應用資料融合策略在磷酸化位置的預測

朱彥煒 游景盛 白嘉祥 吳宥廷
亞洲大學生物資訊系 亞洲大學生物資訊系 亞洲大學資訊工程系 亞洲大學生物資訊系
ywchu@asia.edu.tw csyu@asia.edu.tw badletmehigh@hotmail.com davidwu441987@msn.com

摘要

磷酸化作用一直在蛋白質轉譯後修飾中扮演著重要的角色。因為磷酸化是決定許多酵素與受體是否產生作用的關鍵，進而間接影響著生物體內的訊息傳導以及免疫系統等功能。目前雖然隨著各種鑑定技術的發展使得磷酸化資料有明顯的成長，但許多激酶所作用的磷酸化資料仍然不多。因此，磷酸化預測的工作依然顯得相當的重要。

本研究利用 Netphos、Disphos 及 GPS 三個網站的預測資料，提出一個資料融合策略，去預測蛋白質磷酸化的位置。在最後的預測結果測試上，除了三個磷酸化預測網站之外，也加入單純的投票機制一起比較。而在最後的隨選序列測試結果，依據激酶所分的八個類別中，大部份的評估指標都優於其它方法。

關鍵詞: 資料融合、磷酸化、預測、投票機制

一、前　言

蛋白質磷酸化(Phosphorylation)的後修飾作用可以為蛋白質增加功能基、改變化學性質和蛋白質的結構。而且磷酸化和去磷酸化反應掌控了許多酵素(Enzyme)和受體(Receptor)的作用於否，而酵素跟受體又影響著生物體內的訊息傳導、免疫系統以及消化系統等各種功能。因此，磷酸化反應等於間接影響了這些系統與功能。由此可見在各種不同的蛋白質修飾效果中，磷酸化佔了十分重要的地位。

而蛋白質在磷酸化和去磷酸化反應時都需要特別的酵素來幫忙，其分別為蛋白質激酶(protein kinase)和磷酸酶(phosphatase)。磷酸化反應在激酶的幫助下，會在蛋白質序列上的特定位置的胺基酸進行反應，其中有三種因為含有烴基使其可以被磷酸化的氨基酸，其分別為: Serine、Threonine、Tyrosine 分別簡寫成 S、T 和 Y。三種胺基酸之中則以作用在 S 上的磷酸化較為常見，也有同時作用在 S、T 和 Y 的激酶。

磷酸化的預測也已經被應用在醫學研究、生物科技研究等，在這些範疇之中，磷酸化的預測可以讓各種研究或者投資得到更有效的評估方式。

所以在這種情況下，已經有不少研究利用機器學習的方式，以達到預測磷酸化的效果，而在這些研究之中，因為所使用的演算法、學習方式、資料庫的不同，造成了各種不同的預測結果。也產生出不同預測法對於不同激酶的預測準確度各有所長之現象。

而目前最常見的預測方式則為利用 SVM 來達到磷酸化預測之效果[1]。雖然大部分都利用 SVM 來預測，但是其投入之參數亦有所不同，本篇論文

則整合不同預測網站各有所長的預測法，來做出新的預測，以達到 1+1>2 的預測效果。

二、資料集的建立

我們這次研究中所使用的蛋白質磷酸化數據的來源是從 Phospho.ELM Database 取得蛋白質磷酸化修飾的位置建立起樣本的集合[2]，而此數據來源 Phospho.ELM Database 修飾位置的數據都是經由嚴格的生物實驗室驗證而成。我們將其磷酸化資料引用，並整合成一個三階正規化之後的關聯式資料庫，以利於之後運算使用。此資料庫內包含了有 4422 筆磷酸化蛋白質序列以及有發生催化反應激酶的名稱，我們將研究所需要用到之資料從資料庫中濾出，挑出以下條件之序列以供使用:

・在該序列中曾有過至少一個磷酸化位置其激酶屬於 PKA、PKC、CDK 和 CK2 其中一個群組之成員者。

・序列符合待應用之三個網頁所能接受之長度及格式。

而在這些條件篩選之下，我們得到的可用資料有含 PKA 激酶之序列 336 條、PKC 有 257 條，CDK 有 104 條、CK2 有 249 條。而這些序列裡面含有的 S 胺基酸總計有 5968 個和 T 胺基酸總計有 2943 個可供實驗測試及運算。並使用三個磷酸化預測網頁來進行此次的資料融合測試，這三個磷酸化預測網站分別為 GPS、Disphos 和 Netphos[3,4,5]，並透過程式將可用序列大量傳遞給網站並回收這三個網站的預測結果並且一樣透過正規化放入資料庫內，建置為三張磷酸化預測結果之資料表。

三、實驗方法

我們的目的是希望可以整合三個不同網站之間的預測結果來達到整合優化的效果，期望能達到 1+1>2 的預測效果。所以我們必須將序列投入網站並回收其預測結果，然後再整理出各網站之相關數據如預測準確度(Accuracy)、馬修斯相關係數(MCC)等值回收，並將其整理以供我們整合並做出準確度更高的預測，其訓練及實際測試流程如次頁圖一。實驗一開始我們先把現有的蛋白質序列分類，依照其發生磷酸化處含有的激酶分成實驗常用的 PKA、PKC、CDK 和 CK2 四大類。然後再將此四大類蛋白序列分別利用程式大量投入此四個預測網頁並將其預測結果回收整理至資料庫，而回收的數據則有磷酸化之胺基酸(S、T 和 Y 等)、磷酸化發生之位置以及網站所給的磷酸化分數(用以代表其發生磷酸化可能性之強度，其數值介於 0~1 之間)。

將以上所述之資料回收至資料庫之後，我們再將其資料由原來之四大類再依其胺基酸之不同，分割成 S 和 T 兩大類，故目前我們已擁有八個類別的資料(PKA、PKC、CDK 和 CK2 分割成 S 和 T 總共八種)。分割完成之後，再將其回收回來之資料一一比對實際上已知磷酸化之資料庫，便分別得到了八個種類不同的預測準確度、MCC 等數據。

