Evolution of Protein Phosphorylation for Distinct Functional Modules in Vertebrate Genomes

Zhen Wang,†[,1,2] Guohui Ding,†[,1,3] Ludwig Geistlinger,[4] Hong Li,[1,2] Lei Liu,[1,3] Rong Zeng,*[,1] Yoshio Tateno,*[,5] and Yixue Li*[,1,3]

[1]Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, People's Republic of China

[2]Graduate School of the Chinese Academy of Sciences, Beijing, People's Republic of China

[3]Shanghai Center for Bioinformation Technology, Shanghai, People's Republic of China

[4]Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

[5]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima, Shizuoka, Japan

†These authors contributed equally to this work.

***Corresponding author:** E-mail: zr@sibs.ac.cn; ytateno@genes.nig.ac.jp; yxli@sibs.ac.cn.

Associate editor: Takashi Gojobori

Abstract

Recent publications have revealed that the evolution of phosphosites is influenced by the local protein structures and whether the phosphosites have characterized functions or not. With knowledge of the wide functional range of phosphorylation, we attempted to clarify whether the evolutionary conservation of phosphosites is different among distinct functional modules. We grouped the phosphosites in the human genome into the modules according to the functional categories of KEGG (Kyoto Encyclopedia of Genes and Genomes) and investigated their evolutionary conservation in vertebrate genomes from mouse to zebrafish. We have found that the phosphosites in the vertebrate-specific functional modules (VFMs), such as cellular signaling processes and responses to stimuli, are evolutionarily more conserved than those in the basic functional modules (BFMs), such as metabolic and genetic processes. The phosphosites in the VFMs are also significantly more conserved than their flanking regions, whereas those in the BFMs are not. These results hold for both serine/threonine and tyrosine residues, although the fraction of phosphorylated tyrosine residues is increased in the VFMs. Moreover, the difference in the evolutionary conservation of the phosphosites between the VFMs and BFMs could not be explained by the difference in the local protein structures. There is also a higher fraction of phosphosites with known functions in the VFMs than BFMs. Based on these findings, we have concluded that protein phosphorylation may play more dominant roles for the VFMs than BFMs during the vertebrate evolution. As phosphorylation is a quite rapid biological reaction, the VFMs that quickly respond to outer stimuli and inner signals might heavily depend on this regulatory mechanism. Our results imply that phosphorylation may have an essential role in the evolution of vertebrates.

Key words: protein phosphorylation, evolutionary conservation, functional module, vertebrate genome.

Introduction

Protein phosphorylation (phosphorylation hereafter) is one of the most ubiquitous types of posttranslational modifications (PTMs) (Pawson and Scott 2005). Generally, phosphorylation usually occurs at serine (Ser), threonine (Thr), and tyrosine (Tyr) residues, and the reaction is catalyzed by a variety of kinases (Manning, Whyte, et al. 2002). Through changing the structures and activities of the substrates, phosphorylation can regulate a wide range of biological processes in dynamical manners (Johnson and Barford 1993). With the recent development of high-throughput proteomic techniques based on tandem mass spectrometry, the amount of data on phosphorylation has rapidly been accumulated (Macek et al. 2009). The accumulation of data opens the way for us to conduct research taking all possible phosphorylation events in cellular processes (phosphoproteome) into account and the rising challenge now is to elucidate the underlying functional relevance (Choudhary and Mann 2010).

Comparative genomics can reveal the evolutionary conservation and variation of homologous genes among species, so it has been widely used to provide insights into their functions (Miller et al. 2004). Recently, this strategy also has been applied to phosphoproteomes. If phosphorylated sites (phosphosites) are more conserved than nonphosphosites, we can infer that phosphorylation at the sites may have some important functional implications. In spite of different data sets and strategies, many studies along this line have found evidences for the evolutionary conservation of phosphosites (Boekhorst et al. 2008; Malik et al. 2008; Chen et al. 2010; Nguyen Ba and Moses 2010). In particular, protein structure analyses have revealed that most phosphosites are enriched in disordered regions which, in general, are less conserved than ordered regions (Gnad et al. 2007; Jimenez et al. 2007). Nonetheless, if the

structural propensity is corrected by comparing the phosphosites with their flanking regions, the phosphosites will still show a higher conservation level than the nonphosphosites (Gnad et al. 2007; Nguyen Ba and Moses 2010). Several studies further proposed that in the disordered regions, the conservation of the precise positions of phosphosites might not be required for phosphorylation regulation (Beltrao et al. 2009; Holt et al. 2009; Tan, Bodenmiller, et al. 2009). These results suggest that, although most phosphosites are located in fast-evolving disordered regions with only weak structural constraints, they are still under strong functional constraints. However, there is also a somewhat neutral view on the evolution of phosphorylation. Considering only a small proportion of the phosphosites with known functional effects, Lienhard (2008) proposed that many phosphorylation events are actually nonfunctional because of the nonspecificity of the kinase recognition. Landry et al. (2009) supported this proposal by observing that many phosphosites, especially those with no characterized functions, evolved at a high rate that is comparable with that of nonphosphosites.

Most of the previous studies took all phosphosites as a whole or only focused on their structural propensity. However, we note that as phosphorylation can be involved in a wide range of biological functions, the functional difference might also influence the evolutionary conservation of the phosphosites. Recently, we have publicized a database for PTM research named SysPTM (Li et al. 2009), where we integrated reliable PTM information from numerous public databases and mass spectrometry literature. The phosphorylation sites in the human genome are the most abundant in SysPTM, which covers about 30% of cellular proteins. In the present study, we first classified the human phosphorylated proteins (phosphoproteins) into the functional modules according to the KEGG pathway category (Kanehisa et al. 2008). The category consists of well-annotated pathways and represent our current knowledge about biological functions. Then for each functional module, we investigated the evolutionary conservation of the phosphosites by aligning them with their orthologs in other vertebrates such as mice, rats, chickens, and zebrafishes. As expected, our results suggest that phosphosites in distinct functional modules evolved at different rates.

