

從次世代定序資料探索可能的基因選擇性剪接事件

楊舜安, 鐘文鈺*

國立高雄應用科技大學

資訊工程系

wychung@kuas.edu.tw

摘要

次世代定序技術(Next-Generation Sequencing Technology)產生出高通量的分子序列資料, 而如何快速且有效的分析這龐大資訊則成為計算上的一項新課題。我們在此以預測基因表達量為研究問題。在基因轉錄的過程中, 多個外顯子(exons)可藉由選擇性剪接(alternative splicing), 組合成不同的信使核糖核酸(mRNA), 進而轉譯成序列相異、功能相異的蛋白質。目前雖然已有軟體可進行選擇性剪接的預測, 但多數只提出單一結果。假設從同一mRNA定序得出的短讀取序列值應該相同或十分接近。因此在大量的序列資料中, 以容錯的角度, 從各基因外顯子的讀取數與各種外顯子組合的可能性, 提出評比結果, 並推論出最符合真實情況的組合方式。最後我們用公開的定序資料來測試方法的可行性與實用度。我們的方法將提供一個整體性的測試結果, 希望對基因表達能有更深的了解。

1 緒論

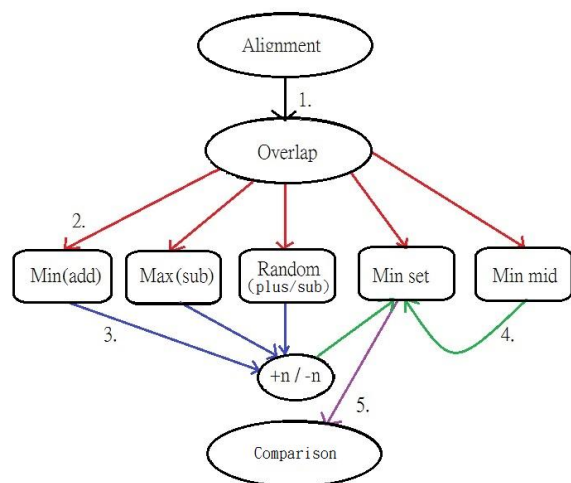
1984年所提出的人類基因體計畫(Human Genome Project, HGP)於1990~2005年間花費30億美元完成人類基因體定序的工作[1]。然而定序的結果只解讀了DNA字串, 尚未能全盤了解基因功能與基因組合的可能。因此目前生物資訊的研究重點之一就在探討基因在不同細胞組織中的表達差異。在人類細胞裡選擇性剪接(Alternative Splicing, AS)是基因轉錄與表達的一個重要步驟。基因中內含子(intron)在基因轉錄成 pre-mRNA 後會被剪接移除, 外顯子(exon)會連結形成mRNA, 爾後再進一步轉譯成蛋白質(protein)。選擇性剪接是指多個外顯子可以產生出不同的組合, 因此最後在轉譯成蛋白質時可能有結構或功能的不同[2]。

同一基因在不同時間, 不同環境下可經由選擇性剪接製造出不同功能的蛋白質。如果能直接從實驗數據中找出可能的不同組合, 讓研究者能夠從機率上掌握到 mRNA表達量的多寡, 從而發覺細胞裡不同的變異情況, 並預測其功能。然而, 要能正確判斷出剪接後產生的正確組合, 是相當困難的事情。本研究自訂數種產生組合的方法, 從中來找出各個可能性, 進而提出最符合真實情況的選擇性剪接事件。

選擇性剪接(AS)是一個從 pre-mRNA 轉錄為 mRNA的過程[3]。因此量化每個外顯子的包含機率一向是推測AS的研究重點。Tophat[4]與SpliceTrap[5]分別提出了不同的技術來研究選擇性剪接可能性。Tophat剪接短讀取序列(short reads)並且標定出在剪接事件中的供體位點(donor site)和受體位點(acceptor site)。通過比較供體位點和受體位點的組合來識別選擇性剪接事件。進一步通過對供體和受體位點的reads計數, 再結合外顯子其他區域的reads數據, 計算選擇性剪接事件之間的比例。SpliceTrap則是量化基因中每個外顯子, 並以兩個或三個外顯子為一集合以動態的方式, 來研究選擇性剪接調控。

