Network Reconstruction and Significant Pathway Extraction Using Phosphoproteomic Data from Cancer Cells

Marion Buffard, Aurélien Naldi, Ovidiu Radulescu, Peter Coopman, Romain Larive, Gilles Gf Freiss

Network reconstruction and significant pathway extraction using phosphoproteomic data from cancer cells

BUFFARD Marion[1,2], NALDI Aurélien[3], RADULESCU Ovidiu[2,#], COOPMAN Peter J[1,#], LARIVE Romain Maxime[4,#,¶] and FREISS Gilles[1,#,¶]

[#] Equal contributing authors

[1] IRCM, Univ Montpellier, ICM, INSERM, Montpellier, France.
[2] DIMNP, Univ Montpellier, CNRS, Montpellier, France.
[3] Computational Systems Biology Team, Institut de Biologie de l'École Normale Supérieure, Centre National de la Recherche Scientifique UMR8197, INSERM U1024, École Normale Supérieure, PSL Université, Paris, France.
[4] IBMM, Univ Montpellier, CNRS, ENSCM, Montpellier, France.

¶ To whom correspondence should be addressed at: Gilles FREISS, IRCM, INSERM U1194, 208 rue des Apothicaires, F-34298 Montpellier cedex 5, France. Tel: + 33 467 61 31 91. Fax: + 33 4 67 61 37 87. E-Mail: gilles.freiss@inserm.fr ; Romain LARIVE, Faculté des Sciences Pharmaceutiques et Biologiques, Laboratoire de Toxicologie du Médicament - Bâtiment K - 1er étage, 15 avenue Charles Flahault - BP 14491, 34093 Montpellier Cedex 5. Tel: + 33 411 75 97 50. Fax: + 33 411 75 97 59. E-Mail: romain.larive@umontpellier.fr

Short title

Reconstructing signaling networks in cancer cells by phosphoproteomic data

Abbreviations

SRMS, Src-related tyrosine kinase lacking C-terminal regulatory tyrosine and N-terminal myristoylation sites
PIK3CA, phosphatidyl-inositol 3-kinase enzyme
KEGG, Kyoto Encyclopedia of Genes and Genomes

Keywords

Data processing and analysis; Phosphoproteomics; Oncogenic signaling; SRMS; PIK3CA

Total number of words: 5825

Abstract

Protein phosphorylation acts as an efficient switch controlling deregulated key signaling pathways in cancer. Computational biology aims to address the complexity of reconstructed networks but overrepresents well-known proteins and lacks information on less-studied proteins. We developed a bioinformatic tool to reconstruct and select relatively small networks that connect signaling proteins to their targets in specific contexts. It enabled us to propose and validate new signaling axes of the Syk kinase. To validate the potency of our tool, we applied it to two phosphoproteomic studies on oncogenic mutants of the well-known PIK3CA kinase and the unfamiliar SRMS kinase. By combining network reconstruction and signal propagation, we built comprehensive signaling networks from large-scale experimental data and extracted multiple molecular paths from these kinases to their targets. We retrieved specific paths from two distinct PIK3CA mutants, allowing us to explain their differential impact on the HER3 receptor kinase. In addition, to address the missing connectivities of the SRMS kinase to its targets in interaction pathway databases, we integrated phospho-tyrosine and phospho-serine/threonine proteomic data. The resulting SRMS-signaling network comprised casein kinase 2, thereby validating its currently suggested role downstream of SRMS. Our computational pipeline is publicly available, and contains a user-friendly graphical interface (http://dx.doi.org/10.5281/zenodo.3333687).

Statement of significance of the study

This study applies and validates a novel bioinformatic pathway extraction and analysis tool on two phosphoproteomic studies on the signaling of oncogenic mutants of the PIK3CA kinase and the SRMS kinase. By combining network reconstruction and signal propagation analysis, we build comprehensive cell signaling networks from substantial experimental data and extract multiple molecular pathways from a kinase to its targets. These various alternatives, ranked by their biological significance, enable us to conceive of molecular hypotheses requiring experimental validation. The results of this study demonstrate here that our framework can be applied to explore substantial amounts of phosphoproteomic data at the network level.

1 Introduction

Aberrant protein phosphorylation contributes to tumor initiation and progression. Despite the development of targeted kinase inhibitors, it remains difficult to predict how tumors will respond to them; which inhibitors to combine; and how to overcome acquired drug resistance. A major shortcoming is the poor molecular understanding of the kinase signaling networks and remains a challenging bioinformatical task.

Pathway-oriented databases, such as KEGG [1–3], Pathway Commons [4] and Reactome [5], contain regulatory relations between proteins, allowing large-scale reconstruction of signaling networks. These databases rely on curation and updating of the interactions. These databases suffer from the overrepresentation of well-studied proteins and the lack of information on less-known proteins. Discovery of signaling pathways and molecular cross-talk is based on experiments and efficient bioinformatic tools that are able to exploit new experimental data and correct the extant biases in the databases. Several tools, such as Netwalker and Pathlinker, were used for the analysis of large-scale networks [6,7]. Netwalker is a software application suite with random walk-based network analysis methods for network-based comparative interpretations of genome-scale data. Pathlinker computes the *k*-shortest simple paths in a network from a source to a target with an option for weighting the edges in the network.

We recently developed a new bioinformatic pipeline that combines the advantages of these two existing methods. Additionally, we integrated our methodology with the reconstruction of a large network composed of the elements of existing database pathways, which are enriched in targets previously identified by phosphoproteomic experiments. This step, prior to network analysis by subnetwork-extraction, avoids the major drawback of the aforementioned over- or underrepresentation. This methodology was applied to the reconstruction and signal propagation analysis of the Syk kinase signaling network in breast cancer cells [8]. The method allows reconstruction of a kinase-related network from the global phosphoproteomic data obtained by mass spectrometry. The input to our method was a list of Syk-dependent differentially tyrosine-phosphorylated proteins [9]. We selected the pathways from existing databases, enriched in Syk-targets, to recreate a global network of signaling proteins. This large network still contains numerous unessential proteins, and we developed a reduction algorithm by selecting the most appropriate potential paths from Syk to its targets. We first associated weights to the interaction network edges. These weights promoted network-directed edges coming from a protein kinase or phosphatase to an identified target and demoted edges with no biological relevance. We then refined these weights by taking into account the topology of the network and optimizing signal propagation, by a random walk with restart (RWR). Subnetworks, related to specific biological processes and based on the Syk-target Gene Ontology, were then extracted. This workflow generated valuable results and allowed us to validate the involvement of Syk in actin-mediated adhesion and motility via cortactin and ezrin.

In this study, we further develop the functionality of our bioinformatic tool by adapting and applying it to two phosphoproteomic studies on the signaling of oncogenic PIK3CA (phosphatidyl-inositol 3-kinase) mutants and the SRMS (Src-related tyrosine kinase lacking C-terminal regulatory tyrosine and N-terminal myristoylation sites) kinase. We optimized the automation of our initial Python code to facilitate the implementation of our bioinformatic method. This approach enables us to retrieve specific signaling molecular paths from two

distinct PIK3CA mutants to the HER3 receptor. We also generated the proximal and distal signaling networks of the SRMS protein tyrosine kinase comprising secondary signaling intermediates by integrating phospho-tyrosine and phospho-serine/threonine proteomic data. We integrate these improvements into our workflow and propose a graphical interface allowing one to apply this bioinformatic pipeline to other phosphoproteomic analyses.

2 Materials and Methods

2.1 Phosphoproteomic data for the bioinformatic workflow

The bioinformatic workflow input is a list of the UniProt Accession Numbers (AC) of proteins that have been identified as differentially phosphorylated (named “targets”) between experimental conditions perturbing the concerned kinase (named “source”).

Identification of specific paths from PIK3CA mutants to the receptor tyrosine kinase HER3 (section 3.2) involved the following: The quantitative phosphoproteomic analyses comparing the control or isogenic breast cancer cell lines that express the E545H or H1047R PIK3CA mutants were performed as reported [10]: After protein extraction, trypsin digestion, and anti-phosphotyrosine immuno-affinity chromatography enrichment, the SILAC-labeled peptides were identified and quantified by LC-MS/MS. The datasets of the protein targets of the E545H or H1047R PIK3CA mutants are displayed in Supplementary Tables S1-4 and were obtained from Supplementary Tables S1-4 of the original work [10]. The sources are the E545H or H1047R PIK3CA mutants.

Reconstruction of the SRMS signaling network by integration of multiple phosphoproteomic data sets (section 3.3) involved: The label-free quantitation-based phosphoproteomic analysis using cells expressing GFP alone (the empty vector control) or cells expressing wild-type GFP-SRMS was performed as described [11]. After protein extraction, the proteins were digested by dual enzymatic digestion (Trypsin/Lys-C) and the phosphopeptides were enriched using TiO2 resin. The dataset containing the indirect targets of SRMS (proteins differentially phosphorylated on serine and threonine) is displayed in Supplementary Table S5 and was obtained from Supplementary Tables S4-5 of the original work [11]. The dataset containing the direct substrates of SRMS is displayed in Supplementary Table S6 and was obtained from Supplementary Table S8 of the original work [12]. The source is SRMS to search the paths from SRMS to the CK2 subunits. The sources are the four CK2 subunits to search the paths from CK2 to the indirect targets of SRMS.

