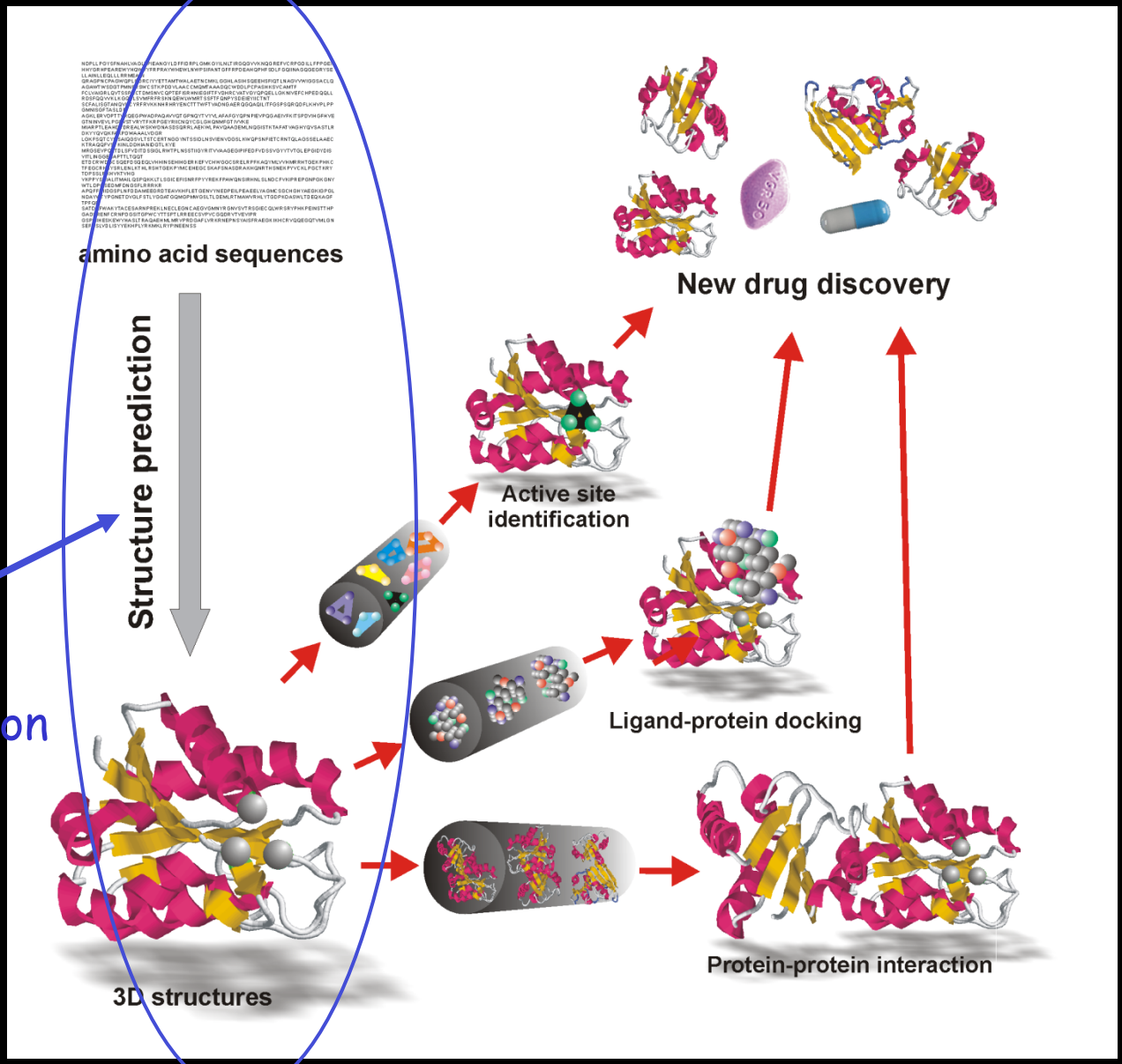
Protein folding & protein structure prediction

Yang Zhang

Department of Computational Medicine & Bioinformatics,
Department of Biological Chemistry
University of Michigan

Table of Contents

Sequence-to-Structure-to-Function Paradigm



Protein structure determination

Milestones of protein structure determination

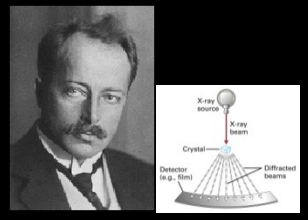
Discover X-ray
(Roentgen)
(NP: 1901)

PDB at BNL
with 2 proteins

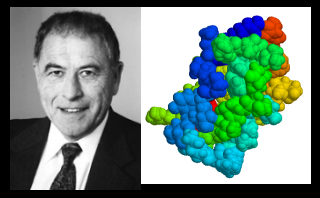
RCSB PDB
PROTEIN DATA BANK

PDB with
184 structures

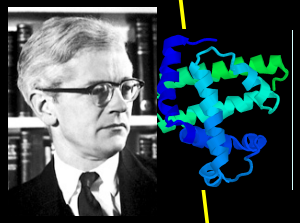
X-ray diffraction
by crystal
(von Laue)
(NP: 1914)



First structure
by NMR
(Wuthrich)
(NP:2002)



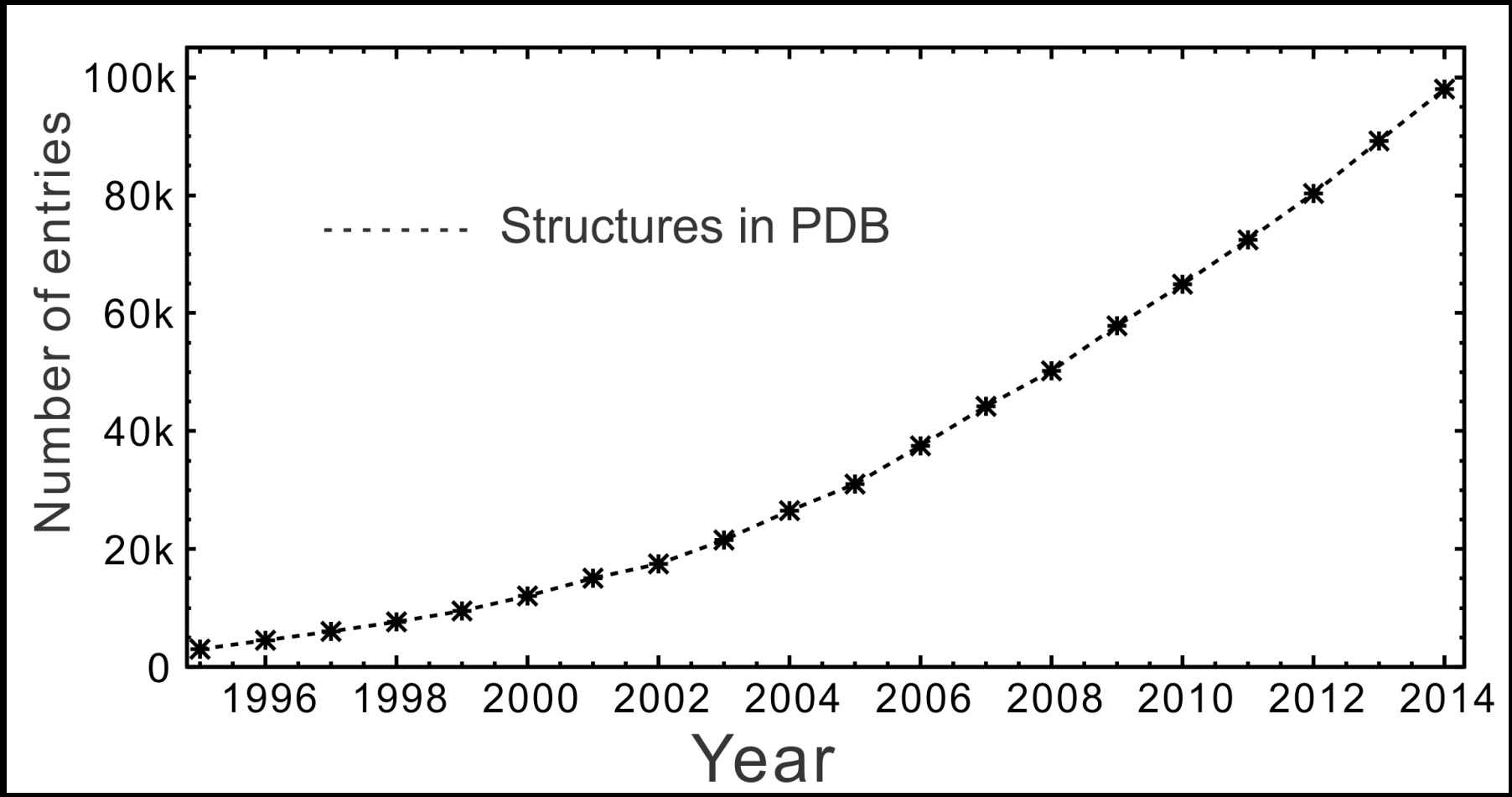
First structure
by x-ray
(Kendrew)
(NP: 1962)



PDB with
103,627
structures

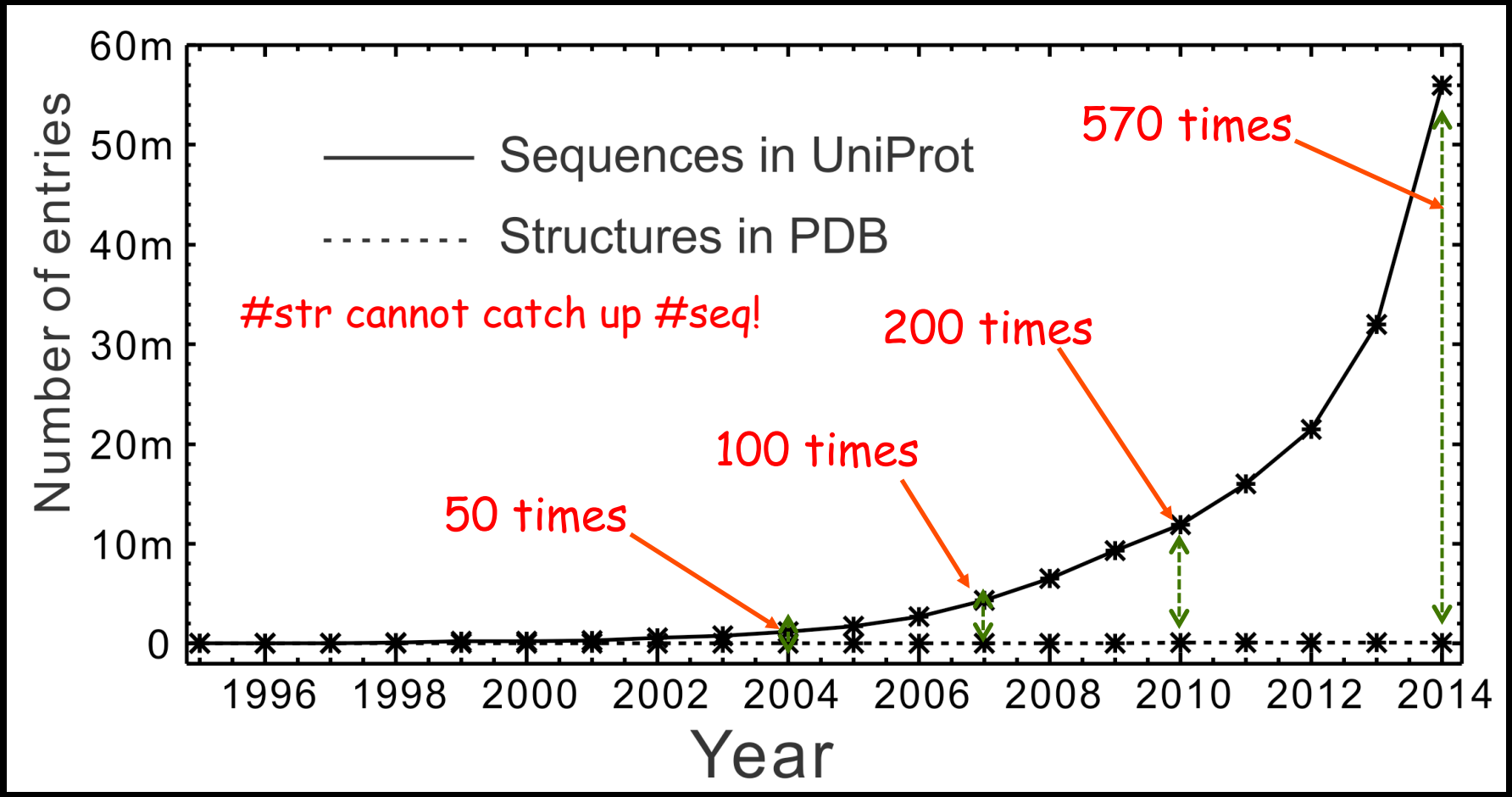
1895 1912 1958 1971 1980 1985 2014

#structure increases rapidly in PDB



30 new protein structures solved per day

#structure lag far behind #sequences

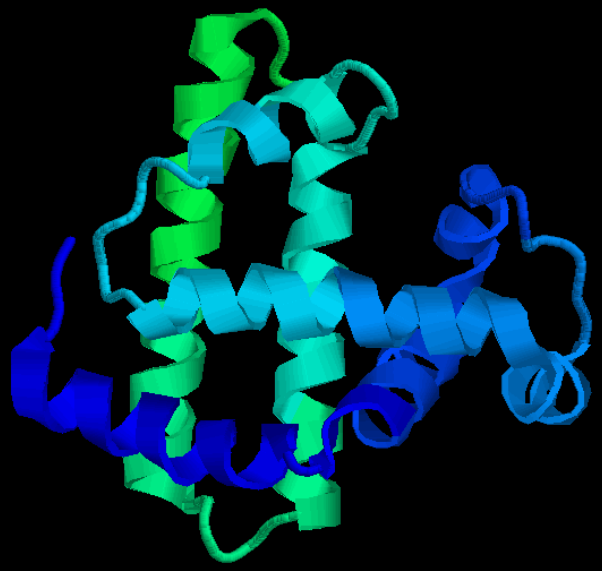


Solving one structure costs ~$250,000-$500,000
Determining one sequence costs ~$1,000-$4,000

Protein structure prediction

MVLSEGEWQLVLHVWAKV
EADVAGHGQDILIRLFKSHP
ETLEKFDRVKHLKTEAEMK
ASEDLKKHGVTVLTALGAIL
KKKGHHEAELKPLAQSHAT
KHKIPIKYLEFISEAIIHVLHS
RHPGNFGADAQGAMNKAL
ELFRKDIAAKYKELGYQG

Is it possible?



3.1 What is protein structure prediction?

Different levels of protein structures

1, Primary amino acid sequences (1D)

MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLE
KFDRVKHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGHHEA
ELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGNFGADA
QGAMNKALELFRKDIAAKYKELGYQG

2, Secondary structure

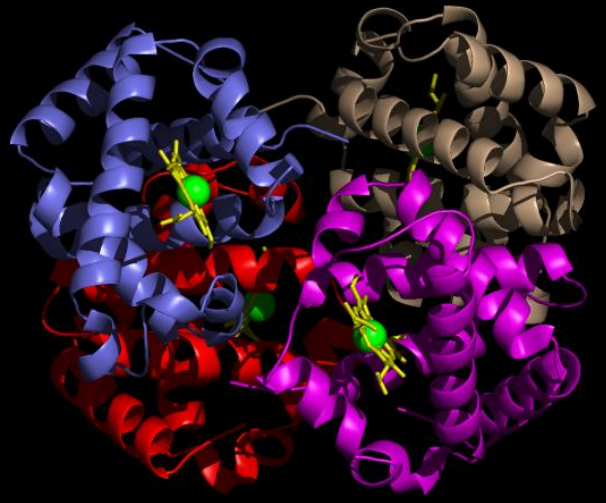
HHHHHHHHHHHHLLLLLEEEEEEEELLLLLLEEEEEEEELLLLEEEEEELLLLHHHHLLLLHHHHHHHH

Different levels of protein structures

3, Tertiary structure

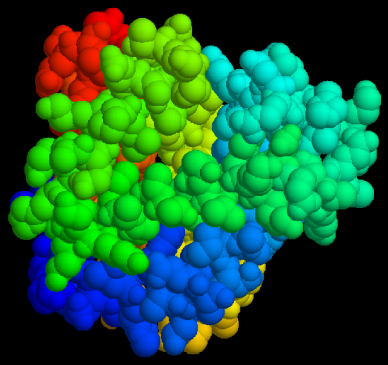
Hemoglobin

4, Quaternary structure

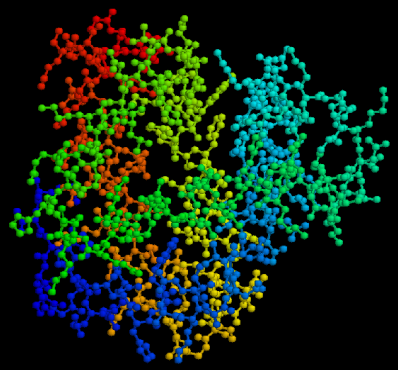


When we talk about protein structure prediction, we usually mean tertiary structure prediction

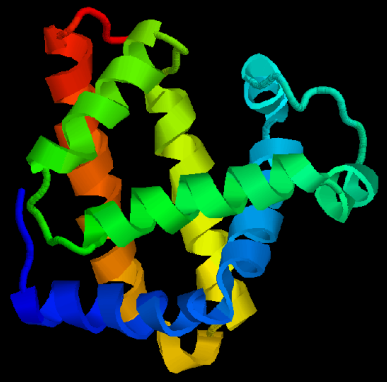
Oxygen transport protein



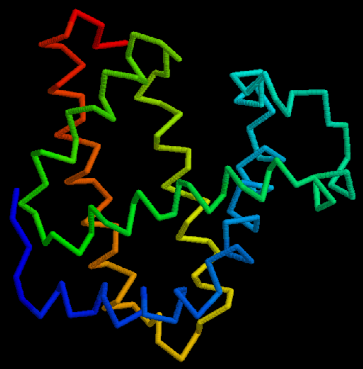
Fullfill



Ball&stick

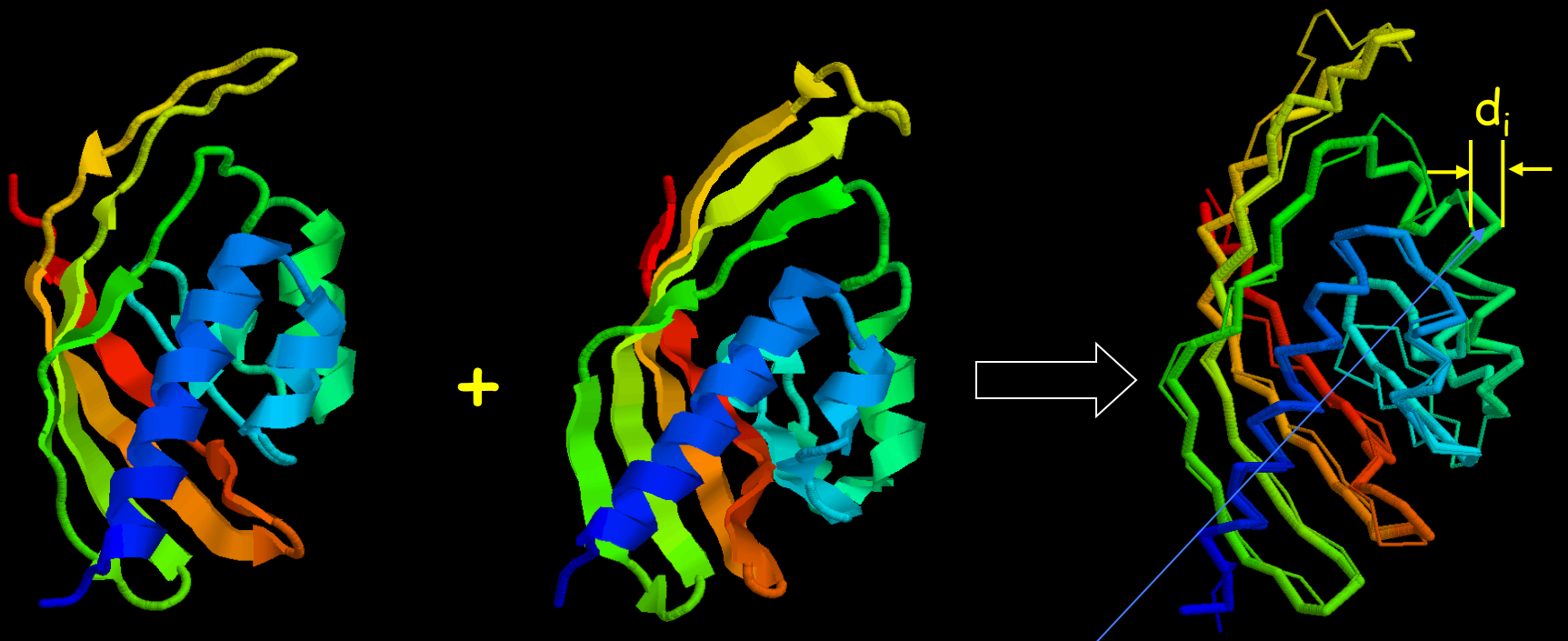


Cartoon



backbone

How to evaluate protein structure prediction?



Computer model

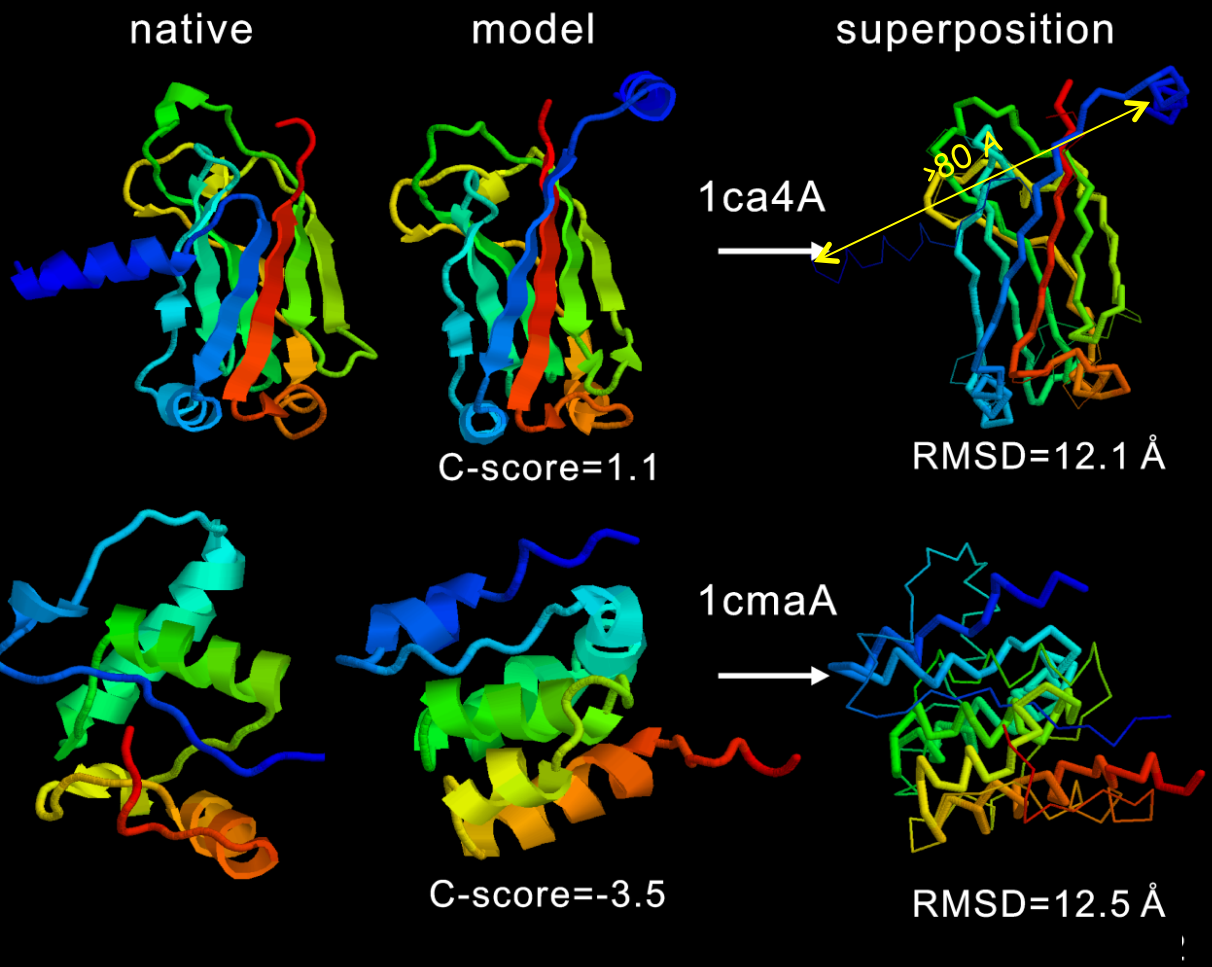
Experimental structure

superposition

$$RMSD = \sqrt{\sum_{i=1}^{N} d_i^2 \Big/ N}$$

Problem of RMSD

$$RMSD = \sqrt{\sum_{i=1}^{N} d_i^2 \Big/ N}$$



A local big difference can result in a high RMSD. This makes it Difficult to distinguish between bad or good predictions

RMSD vs TM-score

RMSD give all distance equal weights

$$RMSD = \sqrt{\sum_{i=1}^{N} d_i^2 \Big/ N}$$

Kabsch, Acta Cryst (1976)

TM-score down-weights larger distance and makes it more sensitive to global topology

