配列のローカル・アラインメントの困難さについて

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

複数の文字列が与えられた時、類似性やエントロピーが最大となるように各文字列から同じ長さの部分文字列を切り出してくる問題はギャップ無しのローカル・マルチプル・アラインメントと呼ばれ、似た機能を持つDNA配列やアミノ酸配列から特徴的な部分を取り出してくるのに有用である。この問題に対し、EMアルゴリズムやギップスサンプリングなどに基づくヒューリスティックなアルゴリズムや分枝限定法によるアルゴリズムが提案されてきたが、最適解を計算する多項式時間アルゴリズムは開発されていなかった。本稿では、エントロピーおよびSP(Sum of Pairs)スコアのいずれのスコアを用いても、この問題がNP困難であることを示す。

Hardness Results on Gapless Local Multiple Sequence Alignment

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

Gapless local multiple sequence alignment is a problem of, given a set of strings, selecting a fixed length substring from each string so that the total similarity among selecting strings is maximized. It is useful for finding characteristic string patterns (i.e., motifs) from DNA or amino acid sequences that have similar biological functions. For this problem, several heuristic algorithms have been developed using such techniques as expectation maximization and Gibbs sampling. Exact algorithms based on branch-and-bound techniques have also been developed. However, no polynomial time algorithm to compute an optimal solution was known. In this short article, we prove that the problem is NP-hard under both information theoretic entropy scoring scheme and SP (sum-of-pairs) scoring scheme.

1 Introduction

Finding "motifs" from DNA or amino acid sequences with similar biological functions is one of well-studied computational problems in molecular biology, where a motif is a characteristic string pattern. *Gapless local multiple alignment* is one of useful tools for finding motifs. Gapless local multiple sequence alignment is a computational problem of, given a set of strings, selecting a fixed length substring from each string so that the total similarity among selecting strings is maximized. For this problem, several heuristic algorithms have been developed using such techniques as beam search [8], expectation maximization [5], and Gibbs sampling [6]. But, these algorithms may miss optimal solutions. In order to find optimal solutions, a branch and bound algorithm and an exact algorithm have been developed [3, 4]. However, in the worst case, they use exponential time. Thus, it is natural to ask whether or not gapless local multiple alignment is NP-hard.

In this short article, we prove that this problem is NP-hard under both information theoretic entropy scoring scheme and SP (sum-of-pairs) scoring scheme, where both scoring schemes are widely used in molecular biology.

2 Hardness Result under Entropy Scoring Scheme

In this section, we prove that gapless local multiple alignment is NP-hard under information theoretic entropy scoring scheme. First we define the problem formally [4]. For simplicity, we consider strings over an alphabet $\Sigma = \{0, 1\}$. We are given a set of strings $\{S_1, S_2, \ldots, S_N\}$ over $\{0, 1\}$, and a length W of the window. From each S_i, we extract a substring of length W. Let $N_{i,0}$ (resp. $N_{i,1}$) be the number of 0's (resp. 1's) in the substring extracted from S_i. Then, the total score (i.e., entropy) is given by

$$\sum_{i=1}^{W} -N_{i,0} \log \frac{N_{i,0}}{N} - N_{i,1} \log \frac{N_{i,1}}{N}.$$

Then, the problem is to extract substrings which minimize the total score (i.e., maximize $\sum_{i=1}^{W}(N_{i,0} \log N_{i,0} + N_{i,1} \log N_{i,1})$).

Here we briefly give an example of the problem. Let $N = 5$ and $W = 4$, and $S_1 = 01011101$, $S_2 = 1110101$, $S_3 = 011110110$, $S_4 = 00011100$, $S_5 = 110011$. Then, the following is an optimal alignment:

```
   WWWW
01011101
   1110101
 011110110
00011100
   110011
```

where WWWW denotes the window (i.e., substrings just below WWWW are selected).

Theorem 1. Gapless local multiple alignment is NP-hard under information theoretic entropy scoring scheme.

