

基因重組之字首反轉排序最佳解上限值問題

吳哲賢、張正佳

中華大學生物資訊學系

{jswu,m09720004}@chu.edu.tw

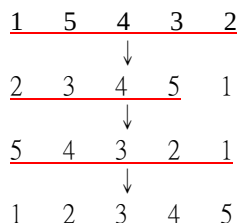
摘要

基因重組之字首反轉排序問題是生物資訊學近來被廣泛研究的主題，又稱為煎餅反轉問題，目的為利用字首反轉方式，找出最少反轉次數。此問題最佳解目前仍為 Open Problem，本論文主要討論最佳解上限值問題。假設煎餅數目為 n ，Gates 及 Papadimitriou 首先於 1979 年提出上限值為 $(5/3)n \doteq 1.6667n$ 。最近 Chitturi 等學者於 2009 年設計採計分數的趨勢函數，推導出 $(18/11)n \doteq 1.6363n$ 上限值，但是極為複雜及不易懂。本論文針對此問題，利用線性規劃方法，簡易且有效率地同樣推導出 $(18/11)n \doteq 1.6363n$ 上限值。

1 簡介

生物演化為遺傳基因序列發生突變，其突變有多種型態。其中基因重組 (Genome Rearrangements) 為兩個不同物種擁有相同基因，但卻有著不同的基因排列順序。利用反轉排序 (Sorting Reversals) 方式，找出從一條基因組反轉至另一條基因組所需要的最少反轉次數。

基因重組之字首反轉 (Prefix Reversals) 排序問題是生物資訊學近來被廣泛研究的主題，又稱為煎餅反轉問題 (Pancake Flipping Problem)，目的為利用字首反轉方式，找出任一序列反轉成完全序列最少反轉次數。此問題最佳解目前仍為 Open Problem，本論文主要討論最佳解上限值問題。圖一為原始序列 15432，利用三次字首反轉方式，轉成完全序列 12345。



圖一：序列 15432 轉成完全序列 12345。

假設煎餅數目為 n ，Gates 及 Papadimitriou 首先於 1979 年[1]，將所有字首反轉情形分成九大類，提出上限值為 $(5/3)n \doteq 1.6667n$ 。最近

Chitturi 等學者於 2009 年[2]，將所有字首反轉情形擴大分成 2220 大類，設計採計分數的趨勢函數 (Potential Function)，推導出 $(18/11)n \doteq 1.6363n$ 上限值，但是極為複雜及不易懂。本論文針對此問題，利用簡易線性規劃 (Linear Programming) 方法，有效率地同樣推導出 $(18/11)n \doteq 1.6363n$ 上限值。

2 線性規劃方法

基因重組之字首反轉排序問題中，獨立字元 (Singleton) 為序列中長度為一不和前後字元相連的獨立字元，以符號 s 表示序列中獨立字元數目；相連區塊 (Block) 為序列中長度超過一的相連字元，以符號 b 表示序列中相連區塊數目；反轉 (Flip) 次數以 f 表示。

考慮所有字首反轉共 2220 大類[3]，圖二為 (X1) 及 (X2) 大類中反轉情況。在 (X1) 大類中， A 及 B 原為獨立字元，字首反轉 f 一次後，相連區塊 b 增加 1，獨立字元 s 減少 2。在 (X2) 大類中， A 為獨立字元， B 為相連區塊字首，字首反轉 f 一次後，相連區塊 b 不變，獨立字元 s 減少 1。

(X1)
 $B_|A_$
 $_B\sim A_$
 $b: 1\ s: -2\ f: 1$

(X2)
 $A_|B\sim_$
 $_A\sim_$
 $b: 0\ s: -1\ f: 1$

圖二：字首反轉 (X1) 及 (X2) 反轉情況。

首先設計線性規劃的目標函數 (Objective Function)，針對所有 2220 大類字首反轉情形，分別給予變數並乘上其反轉次數 f ，累加後取最大值形成目標函數。式(1)為目標函數：

Maximize

obj: $1X1 + 1X2 + 6X31aa + \dots + 6X31cjip + \dots + 6X91cacb + \dots + 6X99h + 6X99i \dots \dots \dots (1)$

接著設計線性規劃限制式 (Constraints)，共

有三條。式(2)為考慮相連區塊變化，起始為 b 最後為 1；式(3)為考慮獨立字元變化，起始為 s 最後為 0；式(4)為考慮相連區塊 b 及獨立字元 s 關係，得到 $s + 2b \leq n$ 。

