Using Multilayer Heterogeneous Networks to Infer Functions of Phosphorylated Sites

Joanne Watson\(^1\), Jean-Marc Schwartz\(^1\) and Chiara Francavilla\(^2\)*,

* Corresponding Author: chiara.francavilla@manchester.ac.uk

\(^1\)Division of Evolution & Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine & Health, University of Manchester, Manchester, M13 9PT

\(^2\)Division of Molecular and Cellular Function, School of Biological Sciences, Faculty of Biology, Medicine & Health, University of Manchester, Manchester, M13 9PT.

Running Title: Phosphorylation Site-centric Functional Analysis
Abbreviations

RWHN – Random walk on heterogeneous network
PTM – Post-translational modification
PPI – Protein-protein interaction
GO – Gene Ontology
PSP – Phosphosite Plus
EGF - Epidermal Growth Factor
TGF-α – Tumour growth factor α
RWR – Random walk with restart
GOBP – Gene Ontology Biological Processes
FDR – False discovery rate
SILAC – Stable Isotope Labelling of Amino Acids in Cell culture
KDE – Kernel Density Estimation

Abstract

Mass spectrometry-based quantitative phosphoproteomics has become an essential approach in the study of cellular processes such as cell signaling. Commonly used methods to analyze phosphoproteomics datasets depend on generic, gene-centric annotations such as Gene Ontology terms and do not account for the function of a protein in a given phosphorylation state. Thus, analysis of phosphoproteomics data is hampered by a lack of phosphorylated site-specific annotations. Here, we propose a method that combines shotgun phosphoproteomics data, protein-protein interactions and functional annotations from ontologies or pathway databases into a heterogeneous multilayer network. Phosphorylation sites are then associated to potential functions using a random walk on heterogeneous network (RWHN) algorithm. We validated our approach using a dataset modelling the MAPK/ERK pathway and were able to associate differentially regulated sites on the same protein to their previously described functions. Random permutation analysis proved that these associations were not random and were determined by
the network topology. We then applied the RWHN algorithm to two previously published datasets; the 
algorithm was able to reproduce the experimentally validated conclusions from the publications, and 
associate phosphorylation sites with both new and known functions based on their regulatory patterns. 
The approach described here provides a robust, phosphorylation site-centric method to analyzing 
phosphoproteomics data and identifying potential context-specific functions for sites with similar 
phosphorylation profiles.

1. Introduction

Phosphorylation is the most studied post-translational modification (PTM) due to its central role in 
cellular regulation. It is thought to be the principal PTM in the human proteome with upwards of 100,000 
predicted phosphorylation sites [1], as well as an essential mediator of protein-protein interaction (PPI) 
and protein function. Transient changes occur at specifically regulated phosphorylation sites, of which 
there may be multiple on each protein. Regulation of phosphorylation is often dependent on 
perturbations such as the activity of extracellular ligands, drug treatment or physical stimuli in the 
extracellular environment [2]. By comparing changes in the phosphoproteome of cells in different 
experimental conditions through mass spectrometry-based phosphoproteomics, phosphorylated sites 
that are key players in cellular processes and functions can be uncovered in an unbiased, high-throughput 
manner [3].

Functional analysis of phosphoproteomics datasets is typically based on gene-centric enrichment of Gene 
Ontology (GO) terms or involvement in known pathways [4]. However, this approach disregards 
information captured by phosphoproteomics data on changes at specific phosphorylated sites, by limiting 
the analysis to the protein level. The modification state of a protein is inherently coupled to its function; 
PTMs alter protein activity, as well as the ability to interact with different sets of proteins. Furthermore, if 
a protein is phosphorylated on multiple sites, each with a different function and regulatory pattern, this 
information is not revealed by gene-centric analysis [5]. For instance, the well-studied signalling protein 
MAPK1 has 18 known phosphorylated sites recorded in the database PhosphositePlus (PSP), however 
only 6 have been annotated to a downstream function[6]. Enrichment analyses that rely on
generalizations based on protein-level or gene-centric descriptions exclude details encoded in the phosphorylation signature. Analyses are thus hampered by the lack of phosphorylation site-specific functional annotations.

Several network-based methods have been proposed to move towards phosphorylation site-specific analyses of phosphoproteomics data. A significant focus has been on the inference of kinase-substrate networks, such as NetworKIN [7], KSEA [8], and IKAP [9], among others. These methods are useful for reconstructing the architecture of signalling intracellular networks, which can be informative for identifying modules of modified proteins involved in cellular processes, but again remain hampered by the lack of site-specific functional annotation [10]. They may also be biased towards the most studied kinases and the exclusion of non-kinase proteins. Rudolph et al. [11] proposed a method to address this issue named PHOTON which identified differentially regulated proteins based on the level of phosphorylation of their binding partners in a high-confidence PPI network, and then used logistic regression to identify the involvement of the phosphorylated proteins in known signalling pathways.

Krug et al. [12] curated a resource containing literature-derived phosphorylated site-specific signatures to assign functions to sites. However, the use of such a resource is limited by the lack of annotation of these signatures to specific cellular processes.

