BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 12 2012, pages 1562–1570
doi:10.1093/bioinformatics/bts195

Sequence analysis

Advance Access publication April 23, 2012

MMFPh: a maximal motif finder for phosphoproteomics datasets

Tuobin Wang [1], Arminja N. Kettenbach [2], Scott A. Gerber [2,3] and Chris Bailey-Kellogg [1,*]

[1]Department of Computer Science, Dartmouth College, Hanover, NH 03755, [2]Department of Genetics, Dartmouth Medical School and [3]Department of Biochemistry, Dartmouth Medical School, Lebanon, NH 03756, USA

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Protein phosphorylation, driven by specific recognition of substrates by kinases and phosphatases, plays central roles in a variety of important cellular processes such as signaling and enzyme activation. Mass spectrometry enables the determination of phosphorylated peptides (and thereby proteins) in scenarios ranging from targeted *in vitro* studies to *in vivo* cell lysates under particular conditions. The characterization of commonalities among identified phosphopeptides provides insights into the specificities of the kinases involved in a study. Several algorithms have been developed to uncover linear motifs representing position-specific amino acid patterns in sets of phosphopeptides. To more fully capture the available information, reduce sensitivity to both parameter choices and natural experimental variation, and develop more precise characterizations of kinase specificities, it is necessary to determine all statistically significant motifs represented in a dataset.

Results: We have developed MMFPh (*M*aximal *M*otif *F*inder for *Ph*osphoproteomics datasets), which extends the approach of the popular phosphorylation motif software Motif-X (Schwartz and Gygi, 2005) to identify all statistically significant motifs and return the maximal ones (those not subsumed by motifs with more fixed amino acids). In tests with both synthetic and experimental data, we show that MMFPh finds important motifs missed by the greedy approach of Motif-X, while also finding more motifs that are more characteristic of the dataset relative to the background proteome. Thus MMFPh is in some sense both more sensitive and more specific in characterizing the involved kinases. We also show that MMFPh compares favorably to other recent methods for finding phosphorylation motifs. Furthermore, MMFPh is less dependent on parameter choices. We support this powerful new approach with a web interface so that it may become a useful tool for studies of kinase specificity and phosphorylation site prediction.

Availability: A web server is at www.cs.dartmouth.edu/~cbk/mmfph/

Contact: cbk@cs.dartmouth.edu

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

Received on November 18, 2011; revised on April 3, 2012; accepted on April 14, 2012

1 INTRODUCTION

Protein phosphorylation plays vital roles in numerous key cellular processes including the regulation of enzyme activation and protein localization and degradation, as well as the propagation of signals through pathways that control higher-level cellular activities such as proliferation, migration, differentiation and death (Cohen, 2000; Ficarro *et al.*, 2002; Manning *et al.*, 2002; Turk, 2008). The importance of phosphorylation is underscored by the fact that 1.5–2.5% of eukaryotic genes encode kinases (Manning *et al.*, 2002).

Investigations into phosphorylation have been greatly enhanced by the advent of high-throughput mass spectrometry, which enables large-scale phosphoproteomics studies identifying peptides phosphorylated under particular conditions (Kettenbach *et al.*, 2011; Matsuoka *et al.*, 2007; Yu *et al.*, 2011). Such sets of identified phosphopeptides in turn provide information about the specificities of the kinases involved, revealing common patterns of amino acids underlying specific kinase-substrate recognition. Databases such as Phospho.ELM (Dinkel *et al.*, 2010), PhosPhAt (Heazlewood *et al.*, 2008) and PHOSIDA (Gnad *et al.*, 2007, 2011) collect sets of identified phosphopeptides for different kinases, and have enabled the development of methods to predict and classify kinase-substrate relationships, e.g. Scansite (Obenauer *et al.*, 2003), NetPhosK (Blom *et al.*, 2004), GPS (Xue *et al.*, 2005), KinasePhos (Wong *et al.*, 2007) and NetPhorest (Miller *et al.*, 2008).

In general, a phosphoproteomics study provides the opportunity to identify kinases that might be present in a sample, as well as to characterize the specificities of the kinases. Since a phosphoproteomics dataset can contain a large number of phosphopeptides produced by a mixture of different kinases, a typical first step is to aggregate and summarize the phosphopeptides in a manner that reveals more general patterns of phosphorylation. Phosphorylation motifs represent common amino acids up- and down-stream from an identified phosphorylation site. For example, if we see a number of phosphopeptides with a proline two residues before a phospho-serine and an arginine immediately following the phospho-serine, we could summarize them with the motif PxSR (where 'x' indicates that any amino acid is allowed). A position-specific scoring matrix could further refine such a motif, indicating the frequencies of amino acids at the non-fixed positions (e.g. if that 'x' tended to be an acidic residue). Matching such summaries against databases then provides evidence that particular kinases have been active, while the amino acid patterns of the motifs give additional insights into the kinase specificities.

Several motif-finding algorithms for phosphoproteomics have been developed, taking advantage of the fact that a motif is anchored by a phosphorylation site, to do better than more general motif finders (e.g. those employed to identify transcription factor motifs). The basic problem was first addressed by Motif-X (Schwartz and Gygi, 2005), which employs a greedy algorithm to incrementally build up a motif from statistically over-represented position/amino acid pairs. Motif-X was demonstrated to identify

*To whom correspondence should be addressed.

protein phosphorylation motifs including validated substrates and new motifs from HeLa cell nuclei mass spectrometry dataset and tyrosine phosphorylation immunoaffinity datesets, and to outperform a number of general-purpose motif-finding algorithms. It has been employed in subsequent studies including sumoylation site prediction (Xue *et al.*, 2006) and phosphoproteomic analysis of organisms including mouse (Villen *et al.*, 2007), *Drosophila* (Zhai *et al.*, 2008) and yeast (Wilson-Grady *et al.*, 2008). MoDL (Ritz *et al.*, 2009) adopted a fundamentally different approach, formulating the problem as one of compactly encoding the phosphorylated peptides, and employing an information-theoretic approach to find a good encoding. It was demonstrated to successfully uncover compact sets of informative motifs, both known and novel, in several published human and mouse phosphoproteomic datasets. Motif-All (He *et al.*, 2011) employed a data mining approach to uncover all motifs that have sufficient support and are statistically significant under an odds ratio assessment. The approach was demonstrated on the PhosPhAt database of *Arabidopsis* phosphorylation sites. Most recently, F-Motif (Chen *et al.*, 2011), an approach that incorporates clustering into the Motif-X-style iterative, greedy selection of motifs and reduction of foreground, was demonstrated to outperform Motif-X and MoDL on four synthetic datasets extracted from Phospho.ELM and a large-scale experimental dataset from mouse.

