Update

Genome Analysis

Weak functional constraints on phosphoproteomes

Christian R. Landry*, Emmanuel D. Levy* and Stephen W. Michnick

Centre Robert-Cedergren, Bio-Informatique et Génomique, Département de Biochimie, C.P. 6128, Succ. Centre-Ville, Montreal, Quebec H3C 3J7, Canada

Owing to their crucial roles in regulating protein function, phosphorylation sites (phosphosites) are expected to be evolutionarily conserved. However, mixed results regarding this prediction have been reported. We resolve these contrasting conclusions to show that phosphosites are, on average, more conserved than non-phosphorylated equivalent residues when their enrichment in disordered regions of proteins is taken into account. Phosphosites of known function are dramatically more conserved than those with no characterized function, indicating that the apparent rapid evolution of phosphoproteomes results from a large fraction of phosphosites being non-functional. Our findings highlight the need to use evolutionary information to identify functional regulatory features such as post-translational modifications of eukaryotic proteomes.

Evolution of post-translational regulatory networks

Understanding the mechanisms and evolution of gene and protein regulation at a network scale is of fundamental importance to both biomedical and evolutionary biology. Protein phosphorylation is a common post-translational modification in eukaryotic cells, in which it serves to regulate protein catalytic activity, localization, stability and interactions (Box 1). These multiple functions have motivated studies aimed at characterizing phosphorylation sites (phosphosites) at the proteome scale in human and in several model organisms. Studies have reported a number of conserved phosphosites compared with non-phosphorylated residues, supporting the notion that phosphosites are under evolutionary constraint because they have key roles in regulating protein function [1,2]. However, it has also been noted that a large number of phosphosites do not have a known function and that, perhaps, in the crowded environment of the cell, random encounters between protein kinases and degenerate recognition motifs in proteins frequently result in non-functional phosphorylation events [3–5]. This view is supported by the observation that some phosphosites show little conservation among species [4]. In an attempt to reconcile the conflicting views of phosphosite evolution described here, we have comprehensively and quantitatively assessed phosphosite conservation by considering protein structural properties and functional data. We show that the phosphoproteome evolves rapidly as a whole, but that phosphosites do evolve slowly when they are functional.

Corresponding author: Michnick, S.W. (stephen.michnick@umontreal.ca).

* These authors contributed equally to this work. Available online 6 April 2009.

High evolutionary turnover and polymorphism of eukaryotic phosphoproteomes

We investigated the evolution of phosphosites in yeasts and vertebrates by compiling a large dataset consisting of 5010 and 7137 high quality phosphosites distributed among 2099 and 2347 groups of orthologous proteins for these two lineages, respectively (Table S1 in the supplementary material online). In each case, we compared the evolution of serine and threonine residues known to be phosphorylated *in vivo* in yeast (*Saccharomyces cerevisiae*) and human (*Homo sapiens*) (pS/pT) with that of control sets of serines and threonines not reported to be phosphorylated in these proteomes (S/T) (Figure S1). Based on observations that some phosphosites are conserved across distant species and regulate protein function [1] (Box 1), we first hypothesized that the phosphoproteome (pS/pT) should exhibit a slower evolutionary turnover than random S/T, that is, pS/pT should be conserved over longer evolutionary periods. Assuming that the total number of phosphosites in these proteomes is at gain-to-loss equilibrium, a slow turnover (i.e. low birth and death rates) would be reflected by an old age of phosphosites. We reconstructed ancestral sequences by maximum likelihood [6] for the yeast and vertebrate lineages (Figure 1a; methods in the supplementary material online for details). We then inferred the age of a phosphosite by the first ancestral node in the tree where the inferred ancestral residue was non-phosphorylatable (other than serines and threonines). Results of these analyses indicate that at least 28% (yeast) and 6.6% (human) of pS/pT have appeared in the past ~100 and ~150 million years (My), respectively. This corresponds to the split between *S. cerevisiae* and the pre-whole-genome duplication hemiascomycota, and between placental mammals and marsupials. Because a phosphosite can only have appeared at the same time or more recently than the appearance of S/T, our estimates of turnover are conservative. We observe a similar turnover for S/T; for example, 29.6% (yeast) and 8.8% (human) in the past ~100 and ~150 My respectively, which supports the view that phosphoproteomes are evolutionarily dynamic.

