Designing succinct structural alphabet

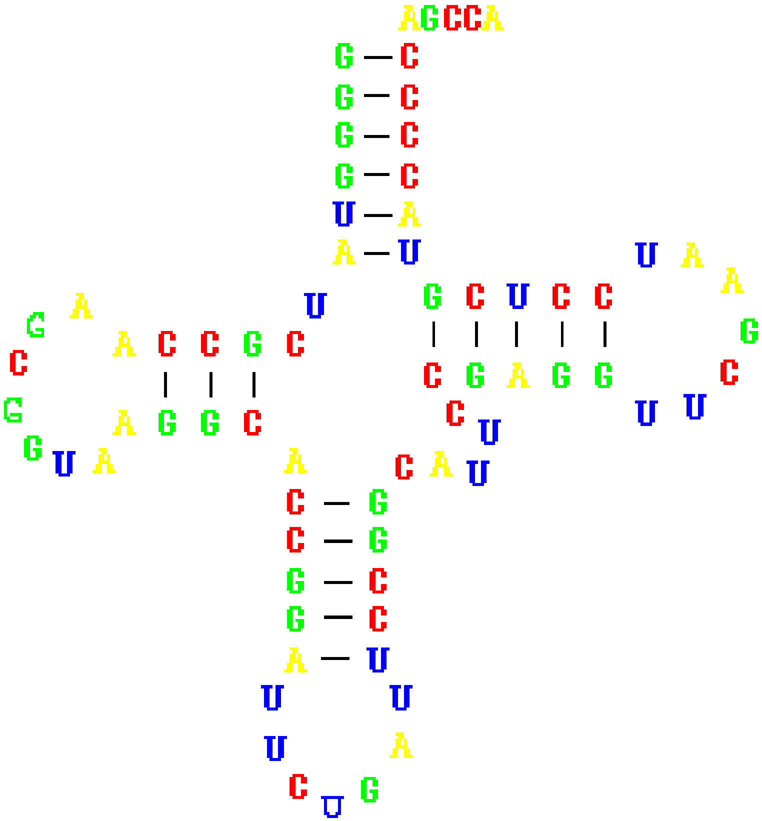
Dongbo Bu

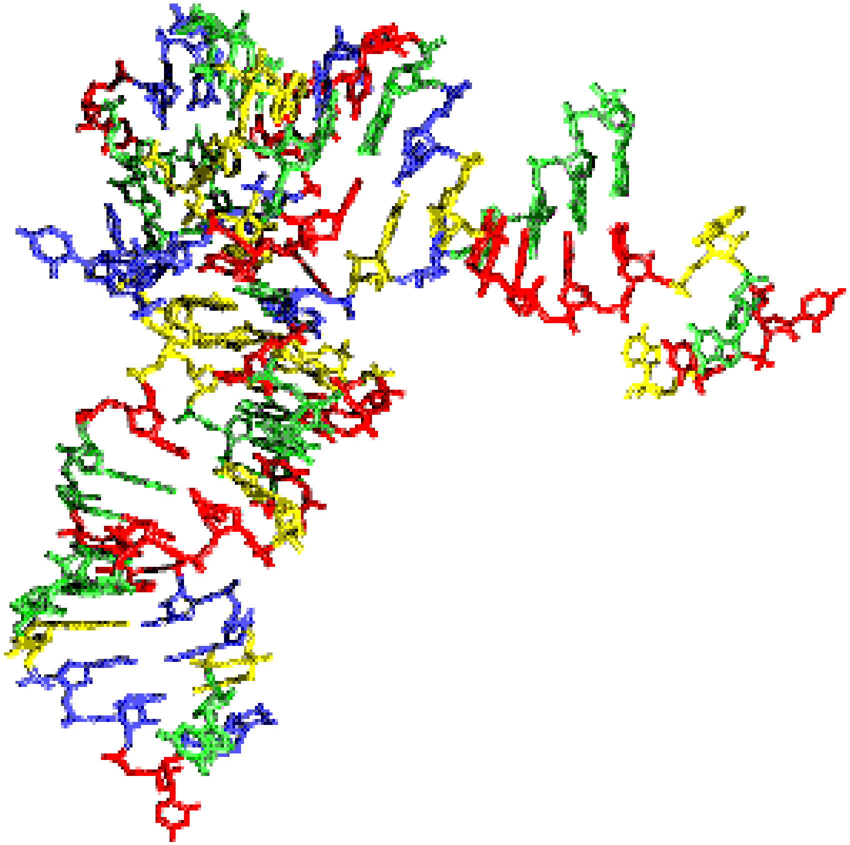
Shuai Cheng Li, Jinbo Xu, Sheng Wang,
Weimou Zheng, Ming Li

Institute of Computing Technology, CAS
Institute of Theoretical Physics, CAS
University of Waterloo, Canada

Protein structure prediction
---from sequence to structure

GGGGUAUCGCCAAGCGGUAAGGCACCGGAUUCUGAUUCCGGCAUUCCGAGGUUCGAAUCCUCGUACCCCAGCCA



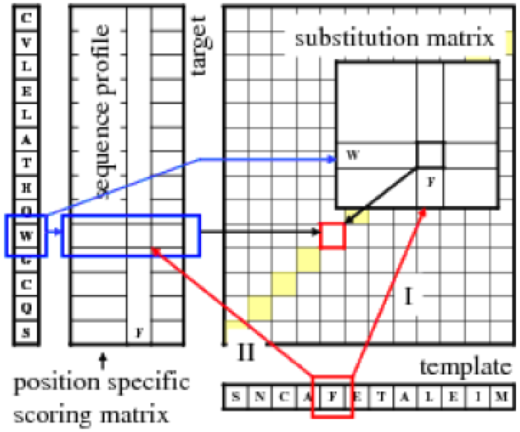


Ways to Obtain Protein Structures

- Wet lab methods: X-ray and NMR
  - $150k per structure
  - 0.5 year
  - Still need computational methods anyways
- Computational methods
  - Homology modelling – PSI-BLAST
  - Threading – RAPTOR
  - **Fragment Assembly (ROSETTA) and Fragment-HMM (FALCON)**.
  - Consensus

Prediction strategy 1: homology modeling

- Basic idea: two proteins usually adopt similar structure if they share similar sequence similarity.
- Technique： sequence-sequence similarity calculation.
- Advantages: can generate accurate predictions for proteins with sequence identity $> 30\%$ against a template.

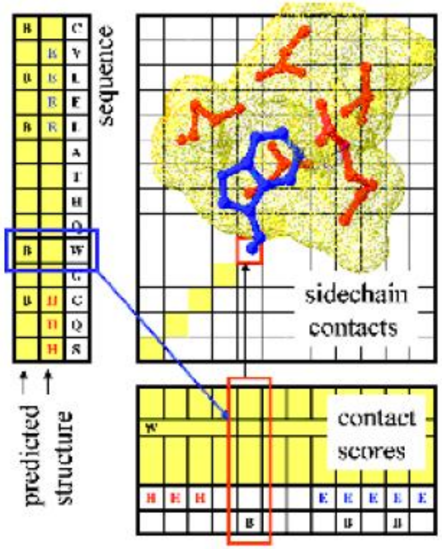


Homology Modelling Tools

- PSI-Blast and PDB-Blast: seq-seq comparison;
- FFAS: profile-profile comparison;
- ORFeus: add SS information to build meta-profile;
- SAM-T99: using HMM to capture relationship between residues, and to generate an accurate profile;

Prediction strategy 2: threading

- Basic idea: structures are usually more conserved than sequence.
- Technique: sequence-structure similarity calculation.
- Advantage: can detect remote-homology.



Threading Tools

- PROSPECT: adopting divide_and_conquer strategy;
- RAPTOR: using ILP optimization technique and SVM-Regression to choose template;
- SPARKS: using structure-driven profile;
- and others, such as mGenThreader, SAM-T02, 3D-PSSM, etc;

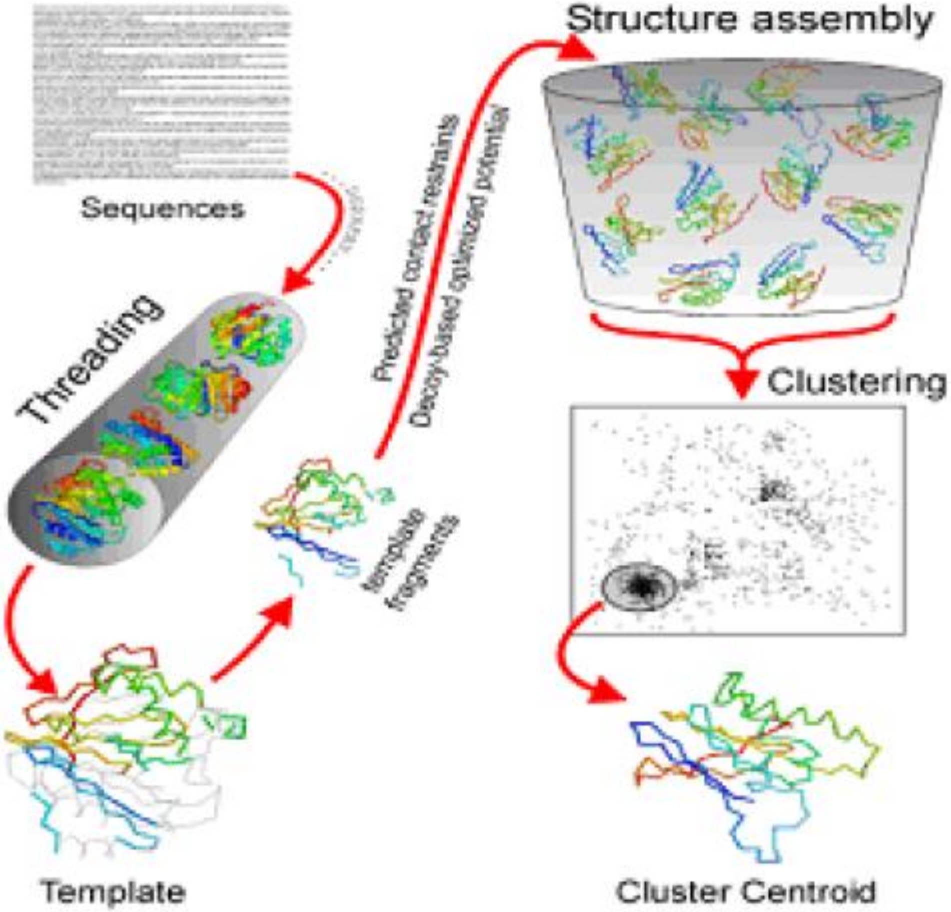
Prediction strategy 3: Ab Initio

- Basic idea: Proteins tends to adopt conformations with the minimal free energy.
- Technique: optimization.
- Advantage: can identify new fold since Ab Inition methods don't rely on templates with known structures.

Ab Initio Tools

- ROSETTA: predicting local structure for 9-mer fragments, and using Monte Carlo to optimization;
- TASSER: using large fragments from threading results as blocks;
- FB5-HMM: using FB5 technique to describe torsion angle preference;
- CRF-Sampler: using Conditional Random Field technique.

TASSER


Sequences
Threading
Template
template fragments
Predicted contact restraints
Decoy-based optimized potential
Structure assembly
Clustering
Cluster Centroid

Some challenges in protein structure prediction

1. Can we represent 3D structure as a sequence of "structural alphabet"?
2. Can we accurately predict structural alphabet for a sequence fragment?
3. Can we efficiently assemble local structures into a full-length structure?

……

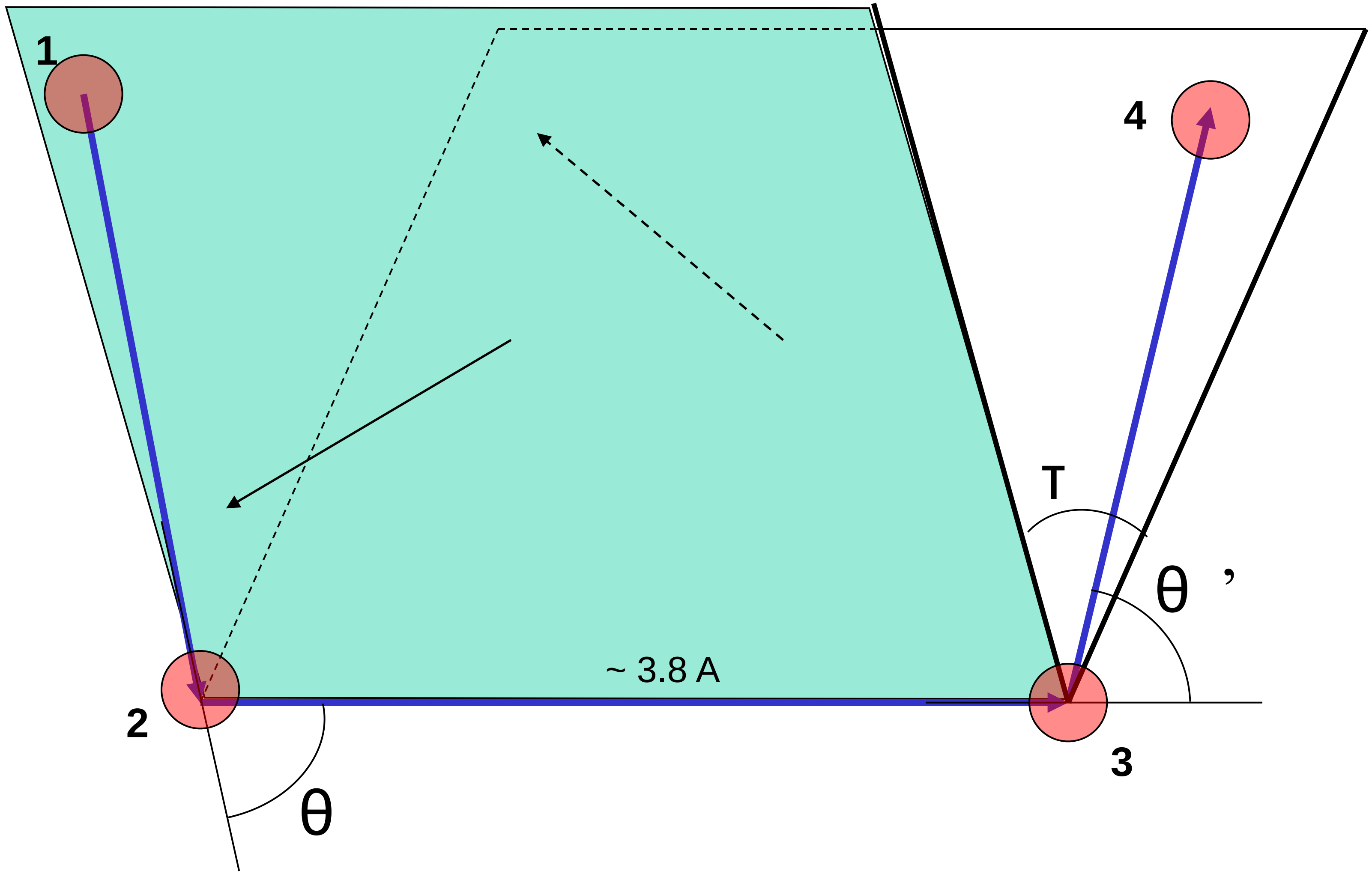
Part I

Conformational LEtter (CLE): representing 3D structure as a sequence

A simple representation of backbone: C-alpha pseudobond angles

θ T θ ’ ↔ $r_1r_2r_3r_4$

Four-residue fundamental unit:



