Content

1. Bioinformatics databases
2. Sequence alignment and database searching
3. Phylogenic tree and multiple sequence alignment
4. Protein structure alignment
5. Protein secondary structure prediction
6. Protein tertiary structure prediction

Protein tertiary structure prediction

杨建益

Email: yangjy@nankai.edu.cn

Webpage: http://yanglab.nankai.edu.cn/

Course: http://yanglab.nankai.edu.cn/teaching/bioinformatics/

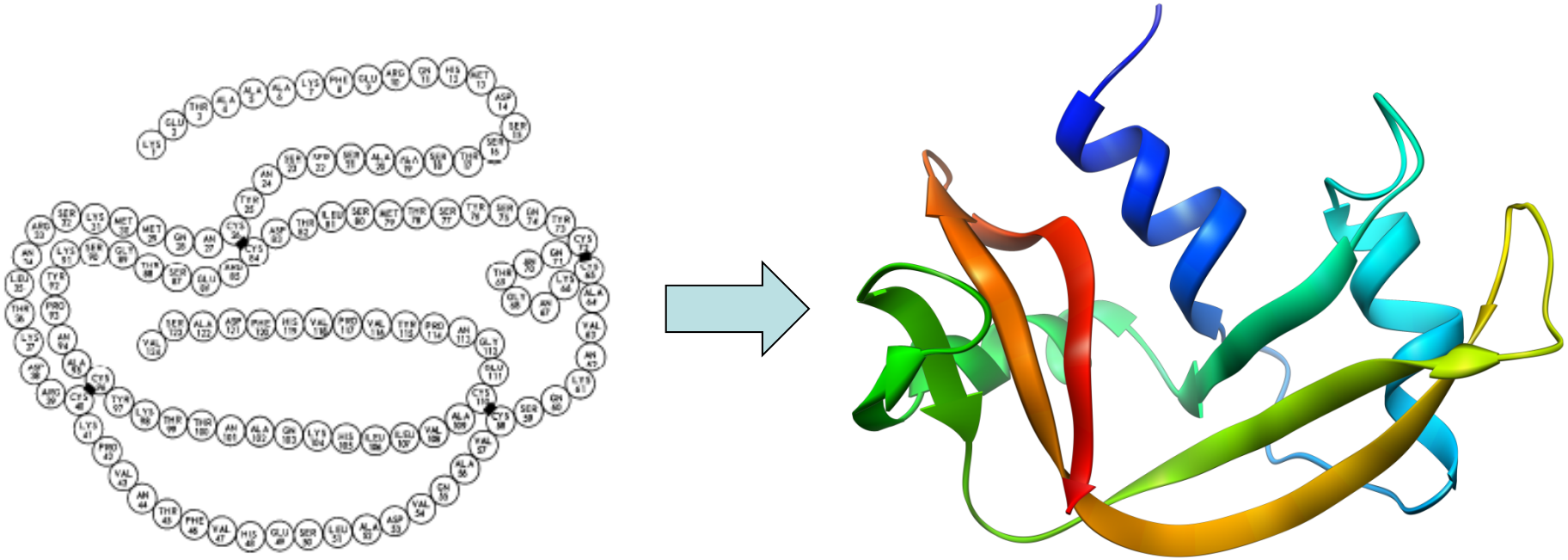
Office: 数学科学学院，419室

Content

- Ab-initio folding
- Homology modeling
  a. Comparative modeling(CM)
  b. Threading (fold recognition)
- Composite approach (I-TASSER)
- Fragment assembly (Rosetta, QUARK)
- CASP

The ground for PSP

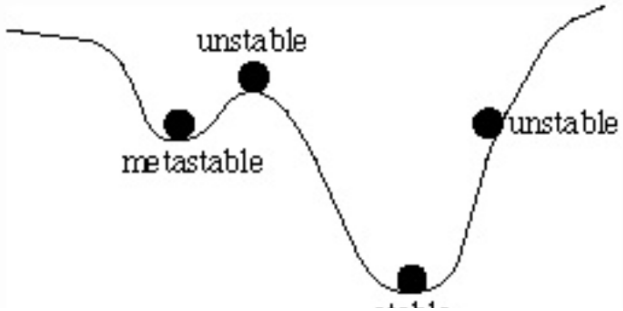
Anfinsen's dogma (1972 Nobel Prize in Chemistry): the native structure is determined only by the protein's amino acid sequence.



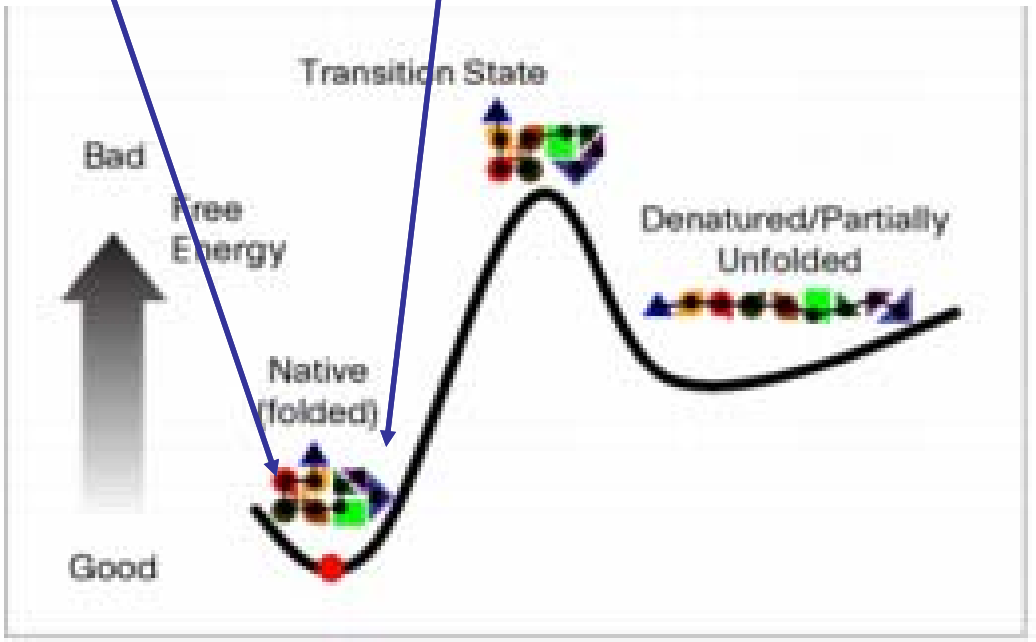
Ribonuclease A

Ab-initio folding

- decide protein structure based on physical principle, i.e. find native structure by searching for structures with the lowest free energy



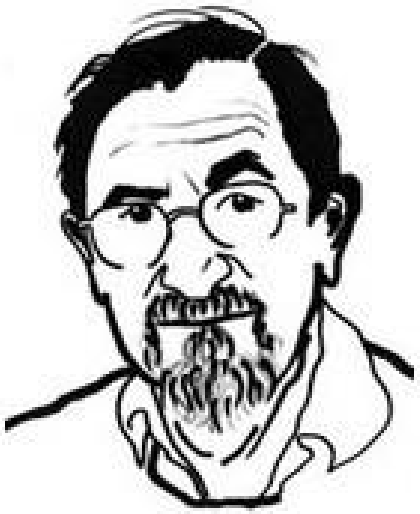
Rule of physics



Ab-initio folding

- Two factors that matter
  a. force field
  b. search engine

The Nobel Prize in Chemistry 2013



© Nobel Media AB

Martin Karplus

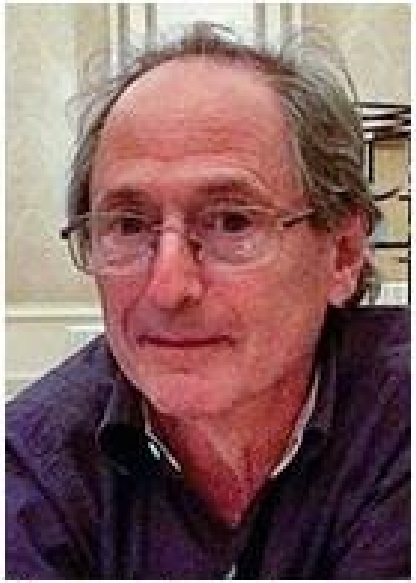


Photo: Keilana via Wikimedia Commons

Michael Levitt

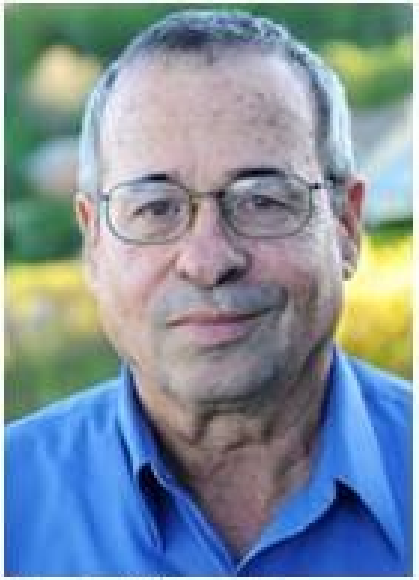


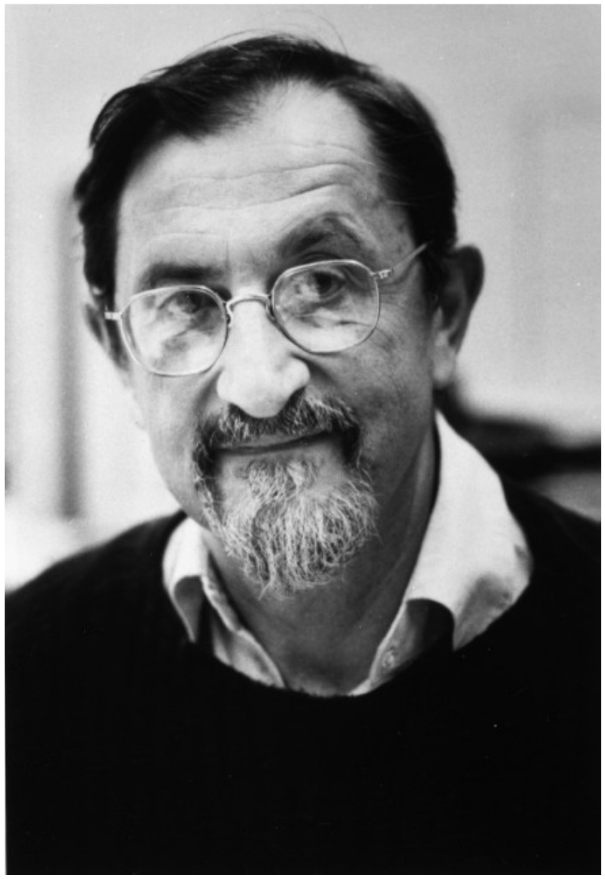
Photo: Wikimedia Commons

Arieh Warshel

for "the development of multiscale models for complex chemical systems"

CHARMM (Martin Karplus):

Chemistry at Harvard Macromolecular Mechanics (CHARMM)

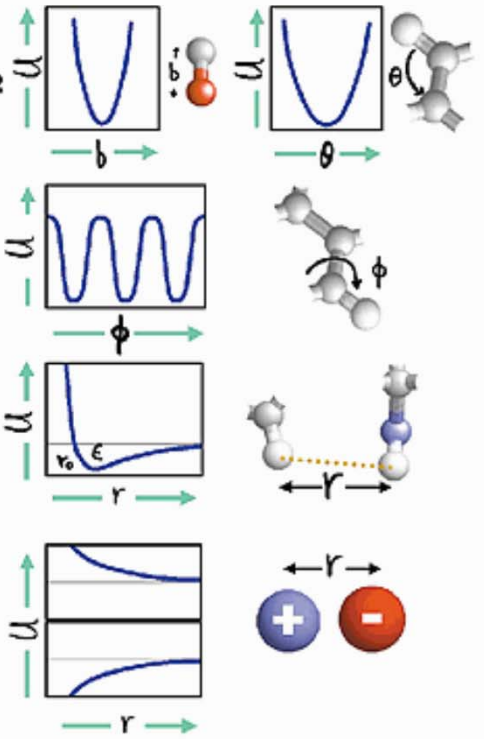


$$U = \sum_{\text{All Bonds}} \tfrac{1}{2}K_b(b-b_o)^2 + \sum_{\text{All Angles}} \tfrac{1}{2}K_\theta(\theta-\theta_o)^2$$

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

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

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



Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983). "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations". J Comp Chem. **4** (2): 187–217.

CHARMM (Martin Karplus):

CHARMM: A program for macromolecular energy, minimization, and dynamics calculations

BR Brooks, RE Bruccoleri, BD Olafson... - Journal of ..., 1983 - Wiley Online Library

Abstract CHARMM (Chemistry at HARvard Macromolecular Mechanics) is a highly flexible computer program which uses empirical energy functions to model macromolecular systems. The program can read or model build structures, energy minimize them by first-or second-derivative techniques, perform a normal mode or molecular dynamics simulation, and analyze the structural, equilibrium, and dynamic properties determined in these calculations. The operations that CHARMM can perform are described, and some ...

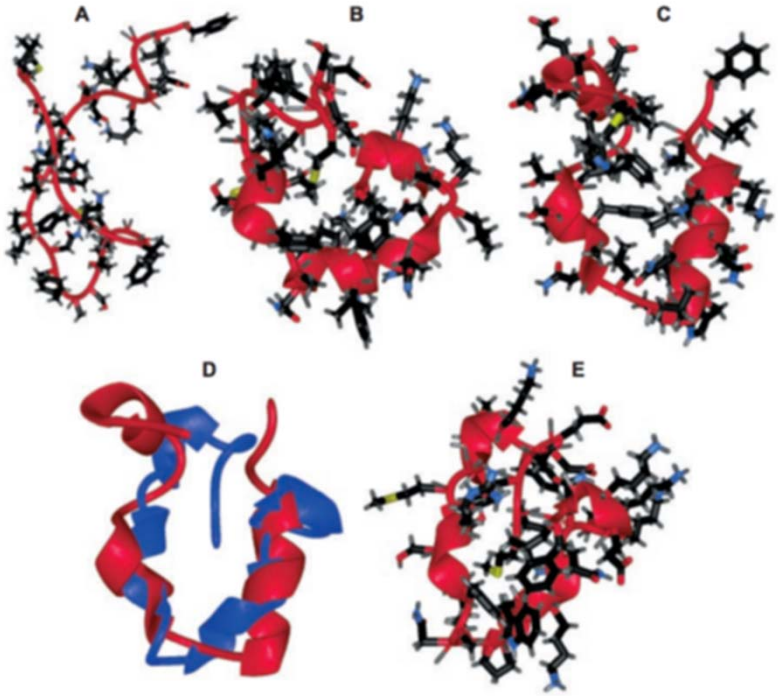
☆ 99 被引用次数：5035 相关文章 所有 6 个版本

AMBER

Assisted Model Building with Energy Refinement (AMBER)

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

1944-2001



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

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

AMBER

A second generation force field for the simulation of proteins, nucleic acids, and organic molecules

WD Cornell, P Cieplak, CI Bayly, IR Gould... - Journal of the ..., 1995 - ACS Publications

We present the derivation of a new molecular mechanical force field for simulating the structures, conformational energies, and interaction energies of proteins, nucleic acids, and many related organic molecules in condensed phases. This effective two-body force field is the successor to the Weiner et al. force field and was developed with some of the same philosophies, such as the use of a simple diagonal potential function and electrostatic potential fit atom centered charges. The need for a 10-12 function for representing hydrogen ...

