Machine Learning of Global Phosphoproteomic Profiles Enables Discrimination of Direct versus Indirect Kinase Substrates

Evgeny Kanshin‡, Sébastien Giguère‡, Cheng Jing‡, Mike Tyers‡§||, and Pierre Thibault‡¶

Mass spectrometry allows quantification of tens of thousands of phosphorylation sites from minute amounts of cellular material. Despite this wealth of information, our understanding of phosphorylation-based signaling is limited, in part because it is not possible to deconvolute substrate phosphorylation that is directly mediated by a particular kinase versus phosphorylation that is mediated by downstream kinases. Here, we describe a framework for assignment of direct in vivo kinase substrates using a combination of selective chemical inhibition, quantitative phosphoproteomics, and machine learning techniques. Our workflow allows classification of phosphorylation events following inhibition of an analog-sensitive kinase into kinase-independent effects of the inhibitor, direct effects on cognate substrates, and indirect effects mediated by downstream kinases or phosphatases. We applied this method to identify many direct targets of Cdc28 and Snf1 kinases in the budding yeast Saccharomyces cerevisiae. Global phosphoproteome analysis of acute time-series demonstrated that dephosphorylation of direct kinase substrates occurs more rapidly compared with indirect substrates, both after inhibitor treatment and under a physiological nutrient shift in wt cells. Mutagenesis experiments revealed a high proportion of functionally relevant phosphorylation sites on Snf1 targets. For example, Snf1 itself was inhibited through autophosphorylation on Ser391 and new phosphosites were discovered that modulate the activity of the Reg1 regulatory subunit of the Glc7 phosphatase and the Gal83 β-subunit of SNF1 complex. This methodology applies to any kinase for which a functional analog sensitive version can be constructed to facilitate the dissection of the global phosphorylation network. Molecular & Cellular Proteomics 16: 10.1074/mcp.M116.066233, 786–798, 2017.

From the ‡Institute for Research in Immunology and Cancer, §Department of Medicine, ¶Department of Chemistry, Université de Montréal, C.P. 6128, Succursale centre-ville, Montréal, Québec, H3C 3J7, Canada

Received December 5, 2016, and in revised form, February 13, 2017

Published, MCP Papers in Press, March 6, 2017, DOI 10.1074/mcp.M116.066233

Author contributions: E.K., P.T., and M.T. designed the research; E.K., S.G., and J.C. performed the research; E.K., S.G., M.D.T., and P.T. analyzed data; and E.K., S.G., M.T., and P.T. wrote the paper.

© 2017 by The American Society for Biochemistry and Molecular Biology, Inc.
This paper is available on line at http://www.mcponline.org

Virtually all cellular behavior is influenced either directly or indirectly by protein phosphorylation (1). The extensive phosphorylation-based signaling network of the cell (2, 3) is governed by a highly interconnected network of protein kinases, phosphatases, and phospho-dependent recognition modules. A central problem in modern biology and drug discovery is therefore to identify the direct in vivo targets of protein kinases. Recent advances in mass spectrometry (MS) have allowed the profiling of 10,000s of phosphosites from few milligrams of cell lysate (4, 5). However, the diversity of kinases in the cell and the organization of kinases into complex regulatory networks precludes the effective deconvolution of large-scale phosphoproteomic datasets into defined signaling hierarchies (6). A particular difficulty arises in isolating the effects of a specific kinase from downstream kinases or phosphatases that form a local signaling network. This situation is further complicated by the propensity of kinases to form a globally connected network (7). Thus, even highly specific perturbation of a given kinase may trigger many ancillary events not associated with the primary set of target substrates. In addition, many phosphorylation events may arise from nonspecific kinase activity, which may contribute substantial nonfunctional noise to the phosphoproteome (8, 9).

Protein kinases must correctly discriminate their cognate substrates and phosphorylation sites within these substrates from a nonspecific background of ~700,000 potentially phosphorylatable residues in the proteome. There are multiple mechanisms responsible for this exquisite specificity, including the structure of the catalytic site, local and distal interactions between the kinase and substrate, the formation of complexes with scaffolding and adaptor proteins that spatially regulate the kinase, systems-level competition between substrates, and error-correction mechanisms (10). The responsibility for the recognition of substrates by protein kinases appears to be distributed among a large number of independent, imperfect specificity mechanisms, many of which are not fully recapitulated by in vitro assays. Conversely, the interpretation of in vivo data is confounded by the complexity and cross-connections of the global signaling network, in which perturbation of a single kinase can alter phosphorylation not only of cognate direct substrates but also through indirect...
effects on many other substrates that may reside in distant regions of the network. Due to these difficulties, the identification of direct kinase substrates requires both in vivo and in vitro studies with multiple levels of cross-validation that is time consuming and often not comprehensive (11, 12).

Genetic approaches enable the manipulation of kinase activities in vivo through either simple gene deletion or the targeted introduction of inactivating or activating point mutations. However, the interpretation of the effects of such mutations on the phosphoproteome is heavily confounded by compensating effects in the network (13). In a more elegant approach, it is possible to engineer many protein kinases to accept cell-permeable bulky ATP analogs that are not able to bind or inhibit wild-type (wt) kinases (14). These engineered kinases, which are referred to as analog-sensitive (as) alleles, provide several key advantages over other genetic techniques for kinase inactivation. In particular, the chemical inhibitor can be applied on short timescales to avoid compensation mechanisms that would otherwise buffer the loss of the target kinase. Both tyrosine and serine/threonine kinases are amenable to this approach, which has been used for selective inhibition of mutant Src-family kinases (14), Abl-family kinases (15), mitogen-activated kinases (16), cyclin-dependent kinases (CDKs) (16), p21-activated kinases (17), and Ca\(^{2+}\)-/calmodulin-dependent kinases. Despite the power of the as-allele approach, the functional integrity of the mutant kinase and the actual target specificity of the inhibitor are crucial considerations. It has been reported that the catalytic activity of some protein kinases is affected by as mutations (15) and unintended off-target effects of the inhibitor on other enzymes (18).

