Protein Structure Prediction using Coarse Grain Force Fields

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

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.

State of the art

In *ab Initio* or *de novo* protein modelling, one seeks to build 3D protein models from scratch rather than modelling them on to known structures. Monte Carlo simulations in their various forms have been a tool that tempts people into 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

The score function may be energy-like or purely statistical and the search method is used to explore the conformational space. The score function and search method are often coupled together and search method is driven by score function to get to native like structures.

Methods

1. Score function:

Our scoring functions, are protein sequence to structure compatilibility/probability functions. These are built by

- A maximally parsimonious Bayesian classification of protein fragments
- Treating a sequence as a set of probability vectors across the resulting classes
- Treating the protein structure as a similar set of probability vectors

$$\text{Probability ratio} = \frac{P(X_N)}{P(X_o)} \qquad (1)$$

$P(X_N)$ probability of new conformation after a Monte Carlo move
$P(X_o)$ probability of old conformation of the structure

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

$X = \{X_1, ..., X_I\}$ the set data instances X_i (fragments)
$\vec{X}_i = \{X_{i1}, ... X_{iK}\}$ the vector of attributes values X_{ik} describing instances 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 parameter values instantiating a p.d.f.
π_j class mixture probability, $\vec{V}_c = \{\pi_1, ..., \pi_J\}$

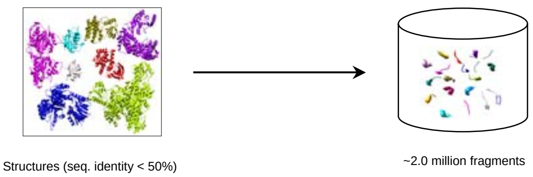

Figure 1: Fragment library used for biased moves

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. The biased moves our search method makes are made by selection from a fragment library generated by chopping up 7000+ structures (which have sequence identity less than 50%) into ~2.0 million fragments (figure 1). Whereas unbiased moves make use of fragments generated from random dihedral angles (figure 2).

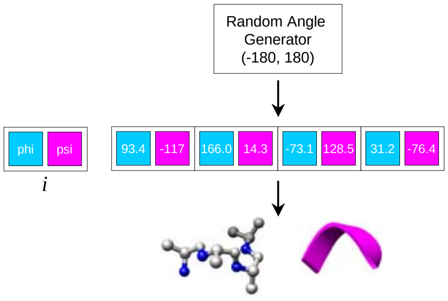

Figure 2: Random fragment used in unbiased moves

Since our score function is based on dihedral angles, the fragments are stored in terms of their dihedral angles rather than Cartesian coordinates and after each successful move, the Cartesian coordinates of the resultant structure are built from its new dihedral angles. Usually, one begins with a random set of angles. The move set consists of fragment replacement either drawn from the fragment library or generated from randomly picked dihedral angles.

The probability ratio (equation 1) calculated from the probabilities of the new and old structures decides the acceptance or rejection of the move as shown in equation 2.

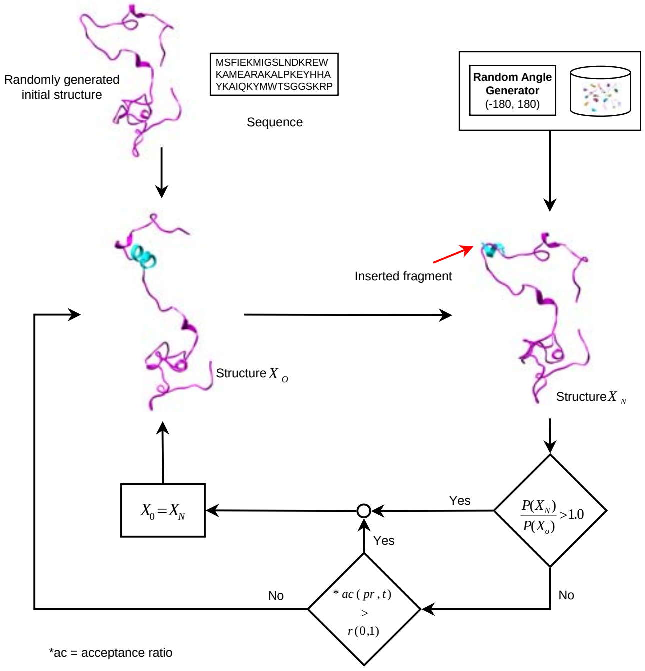

Figure 3: Monte Carlo flow chart

Results

At the moment, our scoring function has little idea of the forces that hold a protein together. This means that it is currently working as a rather impressive secondary structure predictor as shown in figure 4.

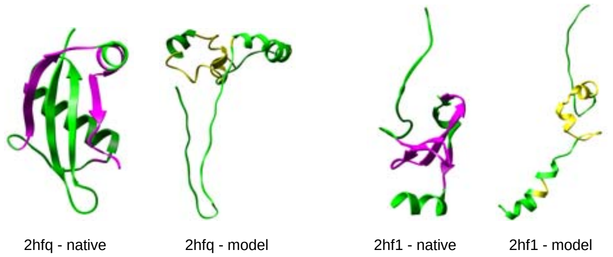

Figure 4: Native structures and their predicted models

We are now introducing solvation/compaction effect, but in a manner which fits to the probability distribution of equation 2.