Evolution of Characterized Phosphorylation Sites in Budding Yeast

Alex N. Nguyen Ba[1,2] and Alan M. Moses*†,[1,2]

[1]Department of Cell and Systems Biology, University of Toronto, Toronto, Canada

[2]Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Canada

†Research performed at the University of Toronto.

***Corresponding author:** E-mail: alan.moses@utoronto.ca.

Associate editor: Claudia Schmidt-Dannert

Abstract

Phosphorylation is one of the most studied and important regulatory mechanisms that modulate protein function in eukaryotic cells. Recently, several studies have investigated the evolution of phosphorylation sites identified by high-throughput methods. These studies have revealed varying degrees of evidence for constraint and plasticity, and therefore, there is currently no consensus as to the evolutionary properties of this important regulatory mechanism. Here, we present a study of high-confidence annotated sites from budding yeast and show that these sites are significantly constrained compared with their flanking region in closely related species. We show that this property does not change in structured or unstructured regions. We investigate the birth, death and compensation rates of the phosphorylation sites and test if sites are more likely to be gained or lost in proteins with greater numbers of sites. Finally, we also show that this evolutionary conservation can yield significant improvement for kinase target predictions when the kinase recognition motif is known, and can be used to infer the recognition motif when a set of targets is known. Our analysis indicates that phosphorylation sites are under selective constraint, consistent with their functional importance. We also find that a small fraction of phosphorylation sites turnover during evolution, which may be an important process underlying the evolution of regulatory networks.

Key words: phosphorylation sites, evolution, prediction.

Introduction

Protein phosphorylation is a ubiquitous posttranslational modification in cells as a means to regulate a variety of cellular processes (Johnson and Hunter 2005). Despite its importance, until recently, few studies had examined the evolution of this regulatory mechanism. Phosphorylation sites are critical functional elements within proteins, and therefore, they are expected to be conserved over evolution. This conservation can be exploited to predict kinase substrate interactions (Budovskaya et al. 2005). However, two recent studies examined the evolution of phosphoregulation in the eukaryotic cell cycle and found evidence for evolutionary changes in the regulatory networks (Jensen et al. 2006; Moses, Liku, et al. 2007). Furthermore, a structural study of phosphorylation sites in mitotic proteins found similar levels of conservation between phosphorylation sites and other similar residues (Jiménez et al. 2007), suggesting no specific constraints on these sites.

With the availability of high-throughput data sets, it has become possible to examine the evolutionary properties of large sets of phosphorylation sites (Macek et al. 2008; Holt et al. 2009; Landry et al. 2009; Yachie et al. 2009). Most studies have found evidence for evolutionary conservation of phosphorylated S/T/Y residues compared with unphosphorylated residues (Gnad et al. 2007; Macek et al. 2008; Malik et al. 2008; Landry et al. 2009). In addition, one high-throughput study compared phosphorylation patterns between distantly related yeast species and quantified the rate of evolution of these patterns (Beltrao et al. 2009). Despite providing evidence for constraint, these studies all identified a large number of phosphorylation sites that were not preserved over evolution. These nonconserved sites may contribute to the large difference of patterns of phosphorylation between species (Beltrao et al. 2009). However, it is also important to consider whether many of the sites contained in high-throughput data sets are not critical to protein regulation; for example, some fraction of sites obtained by mass spectrometry may not be functional sites (Lienhard 2008; Landry et al. 2009). These nonfunctional sites are not expected to be preserved over evolution and therefore may appear as evolutionary changes.

Another important issue is that the alignments used in some of the previous studies include sequences from distantly related species, which creates uncertainty in the analysis because short degenerate motifs such as phosphorylation sites may not be aligned accurately in distant species comparisons (Balla et al. 2006).

Motivated to address these difficulties, we sought to examine the evolution of a large set of high-confidence phosphorylation sites, where we could obtain high-confidence alignments of orthologous protein sequences. To do so, we assembled 249 characterized phosphorylation sites in budding yeast from the literature, where the likely kinase responsible for phosphorylation is known. By examining alignments of protein sequences from closely related species we can explicitly test evolutionary hypotheses about phosphorylation sites using the ratio of nonsynonymous to synonymous substitutions (*Ka*/*Ks*) (Nei and Gojobori 1986). Our results show that the rate of amino acid substitution within the site is lower than the surrounding region and that this property is observed whether the sites appear in structured or unstructured regions of the substrate proteins. As expected, we find that the patterns of substitution in phosphorylation sites are consistent with the specific constraints imposed by the consensus recognition site for the kinase. We also investigate the birth and death and compensation rates of these annotated sites and show that there are evolutionary constraints on the appearance and disappearance of sites in targets of kinases, but only weak constraints on compensation. We also consider the possibility that gain and loss of phosphorylation sites is due to redundancy, but we find no evidence that sites are more likely to be lost or gained in proteins with high number of sites.

Finally, we show that the evolutionary conservation of phosphorylation sites relative to surrounding amino acid sequence can be exploited to improve prediction of kinase substrates or to find the kinase specificity.

Materials and Methods

Alignment of Closely Related Species of Yeasts

Genomic sequences from the four species in our study (*Saccharomyces cerevisiae*, *Saccharomyces bayanus*, *Saccharomyces paradoxus*, and *Saccharomyces mikatae*; Kellis et al. 2003) were obtained from the SGD (SGD project, 2009) and translated open reading frames were aligned using t-coffee (Notredame et al. 2000) at default settings. DNA sequences were aligned using the aligned protein sequences by inserting the gaps from the protein sequence alignments into the cDNA sequences. In all, 86% (5045/5884) of the genes in *S. cerevisiae* were aligned successfully. The amino acid sequences of these species are very similar: 73% of the columns in these alignments have no amino acid differences.

Consensus Sequences of Phosphorylation Sites

The phosphoacceptor for each kinase was aligned, and the flanking sequences were added afterward. We created a seqlogo (Schneider and Stephens 1990; Crooks et al. 2004) for each kinase and set the consensus sequence to start and end where the information content equaled 1.

We defined critical residues as those residues that are likely to be necessary for phosphorylation. These include the phosphoacceptor and residues information content comparable with the phosphoacceptor. We defined degenerate residues to be residues with lower but observable information content and nonspecific residues as residues with marginal information content.

We defined phosphorylation sites as the consensus sequence match of the respective kinase. This includes the phosphoacceptor as well as critical, degenerate, and nonspecific residues as described above.

