

# CMPS 6630: Introduction to Computational Biology and Bioinformatics

Tertiary Structure Prediction

# Tertiary Structure Prediction

---

## ***Why Should Tertiary Structure Prediction Be Possible?***

**Molecules obey the laws of physics!**

**Conformation space is finite**

**Proteins have regular structure**

**Proteins fold into a *small* number of protein folds**

### ***Current Hypothesis:***

Proteins generally adopt the lowest energy conformation

## **Use Chemistry and Physics to model molecular forces!**

### 1) Simulate the Physics of Folding

Start with any conformation and simulate exactly what happens in the cell

### 2) Generate a Scoring Function

Search all conformations and score each according to physics

# Tertiary Structure Prediction

---

## Wants:

- Computationally Feasible
- All atom structures with resolution comparable to experimental methods - for SBDD, Function Analysis
- Score indicating confidence in structure (reliability)

## Why:

- Reduce Time and Cost of Protein Structure Determination
- Solve Structures that are Experimentally Intractable

## Classes of Algorithms:

- Homology Modeling (Comparative Modeling)
- Fold Identification (Protein Threading)
- *Ab-initio* Methods

Target Sequence

Sequence DB

Fold Recognition

Collect Sequences  
Significantly Similar  
to the Target

Threading

Does a  
structure fit  
the sequence?

Yes

No

Is a known  
structure in  
the list

Homology  
Modeling

No

Yes

*Ab initio*

Template Based  
Modeling

Evaluate Model / Experimental Data?

Structure DB

# Critical Assessment of Protein Structure Prediction (CASP)

---

Contest of protein structure prediction

Held every 2 years

Protein Targets: Experimentally solved, but unreleased

## CASP7 - Nov 2006

95 Targets, 124 Domains, ~180 Groups Participating



*7<sup>th</sup> Community Wide Experiment on the*

## **Critical Assessment of Techniques for Protein Structure Prediction**

*Asilomar Conference Center, Pacific Grove, CA  
November 26-30, 2006*

Sponsored by the [US National Library of Medicine \(NIH/NLM\)](#), [National Institute of General Medical Sciences \(NIH/NIGMS\)](#)

Co-sponsored by: [BioSapiens Network of Excellence](#),



# Critical Assessment of Protein Structure Prediction (CASP)

---

Contest of protein structure prediction

Held every 2 years

Protein Targets: Experimentally solved, but unreleased

## **CASP7 - Nov 2006**

95 Targets, 124 Domains, ~180 Groups Participating

### **Questions Addressed:**

- 1) Are the models produced similar to the corresponding experimental structure?
- 2) Is the mapping of the target sequence onto the proposed structure (i.e. the alignment) correct?
- 3) Are comparative models more accurate than can be obtained by simply copying the best template?
- 4) Has there been progress from the earlier CASPs?
- 5) What methods are most effective?
- 6) Where can future effort be most productively focused?

# Critical Assessment of Protein Structure Prediction (CASP)

---

Difficult to score

Many possible criteria

Overall fold vs Regional fold

Secondary structure arrangement

Accuracy of side-chain placement (Coordinate vs Dihedral)

Hydrogen-bonding networks

Was the 'best' structural template identified?

Loop structure

Active sites

## Three Divisions of Algorithms:

- Homology Modeling (Comparative Modeling)
- Fold Identification (Protein Threading)
- *Ab-initio* Methods

# Model Quality - GDT

---

## Global Distance Test

Generate a seed superposition by superimposing all 3, 5, and 7 consecutive  $C_\alpha$  atoms (sliding window)

Starting with an initial set of atom pairs

Obtain the transform (minimizing RMSD) from current set of pairs

Identify all additional atom pairs with distance below threshold

Repeat until no additional atom pairs are added

**Not required to be continuous**

**Use four thresholds (1, 2, 4, 8Å)**

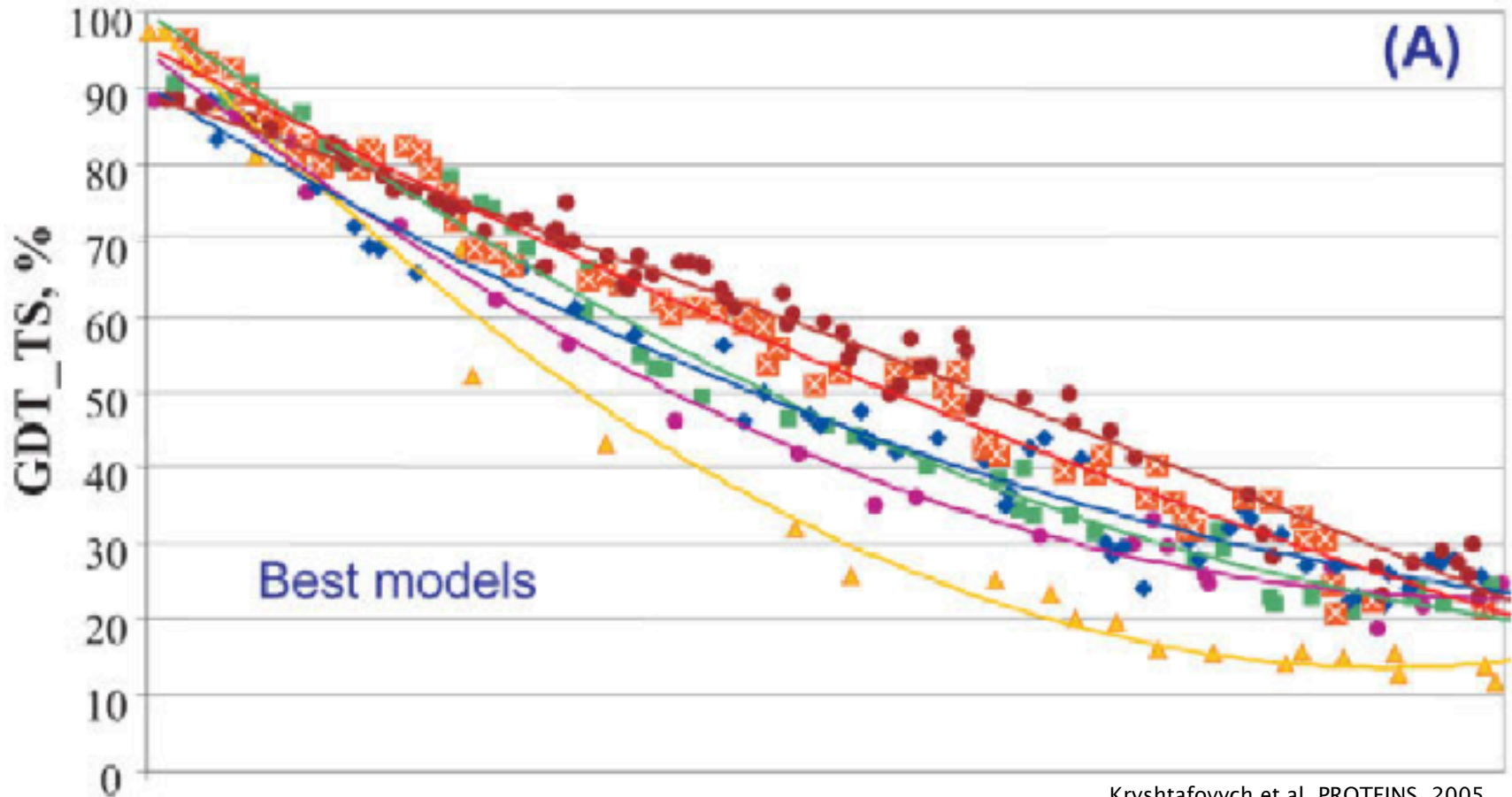
Score based on the maximum number of aligned atom pairs for each threshold

$$\text{GDT} = 1/4 [N1 + N2 + N4 + N8]$$

GDT can be expressed as a percent of total residues



# Best model for each target



Data points:



CASP1



CASP2



CASP3



CASP4



CASP5



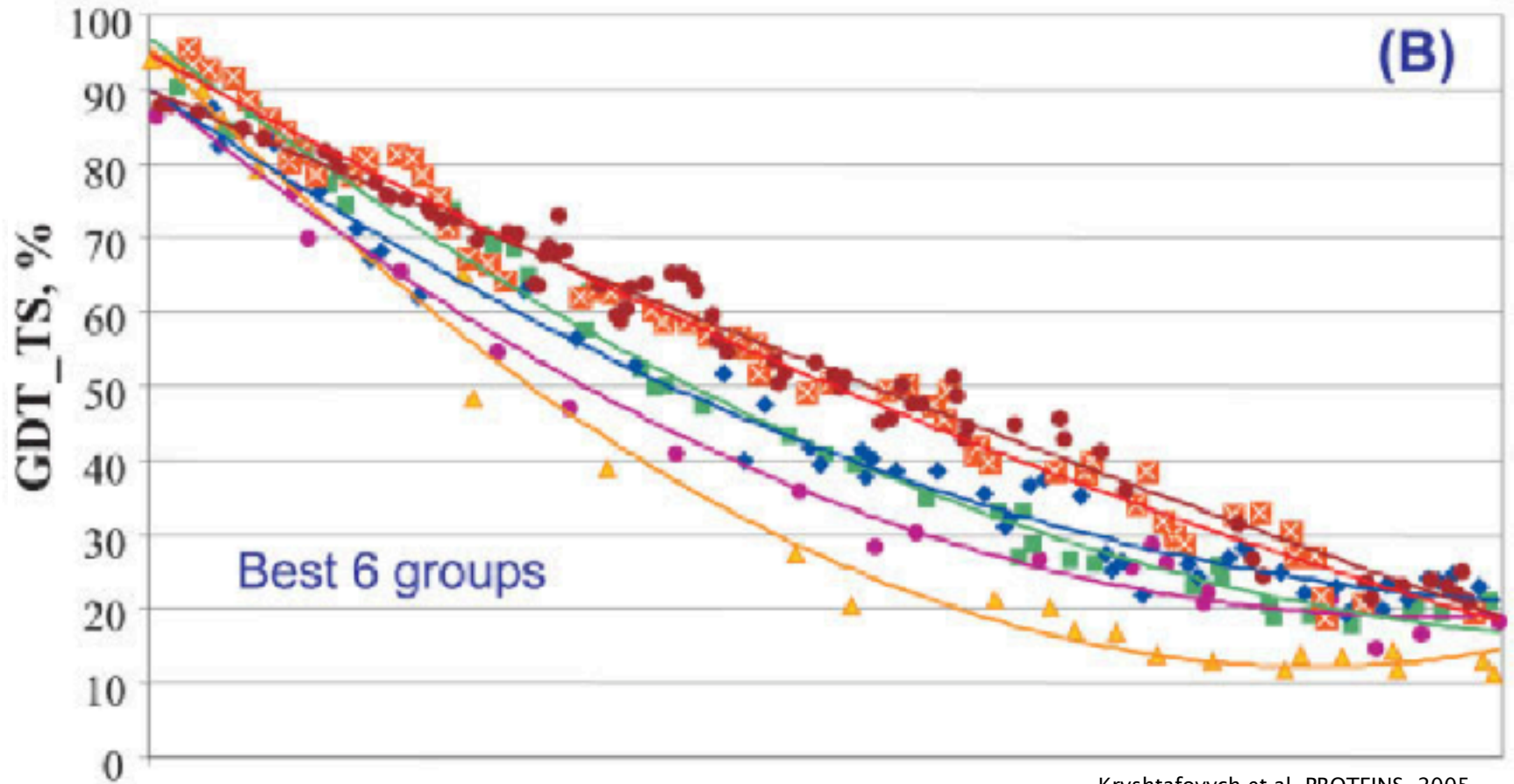
CASP6

Splines:

Increasing Difficulty



# Average of best 6 models



Data points:



CASP1



CASP2



CASP3



CASP4



CASP5



CASP6

Splines:

Increasing Difficulty



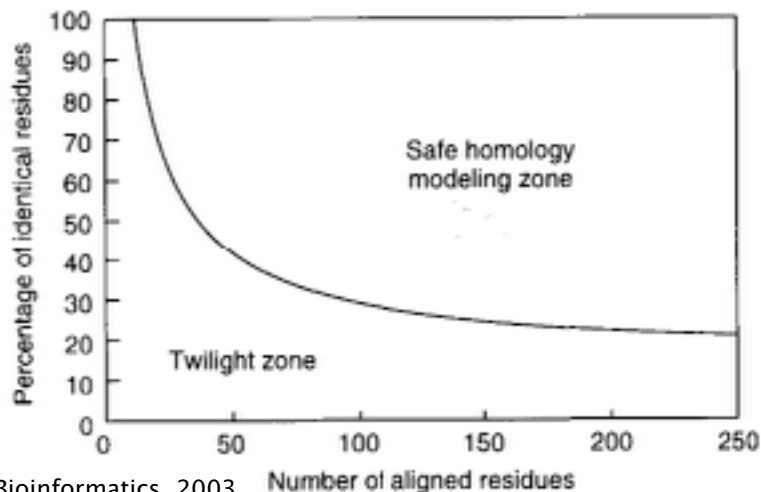
# Homology Modeling

---

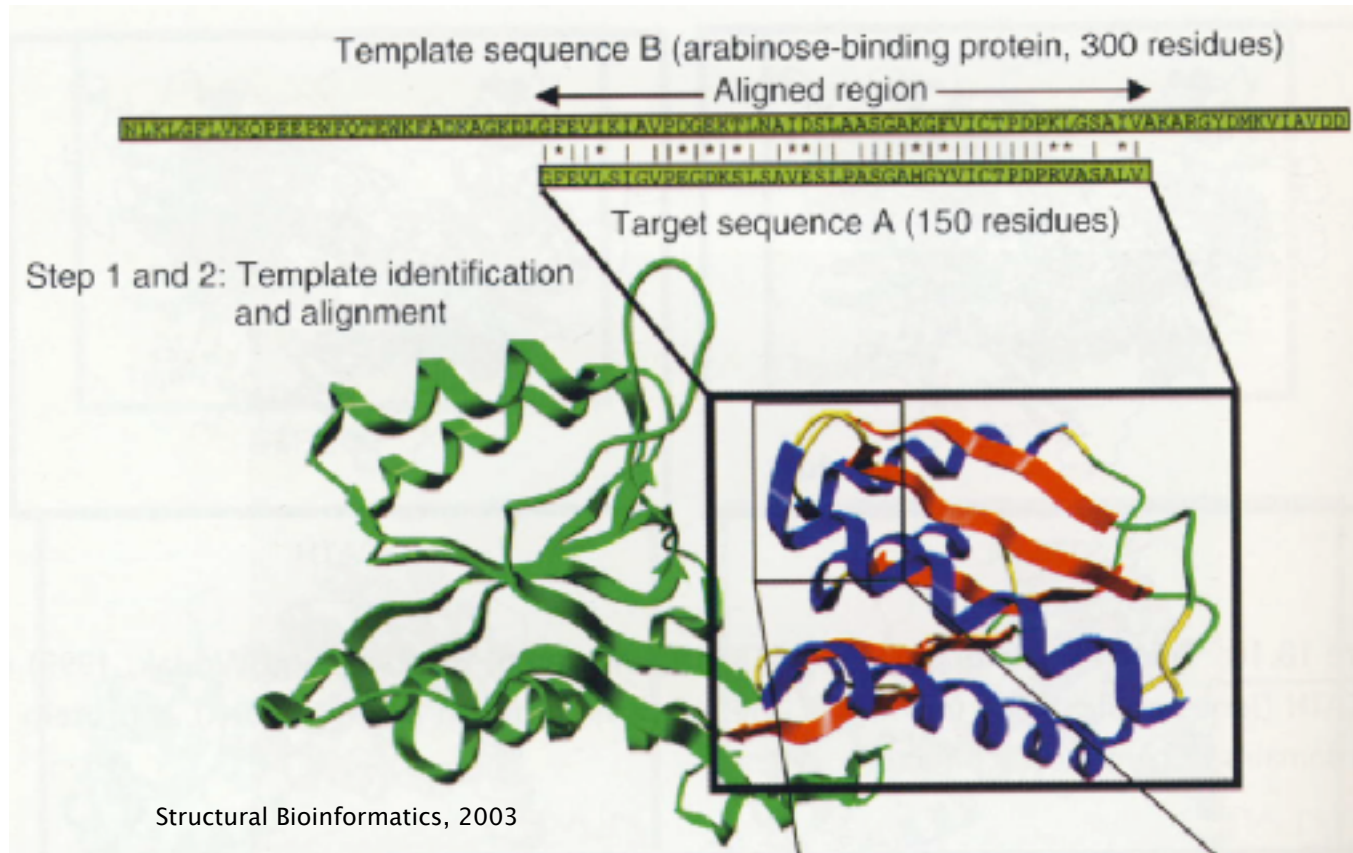
*Most simple (and accurate) of the tertiary structure predictive methods.*