In recent years, heterogeneous or multilayer networks have been used to represent many types of 'omics datasets [13]. These specialized networks are used to describe multiple types of associations with nodes representing different entities. To identify relationships between the different biological layers, random walk algorithms have been applied to these networks. The random walk on heterogeneous network (RWHN) with restart (RWR) method [14] has been particularly popular. Jiang [15] used RWR to prioritize disease candidate genes in a PPI-phenome network; similarly, Soul et al. [16] applied it to a PPI-phenome network to identify disease mechanisms. Recent work has extended the method to multiple layers of biological information, for example to infer disease associated m6A RNA methylation via known gene-disease associations [17]. Similar methodology was used to associate phosphorylation sites recorded in the PSP database to diseases via kinase-substrate interactions [18].
Here, we propose an algorithm that uses RWHN to associate phosphorylated sites to context-specific function via a heterogeneous multilayer network using shotgun phosphoproteomics data. The network combines three layers of information: phosphorylated sites, protein interactions and GO terms. Clustering of phosphoproteomics data is used to find common features within datasets and is generally followed by enrichment analyses. This is based on the assumption that common patterns of phosphorylation based on temporal changes or in response to a particular stimulus, treatment or environmental context, are a likely indicator of involvement in common functions. We utilize this concept in our algorithm by connecting phosphorylated sites that have been clustered together, and therefore share regulatory patterns, within the multilayer network.

To prove the utility of our algorithm, we apply it to a manually-curated validation network and two previously published datasets, describing early signalling events in HeLa cells upon EGF and TGF-α stimulation [19] and phosphorylation-mediated changes in breast cancer cells resistant to lapatinib treatment [20]. We demonstrate that functional annotations can be differentially assigned to phosphorylated sites and this is driven by changes in the context-dependent modification of these sites.

Our method is suitable for use with any phosphoproteomics dataset and could be generalized for data describing other PTMs.

2. Experimental Procedures

A multilayer heterogeneous network was constructed to relate biological function to phosphorylation site via a protein-protein interaction network (Figure 1). Three types of nodes are contained within the network: phosphorylation sites (called ‘sites’ here on in for brevity), proteins and functional annotations (either Gene Ontology Biological Process (GOBP) terms or KEGG pathways). The edges describe five possible associations which are either bipartite (i.e. site to protein, protein to function) or between the same type of node.

2.1 Francavilla et al. Dataset
The results of the phosphoproteomics experiment described in Francavilla et al. [19] were taken from Supplementary Table S1 of their manuscript. The authors had processed the raw mass spectrometry data using MaxQuant [21], filtered identified phosphorylated sites based on a localization probability of greater than 0.75 and normalized remaining data to the ratio of EGFR at 1 min after stimulation with EGF or TGF-α. Prior to multilayer heterogeneous network construction, the data were further filtered to remove sites with two or more missing ratios in their time series and containing at least one SILAC ratio higher than 2 or lower than 0.5 (as described in the methods of Francavilla et al. [19]). Missing data were imputed with random draws from a truncated distribution, as previously described [22], using the impute.QRLIC function from the imputeLCMD R package.

2.2 Ruprecht et al. Dataset

Data from the Ruprecht et al. [20] study were taken from the Supplementary Data file named 'Filtered and normalized phosphoproteome dataset'. The raw mass-spectrometry data had been processed using MaxQuant, filtered to include only identified phosphorylated sites with a localization probability of greater than 0.75, and normalized. Sites that did not show a significant change (FDR <1%) either between the untreated parental and lapatinib-treated resistant (SILAC ratios H/L) or between the lapatinib treated parental and lapatinib-treated resistant conditions (SILAC ratios H/M) experimental conditions were filtered from the data, as described in the methods of Ruprecht et al. [20]. Missing data were imputed with the same method used for the Francavilla et al. dataset [22][22][22][22].

2.3 Construction of Multilayer Network

Phosphorylation Site – Phosphorylation Site Subnetwork

Phosphorylated sites were clustered based on their quantitative values in different experimental conditions using the Fuzzy C-Means [23] or k-means method. Edges were drawn between sites in the same cluster that had a Pearson correlation (R2) between all the data points greater than or equal to 0.99 (Figure 1).

Protein - Protein Subnetwork
The protein-protein interaction network was extracted from STRING [24]. All interactions to the proteins included in the phosphoproteomics datasets with an experimental confidence score of greater than 0.4 were included (Figure 1).

**Function – Function Subnetwork**

We used either GO BP terms or KEGG pathways as functional annotators in this work. If GO BP terms were included in the multilayer network, the GO SemSim package from Bioconductor [25] was used to calculate the semantic similarity of enriched GO BP terms. An edge was drawn between terms with semantic similarity greater than 0.7, as calculated using the Wang method. In the case of KEGG pathways, edges were drawn between pathways that had greater than 70% pairwise similarity in their functional annotation profiles, following the method described in Stoney *et al.* [26] (Figure 1).

**Phosphorylation Site – Protein Bipartite Edges**

Sites and proteins had an edge between them if the residue was found on that protein. Therefore, sites will only have one edge, but proteins will have edges to all the sites found on that protein that were differentially regulated in the data set (Figure 1).

**Protein – Function Bipartite Edge**

We assumed that closely connected nodes in the protein-protein subnetwork would more likely be involved in similar biological processes. Therefore we computed modules of the protein-protein subnetwork using the Louvain method. Proteins from each module were analyzed for enrichment (FDR < 0.05%) of functional annotations (either GO BP terms or KEGG pathways) using the *Enrichr* [27] R interface. This increased the specificity of terms to be included in the network. When GO terms were included, high frequency (annotated to more than 5% of genes) and semantically redundant terms were filtered using the Bioconductor GO SemSim [25] and GO.db packages [28] (Figure 1).

**2.4 Random Walk on Heterogeneous Network**
The heterogeneous network can be represented as an adjacency matrix as follows:

\[
A = \begin{bmatrix}
A_{RR} & A_{RP} & 0 \\
A_{PR} & A_{PP} & A_{PF} \\
0 & A_{FP} & A_{FF}
\end{bmatrix}
\]  

(1)

where: \(A_{RR}\) are site-site associations, \(A_{PP}\) are protein-protein associations, \(A_{FF}\) are function-function associations, \(A_{RP}\) are site-protein associations with \(A_{PR}\) as the transpose and \(A_{PF}\) are protein-function associations with \(A_{FP}\) as the transpose.