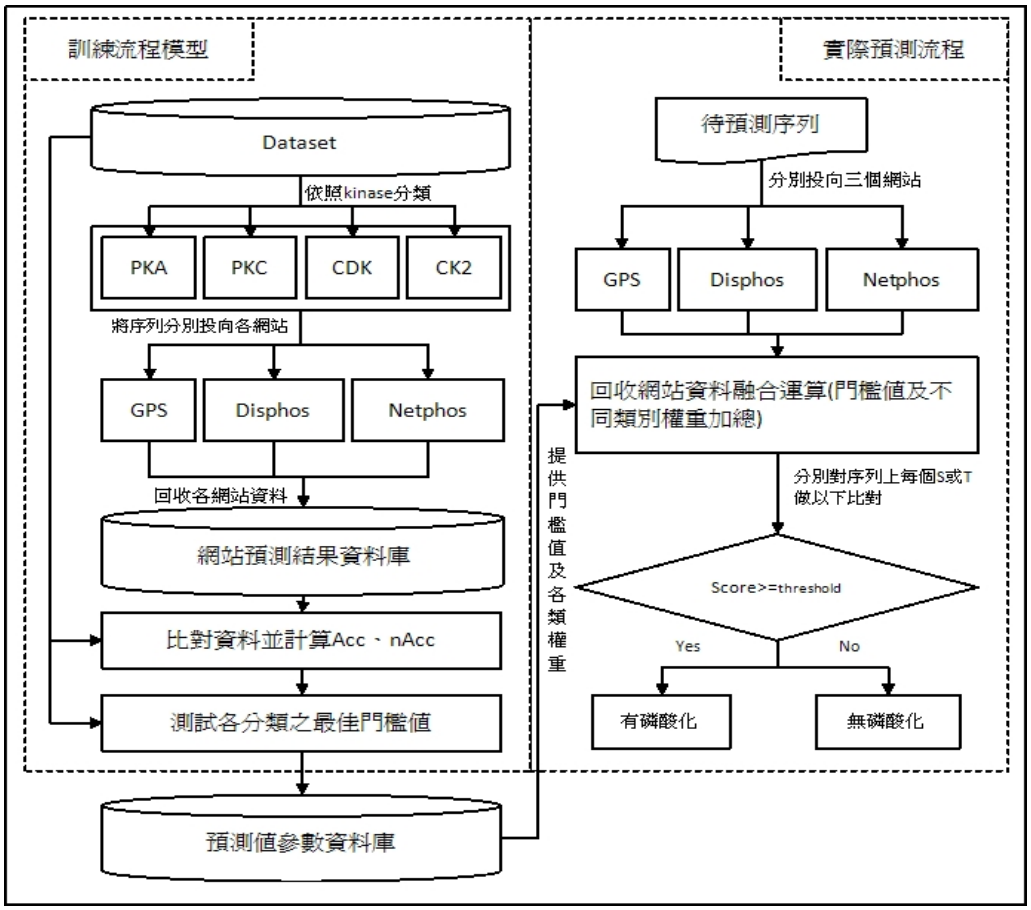

圖 1 預測系統之訓練與預測流程架構圖

當這些資料已經確定之後，我們便可以開始用這些數據建構我們自己的預測方式。而其預測方式為利用三個網站之互相投票而成之磷酸化預測。

而要完成這種投票機制，我們就必須先找到一種方式將三個網站的投票公平化。首先，我們在每一個類別中(PKA_S,PKA_T,PKC_S…等)會需要兩種參數，其一為該網站在此類別中之磷酸化預測的預測準確度(Acc)以及其預測無磷酸化(nAcc)之準確度。

其中 Acc 就是單純的利用是非題的問答方式所取得，一個歸類在 PKA_S 的序列上有著許多的 S 胺基酸，然後在根據已知的磷酸化資料跟其網站預測之資料比對，設 S 為 n 個，而其預測正確之命中個數為 t 個，則其 Acc 值則為 t/n。

而無磷酸化之準確度(nAcc)之取法則為，假設今天所要預測之類別為 PKA_S，則取出所有 PKA_S 類別中之序列上所有的 S 胺基酸位置，並逐一比對網站是否準確的預測該胺基酸位置是否為無磷酸化位置。

設所有待比對之無磷酸化之胺基酸總數為 n，而網站預測其無磷酸化之數量為 t，則該網站於此類別中的無磷酸化準確度(nAcc)則為 t/n，至於此數據之用法如下。

分別取得每個預測網站在每個類別下的數據之後，我們就可以開始進行投票，投票的方式為，若我們要去預測一序列上之 S 是否有磷酸化，則會去查詢三個網站之預測結果，若網站預測其有磷酸化則取其 Acc.值，若

預測無磷酸化則取其 nAcc 值，最後加三個網站之值加總，若其為 nAcc 則為減去。如方程式(1)：

$$\text{Score} = \sum_{i=1}^{n} \text{R} * \text{A}_\text{i} \begin{cases} \text{P}, \text{R} = 1, \text{A}_\text{i} = \text{Acc} \\ \sim\text{P}, \text{R} = -1, \text{A}_\text{i} = \text{nAcc} \end{cases} \quad (1)$$

其中 P 代表的為預測為磷酸化有發生磷酸化，~P 則是預測為無磷酸化，n 代表著網站的數目。

於是經過此運算之後，我們會在要預測的序列上的每個 S 或 T 位置上得到一個 score，至於 score 該大於多少為磷酸化，其門檻值的取法為，在一群已知的磷酸化資料中，取出該類別之序列，並大量以此投票方式算出 score，然後再以 0.01 為區間不斷測試其 Acc.及 Mcc 等數值，而 Mcc 等相關公式如下：

$$\text{Prec.} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)$$

$$\text{Sn.} = \frac{\text{TP}}{\text{TP+FN}} \quad (4)$$

$$\text{Sp.} = \frac{\text{TN}}{\text{TN+FP}} \quad (5)$$

$$\text{Acc.} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (6)$$

$$\text{Mcc.} = \frac{\text{TP}*\text{TN}-\text{FP}*\text{FN}}{(\text{TN+FN})*(\text{TN+FP})*(\text{TP+FN})*(\text{TP+FP})} \quad (7)$$