**Based on two major observations:**

- 1) The structure of a protein is uniquely determined by its amino acid sequence*
- 2) During evolution, structure is more stable and changes more slowly than the associated sequence*



# Homology Modeling

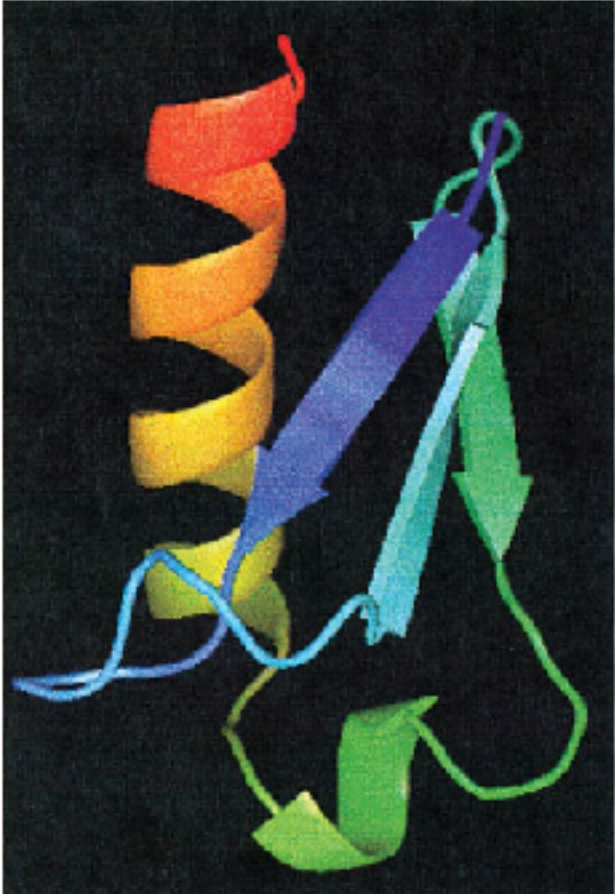


## **Goal:**

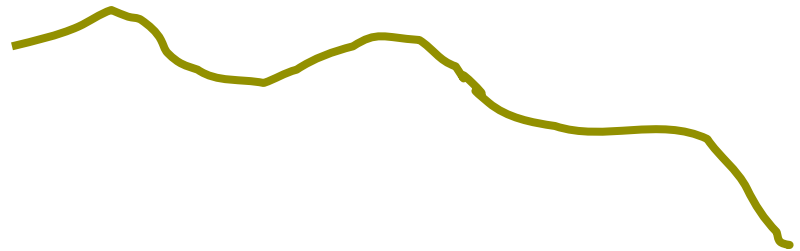
Atomic level model, on par with experimental methods

# Template-Based Modeling

---



**Template**  
(known structure)



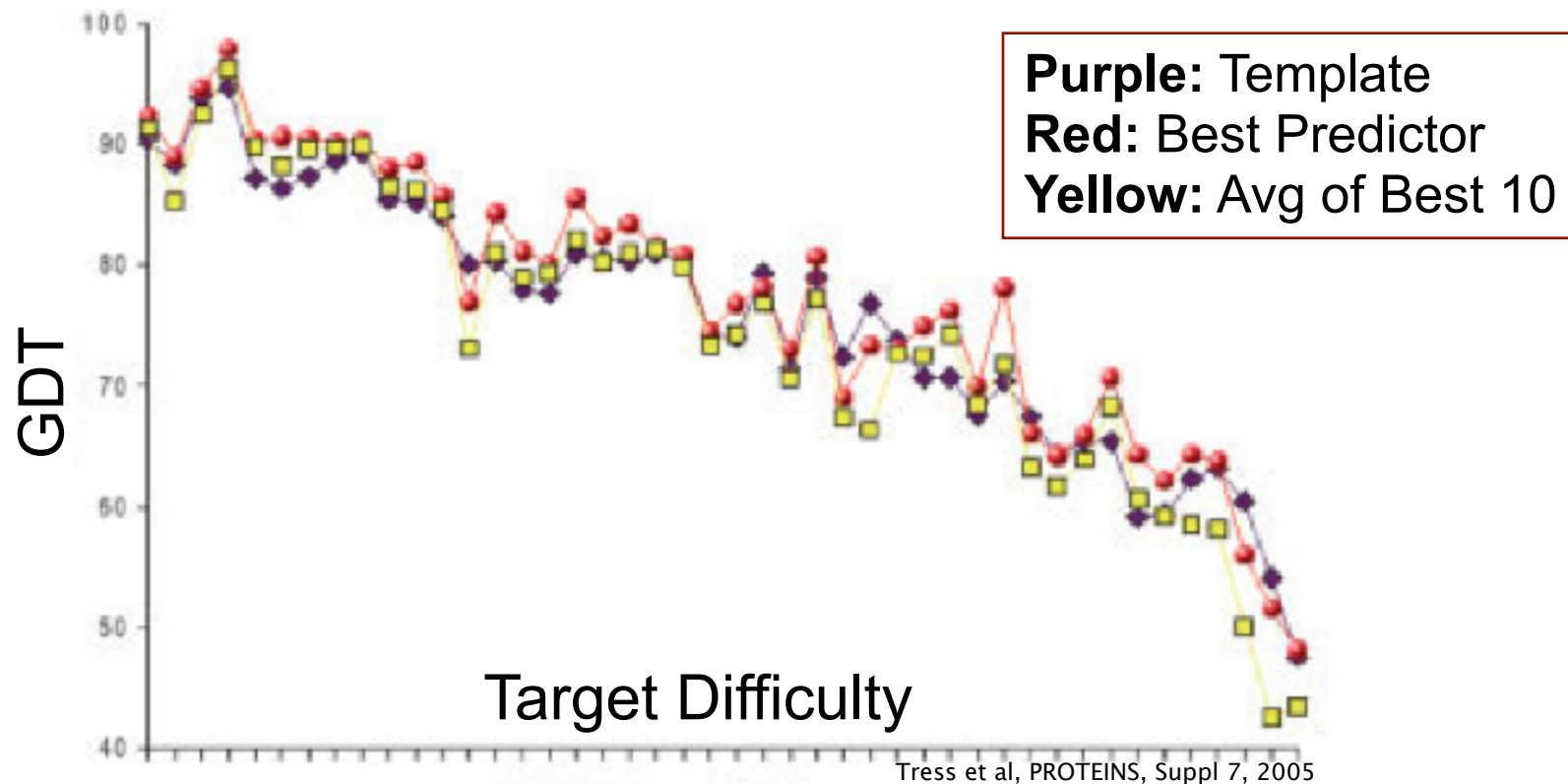
**Target**  
(unknown structure)

## Quality of computer-based models

---

- *"Until 1985, computer-based weather forecasts were less reliable than saying that the weather tomorrow will be like the weather today."*

*“In previous CASP experiments it has been demonstrated that predictors are rarely able to predict models that are closer to the target structure than the structure of the closest template.”* -Tress et al, *PROTEINS*, S7:27-45, 2005



For CASP6, most best models were better than the template. Particularly true for easier structures.

# Homology (Comparative) Modeling

---

*Errors in one stage propagate through to all later stages*

## Seven Steps

1. ***Template recognition and initial alignment***

2. Alignment correction

3. Backbone generation

4. ***Loop modeling***

5. ***Side-chain modeling***

6. ***Model optimization***

7. ***Model validation***

Much room for improvement



- Many Comparative Modeling techniques are the *same*
- Traditional Methods will improve with more solved structures
- Change in methodology likely required for major breakthrough



# Fold Recognition - Threading

---

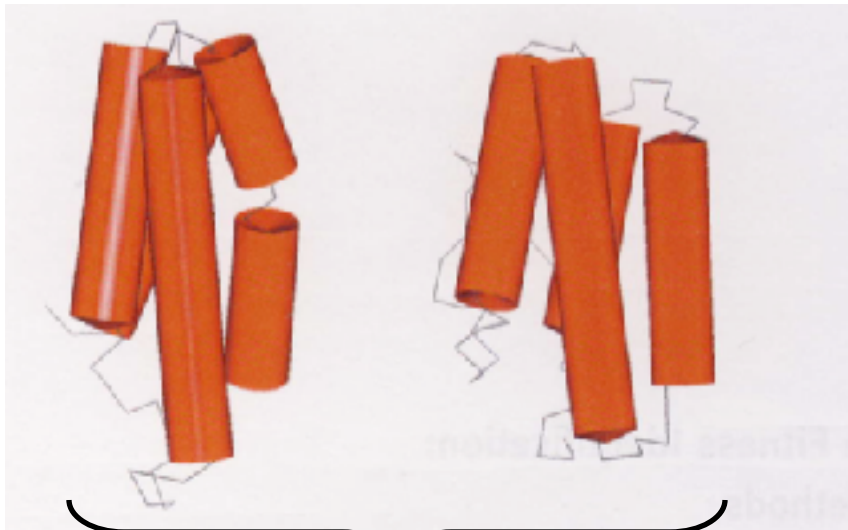
Best chance for structure prediction will use existing structures. Many seemingly unrelated proteins share a similar fold.

*If a homologous structure can not be found by sequence comparison methods, can we identify a structural model by comparing the target sequence directly to known structures?*

## Approach

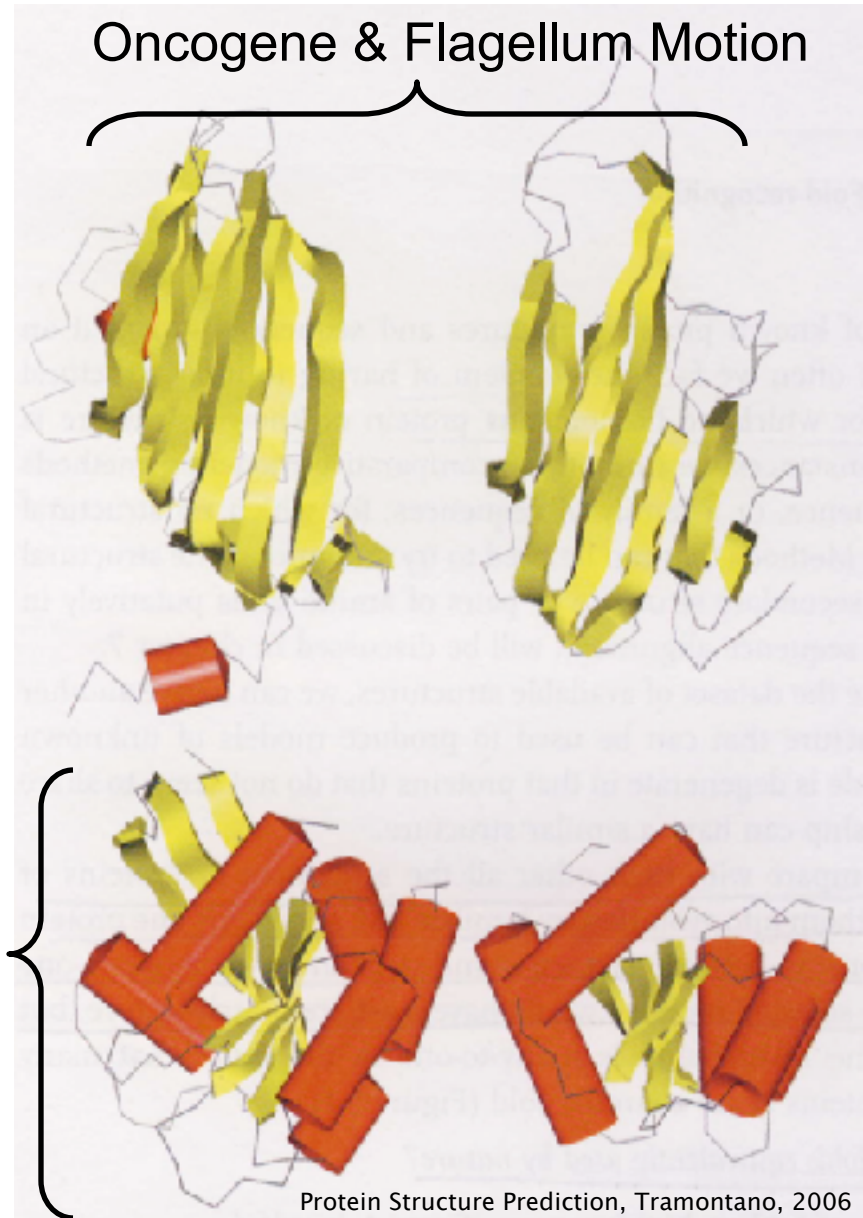
- Rather than asking what is the structure of a target protein, we ask if any known structure can serve as a reasonable model. *Then perform homology modeling.*
- Does not predict 'new' folds (ie. ones we haven't seen).
- Can be considered a '**verification**' step for the harder problem of *ab initio* structure prediction.
- As a field we need to be able to do this.

# Fold Recognition - Threading



Oxygen Transport & Electron Transport

Viral Protein & Immune System Protein



# Fold Recognition - Threading

## Goals:

- Identify a reasonably similar model structure
- Provide an alignment to that structure

...KYFDVALHLINPGLFHVDSTSVALIYKLRTPL...

Start with protein  
structure database

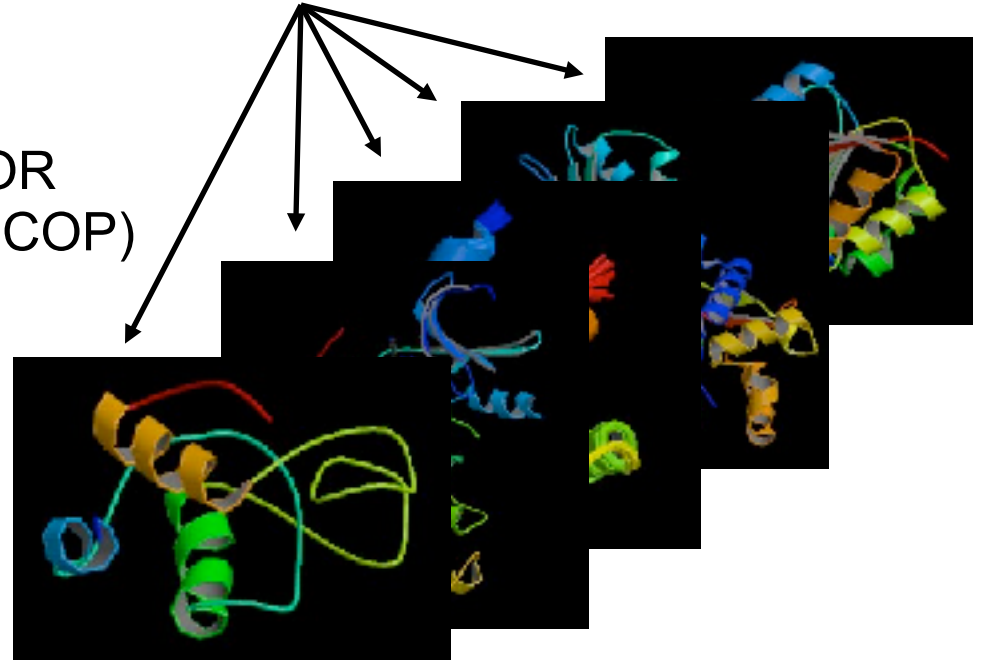
- Complete structure database OR
- Representative folds (CATH, SCOP)

Roughly compute

**$P(\text{model} \mid \text{sequence})$**

Evaluation Function

- Profile-Based Methods
- Threading Methods



# Fold Recognition - Threading

## Goals:

- Identify a reasonably similar model structure
- Provide an alignment to that structure

...KYFDVALHLINPGLFHVDSTSVALIYKLRTPL...

Start with protein  
structure database

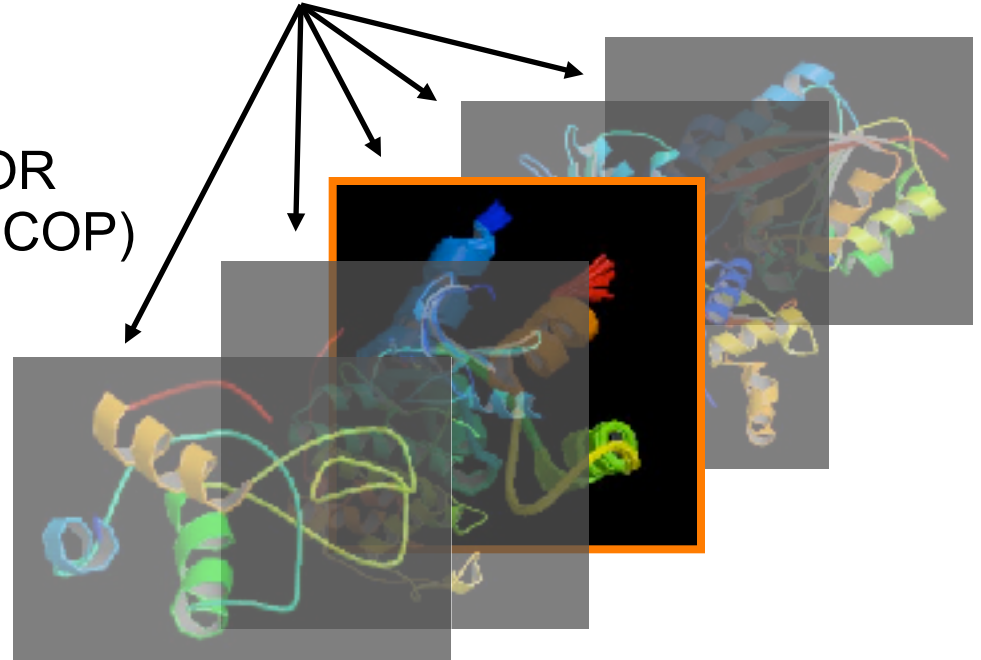
- Complete structure database OR
- Representative folds (CATH, SCOP)

Roughly compute

**P(model | sequence)**

Evaluation Function

- Profile-Based Methods
- Threading Methods



# Fold Recognition - Threading

---

## Goals:

- Identify a reasonably similar model structure
- Provide an alignment to that structure

...KYFDVALHLINPGLFHVDSTSVALIYKLRTPL...

