Tunghai Science Vol. 21: 1–14
July, 2019

基因複製樹問題的 NP-hardness 性質

楊宗頤*　楊敦翔*　謝維華*

摘　要

癌是個動態過程，快速的突變形成複雜的腫瘤基因組，了解其與正常基因組的差異及其演化過程，可幫助我們預測疾病的進展與可能的醫療介入。為探討癌細胞的基因演化過程，El-Kebir et al. 建立基因複製樹問題，並證明其為NP-hard，但我們發現該證明並不正確。我們建立修訂基因複製樹問題，並建立了一個多項式轉換，將修訂最簡約演化樹問題轉換至修訂基因複製樹問題。

關鍵字：癌、基因組重新排列、基因複製向量、演化樹、最簡約假設、NP-hard

一、背　景

在癌與物種演化的研究裡，基因組重新排列（genome rearrangement）是個核心問題。過去 20 年，以計算方法對此問題有些深入的研究（Fertin et al., 2009），但大部分是在探討物種演化。物種演化以百萬年為單位，其基因突變，一代傳一代；然而，一個個體內癌症的基因突變，就只在幾十年內發生。2005 年開始的 The Cancer Genome Atlas 計畫，收集且分類了各種癌症病人的基因突變數據。這些數據的收集，已經是常態性的工作。因此目前有絕佳的機會，利用這些數據，以計算方法了解癌症基因組的演化過程——此為一個未解的問題。

癌是個動態過程，快速的突變，形成複雜的腫瘤基因組。其中，有很多的突變，是整段 DNA 的**刪減**（deletion）或**複製**（amplification），使得腫瘤基因組內的基因個數不斷改變，且與正常基因組不同。了解它們的差異及其演化過程，可幫助我們預測疾病的進展與可能的醫療介入。

*東海大學應用數學系

Schwarz et al. (2014) 訂出基因複製向量（copy-number profile）間的距離，以探討正常基因組與各階段腫瘤基因組的差異，但目前為止並未引起科學界的注意。

以計算方法推論出演化樹（Phylogenetic tree）是物種演化論的根本問題。El-Kebir et al.以基因複製向量間的距離，建立了基因複製樹問題（The copy-number tree problem），以探討癌細胞的基因演化，並證明此問題為 NP-hard [1]。我們發現該證明可能不正確或不完整，因此本論文的目的即在探討該問題的 NP-hardness 性質。

二、NP-completeness 理論

要證明一個最佳化問題（optimization problem）為 NP-hard，我們首先要將其轉為決策問題（decision problem），並證明此決策問題為 NP-complete，如此則原最佳化問題即為 NP-hard。

NP-completeness 理論主要是在探討決策問題（答案為 Yes 或 No 的問題，即為決策問題）[2]。透過以下步驟，可證明一個決策問題 Π 為 NP-complete：

1. 確認 Π 屬於 NP 問題（可由 polynomial time nondeterministic algorithm 解的問題）。
2. 找一個已被證明為 NP-complete 之問題 Π′。
3. 建立一個轉換（transformation）f，可將 Π′ 的例題 I（instance）轉換成 Π 的例題 $f(I)$，並滿足以下若且唯若關係：I 的答案為 Yes 若且唯若 $f(I)$ 的答案為 Yes。
4. 確認 f 為 polynomial transformation（亦即可於多項式時間內完成轉換）。

三、El-Kebir et al.的基因複製樹問題

定義：基因複製向量（copy-number profile，以下簡稱 CNP）

CNP 為一個長度為 n 的向量，其中每一數字依序代表染色體上各種基因的個數。

正常細胞染色體上各種基因的個數為二，因此其 CNP 為 $(2\ 2\ 2 \ldots 2\ 2)$。由於基因的突變，所形成的癌細胞同時存在各種不同的 CNP 數字組合，例如：$(1\ 2\ 5 \ldots 3\ 2)$。了解如何由正常的 $(2\ 2\ 2 \ldots 2\ 2)$ 突變為各種不同 CNP 的演化過程，有助於預測癌的進展與可能的醫療介入。

定義：突變事件 $c = (s, t, b)$

一突變事件 (s, t, b) 代表 CNP 從位置 s 到位置 t，基因的個數都加 b 個，其中若 $b = 1$ 代表該基因複製一個；若 $b = -1$ 則代表該基因刪減一個。