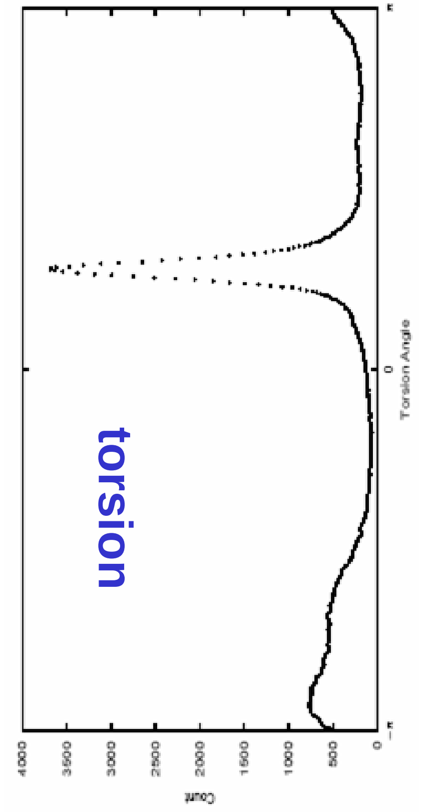
<x,y,z>=>angles distributions => CLE

3.14

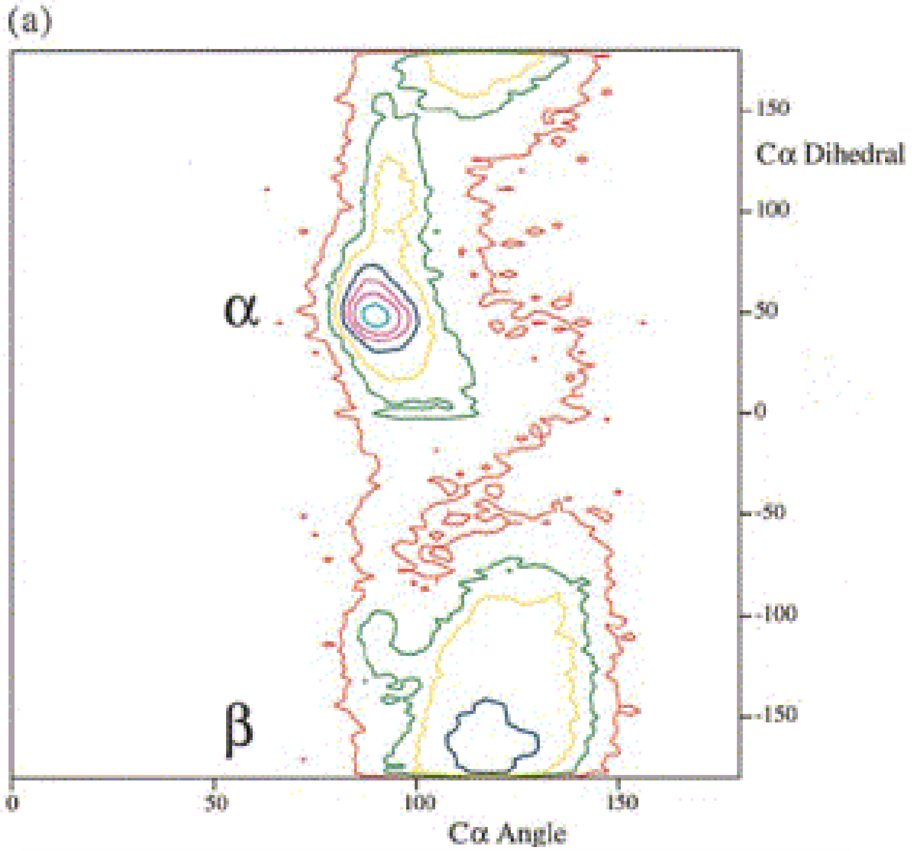
0.9

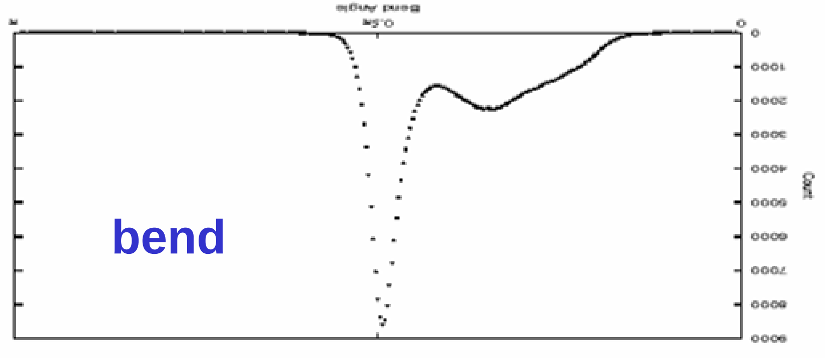
-2.9

-3.14



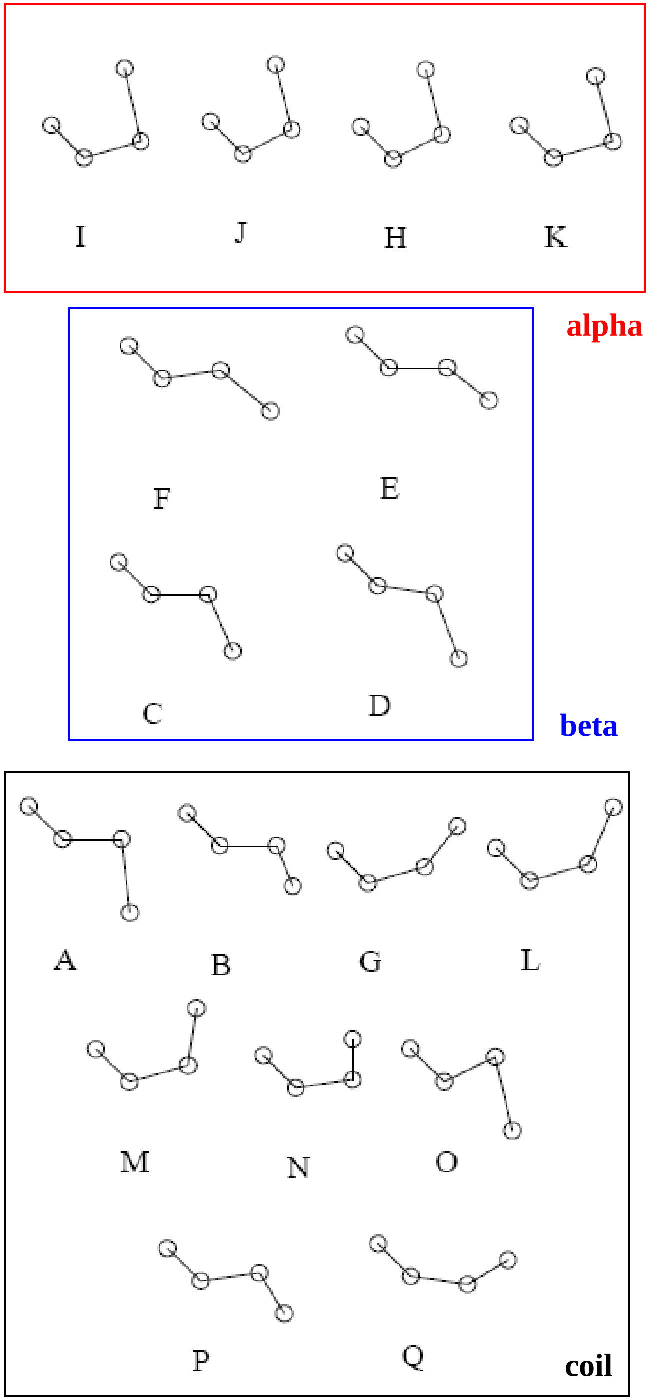
torsion



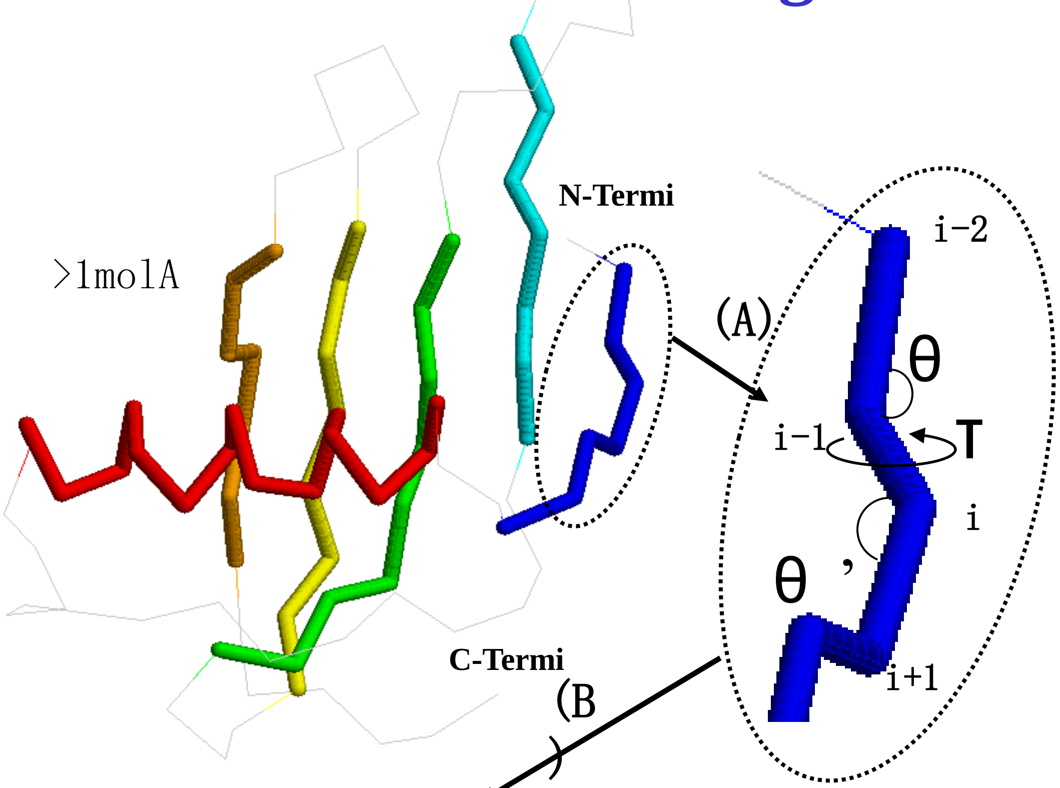


bend

3.14 1.55 1.10 0.0



The transformation from 3D structure to 1D CLE strings



$$i^* = \arg_i \max P(C_i \mid x_k), \quad \text{where} \quad x_k = (\theta, \tau, \theta')$$

$$P(C_i \mid x_k) \propto \pi_i |\Sigma_i|^{-1/2} \exp[-\frac{1}{2}(x_k - \mu_i) \cdot \Sigma_i^{-1}(x_k - \mu_i)]$$

(C)

>1molA

RRFEDECCGAIHHHHHHHHHHHHHHHOMICQEECBLDFQNBFEEEEFEQNNGCP

LDDEEEDEEENOGCEDEEEEEEPKKOGFEDPLDEQBGCCR

CLESUM: Conformational LEtter SUbstitution Matrix

J	37															
H	13	23														
I	16	18	23													
K	13	5	21	49												
N	-2	-34	-11	28	90											
Q	-44	-87	-62	-24	32	90										
L	-32	-62	-41	-1	8	26	74									
G	-21	-51	-34	-13	-8	8	29	69								
M	16	-4	1	12	7	-7	5	21	61							
B	-57	-96	-74	-50	-11	12	-12	13	-13	51						
P	-34	-60	-49	-36	-3	7	-12	5	8	42	66					
A	-23	-45	-31	-19	10	16	-11	-6	-2	20	35	73				
O	-24	-55	-34	5	15	-13	-4	-1	5	-12	4	25	104			
C	-43	-77	-56	-33	-5	29	0	-4	-12	7	4	13	3	53		
E	-93	-127	-108	-84	-43	-6	-21	-22	-47	15	-5	-25	-48	3	36	
F	-73	-107	-88	-69	-32	3	-16	-5	-33	7	0	-20	-30	20	26	50
D	-88	-124	-105	-81	-44	14	-22	-31	-49	13	-10	-17	-42	21	22	21	52
	J	H	I	K	N	Q	L	G	M	B	P	A	O	C	E	F	D

typical helix

evolutionary

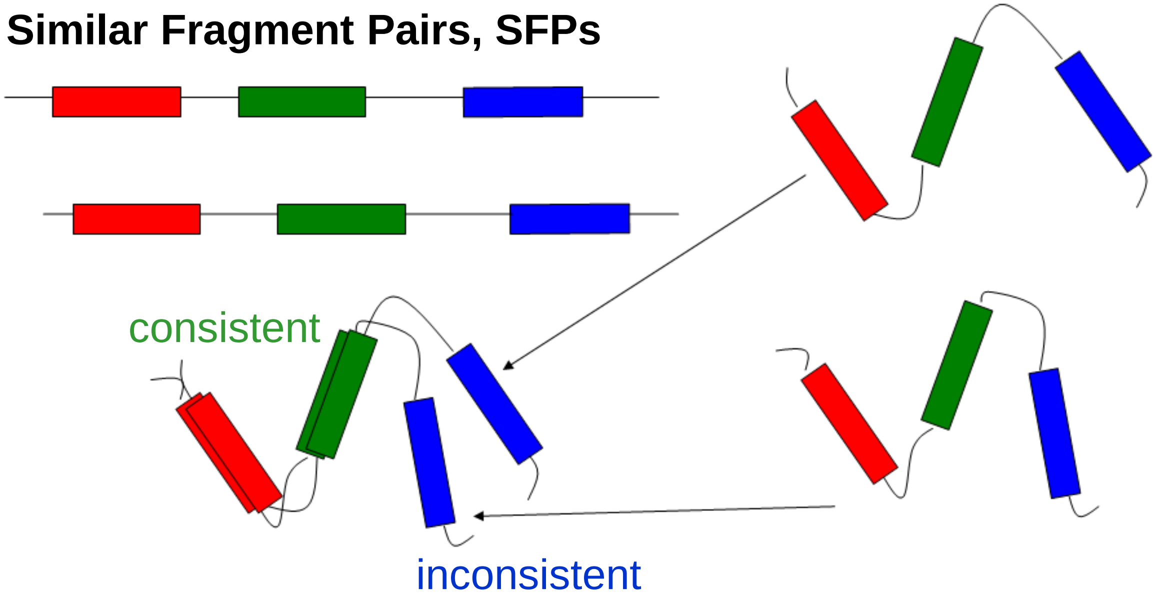
+ geometric

typical sheet

$M_{ij} = 20 * \log_2 (P_{ij}/P_iP_j)$

constructed using FSSP representatives.

