基因複製向量轉換線性時間演算法的修正

楊敦翔*　楊宗頤*　謝維華*

摘　要

Schwarz et al. 訂出基因複製向量間的距離，以探討正常基因組與腫瘤基因組的差距，並提出計算距離的演算法 MEDICC [2]，但並未分析計算複雜度，且在某些情況下，會是指數型態。之後 Zeira et al. 提出線性時間的演算法[1]，但我們發現在推導過程中，有些地方並不正確，確認 Zeira et al. 演算法的結論有誤。另外我們提出新的演算法，對原方法做了若干修正。

關鍵字：癌，基因組重新排列，基因複製向量。

一、背　景

基因組重新排列(genome rearrangement)是物種演化和癌症研究的核心問題，而兩者最大的差異在於歷經的時間長短。物種演化動輒百萬年，而癌症僅有幾十年，相較之下我們能更完整收集癌症基因演變的資料。過去在基因組重新排列方面的工作，大部分在探討物種演化。2006 年開始的美國癌症基因體圖譜計畫 TCGA(The Cancer Genome Atlas)，大規模地蒐集、分類了數萬名病人的癌症基因突變數據。最重要的是，計畫的資料和大部份分析結果皆公開於網路上，可供瀏覽及下載。因此目前有很好的機會，利用這些數據，去分析、計算癌症基因演化的過程。

*東海大學應用數學系

在腫瘤基因組中，DNA 整段的刪減(deletions)和複製(amplifications)，是常見的突變（TCGA Research Network, 2011）[3]。正常基因組每種基因有一對(copies or alleles)，可用一個全為 2 的向量 (2,2,2, ... ,2,2) 代表，稱為基因複製向量(copy number profile)以下簡稱 CNP。因 DNA 整段的刪減和複製，基因個數會改變，CNP 就不會全為 2，譬如 (3,2,5, ... ,1,3)。腫瘤基因組有許多不同的 CNP，瞭解其演化過程，有助於預測疾病的進展和可能的醫療介入。

已有許多方法可以讀出癌症基因組各種不同的 CNP。G 條紋染色法(G-banding)、螢光原位雜合技術(Fluorescence in situ hybridization, FISH)、基於微陣列的比較基因組雜交(array CGH)、深度定序(deep sequencing)。因此獲得癌症基因組的 CNP 並不困難，但要利用它來瞭解癌症的演化過程，仍然是個未解的問題。

科學家提出距離的概念，以比較正常基因組和腫瘤基因組 CNP 的差異。有幾個測定距離的方式：最常見的方法是兩者間的 Euclidean 距離（Schwarz et al., 2014）[2]。Chowdhury et al.訂定 FISH CNP 之間的編輯距離(edit distance)，編輯指的是對單個基因、單個染色體或整個基因組的刪減或複製（Chowdhury et al., 2013, 2014, 2015）[4][5][6]，但計算這個距離所花的時間，隨基因個數，呈指數型態成長。TuMult 演算法則以不同數字的個數，訂為兩個 CNP 間的距離（Letouzé et al., 2010）[7]。

Schwartz et al.的距離模型，則允許整段的刪減或複製（Schwarz et al., 2014）[2]。譬如 CNP (2,2,2, ... ,2,2) 的第一至第三個數字，整段刪減 1，變成(1,1,1, ... ,2,2)。他們提出計算該距離的演算法 MEDICC [2]，但並未分析計算複雜度，且在某些情況下，會是指數型態。之後 Zeira et al.提出動態規劃法計算該距離，並進一步將該法修改成線性時間的演算法[1]。

