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 structure alignment

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

➡ 1. What is structure superposition?

    a. RMSD

    b. TM-score

2. What is structure alignment?

3. Different structure alignment algorithms

    1. DALI

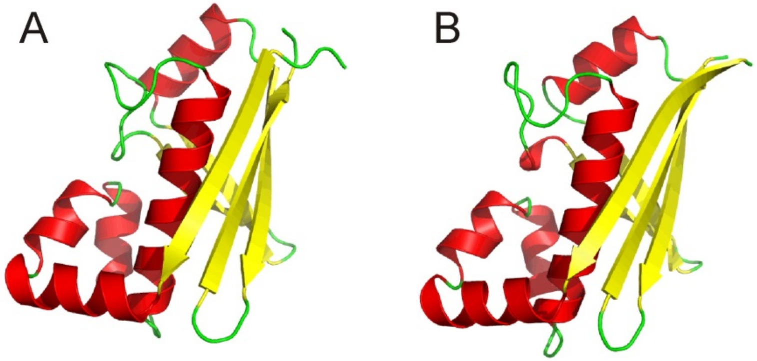
    2. CE

    3. TM-align

4. Multiple protein structure alignment

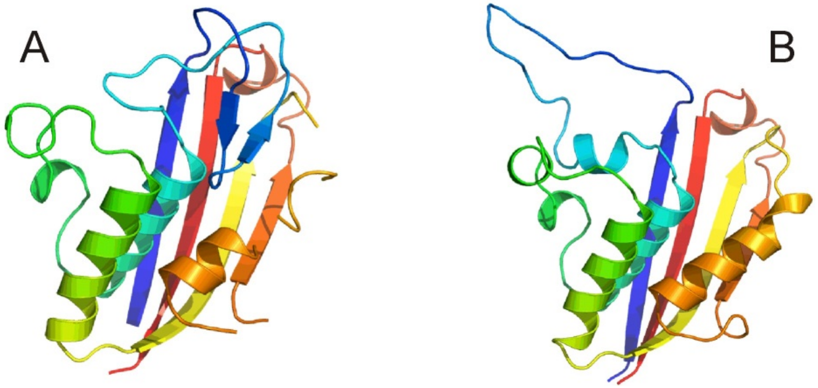
Two types of protein structure comparisons

1. **Structure superposition**: when we know the equivalency of residues



A： AMIVGLGTDIAEIERVEKALARSGENFARRILTDSELEQFHASKQQGRFLAKRFAAKEAASKALGTGIAQGVTFHDFTISHDKLGKPLLILSGQAAELASQLQVENIHLSISDERHYAMATVILERR
B： AMIVGLGTDIAEIERVEKALARSGENFARRILTDSELEQFHASKQQGRFLAKRFAAKEAASKALGTGIAQGVTFHDFTISHDKLGKPLLILSGQAAELASQLQVENIHLSISDERHYAMATVILERR

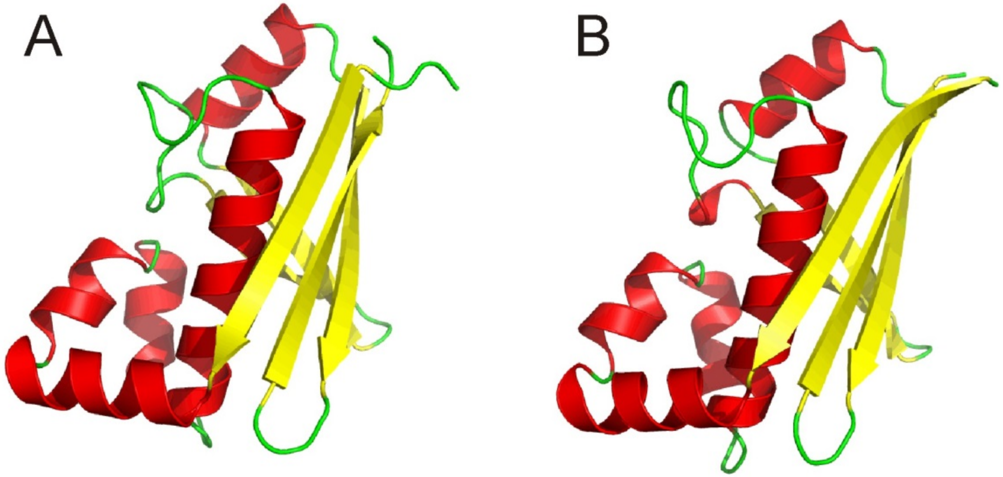
2. **Structure alignment**: when we do NOT know the equivalency of residues



A： ERIGHGFDVHAFGGEGPIIIGGVRIPYEKGLLAHSDGDVALHALTDALLGAAALGDIGKLFPDTDPAFKGADSRELLREAWRRIQAKGYTLGNVDVTIIAQAPKLPHIPQRVFIAEDLGCHDDVNVKATTTEKLGFTGRGEGIACEAVALLIK
B： MIRIGHGFDVHAFGEDRPLIIGGVEVPYHTGFIAHSDGDVALHALTDAILGAAALGDIGKLFPDTDMQYKNADSRGLLREAFRQVQEKGYKIGNVDITIIAQAPKMRPHIDAMRAKIAEDLQCDIEQVNVKATTTEKLGFTGRQEGIACEAVALLIR

What is structure superposition?

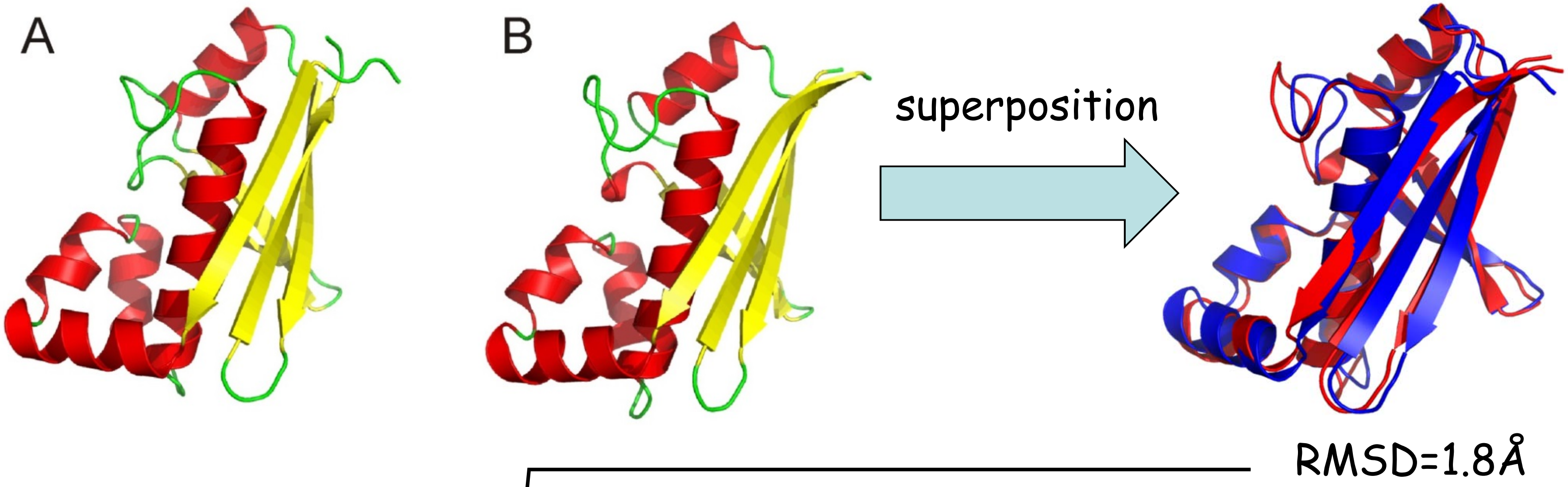
1. **Structure superposition**: when we know the equivalency of residues



A： AMIVGLGTDIAEIERVEKALARSGENFARRILTDSELEQFHASKQQGRFLAKRFAAKEAASKALGTGIAQGVTFHDFTISHDKLGKPLLILSGQAAELASQLQVENIHLSISDERHYAMATVILERR
B： AMIVGLGTDIAEIERVEKALARSGENFARRILTDSELEQFHASKQQGRFLAKRFAAKEAASKALGTGIAQGVTFHDFTISHDKLGKPLLILSGQAAELASQLQVENIHLSISDERHYAMATVILERR

Goal: find the best match (in space) between the equivalent atoms of two structures (after optimal rotation and translation), with the global similarity assessed by a single score.

