Bioinformatics, 32(3), 2016, 424–431
doi: 10.1093/bioinformatics/btv699
Advance Access Publication Date: 1 December 2015
Original Paper

OXFORD

Systems biology

IKAP: A heuristic framework for inference of kinase activities from Phosphoproteomics data

Marcel Mischnik[1,*], Francesca Sacco[2], Jürgen Cox[2], Hans-Christoph Schneider[1], Matthias Schäfer[1], Manfred Hendlich[1], Daniel Crowther[1], Matthias Mann[2] and Thomas Klabunde[1]

[1]Sanofi-Aventis Deutschland GmbH, Frankfurt, Germany and [2]Department of Proteomics and Signal Transduction, Max-Planck-Institute for Biochemistry, Martinsried, Germany

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on 23 April 2015; revised on 28 October 2015; accepted on 25 November 2015

Abstract

Motivation: Phosphoproteomics measurements are widely applied in cellular biology to detect changes in signalling dynamics. However, due to the inherent complexity of phosphorylation patterns and the lack of knowledge on how phosphorylations are related to functions, it is often not possible to directly deduce protein activities from those measurements. Here, we present a heuristic machine learning algorithm that infers the activities of kinases from Phosphoproteomics data using kinase–target information from the PhosphoSitePlus database. By comparing the estimated kinase activity profiles to the measured phosphosite profiles, it is furthermore possible to derive the kinases that are most likely to phosphorylate the respective phosphosite.

Results: We apply our approach to published datasets of the human cell cycle generated from HeLaS3 cells, and insulin signalling dynamics in mouse hepatocytes. In the first case, we estimate the activities of 118 at six cell cycle stages and derive 94 new kinase–phosphosite links that can be validated through either database or motif information. In the second case, the activities of 143 kinases at eight time points are estimated and 49 new kinase–target links are derived.

Availability and implementation: The algorithm is implemented in Matlab and be downloaded from github. It makes use of the Optimization and Statistics toolboxes. https://github.com/marcel-mischnik/IKAP.git.

Contact: marcel.mischnik@gmail.com

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

1 Introduction

Phosphoproteomics is a branch of proteomics that identifies, catalogues and characterizes proteins containing a phosphate group as a post-translational modification. Since intracellular signal transduction is primarily mediated by the reversible phosphorylation of various signalling molecules, data generated by phosphoproteomics approaches is commonly used to study the dynamics of signalling networks (Kirkpatrick *et al.*, 2013; Xia *et al.*, 2013; Zimman *et al.*, 2014). A major challenge in the analysis of such data lies in the high combinatorial complexity of the observed signalling networks. The human genome encodes roughly 540 kinases that phosphorylate 20 000 proteins at probably more than 350 000 phosphosites (Hornbeck *et al.*, 2012; Manning, 2005). Hence, the number of possible targets per kinase tends to be high, and the majority of proteins acting as kinase–targets can be phosphorylated at multiple sites, where each site can potentially be targeted by several kinases, resulting in an intricate network of kinases, target proteins and phosphatases, which remove the attached phosphorylations again. Both kinases and phosphatases can thereby themselves be phosphorylated. The situation gets further complicated by the fact that most known phosphosites can so far not be related to a distinct protein function. Thus, most research up to the present focuses on individual phosphoproteins and kinases or is restricted to clustering and enrichment methods which group the detected sites according to their behaviour and relate the proteins they are localized on to cellular processes like canonical pathways, transcriptional networks or disease mechanisms via statistical testing against databases (Liu and Chance, 2014). Modelling of phosphoproteomics networks is usually limited to few components (<50) and rather includes single phosphosites than whole proteins due to the lack of functional information (Terfve *et al.*, 2012).

To address these challenges, we developed the machine algorithm IKAP (Inference of kinase activities from phosphoproteomics) that estimates the activities of all kinases that are known to phosphorylate at least one phosphosite in a phosphoproteomics dataset. IKAP takes information from the PhosphoSitePlus (PSP) database as input and uses a non-linear optimization routine to minimize a cost function that relates kinase activities and affinities to phosphosite measurements. The predicted activities directly display the engagement of signal transduction pathways such as cAMP/PKA, $Ca^{2+/}$PKC, MTOR or Akt kinase pathways. If data for multiple time points or conditions are available, each kinase receives a unique activity profile which depicts the change of cellular function over all conditions. Furthermore, these profiles can be used as an input for mathematical modelling which enables the usage of protein nodes instead of phosphosite nodes, making the model more meaningful and predictive. Another application lies in the inference of possible new kinase–target links. By correlating the calculated kinase activity profiles to the measured phosphosite profiles the kinases that are most likely to phosphorylate a respective phosphosite can be derived.

To demonstrate the capabilities of IKAP, we applied it to two previously published phosphoproteomics datasets. The first comprises a study of the human cell cycle in HeLaS3 cells (Olsen *et al.*, 2010). Eukaryotic cells replicate by a complex series of evolutionarily conserved events that tightly regulate defined stages of the cell division cycle. Progression through this cycle involves a large number of dedicated protein complexes and signalling pathways, and deregulation of this process is implicated in tumourigenesis. Global analyses of transcriptome dynamics during the cell cycle have been performed in microarray studies. However, in addition to changes in messenger RNA (mRNA) abundance, protein phosphorylation and targeted protein degradation are also important regulators of cell cycle progression. To examine the influence of protein phosphorylations on the progression of the cell cycle, the phosphoproteome was measured through a combination of immobilized metal affinity chromatography (Macek *et al.*, 2009) and stable isotope labelling of amino acids (Ong *et al.*, 2002) at the six phases G1, G1-S, early-S, late-S, G2 and M. Further details can be found in Olsen et al. (2010).

