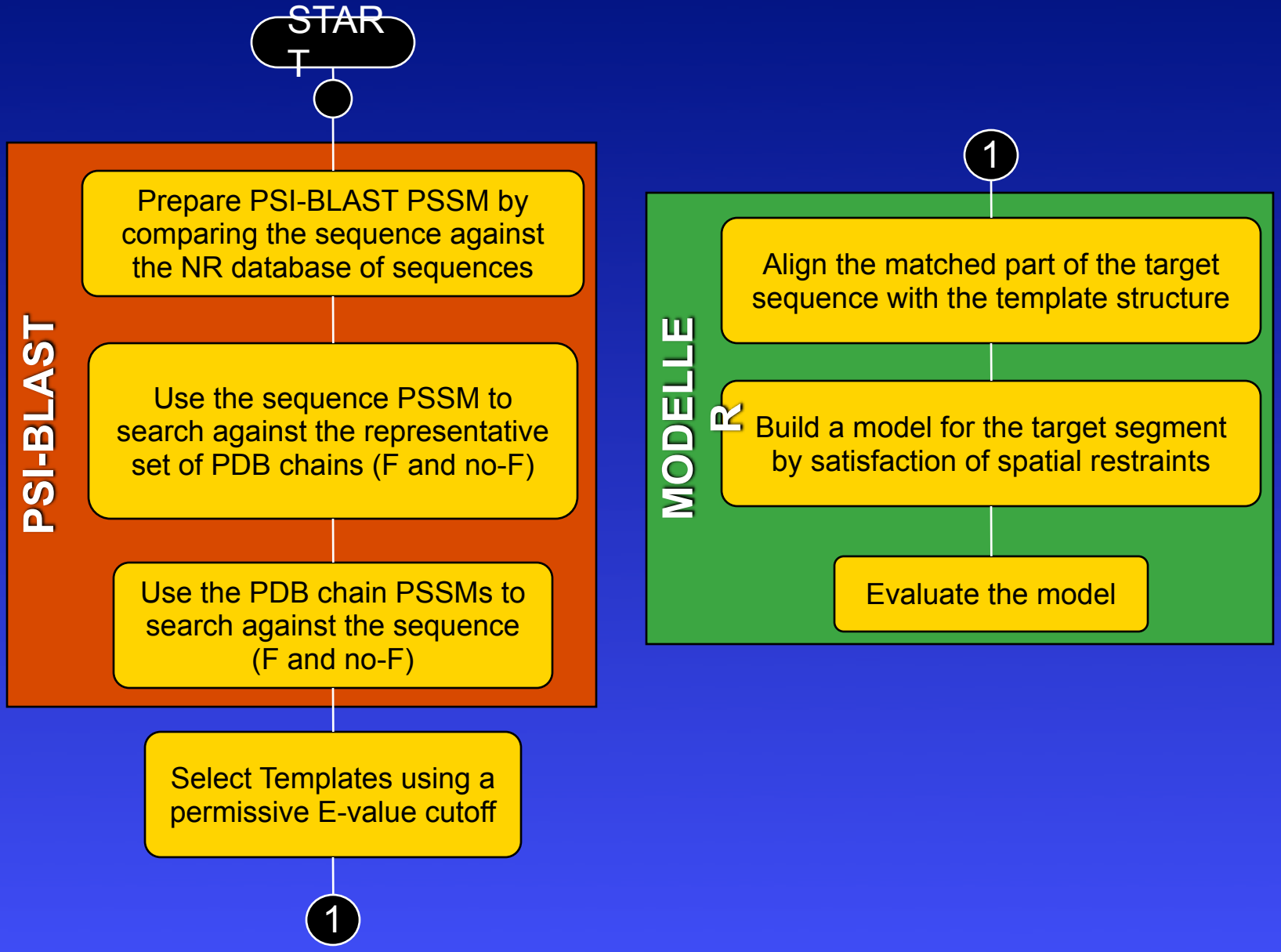


The number of "families" is much smaller than the number of proteins.

Any one of the members of a family is fine.

