利用序列保留性預測磷酸化發生的位置

朱彥煒 游景盛 陳芝瑩 謝昆璋
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 eternitycat@hotmail.com man711126@hotmail.com

摘要

磷酸化是蛋白質中最重要的後轉譯修飾之一，它參與各種不同的生物訊號傳遞的路徑。在傳遞路徑中，蛋白質的磷酸化與去磷酸化反應為真核細胞提供了調控機制，控制了許多酵素和受體的活化與抑制。因此，若能預測磷酸化作用的位置，將對相關的研究有很大的助益。本論文提出了一個新的序列保留性計算法則，它是利用已知磷酸化的片段當作預測的模板，並利用其與非磷酸化的片段求出決定相似度的門檻值。最後實驗預測的結果顯示在馬修斯相關係數及預測準確度都有相當大的改進。

關鍵詞：序列保留性、磷酸化、預測、BLOSUM62

一、前言

磷酸化（Phosphorylation）或稱磷酸化作用，是指在蛋白質或其他類型分子中加入一個磷酸（PO_4）基團，也可定義成「將一個磷酸基團導入一個有機分子」。此作用在生物化學中佔有極重要地位。

許多剛合成出的蛋白質經過化學修飾後才能形成有功能性的蛋白，而這些化學修飾稱之為後轉譯修飾作用（Post-translational modification），有磷酸化（Phosphorylation）、醣化（Glycosylation）、乙醯化（Acetylation）、甲基化（methylation）及加入輔基（Prosthetic groups），這些蛋白質的修飾能為蛋白質改變化學性質、改變蛋白質結構，例如：在二個半胱胺酸中的二個相鄰硫胺基，可能被氧化成雙硫鍵、新生的 NH2-端胺基酸有時被乙醯基化。

以上列出的後轉譯修飾作用裡，以磷酸化最為重要，因為蛋白質的磷酸化對於生物傳遞有很複雜的調控機制。且磷酸化對於真核生物來說[1]，可以算是很重要的一種機制。

蛋白質磷酸化可發生在許多種類的胺基酸（蛋白質的主要單位）上，其中以絲胺酸（Serine）為多，接著是蘇胺酸（Threonine）。而酪胺酸（Tyrosine）則相對較少磷酸化的發生，不過由於經過磷酸化之後的酪胺酸較容易利用抗體來純化，因此酪胺酸的磷酸化作用位置也較廣為了解。

但磷酸化與去磷酸化，需要一些特別的酵素來反應，例如：蛋白激酶(protein kinase)和磷酸酶(phosphatase)，可使得蛋白質磷酸化在特定的序列上的位置的胺基酸進行反應。

而現在預測磷酸化的方法有很多種，例如：(1)SVM(支援向量機)[2]，可分為線性與非線性，主要在輸入的訓練資料(Training Data)中，找出一個可以將資料分隔開最大邊界(Margin)的區分超平面(Separating Hyperplane)，希望可以在不同類別的資料中，找出最大邊

界(Margin)的區分超平面。而目前SVM的預測結果準確性kinase family已經可達到83~95%，而kinase group的準確性也可達76~91%。(2) HMM（隱藏馬可夫模型）[3]，利用不同的序列與激酶組成不同的受質所建立形成的，並將模型實作成預測系統，稱為KinasePhos。(3) K-Nearest Neighbor Algorithm（K近鄰分類演算法）[4]，利用蛋白質的一级序列訊息對不同激酶家族作用的磷酸化位置進行有效的預測, 不但具有快速、自動等優点，還可以對相應的實驗測定進行指令，具有重要的意義。

所以在許多研究蛋白質領域裡，有不少人是將機器學習方法來應用，來求得磷酸化預測的準確值。而在本研究中，則是利用序列保留性的策略來做磷酸化位置的預測，而度量序列保留性是採用用替代矩陣(substitution matrix) BLOSUM62給的值去做比對。

