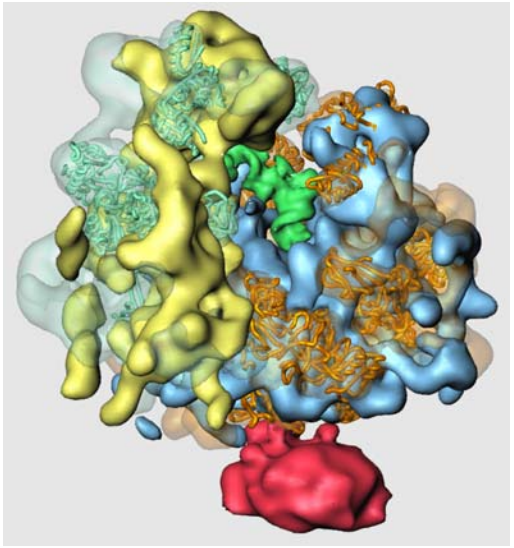
Protein structure modeling and the Protein Structure Initiative

András Fiser



Department of Biochemistry and
Seaver Foundation Center for Bioinformatics
Albert Einstein College of Medicine
New York, USA

Sequence *versus* Structure

GDCAGDFKIWYFGRTLLVAGAKDEFGAIDA

RTLAWYAGHLVAGAKDEFGGDFKIWYFGAI

DFLLVAGAKDEFGKIWYFGGIDAWRTAGDC

HLVAGARTLAFGAIDWYAKDEFGGGDFKIW

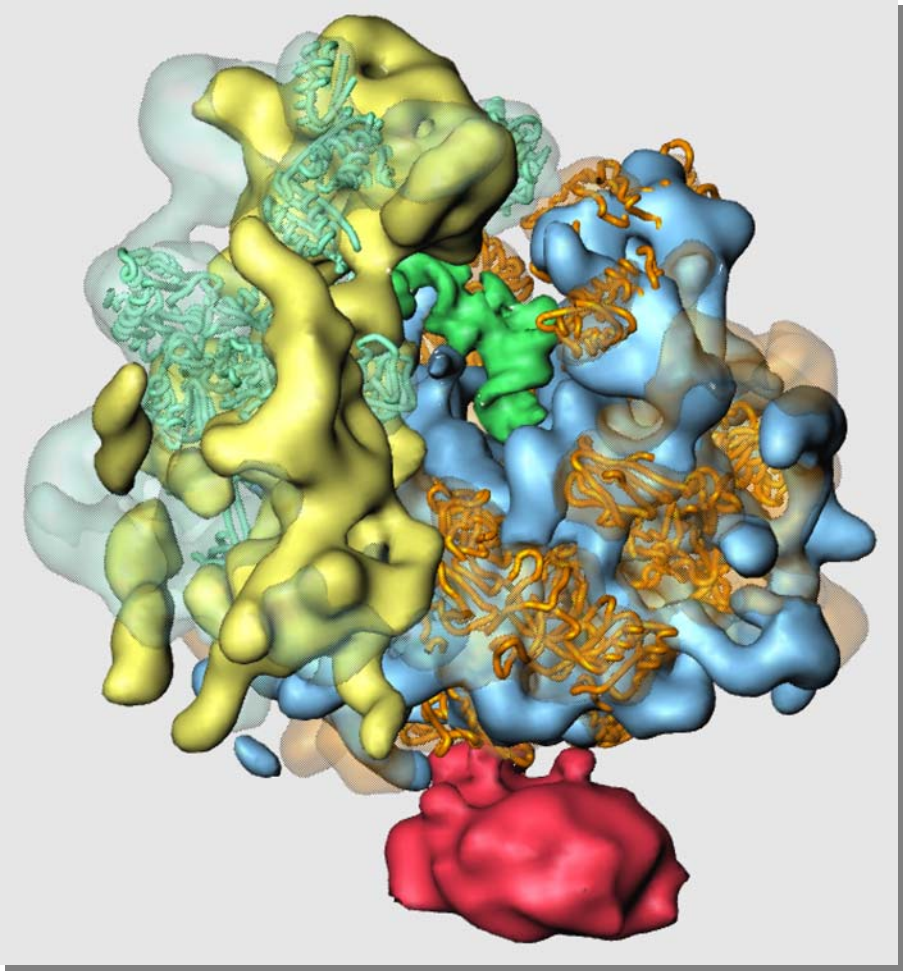
ARTHLVAGFGGGAIDWYFKIWYAKLAFGDE

GCTAGCTTAAGGCCTTCATGATCTTCTGAG

AGGGCTCCTTCATGATAGCTTAAGGCTTAA

AGGCCTTCATGGGGTTAACATATCTTCTGA

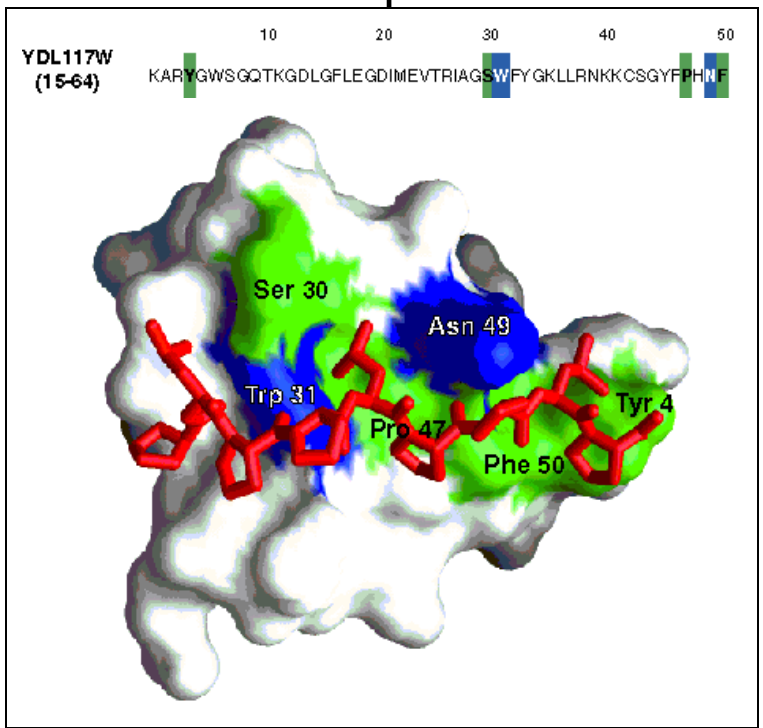
CCTTCATGCTAGCTTAAGGGATCTTAACCG



Why is it useful to know the structure of a protein not only its sequence?

- The biochemical function of a protein is defined by its interactions with other molecules.

- The 3D structure is more informative than sequence because interactions are determined by residues that are close in space but are frequently distant in sequence.



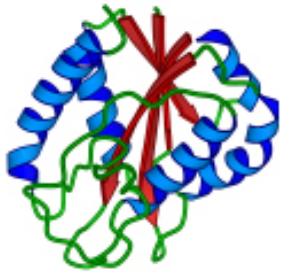
Evolution tends to conserve function and function depends more directly on structure than on sequence, structure is more conserved in evolution than sequence.

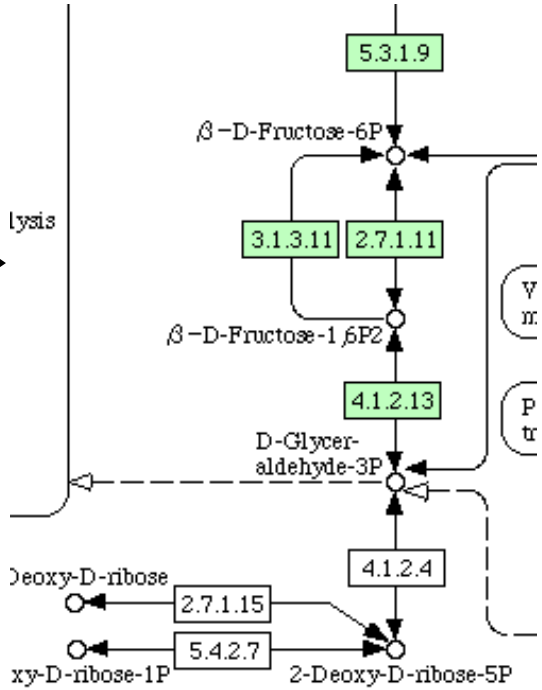
Patterns in space are frequently more recognizable than patterns in sequence.

Function via Structure

Sequence → Structure → Function

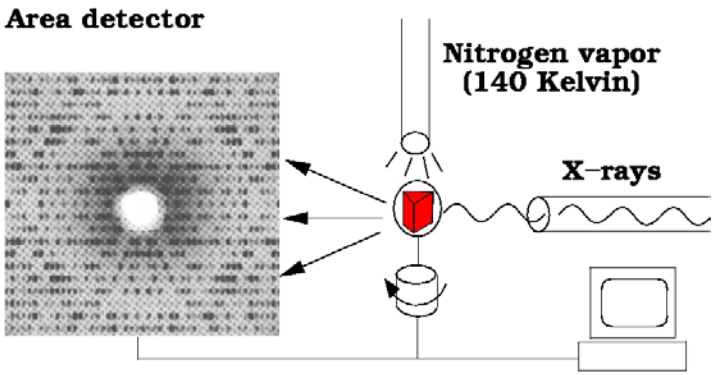
GFCHIKAYTRLIM...

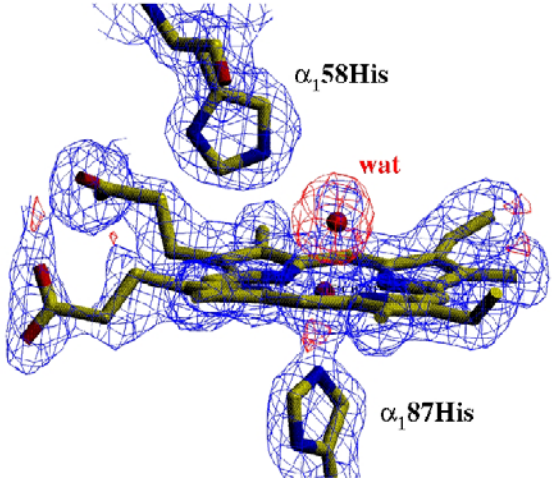




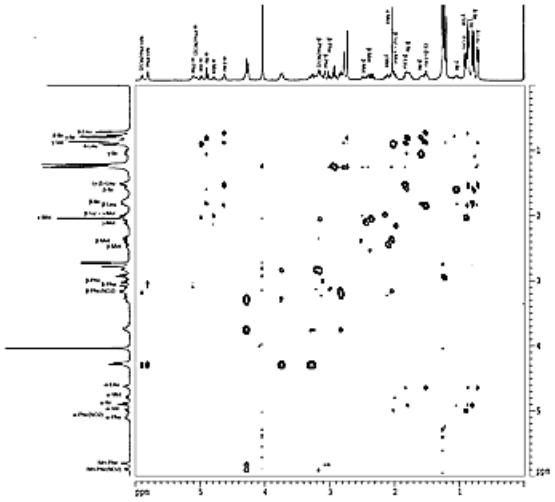
Major experimental tools

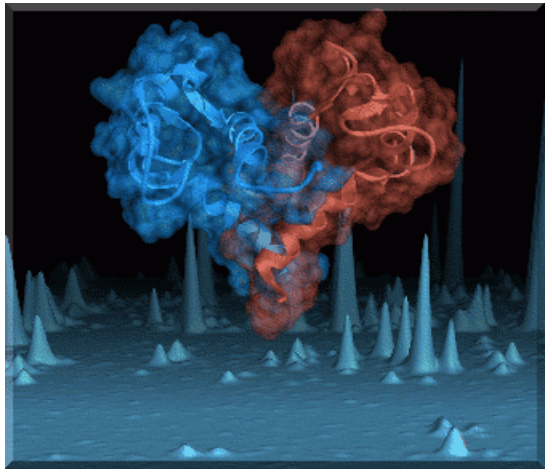
X-ray crystallography





Nuclear Magnetic Resonance





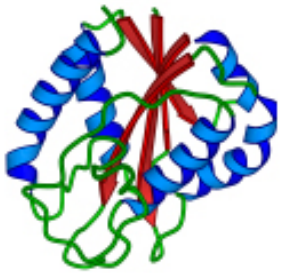
Why Protein Structure Prediction?

	Y 2004	Y 2006
Sequences	1,900,000	millions
Structures	26,000	40,000

We know the experimental 3D structure for about 1% of the protein sequences

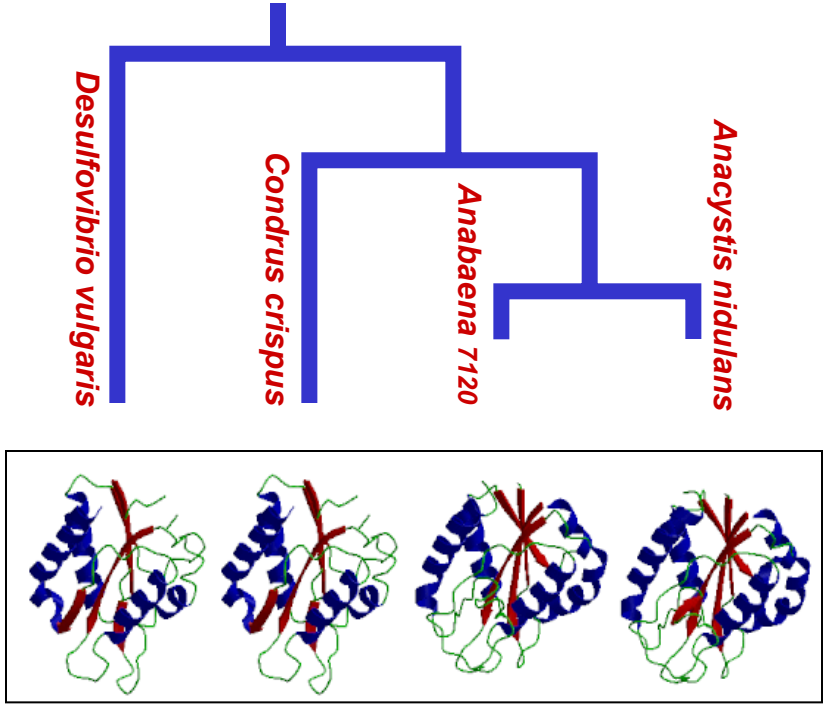
Principles of Protein Structure

GFCHIKAYTRLIMVG...



folding

Ab initio prediction



evolution

Fold Recognition
Comparative Modeling

Protein structure modeling

Ab initio prediction

Applicable to **any** sequence

Not very accurate, and attempted for proteins of <100 residues

Accuracy and applicability are limited by our understanding of the protein folding problem

Comparative Modeling

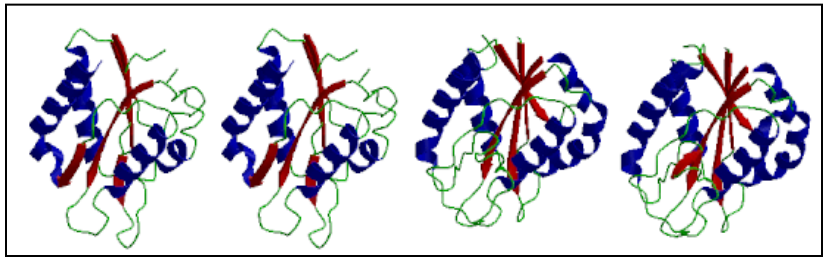
Applicable to those sequences **only** that share recognizable similarity to a template structure