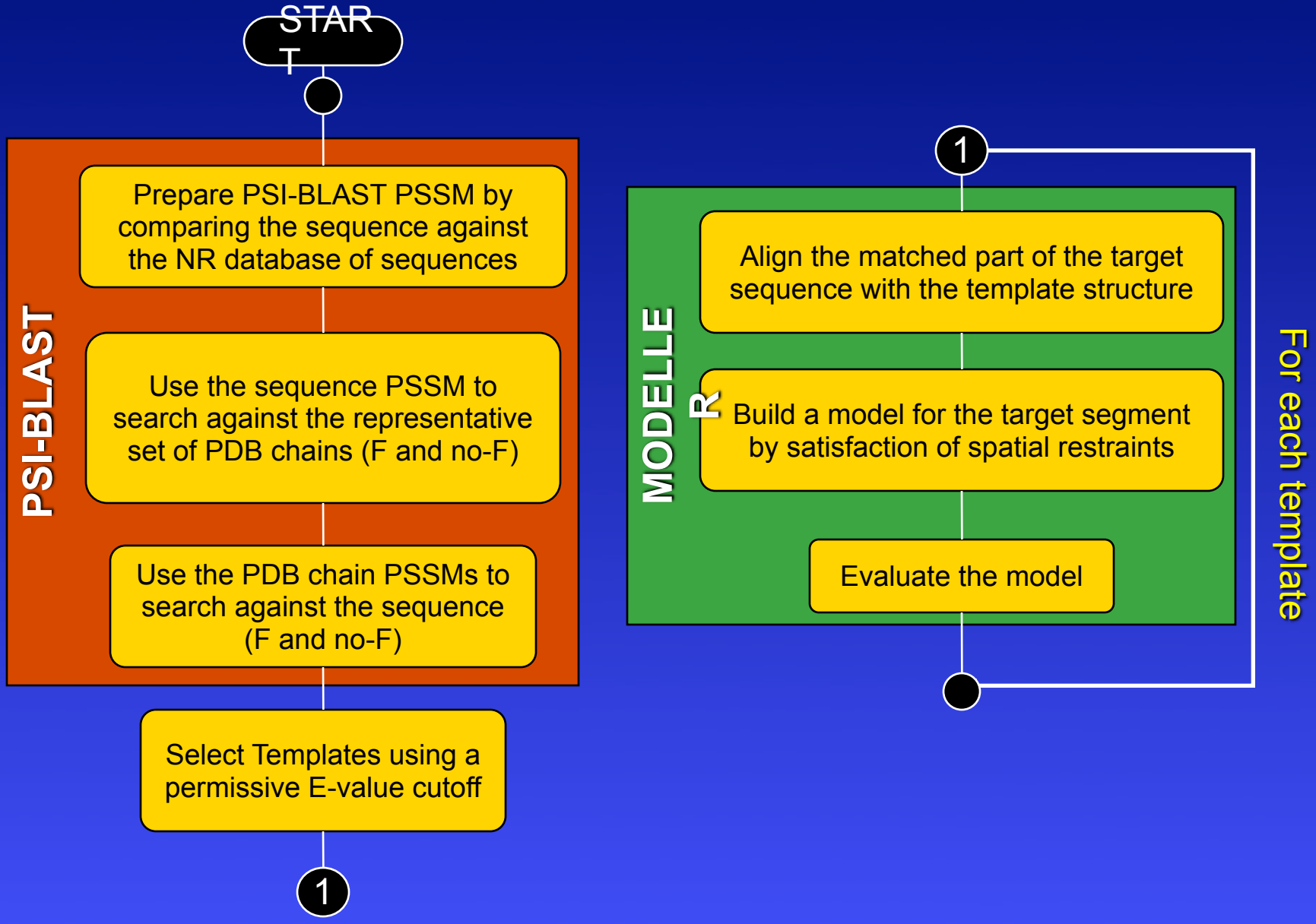
There are ~16,000 30% seq id families
(Vitkup *et al. Nat. Struct. Biol.* **8**, 559, 2001)

