# Using Sequence-Specific Chemical and Structural Properties of DNA to Predict Transcription Factor Binding Sites

Amy L. Bauer[1], William S. Hlavacek[1], and Fangping Mu[1]

***Short Abstract* — An important step in understanding gene regulation is to identify, for each transcription factor, all of its DNA binding sites. Commonly used word-matching methods are based on consensus sequence definitions and position-specific scoring matrices. Here, we take a machine learning approach to transcription factor binding site prediction in which we train classifiers to discriminate between true and false binding sites based on sequence-specific chemical and structural features of DNA. Using *Escherichia coli* data available in RegulonDB and the literature, analysis reveals that our method produces fewer false positives than word-matching approaches without sacrificing accuracy.**

***Keywords* —binding site, *E. coli*, gene regulation, position weight matrix, transcription factor.**

## I. Purpose

One important step in understanding the role of at transcription factor (TF) in gene regulation is to identify its DNA binding sites. Current approaches rely on word matching, such as consensus sequences for preferred nucleotides and position-specific scoring matrices. One drawback of these methods is that they produce a large number of false positive results. The method considered here involves a machine learning approach to TF binding site prediction [1]. A comparison of the new method against popular methods for predicting TF binding sites indicates that the method performs as well as these conventional methods in terms of correctly predicting TF binding sites, but that the method surpasses conventional methods in that it produces a significantly smaller number of false positives.

## II. Method

We train classifiers to discriminate between known true and false binding sites based on sequence-specific chemical and structural features of DNA. We propose a novel representation for TF binding sites using geometric parameters for DNA sequences and features of a sequence's molecular interaction field. We model each DNA base in the context of all possible local neighborhoods to determine features that reflect the three-dimensional structure and the molecular interaction field of each sequence. These features are derived from molecular dynamics calculations and are tabulated. Known binding sites are taken from RegulonDB [2] and used as positive examples. Negative examples are randomly selected from non-coding regions of the *E. coli* genome sequence. Each example is represented as a vector of features and is used to build support vector machine (SVM)-based classification models for the transcription factor. These models are then used to scan the entire non-coding portion of the genome to predict new binding sites.

We apply the method to predict Fis binding sites in *E. coli* and compare our method against the following other transcription factor binding site prediction methods: Berg and von Hippel (BvH), Match, MATRIXSEARCH and QPMEME [3-6]. Experimental assays of Fis binding to DNA in *E. coli* is used to evaluate the accuracy of each method [7]. We find that BvH and MATRIXSEARCH correctly predict .46% binding sites, SVM captures .44%, MATCH .42%, whereas QPMEME only produces a .38% match rate. However, in terms of the number of false positives, our method produces almost 100,000 fewer than the next best method, MATRIXSEARCH.

## III. Conclusion

A novel method for predicting TF binding sites on the basis of the chemical and structural features of DNA produces fewer false positives than word-matching approaches without sacrificing accuracy. These findings provide motivation to further explore and develop the method.

Acknowledgements: This work was funded by NIH grant XX00000.

[1]Theoretical Biology & Biophysics, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, 87545 USA. fmu@lanl.gov

## References

[1] Bauer AL, Hlavacek WS, Mu F "Using Sequence-Specific Chemical and Structural Properties of DNA to Predict Transcription Factor Binding Sites," in preparation.

[2] RegulonDB web site, http://regulondb.ccg.unam.mx/

[3] Berg O, von Hippel P (1987) Selection of DNA binding sites by regulatory proteins. *J. Mol. Biol.* **193**, 723-750.

[4] Kel AE, et al. (2003) MATCH™: A Tool for Searching Transcription Factor Binding Sites in DNA Sequences. *Nucleic Acids Res.* **31**, 3576-3579.

[5] Chen QK, Hertz GZ, Stormo GD (1995) MATRIXSEARCH 1.0: A Computer Program that Scans DNA Sequences for Transcriptional Elements Using a Database of Weight Matrices. *Bioinfiormatics* **11**, 563-566.

[6] Djordjevic M, Sengupta AM, Shraiman BI (2003) A Biophysical Approach to Transcription Factor Biding Site Discovery. *Genome Res.* **13**, 2381-2390.