Protein Structure Prediction: Probabilistic Force Fields

Nasir Mahmood, Andrew Torda

Centre for Bioinformatics, University of Hamburg

Introduction

Protein structure prediction is one of the classic problems from computational chemistry or molecular structural biology. Essentially, one would like to be able to go from the sequence of a protein (easily obtained) to the structure (expensive and often difficult to obtain experimentally).

Our interest has been in devising new purely probabilistic score functions. They make no use of Boltzmann statistics, but instead rely on a mixture of Bayesian probabilities based on normal and discrete distributions. This has an interesting consequence. If one works with a method such as Monte Carlo, one can base the acceptance criterion directly on the calculated probabilities without assuming a Boltzmann distribution.

History

Monte Carlo simulations, in their various forms, have been described by reputable scientists as the path to the simulator's graveyard. This poses the question as to why a rational simulator would venture further into this field.

There are two aspects to this problem:

- the score or quasi-energy function and
- the search method

Method

1. Score function:

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

$$\text{probability ratio} = \frac{P(X_N)}{P(X_o)} \quad (1)$$

$P(X_N)$ probability of new conformation
$P(X_o)$ probability of old conformation

$$P(X|\vec{V},T,S) = \prod_i [\sum_j (\pi_j \prod_k P(X_{ik}|X_i \in C_j, \vec{V}_{jk}, T_{jk}, S))] \quad (2)$$

$X = \{X_1,...,X_I\}$ the set data instances X_i(fragments)
$\vec{X}_i = \{X_{i1},...X_{iK}\}$ attribute vectors X_{ik}describing X_i
i observation number, $i = 1,..., I$
j class number, $j = 1,..., J$
k attribute number, $k = 1,..., K$
l discrete attribute values, $l = 1,..., L$
c inter-class probabilities and parameters
S the set of possible probability density functions (p.d.f.) covering $\vec{V}$, T
$T = T_c, T_1,..., T_J$ the exact functional form of each p.d.f. (usually a Gaussian or multinomial)
$\vec{V} = \vec{V}_c, \vec{V}_1,..., \vec{V}_J$ the set of parameters instantiating p.d.f.
π_j class mixture probability, $\vec{V}_c = \{\pi_1,..., \pi_J\}$

Our score function is purely probabilistic and relies on mixture of Bayesian probabilities by combing sequence, structure and solvation.

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

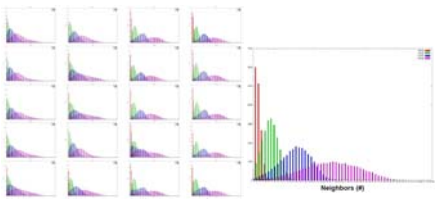


Figure 1: Solvation - Number of neighbours an amino acid has within a certain range is taken as a solvation measure. Four different coloured histograms show neighbour count within four different ranges.

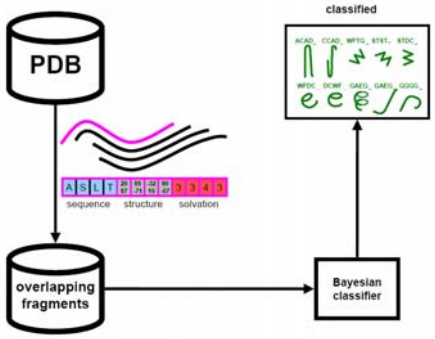


Figure 2: Bayesian classification – overlapping fragments generated from existing structures are classified into a number classes by Bayesian classifier. Each fragment is represented by its sequence, structure (dihedral angles) & solvation.

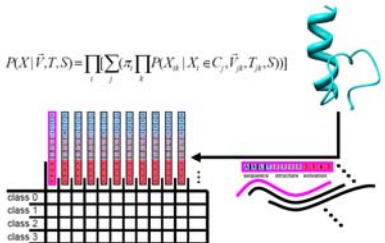


Figure 3: Calculation of probability of a protein structure with the Bayesian classification by taking into account its sequence, structure and solvation.

2. Search Method:

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

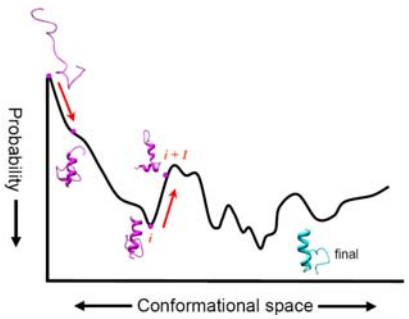


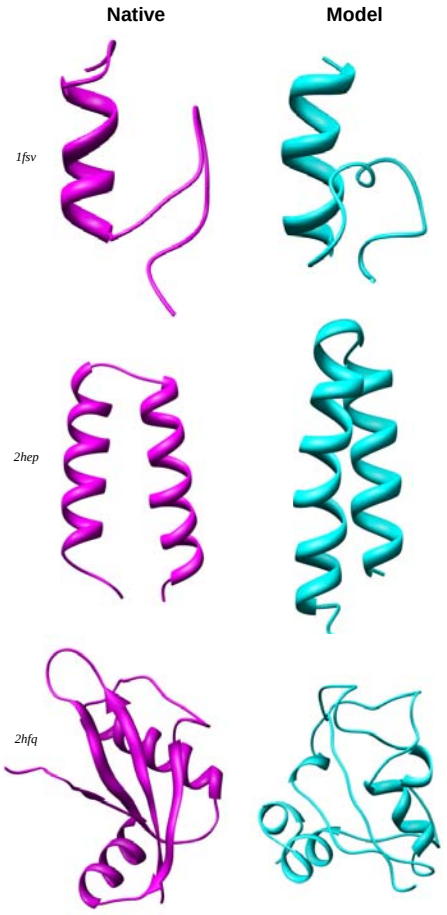
Figure 4: Search method – system starts with a randomly generated structure at high temperature and is gradually cooled down while making (biased or unbiased) moves.

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



Conclusion

The current implementation seems to have a rather good representation of local interactions and works surprisingly well for small proteins. The score function is also being integrated with our existing protein threading machinery for the upcoming CASP8 competition.

We are now working on incorporating simple solvation and hydrogen bond effects into the initial probability calculations to better account for long range interactions.