Start with protein  
structure database

- Complete structure database OR
- Representative folds (CATH, SCOP)

Roughly compute

**P(model | sequence)**

Evaluation Function

- Profile-Based Methods
- Threading Methods



**Template-Based  
Modeling**

# Fold Recognition - Threading

## Threading Methods

- A sequence of amino acids (often including side-chains) is thread through a known structure and may fit or 'click' with a target structure
- Use of ***pairwise potential function***
- Considers ***long-range interactions***
- Stabilizing forces can come from interactions of residues distant in sequence

## Pairwise Potential Function

***Substitution Matrices***

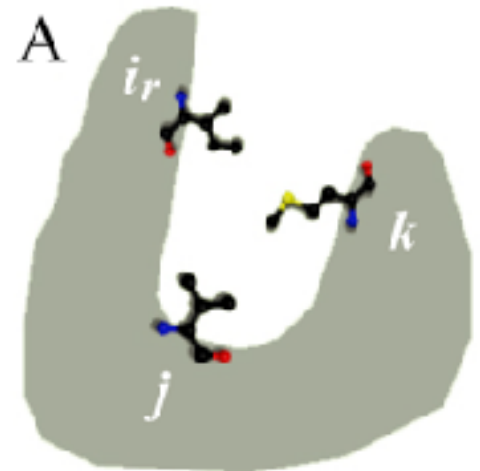
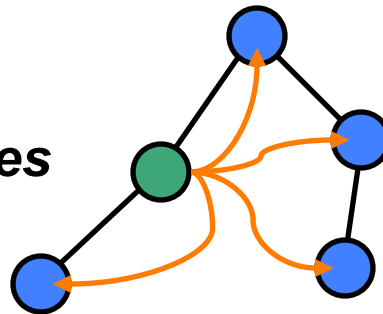
***Secondary Structure Propensities***

***Hydrophobicity***

***Accessibilities***

***Steric fit***

***Likelihood of interacting with neighbour residues***



# Fold Recognition - Threading

## Pairwise Potential Function – Interaction Likelihood

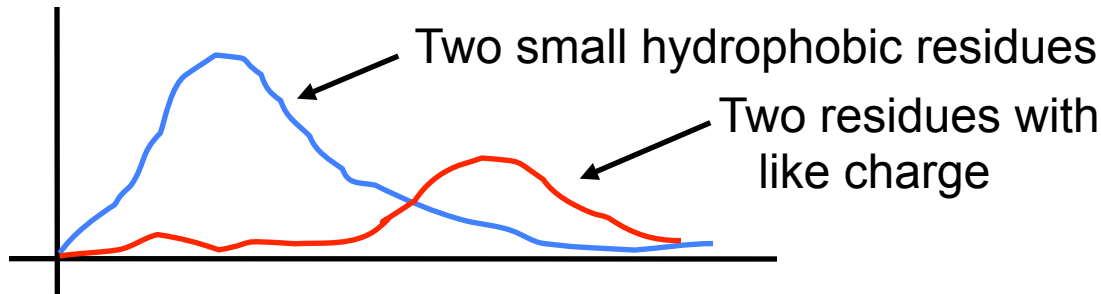
Based on frequency of observing the interaction of two amino acids

### Simple Model:

Count instances of all pairs of AAs occurring within a single threshold distance in DB (i.e., 5Å)

### More Fine Grained Model:

Determine frequency of all AAs occurring as a function of distance  $d$   
Many parameters to estimate



Sparsely observed events are often smoothed

$$f^{ab}(s) \approx \frac{1}{1 + m\sigma} f(s) + \frac{m\sigma}{1 + m\sigma} g^{ab}(s)$$

$m$ : num observations

$f^{ab}(s)$ : corrected frequency

$g^{ab}(s)$ : uncorrected frequency

$f(s)$ : frequency of observing  
any two AA at distance  $s$

$\sigma$ : constant

# Fold Recognition - Threading

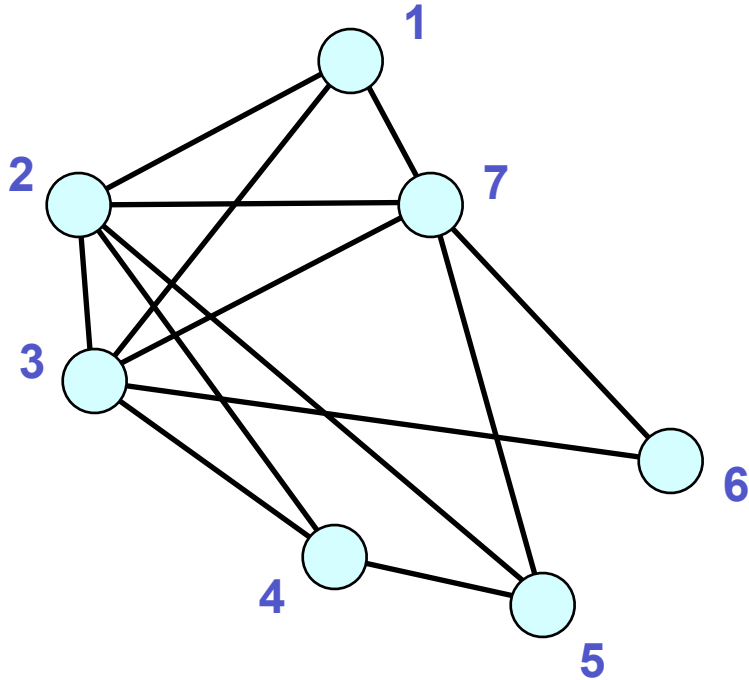
---

## How hard is protein threading?

With pairwise energy function is at least as hard as **MAX-CUT** (which is NP-Complete)



Given graph  $G = (V, E)$ , find a cut  $(S, T)$  of  $V$  with maximum number of edges between  $S$  and  $T$ .





# Fold Recognition - Threading

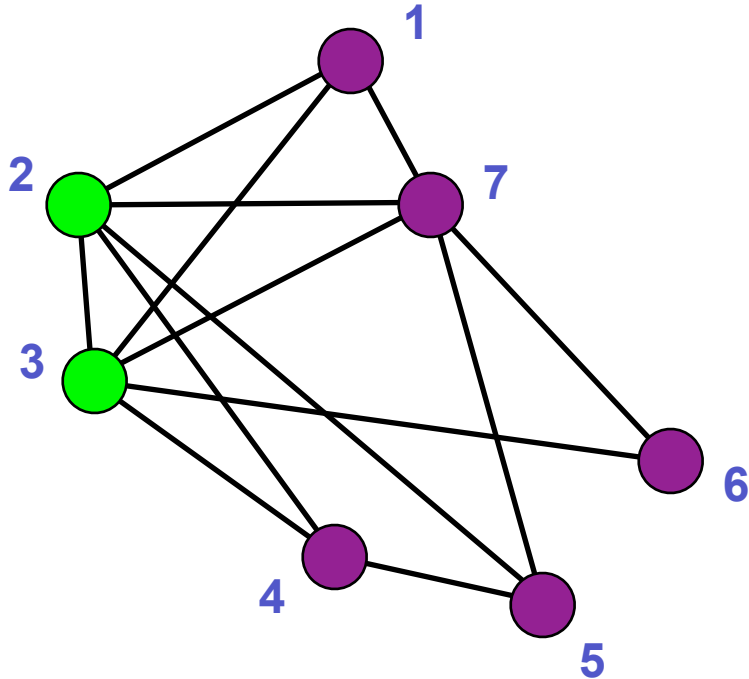
---

## How hard is protein threading?

With pairwise energy function is at least as hard as **MAX-CUT** (which is NP-Complete)



Given graph  $G = (V, E)$ , find a cut  $(S, T)$  of  $V$  with maximum number of edges between  $S$  and  $T$ .



# Fold Recognition - Threading

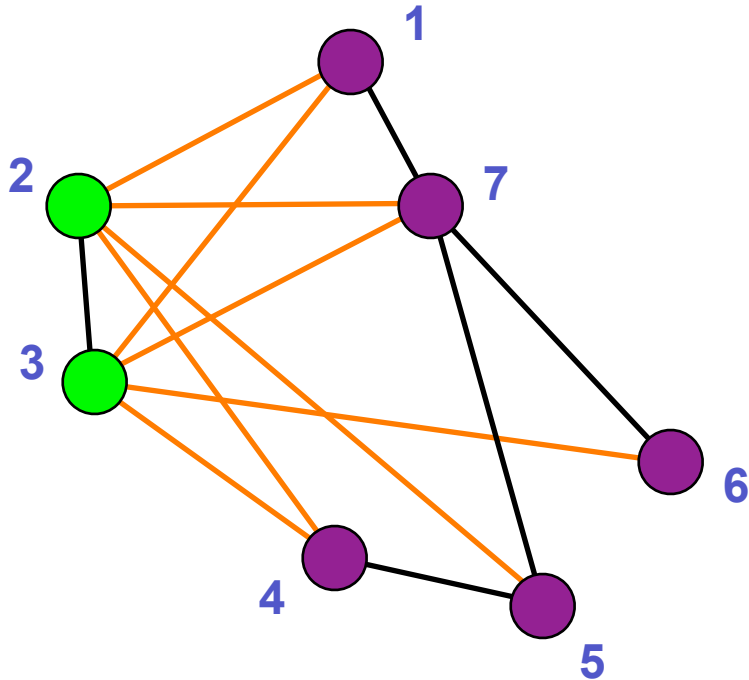
---

## How hard is protein threading?

With pairwise energy function is at least as hard as **MAX-CUT** (which is NP-Complete)



Given graph  $G = (V, E)$ , find a cut  $(S, T)$  of  $V$  with maximum number of edges between  $S$  and  $T$ .



# Fold Recognition - Threading

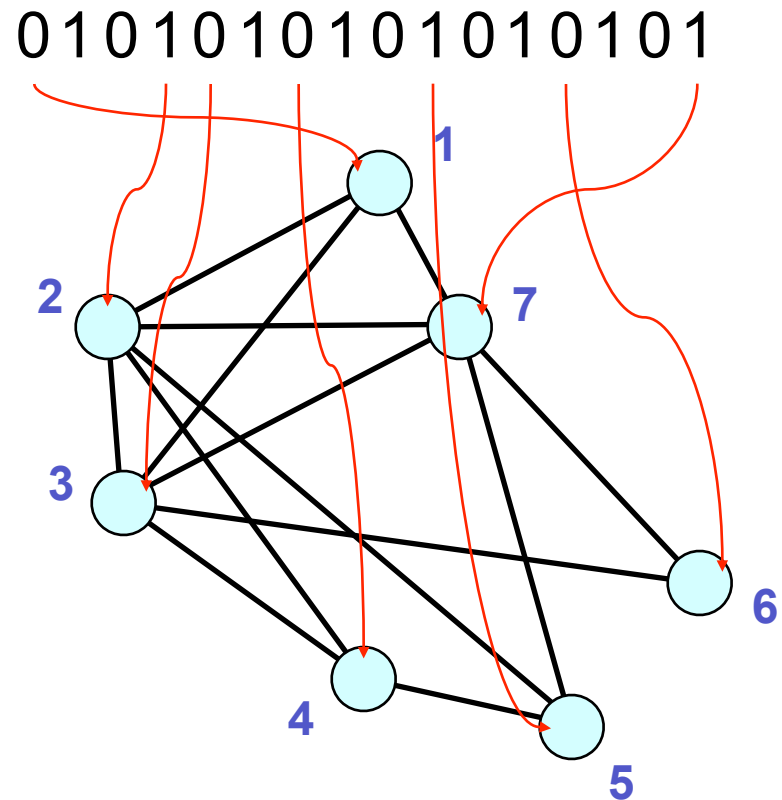
## How hard is protein threading?

With pairwise energy function is at least as hard as **MAX-CUT** (which is NP-Complete)



Given graph  $G = (V, E)$ , find a cut  $(S, T)$  of  $V$  with maximum number of edges between  $S$  and  $T$ .

Consider threading the string  $(01)_{|V|}$  to the graph  $G$ .



# Fold Recognition - Threading

## How hard is protein threading?

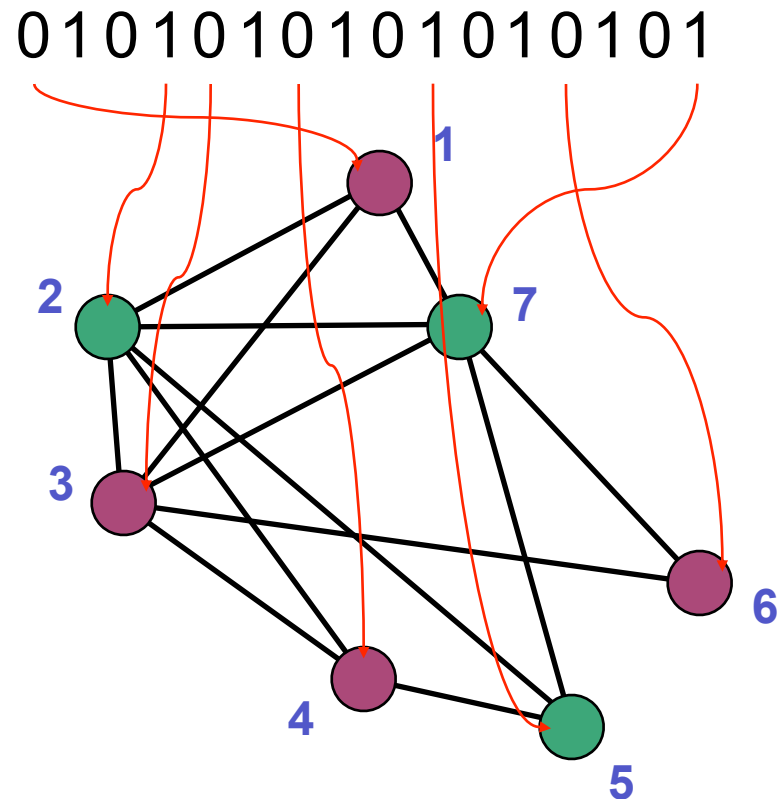
With pairwise energy function is at least as hard as **MAX-CUT** (which is NP-Complete)



Given graph  $G = (V, E)$ , find a cut  $(S, T)$  of  $V$  with maximum number of edges between  $S$  and  $T$ .

$$\text{Score} = \sum_{e_{i,j} \in E} S(e_{i,j})$$

$$S(e_{i,j}) = 1 \text{ if } \text{label}(v_i) \neq \text{label}(v_j) \\ = 0 \text{ if } \text{label}(v_i) = \text{label}(v_j)$$



# Fold Recognition - Threading

## How hard is protein threading?

With pairwise energy function is at least as hard as **MAX-CUT** (which is NP-Complete)

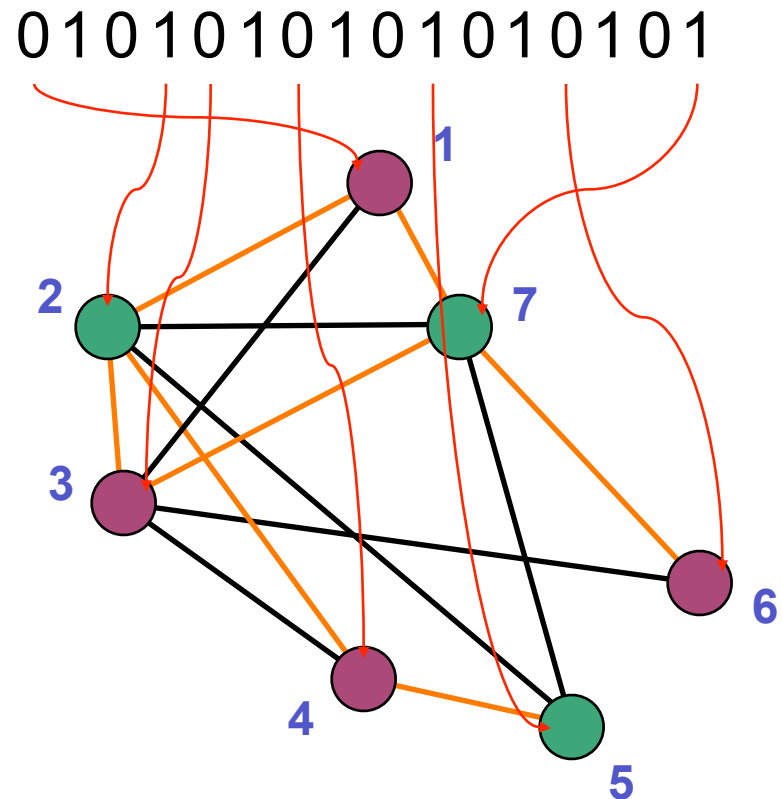


Given graph  $G = (V, E)$ , find a cut  $(S, T)$  of  $V$  with maximum number of edges between  $S$  and  $T$ .