RMSD: root-mean-square-deviation



$$\mathbf{RMSD} = \min_{\substack{U \in M_{3\times 3}, \\ u_0 \in V_3}} \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left\| (U \overrightarrow{x_n} + \overrightarrow{u_0}) - \overrightarrow{y_n} \right\|_2^2}$$

$$s.t. \qquad UU^{\mathrm{T}} = I$$

Unitary transformation to keep structure rigid

Kabsch Algorithm

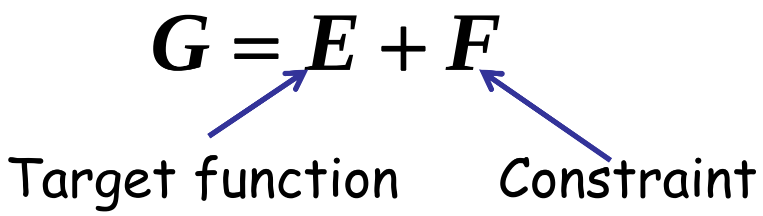
Reference (a mathematical solution to RMSD)
W. Kabsch, A solution for the best rotation to relate two sets of vectors
Acta Cryst (1976) A32: 922-923

Kabsch Algorithm

Lagrange multipliers

$$E = \frac{1}{2}\sum_{n}\left\|U\overrightarrow{x_n} - \overrightarrow{y_n}\right\|_2^2$$

$$F = \frac{1}{2}\sum_{i,j} l_{ij}\left(\sum_{k} U_{kl}U_{kj} - \delta_{ij}\right)$$



$$G = E + F$$

Target function

Constraint

By

$$\frac{\partial G}{\partial U_{ij}} = \sum_{k} U_{ik}\left(\sum_{n} x_{nk}x_{nj} + l_{kj}\right) - \sum_{n} y_{ni}x_{nj} = 0$$

Finally

$$U_{ij} = \sum_{k} b_{ki}a_{kj}$$

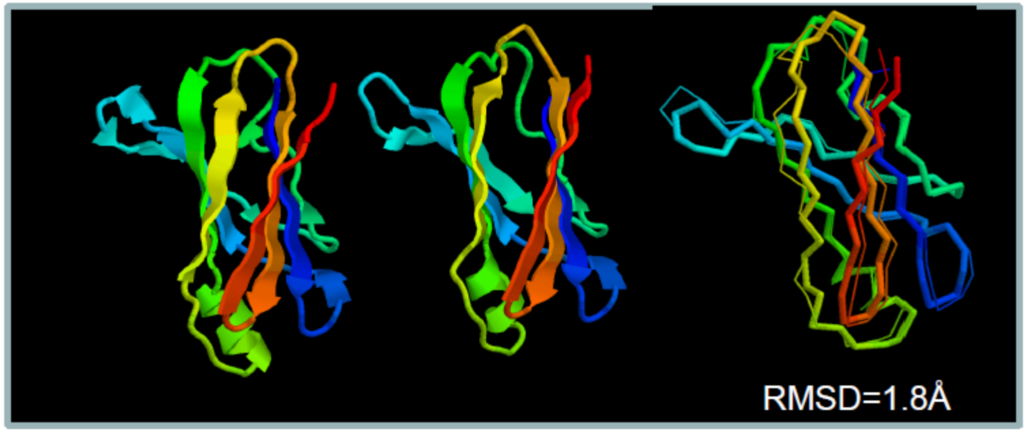
where a_k is the eigenvector of the matrix $R^T R$, $R = X^T Y$

$b_k = \lambda_k R a_k$, (λ_k is the eigenvalue of $R^T R$)

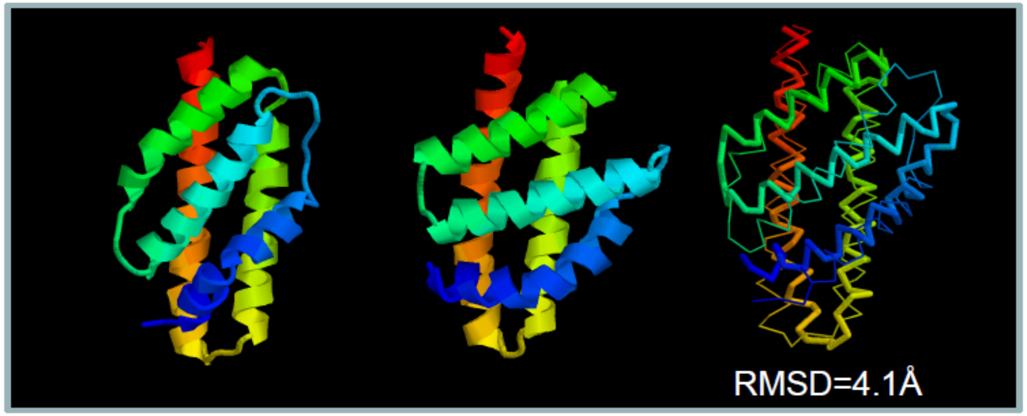
Source code to calculate RMSD: http://zhanglab.ccmb.med.umich.edu/TM-score/RMSD.f

General RMSD values

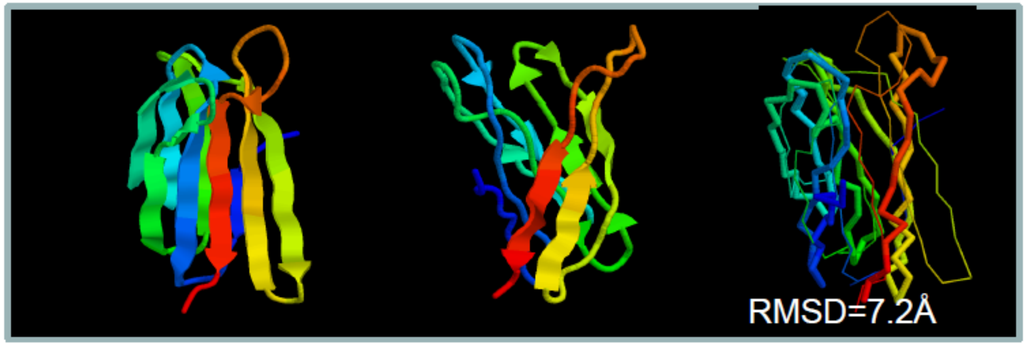
- RMSD in [0, 2Å], high resolution structures of close similarity



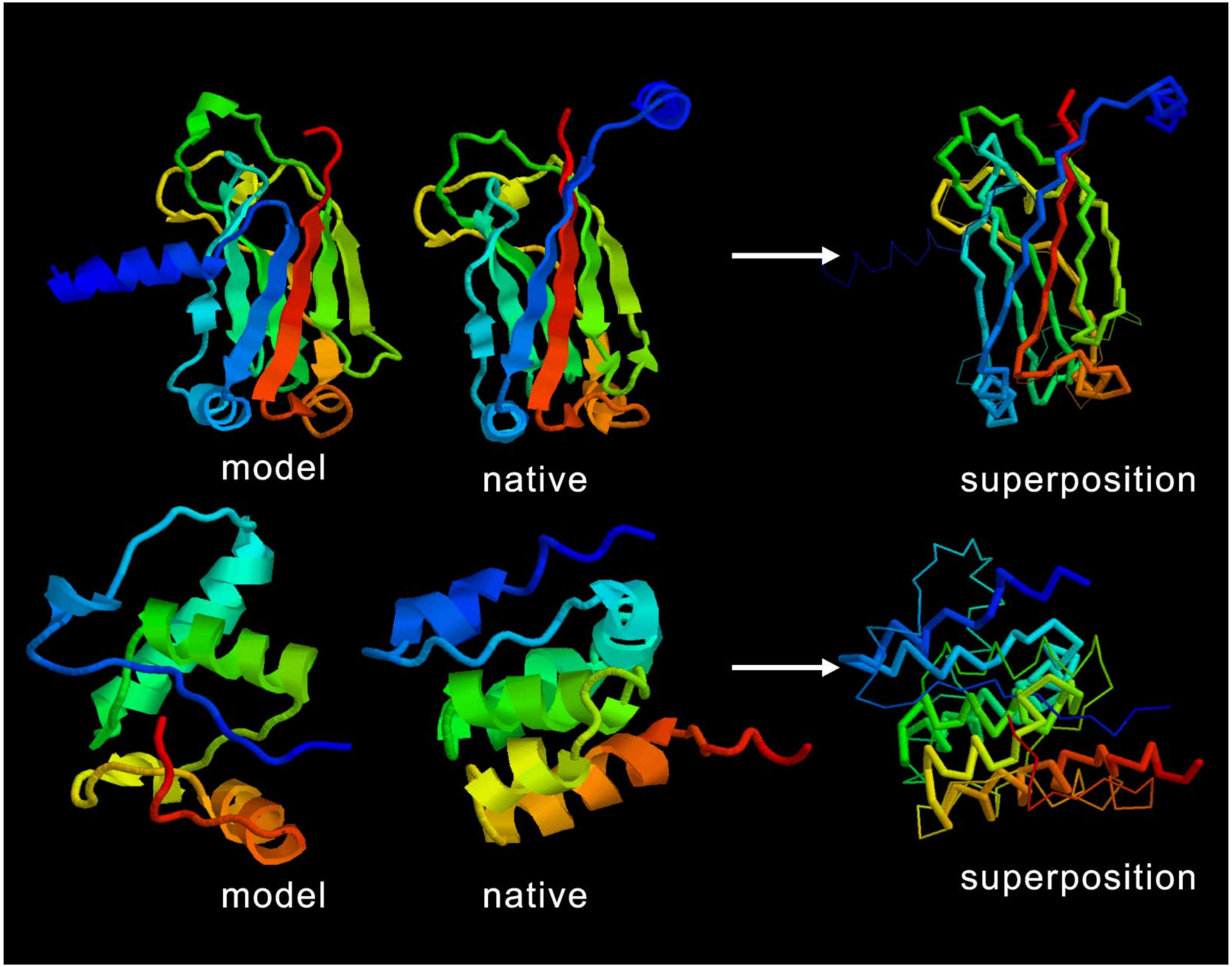
- RMSD in (2Å, 6Å], similar topology, medium resolution



