BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 13 2011, pages 1780–1787
doi:10.1093/bioinformatics/btr291

Sequence analysis Advance Access publication May 6, 2011

Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences

Tzong-Yi Lee*, Zong-Qing Lin, Sheng-Jen Hsieh, Neil Arvin Bretaña and Cheng-Tsung Lu

Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 320, Taiwan

Associate Editor: John Quackenbush

ABSTRACT

Summary: Bioinformatics research often requires conservative analyses of a group of sequences associated with a specific biological function (e.g. transcription factor binding sites, micro RNA target sites or protein post-translational modification sites). Due to the difficulty in exploring conserved motifs on a large-scale sequence data involved with various signals, a new method, MDDLogo, is developed. MDDLogo applies maximal dependence decomposition (MDD) to cluster a group of aligned signal sequences into subgroups containing statistically significant motifs. In order to extract motifs that contain a conserved biochemical property of amino acids in protein sequences, the set of 20 amino acids is further categorized according to their physicochemical properties, e.g. hydrophobicity, charge or molecular size. MDDLogo has been demonstrated to accurately identify the kinase-specific substrate motifs in 1221 human phosphorylation sites associated with seven well-known kinase families from Phospho.ELM. Moreover, in a set of plant phosphorylation data-lacking kinase information, MDDLogo has been applied to help in the investigation of substrate motifs of potential kinases and in the improvement of the identification of plant phosphorylation sites with various substrate specificities. In this study, MDDLogo is comparable with another well-known motif discover tool, Motif-X.

Contact: francis@saturn.yzu.edu.tw

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 7, 2011; revised on April 13, 2011; accepted on May 2, 2011

1 INTRODUCTION

The conservative analysis of a group of sequences associated with a specific biological function such as the investigation of protein post-translational modification (PTM) sites, analysis of micro RNA (miRNA) target sites, or the identification of transcription factor binding sites is often required in bioinformatics research. Sequence logos were first proposed by Tom Schneider and Mike Stephens in 1990 to display patterns in sequence conservation, and to assist in discovering and analyzing such patterns (Schneider and Stephens, 1990). In 2004, a web-based sequence logo generator, WebLogo, was developed for creating graphical representations of the patterns within a multiple sequence alignment (Crooks *et al*., 2004). Several extensions of sequence logo have also been developed, e.g. RNA structure logos (Gorodkin *et al*., 1997), PSSM logos (Fujii *et al*., 2004), energy normalized sequence logos (Workman *et al*., 2005), subfamily logos (Beitz, 2006), 3D sequence logos of DNA or RNA (Bindewald *et al*., 2006) and RNA structural alignment logos (Chang *et al*., 2008). In the case of functionally associated sequence patterns (e.g. protein phosphorylation sites), Two Sample Logo (Vacic *et al*., 2006) has been developed for generating graphical representations of statistically significant position-specific differences in amino acid compositions between phosphorylated sequences and non-phosphorylated sequences. With regard to motif discovery, a wealth of popular approaches have been proposed, such as MEME (Bailey and Elkan, 1994), TEIRESIAS (Rigoutsos and Floratos, 1998), Gibbs motif sampler (Thompson *et al*., 2003) and eMOTIF (Nevill-Manning *et al*., 1998). Additionally, an iterative statistical approach, Motif-X (Schwartz and Gygi, 2005), has been proposed for identifying motifs from large-scale datasets. Motif-X has demonstrated results that outperform other methods based on the identification of phosphorylation motifs.

According to the basic concept of sequence conservation, a sequence logo displays the nucleotide or amino acid composition for each position in a group of aligned signal sequences. However, it is difficult to explore conserved motifs for large-scale sequence data; for instance, a sequence logo for all phosphorylation data involved with various catalytic kinases fails to obviously present the kinase-specific substrate specificity. Thus, a new method is developed, MDDLogo, which applies maximal dependence decomposition (MDD) to cluster a group of aligned signal sequences into subgroups containing statistically significant motifs. Assuming that each position in a set of aligned sequences is a sample of symbols (nucleotides or amino acids) generated according to a probability distribution, a null hypothesis is based on the assumption that the symbol distribution of two positions are mutually independent. MDD is a recursive methodology that adopts the chi-square test to capture the most significant dependencies between each position. In a previous study (Burge and Karlin, 1997), MDD was proposed to group the splice sites during the identification process of splice-site prediction.