二、基因複製向量轉換問題

CNP 為一向量 $V = (v_1, v_2, \ldots, v_n)$，依序代表染色體上各種基因的個數。

我們定義突變事件 $c = (\ell, h, w)$ 其中 $1 \le \ell \le h \le n$，$w \in \{1, -1\}$，來表示 DNA 整段的刪減或複製。其中，$(\ell, h, 1)$是指從 CNP 的位置 ℓ 開始到 h 結束，每個基因個數加 1（複製）；$(\ell, h, -1)$ 是指從 ℓ 開始到 h 結束，每個基因個數減 1（刪減）。譬如下例用了 3 個突變事件，將 CNP S 轉成 CNP T ：

$$S = (1,1,1,1,1)$$
$$\downarrow \quad c_1 = (4,4,-1)$$
$$c_1(S) = (1,1,1,0,1)$$
$$\downarrow \quad c_2 = (2,2,-1)$$
$$c_2(c_1(S)) = (1,0,1,0,1)$$
$$\downarrow \quad c_3 = (1,5,+1)$$
$$T = c_3(c_2(c_1(S))) = (2,0,2,0,2)$$

紅色(實線)是刪減
藍色(虛線)是複製

一系列的突變事件，我們稱作基因複製向量轉換（Copy Number Transformation，簡稱 CNT）。從 CNP S 轉成 CNP T ，有各種可能的 CNT，其中最少可能的突變事件個數訂為 S 與 T 的距離，記作 $dist(S, T)$。那組突變事件個數最少的 CNT（可能非唯一），我們稱為最佳的。如上例 $dist(S, T) = 3$。CNP (2,0,2,0,2)，第二與第四個數字為 0，代表該基因已被刪除，因此已無再刪減或複製的可能，該位置將永遠為 0。另外，不可能用一系列突變事件轉換時，如上例將 $T = (2,0,2,0,2)$ 轉成 $S = (1,1,1,1,1)$，記作 $dist(T, S) = \infty$。

計算兩個 CNP 的距離，即為 CNTP。

給定兩個 CNPs，$S = (s_1, s_2, \ldots, s_n)$，$T = (t_1, t_2, \ldots, t_n)$，其中最大的數記做 $B = max\{max_{i=1}^{n}\{s_i\}, max_{i=1}^{n}\{t_i\}\}$。定義 $u_i = t_i - s_i$。

給定 CNT $C = (c_1, c_2, \ldots, c_m)$，$1 \le j \le n$ ，$w \in \{1, -1\}$，則 CNT C 到第 j 個位置執行 w 的總次數記做 $op(C, w, j)$。

三、Zeira et al. 解 CNTP 的演算法

3.1 Zeira et al. 證明的四個性質[1]。

定義 1.

給定 CNT $C = (c_1, c_2, \ldots, c_m)$，如果刪減全部排在前面，而複製全部排在後面，我們稱 C 為 ordered。

$C = (c_1, c_2, c_3, c_4, c_5, c_6)$
deletions amplifications

性質 1.

存在一個最佳的 ordered CNT。

定義 2.

如右圖顯示，兩個緊鄰的突變事件（紅色線段），可以合成為一個突變事件（藍色線段），距離可少 1。此動作稱為 elongated。

性質 2.

任何最佳的 ordered CNT 都是 elongated。

定義 3.

給定 CNT C ，若 $1 \le i < j \le n$ 且對於所有的 $i < r \le j$ ， $t_r = 0$ ，則 $op(C, -1, j) = \max\{\max\{s_r\}, op(C, -1, i)\}$ 且 $op(C, 1, j) = op(C, 1, i)$。我們稱這種情況為 zero-skipping。

性質 3.

存在最佳的 CNT 是 zero-skipping 且 ordered。

S = (2,1,3,1,2)
T = (1,0,0,0,3)
$\max\{s_r\} = 3$

定義 4.

所有的 $op(C, w, j) \le B$ 。換句話說會有個上界 B 。我們稱這種情況為 bounded。

性質 4.