Fairly accurate, typically comparable to a low resolution X-ray experiment.
Not limited by size

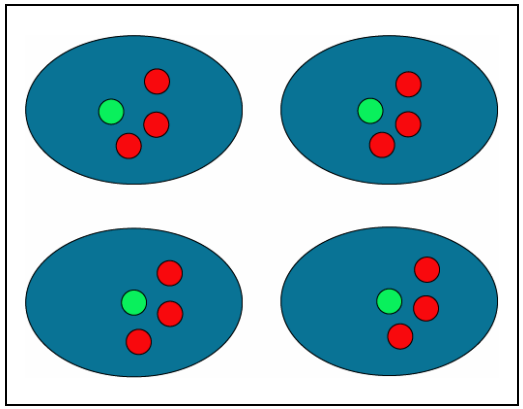
Accuracy and applicability are rather limited by the number of known folds

What makes comparative modeling possible

I A small difference in the sequence makes a small difference in the structure

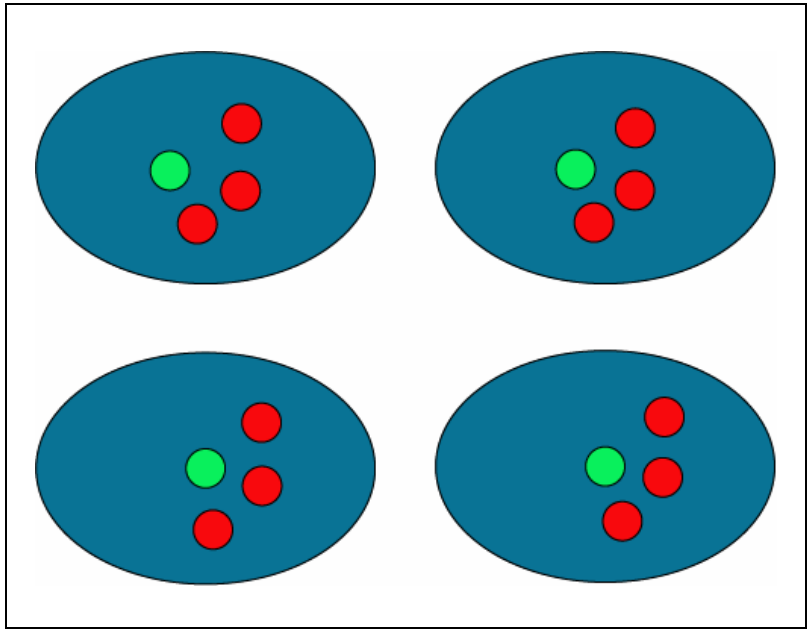


II Protein structures are clustered into fold families



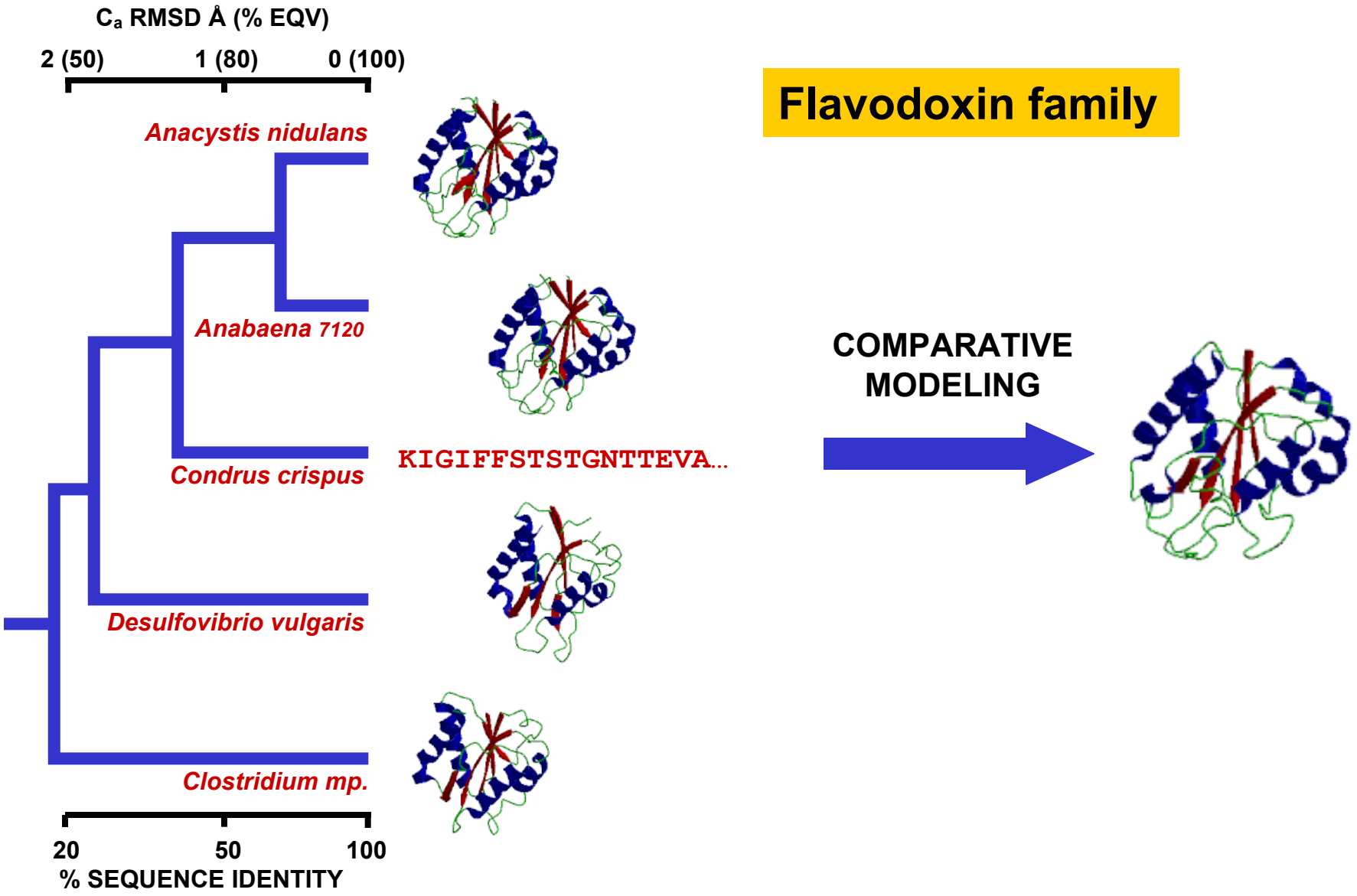
Structural Genomics

Characterize most protein sequences (red) based on related known structures (green).

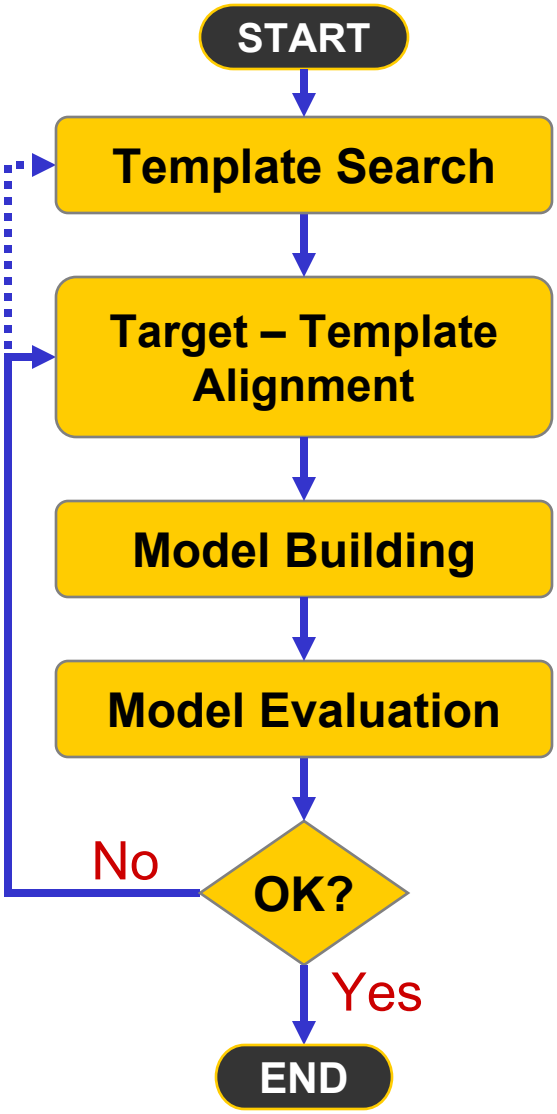


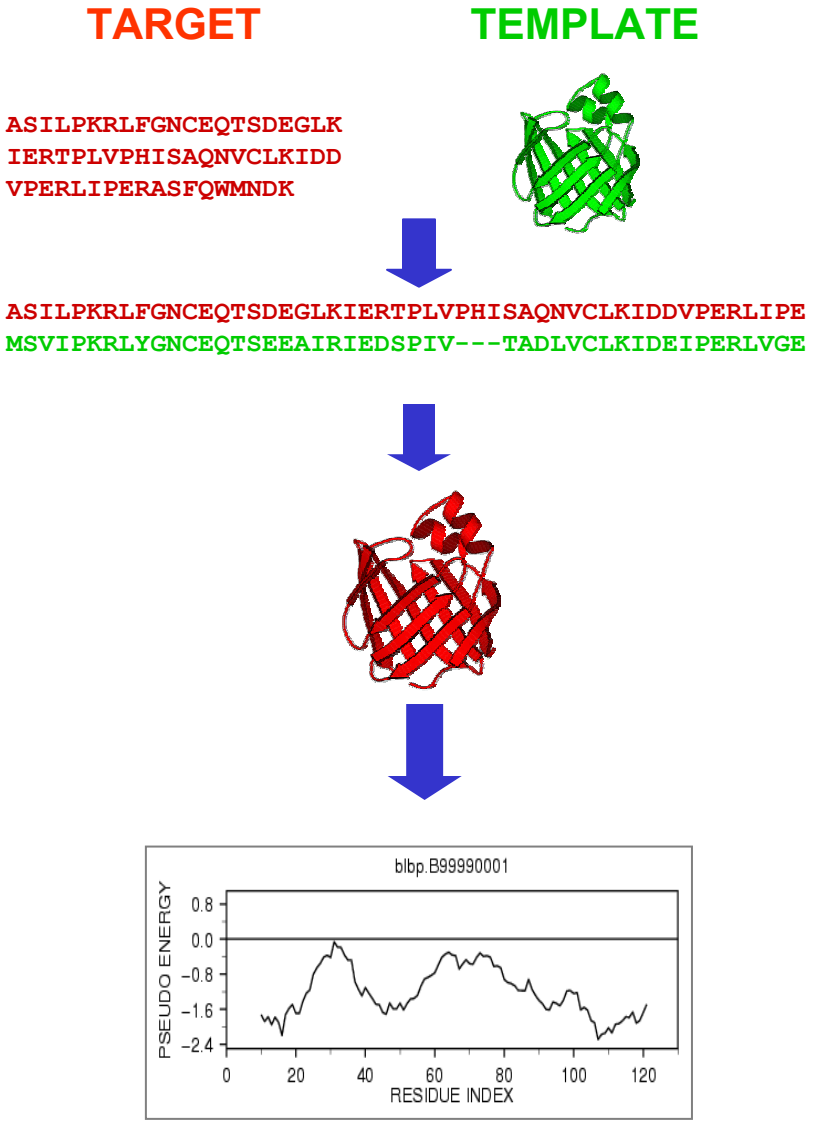
The number of “families” is much smaller than the number of proteins

Comparative Protein Structure Modeling



Steps in Comparative Protein Structure Modeling

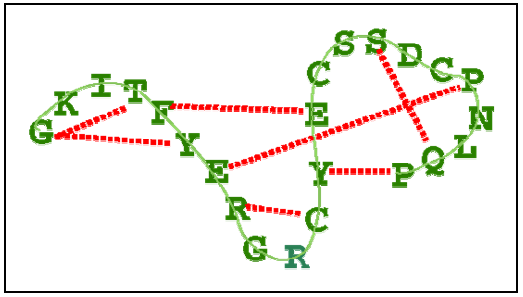


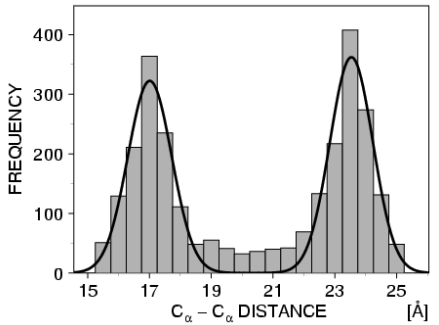


Comparative Modeling by Satisfaction of Spatial Restraints (MODELLER)

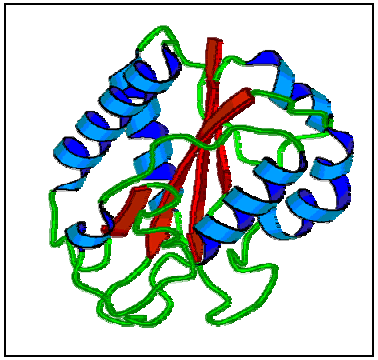
```
3D   GKITFYERGFQGHCYESDC-NLQP...
SEQ  GKITFYERG---RCYESDCPNLQP...
```

1. Extract spatial restraints





2. Satisfy spatial restraints

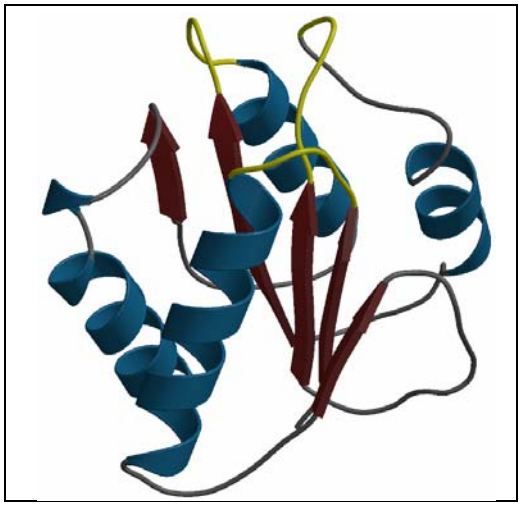


$$F(R) = P \underset{i}{} p_i(f_i/I)$$

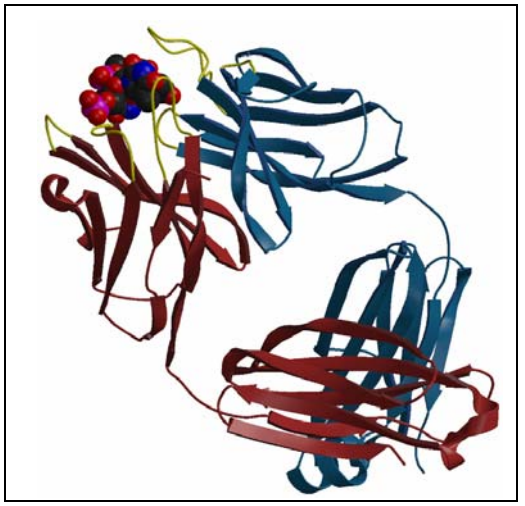
A. Šali & T. Blundell. J. Mol. Biol. 234, 779, 1993.
J.P. Overington & A. Šali. Prot. Sci. 3, 1582, 1994.
A. Fiser, R. Do & A. Šali. Prot Sci. 9, 1753, 2000.