二、資料集的建立

在此研究中，我們所使用的原始資料來源為 Phospho ELM Database[5]取得蛋白質磷酸化修飾的位置建立起樣本的集合，Phospho.ELM Database 修飾位置的數據都是經由嚴格的生物實驗室驗證的。我們將其磷酸化資料引用，並將其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。此資料庫包括 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:

資料集中的 4422 條蛋白質序列，依照其發生磷酸化處所作用的激酶有PKA、PKC、CDK、CK2、EGFR、Lck、ITK、MHCK、ATM、…等，但因為大部份作用的激酶資料太少，無法提供有效的分析與學習。在一般的研究中，僅討論 PKA、PKC、CDK、CK2 這四種激酶。四大類激酶所有包含的序列有：PKA 激酶之序列 336 條、PKC 有 257 條、CDK 有 104 條及 CK2 有 249 條。被磷酸化的胺基酸也有這個問題，只討論數量較多的酪胺酸(Ser)和蘇胺酸(Thr)。我們將四種激酶和二種胺基酸分為 PKA_S、PKA_T、PKC_S、PKC_T、CDK_S、CDK_T、CK2_S 及 CK2_T 八類，分別計算其序列保留性與預測。而這些序列裡面含有的 S 胺基酸總計有 5968 個、T 胺基酸總計有 2943 個可供實驗測試及運算[6]。

利用此資料庫，我們將可進行磷酸化預測比對，將裡面發生磷酸化的位置做為正例標本；未發生磷酸化的位置做為反例的標本。再使用 BLOSUM62 加權去做比對，BLOSUM62 值如圖一所示：

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

圖一、BLOSUM62替代矩陣

三、研究方法

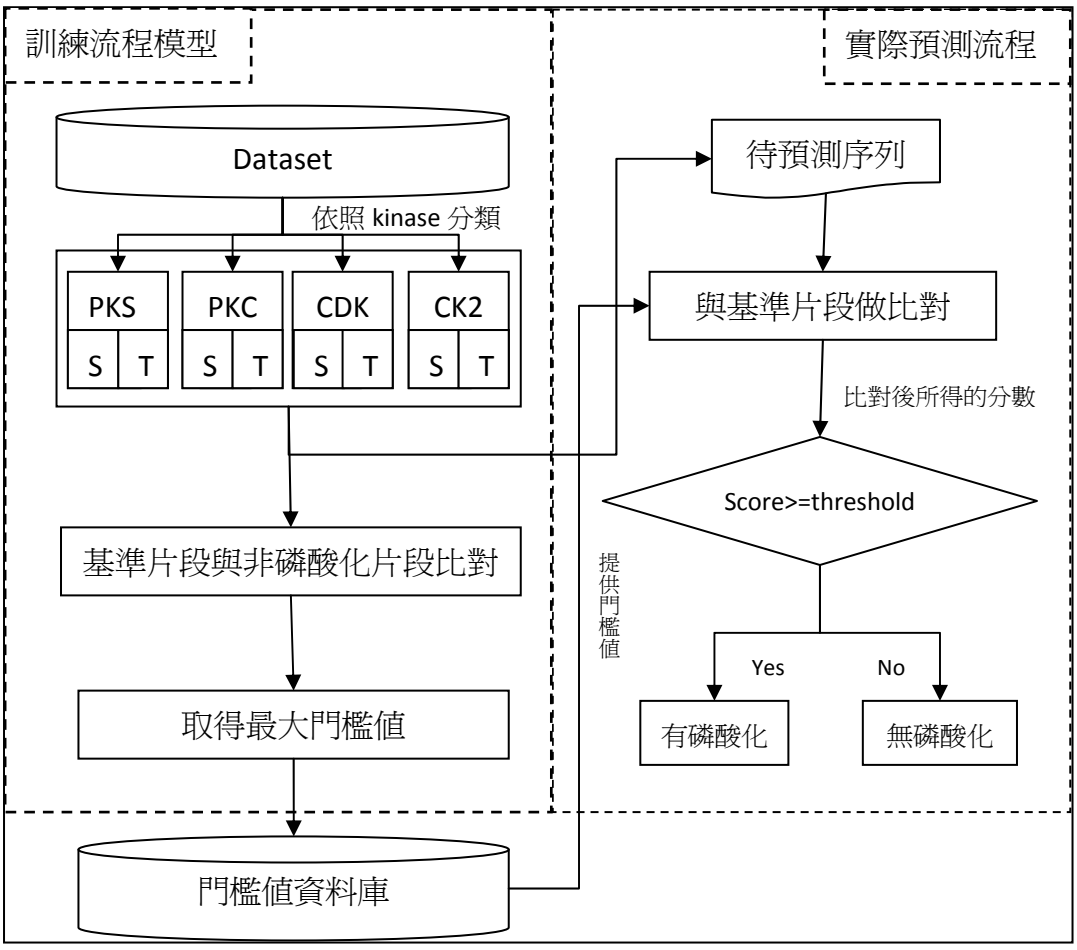
我們假設磷酸化的片段彼此之間是比較相似的，而與非磷酸化的片段相似度較低。因此，當我們視每一條磷酸化片段為比較的基準片段時，也就是用此基準片段來分類磷酸化與非酸化片段。這些基準片段之間，其與所有非磷酸化相似度應該不一定相同，所以我們必需對每一個基準片段之於所有的非磷酸化片段，取一個最大的相似值，每一個基準片段的最大相似值是不同的。若給一個片段，其與基準片段的相似值大於該基準片段的最大相似值，我們才定義該片段與此基準片段相似。這個最大相似值又稱為門檻值，底下我們將說明門檻值是如何設定的(參考圖二)。

