# 局所多重アライメントのための局所探索アルゴリズム：特殊ケースにおける収束性の解析と腫瘍細胞分類への応用

阿久津　達也

東京大学　医科学研究所　ヒトゲノム解析センター

局所多重アライメントは複数の文字列（実際にはDNA配列やアミノ酸配列）が与えられた時、スコアが最大となるように、各文字列から長さの等しい部分文字列を抽出するという問題である。この問題に対して、相対エントロピースコアを用いた単純な局所探索アルゴリズムが知られており、また、計算機実験の結果から数回程度の反復で局所最適解に収束することも知られていた。そこで、本稿では、収束性について理論的考察を行い、非常に特殊な場合（部分文字列の長さが１文字で、かつ、各文字の出現確率が等しい場合）においては多項式回の反復で収束することを示す。一方、このアルゴリズムは局所多重アライメント以外の問題への応用可能である。その例として本稿では遺伝子発現データをもとにした腫瘍細胞分類への応用を示す。

# A Local Search Algorithm for Local Multiple Alignment: Special Case Analysis and Application to Cancer Classification

Tatsuya Akutsu

Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirkanedai, Minato-ku, Tokyo 108-8639, Japan
takutsu@ims.u-tokyo.ac.jp

Local multiple alignment is a problem of locating a region (i.e., a substring) of fixed length from each input protein or DNA sequence so that the score determined from the set of regions is optimized, where the relative entropy score is considered in this paper. For local multiple alignment, a very simple local search algorithm has been known. Previous computational experiments suggested that the algorithm converged very quickly to a local optimal. This paper shows a theoretical result on the convergence rate for a very special case. This paper also shows that the algorithm is useful for other problems. Especially, this paper shows an application of the algorithm to class discovery for cancer classification by gene expression monitoring.

## 1 Introduction

*Local multiple alignment* is one of well-studied problems in bioinformatics [5, 6, 7, 8, 9]. It is also known as the *general consensus patterns* problem or *gapless multiple alignment*. Local multiple alignment is a problem of, given $n$ sequences, locating a region (i.e., a substring) of fixed length from each sequence so that the *score* determined from the set of regions is optimized. Local multiple alignment is useful for finding binding sites, conserved regions and motifs of sequences.

Although several scoring schemes have been proposed, the *relative entropy score* (the average information content score) is widely used. Therefore, this paper considers the relative entropy score. Many studies have been done on local multiple alignment under the relative entropy scoring scheme [1, 5, 6, 8, 9]. Among them, local search algorithms such as the EM algorithm [6] and the gibbs sampling algorithm [7] are very useful. We also proposed a simple local search algorithm (LS, in short) in [1]. It is widely-recognized that making theoretical analysis of the

convergence rate on local search algorithms in bioinformatics (such as EM, Gibbs-sampling) is a very hard task and almost nothing is known. It seems that LS lies in the same situation. This paper makes analysis of the convergence rate of LS for a very special case, in which the length of motif is only one and all types of residues have the same background probability. Even for this very special case, it is not a trivial task to make a theoretical analysis. We show that LS converges to a local optimal within $O(nA)$ steps, where $n$ denotes the number of input sequences and $A$ denotes the size of alphabet $\Sigma$ (i.e., $A = 4$ for DNA sequences, $A = 20$ for protein sequences).

On the other hand, it seems that the algorithms for local multiple alignment can be applied to other problems. Recall that the task of local multiple alignment is to locate similar regions. Locating similar regions is closely related to clustering. Recently, various clustering methods have been applied to analysis of gene expression data [2, 4]. In particular, Golub *et al.* used a kind of clustering algorithm (SOM, self organizing map) for *class discovery* in *cancer cell classification* by gene expression monitoring [4]. Although SOM is useful for class discovery, several parameters must be adjusted manually in order to obtain good clustering results. Moreover, implementation of SOM is not an easy task. So, we applied LS to class discovery. In this paper, we show a preliminary computational result on application of LS to class discovery for cancer classification.

## 2 Algorithm and Analysis

For a string $s$ over an alphabet $\Sigma$, $|s|$ denotes the length of $s$. $s[j]$ is the $j$-th character of $s$. Let $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ be a set of strings. Let $t_i$ be a substring of $s_i$. Let $\#_j(a)$ be the number of the appearances of letter $a$ in the $j$-th column of $t_i$'s (i.e., $\#_j(a) = |\{t_i | t_i[j] = a\}|$). Let $f_j(a)$ be the frequency of letter $a$ in the $j$-th column of $t_i$'s (i.e., $f_j(a) = \frac{\#_j(a)}{n}$). Let $p(a)$ denote the frequency of letter $a$ in the whole genome (i.e., background probability of $a$). Then, local multiple alignment under the relative entropy scoring scheme is defined as follows.