In order to extract motifs containing conserved biochemical properties of amino acids in protein sequences, the set of 20 amino acids is further categorized according to their physicochemical properties, e.g. hydrophobicity, charge or molecular size. After MDD clustering, each subgroup, containing statistically significant motifs, is graphically visualized using WebLogo (Crooks *et al*., 2004). With the exponential increase of protein PTM sites identified

*To whom correspondence should be addressed.

by mass spectrometry, there is a motivation to further understand modification sites by studying its surrounding motifs. A case study done on 1221 human phosphorylation sites associated with seven well-known kinase groups has been used to test the effectiveness of MDDLogo. In addition, MDDLogo has been utilized to investigate known and novel substrate motifs of catalytic kinases in plant phosphorylation data. The identified motifs are used for the computational identification of plant phosphorylation sites with various substrate specificities. Based on a 5-fold cross-validation evaluation, predictive models trained with MDD-clustered subgroups show improvement with its prediction accuracy as compared with those models trained without the application of MDD clustering. This method, in addition to identifying the conserved motifs of identified and as yet unidentified kinases and functional domains, may also eventually be used as a tool to determine potential substrate specificity in proteins of interest.

2 METHODS

2.1 Maximal dependence decomposition

Due to the difficulty of detecting the conserved motifs for a large-scale sequence data, a new program, MDDLogo, is developed. MDDLogo applies MDD to cluster all sequences into subgroups that have statistically significant motifs. MDD was initially proposed to group splice sites during the identification process in splice site prediction (Burge and Karlin, 1997). Huang *et al*. (2005a, b) have incorporated MDD to identify kinase-specific phosphorylation sites; MDD is a recursive methodology which groups a set of aligned signal sequences in order to subdivide a large group into subgroups that capture the most significant dependencies between each position. Figure 1 presents the analytical flowchart of MDD clustering; in addition, a tree-like view of MDD clustering on plant phosphoserine (pSer) sequences is given in Supplementary Figure S1. An amino acid sequence is used as a sample for MDD clustering instead of a nucleotide sequence.

As shown in Figure 1, fragments of amino acids are extracted from plant pSer sequences using a window of length $2n+1$ that is centered on pSer sites. Since n is set to 10, the position Ai is selected from the range of −10 to +10. Then, a contingency table of the amino acids occurrence between two positions Ai and Aj is constructed. In order to extract motifs that have conserved biochemical property of amino acids when doing MDD, the twenty 20 types of amino acids is categorized into five groups: neutral, acid, basic, aromatic, and imino groups, as shown in Supplementary Table S1 (see Supplementary Materials). MDDLogo provides several grouping methods, as well as MDD clustering without the grouping of twenty 20 amino acids. Next, chi-square test $\chi^2(A_i,A_j)$ is adopted to evaluate the dependence of amino acid occurrence between two positions A_i and A_j surrounding the pSer sites. Assuming that each position in a set of aligned sequences is a sample of symbols (nucleotides or amino acids) generated according to the probability distribution, a null hypothesis is based on the assumption that the symbol distributions of two positions are mutually independent. The chi-square test is defined as:

$$\chi^2(A_i,A_j)=\sum_{m=1}^{5}\sum_{n=1}^{5}\frac{(X_{mn}-E_{mn})^2}{E_{mn}} \quad (1)$$

where X_{mn} represents the number of sequences that have the amino acids of group m in position A_i and have the amino acids of group n in position A_j, for each pair (A_i, A_j) with $i \neq j$. E_{mn} is calculated as $\frac{X_{mR}\cdot X_{Cn}}{X}$, where $X_{mR}=X_{m1}+\cdots+X_{m5}$, $X_{Cn}=X_{1n}+\cdots+X_{5n}$, and X denotes the total number of sequences. The two positions A_i and A_j are considered as dependent when their $\chi^2(A_i,A_j)$ value is larger than K. Herein, the value of K is 34.3, corresponding to a cutoff level of $\alpha=0.005$ with $(5-1)\times(5-1)=16$ degrees of freedom (Burge and Karlin, 1997). The testing of null hypothesis is defined as follows:

$$\begin{aligned}&\text{P (null hypothesis is rejected when two positions are dependent)}\\&=\text{P}\,(\chi^2(A_i,A_j)\geq K|\,\text{null hypothesis})=\alpha\end{aligned} \quad (2)$$

Each position A_i is calculated in order to obtain a dependence value, Score(A_i), which is defined as follows:

$$\text{Score}(A_i)=\sum_{j=-10}^{+10}\chi^2(A_i,A_j), j\neq i \quad (3)$$

The position, A_i, with a maximal dependence value of Score(A_i) is applied to cluster sequences, which is based on the amino acid group containing maximal chi-square value in position Ai. According to the distribution of amino acids in position A_i, the sequences containing an amino acid group with the maximal chi-square value are clustered into a subgroup, and the remaining sequences are clustered into another subgroup. MDD clustering is a recursive process that divides all sequences into tree-like subgroups. When applying MDD to cluster sequences, a parameter, i.e. the maximum cluster size, should be set. If the size of a subgroup is less than the cutoff value of maximum cluster size, the subgroup will not be divided any further. The MDD process terminates after all of the subgroup sizes are less than the value of the specified maximum cluster size. After MDD clustering, each subgroup is graphically visualized using WebLogo (Crooks *et al*., 2004) which generates the entropy plot of sequence logo. The conservation of sequences in each subgroup can then be easily examined.

