# Bioinformatics Analysis, Management and Organization of Biological Data related to Post-Translational Regulation.

A thesis submitted to the University of Thessaly, Larisa, Greece, for the degree of Doctor of Philosophy, in the Department of Biochemistry and Biotechnology, School of Health Sciences.


## June 2020

## By Panayotis Vlastaridis

**Ph.D Advisory committee:**

Supervisor: Grigorios D. Amoutzias, Associate Professor of Bioinformatics with emphasis in Microbiology, Department of Biochemistry and Biotechnology, School of Health Sciences, University of Thessaly, Larisa, Greece.

Advisor: Steve Oliver FMedSci, FAAAS, FAAM, FRSB, Emeritus Professor of Systems Biology, Cambridge Systems Biology Centre & Department of Biochemistry, University of Cambridge, UK.

Advisor: Yves Van de Peer, Professor in Bioinformatics and Genome Biology, Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium

# Abstract

Post-translational regulation is an important, fast and energy efficient level of gene regulation that has attracted the focus of many high-throughput technologies in the last 20 years. Post-translational modifications of amino acids and especially protein phosphorylation play a pivotal role at this level of cellular regulation. Accordingly, this thesis focused on publicly available and abundant high-throughput protein phosphorylation and methylation data, in order to develop computational tools and bioinformatics methods and pipelines, with the aim to analyze them and transform raw data into biological knowledge, about the properties of the eukaryotic phosphoproteome. During this thesis, phosphoproteomic and methylproteomic data were mined from the literature. An annotation tool and a database were developed in order to facilitate the mining and storage of these complex data, that were integrated with many other omic and evolutionary data. Statistical analyses of the gathered and filtered data allowed for a reliable estimate of the total number of phosphoproteins and phosphorylation sites in model eukaryotes. Furthermore, a focused and in-depth study of the yeast phosphoproteome revealed its pivotal role in the central metabolism and further identified key metabolic processes of biotechnological importance that may be manipulated in the future, with precision, by mutating key phosphorylation sites. Finally, neural networks were developed to predict phosphorylation and methylation sites and further predict potential meth-phos switches and/or clusters. The tools and analyses that were developed during this thesis may function as the first step towards more advanced tools and methods that will integrate many other post-translational modifications in the future.

**Keywords:** Bioinformatics, evolution, post-translational regulation, protein phosphorylation, protein methylation, databases, neural networks, prediction.

# Τίτλος Διδακτορικής Διατριβής: Βιοπληροφορική ανάλυση, διαχείριση και οργάνωση βιολογικών δεδομένων σχετιζόμενων με τη μετα-μεταφραστική ρύθμιση

## Ιούνιος 2020


## Παναγιώτης Βλασταρίδης

Τμήμα Βιοχημείας και Βιοτεχνολογίας, Σχολή Επιστημών Υγείας, Πανεπιστήμιο Θεσσαλίας

**Τριμελής Συμβουλευτική επιτροπή**

Επιβλέπων: Γρηγόριος Δ. Αμούτζιας, Αναπληρωτής Καθηγητής Βιοπληροφορικής με έμφαση στη Μικροβιολογία, Τμήμα Βιοχημείας και Βιοτεχνολογίας, Σχολή Επιστημών Υγείας, Πανεπιστήμιο Θεσσαλίας.

Steve Oliver FMedSci, FAAAS, FAAM, FRSB, Emeritus Professor of Systems Biology, Cambridge Systems Biology Centre & Department of Biochemistry, University of Cambridge, UK.

