Residue contact and β-sheet Protein Structures Prediction

Jérôme Waldispühl, PhD
School of Computer Science,
McGill Centre for Bioinformatics,
McGill University

http://csb.cs.mcgill.ca

Introduction

Secondary structure prediction methods reached 77% per residue accuracy (E.g. PSI-pred), but performance are weaker on β -strands.

Why?

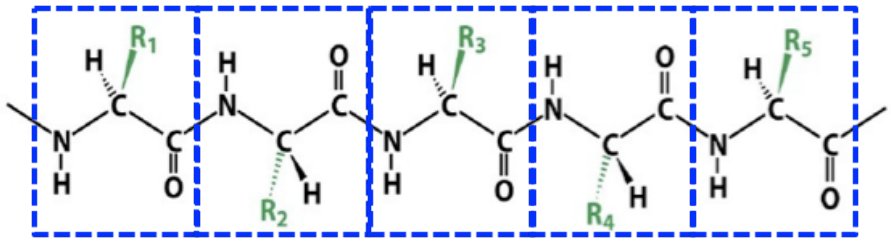
β -sheets are stabilized through long-range contacts. Other, elements (e.g. helices) may benefit of long-range contacts too.

Objective:

Prediction of residue contacts and super-secondary structure.

Protein Structure

Sequence

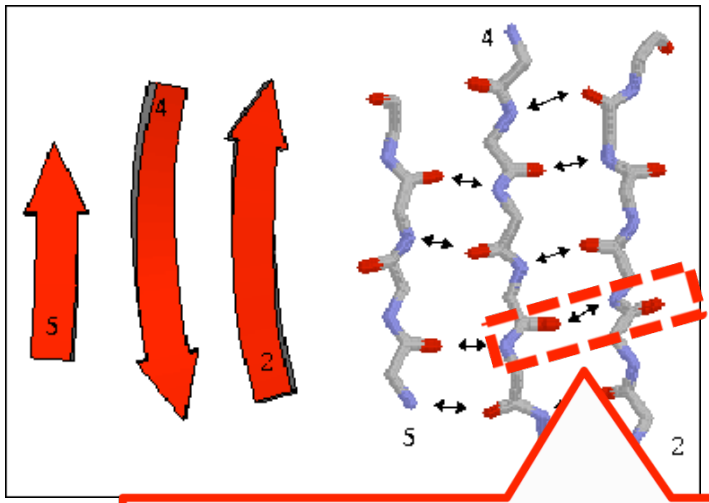


A_1 A_2 A_3 A_4 A_5

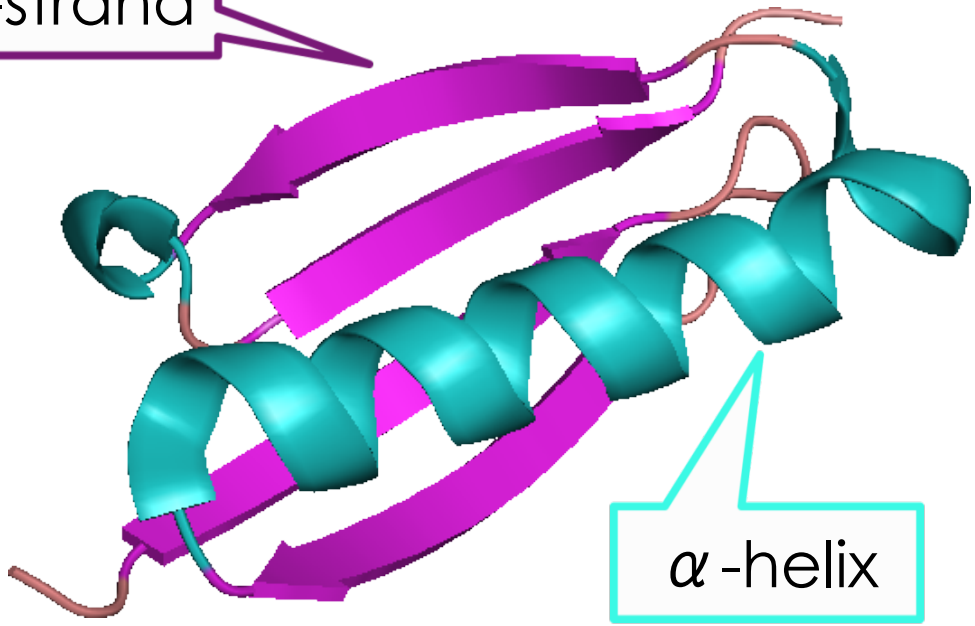
Structure

β -strand

β -sheet

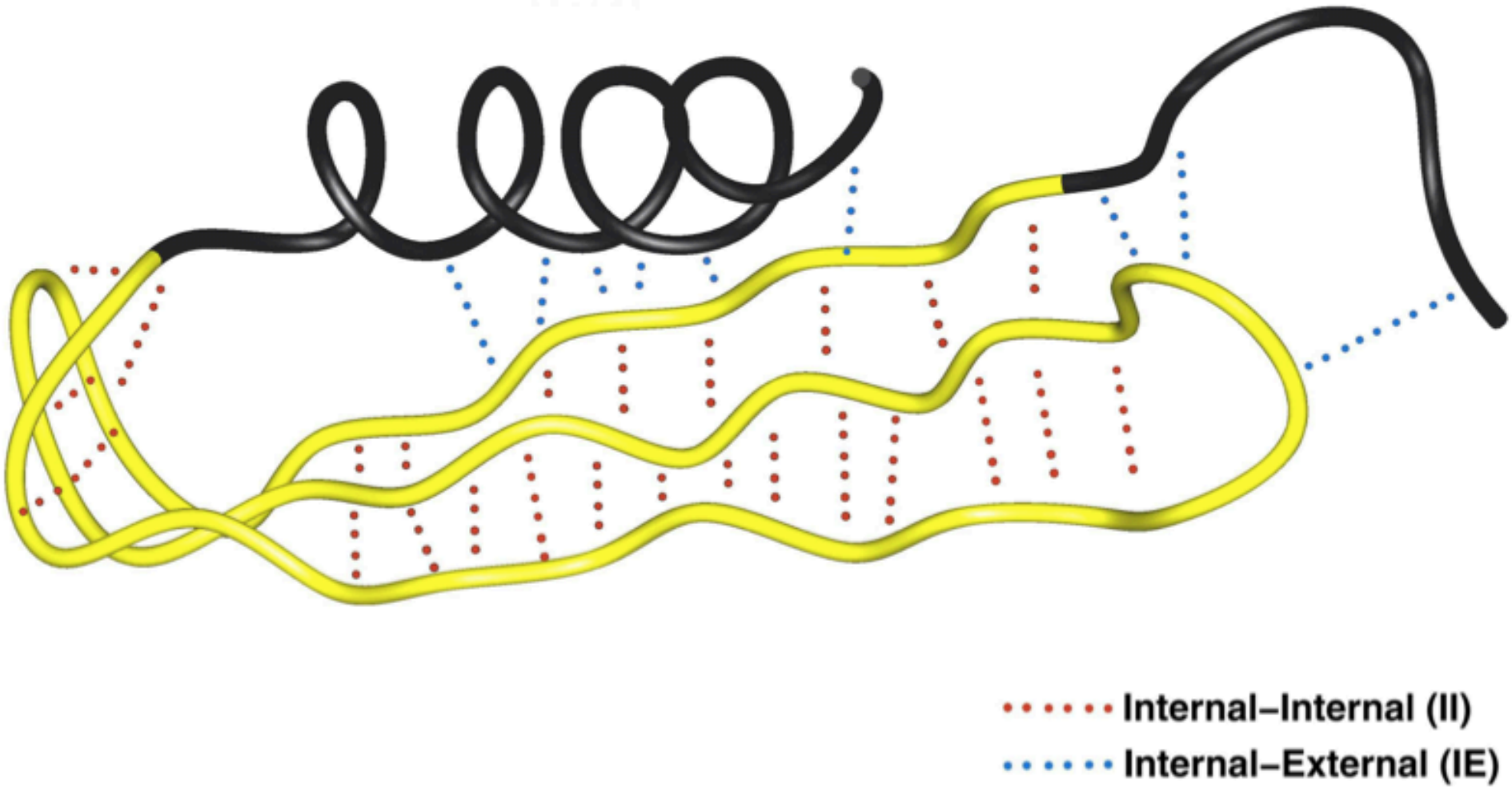


Long-range interactions stabilize β –sheet.

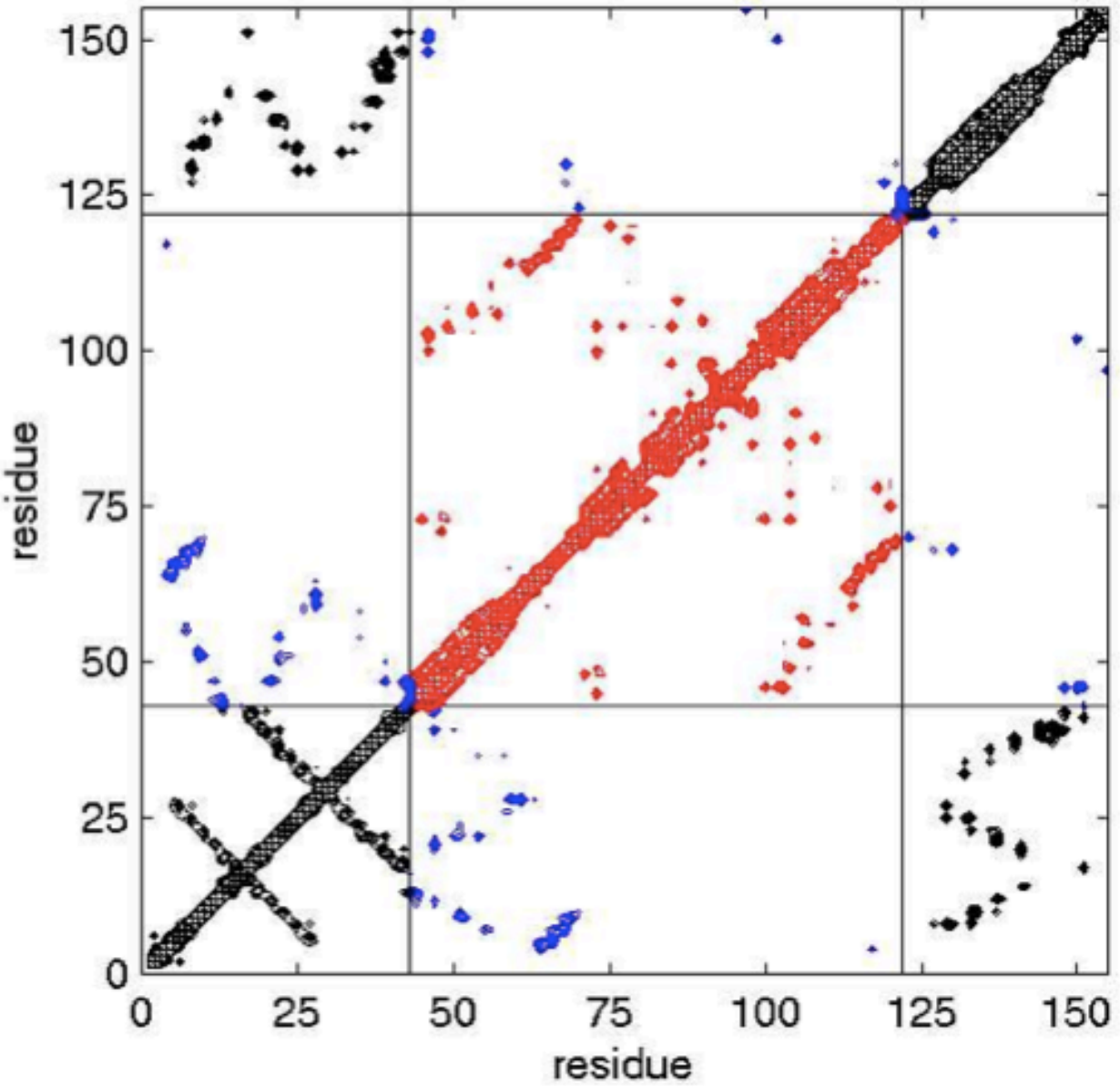


α -helix

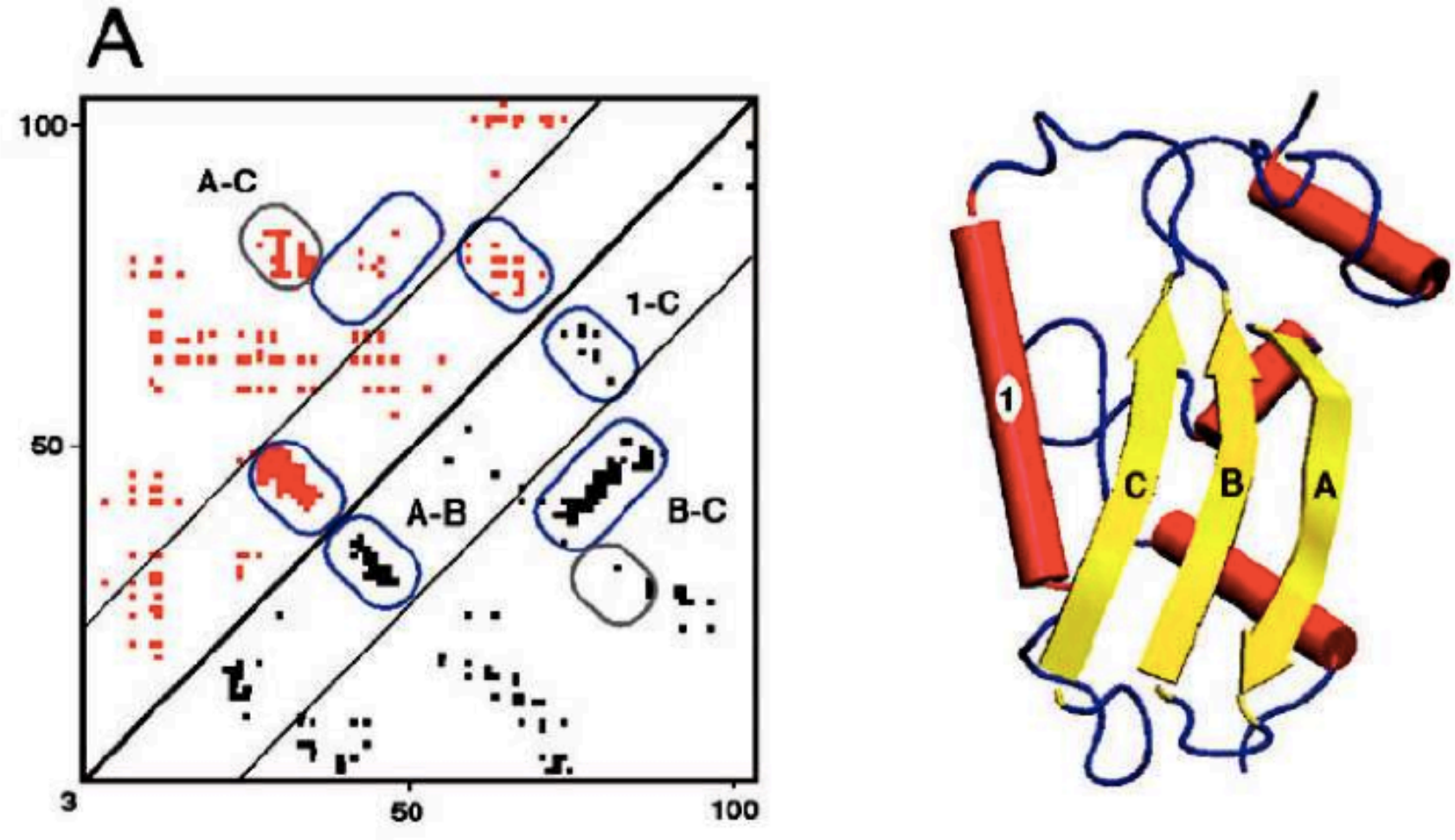
Residue contacts



Contact maps



Interpreting contact maps



Overview

- **Profcon**: Neural Network Approach
- **BetaPro**: Hierarchical Using Neural Networks, Alignments, and Graph Algorithms
- **SMURF**: Markov Random Field
- **tFolder**: Abstract template & ensemble prediction

Neural Network Architecture

Combine secondary structure and solvent accesibility predictions, and multiple sequence alignment information.

- Local information from immediate residue environment,
- Local information from connecting segment,
- Global information.

Local information from immediate residue environment

For each residue i, j involved, we define a window of size 9 centered around them.

Each residue position is characterized by :

- 20 units for the evolutionary informations (frequency of the 20 obtained from a PSI-BLAST multiple sequence alignment at 80% of homology),
- 1 unit N- or C- terminal,
- 4 units for the secondary structure prediction (PROFphd),
- 3 units for the solvent accessibility prediction (PROFphd),
- 1 unit for the conservation weight.

Total : 18 = 522 units + coarse grained biophysical classification of the contact (hydrophobic-hydrophobic, polar-polar, charged-polar, opposite charges, same charge, aromatic-aromatic, others)

Global information

Local information from connecting segment

Design a window for the five consecutive residue that spanned the interval.

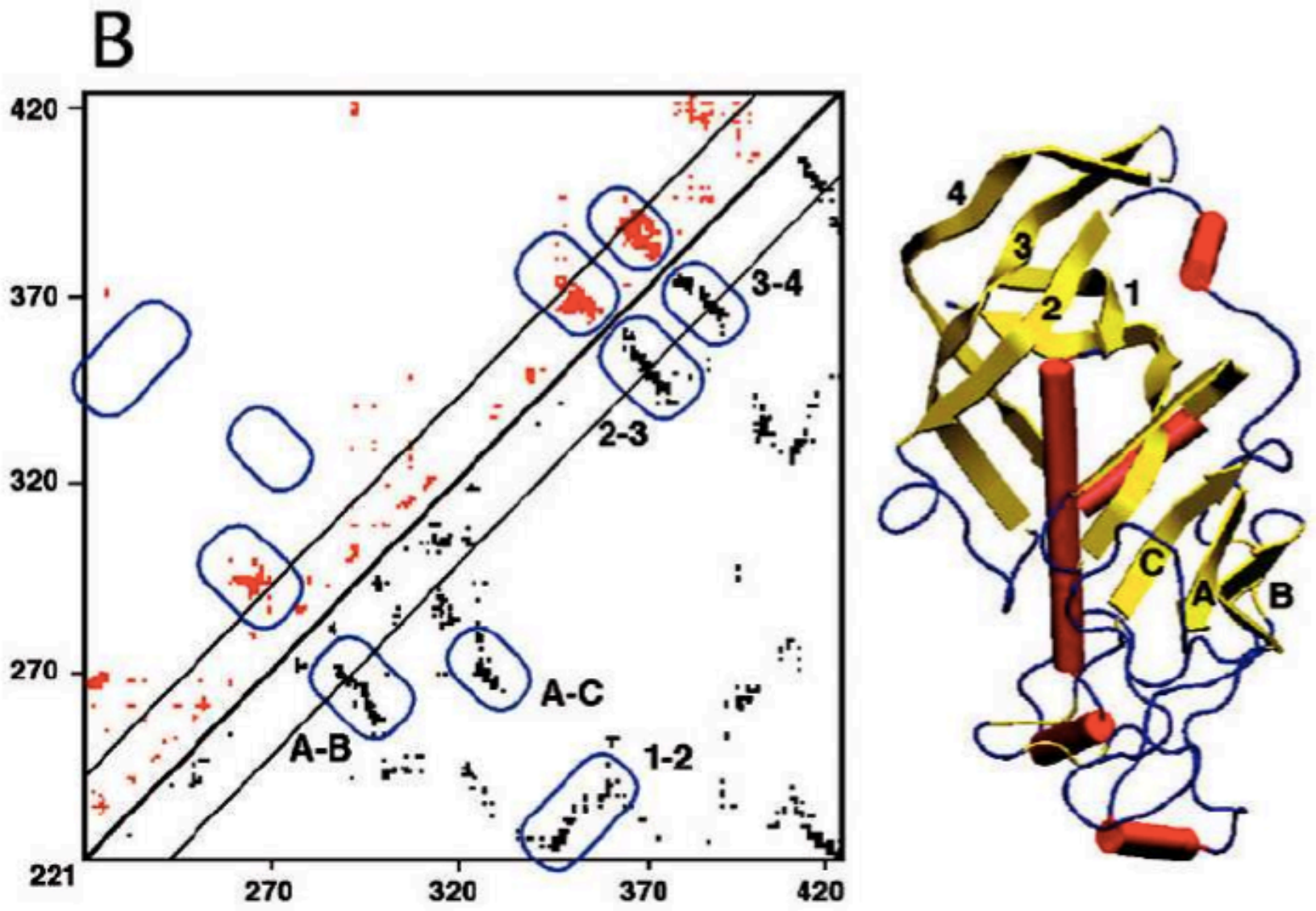
Each residue position is characterized by the same informations as contact residues + length of the segment, amino acid and secondary structure composition, SEG-low-complexity.

Total: 180 units

Global information

23 units for amino acid and secondary structure composition + 4 units for the protein length.