We present here an approach called MMFPh (*M*aximal *M*otif *F*inder for *Ph*osphoproteomics) that pursues the Motif-X goal of identifying motifs comprised of over-represented amino acid/position pairs, but does so by performing a complete search instead of making greedy choices. Thus MMFPh identifies all statistically significant, sufficiently frequent motifs, while Motif-X (and likewise F-Motif) may miss some due to greedy choices and foreground reduction. The greedy approach is justifiable in cases of a few kinases with distinct specificities, though we show that even there it can miss motifs; this problem can only get worse with larger sets of phosphopeptides or with highly overlapped motifs. Furthermore, we show that the complete approach is much more stable over the choice of the key parameter, the minimum occurrence threshold. As its name suggests, MMFPh returns only those motifs that are *maximal*, not subsumed by motifs with more fixed amino acids, as for example $\underline{\mathtt{S}}\mathtt{P}$ would be by $\mathtt{R}\mathtt{x}\underline{\mathtt{S}}\mathtt{P}$. This ensures that in addition to being more sensitive, we are also in some sense more specific. We show that the relative coverage by maximal motifs (occurrences in the phosphopeptides versus the rest of the proteome) can be better than that of the non-maximal ones. However, recognizing that a more general motif (e.g. $\underline{\mathtt{S}}\mathtt{P}$) might capture some of the identified phosphopeptides not matching the more specific motifs (e.g. $\mathtt{R}\mathtt{x}\underline{\mathtt{S}}\mathtt{P}$), we also reassess for the possibility of a *residual* motif among the unmatched phosphopeptides (e.g. those with an $\underline{\mathtt{S}}\mathtt{P}$ but not the up-stream R).

To summarize our contribution, MMFPh is a complete approach, identifying all maximal, statistically significant and sufficiently frequent motifs. Its completeness and maximality stand in contrast to Motif-X and F-Motif, and result in better specificity and sensitivity. MoDL pursues a different goal, but to some extent (as demonstrated by (Chen *et al.*, 2011)) these contrasts carry over, as MoDL can miss important motifs, as well as make specificity and sensitivity trade-offs. While Motif-All also seeks completeness, it employs a different significance assessment from that of Motif-X, rendering it harder to directly assess the importance of completeness. It also does not restrict to maximal motifs and does not identify residual motifs. The conference publication of Motif-All (He *et al.*, 2011) did not rigorously demonstrate its utility. In contrast, we use both synthetic datasets and large-scale experimental datasets to thoroughly substantiate the importance of completeness (including biologically relevant specificities missed by greedy methods), characterize relative coverage of data versus background, and assess stability over parameter choices. In some cases, the differences between MMFPh and Motif-All do not matter much if at all. However, in others, they apparently lead to Motif-All finding many more motifs than MMFPh, and many more than are biologically supported. Furthermore, Motif-All can require substantially more time to find the motifs. Finally, in contrast to both MoDL and Motif-All, we provide a convenient web server so that the wider community may easily find, characterize and visualize motifs present in phosphoproteomics datasets.

2 METHODS

MMFPh takes as input a *foreground* dataset of phosphorylated peptides, along with a corresponding *background* set of phosphorylatable peptides. The goal is to find motifs capturing patterns of amino acids that are over-represented in the foreground relative to the background.

We represent a motif in terms of a set of *fixed* amino acids at nearby positions up- and down-stream from a phosphorylation site. For example, the motif $m=\{(-2,\mathtt{P}),(0,\mathtt{S}),(+1,\mathtt{R})\}$ has a phosphorylated S, with a fixed P two positions up-stream and a fixed R one position down-stream. We index a motif with a position to obtain the fixed amino acid type there; e.g. $m_{-2}=\mathtt{P}$. If a position is not specified in the motif, then any amino acid is acceptable. We write this as x (rather than the more formal $\perp$); e.g. $m_{-1}=\mathtt{x}$. For simplicity, we often write a motif as a string with the phosphorylation site underlined and x in each unspecified position; our example is thus $\mathtt{P}\mathtt{x}\underline{\mathtt{S}}\mathtt{R}$.

We assess three aspects of motif over-representation: frequency, statistical significance and maximality. The first two are as in Motif-X (Schwartz and Gygi, 2005): the motif must appear a sufficient number of times in the foreground, and its fixed amino acids must be surprisingly abundant in the foreground according to a binomial model based on the background. In addition, we require a motif to be *maximal*, in that no other motif is extended from it. For example, if we identified $\mathtt{P}\mathtt{x}\underline{\mathtt{S}}\mathtt{R}$, we would not also identify $\mathtt{P}\mathtt{x}\underline{\mathtt{S}}$ or $\underline{\mathtt{S}}\mathtt{R}$. The two more general motifs might have attained their over-representation from the more specific one, and would also appear in many more background peptides (potential false positives). We would only want to return $\mathtt{P}\mathtt{x}\underline{\mathtt{S}}$ along with $\mathtt{P}\mathtt{x}\underline{\mathtt{S}}\mathtt{R}$ if the former were over-represented when we excluded the peptides of the latter; i.e. if $\mathtt{P}\mathtt{x}\underline{\mathtt{S}}[\neg\mathtt{R}]$ were over-represented ('$\neg$' is logical not, here indicating that any residue other than R is allowed at $+1$). Thus after finding the initial maximal motifs (here, $\mathtt{P}\mathtt{x}\underline{\mathtt{S}}\mathtt{R}$), we reassess the non-maximal motifs to see which can yield maximal *residual* motifs (here, $\mathtt{P}\mathtt{x}\underline{\mathtt{S}}[\neg\mathtt{R}]$ and $[\neg\mathtt{P}]\mathtt{x}\underline{\mathtt{S}}\mathtt{R}$).

Figure 1 summarizes the MMFPh approach, and the following subsections detail the main steps.

2.1 Preprocessing

The foreground is a set of experimentally determined phosphopeptides, each with an indicated phosphorylation site. MMFPh rebuilds and/or truncates each peptide to a specified length of $2d+1$, with d residues (defaulting to 6) up-stream and d more down-stream from the phosphorylation site. Alternatively, a preprocessed set of peptides of this format can be directly provided. MMFPh rebuilds with respect to a specified proteome, by searching for each peptide in a list of proteins. In the case of ambiguity (i.e. two or more proteins contain the same peptide), it simply uses the first. If a reconstructed peptide is duplicated, only one copy is kept.