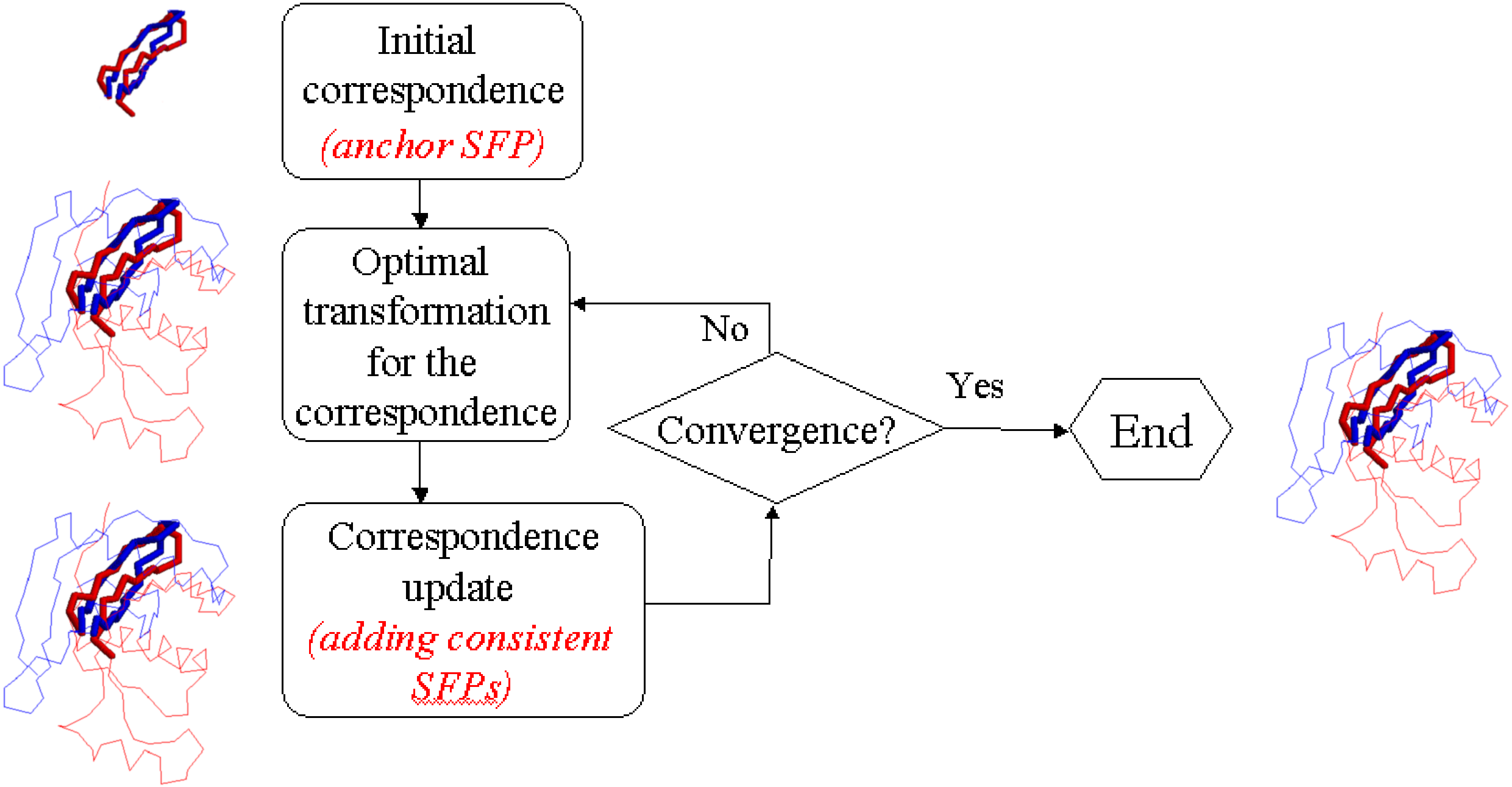
Structures can be aligned efficiently through CLE.



- Alignment: to collect as many consistent SFPs as possible.
- Balance local similarity and global consistency.

Alignment paradigm

“A fast, reliable, and convergent method
for protein structural alignment is not yet available”
*---- by **Patrice Koehl** at Protein Structure Classification 2006*



Initial correspondence *(anchor SFP)*

Optimal transformation for the correspondence

Correspondence update *(adding consistent SFPs)*

Convergence?

No

Yes

End

Two structure alignment strategies

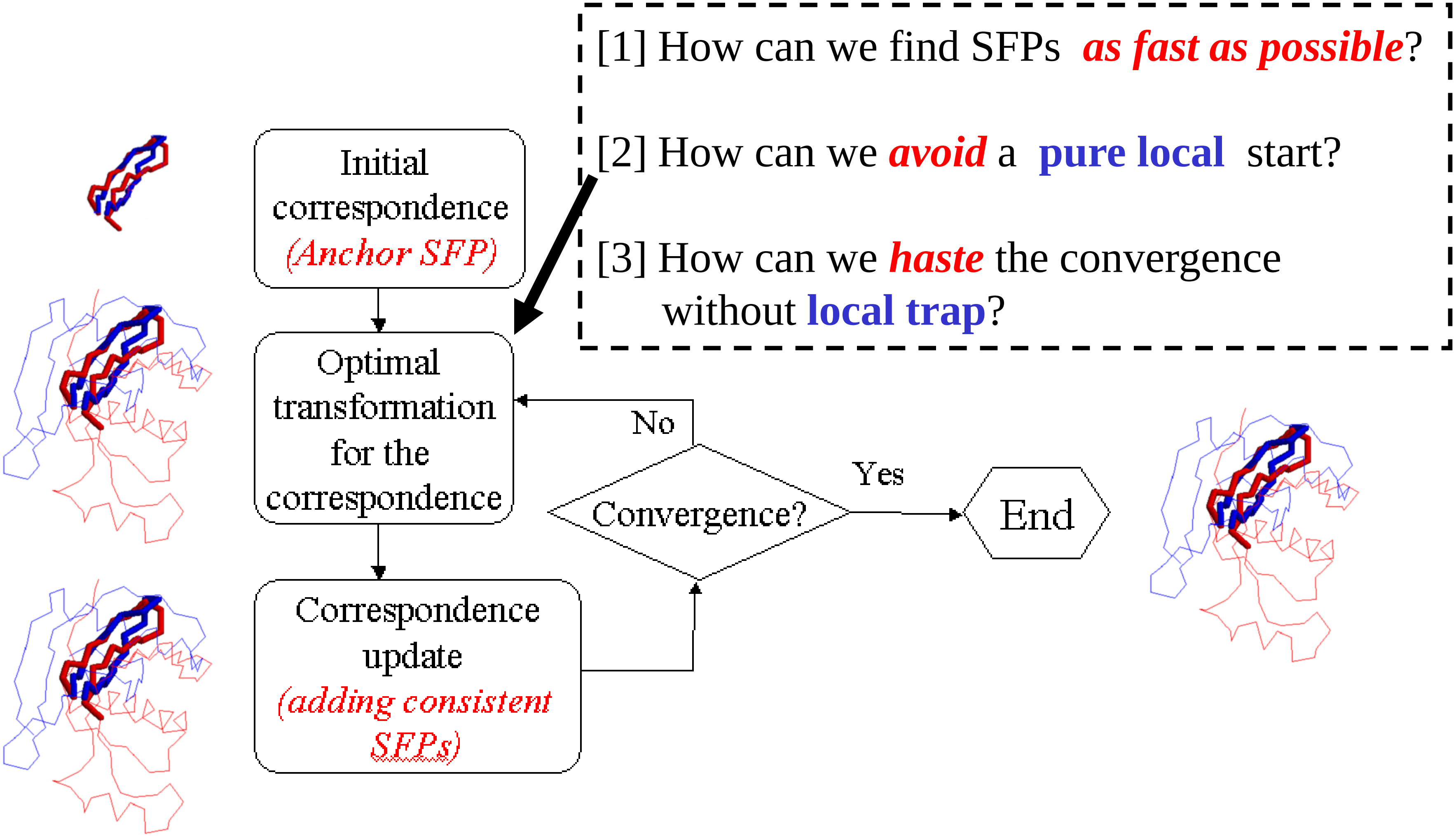
1. Global-consistency-first
Find as much initial ROTMAT as possible, use self-consistent paradigm to each ROTMAT, then select a maximal one with the largest similar path. (like STRUCTAL,PROSUP)
Remark: break the consistent but fast

2. Local-similarity-first
Find as much SFP as possible, then heuristically concatenate them as a larger consistent path as possible. (like DALI,CE)
Remark: retain the consistency but slow

Our algorithm: CLEPAPS

[1] Generate SFPs according to CLESUM score (like pointview 2)
[2] Use top k SFP as initial correspondence + self-consistent iteration (like pointview 1)

Three main problems



[1] How can we find SFPs ***as fast as possible***?

[2] How can we ***avoid*** a **pure local** start?

[3] How can we ***haste*** the convergence
without **local trap**?

Initial correspondence *(Anchor SFP)*

Optimal transformation for the correspondence

Correspondence update *(adding consistent SFPs)*

Convergence?

No

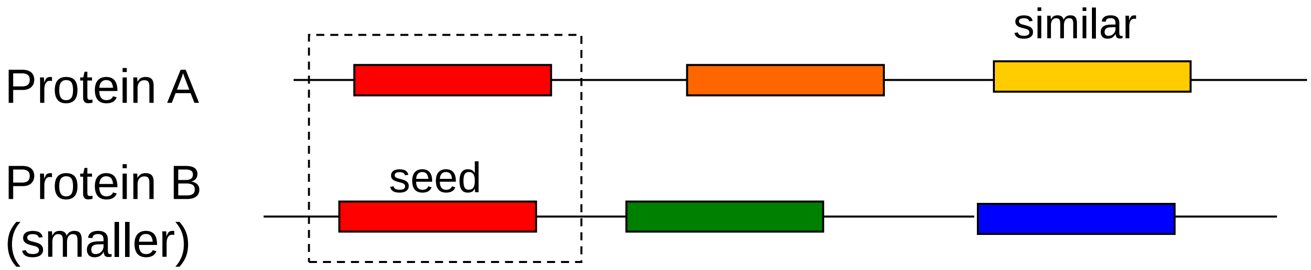
Yes

End

[1] Using CLE and CLESUM to find SFPs

SFP => highly scored string pair

- Fast search for SFPs by **string comparison**

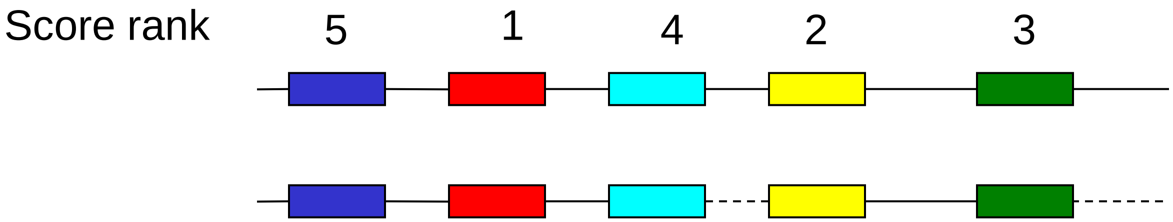


- CLESUM similarity **score → importance of SFPs**

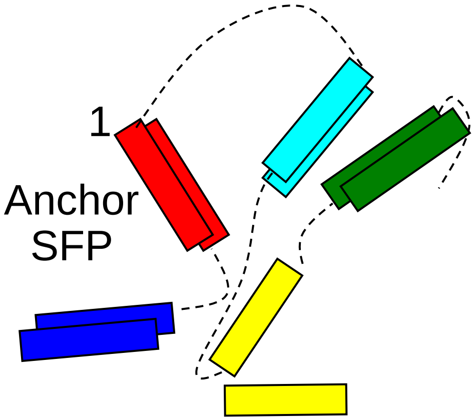
Guided by CLESUM scores, only a few top SFPs need to be examined to determine the superposition for alignment, and hence a reliable greedy strategy becomes possible.

[2] Construct ‘Star-tree’ to select the optimal Anchor-SFP

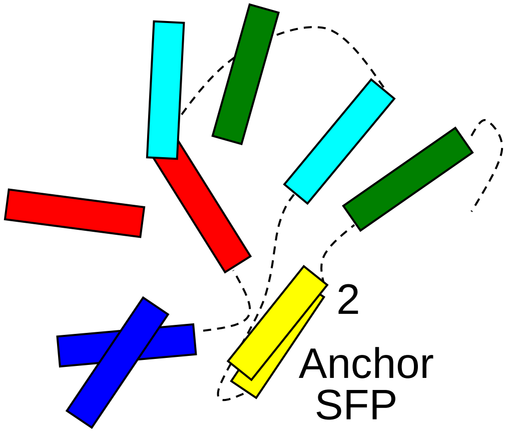
Selection of Optimal Anchor SFP



Example: TopK = 2; TopJ = 5



# of consistent SFPs = 4



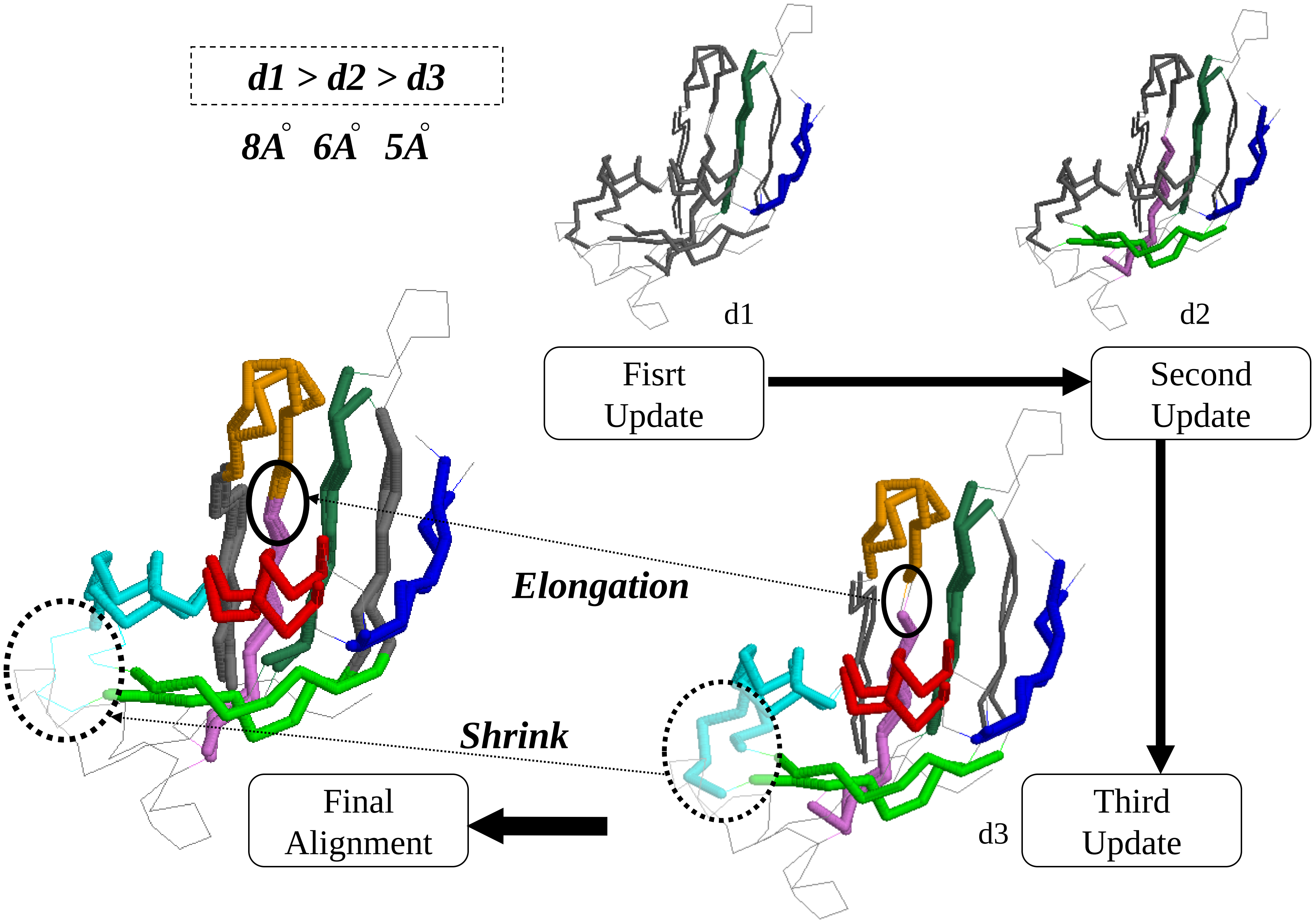
# of consistent SFPs = 1

Top1 SFP is globally supported by three other SFPs, while Top2 SFP is supported only by itself.