Loop Modeling in Protein Structures

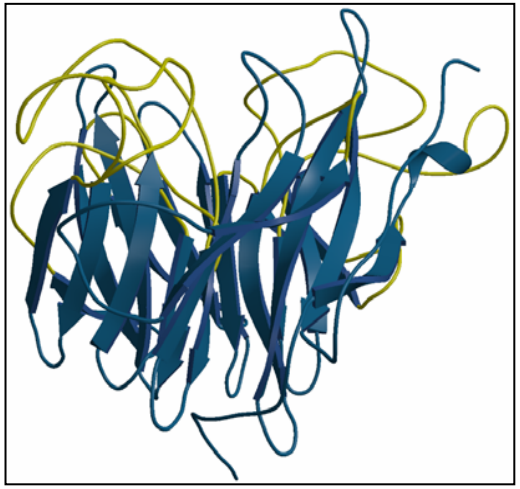
α+β barrel: flavodoxin

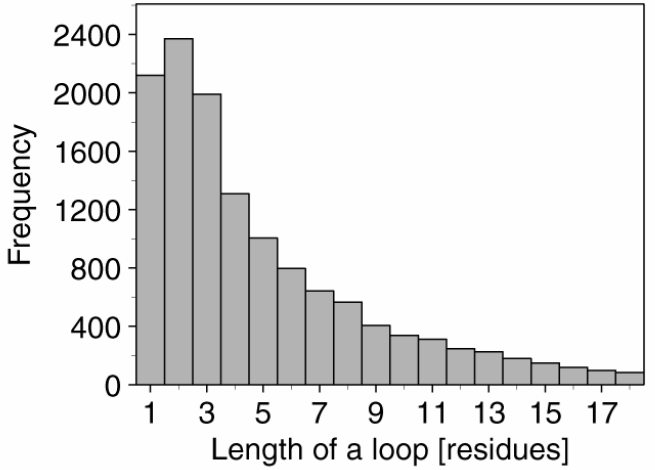


IG fold: immunoglobulin



antiparallel β-barrel



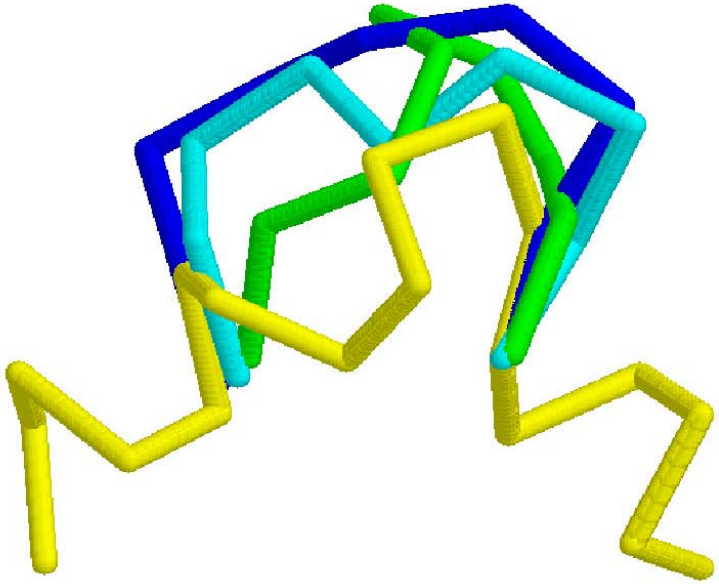


A. Fiser, R. Do & A. Šali, *Prot. Sci.,* **9**, 1753, 2000

Loop modeling strategies

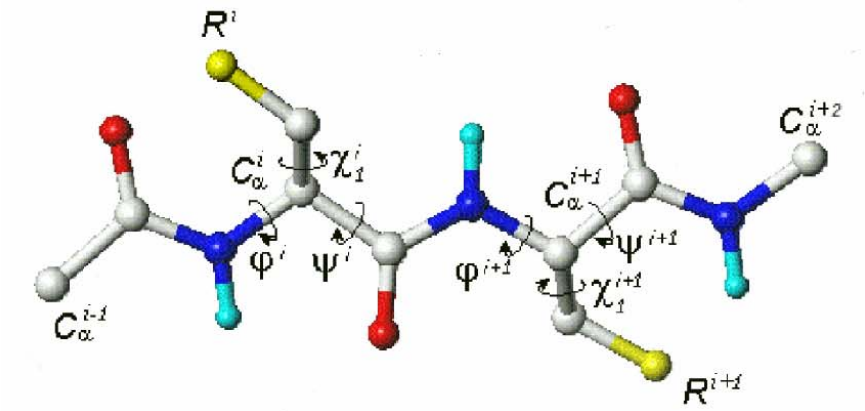
Database search

"Comparative"



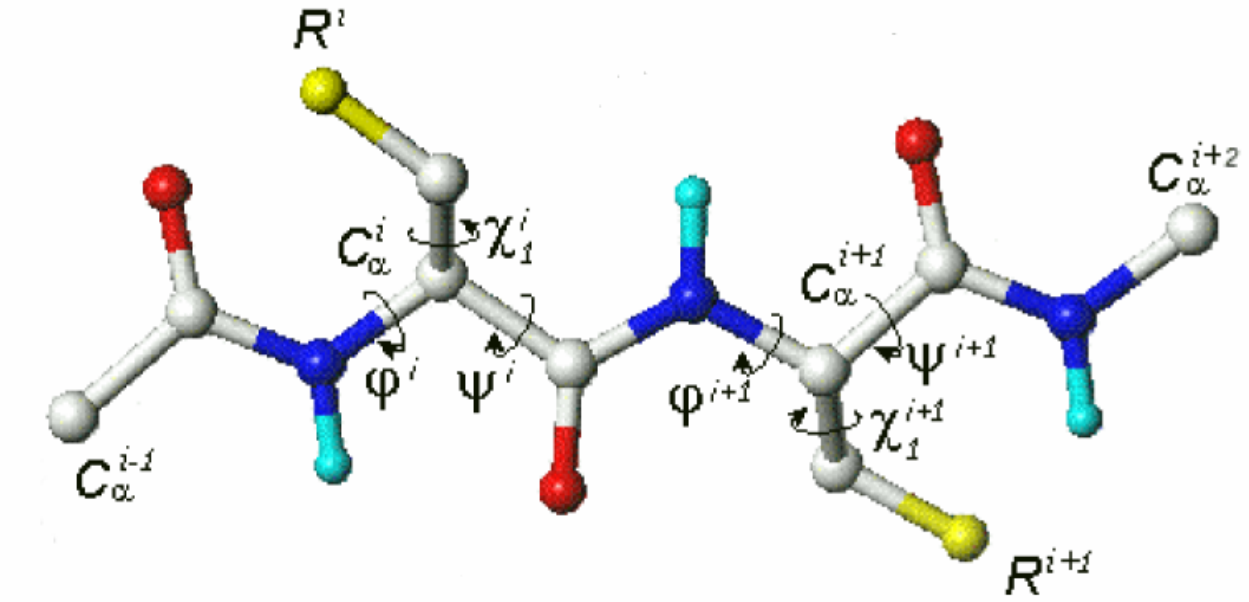
Conformational search

"*ab initio*"



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

Loop Modeling by Conformational Search



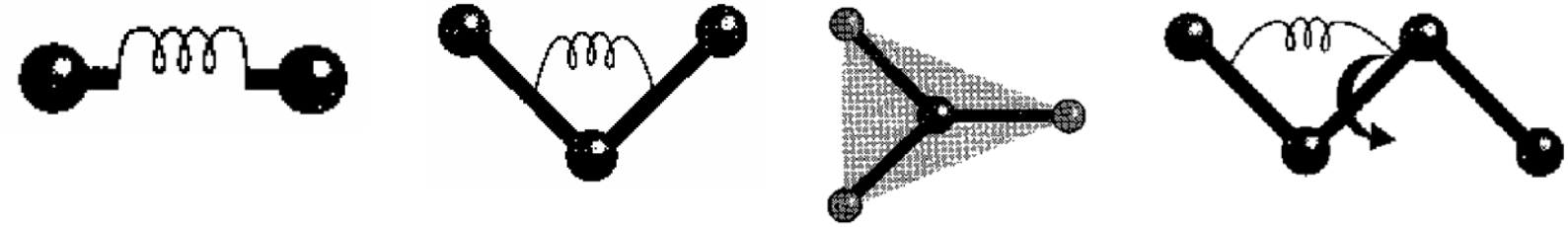
1. Protein representation.
2. Energy (scoring) function.
3. Optimization algorithm.

Energy Function for Loop Modeling

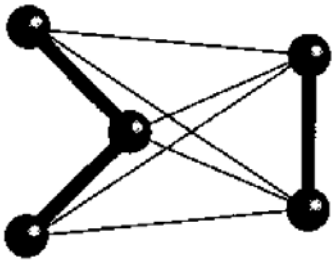
The energy function is a sum of many terms:

1) Statistical preferences for dihedral angles:

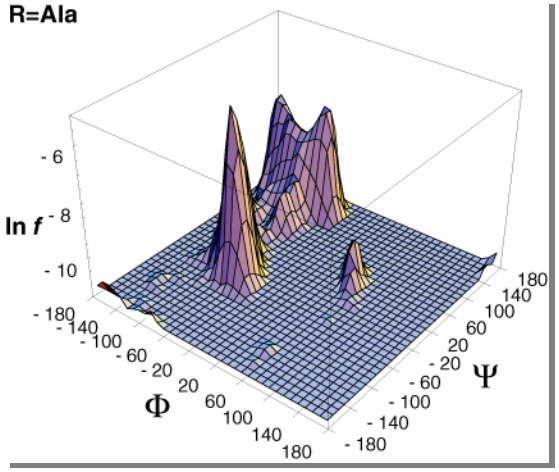
2) Restraints from the CHARMM-22 force field:

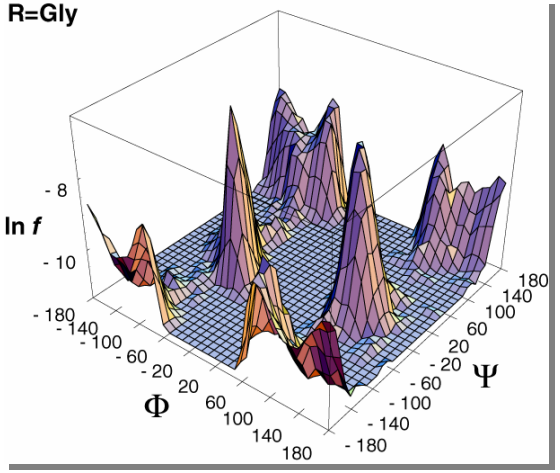


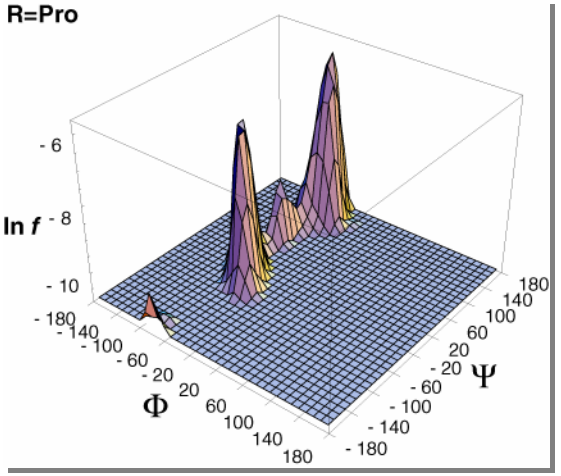
3) Statistical potential for non-bonded contacts:

