

Comparative protein structure modeling of genes and genomes

Marc A. Marti-Renom

Department of Biopharmaceutical Sciences
University of California, San Francisco

Comparative protein structure modeling of genes and genomes

Marc A. Marti-Renom

Department of Biopharmaceutical Sciences
University of California, San Francisco

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

Theory



Experiment

Why protein structure prediction?

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

Theory



Experiment

Why protein structure prediction?

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

Theory



Experiment

Principles of Protein Structure

The principles of protein structure are fundamental to understanding how proteins function in living organisms.

Proteins are large molecules composed of amino acids.

The primary structure of a protein is its linear sequence of amino acids.

The secondary structure refers to the local folding of the polypeptide chain, forming alpha-helices or beta-sheets.

The tertiary structure is the overall three-dimensional conformation of the protein, determined by interactions between distant parts of the chain.

The quaternary structure describes the arrangement of multiple polypeptide chains within a single protein complex.

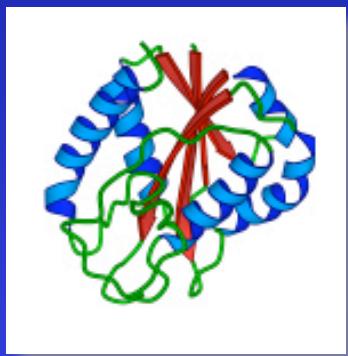
Protein structure is maintained by various non-covalent interactions, including hydrogen bonding, hydrophobic interactions, and ionic bonding.

The principles of protein structure have led to significant advances in biotechnology, including the design of new enzymes and the development of pharmaceuticals.

Understanding protein structure is crucial for addressing many important biological and medical questions.

Principles of Protein Structure

GFCHIKAYTRLIMVG...

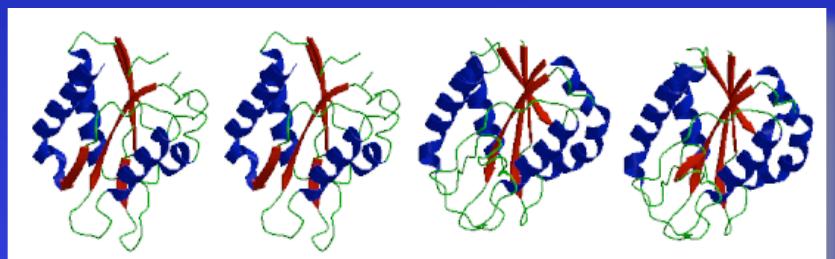
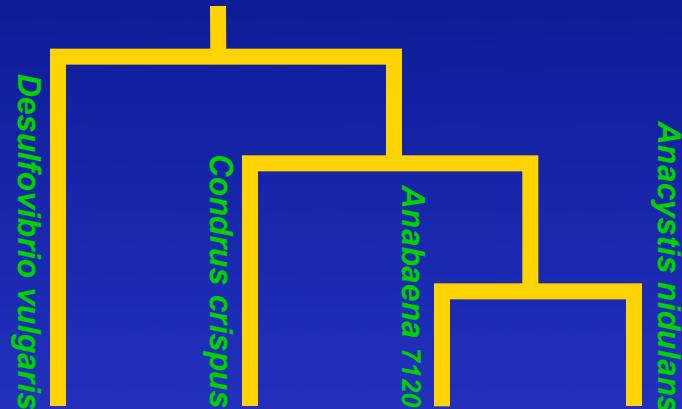
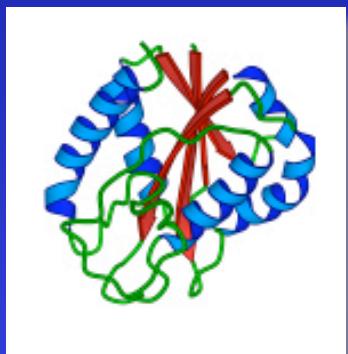


Folding

Ab initio prediction

Principles of Protein Structure

GFCHIKAYTRLIMV...



Folding

Ab initio prediction

Evolution

Threading
Comparative Modeling

Comparative Modeling by Satisfaction of Spatial Restraints (MODELLER)

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

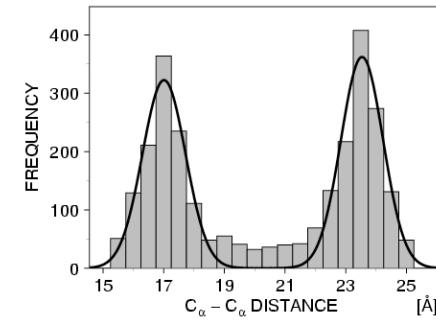
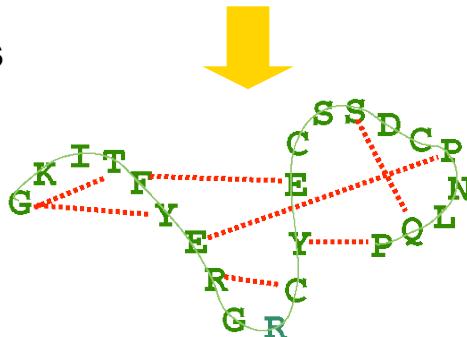
A. Šali & T. Blundell. *J. Mol. Biol.* **234**, 779, 1993.
J.P. Overington & A. Šali. *Prot. Sci.* **3**, 1582, 1994.
A. Fiser, R. Do & A. Šali. *Prot Sci.* **9**, 1753, 2000.

<http://salilab.org/modeller>

Comparative Modeling by Satisfaction of Spatial Restraints (MODELLER)

3D GKTFYERGFQGHHCYESDC-NLQP...
SEQ GKTFYERG---RCYESDCPNLQP...

1. Extract spatial restraints



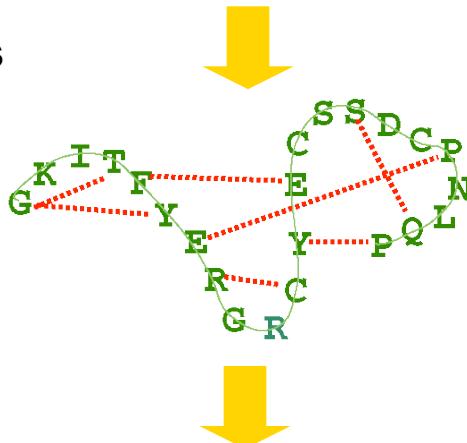
- A. Šali & T. Blundell. *J. Mol. Biol.* **234**, 779, 1993.
J.P. Overington & A. Šali. *Prot. Sci.* **3**, 1582, 1994.
A. Fiser, R. Do & A. Šali. *Prot Sci.* **9**, 1753, 2000.

<http://salilab.org/modeller>

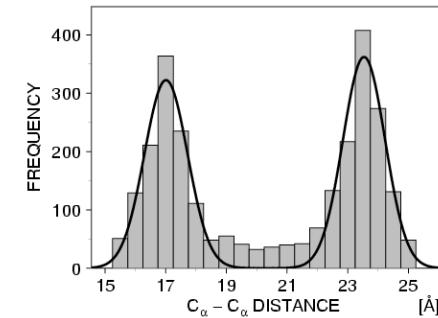
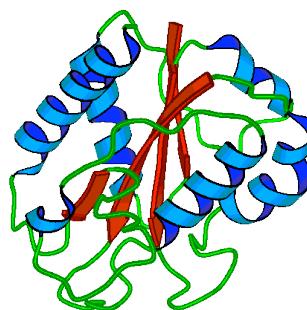
Comparative Modeling by Satisfaction of Spatial Restraints (MODELLER)

3D GKTFYERGFQGHHCYESDC-NLQP...
SEQ GKTFYERG---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.* **9**, 1753, 2000.

<http://salilab.org/modeller>

Steps in Comparative Protein Structure Modeling

START

TARGET

ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVLKIDD
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



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

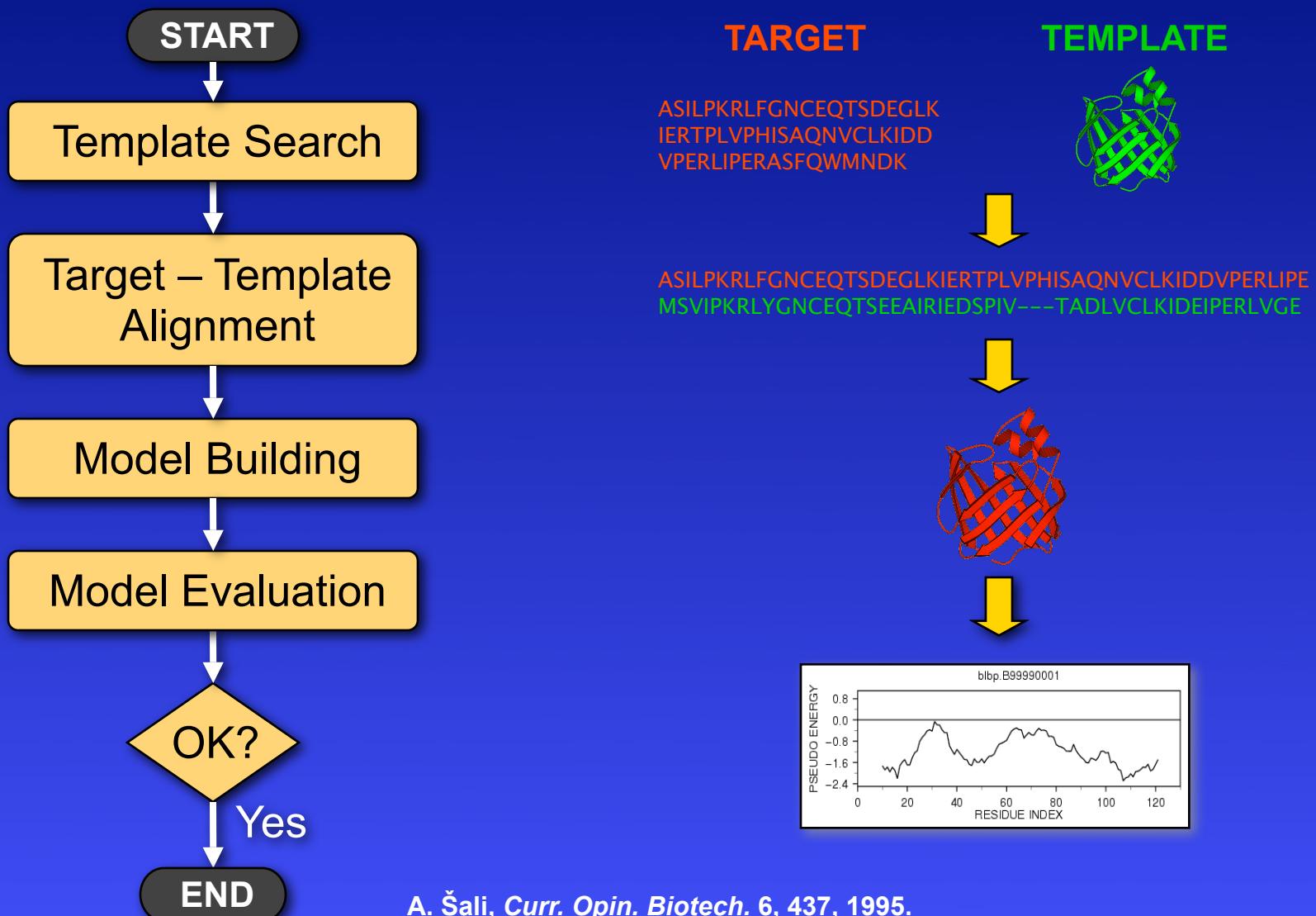


A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.