例子：一個 $n = 6$ 的 CNP $(1\ 1\ 2\ 3\ 3\ 2)$ 可經由兩個突變事件 $(2,4,-1)$ 與 $(1,5,1)$，突變為 $(2\ 0\ 2\ 3\ 4\ 2)$。

上述的例子當中，$(2\ 0\ 2\ 3\ 4\ 2)$ 的第二個位置數字為 0，代表該基因個數已突變為 0，因此已無刪減或複製的可能，亦即任何突變事件皆已無法改變該基因的個數，該位置將一直為 0。

定義：CNP 距離

若一 CNP，可經由一系列的突變事件 $(s_1, t_1, b_1),\ \ldots\ , (s_q, t_q, b_q)$，突變到另一個 CNP，則**最少**可能的突變事件個數 q，即為它們的距離。

定義：基因複製樹問題（The copy number tree problem，以下簡稱 CNT）

以癌細胞當下各個 CNP 為葉（假設長度為 n，且有 k 個），以正常細胞數字全為 2 的 CNP 為根，限定所有 CNP 裡的數都不大於一個給定的常數 e，在最簡約假設（Maximum parsimony assumption）下，找出有根滿二元樹（rooted full binary tree）T，使得

(1) 其內部節點各由一個 CNP 代表

(2) 有邊相連節點的 CNP 距離總和 $\Delta(T)$ 最小。

此模型以有根滿二元樹，描述正常基因組突變為各階段腫瘤基因組的演化過程，不失一般性，因其足以表現其它各種有根樹。因為任何 out-degree 大於 2 的節點，皆可將其改造為 out-degree 為 2 的節點（CNP 距離不改變）；out-degree 等於 1 的內部非根節點，亦可以將其直接剔除（CNP 距離可能更小）。有個特例，即當根的 degree 為 1 時。為了處理此種情況，El-Kebir et al. 加了一片 CNP 全為 2 的葉子，意即正常的 CNP $(2\ 2\ 2 \ldots 2\ 2)$。El-Kebir et al. 證明（加了一片葉子的）CNT 為 NP-hard，此特殊動作扮演著關鍵角色。我們修改了 CNT，容許根的 degree 為 1，即是將此特殊動作剔除。

定義：最簡約演化樹問題（The maximum parsimony phylogeny problem，以下簡稱 MPP）

給定 k 個長度為 n 的二元向量（binary vector），以之為葉，以 0 向量為根，在最簡約假設下，找出一個有根滿二元樹 T，使得

(1) 其內部節點各由一個長度為 n 的二元向量代表

(2) 有邊相連節點的 Hamming distance 總和 $\Delta(T)$ 最小。

MPP 的決策版本已被證明為 NP-complete [3]。El-Kebir et al. 將 MPP 的決策版本，轉換到 CNT 的決策版本，以證明 CNT 為 NP-hard。

3.1 El‑Kebir et al. 的轉換

步驟一：將 MPP 長度為 n 的二元向量（葉）裡的數字 0 全部換成 2。

步驟二：將向量裡的任二數字間卡入 2、 1 交錯，長度為 nk 的向量 Ω。

（形成的向量，即為 CNT 作為葉的複製向量。）

例子：$n = 4,\ k = 3,$ MPP 的二元向量 $(1\ 0\ 0\ 0)$，則轉換成的 CNT 向量為 $(1\ \Omega\ 2\ \Omega\ 2\ \Omega\ 2)$，其中 $\Omega = (2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1\ 2\ 1)$。

3.2 El‑Kebir et al. 的證明

必須證明以下的若且唯若關係：若 MPP 的例子存在一個距離小於等於 h 的有根滿二元樹，若且唯若，轉換過去的 CNT 例子，亦存在一個距離小於等於 $h + W$ 的基因複製樹，其中 $W = \frac{(n-1)nk}{2}$。

（⇒）方向的證明，El-Kebir et al.將 MPP 距離小於等於 h 的滿二元樹保留結構，然後將該樹所有節點的二元向量轉換為 CNP，最後所形成的樹可證明為距離小於等於$h + W$ 的基因複製樹。

（⇐）方向的證明，距離小於等於 $h + W$ 的 CNT 基因複製樹裡，每個內部節點所代表的 CNP，El-Kebir et al.皆假設為上述卡入 Ω 的形式，但這是不存在的假設。因此其證明可能不正確或不完整，有待進一步去釐清，因此我們修訂了 El-Kebir et al. 的 CNT 問題。