[3] Apply ‘Zoom-in’ strategy to avoid **local trap**

d1 > d2 > d3

8Å 6Å 5Å



d1

d2

d3

Fisrt Update

Second Update

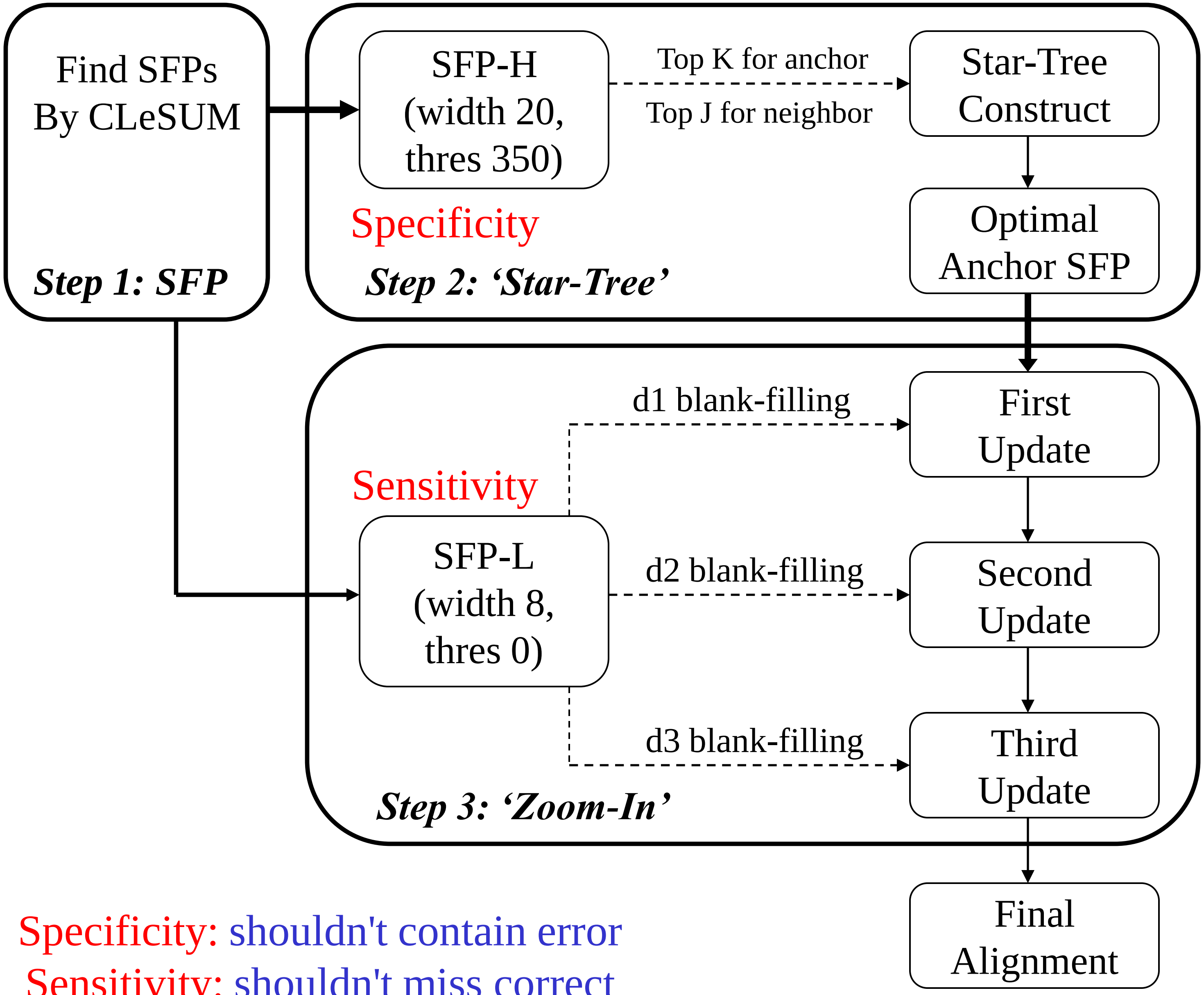
Third Update

Final Alignment

Elongation

Shrink

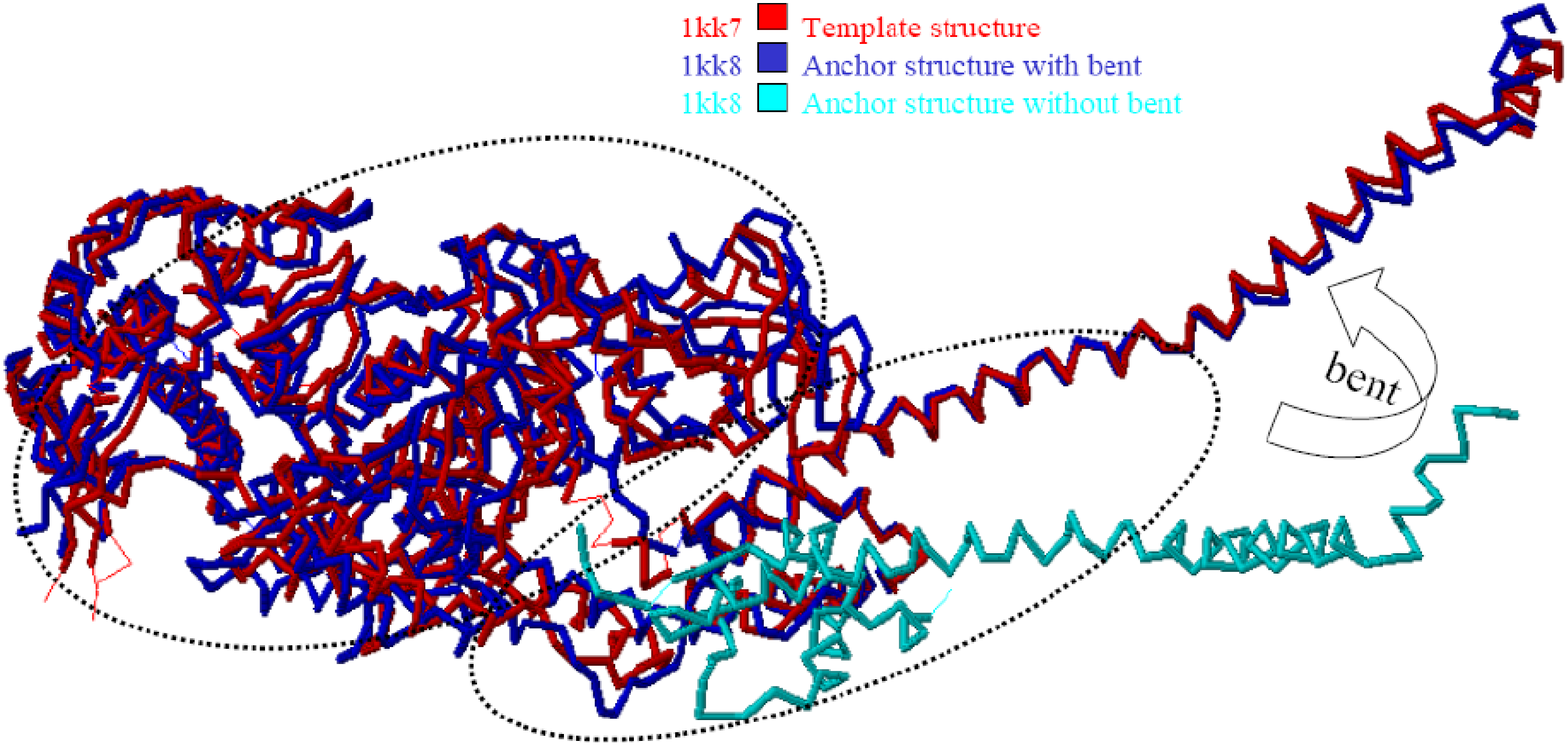
The flow chart of CLEPAPS algorithm



Specificity: shouldn't contain error

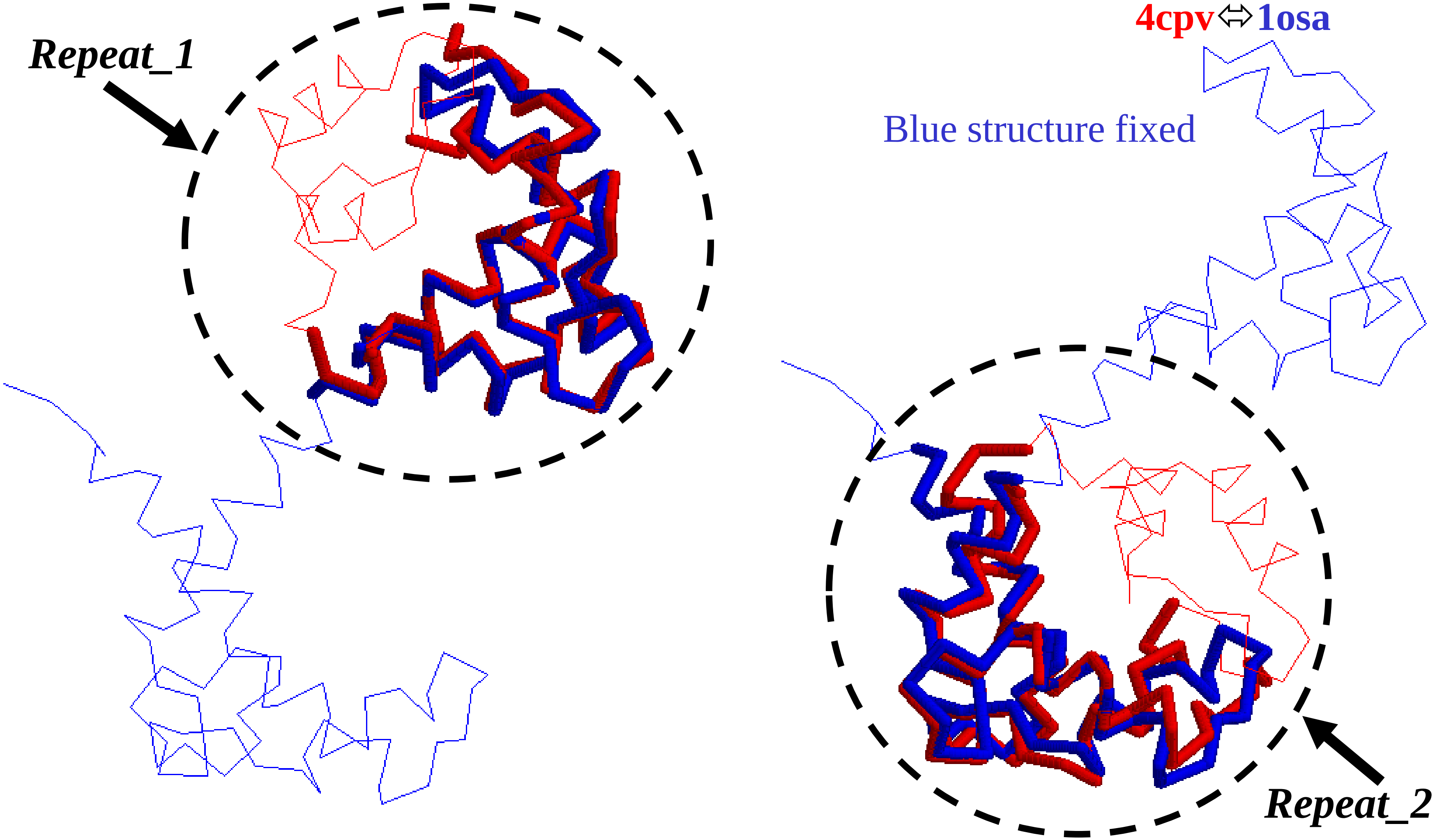
Sensitivity: shouldn't miss correct

Example 1: domain-move


1kk7 🟥 Template structure
1kk8 🟦 Anchor structure with bent
1kk8 🟦 Anchor structure without bent
bent

针对具有双domain的肌动蛋白(Myosin)在不同状态下的联配。红色表示结合ATP的状态，蓝色（青色）表示不结合ATP的状态，青色表示第一次迭代后，domain 2的原始位置，而蓝色表示两次迭代后最终的联配。

Example 2: domain repeats



4cpv⇔1osa

Blue structure fixed

Repeat_1

Repeat_2

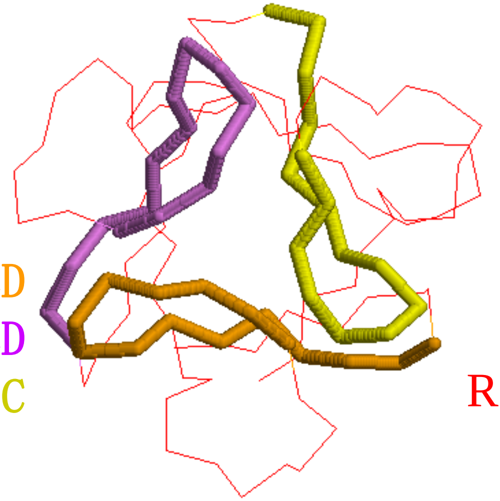
Solution [A]

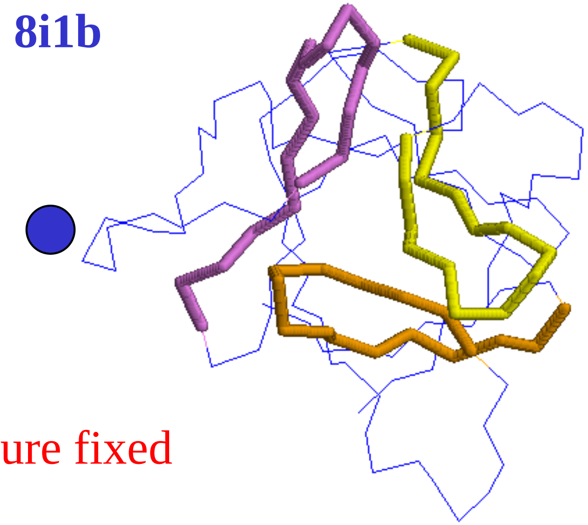
Solution [B]

Example 3: Symmetry

4fgf ⇔ 8i1b

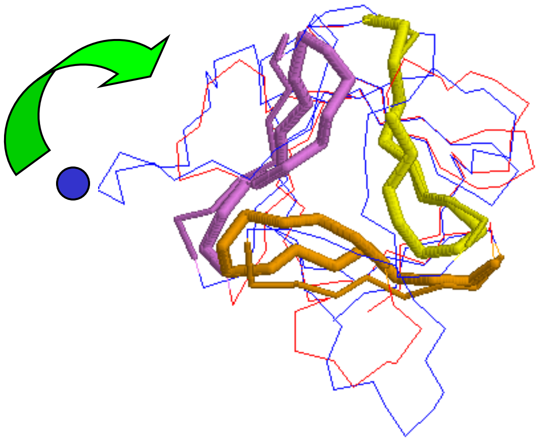
4fgf
OGCCFEFAHOGEED
OGDCEDFAIOGEED
KGFCEDDAJOGCCC



