Bioinformatics, 32(2), 2016, 252–259
doi: 10.1093/bioinformatics/btv550
Advance Access Publication Date: 22 September 2015
Original Paper

OXFORD

Systems biology

Positive-unlabeled ensemble learning for kinase substrate prediction from dynamic phosphoproteomics data

Pengyi Yang[1,2,*], Sean J. Humphrey[3], David E. James[4], Yee Hwa Yang[5] and Raja Jothi[1,2,*]

[1]Systems Biology Section, [2]Epigenetics & Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, National Institutes of Health, RTP, NC 27709, USA, [3]Department of Proteomics and Signal Transduction, Max-Planck-Institute of Biochemistry, Martinsried, Germany, [4]Charles Perkins Centre, School of Molecular Bioscience, Sydney Medical School and [5]School of Mathematics and Statistics, The University of Sydney, NSW 2006, Australia

*To whom correspondence should be addressed.
Associate Editor: Ziv Bar-Joseph

Received on May 12, 2015; revised on August 18, 2015; accepted on September 11, 2015

Abstract

Motivation: Protein phosphorylation is a post-translational modification that underlines various aspects of cellular signaling. A key step to reconstructing signaling networks involves identification of the set of all kinases and their substrates. Experimental characterization of kinase substrates is both expensive and time-consuming. To expedite the discovery of novel substrates, computational approaches based on kinase recognition sequence (motifs) from known substrates, protein structure, interaction and co-localization have been proposed. However, rarely do these methods take into account the dynamic responses of signaling cascades measured from *in vivo* cellular systems. Given that recent advances in mass spectrometry-based technologies make it possible to quantify phosphorylation on a proteome-wide scale, computational approaches that can integrate static features with dynamic phosphoproteome data would greatly facilitate the prediction of biologically relevant kinase-specific substrates.

Results: Here, we propose a positive-unlabeled ensemble learning approach that integrates dynamic phosphoproteomics data with static kinase recognition motifs to predict novel substrates for kinases of interest. We extended a positive-unlabeled learning technique for an ensemble model, which significantly improves prediction sensitivity on novel substrates of kinases while retaining high specificity. We evaluated the performance of the proposed model using simulation studies and subsequently applied it to predict novel substrates of key kinases relevant to insulin signaling. Our analyses show that static sequence motifs and dynamic phosphoproteomics data are complementary and that the proposed integrated model performs better than methods relying only on static information for accurate prediction of kinase-specific substrates.

Availability and implementation: Executable GUI tool, source code and documentation are freely available at https://github.com/PengyiYang/KSP-PUEL.

Contact: pengyi.yang@nih.gov or jothi@mail.nih.gov

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

1 Introduction

Phosphorylation is an essential protein post-translational modification characterized by the precise and reversible addition of a phosphate group, by proteins called 'kinases', to their targets, called 'substrates' (Hunter, 1995). The signaling networks established by protein phosphorylation govern numerous cellular functions such as metabolic homeostasis, cell proliferation, survival and apoptosis (Lemmon and Schlessinger, 2010), and the identification of substrates often reveals new mechanisms by which the cell regulates these processes. The identification of new kinase substrates (i.e. the phosphorylation site on a given protein) is therefore of great biological interest.

Experimental validation of kinase substrates is an expensive and time-consuming process and must therefore be prioritized and performed for only a limited number of candidates. To select appropriate candidates for experimental characterization, *in silico* approaches are widely used to predict candidate substrates of a kinase or family of kinases of interest. To this end, a number of computational approaches have been developed for *de novo* substrate prediction (Miller and Blom, 2009). These range from simple motif-based approaches to sophisticated machine learning algorithms. A list of available methods is reviewed and categorized by Trost and Kusalik (2011) according to the techniques employed for substrate prediction. For example, methods based on sequence motif either contain precompiled regular expression patterns (Amanchy *et al.*, 2007) or rely on calculating position-specific scoring matrices (PSSMs) (Obenauer *et al.*, 2003; Yaffe *et al.*, 2001) from known kinase substrates extracted from the literature or *in vitro* experiments. While motif-based methods are able to sensitively identify substrates that have a similar sequence composition to those used to compile the motif (Miller *et al.*, 2008), they generally extend poorly to more diverse substrates. As increasingly more phosphorylation sites are deposited in public databases such as Phospho.ELM (Dinkel *et al.*, 2011) and PhosphoSitePlus (Hornbeck *et al.*, 2011), machine learning approaches that generalize well to diverse substrates are becoming increasingly necessary. In this category, methods differ in the learning algorithms that have been employed for modeling (Dang *et al.*, 2008; Kim *et al.*, 2004; Trost and Kusalik, 2013; Wong *et al.*, 2007; Xue *et al.*, 2008), the amount of sequence information utilized (Gao *et al.*, 2010; Xue *et al.*, 2010) and the integration of additional information such as protein structure (Hjerrild *et al.*, 2004), colocalization (Linding *et al.*, 2007) or interaction (Horn *et al.*, 2014; Song *et al.*, 2012). While these 'static' features (e.g. protein sequence, structure, co-localization and interaction) are informative, methods based on 'static' features do not account for the dynamic spatio-temporal kinetics that may help accurately define the architecture of cell signaling networks.

Recent advances in liquid chromatography and mass spectrometry has enabled high-throughput protein phosphorylation profiling in time course studies (Choudhary and Mann, 2010; Sabido *et al.*, 2012). Coupled with labeling techniques such as isobaric tagging (e.g. TMT) (Erickson *et al.*, 2014) or stable isotope labeling with amino acids in cell culture (e.g. SILAC) (Olsen *et al.*, 2006) and increasingly label-free-based approaches (Humphrey *et al.*, 2015; Oliveira *et al.*, 2015), tens of thousands of phosphorylation sites can be identified and quantified with high precision on a proteome-wide scale (Olsen and Mann, 2013). Computational approaches that integrate static features with dynamic phosphoproteome will enable kinase-specific substrate prediction in the context of dynamic cellular systems.