2.2 Online databases

UniProt AC mapping from UniProt.org/downloads (2017/02)

HGNC dataset from genenames.org/cgi-bin/statistics (2017/02)

GO ontology from geneontology.org/page/download-ontology (go-basic.obo, 2017/02)

GO annotation from geneontology.org/page/download-annotations (goa_human.gaf, 2017/02)

KEGG: www.kegg.jp, release 84 (2017/10)

Pathway commons: pathwaycommons.org release 8 (2018/01)

2.3 Pathway database selection

We used pathways from the KEGG [1] and Pathway Commons [4] databases. For PIK3CA reconstruction, we first selected the more enriched pathways in the lists of targets (using a Fisher exact test) and included the pathways containing targets not covered by significantly overrepresented pathways from the same database [8]. As SRMS signaling has not been characterized, we kept all the pathways without selection and added the links from SRMS to its identified direct substrates [12]. The selected pathways were combined, resulting in a larger directed network forming the prior-knowledge network. Each node corresponds to a unique protein and edges to all its interactors in the different selected pathways.

2.4 Functional protein annotations

For PIK3CA reconstruction, the components of the network with tyrosine kinases (GO:0004713) and tyrosine phosphatases (GO:0004725) GO terms were annotated as phospho-tyrosine modifiers and we extended the list of phospho-tyrosine modifiers from 123 proteins to 207 manually verified proteins. For SRMS, those proteins present in the serine/threonine kinase activity list (GO:0004674) were annotated as kinase proteins.

2.5 Search path from source to targets

The reconstructed, embedded, large network contains thousands of nodes and edges and billions of path possibilities. We constrained the path research by using an *ad hoc* distance and edge weights.

The path research is based on a weighted near-shortest-path analysis that employs a modified version of Dijkstra's shortest path algorithm. To define the edge weights, we combine functional annotation with random walk for weight refinement.

As targets are differentially phosphorylated, we promote edges from a kinase or phosphatase, in correlation with the functional annotation for each dataset studied, to a target by adding a smaller weight to the corresponding edges (adapted from [8]). Conversely, we demote edges reaching a target identified as differentially phosphorylated but that did not originate from a kinase or phosphatase (for the complete list of *ad hoc* weights used in this study, see the Supp Figure S1).

Random walk analysis allows the weights to be refined by taking into account network topology. This analysis allows the avoidance of multiple paths with exactly the same length and favors plausible paths containing crossroad proteins. We simulated a random walk with return on the network twice; firstly using equal weights for all edges and a secondly using the *ad hoc* weights. The equilibrium node probabilities, in the two cases, are used to modulate the *ad hoc* weights and eliminate biases created by topology (for details, see [8]). Contrary to its usual implementation [14], we do not use the random walk method to prune the network but to refine its initial weights. The final path selection was performed using Dijkstra's algorithm. The Dijkstra algorithm identifies the shortest path from source to every target. As alternative paths can also be interesting, we slightly modified this algorithm from its original form.

In this modified algorithm, not only the shortest paths but also longer paths are accepted. The "overflow", defined as the extra distance measured as percentage of the shortest path, necessary to include near-shortest paths, is a parameter of the method. The overflow is zero for the shortest paths. The choice of shortest paths is sufficient on the first analysis to test that all targets were connected to the source. To refine the analysis for specific targets, the overflow value should be set empirically, by continuously increasing it from zero until new, alternative paths are selected.

2.6 Subnetwork extraction

Sets of alternative paths define subnetworks in the prior large network. Finally, it is also possible to extract subnetworks according to groups of GO terms representing relevant processes and functions, for example cell adhesion and motility [8] or a selected subset of targets. This approach was used to separate networks, by processes and functions, or to reduce the network size to explore, more deeply, alternative paths (e.g. HER3 in the PIK3CA mutant study).

2.7 Network visualization, comparison and analysis

Cytoscape 3.7 (http://www.cytoscape.org/) was used to visualize and explore the networks and to generate figures [15]. For alignment and comparison of the networks obtained for PIK3CA mutants we used DyNet, the Cytoscape plug-in [16]. The parameters were set as follows. Initial layout: Prefuse Force Directed Layout; Treat networks as: Directed networks; Find corresponding nodes by: name; Find corresponding edges by: interaction.
To retrieve information about the putative *in vivo* kinases and the functional effects of phosphorylation on the activity of the target proteins that were identified with differentially phosphorylated peptides, we manually consulted the site-specific annotation database PhosphoSitePlus [17].

3 Results and Discussion

3.1 Improvement of network reconstruction and shortest path analysis

The original Python code (https://github.com/aurelien-naldi/NetworkReconstruct) was modified to optimize the automatic network reconstruction and extraction of all subnetworks within the same improved application program. We also integrated all the pathways from Pathway Commons (Reactome, Panther, and PID) with the KEGG's pathways (Figure 1). The selection of signaling pathways in databases can be expanded as much as is necessary to maximize the path possibilities (step 1). This selection could be necessary when the resulting prior-knowledge network lacks connectivity from source to targets. We then embedded the pathways to create a directed network (step 2). If no selection is applied to step 1, a large network will be generated containing all known database pathways. We also included the possibility of adding protein interactions from experimental data directly to the prior-knowledge network, which is particularly useful if there is a lack of connectivity of the source to the prior-knowledge network (e.g., when the source is poorly described in pathway databases). We applied both options in the case of the SRMS kinase, adding the interactions from SRMS to its substrates (see subsection 3.3 for more details). To detect the most relevant paths and to eliminate unnecessary interactions (step 3), we used the strategy of the weighted shortest path search. This added the possibility of promoting edges according to the type of phosphoproteomic data (see the material and methods). The topology of the network is still taken into account by refining the weight of the protein interactions with the random walk procedure. The subnetwork extraction has also been integrated within the script and can be applied to retrieve those paths from the source to all experimentally identified targets and to those involved in specific cellular processes (based on their Gene Ontology), or to a particular list of proteins of interest (step 4). The overflow, admitting the shortest paths and allowing the inclusion of alternative paths, has also been made modular. This option is important in network biology to understand the etiology of drug mechanisms and drug resistance. The application of this bioinformatic methodology is illustrated below and applied to two phosphoproteomic studies.

3.2 Identification of specific paths from *PIK3CA* mutants to the receptor tyrosine kinase HER3

As a first validation of our improved method for analyzing the molecular paths from a protein to its targets, we selected a study published in *Proteomics* that uses a mass spectrometry-based phosphoproteomic approach to identify unique mediators of the oncogenic *PIK3CA* signaling [10]. *PIK3CA* is an attractive target for cancer therapy because its activity is often dysregulated in cancer. The p110α catalytic subunit of PI3K, encoded by the *PIK3CA* gene, is one of the most frequently mutated oncogenes in breast cancer [18]. Two recurrent oncogenic

"hotspot" mutations, E545H and H1047R, occur in the helical and the kinase domains of the PIK3CA protein [19]. Although the implications of *PIK3CA* mutations in cell transformation have been established [20], the mechanisms of how they lead to increased oncogenic features have not been determined to date . Figure 1 of Blair and colleagues (2015) depicts the experimental strategy applied to analyze the impact of each of these two mutations on the cell signaling of human breast epithelial cells. A differential phosphorylation pattern was observed between the two PIK3CA mutants. The authors focused on their distinct impact on HER3 that is specifically phosphorylated on the Y1328 residue in the presence of the H1047R mutant or on the Y1159 residue in the presence of the E545H mutant. Thus, HER3 could be a molecular intermediate that conducts the signal from the H1047R mutant, but not the E545H mutant, to the MAP kinase pathway.

To identify the specific paths from each of the PIK3CA mutants to the HER3 receptor, we reconstructed the signaling networks of each PIK3CA mutant using quantitative phosphoproteomic comparison with the control condition. To explain the different consequences of the two mutants, we considered the same source, but with different targets, in our network reconstruction process, leading to two distinct prior networks (Supp Tables S1-2). We then applied our analytic method to select the more reliable paths linking the PIK3CA mutants to their targets in each network. The superposition of these two "shortest path" networks revealed paths specific for the E545H PIK3CA (red edges and nodes) or H1047R PIK3CA mutant (green edges and nodes) (Figure 2A). Among the shared targets (white diamonds), some were reachable by paths specific to the E545H PIK3CA mutant (red edges to white diamonds) or to the H1047R PIK3CA mutant (green edges to white diamonds). These properties highlight differences between the signaling networks of each PIK3CA mutant and allow us to formulate molecular hypotheses that can be experimentally verified. Next, we focused on the paths linking PIK3CA to HER3 in each network and found the same path for the two mutants, linking PIK3CA to HER3 through the tyrosine kinases PTK2 (focal adhesion kinase 1) and FYN (Figure 2B). Although this result suggests the indirect impact of PIK3CA on the tyrosine phosphorylation of HER3, it did not explain the differential regulation of HER3 by the two PIK3CA mutants. Enlarging our selection to the near shortest paths, an alternative path through the SRC kinase was detected (Supp Figure S2A-B) but was still shared by the two mutants.