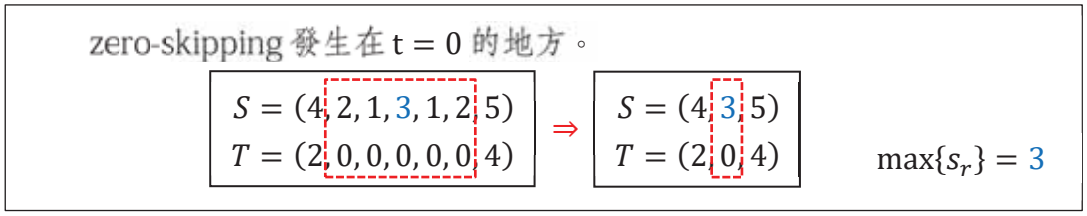
任何一個最佳 zero-skipping、ordered 的 CNT 必為 bounded。

3.2 Zeira et al. 解 CNTP 的動態規劃法 DpCntpAlg

這裡我們介紹 DpCntpAlg，並做部分修改，成為 NewDpCntpAlg。

計算兩個 CNP $S = (s_1, s_2, \ldots, s_n)$ 與 $T = (t_1, t_2, \ldots, t_n)$ 的距離，DpCntpAlg 計算出一個 n 行、$(B+1)$ 列的表 M，其中，$M[j,d]$ 是，在恰好使用 d 個刪減的前提下，$S^j = (s_1, s_2, \ldots, s_j)$ 與 $T^j = (t_1, t_2, \ldots, t_j)$ 的距離，d 的範圍介於 0 和 B 之間。M 表逐行計算，每個 $M[j,d]$ 是利用根據前面某行已經先算出的 $M[i,d]$ 而算出，其中 $i < j$，$0 \le d \le B$。若 CNT 不存在，則 $M[j,d] = \infty$。根據前述四個性質（ordered、elongated、zero-skipping、bounded），計算所需記憶空間，僅為表 M 的兩行[1]。

另外，在 T 連續為 0 的段落，因為**性質 3** 的緣故，僅需留下 Q_j（該段最大的s 值）。我把它和 DpCntpAlg 裡的 $max\{d - d', 0\}$ 合併，在輸入 S 與 T 時，就先將每段 zero-skipping 修正為 $(Q_j, 0)$ 輸入，這樣能簡化計算式子，不用額外計算 $max\{Q_j - max\{d, d'\}, 0\}$。



$M[j,d]$

j	1	2	3 ~ 9	10
S	5	3	3	6
T	9	6	0	8
d = 0	4	4	∞	9
d = 1	6	5	∞	10
d = 2	8	7	∞	11
d = 3	10	∞	7	12
d = 4	12	∞	8	13
...				
d = B	∞	∞	$B+4$	$M[10,B]$

∞ 是因為不可能達到。

例如：(s_2, t_2) 當 $d = 3$ 時，為滿足**命題 1** 而必須先減，但當 T 歸 0 就再也無法出現。因此當 $d = 3$ 時，無法達到 $t = 6$。

$j = 3$~9 時 $t = 0$，滿足**命題 3** 故省略其他不具影響的值，留下 Q_j 為代表。例如：$Q_3 = 3$。

d 最大值是 B，也就是 $max(S)$。

DpCntpAlg 及 NewDpCntpAlg 的符號說明：
$prev(j) < j$ 且 $prev(j)$ 是最靠近 j 的非省略項。
d 是第 j 行的刪減。
d'是第 $prev(j)$ 行的刪減。
$a(j,d)$ 是第 j 行 d 個刪減的複製。
$a(prev(j),d')$ 是第 $prev(j)$ 行 d'個刪減的複製。
$Q_j = max\{s_r\}$ ，其中是 s_r 第 j 行和 $prev(j)$ 的 zero-skipping。

DpCntpAlg：

$M[1,d] = d + a\ （1,d）$
$M[j,d] = \min_{0 \le d' \le B} \{M[prev(j),d'] + max\{d-d',0\} + max\{a(j,d) - a(prev(j),d'),0\}$
$\quad + max\{Q_j - max\{d,d'\},0\}\}$
i. e.
$M[j,d] = \min_{prev(j)的每一項} \{M[prev(j),d'] + (補刪減) + (補複製)$
$\quad + (補\ zero\ skipping\ 刪減)\}$

NewDpCntpAlg：

$M[1,d] = d + a\ （1,d）$
$M[j,d] = \min_{0 \le d' \le B} \{M[prev(j),d'] + max\{d-d',0\} + max\{a(j,d) - a(prev(j),d'),0\}\}$
i. e.
$M[j,d] = \min_{prev(j)的每一項} \{M[prev(j),d'] + (補刪減) + (補複製)\}$