Here, we propose a positive-unlabeled ensemble learning algorithm to integrate kinase recognition motifs with dynamic phosphoproteomics data to predict novel substrates of relevant kinases of interest. Extending on our previous model (Humphrey *et al.*, 2013), the proposed algorithm addresses several key computational challenges faced in model training when using phosphoproteomics data by using a positive-unlabeled learning technique for an ensemble classification algorithm. First, the class distribution of the phosphoproteomics data is highly imbalanced due to the fact that experimentally verified substrates are sparse, resulting in vastly more negative training instances compared with positive instances. We address this by using repeated sampling from the phosphoproteomics data and creating an ensemble of classifiers each trained on a balanced training subset. Second, the model training requires a set of negative instances to be provided. However, defining a true negative set is difficult because unannotated phosphorylation sites comprise a mix of negative and unidentified positive substrates. To overcome this problem, a positive-unlabeled learning procedure is implemented in the ensemble model for improving prediction sensitivity while retaining specificity by correcting for the prediction bias. We studied the behavior of the proposed positive-unlabeled ensemble approach using simulation on synthetic datasets and subsequently applied it to predict substrates of kinases Akt and mTOR (mammalian target of rapamycin) using a time-course phosphorylation dataset generated from insulin-stimulated profiling of cultured adipocytes. Our results indicate that the proposed approach is highly accurate, based on an array of evaluation metrics, and is able to take advantage of both static information from amino acid sequences as well as dynamic information from phosphoproteomic data to predict kinase substrates.

2 Methods

Because the number of experimentally verified kinase substrates (positive set) are often very small compared with the complete set of phosphorylation sites profiled (unannotated set), our learning model (Fig. 1) relies on repeated sampling from the unannotated set and combines each sampling set with the positive set to create balanced subsets (Section 2.2). The base classifiers of support vector machines (SVMs) (Section 2.1) are subsequently trained using these balanced subsets with both static features, extracted from peptide sequences and dynamic features, extracted from phosphoproteomics data (Section 2.4). In addition, since the unannotated set comprises both true negatives and unidentified positives, a positive-unlabeled learning procedure that uses a correction factor estimated from the repartitioned data (Section 2.3) is applied for each base classifier to

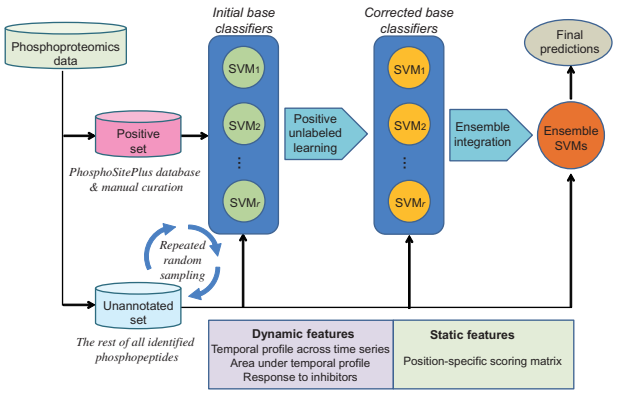

Fig. 1. Schematic flowchart illustrating the positive-unlabeled ensemble learning model for novel substrate prediction from phosphoproteomics data

correct for prediction bias. These corrected base classifiers are then integrated to form the ensemble of SVMs, and the final predictions are made based on the combined probability from all base classifiers.

2.1 Base classifier

Let x_i $(i = 1, ..., n)$ denote phosphorylation sites and let $y_i \in \{-1, 1\}$ be a binary label indicating whether x_i is a substrate ($y_i = 1$) of a kinase or not ($y_i = -1$). We use SVM, implemented by Chang and Lin (2011), to form the bases of an ensemble classifier. In dual formulation, an SVM is expressed as

$$\max_\alpha L(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j k(x_i, x_j)$$

$$\text{subject to}: \quad \sum_{i=1}^{n} y_i \alpha_i = 0, \quad 0 \le \alpha_i \le C,$$

where α_i are Lagrange multipliers and x_i for which $\alpha_i > 0$ are support vectors. The parameter C controls the fraction of support vectors, and the kernel of the classifier is defined by $k(x_i, x_j)$. Specifically, we use nonlinear classifiers with radial basis function kernels, which is expressed as

$$k_\sigma^{\mathrm{RBF}}(x_i, x_j) = \exp(-\frac{1}{\sigma}||x_i - x_j'||^2),$$

where σ is a parameter that controls the width of the radial basis function. When data are linearly inseparable, using SVMs with nonlinear kernel functions can provide better classification accuracy while also retaining the robustness of linear classifiers on background noise. The default parameters were used in this study where $C = 1$ and σ equals to the feature dimension of the data.

2.2 Ensemble model

Kinase substrate prediction can be formulated as constructing a model $f(x)$, such that $f(x) = p(y = 1|x)$ approximates the truth. For most kinases, often there are only a handful of experimentally validated substrates. We denote them as $x \in P$, where P is the set of validated substrates and hence positive. After excluding these positive substrates, the number of remaining phosphorylation sites (denoted by set U) profiled from a large-scale phosphoproteomics studies can easily exceed several thousand. This makes the prediction of kinase substrates from phosphoproteomics data an inherent class imbalanced learning problem, because the number of positive instances are significantly outnumbered by the rest of the profiled but as yet unannotated phosphorylation sites ($P \ll U$). Many popular classification algorithms, including SVMs that were used as the base classifiers of the ensemble, are sensitive to the imbalanced class distribution and would perform poorly without correction of the class distribution of the training instances (Tang *et al.*, 2009).

An effective approach to address this class imbalance is to create an ensemble classifier by sampling from the original dataset and creating base classifiers, each trained on a balanced subset (Yang *et al.*, 2014). This strategy can be modified for solving the imbalanced class problem in our phosphoproteomics application by randomly sampling from U and combining them with members of P to form balanced subsets $(U_1 \cup P, U_2 \cup P, ..., U_r \cup P)$, where r is the number of sampling operations performed. These subsets can then be used to train the base classifiers from which the ensemble is obtained for prediction:

$$p^E(y = 1|x) = \frac{1}{r}\sum_{i=1}^{r} p(y = 1|x; U_i \cup P)$$

The above prediction is an estimate of the combined probability p^E of a given phosphorylation site x to be a positive substrate of a kinase.