According to Blair and colleagues (2015), comparing quantitative differences between the E545H and H1047R PIK3CA mutants allows one to focus on the unique signaling alterations of these two mutations. We therefore refined our analysis by selecting only the phosphoproteomic differences between the E545H and H1047R mutants (Supp Tables 3-4). We reconstructed the PIK3CA mutant networks and searched for paths leading to HER3. The path from the E545H mutant to HER3 remained identical, but the path from the H1047R mutant was profoundly modified, with the MET receptor kinase serving as the final component linked to HER3 (Supp Figure S2C). MET was experimentally identified in the phosphoproteomic screen and the interaction between MET and HER3 has been shown to confer resistance of cancer cells to EGFR pharmacological inhibitors [21]. Nevertheless, the previous step of this path was the interaction between the ligand for the receptor-type KIT kinase (KITLG) and MET, and KITLG has not been described as an activator of MET. This interaction was retrieved from three KEGG pathways as a general mechanism describing the activation of receptor tyrosine kinases by extracellular growth factors (RAS, PI3K-AKT and RAP1). We enlarged the selection to the near shortest paths and, searching for more relevant interactions upstream of MET, we identified the hepatocyte growth factor (HGF) MET ligand that is linked to the H1047R PIK3CA mutant by STAT3, MAPK1 and PTK2 (Figure 2C and

Supp Figure S3). The majority of components in this path were experimentally identified in the phosphoproteomic screen (diamond nodes). We retained this hypothetical path that describes an autocrine and/or paracrine signaling mechanism with the production of extracellular HGF leading to phosphorylation of HER3 by MET. This example clearly demonstrates that our method is useful to retrieve the molecular signaling pathways from protein kinases to their targets, identified by a phosphoproteomic screening, at a level of detail and plasticity for which interesting biological hypotheses may be generated, explored, and tested.

3.3 Reconstruction of the SRMS signaling network by integration of multiple phosphoproteomics data

SRMS is a nonreceptor protein-tyrosine kinase that belongs to the BRK family kinases. While discovered in 1994, little information on the biochemical, cellular and physiopathological roles of SRMS has been reported [22]. SRMS is highly expressed in breast cancers compared to normal mammary cell lines and tissues and SRMS is a candidate serum biomarker for gastric cancer [23,24]. Recently, Goel–and colleagues [11] attempted to uncover SRMS-regulated signaling by identifying differentially phosphorylated peptides on serine or threonine by mass spectrometry. The phosphorylation of these SRMS-indirect targets indicates the regulation of protein-serine/threonine kinase signaling intermediates by SRMS. Phosphorylation motif analysis suggested that casein kinase 2 (CK2) may represent a key downstream target of SRMS [11]. Interestingly, CK2 has been characterized as a crucial player in cancer biology and an attractive target for anticancer drug design [25,26].

To identify the molecular paths linking SRMS to its indirect targets through CK2, we applied our methodological workflow to reconstruct the SRMS-associated network using the proteins differentially phosphorylated on serine and threonine (Supp Table S5). Despite a resulting network of a consistent size (5216 edges and 760 nodes), SRMS was isolated from the major region of this network and only linked to the BRK/PTK6 kinase (Figure 3A and Supp Figure S4). We assumed that this lack of connectivity from SRMS to its signaling network was a consequence of its underrepresentation in signaling databases and searched to add direct protein interactions of SRMS to the network. Goel and colleagues [11] identified novel candidate SRMS substrates using phosphotyrosine antibody-based immunoaffinity purification in large-scale, label-free, quantitative phosphoproteomics and validated a subset of the SRMS candidate substrates by high-throughput peptide arrays [12]. We enriched the set of SRMS targets used for the pathway selection step of the network reconstruction with the SRMS-candidate substrates (steps 1-2 of our methodological workflow) (Supp Table S6). We also added the direct interactions, from SRMS to its substrates, to the set of network interactions, increasing the size of the resulting protein interaction network (6321 edges and 1307 nodes) (Supp Figure S5). Consequently, we searched for the molecular paths from SRMS to its indirect targets. Despite the reconnection of SRMS to the major part of the directed network, only two of its indirect targets were reachable from SRMS (Figure 3B).

Our network reconstruction procedure is based on the generation of a prior-knowledge interaction network composed of the components and interactions described in the public databases of signaling pathways. While such networks are often assembled using complete pathway or interaction databases, we select only the enriched pathways in the list of phosphoproteomic data. Consequently, this restriction reduces the number of irrelevant

interactions and better assesses the relevance of the identified pathways. In the case of SRMS, however, we did not obtain enough coverage to connect SRMS to its indirect targets. For this reason, we enlarged the prior-knowledge interaction network to the components and interactions described in the public databases of all signaling pathways without selecting those enriched in the list of phosphoproteomic data. Then, we searched for the molecular paths from SRMS to its indirect targets in the resulting prior-knowledge interaction network (111736 edges and 8313 nodes). Among the 60 indirect targets of SRMS, 29 were present in this network and 16 were now reachable from SRMS. Since CK2 was identified as one of the major potential SRMS-secondary signaling intermediates, we included CK2 in the list of SRMS-indirect targets and searched for the most relevant molecular paths from SRMS to its indirect targets. In agreement with Goel and colleagues [11], we retrieved CK2 as a candidate intermediate protein-serine/threonine kinase to propagate the signal to the SRMS-indirect targets (Figure 3C). CK2 is composed of two α and two β subunits that appear in our network as CSNK2A1, CSNK2A2, CSNK2A3 and CSNK2B. These subunits are reachable from SRMS through CDK2 (cyclin-dependant kinase 2), a potential SRMS substrate, and propagate the signal to PSMA3 and RAD23A, two indirect SRMS targets, PSMA3 being a known substrate of CK2 [27]. Three other SRMS-candidate substrates are involved in propagating the signal to the SRMS-indirect targets; the CDK1 (cyclin-dependant kinase 1), the GEFs VAV2, and DOK1. Interestingly, CDK1 was also retrieved as a candidate intermediate kinase by Goel and colleagues [11] and DOK1 has been described as an SRMS substrate [23]. To test whether CK2 could propagate the signal from SRMS to all of its indirect targets, we searched the paths from the four CK2 subunits to the SRMS-indirect targets. All of the targets were reachable from each CK2 subunit, suggesting that the role of CK2 as a downstream intermediate of SRMS could be even more prominent. We merged these paths with the path from SRMS to CK2 to obtain a SRMS-signaling network with paths that could confirm the role of CK2 as an intermediate of SRMS (Supp Figure S6). These networks now allow us to formulate molecular hypotheses that require experimental validation to explore the functional role of each of the CK2, CDK2, CDK1 and DOK1 kinases as signaling intermediates of the SRMS kinase.

3.4 Graphical interface

The bioinformatic workflow presents compelling results and provides valuable indications to study protein signaling pathways based on phosphoproteomic data. Used first to study Syk kinase signaling in breast cancer, we demonstrate here that this workflow can be more widely applied to other kinases and other types of data (direct substrates, serine/threonine phosphorylations) with appropriate modifications. Moreover, we have developed a graphical interface that combines the different options and adaptations that were included in this study (Supp Figures S7-8 and Supplementary text) (http://dx.doi.org/10.5281/zenodo.3333687).

4 Concluding Remarks

In this study, we demonstrate that our recently developed bioinformatic pipeline can be generally adapted to other phosphoproteomic datasets and allow us the discovery of candidate mechanisms that explain how signals propagate in large networks of signaling proteins. Further improvements, such as the consideration of the phosphorylated sites and the quantitative phosphoproteomic data, would be necessary to advance towards spatiotemporal dynamic models of signaling and behavior. Taking into account the site of phosphorylation rather than the entire protein would lead to the possibility of predicting the protein kinase upstream of each detected phosphorylation, by analyzing the phosphorylation motifs. Additionally, the impact of phosphorylation on the activity of the identified targets could be retrieved from phosphorylation databases and used to refine the inference of signal propagation. Finally, introducing the quantitative dimension of the phosphoproteomic data would permit the quantification of the static response of the signaling network to specific perturbations, by the bias of, for instance, modular response [28] or static response analysis methods [29]. The results of this study may open the path towards a dynamic description of signaling that uses detailed representations of the interaction mechanisms and can integrate temporal fluctuations at the system level [30].

Acknowledgments

This work was supported by grants from the Plan Cancer (ASC14021FSA), the Ligue Régionale Contre le Cancer (Hérault R18024FF) and the INCa-Cancéropôle GSO (Emergence program, N°2018-E01). MB is a recipient of the Labex EpiGenMed PhD fellowship (an “Investissements d’avenir” program, reference ANR-10-LABX-12-01). The authors declare no conflict of interest.

5 References

[1] M. Kanehisa, S. Goto, *Nucleic Acids Res.* **2000**, *28*, 27.

[2] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, *Nucleic Acids Res.* **2017**, *45*, D353.

[3] M. Kanehisa, Y. Sato, M. Furumichi, K. Morishima, M. Tanabe, *Nucleic Acids Res.* **2019**, *47*, D590.

[4] E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, Ö. Babur, N. Anwar, N. Schultz, G. D. Bader, C. Sander, *Nucleic Acids Res.* **2011**, *39*, D685.

[5] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C. D. Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio, *Nucleic Acids Res.* **2018**, *46*, D649.

[6] K. Komurov, S. Dursun, S. Erdin, P. T. Ram, *BMC Genomics* **2012**, *13*, 282.

[7] A. Ritz, C. L. Poirel, A. N. Tegge, N. Sharp, K. Simmons, A. Powell, S. D. Kale, T. M. Murali, *Npj Syst. Biol. Appl.* **2016**, *2*, 16002.