(一) 設定磷酸化片段之門檻值

每一個訓練資料中的磷酸化片段都會有一個門檻值，其值代表此磷酸化與所有非磷酸化片段的最大相似度。為了計算門檻值，我們先將兩片段相似的分數定義如方程式(1)

$$\text{Score}(x,y)=\sum_{i=1}^{n} \text{blosum62}(xi + yi), \quad n=15 \tag{1}$$

相似的分數即利用 blosum62 替代矩陣(substitution matrix)，計算出此片段中兩兩胺基酸的相似分數之總和。另外，為了相似分數的正規化，我們利用方程式(2)得到這兩片段的相似度。



圖二 、流程圖

$$\text{Similarity}(x,y) = \frac{\text{score}(x,y)}{\text{score}(x,x)} \quad (2)$$

此磷酸化片段最大的相似度即為我們的門檻值，如方程式(3)所示。

$$\text{Threshold}=\max(\text{Similarity}(x_i,y_1), \text{Similarity}(x_i,y_2), \text{Similarity}(x_i,y_3),\ldots\text{Similarity}(x_i,y_n)) \quad (3)$$

(二) 預測方法

我們有了這 552 個門檻值之後，開始著手測試：將所有 PKA 序列裡，隨機取 5 條序列出來當做測式用序列。從這五條序列裡找出所有 S 位置，取前後各 7 個胺基酸，共 15 個胺基酸為一個片段，做為測式片段。將這些測式片段和 552 條基準片段做比對，此步驟做 5 次，共使用 25 條有重複可能性的序列，得 TP、TN、FT、FN 值算出獲取 MCC 值、ACC 值、SN 值、SP 值及 Precision 值。

(三) 評估方式

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

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

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

$$\text{Precision} = \frac{\text{TP}}{(\text{TP+FP})} \quad (7)$$

$$\text{MCC} = \frac{(\text{TP}\times\text{TN}) - (\text{FN}\times\text{FP})}{\sqrt{(\text{TP+FN})\times(\text{TN+FP})\times(\text{TP+FP})\times(\text{TN+FN})}}, \quad (8)$$

