行政院國家科學委員會專題研究計畫　成果報告

多重基因組序列的快速排比方法(3/3)

計畫類別：個別型計畫
計畫編號：NSC91-2213-E-002-129-
執行期間：91年08月01日至92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：完整報告

處理方式：本計畫可公開查詢

中　華　民　國 92年10月29日

行政院國家科學委員會專題研究計畫成果報告

多重基因組序列的快速排比方法(1/3 - 3/3)

計畫編號：NSC 91-2213-E-002-129

執行期限：89 年 8 月 1 日至 92 年 7 月 31 日

執行機關：國立台灣大學資訊工程學系

主持人：國立台灣大學資訊工程學系教授趙坤茂(Kun-Mao Chao)

(email: kmchao@csie.ntu.edu.tw)

中文摘要

隨著基因組定序技術的成熟，愈來愈多的生物基因組序列已被決定出來了，在不久的將來，我們人類的基因組序列初稿也要完成，緊接著將是老鼠、雞、魚…。這些涵蓋該生物所有生命活動遺傳訊息的基因組序列，是我們現階段亟需分析與歸納的資料。透過多重序列排比，我們可以找出生物序列中的保守區域、決定基因規則及推測演化過程。然而，這些基因組序列很大的特色就是它們非常長，即使我們只是比較其中的一些片段，常常也是數以百萬計的鹼基，如果我們以現有的軟體工具來做排比，在計算時間及空間上，都是行不通的。本計畫主要目的就是設計一套可排比多重基因組序列的軟體工具，希望能透過多個基因組序列的比較，協助生物學家探索整個基因體的結構及功能。

我們的構想是以某個基因體序列為基底，首先將其他的基因組序列(或序列片段)很快速地與基底序列做比較，在這比較後，我們可以得到初步的定位。然後我們再將這些序列與基底序列根據這些定位堆疊起來，如此一來，我們就得到了一個較為粗糙的多重序列排比。我們已將這個基本雛型製作好，並進行相關測試。我們也設計了一個改良式的多序列比較分析工具，可精確地計算多個序列的排比分數。此外，我們也擴充了原始排比內排在一起的區段，以便使得排比分數更佳化。

關鍵詞：序列分析、計算基因體學、計算生物學

Abstract

Due to the advancement of genome sequencing technology, more and more genomic sequences have been determined. In the near future, the draft of human genomic sequence will be finished. World-wide sequencing capacity is ramping up to the level of one vertebrate genome per year, and after the human and mouse genomes are completed it will turn to chicken, fish, rat, etc. These data, which essentially encode all the genetic information in life, will soon need to be analyzed and classified. By multiple sequence comparison, we are able to locate the conserved regions in the biological sequences. It can also be used to study gene regulation or even infer evolutionary trees. However, these genomic sequences are usually very long. As the sequences are getting longer and longer, there is no doubt that time-efficient and space-saving strategies for multiple sequence alignments will become more and more important in the near future. The purpose of this project is to design a software tool for aligning multiple genomic sequences. It will be used to explore the structure and function of a whole genome sequence.

Our idea is based on a given genomic sequence. We first use a very fast method to compare other sequences with the base sequence. Then we roughly determine their relative location. By pasting these sequences according to their relativity, a simple multiple sequence alignment can be derived. We have implemented a simple multiple alignment program. We have also implemented an efficient algorithm that can accurately compute the score of a multiple sequence alignment. We have adjusted the bias of the base sequence by extending the segments which were aligned together in the crude alignment.

Keywords：Sequence analysis, computational genomics, computational biology.

We have surveyed the literatures relevant to the multiple sequence alignment problem. In particular, we are interested in the alignment methods dealing with long sequences. In large-scale sequencing projects, the task of converting experimental data into biologically relevant information requires a higher level of abstraction in sequence analysis. Therefore, we have also developed a prototype for genomic sequence visualization tools. A graphic interface allows the user to zoom into any specific area of the resulting alignment.