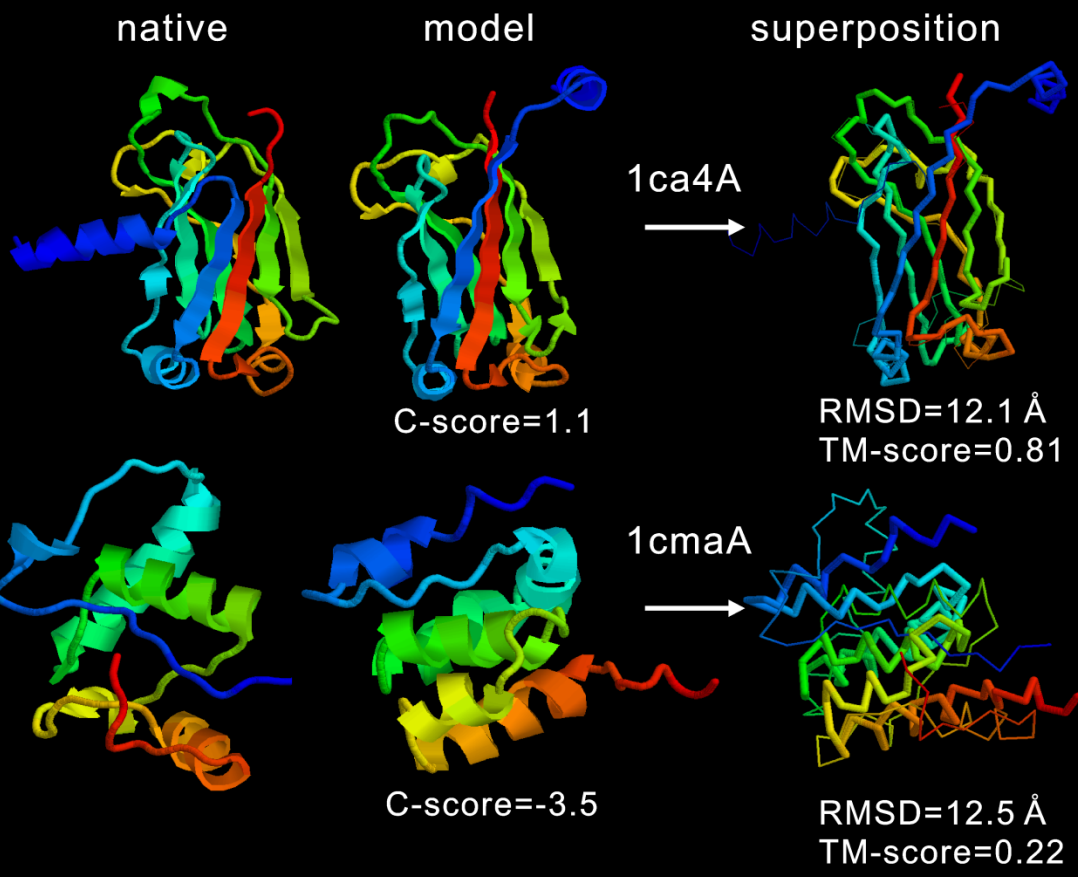
$$TM-score = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\left(d/d_0\right)^2}$$

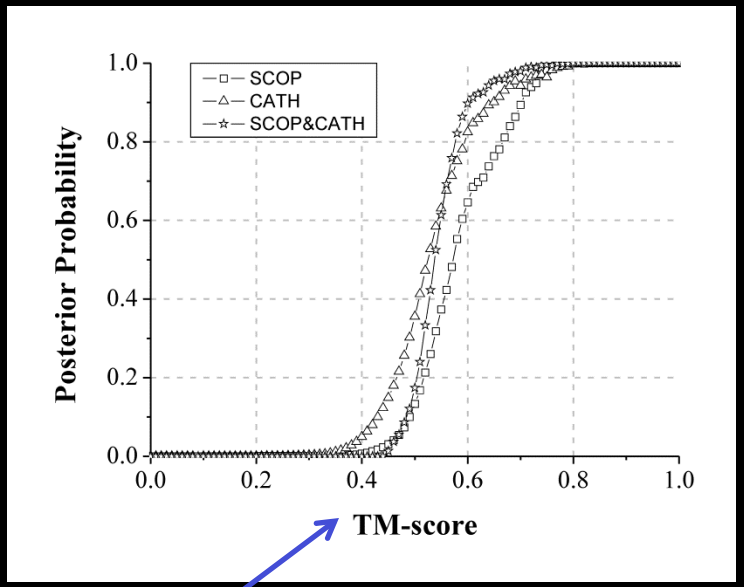
where $d_0 = 1.24\sqrt[3]{N-15} - 1.8$

Zhang and Skolnick, Proteins (2004)

Note: The best RMSD can be calculated analytically by Lagrange multipliers; but the best TM-score can only be calculated by heuristic iteration which is slower.

TM-score is more sensitive to fold than RMSD (since local variation can dramatically influence RMSD values)



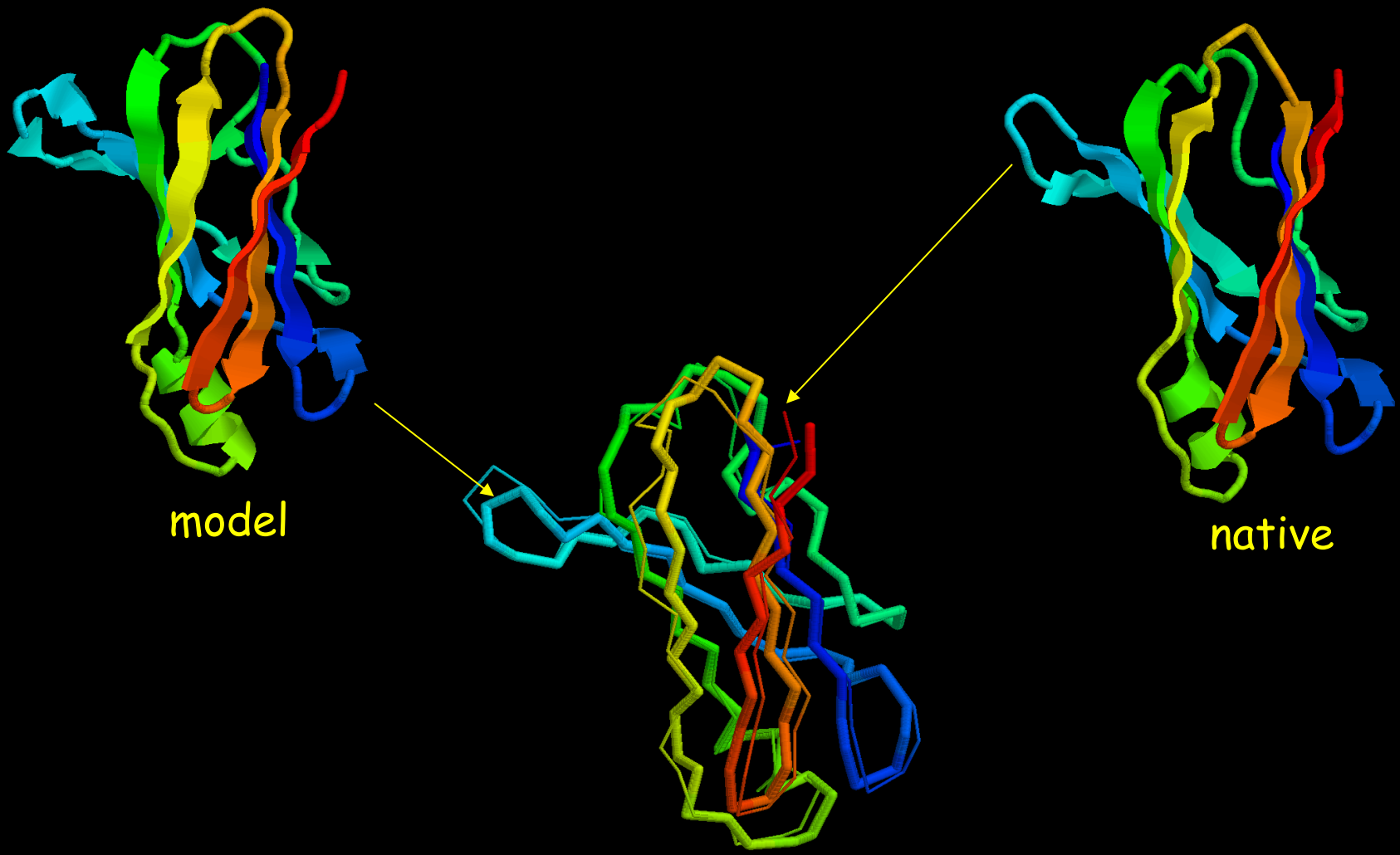


There is a phase transition at TM-score=0.5 for the possibility of the two structures with the same fold.

Xu & Zhang, Bioinformatics 2010

High-resolution prediction

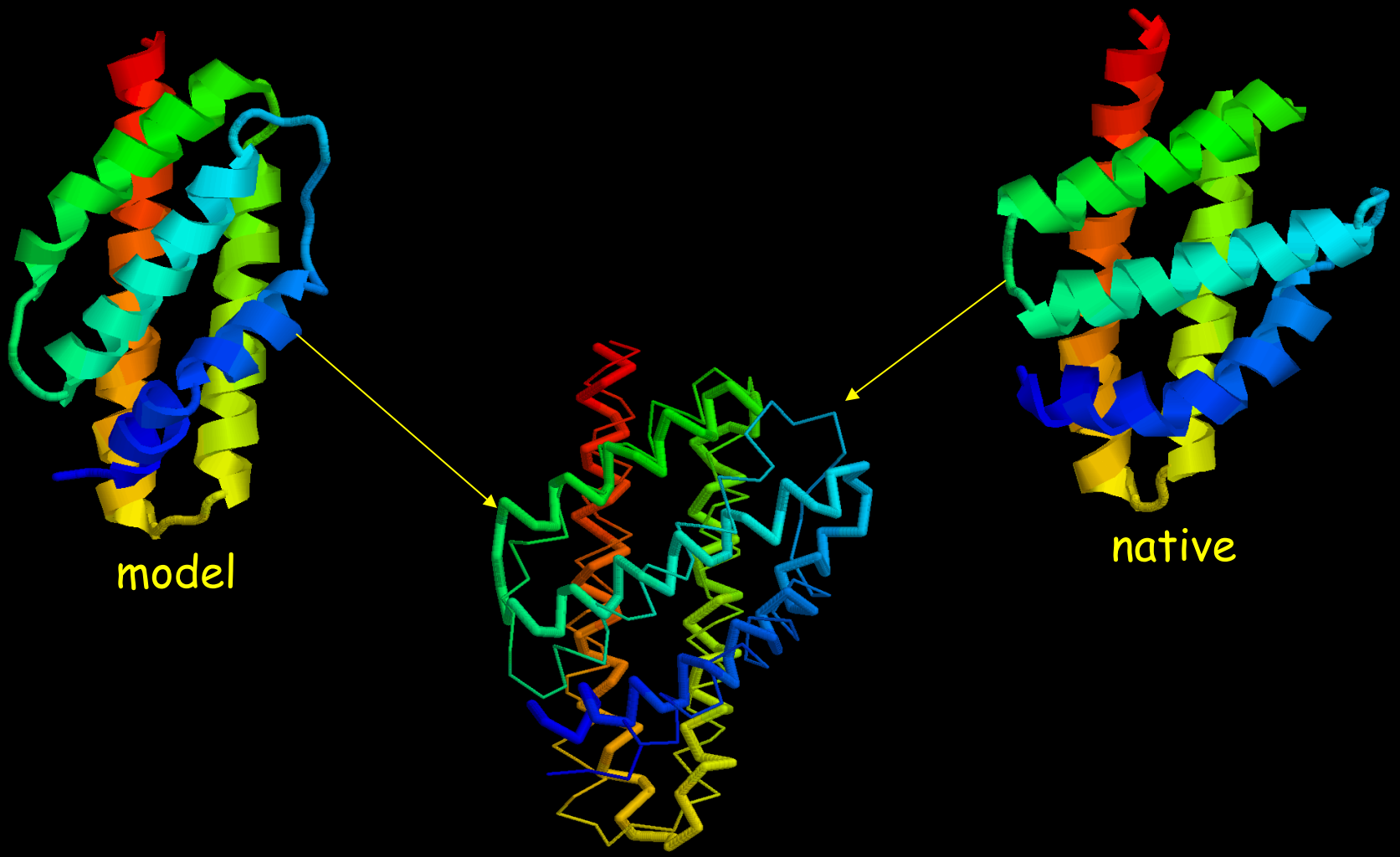
(RMSD<2Å, TM-score>0.8)



Superposition, RMSD=1.8Å, TM-score=0.75

Medium-resolution prediction

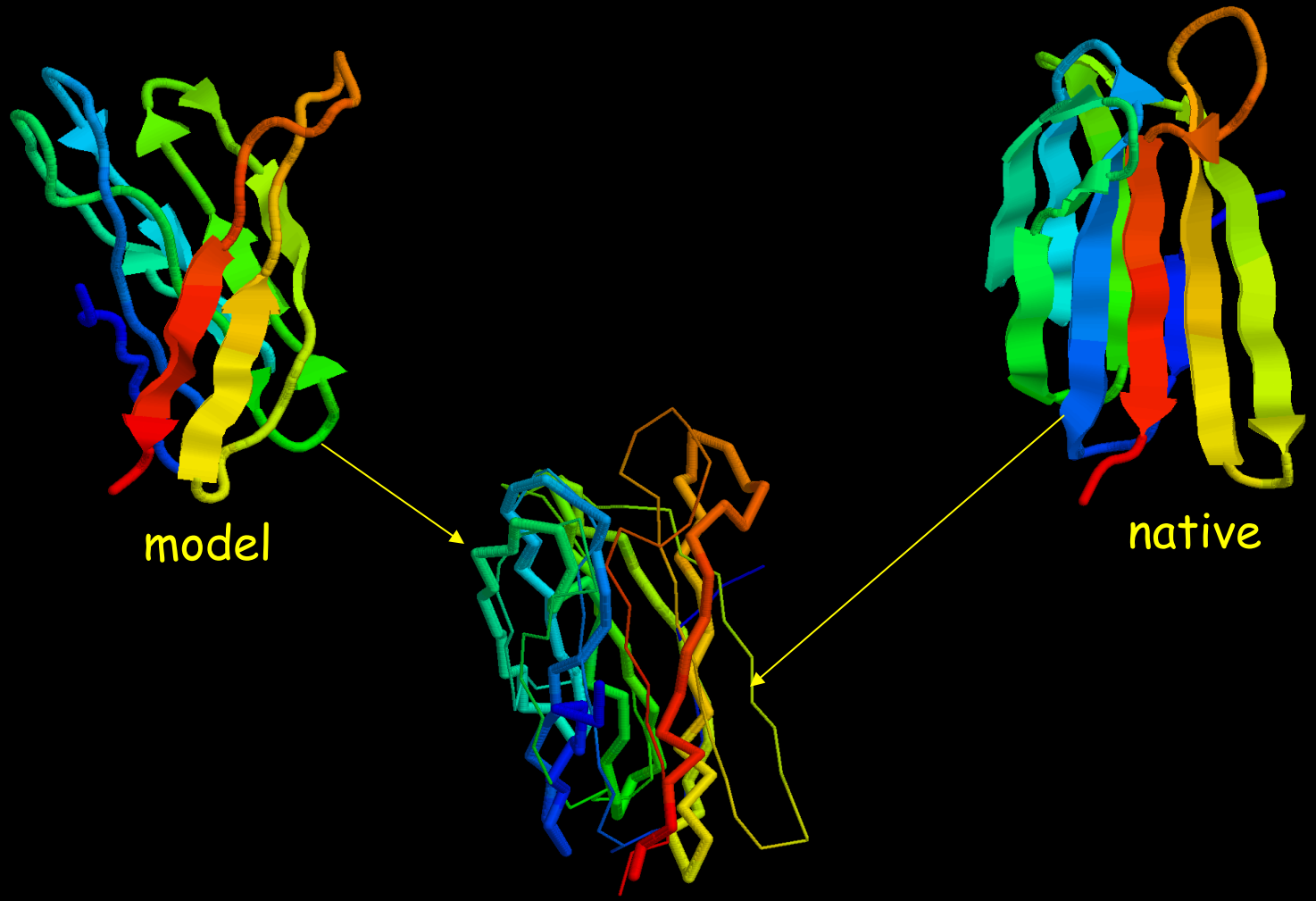
(RMSD~[2Å,6Å], TM-score~[0.4,0.8])



Superposition, RMSD=4.1Å, TM-score=0.5

Low-resolution or wrong prediction

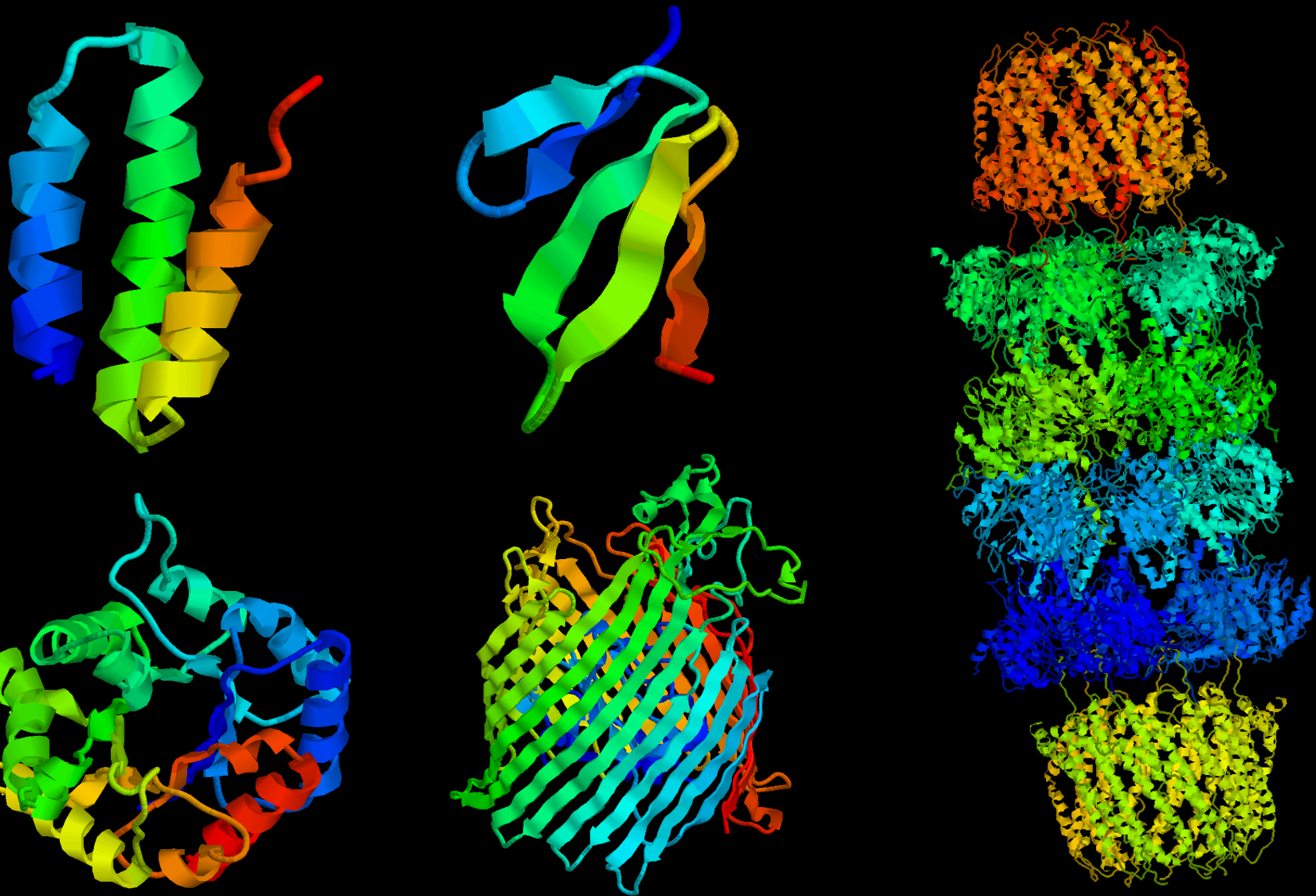
(RMSD>6Å, TM-score<0.4)



Superposition, RMSD=7.1Å, TM-score=0.3

Structure prediction from sequence is not an easy problem

MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRVKHLKTEAEMKASEDLKKHGVTVLTALGAILKKK
GHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGNFGADAQGAMNKALELFRKDIAAKYKELGYQG



Question: What will YOU do to predict protein structure when given an amino acid sequence?

Different protein should use different strategies

There are three methods which can be used to treat for different protein targets

- Ab initio folding
- Comparative modeling (CM)
- Threading

Reference:

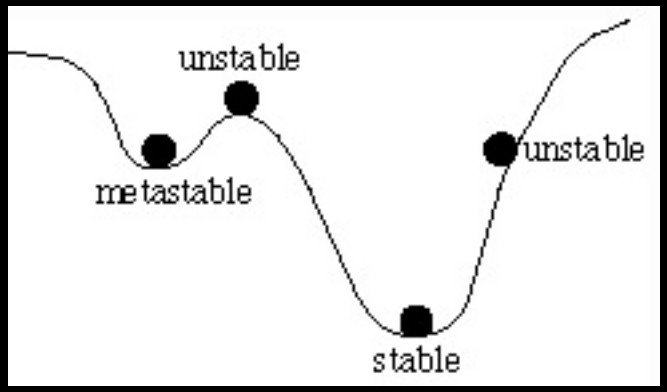
Y Zhang. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18: 342-348 (2008).

Table of Contents

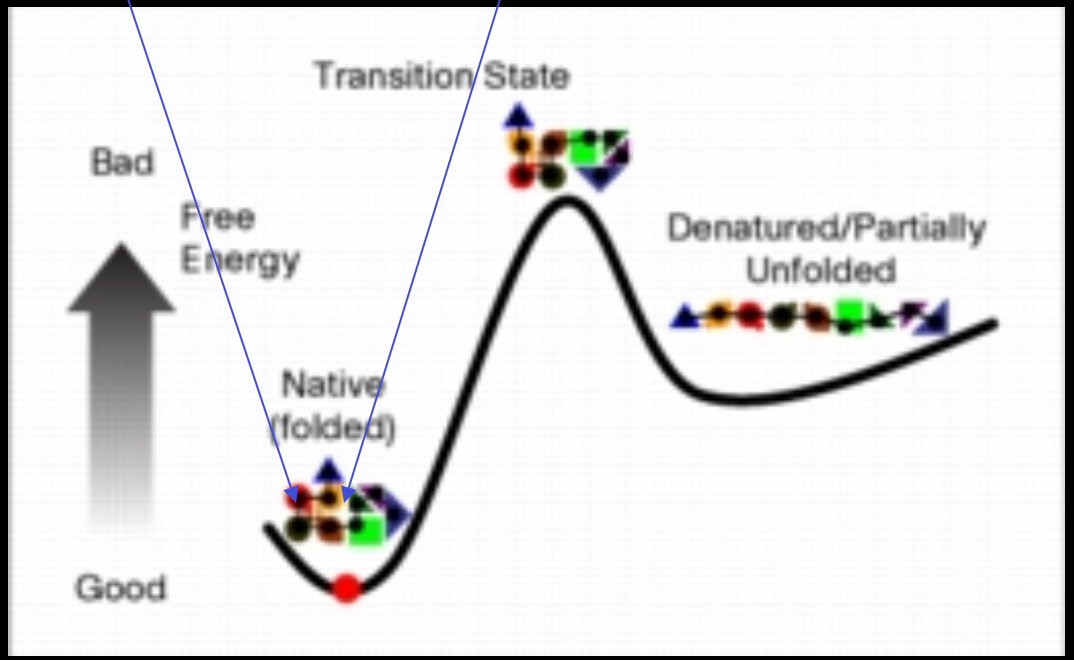
3.2 Method I: Ab initio folding

Definition:

Ab initio folding – decide protein structure based on physical principle, i.e. find native structure by searching for the lowest free energy.



