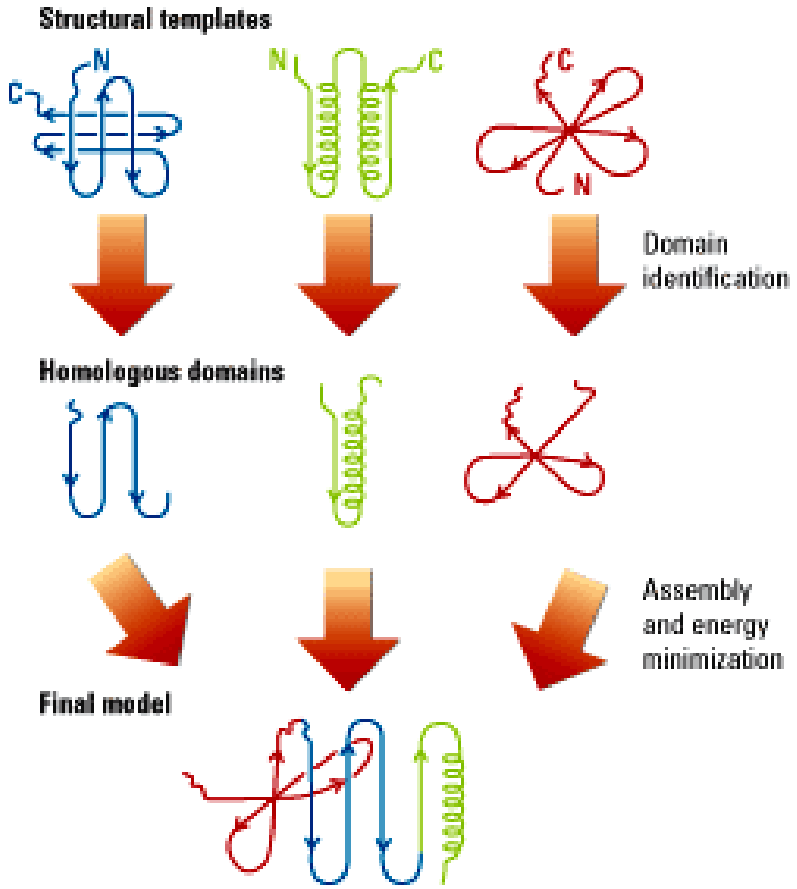
EECS730: Introduction to Bioinformatics

Lecture 13: Protein threading



https://pubs.acs.org/subscribe/archive/mdd/v03/i09/figures/willis-rosetta.gif

Some slides were adapted from Dr. Dong Xu (University of Missouri Columbia)

Protein 3D structure determination (experimentally)

Structure:

Traditional experimental methods:

X-Ray or NMR to solve structures;

generate a few structures per day worldwide

cannot keep pace for new protein sequences

Strong demand for structure prediction:

more than 30,000 human genes;

sequencing a genome becomes routine nowadays.

Unsolved problem after efforts of two decades.

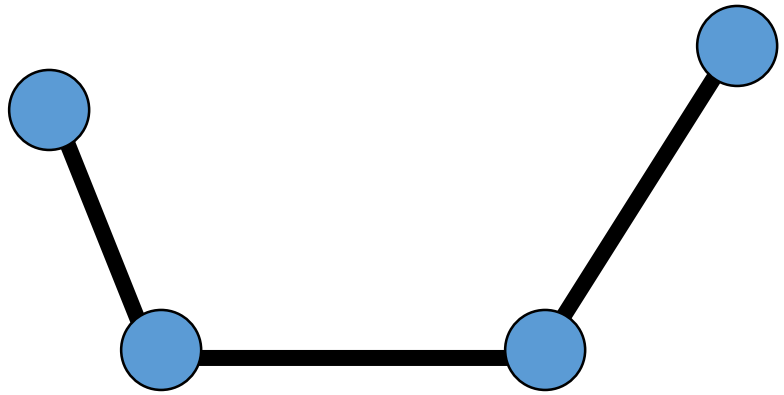
Can we predict protein tertiary structure from sequence?

- Identify distant homologues of protein families
- Predict function of protein with low degree of sequence similarity with other proteins

Ab initio folding

➢**An energy function to describe the protein**

- bond energy
- bond angle energy
- dihedral angel energy
- van der Waals energy
- electrostatic energy



➢**Minimize the function and obtain the structure.**

➢**Not practical in general**

- Computationally too expensive
- Accuracy is poor

Homology modeling

- Sequence is aligned with sequence of known structure, usually sharing sequence identity of 30% or more.
- Superimpose sequence onto the template, replacing equivalent sidechain atoms where necessary.
- Refine the model by minimizing an energy function
- Only applicable when we know the structure of its homolog

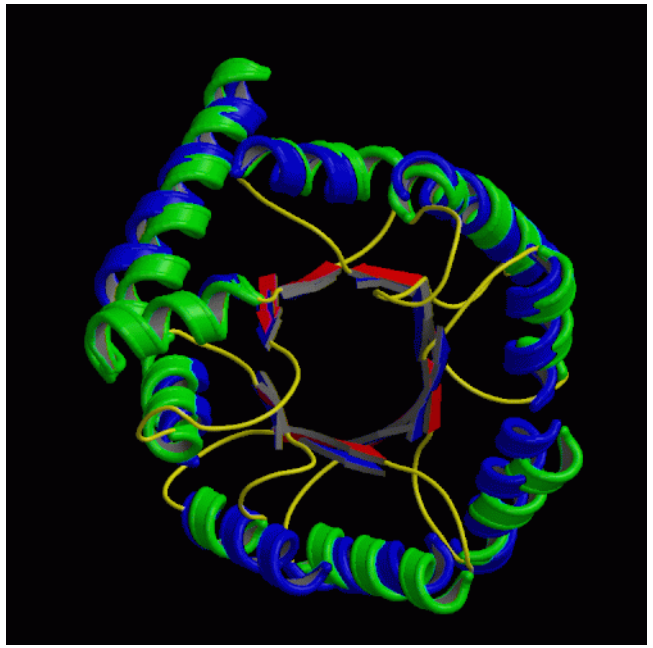
Template-based methods

Structure is better conserved than sequence

Structure can adopt a wide range of mutations.

Physical forces favor certain structures.

Number of fold is limited.
- Currently ~700
- Total: 1,000 ~10,000



TIM barrel

Protein threading

- The number of different folds in nature is fairly small (approximately 1,300).
- 90% of the new structures submitted to the PDB in the past three years have similar structural folds to ones already in the PDB.
- No homology is necessary, indicates the conservation of local structure
- general applicability of template-based modeling methods for structure prediction (currently 60-70% of new proteins, and this number is growing as more structures being solved)
- NIH *Structural Genomics Initiative* plans to experimentally solve ~10,000 “unique” structures and predict the rest using computational methods

Major idea of threading

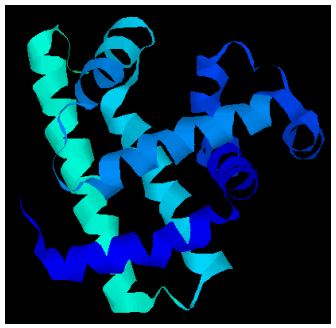
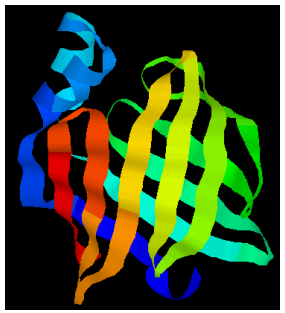
structure prediction through recognizing native-like fold

- o Thread (*align* or *place)* a query protein sequence onto a template structure in "optimal" way
- o Good alignment gives approximate backbone structure

Query sequence

MTYKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEWTYTE

Template set

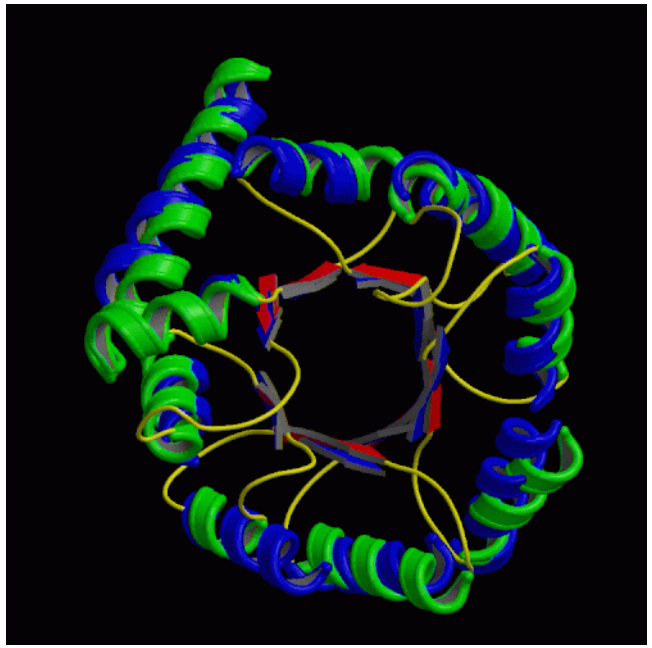
Prediction accuracy: fold recognition / alignment