其中 TP 代表預測為磷酸化且實際亦為磷酸化之樣本個數，TN 代表預測為無磷酸化且實際亦無磷酸化之樣本個數，FP 代表預測為磷酸化但實際卻為無磷酸化之樣本個數，FN 代表預測為無磷酸化但實際卻為磷酸化之樣本個數。Prec.為精確性(Precision)，依照 positive 跟 negative 分為兩種，Sn.為靈敏度(Sensitivity)，Sp.為特異性(Specificity)，Acc.則為預測準確度(Accuracy)，MCC 則為馬修斯相關係數，這些係數越接近 1 則該特質越明顯。

而我們則取其 MCC 之最高位置為該類別之門檻值，如方程式(8)：

$$\text{threshold} = \text{MCC}_{\max_{0\le x\le 1}} \quad (8)$$

從各種不同的門檻值中取得每個類別中的最佳門檻值，而其馬修斯相關係數的優劣則為最優先考量。於是在八個類別中則會有著八個門檻值，當該類別的分數大於此門檻值時我們則預測其為有磷酸化之發生，小於則預測為無磷酸化之發生。

至此，經過網站存取、權重及門檻值的運算，我們所做的磷酸化預測結果正式產生。

四、實驗結果

在整個磷酸化預測系統建構完成之後，為了在整體測試上達到合理且公正的測試結果，我們再從資料庫中分別依八個類別各隨機挑出五組、每組各五條序列，並將這八種測試序列分別投給 GPS、Disphos、Netphos 以及我們自己建構的預測系統，其隨機測試結果如次頁表一，並參考圖二。

表一 Website 欄位中的數值代表著三個網站以及兩組我們所做出的預測系統。其中 Vote 的欄位是我們強調馬修斯相關係數重要性的對照組，他的預測方式與資料融合類似，但是投票時不帶入權重，三個網站每一票比

重相同，意即兩票決定其預測方向。

而圖二的折線圖中的 X 軸代表著五項預測評估之相關係數，Y 軸則是係數的特徵強度，在折線圖中的折線若能表現的越平滑且 Y 軸高度越高，其整體預測表現也越趨優良。

表 1 分組隨機測試結果

	Websites	*Pre*	*Sn*	*Sp*	*Acc*	*CC*
PKA_S	GPS	0.148	0.939	0.939	0.873	0.066
	Disphos	0.176	0.124	0.623	0.623	0.278
	Netphos	0.313	0.496	0.882	0.851	0.312
	Vote	0.396	0.027	0.994	0.916	0.080
	DataFusion	**0.324**	**0.562**	**0.882**	**0.857**	**0.350**
PKC_S	GPS	0.158	0.951	0.951	0.838	0.015
	Disphos	0.233	0.058	0.630	0.630	0.318
	Netphos	0.221	0.632	0.674	0.684	0.222
	Vote	0.256	0.014	0.985	0.864	0.016
	DataFusion	**0.272**	**0.610**	**0.769**	**0.762**	**0.286**
CDK_S	GPS	0.065	0.928	0.928	0.798	-0.044
	Disphos	0.223	0.048	0.508	0.508	0.253
	Netphos	0.293	0.623	0.700	0.692	0.257
	Vote	0.108	0.019	0.985	0.842	0.000
	DataFusion	**0.481**	**0.467**	**0.885**	**0.824**	**0.360**
CK2_S	GPS	0.142	0.947	0.947	0.786	-0.024
	Disphos	0.283	0.037	0.565	0.565	0.310
	Netphos	0.319	0.604	0.688	0.678	0.253
	Vote	0.500	0.025	0.993	0.822	0.073
	DataFusion	**0.376**	**0.553**	**0.794**	**0.754**	**0.305**
PKA_T	GPS	0.000	0.929	0.929	0.855	-0.077
	Disphos	0.202	0.000	0.697	0.697	0.301
	Netphos	0.195	0.752	0.728	0.731	0.283
	Vote	0.000	0.000	0.993	0.914	-0.035
	DataFusion	**0.292**	**0.633**	**0.847**	**0.830**	**0.347**
PKC_T	GPS	0.000	0.944	0.944	0.877	-0.062
	Disphos	0.192	0.000	0.752	0.752	0.308
	Netphos	0.274	0.640	0.848	0.840	0.345
	Vote	0.000	0.000	0.997	0.927	-0.031
	DataFusion	**0.350**	**0.507**	**0.913**	**0.888**	**0.360**
CDK_T	GPS	0.131	0.928	0.928	0.847	0.040
	Disphos	0.181	0.120	0.603	0.603	0.260
	Netphos	0.288	0.429	0.843	0.805	0.233
	Vote	1.000	0.020	1.000	0.905	0.211
	DataFusion	**0.277**	**0.409**	**0.877**	**0.830**	**0.241**
CK2_T	GPS	0.100	0.950	0.950	0.869	-0.006
	Disphos	0.235	0.035	0.731	0.731	0.332
	Netphos	0.451	0.616	0.925	0.897	0.469
	Vote	0.500	0.035	0.997	0.912	0.131
	DataFusion	**0.520**	**0.520**	**0.949**	**0.910**	**0.468**

