Comparative protein structure modeling of genes and genomes

Marc A. Marti-Renom

Department of Biopharmaceutical Sciences
University of California, San Francisco

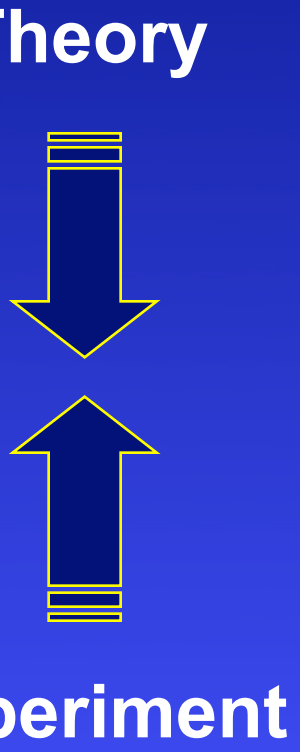
Comparative Modeling…

Why protein structure prediction?

	Y 2003	Y 2005
Sequences	1,000,000	millions
Structures	18,000	50,000

Why protein structure prediction?

	Y 2003
Sequences	1,000,000
Structures	18,000



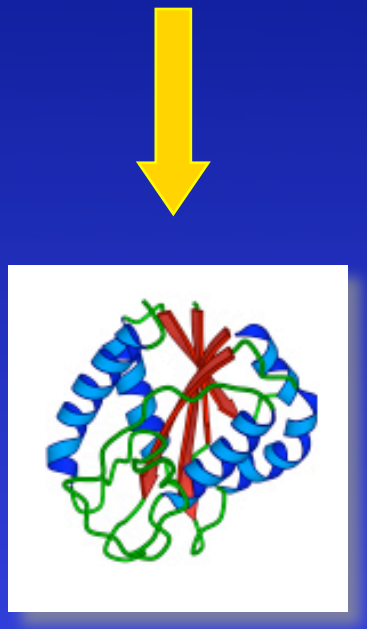
Why protein structure prediction?

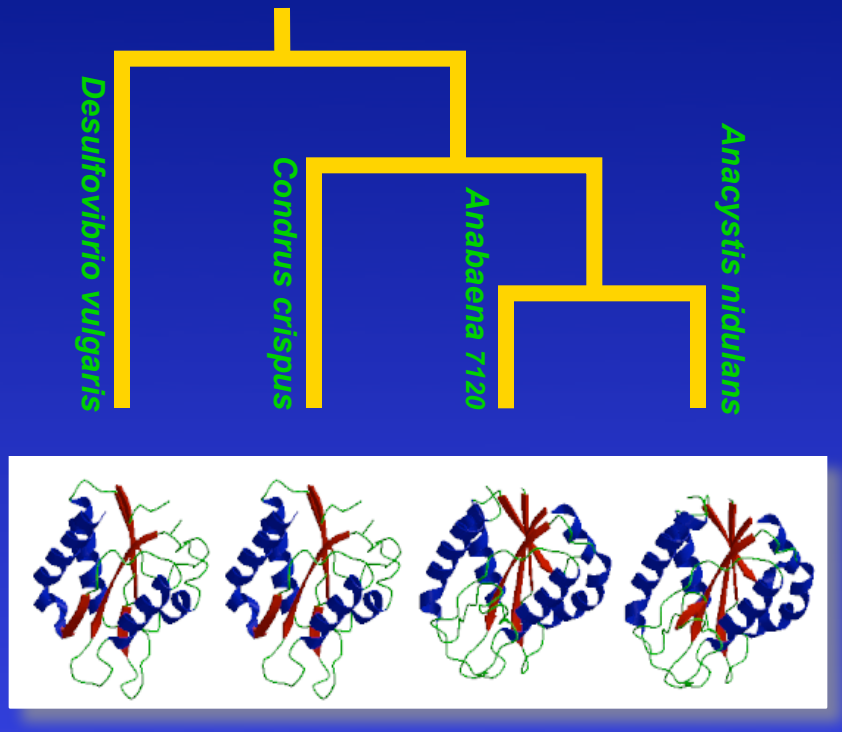
	Y 2003
Sequences	1,000,000
Structures	400,000

http://salilab.org/modbase

Principles of Protein Structure

GFCHIKAYTRLIMVG…





Folding

Ab initio prediction

Evolution

Threading
Comparative Modeling

Comparative Modeling by Satisfaction of Spatial Restraints (MODELLER)

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

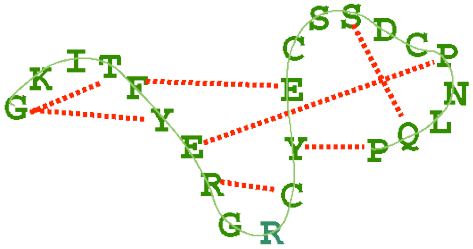
A. Šali & T. Blundell. *J. Mol. Biol.* **234**, 779, 1993.
J.P. Overington & A. Šali. *Prot. Sci*. **3**, 1582, 1994.
A. Fiser, R. Do & A. Šali. *Prot Sci*. **9**, 1753, 2000.

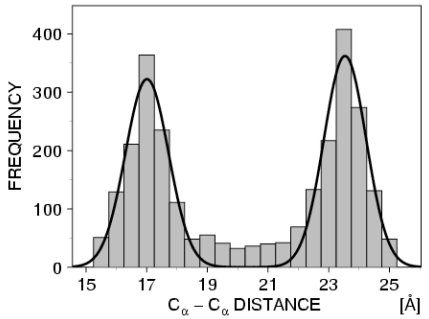
http://salilab.org/modeller

Comparative Modeling by Satisfaction of Spatial Restraints (MODELLER)

3D GKITFYERGFQGHCYESDC–NLQP…
SEQ GKITFYERG–––RCYESDCPNLQP…

1. Extract spatial restraints





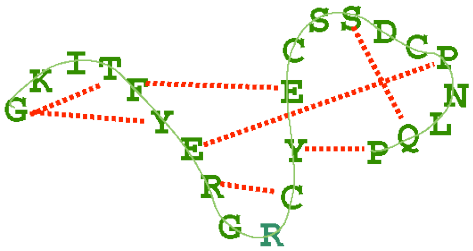
A. Šali & T. Blundell. *J. Mol. Biol.* **234**, 779, 1993.
J.P. Overington & A. Šali. *Prot. Sci.* **3**, 1582, 1994.
A. Fiser, R. Do & A. Šali. *Prot Sci.* **9**, 1753, 2000.

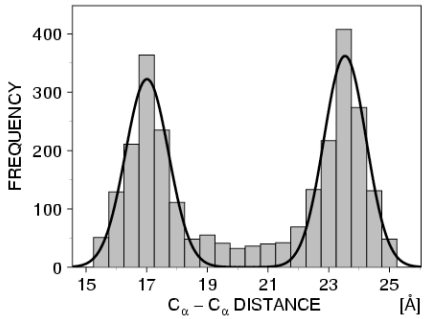
http://salilab.org/modeller

Comparative Modeling by Satisfaction of Spatial Restraints (MODELLER)

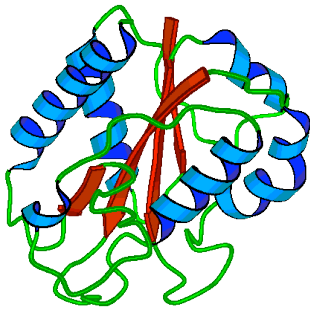
3D GKITFYERGFQGHCYESDC–NLQP…
SEQ GKITFYERG–––RCYESDCPNLQP…

1. Extract spatial restraints





2. Satisfy spatial restraints



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

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.

http://salilab.org/modeller

Steps in Comparative Protein Structure Modeling

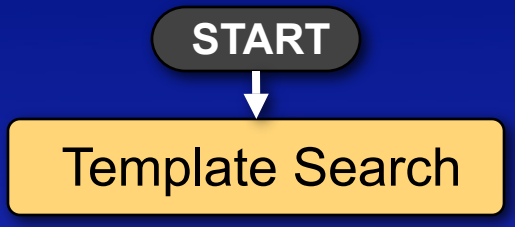
START

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.

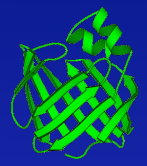
Steps in Comparative Protein Structure Modeling



TARGET

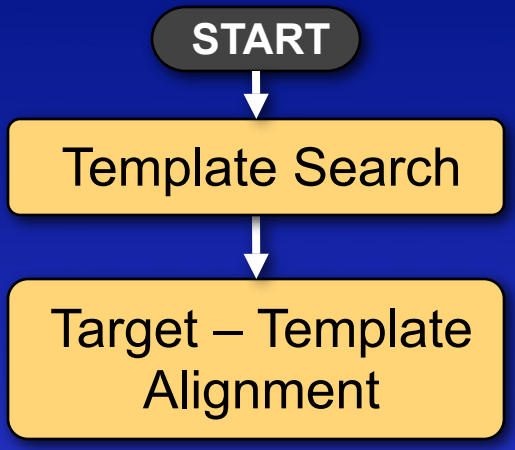
ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCLKIDD
VPERLIPERASFQWMNDK

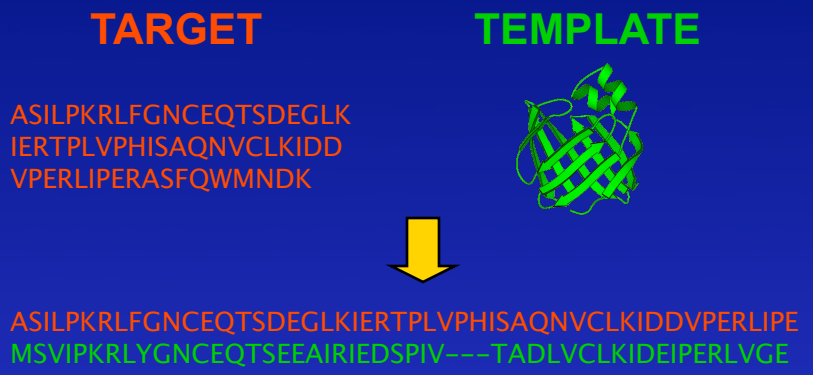
TEMPLATE



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.

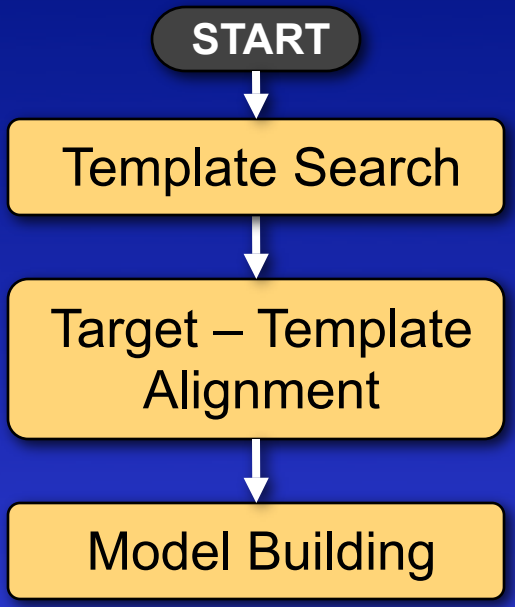
Steps in Comparative Protein Structure Modeling

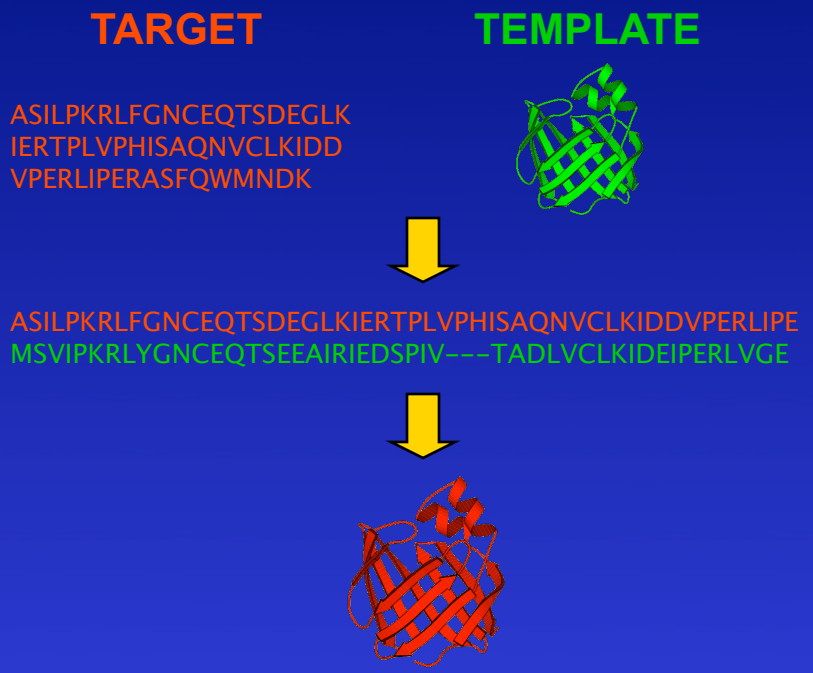




A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.
R. Sánchez & A. Šali, *Curr. Opin. Str. Biol.* 7, 206, 1997.
M. A. Martí-Renom *et al. Ann. Rev. Biophys. Biomolec. Struct.*, 29, 291, 2000.