(Proof) In order to show NP-hardness, we use a polynomial time reduction from MIN 2SAT, which is known to be NP-complete [1]. MIN 2SAT is, given a set of clauses $\{c_1, ..., c_m\}$ over a set of variables $\{x_1, ..., x_n\}$ where each clause consists of two literals (i.e., each clause has one of the following form: $x_i \vee x_j$, $x_i \vee \neg x_j$, $\neg x_i \vee \neg x_j$), to find a truth assignment such that the number of satisfied clauses is minimized.

From an instance of MIN 2SAT, we construct $2n - 3$ strings each of which has the following form

$$S_i \;=\; A.B_i.A.D_i.A$$

where $x.y$ denotes a concatenation of x and y, and A, B_i, D_i are substrings defined below. We let the window size be $W = |A| + |B_i| + |A|$ ($|B_i| = |D_i| = m$).

A is constructed so that A must be aligned with A (otherwise the score would not become the minimum). For example, A can be the following form:

$$(n)\;\; 0000\;\; (n-1)\;\; 0000\;\; (n-2)\;\; 0000\;\; \ldots\;\; 0000\;\; (2)\;\; 0000\;\; (1)\;\; 0000$$

where each (i) denote a binary string representing the number i (using $\log n$ bits) and 0000 denotes the string with $\log n$ 0's.

Let $B_i = b_{i,1}b_{i,2}\ldots b_{i,m}$. For $i = 1, \ldots, n$, we define $b_{i,j}$ as follows: $b_{i,j} = 1$ if x_i appears in c_j, otherwise $b_{i,j} = 0$. For $i = n+1, \ldots, 2n-3$, we define $b_{i,j} = 1$ for all j.

Let $D_i = d_{i,1}d_{i,2}\ldots d_{i,m}$. For $i = 1, \ldots, n$, we define $d_{i,j}$ as follows: $d_{i,j} = 1$ if $\neg x_i$ appears in c_j, otherwise $d_{i,j} = 0$. For $i = n+1, \ldots, 2n-3$, we define $d_{i,j} = 1$ for all j.

Then, we can consider the following correspondence:

- $x_i = 1 \iff A \cdot B_i \cdot A$ is selected from S_i as a substring of length W,
- $x_i = 0 \iff A \cdot D_i \cdot A$ is selected from S_i as a substring of length W.

Based on this, we have the following correspondence:

(i) c_j is not satisfied $\iff N_{j,0} = n$ and $N_{j,1} = n - 3$,

(ii) c_j is satisfied $\iff (N_{j,0} = n - 1$ and $N_{j,1} = n - 2)$ or $(N_{j,0} = n - 2$ and $N_{j,1} = n - 1)$,

where we only consider columns corresponding to B_i's and D_i's. It is easy to see that the score for case (i) is lower than the score for case (ii).

Therefore, the total score is minimized if and only if the number of satisfied clauses is minimized. Since the above construction can be done in polynomial time, the theorem holds. □

Here we give an example of a construction.

MIN 2SAT: $\{x_1 \vee \neg x_2, \neg x_2 \vee x_3, \neg x_1 \vee x_4\}$.

Constructed strings: $S_1 = A.100.A.001.A$, $S_2 = A.000.A.110.A$, $S_3 = A.010.A.000.A$, $S_4 = A.001.A.000.A$, $S_5 = A.111.A.111.A$.

Then, the followings are optimal alignments:

```
      WWWWWWWW              WWWWWWWW              WWWWWWWW
A.100.A.001.A         A.100.A.001.A         A.100.A.001.A
      A.000.A.110.A         A.000.A.110.A         A.000.A.110.A
A.010.A.000.A         A.010.A.000.A     A.010.A.000.A
A.001.A.000.A               A.001.A.000.A     A.001.A.000.A
A.111.A.111.A         A.111.A.111.A     A.111.A.111.A
```

Note that an alignment in the left part corresponds to a case of $x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 0$, $c_1 = 0, c_2 = 0, c_3 = 1$, an alignment in the middle part corresponds to a case of $x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 1$, $c_1 = 0, c_2 = 0, c_3 = 1$, and an alignment in the right part corresponds to a case of $x_1 = 1, x_2 = 1, x_3 = 0, x_4 = 0$, $c_1 = 1, c_2 = 0, c_3 = 0$.