從表一及圖二之資料可見，Vote 在預測結果的預測準確度上明顯高於其他預測方式，但是其靈敏度及馬修斯相關係數明顯遠低於其他預測方式，所以 Vote 高準確度的原因只是保守的不做出預測，犧牲靈敏度及馬修斯相關係數來達到高預測準確度，但是這種利用有磷酸化數量遠低於無磷酸化數量的方式，來達到取巧的高預測準確度卻可以由其靈敏度及馬修斯相關係數的低落看出其預測方式之所以不可取之理由。

而 Data Fusion 欄代表我們經由資料融合所得到的預測結果，從表中可見，經過資料融合之後所得到之精確性(Precision)、準確度(Accuracy)以及馬修斯相關係數(MCC)大多明顯優於其他網站之預測結果，而其中少數 Mcc 或者準確度較低的情況，卻也能達到平衡的狀態，在該項目不明顯低於其他網站預測結果的情況下，也能讓另外幾項數據保持在水平之上，而不會有犧牲其中一項數據而提升另一項數據的情況產生。

而在數據細節分析時，我們發現 GPS 在整體預測表現上較弱，但是卻發現他在預測無磷酸化時的準確度明顯優於其他網站，而從我們的融合結果上也能發現我們在預測無磷酸化的準確度上也優於另外兩個網站，這結果極有可能是融合進了 GPS 在預測無磷酸化上的優勢，也因為權重比例跟門檻值的把關，而不會讓我們在預測磷酸化位置上的表現被拉低，甚至提高了我們在整體的預測表現，在最佳的表現案例中，我們在預測準確度不低於其他網站表現下的最佳案例中，馬修斯相關係數的結果高出最佳網站表現 0.1 之多、精確性更高出 0.19。

五、總結與討論

一開始我們所提出的構想，是由於不同網站對於磷酸化預測使用之演算法之不同且其強項也不同之假設下所進行資料融合，以期網站之間能互相截長補短，將不同網站之資料結合，使其預測之最高點能突破不同網站中最佳的一個預測結果，達到 1+1>2 的

互補效應。

而最後就整個測試結果來論，對於三個網站的預測結果所做出的資料融合,經由每個網站不同的預測方式，所提供的不同的預測資料融合運算，明顯提升了我們在這個蛋白質磷酸化資料庫內的預測品質表現，不僅綜合表現優於三個網站，連在大部份的單項表現上亦高於其他網站，也達到了我們當初所要達到的 1+1>2 的資料融合目的。

至於未來可望再配合更多不同的網站來引入融合，來達到更多元性、多面性的評估能力。甚至除了單純的計算測試的融合方式外，我們還打算接著利用各種人工智慧(例如類神經網路、基因演算法等)的方法來應用這些網站所給予的數據，去找到更好的預測方法跟更有效率的數據應用，來達到更高的預測效果。

誌謝

感謝國科會計畫 NSC 96-2218-E -468-001 對此研究工作之補助。

六、參考文獻

[1] Jong Hun Kim,Juyoung Lee, Bermseok Oh, Kuchan Kimm and ln-Song Koh," Prediction of phosphorylation sites using SVMs", J.H. Kim et al. Vol.20, pages 3179–3184, 2004.