To obtain a more quantitative estimate of phosphosites evolution, we then tested the hypothesis that pS/pT globally evolve slower on average than S/T by comparing their relative rate of evolution along these phylogenetic trees. We concatenated 2099 (yeast) and 2347 (human) groups of orthologous proteins into two proteome-wide multiple alignments to obtain proteome-wide rates. We inferred the rates of amino acid replacement at every position using a Bayesian method as implemented in rate4site (http://www.tau.ac.il/~itaymay/cp/rate4site.html) [7] (Figure 1c and Figure S1). The alignments confirm that phosphoproteomes have

Box 1. The role of protein phosphorylation in protein and cellular regulation

Eukaryotic proteins are phosphorylated on serines, threonines and tyrosine residues by serine/threonine and tyrosine kinases. The reversible addition of a phosphate group to these amino acids at specific positions on a substrate affects several of the properties of the protein and, thus, its function. Protein kinases recognize their substrates through different mechanisms, including the recognition of linear peptide motifs in binding grooves proximal to the kinase active site, the interaction between sites distal from the active site on the kinase with the substrate protein and through association via other substrate-binding subunits, adaptors or scaffold proteins [3]. Protein kinases constitute 2–4% of all genes in eukaryotic genomes [17] and up to 30% of proteins are phosphorylated [18]. Regulation of the function of the substrate can be achieved through allosteric conformational changes, by modifying the accessibility of enzyme active sites or simply through modification of its bulk electrostatic charge [16]. The regulation of protein activity by protein kinases has an important role at the cellular level. For instance, protein kinases can phosphorylate transcription factors to modulate their transcriptional activity, through modifying their oligomerization state or localization in the cells. In mammals, tyrosine phosphorylation of the transcription factor STAT1 modifies its dimerization interface, inducing cytosol to nuclear translocation and DNA-binding activity [19]. In yeast grown under rich condition, the transcription factor Pho4 is phosphorylated on multiple sites and is concentrated in the cytoplasm. Upon phosphate starvation, Pho4 is dephosphorylated and translocates to the nucleus, inducing the expression of specific genes required to respond to this environmental stress [20]. Protein phosphorylation is also the mechanism of information transfer in signal transduction pathways. For instance, several cell-surface receptors transfer information within the cells through cascades of mitogen-activated protein kinases (MAPK), in which a first MAPK sequentially phosphorylates and activates a second MAPK, followed in turn by sequential phosphorylation of one or more additional MAPKs. These cascades are responsible for the control of gene expression, cell proliferation and, also, programmed cell death. For instance, in yeast, MAP kinase cascades regulate diverse cellular processes such as osmotic response and response to mating pheromones [21]. Finally, protein phosphorylation can also regulate the activity of specific enzymes [22].

The prime role of protein phosphorylation in cell signaling and regulation has spurred the development of powerful genomic and proteomic tools aimed at mapping phosphorylation sites, measuring their dynamics in response to stimuli as well as kinase–substrate interactions, both experimentally and computationally. Phosphoproteomics aims to provide a comprehensive view of phosphoproteomes. Typical experiments involve the enrichment of complex peptide mixtures for peptides containing phosphorylated residues using chemical modifications or direct enrichment by affinity-based methods, followed by the identification of peptides and the localization of the modified residues by mass-spectrometry [23]. These approaches are usually very stringent, with estimated false-positive rates of <5%, and are highly reproducible [23]. From an evolutionary perspective, such approaches enable one to carry out genome-wide studies of the conservation and divergence of phosphosites.

evolved rapidly when considered as a whole. The median rate of yeast phosphosite evolution is ~26% above that of randomly sampled S/T and, for human, rates are similar (Figure 1c). It has previously been observed that phosphosites frequently appear outside functional domains and in disordered regions of proteins [8,9]. Because we know that disordered regions evolve faster than ordered regions [10], our observation that the phosphoproteome has a high turnover could result from the over-representation of phosphosites in these fast evolving regions. Indeed, we found that 79% and 84% of phosphosites occur in disordered regions for yeast and human, respectively, which is similar to what was