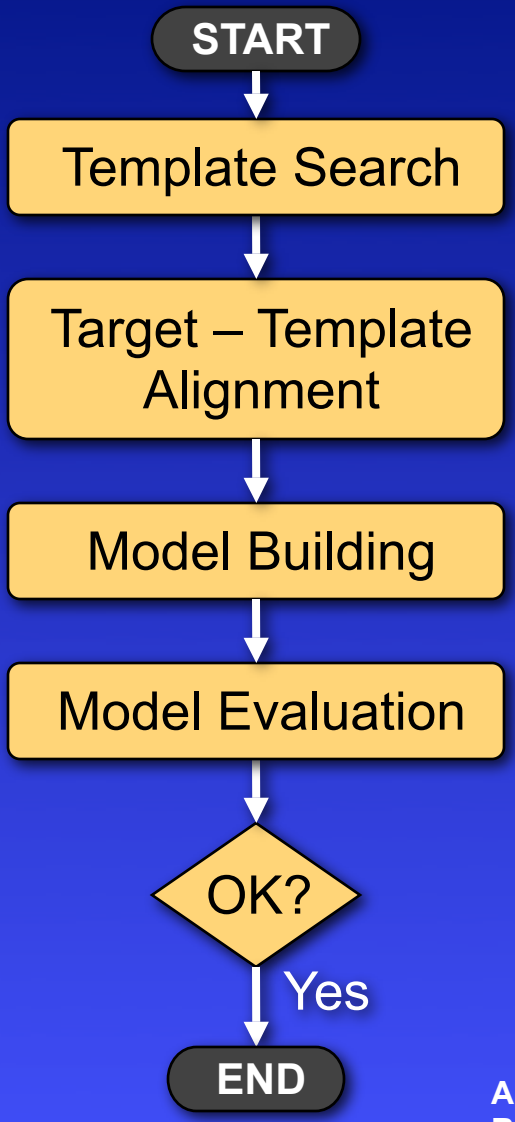
Steps in Comparative Protein Structure Modeling

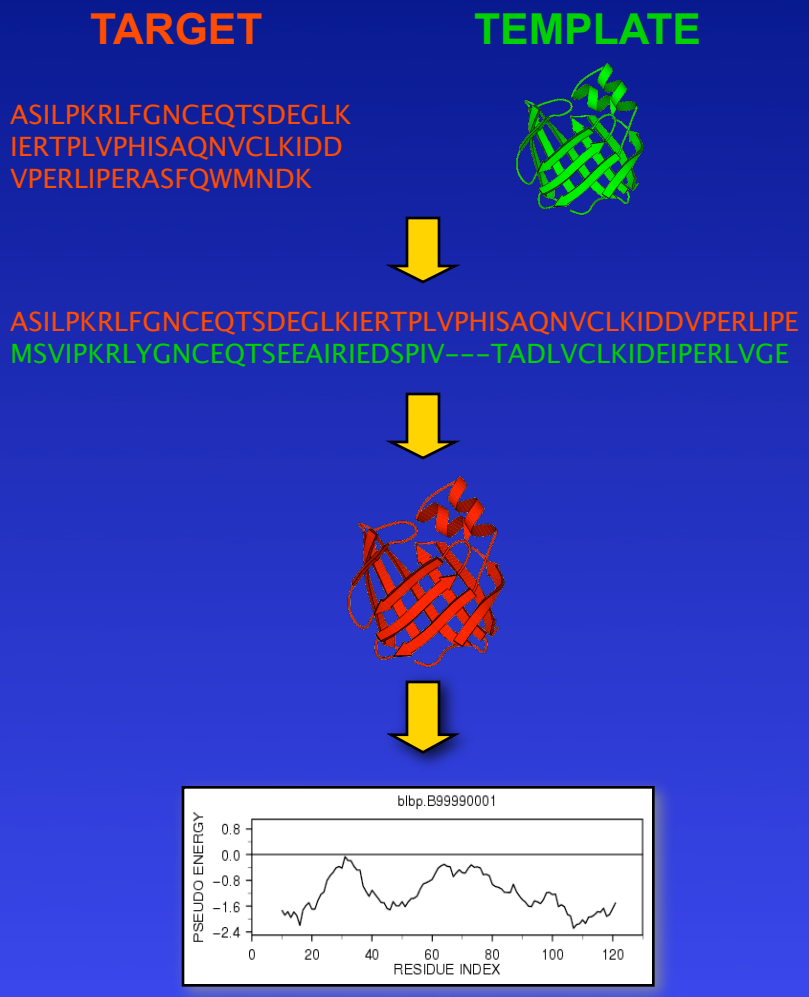




A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.
R. Sánchez & A. Šali, *Curr. Opin. Str. Biol.* 7, 206, 1997.
M. A. Martí-Renom *et al. Ann. Rev. Biophys. Biomolec. Struct.*, 29, 291, 2000.

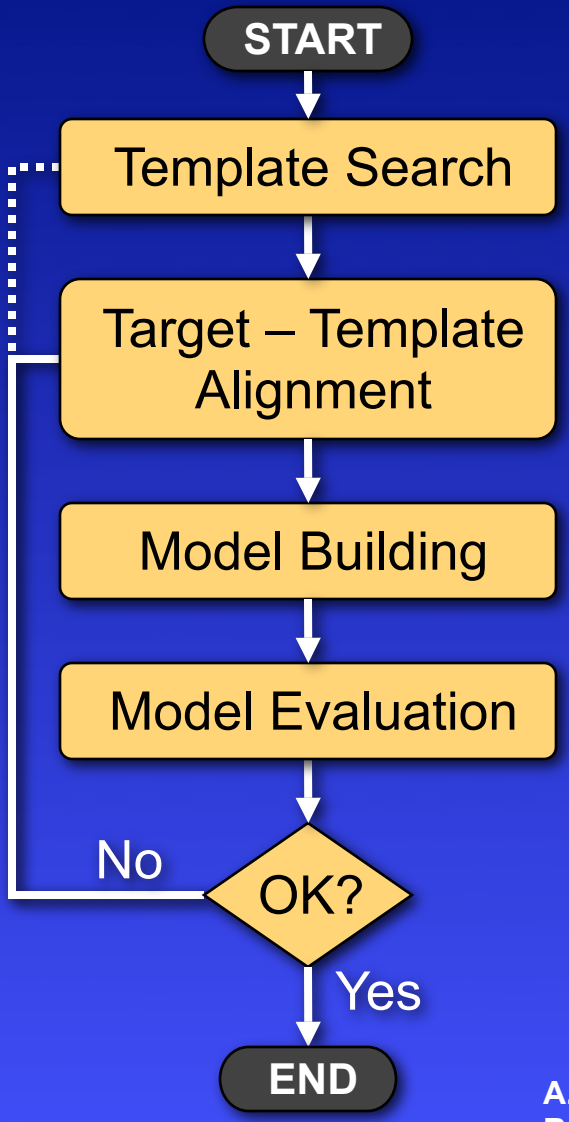
Steps in Comparative Protein Structure Modeling

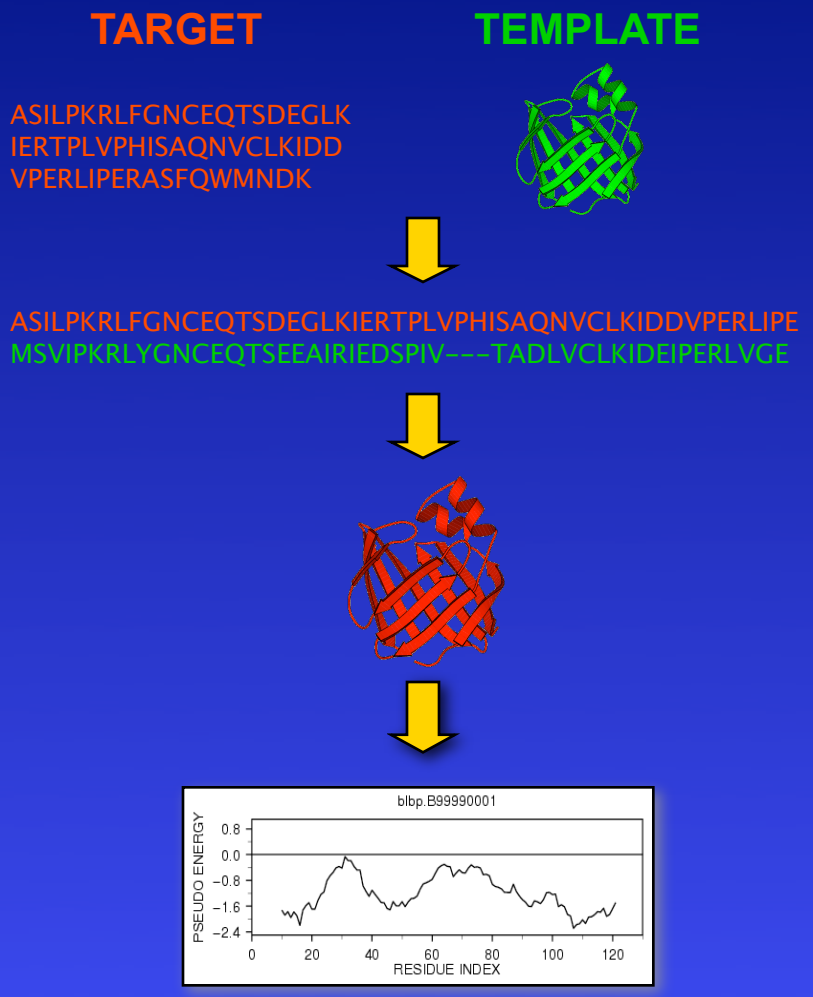




A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.
R. Sánchez & A. Šali, *Curr. Opin. Str. Biol.* 7, 206, 1997.
M. A. Martí-Renom *et al. Ann. Rev. Biophys. Biomolec. Struct.*, 29, 291, 2000.

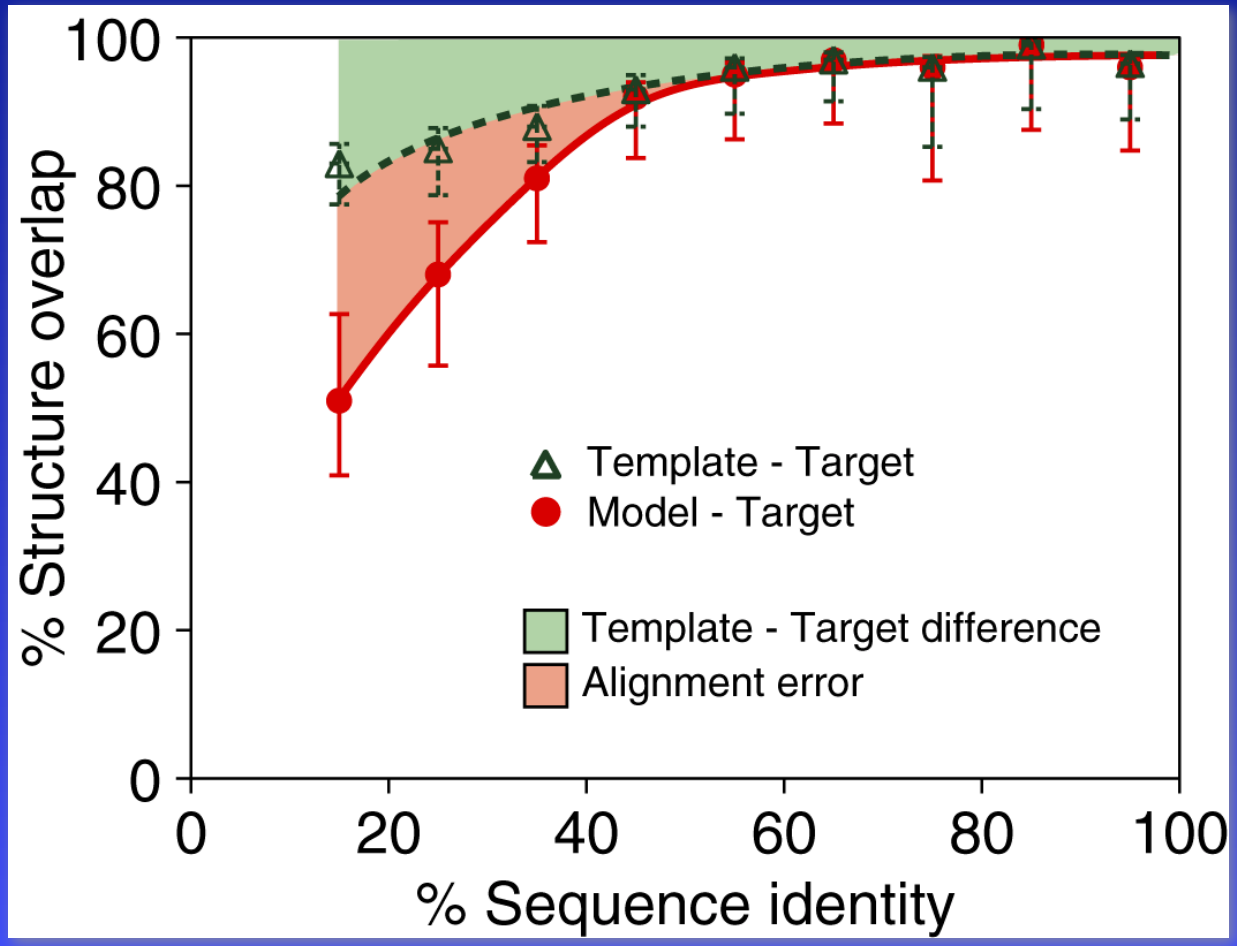
Steps in Comparative Protein Structure Modeling





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 as a Function of Target-Template Sequence Identity