As described in previous work [14], a transition matrix (\(M\)) was calculated for use in the first stage of the algorithm:

\[
M = \begin{bmatrix}
M_{RR} & M_{RP} & 0 \\
M_{PR} & M_{PP} & M_{PF} \\
0 & M_{FP} & M_{FF}
\end{bmatrix}
\]  

(2)

The bipartite inter-subgraph transition matrices (\(M_{RP}\) and \(M_{PF}\)) were calculated as:

\[
(M_{RP})_{l,j} = \begin{cases} 
\lambda (A_{RP})_{l,j} \sum_j (A_{RP})_{l,j}, & \text{if } \sum_j (A_{RP})_{l,j} \neq 0 \\
0, & \text{otherwise}
\end{cases}
\]  

(3)

\[
(M_{PF})_{l,j} = \begin{cases} 
\lambda (A_{PF})_{l,j} \sum_j (A_{PF})_{l,j}, & \text{if } \sum_j (A_{PF})_{l,j} \neq 0 \\
0, & \text{otherwise}
\end{cases}
\]  

(4)

where \(\lambda\) is the transition probability (i.e. the likelihood of the walker moving between two layers of the network).

The intra-subgraph transition matrices (\(M_{RR}, M_{PP}\) and \(M_{PP}\)) were calculated as:
\[
(M_{RR})_{i,j} = \begin{cases} 
(A_{RR})_{i,j} / \sum_j (A_{RR})_{i,j}, & \text{if } \sum_j (A_{RP})_{i,j} = 0 \\
(1 - \lambda) (A_{RR})_{i,j} / \sum_j (A_{RR})_{i,j}, & \text{otherwise}
\end{cases}
\]

\[
(M_{PF})_{i,j} = \begin{cases} 
(A_{PF})_{i,j} / \sum_j (A_{PF})_{i,j}, & \text{if } \sum_j (A_{RP})_{i,j} = 0 \\
(1 - \lambda) (A_{PF})_{i,j} / \sum_j (A_{PF})_{i,j}, & \text{otherwise}
\end{cases}
\]

\[
(M_{FF})_{i,j} = \begin{cases} 
(A_{FF})_{i,j} / \sum_j (A_{FF})_{i,j}, & \text{if } \sum_j (A_{FP})_{i,j} = 0 \\
(1 - \lambda) (A_{FF})_{i,j} / \sum_j (A_{FF})_{i,j}, & \text{otherwise}
\end{cases}
\]

RWHN is a ranking algorithm; nodes are ranked based on the probabilities of finding the random walker at a given node in the steady state, having started at a given seed node or set of seed nodes. In this work we set the seed nodes to be those sites belonging to a particular cluster. The probability of finding the random walker at each node for each step is calculated based on the iterative equation:

\[
P_{s+1} = (1 - r) \times M^T \times P_s + r \times P_0
\]

where \( r \) is the restart probability (set to 0.7 as described in Kohler et al. [29]), \( P_0 \) is the initial probability vector and \( P_s \) is the probability vector at step \( s \). \( P_0 \) was calculated such that all seed nodes were given equal probabilities with their sum equal to 1. All other nodes in the site-site subnetwork were assigned an initial probability of 0. Nodes in other subnetworks were assigned equal probabilities with their sum equal to 1 and weighted with the tunable parameters \( \eta_P \) and \( \eta_F \), that are used to weight the influence of each layer.
The output of the algorithm is a ranked list of all the nodes, based on the probability of finding the
random walker at each node in the steady state. This list is filtered by retaining functional annotations
and removing proteins and sites.

We implemented the algorithm in R using packages available from Bioconductor and CRAN and the
source code is available at www.github.com/jowatson2011/RWHN_phosphoproteomics. RWHN on the
validation network took less than 30 seconds to run however the larger experimental data sets took
several hours on a moderately powerful computer (32GB RAM, Intel i7 Processor).

3. Results

3.1 Overview of Algorithm

To associate phosphorylated sites of unknown function to potential cellular functions we developed an
algorithm to apply to shotgun phosphoproteomics data. First, in order to identify context-specific roles of
proteins based on the regulation of their phosphorylated sites, a multilayer heterogeneous network is
constructed. This represents three layers of biological entities and information: phosphorylated sites,
proteins-protein interaction, and biological function (Figure 1). We then apply a ranking algorithm,
RWHN, which ranks nodes of each layer based on: (i) distance from the phosphorylated sites of interest,
which are assigned as ‘seed’ nodes and (ii) the topology of the multilayer heterogeneous network [14].
Functions that are highly ranked can be considered more correlated with a set of seed nodes.

In the multilayer heterogeneous network, edges are drawn between sites based on similarity in the
regulatory pattern of their phosphorylation which we determine using k-means or fuzzy C-means
clustering. For the protein layer, a PPI network is constructed of the phosphorylated proteins and their
interactors using the STRING database [24]. Interactors are included to account for non-phosphorylated
proteins, or those below the limit of detection of the experiment [30]. In order to get a comprehensive
and specific list of functional annotations, we calculated closely connected groups of proteins in the PPI
network using module detection and performed functional enrichment analysis on these modules. Edges
between functional annotations terms are drawn based on functional similarity and overlap [25,26].
The RWHN algorithm simulates a walker moving from a starting node(s) (called a seed) and then node to node through the multilayer network. Each step is influenced by the probability of transition to another layer \( \lambda \), the weighting of the protein and function layers \( \eta_P \) and \( \eta_F \) and the probability of restart (that is, teleportation back to the seed node(s), r). The output of RWHN consists in a list of ranks for all the nodes in the network based on the likelihood of the walker reaching that node.