Supplementary Figure S1 presents an example of MDD clustering on a set of *Arabidopsis Thaliana* phosphoserine (pSer) sequences collected from the TAIR database (Huala *et al*., 2001). Using all 2506 experimental pSer sequences, it can be observed that position +1 has the maximal dependence value with a statistically significant occurrence of imino amino acid,

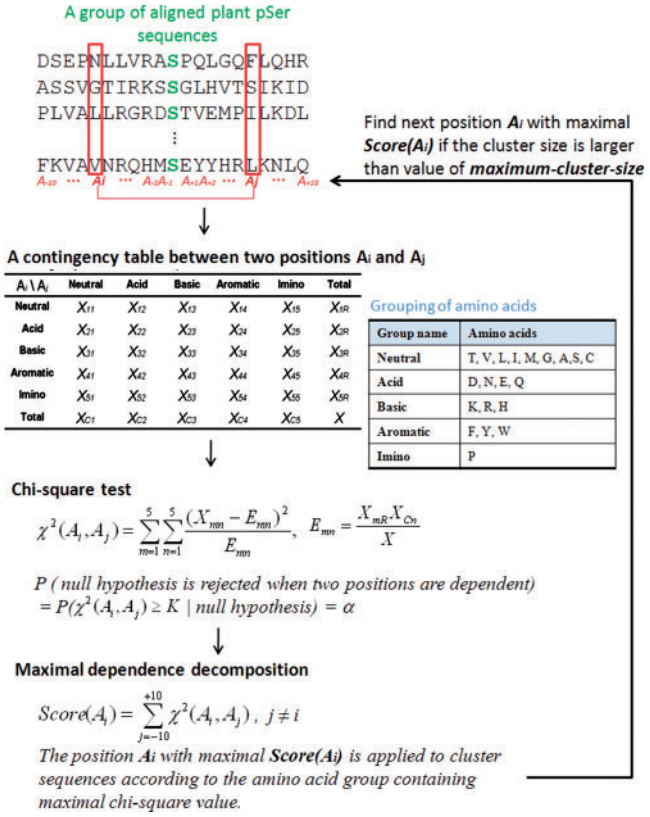


Fig. 1. The analytical flowchart of MDD clustering.

i.e. proline (P); Thus, all of the data can be divided into two subgroups: one subgroup (624 sequences) having the occurrence of imino amino acids in position +1 and the other subgroup (1882 sequences) not having the occurrence of imino amino acids in position +1. Because the value of maximum cluster size is set to 500, these two subgroups would be further clustered. In the left-hand tree, the subgroup of 624 sequences is further divided into two subgroups: one subgroup (236 sequences) having the occurrence of basic amino acids in position −3 and the other (388 sequences) which does not have the occurrence of basic amino acids in position −3. After this, further division of subgroups using MDD clustering is terminated in the left-hand tree because the data size of the resulting subgroups is now less than 500 sequences. With regard to the right-hand tree, the subgroup of 1882 sequences is further divided into two subgroups: one subgroup (786 sequences) having the occurrence of acid amino acids in position +3 and the other subgroup (1096 sequences) which does not have the occurrence of acid amino acids in position +3. The resulting subgroups in the right-hand tree contain over 500 sequences; thus, it is further clustered using MDD. Finally, the recursive MDD clustering is terminated when each subgroup in the leaf node has less than 500 sequences.

2.2 Construction of profile hidden Markov model

In this work, profile hidden Markov model (HMM) is applied to learn a predictive model from the sequences of each optimized MDD-clustered subgroup. An HMM describes a probability distribution over a potentially infinite number of sequences (Eddy, 1998). It can be used to detect distant relationships between amino acids sequences. Here, we use the software package HMMER version 2.3.2 (Eddy, 1998) to build the profile HMMs, to calibrate the HMMs and to search the putative functional sites against the protein sequences. HMM builds a model based on the positive instances of a class; thus, in this study, only positive data were considered to build a model. In the case of protein ubiquitylation sites, the MDD-clustered subgroups of all ubiquitylation data are taken as training sets to learn the HMMs. One HMM is built for each MDD-clustered subgroup.

For each model of the MDD-clustered subgroups, a threshold parameter is selected as a cutoff value in identifying potential positive data from a query (Eddy, 1998). Next, in order to search the hits of a HMM, HMMER returns both a bit score and an expectation value (E-value). The bit score is the base two logarithm of the ratio between the probability that the query sequence is a significant match and the probability that it is generated by a random model. The E-value represents the expected number of sequences with a score greater than or equal to the returned HMMER bit scores. A search with the HMMER bit score greater than the threshold parameter is taken as a positive prediction. While decreasing the bit score threshold favors finding true positives, increasing the bit score threshold favors finding true negatives.

