以生物資訊方式偵測染色體異常

Bioinformatics Approach to Detect Chromosomal Abnormality

許芳榮
逢甲大學資訊工程學系
生醫資訊暨生醫工程碩士學位學程
Email:frhsu@fcu.edu.tw

梁容慈
逢甲大學資訊工程學系
Email:M9702799@fcu.edu.tw

***摘要*—人類序列在 DNA 複製過程中發生重新排列組合，將會影響基因的表現，進而成為遺傳工程上重要的訊息。近年來，染色體異常的偵測是一個重要的議題。異常的現象包括了轉位(translocations)、刪除(deletions)、反轉(inversions)，造成許多染色體的併發症狀以及許多類型的癌症。此篇的目的在於偵測哪些基因容易發生異常。資料來源為人類的 EST 序列，數量約為八百萬條。經過判斷、比對、分析、以及動態規劃所過濾出的染色體異常的 EST 數目將近一萬五千多筆。其中轉位、刪除、反轉事件分別有五千多筆、九千多筆、以及一千多筆。**

***關鍵詞*—轉位、染色體異常**

***Abstract*—Since there occurs an unavoidable rearrangement of human DNA sequence during its transcription ,such a rearrangement will influence the expression of the protein amino acid sequence, which becomes important information on genetic engineering. The identification of chromosome abnormalities is an essential part of chromosome syndromes and many types of cancer involved in translocations, deletions, and inversions. The purpose of this paper is identification which genes prone to occurs abnormalities. Identification scope of all human's EST sequence, quantity approximately are 8,000,000, will undergo judgment, alignment, analysis, as well as dynamic programming, will filter nearly 15,000, in which translocation and deletion and inversion event about 5000 and 9000 and 1000 respectively.**

***keywords*—translocation、Chromosomal abnormality**

一、緒論

染色體變異會造成癌症、視力與聽覺上的障礙、外觀上的異常的現象，及身心發展遲緩。Theodor Boveri 是第一個提出腫瘤的形成可能來自於染色體的不穩定。1960 年 Nowell 和 Hungerford 發現了與慢性骨髓性白血病(CML)有關的費城染色體，並且在之後研究指出是由於染色體 9 號以及 22 號發生轉位而產生的[6,9]。染色體轉位一般發生在減數分裂時，產生了非同源染色體的片段重新排列的異常現象，當這不正常的染色體轉位沒有遺失任何 DNA 時，稱為平衡轉位(balance translocation)，反之有遺失 DNA 或是多出 DNA，稱為不平衡轉位(unbalance translocation)。

癌症細胞會造成基因體不穩定，使得染色體位置發生錯誤的重組。錯誤的重組會導致刪除、轉向、轉位，造成潛在的致癌基因被活化或者因此喪失抑制癌症的基因。雖然我們尚未明白這中間的過程，但是我們知道 DNA 斷裂是造成一連串行為的第一個事件。

DNA 的斷裂來自於內源性損傷像是正常代

謝的副產物活性氧分子攻擊導致的損傷或是外源性損傷由外部因素引起，像是太陽的紫外射線。

近年來轉位陸陸續續被發現，所以有人也因此將與癌症有關的轉位事件建構成資料庫供人方便查詢。例如 TICdb[4]資料庫共有 1445 個切位，而參與轉位的基因共有 310 個基因，另外也指出發現的轉位切位大部份都落於 intron 上。

目前的相關研究指出由於轉位所引起的癌症細胞中有超過一半以上的是比對到脆裂點(fragile site)，而脆裂點指的是染色體上非隨機在細胞分裂中期(metaphase)容易發生斷裂的位置。由於目前只有121個不同的脆裂點被辨識出來，所以其中不免有已被證實參與許多轉位事件的切位但卻沒落於脆裂點上[2]。另外一項研究指出迴文序列(palindrome)導致序列發生被刪除或是轉位，其中的原因在於迴文序列容易形成十字的二級結構，而十字的二級結構是不穩定的結構，因此增加了染色體發生斷裂的機會[10]。