空間複雜度為 $O(B)$

原因：對於 $M[j,d]$ 的計算，僅需要保持位置 $prev(j)$ 的那一行，因此，空間複雜度為 $O(B)$ 。

計算複雜度為 $O(nB^2)$

原因：表 M 有 n 行、$(B+1)$ 列，其中 B 為 S 的最大值。因此共有 $n(B+1)$個數字需要計算。每個數字都利用前一行來計算，計算 $(B+1)$ 次後得到。因每次計算皆為常數時間，故計算複雜度為 $O(nB^2)$。

3.3 Zeira et al. 解 CNTP 的線性時間演算法 LinearCntpAlg

LinearCntpAlg 將 DpCntpAlg 表 M 中的每一行，用一個 piecewise linear function 來表示，使計算複雜度減少至 $O(n)$。但我們發現原論文[1]的程式在執行上會出現 $M[i,d] > M[j,d']$ 的狀況，其結論很可能有誤。因此我們依其原來做法，地毯式地分析每種狀況，建立出新的演算法 GreedyCntpAlg。最終證明最佳值 $best_j$ 來自最佳值 $best_i$，其中 $i = prev(j)$，換句話說每行的最佳值都來自前一行的最佳值！

在開始詳細案例分析前，我們先利用 Excel 撰寫了程式[附錄]，並用隨機值(random value)生成隨機的 S 和 T ，執行 DpCntpAlg 和貪婪演算法(Greedy algorithm)來比對結果。我發現每次執行結果都相同，因此最佳值 $best_j$ 來自最佳值 $best_i$ 的可能性很大！

令 $1 \le r \le n$ 使得 $T_r \ge 0$。另設 $C = (c_1, c_2, \dots, c_m)$為 CNT，使得 $C(S) = T$。則 $u_r = op(C, 1, r) - op(C, -1, r)$。$u_r > 0$ 則表示基因數增加，反之表示減少。

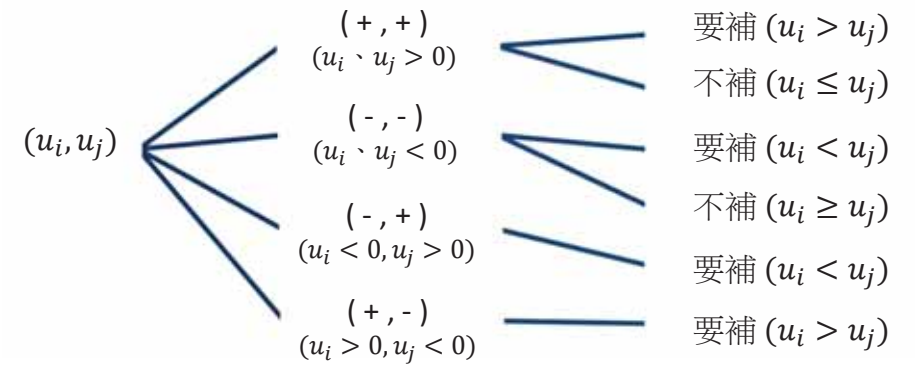
GreedyCntpAlg：

$$best_j = best_i + max\big(max(-u_j, 0) - max(-u_i, 0), 0\big) + max\big(max(u_j, 0) - max(u_i, 0), 0\big)$$

這個式子會比 DpCntpAlg 有更多的判斷，目的是為了區分輸入值$(\mathrm{s}_i, \mathrm{t}_i)$、$(\mathrm{s}_j, \mathrm{t}_j)$ 之間的關係。首要判斷 $u > 0$ 或 $u < 0$，也就是判斷該片段基因數是增加或減少（$u = 0$ 不列入考量），接著將增加的部分跟增加的部分相減，也就是判斷是否比前項增加更多；反之判斷是否比前項減少更多。