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

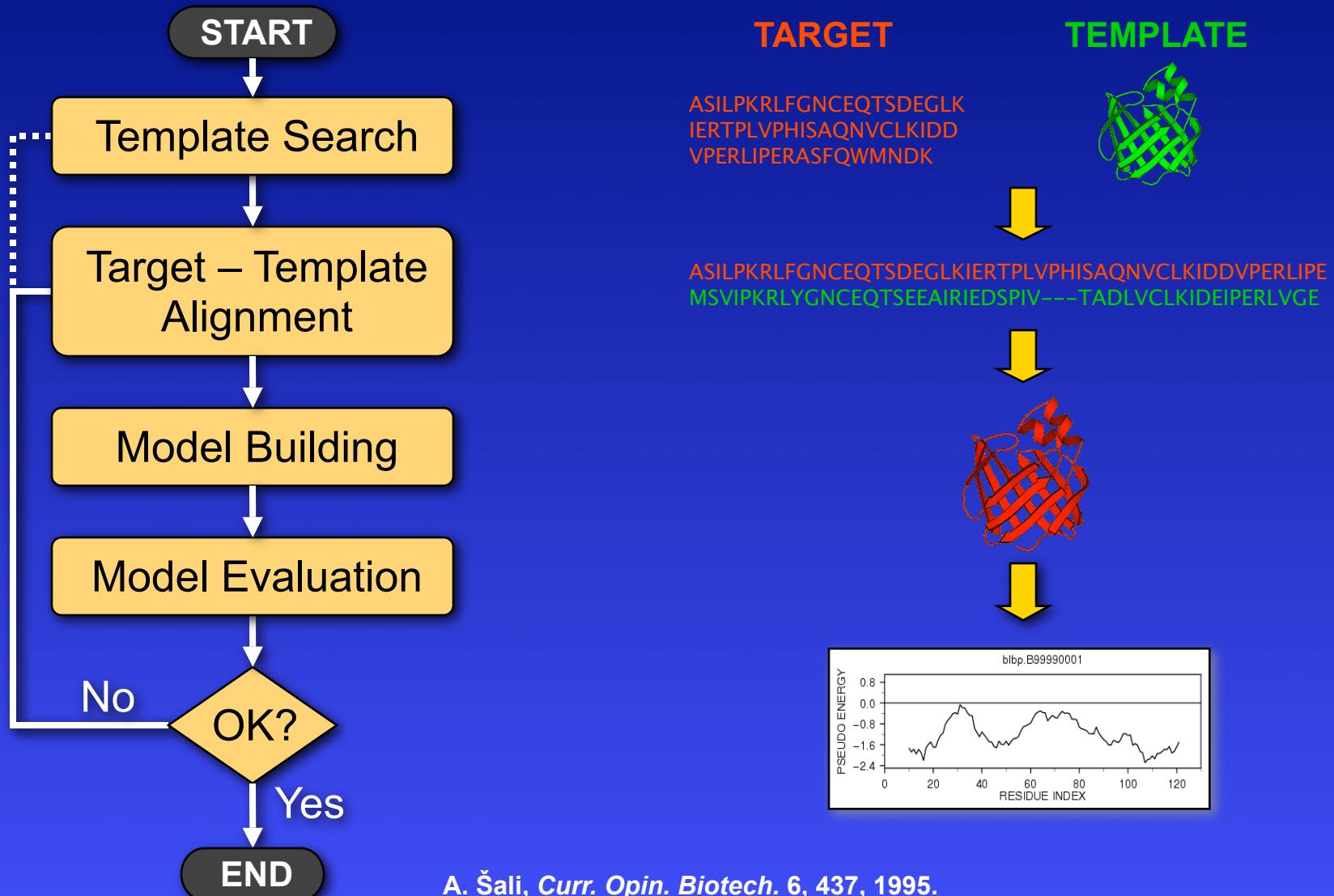


A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.

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

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

Steps in Comparative Protein Structure Modeling

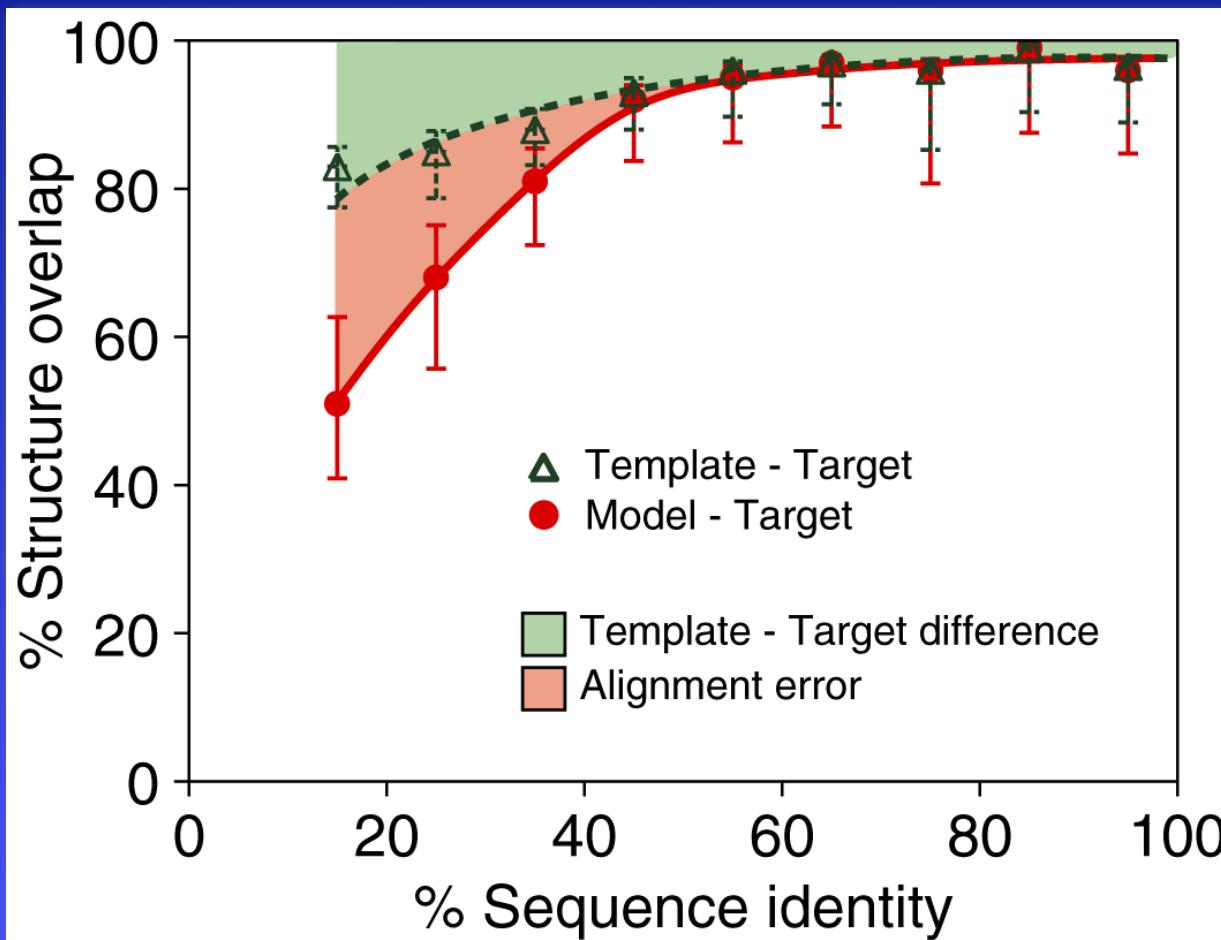


A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.

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

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

Model Accuracy as a Function of Target-Template Sequence Identity



Typical Errors in Comparative Models

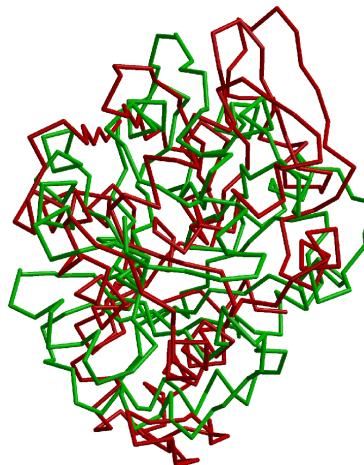
Typical Errors in Comparative Models

MODEL

X-RAY

TEMPLATE

Incorrect template



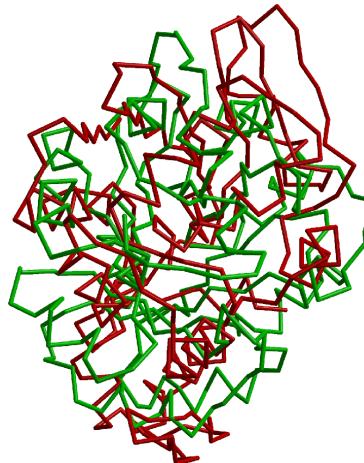
Typical Errors in Comparative Models

MODEL

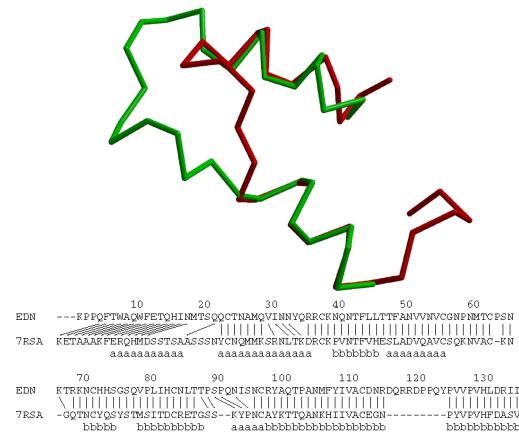
X-RAY

TEMPLATE

Incorrect template



Misalignment



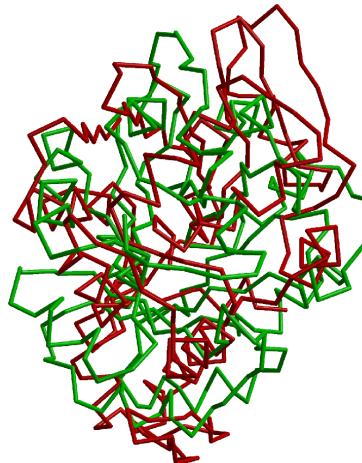
Typical Errors in Comparative Models

MODEL

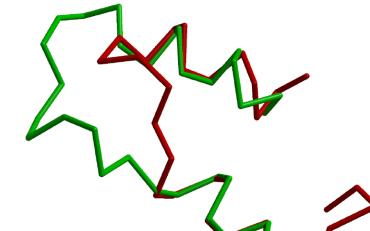
X-RAY

TEMPLATE

Incorrect template



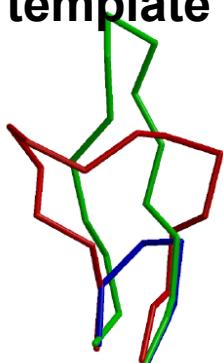
Misalignment



EDN	---	KPPQFTWAAQWFETQHINMTSQQCTNAMQVINNYQRCKHQNTFLLTTFANVVNCGNPNMTCPSN
7RSA	KETAAKFERQHMDSSTSAASSNNYCNQMKSRNLTKDRCKFVNTPVHESLADVQAVCQRNVAC-KN	
	aaaaaaaaaaaa	aaaaaaaaaaaa
	aaaaaaaaaaaa	bbbbbbbbb
	aaaaaaaaaaaa	aaaaaaaaaaa
EDN	70	-----KTKRKHHSQGVPLIHGNLTTESPQNISNCRYAQTFANNFYIVACDNRQRRDPFQYFWPVHLDRII
7RSA	80	-GQTCVQSYSTMSTTDCEETGS-----KYPNCAYKTTQANKHIIVACEGN-----PVPVPHFDASV
	90	bbbybb
	100	bbbbbbbbb
	110	aaaaabbbbbb
	120	bbbbb
	130	bbbbb

Region without a

template



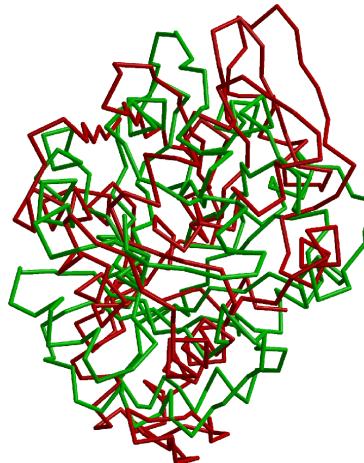
Typical Errors in Comparative Models

MODEL

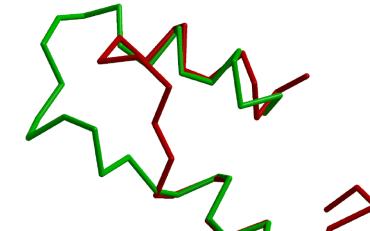
X-RAY

TEMPLATE

Incorrect template



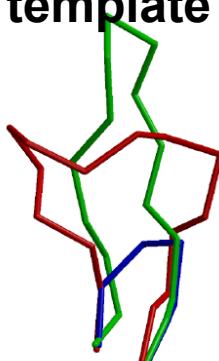
Misalignment



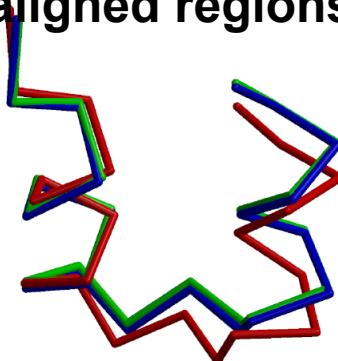
EDN	---	KPPQFTWAAQWFETQHINMTSQQCTNAMQVINNYQRRCKHQNTFLLTTFANVVNCGNPNMTCPSN
7RSA	KETAAFFERQHMDSSTSAASSNNYCNQMKSRNLTKDRCKFVNTPVHESLADVQAVCQRNVAC-KN	
	aaaaaaaaaaaaaa	
	aaaaaaaaaaaaaa	
	bbbbbbbbb	
	aaaaaaaaaaaaaa	