☆ 99 被引用次数：12386 相关文章 所有 29 个版本

GROMACS

GROningen MAchine for Chemical Simulations (GROMACS)

GROMACS: a message-passing parallel molecular dynamics implementation

HJC Berendsen, D van der Spoel... - Computer Physics ..., 1995 - Elsevier

A parallel message-passing implementation of a molecular dynamics (MD) program that is useful for bio (macro) molecules in aqueous environment is described. The software has been developed for a custom-designed 32-processor ring GROMACS (GROningen MAchine for Chemical Simulation) with communication to and from left and right neighbours, but can run on any parallel system onto which aa ring of processors can be mapped and which supports PVM-like block send and receive calls. The GROMACS software consists of a ...

☆ 99 被引用次数：5427 相关文章 所有 16 个版本

Defects of *ab-initio* folding

- Only works for small proteins (i.e., < 200 AAs)
- Slow
- Accuracy is low

Content

- Ab-initio folding
- Homology modeling
  a. Comparative modeling(CM)
  b. Threading (fold recognition)
- Composite approach (I-TASSER)
- Fragment assembly (Rosetta, QUARK)
- CASP

Homology modeling

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

A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme

WJ Browne, ACT North, DC Phillips, K Brew... - Journal of molecular ..., 1969 - Elsevier

Bovine α-lactalbumin and hen egg-white lysozyme have closely similar amino acid sequences. A model of α-lactalbumin has been constructed on the basis of the main chain conformation established for lysozyme. The side chain interactions of lysozyme are listed (Table 2) and the consequences of the side chain replacements in α-lactalbumin examined. Changes in internal side chains are generally interrelated in a convincing manner, suggesting that the model is largely correct, but there are some regions where it has not ...

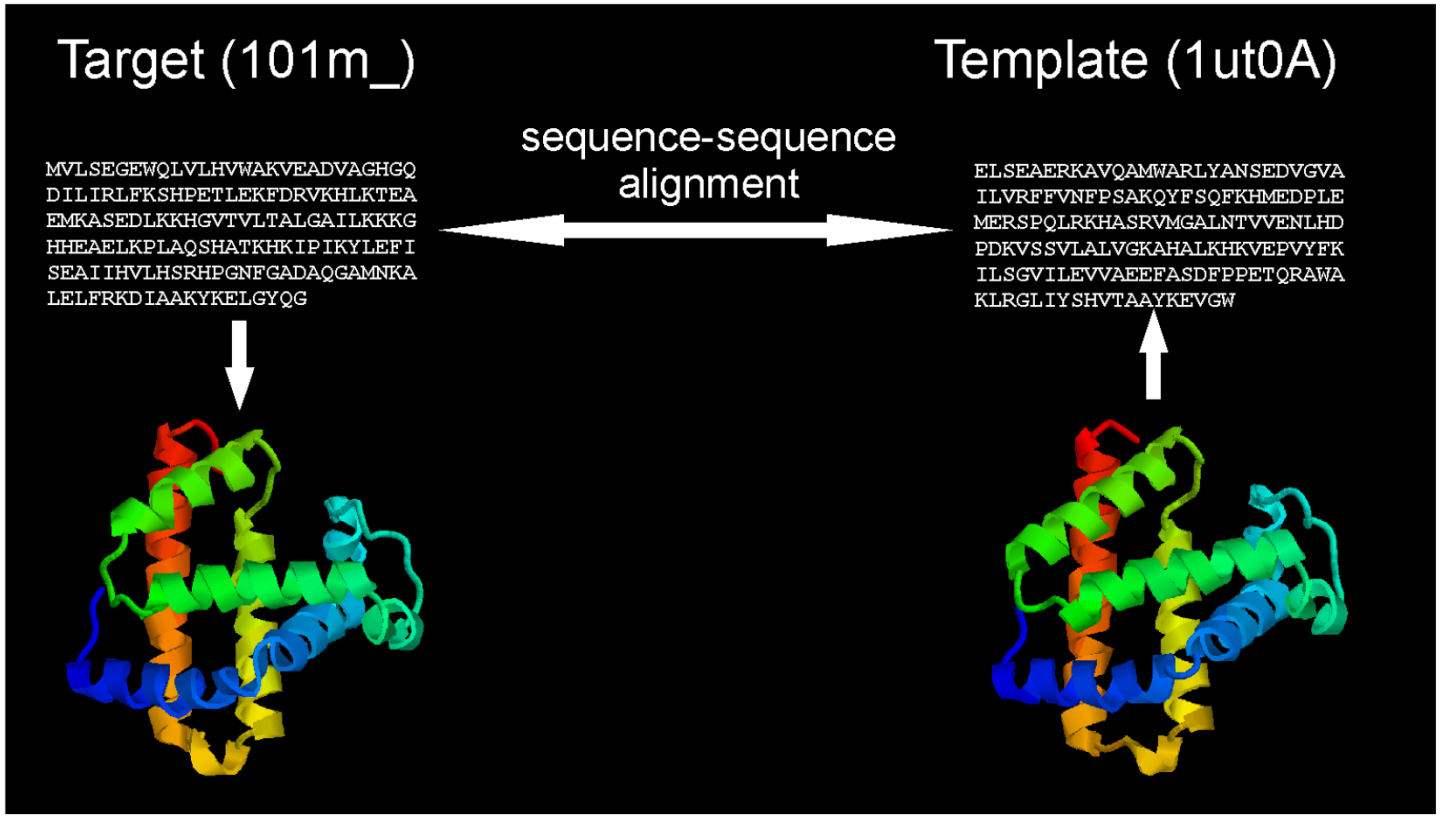
☆ 99 被引用次数：539 相关文章 所有 8 个版本

Homology modeling

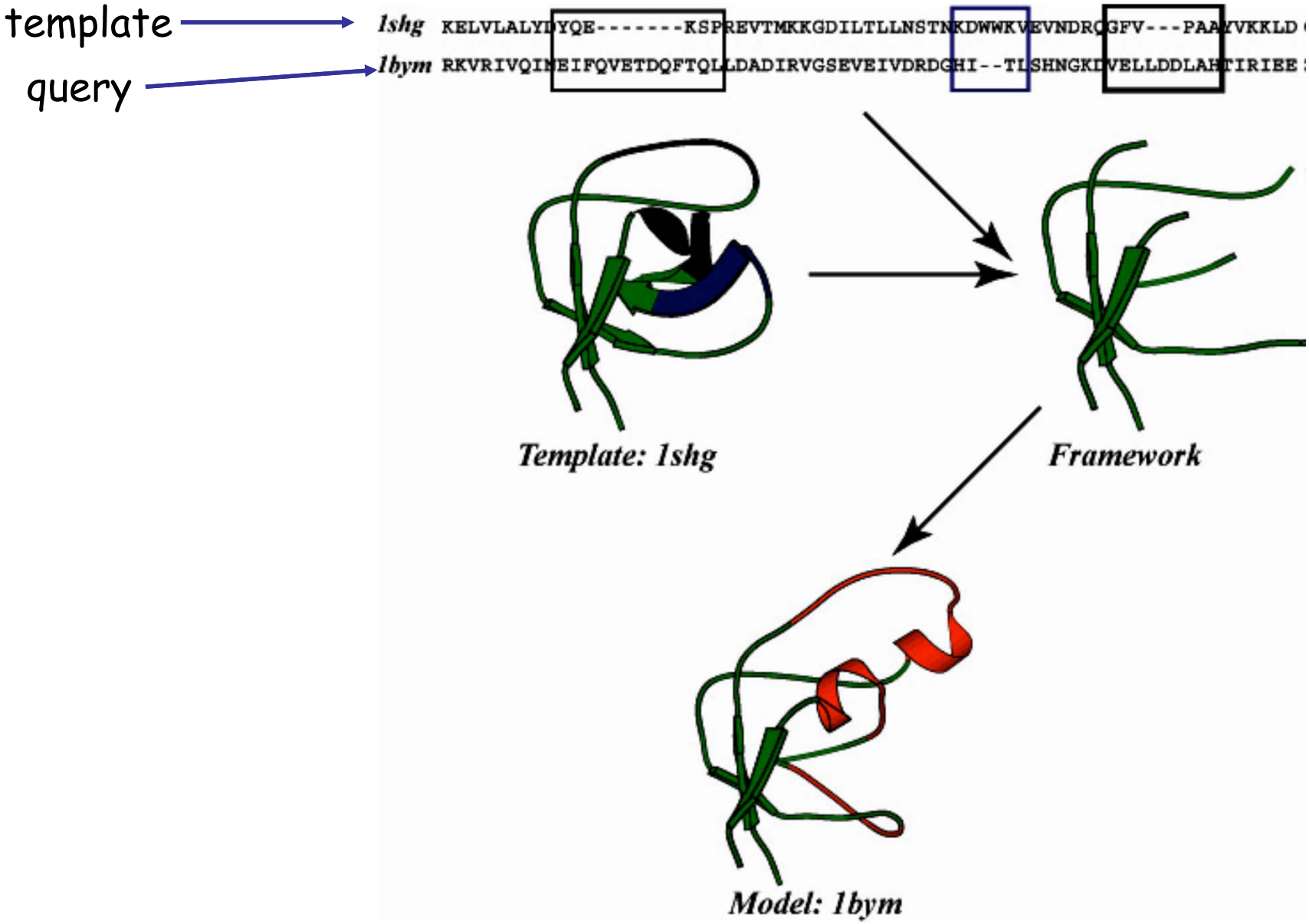
Step 1: query-templates alignment (the key)

Step 2: Obtain target coordinates from template based on alignments

Step 3: Generate models (add missing regions & side chains)



An example of homology modeling



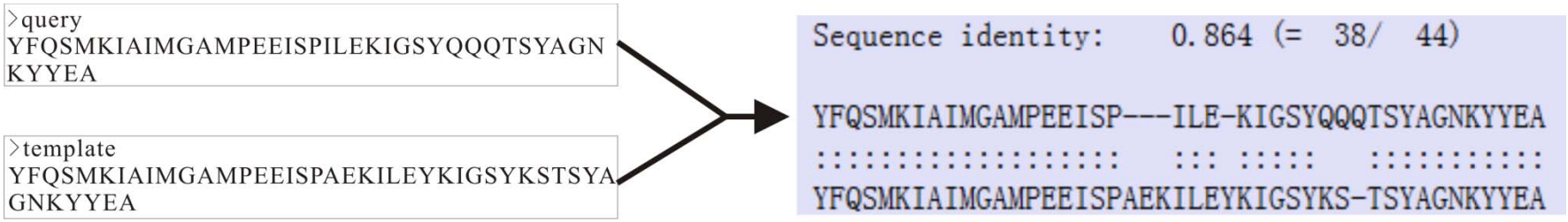
How to build query-template alignment?

- Sequence-sequence alignment (NW-DP)
- Sequence-profile alignment (PSI-BLAST, HMMER)
- Profile-profile alignment (FFAS, HHsearch)
- Threading (fold recognition)

Comparative modeling

Easy: similar sequences -> similar structures (>50%) sequence-sequence alignment

NW-DP, FASTA (1988), BLAST (1990)



Comparative modeling

Medium: similar profiles -> similar structures (>25%) profile-sequence/profile alignment

PSI-BLAST(1997), HHMER(1995), HHsearch (2005), FFAS(2003)

Query profile

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	S	4	-3	-2	-3	-2	-2	-2	-2	-3	-3	-4	-2	-3	-4	-3	4	3	-4	-3	-2
2	R	-1	4	0	-3	-5	0	-2	-2	-3	-5	-5	6	-4	-5	-2	-1	0	-5	-4	-4
3	R	2	3	-2	-2	-4	-1	0	-3	-3	-2	-3	4	-2	-5	2	0	1	-5	-4	-2
4	S	1	-2	-2	-3	-4	-1	-2	-2	-2	-3	-3	3	-3	-4	5	0	2	-5	-4	-1
5	A	2	-1	-1	-2	-3	-1	-2	-2	-2	-1	-2	0	-1	-4	4	2	0	-5	-4	0
6	S	1	-1	1	2	-3	-1	-1	-2	0	-3	-3	-1	-3	-4	2	3	3	-5	-4	-3
7	H	-4	-2	-2	-5	-2	-1	-3	2	10	-4	-3	-1	-5	-5	-6	-2	-4	-6	0	-6
8	P	-1	-6	-6	-6	-7	-6	-5	-5	-4	-7	-5	-4	-7	-4	8	-2	-3	-8	-7	-7
9	T	-2	-1	-2	-3	-5	-3	-4	-3	-3	-5	-6	2	-5	-6	6	3	3	-6	-5	-4
10	Y	-3	-6	-6	-7	-4	-5	-6	-7	-1	-2	-2	-4	2	0	-6	-4	1	2	8	3
11	S	0	0	0	0	0	0	0	-2	-2	1	0	0	0	0	0	2	0	-4	-1	-1
12	E	0	-1	0	4	-5	1	4	-3	-2	-3	-4	0	-3	-4	-3	-2	-2	2	-4	0

Template profile

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	D	-4	-3	2	6	-6	-1	5	-4	-3	-6	-6	-2	-5	-6	-4	-2	-3	-6	-5	-5
2	R	-2	3	1	1	-3	1	1	-2	6	-2	-4	0	-2	-3	-2	0	0	-4	0	-2
3	V	0	-3	-4	-5	-3	-3	-4	-5	0	-2	-2	-4	-3	3	-4	0	0	0	7	2
4	P	-5	-2	-1	8	-7	-3	1	-5	-2	-6	-5	-4	-6	-7	0	-3	-3	-7	-6	-3
5	A	3	-5	-4	-4	-1	-4	-4	-4	-3	1	-2	-4	-2	0	-5	-3	-2	-2	3	5
6	L	-2	-5	-6	-6	1	-5	-5	-6	-6	5	2	-5	-1	-3	-5	-4	-3	-5	-4	5
7	V	-3	-6	-6	-7	-4	-6	-6	-7	-7	5	-2	-6	-2	-4	-6	-5	-4	-6	-5	6
8	I	-3	-6	-6	-6	-2	-6	-6	-7	-6	6	-1	-6	-2	-4	-6	-5	-3	-6	-5	5
9	G	-4	-6	-4	-5	-7	-6	-6	8	-6	-8	-8	-6	-7	-7	-6	-4	-6	-7	-7	-7
10	S	2	-4	-3	-4	-2	-4	-4	0	-4	-5	-4	-4	-5	-4	-4	6	1	-6	-5	-3
11	G	-4	-6	-4	-5	-7	-6	-6	8	-6	-8	-8	-6	-7	-7	-6	-4	-6	-7	-7	-7
12	Y	2	-4	-4	-4	2	-3	-3	-3	-3	0	-1	-4	-1	4	2	0	-1	1	3	0