Ka and *Ks* Calculation

To calculate the rate of synonymous (or nonsynonymous) substitution, *Ks* (or *Ka*), we calculated the number of synonymous (or nonsynonymous) substitutions and divided by the number of synonymous (or nonsynonymous) sites. This calculation was done either on individual columns of alignments or on the "site" and "flank." To calculate the number of substitutions, we used the maximum parsimony algorithm (Durbin et al. 1998) with no weighting on the amino acid sequence, and to calculate the number of synonymous or nonsynonymous sites, we used the method presented by Nei and Gojobori (1986).

Error bars were obtained by nonparametric bootstrapping of 1,000 samples with a 95% confidence interval (Nei and Kumar 2000). *P* value from bootstrap analysis is obtained by counting the number of times the *Ka*/*Ks* of the flanking region is slower than the site divided by the number of samples. Significance is assessed at *P* value <0.05.

A Likelihood Ratio Test of Two Rates of Substitution Against One

We sought to test the hypothesis that the phosphorylation site evolved at a slower rate than its flanking region. To do so, we compared that hypothesis against the null hypothesis that the whole region (site and flank) evolved at a constant rate using a likelihood ratio test (LRT). Formally:

$$\mathrm{logLR} = \log \frac{f(x|\lambda)}{f(x_{\mathrm{flanks}}|\lambda_{\mathrm{flanks}})f(x_{\mathrm{site}}|\lambda_{\mathrm{site}})},$$

where x is the observed number of substitutions at a given position, and λ is the rate of evolution, and where $\lambda_{\mathrm{flank}} \geq \lambda_{\mathrm{site}}$. We assumed that substitutions occurred following a Poisson process.

$$f(x|\lambda) = \frac{\lambda^{x} e^{-\lambda}}{x!}.$$

Assuming a single *Ks* rate over the whole region, the likelihood ratio only depends on the amino acid substitution rate, *Ka*. The maximum likelihood estimate of *Ka* is simply the number of nonsynonymous substitutions divided by the number of nonsynonymous sites.

After some algebra, the log likelihood ratio is

$$\mathrm{logLR} = S_{\mathrm{flank}} \log \frac{\lambda}{(\lambda_{\mathrm{flank}})} + S_{\mathrm{site}} \log \frac{\lambda}{(\lambda_{\mathrm{site}})},$$

where S is the number of nonsynonymous substitutions, and λ is the maximum likelihood estimate of the rate of amino acid substitution (*Ka*).

The LRT statistic is given by LRT=2logLR. Under the null hypothesis, this statistic follows the χ^2 distribution with degrees of freedom equal to 1. Significance is assessed at *P* value <0.05.

Structure

To assess if a site was present in a structured or unstructured region, we used the method presented by Uversky et al. (2000) (Prilusky et al. 2005), first by removing the annotated phosphorylation site from the protein segment and then using a window of 50 amino acids centered on the region of the phosphorylation site.

Turnover Rate Calculation

Turnover rate is defined by both death and birth rates. We defined death rate as the number of sites disappearing from the inferred ancestral sequence divided by the number of initial sites in the ancestral sequence, and we defined birth rate as the number of sites appearing along the lineage leading to *S. cerevisiae* after the divergence with *S. mikatae* or *S. paradoxus* per 1,000 residues of the ancestral sequence.

Significance of birth and death rate differences was assessed using a two-tailed Fisher's exact test by summing the probability of more extreme possible observations. A *P* value $<$0.05 was assessed as significant.

Site compensation was defined as a pair of nonconserved phosphorylation site and birth within the same lineage, within a local region of the protein. The distance allowed for the site birth was halfway until the next predicted site within *S. cerevisiae*.

Site and Target Prediction

For various controls and kinase target prediction, we predicted phosphorylation sites using a profile hidden Markov model (HMM) obtained from our initial alignments of sites. Profile HMMs have been used in the past to predict protein domains and model linear states that approximates a consensus sequence (e.g., Pfam; Finn et al. 2008). Although this may not be needed for kinases such as Mec1p, which follow a strict consensus sequence, other kinases such as cyclin-dependent kinase (CDK) have "weak" and "strong" consensus matches that offer more leeway in their recognition signal. Because HMMbuild (Eddy 1998) was found to be more reliable for longer sequences than most phosphorylation recognition signal, we built a similar model using a single Dirichlet prior that fitted most with one of the Dirichlet mixture for pseudocount (Sjölander et al. 1996). We then used the posterior algorithm and a threshold that validated most of our annotated sites to predict putative sites.

For CDK, kinase target prediction was assessed using the proteins identified as substrates by Ubersax et al. (2003) as positives and the remaining proteins tested by Ubsersax et al. as negatives (nontargets). We calculated the positive predictive power by counting the number of positive proteins above an LRT threshold divided by the total number of proteins above the same threshold among the two sets. For Mec1, kinase target prediction was assessed using the proteins with at least one characterized Mec1 phosphorylation site as positives and all proteins with characterized phosphorylation sites for other kinases, but not Mec1, as negatives. Although some of these proteins may indeed be targets of Mec1, we hoped that by using this set as negatives, we would reduce the effect of the bias that is induced by researchers when they choose which proteins to study.

Results

A Set of Functional Phosphorylation Sites in *S. cerevisiae* for Which the Kinase is Known

In order to study the evolutionary properties of phosphorylation sites, we searched the literature for experimentally verified phosphorylation sites where the kinase had been identified in low-throughput experiments. Although there is no single experiment that conclusively shows that a specific site is phosphorylated by a specific kinase in vivo, we chose to include phosphorylation sites where 1) site-specific mutagenesis on the phosphoacceptor site (usually S/T/Y to A to create nonphosphorylatable mutants) has revealed a functional role for the site or group of sites or 2) low-throughput identification of phosphosites by mass spectrometry had identified sites. The vast majority of the sites included have been confirmed by site-specific mutagenesis. In addition, we required that each site has evidence for the specific kinase responsible, either by in vitro experiments showing phosphorylation of the site by that kinase or by in vivo experiments showing that the phosphorylation or mutant phenotype depended on a particular kinase. Because kinases usually recognize a short degenerate consensus sequence around the phosphorylated residue (Miller and Blom 2009), knowing the identity of the kinase responsible allows us to accurately define the extent of expected conservation around the phosphorylation site. This contrasts with previous studies on phosphoevolution where phosphorylation sites have been obtained with high-throughput mass spectrometry (Macek et al. 2008; Holt et al. 2009; Landry et al. 2009; Yachie et al. 2009) and where the kinase was unknown. In those studies, the evolutionary properties of only the phosphoacceptor sites can be studied.