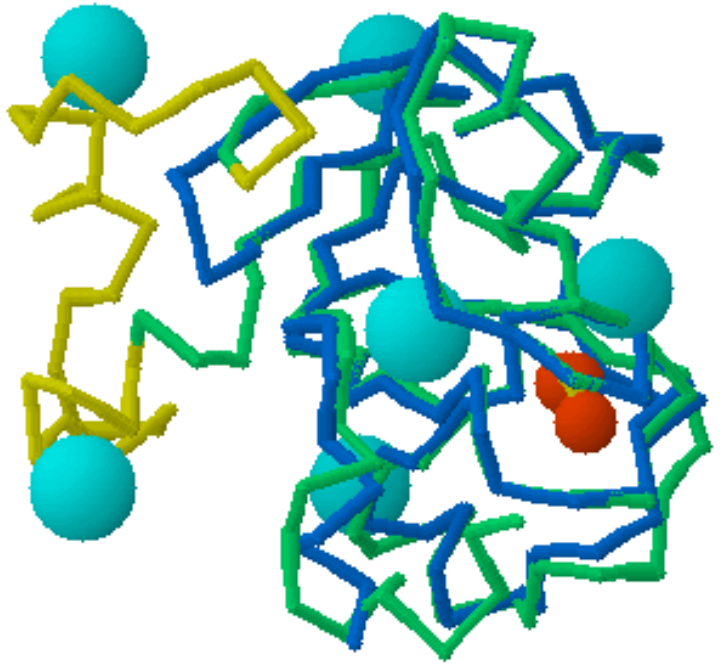
Major components of threading

- Template library
- Scoring function
- Alignment
- Confidence assessment

Template and fold

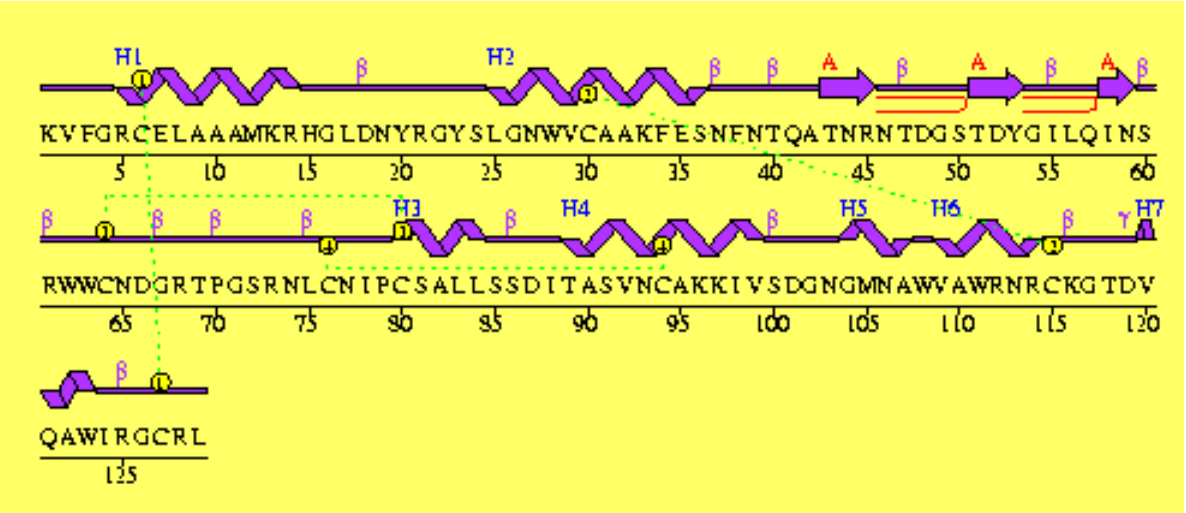
Non-redundant representatives through structure-structure comparison

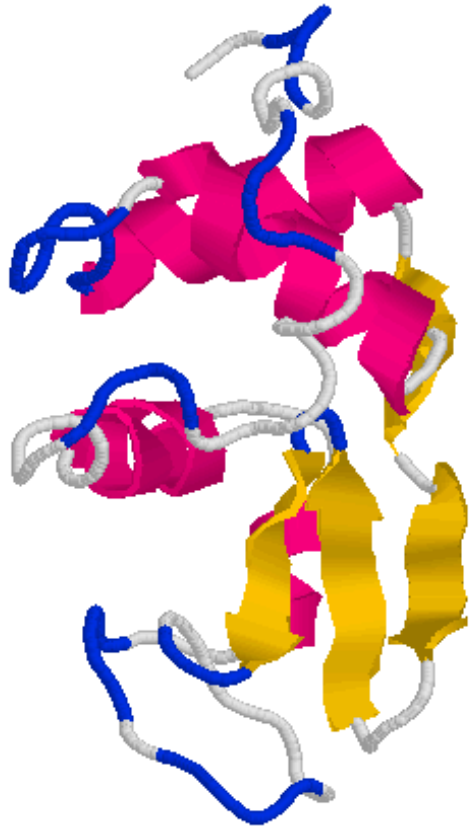




Secondary structures and their arrangement

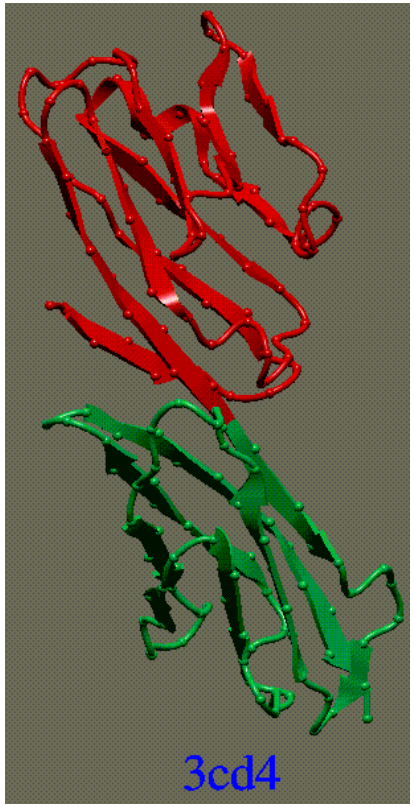
Core of a template



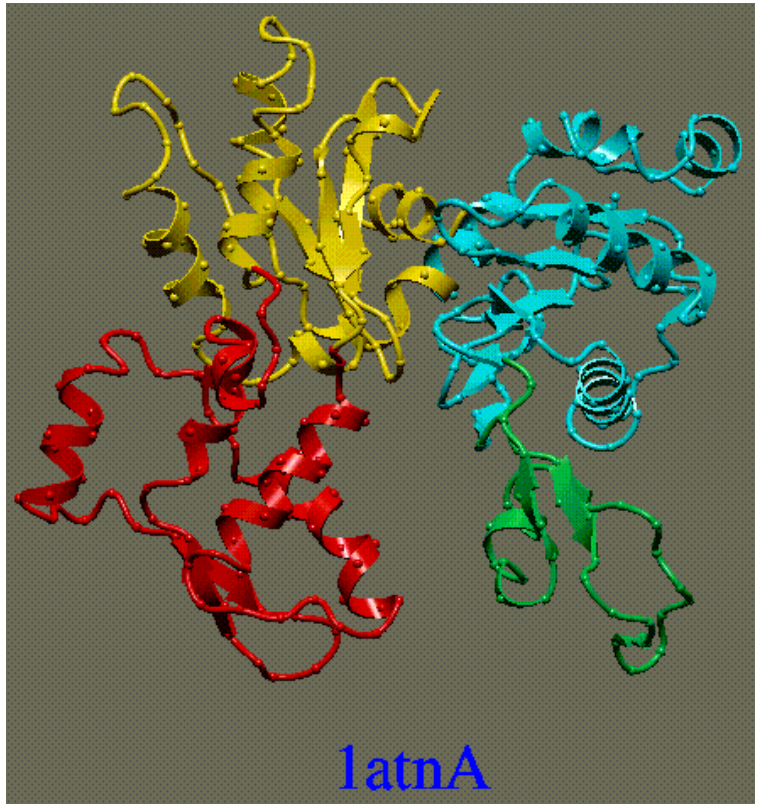


Core secondary structures:
α-helices and β-strands

Chain/domain library



glycoprotein



actin

Domain may be more sensitive but depends on correct partition

Available library databases

- SCOP: http://scop.mrc-lmb.cam.ac.uk/scop/
  (domains, good annotation)
- CATH: http://www.biochem.ucl.ac.uk/bsm/cath/
- CE: http://cl.sdsc.edu/ce.html
- Dali Domain Dictionary: http://columba.ebi.ac.uk:8765/holm/ddd2.cgi
- FSSP: http://www2.ebi.ac.uk/dali/fssp/
  (chains, updated weekly)
- HOMSTRAD:
- http://www-cryst.bioc.cam.ac.uk/~homstrad/
- HSSP: http://swift.embl-heidelberg.de/hssp/