Threading (fold recognition)

Hard: dissimilar profiles -> similar structures (< 25%, twilight zone) sequence-structure alignment (Threading, fold-recognition, remote-homology detection)

A method to identify protein sequences that fold into a known three-dimensional structure

JU Bowie, R Luthy, D Eisenberg - Science, 1991 - science.sciencemag.org

The inverse protein folding problem, the problem of finding which amino acid sequences fold into a known three-dimensional (3D) structure, can be effectively attacked by finding sequences that are most compatible with the environments of the residues in the 3D structure. The environments are described by:(i) the area of the residue buri and inaccessible to solvent;(ii) the fraction of side-chain area that is covered (O and N); and (iii) the local secondary structure. Examples of this 3D profile

☆ 99 被引用次数：2701 相关文章 所有 17 个版本

Threading (fold recognition)

A new approach to protein fold recognition

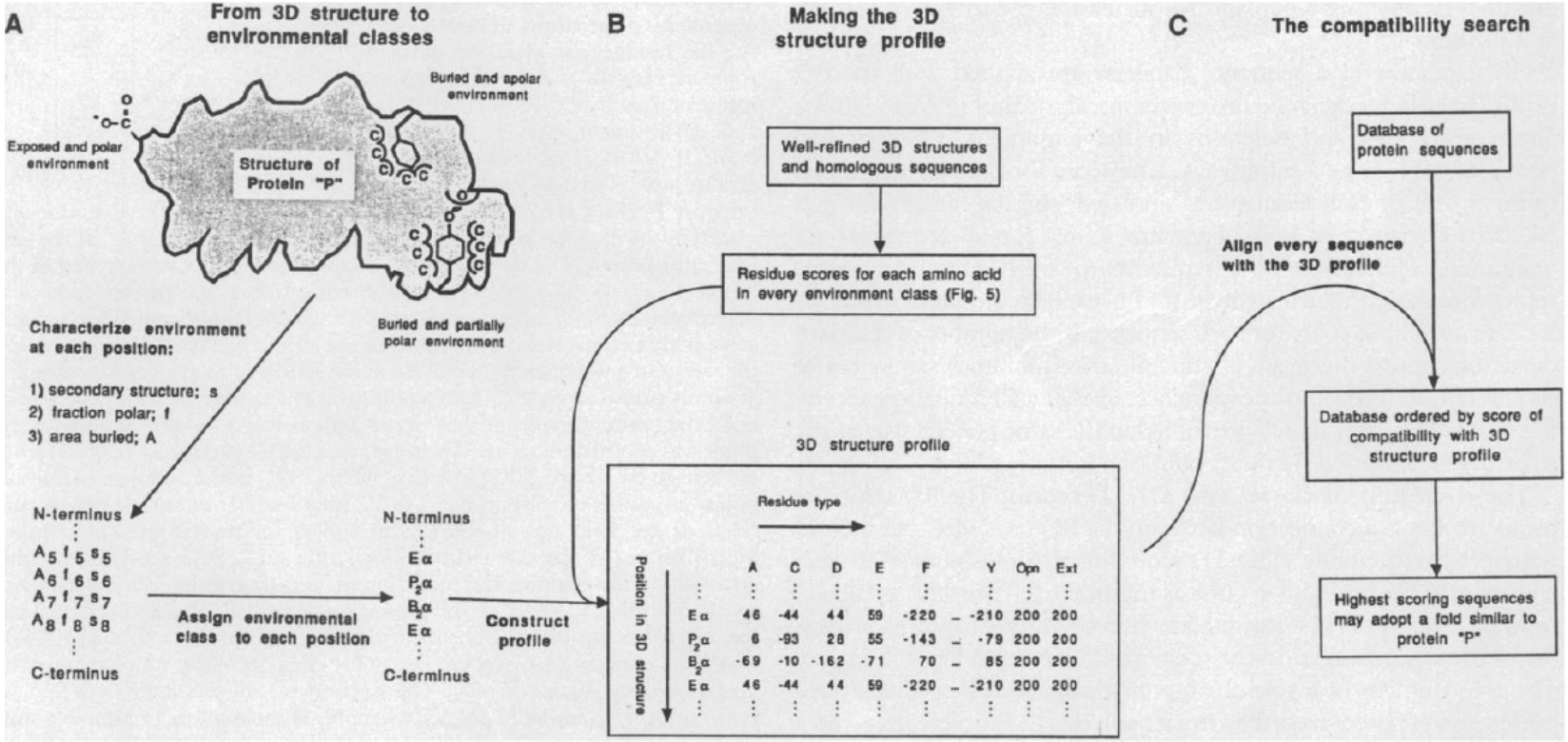
DT Jones, WR Taylort, JM Thornton - Nature, 1992 - nature.com

THE prediction of protein tertiary structure from sequence using molecular energy calculations has not yet been successful; an alternative strategy of recognizing known motifs 1 or folds 2–4 in sequences looks more promising. We present here a new approach to fold recognition, whereby sequences are fitted directly onto the backbone coordinates of known protein structures. Our method for protein fold recognition involves automatic modelling of protein structures using a given sequence, and is based on the frameworks of known protein …

☆ 99 被引用次数：1430 相关文章 所有 7 个版本

Threading (fold recognition)

David Eisenberg *Science, 1991.*



Threading-MUSTER

$$\begin{aligned}\mathrm{Score}(i,j) &= \sum_{k=1}^{20}\left(Pc_{\mathrm{q}}(i,k)+Pd_{\mathrm{q}}(i,k)L_{\mathrm{t}}(j,k)/2+c_1\delta(s_{\mathrm{q}}(i),s_{\mathrm{t}}(j))\right.\\ &+c_2\sum_{k=1}^{20}Ps_{\mathrm{t}}(j,k)L_{\mathrm{q}}(i,k)+c_3\left(1-2\left|SA_{\mathrm{q}}(i)-SA_{\mathrm{t}}(j)\right|\right),\\ &+c_4\left(1-2\left|\varphi_{\mathrm{q}}(i)-\varphi_{\mathrm{t}}(j)\right|\right)+c_5\left(1-2\left|\phi_{\mathrm{q}}(i)-\phi_{\mathrm{t}}(j)\right|\right)\\ &\left.+c_6M(AA_{\mathrm{q}}(i),AA_{\mathrm{t}}(j))+c_7\right.\end{aligned}$$

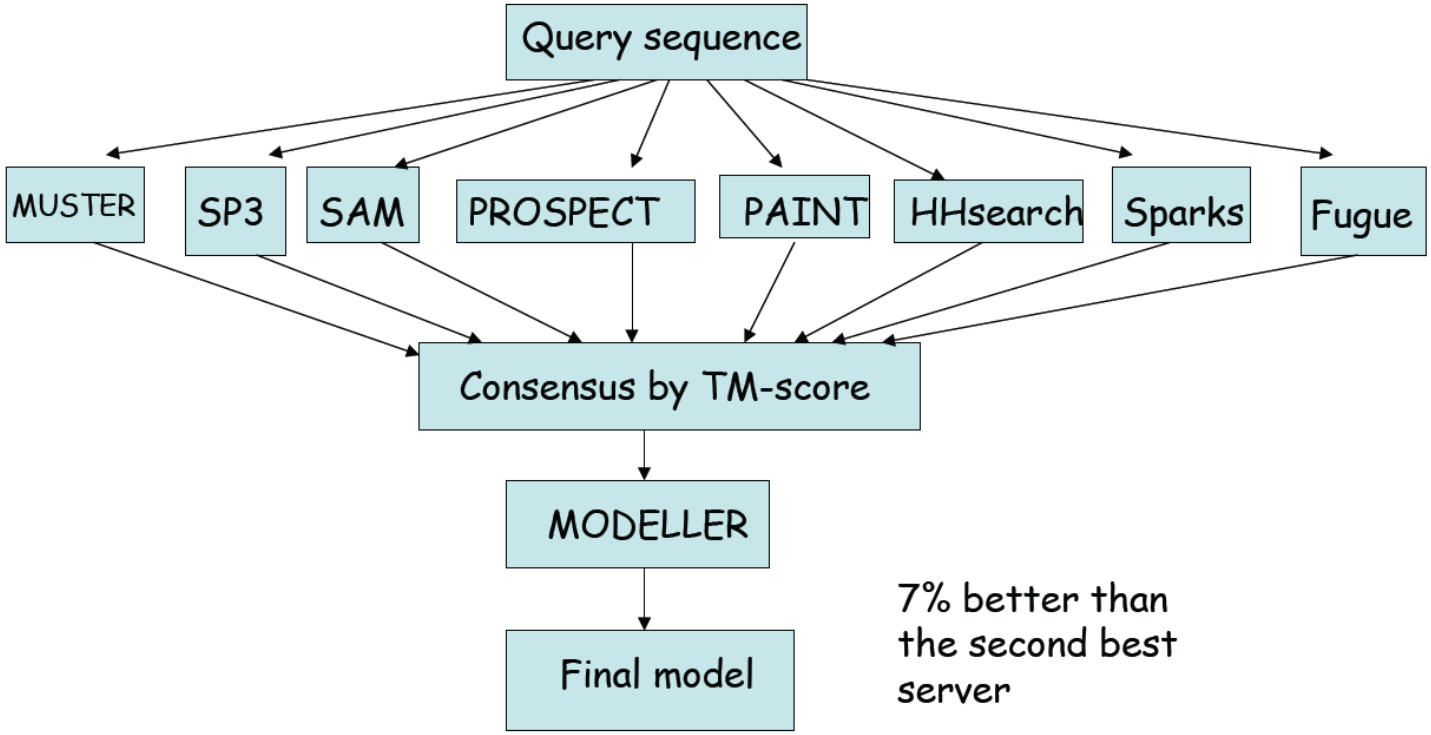
MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information

S Wu, Y Zhang - Proteins: Structure, Function, and ..., 2008 - Wiley Online Library

We develop a new threading algorithm MUSTER by extending the previous sequence profile–profile alignment method, PPA. It combines various sequence and structure information into single-body terms which can be conveniently used in dynamic programming search:(1) sequence profiles;(2) secondary structures;(3) structure fragment profiles;(4) solvent accessibility;(5) dihedral torsion angles;(6) hydrophobic scoring matrix. The balance of the weighting parameters is optimized by a grading search based on the average TM ...

☆ 99 被引用次数：273 相关文章 所有 10 个版本

Threading-LOMETS



7% better than the second best server

LOMETS: a local meta-threading-server for protein structure prediction
S Wu, Y Zhang - Nucleic acids research, 2007 - academic.oup.com
We developed LOMETS, a local threading meta-server, for quick and automated predictions of protein tertiary structures and spatial constraints. Nine state-of-the-art threading programs are installed and run in a local computer cluster, which ensure the quick generation of initial threading alignments compared with traditional remote-server-based meta-servers. Consensus models are generated from the top predictions of the component-threading servers, which are at least 7% more accurate than the best individual servers based on TM ...
☆ 99 被引用次数：552 相关文章 所有 17 个版本

How to build full-length models quickly?

MODELLER

Widely used for quick build of full-length structure models based on sequence-template alignment

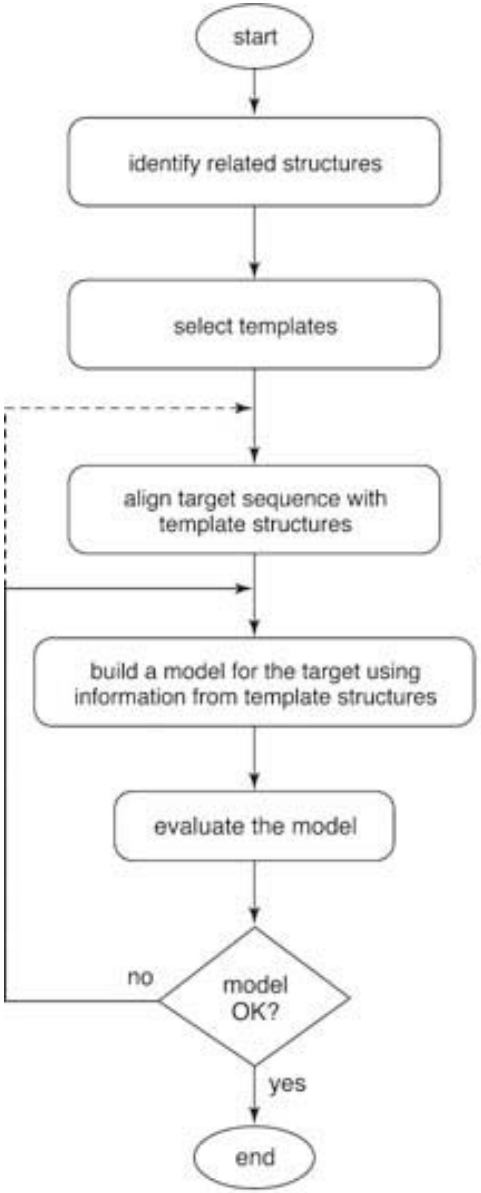
Comparative protein modelling by satisfaction of spatial restraints

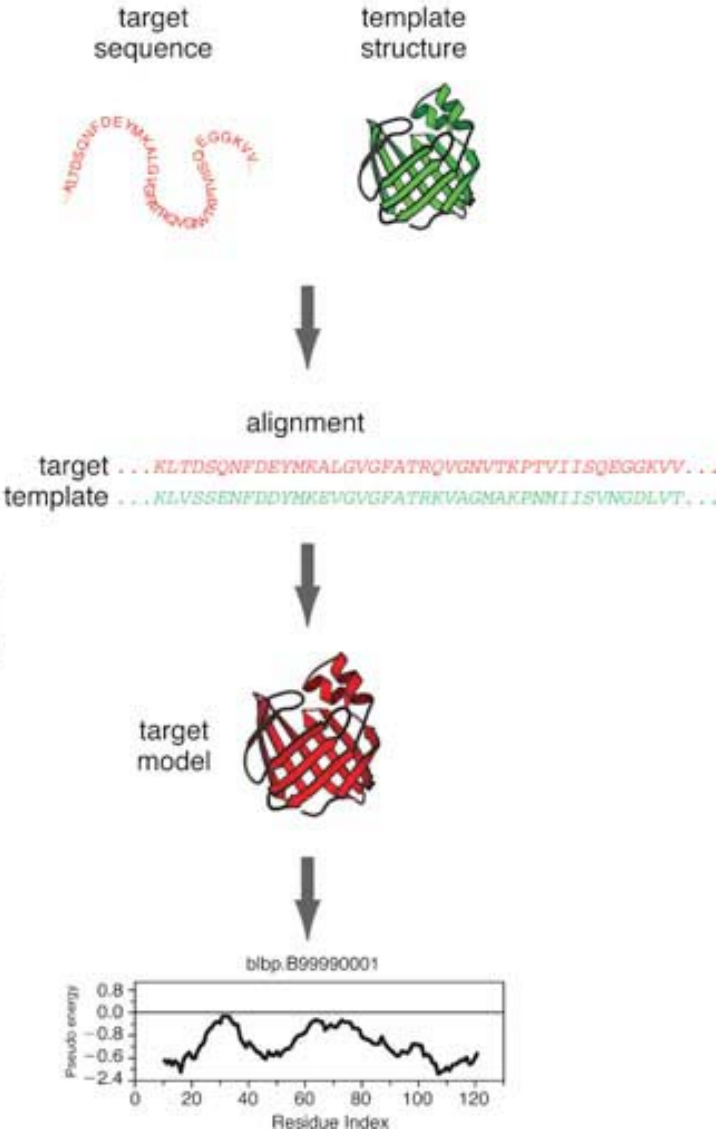
A Šali, TL Blundell - Journal of molecular biology, 1993 - Elsevier

We describe a comparative protein modelling method designed to find the most probable structure for a sequence given its alignment with related structures. The three dimensional (3D) model is obtained by optimally satisfying spatial restraints derived from the alignment ...