本篇目的在於找出基因體中哪些基因以及位置容易發生斷裂，導致錯誤的黏接以致於活化潛在的致癌基因、喪失抑制癌症的基因或是其他的疾病。

二、染色體變異比對問題

染色體的變異分為下列三種類型

(一) 轉位(translocation)

當序列經過斷裂、重新與非同源染色體接合後(圖 2.1)，造成整條序列對齊到基因體分數偏低，當經過適當的裁切，左右兩序列分別完整的重新對齊到基因體上，所表現的特徵在於使得當兩條新序列重新定位到不同的染色體上。

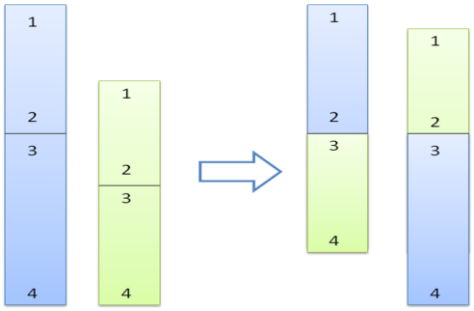

圖 2.1 translocation

(二) 刪除(deletion)

人類共有 23 對染色體(包括 1 對 XX 或 XY 的性染色體)，若任何染色體上發生了片段的缺失，都可能造成患者出現異常的臨床表徵(圖 2.2)。當序列發生片段的遺失時，會造成序列比對分數偏低，在經過適當的裁切，左右兩序列重新對其到基因體上，所表現的特徵在於此兩條新序列會被定位到同一條染色體上且方向一致。

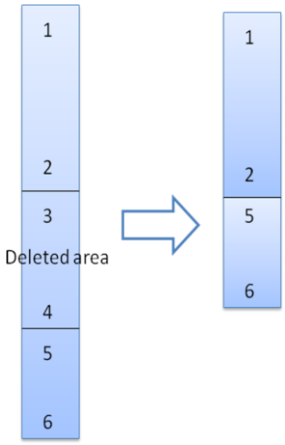

圖 2.2 delete

(三) 反轉(inversion)

當序列經過斷裂、重新接合後，也可能發生序列方向相反的情形(圖 2.3)，造成整條序列對齊到基因體分數偏低，在經過適當的裁切，左右兩序列重新對其到基因體上，所表現的特徵在於兩條新序列會被定位到同一條染色體卻是不同的方向(其中一條是正股，另一條是反股)。

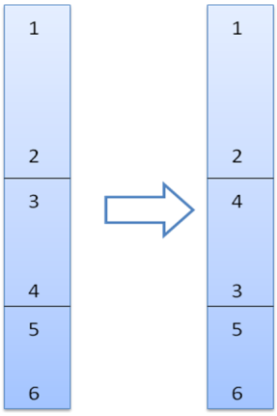

圖 2.3 inversion

當染色體發生染色體變異，有兩件事件是我們所觀察到的現象，事件一:若在融合片段產生EST，該EST比對到染色體時將無法完全比對，事件二:若我們將EST由中央裁切成兩段序列，其中必定有一段序列能完全的比對到染色體上。而以下為我們所定義出染色體變異的問題。

給定基因體序列G= $(g_1,g_2, \ldots g_m)$以及一條EST序列E= $(e_1,e_2,\ldots e_n)$;而S(G,E)為序列E比對到序列G的比對分數，我們希望找到一個位置i使得$E_1=(e_1,\ldots,e_i)$、$E_2=(e_{i+1},\ldots e_n)$ 且 $(S(G,E_1)+S(G,E_2))$比對分數為最大值。