四、修訂 CNT

定義：修訂 CNT（The revised copy-number tree problem，以下簡稱 RCNT）

以癌細胞當下各個**不同**的 CNP 為葉（假設長度為 n，且有 k 個），以正常細胞數字全為 2 的 CNP 為根，限定所有 CNP 裡的數都不大於一個給定的常數 e，在最簡約假設下，找出有根滿二元樹 T，**但根的 degree 可為 1 或 2**，使得

(1) 其內部節點各由一個 CNP 代表

(2) 有邊相連節點的 CNP 距離總和 $\Delta(T)$ 最小。

El-Kebir et al.建構的 CNT，並沒有要求任兩片葉子的 CNP 必須不同。其實若兩片葉子有相同的 CNP，我們可以透過合併或移除，將問題轉成任兩片葉子的 CNP 皆不同，而距離總和只會更小（如圖一）。我們的目的在探討癌細胞的基因演化過程，因此以癌細胞當下各個不同的 CNP 為葉。兩片葉子的 CNP 相同，無助於探討此問題，且出現的機率極低。因此 RCNT 要求任兩片葉子的 CNP 必須不同。

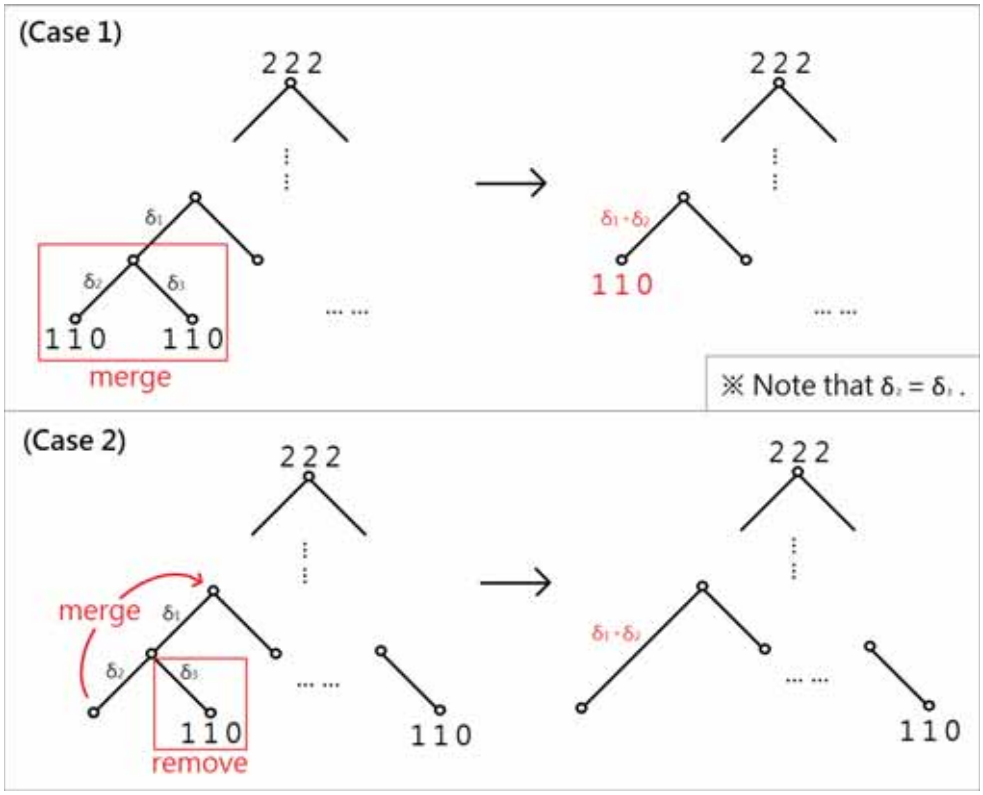

圖 1：當有二片葉子 CNP 相同時，可透過上述方法加以合併，達到減少 $\Delta(T)$ 的效果。

要驗算（polynomial time nondeterministic algorithm 的 checking stage）有根滿二元樹有邊相連節點的 CNP 距離總和，很明顯可於多項式時間內完成，因此 RCNT 問題的決策版本屬於 NP。