☆ 99 被引用次数：10127 相关文章 所有 18 个版本

MODELLER


start
identify related structures
select templates
align target sequence with template structures
build a model for the target using information from template structures
evaluate the model
model OK?
no
yes
end


target sequence
template structure
alignment
target ...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
template ...KLVSSENFDDYMKEVGVGFATRKVAGMAKPNMIISVNGDLVT...
target model
blbp.B99990001
Pseudo energy
0.8
0.0
−0.8
−0.6
−2.4
0
20
40
60
80
100
120
Residue Index

Content

- Ab-initio folding
- Homology modeling
  a. Comparative modeling(CM)
  b. Threading (fold recognition)
- Composite approach (I-TASSER)
- Fragment assembly (Rosetta, QUARK)
- CASP

I-TASSER: a composite approach

[HTML] I-TASSER server for protein 3D structure prediction

Y Zhang - BMC bioinformatics, 2008 - bmcbioinformatics.biomedcentral ...

Prediction of 3-dimensional protein structures from amino acid sequences represents one of the most important problems in computational structural biology. The community-wide Critical Assessment of Structure Prediction (CASP) experiments have been designed to ...

☆ 99 被引用次数：2895 相关文章 所有 22 个版本 »

I-TASSER: a unified platform for automated protein structure and function prediction

A Roy, A Kucukural, Y Zhang - Nature protocols, 2010 - nature.com

The iterative threading assembly refinement (I-TASSER) server is an integrated platform for automated protein structure and function prediction based on the sequence-to-structure-to-function paradigm. Starting from an amino acid sequence, I-TASSER first generates three ...

☆ 99 被引用次数：3392 相关文章 所有 15 个版本

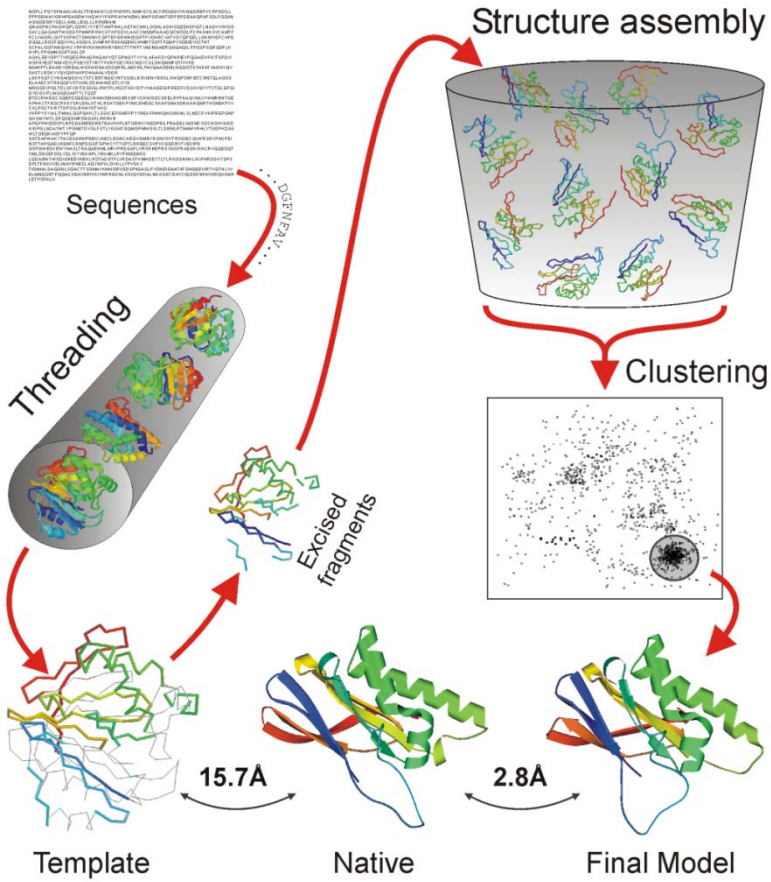
The I-TASSER Suite: protein structure and function prediction

J Yang, R Yan, A Roy, D Xu, J Poisson, Y Zhang - Nature methods, 2015 - nature.com

Assignment of structure and function to all genes and gene products (such as proteins) of all organisms represents a major challenge in this postgenomic era. Here we present the I-TASSER Suite (http://zhanglab. ccmb. med. umich. edu/I-TASSER/download/), a stand-alone ...

☆ 99 被引用次数：1288 相关文章 所有 8 个版本

TASSER (Threading ASSEmbly Refinement)



Automated structure prediction of weakly homologous proteins on a genomic scale

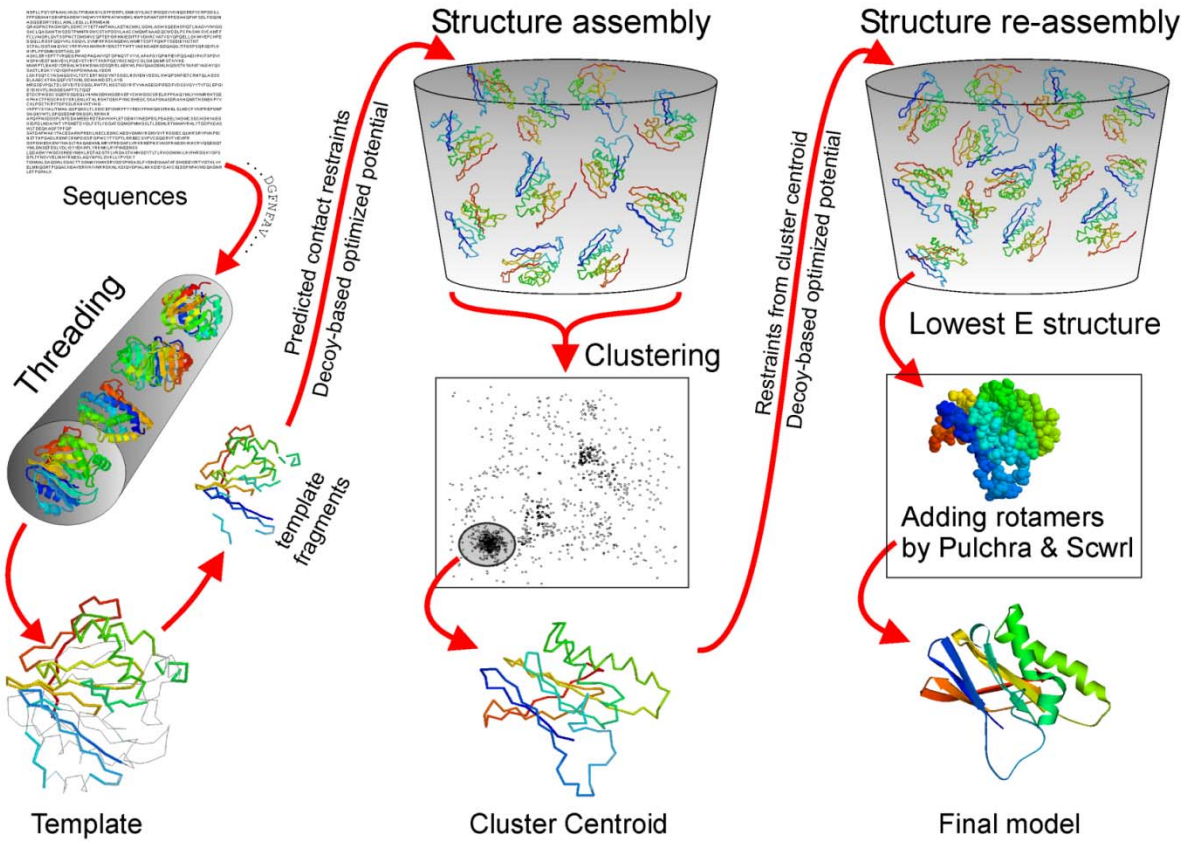
Y Zhang, J Skolnick - ... of the National Academy of Sciences ..., 2004 - National Acad Sciences

We have developed tasser, a hierarchical approach to protein structure prediction that consists of template identification by threading, followed by tertiary structure assembly via the rearrangement of continuous template fragments guided by an optimized C α and side ...

☆ 99 被引用次数：343 相关文章 所有 18 个版本

I-TASSER algorithm

Iterative Threading ASSEmbly Refinement



[HTML] Ab initio modeling of small proteins by iterative TASSER simulations

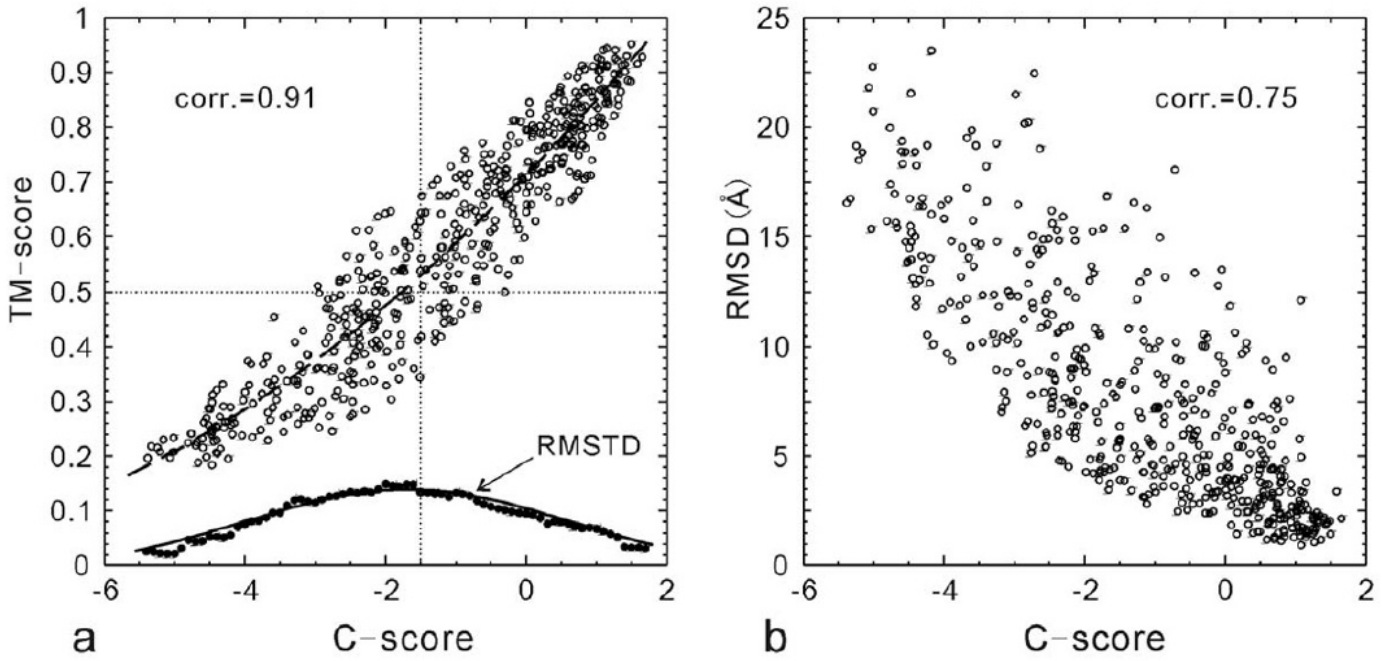
S Wu, J Skolnick, Y Zhang - BMC biology, 2007 - bmcbiol.biomedcentral.com

Predicting 3-dimensional protein structures from amino-acid sequences is an important unsolved problem in computational structural biology. The problem becomes relatively easier if close homologous proteins have been solved, as high-resolution models can be built by aligning target sequences to the solved homologous structures. However, for sequences without similar folds in the Protein Data Bank (PDB) library, the models have to be predicted from scratch. Progress in the ab initio structure modeling is slow. The aim of this ...