EDN	70	EKKKCHHSQGVPLIHGNLTTESPQNISNCRYAQTFANNFYIVACDNRQQRDPFQYFWPVHLDRII
7RSA	80	-GQTCVQSYSTMSITTDCEETGS--KYPNCAYKTTQANKHIIVACEGN-----PVPVPHFDASV
	90	bbbbb
	100	bbbbb
	110	aaaaaabbbbbb
	120	bbbbb
	130	bbbbb

Region without a
template



Distortion in correctly
aligned regions



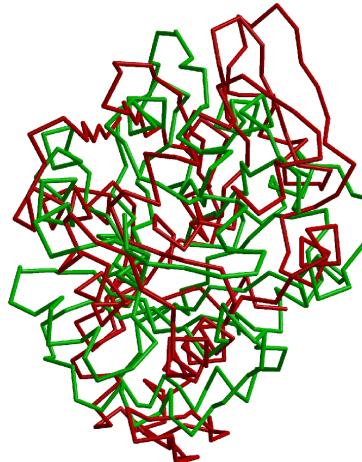
Typical Errors in Comparative Models

MODEL

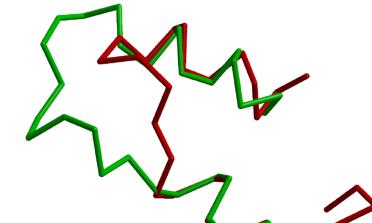
X-RAY

TEMPLATE

Incorrect template



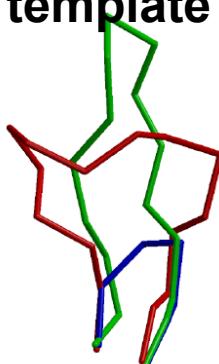
Misalignment



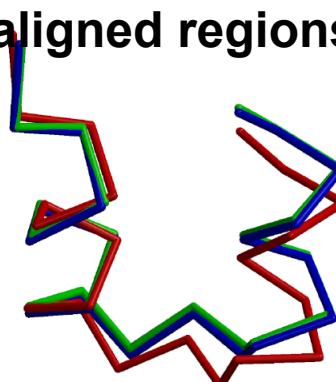
EDN ---KPPQFTWAAQMFWTQHINMTSQQCTNAMQVINNNYQRCKHQNTFLLTFFANVVNCGNPNMTCPSN
7RSA KETAAKFPERQHMDSSTSAAASSSSNYCNQMMKSRNLTKDRCKFVNITPVIESLADVQAVCQQRNVAC-KN
aaaaaaaaaaaaaaa
aaaaaaaaaaaaaa bbbbbbbb aaaaaaaaaa

EDN ETKRNCHHSQGVPLIHGNLTTESPQNISNCRVQAQTANMFIVVACCDNRQQRDPFQYFWPVPHLDRII
7RSA -GQTCVQSYSTMSTTDCEETGS--KYPNCAYKTTQANKHIIVACEGN-----PWFVPHFDASV
bbbbb bbbbbbbbbb aaaaaaaaaaaaaaaa bbbbbbbbbb

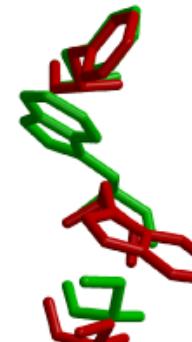
Region without a template



Distortion in correctly aligned regions



Sidechain packing

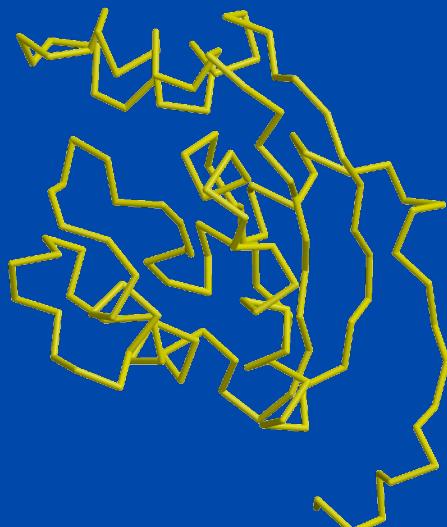


Model Accuracy

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

HIGH ACCURACY

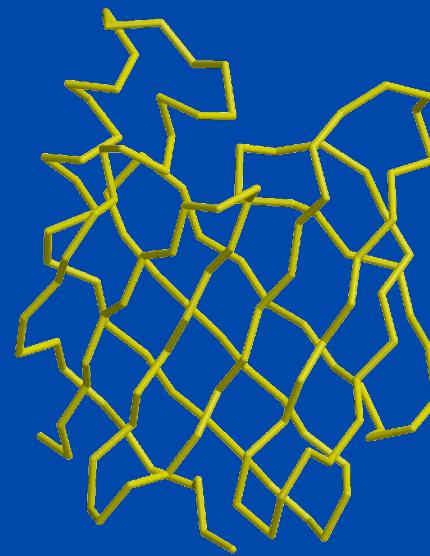
NM23
Seq id 77%



X-RAY

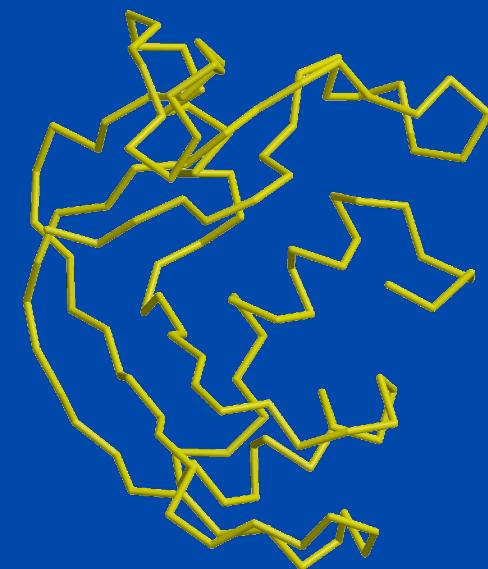
MEDIUM ACCURACY

CRABP
Seq id 41%



LOW ACCURACY

EDN
Seq id 33%

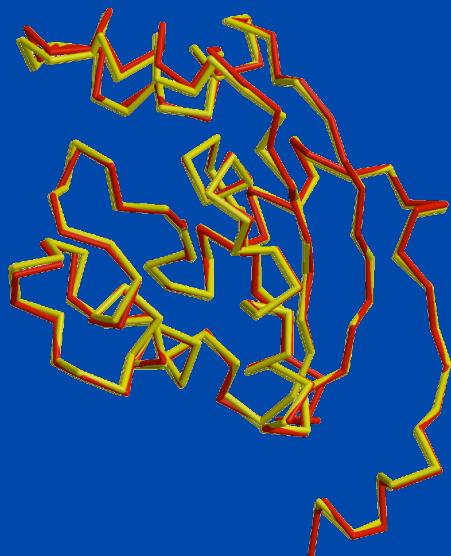


Model Accuracy

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

HIGH ACCURACY

NM23
Seq id 77%
 $C\alpha$ equiv 147/148
RMSD 0.41 Å

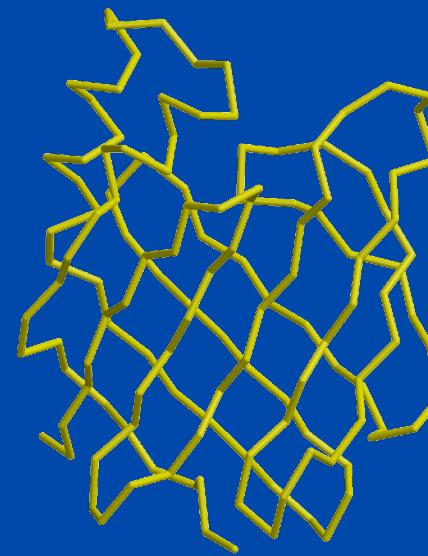


Sidechains
Core backbone
Loops

X-RAY / MODEL

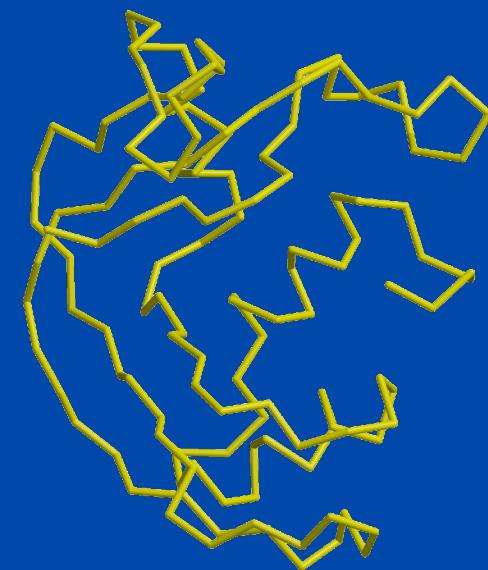
MEDIUM ACCURACY

CRABP
Seq id 41%



LOW ACCURACY

EDN
Seq id 33%

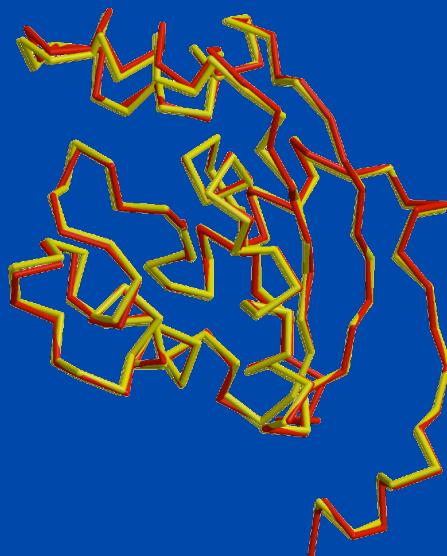


Model Accuracy

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

HIGH ACCURACY

NM23
Seq id 77%
 $C\alpha$ equiv 147/148
RMSD 0.41Å

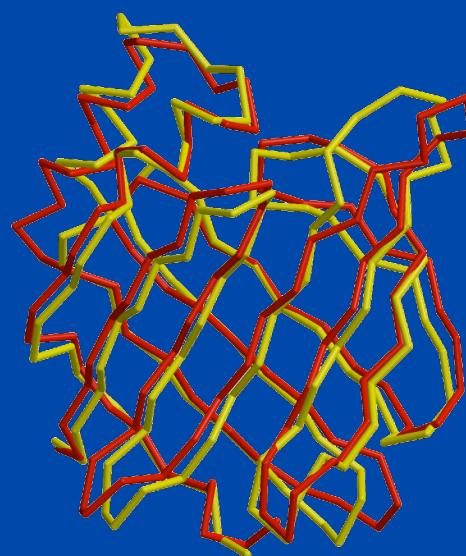


Sidechains
Core backbone
Loops

X-RAY / MODEL

MEDIUM ACCURACY

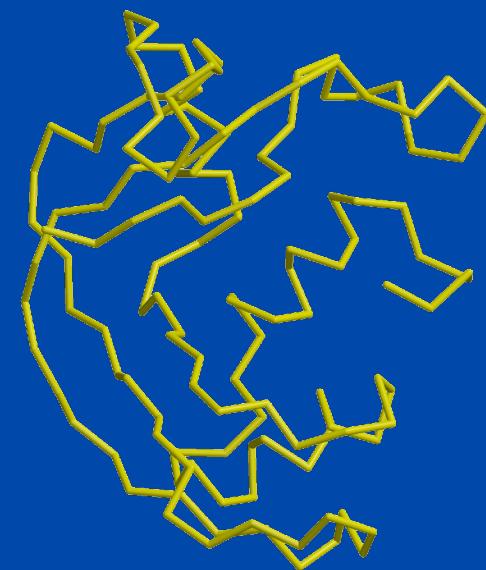
CRABP
Seq id 41%
 $C\alpha$ equiv 122/137
RMSD 1.34Å