三、方法

(一) 資料來源

表現序列標籤(Expressed Sequence Tag, EST)是 cDNA 序列片段，雖然只是 cDNA 的片段序列，但具快速、低廉且直觀的優點。dbEST 可免費下載於 NCBI ftp://ftp.ncbi.nih.gov/repository/dbEST/，本研究所使用的人類序列版本為 Build 36.3，而目前人類的 EST 數量為 8,296,241 條。

(二) 方法階段

第一階段:過濾資料

將不是由染色體異常所產生的 EST 過濾出，過濾的依據為若比對分數大於 80%則過濾掉。由於使用一種比對工具無法確保每條 EST 都有最佳的比對方式，所以這邊使用兩種比對工具分別為 Mugup[5]以及 Gmap[11]。另外由於 Mugup 利用多層單標籤的概念，改善了傳統較慢的比對方法，因此一開始使用 Mugup 將所有的 EST 進行比對，此時比對分數小於 80%之 EST 數量為約 75 萬筆，接著將這 75 萬筆 EST 再利用 Gmap 做第二次的過濾，把比對分數大於 80%的 EST 過濾掉，最後剩下的資料為約 73 萬筆之 EST。

在這 73 萬筆資料中除了轉位、刪除、反轉之外還有許多未知的現象導致低的比對分數，另外由於 EST 是人工所產生出的資料，所以也有人為汙染的情況，因此過濾出較有可能由染色體異常產生的 EST，接著再進行找尋最佳切位。

將 EST 利用二元搜尋的概念由序列中間剪裁成兩段相同比例的序列，接著將兩段序列使用 Mugup 進行比對，依比對分數 90%為基準，會有下列四種情況產生。最後過濾出的資料量由原本的 73 萬減少為約 11 萬筆 EST 序列。

情況一:兩段序列皆大於 90%，表示最佳切位靠近於此 EST 中央的位置，因此將此 EST 納入研究的檔案中。

情況二:左邊序列大於 90%而右邊序列小於 90%，遇到這種情況則繼續二元搜尋由右邊序列中間剪裁成右邊 1/4 序列接著進行比對直到比對分數大於 90%或是序列的長度小於 70bp，若比對分數大於 90%表示此 EST 序列最佳切位靠近右半部，而若是序列長度小於 70bp 則把該筆 EST 過濾掉，因為太短的序列比對會失去它的意義。

情況三: 左邊序列小於 90%而右邊序列大於 90%遇到這種情況同樣繼續二元搜尋將左邊序列中間剪裁成左邊 1/4 序列接著進行比對直到比對分數大於 90%或是序列的長度小於 70bp 若比對分數大於 90%表示此 EST 序列最佳切位靠近左半部，而若序列長度小於 70bp 則過濾掉此 EST。

情況四:兩段序列皆小於 90%，依照觀察到染色體變異的事件中得知必定有一段序列能完全的比對到染色體上，故遇到情況四則將該 EST 序列過濾掉。

第二階段:找尋最佳切位

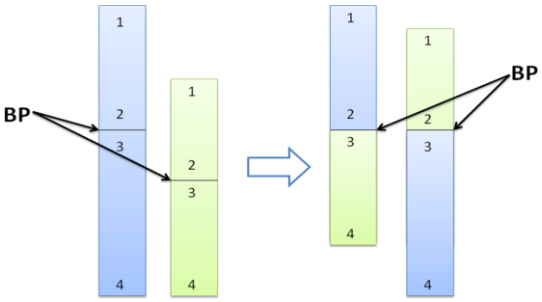

圖 3.1 找尋 translocation 的切位(BP)

由第一階段我們僅得知最佳切位所位於的片段，而此階段則是要尋找出最佳的切位。如圖 3.1 找尋出最佳切位使得原本的 EST 序列裁切為藍色以及綠色，並且兩段序列分別進行比對，分別能夠完全對應到個別的染色體上。

步驟一:根據上一階段得知最佳切位所位於的片段，依照該片段的比對分數裁切出該片段之一段序列(紅線)，可以幫助我們取出恰當的序列，接著將該序列使用 Mugup 比對。