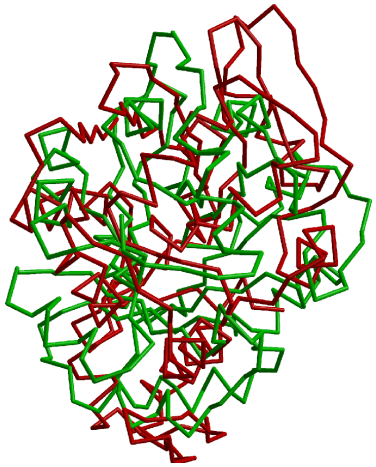
Typical Errors in Comparative Models

MODEL

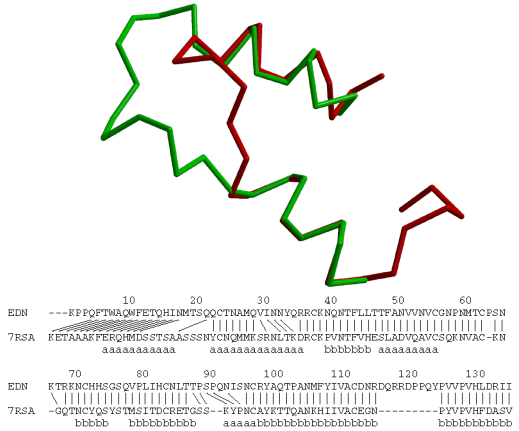
X-RAY

TEMPLATE

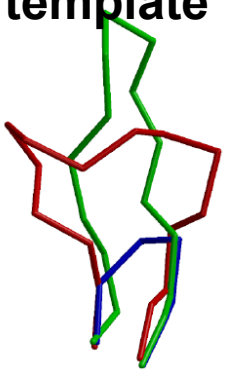
Incorrect template



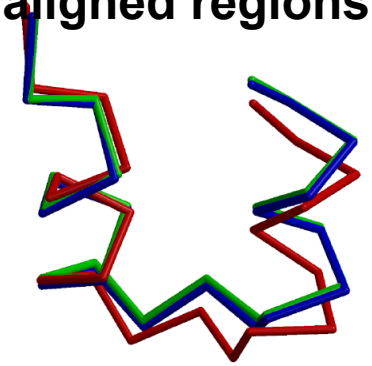
Misalignment



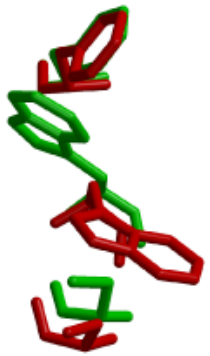
Region without a template



Distortion in correctly aligned regions



Sidechain packing

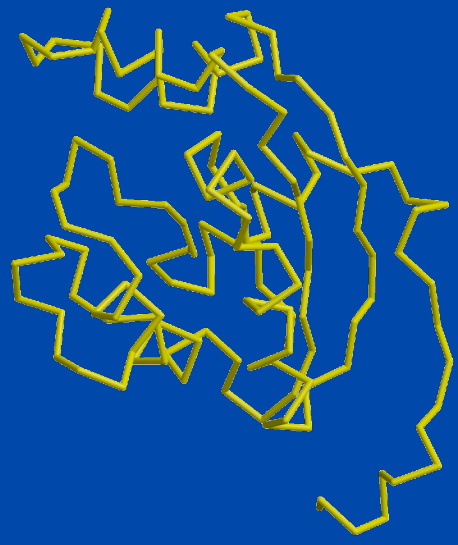


Model Accuracy

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

HIGH ACCURACY

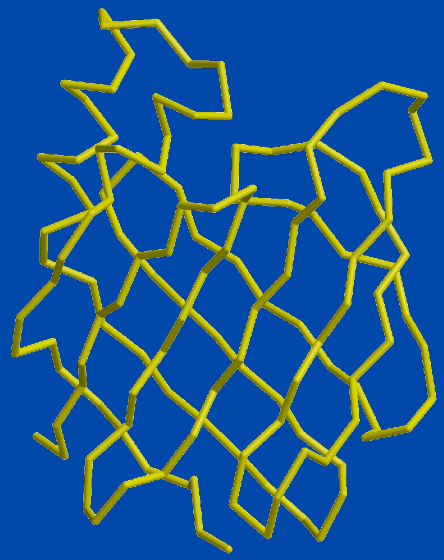
NM23
Seq id 77%



X-RAY

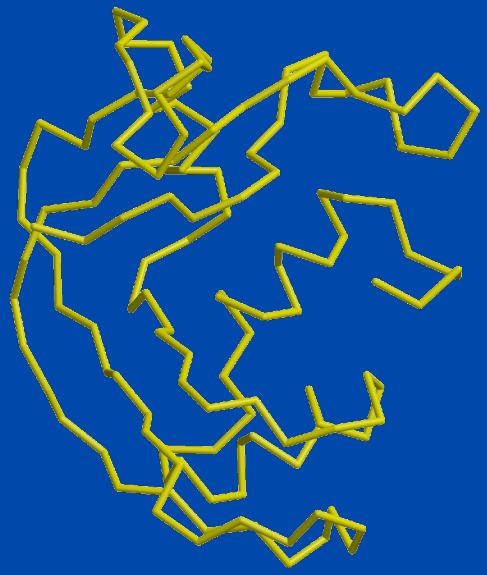
MEDIUM ACCURACY

CRABP
Seq id 41%



LOW ACCURACY

EDN
Seq id 33%

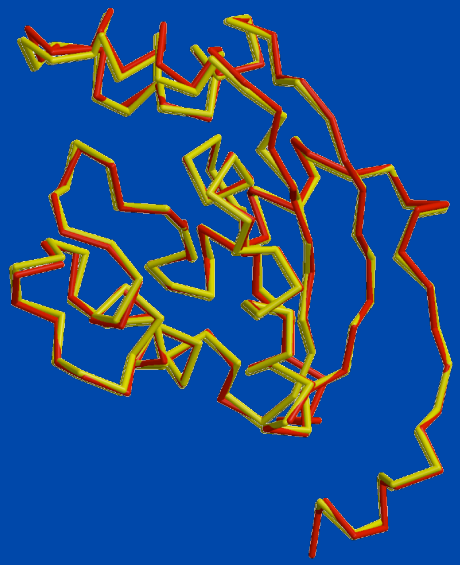


Model Accuracy

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

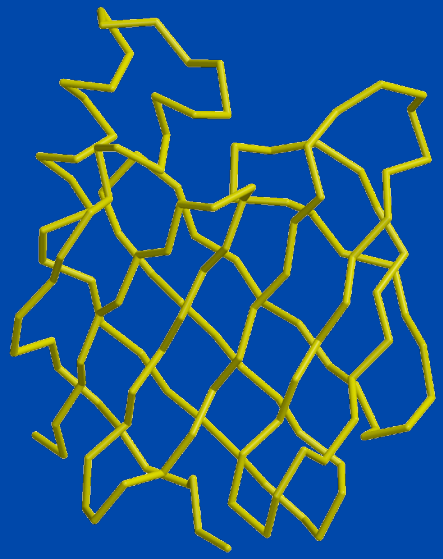


Sidechains
Core backbone
Loops

X-RAY / MODEL

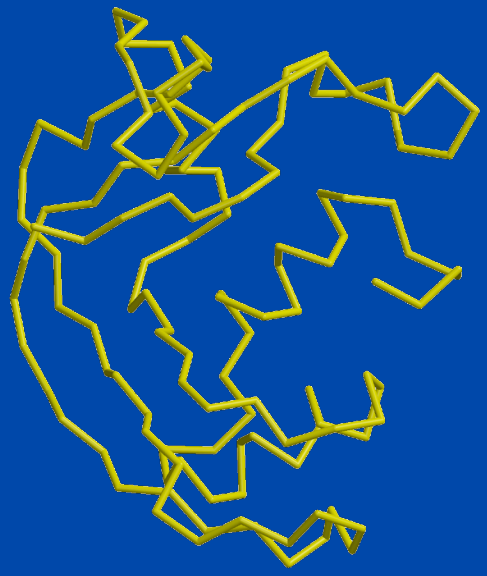
MEDIUM ACCURACY

CRABP
Seq id 41%



LOW ACCURACY

EDN
Seq id 33%

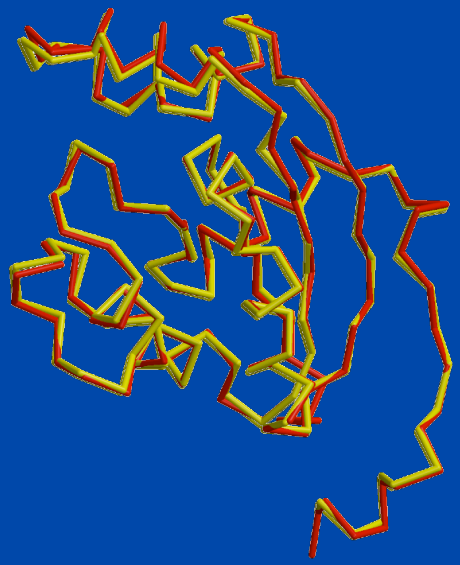


Model Accuracy

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

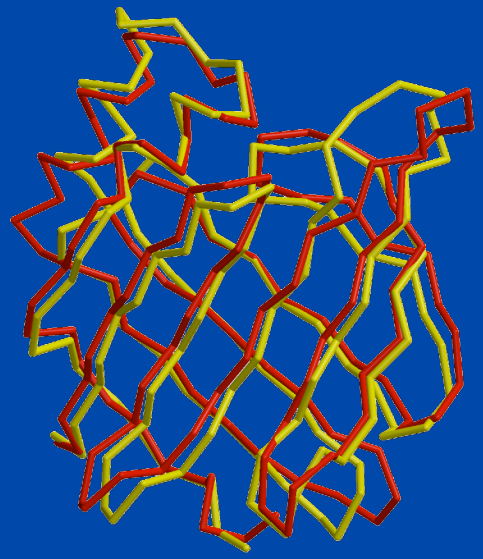


Sidechains
Core backbone
Loops

X-RAY / MODEL

MEDIUM ACCURACY

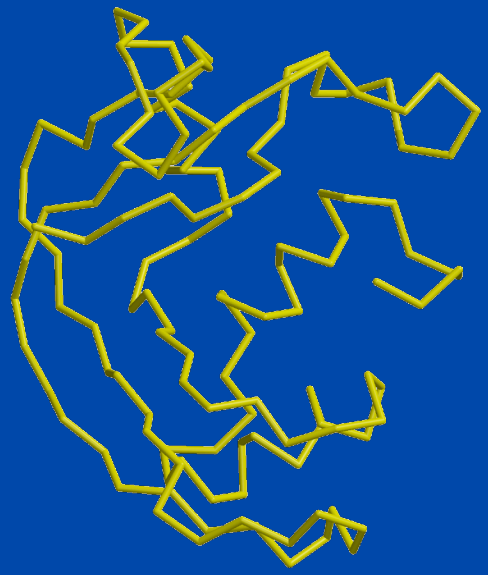
CRABP
Seq id 41%
Cα equiv 122/137
RMSD 1.34Å



Sidechains
Core backbone
Loops
Alignment

LOW ACCURACY

EDN
Seq id 33%

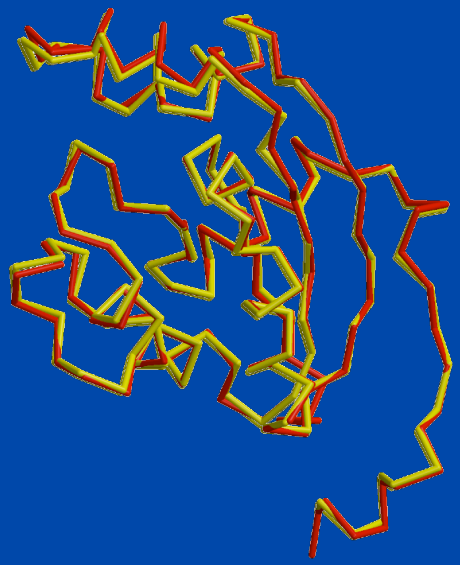


Model Accuracy

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

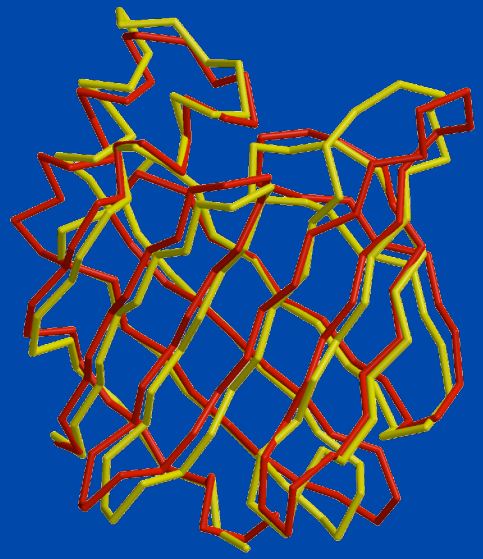


Sidechains
Core backbone
Loops

X-RAY / MODEL

MEDIUM ACCURACY

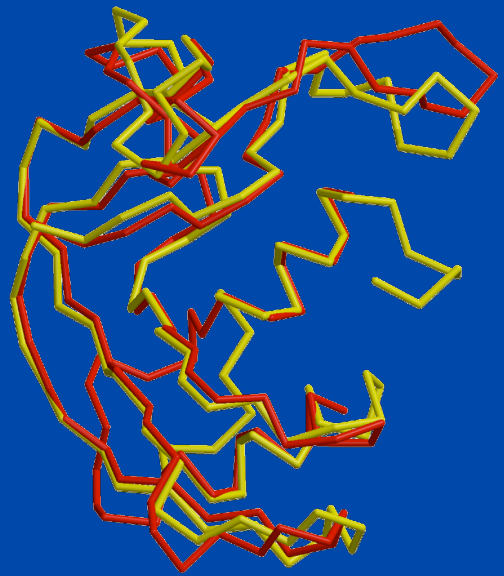
CRABP
Seq id 41%
Cα equiv 122/137
RMSD 1.34Å