3 Hardness Result under Sum-of-Pairs Scoring Scheme

In this section, we briefly show an NP-hardness result for SP (sum-of-pairs) scoring scheme. In SP scheme, we are given a function $s(x, y)$ from $\Sigma \times \Sigma$ to a set of reals, where $s(x, y) = s(y, x)$. Let $s_i^1 s_i^2 \ldots s_i^W$ be the substring selected from S_i. Then, the total score is defined by

$$\sum_{k=1}^{W} \sum_{i<j} s(s_i^k, s_j^k).$$

Note that, in this case, the problem is defined as a maximization problem and thus the score should be maximized.

Theorem 2. Gapless local multiple alignment is NP-hard under SP (sum-of-pairs) scoring scheme.

(Proof) We use a reduction from MAXIMUM INDEPENDENT SET as in [7]. MAXIMUM INDEPENDENT SET is, given an undirected graph (V, E) and an integer K, decide whether or not there exists $U \subseteq V$ such that $|U| \geq K$ and $(\forall u, v \in U)(\{u, v\} \notin E)$.

From an instance of MAXIMUM INDEPENDENT SET ($|V| = n, |E| = m$), we construct the following set of sequences over $\Sigma = \{-, a, b, 0, 1\}$: (i) for each $i = 1, \ldots, K$, we construct $A_i = (-)^{2n}\, a\, (-)^{2n}$ where $(-)^{2n}$ means $2n$ '$-$'s; (ii) for each $v_i \in V$, we construct a string V_i as below:

```
                        vi                                  vj1 vj2  vj3
                         |                                   |   |    |
V_i: bbbbbbbbbbbbb 1111011111111 bbbbbbbbbbbbb 1101110111101 bbbbbbbbbbbbb
      length n       length n       length n       length n       length n
```

where $v_{j_1}, \cdots, v_{j_h}$ are adjacent nodes of v_i.

We define score function by $s(a, 1) = s(1, a) = 1$, $s(a, -) = s(-, a) = 1$, otherwise $s(x, y) = 0$. We define $W = 3n$.

Then, the score of an alignment is at least $Kn + K(K-1)$ if and only if there exists an independent set of size K. □

4 Concluding Remarks

We proved that gapless local multiple alignment is NP-hard under both information theoretic entropy scoring scheme and SP (sum-of-pairs) scoring scheme. This result gives a justification for using heuristic algorithms that were previously developed. From a theoretical viewpoint, development of an approximation algorithm with guaranteed approximation ratio (especially, under information theoretic entropy scoring scheme) is important future work.

Acknowledgement

The author would like to thank Dr. Paul Horton for valuable discussions.

References

[1] M.R. Garey and D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, Freeman, San Francisco (1979).

[2] D. Gusfield, Efficient methods for multiple sequence alignment with guaranteed error bounds, *Bulletin of Mathematical Biology* **55** (1993) 141–154.

[3] P. Horton, A branch and bound algorithm for local multiple alignment, *Proc. Pacific Symp. Biocomputing '96* (PSB96) (1996) 368–383.

[4] P. Horton, New algorithms for multiple sequence alignment, *Technical Report* **98-MPS-22-6**, Information Processing Society of Japan (1998).

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

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

[7] D. Maier, The complexity of some problems on subsequence problem over binary alphabet is NP-complete, *J. ACM* **25** (1978) 322-336.

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

[9] L. Wang and T. Jiang: On the complexity of multiple sequence alignment, *Journal of Computational Biology* **1** (1994) 337–348.