To optimize the algorithm, we ran RWHN with each of these parameters \( (\lambda, r, \eta_P \text{ and } \eta_F) \) tested over a range of values \( (0.2 - 0.8) \), changing one whilst setting all others to 0.5 (Figure S1). Performance was decidedly stable over the range of \( \eta_P \) and \( \eta_F \) however altering \( \lambda \) and \( r \) resulted in a spread of ranks assigned to a term depending on the parameter value. Previous work has suggested an ideal \( \lambda \) and \( r \) of 0.7 \cite{14,29}, whilst \( \eta_P \) and \( \eta_F \) were set to 0.7 and 0.3 respectively to prioritize movement in the protein layer and reduce the number of terms having the same rank regardless of seed.

### 3.2 Validation Data Case Study

MAPK/ERK signalling has been well studied and the temporal phosphorylation status of component proteins in response to growth factor stimulus are relatively established \cite{31,32} (Figure 2A). Based on this general understanding, we compiled a simple validation dataset of phosphorylation dynamics at particular sites of the main signalling proteins within the pathway (Figure 2B). Sites were chosen based on whether their phosphorylation is known to activate or inactivate protein activity, as recorded in the PSP database \cite{6}. We clustered the data into 5 clusters, referred to here as clusters 1-5, using the fuzzy c-means algorithm (Figure 2B). The number of clusters was selected based on the number of temporal trends we could visually identify in the data. The features of the multilayer network that was constructed are summarized in Table 1. Enrichment of GOBP terms was used to form the functional annotation layer of the network.

We ran the RWHN algorithm over the multilayer heterogeneous network with seed nodes set to all the sites belonging to one of the clusters; this was then repeated for each cluster (Figure 3A). The highest ranked GOBP term was the same (“Protein autophosphorylation”) regardless of the seed nodes; however, it was possible to differentiate between the five clusters based on the ranking of terms below the first one.
For example, the inactivating phosphorylation sites of MAP2K1 and MAP2K2 (S212 and S216 respectively) are clustered separately to the activating sites (MAP2K1_S218, MAP2K1_S222, MAK2K2_S222 and MAP2K2_S226), in clusters 1 and 3 respectively. The ranking of the term “regulation of mitotic cell cycle” is higher for cluster 1 than cluster 3. This is consistent with annotation in PSP, showing that whilst activating MAP2K1/MAP2K2 sites are positive regulators of the cell cycle, MAP2K1 and MAP2K2 are negative regulators of the cellular behaviors downstream of the activating phosphorylation. Another example is the highly specific term “regulation of cysteine-type endopeptidase activity involved in apoptotic process”, which is ranked highest when the seed nodes are set to cluster 2. The process is annotated to two of the proteins in this cluster (RPS6KA3 and RAF1) as well as other well-studied signalling kinases found in the PPI layer of the multilayer heterogeneous network (SRC, AKT1, MEK5, BAD and MAP3K5). Higher association with this cluster could indicate a particular role for RAF1-S289 in this process, compared to the other RAF1 phosphorylated sites included in this network.

To assess the robustness of the algorithm, we performed a random permutation control. RWHN was run one hundred times with the seeds set as before, random permutations of the subnetworks and bipartite edges maintained. Kernel Density Estimation (KDE) was calculated to assess how often each GO term occurred at each rank in the 100 random permutations for each set of seed nodes. In each case, there were a subset of terms that were more likely to have a rank greater than 50, however none strongly correlated with the actual rankings for each set of seed nodes (Figure 3B). This confirms that the rankings are not random, but primarily determined by the network topology.

For a given set of seed nodes, the algorithm is capable of giving high ranks to functions with known associations to those sites. It is also capable of predicting non-random, reasonable functions for sites of unknown function in this context.

3.3 Experimental Case Study: Dissecting EGF- and TGF-α-Induced Dynamic Phosphorylation

We next wanted to test the performance of the algorithm on an already available and complex phosphoproteomics dataset. The phosphoproteomics data retrieved from Francavilla et al. [19] describes the effect in HeLa cells of stimulation with EGF or TGF-α over a period of 90 minutes (with time points at
In this study, one of the findings was that the two ligands activating the same cell-surface receptor, regulated phosphorylation dynamics in a different manner, resulting in opposite cellular responses. Phosphorylated sites from Supplementary Table S1 of the original publication were filtered based on regulation by EGF or by TGF-α and divided based on which stimuli the regulation was dependent on. The separated data were clustered into 6 clusters using the fuzzy C-means method, as per the original publication, with each cluster representing a distinct dynamic profile of phosphorylation (Figure S2). A multilayer heterogeneous network was constructed for both sets of data[19][19][19], described in Table 2 and Table 3 respectively.

Several terms that were biologically relevant were differentially ranked in the EGF and TGF-α networks (Figure 4A-B). Consistent with the original publication, the term ‘regulation of ERK1 and ERK2 cascade’ was highly ranked when seeds were set to sites in cluster 3 and 5 (representing EGF transient responders) for the EGF network and the term ‘regulation of MAPK cascade’ for clusters 4 and 5 (representing TGF-α cycling or sustained responders) for the TGF-α network. We used these terms as positive controls, due to the process being well studied and many sites within the cascade having previously documented function we could use to verify the robustness of our approach.