Given that kinase inactivation will shift the equilibrium of cognate substrates to the dephosphorylated state, due to the action of constitutive phosphatases, a combination of quantitative phosphoproteomics and selective chemical inhibition can in principle be used to discover in vivo kinase targets on a large-scale. This approach has for example been used to identify Cdc28 substrates in S. cerevisiae (19). Based on criteria of at least twofold dephosphorylation in response to kinase inhibition and the presence of a full CDK consensus motif, a total of 547 phosphorylation sites have been assigned to 308 candidate Cdc28 substrates (19). Unexpectedly, a significant proportion of the quantified sites on these substrates were up-regulated following Cdc28 inhibition, which suggests the influence of feedback loops that may activate additional kinases and/or inactivate phosphatases. Notably, some of these up-regulated phosphosites were even located within the CDK consensus sequence. Additional criteria are thus needed to resolve the spectrum of direct in vivo substrate classes for any given kinase.

To address the above issues with kinase substrate identification, we set out to develop a method for unbiased large-scale identification of direct in vivo kinase substrates based on a combination of specific kinase inhibition, large-scale phosphoproteomics, and machine learning approaches (Fig. 1a). The point of departure for our strategy is the selective inhibition of an as-form of a kinase of interest by the chemical inhibitor 1NM-PP1 (20). Upon treatment with the inhibitor, we expect changes in phosphorylation in response to the inhibitor itself (kinase independent) and changes that result from inhibition of the as kinase (kinase dependent). The latter may arise from either dephosphorylation of direct substrates or phosphorylation/dephosphorylation of indirect substrates. The latter substrates would represent downstream targets that are not directly phosphorylated by a kinase, for instance when the targeted kinase modulates the activity of another kinase or phosphatase, which in turn affects the phosphorylation of other proteins. To discriminate between kinase-independent, direct, and indirect substrate classes, we developed and validated a machine learning strategy that uses the support vector machine (SVM) (21) and ideas from positive and unlabeled learning (22). A proof of principle using the well-studied Cdc28 kinase in yeast (23) allowed us to identify many novel substrates and obtain a deeper understanding of kinase–substrate relationships. To demonstrate applicability to other less-characterized kinases, we then use our approach to identify direct targets of Snf1, a kinase that is required for transcription of glucose-repressed genes, thermotolerance, sporulation, and peroxisome biogenesis in yeast (24, 25). Our integrated approach should enable the identification of direct substrates for any kinase for which a functional as allele can be engineered. The systematic application of this method should help reveal the true structure of the global phosphorylation-based network and provide insights into biological regulation.

**EXPERIMENTAL PROCEDURES**

**Experimental Design and Statistical Rationale**—Inhibition of CDK\(^{1+}\) and SNF1 with 1NM-PP1 was performed as six independent treatments, every time using reverse triple SILAC labeling (see supplemental Fig. S1). To find significantly affected phosphosites, t test pValue threshold was chosen based on false discovery rate (FDR) <5\% using permutation-based correction. For all dynamic (time course) experiments, only one replicate was used for analysis and significance was calculated based on the shape of the entire kinetic profile rather than individual measurements (26).

**Yeast Strain Construction**—The S. cerevisiae strains used in this study were isogenic with the S288C background. Single deletion

\(^{1}\) The abbreviations used are: CDK1, cyclin-dependent kinase; Cdc28, cyclin-dependent kinase 1; as, analog-sensitive alleles; 1NM-PP1, 1-(1,1-dimethylpropyl)-3-(1-naphthalenylmethyl)-1H-pyrazolo[3, 4-d]pyrimidin-4-amine; OD, optical density; SILAC, stable isotope labeling with amino acids in culture; TRIS, tris(hydroxymethyl)aminomethane; FDR, false discovery rate; SVM, support vector machine; GS, generic string; YPD, yeast extract peptone dextrose; AGC, automatic gain control; PPI, protein-protein interaction(s); GO -gene ontology; STRING (Search Tool for the Retrieval of Interacting; Genes/Proteins); SCB-binding factor, SBF; MCB-binding factor, MBF; SCB, Swi4/6 cell cycle box; MCB, MluI cell cycle box; SNF1, Sucrose NonFermenting 1.
strains, in which open reading frames were replaced by the kanMX cassette, were obtained from H. Bussey (McGill University, Montreal, Canada), Research Genetics and EUROSCARF. For SILAC labeling, the wt strain YAL6B (MATα his3Δ leu2Δ met15Δ ura3Δ lys1::KanMX6 arg4::KanMX4) was provided by Ole Jensen (University of Southern Odense, Denmark). To construct the cdc28-asi strain used for SILAC analysis, a mutation resulting in an F88G substitution was introduced by sequencing the entire open reading frame. The snf1-asi strain for SILAC labeling was constructed by integration of an 1NM-PP1 to a final concentration of 10 μM Tris, pH 8.0, supplemented with HALT phosphatase inhibitor mixture (Pierce). Samples were centrifuged at 40,000 × g for 10 min, and the supernatants were transferred into clean tubes prior to determination of protein concentrations by bicinchoninic acid assay (Thermo Fisher Scientific). Disulfide bonds were reduced by adding dithiothreitol to a final concentration of 5 mM, and samples were incubated at 56 °C for 30 min. Reduced cysteines were alkylated by adding iodoacetamide to 15 mM and incubating for 30 min in the dark at room temperature. Alklyation was quenched with 5 mM dithiothreitol for 15 min. Samples were diluted sixfold with 20 mM TRIS, pH 8, containing 1 mM CaCl2 prior to overnight digestion at 37 °C with trypsin (Sigma-Aldrich) using an enzyme to substrate ratio of 1:50 (w/w). Tryptic digests were acidified with 1% formic acid (FA), centrifuged (20,000 × g 10 min) and desalted on Oasis HLB cartridges (Waters) according to manufacturer instructions. Peptide eluates were snap-frozen in liquid nitrogen, lyophilized in a speedvac centrifuge, and stored at −80 °C.

