Predicting Gibbs Free Energy Using Statistical Information and Machine Learning

Gina El Nesr, Doug Barrick
The Johns Hopkins University, Dept. of Biophysics

Introduction

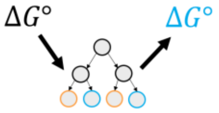
Our ability to measure the Gibbs free energy change of mutations in each protein does not keep pace with the increase in new sequence data. With the already available stability data, there have been some attempts at developing predictors of protein stability; however, none have attempted to use statistical information of the protein as predictive features in their algorithms.

Recently, consensus sequences have been found to be more stable and biologically active, suggesting that each residue's frequency could be a valuable predictor in protein stability. By using information intrinsic to both the protein's sequence and structure, we use machine learning algorithms that predict the stability of a protein. A successful predictor would also shed light on characteristics of the protein are important to its stability.

Machine Learning:

Random Forests

Random forests (RF) are an ensemble supervised learning method used for either classification or regression. Using multiple decision trees, a random forest outputs the mean of the classes that each individual tree predicted in a classification problem.



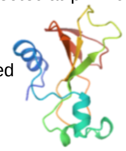
Performance is dependent on the features the RF uses for classification. If the features it is given do not yield any predictive power, their performance will tend to be low. If features have some predictive power, we can generate a ranking of their importance; however, their behavior is a "black box'". Here, we attempt to solve a classification problem of if a protein mutation is stabilizing or destabilizing relative to its mean ΔGº.

Datasets

[1]Staphylococcal nuclease
PDB: 3BDC

Due to the supervised nature of a RF, we need data from proteins whose ΔGº of mutation has been experimentally obtained (with accuracy). ΔGºs must have been collected at pH 7.0.

Two datasets were used (obtained from ProTherm). The Snase dataset contained 514 mutations with known ΔGº. The Barnase dataset contained 200 mutations with known ΔGº.


[1]Barnase
PDB: 1BNI

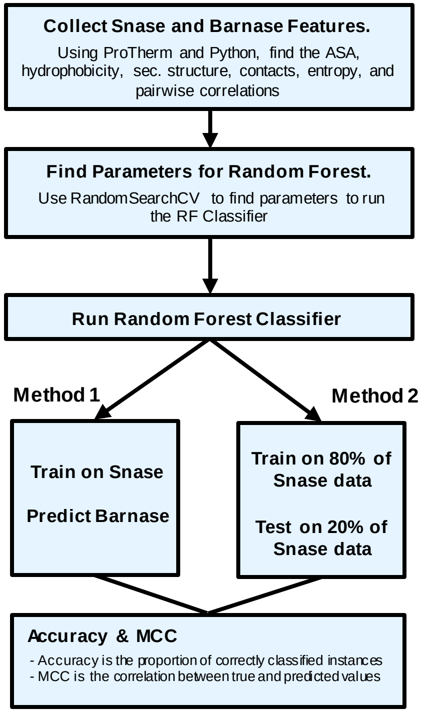
Feature Selection

Success of a RF depends on the features it trains upon. We test upon a subset of structural (obtained from ProTherm) and statistical information from each protein and its mutation.

- **Accessible Surface Area (ASA):** the solvent accessible SA of the wildtype residue
- **Hydrophobicity**: the change of hydrophobicity using the Octanol scale for each residue
- **Secondary Structure**: information on whether the residue being mutated is on a helix, sheet, or coil
- **Residue Contacts**: the change in contacts within 5.2 Angstrom radius

- **Statistical Entropy**: the change in statistical entropy as defined by $S = k_b(\ln p_f - \ln p_i)$. The probabilities p are derived from a multiple sequence alignment and indicate the statistical frequency of either the wildtype or mutant amino acid existing at that position.
- **Pairwise Correlations**: from the Direct Coupling Analysis (DCA) and Hopfield-Potts model, we can calculate the change in pairwise correlations between the amino acids being mutated with all other amino acids in the protein.

Methods



Collect Snase and Barnase Features.
Using ProTherm and Python, find the ASA, hydrophobicity, sec. structure, contacts, entropy, and pairwise correlations

Find Parameters for Random Forest.
Use RandomSearchCV to find parameters to run the RF Classifier

Run Random Forest Classifier

Method 1

Train on Snase

Predict Barnase

Method 2

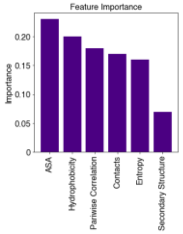
Train on 80% of Snase data

Test on 20% of Snase data

Accuracy & MCC
- Accuracy is the proportion of correctly classified instances
- MCC is the correlation between true and predicted values

Results

The Random Forest Classifier generates values for feature importance.



Method 1:

Accuracy	/0.67
MCC	0.35

Method 2:

Accuracy:	0.78
MCC:	0.59

. This is within the range of other published accuracies but uses less structural information.

Limitations

- The RFC was trained and tested on a limited amount of data. This is due to the unavailability of high quality, experimental ΔG information. Including a larger variety of proteins and protein type may produce a classifier that is more specific and more accurate.
- The classification of stabilizing vs destabilizing was based on the mean of the free energies for each protein. This is different than drawing the boundary at ΔG = 0.
- Datasets would require information about contacts, which would require a crystal structure of the protein to enumerate them. However, this feature may not be necessary for good classification.

Conclusions

Training and testing a Random Forest Classifier on structural and statistical information yields accuracies between 0.67 and 0.78. This is within the range of other published accuracies but uses less structural information. With the elimination of contacts but the inclusion of statistical information, this information suggests that a better machine learning algorithm could be generated that does not require more than the sequence information of the protein. **The feature importance also suggests that sequence information is a good predictor of the stability of a protein**. This agrees with recent work on consensus sequences that uses statistics.

Future Direction

Future research would suggest turning this classification problem into a regression problem to predict the value of the ΔG. Removing contact information would result in a machine learning algorithm purely based on statistical information. With more data, a neural network could be used (but this is limited to experimental results).