Materials and Methods

Data Collection and Preprocessing

First, human phosphorylated proteins and sites were retrieved from SysPTM v1.1 (Li et al. 2009). As the protein sequences in SysPTM may come from different data resources, only those consistent with NCBI RefSeq (release 30) (Sayers et al. 2009) were preserved. Next, the phosphoproteins were mapped to KEGG pathways and functional categories (release 50) (Kanehisa et al. 2008), and a phosphoprotein may be assigned to multiple categories. A total of 1,710 nonredundant phosphoproteins with 7,613 phosphosites were preserved for subsequent analysis. Finally, the orthologs of the phosphoproteins in mice, rats, chickens, and zebrafishes were downloaded from NCBI HomoloGene (build 61) (Sayers et al. 2009). The program ClustalW (Chenna et al. 2003) was used to align the sequences for every ortholog family, and the orthologous sites of the human phosphosites were extracted from the alignments (supplementary table S1, Supplementary Material online). In addition, we also searched several nonvertebrate genomes (*Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Saccharomyces cerevisiae*) for the orthologs of the human phosphoproteins in HomoloGene (Sayers et al. 2009).

Estimating Evolutionary Distance of Phosphosites

All phosphosites belonging to the same functional category were concatenated together. A fraction of the phosphosites may be functionally dependent (Cohen 2000) but as they were sampled from a lot of different proteins, they can be treated as independent sites in evolution for simplification. The evolutionary conservation of the phosphosites was evaluated through pairwise comparison with their orthologous sites in the other four vertebrates. The Poisson distance was adopted to measure the degree of conservation, which can correct multiple substitutions at a site and has the linear relationship with time (Nei and Kumar 2000). Specifically, supposing for a functional module i, the proportion of different residues between the human phosphosites and their orthologous sites in a species s is p_{is}. Then the Poisson distance d_{is} can be estimated as

$$d_{is} = -\ln(1 - p_{is}),$$

and the estimate variance is given by

$$\mathrm{var}(d_{is}) = \frac{p_{is}}{(1 - p_{is})n_{is}},$$

, where n_{is} is the total number of the phosphosites (excluding indels) (Nei and Kumar 2000). When the conservation of phosphosites between two functional modules are compared, the larger the Poisson distance, the higher the evolutionary rate and the lower the evolutionary conservation. To test the difference in the Poisson distance between two functional modules i and j statistically, we defined the z-score

$$z_{ij,s} = \frac{d_{is} - d_{js}}{\sqrt{\mathrm{var}(d_{is}) + \mathrm{var}(d_{js})}},$$

which follows the standard normal distribution approximately under the null hypothesis. This static is based on the comparison with only one species s. To assess the difference across all the species between the module i and j, we further defined

$$z_{ij}^2 = \sum_{s=1}^{4} z_{ij,s}^2,$$

which follows a chi-square distribution with the degree of freedom equal to 4 under the null hypothesis (as we considered four vertebrates in this study).

Contrasting Phosphosites with Their Flanking Regions

Comparing the evolutionary conservation of phosphosites only is not enough to infer the functional essentiality of

phosphorylation because the sites might also be conserved for many other reasons other than phosphorylation. These potential biases were usually removed by contrasting the conservation of phosphosites with that of their flanking regions (Gnad et al. 2007; Nguyen Ba and Moses 2010). As in previous studies (Nguyen Ba and Moses 2010), the flanking region was defined as the ten residues centered on a phosphosite. All flanking regions in a functional module were extracted and combined together, and the Poisson distance of the flanking regions was estimated with the same method as we did for the phosphosites (supplementary table S2, Supplementary Material online). The *z*-score was also used to assess whether the phosphosites and the flanking regions evolved at the same rate, which is calculated as

$$z'_{is} = \frac{d_{is} - d'_{is}}{\sqrt{\mathrm{var}(d_{is}) + \mathrm{var}(d'_{is})}},$$

where d_{is} and d'_{is} are the Poisson distance for the phosphosites and the flanking regions, respectively. Under the null hypothesis, the *z*-score follows the standard normal distribution approximately. Also, to test the conservation of the phosphosites over the flanking regions across all the species for the module *i*, we calculated

$$z'^2_i = \sum_{s=1}^{4} z'^2_{is},$$

which follows a chi-square distribution with the degree of freedom equal to 4.

Protein Structures and Known Functional Phosphosites

In the following analysis, we classified the phosphosites according to three criteria and investigated their distribution and evolutionary conservation separately. The criteria included: 1) the type of phosphorylated residues (Ser/Thr or Tyr), 2) the local protein structures (disordered or ordered regions), and 3) whether the function of a phosphosite was known or not. The protein structures were predicted by the disEMBL program (v1.5) (Linding et al. 2003), and the disordered region was defined as either loops/coils or hot loops as given by the result of the program. The known functional phosphosites were predicted by NetworKIN 2.0 (Linding et al. 2007). They were inferred from high-confidence kinase–site interactions.

Results

Phosphoproteins in Distinct Functional Modules

We mapped 1,710 nonredundant phosphoproteins with 7,613 phosphosites in the human genome to the KEGG pathways (Kanehisa et al. 2008). The pathways were further grouped into larger functional categories. We focused on the five major functional categories in KEGG, namely metabolism, genetic information processing, environmental information processing, cellular processes, and human diseases (fig. 1*A*). The metabolism category contains metabolic pathways of cellular molecules, and the genetic information processing category contains pathways related to replication, transcription, and translation (supplementary table S1, Supplementary Material online). As we found that more than 75% of the phosphoproteins in the two categories have orthologs in nonvertebrate genomes (fig. 1*B*), we defined them as the basic functional modules (BFMs). The environmental information processing category contains signal transduction pathways that respond to environmental stimuli. The cellular processes category also contains signaling pathways that underlie more complex organic systems, such as developmental, immune, and