[2] Francesca Diella, Scott Cameron, Christine Gemünd, Rune Linding, Allegra Via, Bernhard Kuster, Thomas Sicheritz-Pontén, Nikolaj Blom and Toby J Gibson, " Phospho.ELM: A database of experimentally verifiedphosphorylation sites in eukaryotic proteins", BMC Bio- formatics,2004.

[3] http://bioinformatics.lcd-ustc.org/gps_web/

[4] http://core.ist.temple.edu/pred/

[5] http://www.cbs.dtu.dk/services/NetPhos/

[6] 蔡津津、趙杰煜、王樂珩, "AproPhos:基於 AdaBoost 方法的蛋白質磷酸化修飾預測系統",微電子學與計算機, 第 24 卷,第 7 期, pp.35-39, 2007.

[7] Nikolaj Blom, Thomas Sicheritz Ponten, Ramneek Gupta, Steen Gammeltoft and Soren Brunak" Prediction of post-translational glycolsylation and phosphorylation of proteins from the amino acid sequence" N. Bl- om et al.1, Vol.4,pp. 1633–1649, 2004.

[8] Y. Xue, F. Zhou, M. Zhu, G. Chen and X.Yao."GPS: a comprehensive www server for phosphorylation sites prediction",Nucleic Acids Research,Vol.33,w184-w187,2005.

[9] Genetha A. Gray, Pamela J. Williams and Kenneth L. Sale," Disparate Data Fusion for Protein Phosphorylation Prediction", Proceedings of the AI/DM Pre-Conference Work- shop at the INFORMS Annual Meet- ing,2006.

[10] Nick Littlestone and Manfred K. Warmuth,"The Weighted Majority Algorithm", Foundations of Computer Science ,IEEE, pp.256-261, 1989.

[11] Shai BenDavid, Johannes Gehrke and Reba Schuller,” A Theoretical Framework for Learning from a Pool of Disparate Data Sources”, Proceedings of the eighth ACM SIGKDD international conference, pp.443-449,2002.

[12] SUNG K. AHN,” A bioinformatics -based approach for the prediction and identification of novel proteins potentially involved in phosphorrylation signalling pathways”, INTERNATIONAL JOURNAL OF MOLECULAR MEDICINE, Vol12, pp. 391-397, 2003.

[13] Nikolaj Blom, Steen Gammeltoft and Sùren Brunak,” Sequence and Structure-based Prediction of ukaryotic Protein Phosphorylation Sites”, Mol. Biol.Vol.294,pp. 1351－1362,1999.

[14] Jeffrey J. Saucerman, Jin Zhang, Jody C. Martin, Lili X. Peng, Antine E. Stenbit, Roger Y. Tsien,and Andrew D. McCulloch,” Systems analysis of PKA-mediated phosphorylation gradients in live cardiac myocytes”, PNAS,Vol. 103, pp.12923–12928,2006.

[15] Akira Kikuchi, Shosei Kishida and Hideki Yamamoto,” EXPERIMENTAL and MOLECULAR MEDICINE”, Vol.38,pp. 1-10, 2006.

[16] Wei Guo,Feng Wei, Shiping Zou, Meredith T. Robbins,Shinichi Sugiyo,Tetsuya Ikeda,Jian-Cheng Tu,Paul F. Worley,Ronald Dubn Dubner,and Ke Ren,”Group I Metabotropic Glutamate Receptor NMDA Receptor Coupling and Signaling Cascade Mediate Spinal Dorsal Horn NMDA Receptor 2B Tyrosine Phosphorylation Associated with Inflammatory Hyperalgesia”, The Journal of Neuroscience, Vol.24(41), pp.9161–9173, 2004.

[17] James R. A. Hutchins,Dina Dikovskaya , and Paul R. Clarke,” Regulation of Cdc2/Cyclin B Activation in Xenopus Egg Extracts via Inhibitory Phosphorylation of Cdc25C Phosphatase by Ca2_/ Calmodium-dependent Kinase II”, Molecular Biology of the Cell, Vol.14, pp. 4003–4014, 2003.