Sidechains
Core backbone
Loops
Alignment

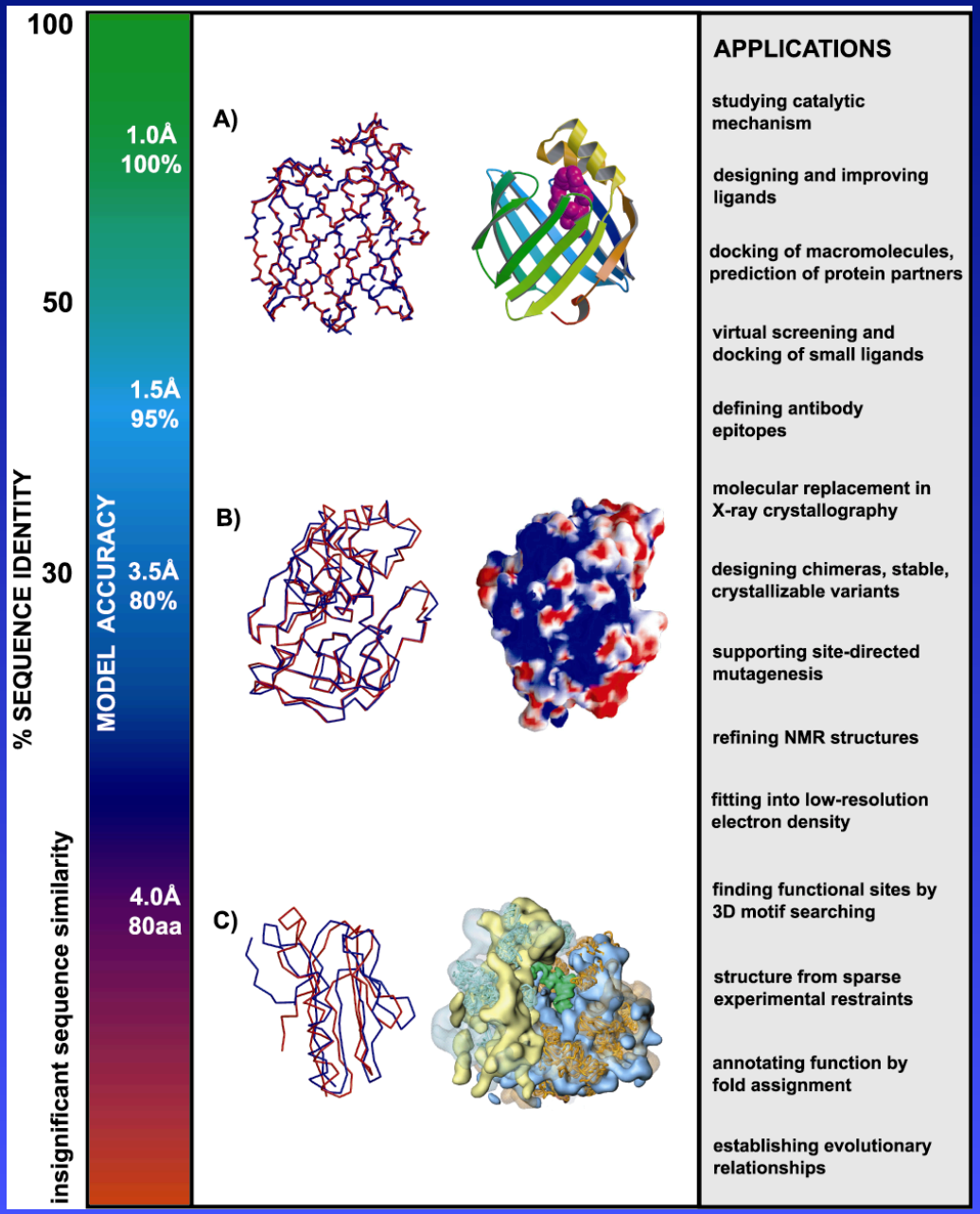
LOW ACCURACY

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



Sidechains
Core backbone
Loops
Alignment
Fold assignment

Applications of Comparative Models



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

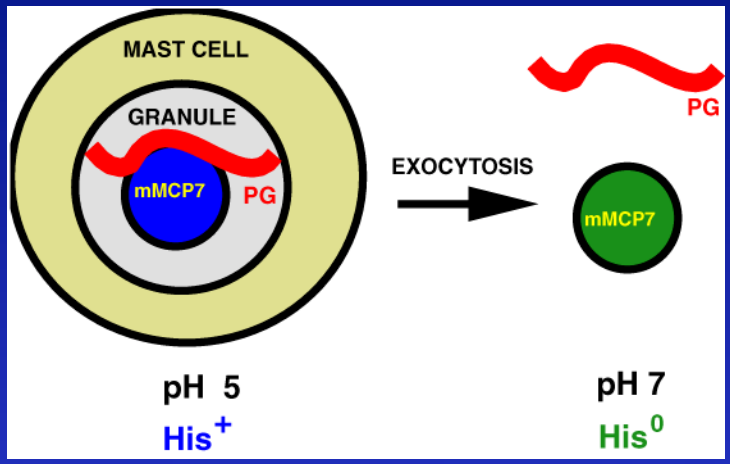
A. Sali & J. Kuriyan. *TIBS* **22**, M20, 1999.

genes…

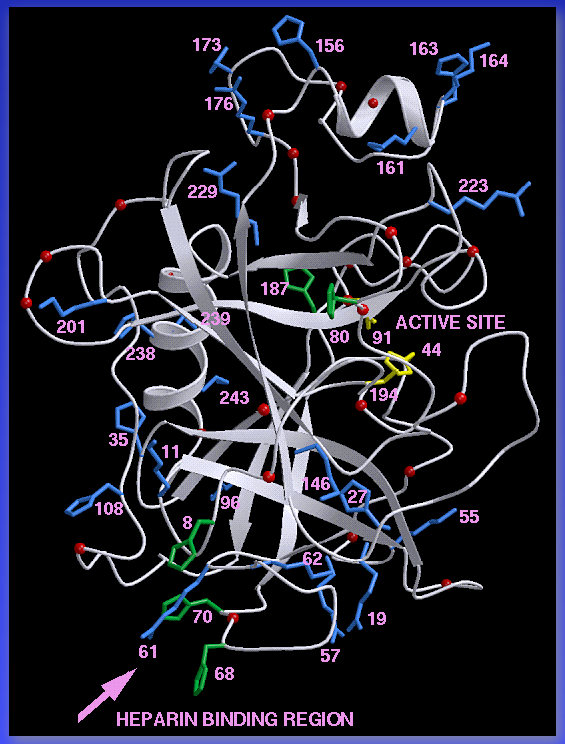
Do mast cell proteases bind proteoglycans? Where? When?

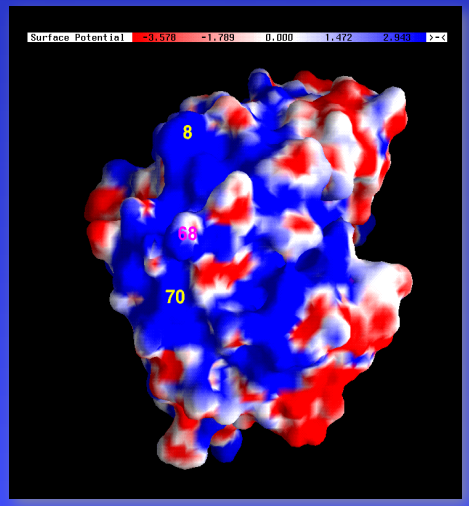
Predicting features of a model that are not present in the template

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.

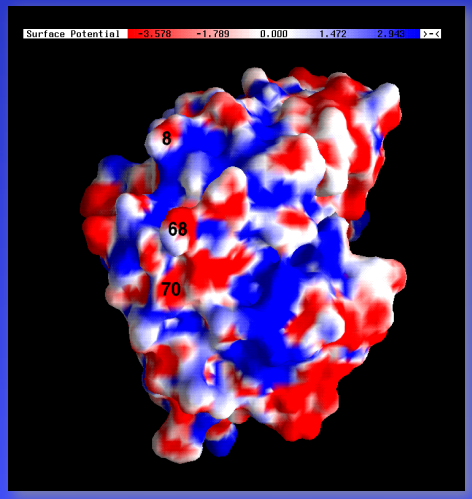


Huang *et al*. *J. Clin. Immunol*. **18**,169,1998.
Matsumoto *et al*. *J.Biol.Chem*. **270**,19524,1995.
Šali *et al*. *J. Biol. Chem*. **268**, 9023, 1993.





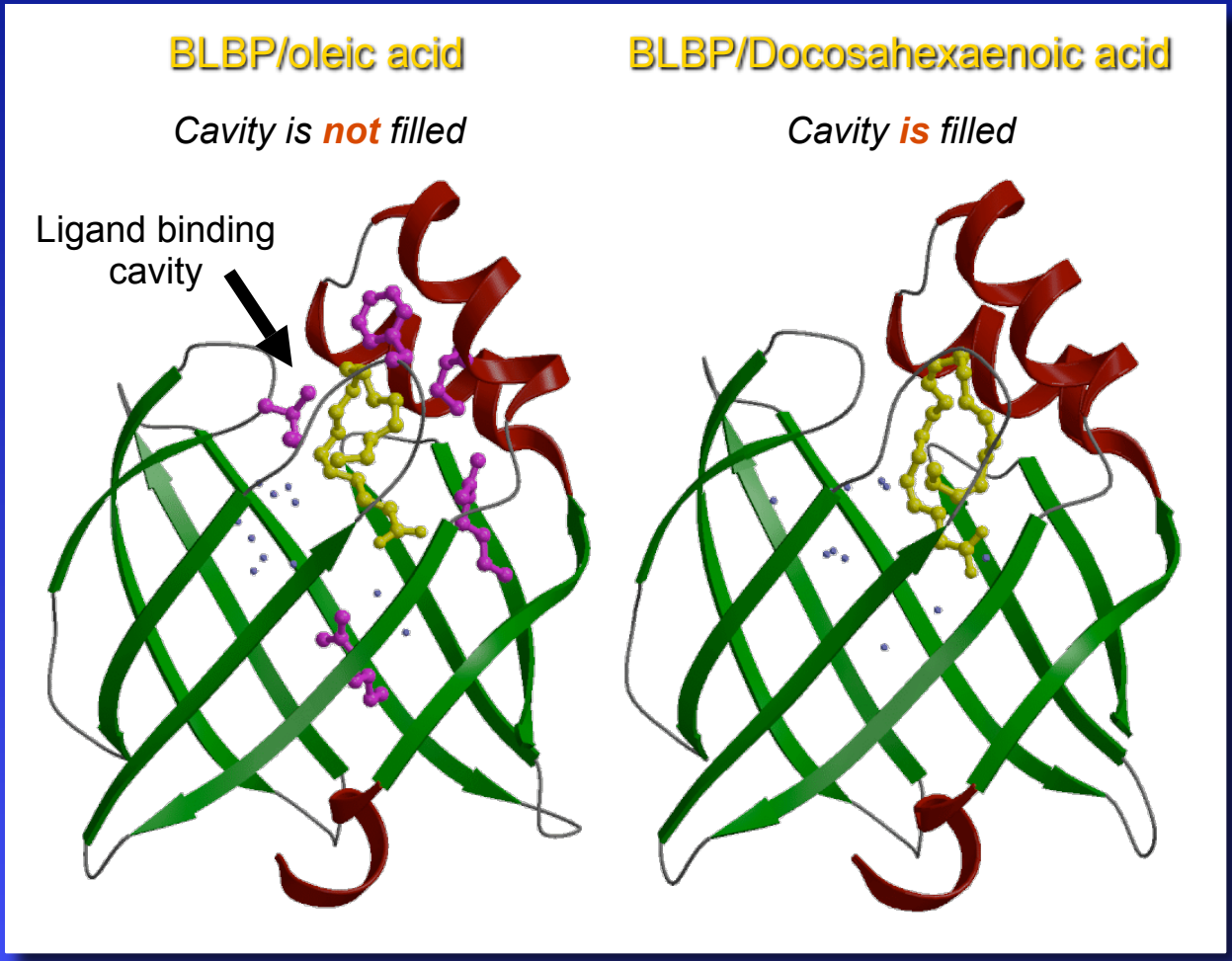
Native mMCP-7 at *p*H=5 (His^{+})



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

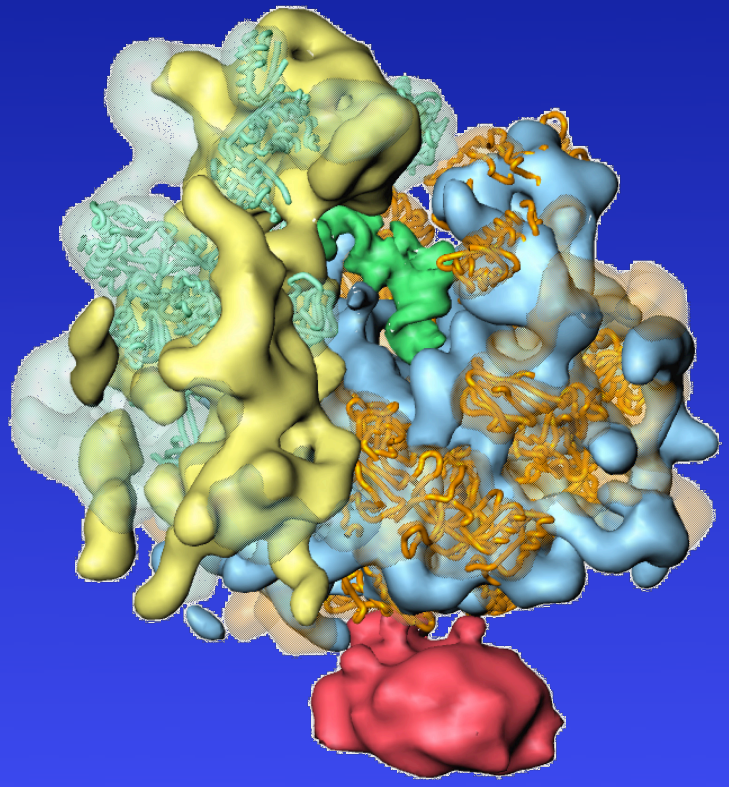


1. BLBP binds fatty acids.
2. Build a 3D model.
3. Find the fatty acid that fits most snuggly into the ligand binding cavity.