MODPIPE: Large-Scale Comparative Protein Structure Modeling



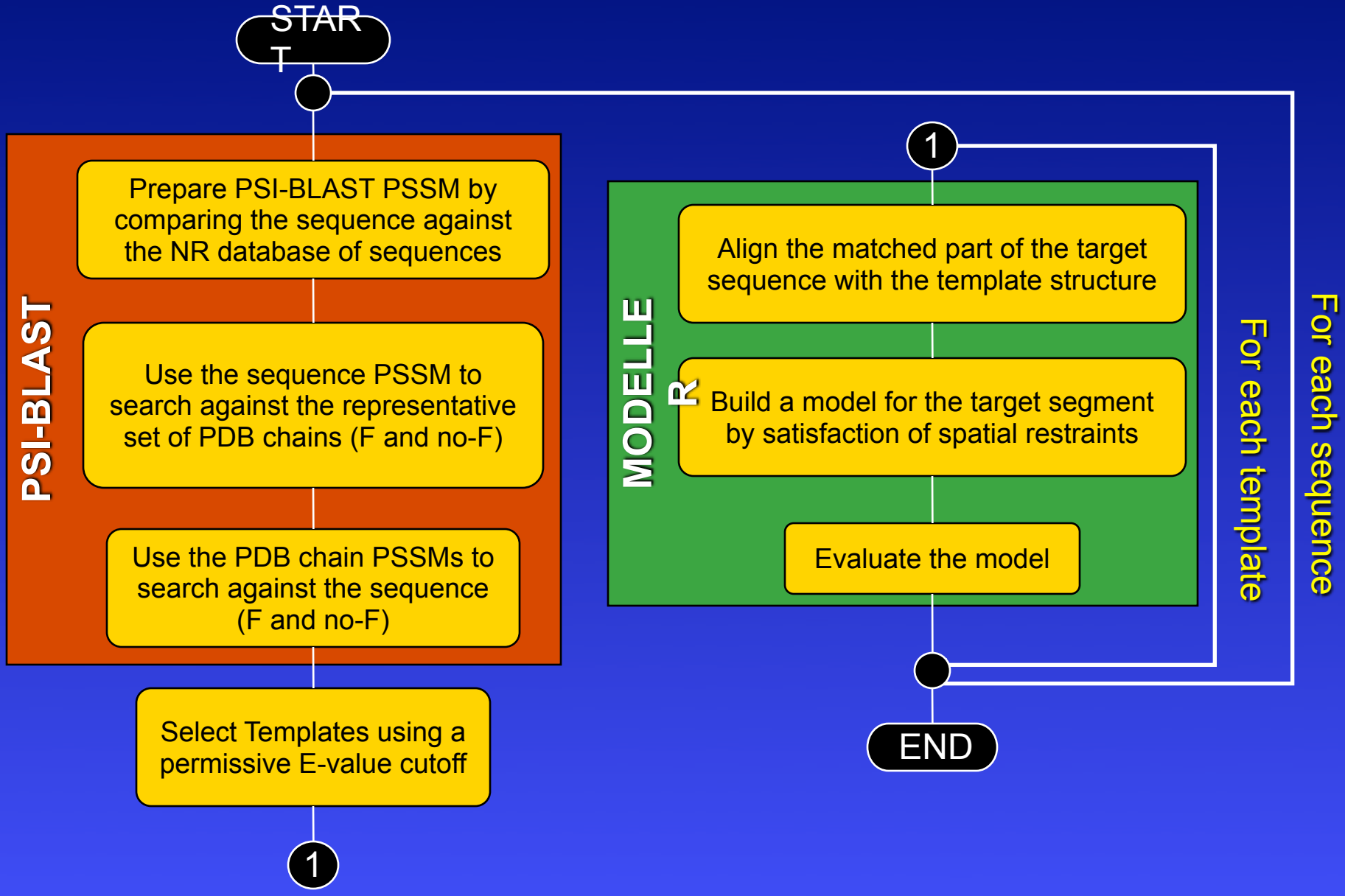
R. Sánchez & A. Šali, *Proc. Natl. Acad. Sci. USA* **95**, 13597, 1998
R. Sánchez, F. Melo, N. Mirkovic, A. Šali, in preparation

MODPIPE: Large-Scale Comparative Protein Structure Modeling



R. Sánchez & A. Šali, *Proc. Natl. Acad. Sci. USA* **95**, 13597, 1998
R. Sánchez, F. Melo, N. Mirkovic, A. Šali, in preparation

MODPIPE: Large-Scale Comparative Protein Structure Modeling



R. Sánchez & A. Šali, *Proc. Natl. Acad. Sci. USA* **95**, 13597, 1998
R. Sánchez, F. Melo, N. Mirkovic, A. Šali, in preparation

Comparative modeling of the TrEMBL database

Unique sequences processed: 733,239

Sequences with fold assignments or models: 415,937 (57%)