The second dataset is a study of insulin signalling dynamics in the mouse liver (Humphrey *et al.*, 2015). Here, the phosphoproteome was measured after 0, 0.5, 1, 2, 3, 4, 6 and 10 min after stimulation with insulin.

2 Methods

The algorithm consists of a set of Matlab functions that have to be sequentially executed. The package including explanations for use can be found in the Supplementary Material and in our code repository at github. All scripts were written in Matlab R2014b, but were tested on earlier versions down to R2012a. The algorithm makes use of functions from the Optimization and Statistics toolboxes for parameter estimation and statistical analysis. It is divided into three parts, where in the first part kinase activities and affinity parameters are estimated from the measured data using known interactions, in the second the calculated kinase profiles are used to infer possible new kinase–target links, which are validated through database, motif and literature information in the third part. All experimental methods can be found in Olsen *et al.* (2010).

2.1 Part 1: estimation of kinase activities

The algorithm assumes as an input a data sheet consisting of protein or gene names, the sequences of the measured peptides and t columns with data values.

1. The first function performs a PSP database search for all sequences in the dataset and outputs the kinases that are known to phosphorylate each sequence in a column next to the data values. The PSP database (Kinase–substrate dataset) can be downloaded from http://www.phosphosite.org/staticDownloads.do.

2. In the next step, the kinases found can be filtered by proteomics data to ensure only those kinases that are present in the cell type under investigation are included. This step is optional. Afterwards, the original dataset is filtered for those phosphosites that have at least one annotated kinase in PSP which is present in the proteome. If no proteome data is available, the filtering is done with respect to all kinases from the first step. We end up with a reduced dataset *data_red* and the string array *kin* that contains the kinase names.

3. The third step generates a truth table for the kinase–phosphosite interactions in the reduced dataset. It produces the $n \times m$ matrix A with $A(i,j) \in \{0,1\}$, n being the number of phosphosites in the reduced dataset and m the number of kinases in *kin*.

4. Now that we have produced *data_red*, *kin* and A, we can use these variables to perform the optimization task of finding the kinase activities which best explain our data. In the most simple case, the activity k of kinase l at time point i can be defined as the product of the kinase's abundance a, which is mostly known by proteomics data, and its mobilization m, which is dependent on post-translational regulation.

$$k_{li} = a_{li} * m_{li}. \tag{1}$$

The effect e a certain kinase has on a phosphosite j can in turn be regarded as the product of its activity and the affinity p the kinase has to the respective phosphosite.

$$e_{jli} = k_{lji} * p_{jl}. \tag{2}$$

Thus, we can define the value of phosphosite j as the mean of all effects acting on it. Since the affinities of various kinases to phosphosite j are different and both activities and affinities would become structurally non-identifiable if calculated for one single condition, the mean is computed by dividing the summed effects by the sum of affinities instead of the number of kinases. Hence, optimization is done by minimizing the cost function

$$C = \sum_{i=1}^{t} \sum_{j=1}^{n} \left(\frac{1}{\sum p} \sum_{l=1}^{m} (K_{lji} \ * p_{jl}) - P_{ji} \right)^2 \tag{3}$$

with t being the number of time points measured, n the number of phosphosites in the reduced dataset, m the number of kinases that phosphorylate site j, K the activity of kinase l for j at time point i, p the affinity of kinase l for site j and P the measured value of site j at time point i. Figure 1 gives an exemplified overview of how

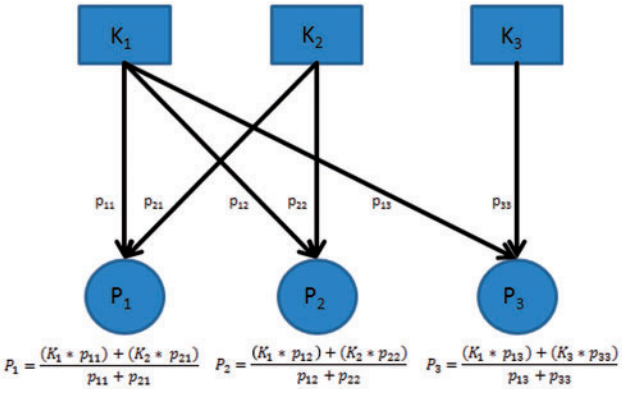
K_1
K_2
K_3
p_{11}
p_{21}
p_{12}
p_{22}
p_{13}
p_{33}
P_1
P_2
P_3
$P_1 = \frac{(K_1 * p_{11}) + (K_2 * p_{21})}{p_{11} + p_{21}}$
$P_2 = \frac{(K_1 * p_{12}) + (K_2 * p_{22})}{p_{12} + p_{22}}$
$P_3 = \frac{(K_1 * p_{13}) + (K_3 * p_{33})}{p_{13} + p_{33}}$

Fig. 1. Exemplified model structure. At a given time point, each phosphosite *P* is modelled as the sum of the products of the assigned kinase activities *K* and the respective affinity parameters *p*, divided by the sum of the affinity parameters

phosphosites are modelled at a given time point. The affinity parameters are optimized globally with respect to all measured time points, whereas the kinase activities are estimated for each time point separately. The function makes use of the Matlab built-in function *fmincon* which is part of the Optimization toolbox and uses a trust region related routine to find a parameter set that minimizes our cost function (3) applying parameter bounds. Bounds were set to 0 and 50. Besides the model-to-data-distance of a given parameter set, Step 4 also calculates the parameters' gradients, which are then forwarded to *fmincon*. In addition to the variables *data_red, kin* and *A*, a fourth input is the desired number of iterations. Since for each iteration, the starting values for both kinase activities and affinity parameters are randomized, a higher number of iterations yields a better coverage of the parameter space and thus a better chance to find the global optimum. The outputs of Step 4 are the arrays *K* and *AP* which contain the optimal kinase activities and affinities as well as the minimal cost.