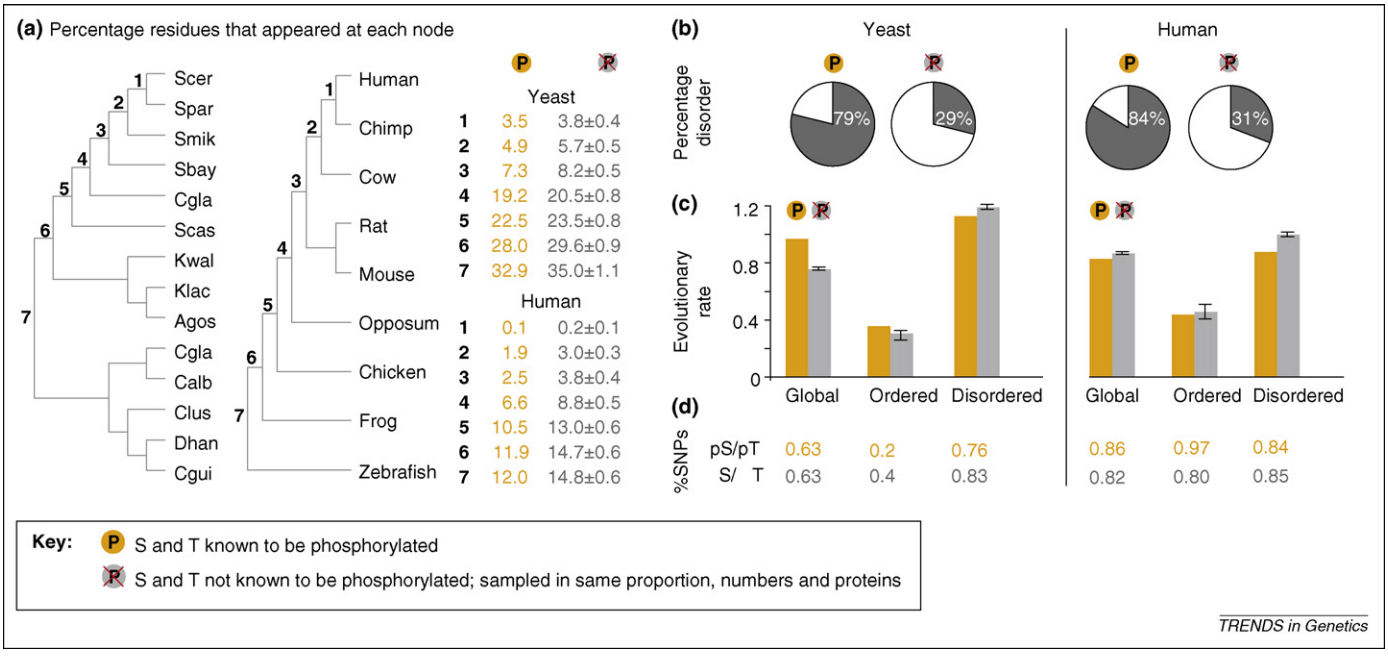

Figure 1. Phosphoproteomes of yeast and human evolve at a rate comparable to non-phosphorylated residues. **(a)** Ancestral sequences were reconstructed and the time at which each residue 'appeared' in the yeast or human proteomes was inferred. **(b)** Phosphosites are over-represented in unstructured regions of eukaryotic proteomes. **(c)** Rate of evolution of phosphosites along the yeast and human lineages as inferred by rate4site (http://www.tau.ac.il/~itaymay/cp/rate4site.html). The overall median rates are presented along with the rates for the ordered and disordered regions. Rates of phosphosites are compared with a random set of S/T. The set is controlled for the representation of serines and threonines and for the relative representation of proteins in the phosphoproteomes (see supplementary material online). The over-representation of phosphosites in disordered regions increases the overall rate of phosphoproteome evolution. That is, fast evolving phosphosites in disordered regions contribute ~80% to the overall rate, whereas randomly sampled S/T contribute only ~30% to it. This is why yeast phosphosites seem to be less conserved than the control set, but become more conserved as we take disorder into account. Error bars indicate two standard deviations estimated from 50 control sets. **(d)** Phosphosites are not subject to a significant reduction of polymorphism in yeast and human populations compared with S/T.

observed for a smaller set of mouse proteins [8]. These percentages are over twofold that of the global proportion of S/T in unstructured regions of these proteomes, confirming a strong enrichment of pS/pT in these regions (Figure 1b; $p \sim 0$, $Z_{score} = 63$). The difference in rate between 'ordered' and 'disordered' phosphosites is also significant, being over threefold in yeast and twofold in human (Figure 1c). Thus, the order–disorder dichotomy explains the overall high rate, whereby fast evolving phosphosites contribute 79% (yeast) and 84% (human) to the global rate, even after controlling for the lower alignment quality of unstructured regions (Table S3).