Sidechains
Core backbone
Loops
Alignment

LOW ACCURACY

EDN
Seq id 33%

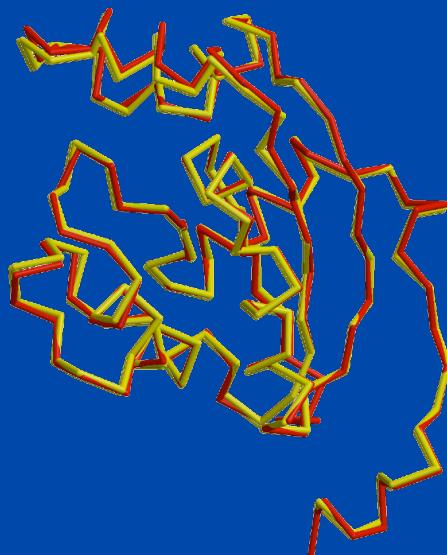


Model Accuracy

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

HIGH ACCURACY

NM23
Seq id 77%
 $C\alpha$ equiv 147/148
RMSD 0.41Å

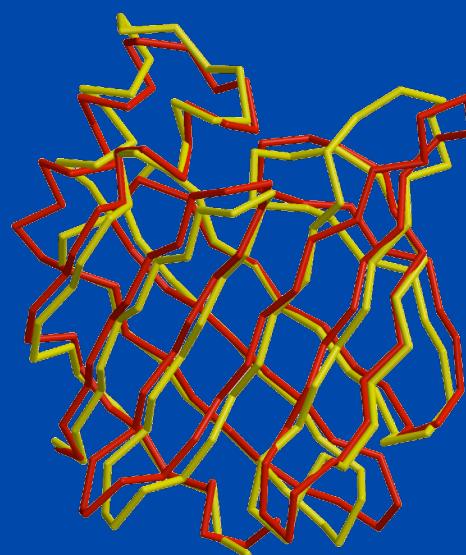


Sidechains
Core backbone
Loops

X-RAY / MODEL

MEDIUM ACCURACY

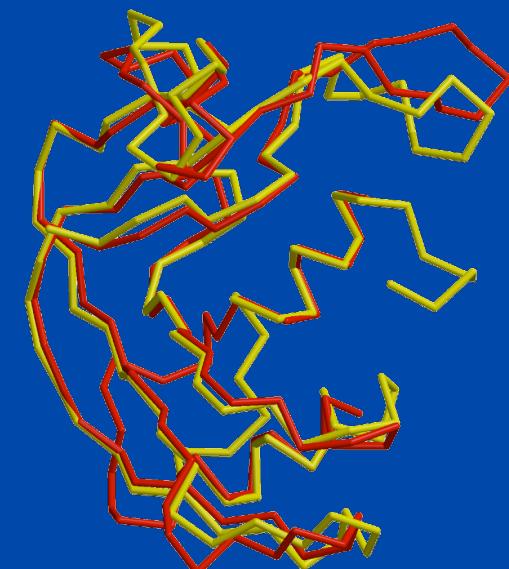
CRABP
Seq id 41%
 $C\alpha$ equiv 122/137
RMSD 1.34Å



Sidechains
Core backbone
Loops
Alignment

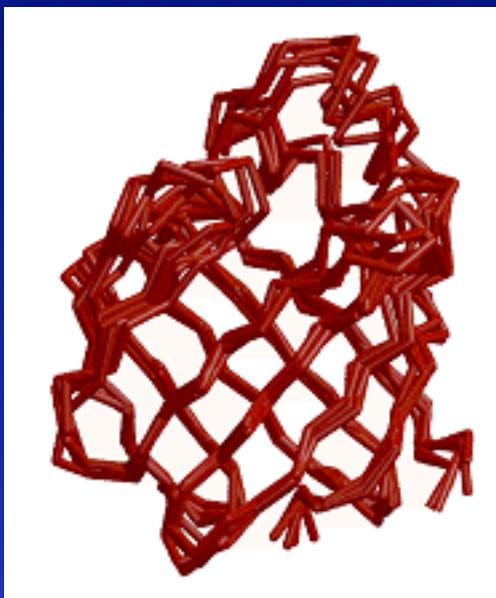
LOW ACCURACY

EDN
Seq id 33%
 $C\alpha$ equiv 90/134
RMSD 1.17Å



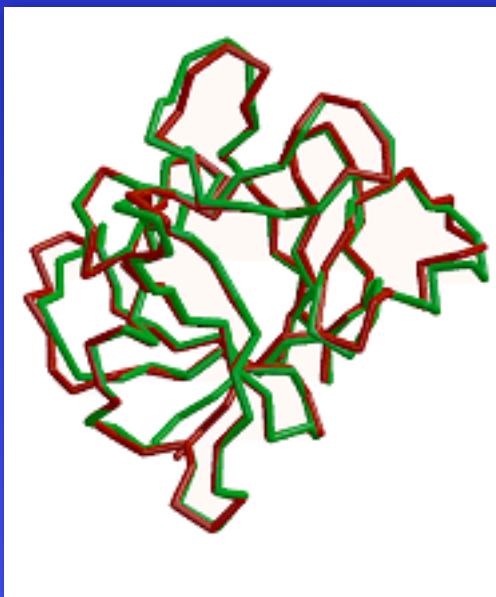
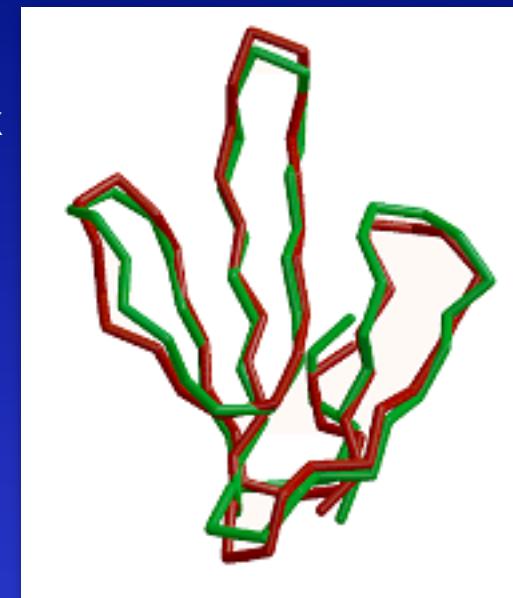
Sidechains
Core backbone
Loops
Alignment
Fold assignment

“Biological” significance of modeling errors



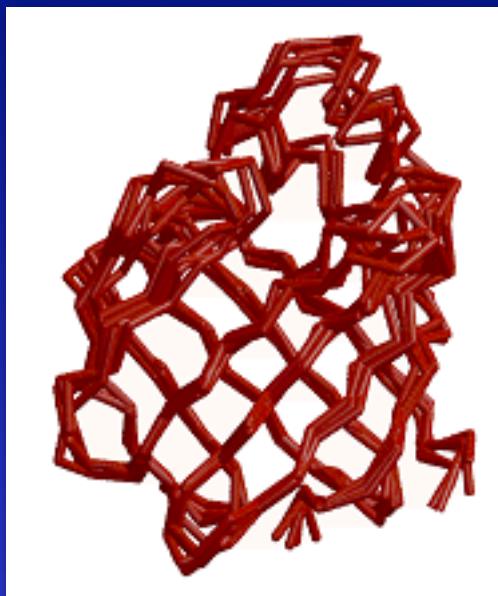
NMR
Ileal lipid-binding protein
1eal

NMR – X-RAY
Erabutoxin 3ebx
Erabutoxin 1era



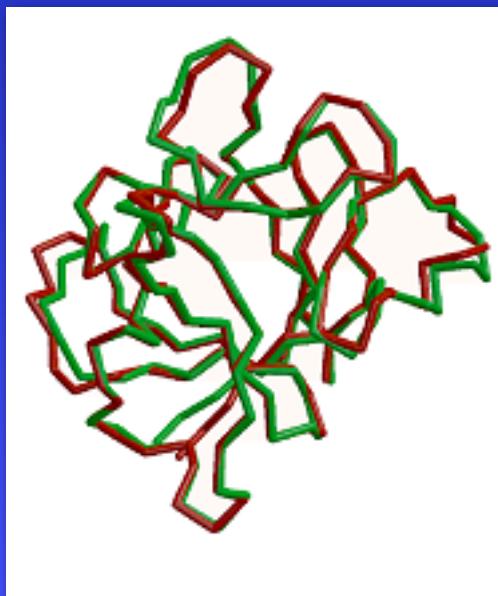
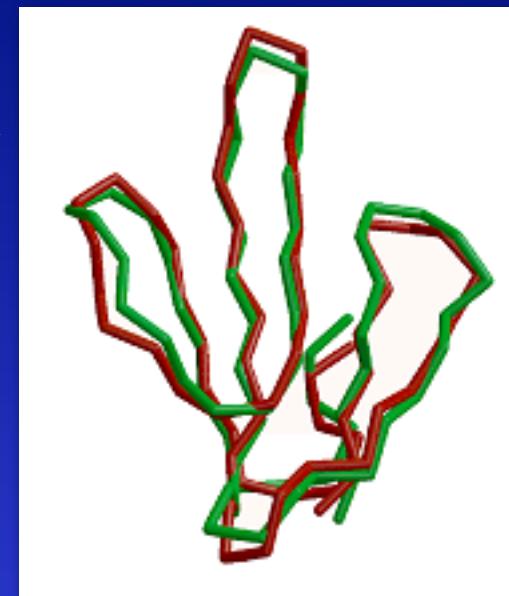
X-RAY
Interleukin 1 β 41bi (2.9 \AA)
Interleukin 1 β 2mib (2.8 \AA)

“Biological” significance of modeling errors



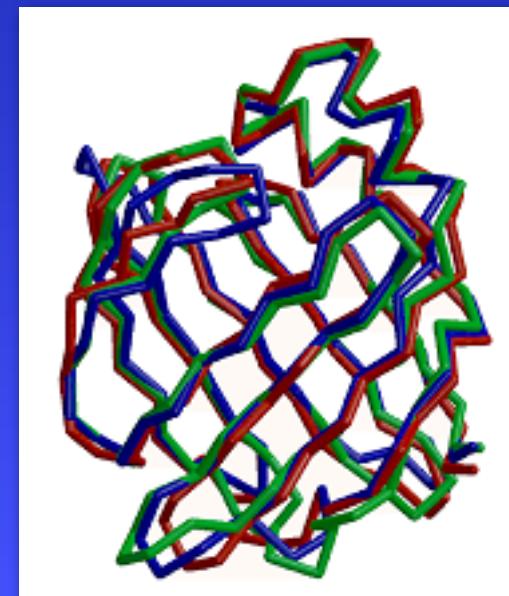
NMR
Ileal lipid-binding protein
1eal

NMR – X-RAY
Erabutoxin 3ebx
Erabutoxin 1era

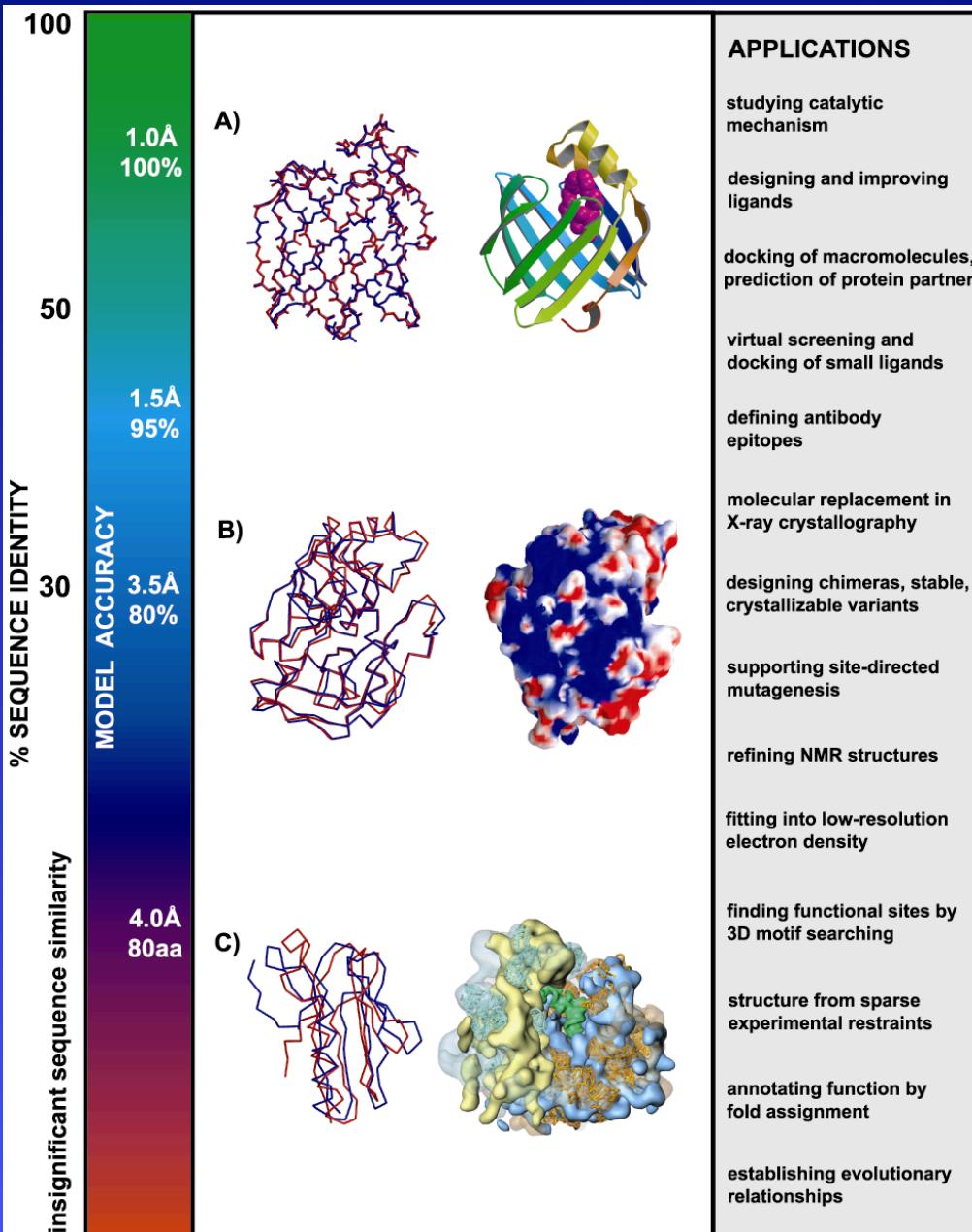


X-RAY
Interleukin 1 β 41bi (2.9 \AA)
Interleukin 1 β 2mib (2.8 \AA)

CRABPII 1opbB
FABP 1ftpA
ALBP 1lib
40% seq. id.



