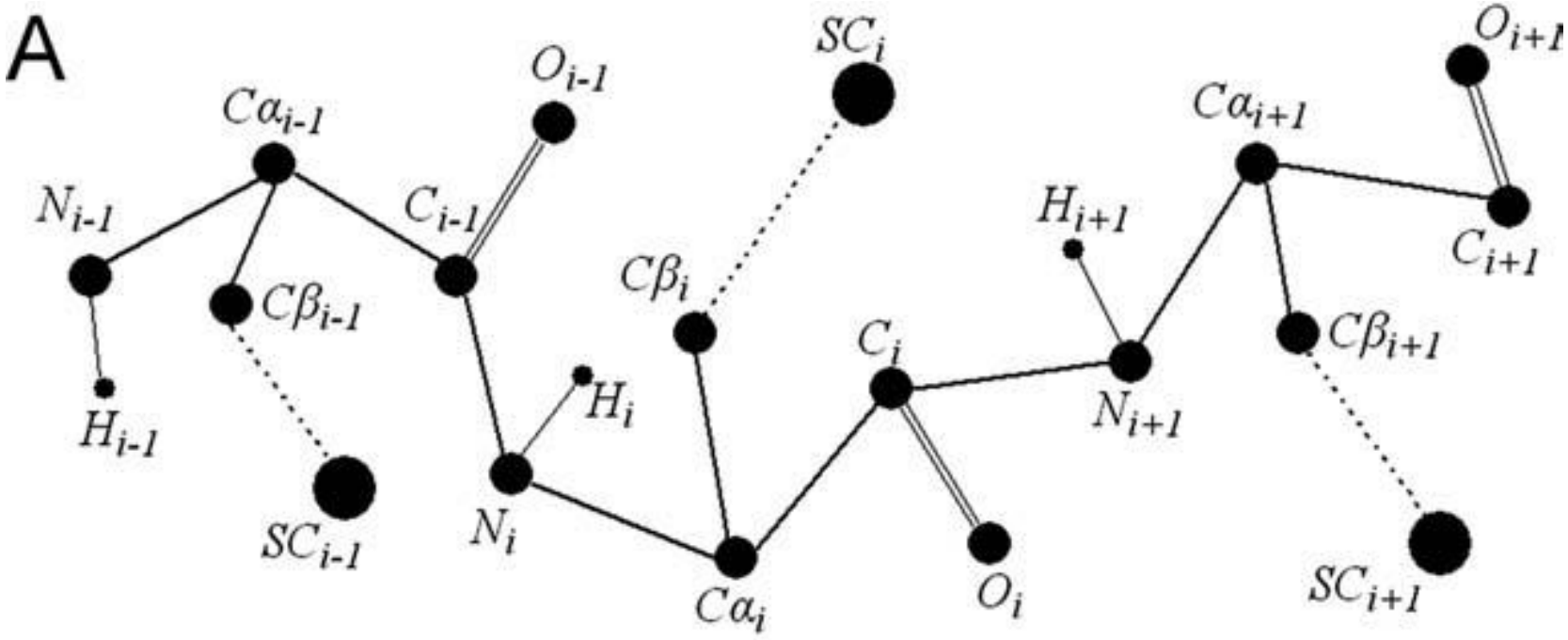
Lecture 26-

Protein Folding

Model representation



Identification of protein fold

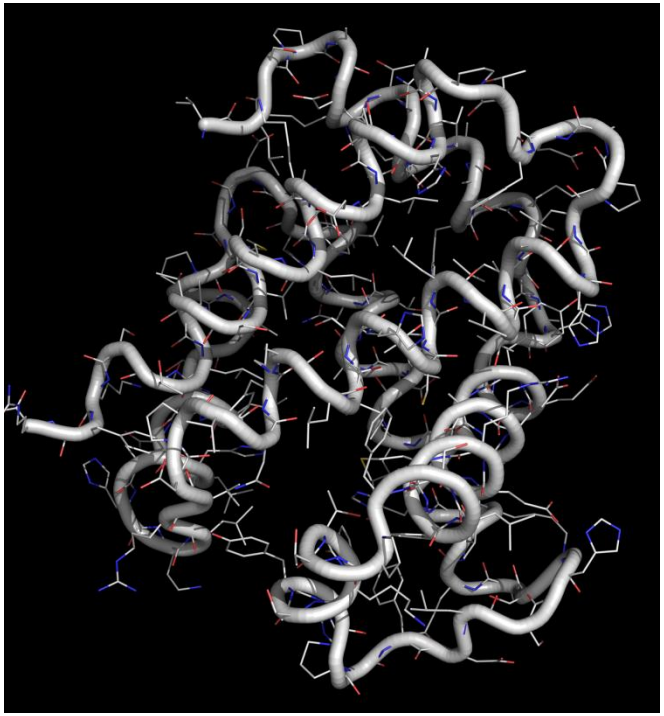
>1A00:A|PDBID|CHAIN|SEQUENCE
VLSPADKTNVKAAWGKVGAHAGEYGAEALERM
FLSFPTTKTYFPHFDLSHGSAQVKGHGKKVAD
ALTNAVAHVDDMPNALSALSDLHAHKLRVDPV
NFKLLSHCLLVTLAAHLPAEFTPAVHASLDKF
LASVSTVLTSKYR