Subject to

$$cl: b + 1X1 + 0X2 + 1X31aa + \dots + X31cjip + \dots - 2X91cacb + \dots - 2X99h - 2X99i = 1 \text{ -----(2)}$$

$$c2: s - 2X1 - 1X2 - 5X31aa + \dots - 5X31cjip + \dots - X91cacb + \dots - 1X99h - 1X99i = 0 \text{ -----(3)}$$

$$c3: s + 2b \leq n \text{ -----(4)}$$

表一及表二分別為式(2)及式(3)的完整表示式。最後執行結果為式(5)至式(9)：

$$X31cjip = (2n + 1) / 11 \text{ -----(5)}$$

$$X91cacb = (n - 5) / 11 \text{ -----(6)}$$

$$s = n \text{ -----(7)}$$

$$\text{Other variables are all zero} \text{ -----(8)}$$

$$obj = (18n - 24) / 11 \text{ -----(9)}$$

根據結果可知利用字首反轉方式，最少反轉次數上限值為 $(18n - 24) / 11 \doteq (18/11)n \doteq 1.6363n$ 。其中起始狀態相連區塊 b 為 0，獨立字元 s 為 n 。字首反轉 2220 大類中， $X31cjip$ 執行次數為 $(2n + 1) / 11$ 以及 $X91cacb$ 執行次數為 $(n - 5) / 11$ ，其餘皆為 0。在圖三中為 $X31cjip$ 以及 $X91cacb$ 反轉情況。

(X31cjip)

B ~ A | I ~ C _ J _ T ~ _ R _ S H _
 → A ~ _ | B I ~ C _ J _ T ~ _ R _ S H _
 → _ ~ B I ~ C _ J _ T ~ _ R _ S H _
 → C ~ _ I | B ~ _ _ J _ T ~ _ R _ S H _
 → I _ ~ _ _ J _ T ~ _ R _ S | H _
 → S _ | R _ ~ T _ J _ _ ~ _ I ~ H _
 → _ S ~ R _ ~ T _ J _ _ ~ _ I ~ H _
 b: 1 s: -5 f: 6

(X91cacb)

B ~ C _ | R ~ _ D I _ ~ A _ S H _ ~ T _
 → _ C ~ B R ~ _ D I _ ~ A | _ S H _ ~ T _
 → A ~ _ I D _ ~ R | B ~ C _ _ S H _ ~ T _
 → R ~ _ D I _ ~ C _ _ | S H _ ~ T _
 → _ _ C ~ _ I D _ ~ S H _ ~ T | _
 → T ~ _ H | S ~ _ D I _ ~ C _ _ _
 → H _ ~ _ D I _ ~ C _ _ _
 b: -2 s: -1 f: 6

圖三：字首反轉(X31cjip)及(X91cacb)反轉情況。

3 實驗結果

在 Gates 和 Papadimitriou 演算法中字首反轉總共分為九大類，首先目標函數為式(10)：

Maximize

$$obj: X1 + X2 + 4X3 + X4 + 2X5 + X7 \text{ ----- (10)}$$

接著限制式共有四條：式(11)為變數關係 $X3 = X31 + X32 + X33 + X34$ ；式(12)為考慮相連間隔(Adjacencies)變化，相連間隔為相連區塊中所有相連間隔數目，以 a 表示，起始為 a 最後為 $n - 1$ ；式(13)為考慮相連區塊變化，起始為 b 最後為 1；式(14)為考慮相連區塊 b 及相連間隔 a 關係，得到 $b \leq a$ 。

Subject to

$$cl: X3 = X31 + X32 + X33 + X34 \text{ -----(11)}$$

$$c2: a + X1 + X2 + 2X31 + 3X32 + 3X33 + 3X34 + X4 + X5 + X7 = n - 1 \text{ -----(12)}$$

$$c3: b + X1 - X31 - X33 - 2X34 - X5 - X7 = 1 \text{ --(13)}$$

$$c4: b \leq a \text{ -----(14)}$$

在式(12)中假設煎餅數目 n 為 100，最後執行結果為 164.33，符合 $(5n - 7) / 3 \doteq (5/3)n$ 上限值。接著針對上一節所設計的線性規劃方法，在式(4)中假設煎餅數目 n 為 100，最後執行結果為 161.45，同樣符合符合 $(18n - 24) / 11 \doteq (18/11)n$ 上限值。

4 結論

假設煎餅數目為 n ，Gates 及 Papadimitriou 首先提出上限值為 $(5/3)n$ 。接著 Chitturi 等學者設計採計分數的趨勢函數，推導出 $(18/11)n$ 上限值，但是極為複雜及不易懂。本論文針對此問題，利用線性規劃方法，簡易且有效率地同樣推導出 $(18/11)n$ 上限值。

參考文獻

- [1] W. Gates and C. Papadimitriou, "Bounds for sorting by prefix reversal", Discrete Math., vol. 27, pp.47-57, 1979.
- [2] B. Chitturi, W. Fahle, Z. Meng, L. Morales, C. Shields, I. Sudborough, and W. Voit, "An $(18/11)n$ upper bound for sorting by prefix reversals", Theoretical Computer Science, vol. 410, pp. 3372-3390, 2009.
- [3] <http://www.utdallas.edu/~hal/PancakeSolution/>.

表一：式(2)的完整表示式。(續)

[illegible]

表二：式(3)的完整表示式。

[illegible]

表二：式(3)的完整表示式。(續)

[illegible]