Applications of Comparative Models



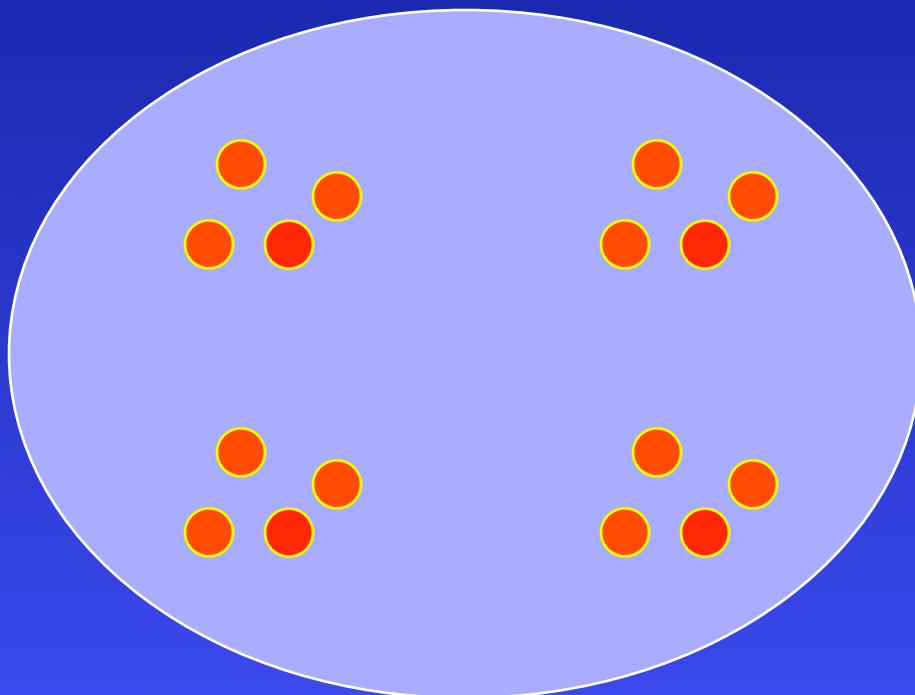
D. Baker & A. Sali.
Science **294**, 93, 2001.

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

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.

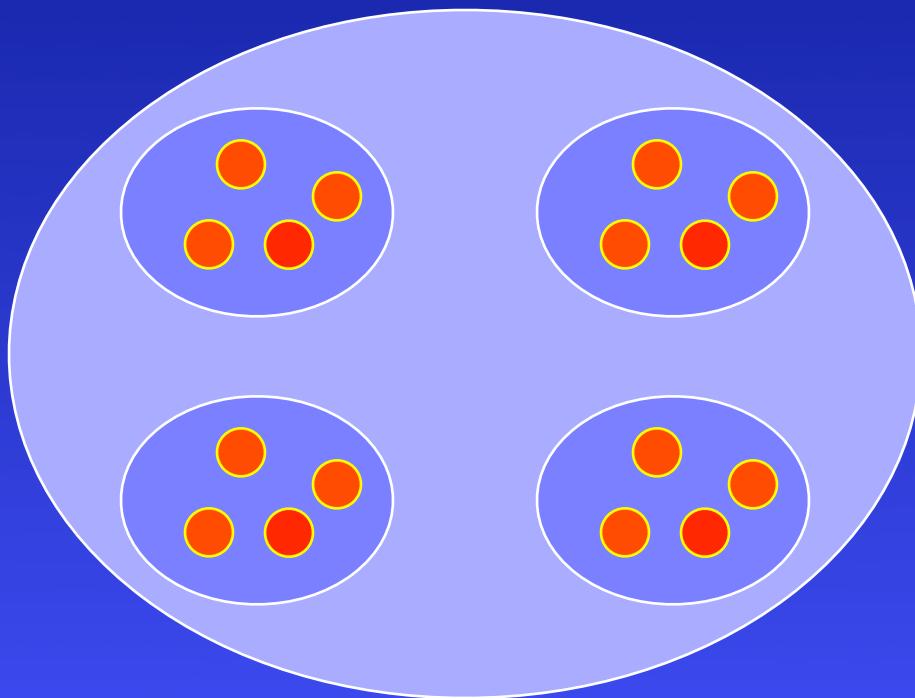
Characterize most protein sequences based on related known structures.



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.

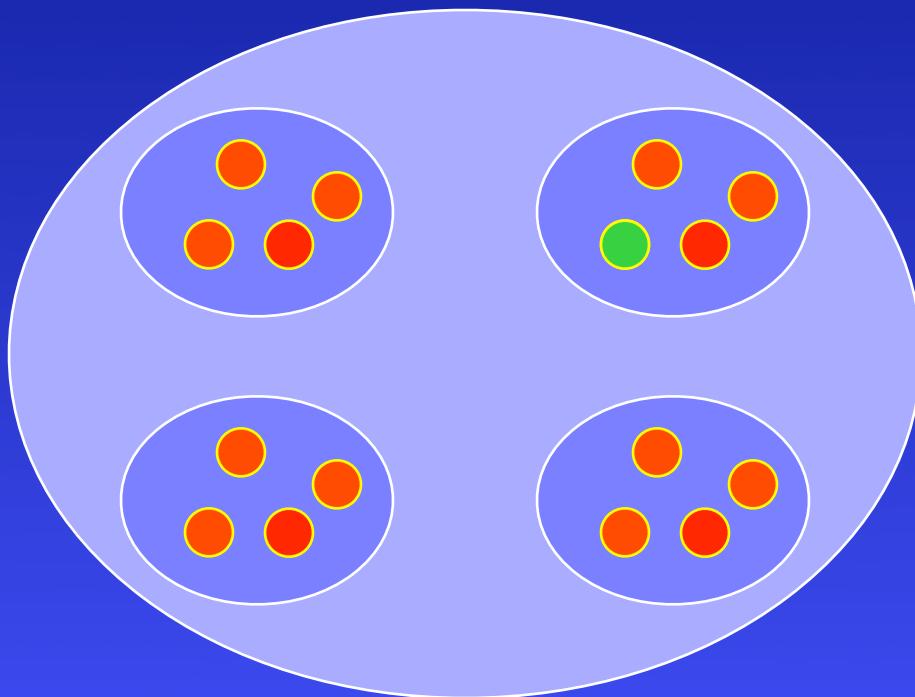
Characterize most protein sequences based on related known structures.



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.

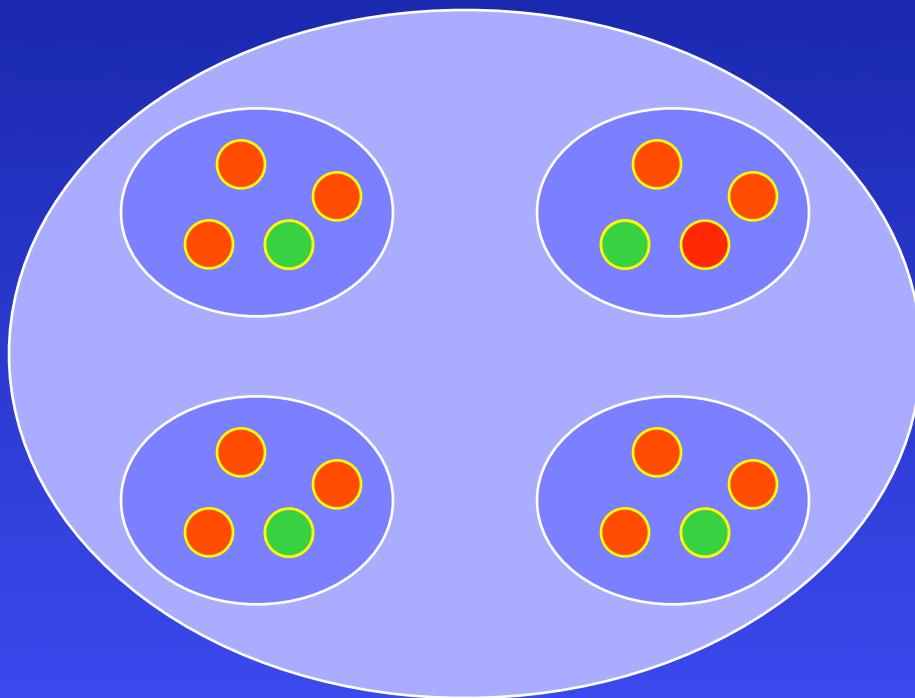
Characterize most protein **sequences** based on related known structures.



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.

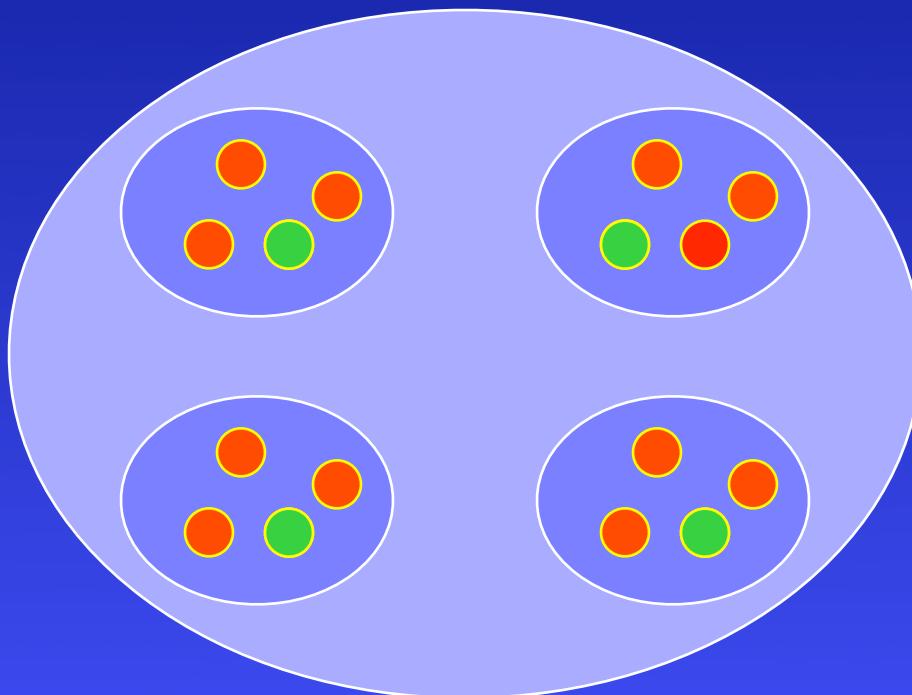
Characterize most protein **sequences** based on related known structures.



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.



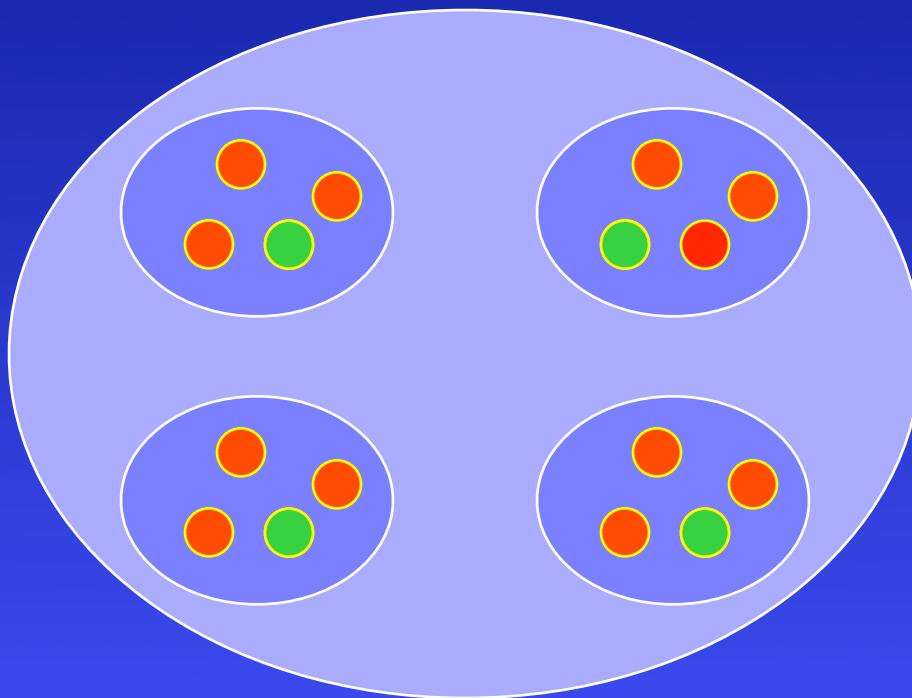
The number of “families” is much smaller than the number of proteins.