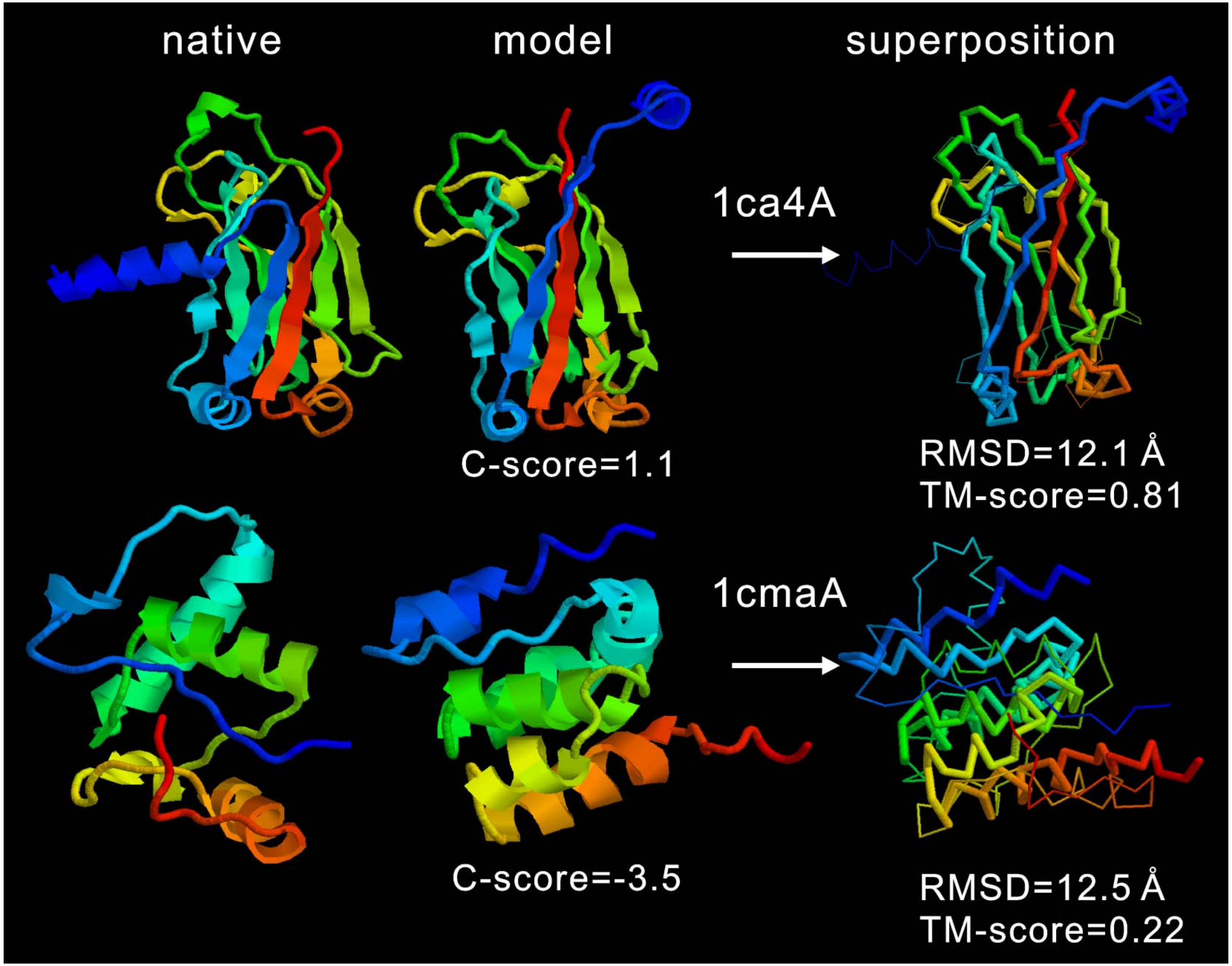
- RMSD > 6Å, different topology, low resolution



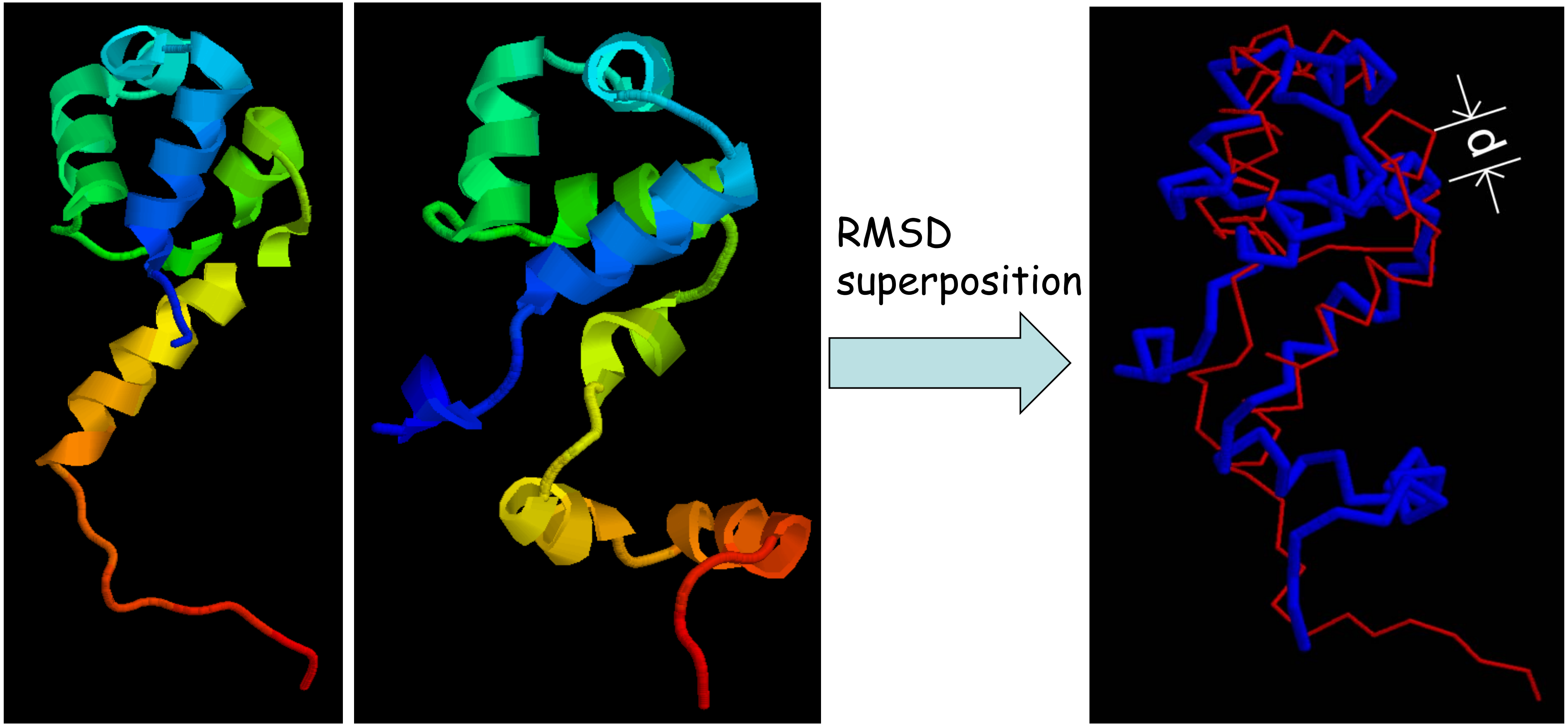
Problem of RMSD



Problem of RMSD



Problem of RMSD



RMSD superposition

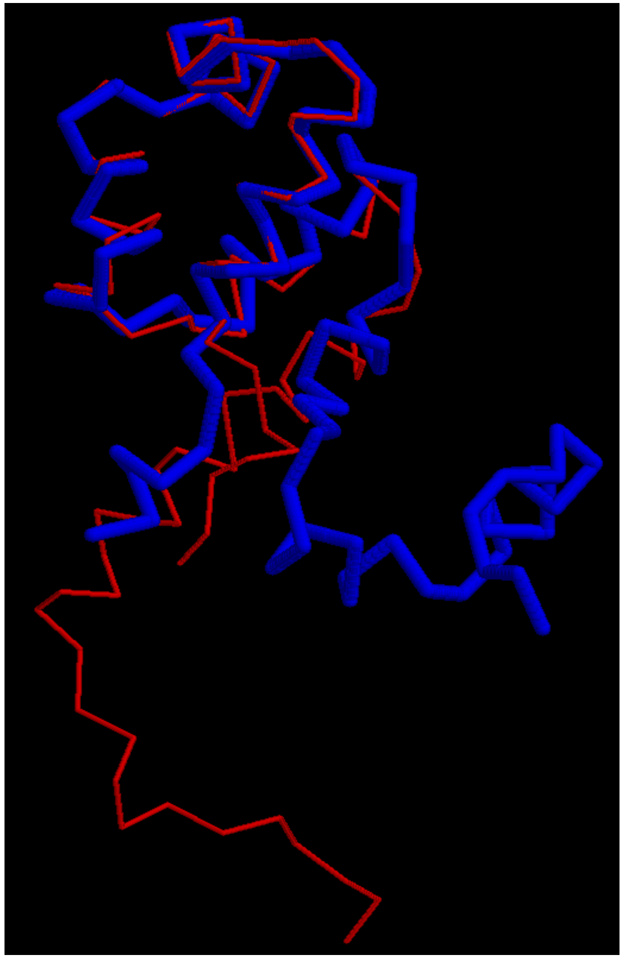
RMSD=9.1Å

TM-score

Reference: Yang Zhang, Jeffrey Skolnick. A scoring function for the automated assessment of protein structure template quality. Proteins, vol 57, 702 (2004).

