BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 2 2015, pages 194–200
doi:10.1093/bioinformatics/btu598

Sequence analysis

Advance Access publication September 17, 2014

Proteomic analysis and prediction of human phosphorylation sites in subcellular level reveal subcellular specificity

Xiang Chen[1], Shao-Ping Shi[1,2], Sheng-Bao Suo[1], Hao-Dong Xu[1] and Jian-Ding Qiu[1,3,*]

[1]Department of Chemistry, Nanchang University, Nanchang 330031, [2]Department of Mathematics, Nanchang University, Nanchang 330031 and [3]Department of Materials and Chemical Engineering, Pingxiang College, Pingxiang 337055, P.R. China

Associate Editor: John Hancock

ABSTRACT

Motivation: Protein phosphorylation is the most common post-translational modification (PTM) regulating major cellular processes through highly dynamic and complex signaling pathways. Large-scale comparative phosphoproteomic studies have frequently been done on whole cells or organs by conventional bottom-up mass spectrometry approaches, i.e at the phosphopeptide level. Using this approach, there is no way to know from where the phosphopeptide signal originated. Also, as a consequence of the scale of these studies, important information on the localization of phosphorylation sites in subcellular compartments (SCs) is not surveyed.

Results: Here, we present a first account of the emerging field of subcellular phosphoproteomics where a support vector machine (SVM) approach was combined with a novel algorithm of discrete wavelet transform (DWT) to facilitate the identification of compartment-specific phosphorylation sites and to unravel the intricate regulation of protein phosphorylation. Our data reveal that the subcellular phosphorylation distribution is compartment type dependent and that the phosphorylation displays site-specific sequence motifs that diverge between SCs.

Availability and implementation: The method and database both are available as a web server at: http://bioinfo.ncu.edu.cn/SubPhos.aspx.

Contact: jdqiu@ncu.edu.cn

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

Received on April 5, 2014; revised on August 17, 2014; accepted on August 26, 2014

1 INTRODUCTION

Protein phosphorylation is a reversible post-translational modification (PTM) regulating major cellular processes such as cell division, growth and differentiation through highly dynamic and complex signaling pathways. Approximately one-third of proteins encoded by the human genome are presumed to be phosphorylated during their life cycle (Manning *et al.*, 2002; Olsen *et al.*, 2006). Mass spectrometry (MS) has been successfully used to identify protein phosphorylation in specific pathways and for global phosphoproteomic analysis (Boersema *et al.*, 2010). However, phosphoproteomic approaches do not evaluate the subcellular localization of the phosphorylated forms of proteins, which is an important factor for understanding the roles of protein phosphorylation on a global scale and the function of protein phosphorylations in regulating biological processes. Unfortunately, this understanding is limited by conventional MS technology to identify protein phosphorylations and their subcellular localization (Chan *et al.*, 2010).

Subcellular phosphoproteomics still represents a major analytical challenge, as only a few studies were reported over the past 5 years although efficient phosphopeptide enrichment methods were available since Zhou *et al.* (2010). This is in stark contrast to the large number of subcellular proteomics studies that have been reported over the same time period, reviewed by Brunet *et al.* (2003), Dreger (2003) and Yates *et al.* (2005). Several examples of subcellular phosphoproteomics were chosen to illustrate how this emerging field already uncovered important biological paradigms as shown by Trost *et al.* (2010). Although subcellular phosphoproteomics has the potential to uncover new regulatory pathways, the in-depth mapping of protein phosphorylation at the subcellular level and the further understanding of their biological significance require independent methods.

Computational methods for identifying phosphorylation sites have become increasingly popular, which can predict potential targets to significantly reduce the number of candidates that need to be verified by MS. A recent review by Trost and Kusalik (2011) described a comprehensive list of these methods. Most current predictors focus on organism-specific or kinase-specific phosphorylation sites, and such predictors do not take into account for specific subcellular compartment (SC). Every subcellular context is highly dynamic because the expressed proteins, their abundance and their post-translational modifications (including phosphorylation) depend on the physiological state of the cell (Hjerrild and Gammeltoft, 2006). Therefore, annotating the subcellular phosphoproteome is important as can be viewed from the following four aspects. (i) It can offer helpful clues or insights about their functions; particularly, one of the fundamental goals in proteomics and cell biology is to identify the functions of proteins in the context of a specific compartment. (ii) It can indicate in what kind of and how subcellular contexts the proteins interact with other molecules and with each other; this is particularly pivotal for the in-depth study of *in vivo* phosphorylation networks, one of the current hot topics in phosphoproteomics. (iii) It can help our understanding of the intricate phosphorylation pathways that regulate biological processes at the subcellular level (Ehrlich *et al.*, 2002; Glory and Murphy, 2007), and hence, it is indispensable for many studies in system

*To whom correspondence should be addressed.

biology. (iv) It is extremely useful for identifying and prioritizing drug targets (http://www.biocompare.com/Editorial-Articles/41619-subcellular-targeting-of-proteins-and-drugs/) during the process of drug development.