EGF cluster 3 contains two EGFR sites described as activating in the PSP database (EGFR_Y1197, EGFR_Y1110), as well as sites belonging to EGFR downstream proteins SHC1 (SHC1_Y427, SHC1_Y349, SHC1_Y350) and PTPN11 (PTPN11_Y584). Cluster 5 contains MAPK1 activating sites (MAPK1_T185, MAPK1_Y187), serving as a positive control for this association, and two sites associated with EGFR internalization in PSP (EGFR_S991, EGFR_S693). Interestingly, cluster 1 (representing EGF cycling responders) contains a tyrosine residue that is also associated with regulation of EGFR internalization (EGFR_Y998) as well as the RAB7 regulatory site experimentally investigated in the original publication (RAB7_Y183) and MAPK3 activating sites (MAPK3_T202, MAPK_Y204). When seed nodes are set to cluster 1 there are higher proportion of terms related to protein localization highly ranked. This potentially indicates nuanced roles between tyrosine and serine/threonine residues on EGFR and MAPK1/MAPK3 activation in receptor internalization. For the TGF-α network, only one site (EGFR-Y1197, activating) was found in one of the two TGF-α clusters associated with MAPK regulation (cluster...
4). RAB7A_Y183 and RAB11B, important regulators of the receptor recycling investigated in the paper, were also found in this cluster. When seed nodes were set to this cluster, terms associated with cell migration were ranked higher (e.g. ‘positive regulation of cellular migration’). This was in line with one of the conclusions of the paper, demonstrating a role for recycling and migration upon TGF-α stimulation.

When seed nodes were set to sites in cluster 3, the term ‘regulation of receptor mediate endocytosis’ was ranked high; this cluster contained EGFR sites Y998 and Y1172, both associated to receptor internalization in PSP, and SH3BP4_S279, a known player in this process [33]. Taken together, the functions that were highly ranked were related to the biological functions of the sites set as seed nodes.

The algorithm captured the experimentally verified conclusion from the original paper, where EGFR stimulated by TGF-α is associated with receptor internalization/recycling and a strong migratory phenotype, whilst EGFR stimulated with EGF is degraded with fewer potent changes in cell behavior.

To consider a single protein example, phosphorylated sites on RAF1 were clustered for the EGF data depending on whether they are activating (S289, S301 in cluster 6, S296 in cluster 2) or inactivating (S259, a 14-3-3 binding site [34], in cluster 5) sites, as annotated in PhosphositePlus (marked in Figure S2). In the EGF network, seeds nodes set as clusters 2 or 6 resulted in ‘positive regulation of cellular process’ being ranked higher than seed nodes as cluster 5 sites. These clusters also were able to rank ‘regulation of PI3K signalling’ and ‘regulation of protein kinase B signaling’ highly, picking up on RAF1 crosstalk with other pathways when active [35,36]. In particular, the cluster containing the sites RAF1_S289 and S301 ranked ‘regulation of lipid kinase activity higher’. Cluster 5 ranked ‘regulation of ERK1 and ERK2 cascade’ higher than cluster 2 and 6, however it also counterintuitively ranked ‘positive regulation of MAPK cascade’ higher.

The TGF-α data were also clustered so that RAF_S296 was separated (cluster 3) from the other activating sites (cluster 5), but the negatively regulating site was not quantified in this experimental condition. As in the EGF network, cluster 5 ranked ‘inositol lipid-mediated phosphorylation’ and ‘regulation of lipid kinase activity’ higher than cluster 3 did, indicating a stronger correlation between RAF1_S289 and RAF1_S301 phosphorylation and lipid kinase pathway crosstalk. These results suggest the algorithm is capable of assigning functional terms to differentially regulated phosphorylation site and validates that it can distinguish between sites of known function.
3.4 Experimental Case study: Phosphorylation-dependent response to Lapatinib treatment and resistance in breast cancer

To verify the applicability of the algorithm with non-temporal shotgun phosphoproteomics data, we applied it to the experimental dataset from Ruprecht et al. [20] describing changes in the phosphoproteome upon treatment with the breast cancer drug lapatinib in sensitive ('parental' BT-474) or lapatinib-resistant (BT-474-J4) cells. The paper uncovered and experimentally validated the role of several metabolic enzymes and signaling pathways that were driving lapatinib resistance. In particular, proteins that form the spliceosome, those involved in glycolysis and glycogen catabolism and PI3K/AKT/mTOR pathway members were differentially phosphorylated in lapatinib-resistant cells compared to parental lapatinib-treated cells.

We first constructed a multilayer heterogeneous network as described in Experimental Procedures which included 1603 phosphorylated sites that showed significant changes in response to lapatinib treatment in either the parental or resistant cell lines (Table 4). Data were grouped into 5 clusters (referred to as clusters 1 – 5) using the k-means method. The number of clusters to use was determined using the elbow plot method [37] (Figure S3A), confirming that 5 clusters was sufficient to capture all features found in the data. Indeed, sites sharing similar regulation profiles between the resistant and parental cells were clustered together (Figure S3B). As the original publication used KEGG pathways rather than GO terms, we incorporated these in the functional-annotation layer of our network to investigate the flexibility of our approach. As with the Francavilla et al. [19] data, terms that were ranked the same regardless of the seed nodes were excluded from further analysis. There were 26 such pathways out of the total 149 in the network; these included ErbB signaling, MAPK signaling, ‘pathways in cancer’ and various specific cancer pathways (including breast cancer).