5. In this optional step, we test the identifiability of the calculated activities. This is done by exploring the χ^2 space along the dimension of the respective kinase activity. If the resulting profile exhibits a minimum, the activity is identifiable on the basis of the measured data.

2.2 Part 2: inference of kinase–target links

In this part, the calculated activity profiles are applied to infer the most likely kinases for each phosphosite by means of correlation coefficients.

Our assumption is that the phosphorylation state of a site is correlated to the activity of the kinase that dominates the phosphorylation of this site. In case of one dominating kinase for a site the phosphorylation state of the site and the activity of the kinase result in a high correlation coefficient. A high and significant correlation therefore points to a kinase–substrate link. In case of sites that are phosphorylated by several different kinases, the phosphosphorylation time profile is the sum of the activity profiles of these kinases. If the contributions of the kinases to the phosphorylation of the sites are in a similar range a correlation will be less obvious. It should be pointed out, however, that the sum of the activity profiles of one set of kinases might mimic the activity profile of other kinases. Potentially, this could lead to the detection of false positive kinase–substrate links.

1. Using the function *ComputeDistances*, we calculated *P*-values for all kinase–phosphosite pairs in the complete dataset by computing the correlation coefficient between the estimated kinase profiles and the measured phosphosite values. The function generates the $n \times m$ matrix *dist*. Each entry represents a *P*-value for the respective kinase–target interaction indicating the probability of getting a correlation as large as the observed value by random chance, when the true correlation is zero. Since not every phosphosite has a valid measurement value at all applied conditions, the function also outputs a vector containing the number of valid values for all phosphosites. This can be important when comparing the different *P*-values to each other.

2. The second step generates the cell array *psig*, which includes all significant kinase–phosphosite links and their respective *P*-values, taking the newly produced distance matrix *dist*, the kinase list *kin* and the desired false discovery rate as an input.

2.3 Part 3: validation functions

Now that we have derived the most likely kinases for each phosphosite, we validate the newly found interactions by means of database information from IPA (QIAGEN Redwood City, www.qiagen.com/ingenuity) and MetaBase (Thomson Reuters, http://thomsonreuters.com/metabase) as well as motif information from NetworKIN (Horn *et al.*, 2014). Since the latter is based on the STRING database, this also covers publicly available information at this point. Before we start to validate our links, Part 2 should be repeated with a dataset from which the pairs in the reduced dataset are removed to guarantee for an unbiased validation. This can be done by the optional function *MakeDataVal*, which produces *data_val*.

1. This step produces a list *pnsig* which has the same size as *psig* and contains randomly taken kinase–phosphosite links for comparison.

2. To validate our significant links, we established a list of all kinase–target links which are present in the IPA and MetaBase databases. Those links are not phosphosite-specific but rather limited to interactions of kinases with entire proteins. However, if a certain kinase–protein link can be found in the databases and a phosphosite located on that protein has a low IKAP *P*-value with respect to the respective kinase, the interaction is likely to be realized within the investigated cell type. The function compares the database list with both *psig* and *pnsig*, counts the agreements and calculates a *P*-value by means of a fisher exact test. Since *pnsig* has been generated randomly, Steps 1 and 2 should be repeated several times with taking the mean of all obtained *P*-values.

3. The second validation consists in checking the likelihood of the found interactions in a motif-based manner via NetworKIN (Horn *et al.*, 2014), which is an integrated platform for modelling kinase signalling networks by combining sequence specificity with cellular context from the STRING database (Franceschini *et al.*, 2012). Thus, here we replenish our database validation from Step 4 by publicly available resources. This is again done for both *psig* and *pnsig*. In this case, the algorithm calculates two different *P*-values. The first is based on the summed likelihoods of all links in *psig* and *pnsig*, respectively, and the second is computed on the number of significant likelihoods (≥1) in both lists. Both *P*-values are derived by a fisher exact test.

3 Results

3.1 Calculation of kinase activities and most likely pairs

We started our analysis with the HeLaS3 phosphoproteome data (Supplementary Material, original data mit.xls) taken from Olsen *et al.* (2010) and the mouse liver data taken from Humphrey *et al.* (2015) (Supplementary Material, data-mit.mat).

In the case of HeLaS3 data, where proteomics measurements were available, the measured phosphopeptide values are normalized with respect to the corresponding changes in protein abundance to determine changes in phosphorylation state rather than changes in protein abundance. Here, all values are given as $\log_2$ and lie within the interval $[-15,15]$. The data contains 24 714 phosphosites in total which were measured at the six cell cycle stages G1, G1/S, early S, late S, G2 and M. Figure 2 depicts the overall workflow.

