

行政院國家科學委員會專題研究計畫 期中進度報告

比較相似度間斷序列的動態規劃演算法(1/3)

計畫類別：個別型計畫

計畫編號：NSC92-2213-E-002-059-

執行期間：92年08月01日至93年07月31日

執行單位：國立臺灣大學資訊工程學系暨研究所

計畫主持人：趙坤茂

報告類型：精簡報告

處理方式：本計畫可公開查詢

中 華 民 國 93 年 5 月 31 日

行政院國家科學委員會補助專題研究計畫 成果報告
v 期中進度報告

比較相似度間間斷斷序列的動態規劃演算法(1/3)

計畫類別： 個別型計畫 整合型計畫

計畫編號：92-2213-E-002-059-

執行期間： 92 年 8 月 1 日至 93 年 8 月 1 日

計畫主持人： 趙坤茂

共同主持人：

計畫參與人員：

成果報告類型(依經費核定清單規定繳交)： 精簡報告 完整報告

本成果報告包括以下應繳交之附件：

赴國外出差或研習心得報告一份

赴大陸地區出差或研習心得報告一份

出席國際學術會議心得報告及發表之論文各一份

國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、
列管計畫及下列情形者外，得立即公開查詢
涉及專利或其他智慧財產權，一年二年後可公開查詢

執行單位：台灣大學資訊工程系

中 華 民 國 93 年 5 月 30 日

(一) 計畫中文摘要

關鍵字：演算法、生物序列分析、序列排比、基因組分析

同源序列有時候幾個區域局部相似；而又有幾個區域不相似。因此，如果不同的區域遠比相似區域長的話，同源序列的整體相似度可能會很低。

在這個計畫裡，我們提出了一個新的擴充化的排比演算法，它將可用來比較相似度間間斷斷的序列，整理出一個由相異區域隔開的相似區域串列。我們所發展出的動態規劃演算法將只需要與兩個序列長度乘積成正比的時間，以及與兩個序列長度總和成正比的空間。

在第一年裡，我們設計了新排比分數計算的遞迴關係式，並且發展快速有效的方法來實際算出最佳排比，我們並不會儲存所有的矩陣資料，因為這將耗費太多的計算空間，而是藉由有效的回朔方式來達成省空間的需求。所有的方法我們都用 C 語言寫成一個整合式的軟體工具，它可以找低相似度的同源序列。

(二) 計畫英文摘要

Keywords: Algorithms, biomolecular sequence analysis, sequence alignment, comparative genomics.

Homologous sequences are sometimes similar over some regions but different over other regions. Homologous sequences have a much lower global similarity if the different regions are much longer than the similar regions.

In this project, we propose a generalized global alignment algorithm for comparing sequences with intermittent similarities, an ordered list of similar regions separated by different regions. A dynamic programming algorithm will be designed to compute an optimal general alignment in time proportional to the product of sequence lengths and in space proportional to the sum of sequence lengths.

In the first year, we have formulated the recurrences for generalized global alignments, and developed efficient algorithm for computing them. All the algorithms are implemented in C as an integrated software tool. This tool extends the ability of standard global alignment programs to recognize homologous sequences of lower similarity.

報告內容

Global alignment algorithms are intended for comparing two sequences that are entirely similar (Needleman and Wunsch, 1970; Sellers, 1974; Wagner and Fisher, 1974; Waterman et al., 1976; Gotoh, 1982; Myers and Miller, 1988). Local alignment algorithms are intended for comparing sequences that contain locally similar regions (Smith and Waterman, 1981; Gotoh, 1982; Waterman and Eggert, 1987; Pearson and Lipman, 1988; Altschul et al., 1990; Huang and Miller, 1991; Burkhardt et al., 1999; Kurtz and Schleiermacher, 1999; Ma et al., 2002). Those methods are very useful in analysis of DNA and protein sequences. In this project, we generalize the global alignment algorithms to compare sequences with intermittent similarities.

Homologous sequences are sometimes similar over some regions but different over other regions. For example, homologous genomic DNA sequences from related organisms such as

human and mouse are usually similar over exon regions but different over intron regions. Homologous protein sequences are sometimes similar over some conserved domains but different over variable regions. If different regions are much longer than similar regions, homologous sequences are not similar over their entire lengths. We plan to develop a generalized global alignment model to address sequences with intermittent similarities and design a dynamic programming algorithm for computing an optimal general alignment of two sequences. The algorithm runs in time proportional to the product of sequence lengths and in space proportional to the sum of sequence lengths.