[8] A. Naldi, R. M. Larive, U. Czerwinska, S. Urbach, P. Montcourrier, C. Roy, J. Solassol, G. Freiss, P. J. Coopman, O. Radulescu, *PLoS Comput. Biol.* **2017**, *13*, e1005432.

[9] R. M. Larive, S. Urbach, J. Poncet, P. Jouin, G. Mascré, A. Sahuquet, P. H. Mangeat, P. J. Coopman, N. Bettache, *Oncogene* **2009**, *28*, 2337.

[10] B. G. Blair, X. Wu, M. S. Zahari, M. Mohseni, J. Cidado, H. Y. Wong, J. A. Beaver, R. L. Cochran, D. J. Zabransky, S. Croessmann, D. Chu, P. V. Toro, K. Cravero, A. Pandey, B. H. Park, *PROTEOMICS* **2015**, *15*, 318.

[11] R. K. Goel, M. Meyer, M. Paczkowska, J. Reimand, F. Vizeacoumar, F. Vizeacoumar, T. T. Lam, K. E. Lukong, *Proteome Sci.* **2018**, *16*, 16.

[12] R. K. Goel, M. Paczkowska, J. Reimand, S. Napper, K. E. Lukong, *Mol. Cell. Proteomics MCP* **2018**, *17*, 925.

[13] R. K. Goel, M. Meyer, M. Paczkowska, J. Reimand, F. Vizeacoumar, F. Vizeacoumar, T. T. Lam, K. E. Lukong, *Proteome Sci.* **2018**, *16*, 16.

[14] K. Komurov, M. A. White, P. T. Ram, *PLoS Comput. Biol.* **2010**, *6*, DOI: 10.1371/journal.pcbi.1000889.

[15] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, *Genome Res.* **2003**, *13*, 2498.

[16] I. H. Goenawan, K. Bryan, D. J. Lynn, *Bioinformatics* **2016**, *32*, 2713.

[17] P. V. Hornbeck, B. Zhang, B. Murray, J. M. Kornhauser, V. Latham, E. Skrzypek, *Nucleic Acids Res.* **2015**, *43*, D512.

[18] K. E. Bachman, P. Argani, Y. Samuels, N. Silliman, J. Ptak, S. Szabo, H. Konishi, B. Karakas, B. G. Blair, C. Lin, B. A. Peters, V. E. Velculescu, B. H. Park, *Cancer Biol. Ther.* **2004**, *3*, 772.

[19] B. Karakas, K. E. Bachman, B. H. Park, *Br. J. Cancer* **2006**, *94*, 455.

[20] A. G. Bader, S. Kang, P. K. Vogt, *Proc. Natl. Acad. Sci.* **2006**, *103*, 1475.

[21] J. A. Engelman, K. Zejnullahu, T. Mitsudomi, Y. Song, C. Hyland, J. O. Park, N. Lindeman, C.-M. Gale, X. Zhao, J. Christensen, T. Kosaka, A. J. Holmes, A. M. Rogers, F. Cappuzzo, T. Mok, C. Lee, B. E. Johnson, L. C. Cantley, P. A. Jänne, *Science* **2007**, *316*, 1039.

[22] N. Kohmura, T. Yagi, Y. Tomooka, M. Oyanagi, R. Kominami, N. Takeda, J. Chiba, Y.

Ikawa, S. Aizawa, *Mol. Cell. Biol.* **1994**, *14*, 6915.
[23] R. K. Goel, S. Miah, K. Black, N. Kalra, C. Dai, K. E. Lukong, *FEBS J.* **2013**, *280*, 4539.
[24] M.-W. Yoo, J. Park, H.-S. Han, Y.-M. Yun, J. W. Kang, D.-Y. Choi, J. won Lee, J. H. Jung, K.-Y. Lee, K. P. Kim, *PROTEOMICS* **2017**, *17*, 1600332.
[25] J. H. Trembley, G. Wang, G. Unger, J. Slaton, K. Ahmed, *Cell. Mol. Life Sci. CMLS* **2009**, *66*, 1858.
[26] I. M. Hanif, I. M. Hanif, M. A. Shazib, K. A. Ahmad, S. Pervaiz, *Int. J. Biochem. Cell Biol.* **2010**, *42*, 1602.
[27] S. Bose, F. L. L. Stratford, K. I. Broadfoot, G. G. F. Mason, A. J. Rivett, *Biochem. J.* **2004**, *378*, 177.
[28] B. N. Kholodenko, A. Kiyatkin, F. J. Bruggeman, E. Sontag, H. V. Westerhoff, J. B. Hoek, *Proc. Natl. Acad. Sci.* **2002**, *99*, 12841.
[29] Radulescu Ovidiu, Lagarrigue Sandrine, Siegel Anne, Veber Philippe, Le Borgne Michel, *J. R. Soc. Interface* **2006**, *3*, 185.
[30] M. Buffard, O. O. Ortega, C. F. Lopez, O. Radulescu, *JOBIM Meet. Abstr. A34* **2017**.

Figure legends

Figure 1. Workflow of the network construction and signal propagation analysis.

This workflow allows us to uncover potential signaling paths, from a kinase of interest to a list of proteins identified by phosphoproteomic experiments. Step1: Select pathways from KEGG and Pathway Commons databases. Step 2: Embed the selected pathways to create a prior-knowledge interaction network. Step 3: Search for paths from the source to its experimentally identified targets by a combination of weighted shortest paths and random walk methods. Step 4: Focus on the more biologically relevant paths to a subset of targets or to a unique target.

Figure 2. Identification of specific paths from the *PIK3CA* mutants to the receptor tyrosine kinase HER3.

The protein interaction networks are composed of nodes and edges. Nodes represent the proteins whose diamond or rounded rectangle shape correspond to the experimentally identified targets or to the proteins of the pathway databases, respectively. The edges of the networks represent the protein interactions whose target arrow shape corresponds to the sign of the interaction (Delta, positive interaction; T, negative interaction; Circle, unknown consequence).
(A) Alignment and comparison of the signaling networks of the E545H and H1047R PIK3CA mutants obtained from the quantitative phosphoproteomic comparison of each mutant with the control condition. The source of the signal (PIK3CA) is displayed in yellow. Red edges and nodes are specific for the E545H mutant. Green edges and nodes are specific for the H1047R mutant. White nodes and gray edges are common to both networks.
(B) Subnetwork of the signal propagation from the PIK3CA mutants to HER3 extracted from the signaling networks of the E545H and H1047R mutants obtained from the quantitative phosphoproteomic comparison of each PIK3CA mutant with the control condition.
(C) A subset of the near shortest paths from the H1047R to HER3 extracted from the signaling network of the H1047R mutants obtained from the quantitative phosphoproteomic differences between the E545H and H1047R mutants.

Figure 3. Reconstruction of the SRMS-signaling network by integration of multiple phosphoproteomics data

The protein interaction networks are composed of nodes and edges. Green nodes with rounded rectangle shapes represent the proteins experimentally identified as potential direct

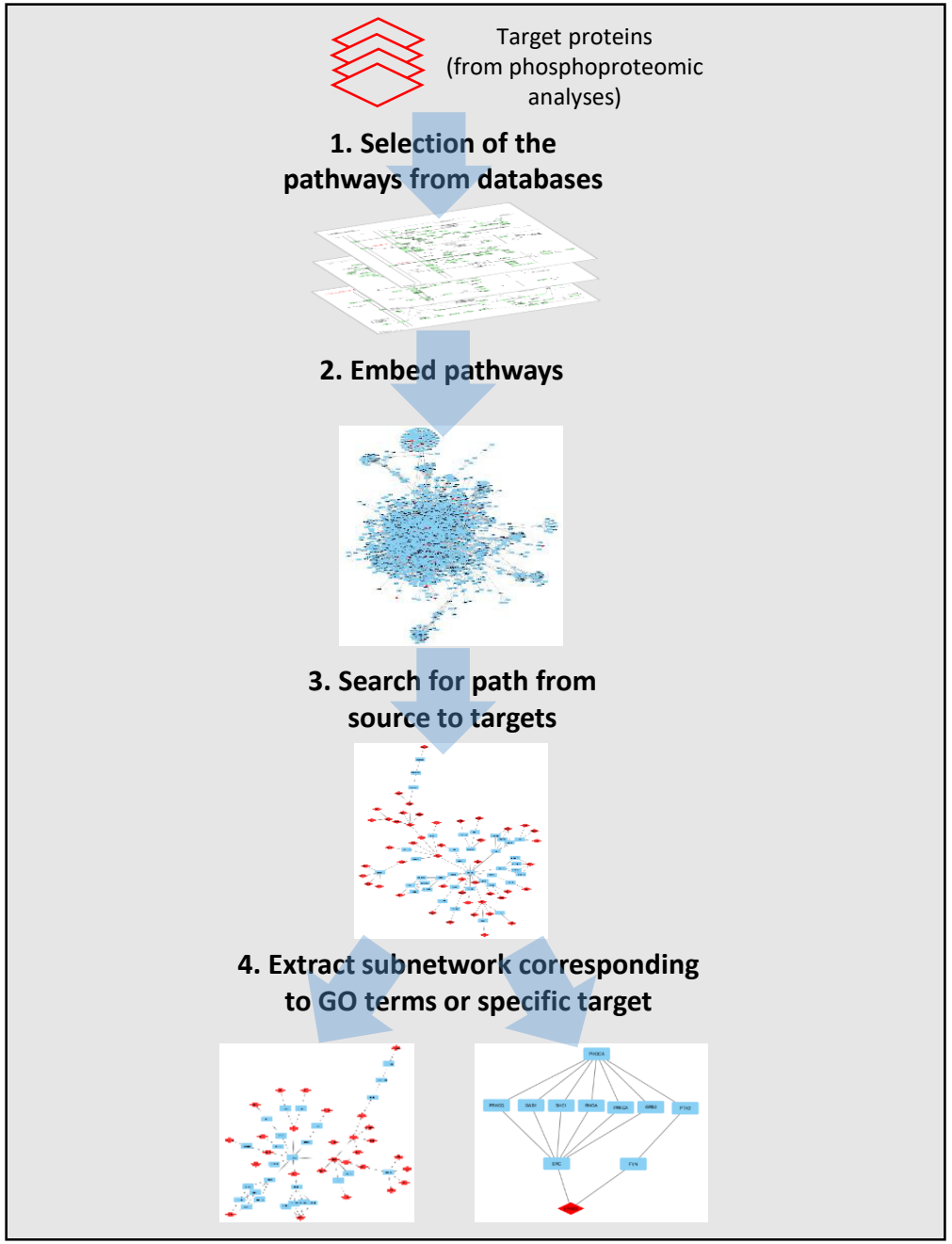
substrates of SRMS (phosphorylated on tyrosine residues). Red diamond-shaped nodes represent the experimentally identified indirect targets of SRMS (phosphorylated on serine/threonine residues). The edges of the networks represent the protein interactions whose target arrow shape corresponds to the sign of the interaction (Delta, positive interaction; T, negative interaction; Circle, unknown consequence).