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.

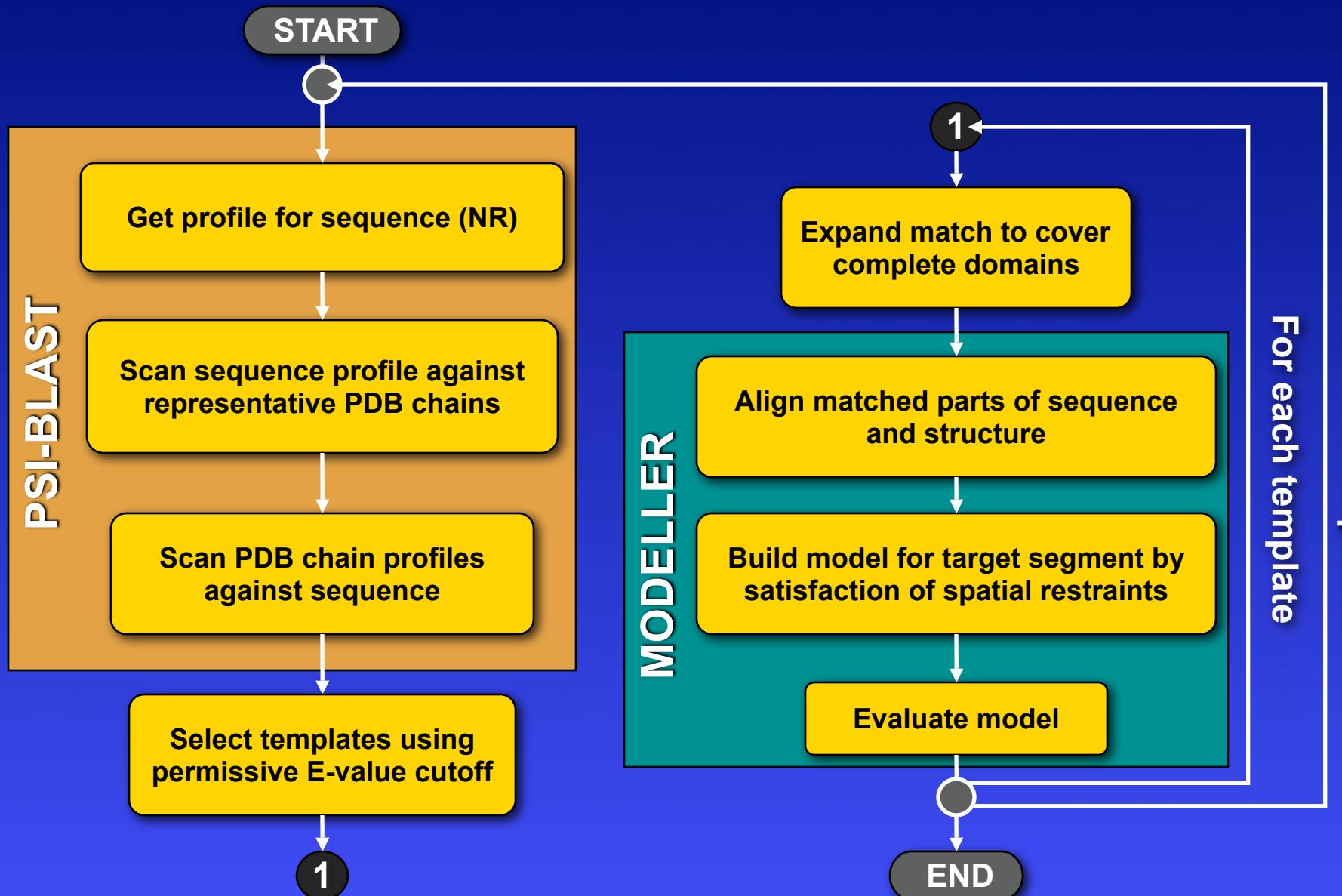


The number of “families” is much smaller than the number of proteins.

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)

MODPIPE: Large-Scale Comparative Protein Structure Modeling



Modeling with NY-SGRC structures

Mod
Web

Server for Comparative Modeling of Genes and Genomes



Please choose input type:

Sequence Structure

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

Netscape: ModBase: Login Form

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Location: <http://pipe.rockefeller.edu/modbase-cgi/index.cgi> What's Related

WebMail Calendar Radio People Yellow Pages Download Customize...

 **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 comparative protein structure models.
Ursus Pappa, Narayanan Eswar, Andrej Sali, Venkatesh A. Iyer, Andrei Sali.
Nucleic Acids Res. **30**, 255–259, 2002.

 MODBASE is maintained by Ursus Pappa in the group of Andrei Sali, Laboratories of Molecular Biophysics, Fels Family Center for Biochemistry and Molecular Biology, Temple University, 1230 York Ave, New York, NY 10032. Please address all inquiries to modbase@fagar.rockefeller.edu

[The Rockefeller University](#)

MODBASE Contents

837,698 Reliable Models or PSI-BLAST Fold Assignments for domains in 415,337 proteins. Last Update on 04/03/02. [MODBASE statistics](#)

Search for Models

Enter SwissProt/TrEMBL/GenBank/PDB identifier or description:

[Advanced Search](#)

Login [HELP](#)

[Academic login](#) [User login](#) [Logout](#)

Current logins: *modbase*

Some datasets are accessible freely without a login (e.g., the "Protein Data Bank" datasets). Other datasets are available to academic users only (e.g., our "SP3/TR" model set). And some datasets require a specific username and password. For commercial access to the models, please contact [Structural Genomics Inc.](#)

Notes

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

100% Bookmarks Location: http://pipe.rockefeller.edu/modbase-cgi/query_results.cgi?pub



<http://salilab.org/modbase>

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

8/9/02

Netscape: ModBase: Search Form

File Edit View Go Communicator

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Location: http://pipe.rochester.edu/modbase-cgi/search_form.cgi What's This?

WebMail Calendar Radio People Yellow Pages Download Customize...

ModBase

Database of Comparative Protein Structure Models

User: Academic User [Change User](#)

[RESET]

SEARCH for Models **SEARCH for Sequences**

[HELP]

DATASET SELECTION

Datasets:

- SPTR-2002
- SPTR-2001
- nysgrc_1M6
- nysgrc_1M6t
- nysgrc_1I9a
- nysgrc_1I4

[HELP]

SEARCH BY PROPERTIES

[HELP]

All

Organism or

Sort matching models by

[HELP]

MODEL SEARCH BY PROPERTY RANGES

[HELP]

()

<input type="text" value="["/> <input type="text" value="]"/>			
lower limit	upper limit	lower limit	upper limit
<input type="text" value="["/> <input type="text" value="]"/>			
lower limit	upper limit	lower limit	upper limit

SEARCH BY SEQUENCE SIMILARITY

[HELP]

Protein Sequence:

Search Summary

SUMMARY		Search Criteria			
Keywords	dfr	Category	-	Properties	(% Seq. Ident. and Model Size and Model Score)
Ranges (min-max)	-30	-	-	-	-
Values	<hr/>				
Minimum	8.00	82	0.01		
Average	25	181	0.79		
Maximum	30.00	492	1.00		

Many companies that market their products without test cannot not be considered.

29 matches were found using the specified search criteria. Click on the links in the table header to re-sort your output.

TARGET						MODEL DATA				TEMPLATE				
Model/Fold Reliability	Sequence based View	Select Sequence Database Links	Database Description	Organism	Protein Size	Modeled Segment	Size	Seq Id (%)	E-value	Model Score	PDB code	Template based View	Segment	Description
Green	Sequence based View	■ SB 037452	DIVIHYDROFOLATE REDUCTASE TYPE VIII (EC 1.5.1.5) (DHFR TYPE IIC) Dataset: PDB+ PRD000M	Escherichia coli Shigella sonnei	169	1-165	165	30.00	4e-32	1.00	1dfr_A		1-158	DIVIHYDROFOLATE REDUCTASE
Green	Sequence based View	■ SB 027392	BIFUNCTIONAL CHYDROXYLATE REDUCTASE-THYMIDYLATE SYNTHASE (EC 1.5.1.15-1.4.1.11) (BCHRS) (DHFR TYPE EC 1.5.1.5) Dataset: SPTR-2001 PRAM PRODOM	Lemnaceae major	520	24-231	208	30.00	2e-49	1.00	1dfr_B		1-166	DIVIHYDROFOLATE REDUCTASE (EC 1.5.1.5)-CHYDROXYMETHYL FOLATE Reductase
Green	Sequence based View	■ SB 020712	BIFUNCTIONAL CHYDROXYLATE REDUCTASE-THYMIDYLATE SYNTHASE (DHFR-TSDE INCLUDES: DHYDROFOLATE REDUCTASE-THYMIDYLATE SYNTHASE (EC 1.5.1.5-1.4.1.11)) Dataset: SPTR-2001 PRAM PRODOM	Plasmodium chabaudi	583	21-241	221	30.00	3e-39	1.00	1dfr_C		2-193	DIVIHYDROFOLATE REDUCTASE
Green	Sequence based View	■ SB 028924	BIFUNCTIONAL CHYDROXYLATE REDUCTASE-THYMIDYLATE SYNTHASE (DHFR-TSDE INCLUDES: DHYDROFOLATE REDUCTASE-THYMIDYLATE SYNTHASE (EC 1.5.1.5-1.4.1.11)) Dataset: SPTR-2001 PRAM PRODOM	Plasmodium vivax	623	35-237	203	30.00	1e-37	1.00	1dfr_D		14-203	DIVIHYDROFOLATE REDUCTASE
Sequence based View	Select Sequence Database Links	Database Description				Organism			Protein Size Modeled Segments - Schema					
Green	Sequence based View	■ SB 037452	DIVIHYDROFOLATE REDUCTASE TYPE VIII (EC 1.5.1.5) (DHFR TYPE IIC) Dataset: PDB+ PRD000M	Escherichia coli Shigella sonnei	169	Protein Size Modeled Segments - Schema				Protein Size Modeled Segments - Schema				
Green	Sequence based View	■ TR 059100	DHFRP PROTEIN (FRAGMENT) Dataset: PDB+ PRD000M	Homo sapiens	121	Protein Size Modeled Segments - Schema				Protein Size Modeled Segments - Schema				

Database Synonyms for this Sequence (100% Sequence Identity)			
TrEMBL	Q93V12	Salmonella typhimurium	Dihydrofolate reductase
GI	1333727	Plasmid pLM0229	dfr1 product (AA 1 - 157)
GI	897050	Salmonella typhimurium	dihydrofolate reductase
GI	95755	Escherichia coli plasmid pLM0229	S1176 dihydrofolate reductase (EC 1.5.1.3) locus - Escherichia coli plasmid pLM0229

Model Data

PSI-BLAST Fold Assignments (left half) and Reliable Models (right half) are indicated in green.

* Indicates an E-value from an unfiltered PSI-BLAST search when a filtered search does not result in a significant match.