Properties of template

- **Residue type / profile**
- **Secondary structure type**
- **Solvent assessibility**
- **Coordinates for Cα / Cβ**

```
RES 1 G  156  S  23   10.528  -13.223    9.932   11.977  -12.741   10.115
RES 5 P  157  H 110   12.622  -17.353   10.577   12.981  -16.146   11.485
RES 5 G  158  H  61   17.186  -15.086    9.205   16.601  -15.457   10.578
RES 5 Y  159  H  91   16.174  -10.939   12.208   16.612  -12.343   12.727
RES 5 C  160  H   8   12.670  -12.752   15.349   14.163  -13.137   15.545
RES 1 G  161  S  14   15.263  -17.741   14.529   15.022  -16.815   15.733
```

Scoring Function
(similarity between a sequence and a template)

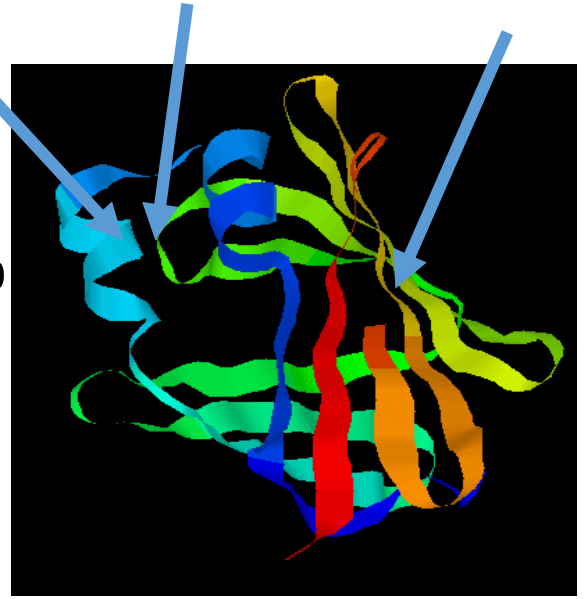
- **Physical energy function: too sensitive**
  - bond energy
  - van der Waals energy
  - electrostatic energy…

- **Knowledge-based scoring function**
  (derived from known sequence/structure)

- **Two types of functions correlate each other**

Scoring Function

…YKLILNGKTKGETTTEAVDAATAEKVFQYANDNGVDGEW…



How preferable to put two particular residues nearby: E_p (pairwise term)

Alignment gap penalty: E_g

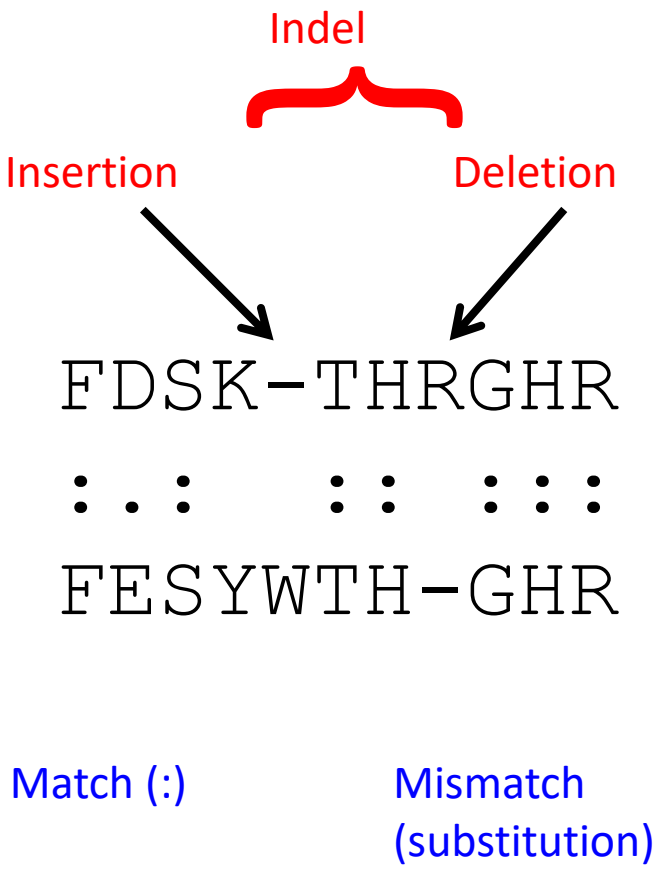
How well a residue align to another residue on sequence: E_m (mutation term)

How well a residue fits a structural environment: E_s (singleton term)

Total energy: E_m + E_p + E_s + E_g

Describe how sequence fit template

Sequence conservation



- Close homolog: high cutoffs for BLOSUM (up to BLOSUM 90) or lower PAM values
  BLAST default: BLOSUM 62
- Remote homolog: lower cutoffs for BLOSUM (down to BLOSUM 10) or high PAM values (PAM 200 or PAM 250)
  A threading best performer: PAM 250

Structure-based score

- Structure provides additional (independent) information
- Free energy (score) and distribution in thermal equilibrium (known protein structures)
- Preference model of characteristics
- Derive parameters for structure-based score using a non-redundant protein structure database (FSSP)

Singleton scores

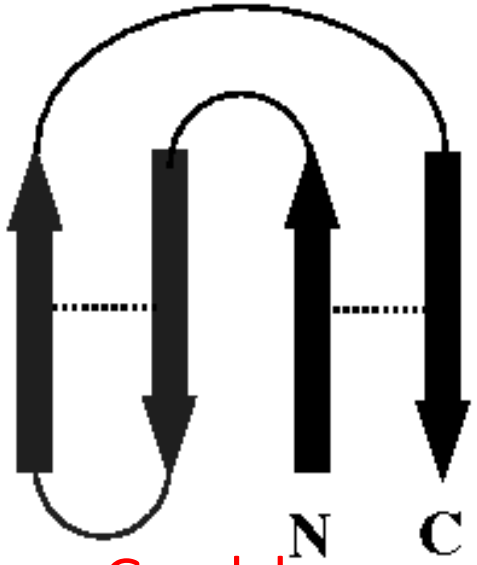
- A single residue's preference in a specific structural environments.
  - secondary structure
  - solvent accessibility
- Compare actual occurrence against its "expected value" by chance

$$e_{single}(i, ss, sol) = -k_B T \log \frac{N(i, ss, sol)}{< N(i, ss, sol) >}$$

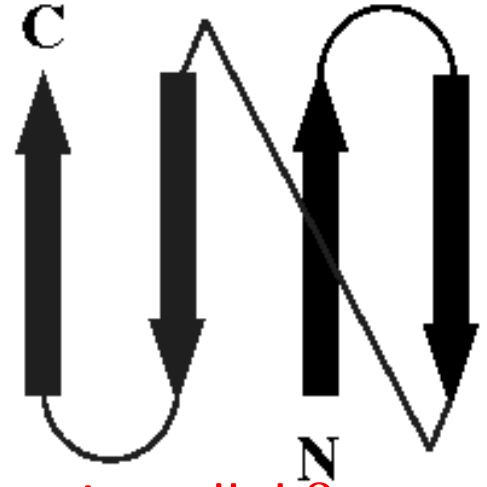
$$< N(i, ss, sol) > = \frac{N(i)\, N(ss)\, N(sol)}{N^2}$$