四、案例分析

光是執行數次程式做比較，不足以證明最佳值 $best_j$ 來自最佳值 $best_i$，因此我將每種可能的情況列舉出來：（判別 $u > 0$ 或 $u < 0$；$u_i > u_j$ 或 $u_i < u_j$）



設起始位置（第一個非 ∞ 的位置）為 F，終點位置（最後非 ∞ 的位置）為 E。

當 $u > 0$，則 $F = 0,\ E = s - 1$，總個數 $E - F + 1 = s$

當 $u < 0$，則 $F = -u,\ E = \begin{cases} s-1 & ,t \neq 0 \\ \infty & ,t = 0 \end{cases}$，總個數 $E - F + 1 = \begin{cases} t & ,t \neq 0 \\ \infty & ,t = 0 \end{cases}$

設 $\mathrm{k} = \mathrm{u}_j - \mathrm{u}_i$，用來表示後項需要而外補足的數量。

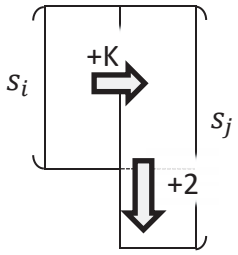
設 $D_r = M[r, d+1] - M[r, d]$，用來表示上下項的差。$|D_r|_\alpha$ 表示 $D_r = \alpha$ 的個數。（註：若 $F \leq r \leq E$ ，則 $\alpha = 1\ or\ 2$，且 $D_r = 1$ 的區塊在 $D_r = 2$ 的區塊之上）

（+ , +）都增加的時候：$(u_i \text{、} u_j > 0)$

分為要補複製 $(u_i < u_j)$、不補複製 $(u_i > u_j)$

I. $u_i < u_j$

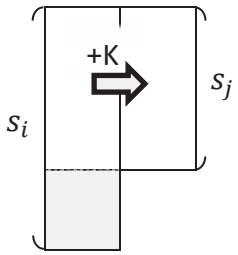
a. $E_i < E_j$



$$\begin{cases} M[i,d]+k=M[j,d] & ,d \le s_i \\ M[j,d]+2=M[j,d+1] & ,s_i < d \le s_j \\ M[j,d]=\infty & ,s_j < d \end{cases}$$

$\Rightarrow best_j = best_i + k$

b. $E_i \geq E_j$

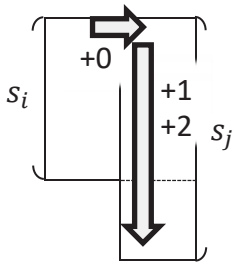


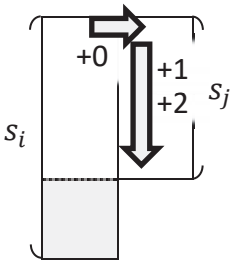
$$\begin{cases} M[i,d]+k=M[j,d] & ,d \le s_j \\ M[j,d]=\infty & ,s_j < d \end{cases}$$

$\Rightarrow best_j = best_i + k$

II. $u_i \geq u_j$

a. $E_i < E_j$ ， $E_i \geq E_j$





$$\begin{cases} M[i,1]=M[j,1] \\ |D_i|_1 - k = |D_j|_1 \\ M[j,d]=\infty & ,s_j < d \end{cases}$$

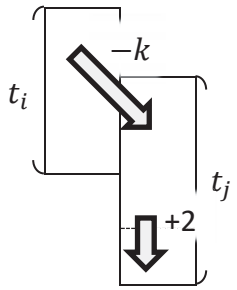
$\Rightarrow best_j = best_i$

（-，-）都減少的時候：$(u_i \text{、} u_j < 0)$

分為要補刪減 $(u_i > u_j)$、不補刪減 $(u_i < u_j)$

I. $\mathbf{u_i > u_j}$

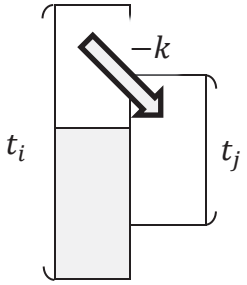
a. $E_i < E_j - k$



$$\begin{cases} M[j,d] = \infty & ,d < -u_j \\ M[i,d] - k = M[j,d-\mathrm{k}] & ,-u_j \le d \le s_i - k \\ M[j,d] + 2 = M[j,d+1] & ,s_i - k < d \le t_j \\ M[j,d] = \infty & ,t_j < d \end{cases}$$