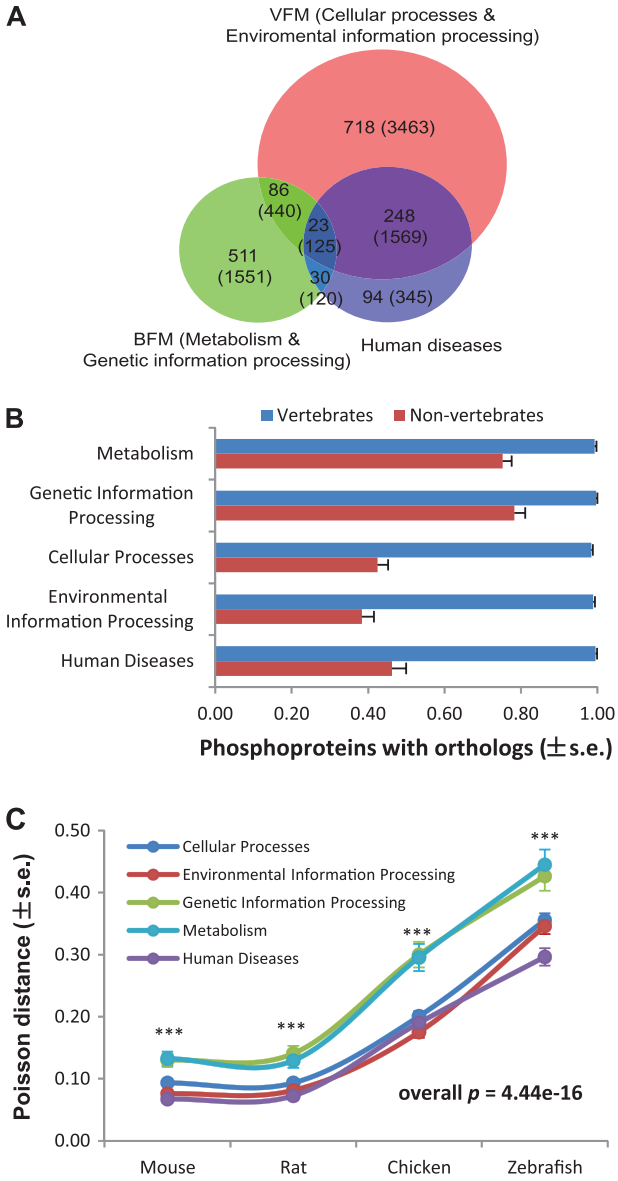

FIG. 1. Human phosphoproteins in the KEGG functional categories. (*A*) Number and overlap of human phosphoproteins among the functional categories. The number of phosphosites is indicated in brackets. (*B*) Proportion of phosphoproteins with orthologs in the genomes of vertebrates (mouse, rat, chicken, and zebrafish) and nonvertebrates (*Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Saccharomyces cerevisiae*). (*C*) Poisson distance of the phosphosites for the functional categories in vertebrate genomes (SE, standard error; $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$).

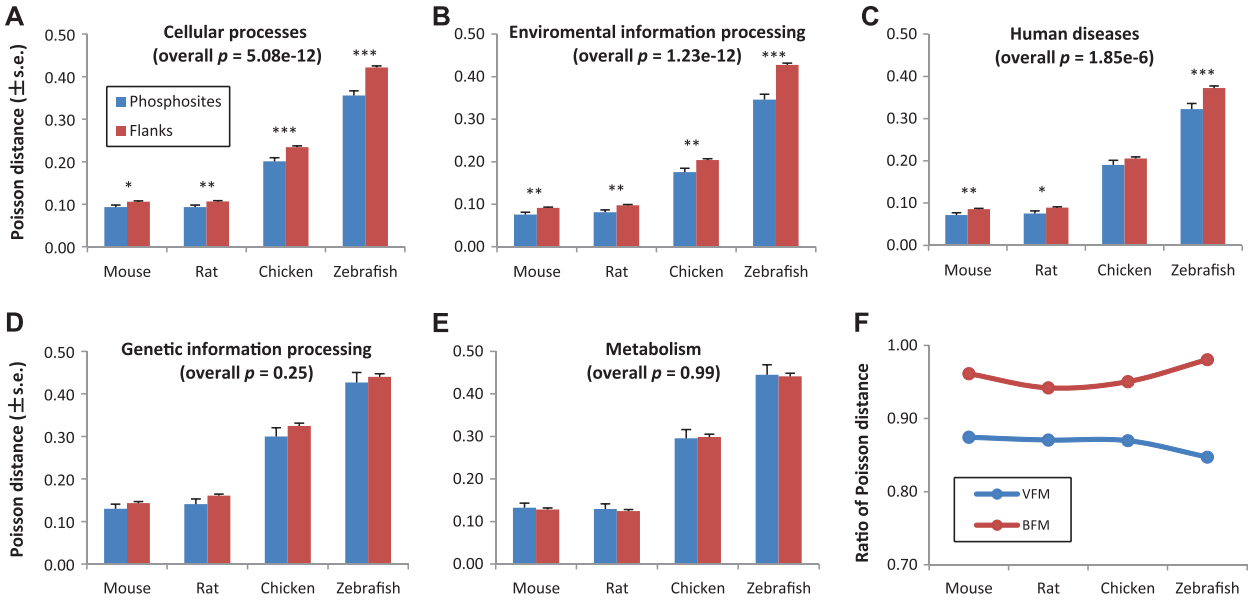

FIG. 2. (*A*–*E*) Comparison of the Poisson distance between the phosphosites and their flanking regions (SE, standard error; $*P < 0.05$; $**P < 0.01$; $***P < 0.001$). (*F*) Ratio of the Poisson distance between the phosphosites and their flanking regions.

nervous systems (supplementary table S1, Supplementary Material online). For the phosphoproteins in these two categories, we found less than 45% of them have orthologs in nonvertebrate genomes (fig. 1*B*), which is significantly lower than the percentage in the BFMs ($P < 2.2 \times 10^{-16}$, chi-square test). Thus, we defined the two categories as the vertebrate-specific functional modules (VFMs). Figure 1*A* shows that there are more phosphoproteins and phosphosites in the VFMs than in the BFMs. Interestingly, most phosphoproteins related to human diseases were found in the VFMs (fig. 1*A*).

Next, we investigated the evolution of the phosphosites for the five functional categories in the vertebrate genomes. We pairwisely compared the human phosphosites with their orthologous sites in mice, rats, chickens, and zebrafishes and computed the Poisson distance between each pair of the comparison (see Materials and Methods). The results show that the phosphosites involved in the VFMs evolved at a slower rate than those in the BFMs across all the vertebrates (fig. 1C). The phosphosites in human diseases also display a similar evolutionary rate to those in the VFMs. To validate the results, we further performed the *z*-score test to compare the Poisson distance of the phosphosites between the VFMs and BFMs (see Materials and Methods). The small *P* values for all the species demonstrated that the phosphosites in the VFMs evolved significantly more slowly than those in the BFMs during the evolution of vertebrates (fig. 1C).