步驟二:依據 Mugup 比對的結果檔，可以觀察出由哪個位置開始比對分數開始往下降(黃線)或往上升，此時已經非常接近最佳切位了。

步驟三:由於最佳切位位於比對分數下降點或上升點附近，因此將左序列往右延長 10bp 並且右序列往左延長 10bp overlap，總共在比對分數下降點或上升點兩側延伸 20bp，並且使用 Mugup 比對兩段延伸的序列(綠線)。

步驟四:依據 Mugup 比對結果檔，得知兩段序列所比對到之 contig 位置，左側延伸序列往右擷取長度為 40bp 之 contig 序列，而右側延伸序列往左擷取長度為 40bp 之 contig 序列(黑色虛線框)。

步驟五:左右兩側序列分別進行動態規劃，EST 長度為 20bp，contig 長度為 40bp。

步驟六:左右兩側序列分別進行動態規劃後分別得到 10 個比對分數，左右序列比對分數相加會有十筆比對分數，選擇最大的分數將會是最佳切位。圖 3.2 為其中一件事件之流程示意圖。

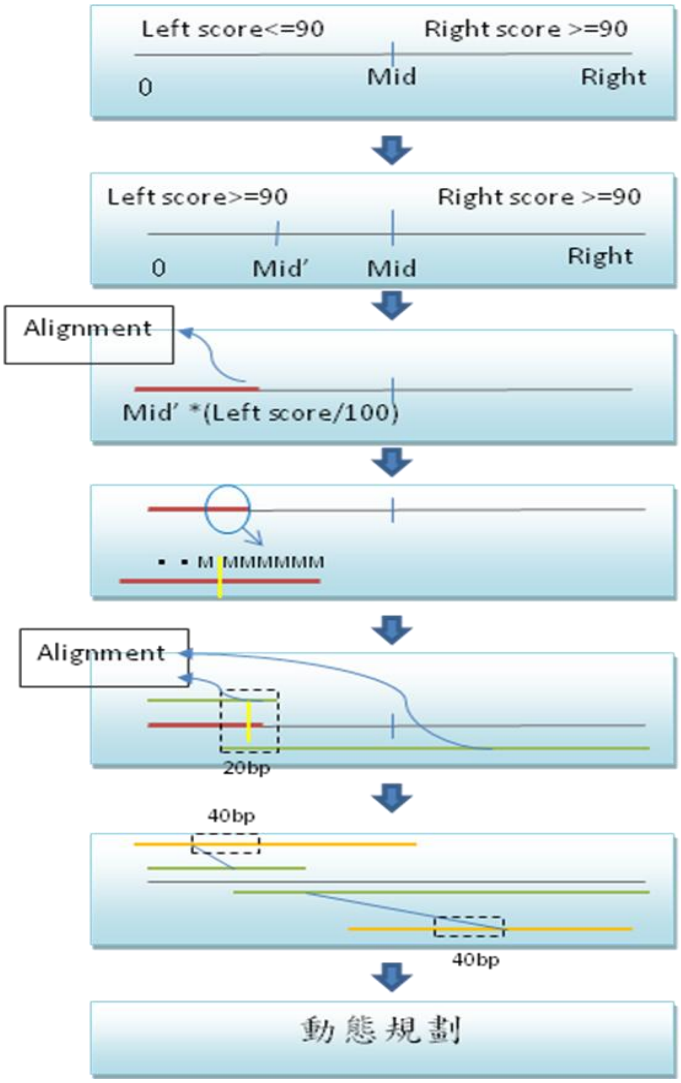

圖 3.2 最佳切位流程示意圖

第三階段:判斷事件

依照轉位、刪除、反轉之事件特徵判別該 EST 是屬於哪一種事件。

(三) 結果驗證

藉由人工的方式產生 479 筆轉位事件 816 筆刪除事件 925 筆反轉事件，該 EST 檔案中除了序列的資訊外另外還有正確的 EST 切位，因此使用我們的演算法去執行此檔案將可以驗證我們的

