應用資料探勘探討轉錄因子與選擇性剪切的潛在模式

吳旻昇(醫工 F94036089)、林子傑(生科 C54049032)

摘要

選擇性剪切 (alternative splicing)是真核生物基因表達非常重要的一環，透過組合不同的外顯子 (exon) 以及內含子 (intron) 形成不同的 mRNA 並轉譯 (translate) 成不同的蛋白質。過去的的研究多專注於探討位於基因內的元件 (e.g. exon, intron) 如何影響選擇性剪切，而較少研究一些距離選擇性剪切事件較遠的元件 (i.e. 啟動子) ，因此我們在這個專題是想要了解一些基因外的元件是否可以影響選擇性剪切的模式，我們使用了深度學習 (Deep Learning) 與基於決策樹的 XGBoost 來尋找啟動子上的轉錄因子與選擇性剪切事件的關聯並進行預測，並且發現了一些潛在的關鍵因子。

背景與動機

在人類基因體計劃 (Human Genome Project) 完全定序完成後，生物學家發現人類的基因數量遠比之前猜測的少，蛋白編碼基因 (protein-coding gene) 僅有不到 2 萬個，但是人類的蛋白多樣性遠遠高於基因的數量，這引起了大家對於選擇性剪切 (Alternative splicing) 的興趣，選擇性剪切可 以從一個基因產生多種的蛋白，而不同的蛋白可以執行不同的功能以應對生物所面臨的不同環境，因此瞭解生物是如何調控選擇性剪切是一個相當重要的問題。

過去的研究大多只關注在內含子與外含子上面的一些元件，對於距離基因較遠的元件 (e.g. promoter, enhancer) 較少關注，因此我們想要瞭解這些元件是否對於選擇性剪切的模式 (alternative splicing pattern) 有造成影響，以及不同元件對於其的貢獻強度。在這個專案中主要是想要探啟動子上的轉錄因子結合對於選擇性剪切的影響，是不是存在一群轉錄因子其在啟動子的結合，就會造成選擇性剪切模式的改變。

因此我們結合了 2 種不同的次世代定序嘗試來回答這個問題，DNA 平時受到組蛋白以及其他蛋白的包裹，平時不會裸露出來，而受到轉錄因子結合的區域則會暴露出來，透過 FAIRE-Seq 我們可以得到這些裸露的區域 (chromatin-accessible region)的資訊，並且我們可以透過 matrix-scan 來辨識裸露區域中的轉錄因子結合位點。另外在選

擇性剪切的部分我們透過 RNA-Seq 的資料來瞭解一個基因其所製造的 mRNA 是屬於那一種剪切模式，而資料主要是來自於 NIH GEO 資料庫。

研究方法

資料品質過濾 (QC)

因為次世代定序產生的序列資料相較於傳統定序準確度更低，所以我們必須去除掉低品質的序列資料，避免錯誤的定序資料對後續分析的影響。我們使用 Trimmomatic 來對原始序列 (Reads) 資料進行序列資料檢查與篩選，Trimmomatic 是針對 illumina 定序平臺開發的工具[1]，除了可以進行序列品質檢查之外也能去除在殘留在 Reads 中的引子片段，這可以避免在後續進行 Mapping 時的 mis-alignment。

針對 Single End 定序使用一下參數：`SE -phred33 input.fq.gz output.fq.gz ILLUMINACLIP:TruSeq3-SE:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36`
針對 Paired End 定序使用一下參數：`PE -phred33 input_forward.fq.gz input_reverse.fq.gz output_forward_paired.fq.gz output_forward_unpaired.fq.gz output_reverse_paired.fq.gz output_reverse_unpaired.fq.gz ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36`

RNA-Seq 資料預處理

在 QC 之後我們將 RNA-Seq 所得到的 Reads 比對 (mapping) 回果蠅的基因組序列，由於次世代定序所產生的資料相當龐大，我們使用 HISAT2 來進行 mapping，HISAT2 使用 BWT 轉換 (Burrows-Wheeler transform) 以及 FM 索引 (Ferragina-Manzini index) 來加速 mapping 的速度[2]，並且同時處理 RNA 的剪接造成的 Junction Reads。在 mapping 回果蠅的基因組後，使用 MISO 來辨識不同的剪切模式佔整體 RNA transcript 的比例，MISO 使用了 Mixture Model 來提升估計的準確性，並計算出 PSI 值以供後續分析[3]，PSI 值表示細胞傾向於製造哪一種的剪切，因為其值為一比例因此介於 0~1 之間。