PROFcon@CASP6



PROFcon Results

Table 1. Benefit from using connecting segments

Method[a]	Sequence separation[b]	Nprot[c]	Acc[d]	Err^{e}_{Acc}	Cov[d]	Err^{e}_{Cov}
Local only	6	633	29.7	0.5	8.6	0.1
PROFcon	6	633	32.4	0.5	9.8	0.2
Local only	24	621	19.5	0.5	8.8	0.3
PROFcon	24	621	20.0	0.5	9.4	0.3

Table 2. Improvement through evolutionary information

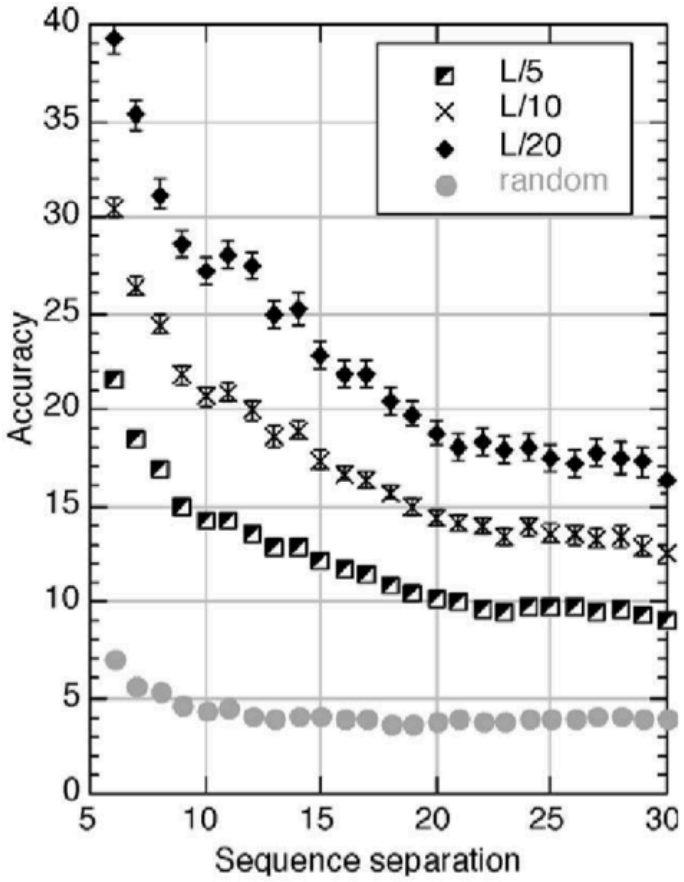
Nali[a]	Sequence separation[b]	Nprot[b]	Acc[b]	Err^{b}_{Acc}	Cov[b]	Err^{b}_{Cov}
0–14	6	138	23.0	1.0	9.5	0.8
15–49		123	31.0	1.0	9.8	0.5
50–199		187	35.6	0.9	9.7	0.3
≥200		185	37.0	0.9	10.2	0.4
0–14	24	132	13.2	0.7	10.0	1.0
15–49		120	19.0	1.0	9.1	0.7
50–199		185	21.5	0.9	9.0	0.4
≥200		184	24.2	0.9	9.5	0.4

[a] Number of proteins in multiple sequence alignment used to extract evolutionary profiles.
[b] As in Table 1; note all values compiled for the first $L/2$ predictions (Methods).

PROFcon Results: Length & sequence separation

Table 3. Performance versus protein length

L^a	Sequence separation[b]	Nprot[b]	Acc[b]	Err^b_{Acc}	Cov[b]	Err^b_{Cov}
≤100	6	78	31.0	2.0	18.0	1.0
101–200		230	32.5	0.9	9.7	0.3
201–300		191	33.9	0.9	8.6	0.2
301–400		134	31.0	1.0	7.1	0.1
All	6	633	32.4	0.5	9.8	0.2
≤100	24	66	19.0	1.0	22.0	2.0
101–200		230	18.5	0.7	8.8	0.3
201–300		191	22.0	1.0	8.2	0.3
301–400		134	19.4	0.8	6.2	0.2
All	24	621	20.0	0.5	9.4	0.3



PROFcon Results: Structural classes

Table 4. Performance differs between structural classes

SCOP class	Sequence separation[a]	Nprot[a]	Acc[a]	Err^{a}_{Acc}	Imp[a]	Err^{a}_{Imp}	Cov[a]	Err^{a}_{Cov}
All-alpha	6	131	24.0	1.0	14.3	0.8	11.0	0.7
All-beta		103	35.0	1.0	13.0	0.7	7.8	0.3
Alpha + beta		119	36.0	1.0	14.8	0.6	10.0	0.4
Alpha/beta		169	35.9	0.8	19.4	0.5	8.7	0.2
All-alpha	24	128	13.5	0.6	9.4	0.6	9.9	0.8
All-beta		103	17.0	1.0	8.3	0.6	6.6	0.4
Alpha + beta		118	17.0	1.0	8.8	0.5	7.8	0.5
Alpha/beta		169	29.0	1.0	18.5	0.6	10.0	0.3

[a] As in Table 1; note all values compiled for the first $L/2$ predictions (Methods).

PROFcon Discussion

- Connecting segment are very informative for contact formation,
- Evolutionary profile are crucial,
- Contact density dependent on type of protein,
- Similar accuracy but better performance for short proteins,
- α worst and α / β best,
- Of 50 % within 2 residues of observed contact,
- Correct for core, hydrophobic and regular secondary structure.
- **Limitations** : very large training set required (400,000 contacts).

Overview

- **Profcon**: Neural Network Approach
- **BetaPro**: Hierarchical Using Neural Networks, Alignments, and Graph Algorithms
- **SMURF**: Markov Random Field
- **tFolder**: Abstract template & ensemble prediction

Three-Stage Prediction of Beta-Sheets

- Stage 1

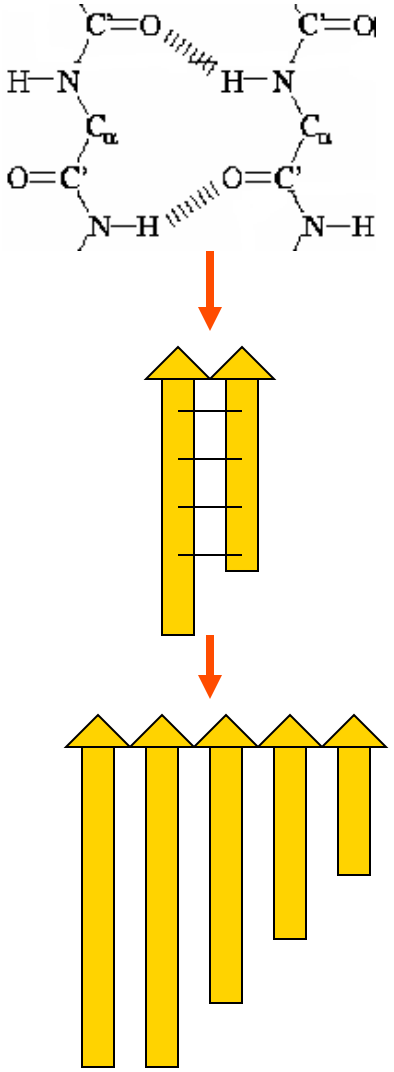
  Predict beta-residue pairing probabilities using 2D-Recursive Neural Networks (2D-RNN, Baldi and Pollastri, 2003)

- Stage 2

  Use beta-residue pairing probabilities to align beta-strands

- Stage 3

  Predict beta-strand pairs and beta-sheet architecture using graph algorithms



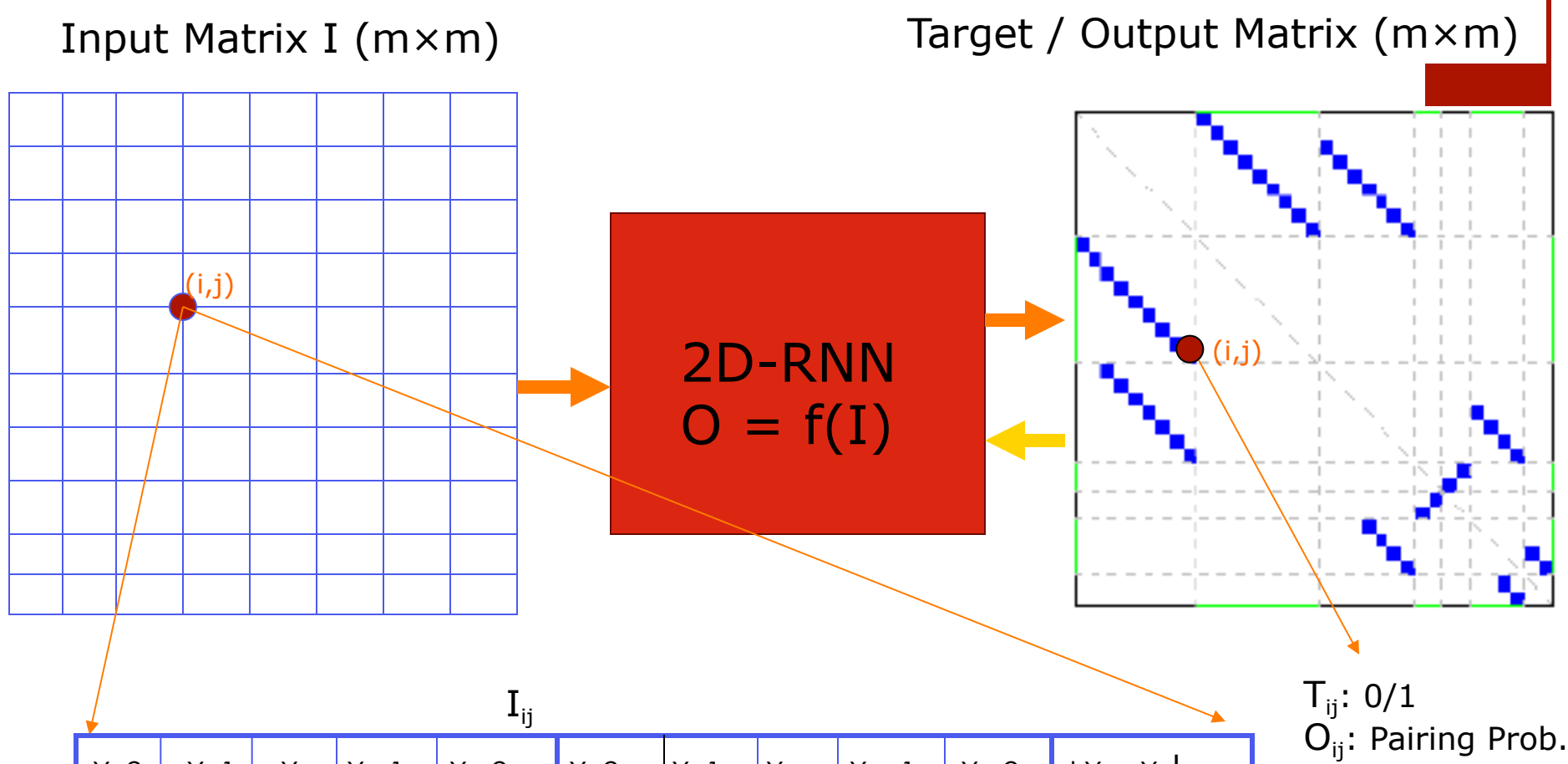
Dataset and Statistics

- Extract proteins with high resolution from Protein Data Bank (Berman et al., 2000)
- Use DSSP (Kabsch and Sander, 1983) to assign intra-chain beta-sheet structure
- Use UniqueProt (Mika and Rost, 2003) to reduce redundancy
- Use PSI-BLAST (Altschul et al., 1997) to generate profiles

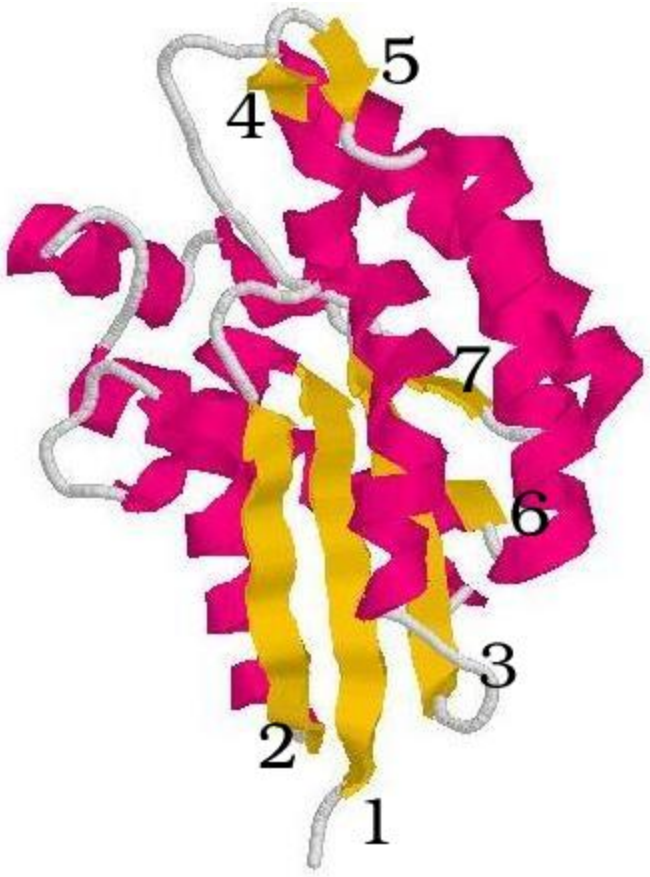
Statistics

	Num
Protein Chains	916
Beta Residues	48,996
Beta Residue Pairs	31,638
Beta Strands	10,745
Beta Strand Pairs	8,172
Beta Sheets	2,533

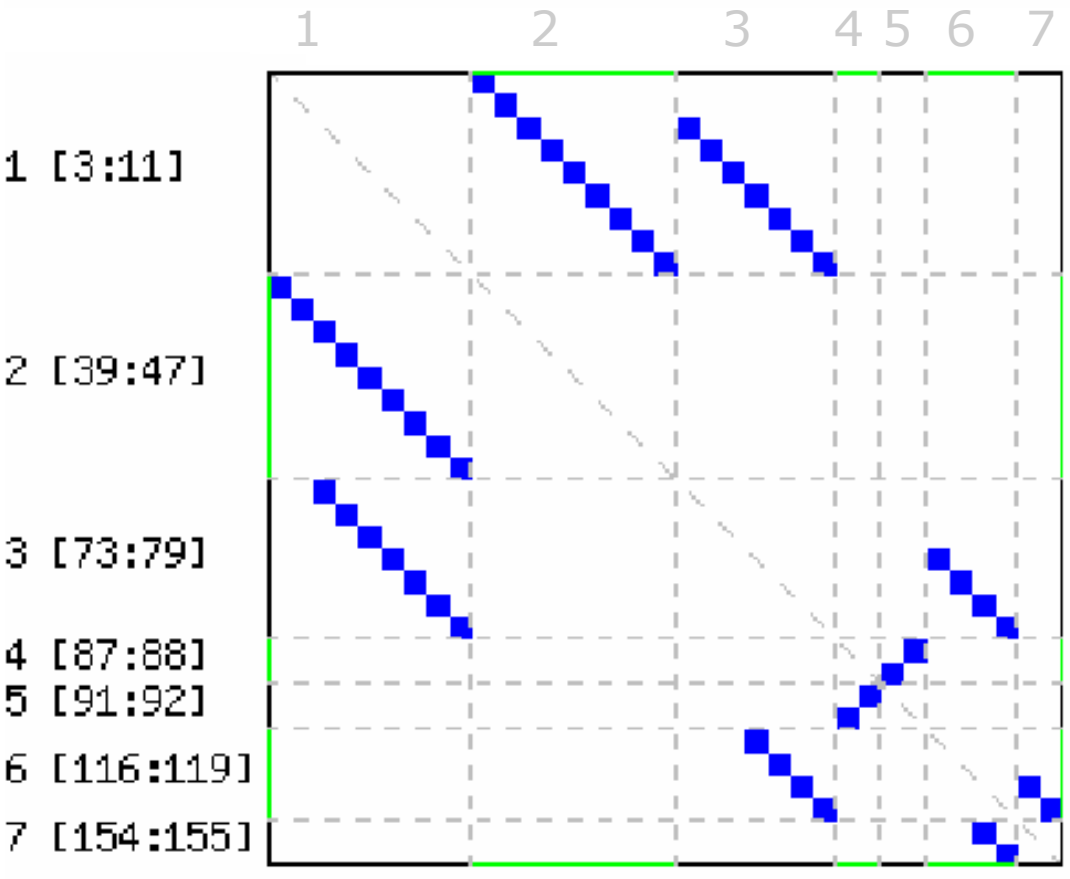
Stage 1: Prediction of Beta-Residue Pairings Using 2D-RNN



An Example (Target)

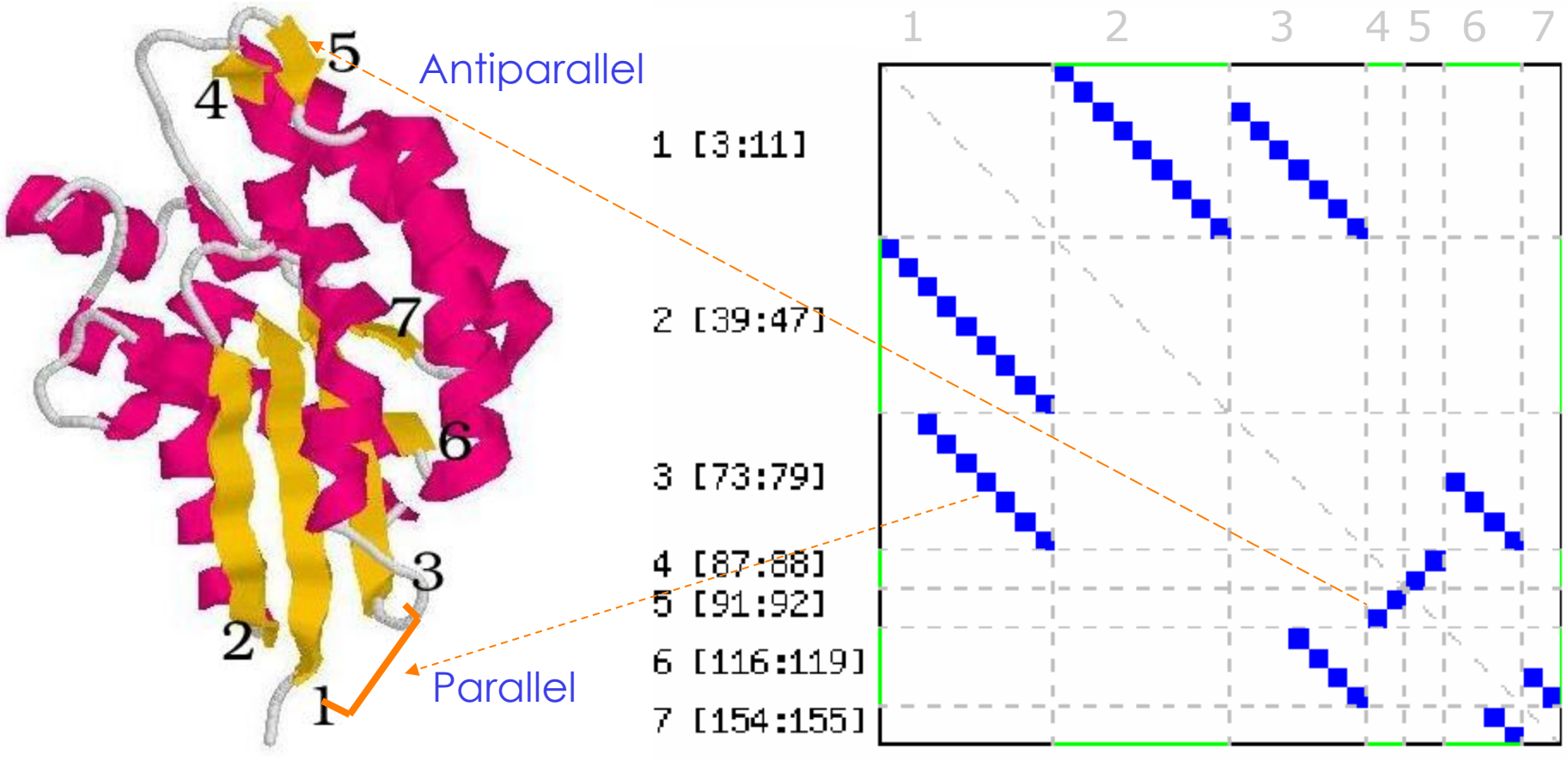


Protein 1VJG



Beta-Residue Pairing Map (Target Matrix)

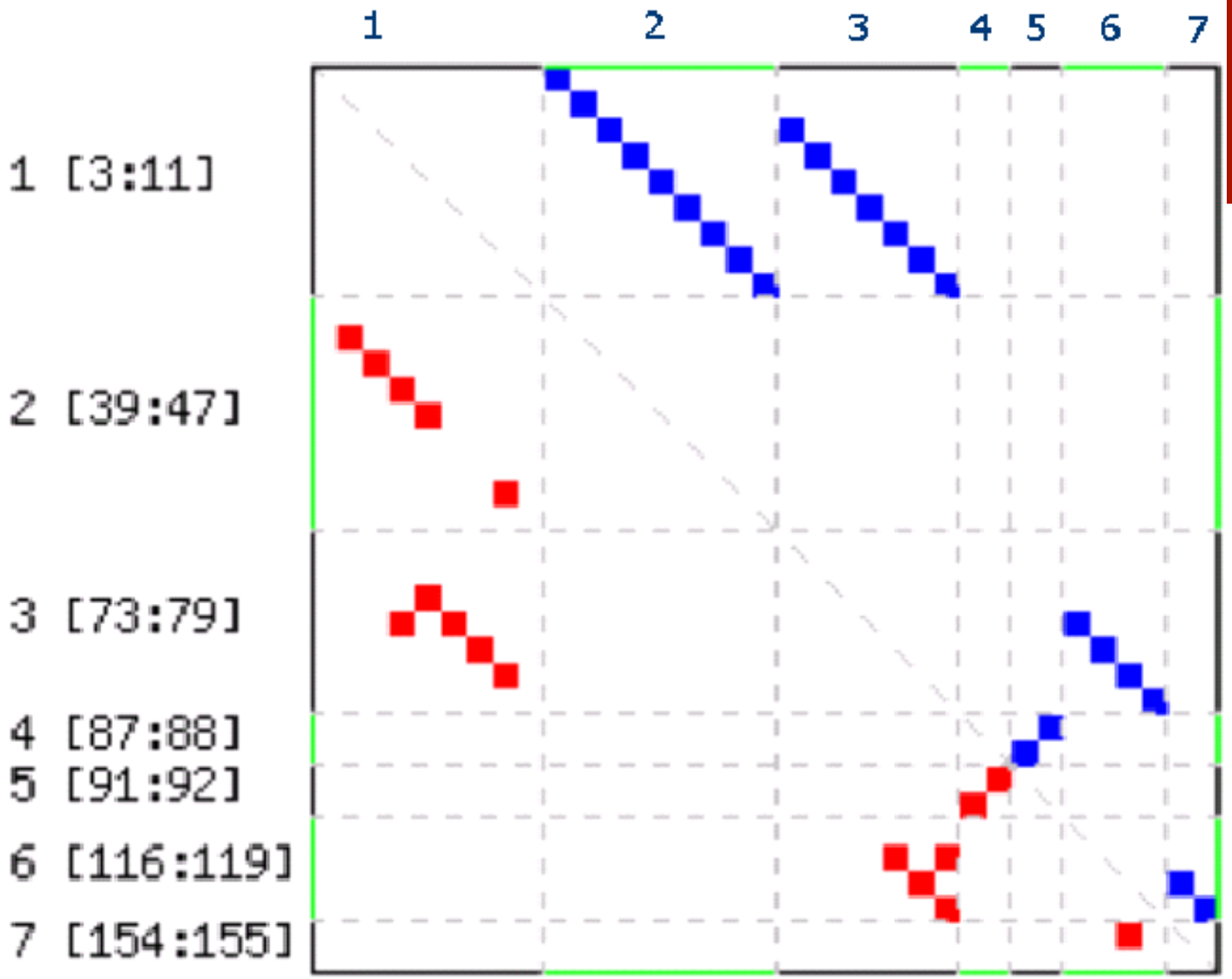
An Example (Target)



