

Building Common Patterns for Transcription Factor Binding Sites of *Escherichia coli*

Wei Jiang Li

wjlee@bionfo@hotmail.com

Hiroyuki Kurata

kurata@bse.kyutech.ac.jp

Department of Biochemical Engineering and Science, Kyushu Institute of Technology,
680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

Keywords: transcription factor, DNA binding site, pattern

1 Introduction

Only a small part of transcription factors (TFs) and their DNA binding sites have been experimentally identified. A number of algorithms were developed to identify additional members of known binding site families [2], or even to infer new binding sites recognized by unknown TFs [4, 5]. Whether cross-species information of orthologous genes is used or not, most methods are essentially based on the sequence similarities within same binding site families.

It can be found, however, that for some TFs, the corresponding binding sites have diverse sequence patterns. Commonly used recognition matrix method may produce a lot of false-positive predictions for these TFs. The binding sites are indeed a sort of so unique sequences that the TFs can successfully recognize. Are there common characteristics (words, patterns) shared by many binding sites belonging to different TFs? Can the binding sites be divided into small building blocks?

It is known that there are shared motifs in eukaryotic regulatory sequences [1]. In *E. coli* genome, it can also be easily found by a simple search that some long words occur in different binding sites. But individual words are not adequate to describe these common features, since the most active words appear in only a very limited number of binding sites. Therefore we need to extend individual words into patterns comprised of several related individual words. It should be noticed that building patterns is in fact a tradeoff between adaptability and specificity: We need patterns to appear in various binding sites as often as possible. On the other hand, a pattern should be distinct enough to contribute to the characteristics of binding sites. This work proposes a method to build such patterns.

2 Method and Results

2.1 Data

Binding site sequence data were taken from [3] (http://arep.med.harvard.edu/ecoli_matrices/). There are now 68 known TFs and 802 corresponding binding sites from 10bp to 50bp long. To prevent possible errors, we selected the binding sites that occur only once in the whole genome. There are 757 such sites. *E. coli* (strain K12) genome sequence data were taken from GenBank (U00096). The upstream regions are 450bp long from the starting codons.

2.2 Method

Generally, the shorter a pattern is, the more adaptable it is. But the specificity of a short pattern is too low to reflect the features of binding sites. In this work we chose the pattern length to be 7. Each pattern can be seen as a regulation to describe the binding site sequences. We construct more adaptable (complex) patterns from less adaptable (simple) ones by merging simple patterns gradually. At the beginning, the patterns are simply the individual words that occur in the binding sites. We want to merge these patterns to simplify the regulations.

Patterns are denoted using International Union of Pure and Applied Chemistry symbols. Two patterns are merged by simply apply the “OR” operation to each elements of the patterns. For

example, merging GAACATG and TAACATA yields pattern KAACATR (K=G or T, R=A or G). We define a pattern's significance by

$$S(\sigma) = N_b(\sigma)/N(\sigma)$$

where $N_b(\sigma)$ and $N(\sigma)$ are the numbers of occurrences of pattern σ in all the binding sites and in the all upstream regions, respectively. A pattern with high significance means that it tends to appear in binding sites. Or, in other words, a sequence fragment containing a high significance pattern has a high possibility to be a binding site.

We use a two stage algorithm to build the patterns. At the first stage, we look for every possibility to produce an "absolutely" better pattern by merging two existing patterns. The algorithm is as follows.

1. For each pair of patterns σ_i and σ_j , calculate the significance of merged pattern $S(\sigma_{ij})$, record this value if $S(\sigma_{ij}) \geq \max(S(\sigma_i), S(\sigma_j))$
2. If there is no recorded value, stop searching;
3. Merge the patterns that yield the pattern with highest recorded value;
4. Eliminate redundant patterns and repeat these steps.

At the second stage, we apply the same algorithm but with more relaxed merging criterion: record the significance value if $S(\sigma_{ij}) \geq \max(0.02, 0.8\max(S(\sigma_i), S(\sigma_j)))$.

2.3 Results

We began with a comprehensive table of all 7-letter words that appear in the known binding sites. The number of such words is 7701. But most of these words appear only once in all binding sites. After the first stage merging, these individual words were merged into 4317 patterns. When the second stage merging finished, the number of patterns were further compressed to 3438. In this final set of patterns, there are many independent words that appear very few times in all binding sites and they are so rigid that no other patterns can absorb them under the relaxed merging criterion. If the patterns with either low significances (<0.02) or small occurrence numbers in binding sites (<10) are excluded, there are only 994 patterns. All binding sites (but only 3 exceptions) contain at least two patterns (overlap allowed) in this set.

3 Discussion

This work focuses on common building blocks of unaligned binding sites. We have deduced a small set of patterns that can be used to describe the most known binding sites. It shows that various binding sites can be delineated in a united framework. The common features are useful in predicting new binding sites, even for unknown TFs.

References

- [1] Bussemaker, H.J., Li, H., and Siggia, E.D., Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis, *Proc. Natl. Acad. Sci. USA*, 97:10096–10100, 2000.
- [2] Collado-Vides, J., Moreno-Hagelsieb, G., and Medrano-Soto, A., Microbial computational genomics of gene regulation, *Pure and Applied Chemistry*, 74:889–895, 2002.
- [3] Robison, K., McGuire, A.M., and Church, G.M., A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K12 genome, *J. Mol. Biol.*, 284:241–254, 1998.
- [4] Thieffry, D., Salgado, H., Huerta, A.M., and Collado-Vides, J., Prediction of transcription regulatory sites in the complete genome of *Escherichia coli*, *Bioinformatics*, 14:391–400, 1998.
- [5] Van Nimwegen, E., Zavolan, M., Rajewsky N., and Siggia, E.D., Probabilistic clustering of sequences: Inferring new bacterial regulons by comparative genomics, *Proc. Natl. Acad. Sci. USA*, 99:7323–7328, 2002.