4/03/02 ~4 weeks on 500 Pentium III CPUs

Comparative modeling of the TrEMBL database

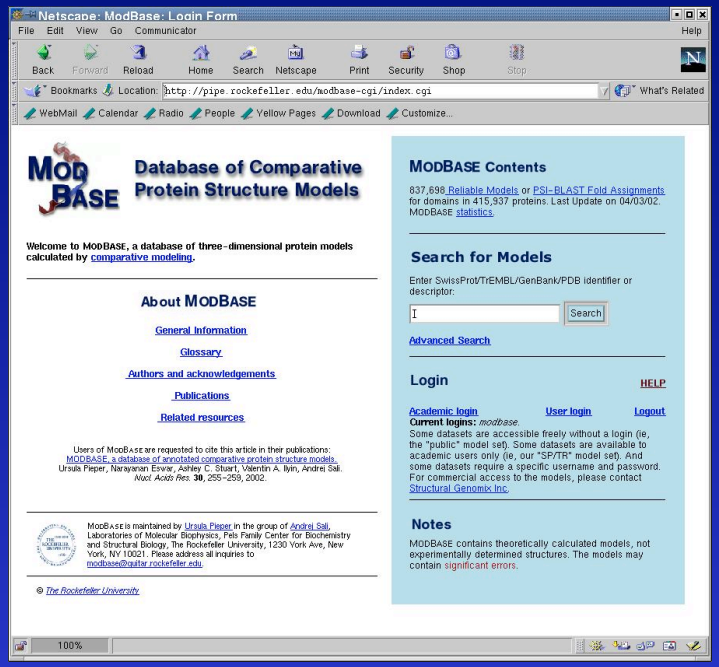
Unique sequences processed: 733,239

Sequences with fold assignments or models: 415,937 (57%)

70% of models based on <30% sequence identity to template.

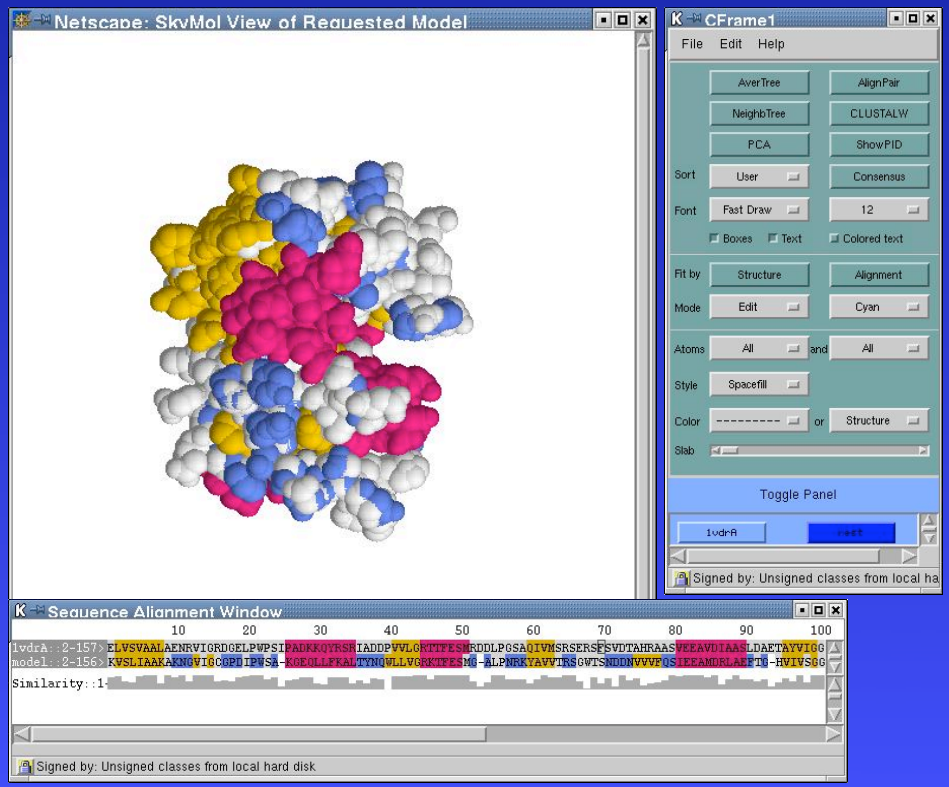
On average, only a domain per protein is modeled
(an “average” protein has 2.5 domains of 175 aa).

4/03/02 ~4 weeks on 500 Pentium III CPUs



http://guitar.rockefeller.edu/modbase

Pieper *et al.*, Nucl. Acids Res. 2002.