FAIRE-Seq 資料預處理

在 QC 之後我們同樣將 FAIRE-Seq 所得到的 Reads 比對 (mapping) 回果蠅的基因組序列，因為 FAIRE-Seq 是 DNA 序列，不會產生 Junction Reads，因此我們使用 Bowtie2 來進行 mapping，Bowtie2 何 HISAT2 一樣使用了 BWT 轉換來加速 mapping 的速度[4]，在 mapping 之後我們利用 MACS2 來辨識 DNA 上的裸露區域[5]，因為 FAIRE-Seq 只能知道轉錄因子粗略的結合區域，我們使用 RSAT 工具中的 matrix-scan 搭配 CIS-BP 的 motif 資料庫來找出轉錄因子精確的結合位點[6, 7]。

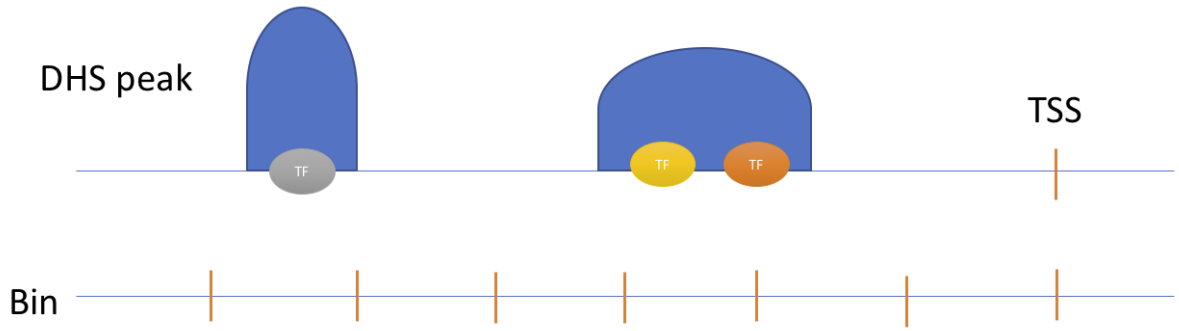
資料探勘

我們是用了兩種不同的方法分別來探討 TF 與選擇性剪接的關係，第一個我們使用了深度學習，深度學習是近年來相當熱門的方法，因為它強大的模型表示力，所以在大多數的問題中都取得了非常好的預測效果，我們分別使用了 Fully-Connected DNN (Deep Neural Networks) 與 CNN (Convolutional Neural Networks) 兩種不同的網路架構。但是因為為們不只是想要預測而已，我們還希望知道那些 TF 是重要的，可能會調控基因的選擇性剪切，我們另外使用基於決策樹的 XGBoost 來訓練模型，因此可以透過每個特徵 (TF) 的資訊增益 (Information Gain) 來瞭解 TF 的重要性，並且給於我們一些因果關係的 insight，在模型中我們使用 FAIRE-Seq 的 TF 結合資訊作為特徵，RNA-Seq 的 PSI 值作為反應變數 (response variable)。

在 CNN 網路中我們使用了一個 1D Convolution Layer (filter size = 7, filter num = 128)，並在 Flatten 之後接兩層 Fully-Connected Layer (size = 128, 256)，最後接 SoftMax 的 Output Layer。在 DNN 的部分我們使用了四層的 Fully-Connected Layer (size = 128, 256, 512, 1024)，後接 SoftMax 進行輸出。在兩個網路中我們使用 ELU 作為激活函數並且搭配 Batch Normalization (BN) 避免死掉的神經元，為了避免過度擬合我們使用 Dropout 來隨機關閉一些神經元，Dropout 的比例在 CNN 以及 DNN 前兩層為 0.5，DNN 後兩層為 0.25，除了 Dropout 我們還使用了 L2 正規化來防止模型過擬合，在訓練模型時是用 Adam 優化器。在 XGBoost 中，我們使用了 500 個 estimator 以及限制決策樹的深度為 sqrt(n_TF) 來避免過度擬合，最後利用資訊增益來獲取 TF 的 Feature importance。

