2008－BIO－13　(1)
2008／6／26

反復改善法によるマルチプルアラインメントスコアの統計的比較検証

若津 大悟[†]，名嘉村 盛和[††]，岡崎 威生[††]

[†] 琉球大学理工学研究科情報工学専攻　[††] 琉球大学工学部情報工学科

反復改善法はマルチプルアラインメントの解の質を上げるために用いられる有用な手法であり、様々なアルゴリズムが提案されている。本研究では、それらのアルゴリズムの比較検証をおこなう。BAliBASE データを用いて、各アルゴリズムの Sum-of-Pairs(SP) スコア、Log Expectation(LE) スコア、BAliBASE スコアの値のふるまいを統計的観点から考察し、各手法の配列種類に対する有用性を考察する。

Statistical Comparative Study of Multiple Sequence Alignment Score with Iterative Algorithms

Daigo WAKATSU[†], Morikazu NAKAMURA[††], Takeo OKAZAKI[††]

[†] Information Engineering Course, Graduate School of Engineering and Science, University of the Ryukyus
[††] Department of Information Engineering, Faculty of Engineering, University of the Ryukyus

Iterative algorithms is a useful method to improve the alignment results. In this paper, we evaluate several different iterative algorithms by comparing the statistical results. There are five iterative algorithms Remove First(RF) and Best-First(BF), Random(RD), Tree-based(Tb) and Tree-based Splitting(TbS) iterative algorithms. We compared these algorithms using four alignment scores, Sum-of-Pairs(SP) Score and Log Expectation(LE) Score, BAliBASE SP(BSP) Score, BAliBASE TC(BTC) score. We observed the behavior of these scores from the point of view of the Cumulative Frequency and other statistics.

1 Introduction

Progressive alignment[1] is the most widely used heuristic approach to align a large number of sequences. The multiple sequence alignment is built up progressively by aligning pairs of sequences followed by pairs of alignments/profiles. The guide tree determines the order in related being aligned first. However, once an error occurs in the alignment process, that error can never be corrected. Iterative algorithms solves this problem. By applying Dynamic Programming to partially aligned sequences iteratively, their alignment quality is improved. Such an iterative strategy requires heuristic search methods to solve practical alignment problems.

Hirosawa *et. al.*[2] investigated the performance of different iterative algorithms. They tested effectiveness about each algorithms using Sum-of-Pairs score for improvement of each alignment results. They used a group of 30 protein kinase sequences as the basis for their evaluations. Wallace *et. al.*[3] systematically tested several different iterative algorithms by comparing the results on sets of alignment test cases using HOMSTRAD database of structure-based alignments[4] . They tested that iterative algorithms were an effective way of improving the performance of progressive alignment programs, and they proved it can be used to improve existing alignment software. It was found to be even more powerful when it was directly incorporated into a progressive alignment scheme.

In this paper, we revisited these important work. There are several iterative algorithms using several scores, and several characteristics of data set types. We focused characteristics of data set types, and tested the performance of each alrotighms depending on the sequence types. We examined the comprehensive analysis in several alignment strategies. To compare all strategies in the statistical view, we discussed the best performance strategies for the each data set types.

2 Benchmark Data Set

BAliBASE 3.0 alignment benchmark data set[5] was used to compare the performance of different alignment algorithms. The BAliBASE benchmark data set contains 218 reference alignments, and divided into 6 different reference sets, each with different characteristics.

Reference 1-1 contains alignments of equi-distant very divergent sequences(20% identity) with 38 alignment sets. Reference 1-2 contains alignments of equi-distant medium to divergent sequences with 44 alignment sets. Reference 2 contains families aligned with a highly divergent “orphan” sequences with 41 alignment sets. Reference

3 contains subgroups with 25% residue identity between groups with 30 alignment sets. Reference 4 contains sequences with N/C-terminal extensions with 49 alignment sets. Reference 5 contains sequences with large internal insertions with 16 alignment sets.