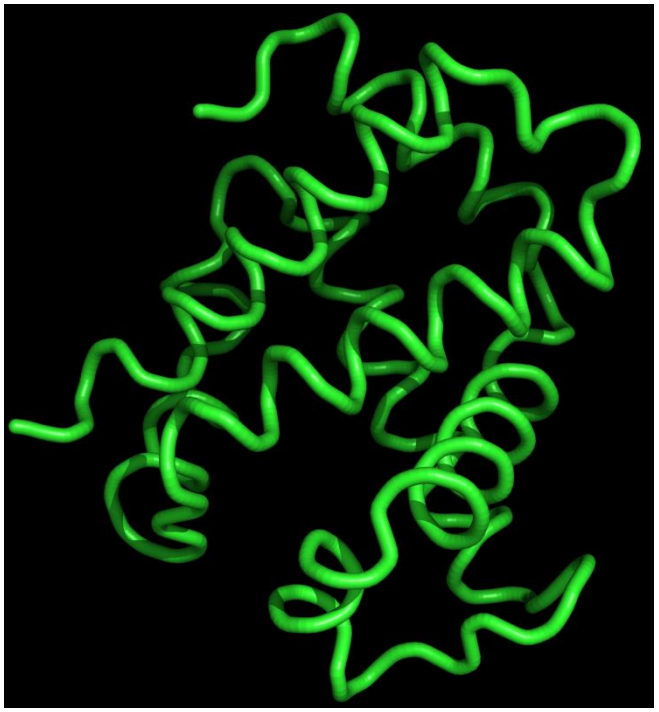
Side chain prediction

- Most of the modeling methods first predict backbone.