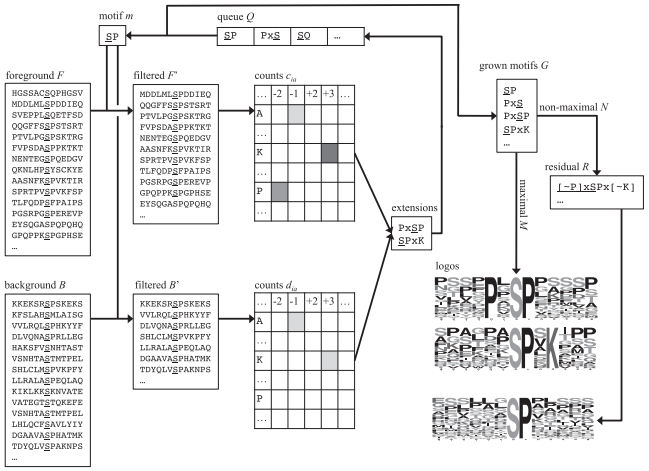

Fig. 1. Maximal motif finder for phosphoproteomics datasets (MMFPh). The foreground and background peptides have been preprocessed to peptides of uniform length, in a fixed window around the (potential) phosphorylation sites. An iterative cycle grows motifs by testing each possible extension by fixing one more amino acid in a motif discovered in a previous cycle. In the example, we are attempting to extend the motif $\underline{\mathtt{S}}\mathtt{P}$. We see among the foreground peptides containing this motif some A at -1 (indicated by the yellow cell), even more P at -2 (orange cell) and still more K at $+3$ (red cell). The extensions $(-2,\mathtt{P})$ and $(+3,\mathtt{K})$ are over-represented relative to the expected number predicted by a binomial distribution based on the filtered background, but $(-1,\mathtt{A})$ is not. In general, all frequent, statistically significant extensions are added to the queue. We return all maximal motifs (those for which there is no extension), and reassess the non-maximal to find maximal residual motifs (those that are frequent and statistically significant among the peptides that do not contain their extensions)

The background is a set of peptides that are potentially phosphorylatable, each again of length $2d+1$ and centered at a serine, threonine or tyrosine. Typically the background set is constructed by scanning a specified proteome for all such peptides. Alternatively, an empirical background can be provided, e.g. a set of peptides collected from other mass spectrometry runs. Empirical background peptides are reconstructed as described for the foreground. For both proteome-based and empirical backgrounds, only a single unique copy of each peptide is kept.

2.2 Motif growing

Figure 1 summarizes the flow of our algorithm; detailed pseudocode is provided in the Supplementary Material (Algorithm 1). We maintain a queue Q of motifs to grow, initially just the center (phosphorylated) serine, threonine or tyrosine. At each iteration, we dequeue a motif m and filter the foreground and background to those peptides containing it (i.e. matching each of its fixed positions), giving F' and B'. We then consider each possible extension of m to include an amino acid a at a non-fixed position i, and check whether the extension is sufficiently frequent and statistically significant.

Frequency: the number c_{ia} of occurrences of amino acid a at position i in the filtered foreground F' must be sufficiently large, at least a user-specified threshold θ_{occ}.

Statistical significance: the probability of observing c_{ia} occurrences of a at i in F' must be sufficiently small, at most a user-specified threshold θ_{sig}, according to a probability model derived from the distribution of amino acids at i in the filtered background B'. We employ the same binomial probability model as Motif-X (Schwartz and Gygi, 2005); we write it here as:

$$P(x;n,p)=\sum_{k=x}^{n}\binom{n}{k}p^k(1-p)^{n-k} \tag{1}$$

where x is the number of occurrences out of n (here c_{ia} out of $|F'|$), with a background probability of p (here the background count d_{ia} out of $|B'|$). We note that the threshold θ_{sig} should be set sufficiently low to account for multiple hypothesis testing.

If these tests are satisfied, the extended motif is added to the queue unless it has already been there. Note that there are multiple ways to grow to a motif of more than one fixed position (e.g. $\underline{\mathtt{S}} \to \mathtt{Px}\underline{\mathtt{S}} \to \mathtt{Px}\underline{\mathtt{S}}\mathtt{R}$ or $\underline{\mathtt{S}} \to \underline{\mathtt{S}}\mathtt{R} \to \mathtt{Px}\underline{\mathtt{S}}\mathtt{R}$). If none of the possible extensions is added to the queue, the original motif is maximal; otherwise, it is non-maximal and will be reassessed (below) for a possible residual motif upon excluding its extensions.

In summary, MMFPh explicitly considers all extensions at each iteration, and evaluates them with respect to all the foreground and background peptides that contain the motif. Thus, unlike the greedy approach of Motif-X, it is complete and guaranteed to find all significant maximal motifs.

2.3 Residual motif identification

As discussed above, we only want to return maximal motifs, as they are more specific than the motifs from which they are extended. But the more general non-maximal motifs may also appear in some foreground peptides that do not contain the more specific extensions. To account for that possibility, we reassess the non-maximal motifs to see if, after excluding the peptides containing their extensions, they are frequent and statistically significant (the same criteria as in the growing algorithm). Detailed pseudocode is provided in the Supplementary Material (Algorithm 2).

To reassess a non-maximal motif, we first filter the foreground and background to contain only those residues not containing its (maximal) extensions. We then essentially re-do the growing of the motif, ensuring that each extension used to build it remains sufficiently frequent and statistically significant among the remaining peptides. We look at each such extension (position and amino acid) separately; if any is valid, then the motif is residual. The test is the same as in the growing algorithm—filter the foreground and background and then count and test the number of occurrences of the extension.

We store maximal residual motifs in a separate set from maximal non-residual motifs. In the results, we either write them using the 'not' notation illustrated above, or rely on the context of the extensions to indicate which fixed amino acids are eliminated.

2.4 Scoring

In sorting through the identified motifs, it is helpful to have an overall score assessing each. While Motif-X scores a motif by summing the negative logarithm of the binomial probability of each of its fixed positions, that is not appropriate here because, as we have discussed, there may be multiple ways to grow a motif and thus multiple different such scores. Instead, we use a score that is similar to log-odds (Krogh, 1998), assessing the degree of over-representation of the motif in the foreground relative to the background. Rather than estimating position-specific probabilities, we simply use the numbers of occurrences of the entire motif in the foreground and background:

$$\text{Score}(m)=\log\frac{F'/F_{\text{c}}}{B'/B_{\text{c}}} \tag{2}$$

where, as above, F' and B' are the filtered foreground and background, respectively, according to the motif m. Since we separately consider S, T and Y motifs, we filter the initial foreground according to the center residue (F_c and B_c).