3 Multiple Sequence Alignment Algorithms

Six multiple alignment algorithms were selected for comparison.[2) 3)]

Progressive Alignment algorithms(PA) utilizes the algorithm of Needleman and Wunsch[6)] iteratively for the pairwise alignment in order to obtain a multiple sequence alignment. The distance of each pair is calculated by the pairwise alignment with Dynamic Programming, and the distance matrix that shows the distance of all sequence pairs is constructed. The guide tree is constructed from the distance matrix.

Remove First iterative algorithms(RF) has simple iterative strategy. In each iteration step, one sequence is removed from the alignment, and realigned to the remaining alignment. If the alignment result is better, it is kept and used as input for the next iteration. The iteration cycle is terminated if the alignment score converges, or upon completing $2N^2$ iterations. N is the number of sequences.

Random iterative algorithms(RD) uses the strategy that the alignment is split randomly into two sets of sequences, which are realigned. If the score is improved, the alignment result is kept.

Best-First iterative algorithms(BF) uses the strategy that in each iteration cycle every sequence is removed and realigned to the rest. The alignment with the best score is kept as input for the next iteration. Again, the iteration cycle is terminated if the alignment score converges.

Tree-based iterative algorithms(Tb) uses the strategy that the alignment improvement algorithms are also incorporated into a progressive alignment strategy. Every time two profiles are combined, the alignment result is refined using one of the iterative algorithms described above.

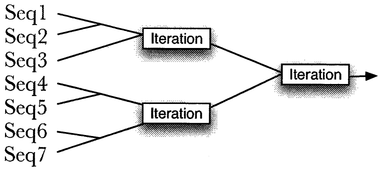

Figure 1: Tree-based Iterative Algorithms

Tree-based Splitting iterative algorithms(TbS) is an extension of the tree-based iterative algorithm. The sequences are split into smaller subsets of a predefined maximum size using a tree.

4 Scores

To assess the performance of these algorithms in this study, there are four different scores.

The Sum-of-Pairs(SP) measure is a well known scoring function for multiple sequence alignment. To calculate the score of a multiple sequence alignment, where the score of each pair of rows of the multiple alignment is added up to form to overall score. Consider a multiple sequence alignment A of length l composed on N nucleotide or amino acid sequences, The SP score of A, SP(A) is defined by:

$$\mathbf{SP}(A) = \sum_{j=2}^{N} \sum_{k=1}^{j-1} S_{j,k} \tag{1}$$

where $S_{j,k}$ is the score associated with the pairwise alignment between the jth and $k$the sequences within A.

The Log Expectation (LE) Scoring function is used by Muscle[9)] .

$$\mathbf{LE}^{xy} = (1 - f_G^x)(1 - f_G^y)\, log \sum_i \sum_j \frac{f_i^x f_j^y p_{ij}}{p_i p_j} \tag{2}$$

Where p_i is a background probability of amino acid i, p_{ij} is a joint probability of i and j being aligned. f_i^x is a observed frequency of $i(j)$ in column x. f_G^x is an observed frequency of gaps in that column at position x in the 1st profile.

These SP and LE scoring function are used in the iterative algorithms to measure iteration judgement steps.

The BAliBASE score gives a true and estimated multiple sequence alignment, the accuracy of the estimated alignment is usually computed using the BAliBASE Sum-of-Pairs(BSP). BSP is a measure of the number of correctly aligned residue pairs divided by the number of aligned residue pairs in the true alignment.

5 Statistical View for Comparison

We examined total twelve alignment strategies: the combination of six multiple sequence alignment strategies (PA, RF, RD, BF, Tb, TbS) and two scoring funcions(SP, LE).

The multiple alignment programs ClustalW (v1.81) and Muscle (v3.7) was used to generate alignments as inputs for the iterative algorithms. We used SAGA (v0.95)[10)] 's SP scoring function and Muscle's LE scoring function to evaluate in the iteration judgement step.

To compare all strategies in the statistical view, the distributions of each scores are important. For the estimator of the distribution function, we utilize the Cumulative Frequency(CF). In the CF of each scores, it is preferable that the frequency is low at the low score and the frequency increases rapidly at the high score. We consider the performance of CF of each BAliBASE references and each scores.

