GENOME-SCALE PREDICTION OF TRANSCRIPTION FACTORS AND THEIR TARGETS

Sarai A.1, Ahmad Sh.1, Gromiha M.M.2, Kono H.3

1 Dept. Bioscience & Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, 820-8502 Japan; 2Computational Biology Research Center, AIST, 2-41-6 Koto-ku, Tokyo 135-0064 Japan; 3Neutron 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

* Corresponding author: e-mail: sarai@bsc.kyutech.ac.jp

Keywords: transcription factor, target genes, genome

Summary

Motivation: 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. Identifying transcription factors and finding their target genes at the genome level will lay a basis for the analysis of the gene regulatory network.

Results: We have developed an algorithm for predicting DNA-binding proteins based on sequence and structural information. We have achieved accuracies higher than 80%. For the prediction of target sites, we have used a knowledge-based approach, utilizing rapidly increasing structural data of protein-DNA complexes to derive empirical potential functions for the specific interactions between bases and amino acids as well as for DNA conformation, by the statistical analyses of the structural data. These statistical potentials were used to quantify the specificity of protein-DNA recognition, which enabled us to establish the structure-function relationship of transcription factors, such as the effects of binding cooperativity on target recognition. The method was applied to yeast genome sequences, and we could identify target genes of transcription factors successfully. We are also developing an integrated genome-scale prediction system by combining various kinds of methods.

Availability: The prediction system will be made available to the public soon.

Introduction

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. However, the targets of transcription factors are largely unknown. Thus, understanding the molecular mechanism of protein-DNA recognition and its application to genome-wide prediction are essential for the analysis of gene regulation network. We have been developing methods for predicting transcription factors and their targets based on various kinds of information ranging from sequence to structure. Here we present a strategy for genome-scale prediction of transcription factors and their targets based on a combination of these methods.

Methods

In order to predict whether the new protein products derived from genome analysis bind to DNA or not if they do not have known homologues, we have developed a prediction method based on sequence and structural information (Ahmad et al., 2003; Ahmad, Sarai, 2004). For the prediction of targets of transcription factors, we made a statistical analysis of structural database of protein-
DNA complex, and derived empirical potential functions for the specific interactions between bases and amino acids (Kono, Sarai, 1999; Sarai et al., 2001; Kono, Sarai, 2003; Selvaraj et al., 2002). Then, we used a sequence-structure threading to examine the relationship between structure and specificity in protein-DNA recognition. By threading a set of random DNA sequences onto the template structure, we calculated the Z-score of the specific sequences against the random sequences, which represent the specificity of the complex. By threading real genome sequences, we can predict target sites of transcription factors at the genome scale. In addition to this so-called direct or “intermolecular” readout mechanism, by which proteins recognize DNA sequence through the direct contact between amino acids and base pairs, we also evaluated the fitness of DNA sequence against DNA structure to examine the role of so-called indirect or “intramolecular” readout mechanism (hereafter we use the terms “intermolecular” and “intramolecular” readout mechanisms). For this purpose, we have derived statistical potential functions for conformational energy of DNA from the same set of protein-DNA complex structures (Sarai et al., 2001; Gromiha et al., 2004). In another approach, we have analyzed protein-DNA recognition by computer simulations. We have used various kinds of methods: Monte Carlo simulation of base amino acid interactions (Pichierri et al., 1999; Sayano et al., 2000; Yoshida et al., 2002); molecular dynamics of DNA conformation; empirical calculations of interaction energy of protein-DNA complex (Oobatake et al., 2003); molecular dynamics/free energy calculations of protein-DNA complex (Saito, Sarai, 2003); and docking simulation of protein-DNA binding and sliding.

Results and Discussion

In order to predict DNA-binding proteins, we have used amino acid composition, sequence information, solvent accessibility surface information (Ahmad et al., 2003) and electrostatic information such as charge and dipole moment (Ahmad, Sarai, 2004). We have achieved accuracies higher than 80%. These simple and yet accurate methods, together with homology modelling of structures, enable us to predict DNA-binding proteins at the genome scale quickly.

We have derived empirical potential functions for the base-amino acid interactions from the analysis of protein-DNA complex structures. We have compared the structures of cognate and non-cognate protein-DNA complex structures in order to test our method and to understand what is important for specific binding and what is different between them. The statistical potentials could distinguish the two structures as differences in the Z-scores as well as statistical potentials (Kono, Sarai, 1999; Selvaraj et al., 2002). Thus, the subtle differences in specificity of these structures could be detected by our method. We also applied this method to examine the relationship between structure and specificity in cooperative protein-DNA binding. The effect of cooperative binding was examined by comparing the monomer and heterodimer complexes of MAIα1/α2 (Kono, Sarai, 1999), MCM1/MATα2 and NFAT/AP-1 transcription factors. We found that the heterodimer binding enhances the specificity in a non-additive manner. This result indicates that the conformational changes introduced by the heterodimer binding play an important role in enhancing the specificity.

We can calculate Z-score for the intramolecular recognition due to sequence-dependent DNA conformation in the same way as for the intermolecular recognition. By comparing both the Z-scores we can assess the relative contributions of intermolecular and intramolecular readout mechanisms. We have analyzed various protein-DNA complexes systematically, and found that both the intermolecular and intramolecular mechanisms make significant contribution to the specificity (Sarai et al., 2001; Gromiha et al., 2004). The relative contributions depend on the types of DNA-binding proteins. Because both the potentials are independent quantities, they can be summed up to calculate the total energy and used to find target sites, although a weighting factor needs to be determined as the two potentials were derived from different statistics. We found that the Z-score was indeed enhanced compared with individual Z-scores for intermolecular
or intramolecular readout alone. This result indicates that the energies of the intermolecular and intramolecular readouts contain independent information that in combination enhances the specificity of the recognition. We used the combined energy for threading against DNA sequences to make target predictions for transcription factors.

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 experimentally-verified binding sites of the transcription factor MATa1/α2 in the promoter of HO gene successfully (Kono, Sarai, 1999). We have also attempted to identify target sites and genes of MCM1/MATα2 in the whole yeast genome. The target genes of this transcription factor have been identified in yeast genome experimentally (Zhong, Vershon, 1997). The predicted target genes were ranked by the Z-score and compared with experimental data. The target genes identified positively by experiment were ranked high in the list, and the experimentally negative genes were ranked low (Sarai et al., 2004). 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.

In order to complement the knowledge-based approach described above, we also performed various kinds of computer simulations (Pichierri et al., 1999; Sayano et al., 2000; Yoshida et al., 2002; Oobatake et al., 2003; Saito, Sarai, 2003). These simulations could derive interaction potentials between bases and amino acids equivalent to the statistical potentials, and provided insight into the molecular mechanism of specificity in protein-DNA recognition. We are now combining knowledge-based approach and computer simulations, with other methods based on sequence information and experimental binding data (Deng et al., 1996) and genome annotation information, to develop an automated prediction system. Such integrated prediction system with combination of different kinds of methods would become a powerful tool for analyzing transcription factors and for providing insight into the mechanism of gene expression regulation.

References