2.5 Clustering

We use hierarchical clustering to organize the motif results and help show 'meta-motif' patterns. Each pair of motifs is globally aligned by the algorithm of (Needleman and Wunsch, 1970), forcing phosphorylation sites to align to each other. We employ a scoring model that includes an independent gap penalty (default -8) and standard substitution matrix (e.g. Blosum-62) extended with scores for the 'x' positions (default: x to x scores 0, while x

to a fixed amino acid scores −4). We then employ average-linkage clustering based on these pairwise alignments.

2.6 Implementation and web server

The MMFPh motif-identification algorithm is implemented in platform-independent Java SE 1.6. Post-processing scripts generate logos (Schneider and Stephens, 1990) with WebLogo (Crooks *et al.*, 2004) and visualize the hierarchical clusters with Jalview (Waterhouse *et al.*, 2009). We have implemented a web interface, available at www.cs.dartmouth.edu/~cbk/mmfph/, via which users can specify a foreground dataset and set the parameters controlling the motif identification and post-processing.

3 RESULTS

To demonstrate MMFPh's utility in uncovering motifs, we present a series of tests with both synthetic and experimental data.

3.1 Synthetic datasets

We first applied MMFPh to synthetic datasets so that we know a form of 'ground truth'—the specific kinases giving rise to the phosphorylated peptides. To this end, we use the benchmark synthetic datasets from the Motif-X and F-Motif papers, augmenting the results provided there with results for MMFPh and the other new method Motif-All. We compare and contrast the results from the different methods below, and provide lists of identified motifs in Supplementary Material 1 (Spreadsheet 1). As we discussed, MoDL (Ritz *et al.*, 2009) pursues a somewhat different problem and thus would require a different metric for comparison, so we do not benchmark against it here. For consistency, we use the backgrounds specified in the F-Motif paper. We use the default 10^{-6} as the significance threshold, 20 as the occurrence threshold for MMFPh, Motif-X and Motif-All, and $G = 15$, $T = 15$ for F-Motif.

3.1.1 Five designed motifs Five motifs (RxSxxP, RxSxxI, KSxxxI, TVxSxE and DxxSQxN) were planted in a foreground in a manner such that other discovered motifs can be treated as false positives (Schwartz and Gygi, 2005). As Table 1 summarizes, MMFPh outperforms the other methods, finding all the true positives and only six false positives. F-Motif missed a number of the true positives and found a large number of false positives. Motif-All found even more false positives, 11 of which are extensions of the planted motifs but 87 of which are unrelated to the planted motifs. In addition, Motif-All took 27 min to discover these motifs, whereas MMFPh required only 15 sec.

3.1.2 Phospho.ELM single-kinase datasets Separate foreground datasets for four different kinases (PKA, PKC, CK2 and CDK) were generated from Phospho.ELM, using either the substrates reported for all species or restricting them to human proteins (Chen *et al.*, 2011). Following our standard protocol, we removed duplicates in these datasets before performing motif analysis. Corresponding all-species or human-only backgrounds were used.

Table 1. Recovery of five planted motifs

Method	True positives	False positives
MMFPh	5	6
Motif-X	5	7
Motif-All	5	98
F-Motif	2	26

Table 2 enumerates the numbers of different motifs discovered by the different methods. In almost every dataset, MMFPh found additional novel or more specific motifs than Motif-X and F-Motif. For example, for PKA, MMFPh found RxRxxS, which was missed by Motif-X and F-Motif due to the greedy approach and foreground reduction, eliminating peptides covered by the motif RRxS which happened to have been found before RxRxxS. Similarly, for CDK, MMFPh found PxSP, which Motif-X missed by finding SPxK beforehand. (In both cases, MMFPh found both motifs.) For PKC, MMFPh found two previously reported motifs missed by the others: SxxR and SF (Nishikawa *et al.*, 1997), and for CK2, it found more specific extensions of SxxE and several additional novel motifs with down-stream D and E. Overall, MMFPh missed eight motifs found by Motif-X or F-Motif, four of which were replaced by more specific extensions and the rest of which were deemed statistically insignificant (and found by the other methods due to the order-dependent foreground reduction). Motif-All found all the motifs from the other methods, but in the all-species dataset also found a very large number of extra motifs (as we observed with the designed motifs). Most of these do not appear to be particularly informative; e.g. whereas other methods found at most 3 motifs for CDK, Motif-All found 29, and for CK2, almost half of their 33 motifs are with a single up- or down-stream D or E, deemed insignificant under our scoring.

3.1.3 Phospho.ELM mixture A foreground was constructed (Schwartz and Gygi, 2005) from 43 ATM substrates, 184 Casein II substrates, 41 CaMK II substrates and 30 MAPK substrates, as deposited in Phospho.ELM. With either a human-only or all-species background, Motif-X was reported to find six motifs and F-Motif seven motifs.

With either background, MMFPh found 11 motifs, including 5 more specific extensions of the motif SxxE, a potential Casein II motif with the critical E at position 3 and additional acid residues down-stream (Kuenzel *et al.*, 1987). MMFPh did not find the motifs

Table 2. Differences among motifs found by different methods on the single-kinase datasets

All species

	MMFPh	Motif-X	F-Motif	Motif-All
MMFPh		1/6/14/1	1/6/13/1	0/0/0/0
Motif-X	0/0/1/0		0/0/0/0	0/0/0/0
F-Motif	0/1/1/3	0/1/1/3		0/0/0/0
Motif-All	26/4/16/26	27/10/29/27	27/9/28/24	

Human only

	MMFPh	Motif-X	F-Motif	Motif-All
MMFPh		1/2/10/1	1/2/10/0	0/0/0/0
Motif-X	1/0/0/0		0/0/0/0	0/0/0/0
F-Motif	1/0/0/0	0/0/0/1		0/0/0/0
Motif-All	5/1/4/6	5/3/14/7	5/3/14/16	

Each cell lists the numbers of motifs found by the method of its row but not by the method of its column, in the order PKA/PKC/CK2/CDK.