**Phosphopeptide Isolation and Fractionation**—Tryptic digests were subjected to enrichment on TiO2 beads as described previously (28). Sample loading, washing, and elution steps were performed using custom StageTips (29, 30) made from 200 μl pipette tips containing an SDB-XC matrix (3 μl) and black TiO2 beads. We equilibrated TiO2 material in 250 mM lactic acid 70% acetonitrile 3% TFA; the same buffer was used for sample loading. After extensive washing steps, retained phosphopeptides were displaced from TiO2 with 500 mM phosphate buffer at pH = 7. Peptides were desalted in 50 μl of 1% FA directly on SDB-XC frits and subsequently eluted using 50 μl of 50% acetonitrile 1% FA. Eluates were dried in a speedvac and stored at −80 °C. To increase phosphoproteome coverage prior to MS analysis, phosphopeptides were fractionated offline by strong cation exchange chromatography. Peptides were solubilized in 100 μl of loading buffer (0.2% FA 10% acetonitrile) and loaded on StageTips containing 10 μg of poly-sulfoethyl-A strong cation exchange phase (5 μm 300A, Canada Life Science). Columns were washed with 50 μl of loading buffer, and peptides were eluted in five separate 100 μl salt steps of 25, 50, 80, 150, and 250 mM NaCl. All fractions (including flow-through) were collected, dried in a speedvac, and resuspended in 10 μl of 4% FA prior to MS analysis.

**Mass Spectrometry**—Enriched phosphopeptide extracts were analyzed by LC-MS/MS using a Proxeon nanoflow HPLC system coupled to a tridrib fusion mass spectrometer (Thermo Fisher Scientific). Each sample was loaded and separated on a reverse-phase analytical column (18 cm length, 150 μm inner diameter) (Jupiter C18, 3 μm, 300 Å, Phenomenex) packed manually. LC separations were performed at a flow rate of 0.6 μl/min using a linear gradient of 5–30% aqueous acetonitrile (0.2% FA) in 106 min. MS spectra were acquired with a resolution of 60,000. The TopSpeed (maximum number of sequencing events within 5 s window) method was used for data-dependent scans on the most intense ions using high energy dissociation. AGC target values for MS and MS/MS scans were set to 5e5 (max fill time 200 ms) and 5e4 (max fill time 200 ms), respectively. The precursor isolation window was set to m/z 1.6 with a high energy dissociation normalized collision energy of 25. The dynamic exclusion window was set to 30 s.

**Data Processing and Analysis**—MS data were analyzed using MaxQuant (31, 32) software version 1.3.0.3 and searched against the SwissProt subset of the S. cerevisiae uniprot database (http://www.uniprot.org) containing 6,630 entries (November 2013). A list of
248 common laboratory contaminants included in MaxQuant was also added to the database as well as reversed versions of all sequences. The enzyme specificity was set to trypsin with a maximum number of missed cleavages set to 2. Peptide identification was performed with an allowed initial precursor mass deviation up to 7 ppm and an allowed fragment mass deviation of 20 ppm with subsequent nonlinear mass recalibration (33). Phosphorylation of serine, threonine, and tyrosine residues was searched as variable modification; carbamidomethylation of cysteines was searched as a fixed modification. The FDR for peptide, protein, and site identification was set to 1% and was calculated using decoy database approach. The minimum peptide length was set to 6, and the “peptide requantification” function was enabled. The option match between runs (1-min time tolerance) was enabled to correlate identification and quantitation results across different runs. In addition to an FDR of 1% set for peptide, protein, and phosphosite identification levels, we considered only phosphosites for which localization confidence was higher than 75%. Relative quantification of the peptides against their heavy-labeled counterparts was performed with MaxQuant using area under (31, 32).

Detailed information on peptide and protein identifications is available in MaxQuant output tables provided along with the raw data at [http://www.peptideatlas.org/](http://www.peptideatlas.org/) (Identifier PASS00888; password: ZD2377ro). Annotated MS/MS spectra can be accessed through MS-Viewer (34) ([http://msviewer.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msviewer](http://msviewer.ucsf.edu/prospector/cgi-bin/msform.cgi?form=msviewer)) with the following search keys: noj5x15byw (Cdc28 inhibition time course), hqgdudq8g (nutrient shift time course), teobkhfofbl (Cdc28 15-min inhibition), and jri18xoti1 (Snf1 15-min inhibition).