定義：修訂 MPP 問題（The revised maximum parsimony phylogeny problem，以下簡稱 RMPP）

給定 k 個長度為 n 的二元向量，以之為葉，以 0 向量為根，在最簡約假設下，找出一個有根滿二元樹 T，**但根的 degree 可為 1 或 2**，使得

(1) 其內部節點各由一個長度為 n 的二元向量代表

(2) 有邊相連節點的 Hamming distance 總和 $\Delta(T)$ 最小。

此論文的主要結果，在將 RMPP 的決策版本，轉換到 RCNT 的決策版本，然後證明 NP-completeness 理論所要求的若且唯若關係。

4.1 轉換

我們將RMPP決策問題的二元向量 $b_i = (b_{i,1} \quad b_{i,2} \quad \dots \quad b_{i,n}),\ 0 \le i \le k$，透過轉換 Φ，做出 RCNT 決策問題的 CNP 如下：

$$c_i = \Phi(b_i) = (\phi(b_{i,1}) \quad \Omega \quad \phi(b_{i,2}) \quad \Omega \ \dots \ \Omega \quad \phi(b_{i,n}))$$

其中

$$\phi(b_{i,s}) = \begin{cases} 1, & if\ b_{i,s} = 1 \\ 2, & otherwise \end{cases},\ \text{for } 0 \le i \le k,\ 0 \le s \le n$$

此轉換將 b_i 向量的數字之間皆卡入一個向量

$$\Omega = (2 \quad 1 \quad 2 \quad 1 \dots 1 \quad 2)$$

其中 Ω 的長度 $= \begin{cases} nk\ ,\ if\ n\ and\ k\ are\ both\ odd \\ nk+1\ ,\ otherwise \end{cases}$

Ω 的最後一個數字，我們設計為 2（El-Kebir et al.並無此要求）。這些 Ω 所在的位置，以下稱作牆。令 $W = \frac{(nk+1)(n-1)}{2}$，則任一葉子中所有 Ω 內數字 1 的個

數小於 W（見下表）。因為由 Φ 轉換過來的向量，其數字非 1 即 2，並形成 RCNT 之葉，而根為數字全為 2 的向量，因此在最簡約假設下，我們令 $e = 2$。

計算函數Φ，最多需要 $(nk+1)(n-1)+nk$ 個步驟，因此其計算複雜度為 $O(n^2k)$，故Φ為 polynomial transformation。

nk	odd	even
Ω 中數字 1 的個數	$\frac{nk-1}{2}$	$\frac{nk}{2}$
Ω 中數字 2 的個數	$\frac{nk+1}{2}$	$\frac{nk}{2}+1$

4.2 證明

RMPP 決策問題的一個例題 $b_1, \dots, b_k$，存在一棵樹 T，$\Delta(T) \le h$	$\Leftrightarrow$	轉換過來的 RCNT 例題 $c_1, \dots, c_k$，存在一棵樹 T'，$\Delta(T') \le h + W$

（⇒）我們分為T的根 degree 為 1 與 degree 為 2 兩種情況來探討：

Case 1：根的 degree 為 1（亦即只有一個邊從根連出來）

保留 T 的結構，但經由 Φ 將 T 所有的內部節點與葉子做轉換；根的部分則全部換成數字 2，並補足其長度。如此形成我們要的 T'，其根與下接的 CNP，距離最多增加 W（為了將根牆內的數字 2 轉成數字 1），其它距離總和仍為$\Delta(T)$，沒有改變，因此 $\Delta(T') \le h + W$。

Case 2：根的 degree 為 2（亦即有兩個邊從根連出來）

保留 T 的結構，但從其根往上接一個新的根，其 CNP 訂為$(2\ 2\ 2 \dots 2\ 2)$，形成我們要的 T' 的結構。然後我們經由 Φ 將 T 的原有節點（包含根與葉子）做轉換。因此 T' 的根與下接的 CNP，距離 $\le W$，而其它距離總和仍為 $\Delta(T)$，故 $\Delta(T') \le h + W$。 ■

（⇐） 我們分為 $h \geq nk-k+1$ 與 $h < nk-k+1$兩個情況來探討：

Case 1：$h \geq nk-k+1$

RMPP 二元向量 $b_1, \ldots, b_k$兩兩皆不相同，所以這些向量合起來至少有 $k-1$ 個數字 0，亦即最多只有 $nk-(k-1)$ 個數字 1。所以存在一棵樹 T，其內部節點皆為 0 向量， $\Delta(T) \leq nk-(k-1) = nk-k+1 \leq h$。

