Motif Analysis of Yeast Transcriptional Regulatory Regions

Tetsushi Yada [1] yada@tokyo.jst.go.jp
Yasushi Totoki [2] totoki@imslab.co.jp
Kenta Nakai [3] nakai@imcb.osaka-u.ac.jp

[1] Japan Science and Technology Corporation (JST)
5-3 Yonbancho, Chiyoda-ku, Tokyo 102, Japan

[2] Information and Mathematical Science Laboratory, Inc.
2-43-1 Ikebukuro, Toshima-ku, Tokyo 171, Japan

[3] Institute for Molecular and Cellular Biology, Osaka University
1-3 Yamada-oka, Suita 565, Japan

Most of the genetic information can be conveniently classified into two categories: the information on the amino acid sequences and the information on their regulation. With the recent progress of large-scale sequencing efforts, a large number of hypothetical amino acid sequences have been determined. However, the computational study of their regulatory sequences is still in its infant stage. To overcome this situation, we presented a new scheme to predict the function of open reading frames, identified in the genomes of *B. subtilis* [5] and *E. coli* through their sigma-factor dependency. Since eukaryotic cells do not have sigma-factors, something different approach should be explored for the analyses of eukaryotic ORFs. We report here our recent attempt in this direction.

Last year, we introduced the YEBIS system, which can be used to extract a set of motifs from unaligned DNA sequences without any prior knowledge [4]. Two of the strongness of YEBIS are its high speed and its capacity to treat with many sequences at a time. Moreover, the extracted motifs are represented in a hidden Markov model, which allows more flexible representation than other existing methods, such as the regular expression and the weight matrix. In addition, YEBIS outputs a set of motifs in a single run, which is practical because users do not know how many motifs exist in a given set of sequence data. Recently, we further improved its algorithm to detect conserved motifs using recently compiled benchmark data sets [1]. As a result, YEBIS shows comparative or sometimes better sensitivity compared with other programs but it can still process much more sequences quickly (Yada, *et al.*, submitted).

The above-mentioned feature of YEBIS, *i.e.*, its high capacity to process a large number of input sequences, allows us to exploit a new approach to discover any *cis*-elements that characterize a set of genes which are likely to be regulated by some common factors. First, two sets of genes are prepared: one set is related to a specific function, say, amino acid metabolism, while the other is known to be unrelated to its function. We used the functional classification of yeast genes by MIPS [2]. For both data sets, the upstream sequences of length 600bp were extracted and were applied to YEBIS. Then, the obtained two sets of motifs were compared using the group-to-group dynamic programming method, and the motifs which appear in the former set only were selected. Thus, we call this approach '*In silico* subtraction assay'. The extracted motifs were tested whether they are previously characterized ones or not by comparing with the known binding sites stored in the TRANSFAC database [3]. In Table 1, some examples of extracted motifs are shown. Further experiments using other datasets are also ongoing. Because of the recent progresses of both the DNA chip technology and the functional genome projects, there will be a number of works that systematically determine the expression profile of various genes. Thus, our approach should be useful to interpret these results from the sequence data.

Table 1: Typical Motifs Extracted from Various Functional Groups of Yeast Genes. Motifs are shown in a simpler representation than original HMM. The last column indicates the name of known transcription factors which have similar sequence specificity.

Group	Examples of Significant Motifs	Simil. Fact.
Metabolism	C_{100} T_{100} C_{60} s_{99} A_{100} G_{100} A_{56} n n w_{69}	CAR1
	T_{59} y_{75} T_{87} C_{100} C_{85} G_{73} C_{100} G_{58} G_{57}	HSTF, CUP2
Energy	y_{70} n y_{68} C_{62} T_{92} T_{92} T_{70} C_{100} T_{100} T_{100} T_{59} n w_{68}	AP-1
Cell growth	A_{84} A_{86} A_{100} C_{100} G_{100} C_{100} G_{100} T_{65}	DSC1
Transcription	A_{62} G_{83} A_{97} A_{89} G_{74} A_{100} G_{100} G_{100}	UME6
Protein synth.	T_{56} A_{62} C_{63} A_{85} T_{56} C_{98} C_{85} G_{63} T_{100} A_{100} C_{100} ...	RAP1
	... A_{100} T_{80} y_{76} T_{56}	
Protein destin.	G_{100} C_{100} G_{100} A_{100} T_{100} G_{100} A_{100}	BUF
Cell. organization	A_{65} A_{94} T_{100} A_{100} A_{89} T_{100} A_{100} T_{80} A_{59}	MAL63
	G_{100} C_{100} T_{100} G_{100} C_{89} T_{67} G_{89}	CAR1
	w_{76} C_{100} m_{88} G_{100} T_{86} G_{100} G_{100} A_{59} r_{65}	DAL82
Protein kinases	A_{73} A_{100} A_{100} A_{87} A_{100} G_{100} A_{72} A_{83}	UME6

Acknowledgments

This work was carried out as a part of ALIS (Advanced Lifescience Information System) project for genome analysis by Japan Science and Technology Corporation (JST). KN was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas, "Genome Science", from The Ministry of Education, Science, Sports, and Culture in Japan.

References

[1] Frech, K., Quandt, K., and Werner, T., "Software for the analysis of DNA sequence elements of transcription," *Comput. App. Biosci.*, 13:89–97, 1997.

[2] Mewes, H. W., Albermann K, Bähr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S. G., Pfeiffer, F., and Zollner, A., "Overview of the yeast genome," *Nature*, 387(6632 Suppl):7–65, 1997.

[3] Wingender, E., Dietze, P., Karas, H., and Knüppel, R., "TRANSFAC: a database on transcription factors and their DNA binding sites," *Nucl. Acids Res.*, 24:238–241, 1996.

[4] Yada, T., Totoki, Y., Ishikawa, M., and Asai, K., "Automatic discovery of hidden markov representations for functional sites within DNA sequences," *Genome Informatics 1996*, pp.210–211, Univ. Acad. Press Inc., Tokyo, 1996.

[5] Yada, T., Totoki, Y., Ishii, T., and Nakai, K., "Functional prediction of *B. subtilis* genes from their regulatory sequences," *Intell. Syst. Mol. Biol.*, 5:354–357, 1997.