The results of RWHN applied to the multilayer heterogeneous network (Figure 5A) show that when the seed nodes are set to sites belonging to clusters 3, 4, or 5, which respectively represent sites that are: rescued in resistance, decreased in resistance, or substantially increased phosphorylation upon lapatinib treatment, the spliceosome pathway was highly ranked. Clusters 3 and 4 also highly ranked the mTOR
pathway and metabolism-related pathways (such as ‘carbon metabolism in cancer’). Cluster 4 highly ranked ‘glycolysis’ specifically, along with ‘transcriptional misregulation’, indicating alternate and nuanced roles for different SRRM2 phosphorylated sites and its association with non-canonical pathways. Metabolic terms were also highly ranked when the seed nodes were set to cluster 1 (representing increased phosphorylation in resistant cells), with terms such as ‘glycolysis/gluconeogenesis’, ‘pentose phosphate pathway’ and ‘starch and sucrose metabolism’ ranked higher. This cluster contains the ALDOA-S39 site highlighted in the original publication and, interestingly, the activating S17 on SRC; this could indicate crosstalk between the ErbB/MAPK and metabolic pathways to be experimentally investigated. Cluster 2 (representing decreased phosphorylation upon lapatinib-treatment) contains several of the glycogen catabolism sites highlighted in the paper. There is no KEGG pathway associated with this process, however this cluster highly ranked the mTOR pathway, adding more evidence for the link validated in the paper between mTOR signalling, metabolic rewiring and lapatinib resistance. This cluster also contained several of the key spliceosome sites highlighted by Ruprecht et al. [20] (SRRM2_S1132, SRRM2_S1987, HNRNPU_S59), indicating parallel- or cross-regulation of the ‘transcriptional misregulation in cancer’ pathway with the metabolic processes. Sites within this cluster could be investigated as therapeutic targets.

Much of the investigation by Ruprecht et al. [20] was done purely on the sites found to be regulated in lapatinib-treated resistant cells compared to the untreated parental cells. We therefore constructed a second network using only those sites that were regulated in lapatinib-treated resistant cells (Table 5). These were grouped into four clusters (referred to as clusters 1 – 4) by K-means clustering as before (S3A,C) and KEGG terms were used again in the functional annotation layer. Of the 156 pathways included in the network, 23 were ranked identically regardless of node; these were largely the same as those in common in the first network but did not include ‘ErbB signaling’.

As described in Figure 5B, two sets of seed nodes (clusters 1 and 4, representing sites downregulated and moderately upregulated) tended to rank metabolic terms more highly (e.g. ‘starch and sucrose metabolism’, ‘galactose metabolism’, ‘pentose phosphate pathway’, ‘purine metabolism’), with clusters 2 and 3 (representing sites slightly and significantly upregulated, respectively) tending to rank other cellular processes and signalling pathways more highly. However, ‘mTOR’, ‘Hippo’ and ‘cgmp-pkg’ signaling were
highly ranked in clusters 2, 3 and 4, indicating some cross-regulation between these pathways and the
metabolism-associated sites in cluster 4. This is corroborated by the inclusion LDHA_Y10 in cluster 4,
discussed by Ruprecht et al. to be an activating site which is known to be phosphorylated by oncogenic
kinases such as FGFR and ABL. Cluster 2 contains ABL1_S737; this site has not been recorded in PSP, but
could be investigated as a regulator of metabolic rewiring in lapatinib resistance.

When sites found in cluster 1 are set as seed nodes, the pathways 'glycolysis/gluconeogenesis' and
'glucagon signalling' are ranked high alongside 'spliceosome', adding further support for the relationship
between these processes and lapatinib-resistance. Cluster 1 contains ENO1-Y44, which was validated in
the original publication as a marker for sustained glycolysis and a target for therapy, PYGB-T59, which is
involved in glycogen catabolism, and sites of various spliceosome and ribosomal proteins such as RPS6-
S240 and SRRM2-S2123. Furthermore, the high rank of ErbB signalling for this cluster indicates further
that it may contain therapeutically relevant sites, as the BT-474 cell-line are HER2 (ERBB2) overexpressing.
An avenue of investigation could be the specific roles the sites in this cluster, such as ERBB2-Y1233 and
ERBB2 interactors YAP1-S164 and RPS6-S240, may have on signalling and metabolism in the context of
lapatinib resistance in HER2-postivive breast cancer.

By comparing to the experimentally validated conclusions of Ruprecht et al., we have extracted meaningful
associations between specific phosphorylated sites and their functions in driving lapatinib resistance. The
algorithm also highlighted several potential sites, such as SRC_S17, that may serve as hubs or intersections
of crosstalk between key signalling pathways and the rewired metabolic processes.

4. Discussion

Phosphorylation has an important impact on protein function and thus cellular behavior. Within a network
of kinase-substrate interactions, phosphorylation modulates the flow of information and regulation of
disparate processes throughout the cell. Understanding the impact of phosphorylation on protein function
traditionally required the experimental investigation and manipulation of individual sites, however with
the advent of high throughput phosphoproteomics, thousands of novel phosphorylation sites of unknown
function have been discovered. As there is limited experimentally validated information available on how
regulation at individual phosphorylated sites impacts the downstream cellular output, current methods for analyzing high-throughput phosphoproteomics data rely on general descriptors of phosphorylated protein function. These often do not account for the specific role of a given phosphorylated site. These methods, valuably used in transcriptomics or proteomics analysis, include GO or pathway enrichment analyses or PPI network construction with differentially phosphorylated proteins. Their use in analyzing phosphoproteomics data disregards site-specific or multiple-site regulation, masking the roles of proteins in a particular modification state [10,12]. Here, we developed and tested a method that associates phosphorylated sites to potential function using RWHN. By incorporating the pattern of phosphorylation upon perturbation, we consider more of the information available in phosphoproteomics datasets and establish the phosphorylated sites as the key drivers in the functional analysis. The algorithm is capable of recapturing experimentally proven functions of phosphorylated sites in a non-gene centric manner, improving the accuracy of analysis and extracting more information from phosphoproteomics data.