Our inter-species comparisons are supported by the evolution of individuals within yeast and human populations. We examined two entire genomes of natural strains of yeast [11] and found that phosphosites are as likely to be polymorphic as non-phosphorylated residues, with 0.62% of pS/pT being polymorphic, compared with 0.63% for S/T (Fisher's exact test, $p = 0.92$) (Figure 1d, Table S2). Thus, functional constraints imposed by protein phosphorylation on phosphosites, if indeed there are any, do not significantly limit their polymorphism in natural populations. There is also no clear depletion of polymorphisms at phosphorylated compared with

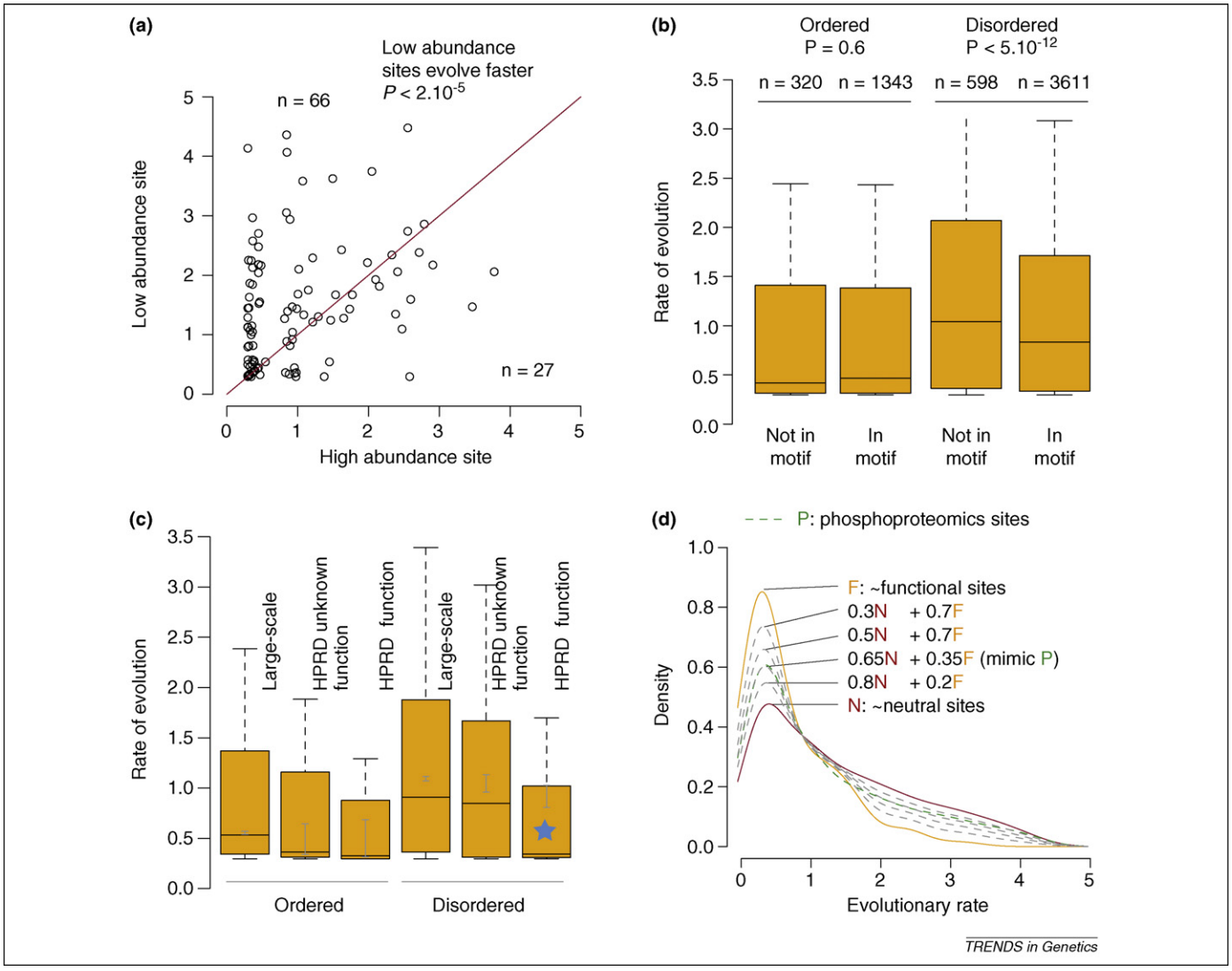

Figure 2. Function constrains the evolution of phosphosites. **(a)** Highly phosphorylated sites tend to be more conserved. Among pairs of phosphosites occurring on the same peptide, one site was assigned a high abundance and the other a low abundance according to the number of times they were identified (the number of times *N* a site is detected). We consider pairs where $N_{high}/N_{low} > 2$ and $N_{high} > 4$. Among these pairs, the low abundance site evolves faster than the high abundance one. **(b)** Human phosphosites that can be confidently associated with a kinase consensus motif evolve slower on average than those that cannot be confidently assigned. Note that, although significant, the difference is small, which puts into question the use of motifs to identify functional sites. The boxes correspond to the upper and lower quartiles and the dashed vertical lines extend to the most extreme data points no more than 1.5 times the inter-quartile range. For clarity, outliers are not shown. **(c)** Sites for which functional evidence exists evolve slower on average than other sites. 'Large scale' contain phosphosites from phosphoproteomic experiments (Table S1); 'HPRD unknown function' are sites contained in the curated HPRD database for which no evidence for a functional role is reported. Phosphosites in the 'HPRD function' category have been assigned a function, for instance by site-directed mutagenesis and functional assays. The gray error bars indicate the median evolutionary rate of a control set of S/T among the same proteins. The largest difference in evolutionary rate is seen in disordered regions (blue star), suggesting that these regions are particularly prone to non-functional phosphorylation events. **(d)** Estimating the prevalence of non-functional phosphorylation events in disordered regions. We plot the rate density distributions of three categories of sites: (i) neutral sites (red), approximated by random non-phosphorylated serines in disordered regions; (ii) functional sites (yellow), approximated by functional sites from HPRD in disordered regions; and (iii) phosphoproteomics sites (Table S1). We then searched for the combination of (i) and (ii) that best approximates (iii). This indicates that ~65% of phosphoproteomic sites in disordered regions evolve indistinguishably from non-phosphorylated sites and might not be functional. We did not carry out this analysis for ordered regions because it is difficult to define a 'neutral' population of sites for these.