We first compare the selected genomic sequence with all other given sequences. Then we develop a simple pasting program for converting these pairwise alignments into a tentative multiple sequence alignment. The pairwise alignments provide the information about the possible coherent multiple alignment columns in sequences. What we do here is more or less a pile-up procedure for aligning all sequences together. We first use a very fast method to compare other sequences with the base sequence. Then we roughly determine their relative location. By pasting these sequences according to their relativity, a crude multiple sequence alignment can be derived.

To improve the quality of the multiple sequence alignment, a round-robin iterative improvement of a multiple alignment will be initiated in the next year. The improved alignment tool will be used to test some real-world data.

We comprise software dedicated to the visualization of resulting alignments so that more biological meaningful information can be extracted. It will provide users a reliable data management system which allows the user to manipulate both the sequences as well as the resulting alignment. It will be a framework that allows several tools to work together in a cooperative way under the user's control. Automatic annotation of the alignment will give the users more valuable information.

To improve the quality of the multiple sequence alignment, a round-robin iterative improvement of a multiple alignment is initiated. We start by pasting the alignments together, then repeatedly (1) delete an aligned fragment and (2) align that fragment with the remainder of the multiple alignment (using a variant of our yama2 procedure where we need to optimize based on the fact that one of the two alignments must be a single sequence). The improved alignment tool will be used to test some real-world data.

We continue improving the alignment tool by other approaches. Specifically, we adjust the bias of the base sequence by extending the segments which were aligned together in the crude alignment. That way, we are able to compensate the situations where the segments are more similar to each other (longer local alignments) than they are to the base genomic sequence. The local alignments we find by iteratively improving the crude alignment created from the pairwise alignments with the base genomic sequence encompass these longer alignments in some way.

參考文獻

[1] Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990) A basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.

[2] Altschul, S. and Lipman, D. (1989) Trees, stars, and multiple biological

sequence alignment. *SIAM J. Appl. Math.* **49**, 197-209.

[3] Altschul, S., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389-3402.

[4] Bassett, Jr. D.E., Eisen, M. B. and Boguski, M. S. (1999) Gene expression informatics – it's all in your mine. *Nature Genetics Supplement* 21, 51-55.

[5] Chao, K. -M. (1999) Calign: aligning sequences with restricted affine gap penalties. *Bioinformatics*, 15, 298-304.

[6] Ephremides, A. and Hajek, B. (1998) Information theory and communication networks: an unconsummated union. *IEEE Transactions on Information Theory* **44**, 2416-2434.

[7] Eppstein, D., Gaili, Z., Giancarlo, R. and Italiano, G. (1992) Sparse dynamic programming I: linear cost functions. *Journal of the ACM* **39**, 519-545.

[8] Feng, D. and Doolittle, R. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25**, 351-360.

[9] Gusfield, D. (1997) Algorithms on strings, trees, and sequences: computer science and computational biology. *Cambridge University Press*.

[10] Lenhof, H. Morgenstern, B. and Reinert, K. (1999) An exact solution for the segment-to-segment multiple sequence alignment problem. *Bioinformatics* 15, 203-210.

[11] Medigue, C., Rechenmann, F., Danchin, A. and Viari, A. (1999) Imagene: an integrated computer environment for sequence annotation and analysis. *Bioinformatics* 15, 2-15.

[12] Morgenstern, B., Dress, A., and Werner, T. (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci.* **93**, 12098-12103.

[13] Morgenstern, B., Frech, K., Dress, A. and Werner, T. (1998) DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics* **14**, 290-294.

[14] Mott, R. (1999) Local sequence alignments with monotonic gap penalties. *Bioinformatics* 15, 455-462.

[15] Setubal, J. and Meidanis, J. (1997) Introduction to computational molecular biology. *PWS Publishing Company*.

[16] Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673-4680.

[17] Z. Zhang, P. Berman and W. Miller (1998) Alignments without low-scoring regions. *J. Computational Biology* 5, 197-210.