2 研究方法

研究方法主要包含三大部分: 實驗前處理、重疊判斷與容錯計數的輸出。最後我們比較各種可能性出現的機率。實驗流程圖由圖一表示。實驗前處理首先是序列比對(sequence alignment)。所謂序列比對指將兩個或多個序列排列在一起, 標明其相似之處[3]。本實驗使用Bowtie[4]得出基因表達時各短序列的位置。接著做重疊判斷。若重疊超過某一特定的百分比, 則表示此外顯子有表達, 且將表達值累加。



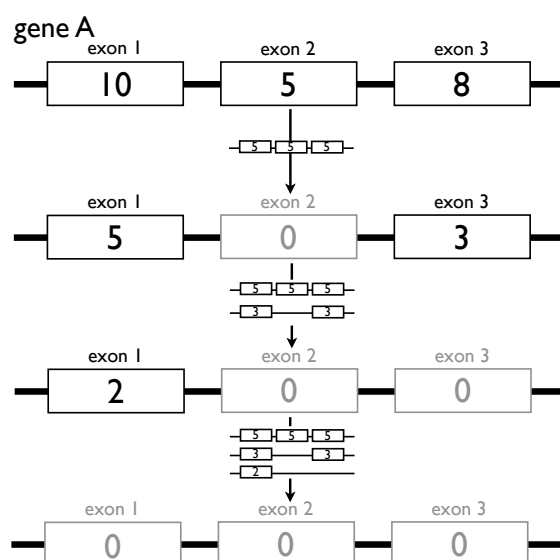
圖一、流程圖。

最後將與外顯子與重疊的短序列，來做組合的分析。在此是以實現選擇性剪接的觀念，提出六種方法，分別為1. 最小外顯子集合(Minimal), 2. 分割基因(Minimal middle), 3. 最大外顯子遞減(Maximal subtraction), 4. 最小外顯子遞增(Minimal addition), 5. 隨機外顯子遞增(Random plus), 6. 隨機外顯子遞減(Random minus)。其中第2到第6種方法分別計算後，在以最小外顯子集合的假設進行預測。

首先，最小外顯子集合(圖二)是從基因中找到最少短序列讀取的外顯子，將基因中每個外顯子讀取數減去此外顯子的值，表示在該基因中有一組選擇性剪接包含所有外顯子，且每個外顯子均擁有此最小短序列讀取值。重複此步驟，直到此基因沒有任何剩餘的表達值。在此包含不同最小短序列讀取值所有的組合，皆為可能的異構體(isoform)。因為我們假設在該基因所有剪接異構體中，包含最多外顯子的組合是最有可能的情況。所以以找出最小短序列讀取為優先選擇，讓一選擇性剪接的可能性包含所有的外顯子。因為抓取到最小短序列讀取外顯子在全域時與局部時會有所不同，所以我們分割基因，從基因中間分割為兩部分，再分別執行最小外顯子集合方法；從全域與局部的方式去探討外顯子的可能性。在此我們再應用最小外顯子集合來推斷可能的結果。另外，最大外顯子遞減、最小外顯子遞增則是從基因中各外顯子找最大與最小的短序列讀取外顯子去做遞減與遞增的運算(此為定序或比對產生的錯誤)後，再分別執行最小外顯子集合的方

