雙序列比對工具：BLAST 之分析與改進

吳哲賢
中華大學資訊工程研究所
jswu@chu.edu.tw

江欣倩
中華大學資訊工程研究所
m9002025@chu.edu.tw

摘要

本篇論文針對現今應用最為廣泛的序列比對(sequence alignment)工具：BLAST，做為研究之對象。對 BLAST 比對後的結果，設計簡單且快速的演算法，有效率地提昇序列比對分數。首先我們將設計兩種改進方法：內部改進法及外部改進法，然後針對 BLAST 比對結果不佳的序列，分析他們提昇比對分數的情況。最後實驗結果得知：內部改進方法提昇比對分數的空間有限；然而外部改進方法有顯著的分數提昇，表示外部改進方法是一個有效率改進 BLAST 的工具。

關鍵詞：sequence alignment、BLAST

一、簡介

生物體為求適者生存，不斷演進，而生物演化的動力便是來自基因序列的突變(mutation)，序列由於突變而發生替代(substitution)、插入(insertion)、刪除(deletion)等現象，而造就了相異的生物序列。序列比對技術，在生物科技中是最基本，也是非常重要的工具。Needleman 及 Wunsch[4]在 1970 年提出利用動態規劃(dynamic programming)的技巧來進行計分，研發出全域序列比對(global sequence alignment)之最佳解演算法。假設兩序列長度為 M 及 N，其時間複雜度為 O(MN)。在 1981 年，由 Smith 及 Waterman[5]兩位設計局部序列比對(local alignment)演算法，目的為找出兩序列間部分片段比對的最佳結果，其時間複雜度亦為 O(MN)。由於一般序列長度極大，平方時間複雜度的演算法不實用，於是有人提出線性時間複雜度的 Heuristic 演算法。雖然其結果並非最佳解，但能在可接受的線性時間內完成序列比對。較具代表性的為 1985 年由 Lipman 與 Pearson 提出的 FAST (Fast Alignment Sequence Tools)[3]，及 1990 年由 Altschul, Gish, Miller, Myers，及 Lipman 研發出的 BLAST (Basic Local Alignment Search Tools)[1]。

BLAST 是現今最為廣範使用，且功能最為強大的序列比對工具。本篇論文便針對 BLAST，做為研究之對象。較具代表性的網站就是 NCBI，當中除了提供序列資料庫查尋外，也提供 BLAST 比對的功能。由於核？酸與胺基酸之間有相互轉譯的關係，故 BLAST 提供各種型態的查尋序列及資料庫的功能。在 1990 年所提出之原始 BLAST，為求迅速找到相似片段，所以演算法中並不考慮間隔的發生，但也因此比對後的結果和最佳解的誤差極大。於是在 1997 年由 Altschul，Madden，Jinghui Zhang，Miller，及 Lipman[2]等人將原始 BLAST 做了運用及改進，考慮到間隔問題，設計了一個新的版本，稱為 Gapped BLAST。其與原始 BLAST 有兩點不同之處：(一)Gapped BLAST 使用 Two-Hit 方法，(二)Gapped BLAST 利用動態規劃做延伸動作。現今廣範使用的 BLAST，已都是新版的 Gapped BLAST。Gapped BLAST 因為已經考慮到間隔問題，所以比對後的結果和最佳解的誤差極小。但仍有些序列經過 Gapped BLAST 所得到的結果和最佳解的誤差仍大，表示其中還有改進的空間，而我們便是以 Gapped BLAST 誤差大的情況，做為改進研究的對象。

改進演算法可分為內部改進方法及外部改進方法兩步驟。首先對 BLAST 比對結果曲線取一寬度，利用全域序列比對方法，作帶狀動態規劃分析，嘗試提昇比對分數(內部改進方法)。接著對 BLAST 比對結果曲線兩端，運用外部延伸技巧，期望找出 BLAST 未能包含的相似序列片段，再度提高比對分數(外部改進方法)。我們的實驗結果得知：內部改進方法提昇比對分數的空間有限；然而對 BLAST 比對結果較差的情況，外部改進方法有顯著的分數提昇，表示外部改進方法是一個有效率改進 BLAST 的工具。最後我們建議在 BLAST 比對後，再進行外部改進方法，是一個簡單快速且有效率提昇序列比對分數的演算法。