特徵萃取

因為我們要研究的啟動子上的轉錄引子 (TF) 與選擇性剪切的關係，所以我們定義啟動子的範圍為轉錄起始點的上游 2000 bp 到下游的 500 bp，因此特徵資料的格式是一個 2500 * n_TF 的矩陣 (n_TF = 307)，每一個 row 代表該 TF 在啟動子上的結合狀態，1 代表該位置有 TF 結合，0 則表示沒有，特徵的 Layout 如圖一所示。



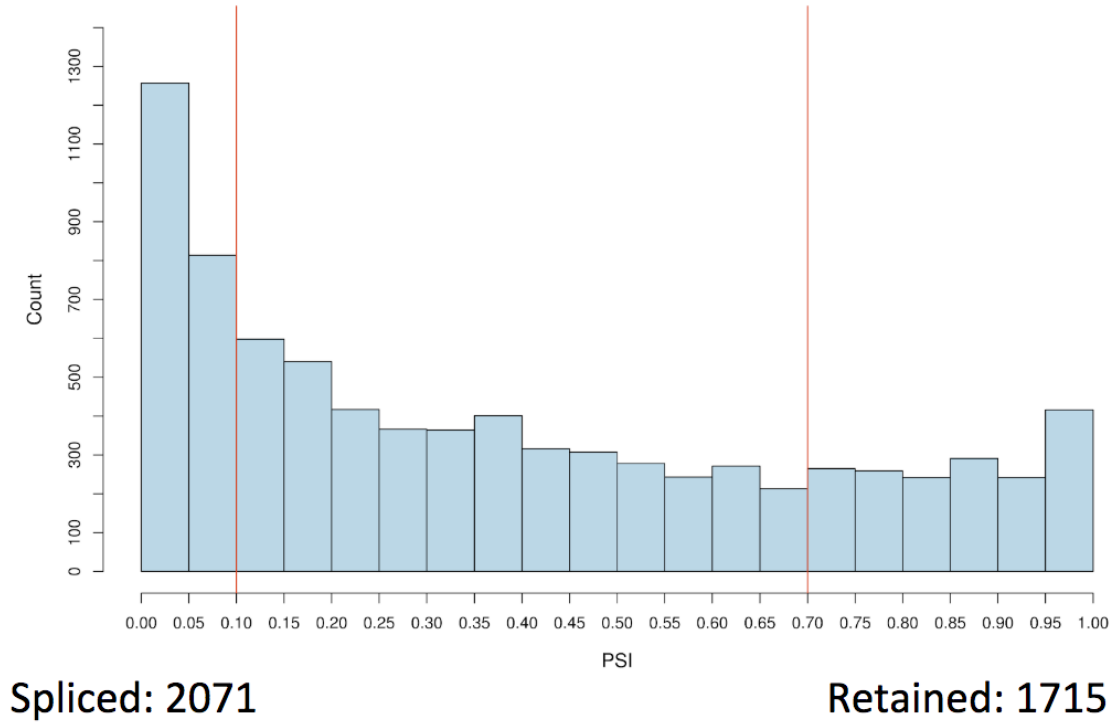
TF1	0	1	0	0	0	0	0	0
TF2	0	0	0	0	1	0	0	0
TF3	0	0	0	0	1	1	0	0
TF4	0	0	0	0	0	0	0	0

圖一、特徵的 Layout

另外在 XGBoost 的部分，因為我們只對 TF 的重要程度感到興趣，我們去除掉了位置上的資料，只保留 TF 是否結合的資訊，如此一來可以減少訓練所需要的記憶體消耗以及時間。

PSI 值的二元化

由於我們有興趣的是 TF 有沒有影響選擇性剪切，進而影響 PSI 值，因此準確預測 PSI 值並不是我們的目標，因此我們將 PSI 分成兩個類別 H/L 分別表示，傾向於剪切 Splicing/保留 Retained，如此也比較方便我們解釋結果。為了避免訓練資料不平衡，而導致兩個類別的預測能力有所落差，我們觀察了 PSI 值的分佈後以> 0.7 為 H 類別，< 0.1 為 L 類別，以達到兩個類別的訓練資料數量相近。圖二為訓練資料 PSI 值的分佈狀況。