2.3 Performance evaluation of predictive models

The predictive performance of the constructed models is evaluated by performing k-fold cross-validation (Ron, 1995). The original data are divided into k subgroups by splitting each dataset into k approximately equal sized subgroups. In each round of the cross-validation process, a subgroup is regarded as the test set, and the remaining $k-1$ subgroups are regarded as the training set. The cross-validation process is repeated k rounds, with each of k subgroups used as the test set in turn. Then, the k results are combined to produce a single estimation. The advantage of k-fold cross-validation is that all original data are regarded as both training set and test set, and that each data is used for test exactly once (Chen *et al*., 2010). In this study, the value of k is set to 5. Next, the following measures of predictive performance of the trained models are calculated: Precision (Pre) = TP/(TP + FP), Sensitivity (Sn) = TP/(TP + FN), Specificity (Sp) = TN/(TN + FP), and Accuracy (Acc) = (TP + TN)/(TP + FP + TN + FN), where TP, TN, FP and FN denote the number of true positives, true negatives, false positives and false negatives, respectively. Along with a 5-fold cross-validation evaluation, different values for the HMMER bit score were also tested in order to obtain the optimal threshold parameter for predicting query sequences. The HMMER bit score from the range of −10 to 0 was each tested as in order to obtain the best threshold value for each MDD-clustered subgroup. The performance of the models learned by HMMER given different threshold values of bit score is then compared.

3 UTILITY OF MDDLOGO

3.1 Identification of kinase-specific motifs in human phosphorylation sites

Protein phosphorylation catalyzed by kinases plays crucial regulatory roles in intracellular signal transduction (Hubbard and Cohen, 1993). Manning *et al*. (2002) have identified 518 human kinase genes, the so-called 'kinome', that provides a starting point for comprehensive analysis of kinase-specific substrate site specificity. Although mass spectrometry-based proteomics have enabled the increasing number of *in vivo* phosphorylation sites (Aebersold and Mann, 2003), only about 20% phosphorylation sites have the annotation of catalytic kinases (Lee *et al*., 2011). In order to fully investigate how protein kinases regulate the intracellular processes, it is necessary to accurately identify the substrate specificity of various kinases. This case study applies MDDLogo to identify the kinase-specific motifs for a large-scale data of protein phosphorylation sites in humans. A total of 1221 human phosphorylation sites involved with seven kinase groups are extracted from Phospho.ELM (Diella *et al*., 2008). Supplementary Table S2 shows the 232, 89, 280, 248, 226, 241 and 67 human phosphorylation sites for PKA, PKB, PKC, MAPK, CDC2, CK2 and ATM kinase groups, respectively, as well as their substrate-site motifs. Fragments of amino acids are extracted from 1221 phosphorylation sites using 11mer window length which is centered on phosphorylation sites.

Supplementary Figure S2 presents the user interface as well as the MDD-clustered results of 1221 human phosphorylation sites. As the grouping of amino acids is defined as given in Supplementary Table S1 and the maximum cluster size is set to 140 (Supplementary Fig. S2A), a total of 12 subgroups are obtained along with the number of sequences, the position with the maximal dependence value, property, chi-square score and sequence logo of MDD-clustered motif (Supplementary Fig. S2B). Moreover, a tree view of MDD clustering is presented in Figure 2. Using all 1221 sequences, it is observed that position +1 has the maximal dependence value with the statistically significant occurrence of imino amino acid, i.e. proline (P).;hus, all data can be divided into two subgroups: one subgroup (338 sequences) having the occurrence of imino amino acid in position +1 and another subgroup (883 sequences) which does not have the occurrence of imino amino acid in position +1. Because both subgroups contain more than 140 sequences (the value of maximum cluster size), these two subgroups are further clustered using MDD. The recursive MDD clustering is terminated after each subgroup in leaf node (marked in red box of Fig. 2) has less than 140 sequences. It is observed in Figure 2 that the sequences in the left-hand tree are associated with the proline-directed kinase groups, such as MAPK and CDC2. In the right-hand tree, the phosphorylation sites are mainly involved in basic (PKA, PKB and PKC) and acid (CK2 and ATM) kinase groups. To verify the results of MDD clustering, the 12 MDD-clustered motifs are