在(4)到(7)的公式中，FN（False Negative）表示失敗的預測將對的預測成錯的次數，TP（True Positive）表示成功的預測將對的預測成對的次數，FP（False Positive）表示失敗的預測將錯的預測成對的次數，TN（True Negative）表示成功的預測將對的預測成錯的次數。公式(4)Sn（sensitivity）值為靈敏度；公式(5)Sp（specificity）值為特異性；公式(6)Acc（Accuracy）表對的預測為對的除以預測為對的數量，所以稱為準確度；公式(7)Precision表示針對預測有磷酸化地方的精準度；公式(8)MCC為馬修斯相關係數。

表一、分組隨機測試結果

		Mcc	*Acc*	*Precision*	*Sn*	*Sp*
PKA_S	**SC**	**0.296**	**91.503**	**0.455**	**0.250**	**0.974**
	GPS	0.056	87.052	0.138	0.114	0.937
	disphos	0.272	64.162	0.169	0.871	0.621
	netphos	0.309	86.301	0.294	0.493	0.896
PKC_S	**SC**	**0.293**	**87.596**	**0.422**	**0.306**	**0.947**
	GPS	0.003	84.669	0.115	0.056	0.946
	disphos	0.339	64.111	0.228	0.931	0.605
	netphos	0.252	70.383	0.223	0.669	0.708
CDK_S	**SC**	**0.333**	**86.617**	**0.545**	**0.293**	**0.960**
	GPS	-0.037	80.425	0.093	0.044	0.929
	disphos	0.248	51.094	0.211	0.902	0.447
	netphos	0.238	70.544	0.261	0.596	0.723
CK2_S	**SC**	**0.250**	**83.399**	**0.522**	**0.203**	**0.962**
	GPS	-0.012	79.862	0.149	0.041	0.953
	disphos	0.298	55.206	0.263	0.913	0.479
	netphos	0.237	70.825	0.302	0.552	0.740
PKA_T	**SC**	**0.471**	**92.991**	**0.600**	**0.429**	**0.975**
	GPS	-0.076	85.748	0.000	0.000	0.934
	disphos	0.311	71.729	0.201	0.829	0.707
	netphos	0.283	73.364	0.198	0.743	0.733
PKC_T	**SC**	**0.383**	**92.927**	**0.447**	**0.396**	**0.966**
	GPS	-0.063	88.049	0.000	0.000	0.941
	disphos	0.340	77.195	0.201	0.849	0.767
	netphos	0.353	85.732	0.261	0.660	0.871
CDK_T	**SC**	**0.482**	**92.100**	**0.644**	**0.422**	**0.975**
	GPS	0.042	84.740	0.141	0.111	0.927
	disphos	0.264	61.364	0.183	0.856	0.588
	netphos	0.225	81.818	0.247	0.422	0.861
CK2_T	**SC**	**0.203**	**89.691**	**0.355**	**0.180**	**0.968**
	GPS	-0.027	86.451	0.057	0.033	0.947
	disphos	0.311	71.870	0.215	0.803	0.710
	netphos	0.465	89.691	0.446	0.607	0.926

四、實驗結果與分析

在建立門檻值資料庫後，依八種不同的類別之中各隨機挑出五組、每組各五條序列，並將這八種測試序列分別投給 GPS[3]、Disphos、Netphos 以及我們的預測方法(SC)，其隨機測試結果如下表一，並參考圖二。

由測試結果得知，針對絲胺酸的部份，我們在 PKA、PKC 與 CK2 的 ACC 值與 Precision 值都優於其他的網站，但 MCC 值略低。唯 CDK 的 MCC 值、ACC 值及 Precision 值皆優於其他的網站。

而在蘇胺酸方面，除了 CK2 的各項數值偏低，劣於其它網站外，其他的類別 PKA、PKC 與 CDK 的 ACC 值、MCC 值與 Precision 值全都優於其他網站，以整體論，我們所使用的方法提升了預測磷酸化的準確率。