[18] Patrick Viatour, Marie-Paule Merville, Vincent Bours and Alain Chariot,” Protein Phosphorylation as a Key Mechanism for the Regulation of BCL-3 Activity”, Vol.3, pp.1498 –1501, 2004.

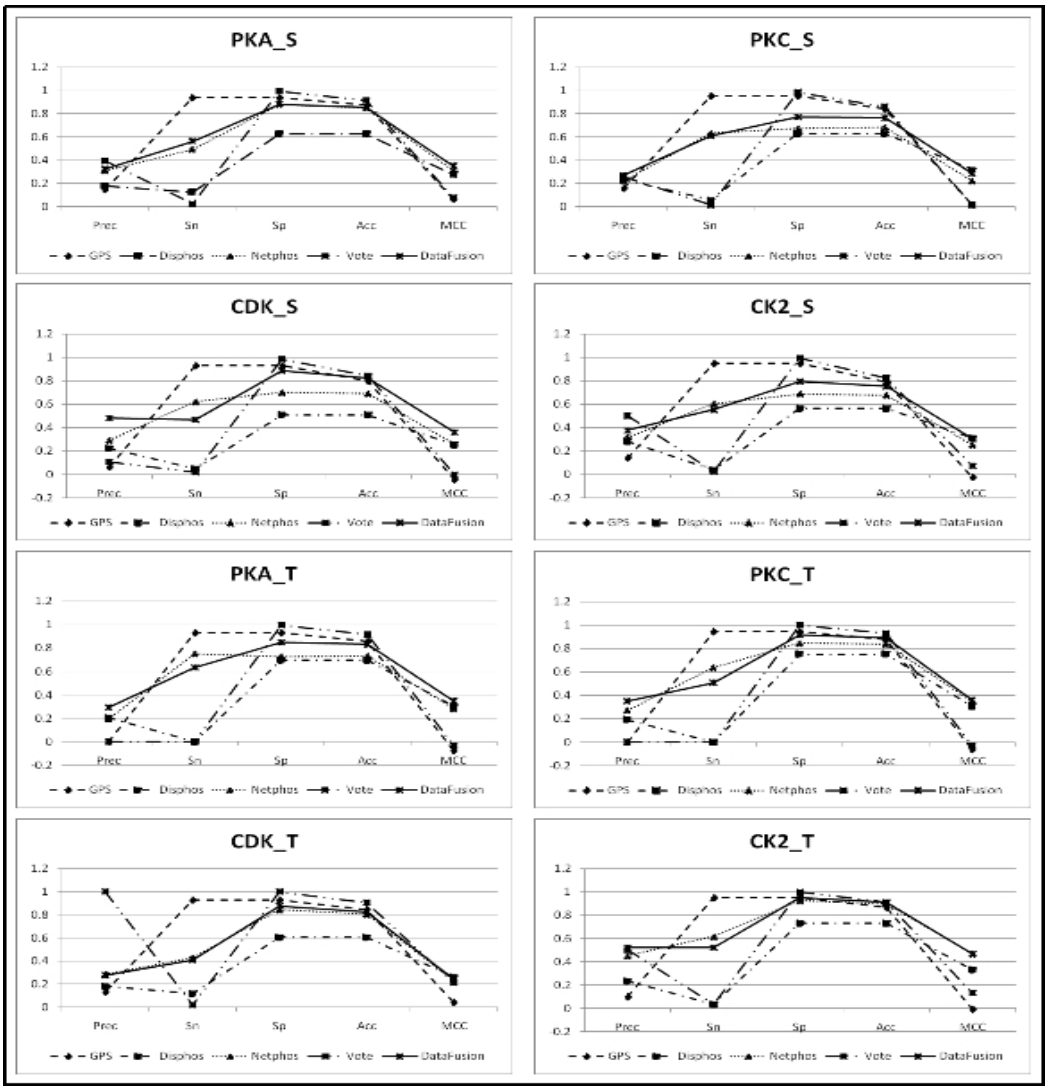

圖 2 各種預測方式在各數據上的表現