To efficiently accelerate development of the highly complex subcellular phosphoproteomic, an integrated platform combining experimentally data querying and unknown data annotation is highly demanded. Here, we developed a platform that provides both a searchable online database and a computational tool to efficiently and reliably accumulate the subcellular phosphoproteome for further experimental investigation. In this work, we report the most thorough characterization of subcellular phosphoproteome in humans to date. Originally, reliable experimental phosphoproteomic data with verified information of subcellular localization in humans were collected from several sources and used to profile subcellular phosphoproteome. Not only do we find that most phosphorylation proteins are uniquely resided in a specific SC, but also we show that the distribution of phosphorylated proteins in SCs is compartment-specific. Functional enrichment analysis and protein–protein network analysis reveal that the phosphorylation signaling pathways of SCs have higher specialization. Moreover, our large datas*et al*lows us to delineate type-specific phosphorylation sequence motifs contrary to general phosphoproteome, and we show that there are sequence motifs of specific SCs. Overall, our observations highlight compartment-specific phosphorylation signaling pathways, which stress the importance of mapping protein phosphorylation in the physiologically relevant SC. Later, we developed a bioinformatics tool termed SubPhosPred, which combines a novel discrete wavelet transform (DWT) algorithm with a support vector machine (SVM) approach to identify phosphorylation sites for different SCs in humans. As one innovative character of our method, the most attractive character of wavelet transform is the ability to elucidate simultaneously both the spectral and temporal information (Mori *et al.*, 1996) that was used for encoding as features for PTM prediction. Cross-validation tests show that the DWT algorithm can boost predictive performance and obtain encouraging prediction results for each compartment. Additionally, the independent test demonstrates that the proposed method outperforms Musite (Gao, 2010) when the customized models use the same training datasets as SubPhosPred. For SubPhosPred, we have trained eight compartment-specific phosphorylation prediction models [cell membrane (CM), nucleus (NU), cytoplasm (CY), mitochondrion (MI), Golgi apparatus (GA), endoplasmic reticulum (ER), secretion (SE) and lysosome (LY)] using datasets from our database (SubPhosDB). Finally, the platform-integrated SubPhosDB database and SubPhosPred predictor are freely available for academic research at: http://bioinfo.ncu.edu.cn/SubPhos.aspx.

2 METHODS

2.1 Data collection

Phosphorylation data for *Homo sapiens* from several sources including UniProt/Swiss-Prot (version 55.0), Phospho.ELM (version 8.0), PHOSIDA (version 1.0), HPRD (version 7.0) and PhosphoSite (9-Oct-2012) were collected as shown in Supplementary Table S1. After removing the redundant data among these databases, the data contain 137 153 experimental verified phosphorylation sites within 17 297 phosphoproteins. Furthermore, the data pertaining to subcellular localization were extracted from the UniProt/Swiss-Prot database released on October 9, 2012. Sequence annotated as ambiguous or uncertain localization terms (such as "potential", "probable", "probably", "maybe", or "by similarly") were excluded, where 10 265 phosphorylated proteins with experimental verified information of subcellular localization were obtained for different SCs. In addition, the experimental verified localization information of corresponding kinases was also extracted. The statistical result of the numbers of phosphorylation proteins and corresponding kinases for different SCs was listed in Supplementary Table S2. We integrated these datasets as a free online database termed SubPhosDB for the biological research community.

2.2 Building the classifier

2.2.1 SVM learning As a machine-learning method of binary classification, SVM aims to find a regulation that best maps each member of a training set to the correct classification (Vapnik, 1999), and SVM has been used for a variety of classification/prediction tasks relating to protein bioinformatics. Using the feature encoding of phosphorylated sequence, the SVM was trained to distinguish phosphorylation and non-phosphorylation sites for different SCs. The implemented SVM algorithm was LIBSVM (A library for support vector machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm), and the applied kernel function was the radial basis function. To maximize the performance of the SVM algorithm, the grid search method was applied to tune the parameters.

2.2.2 Training sets As previously described (Trost and Kusalik, 2011), the experimentally verified phosphorylation sites were regarded as positive data, whereas all the other non-phosphorylated serine/threonine/tyrosine (S/T/Y) residues were taken as negative data, respectively (Supplementary Table S3 and Supplementary Ep1). In machine-learning problems, imbalanced datasets occur when one class has a significantly different number of instances than another class and can significantly affect the accuracy of some learning methods (Japkowicz and Stephen, 2002). In the context of phosphorylation site prediction, positive data are vastly outnumbered by negative data. To correct this imbalance, for each compartment as well as their each site type, the number of positive sites was determined, and an equal number of negative sites were randomly chosen from the negative training data.

2.2.3 Features and DWT Local sequence clusters (LSC) often exist around phosphorylated sites because the sites of the same kinase family or kinase often share similar patterns in local sequences (Kennelly and Krebs, 1991). Additionally, amino acid pair compositions (AAPC) could reflect the characteristics of the residues surrounding phosphorylated sites, and it has been successfully used for predicting phosphorylation sites (Zhao *et al.*, 2012). Therefore, we took into account similarity scores and amino acid pair compositions of the phosphorylated sequence to convert these training sets into numerical series. The detailed procedures of feature representation are described in the Supplementary Ep2. After obtaining the numerical sequences of training sets, the feature wavelet coefficients of each query sequence were extracted by using the DWT algorithm to optimize each feature (Lu *et al.*, 2004). Over the past several years, we developed a series of DWT algorithms mainly for the prediction of protein function (Qiu *et al.*, 2009; Shi *et al.*, 2011). In this work, we first refined the DWT algorithm for the prediction of the PTM sites in proteins (the calculation procedures are described in the Supplementary Ep3). To evaluate the stability of each feature, 10 training sets were constructed by selecting randomly 10 times for negative samples to match the positive ones in the training sets.