二、內部改進方法

在 Heuristic 演算法中，有一方法可以快速找出兩序列中相似片段，那就是帶狀動態規劃(banded dynamic programming)。假設兩長度相近的

序列 S 與 T，長度為 M 及 N，若其最佳比對結果中含有少量的間隔 則其結果呈現在動態規劃中會接近中央對角線，如圖 1 所示。

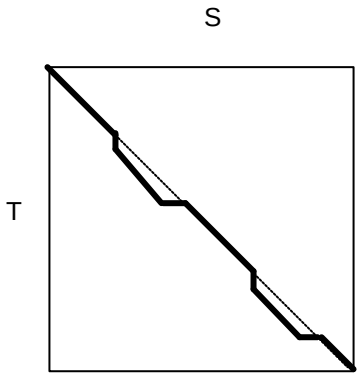

圖 1. S 與 T 序列最佳比對結果接近中央對角線

若要對此種序列進行比對，並不需要如局部序列比對中計算出 M × N 的矩陣，只須針對角線取一寬度 K 值包含最佳比對結果曲線，在此帶狀中去進行動態規劃計算，便可得到最佳比對解，如圖 2，而其時間複雜度只須線性時間 O(KN)。

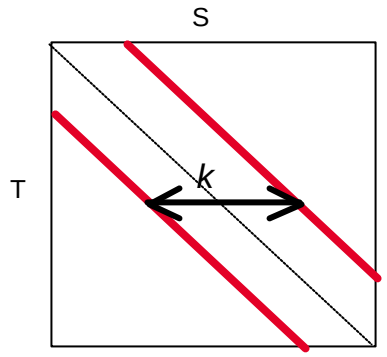

圖 2.帶狀動態規劃：針對中央對角線取一寬度 K，在此帶狀中進行計分運算, 即可找出最佳比對結果

但此種方法的缺點在於, 若兩序列長度並不相近, 或是兩序列中相似片段並不分佈在中央對角線上，則帶狀動態規劃便無法得到最佳比對解。

接著描述如何進行內部改進方法。在執行 BLAST 後，可以得到序列比對結果曲線，以此曲線為中央線，取一寬度 R 值，在此彎曲帶狀中利用動態規劃做全域序列比對，如圖 3。