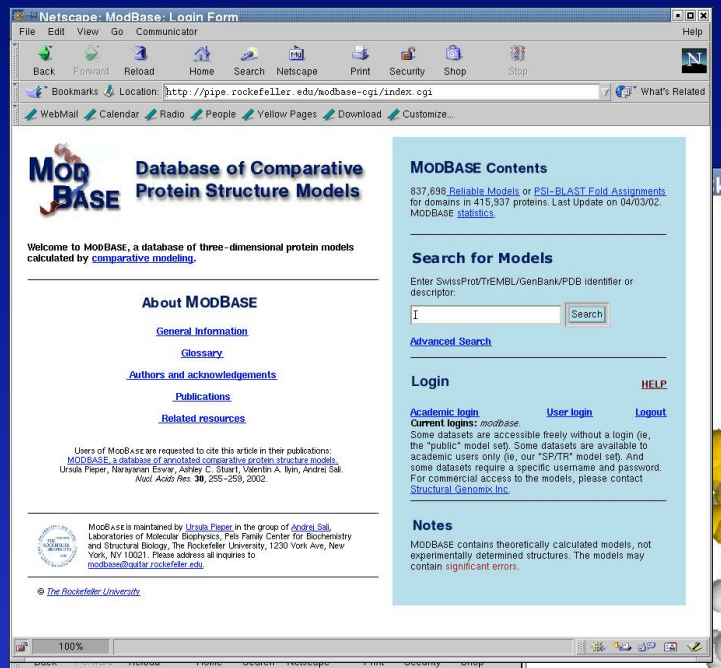


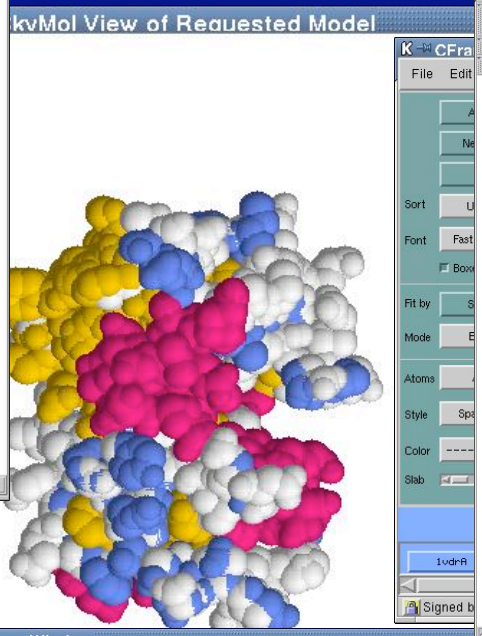
http://guitar.rockefeller.edu/modview

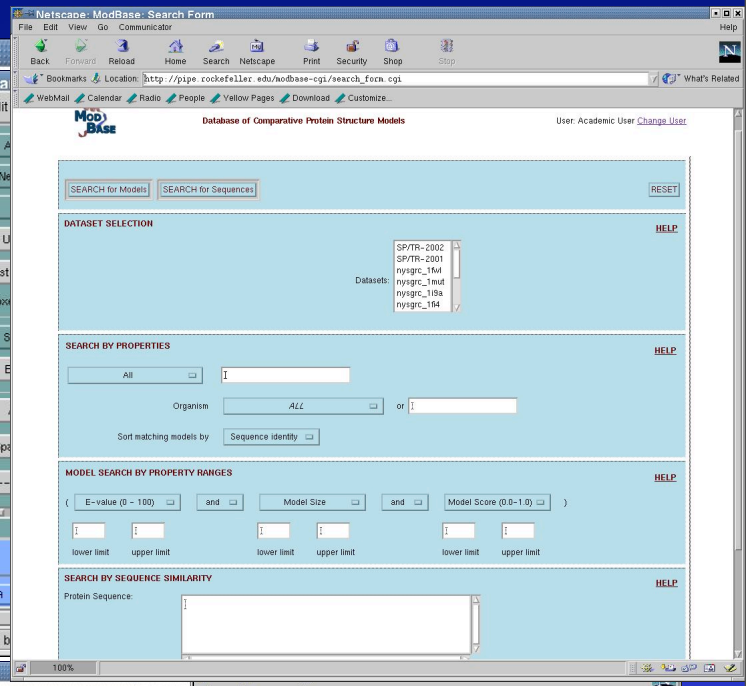
Ilyin *et al.*, 2002 *(in press)*.

