Evaluation of a Method for Predicting Transcription Factors Using Motif-Search Programs

Goro Terai [1,2] terai@gic.intec.co.jp
Takeshi Mizuno [1] mtakeshi@gic.intec.co.jp
Toshihisa Takagi [2] takagi@ims.u-tokyo.ac.jp

[1] Genome Informatics Center, INTEC Inc., 1-3-3 Shinsuna, Koto-ku, Tokyo 136-8637, Japan
[2] Laboratory of Genome Database, Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

1 Introduction

To understand biological phenomena at the molecular level, such as cell-cycle, differentiation, development, proliferation, and so on, it is important to elucidate the mechanism for transcriptional control of each gene, which is involved in their phenomena. Discovery of transcription factors helps us to elucidate the new transcriptional control mechanisms that have not yet been characterized. Although transcription factors can be predicted based on the information derived from homology search using computational tools such as BLAST and FASTA, homology search fails to predict the novel transcription factors that have no significant homology to known transcription factors. Therefore we are evaluating a method for predicting transcription factors based on the existence of local structure such as a DNA-binding domain and/or a transactivation domain, which is often too short to be detected by homology search.

2 Method

1. Overview of our research
   At the first step, we list all candidate transcription factors that might be predicted to include at least one of eleven types of DNA-binding domains, all of which is known to be present in transcription factors. Then we confirm

   (a) how many known transcription factors are listed among the total number of known transcription factors, i.e. sensitivity, and

   (b) how many known transcription factors are included in all listed proteins whose biological functions have been already identified, i.e. specificity.

   Here we searched all proteins, which are encoded in *Saccharomyces cerevisiae* genome, by the three motif-search programs, HMMER [1], PSI-BLAST [2], and PROSITE [3]. And then we evaluated their results separately in terms of sensitivity and specificity.

2. Determining a threshold
   In using HMMER and PSI-BLAST, we can set a threshold. Since it is uncertain what score should be used as a threshold for each DNA-binding domain, we check the result with different threshold and decide an adequate threshold artificially (Fig. 1).

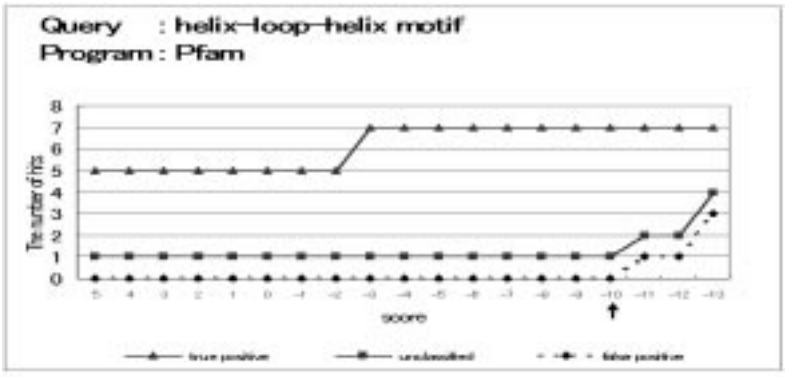

Figure 1: Determining a threshold. In this case we decided that score −10 is adequate.

Table 1: The number of proteins predicted to include at least one DNA-biding domain by three motif-search programs.

	Known transcription factors (total number 130)	Known proteins other than transcription factors (total number 3068)	Unknown proteins (total number 3160)
HMMER	98	24	64
PSI-BLAST	95	23	54
PROSITE	90	13	61

3 Result

As calculated from the data shown in Table 1, the sensitivity is from 0.70 to 0.75 and the specificity is from 0.80 to 0.87.

In terms of both sensitivity and specificity, we conclude that this method is useful to predict transcription factors. Hereafter we will evaluate a method for predicting transcription factors based on the existence of a transactivation domain.

References

[1] Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D., and Sonnhammer, E.L., Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins, *Nucleic Acids Res.*, 27:260–262, 1999.

[2] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 25:3389–402, 1997.

[3] Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A., The PROSITE database, its status in 1999, *Nucleic Acids Res.*, 27:215–219, 1999.