Ab initio structure prediction

Kim T. Simons, Charles Kooperberg, Enoch Huang and David Baker
“Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions”
J. Mol. Biol., vol. 268, 209-225 (1997).

A simulated annealing procedure needs:

- method for generating structures
- scoring function

Generating structures

Three-dimensional structures are generated by splicing together fragments of proteins of known structure with similar local sequences.

Earlier studies showed a strong correlation between local sequence and local structure of nine residue fragments.

For each segment of length 9 in the sequence being folded, the 25 nearest sequence neighbours in the structure database were identified.

The conformation of each of these segments was adjusted to give ideal bond lengths and angles.

The percentage of neighbours structurally similar to the true structure is greater when multiple sequence information is available.

Estimating P(structure)

In fold recognition, we can assume that each known fold (a finite set) is equally probable.

However, when considering a vast number of synthesised conformations, many of which are highly improbable, we need some way of assessing the feasibility of each conformation.

Simons et al. (1997) suggest a simple approach in which P(structure) is zero if atoms overlap, and otherwise P(structure) is related to the compactness of the structure, measured by the “radius of gyration”.

The radius of gyration is defined as the square root of the mass average of r_i^2 for all of the mass elements.

Estimating P(sequence|structure)

Similar to scoring a sequence-fold match when threading.

Profiles:

$$\prod_i P(aa_i \mid E_i)$$

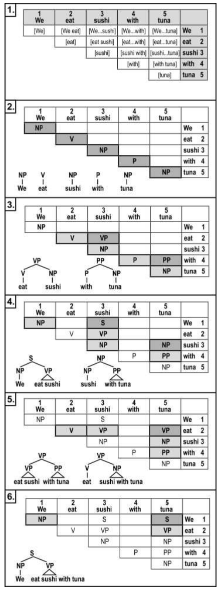
Pairwise potentials:

$$\prod_{i<j} P(aa_{i,} aa_j \mid r_{ij})$$

Simons et al. (1997):

$$\prod_i P(aa_i \mid E_i) \times \prod_{i<j} \frac{P(aa_{i,} aa_j \mid r_{ij,} E_{i,} E_j)}{P(aa_i \mid r_{ij,} E_{i,} E_j) P(aa_j \mid r_{ij,} E_{i,} E_j)}$$

The CKY algorithm — natural language



Parsing natural language vs. folding a protein

Parsing natural language:

a) start with one-dimensional string of words;

b) consider all possible topologies representing possible relationships among words and phrases;

c) chooses the one that conveys the correct single meaning of the sentence.

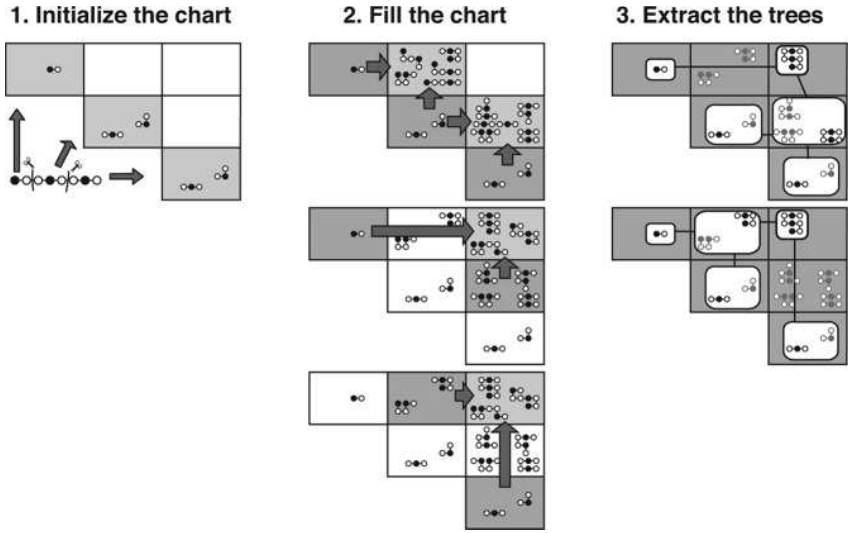
Folding a protein:

a) start with one-dimensional string of amino acid residues;

b) consider all possible topologies representing possible native substructures of a protein;

c) chooses the one that has the global minimum free energy.

The CKY algorithm — protein structure



Zipping and Assembly Mechanism by Dynamic Programming

- A variant of the CKY (Cocke-Kasami-Younger) algorithm
- Zipping and Assembly Mechanism by Dynamic Programming (ZAMDP)

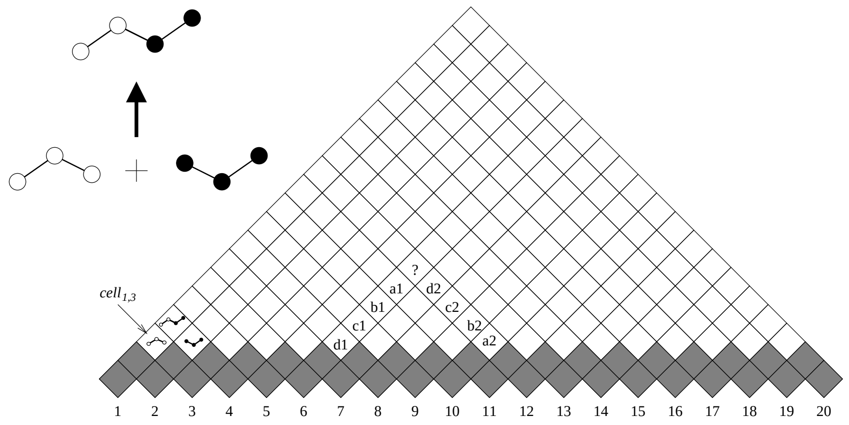
Unlike standard CKY, ZAMDP does not use a grammar, but simply concatenates adjacent chain fragments like pieces of a jigsaw puzzle, and explores all their local configurations. When two pieces are brought together, we search all the viable ways they can be configured and keep only those having lowest energies.

[Dill et al., 2007]

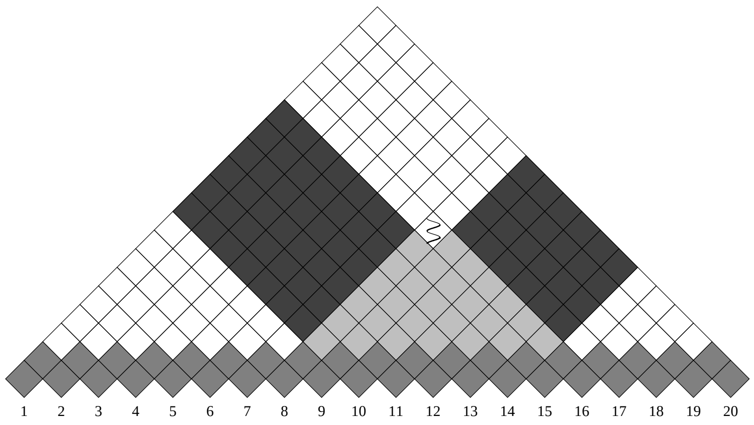
Claims made for ZAMDP method

- local-first-global-later explains quick folding, and avoidance of vast stretches of conformational space
- reflects parallel nature of physical kinetics
- captures relationship between contact order (whether contacts are mainly local or mainly non-local) and folding rate
- identifies slow- and fast-folding proteins, and slow- and fast-folding routes

Zipping and assembly



Zipping and assembly


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20