We focused on seven kinases for which we could define a consensus sequence: CDK (or Cdc28p), Mec1p, CKII (Cka1p/Cka2p/Ckb1p/Ckb2p), Prk1p, Ipl1p, protein kinase A (PKA) (or Tpk1p/Tpk2p/Tpk3p), and Pho85p (see table 1). We manually aligned all the sites for each kinase and determined the extent of the consensus sequences based on the information content. These consensus sequences are represented as seqlogos (Schneider and Stephens 1990; Crooks et al. 2004) in figure 1. We refer to these sites as "annotated" phosphorylation sites, and we believe that they represent a high-confidence set of bona fide phosphorylation sites in budding yeast. These sites will be made available through a publicly available website (A.N.N.B., A. Hussin, A. Pogoutse, A.M.M., in preparation). A complete table of these sites and references can be found as supplementary table S1 (Supplementary Material online).

There is Evidence of Conservation of Phosphorylation Sites

We first sought to test for evidence of evolutionary constraint on the annotated phosphorylation sites. To perform our analysis, we aligned orthologous proteins from four closely related species of yeasts (*S. cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. bayanus*, see Materials and Methods). We

Table 1. Summary of the Sites Included in Our Analysis. *Ka*/*Ks* was Calculated as in the Materials and Methods. The LRT is the Likelihood Ratio Statistic and the *P* Value is Given Following a χ^2 with a Degree of Freedom Equal to 1 (see Materials and Methods). *P* Value From Bootstrap Analysis is Given by the Number of Times the *Ka*/*Ks* of the Flank was Observed to be Higher than the Site in 1,000 Nonparametric Bootstraps (see Materials and Methods). Double Asterisks Show Strong Significance ($P < 0.01$), and Single Asterisk Shows Significance ($P < 0.05$).

Kinase	Number of Phosphorylation Sites[a]	*Ka*/*Ks* in Site	*Ka*/*Ks* in Flank	LRT	*P* Value of LRT	*P* Value of Bootstrap
CDK	114	0.071	0.109	19.62	9.4×10^{-06}**	0.005 **
Mec1p	47	0.062	0.173	24.26	8.4×10^{-07}**	<0.001**
Ipl1p	18	0.14	0.157	0.008	0.93	0.37
CKII	17 (27)	0.05	0.01	3.06	0.08	0.003**
Prk1p	20 (23)	0.034	0.06	4.94	0.03*	0.128
PKA	14 (18)	0.126	0.057	0	1	0.986
Pho85p	19 (22)	0.025	0.062	5.64	0.018*	0.031*
Total	249 (269)	0.069	0.118	64.36	1×10^{-15}**	<0.001**

[a] In parenthesis is the total number of sites found in the literature. Our analysis excluded overlapping sites.

then used maximum parsimony (Durbin et al. 1998) (see Materials and Methods) to calculate the rate of amino acid and synonymous substitution (*Ka* and *Ks*) (Nei and Gojobori 1986) in the phosphorylation sites (termed "site"). Because phosphorylation sites occur preferentially in unstructured regions of proteins (Gnad et al. 2007; Landry et al. 2009), and phosphoproteins evolve more slowly than other proteins (Gnad et al. 2007), comparing them with a random sample of sites can be misleading. We therefore compared the rates of evolution in the characterized phosphorylation sites with five amino acids on each side (termed "flank"). We use the flanking region to control for structured and unstructured segments of proteins, as well as different rates of protein evolution. To explicitly test for a difference in substitution rate between the sites and flanks, we performed an LRT to compare the hypothesis that the site evolves at a different rate than the flanking region with the hypothesis that a single rate of evolution explains the patterns in both classes (see Materials and Methods). In each case we also performed a nonparametric bootstrap to confirm the significance of our results (see table 1).

Using this method, we found that there is a significant reduction in amino acid substitution rate within the phosphorylation sites as compared with the flanking regions (*Ka*/*Ks* 0.069 vs. 0.118 for sites and flanks, respectively, LRT = 64.36, *P* value $< 10^{-14}$; fig. 2*a* and table 1). This indicates that phosphorylation sites evolve under specific evolutionary constraint relative to the regions in which they occur in proteins.

Because we defined a phosphorylation site to include more information than the phosphoacceptor, we investigated the *Ka*/*Ks* ratio of amino acids defined as critical (substitution would very likely prevent phosphorylation),

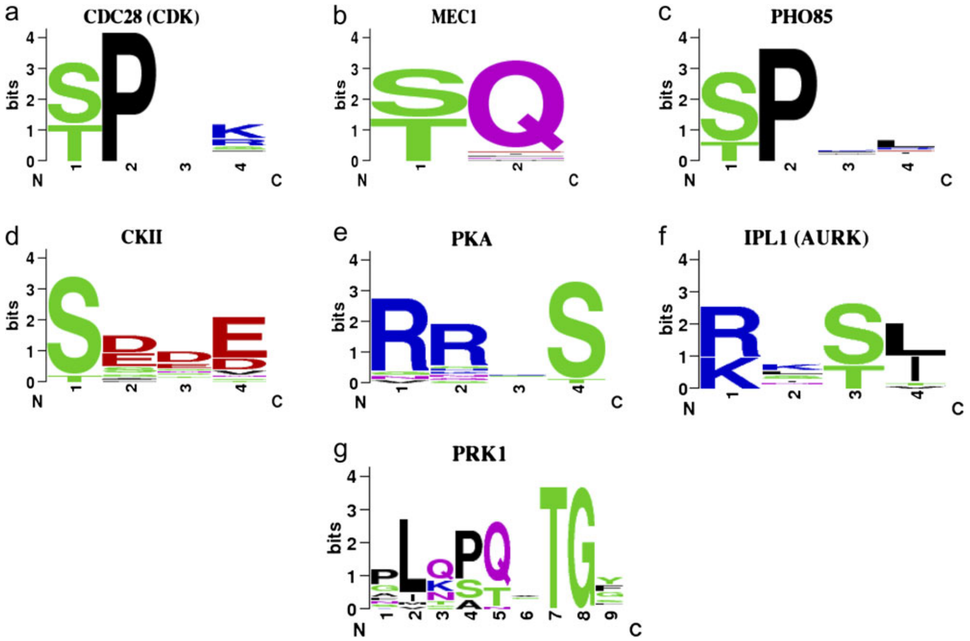

Fig. 1. Sequence logos of aligned annotated sites of seven kinases. (*a*–*g*) Annotated sites from each kinases were aligned along with their flanking regions, and boundaries were chosen where the information content was >1.

