Transcription Factor Binding Sites:
Position-Specific and Position-Dependent Modeling

Mark Connell[1,5] and Panayiotis V. Benos[2,3,4]

[1]Bioengineering and Bioinformatics Summer Institute, [2]Center for Computational Biology and Bioinformatics, [3]Department of Human Genetics, [4]University of Pittsburgh Cancer Institute, University of Pittsburgh, Pittsburgh, PA 15261
[5]Department of Biomedical Engineering, Duke University, Durham, NC 27708

6/17/04

Transcription of a gene is cued by the binding of a protein to a binding site. Transcription factors bind to specific binding sites (TFBSs), of which there may be many for a single transcription factor (1). However, these TFBSs often exhibit a considerable amount of variability, as the sequences consist of similar nucleotides rather than complete replicas of a TFBS (1). The ability to model these sites is important in determining other possible TFBSs given the rapid increase of sequenced DNA (2).

A common model for TFBSs is the position-specific scoring matrix (PSSM), in which the preference for each nucleotide at each site is recorded (3). This model assumes that each position in the binding site contributes equally to the overall protein-DNA binding affinity, known as the additivity assumption (4). Benos et al (4) have shown that for the most part, that additivity assumption "provides a very good approximation of the true nature of the specific protein-DNA interactions", and are thus useful. While this model does not represent complex dependencies, it requires a lesser amount of data and is efficient (3).

To discover possible TFBSs given a genomic sequence, information content can be used to judge the binding energies for a collection of sites, as shown by Berg and von Hippel (5). Using relative information content provided a more accurate means for measuring binding energies by accounting for the genomic base probabilities (2).

In my research, I aim to use mutual information content to explore positional dependencies using sets of known transcription factors. Mutual information content (MIC), a representation of entropy, judges whether or not two random events are

related. In the case of TFBSs, the MIC compares the frequency of a dinucleotide (one from each position) to the probability of each of the nucleotides at their respective positions. Thus, the MIC will detect occurrences when the frequency of a dinucleotide is greater than simply the probability of each base multiplied.

In order to understand the significance of MIC values, p-values will be determined for each. To see whether or not the null hypothesis (nucleotides are independent) holds, a distribution will be constructed using the probabilities of each base at a given position from the known binding sites. Using these probabilities, many sets of dinucleotides will be made and the MIC will be calculated for those positions. Using these values as a distribution, the initial MIC value for the known binding sites will be used to find the p-value for that MIC value; depending on the results, an appropriate p-value will be chosen. Using this information, a model composed of both PSSM characteristics and conditional probabilities for those positions with significant MIC values.

Finally, this model will then be tested using genomic sequences. This representation's efficiency will be judged by its ability to identify known binding sites. Selectivity and specificity of the model will be calculated and compared to other models (PSSM) to gauge the model's improvements.

Ideally, this representation will locate all known TFBSs and detect unknown sites as well, while at the same time limiting the count of false positives. The representation should strongly detect actual TFBSs while ignoring non-binding sites. This representation can then be used to scan genomic sequences in search of other possible binding sites along the sequence. While false positives (for TFBSs) will occur biologically and due to inaccurate modeling, ideally these false positives can be solely credited to biological characteristics.

References

1. King,O.D., and Roth,F.P. (2003) A non-parametric model for transcription factor binding sites. *Nucleic Acids Res.*, **31** e116 (1-8).
2. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23
3. Barash,Y., Elidan,G., Friedman,N. and Kaplan,T. (2003) Modeling dependencies in protein-dna binding sites. In Vingron,M., Istrail,S., Pevzner,P. and Waterman,M. (eds), *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*. ACM Press, New York, NY, in press. Available at http://www.cs.huji.ac.il/labs/compbio/TFBN/
4. Benos,P., Bulyk,M.L. and Stormo,G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, **30**, 4442-4451.
5. Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723-750.