Prediction of Transcription Factor Binding Sites in *Saccharomyces cerevisiae*

Takuya Ueno  
f673009t@bio.kyutech.ac.jp

Hidetoshi Kono  
kono.hidetoshi@jaea.go.jp

Sarai Akinori  
sarai@bio.kyutech.ac.jp

1  Kyushu Institute of Technology, Graduate School of Computer Science and Systems Engineering, 680-4, Kawazu, Iizuka, Fukuoka 820-8502, Japan  
2  Computational Biology Group, Neutron Biology Research Center, Quantum Beam Science Directorate, Japan Atomic Energy Agency (JAEA) 8-1, Umemidai, Kizu-cho, Souraku-gun, Kyoto 619-0215, Japan  
3  PRESTO, Japan Science and Technology Agency (JST), Japan

Keywords: transcription factor binding site, binding site prediction, matrix, database

1 Introduction

Now that the complete genome sequences were determined for more than 500 species, it becomes possible to analyze the function at the genomic scale. One of the most vital functions in organisms is the regulation of gene expression achieved by a network of transcription factors and their target genes. In recent years, large amounts of sequence and structural information for transcription factors and the target genes have emerged from various ongoing genome projects. However, the mechanisms by which the transcription factors identify their target genes are not well understood yet. To achieve better understanding of the gene regulation mechanism, it is vital to elucidate which gene is controlled by which transcription factor by identifying binding sites of transcription factors. We have developed the methods to predict binding sites of transcription factors on the basis of structural information and computer simulations. In order to apply these methods to a genome-wide prediction, we conducted a pilot study for the yeast genome. For this purpose, we first constructed a database, to which we put together various kinds of data related to gene regulation in the yeast genome. We also integrate the predicted binding sites for each transcription factor into the database, and analyze the relationship with other annotation data in the database.

2 Method

We collected various kinds of data related to gene regulation in the yeast genome, including ORF, transcription factor sequences and features, structural information, known binding sites, microarray data and ChIP-chip data. We used MySQL and PHP to integrate the data and to provide a Web interface. Users can search various fields and analyze the relationship among the data. The database has links to other databases such as TRANSFAC [1] and SGD (Saccharomyces Genome Database) [2]. Furthermore, the database has links with the tools to predict transcription factors and target sites.

We have implemented several methods for the binding-site prediction. For the sequence-based prediction, we used a tool, Match, which uses position-specific weight matrices in TRANSFAC. The public version of TRANSFAC contains a small number of the matrices. Also, some matrices are based on a few sequences of known binding sites. In order to increase the number of matrix, we collected similar binding sites by BLAST using a known binding site and created the position-specific weight matrices, or profile, of the binding site for particular transcription factors, by using a sequence analysis package GCG [3]. We used these profiles to predict the binding sites of transcription factors. We also used the structure-based method for target prediction. 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 of protein-DNA complexes [4]. Then these statistical potentials are used to evaluate the fitness of sequences to the complex structures of particular transcription factors by a so-called combinatorial threading procedure, in which the original target DNA sequence in the protein-DNA complex is replaced by other sequences. When the threading procedure is applied to a real genome sequence, we can identify potential binding sites of transcription factors. Eventually, we combine these methods to enhance the accuracy of the prediction. The predicted binding sites are incorporated into the database.

3 Results and Discussion

There are only 23 position-specific weight matrices for the transcription factors of *Saccharomyces cerevisiae* in the public version of TRANSFAC. Thus, we built the profile matrices by the above-mentioned method. We could create 312 kinds of profiles for 85 transcription factors. For those transcription factors with structural information, we have used the structure-based method to predict their binding sites. In yeast, about one third of genes have some kind of structural information. If homologous structures are available for a transcription factor, we can first predict its structure (only Cα coordinates need to be predicted), and use the model structure to predict its binding sites. Then, we have combined this method with the sequence-based method to enhance the accuracy of the prediction. We have implemented these predicted data into the database.

We have developed the database of gene regulation in yeast genome as a basis for the pilot study on the transcriptional regulation at the genome scale. We are also developing a system to predict binding sites and target genes of transcription factors. We integrate the predicted results into the database and compare them with the experimental data such as microarray and ChIP-chip data. We will make the resources to the public. These database and prediction tools would help us to analyze the context of binding sites in the promoter region in a systematic manner, and to decipher the logic of transcriptional regulation.

References


