

Transcription factor binding sites discovery by structural and molecular interaction field features

Fangping Mu^{1*}, William S. Hlavacek^{2*}

1. Theoretical Biology and Biophysics Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

2. Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

*email: {fmu, [wish](mailto:wish@lanl.gov)}@lanl.gov

An important step in characterizing a genetic regulatory network is to identify, for each transcription factor (TF), its DNA binding sites. Commonly used approaches involve consensus sequences and position-specific scoring matrices, in which DNA binding sites are represented as letter sequences. However, letter sequences are not designed or intended to capture the chemistry and physics of protein-DNA interaction.

TFs usually recognize specific DNA sequences through direct contact between bases and amino acids. TFs can also recognize DNA conformation. To model DNA sequence-dependent conformation, we use six rigid-body parameters to describe the geometry of complementary base pairs and sequential base pairs steps. Nearby bases will affect these parameters. We determine 3D structures for all 3-mers and 4-mers embedded within flanking GC sequences using the NAMD software tool for molecular dynamics based on CHARMM27 force field parameters. The geometry parameters for the middle base are estimated through the average structures. To characterize the sequence-dependent molecular interaction field around DNA, we place a small probe at specific location and the interaction energy between DNA and the probe can be estimated using the molecular force field. Using the average structures of all 3-mers from molecular simulations, we define S as the space around the middle base (see figure 1) and place probes at different locations within S . We recorded the minimum interaction energy when probes are moved within S , which is defined as,

$$P_i = \min_{r \in S} \Phi(r)$$

and the interaction score as,

$$Q_i = \int_{\Delta S_i} \Phi(r) dA(r)$$

where $\Phi(r)$ is the potential at point r . The integration will be performed over the space outside of the DNA backbone, which is defined by molecular surface. We use the GRID software tool to estimate P_i and Q_i using different probes, such as alkyl hydroxyl OH group, methyl CH₃ group, aliphatic neutral amide group, etc. For middle base within all possible 3-mers, the 6 geometrical parameters and the probe-based parameters are computed and tabulated. For the middle base steps within all possible 4-mers, the 6 geometrical parameters are also tabulated.

For a given set of known binding sites for a transcription factor, the sites are transformed into weight vectors for both strains, which are positive examples. We then randomly picked

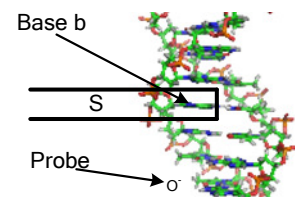


Figure 1. Computation of molecular interaction potential

background sequences from the genome, and transformed these sites into negative examples. Support vector machine (SVM) will be used to build classifiers to distinguish positive from negative examples, and the classification model will be used to scan for novel binding sites.

The method has been implemented for 55 TFs of *Escherichia coli* documented in ReglonDB to have greater than four binding sites. Negative examples are taken from the non-coding sequences of the *E. coli* genome. ~0.2% of the negative examples, randomly chosen, are used in training. SVM with linear kernels are built and these models are then used to predict new binding sites in the genome. For comparison, we also implemented four other methods: the method of Berg and von Hippel (BvH), Match, MATRIXSEARCH and QPMEME. Compared to these methods, the new method produces fewer expected false positives (figure 2).

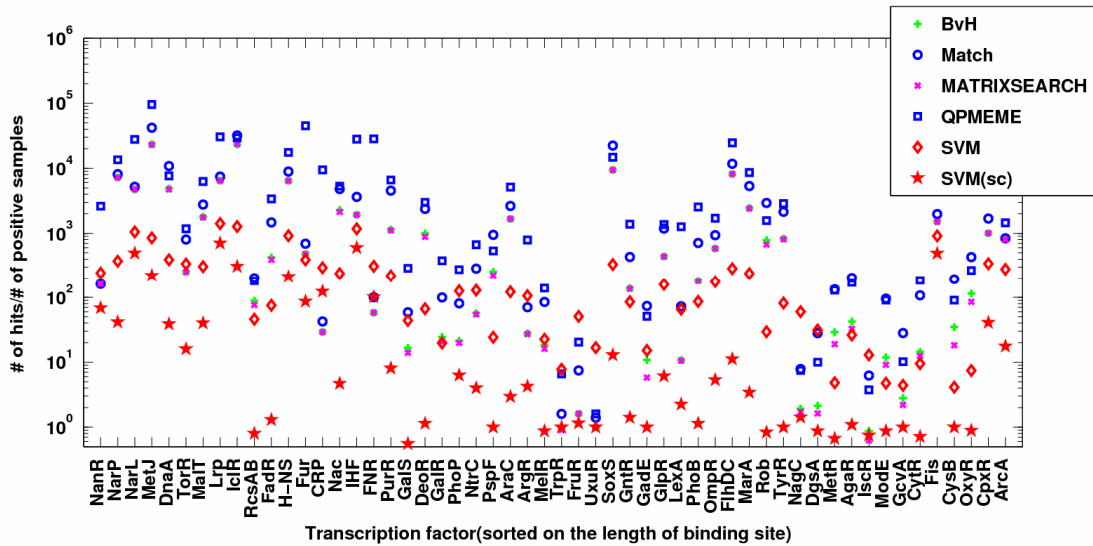


Figure 2. Comparison of four consensus sequences and position-specific scoring method with the proposed method. For different TFs, whose name is displayed in x-axis, # of hits/# of positive samples is plotted. The cutoff values of BvH, Match and MATRIXSEARCH are the mean scores of positive samples. QPMEME is implemented as proposed in the original paper. Two cutoff values are used: SVM ($wx + b > 0$) and SVM(sc) ($wx + b > 1$).