

基因重組字首反轉排序問題：利用整數規劃分析

吳哲賢，侯玉松，蕭劭彥

中華大學生物資訊學系

jswu@chu.edu.tw, yshou@chu.edu.tw, m10420002@chu.edu.tw

摘要

基因重組字首反轉排序問題，又稱為煎餅反轉問題，目的在找出由一組原始基因反轉排序至完全基因序列所需最少反轉次數，此問題最佳解目前仍為 Open Problem。假設序列長度為 n ，Gates 及 Papadimitriou 利用線性規劃提出反轉次數上限值為 $(5n+5)/3$ ，但反轉次數應該為整數而非實數，因此線性規劃分析方法是錯誤的。本論文針對此問題，利用整數規劃方法，重新推導並更新基因重組字首反轉排序次數，當 n 為 $3k$ ， $3k+1$ 以及 $3k+2$ ， $k \in \mathbb{Z}^+$ ，整數上限值分別為 $5k+1$ ， $5k+3$ 以及 $5k+5$ 。

1 簡介

基因重組(Genome Rearrangements)問題是兩個不同物種擁有相同基因，但不同的基因排列順序，可以互相演化。本論文研究利用字首反轉排序(Sorting by Prefix Reversals)方法，找出由一組原始基因反轉排序至完全基因序列所需最少反轉次數，又稱煎餅反轉問題(Pancake Flipping Problem)[1-5]。圖一為原始序列 52134，利用三次字首反轉方式，轉成完全序列 12345。

5	2	1	3	4
4	3	1	2	5
2	1	3	4	5
1	2	3	4	5

圖一：序列 52134 字首反轉成序列 12345

此問題之最少反轉次數最佳解，到目前仍然是 Open Problem。假設序列長度為 n ，Gates 及 Papadimitriou[1]於 1979 年，利用線性規劃(Linear Programming)方法，推導出反轉次數上限值為 $(5n+5)/3$ 。然而反轉次數為整數並非實數，之前線性規劃分析方法是錯誤的。因此本論文針對此問題，利用整數規劃(Integer Programming)方法，推導出更新的正確基因重組字首反轉排序整數上限值。

2 線性規劃方法

Gates 和 Papadimitriou[1]演算法中，線性規劃列式如下： n 為原始序列長度， x_1 、 x_2 、 x_3 、 x_4 、 x_5 及 x_7 分別代表不同類型的字首反轉， b

為原始序列區塊(Blocks)數目， a 為原始序列相鄰字元(Adjacencies)數目。其中字首反轉 x_3 ，又可以細分成四種不同類型的字首反轉： x_{31} 、 x_{32} 、 x_{33} 及 x_{34} 。

$$\text{Max } z = x_1 + x_2 + 4x_3 + x_4 + 2x_5 + x_7$$

s.t.

$$b + x_1 - x_{31} - x_{33} - 2x_{34} - x_5 - x_7 = 1,$$

$$n - 1 = a + x_1 + x_2 + 2x_{31} + 3x_{32} + 3x_{33} + 3x_{34} + x_4 + x_5 + x_7,$$

$$x_3 = x_{31} + x_{32} + x_{33} + x_{34},$$

$$b \leq a.$$

把 $b \leq a$ 代入第二個限制式，列式修正為：

$$\text{Max } z = x_1 + x_2 + 4x_3 + x_4 + 2x_5 + x_7$$

s.t.

$$b + x_1 - x_{31} - x_{33} - 2x_{34} - x_5 - x_7 = 1,$$

$$x_1 + x_2 + 2x_{31} + 3x_{32} + 3x_{33} + 3x_{34} + x_4 + x_5 + x_7 + b \leq n - 1,$$

$$x_3 = x_{31} + x_{32} + x_{33} + x_{34}.$$

利用對偶方法推導出結果為：

$$z = (5n - 7)/3$$

$$x_1 = (n + 1)/3$$

$$x_3 = x_{31} = (n - 2)/3$$

$$x_2 = x_4 = x_5 = x_7 = b = 0$$

最後 z 值要再加上反轉成完全序列最多需額外 4 次字首反轉，所以基因重組字首反轉排序上限值為 $(5n-7)/3 + 4 = (5n+5)/3$ 。

3 整數規劃方法

上述線性規劃計算結果發現， z 值及各 x_i 變數值定義為反轉次數，但是結果並非整數。現在我們把 x_3 用 $x_{31} + x_{32} + x_{33} + x_{34}$ 代入並利用線性規劃 Big M 方法加入變數 A 及 c ， M 代表極大值，重新列式如下：

$$\text{Max } z = x_1 + x_2 + 4x_{31} + 4x_{32} + 4x_{33} + 4x_{34}$$

$$+ x_4 + 2x_5 + x_7 - M \cdot A$$

s.t.