$$\text{Score} = \sum_{e_{i,j} \in E} S(e_{i,j})$$

$$S(e_{i,j}) = 1 \text{ if } \text{label}(v_i) \neq \text{label}(v_j) \\ = 0 \text{ if } \text{label}(v_i) = \text{label}(v_j)$$

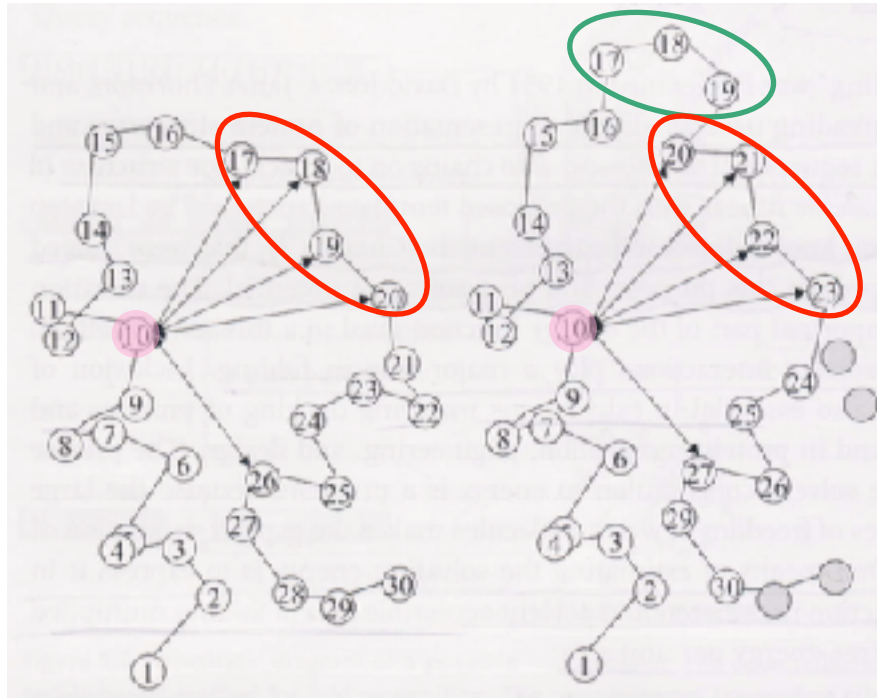
$$\text{Score} = 7$$



# Fold Recognition - Threading

## Threading Methods: Pairwise Potential Function

Breaks dynamic programming independence assumption!  
The score obtained for a match depends on other positions in the alignment.



# Fold Recognition - Threading

---

## Threading Methods: Pairwise Potential Function

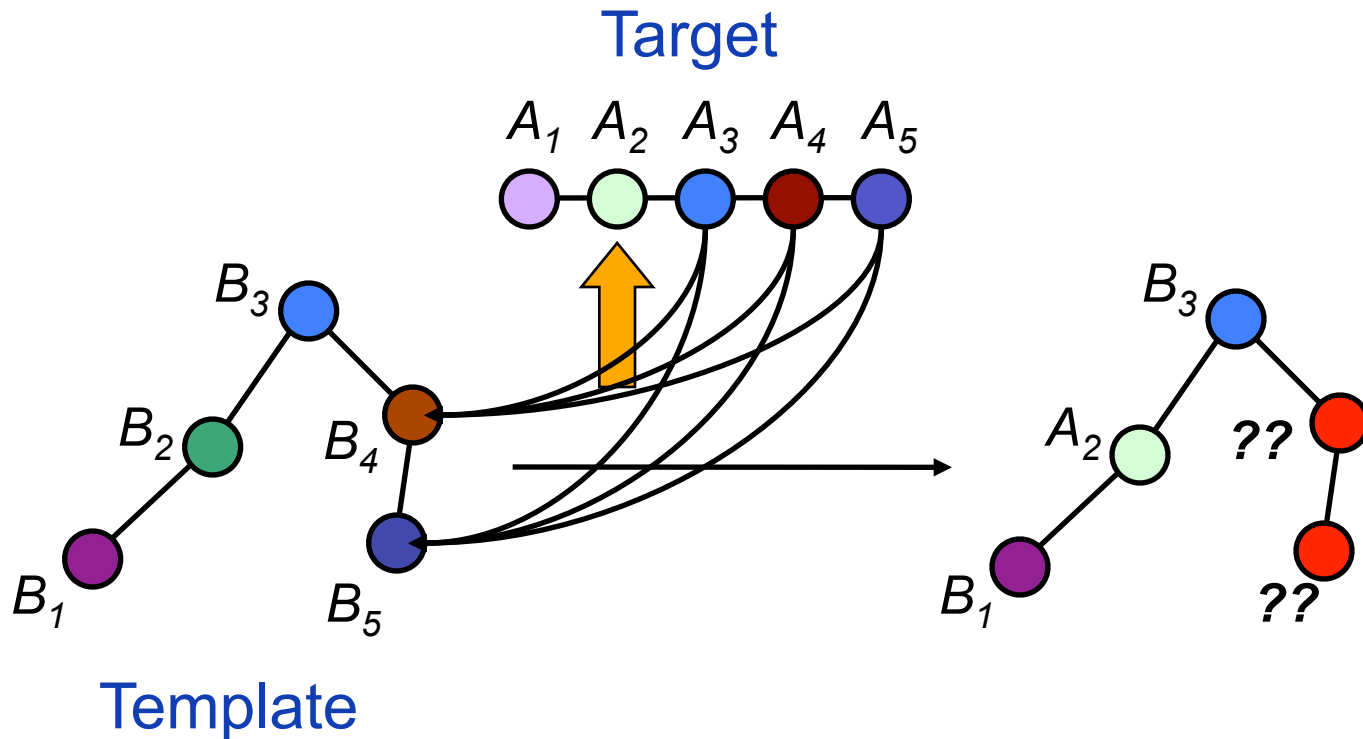
Breaks dynamic programming independence assumption!  
The score obtained for a match depends on other positions in the alignment.

### Possible Work-Arounds:

Alignment techniques that can handle non-local interactions  
Likely not guaranteed polynomial-time  
Double dynamic programming  
Various Approximations

# Fold Recognition - Threading

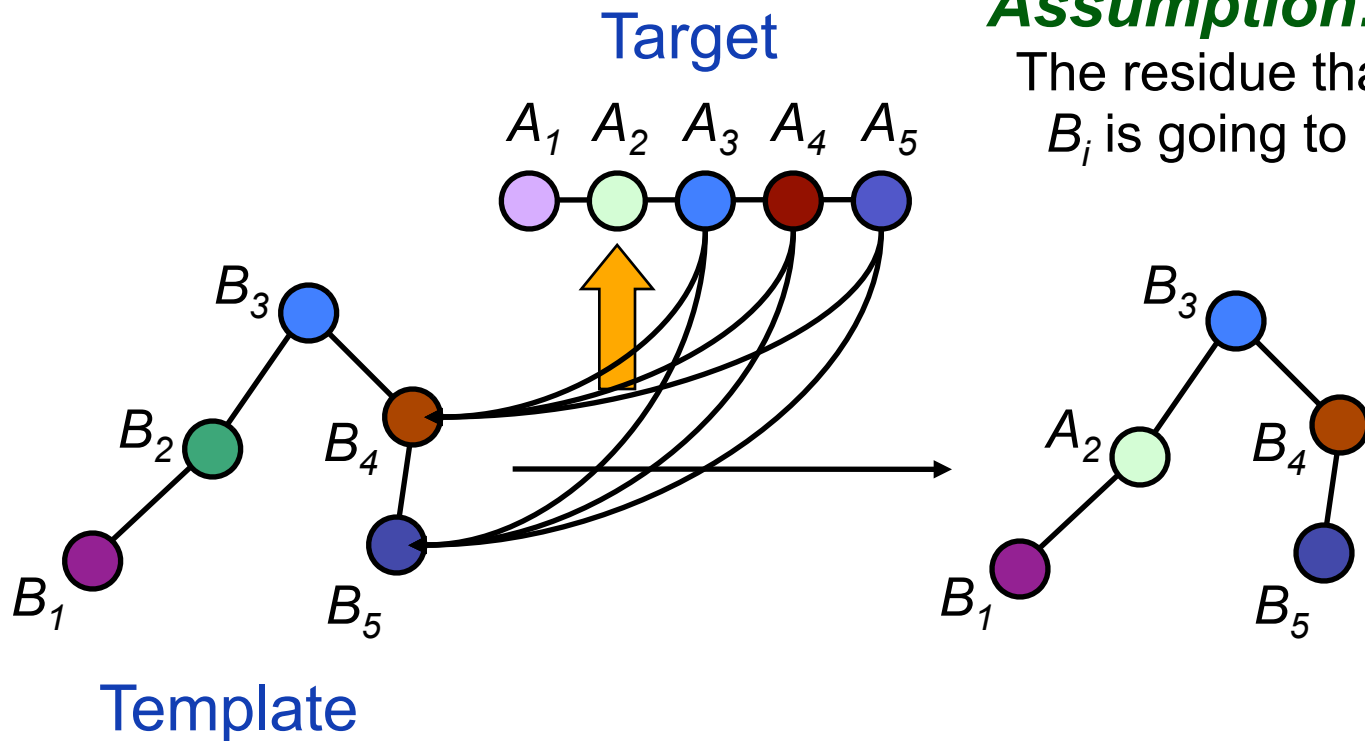
## Threading Methods





# Fold Recognition - Threading

## Threading Methods



### **Assumption:**

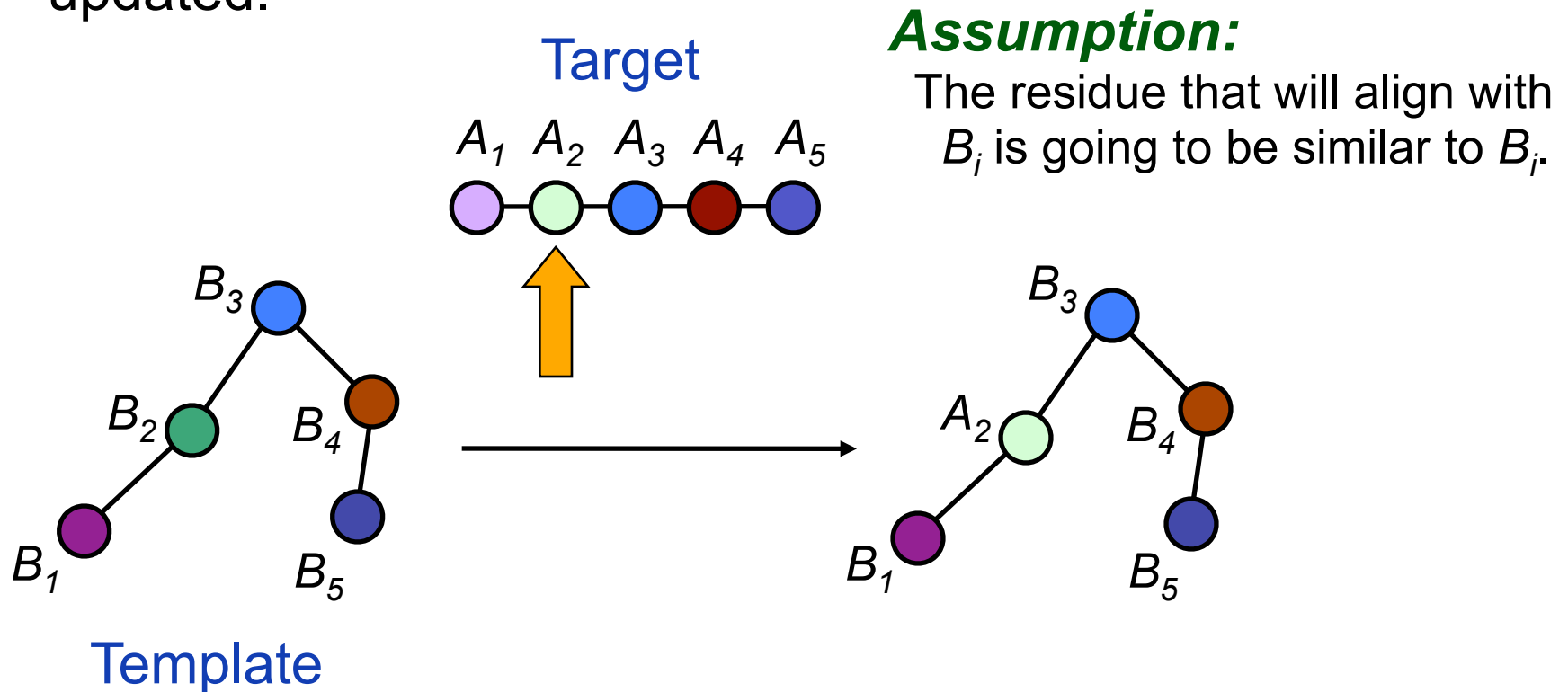
The residue that will align with  $B_i$  is going to be similar to  $B_i$ .

# Fold Recognition - Threading

## Threading Methods

### Frozen Approximation

Interaction partners for residue  $i$  are frozen to be the same as in the **template** (i.e., not the target) and are iteratively updated.



# Fold Recognition - Threading

## Threading Methods

Consider threading just **Core** elements

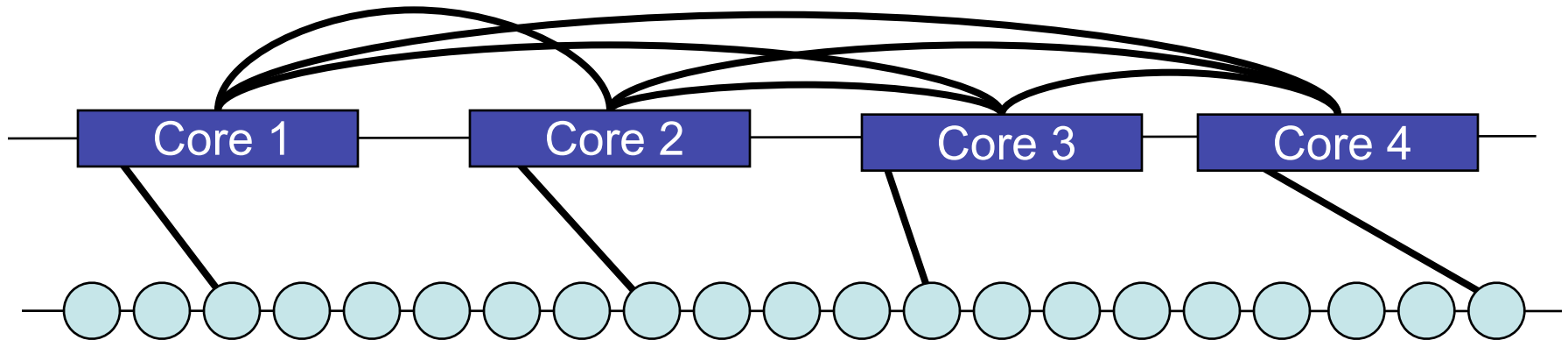
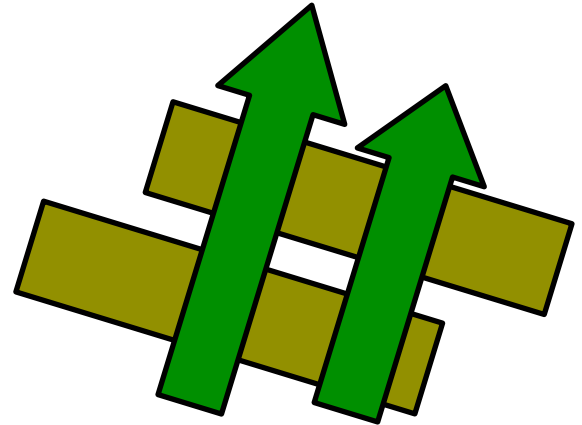
Maintain topology (ordering)

Gaps are not allowed within a core

So threading only needs to specify a start position for each core

Threading Score:

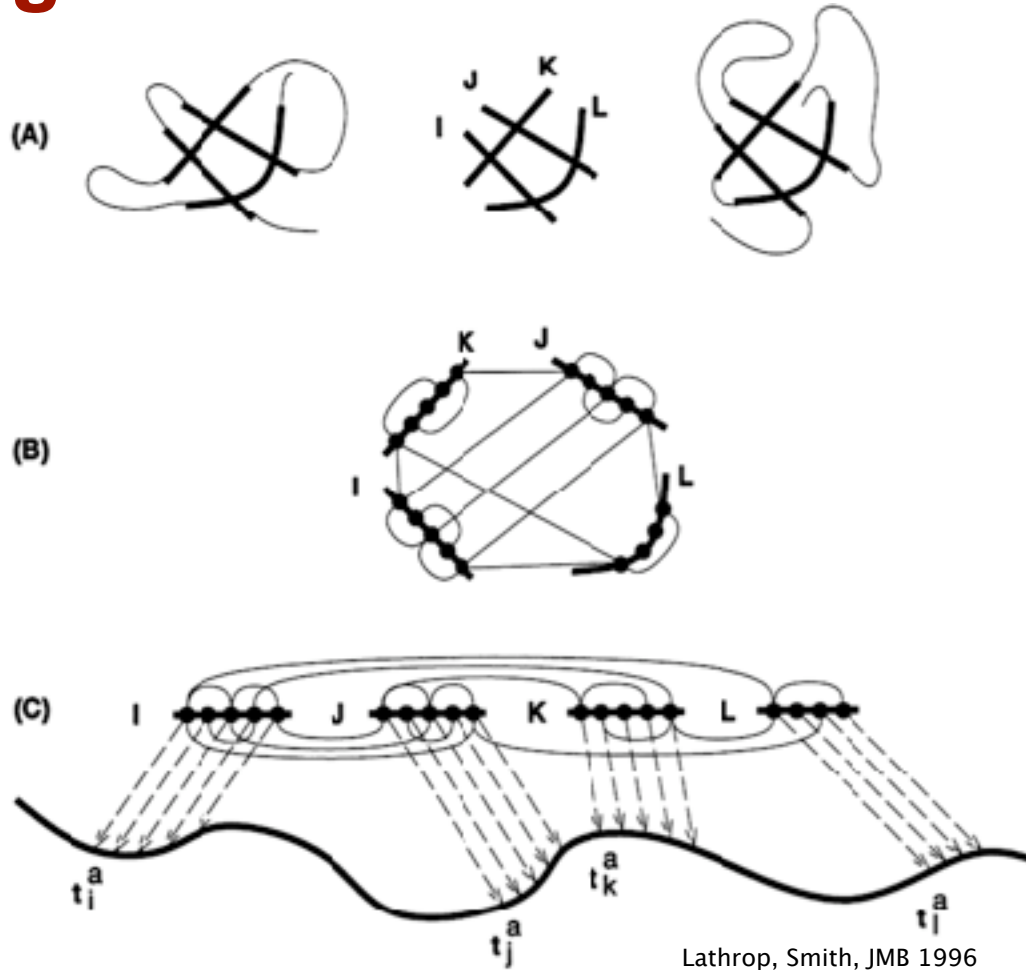
Sum of self-, pairwise-, and gap terms



# Fold Recognition - Threading

## Threading Methods

## Branch-and-Bound Search

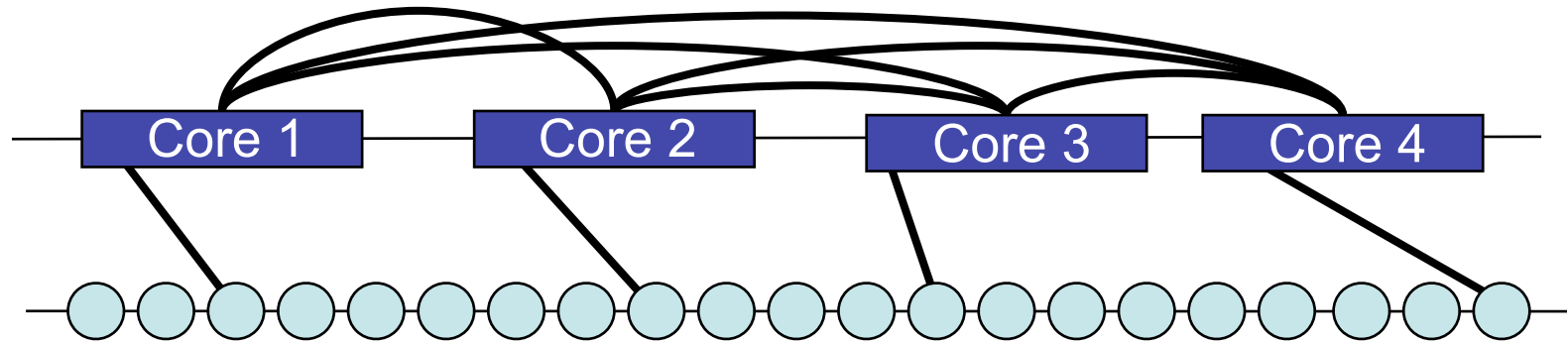


Lathrop & Smith, "Global Optimum Protein Threading with Gapped Alignment and Empirical Pair Score Functions" *Journal of Molecular Biology*, 1996

# Fold Recognition - Threading

## Threading Methods

## Branch-and-Bound Search



$m$  core elements  $C = \{C_1, C_2, \dots, C_m\}$  with lengths  $c_1, \dots, c_m$

$m$  loops,  $z_0, \dots, z_{m-1}$ , loop  $z_i$  connects  $C_i$  and  $C_{i+1}$

Loops have minimum / maximum lengths  $l^{min}, l^{max}$

Protein sequence  $\mathbf{a}$  has  $n$  amino acids  $a_1, \dots, a_n$

A threading  $\mathbf{t}^a = (t^a_1, t^a_2, \dots, t^a_m) : t^a_i$  is the starting AA index of  $C_i$

Each  $t^a_i$  is bounded:

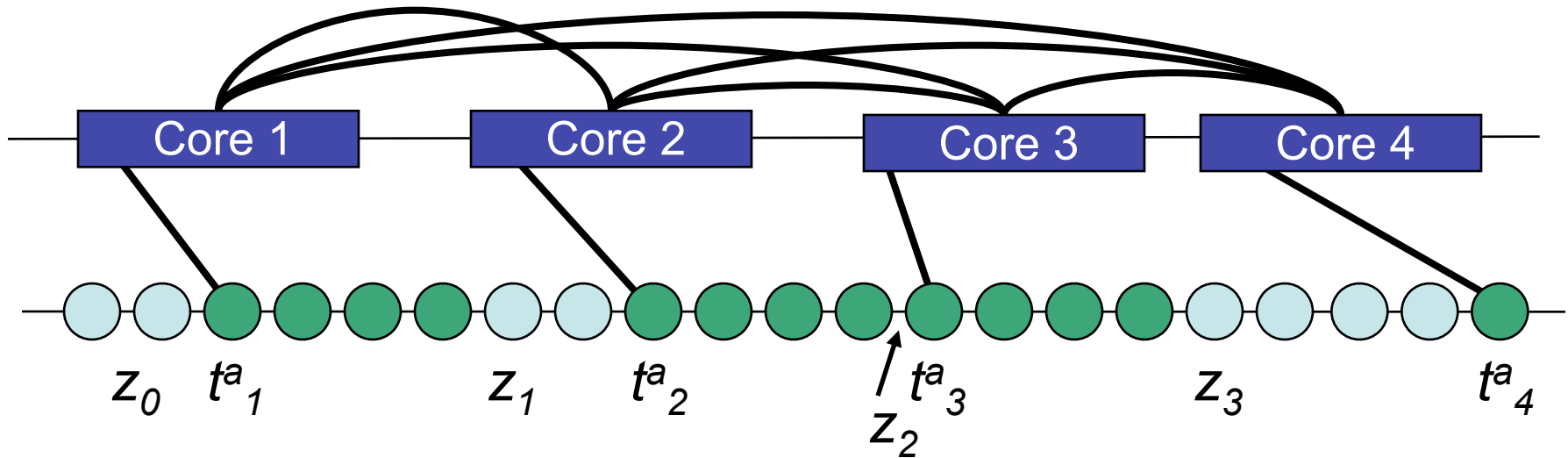
$$1 + \sum_{j < i} (c_j + l^{min}) \leq t^a_i \leq n + 1 - \sum_{j \geq i} (c_j + l^{min})$$

$$t^a_i + c_i + l^{min} \leq t^a_{i+1} \leq t^a_i + c_i + l^{max}$$

# Fold Recognition - Threading

## Threading Methods

## Branch-and-Bound Search



$$\text{score}(t_a) = \underbrace{\sum_i E(i, t_i^a)}_{\text{Self-Energy}} + \underbrace{\sum_i \sum_{j>i} E(i, j, t_i^a, t_j^a)}_{\text{Pairwise-Energy}}$$

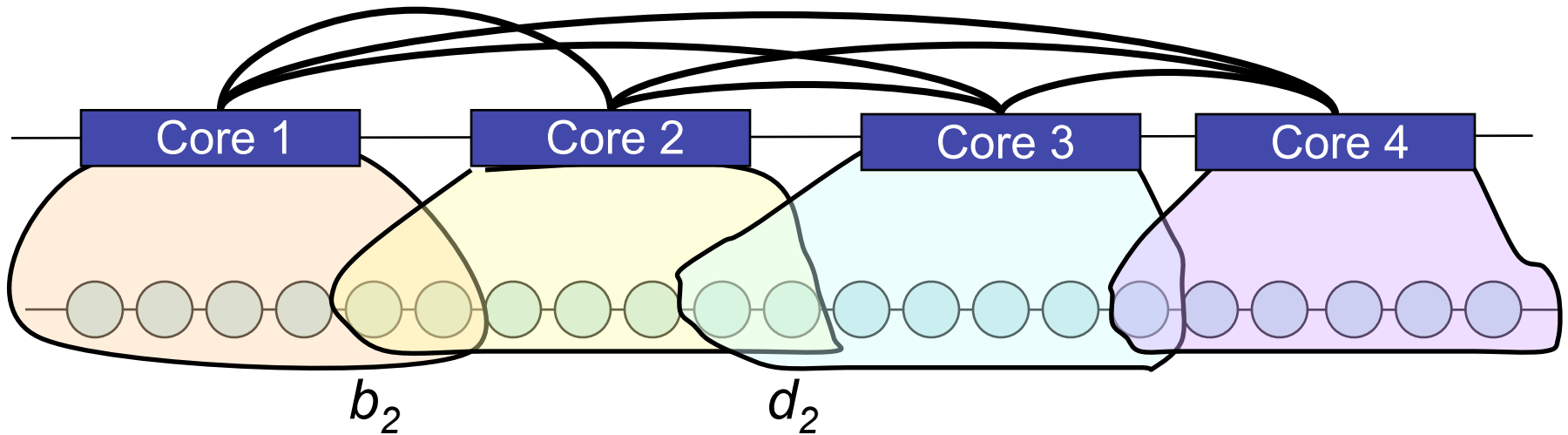
**Self-Energy:**  
of core  $C_i$  at  
position  $t_i^a$

**Pairwise-Energy:**  
between core  $C_i$  at  $t_i^a$   
and core  $C_j$  at  $t_j^a$

# Fold Recognition - Threading

## Threading Methods

## Branch-and-Bound Search



Each core has a limited range of possible alignment positions.

## Branch-and-Bound Search

Keep track of allowable threading space  $T$ .

For each core,  $C_i$ , store the lower  $b_i$  and upper  $d_i$  placement limits

Compute lower bounds on the score for each region.

Explore regions with lowest minimum bound.

# Fold Recognition - Threading

## Threading Methods

## Branch-and-Bound Search

### A Lower Bound on Threading Score over Threadings $T$

$$\begin{aligned} \min_{t \in T} \text{score}(t) &= \min_{t \in T} \sum_i \left[ E(i, t_i) + \sum_j E(i, j, t_i, t_j) \right] \\ &\geq \sum_i \left[ \min_{b_i \leq x \leq d_i} E(i, x) + \sum_{j > i} \min_{\substack{b_j \leq y \leq d_j \\ b_i \leq z \leq d_i}} E(i, j, y, z) \right] \end{aligned}$$

**Polynomial number of pairwise terms**

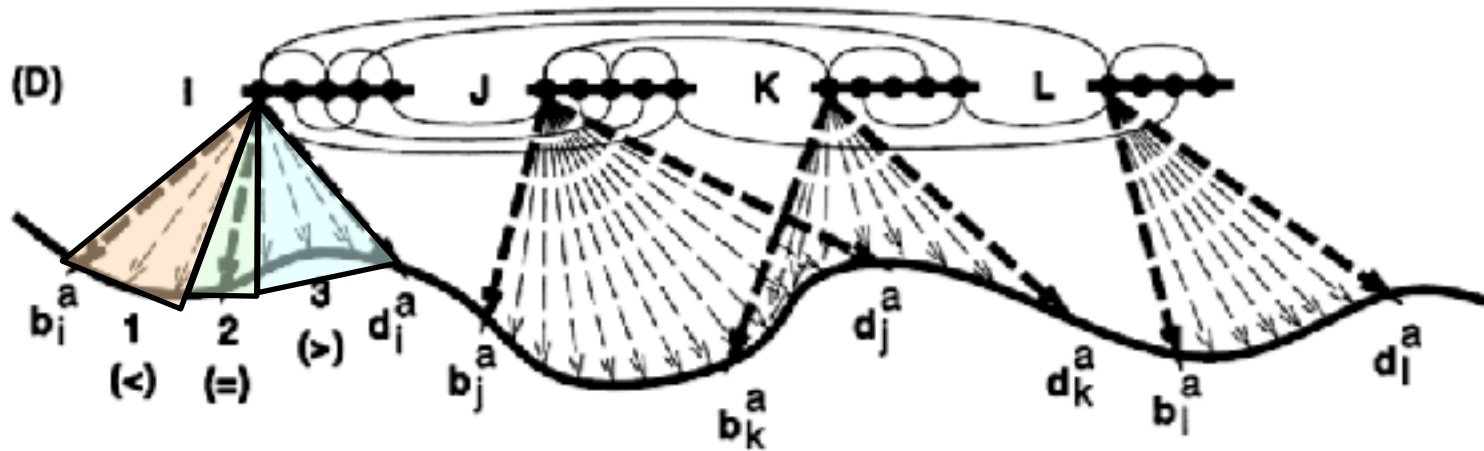
**Exponential number of threadings**



# Fold Recognition - Threading

## Threading Methods

## Branch-and-Bound Search



Lathrop, Smith, JMB 1996

$$(b_i^a = b_i^a, d_i^a = 1), (b_j^a = b_j^a, d_j^a = d_j^a), (b_k^a = b_k^a, d_k^a = d_k^a), (b_l^a = b_l^a, d_l^a = d_l^a)$$

$$(b_i^a = 2, d_i^a = 2), (b_j^a = b_j^a, d_j^a = d_j^a), (b_k^a = b_k^a, d_k^a = d_k^a), (b_l^a = b_l^a, d_l^a = d_l^a)$$

$$(b_i^a = 3, d_i^a = d_i^a), (b_j^a = b_j^a, d_j^a = d_j^a), (b_k^a = b_k^a, d_k^a = d_k^a), (b_l^a = b_l^a, d_l^a = d_l^a)$$

# Fold Recognition - Threading

---

## Threading Methods

## Branch-and-Bound Search

### Algorithm:

#### *Initialization:*

- Compute a lower bound for the set of all threadings
- Initialize a sorted list to contain one entry (the set of all threadings with its lower bound)

#### *Iteration:*

- Remove from the list the set having the lowest lower bound
- If the set contains only a single threading, stop, and announce success  
This is a global optimum threading
- Otherwise, split the the set into smaller subsets (split the core with the largest  $d_i - b_j$ )
- Compute a lower bound for each new subset
- Merge the new subsets into the list, sorted by lower bound

***They have identified the global optimum in search spaces as large as  $10^{31}$  threadings -- at rates of  $10^{28}$  equivalent threadings per second.***