五、討論與未來展望

所有重要的生物調節系統都多少與蛋白磷酸化有些關連。甚而有人推測在我們身體中有多於一千種不同的激酶，分別負責體內各種細胞間訊號傳遞的蛋白質磷酸化反應。研究各種激酶的特性與其負責調控的反應，並試圖了解不同激酶間的相互影響。

本論文先提出一個資料保留性的策略，來改進蛋白質磷酸化預測的正確率。一開始是利用片段比對來獲取門檻值以作為判斷是否為磷酸化的一個基準，進而提升整體的預測正確率。另外，我們將自己預測的結果數據與其他三個網站相比較，絕大部分的預測結果品質都較佳，但唯獨CK2_T的結果不如我們所預期，再接下來的研究，我們將會觀察CK2_T中的序列與結構，探討與研究預測不佳的原因。另外，替代矩陣BLOSUM62在我們所提的策略中扮演相當重要的角色，未來我們將分析不同的替代矩陣是否對我們所提出的策略有不同的影響

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] Nikolaj Blom, Steen Gammeltoft and Sùren Brunak,” Sequence and Structure-based Prediction of Eukaryotic Protein Phosphorylation Sites”, J. Mol. Biol,pp.1351-1362,1999.

[2] Jong Hun Kim, Juyoung Lee, Bermseok Oh, Kuchan Kimm and InSong Koh,” Prediction of phosphorylation sites using SVMs”, J.H.Kim et al.Vol.20,pp.3179-3184,2004.

[3] Hsien-Da Huang , Tzong-Yi Lee , Shih-Wei Tzeng , Li-Cheng Wu , Jorng-Tzong Horng , Ann-Ping Tsou , Kuan-Tsae Huang,” Incorporating Hidden Markov Models for Identifying Protein Kinase-specific Phosphorylation Sites”, Journal of Computational Chemistry, Vol.26,pp.1032-1041,2005.

[4] Ao Li, Lirong Wang, Yunzhou Shi, Minghui Wang, Zhaohui Jiang and Huanqing Feng,” Phosphorylation Site Prediction with A Modified

k-Nearest Neighbor Algorithm and BLOSUM62 Matrix",IEEE Engineering in Medicine and Biology 27th Annual Conference,pp.1-4.2005.

[5] 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 verified phosphorylation sites in eukaryotic proteins",BMC Bioinformatics,2004.

[6] T.K. Attwood," The quest to deduce protein function from sequence: the role of pattern databases", The International Journal of Biochemistry & Cell Biology,pp.139-155,2000.

[7] 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.