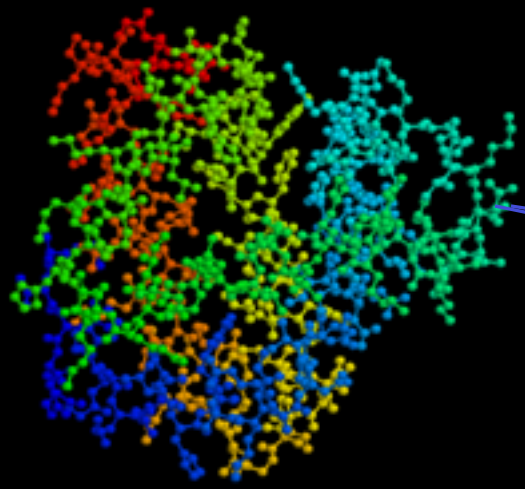
Rule of physics



Two factors that matter: force field, search engine

CHARMM (Martin Karplus):

An example of *ab intio* protein folding



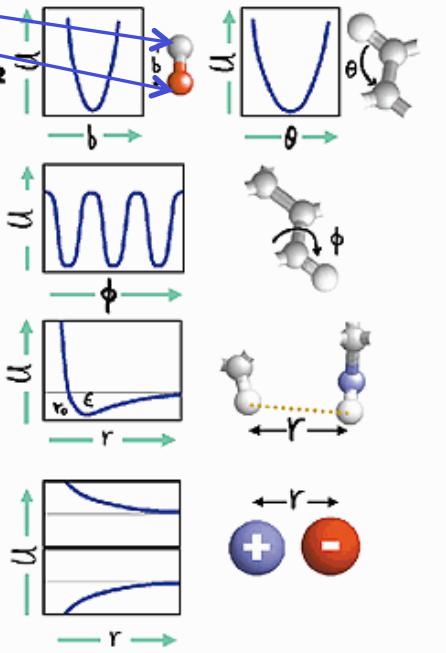
Target protein

$$U = \sum_{\text{All Bonds}} \tfrac{1}{2}K_b(b-b_o)^2 + \sum_{\text{All Angles}} \tfrac{1}{2}K_\theta(\theta-\theta_o)^2$$

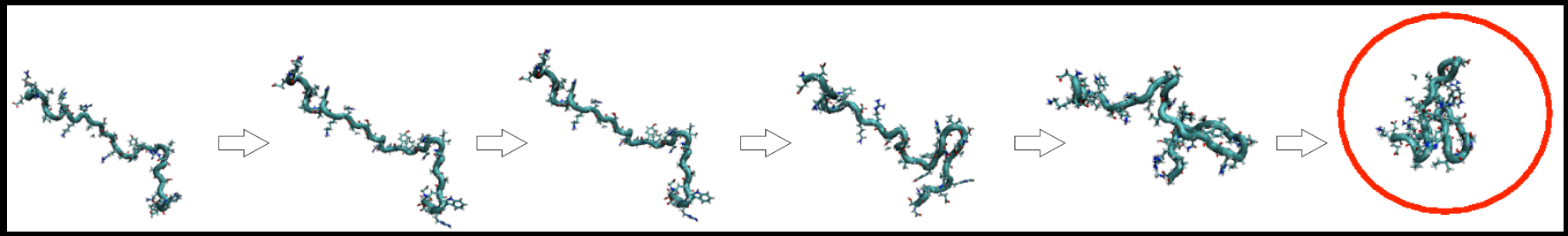
$$+ \sum_{\text{All Torsion Angles}} K_\phi[1-\cos(n\phi+\delta)]$$

$$+ \sum_{\text{All nonbonded pairs}} \varepsilon\left[\left(r_o/r\right)^{12} - 2\left(r_o/r\right)^{6}\right]$$

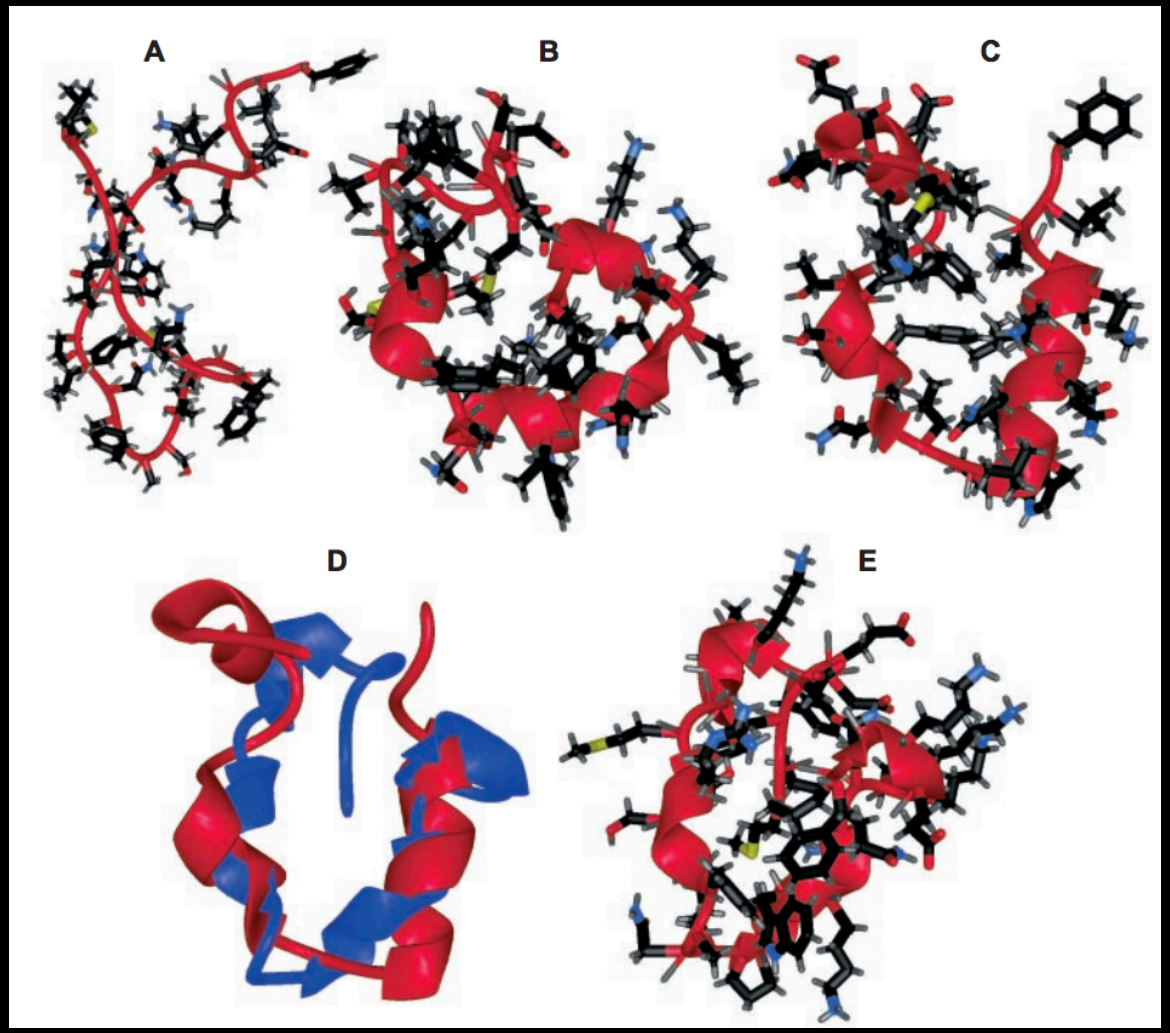
$$+ \sum_{\text{All partial charges}} 332 q_i q_j / r$$



Fold a protein by Newton's law of motion: F=ma



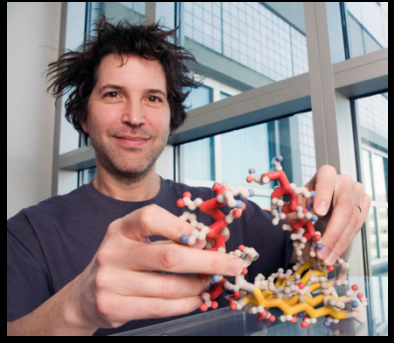
AMBER – the first success in folding a small protein using MD (Peter Kollman)

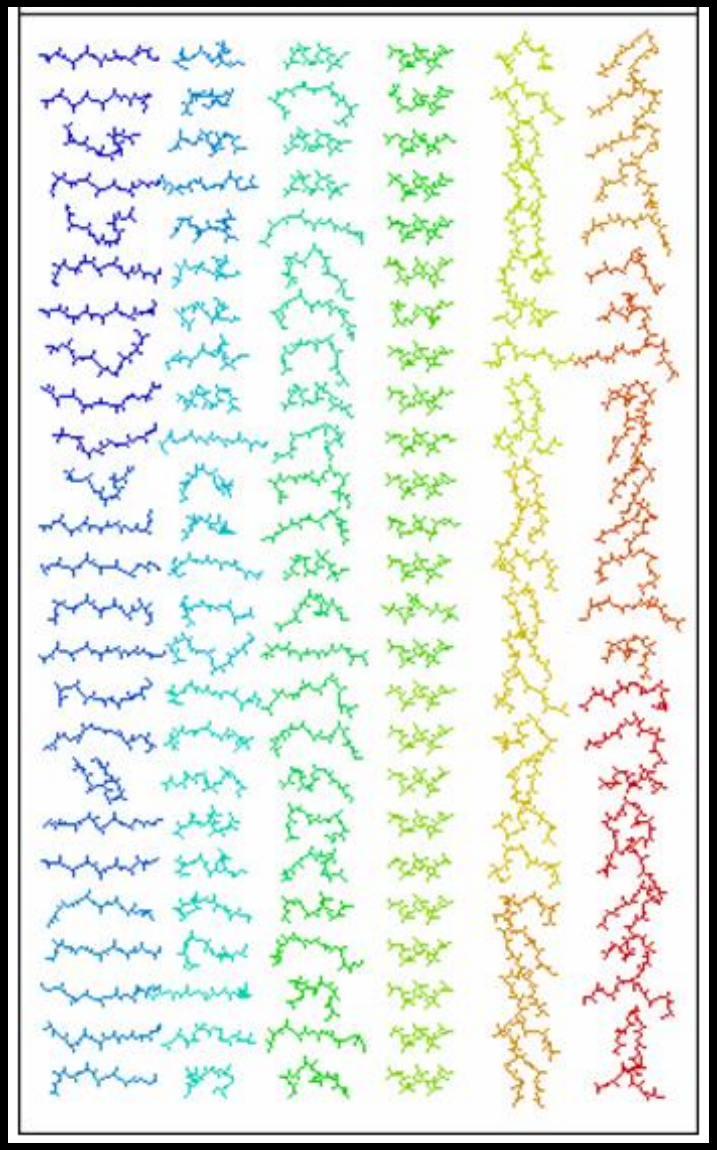


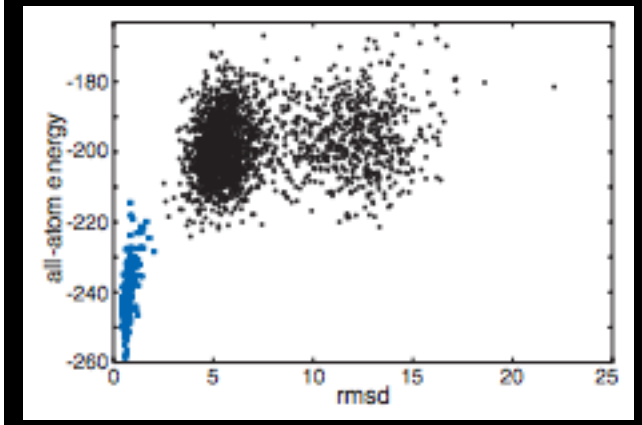
- 36-mer
- 4.5 Å
- Parallel computing
- Two months
- 150 nanoseconds

Duan, Kollman, (1998), Science, 282, 740-744

Rosetta – fold protein structure by fragment assembly







Simulated annealing
Monte Carlo simulation

Baker et al JMB 1997

QUARK: Ab initio protein structure prediction using continuous fragments

Dong Xu

MPRGRSTGRKGRDTEKERSRSPLRAPGGSDGPST
RAGCGAGPCQLSSPIKRGGVRGQPAKCHGKYTTT
AEGLTALLNRRHSPRTSNEGRWMNGVMAVNLSKW
DTEKERSRSPLRAPGGSDGPSTRAGCGAGPCQLS
SPIKRGGVRGQPAKCHGKYTTTAEGLTAL

Query sequence

Gapless threading through PDB

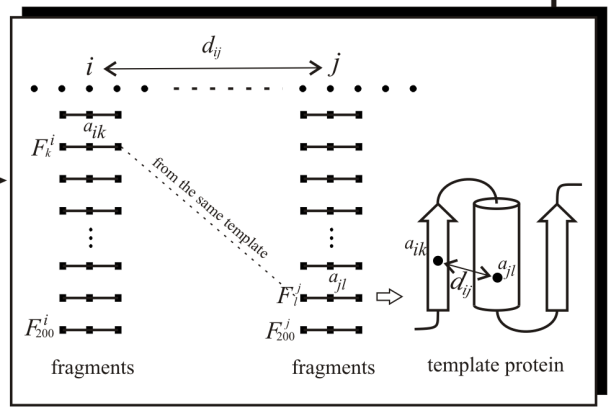
pos=1 ··· n/4 ··· n/2 ··· 3n/4 ··· n

L=1 L=5 L=10 L=15 L=20

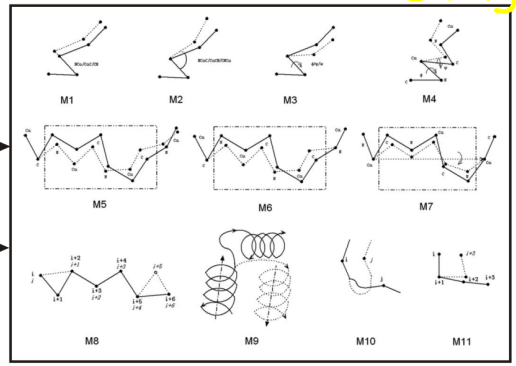
200 fragments with 1-20 AA created for each position

Physics- and knowledge-based potentials:

- hydrogen-bonding
- van der Waals
- solvation
- Coulomb
- backbone-torsion
- bond-length
- bond-angle
- atomic distance
- strand packing

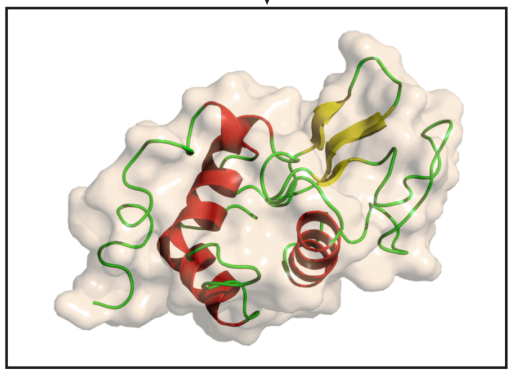


Distance profile restraints



Replica-Exchange Monte Carlo Simulations

Clustering by SPICKER



Final model

Xu and Zhang. Proteins (2012)

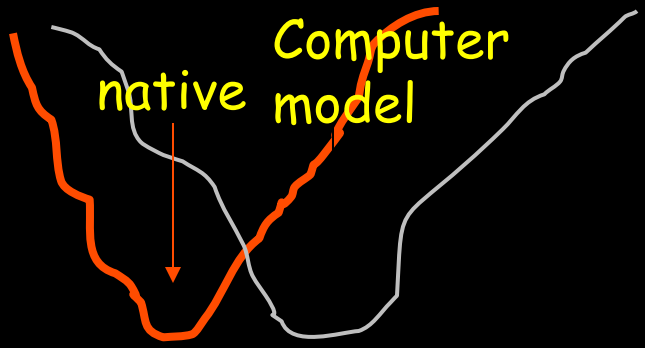
Summary of ab initio modeling

Where we are?

1. We can only fold very small proteins (<100 residues)
2. two much time consuming (up to 1000 CPU years)
3. Error is big (5-20 Angstroms)

What is the challenge?

1. force field is not accurate



2. Search engine can not find the lowest free energy state

There are many local minima

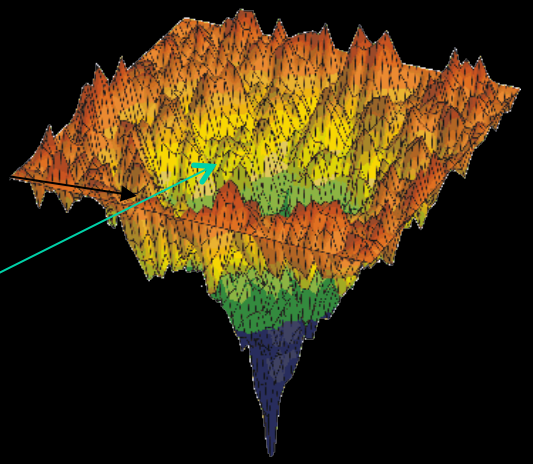
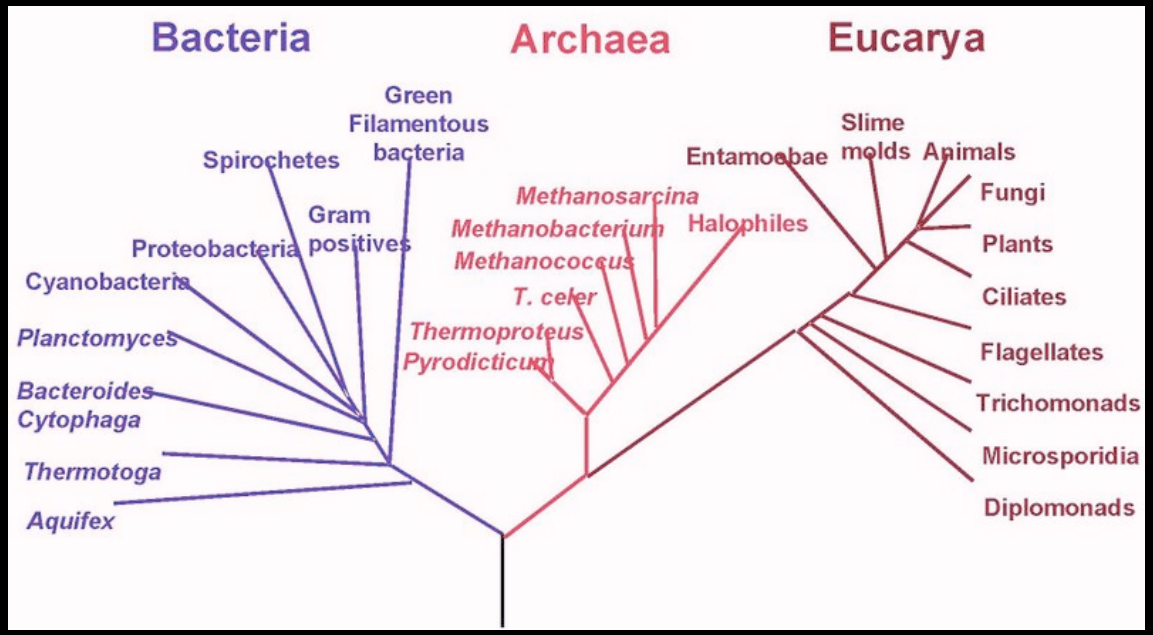


Table of Contents

2.2 Method II: Homologous modeling

Definition:

Homologous modeling is to derive the structure of unknown protein based on the solved protein structures. The equivalency between two proteins is established by sequence homology comparison.

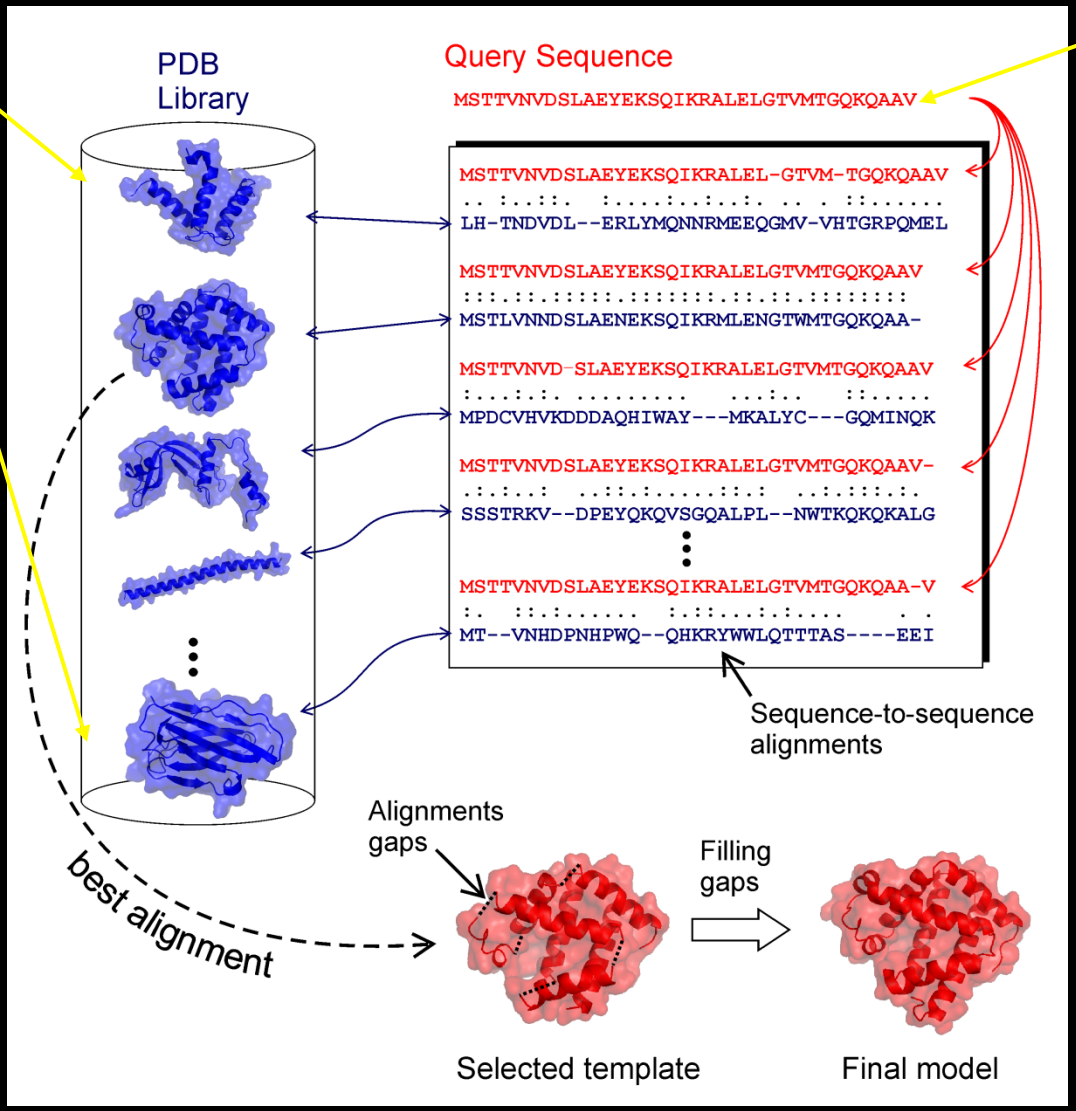


The idea: similar sequences take similar structures due to evolution

Method II: Homologous modeling

Solved

Unknown



Procedure of HM

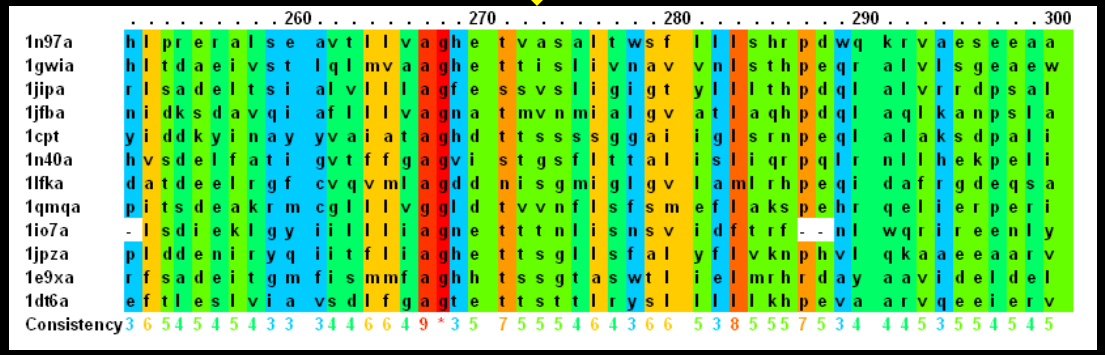
1. *Find template*
2. *Align target sequence with template*
3. *Generate model:*
   - *add loops*
   - *add sidechains*
4. *Refine model*

Routine method is sequence-sequence alignment based on mutation scores.