# Fold Recognition - Threading

## Threading Methods

## Branch-and-Bound Search

Protein number	PLDB code	Protein length	Number of core segments	Search Space Size	Number of search iterations	Total (search-only) seconds	Equivalent threadings per iteration	Equivalent threadings per second
1	256b	106	5	6.19e + 3	6	1 (1)	1.03e + 3	6.19e + 3
2	1end	137	3	4.79e + 4	6	1 (1)	7.98e + 3	4.79e + 4
3	1nrb	129	4	5.89e + 4	7	1 (1)	8.41e + 3	5.89e + 4
4	2mhr	118	4	9.14e + 4	7	1 (1)	1.31e + 4	9.14e + 4
5	351c	82	4	1.12e + 5	5	1 (1)	2.24e + 4	1.12e + 5
6	1bgc	174	4	1.63e + 5	6	1 (1)	2.72e + 4	1.63e + 5
7	1ubq	76	5	1.70e + 5	6	1 (1)	2.83e + 4	1.70e + 5
8	1mbd	153	8	1.77e + 5	10	1 (1)	1.77e + 4	1.77e + 5
9	1lis	136	5	5.02e + 5	7	1 (1)	7.17e + 4	5.02e + 5
48	1apa	266	14	3.56e + 17	141	18 (6)	2.52e + 15	1.98e + 16
49	4tgl	269	14	5.86e + 18	361	22 (7)	1.62e + 16	2.66e + 17
50	5trn	316	14	6.51e + 18	164	28 (7)	3.97e + 16	2.32e + 17
51	1lec	242	15	7.01e + 18	320	26 (12)	2.19e + 16	2.70e + 17
52	1nar	290	17	2.33e + 19	3984	208 (183)	5.85e + 15	1.12e + 17
53	1sfl	273	15	4.36e + 19	541	32 (13)	8.05e + 16	1.36e + 18
54	5cpu	307	16	1.22e + 20	1089	72 (50)	1.12e + 17	1.69e + 18
55	9api	384	17	1.95e + 22	290	57 (25)	6.71e + 19	3.41e + 20
56	2had	310	19	2.57e + 22	4027	201 (179)	6.39e + 18	1.28e + 20
57	2cpp	414	20	6.37e + 24	3068	205 (164)	2.08e + 21	3.11e + 22
58	6taa	478	23	9.63e + 31	4917	1409 (1267)	1.96e + 28	6.83e + 28

Lathrop, Smith, JMB 1996

Self-Threadings

# Fold Recognition - Threading

---

## Differences Between Fold Recognition Algorithms

- **Protein Model and Interaction Description**  
The full three-dimensional structure is often simplified
- **Energy Parameterization**  
Energy functions not as sophisticated as we'll see in molecular simulation
- **Alignment Algorithms**  
Dynamic Programming with Frozen Approximation  
Double Dynamic Programming  
Monte Carlo Minimization  
Branch-and-Bound

## Limitations

- Fold Recognition algorithms will return the fold that minimizes the energy function or maximizes the alignment score - but that doesn't mean the identified model is correct.
- Identified model structure is often not as good as in homology modeling

# Ab initio

---

When no structural model can be identified a model must be constructed from first principles

**Molecules obey the laws of physics!**

**Proteins generally adopt the lowest energy conformation**

**Conformation space is finite**

**Proteins fold into a *small* number of protein folds**

## **Current Challenges:**

- Potential functions have limited accuracy

- Conformational search space is huge

*Many methods use reduced representations, simplified potentials, coarse search strategies, simplified solvent models, but...*

# *Ab initio - Molecular Dynamics*

---

When no structural model can be identified a model must be constructed from first principles

**Molecules obey the laws of physics!**

**Proteins generally adopt the lowest energy conformation**

**Conformation space is finite**

**Proteins fold into a *small* number of protein folds**

## **Current Challenges:**

- Potential functions have limited accuracy

- Conformational search space is huge

*Many methods use reduced representations, simplified potentials, coarse search strategies, simplified solvent models, but...*

# Folding@Home

---

When no structural model can be identified a model must be constructed from first principles

**Molecules obey the laws of physics!**

**Proteins generally adopt the lowest energy conformation**

**Conformation space is finite**

**Proteins fold into a *small* number of protein folds**

## **Current Challenges:**

- Potential functions have limited accuracy

- Conformational search space is huge

*Many methods use reduced representations, simplified potentials, coarse search strategies, simplified solvent models, but...*



# Ab initio

---

When no structural model can be identified a model must be constructed from first principles

**Molecules obey the laws of physics!**

**Proteins generally adopt the lowest energy conformation**

**Conformation space is finite**

**Proteins fold into a *small* number of protein folds**

## **Current Challenges:**

- Potential functions have limited accuracy
- Conformational search space is huge

*Many methods use reduced representations, simplified potentials, coarse search strategies, simplified solvent models, but...*

*When a new fold is discovered, it is often composed of common structural motifs at the fragment level*

# Ab initio

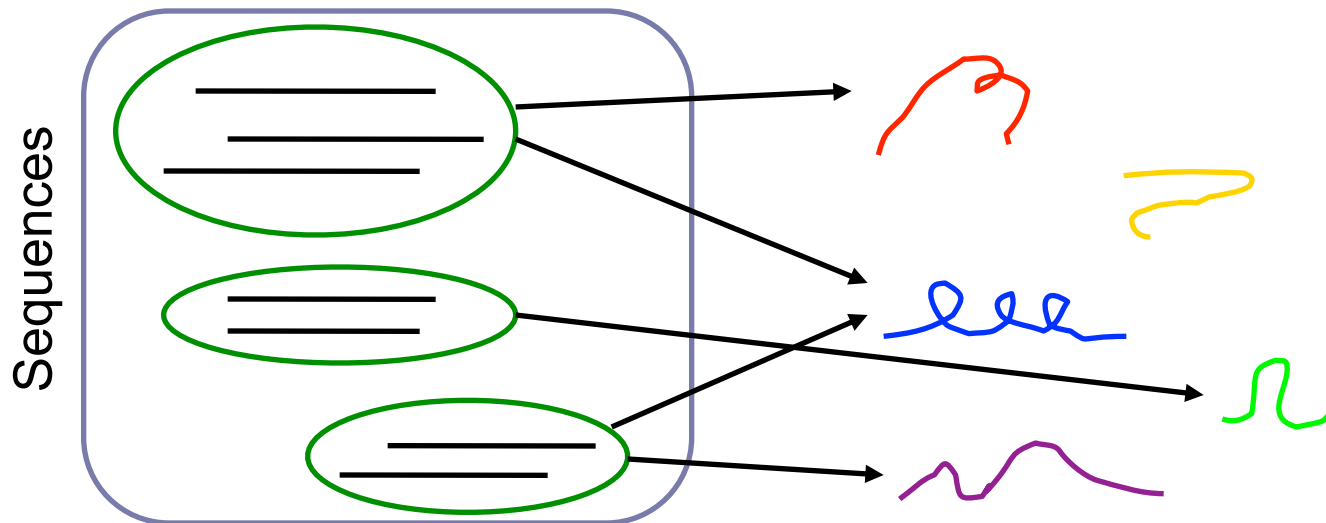
---

## Fragment Based Methods

- Local structure profiles from PDB database

### **Assumption:**

*The distribution of conformations sampled by a short sequence (fragment) is approximated by the distribution of conformations of the same sequence and closely related sequences in the structure database.*



# *Ab initio*

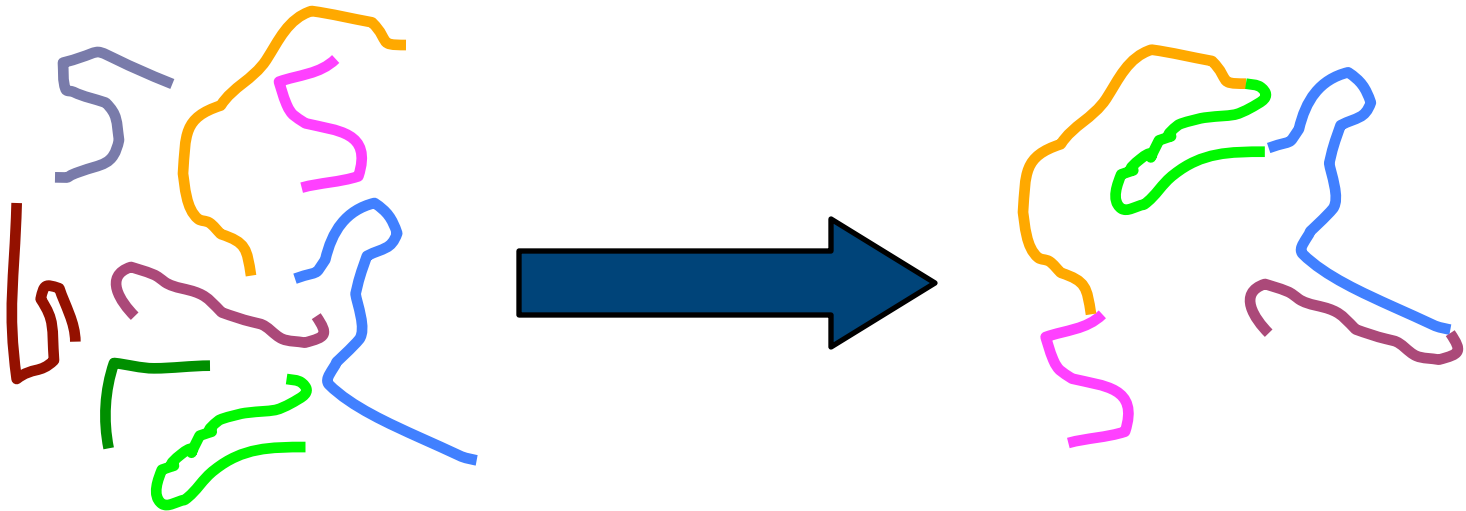
---

## Fragment Based Methods

### Method

- Split the sequence into fragments
- Search the DB for structures having similar sequence
- Use optimization technique to find best combination of fragments

Generate a small alphabet of structural fragments that can be used to construct any protein:



# Ab initio

---

## Rosetta

Performs structure prediction in a probabilistic framework

Fragment library mined from PDB

Two scoring functions

**Coarse-grained**, probabilistic-based scoring function in initial stages

**Fine-grained**, more physically realistic, atomic-level in later stages

Tries to find structure that optimizes:

$$\Pr[M | S] = \frac{\Pr[S | M] \Pr[M]}{\Pr[S]}$$

$M$ : model structure,  $S$ : sequence

Marginal  $P(S)$  assumed 1.

$P(M)$  and  $P(S|M)$  from above scoring functions

# *Ab initio*

---

## Rosetta

### Fragment Selection

- All 9-residue windows in the query are scored against all windows in their non-redundant database of high-resolution structures (<50% sequence identity).
- Sequence profiles for both the query and DB subsequence are generated by PSI-BLAST and compared.
- Predicted secondary-structure of the query is compared with the DSSP computed SS of the known structure.
- A ranked list of top fragments in each sequence window is maintained.
- Up to 200 structural fragments are maintained for each 9-residue window in the query sequence

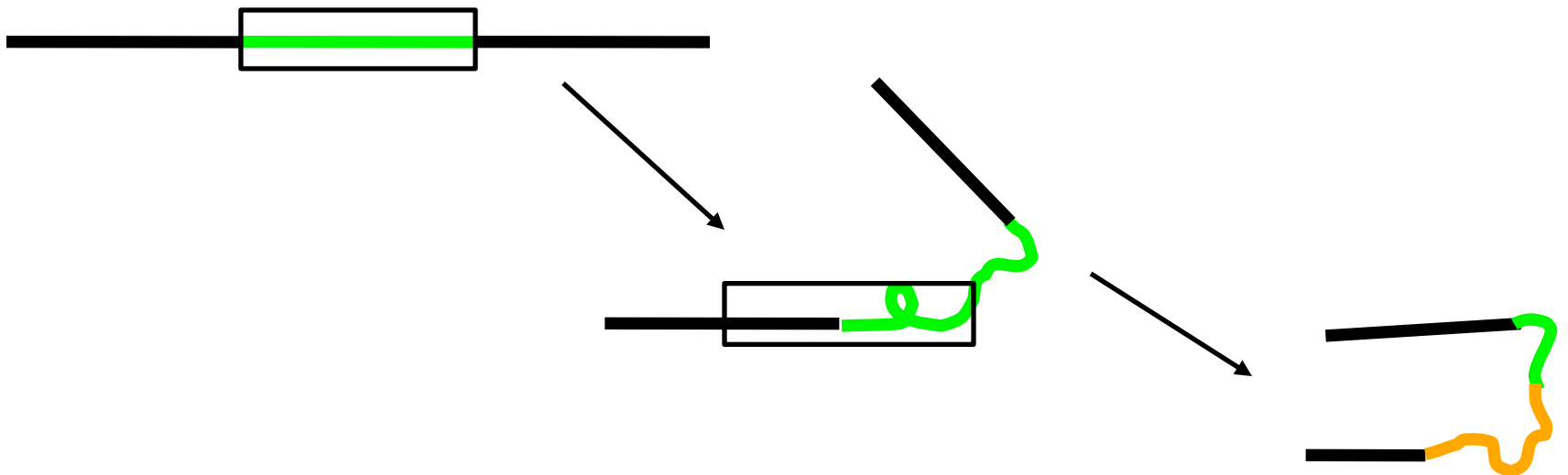
# Ab initio

---

## Rosetta

### Fragment Assembly -- Monte Carlo Search

- Protein starts as extended sequence
- A 9-residue fragment window is chosen at random and is swapped (structurally) with a fragment from the DB
- Maintains an ensemble of conformations with high posterior probabilities



# Ab initio

---

## Rosetta

### Fragment Assembly -- Monte Carlo Search

Repeat  $k$  times - each time starting from extended configuration

Stage of Search	Terms Included in Scoring Function
Initial (until all AA replaced)	steric term only
First 2,000 Steps	steric, helix packing, sheet arrangement, electrostatics, solvation, 30% strand hydrogen bonding
Next 20,000 Steps	sterics, helix packing, sheet arrangement, electrostatics, solvation, strand hydrogen-bonding, 50% side-chain solvation
Last 6,000 Steps	all terms

**Observation:** After initial collapse, any fragment swap is likely to create a clash with neighbouring residues or break favorable contacts. We need for something a bit more gentle.

# *Ab initio*

---

## Rosetta

### Fragment Assembly -- Monte Carlo Search

Generate multiple structures within the *radius-of-convergence* of the native free-energy minima.

#### Fine-Grained Refinement

- Optimization of side-chain rotamers
- Gradient descent minimization after backbone modification
- Random torsion angle refinement
- Rapid torsion angle optimization to offset global backbone perturbations

#### Cluster results:

- Best cluster has greatest number of conformations within 4Å RMSD of the center
- Representative structures taken from each of the best  $k$  clusters and returned to user

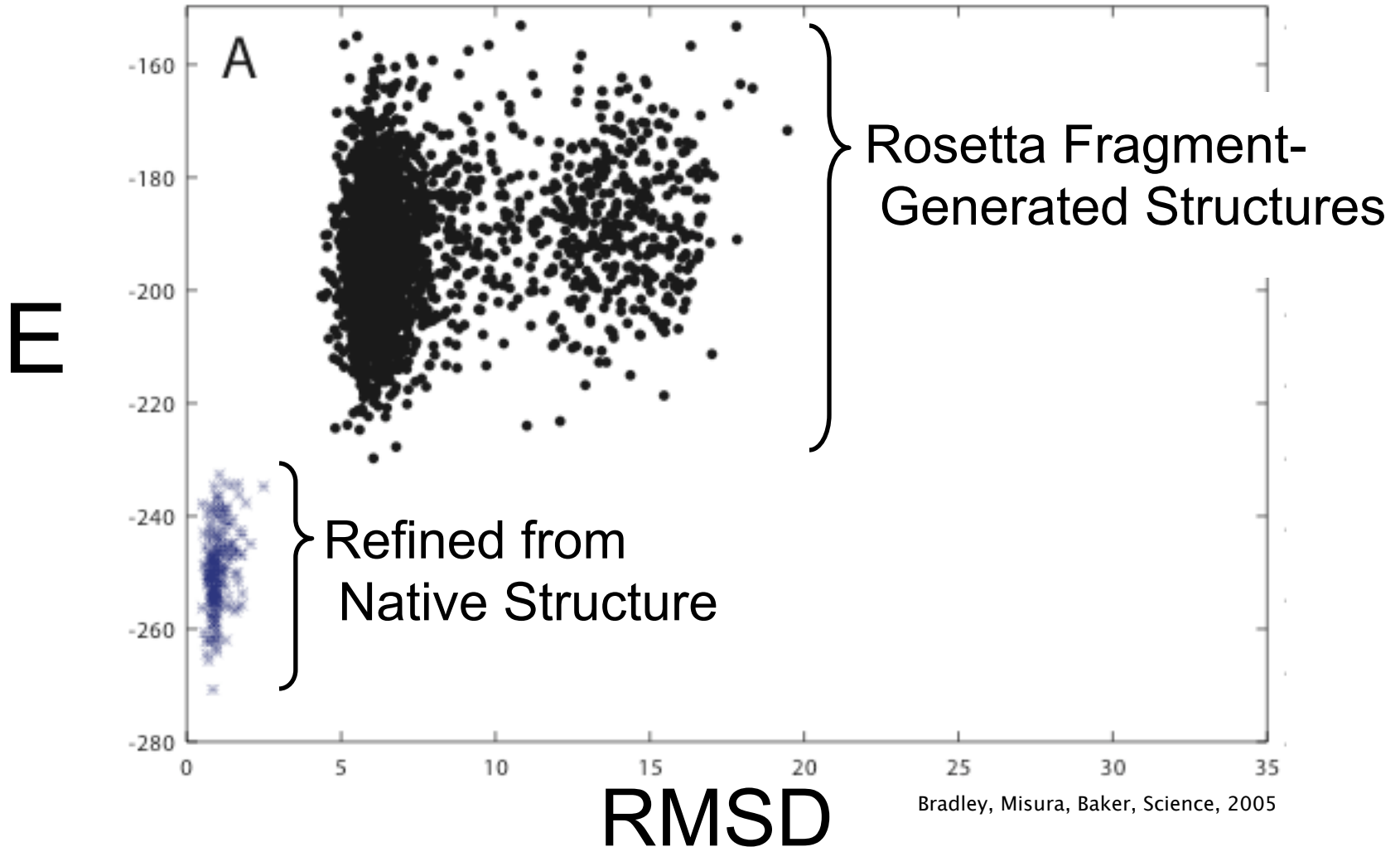


# Rosetta Video

[http://boinc.bakerlab.org/rosetta/rah\\_movies/1ubi\\_copyright.wmv](http://boinc.bakerlab.org/rosetta/rah_movies/1ubi_copyright.wmv)