演算法是否能夠找到 EST 最佳切位。

由執行的結果發現，最佳切位並非絕對唯一， 某些 EST 在許多位置都屬於最佳切位，也就是說有某些位置可以屬於左邊的序列也可以屬於右邊的序列。而檔案中所有之 EST 都可由我們的演算法找到最佳切位，因此可證明出結果的可靠性。

四、結果與討論

分析結果轉位、刪除、以及反轉數目如表格 4.1 所示:

表 4.1 各個染色體異常事件次數

染色體異常事件	EST 發生筆數
translocation	5018
deletion	9339
inversion	1362

(一) 轉位

轉位事件中依照基因來做分析，得知在哪些基因容易發生斷裂而引起轉位現象。總共有 3978 個已知基因被尋找出來，在這僅列出前六筆資料(表 4.2)，依照發生斷裂的次數統計出來，進一步分析斷裂次數最多的基因。

基因 EIF3F 在轉位的事件中發生最頻繁，205 筆事件中其中有 165 筆是由基因 EIF3F 以及基因 LOC339799 所組成，EIF3F 在轉譯初始時扮演重要的角色，EIF3F 基因功能在癌細胞中是不明顯的，在胰臟癌中 EIF3F 的表現量下降，而造成胰臟癌細胞發生細胞自毀[1]，因此 EIF3F 發生斷裂進而造成轉位，使得該基因無法正常表現造成胰臟癌。圖 4.1、4.2 分別為基因 EIF3F 及 LOC339799 斷裂點的分布。

另外我們發現某些基因特別容易與別的基因相互連接，包含了 FAM60A、SCYL1、COL1A1、HNRNPC、HSP90AA1、IGF2、NFIC、UBB、LOC728658、LOC100008588 等等。

(二) 刪除

總共有 1834 個已知基因被尋找出來，在這僅列出前六筆發生刪除事件的 EST 筆數資料(表 4.3)，基因 MRFAP1 所位於的位置 4p16.1 已被證實該位置發生刪除會造成 Wolf-Hirschhorn 症候群[8]。其臨床特徵包括生長遲緩、智能障礙、心智發育遲緩、小腦症、顏面畸形、骨骼發育異常等等。而基因 OR4F16 位於 1p36.3 處，有關該位置的研究指出 Monosomy 1p36 deletion(或稱 1p36 deletion)[3]是指單一條的第一號染色體，在短臂末端 36 的位置上發生了缺失，其缺失的位置可能從 1p36.13 到 1p36.33 不等。臨床上稱為 1p36 缺失症候群，為目前最常見的先天染色體缺失的症候群之一。發生率約為 1/5000，男女比例相等。1p36 缺失症候群的患者，常出現中度至重度不等的身心發展遲緩、視力與聽覺上的障礙，及外觀上的異常，而患者的臨床表徵的嚴重度，將因缺失的情形不同而有程度之別。圖 4.3 為基因 MRFAP1 斷裂點的分布。

