Multiple Alignment of Biological Sequences Containing Tandem Repeat

Hidetoshi Kitada [1] kitada@huie.hokudai.ac.jp
Kazuhiro Tono [1] herzen@huie.hokudai.ac.jp
Masahito Yamamoto [1] masahito@huie.hokudai.ac.jp
Tamotsu Mitamura [1] mitamura@huie.hokudai.ac.jp
Azuma Ohuchi [1] ohuchi@huie.hokudai.ac.jp
Toshio ohyanagi [2] ohyanagi@shs.sapmed.ac.jp
Norio Matsushima [2] matusima@shs.sapmed.ac.jp

[1] Division of Systems and Information Engineering, Faculty of Engineering, Hokkaido University
Kita 13, Nishi 8, Kita-ku, Sapporo, Hokkaido 060, Japan

[2] School of Health Sciences, Sapporo Medical University
Minami 1, Nishi 17, Chuou-ku, Sapporo, Hokkaido 064, Japan

Abstract

Multiple sequence alignment can be a useful technique for studying molecular evolution and analyzing sequence-structure relationships. The multiple sequence alignment problem may be formulated as an optimization problem and various approaches to solving it have been applied in the past. But when the sequence contains tandem repeat, the conventional methods cannot yield the biologically significant alignment. We present a method for the multiple alignment of sequences containing tandem repeat.

1 Introduction

Recently, it was shown that many protein sequences contain subsequences which display similarity. The similarities between the subsequences are a result of gene duplication of some subsequence and/or insertion, deletion and substitution of a base. Similar subsequences continuously repeated is called a "tandem repeat". A basic pattern of tandem repeat is called a "repeat unit". Tandem repeats in biological sequences represent a significant fraction of the total sequence database. For analyzing these tandem repeats, multiple sequence alignment is very helpful. However, the conventional method cannot yield the biologically significant alignment for sequences with tandem repeats. We presents a multiple alignment method for biological sequence containing tandem repeat.

2 Method

Tandem repeats are considered as the result of gene duplication of some subsequence. Some repeat unit forms a tertiary substructure. In consideration of them, the frame of repeat unit must not be divided in the process of an alignment. For example, when S_1 = `ABCDABCDABCEAFCD` , S_2 = `ABCDABCDABCDABCE` and the consensus repeat unit (although not perfectly conserved in any position) should be `ABCD`, then an alignment of S_1 and S_2 by the conventional method is shown as a follow.

```
ABCD ABCD ABCE- AF-CD-
ABCD ABCD ABC-D A-BC-E
```

But the above alignment is not suitable from the biological point of view. The biologically significant alignment can be obtained by comparing each of the repeat units. An example of such an alignment is shown as a follow.

```
ABCD ABCD ---- ABCE AFCD
ABCD ABCD ABCD ABCE ----
```

To obtain such an alignment, our proposed method takes the following steps. 1) search the sites of repeat units, 2) calculate the cost function comparing each of the repeat units, 3) treat repeat units as single characters and align sequences.

In Step 1, the start and end sites of all repeat units in all input sequences are specified. In Step 2, The cost function $rsub(A, B)$ of substituting a repeat unit B for a repeat unit A is calculated. $rsub(A, B)$ is defined by an optimal alignment cost of A and B. $rsub$ is calculated by the method based on Dynamic Programming [1]. In addition, the cost function of substituting a repeat unit for a single character is also calculated by the function, that is, A is a repeat unit and B is a repeat unit whose length is 1. In Step 3, all repeat units is replaced single characters. For example, in the sequence $S_1 = a_1a_2...a_{i-1}a_ia_{i+1}a_{i+2}a_{i+3}a_{i+4}a_{i+5}a_{i+6}...$, let a subsequence $a_ia_{i+1}a_{i+2}$ be a repeat unit $r_{1,1}$ and $a_{i+3}a_{i+4}a_{i+5}$ be a repeat unit $r_{1,2}$. Then an ordinary alignment algorithm can be applied to a sequence $a_1a_2...a_{i-1}r_{1,1}r_{1,2}a_{i+6}...$ by use of $rsub$. This procedure can also apply to sequences whose some different tandem repeats are located at intervals.

References

[1] S. B. Needlemann, C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequences of two proteins." *Journal of Molecular Biology*, Vol. 48, pp. 444-453, 1970.