(A) SRMS subnetwork isolated from the prior-knowledge network obtained from embedding the database pathways enriched in the list of SRMS-indirect targets.

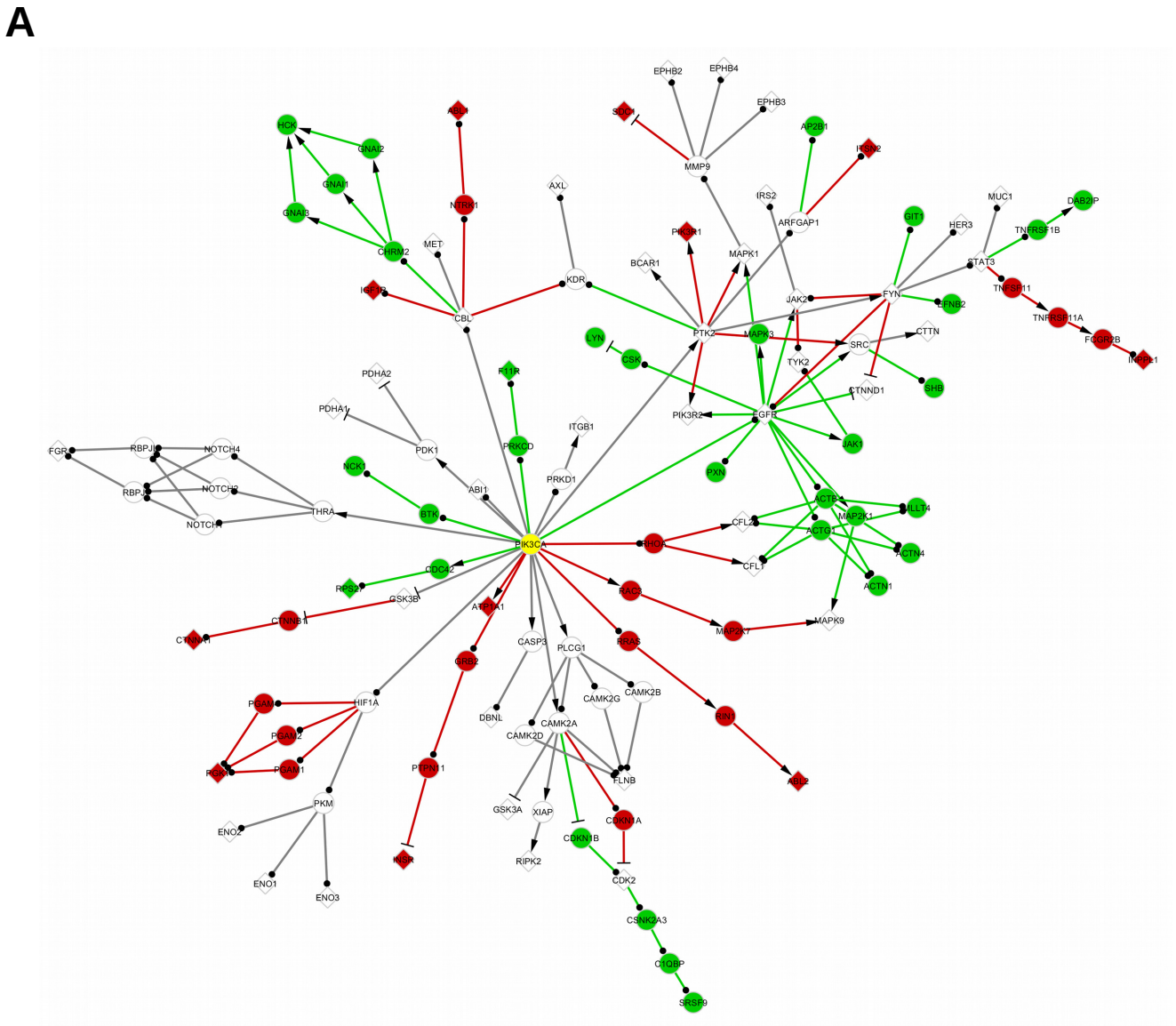
(B) Subnetwork of the signal propagation from SRMS to its direct substrates (green round rectangles) and to its indirect targets (red diamonds). This subnetwork is extracted from the prior-knowledge network enriched with the -direct SRMS substrates.

(C) Subnetwork of the signal propagation from SRMS to its direct substrates (green round rectangles) and to its indirect targets (red diamonds). This subnetwork is extracted from the prior-knowledge network enlarged to the components and interactions described in the public databases of all signaling pathways. The CK2 subunits are light blue in color.

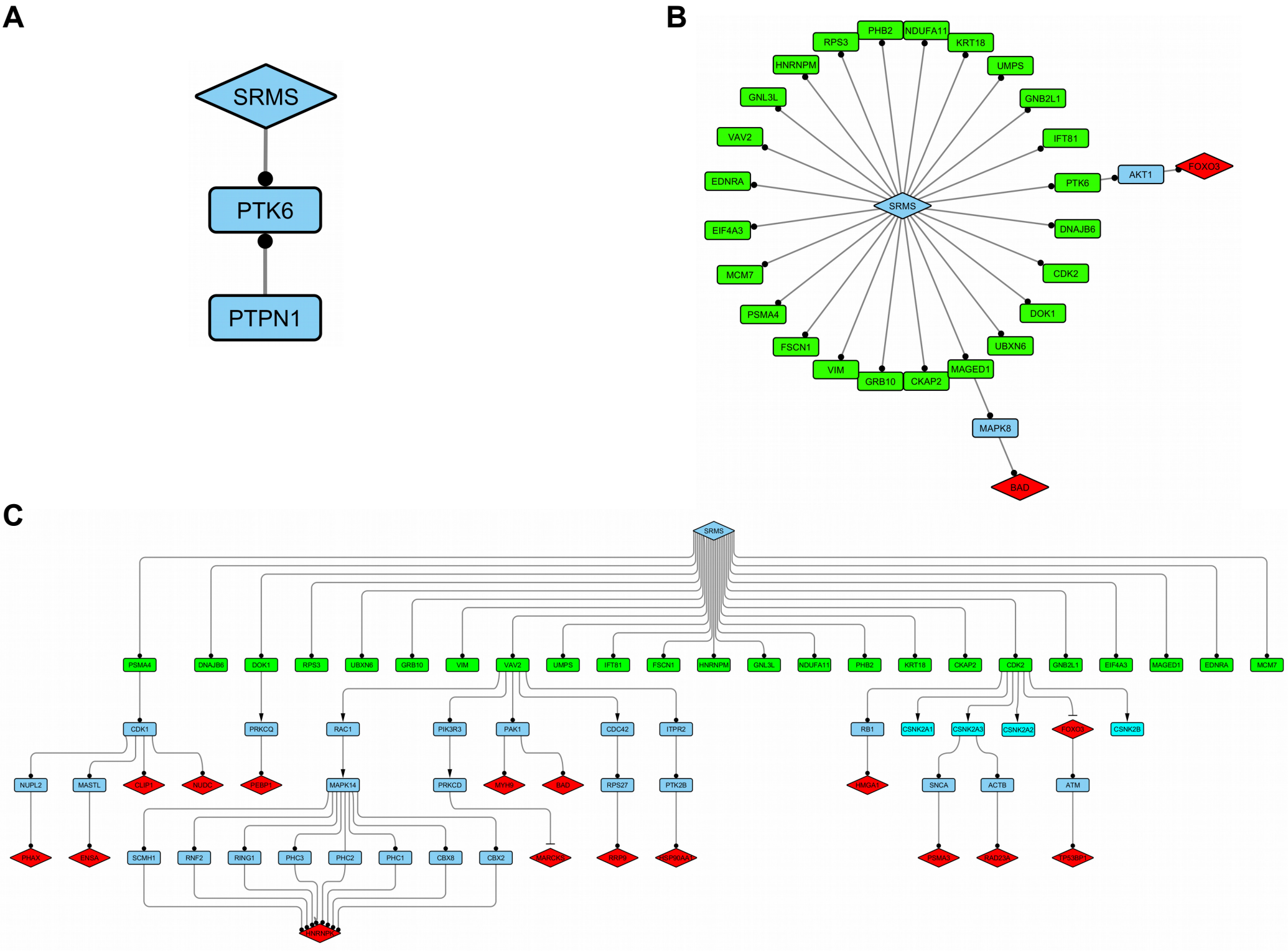
Buffard, figure 1.