PSI-BLAST: profile-sequence alignment search

Query sequence: MAPGPEIFKEQSVPFKMRIRGTVNGKKVTITGQGSGDARTGKMRGKWMSYAAYRV

Sequence database:



Template library

Template structure

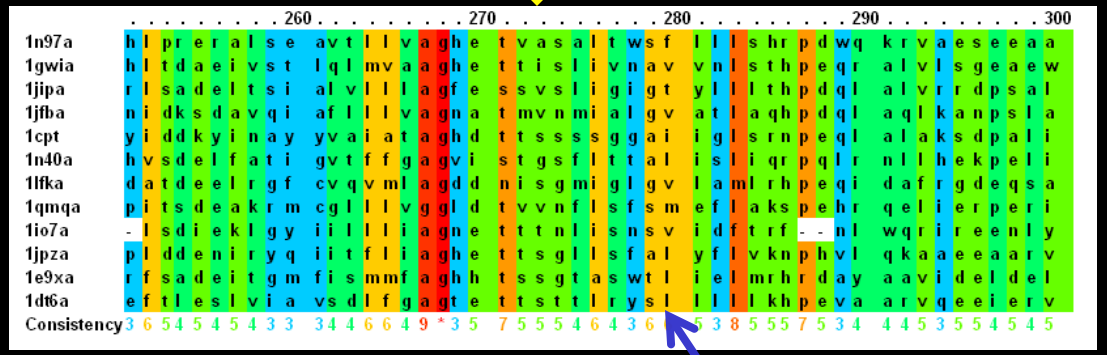
S Altschul et al.
Nucl. Acids Res., 25 (17), 3389-3402

Profile-profile alignment

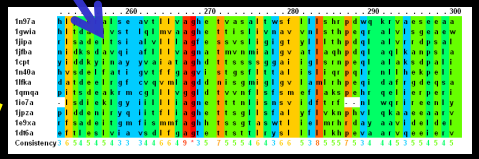
Query sequence: MAPGPEIFKEQSVPFKMRIRGTVNGKKVTITGQGSGDARTGKMRGKWMSYAAYRV

Sequence database:

Sequence profile:



Template library



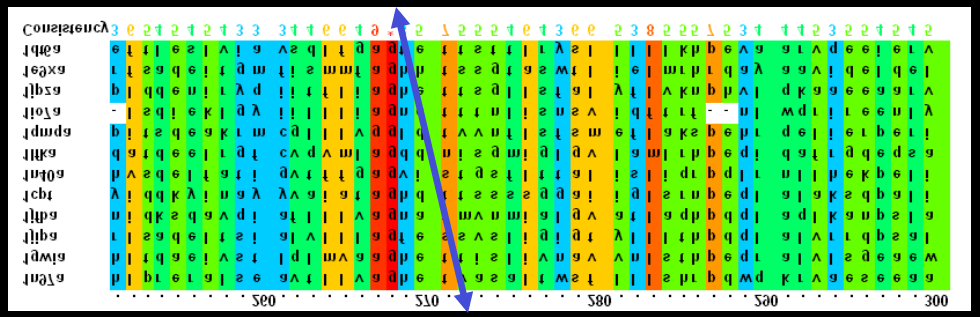
Template structure

Profile-profile alignment: A better way to generate more accurate alignments

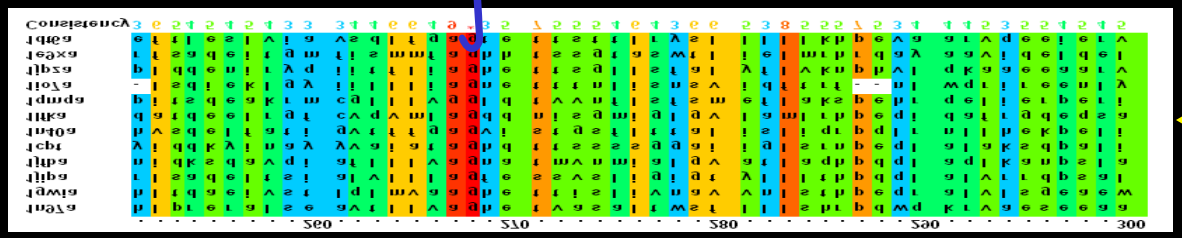
Profile-profile

Secondary structure

$$Score(i,j) = \sum_{m=1}^{20}\sum_{n=1}^{20} F(i,m)F(j,n)B(A_m,A_n) + \delta(ss_i, ss_j)$$

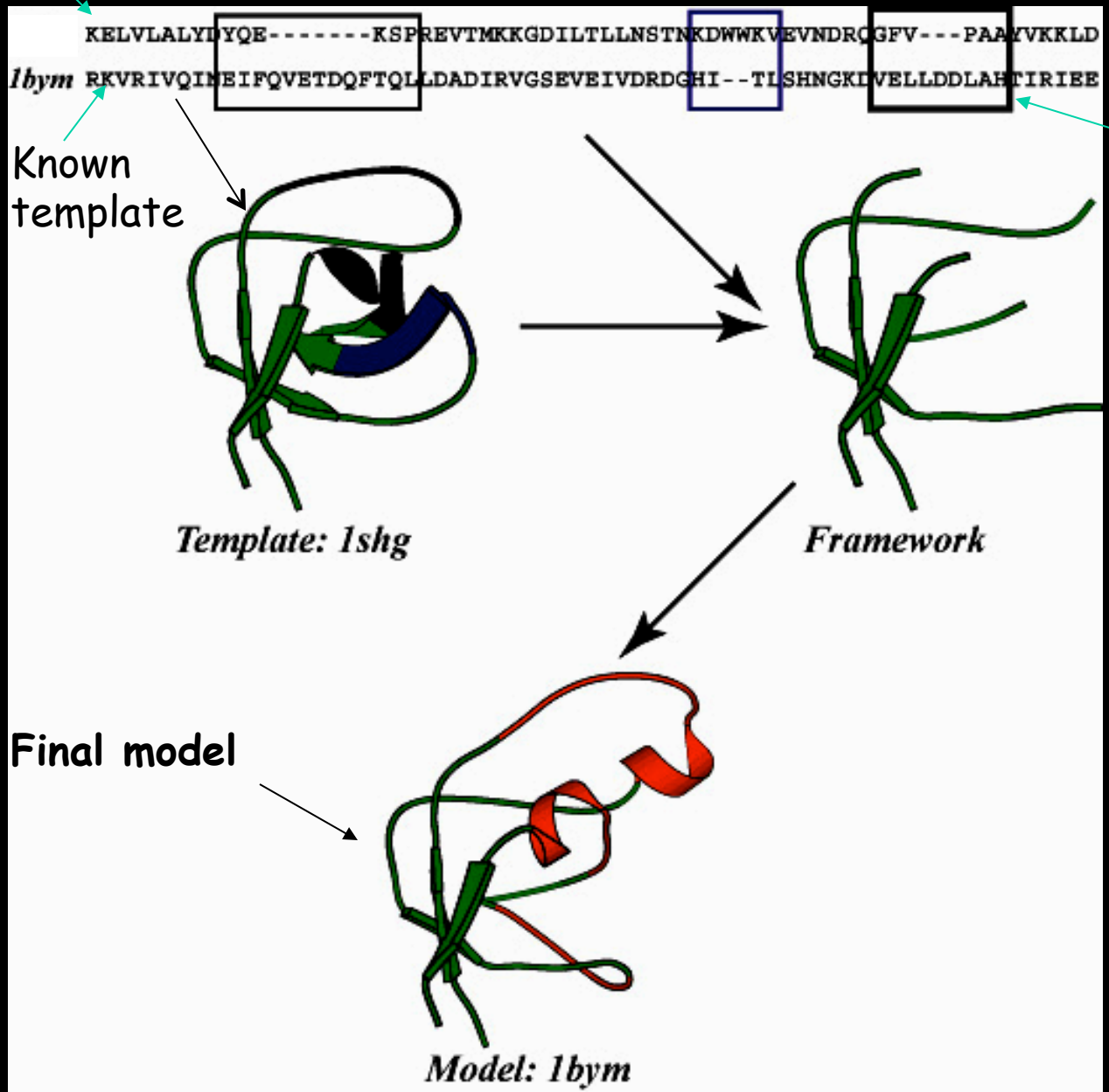


Multiple sequence alignment of target



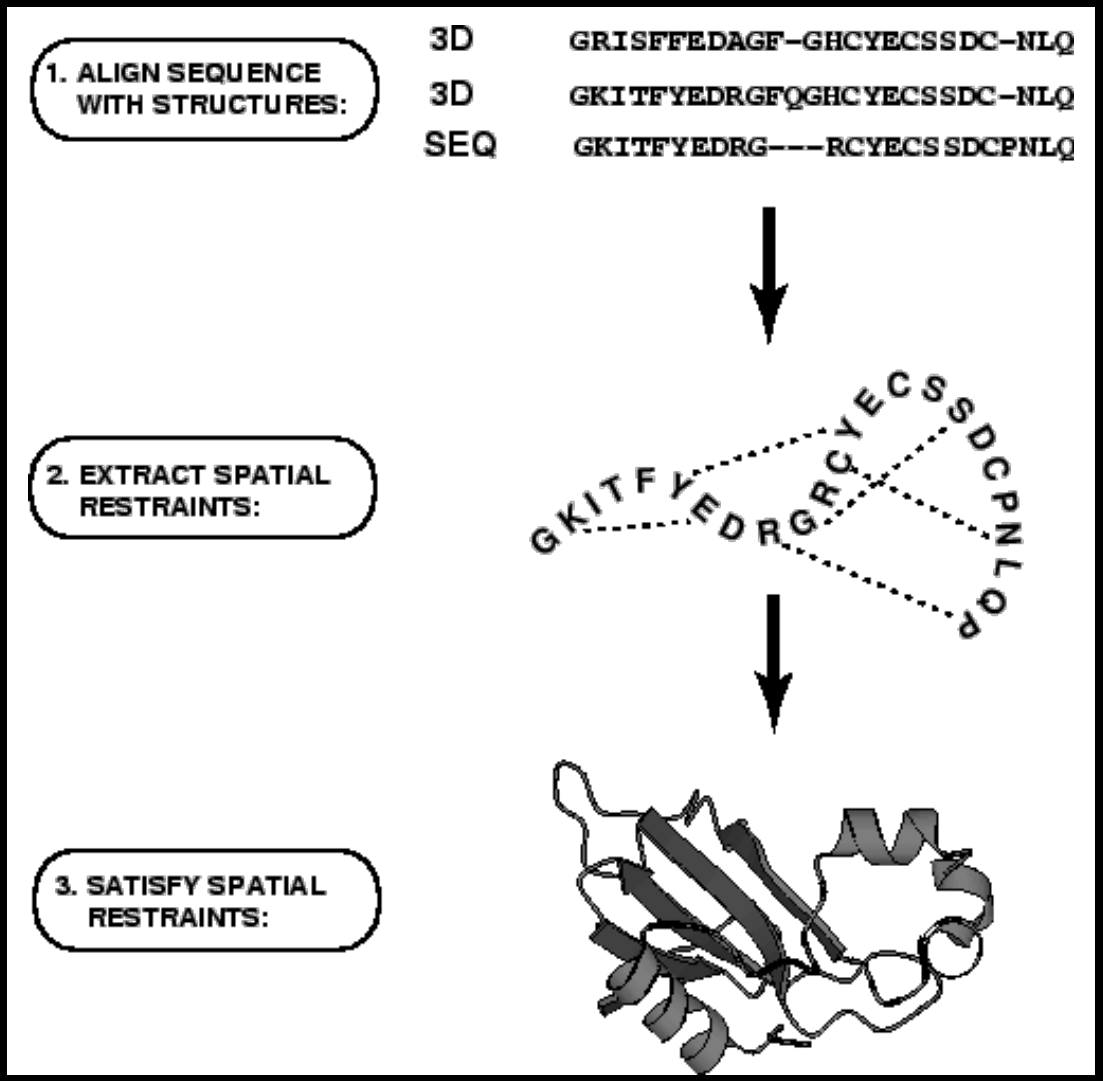
MSA of template

A real example of homologous modeling



Note: you can only copy the structure of aligned regions. So the key of modeling accuracy is the alignment accuracy.

MODELLER: Homologous modeling tool



A Sali, TL Blundell (1993) J. Mol. Biol. 234, 779-815.

Summary of homologous modeling

Merit:

When an appropriate template is identified, the accuracy can be very high. The error can be 1~2 Angstroms!

Demerit:

It can not predict structure if there is no homologous proteins in PDB

Key point:

How to correctly align two sequences is the key for the quality of final models.

Table of Contents

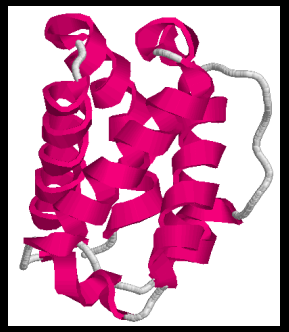
2.3 Method III: Fold recognition

Fact: Many proteins have different but with the same fold

LHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRVKHLKTEAEMKASEDLKK
HGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSR
HPGNFGADAQGAMNKALELFRKDIAAKYKELGYQG

Only 10% of residues
are identical

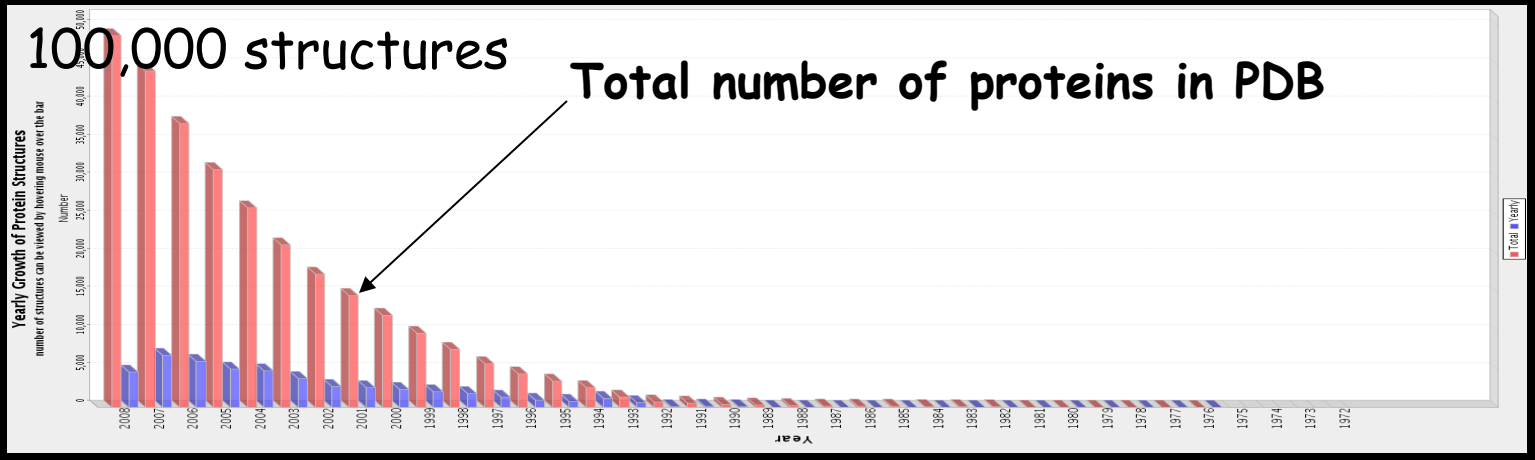
TEGYYTIGIGHLLTKSPSLNAAAKSELDKAIGRNTNGVITKDEAEKLFNQDVDA
AVRGILRNAKLKPVYDSLDAVRRAALINMVFQMGETGVAGFTNSLRMLQQKRW
DEAAVNLAKSRWYNQTPNRAKRVITTFRTGTWDAYK

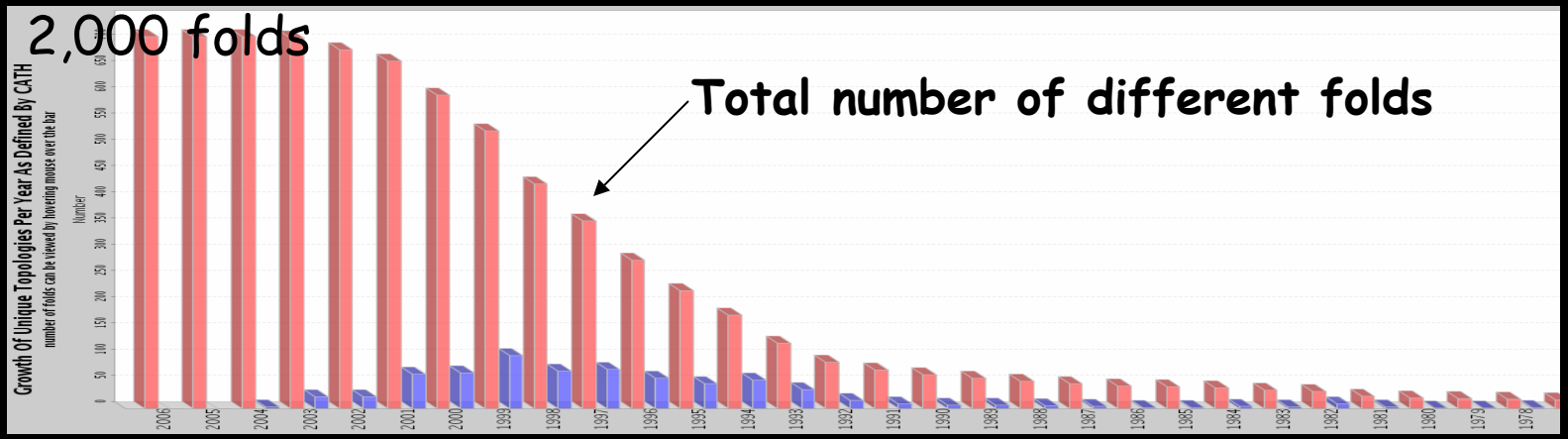


Problem of homologous modeling: Structure templates cannot be detected based on sequence alignment, when sequence identity is low

Summary of PDB Library

Fact: Number of folds in nature is limited





100,000 proteins in PDB have only 2,000 different structures.
Many proteins of different sequence but with similar structure

Method III: Fold recognition (threading)

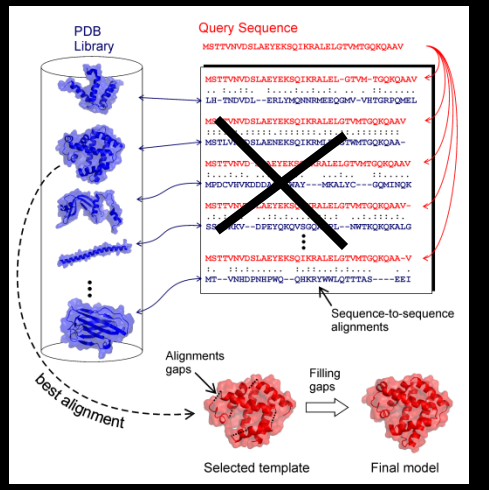
Definition:

Fold recognition (also called threading) is to identify the template proteins which have similar structure to the query but may have different sequences.

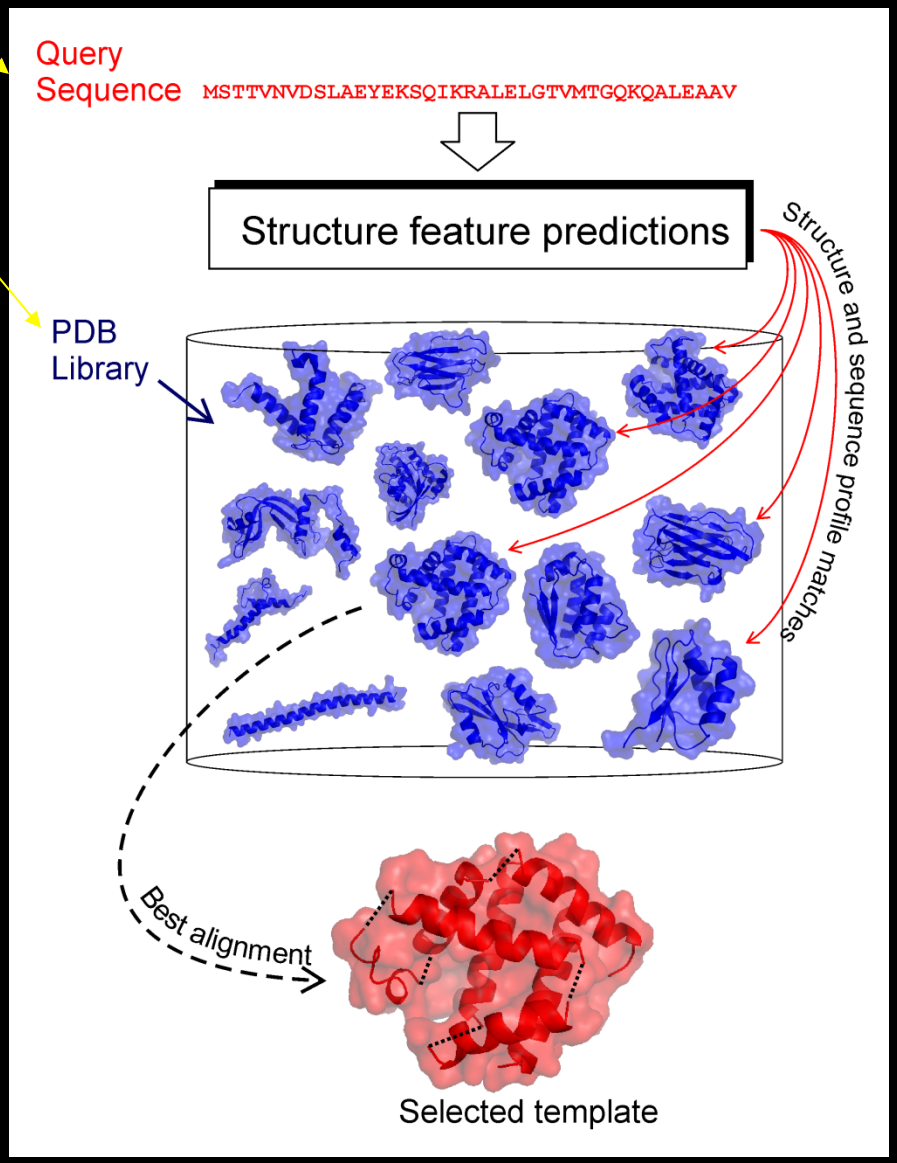
Method III: Fold recognition (threading)

unknown

solved



CM



The idea is to directly align sequence on the 3D structure

MUSTER: thread sequence through PDB using multiple structure information

Profile & SS

$$\mathrm{Score}(i,j) = \sum_{k=1}^{20} (Pc_q(i,k) + Pd_q(i,k)L_t(j,k)/2 + c_1\delta(s_q(i), s_t(j))$$

Structure profile

$$+ c_2 \sum_{k=1}^{20} Ps_t(j,k)L_q(i,k) + c_3(1 - 2|SA_q(i) - SA_t(j)|)$$

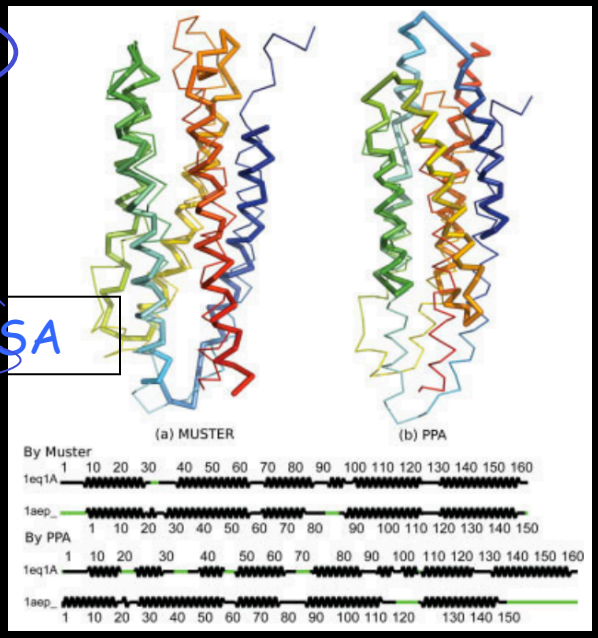
SA

$$+ c_4(1 - 2|\varphi_q(i) - \varphi_t(j)|) + c_5(1 - 2|\phi_q(i) - \phi_t(j)|)$$

Backbone torsion

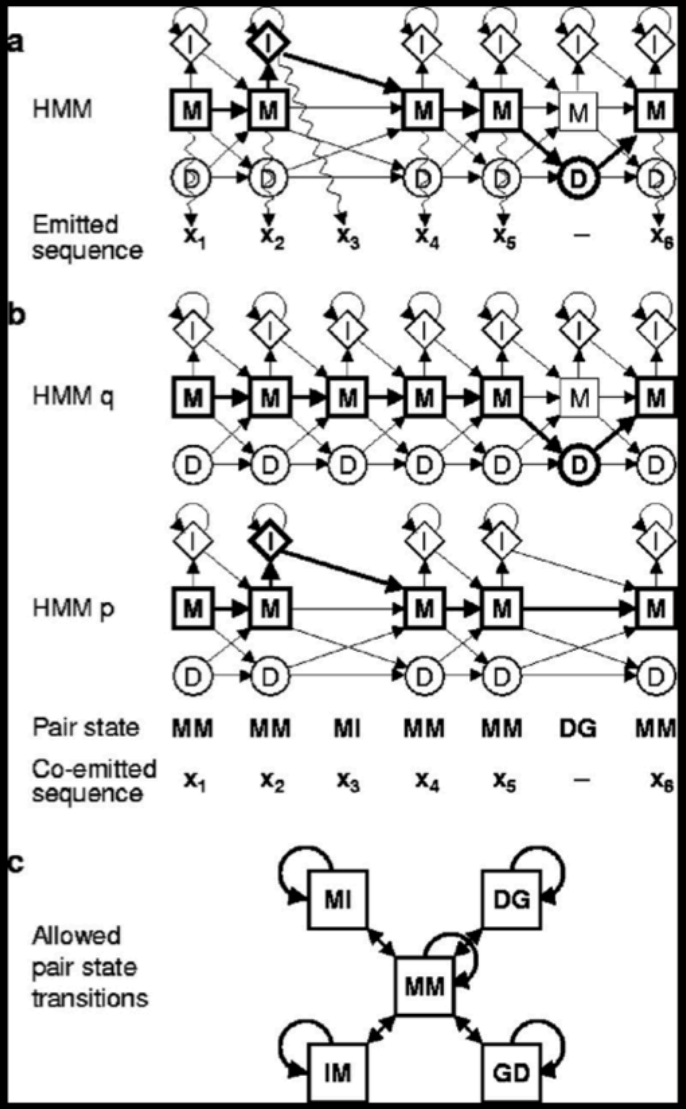
$$+ c_6 M(AA_q(i), AA_t(j)) + c_7$$

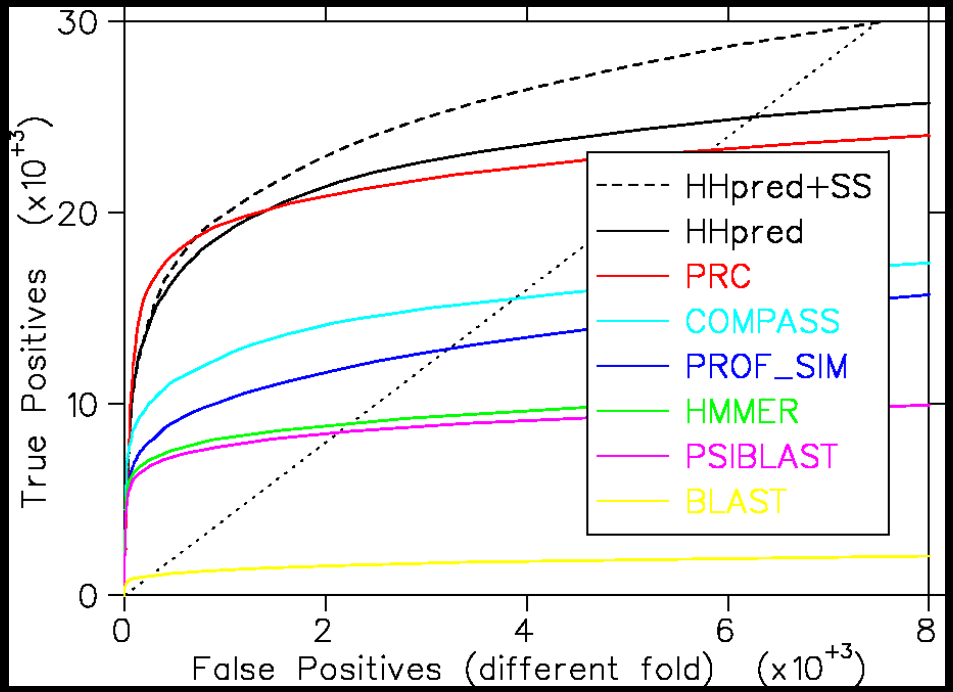
Hydrophobicity



Wu & Zhang, Proteins (2008)

HHsearch: detecting template by hidden Markov alignment





J Soding (2005) Bioinformatics 21, 951-960

Summary of Threading

Merit:

When an appropriate template is identified, the accuracy can be very high. The error can be 2~4 Angstroms!