By applying IKAP Part 1, we at first obtained a reduced dataset comprising 1069 phosphosites for which at least one kinase is known from PSP. Filtering of the obtained kinases by proteomics led to a list comprising 118 kinases which are expressed in HeLaS3 and known to phosphorylate a peptide in our dataset at the respective position. By applying Step 4 with an iteration number of 100, we estimated the activities of these 118 kinases within the six cell cycle stages. As quoted before, affinity parameters were estimated globally with respect to all six conditions, whereas kinase activities were fitted for each condition separately. Parameter bounds were set to 0.1 and 10 for affinity parameters and -15 and 15 for activities. The latter thus lie in the same range as the phosphosite measurements. To match units and for the sake of clarity, we can modify Equation (1) to:

$$k_{li} = \log_2\left(\frac{m_{li}}{a_{li}}\right). \tag{4}$$

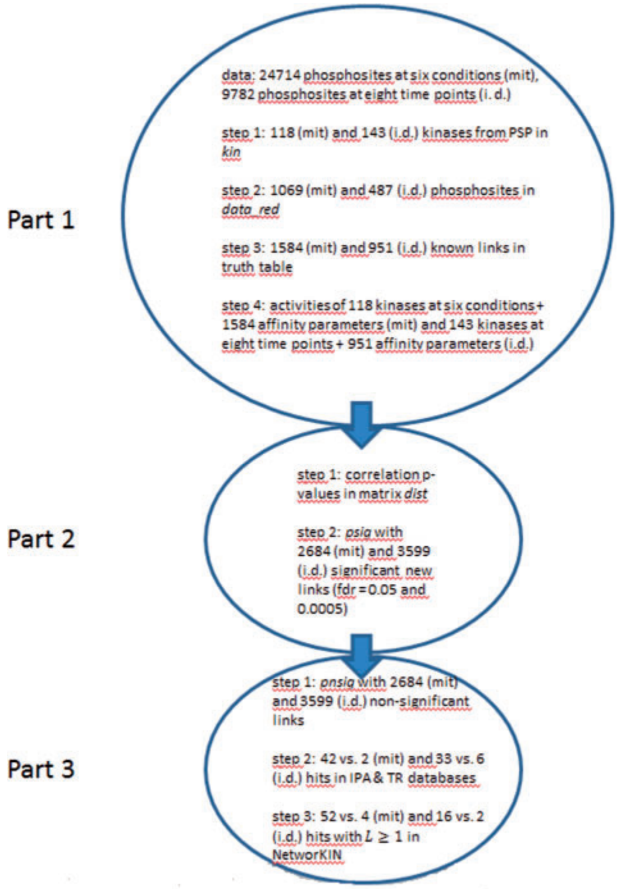
Part 1
data: 24714 phosphosites at six conditions (mit), 9782 phosphosites at eight time points (i. d.)
step 1: 118 (mit) and 143 (i.d.) kinases from PSP in *kin*
step 2: 1069 (mit) and 487 (i.d.) phosphosites in *data_red*
step 3: 1584 (mit) and 951 (i.d.) known links in truth table
step 4: activities of 118 kinases at six conditions + 1584 affinity parameters (mit) and 143 kinases at eight time points + 951 affinity parameters (i.d.)
Part 2
step 1: correlation p-values in matrix *dist*
step 2: *psig* with 2684 (mit) and 3599 (i.d.) significant new links (fdr = 0.05 and 0.0005)
Part 3
step 1: *pnsig* with 2684 (mit) and 3599 (i.d.) non-significant links
step 2: 42 vs. 2 (mit) and 33 vs. 6 (i.d.) hits in IPA & TR databases
step 3: 52 vs. 4 (mit) and 16 vs. 2 (i.d.) hits with *L* ≥ 1 in NetworKIN

Fig. 2. Workflow with quantified results for mitosis data (mit) and insulin dynamics data (i.d.). In Part 1, kinase activities and affinity parameters were estimated. Part 2 comprises the search for new kinase–target links via correlation-derived *P*-value. In Part 3, the newly found links are validated by database and motif information

The mobilization factor m is now comparable to a raw phosphosite value. However, since we estimate k, the algorithm is not affected by this change.

Fitting led to an overall χ^2/N value of 1.22 for $N = 6414$ data points (1069 phosphosites at 6 conditions) and 1777 parameters [1584 affinities and 708 kinase activities (118 kinases at 6 conditions)], using a SD of 1.5.

Table 1 shows the five most and least active kinases at each cell cycle stage. In all stages, we see at least one member of the MAP-kinase pathway among the five most up-regulated kinases. This suggests this pathway to be particularly active in HeLaS3 cells. In Figure 3, the mean kinase activity is displayed. We can see that the overall kinase activity is highest at G1 and G2, and lowest during the S and M phases, which is what we would expect (Fig. 3A).

IKAP Part 2.1 produces a distance matrix which displays *P*-values for all kinase–phosphosite pairs. Step 2.2 finally shows all significant pairs (2684 with false discovery rate = 0.05). Figure 4 exemplary depicts the profiles of CDK1, CDK4, GRK5 and NEK2, each along with one of their most likely targets. An activity value of zero indicates no post-translational regulation to be present. Here, the mobilization factor is zero. Higher activity values imply a post-translational up-regulation, whereas lower values suggest a post-translational down-regulation. The activity of CDK1 is high during mitosis but low during S phase (Fig. 4A), which is what we would expect (Morgan *et al.*, 2007; Pruitt *et al.*, 2014). CDK4 on the other hand shows its highest activity in G1 followed by a drop and a second onset in late S phase (Fig. 4B). Here, we would expect a single peak within the G1 and S phases (Morgan *et al.*, 2007; Pruitt *et al.*, 2014) indicating that cell cycle control is altered in HeLaS3 cells. The GRK5 protein phosphorylates the activated forms of G protein-coupled receptors thus initiating their deactivation. Its activity should be highest within the G1 phase and lowest during the S-phase (Pruitt *et al.*, 2014), which is confirmed by our approach (Fig. 4C). NEK2 is a protein which localizes to the centrosome, and is undetectable during G1 phase, but accumulates progressively throughout the S phase (Pruitt *et al.*, 2014), which is consistent with our model as well (Fig. 4D). All kinase activations can be found in the Supplementary Material.