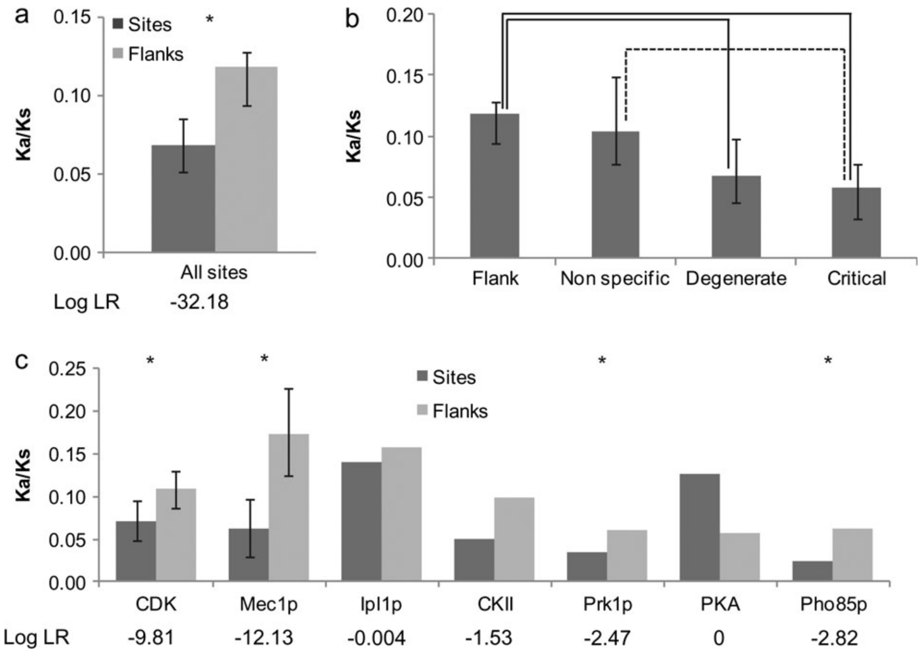

FIG. 2. Difference in *Ka*/*Ks* ratio between sites and their flanking regions. (*a*) *Ka*/*Ks* ratio of annotated sites and their flanking regions. (*b*) *Ka*/*Ks* ratio of amino acids denoted as critical, degenerate, or nonspecific and the flanking region of each site. Lines between each bar show significant differences under the LRT. (*c*) *Ka*/*Ks* ratio of annotated sites from each kinase. In all graphs, the error bars are obtained from a 95% confidence interval from a nonparametric bootstrap of 1,000 replicates. Error bars are only shown for cases with at least 40 phosphorylation sites. Significance from the LRT is shown as an asterisk ($P < 0.05$).

degenerate (substitution may lower phosphorylation affinity), or nonspecific (substitution is unlikely to impact phosphorylation). We categorized each position in the recognition motifs based on the information content (see Materials and Methods). We calculated the *Ka* and *Ks* at each position of the phosphorylation site consensus sequence and binned the sites according to the categorization. As expected, the *Ka*/*Ks* ratio of the amino acids defined as critical (0.057) is lower than degenerate amino acids (0.068), which are lower than nonspecific amino acids (0.104, fig. 2*b*).

Performing the analysis on each kinase independently reveals a lower *Ka*/*Ks* ratio in the sites versus the flanking residues for all the studied kinases but PKA (fig. 2*c*). The LRT indicates that most of the sites of kinases have a significantly lower rate of substitution than the flanking residues.

It is possible that this lower rate of substitution is due to the difference in frequency of particular amino acids within the consensus sequences, as compared with the flanking regions. To confirm that this could not explain our results, we used as a negative control 514 proteins which were shown not to be targets of CDK by Ubersax et al. (2003). We randomly sampled an equivalent number of sequences matching the CDK consensus from these "nontargets" and computed the *Ka*/*Ks* ratio as well as the LRT described above (see Materials and Methods). On average, these nontargets showed LRT statistics of 1.8 (SD = 2.23), much less than the LRT = 19.62 observed for the annotated sites. This indicates that the differences in amino acid composition between the sites and flanks cannot account for the large LRT statistics that we have observed in the annotated sites. We note that the nontargets include some fraction of false negatives, and therefore, this can be regarded as a conservative estimate for the contribution of the difference in residue frequencies. Therefore, at least for CDK sites, the lower rate of substitution within the site compared with the flanking region was not due to amino acid content as our negative and annotated sets show dramatically different results although having similar amino acid content.

We note that most of the sites in our data set appear in unstructured regions of proteins (75% unstructured and 25% structured), and a previous study (Landry et al. 2009) has stressed the importance of studying the context of the phosphorylation site in evolutionary analyses. However, we found that the constraint observed above is similar in both structured and unstructured regions (fig. 3).

Phosphorylation Site Turnover

Previous studies have shown that phosphoregulation may change over evolution (Moses, Liku, et al. 2007), and consistent with this, alignments of phosphorylation sites over long evolutionary distances show evidence of change (Holt et al. 2009; Tan et al. 2009). One possible explanation is that the sites may not be required to stay at a particular location in a protein and therefore may shift position over evolution (Moses, Liku, et al. 2007; Holt et al. 2009; Tan et al. 2009), especially in unstructured regions (Brown et al. 2002). Another explanation is that proteins with multiple sites may lose or gain a few sites without changing the regulation of the protein (Moses, Liku, et al. 2007; Serber and Ferrell 2007). In both these cases, functional phosphorylation site

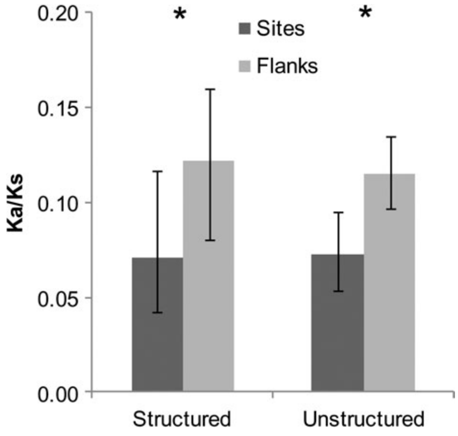

FIG. 3. Difference in *Ka/Ks* ratio between sites and their flanking regions categorized by structured and unstructured regions. *Ka/Ks* ratio and results of the LRT of annotated sites and their flanking regions for annotated sites separated by structured or unstructured regions. Error bars were obtained from a 95% confidence interval from a nonparametric bootstrap of 1,000 replicates. Significance from the LRT is shown as an asterisk ($P < 0.05$).

turnover does not impact protein function. However, a third possibility is that phosphorylation sites identified in high-throughput experiments may be nonfunctional, and the evolutionary changes we observe are simply due to the loss of nonfunctional residues.