Conservation of Phosphosites Versus Their Flanking Regions

Although the phosphosites in the VFMs evolved more slowly than those in the BFMs, we could not attribute it to the difference in the functions of phosphorylation immediately because of the possibility that the phosphosites can also be conserved for many other reasons rather than phosphorylation alone. For example, the conservation of the phosphosites may be influenced by protein dispensability, expression abundance, and others (Pal et al. 2006). More importantly because phosphosites occur preferentially in disordered regions, the local structural constraints also have a large impact on the conservation of phosphosites (Gnad et al. 2007; Jimenez et al. 2007). In fact, the higher conservation of phosphosites over nonphosphosites can only be pronounced when the local protein structures are taken into account (Gnad et al. 2007). Following Nguyen Ba and Moses (2010), we selected the ten flanking residues around a phosphosite as the background to correct the potential biases (supplementary table S2, Supplementary Material online). We computed the Poisson distance of the flanking regions by the same method as we did for the phosphosites and performed the *z*-score test to examine whether the phosphosites are more conserved than their flanking regions. In addition to obtaining the *P* value for each individual species, we also developed a method to gain an overall *P* value to assess the statistical significance across all the vertebrates (see Materials and Methods).

Figure 2 shows the results of the comparison between the phosphosites and their flanking regions for the five major functional categories mentioned above. The phosphosites in cellular processes, environmental information processing, and human diseases were turned out to evolve significantly more slowly than their flanking regions (overall $P \leq 1.85 \times 10^{-6}$). On the contrary, the phosphosites in genetic information processing and metabolism were shown to evolve at a rate comparable with their flanking regions (overall $P \geq 0.25$). To make the contrast more clear, we computed the ratio of the Poisson distance between the phosphosites and their flanking regions (fig. 2*F*). It is shown that the ratio for the VFMs is also lower than that for BFMs,

Table 1. Subcategories with Phosphosites Significantly More Conserved Than Flanking Regions (FDR < 0.05).

Subcategory	Major Category	Overall *P* Value	FDR
Behavior	Cellular processes	8.28×10^{-13}	2.65×10^{-11}
Immune system	Cellular processes	1.88×10^{-09}	3.01×10^{-08}
Signal transduction	Environmental information processing	1.42×10^{-08}	1.51×10^{-07}
Signaling molecules and interaction	Environmental information processing	4.02×10^{-08}	3.22×10^{-07}
Cancers	Human diseases	7.35×10^{-05}	4.70×10^{-04}
Transcription	Genetic information processing	4.42×10^{-04}	2.36×10^{-03}
Immune disorders	Human diseases	5.11×10^{-03}	2.34×10^{-02}
Endocrine system	Cellular processes	7.37×10^{-03}	2.95×10^{-02}
Metabolic disorders	Human diseases	1.26×10^{-02}	4.48×10^{-02}

NOTE.—FDR, false discovery rate.

and the tendency is rather stable across all the species. These results strongly suggest that the phosphosites in the VFMs have been subject to stronger selective constraints than those in the BFMs.

Each of the five major functional categories in KEGG is composed of several subcategories. In order to investigate the functional modules in more detail, we also compared the evolutionary conservation between the phosphosites and their flanking regions for each subcategory (table 1). Again, most subcategories in which the phosphosites are significantly more conserved than their flanking regions are within the VFMs and human diseases, such as immune system, signal transduction, and cancers. The only exception is transcription, which belongs to the genetic information processing category, but it contains more conserved phosphosites than corresponding flanking regions.

Evolutionary Pattern of Phosphorylated Ser, Thr, and Tyr

Next, we investigated the evolutionary pattern for the three phosphorylated residues (Ser, Thr, and Tyr) separately. The three residues differ in two main aspects. First, although Ser and Thr are highly hydrophilic, Tyr contains a hydrophobic ring (Jimenez et al. 2007). Second, although Ser and Thr usually share the same group of kinases, Tyr can only be catalyzed by more specific kinases (Ubersax and Ferrell 2007). In fact, although Ser/Thr kinases are present in all eukaryotes, Tyr kinases can only be found in metazoans (Manning, Plowman, et al. 2002). Moreover, a recent study suggested that the specificity of Tyr phosphorylation is associated with the metazoan complexity (Tan, Pasculescu, et al. 2009). In our data set, we also found that the proportion of phosphorylated Tyr is larger in the VFMs than in the BFMs (fig. 3*A*, $P = 3.15 \times 10^{-15}$, chi-square test), implying that Tyr phosphorylation plays more important roles in the VFMs than BFMs.

We then computed the Poisson distance for the phosphorylated Ser/Thr and Tyr, respectively and compared their evolutionary conservation between the BFMs and VFMs. We found that both phosphorylated Ser/Thr (overall $P = 2.62 \times 10^{-9}$) and Tyr (overall $P = 0.02$) residues are more conserved in the VFMs than in the BFMs (fig. 3*B* and *C*). In addition, both types of the phosphorylated residues show a higher conservation level than their flanking regions only in the VFMs (fig. 3*D*–*G*), though this is more significant for the phosphorylated Tyr (overall $P = 9.37 \times 10^{-52}$) than for the phosphorylated Ser/Thr (overall $P = 2.38 \times 10^{-3}$). Taken together, these results suggest that our finding that the phosphosites involved in the VFMs are under stronger selective constraints than those in the BFMs holds for both phosphorylated Ser/Thr and Tyr residues.

Impact of Local Protein Structures

It has been demonstrated that most phosphosites are located in disordered protein surfaces, which enable them to be more accessible to kinases and other interaction partners (Iakoucheva et al. 2004; Gnad et al. 2007; Jimenez et al. 2007). Only a few proportion of phosphosites are buried in ordered regions, and their phosphorylation often involves conformational rearrangements (Pawson and Scott 2005; Jimenez et al. 2007). As mentioned in Introduction, the relationship between the local protein structures and the evolution of phosphosites has widely been discussed. Generally, the disordered regions of a phosphoprotein diverged at a higher rate than the ordered regions (Gnad et al. 2007; Jimenez et al. 2007; Lin et al. 2007), but the phosphosites in the disordered regions can remain conserved to keep their functions intact (Gnad et al. 2007; Holt et al. 2009; Nguyen Ba and Moses 2010). Although we found that the phosphosites in the VFMs are more conserved than those in the BFMs, it may simply be because the phosphosites are more enriched in the disordered regions in the BFMs than in the VFMs. To rule out this possibility, we predicted the local protein structures for the phosphosites by disEMBL (Linding et al. 2003). Figure 4*A* indicates that the phosphorylated Ser/Thr residues have little difference in the local structures between the BFMs and VFMs ($P = 0.73$, chi-square test), and there is even a higher fraction of phosphorylated Tyr residues enriched in the disordered regions in the VFMs than in the BFMs ($P = 0.02$). Therefore, the local protein structures cannot account for the difference in the evolutionary conservation of the phosphosites between the BFMs and VFMs.