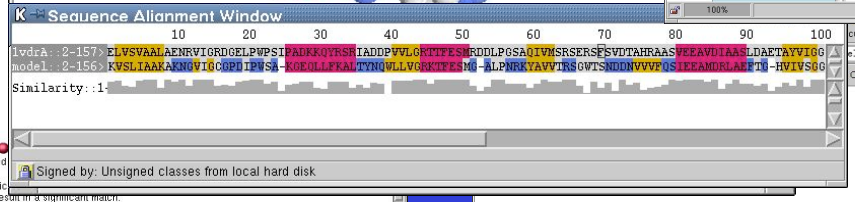
http://guitar.rockefeller.edu/modbase

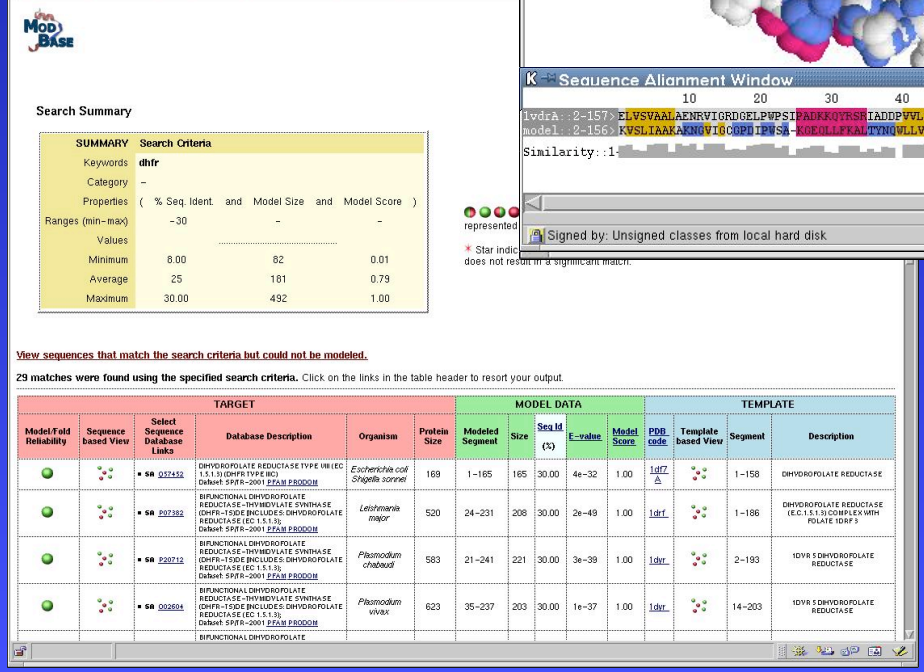
Pieper *et al.*, Nucl. Acids Res. 2002.

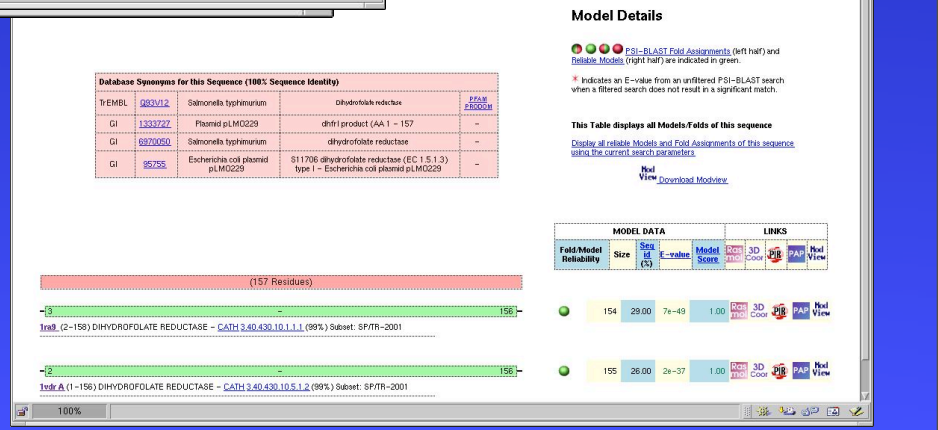












Conclusions

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 57% of the proteins (25% 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 57% of the proteins (25% 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 57% of the proteins (25% 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 in approximately 5 years, because of **structural genomics**.

Acknowledgments

Burroughs Wellcome Fund
The Rockefeller University Presidential Fellowship

Andrej Šali
Frank Alber
Narayanan Eswar
András Fiser
Valentin Ilyin
Bozidar Yerkovich
Bino John
M. S. Madhusudhan
Linda McMahan
Nebojša Mirković
Ursula Pieper
Andrea Rossi
Ash Stuart

http://guitar.rockefeller.edu