☆ 99 被引用次数：469 相关文章 所有 23 个版本 »

I-TASSER server

$$\text{C-score} = \ln\left(\frac{M}{M_{tot}} \cdot \frac{1}{\langle \text{RMSD} \rangle} \cdot \frac{\prod_{i=1}^{4} Z(i)}{\prod_{i=1}^{4} Z_0(i)} \right)$$



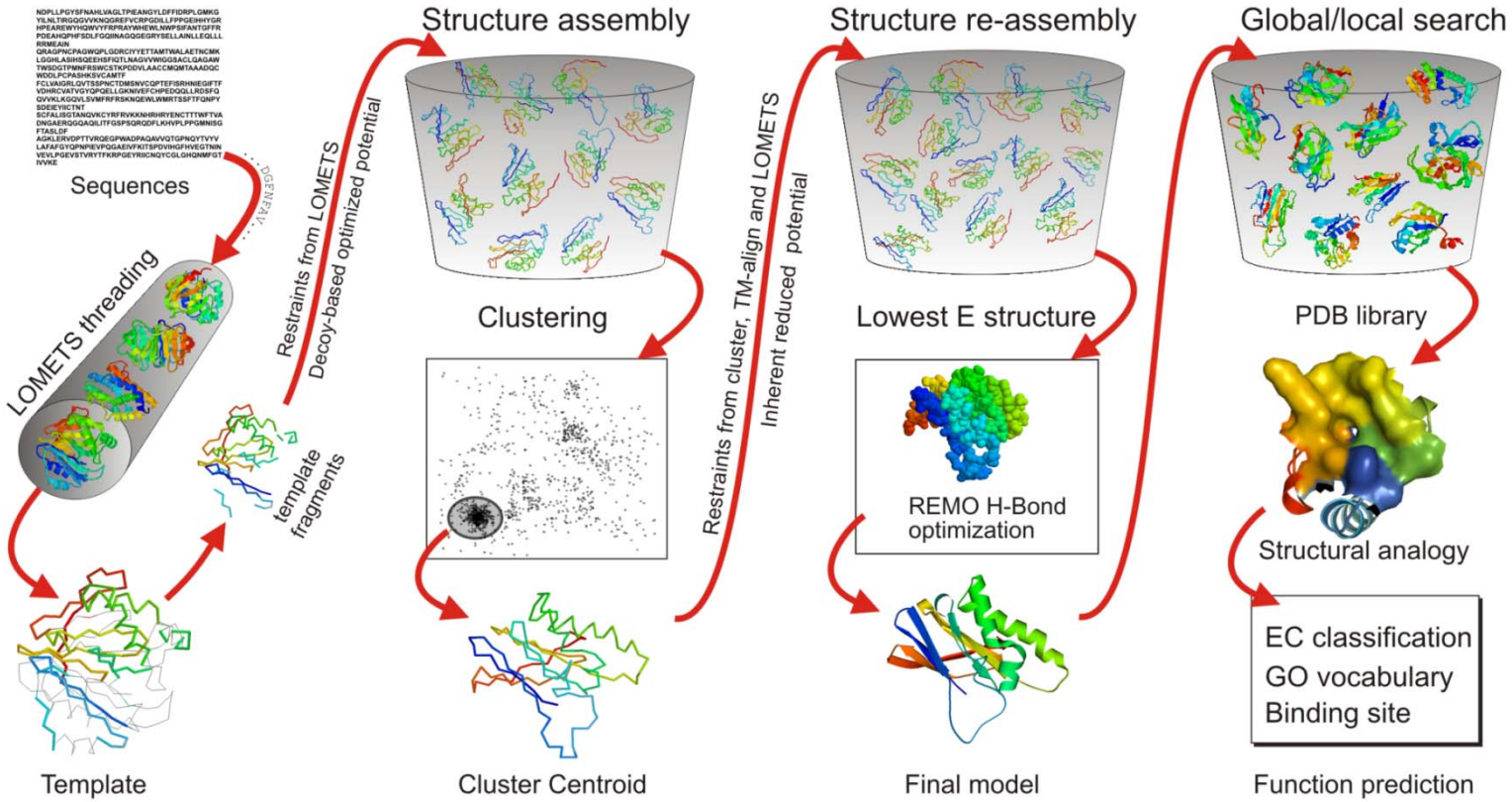
[HTML] I-TASSER server for protein 3D structure prediction

Y Zhang - BMC bioinformatics, 2008 - bmcbioinformatics.biomedcentral ...

Prediction of 3-dimensional protein structures from amino acid sequences represents one of the most important problems in computational structural biology. The community-wide Critical Assessment of Structure Prediction (CASP) experiments have been designed to ...

☆ 99 被引用次数：2895 相关文章 所有 22 个版本 »

I-TASSER



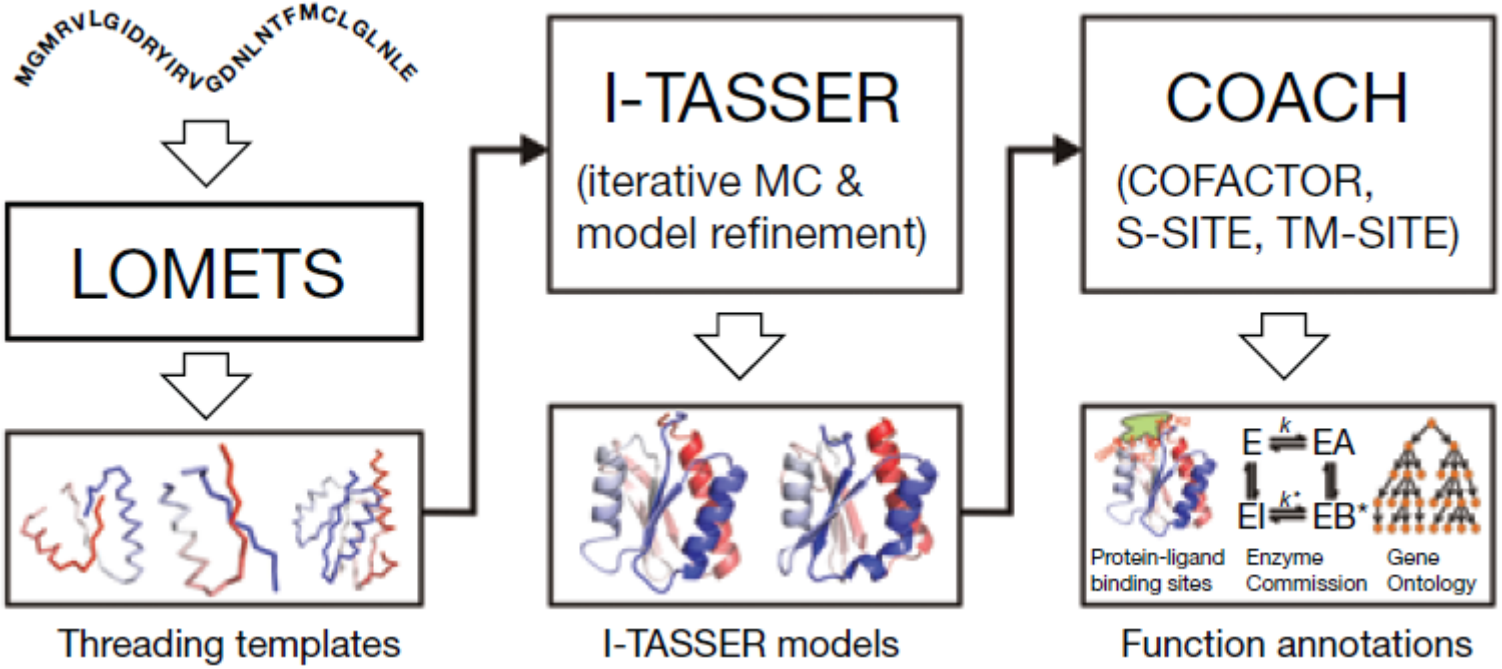
I-TASSER: a unified platform for automated protein structure and function prediction

A Roy, A Kucukural, Y Zhang - Nature protocols, 2010 - nature.com

The iterative threading assembly refinement (I-TASSER) server is an integrated platform for automated protein structure and function prediction based on the sequence-to-structure-to-function paradigm. Starting from an amino acid sequence, I-TASSER first generates three ...

☆ 99 被引用次数：3392 相关文章 所有 15 个版本

I-TASSER Suite



The I-TASSER Suite: protein structure and function prediction

J Yang, R Yan, A Roy, D Xu, J Poisson, Y Zhang - Nature methods, 2015 - nature.com

Assignment of structure and function to all genes and gene products (such as proteins) of all organisms represents a major challenge in this postgenomic era. Here we present the I-TASSER Suite (http://zhanglab. ccmb. med. umich. edu/I-TASSER/download/), a stand-alone ...

☆ 99 被引用次数：1288 相关文章 所有 8 个版本

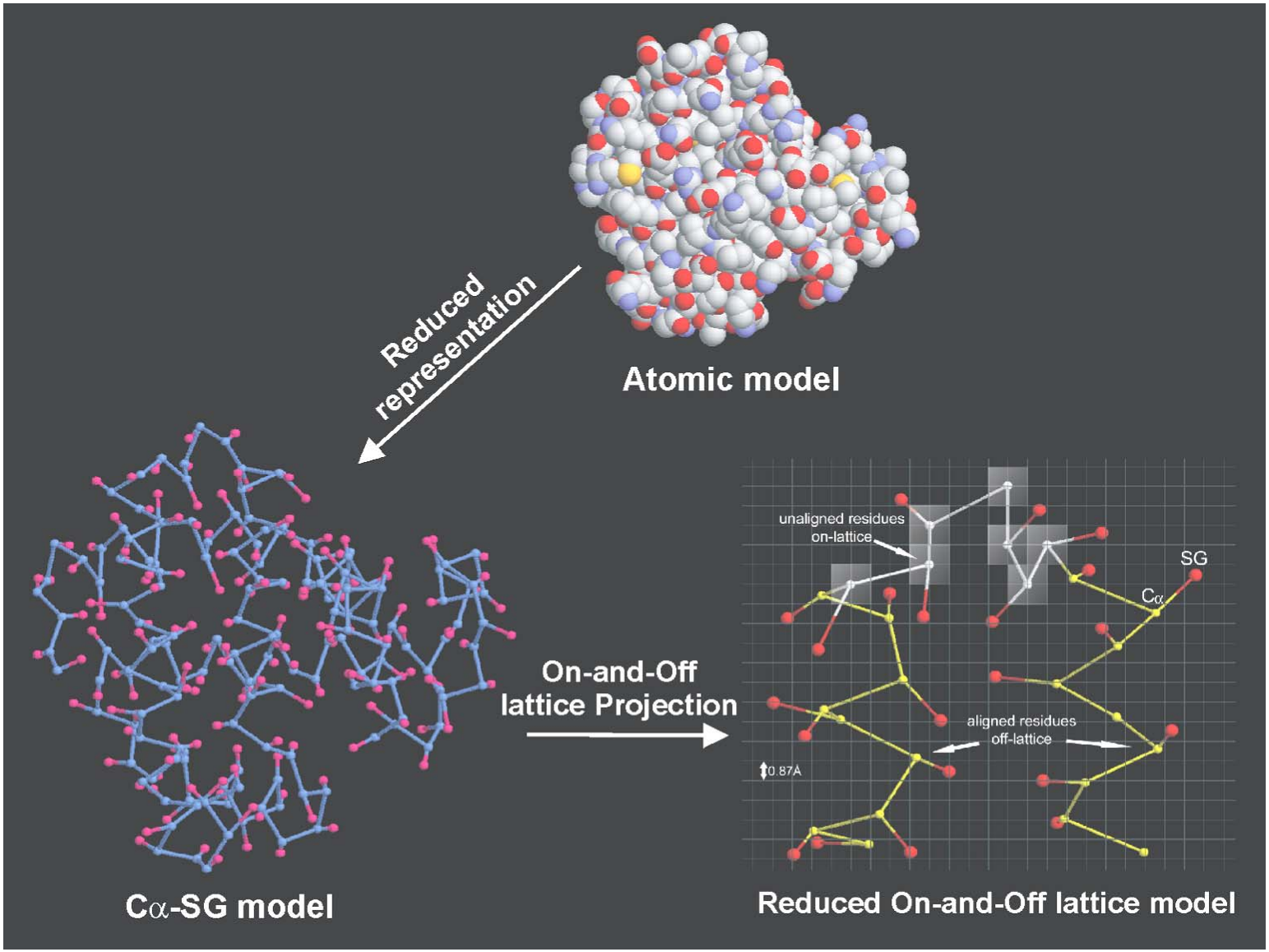
I-TASSER force field

Three sources (25 terms):

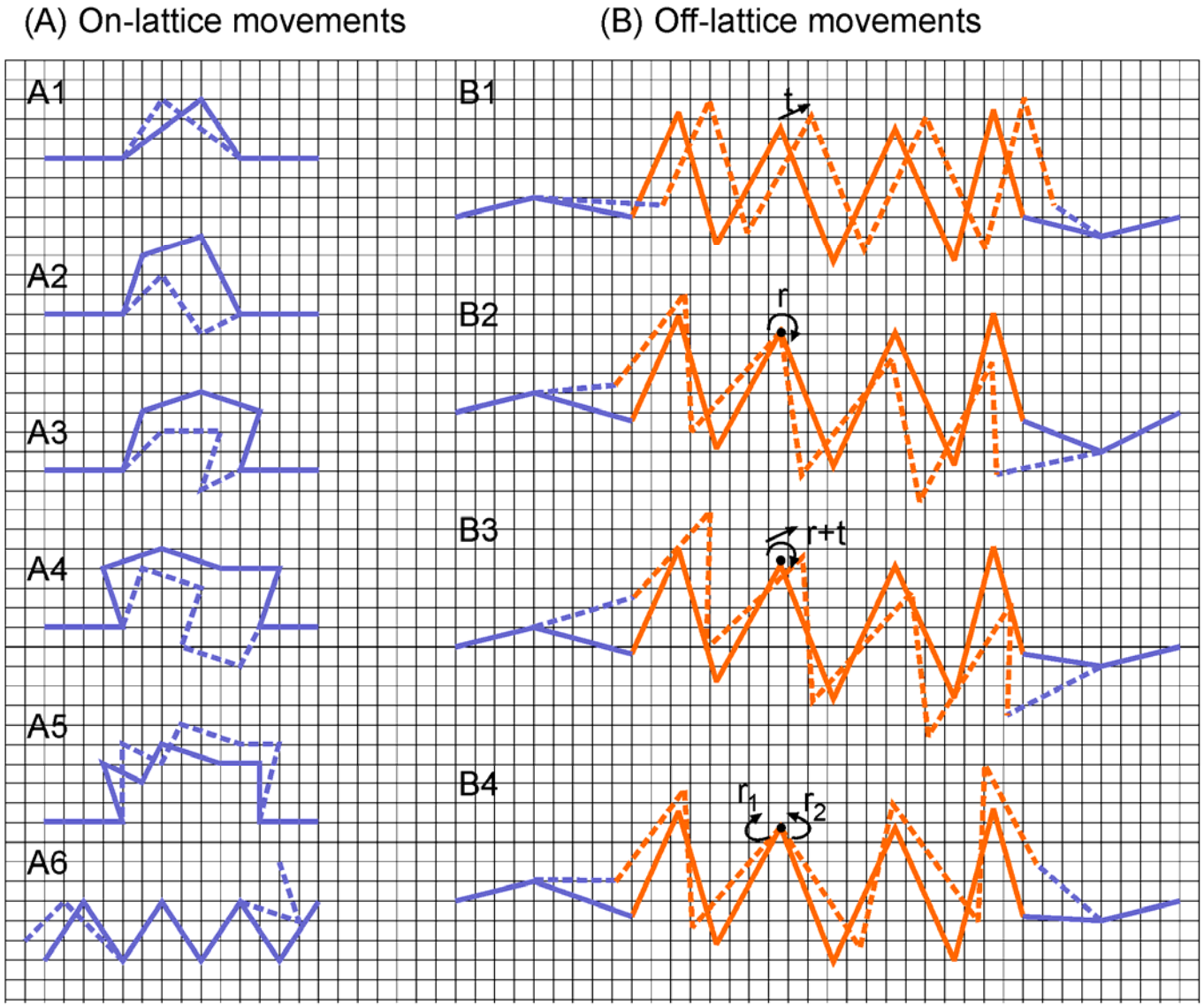
- Statistical terms from PDB library
  - Hydrophobicity/hydrophilicity
  - H-bond
  - Short-range Ca/SG correlations
- Propensity to predicted secondary structure (PSIPRED)
  - Short-range restraints
  - Protein-like
- Template-based terms
  - Long-range contacts
  - Ca-distance restraints
  - pair-potential

