Using Structure and Sequence Information for Predicting Transcription Factor Binding Sites

Kaplan, T.[1,2], Friedman, N.[1] and Margalit, H.[2]

[1] School of Computer Science and Engineering, Hebrew University of Jerusalem

[2] Institute of Microbiology, Hebrew University of Jerusalem

In recent years, vast amounts of genomic data are being accumulated at a rapid rate. These data open new avenues for investigation of transcription regulation, and in particular for identifying DNA binding sites of regulatory proteins. The current approaches for predicting transcription factor binding sites are based on multiple alignments of already known sites. Here we propose a structure-based approach for predicting novel binding sites, applicable even to newly discovered proteins.

Our approach begins with solved protein-DNA complexes of some family of transcription factors. Based on the solved complexes, a structural binding model is derived. This model, together with the sequences of DNA-binding proteins and their DNA targets, allows accurate characterization of amino acid-nucleotide interactions. We use an Expectation-Maximization (EM) algorithm to simultaneously determine the exact binding location for each protein-DNA pair, and to estimate the parameters of the amino acid-nucleotide interactions. These parameters can now be used for prediction. Given a transcription factor of that family, the method allows the determination of a specific nucleotide profile of its binding site, even in the absence of previously known targets.

Here the method is applied to the Cys2His2 zinc-finger family. We show the compatibility of our parameters with experimental results, and demonstrate the predictive power of the algorithm.