2.3 Positive-unlabeled learning in ensemble

One complication in applying the ensemble model described above is that U is an unannotated set that comprises a mixture of unidentified substrates of a given kinase as well as negatives. This is known as the positive-unlabeled learning problem where only a small set of positive instances are known, while the rest of the data contains both positive and negative instances but are unlabeled (Letouzey *et al.*, 2000). The ensemble model described above makes a naive assumption by treating all members of U (that is, $x \notin P$) as negative instances for sampling and training the base classifiers. The model would therefore predict conservatively and may penalize heavily on novel substrates because they may be selected as negative instances in model training. This is undesirable since our goal is precisely to discover from U the set of novel substrates.

It has been shown that by multiplying a correction factor, estimated from a data repartitioned approach, one can correct for the prediction bias on the unannotated set (Elkan and Noto, 2008). Here, we extend this positive-unlabeled learning technique for our ensemble classifier by estimating the correction factors for individual base classifiers and subsequently correcting their prediction bias in ensemble classification. Let $s = 1$ if the instance x is labeled, and $s = -1$ otherwise. Since only experimentally validated substrates are labeled, $y = 1$ is certain when $s = 1$, whereas y could be 1 or -1 when $s = -1$. Therefore, by treating all labeled instances as positive and all unlabeled ones as negative, we are in fact learning $g(x) = p(s = 1|x)$ and using this as approximation to $p(y = 1|x)$. Elkan and Noto (2008) have mathematically proved that $p(y = 1|x) = p(s = 1|x)/c$, where $c = p(s = 1|y = 1)$ and by reserving a subset of validation data from the training data the value of c can be estimated as

$$c = \frac{1}{m}\sum\nolimits_{x \in P^*} g(x),$$

where P^* is a set of labeled instances in the validation dataset and m is the cardinality of P^*.

For our ensemble classifier, we have, for each base classifier, $g_i(x) = p(s = 1|x; U_i \cup P)$, $(i = 1, ..., r)$. Therefore, a correction factor can be estimated by reserving a validation dataset for each base classifier $g_i(x)$ and ensemble model with prediction correction can then be expressed as

$$p_c^E(y = 1|x) = \frac{1}{r}\sum_{i=1}^{r} p(s = 1|x; U_i \cup P)/c_i$$

$$c_i = \frac{1}{m_i}\sum\nolimits_{x \in P_i^*} g_i(x).$$

The corrected and combined probability p_c^E is the estimation of x being a positive substrate of a given kinase.

2.4 Static and dynamic feature extraction from phosphoproteomics data

We demonstrate the proposed approach on a previously published phosphoproteomics dataset where insulin stimulation of 3T3-L1 adipocytes was performed to quantify the insulin signaling

phosphoproteome in adipocyte cells (Humphrey *et al.*, 2013). The phosphopeptides were quantified in biological triplicates and included a serum-starved ('basal') control and a time-course of insulin treatment (at 15 s, 30 s, 1 min, 2 min, 5 min, 10 min, 20 min and 60 min), as well as cells treated with or without PI3K/mTOR or Akt inhibitors (LY294002 and MK2206, respectively) applied for 30 min prior to a 20-min insulin treatment. Data were filtered to select phosphorylation sites that are quantified in at least one of the three biological replicates in each time point and the basal condition. This resulted in 12 289 phosphorylation sites.

For each kinase of interest, a positive training set of annotated substrates was curated from a combination of PhosphoSitePlus (Hornbeck *et al.*, 2011) database and manual extraction from literature. Next, the PSSM was calculated for each kinase of interest using the six amino acids flanking the actual phosphorylation site (total length: 13) of its annotated substrates as

$$M_{k,j} = \frac{1}{l}\sum_{i=1}^{l} I(a_{i,j} = k),$$

where l is the number of annotated substrates, j is the sequence window size and k is the set of 20 amino acids. All identified phosphorylation sites are then scored against $M_{k,j}$ to determine their similarity to the annotated substrates.

For dynamic features from phosphoproteome profiling, we calculate, for each phosphorylation site, the log 2-fold change at each time point compared with basal control and the log 2-fold change before and after inhibitor treatments. In addition, we extracted several secondary features for each phosphorylation site from its temporal pattern. These include the mathematical mean of the log 2-fold change at each time point compared with basal control, the area under the temporal profile calculated as follows:

$$\text{Area} = \frac{1}{2}\sum_{i=1}^{t}(y_{i-1} + y_i),$$

and the goodness of a polynomial (degree of two) curve fit measured by F-test as follows:

$$F = \frac{\sum_{i=1}^{t}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{t}\varepsilon_i},$$

where t is the number of time points, y_i is the standardized log 2-fold change at time point i, $\hat{y}_i$ is the fitted value of y_i and ε_i is the unfitted residuals from the model at time point i.

2.5 Performance evaluation and comparison

To compare the performance of our models and other commonly used kinase-specific substrate prediction tools, we used an array of evaluation metrics including sensitivity (Se), specificity (Sp), F_1 score, geometric mean (GM) and Matthews correlation coefficient (MCC) defined as follows:

$$\text{Se} = \frac{\text{TP}}{\text{TP}+\text{FN}}; \quad \text{Sp} = \frac{\text{TN}}{\text{FP}+\text{TN}};$$

$$\text{F}_1 = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}; \quad \text{GM} = \sqrt{\frac{\text{TP}}{\text{TP}+\text{FN}} \times \frac{\text{TP}}{\text{TP}+\text{FP}}};$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN} - \text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})\times(\text{TP}+\text{FN})\times(\text{TN}+\text{FP})\times(\text{TN}+\text{FN})}},$$

where TP, TN, FP and FN denote the number of true positives, true negatives, false positives and false negatives, respectively.

Beside using metrics defined above for evaluating a single classification threshold, we also generated Precision/Recall curves to evaluate the performance of the model trained using both motif and phosphoproteomics data and those trained using only motif or phosphoproteomics data across a range of threshold by varying the classification cutoff c as follows:

$$\text{Precision(c)} = \frac{\text{TP(c)}}{\text{TP(c)}+\text{FP(c)}}; \quad \text{Recall(c)} = \frac{\text{TP(c)}}{\text{TP(c)}+\text{FN(c)}}.$$

Methods with predefined prediction thresholds that cannot be adjusted to an arbitrary cutoff are represented as single points in the Precision/Recall comparison plot.