Definition of TM-score

TM-score superposition:



TM-score=0.61

$$\text{TM-score} = \max \frac{1}{L} \sum_{i=1}^{N_{ali}} \frac{1}{1+\left(d_i / d_0\right)^2}$$

$$d_0 = \begin{cases} 1.24 \times \sqrt[3]{L-15} - 1.8, & \text{if } L > 21 \\ 0.5, & \text{otherwise} \end{cases}$$

0<TM-score≤1

Reference: Y. Zhang and J. Skolnick, Proteins 27 (2004) 702-710.

TM-score is length-independent

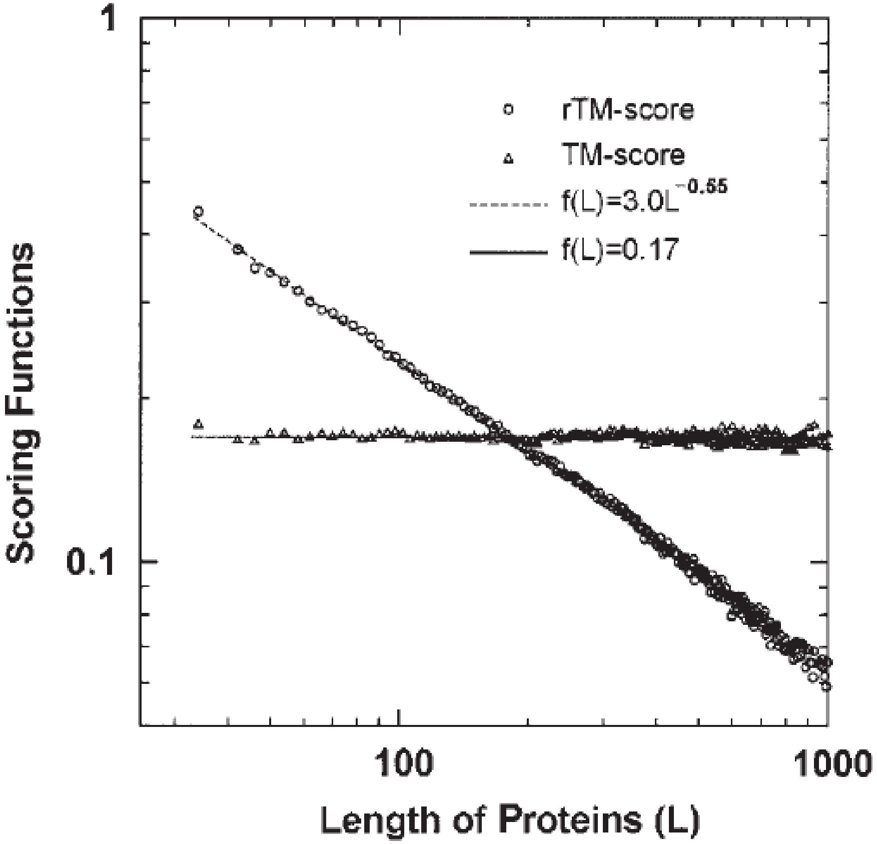
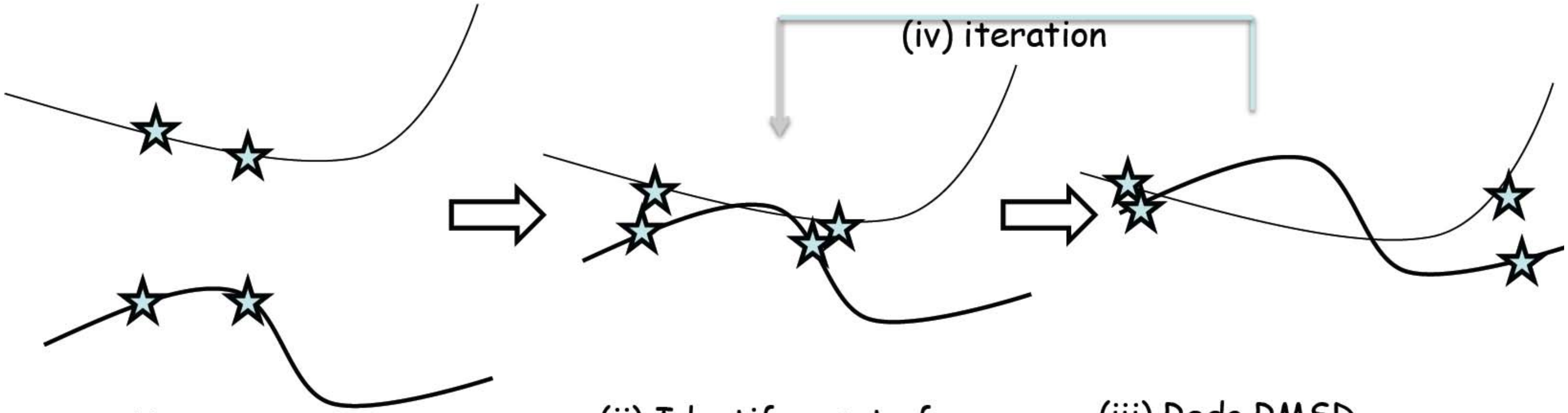


Fig. 3. The average ‘raw TM-score’ (rTM-score) and TM-score of random protein pairs as a function of protein size. For the rTM-score, d_0 = 5 Å; for the TM-score, d_0 is defined as in eq. (5). The data are calculated from all pairs of 3656 PDB structures of <30% sequence identity. The statistical error bars are smaller than the size of the points. The dashed line is a nonlinear least square Marquardt–Levenberg fit of the rTM-score data to a power-law equation *f*(*L*), where *L* is the length of the smaller protein of the corresponding structure pairs. The solid line denotes the horizontal line of TM-score = 0.17.

How to calculate TM-score?

How to analytically calculate the value of TM-score is still an open question.

1. Heuristic iterations from a random segments:



(iv) iteration

(i) RMSD superposition of an initial segment

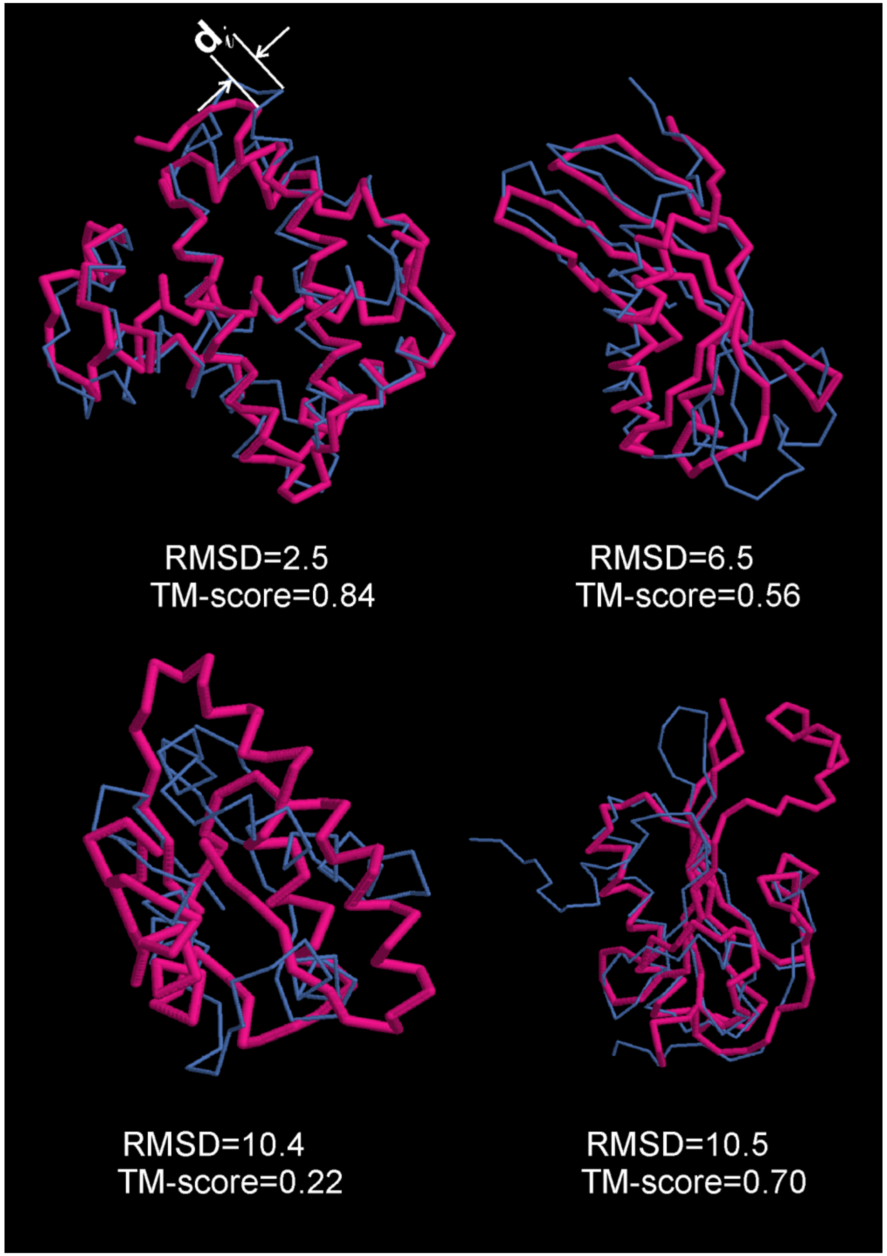
(ii) Identify a set of residue pairs with $d<d_0$

(iii) Redo RMSD superposition based on the new residue set

2. Start from different positions and then select the superposition with the highest TM-score

Concept of RMSD and TM-score

TM-score is more sensitive to the global topology



Content

1. What is structure superposition?
    a. RMSD
    b. TM-score
2. ➡ What is structure alignment?
3. Different structure alignment algorithms
    1. DALI
    2. CE
    3. TM-align
4. Multiple protein structure alignment

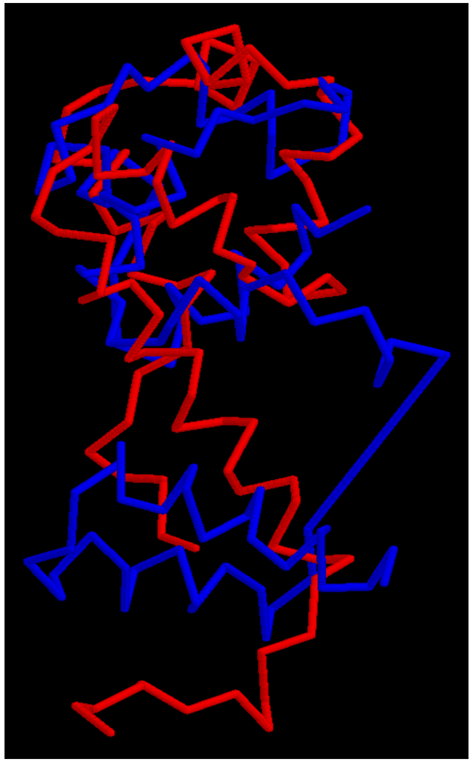
What is structure alignment?