Pairwise Energy

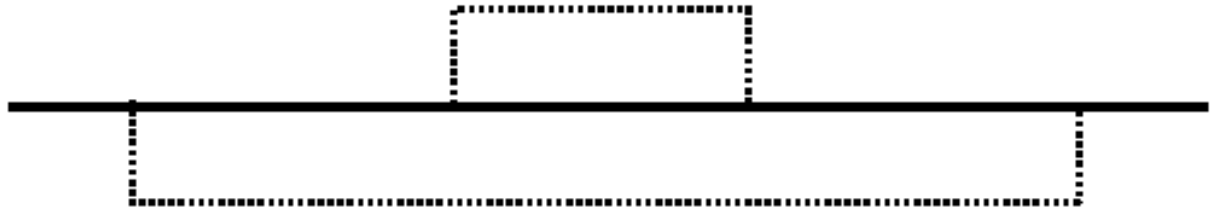
- More reliable than single amino acid's preference
- Use probabilities of the three secondary structure states (α-helices, β-strand, and loop)
- May have a risk of over-dependence on secondary structure prediction



Greek key



4-antiparallel β-strand



Pairwise energy for fold differentiation

Singleton score matrix

	Helix			Sheet			Loop		
	Buried	Inter	Exposed	Buried	Inter	Exposed	Buried	Inter	Exposed
ALA	-0.578	-0.119	-0.160	0.010	0.583	0.921	0.023	0.218	0.368
ARG	0.997	-0.507	-0.488	1.267	-0.345	-0.580	0.930	-0.005	-0.032
ASN	0.819	0.090	-0.007	0.844	0.221	0.046	0.030	-0.322	-0.487
ASP	1.050	0.172	-0.426	1.145	0.322	0.061	0.308	-0.224	-0.541
CYS	-0.360	0.333	1.831	-0.671	0.003	1.216	-0.690	-0.225	1.216
GLN	1.047	-0.294	-0.939	1.452	0.139	-0.555	1.326	0.486	-0.244
GLU	0.670	-0.313	-0.721	0.999	0.031	-0.494	0.845	0.248	-0.144
GLY	0.414	0.932	0.969	0.177	0.565	0.989	-0.562	-0.299	-0.601
HIS	0.479	-0.223	0.136	0.306	-0.343	-0.014	0.019	-0.285	0.051
ILE	-0.551	0.087	1.248	-0.875	-0.182	0.500	-0.166	0.384	1.336
LEU	-0.744	-0.218	0.940	-0.411	0.179	0.900	-0.205	0.169	1.217
LYS	1.863	-0.045	-0.865	2.109	-0.017	-0.901	1.925	0.474	-0.498
MET	-0.641	-0.183	0.779	-0.269	0.197	0.658	-0.228	0.113	0.714
PHE	-0.491	0.057	1.364	-0.649	-0.200	0.776	-0.375	-0.001	1.251
PRO	1.090	0.705	0.236	1.249	0.695	0.145	-0.412	-0.491	-0.641
SER	0.350	0.260	-0.020	0.303	0.058	-0.075	-0.173	-0.210	-0.228
THR	0.291	0.215	0.304	0.156	-0.382	-0.584	-0.012	-0.103	-0.125
TRP	-0.379	-0.363	1.178	-0.270	-0.477	0.682	-0.220	-0.099	1.267
TYR	-0.111	-0.292	0.942	-0.267	-0.691	0.292	-0.015	-0.176	0.946
VAL	-0.374	0.236	1.144	-0.912	-0.334	0.089	-0.030	0.309	0.998

Amino acids side chain properties

Neutral Hydrophobic
Alanine
Valine
Leucine
Isoleucine
Proline
Tryptophane
Phenylalanine
Methionine

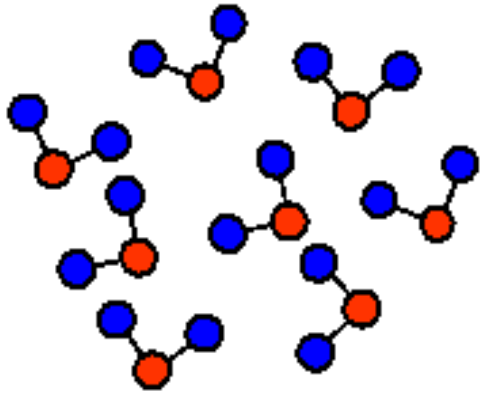
Neutral Polar
Glycine
Serine
Threonine
Tyrosine
Cysteine
Asparagine
Glutamine

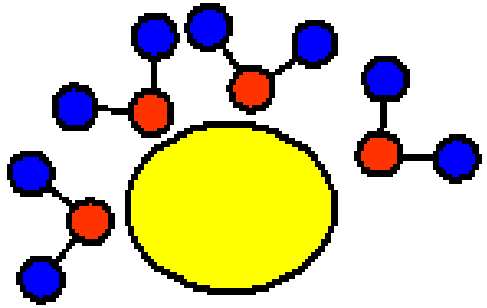
Acidic
Aspartic Acid
Glutamic Acid

Basic
Lysine
Arginine
(Histidine)

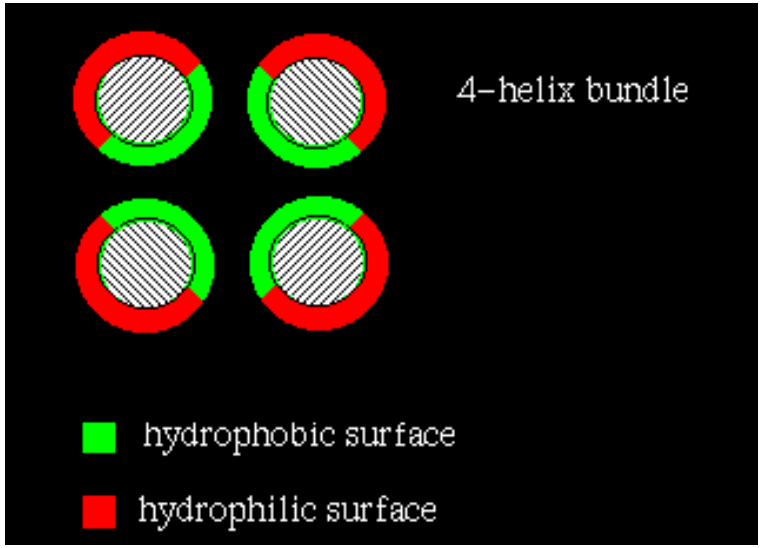
Hydrophobic Effects: Main Driving Force for Protein Folding

Water molecules in bulk water are mobile and can form H-bonds in all directions.





Hydrophobic surfaces don't form H-bonds. The surrounding water molecules have to orient and become more ordered.



Pairwise score

- Preference for a pair of amino acids to be close in 3D space.
- How close is close?
  - Distance dependence
  - 7-8 Armstrong between C_β
- Observed occurrence of a pair compared with it “expected” occurrence

