

Target Prediction of Transcription Factors: Application of Structure-Based Method to Yeast Genome

Akinori Sarai¹

sarai@rtc.riken.go.jp

Samuel Selvaraj²

selvaraj@rtc.riken.go.jp

Michael M. Gromiha³

gromiha@rtc.riken.go.jp

Joerg-Gerald Siebers¹

siebers@rtc.riken.go.jp

Ponraj Prabakaran¹

praba@rtc.riken.go.jp

Hidetoshi Kono⁴

kono@apr.jaeri.go.jp

¹ RIKEN Tsukuba Institute, 3-1-1 Koyadai, Tsukuba 305-0074, Japan

² Bharathidasan University, Tiruchirapalli 620 024, Tamilnadu, India

³ Computational Biology Research Center, AIST, 2-41-6 Koto-ku, Tokyo 135-0064, Japan

⁴ Department of Health Physics, JAERI, 8-1, Umemidai, Kizu-cho, Souraku-gun, Kyoto 619-0215, Japan

Keywords: transcription factor, target genes, structural information, yeast genome

1 Introduction

Complete genome sequences of many organisms have become available and the functional analysis of genomes will be a target of intensive research. Gene regulation in higher organisms is one of the most important biological functions, and it is achieved by a complex system of transcription factors. Transcription factors usually bind to multiple target sequences and regulate multiple genes in a complex manner. Finding target genes for transcription factors at the genome level will lay a basis for the analysis of the gene regulatory network. We have been developing methods for predicting target sequences of transcription factors. Structure-based methods, which utilize structural data of protein-DNA complexes, are promising candidates for DNA target prediction. Here we apply this method to the target prediction of transcription factors in the whole yeast genome.

2 Methods

The structure-based method is based on the analysis of a structural database of protein-DNA complexes. We derived empirical potential functions for the specific interactions between bases and amino acids from the statistical analysis of the structural data [1]. Then these statistical potentials were used to evaluate the fitness of sequences to the complex structures of particular transcription factors by a combinatorial threading procedure similar to the fold recognition of protein structures, i.e., finding amino acid sequences that fold into a particular structure. The summation of individual potential functions between bases and amino acids in a particular protein-DNA complex gives a total interaction energy. By threading a set of random DNA sequences through the template structure, we could calculate the Z-score of the specific sequence against the random sequences, which represent the specificity of the complex. When the threading procedure was applied to the real genome sequence, we could find potential target sites of transcription factors.

The above method is based on the direct readout mechanism, in which protein recognizes DNA sequence through the direct contact between amino acids and base pairs. On the other hand, substitutions of those base pairs not in contact with amino acids often affect binding affinity, indicating that protein may recognize DNA sequence through sequence-dependent conformation and properties

of DNA. This in-direct readout mechanism may significantly contribute to the specificity of protein-DNA recognition. Thus, we have attempted to quantify the specificity due to this in-direct readout mechanism [4]. In order to estimate the conformational energy of DNA, we considered six coordinates (three translational and three rotational coordinates) to describe the conformation of each base step. We approximated the conformational energy by harmonic potentials along each coordinate. The equilibrium conformation and empirical force constants were determined by the analysis of protein-DNA complex structures [2]. By using these potentials, the total conformational energy of the DNA was estimated for a given structure, and the threading procedure was used to evaluate the fitness of the sequence to the DNA structure. We then calculated the Z-score for the in-direct recognition in the same way as for the direct recognition.

Both energies from direct and in-direct readout mechanisms were summed up to give a total interaction energy of the protein-DNA complex. However, because the direct and in-direct readout energies are based on different statistics, we need to combine them with a weighting coefficient. The weighting factor was determined by maximizing the total Z-score. Because both the potentials are independent quantities, we expect that the total Z-score increases compared to the individual contributions. Finally, we used the total interaction energy for threading the real yeast genome sequences to find target sites.

3 Results and Discussion

MATa2/MCM1 is a pair of transcription factors involved in the determination of mating types of yeast. The structure of the protein-DNA complex was determined and some experimental analyses have been carried out for determining the target genes. Thus, we used MATa2/MCM1 as an example for target prediction. We used 63 non-redundant protein-DNA complexes to derive the statistical pair potentials. We calculated the Z-scores for direct and in-direct readout mechanisms for the MATa2/MCM1/DNA complex. We summed the two potential energies to calculate the total energy. The weighting coefficient was varied and the total Z-score was calculated from the total energy. The combination of in-direct readout potential energy indeed increased the Z-score, and we took the weighting coefficient that gave the largest Z-score.

The promoter regions of candidate target genes were examined by threading and a Z-score was calculated. The target genes were ranked by the Z-score and compared with experimental data [6]. The target genes identified positively by experiment were ranked high in the list, and the experimentally negative genes were ranked low. Separation between the positive and negative genes was not perfect but they were segregated by a certain threshold Z-score value. The total Z-score gave better separation than that of direct contribution alone.

The present result shows that the structure-based method can be used for predicting target genes of transcription factors in the yeast genome. The present method complements other prediction methods such as sequence-based methods, and, together with genome annotation information, the accuracy will be significantly enhanced by reducing false positives and false negatives. The structure-based method relies on known structural data of the protein-DNA complex, but the structure can be predicted by using structure prediction methods like homology modeling. Thus, in combination with the structure prediction methods, its range of application can be significantly enhanced, and it is the only approach that enables one to predict targets of new transcription factors without carrying out further experiments. Another merit of structure-based method is that it enables us to examine the quantitative relationship between structure and specificity in protein-DNA recognition, such as the effects of cooperativity [1, 3], cognate/non-cognate binding [1, 5], asymmetric binding [5], and DNA deformation on the specificity. Furthermore, in the light of ever increasing structural data and progress in structural genomics, the approach is of growing interest.

References

- [1] Kono, H. and Sarai, A., Structure-based prediction of DNA target sites by regulatory proteins, *Proteins*, 35:114–131, 1999.
- [2] Olson, W.K., Gorin, A., Lu, X., Hock, L., and Zhurkin, V., DNA sequence-dependent deformability deduced from protein-DNA crystal complexes, *Proc. Natl. Acad. Sci. USA*, 95:11163–11168, 1998.
- [3] Sarai, A. and Kono, H., DNA-protein interactions: Target predictions, in *Handbook of Computational Biology*, Marcel Dekker Inc., New York, 2002.
- [4] Sarai, A., Selvaraj, S., Gromiha, M.M., Siebers, J.-G., Prabakan, P., and Kono, H., Target prediction of transcription factors: Refinement of structure-based method, *Genome Informatics*, 12:384–385, 2001.
- [5] Selvaraj, S., Kono, H., and Sarai, A., Specificity of protein-dna recognition revealed by structure-based potentials: Symmetric/asymmetric and cognate/noncognate binding, *J. Mol. Biol.*, in press.
- [6] Zhong, H., McCord, R., and Vershon, A.K., Identification of target sites of the a2-Mcm-1 repressor complex in the yeast genome, *Genome Res.*, 9:1040–1047, 1999.