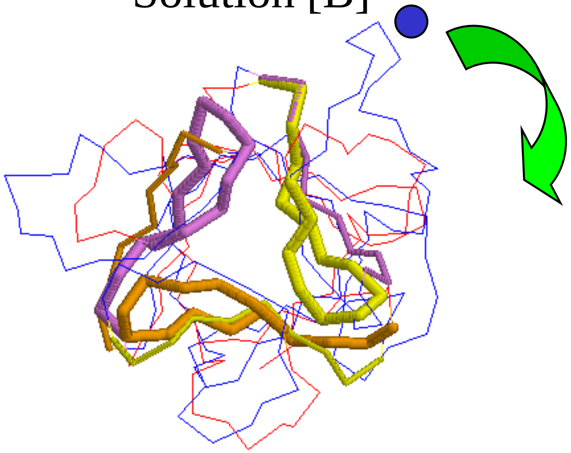


Red structure fixed

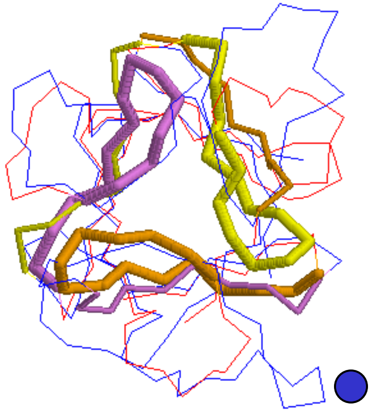
Solution [A]



Solution [B]



Solution [C]



针对具有三度对称性的两个蛋白的联配。红色蛋白是4FGF，蓝色蛋白是8I1B。通过三次迭代结果我们可以看清其具备的对称性。分析4FGF中三段处于对称位置的结构码（分别用紫色，橙色，黄色标记）也可以发现其相似性。

Summary

CLEPAPS →

1, Fast search for SFPs by merely string comparison

2, Width 20 for specificity + width 8 for sensitivity

3, Optimal Anchor SFP selected by checking consistency

4, Avoid Local Trap by ’Zoom-in’

CLEFAPS →

1, Seed-extension, 6-9 for SFP-L and 9-18 for SFP-H

2, Consistency parameters self-adaptive with the input length

3, Using some detailed refinement for ‘Zoom-in’

4, Introduce amino-acid information to improve SFP quality

Result →

1, Both CLEPAPS and CLEFAPS runs 50-200 times faster than others

2, CLEPAPS got similar alignment length with reference,
while CLEFAPS got similar alignment accuracy.

Part II

FRazor: accurately predict local structure for a sequence fragment

Fragment Based Protein Structure Prediction

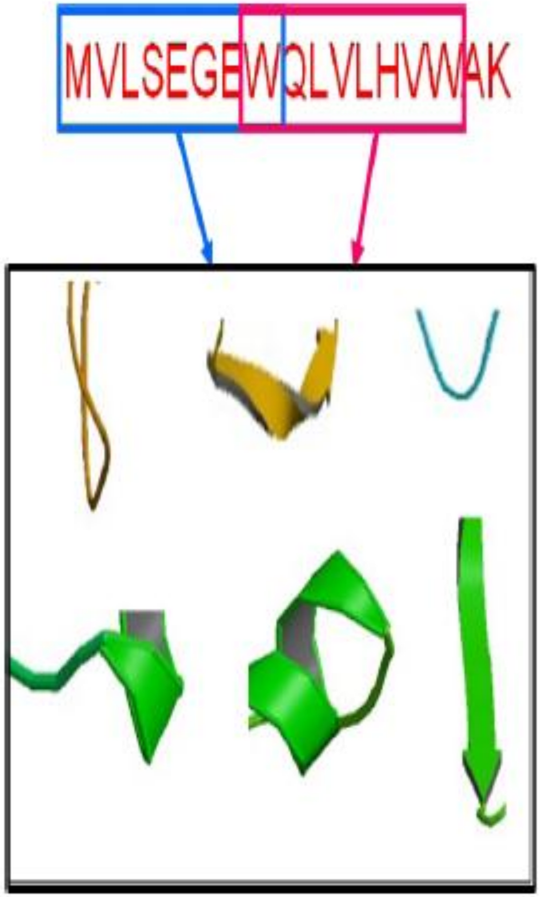
Fragment-based protein structure prediction is done in two major steps:

- **Identify the building blocks, which are fragments of known structures.**
- Construct or sample the protein structure with those building blocks using some search or simulation algorithms.

Fragment Libraries

- Non-Position Specific Libraries:
  - Number for fragments: Vary from dozens to hundreds.
  - Length of fragments: Fixed or variable lengths. Typically, no more than nine [Fidelis et al 1994].
- Position Specific Libraries: ROSETTA

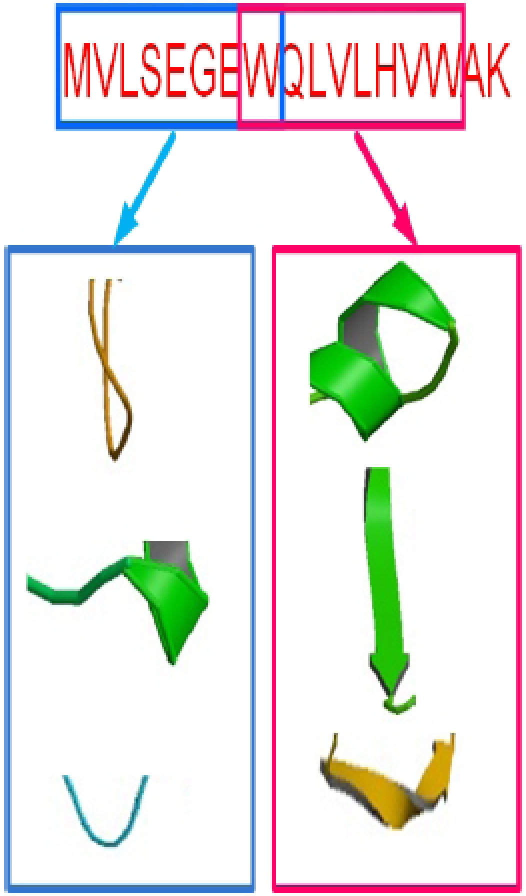
$$DISTANCE = \sum_{i=1}^{\ell} \sum_{aa=1}^{20} |S(aa,i) - X(aa,i)|$$



Fragment Libraries

- Non-Position Specific Libraries:
  - Number for fragments: Vary from dozens to hundreds.
  - Length of fragments: Fixed or variable lengths. Typically, no more than nine [Fidelis et al 1994].
- Position Specific Libraries: ROSETTA

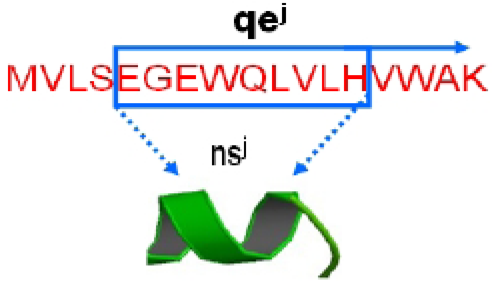
$$DISTANCE = \sum_{i=1}^{\ell} \sum_{aa=1}^{20} |S(aa, i) - X(aa, i)|$$



Problem Statement–Notations

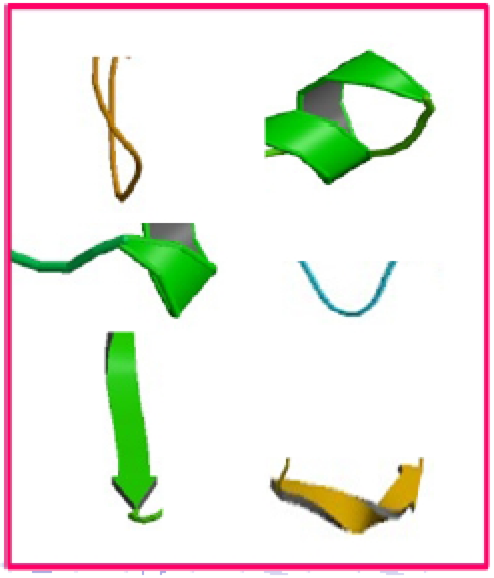
Target sequence t of length n parsed into sequence segments:

- A sliding window of a fixed length ℓ and step size 1 is used.
- These segments are denoted as $qe^1, qe^2, \ldots, qe^p,\ p = n - \ell + 1$.
- denote the *native structural* fragments as $ns^1, ns^2, \ldots, ns^p$.



Structural Space:

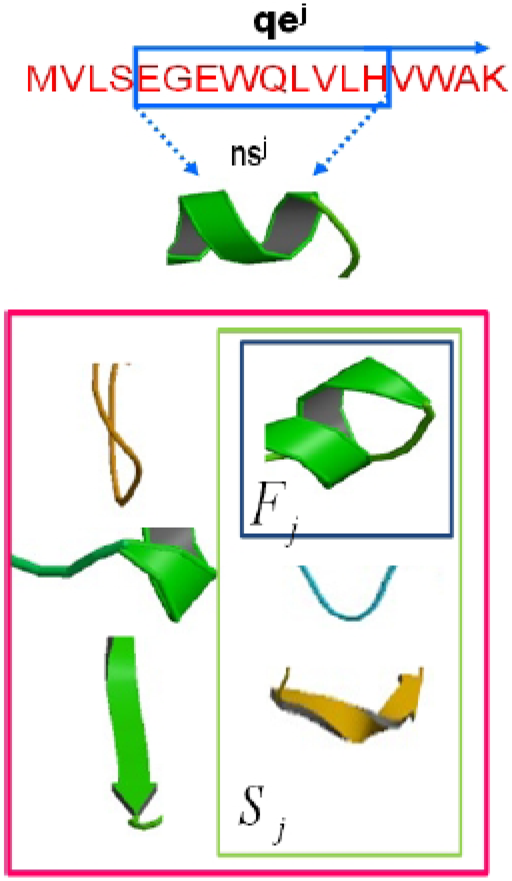
- Structural fragments to select the structural candidates for sequence segments.
- Denoted as $\mathcal{S} = \{se^1, se^2, \ldots, se^q\}$.



Problem Statement

Problem Definition: Given sequence segment qe^j, integer k and k', $k' \leq k$ and a threshold θ, to select a set of structural fragments, denoted as $\mathbb{S}_j$, such that:

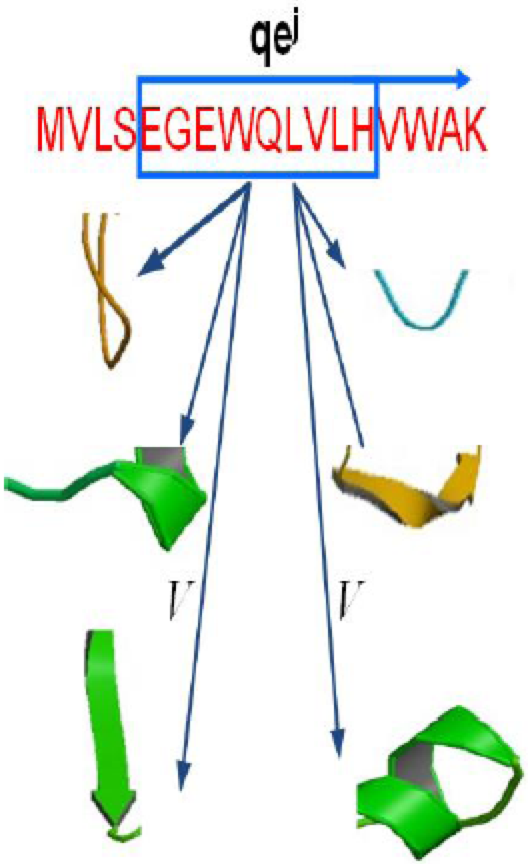
- $|\mathbb{S}_j| = k$.
- $\exists F_j \subset \mathbb{S}_j$ with $|F_j| \geq k'$.
- $\forall s \in F_j$, $dist(s, ns^j) \leq \theta$, $dist$ is the Cα root-mean-squared deviation.



Generalized Linear Model–Motivation

Between each structural fragment se^i in and each sequence segment qe^j, a feature vector is computed:

- Denote the feature vector as: $\mathcal{V}^{i,j} = \langle v_1^{i,j}, \ldots, v_d^{i,j} \rangle$, $d = 4 \times 9$, $1 \leq i \leq q$ and $1 \leq j \leq p$
- It measures how well se^i and qe^j match
- Each entry in $\mathcal{V}^{i,j}$ may be a linear or nonlinear scoring function.
- We label $\mathcal{V}^{i,j}$ with $+1$ if $dist(se^i, ns^j) \leq \theta$, and -1 otherwise.



Formulation–Linear Model

A general linear model has the form [Bishop 06]:

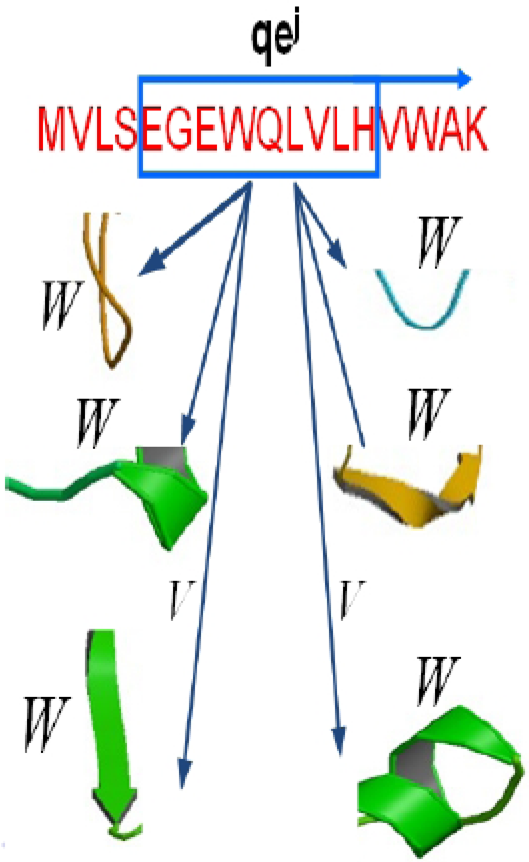
$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{k=1}^{M} w_k \phi_k(\mathbf{x})$$

- $\mathbf{w} = (w_0, ..., w_M)^T$ **w** is the *weight vector or the parameters* to train, and w_0 is called a bias parameter and used for any fixed offset in the data.
- **x** is the input data.
- $(1, \phi_1, ..., \phi_M)^T$ are the *basis functions.* The basis functions are generally nonlinear and are applied to the original data variables.
- $y(\mathbf{x}, \mathbf{w})$ is a nonlinear function of the input variables due to the non-linearity of the basis functions.