2.2.4 Performance evaluation We first developed a predictor for the prediction of phosphorylation site in a specific subcellular proteome.

Therefore, it is difficult to compare it with other existing tools. Interestingly, Gao *et al.* (Gao, 2010) presented a novel software tool known as Musite that provides a unique functionality for training customized prediction models from users' own data. Hence, we should use the customized model from Musite to predict phosphorylation sites for further evaluating the performance of SubPhosPred. The comparing method and evaluation criteria are expatiated in Supplementary Ep4.

2.3 Functional enrichment analysis

Gene Ontology (GO) and pathway enrichment analysis were performed using the functional annotation tool of the DAVID bioinformatics resources (Huang da *et al.*, 2009) (http://david.abcc.ncifcrf.gov/home.jsp). According to the two-sided category of Fisher's exact test, a *P*-value <1.00E-2 (adjusted for multiple comparisons) was considered statistically significant. Enriched terms were sorted by *P*-value. To show diverse processes enriched in our data, redundant or highly similar terms were removed.

2.4 Phosphorylation network analysis

Phosphorylation network analysis was performed using protein interaction data from the STRING database (Franceschini *et al.*, 2013) (http://string-db.org/). Only interactions with a score >0.7 were represented in the networks. Cytoscape version 2.8 (Shannon *et al.*, 2003) was used for visualization of protein interaction networks. NetworkAnalyzer plug-in for Cytoscape software (Assenov *et al.*, 2008) was used to calculate the topological parameters of the subcellular phosphorylation networks.

2.5 Sequence motif analysis

It is well known that phosphorylated sites are more conserved than non-phosphorylated sites, and there are many sequence motifs in the vicinity of the phosphorylated sites (Olsen *et al.*, 2006). To investigate whether sequence motif exists for specific subcellular phosphoproteome, we performed an enrichment analysis of short-linear motifs for every dataset of SCs using the Motif-x software (Schwartz and Gygi, 2005) (http://motif-x.med.harvard.edu/). Default parameters were used for this analysis.

3 RESULTS

3.1 SC distribution of phosphorylation in humans

Phosphorylation proteome distribution. Each SC is irreplaceable in cells. Substantial evidence has confirmed that subcellular phosphoproteomes play an essential role in a variety of cellular processes via phosphorylation-mediated signaling transduction (Chan *et al.*, 2010; Rindress *et al.*, 1993; Trost *et al.*, 2010). To investigate the distribution of phosphorylation in humans across SCs, we evaluated the phosphorylated proteins as well as corresponding kinases according to their SC annotations. As mentioned earlier, we compiled the subcellular phosphoproteome database that has collected 10 265 experimentally phosphorylated proteins and corresponding 340 kinases from several public databases. After initial statistics of these data about SC distribution of phosphorylation, the results are shown in Figure 1. Both phosphorylated proteins and corresponding kinases are unevenly found across SCs that mostly distribute in common SCs. The majority of kinases reside in the CM, the CY and the NU, with each one containing >30% of all kinases. The MI, GA and ER account for ~5% of all kinases, and the other SCs, including cytoskeleton, cell junction, synapse, peroxisome, centrosome, microsome, melanosome and so on, harbor ~30%. Interestingly, also for phosphorylated proteins, this proportion is almost consistent with the corresponding protein kinases (Fig. 1A and B). At the global level, we also analyzed the number of different distributions of SCs for all phosphorylated proteins and corresponding kinases from our compiled database. As a result, phosphorylated proteins mostly reside in one SC, where ~64% of phosphorylated proteins are seen in a unique SC. Instead, there are ~65% of the kinases that were observed in more than one SC (Fig. 1C and D). Meanwhile, we analyzed the distributions of the phosphorylated proteins and kinases across SCs. Although the proportion of unique localization slightly varies in SCs, the general tendency of both global cell level and subcellular level was identical (Fig. 1E and F). This result clearly indicates that kinases are involved in a broad array of physiological functions. But as evident from Figure 1F, the phosphorylated proteins have higher specialization than the corresponding kinases for phosphorylation signaling pathways in specific SCs.

Additionally, for two different levels, all phosphorylated proteins in compartments and unique phosphorylated proteins in one compartment, we computed the average number of phosphosites observed in eight SCs and the global cell (Fig. 1G). At these two levels, we identified an average of 7.15 or 11.94 phosphosites, respectively, per protein in the global cell, but the number greatly varies across different SCs. SE exhibits the lowest number of phophosites with an average of 1.16 or 4.25 sites per protein. This may explain why many previous studies rarely identified phosphorylation sites from SE, although this is not the compartment with the lowest number of phosphorylated proteins (Fig. 1B). Also, the average number of different SCs or the global cellular compartment observably separates the two different levels. For example, LY proteins contain on average 22.85 phophosites from all LY compartments, and only 3.13 phophosites are identified from unique proteins in LY. These results illustrate that subcellular phosphorylation distribution is compartment type dependent and possesses relatively its own phosphorylation signaling network.