SP (MAPK) and RxxS (CamK II). This is due to the imbalance in the number of occurrences of different kinase-specific peptides in this artificially constructed dataset—it does not include consistently representative levels of the diversity of peptides for each, leading to statistical insignificance of the less-represented motifs within the entire foreground. We note that the greedy reduction step enabled recovery of these under-represented motifs only because their specifities are distinct. Motif-All found 22 motifs, in either background, and missed SP. Most of its additional motifs, which were insignificant under the binomial model, simply placed a single D or E at the various up- and down-stream positions.

3.2 Experimental phosphoproteomics datasets

We have applied MMFPh to a variety of phosphoproteomics datasets. We summarize here results on three previously studied datasets from other labs, and then do a more detailed case study on one of our own datasets. Detailed motif lists are provided in Supplementary Materials 2 (Spreadsheet 2 for data from other labs; Spreadsheet 3 for our case study).

We compare against the well-established Motif-X method (Schwartz and Gygi, 2005), to enable evaluation of the importance of complete versus greedy searches (the key difference between the two approaches). To ensure that we control for implementation details other than the search method, we implemented an option in MMFPh to enable it to perform the Motif-X greedy search; we call this version GrMFPh (*Gr*eedy *M*otif *F*inder for *Ph*osphoproteomics datasets). We have used GrMFPh with a variety of different datasets and found only minor differences between its results and those of Motif-X, presumably due to different background proteomes and some small undocumented implementation details. Here we use a background based on the Human IPI database (http://www.ebi.ac.uk/IPI/IPIhuman.html).

3.2.1 Distinct phosphorylation sites of CDK1 substrates We tested MMFPh and Motif-X on the dataset from (Holt *et al.*, 2009), which had 547 distinct phosphorylation sites (15mers). We set θ_{occ}=20 and θ_{sig}=10^{-6}. Table 3 (right) lists the motifs found by the two methods: GrMFPh found eight motifs and MMFPh found all those plus five additional ones extending the fairly general SP to be more specific (still deeming the residual motif to be significant). The minimal consensus motifs reported in the paper, SP, TP, SPxK and TPxK, were found, but an additional preference for R down-stream was not found.

3.2.2 DNA damage-regulated phospho-SQ and TQ sites We rebuilt the 905 DNA damage-regulated phospho-SQ and TQ sites from (Matsuoka *et al.*, 2007) to 13mers and applied both GrMFPh and MMFPh at θ_{occ}=20 and θ_{sig}= 10^{-6}. Table 3 (right) lists the identified motif. GrMFPh found eight motifs. Of these, MMFPh returned three as maximal significant, identified more specific extensions for another three while deeming the general motif to be significant as a residual motif, and replaced the final two with more specific extensions (deeming the more general motif not to be significant as a residual motif). In particular, MMFPh motifs provide more specific characterizations of up- and down-stream D and E (e.g. TQE in addition to TQ), along with enrichment of S around SQ sites (e.g. SxSQ and SQGxS) as described in the original publication.

3.2.3 TCR-responsive phosphorylation sites We compared GrMFPh and MMFPh on a compilation of TCR-responsive phosphorylation sites from (Mayya *et al.*, 2009). In that paper, the foreground and background were examined for occurrences of putative motifs for seven kinases. In contrast, Motif-X and MMFPh work in the 'opposite' direction, extracting significant motifs from the given dataset. After preprocessing the dataset of 21mers, we obtained 5297 S-centered, 1028 T-centered and 164 Y-centered unique phosphorylation sites. To match the fairly general motifs from the paper, we increased the occurrence threshold to ∼6% of the foreground size, to θ_{occ}=300 for S-centered and θ_{occ}=60 for T-centered peptides. No Y-centered motifs were found, even with an occurrence threshold of 10.

Table 3. GrMFPh and MMFPh motifs for two different experimental datasets

(Holt *et al.*, 2009)		(Matsuoka *et al.*, 2007)	
GrMFPh	MMFPh	GrMFPh	MMFPh
SPxK	SPxK	GSQ	GSQ
SPxxN	SPxxN	ExxxSQ	
SPI	SPI		EExxxSQ
NxxSP	NxxSP	SQG	
NxxxxxSP	NxxxxxSP		ExxxSQG
SP	SP		SQGS
	NxSP	SQD	SQD
	NxxxxxxSP	SQE	SQE
	SPxxxxN		GSQE
	SPxxxK	SQxE	SQxE
	SxxxSP	SQ	SQ
TP	TP		SQxSQ
TPxK	TPxK		SQxxxE
			ExxxxxSQ
		TQ	TQ
			TQE

GrMFPh found 11 motifs (6 S-centered and 5 T-centered), whereas MMFPh found 72 motifs (55 S-centered and 17 T-centered), including 10 of the GrMFPh ones (deeming the final motif found by GrMFPh to be insignificant, perhaps an artifact of the greedy reduction step). The GrMFPh motifs cover 10 of the putative motifs explored in the paper, whereas the MMFPh motifs cover 12. All other putative motifs are either statistically insignificant or contain few instances in the dataset. The additional MMFPh motifs include more specific extensions with up- and down-stream D and E. They also include several motifs that are potentially representative of other kinases; e.g. DxxS of CK1 (Pulgar *et al.*, 1999), SPxxS of GSK-3β (Fiol *et al.*, 1987; Hardt and Sadoshima, 2002) and RxxxS of (Arora *et al.*, 2010). Moreover, MMFPh's results cover all the S-centered dataset and 75% of T-centered dataset, whereas Motif-X covers only 78% and 65%, respectively.

3.2.4 NCI-H23 non-small cell lung cancer cells dataset Finally, we perform a case study analysis of a set of 14 769 phosphopeptides from NCI-H23 non-small cell lung cancer cells (Kettenbach and Gerber, 2011). After preprocessing, this represents 9817 unique 13mers. We used the default significance threshold of 10^{-6}. We first present some motifs discovered at an occurrence threshold of 100 (roughly 1% of the dataset), and then characterize trends over a range of thresholds.