圖二、PSI 值的分佈

交叉驗證

為了瞭解模型是否有效，以及是否有過擬合的現象發生，我們對模型進行 K-fold cross-validation，在 DNN 以及 CNN 中我們進行 10-fold CV，在 XGBoost 中使用 5-fold CV 來驗證模型。

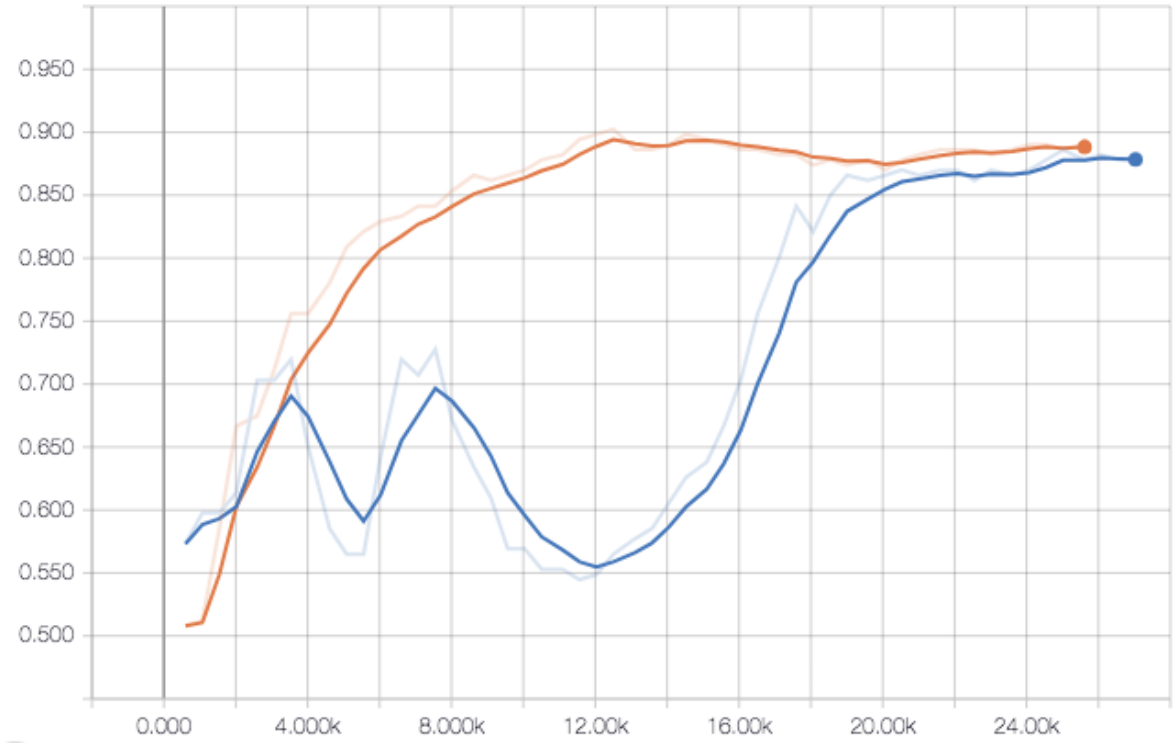
分析結果

資料蒐集

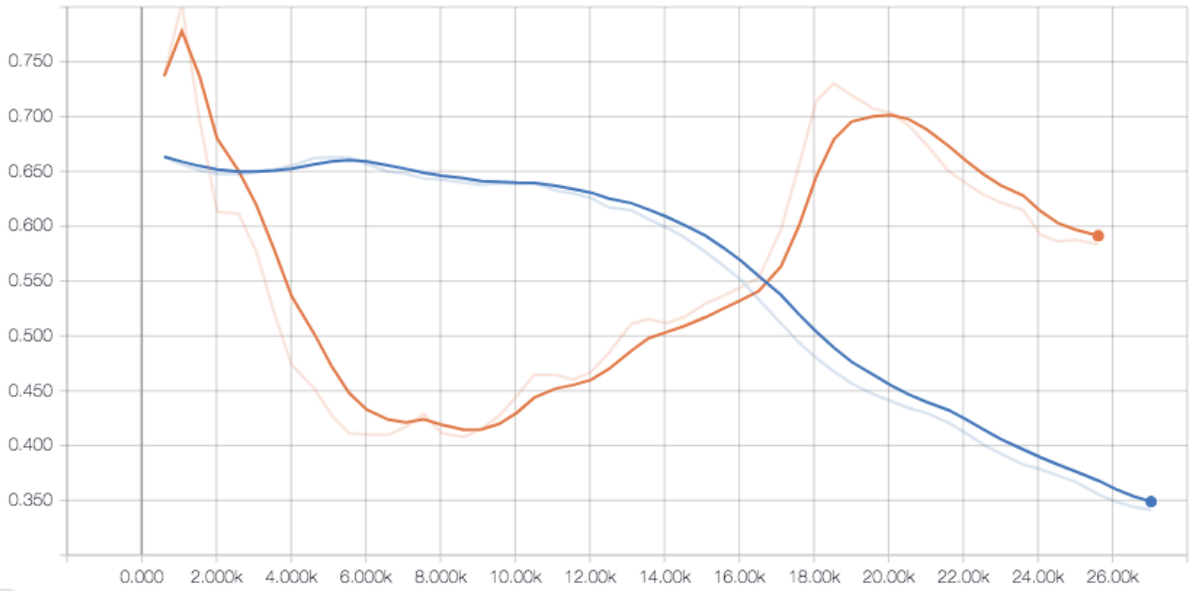
我們主要資料來源是 NIH NCBI GEO 資料庫，其 Data Accession 是 GSE38727、GSE40739、GSE62558 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38727, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40739, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62558)[8-10]，使用的基因組序列版本為 BDGP6.90，經過 QC 與資料前處理、PSI 二元化後，總共剩餘 3786 筆資料，其中 Spliced (PSI L)為 2071 筆，Retained (PSI H)為 1715 筆。

深度學習結果

在深度學習中我們將資料集分成，2832 訓練、246 驗證、708 測試，在兩個模型的訓練結果 DNN 表現的較 CNN 更好，從訓練過程來看 DNN 最終達到 9 成的準確率，而 CNN 只有約 87%，在 Cross-Entropy Lost 的部分也是 DNN 較 CNN 小，DNN 約只有 CNN 的一半，顯見在這個問題中 DNN 應該是較好的模型。