In addition to CF, we focused the average scores and maximum, minimum, variance score. It is preferable that the average and maximum and minimum score are high, and the variance value is low.

6 Experimental Results

The experimental results for Reference 1-1 and Reference 2 of BAliBASE summarized below.

Figure2 shows the CF of each scores for Reference 1-1. Using either LE and SP scoring function in the each strategy, the results of all SP score gave similar distributions. In the LE score, using SP scoring function in the process of iteration made the results of LE score worse. The result of BAliBASE SP score showed PA using LE had best performance. Using LE in RF, RD, BF gave good performance. Tb using SP had the worst of the all strategies. In the statistics, the average scores of PA using LE and RF using LE, BF using LE gave good scores. Tb using SP had low variance value but the average score was the worst.

Figure3 shows the CF of each scores for Reference 2. In the case of SP score, RD using SP had best performance. And we know that there were almost same distribution. In the LE score, RD using SP had a bad performance, Tb using LE had a good performance. In the BAliBASE SP score, Using LE PA, Rf, RD, BF had good performance, RD using SP had the worst performance. In the statistics, the average scores of PA using LE, RD using LE, BF using LE score gave good scores. Tb using LE had low variance value and others statistics value had relatively good.

7 Conclusion

We noticed that the PA using LE scoring function is a good strategy to consider the efficiency for the computational effort. BF using LE in the of Reference 1-1 and RD using LE in the Reference 2 were the best strategies. We got different results in the Reference 1-1 and Reference2. It means that it is better to change the strategy depending on the sequence types. Also, we found that using the LE scoring function in the iteration judgement steps of each iterative algorithms gave good scores. Using LE scoring function almost outperforms SP scoring function.

References

1) Taylor, W.R., *Multiple sequence alignment by a pairwise algorithm,* Comput. Appl. Biosci., No.3, pp.81–87 (1987)

2) Hirosawa M., Totoki Y., Hoshida M., Ishikawa M., *Comprehensive Study on iterative algorithms of multiple sequence alignment,* CABIOS, Vol.11 NO.1, pp.13–18 (1995)

3) Wallace I.M. , O'Sullivan O. and Higgins D.G. , *Evaluation of Iterative Alignment algorithms for multiple alignment,* BIOINFORMATICS, Vol.21, No.8, pp.1408–1414 (2005)

4) Mizuguchi K., *HOMSTRAD: a database of protein structure alignments for homologous families.,* Protein Sci., Vol.7, pp.2469–2471 (1998)

5) Thompson J.D., Koehl P., Ripp R., Poch O., *BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark.,* PROTEINS: Structure, Function, and Bioinformatics Vol.61, pp.127–136 (2005)

6) Needleman S.B., and Wunsch C.D., *A general method applicable to the search for similarities in the amino acid sequence of two proteins,* J. Mol. Biol., Vol.48, pp.443–453 (1970)

7) Gotoh O. *A weighting system and algorithm for aligning many phylogenetically related sequences.,* Comput. Appl. Biosci, Vol 11, pp.543–551, (1995)

8) Thompson, J.D., Higgins, D.G. and Gibson, T.J., *ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice,* Nucleic Acids Res., Vol. 22, pp.4673–4680 (1994)

9) Edgar R.C., *MUSCLE: Multiple sequence alignment with high accuracy and high throughput,* Nucleic Acids Res., Vol 32, pp.1792–1797 (2004)

10) Notredame C. and Higgins D.G., *SAGA: sequence alignment by genetic algorithm,* Nucleic Acids Research, Vol 24, pp.1515–1524 (1996)

