今期から始まった遠隔授業の受講者の方々にも対応するため、早めにファイルを置いていますが、授業直前に内容を修正する可能性があります。2010年4月19日

#14

Protein Structure Prediction

Topics:

- Needs for Protein Structure Prediction
- Preparation: Protein Structure *Comparison*

  RMSD, RMSDd, double dynamic programming
- Structure Prediction: simple Lattice model
- Homology modeling

  Modeller, Swiss-model, SCWRL
- Threading method

  Sippl (Threading) , Bowie-Eisenberg (3D-1D), Jones (Double DP)


SPring-8 synchrotron facility
X-ray
Diffraction
Analysis

RIKEN GSC NMR facility

Nucleic
Magnetic
Resonance
Spectroscopy

Protein Data Bank

- Tertiary (3-D) structure archive of proteins, DNAs, and complexes.
  58,236 entries (as of 16h June, 2009)

PROTEIN DATA BANK

PDB Current Holdings Breakdown

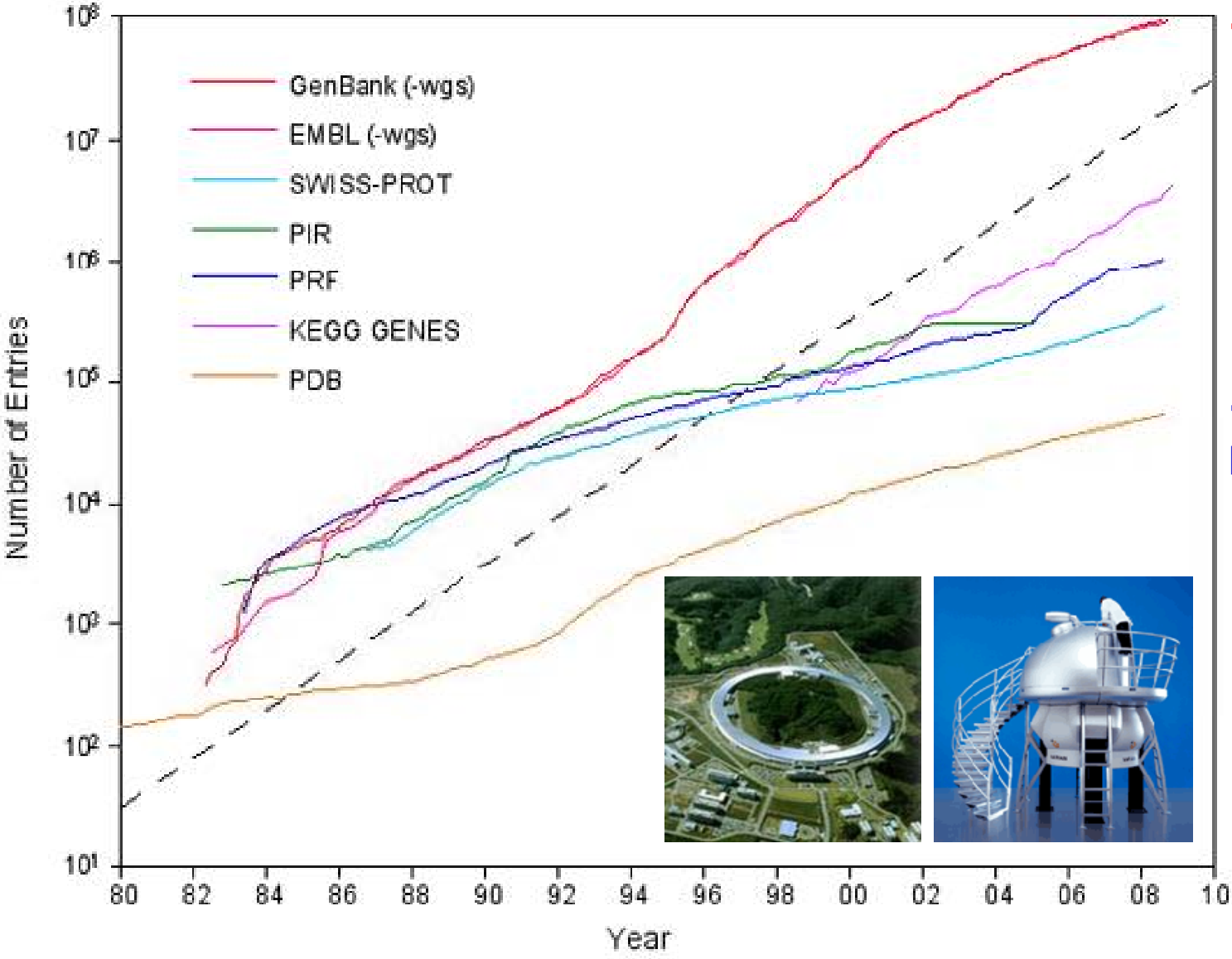
		Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
Exp. Method	X-ray	46626	1147	2163	17	49953
	NMR	6886	856	146	6	7894
	Electron Microscopy	168	16	59	0	243
	Hybrid	14	1	1	1	17
	Other	112	4	4	9	129
	Total	53806	2024	2373	33	58236

http://www.rcsb.org

- history

1971 Started at Brookhaven National Laboratory
1998 Moved to RCSB (Research Collaboratory for Structural Bioinformatics)
2006 wwPDB (The Worldwide Protein Data Bank) by US, Europe, and Japan.

Growth of Sequence and 3D Structure Databases



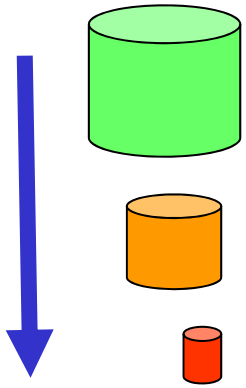
100-Fold / 10yr
DNA seq.

10-Fold / 10yr
Protein Strct.

Kyoto University GenomeNet WWW site (Database Growth)

Huge Cost Gaps among Data Types

- **DNA Seq**. information is **inexpensively** read by sequencers
- **Protein 3-D structure** information is **extremely expensive**



type	DB	version	#data entries
DNA	nr-nt	2009-07-24	108,589,156
			← 1/12
Protein	nr-aa	2009-07-24	9,080,960
			← 1/150
3-D	pdb	2009-07-20	58,886

2000 : 1

Bridging by Bioinformatics?
Not by expensive experiments.