We decided to test whether we could observe the microevolutionary steps that underlie the changes in functional phosphoregulatory networks. We therefore sought to quantify the rate of turnover of phosphorylation sites within our set of annotated sites. We considered sites that contained substitutions of the critical residues (see Materials and Methods) to be nonconserved. Of the 249 functional sites in our set, we found 22 (8.8 ± 1.8%) that were not conserved in the alignments of the closely related species studied here. We confirmed that these nonconserved sites were not due alignment errors or missing data (supplementary data and supplementary table S2, Supplementary Material online).

To test for evidence of selection influencing the rates of phosphorylation site turnover, we took advantage of the CDK unbiased "nontarget" set (Ubersax et al. 2003). If selection acts to preserve functional phosphorylation sites, we predict that characterized sites should be lost at a slower rate than similar sequences in the nontargets. On the other hand, if phosphorylation sites were recently added by positive selection, we would expect to see a faster rate of characterized phosphorylation site gain relative to the appearance of matches to the consensus sequence in proteins we know not to be targets. Another hypothesis for the observation of turnover is that constraint at the individual site is superseded by the constraint that total number of sites should be conserved. Therefore, if selection is acting to preserve the total number of sites in a protein, a "death" can be compensated by a "birth" nearby on the same lineage. To test these hypotheses, we compared the rate of birth (fig. 4*a* for an example), death (fig. 4*b* for an example), and compensation (fig. 4*c*) from our annotated CDK sites with the consensus sequences in the set of unbiased negative proteins (nontargets) from Ubersax et al. We chose the set of unbiased nontargets because the birth rate obtained from the total negative set would be biased, as these proteins were chosen to be tested on the basis of the presence of a consensus sequence.

We counted birth as sites appearing in the lineage leading to *S. cerevisiae* after the divergence with *S. mikatae* or *S. paradoxus* and deaths where sites disappeared in either *S. mikatae* or *S. paradoxus*, and compensation as a pair of birth and death within the same lineage. Doing so, we found that in our annotated sites, both the birth rate and the death rate were lower than in the set of unbiased nontargets (0.019 vs. 0.075 for deaths, $P = 0.04$, and 0.14 vs. 0.55 for births, $P < 0.01$, Fisher's exact test, fig. 4*d*). Although we observed an increased rate of compensation within our annotated set, it was not found to be significant (0.5 compensation/turnover vs. 0.2 compensation/turnover P value = 0.14, Fisher's exact test, fig. 4*d*). To control for the possibility that this birth and death rate difference is due to the difference in the amount of structured or unstructured regions in both sets, we also calculated the birth and death rates on only unstructured regions or structured regions. Doing so, we found similar results: Both the birth and death rates are lower in the annotated set whether or not we look at structured or unstructured regions (data not shown).

Taken together, this analysis indicates that functional sites are under selective constraint to be preserved and that the bona fide targets of the kinase are also less likely to spawn new sites, suggesting selection against spurious matches to the consensus. We propose that in real targets, the appearance of new sites is more likely to disrupt protein function (e.g., inappropriate phosphorylation of a protein domain) than in the nontargets where consensus matches are likely not to be phosphorylated (e.g., because the kinase is never localized close to the substrate).

If selection acts on the number of phosphorylation sites, rather than the specific residues, loss or gain of sites may be permissive in proteins with a high number of phosphorylation sites. We found that the average 8.8% site turnover was seen across all proteins regardless of their site count and found no significance with a simulation of random turnover event (fig. 5). We tested our statistical power to observe significance in this test and found that, in a simulation where we assumed a site was *n* times more likely to be lost in a protein with *n* sites than in a protein with a single site, we did have a large enough sample size to detect this effect.

Conservation of Phosphorylation Sites Can Improve Kinase Target and Specificity Predictions

Two important challenges in computational biology are predicting kinase substrates based on kinase specificity (Kobe et al. 2005; Turk 2008; Miller and Blom 2009) and predicting kinase specificity given a set of known substrates (Schwartz and Gygi 2005). We observed that

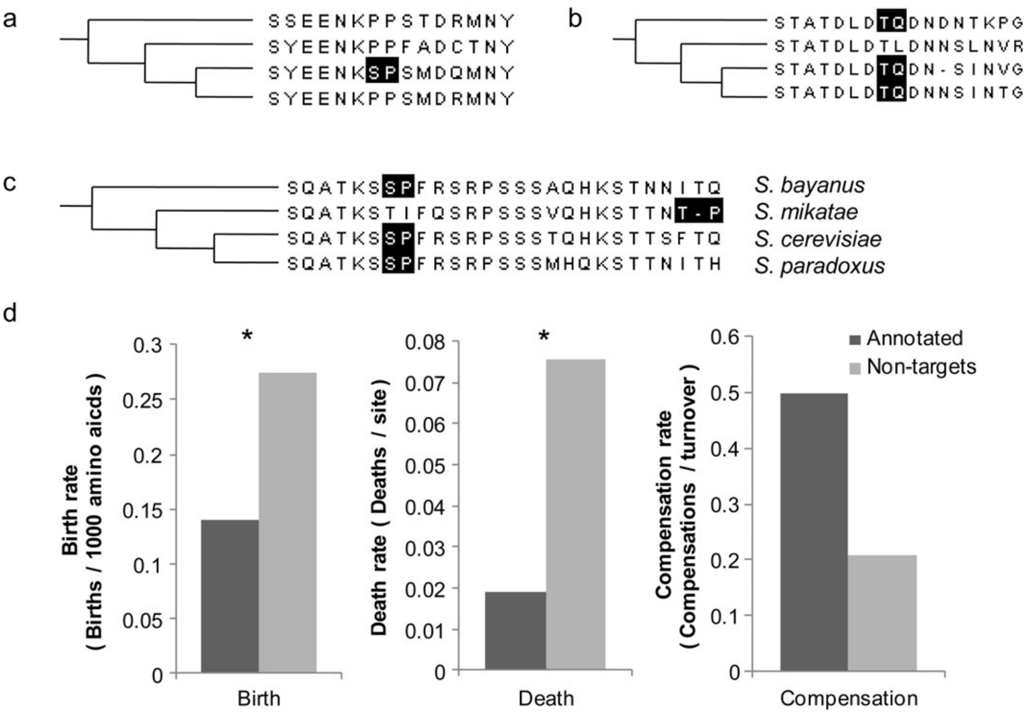

Fig. 4. Birth and death rate comparison. (*a*) Example of a site birth. Site shown is a CDK site in Cnm67p. (*b*) Example of a site death. Site shown is a Mec1p site in Mrc1p. (*c*) Example of site death compensated by a birth. Site shown is a CDK site in Tgl4p. (*d*) Birth, death, and compensation rates of our annotated CDK sites were compared with the consensus sequences appearing in the set of unbiased nontargets by Ubersax et al. Significance from a Fisher's exact test is shown as an asterisk ($P < 0.05$).

experimentally confirmed phosphorylation sites had a lower rate of substitutions than their flanking region. We sought to see if this information could be used to improve kinase target prediction or if it could uncover specific recognition motifs. To perform this analysis, we attempted to predict targets and specificity of CDK and Mec1p, the kinases for which we had the most available data.