$$b + x_1 - x_{31} - x_{33} - 2x_{34} - x_5 - x_7 + A = 1,$$

$$x_1 + x_2 + 2x_{31} + 3x_{32} + 3x_{33} + 3x_{34}$$

$$+ x_4 + x_5 + x_7 + b + c = n - 1.$$

計算結果如下，執行過程在表一：

$$\begin{aligned}
z &= (5n - 7)/3 \\
x_1 &= (n + 1)/3 \\
x_{31} &= (n - 2)/3 \\
x_2 = x_{32} = x_{33} = x_{34} = x_4 = x_5 = x_7 &= A = b = c = 0
\end{aligned}$$

計算結果發現， z 值及各 x_i 變數值並非整數。考慮 n 的三種狀況： $3k$, $3k+1$, 以及 $3k+2$, $k \in \mathbb{Z}^+$ ，利用分枝界定法(Branch and Bound)及線性規劃方法來推導出整數的 z 值及各 x_i 變數。
[Case I: $n = 3k$]

令 $n = 3k$ ，利用線性規劃計算結果如下：

$$\begin{aligned}
z &= 5k - (7/3) \\
x_1 &= k + (1/3) \\
x_{31} &= k - (2/3) \\
x_2 = x_{32} = x_{33} = x_{34} = x_4 = x_5 = x_7 &= A = b = c = 0
\end{aligned}$$

結果中可發現， z 值及各 x_i 變數值並非為整數。接著如圖二利用分枝界定法，最後推導出 z 值及各 x_i 變數值的整數解有兩組如下。

$$\begin{aligned}
\text{(i)} \quad z &= 5k - 3 \\
x_1 &= k + 1 \\
x_{31} &= k - 2 \\
x_5 &= 2 \\
x_2 = x_{32} = x_{33} = x_{34} = x_4 = x_7 &= A = b = c = 0
\end{aligned}$$

$$\begin{aligned}
\text{(ii)} \quad z &= 5k - 3 \\
x_1 &= k \\
x_{31} &= k - 1 \\
x_4 &= 1 \\
x_2 = x_{32} = x_{33} = x_{34} = x_5 = x_7 &= A = b = c = 0
\end{aligned}$$

比較特別的結果是(i) $x_5 = 2$ ，以及(ii) $x_4 = 1$ ，多了 Gates 和 Papadimitriou 原先分析結果所沒有的字首反轉類型。因為 x_2 和 x_4 在列式中條件一致，因此得到額外的第三組解：

$$\begin{aligned}
\text{(iii)} \quad z &= 5k - 3 \\
x_1 &= k \\
x_{31} &= k - 1 \\
x_2 &= 1 \\
x_{32} = x_{33} = x_{34} = x_4 = x_5 = x_7 &= A = b = c = 0
\end{aligned}$$

[Case II: $n = 3k+1$]

令 $n = 3k+1$ ，利用線性規劃計算結果如下：

$$\begin{aligned}
z &= 5k - (2/3) \\
x_1 &= k + (2/3) \\
x_{31} &= k - (1/3) \\
x_2 = x_{32} = x_{33} = x_{34} = x_4 = x_5 = x_7 &= A = b = c = 0
\end{aligned}$$

結果中可發現， z 值及各 x_i 變數值並非為整數。利用圖三的分枝界定法，最後推導出 z 值及各 x_i 變數值的整數解如下。

$$\begin{aligned}
z &= 5k - 1 \\
x_1 &= k + 1 \\
x_{31} &= k - 1 \\
x_5 &= 1 \\
x_2 = x_{32} = x_{33} = x_{34} = x_4 = x_7 &= A = b = c = 0
\end{aligned}$$

發現比較特別的是 $x_5 = 1$ ， x_5 類型的字首反轉出現反轉一次。

[Case III: $n = 3k+2$]

令 $n = 3k+2$ ，利用線性規劃計算結果如下，可發現 z 值及各 x_i 變數值皆為整數，不需要再利用分枝界定法。

$$\begin{aligned}
z &= [5(3k+2) - 7]/3 = 5k + 1 \\
x_1 &= [(3k+2) + 1]/3 = k + 1 \\
x_{31} &= [(3k+2) - 2]/3 = k \\
x_2 = x_{32} = x_{33} = x_{34} = x_4 = x_5 = x_7 &= A = b = c = 0
\end{aligned}$$