Statistics information from:
www.genome.ad.jp/dbget-bin/binfo/

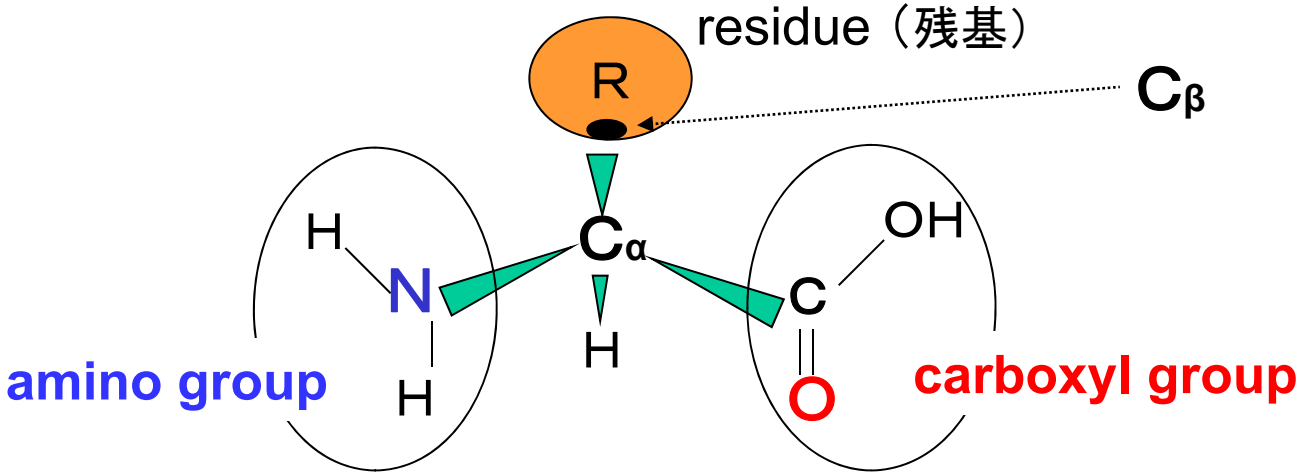
DNA Sequencer

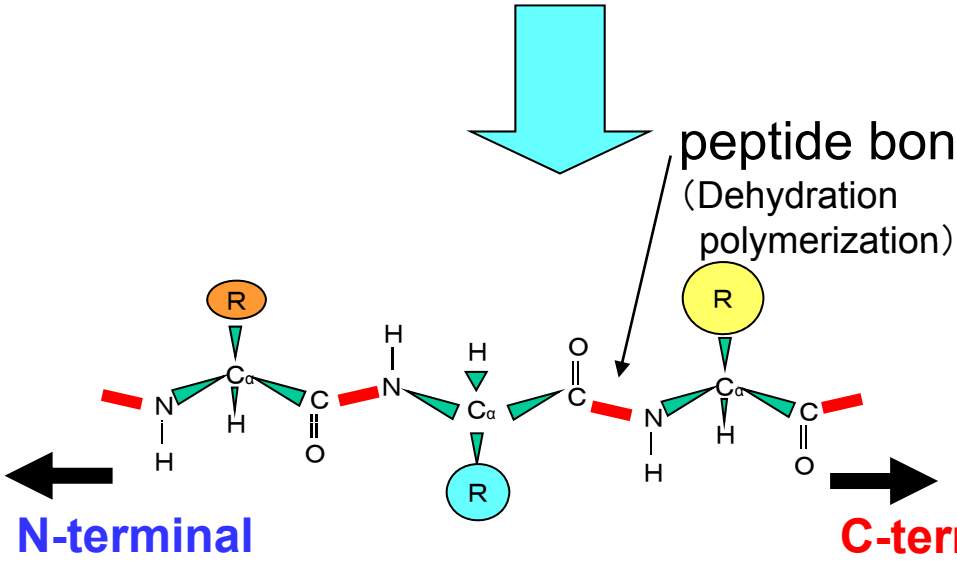
Structure Analysis（X-ray、NMR）

Protein Structure

Review

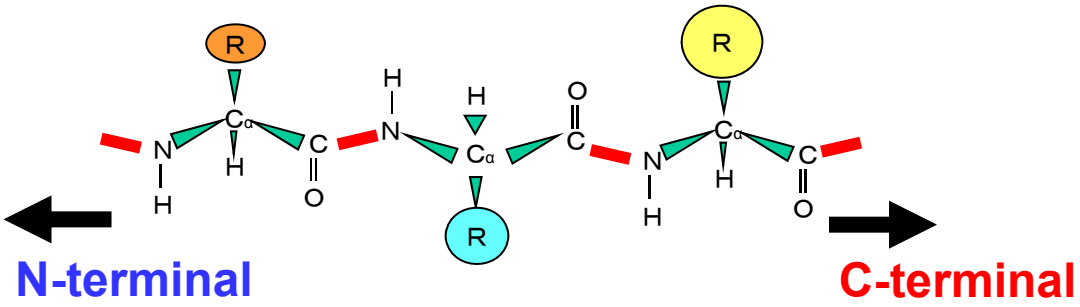
・Amino acid as a building block

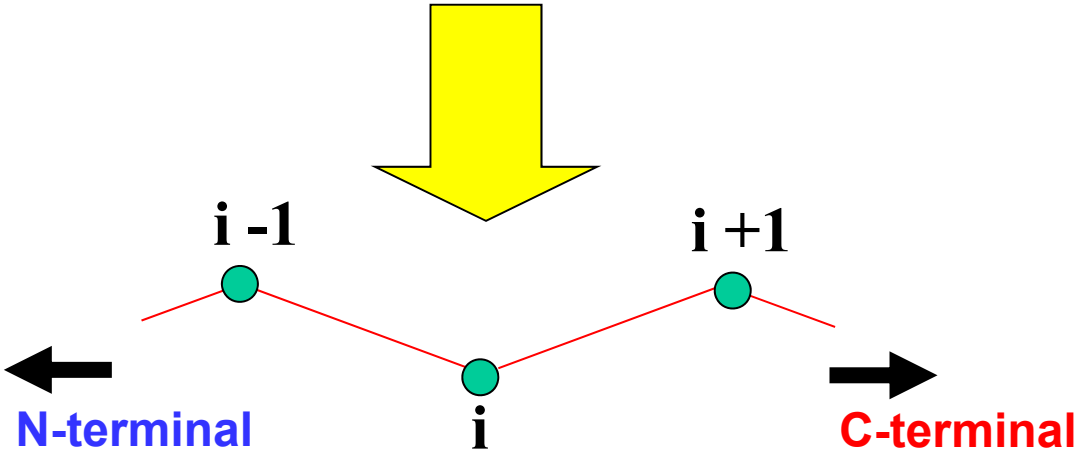


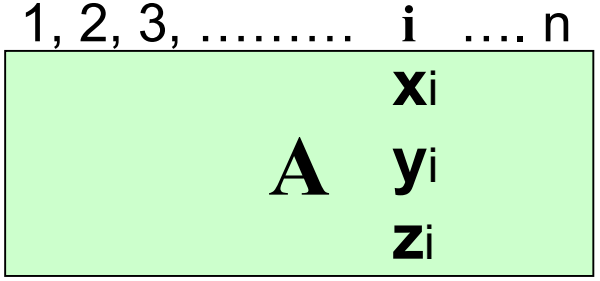


・S-S (disulfide bridge)
・Post-translational modification
acetylation,
phosphorylation,
sugar chain addition
・Dimerization, etc.

C_α coordinates: a simplified representation of main chain







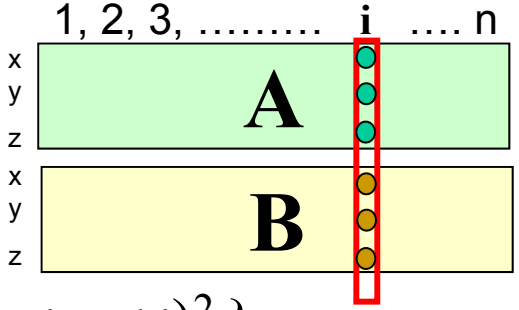
3 × n Matrix

RMSD

RMSD (Root Mean Square Deviation)

$$\mathrm{RMSD}(A, B) = \mathrm{SQRT}\left\{ \frac{1}{n} \cdot \sum_{i=1}^{n} (\mathbf{A}i - \mathbf{B}i)^2 \right\}$$

$$= \mathrm{SQRT}\left\{ \frac{1}{n} \cdot \sum_{i=1}^{n} (xai - xbi)^2 + (yai - ybi)^2 + (zai - zbi)^2 \right\}$$



In bioinformatics, “RMSD” usually means for the following “least” RMSD between superposed (protein) coordinates A and B.

$$\text{least-RMSD}(A, B) = \min_{\boldsymbol{U}} \mathrm{SQRT}\left\{ \frac{1}{n} \cdot \sum_{i=1}^{n} (\boldsymbol{U}\,\mathbf{A}i - \mathbf{B}i)^2 \right\}$$

***U*: orthonormal (rotational) matrix**
正規直交（回転）行列

where coordinates **A, B** are centered beforehand.

Best rotation matrix (1)

Kabsch's method

W. Kabsch: "A Solution for the Best Rotation to Relate Two Sets of Vectors", *Acta Crystallographica*, ***32***, 922-923 (1976).

Covariance matrix (3×3)

$$C = A\,B^T$$

Perform Singular Value Decomposition

$$C = V\,S\,W^T$$

where S is composed of eigenvalues(固有値) λ_1, λ_2, and λ_3

The best rotation Matrix U is obtained as

$$U = W\ \ V^T$$

(if det(C)<0, then mirror reflection is also needed)

Best rotation matrix (2)

Quaternion method

E. Coutsias, C. Seok, K. Dill: “Using quaternions to calculate RMSD”, Journal of Computational Chemistry, **25**, 1849-1857 (2004).