We next examined whether the phosphosites in the VFMs are more conserved than those in BFMs when the local protein structures are considered. In the disordered regions, it is evident that the phosphosites in the VFMs evolved more slowly than those in the BFMs (overall $P = 1.11 \times 10^{-16}$, fig. 4*B*). The phosphosites in the disordered regions also show a higher conservation level than

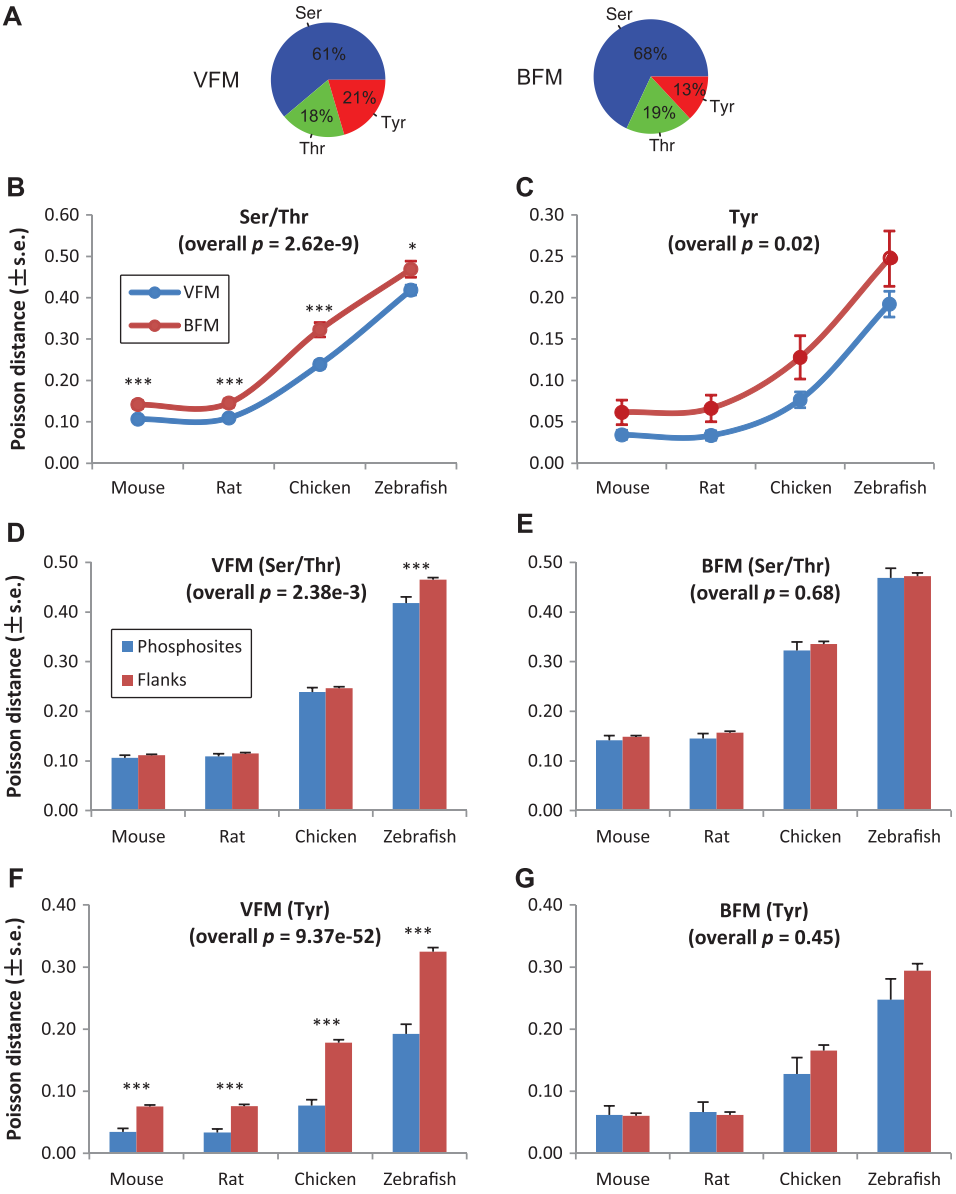

FIG. 3. Evolutionary conservation of phosphorylated Ser/Thr and Tyr residues. (*A*) Distribution of the residues in the VFMs and BFMs. (*B*–*C*) Poisson distance of the residues for the VFMs and BFMs. (*D*–*G*) Comparison of the Poisson distance between the residues and their flanking regions (SE, standard error; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$).

their flaking regions in the VFMs (overall $P = 7.12 \times 10^{-14}$) but not in the BFMs (overall $P = 0.96$, fig. 4*D* and *E*). In the ordered regions, however, the phosphosites in the BFMs and VFMs evolved at a comparable rate (overall $P = 0.55$, fig. 4*C*). In addition, even the phosphosites in the VFMs do not show a significantly higher conservation level than their flanking regions in the ordered regions (overall $P = 0.08$, fig. 4*F* and *G*). These results further suggest that, at least for the phosphosites in the disordered regions, the difference in the evolutionary rate between the BFMs and VFMs should not be ascribed to the local protein structures.

Distribution of Known Functional Phosphosites

Because the local protein structures cannot account for the higher conservation of phosphosites in the VFMs than in the BFMs, it is more likely that the functional constraints on the phosphosites are different for the two groups of modules. For the phosphosites with known functions, it has been demonstrated that they are under strong functional constraints (Landry et al. 2009; Nguyen Ba and Moses 2010). However, the constraints are weak for those with no characterized functions, which might be because a fraction of them have no function at all (Lienhard 2008; Landry et al. 2009). To investigate whether the functional effects of the phosphosites are different between the VFMs and BFMs, we retrieved known functional phosphosites from the NetworKIN repository (Linding et al. 2007), which were predicted from high-confidence kinase–site interactions. Figure 5*A* indicates that the proportion of known functional phosphosites is always larger in the VFMs than in the BFMs ($P < 0.01$, chi-square test),