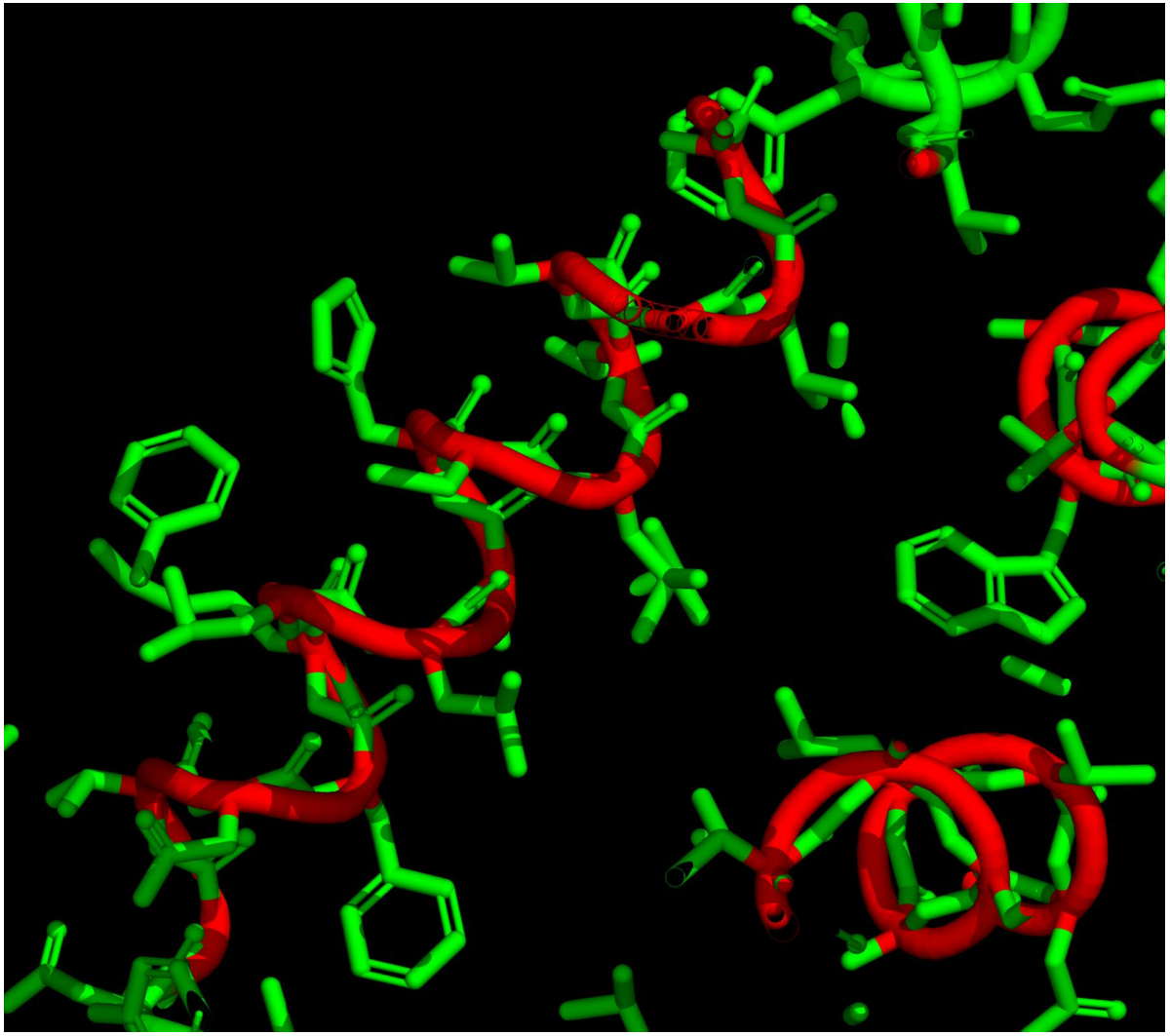
- Side chains are optimized after fixing backbone folds.

Side Chain Fitting





Side Chain Fitting



In silico folding

- ***Ab initio* protein modeling**
- **Comparative Protein Modeling/Template Based Modeling**
  - Homology modeling
  - Protein threading

Homology Modeling

- Homology modeling relies on the identification of one or more known protein structures likely to resemble the structure of the query sequence, and on the production of an alignment that maps residues in the query sequence to residues in the template sequence.

- The quality of the homology model is dependent on the quality of the sequence alignment and template structure. Homology modeling can produce high-quality structural models when the target and template are closely related.

Homology Modeling - Steps

- Template selection and Target-template alignment
  - Pairwise sequence alignment (BLAST, NW etc.)
  - Multiple sequence alignment (PSI-BLAST using PSSM)
  - 3D-1D alignment
- Model construction
  - Fragment assembly
  - Segment matching
  - Satisfaction of spatial restraints
  - Loop modeling
- Model assessment
  - Structural comparison methods
  - RMSD/TM-Score

MODELLER

Homology Modeling - Drawbacks

- Loop modeling
- Low sequence identity
- Larger gaps in the alignments
- Side chain packing/positioning

Protein threading

Protein threading/fold recognition

- Protein threading/fold recognition is used to model those proteins which have the same fold as proteins of known structures, but do not have homologous proteins with known structure.
- It differs from the homology modeling method of structure prediction as it is used for proteins which do not have their homologous protein structures deposited in the Protein Data Bank (PDB).
- Threading works by using statistical knowledge of the relationship between the structures deposited in the PDB and the sequence of the protein which one wishes to model.

Protein threading/fold recognition

Protein threading is based on two basic observations:

1. The number of different folds in nature is fairly small (~ 1300).

2. 90% of the new structures submitted to the PDB in the past five years have similar structural folds to ones already in the PDB.

Protein threading/fold recognition

Method

The construction of a structure template database

1. Select protein structure templates from PDB/FSSP/SCOP/CATH
2. Remove protein structures with high sequence similarities.

