64. A Heuristic Approach for Finding Optimal Motifs for Protein Binding Sites

Osamu Maruyama [1], Daisuke Shinozaki [2], Satoru Kuhara [3]

Keywords: motif, word-enumeration, pruning, co-regulated genes

1 Introduction

One of the important issues in bioinformatics is to find potential promoter regulatory elements in the upstream regions of co-regulated genes. Those regulatory elements are so-called "transcription factor binding sites", which are important factors in understanding the mechanisms of regulation of transcription and gene expression. For each transcription factor, the sequences of its binding sites are often approximately conserved across the upstream regions of the regulated genes. A *motif* means a pattern consistent with different binding sites of a transcription factor. Determined genome sequences of various organisms and microarrays let us have a great number of collections of putatively co-regulated genes. Thus, it is urgently needed to establish a method for finding convincing motifs computationally.

There are three major models for motifs for binding sites: position weight matrix, oligonucleotide with some mismatches, and pattern over IUPAC nucleic acid codes. Position weight matrices are mainly used in probabilistic methods, especially, methods based on Expectation Maximization (EM) or Gibbs sampling. The methods for finding significant oligonucleotides with some mismatches have been investigated in [1]. In [4, 3], some methods for finding optimal patterns over IUPAC nucleic acid codes with *specific* score functions have been proposed.

2 Method

In this work, we have investigated a heuristic word-enumeration method for finding optimal pattern over IUPAC nucleic acid codes, which works for any conic functions defined in [2]. It should be noted here that most of the reasonable score functions like the entropy information gain are conic. In the algorithm, a pruning technique [2] is employed for pruning sub-spaces of such patterns. Our algorithm enumerates those patterns from specific to general. First, the algorithm extracts all of the oligonucleotides of a specific length, and generates new patterns over IUPAC nucleic acid codes from existing ones while pruning insignificant patterns.

3 Computational Experiments

We apply our method to co-regulated genes of yeast, and report successful experimental results which detect several known transcription factor binding sites in a promoter database of the yeast Saccharomyces cerevisiae (SCPD)[5].

[1] Faculty of Mathematics, Kyushu University, Fukuoka 812-8581, Japan. om@math.kyushu-u.ac.jp
[2] Faculty of Mathematics, Kyushu University, Fukuoka 812-8581, Japan. ma202035@math.kyushu-u.ac.jp
[3] Graduate School of Genetic Resources Technology, Kyushu University, Fukuoka 812-8581, Japan. kuhara@grt.kyushu-u.ac.jp

Acknowledgments

This study was supported from the Research for the Future Program of the Japan Society for the Promotion of Science.

References

[1] Eskin, E., and Pevzner, P. A. 2002. Finding composite regulatory patterns in DNA sequences. *Bioinformatics 18*, supplement 1, S354–S363.

[2] Shinohara, A., Takeda, M., Arikawa, S., Hirano, M., Hoshino, H., and Inenaga, S. 2001. Finding best patterns practically. In *Progress in Discovery Science*, S. Arikawa and A. Shinohara, Eds., Springer-Verlag Berlin Heidelberg, pp. 307–317.

[3] Sinha, S. Discriminative motifs. 2002. In *Proceedings of the 6th Annual International Conference on Computational Biology*, pp. 291–298.

[4] Sinha, S., and Tompa, M. 2000. A statistical method for finding transcription factor binding sites. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, pp. 344–354.

[5] Zhu, J., and Zhang, M. Q. 1999. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics 15*, 7/8, 607–611.