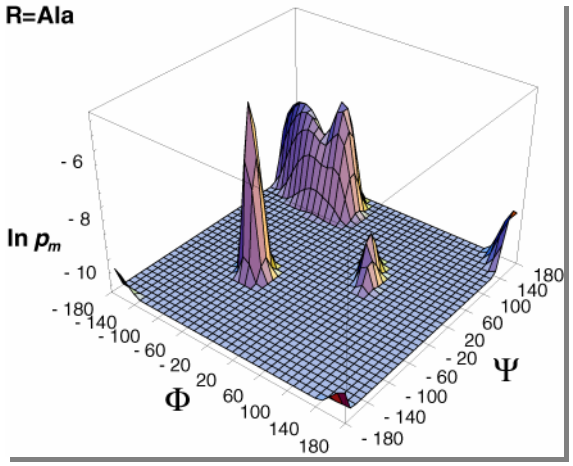


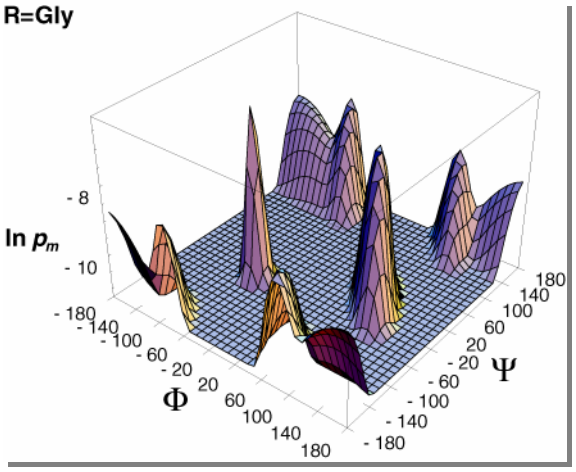
Mainchain Terms for Loop Modeling

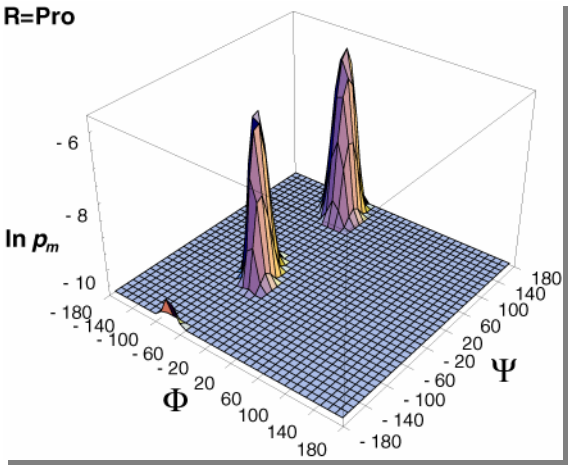




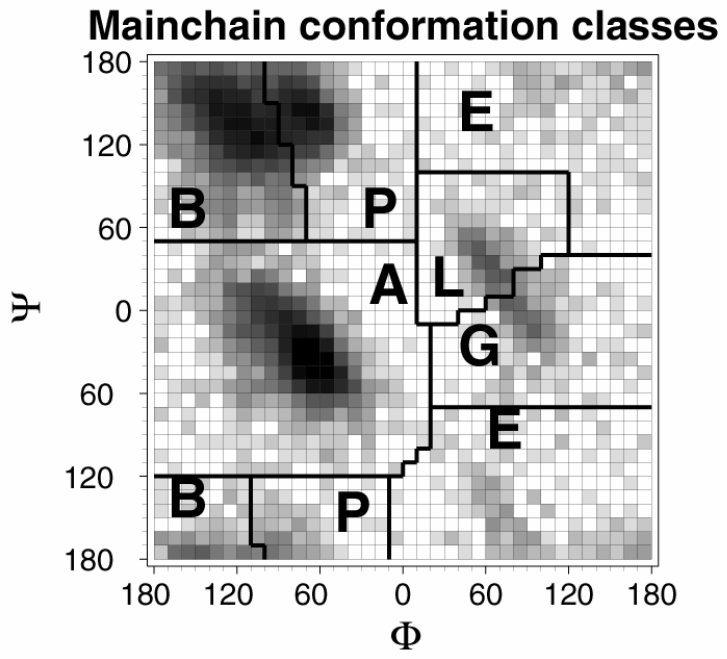


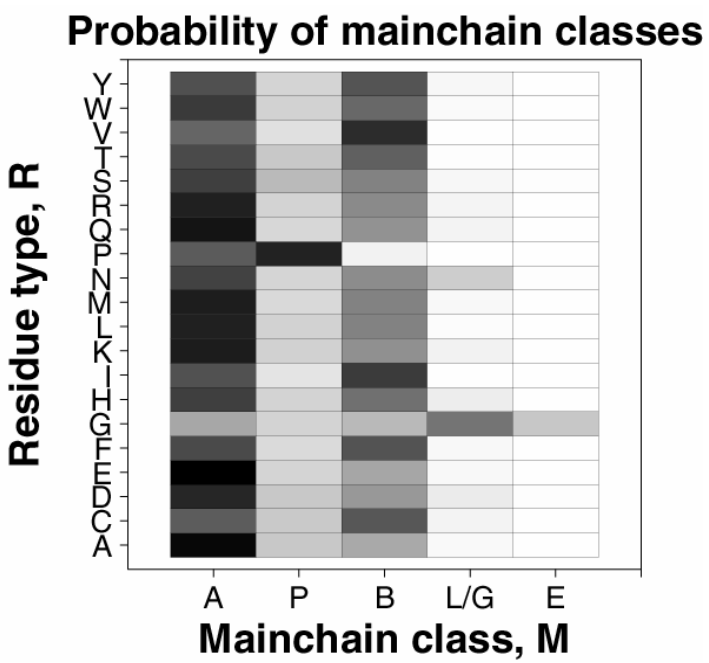






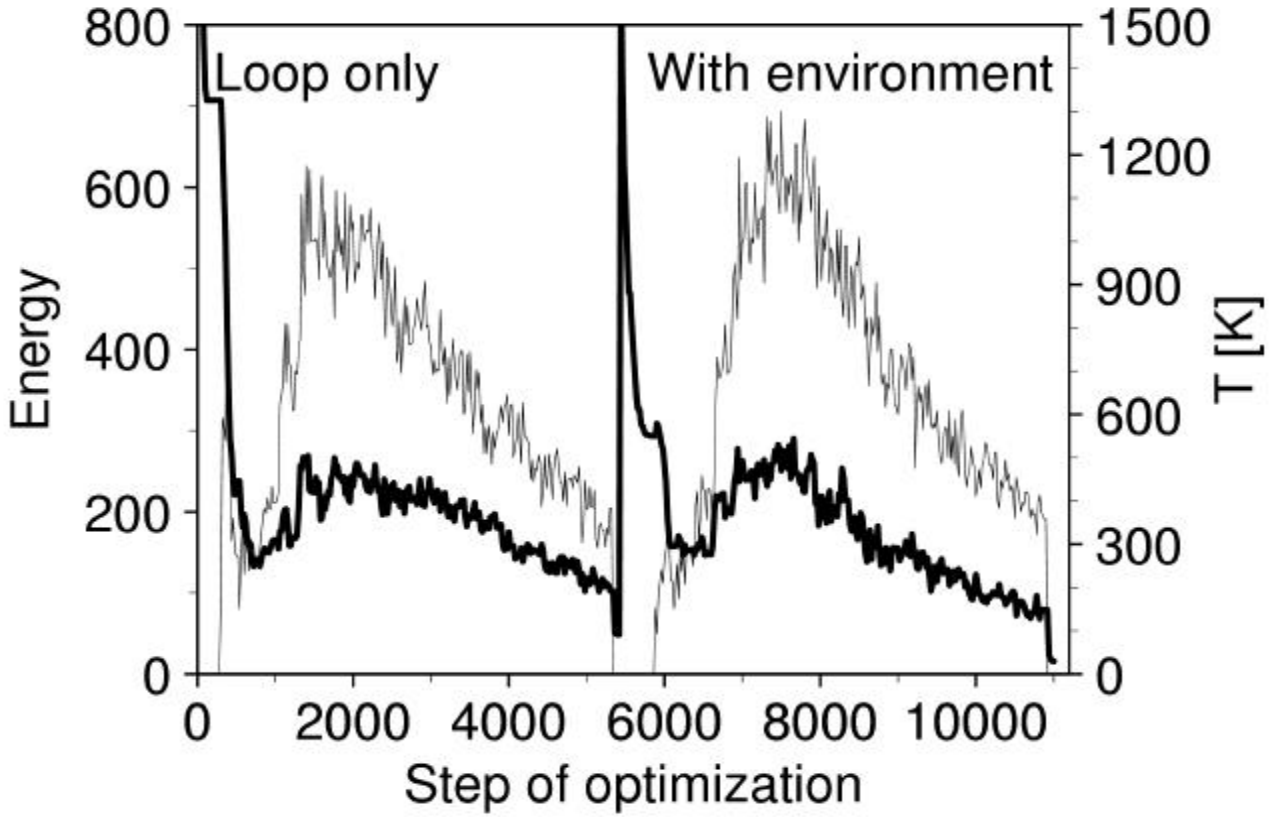
Mainchain terms for loop modeling





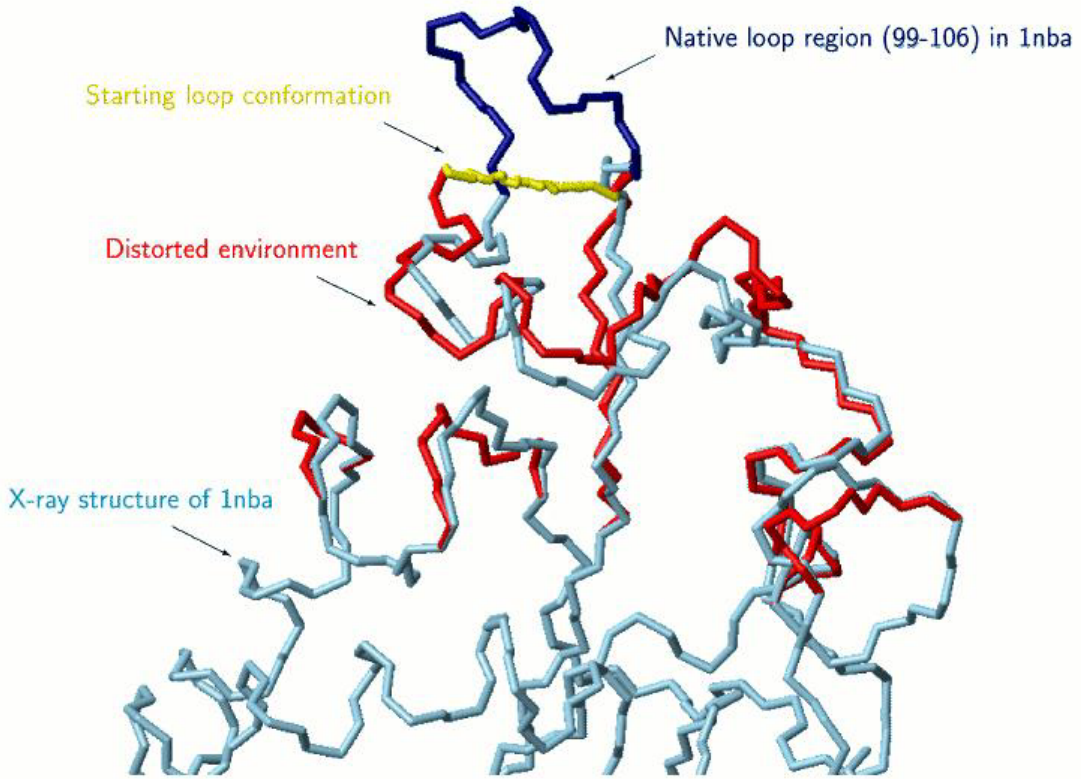
Optimization of Objective Function

Many different combinations of objective function terms were explored. The objective function was optimized with a combination of conjugate gradients method and molecular dynamics simulation with simulated annealing in a two step process: without and with the environment.

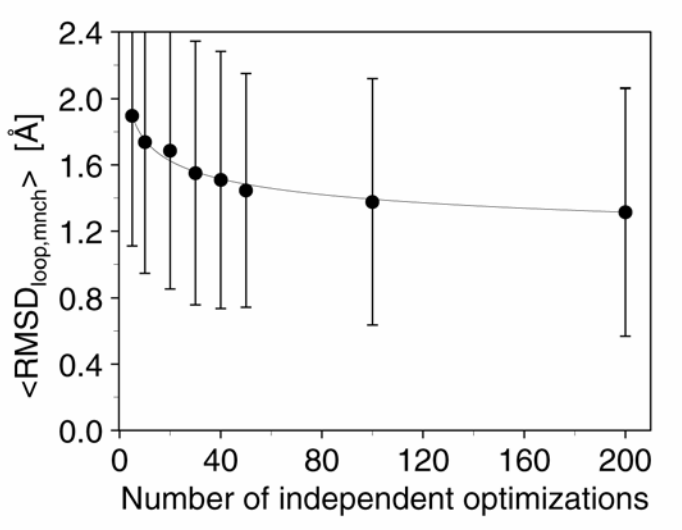


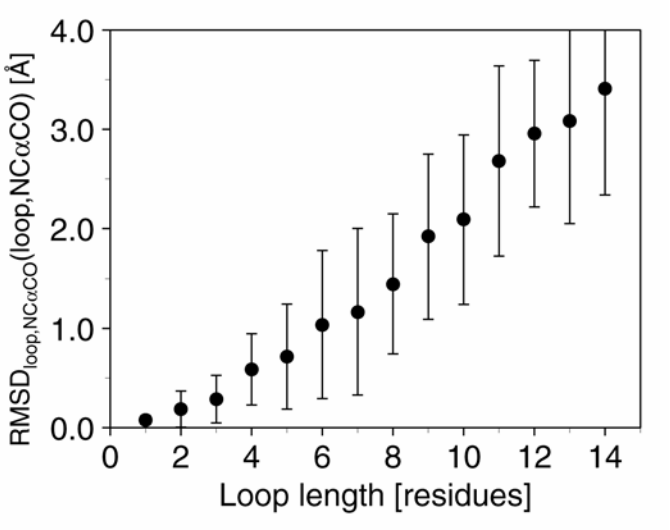
Optimization of Objective Function

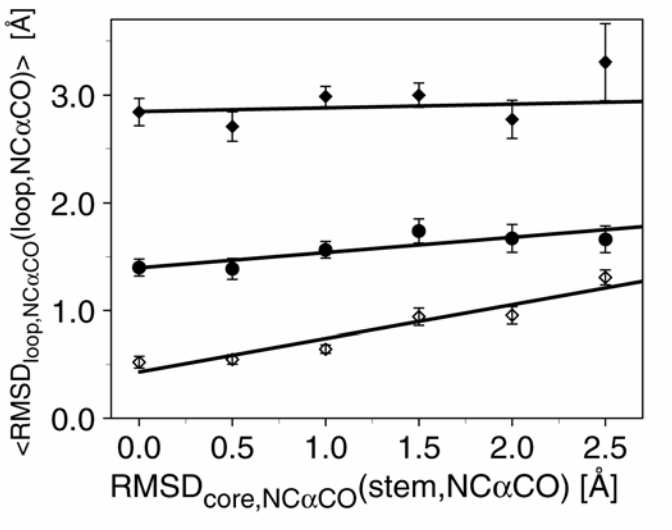
- Test set: 40 randomly selected loops of known structures, for each length from 1 to 14 residues.
- Starting conformation: Loop atoms were spaced evenly on a line spanning the two anchor regions, then randomized by $\pm$ 5 Å.
- To simulate real comparative modeling situations, performance of the loop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



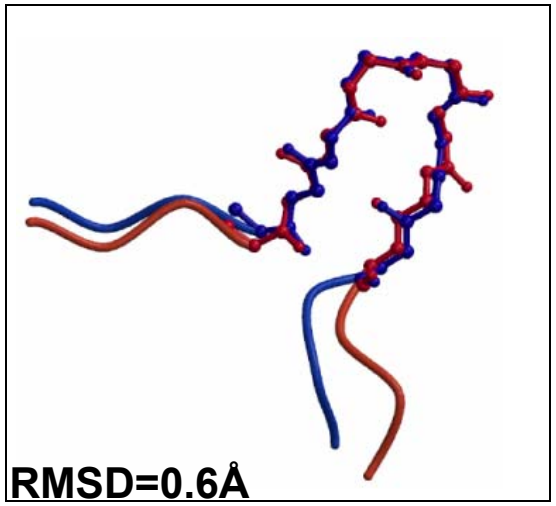
Accuracy of loop models





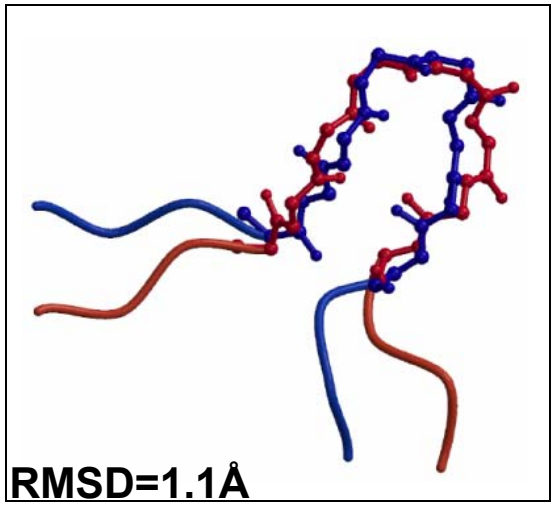


Accuracy of Loop Modeling



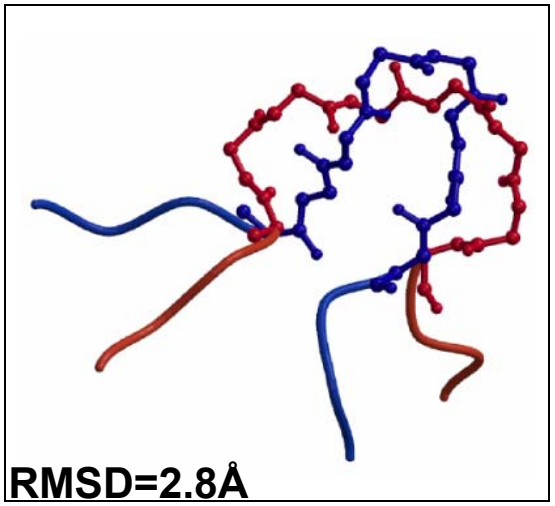
HIGH ACCURACY (<1Å)

50% (30%) of 8-residue loops



MEDIUM ACCURACY (<2Å)