In the simulation study, we repeatedly generated 10 simulation datasets and for each dataset we trained and compared the prediction accuracy of a single SVM model, an uncorrected ensemble model and a corrected ensemble model using the evaluation metrics described above. To evaluate the individual contributions of the 'static' motif information and the dynamic phosphoproteome information, we created partial models using either motif or phosphoproteomics data and compared these with the model trained with both using the dynamic phosphoproteomics data. In addition, we also compared the performance of our approach with other commonly used kinase-specific substrate prediction tools including GPS 3.0 (Xue *et al.*, 2008), iGPS 1.0 (Song *et al.*, 2012), KinasePhos 1.0 (Huang *et al.*, 2005), KinasePhos 2.0 (Wong *et al.*, 2007), NetPhosK 1.0 (Hjerrild *et al.*, 2004) and NetworKIN 3.0 (Horn *et al.*, 2014). We used the default/suggested parameters for each prediction tool. Specifically, the prediction specificity of HMM is set to 0.9 for KinasePhos 1.0, the prediction specificity of SVM is set to 0.8 for KinasePhos 2.0, the minimum score of 2.0 and the max difference of 4.0 is set for NetworKIN 3.0, the prediction threshold is set to 0.5 for NetPhosK 1.0, high for GPS 3.0 and high for iGPS 1.0. For calculating sensitivity, the annotated substrates were treated as positive instances, and leave-one-out cross-validation was used for our model evaluation. For calculating specificity, since we do not know the set of true negative phosphorylation sites, we assumed that only a very small fraction of phosphorylation sites in the unlabeled set of U are positive instances with most of them being negative instances. On the basis of this assumption, we randomly sampled five sets of phosphorylation sites from all profiled phosphorylation sites, excluding annotated substrates of Akt and mTOR and evaluated each method on each of the five negative sets. The performance of each method is reported as the mean plus and minus the standard deviation.

3 Results

3.1 Evaluation of positive-unlabeled ensemble learning on synthetic data

We generated synthetic datasets to analyze the behavior of different classification models in classifying data with highly imbalanced class distribution and with positive and unlabeled instances. Figure 2a shows a typical example where a synthetic dataset is simulated with 10 positive instances (red points) with two features both generated from a Gaussian distribution $\mathcal{N}(6.5, 1)$ and 10 500 unlabeled instances, of which 500 are unannotated positive instances (dark blue points) and the remaining 10 000 are negative instances (cyan points) generated from a Gaussian distribution $\mathcal{N}(4.5, 1)$. In this example, a single SVM would predict every instance, including the 10 positives, as a negative (purple region in Fig. 2b). This is because the negative class is significantly over-represented in the training data.

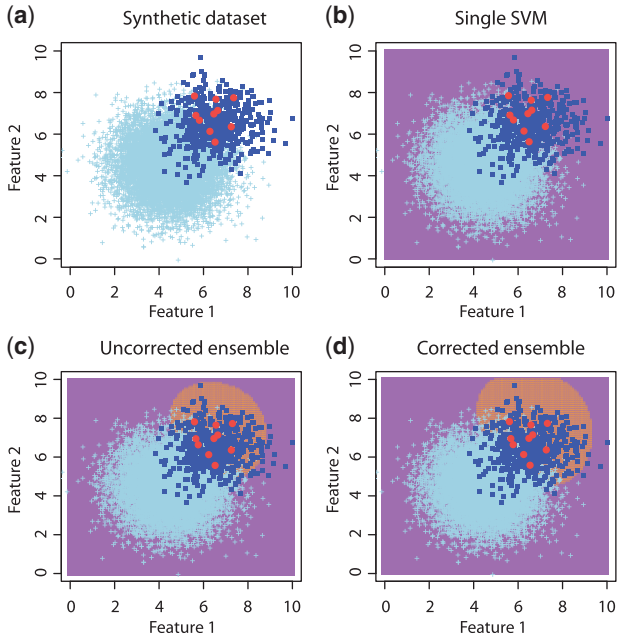

Fig. 2. Synthetic data for analyzing the behavior of classification models. (**a**) A synthetic dataset containing 10 positive examples (red filled circles) and 10 500 unlabeled examples of which 500 are unlabeled positive examples (dark blue filled squares) and the remaining 10 000 are negative examples (cyan +s). (**b**) Negative prediction region (purple rectangle) from a single classifier of SVM. (**c**) Negative prediction region (purple rectangle) and positive prediction region (orange oval) from an ensemble of SVMs without positive-unlabeled learning (ensemble size: 50). (**d**) Negative prediction region (purple) and positive prediction region (orange oval) from an ensemble of SVMs with positive-unlabeled learning (ensemble size: 50)

In comparison, an ensemble model without positive-unlabeled learning correction is able to correctly classify 10 positive instances as well as a large proportion of unlabeled positive instances (orange in Fig. 2c). By applying positive-unlabeled learning correction for the base classifiers, an ensemble model is able to correctly classify even more unlabeled positive instances (orange in Fig. 2d)

To verify if corrected ensemble models (by positive-unlabeled learning) perform consistently better than the single SVMs or the uncorrected ensemble models, we repeated the above simulation 10 times, creating 10 synthetic datasets and evaluated each model based on multiple performance metrics. As listed in Table 1, single SVMs always predict all instances as negatives. When it comes to ensemble models with and without positive-unlabeled learning correction, corrected ensemble models significantly improve prediction sensitivities compared with uncorrected ensemble models while retaining comparable specificities. The overall performance of corrected ensemble models is also better than uncorrected ensemble models according to F_1 score, GM or MCC.

3.2 Application of positive-unlabeled ensemble learning on insulin-activated phosphoproteome

We demonstrate the application of the proposed method using the mass spectrometry-based phosphoproteomics data generated from insulin-activated time-series studies of 3T3-L1 adipocyte cells (Humphrey *et al.*, 2013).

