Multiple sequence alignment by combining incomplete blocks of similar segments

K. Suzuki Y. Akiyama M. Kanehisa

suzuken@kuicr.kyoto-u.ac.jp akiyama@kuicr.kyoto-u.ac.jp kanehisa@kuicr.kyoto-u.ac.jp

Institute for Chemical Research, Kyoto University
Gokasho, Uji, Kyoto 611 Japan

Abstract

We have developed a novel method for multiple sequence alignment based on combinatorial selection of similar block candidates. Our method resembles manual multiple alignment performed by biologists. The method is more feasible for finding functional motifs than previous multiple alignment algorithms that are extensions of pairwise alignments. We employed a Hopfield neural network technique so that the method can cope with the combinatorial explosion in examining a large number of "incomplete" block candidates.

1 Introduction

Traditional multiple sequence alignment algorithms use the strategies extending pairwise alignments: tree based methods [1] and randomized iterative strategies [2]. Such computational methods sometimes miss common sequence motifs with biological significance.

In contrast, manual multiple alignments often start with segments or motif candidates and take the following steps. 1) search putative functional motifs or homologous segments, 2) fix these segments as anchor points, and 3) align intermediate sequences. This segment-based method often gives biologically more proper results than computational methods.

As a computational segment-based method, Vingron and Argos [3] developed a multiple alignment program, but their method treats only exact match segments and easily misses weakly similar segments. Gracy and Sallantin [4] developed a method considering not only exact matches but also similar segments. These two previous proposals were based on segments shared by all the sequences, which we call complete blocks. If incomplete blocks, i.e., segments shared by not all the sequences, are considered, the method will produce more proper alignments but the time complexity will also increase explosively. Thus previous methods for computational segment-based multiple alignment utilized only complete blocks.

Here in order to consider incomplete blocks within a reasonable time, we employ a Hopfield neural network technique[5] to solve the combinatorial optimization problem.

鈴木健二、秋山　泰、金久　實：京都大学化学研究所、〒 611 京都府宇治市五ヶ庄

2 Algorithm

Any segment-based multiple alignment method requires a pre-processing for finding similar segment blocks. We use Sagot's algorithm[6] and generate about 100-1500 candidate blocks typically. The main process selects the combinations of blocks that do not contradict each other. We employ the Hopfield neural network for this step. We construct an evaluation function $E(\overrightarrow{x})$ as follows:
(1) the higher homology score a block has, the more this block is feasible,
(2) blocks must not interfere with each other, and
(3) the more parallel two blocks are, the more these blocks are selected simultaneously.

$$E(\overrightarrow{x}) = -\sum_{i=1}^{M} e_i \cdot x_i + \frac{\max(|e_i|)}{2} \sum_{i=1}^{M}\sum_{j=1}^{M} c_{ij} \cdot x_i \cdot x_j - \alpha \sum_{i=1}^{M}\sum_{j=1}^{M} f_{ij} \cdot x_i \cdot x_j$$

with the binary variable:

$$x_i = \begin{cases} 0, & \text{for candidate } (i) \text{ not selected} \\ 1, & \text{for candidate } (i) \text{ selected} \end{cases} \quad (1 \le i \le M)$$

the similarity score:

$$e_i = \frac{1}{2(N-1)} \sum_{s_1=1}^{N} \sum_{\substack{s_2 \neq s_1 \\ s_2=1}}^{N} \sum_{r=1}^{L} \text{similarity}(\text{residue}(i, s_1, r), \text{residue}(i, s_2, r))$$

the degree of interference:

$$c_{ij} = \begin{cases} 0, & \text{for candidate } (i) \text{ and candidate}(j) \text{ not interfere} \\ 1, & \text{for candidate } (i) \text{ and candidate}(j) \text{ interfere} \end{cases} \quad (1 \le i, j \le M)$$

and the degree of parallelism:

$$f_{ij} = \sqrt{\text{variance}(\text{segment location}(i, s, 1) - \text{segment location}(j, s, 1))}$$

where L is segment length, M is number of candidates, and N is number of sequences.

It was thus possible to develop a segment-based multiple alignment considering incomplete blocks. An experimental system was implemented on SPARCstations. However, at the moment the pre-processing requires a long computation time, which will be improved in the future.

Acknowledgement

This work was supported by the Grant-in-Aid for Scientific Research on Priority Area 'Genome Informatics' from The Ministry of Education, Science, Sports and Culture in Japan.

References

[1] Feng, D-F., Doolittle, R. F., *J. Mol. Evol.*, **25**, 351-360, (1987).

[2] Berger, M. P., Munson, P. J., *Comput. Applic. Biosci.*, **7**, 479-484, (1991).

[3] Vingron, M., Argos, P., *Comput. Applic. Biosci.*, **5**, 115-121, (1989).

[4] Gracy, J., Sallantin, J., *Proc. Genome Informatics Workshop 1994*, 100-109, (1994).

[5] Hopfield, J.J., Tank, D. W., *Biol. Cybern.*, **52**, 141-152, (1995).

[6] Sagot, M-F., Viari, A., Soldano, H., *Proc. ISMB-95*, 322-331, (1995).