法。由於基因中最大與最小短序列讀取數，各別代表每個基因裡外顯子最大與最小的表達量，假設基因的每個外顯子表達量應該趨向一致(平均值)，此極值為最有可能產生錯誤，即最大讀取數可能比現值更小，而最小讀取數可能比現值更大。因此我們對相對應的數值做遞減或遞增，藉此相互比較再推斷找出可能的結果。最後隨機外顯子遞增、隨機外顯子遞減方式，是從基因中隨機挑選外顯子分別去執行遞增與遞減，接著執行最小外顯子集合的方法。由於Maximal subtraction與Minimal addition分別只對最大及最小的短序列讀取外顯子運算，但任一外顯子的讀取值也有可能不正確，於是以隨機的方式來對外顯子做運算後，加強不同於Maximal subtraction與Minimal addition的非固定性，藉此來找出差異。在做完最大外顯子遞減、最小外顯子遞增、隨機外顯子遞增、隨機外顯子遞減四種方法的運算後，再分別做最小外顯子集合來判斷可能的組合。

最後是將這六種方法產生出的所有可能性，計算出不同組合在每次步驟出現的機率，算出所有組合所產生的次數以及在六種方法所出現的機率。若同一組合在不同方法下重複出現，即機率較高，則我們相信此組合為符合資料的可能性相對較高，因此該組合在選擇性剪接中為最符合現實情況的方式。



圖二、以最小外顯子集合方式來預測選擇性剪接事件。方框為外顯子(exons)，其中的數字為比對後的短序列讀取數。以最少的短序列讀取數為基準點，逐步找出可能的的外顯子組合。

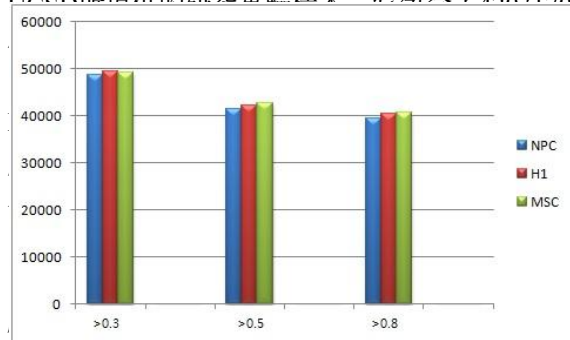
3 實驗結果

首先使用由已定序完成的人類第22號染色體[6]，神經前導細胞(Derived Neuronal Progenitor Cultured Cells)，胚胎幹細胞(H1 Cell Line)與間葉幹細胞(Mesenchymal Stem Cells)[6]來做此研究方法的測試。樣本分別如下：

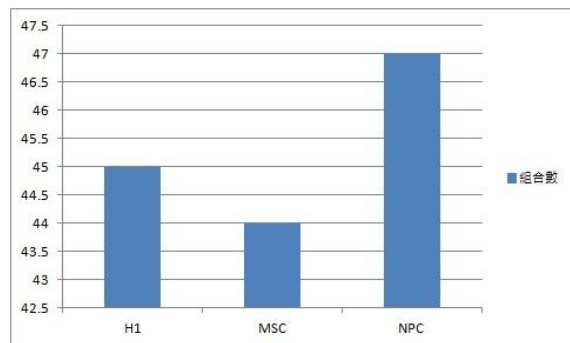
hg19_chr22 這組資料為人類第22號染色體，是由UCSC所提供已定序的數據。它包含了Gene ID與17817個外顯子位置。

Derived Neuronal Progenitor Cultured Cells 這組資料為神經前導細胞，是由UCSD所提供的研究實驗樣本，它包含了5937373個短序列讀取。

H1 Cell_Line 這組資料為胚胎幹細胞，同樣由UCSD所提供的研究實驗樣本，它包含了1002170



出三組數據表示各基因中外顯子包含的短序列讀取數，爾後將對此去做選擇性剪接的分析出可能的組合以及出現的機率。



圖三、由實驗數據得知在神經前導細胞，胚胎幹細胞，間葉幹細胞中，擁有最多組合的基因為NM_058004。以上為該基因在三種細胞中的可能選擇性剪接的異構體個數。

平均而言，在胚胎幹細胞中，每個基因有13.2個異構體；在神經前導細胞與間葉幹細胞中，此數字分別為10.4與12.8。間葉幹細胞較神經前導細胞更接近原始的胚胎幹細胞，其分化的時間比較短。從異構體個數來觀察，似乎也符合生物實驗結果。