3.2.1 Kinase selection and feature extraction

Before applying the proposed approach for kinase-specific substrate prediction, we need to determine kinases that are of interest and

Table 1. Comparison of single SVM, uncorrected ensemble and corrected ensemble by positive-unlabeled learning using synthetic datasets

	Single SVM	Uncorrected ensemble	Corrected ensemble
Se (%)	0.0 ± 0.0	83.6 ± 2.9	90.2 ± 2.5
Sp (%)	100.0 ± 0.0	92.4 ± 1.1	90.9 ± 2.5
F_1 (%)	0.0 ± 0.0	87.9 ± 1.3	90.1 ± 0.6
GM (%)	NaN	88.0 ± 1.2	90.2 ± 0.6
MCC	NaN	0.774 ± 1.9	0.803 ± 1.1

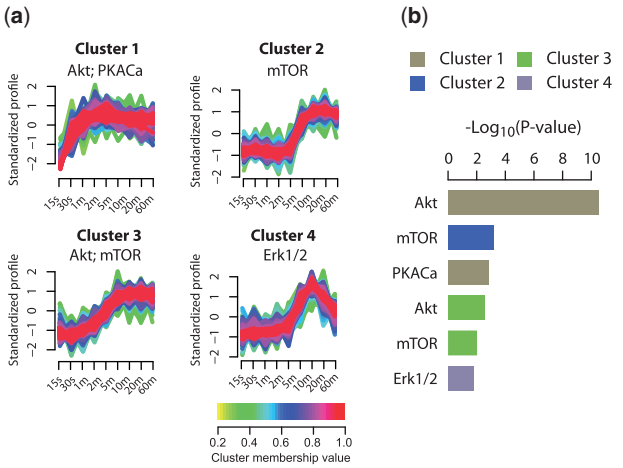

Fig. 3. Kinase substrate enrichment analysis. (**a**) Temporal profile clusters with kinases whose substrates are enriched. Color gradient correspond to the cluster membership score generated by fuzzy *c*-means clustering. (**b**) Enrichment based on Fisher's exact test ($-\log_{10}(p)$) for each kinase. Each shade of color corresponds to a cluster index

relevant to insulin-activated phosphoproteomics data. Typically, kinases that are activated or inhibited in the context of the experiment performed are of particular interest. Since many substrates of a given kinase have similar temporal kinetics, clustering phosphorylation sites into distinct clusters and testing for enrichment of known substrates of each kinase (obtained from an annotation database such as PhosphoSitePlus) (Hornbeck *et al.*, 2011) can facilitate the identification of kinases that are perturbed during the time course and/or by the treatments in the phosphoproteomics experiments. To this end, we filtered the set of all phosphorylation sites by selecting only those that have an associated gene product and are differentially phosphorylated in at least one of the eight time points profiled, as determined using a moderated *t*-test (Gentleman *et al.*, 2005) (FDR < 0.05). This resulted in 3178 regulated phosphorylation sites. We then applied knowledge-based cluster evaluation (CLUE) (Yang *et al.*, 2015) to determine the optimal clustering of the phosphorylation sites and identified kinases whose substrates are enriched within each cluster. Using CLUE, we identified four clusters enriched with known substrates of a set of kinases (Fig. 3a). Notably, Akt and mTOR substrates are the most highly enriched (Fig. 3b) based on Fisher's exact test. We therefore selected Akt and mTOR as kinases of interest for substrate prediction.

After selecting Akt and mTOR as kinases of interest, we used a combination of manual curation and database extraction to compile lists of annotated substrates for the Akt and mTOR kinases. We found a total of 22 and 26 substrates for Akt and mTOR, respectively, and used these as positive instances for model training in subsequent analyses (Supplementary Table S1). We next constructed

PSSMs using sequence recognition sites of the known Akt and mTOR substrates (Fig. 4a) and generated consensus motifs (Thomsen and Nielsen, 2012) for learning. Next, we extracted temporal information for the known Akt and mTOR substrates from the phosphoproteomics data (Fig. 4b–d). Boxplots of the log 2-fold change of insulin stimulation versus basal control across the eight time points indicate that Akt substrates responded to insulin activation significantly faster than those of mTOR (Fig. 4b). This is confirmed by calculation of the area under the curve (see Section 2.4), where Akt substrates generally have very large values (Fig. 4c). This temporal latency is in agreement with the known cellular topology of the insulin signaling network in which Akt is proximal to the insulin receptor and is rapidly activated by multisite phosphorylation (Humphrey and James, 2012), while mTORC1 is downstream of Akt and relies on the multiple discrete steps including the phosphorylation of TSC2 by Akt, ultimately leading to the activation of the kinase (Laplante and Sabatini, 2012). Interestingly, the area under the curve values for mTOR substrates appears to be bimodal suggesting a subset of mTOR substrates may respond to insulin activation more slowly than the others (Fig. 4c). This is consistent with the enrichment analysis (Fig. 3) where mTOR is found to be enriched in both a relatively faster cluster (cluster 3) and a slower cluster (cluster 2). Figure 4d shows the log 2-fold change of phosphorylation level in insulin stimulation with and without prior treatment of inhibitors. The phosphorylation levels of both Akt and mTOR substrates appear to be inhibited by both LY and MK inhibitors although to different degrees.

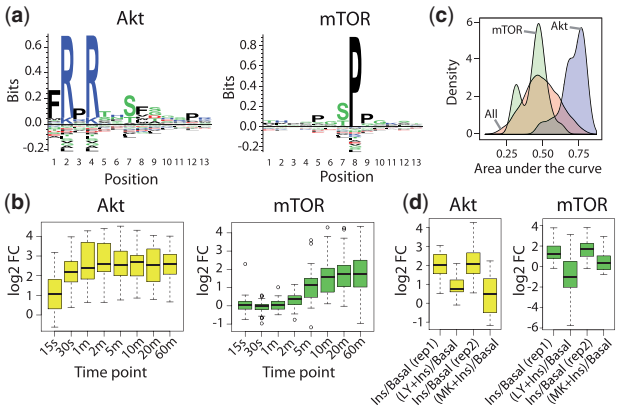

Fig. 4. Feature extraction. (**a**) Recognition motifs of Akt and mTOR kinases generated from PSSMs of known Akt and mTOR substrates. (**b**) Log 2-fold change of insulin stimulated phosphorylation level from each time point compared with basal for defined Akt and mTOR substrates. (**c**) Distribution of area under the curve values calculated from the temporal profile of all phosphorylation sites (red), defined Akt substrates (blue) and defined mTOR substrates (green). (**d**) Log 2-fold change of insulin stimulated phosphorylation level at 20 min with and without prior treatment of PI3K/mTOR inhibitor (LY) or Akt inhibitor (MK) compared with basal