Quaternion (四元数)

$$r = w + x\,i + y\,j + z\,k$$

$$i^2 = j^2 = k^2 = -1$$
$$i \cdot j = k, \quad j \cdot k = i, \quad k \cdot i = j$$
$$j \cdot i = -k, \quad k \cdot j = -i, \quad i \cdot k = -j$$

Rotation by unit quaternion

$$r = x\,i + y\,j + z\,k$$

$$q = a + b\,i + c\,j + d\,k$$
$$q^* = a - b\,i - c\,j - d\,k$$

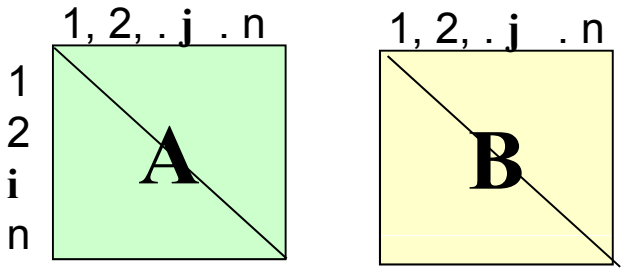
$$r' = q\ r\ q^* \quad \text{(rotation by unit quaternion)}$$

$$\max_{q} \sum_{i=1}^{n} (q\,\mathbf{A}i\,q^* - \mathbf{B}i)^2$$

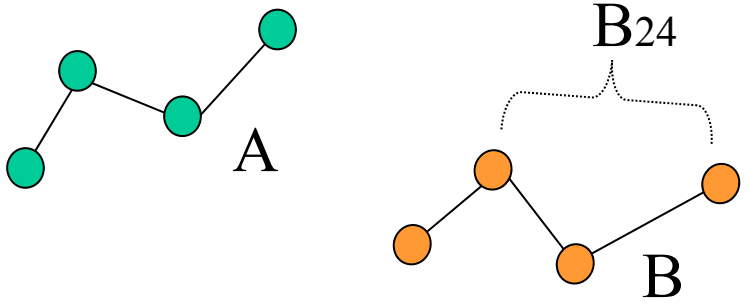
distance based RMSD

RMSDd

$$\mathrm{RMSD}(A, B) = \mathrm{SQRT}\left\{ \frac{1}{n\,(n-1)} \cdot \sum_{i=1}^{n} \sum_{j=1}^{n} (\mathbf{A}ij - \mathbf{B}ij)^2 \right\}$$



distance matrices



advantage:

1. Robust for outliers (while normal RMSD tends to be influenced)
2. Easy to calculate

disadvantage:

1. Mirror reflection images cannot be distinguished.

DALI

Structure comparison based on RMSDd

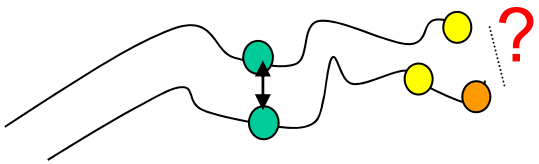
Lisa Holm, Chris Sander: *"Protein structure comparison by alignment of distance matrices"*, J Mol Biol, **233**, 123-138 (1993

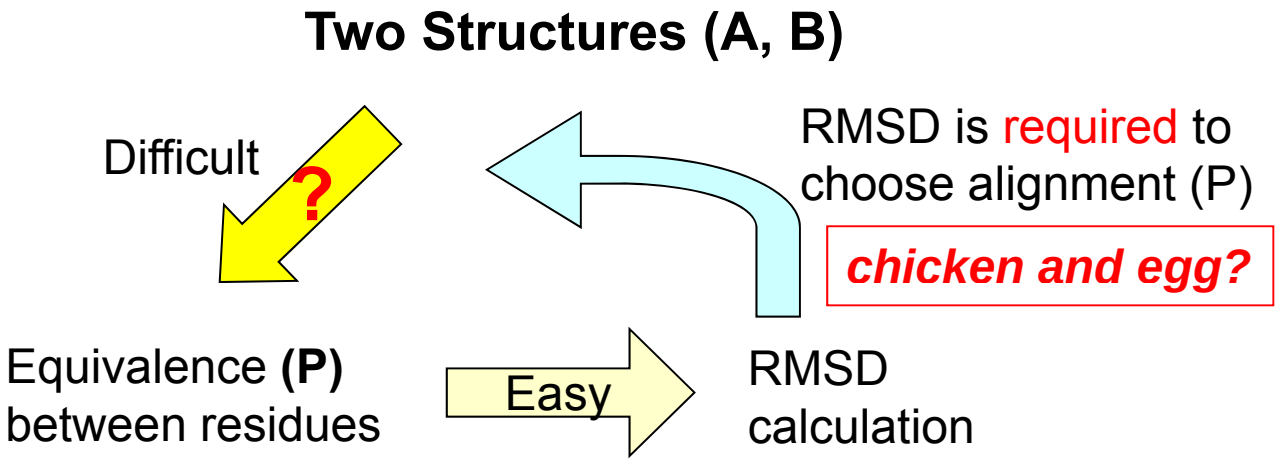
DALI server (**D**istance matrix **ALI**gnment)

http://www2.ebi.ac.uk/dali/

- Not a direct calculation with whole protein length.
- Partial comparisons are done, and then combined.
- DALI system is used to define FSSP database (shown later).

Structure Alignment



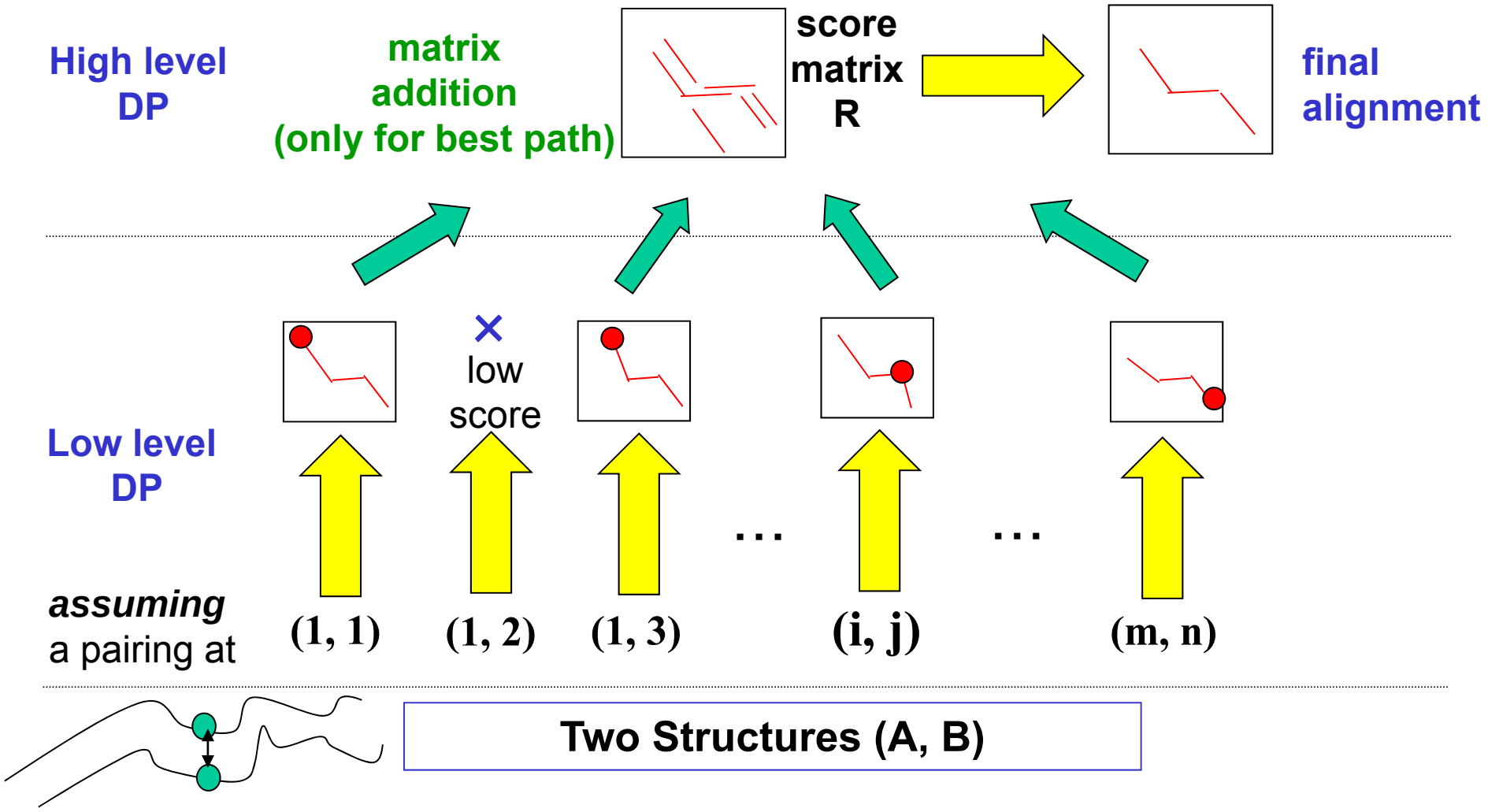


A simple iterative approach

1) start from initial given alignment P.
2) calculate least RMSD rotation, based on the alignment P.
3) measure residue distances and make a score table Rij.
4) perform dynamic programming between (A,B) with Rij.
5) get new alignment P'. (if P' is not enough, P=P' and go to 2)

S.T. Rao, M.G. Rossman: "*Comparison of super-secondary structures in proteins*," J. Mol. Biol., **76**, 211-256 (1973).