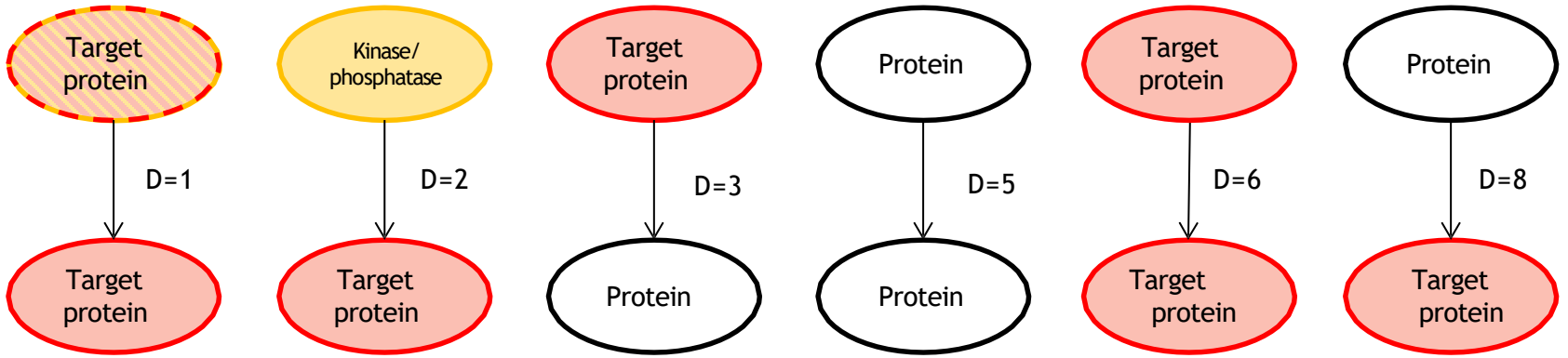


Buffard, Figure 2



Buffard, Figure 3



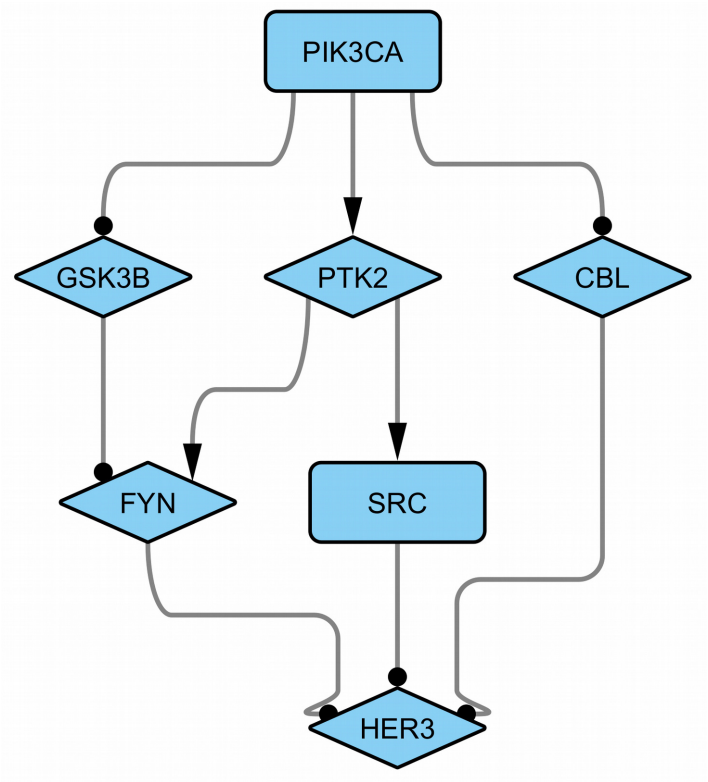


: Differentially phosphorylated proteins = target proteins

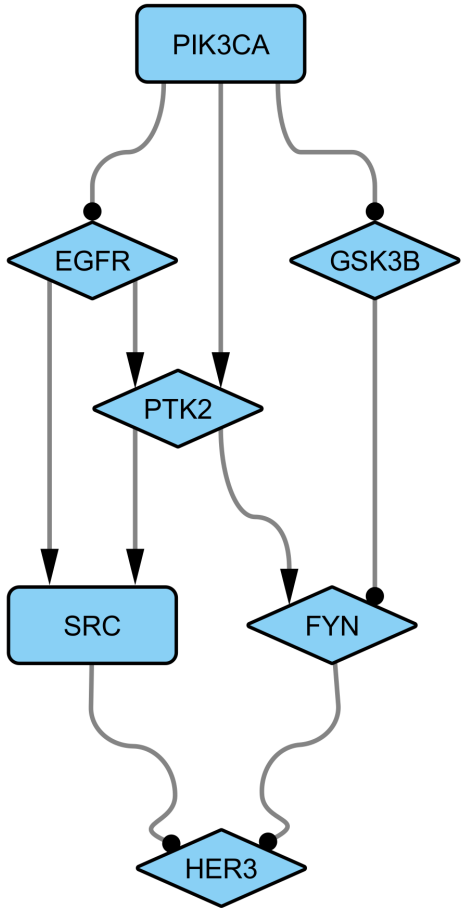
: Kinase or phosphatase protein

: Protein from databases not identified in dataset with no kinase or phosphatase activity