To test whether the tendency toward a specific signaling network in the SC is relatively independent by diversity protein abundance, we estimated the relative functional enrichment for all phosphorylated proteins using the DAVID bioinformatics resources (see Supplementary Table S4). Also, we created the interaction networks observed in phosphorylated protein in the different compartments using Cytoscape software, and determined their four topological measures using the NetworkAnalyzer plug-in (see Supplementary Ep5). As expected, the analysis of the GO annotation between the phosphorylated proteins found in the different compartments clearly indicates that different GO biological process and molecular function subcategories are enriched in different SCs (*P*-value < 1.00E-6, Supplementary Fig. S1). Analyzing the phosphorylation signaling pathway per SC also revealed similar results (*P*-value < 1.00E-4, Supplementary Fig. S1). Moreover, compartment-specific phosphorylated interaction networks show uneven clustering features among themselves or compared with random phosphorylated networks in humans (Supplementary Fig. S2).

Phosphorylation type distribution. There are three major protein kinase types, including serine, threonine and tyrosine kinases (STYKs), which exist in different SCs to regulate

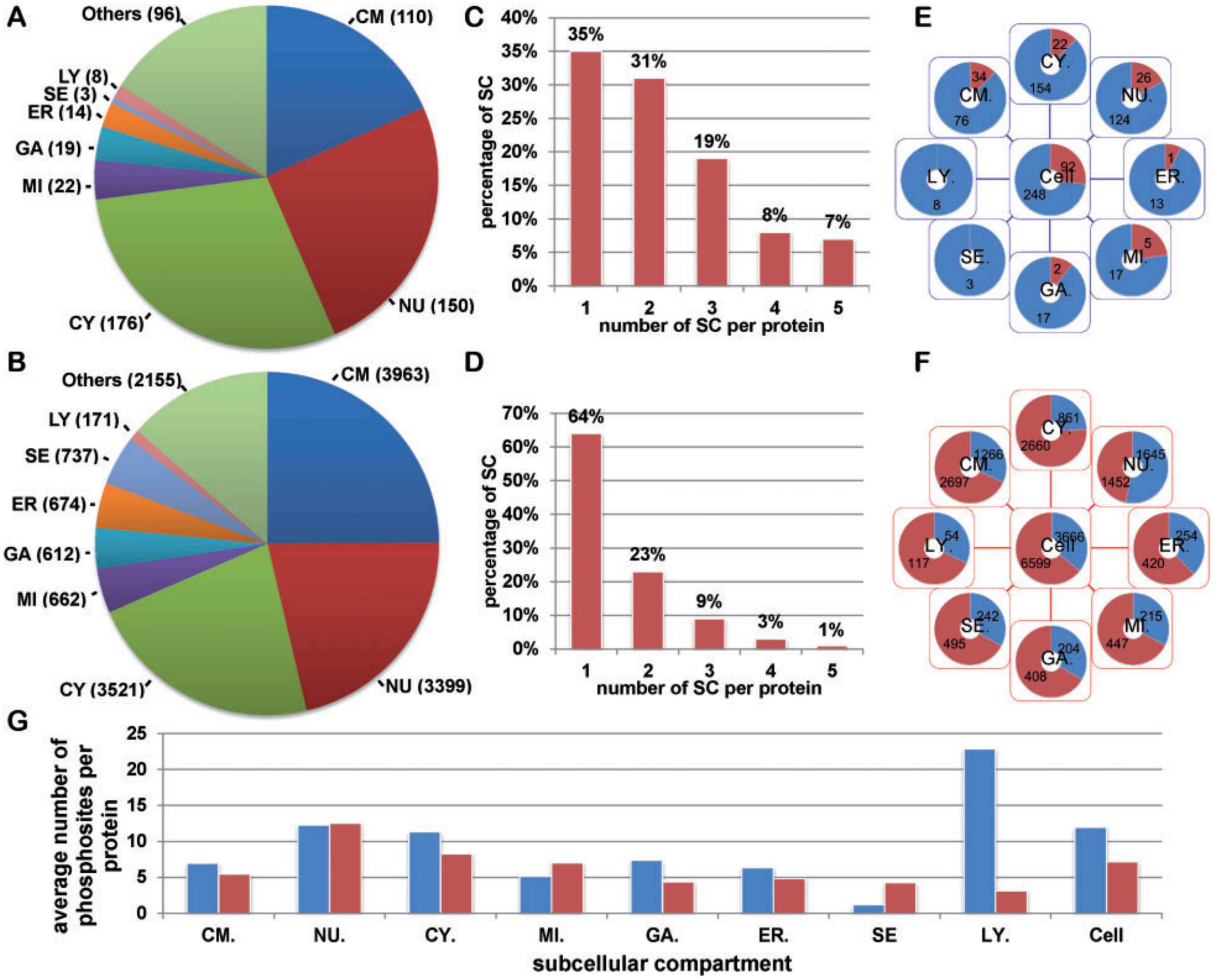

Fig. 1. Statistics of subcellular phosphoproteome in humans. (**A**) The pie chart represents the SC distribution of the kinases in humans. (**B**) The pie chart represents the SC distribution of phosphorylated proteins in humans. (**C**) The histogram represents the distribution of the number of SCs per kinases. (**D**) The histogram represents the distribution of the number of SCs per phosphorylated protein. (**E**) The ring chart represents the distribution of unique kinases (red fraction) relative to all kinases residing in a specific SC. (**F**) The ring chart represents the distribution of unique phosphorylated proteins (red fraction) relative to all phosphorylated proteins residing in a specific SC. (**G**) The histogram represents the average number of phosphosites per phosphorylated proteins in humans (the red/blue represents the proteins uniquely/globally resided in a specific SC)

