Analysis of Pairwise Dependency Information Content for Representing and Searching for Transcription Factor Binding Sites

2013 NSF Bio-Grid
REU Research Fellows
University of Connecticut
Rachel Busch (University of Massachusetts Amherst)
Jacob Mauthe (University of Maine, Orono)
Erin Penrod (University of Tennessee at Chattanooga)

Transcription factor binding sites are areas in a genome in which the sequence of DNA bases attract transcription factors, which then induce a specific gene to be transcribed. There are many different transcription factors, and each one has binding sites with which it can interact. Binding sites for a specific transcription factor are often the same in length, and follow a general pattern of bases. Since they do not follow an exact and specific order of bases, it can be difficult to identify transcription factor binding sites in segments of DNA where the function of said DNA is unknown. In an attempt to correctly identify unknown sequences as binding sites, the machine learning project compares the sequence to be tested against a list of known binding sites. The ability of the algorithm to discern likely binding sites will aid in narrowing down the possibilities for the wet lab biologists to test.

Scoring methods to evaluate transcription factors have already been used in the past. Osada et al used the consensus, centroid, position-specific scoring matrices, and Berg and von Hippel methods to assess transcription factors (Osada, Comparative analysis of methods for representing and searching for transcription factor binding sites). Osada factored in an entropy function in the form of an information content score as a multiplier to more accurately determine the effectiveness of different methods, which made consensus method competitive with the others. However, the approach was limited to binding sites all of the same length, and pairwise dependencies were not thoroughly explored. This project intends to use a new entropy function and investigate the predictive power that can be found when including pairwise dependencies.

The process of making a scoring matrix and then using leave-one-out cross-validation to construct ROC curves showing the accuracy of different sensitivities will be used. The same known binding sites used by Osada will be used. His dataset consists of all the known Escheria coli transcription factor binding sites for Saccharomyces cerevisiae, Drosophila melanogaster, mouse, and human, a total of 35 transcription factors and 410 known binding sites. As a result, known sites of the same length will be used. The known negatives will be the known binding sites of other transcription factors, so they will not be the same length as the binding sites that made up the scoring matrix. To account for this, they will be processed through the scoring matrix finding every possible score and taking the highest score. This project will be testing new information content which will see if there is there is a pairwise conditional dependence. In other words, for each pair we will be testing to see if the second base is determined or influenced by the first base. To test this, we will be modifying the source code used by Osada. This research will either support or disprove our assumption of pairwise dependencies using conditional probability. This will narrow down the selection of possible binding sites for the wet lab to test, streamlines the process of identifying definite transcription factor binding sites.