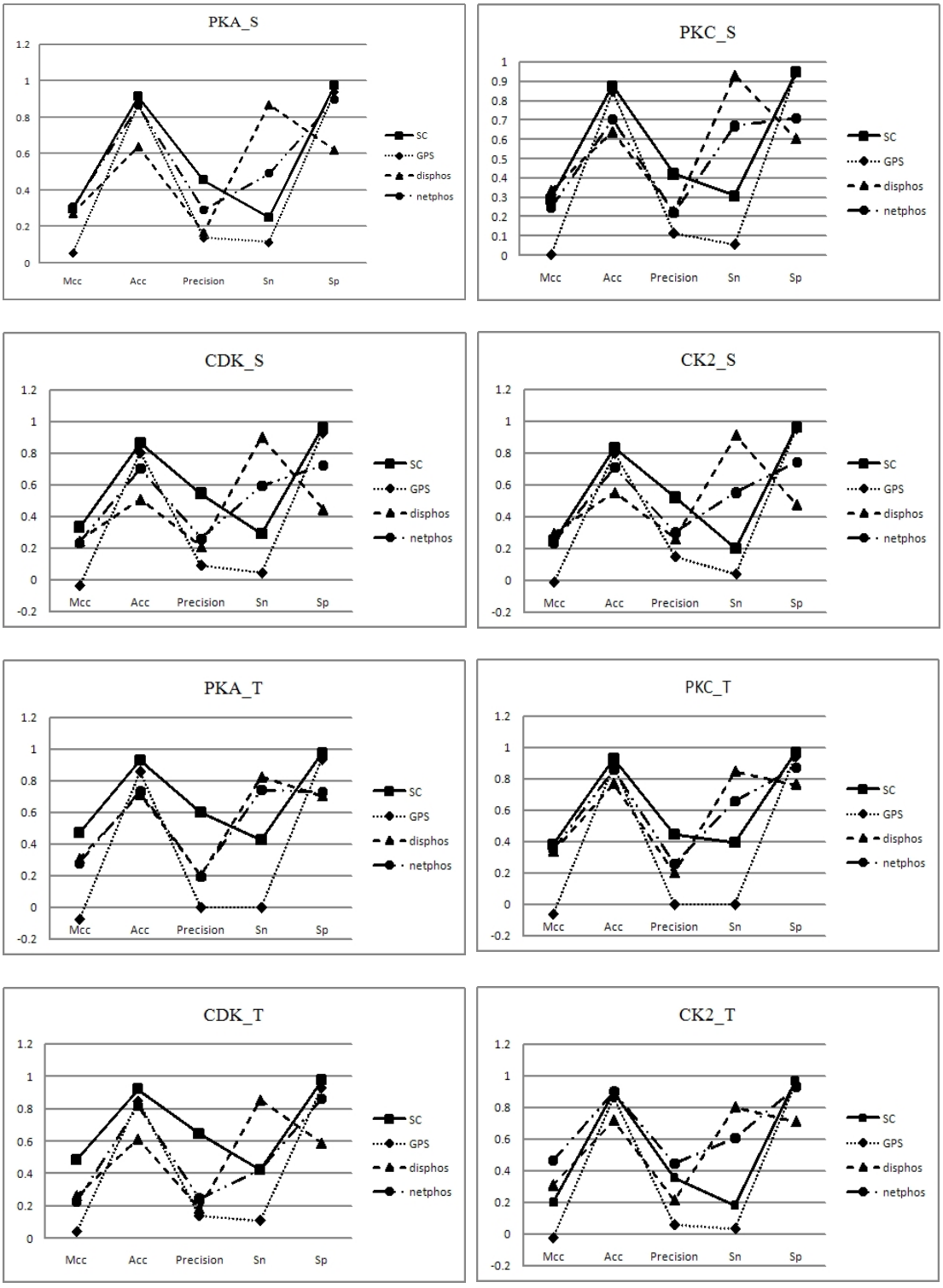
[8] Nikolaj Blom, Thomas Sicheritz-Ponten, Ramneek Gupta, Steen Gammeltoft and Soren Brunak," N. Blom et al.pp.1633-1649.2004.

[9] Yu Xue, Fengfeng Zhou, Minjie Zhu, Kashif Ahmed, Guoliang Chen and Xuebiao Yao," GPS: a comprehensive www server for phosphorylation sites prediction", Nucleic Acids Research,Vol.33,W184-W187,2005.

[10] Hsien-Da Huang, Tzong-Yi Lee, Shih-Wei Tzeng and Jorng-Tzong Horng," KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites", Nucleic Acids Research, Vol.33, W226-W229,2005.

[11] Peter V. Hornbeck, Indy Chabra, Jon M. Kornhauser, Elzbieta Skrzypek and Bin Zhang," PhosphoSite: A bioinformatics resource dedicated to physiological protein Phosphorylation", P. V. Hornbeck *et al.*pp.1551-1561,2004.

[12] Nick Littlestone and Manfed K. Warmuth,"The Weighted Majority Algorithm",IEEE,pp.256-261,1989.



圖三、各種預測方式在各數據上的表現