Double Dynamic Programming



C. Orengo and W. Taylor: "*SSAP: Sequential Structure Alignment Program for Protein Structure Comparison*", Methods in Enzymology **266**, 617-635 (1996).

Algorithm for DDP

R := { 0 }

for each pair (Ai, Bj) **do**
  force pairing between Ai and Bj
  compute the low level scoring matrix $R^{(i,j)}$
  (score, P) := DP (A, B) with $R^{(i,j)}$
  forall (Ap, Bq) in P
    do Rpq := Rpq + $R^{(i,j)}$ pq
end

(s, p) := DP(A, B) with R (High level DP)

Cited and modified from “Protein structure comparison and structure patterns” by Ingvar Eidhammer and Inge Jonassen

Geometric Hashing

- Suitable to quickly search similar sub-structures from a large-scale (protein) structure database. In order to do this, a large “hash table” is pre-calculated.

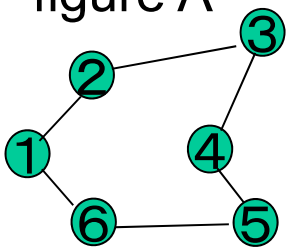
- Originally developed as a 3-D model comparison method in computer vision study.

- Does not care for residue number.
  Only compares vertex positions.

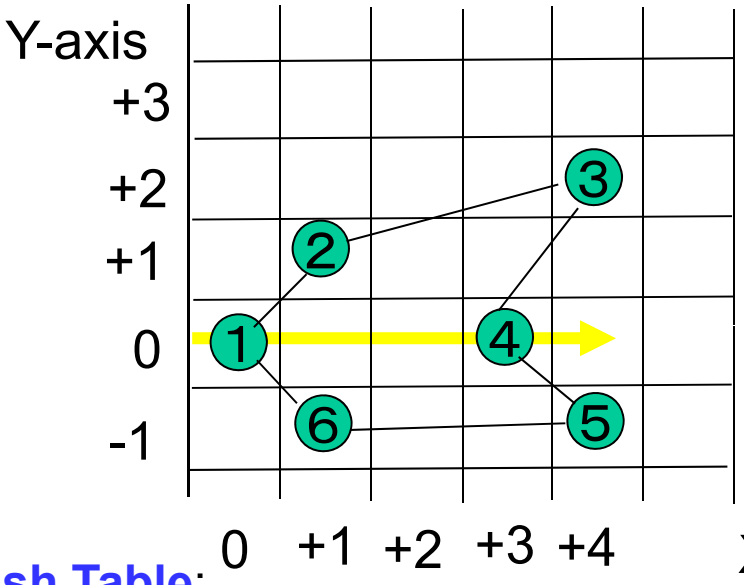
R. Nussinov and HJ Wolfson:
“Efficient Detection of Three Dimensional Structural Motifs in Biological Macromolecules by Computer Vision Techniques”, Proc. Nat’l Acad. Sci., **88**, 10495-10499 (1991).

Geometric Hashing (example)

figure A



For example,
(1) as the origin, (1)-(4) as x-axis.
This is called "(1, 4) transform".



(0, 0) hit !
(1, 1) hit !
(4, 2) hit !
(3, 0) hit !
(4, -1) hit !
(1, -1) hit !

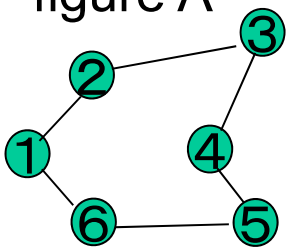
Hash Table:
detailed coordinates within mesh are ignored.

For a figure with n points, there are $z = n(n-1)/2$ different "transform" exist.

Figure A has 6 points.
Thus Z = 6 (6-1) / 2 = 15
15 different "transform" should be tested.

The number,
z = n (n-1) / 2, is much smaller than that of possible smooth analogue transforms.

figure A

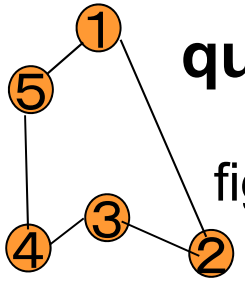


All possible transform are tested, and the hit results are stored in the same table.

Y-axis	0	+1	+2	+3	+4	x-axis
+3				A6, 5		
+2	A6, 5				A1, 4	
+1		A1, 4	A6, 5			
0	A1, 4 A6, 5			A1, 4 A6, 5		
-1		A1, 4			A1, 4	

query

figure B



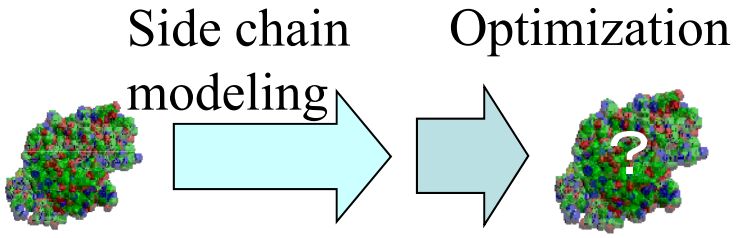
All possible z= m(m-1)/2 transform are sequentially tested. Now for example (1,3).

Now (1,3) map are compared with the pre-calculated hash table of figure A.
B(1,3) == A(1.4) ◎

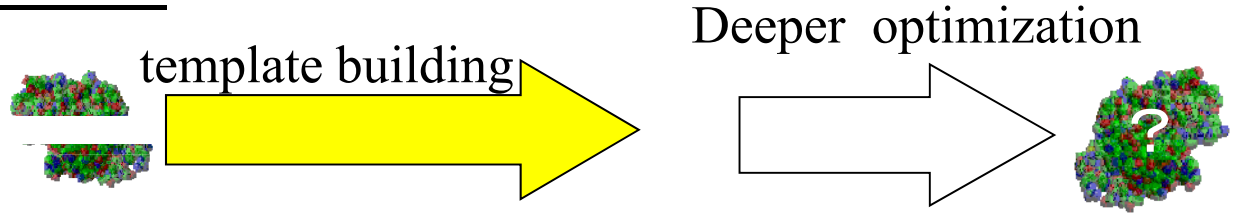
pre-calculated **hash table** can be an overlay of thousands different figures in a database. all figures are searched at once.

Three approaches for structure prediction

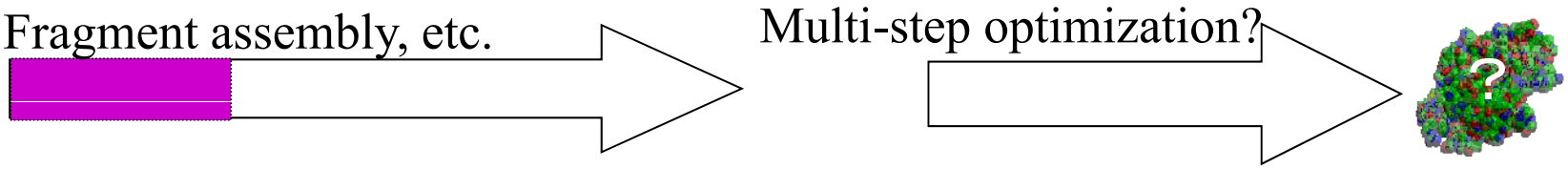
- Homology modeling



- Fold recogniton

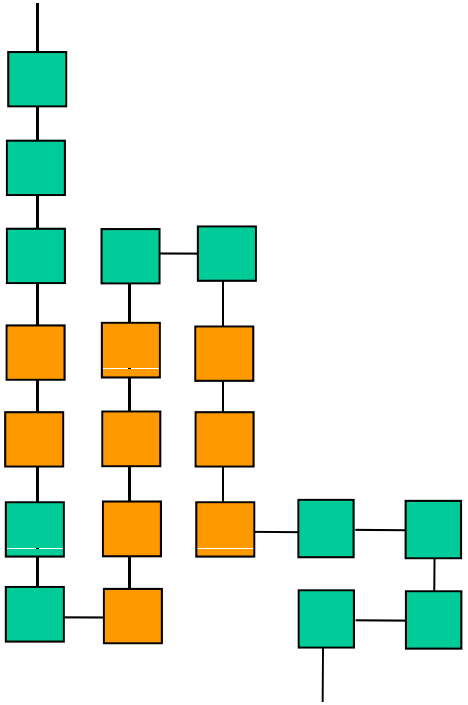


- New fold *("ab initio")*



Lattice model

- Extremely simplified model
- Easy to find out the minimum energy conformation



HP model
Hydrophobic ( ) vs. Polar ( )

Lattice: cubic, tetrahedron, etc.