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

Some Models Can Be Used in Docking to Density Maps

(Yeast Ribosomal 40S subunit)

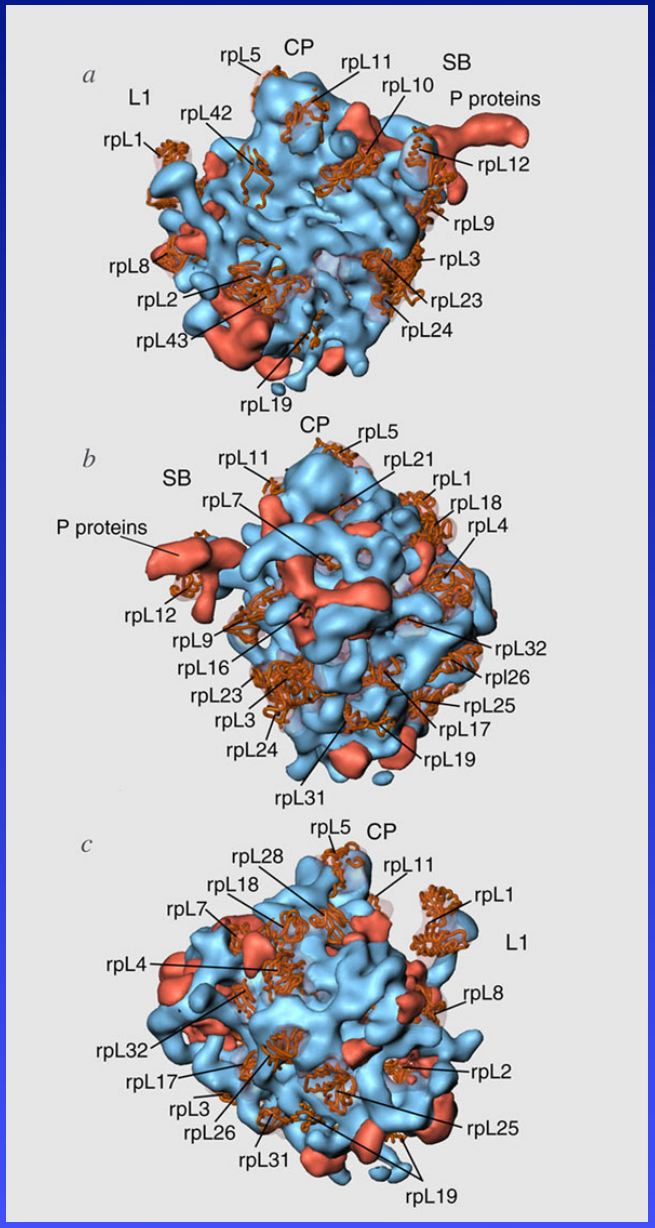


Small 30S subunit from *Thermus thermophilus*
Large 50S subunit from *Haloarcula marismortui*

Docking of comparative models into the cryo-EM map.

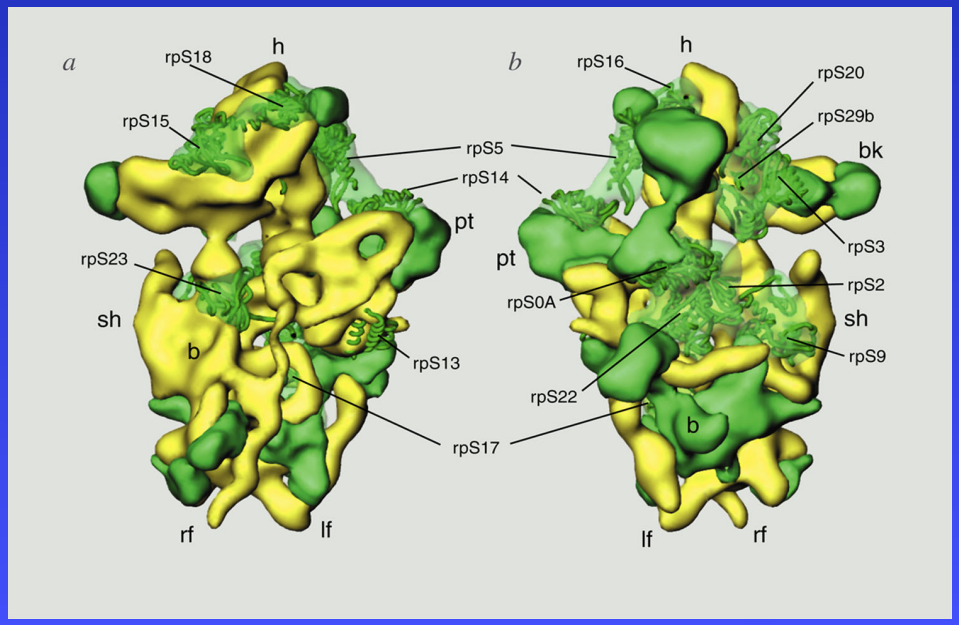
Spahn *et al.* 2001 Cell **107**:373-386

60S Subunit



- 43 proteins could be modeled on 20-56% seq.id. to a known structure.
- The coverage of the models ranges from 34-99%.
- Models were manually docked into the 15Å cryo-electron density map.
- The solid orange in the 60S subunit and the solid green in the 40S subunit correspond to proteins without known bacterial homologs.

40S Subunit

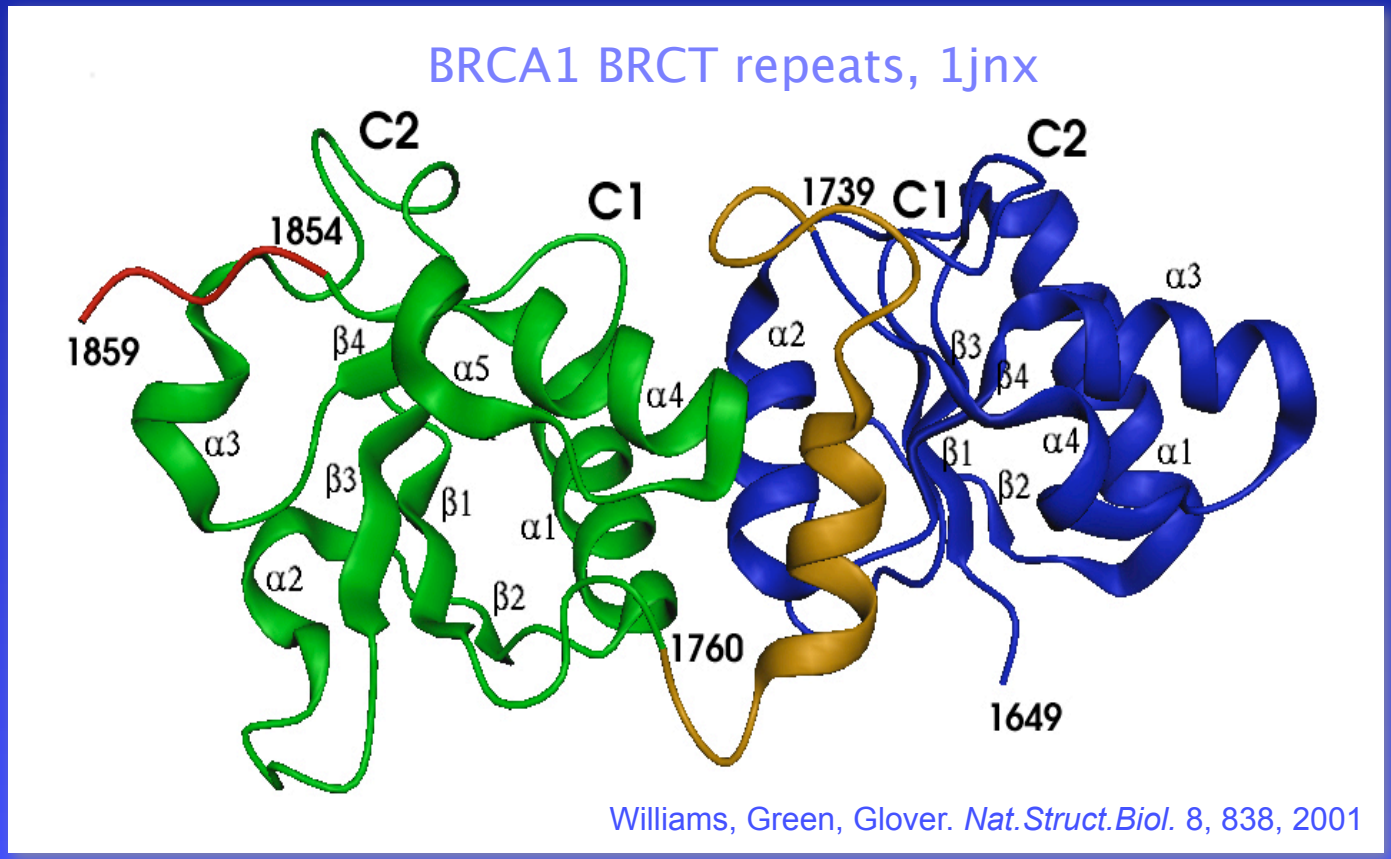


Structural analysis of missense mutations in human BRCA1 BRCT domains

Nebojsa Mirkovic, Marc A. Marti-Renom, Andrej Sali

Alvaro N.A. Monteiro (Sprang Center, Cornell U.)

Human BRCA1 and its two BRCT domains



CONFIDENTIAL

BRACAnalysis™

Comprehensive BRCA1-BRCA2 Gene Sequence Analysis Result

Strang Cancer Prevention Center
428 E 72nd St
New York, NY 10021

SPECIMEN	
Specimen Type:	Blood
Draw Date:	n/a
Accession Date:	Oct 27, 2000
Report Date:	Nov 17, 2000

PATIENT	
Name:	
Date of Birth:	Feb 02, 1953
Patient ID:	
Gender:	Female
Accession #:	00019998
Requisition #:	56694

Physician

Test Result

Gene Analyzed	Specific Genetic Variant
BRCA2	H2116R
BRCA1	None Detected

Interpretation

GENETIC VARIANT OF UNCERTAIN SIGNIFICANCE

The BRCA2 variant H2116R results in the substitution of arginine for histidine at amino acid position 2116 of the BRCA2 protein. Variants of this type may or may not affect BRCA2 protein function. Therefore, the contribution of this variant to the relative risk of breast or ovarian cancer cannot be established solely from this analysis. The observation by Myriad Genetic Laboratories of this particular variant in an individual with a deleterious truncating mutation in BRCA2, however, reduces the likelihood that H2116R is itself deleterious.

Authorized Signature:

Laboratory Director Medical Director

CONFIDENTIAL

BRACAnalysis™
Comprehensive BRCA1-BRCA2 Gene Sequence Analysis Result

Strang Cancer Prevention Center
428 E 72nd St
New York, NY 10021

Physician:

SPECIMEN	
Specimen Type:	Blood
Draw Date:	n/a
Accession Date:	Oct 27, 2000
Report Date:	Nov 17, 2000

PATIENT	
Name:	
Date of Birth:	Feb 02, 1953
Patient ID:	
Gender:	Female
Accession #:	00019998
Requisition #:	56694

Test Result

Gene Analyzed	Specific Genetic Variant
BRCA2	H2116R
BRCA1	None Detected

Interpretation

GENETIC VARIANT OF UNCERTAIN SIGNIFICANCE

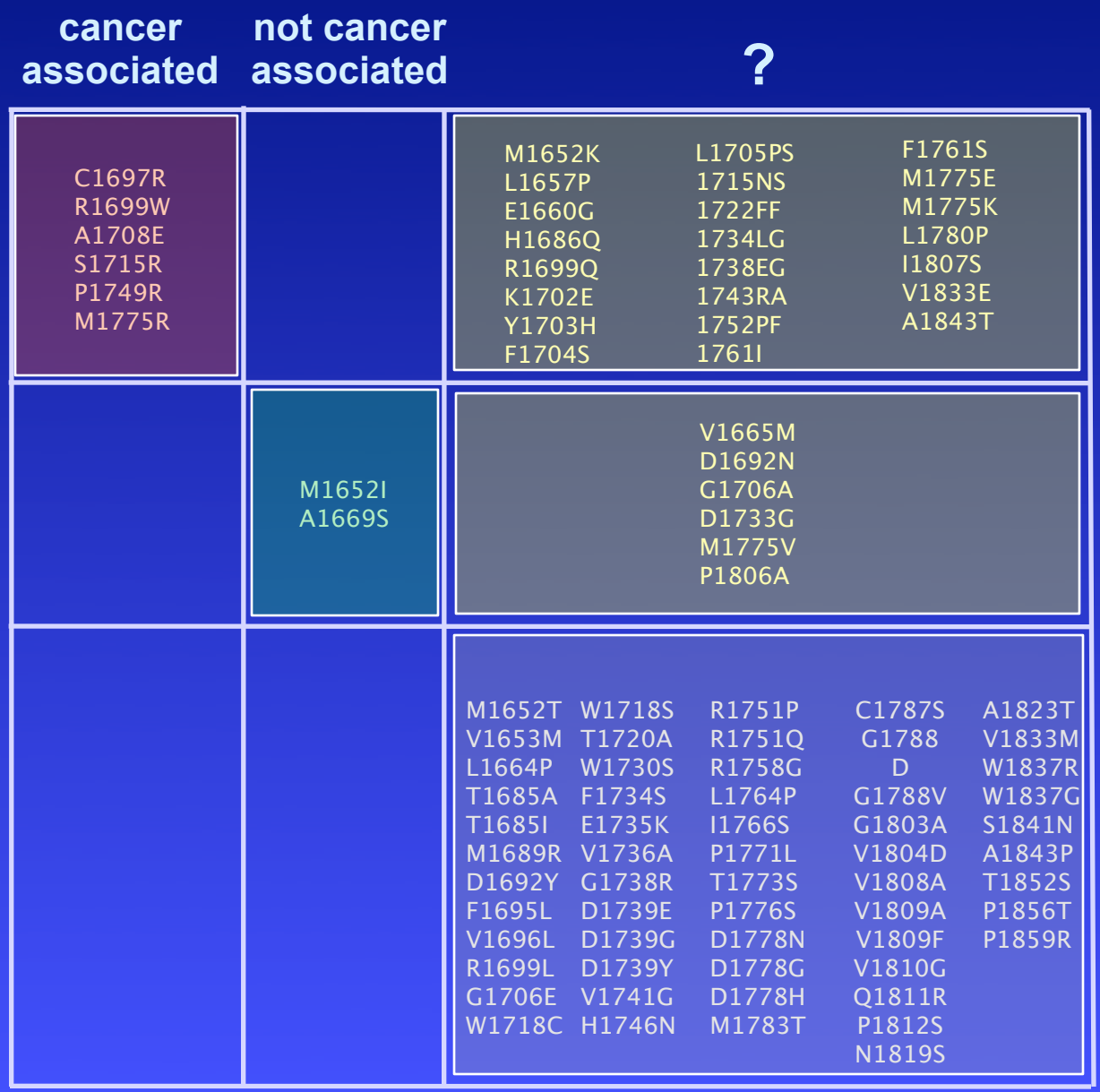
The BRCA2 variant H2116R results in the substitution of arginine for histidine at amino acid position 2116 of the BRCA2 protein. Variants of this type may or may not affect BRCA2 protein function. Therefore, the contribution of this variant to the relative risk of breast or ovarian cancer cannot be established solely from this analysis. The observation by Myriad Genetic Laboratories of this particular variant in an individual with a deleterious truncating mutation in BRCA2, however, reduces the likelihood that H2116R is itself deleterious.