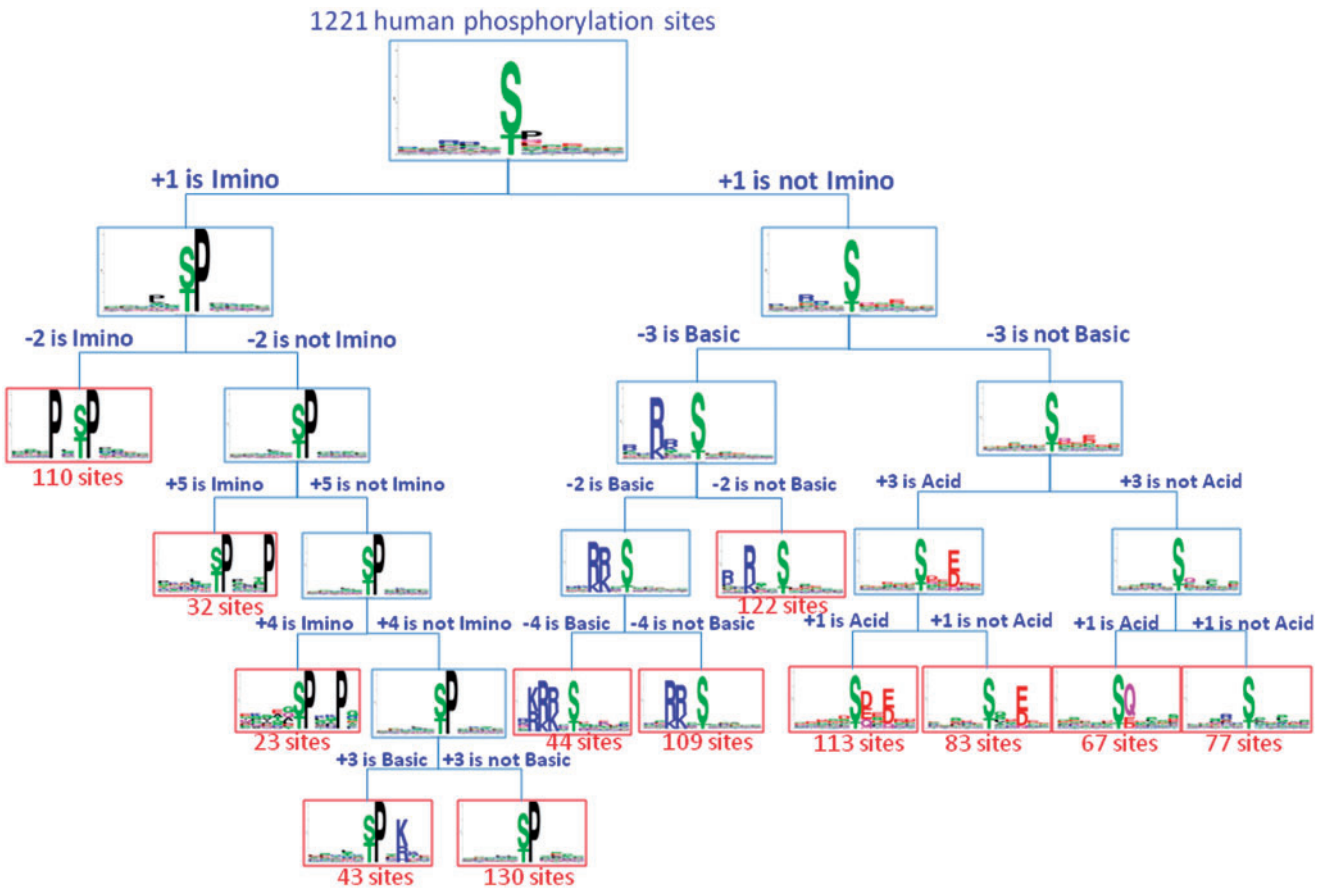


Fig. 2. A tree view of MDD clustering on 1221 human phosphorylation sites with 11mer sequence length.

mapped to known kinase-specific motifs from Phospho.ELM. As shown in Supplementary Table S3, group numbers 4, 5, 6, 8, 9 and 11 of MDD clustering are mapped to CDC2, MAPK, PKB, PKA, CK2 and ATM kinase groups of Phospho.ELM, respectively. In particular, the motif of Group 7 is similar to the substrate motif of PKC iota which is a subfamily of PKC group. Additionally, Group 1 is mapped to the ERK motif which is a subfamily of MAPK group. Interestingly, two novel MDD-clustered motifs (group number two and three) are slightly similar to two known kinase motifs in humans, PDK1 (Phosphoinositide-dependent protein kinase) and TTBK (Tau Tubulin Kinase). It follows from what has been demonstrated that MDDLogo could accurately identify the kinase-specific motifs from a large-scale phosphopeptides.

Given the difficulty in exploring conserved motifs in a large-scale sequence data, this work presents a new approach MDDLogo. MDDLogo applies MDD to cluster a group of aligned signal sequences into subgroups having statistically significant motifs. With regard to the method for extracting phosphorylation motifs, an iterative statistical approach, Motif-X (Schwartz and Gygi, 2005), has been proposed. To compare the results, Motif-X is also applied to detect the motifs from the 1221 human phosphorylation sites. With a significance value of 0.000001, a total of 19 motifs have been identified by motif-x, as shown in Supplementary Table S4. Table 1 presents a comparison of the detected motifs for the 1221 human phosphorylation sites using MDDLogo and Motif-X. Although Motif-X could identify the motifs that are matched to seven well-known kinase families, it is observed that several motifs need to be combined in order to match the substrate motif of a known kinase. For instance, Groups 1 (xxRRxSxxxxx) and 7 (xxRKxSxxxxx) resulting from Motif-X need to be combined to match the substrate motif (xx[R/K][R/K]xSxxxxx) of PKA group. The need for combining motifs also occurs in the PKB and the CK2 kinase groups. However, two motifs (Groups 15 and 16) that are not detected using MDDLogo are matched to the substrate motif of PKC eta, which is a subfamily of PKC group. The case study done using 1221 human phosphorylation sites involved with seven well-known kinase families demonstrates the effectiveness of MDDLogo.