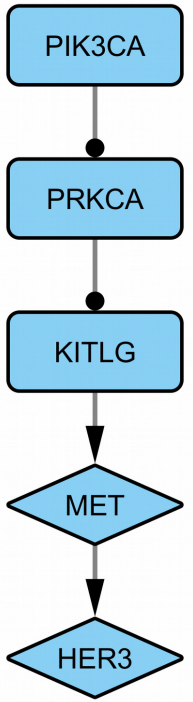
A


PIK3CA
GSK3B
PTK2
CBL
FYN
SRC
HER3

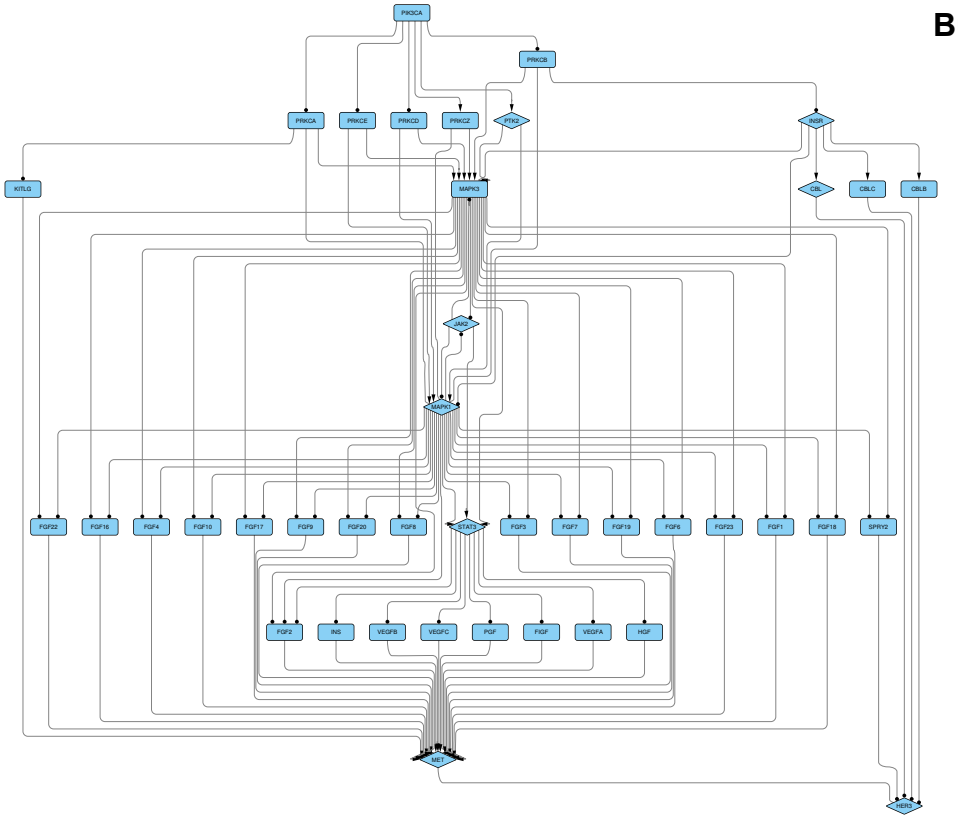
B


PIK3CA
EGFR
GSK3B
PTK2
SRC
FYN
HER3

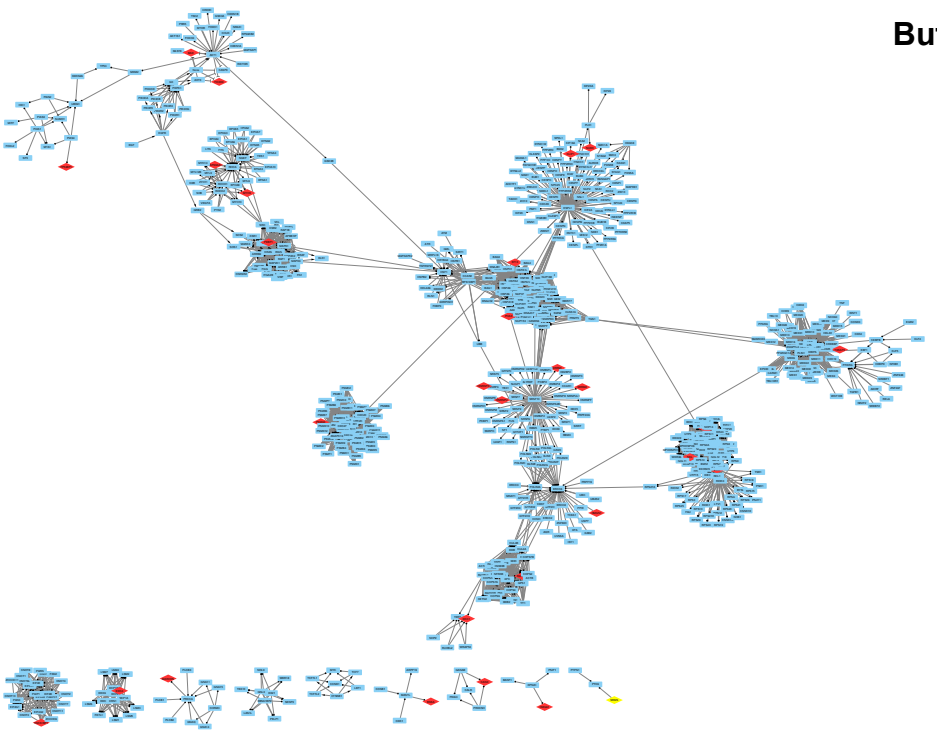
C


PIK3CA
PRKCA
KITLG
MET
HER3

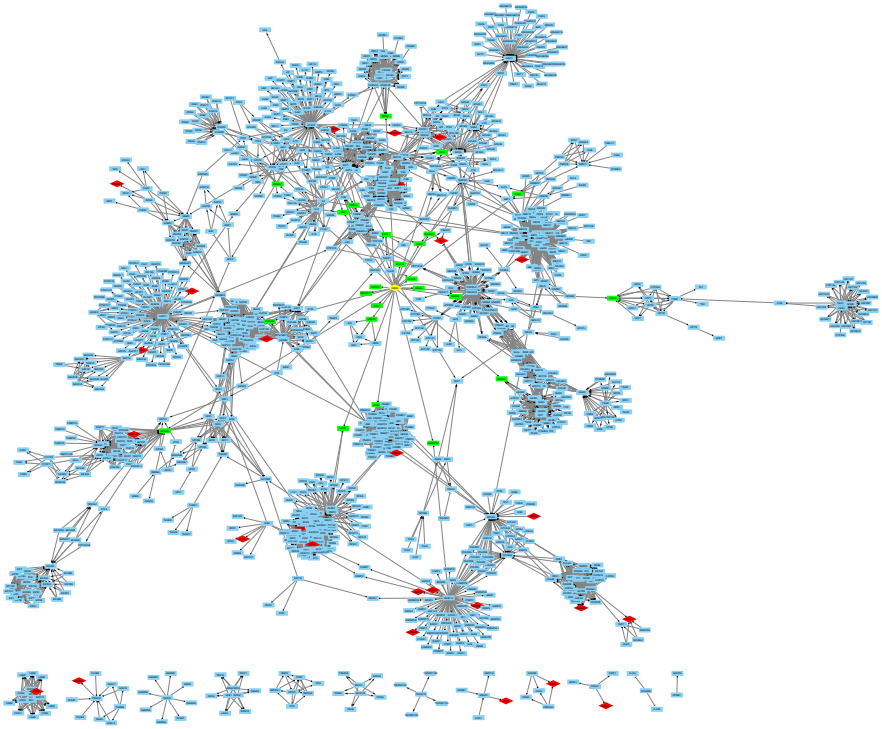
Buffard, Figure S3



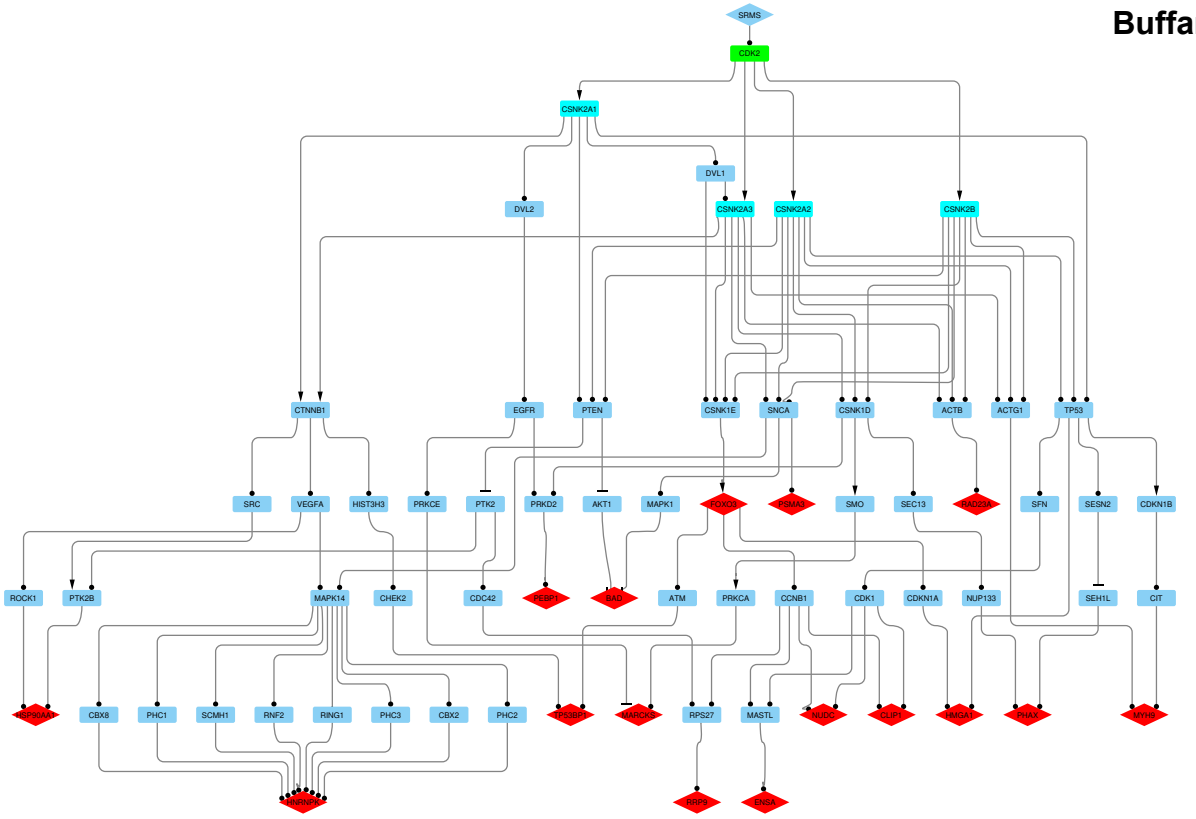
Buffard, Figure S4



Buffard, Figure S5



Buffard, Figure S6



Graphical User Interface

Inputs

Workflow

Buffard, figure S7

Phos2Net

Choose target file (with phosphorylated proteins) : ? Browse...

Choose output folder : Browse...

Choose database(s) : KEGG, Pathway Commons

Choose pathway's selection mode : ? Enriched pathways, All pathways

Add source direct substrates : ?

Enter source node :

Weight the edges, enter list of : ?

Kinases/phosphatases — Tyrosine kinase, Tyrosine phosphatase

Add specific proteins to promote

Add overflow (in % of the shortest path length) ? 0

Network extraction, enter list of : ?