Node: single residue

Energy (only between neighboring nodes):
1) steric, 2) hydrophobic, 3) hydrogen-bond, etc.

tetrahedron lattice model (green) vs. native BPTI structure (red)

D. Hinds and M. Levitt:"A lattice model for protein structure prediction at low resolution", PNAS, 89, 2536-2540 (1992).

Homology modeling

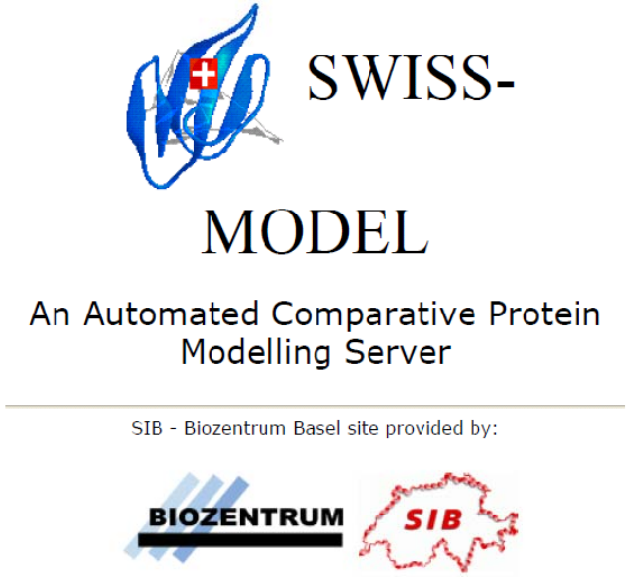
- MODELLER by Andrej Sali

Main chain: template
Side chain: modeling

http://www.salilab.org/modeller/

A. Sali, T.L. Blundell: “Comparative protein modelling by satisfaction of spatial restraints”, J. Mol. Biol. 234, 779-815 (1993).

- SWISS-MODEL by SIB



http://swissmodel.expasy.org/

Schwede T, Kopp J, Guex N, and Peitsch MC: “SWISS-MODEL: an automated protein homology-modeling server”, Nucleic Acids Research, 31, 3381-3385 (2003).

Side chain prediction

Input: Main chain structure
Output: Side chain structure (optimized rotamer arrangement)

Rotamer Library:
Collection of frequently observed side chain dihedral angles (for each residue, with/without main-chain dependent situation)

Combinatorial Optimization:
Maximizing total fitness of mutual relation among selected rotamers. Algorithms like **DEE** (Dead-End Elimination), etc.

- SCWRL (Library & Software) by Roland Dunbrack

http://dunbrack.fccc.edu/SCWRL3.php

A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack, Jr. : “A graph theory algorithm for protein side-chain prediction”, Protein Science, 12, 2001-2014 (2003).

How many protein folds exist?

estimation of protein fold numbers

- G.E. Schulz: *Angew. Chem. Int. Ed. Eng.* (1981) ~ 500
- C. Chothia: *Nature*, 357, 543-544. (1992) ~1000

"One Thousand Families for the Molecular Biologist"

At 1992, there were 120 known unique folds in PDB.

At the same time, one quarter of known protein sequences were predicted to have one of 120 known folds, while other 3/4 are unknown.

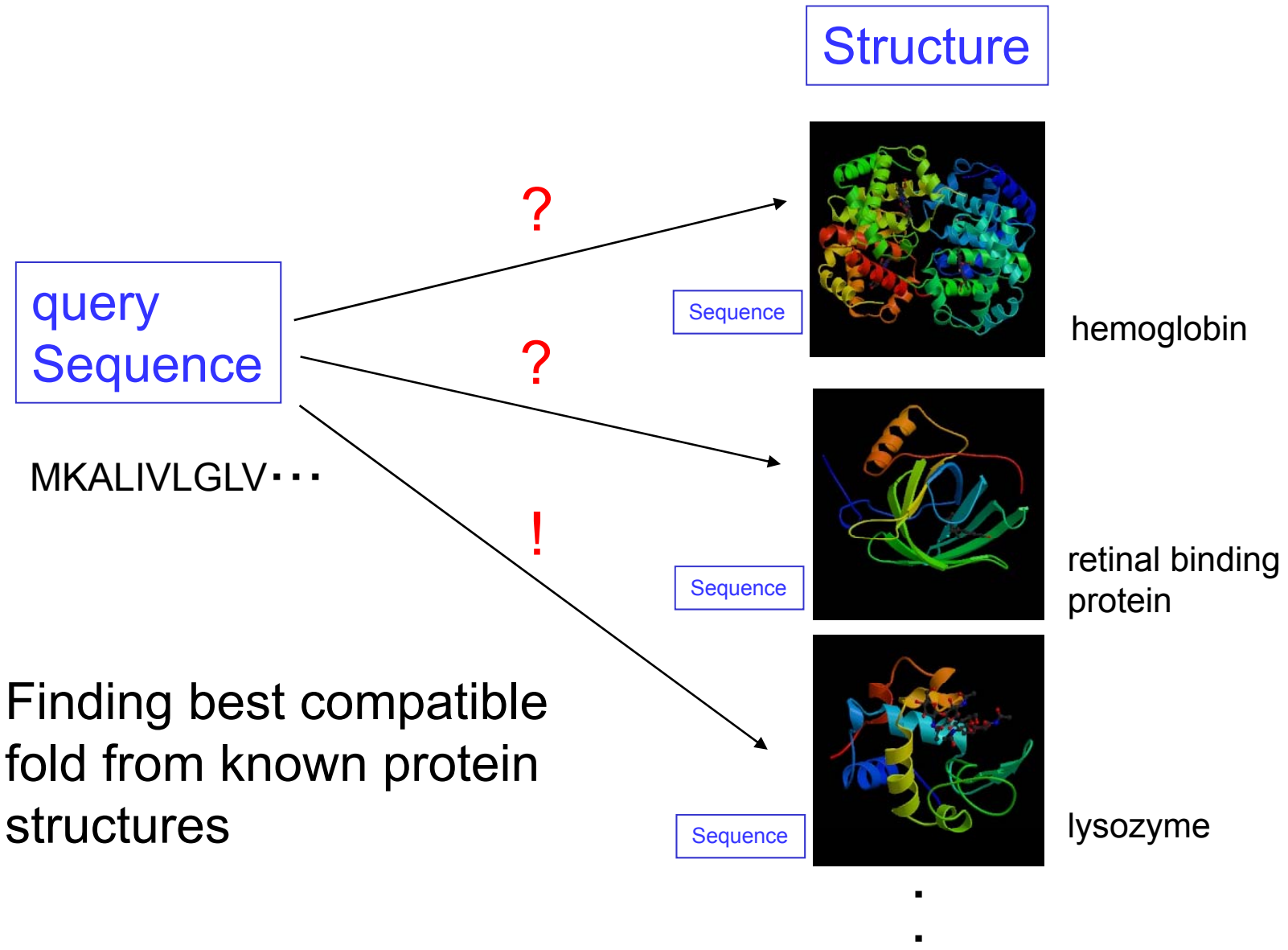
Chothia estimated that only 1/3 of existing proteins had been sequenced at the time. Thus, the first rough estimation for unique folds are given by

$120 \times 4 \times 3 = 1440$

Sequence homology search method must have a limited sensitivity, thus, finally he estimated as: $120 \times (4 / 1.25) \times (3 / 1.25) = 920$.

(Recently researchers are believing > 3000 folds? exist)

Fold Recognition

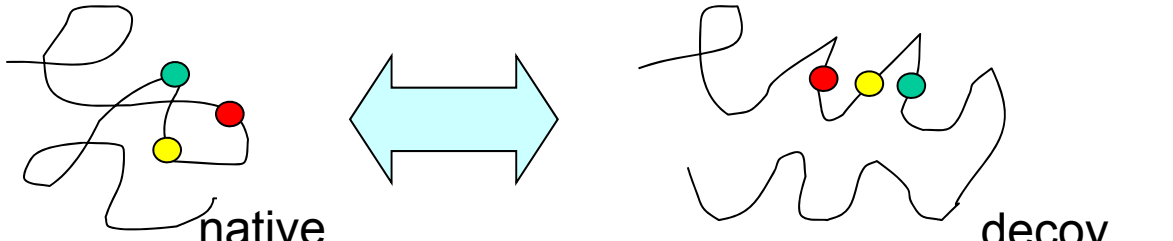


Finding best compatible fold from known protein structures

Potential-based Threading

Recognizing native structure by calculating pseudo energy potential among residues with "threading" the sequence into a 3D model.