In the insulin signalling dataset, a total of 9782 phosphosites was measured with respect to eight time points (0, 0.5, 1, 2, 3, 4, 6 and 10 min after stimulation with insulin). The values are given as $\log_2$ and lie within the interval [0,50]. IKAP Part 1 leads to a reduced dataset containing 487 phosphosites for which a kinase is known and 143 kinases that are known to phosphorylate at least one site in the data (Steps 1 and 2). After calculating the truth table (Step 3), the kinase activities and affinity parameters were estimated with an iteration number of 100 (Step 4). For affinity parameters, we chose the same bounds as for the mitosis experiment (0.1 and 10), whereas bounds for kinase activities were set to 0 and 50 to match the range of measured values. Here, no proteome normalization was performed and Equation (1) stays in its basic form. Fitting led to an overall χ^2/N value of 1.51 for $N = 3896$ data points (487 phosphosites at 8 time points) and 2095 parameters [951 affinities and 1144 kinase activities (143 kinases at 8 time points)], using a SD of 2.

The five most and least active kinases at each time point compared with baseline are given in Table 2. After 30 s and 1 min, we see a strong increase in the activity of Nemo-like-kinase, which is an atypical member of the MAPK family. Its activation mechanism and downstream targets are still not well characterized. Thus, it could be an interesting target of further investigations. TBK1, whose activity rises after 1 min, can mediate NFKB activation in response to certain

Table 1. Mostly up- and down-regulated kinases at each cell cycle stage and their calculated activities

G1 ↑	G1/S ↑	Early S ↑	Late S ↑	G2 ↑	M ↑
CDC42BPA 13.67	IRAK1 15	PTK6 14.07	MAP2K7 15	IRAK1 14,99	RET 10,35
PKN1 12.82	NLK 15	PDGFRB 12.59	NLK 8.18	NLK 14,99	PTK6 10,06
CIT 12.76	MARK3 13.57	MAP2K6 10.49	YES1 6.98	MAP3K1 12,16	MAP2K3 8,68
ARAF 11.85	MAP2K1 10.46	YES1 8.69	IRAK1 6.37	MAP2K3 9,97	CDC42BPA 8,53
ILK 10.63	DAPK3 9.69	CDC42BPA 8.36	CSNK1G2 5.86	CSNK1G2 7,26	PDGFRB 6,04
G1 ↓	G1/S ↓	Early S ↓	Late S ↓	G2 ↓	M ↓
CAMK2G−15	MAP2K7−15	ROCK2−15	MAP3K1−7.26	MAP2K7−15	ROCK2−15
MAP2K7−15	PTK6−15	MST4−15	PKN1−4.31	DAPK3−6.02	MST4−15
DAPK3−12.61	CDC42BPA−11.68	MAP2K7−12.66	CAMK2G−3.15	ARAF−4.34	MAP2K6−9.33
PTK6−10.85	CAMK2G−9.78	RET−10.99	MAP2K6−2.83	MAP2K6−3.86	DAPK3−3.92
RET−10.58	ILK−8.64	MAP2K3−8.33	CDK6−2.81	PRKAA2−3.27	MAP2K7−2.94

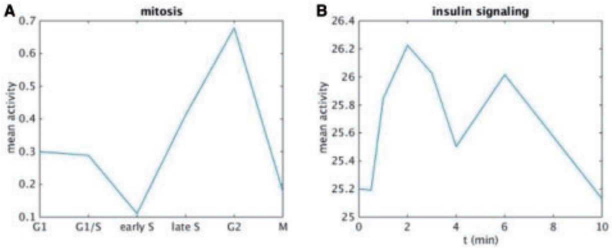

Fig. 3. Mean kinase activity during the cell cycle stages in HeLaS3 cells (**A**) and after insulin stimulation in mouse hepatocytes (**B**)

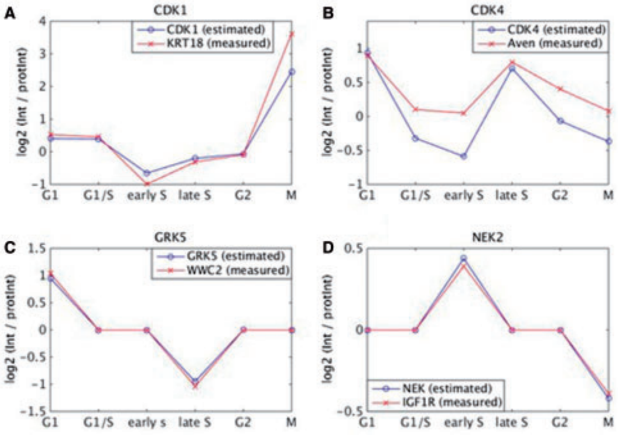

Fig. 4. Estimated kinase profiles in the HeLaS3 experiment, each in combination with one of its most likely targets. Original measurement values were given as $\log_2$ ratios of phosphoproteome intensity and proteome intensity. Kinase activities correspond to $\log_2$ ratios of mobilization and proteome intensity. Lines with circles = kinases, lines with crosses = targets

growth factors. This result suggests a crosstalk between the classical insulin receptor and the NFKB pathway and could be another rewarding target for future experiments. The insulin receptor shows up among the 10 mostly up-regulated kinases at all examined time points, which is what we would expect. The strongest down-regulation affects kinases which are involved in cytoskeletal rearrangements, like MYLK, ROCK2 and LIMK2.