**Local Multiple Alignment**: Given a set $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ of sequences, and an integer $L$, find a substring $t_i$ of length $L$ from each $s_i$, maximizing the score of

$$score(t_1, \ldots, t_n) = \frac{1}{L} \sum_{j=1}^{L} \sum_{a \in \Sigma} f_j(a) \log \frac{f_j(a)}{p(a)}.$$

The following local search algorithm (LS) was proposed in [1].

(1) Select a substring $t_i$ from each $s_i$ at uniformly random.
(2) Let $f_j(a)$ be the frequency of letter $a$ in the $j$-th column of $t_i$'s.
(3) For each $i$, find a substring $t'_i$ of $s_i$ maximizing $s(t'_i) = \sum_{j=1}^{L} \log \frac{f_j(t'_i[j])}{p(t'_i[j])}$.
(4) Replace $(t_1, \ldots, t_n)$ with $(t'_1, \ldots, t'_n)$.
(5) Repeat (2)-(4) until reaching a local optimum.

It should be noted that each iteration can be done in linear time (i.e., $O(\sum |s_i|)$ time). Although LS is introduced explicitly in [1], LS has some similarities with the GIBBS sampling algorithm and the EM algorithm, where similarities and differences are discussed in [5]. We obtained the following theoretical results on this algorithm.

**Proposition 1.** The number of iteration steps is $O(m^n)$, where $m = \max_i |s_i|$.

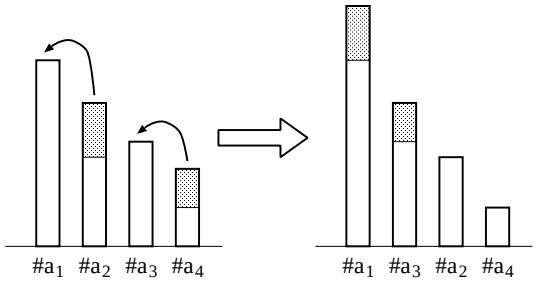


Figure 1: Illustration of changes of $\#a_i$ before and after steps (3)-(4). $\#a_i$ may decrease. But $\sum_{i=1}^{k} \#a_{\pi_i}$ does not decrease.

(Proof) It is not difficult to show that the score increases monotonically in LS. Since the number of possible alignments is $O(m^n)$, the proposition holds. □

Next, we consider a very special case:$L = 1$ and $p(a) = \frac{1}{A}$ for all $a \in \Sigma$. In this case, detailed values of the scores are not important. Let $\#a_i = \#_1(a_i)$. Let $\pi$ be a permutation of $\{1, 2, \ldots, n\}$ such that $\#a_{\pi_1} \geq \#a_{\pi_2} \geq \ldots \geq \#a_{\pi_n}$. Then, it is easy to see that LS selects $t'_i$ in the following way, where we assume without loss of generality that $\#a_{\pi_1} > \#a_{\pi_2} > \ldots > \#a_{\pi_n}$. If $s_i$ contains letter $a_{\pi_1}$, LS selects $a_{\pi_1}$ as $t'_i$ in step (3). If $s_i$ does not contain $a_{\pi_1}$ but contains $a_{\pi_2}$, LS selects $a_{\pi_2}$. If $s_i$ does not contain $a_{\pi_1}, \ldots, a_{\pi_k}$ but contains $a_{\pi_{k+1}}$, LS selects $a_{\pi_{k+1}}$. ¿From this property and the fact that the score increases monotonically, we have:

**Observation 1.** The same $\pi$ does not appear more than twice in LS.

¿From this observation, we have:

**Proposition 2.** The number of iteration steps is $O(A!)$ if $L = 1$ and $p(a) = \frac{1}{A}$.

Proposition 2 is interesting because the bound is not affected by $n$. If $A$ is small (e.g., in a case of DNA sequences), $A!$ is not so large. However, $A!$ will be very large if $A$ is not small (e.g., in a case of protein sequences). So, we need other bounds.