Protein 1VJG

Beta-Residue Pairing Map (Target Matrix)

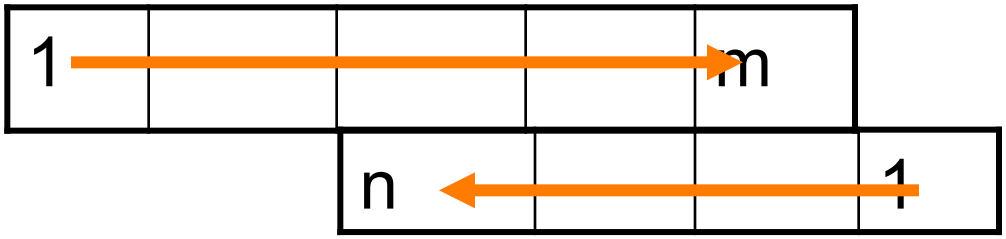
An Example (Prediction)



Stage 2: Beta-Strand Alignment

- Use output probability matrix as scoring matrix
- Dynamic programming
- Disallow gaps and use the simplified search algorithm

Antiparallel



Parallel

Total number of alignments = 2(m+n-1)

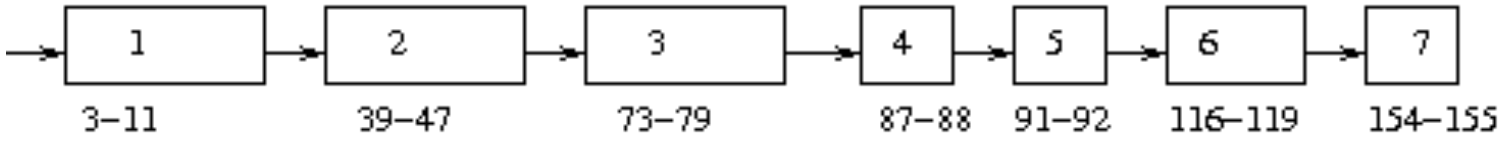
Strand Alignment and Pairing Matrix

- The alignment score is the sum of the pairing probabilities of the aligned residues
- The best alignment is the alignment with the maximum score
- Strand Pairing Matrix

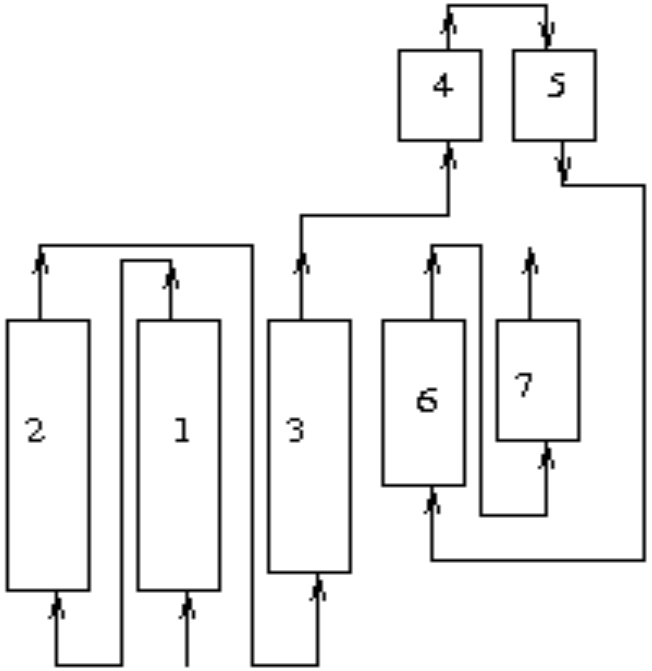
	1	2	3	4	5	6	7
1	0						
2	1.3	0					
3	.94	.37	0				
4	.02	.02	.04	0			
5	.02	.02	.03	1.9	0		
6	.10	.05	.74	.04	.04	0	
7	.02	.02	.03	.02	.02	.20	0

Strand Pairing Matrix of 1VJG

Stage 3: Prediction of Beta-Strand Pairings and Beta-Sheet Architecture (Constraints)

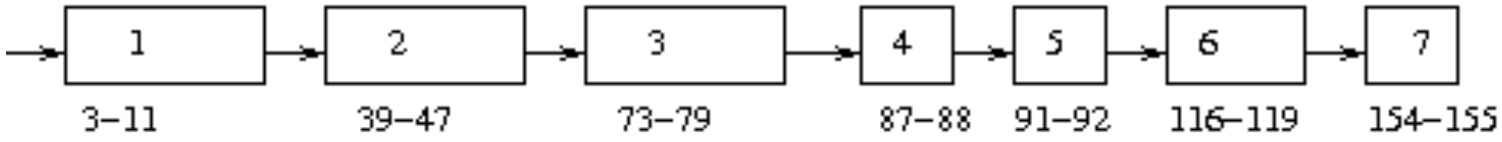


(a) Seven strands of protein 1VJG in sequence order

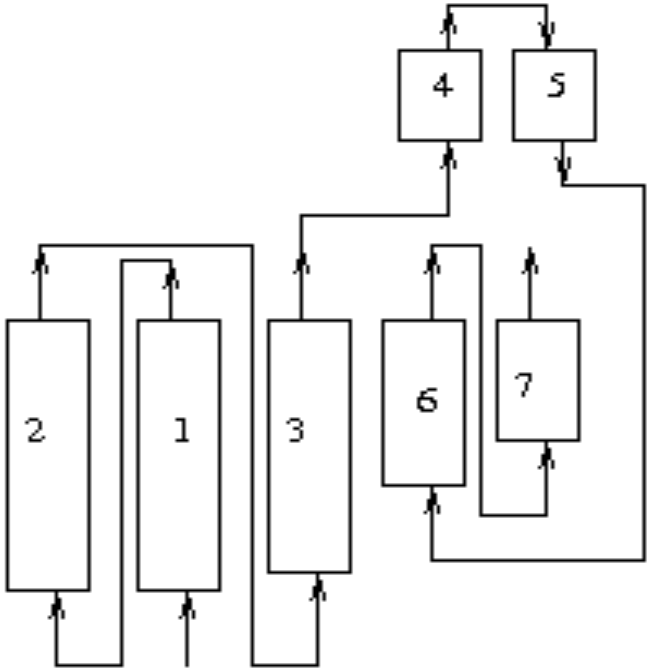


(b) Beta-sheet topology of protein 1VJG

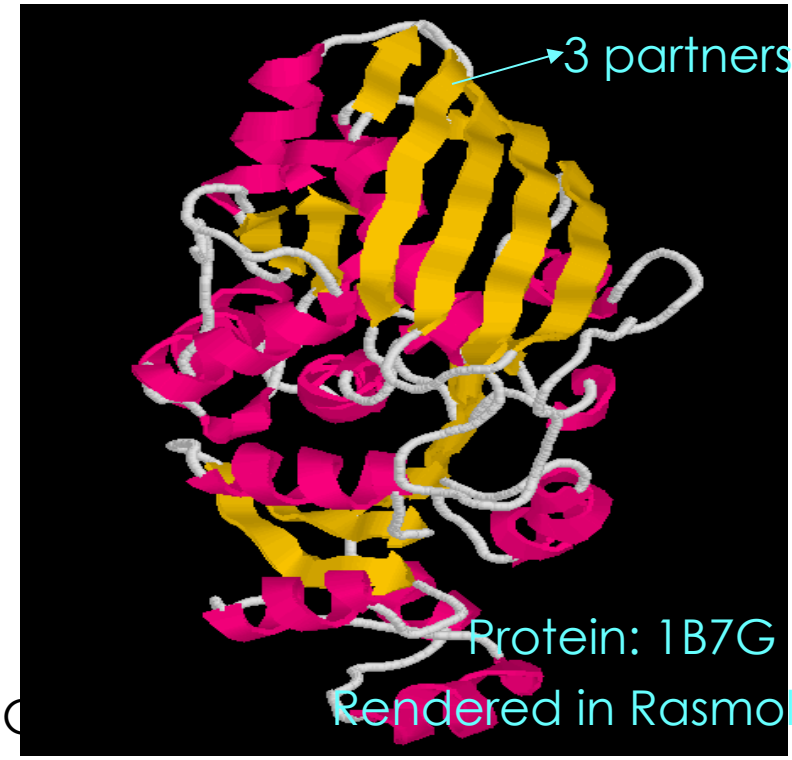
Stage 3: Prediction of Beta-Strand Pairings and Beta-Sheet Architecture (Constraints)



(a) Seven strands of protein 1VJG in sequence order



(b) Beta-sheet topology of protein 1VJG

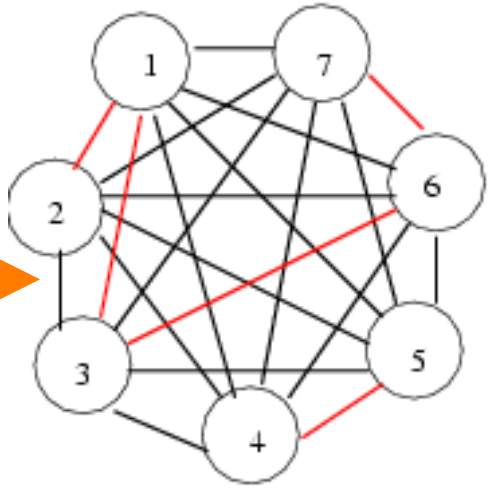


Minimum Spanning Tree Like Algorithm

Strand Pairing Graph (SPG)

	1	2	3	4	5	6	7
1	0						
2	1.3	0					
3	.94	.37	0				
4	.02	.02	.04	0			
5	.02	.02	.03	1.9	0		
6	.10	.05	.74	.04	.04	0	
7	.02	.02	.03	.02	.02	.20	0

Strand Pairing Matrix



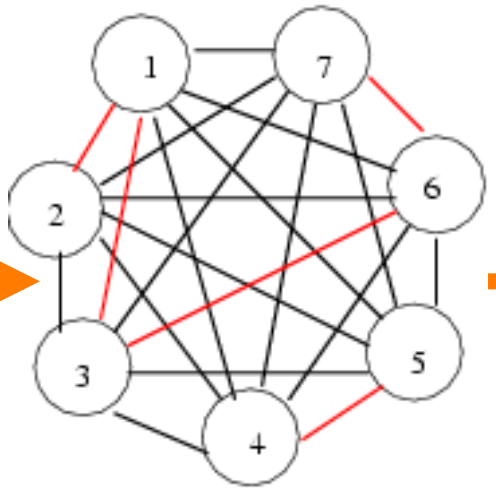
(a) Complete SPG

Minimum Spanning Tree Like Algorithm

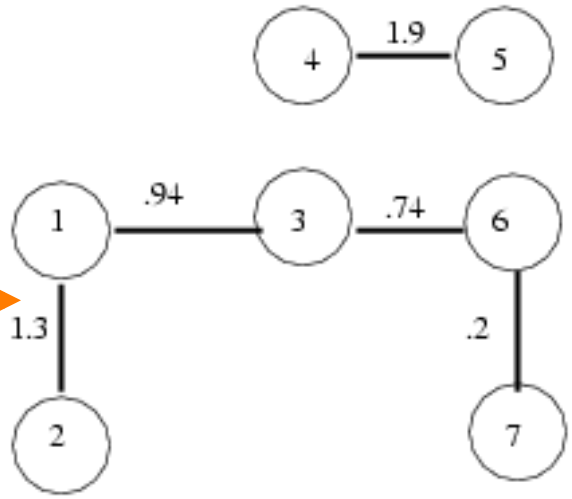
Strand Pairing Graph (SPG)

	1	2	3	4	5	6	7
1	0						
2	1.3	0					
3	.94	.37	0				
4	.02	.02	.04	0			
5	.02	.02	.03	1.9	0		
6	.10	.05	.74	.04	.04	0	
7	.02	.02	.03	.02	.02	.20	0

Strand Pairing Matrix



(a) Complete SPG



(b) True Weighted SPG

Goal: Find a set of connected subgraphs that maximize the sum of the alignment scores and satisfy the constraints

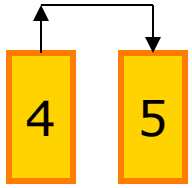
Algorithm: Minimum Spanning Tree Like Algorithm

An Example of MST Like Algorithm

	1	2	3	4	5	6	7
1	0						
2	1.3	0					
3	.94	.37	0				
4	.02	.02	.04	0			
5	.02	.02	.03	1.9	0		
6	.10	.05	.74	.04	.04	0	
7	.02	.02	.03	.02	.02	.20	0

Strand Pairing Matrix of 1VJG

Step 1: Pair strand 4 and 5

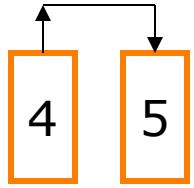


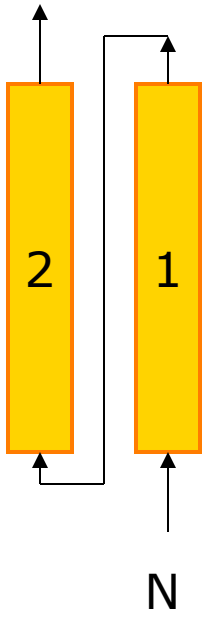
An Example of MST Like Algorithm

	1	2	3	4	5	6	7
1	0						
2	1.3	0					
3	.94	.37	0				
4	.02	.02	.04	0			
5	.02	.02	.03	1.9	0		
6	.10	.05	.74	.04	.04	0	
7	.02	.02	.03	.02	.02	.20	0

Strand Pairing Matrix of 1VJG

Step 2: Pair strand 1 and 2



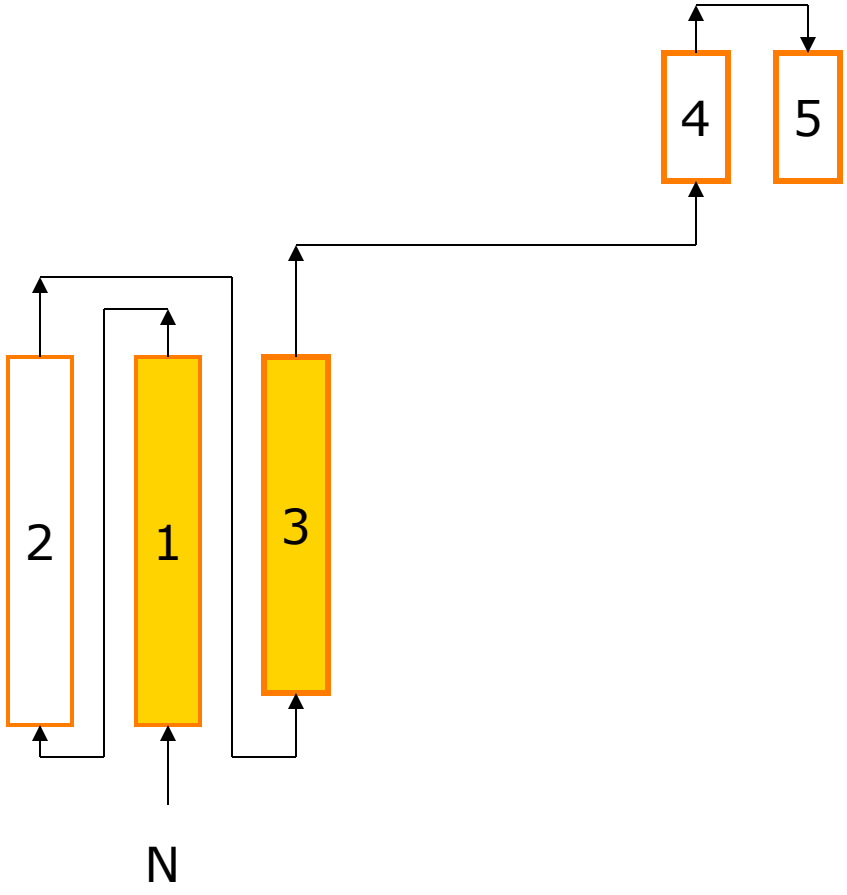


An Example of MST Like Algorithm

	1	2	3	4	5	6	7
1	0						
2	1.3	0					
3	.94	.37	0				
4	.02	.02	.04	0			
5	.02	.02	.03	1.9	0		
6	.10	.05	.74	.04	.04	0	
7	.02	.02	.03	.02	.02	.20	0

Strand Pairing Matrix of 1VJG

Step 3: Pair strand 1 and 3

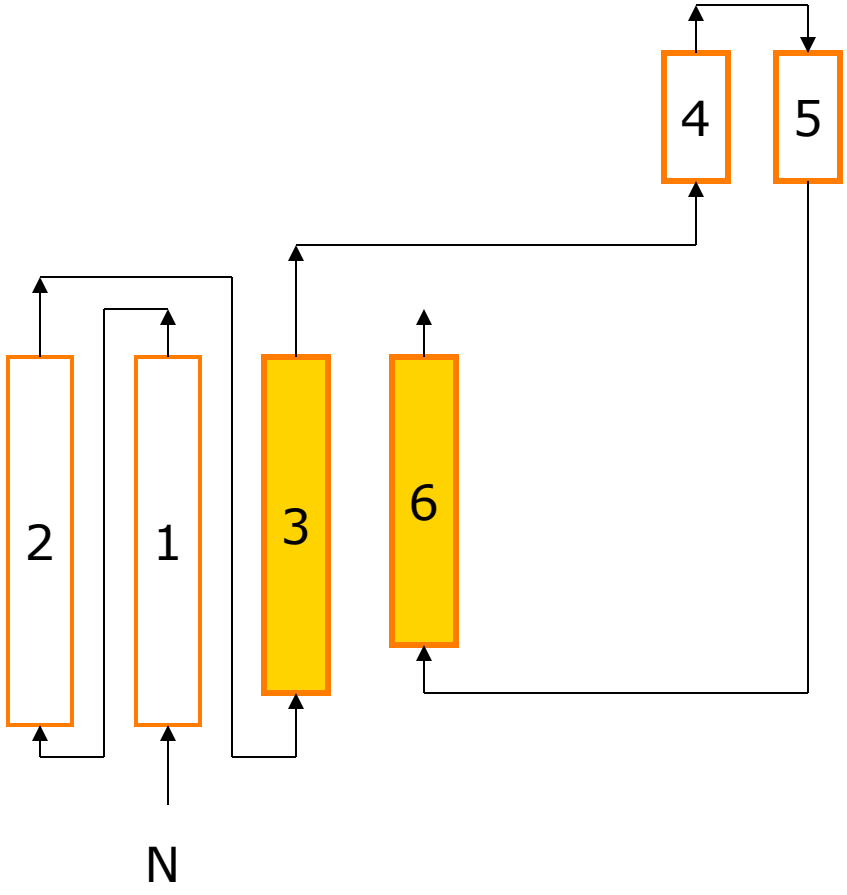


An Example of MST Like Algorithm

	1	2	3	4	5	6	7
1	0						
2	1.3	0					
3	.94	.37	0				
4	.02	.02	.04	0			
5	.02	.02	.03	1.9	0		
6	.10	.05	.74	.04	.04	0	
7	.02	.02	.03	.02	.02	.20	0

Strand Pairing Matrix of 1VJG

Step 4: Pair strand 3 and 6

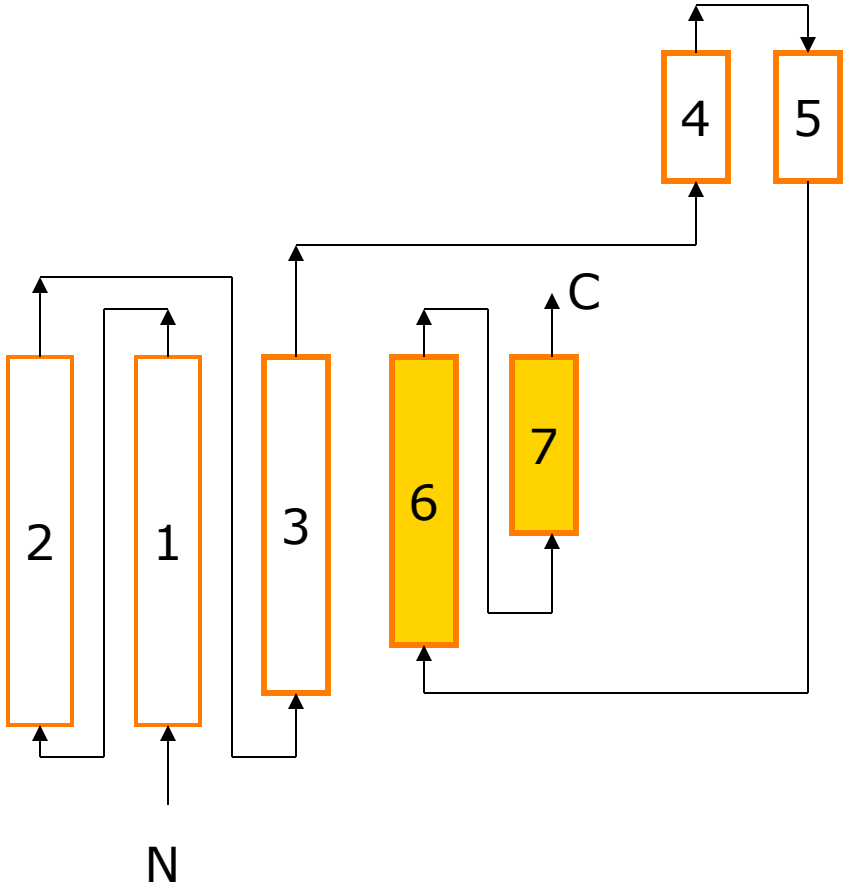


An Example of MST Like Algorithm

	1	2	3	4	5	6	7
1	0						
2	1.3	0					
3	.94	.37	0				
4	.02	.02	.04	0			
5	.02	.02	.03	1.9	0		
6	.10	.05	.74	.04	.04	0	
7	.02	.02	.03	.02	.02	.20	0