3.2 Investigation of known motifs and novel motifs in plant phosphorylation sites

Protein phosphorylation is an important PTM that regulates various cellular processes not only in humans but also in plants. However, information regarding protein kinases that phosphorylate substrates in plants is very limited (Gao *et al.*, 2009). According to the collection of experimentally verified plant phosphorylation data from TAIR9 database (Huala *et al.*, 2001), each phosphorylation site is not annotated with its catalytic kinase. Thus, MDDLogo is adopted

Table 1. Comparison of the detected motifs for 1221 human phosphorylation sites between MDDLogo and Motif-X

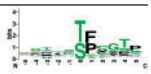
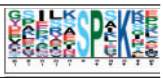
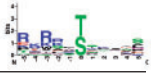
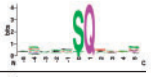
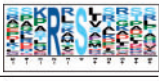
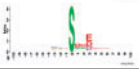
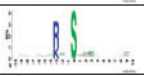
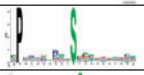
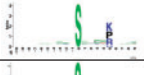
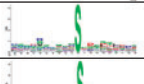
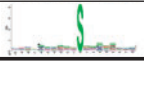
MDDLogo			Motif-X			Phospho.ELM	
Group number	Number of sequences	Motif	Group number	Number of sequences	Motif	Matched kinase	Entropy plot of kinase motif
1	110		2	64		ERK	
			17	41			
2	32		-	-	-	PDK1	
3	23		-	-	-	TTBK	
4	43		4	22		CDC2	
5	130		6	122		MAPK	
			18	73			
6	122		3	69		PKB	
			12	49			
			14	35			
			19	42			
7	44		-	-	-	PKC iota	
8	109		1	77		PKA	
			7	28			
9	113		5	30		CK2	
			10	20			
10	83		9	81		CK2	
			13	39			
11	67		11	65		ATM	
12	77		8	83		Aurora	
-	-	-	15	30		PKC eta	
-	-	-	16	21		PKC eta	

Table 2. The combined groups of MDD-clustered subgroups for 2506 sequences of phosphoserine (pSer) in plants

Group	Number of data	Entropy plot of motif	Matched kinase from Phospho.ELM	Entropy plot of kinase motif from Phospho.ELM
S1	624		MAPK	
S2	786		CK2	
S3	355		A novel motif	
S4	230		CAMK2	
S5	77		A novel motif	
S6	93		A novel motif	
S7	109		A novel motif	
S8	41		A novel motif	
S9	191		A novel motif	

to investigate the substrate specificity of various plant kinases based on amino acid sequences. A total of 2506, 378 and 108 experimental phosphorylation sites for serine (pSer), threonine (pThr) and tyrosine (pTyr), respectively, are extracted from TAIR9. To comprehensively explore the substrate specificity, a longer length of a window centered on the phosphorylation site is applied. The flanking amino acids ($-10\sim+10$) of the non-redundant phosphorylation sites are graphically visualized as entropy plots of sequence logo using WebLogo (Crooks, *et al*., 2004; Schneider and Stephens, 1990). As presented in Supplementary Table S5, the entropy plots indicate that pSer, pThr and pTyr have no conserved amino acids.

Applying MDDLogo resulted to 23 subgroups in pSer, six subgroups in pThr, and six subgroups in pTyr. The conservation of flanking amino acids in each MDD-clustered subgroup is represented using an entropy plot in the sequence logo. Based on these sequence logo representations, the resulting MDD-clustered subgroups could be further analyzed in order to combine subgroups with highly similar amino acid motifs. This process allowed us to categorize similar subgroups in pSer which resulted into nine groups (Table 2), while pThr (Supplementary Table S6) and pTyr (Supplementary Table S7) maintained six groups. However, subgroups S9, T6 and Y6 are observed to show no significantly conserved amino acid at any position. In order to show that each MDD-generated motif has potential kinase-specific substrate specificity, each MDD-generated motif is matched against the known kinase-specific motifs. To obtain known kinase-specific motifs, all available kinase-specific phosphorylation sites were obtained from Phospho.ELM (Diella *et al*., 2008), which is a well-known database containing experimentally kinase-specific phosphorylation data from multiple organisms. According to a chi-square test on the dependence of five amino acid groups in flanking positions of plant pSer, the most featured motif is the group that contains conserved proline (P) at position +1. This is matched to MAPK kinase family of non-plant organisms in Phospho.ELM. Furthermore, it is observed that S2 contains a conserved glutamic acid (G) and aspartic acid (D) at positions +1, +2 and +3. A similar conservation is seen in the CK2 kinase family of non-plant organisms in Phospho.ELM. Thus, the pSer group S2 is matched to CK2. Additionally, the conserved arginine (R) and lysine (K) at position -3 in the pSer group S4 is observed to be similar with the CAMK2 kinase family of non-plant organisms in Phospho.ELM; therefore, S4 is matched to CAMK2. Overall, as presented in Table 2, plant pSer groups S1, S2 and S4 are matched to kinase families MAPK, CK2 and CAMK2, respectively. The remaining pSer groups which contains significantly conserved amino acids at specific positions but have no matching kinases are considered to be novel motifs of potential kinases in plants.