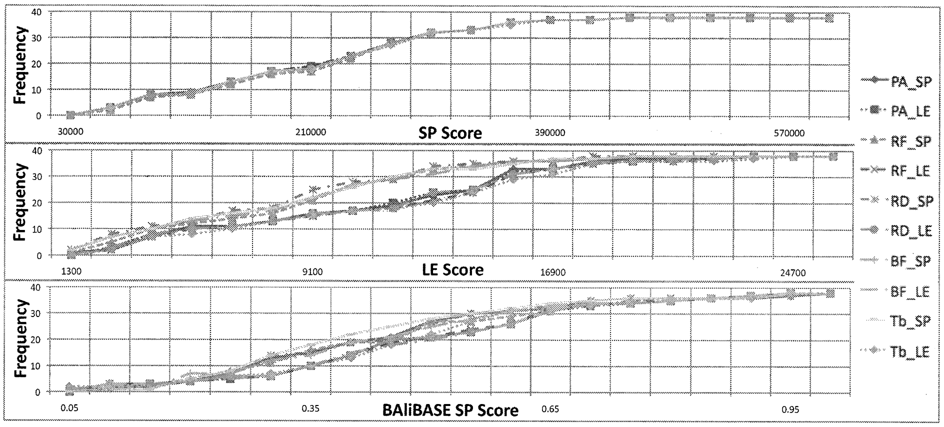

Figure 2: Cumulative Frequency of Reference 1-1

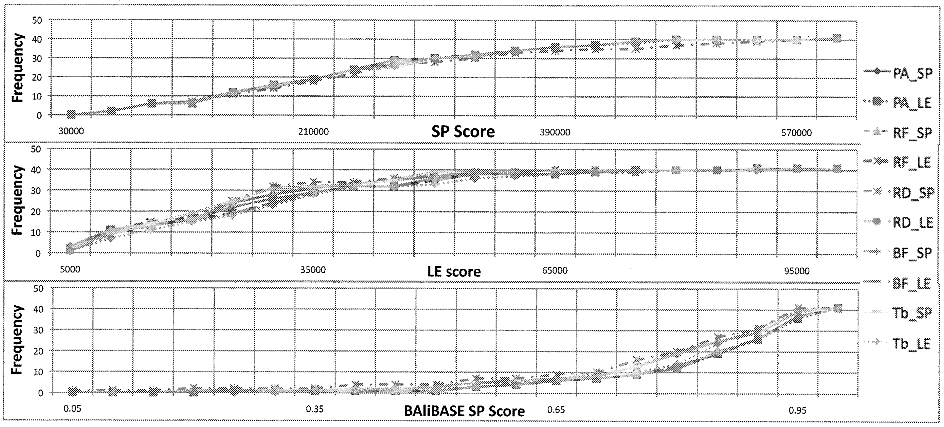

Figure 3: Cumulative Frequency of Reference 2

Table 1: Statistical Value of BAliBASE SP Score

(RV11)	**PA**		**RF**		**RD**		**BF**		**Tb**	
	LE	*SP*	*LE*	*SP*	*LE*	*SP*	*LE*	*SP*	*LE*	*SP*
Avg.	**0.465**	0.415	0.464	0.418	0.463	0.407	0.463	0.416	0.452	0.390
Max.	0.904	**0.961**	0.904	0.939	0.904	**0.961**	0.904	0.939	0.904	0.939
Min.	0.073	**0.084**	0.07	0.024	0.07	0.00	0.07	0.024	0.013	0.041
Var.	0.044	0.041	0.044	0.044	0.044	0.045	0.043	0.042	0.043	**0.039**

(RV20)	**PA**		**RF**		**RD**		**BF**		**Tb**	
	LE	*SP*	*LE*	*SP*	*LE*	*SP*	*LE*	*SP*	*LE*	*SP*
Avr.	**0.806**	0.765	0.804	0.764	0.804	0.726	0.805	0.765	0.782	0.741
Max.	**0.972**	0.954	**0.972**	0.954	0.969	0.933	**0.972**	0.954	0.961	0.953
Min.	0.217	0.166	0.21	0.167	0.217	0.001	0.212	0.166	**0.332**	0.147
Var.	0.023	0.028	0.023	0.028	0.023	0.043	0.023	0.028	**0.021**	0.031