To prove the utility of our algorithm we applied it to a validation network, which represented a simple model of phosphorylation dynamics in the MAPK/ERK pathway, and to two previously published phosphoproteomics dataset describing temporal cellular signalling events and the impact of resistance to drug treatment in breast cancer, respectively. Our algorithm successfully distinguished between differentially regulated phosphorylated sites from the same protein and associated them to both known and previously uninvestigated functions. For example, when applying this approach to the data from Francavilla et al. [19], we can recapture the roles of EGFR sites with functions described in PSP. We also associate EGFR sites to the phenotypes described in the paper. In the case of EGF-driven receptor localization, EGFR_Y998 associated with the RAB7_Y183 site described in the original publication as an indicator of the EGF-EGFR response, and was part of a cluster that ranked the term ‘establishment of protein localization to organelle’ highly. On the contrary, TGF-α induced receptor recycling and migration were associated here with EGFR sites Y998, Y1172 and Y1197. Francavilla et al. associated these sites with their functions using a combination of gene-centric approaches and experimental validation. Here, we use a single site-centric algorithm to extract these associations directly from the phosphoproteomics dataset, demonstrating the power of our approach in narrowing down candidates for further functional studies. By
associating ligand-dependent regulation of EGFR sites, it is possible to disentangle the multifaceted role EGFR plays in regulating cellular signalling networks and downstream cellular behaviors.

From the Ruprecht et al. data, we can use the example of SRRM2. This component of the spliceosome is known to be highly phosphorylated, with 675 phosphorylation sites recorded in Phosphosite Plus. At the time of writing none of these sites were ascribed to a function in the database, despite the importance of phosphorylation and dephosphorylation events in orchestrating splicing events [38]. It is also regularly mutated (>5% of cases in TGCA) in lung, stomach, bladder, endometrial and colorectal cancers [6]. Ruprecht et al. uncovered a new role for the spliceosome in modulating lapatinib-resistance, highlighting several sites on spliceosome proteins (including SRRM2_S1132, S1987 and S970) that were differentially modified in resistant and parental cell populations. Previous work has described how SRRM2-depletion in HER2-positive breast and ovarian cancer cells reduced the rate of migration [39]; the spliceosome plays myriad roles in the breast cancer environment, but site-specific analysis is lacking. Using our algorithm, we can begin to uncover the nuanced modification-specific roles of proteins such as SRRM2 in metabolic rewiring. Our analysis showed that SRRM2 sites clustered with sites belonging to proteins involved in glycogen metabolism. When sites in this cluster were set as seed nodes metabolic and transcriptional processes were upregulated along with mTOR signalling. The sites in this cluster could be investigated as points of crosstalk between these different cellular processes. This example highlights the value of our approach in generating hypotheses of modified protein function and signaling pathway crosstalk.

Although the associations predicted here between sites and functions cannot be interpreted as causative, phosphorylation sites can act as indirect, and context-dependent, regulators in cellular processes. This is demonstrated by the clustering of SRC_S17, a protein that is typically associated with signalling downstream of cell-surface receptors, with a cluster that ranked metabolic pathways higher than cellular processing or signalling pathways. Critical to this is a robust method of defining clusters, to capture meaningful regulatory patterns at different phosphorylated sites. Here we define clusters using established techniques (e.g. elbow method for K-means) or visually identifying those with differing time-resolved behaviors. By integrating clustering of quantitative phosphoproteomics data and multilayer network
construction our algorithm closely associates sites with similar behaviors to predict the influence they have on downstream cellular processes.

Multilayer heterogeneous networks are increasingly being used to integrate ‘omic data; here, their use allows phosphoproteomics data to be incorporated with PPI networks and functional annotations, overcoming the issue of considering phosphoproteomics data primarily on the protein level. A potential drawback of our approach is the reliance on large semi-curated resources such as GO or STRING. For instance, different clusters may have many terms or pathways in common given the involvement of many proteins in the same biological functions and the high proportion of frequently used GO terms [40]. In the work described here, we encounter this issue with the RAF1 inhibitory site S259, which was part of a cluster that resulted in ‘positive regulation of MAPK cascade’. We theorize that this may be rectified if the data were clustered into more groups, in order to capture more nuanced phosphorylation patterns. However, enrichment of non-specific terms or false positives remains an issue when analyzing high-throughput ‘omics data by other commonly used methods too. We have attempted to limit this effect by only including edges in the PPI network with a STRING experimental confidence score of greater than 0.4. This could be taken further by filtering for only high confidence functional annotations or incorporating annotations from multiple sources.

An important application of this work could be the opportunity to revise databases like PhosphositePlus, in particular if done in combination with the recently described PTMsigDB phosphorylated-site signature library [12]. There is a pressing need in the phosphoproteomics community to have accurate and context-specific information identified sites, particularly those that are regularly detected in phosphoproteomics experiments. We also anticipate that this method could be used routinely in laboratories generating mass spectrometry-based phosphoproteomics data. Our method is flexible enough to be used with any discovery phosphoproteomics data that describes a change between conditions. This is an improvement on the previously published use of RWHN using multiple sources of phosphoproteomics data to uncover disease-dependent regulation [18]. Moreover, it could easily be generalized to any post-translational modification proteomics dataset, as it incorporates readily available PPI and functional annotation data, as demonstrated here. The fundamental aspect would be maintained with any of these expansions: specific
patterns of regulation at modified sites dictate movement through the multilayer heterogeneous network. It can also be noted that although the experimental datasets used in this work are based on SILAC quantification, our method is not dependent on a specific quantitative phosphoproteomics method.