$$e_{pair}(i,j) = -k_B T \log \frac{M(i,j)}{< M(i,j) >}$$

$$< M(i,j) > = \frac{M(i)\, M(j)}{M}$$

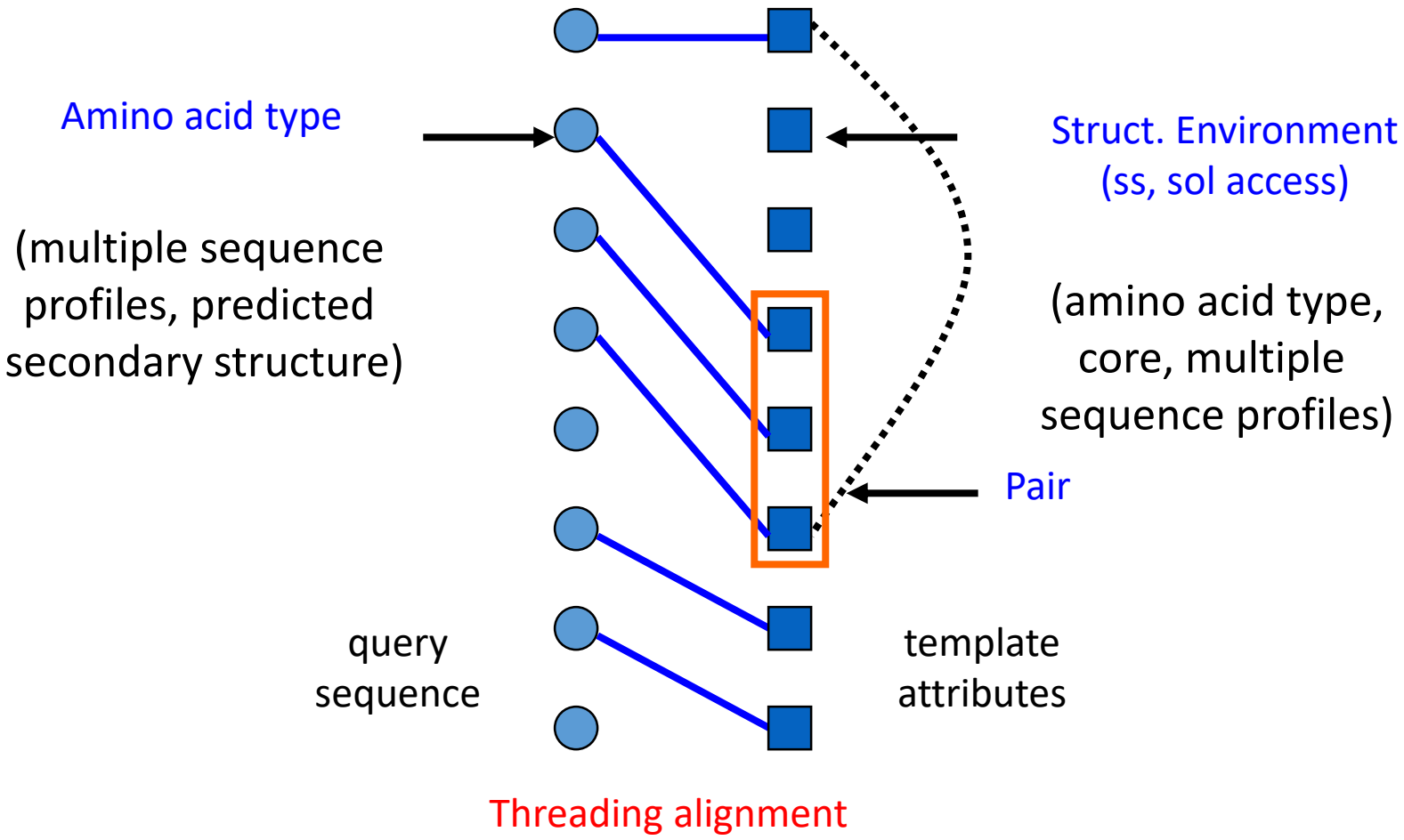
Pairwise score parameters

pairwise potential in unit of 0.001

distance cutoff used -- 7A

ALA	-140																		
ARG	268	-18																	
ASN	105	-85	-435																
ASP	217	-616	-417	17															
CYS	330	67	106	278	-1923														
GLN	27	-60	-200	67	191	-115													
GLU	122	-564	-136	140	122	10	68												
GLY	11	-80	-103	-267	88	-72	-31	-288											
HIS	58	-263	61	-454	190	272	-368	74	-448										
ILE	-114	110	351	318	154	243	294	179	294	-326									
LEU	-182	263	358	370	238	25	255	237	200	-160	-278								
LYS	123	310	-201	-564	246	-184	-667	95	54	194	178	122							
MET	-74	304	314	211	50	32	141	13	-7	-12	-106	301	-494						
PHE	-65	62	201	284	34	72	235	114	158	-96	-195	-17	-272	-206					
PRO	174	-33	-212	-28	105	-81	-102	-73	-65	369	218	-46	35	-21	-210				
SER	169	-80	-223	-299	7	-163	-212	-186	-133	206	272	-58	193	114	-162	-177			
THR	58	60	-231	-203	372	-151	-211	-73	-239	109	225	-16	158	283	-98	-215	-210		
TRP	51	-150	-18	104	52	-12	157	-69	-212	-18	81	29	-5	31	-432	129	95	-20	
TYR	53	-132	53	268	62	-90	269	58	34	-163	-93	-312	-173	-5	-81	104	163	-95	-6
VAL	-105	171	298	431	196	180	235	202	204	-232	-218	269	-50	-42	46	267	73	101	107	-324
	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL

Formulation of the threading problem



Mathematical formulation of threading problem

For a sequence $S = S_1.....S_n$ and a template $T = T_1.....T_m$, find an alignment $(\overline{S}, \overline{T})$:

$$\begin{array}{cccccc} \overline{S}_1 & \overline{S}_2 & ..... & \overline{S}_p \\ | & | & & | \\ \overline{T}_1 & \overline{T}_2 & ..... & \overline{T}_p, \end{array}$$

to minimize an energy function

$$E_{total}(\overline{S}, \overline{T}) = \sum_i E_{singleton}(\overline{S}_i, \overline{T}_i) + \sum_{(\overline{T}_i, \overline{T}_j) \in \mathbf{PAIRS}} E_{pair}(\overline{S}_i, \overline{S}_j, \overline{T}_i, \overline{T}_j)$$

where PAIRS is the set of interacting pairs, $\max\{n, m\} \leq p \leq n+m$.

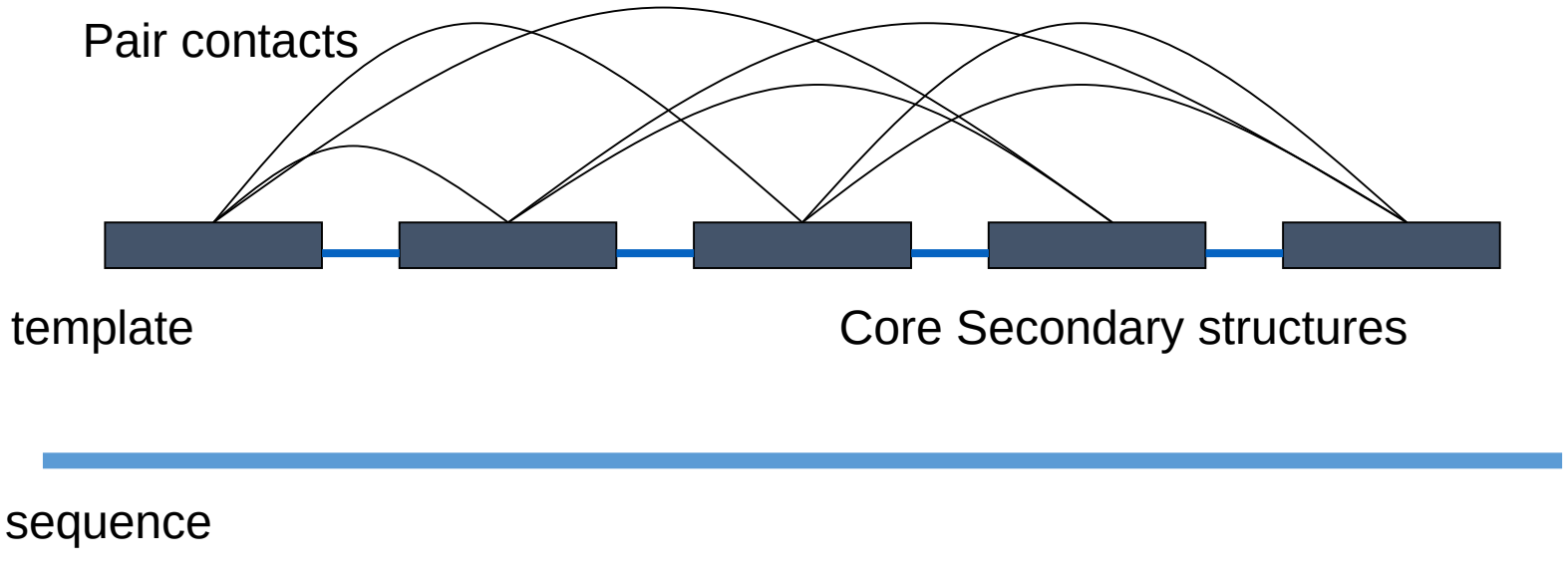
Compare the minimum energy between different templates:

$$E_{total}(S, T) = \min_{(\overline{S}, \overline{T})} E_{total}(\overline{S}, \overline{T}).$$