上述三種結果，再加上轉成完全序列最多需額外 4 次字首反轉，最後整理出表二的基因重組字首反轉排序上限值的整數解：(i) $n=3k$ ，上限值為 $(5k-3) + 4 = 5k+1$ ；(ii) $n=3k+1$ ，上限值為 $(5k-1) + 4 = 5k+3$ ；(iii) $n=3k+2$ ，上限值為 $(5k+1) + 4 = 5k+5$ ， $k \in \mathbb{Z}^+$ 。

4 結論

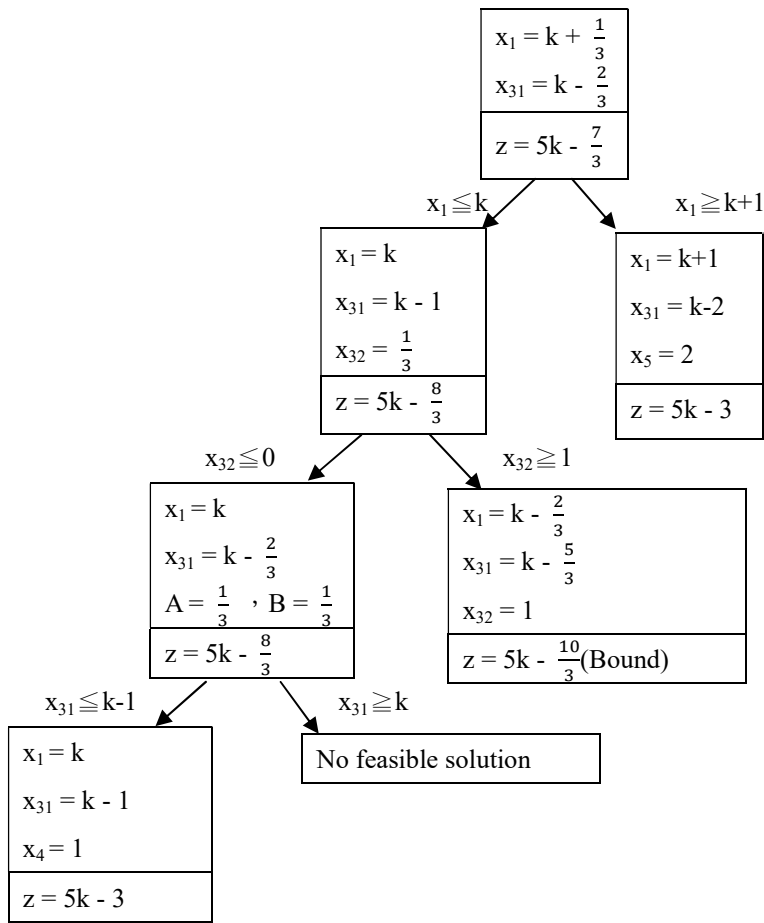
Gates 及 Papadimitriou 利用線性規劃提出上限值為 $(5n+5)/3$ ，但反轉次數為整數並非實數，因此線性規劃分析方法是錯誤。本論文利用整數規劃方法，重新推導並更新正確的基因重組字首反轉排序整數上限值。

參考文獻

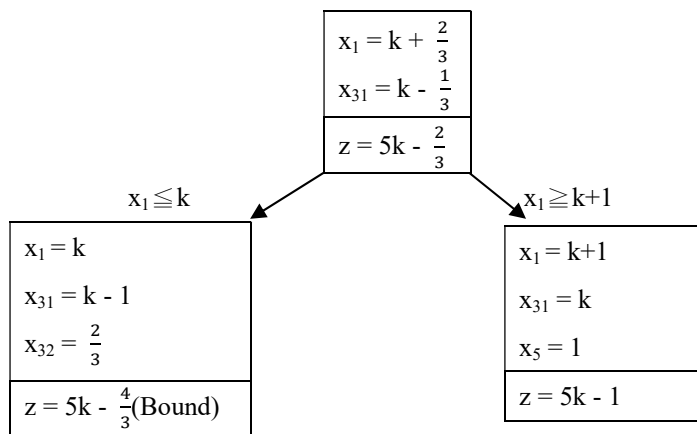
- [1] W. Gates and C. Papadimitriou, "Bounds for sorting by prefix reversal", Discrete Math., vol. 27, pp.47-57, 1979.
- [2] B. Chitturi, W. Fahle, Z. Meng, L. Morales, C. Shields, I. Sudborough, and W. Voit, "An $(18/11)n$ upper bound for sorting by prefix reversals", Theoretical Computer Science, vol. 410, pp. 3372-3390, 2009.
- [3] 吳哲賢、張正佳, "基因重組之字首反轉排序最佳解上限值問題", The 27th Workshop on Combinatorial Mathematics and Computational Theory, pp 68-74, 2010.
- [4] 吳哲賢, "基因重組之燃燒煎餅圖上單壞點容錯繞線分析", 2012 國際台灣生物暨醫學資訊聯合研討會, pp 256-261, 2012.
- [5] Jer-Shyan Wu, Yu-Song Hou and Wei-Qian Cai, "Genome rearrangements: efficient algorithm in pancake graphs with one faulty node", The 32th Workshop on Combinatorial Mathematics and Computational Theory, pp 130-134, 2015.