3.2.2 Determining effective ensemble size

A key parameter in designing an effective ensemble of classifiers is the number of base classifiers used for creating the final ensemble model ('ensemble size'). We tested a range of ensemble sizes for predicting Akt and mTOR substrates, respectively, and calculated the correlation of prediction scores for all phosphorylation sites. This procedure was repeated 10 times to obtain the variance introduced by using a positive-unlabeled ensemble model with a given ensemble size. As shown in Figure 5, the correlation from each individual prediction increases with the ensemble size and plateaus at size of around 30. Since ensemble models with smaller size yield more variable predictions while larger size yield more consistent predictions (Fig. 5), we made a conservative decision to use an ensemble size of 50 for subsequent ensemble models.

3.2.3 Predicting Akt- and mTOR-specific substrates

After extracting learning features and determining the effective ensemble size, we compared the proposed Positive-Unlabeled Ensemble Learning models that use both motif and phosphoproteome data for prediction ('PUEL') with those that use only the motif ('-Motif') or the

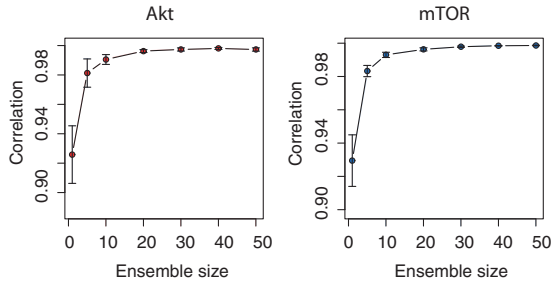

Fig. 5. Ensemble size estimation. Positive-unlabeled ensemble models with a range of ensemble sizes are tested for predicting Akt and mTOR substrates, respectively. For each ensemble size, 10 ensemble models are created using random sampling and the mean and the standard deviation of the correlation from these 10 predictions is plotted

Table 2. Evaluation of kinase-specific substrate prediction for Akt and mTOR

	Akt					mTOR				
	Se (%)	Sp (%)	F_1 (%)	G mean (%)	MCC	Se (%)	Sp (%)	F_1 (%)	G-Mean (%)	MCC
PUEL	100.0	99.1 ± 2.0	99.6 ± 1.0	99.6 ± 0.9	0.991 ± 0.019	88.5	86.4 ± 7.2	86.4 ± 5.4	88.6 ± 2.7	0.749 ± 0.068
Motif	90.9	98.2 ± 2.5	94.4 ± 1.2	94.4 ± 1.3	0.894 ± 0.026	80.8	90.9 ± 3.2	85.7 ± 1.2	85.9 ± 1.3	0.715 ± 0.033
Phospho	90.9	93.6 ± 5.2	92.2 ± 2.4	92.3 ± 2.4	0.847 ± 0.053	80.8	78.2 ± 7.5	81.1 ± 2.6	81.2 ± 2.6	0.590 ± 0.072
EL	95.4	99.1 ± 2.0	97.2 ± 1.0	97.3 ± 1.0	0.946 ± 0.021	84.6	87.3 ± 7.5	86.6 ± 2.8	86.7 ± 2.8	0.718 ± 0.073
GPS 3.0	100.0	77.3 ± 8.5	89.9 ± 3.4	90.4 ± 3.0	0.795 ± 0.070	61.5	70.9 ± 9.4	66.2 ± 2.9	66.5 ± 3.3	0.326 ± 0.098
iGPS 1.0	45.5	91.8 ± 6.7	59.3 ± 2.6	62.4 ± 3.9	0.425 ± 0.092	0.0	100.0 ± 0.0	0.0 ± 0.0	NaN	NaN
KinasePhos 1.0	86.4	77.3 ± 5.6	82.6 ± 2.2	82.8 ± 2.1	0.640 ± 0.052	—	—	—	—	—
KinasePhos 2.0	72.7	94.5 ± 5.9	81.7 ± 2.7	82.4 ± 3.1	0.691 ± 0.069	—	—	—	—	—
NetPhosK 1.0	90.9	97.3 ± 2.5	93.9 ± 1.2	94.0 ± 1.3	0.884 ± 0.027	—	—	—	—	—
NetworKIN 3.0	86.4	99.1 ± 2.0	92.2 ± 1.0	92.5 ± 1.1	0.862 ± 0.022	—	—	—	—	—

'PUEL': prediction using both motif and phosphoproteome for positive-unlabeled ensemble learning; '-Motif': prediction using motif only; '-Phospho': prediction using phosphoproteome only; 'EL': prediction using ensemble learning [as described in Humphrey *et al.* (2013)] but without positive-unlabeled learning technique. The calculation of each evaluation metrics are described in Section 2.5.

phosphoproteome data ('-Phospho'). As can be seen from Table 2, the performance of the models using both types of information performed better in both Akt and mTOR substrate prediction. This is especially evident in the case of mTOR, where the sensitivity of the PUEL is 88.5%, while for the models trained on only motif or phosphoproteome data, this is 80.8%. It is evident that PUEL performs better across the range of classification thresholds for mTOR and Akt, based on Precision/Recall curves (Fig. 6a). These results demonstrate the complementary nature of static sequence motifs and dynamic phosphoproteome data in kinase substrate prediction. It is also evident that PUEL outperforms the ensemble model without positive-unlabeled learning ('EL') (Table 2), which demonstrates a contribution by positive-unlabeled ensemble learning toward kinase substrate prediction.