**Machine Learning**—We devised an algorithm to allow the learning of a molecular recognition model despite the lack of validated substrates. We let \( x \) be a phosphosite and \( y \) be a binary label equal to 1 if \( x \) is a direct kinase target (positive example) and 0 otherwise (negative example). Additionally, let \( s = 1 \) if \( x \) is a known direct target (or not), and \( s = 0 \) if it is unclear. From our experimental strategy, only negative examples are bona fide, meaning that \( y = 0 \) is certain when \( s = 1 \), but \( y \) can either be 0 or 1 when \( s = 0 \). Formally, we have that \( P(y = 1|x, y = 1) = 0 \) and we want to learn a model function \( f(x) \) that approximates \( P(y = 1|x) \) as closely as possible. This learning paradigm is known as negative and unlabeled learning; its twin problem, positive and unlabeled learning, has been studied previously (27). The data generating distribution, \( f(x) \) can be derived from \( g(x) = P(s = 1|x) \). This implies that it is sufficient to learn a model that discriminates down-regulated Cdc28-related phosphosites from the negative set to derive a model that can discriminate direct Cdc28 phosphosites. The binary learning set for \( g(x) \) was thus composed of up-regulated phosphosites (\( s = 1 \), certainly negative) and down-regulated Cdc28-related phosphosites (\( s = 0 \), unclear if positive or negative). To approximate \( g(x) \), we choose the SVM (21) for its ability to produce probabilistic outputs, its effectiveness at handling noisy data, and its compatibility with kernels. In proteomics, examples are in the form of peptide sequences while virtually all learning algorithms need examples in the form of real valued feature vectors. String kernels have been specifically designed to address this limitation (35). Such kernels can implicitly map strings of amino acids to an arbitrary feature vector to allow the computation of the dot product between any two strings. Kernel methods, such as the SVM, rely solely on the dot product, in contrast with the actual vectors. This approach is well suited for learning models from peptides and other biomolecules. In this study, peptides were represented using a fixed window of 15 amino acids centered on the phosphorylated residue. Residues with overlapping windows were considered separately. In addition, the generic string (GS) kernel (36) was chosen as kernel for the amino acids windows because it can account for the properties of amino acids and the flexibility of the substrates. In this study, the Blocks Substitution Matrix was used to compare amino acids in the GS kernel. The GS kernel has been shown to produce models with improved accuracies for a variety of peptide-related learning problems. In particular, this kernel has produced models that are more complex and accurate than position-specific weight matrix and motifs (37). However, this improved accuracy comes at the cost of producing models that are very hard to manually interpret. 10-fold cross-validation and the F1-score (harmonic mean of precision and recall) were used to determine the optimal parameters for the SVM and the GS kernel. With these parameters, a model was learned using all but one phosphosite, and the probability that this left-out site was a direct substrate was predicted. This leave-one-out prediction scheme was repeated for each phosphosite. Following Lemma 1 from (27), \( f(x) = P(y = 1|x) = g(x)/c \) was obtained by applying a constant transformation to \( g(x) = P(s = 1|x) \), effectively predicting the probability that a site \( x \) is directly phosphorylated by the kinase of interest. A more detailed description of the machine learning algorithm is available as a Supplemental Note.

**NetworkKin, GO, and PPI Network Analyses**—Prediction of kinase–substrate relationships were achieved using KinomeXplorer ([http://networkkin.info](http://networkkin.info)) with default thresholding criteria (38). Gene ontology enrichment analyses were performed against whole *S. cerevisiae* proteome as background using the Database for Annotation, Visualization and Integrated Discovery bioinformatics resources (39, 40). All protein–protein interaction networks were built in STRING (41, 42). All interaction predictions were based on experimental evidences with the minimal confidence score of 0.9, which is considered as the highest confidence filter in STRING. Results were visualized in Cyto-scape (43, 44, 45).

**Data Availability**—The mass spectrometry raw data from this publication along with MaxQuant output files and fasta files have been submitted to the Peptide Atlas database ([http://www.peptideatlas.org/](http://www.peptideatlas.org/)) and assigned the identifier PASS00888 (Password: ZD2377ro).

**RESULTS**

**Inhibition of Cdc28-as Results in Phosphorylation and Dephosphorylation of a Large Number of Substrates**—We used quantitative SILAC-based phosphoproteomics to study the global effect of inhibition of an as-form of Cdc28 (Cdc28-as) upon treatment with 1NM-PP1. Cells were grown at 30 °C in media containing light (0Lys, 0Arg), medium (4Lys, 6Arg), or heavy (8Lys, 10Arg) isotopic forms of lysine and arginine then treated with 10 μM 1NM-PP1 or vehicle (DMSO) for 15 min. Cells were rapidly harvested by addition of TCA (46) and the SILAC channels for each culture combined before lysis in denaturing urea buffer ([supplemental Fig. S1a](https://www.mcponline.org/)). Extracted proteins were digested into peptides with trypsin and phosphopeptides were concentrated on TiO2 affinity resin with subsequent offline strong cation exchange fractionation and LC-MS/MS analyses on an LTQ-Orbitrap Fusion instrument ([supplemental Fig. S1b](https://www.mcponline.org/)). We performed six independent treatments and only phosphosites quantified in at least four replicates were considered for subsequent analysis ([supplemental Table S1](https://www.mcponline.org/), Fig. 1b). As a result, we obtained 1,008 down-regulated and 720 up-regulated phosphosites, corresponding to ~28% of all quantified sites (FDR<5%). The observed ratio between phosphorylated and dephosphorylated phospho-sites in our dataset was consistent with a previous study (19).
As expected, we found a relative enrichment of the CDK consensus motif among dephosphorylated phosphosites (Fig. 1b). The large fraction of up-regulated sites indicated significant effects due to 1NM-PP1 off-target inhibition and/or the occurrence indirect substrates as a result of regulation of other kinases/phosphatases by Cdc28. We thus sought to improve our experimental and data analysis workflow to filter out contributions from these nonspecific or indirect effects.