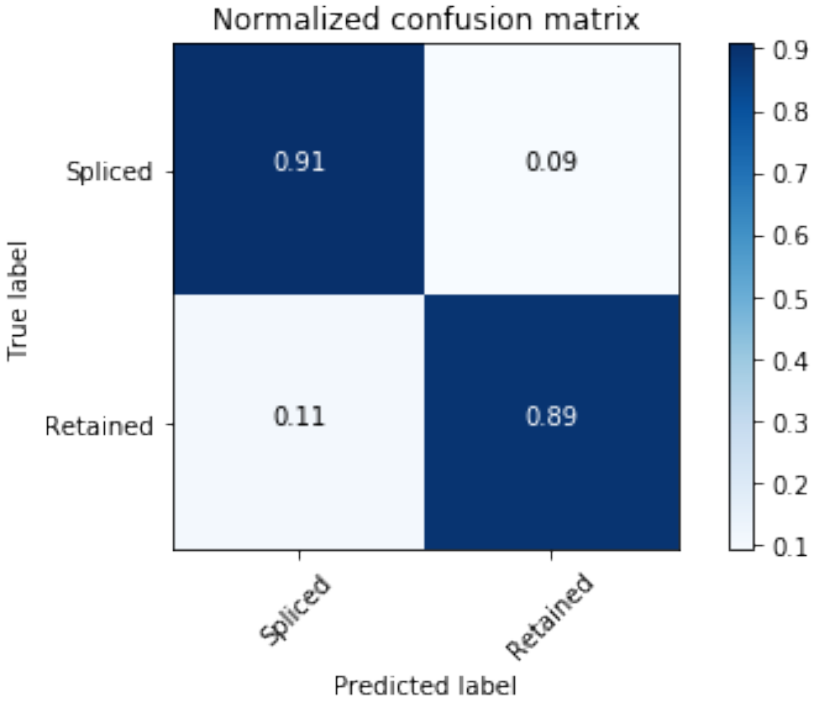


圖三、訓練過程驗證集的準確率變化

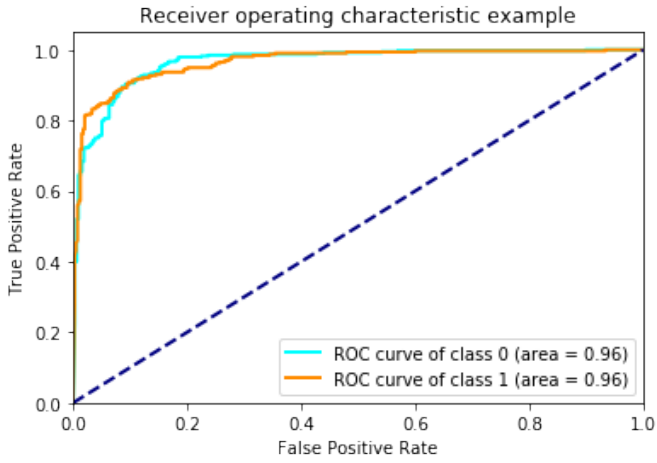


圖四、訓練過程中驗證集的 Loss 變化

因為 DNN 的表現較 CNN 更好，所以之後的模型評估就以 DNN 為主，在混淆矩陣中，Spliced/Retained 兩個類別分別有 0.91 與 0.89 的 recall 率，並沒有明顯的不平衡現象，因此在挑選 PSI 的二元化條件是有用的。DNN 模型的 Recall rate 達到 90%，並且 Precision 也達到 89.9%，顯示我們的模型具有良好的預測能力。我們也計算 AUROC 來評估模型的 Sensitivity 與 Specificity，由圖五發現兩個類別的 ROC 都達到 0.96，顯示我們的模型能夠兼顧 Sensitivity 與 Specificity。最後在交叉驗證中 DNN 模型的平均準確率 90.16 +- 0.96%，模型 Loss 為 0.631 +- 0.104，與前面的測試想去不遠顯見我們的模型是有效的。