Similar to the matches in pSer groups, pThr groups T1 and T3 are matched to kinase families MAPK and CK2, respectively for having similar conserved amino acids at the same position. Due to the insufficient data and unobvious kinase motifs of pTyr in Phospho.ELM, the pTyr groups in Table S7 have no matching kinases. Moreover, as observed in the unmatched pSer groups, the remaining groups in pThr and pTyr contains significantly conserved amino acids at specific positions, therefore, these are also considered to be novel kinase motifs in plants.

3.3 Improving the prediction of plant phosphorylation sites

With regard to the computational identification of plant phosphorylation sites, PhosPhAt (Heazlewood *et al*., 2008) has utilized a set of 802 experimentally validated pSer sites to develop a classifier of support vector machine (SVM) for identifying pSer sites in *Arabidopsis Thaliana*. More recently, Gao *et al*. (2009) incorporated protein sequence information and protein disordered regions, and integrated *k*-nearest neighbor and SVM for predicting phosphorylation sites. However, these two works mainly focus on substrate specificity as phosphorylation data in plant species are not well annotated in terms of kinase-specific substrate specificity. There is a need to investigate potential catalytic kinases in plants and utilize this information for predicting kinase-specific plant protein phosphorylation sites. Thus, MDD-clustered motifs are utilized.

According to a 5-fold cross-validation evaluation using an optimized threshold for the HMMER bit score, the predictive performance of HMM using MDD clustering performs significantly better than the single HMM of pSer and pThr. As shown in Table 3, the single HMM for pSer yields a precision rate of 49.5%, a sensitivity rate of 58.6%, a specificity rate of 70.0% and an accuracy rate of 66.2%. On the other hand, the performance of HMM for MDD-clustered pSer sites yield a precision rate of 71.3%, a sensitivity rate of 81.9%, a specificity rate of 81.9% and an accuracy rate of 81.9%. With regard to pThr, using a single HMM yields a precision rate of 45.4%, a sensitivity rate of 60.5%, a specificity rate of 63.4% and an accuracy rate of 62.5%. On the other hand, the performance of HMM for MDD-clustered pThr sites yield a precision rate of 64.7%, a sensitivity rate of 78.8%, a specificity rate of 78.4% and an accuracy rate of 78.5%.

In the case of pTyr, however, there is no significant improvement in MDD-clustered HMMs. Single HMM yields a precision rate of 75.4%, a sensitivity rate of 90.6%, a specificity rate of 84.7% and an

Table 3. The predictive performance of MDD-clustered models comparing with the model without MDD clustering

Method	Number of dataset			Pre	Sn	Sp	Acc
Single HMM	pSer	Positive data	2506	49.5	58.6	70.0	66.2
		Negative data	5012				
	pThr	Positive data	378	45.4	60.5	63.4	62.5
		Negative data	756				
	pTyr	Positive data	108	75.4	90.6	84.7	86.6
		Negative data	216				
MDD-clustered HMMs	pSer	Positive data	2506	71.3	81.9	81.9	81.9
		Negative data	5012				
	pThr	Positive data	378	64.7	78.8	78.4	78.5
		Negative data	756				
	pTyr	Positive data	108	75.5	92.6	83.8	86.7
		Negative data	216				

accuracy rate of 86.6%. On the other hand, the performance of HMM for pThr with MDD clustering yields a precision rate of 75.5%, a sensitivity rate of 92.6%, a specificity rate of 83.8% and an accuracy rate of 86.7%. This result may be due to an insufficient pTyr dataset used in this work. It is observed that MDDLogo could improve the identification of plant phosphorylation sites when the data size is sufficient.

4 CONCLUSION

MDDLogo has been demonstrated to accurately identify kinase-specific substrate motifs for 1221 human phosphorylation sites associated with seven well-known kinase families from Phospho.ELM. The ability of MDDLogo to discover motifs is comparable with another well-known motif discovery tool, Motif-X.