**1NM-PP1 Itself Triggers Many Phosphorylation Events Independent of Cdc28-as Activity**—To evaluate the effect of 1NM-PP1 on the phosphoproteome, we repeated the inhibitor and control treatment with a wt strain that did not contain the cdc28-as mutation (Fig. 2a). Surprisingly, we observed significant changes in the abundance of 757 phosphosites (~13% of the total quantified phosphosites) with an approximately equal number of up-regulated and down-regulated sites. This result suggests that 1NM-PP1 itself caused substantial changes in the phosphoproteome in the absence of any as-kinase allele. To understand the biological basis for the effect of 1NM-PP1 alone, we extracted sequence motifs from affected phosphosites. Up-regulated phosphosites did not share any obvious motifs, whereas down-regulated sites were enriched for basophilic kinase motifs containing basic amino acids (R/K) in the -2/-3 positions (Fig. 2b). We attempted to link these dephosphorylated sites with putative kinases using KinomeXplorer (38), a platform that integrates an improved NetworKIN (47) algorithm for modeling kinase signaling networks by combining linear motif analysis with protein–protein interactions data. We found that the majority of dephosphorylated sites were associated with PKA activity (Fig. 2b). This result suggested that 1NM-PP1 either directly inhibits PKA catalytic subunits or decreases PKA activity through an indirect mechanism. Notably, each PKA catalytic subunit in yeast (Tpk1, Tpk2, Tpk3) contains a methionine residue in the gatekeeper position, and hence, their ATP-binding pocket cannot accommodate the bulky 1NM-PP1 moiety, suggesting that 1NM-PP1 does not directly inhibit endogenous PKA activity. It is possible that 1NM-PP1 produces a stress response that causes a subsequent decrease in PKA activity (26, 48, 49).

Consistent with this possibility, gene ontology analysis of the phosphosites altered in response to 1NM-PP1 revealed an enrichment for proteins involved in the response to osmotic stress (such as Sln1, Ssk1, Fps1) along with amino acid transporters (Avt1/3/4, Tat1) that might help to remove 1NM-PP1 from the cell (supplemental Fig. S2). Regardless of the precise mechanism, the significant off-target effect of the inhibitor required reinterpretation of the as-inhibition data.

**Normalization of 1NM-PP1 Off-Target Phosphorylation Allows Definition of Cdc28-Dependent Phosphorylation**—As a next step, we compared inhibition of both cdc28-as and wt strains to select phosphosites that were significantly (FDR<5%) regulated in as- compared with wt cells. This subtractive approach yielded a list of phosphorylation sites that were definitively associated with the loss of Cdc28 activity (Fig. 3a). In this refined dataset, the majority of the affected phosphosites were dephosphorylated (523 down-regulated sites versus 207 up-regulated sites), as expected for bona fide Cdc28 substrates. Importantly, removal of the nonspecific 1NM-PP1-dependent events lowered the total number of regulated sites (and complexity of the dataset) by ~2.5-fold compared with the initial dataset. This normalization also substantially increased the enrichment CDK consensus motifs and known Cdc28 substrates reported in the PhosphoGRID database (http://www.phosphogrid.org/) (50, 51) among the dephosphorylated targets (Figs. 3b and 3c). However, the observed increase in phosphorylation on a substantial number of phosphosites also indicated the presence of indirect Cdc28 substrates in the dataset. As such indirect substrates might become either phosphorylated or dephosphorylated on Cdc28 inhibition, we inferred that the dephosphorylated sites were a mixture of direct and indirect substrates of Cdc28.

**Machine Learning Separates Direct and Indirect Cdc28-Dependent Phosphorylation Events**—As it was not possible to design a purely experimental strategy to filter out indirect substrates from actual down-regulated Cdc28-regulated targets, we used an SVM approach to computationally partition phosphosites into direct and indirect classes. In our setting, a typical supervised learning problem consists of learning a
PKA (Tpk1, Tpk2, and Tpk3). (non-) strain was treated with 10 μM 1NM-PP1 for 15 min and compared with DMSO control treatment. Significant changes were observed on ~13% of detected phosphosites. (b) MotifX analysis revealed basophilic kinase motifs among the dephosphorylated sites in a 1NM-PP1 treated wt strain. NetworKIN analysis also suggested that ~60% of dephosphorylated sites were predicted substrates of the basophilic kinase PKA (Tpk1, Tpk2, and Tpk3).

Fig. 2. 1NM-PP1 causes off-target phosphorylation events. (a) In order to estimate the effect of 1NM-PP1 treatment alone a wild type (non-as) strain was treated with 10 μM 1NM-PP1 for 15 min and compared with DMSO control treatment. Significant changes were observed on ~13% of detected phosphosites. (b) MotifX analysis revealed basophilic kinase motifs among the dephosphorylated sites in a 1NM-PP1 treated wt strain. NetworKIN analysis also suggested that ~60% of dephosphorylated sites were predicted substrates of the basophilic kinase PKA (Tpk1, Tpk2, and Tpk3).

binary classification model from a set of validated substrates and negative controls. Such a dataset would allow the learning of a statistical model capable of differentiating direct from indirect substrates. However, de novo studies like ours do not contain validated training examples needed to learn a conventional model. We circumvent this problem by using a binary learning approach similar to positive and unlabeled learning (22).