表 4.2 發生轉位的基因次數

基因	EST 發生轉位筆數
EIF3F	205
LOC339799	166
PDE3A	100
LOC729708	74
TPI1	68
CRYGD	62

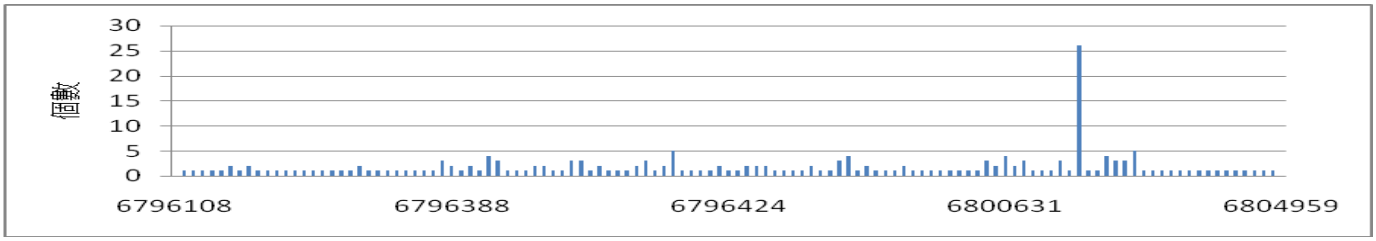

圖 4.1 基因 EIF3F 斷裂點分布圖

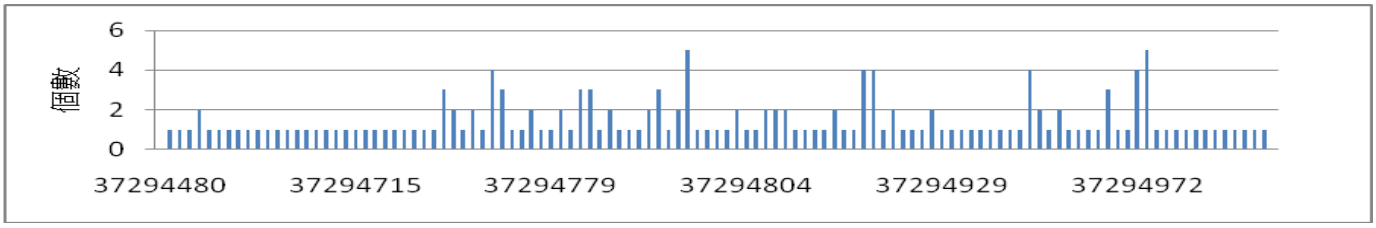

圖 4.2 基因 LOC339799 斷裂點分布圖

表 4.3 發生刪除的基因次數

基因	EST 發生刪除筆數
MRFAP1	417
OR4F16	112
HSPB1	98
MRFAP1L1	86
RANBP5	73
C1orf144	70

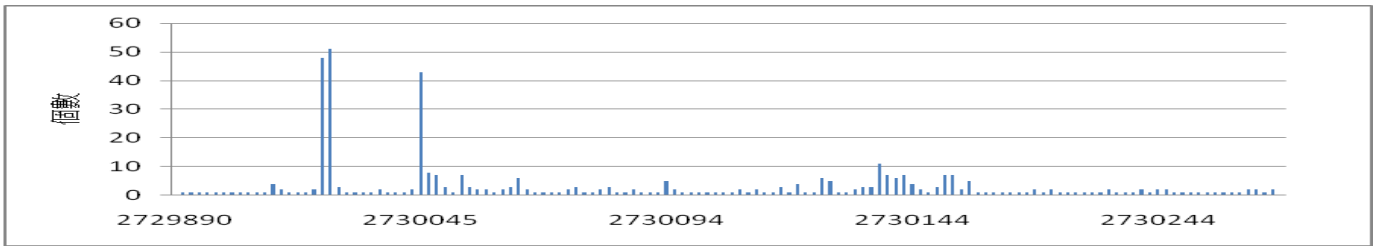

圖 4.3 基因 MRFAP1 斷裂點分布圖

表 4.4 發生反轉的基因次數

基因	EST 發生反轉筆數
HSF1	159
NPEPPS	70
LOC440434	67
ARMC9	36

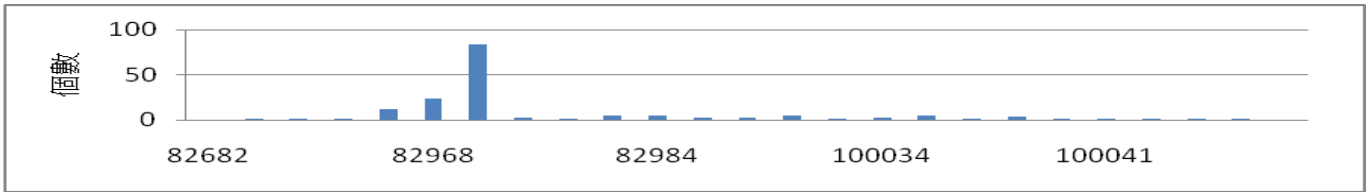

圖 4.4 基因 HSF1 斷裂點分布圖

(三) 反轉

總共有 902 個已知基因被尋找出來，在這僅列出前四筆發生反轉事件的 EST 筆數資料(表 4.4)，基因 HSF1 位於 8q24.3，該位置在 M5 型急性骨隨性白血病患中發現有反轉的現象[7]，圖 4.4 為基因 HSF1 斷裂點的分布，可看出斷裂處容易發生 82970 處。

五、結論

TICdb 是蒐集與癌症相關的轉位事件，而本篇所找尋出來的現象不僅是會造成不同型式的癌症也會造成其他各種疾病包涵中度至重度不等的身心發展遲緩、視力與聽覺上的障礙，及外觀上的異常的現象。總資料量約一萬五千多筆，其中轉位、刪除、反轉事件分別有五千多筆、九千多筆、以及一千多筆。

六、參考文獻

[1] Adriana Doldan, Anupama Chandramouli, Renee´ Shanas, Achyut Bhattacharyya, John T. Cunningham, Mark A. Nelson, and Jiaqi Shi, “Loss of the Eukaryotic Initiation Factor 3f in Pancreatic Cancer”, MOLECULAR CARCINOGENESIS, 2008.

[2] Allison A Burrow, Laura E Williams, Levi CT Pierce and Yuh-Hwa Wang, “Over half of breakpoints in gene pairs involved in cancer-specific recurrent translocations are mapped to human chromosomal fragile sites”, BMC Genomics,2009.

[3] Doctor Anne Slavotinek , “Chromosome 1p36 deletions”, Orphanet encyclopedia, 2003.

[4] Francisco J Novo, Iñigo Ortiz de Mendíbil and José L Vizmanos, “TICdb: a collection of gene-mapped translocation breakpoints in cancer”, BMC Genomics, 2007.