Demerit:

It can not predict structure if there is no similar structure solved in PDB

Key point:

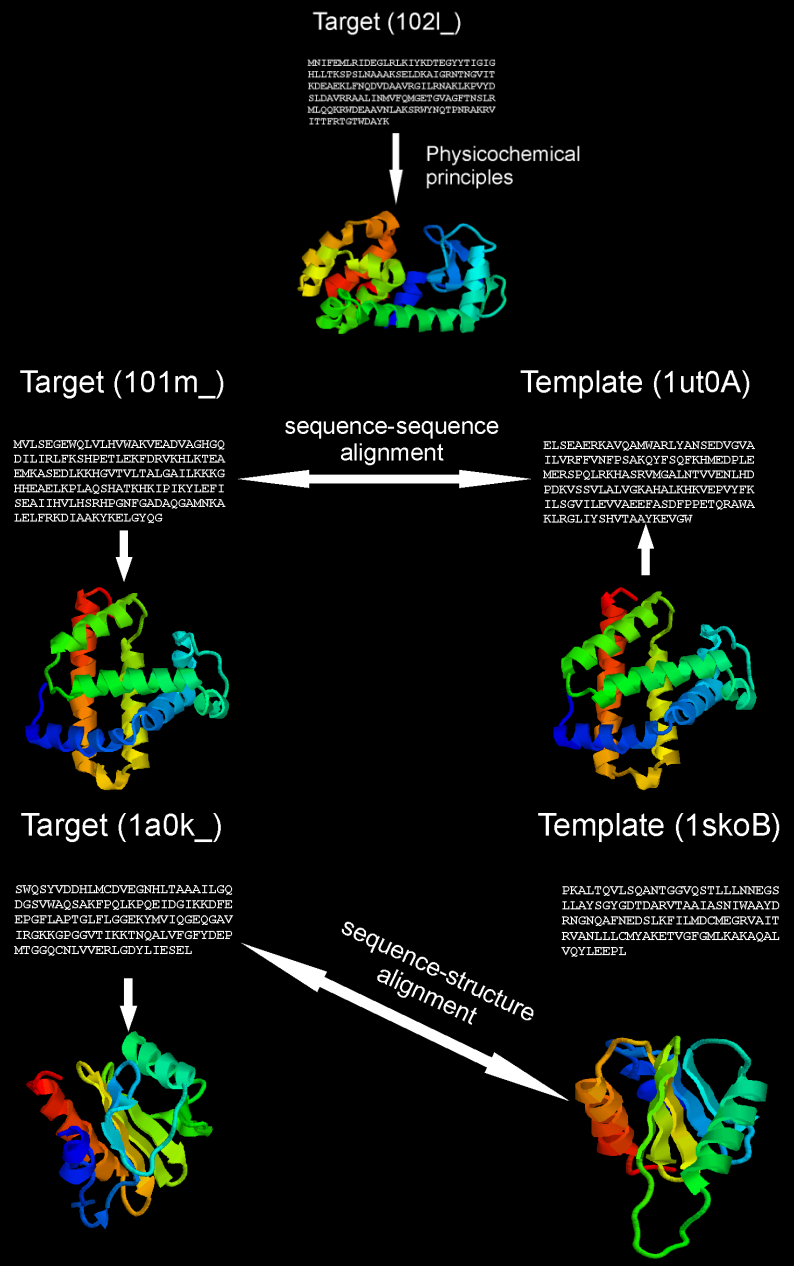
Good template and good alignment is the key

Summary of all three methods

- ***Ab initio* modeling**
  (only for small proteins)
  (less reliable, <6 *Å is good*)

- **Comparative modeling**
  (*homology* >~35%)
  (RMSD~1-2 Å)

- **Threading**
  (needs templates)
  (RMSD>2Å)



Summary of all three methods

Same sequence same structure

Different sequence same structure

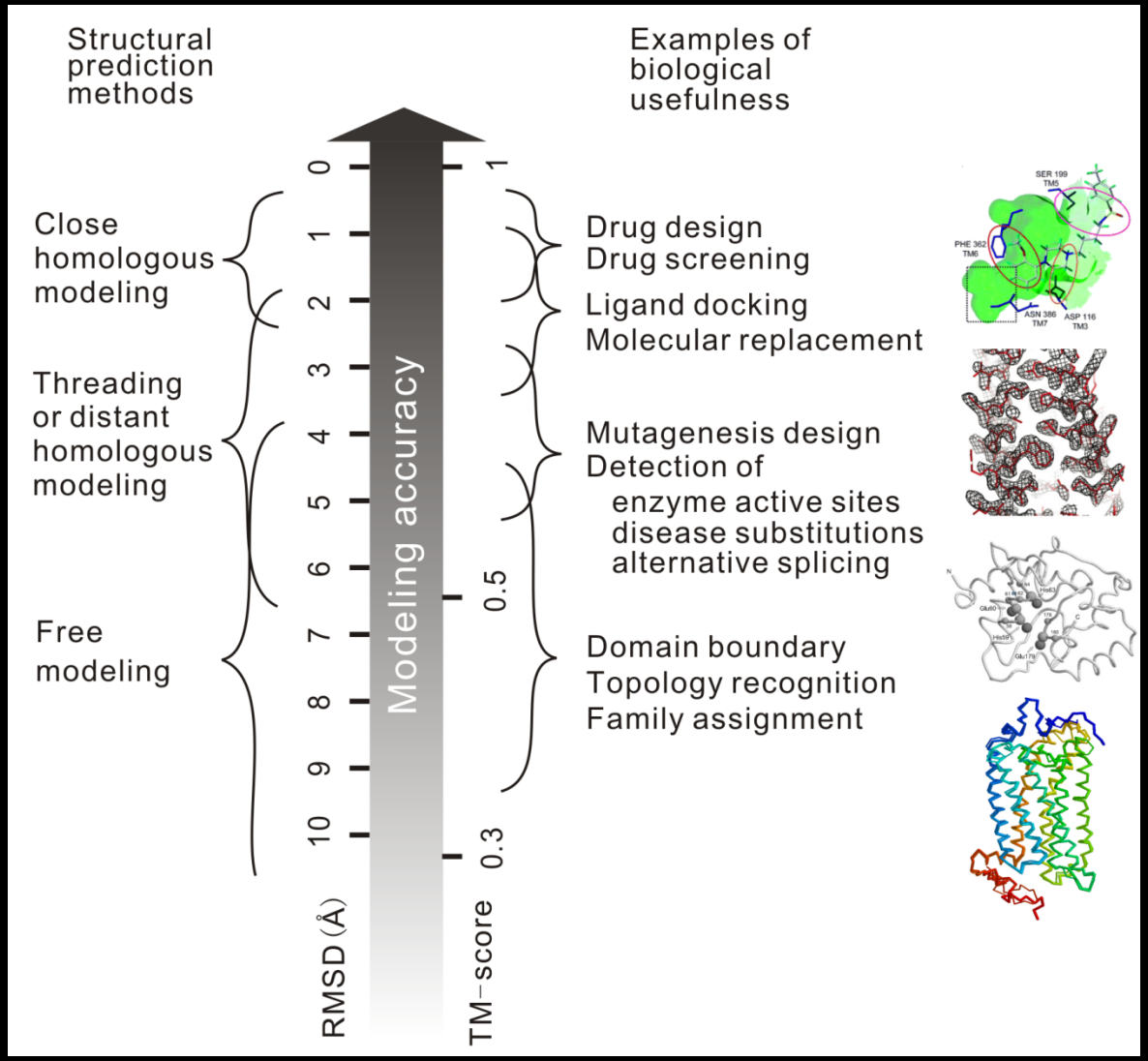
	Methods based on	Homologous template	Analog template	Protein length	Accuracy
Ab initio folding	Physical principles	No need	No need	Small proteins	>3-5Å (low)
Homologous modeling (comparative modeling)	Evolutionary relationship	Need	NA	Any size	>1Å (high)
Fold recognition (threading)	Sequence-structure comparison	No need	Need	Any size	>2Å (high)

Summary: Methods vs Accuracy

Experimental resolution

Alignment is key error on loop

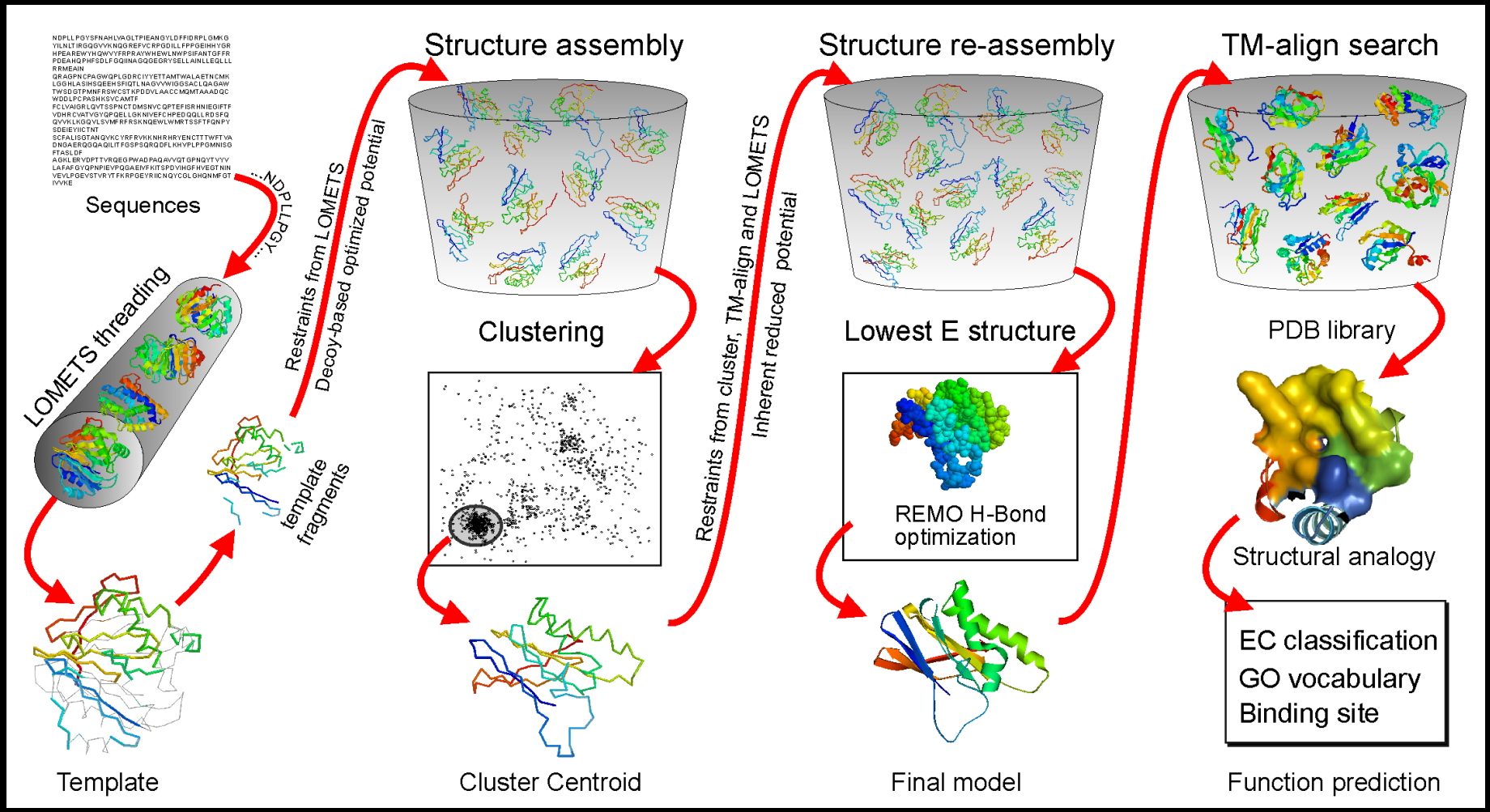
Low resolution for small proteins



Y Zhang Curr Opin Str Biol (2009)

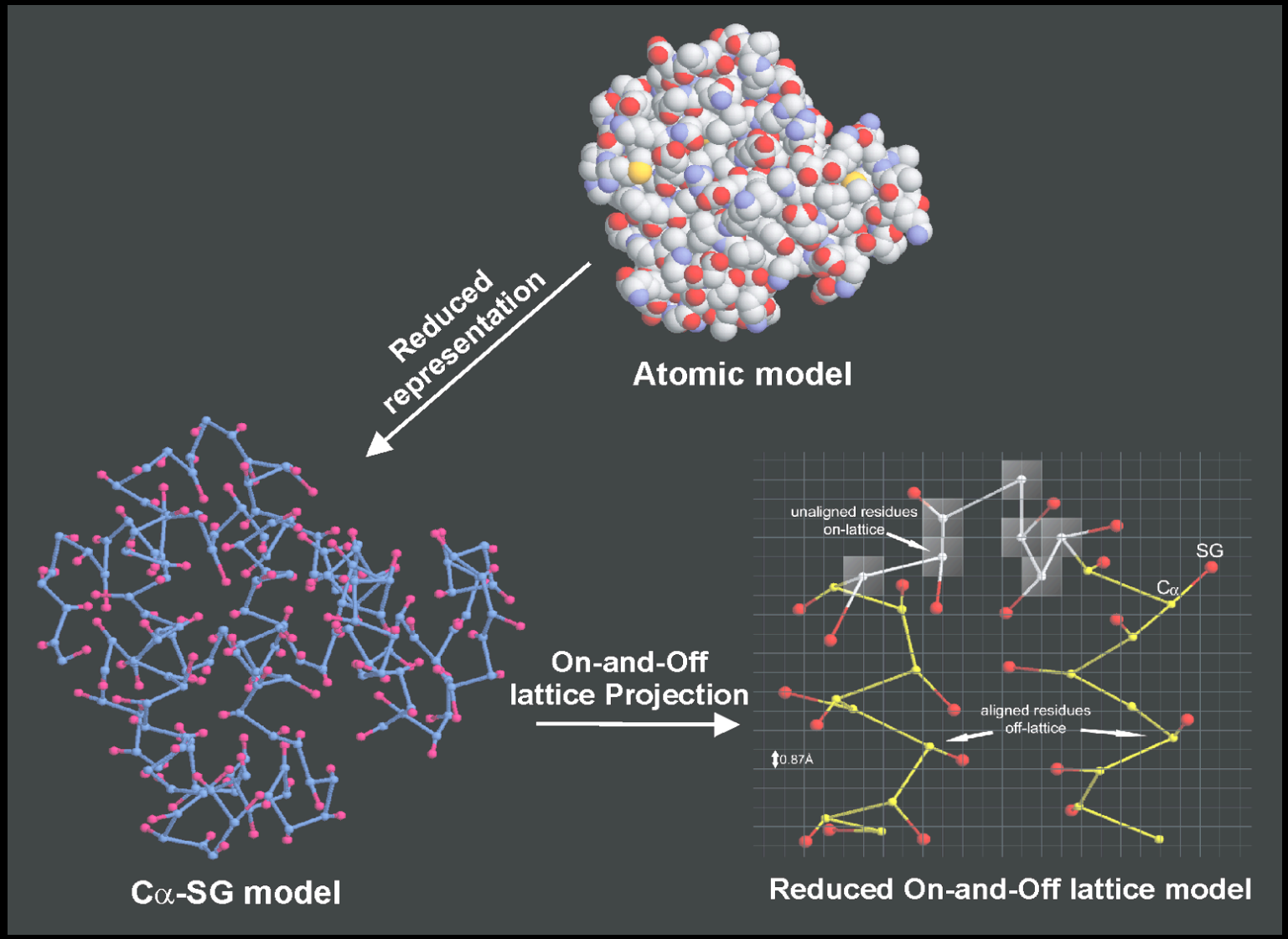
Table of Contents

I-TASSER: Iterative Threading ASSEmbly Refinement



Roy, Kucukural, Zhang. Nature Protocol (2010)

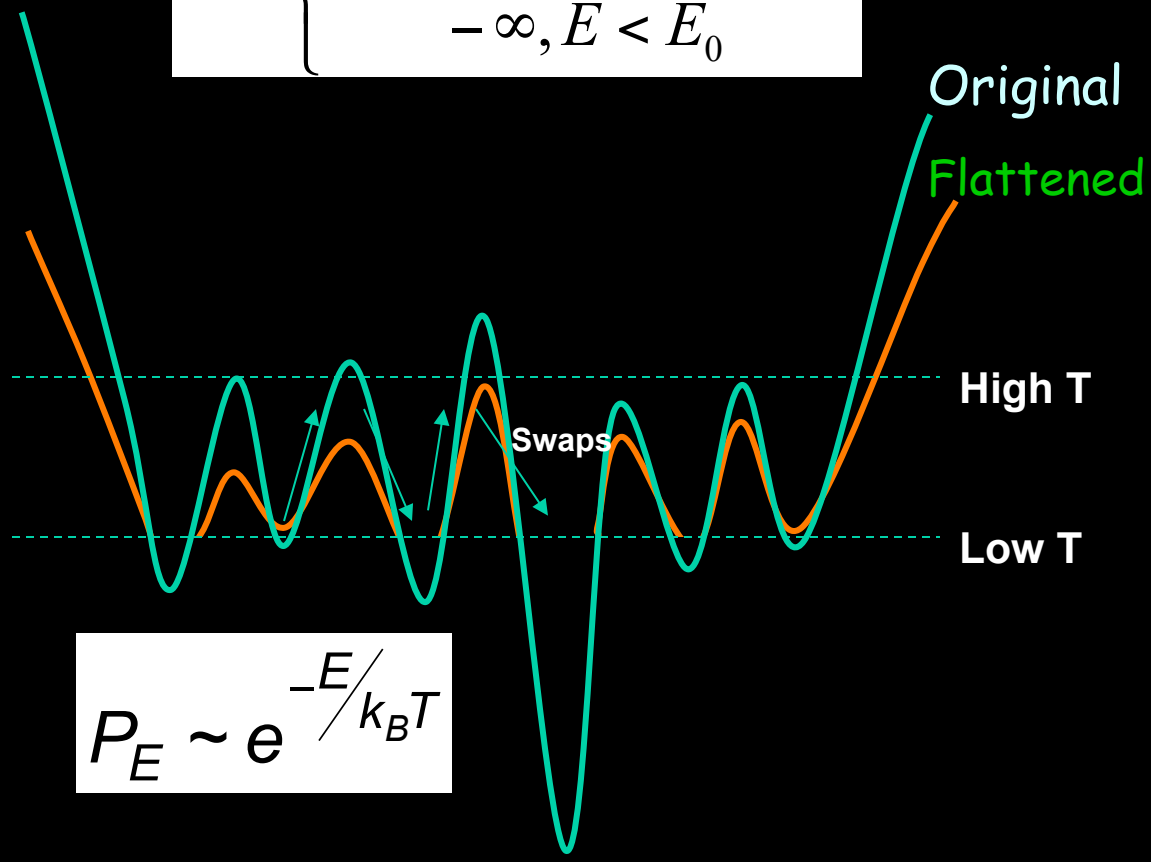
On-and-Off lattice model



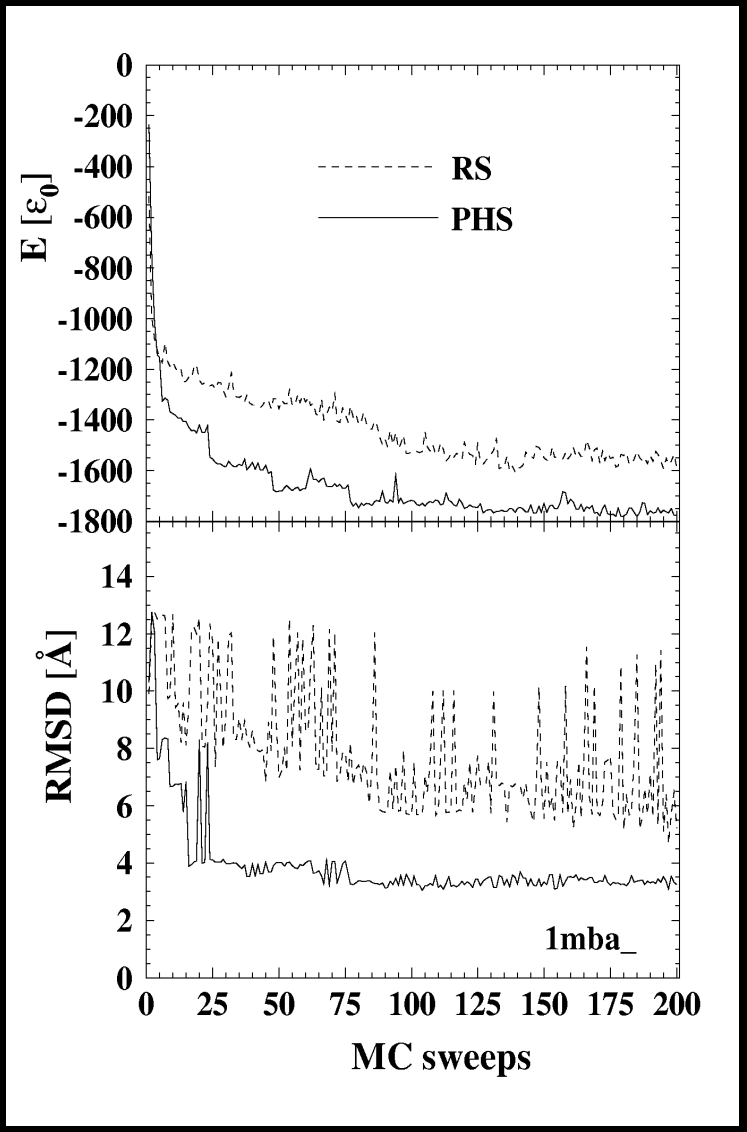
- Reduce CPU time
- Retain the accuracy of well-aligned fragments

Monte Carlo simulation for conformational search

$$E' = \begin{cases} arcsh(E - E_0), E > E_0 \\ -\infty, E < E_0 \end{cases}$$



$$P_E \sim e^{-E/k_B T}$$



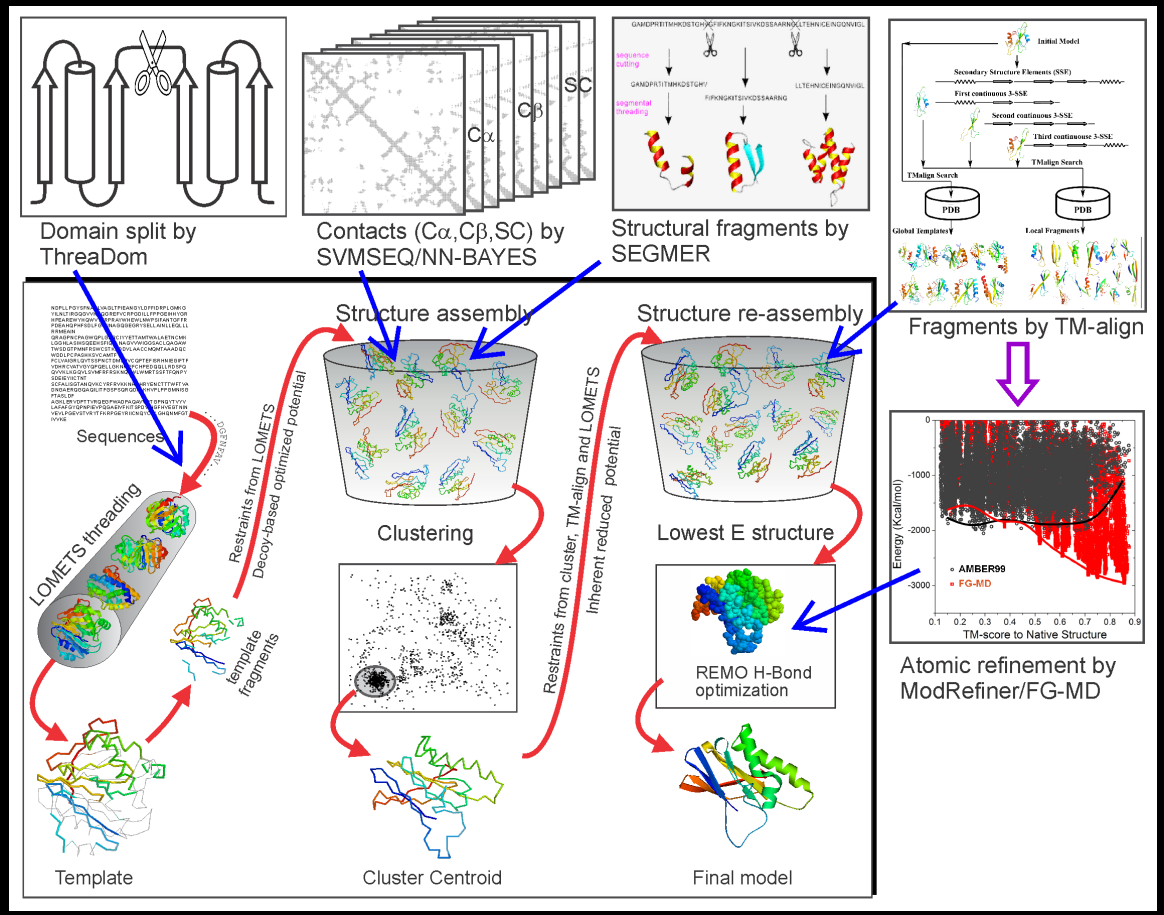
Zhang et al. *Proteins* **48**, 192 (2002)