To filter out indirect substrates from down-regulated by Cdc28 phosphosites, we used a mixture of direct and indirect Cdc28 targets (dephosphorylated phosphosites) as potential positive examples along with negative controls (up-regulated phosphosites across all experiments) to build a model (see Experimental Procedures for details). Accuracy was estimated by 10-fold cross-validation on the negative control group (Fig. 4a) with a value of 93.2% that suggests a good separation between indirect and direct Cdc28 substrates. Notably, despite the model having no prior information on the CDK consensus sequence and known Cdc28 targets (inferred from www.phosphogrid.org), phosphosites predicted as direct substrates were enriched for both of these independent criteria. To further confirm that the machine learning approach made biologically relevant predictions, we used the Yeast Kinase Interaction Database (52) (KID, http://www.moseslab.csb.utoronto.ca/ Kid/), which contains high- and low-throughput data relevant to Cdc28. To independently benchmark our results against other experimental studies, we extracted all Cdc28 interacting partners from the KID database and computed the overlap with our data. This analysis revealed a much better correlation between KID and our direct Cdc28 target set compared with the indirect target set (Fig. 4b). We stress that our model did not use any prior information from KID or any other previous datasets.

Among 260 phosphosites predicted to be directly phosphorylated by Cdc28, we found a number of known substrates. Notable examples included components of the SBF and MBF G1/S transcription complexes (supplemental Fig. S3) and a phosphorylation site on Ser429 of the phosphatase Cdc14 involved in feedback regulation of Cdc28 activity (53). Our data also revealed some Cdc28 targets implicated in cytoskeletal dynamics (supplemental Fig. S4). Interestingly, on many Cdc28 targets, we identified multiple phosphosites that corresponded to the direct and indirect classes. These substrates may represent examples of complex regulation whereby a substrate is first directly phosphorylated by Cdc28, followed by indirect phosphorylation events mediated by downstream kinases/phosphatases (supplemental Fig. S4a). In this regard, we note that inhibition of Cdc28 triggers changes in the phosphorylation status of 33 kinases and phosphatases, which may explain the dramatic contribution (~50%) of indirect phosphorylation events to our dataset (supplemental Fig. S5).

Direct Cdc28 Sites Are Dephosphorylated Faster Than Indirect Sites—We investigated possible correlations between experimentally measured parameters and the probability of a phosphosite being a direct Cdc28 substrate. We found a dependence between the score generated by the model, that is, the probability that the phosphosite is direct, and the observed fold change in phosphosite abundance between control and inhibitor-treated conditions (Pearson r = 0.49, Fig.
Direct Cdc28 sites were thus significantly more downregulated compared with indirect sites (Wilcoxon test p < 10^{-8}, Fig. 5b). It is important to highlight that no quantitative information was used to train the model (see Supplemental Note for additional information on the machine learning workflow).

Although multiple parameters underlie the kinetics of phosphorylation/dephosphorylation (54, 55), we tested the simple hypothesis that direct substrates might be dephosphorylated more rapidly compared with indirect targets, which would result in higher absolute fold change values upon kinase inhibition. To test this notion, we performed kinetic experiments by determining global phosphorylation changes every 1 min following 1NM-PP1 treatment for a total duration of 15 min (Fig. 5c and supplemental Table S1). We obtained a 95% overlap between this dynamic dataset and our original single time point datasets. We assigned high-quality kinetic profiles, defined as sites being quantified in at least 14 of 16 measurements across the 15-min time series, for a majority of the Cdc28-associated direct and indirect phosphosites. We used this kinetic data to plot averaged profiles for dephosphorylated direct (n = 237) and indirect (n = 213) Cdc28 targets (Fig. 5c). These results confirmed the predicted faster dynamics of dephosphorylation for direct compared with indirect targets. We did not observe an obvious temporal delay between upregulated phosphorylation events, in agreement with previous high-resolution dynamic studies (26, 56, 57).

Identification of Direct Snf1 Targets—To test our approach on a less well-characterized kinase, we applied the same workflow to study the Snf1 kinase, which regulates cell growth and metabolism under nutrient-limiting conditions. Snf1 is the founding member of the SNF1/AMPK family of protein kinases, which is highly conserved in eukaryotes and required for energy homeostasis in mammals, plants, and fungi (58). In S. cerevisiae, Snf1 allows cells to respond to declining levels of glucose by promoting respiratory metabolism, glycogen accumulation, gluconeogenesis, autophagy, the glyoxylate cycle, and peroxisome biogenesis (59, 60, 61, 62).

We constructed strain bearing an snf1-as allele and repeated inhibition with 1NM-PP1 under the same conditions as for the cdc28-as experiment (10 μM 1NM-PP1 for 15 min). To maintain the activity of Snf1, we used culture medium with ethanol as a carbon source. In total, this experiment yielded 12840 -4 -2 0 2 4 1NM-PP1/Control, [log2] pValue, [-log10] FDR<5%