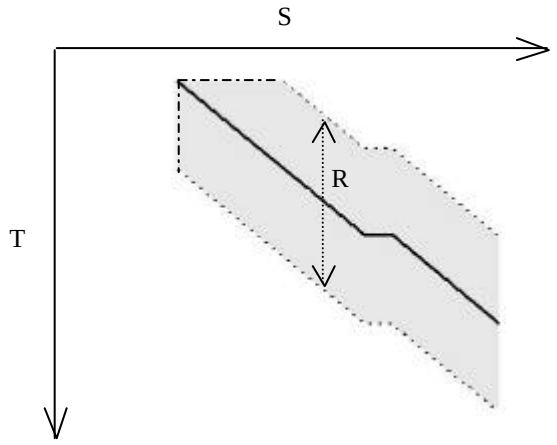

圖 3. 內部帶狀動態規劃示意圖：圖中實線為 Gapped BLAST 所找到的序列片段, 在動態規劃中所呈現的曲線，延此曲線取一寬度 R 就是圖中灰色部份

動態規劃中計分方法與全域序列比對是一樣的，唯一不同的是當點位在臨界邊上時，我們選擇只考慮兩個方向的計分，可從圖 4 中了解。

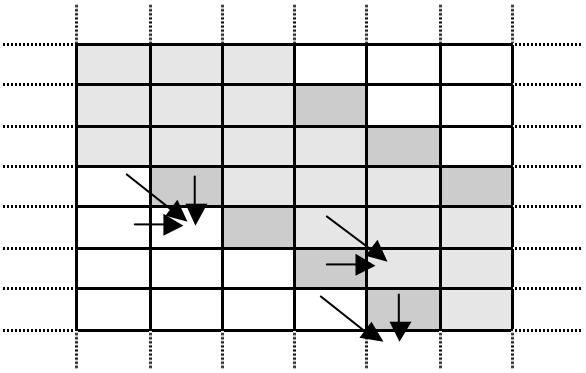

圖 4. 帶狀動態規劃計分圖

此步驟經分析，其時間複雜度為 O(RM)，M 值表示 Gapped BLAST 所得到的序列片段之長度。

三、外部改進方法

本篇論文除了針對 BLAST 所找出序列片段做內部帶狀改進之外, 同時也對此結果片段之外部作延伸改進。我們再度應用到帶狀動態規劃的理念，以 BLAST 所找出之片段，做前後對角線帶狀計分，同時給予兩參數值，K 值與 L 值，K 值代表帶狀動態規劃所取之寬度, L 值則表示在計分過程中，分數未增加次數之值。為求能同時擁有較佳的靈敏度及時間複雜度, 我們選擇考慮間隔的帶狀動態規劃來做延伸，而不是單針對對角線做延伸，並且以參數 L 值控制其延伸的幅度，如圖 5 所示，

圖中細虛線表是 BLAST 所找出之片段，前後延伸可以控制 K 及 L 兩參數值來調整其時間及靈敏度。

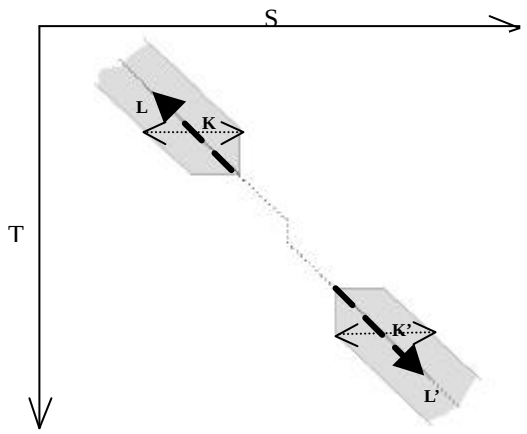

圖 5. 外部延伸示意圖

做前後對角線帶狀帶狀動態規劃延伸計分過程中，記錄每一列比對之最佳分數，並和最高分作比較。若每一列所記錄的最高分數，連續 L 次都無法取代最佳分數，便停止延伸，表示已經 L 列分數無改善。

四、實驗數據

表 1 是 1997 年由 Altschul，Madden，Jinghui Zhang，Miller，Lipman[2]等人提出 Gapped BLAST 論文中，針對一些序列在 SWISS-PROT 資料庫中，設定 E-value 值為 0.01，利用三種方法：Smith-Waterman，Original BLAST，及 Gapped BLAST 所找到的相似序列個數。其中可看出 Gapped BLAST 明顯改善 Original BLAST 結果，但是序列 P01111、P10318、及 P14942，利用 Gapped BLAST 和 Smith-Waterman 搜尋的結果差距仍大。於是我們針對這三個序列再做深入研究，找出 20 個 Gapped BLAST 未能找到而 Smith-Waterman 找到的序列做為實驗分析的對象。

表 2 是內部改進方法之實驗結果。從實驗結果的表格中首先可看出這 20 個序列，利用 Gapped BLAST 比對結果的分數，和 Smith-Waterman 最佳解的結果差距很大，我們的目的就是要改進比對的分數，縮小和最佳解的差距。接著觀察內部改進方法比對的分數和 Gapped BLAST 差距不大，顯示內部改善並無太大的效果。