神經前導細胞	
NM_001008496(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21)	0.666667
NM_001008496(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20)	0.166667
NM_001008496(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21)	0.166667
胚胎幹細胞	
NM_001008496(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44)	0.333333
NM_001008496(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21)	0.166667
NM_001008496(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21)	0.166667
NM_001008496(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21)	0.166667
NM_001008496(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21)	0.166667
間葉幹細胞	
NM_001008496(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21)	0.666667
NM_001008496(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21)	0.333333
NM_001008496(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21)	0.166667
NM_001008496(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21)	0.166667
NM_001008496(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21)	0.166667
NM_001008496(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21)	0.166667

圖四、分別為神經前導細胞，胚胎幹細胞，間葉幹細胞所產生的外顯子組合以及組合出現機率。第一欄為RefSeq ID (NM_XXX); 第二欄為選擇性剪接的外顯子組合，每個外顯子依照其在基因體上的位置編號; 最後一個欄位為出現機率。

4 結論

本研究首先利用最大/最小序列讀取數的外顯子做遞減/遞增運算後判斷出現的組合，從中假設找到外顯子所包含序列讀取數些微的誤差。再接著加入隨機外顯子做遞增/遞減後加強結果的可靠性。最後嘗試最小序列讀取數的外顯子，利用分割基因以及最大/最小序列讀取數、隨機外顯子做遞增/遞減產生的方式來產生基因中出現各種組合。從多種運算結果去找出各組合的機率，假設較高機率的組合在基因中出現的可能會是較高的。由於在龐大的實驗數據中尚未做運算與判斷時，我們並不清楚哪段外顯子組合是較有可能形成有表現的基因，於是採用了此方法來做運算及判斷，藉此讓數據能提供我們有效的資訊。

之後我們將把實驗數據擴大到人類全基因組，然後與已知的alternative splicing forms做比較。

未來我們希望能由找出各外顯子關聯性，來判斷組合所產生的機率，藉此讓各數據更為精確，數據包含做完運算的短序列讀取數、外顯子的組合、所產生組合的機率，讓研究人員能更清楚知

道不同實驗細胞每個基因所產生的組合中的差異性。另外在實驗中以更多變數來紀錄更多不同的數據，包括實驗次數、最短序列讀取數等實驗相關數據，藉此能使研究更有準確的判斷。另外，由於本實驗提供了大量的數據結果來提供研究人員參考，若能在研究中加入圖形化的辨識，如AGSG[7]與TopHat以圖形化的方式來呈現給研究人員，相信在視覺上對基因的判斷能更加清楚，更能提昇實驗上的收獲。

References

- [1] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome, *Nature*, 2001.
- [2] Benjamin Lewin. Genes IX. *Jones and Bartlett Publishers*, 2008.
- [3] Elaine R. Mardis. The impact of next-generation sequencing technology on genetics, *Cell*, 2008.
- [4] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nature Protocols*, 2012.
- [5] Jie Wu, Martin Akerman, Shuying Sun, W. Richard McCombie, Adrian R. Krainer and Michael Q. Zhang. SpliceTrap: a method to quantify alternative splicing under single cellular conditions, *Bioinformatics*, 2011.
- [6] Wei Xie, Matthew D. Schultz, Ryan Lister, Zhonggang Hou, Nisha Rajagopa, Pradipta Ray, John W. Whitaker, Shulan Tian, R. David Hawkins, Danny Leung and Michael Q. Zhang. Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells, *Cell*, 2013.
- [7] Durgaprasad Bollina, Bernett T. K. Lee, Tin Wee Tan and Shoba Ranganathan. ASGS: an alternative splicing graph web service. *Nucleic Acids Research*, 2006.