Case 2：$h < nk-k+1$

當 $n < 3$ 的時候，複雜度低，可以窮舉法（Exhaustive search）列出所有可能情況；又討論 NP-completeness 性質時，我們在意的是當變數很大的情況。因此不失一般性，我們令 $n \geq 3$，則

$$h < nk-k+1 < nk+1 = \frac{(nk+1)(3-1)}{2} \leq \frac{(nk+1)(n-1)}{2} = W。$$

T' 根的 degree 必為 1，否則因從根往下分成兩邊，各至少連一片葉子，因此 $\Delta(T') \geq 2W \geq h+W$（從根至任一片葉子，至少需 W 個突變事件，以便將根牆內的數字 2 轉成數字 1）。

接下來，證明以下敘述：存在另一棵樹 T″，其根的 degree 為 1，內部節點的牆皆為 Ω 的形式，且 $\Delta(T'') \leq h+W$（T′的牆未必是 Ω 的形式）。

***Claim* 1.** T' 內部節點（與根相連的節點除外）的任一牆，不會全部都是 1 或全部都是 2。

〈**證明**〉我們使用反證。若內部節點有一牆全部都是 1 或全部都是 2。該節點有一上接節點，並下接至另一節點。從根出發，經過該節點至葉子，距離至少為 W。原節點下接至少 2 片葉子，且葉子的牆皆為 Ω 的形式。考慮以下四種情況：

(1) 若有一牆全部都是 **1**，且 $\boldsymbol{nk}$ **為奇數**。

為了形成下接葉子所需的 Ω 形式，必須將牆裡的 $\frac{nk+1}{2}$ 個 1 全部轉為數字 2，

因而

$$\Delta(T') \geq \frac{nk+1}{2} + \frac{nk+1}{2} + W = nk + 1 + W > h + W$$

因此推得矛盾。

(2) 若有一牆全部都是 **1**，且 ***nk* 為偶數**。

為了形成下接葉子所需的 Ω 形式，必須將牆裡的 $\left(\frac{nk}{2}+1\right)$ 個 1 全部轉為數字 2，因而

$$\Delta(T') \geq \left(\frac{nk}{2}+1\right) + \left(\frac{nk}{2}+1\right) + W = nk + 2 + W > h + W$$

因此推得矛盾。

(3) 若有一牆全部都是 **2**，且 ***nk* 為奇數**。

為了形成下接葉子所需的 Ω 形式，必須將牆裡的 $\frac{nk-1}{2}$ 個 2 全部轉為數字 1，因而

$$\Delta(T') \geq \frac{nk-1}{2} + \frac{nk-1}{2} + W = nk - 1 + W > h + W$$

因此推得矛盾。

(4) 若有一牆全部都是 **2**，且 ***nk* 為偶數**。

為了形成下接葉子所需的 Ω 形式，必須將牆裡的 $\frac{nk}{2}$ 個 2 全部轉為數字 1，因而

$$\Delta(T') \geq \frac{nk}{2} + \frac{nk}{2} + W = nk + W > h + W$$

因此推得矛盾。

二個 CNP 的距離，Zeira et al. 已證明可由**刪減**全在前、**複製**全在後的系列突變事件決定（稱作 ordered）[4]。

***Claim* 2**. 除了根和與其相連的節點外，T'的複製或刪減事件，最多包含一個非牆的位置。

〈**證明**〉(1) 若**複製**事件跨越至 2 個以上的非牆位置：

根據 ***Claim* 1**.，任一牆，數字不會全部都是 1。要**複製**前，必須將牆裡的所有數字 2 **刪減**至 1（$e = 2$）。然而，在**複製**之後為滿足 ***Claim* 1**.，牆裡的若干位置必需**刪減**至 1，此處違背了 Zeira et al. 所提出的 Ordered 性質。故此種情況不會發生。

(2) 若**刪減**事件跨越至 2 個以上的非牆位置：

根據 ***Claim* 1**.，任一牆，數字不會全部都是 2。要**刪減**前，必須先將牆裡的所有數字 1 **複製**至 2，然後**刪減**。此處違背了 Zeira et al. 所提出的 Ordered 性質。故此種情況不會發生。