Manfred Sippl



M. Sippl :

"Calculation of Conformational Ensembles from Potentials of Mean Force,- An Approach to the Knowledge-based Prediction of Local Structures in Globular Proteins- ", J. Mol. Biol., 213, pp.859-883 (1990).

M. Hendlich, P. Lackner, S. Weitckus, …. and M. Sippl :

"Identification of Native Protein Folds Amongst a Large Number of Incorrect Models, - The Calculation of Low Energy Conformations from Potentials of Mean Force ", J. Mol. Biol., 216, pp.167-180 (1990).

Inverse Folding

J. Bowie, R. Luthy, D. Eisenberg: "A Method to Identify Protein Sequences That Fold into a Known Three Dimensional Structure", Science, 253, 164-170 (1991).

Inverse Folding problem

Find amino acid sequences that are compatible to a given 3-D structure.

Search the most compatible sequence with residue "environment" in a given protein 3-D structure. "Environment" includes:

(1) **A**: surface area which buried in the protein, and not exposed.

(2) **f**: fraction of side chain surface area covered by polar atoms (O, N)

(3) **s**: local secondary structure

Four examples are shown in this paper.

- globins
- cyclic AMP receptor-like proteins
- periplasmic binding proteins
- actins

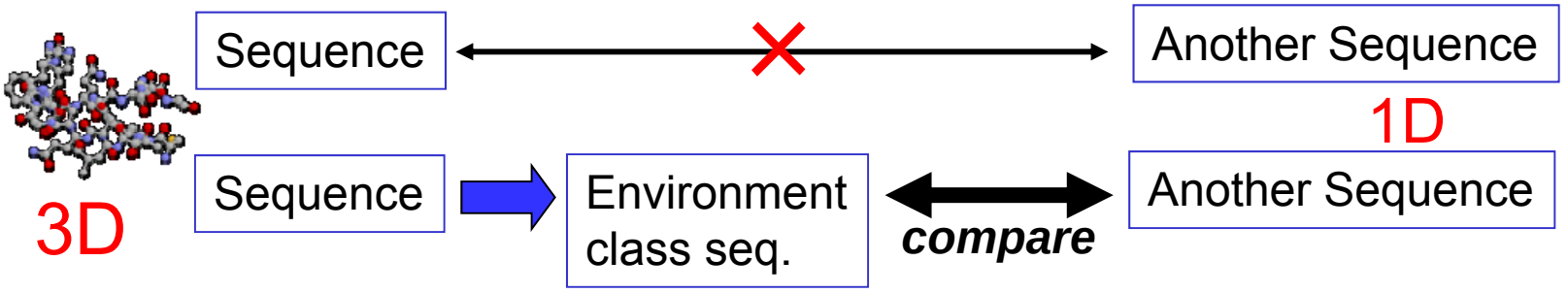
David Eisenberg

3D-1D score

3D compatibility search

N - X1- X2- X3- X4- … Xn- C (amino acid sequence)

E_{α}- $P_{2\alpha}$- $B_{2\alpha}$- E_{α}- … (environment class)



3D-1D score:

$$\text{score} = \sum_{i=1}^{n} f(\text{class}_i, \text{residue}_i)$$

Best alignment between the 3-D sequence and another sequence can be efficiently calculated with **Dynamic Programming** technique

3D Profile

Position in fold	Environment class	Amino acid type A	C	D	E	F	G	...	R	S	T	V	W	Y	Gap penalty Opn	Ext
1	E	12	-48	22	3	-190	113	...	-32	32	12	-91	-214	-94	2	0.02
2	B_2	-68	-5	-128	-135	185	-160	...	-80	-117	-78	60	102	112	2	0.02
3	E α	46	-44	44	59	-220	68	...	-34	15	-17	-110	-135	-210	200	200
4	P_2α	6	-93	28	56	-143	-50	...	50	-18	-5	-48	-114	-79	200	200
5	E α	46	-44	44	59	-220	68	...	-34	15	-17	-110	-135	-210	200	200
6	P_2α	6	-93	28	56	-143	-50	...	50	-18	-5	-48	-114	-79	200	200
7	B_2α	-68	-10	-152	-71	90	-149	...	6	-147	-150	68	50	86	200	200
8	E α	46	-44	44	59	-220	68	...	-34	15	-17	-110	-135	-210	200	200
9	P_2α	6	-93	28	56	-143	-50	...	50	-18	-5	-48	-114	-79	200	200
10	B_1α	-68	-73	-197	-174	132	-253	...	-167	-273	-129	66	100	18	200	200
.	.	.	.	.	.	.	.		.	.	.		.	.	.	.
.	.	.	.	.	.	.	.		.	.	.		.	.	.	.
.	.	.	.	.	.	.	.		.	.	.		.	.	.	.

・Fig. 3　3D profile example

3D profile of sperm whale myoglobin（original sequence length is 153）
The 3D-1D scores in this table are calculated (x 100) from probability in Fig. 5.
Heavy penalty applies for opening gap in a helix region at position 3-10.

Double DP technique for Threading

D. Jones, W. Taylor, J. Thornton: "A new approach to protein fold recognition", Nature, 358, pp.86-89 (1992).

David Jones

Difficulty of Sippl's approach:

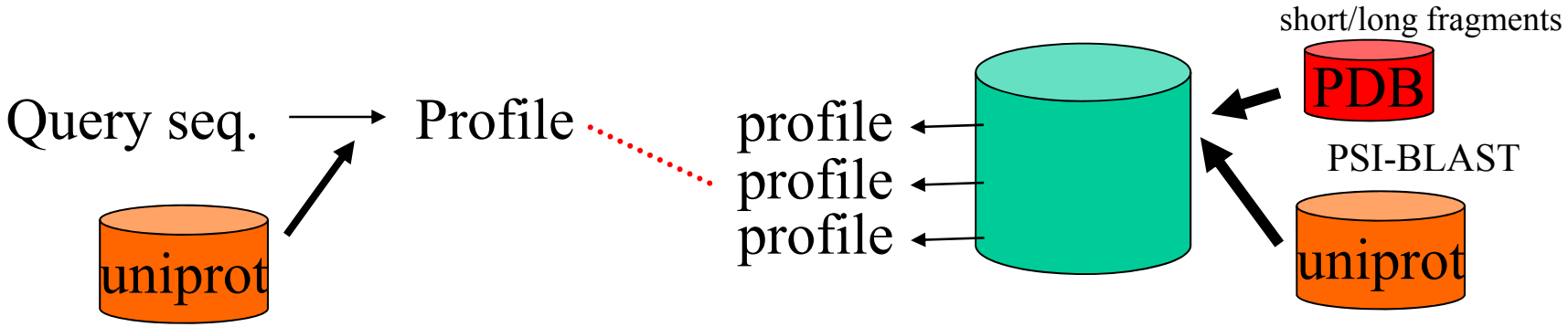
Gap (insertion/deletion) between model and query should be considered.
Pairwise potential can be calculated only after determining residue positions.
Alignment and evaluation phases are like a "Chicken and Egg" problem.