Formulation–Notations

- Feature vector $\mathcal{V}^{i,j} = \langle v_1^{i,j}, \ldots, v_d^{i,j} \rangle$
  - to measure the similarity between a structural fragment se^i and sequence segment qe^j
  - $-1 \leq v_l^{i,j} \leq 1$.
- Each structural fragment se^i is associated with a weight vector $\mathcal{W}^i = \langle w_1^i, \ldots, w_d^i \rangle$.
- distance between se^i and qe^j is: $\mathcal{D}^{i,j} = \sum_{l=1}^{d} w_l^i v_l^{i,j}$

Objective: To adjust $\mathcal{W}^i = \langle w_1^i, \ldots, w_d^i \rangle$ so that some "native-like" structure for qe^j, $\mathcal{D}^{n_j,j}$ is



Formulation

For $1 \leq i \leq q$, indexing the structural space and $1 \leq j \leq p$, indexing the sequence segments, the ILP is as follows:

$$\begin{aligned}
&\min \sum_{j=1}^{p} g_j && \\
\mathcal{D}^{n_j,j} - \mathcal{D}^{i,j} &\leq d_{n_j,i,j}(2+\epsilon) - \epsilon, && n_j \in \mathcal{Q}^j, i \notin \mathcal{Q}^j, \forall j \\
\sum_{1 \leq i \leq q, i \notin \mathcal{Q}^j} d_{n_j,i,j} &\leq k - 1 + f_{n_j,j}(q - (k-1)), && n \in \mathcal{Q}^j, \forall j \\
\sum_{n_j \in \mathcal{Q}^j} f_{n_j,j} &\leq |\mathcal{Q}^j| - 1 + g_j, && \forall j \\
\sum_{l=1}^{d} w_l^j &\leq 1, && \forall j \\
d_{n_j,i,j}, f_{n_j,j}, g_j &\in \{0,1\}, && w_i^j \in [0,1]
\end{aligned}$$

$V^{i,j}$, totally four types of basis functions are defined, each of which contains nine items:

- Mutation Scores.

$$\sum_{aa=1}^{20} S(aa,i) \times \log \frac{X(aa,i)}{S(aa,i)} \tag{1}$$

- Secondary Structure Score.
  - If the secondary structure type of $se[i]$ is α-helix, then we use α_i
  - If the secondary structure type of $se[i]$ is β-sheet, then we use β_i
  - If it is loop, we just use l_i.
- Contact Capacity Score.
- Environmental Fitness Score.

Basis Functions–Heuristics to solve the formulation

Cplex is used to solve the problem. Following heuristics are used when problem size is large and to avoid over-fitting.

- Parameters are justified for each (sub) structural fragment.
- Combine the parameters from each fragments.

Our data set consists of three parts:

- Structural Space: the collection of structural fragments to select the candidate structural fragments. It is made from 40 protein chains
- Training Set: the fragments used to justify parameters. it is made from 30 protein chains. The proteins for Structure Space and Training Set are both from a non-homologous (less than 30% homology) list with resolution $< 2\AA$, dated on March 26^{th}, 2006.
- Testing Set, proteins for evaluating our method Proteins from CASP 7.

Position Coverage for CBM vs. FRazor's Score Function

	α-Helix		β-Sheet		Loop		Overall	
θ (Å)	CBM	FRazor	CBM	FRazor	CBM	FRazor	CBM	FRazor
0.5	94.2	95.1	10.0	37.6	26.6	38.7	49.4	55.1
1	98.2	98.6	56.4	89.6	55.5	78.1	72.2	88.2
1.5	99.7	99.7	89.3	98.2	81.3	93.3	89.9	96.7
2	100	100	99.7	99.8	96.9	98.9	98.6	99.4
2.5	100	100	99.9	99.9	99.7	99.7	99.8	99.8
3	100	100	100	100	99.9	100	99.9	100
3.5	100	100	100	100	100	100	100	100

position coverage: The percentage of the positions which are correctly predicted.

Position Coverage for Threshold Value as 1Å.

	α-Helix		β-Sheet		Loop		Overall	
k	CMB	FRazor	CMB	FRazor	CMB	FRazor	CMB	FRazor
5	90.5	96.6	34.2	65.6	40.3	59.8	60.7	75.1
10	97.2	97.5	42.4	79.1	46.1	67.9	65.1	81.5
15	97.8	99.3	49.5	82.1	50.6	70.5	68.6	85.0
20	98.1	98.0	53.6	85.1	53.5	73.0	70.8	86.4
25	98.2	98.6	56.4	89.6	55.5	78.1	72.2	86.4
30	98.3	98.7	59.9	90.8	57.4	79.6	73.6	88.2
35	98.5	98.8	61.5	92.0	58.5	81.1	74.5	90.0
40	98.7	99.0	63.5	92.9	59.5	82.3	75.4	90.8

Customized Fragment Lists vs. Independent (Kolodny) Fragment Libraries

	Fragment Coverage (%)		Local Fit Score (Å)	
L or *k*	KFL	FRazor	KFL	FRazor
25	–	45.3	–	0.763
50	36.2	40.5	0.754	0.667
100	40.7	55.7	0.673	0.589
150	43.3	58.6	0.633	0.554
200	44.0	60.4	0.603	0.531
250	46.3	61.8	0.585	0.515

Decoy quality comparison between ROSETTA and FRazor

Target Protein			ROSETTA			FRazor		
PDB code	L	α, β	<6.0Å(%)	Best	Avg.	<6.0Å(%)	Best	Avg.
1FC2	43	2,0	20.5	2.59	7.3	38.6	2.60	6.4
1ENH	54	2,0	39.5	3.06	7.3	53.8	2.61	6.4
2GB1	56	1,4	89.8	1.88	4.3	90.6	2.04	4.4
2CRO	65	5,0	40.6	3.02	6.7	67.2	2.57	5.8
1CTF	68	3,3	9.2	3.42	9.1	11.0	3.14	8.4
4ICB	76	4,0	2.8	4.74	9.4	2.6	4.81	9.6

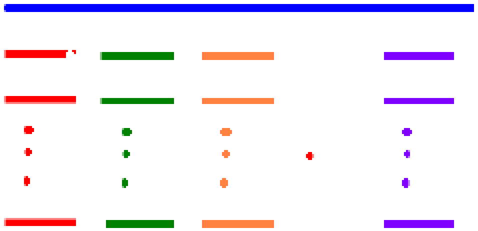
The % of good decoys were improved for five out of six targets.
The average RMSD values were improved for four out of six targets.
The best RMSD values were improved for three out of six targets.

Part III

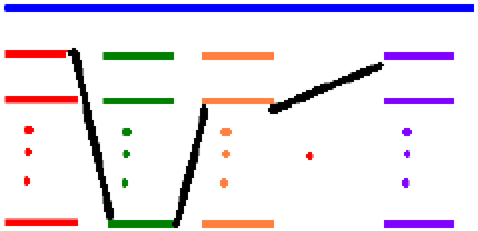
FALCON: assemble local fragments into full-length structure through sampling

ROSETTA

- Basic idea: predicting 200 local structure candidates for each 9-mer fragment, and then assembling the local structures into a full-length structure.
- Technique: using Monte Carlo technique to optimize an energy function.
- Pros and cons: the discreteness of search space implies the failure to cover the continuous conformation space. A small error of torsion angle usually incur a great RMSD.



(g) Predicting 200 local structures for each 9-mer fragment.



(h) Assembling into a full-length structure.

Our method: Fragment-HMM

Biological Insight: protein structure is result of the combination of two types of interactions:

- Local Interaction: forming local structural preference;
- Global Interaction: put all local structures into their correct positions to minimize energy.

Questions:

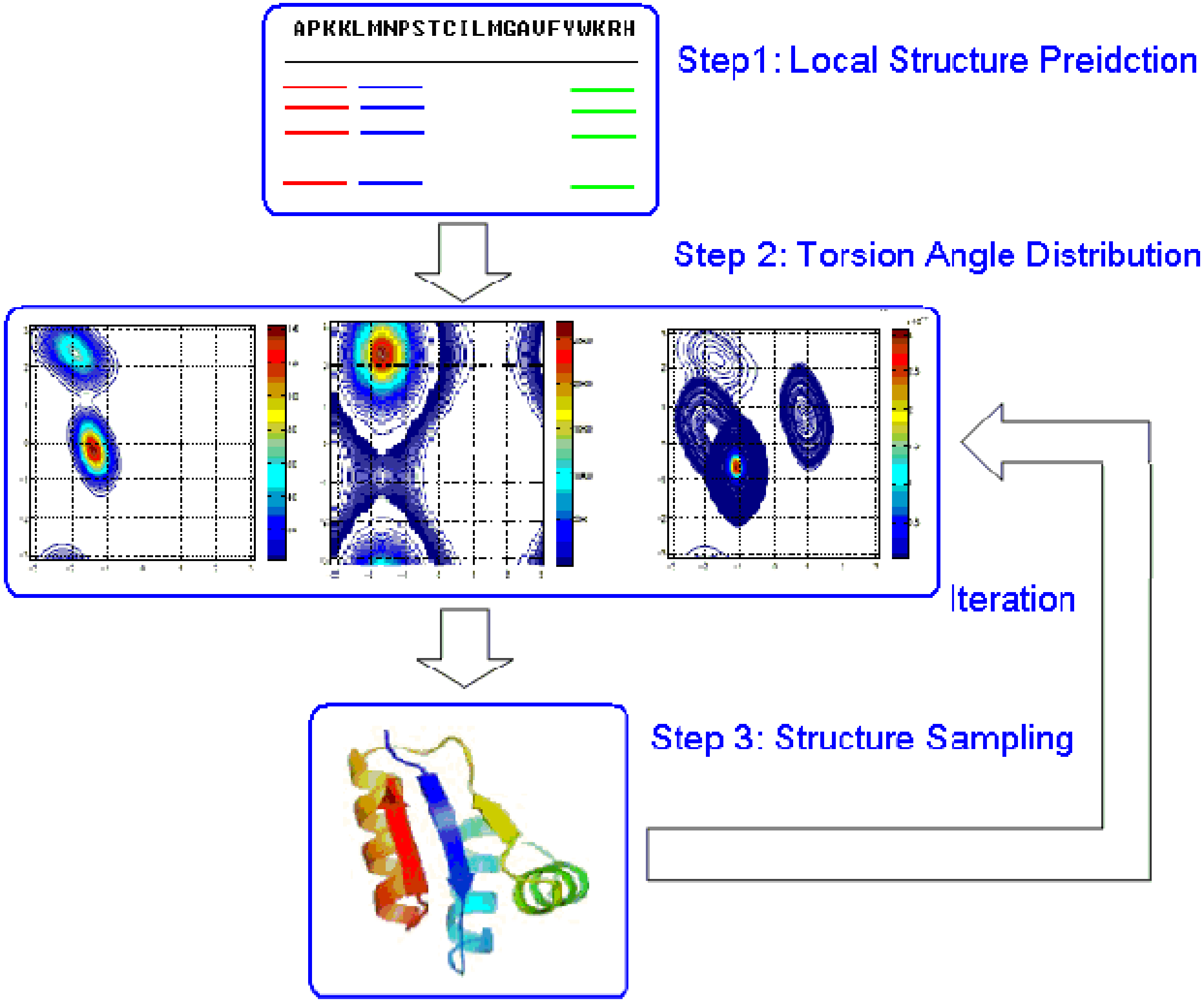
- How to describe the local structural preference?
- How to capture the dependence generated by long-distance interactions?

FALCON's paradigm

Basic idea: sampling out a full-length structure that meets local structural preference of all amino acids. More specifically,

1. for residue i, we use *Cosine* models to describe the distribution of its torsion angle (ϕ_i, ψ_i);.
2. we employ a position specific HMM to describe the dependence between neighboring residues;
3. after training the position specific HMM, we sample out (ϕ_i, ψ_i) for residue i, and use an energy function to evaluate the generated decoys;
4. finally the generated decoys are fed back to generate more accurate torsion angle estimation until convergence.

FALCON's schema


APKKLMNPSTCILMGAVFYWKRH
Step1: Local Structure Preidction
Step 2: Torsion Angle Distribution
Iteration
Step 3: Structure Sampling

Technique 1: Cosine Model

The probability density function of *Cosine* model is specified by five parameters κ_1, κ_2, κ_3, μ and ν:

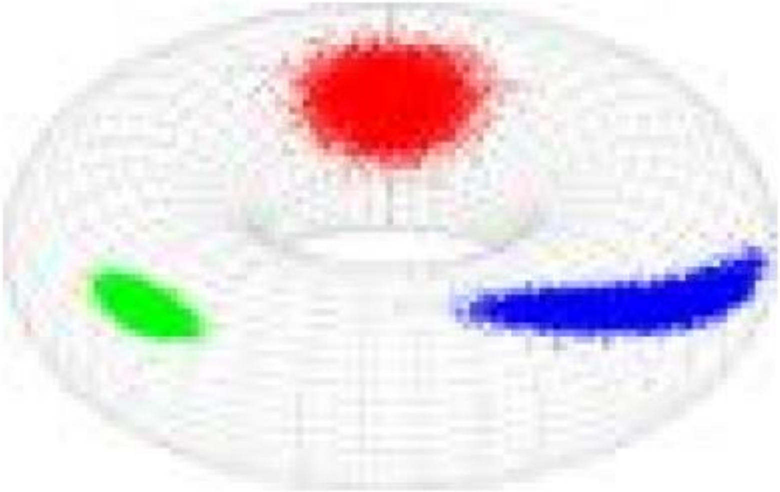
$$f(\phi, \psi) = c(\kappa_1, \kappa_2, \kappa_3) e^{\kappa_1 \cos(\phi-\mu) + \kappa_2 \cos(\psi-\nu) + \kappa_3 \cos(\phi-\mu-\psi+\nu)}$$

where μ is the mean value of ϕ, ν is the mean value of ψ, and $c(\kappa_1, \kappa_2, \kappa_3)$ is a normalization constant:

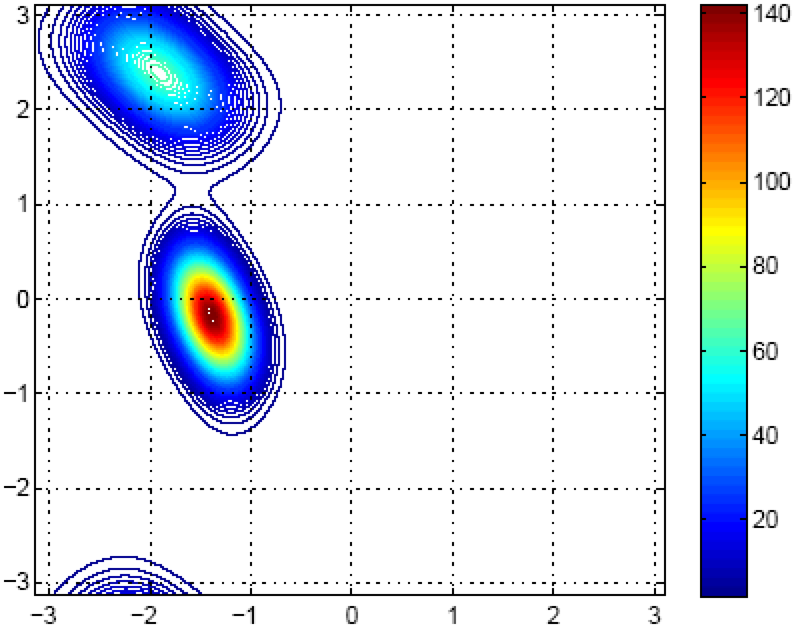
$$c(\kappa_1, \kappa_2, \kappa_3)^{-1} = (2\pi)^2 \left\{ I_0(\kappa_1) I_0(\kappa_2) I_0(\kappa_3) + 2 \sum_{p=1}^{\infty} I_p(\kappa_1) I_p(\kappa_2) I_p(\kappa_3) \right\}$$

in which $I_r(\kappa)$ is the modified Bessel function of the first kind and order r.