A number of fast comparison programs have been developed specially for comparing homologous and syntenic genomic DNA sequences (Delcher et al., 1999; Jareborg et al., 1999; Batzoglou et al., 2000; Schwartz et al., 2000) Our program is more sensitive than the fast comparison programs because it searches the entire solution space and produces an optimal solution. The improved sensitivity of optimal alignment algorithms over fast comparison algorithms is confirmed by Pearson (1995) in a comprehensive study. Thus, it is suitable for comparing short sequences such as protein sequences.

Let $A = a_1 a_2 \dots a_m$ and $B = b_1 b_2 \dots b_n$ be two sequences of lengths m and n . A generalized global alignment (or general alignment) of A and B consists of substitutions, gaps, and difference blocks. A substitution associates a residue of A with a residue of B . A gap consists of only residues from one sequence with each residue associated with the symbol '-'. A difference block consists of residues from one or two sequences with each residue associated with the symbol '+'. Let (a, b) be the score of a substitution involving residues a and b . The score of a gap of length k is $-(q + k \times r)$, where nonnegative numbers q and r are gap-open and gap-extension penalties, respectively. The score of a difference block is $-d$, where nonnegative number d is a constant penalty for each difference block. The score of a general alignment is the sum of scores of each substitution, each gap, and each difference block in the alignment. An optimal general alignment is one with the maximum score.

An algorithm for computing an optimal general alignment of A and B is developed using dynamic programming. Define $S(i, j)$ to be the maximum score of general alignments ending at positions i and j of A and B . Then $S(m, n)$ is the score of an optimal general alignment of A and B . To compute the matrix S efficiently, three additional matrices are introduced. Define $H(i, j)$ to be the maximum score of general alignments ending with a difference block at positions i and j of A and B . Similarly, define $D(i, j)$ for general alignments that end with a deletion gap and $I(i, j)$ for general alignments that end with an insertion gap. In the first year, we have formulated the recurrences for generalized global alignments, and developed efficient algorithm for computing them.

$$\begin{aligned}
S(0, 0) &= 0, \\
S(i, 0) &= \max\{D(i, 0), H(i, 0)\} \text{ for } i > 0, \\
S(0, j) &= \max\{I(0, j), H(0, j)\} \text{ for } j > 0, \\
S(i, j) &= \max\{S(i - 1, j - 1) + \sigma(a_i, b_j), D(i, j), I(i, j), H(i, j)\} \\
&\quad \text{for } i > 0 \text{ and } j > 0.
\end{aligned}$$

$$\begin{aligned}
D(0, j) &= S(0, j) - q \text{ for } j \geq 0, \\
D(i, 0) &= D(i - 1, 0) - r \text{ for } i > 0, \\
D(i, j) &= \max\{D(i - 1, j) - r, S(i - 1, j) - q - r\} \\
&\quad \text{for } i > 0 \text{ and } j > 0.
\end{aligned}$$

$$\begin{aligned}
I(i, 0) &= S(i, 0) - q \text{ for } i \geq 0, \\
I(0, j) &= I(0, j - 1) - r \text{ for } j > 0, \\
I(i, j) &= \max\{I(i, j - 1) - r, S(i, j - 1) - q - r\} \\
&\quad \text{for } i > 0 \text{ and } j > 0.
\end{aligned}$$

$$\begin{aligned}
H(i, j) &= -d \text{ for } i = 0 \text{ or } j = 0, \\
H(i, j) &= \max\{H(i, j - 1), H(i - 1, j), S(i, j - 1) - d, S(i - 1, j) - d\} \\
&\quad \text{for } i > 0 \text{ and } j > 0.
\end{aligned}$$

Consider a largest-scoring alignment $P(i, j)$ ending with a difference block at positions i and j of A and B . The score of the alignment is $H(i, j)$. Note that either residue i of A or residue j of B is part of the difference block. Assume that residue i is part of the difference block. Let $P(i-1, j)$ denote the entire portion of $P(i, j)$ before residue i . If the difference block consists only of residue i , then the alignment $P(i-1, j)$ ends with a substitution or gap and its score has to be $S(i-1, j)$. Thus, the score of the alignment $P(i, j)$ is equal to the score of the alignment $P(i-1, j)$ minus the penalty for the difference block, that is, $H(i, j) = S(i-1, j) - d$. If the difference block consists of residue i and other residues, then the alignment $P(i-1, j)$ also ends with a difference block and its score has to be $H(i-1, j)$. Thus, the score of the alignment $P(i, j)$ is equal to the score of the alignment $P(i-1, j)$, that is, $H(i, j) = H(i-1, j)$ because the penalty for the difference block is already included in $H(i-1, j)$. Similarly, if residue j is part of the difference block, we can show that either $H(i, j) = S(i, j-1) - d$ or $H(i, j) = H(i, j-1)$.