- choice 1: Use sequence alignment between model and query sequence.
  → usually difficult, because high homology cannot be expected.
- choice 2: Give up "pairwise" potential.
  → use "environment" or such (Eisenberg's approach)
- choice 3: Give up real "pairwise" calculation, and use approximation.
  → use "frozen approximation".
  (calculate potential between query residue and model residue)
- choice 4: Try to give good alignment between 3-D model and query.
  → **use "double DP" technique** (This paper)

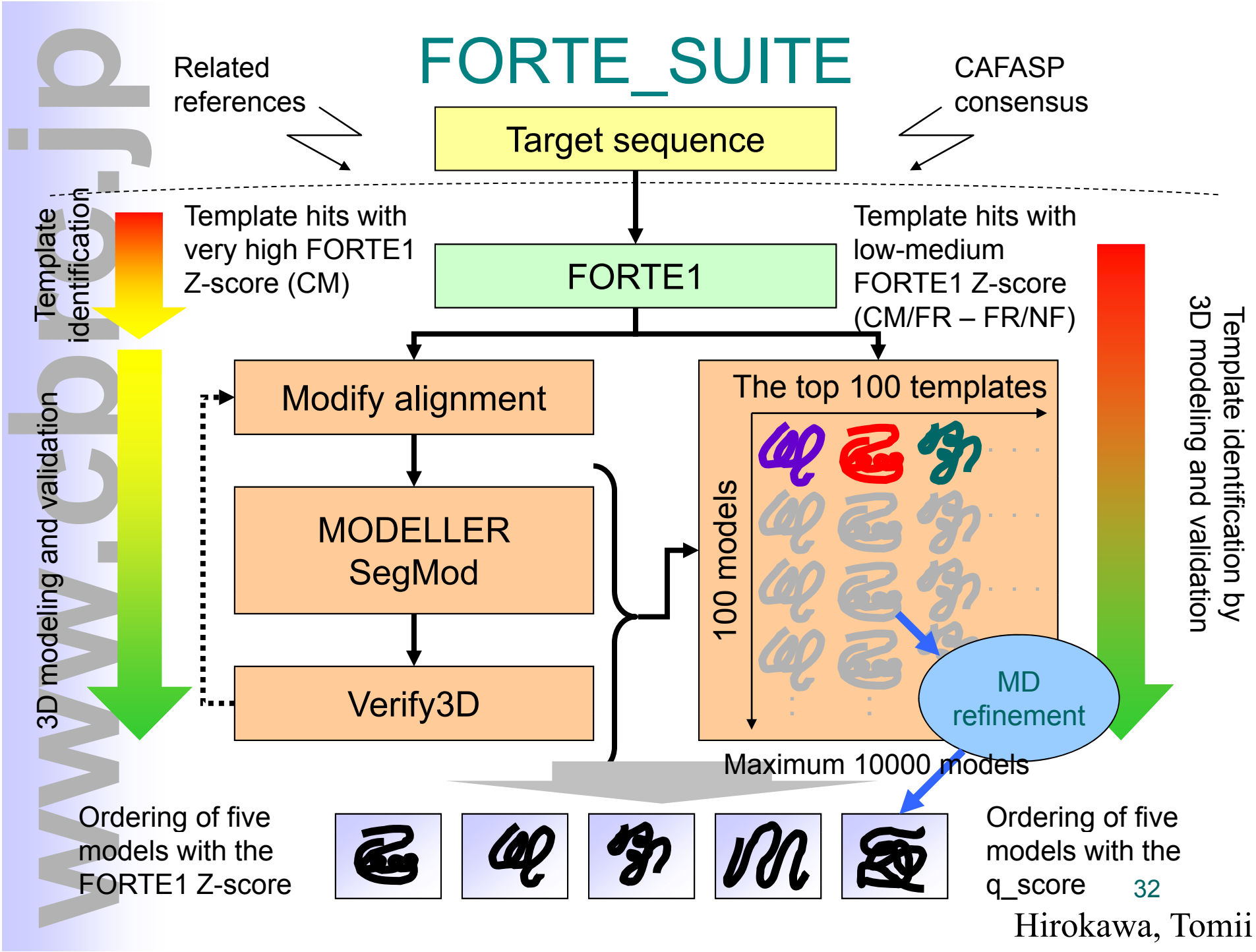
Fold Recognition method: FORTE

FORTE

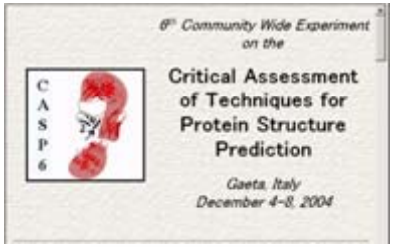
Tomii, K. & Akiyama, Y. :*Bioinformatics*, 20, (2004).



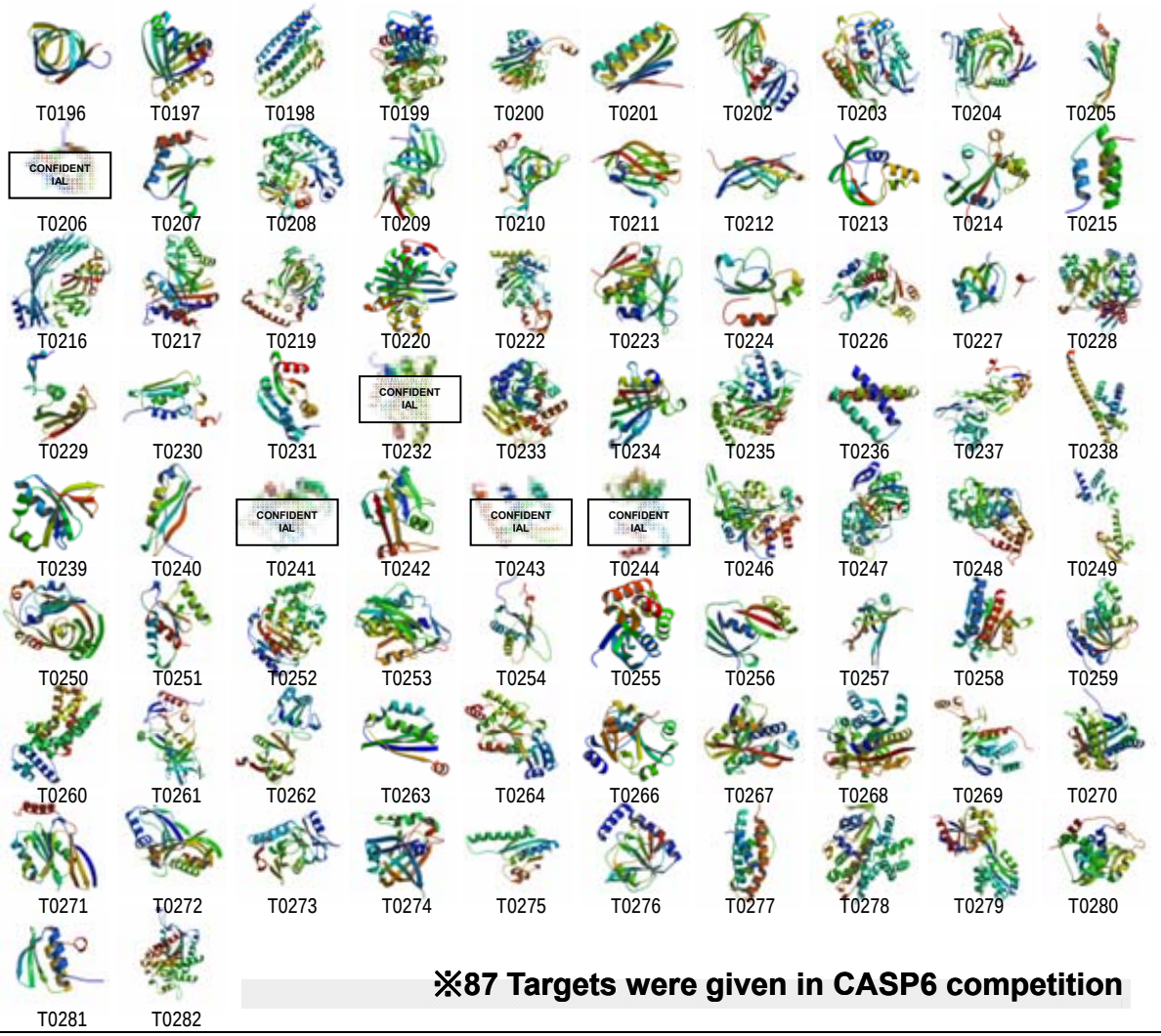
- Profile-Profile comparison based (not 3D-1D)
- Correlation coefficient for vector similarity score
- Frequent update of protein structure profiles
- Fully parallelized prediction (FORTE-SUITE)



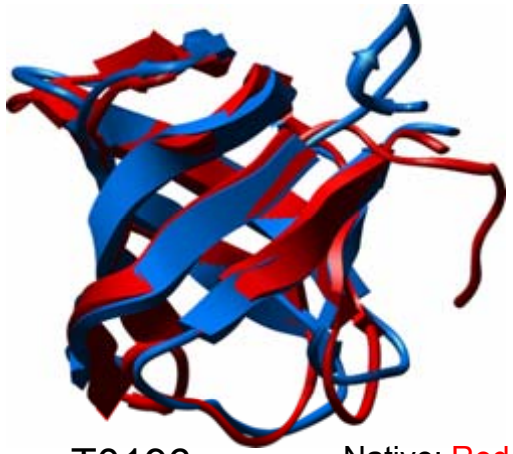
Hirokawa, Tomii



International competition CASP6 (2004)