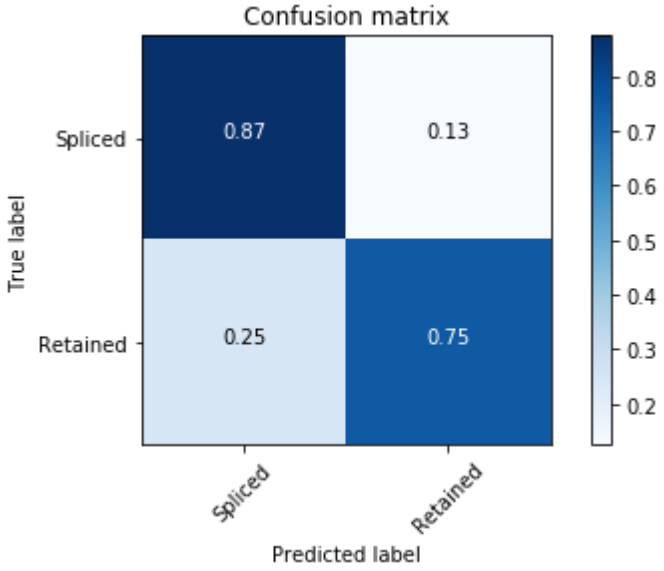
圖五、DNN 模型在測試集上的混淆矩陣



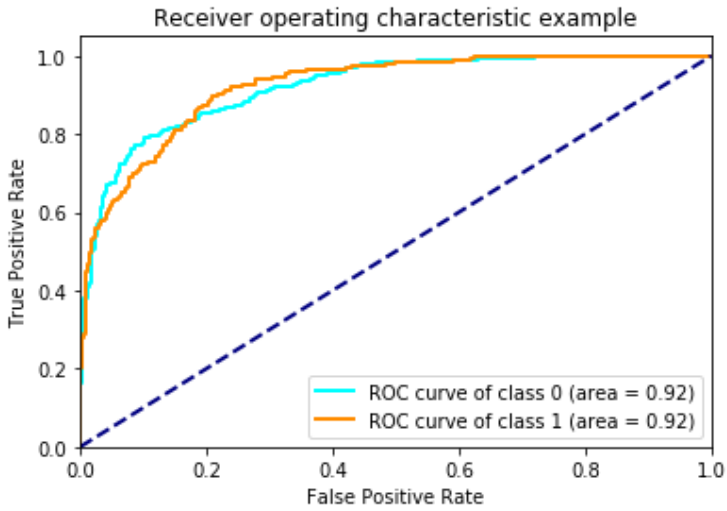
圖六、DNN 模型的 ROC 曲線

XGBoost 模型結果

為了能夠理解 TF 與選擇性剪切之間的關係，我們利用 XGBoost 來建構模型並且利用 Feature Importance 的方法，來找出影響選擇性剪切的重要 Feature (TF)。在訓練之前資料依 0.8/0.2 的比例進行分割，切成訓練集以及測試集，在訓練後的 XGBoost 模型中，準確率到達 81.66%，效果略遜於深度學習，其混淆矩陣如圖七所示，其不平衡的問題相較深度學習更為嚴重，但是仍在可以接受的範圍，其 AUROC 也有 0.92，可見 XGBoost 的模型能夠兼顧 Sensitivity 與 Specificity。為了驗證模型的有效性我們也利用 5-fold CV 來進行驗證，其平均準確率在 83.44 +- 0.93%，因此我們的模型是有效的，並沒有和測試集的測試結果相差太大。