phosphorylation signaling networks and control subcellular activities. A number of kinases can often phosphorylate different residues, while some other kinases phosphorylate only one unique type of residue (Fig. 2A). It is reported in Figure 2A that the overlap between serine and threonine kinases is far greater than between serine and tyrosine kinases and between threonine and tyrosine kinases. Another report of the phylogenetic relationship among the non-overlapping STYKs from Figure 2A also suggests that tyrosine kinases are distinguished from both serine and threonine kinases, as they are evolutionarily conserved (Fig. 2B). Actually, different STYKs play various roles in phosphorylation signaling networks that influence catalysis, subcellular localization, regulation and other functions of target proteins (Supplementary Fig. S3). Hence, it is necessary to investigate the SC distributions of both STYKs and phosphorylated residue types for further observing the specificity of subcellular phosphoproteome. Our investigations showed that phosphorylated protein increases approximately as the number of kinases increases for each compartment except for the SE compartment that drifts away anywhere in the cell (Fig. 2C). This investigation likely reflects that compartment-specific phosphorylation regulates predominantly by compartment-specific kinases. It also may explain why the previous survey observed functional specificity from the compartment-specific phosphoproteome (Fig. 3), although the proteome resides exactly in the cell. In addition, an interesting observation is that as the subcellular STYKs distribute variously, a similar distribution of phosphorylated residue type is displayed for the different SCs (Fig. 2C). The result likely reflects that the kinase resided in different SCs and plays the different extent of phosphorylated

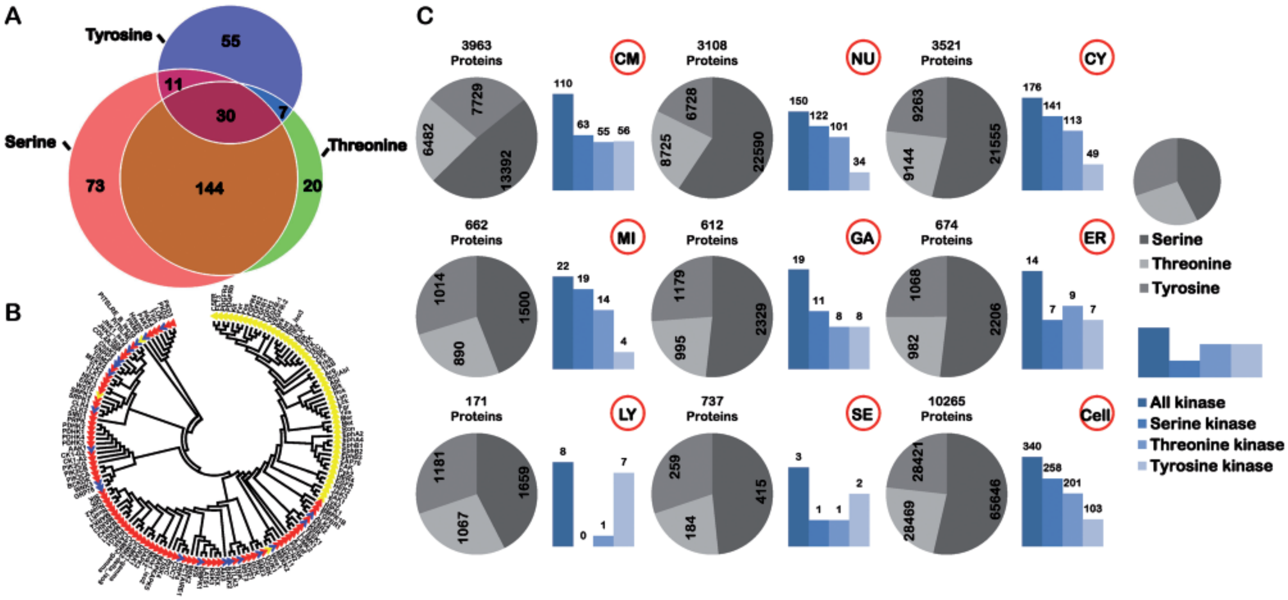

Fig. 2. Distribution of phosphorylation types in human cell. (**A**) Overlap across kinases of serine, threonine and tyrosine. Numbers in the Venn diagram represent the number of kinases in the fragment. (**B**) Phylogenetic analysis of the non-overlapping kinases of serine, threonine and tyrosine from Figure 2A. The trilateral represents kinases of serine (red), threonine (yellow) and tyrosine (blue). (**C**) The pie diagram shows the number of phosphorylated sites on serine, threonine and tyrosine; the histogram shows the number of kinases of serine, threonine and tyrosine; the numbers above per diagrams represent the number of phosphorylated proteins

role, and this extent is compartment-specific. For example, the fluctuant distributions of tyrosine kinase do not significantly impact the distributions of phosphorylated type.