Authorized Signature:

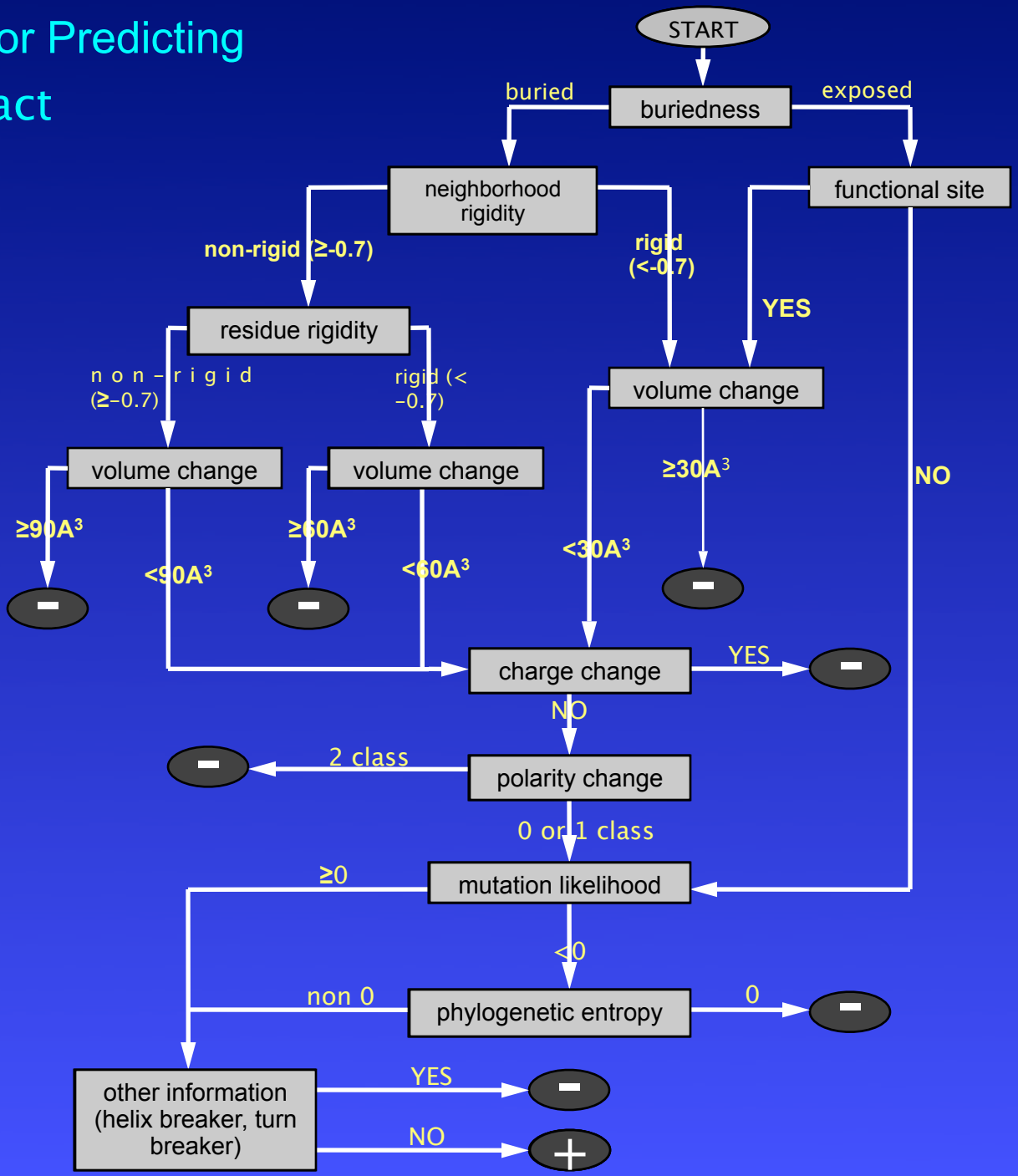
Laboratory Director Medical Director

These test results should only be used in conjunction with the patient's clinical history and any previous analysis of appropriate family members. It is strongly recommended that these results be communicated to the patient in a setting that includes appropriate counseling. The accompanying Technical Specifications summary describes the analysis, method, performance characteristics, nomenclature, and interpretive criteria of this test. This test may be considered investigational by some states. This test was developed and its performance characteristics determined by Myriad Genetic Laboratories. It has not been reviewed by the U.S. Food and Drug Administration. The FDA has determined that such clearance or approval is not necessary.

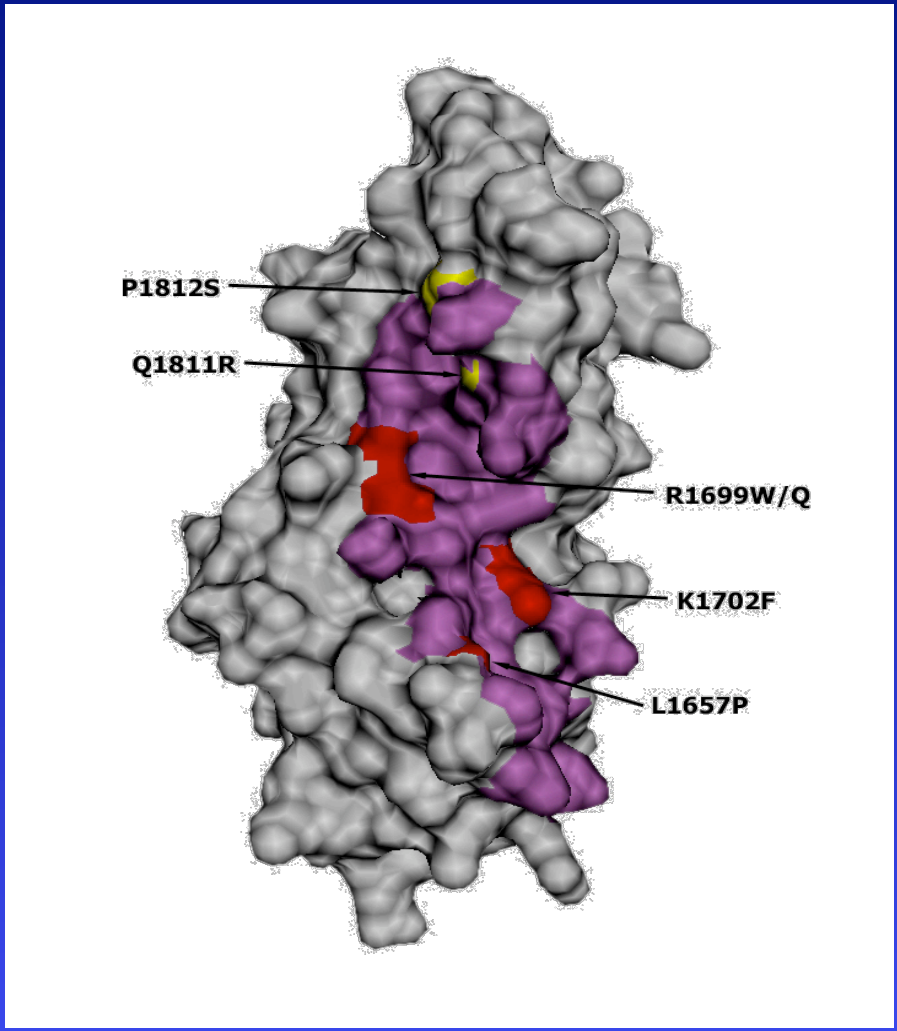
Missense Mutations in BRCT Domains by Function



"Decision" Tree for Predicting Functional Impact of Genetic Variants



Putative Binding Site on BRCA1



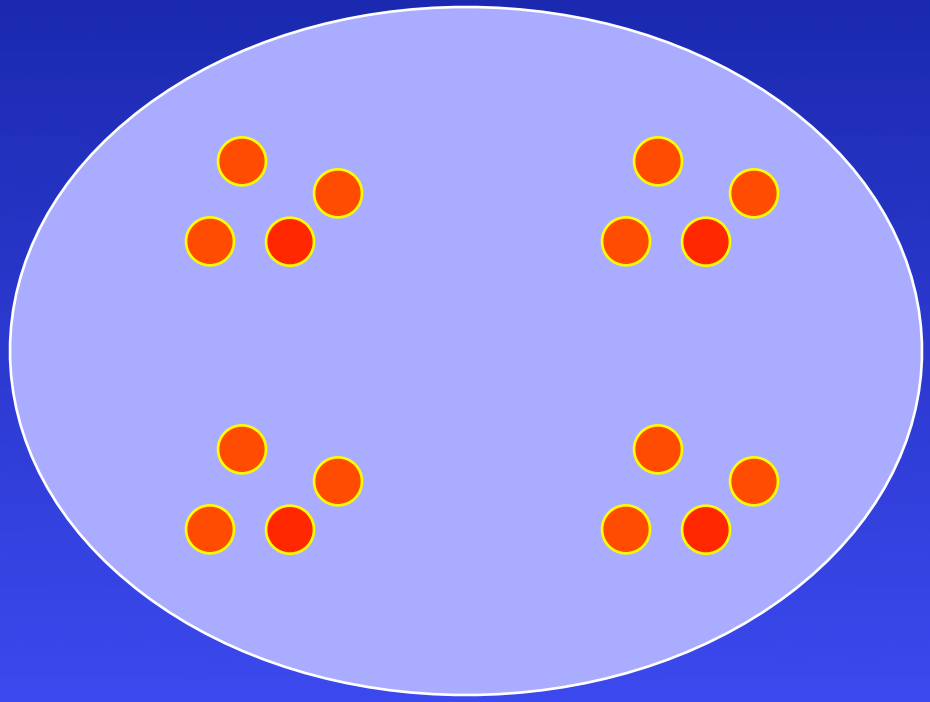
RMSMVVSGLTPEEFMLVYKFARKHHITLTNLITEETTHVVMKTDAEFVCERTLKYFLGIAGGKWVVSYFWVTQSIKERK
MLNEHDFEVRGDVVNGRNHQGPKRARESQDRKIFRGLEICCYGPFTNMPTDQLEWMVQLCGASVVKELSSFTLGTGVHP
IVVVQPDAWTEDNGFHAIGQMCEAPVVTREWVLDSVALYQCQELDTYLIPQIP

genomes…

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

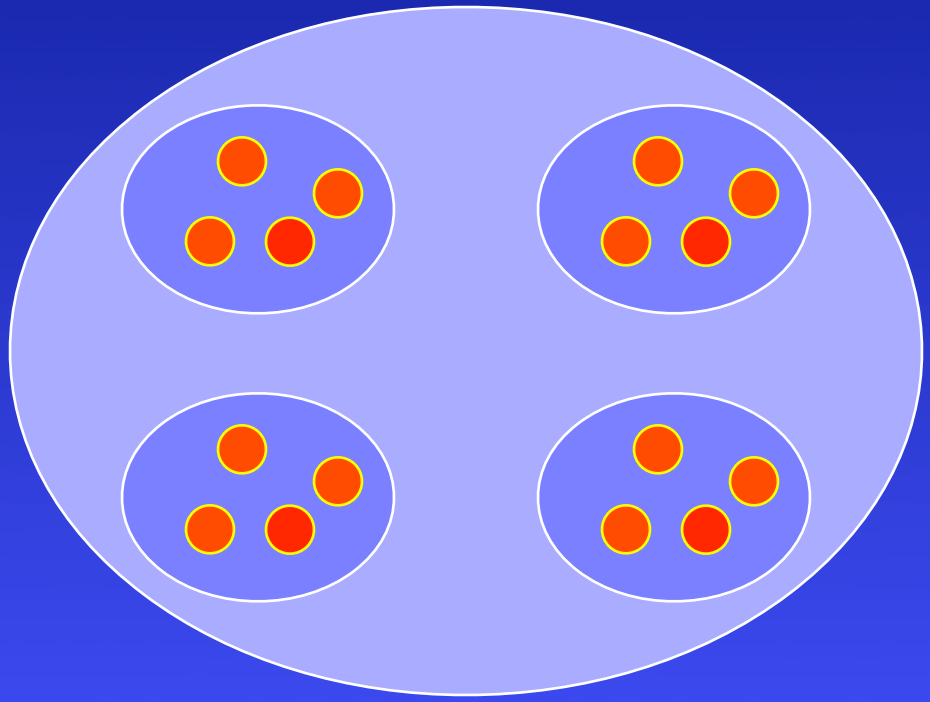
Characterize most protein sequences based on related known structures.