Next, look at Fig. 1. ¿From Fig. 1, you can see that decrease of $\#a_{\pi_k}$ contributes to increase of $\#a_{\pi_h}$ such that $h < k$. That is, decrease of $\#a_{\pi_k}$ does not contribute to increase of $\#a_{\pi_h}$ such that $h > k$. Therefore, it is seen that, for any $k$, $\sum_{i=1}^{k} \#a_{\pi_i}$ does not decrease. Since $\sum_{i=1}^{k} \#a_{\pi_i} \leq n$ holds for all $k = 1, \ldots, A$ and $\sum_{i=1}^{k} \#a_{\pi_i}$ must increase for at least one $k$ until LS reaches a local optimal (otherwise, LS reaches a local optimal), the number of iteration steps is bounded by $O(nA)$.

**Proposition 3.** The number of iteration steps is $O(nA)$ if $L = 1$ and $p(a) = \frac{1}{A}$.

We can also make an example in which $\Omega(n)$ steps are required by using an alphabet of size $\Theta(\sqrt{n})$.

## 3 Application to Class Discovery for Cancer Classification

We applied local multiple alignment to class discovery for cancer cells. For that purpose, we considered the problem of finding a most significant cluster. Assume that each gene expression level is rounded to 0 (low) or 1 (high) using appropriate threshold values. Then, we define the problem of finding a most significant cluster as follows.

**Most Significant Cluster**: Given a set $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ of sequences of length $m$ over $\Sigma = \{0, 1\}$ and an integer $k$, find a set of sequences $\{s_{i_1}, \ldots, s_{i_k}\} \subseteq \mathcal{S}$ maximizing the relative entropy score.

In this definition, $s_i$ corresponds to an expression pattern of $i$-th gene ($g_i$), and $s_i[j]$ corresponds to the (rounded) expression level of gene $g_i$ for sample $j$. It is expected that the most significant cluster contains useful information for class distinction.

As in local multiple alignment [1], this problem is NP-hard. Since this problem is quite similar to local multiple alignment, we developed the following local search algorithm by modifying the algorithm in Section 2.

(1) Select $k$ sequences $s_{i_1}, \ldots, s_{i_k}$ from $\mathcal{S}$ at uniformly random.
(2) Let $f_j(a)$ be the frequency of letter $a$ in the $j$-th column of $s_{i_h}$ $(h = 1, \ldots, k)$.
(3) Select sequences $s_{i'_1}, \ldots, s_{i'_k}$ from $\mathcal{S}$ which have $k$ best relative entropy scores.
(4) Replace $(s_{i_1}, \ldots, s_{i_k})$ with $(s_{i'_1}, \ldots, s_{i'_k})$.
(5) Repeat (2)-(4) until reaching a local optimum.

We made a preliminary computational experiment using a data set of real expression patterns obtained by Golub *et al* [4]. The algorithm was applied to two cases: classification of ALL(acute lymphoblastic leukemia) and AML (acute myeloid leukemia) and classification of B-cell ALL and T-cell ALL. For each case, a set of 25 genes was selected as a most significant cluster. In the case of classification of ALL and AML, only 4 samples among 38 samples were misclassified even if classification was done by using the simple majority voting. In the case of classification of T-cell ALL and B-cell AML, only 1 sample among 27 samples was misclassified.

## References


[1] T. Akutsu, H. Arimura, S. Shimozono. On approximation algorithms for local multiple alignment. *Proc. 4th ACM Int. Conf. Computational Molecular Biology*, 1–7, 2000.

[2] J. L. DeRisi, V. R. Lyer, P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.

[3] R. Durbin, S. Eddy, A. Krogh, G. Mitchison. *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[4] T. R. Golub *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[5] P. Horton. Alignment vs. sum of all alignments scoring for motif extraction. *Proc. 6th SIGMPS Symposium*, IPSJ, 2000.

[6] C. E. Lawrence, A. A. Reilly. An expectation maximization (EM) algorithm for identification and characterization of common sites in unaligned biopolymer sequences. *PROTEINS: Structure, Function, and Genetics*, 7:41–51, 1990.

[7] C. E. Lawrence *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262:208–214, 1993.

[8] M. Li, B. Ma, L. Wang. Finding similar regions in many strings. *Proc. 30th ACM Symp. Theory of Computing*, 473–482, 1999.

[9] G. Stormo, G. W. Hartzell. Identifying protein-binding sites from unaligned DNA fragments. *Nucleic Acids Research*, 86:1183–1187, 1989.