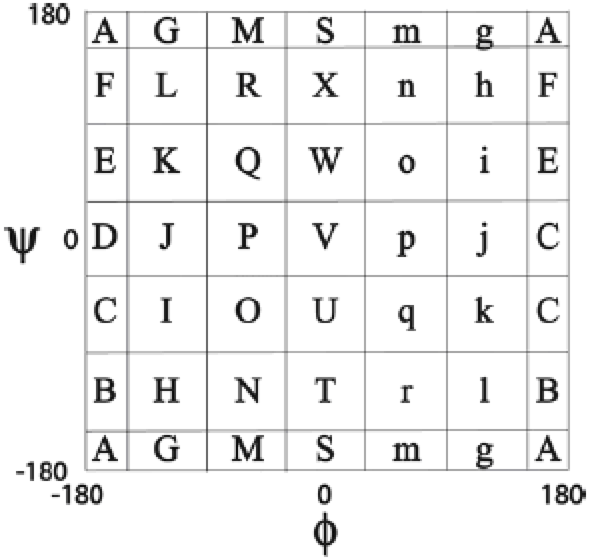
von Mises distribution



An Example: local structural preference of K13 of protein 1CTF

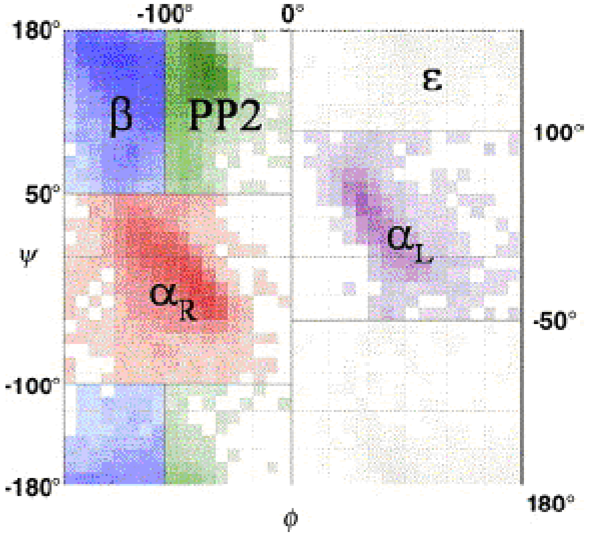


Other approaches to describe local preference.



A	G	M	S	m	g	A
F	L	R	X	n	h	F
E	K	Q	W	o	i	E
D	J	P	V	p	j	C
C	I	O	U	q	k	C
B	H	N	T	r	l	B
A	G	M	S	m	g	A

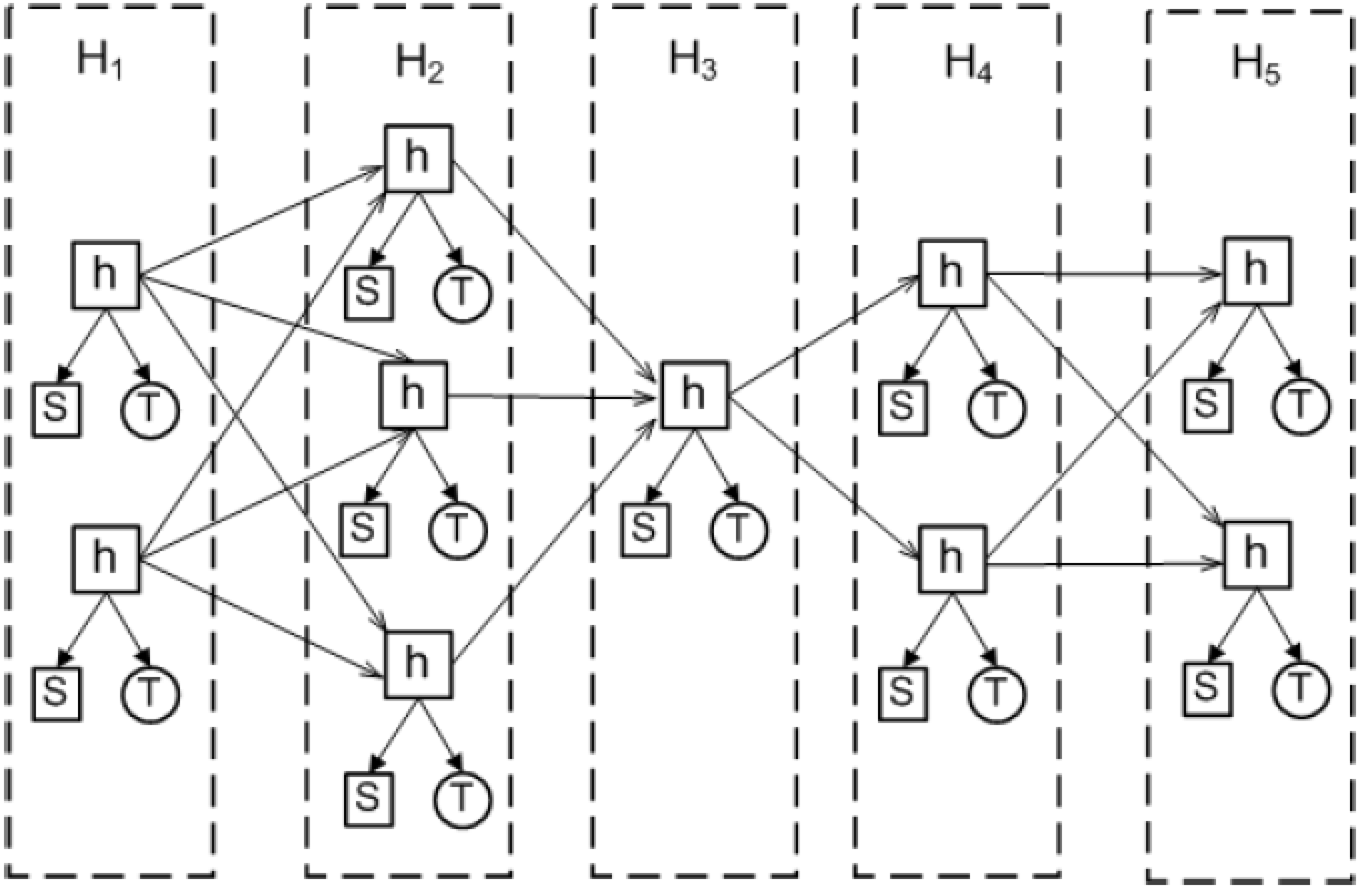
(i) Gong's method



(j) Shortle's method

Technique 2: Position-specific HMM

Model Topology: residues have specific hidden nodes and transition probabilities.



Structure of Fragment-HMM

- Hidden node: each hidden node corresponds to a *Cosine* model;
- Transition probability:

$$\Pr.(h' \in H_{i+1} | h \in H_i) \quad = \quad \frac{\Pr.(h' \in H_{i+1}, h \in H_i)}{\sum_{h' \in H_{i+1}} \Pr.(h \in H_i, h' \in H_{i+1})}$$

where

$$\Pr.(h \in H_i, h' \in H_{i+1}) = \sum_{q \in \mathcal{F}} \frac{g_h(q) g_{h'}(q)}{\sum_{h \in H_i, h' \in H_{i+1}} g_h(q) g_{h'}(q)}$$

Technique 3: Primal_Dual optimization technique

$$
\begin{aligned}
& minE(\phi_1, \psi_1, \phi_2, \psi_2, ..., \phi_n, \psi_n) \\
s.t. \quad & (\phi_i, \psi_i) c.t. f_i
\end{aligned}
$$

- Primal step: sampling the full-length structure based on distribution f_i, and using energy function to direct search process and to meet global interaction requirements;
- Dual step: re-calculating the distribution from the generated good decoys, i.e., reshaping local structural biases;

Advantage over Monte Carlo/Local Search: search space can be significantly narrowed down.

Essence: problem transformation

Discrete optimization => sampling approach to continous optimization problem

$$\begin{aligned} minE(x_1, x_2, ..., x_n) \\ s.t.\ x_i \in S_i, \|S_i\| = 200 \end{aligned}$$

➡

$$\begin{aligned} &\text{sample } x_1, x_2, ..., x_n \\ &p(x_1, x_2, ..., x_n) = \tfrac{1}{Z} e^{-E(x_1, x_2, ..., x_n)} \\ &x_i \sim f(x; \theta_i) \end{aligned}$$

Experimental results

Data Set

- We use the six proteins that were used in previous studies : Protein A (code 1FC2), Homeodomain (code 1ENH), Protein G (code 2GB1), Cro repressor (code 2CRO), Protein L7/L12 (code 1CTF) and Calbidin (code 4ICB).
- We further test FALCON on eight larger proteins with over 100 residues.

Result 1: How many states can a residue adopt?

Table: The number of Cosine models per residue. Column 2 is length. Column 3 is the number of α-helices and β-strand. Column 4-7 are numbers of residues with 1,2,3,4 *Cosine* models, respectively. Column 8 is the average number of *Cosine* models per residue.

Target Protein			# Residue				
Name,PDB code	L	α, β	1	2	3	4	Ave.
Protein A, 1FC2	43	2,0	12	25	3	2	1.66
Homeodomain, 1ENH	54	2,0	24	24	6	0	1.21
Protein G, 2GB1	56	1,4	28	21	7	0	1.63
Cro repressor, 2CRO	65	5,0	52	12	1	0	1.22
Protein L7/L12, 1CTF	68	3,3	50	14	3	1	1.34
Calbidin, 4ICB	76	4,0	47	23	3	3	1.50

how many structural conformation can a protein adopt?

The number of possible protein conformations (or search space):

- $C = 200^n$ by ROSETTA.
- $C = 75^n$ by Hamelryck *et al.*
- $C = 1.66^n$ by FALCON.
- $C = 1.6^n$ by Sims and Kim, Dill, *et al.*

This observation suggests that:

1. Local structural preference significantly limit the number of possible structural conformations.
2. It is possible to sample a native-like structure since the search space is significantly narrowed donw.

Result 2: Discrete optimization vs. continuous optimization, which one is better?

Table: Decoy quality of ROSETTA and FALCON. Column 2-3: RMSD of the best decoy ($\mathring{A}$) and percentage of the good decoys (RMSD$<$ 6$\mathring{A}$) for ROSETTA. Column 4-5: corresponding values for FALCON.

Target Protein	ROSETTA		FALCON	
	Best	<6.0Å(%)	Best	<6.0Å(%)
Protein A, 1FC2	2.82	80.2	2.64	94.3
Homeodomain, 1ENH	1.52	94.4	1.81	92.8
Protein G, 2GB1	2.21	53.7	2.18	93.4
Cro repressor, 2CRO	2.56	70.4	2.48	75.8
Protein L7/L12, 1CTF	1.44	14.3	0.56	25.6
Calbidin, 4ICB	3.87	19.9	2.93	46.3

The decoy quality increases as iteration proceeds.

Energy function was used to capture global interactions, and therefore may help to reshape the local biases.

Table: RMSD distribution over iterations for protein 2CRO. Col. 2-7: Percentages of decoys with RMSD values in the corresponding intervals.

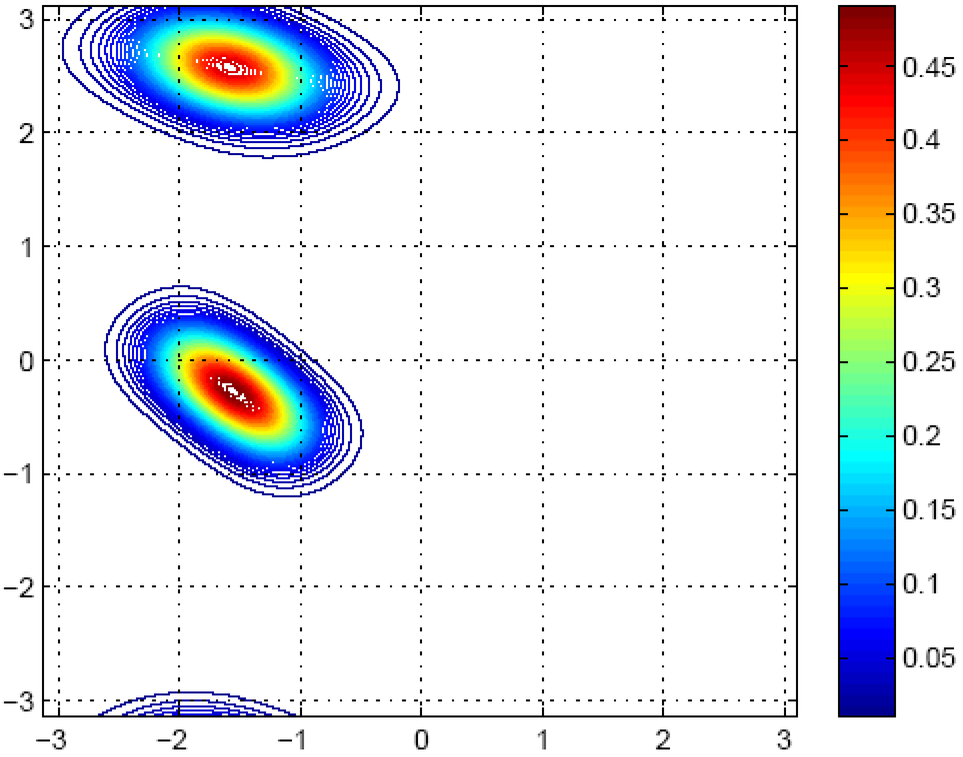
	#Iterations					
RMSD (Å)	1	2	3	4	5	6
$[0, 3)$	0.1	0	0.1	0.1	0	0
$[3, 4)$	22.8	47.2	75.3	87.9	94.7	94.9
$[4, 5)$	41.5	45.4	24.5	12.0	5.3	5.1
$[5, 6)$	11.4	4.7	0.1	0	0	0
$[6, 7)$	8.5	0.8	0	0	0	0
$[7, \infty)$	15.7	1.5	0	0	0	0

The "Good decoy" ratio also increases as iteration proceeds, and can reach 100% on the six proteins.

Table: Percentage of good decoys with RMSD below 6Å after each iteration.