3.2 Phosphorylation sequence motifs are specific for SCs

It is well known that phosphorylation has clear sequence motifs. To investigate whether the specificity of a phosphorylation motif exists in different compartments, we estimated the sequence preferences for compartment-specific phosphoproteome. Before the compartment-specific analysis, heatmap analysis of all identified phosphorylation sites reveals general preferences for amino acid residues surrounding the phosphorylated sites (Fig. 3A). As also previously reported (Gao, 2010), we find that the amino acids Pro, Arg, Asp, Glu, Ser, Lys and Gly are enriched in the surrounding sequences for phosphorylated serine and threonine sites, whereas Cys, Trp, Tyr, Phe, Ile, Met, Leu, His, Thr and Val are depleted. For phosphorylated tyrosine sites, Asp, Glu, Pro, Ser and Gly are enriched, whereas Trp, Cys, Phe, Leu, His, Met and Ile are depleted. All phosphorylated types preferentially occur in serine-rich regions, with a tendency toward neutral residues in the immediate surroundings of the modified site.

Additionally, position-specific preferences include Pro residues at amino acid position +1 relative to the phosphorylated serine and threonine residues and Ser and Leu residues at amino acid positions +1 and +3 relative to phosphorylated tyrosine residue. We next investigated whether these sequence preferences are similar across different SCs. To visualize compartment-specific sequence motifs, the Motif-x software (Schwartz and Gygi, 2005) was used to explore compartment-specific datasets for analyzing the 12 residues flanking the modified site for overrepresentation of specific amino acids relative to the human phosphoproteome background distribution (P-value < 1.00E-5). This analysis reveals that the sequence motifs differ for SCs (Fig. 3B). The preference for Pro residues at amino acid position +1 is general to phosphorylated serine and threonine in most compartments, but on CM proteins, there are many clearly preferences, such as the motif Rxx[S]P (where S is the phosphorylated site and x can be any residue, and the motif is preferentially phosphorylated by a known Pro-directed kinase enriched in CM). Mitochondrial proteins have a preference for hydrophilic residue (Lys and Gly) in the upstream of the phosphorylated tyrosine, but they additionally show a slight preference for hydrophobic Ala residue at position –4. Proteins that reside in either SE/GA also have the general preference for hydrophilic residues at positions in the vicinity of the phosphorylated sites. The most distinct sequence motif is evident on NU proteins, where there are many strong preferences for serine-rich regions in positions –4 to +4, such as RSx[S]xS, RSx[S]P and so on. Because a number of compartment sequence motifs differ substantially from the previously reported motif for phosphorylation (Chen *et al.*, 2011), the guess about compartment-specific phosphorylation is proven true. It seems hardly surprising that some kinases reside solely in a specific compartment (Fig. 1). All identified compartment-specific sequence motifs are summarized in Supplementary Table S5 as a source.

3.3 Development of SubPhosPred for predicting phosphorylation sites of subcellular proteomes

The importance of compartment-specific mapping of post-translational modifications of proteins is underscored by the

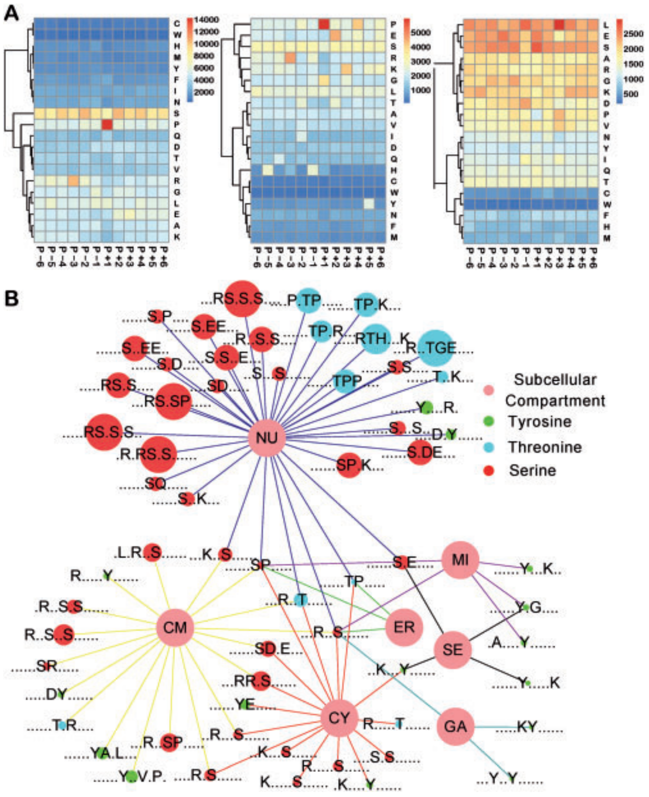

Fig. 3. Sequence motifs of phosphorylated sites. (**A**) Heatmap indicating preference of amino acids in positions –6 to +6 from phosphorylated serine (left), threonine (center) and tyrosine (right) in human proteins. (**B**) It shows across networks of sequence motif (*P*-value < 1.00E-5) found in compartment-specific phosphorylated sequence. The node size increases with the score calculated by Motif-x of sequence motif increase. The node color represents phosphorylated serine (red), threonine (blue) and tyrosine (green)