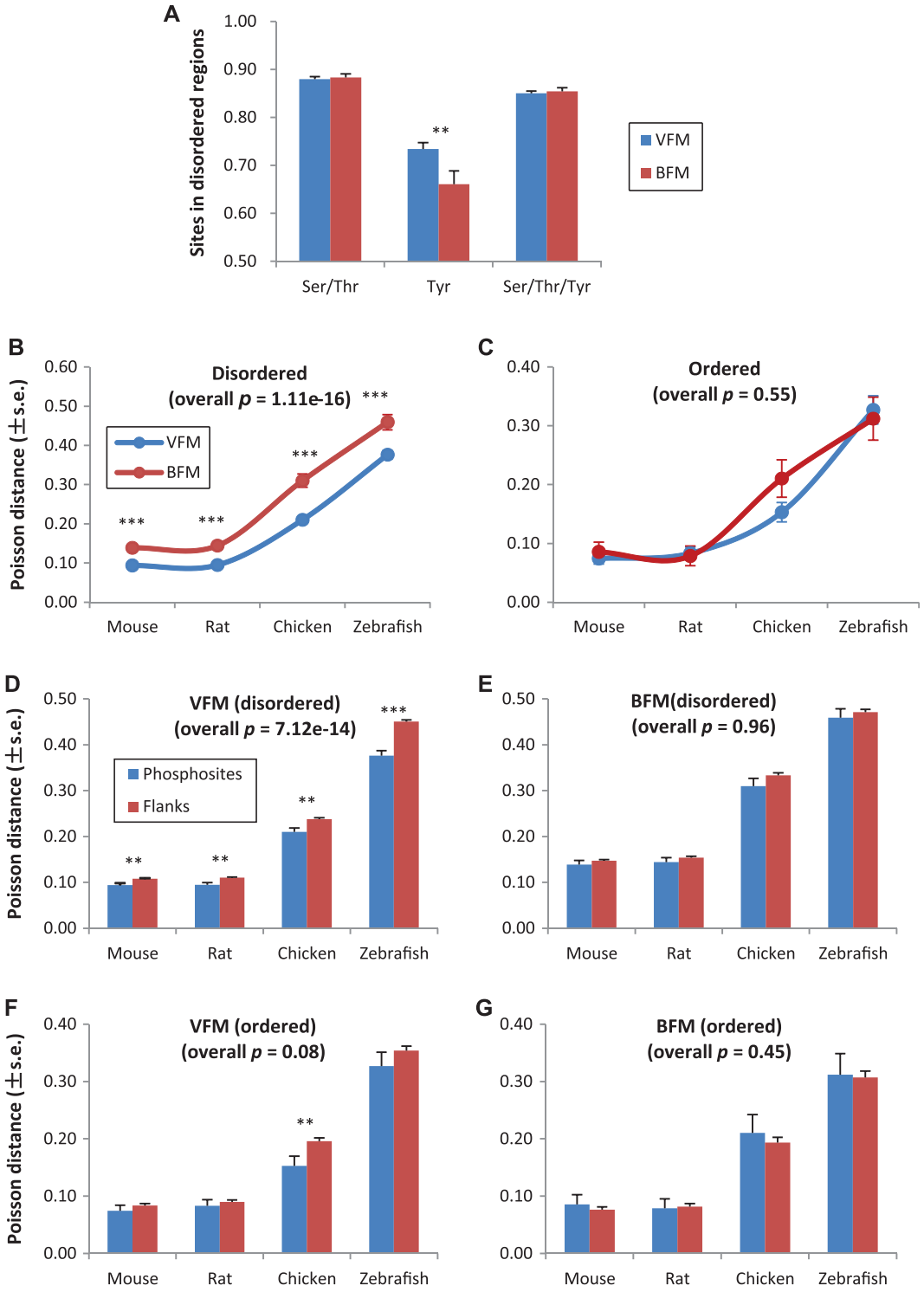

Fig. 4. Impact of local protein structures. (*A*) Proportion of phosphosites in disordered regions. (*B*–*C*) Poisson distance of the phosphosites in disordered and ordered regions. (*D*–*G*) Comparison of the Poisson distance between the phosphosites and their flanking regions in disordered and ordered regions (SE, standard error; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$).

regardless of the phosphorylated residues and the local protein structures.

We then computed the Poisson distance for the phosphosites with known and unknown functions separately and found that in both the BFMs and VFMs, the known functional phosphosites have lower evolutionary rates than the unknown ones (overall $P < 0.01$, fig. 5*B* and *C*). The known functional phosphosites are also more conserved than their flanking regions either in the VFMs (overall $P = 2.11 \times 10^{-27}$) or in the BFMs (overall $P = 5.15 \times 10^{-4}$, fig. 5*D* and *E*). These results are in agreement with the strong functional constraints of the phosphosites with