### FIG. 3. Cdc28-specific phosphorylation sites.

(a) Subtraction of phosphosites due to 1NM-PP1 alone from a cdc28-as inhibition experiment revealed targets whose abundance was affected by the change in Cdc28 activity. In total ~12% of all identified phosphosites were regulated by Cdc28 with strong preference for dephosphorylation (523 dephosphorylated versus 207 phosphorylated sites). The increase in phosphorylation on 207 substrates indicates existence of indirect downstream effectors of Cdc28 inhibition. (b) Removal of 1NM-PP1 off-target phosphorylation events (1NM-PP1) from the Cdc28as results (global) improved the quality of the dataset as indicated by dramatic increase of the ratio of known Cdc28 substrates and sites with a CDK consensus motif between down-regulated and up-regulated peptides (Cdc28). (c) NetworKIN analysis revealed enrichment of CDK substrates among dephosphorylated phosphosites.
408 down-regulated and 442 up-regulated phosphosites (supplemental Table S2). In this Snf1 experiment, we noticed a much smaller off-target effect of 1NM-PP1 as only 3% of the quantified sites were affected by 1NM-PP1 alone compared with 13% in the Cdc28 experiment. This difference may be explained by the use of ethanol versus glucose as a carbon source since the major effect of 1NM-PP1 in glucose medium appears to be mediated by an indirect decrease in PKA activity (Fig. 2b). In the absence of glucose, PKA activity would be significantly reduced, and therefore, an additional inhibition with 1NM-PP1 alone would produce a smaller relative change in the phosphoproteome. As for Cdc28, correction for the non-specific effects of 1NM-PP1 on the phosphoproteome allowed us to identify Snf1-specific events (Fig. 6a) that were represented by both phosphorylation (73 sites) and dephosphorylation (225 sites) events. Machine learning was used to separate between direct (134) and indirect (78) Snf1 substrates among the dephosphorylated sites with an estimated cross-validation accuracy of 93%. Notably, we found the same correlation between the fold change decrease in phosphorylation and the probability generated by the model as for Cdc28 (Pearson $r = 0.54$, supplemental Figs. S6a and S6b). Snf1 direct substrates were also dephosphorylated more rapidly than indirect substrates. We were not able to find a consensus linear motif among the indirect Snf1 substrates, whereas for direct Snf1 targets we identified a motif in which the phosphorylated residue was followed by a leucine residue at the $+4$ position. Gene ontology analysis of the predicted direct substrates revealed an enrichment for proteins implicated in sugar metabolism and growth regulation (Fig. 6b), consistent with the known role of Snf1. Finally, we detected phosphorylation on many proteins previously associated with Snf1-mediated signaling events (Fig. 6c). Of particular note, we identified direct phosphorylation sites on the well-known Snf1 targets Mig1 (63, 64), Hxk1/2 (65, 66), and Cat8 (67, 68), including previously undocumented sites. We also observed
Extracting Direct Kinase Substrates from Phosphoproteomic Data

Fig. 5. Predicted direct Cdc28 substrates are dephosphorylated more rapidly than predicted indirect substrates. (a) Correlation between extent of site dephosphorylation and SVM score. (b) Extent of dephosphorylation was larger for direct versus indirect substrates (pValue = 1e-8, Wilcoxon test). (c) Kinetic measurements of phosphorylation changes upon Cdc28 inhibition over a 15-min time course at 1 min intervals revealed differential dephosphorylation dynamics between direct versus indirect Cdc28 targets.

Phosphorylation on S241 of adenylate cyclase, which has recently been reported to be a direct Snf1 substrate (69, 70). We found both direct (Sip1) and indirect (Sip1, Gal83, and Sip2) phosphorylation on all β-subunits of SNF1 complex, which play a role in determining Snf1 substrate specificity and subcellular localization (24). In an interesting new regulatory connection, we identified Ser826, Ser75, and Ser925 as direct Snf1 sites on the Reg1 regulatory subunit of the Glc7 (PP1) phosphatase that inactivates Snf1, suggesting a potential feedback regulatory mechanism. Finally, we identified a new direct autophosphorylation site on Ser391 of Snf1 itself, suggestive of yet another regulatory mechanism.

Physiological Regulation of Direct Snf1 Sites—To demonstrate that the Snf1 sites were also affected under physiological conditions, we undertook a large-scale phosphoproteomic study of a nutrient shift that would alter endogenous Snf1 activity. A wt non-as strain was grown in the presence of ethanol as a carbon source and then transferred to the glucose-containing medium to cause the physiological inactivation of Snf1. Changes in the phosphoproteome were monitored every 2 min over a 30-min time course. In total, we were able to monitor the dynamic changes on ~4,000 phosphosites with a minimum representation of nine measurements across the 16-point time series (supplemental Table S3). In this physiological nutrient shift experiment, we observed changes on 42 phosphosites out of the 134 that were identified as direct Snf1 substrates. All of these sites exhibited dephosphorylation upon the nutrient shift, consistent with the loss of Snf1 activity. We also observed dynamic changes for 30 phosphosites out of 78 that were classified as indirect Snf1 substrates. When we averaged normalized kinetic profiles for direct and indirect sites in the nutrient response, we observed the same trend as for Cdc28 inhibition, namely that dephosphorylation of direct targets was faster than on indirect direct targets (supplemental Fig. S6b). This result in the context of a physiological response provided support for the machine-learning-based classification of phosphorylation sites.