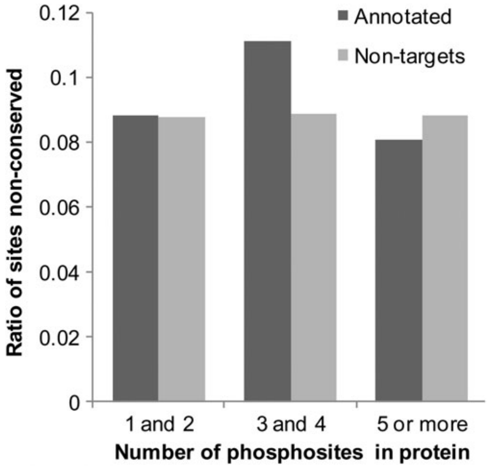

Fig. 5. Turnover in proteins with different number of sites. Percentage of nonconserved site within proteins of different number of sites. Significance of the different distributions was assessed with a χ^2 test at a significance level of 5% with degrees of freedom equal to 2.

We first attempted to predict kinase targets. For CDK, our set of positives and negatives were obtained from the set of CDK targets of Ubersax et al. using a CDK-as1 allele (Ubersax et al. 2003). To incorporate the slower rate of evolution of phosphorylation sites into sequence-based prediction, we applied the LRT described above to the matches to the CDK consensus sequence in each particular protein. Running our LRT on individual proteins, we observe that the likelihood ratio alone is a strong predictor of targets, yielding significant positive predictive value ($P < 0.05$, Fisher's test compared with the consensus sequence alone). Because proteins with more matches to the CDK consensus are more likely to represent bona fide targets of this kinase (Moses, Hériché, and Durbin 2007), we also analyzed targets separately depending on the number of matches to the full consensus.

We find that the improvement in positive predictive value is more pronounced when the number of full consensus matches within the protein is lowest (fig. 6*a*). This is likely due to the strong predictive power achieved in the case of large numbers of consensus sites in the absence of evolutionary information. The evidence for constraint on phosphorylation sites improves prediction in the cases where consensus sites alone provide poor predictive power.

We then performed a similar analysis on Mec1p targets. Our set of positives were the proteins with annotated Mec1p sites, and our set of negatives were protein targets of other kinases within our initial data set. Similar to CDK, we observe that the LRT on individual protein is a strong predictor of targets with significant positive predictive

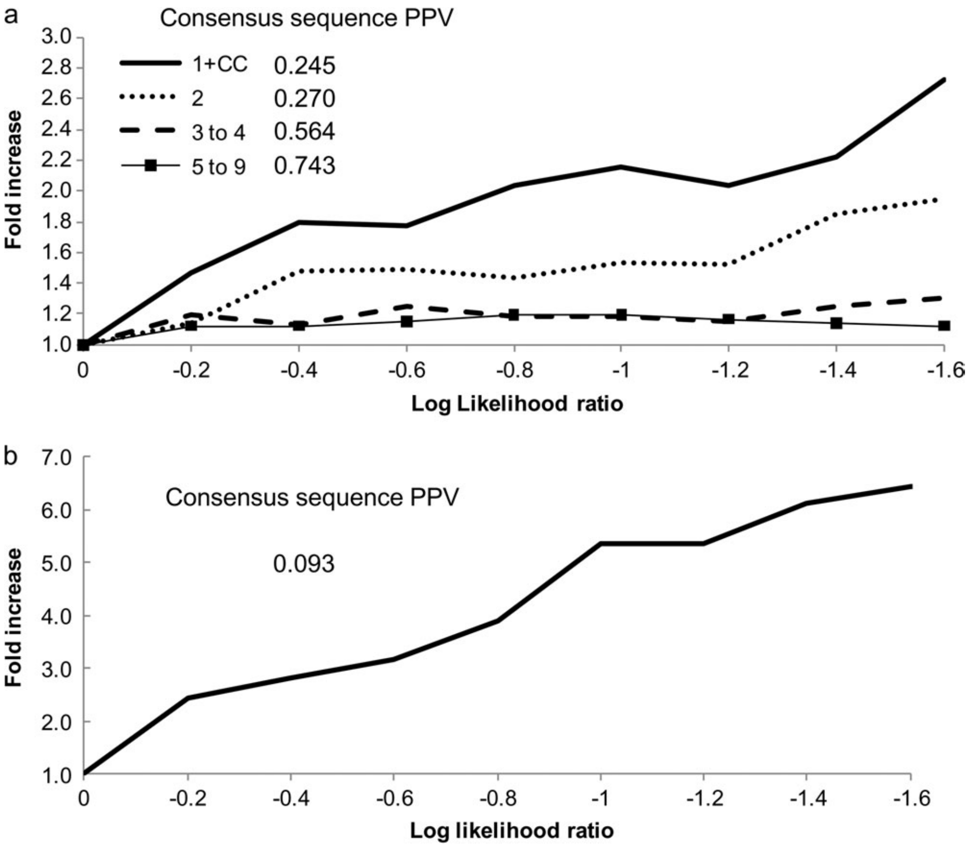

Fig. 6. Improved target prediction using the LRT. (*a*) Ratio of the positive predictive value of the LRT against the consensus sequence alone on the Ubersax et al. set of negative and positive targets of CDK. The number of matches to the full CDK consensus sequence was used to separate the set in multiple categories. (*b*) Ratio of positive predictive value of the LRT against the consensus sequence alone on our set of proteins with annotated phosphorylation site. Positive targets were genes that had annotated Mec1p sites, and nontargets were the rest of the proteins with annotated sites having a Mec1p consensus sequence.

value ($P < 0.05$, Fisher's test compared with the consensus sequence alone) (fig. 6*b*).

We next tested whether the evolutionary information could be used to search for the kinase recognition motif. For a given set of kinase targets, we identified all k-mers that included a serine or a threonine within unstructured regions and tested their conservation compared with their flanking region using the LRT. For k-mers that are found at least three times in the unstructured regions of the substrates, we found that the LRTs are sufficient to uncover both the CDK recognition motif ([ST]-P, indicated with circles in fig. 7*a*) and the Mec1p recognition motif ([ST]-Q, indicated with circles in fig. 7*b*). As a negative control, we performed a similar analysis on the non-CDK targets described above and did not recover the CDK consensus (data not shown).