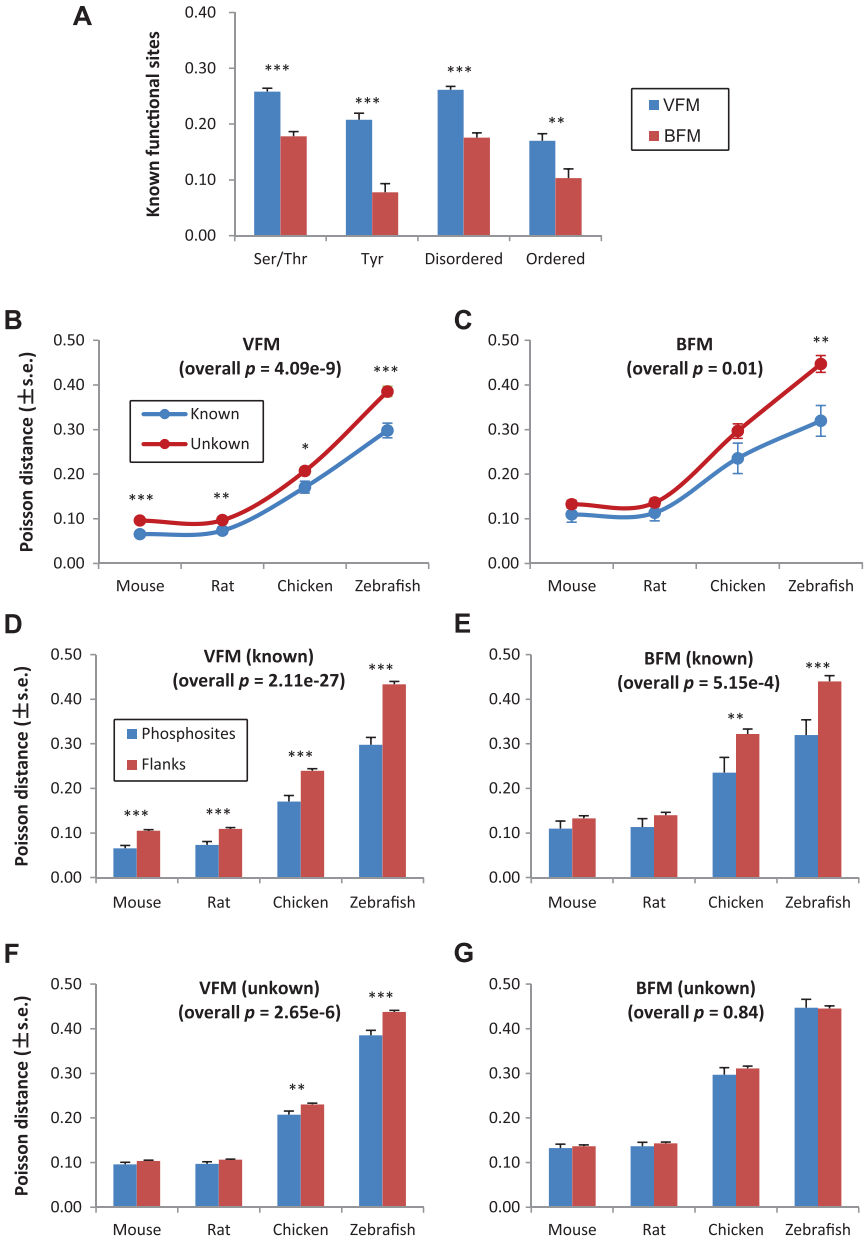

FIG. 5. Distribution of known functional phosphosites. (*A*) Proportion of phosphosites with known functions. (*B*–*C*) Poisson distance of the phosphosites with known and unknown functions. (*D*–*G*) Comparison of the Poisson distance between the phosphosites and their flanking regions (SE, standard error; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$).

known functions. However, the unknown functional phosphosites show a higher conservation level than their flanking regions only in the VFMs (overall $P = 2.65 \times 10^{-6}$) but not in the BFMs (overall $P = 0.84$, fig. 5*F* and *G*). This result implies that there is still a large fraction of phosphosites with potential functions in the VFMs.

Discussion

In addition to the local protein structures (Gnad et al. 2007; Jimenez et al. 2007; Holt et al. 2009) and whether the phosphosites have characterized functions or not (Landry et al. 2009; Nguyen Ba and Moses 2010), our study has revealed that the evolution of phosphosites is also influenced by the functional modules they are involved in. In the vertebrate genomes, the phosphorylated functional modules can be divided into two broad groups. The first group contains fundamental biological functions, such as metabolic and genetic processes. These functions are shared by both vertebrates and nonvertebrates, thus named the BFMs (fig. 1*B*). In contrast, the second group contains more VFMs, such as signal transductions and responses to inner and outer stimuli (fig. 1*B*). These functions may improve the complexity of vertebrates and make the vertebrates more adaptive to diverse environments.

On the basis of the division, we have demonstrated that the phosphosites are generally more conserved in the VFMs than in the BFMs during the evolution of vertebrates. First, the evolutionary rate of the phosphosites in the VFMs is

significantly lower than that in the BFMs (fig. 1C). Secondly, although the phosphosites in the VFMs show a significantly higher conservation level than their flanking regions, those in the BFMs do not (fig. 2). Thirdly, the phosphosites in the disordered regions also evolved significantly more slowly in the VFMs than in the BFMs, which cannot be explained by the difference in the structural constraints (fig. 4). Finally, there is a larger fraction of phosphosites with known functions in the VFMs than in the BFMs (fig. 5). Taken together, our findings suggest that the difference in the evolutionary rate of the phosphosites for the functional modules is due to the difference in the functional constraints of phosphorylation. In other words, phosphorylation may play more dominant roles for the VFMs than for the BFMs.

To rationalize our suggestion, we note that protein activities are regulated at the transcriptional, posttranscriptional, and posttranslational levels, to the last of which phosphorylation belongs (Malbon et al. 1990). Among the three levels of regulations, phosphorylation may be the quickest regulatory mechanism. For example, the *G* protein–coupled receptors can respond to new stimuli within milliseconds to minutes via phosphorylation, whereas the change in the expression of the receptors may require several hours (Pitcher et al. 1998). In this sense, the VFMs aiming to process quick signals and unexpected stimuli may heavily rely on the phosphorylation regulation. On the contrary, for the BFMs requiring routine and stable reactions, the transcriptional regulation may be more dominant than the phosphorylation regulation. Because vertebrates are characterized by their improved organismal complexity and adaptive capacity to diverse environments, our findings imply that protein phosphorylation may play an essential role in the evolution of vertebrates. It is interesting that besides the large-scale genomic revolution, such as whole-genome duplications (Wang et al. 2009), the small-scale modifications, such as protein phosphorylation might also make a great contribution to the macroevolution events.

Although phosphorylation may be more dominant for the VFMs than for the BFMs, our results do not imply that phosphorylation is not important for the regulation of fundamental functions. In fact, there are many examples suggesting that phosphorylation regulates metabolic and genetic information processing pathways (Karin 1994). For instance, the AMP-activated protein kinase (AMPK) stimulates ATP-producing pathways and inhibits ATP-consuming pathways (Hardie 2007). In this case, although the metabolic enzymes are initially regulated via phosphorylation, the longer term effects are still achieved via transcriptional regulation (Hardie 2007). More importantly, the targets of AMPK are also regulated by the insulin signaling pathway in the whole map in KEGG, which seems specific to vertebrates. In our data set, we have found that in the BFMs, only the phosphosites involved in transcription show a significantly higher conservation level than their flanking regions (table 1). However, because many transcription factors are actually located at the downstream of signaling pathways (Karin 1994), our conclusion does not contradict to the fact that phosphorylation participates in regulating the BFMs.

We have also found that most phosphosites related to human disease pathways are in the VFMs and remain conserved in vertebrates (figs. 1 and 2). This finding could be explained by the association between the disturbance of signaling systems and human diseases, such as cancers. This result is also consistent with a recent report (Tan, Bodenmiller, et al. 2009), though we used quite different approaches.