Structure superposition

123······789

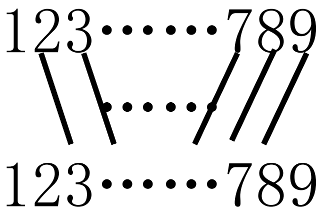
|||······|||

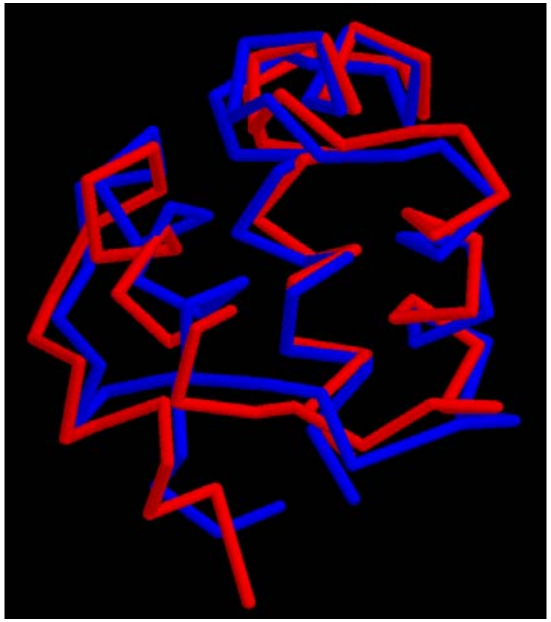
123······789



Alignment is given

Structure alignment





Alignment should be identified based on structure match

Structure Alignment: Issues

Theoretical Issues

- NP-hard geometric problem
- Heuristics needed
- No unique solution

Methodological Issues

Choices:

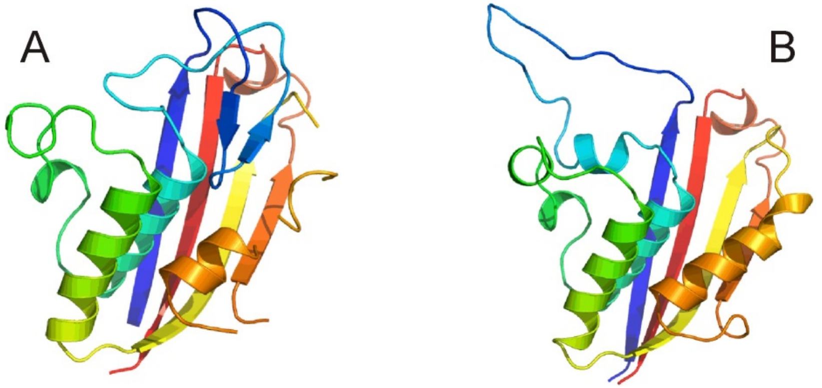
- Structure Representation
- Scoring function
- Search algorithm

Goals Desirable

- Automatic
- Discriminating
- Fast

What is structure alignment?

structure alignment: when we do not know the equivalency of residues



A： ERIGHGFDVHAFGGEGPIIIGGVRIPYEKGLLAHSDGDVALHALTDALLGAAALGDIGKLFPDTDPAFKGADSRELLREAWRRIQAKGYTLGNVDVTIIAQAPKLPHIPQRVFIAEDLGCHDDVNVKATTTEKLGFTGRGEGIACEAVALLIK
B： MIRIGHGFDVHAFGEDRPLIIGGVEVPYHTGFIAHSDGDVALHALTDAILGAAALGDIGKLFPDTDMQYKNADSRGLLREAFRQVQEKGYKIGNVDITIIAQAPKMRPHIDAMRAKIAEDLQCDIEQVNVKATTTEKLGFTGRQEGIACEAVALLIR

Two steps:

Step 1: find alignment
Step 2: calculate superposition score (e.g. RMSD or TM-score)

Content

1. What is structure superposition?
   a. RMSD
   b. TM-score
2. What is structure alignment?
3. Different structure alignment algorithms
   1. DALI
   2. CE
   3. TM-align
4. Multiple protein structure alignment

Methods for structure alignment

DALI:
Holm and Sander. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993, 233: 123-28

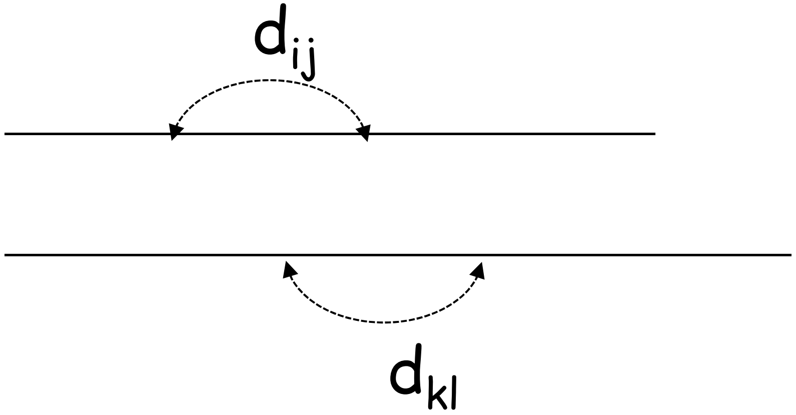
CE:
Shindyalov and Bourne, Protein structure alignment by incremental combinatorial extension (CE) of optimal path. Prot Eng, 1998, 11 739-747

TM-align:
Yang Zhang, Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research, vol 33, 2302 (2005).

Two ways of scoring structural similarity

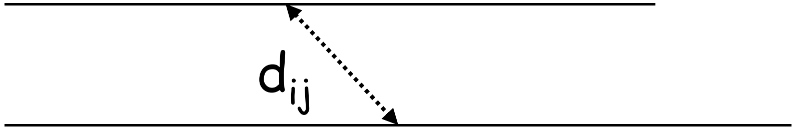
Intra-distance

$$S = \sum | d_{ij} - d_{kl} |$$



Inter-distance

$$S = \sum d_{ij}$$



Many search methods

- Heuristic growing
- Monte Carlo
- Dynamics programming

Content

1. What is structure superposition?
    a. RMSD
    b. TM-score
2. What is structure alignment?
3. Different structure alignment algorithms
    ➡ 1. DALI
    2. CE
    3. TM-align
4. Multiple protein structure alignment

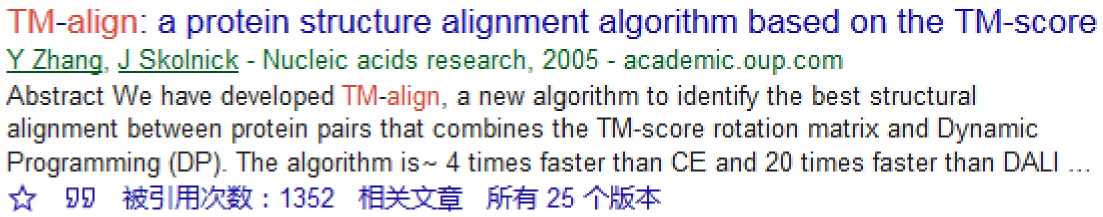
Content

1. What is structure superposition?
    a. RMSD
    b. TM-score
2. What is structure alignment?
3. Different structure alignment algorithms
    1. DALI
    2. CE
    3. ➡ TM-align
4. Multiple protein structure alignment

TM-align

Reference:

Yang Zhang, Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research, vol 33, 2302 (2005).