References:

- [1] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. “Basic local alignment search tool.” *J. Mol. Biol.*, 215, 403-410, 1990.
- [2] Ansari-Lari, M.A., Oeltjen, J.C., Schwartz, S., Zhang, Z., Muzny, D.M., Lu, J., Gorrell, J.H., Chinault, A.C., Belmont, J.W., Miller, W. and Gibbs, R.A. “Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse

chromosome 6.” *Genome Res.*, 8, 29-40, 1998.

- [3] Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B. and Lander, E.S. “Human and mouse gene structure: comparative analysis and application to exon prediction.” *Genome Res.*, 10, 950-958, 2000.
- [4] Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O. and Salzberg, S.L. “Alignment of whole genomes.” *Nucleic Acids Res.*, 27, 2369-2376, 1999.
- [5] Eddy, S.R. “Profile hidden Markov models.” *Bioinformatics*, 14, 755-763, 1998.
- [6] Gotoh, O. “An improved algorithm for matching biological sequences.” *J. Mol. Biol.*, 162, 705-708, 1982.
- [7] Hirschberg, D.S. “A linear space algorithm for computing maximal common subsequences.” *Commun. Assoc. Comput. Mach.*, 18, 341-343. 1975.
- [8] Huang, X. “On global sequence alignment.” *Comput. Appl. Biosci.*, 10, 227-235, 1994.
- [9] Huang, X. and Miller, W. “A time-efficient, linear-space local similarity algorithm.” *Adv. Appl. Math.*, 12, 337-357, 1991.
- [10] Jareborg, N., Birney, E. and Durbin, R. “Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs.” *Genome Res.*, 9, 815-824, 1999.
- [11] Ma, B., Tromp, J. and Li, M. “PatternHunter: faster and more sensitive homology search.” *Bioinformatics*, 18, 440-445.
- [12] Myers, E.W. and Miller, W. “Optimal alignments in linear space.” *Comput. Applic. Biosci.*, 4, 11-17, 1988.
- [13] Needleman, S.B. and Wunsch, C.D. “A general method applicable to the search for similarities in the amino acid sequences of two proteins.” *J. Mol. Biol.*, 48, 443-453, 1970.
- [14] Pearson, W.R. “Comparison of methods for searching protein sequence databases.” *Protein Science*, 4, 1145-1160, 1995.
- [15] Pearson, W.R. and Lipman, D. “Improved tools for biological sequence comparison.” *Proc. Natl. Acad. Sci. USA*, 85, 2444-2448, 1988.
- [16] Sellers, P.H. “On the theory and computation of evolutionary distances.” *SIAM J. Appl. Math.*, 26, 787-793, 1974.
- [17] Smit, A and Green, P. “<http://ftp.genome.washington.edu/RM/RepeatMasker.html>”, 1997.
- [18] Smith, T.F. and Waterman, M.S. “Identification of common molecular subsequences.” *J. Mol. Biol.*, 147, 195-197, 1981.
- [19] Thompson, J.D., Higgins, D.G. and Gibson, T.J. “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.” *Nucleic Acids Res.*, 22, 4673-4680, 1994.
- [20] Wagner R.A. and Fischer, M.J. “The string-to-string correction problem.” *J. Assoc. Comput. Mach.*, 21, 168-173, 1974.
- [21] Waterman, M.S. and Eggert, M. “A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons.” *J. Mol. Biol.*, 197, 723-728, 1987.
- [22] Waterman, M.S., Smith, T.F. and Beyer, W.A. “Some biological sequence metrics.” *Adv. Math.*, 20, 367-387, 1976.
- [23] Schwartz, S., Zhang, Z., Frazer, K., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R. and Miller, W. “PipMaker--A web server for aligning two genomic DNA sequences.” *Genome Res.*, 10, 577-586, 2000.