Although both phosphorylated Ser/Thr and Tyr residues are more conserved in the VFMs than in the BFMs, a higher fraction of phosphorylated Tyr residues is found in the VFMs (fig. 3A). There is also a higher fraction of phosphorylated Tyr residues located in the disordered regions in the VFMs than BFMs (fig. 4A). These results are in agreement with the study of kinase evolution, which suggests that Tyr kinases are largely expanded during the evolution of complex metazoans but not in unicellular organisms (Manning, Plowman, et al. 2002). However, it remains a mystery about the evolutionary advantage that the Tyr phosphorylation brings to the organisms. Recently, Tan, Pasculescu, et al. (2009) found that the total number of Tyr residues were decreased with the increase in genome complexity. They argued that the Tyr loss could eliminate deleterious phosphosites and potential noises in the signaling systems (Tan, Pasculescu, et al. 2009). This finding might provide an explanation for this issue in part.

Supplementary Material

Supplementary tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We acknowledge the anonymous reviewers for their constructive comments. This work was supported by grants from National High-Tech R & D Program (863): 2006AA02Z334, 2007DFA31040; State key basic research program (973): 2006CB910705, 2010CB529206, 2010CB529200; Research Program of CAS: KSCX2-YW-R-112; and Special Start-up Fund for Chinese Academy of Sciences President Award Winner (to G.D.).

References

Beltrao P, Trinidad JC, Fiedler D, Roguev A, Lim WA, Shokat KM, Burlingame AL, Krogan NJ. 2009. Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. *PLoS Biol.* 7:e1000134.

Boekhorst J, van Breukelen B, Heck AJ, Snel B. 2008. Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. *Genome Biol.* 9:R144.

Chen SC, Chen FC, Li WH. 2010. Phosphorylated and non-phosphorylated serine and threonine residues evolve at different rates in mammals. *Mol Biol Evol.* doi:10.1093/molbev/msq1142

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* 31:3497–3500.

Choudhary C, Mann M. 2010. Decoding signalling networks by mass spectrometry-based proteomics. *Nat Rev Mol Cell Biol*. 11:427–439.
Cohen P. 2000. The regulation of protein function by multisite phosphorylation—a 25 year update. *Trends Biochem Sci*. 25:596–601.
Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, Mann M. 2007. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol*. 8:R250.
Hardie DG. 2007. AMP-activated/SNF1 protein kinases: conserved guardians of cellular energy. *Nat Rev Mol Cell Biol*. 8:774–785.
Holt LJ, Tuch BB, Villen J, Johnson AD, Gygi SP, Morgan DO. 2009. Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science* 325:1682–1686.
Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*. 32:1037–1049.
Jimenez JL, Hegemann B, Hutchins JR, Peters JM, Durbin R. 2007. A systematic comparative and structural analysis of protein phosphorylation sites based on the mtcPTM database. *Genome Biol*. 8:R90.
Johnson LN, Barford D. 1993. The effects of phosphorylation on the structure and function of proteins. *Annu Rev Biophys Biomol Struct*. 22:199–232.
Kanehisa M, Araki M, Goto S, et al. (11 co-authors). 2008. KEGG for linking genomes to life and the environment. *Nucleic Acids Res*. 36:D480–D484.
Karin M. 1994. Signal-transduction from the cell-surface to the nucleus through the phosphorylation of transcription factors. *Curr Opin Cell Biol*. 6:415–424.
Landry CR, Levy ED, Michnick SW. 2009. Weak functional constraints on phosphoproteomes. *Trends Genet*. 25:193–197.
Li H, Xing X, Ding G, Li Q, Wang C, Xie L, Zeng R, Li Y. 2009. SysPTM—a systematic resource for proteomic research of post-translational modifications. *Mol Cell Proteomics*. 8:1839–1849.
Lienhard GE. 2008. Non-functional phosphorylations? *Trends Biochem Sci*. 33:351–352.
Lin YS, Hsu WL, Hwang JK, Li WH. 2007. Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Mol Biol Evol*. 24:1005–1011.
Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. 2003. Protein disorder prediction: implications for structural proteomics. *Structure* 11:1453–1459.
Linding R, Jensen LJ, Ostheimer GJ, et al. (21 co-authors). 2007. Systematic discovery of in vivo phosphorylation networks. *Cell* 129:1415–1426.
Macek B, Mann M, Olsen JV. 2009. Global and site-specific quantitative phosphoproteomics: principles and applications. *Annu Rev Pharmacol Toxicol*. 49:199–221.
Malbon CC, Hadcock JR, Rapiejko PJ, Ros M, Wang HY, Watkins DC. 1990. Regulation of transmembrane signalling elements: transcriptional, post-transcriptional and post-translational controls. *Biochem Soc Symp*. 56:155–164.
Malik R, Nigg EA, Korner R. 2008. Comparative conservation analysis of the human mitotic phosphoproteome. *Bioinformatics* 24: 1426–1432.
Manning G, Plowman GD, Hunter T, Sudarsanam S. 2002. Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci*. 27:514–520.
Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. 2002. The protein kinase complement of the human genome. *Science* 298:1912–1934.
Miller W, Makova KD, Nekrutenko A, Hardison RC. 2004. Comparative genomics. *Annu Rev Genomics Hum Genet*. 5: 15–56.
Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. New York: Oxford University Press.
Nguyen Ba AN, Moses AM. 2010. Evolution of characterized phosphorylation sites in budding yeast. *Mol Biol Evol*. 27:2027–2037.
Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet*. 7:337–348.
Pawson T, Scott JD. 2005. Protein phosphorylation in signaling—50 years and counting. *Trends Biochem Sci*. 30:286–290.
Pitcher JA, Freedman NJ, Lefkowitz RJ. 1998. G protein-coupled receptor kinases. *Annu Rev Biochem*. 67:653–692.
Sayers EW, Barrett T, Benson DA, et al. (33 co-authors). 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 37:D5–D15.
Tan CS, Bodenmiller B, Pasculescu A, Jovanovic M, Hengartner MO, Jorgensen C, Bader GD, Aebersold R, Pawson T, Linding R. 2009. Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Sci Signal*. 2:ra39.
Tan CS, Pasculescu A, Lim WA, Pawson T, Bader GD, Linding R. 2009. Positive selection of tyrosine loss in metazoan evolution. *Science* 325:1686–1688.
Ubersax JA, Ferrell JE Jr.. 2007. Mechanisms of specificity in protein phosphorylation. *Nat Rev Mol Cell Biol*. 8:530–541.
Wang Z, Ding G, Yu Z, Liu L, Li Y. 2009. Modeling the age distribution of gene duplications in vertebrate genome using mixture density. *Genomics* 93:146–151.