TM-align: a protein structure alignment algorithm based on the TM-score

Y Zhang, J Skolnick - Nucleic acids research, 2005 - academic.oup.com

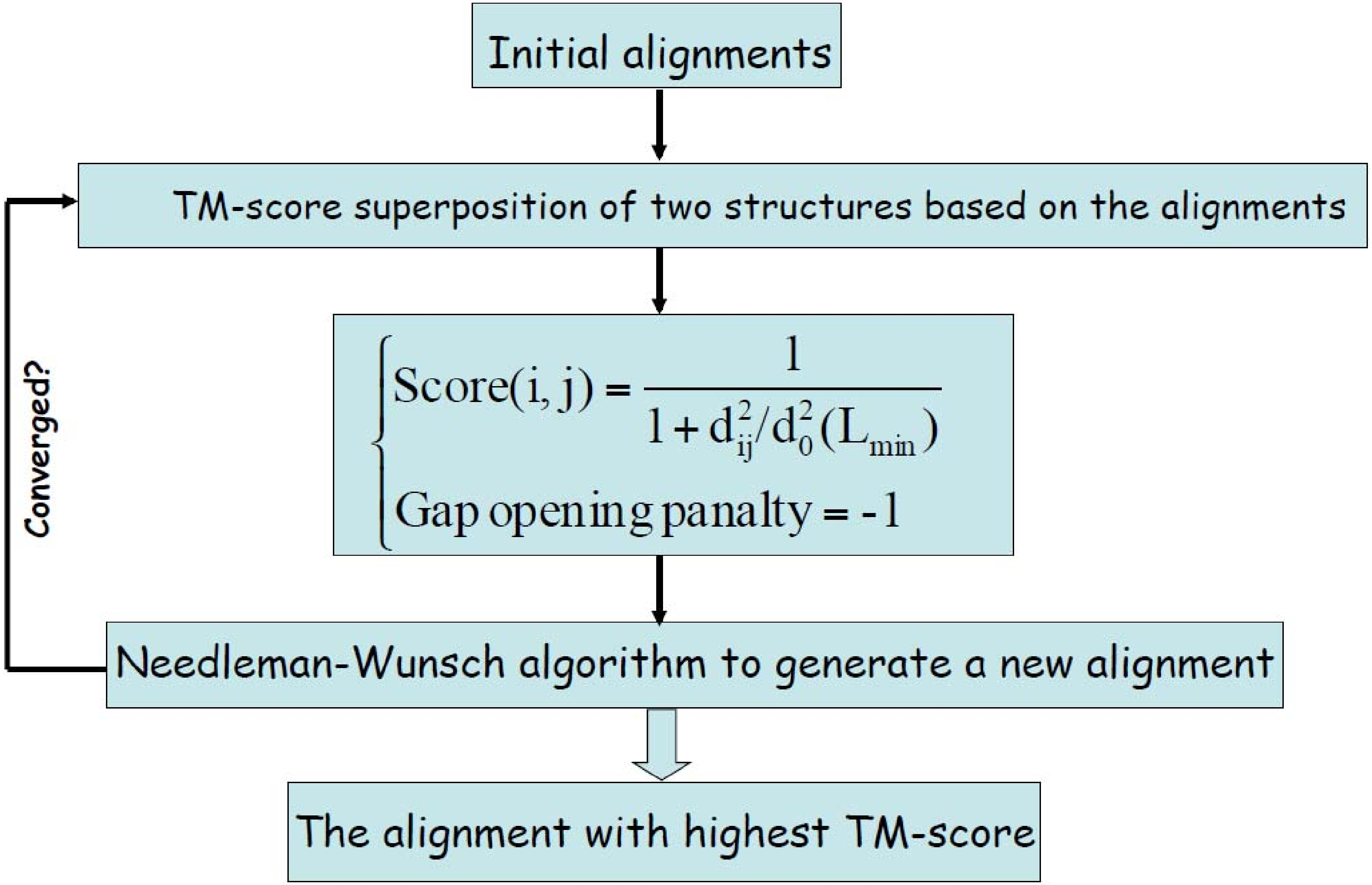
Abstract We have developed TM-align, a new algorithm to identify the best structural alignment between protein pairs that combines the TM-score rotation matrix and Dynamic Programming (DP). The algorithm is~ 4 times faster than CE and 20 times faster than DALI ...

☆ 99 被引用次数：1352 相关文章 所有 25 个版本

TM-align

Objective function: TM-score

Method: Heuristic-based iterative DP



Initial alignments

TM-score superposition of two structures based on the alignments

$$\begin{cases} \text{Score}(i,j) = \dfrac{1}{1 + d_{ij}^2/d_0^2(L_{\min})} \\ \text{Gap opening panalty} = -1 \end{cases}$$

Needleman-Wunsch algorithm to generate a new alignment

The alignment with highest TM-score

TM-align

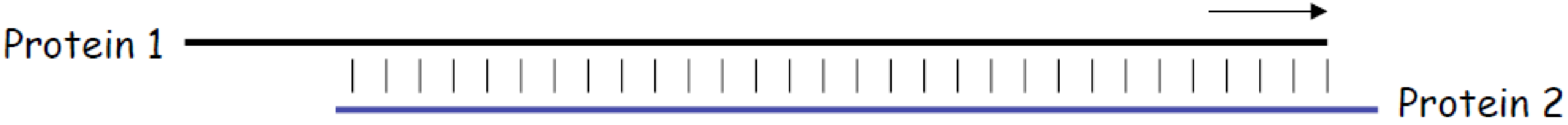
How to generate initial alignments?

1, Needleman-Wunsch alignment of secondary structure

$$Score_{SS}(i,j)=\begin{cases}1, & if\ \ S_i = S_j \\ 0, & \text{otherwise}\end{cases}$$

and gap opening penalty = -1

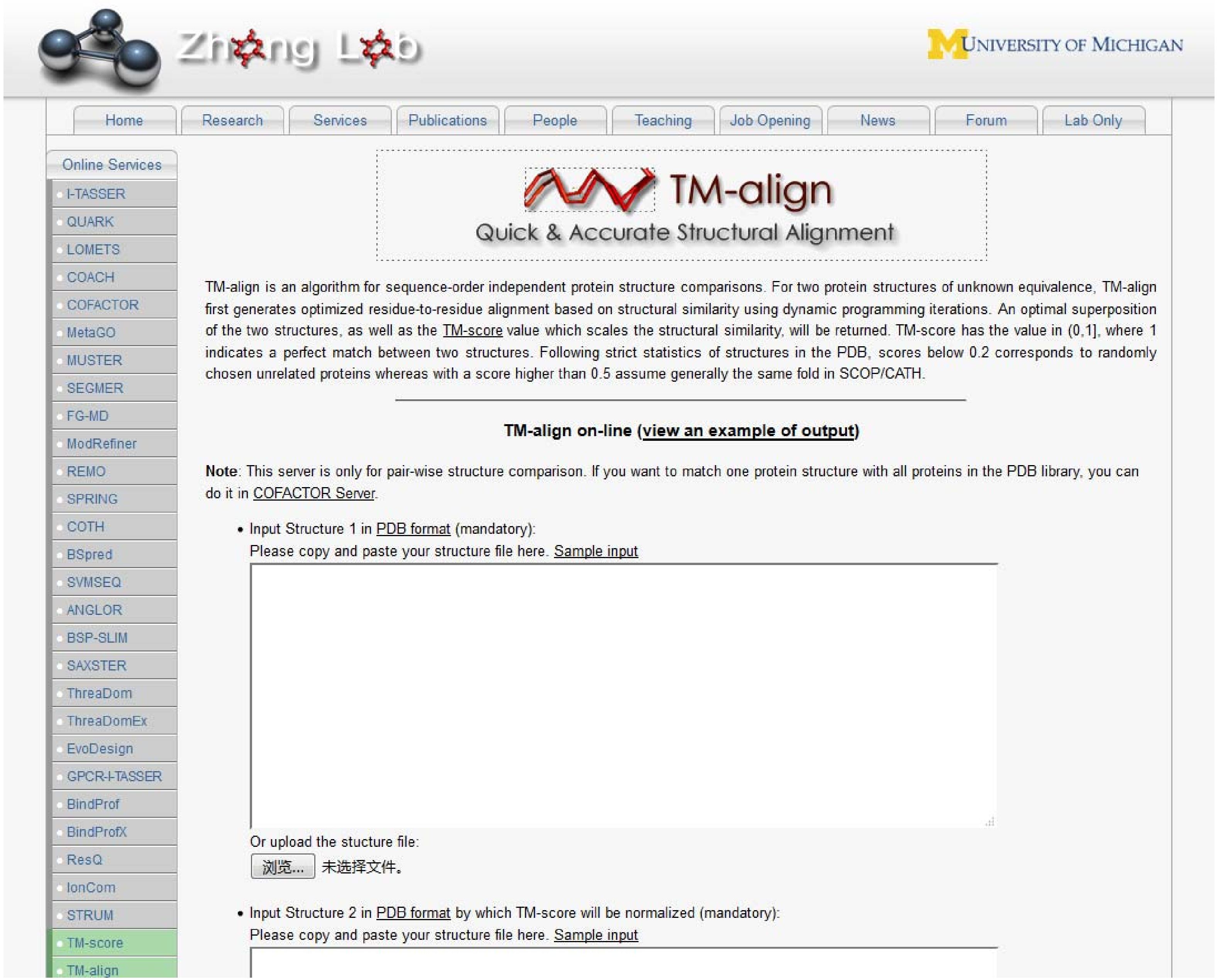
2, Best TM-score from gapless alignment of two target structures:



3, Needleman-Wunsch alignment of combined scores:

$$Score(i,j)=\frac{1}{1+\left(d_{ij}/d_0\right)^2}\ +\text{Score}_{SS}\,\text{(i,j)}$$

where d_{ij} is the distance matrix found from 2 https://zhanglab.ccmb.med.umich.edu/TM-align/



TM-align

Quick & Accurate Structural Alignment

TM-align is an algorithm for sequence-order independent protein structure comparisons. For two protein structures of unknown equivalence, TM-align first generates optimized residue-to-residue alignment based on structural similarity using dynamic programming iterations. An optimal superposition of the two structures, as well as the TM-score value which scales the structural similarity, will be returned. TM-score has the value in (0,1], where 1 indicates a perfect match between two structures. Following strict statistics of structures in the PDB, scores below 0.2 corresponds to randomly chosen unrelated proteins whereas with a score higher than 0.5 assume generally the same fold in SCOP/CATH.