Strand Pairing Matrix of 1VJG

Step 5: Pair strand 6 and 7



A New Fold Example (Last CASP)

1S12 (94 residues)

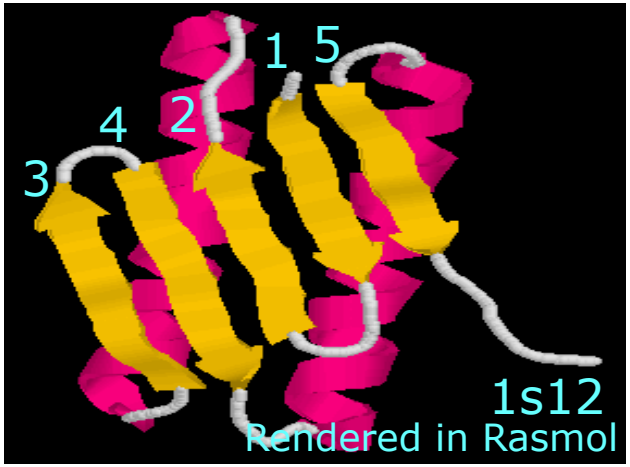
True secondary structure

CEEEEECCCEEEEECCCCCHHHHHHHHHHHHHHHHHHHCCCEEEEEECCEEEEEECCCCHHHHHHHHHHHHHHHHHHHCCCCEEEEECCCCCC

Predicted secondary structure by SSpro (Pollastri, et al., 2002)

CEEEEEECCEEEECCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHEHHCCCCEEEEHHHHHHHHHHHHHHHHHHHHHHHCCCCEEEEEEECCC

Beta Sheet Topology



Strand Pairing Matrix

	1	2	3	4	5
1	0	1.71	.05	.29	.33
2		0	.06	.41	.12
3			0	.22	.04
4				0	.53
5					0

True: 1-2, 2-4, 3-4, 1-5

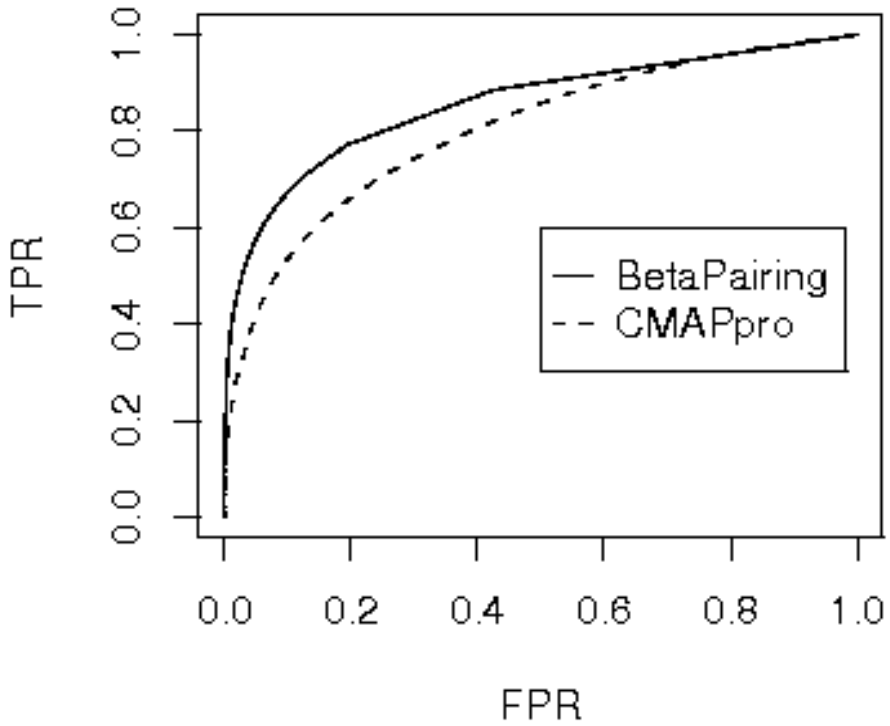
Predicted: 1-2, 2-4, 3-4, 4-5

Beta-Residue Pairing Results

Method	Specificity/ Sensitivity	Ratio of Improvement	ROC Area	TPR at 5% FPR
BetaPairing	41%	17.8	0.86	58%
CMAPpro (Pollastri and Baldi, 2002)	27%	11.7	0.80	42%

The accuracy of random algorithm is 2.3%.

ROC Plot



Strand Pairing Results

- Naïve algorithm of pairing all adjacent strands
  - Specificity = 42%
  - Sensitivity = 50%
  - All strand pairs are local strand pairs.

- MST like algorithm
  - Specificity = 53%
  - Sensitivity = 59%
  - >20% correctly predicted strand pairs are non-local strand pairs.

Strand Alignment Results

On the correctly predicted strand pairs

	Paring Direction	Alignment
Accuracy	93%	72%

On all native strand pairs

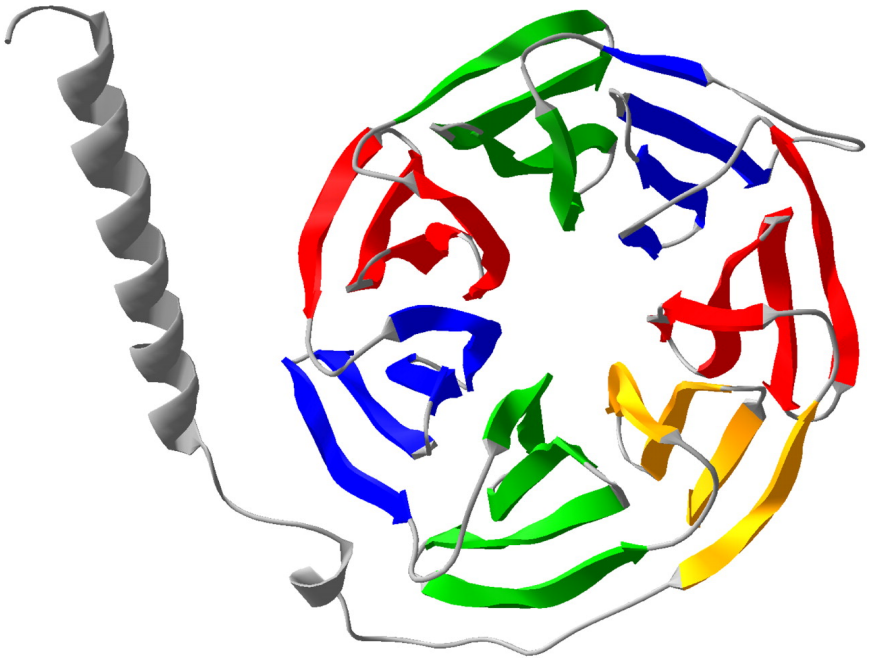
	Pairing Direction	Alignment
Accuracy	84%	66%

- The accuracy of pairing direction is 15% higher than that of the base-line algorithm.
- The alignment accuracy is significantly higher than previous methods.

Overview

- **Profcon**: Neural Network Approach
- **BetaPro**: Hierarchical Using Neural Networks, Alignments, and Graph Algorithms
- **SMURF**: Markov Random Field
- **tFolder**: Abstract template & ensemble prediction

Homology recognition of β-propeller



Seven-bladed β-propeller.

Before:

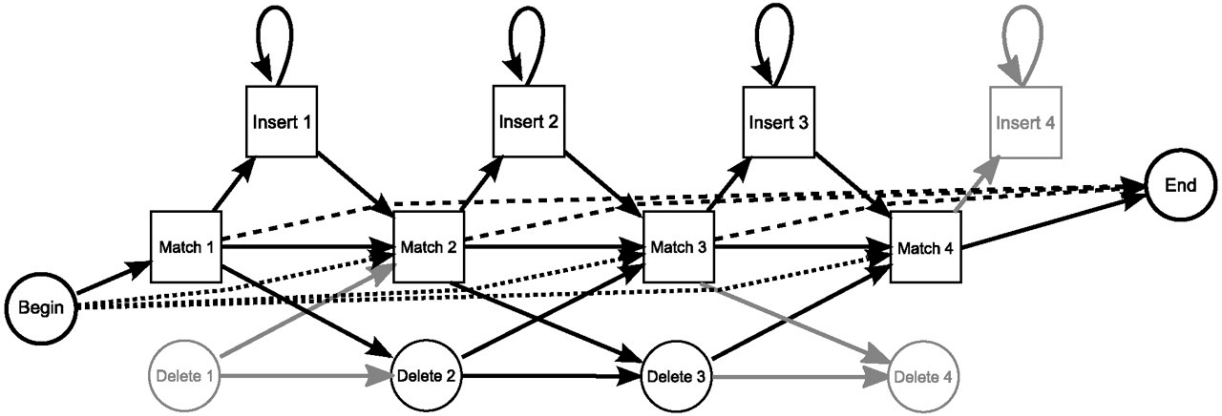
```
Seq1    ...VVDGD-ALLV--GFSEGSVN-YLYDG-GET-KLR--ING...
Seq2    ...VVDGDK--LLV-GFSEGSLQ-SMYDS-GETVKLR--ING...
Seq3    ...LD-GDLIA--FVS----RGQAFIQDSVGTYVL--KVL--...
Seq4    ...VI-GDL--IAFVS----RGY----DSVGTYVLKV--L--...
Seq5    ...VI-GDL--IAF-S----AGY--IQDSVGTY-LKV--L--...
                1  2345                   5432  1
```

After:

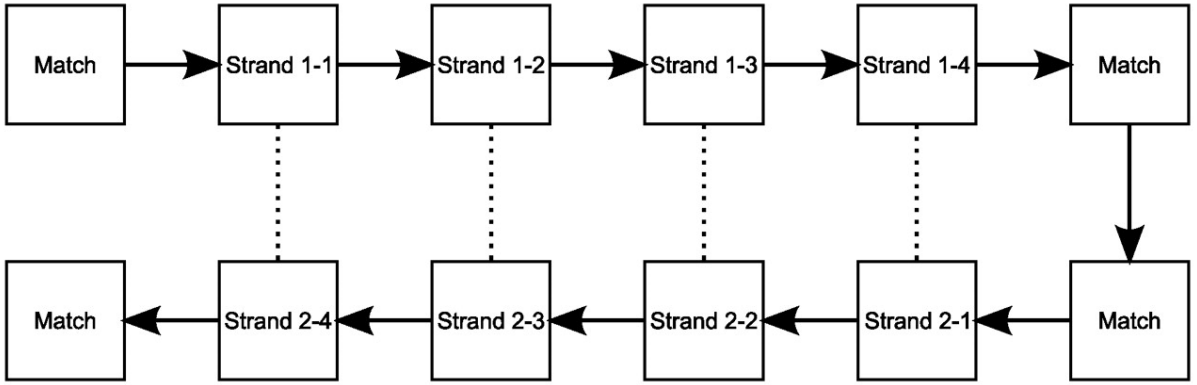
```
Seq1    ...VVDGD-  LV--GFSEGSVN-YLYDG-GET-KLR  ING...
Seq2    ...VVDGDK  LLV-GFSEGSLQ-SMYDS-GETVKLR  ING...
Seq3    ...LD-GDL  --FVS----RGQAFIQDSVGTYVL--  L--...
Seq4    ...VI-GDL  IAFVS----RGY----DSVGTYVLKV  L--...
Seq5    ...VI-GDL  IAF-S----AGY--IQDSVGTY-LKV  L--...
                1  234                    432  1
```

Modeling β -sheets with Markov Random Fields

HMM:



MRF:



(MRFs can model Long range dependencies of beta-strands)

Results

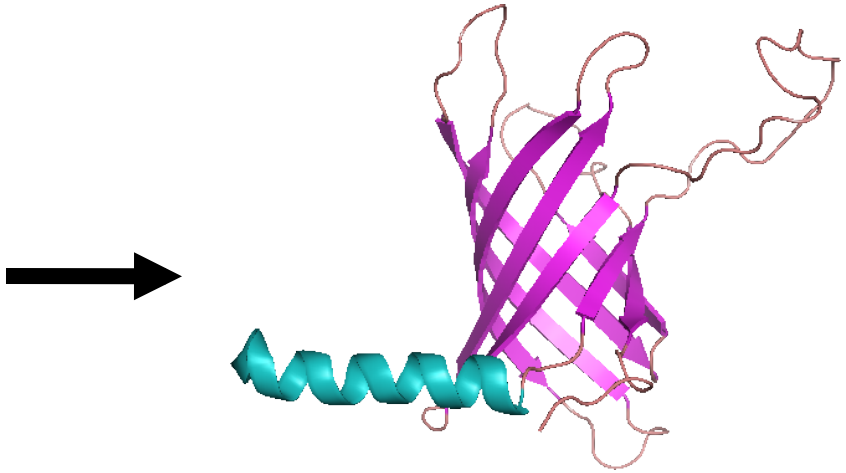
	Six-bladed			Seven-bladed			Eight-bladed		
TN	HMMER	SMURF(P)	SMURF	HMMER	SMURF(P)	SMURF	HMMER	SMURF(P)	SMURF
97%	52	20	**80**	80	23	**87**	0	**40**	0
96%	56	24	**80**	80	33	**87**	20	**40**	**40**
95%	64	36	**80**	87	47	**93**	20	**40**	**40**
94%	68	36	**84**	90	53	**93**	40	**60**	40
93%	68	48	**84**	90	53	**97**	40	**100**	40
92%	68	60	**88**	90	57	**97**	40	**100**	40
91%	68	60	**92**	90	57	**97**	40	**100**	40
90%	68	60	**92**	93	57	**100**	60	**100**	**100**

Overview

- **Profcon**: Neural Network Approach
- **BetaPro**: Hierarchical Using Neural Networks, Alignments, and Graph Algorithms
- SMURF: Markov Random Field
- **tFolder**: Abstract template & ensemble prediction

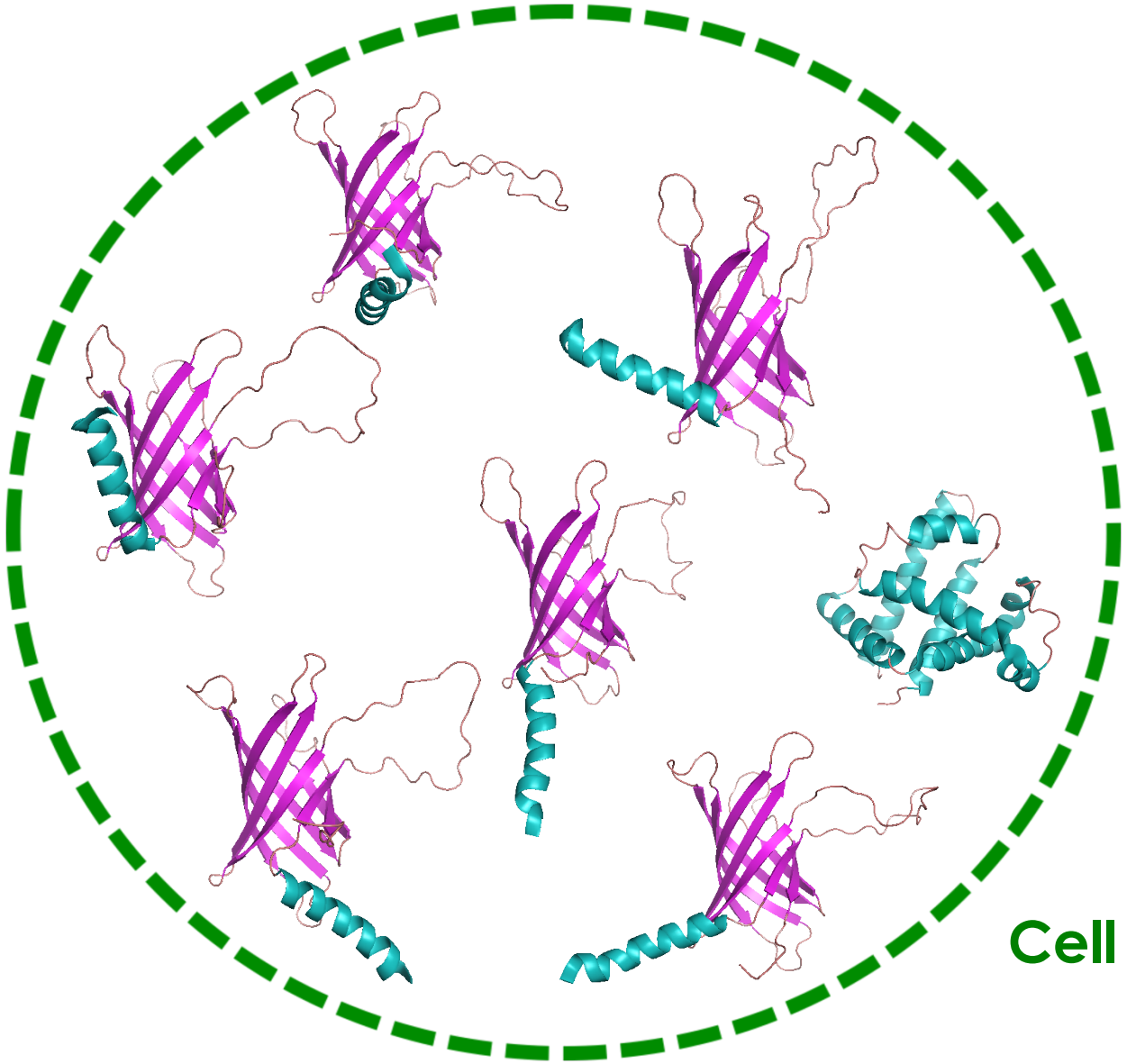
Classical View of Structural Biology

ATSTVTGGYAQSDAQGQMNK
MGGFNLKYRYEEDNSPLGVIGSF
TYTEKSRTASSGDYNKNQYYGITA
GPAYRINDWASIYGVVGVGYG
KFQTTEYPTYKNDTSDYGFSYGA
GLQFNPMENVALDFSYEQSRIRS
VDVGTWIAGVGYRF



Modern View

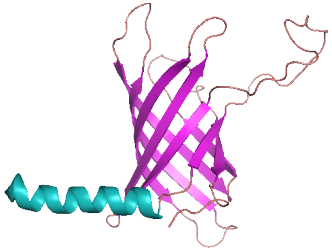
ATSTVTGGYAQSDAQGQM
NKMGGFNLKYRYEEDNSPL
GVIGSFTYTEKSRTASSGDYN
KNQYYGITAGPAYRINDWA
SIYGVVGVGYGKFQTTEYPT
YKNDTSDYGFSYGAGLQFN
PMENVALDFSYEQSRIRSVD
VGTWIAGVGYRF



Ensemble modeling

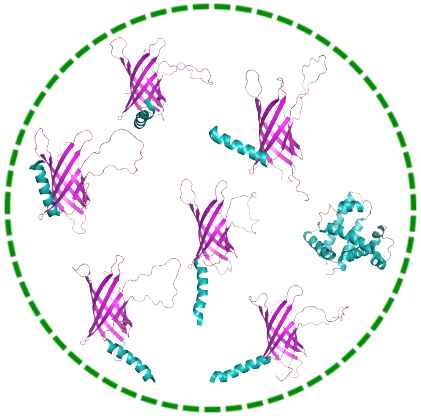
- Single structure prediction - Classical view.

ATSTVTGGYAQSDAQGQMNKMG
GFNLKYRYEEDNSPLGVIGSFTYTEKSR
TASSGDYNKNQYYGITAGPAYRIND
WASIYGVVGVGYGKFQTTEYPTYKN
DTSDYGFSYGAGLQFNPMENVALDF
SYEQSRIRSVDVGTWIAGVGYRF



- Ensemble prediction - Modern view.

ATSTVTGGYAQSDAQGQMNKMG
GFNLKYRYEEDNSPLGVIGSFTYTEKSR
TASSGDYNKNQYYGITAGPAYRIND
WASIYGVVGVGYGKFQTTEYPTYKN
DTSDYGFSYGAGLQFNPMENVALDF
SYEQSRIRSVDVGTWIAGVGYRF



Our approach: Compute a realistic ensemble representation of the structure landscape.

Outline

Ensemble approach enables:

1. Protein structure prediction.
   "Ensembles provide realistic structure prediction"
   (Waldispühl *et al.*, Proteins, 2008; O'Donnell *et al.*, ISMB 2011)

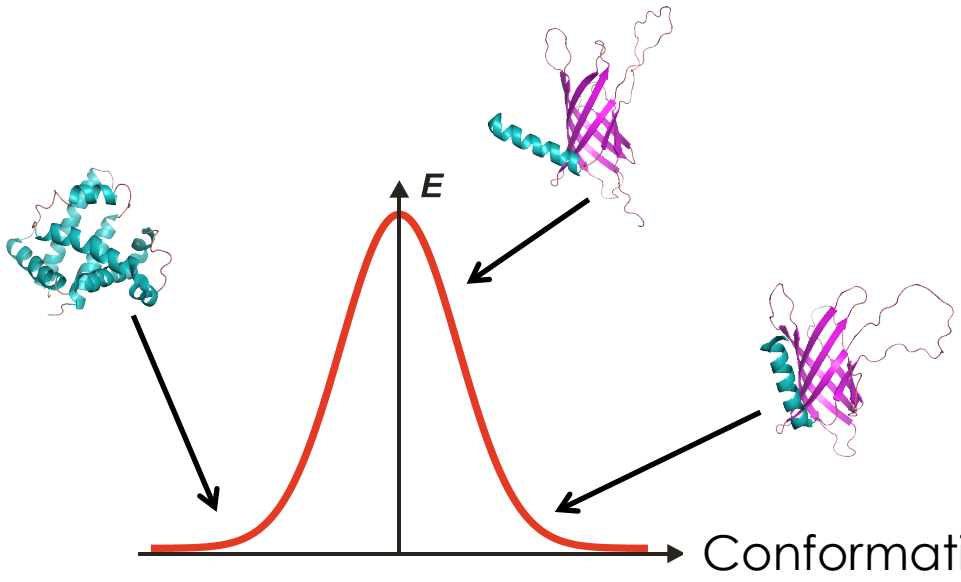
2. Folding pathway prediction.
   "Ensembles enable fast prediction of folding dynamics"
   (Shenker *et al.*, RECOMB, 2011)

Modeling ensembles of structures

Gibbs Measure:

$$P(S) \propto e^{\frac{-E(S)}{RT}}$$

Structure probability is proportional to the exponential of its energy.



