Methods for Predicting Target Sites of Transcription Factors

Akinori Sarai [1]
sarai@rtc.riken.go.jp

Hidetoshi Kono [3]
kono@sas.upenn.edu

Michael Gromiha [1]
gromiha@rtc.riken.go.jp

Fabio Pichierri [2]
fabio@atlas.riken.go.jp

Kenji Sayano [4]
sayano@etl.go.jp

Misako Aida [5]
maida@ipc.hiroshima-u.ac.jp

[1] Life Science Center (Tsukuba)
[2] Computational Science Laboratory (Wako)
RIKEN, 3-1-1 Koyadai, Tsukuba 305-0074 and 2-1 Horosawa, Wako 351-0198 Japan
[3] University of Pennsylvania, 231 South 34 St. Philadelphia PA 19104, USA
[4] Electrotechnical Laboratory, Umezono, Tsukuba 305-8568, Japan
[5] Hiroshima University, Higashi-hiroshima 739-8526, Japan

1 Introduction

Gene regulation in higher organisms is achieved by a complex system of transcription factors. Explosive amount of sequence information and identified transcription factors are presenting a great challenge to bioinformatics. Transcription factors usually bind to multiple target sequences and regulates multiple genes. Finding potential target sites in a vast sequence space is a multiple-minimum problem, similar to the one in protein folding. Thus, similar algorithm may be applied. We are developing new methods for predicting the target sequences. Here we describe the methods and discuss comparison among different methods.

2 Methods

The method for predicting the target sites may be classified according to whether it uses structural information or not. Here, the following four methods are considered:

(1) **Sequence-based method.** It relies on sequence information obtained from known binding sequences. Or, the consensus sequences are derived by random-oligo screening. From the alignment of collected sequences, "weight matrix" is constructed. Then, the weight matrices are used to scan the database for finding potential target binding sites.

(2) **ΔG-based method.** This is based on experimental measurements of binding between protein and DNA. The binding-affinity data for systematic single-base mutations to consensus binding site can be used to derive matrices similar to the weight matrices in the sequence-based method.

(3) **Structure-based method.** This is based on the analysis of structural database of protein-DNA complex. We can derive empirical potential functions for the specific interactions between bases and amino acids from the statistical analysis [2]. Then these statistical potentials are 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.

(4) ***Ab-initio* method.** This method does not rely on any experimental data, but it is based on computer simulations to derive contact potential between bases and amino acids.

3 Results an Discussion

Currently, the sequence-based method is the most commonly used method for the target prediction. The method is quite straightforward but its validity severely depends on the quality of the sequence information. On the other hand, ΔG-based weight matrices will be more reliable than the sequence-based weight matrices, because the data are based on physical interactions. We have compared both the matrices for the case of c-Myb, for which both the data are available. The correlation between both the matrices is very good when the number of aligned sequences is large (the correlation coefficient is larger than 0.9 as the number of sequences is larger than 30). However, the correlation is monotonously decreasing as the number of aligned sequences is decreasing. Thus, sufficient number of sequences will be required for making reliable prediction based on the sequence-based method. The application of ΔG-based method to particular transcription factors showed some success [1]. Nevertheless, it has limitations, because of the complexities in transcription factor system such as cooperativity.

The application of structure-based method to several transcription factors showed that the expected specificity for some transcription factors is quite good but poorer for others. In general, the accuracy of structure-based method for the target prediction is still limited because of the limited numbers of available structural data. However, the power of this method is that we can examine the effects of DNA deformation, cooperativity and other structural effects on the specificity in a quantitative manner [2]. In fact, we could show by using this method that the cooperative binding of two transcription factors and DNA deformation increase the specificity significantly [2]. Also, increase in the structural data will make this method promising. This method can also be applied to proteins of unknown structure having substantial sequence similarity to known proteins, on the basis of which structures can be modeled and binding sites can be predicted.

In order to complement the structure-based method, we are now developing the *ab-initio* method, which derives contact potentials between bases and amino acids by computer simulations. The computer simulations, which consider structural flexibility and interaction redundancy, would require intensive computation time. The interaction “free-energy maps” derived from the calculations for different pairs of base and amino acid have shown different specificity [3]. These data for all the combination of bases and amino acids can be eventually used for the prediction of target sequences.

All the methods discussed here use different algorithms for the prediction, and contain complementary information with one another. Thus, combined together, they would provide a powerful tool for predicting target binding sites and target genes of transcription factors.

References

[1] Deng, Q.-L., Ishii, S., and Sarai, A., Binding-site analysis of c-Myb: screening of potential binding sites by the mutational matrix derived from systematic binding affinity measurements, *Nucleic Acids Res.*, 24:766–774, 1996.

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

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