[5] F.R. Hsu and J.F. Chen, “Alignment ESTs to Genome Using Multi-Layer Unique Markers”, IEEE Computer Society Bioinformatics (CSB’03),p.564

[6] J.D Rowley, “A new consistent chromosome abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining”, Nature ,243(1973) ,290-293

[7] Max Chaffanet , Marie-Joëlle Mozziconacci , Francisca Fernandez , Danielle Sainty , Marina Lafage-Pochitaloff, Daniel Birnbaum , Marie-Josèphe Pébusque , “ A case of inv(8)(p11q24) associated with acute myeloid leukemia involves the MOZ and CBP genes in a masked t(8;16)”, Genes, Chromosomes and Cancer 26,161-165(1999).

[8] N. Simon Thomas, Miranda Durkie, Berendine Van Zyl, Richard Sanford, Gemma Potts, Sheila Youings, Nicholas Dennis and Patricia Jacobs “Parental and chromosomal origin of unbalanced de novo structural chromosome abnormalities in man”, Human Genetics,444-450,2006.

[9] P.C Nowell, D.A. Hungerford, “ A minute chromosome in human chronic granulocytic leckemia”, Science, 132(1960),1497.

[10] Takema Kato, Hidehito Inagaki, Hiroshi Kogo, Tamae Ohye, Kouji Yamada,Beverly S. Emanuel and Hiroki Kurahashi, “Two different forms of palindrome resolution in the human genome: deletion or translocation”, Human Molecular

Genetics, 1184–1191,2008.
[11] Thomas D. Wu and Colin K. Watanabe, “GMAP: a genomic mapping and alignment program for mRNA and EST sequences”, Bioinformatics,1859-1875,2005.