Subset of target(s) :

GO terms associated categories : cell adhesion and motility, cell growth and death

Submit

a

b

c

d

e

Target proteins (from phosphoproteomic analyses)

1. Select pathways from databases

2. Embed pathways

3. Search for paths from source to targets

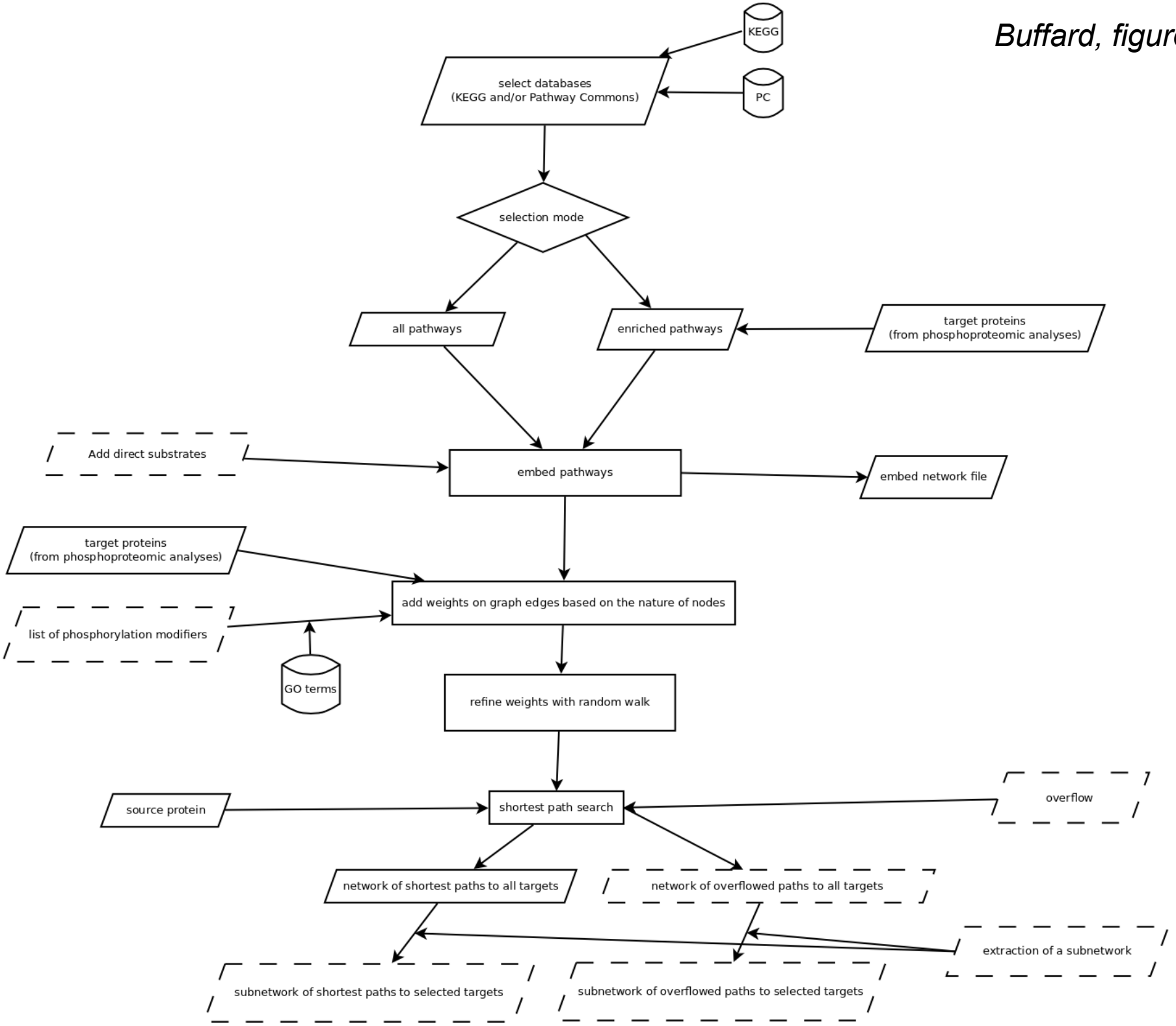
4. Extract subnetwork co to GO terms or specif

Outputs

f

Cytoscape readable and modifiable files

Buffard, figure S8



Supporting Information to Buffard, et al.

Supplementary Text

Graphical interface

In this section, we briefly present the different implemented options for the graphical interface (Suppl Figure 7). First, the user selects the data file (Browse button) containing the UniProt AC of the differentially phosphorylated proteins (step a) and the output folder. Then, the user chooses (1) the pathway databases by ticking the boxes allowing the choice of KEGG, Pathway Commons or both; (2) the pathway selection mode, by ticking radio buttons (the “all pathways” option will embed all pathways from the selected databases) (step b). To avoid the lack of connectivity of the source in the prior-knowledge network, we also included the possibility of directly adding protein interactions (e.g. with a kinase and its substrates). The protein components of these added interactions will be taken into account for the pathway selection only if the “enriched pathways” mode has been selected (step c). Next, the user sets all the parameters for the shortest path search from a source to the list of the differentially phosphorylated proteins: the source node (e.g., UniProt AC Q9H3Y6 for SRMS), the list of proteins for promoted edges (a personalized list and/or a selected pre-established list of kinase/phosphatase). The “overflow option” allows one to retrieve the near-shortest paths instead of the “strict” shortest paths to generate a set of alternatives, selecting all paths for which the total distance is up to xx% higher than that of the shortest path (step d). Finally, the “shortest path” extraction can be tuned by selecting a subset of targets and/or categories (regrouped GO terms) (step e). We have already included some GO term groups representing relevant processes and functions in cancer. All generated networks are able to be explored and manipulated using Cytoscape.

Legends to supplementary figures

Supplementary Figure 1. List of *ad hoc* weights of edges

An *ad hoc* weight is introduced on each edge based on the nature of the source and target node. “Normal” edges have a distance of 5 (d = 5), edges coming out of identified proteins (d = 3), edges reaching an

identified protein while coming out of a tyrosine kinase or phosphatase (d = 2) or combining these two conditions (d = 1). Edges reaching a target identified as differentially phosphorylated, but which did not come from a kinase or phosphatase (d = 8), even if they came out of another identified protein (d = 6).

Supplementary Figure 2. Identification of specific paths from the *PIK3CA* mutants to the receptor tyrosine kinase HER3.

The protein interaction networks are composed of nodes and edges. The nodes represent the proteins whose diamond or rounded rectangle shape correspond, respectively, to the experimentally identified targets or to the proteins of the pathway databases. The edges of the networks represent the protein interactions whose target arrow shape corresponds to the sign of the interaction (Delta, positive interaction; T, negative interaction; Circle, unknown sign).
(A) Subnetwork of the signal propagation from the E545H mutant to HER3, extracted from the signaling network obtained from the quantitative phosphoproteomic comparison between the E545H mutant and the control condition. This subnetwork contains all the near-shortest paths allowing a 20% overflow.
(B) Subnetwork of the signal propagation from the H1047R mutant to HER3, extracted from the signaling network obtained from the quantitative phosphoproteomic comparison between the H1047R mutant and the control condition. This subnetwork contains all the near-shortest paths allowing a 20% overflow.
(C) Subnetwork of the signal propagation from the H1047R mutant to HER3, extracted from the signaling network of this mutant obtained from the quantitative phosphoproteomic differences between the E545H and H1047R mutants.

Supplementary Figure 3. Signal propagation from the H1047R PIK3CA mutant to HER3.

Subnetwork of the signal propagation from the H1047R mutant to HER3, extracted from the signaling network of this mutant obtained from the quantitative phosphoproteomic differences between the E545H and H1047R mutants. This subnetwork contains all the near-shortest paths allowing a 20% overflow.
The protein interaction networks are composed of nodes and edges. Nodes represent the proteins whose diamond or rounded rectangle shape correspond, respectively, to the experimentally identified targets or to the proteins of the pathway databases. Edges of the networks represent the protein interactions whose target arrow shape corresponds to the sign of the interaction (Delta, positive interaction; T, negative interaction; Circle, unknown consequence).

Supplementary Figure 4. SRMS prior-knowledge network obtained from embedding the database pathways enriched in the list of SRMS indirect targets.

The protein interaction networks are composed of nodes and edges. The red diamond-shaped nodes represent the experimentally identified indirect targets of SRMS (phosphorylated on serine/threonine residues). The edges of the networks represent the protein interactions whose target arrow shape corresponds to the sign of the interaction (Delta, positive interaction; T, negative interaction; Circle, unknown consequence).

Supplementary Figure 5. SRMS prior-knowledge network enriched with the SRMS-indirect targets and the SRMS direct substrates.

The protein interaction networks are composed of nodes and edges. The green rounded rectangle-shaped nodes represent the proteins experimentally identified as potential direct tyrosine-phosphorylated SRMS substrates. Red diamond-shaped nodes represent the experimentally identified indirect targets of SRMS (phosphorylated on serine/threonine residues). Edges of the networks represent the protein interactions whose target arrow shape corresponds to the sign of the interaction (Delta, positive interaction; T, negative interaction; Circle, unknown consequence).

Supplementary Figure 6. Subnetwork of the paths from SRMS to its indirect targets and through the four CK2-subunits.

The protein interaction networks are composed of nodes and edges. The green rounded rectangle-shaped nodes represent the proteins experimentally identified as potential direct tyrosine-phosphorylated SRMS substrates. The red diamond-shaped nodes represent the experimentally identified indirect targets of SRMS (phosphorylated on serine/threonine residues). The CK2 subunits are light blue in color. Edges of the networks represent the protein interactions whose target arrow shape corresponds to the sign of the interaction (Delta, positive interaction; T, negative interaction; Circle, unknown consequence).

Supplementary Figure 7. Graphical interface

Outline of the graphical interface steps and options. Steps (a) to (e) must be entered by the user as inputs for the different corresponding steps of the workflow (black arrows). Step (f) represents the workflow output results, creating files corresponding to the different subnetworks created by the user's entries and choices (in step (e)). These files can be visualized and manipulated with Cytoscape (cytoscape.org).

Supplementary Figure 8. Diagrammatic representation of the algorithm

The algorithm is dependent on the user input. Input and output are represented by parallelogram, alternatives by diamond. Optional inputs and outputs are represented by dashed lines.