An Inspection of the Multiple Alignment Method with use of a Genetic Algorithm

Yoshitomo Harada
si97014@si.hirosaki-u.ac.jp

Masato Wayama
wayama@cc.hirosaki-u.ac.jp

Toshio Shimizu
slsimi@si.hirosaki-u.ac.jp

Department of Information Science , Faculty of Science,
Hirosaki University,
Bunkyo-cho 3 , Hirosaki 036 Japan

1 Introduction

We proposed a method for amino acid sequence alignment problems by using a genetic algorithm[1][2]. The resulted alignments obtained by applying our method to the data sets of rather small number of comparatively short sequences with higher similarity were satisfactory ones, even better than those of CLUSTALW[3] with respect to e.g., the alignment score, the numbers of inserted gaps and matches. For the data sets including large number of sequences with less similarity, however, rather poor alignment results were obtained. In order to find out major impediments in the method and to improve it further, it seems to be necessary to investigate the performance of the method by applying it to various data sets of different characteristics such as the number of sequences, sequence length and similarity, etc..

2 Methods and Data

A sequence was encoded as a bit string in our method, and an alignment was expressed with an $N \star M$ matrix which is a vertical alignment of binary strings. We employed the liner ranking method[4] for the reproduction process. The alignment score was calculated by the "Sum-of-the-Pairs" method according to fitness (alignment score) [5]. We used a form of crossover known as "uniform crossover" ("window-frame" crossover). In addition to crossover, four different mutations were used : "continuous-gap-shift", "continuous-gap-extension", "gap-block-shift" and "gap-block-extension" operators. The population size was set to 1000 and the GA ran for 5000 generations. The details of the method was described previously[1][2].

We prepared the 13 data sets of various sequence length, the number of sequence and similarities, as summarized in Table 1. Serine proteinase sequences were chosen from PDB and the resting 12 data sets were from SWISS-PROT.

3 Results and Discussion

We obtained the alignment results comparable in the alignment quality to CLUSTALW with regard to gap insertion pattern, the number of inserted gaps, the number of matches and the alignment score, for all the data sets except Try11 and Spr12, as shown in Table 1. The size of problem space of the data sets given good results by our method correspond to protein families in which $l_s \star n_s$ (defined in Table 1) is approximately less equal than 2000, and the other two data sets to greater equal than 2000. We are currently investing the cause of this behavior. Another problem in our method is that it takes much longer time than CLUSTALW to obtain an alignment. We are planning to parallelize the GA process to get alignments more faster for practical use.

Table 1: Data Sets

data set	protein name	average sequence length(l_s)	number of sequences (n_s)	similarity	$l_s \star n_s$
Tox5	toxin	67	5	16.9	335
Tox10	toxin	67	10	14.1	670
Fla4_1	flavodoxin	162	4	8.4	648
Fla4_2	flavodoxin	173	4	28.7	692
Fla5	flavodoxin	164	5	5.9	820
Fla8	flavodoxin	165	8	3.7	1320
Tim4	TIM	245	4	19.5	980
Tim5	TIM	247	5	18.6	1235
Tim8	TIM	247	8	16.7	1976
Apo7	apolipoprotein	285	7	8.4	1995
Try4	trypsin	251	4	21.1	1004
Try11	trypsin	247	11	9.7	2717
Spr12	serine proteinase	219	12	3.9	2628

Table 2: Comparison of alignment results between our method and CLUSTALW

	matches		inserted gaps		score	
data sets	GA	CLUSTALW	GA	CLUSTALW	GA	CLUSTALW
Tox5	15	13	50	50	1264	814
Tox10	11	11	102	110	5652	4046
Fla4_1	11	15	55	68	843	973
Fla4_2	46	51	53	41	2405	2683
Fla5	9	11	91	99	1612	1903
Fla8	5	7	170	216	4728	5312
Tim4	47	49	43	47	2045	2272
Tim5	48	48	57	57	3895	3884
Tim8	41	44	116	124	13541	13753
Apo7	17	27	261	239	8556	10464
Try4	52	57	98	62	2661	2887
Try11	0	28	279	452	1111	16962
Spr12	0	11	266	754	-7776	8491

References

[1] Wayama, M., Takahashi, K. and Shimizu, T., *Proc. Genome Informatics Workshop 1995*, Universal Academy Press, pp. 122-123, 1995.

[2] Isokawa, M., Wayama, M. and Shimizu, T., *Proc. Genome Informatics Workshop 1996*, Universal Academy Press, pp. 176-177, 1996.

[3] Thompson, J. D., Higgins, D. G. and Gibson, T. J. *Nucleic Acids Research*, Vol.22, pp. 4673-4680.

[4] Baker, J. E., *Proceeding of the First International Conference on Genetic Algorithm*, pp. 164-170, 1985.

[5] Altschul, S. F. and Lipman, D. J., *SIAM J. Appl. Math.*, pp. 197-209, 1989