$$E = \sum_{i=1}^{25} w_i E_i$$

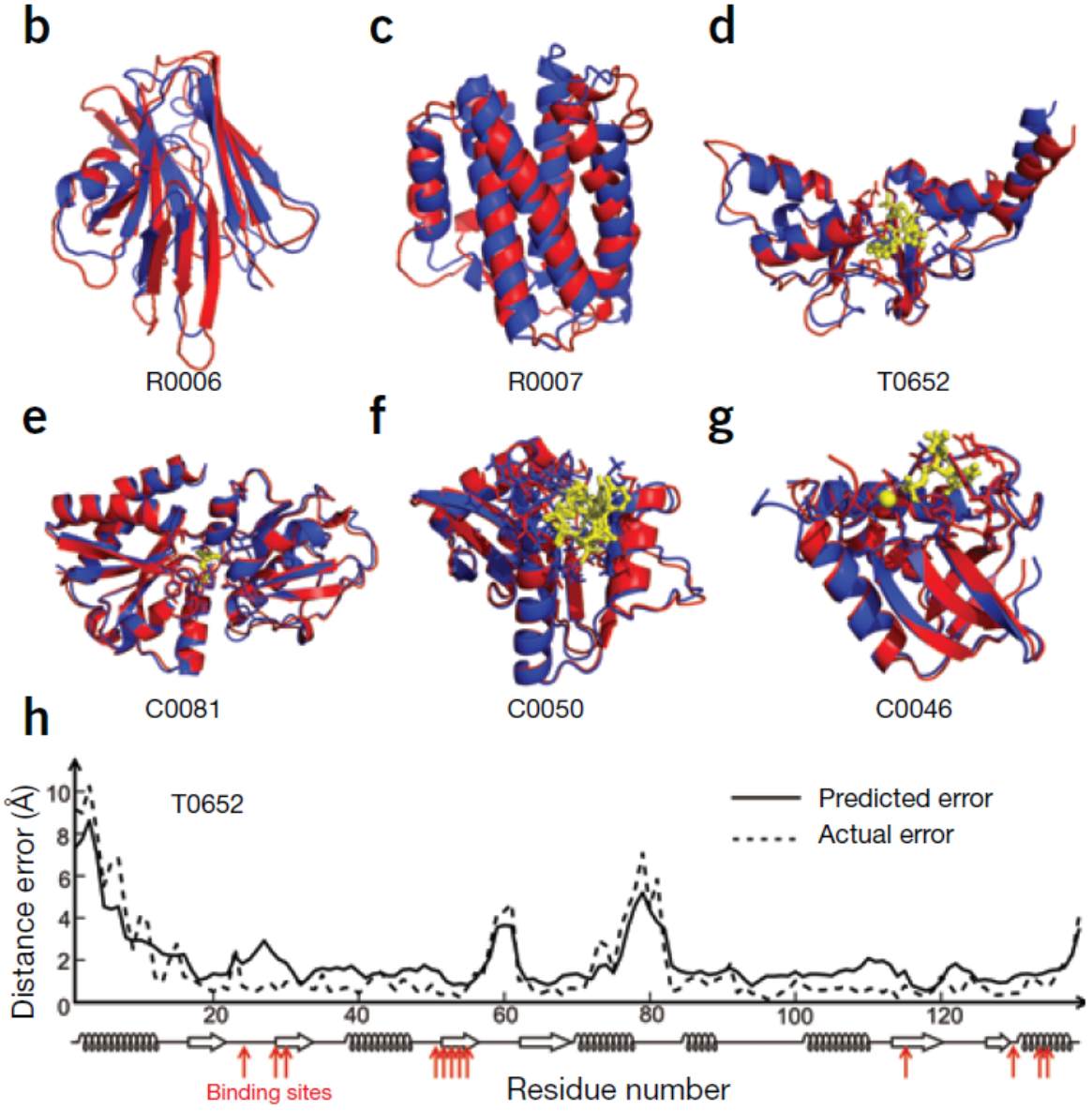
CAS model and on-and-off lattice model



Movement

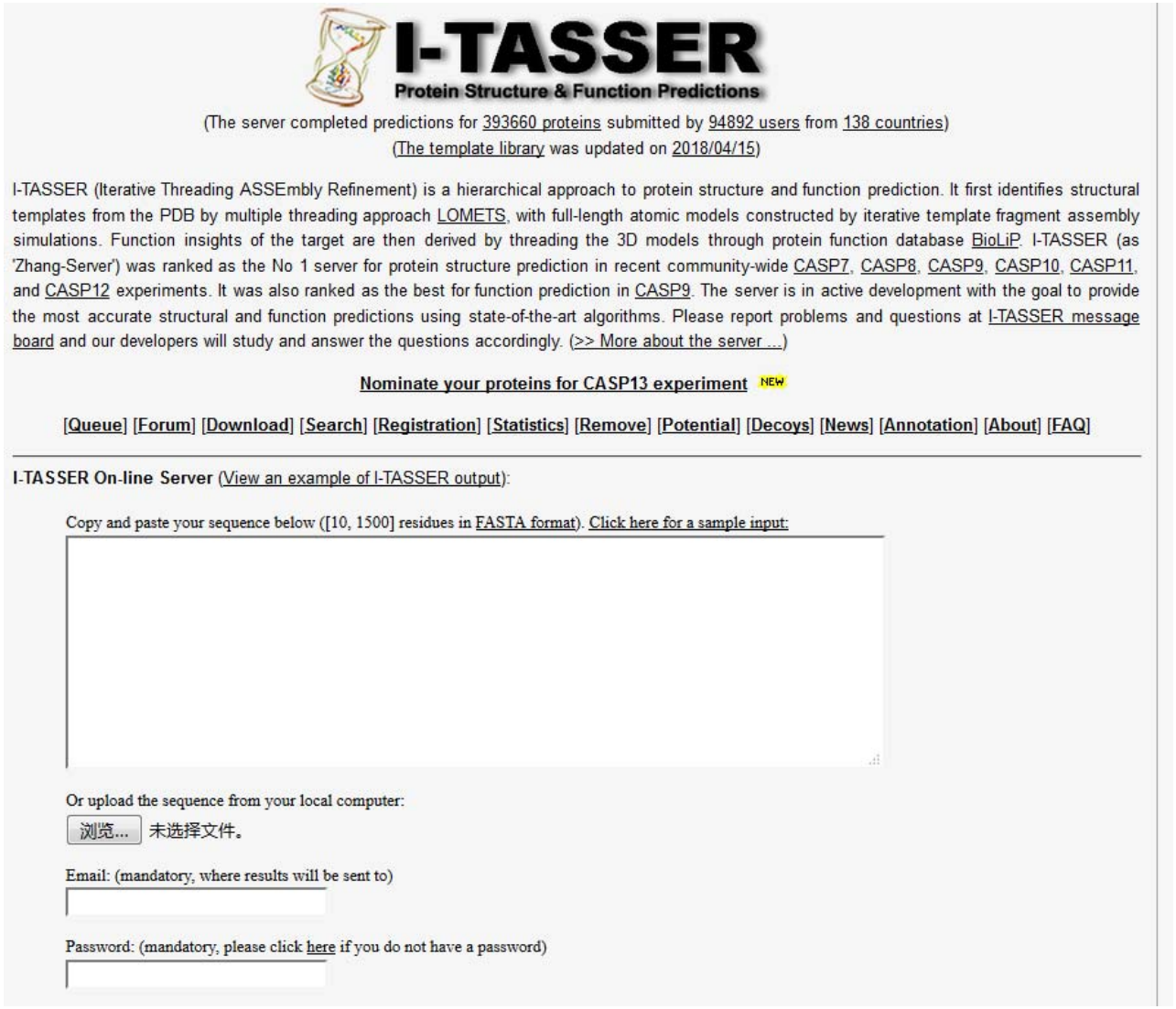

(A) On-lattice movements
(B) Off-lattice movements
A1
A2
A3
A4
A5
A6
B1
t
B2
r
B3
r+t
B4
r1
r2

I-TASSER results



Use I-TASSER server

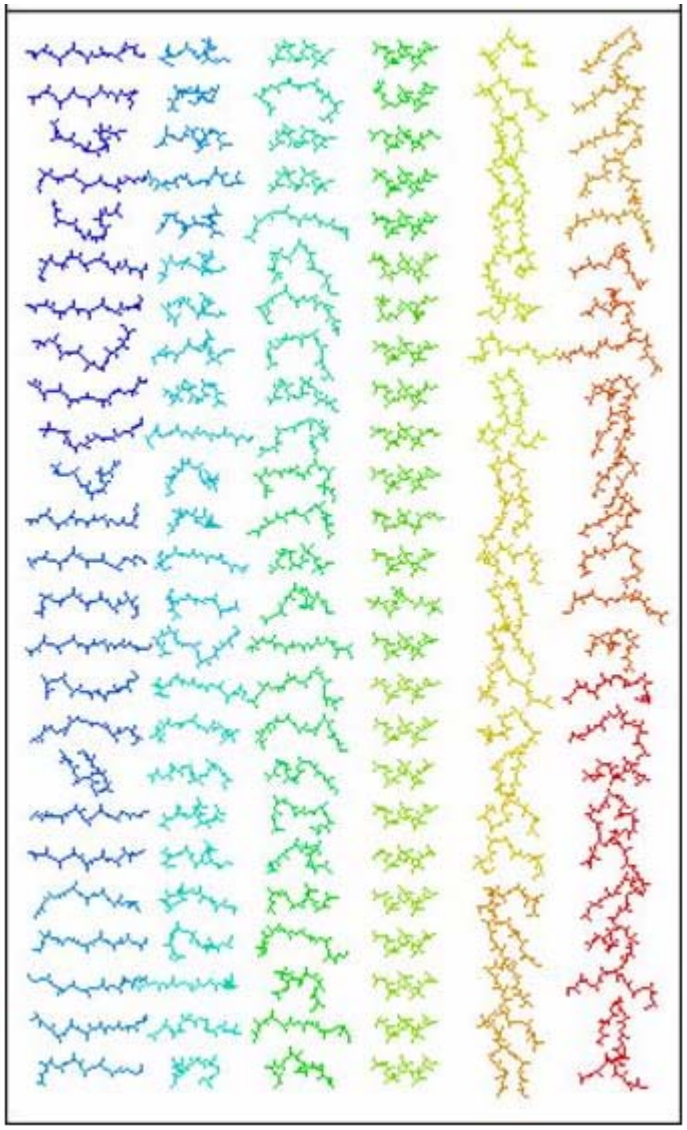
https://zhanglab.ccmb.med.umich.edu/I-TASSER/



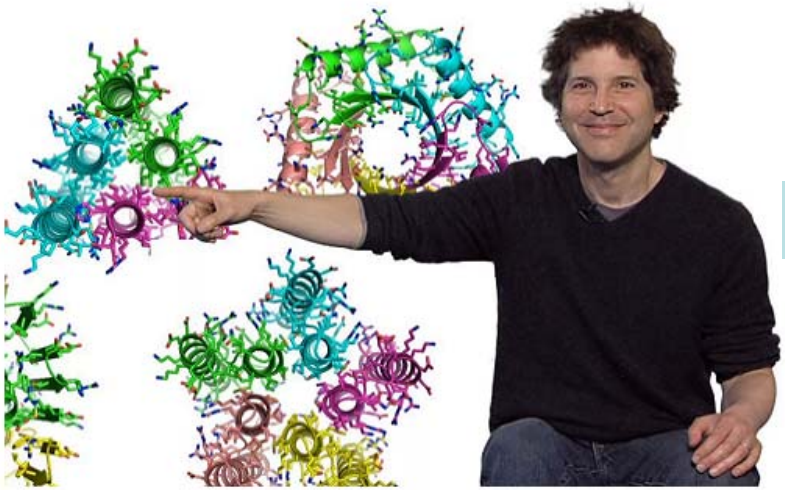
Content

- Ab-initio folding
- Homology modeling
  a. Comparative modeling(CM)
  b. Threading (fold recognition)
- Composite approach (I-TASSER)
- Fragment assembly (Rosetta, QUARK)
- CASP

Rosetta: fold protein structure by fragment assembly

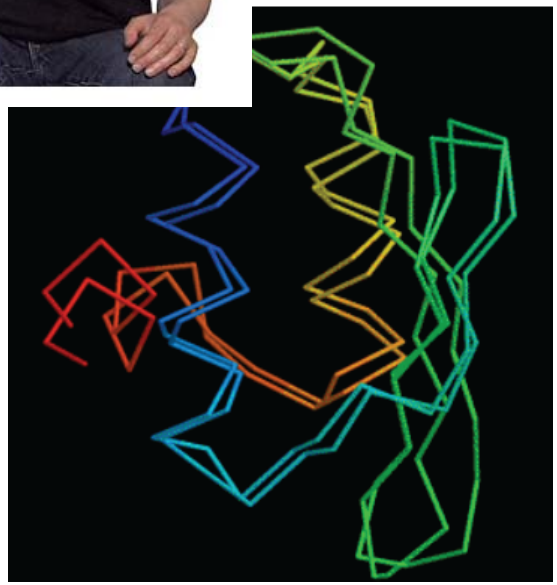


Fragment library



David Baker

Simulated annealing,
Monte Carlo simulation



Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions1

KT Simons, C Kooperberg, E Huang, D Baker - Journal of molecular biology, 1997 - Elsevier

We explore the ability of a simple simulated annealing procedure to assemble native-like structures from fragments of unrelated protein structures with similar local sequences using Bayesian scoring functions. Environment and residue pair specific contributions to the ...

