Methylation of Prostate Cancer Predict

專題編號：113-CSIE-S010

執行期限：112 年第 1 學期至 113 年第 1 學期

指導教授：白敦文

專題參與人員：　110820004　黃新哲

110820051　劉秉逸

110590064　劉韶軒

1、 摘要

本研究旨在利用 DNA 甲基化實現前列腺癌的早期預測。DNA 甲基化是影響基因功能與表現的生物機制，一般認為在癌症的發生中發揮重要作用。

本研究利用癌症基因組圖譜(TCGA)的甲基化數據，開發出一套基於機器學習的特徵篩選系統，用以篩選出潛在的癌症預測基因，以解決前列腺癌的早期檢測準確性與成本問題。

研究最終找到兩個基因(BMP7、CCDC8)。在測試中，能良好的預測前列腺癌的發生，達成研究目標。

2、 緣由與目的

前列腺癌隨著人口老化和高脂肪飲食等因素影響下，已日漸成為男性健康一大議題，然而前列腺癌的診斷通常需要透過觸診或切片診斷，不容易在出現病症前被察覺。

DNA 甲基化是影響基因功能與表現的重要機制。透過分析病人的甲基化數據，我們可以篩選出少量具代表性的基因，用於前列腺癌的預測。從而降低預測成本，有助於前列腺癌預測的普及，提升癌症的早期預警能力。

3、 研究報告內容

一、　資料預處理

首先，從 TCGA 取得前列腺癌甲基化數據，使用 ChAMP 套件進行基礎分析[1]，取得各點位的甲基化數值(Beta 值)

二、　統計分析

(一) 離群值去除

去除離群值並計算 Normal 與 Tumor 樣本在 Beta 值上的差距(ΔBeta)和 p-value。個基因僅保留 ΔBeta 最佳的點位，依 ΔBeta 和 p-value 閾值將基因區分為：

- Hyper (圖 2 紅色點位)與 ΔBeta 正相關
- Hypo (圖 2 藍色點位)與 ΔBeta 負相關

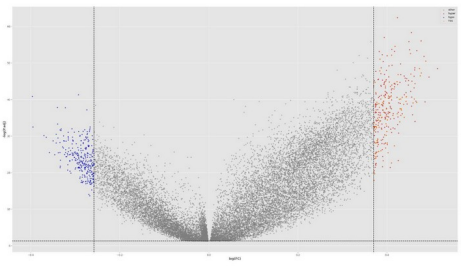

圖 1. ΔBeta 與 p-value 呈現火山圖

(二) 尋找最佳切點

對 hyer、hypo 資料使用混淆矩陣計算最佳切點，以區分 Normal 與 Tumor 樣本。並且利用 F1-score 以及 AUC/ROC 進行測試，移除表現不佳的基因。

三、　模型訓練

利用模型訓練以篩選基因。透過 SMOTE 平衡資料[2]，以解決 Normal 與 Tumor 樣本資料不平衡。

訓練 Random　Forest、XGBoost 分群模型，並調整參數。

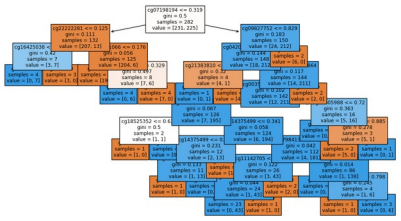

圖 2. Random Forest 首棵樹的視覺化結果

使用 RFECV 篩選模型特徵，並將結果交集，以確保篩選出的基因具有代表性，而非僅適用於某個特定模型。

四、　基因功能性分群

以基因功能性作為分群。使用 GOSemSim 套件的 Wang 方法[2,3]，對候選的基因計算 GO Term 的相似度矩陣，並以 hierarchical clustering 分群。

最終由各群選出基因，相互組合出最佳搭配。

五、　癌症期數分析

將資料以不同期數個別分析，避免基因在早期顯著表現後，受晚期影響而被過早剔除。

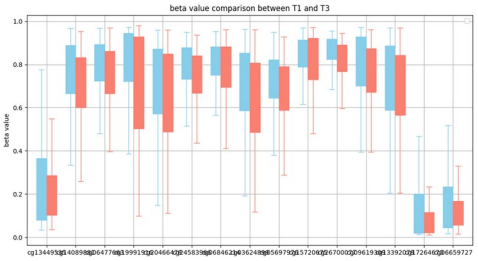

圖 3. 前期(藍)後期(紅)甲基化程度分布比較圖表

以早期資料(一、二期)進行實驗流程，篩選出對癌症早期具代表性的基因。

六、　實驗結果分析

利用外部資料集檢測挑選出來的基因是否具有良好的預測性。在 450k 資料集中使用 3 個基因能達到 90%以上的準確度。在 850k 資料集中使用 3 個基因能達到 75%左右的準確度。PIA 為前列腺的萎縮性病變，研究表示 PIA 可能與早期前列腺癌有因果關系。在含有 PIA 的資料集中使用 3 個基因能達到 74%左右的準確度。