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

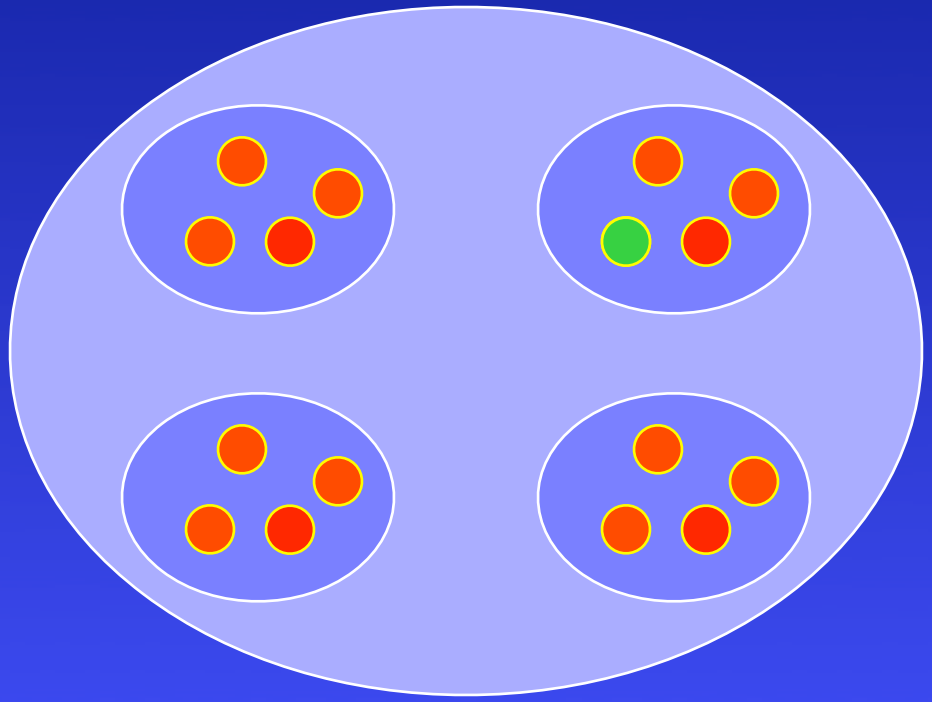
Characterize most protein sequences based on related known structures.



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

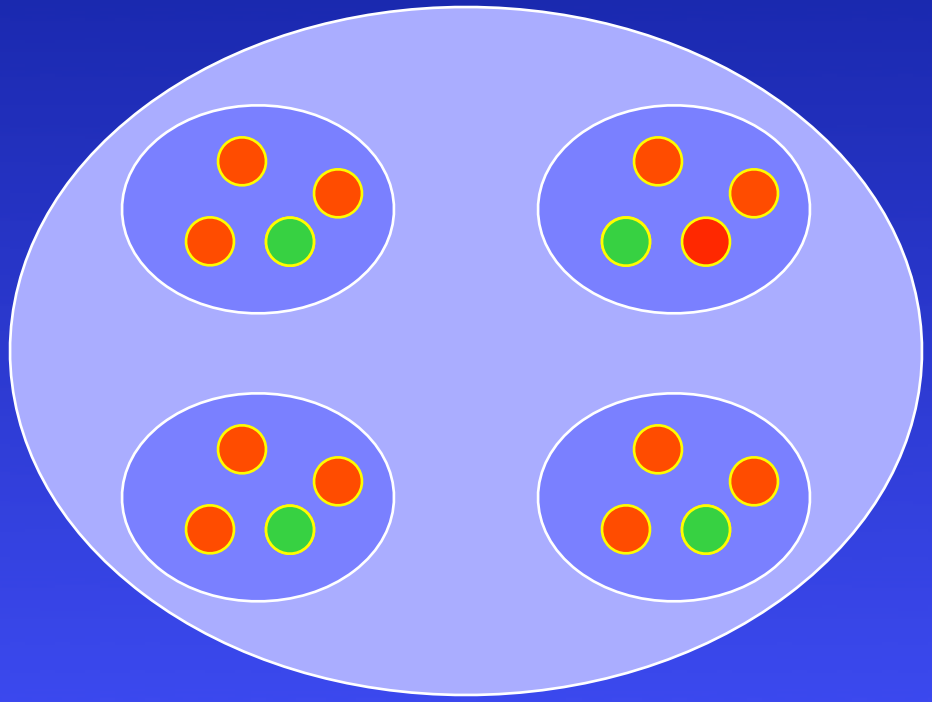
Characterize most protein sequences based on related known structures.



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

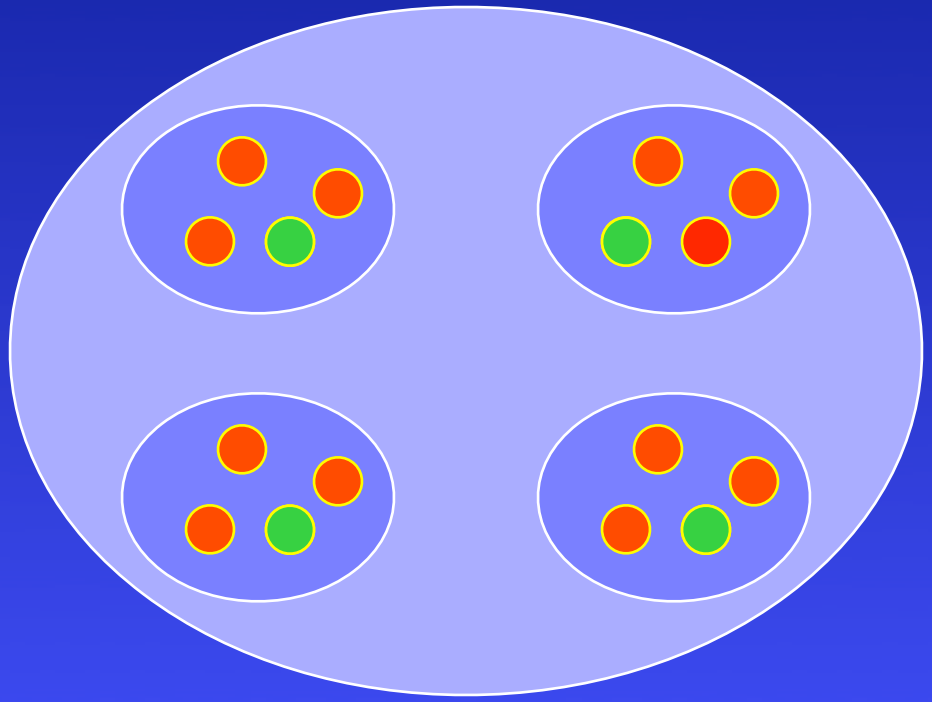
Characterize most protein sequences based on related known structures.



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

Characterize most protein sequences based on related known structures.



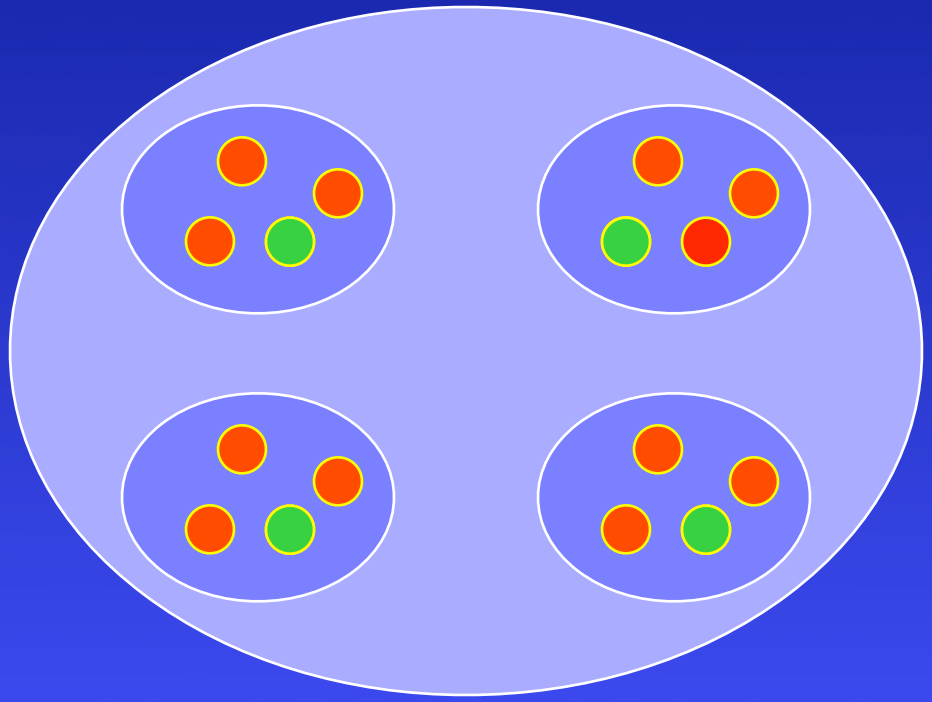
The number of “families” is much smaller than the number of proteins.

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

Characterize most protein sequences based on related known structures.



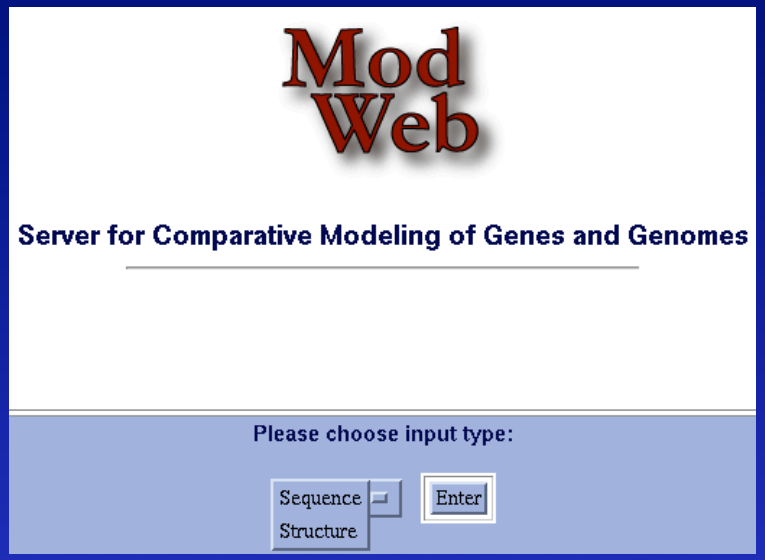
The number of "families" is much smaller than the number of proteins.

Any one of the members of a family is fine.

There are ~16,000 30% seq id families (90%)
(Vitkup *et al. Nat. Struct. Biol.* **8**, 559, 2001)

Modeling with NY-SGRC structures

http://salilab.org/modweb/

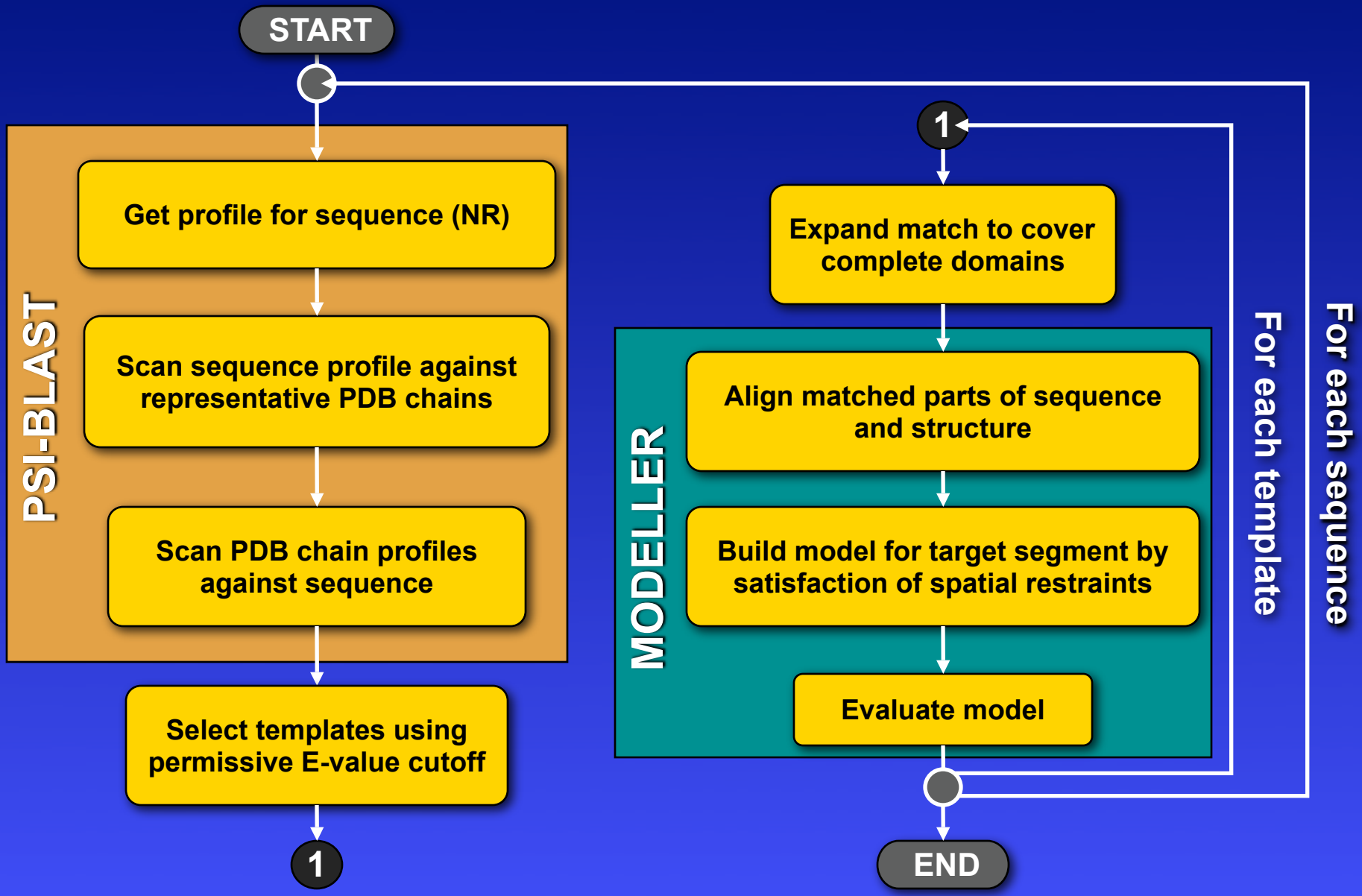


June 2001

Protein Name	GI or Swissprot Code	Length	# Acceptable Models	Min. Seq. ID	Max. Seq. ID	# Models >50% Seq. ID	# Models 30–50% Seq. ID	# Models <30% Seq. ID
Yeast hypothetical protein	P38197	230	46	21	45	0	35	11
PNP oxidase	P38075	205	43	18	58	1	33	9
Yeast hypothetical protein	P49954	271	234	15	54	4	40	190
T. maritima L-threonine acetaldehyde-lyase	GI 4982322	342	2098	9	51	2	27	2069
Hypothetical esterase	P40363	288	157	9	48	0	13	144
Hypothetical protein	P40165	214	35	19	86	1	6	28
mevalonate diphosphate decarboxylase	GI 1292890	391	121	10	68	2	18	101
yeast glutathione synthetase	GI 2198534	419	26	30	42	0	26	0

Bonanno *et al*. *Proc.Natl.Acad.Sci.USA* **98**, 12896, 2001.
Chance *et al*. Protein Science **11**, 723, 2002.