40% (48%) of 8-residue loops

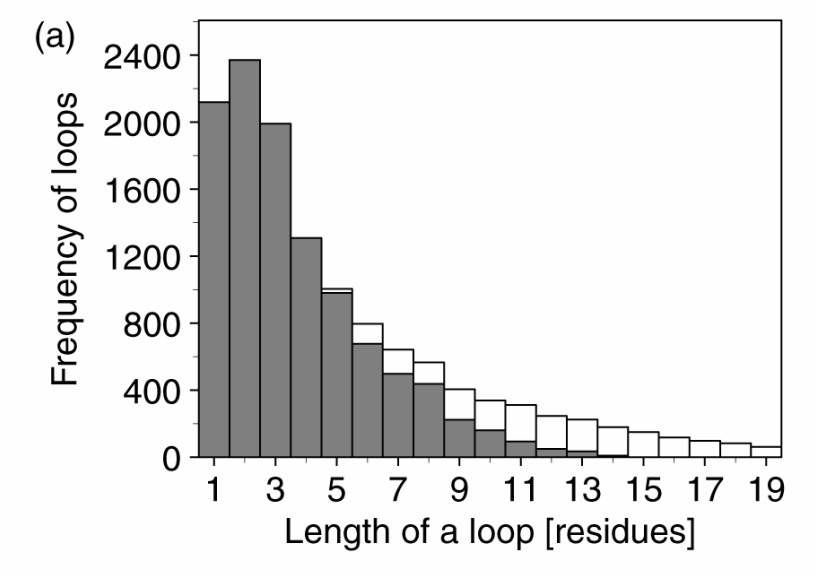


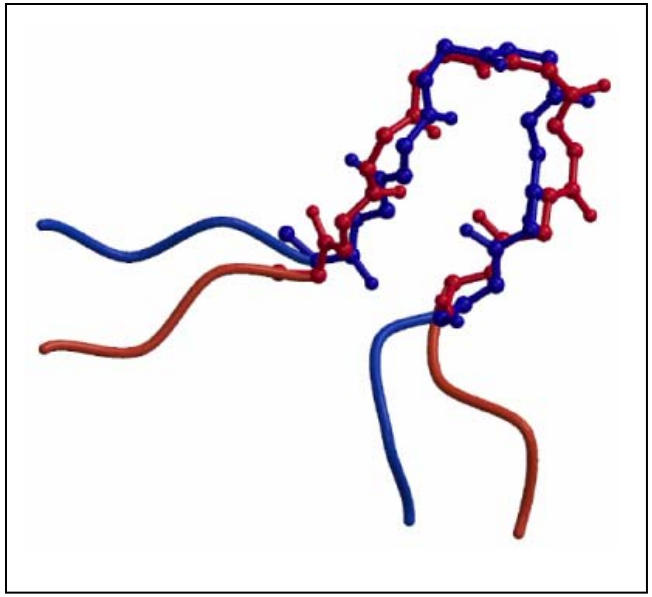
LOW ACCURACY (>2Å)

10% (22%) of 8-residue loops

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

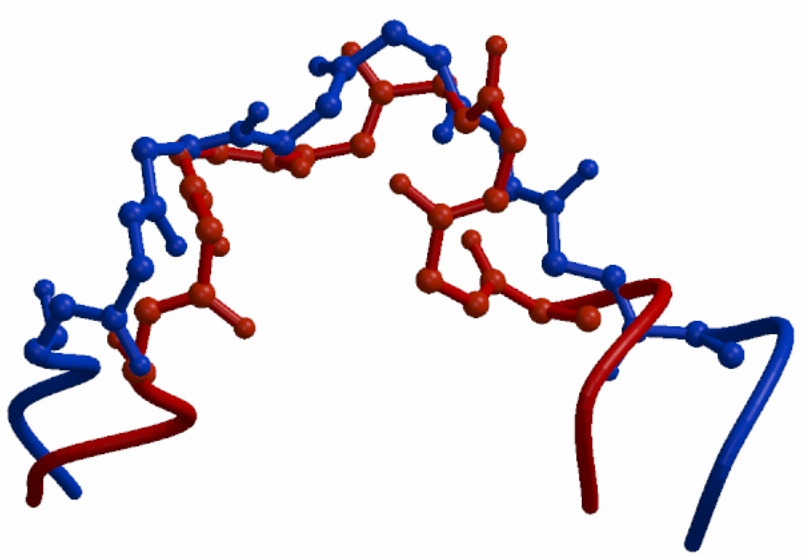
Fraction of Loops Modeled With at Least Medium Accuracy



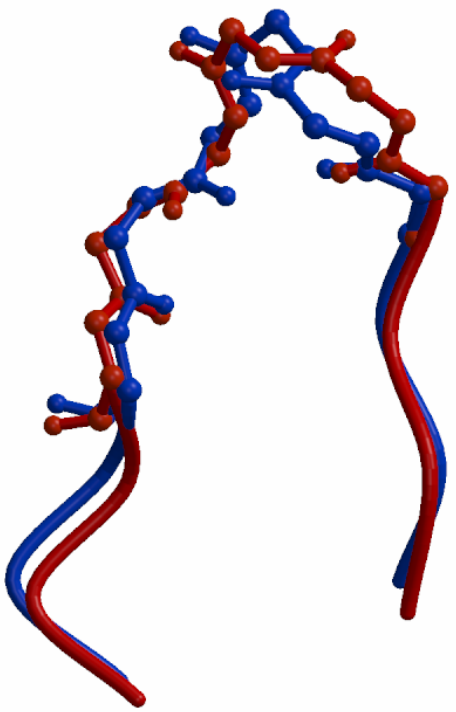


Problems in Practical Loop Modeling

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

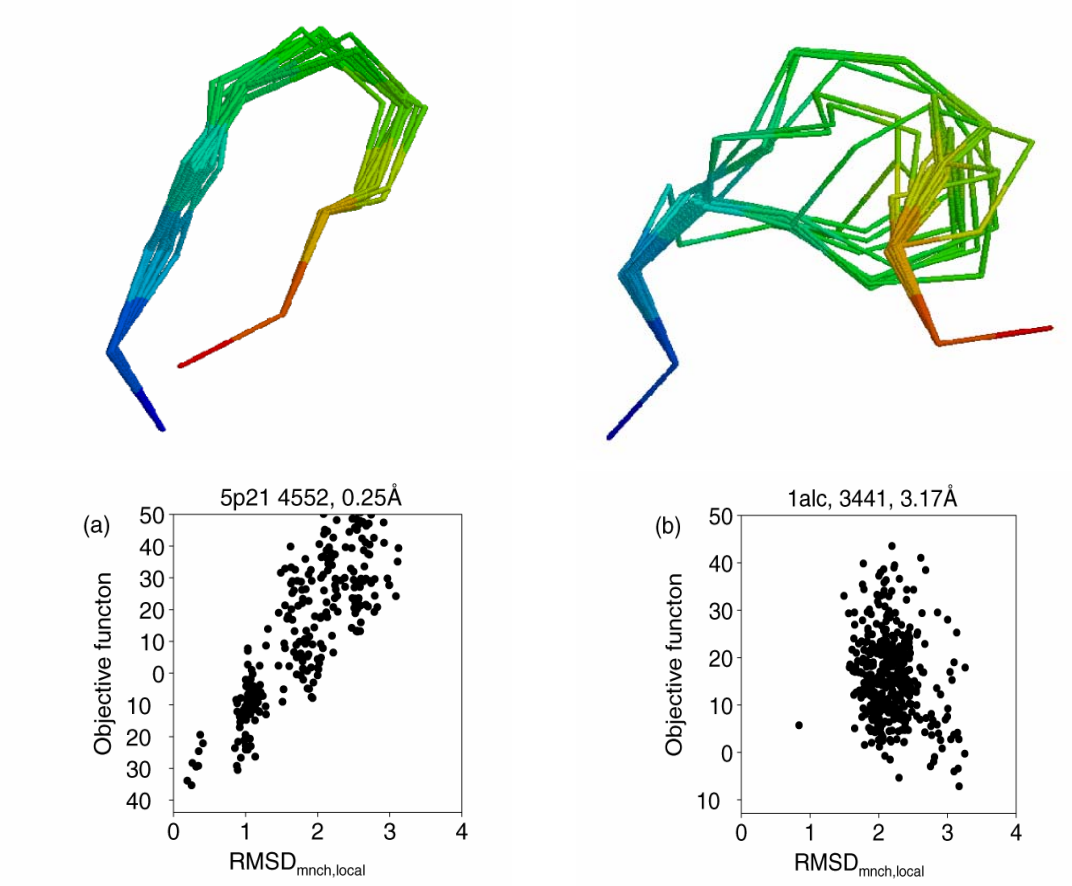


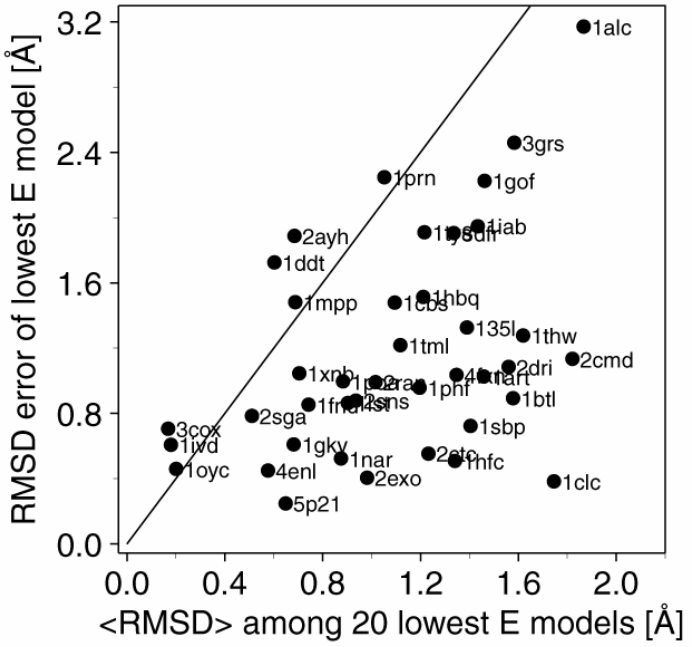
T0076: 46-53
$RMSD_{mnch}$ loop = 1.37 Å
$RMSD_{mnch}$ anchors = 1.52 Å



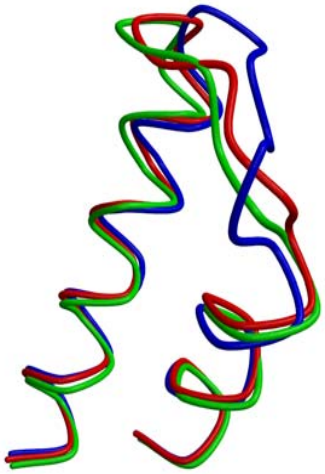
T0058: 80-85
$RMSD_{mnch}$ loop = 1.09 Å
$RMSD_{mnch}$ anchors = 0.29 Å

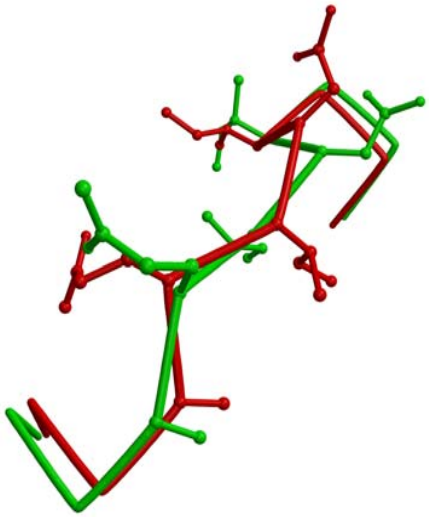
Assessing Accuracy of Loop Models

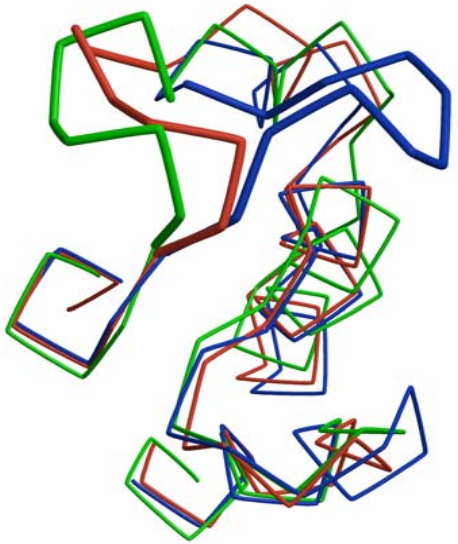




Examples from CASP 3/4





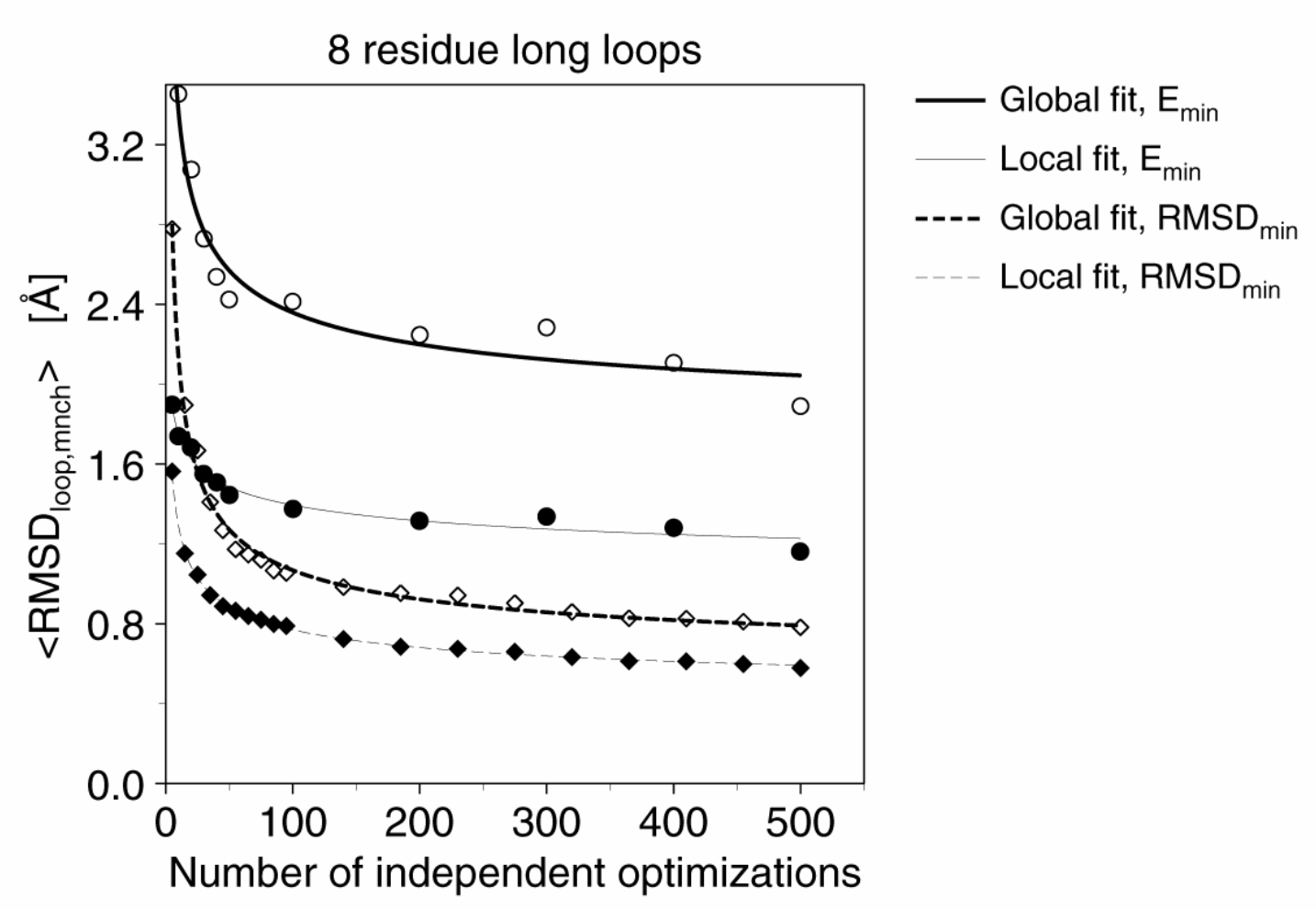


RMSD	3-3 anchor residues	global MNCH	global ALL	local MNCH	local ALL
X-ray/model	0.76	1.28	2.48	1.05	1.90
X-ray/template	2.38	2.69	4.22	1.94	3.53
model/template	2.10	2.75	4.29	1.64	2.92