表一：利用線性規劃 Big M 方法計算 Gates 和 Papadimitriou 演算法結果

c_B	$\begin{matrix} c_j \\ x_B \end{matrix}$	1	1	4	4	4	4	1	2	1	0	0	-M	b_i	$\frac{b_i}{a_{ij}}$
		x_1	x_2	x_{31}	x_{32}	x_{33}	x_{34}	x_4	x_5	x_7	b	c	A		
-M	A	1	0	-1	0	-1	-2	0	-1	-1	1	0	1	1	$\nearrow$
0	c	1	1	2	3	3	3	1	1	1	1	1	0	n-1	
	z_j	-M	0	M	0	M	2M	0	M	M	-M	0	-M	-M	$\nless 0$
	c_j-z_j	1+M	1	4-M	4	4-M	4-2M	1	2-M	1-M	M	0	0	0	
1	x_1	1	0	-1	0	-1	-2	0	-1	-1	1	0	1	1	$\nearrow$
0	c	0	1	3	3	4	5	1	2	2	0	1	-1	n-2	
	z_j	1	0	-1	0	-1	-2	0	-1	-1	1	0	1	1	$\nless 0$
	c_j-z_j	0	1	5	4	5	6	1	3	2	-1	0	-M-1	0	
1	x_1	1	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{6}{5}$	$\frac{3}{5}$	0	$\frac{2}{5}$	$-\frac{1}{5}$	$-\frac{1}{5}$	1	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{2n+1}{5}$	$\nearrow$
4	x_{34}	0	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	1	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	0	$\frac{1}{5}$	$-\frac{1}{5}$	$\frac{n-2}{5}$	
	z_j	1	$\frac{6}{5}$	$\frac{13}{5}$	$\frac{18}{5}$	$\frac{19}{5}$	4	$\frac{6}{5}$	$\frac{7}{5}$	$\frac{7}{5}$	1	$\frac{6}{5}$	$-\frac{1}{5}$	$\frac{6n-7}{5}$	$\nless 0$
	c_j-z_j	0	$-\frac{1}{5}$	$\frac{7}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	0	$-\frac{1}{5}$	$\frac{3}{5}$	$-\frac{2}{5}$	-1	$-\frac{6}{5}$	$-M+\frac{1}{5}$	0	
1	x_1	1	$\frac{1}{3}$	0	1	$\frac{1}{3}$	$-\frac{1}{3}$	$\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$	1	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{n+1}{3}$	
4	x_{31}	0	$\frac{1}{3}$	1	1	$\frac{4}{3}$	$\frac{5}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	0	$\frac{1}{3}$	$-\frac{1}{3}$	$\frac{n-2}{3}$	
	z_j	1	$\frac{5}{3}$	4	5	$\frac{17}{3}$	$\frac{19}{3}$	$\frac{5}{3}$	$\frac{7}{3}$	$\frac{7}{3}$	1	$\frac{5}{3}$	$-\frac{2}{3}$	$\frac{5n-7}{3}$	
	c_j-z_j	0	$-\frac{2}{3}$	0	-1	$-\frac{5}{3}$	$-\frac{7}{3}$	$-\frac{2}{3}$	$-\frac{1}{3}$	$-\frac{4}{3}$	-1	$-\frac{5}{3}$	$-M+\frac{2}{3}$	≤ 0	



圖二： $n=3k$ ，利用分枝界定法及線性規劃方法推導出整數的 z 值及各 x_i 變數



圖三： $n=3k+1$ ，利用分枝界定法及線性規劃方法推導出整數的 z 值及各 x_i 變數

表二：基因重組字首反轉排序上限值之整數解

n	3k	3k+1	3k+2
Solutions	(i) $z = 5k - 3$ $x_1 = k + 1$ $x_{31} = k - 2$ $x_5 = 2$ (ii) $z = 5k - 3$ $x_1 = k$ $x_{31} = k - 1$ $x_4 = 1$ (iii) $z = 5k - 3$ $x_1 = k$ $x_{31} = k - 1$ $x_2 = 1$	$z = 5k - 1$ $x_1 = k + 1$ $x_{31} = k - 1$ $x_5 = 1$	$z = 5k + 1$ $x_1 = k + 1$ $x_{31} = k$
Flips	5k+1	5k+3	5k+5