To assess the performance of our approach in relation to that of other approaches, we evaluated substrate prediction accuracy for Akt and mTOR kinases using six popular kinase substrate prediction tools (Table 2). For Akt substrate prediction, our ensemble approach achieved the highest sensitivity and specificity. NetworKIN and NetPhosK 1.0 performed comparably with high specificity but with relatively low sensitivity. Compared with KinasePhos 2.0, KinasePhos 1.0 offers good sensitivity but lower specificity. GPS 3.0 has a perfect prediction sensitivity for Akt substrate prediction. However, this seems to be achieved by sacrificing specificity. iGPS 1.0 improves the specificity compared with GPS 3.0 but at the cost of sensitivity. For mTOR substrate prediction, GPS 3.0 did not perform as well as our approach. Its variant iGPS 1.0 did not predict any tested phosphorylation sites to be a positive substrate of mTOR. The pre-trained models used by KinasePhos 1.0/2.0, NetPhosK 1.0 and NetworKIN 3.0 preclude them from making substrates predictions for mTOR. Overall, the positive-unlabeled ensemble model performed better than other methods in predicting both Akt and mTOR substrates according to the evaluation metrics (Table 2) as well as the Precision/Recall curves (Fig. 6a).

Figure 6b shows the prediction scores (*z*-axis) of Akt and mTOR substrates, respectively, using the positive-unlabeled ensemble model with both motif features and phosphoproteome features. The contribution of motif feature and the phosphoproteome features are plotted on *x*-axis and *y*-axis. A complementarity pattern of motif score and phosphoproteome score can be seen in both Akt and mTOR

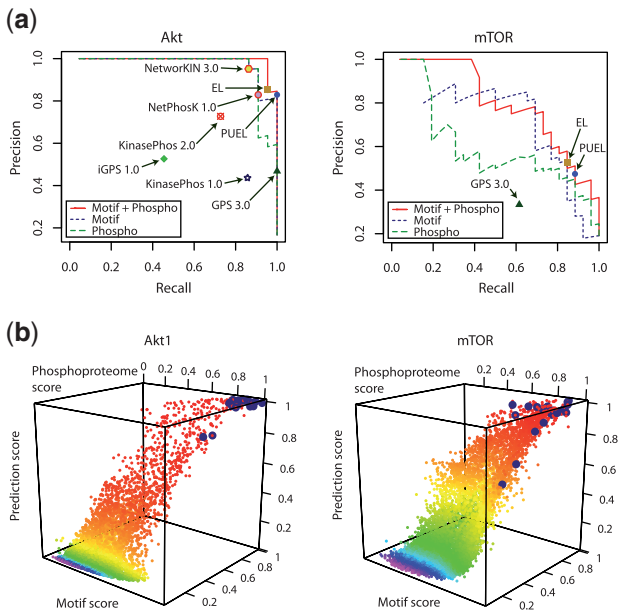

Fig. 6. (**a**) Precision/Recall curves for performance assessment of Akt and mTOR substrate prediction. Methods with predefined prediction thresholds are represented as single points. (**b**) Prediction of Akt and mTOR substrates using positive-unlabeled ensemble approach. Prediction scores (probabilities) from the model are plotted on *y*-axis. Prediction scores using sequence motif only or phosphoproteomic data only are plotted on *x*-axis and *z*-axis, respectively. Each point corresponds to a phosphorylation site profiled by the phosphoproteomics experiments and is rainbow colored by the value of prediction score. The larger points highlighted in dark blue (top-right) are the annotated substrates of Akt and mTOR, respectively

Table 3. List of top-20 predicted Akt and mTOR substrates

	Akt predictions						mTOR predictions					
Rank	Gene symbol	Site	Known	Full model	Motif	Phospho	Gene Symbol	Site	Known	Full model	Motif	Phospho
1	Tsc2	981	✓	0.999	0.999	0.999	Patl1	184	✓	0.999	0.998	0.996
2	Foxo1	316	✓	0.996	0.988	0.989	Grb10	503	✓	0.993	0.998	0.987
3	Irs2	303		0.995	0.983	0.974	Ulk2	772		0.993	0.873	0.984
4	Cep131	114		0.994	0.910	0.999	C2cd5	295		0.987	0.873	0.975
5	Gsk3a	21	✓	0.994	0.995	0.990	Ulk2	781		0.985	0.827	0.973
6	Ndr3	331		0.993	0.997	0.986	Wdr91	257		0.984	0.608	0.993
7	Irs1	265	✓	0.992	0.956	0.988	Lip1	444		0.983	0.710	0.968
8	Pfkfb2	486	✓	0.991	0.997	0.990	Maf1	68	✓	0.980	0.992	0.964
9	Flnc	2234		0.990	0.886	0.997	Oxr1	62		0.973	0.994	0.961
10	Fkhr2	252	✓	0.988	0.945	0.989	Znf503	107		0.971	0.547	0.964
11	Tsc2	939	✓	0.988	0.998	0.980	Tcfeb	167		0.964	0.589	0.970
12	Gsk3b	9	✓	0.987	0.994	0.981	Ulk1	450		0.964	0.603	0.990
13	Rtkn	520		0.987	0.977	0.978	Dock1	1772		0.963	0.643	0.967
14	Tsc2	1466	✓	0.985	0.983	0.989	Mapk6	704		0.960	0.647	0.971
15	Uhrf1bp1	430		0.982	0.637	0.987	Maf1	60	✓	0.960	0.972	0.938
16	As160	324		0.982	0.983	0.987	Rhbdf2	295		0.960	0.827	0.960
17	Cables1	272	✓	0.981	0.988	0.981	Pik3r4	905		0.957	0.772	0.957
18	Kank1	325		0.980	0.886	0.983	Mtus1	505		0.957	0.520	0.995
19	Fam13a	322		0.980	0.945	0.972	Tbc1d10b	693	✓	0.957	0.964	0.931
20	Mllt4	1802		0.978	0.464	0.982	Lpin1	323		0.955	0.620	0.988

Full model: prediction using both motif and phosphoproteome; Motif: prediction using motif only; Phospho: prediction using phosphoproteome only.

predictions. These results argue for the utilization and integration of dynamic features extracted from phosphoproteomics data with traditional motif features. Table 3 lists the top 20 substrates predicted for Akt and mTOR, with Supplementary Table S2 containing the full list of predictions.