The design of the scoring function

1. Design a good scoring function to measure the fitness between target sequences and templates.
2. A good scoring function should contain mutation potential, environment fitness potential, pairwise potential, secondary structure compatibilities, and gap penalties.

Protein threading/fold recognition

Method

Threading alignment

1. Align the target sequence with each of the structure templates by optimizing the designed scoring function.
2. This step is one of the major tasks of all threading-based structure prediction programs that take into account the pairwise contact potential.
3. Alternatively, look for a dynamic programming algorithm.

Threading prediction

1. Select the threading alignment that is statistically most probable as the threading prediction.
2. Construct a structure model for the target by placing the backbone atoms of the target sequence at their aligned backbone positions of the selected structural template

Protein Threading

- Given a library of possible protein folds and an amino acid sequence find the fold with the best sequence -> structure alignment (threading)

- Evolution depends on designability to preserve function under mutation
- Estimate only different protein structures exist in nature (Chothia,1992)

Four components of threading

1. A library of protein folds (templates)
2. A scoring function to measure the fitness of a sequence → structure alignment
3. A search technique for finding the best alignment between a fixed sequence and structure
4. A means of choosing the best fold from among the best scoring alignments of a sequence to all possible folds

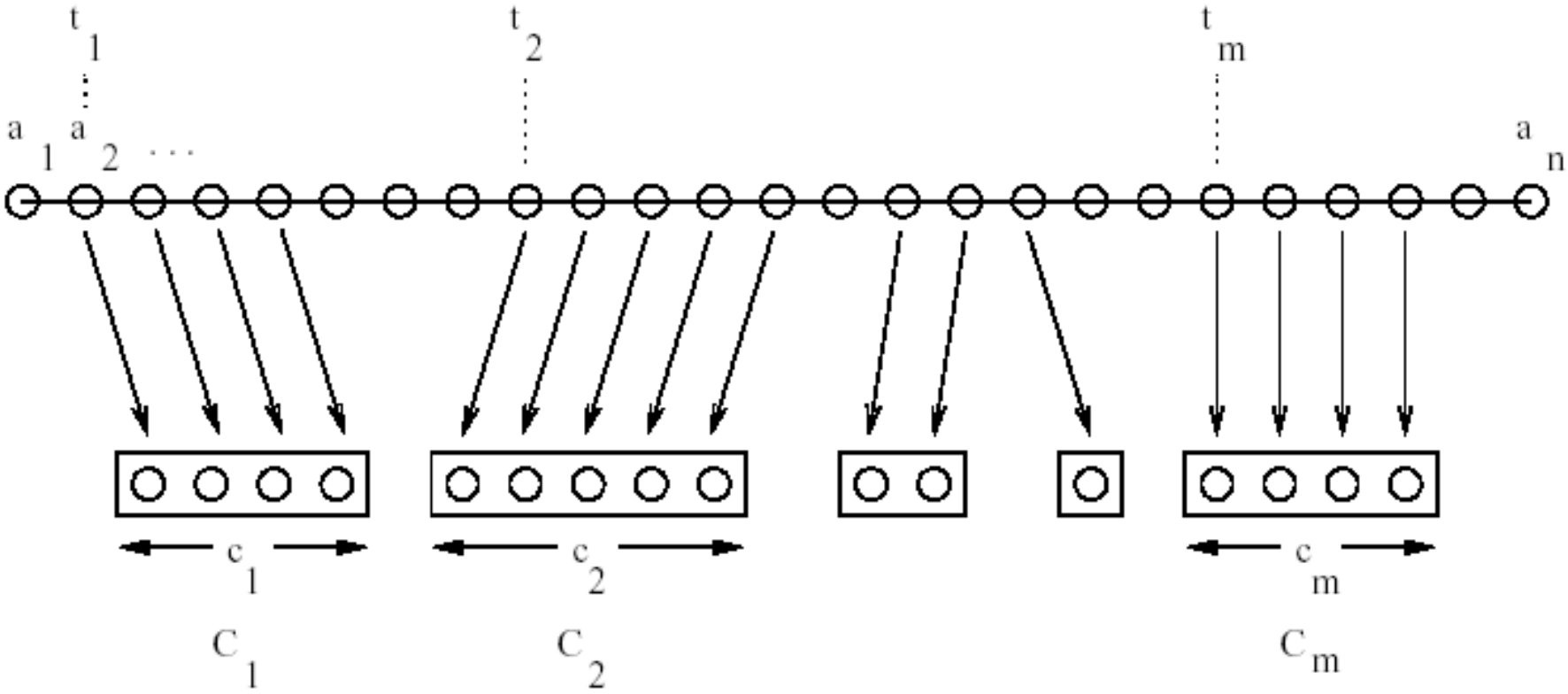
Scoring Schemes for Sequence to Structure Alignments