PROSPECT algorithm summary

Formulation

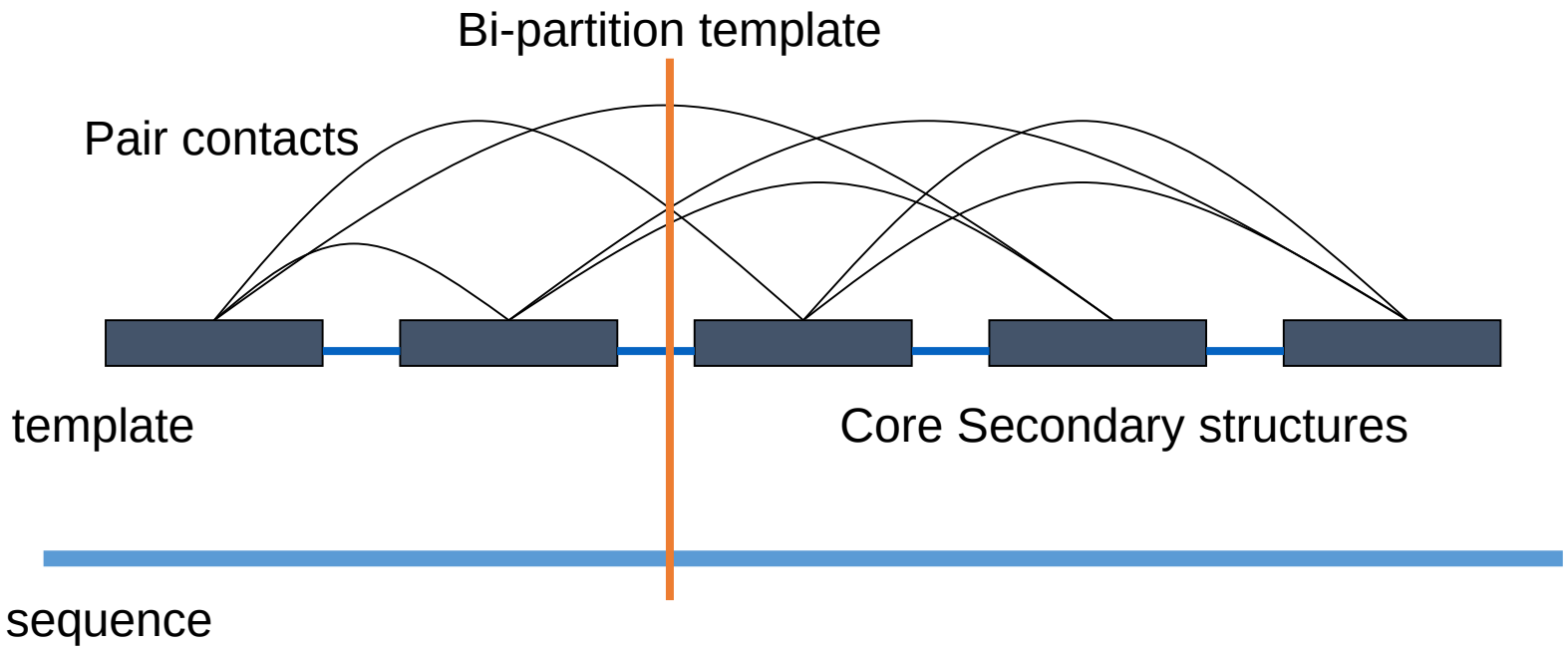
- No gap for core alignment
- Pariwise interactions only between cores



PROSPECT algorithm

Divide-and-conquer algorithm:

- repeatedly bi-partition template into sub-structures till cores
- merge partial alignments into longer alignments *optimally*



PROSPECT pseudo-code

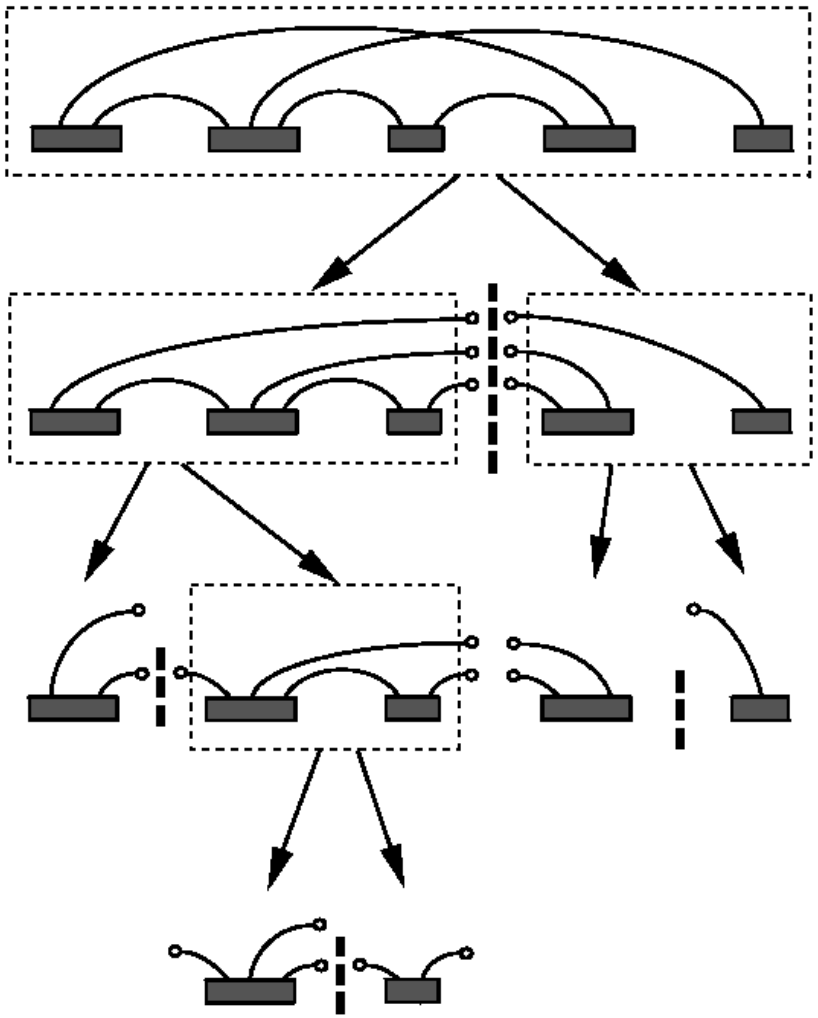
Procedure THREADING$(s[i,j], t[p,q], L_{p,q}, P_{p,q})$

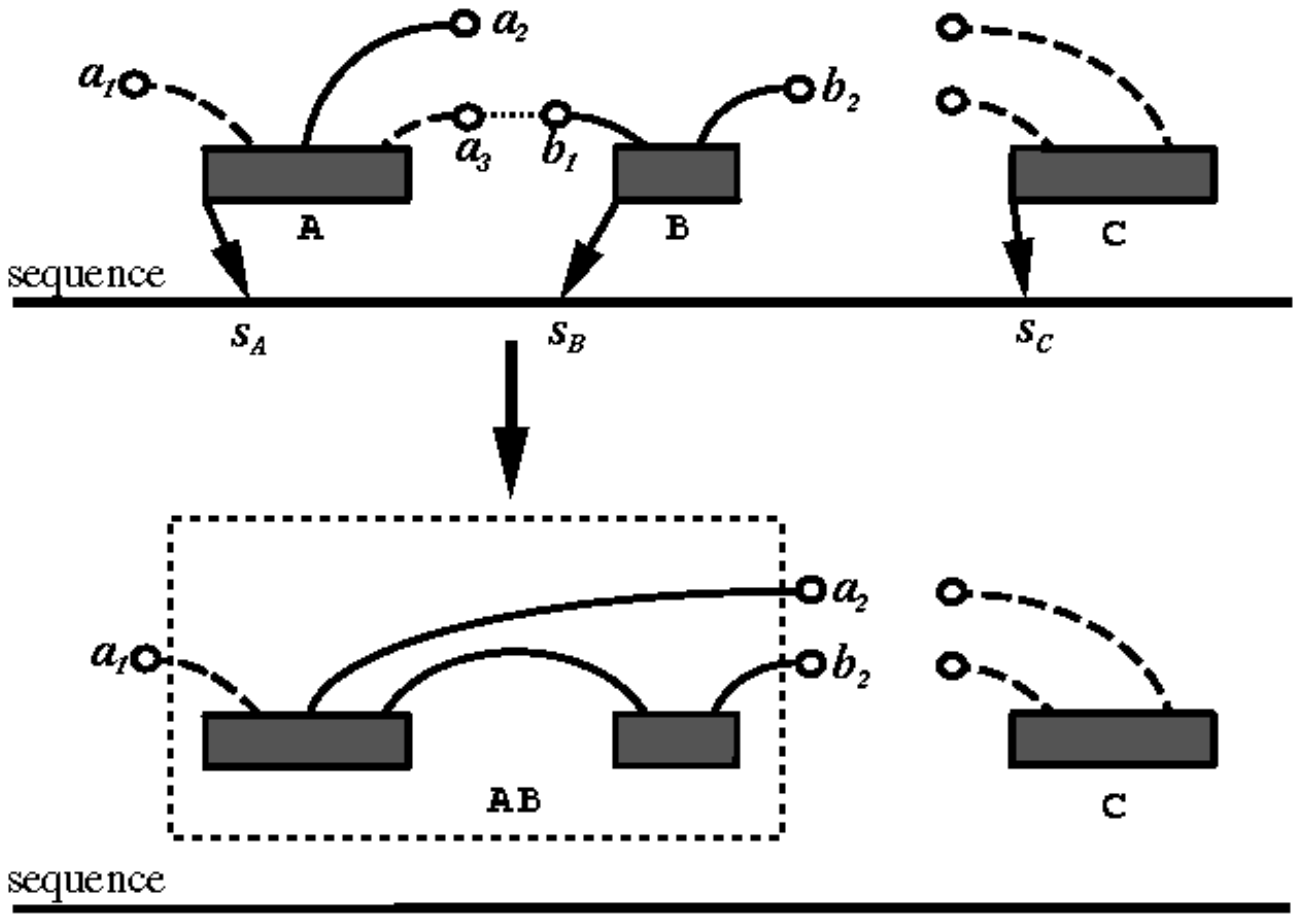
1. **if** $t[p,q]$ is a core **then**
2. **return** optimal sequence-alignment score between $s[i,j]$ and $t[p,q]$, added by the pairwise contact energies involving an element of $t[p,q]$ and an element outside $t[p,q]$;
3. **else**
4. score $\leftarrow +\infty$;
5. **for** each $k_1, k_2 \in [i, j-1]$ with $k_1 < k_2$ **do**
6. **for** each possible set P_{p,r_1-1} and each possible set $P_{r_2+1,q}$ that are consistent with $P_{p,q}$ **do**
7. $\text{score}_1 \leftarrow$ **THREADING**$(s_{i,k_1}, t_{p,r_1-1}, L_{p,r_1-1}, P_{p,r_1-1})$;
8. $\text{score}_2 \leftarrow$ **THREADING**$(s_{k_2,j}, t_{r_2+1,q}, L_{r_2+1,q}, P_{r_2+1,q})$;
9. $\text{score}_0 \leftarrow$ optimal sequence-alignment score between $s[k_1+1, k_2-1]$ and $t[r_1, r_2]$;
10. **if** score $>$ $\text{score}_0 + \text{score}_1 + \text{score}_2$ **then** score $\leftarrow \text{score}_0 + \text{score}_1 + \text{score}_2$;
11. **return** score.