substantial distinctions we find for phosphorylation specificity and patterns across SCs. Based on this observation and state-of-the art machine-learning principles, we presented a novel tool termed SubPhosPred which specifically designed for compartment-specific phosphorylation site predictions. We sorted human phosphoproteomics data from SubPhosDB into multiple SCs and used them to train prediction models by an SVM learning approach that integrates DWT algorithm and two feature extractions (AAPC and LSC). For the performance optimizing of the model construction, the detailed processes and results are illustrated in Supplementary Ep6.

To evaluate the performance of SubPhosPred, the cross-validation process was performed for all SCs. Here we performed test of 10-fold cross-validation on each type of phosphorylation for 10 training sets of each compartment. We then calculated the values of corresponding evaluation criteria for each training set as shown in Supplementary Table S6 and plotted the receiver-operating characteristic curves as shown in Supplementary Figure S4. The results show in SubPhosPred that the coupled features used in the DWT algorithm yield more accurate predictions as expected.

3.4 Comparison with other prediction tools

As mentioned in the ‘Methods Compared’ section, the performance of SubPhosPred was further evaluated by comparing a novel tool Musite with an independent test. As sufficient training data are required for training a model in Musite, only three SCs (CM, NU and CY) tests were performed to make a comparison as shown in Supplementary Table S7. Musite-1 (tested using the human general prediction model from the pre-trained models in Musite), Musite-2 (tested using the customized prediction model of CM, NU and CY) and our method exhibited satisfying performance, but our method SubPhosPred has greatly improved for all thresholds (high, medium and low). Moreover, an expected observation from Supplementary Table S7 is that Musite-2 has slightly better prediction performance than Musite-1. This observation illustrates that the classification performance of the model trained using correlated proteins in a specific subcellular context outperformed the model trained using proteins in a general context.

4 DISCUSSION

Although SCs share a partly independent phosphorylation network according to our analysis, the protein composition of a specific compartment is not static and undergoes dynamic changes following interactions with other SCs. Statistical results from kinase and phosphorylated protein data have exhibited coincident results that the kinases and the phosphorylated proteins concurrently resided in different SCs (Supplementary Fig. S5, for the detailssee Supplementary Tables S8 and S9). Interestingly, the co-localization distribution is similar between phosphorylated proteins and kinases (Supplementary Fig. S5). This may explain why we can identify phosphorylation sites in different compartment-specific models by using the SubPhosPred predictor. In addition, for the phosphorylation cross talk across SCs, there are still at least three possible explanations for this: (i) all kinases are synthesized in the CY and may phosphorylate CY proteins before entering various SCs, (ii) kinases from a particular SC may have access to substrates from other SCs during mitosis when the subcellular membrane is absent and (iii) many kinases may dissociate between SCs. This is exemplified by Jnk1 (MAPK10) and Jnk2 (MAPK9), which, on activation, translocate from the CY to the NU or the perinuclear region (Mizukami *et al.*, 1997; Whitmarsh *et al.*, 2001). Despite diverse reasons, the predominantly phosphorylated cross talk occurs because of the co-localizations between kinases and phosphorylated proteins. Hence, this does not affect compartment-specific network independence itself. An observation for comparing the phosphorylation network in NU with in the cell was shown in Supplementary Figure S6A. According to the topological calculation of networks, the network parameters including the average number of neighbors, the network centralization and the network density clearly reveal that the NU network is highly connected with stronger robustness against the network in human cell. It means that compartment-specific phosphorylation subnetwork is self-governed in nature, which would explain why the phosphorylation prediction on subcellular context significantly leads to precision improvement (for SubPhosPred *P*-value < 3.21E-06, for Musite *P*-value < 5.01E-02, see

Supplementary Fig. S6B). Actually, the independence of the network between SCs has become a common view that widely applies to constructing the training set for the prediction of protein–protein interaction (Jansen *et al.*, 2003; Rhodes *et al.*, 2005). In summary, these results again underscore the specificity of a compartment-specific network (kinase–substrate interaction network).

For several years, phosphoproteomics has moved far beyond a simple catalog of phosphorylation sites and is contributing to important cell biology discoveries by unveiling the dynamic changes in protein phosphorylation regulating numerous cellular functions. Subcellular phosphoproteomics also has enormous potential to uncover new regulatory pathways (Kislinger *et al.*, 2006), while the verification of these new findings or the further learning of their biological significance often requires independent methods. Annotation of phosphorylation sites in intact subcellular proteomes as pivotal step would further advance our understanding of compartment-specific phosphorylation. To this end, we present first a novel platform for annotating subcellular phosphoproteome in humans. Although it does not directly address these issues, our present work provides a foundation for subsequent studies by demonstrating effective methods for large-scale multi-compartment surveys of phosphorylation. Furthermore, this phosphoproteomic profiling can also serve as a basis of comparison to explore changes in phosphorylation that occur in many physiological and pathological states.