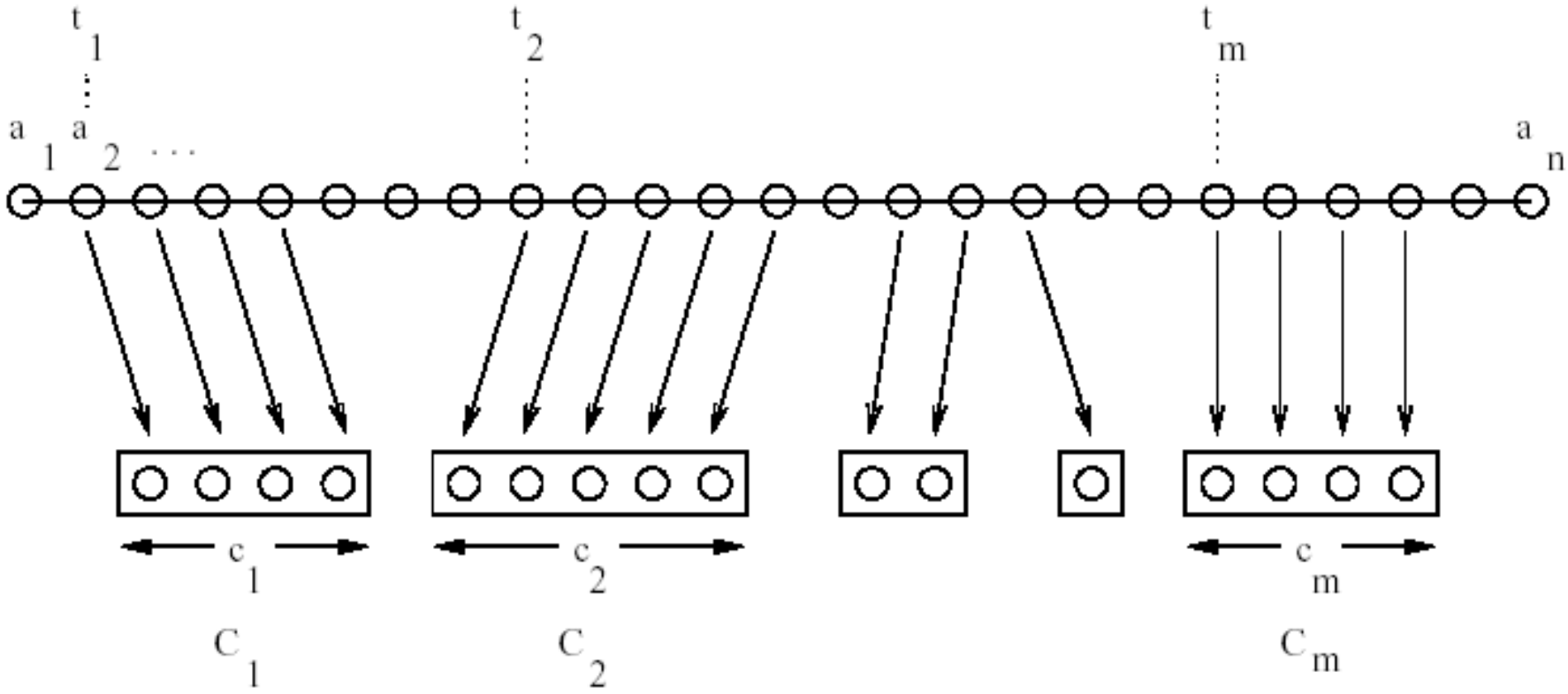
- The scoring scheme for a particular threading of a sequence onto a structure measures the degree to which
  - Environmental preferences are satisfied
  - Different amino acid types prefer different environments
  - Structural preferences: helix, sheet, not exposed to solvent
  - Pairwise interactions with neighbouring amino acids

