The Evolution of Logic Circuits for the Purpose of Protein Contact Map Prediction (Short Abstract)

Samuel D. Chapman[1], Christoph Adami[2], Claus O. Wilke[3], and Dukka B. KC[1*]
[1]Department of Computational Science and Engineering, North Carolina A&T State University, Greensboro, NC
[2]Michigan State University, East Lansing, MI
[3]Department of Integrative Biology, The University of Texas, Austin, TX
[4]Corresponding Author
Email: sdchapma@aggies.ncat.edu, adami@msu.edu, wilke@austin.utexas.edu, dbkc@ncat.edu

Proteins are important biological molecules that perform many functions in an organism. These molecules are composed of a string, or sequence, of amino acids comprised of a 20-letter "alphabet" of amino acids. This sequence largely determines the structure of a protein, and structure in turn plays a large part in determining a protein's function. Thus, determining the structure of a protein can aid in understanding its function and also in using this knowledge in applications such as elucidating evolutionary relationships and drug discovery.

Predicting protein structure from sequence remains a major open problem in protein biochemistry. One component of predicting complete structures is the prediction of inter-residue contact patterns (contact maps). In this work, we present a novel method for protein contact map prediction based on the evolution of deterministic *Markov networks*, or deterministic logic circuits. Our method is composed of two main parts. First, the Markov networks act as the predictors of the contact map. That is, they take feature data from the proteins, which are used as input, and as output produce the contact predictions. Second, the training of the Markov networks is done using an evolutionary algorithm that evolves the Markov networks in a population according to a fitness function, which scores them on how well they predict the contact map of a set of training examples. Over time, the Markov networks in the population become better at predicting contact maps, and at the end of a run, the highest-fitness individual from a population is tested on a testing set of contact examples.

The feature data used is composed of several hundred thousand training and testing examples that each have 688 features associated with them. Each of these examples represent a single possible contact between a pair of amino acids, and the features are either continuous (can take any value), or binary (are a zero or one). Because Markov networks must take binary values as input, we use several different *encodings* (treatments) that each transform the input features in a different way. These encodings split the continuous features into a set number of bins based on the range of values; these included splits of 4, 10, and 16, that used that many bits per feature, and splits of 4 and 16 compressed into 2 and 4 bits/feature, respectively. For each encoding, we evolve 60 different populations of Markov networks for 100,000 updates (evolutionary mutation/reproduction cycles) and take the highest-fitness individual from each population, creating a *committee* of networks that predict each testing example. The final answer on each testing example is the majority vote of the 60 committee members.

To our knowledge, the evolution of Markov networks (deterministic logic circuits) has not been used in contact map prediction, or in any Bioinformatics problem. We show that this evolution of Markov networks is capable of producing networks that can predict a protein's contact map with reasonable performance, based on a measure, Fmax, that takes into account both the proportion of correctly-predicted contacts (specificity) and total correct contacts (sensitivity). We examine the performance differences between different encodings, the number of updates of the evolutionary runs, and the number of committee members used. Because of the successful demonstration of our method, we expect that in the future, improvements can be made, such as by using a better evolutionary algorithm, encoding, or feature type.

During evolution, Markov networks can evolve to use any of the 688 features as input. Thus, we are able to determine which features any network uses during evolution. Because we evolve 60 different populations for each treatment, we can count the number of networks that evolved to recognize each feature and use this as a proxy to determine the importance of a feature in determining a contact map. We look more closely at the results from one treatment and show that certain features are represented much more than others in the 60 members.

In addition, we take the most-represented features and rerun this treatment using only those features, ignoring the rest. The resulting performance from the new run is almost as good as that using all 688 features. This observation shows that evolution of Markov networks can, in an unbiased manner, produce networks that recognize relevant (salient), useful features from a dataset. Choosing these salient features could help to remove extraneous features, a task that would otherwise be computationally intensive. Also, one could use this feature information as a prediction tool to guess which kinds of features are useful or not. Furthermore, this strategy can in principle be used on other types of data and problems in addition to contact map prediction.