表 3 是外部改進方法之實驗結果。20 個序列比對結果中，計分矩陣選用 BLUSOM62，間隔處罰函數採用仿射性間格處罰(affine gap penalty)函數，間格扣分參數 h 值給予 10，空白扣分參數 g 值給予 1，帶狀動態規劃寬度值 K 給予 5，分數未增加次數值 L 給予 40。實驗結果得到在 20 筆資料中，17 筆分數有改進，顯示分數改善情況顯著。

假設我們採取改進比例及正確率公式如下。

$$\text{改進百分比} = \frac{\text{延伸改進分數}}{\text{Gapped BLAST分數值}} * 100\%$$

$$\text{正確率} = \frac{\text{改進後總分}}{\text{Smith-Waterman分數}} * 100\%$$

由此我們可計算出實驗結果數據之平均改進分數百分比可達 18%，正確率更可高達 97.65%。由以上實驗所得數據，可以証實外部改進方法，確實是一個簡單快速且有效率提昇序列比對分數的演算法。

五、結論

BLAST 已是現今，應用最為廣泛序列比對工具。在 BLAST 比對後，再進行本篇論文中，我們提出的既簡單又快速的外部改進方法，可以有效率地改善及提昇，BLAST 序列比對分數和 Smith-Waterman 最佳解差距過大的情況。

Protein family	Query	Smith-Waterman	Original BLAST	Gapped BLAST
Serine protease	P00762	275	273	275
Ras	P01111	429	419	421
Globin	P02232	28	26	28
Interferon a	P05013	53	53	53
Alcohol dehydrogenase	P07327	205	205	205
Histocompatibility antigen	P10318	119	88	112
Cytochrome P450	P10635	211	197	211
Glutathione transferase	P14942	122	102	109
H+-transporting ATP synthase	P20705	198	191	198

表 1. 搜尋 SWISS-PROT 資料庫結果之對照表

Sequence 1	Sequence 2	Gapped BLAST	Smith-Waterman	內部改進方法
P01111	P32559	85	118	85
	P40617	80	99	80
	P35295	85	109	85
	P52198	101	127	101
	P49703	81	106	81
	P38987	87	107	87
	Q9NX57	82	106	82
	Q92737	101	133	101
	P56559	70	88	70
P10318	O00214	102	110	102
	O46631	100	105	100
P14942	O04437	70	85	70
	Q93112	84	103	84
	Q9WVL0	105	114	105
	Q9VG93	91	104	91
	P28342	94	102	94
	O43708	106	120	106
	P42860	62	76	62
	P28338	66	80	66
	P20432	69	85	69

表 2. **內部改進方法之實驗結果**

Sequence 1	Sequence 2	Gapped BLAST	延伸改進分數	改進後	總	Smith-Waterman
P01111	P32559	85	33	118	118	
	P40617	80	19	99	99	
	P35295	85	24	109	109	
	P52198	101	26	127	127	
	P49703	81	22	103	106	
	P38987	87	20	107	107	
	Q9NX57	82	24	106	106	
	Q92737	101	32	133	133	
	P56559	70	18	88	88	
P10318	O00214	102	0	102	110	
	O46631	100	0	100	105	
P14942	O04437	70	6	76	85	
	Q93112	84	17	101	103	
	Q9WVL0	105	0	105	114	
	Q9VG93	91	13	104	104	
	P28342	94	5	99	102	
	O43708	106	10	116	120	
	P42860	62	11	73	76	
	P28338	66	11	77	80	
	P20432	69	16	85	85	

表 3. **外部改進方法之實驗結果**

六、參考文獻

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic Local Alignment search Tool", J. Mol. Biol., 215, pp.403-410, 1990.

[2] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res., 25, pp.3389-3402, 1997.

[3] D. J. Lipman and W. R. Pearson, "Rapid and sensitive protein similarity search", Science, 227, pp.1435-1441, 1985.

[4] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequences of two proteins", J. Mol. Biol., 147, pp.195-197, 1970.

[5] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences", J. Mol. Biol., 147, pp.195-197, 1981.