Structure-Based Target Prediction of Transcription Factors

Akinori Sarai[1] **Hidetoshi Kono**[2]
sarai@rtc.riken.go.jp kono@sas.upenn.edu

[1] RIKEN Life Science Center, 3-1-1 Koyadai, Tsukuba 305-0074 Japan
[2] Department of Chemistry, University of Pennsylvania, 231 South 34 St., Philadelphia PA 19104 USA

Keywords: protein-DNA recognition, base-amino acid interaction, statistical potential, sequence-structure threading, structure-sepcificity relationship, computer simulation, target prediction

1 Introduction

Regulation of gene expression in higher organisms is achieved by a complex network of transcription factors and their target genes. Explosive amount of sequence information of genes and transcription factors from genome analyses are presenting a great challenge to bioinformatics. Finding target genes for transcription factors at the genome level will lay a basis for the analysis of gene regulatory network. Sequences contain rich amount of biological information, and comparison of known binding sequences is among the most commonly used methods for the target prediction. However, the accuracy of this method is rather limited, dependent on the quality of sequence information used. Furthermore, because transcription factors usually bind to multiple target sequences and regulate multiple genes in a cooperative manner, the target prediction is a rather complicated problem. In order to tackle such problem, we need to utilize as much information as possible. Structural data on protein-DNA complex contain valuable functional information as well. Due to the progress of X-ray crystallography and NMR spectroscopy techniques, structural data on the protein-DNA complex have been rapidly increasing and more than 200 complexes have been registered in Protein Data Bank. Here, we show that these structural data can be used for the target prediction. Because the structural information is independent of sequence information, it can complement the sequence-based method.

2 Method and Results

The structure-based method is based on the statistical analysis of structural database of protein-DNA complex. Empirical potential functions for the specific interactions between bases and amino acids can be derived and used to evaluate the fitness of sequences to the complex structures of particular transcription factors by a combinatorial threading procedure similar to the protein structure prediction [1].

The threading procedure was used to find target sites of transcription factors in real genome sequences. As an example of such applications, we could identify the binding sites of a transcription factor MATa1/α2 in the promoter of *HO* gene successfully.

We can also apply this method to examine the relationship between structure and specificity of target recognition quantitatively. We have examined the effects of DNA deformation, cooperativity and other structural effects on the specificity, and shown that those structural effects have significant consequences to the specificity.

The accuracy of this method for the target prediction is still limited because of the limited numbers of available structural data. In order to complement the method, we are also using computer simulations to derive contact potential between bases and amino acids [2]. The computer simulations, which consider structural flexibility and redundancy in the interaction between amino acid and base pair, require intensive computation time. Thus, we are developing efficient algorithms such as multicanonical Monte Carlo for the sampling of interaction configurations. The interaction "free-energy maps" derived from the calculations for different pairs of base and amino acid have shown different specificity. These data for all the combination of bases and amino acids can be eventually used for the prediction of target sequences. The increase in the structural data, together with the information from computer simulation, will make the structure-based method promising for the accurate target prediction of transcription factors.

References

[1] Kono, H. and Sarai, A., Structure-based prediction of DNA target sites by regulatory proteins, *Proteins*, 35(1):114–131, 1999.

[2] Pichierri, F., Aida, M., Gromiha, M., and Sarai, A., Free energy maps of base-amino acid interaction for protein-DNA recognition, *J. Am. Chem. Soc.*, 121:6152–6157, 1999.