We have proposed a site-centric approach to analyze phosphoproteomics data, that provides a robust alternative to gene-centric methods of analysis. We integrated clustered quantitative phosphoproteomics data, a context-specific PPI network and functional annotations into a multilayer, heterogeneous network and use the RWHN method to predict the functions of phosphorylation sites with similar regulatory patterns. Using our algorithm, we extracted experimentally validated associations between phosphorylated sites and their role in cellular processes which could not be captured using the typical gene-centric methods used in previous papers. Moreover, our algorithm has the potential to be used by researchers in predicting novel site-function associations and generating hypotheses to be experimentally validated.

5. Acknowledgements

JW is supported by a Biotechnology and Biological Sciences Research Council (BBSRC) Doctoral Training Programme grant (BB/M011208/1). CF is supported by a Wellcome Trust Sir Henry Dale Fellowship (107636/Z/15/Z), the Biotechnology and Biological Sciences Research Council (BB/R015864/1), and the Medical Research Council (MR/T016043/1). We would like to thank current and former members of the Schwartz and Francavilla labs for their thoughtful and insightful discussion on this work.

6. Data Availability

Data and source code can be found at github.com/JoWatson2011/RWHN_Phosphoproteomics

7. References


10. Needham EJ, Parker BL, Burykin T, James DE, Humphrey SJ. Illuminating the dark


8. Tables

Table 1. Multilayer heterogeneous network, constructed from MAPK/ERK validation data. GOBP terms were included as functional annotations.

<table>
<thead>
<tr>
<th>Subnetwork</th>
<th>Edges</th>
<th>Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site-site</td>
<td>28</td>
<td>19</td>
</tr>
<tr>
<td>Protein-protein</td>
<td>932</td>
<td>313</td>
</tr>
<tr>
<td>Function-function</td>
<td>176</td>
<td>71</td>
</tr>
<tr>
<td>Site-protein</td>
<td>19</td>
<td>19 sites, 8 proteins</td>
</tr>
<tr>
<td>Protein-function</td>
<td>1616</td>
<td>251 proteins, 161 functions</td>
</tr>
</tbody>
</table>

Table 2. Multilayer heterogeneous network, constructed from Francavilla et al. [19] EGF-regulated phosphorylated sites. GOBP terms were included as functional annotations.
Table 3. Multilayer heterogeneous network, constructed from Francavilla et al. [19] TGFα-regulated phosphorylated sites. GOBP terms were included as functional annotations.

<table>
<thead>
<tr>
<th>Subnetwork</th>
<th>Edges</th>
<th>Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site-site</td>
<td>4259</td>
<td>729</td>
</tr>
<tr>
<td>Protein-protein</td>
<td>17476</td>
<td>4181</td>
</tr>
<tr>
<td>Function-function</td>
<td>24</td>
<td>17</td>
</tr>
<tr>
<td>Site-protein</td>
<td>729</td>
<td>729 sites, 437 proteins</td>
</tr>
<tr>
<td>Protein-function</td>
<td>190</td>
<td>42 proteins, 53 functions</td>
</tr>
</tbody>
</table>
Protein-function 710 107 proteins, 138 pathways

Table 5. Multilayer heterogeneous network, constructed from Ruprecht et al. [20] lapatinib-regulated phosphorylated sites in lapatinib-resistant cell line. KEGG pathways were included as functional annotations.

<table>
<thead>
<tr>
<th>Subnetwork</th>
<th>Edges</th>
<th>Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site-site</td>
<td>707785</td>
<td>1920</td>
</tr>
<tr>
<td>Protein-protein</td>
<td>39506</td>
<td>6337</td>
</tr>
<tr>
<td>Function-function</td>
<td>15</td>
<td>24</td>
</tr>
<tr>
<td>Site-protein</td>
<td>1920</td>
<td>1920 sites, 1093 proteins</td>
</tr>
<tr>
<td>Protein-function</td>
<td>883</td>
<td>134 proteins, 142 pathways</td>
</tr>
</tbody>
</table>

9. Figures & Figure Legends

Figure 1. Overview of multilayer heterogeneous network construction from phosphoproteomics data.
Figure 2. Construction of validation data based on the MAPK/ERK pathway. A) Traditional representation of the MAPK/ERK pathway. B) A validation dataset was constructed based on the MAPK/ERK pathway, simulating phosphorylation dynamics over time after pathway activation. The sites were clustered using fuzzy c-means clusters. These data were used to construct a heterogeneous multilayer network.
**Figure 3.** Output of RWHN run on validation data. A) Top ranked terms from each set of seed nodes (equivalent to the clusters, inset) when RWHN was run on the validation network. The top 20 ranked terms from each set of seed nodes is shown. B) The validation network was permuted 100 times and RWHN run over each permutation with seed nodes set to sites found in clusters 1 to 5. KDE was calculated to estimate how frequently each term appeared at each rank in the random networks. If the random networks generated the same result as the actual network we would expect the KDE plots to closely resemble the trend of the true rank.
Figure 4. Output of the RWHN algorithm run on the EGF (A) and TGF-α (B) networks constructed from the Francavilla et al. [19] dataset. The seed nodes were set to the clustered phosphorylated sites. The top 20 ranked terms for each set of seed nodes are shown.
Figure 5. Output of the RWHN algorithm run on the total (lapatinib regulated sites in lapatinib-resistant and parental cell lines) (A) and lapatinib-resistant only (B) networks from the Ruprecht et al. [20] data with KEGG pathways. The seed nodes were set to the clustered phosphorylated sites. The top 30 ranked pathways for each set of seed nodes are shown.