4 Conclusion

In this study, we developed a positive-unlabeled ensemble learning approach that enables the accurate prediction of kinase-specific substrates by integrating motif information derived from amino acid sequences surrounding the phosphorylation recognition sites with dynamic phosphorylation patterns quantified in large-scale time-series phosphoproteomics studies. We extended the positive-unlabeled learning techniques for ensemble learning and illustrated how such an application can improve the prediction sensitivity while retaining high specificity for kinase-substrate prediction. Finally, we demonstrated the complementary nature of the static features extracted from sequence motif and the dynamic features extracted from phosphoproteomics data. Compared with traditional sequence motif-centric methods, our study suggests an alternative approach for predicting kinase-specific substrates that incorporates and exploits the wealth of physiological information captured by large-scale phosphoproteomics studies.

With the recent development of high-throughput methods promising to streamline the generation of large-scale phosphoproteomics data (Humphrey *et al.*, 2015), we anticipate that the methods described here will become an increasingly valuable approach for enabling prediction of kinase substrates, while preserving the dynamic and cellular context of the biological system being studied.

Funding

This work was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (1ZIAES102625 to R.J.).

Conflict of Interest: none declared.

References

Amanchy,R. *et al.* (2007) A curated compendium of phosphorylation motifs. *Nat. Biotechnol.*, **25**, 285–286.

Chang,C.-C. and Lin,C.-J. (2011) LibSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)*, **2**, 27.

Choudhary,C. and Mann,M. (2010) Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.*, **11**, 427–439.

Dang, T.H. *et al.* (2008) Prediction of kinase-specific phosphorylation sites using conditional random fields. *Bioinformatics*, **24**, 2857–2864.

Dinkel,H. *et al.* (2011) Phospho. elm: a database of phosphorylation sitesupdate 2011. *Nucleic Acids Res.*, **39**(suppl 1), D261–D267.

Elkan,C. and Noto,K. (2008) Learning classifiers from only positive and unlabeled data. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Las Vegas, Nevada, USA, pp 213–220.

Erickson,B.K. *et al.* (2014) Evaluating multiplexed quantitative phosphopeptide analysis on a hybrid quadrupole mass filter/linear ion trap/orbitrap mass spectrometer. *Anal. Chem.*, **87**, 1241–1249

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

Gentleman,R. *et al.* (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York.

Hjerrild,M. *et al.* (2004) Identification of phosphorylation sites in protein kinase A substrates using artificial neural networks and mass spectrometry. *J. Proteome Res.*, **3**, 426–433.

Horn,H. *et al.* (2014) Kinomexplorer: an integrated platform for kinome biology studies. *Nat. Methods*, **11**, 603–604.

Hornbeck,P.V. *et al.* (2011) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.*, **40**, D261–D270.

Huang,H.-D. *et al.* (2005) Kinasephos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res.*, **33**(suppl 2), W226–W229.

Humphrey,S. *et al.* (2015) High-throughput phosphoproteomics reveals *in vivo* insulin signaling dynamics. *Nat. Biotechnol*, **33**, 990–995.

Humphrey,S.J. and James,D.E. (2012) Uncaging akt. *Sci. Signal.*, **5**, pe20.

Humphrey,S.J. *et al.* (2013) Dynamic adipocyte phosphoproteome reveals that akt directly regulates mtorc2. *Cell Metab.*, **17**, 1009–1020.

Hunter,T. (1995). Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling. *Cell*, **80**, 225–236.

Kim,J.H. *et al.* (2004) Prediction of phosphorylation sites using SVMs. *Bioinformatics*, **20**, 3179–3184.

Laplante,M. and Sabatini,D.M. (2012) mTor signaling in growth control and disease. *Cell*, **149**, 274–293.

Lemmon,M.A. and Schlessinger,J. (2010) Cell signaling by receptor tyrosine kinases. *Cell*, **141**, 1117–1134.

Letouzey,F. *et al.* (2000) Learning from positive and unlabeled examples. In: Arimura,H. *et al.* (eds.), *Algorithmic Learning Theory*. Springer, Berlin Heidelberg, pp. 71–84.

Linding,R. *et al.* (2007) Systematic discovery of in vivo phosphorylation networks. *Cell*, **129**, 1415–1426.

Miller,M.L. and Blom,N. (2009) Kinase-specific prediction of protein phosphorylation sites. In: de Graauw,M. (ed), *Phospho-Proteomics*. Springer, pp. 299–310.

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

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

Oliveira,A.P. *et al.* (2015) Dynamic phosphoproteomics reveals torc1-dependent regulation of yeast nucleotide and amino acid biosynthesis. *Sci. Signal.*, **8**, rs4.

Olsen,J.V. and Mann,M. (2013) Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol. Cell. Proteomics*, **12**, 3444–3452.

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

Sabido,E. *et al.* (2012) Mass spectrometry-based proteomics for systems biology. *Curr. Opin. Biotechnol.*, **23**, 591–597.

Song,C. *et al.* (2012) Systematic analysis of protein phosphorylation networks from phosphoproteomic data. *Mol. Cell. Proteomics*, **11**, 1070–1083.

Tang,Y. *et al.* (2009) SVMs modeling for highly imbalanced classification. *IEEE Trans. Syst. Man Cybern. B Cybern.*, **39**, 281–288.

Thomsen,M.C.F. and Nielsen,M. (2012) Seq2logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.*, **40**, W281–W287.

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

Trost,B. and Kusalik,A. (2013) Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights. *Bioinformatics*, **29**, 686–694.

Wong,Y.-H. *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**(suppl 2), W588–W594.

Xue,Y. *et al.* (2008) Gps 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol. Cell. Proteomics*, **7**, 1598–1608.

Xue,Y. *et al.* (2010) Gps 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. *Protein Eng. Des. Sel*, **24**, 255-260.

Yaffe,M.B. *et al.* (2001) A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat. Biotechnol.*, **19**, 348–353.

Yang,P. *et al.* (2014) Sample subset optimization techniques for imbalanced and ensemble learning problems in bioinformatics applications. *IEEE Trans. Cybern.*, **44**, 445–455.

Yang,P. *et al.* (2015) Knowledge-based analysis for detecting key signaling events from time-series phosphoproteomics data. *PLoS Comput. Biol.*, **11**, e1004403.