Mutational Analysis of Snf1 Phosphosites—We evaluated the functional importance of a number of both direct and indirect Snf1 phosphorylation sites by mutational analysis. Nonphosphorylatable alanine mutations were introduced into the endogenous loci for a total of 16 phosphosites in 8 Snf1 substrates within the immediate Snf1 signaling network. We assessed the effect of either single or combined mutations on culture doubling time in log phase, lag time for re-entry into log phase from stationary phase, and cell size on glucose, galactose, or glycerol medium. Strikingly, we were able to detect one or more growth-related phenotypes for each phosphosite mutant generated, suggesting that both direct and indirect phosphorylation sites contribute to the function of the Snf1 signaling network (supplemental Table S4). For example, mutation of phosphosites on each of the β-subunits of the SNF1 complex (Gal83, Sip1, and Sip2) caused phenotypes of differing severities. Gal83 is the most abundant β-subunit and targets SNF1 into the nucleus where it activates transcription factors Sip4 and Cat8 and deactivates Mig1. Mutation of the S44 site on Gal83 caused a decrease in growth rate and cell size on glycerol medium when compared with the wt control strain but without any difference when the strains were grown on glucose medium (supplemental Figs. S7a and S7b). This result illustrates the functional importance of an indirect Snf1 phosphosite. Mutation of other mapped sites on the Sip1 and Sip2 β-subunits caused a more modest effect on the lag time for log phase re-entry without a clear effect on doubling time or cell size (supplemental Table S4). In another example, mutation of the Ser925 and Ser96/Ser75 direct phosphosites on Reg1 both caused a significant delay in growth rate and a smaller cell size on glycerol medium. Interestingly, these phenotypes on the nonfermentable carbon source were even more severe than in the corresponding reg1Δ deletion strain (supplemental Table S4). A particularly notable direct phosphosite was mapped to S391 on Snf1 itself. Although mutation of this previously unknown site did not cause an obvious phenotype in isolation, the S391A mutation partially suppressed the loss of Snf4, which is an essential activating subunit of the SNF1 kinase complex (supplemental Fig. S7c). The SNF1(S391A) mutation thus bypassed the pronounced size defect of an snf4Δ deletion strain on both glucose and glycerol medium. In con-
**FIG. 6. Identification of direct Snf1 substrates.** (a) Application of global phosphorylation profiles and machine learning to identify direct targets of the Snf1 kinase. After removal of 1NM-PP1 inhibitor off-target phosphorylation events, significant regulation was observed for ~5% of all detected sites with a strong prevalence of dephosphorylation. SVM model allowed separation of direct (134) and indirect (78) Snf1 targets among all dephosphorylated targets. (b) MotifX analysis revealed enrichment of a sequence motif where the phosphorylated residue was followed by leucine in the +4 position for direct Snf1 substrates. No significantly enriched motifs were obtained in a similar analysis of indirect Snf1 targets. Gene ontology analysis revealed enrichment of terms associated with sugar metabolism and regulation of cell growth. (c) Many known components of the local Snf1 signaling network were identified as either direct or indirect targets.
trast, an snf1Δ snf4Δ double mutant grows as poorly on a non-fermentable carbon source as either single deletion (71). From these results, we postulate that the SnfT(S391A) allele encodes a hyperactive form of the kinase that is partially Snf4 independent and that autophosphorylation of Ser391 might be a feedback mechanism to limit Snf1 activity.

**Discussion**

Our study addresses a straightforward but important biological question, namely, why does selective inhibition of a particular kinase result in both an increase and decrease in phosphorylation on a large number of proteins? To explain this phenomenon, we show that the global phosphorylation response to selective kinase inhibitors arises from three distinct contributions: non-specific off-target effects of the inhibitor, direct substrates of the kinase, and indirect signaling events that arise downstream of the primary kinase. We developed a new method to disentangle these different contributions and performed a benchmark analysis on an as allele of the Cdc28 cell cycle regulatory kinase in S. cerevisiae. This analysis revealed an unexpectedly large effect of the 1NMP1 inhibitor on the phosphoproteome. Correction for these off-target events then allowed us to obtain a set of Cdc28-specific phosphorylation sites. Many of these sites were inferred to be indirect consequences of Cdc28 inhibition because a significant number exhibited increased phosphorylation upon Cdc28 inhibition, indicative of downstream signaling effects. To deconvolute direct from indirect phosphorylation sites, we developed a machine learning model with strong discriminatory power, as assessed against independent datasets. Consistently, we demonstrated higher experimental fold changes and faster dephosphorylation kinetics for direct substrates compared with indirect substrates. Finally, we successfully applied our workflow to identify direct substrates of the less-well-characterized kinase Snf1 in S. cerevisiae. We identified many new direct and indirect phosphorylation sites on Snf1 substrates, for which a significant fraction of the direct sites exhibited faster dephosphorylation during a physiological nutrient shift that inactivates Snf1. Mutational analysis of a number of both direct and indirect Snf1 sites revealed new regulatory features of the Snf1 signaling network, including a potential new autoinhibition mechanism on Snf1 itself. This combined methodology based on specific kinase inhibition, appropriately controlled global phosphoproteome profiles, and machine learning may in principle be applied to any kinase for which an analog sensitive form can be engineered. The identification of direct in vivo kinase substrates in an unbiased manner will provide the foundation for deconvolution of the labyrinthine phosphoregulatory architecture of the cell.

**Acknowledgments**—The Institute for Research in Immunology and Cancer receives infrastructure support from the Canadian Center of Excellence in Commercialization and Research, the Canadian Foundation for Innovation, and the Fonds de Recherche du Quebec-Sante (FRQS). Proteomics analyses were performed at the Center for Advanced Proteomic Analyses, a Node of the Canadian Genomic Innovation Network supported by the Canadian Government through Genome Canada.

* This work was supported by Canada Research Chairs in Proteomics and bioanalytical mass spectrometry (PT) and in Systems and Synthetic Biology (MT) and by Genome Canada Bioinformatics grant (MT, PT), the National Science and Engineering Research Council (311598, PT); the Canadian Institutes for Health Research (MOP 126129 to MT, and MOP 366608 to MT and PT), the National Institutes of Health (R01RR024031 to MT), the Ministère de l’enseignement supérieur, de la recherche, de la science et de la technologie du Québec through Génome Québec (MT), and the Quebec Consortium for Drug Discovery (MT).

[1] To whom correspondence should be addressed: Tel.: 514-343-6910, Fax: 514-343-6843. E-mail: pierre.thibault@umontreal.ca; Tel.: 514 343-6688 and E-mail: md.tyers@umontreal.ca.

**References**


64. Karunanithi, S., and Cullen, P. J. (2012) The filamentous growth MAPK pathway responds to glucose starvation through the Mig1/2 transcriptional repressors in *Saccharomyces cerevisiae*. *Genetics* 192, 869–887