This defines a *Boltzmann distribution* enabling:

- to compute statistics.
- to sample Structures.

Outline

Ensemble approach enables:

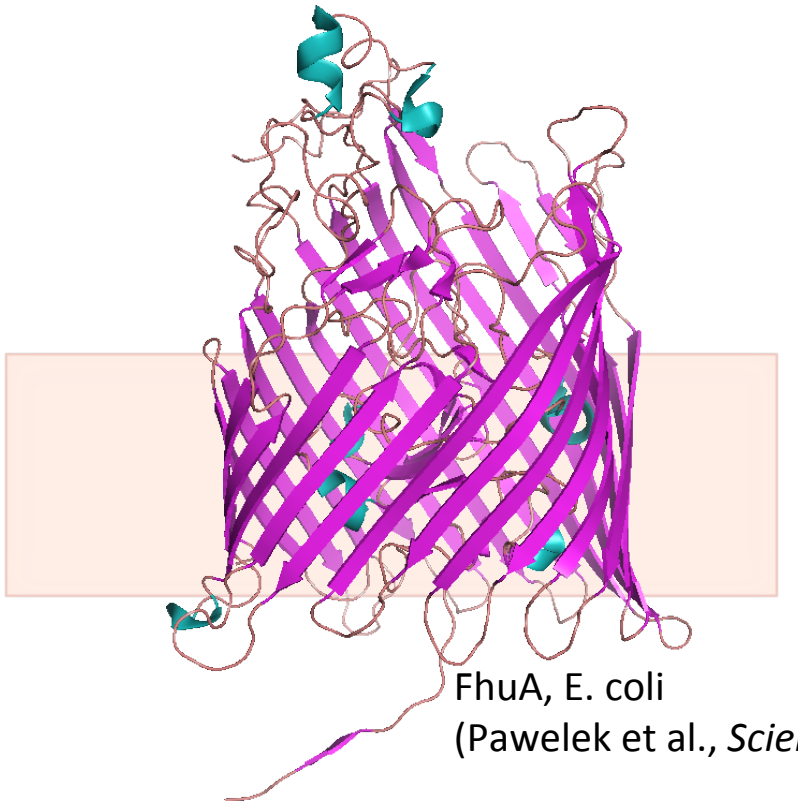
1. **Protein structure prediction.**
   "Ensembles provide realistic structure prediction"
   (Waldispühl *et al.*, Proteins, 2008; O'Donnell *et al.*, ISMB 2011)

2. Folding pathway prediction.
   "Ensembles enable fast prediction of folding dynamics"
   (Shenker *et al.*, RECOMB, 2011)

Transmembrane β -barrel proteins

- Found in outer-membranes.
- Wide variety of functions.

Ompx, E. Coli
(Vogt & Schulz,1999)

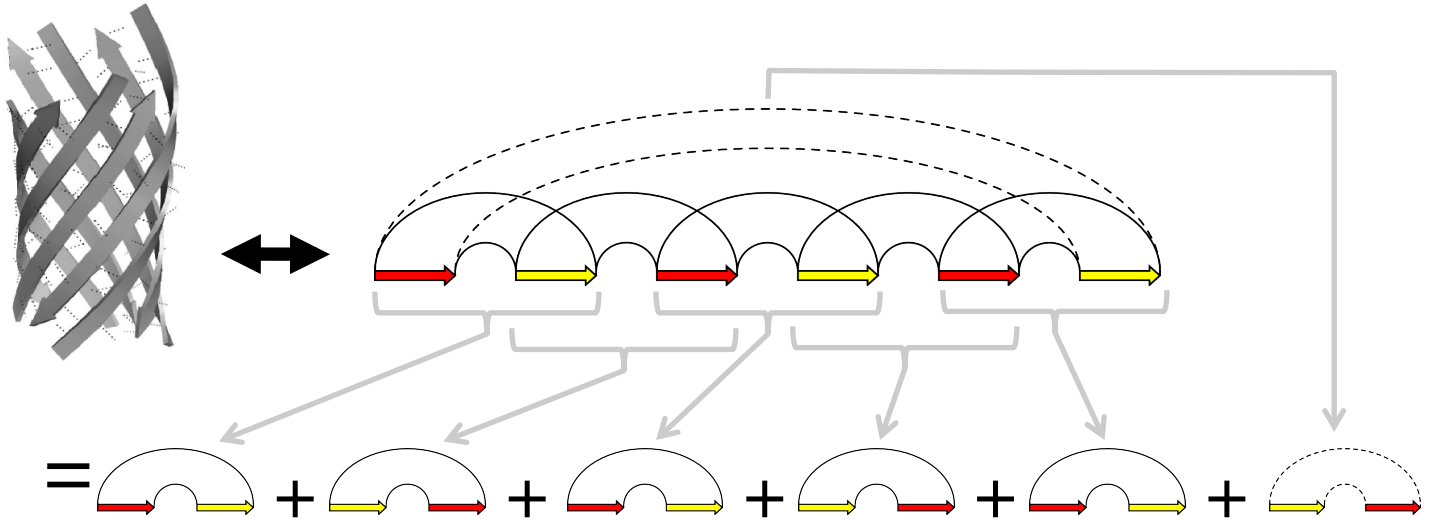


FhuA, E. coli
(Pawelek et al., *Science*, 2006)

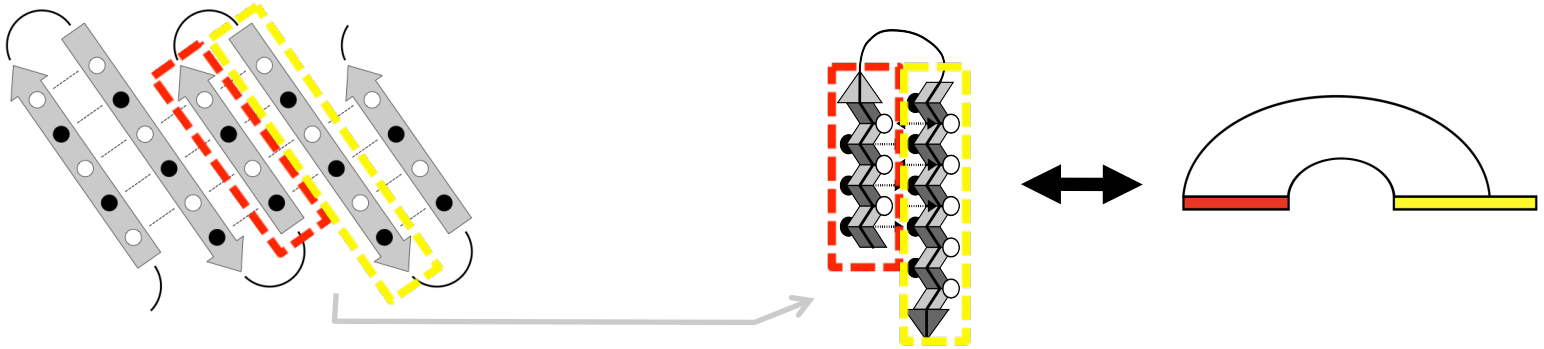
- Difficult to solve with X-Ray/NMR techniques,
  - Only few non-homologous structures in PDB.

Modeling β-barrels

1. The barrel is decomposed in a sum of β–strand pairs.



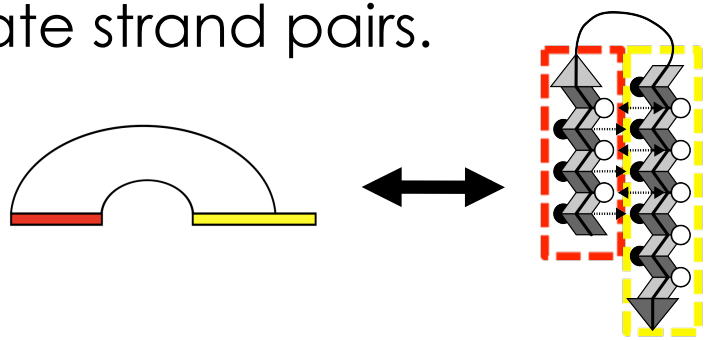
2. Inclination modeled using strand extensions.



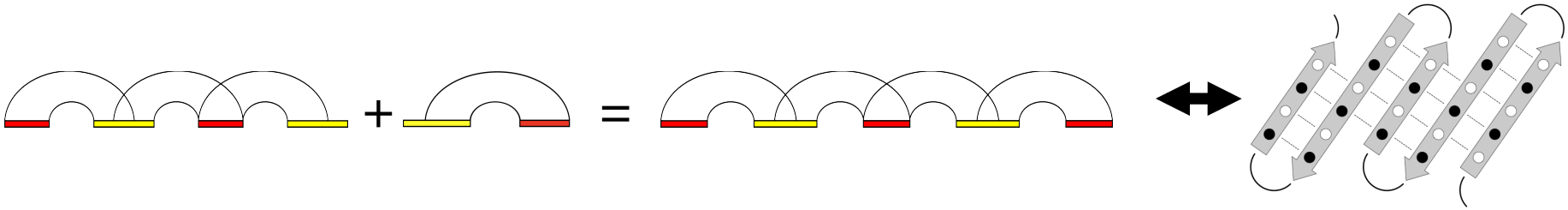
(Waldispuhl et al., *Proteins*, 2006)

Exploring the TMB folding landscape

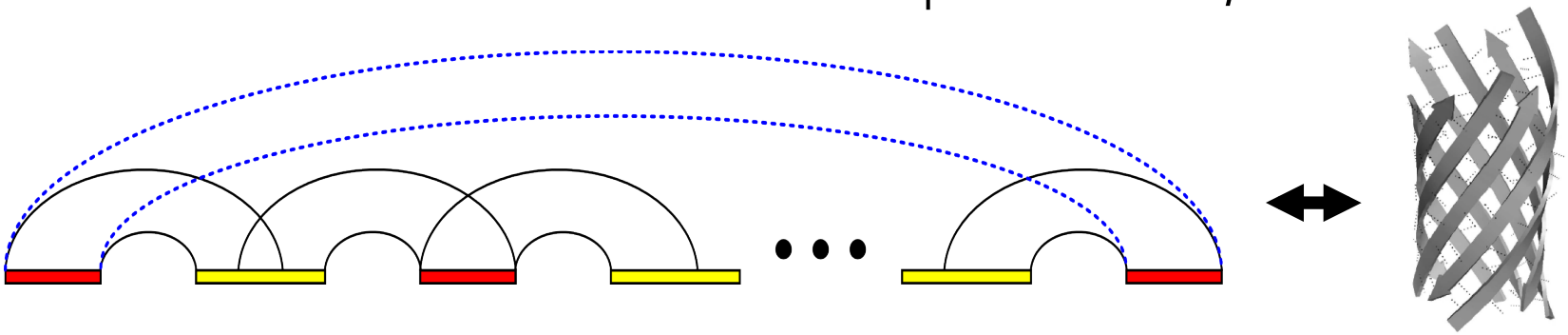
1. **Initialization:** create strand pairs.



2. **Chaining:** concatenate strand pairs to build β-sheets.



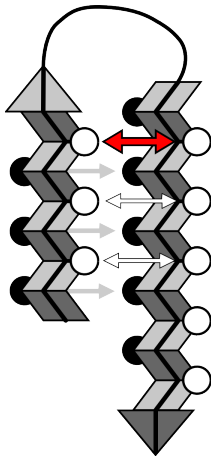
3. **Closure:** Pair first and last strand pair of the β-sheet.



(Waldispuhl et al., *Proteins*, 2006)

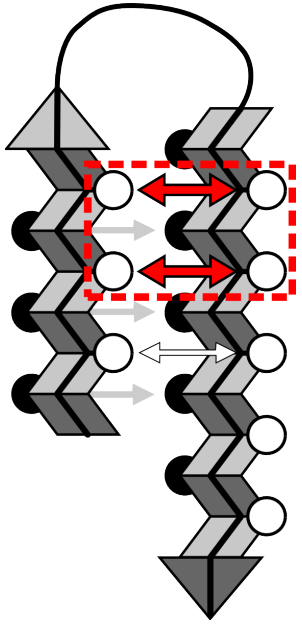
Energy model

- Classical approach: Residue contacts.



Energy of the structure is the sum of the energies of all contacts.

- Our new Concept: **Stacking pairs**.



$$E(i,j,x \mid i+2, j+2) = -RT\log\left(p_{i,j,x \mid i+2,j+2}\right) - RT\log\left(Q_{tmb}\right)$$

- Computed from globular proteins
- Distinguish environment

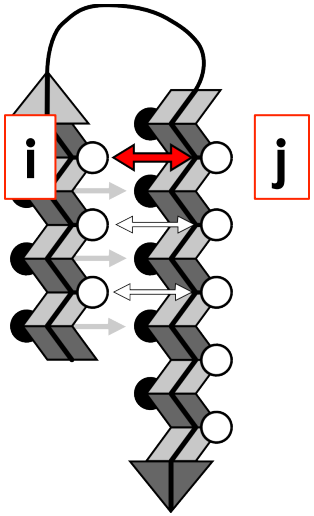
(Waldispuhl et al., *Proteins*, 2008)

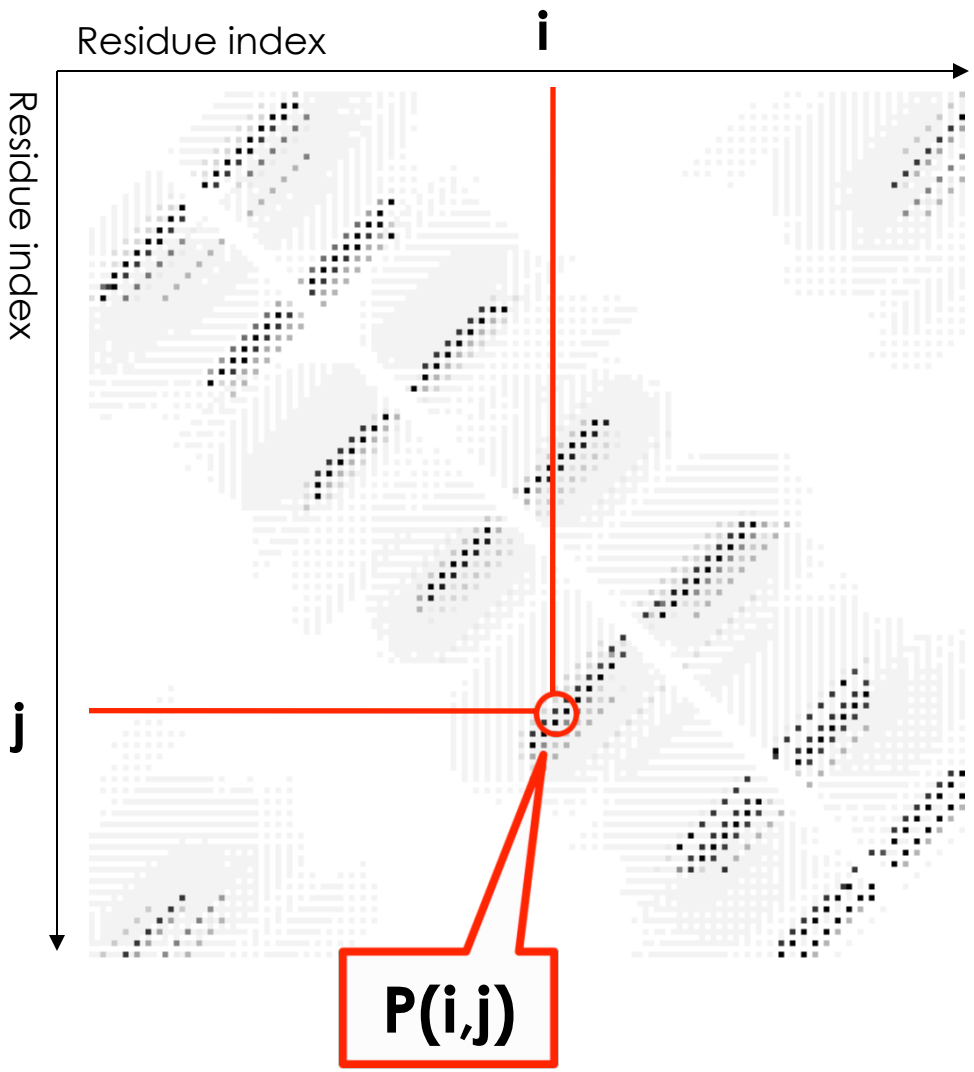
Stochastic Contact Maps

Contact probability:

$$p(i,j) \propto \sum_{(i,j)\in S} e^{\frac{-E(S)}{RT}}$$

Backtrack dynamic tables $O(n^3)$.

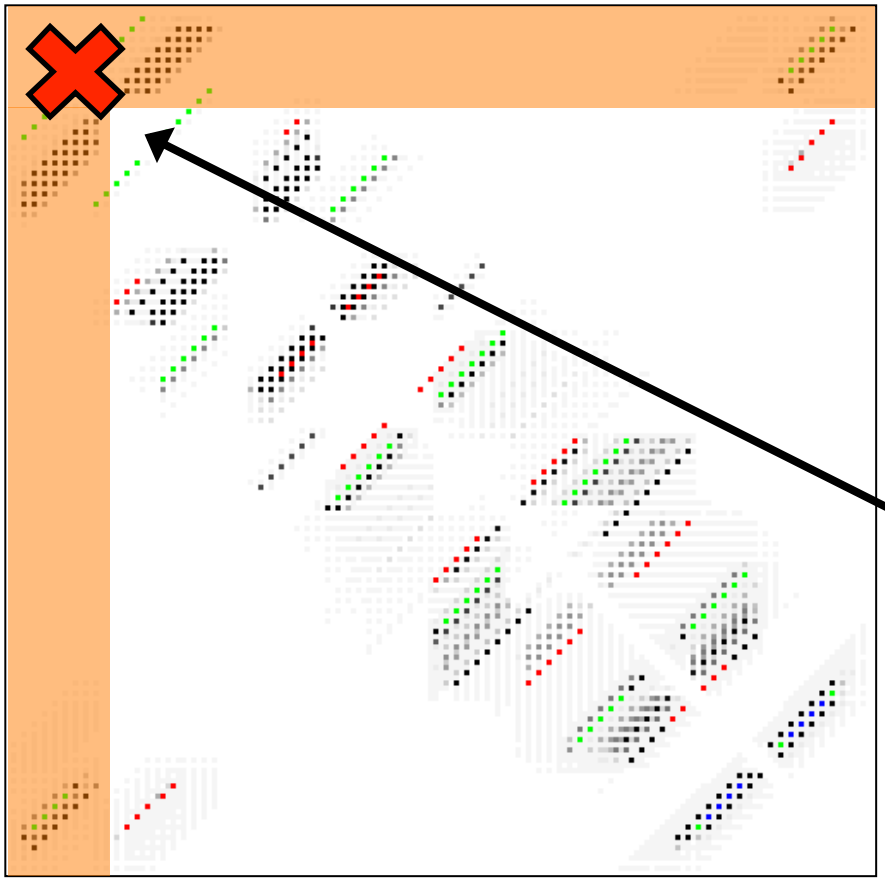


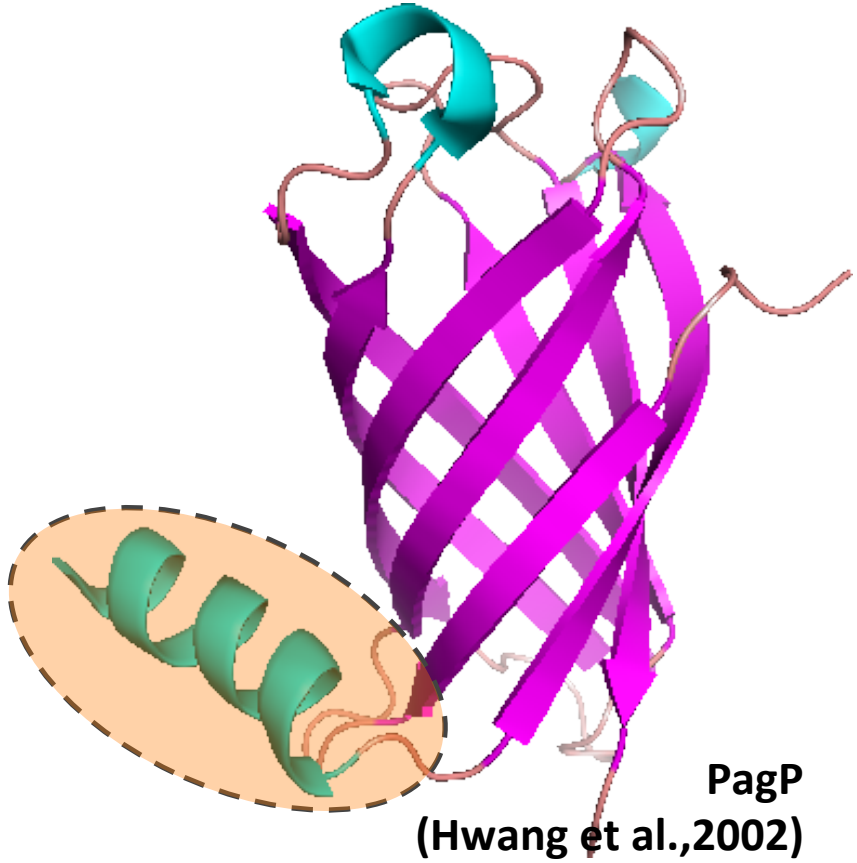


Upper triangle: Membrane
Lower triangle: Channel

(Waldispuhl et al., *Proteins*, 2008)

Stochastic Contact Maps





PagP
(Hwang et al.,2002)

Red: Crystal structure
Green: Single structure prediction

Initial α-helix stabilizes PagP
(Huysmans et al,2007)

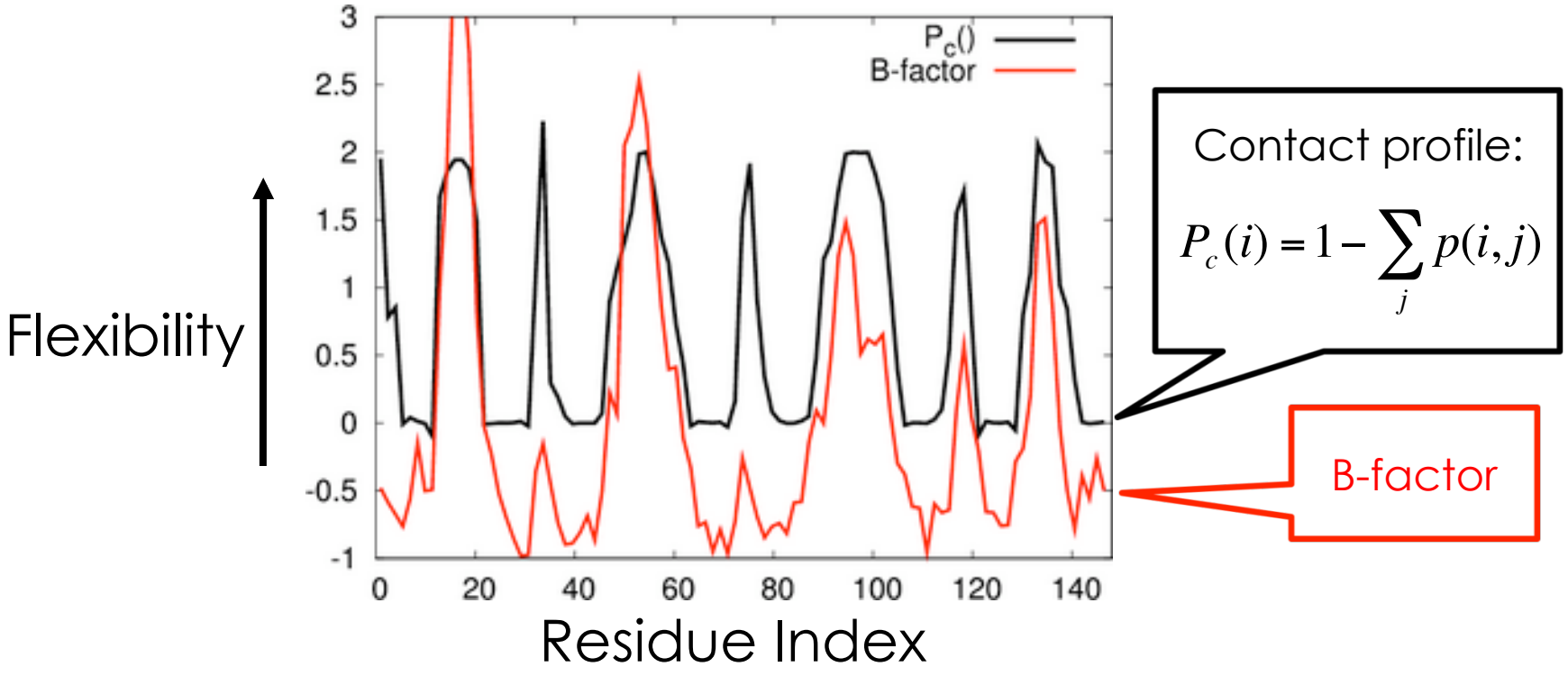
Comprehensive representation of ensemble of folds

(Waldispühl et al., *Proteins*, 2008)

Contact Profiles

B-factor: experimental measure of the flexibility of residues.