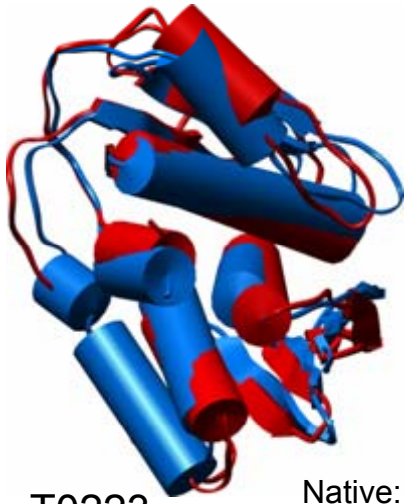


※87 Targets were given in CASP6 competition



T0196

Native: Red
Predicted: Blue



T0223

Native: Red
Predicted: Blue

Prediction

By CBRC-3D team

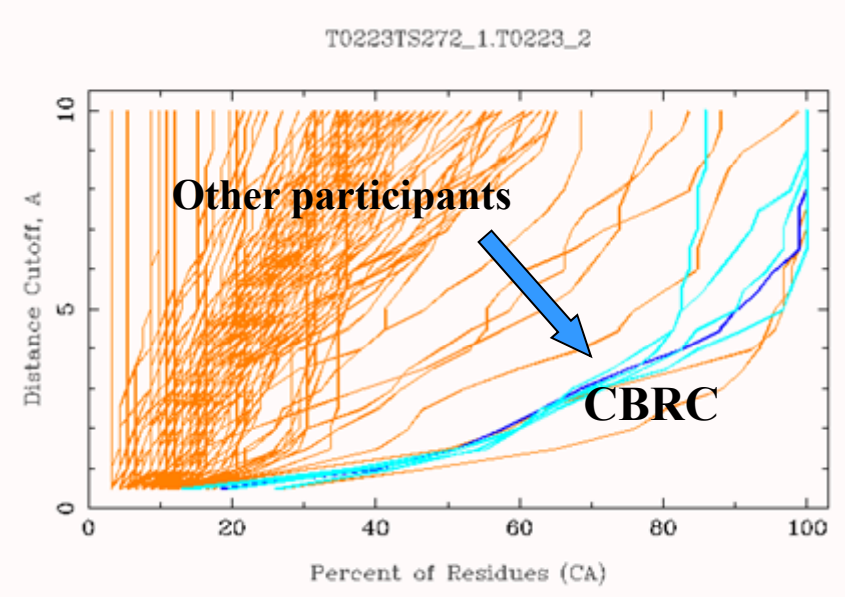
MNIFEAIENRHSVRDFLERKMPERVKDDIENLLVKFITKKLDWKINLSSFPSYIYAKAEK
HFDELVEYGFQGEQIVLFLTAQGFGTCWMARSPHPDVPYIIVFGYPRTRNFTRKRRPITS
FLENDLEELPPEIVKIVEMTILAPSALNRQPWKIKYTGGELCISSERPVDLGIALSHAYL
TAREIFKREPVIQKRGEDTYCLILNP

T0223 206 AA

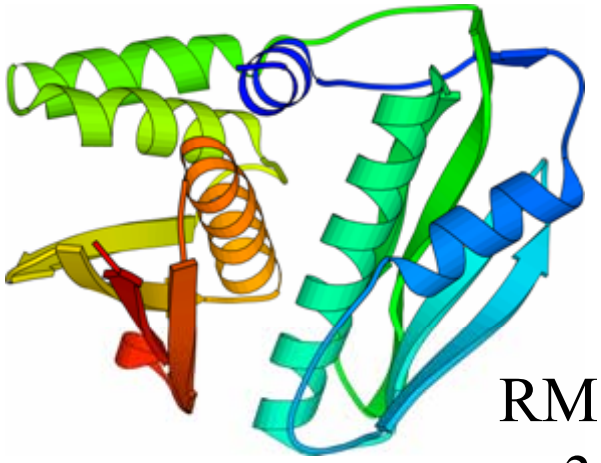
Putative Nitroreductase, *T. maritima*

CM/hard and FR/H

(hard for comparative modeling)

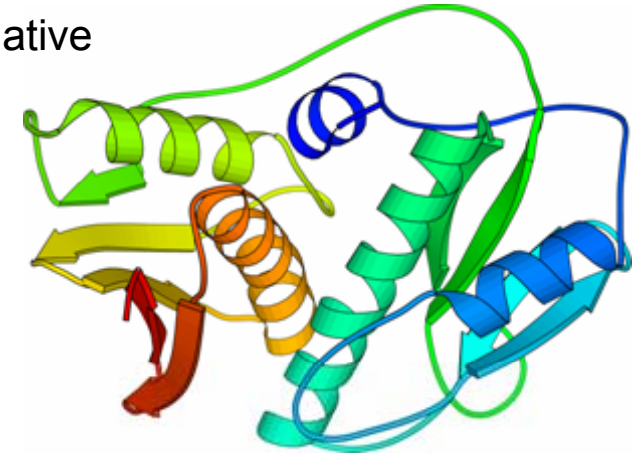


CBRC-3D (1st)



RMSD
3.68 Å

Native



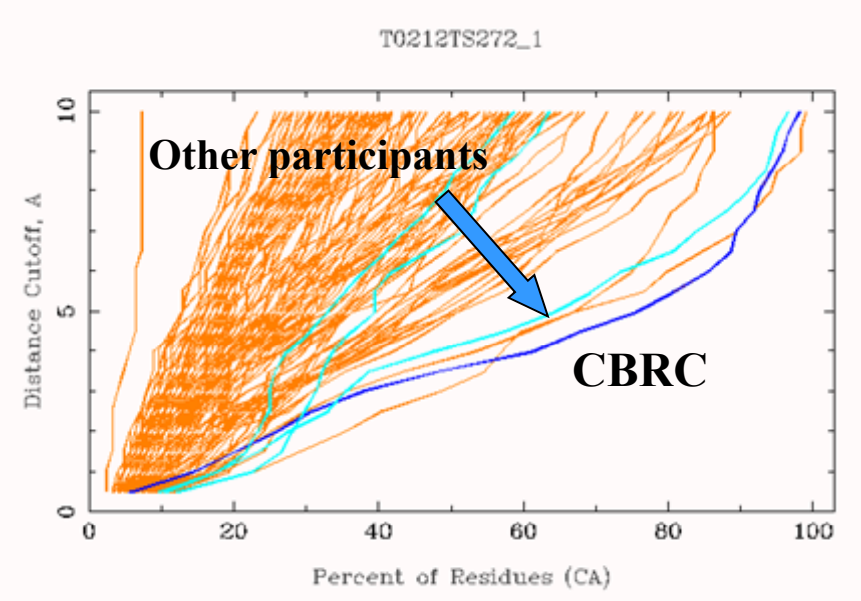
MSALDNSIRVEVKTEYIEQQSSPEDEKYLFSYTITIINLGEQAAKLETRHWIITDANGKT
SEVQGAGVVGETPTIPPNTAYQYTSGTVLDTPFGIMYGTYGMVSESGEHFNAIIKPFRLA
TPGLLH

T0212 126 AA

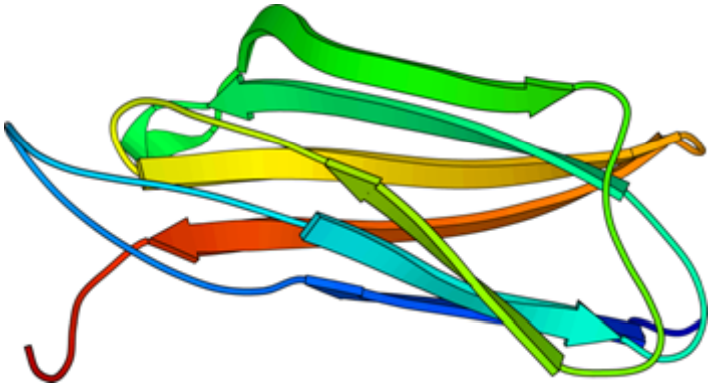
SOR45, *Shewanella oneidensis*

FR/A

(difficult for fold recognition)



CBRC-3D (1st)



RMSD
5.07 Å

Native