This indicates that the evolutionary conservation can be complementary to the information about the number of consensus sites when predicting kinase substrates and that it can help in predicting kinase recognition motifs.

Discussion

Our study differs from previous studies in four main methodologies. First, the phosphorylation sites in our study were known to be functional. Second, we only included closely related species of yeasts. Third, we included important residues other than the phosphosite in the evolutionary rate calculations. Finally, the phosphosites were categorized by their respective kinases in order to test for differences between kinases. Our methodology was chosen to ensure that we obtained reliable alignments and to ensure that our analysis did not include falsely labeled phosphorylation sites. Thus, we have higher confidence in the alignments and the studied sites, but our conclusions are based on less data and substitutions, which are necessary to infer evolutionary properties.

Nevertheless, our analysis of the annotated phosphorylation sites in *S. cerevisiae* yielded several results which have been suggested in other studies (Gnad et al. 2007; Macek et al. 2008; Malik et al. 2008; Beltrao et al. 2009; Holt et al. 2009; Landry et al. 2009; Yachie et al. 2009): On average, phosphorylation sites show evidence of functional constraint, but individual sites appear to turnover during evolution. We note that the number of sites in our study is much smaller than many of the previous studies (and represents a small fraction of the total number of phosphorylation sites in the yeast proteome). Furthermore, as we only studied seven kinases, we note that it might not be possible to generalize our study to the whole phosphoproteome. Indeed, results vary between kinases: For example, a

	k-mer	LRT
●	SP	92.8727
●	TP	22.6324
●	PSP	21.7248
●	SPK	19.031
●	SPKK	16.2431
●	TPTK	13.3427
●	TPSK	12.55
●	SPVK	12.2602
	SSSY	11.7418
●	KTPS	11.7328
	SQ	11.4341
●	LSP	11.3636
●	KSPE	11.35
●	GSSP	11.1328
●	SSPVK	10.9455
●	TPRR	10.7002
	SSSSL	10.6253
●	KSP	10.3049
	TKQ	10.127
●	STPTK	9.99264
●	SPLK	9.33111
	SQGS	8.97087
	SSS	8.82163
●	SPV	8.6286
●	TPS	8.58965
●	FTPR	8.51738

b

	k-mer	LRT
●	TQ	13.6924
	SD	9.9984
	SNS	6.95316
●	SQ	6.9488
●	DTQ	6.58132
	DSD	4.93539
	DSE	4.77475
	TL	4.31038

FIG. 7. Kinase specificity prediction using the LRT. (*a*) K-mers with serines or threonines ranked by their likelihood ratio statistics in the unstructured regions of CDK targets. Black circles are k-mers that fit the known CDK consensus sequence. (*b*) K-mers with serines or threonines ranked by their likelihood ratio statistics on Mec1p targets. Black circles are k-mers that fit the known Mec1p consensus sequence.

PKA sites did not show evidence for constraint relative to their flanking sequences. At least in the case of CDK, however, our results seem to be generalizable as we observe similar results (supplementary data, Supplementary Material online) on putative phosphorylation sites in a large set of CDK targets (Ubersax et al. 2003).

In addition, conservation of sites within very closely related species of yeasts has been observed for sites within targets of CDK (Holt et al. 2009), and the lower death rate in annotated CDK sites that we observed is consistent with the observation that enrichment of sites is also maintained over evolution (Holt et al. 2009).

In another study, it had been shown that phosphorylated residues from high-throughput data in yeast do not appear to be constrained when compared with nonphosphorylated serines/threonines/tyrosines (Landry et al. 2009). Further analysis from that study showed that constraint was significantly observed when comparing phosphorylated residues with known function instead in a human data set. Consistent with this, we found that constraint could not be observed by comparing the phosphoacceptor with its flanking residues in their yeast high-throughput data set (supplementary data, Supplementary Material online). Because our sites were all shown to be functional, our analysis confirms the idea that functional sites are more likely to be preserved (Budovskaya et al. 2005; Landry et al. 2009).

The above indicates that evolutionary information can therefore be used to infer functional phosphorylation sites in proteins known to be phosphorylated by a certain kinase. Indeed, some studies in the past have taken advantage of this information (Wang et al. 2005; Koch et al. 2009) for their protein of interest. Evolutionary conservation has also been used systematically to predict novel substrates of PKA (Budovskaya et al. 2005). Although we did not devise a method to predict targets of kinases, we confirmed that, as a proof of concept, simple evolutionary constraint improves predictive power in proteins with small numbers of full consensus sites.

Furthermore, kinase specificity has been predicted in the past using enrichment of linear motifs (Schwartz and Gygi 2005) as proteins phosphorylated by a kinase often share a common recognition motif around the phosphoacceptor. We also have shown that because these motifs tend to be conserved, evolutionary information can help in predicting the kinase recognition motif in the case where the substrates are known.

Although conservation seems to be a general feature of functional phosphorylation sites, by looking at well-characterized sites in closely related species, we quantified the process of evolutionary change in phosphorylation sites: We found that ∼9% of phosphorylation sites were not conserved in the closely related species considered here. Turnover of phosphorylation sites could be consistent either with redundancy of sites in unstructured multiply phosphorylated regions or with changes in phosphoregulatory networks (Moses, Liku, et al. 2007; Beltrao et al. 2009; Tan et al. 2009).

Consistent with the hypothesis that the total number of sites within a protein is conserved, and the individual sites are free to turnover, we observed examples of compensation within our annotated set, although we could not find statistical evidence for selection on this process. Similarly, we could not find any evolutionary evidence for redundancy of phosphorylation sites in multiply phosphorylated proteins.

On the other hand, it has been proposed that changes in regulatory networks are enabled by the presence of phosphorylation sites in unstructured regions that evolve rapidly (Collins 2009; Holt et al. 2009). However, our data showed similar constraints on phosphorylation sites in structured and unstructured regions, which is in agreement with previous results regarding functional sites (Landry et al. 2009). Another explanation for the prevalence of phosphorylation sites within unstructured regions is that phosphorylation of structured regions is more likely disrupt function. This is supported by the fact that there are fewer consensus matches in structured regions in the targets than expected (0.1846 in annotated targets vs. 0.2853 per 1,000 residues in nontargets). However, we also observed a reduction in the birth rate of phosphorylation consensus sites in bona fide targets relative to the nontargets, indicating that spurious phosphorylation sites may disrupt function even in the unstructured regions. Understanding the constraints on the organization of regulatory sequences in proteins is an important area for further research.