S. pombe contractile ring protein Cdc4p.
33 % seq id. 8 residue long loop,
$RMSD_{global}$: 3.64, $RMSD_{local}$:1.36Å

Adding solvent effect to loop modeling

Accuracy of loop models as a function of amount of optimization



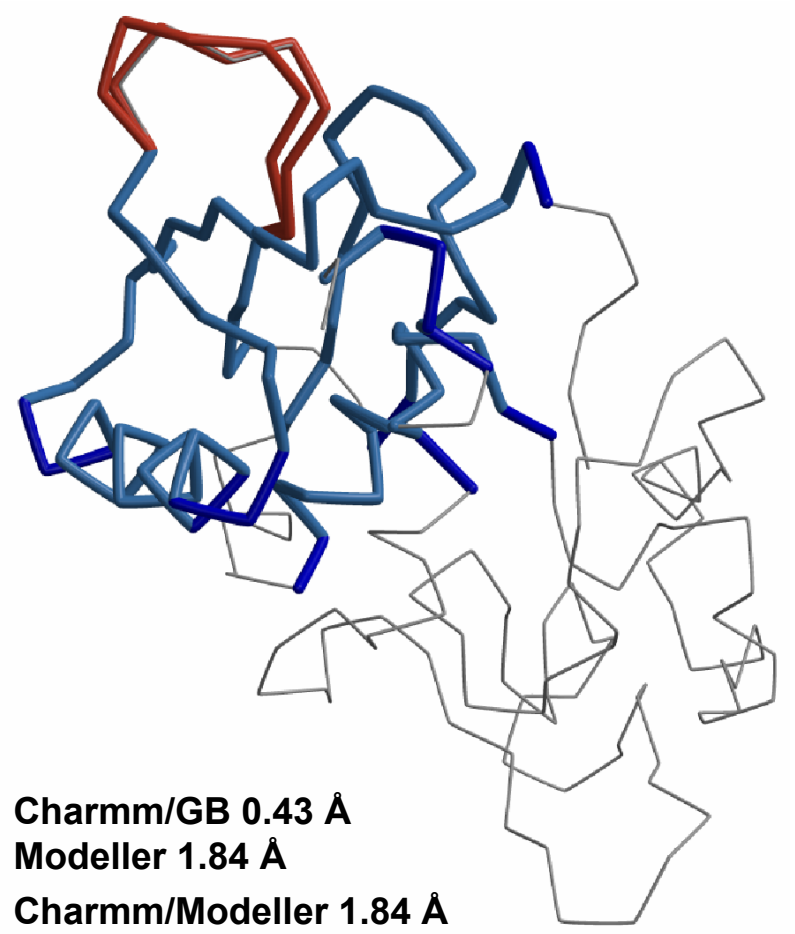
Refining with CHARMM/GB

sample loop models

Select loop models

Select loop and its environment

Minimize loop and its environment with CHARMM/GB

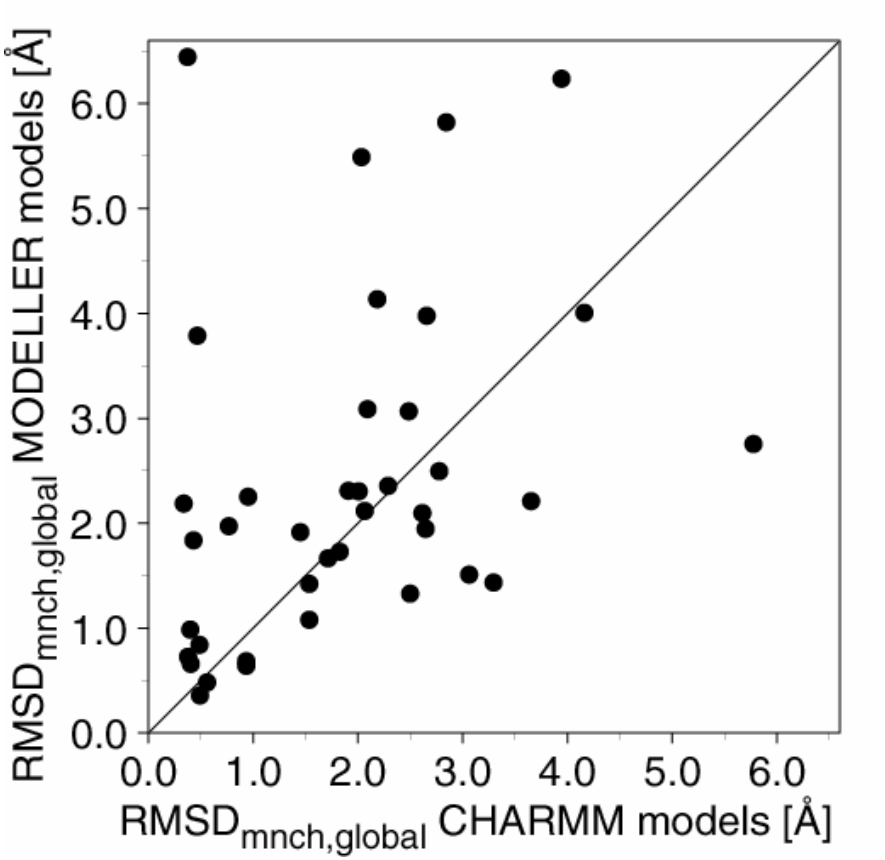


Charmm/GB 0.43 Å
Modeller 1.84 Å
Charmm/Modeller 1.84 Å

50 steps of the steepest descent relaxation, followed by 2000 steps of ABNR minimization or until convergence (D < 10-4 kcal/mol).