PROSPECT algorithm

- The algorithm first calculates the alignment score two partitions. Since we assume that there is no alignment gap within a core alignment, this score can be calculated by simply adding the singleton scores.
- The calculation of the pair score is tricky since we do not know which sequence positions are aligned to the cores at the other ends of the open links. To overcome this, we simply consider all possible legal alignments of these cores.
- Note that not every combination of the alignments of these cores makes a legal (overall) alignment since some of them may
  - violate the relative order of these cores (e.g., the first core is aligned to a sequence position that is to the right of the aligned sequence position of the fourth core);
  - overlap with each other;
  - violate the allowed minimum and maximum length difference in loop alignments (we allow a user to specify these numbers in PROSPECT).

PROSPECT algorithm



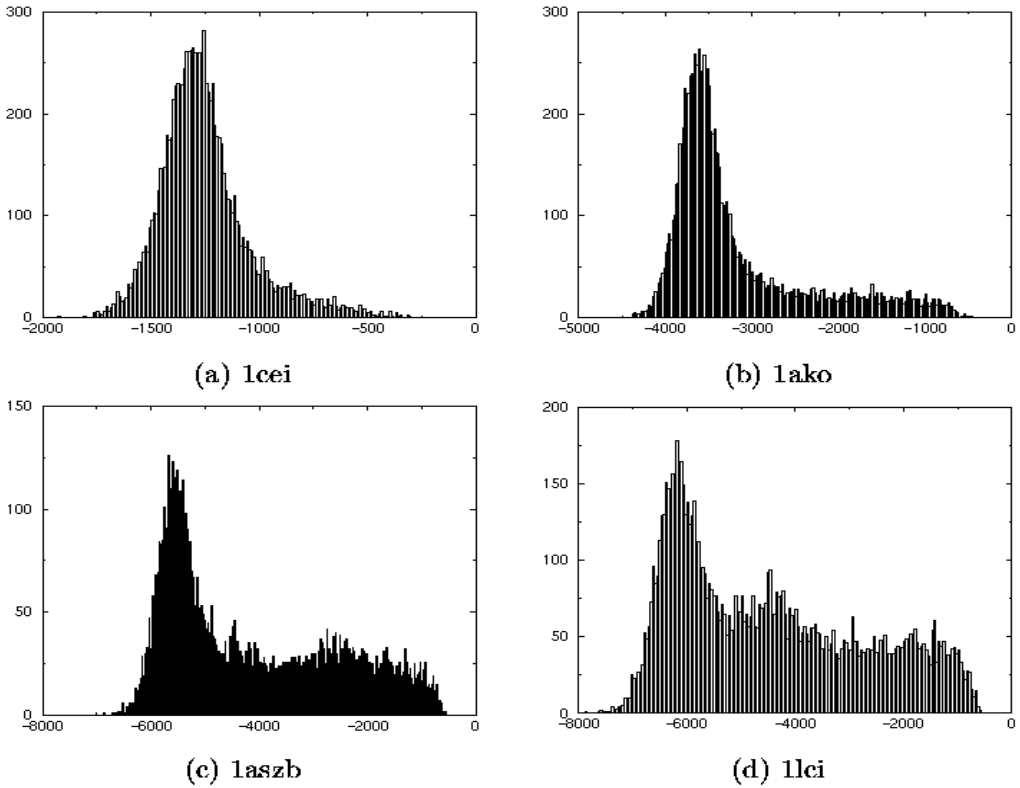


Threading scores

- A confidence score is need to normalized raw threading score
- Z-score through random shuffling

$$\text{z-score} = \frac{\text{score} - \text{ave\_score}}{\text{standard\_dev}}$$

- Using known correct pairs for training (neural networks / SVM)



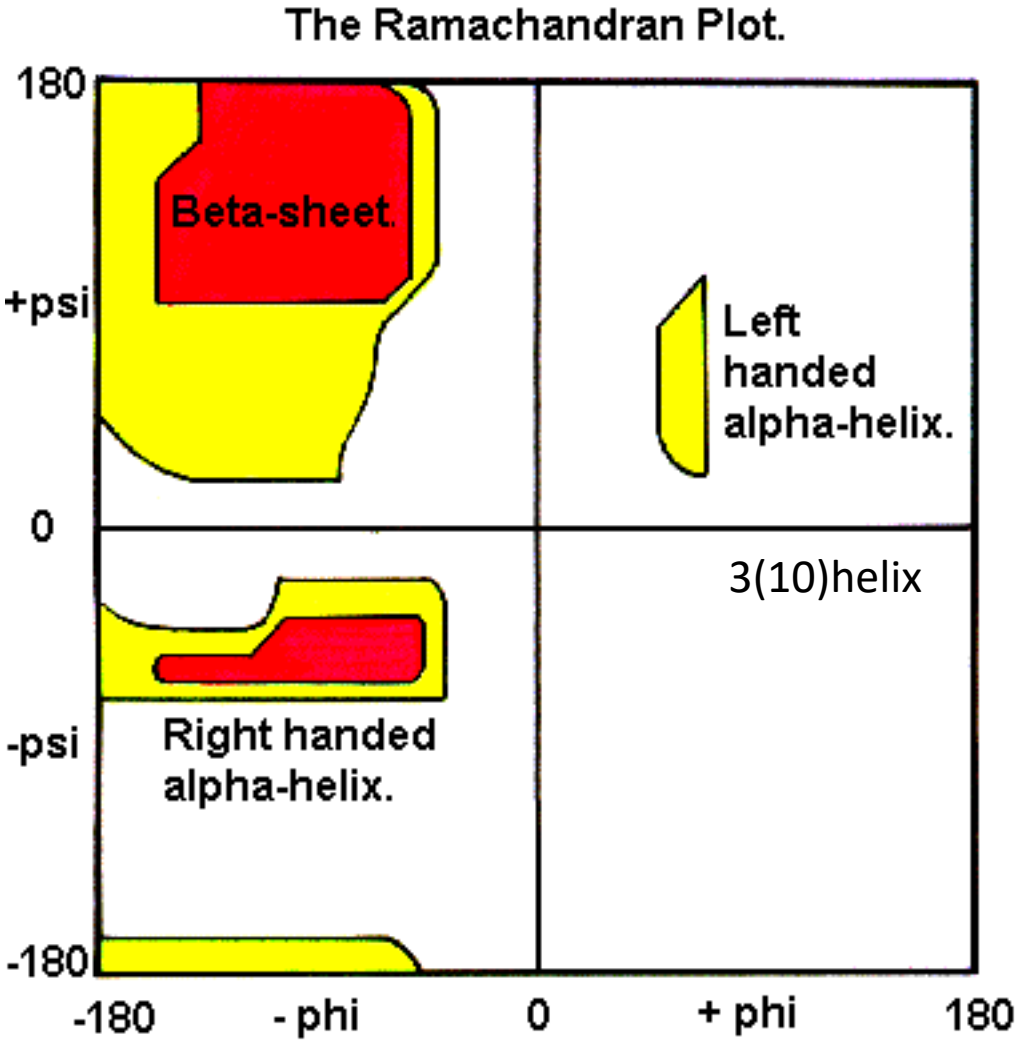
Performance

TABLE VI. Threading Performance With Predicted Secondary Structures[†]