TM-align on-line (view an example of output)

Note: This server is only for pair-wise structure comparison. If you want to match one protein structure with all proteins in the PDB library, you can do it in COFACTOR Server.

- Input Structure 1 in PDB format (mandatory):
  Please copy and paste your structure file here. Sample input

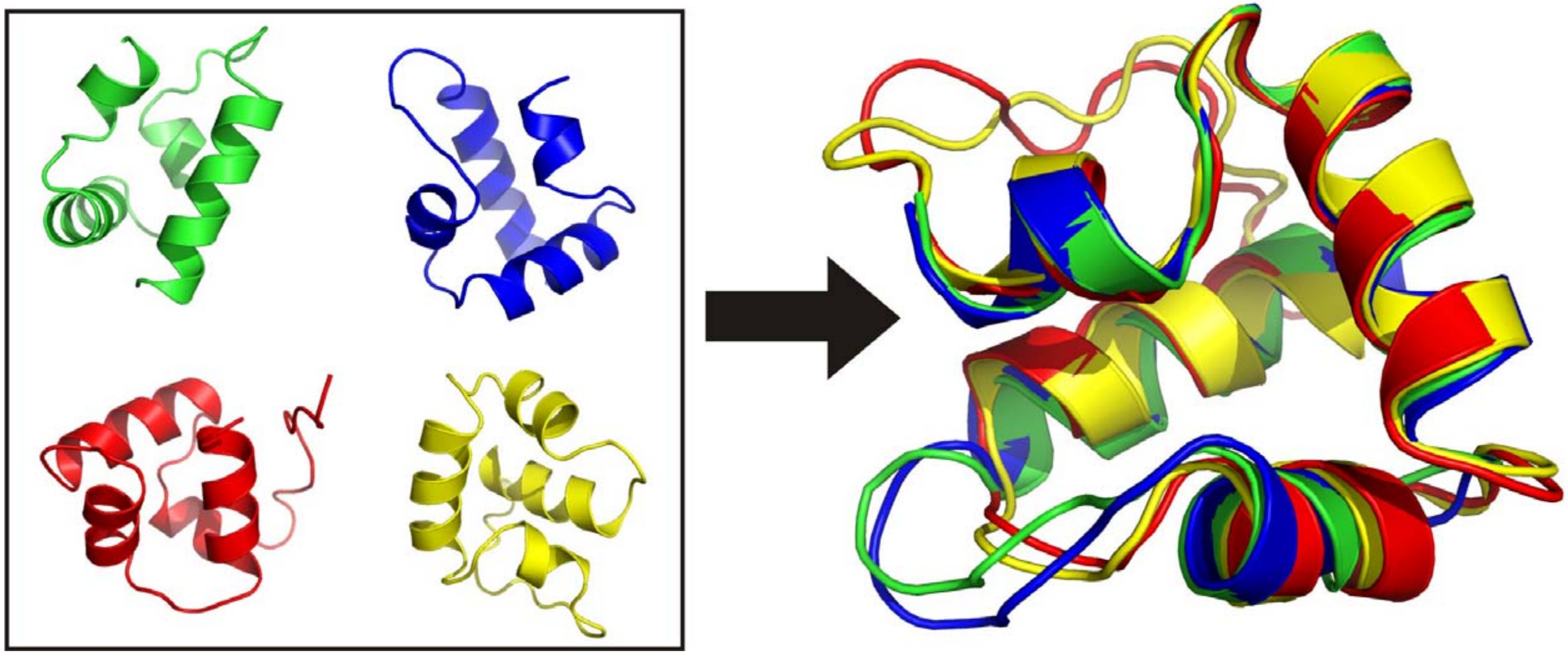
  Or upload the structure file:
  浏览... 未选择文件。

- Input Structure 2 in PDB format by which TM-score will be normalized (mandatory):
  Please copy and paste your structure file here. Sample input

Content

1. What is structure superposition?
   - a. RMSD
   - b. TM-score
2. What is structure alignment?
3. Different structure alignment algorithms
   1. DALI
   2. CE
   3. TM-align
4. Multiple protein structure alignment

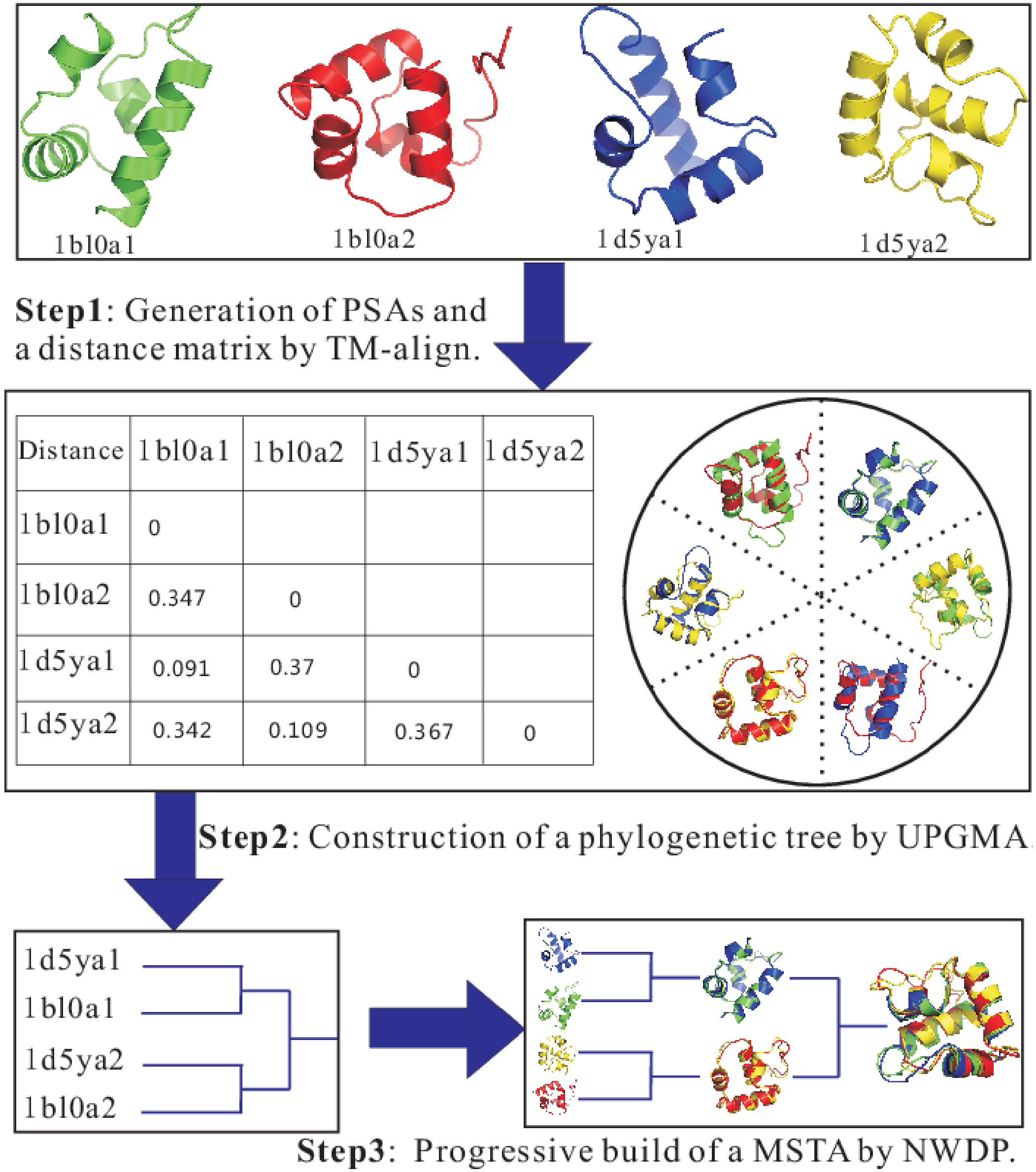
Multiple protein structure alignment



- Matt
- MAMMOTH-mult
- MUSTANG
- MultiProt

...

mTM-align



Step1: Generation of PSAs and a distance matrix by TM-align.

Distance	1bl0a1	1bl0a2	1d5ya1	1d5ya2
1bl0a1	0			
1bl0a2	0.347	0		
1d5ya1	0.091	0.37	0	
1d5ya2	0.342	0.109	0.367	0

Step2: Construction of a phylogenetic tree by UPGMA.

Step3: Progressive build of a MSTA by NWDP.