Refining models with CHARMM/GB

Improving ranking

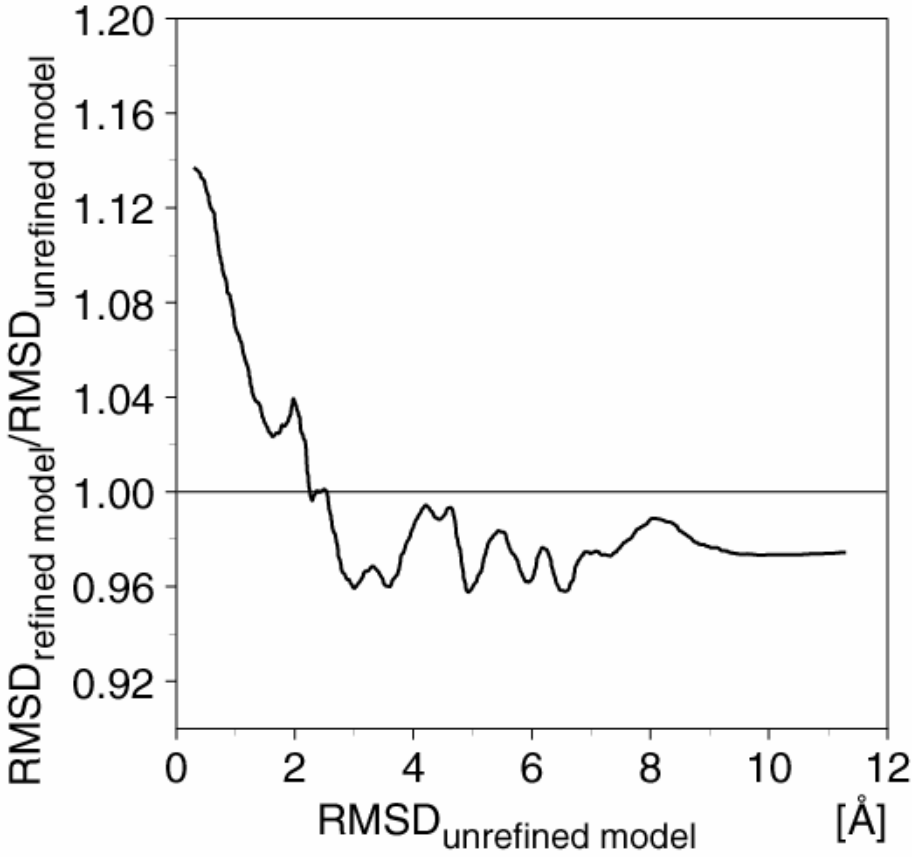


<rmsd$_{GLOBAL/LOCAL}$>

CHARMM/GB 1.87 /1.07 Å

MODELLER 2.36 /1.29 Å

Improving model quality



55%-63%

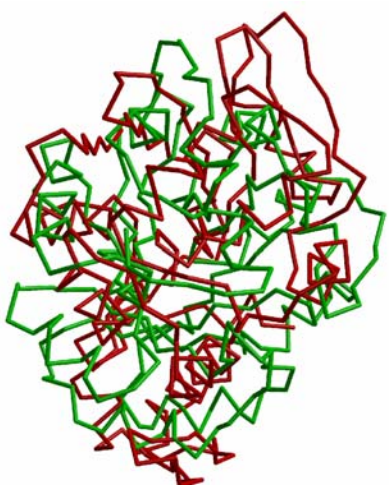
45%-37%

A. Fiser, M. Feig, C.L. Brooks, A. Sali (2002) *Acc. Chem. Res.*

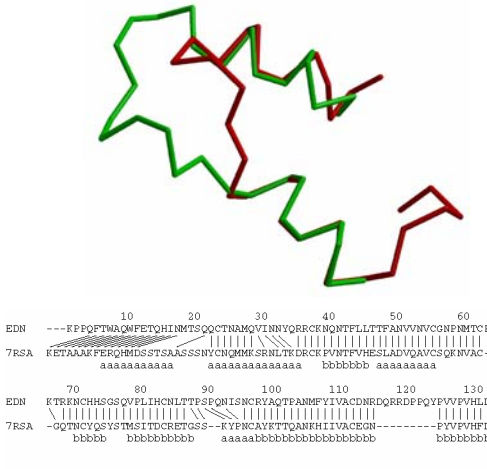
Typical Errors in Comparative Models

MODEL
X-RAY
TEMPLATE

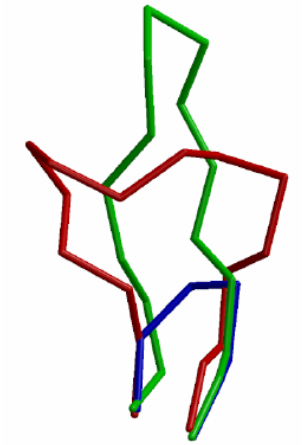
Incorrect template



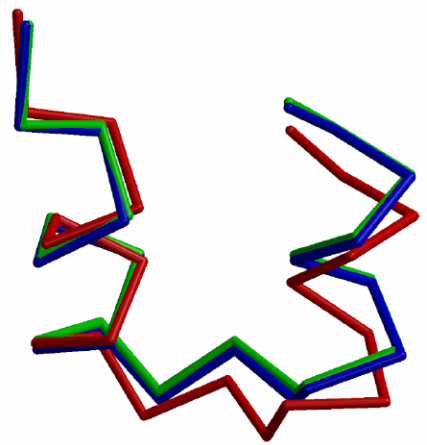
Misalignment



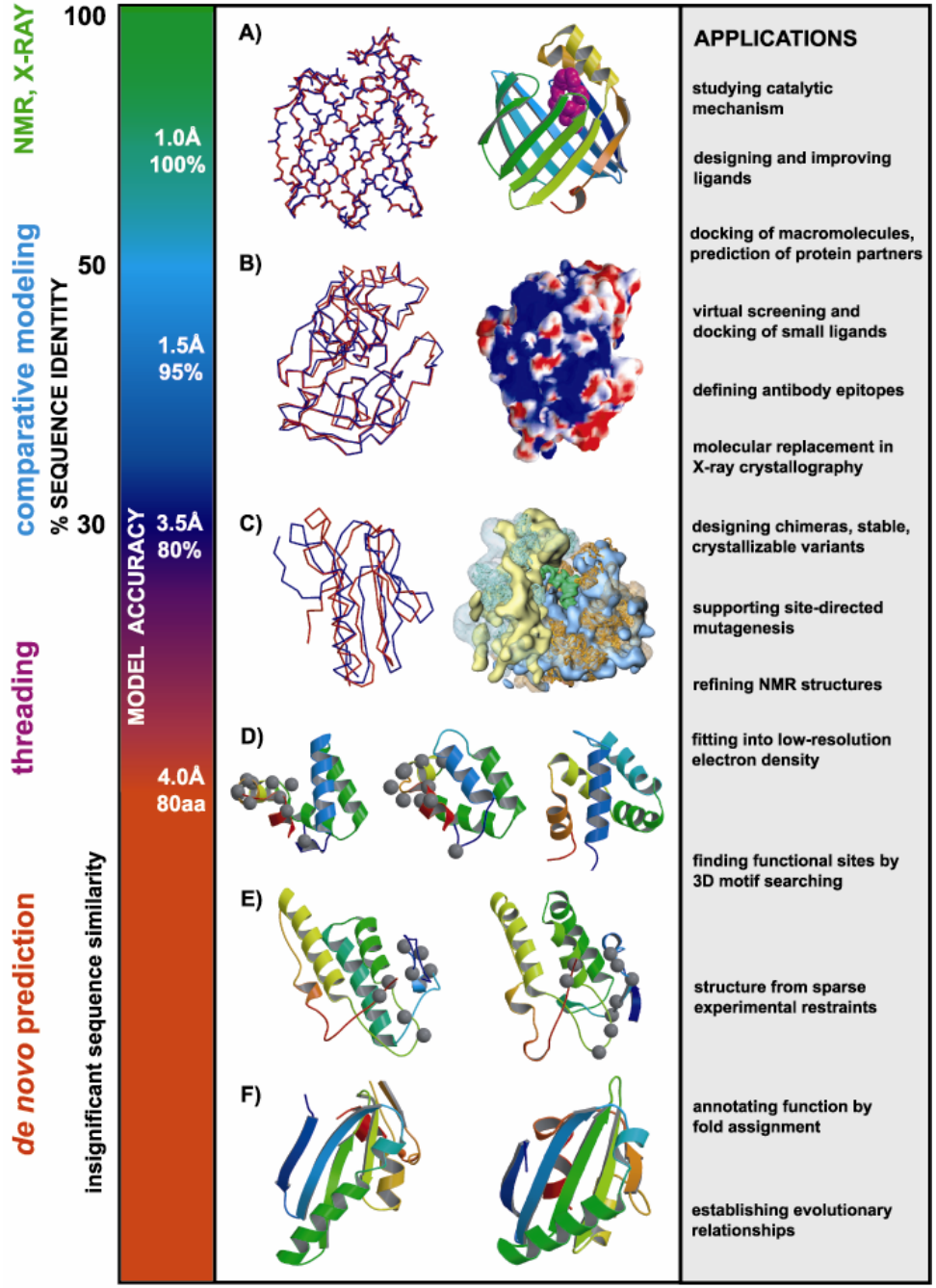
Region without a template



Distortion in correctly aligned regions



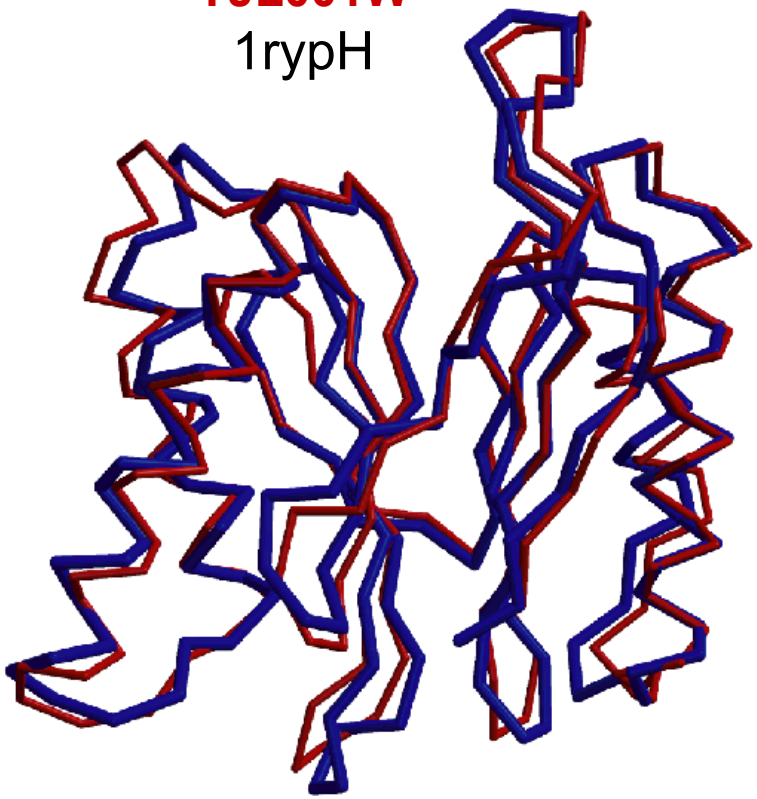
Side chain packing



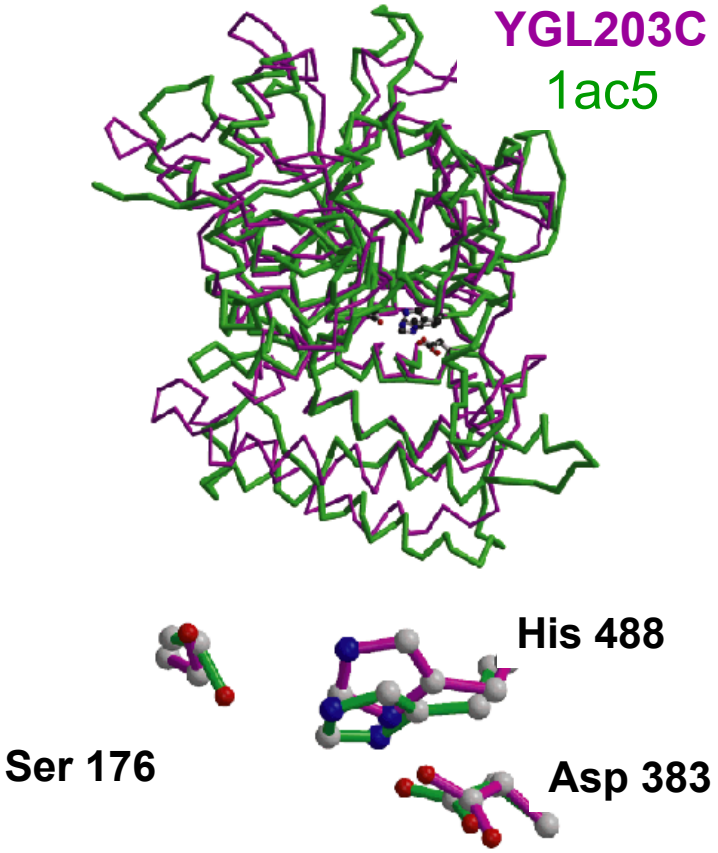
D. Baker & A. Sali. Science, 2001

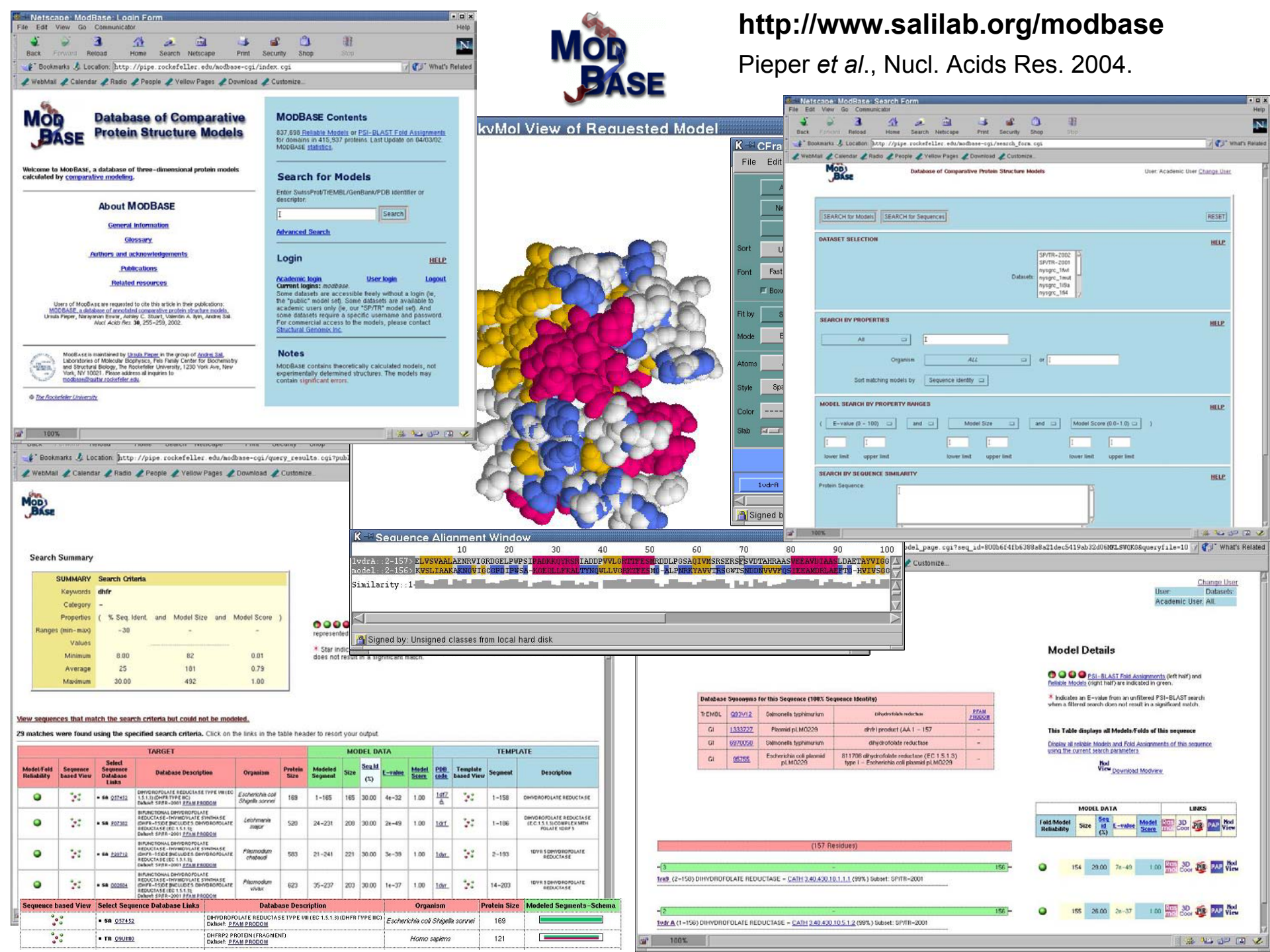
Some Models Can Be Surprisingly Accurate (in Some Regions)

24% sequence identity

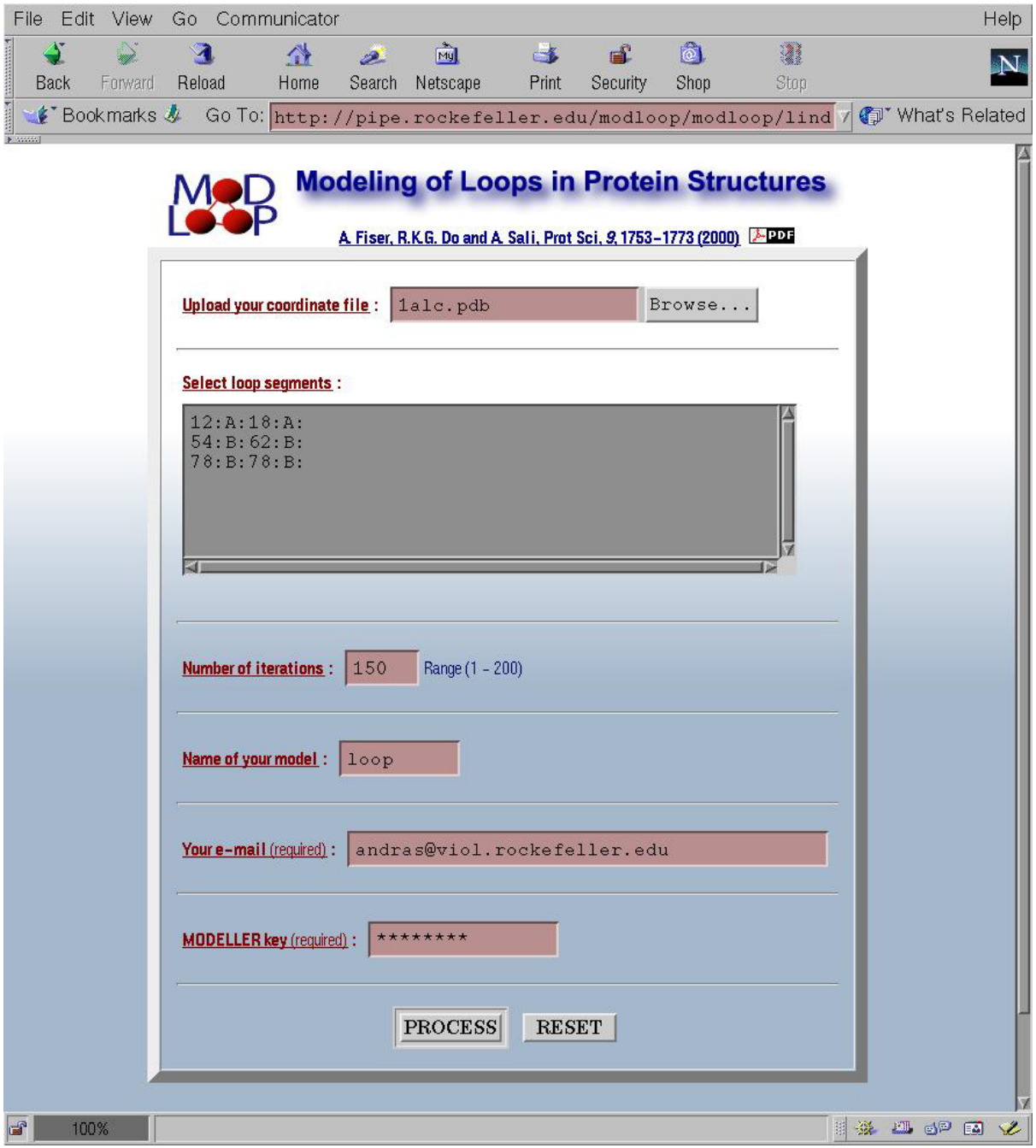


25% sequence identity




http://www.salilab.org/modbase
Pieper et al., Nucl. Acids Res. 2004.
MODBASE
Database of Comparative Protein Structure Models
MODBASE Contents
Search for Models
About MODBASE
Login
Notes
kvMol View of Requested Model
Sequence Alignment Window
Search Summary
Model Details

http://www.fiserlab.org/



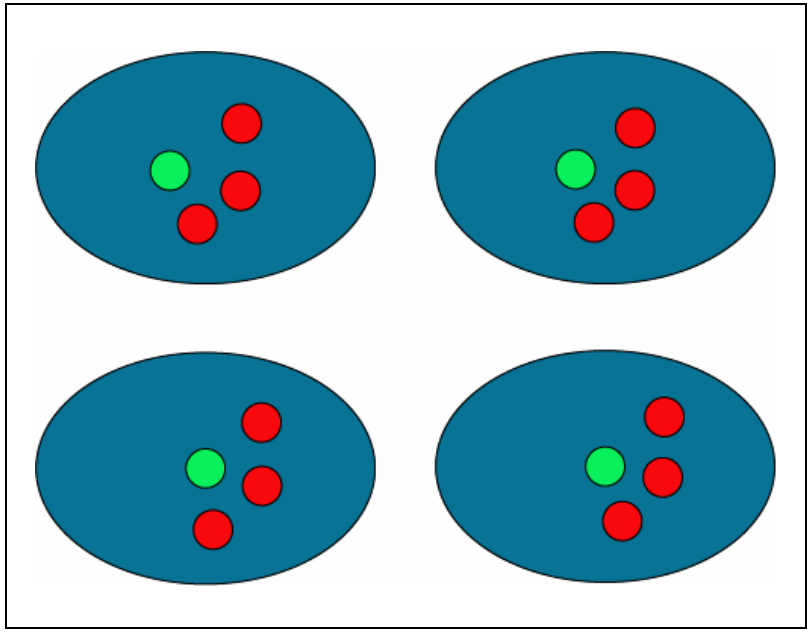
Fiser, A. and Sali, A. *Bioinformatics.* (2003) **18**(19) 2500-01

Applications I.

Structural genomics

Structural Genomics

Characterize most protein sequences (red) based on related known structures (green).



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

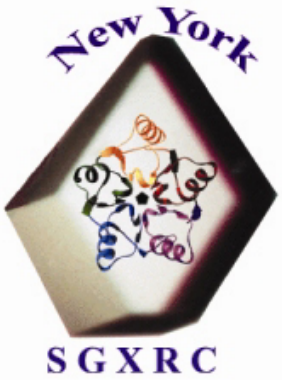
Structural Genomics

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

- **Size of the problem:**
  - There are a few thousand domain fold families.
  - There are ~20,000 sequence families (30% sequence id).

- **Solution:**
  - Determine protein structures for as many different families as possible.
  - Model the rest of the family members using comparative modeling

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



New York Structural Genomics Research Consortium

Mission Statement

To develop and use the technology for high-throughput structural and functional studies of proteins.