Set	1–6%	7–9%	9–12%	13–15%	16–18%	19–21%	22–24%	25–27%	28–30%	Overall
Superfamily-train	68%	22%	58%	64%	73%	87%	89%	84%	89%	74%
Superfamily-test	39%	38%	49%	67%	67%	82%	89%	93%	98%	73%
Fold-family-train	0%	12%	26%	31%	29%	60%	90%	—	—	29%
Fold-family-test	80%	55%	22%	35%	17%	20%	92%	—	—	28%

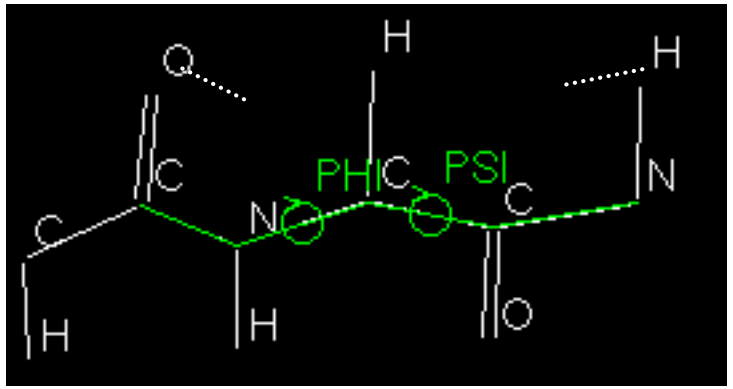
[†]Each column represents a different range of sequence identity level. Each row represents the averaged alignment accuracy among pairs with a particular level of sequence identity for each of the four sets: the training and testing sets of the superfamily set, and the training and testing sets of the fold-family set.

Rosetta Stone Approach (mini-threading)



RADFGHYPL

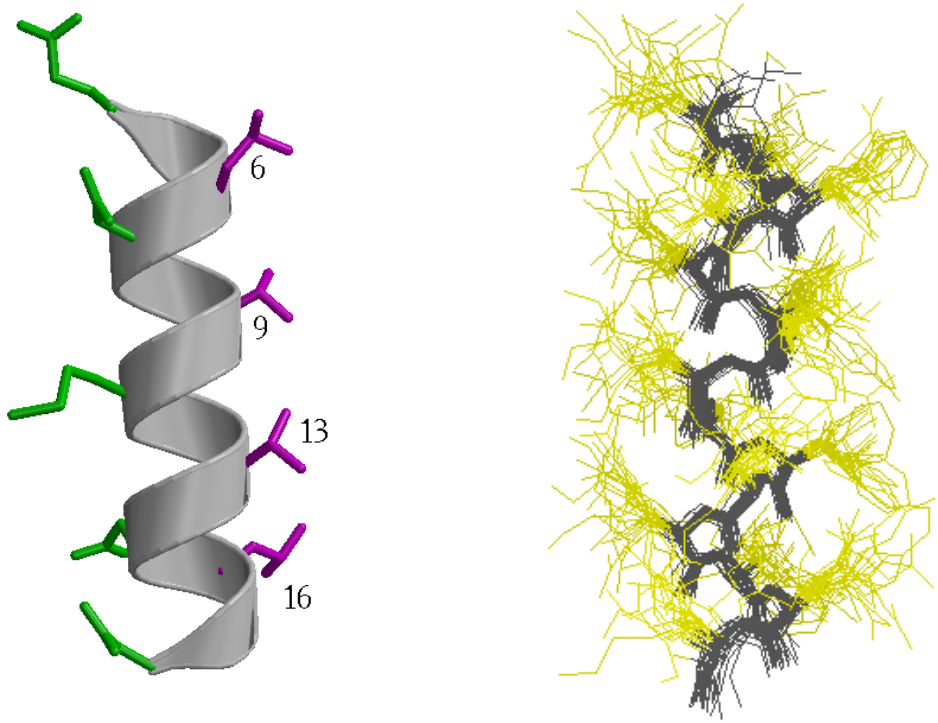
(local sequence)



Protein structure

Micro sequence-structure relationship

Some sequence patterns strongly correlate with protein structure at the local level

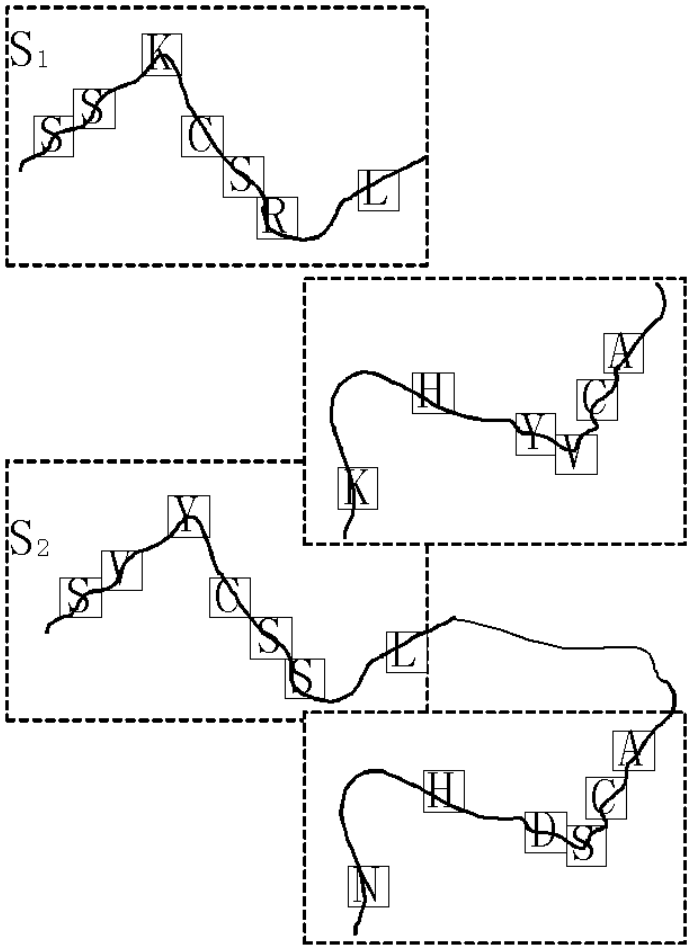


amphipathic helix

Mini-threading

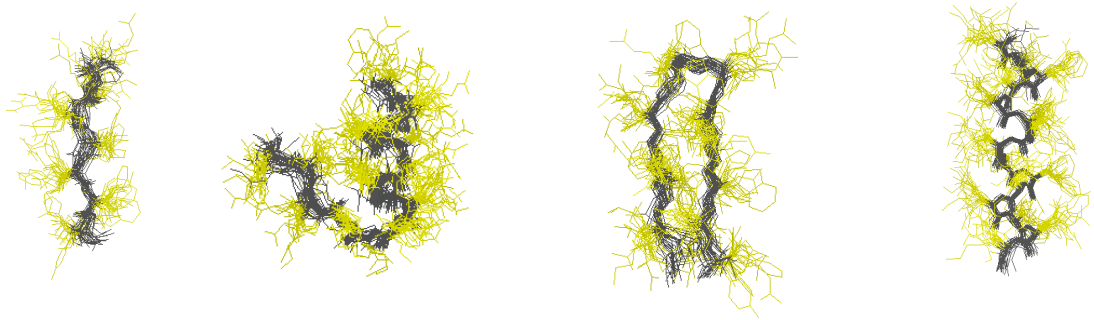
```
SVKCSRL
|  |||||
SSKCSRL

SVKCSRL
||  ||  |
SVYCSSL
```



Similar sequence → Similar structural segment

Model building



-Search for compatible fragments of short sequences in structure database (9-mer)

-Build phi-psi angle distributions

-Use Monte Carlo simulated annealing to assemble the fragments

-Scoring functions are used to select best models (~1000)

-Clustering the model to choose the best one