An example from 1mba_

New progress in I-TASSER development

Major problem of I-TASSER

- Lack of long-range restraint to guide *ab initio* folding
- Loss of atomic details due to reduced modeling



Yang, Yan, Roy, Xu, Poisson, Zhang. Nature Methods (2015)

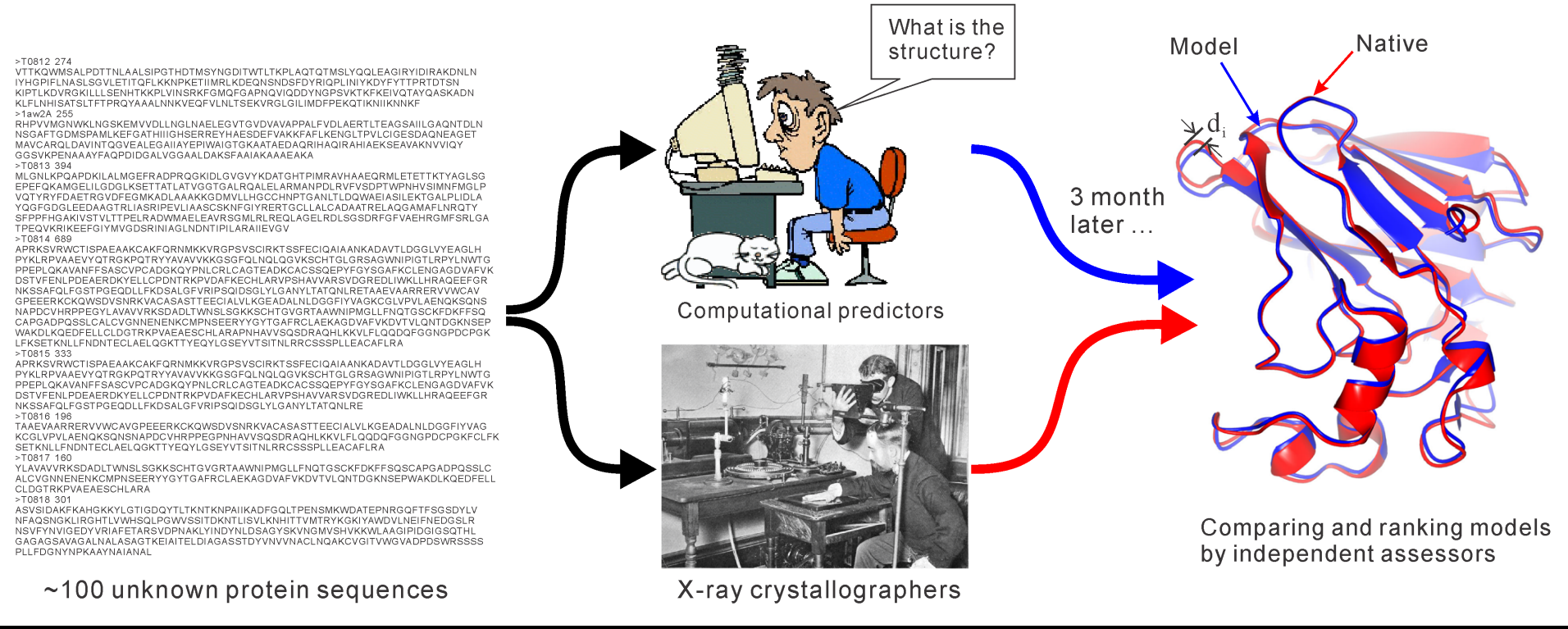
Many labs work on developing methods for protein structure prediction

Name	Institution	Software	Method
Baker	U Washington, USA	ROSETTA	Ab initio/threading
Eisenberg	UCLA, USA	BE	Threading
Elofsson	Stockholm U, Sweden	Pcons	Meta-server
Honig	Columbia U, USA	Jackal	Homologous modeling
Jones	U Coll London, UK	Mgenthreader	Threading
Karplus	Harvard U, USA	CHARMM	Ab initio
Levitt	Stanford U, USA	KoBaMIN	Ab initio/refinement
Sali	UCSF, USA	MODELLER	Homologous modeling
Scheraga	Cornell U, USA	UNRES	Ab initio
Shaw	D.E.Shaw, USA	MD	Ab initio
Skolnick	Georgia Tech, USA	TASSER	Ab initio/threading
Soding	Gene Center Munich, Germany	HHsearch	Threading
Sternberg	Imper Coll London, UK	Phyre	Threading
Zhang	U Michigan, USA	I-TASSER	Ab initio/threading/refinement

And many other methods ……Question: What works and what does not work?

Table of Contents

CASP: Olympic Games in Protein Structure Prediction



"High scoring groups in this competitive experiment are considered the *de facto* standard-bearers for what is the state of the art in protein structure prediction" (http://www.wikipedia.org)

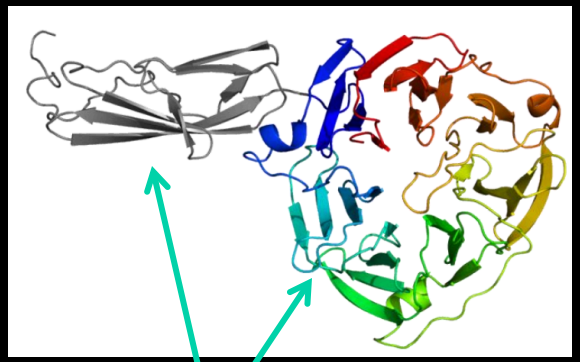
A history of CASP experiments

- CASP1 (1994), 35 groups, 33 proteins
- CASP2 (1996), 152 groups, 42 proteins
- CASP3 (1998), 120 groups, 43 proteins
- CASP4 (2000), 160 groups +38 servers, 43 proteins
- CASP5 (2002), 187 groups +72 servers, 67 proteins
- CASP6 (2004), 201 groups +65 servers, 64 proteins
- CASP7 (2006), 209 groups +98 servers, 100 proteins
- CASP8 (2008), 112 groups +121 servers, 128 proteins
- CASP9 (2010), 151 groups +140 servers, 129 proteins
- CASP10 (2012), 122 groups+94 servers, 114 proteins
- CASP11 (2014), 123 groups+84 servers, 100 proteins
- CASP12 (2106), 108 groups+80 servers, 82 proteins

Results and procedure can be seen at: http://predictioncenter.org/

CASP11 in number

- Number of human expert groups: 123
- Number of automated servers: 84
- Number of targets/domains: 126



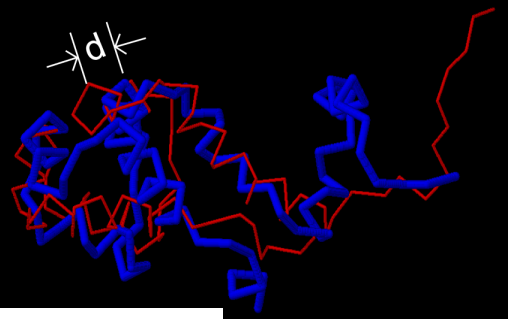
Domains are assessed individually

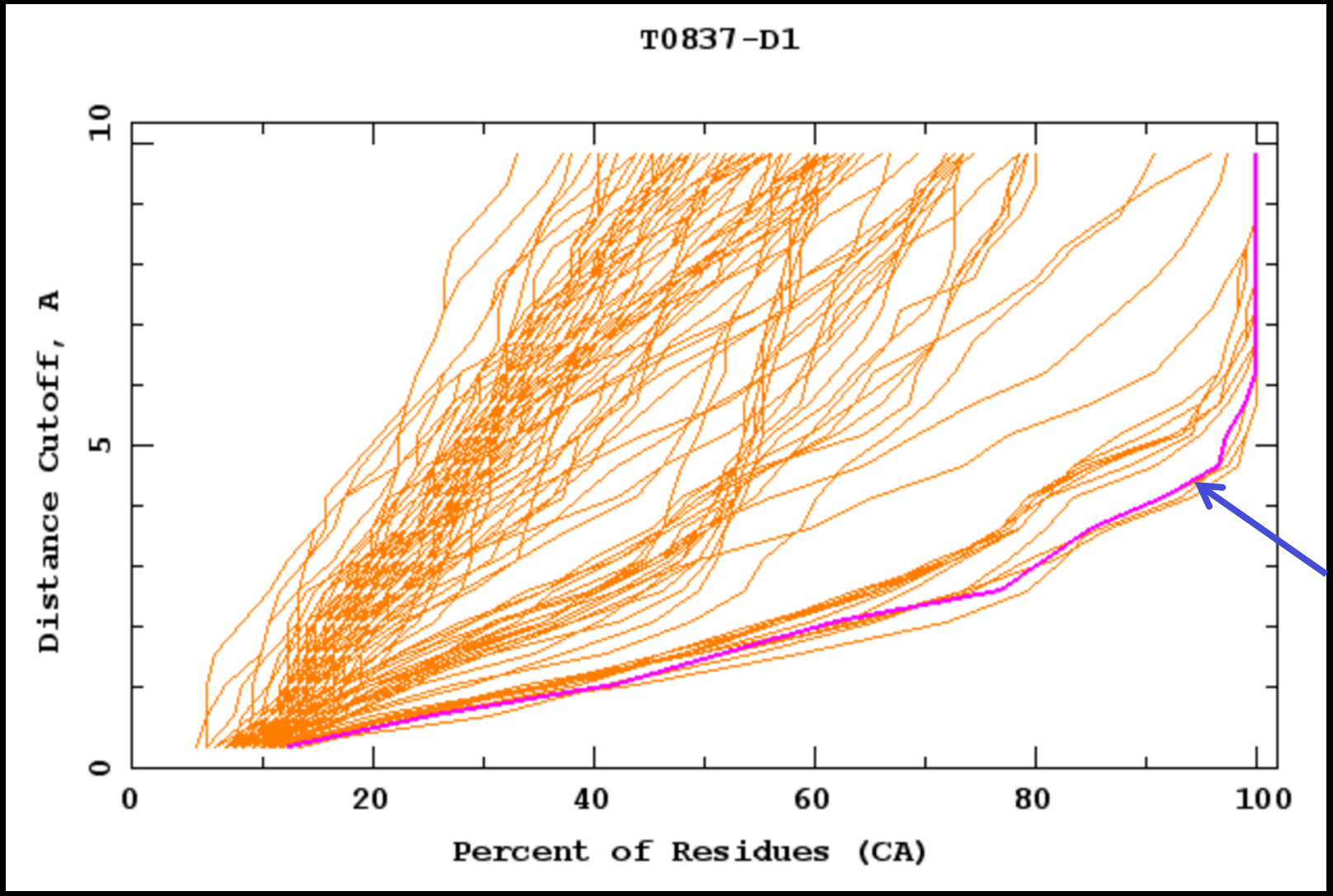
Three categories:

81 TBM: Template based modeling targets

45 FM: Free modeling targets

How does CASP assess model quality?





A summary ranking on all targets/domains

#	GR code	GR name	Domains Count	SUM Z-score (>-2.0)	Rank SUM Z-score (>-2.0)	AVG Z-score (>-2.0)	Rank AVG Z-score (>-2.0)	SUM Z-score (>0.0)	Rank SUM Z-score (>0.0)	AVG Z-score (>0.0)	Rank AVG Z-score (>0.0)
1	204	Zhang	78	76.4117	1	0.9796	1	77.9365	1	0.9992	1
2	169	LEE	78	68.7497	2	0.8814	2	73.7065	2	0.9450	2
3	290	MULTICOM	78	66.7849	3	0.8562	3	70.4849	4	0.9037	4
4	044	LEER	78	66.5034	4	0.8526	5	72.5600	3	0.9303	3
5	277	Zhang-Server	78	65.9858	5	0.8460	6	70.1240	5	0.8990	5
6	425	Seok-refine	78	63.2947	6	0.8115	7	67.6359	6	0.8671	8
7	499	QUARK	78	59.5585	7	0.7636	10	63.4839	11	0.8139	14
8	065	Jones-UCL	75	58.0721	8	0.8543	4	67.3873	7	0.8985	6
9	042	TASSER	78	56.5341	9	0.7248	11	60.3926	13	0.7743	21
10	338	ProQ2	78	56.3264	10	0.7221	12	61.6675	12	0.7906	17
11	132	ProQ2-refine	78	55.5291	11	0.7119	15	60.3740	14	0.7740	22
12	333	Kiharalab	76	55.0840	12	0.7774	9	66.3332	8	0.8728	7
13	347	Wallner	78	54.3184	13	0.6964	17	59.5259	15	0.7632	23
14	358	Skwark	78	53.0744	14	0.6804	19	63.7288	10	0.8170	13
15	067	CNIO	78	51.5664	15	0.6611	22	64.6887	9	0.8293	9
16	282	PML	77	48.9282	16	0.6614	21	58.6265	17	0.7614	24
17	144	Mufold	78	46.3855	17	0.5947	26	53.2095	22	0.6822	29
18	438	QA-RecombineIt_H	74	43.2909	18	0.6931	18	58.2607	19	0.7873	18
19	241	SHORTLE	75	42.5868	19	0.6566	23	58.5716	18	0.7810	19
20	364	QA-RecombineIt_WFH	72	39.8400	20	0.7200	13	59.4192	16	0.8253	11
21	064	BAKER	78	39.7843	21	0.5101	30	55.3932	21	0.7102	28
22	162	McGuffin	78	35.8029	22	0.4590	32	44.0066	28	0.5642	36
23	038	nns	78	35.5076	23	0.4552	33	46.5890	25	0.5973	34
24	482	wfMix-KPa	72	35.1184	24	0.6544	24	57.5703	20	0.7996	16
25	434	QA-RecombineIt_H2	72	31.8436	25	0.6089	25	51.9407	24	0.7214	27
26	056	wfMix-KPb	72	29.9382	26	0.5825	27	52.5422	23	0.7298	26
27	317	keasar	78	25.0417	27	0.3210	40	43.7018	29	0.5603	37

A history of CASP experiments

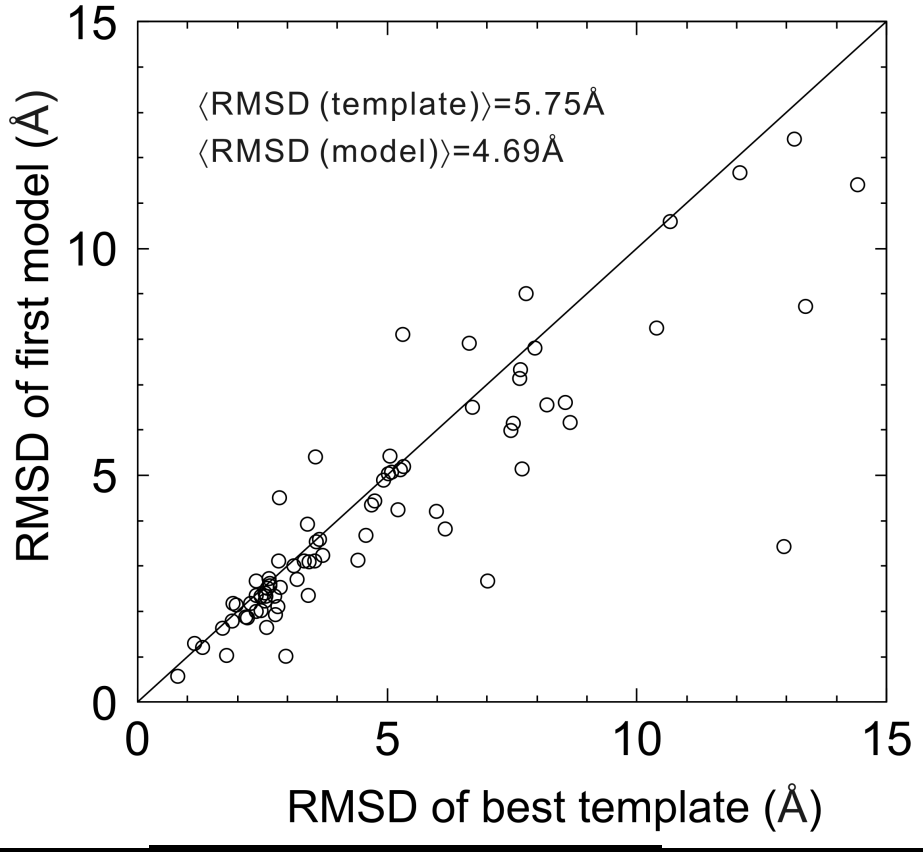
- CASP1 (1994), 35 groups, 33 proteins
- CASP2 (1996), 152 groups, 42 proteins
- CASP3 (1998), 120 groups, 43 proteins
- CASP4 (2000), 160 groups +38 servers, 43 proteins
- CASP5 (2002), 187 groups +72 servers, 67 proteins
- CASP6 (2004), 201 groups +65 servers, 64 proteins
- CASP7 (2006), 209 groups +98 servers, 100 proteins
- CASP8 (2008), 112 groups +121 servers, 128 proteins
- CASP9 (2010), 151 groups +140 servers, 129 proteins
- CASP10 (2012), 122 groups+94 servers, 114 proteins
- CASP11 (2014), 123 groups+84 servers, 100 proteins
- CASP12 (2106), 108 groups+80 servers, 82 proteins

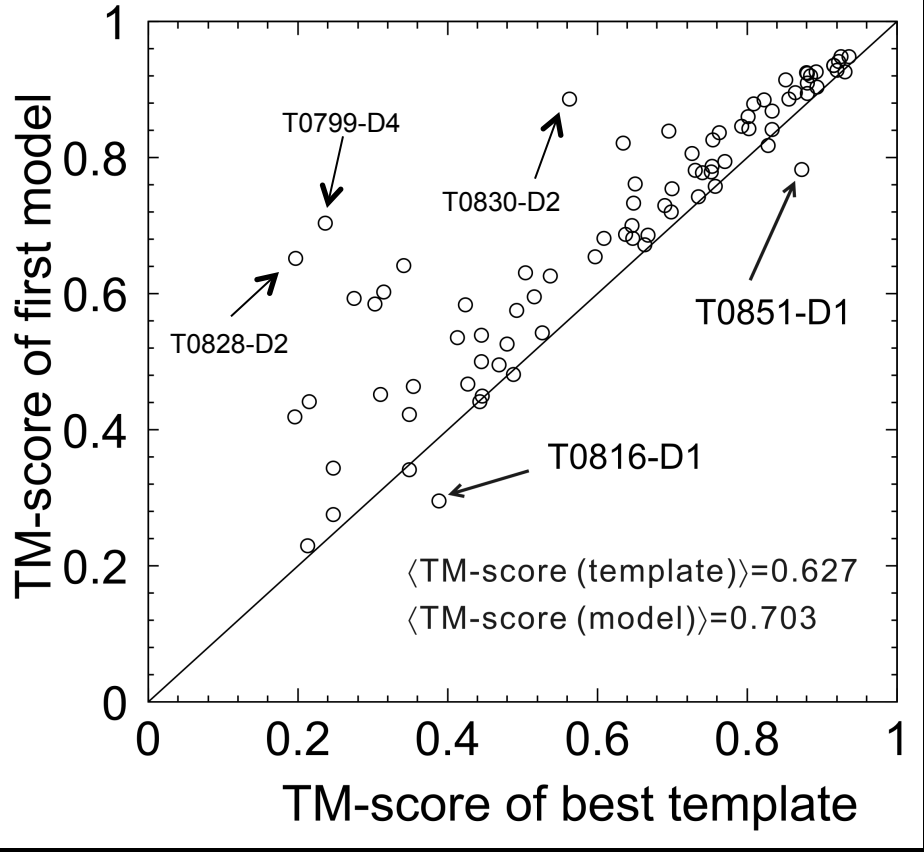
Results and procedure can be seen at: http://predictioncenter.org/

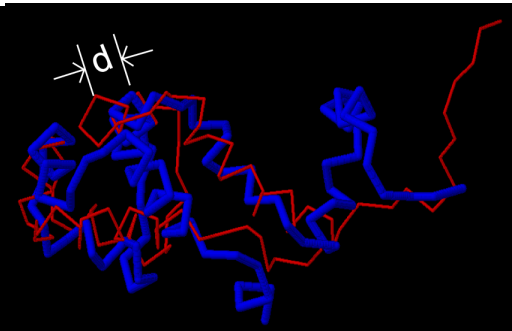
Template based modeling in CASP11 (TBM)

GOAL: how to identify the best template and how to refine the template closer to the native

First Zhang-server model vs best LOMETS templates (82 domain/targets)



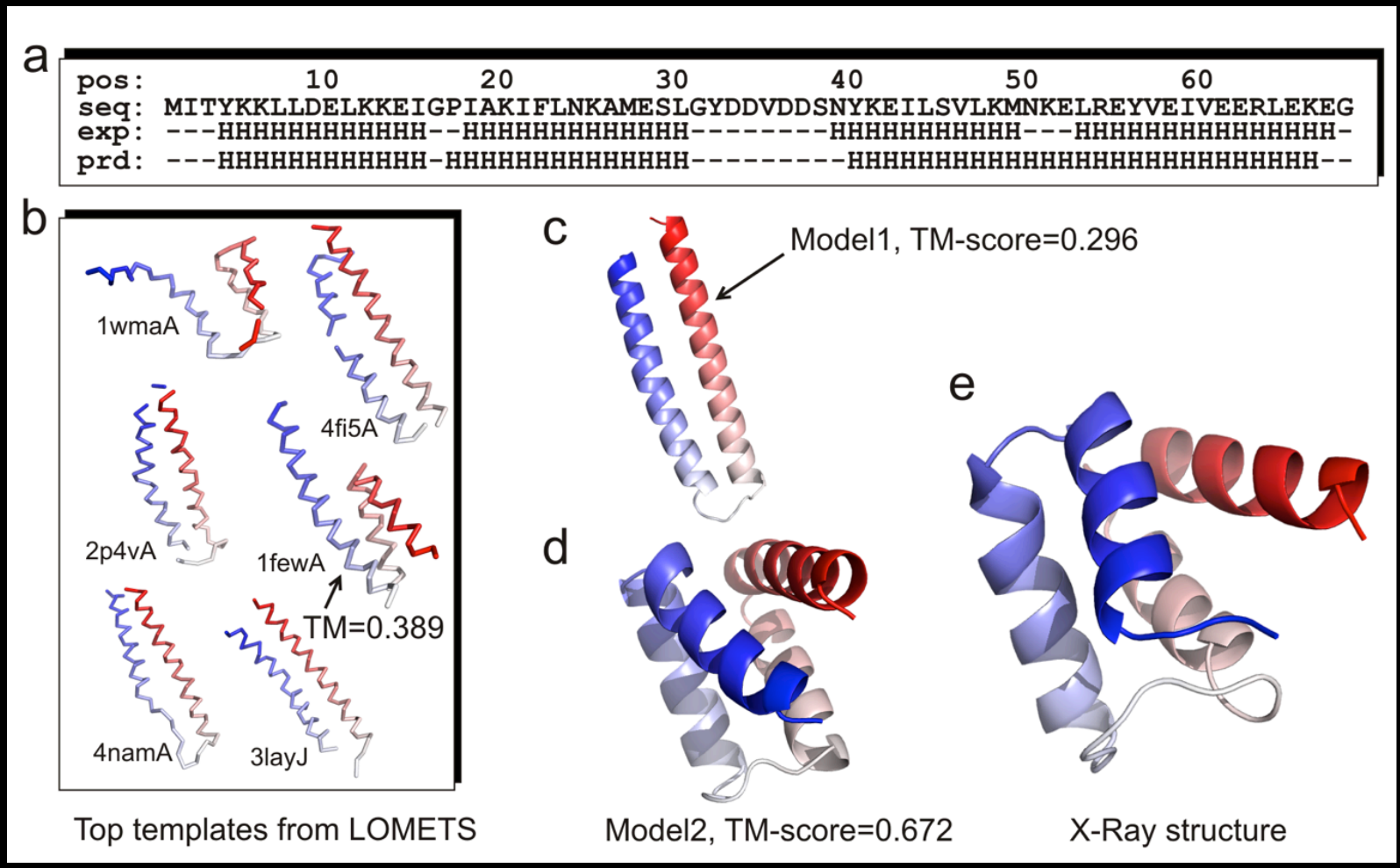




$$\mathrm{RMSD} = \sqrt{\frac{1}{L}\sum_{i=1}^{L} d_i^2}$$

$$\text{TM-score} = \frac{1}{L}\sum_{i=1}^{L_{ali}} \frac{1}{1 + d_i^2 / d_0^2}, \qquad d_0 = 1.24\sqrt[3]{L-15} - 1.8$$

