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

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.

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(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):

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