Per residue contact probability correlates with B-factor



Direct Prediction of experimental measures.

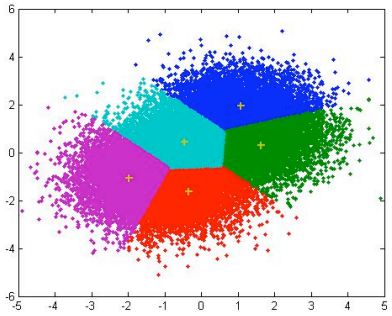
(Waldispühl et al., *Proteins*, 2008)

3D Structure Construction Pipe-line

Sample contacts (partiFold)

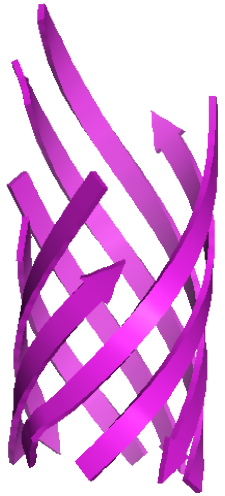
1	A	.	0	0
2	T	.	0	0
3	S	.	0	0
4	T	.	0	0
5	V	M	30	148
6	T	C	29	147
7	G	M	28	146
8	G	C	27	145
9	Y	M	26	144
10	A	C	25	143
11	Q	M	24	142
12	S	C	23	141
13	D	.	0	0
14	A	.	0	0

Cluster (k-means)



NEW

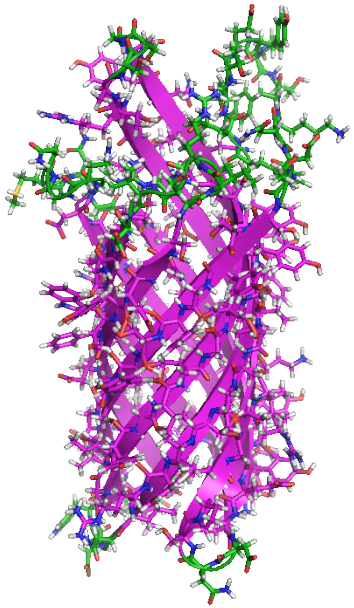
Build backbone (partiFold2pdb)



NEW

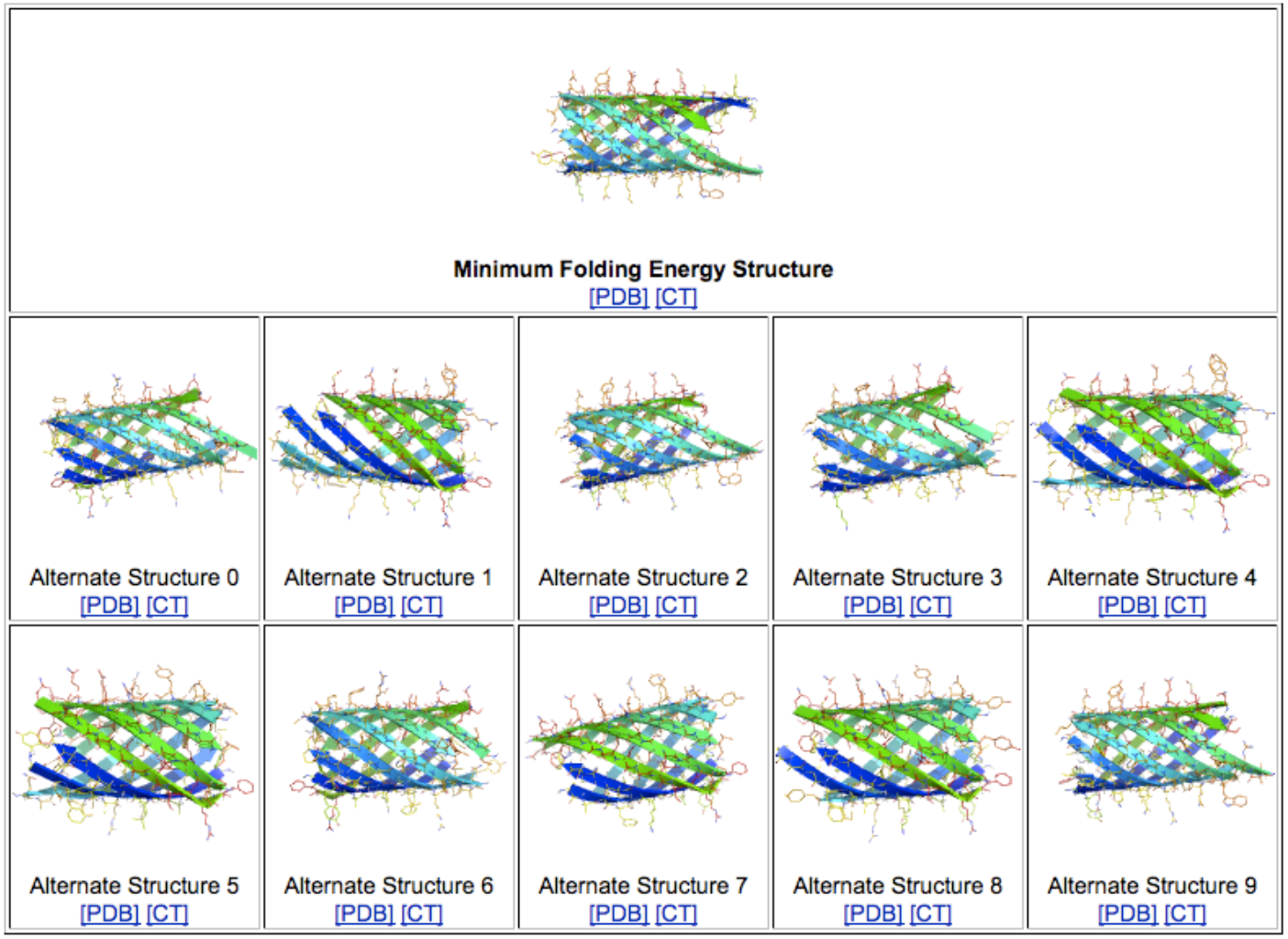
Add loops (Spanner)

Add side-chains (TreePack,SCWRL)



(Waldispuhl et al., *in preparation*)

http://partiFold.csail.mit.edu



(Waldispuhl et al., *in preparation*; Waldispuhl et al., *NAR, 2006*)

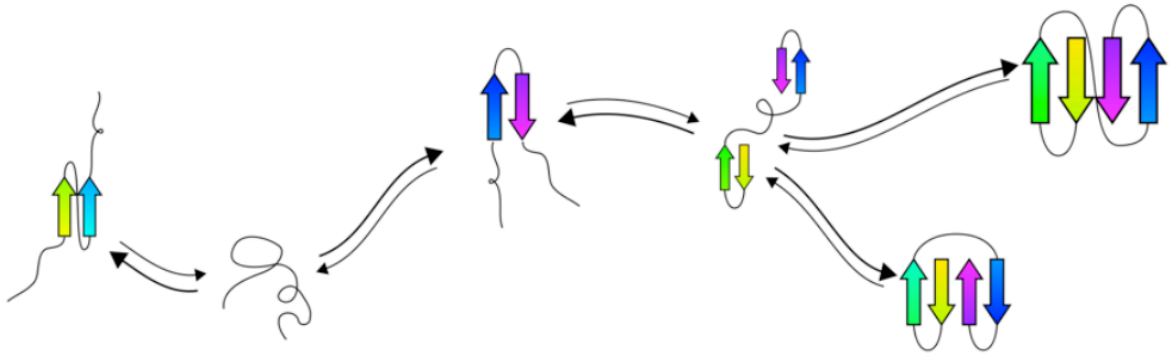
Outline

Ensemble approach enables:

1. Protein structure prediction.
   "Ensembles provide realistic structure prediction"
   (Waldispühl *et al.*, Proteins, 2008; O'Donnell *et al.*, ISMB 2011)

2. **Folding pathway prediction.**
   "Ensembles enable fast prediction of folding dynamics"
   (Shenker *et al.*, RECOMB, 2011)

Objective



Predicting coarse grained β-sheet protein folding pathways in a couple of minutes on your laptop.

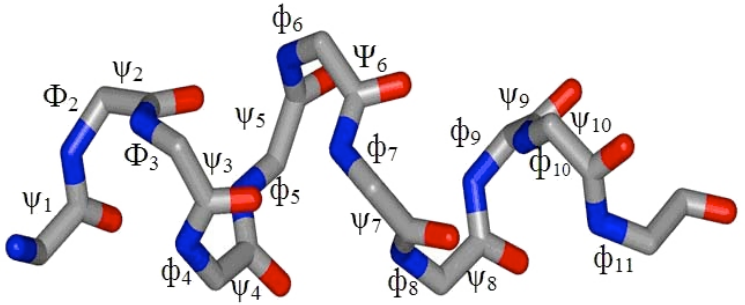
Vs.

(Shenker *et al.*, *RECOMB, 2011*)

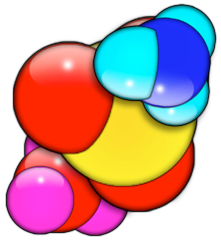
Previous works (1)

Molecular dynamics (GROMACS, AMBER, CHARMM):

+ Accurate
+ High-resolution structure
+ Use classical physics laws
- Slow
- Fragile
- Limited to small polypeptides and short folding time



Distributed computing (Folding@home):



+ More powerful
+ High-resolution
- Not flexible
- Remains limited to small sequences (40 residues)

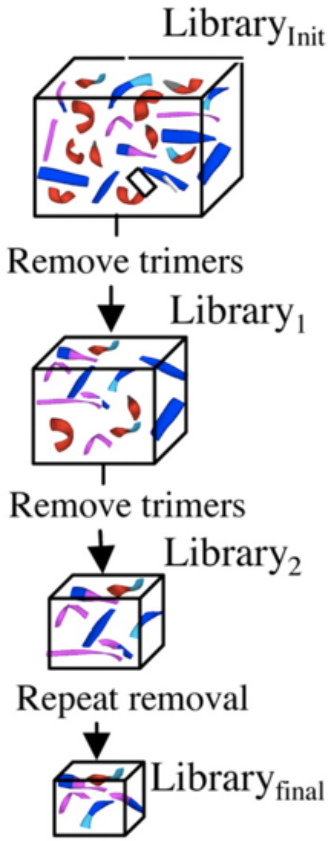
Previous works (2)

Motion planning and unfolding pathways (Amato *et al.*):

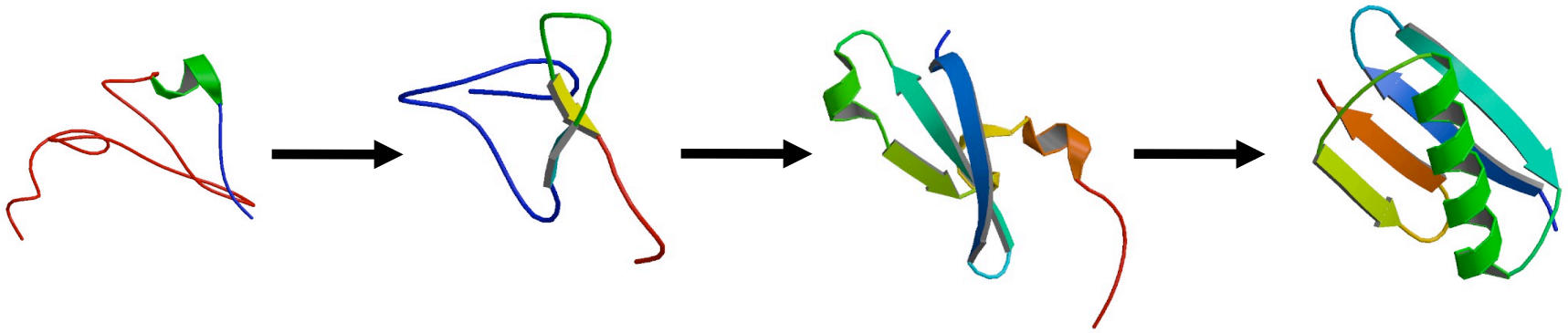
+ Fast
+ High resolution
- Require the knowledge of the native 3D structure
- Cannot be used to study misfolding

Iterative fixing (DeBartolo *et al., 2009*)

+ Fast
+ Accurate
+ High resolution
- Do not explore the full conformational landscape



Classical approach: Folding Simulation



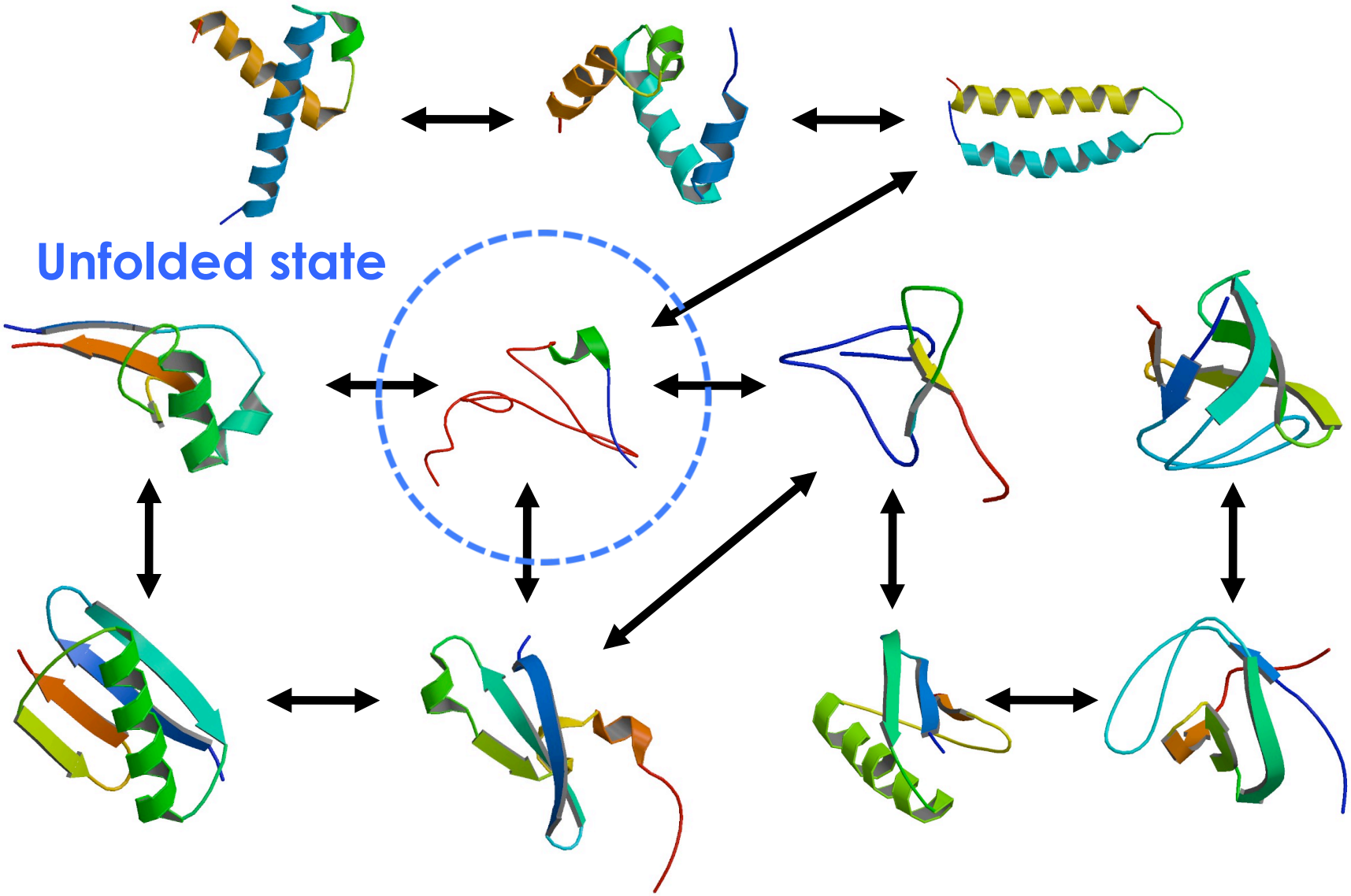
t_0 $t_1=t_0+\Delta t$ $t_2=t_1+\Delta t$ $t_3=t_2+\Delta t$

Principle: Calculate a new structure from current structure

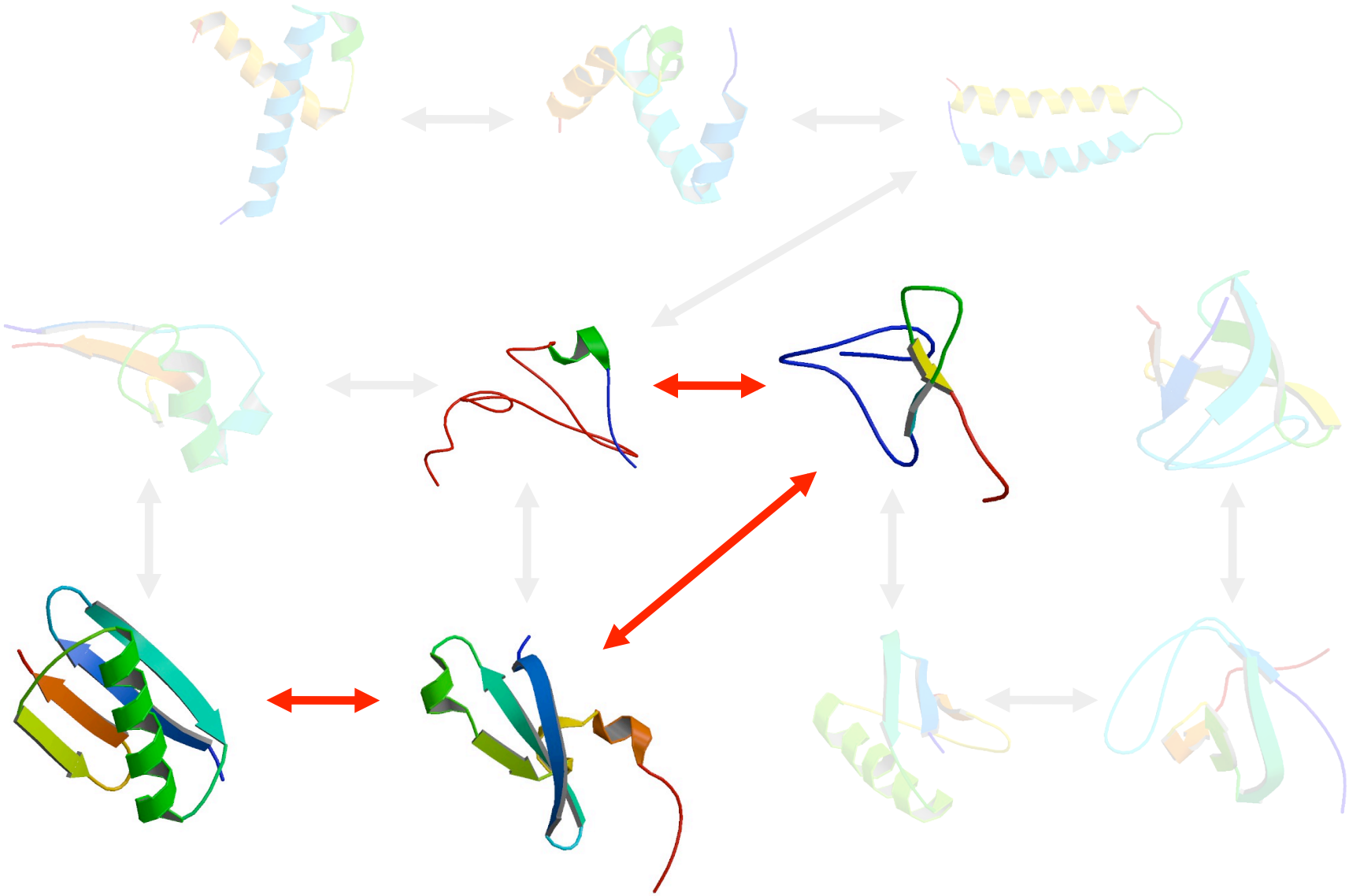
Drawbacks:

- No global view.
- Require to simulate multiple trajectories (often similar).
- Difficulties to incorporate long-range interactions and thus β-sheets.