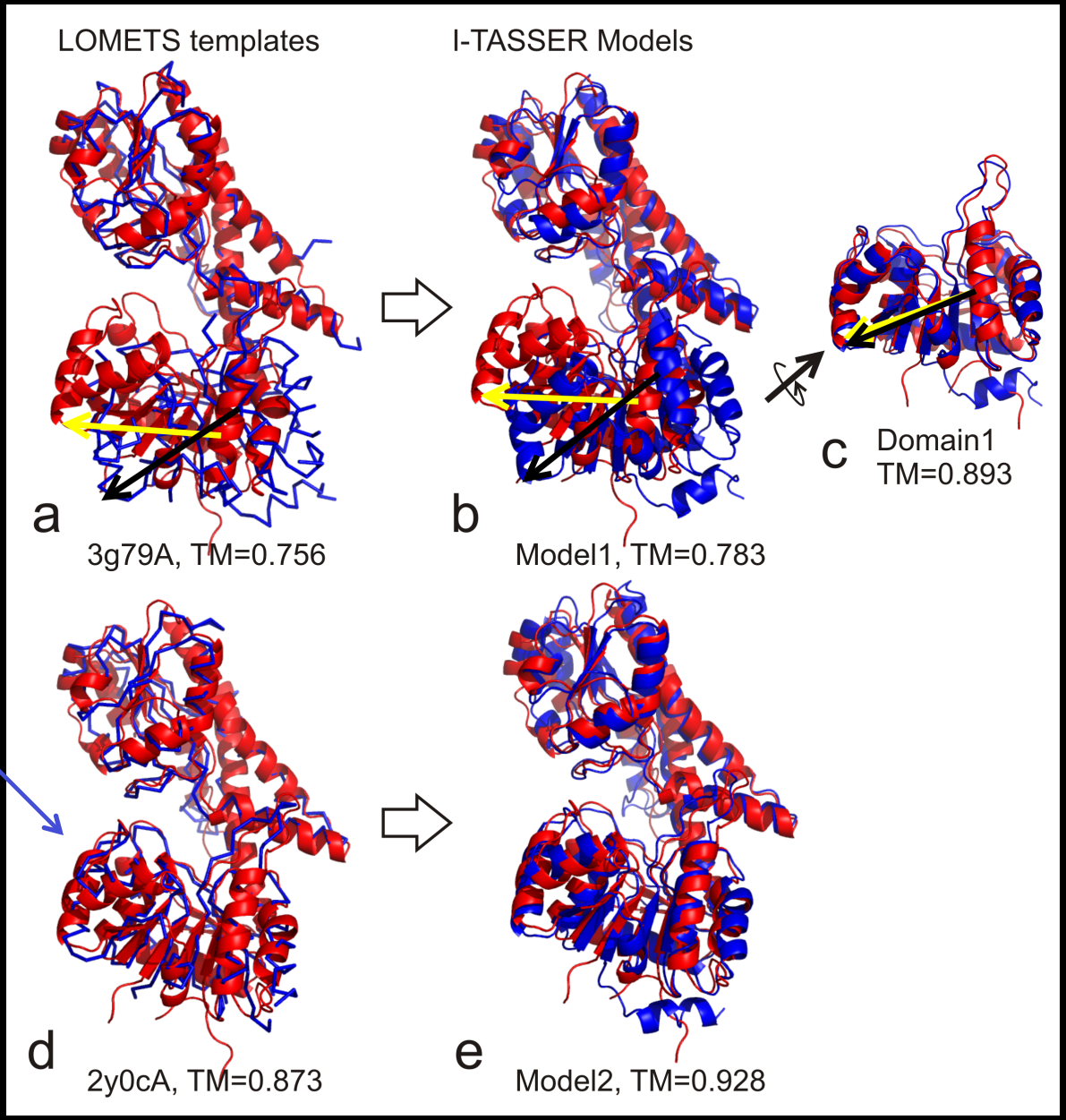
T0816-D1: Problem in model selection



Why did I-TASSER selects the wrong template as the first model?

1, SS prediction has a long helix at C-terminal (favor the two-helix bundle model)
2, consensus threading templates prefers the two-helix bundle model

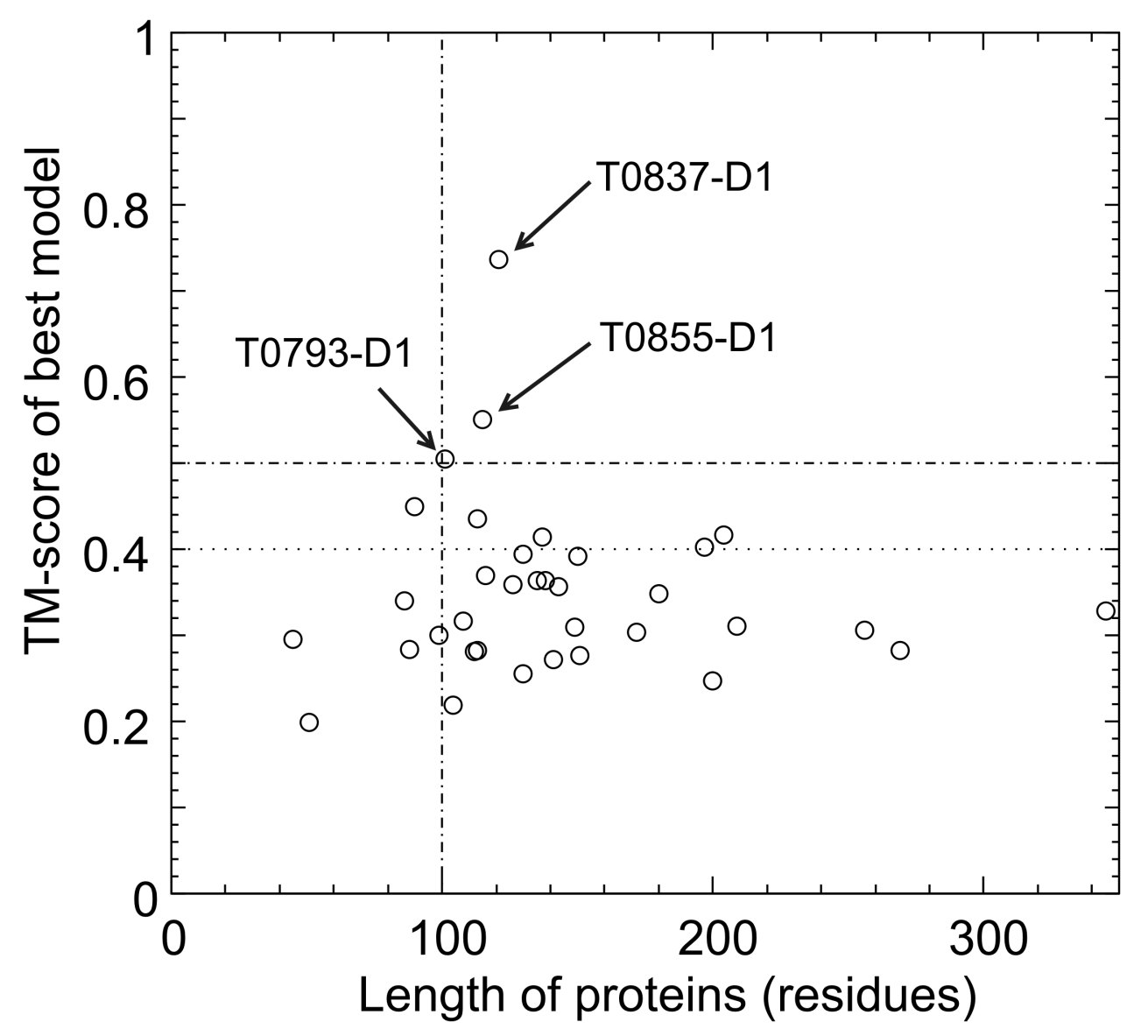
T0851-D1: Problem in domain orientation



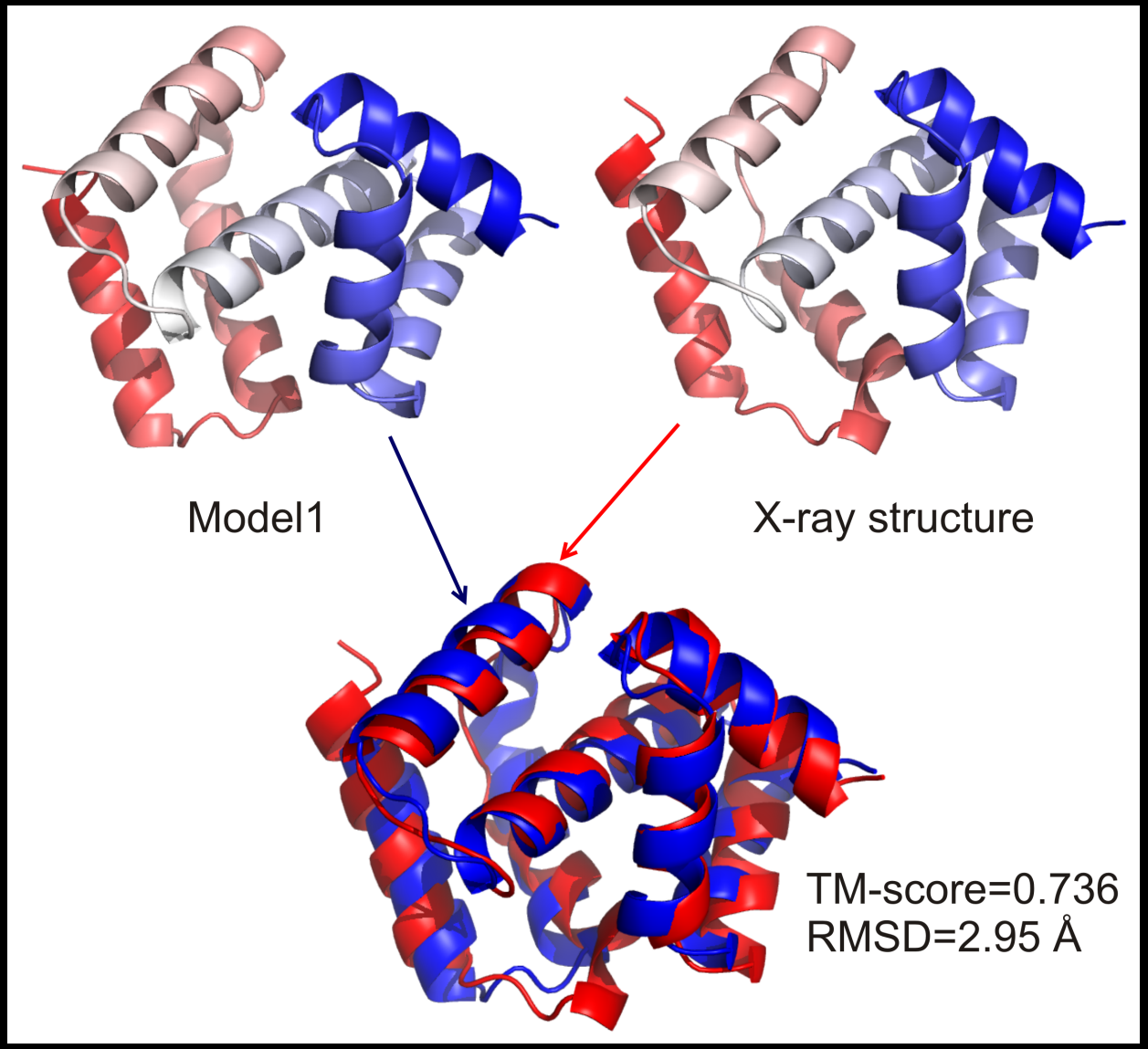
Free modeling in CASP11 (FM)

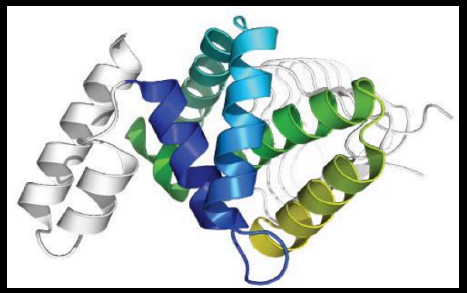
GOAL: how to construct correct fold from scratch

Summary of Free Modeling



QUARK modeling of T0837-D1 (128 AA)



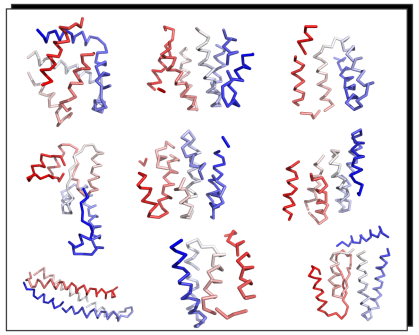


Closest structure in PDB (3mc4)

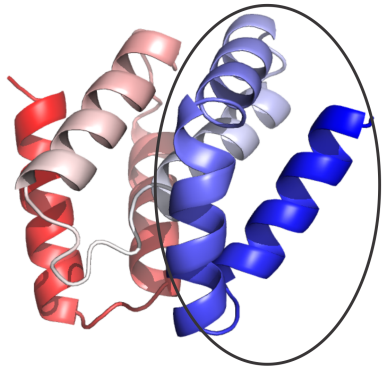
Assessor's comment: T0837-D1_499_1 represents the FM model with biggest improvement from PDB templates in CASP11 experiment.

Why did QUARK do so well on T0837-D1?

Threading fragments
Best TM-score=0.304



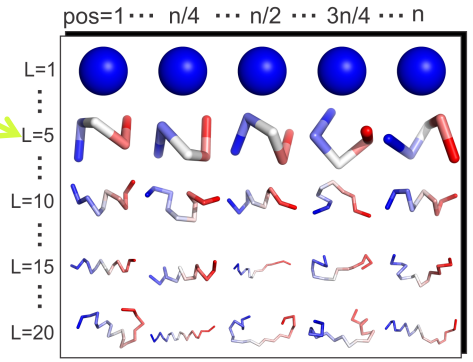
(A) Top templates, best TM-score=0.304



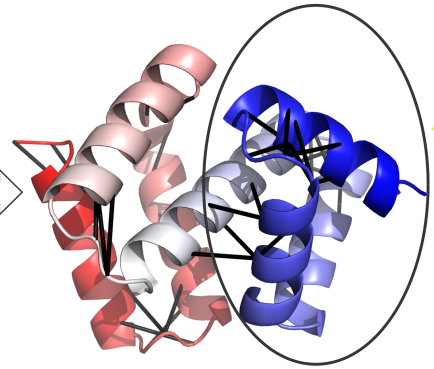
I-TASSER model1, TM-score=0.428

Fragment assembly
by I-TASSER
TM-score=0.428

QUARK fragments
RMSD ~ 0.1-2.6Å



Top fragments at each position

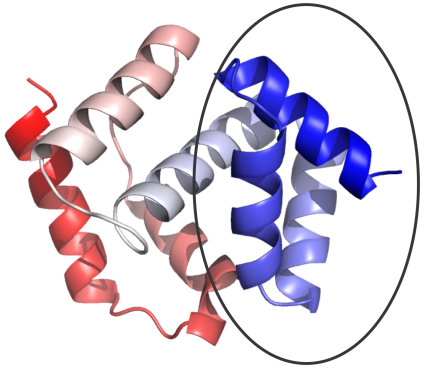


QUARK model1, TM-score=0.736

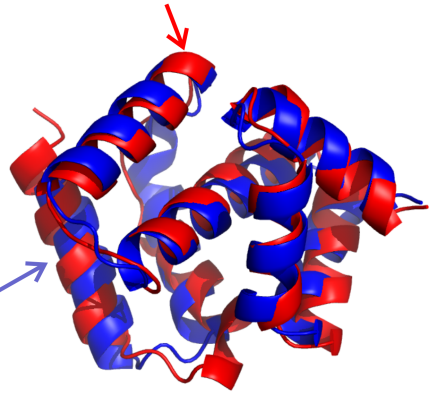
QUARK contact
prediction with
dist_error=0.9Å

(B)

X-ray structure

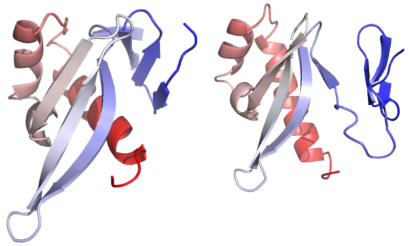


(C) X-ray structure

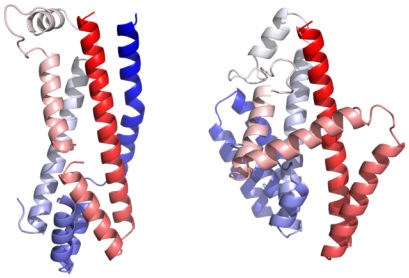


Superposition of QUARK model1 (blue)
and X-ray structure (red), RMSD=2.9Å

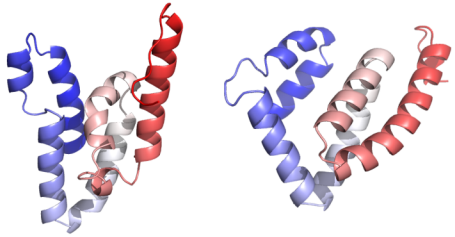
Other examples of FM models with TM-score > 0.4



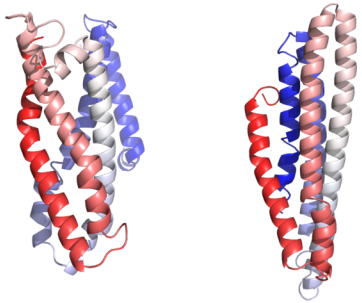
T0793-D1, L=101
TM-score=0.505
RMSD=8.4 Å



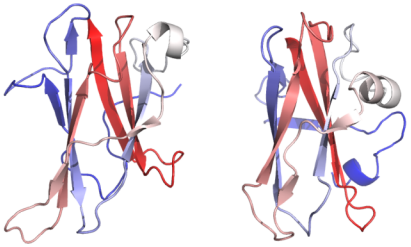
T0831-D2, L=197
TM-score=0.402
RMSD=15.7 Å



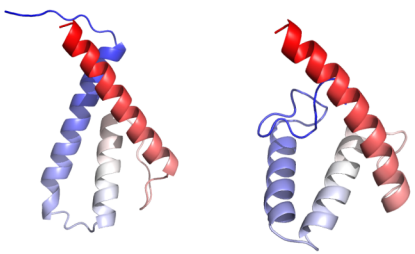
T0810-D1, L=113
TM-score=0.435
RMSD=8.1 Å



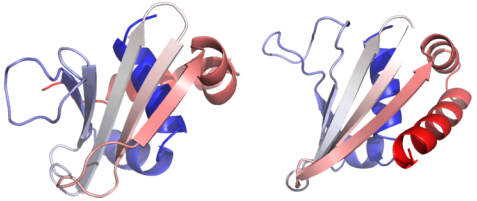
T0836-D1, L=204
TM-score=0.417
RMSD=12.1 Å



T0814-D1, L=137
TM-score=0.414
RMSD=8.4 Å



T0820-D1, L=90
TM-score=0.449
RMSD=8.5 Å



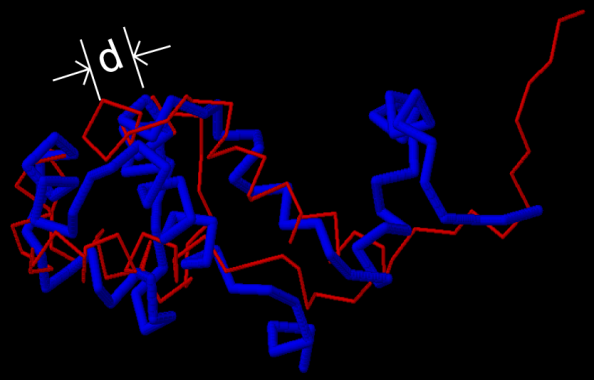
T0855-D1, L=115
TM-score=0.551
RMSD=7.7 Å

Overall QUARK folds 4, I-TASSER folds 8 FM targets with TM>0.4

Top 10 groups in CASP11 on 126 targets

Data from http://predictioncenter.org/casp11/zscores_final.cgi

Groups	GDT (Z-score)	Institutions
Zhang-Server	6110 (132.4)	University of Michigan, USA
QUARK	6074 (125.5)	University of Michigan, USA
nns	5750 (77.7)	KIAS, Korea
Myprotein-me	5582 (68.7)	Aalto University, Finland
Baker-Rosetta	5542 (68.1)	University of Washington, USA
MulticonConst	5562 (60.8)	University of Missouri, USA
Tasser-VMT	5443 (43.6)	Georgia Institute of Technology, USA
RaptorX	5503 (31.3)	Toyota Institute at Chicago, USA
HHPredA	5377 (22.0)	Ludwig-Maximilians University, Germany
Falcon_topo	5215 (17.2)	Waterloo University, Canada



$$GDT = \frac{1}{4L}\left(n_{d<1} + n_{d<2} + n_{d<4} + n_{d<8}\right)$$

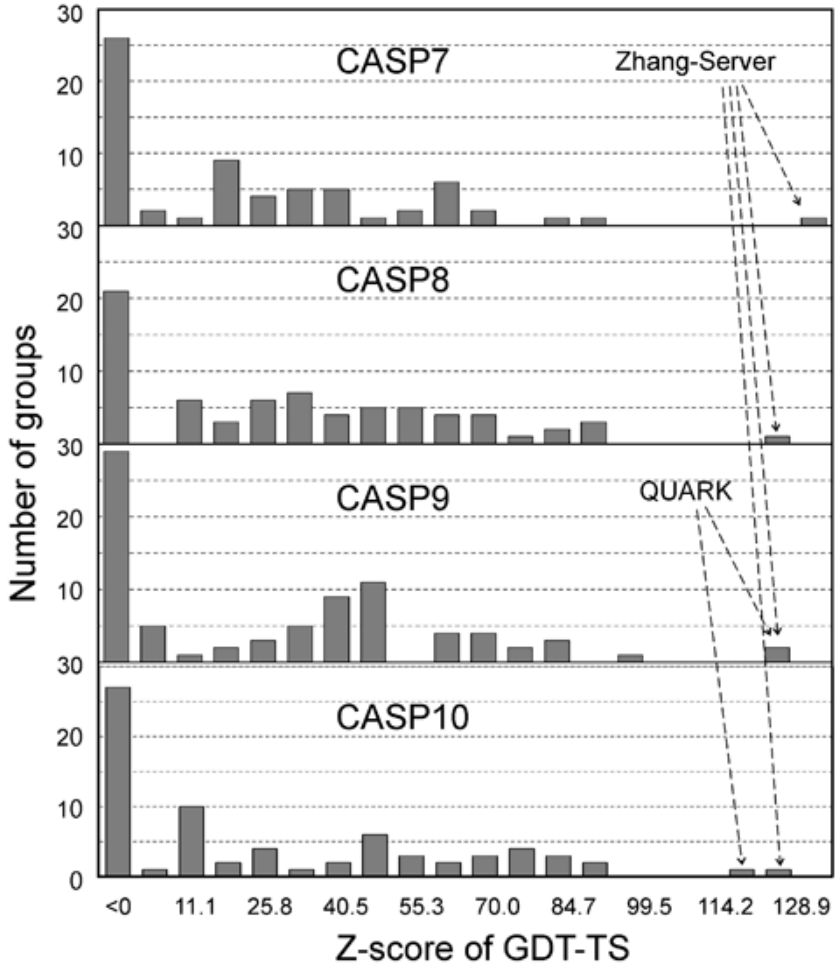
$$Z - score = \frac{GDT_{group} - \langle GDT \rangle}{\sigma}$$