$\Rightarrow best_j = best_i - k$

b. $E_i \ge E_j - k$

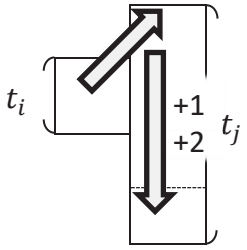


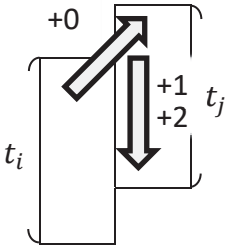
$$\begin{cases} M[j,d] = \infty & ,d < -u_j \\ M[i,d] - k = M[j,d-\mathrm{k}] & ,-u_j \le d \le s_i - \mathrm{k} \\ M[j,d] = \infty & ,s_i - k < d \end{cases}$$

$\Rightarrow best_j = best_i - k$

II. $\mathbf{u_i \le u_j}$

a. $E_i < E_j - k$ ， $E_i \ge E_j - k$





$$\begin{cases} M[j,d] = \infty & ,d < -u_j \\ M[i,-u_i] = M[j,-u_j] & \\ |D_i|_1 + k = |D_j|_1 & \\ M[j,d] = \infty & ,t_j < d \end{cases}$$

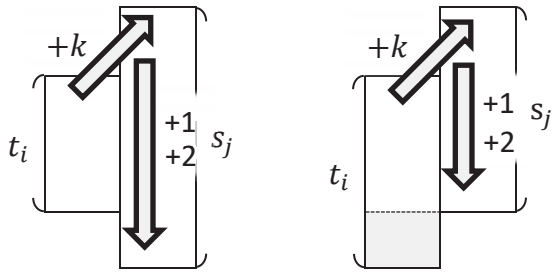
$\Rightarrow best_j = best_i$

（- , +）先減再增的時候：$(u_i < 0, u_j > 0)$

都要補複製 $(u_i < u_j)$

I. $u_i < u_j$

a. $E_i < E_j$ ， $E_i \geq E_j$



$$\begin{cases} M[i,-u_i]+k=M[j,1] \\ |D_i|_1-u_i=|D_j|_1 \\ M[j,d]=\infty \quad ,s_j<d \end{cases}$$

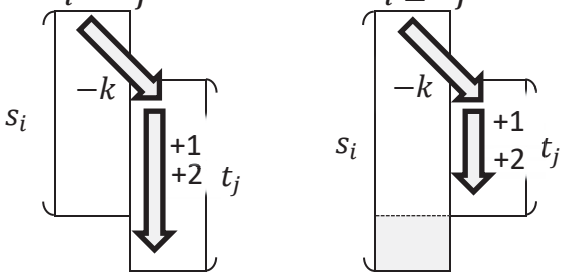
$\Rightarrow best_j = best_i + k$

（+ , -）先增再減的時候：$(u_i > 0, u_j < 0)$

都要補刪減 $(u_i > u_j)$

I. $u_i > u_j$

a. $E_i < E_j$ ， $E_i \geq E_j$



$$\begin{cases} M[i,1]-k=M[j,-u_j] \\ |D_i|_1+u_i=|D_j|_1 \\ M[j,d]=\infty \quad ,t_j<d \end{cases}$$

$\Rightarrow best_j = best_i - k$

※ 透過以上的分析可看出 $best_j$ 確實來自 $best_i$。

(u_i, u_j)	$(+,+)$	$(-,-)$	$(-,+)$	$(+,-)$
$u_i > u_j$	$best_j = best_i$	$best_j = best_i - k$		$best_j = best_i - u_j$
$u_i = u_j$	$best_j = best_i$	$best_j = best_i$		
$u_i < u_j$	$best_j = best_i + k$	$best_j = best_i$	$best_j = best_i + u_j$	

i. e. $bast_j = bast_i +$ (補刪減) + (補複製)