In Figure 3B, the mean kinase activity is plotted in dependence of time. The trajectory shows a biphasic behaviour with a first peak after 2 min and a second after 6 min. This could be due to autocrine signalling or other time-delayed feedback-loops.

As in the HeLaS3 example produces IKAP Step 2.1 a distance matrix which displays *P*-values for all kinase–phosphosite pairs. Step 2.2 extracts all significant links. Here, a false discovery rate of 0.0005 was applied. Figure 5 exemplary depicts the profiles of GRK6, CDK5, PLK3 and MAP2K4, each along with one of their most likely targets. In this dataset, the activities of many kinases show either a peak or a drop within the first minute after stimulation, followed by a return to basal levels. Thus, insulin seems to have a rather short lasting effect in murine hepatocytes.

3.2 Identifiability analysis

To test whether the optimization results are unique, we performed an identifiability analysis by means of profile likelihood examination (Raue *et al.*, 2009). For each condition and each kinase, the computed kinase activity was increased and decreased in a stepwise manner. The new value was then fixed and the other parameters reoptimized to get an impression of the χ^2 landscape along the dimension of the respective kinase activity. All kinase activities displayed a minimum within the χ^2 space and were thus found to be identifiable. Figure 6 exemplary shows the χ^2 profiles of the kinases CDK1, CDK4, GRK5 and NEK2 at G1 of the HeLaS3 dataset (Fig. 6A) and of AKT1, MAPK1, MTOR and TBK1 at 0.5 min of the mouse liver dataset. In some cases, like the CDKs, an activity increase has more severe effects than a decrease, whereas in other cases, like NEK2, the activity can vary within a certain interval without larger effects.

3.3 Validation

By exerting IKAP Part 3, we validated our approach through literature and motif information. After producing the lists *psig* and *pnsig*, we tested both lists against all known kinase–target pairs of the IPA and MetaBase databases and computed a *P*-value by applying a fisher exact test to the number of hits in both lists (Step 2 IKAP Part 3). This was done five times, each time varying the list *pnsig*. In the HeLaS3 example, the number of hits for *psig* turned out to be 42 out of 2684, whereas the average number of hits for *pnsig* was just 2 out of 2684. This led to a *P*-value of $9.7278*10^{-11}$ (pDB Mit, Table 3). In the mouse liver example, *psig* showed 33 hits out of 3599 significant kinase–target pairs. Compared with *pnsig*, this leads to a *P*-value of $1.3605*10^{-5}$ (pDB Ins). Estimating kinase activity profiles and correlating those to phosphosite measurements thus produces a significantly higher amount of literature-known kinase–target pairs than taking random combinations of kinases and targets.

Step 3 finally utilizes the amino acid sequence of the phosphorylated peptide to infer a likelihood for the kinase–target pair under

Table 2. Mostly up- and down-regulated kinases in mouse hepatocytes after insulin stimulation (activity fold change compared with control)

0.5 min ↑	1 min ↑	2 min ↑	3 min ↑	4 min ↑	6 min ↑	10 min ↑
NLK 49.9	NLK 24.83	TBK1 21.84	TBK1 19.44	ILK 17.35	LCK 19.21	TBK1 17.42
PASK 49.9	TBK1 18.46	ILK 18.05	YES1 16.63	YES1 16.92	TBK1 18.78	ILK 12.51
MARK1 46.5	AURKA 15.87	PDGFRA 16.02	MAP4K5 11.25	NLK 15.53	PDGFRA 14.68	LCK 12.27
MAP4K5 31.2	ILK 13.68	MAP4K5 13.74	ILK 10.71	LCK 14.16	ILK 13.61	YES1 10.72
DAPK3 29.8	INSR 7.23	FYN 12.58	PDGFRA 9.38	TBK1 13.71	FYN 11.38	PIM3 9.11
0.5 min ↓	1 min ↓	2 min ↓	3 min ↓	4 min ↓	6 min ↓	10 min ↓
MAPK12−45.62	CAMK4−34.14	MAP3K6−11.17	CDC42BP−9.43	CAMK4−49.99	CDC42BP−11.99	CAMK4−40.88
RPS6KB2−37.41	CDC42BP−16.88	CDC42BP−10.57	MAP3K6−8.86	MAP3K6−12.98	DAPK1−6.57	CDC42BP−21.06
ROCK2−32.11	LIMK2−15.11	LIMK2−8.24	PRKCH−5.10	TESK2−9.26	TESK1−5.96	MAP3K4−10.75
DAPK1−24.44	PRKCH−5.85	PRKCH−6.47	DAPK3−4.53	MYLK−7.18	LIMK2−5.65	TESK1−5.61
MYLK−22.94	DAPK3−5.61	DAPK3−5.26	IGF1R−4.20	CIT−6.07	PRKCH−5.32	MAP3K7−5.27

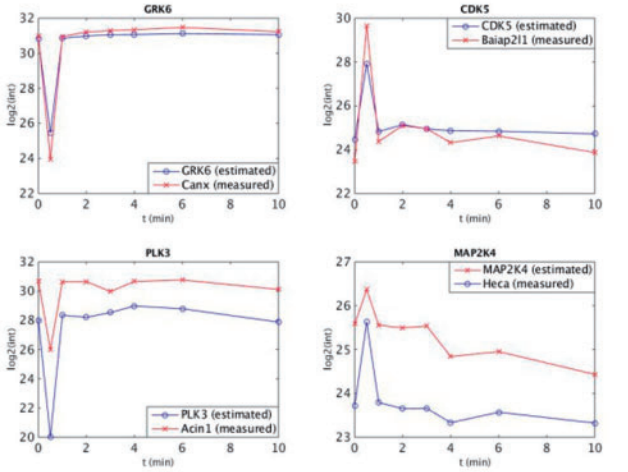

Fig. 5. Esitmated kinase profiles in the mouse liver experiment, each in combination with one of its most likely targets. Original measurement values and kinase activities are given as $\log_2$ intensities. Lines with circles = kinases, lines with crosses = targets