Discovered motifs at an occurrence threshold of 100 One of the reasons to find motifs is to identify which kinases may be active

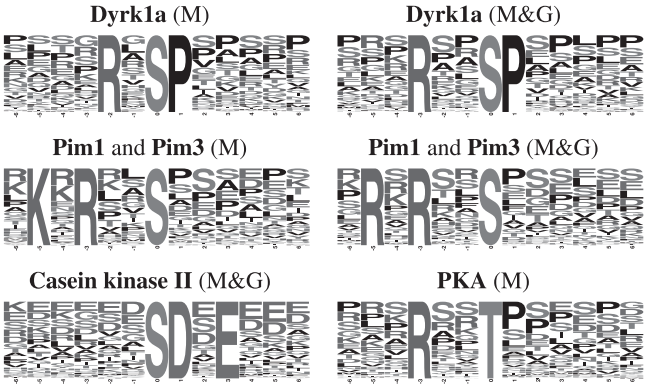

Fig. 2. Logos for example motifs that were found by MMFPh (marked 'M') and sometimes GrMFPh ('G') in the NCI-H23 dataset and that correspond to published motifs for particular kinases

Fig. 3. Logos for two residual motifs identified by MMFPh

within the conditions underlying a dataset. MMFPh finds a number of motifs that are supported by previous literature as capturing the specificities of particular kinases; some are illustrated in Figure 2. For example, the substrate of Dyrk1a is believed to include an R at positions −2 or −3 and a P at position +1 (Campbell and Proud, 2002; Himpel *et al.*, 2000). Pim1 and Pim3 both prefer to phosphorylate S and T with R or K at positions −3 and −5 (Hutti *et al.*, 2004; Miller *et al.*, 2008). Casein kinase II has been found to phosphorylate peptides with the motif SDxE (Tuazon and Traugh, 1991). An RxxT motif consistent with protein kinase A (PKA) substrates (Shabb, 2001) was also found. We note that GrMFPh misses three of these seven motifs.

In addition to growing maximal motifs, MMFPh reassesses motifs that have been extended, to see if the residual motifs (explaining the peptides not containing its extensions) are maximal. Most such motifs (e.g. SP, RxxS and SxE) are deemed insignificant with respect to their more specific extensions. However, one interesting residual motif (Figure 3, left) is the MAPK1 motif PxSP, also extended to SxPxSP. This motif is highly overlapped with the two Dyrk1a motifs, which indicates that MMFPh has the ability to identify similar motifs. Another interesting residual motif (Figure 3, right) is SxxxSP, which was extended to SPxxSP. (Note that GrMFPh does not find either of these.) In addition to the P, the +1 position also has a (smaller) preference for R and L, rendering the residual motif significant. These motifs may be capturing substrates of GSK-3β, which phosphorylates such peptides in various transcription factors, such as c-Jun, c-myc, L-myc and the microtubule-associated protein tau (Chen *et al.*, 1999; Hanger *et al.*, 1992; Litersky *et al.*, 1996; Yang *et al.*, 1993).

Aggregate characterization The complete set of phosphopeptides matching a motif, as well as newly discovered extensions of a previously described motif, may help refine our understanding of the specificity of various kinases. The richness of the MMFPh motifs relative to the GrMFPh motifs supports this endeavor. MMFPh found a total of 225 motifs, listed in Supplementary Materials 3 (Spreadsheet 3). Figure 4 summarizes these motifs within a hierarchical clustering (Blosum-62, x:x=0, x:fixed=−4, gap=−8), providing logos for some of the sub-trees. For example, our Pim1/Pim3 and PKA motifs (Figure 2) are in the second cluster, along with other motifs that have up-stream R/K preferences and no down-stream ones. Our Dyrk1a motif is in the following cluster, along with others that prefer P at +1 and have some up-stream preferences. The final two clusters concentrate on down-stream specificities, such as with our casein kinase II motif.

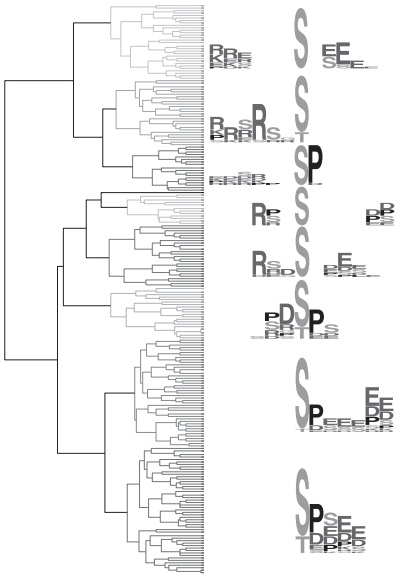

Fig. 4. Hierarchical clustering of MMFPh motifs from the NCI-H23 dataset

We compared the motifs identified by MMFPh with those identified by GrMFPh. Table 4 summarizes the number of motifs found by each algorithm (and of those, how many the other algorithm also found), categorized by number of fixed positions in addition to the phosphorylation site. The vast majority of the 225 MMFPh motifs at θ_{occ}=100 have 2 fixed positions; 18 have 1 (all T-centered); and 3 have 3. GrMFPh finds a total of 37 motifs, roughly equally split between one and two fixed positions. GrMFPh finds only 3 of MMFPh's one-position motifs, 21 of its two-position motifs, and none of its three-position motifs. Whereas MMFPh only returns a few of GrMFPh's one-position motifs (all T-centered), that is because it extended all of the one-position S-centered motifs to more specific two-position motifs, and found only five residual motifs (TxP, PxSP, SPxxxP, SPxxS and SxxxSP) to be significant under reassessment. The single two-position GrMFPh motif missed by MMFPh is SPxxR. Upon further investigation, this appears to be an artifact of extending SP with the greedy method. While only 238 of the 3225 SP peptides contain the R (failing the significance test), the greedy reduction step has eliminated many of the peptides that match motifs discovered during previous iterations, leaving 112 of 939 SP peptides that contain the R, which is significant.

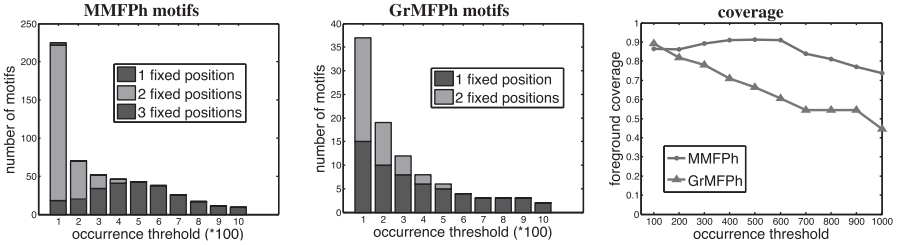

Fig. 5. Effects of different occurrence thresholds on (left, middle) number of motifs and (right) foreground coverage

Table 4. Summary of MMFPh and GrMFPh results for the NCI-H23 dataset at two occurrence thresholds

	θ_{occ}=100			θ_{occ}=200		
	1	2	3	1	2	3
MMFPh	18	204	3	20	50	0
and GrMFPh	3	21	0	4	9	0
GrMFPh	15	22	0	10	9	0
and MMFPh	3	21	0	4	9	0

Each cell indicates a number of motifs. Columns specify numbers of fixed amino acids. Rows specify MMFPh motifs, MMFPh motifs also found by GrMFPh, GrMFPh motifs and GrMFPh motifs also found by MMFPh.