☆ 99 被引用次数：1308 相关文章 所有 29 个版本

Rosetta force field (15 energy terms)

Name	Description (putative physical origin)	Functional form
env[b]	Residue environment (solvation)	$\sum_i -\ln\left[P(\mathrm{aa}_i\|\mathrm{nb}_i)\right]$
pair[b]	Residue pair interactions (electrostatics, disulfides)	$\sum_i \sum_{j>i} -\ln\left[\frac{P(\mathrm{aa}_i,\mathrm{aa}_j\|s_{ij}d_{ij})}{P(\mathrm{aa}_i\|s_{ij}d_{ij})P(\mathrm{aa}_i\|s_{ij}d_{ij})}\right]$
SS[d]	Strand pairing (hydrogen bonding)	$\mathrm{SchemeA}: \mathrm{SS}_{\phi,\vartheta}+\mathrm{SS}_{hb}+\mathrm{SS}_{d}$ $\mathrm{SchemeB}: \mathrm{SS}_{\phi,\vartheta}+\mathrm{SS}_{hb}+\mathrm{SS}_{d\sigma}$ where
sheet[e]	Strand arrangement into sheets	$-\ln\left[P(n_{\mathrm{sheets}}n_{\mathrm{lonestrands}}\|n_{\mathrm{strands}})\right]$
HS	Helix–strand packing	$\sum_m \sum_n -\ln\left[P(\phi_{mn},\psi_{mn}\|sp_{mn}d_{mn})\right]$
rg	Radius of gyration (vdw attraction; solvation)	$\sqrt{\langle d_{ij}^2\rangle}$
cbeta	Cβ density (solvation; correction for excluded volume effect introduced by simulation)	$\sum_i \sum_{sh} -\ln\left[\frac{P_{\mathrm{compact}}(\mathrm{nb}_{i,sh})}{P_{\mathrm{random}}(\mathrm{nb}_{i,sh})}\right]$
vdw[g]	Steric repulsion	$\sum_i \sum_{j>i} \frac{\left(r_{ij}^2-d_{ij}^2\right)^2}{r_{ij}};\ d_{ij}<r_{ij}$

Name	Description (physical origin)	Functional form
rama	Ramachandran torsion preferences	$\sum_i -\ln\left[P(\phi_i,\psi_i\|\mathrm{aa}_i,\mathrm{ss}_i)\right]$
LJ[c]	Lennard–Jones interactions	$\sum_i \sum_{j>i} \begin{cases} \left[\left(\frac{r_{ij}}{d_{ij}}\right)^{12}-2\left(\frac{r_{ij}}{d_{ij}}\right)^{6}\right]e_{ij}, & \text{if } \frac{d_{ij}}{r_{ij}}>0.6 \\ \left[-8759.2\left(\frac{d_{ij}}{r_{ij}}\right)+5672.0\right]e_{ij}, & \text{else} \end{cases}$
hb[f]	Hydrogen bonding	$\sum_i \sum_j \left(-\ln\left[P(d_{ij}\|h_j\mathrm{ss}_{ij})\right]\right.$ $-\ln\left[P(\cos\theta_{ij}\|d_{ij}h_j\mathrm{ss}_{ij})\right]$ $\left.-\ln\left[P(\cos\psi_{ij}\|d_{ij}h_j\mathrm{ss}_{ij})\right]\right)$
solv	Solvation	$\sum_i \left[\Delta G_i^{\mathrm{ref}}-\sum_j\left(\frac{2\Delta G_i^{\mathrm{free}}}{4\pi^{3/2}\lambda_i r_{ij}^2}e^{-d_{ij}^2}V_j+\frac{2\Delta G_i^{\mathrm{free}}}{4\pi^{3/2}\lambda_j r_{ij}^2}e^{-d_{ij}^2}V_i\right)\right]$
pair	Residue pair interactions (electrostatics, disulfides)	$\sum_i \sum_{j>i} -\ln\left[\frac{P(\mathrm{aa}_i,\ \mathrm{aa}_j\|d_{ij})}{P(\mathrm{aa}_i\|d_{ij})P(\mathrm{aa}_i\|d_{ij})}\right]$
dun	Rotamer self-energy	$\sum_i -\ln\left[\frac{P(\mathrm{rot}_i\|\phi_i,\psi_i)P(\mathrm{aa}_i\|\phi_i,\psi_i)}{P(\mathrm{aa}_i)}\right]$
ref	Unfolded state reference energy	$\sum_{\mathrm{aa}} n_{\mathrm{aa}}$

Rosetta benchmark results

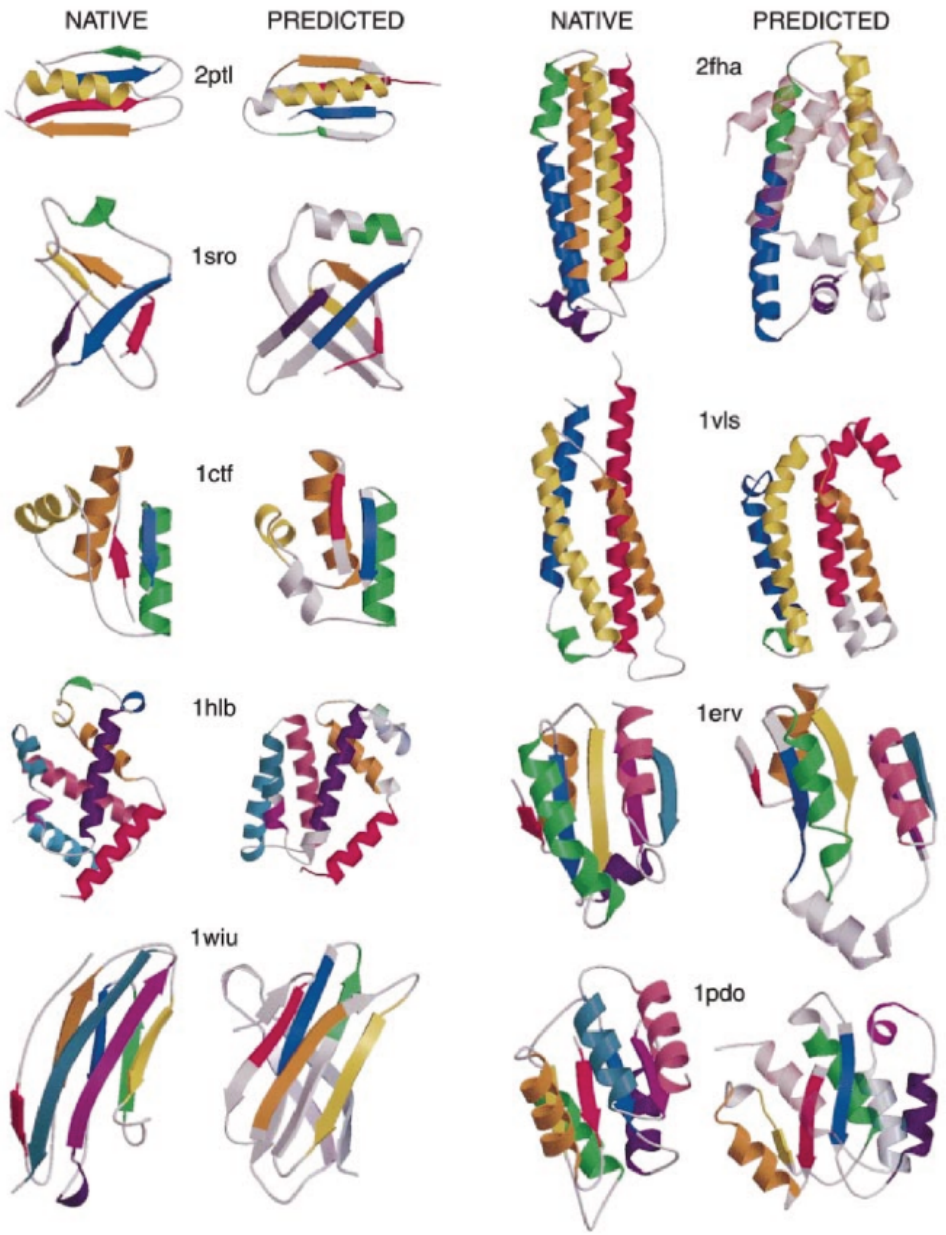


Table 2. Summary of predictions

Protein class	Total	<5 Å top five	<5 Å set
All	172	32	71
α	65	18	40
β	36	2	11
α/β	71	12	30
Small all	30	14	24
Small α	10	8	10
Small β	8	1	7
Small α/β	12	5	7
Medium all	127	18	56
Medium α	46	10	29
Medium β	27	1	4
Medium α/β	54	7	23
Large all	15	0	1
Large α	9	0	1
Large β	1	0	0
Large α/β	5	0	0

Best in top5 models

Best model

Figure 3. Comparison of native and predicted structures. The left column depicts the native structure and the right column is the best cluster center as identified in Table 1. The coloring of the secondary structural elements is demarcated by the native secondary structure assignment. Beginning with the N terminus, the coloring scheme is red, orange, yellow, green, blue, indigo, violet, turquoise and cyan. Images were prepared using Molscript (Kraulis, 1991) and Raster3d (Merritt & Bacon, 1997).

Foldit

Predicting protein structures with a multiplayer online game

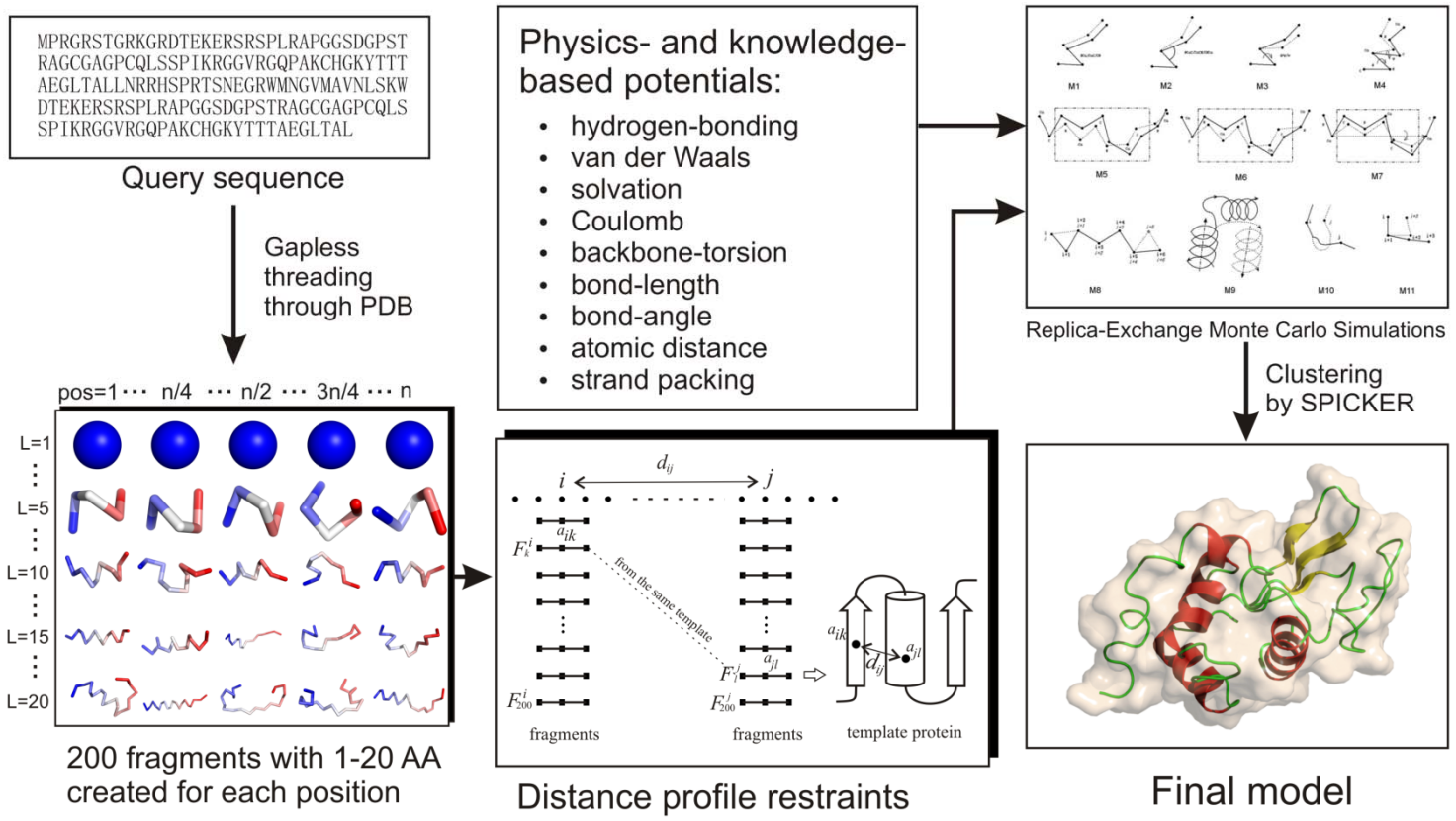
Seth Cooper[1], Firas Khatib[2], Adrien Treuille[1,3], Janos Barbero[1], Jeehyung Lee[3], Michael Beenen[1], Andrew Leaver-Fay[2]†, David Baker[2,4], Zoran Popović[1] & Foldit players

People exert large amounts of problem-solving effort playing computer games. Simple image- and text-recognition tasks have been successfully 'crowd-sourced' through games[1–3], but it is not clear if more complex scientific problems can be solved with human-directed computing. Protein structure prediction is one such problem: locating the biologically relevant native conformation of a protein is a formidable computational challenge given the very large size of the search space. Here we describe Foldit, a multiplayer online game that engages non-scientists in solving hard prediction problems. Foldit players interact with protein

retaining the deterministic Rosetta algorithms as user tools. We developed a multiplayer online game, Foldit, with the goal of producing accurate protein structure models through gameplay (Fig. 1). Improperly folded protein conformations are posted online as puzzles for a fixed amount of time, during which players interactively reshape them in the direction they believe will lead to the highest score (the negative of the Rosetta energy). The player's current status is shown, along with a leader board of other players, and groups of players working together, competing in the same puzzle (Fig. 1, arrows 8 and 9). To make the game approachable by players with

Author Contributions All named authors contributed extensively to development and analysis for the work presented in this paper. Foldit players (more than 57,000) contributed extensively through their feedback and gameplay, which generated the data for this paper.

QUARK: Ab-initio protein structure prediction using continuous fragments



Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field

D Xu, Y Zhang - Proteins: Structure, Function, and ..., 2012 - Wiley Online Library

Ab initio protein folding is one of the major unsolved problems in computational biology owing to the difficulties in force field design and conformational search. We developed a novel program, QUARK, for template-free protein structure prediction. Query sequences are first broken into fragments of 1–20 residues where multiple fragment structures are retrieved at each position from unrelated experimental structures. Full-length structure models are then assembled from fragments using replica-exchange Monte Carlo simulations, which are ...