This Table displays all Models/Folds of this sequence

using the current search parameters

Mod View Download Now

MODEL DATA					LINKS				
Fold Model Reliability	Size	Seq Id (%)	E-value	Model Score	RCS	3D Coor	PDB	PAP	Nxt View
0.62	154	29.00	7e-49	1.00					
0.62	155	26.00	2e-37	1.00					

Comparative modeling of the TrEMBL database

Unique sequences processed: 733,239

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

Comparative modeling of the TrEMBL database

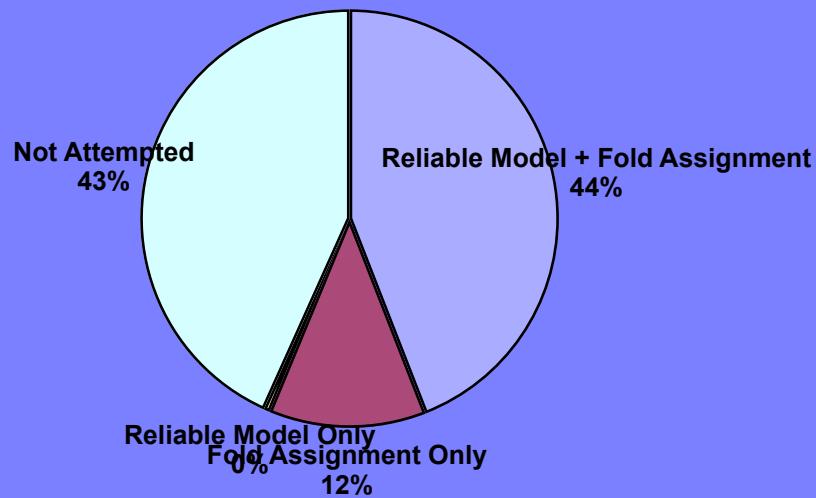
Unique sequences processed: 733,239

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

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

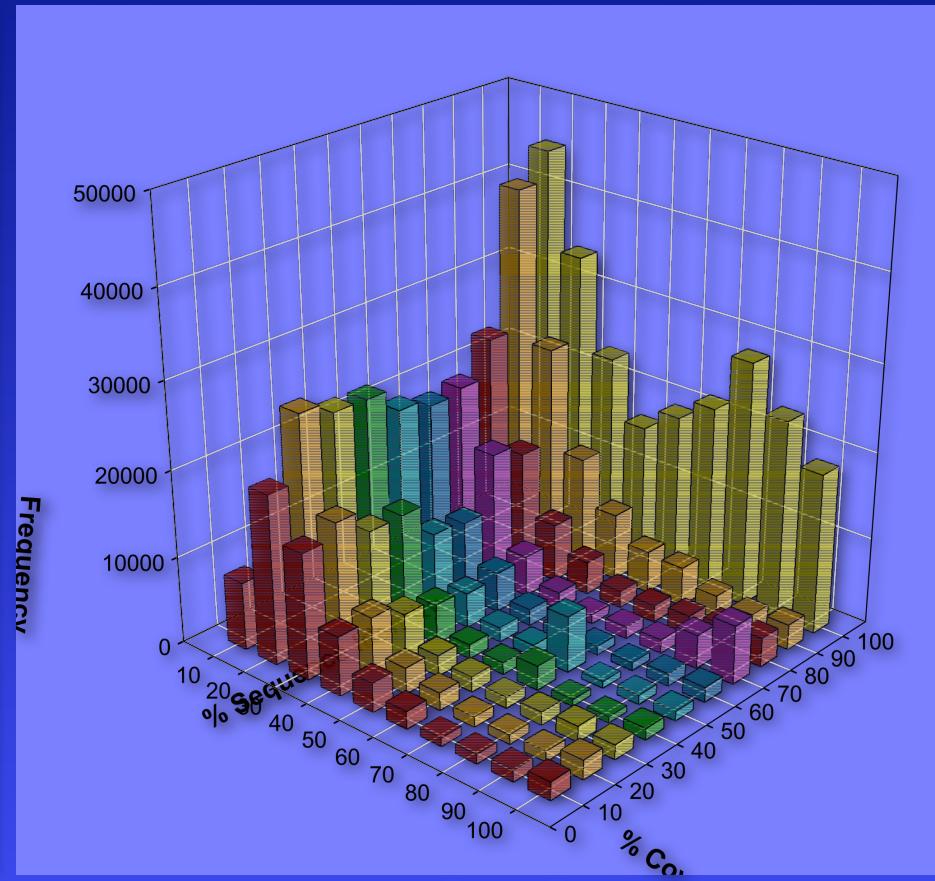
On average, only a domain per protein is modeled
(an “average” protein has 2.5 domains of 175 aa).

Modeling Coverage of the Sequence Space



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

Reliable Model: **Model Score ≥ 0.7**

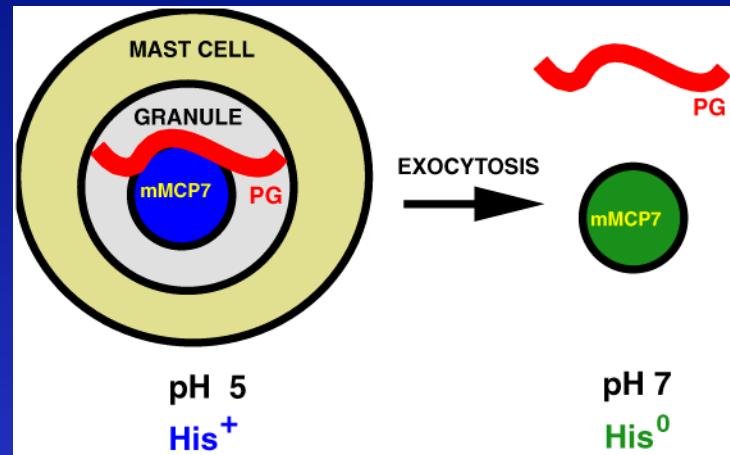
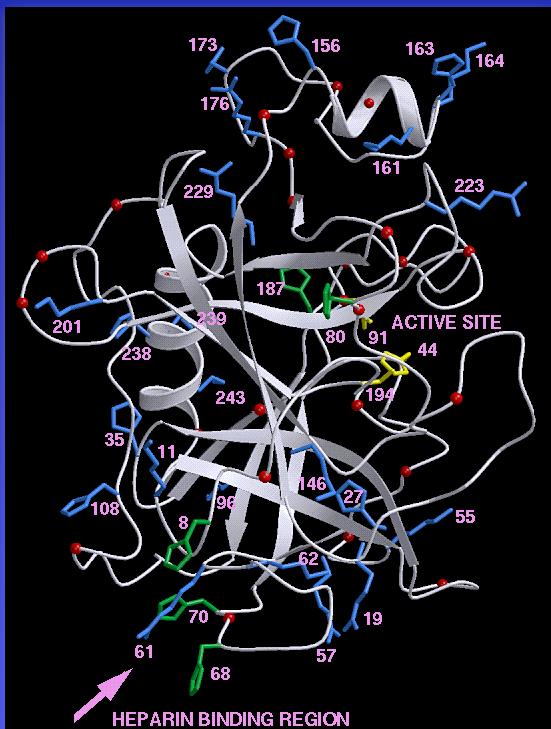


Examples...

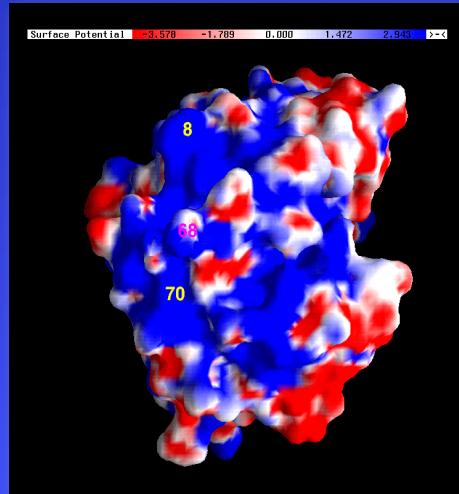
Do mast cell proteases bind proteoglycans? Where? When?

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

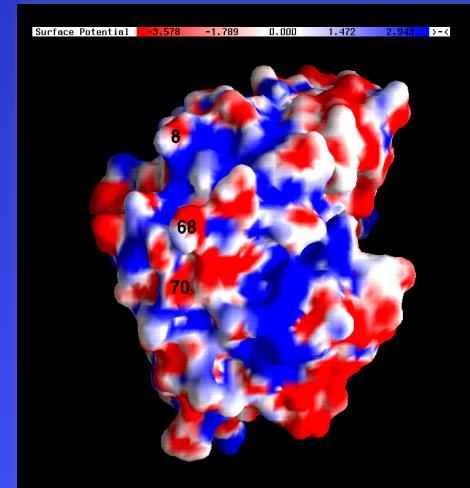
1. mMCPs bind negatively charged proteoglycans through electrostatic interactions?
2. Comparative models used to find clusters of positively charged surface residues.
3. Tested by site-directed mutagenesis.



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.



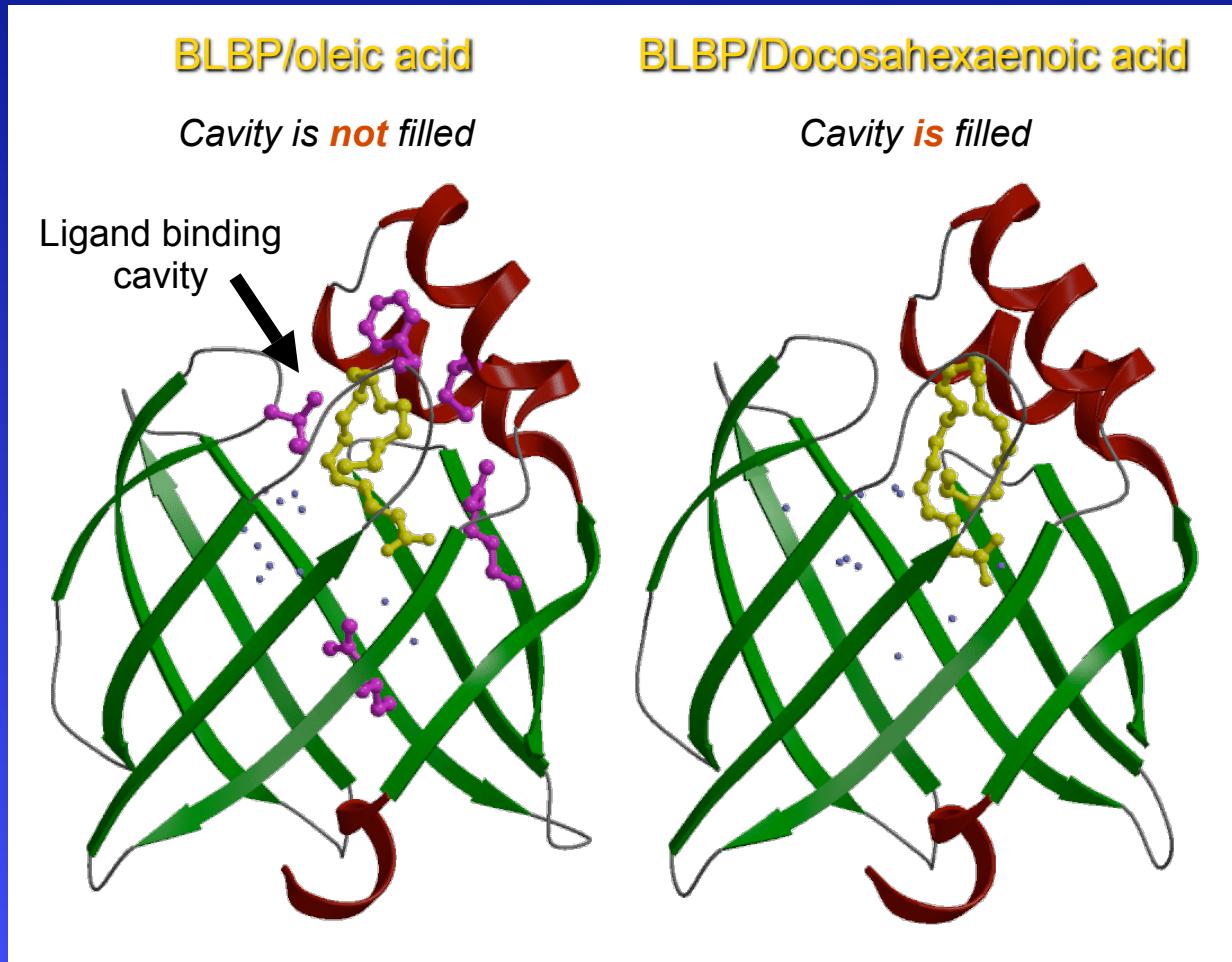
Native mMCP-7 at pH=5 (His⁺)



Native mMCP-7 at pH=7 (His⁰)