non-phosphorylated residues in human populations (Figure 1d). Interestingly, these analyses show that the conservation of an S/T residue is much more influenced by the region in which the residue is found than by whether it is a phosphosite or not; why is this? A simple explanation could be that a fraction of phosphosites could be non-functional. As such, they could mutate freely and their evolutionary rate would then depend mostly on the region in which they lie. Therefore, we next examined this possibility.

How many non-functional phosphosites are there?

To investigate the possibility that the high phosphoproteome turnover results from a large number of phosphosites being non-functional, we assessed whether or not the functionality of phosphosites covaries negatively with their rate of evolution. We used two proxies for phosphosite functionality. First, because non-functional phosphorylations are likely to represent off-target interactions, they should be rare molecular events and thus should be in lower abundance than functional sites. We thus compared the rate of evolution of high and low stroichiometric phosphosites (see supplementary material online). We found that phosphosites that are more phosphorylated evolve slower on average than less abundant phosphosites for both yeast ($r_{high} = 0.95$, $r_{low} = 1.08$, $p < 7.10^{-4}$, paired-Wilcoxon test, Figure S2) and human ($r_{high} = 0.44$, $r_{low} = 1.18$, $p < 2.10^{-5}$; Figure 2a). Second, we predicted that if rapidly evolving sites result from non-functional, supposedly non-specific, encounters with kinases, they should be under-represented in consensus kinase recognition motifs. To test this hypothesis, we used the NetPhorest (http://netphorest.info/) classifier to predict which kinase phosphorylates each human phosphosite [12]. The confidence score provided for each assignment was used to separate phosphosites into two categories, based on whether a kinase could be confidently assigned or not (Figure 2b, Figure S3). As expected, phosphosites in known motifs are more conserved, but only if they are also present in a disordered region ($r_{motif+disorder} = 0.83$, $r_{nomotif+disorder} = 1.04$, $p < 5.10^{-11}$, Wilcoxon test). Surprisingly, sites in ordered regions do not seem to be affected by their presence within a motif ($r_{motif+order} = 0.47$, $r_{nomotif+order} = 0.42$, $p = 0.6$, Wilcoxon test). This indicates that filtering phosphosites according to their presence within motifs should be applied primarily to sites also present in disordered regions.