In general, motif identification seeks to compactly represent the common amino acid patterns in the foreground dataset. We say that a phosphopeptide is *covered* if it matches the pattern of at least one motif; the number of covered phosphopeptides then serves as a measure of sensitivity. For the NCI-H23 dataset, 87% of the phosphopeptides are covered by MMFPh motifs, while slightly more, 90%, are covered by GrMFPh motifs. This is because MMFPh finds more specific motifs; as we discussed above, it extends 12 of the GrMFPh one-position motifs to two-position motifs and deems the residual motifs not to be maximal upon reassessment. Supplementary Table 1 provides coverage statistics for all the non-maximal GrMFPh motifs versus their MMFPh extensions. In most cases, the extensions cover roughly the same set of foreground peptides. For example, the non-maximal GrMFPh motif SxxxE has 17 MMFPh extensions; these include 3 each at positions +1 and +2, 2 each at positions +5 and +6, and one extension at each of various other positions. A total of 936 of the 963 original phosphopeptides match at least one of these 17 extensions.

We can likewise characterize the coverage of background peptides. As might be expected, in general, the extended MMFPh motifs cover fewer background peptides than the non-maximal motifs from which they are extended (Supplementary Table 1). In cases where MMFPh extensions cover less foreground, such as KxxS which is extended only to KxxSP, the marked improvement in reduced background coverage might be worth it, here giving a roughly five-fold improvement in the ratio between the size of the covered foreground and that of the covered background.

Stability analysis Both MMFPh and GrMFPh require setting a minimum occurrence threshold θ_{occ} for the number of phospopeptides covered by a motif. For MMFPh, this is with respect to all foreground peptides; with GrMFPh, for those remaining uncovered by previously identified motifs. Consequently, the choice of this parameter has relatively little effect on MMFPh compared with GrMFPh. We ran both MMFPh and GrMFPh with occurrence thresholds ranging from 100 (≈1% of the foreground) to 1000 (≈10%). Figure 5 summarizes the trends. We see that requiring more occurrences shifts MMFPh from motifs that have more fixed positions to those that have fewer fixed positions, but that cover roughly the same foreground phosphopeptides. The coverage is extremely stable over most of the range, only decreasing much with very large thresholds (as would be expected). GrMFPh, on the other hand, shows a strong dependence on the exact setting of the parameter. It find mostly motifs with a single fixed position, and simply finds fewer and fewer of them with an increased threshold, and thereby rapidly degrades in foreground coverage.

In practice, it may be hard to decide *a priori* upon a proper value for θ_{occ}. An advantage of the MMFPh approach is that identified maximal motifs are independent in terms of their coverage. Thus maximal motifs identified at one occurrence level persist at any higher level, up to their actual number of occurrences. (Residual motifs are, however, dependent, and must be reassessed.) Thus our implementation supports a scan over a range of thresholds, presenting trends as in Figure 5, and allowing the user to drill down and examine motifs at a level that strikes a desired balance in the trends (number of motifs, fixed positions).

In this dataset, the 200-occurrence level makes a good contrast with the 100-occurrence level that we have characterized so far. As detailed in Table 4, most of the two-position motifs do not meet the higher threshold and are replaced by one-position motifs, all of the three-position motifs are dropped in favor of two-position motifs, and some of the one-position are dropped. Some new motifs are introduced; e.g. SxD had 9 extensions at θ_{occ} = 100 covering 652 of its 772 instances and thus did not pass reassessment, but it had 0 extensions at 200. Whereas the 200-occurrence motifs are less specific than the 100-occurrence ones, they still cover the foreground equally well (Figure 5, right). Finally, we again find (Supplementary Table S2) that the maximal MMFPh motifs cover most of the foreground covered by the non-maximal GrMFPh motifs that they extend, and often much less of the background. In particular, whereas non-maximal SP has a large number of extension with about the same coverage, most of the other non-maximal motifs have only a few extensions, and sacrifice a bit of foreground coverage for a 2- to 4-fold improvement in the ratio of foreground to background coverage.

We also studied the stability of the motif-finding results under variation in the statistical significance level or the background

proteome. The results are as would be expected and easily summarized, thus provided in the Supplementary Material (Fig. S1). The tighter the significance threshold, the lower the foreground coverage. GrMFPh is relatively more stable to θ_{sig} as it tends to find motifs with just one fixed position at a very tight threshold. Subsetting the background even to 25% of the original proteome has little effect on the motifs found and their coverage (Supplementary Fig. S2).

4 CONCLUSION

We have developed and demonstrated a new method, MMFPh, that identifies all significant maximal motifs in a phosphoproteomics dataset, to summarize the data and help identify involved kinases and characterize their specificities. Tests with both synthetic and experimental datasets demonstrate the importance of employing a complete search rather than a greedy search, obtaining better specificity, sensitivity and stability to parameter choices. To enable the wider community to take advantage of this approach, we provide a web server that finds motifs, aggregates and summarizes the supporting peptides, and presents logos and hierarchical clusters to aid analysis.

ACKNOWLEDGEMENTS

The authors would like to thank Jeffrey Milloy and Jason Gilmore for helpful discussions.

Funding: NSF grant IIS-0905206 (in part) and NIH grant P20-RR018787 (in part) for the IDeA Program of the National Center for Research Resources.

Conflict of Interest: none declared.

REFERENCES

Arora,G. *et al.* (2010) Understanding the role of PknJ in mycobacterium tuberculosis: biochemical characterization and identification of novel substrate Pyruvate Kinase A. *PLoS ONE*, **5**, e10772.

Blom,N. *et al.* (2004) Prediction of post-translational glycosylation and phosphrylation of proteins from the amino acid sequence. *Proteomics*, **4**, 1633–1649.

Campbell,L.E. and Proud,C.G. (2002) Differing substrate specificities of members of the DYRK family of arginine-directed protein kinases. *FEBS Lett.*, **510**, 31–36.