與根相連節點的牆，與其左右緊鄰的兩個非牆位置，若全為數字 1，代表其下接的兩個節點相對應的兩個非牆位置有若干數字為 1。又***Claim* 1**.的關係，該兩個下接節點的牆，至少需有一個突變事件，將牆內的數字 1 轉為數字 2。我們可用其它突變事件取代上述的突變事件，使得與根相連節點下接的兩個節點的兩個非牆位置數字不變，牆內數字不全為 1，且 $\Delta(T')$ 不會增加。

T' 內所有涵蓋非牆位置的**刪減**與**複製**事件，我們將其突變範圍皆縮短至只對非牆位置。因此，T'的所有突變事件，變成兩類如下，其個數和 $\leq h + W$：

(1) 突變範圍只在牆裡

(2) 突變範圍只對一個非牆位置

由根 (2 2 2 …2 2) 突變至含有 Ω 的葉子，至少要有 W 個突變。因第(2)類突變在範圍縮短前，無助於此，實由第(1)類突變加上若干範圍縮短前的突變達成此目的，故第(1)類突變次數 $\geq W$，因此第(2)類突變次數 $\leq h$。

我們將 T'所有牆的數字換成 Ω（根除外，仍為(2 2 2 …2 2)），形成新樹T''，$\Delta(T'')$ 等於第(2)類突變的個數加上 W（由根突變至其下接節點），所以$\Delta(T'') \leq h + W$。

最後，將T''所有的 Ω 移除（根則移除相對應位置的 2），並將T''的所有數字 2 全部換成數字 0，我們即得到 RMPP 的一棵樹 T，其 Hamming distance 總合為 $\Delta(T) = \Delta(T'') - W \leq h$。

■

五、結論

El-Kebir et al.的 CNT 問題，特別加了一片(2 2 2 … 2 2)的葉子，一是為了形成滿二元樹（在其⇒方向的證明）；另一是為了證明 CNT 為 NP-complete（在其⇐方向的證明），但我們發現該證明並不正確。我們的目的在探討癌細胞的基因演化過程，因此應以癌細胞當下各個不同的 CNP 為葉。加了一片(2 2 2 … 2 2)的未突變葉子，違背此目的。且由正常的根(2 2 2 … 2 2)，經過一系列突變後，回復至正常(2 2 2 … 2 2)的葉子，機率極低。

我們剔除該片葉子，將 CNT 修改為 RCNT。此論文的主要結果，在將 RMPP 的決策版本，轉換到 RCNT 的決策版本，然後證明 NP-completeness 理論所要求的若且唯若關係。

RMPP 的決策問題是否為 NP-complete，有待證明。若答案是肯定的，則 RCNT 問題即為 NP-hard。

六、參考文獻

1. Mohammed El-Kebir, Benjamin J. Raphael, Ron Shamir, Roded Sharan, Simone Zaccaria, Meirav Zehavi and Ron Zeira（2017）。Complexity and algorithms for copy‑number evolution problems。Algorithms for Molecular Biology。https://doi.org/10.1186/s13015-017-0103-2。

2. Garey, Michael R./ Johnson, David S.（1979）。Computers and Intractability: A Guide to the Theory of NP-completeness。W H Freeman & Co。

3. Foulds LR, Graham RL.（1982）。The Steiner problem in phylogeny is NP-complete。Advances Applied Mathematics。https://www.sciencedirect.com/science/article/pii/S0196885882800043。

4. Zeira R, Zehavi M, Shamir R（2017）。A Linear-Time Algorithm for the Copy Number Transformation Problem。Journal of Computational Biology。https://www.ncbi.nlm.nih.gov/pubmed/28837352。

The NP-hardness property of the copy-number tree problem

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

Abstract

Cancer is a dynamic process. Rapid mutations form complex tumor genomes. Understanding their differences with normal genomes and their evolution can help predict the evolution of the disease and possible medical interventions. To predict the evolution of cancer cells, El-Kebir et al. established the copy-number tree problem and proved it to be NP-hard. We found that the proof is not exactly correct. Thus, we established the revised copy-number tree problem and transformed the revised maximum parsimony phylogeny problem to the revised copy-number tree problem by a polynomial transformation.

Keywords: Cancer, Genome rearrangement, Copy number profile, Phylogeny, Maximum parsimony, NP-hard

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