Developing Genome-Scale Prediction System for Transcription Factors and Their Targets

Akinori Sarai[1] sarai@bse.kyutech.ac.jp
Shandar Ahmad[1] shandar@bse.kyutech.ac.jp
Michael M. Gromiha[2] michael-gromiha@aist.go.jp
M. D. Shaji Kumar[1] shaji@bse.kyutech.ac.jp
Abdulla Bava[1] bava@bse.kyutech.ac.jp
Hidetoshi Kono[3] kono@apr.jaeri.go.jp

[1] Dept of Biochemical Eng. & Sci., Kyushu Institute of Technology, 680-4 Kawazu, Iizuka 820-8502, Japan
[2] Computational Biology Research Center, AIST, 2-41-6 Koto-ku, Tokyo 135-0064, Japan
[3] Neutron Science Research Center and Center for Promotion of Computational Science and Engineering, Japan Atomic Energy Research Institute, 8-1 Umemidai, Kizu-cho, Soraku, Kyoto 619-0215, Japan

Keywords: transcription factor, target genes, genome

1 Introduction

Complete genome sequences of many organisms have become available and the functional analysis of genomes is 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. The genome analyses show that in most species about a half of the genes is of function unknown. Many of the genes may turn out to code for transcription factors. Subsequent functional analyses of transcription factors involve identification of their target genes. Transcription factors usually bind to multiple target sequences and regulate multiple genes in a complex manner. Thus, identifying transcription factors and finding their target genes at the genome level will lay a basis for the analysis of the gene regulatory network. We have been developing methods for predicting transcription factors and their targets based on various kinds of information ranging from sequence to structure. We are developing an integrated genome-scale prediction system by combining those methods together.

2 Strategy

Figure 1 illustrates the prediction scheme. New genes obtained from genome analyses are filtered through a module for predicting transcription factors. The module carries out a simple homology search against sequence database, and those genes without known homologues are further analyzed by a prediction algorithm based on sequence and structural information [1]. Currently, the sequence-based method, which relies on sequence information obtained from known binding sequences, is widely used for the target prediction. The method is quite straightforward but its accuracy solely depends on the quality of the sequence information. We have developed an alternative method, which is based on the analysis of a structural database of protein-DNA complexes [2]. In this method, we first derive statistical potentials for the specific interactions between bases and amino acids and for the sequence-dependent conformation of DNA, corresponding to direct and indirect readout mechanisms, respectively, from the statistical analysis of the structural data [2, 3]. Then

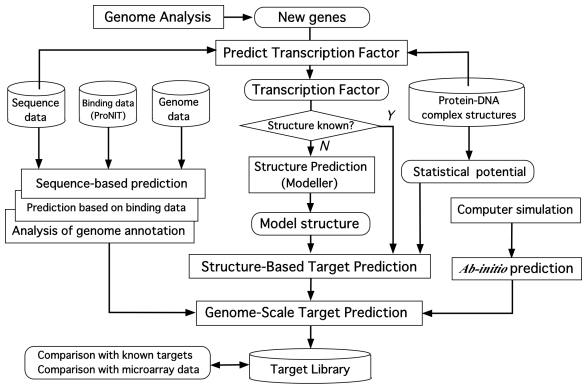

Figure 1: Strategy for genome-scale prediction of transcription factors and their targets.

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. When the threading procedure is applied to a real genome sequence, we can identify potential target sites of transcription factors. Even for transcription factors with unknown structures, we can still use this procedure based on the modeled structure (only C_a coordinates are required in the prediction scheme) if homologous structure is available. This opens up a possibility of predicting target genes of new transcription factors without conducting further experiments. The structure-based method can be combined with other methods; computer simulation of protein-DNA binding, sequence-based method, analysis of experimental binding data, and analysis of various genome annotation data, in order to enhance the accuracy adequate for the genome-scale prediction. The predicted target sites and target genes of transcription factors will be stored in a target library. Then, these data can be used for comparison with the experimentally known target genes and for the systematic analysis of gene expression data from microarrays.

3 System Development

Developing the prediction system includes the following steps: 1) Systematic search against sequence database to find genes homologous to known transcription factors. 2) Prediction of transcription factors if genes have no homology with known transcription factors. 3) Updating protein-DNA complex structure database: First we collect protein-DNA complex structures, carry out all-against-all comparison of protein sequences, and cluster them according to the similarity. One representative complex is selected from each cluster by considering the specificity of interactions involved. These representative complexes constitute a non-redundant database for further analysis. 4) Generation of statistical potential. This module automatically generates statistical potential between bases and amino acids by extracting information about their interactions from the protein-DNA complex database. 5) Structure prediction of transcription factors unless available. Any homologue of the transcription factors of unknown structure is searched for in the protein-DNA complex database. If homologues are found, the structure is modeled and the model structure is fitted into the template complex structure. 6) The protein-DNA complex structure is used as a template for threading procedure, calculating interaction energy for any DNA sequence. 7) The threading procedure is applied to genome sequences to find potential targets of transcription factors. 8) Sequence-based prediction module for the prediction of target sites using sequence information (pattern and profile). In addition, experimental binding data are used, if available, for more accurate prediction. 9) Incorporating available genome annotation information (contextual information such as binding sites of other transcription factors; functional information of target genes; pathway information of protein products from target; evolutional information from comparative analysis of target genes among different species). 10) Developing an integrated Web interface for combining all the analysis modules.

A pilot study on yeast genome has shown that the structure-based method can identify target genes of transcription factors correctly. An integrated prediction system with combination of different kinds of methods would become a powerful tool for analyzing transcription factors and gene network and for providing insight into the mechanism of gene expression regulation.

References

[1] Ahmad, S., Gromiha, M.M., and Sarai, A., Role of composition, sequence and structural information in DNA-binding: analysis and prediction *Bioinformatics*, in press.

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

[3] 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.