Dong et al, **Bioinformatics**, 2017

Scoring in mTM-align

- Pairwise structure similarity: TM-score

$$\text{TM-score} = \max \frac{1}{L} \sum_{i=1}^{N_{ali}} \frac{1}{1+\left(d_i / d_0\right)^2}, \qquad d_0 = \begin{cases} 1.24 \times \sqrt[3]{L-15} - 1.8, & \text{if } L > 21 \\ 0.5, & \text{otherwise} \end{cases}$$

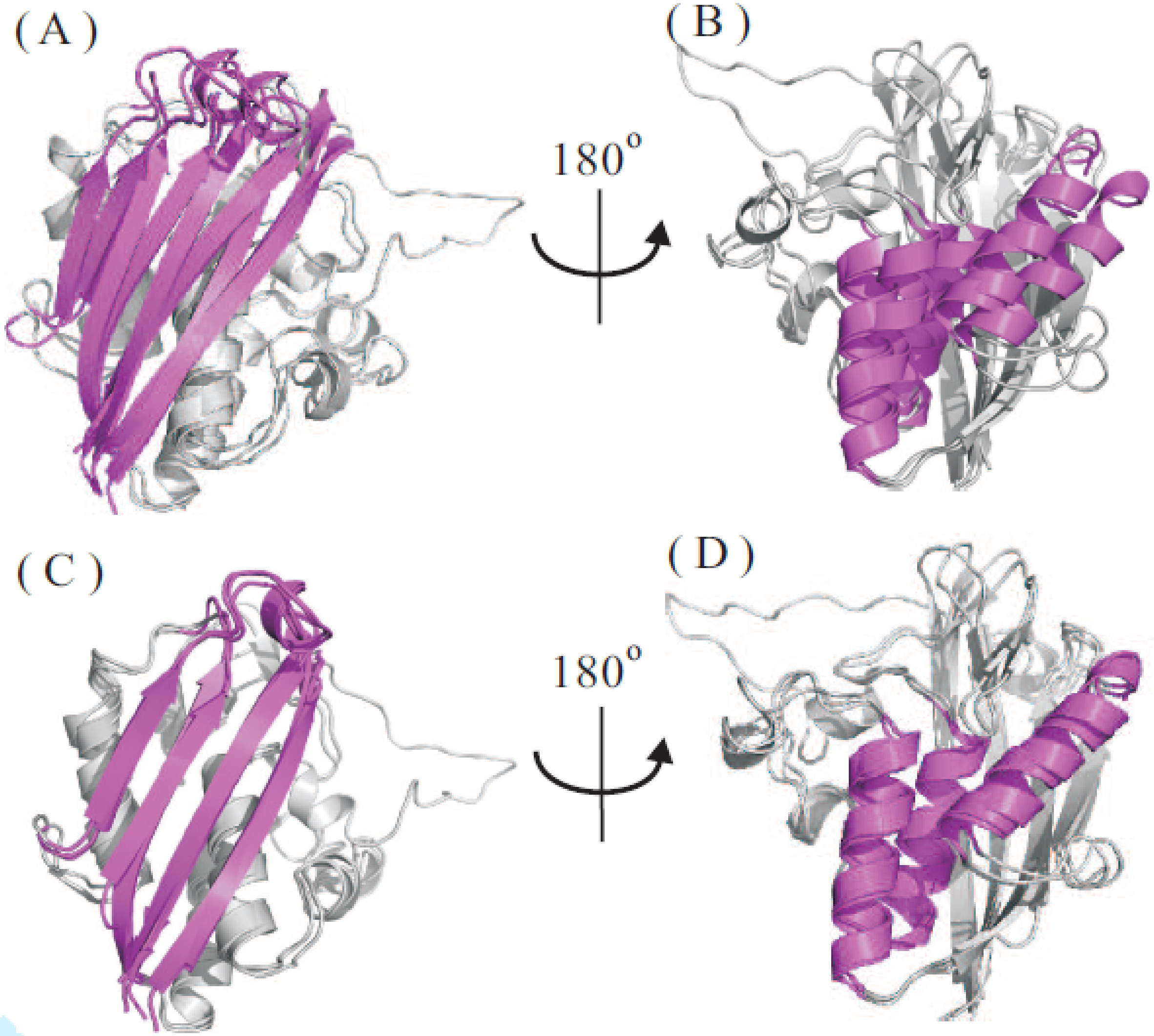
- Scoring in dynamic programming: $S(i, j) = \sum_{m=1}^{M} \sum_{n=1}^{N} s(i_m, j_n)$

$$s(i_m, j_n) = \begin{cases} 1/(1+(\dfrac{d(i_m, j_n)}{d_0})^2), & \text{if } d(i_m, j_n) < d_{cut} \\ s, & \text{otherwise} \end{cases}$$

$$s = \begin{cases} -0.1 \times (1 - e^{d_{cut} - d(i_m, j_n)}), & \text{if } \overline{\text{TM-score}} > 0.5 \\ -0.1 \times (1 - e^{d_{cut} - d(i_m, j_n)}) / b(i_m, j_n), & \text{otherwise} \end{cases}$$

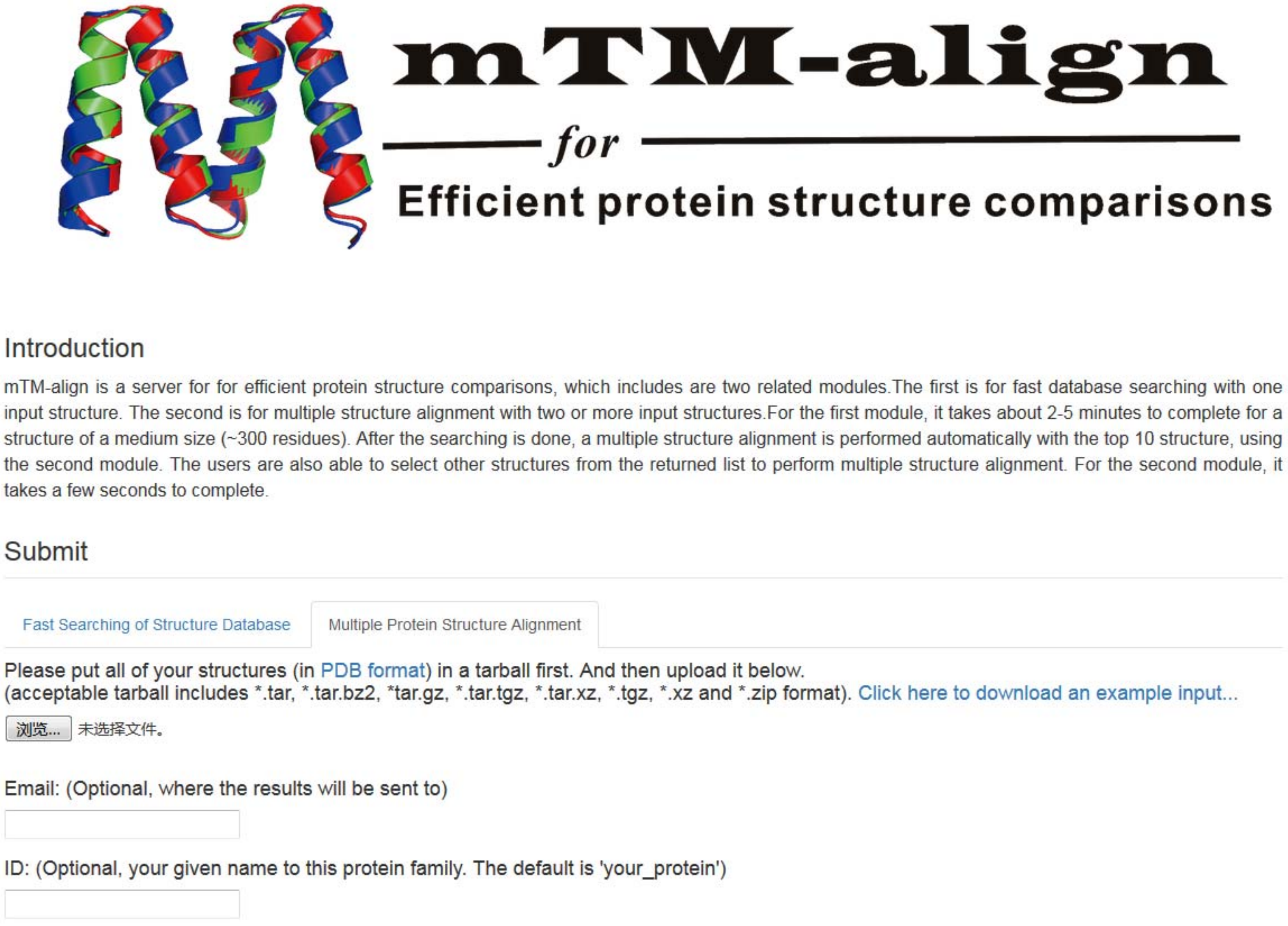
Comparison with manual alignments

A,B: manual
C,D: mTM-align



HOMSTRAD 'YgbB' family

http://yanglab.nankai.edu.cn/mTM-align/



mTM-align

for

Efficient protein structure comparisons

Introduction

mTM-align is a server for for efficient protein structure comparisons, which includes are two related modules.The first is for fast database searching with one input structure. The second is for multiple structure alignment with two or more input structures.For the first module, it takes about 2-5 minutes to complete for a structure of a medium size (~300 residues). After the searching is done, a multiple structure alignment is performed automatically with the top 10 structure, using the second module. The users are also able to select other structures from the returned list to perform multiple structure alignment. For the second module, it takes a few seconds to complete.

Submit

Fast Searching of Structure Database
Multiple Protein Structure Alignment

Please put all of your structures (in PDB format) in a tarball first. And then upload it below.
(acceptable tarball includes *.tar, *.tar.bz2, *tar.gz, *.tar.tgz, *.tar.xz, *.tgz, *.xz and *.zip format). Click here to download an example input...

Email: (Optional, where the results will be sent to)

ID: (Optional, your given name to this protein family. The default is 'your_protein')

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
7. Protein function prediction

Papers to read

Secondary structure prediction

Jones, D., 1999. Protein secondary structure prediction based on position-specific scoring matrices. **J. Mol. Biol** 292, 195-202.

Threading

J. U. Bowie, R. Luthy, D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. **Science**. (1991) 253:164-170.

Rosetta

K. T. Simons, C. Kooperberg, E. Huang, D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. **J Mol Biol**. 1997 Apr 25;268(1):209-25.