Our Approach: Compute Energy Landscape

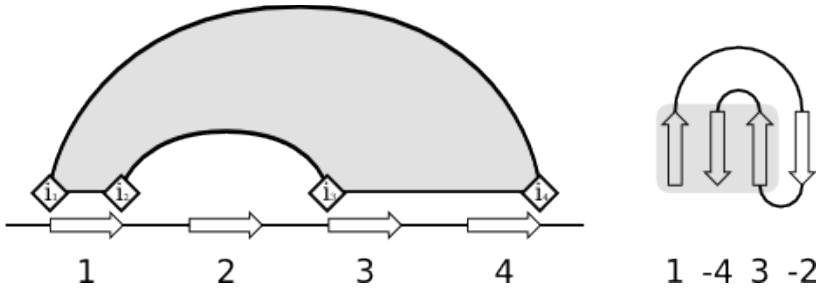


Our Approach: Predict Folding Pathway

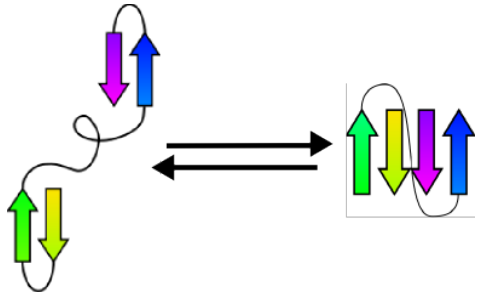


Our approach

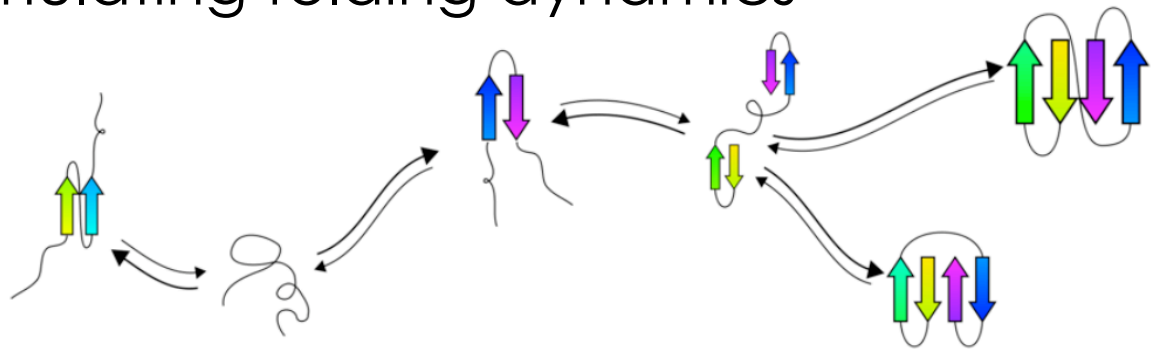
1. Computing an approximate energy landscape



2. Estimating transition rates between structures



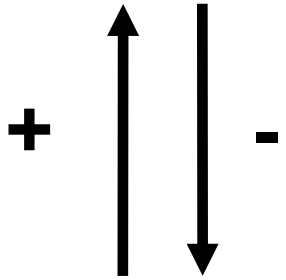
3. Simulating folding dynamics

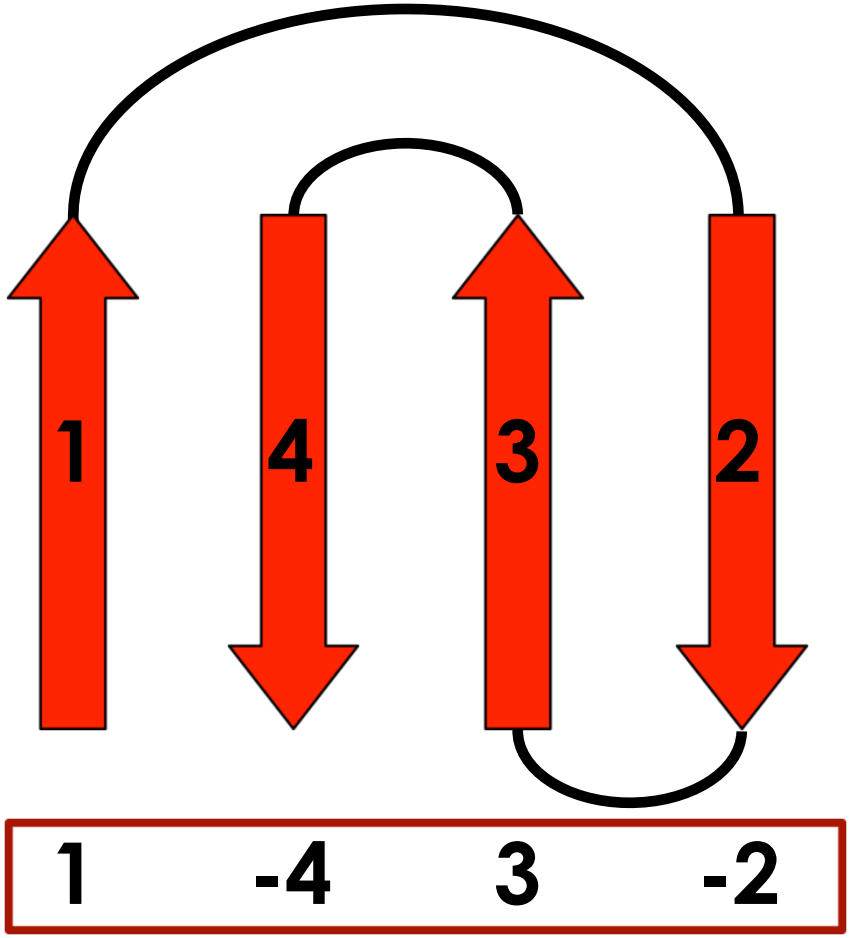


(Shenker *et al.*, *RECOMB, 2011*)

Modeling β -sheets

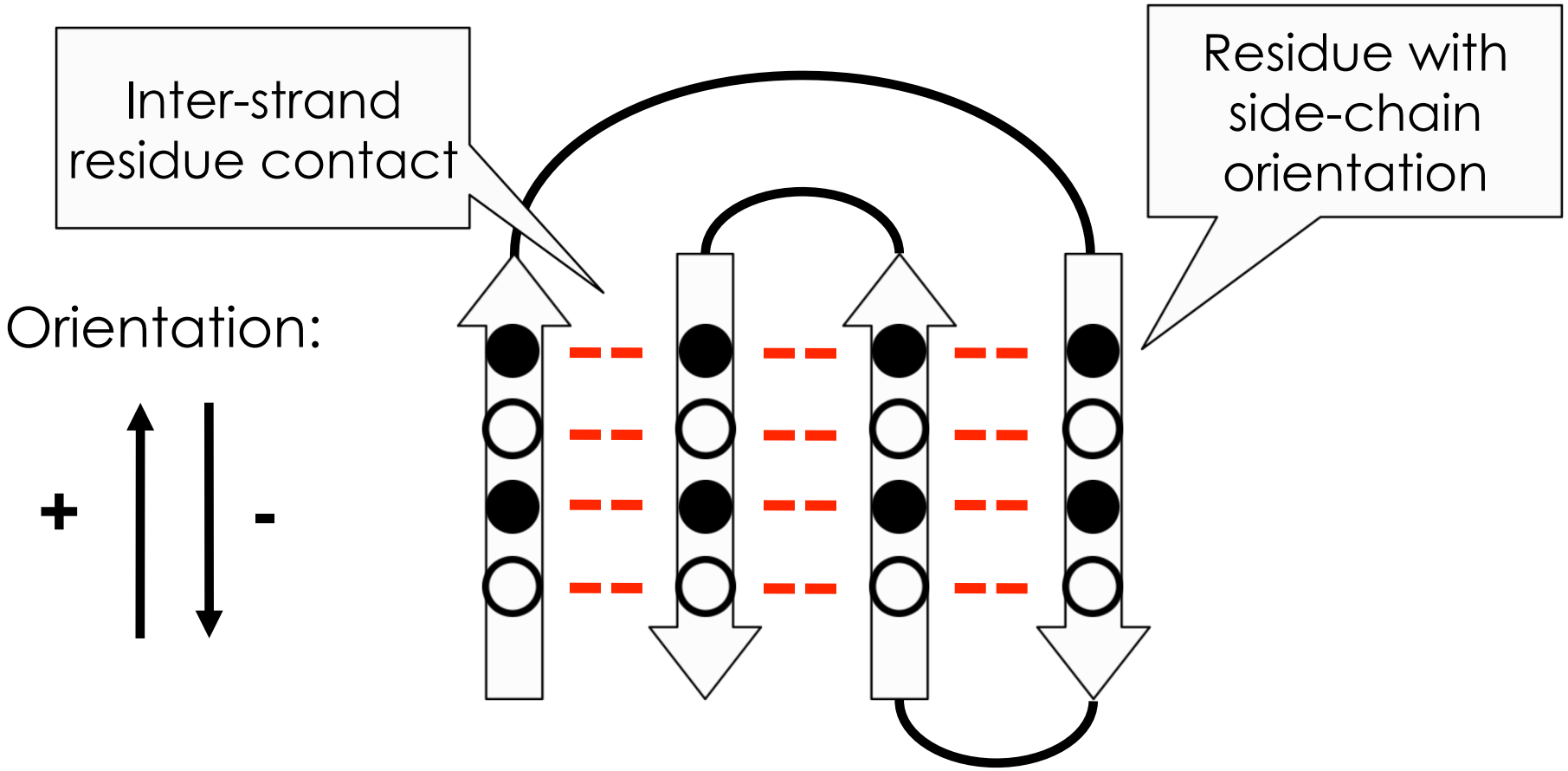
Orientation:





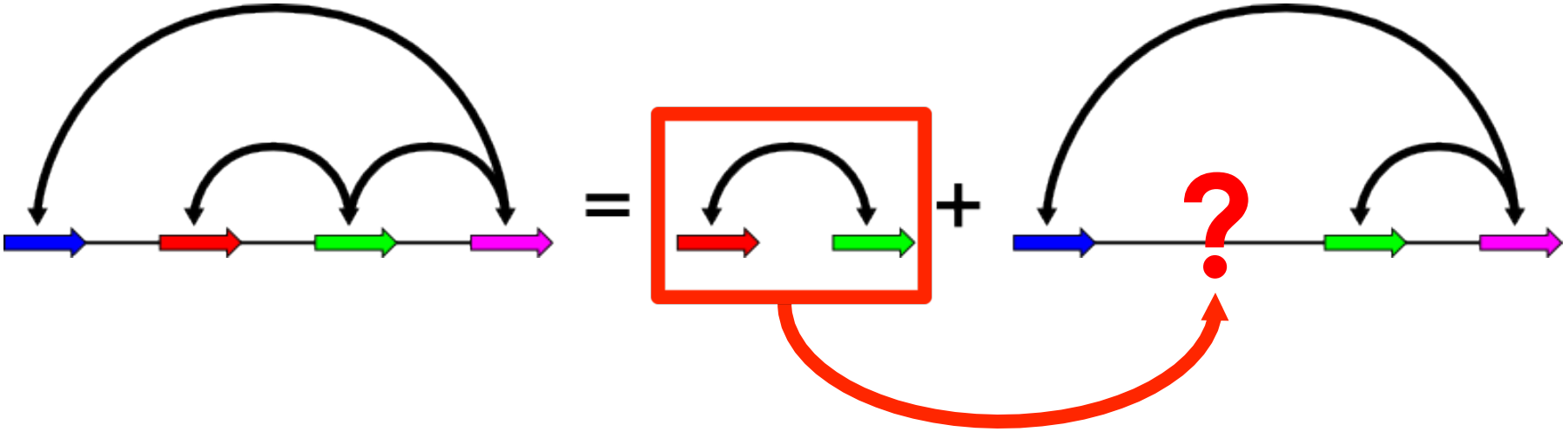
β -sheet topologies modeled as signed permutations

Modeling β -sheets

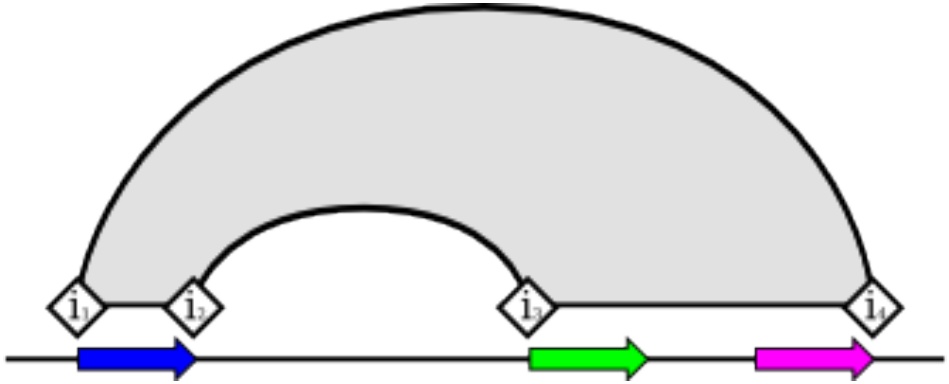


β -sheet model explicitly incorporates inter-strand residue interaction with side-chain orientation.

Recursive enumeration of β -sheets



Expand dynamic tables to allow strand insertion:



Signed permutations define the order of strand insertions.

Contact prediction benchmark

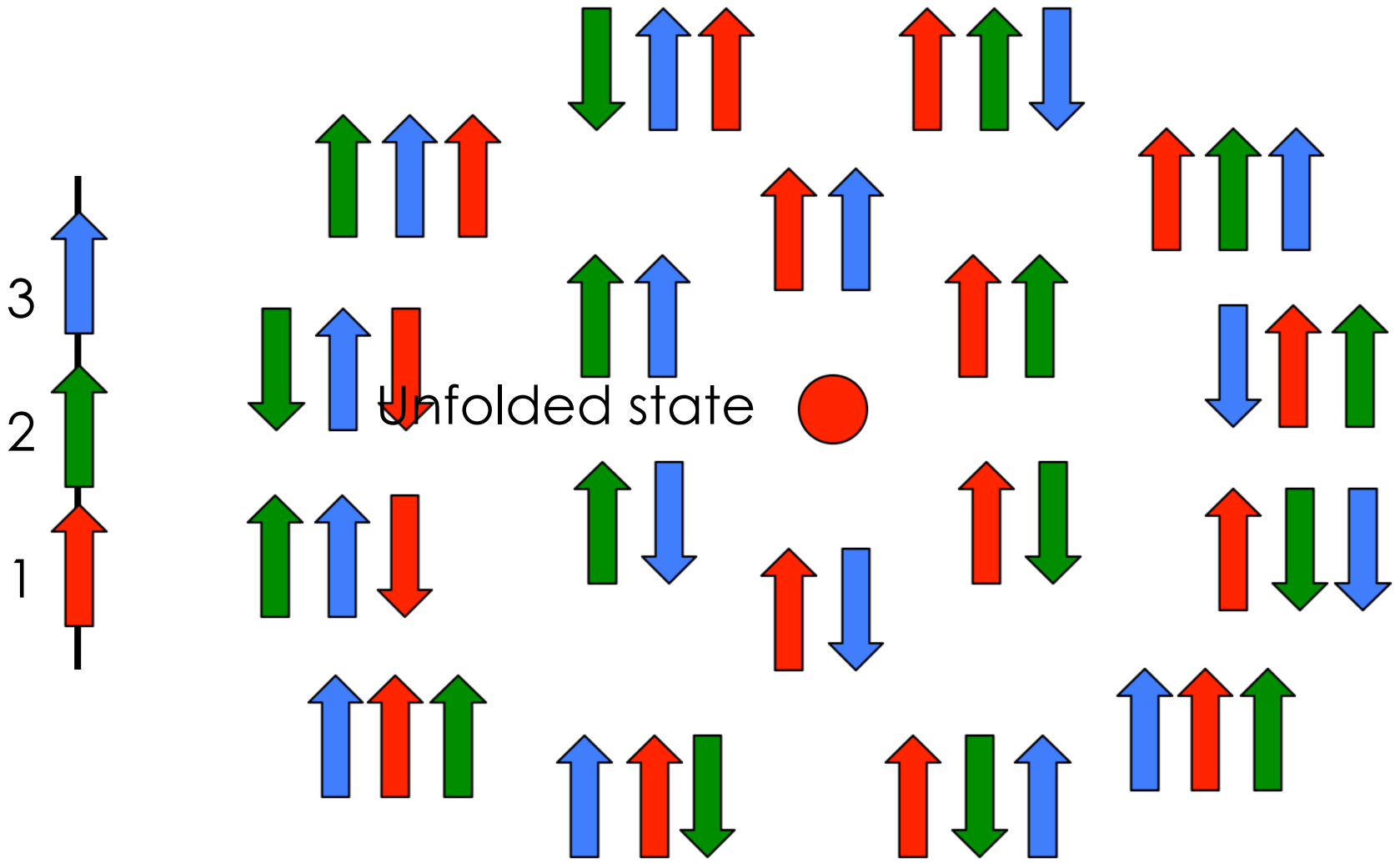
Benchmark 12 non-homologous protein data set

Separation	>= 12		>= 24	
Method	Accuracy	Coverage	Accuracy	Coverage
tFolder	0.27	0.27	0.23	0.28
BETApro	0.22	0.40	0.05	0.14
SVMcon	0.32	0.31	0.24	0.21

Accuracy=correct/total predicted Coverage=correct/total native

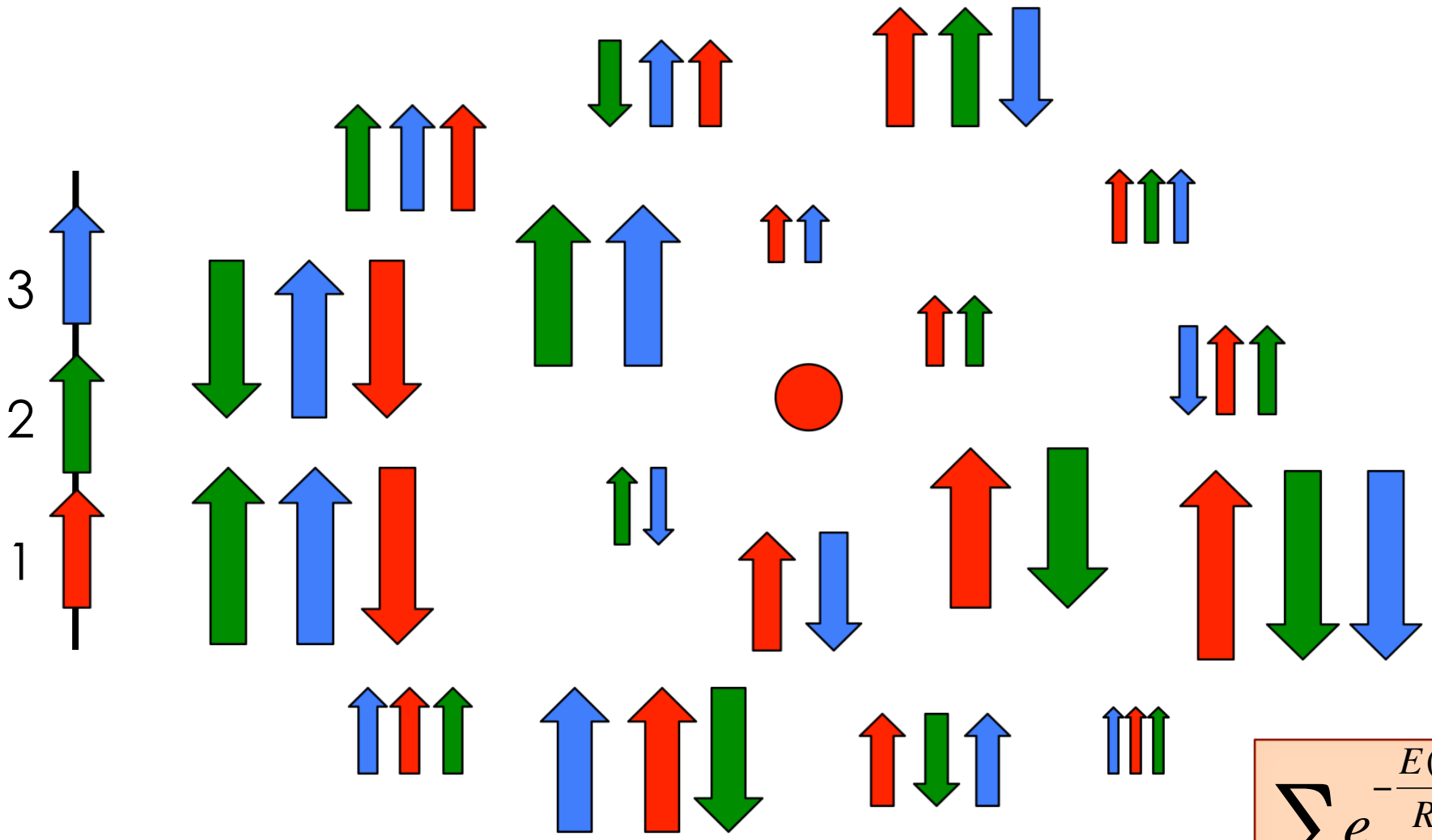
Performance not affected by residue sequence separation

β -sheet energy landscape



Enumerate all topologies (i.e. permutations).