The importance of phosphorylation has been indicated in the regulation of protein functions and cell signaling in plants, but the state of research in this field is hindered by experimental difficulties, especially for the investigation of substrate specificity in various catalytic kinases. Due to the limitation in experimental data regarding protein kinases that phosphorylate substrates in plants, MDDLogo has been adopted to investigate the substrate specificity of various plant kinases. After the application of MDD clustering, a total of 9, 6 and 6 groups are detected for pSer, pThr and pTyr, respectively. Interestingly, some of the groups are similar to the substrate motifs of known kinase groups from the non-plant organisms in Phospho.ELM. Although the newly identified motifs could not be experimentally verified, it is important how MDDLogo can help biologists to further investigate the substrate specificity of potential catalytic kinases in plants. A more noteworthy benefit is that MDD-clustered motifs can be applied to improve the predictive power of computationally identifying plant phosphorylation sites with various substrate specificities, especially for pSer and pThr sites with a sufficient data size. Furthermore, the acquisition of additional experimentally verified plant phosphorylation data will allow the re-calibration of more robust MDD clusters.

In analyzing protein sequences, the grouping of amino acids can be defined by the users. The user can group amino acids according to various physicochemical properties. Basically, MDD clustering without the grouping of 20 amino acids is provided.;everal major grouping methods have been given. Also, in applying MDD clustering, it is important to optimize the value of the maximum cluster size. In general, MDD clustering could help users investigate the conserved motifs behind a large-sized sequence data. However, in some cases, more groups could produce more noise. Therefore, users need to optimize the value of the maximum cluster size to produce more meaningful clusters from a set of uncharacterized sequence data. MDDLogo can be freely accessed at http://csb.cse.yzu.edu.tw/MDDLogo/; a stand-alone program is available as well. DNA and RNA sequences are both allowed for use in MDDLogo.

Funding: The authors sincerely appreciate the National Science Council of the Republic of China for financially supporting this research under Contract Numbers of NSC 99-2320-B-155-001. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: none declared.

REFERENCES

Aebersold,R. and Mann,M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Beitz,E. (2006) Subfamily logos: visualization of sequence deviations at alignment positions with high information content. *BMC Bioinformatics*, **7**, 313.

Bindewald,E. *et al.* (2006) CorreLogo: an online server for 3D sequence logos of RNA and DNA alignments. *Nucleic Acids Res.*, **34**, W405–W411.

Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.

Chang,T.H. *et al.* (2008) RNALogo: a new approach to display structural RNA alignment. *Nucleic Acids Res.*, **36**, W91–W96.

Chen,S.A. *et al.* (2010) Incorporating significant amino acid pairs to identify O-linked glycosylation sites on transmembrane proteins and non-transmembrane proteins. *BMC Bioinformatics*, **11**, 536.

Crooks,G.E. *et al.* (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Diella,F. *et al.* (2008) Phospho.ELM: a database of phosphorylation sites–update 2008. *Nucleic Acids Res.*, **36**, D240–D244.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Fujii,K. *et al.* (2004) Kinase peptide specificity: improved determination and relevance to protein phosphorylation. *Proc. Natl Acad. Sci. USA*, **101**, 13744–13749.

Gao,J. *et al.* (2009) A new machine learning approach for protein phosphorylation site prediction in plants. *Lect. Notes Comput. Sci.*, **5462/2009**, 18–29.

Gorodkin,J. *et al.* (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res.*, **25**, 3724–3732.

Heazlewood,J.L. *et al.* (2008) PhosPhAt: a database of phosphorylation sites in Arabidopsis thaliana and a plant-specific phosphorylation site predictor. *Nucleic Acids Res.*, **36**, D1015–D1021.

Huala,E. *et al.* (2001) The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.*, **29**, 102–105.

Huang,H.D. *et al.* (2005a) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res.*, **33**, W226–W229.

Huang,H.D. *et al.* (2005b) Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites. *J. Comput. Chem.*, **26**, 1032–1041.

Hubbard,M.J. and Cohen,P. (1993) On target with a new mechanism for the regulation of protein phosphorylation. *Trends Biochem. Sci.*, **18**, 172–177.

Lee,T.Y. *et al.* (2011) RegPhos: a system to explore the protein kinase-substrate phosphorylation network in humans. *Nucleic Acids Res.*, **39**, D777–D787.

Manning,G. *et al.* (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.

Nevill-Manning,C.G. *et al.* (1998) Highly specific protein sequence motifs for genome analysis. *Proc. Natl Acad. Sci. USA*, **95**, 5865–5871.

Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics*, **14**, 55–67.

Ron,K. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, **2**, 1137-1143.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Schwartz,D. and Gygi,S.P. (2005) An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat. Biotechnol.*, **23**, 1391–1398.

Thompson,W. *et al.* (2003) Gibbs recursive sampler: finding transcription factor binding sites. *Nucleic Acids Res.*, **31**, 3580–3585.

Vacic,V. *et al.* (2006) Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*, **22**, 1536–1537.

Workman,C.T. *et al.* (2005) enoLOGOS: a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res.*, **33**, W389–W392.