# *Ab initio*

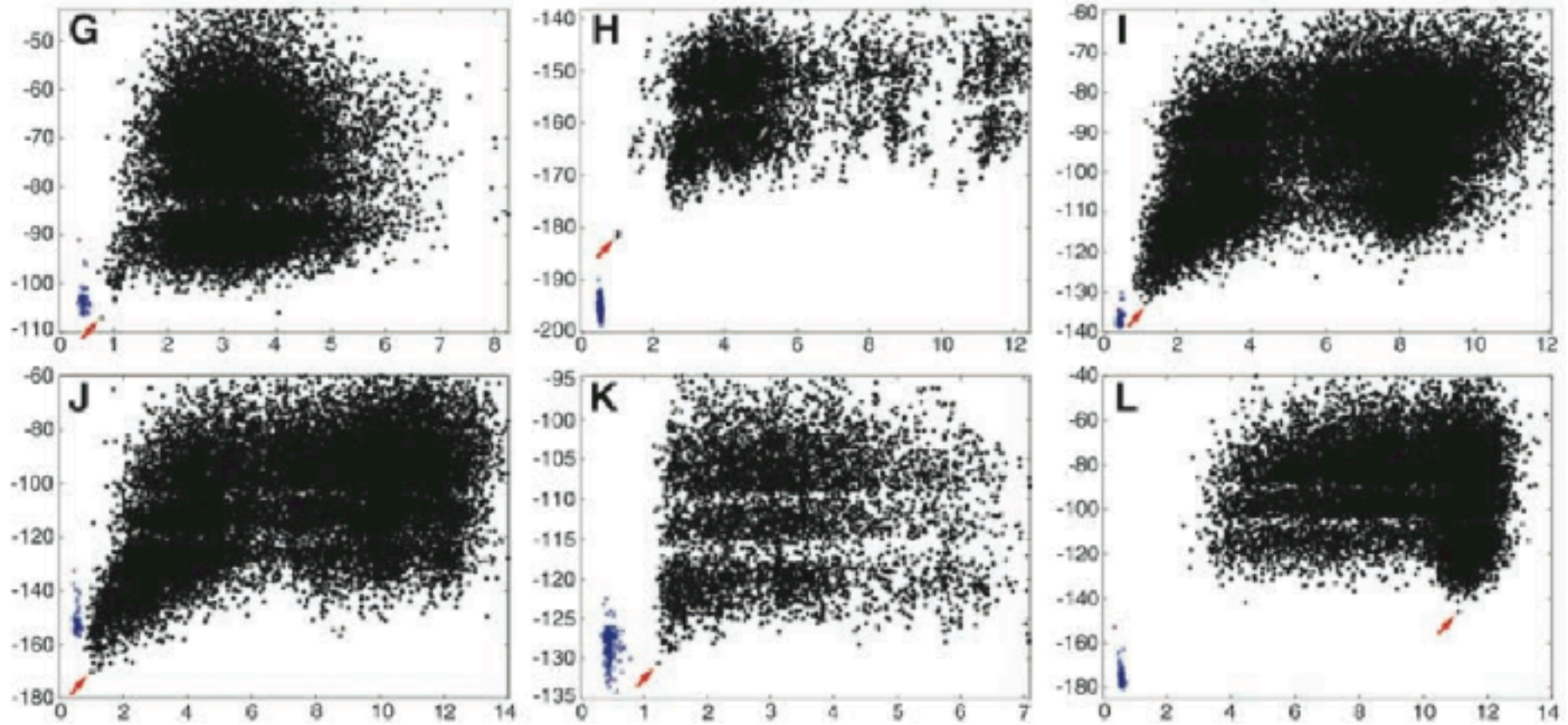
## Rosetta



Bradley, Misura, Baker, Science, 2005

# *Ab initio*

## Rosetta



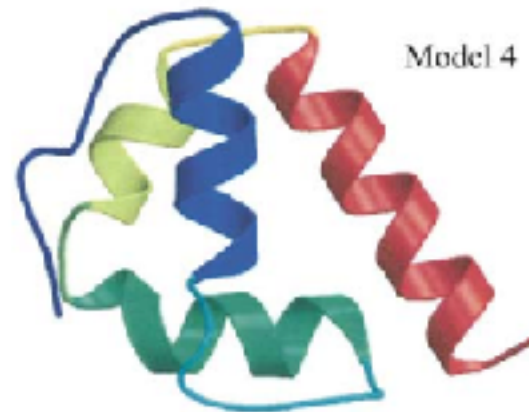
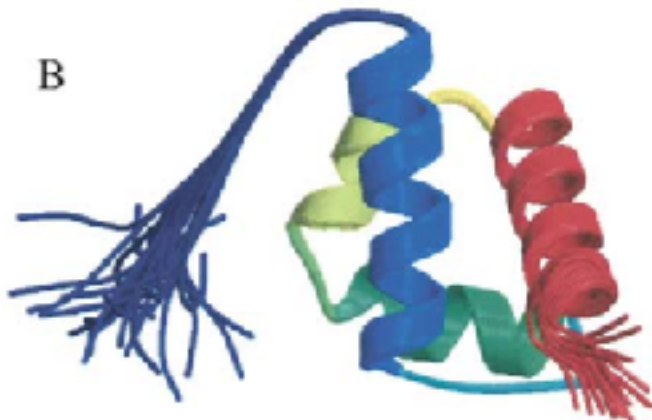
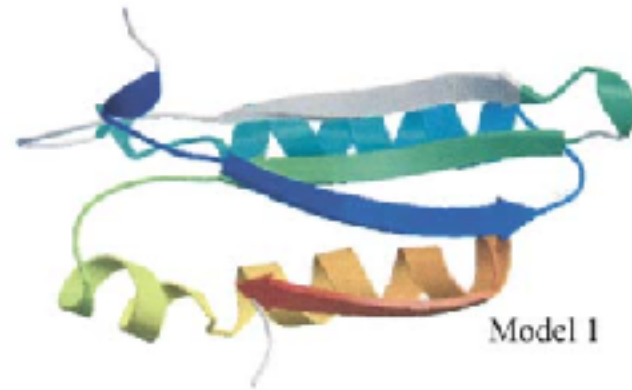
Bradley, Misura, Baker, Science, 2005

# *Ab initio*

## Rosetta - *Not necessarily typical results*

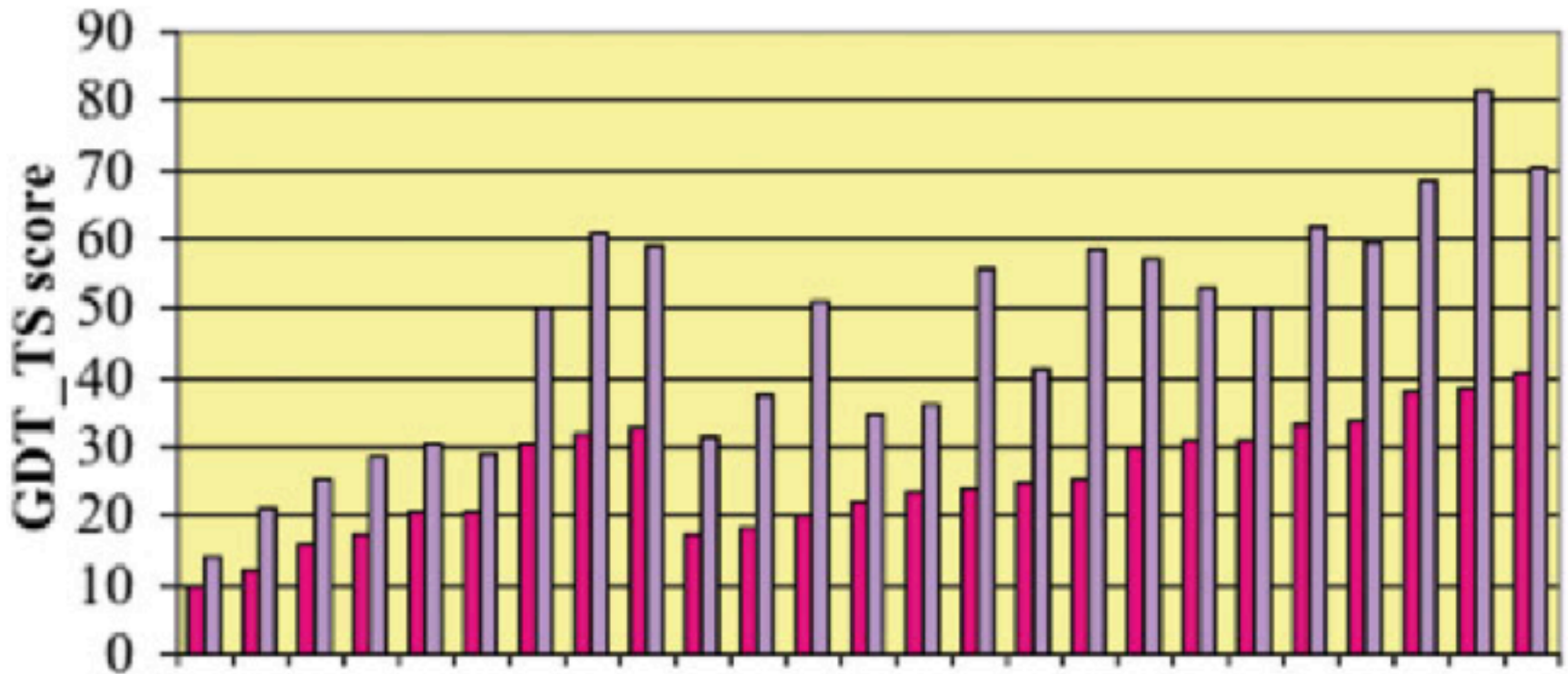
Experimental

Model



## GDT scores for New Fold Targets

■ Average ■ Best



Vincent, Tai, Sathyanarayana, Lee, PROTEINS, S7, 2005

# Target

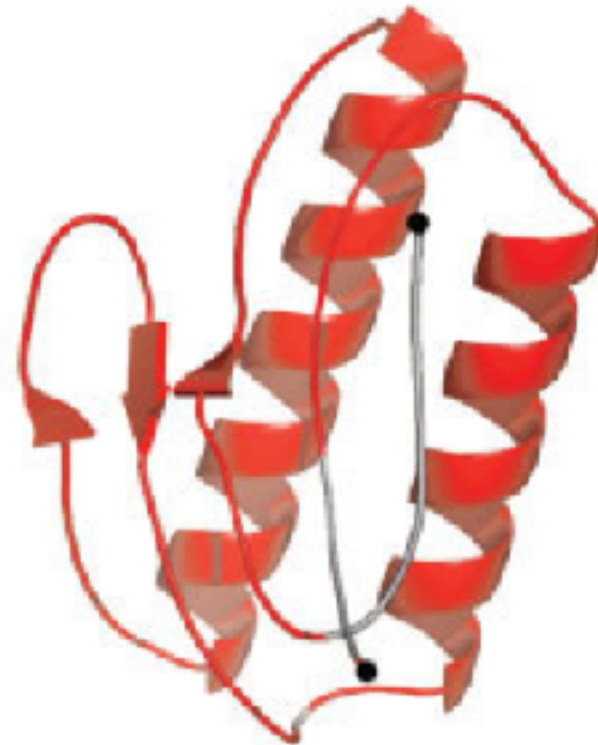
Target ID <sup>a</sup>	Size (AA)	Class <sup>b</sup>	Average <sup>c</sup> GDT_TS	Best <sup>c</sup> GDT_TS
T0216_2	213	NFh	9.9	14.0
T0216_1	213	NFh	12.0	21.3
T0241_1	117	NFh	16.2	25.4
T0241_2	119	NFh	17.3	28.8
T0242	115	NFh	20.5	30.7
T0238	153	NFh	20.8	29.3
T0248_2	87	NFe	30.6	50.0
T0201	90	NFe	32.1	61.2
T0209_2	73	NFe	33.5	59.2
T0209_1	130	FR/A	17.4	31.5
T0273	186	FR/A	18.1	37.5
T0198	221	FR/A	20.2	51.1
T0272_2	122	FR/A	21.9	34.6
T0199_3	82	FR/A	23.7	36.3
T0212	119	FR/A	23.8	55.8
T0239	98	FR/A	25.0	41.3
T0272_1	85	FR/A	25.5	58.5

## Very good prediction

T0201



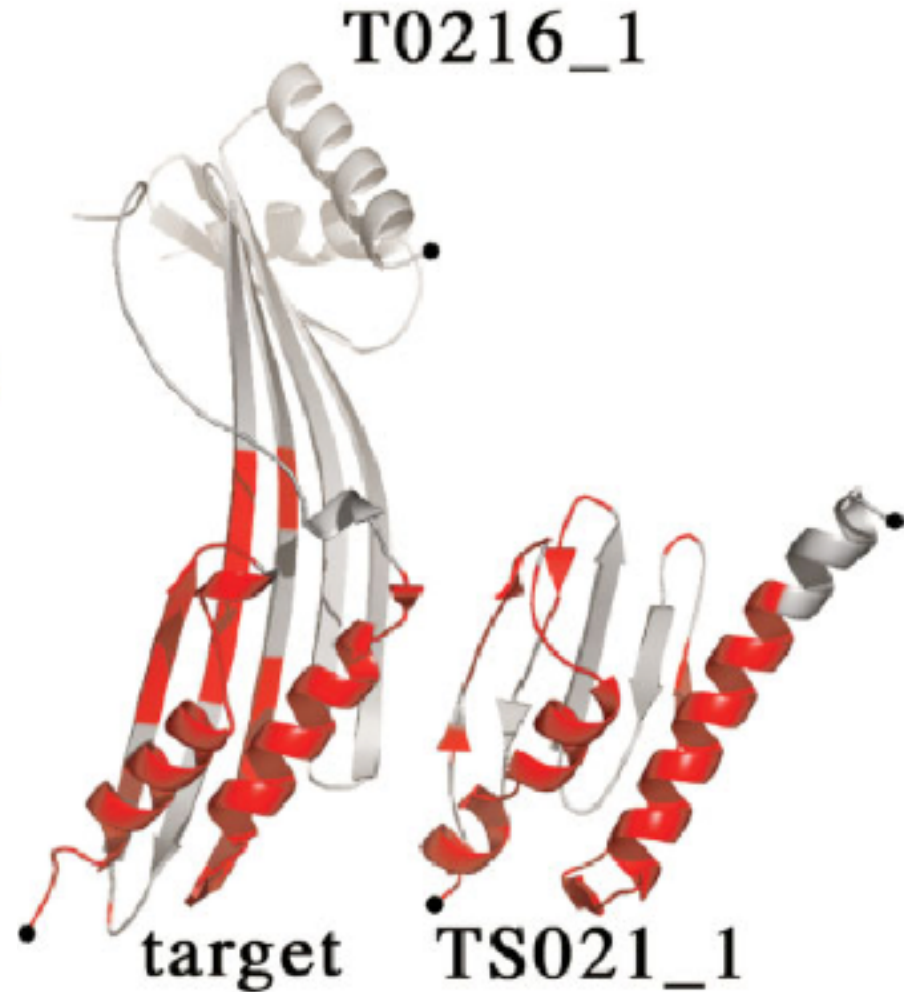
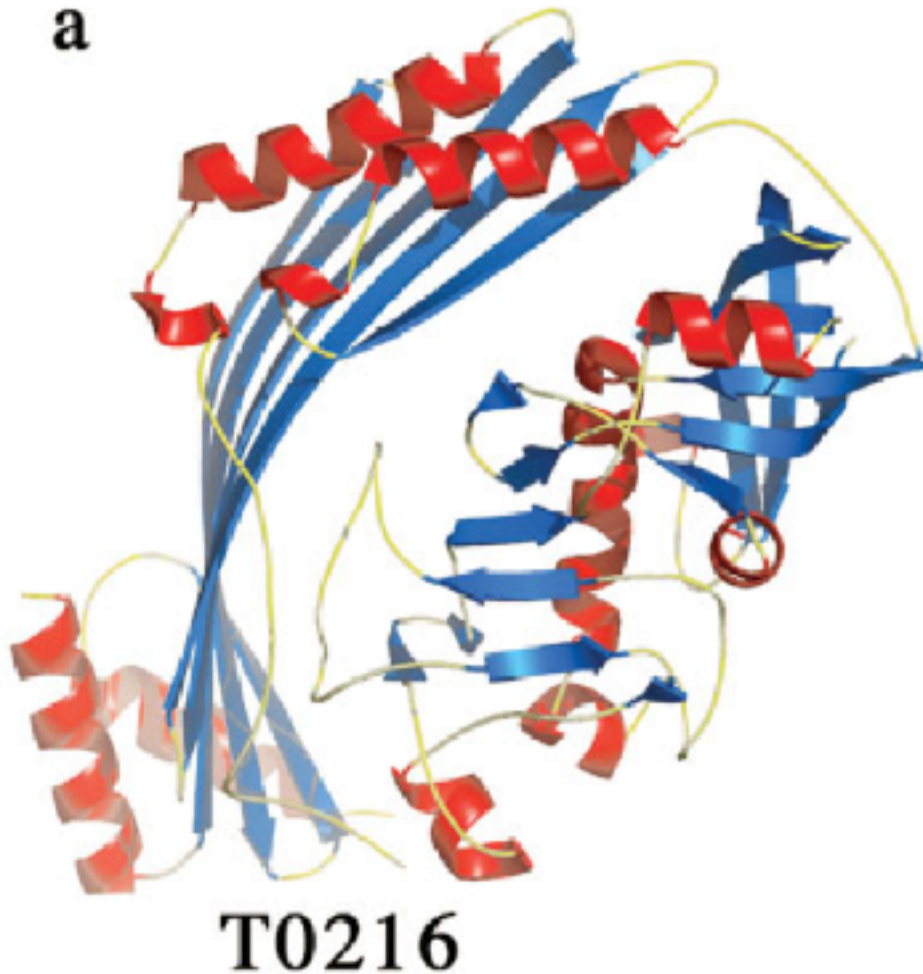
target