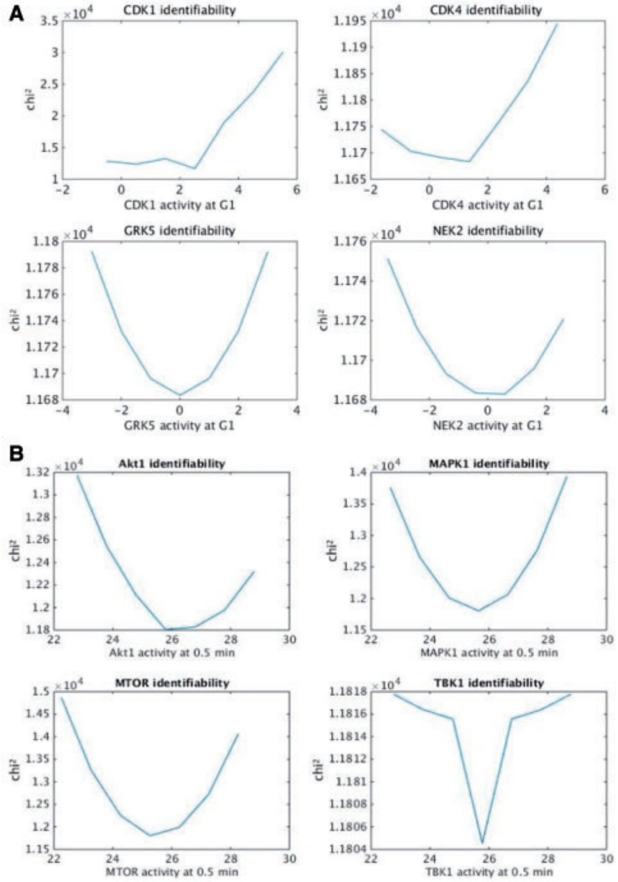

Fig. 6. Chi^2 space in the dimensions of CDK1, CDK4, GRK5 and NEK2 activities at G1 of HeLaS3 dataset (**A**) and of AKT1, MAPK1, MTOR and TBK1 at 0.5 min of liver dataset (**B**). All Chi^2 courses display a minimum, the kinase activities can thus be determined uniquely using the applied data

investigation via NetworKIN (Horn *et al.*, 2014). A likelihood ≥ 1 can thereby be regarded as significant. This step is again applied several times to both *psig* and *pnsig* to be able to calculate two different *P*-values on the basis of fisher exact tests. On the one hand, a *P*-value for the summed likelihoods of *psig* and *pnsig* is calculated in each iteration (pSUM). On the other, a *P*-value accounting for the number of significant likelihoods in *psig* and *pnsig* is computed (pNUM). Table 3 gives an overview of all three *P*-values generated by Steps 2 and 3 and displays the respective input numbers. pSUM turned out to be $2.5661*10^{-28}$ whereas pNUM came down to $8.8961*10^{-12}$ for the HeLaS3 example. For the mouse liver example, motif validation led to *P*-values of 0.0179 (pSUM Ins) and 0.0013 (pNUM Ins). Even though these are substantially higher than those for the HeLaS3 data, applying an alpha of 0.05 for the fisher exact test leads to significant test results in all cases. Thus, on the basis of sequence information the kinase–target links identified by our algorithm are substantially more likely than random pairs. To replenish our validation through findings that are not included in one of the three databases but nonetheless appear in literature, we performed a Linguamatics search (Bandy *et al.*, 2009) for both *psig* and *pnsig* links. This led us to 216 vs 122 hits resulting in a *P*-value (pLIN) of $1.4694*10^{-7}$. If a significance level of 0.01 is applied, all four *P*-values lie far below this threshold, indicating the algorithm to be beneficial. However, since the algorithm relies on correlations of the time courses of single kinases and phosphosites, whereas in the cell the time profile of a phosphosite is likely to be the result of several kinases phosphorylating it, the method can only capture those links where one kinase is predominant. In addition, it should be remarked that the overall coverage of significant links and validated links is a mere 1.56% for database-based, 1.93% for motif-based and 8.04 for literature-based validation [42, 52 and 216 of 2684 significant links in total (HeLaS3 example)]. The links found

Table 3. ***P*-values resulting from validation through databases (IPA and MetaBase), motif information (NetworKIN) and literature (Linguamatics) for HeLaS3 data (Mit) and mouse liver data (Ins)**

Type	*psig*	*pnsig* (mean)	Value
pDB Mit	42	2	9.7278×10^{-11}
pSUM Mit	226	52	2.5661×10^{-28}
pNUM Mit	52	4	8.8961×10^{-12}
pLIN Mit	216	122	1.4694×10^{-7}
pDB Ins	33	6	1.3605×10^{-5}
pSUM Ins	70	44	0.0179
pNUM Ins	16	2	0.0013
pLIN Ins	312	192	3.3889×10^{-8}

by the algorithm should therefor in the first place be regarded as guidelines for experimental verification. Nevertheless, we regarded a statistical validation of newly found kinase–target links to be a more reasonable means of verification than a comparison of kinase activities to autophosphorylation sites, since on the one hand in the most of all cases they can be phosphorylated by several other kinases as well, and on the other the activity of a kinase is a function of all its sites.