☆ 99 被引用次数：479 相关文章 所有 14 个版本

QUARK force field

$$E_{\text{tot}} = E_{\text{prm}} + w_1 E_{\text{prs}} + w_2 E_{\text{ev}} + w_3 E_{\text{hb}} + w_4 E_{\text{sa}} + w_5 E_{\text{dh}} + w_6 E_{\text{dp}} + w_7 E_{\text{rg}} + w_8 E_{\text{bab}} + w_9 E_{\text{hp}} + w_{10} E_{\text{bp}}$$

1, Backbone atomic pair-wise potential

$$E_{\text{prm}}(i,j,r_{ij}) = -RT \log\left(\frac{N_{\text{obs}}(i,j,r_{ij})}{r_{ij}^{\alpha} N_{\text{obs}}(i,j,r_{\text{cut}})}\right)$$

2, Side-chain center pair-wise potentials

$$E_{\text{prs}}(i,j,r_{ij}) = -RT \log\left(\frac{N'_{\text{obs}}(i,j,r_{ij})}{r_{ij}^{\alpha'} N'_{\text{obs}}(i,j,r_{\text{cut}})}\right)$$

3, Excluded volume

$$E_{\text{ev}}(i,j,r_{ij}) = \begin{cases} (vdw(i)+vdw(j))^2 - r_{ij}^2 & \text{if } r_{ij} < vdw(i)+vdw(j) \\ 0 & \text{else} \end{cases}$$

4, Hydrogen bonding

$$E_{\text{hb}}(i,j,T_k) = \sum_{l=1}^{n_k} \frac{(f_l(i,j)-\mu_{kl})^2}{2\delta_{kl}^2}, \qquad n_k = \begin{cases} 4 & k=1,2 \\ 3 & k=3,4 \end{cases}$$

5, Solvent accessibility

$$E_{\text{sa}} = \sum_{i=1}^{L} |s_i - s_i^E| \qquad s_i = 1 - w \sum_{d(G_i,G_j)<9\text{Å}} \frac{A_{aa(j)}}{d^2(G_i,G_j)}$$

6, Backbone torsion potential

$$E_{\text{dh}} = -\sum_{i=2}^{L-1} \log(P(\phi_i,\psi_i|aa(i),ss(i)))$$

7, Fragment-based distance profile

$$E_{\text{dp}} = -\sum_{(i,j)\in S_{\text{dp}}} \log(N_{i,j}(d_{ij}))$$

8, Radius of gyration

$$E_{\text{rg}} = \begin{cases} 0 & r_{\min} \le r \le r_{\max} \\ (r_{\min}-r)^2 & r < r_{\min} \\ (r-r_{\max})^2 & r > r_{\max} \end{cases}$$

9, Strand-helix-strand packing

$$E_{\text{bab}} = \begin{cases} E_{\text{pen}} & \text{left}-\text{handed} \\ 0 & \text{else} \end{cases}$$

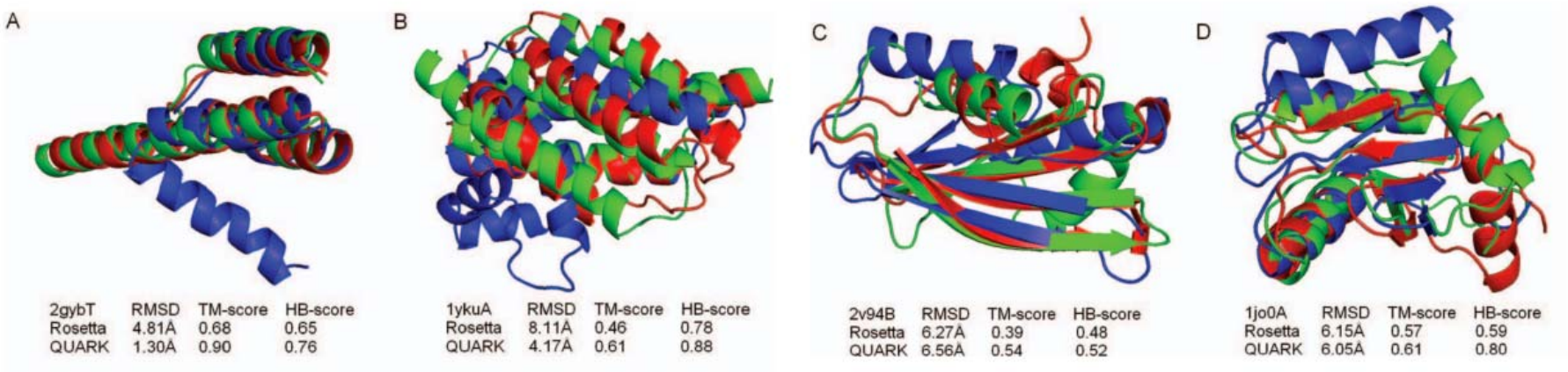
10, Helix packing

$$E_{\text{hp}}(i,j) = -\log(P(d_{ij},\ \phi_{ij}))$$

11, Solvent accessibility

$$E_{\text{bp}}(i,j) = -\log(P(aa(i),aa(j),T_{ij}))$$

Benchmark result of QUARK vs. Rosetta



Native: Red; Rosetta: Blue; QUARK: Green

Ab initio Modeling Results by QUARK and Rosetta

		First (best in top five) cluster center model					
		RMSD	TM-score	GDT-TS	MaxSub	HB-score	Time (h)
51 small proteins with [70–100) residues	Rosetta	10.1 Å (8.5 Å)	0.350 (0.393)	0.381 (0.418)	0.291 (0.337)	0.442 (0.491)	25.0
	QUARK	9.1 Å (7.7 Å)	0.404 (0.441)	0.428 (0.466)	0.349 (0.389)	0.503 (0.538)	37.7
	QUARK-h[a]	6.4 Å (4.6 Å)	0.585 (0.667)	0.602 (0.691)	0.552 (0.635)	0.610 (0.681)	37.7
94 medium proteins with [100–150] residues	Rosetta	13.0 Å (11.5 Å)	0.317 (0.346)	0.293 (0.318)	0.224 (0.247)	0.410 (0.453)	63.3
	QUARK	12.5 Å (10.7 Å)	0.334 (0.374)	0.310 (0.342)	0.237 (0.268)	0.471 (0.504)	79.5
	QUARK-h[a]	8.6 Å (6.6 Å)	0.491 (0.541)	0.439 (0.483)	0.362 (0.410)	0.449 (0.503)	79.5

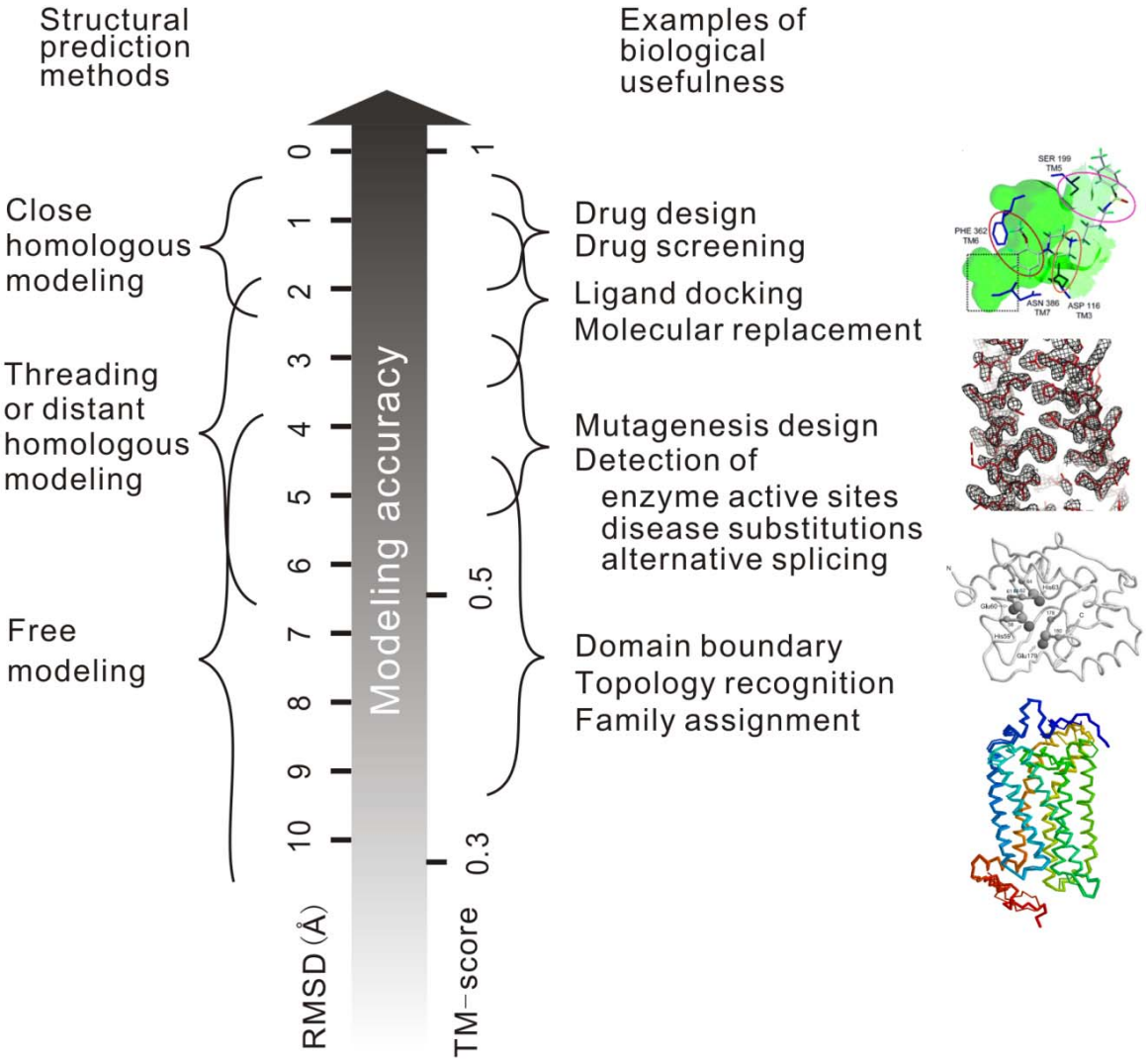
[a]QUARK simulation using fragments without removing homologous templates to the query sequence. However, the target proteins themselves, if existing in the library, were excluded.

Summary: Methods vs. Accuracy

Experimental resolution

Alignment is key, error on loop

Low resolution for small proteins



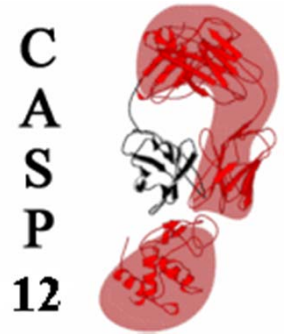
Y Zhang, Curr Opin Str Biol 19: 145 (2009)

Content

- Ab-initio folding
- Homology modeling
  a. Comparative modeling(CM)
  b. Threading (fold recognition)
- Composite approach (I-TASSER)
- Fragment assembly (Rosetta, QUARK)
- CASP

CASP

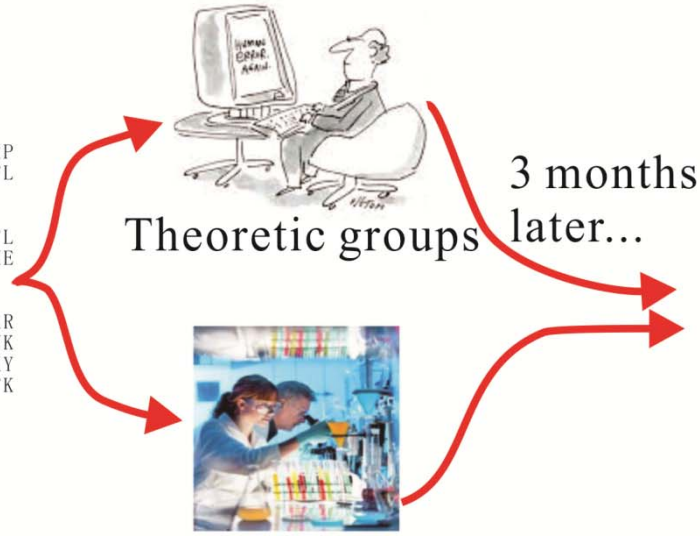
http://predictioncenter.org/index.cgi

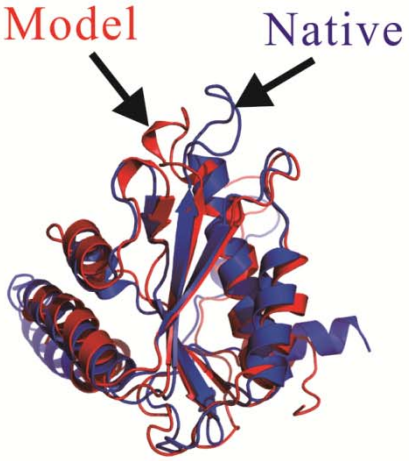


- Critical Assessment of protein Structure Prediction
- CASP1(1994)—CASP12(2016)

A double blind test

>T0859 AP205, Acinotebacter Phage AP205, 133 residues
GSMANKPMQPITSTANKIVWSDPTRLSTTFSASLLRQRVKVGIAELNNVSGQYVSVYKRP
APKPEGGADAGVIMPNENQSIRTVISGSAENLATLKAEWETHKRNVDTLFASGNAGLGFL
DPTAAIVSSDTTA
>T0860 fibre head domain, Raptor adeovirus 1, 137 residues
VSYSDGHFLTKSGGVINFRKTRVTSITITILGNYGLRVVNGELQNTPLTFKGADFKSSTL
KDELLIPLEGAVQLNTAPSTALCIFITTDHVYRELCMMQFLTDVDKTPFLVVLRSESKHE
TIQYMHIVTVHPFLSLT
>T0863 STRA6, Danio rerio, 670 residues
MSAETVNNYDYSDWYENAAPTKAPVEVIPPCDPTADEGLFHICIAAISLVVMLVLAILAR
RQKLSDNQRGLTGLLSPVNFLDHTQHKGLAVAVYGVLFCKLVGMVLSHHPLPFTKEVANK
EFWMILALLYYPTLYYPLLACGTLHNKVGYVLGSLLSWTHFGILVWQKVDCPKTPQIYKY
YALFGSLPQIACLAFLSFQYPLLLFKGLQNTETANASEDLSSSYYRDYVKKILKKKKPTK
ISSSTSKPKLFDRLRDAVKSYIYTPEDVFRFPLKLAISVVVAFIA





Meeting...
Comparing...
Ranking...