MODPIPE: Large-Scale Comparative Protein Structure Modeling



R. Sánchez & A. Šali, *Proc. Natl. Acad. Sci. USA* **95**, 13597, 1998.
N. Eswar, M. Marti-Renom, M.S. Madhusudhan, B. John, A. Fiser, R. Sánchez, F. Melo, N. Mirkovic, A. Šali.

Comparative modeling of the TrEMBL database

Unique sequences processed: 733,239

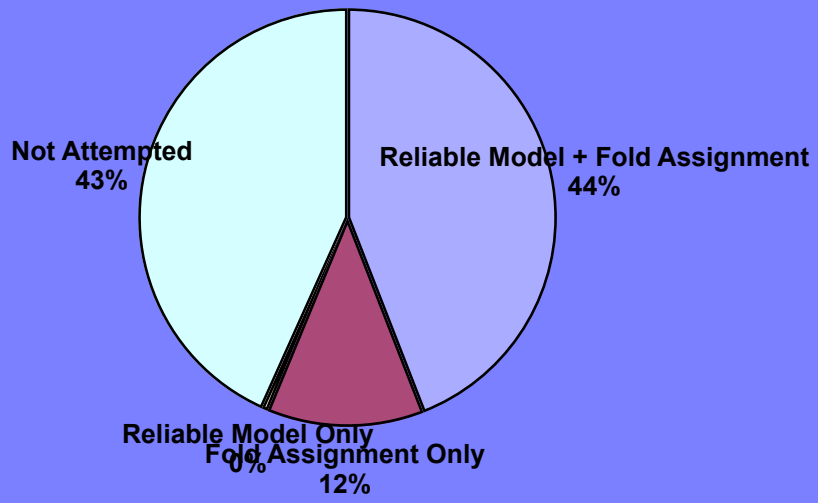
Sequences with fold assignments or models: 415,937 (57%)

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).

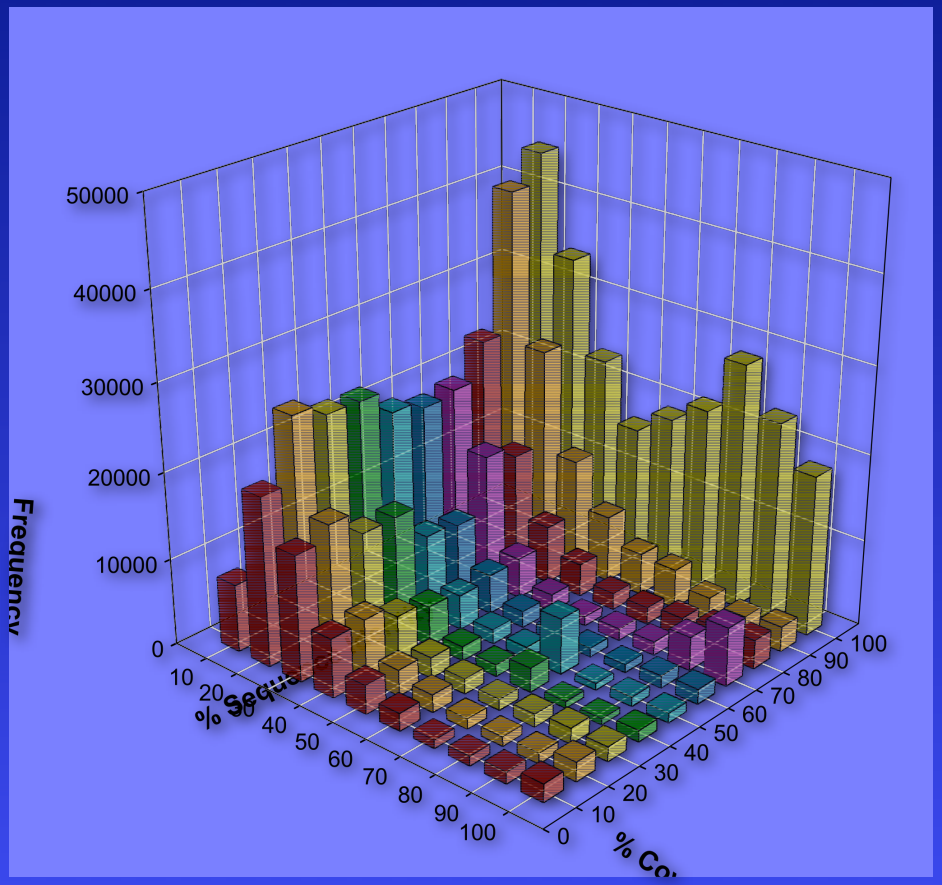
4/03/02 ~4 weeks on 500 Pentium III CPUs

Modeling Coverage of the Sequence Space



Fold assignment: PSI-BLAST E-value $\leq 10^{-4}$

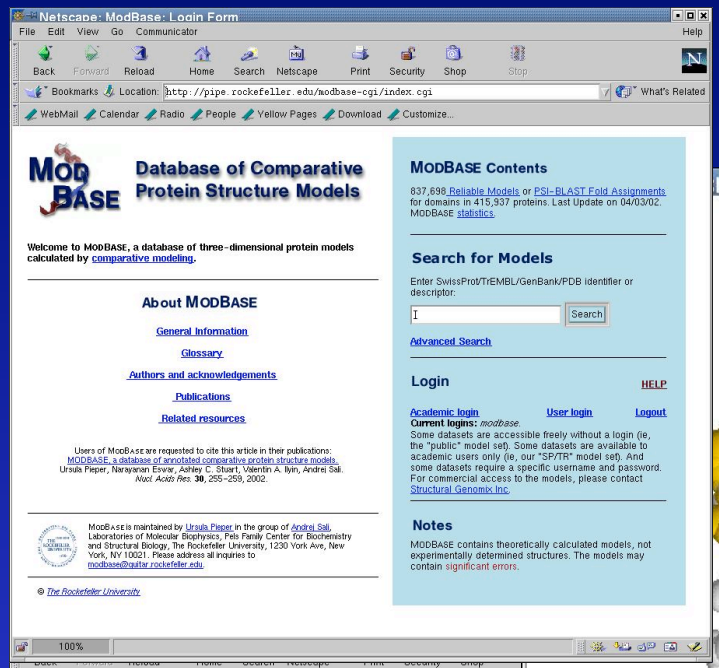
Reliable Model: Model Score ≥ 0.7

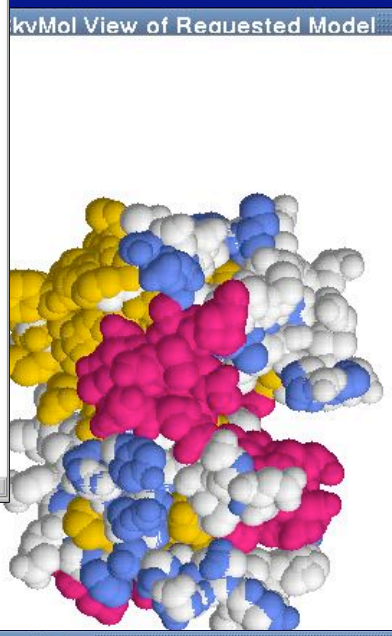


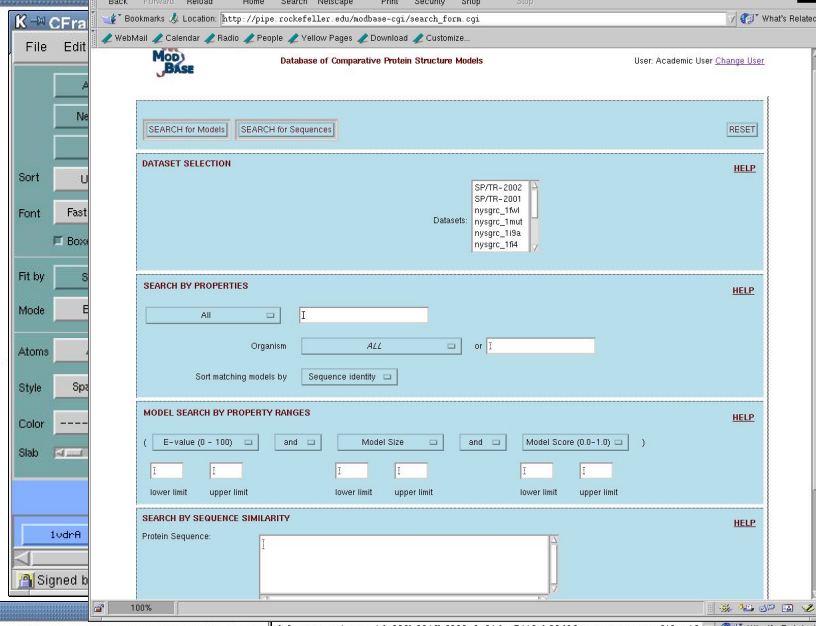
http://salilab.org/modbase

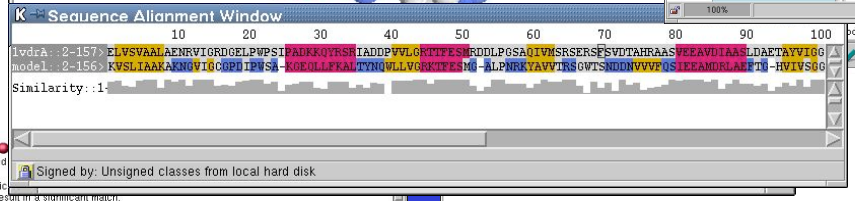
Pieper *et al.*, Nucl. Acids Res. 2002.

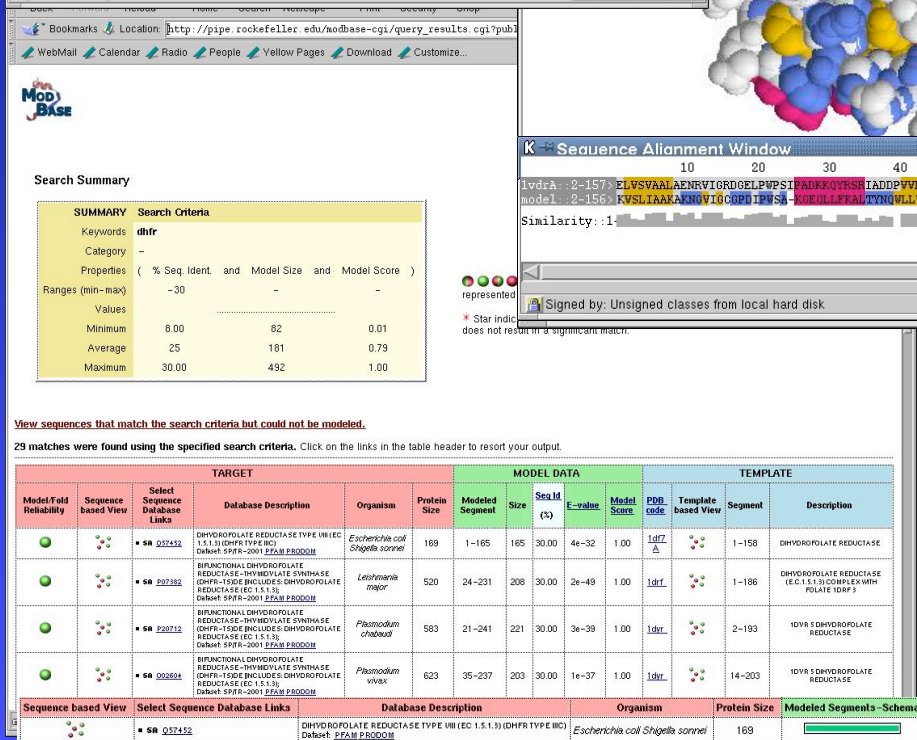
8/9/02

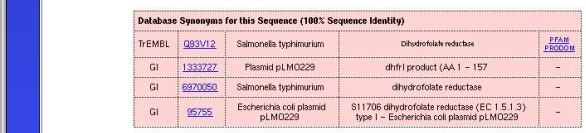


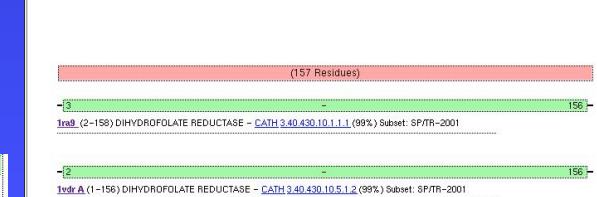


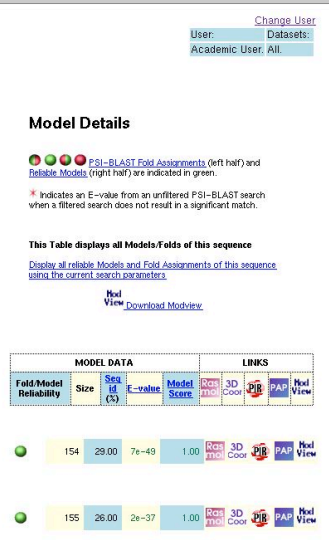












Conclusions

- ✓ At present, useful 3D models can be obtained for domains in ~ 55% of the proteins (25% of domains).
- ✓ Sampling at >30% sequence identity level.
- ✓ Completeness in structural coverage.
- ✓ Application to biological problems.

Acknowledgments

Andrej Sali

Frank Alber
Fred Davis
Damien Devos
Narayanan Eswar
Rachel Karchin
Libusha Kelly
Michael F. Kim
Dmitry Korkin
M. S. Madhusudhan
Nebosja Mirkovic
Ursula Pieper
Andrea Rossi
Min-yi Shen
Maya Topf

http://www.salilab.org