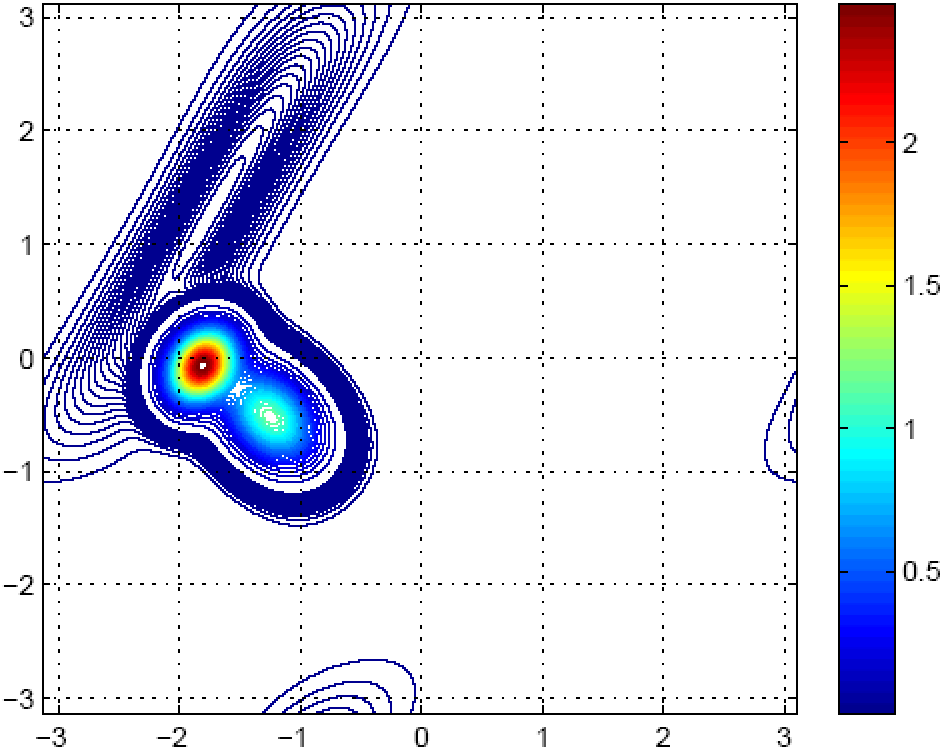
Target Protein	# Iterations					
	1	2	3	4	5	6
Protein A, 1FC2	94.3	98.5	100	100	100	100
Homeodomain, 1ENH	92.8	95.0	96.9	100	100	100
Protein G, 2GB1	93.4	96.4	100	100	100	100
Cro repressor, 2CRO	75.8	97.3	100	100	100	100
Protein L7/L12, 1CTF	25.6	68.8	97.0	100	100	100
Calbidin, 4ICB	46.3	90.5	99.3	100	100	100

Reshaping Local Bias: Iteration 1



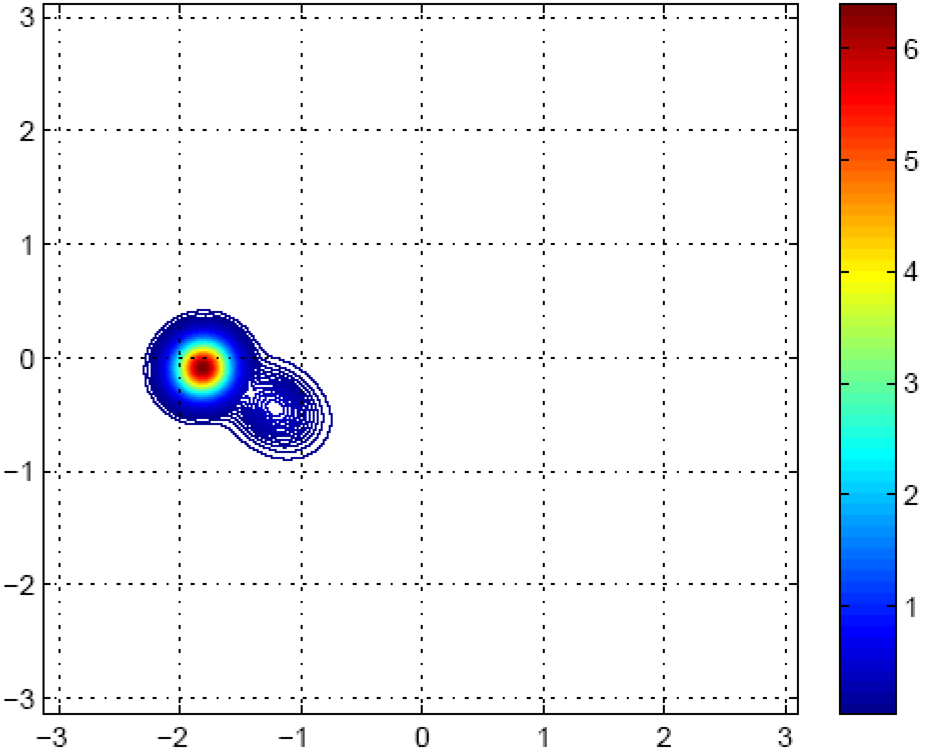
(k) Iteration #1: Two Cosine models centered at $(-1.55, -0.28)$ and $(-1.58, 2.57)$.

Reshaping Local Bias: Iteration 2



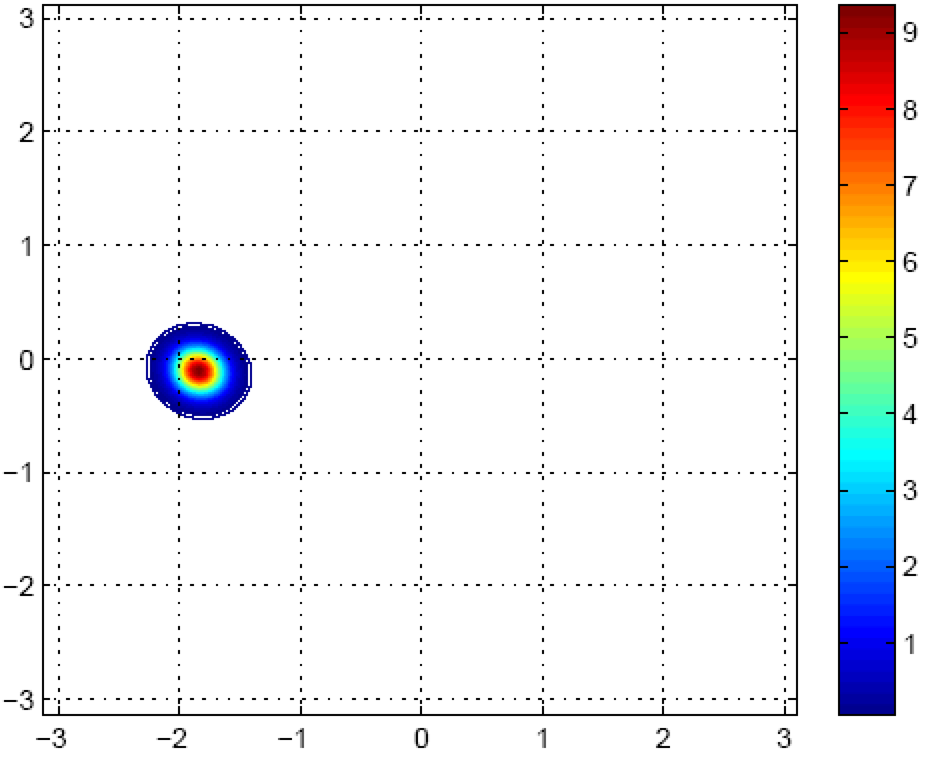
(l) Iteration #2: Three Cosine models centered at $(-1.25, -0.52)$, $(-1.75, 1.26)$, and $(-1.82, -0.07)$.

Reshaping Local Bias: Iteration 3



(m) Iteration #3: Two Cosine models centered at $(-1.22, -0.44)$ and $(-1.82, 0.09)$.

Reshaping Local Bias: Iteration 4



(n) Iteration #4: One Cosine model centered at (−1.84, −0.11).

Result 4: the quality of the final prediction results.

Table: Quality of the finally reported decoys of ROSETTA and FALCON for the six benchmark proteins. Column 2-3: RMSD ($\AA$) of the finally chosen decoys of ROSETTA and FALCON.

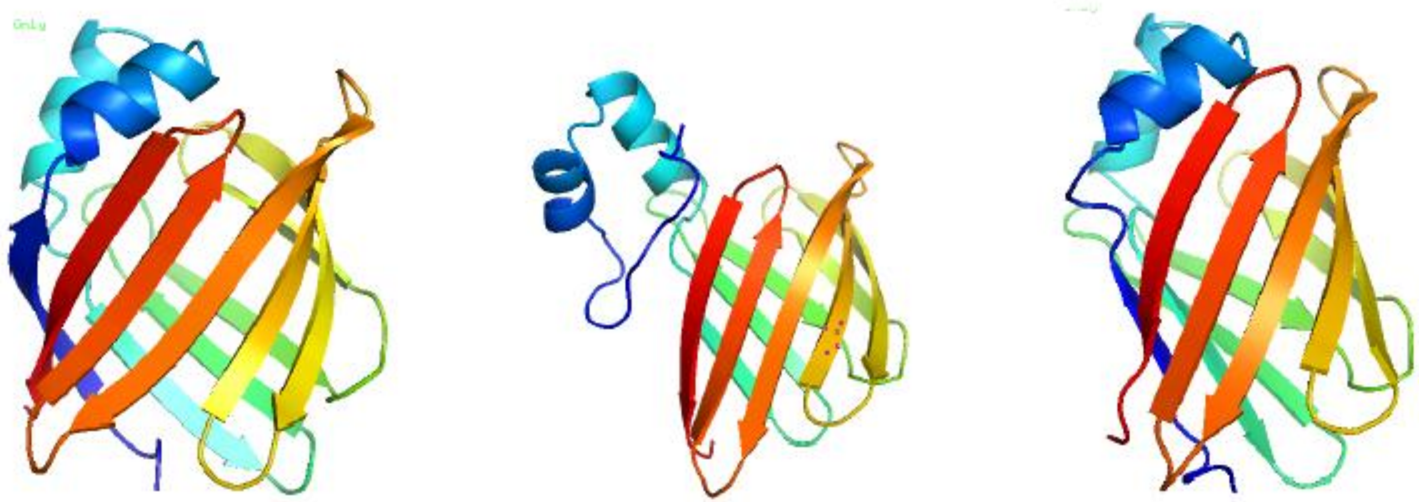
Target Protein	ROSETTA	FALCON
Protein A, 1FC2	3.660	3.652
Homeodomain, 1ENH	2.717	2.464
Protein G, 2GB1	2.755	3.323
Cro repressor, 2CRO	3.997	3.477
Protein L7/L12, 1CTF	8.327	3.035
Calbidin, 4ICB	4.866	4.770

Experimental results on CASP-7 targets.

Table: Quality of the finally reported decoys of ROSETTA and FALCON for eight larger proteins from CASP7 free modeling targets. Column 4-5: RMSD ($\AA$) of the finally chosen decoys of ROSETTA and FALCON.

Target Protein	PDB Entry	Length	ROSETTA	FALCON
T0283	2HH6	112	11.544	11.083
T0300	2H3R	102	7.557	9.282
T0307	2H5N	133	14.822	16.343
T0327	2HGC	102	9.394	11.149
T0350	2HC5	117	10.635	7.406
T0354	2ID1	130	11.254	8.085
T0361	2HKT	169	20.009	12.225
T0373	2HR3	147	19.097	14.224

Example 1: prediction results for 1IFB (RMSD$= < 2.0\mathring{A}$)



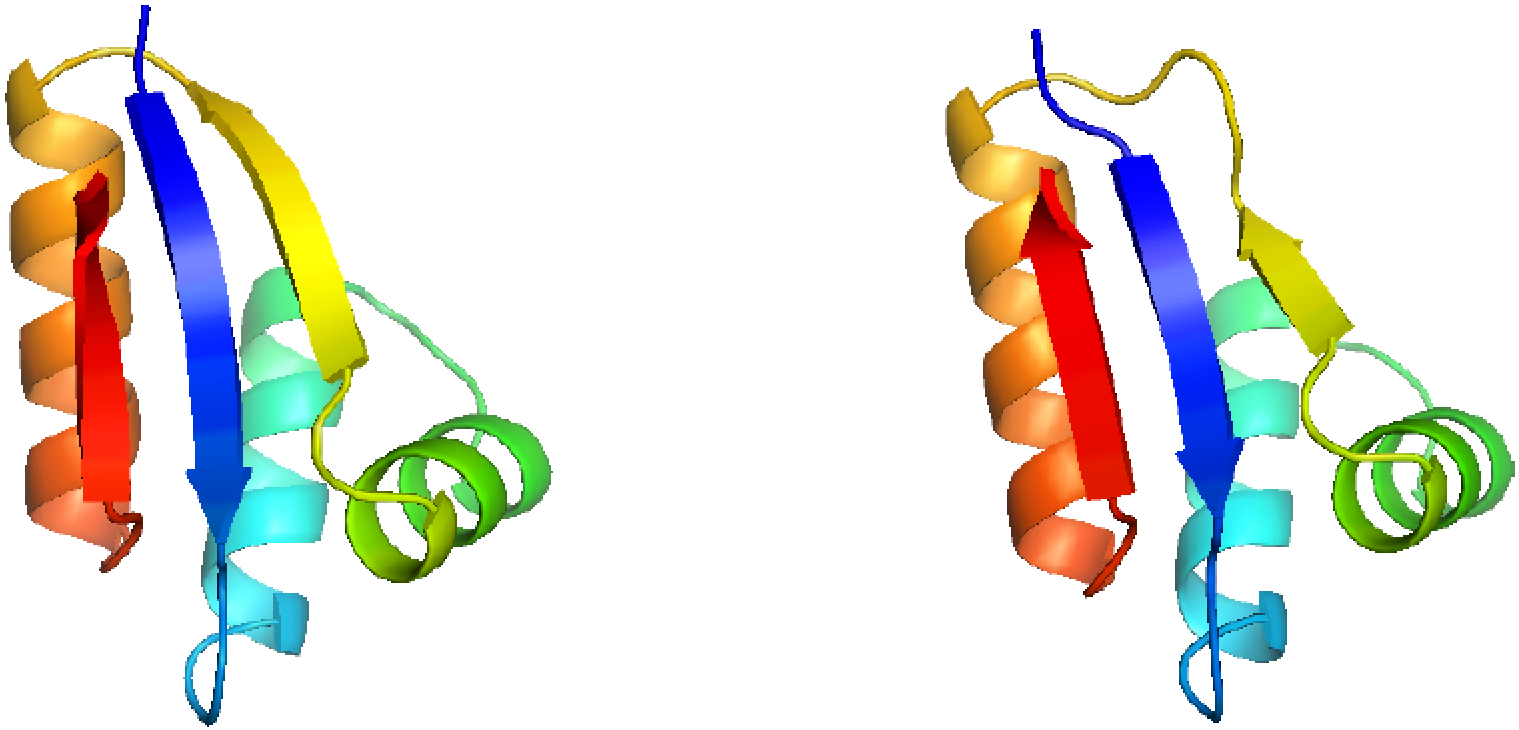
(o) The native structure of protein 1IFB.

(p) The final decoy reported by FALCON.

(q) The best decoy reported by FALCON.

Figure: The Native Structure, the Final Decoy, and the Best Decoy Reported by by FALCON.

Example 2: prediction result for 1CTF (RMSD: 0.557Å)



(a) The native structure of protein 1CTF.

(b) The predicted structure of protein 1CTF.

Figure: The Native Structure and the Best Decoy Predicted by FALCON. The RMSD is 0.557Å.

Summary

1. discrete optimization$\Rightarrow$ continuous optimization: in principle can explore all the conformation space;
2. Monte Carlo$\Rightarrow$Primal Dual: the search space is reduced from $O(200^n)$ to $O(1.66^n)$; thus, the probability to sample a native-like conformation is increased.
3. FALCON was ranked 3rd in the FR-Hard category in CASP8.

Ongoing...

- Identifying sampling bottlenecks: some fragments are forced segments since their local structural preference is changed by global interaction.
- Improving predictions for proteins with complex topology.
- Designing a more accurate energy function.

Acknowledgement

- Collaborators
  - Shuai Cheng Li
  - Jinbo Xu
  - Sheng Wang
  - Weimou Zheng
  - Ming Li
- Grants:
  - NSFC
- Students
  - Xiongying Yuan
  - Mingfu Shao
  - Chao Wang
  - Haicang Zhang
  - Chunlin Huang