Participating Research Groups

Albert Einstein College of Medicine
Mark R. Chance
Steve Almo
Anne Bresnick
Andras Fiser

Brookhaven National Laboratory
Robert Sweet
Jian-Sheng Jiang
F. William Studier
S. Swaminathan

Columbia University
Lawrence Shapiro

Structural Genomix, Inc
Stephen K. Burley

The Rockefeller University
Terry Gaasterland

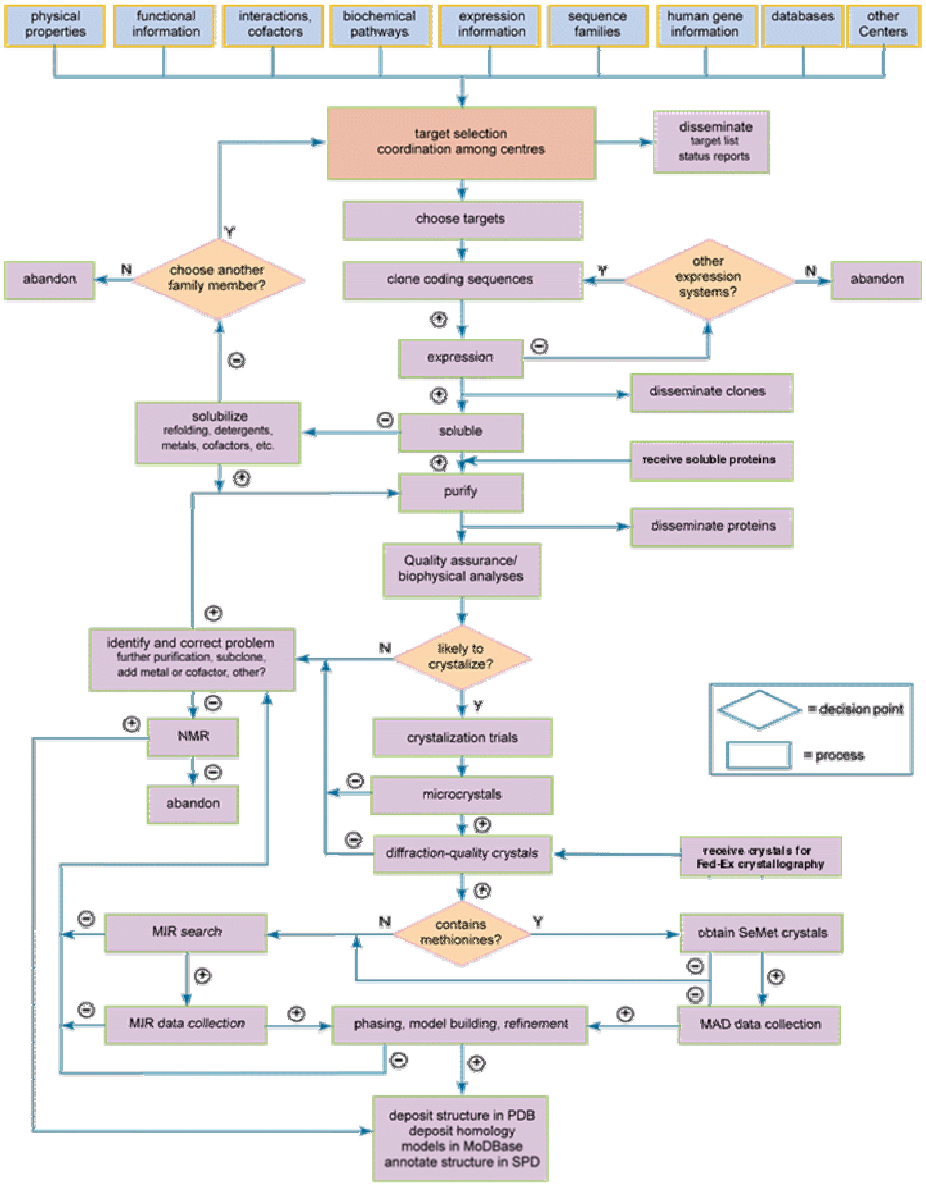
University of California, San Francisco
Andrej Sali

Weill Medical College of Cornell University
Christopher Lima

Public Target Information

Public Target Progress Report
Download: Public Target Progress Report in XML Format

Home Proposal Publications Flowchart IceDB Tools Contact


physical properties
functional information
interactions, cofactors
biochemical pathways
expression information
sequence families
human gene information
databases
other Centers
target selection coordination among centres
disseminate target list status reports
choose targets
clone coding sequences
other expression systems?
N
abandon
Y
choose another family member?
N
abandon
expression
disseminate clones
soluble
solubilize refolding, detergents, metals, cofactors, etc.
receive soluble proteins
purify
disseminate proteins
Quality assurance/ biophysical analyses
likely to crystalize?
N
identify and correct problem further purification, subclone, add metal or cofactor, other?
Y
= decision point
= process
NMR
abandon
crystalization trials
microcrystals
diffraction-quality crystals
receive crystals for Fed-Ex crystallography
contains methionines?
N
Y
MIR search
obtain SeMet crystals
MIR data collection
phasing, model building, refinement
MAD data collection
deposit structure in PDB
deposit homology models in MoDBase
annotate structure in SPD

Comparative Protein Structure Modeling with NYSGXRC Structures (January 23, 2003)

NYSGXRC SOLVED-STRUCTURE TEMPLATE					MODBASE	NYSGXRC ACCEPTABLE MODEL DATA					
Target_ID	Protein Name / Comment	GI or Swissprot Code	PDB Code	Protein Size	Total Models	Total Models	Min. Seq. ID	Max. Seq. ID	# Models >50% Seq. ID	# Models 30-50% Seq. ID	# Models <30% Seq. ID
T9	Similar to putative GTP-binding protein	O13998	1NI3	392	1253	76	13	63	6	40	30
T132	Putative cell cycle protein mesJ	P52097	1NI5	432	870	150	11	56	1	15	134
T503	Conserved hypothetical protein YDCE from Bacillus Subtilis	P96622	1NE8	116	52	36	21	81	9	10	17
TBA	unc78	TBA	TBA	1315	693	0	0	0	0	0	0
TBA	phlp1, a Major Timothy Grass Pollen Allergen	TBA	1N10	228	191	162	20	92	18	53	91
T746	phlp6	P43215	TBA	209	249	11	25	47	0	10	1
P089	Hypothetical 32.1 kDa protein in ADH3-RCA1 intergenic region	Q04299	1NJR	284	1	0	0	0	0	0	0
P096	Hypothetical 28.8 kDa protein in PSD1-SKO1 intergenic region	P53889	1NKQ	259	301	161	14	43	0	82	79
TBA	putative_thioesterase_(comA) polymer	TBA	TBA	138	479	147	9	89	17	16	114
TBA	hypothetical_protein_(yqeU) polymer	TBA	TBA	241	85	76	18	81	6	16	54
T299	URACIL-DNA GLYCOSYLASE FROM T. MARITIMA (Hypothetical protein TM0511)	Q9WYY1	1L9G	192	110	84	13	49	0	49	35
P007	Hypothetical 29.1 kDa protein in URA7-POL12 intergenic region	P38197	1B54	257	53	44	27	43	0	34	10
P008	Pyridoxamine 5'-phosphate oxidase	P38075	1CI0	228	1374	1266	8	99	103	33	1130
P018	Hypothetical 32.5 kDa protein YLR351C	P49954	1F89	291	302	251	13	54	1	38	212
P044a	L-allo-threonine aldolase	GI: 4982322	1JG8	343	1049	923	10	46	0	13	910
P068	Hypothetical 33.9 kDa esterase in SMC3-MRPL8 intergenic region	P40363	TBA	299	804	119	10	52	1	14	104
P097	Hypothetical 27.5 kDa protein in SPX19-GCR2 intergenic region	P40165	1JZT	246	1401	6	14	41	0	4	2
P100	Diphosphomevalonate decarboxylase	P32377	1FI4	396	422	139	9	68	2	26	111
P102	Glutathione synthetase Apo	Q08220	1M0T	491	140	25	30	42	2	23	0
P102a	Glutathione synthetase Lig	Q08220	1M0W	491	33	30	32	39	0	30	0
P109a	Isopentenyl-diphosphate delta-isomerase (IPP isomerase)	GI: 6225535	1I9A	182	947	417	10	72	5	18	394
P111a	Translation initiation factor 6	Q60357	1G61	228	37	35	29	46	0	34	1

IceDB Report Display

Target ID: P097
Target Iteration: 1

Go to Upload Report Page
Download Report as PDF Format
To make changes to other experimental types:
Go to Add/Edit/Upload Data - Start Page

Target Identifier: P097

Protein Name: HYPOTHETICAL 27.5 KDA PROTEIN IN SPX19-GCR2 INTERGENIC REGION

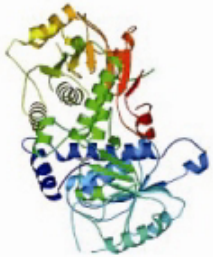
Organism: Saccharomyces cerevisiae

PDB Code: 1JZT

Rationale for Target Selection: Unknown function and represents a domain from a number of protein families (ProDom accession PD005835).

Method of Structure Determination: Se-Met MAD method and NCS density averaging.

Structure Description: The structure of P097 is a three-layer a-b-a sandwich. The two molecules related by NCS in the asymmetric unit form a tightly packed dimer. Each monomer consists of eight b-strands and nine a-helices. The order of b-strands is 32145678 according to the SCOP classification.



Comparisons of Structurally Similar Proteins in PDB: P097 represents an unusual Rossmann fold. A typical NAD-binding Rossmann fold should have six b-strands forming an open twisted parallel b sheets in the middle and two a-helices on both sides (3LDH). The secondary structure of the Rossmann fold is b1-aA-b2-aB-b3-aC-b4-aD-b5-aE-b6 and the order of b-strands is 321456 (SCOP). The NAD-binding proteins usually consist of two domains: a dinucleotide (or

Structural Genomics Initiatives **Structural Genomics Link at the PDB**

TargetDB

Target Search for Structural Genomics

TargetDB is a target registration database that was originally developed to provide registration and tracking information for NIH P50 structural genomics centers. TargetDB has now been expanded to include target data from worldwide structural genomics and proteomics projects. The scope of TargetDB is to provide timely status and tracking information on the progress of the production and solution of structures.

Sequences from the NIH P50 and other structural projects have been loaded into the TargetDB database and can be searched using the form below. TargetDB is updated weekly. All targets are available for download in XML format here.

A new Target Status Query Feature is now available, please click here.

Target sequence lists are also maintained at the following sites:

| BIGS | BSGC | BSGI | JCSG | MCSG | MSGP | NESG | NYSGRC | OPPF | PSF | RIKEN | S2F | SECSG | SGPP | TB | YSG |

Using the Target Search Form:

- Enter text and/or select menu options in the form below to define the desired target search, select a result format, and press the SUBMIT button to execute the query.
- All form attributes are optional. If no options are entered a query will return all of the NIH target entries in the database.
- Click on any attribute name for an explanation and examples of the attribute.
- For a FASTA sequence comparison, enter the one-letter code sequence into the sequence text box.

Project Target ID:

Status: Any

Site: All

Include Data From: All Structural Genomics Centers

Target Data Updated: after Month Day Year
before Month Day Year

Protein Name:

Source Organism:

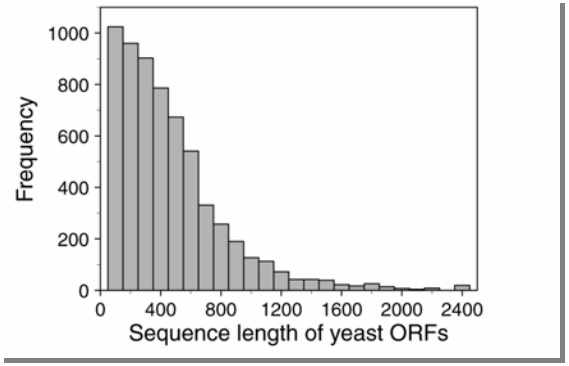
Limitations of structural genomics: Quaternary structure

Proteins are modular, ~2.7 domains per protein.

Evolution shuffles domains.

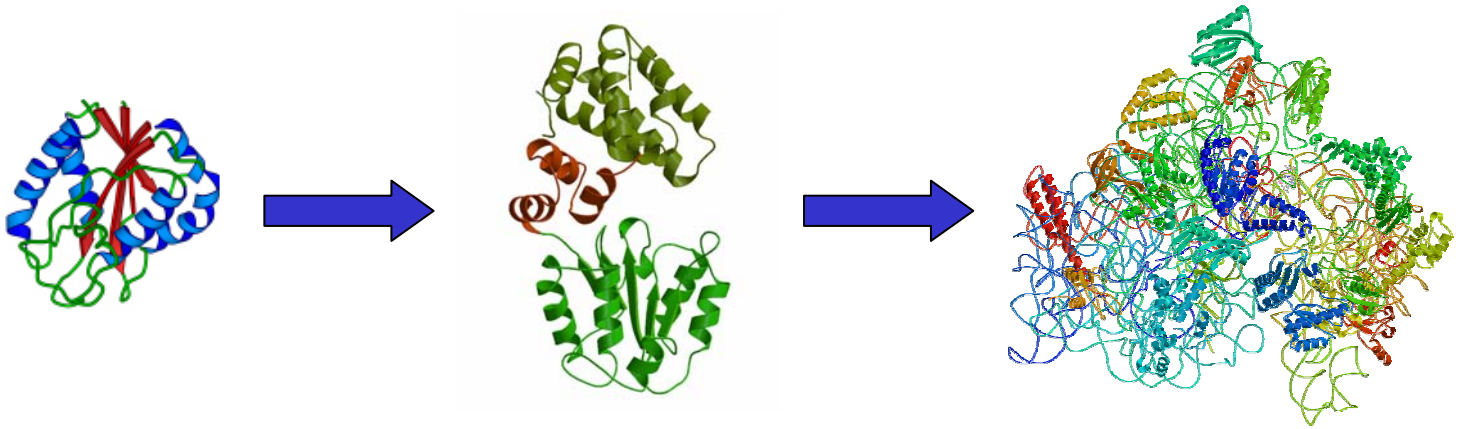
24% of domains/residues in 57% of proteins are modeled

Structural genomics is determining structures of domains, usually not proteins, definitely not assemblies.



Average Length		
Protein	Domain	Model
472	175	192

Thus, there is a great need for methods for docking of domains into proteins and of proteins into macromolecular assemblies.



domains proteins assemblies

Acknowledgements

- Andrej Sali (Rockefeller/UCSF)

- Charlie Brooks (Scripps)
- Michael Feig (Scripps)

- Steve Burley (Rockefeller/SGX)
- Steve Almo (Albert Einstein Coll. Of Med.)
- Mark Chance (Albert Einstein Coll. Of Med.)

Lab
- Guiping Xu
- Eduardo Fajardo
- Dmitry Rykunov
- Brajesh Kumar Rai
- Narcis Fernando-Fuentes
- Rotem Rubinstein

Reviews

Fiser, A.
Protein structre modeling in the proteomcis era
Exp. Rev. in Proteomics . (2004) 1, 89-102

Fiser, A. and Sali, A.
MODELLER: Calculating and refining homology models
Methods Enzymol. (2003) 374,463-493

Fiser, A., Sanchez, R., Melo, F. and Sali, A.
Comparative protein structure modeling.
in *Computational Biochemistry and Biophysics*, (2001) pp. 275-312,
Marcel Dekker. Eds. M. Watanabe, B. Roux, A. MacKerell, and O. Becker.