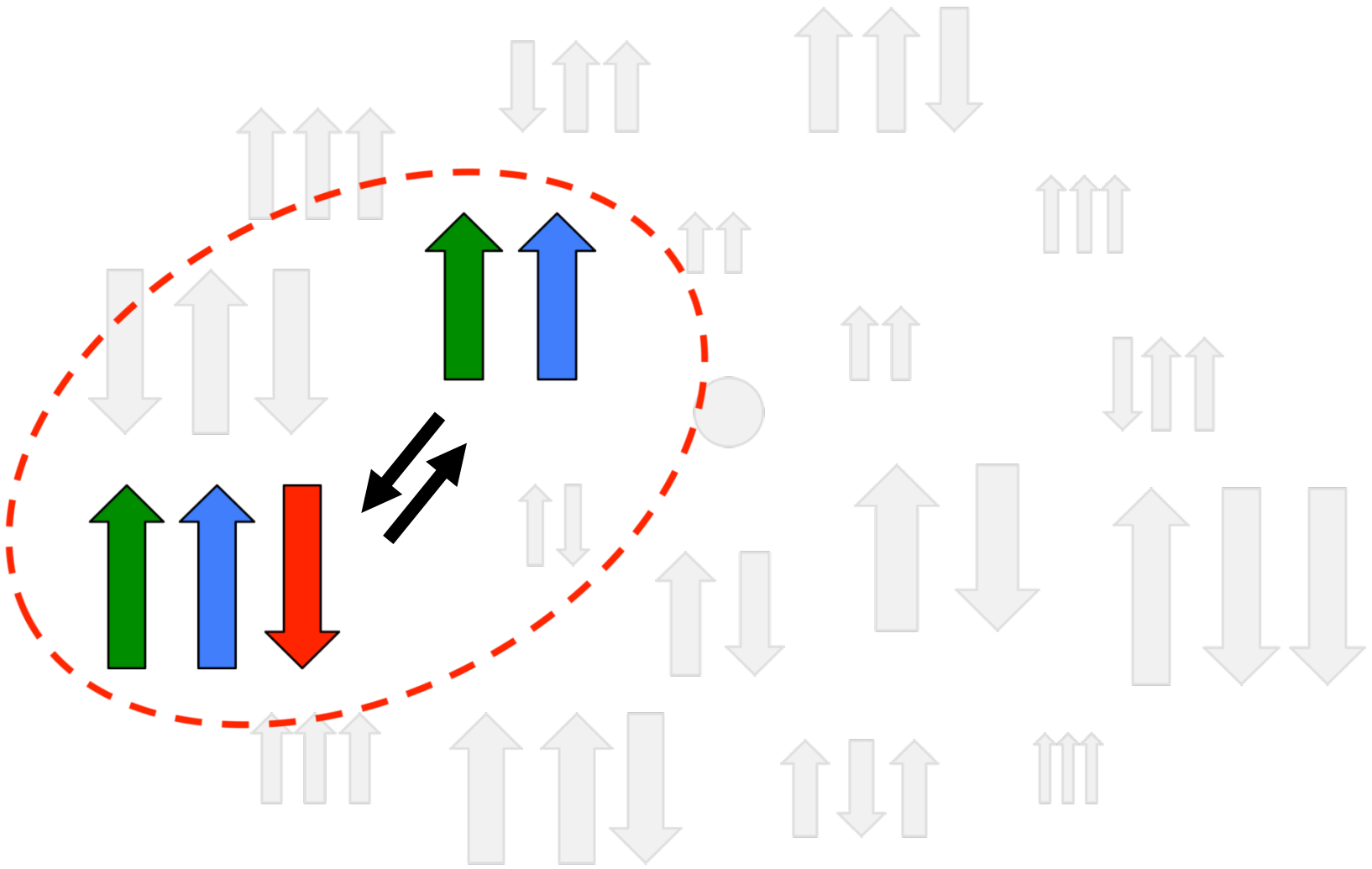
β -sheet energy landscape



Weight topologies using the energy ensemble:

$$\sum_{S \in T} e^{-\frac{E(S)}{RT}}$$

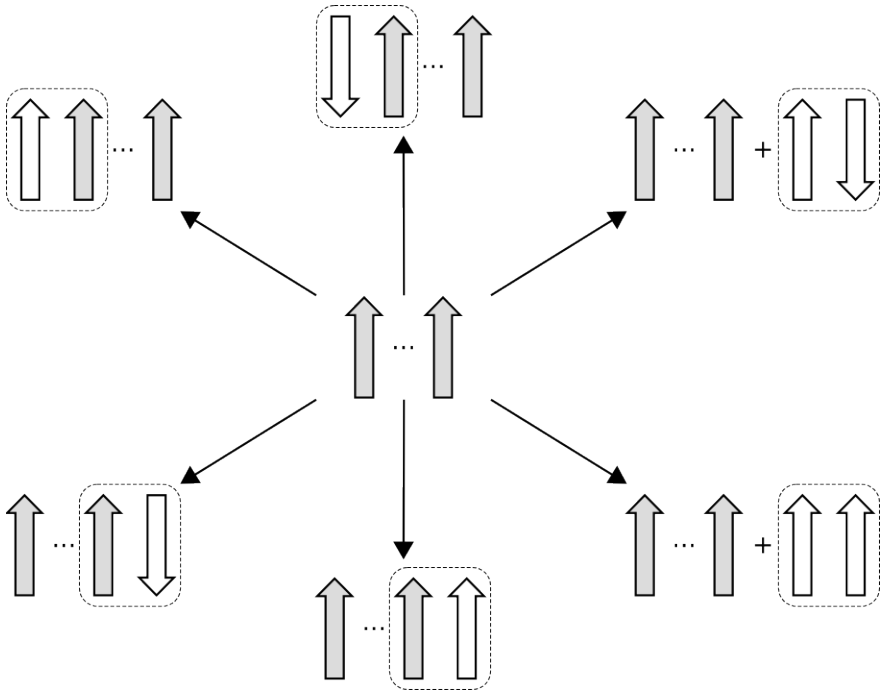
Connecting the structural states



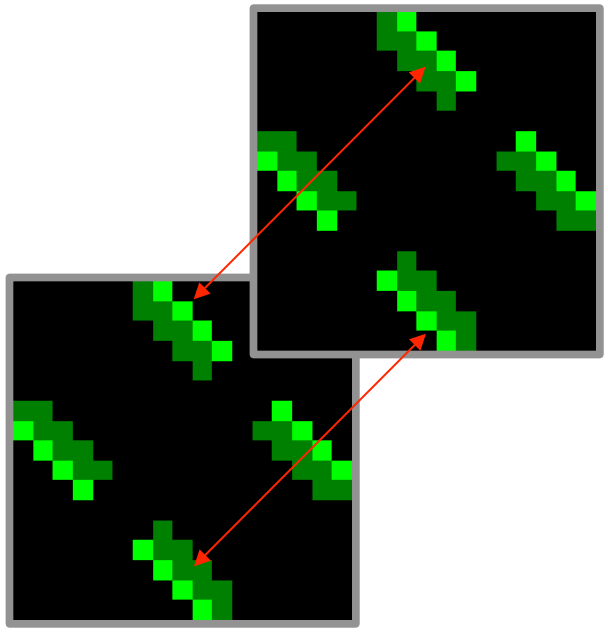
How to **estimate** the transition rates between **compatible** β -sheets?

Enabling state transitions

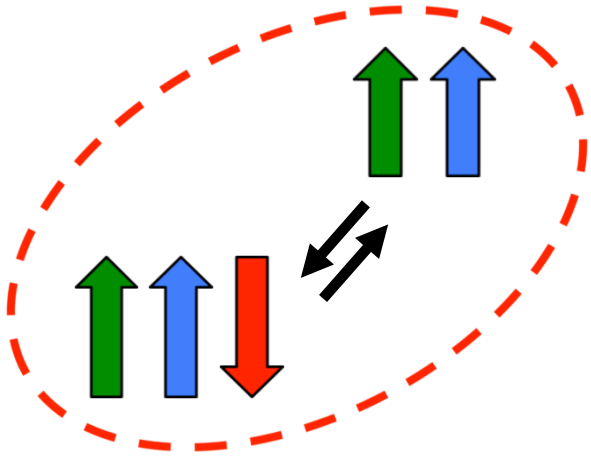
1. Must have compatible topology



2. Must share contacts



Estimating transition rates



$$E(\chi_i) = \sum_{x \in \chi_i} E(x)$$: **Ensemble energy**

$$\Delta G_{ij} = E(\chi_i) - E(\chi_j)$$

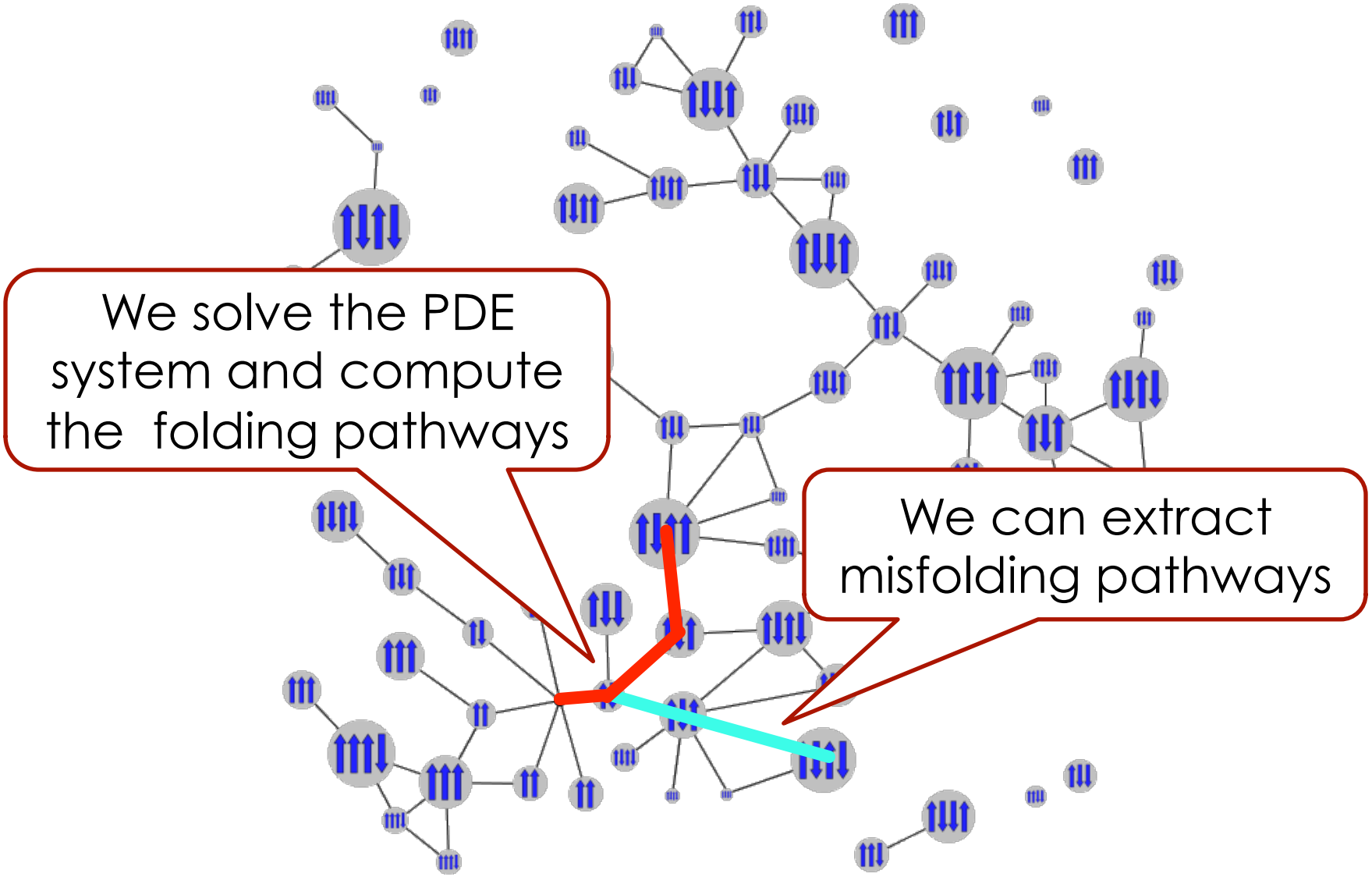
$$E(\chi_j) = \sum_{x \in \chi_j} E(x)$$

Kawasaki's rule: $r_{ij} = r_0 e^{-\frac{\Delta G_{ij}}{2RT}}$

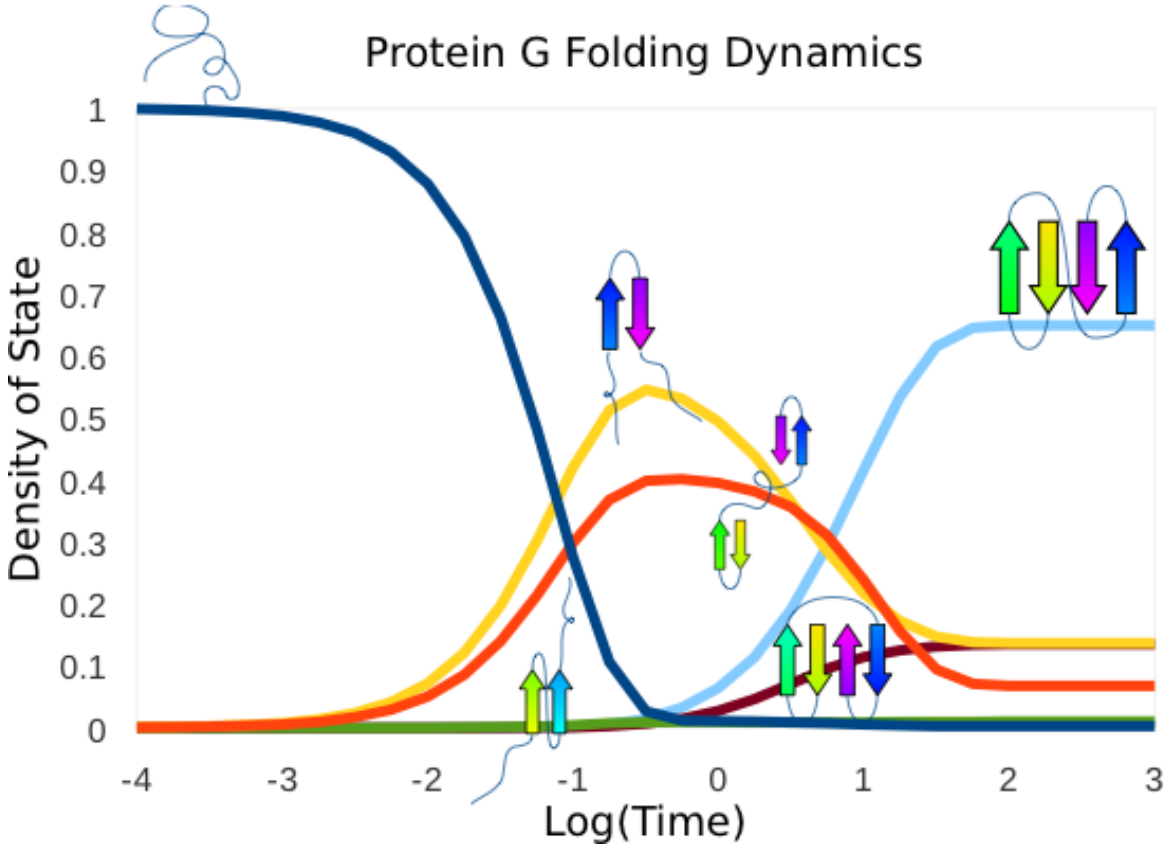
Transition rates are defined with differential equations:

$$\frac{dp_i}{dt} = \sum_{j \in X} r_{ij} p_j(t) \quad \text{with} \quad r_{xx} = -\sum_{y \neq x} r_{yx}$$

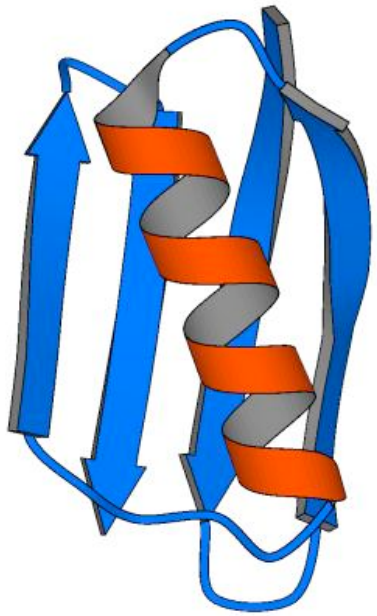
Coarse grained Energy Landscape



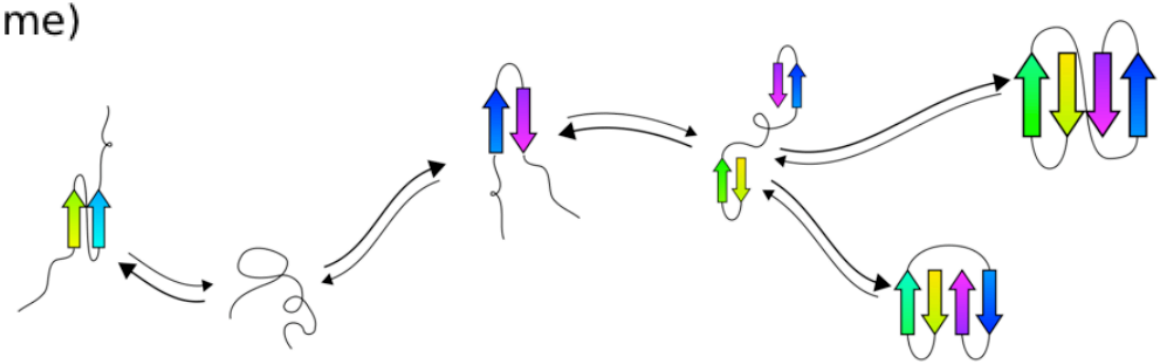
Folding Simulation of Protein GB1



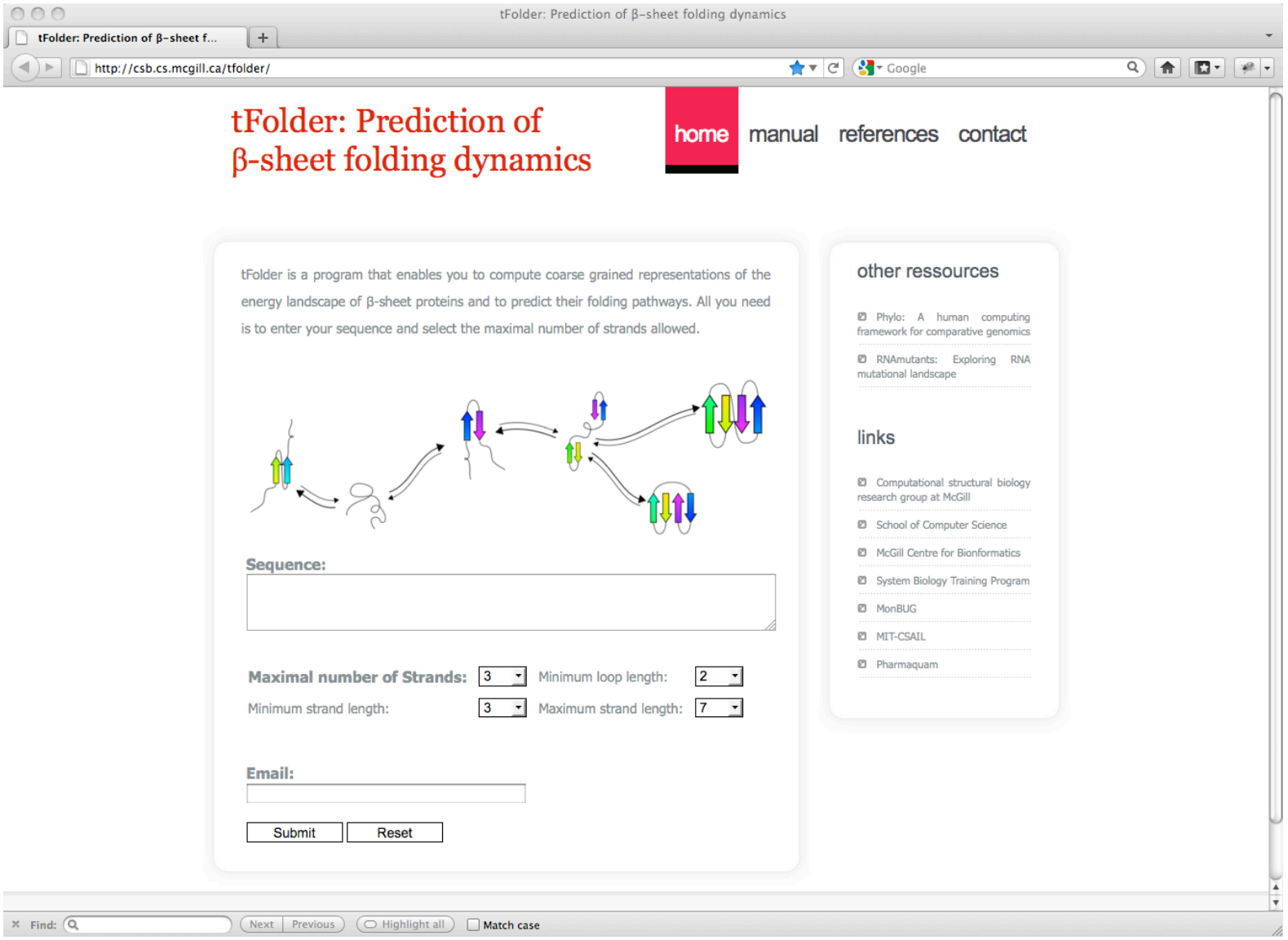
X-ray structure:



Predicted Pathway:



http://csb.cs.mcgill.ca/tfolder



Acknowledgments

McGill
- Solomon Shenker
- David Becerra

MIT
- Bonnie Berger
- Srinivas Devadas
- Charles W. O'Donnell

Ecole Polytechnique
- Jean-Marc Steyaert
- Philippe Chassignet
- Yann Ponty

Boston College
- Peter Clote