表 1. 外部資料集測試結果

參考文獻

[1] A. B. Smith, C. D. Jones, and E. F. Roberts, “Article Title,” *Journal*, Vol., No., pp. 1-10, Date.

[2] 著者姓名，「中文期刊論文篇名」，**中文期刊名**，卷，期，發行年次，頁次。

https://www.sciencedirect.com/science/article/pii/S0960076016301054

GAPDH:https://link.springer.com/article/10.1007/s12094-012-0924-x，引用 140 次 2013

BMP7:https://www.sciencedirect.com/science/article/pii/S0002944010620342?casa_token=KrdQTPfw8NwAAAAA:oIvL2kGP_s5_MlFRd_Gk9xvFigUJga9WSQkN_hVYYI8KEguQj_NtDcYge_ewRJFUSQ3h_oDqtLE，引用 246 次，007

NR1H3:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5381174/，被引用 16 次 2017

CDH23:https://www.science.org/doi/full/10.1126/sciadv.aaw6710，被引用 35 次 2019

RBP1:https://jcp.bmj.com/content/57/8/872.abstract，被引用 76 次 2004

VIPR2:https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2022.852358/full ， 被 引 用 7 次 2022

	AJCC Prostate Cancer Stage Groupings													
	Stage I		Stage IIa			Stage IIb			III	Stage IV				
Jewett-Whitmore stage	A1		A2, B0-2						C1-3	D1		D2		
TNM stage	T1a-c N0M0	T2a N0M0	T1a-c N0M0	T1a-c N0M0	T2a N0M0	T2a N0M0	T2b N0M0	T2c N0M0	T1-2 N0M0	T1-2 N0M0	T3a-b N0M0	T4 N0M0	Any T N1M0	Any T Any N M1
Gleason score	≤ 6	≤ 6	7	≤ 6	≤ 6	7	≤ 7	Any	Any	≥ 8	Any	Any	Any	Any
PSA level (ng/ml)	< 10	< 10	< 20	10-19.9	10-19.9	< 20	< 20	Any	≥ 20	Any	Any	Any	Any	Any

Dataset	CpG Sites	Percentage
450k_GSE157272(No PIA)	('cg18759209', 'cg03576469', 'cg23740882')	92.86%
	('cg18759209', 'cg03576469', 'cg27223047', 'cg24530250')	89.29%
450k_GSE157272	('cg18759209', 'cg03576469', 'cg23740882')	74.29%
	('cg18759209', 'cg03576469', 'cg27223047', 'cg24530250')	71.43%
450k_GSE47915	('cg18759209', 'cg03576469', 'cg15323528')	100.00%
	('cg18759209', 'cg03576469', 'cg15323528', 'cg08151731')	100.00%
450k_GSE112047	('cg18759209', 'cg03576469', 'cg27223047')	100.00%
	('cg18759209', 'cg03576469', 'cg27223047', 'cg19300568')	100.00%
850k_GSE183040	('cg18759209', 'cg24530250', 'cg15229124')	75.44%
	('cg18759209', 'cg24530250', 'cg15229124', 'cg08378442')	63.16%

七、 結論

我們目前找到兩個較穩定的點位，在測試中有著良好的預測性。

- BMP7：cg18759209
- CCDC8：cg03576469

目前對整體 450k 資料的 F1-score 已經達到 90%。實驗結果以能透過少量基因數據檢測出準確的結果，以此達到降低成本、提高效率的目標。

此外我們發現甲基化數據在癌症的不同階段中表現出不一樣的分布，未來有望能以此為方向，提高早期癌症的預測準確度。

參考文獻

[1]Tian Y., Morris T.J., Webster A.P. et al. (2017). ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* 33, 3982–3984.

[2]Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

[3] Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products.

[4]Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., & Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. Bioinformatics, 23(10), 1274–1281