Chen,G. *et al.* (1999) The mood-stabilizing agent valproate inhibits the activity of glycogen synthase kinase-3. *J. Neurochem.*, **72**, 1327–1330.

Chen,Y. *et al.* (2011) Discovery of protein phosphorylation motifs through exploratory data analysis. *PLoS ONE*, **6**, e2002.

Cohen,P. (2000) The regulation of protein function by multisite phosphorylation—a 25 year update. *Trends Biochem. Sci.*, **25**, 596–601.

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

Dinkel,H. *et al.* (2010) Phospho.ELM: a database of phosphorylation sites - update 2011. *Nucleic Acids Res.*, **36**, 240–244.

Ficarro,S. *et al.* (2002) Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nat. Biotechnol.*, **20**, 301–305.

Fiol,C.J. *et al.* (1987) Formation of protein kinase recognition sites by covalent modipcation of the substrate. Molecular mechanism for the synergistic action of casein kinase II and glycogen synthase kinase 3. *J. Biol. Chem.*, **262**, 14042–14048.

Gnad,F. *et al.* (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.*, **8**, R250.

Gnad,F. *et al.* (2011) Phosida 2011: the posttranslational modification database. *Nucleic Acids Res.*, **39**, 253–260.

Hanger,D. *et al.* (1992) Glycogen synthase kinase-3 induces Alzheimer's disease-like phosphorylation of tau: generation of paired helical filament epitopes and neuronal localisation of the kinase. *Neurosci. Lett.*, **147**, 58–62.

Hardt,S. and Sadoshima,J. (2002) Glycogen synthase kinase-3?: a novel regulator of cardiac hypertrophy and development. *Circ. Res.*, **90**, 1055–1063.

Heazlewood,J. *et al.* (2008) Phosphat: a database of phosphorylation sites in arabidopsis thaliana and a plant-specific phosphorylation site predictor. *Nucleic Acids Res.*, **36**, 1015–1021.

He,Z. *et al.* (2011) Motif-All: discovering all phosphorylation motifs. *BMC Bioinform.*, **12** (Suppl. 1), S22.

Himpel,S. *et al.* (2000) Specificity determinants of substrate recognition by the protein kinase DYRK1A. *J. Biol. Chem. Meth.*, **275**, 2431–2438.

Holt,L. *et al.* (2009) Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science*, **325**, 1682–1686.

Hutti,J. *et al.* (2004) A rapid method for determining protein kinase phosphorylation specificity. *Nat. Meth.*, **1**, 27–29.

Kettenbach,A. and Gerber,S. (2011) Rapid and reproducible single-stage phosphopeptide enrichment of complex peptide mixtures: application to general and phosphotyrosine-specific phosphoproteomics experiments. *Anal. Chem.* **83**, 7635–7644. http://www.ncbi.nlm.nih.gov/pubmed/21899308.

Kettenbach,A. *et al.* (2011) Quantitative phosphoproteomics identifies substrates and functional modules of Aurora and Polo-like kinase activities in mitotic cells. *Sci. Signal*, **4**, rs5.

Krogh,A. (1998) *Computational Methods in Molecular Biology*. Elsevier Science, Denmark.

Kuenzel,E. *et al.* (1987) Substrate specificity determinants for casein kinase I1 as deduced from studies with synthetic peptides. *J. Biol. Chem.*, **262**, 9136–9140.

Litersky,J. *et al.* (1996) Tau protein is phosphorylated by cyclic AMP-dependent protein kinase and calcium/calmodulin-dependent protein kinase II within its microtubulebinding domains at Ser-262 and Ser-35. *J. Biochem.*, **316**, 655–660.

Manning,G. *et al.* (2002) Evolution of protein kinase signaling from yeast to man. *Trends Biochem. Sci.*, **27**, 514–520.

Matsuoka,S. *et al.* (2007) ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science*, **316**, 1160–1166.

Mayya,V. *et al.* (2009) Quantitative phosphoproteomic analysis of T cell receptor signaling reveals system-wide modulation of protein-protein interactions. *Sci. Signal*, **2**, ra46.

Miller,M.L. *et al.* (2008) Linear motif atlas for phosphorylation-dependent signaling. *Sci. Signal*, **1**, ra2.

Needleman,S. and Wunsch,C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Nishikawa,K. *et al.* (1997) Determination of the specific substrate sequence motifs of protein kinase C isozymes. *J. Biol. Chem.*, **272**, 952–960.

Obenauer,J. *et al.* (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.

Pulgar,V. *et al.* (1999) Optimal sequences for non-phosphate-directed phoshorylation by protein kinase CK1 (casein kinase-1) - a re-evaluation. *Eur. J. Biochem.*, **260**, 520–526.

Ritz,A. *et al.* (2009) Discovery of phosphorylation motif mixtures in phosphoproteomics data. *Bioinformatics*, **25**, 14–21.

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

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

Shabb,J. (2001) Physiological substrates of cAMP-dependent protein kinase. *Chem. Rev.*, **101**, 2381–2411.

Tuazon,P. and Traugh,J. (1991) Casein kinase I and II–multipotential serine protein kinases: structure, function, and regulation. *Adv. Second Messenger Phosphoprotein Res.*, **23**, 123–164.

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

Villen,J. *et al.* (2007) Large-scale phosphorylation analysis of mouse liver. *PNAS*, **104**, 1488–1493.

Waterhouse,A. *et al.* (2009) Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.

Wilson-Grady,J. *et al.* (2008) Phosphoproteome analysis of fission yeast. *J. Proteome Res.*, **7**, 1088–1097.

Wong,Y. *et al.* (2007) KinasePhos 2.0 - a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.*, **35**, 588–594.

Xue,Y. *et al.* (2005) GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Res.*, **33**, 184–187.

Xue,Y. *et al.* (2006) SUMOsp: a web server for sumoylation site prediction. *Nucleic Acids Res.*, **34**, 254–257.

Yang,S. *et al.* (1993) Protein kinase FA/GSK-3 phosphorylates tau on Ser235-Pro and Ser404-Pro that are abnormally phosphorylated in Alzheimer's disease brain. *J. Neurochem.*, **61**, 1742–1747.

Yu,Y. *et al.* (2011) Phosphoproteomic analysis identifies Grb10 as an mTORC1 substrate that negatively regulates insulin signaling. *Science*, **332**, 1322–1326.

Zhai,B. *et al.* (2008) Phosphoproteome analysis of Drosophila melanogaster embryos. *J. Proteome Res.*, **7**, 1675–1682.