TS021\_4

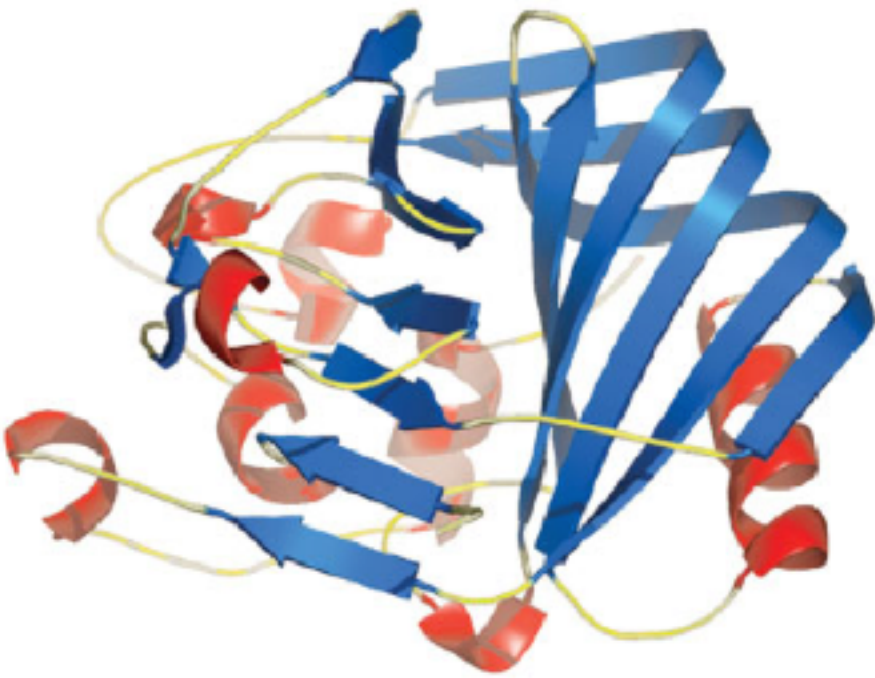
## Poor prediction

a



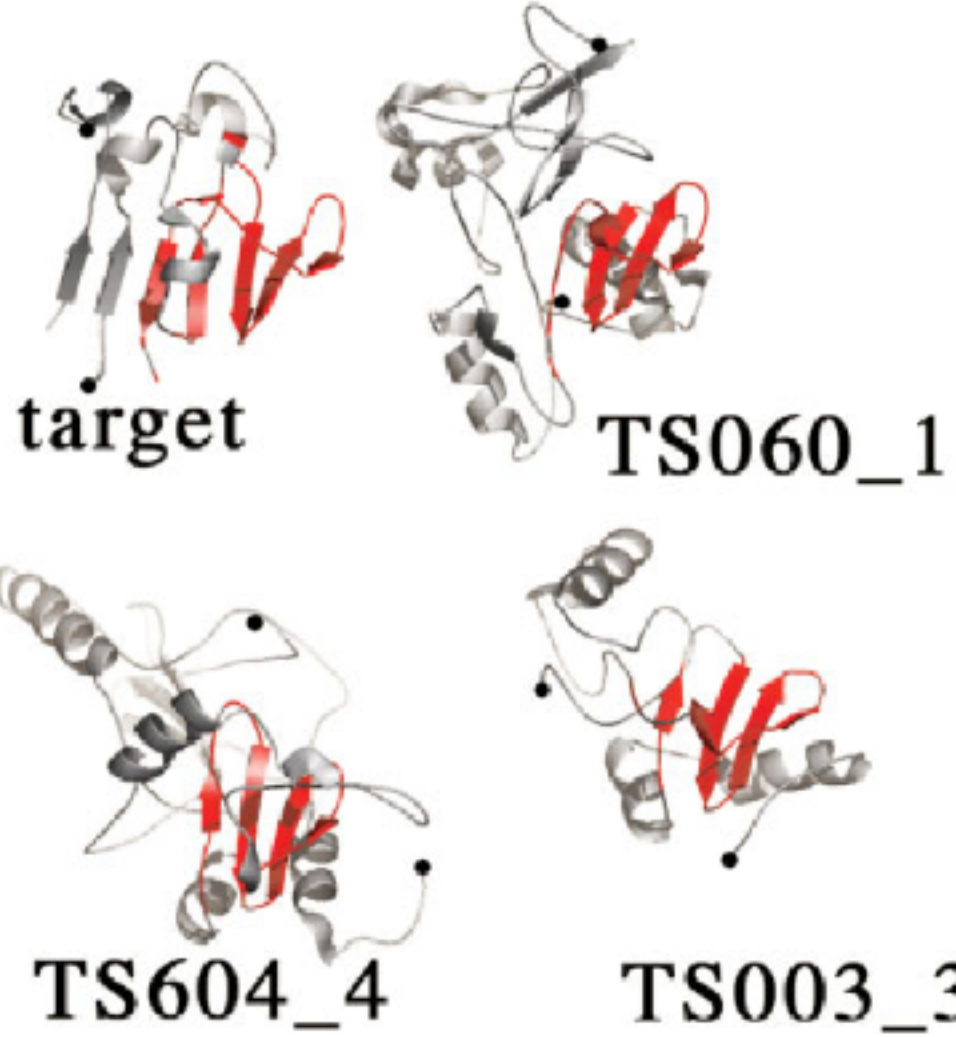


## Poor prediction



T0241

T0241\_1



target

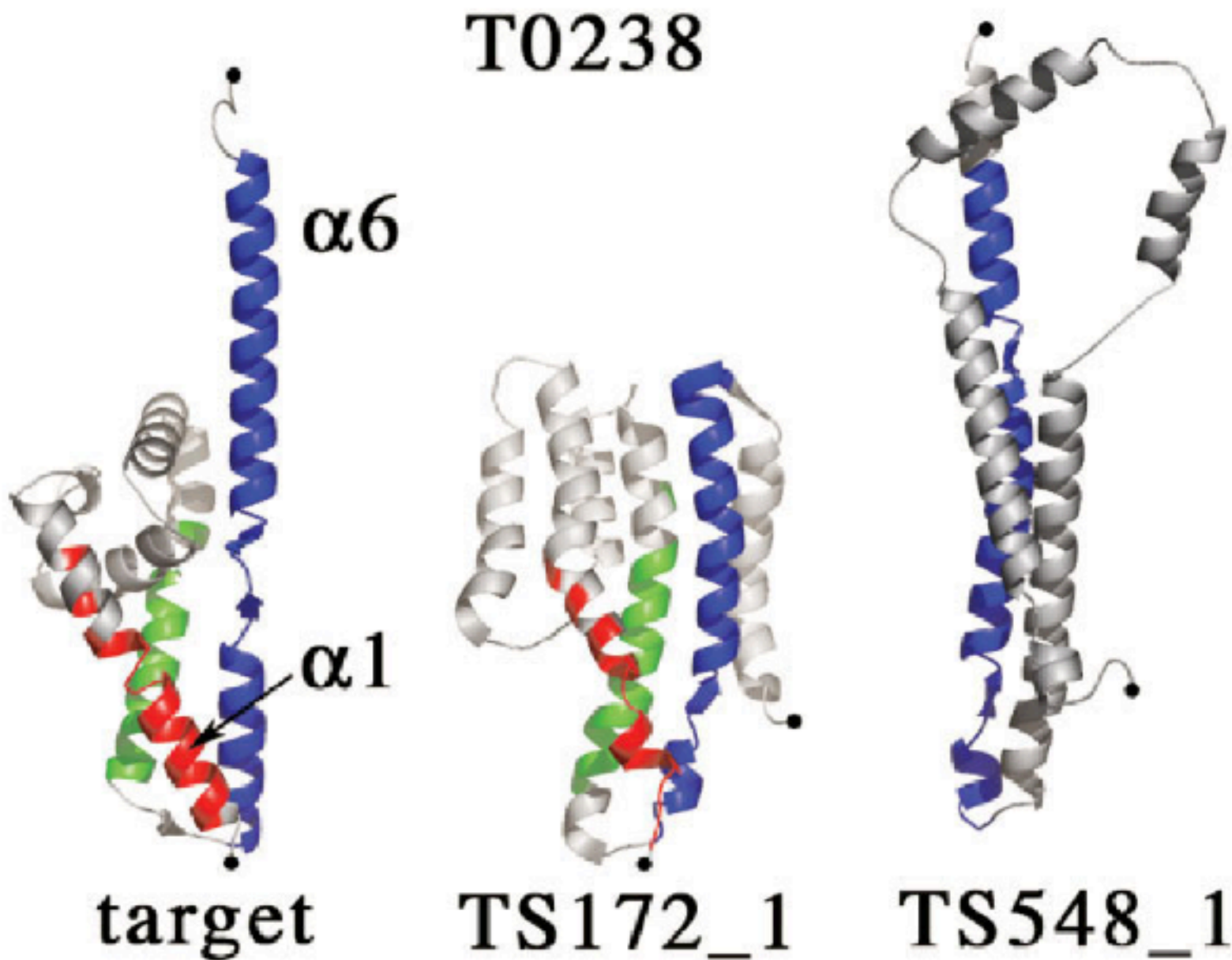
TS060\_1

TS604\_4

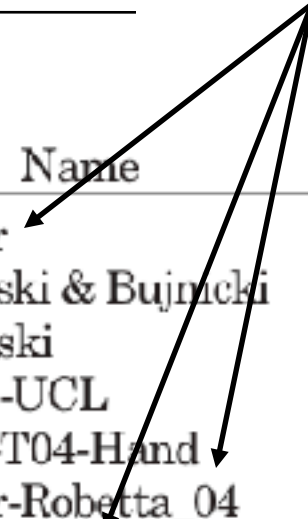
TS003\_3

T0242





## Rosetta



ID	Name	Visually best <sup>a</sup>			Among top five <sup>b</sup>		
		NFh	NFe	FRA	NFh	NFe	FRA
100	Baker	1	0	5	4	2	4
021	Kolinski & Bujnicki	1	2	0	4	2	4
450	Ginalski	0	2	0	1	4	4
003	Jones-UCL	2	0	0	1	2	3
166	SAM-T04-Hand	—	—	—	2	4	0
604	Baker-Robetta_04	1	0	0	2	0	2
101	Baker-Robetta	0	2	0	2	2	0
160	Keasar	—	—	—	0	4	1
052	Rokky	—	—	—	2	0	0
060	Bilab	1	0	0	2	0	0
176	Skolnick-Zhang	—	—	—	0	0	2
501	Mcon	0	1	0	0	0	1
113	Pmodeller5	1	0	0	—	—	—
172	ProteinShop	1	0	0	—	—	—
035	GeneSilico	0	0	1	—	—	—
089	KIAS	0	0	1	—	—	—
157	3D-Jigsaw	0	0	1	—	—	—

# Rosetta @ home

<http://www.rosettaathome.org/>

**Rosetta@home**

Protein Folding, Design, and Docking

[What is Rosetta@home?](#)



HHMI  
HOWARD HUGHES MEDICAL INSTITUTE

UNIVERSITY OF  
WASHINGTON



**Rosetta@home** needs your help to determine the 3-dimensional shapes of proteins in research that may ultimately lead to finding cures for some major human diseases. By running the Rosetta program on your computer while you don't need it you will help us speed up and extend our research in ways we couldn't possibly attempt without your help. You will also be helping our efforts at designing new proteins to fight diseases such as HIV, Malaria, Cancer, and Alzheimer's (See our [Disease Related Research](#) for more information). Please [join us](#) in our effort! **Rosetta@home is not for profit.**

Site search

## Join Rosetta@home

1. [Rules and policies](#)
2. [System requirements](#)
3. [Download, install, and run BOINC](#)  
(enter the project URL: <http://boinc.bakerlab.org/rosetta/>)
4. [A welcome from David Baker](#)

## About

- [10 reasons why users crunch Rosetta@home](#)
- [Quick Guide to Rosetta@home and its Graphics](#)
- [Rosetta@home FAQ](#)
- [Rosetta@home Science FAQ](#)
- [Disease Related Research](#)
- [Research Overview](#)
- [News & Articles about Rosetta](#)
- [David Baker's Rosetta@home Journal](#)
- [Technical news](#)

## Returning participants

- [Your account](#) - view stats, modify preferences
- [Results](#) - view your results
- [Teams](#) - create or join a team
- [Applications](#)
- [Server Status](#)
- [Add-ons](#)
- [How to view your structure predictions](#)

## Community

## User of the day



XENO

I believe: Conservative thinking is destroying the United States.

The Roman Senate ended up being overthrown by Caesar because they...

Feb 04, 2007

**Predictor of the day:** Congratulations to [goingharris](#) (Team [Eastern Michigan University--CompSci](#)) for predicting the lowest energy structure for workunit `ls018_BOINC_ABRELAX_SAVE_ALL_GUT_hum009_1408_0_0 Lem.sc1`

...more

## News

Feb 2, 2007

**Outage Resolved:** The project is back online.

Feb 1, 2007

Our server that runs the validator and a couple assimilators went down last night. We've switched to another server but it may take some time to catch up and grant credit to pending/completed work units. **Outage Notice:** Tomorrow at around 2pm PST, Friday, the project will be down for maintenance for 2-3 hours.

Jan 29, 2007

Rosetta@home has been updated to version 5.45. The new version has some fixes in the graphics to make it significantly more stable and, as a result, sidechains and protein rotation has been turned back on. Some science related updates are also included. For details, see [this thread](#).

## Server Status as of 4 Feb 2007 23:18:02 UTC

[ Schedule running ]    [ Queue: 19,784 ]  
In progress: 228,412  
Successes (last 24h): 167,827  
Users: [↓](#) (last day [↑](#)): 105,882 (1-190)  
Hosts: [↓](#) (last day [↑](#)): 261,570 (1-195)  
Credits (last 24h): 1,800,217  
Total credits: 1,347,642,631  
TaskLOPS (all sites): 88,282

[XML](#) Available as an [RSS feed](#).

# Ab initio

---

- Clearly a more challenging problem
- Predictive accuracy is not as good as other methods
- 5 - 20 Å RMSD, Not sufficient for SBDD, but:
  - May recognize more distant structural homologues
  - May recognize structural or functional motifs
  - May be useful to assist in experimental structure determination
  - May benefit from more deterministic search

## Why?

- Quality of Search, Quality of Scoring Function
- Is the folding process important? Are we simulating folding?
- MC search in ROSETTA may have hard time climbing hills
  - Low energy states that can only be achieved by multiple uphill moves are not likely to be achieved.

Do we consider the *correct* regions of conformation space?  
Can we recognize the correct conformation when we see it?