ACKNOWLEDGEMENT

The authors thank A. Burlingame for suggestions for this article.

Funding: This work was supported by Program for New Century Excellent Talents in University (NCET-11-1002); and the National Natural Science Foundation of China (21305062, 21175064).

Conflict of interest: none declared.

REFERENCES

Assenov,Y. *et al.* (2008) Computing topological parameters of biological networks. *Bioinformatics*, **24**, 282–284.

Boersema,P.J. *et al.* (2010) In-depth qualitative and quantitative profiling of tyrosine phosphorylation using a combination of phosphopeptide immunoaffinity purification and stable isotope dimethyl labeling. *Mol. Cell. Proteomics*, **9**, 84–99.

Brunet,S. *et al.* (2003) Organelle proteomics: looking at less to see more. *Trends Cell Biol.*, **13**, 629–638.

Chan,L.-S. *et al.* (2010) Differential phosphorylation of dynamin I isoforms in subcellular compartments demonstrates the hidden complexity of phosphoproteomes. *J. Proteome Res.*, **9**, 4028–4037.

Chen,Y.-C. *et al.* (2011) Discovery of protein phosphorylation motifs through exploratory data analysis. *PLoS One*, **6**, e20025.

Dreger,M. (2003) Subcellular proteomics. *Mass Spectrom Rev.*, **22**, 27–56.

Ehrlich,J.S. *et al.* (2002) Spatio-temporal regulation of Rac1 localization and lamellipodia dynamics during epithelial cell-cell adhesion. *Dev. Cell*, **3**, 259–270.

Franceschini,A. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**, D808–D815.

Gao,J. (2010) Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol. Cell. Proteomics*, **9**, 2586–2600.

Glory,E. and Murphy,R.F. (2007) Automated subcellular location determination and high-throughput microscopy. *Dev. Cell*, **12**, 7–16.

Hjerrild,M. and Gammeltoft,S. (2006) Phosphoproteomics toolbox: computational biology, protein chemistry and mass spectrometry. *FEBS Lett.*, **580**, 4764–4770.

Huang da,W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Jansen,R. *et al.* (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.

Japkowicz,N. and Stephen,S. (2002) The class imbalance problem: a systematic study. *Intell. Data Anal.*, **6**, 429–449.

Kennelly,P.J. and Krebs,E.G. (1991) Consensus sequences as substrate-specificity determinants for protein-kinases and protein phosphatases. *J. Biol. Chem.*, **266**, 15555–15558.

Kislinger,T. *et al.* (2006) Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell*, **125**, 173–186.

Lu,X.Q. *et al.* (2004) Maximum spectrum of continuous wavelet transform and its application in resolving an overlapped signal. *J. Chem. Inf. Comp. Sci.*, **44**, 1228–1237.

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

Mizukami,Y. *et al.* (1997) A novel mechanism of JNK1 activation – nuclear translocation and activation of JNK1 during ischemia and reperfusion. *J. Biol. Chem.*, **272**, 16657–16662.

Mori,K. *et al.* (1996) Prediction of spalling on a ball bearing by applying the discrete wavelet transform to vibration signals. *Wear*, **195**, 162–168.

Olsen,J.V. *et al.* (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, **127**, 635–648.

Qiu,J.-D. *et al.* (2009) Using support vector machines for prediction of protein structural classes based on discrete wavelet transform. *J. Comput. Chem.*, **30**, 1344–1350.

Rhodes,D.R. *et al.* (2005) Probabilistic model of the human protein-protein interaction network. *Nat. Biotechnol.*, **23**, 951–959.

Rindress,D. *et al.* (1993) Organelle-specific phosphorylation - identification of unique membrane phosphoproteins of the endoplasmic-reticulum and endosomal apparatus. *J. Biol. Chem.*, **268**, 5139–5147.

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

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Shi,S.-P. *et al.* (2011) Identify submitochondria and subchloroplast locations with pseudo amino acid composition: approach from the strategy of discrete wavelet transform feature extraction. *Biochim. Biophys. Acta*, **1813**, 424–430.

Trost,B. and Kusalik,A. (2011) Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics*, **27**, 2927–2935.

Trost,M. *et al.* (2010) Subcellular phosphoproteomics. *Mass Spectrom Rev.*, **29**, 962–990.

Vapnik,V.N. (1999) An overview of statistical learning theory. *IEEE T. Neural. Networ.*, **10**, 988–999.

Whitmarsh,A.J. *et al.* (2001) Requirement of the JIP1 scaffold protein for stress-induced JNK activation. *Gene Dev.*, **15**, 2421–2432.

Yates,J.R. III *et al.* (2005) Proteomics of organelles and large cellular structures. *Nat. Rev. Mol. Cell Biol.*, **6**, 702–714.

Zhao,X. *et al.* (2012) Prediction of protein phosphorylation sites by using the composition of *k*-spaced amino acid pairs. *PLoS One*, **7**, e46302.

Zhou,H. *et al.* (2010) Analysis of the subcellular phosphoproteome using a novel phosphoproteomic reactor. *J. Proteome Res.*, **9**, 1279–1288.