4 Discussion

Phosphorylation of serine, threonine and tyrosine plays significant roles in cellular signal transduction and in modifying multiple protein functions. Phosphoproteins are coordinated and regulated by a network of kinases, phosphatases and phospho-binding proteins, which modify the phosphorylation states, recognize unique phosphopeptides or target proteins for degradation. Detailed and complete information on the structure and dynamics of these networks is required to better understand fundamental mechanisms of cellular processes and diseases. High-throughput technologies have been developed to investigate phosphoproteomes in model organisms and human diseases. Among them, mass spectrometry-based technologies are the major platforms and have been widely applied, which has led to explosive growth of phosphoproteomic data in recent years. New bioinformatics tools are needed to analyse and make sense of these data. Moreover, most research has focused on individual phosphoproteins and kinases. To gain a more complete knowledge of cellular processes, systems biology approaches, including pathways and networks modelling, have to be applied to integrate all components of the phosphorylation machinery, including kinases, phosphatases, their substrates and phospho-binding proteins.

We developed here a machine-learning algorithm that estimates the activities of all kinases which are known to phosphorylate sequences occurring in a phosphoproteomics dataset. The algorithm thus reduces the dimensions of the parameter space to the number of kinases, and renders the data easier to display and analyse. The estimated activities can then be used for modelling or inference of new kinase–target links.

Previous approaches to infer kinase activities from phosphoproteomic data only rely on the number of targets in the sample, without taking the detected amounts of phosphorylation into account (Qi *et al.*, 2014), or solely deliver relative activities from one condition to another (Casado *et al.*, 2013) and do not include affinity parameters for kinase–target interactions. IKAP however calculates the absolute activities for each condition separately and thus provides a more precise picture of pathway engagement in biological samples. Furthermore, in opposition to previous approaches, an arbitrary number of kinases can be assigned to each phosphosite, which makes the approach more realistic compared with sorting phosphosites into distinct kinase groups and considers uncertainties in kinase specificities for certain targets. Especially in the case of time-resolved data the computed kinase activities are directly applicable as an input to subsequent dynamical modelling. The calculated affinities can thereby be used as starting points for parameter calibration, and through the use of identifiability analysis it is possible to determine how reliable the computed activities are. The derivation of novel kinase–target pairs as it is described in (Imamura *et al.*, 2014) is an *in vitro* method that has to be performed for each kinase individually. Thus, it is much more time consuming. Furthermore, IKAP provides a ready-to-use framework, whereas previous approaches need to be implemented by the user from scratch.

Although the approach does not include phosphatases, both the estimated kinase activities and the newly derived links proved to be reliable in literature- and motif-based validation. Our conclusion is that, at least in HeLaS3 cells and murine hepatocytes, phosphatase activity could rather be considered as a factor which removes phosphorylations at a constant rate than a mechanism which is under sophisticated regulation. Since this can be different in other cell types, the findings that are generated through the application of our method should always be critically appraised and validated. The algorithm represents a step towards a system-wide understanding of signal transduction processes within a cell. In the near future, we will further test it on other phosphoproteomics datasets and implement versions in other programming languages like Python and R to make it available to a larger community of scientists.

Funding

The project was funded by Sanofi-Aventis Deutschland GmbH, Frankfurt, Germany

Conflict of Interest: none declared.

References

Bandy,J. *et al*. (2009) Mining protein-protein interactions from published literature using Linguamatics I2E. *Methods Mol. Biol*., **563**, 3–13.

Casado,P. *et al*. (2013) Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci. Signal*., **6**, rs6.

Franceschini,A. *et al*. (2012) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*., **41**, D808–D815.

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

Hornbeck,P.V. *et al*. (2012) 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.

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

Imamura,H. *et al*. (2014) Large-scale identification of phosphorylation sites for profiling protein kinase selectivity. *J. Proteome Res*., **13**, 3410–3419.

Kirkpatrick,D.S. *et al*. (2013) Phosphoproteomic characterization of DNA damage response in melanoma cells following MEK/PI3K dual inhibition. *Proc. Natl. Acad. Sci. USA*, **110**, 19426–19431.

Liu,Y. and Chance,M.R. (2014) Integrating phosphoproteomics in systems biology. *Comput. Struct. Biotechnol. J*., **10**, 90–97.

Macek,B. *et al*. (2009) Global and site-specific quantitative phosphoproteomics: principles and applications. *Annu. Rev. Pharmacol. Toxicol*., **49**, 199–221.

Manning,G. (2005) Genomic overview of protein kinases. *WormBook*, **13**, 1–19. Review. PMID 18050405.

Morgan,D.O. *et al*. (2007) *The Cell Cycle: Principles of Control*, 1st edn. New Science Press, London.

Olsen,J.V. *et al*. (2010) Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci. Signal.*, **3**, ra3.

Pruitt,K.D. *et al*. (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res*., **42**, D756–D763.

Qi,L. *et al*. (2014) Systematic analysis of the phosphoproteome and kinase-substrate networks in the mouse testis. *Mol. Cell. Proteomics*, **13**, 3626–3638.

Raue,A. *et al*. (2009) Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, **25**, 1923–1929.

Terfve,C. *et al*. (2012) Modeling signaling networks using high-throughput phospho-proteomics. *Adv. Exp. Med. Biol*., **736**, 19–57.

Xia,L. *et al*. (2013) Phosphoproteomics study on the activated PKCδ-induced cell death. *J. Proteome Res*., **12**, 4280–4301.

Zimman,A. *et al*. (2014) Phosphoproteomic analysis of platelets activated by prothrombotic oxidized phospholipids and thrombin. *PLoS One*, **9**, e84488.