Prediction of Cooperative Binding Sites for Multiple Transcription Factors

Yifei Wang wang@rtc.riken.go.jp
Akinori Sarai sarai@rtc.riken.go.jp
The Institute of Physical and Chemical Research (RIKEN)
3-1-1, Koyadai,Tsukuba, Ibaraki 305-0074, Japan

1 Introduction

The precise spatial and temporal regulation of gene expression is achieved by sequence- specific recognition of target genes with multiple transcription factors. The failure in their action could cause many diseases. In order to understand the mechanisms of such a complex system, we need to study the action of transcription factor not only at the molecular level but also at the level of a network of assembled molecules. But, it is difficult to reveal the mechanism of the synergistic regulation of gene expression by multiple transcription factors only using traditional experiments. Thus, developing computer methods to predict the cooperative binding sites and target genes of multiple transcription factors would be useful not only for understanding the mechanism of regulation of gene expression but also for designing experiments. We have developed a computer program called BSS for predicting the potential cooperative binding sites target genes for multiple transcription factors.

2 System and Method

BSS consists of DNA sequences database, recognition matrix, search module, and integration module. The source code for BSS is written in C.

DNA sequences database: DNA sequence containing regulatory regions are stored in this database. The files are in GenBank format. The database can be linked to GenBank database in network environment.

Recognition matrices database: We collected "mutation matrix" or "weight matrix", which are used for the binding-site search. Those matrices are derived from systematic mutagenesis and binding measurements [1,2], from random-sequence selection experiment [3], and sequence alignment of actual binding sequences [4]. All the data are translated into the binding free energy change $\Delta\Delta G$ (see below). Now the database contains more than 200 matrices.

Searching module: This is a core module of BSS for searching the binding sites and target genes for transcription factors from DNA sequence database. In this method, we use the the binding free energy change $\Delta\Delta G$ as the weight matrix. The probability of occurrence of base is converted to $\Delta\Delta G$ by Boltzmann distribution. The total binding free energy change $\Delta\Delta G_{\rm tot}$ can be calculated by moving the window along the DNA sequence [2]. In order to minimize false negatives and false positives, a threshold g is introduced with the following criterion:

if $\langle\Delta\Delta G_{\rm tot}\rangle - \Delta\Delta G_{\rm tot} > g$, the DNA segment will be a potential binding site;

if $\langle\Delta\Delta G_{\rm tot}\rangle - \Delta\Delta G_{\rm tot} < g$, the DNA segment will be not a potential binding site.

Two modes of determining threshold are provided: In the static mode, the threshold value is determined empirically by user. In the dynamic mode, the threshold value is determined automatically with optimization techniques. In the calculation, both strands of the DNA sequence are screened simultaneously.

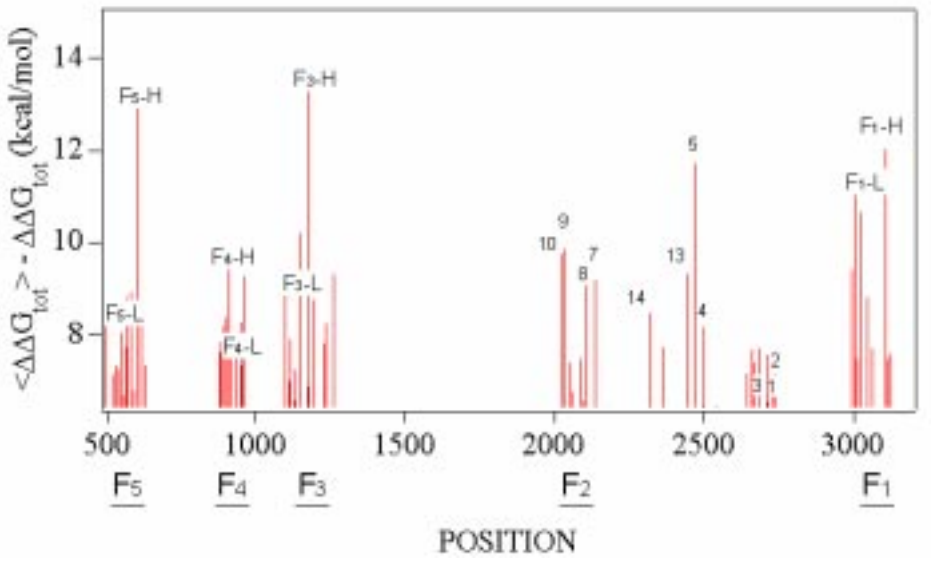

Figure 1: Possible c-Myb binding sites in the c-myc promoter region predicted by BSS.

Integration module: A two-step strategy is used for handling multiple transcription factors. First, the search is executed one by one when the multiple transcription factors were selected. Then, all of the search results are integrated in this module. The integrated result for multiple transcription factors is saved in a output file.

3 Results and Discussion

Three tests were executed in order to assess predictive capacity and reliability of the method. (1) The c-*myc* sequence was screened by BSS using the recognition matrix of the c-Myb oncoprotein. Fig. 1 shows the predicted results. The experimentally known binding sites are marked by alphabets and numbers. This result shows that the method is efficient for searching potential binding sites of transcription factor. (2) The known target genes of c-Mby were screened by BSS. The ratio between the number of identified binding sites and the number of predicted binding sites was about 0.7. (3) To examine the search capability for multiple transcription factors, we searched potential cooperative binding sites and target genes for c-Myb, Sap-92 and Elk-93 using BSS. The results identified some known binding sites (data not shown). BSS is a useful tool for predicting cooperative binding sites and target genes of multiple transcription factors. It will help us get insight into cooperative regulation gene expression by multiple transcription factors. We are currently preparing the network version of BSS.

References

[1] Sarai, A. and Takeda, Y., λ repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically, *Proc. Natl. Acad. Sci. USA*, 86:6513–6517, 1989.

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

[3] Weston, K., Extension of the DNA bind consensus of the chicken c-Myb and v-Myb proteins, *Nucleic Acids Res.*, 20:3043–3049, 1992.

[4] Frech, K., Quandt, K., and Werner, T., Finding protein-binding sites in DNA sequences: the next generation, *TIBS*, 22:103–104, 1997.