Monte Carlo Simulations without Herrn Boltzmann

Nasir Mahmood, Andrew Torda
Centre for Bioinformatics, University of Hamburg, Germany
mahmood@zbh.uni-hamburg.de

Introduction

Must one obsequiously grovel before the grave of Herrn Boltzmann in order to play protein structure prediction? If one has a statistical model for aspects of protein structure, one could use this as a probabilistic framework. This leads to a method which

- does not rely on the Boltzmann relation
- performs Monte Carlo directly on a probability surface without using any explicit energy terms

Philosophy

Probabilities are calculated from the product of several distributions with

- simple and bivariate Gaussians for a solvation term and the backbone angles
- multi-way discrete distributions for the sequence of a protein The distribution parameters come from surveying the protein data bank, based on small fragments.

We knew at the start that the method has a great weakness in the treatment of long-range interactions.

Method

1. Score function:

Unlike most Monte Carlo methods we do not use an energy, but calculate probabilities (or ratio of probabilities) directly:

$$\frac{P^{stat}(c_i)}{P^{stat}(c_j)} \bullet e^{-w\frac{\Delta E_{hb}}{kT}} \quad (1)$$

$P^{stat}(c_i)$	probability of new formation
$P^{stat}(c_j)$	probability of old formation
E_{hb}	hydrogen bond energy
k	Boltzmann constant
T	temperature
w	weight factor for hydrogen bonding

$$P^{stat(c_x)} = \left(\prod_{i=1}^{N_{fragments}} \sum_{j=1}^{N_{classes}} w_{jif_i^{seq}} P_j\left(f_i^{struct,slv}\right) \right)^{\frac{1}{N_{fragments}}} \quad (2)$$

$$E_{hb} = q_1 q_2 \left(\frac{1}{r(ON)} + \frac{1}{r(CH)} - \frac{1}{r(OH)} - \frac{1}{r(CN)} \right) \bullet f \quad (3)$$

There is no explicit energy or score function, merely probabilities for descriptors associated with fragments of proteins. These have come from a probabilistic (fuzzy) classification of protein fragments. The only ugly component is a rather ad hoc hydrogen bonding term.

The statistical models used to model information coming from sequence, structure and solvation are:

1. sequence – multi-way Bernoulli
2. structure – bivariate Gaussian
3. solvation – simple Gaussian

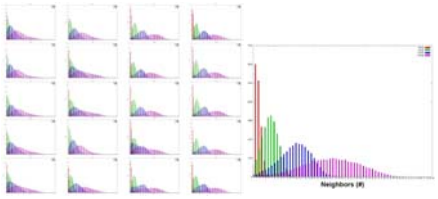


Figure 1: Solvation - Number of neighbours an amino acid has within a certain range is taken as a solvation measure. Four different coloured histograms show neighbour count within four different ranges.

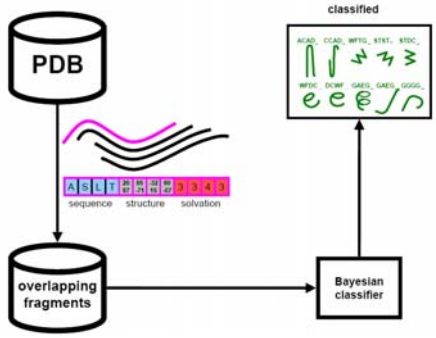


Figure 2: Bayesian classification – overlapping fragments generated from existing structures are classified into a number of classes by Bayesian classifier. Each fragment is represented by its sequence, structure (dihedral angles) & solvation.

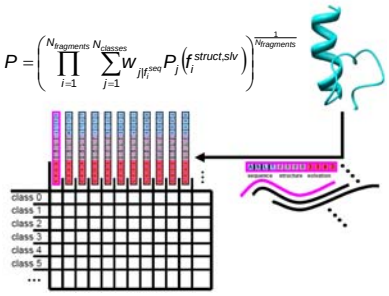


Figure 3: Calculation of probability of a protein structure with the Bayesian classification by taking into account its sequence, structure and solvation.

2. Search Method:

We are using simulated annealing Monte Carlo as a search method to find probable structural arrangement of a given amino acid sequence.

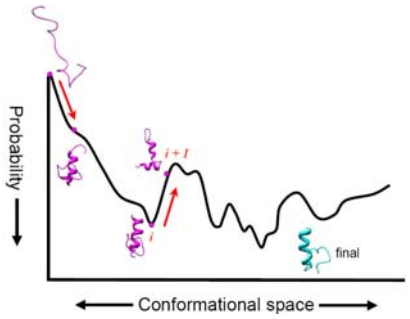


Figure 4: Search method – system starts with a randomly generated structure at high temperature and is gradually cooled down while making (biased or unbiased) moves.

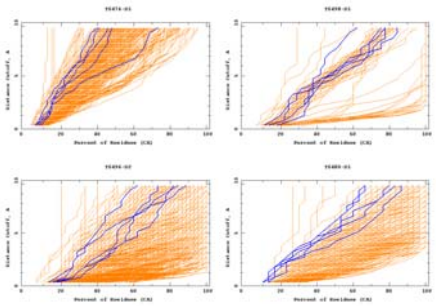
The search method makes two kinds of moves: 1) biased moves made by drawing a fragment from a fragment library generated from existing protein structures and 2) completely unbiased moves.

Internally, the score function is based on dihedral angles, Cartesian coordinates and sequence description, so there is some computational work involved in moving between representations.

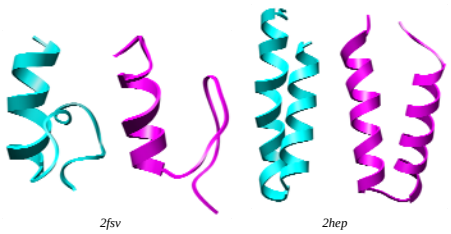
The acceptance criterion depends solely upon the probability ratio (equation 1) calculated from the probabilities of the new and old structures.

Results

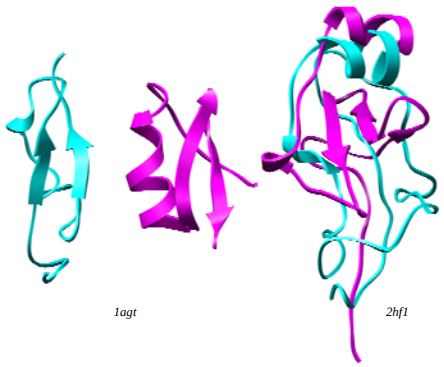
CASP8: Here are few targets where our method did rather well.



Not CASP8:



2fsv *2hep*



1agt *2hf1*

Conclusions:

The code evolved during CASP and continues to change. It seems to be a rather effective (albeit ridiculously expensive) secondary structure predictor which usually produces protein like structures and often quite persuasive structures.

Web server: http://cardigan.zbh.uni-hamburg.de/~mahsch/mahmood/