A machine learning based approach to side chain optimization

Sabareesh Subramaniam¹, Sriraam Natarajan² and Alessandro Senes¹

¹Dept of Biochemistry, University of Wisconsin-Madison and ²Translational Sciences Institute, Wake Forest University, Winston-Salem, NC

Abstract
Side chain optimization refers to the problem of repacking side chain atoms on a fixed backbone so as to minimize the energy of the resultant structure. It is typically performed as a search over the combinatorial space of conformations for all the positions in the backbone. The finite set of representative conformations sampled for each amino acid type is called a “conformer library”. Optimization procedures do not take into account the fact that each position in a protein backbone has different sampling (number of conformations) needs, for example, solvent exposed positions require less sampling than positions buried in the core of the protein. The key contribution of this work is a method to distribute conformations among different positions in a protein backbone based on their sampling needs using machine learning. Our results demonstrate that this strategy helps to redistribute sampling efficiently and helps achieve lower energies.

Sidechain Optimization

A side chain optimization problem is defined by a template (backbone) and a set of energy functions, which define a continuum energy landscape covering all possible theoretical side chain conformations.

Using a conformer library leads to a combinatorial search space

At each position in the backbone, the sidechain is allowed to assume any one of a finite number of conformations. This set of conformations is called a conformer/rotamer library. The conformer library is typically constructed using statistics from various proteins. Sidechain optimization thus reduces to the problem of deciding the conformation at each position in the backbone from among the conformations in the library.

The number of possible conformations grows combinatorially with the number of conformers at each position. Typically, the same number of conformers are assigned to all positions containing a particular amino acid sidechain. In the figure above, all TRPs have 10 conformers, all VALs have 3 conformers and all GLUs have 9 conformers.

The combinatorial complexity may be reduced by predicting the sampling requirement for each side chain based on its immediate environment.

To achieve good energy, a buried position may need a very specific conformation whereas an exposed position may have multiple near isoenergetic solutions. Thus, every position has different sampling requirements.

If we can predict the sampling requirement of each position, we can then re-allocate sampling optimally, reducing the combinatorial complexity and/or achieving better energy.

Using the labels predicted by the machine learning algorithms helps achieve lower energies than an equivalent unbiased method.

The following classification algorithms, as implemented in WEKA, were used 1) Bagging 2) CART 3) LogitBoost 4) NaiveBayes

The labels predicted by all the classifiers were used to perform sidechain optimization and the energies obtained were compared against an unbiased sampling of comparable size. All the classifiers were able to achieve lower energies than the unbiased scheme in a majority of the test set proteins.

References