$n_{d<x}$: number of residues with d below x Angstroms

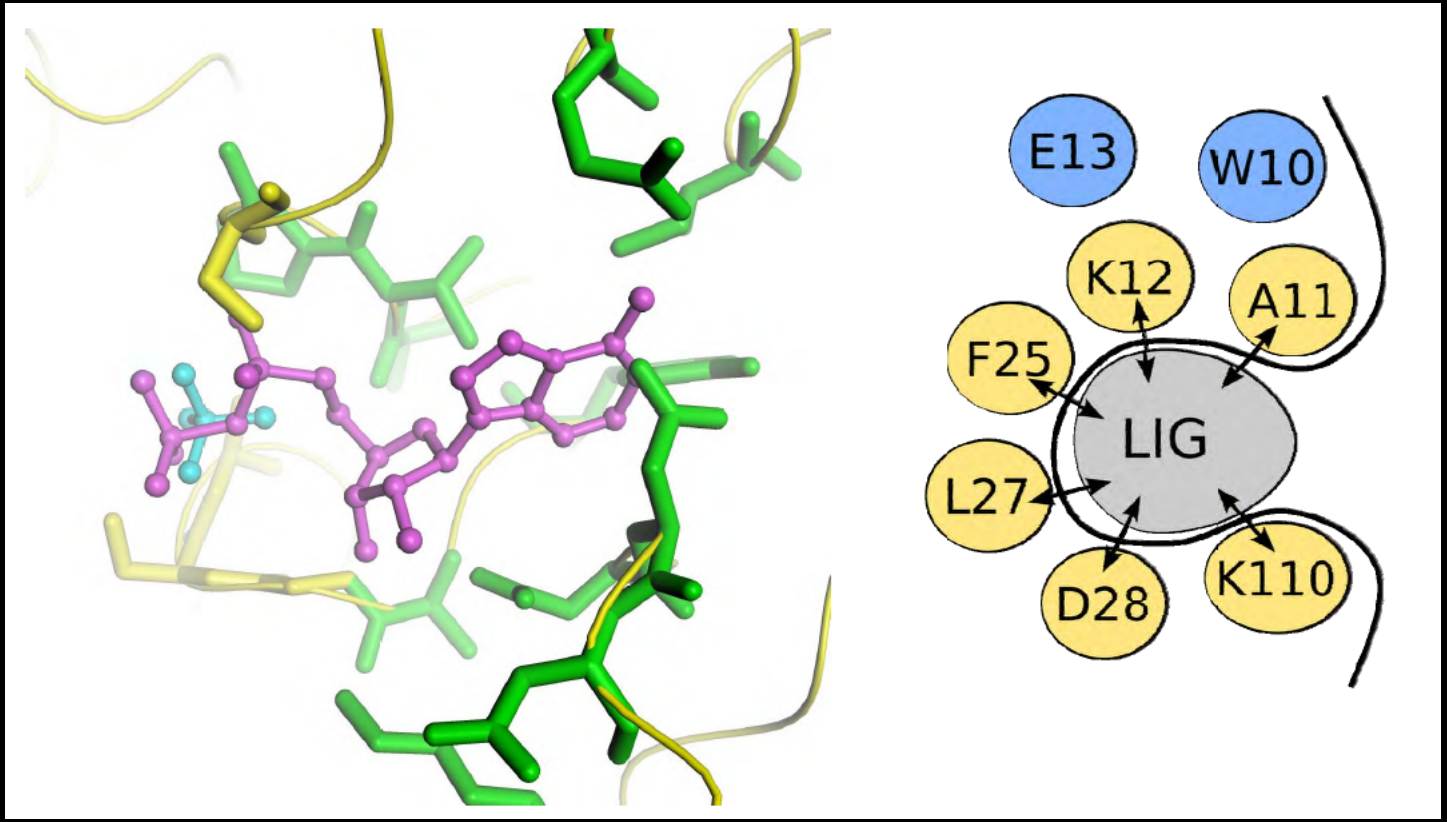
Summary of protein structure prediction in previous CASP experiments (7th-10th)

CASP7 (124 targets, 98 groups)		CASP8 (164 targets, 122 groups)	
Groups	GDT (Z-score)	Groups	GDT (Z-score)
Zhang-Server	7604 (112.4)	Zhang-Server	11217 (124.8)
HHpred2	7194 (63.8)	Raptor	10834 (93.4)
Pmodeller6	7169 (82.3)	Pro-sp3-Tasser	10786 (95.4)
Circle	7109 (63.6)	Baker-Robetta	10727 (94.2)
Baker-Robetta	7087 (77.4)	Phyre_de_novo	10723 (84.7)
MetaTasser	7077 (68.1)	Multicomcluster	10639 (79.3)
Raptor-Ace	6970 (55.7)	MUProt	10548 (71.4)
SP3	6938 (47.4)	Hhpred4	10495 (67.2)
Beautshot	6926 (50.6)	GSKudlatyPred	10483 (73.9)
Uni-Eid-Bnmx	6913 (45.9)	FAMSD	10439 (65.5)

CASP9 (147 targets, 140 groups)		CASP10 (127 targets, 122 groups)	
Groups	GDT (Z-score)	Groups	GDT (Z-score)
Zhang-Server	9226 (96.9)	Zhang-Server	7597 (104.0)
QUARK	9213 (100.6)	QUARK	7546 (97.5)
RaptorX-MSA	9081 (85.2)	RaptorX-ZY	7339 (79.2)
Seok-Server	8843 (66.8)	HHpredA	7244 (68.1)
HHpredA	8751 (54.9)	PMS	7237 (74.2)
MulticomRefine	8749 (64.4)	Baker-Rosetta	7225 (79.3)
Chunk-Tasser	8650 (59.7)	Tasser-VMT	7188 (68.2)
Phyre2	8647 (52.7)	PconsM	7094 (66.9)
Gws	8545 (55.8)	MulticonNovel	7078 (57.7)
Baker-Robetta	8521 (61.3)	MUfold-Servr	6964 (39.1)

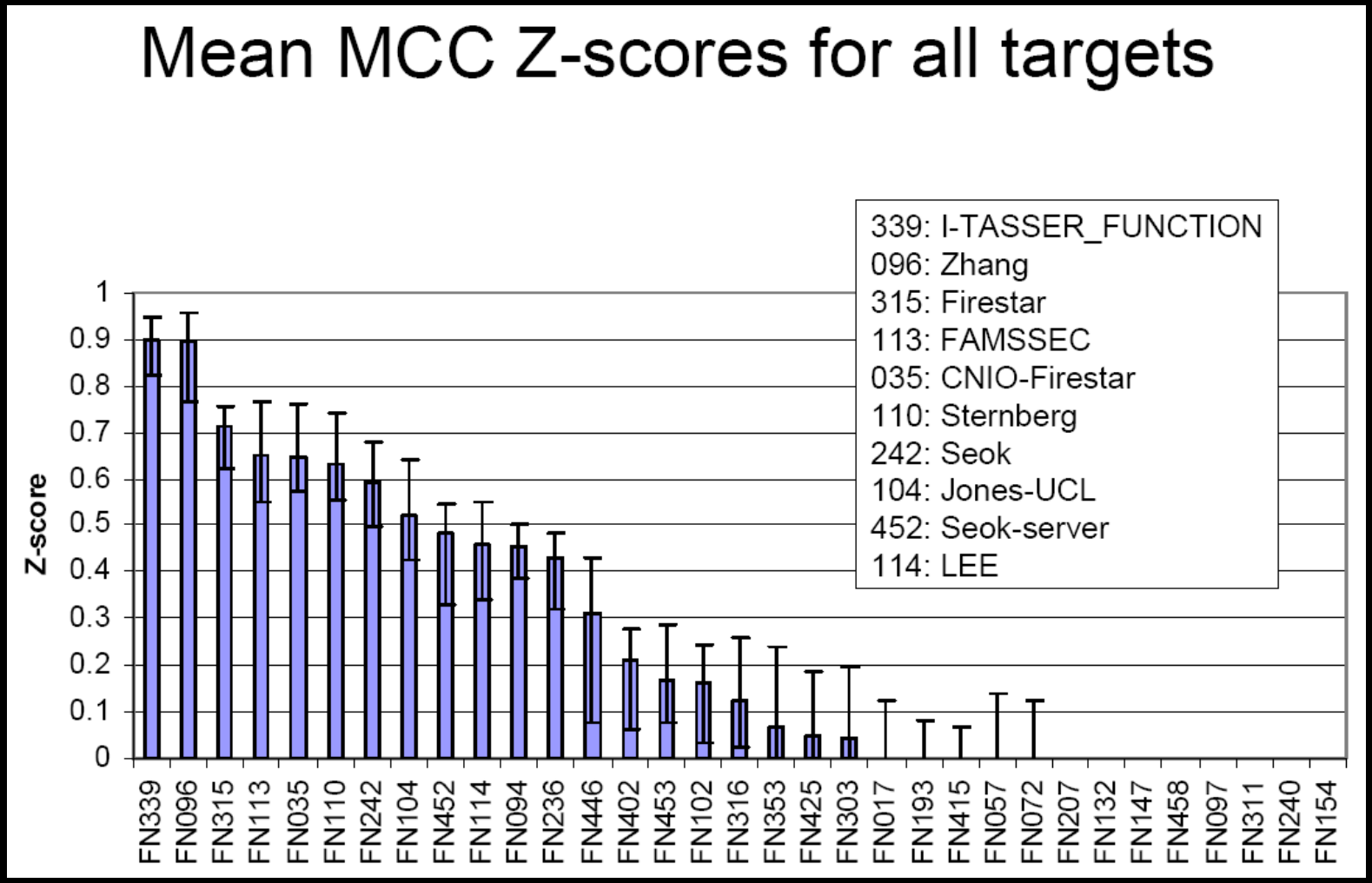


Function prediction: A new category in CASP9



$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Top 10 groups in CASP9 (protein function prediction)



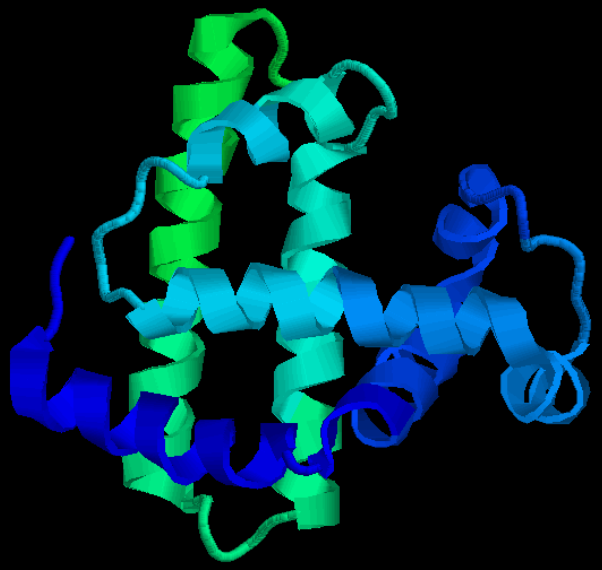
Demonstrate efficiency of structure-based function prediction.

Table of Contents

Protein structure prediction Problem

MVLSEGEWQLVLHVWAKV
EADVAGHGQDILIRLFKSHP
ETLEKFDRVKHLKTEAEMK
ASEDLKKHGVTVLTALGAIL
KKKGHHEAELKPLAQSHAT
KHKIPIKYLEFISEAIIHVLHS
RHPGNFGADAQGAMNKAL
ELFRKDIAAKYKELGYQG

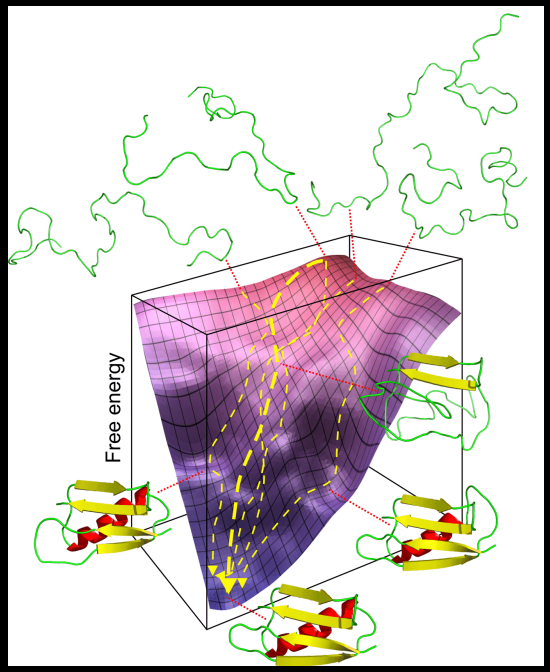
Is it possible?

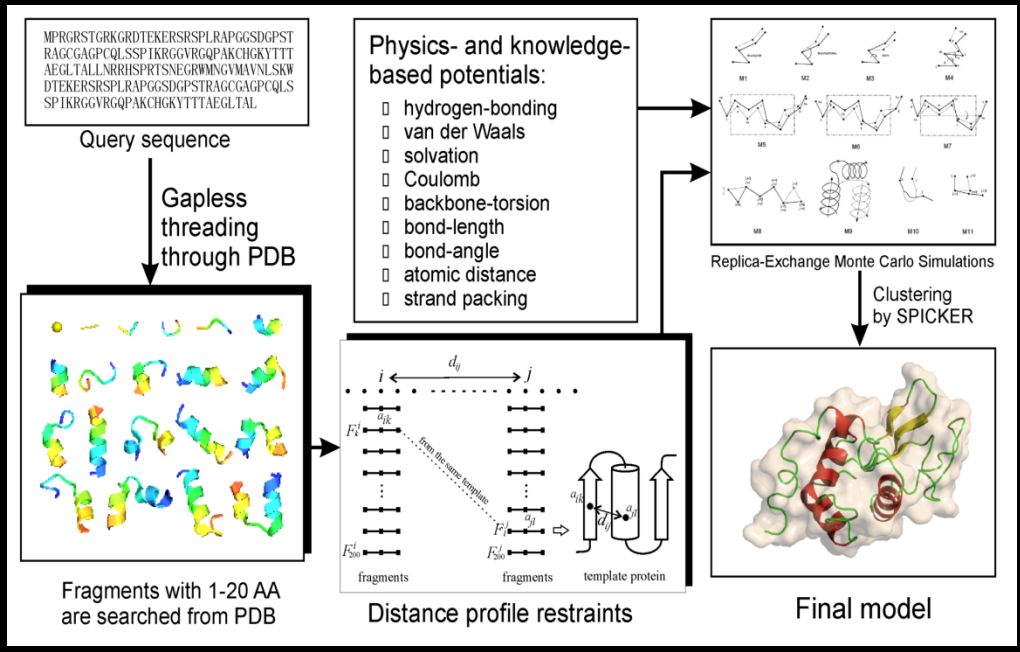


Has the problem solved?

What remain difficult?

I. We can not fold a protein without using templates (*ab initio* folding)



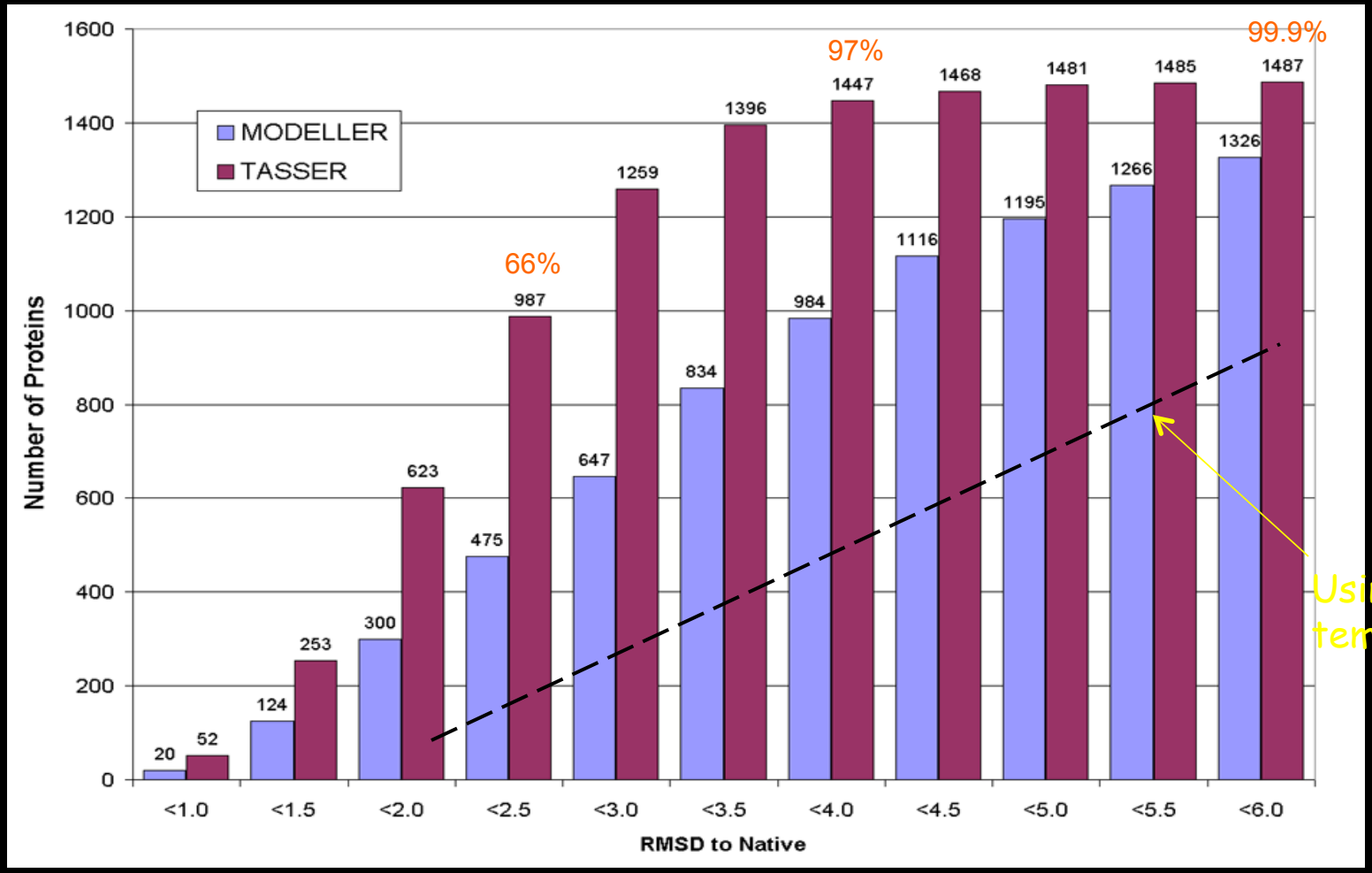


Pure physics based methods do not work for protein structure prediction

The best method is pseudo-*ab initio* technique but work best only for small alpha-proteins.

What remain difficult?

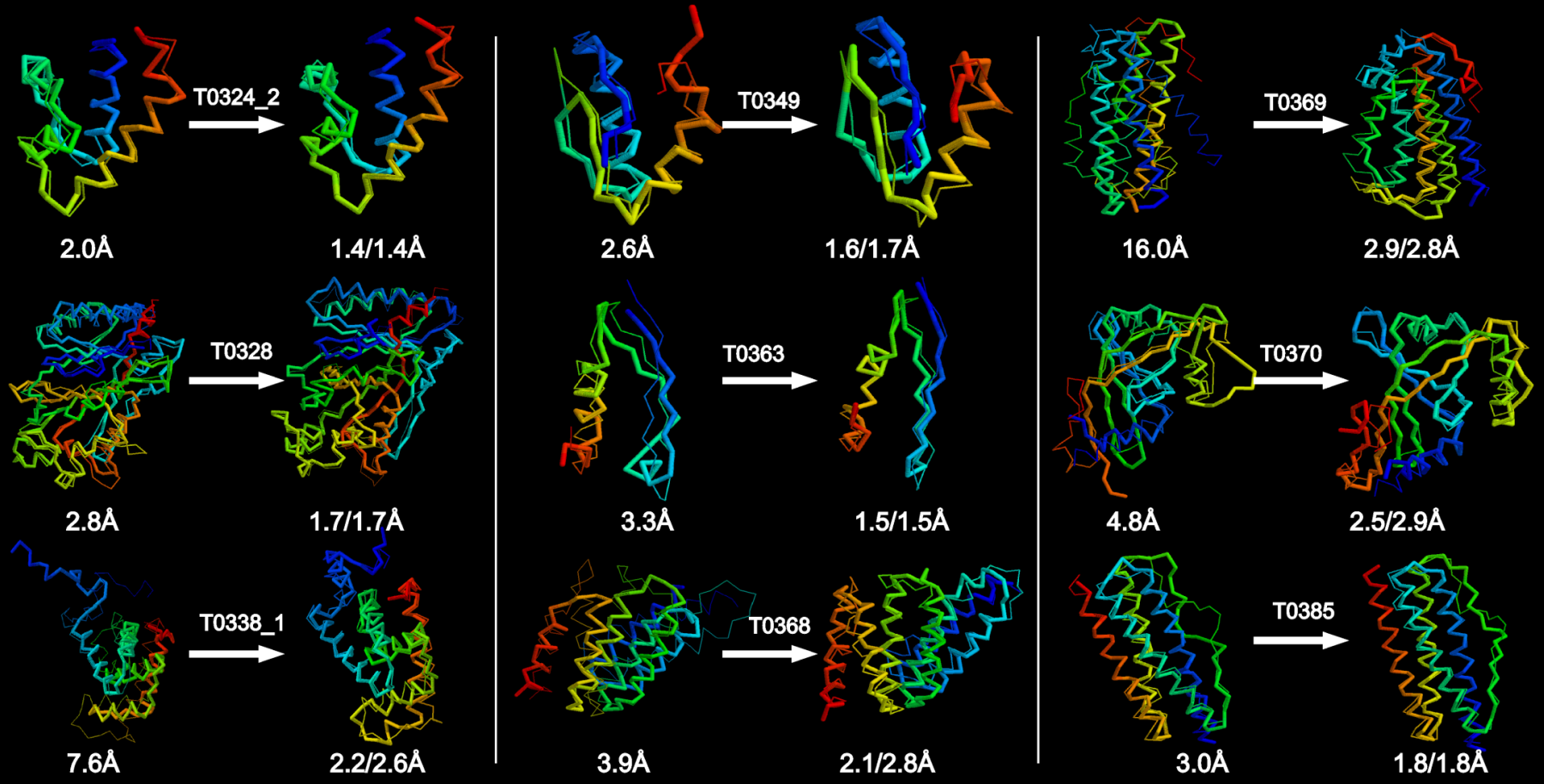
II. We cannot identify the best, non-homologous templates from PDB



Zhang, Skolnick, PNAS (2005)

What remain difficult?

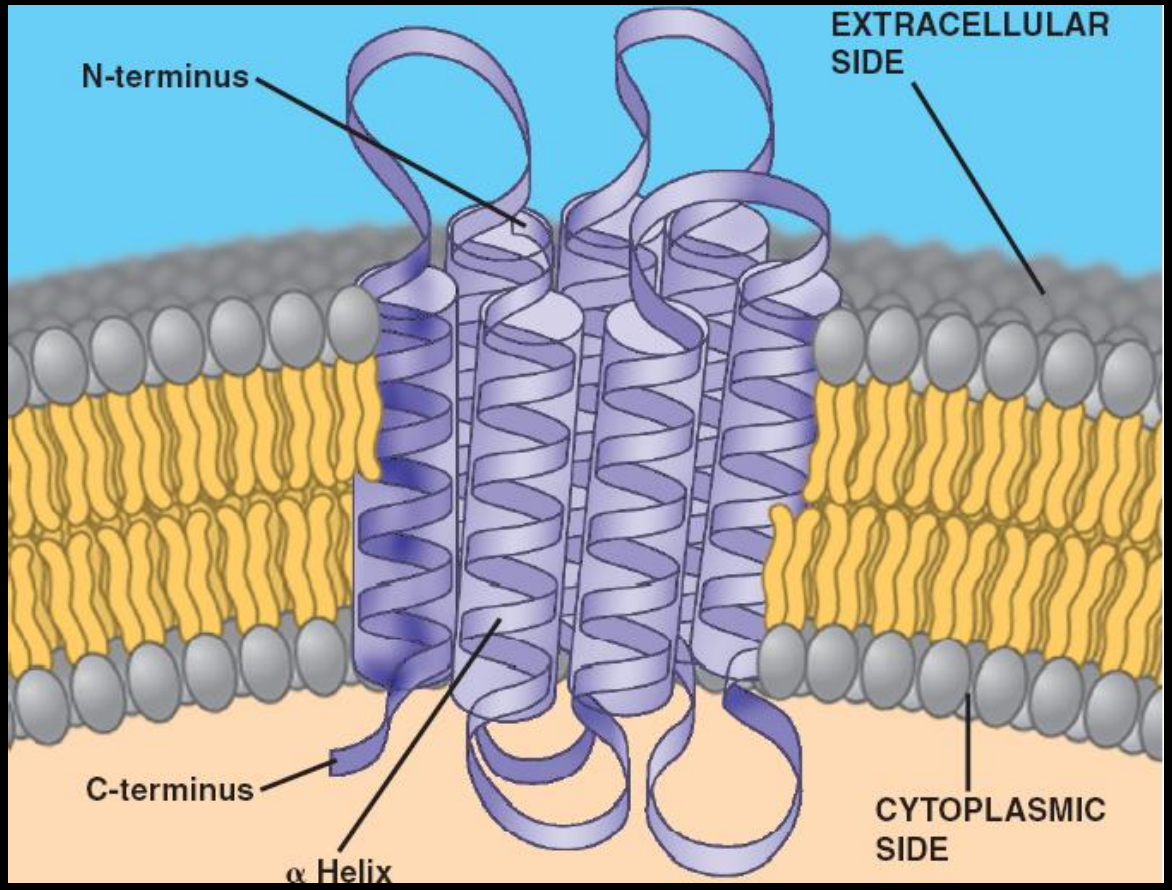
III. How to refine the templates to atomic resolution?



- Drug screening usually requires ~1-2 Å models
- Current refinement rely on multiple templates
- Physics-based refinement does not work well

What remain difficult?

IV. We cannot model the structure of membrane proteins



1, the major challenge is the lack of homology templates
2, model interactions with membrane

Acknowledgements

Protein structure prediction

- Dong Xu
- Jouko Virtanen
- Baoji He
- Golam Mortuza
- Wei Zheng

Structure-based function prediction

- Ambrosh Roy
- Jianyi Yang
- Wallace Chan
- Chengxin Zhang

Protein-protein interaction

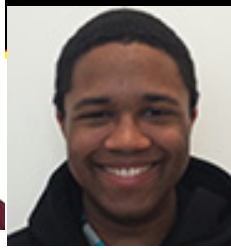
- Srayanta Mukherjee
- Aysam Guerler
- Brandon Govindarajoo

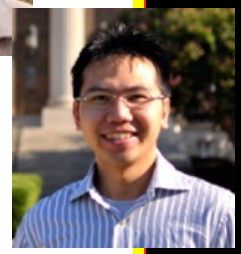
SNP mutation and cancer

- Lijun Quan
- Gulzhan Raiymbek

Protein design

- David Shultis
- Jeffrey Brender
- Jarrett Johnson
- Chengyuan Wang
- Jiong Li
- Xiaoqiong Wei
- Dani Setiawan
- Yang Cao





System Admin

- Jonathan Poisson

Funding support:

- NIH R01 GM083107
- NIH R01 GM116960
- NIH SBIR GM119985
- NSF DBI 1564756

Thank you!