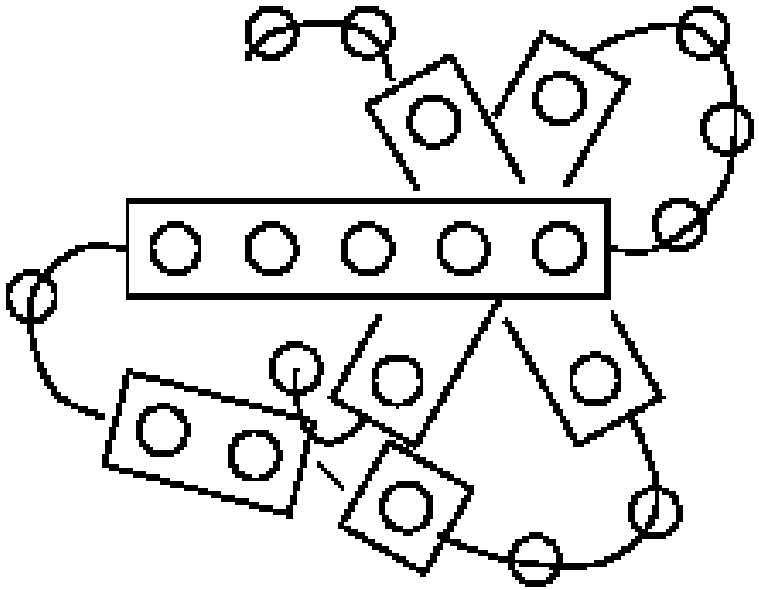
Formal Statement of the Protein Threading Problem

- ***C*** is a protein core having n segments C_i representing a set of contiguous amino acids. Let c_i be the length of C_i
- Sequence $\boldsymbol{a} = a_1 a_2 \ldots a_n$ of amino acids

A threading $t = (t_1, t_2, t_3, .. t_m)$ is a list of indices of **a**, where t_i indicates amino acid a_{ti} occupies the first position in core segment C_i. Valid threading are non-overlapping and sequential; thus t is subject to $1 \leq t_1$, $t_i + c_i \leq t_i + 1$, $t_m + c_m \leq n + 1$. $a_{ti} \rightarrow C_{i,1}$




t 1
t 2
t m
a 1
a 2
a n
c 1
c 2
c m
C 1
C 2
C m



Scoring functions:

- $f(i,t_i)$ = degree to which the structural preferences of core segment C_i are satisfied.
- $g(i,j,t_i,t_j)$ = pairwise interactions of core segments C_i an C_j
- The score of a threading $\boldsymbol{t}$ is the sum

$$\Sigma_i \, f(i,t_i) + \Sigma_{i<j} \, g(i,j,t_i,t_j)$$

Decision problem:

Given $\boldsymbol{a}$, c_1, c_2, …, c_m, f and g, and a number x, does there exists a threading $\boldsymbol{t}$ with score x or greater?

Optimization problem:

Given $\boldsymbol{a}$, c_1, c_2, …, c_m, f and g, find a threading $\boldsymbol{t}$ with maximal score.

Current limitations to protein threading

- Statistical problems
- Definition of neighbor and /or pairwise contact environments:
  - Energetic neighbor? Or Contact neighbor

Computational Complexity of Finding an Optimal Alignment

- The complexity of the protein threading problem depends on whether:

  (i) Variable-length gaps are allowed in alignments

  (ii) the scoring function for an alignment incorporates pairwise interactions between amino acids

  - Property(I) makes the search space exponential in size to the length of the sequence
  - Property(II) forces a solution to take non-local effects into account

Any protein threading scheme with both properties is NP-complete (3-SAT Lathrop 1994) (MAX-CUT Akutsu,Miyano 1999)

Thus all protein threading approaches can be divided into four groups:

1. No variable length gaps allowed
2. No pairwise interactions considered in scoring function
3. No optimal solution guarantee
4. Exponential runtime

Comparison with homology modeling

- Both are TBM, no boundary in terms of prediction techniques.
- But the protein structures of their targets are different.
  - Homology modeling (HM) is for those targets which have homologous proteins with known structure (usually/maybe of same family),
  - Protein threading (PT) is for those targets with only fold-level homology.
  - Thus, HM is for "easier" targets and PT is for "harder" targets.
- Homology modeling treats the template in an alignment as a sequence, and only sequence homology is used for prediction.
- Protein threading treats the template in an alignment as a structure, and both sequence and structure information extracted from the alignment are used for prediction.