Finally, to directly estimate how functionality could influence the rate of phosphosite evolution, we examined abstracts from the literature associated with >500 phosphosites obtained from the Human Protein Reference Database (HPRD; www.hprd.org). We then compared the rate of evolution of pS/pT that were shown to have a functional role by direct or indirect evidence (e.g. site-directed mutagenesis and functional assay) with that of phosphosites reported in this database without evidence for a functional role. We found that among the HPRD phosphosites, those characterized as having specific functions do evolve significantly slower on average than those reported without functional characterization ($r_{function} = 0.33$, $r_{unknown\_func} = 0.82$, $p < 10^{-4}$, Wilcoxon test, Figure 2c). Interestingly, functional sites evolve at a relatively similar rate regardless of whether they occur in disordered or ordered regions ($r_{func+order} = 0.33$, $r_{func+disorder} = 0.34$, $p = 0.4$). As a result, the difference in evolutionary rate seen for functional sites is strongest in disordered regions ($r_{func+disorder} = 0.34$, $r_{unknown\_func+disorder} = 0.85$), suggesting that they are particularly rich in non-functional sites. To estimate what fraction of non-functional sites could be present in the current phosphoproteomes, we examined what mixture of rate distribution of functional and non-functional sites would be consistent with the distribution observed in unstructured regions. We estimate that as much as 65% of sites detected by phosphoproteomic experiments are not under stronger constraints than random S/T in disordered regions, which suggests that they might be non-functional (Figure 2d) or that the loss of such phosphosites could be compensated for by other mechanisms.

Concluding remarks

Overall, phosphoproteomes evolve rapidly at a rate comparable to that of non-phosphorylated residues. Our results suggest that this high rate is largely explained by two linked factors: first, most phosphosites occur in disordered regions and these evolve rapidly; second, experimental setups used to characterize phosphosites are highly sensitive and detect a fraction of sites that could have no function. These non-functional phosphosites might result from the off-target activity of protein kinases. Interestingly, disordered regions in particular seem to be subject to this 'noisy phosphorylation', possibly because residues in these regions are more accessible to kinases and because recognition of phosphorylation motifs in these regions is less dependent on the tertiary structure, making them more permissive substrates [13]. Indeed, disordered proteins are substrates for twice as many kinases as structured proteins in *in vitro* kinase assays [14].

It is often assumed that protein–protein interactions, such as kinase–substrate interactions, are highly specific. However, it is unclear how much noise is tolerated by the cell and how many non-functional interactions there are. Our results support our recent proposal that non-functional interactions could be important and ought to be taken into consideration [15]. Another possibility, which we did not investigate here, is that the function of phosphosites might be dependent on the structure of the region in which they are found. For instance, phosphosites could be used to regulate bulk electrostatic charges in these regions [16], such that the position of pS/pT is less important and can be compensated by other sites in the region. This would result in a faster rate of pS/pT evolution in disordered regions. Under this scenario, the number of phosphosites in a region would tend to be more conserved than their position. Thorough comparative analyses of closely related species combined with functional experiments will provide estimates for the importance of this mechanism.

Our results stress the need to consider phosphosites relative to their structural environment (e.g. order or disorder), and especially for those sites in disordered regions, orthogonal information should be used to maximize the representation of functional sites. Other types of

information could also be used, for instance, kinase recognition motifs and the abundance of phosphosites. Because phosphosites do evolve slowly when they are functional, the evolutionary rate can also provide useful orthogonal information. For this purpose, we provide the rate of evolution of all the phosphosites and their predicted kinases analysed here, such that they can be used to choose the most likely functional phosphosites candidates (http://tinyurl.com/phosphoevo). Ultimately, a combination of functional and evolutionary analyses will simplify the problem of determining what fraction of species-specific phosphosites contributes to their phenotypic differentiation and will help prioritize research aimed at understanding the molecular bases of human disorders involving dysregulation of protein phosphorylation.