Nevertheless, our evidence for "turnover" of characterized phosphorylation sites provides the microevolutionary material for the plasticity in regulatory networks that has been observed over longer evolutionary timescales (Jensen et al. 2006; Moses, Liku, et al. 2007; Beltrao et al. 2009). This plasticity in regulatory networks has also been seen in other classes of regulatory elements, namely in transcription factor binding sites. It is interesting to note that the fraction of nonconserved phosphorylation sites found in our study (>8.8% species divergent by 5–20 My; Kellis et al. 2003) seems proportional to the fraction of nonconserved functional transcription factor–binding sites when comparing humans to mouse, where 36–40% of sites turnover (Dermitzakis and Clark 2002) during the 65–75 My (Waterson et al. 2002) separating those species. Our finding that the positions with high information content in phosphorylation sites shower fewer substitutions also mirrors the patterns of evolution seen in transcription factor–binding sites (Moses et al. 2003). Therefore, we speculate that conservation with a small rate of turnover is likely to be a general feature of many other classes of regulatory elements.

Supplementary Material

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

Acknowledgments

A.M.M. and A.N.N.B. are supported by Natural Sciences and Engineering Research Council of Canada Discovery grant. This research was supported by an infrastructure grant from the Canadian Foundation for Innovation to A.M.M. We thank Dr Nicholas Provart for providing us with a web server to store our data. We also thank Dr Philip M. Kim for comments on the manuscript.

References

Balla S, Thapar V, Verma S, et al. (14 co-authors). 2006. Minimotif Miner: a tool for investigating protein function. *Nat Methods.* 3(3):175–177.

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(6):e1000134.

Brown CJ, Takayama S, C AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol.* 55(1): 104–110.

Budovskaya YV, Stephan JS, Deminoff SJ, Herman PK. 2005. An evolutionary proteomics approach identifies substrates of the cAMP-dependent protein kinase. *Proc Natl Acad Sci U S A.* 102(39):13933–13938.

Collins MO. 2009. Evolving cell signals. *Science* 325(5948): 1635–1636.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res.* 14:1188–1190.

Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol.* 19(7):1114–1121.

Durbin R, Eddy SR, Krogh A, Mitchison G. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press.

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* 14(9):755–763.

Finn RD, Tate J, Mistry J, et al. (11 co-authors). 2008. The PFAM protein familes database. *Nucleic Acids Res.* 36 (Database):D281–D288.

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(11):R250.

Holt LJ, Tuch BB, Villén J, Johnson AD, Gygi SP, Morgan DO. 2009. Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science* 325(5948):1682–1686.

Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, Bork P. 2006. Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature* 443(7111):594–597.

Jiménez 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(5):R90.

Johnson SA, Hunter T. 2005. Kinomics: methods for deciphering the kinome. *Nat Methods.* 2(1):17–25.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423(6937):241–254.

Kobe B, Kampmann T, Forwood JK, Listwan P, Brinkworth RI. 2005. Substrate specificity of protein kinases and computational prediction of substrates. *Biochim Biophys Acta.* 1754(1–2):200–209.

Koch R, Ledermann R, Urwyler O, Heller M, Suter B. 2009. Systematic functional analysis of Bicaudal-D serine phosphorylation and intragenic suppression of a female sterile allele of BicD. *PLoS One.* 4(2):e4552.

Landry CR, Levy ED, Michnick SW. 2009. Weak functional constraints on phosphoproteomes. *Trends Genet.* 25(5):193–197.

Lienhard GE. 2008. Non-functional phosphorylations? *Trends Biochem Sci.* 33(8):351–352.

Macek B, Gnad F, Soufi B, Kumar C, Olsen JV, Mijakovic I, Mann M. 2008. Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. *Mol Cell Proteomics.* 7(2):299–307.

Malik R, Nigg EA, Körner R. 2008. Comparative conservation analysis of the human mitotic phosphoproteome. *Bioinformatics* 24(12):1426–1432.

Miller ML, Blom N. 2009. Kinase-specific prediction of protein phosphorylation sites. *Methods Mol Biol.* 527:299–310, x.

Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB. 2003. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol Biol.* 3:19.

Moses AM, Hériché JK, Durbin R. 2007. Clustering of phosphorylation site recognition motifs can be exploited to predict the targets of cyclin-dependent kinase. *Genome Biol.* 8(2):R23.

Moses AM, Liku ME, Li JJ, Durbin R. 2007. Regulatory evolution in proteins by turnover and lineage-specific changes of cyclin-dependent kinase consensus sites. *Proc Natl Acad Sci U S A.* 104(45):17713–17718.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* 3(5):418–426.

Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. New York: Oxford University Press.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol.* 302(1):205–217.

Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Sussman JL. 2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* 21(16): 3435–3438.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18(20): 6097–6100.

Schwartz D, Gygi S. 2005. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat Biotechnol.* 23(11):1391–1398.

Serber Z, Ferrell JEJ. 2007. Tuning bulk electrostatics to regulate protein function. *Cell* 128(3):441–444.

Saccharomyces Genome Database [Internet]. SGD project [cited 2009 Sept 16]. Available from ftp://ftp.yeastgenome.org/yeast/.

Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D. 1996. Dirichlet mixtures: a method for improving detection of weak but significant protein sequence homology. *Comput Applic Biosci.* 12:327–345.

Tan CS, Bodenmiller B, Pasculescu A, Jovanovic M, Hengartner MO, Jørgensen C, Bader GD, Aebersold R, Pawson T, Linding R. 2009. Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. *Sci Signal.* 2(81):ra39.

Turk BE. 2008. Understanding and exploiting substrate recognition by protein kinases. *Curr Opin Chem Biol.* 12(1):4–10.

Ubersax JA, Woodbury EL, Quang PN, Paraz M, Blethrow JD, Shah K, Shokat KM, Morgan DO. 2003. Targets of the cyclin-dependent kinase Cdk1. *Nature* 425(6960):859–864.

Uversky VN, Gillespie JR, Fink AL. 2000. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins* 41(3):415–427.

Wang X, Goshe MB, Soderblom EJ, Phinney B, Kuchar JA, Li J, Asami T, Yoshida S, Huber SC, Clouse SD. 2005. Identification and functional analysis of in vivo phosphorylation sites of the Arabidopsis BRASSINOSTEROID-INSENSITIVE1 receptor kinase. *Plant Cell.* 17(6):1685–1703.

Waterson RH, Mouse Genome Sequencing Consortium, Lander ES. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520–562.

Yachie N, Saito R, Sugahara J, Tomita M, Ishihama Y. 2009. In silico analysis of phosphoproteome data suggests a rich-get-richer process of phosphosite accumulation over evolution. *Mol Cell Proteomics.* 8(5):1061–1071.