An improved method of finding over-represented sequence motifs in sets of DNA sequences.

Tadashi Imanishi[1], Hiroki Hokari[2], Motohiko Tanino[3], Jun-ichi Takeda[4], Taichiro Sugisaki[5] and Shin Nurimoto[6]

Keywords: sequence motifs, OSMO-finder, computer simulation

1 Introduction.

Aside from gene coding sequences, eukaryotic genomes contain important functional sequences such as regulatory elements of gene expression and chromosomal replication. However, many of them remain uncharacterized. Finding functional sequence motifs in genomic sequences is thus a very important and urgent issue in bioinformatics. We thus developed a method for finding sequence motifs with the aim of discovering unknown functional sequence motifs in genomes, and we implemented the method in a computer program called OSMO-finder (Over-represented Sequence MOtif finder)[1]. This is a heuristic method of finding sequence motifs that appear frequently in sets of DNA sequences. We have extensively improved the algorithm of OSMO-finder and examined its efficiency of finding sequence motifs by using computer simulations. In this paper, we describe the outline of the improved algorithm and the results of computer simulations.

2 An improved algorithm of OSMO-finder.

The purpose of the method is to find a consensus DNA sequence (motif) without gaps that are over-represented in a set of unaligned sequences. Motifs can be evaluated based on the properties including the length, the number of occurrences (L) in the data set, and the level of sequence conservation (the number of nucleotide mismatches allowed). In OSMO-finder, we evaluate the identified motifs by the exact probability (P-value) that the motifs appear L times or more in random sequences of the same size as the data set. OSMO-finder first finds "seeds" of motifs of a given length W that appear frequently, allowing some mismatches, in a set of DNA sequences. It then searches for the optimum motif by calculating P-values for variously modified motifs by extending the seeds and changing the levels of sequence conservation.

[1] Biological Information Research Center, National Institute of Advanced Industrial Science and Technology. Time24 Bldg. 10F, Aomi 2-45, Koto-ku, Tokyo 135-0064, Japan. E-mail: imanishi@jbirc.aist.go.jp
[2] Mitsui Knowledge Industry Co., Ltd. Harmony Tower 21F, Honcho 1-32-2, Nakano-ku, Tokyo 164-8721, Japan. E-mail: hhokari@jbirc.aist.go.jp
[3] Japan Biological Information Research Center, Japan Biological Informatics Consortium. Time24 Bldg. 10F, Aomi 2-45, Koto-ku, Tokyo 135-0064, Japan. E-mail: mtanino@jbirc.aist.go.jp
[4] Biological Information Research Center, National Institute of Advanced Industrial Science and Technology. Time24 Bldg. 10F, Aomi 2-45, Koto-ku, Tokyo 135-0064, Japan. E-mail: jtakeda@jbirc.aist.go.jp
[5] Mitsui Knowledge Industry Co., Ltd. Harmony Tower 21F, Honcho 1-32-2, Nakano-ku, Tokyo 164-8721, Japan. E-mail: sugisaki@hydra.mki.co.jp
[6] Mitsui Knowledge Industry Co., Ltd. Harmony Tower 21F, Honcho 1-32-2, Nakano-ku, Tokyo 164-8721, Japan. E-mail: nurimoto@hydra.mki.co.jp

Finally, the motif with the smallest *P*-value is chosen as the optimum one. The program is written in JAVA and can be run under various operating systems.

3 Computer simulation.

To examine how efficiently OSMO-finder can discover sequence motifs in sets of DNA sequences, we conducted computer simulation using artificial sequences with hiddensequence motifs. We generated 10 artificial DNA sequences of 600 base pairs (bps) and inserted 10 or 20 sequence motifs of 15 bps. We then run the OSMO-finder (version December 2003) and measured the efficiency of finding hidden motifs. The test was repeated 5 times. The efficiency was measured by the level of overlap of correct and predicted motifs. As a result, it appeared that OSMO-finder can discover hidden motifs efficiently (Table 1). For example, 81% of the hidden motifs were successfully identified when there are 20 motifs with 10% sequence diversity from a consensus sequence. We also compared the efficiency of OSMO-finder with that of MEME (version 3.0.4) with "Two-Component Mixture" (TCM) option[2]. These two programs showed comparable efficiency. In particular, OSMO-finder showed better performance when the level of sequence conservation of motifs is low. These results clearly demonstrate the usefulness of OSMO-finder in identifying functional sequence motifs from genomic sequences.

number of hidden motifs in the data	diversity of sequence motifs	OSMO-finder	MEME (TCM)
10	0%	0.718	0.940
10	10%	0.731	0.819
10	20%	0.242	0.017
20	0%	1.000	1.000
20	10%	0.811	0.920
20	20%	0.346	0.030

Table 1. Efficiency of finding hidden sequence motifs by OSMO-finder and MEME (TCM).

References

[1] Imanishi T., Shikanai T., Nurimoto S. and Sugisaki T. 2003. A new method of finding functional sequence motifs and its application to human GC-rich genomic sequences. In Spang R., Beziat P. and Vingron M. editors, Currents in Computational Molecular Biology 2003 (RECOMB 2003, Berlin), pp. 61-62.

[2] Bailey, T.L. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36.