計算複雜度為 $O(n)$

證明我們所建的演算法 GreedyCntpAlg 之計算複雜度，與 **Zeira et al.** 建立的演算法 LinearCntpAlg 之計算複雜度想達到的目標，同為多項式時間 $O(n)$。

原因：最佳值 $best_j$ 來自前一項的最佳值 $best_i$，故每行只需計算最佳值，每次計算是常數時間 (輸入的時間 ≈ 輸出的時間)。換句話說當目標基因全長為 n，而計算最佳值的時間為常數 K 時，此時計算時間等於 nK，因此計算複雜度為多項式時間 $O(n)$。

五、結論

我們在原論文的基礎之上，修正其方法，並分析所有的狀況。修正的地方有：

1. 將 zero-skipping 的部分，在輸入 S 與 T 時提前做修正，這樣便能和刪減一起討論，能讓後續證明更加簡明。

2. 原論文的 LinearCntpAlg 含有許多未知數，並在後續分析狀況時的分類不夠精細。我們接續其想法，提出新的演算法 GreedyCntpAlg，並討論每種狀況，證明 $best_j$ 確實來自 $best_i$。

正常細胞與突變細胞的距離，確實能夠在線性時間內算出。距離愈大是否代表距離癌愈近，有待進一步用真實數據分析。

六、參考文獻

1. Zeira R., Zehavi M., Shamir R. 2017。A Linear-Time Algorithm for the Copy Number Transformation Problem。
https://www.ncbi.nlm.nih.gov/pubmed/28837352

2. Schwarz R.F., Trinh A., Sipos B., Brenton J.D., Goldman N., Markowetz F. 2014。Phylogenetic quantification of intra-tumour heterogeneity。
https://www.ncbi.nlm.nih.gov/pubmed/24743184

3. TheCancer GenomeAtlas Research Network. 2011。Integrated genomic analyses of ovarian carcinoma。
https://www.ncbi.nlm.nih.gov/pubmed/21720365

4. Chowdhury S.A., Shackney S.E., Heselmeyer-Haddad K., Ried T., Schäffer A.A., Schwartz R. 2013。Phylogenetic analysis of multiprobe fluorescence in situ hybridization data from tumor cell populations。
https://www.ncbi.nlm.nih.gov/pubmed/23812984

5. Chowdhury S.A., Shackney S.E., Heselmeyer-Haddad K., Ried T., Schäffer A.A., Schwartz R. 2014。Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics。PLoS Comput. Biol。
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003740

6. Chowdhury S.A., Gertz E.M., Wangsa D., Heselmeyer-Haddad K., Ried T., Schäffer A.A., Schwartz R. 2015。Inferring models of multiscale copy number evolution for single-tumor phylogenetics。
https://www.ncbi.nlm.nih.gov/pubmed/26072490

7. Letouzé E., Allory Y., Bollet M.A., Radvanyi F., Guyon F. 2010。Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis。
https://www.ncbi.nlm.nih.gov/pubmed/20649963

七、附錄

程式（Excel）：DpCntpAlg、GreedyCntpAlg、LinearCntpAlg。

https://drive.google.com/drive/folders/1b2Mkk1xuff0OX2f92TvyqjdYUKvexrj5?usp=sharing

Correction of the linear-time alogorithm for the Copy Number Transformation Problem

Dun-Siang Yang* Tsung-Yi Yang* Wei-Hua Hsieh*

Abstract

Schwarz et al. set the distance between copy number profiles to explore the gap between normal genome and tumor genome. They proposed the MEDICC algorithm for calculating distance[2], but did not analyze its complexity. However, in some cases, MEDICC would be exponential time. After that, Zeira et al. proposed a linear time algorithm [1], but we found the process that leading out the result has something wrong. So we propose a new algorithm and make some changes to the original method.

Keywords: Cancer, Genome rearrangement, Copy number profile

*Department of Applied Mathematics, Tunghai University, Taichung 40704, Taiwan