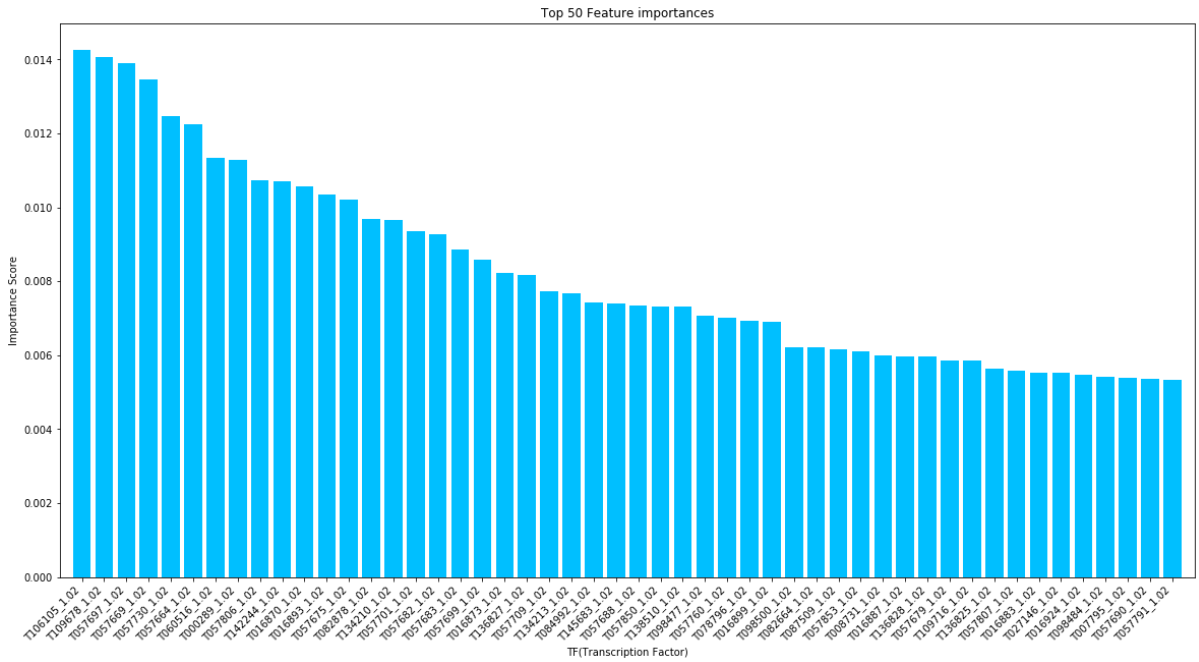


圖七、XGBoost 模型在測試集上的混淆矩陣



圖八、XGBoost 模型的 ROC 曲線以及面積

在 Feature Importance 的部分，為了避免提升 Robustness，避免訓練是 Stochastic，我們取 5 個 Fold 的結果做 Rank 的平均，再做排序挑選出重要的 TF。前 50 個重要 TF 的 Importance Score 與其長條圖如圖九所示。我們挑選 Top 10 的 TF 來進行文獻搜索，發現排名第九的 TF——CTCF (T057806_1)，在過去已經有多篇文獻發現能夠影響選擇性剪切，當 CTCF 結合之後 RNA 聚合酶的合成速度就會放慢，而使得進行剪切的 Spliceosomes 有機會進行作用進而切除 Exon[11]，因此這也增加了 TF 影響選擇性剪切的可信度。



圖九、Top 50 之重要 Feature (TF)

1. T106105_1.02 (0.014262) (Hsf)
2. T109678_1.02 (0.014078) (Adf1)
3. T057697_1.02 (0.013909) (lola)
4. T057669_1.02 (0.013463) (hb)
5. T057730_1.02 (0.012457) (jim)
6. T057664_1.02 (0.012262) (Cf2)
7. T060516_1.02 (0.011338) (rn)
8. T000289_1.02 (0.011273) (br)
9. T057806_1.02 (0.010726) (CTCF, CCCTC-binding factor)
10. T142244_1.02 (0.010704) (Mad)

表一、Top 10 之重要 Feature (TF)及其名稱

結論

從本專題的兩種模型——深度學習以及 XGBoost 中我們發現資料探勘演算法可以從 TF 的結合狀態資訊與選擇性剪切中發現資料的潛在模式，並且能夠良好的預測其行為，這顯示出 TF 的結合確實能夠調控選擇性剪切的模式，我們在 XGBoost 模型的特徵重要度分析中可以也發現一些過去已經被證實有調控選擇性剪切的 TF，這顯示

我們的結果具有一定程度的可信度，能夠和實驗結果互相印證，同時我們也發現沒有少數 Dominate 的 TF 主宰整個選擇性剪切的結果，重要度呈現緩慢下降的趨勢，顯示出 Combinatorial Effect 可能是存在的，一群 TF 一起調控一個基因的選擇性剪切。綜上所述，我們的模型預測了一個潛在的調控選擇性剪切的機制，並且提供生物學家線索來探討調控選擇性剪切的因子。

參考文獻

1. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data**. *Bioinformatics* 2014, **30**(15):2114-2120.
2. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements**. *Nat Methods* 2015, **12**(4):357-360.