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

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



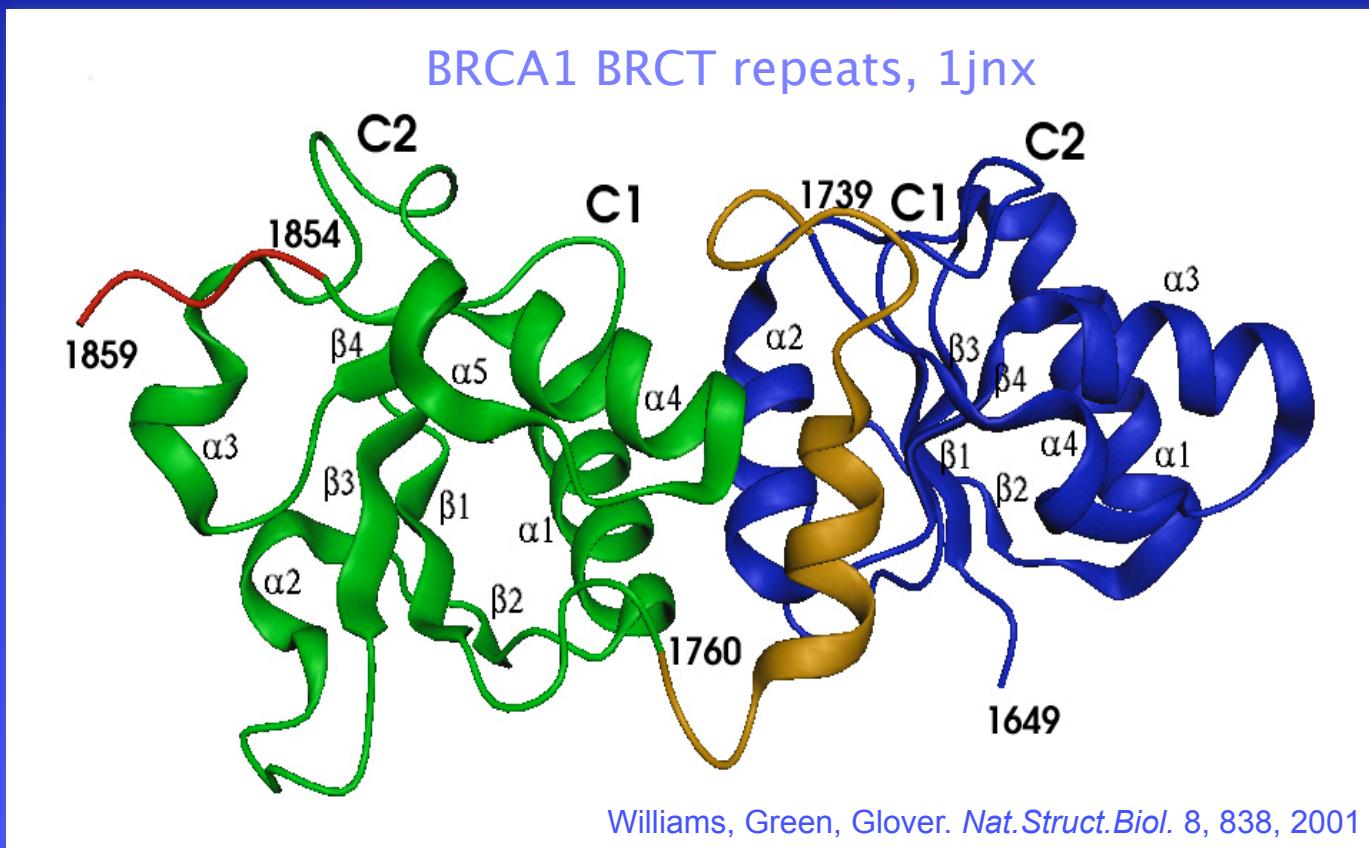
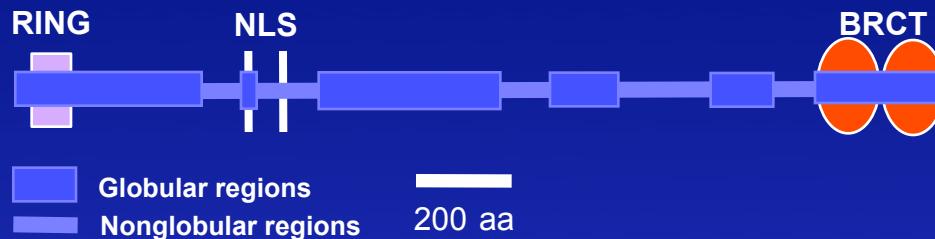
1. BLBP binds fatty acids.
2. Build a 3D model.
3. Find the fatty acid that fits most snuggly into the ligand binding cavity.

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

Nieco Singer, MS 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 #: 56594
Physician: Fred Gilbert, MD		

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:

Brian E. Ward, Ph.D.
Laboratory Director

Thomas S. Frank, M.D.
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.

CONFIDENTIAL



BRACAnalysis™
Comprehensive BRCA1-BRCA2 Gene Sequence Analysis Result

Nicole Singer, MS 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 #: 56594
Physician: Fred Gilbert, MD		

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:

Brian E. Ward, Ph.D.
Laboratory Director

Thomas S. Frank, M.D.
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

no transcription activation

transcription activation

?

cancer associated
not cancer associated

?

C1697R
R1699W
A1708E
S1715R
P1749R
M1775R

cancer associated
not cancer associated

M1652K
L1657P
E1660G
H1686Q
R1699Q
K1702E
Y1703H
F1704S

L1705PS
1715NS
1722FF
1734LG
1738EG
1743RA
1752PF
1761I

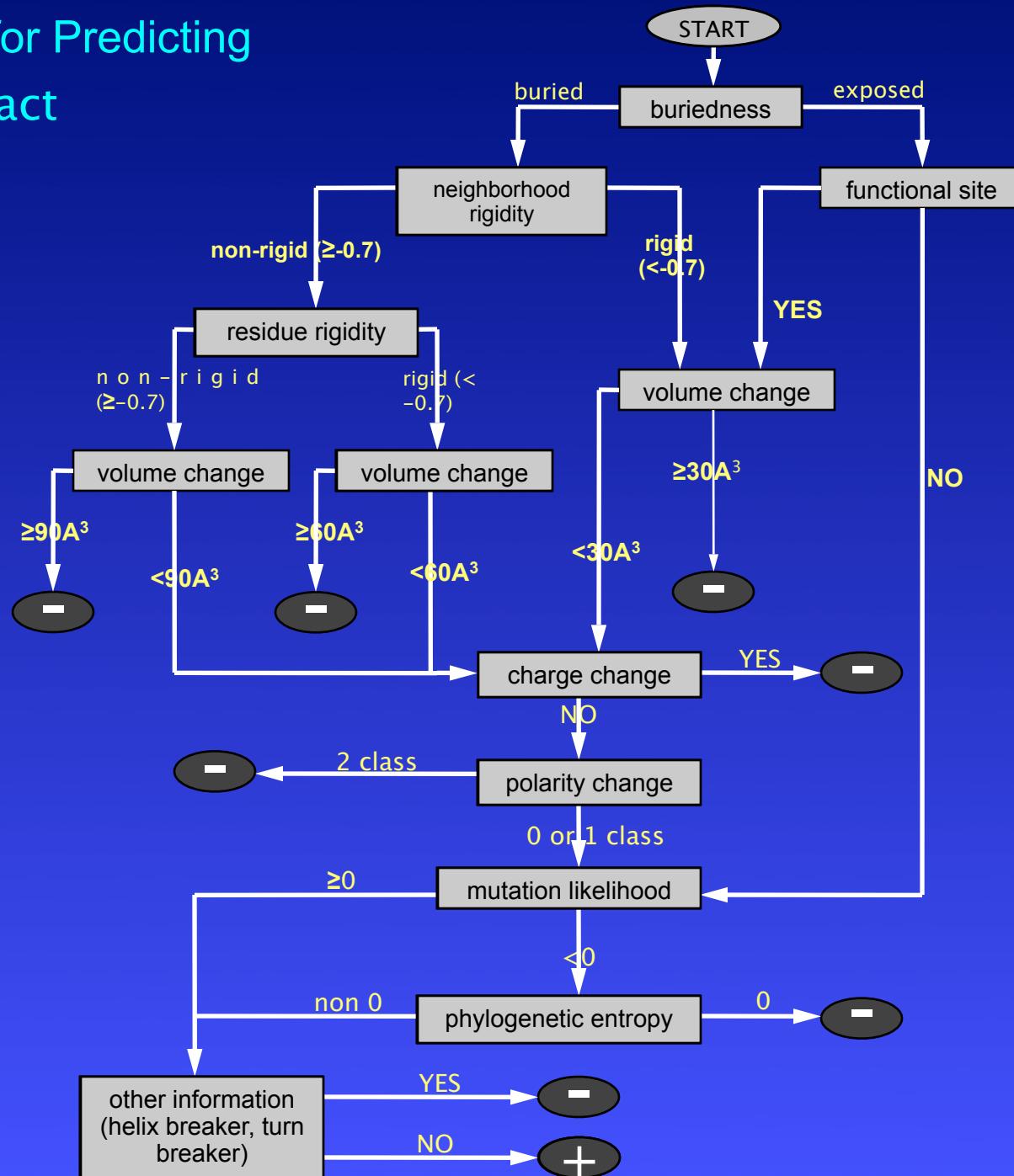
F1761S
M1775E
M1775K
L1780P
I1807S
V1833E
A1843T

M1652I
A1669S

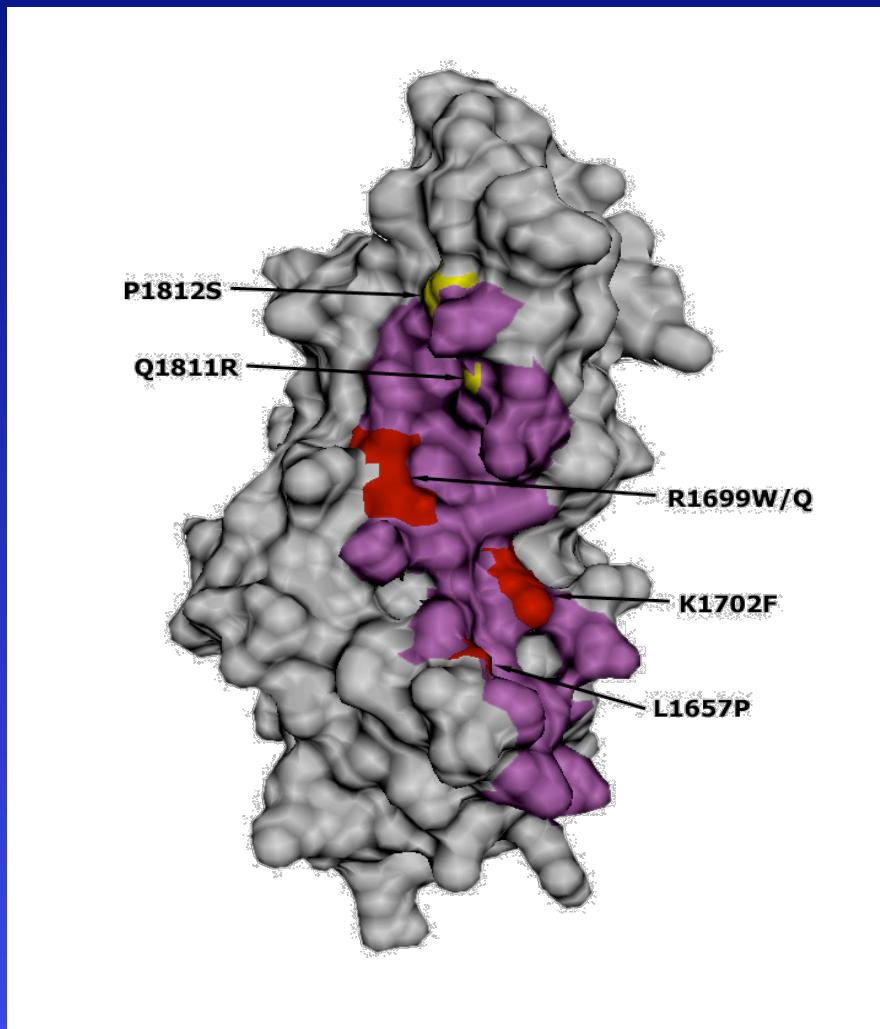
V1665M
D1692N
G1706A
D1733G
M1775V
P1806A

M1652T V1653M L1664P T1685A T1685I M1689R D1692Y F1695L V1696L R1699L G1706E W1718C
W1718S T1720A W1730S F1734S E1735K V1736A G1738R D1739E D1739G D1739Y V1741G H1746N
R1751P R1751Q R1758G L1764P I1766S P1771L T1773S P1776S D1778N D1778G D1778H M1783T
C1787S G1788 D G1788V G1803A V1804D V1808A V1809A V1809F V1810G Q1811R P1812S
A1823T V1833M W1837R W1837G S1841N A1843P T1852S P1856T P1859R N1819S

“Decision” Tree for Predicting Functional Impact of Genetic Variants



Putative Binding Site on BRCA1



RMSMV**VSGLTPEEFMLVYK**FARKHHITLTNLITEETTHVVMKTDAEV**CERTLKYFLGIAGGKwVVSYFWVTQSIKERK**
MLNEHDFEVRGDVVNGRNHQGPKRARESQDRKIFRGLEICCYGPFT**TNMP**TDQLEWMVQLCGASVV**KELSSFTLGTGVHP**
IVVVQPDAWTEDNGFHAI**GQMCEAPVVTREWVLDSVALYQCQELDTYLIPQIP**

Conclusions

Conclusions

- ✓ At present, useful 3D models can be obtained for domains in ~ 55% of the proteins (25% of domains).

Conclusions

- ✓ At present, useful 3D models can be obtained for domains in ~ 55% of the proteins (25% of domains).

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.

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.

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.

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.

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
Bino John
Dmitry Korkin
M. S. Madhusudhan
Nebosja Mirkovic
Ursula Pieper
Andrea Rossi
Min-yi Shen
Maya Topf

<http://www.salilab.org>