Acknowledgements

We thank S. De for suggesting the evolutionary rate calculation methodology, I. Wapinski for providing yeast orthogroups and alignments, H. Zhou for providing the redundant list of yeast phosphopeptides, P. Fontanillas, S. Teichmann, N. Lartillot, M. Zilversmit, N. Aubin-Horth and the three anonymous reviewers for helpful comments on the manuscript, and H. Philippe for helpful discussions. This research was supported by the CIHR (MOP-GMX-152556) to S.W.M. C.R.L. and E.D.L. are CIHR and EMBO Postdoctoral fellows, respectively, and S.W.M. is the Canada Research Chair in Integrative Genomics.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2009.03.003.

References

1 Boekhorst, J. *et al.* (2008) Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. *Genome Biol.* 9, R144
2 Gnad, F. *et al.* (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.* 8, R250
3 Ubersax, J.A. and Ferrell, J.E., Jr (2007) Mechanisms of specificity in protein phosphorylation. *Nat. Rev. Mol. Cell Biol.* 8, 530–541
4 Lienhard, G.E. (2008) Non-functional phosphorylations? *Trends Biochem. Sci.* 33, 351–352
5 Malik, R. *et al.* (2008) Comparative conservation analysis of the human mitotic phosphoproteome. *Bioinformatics* 24, 1426–1432
6 Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol* 24, 1586–1591
7 Mayrose, I. *et al.* (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.* 21, 1781–1791
8 Collins, M.O. *et al.* (2008) Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder. *Mol. Cell. Proteomics* 7, 1331–1348
9 Nuhse, T.S. *et al.* (2004) Phosphoproteomics of the *Arabidopsis* plasma membrane and a new phosphorylation site database. *Plant Cell* 16, 2394–2405
10 Brown, C.J. *et al.* (2002) Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* 55, 104–110
11 Doniger, S.W. *et al.* (2008) A catalog of neutral and deleterious polymorphism in yeast. *PLoS Genet.* 4, e1000183
12 Miller, M.L. *et al.* (2008) Linear motif atlas for phosphorylation-dependent signaling. *Sci. Signal.* 1, ra2
13 Tompa, P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.* 579, 3346–3354
14 Gsponer, J. *et al.* (2008) Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. *Science* 322, 1365–1368
15 Levy, E.D. *et al.* (2009) How perfect can protein interactomes be? *Sci. Signal.* 2, pe11
16 Serber, Z. and Ferrell, J.E., Jr (2007) Tuning bulk electrostatics to regulate protein function. *Cell* 128, 441–444
17 Manning, G. *et al.* (2002) Evolution of protein kinase signaling from yeast to man. *Trends Biochem. Sci.* 27, 514–520
18 Cohen, P. (2000) The regulation of protein function by multisite phosphorylation–a 25 year update. *Trends Biochem. Sci.* 25, 596–601
19 Wenta, N. *et al.* (2008) Tyrosine phosphorylation regulates the partitioning of STAT1 between different dimer conformations. *Proc. Natl. Acad. Sci. U. S. A.* 105, 9238–9243
20 O'Neill, E.M. *et al.* (1996) Regulation of PHO4 nuclear localization by the PHO80-PHO85 cyclin-CDK complex. *Science* 271, 209–212
21 Martin, H. *et al.* (2005) Protein phosphatases in MAPK signalling: we keep learning from yeast. *Mol. Microbiol.* 58, 6–16
22 Rider, M.H. *et al.* (1992) The two forms of bovine heart 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase result from alternative splicing. *Biochem. J* 285, 405–411
23 Bodenmiller, B. *et al.* (2007) Reproducible isolation of distinct, overlapping segments of the phosphoproteome. *Nat. Methods* 4, 231–237

0168-9525/$ – see front matter
doi:10.1016/j.tig.2009.03.003 Available online 6 April 2009

Upcoming conferences

May 13–17: Chromatin and Epigenetics EMBL, Heidelberg, Germany
http://www-db.embl.de/jss/EmblGroupsOrg/conf_112

May 23–26: European Human Genetics Conference, Vienna, Austria http://www.eshg.org/eshg2009/

June 3–7, 2009: SMBE Annual Meeting: Darwin to the Next Generation Iowa City, Iowa. http://ccg.biology.uiowa.edu/smbe/

September 16–19: Annual Conference of the German Genetics Society Cologne, Germany http://www.genetics2009.de/

October 25–30: International Congress on Plant Molecular Biology St. Louis, MO, USA www.ipmb2009.org