3. Katz Y, Wang ET, Airoldi EM, Burge CB: **Analysis and design of RNA sequencing experiments for identifying isoform regulation**. *Nat Methods* 2010, **7**(12):1009-1015.
4. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**(4):357-359.
5. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W *et al*: **Model-based analysis of ChIP-Seq (MACS)**. *Genome Biol* 2008, **9**(9):R137.
6. Turatsinze JV, Thomas-Chollier M, Defrance M, van Helden J: **Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules**. *Nat Protoc* 2008, **3**(10):1578-1588.
7. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K *et al*: **Determination and inference of eukaryotic transcription factor sequence specificity**. *Cell* 2014, **158**(6):1431-1443.
8. McKay DJ, Lieb JD: **A common set of DNA regulatory elements shapes Drosophila appendages**. *Dev Cell* 2013, **27**(3):306-318.
9. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A: **Genome-wide quantitative enhancer activity maps identified by STARR-seq**. *Science* 2013, **339**(6123):1074-1077.
10. Potier D, Davie K, Hulselmans G, Naval Sanchez M, Haagen L, Huynh-Thu VA, Koldere D, Celik A, Geurts P, Christiaens V *et al*: **Mapping gene regulatory networks in Drosophila eye development by large-scale transcriptome perturbations and motif inference**. *Cell Rep* 2014, **9**(6):2290-2303.
11. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S: **CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing**. *Nature* 2011, **479**(7371):74-79.