Yves Van de Peer, Professor in Bioinformatics and Genome Biology, Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium

# Περίληψη

Η μετα-μεταφραστική ρύθμιση αποτελεί ένα σημαντικό, γρήγορο και ενεργειακά αποδοτικό επίπεδο της κυτταρικής ρύθμισης, για το οποίο έχουν αναπτυχθεί πολλές ομικές τεχνολογίες μεγάλης κλίμακας. Η μετα-μεταφραστική τροποποίηση των αμινοξέων και ειδικότερα η πρωτεϊνική φωσφορυλίωση και μεθυλίωση έχουν κεντρικό ρόλο. Αυτή η διδακτορική διατριβή εστίασε σε δημοσιευμένα δεδομένα φωσφο-πρωτεωμικής και μεθυλ-πρωτεωμικής με σκοπό να αναπτύξει υπολογιστικά εργαλεία και βιοπληροφορικές μεθόδους/αναλύσεις που θα μπορούν να τα αναλύσουν και να εξάγουν γνώση για τις ιδιότητες αυτού του επιπέδου ρύθμισης. Κατά την διάρκεια αυτής της διατριβής, συλλέχθησαν, φιλτραρίστηκαν, αποθηκεύτηκαν και οργανώθηκαν δημοσιευμένα δεδομένα, με τη βοήθεια ενός υπολογιστικού εργαλείου διαχείρισης της Βιβλιογραφίας και μιας βάσης δεδομένων που ανέπτυξα. Επιπλέον, και άλλα ομικά και εξελικτικά δεδομένα ενσωματώθηκαν με σκοπό να πραγματοποιηθούν βιοπληροφορικές αναλύσεις σε βάθος. Στατιστικές αναλύσεις επέτρεψαν να εκτιμηθεί το σύνολο των πρωτεϊνών και των αμινοξέων ενός ευκαρυωτικού οργανισμού που υφίστανται φωσφορυλίωση. Μια εις βάθος βιοπληροφορική ανάλυση επέτρεψε να αποκαλυφθεί η σημασία της πρωτεϊνικής φωσφορυλίωσης στην ρύθμιση του κεντρικού μεταβολισμού του ζυμομύκητα *S. cerevisiae* όπως επίσης και οι θέσεις φωσφορυλίωσης με πιθανές βιοτεχνολογικές εφαρμογές, σε περίπτωση στοχευμένης μετάλλαξής τους. Επιπλέον, αναπτύχθηκαν νευρωνικά δίκτυα για την πρόβλεψη θέσεων φωσφορυλίωσης, μεθυλίωσης, καθώς επίσης και συνδυαστικών μοριακών διακοπτών. Τα εργαλεία και οι μέθοδοι/αναλύσεις που αναπτύχθηκαν/εφαρμόστηκαν κατά την πραγματοποίηση αυτής της διατριβής δύνανται να εξελιχθούν ώστε να επιτρέψουν την ενσωμάτωση επιπλέον μετα-μεταφραστικών τροποποιήσεων στο μέλλον.

Λέξεις κλειδιά: Βιοπληροφορική, εξέλιξη, μετα-μεταφραστική ρύθμιση, πρωτεϊνική φωσφορυλίωση, πρωτεϊνική μεθυλίωση, βάσεις δεδομένων, νευρωνικά δίκτυα, πρόβλεψη.

# Acknowledgements

I thank my Lord Jesus Christ and the Most Holy Theotokos.
My deepest gratitude to my Supervisor Dr. Grigorios Amoutzias and the other two distinguished members of my advisory committee, Prof. Steve Oliver and Prof. Yves Van de Peer, whose expertise, patience and understanding have been very important to complete this thesis. Also, I would like to thank all my co-authors and lab colleagues for their contribution to this research.
To my wife and two daughters that have supported me without any expectations and however the difficulties. To all the rest of the family and friends who have missed me and my support, due to this engagement.

# Preface

I started working in the field of Post-translational regulation when I was employed between 2014-2015 as a computer and database scientist at the FAB-PHOS project of the ARISTEIA II Action of the ”OPERATIONAL PROGRAMME EDUCATION AND LIFELONG LEARNING” that is co-funded by the European Social Fund and National Resources (code 4288 to GDA). The project was directed by Dr. Grigorios Amoutzias, Assistant Professor of Bioinformatics in Genomics, at the Department of Biochemistry and Biotechnology, University of Thessaly, Greece. Very soon, I was fascinated by the project and discussed with Dr. Amoutzias the possibility of a PhD thesis on this subject. The thesis started in November 2014.

# Contents

# Chapter 1

# Introduction of the Thesis

## Preface

This is an Introduction into the topic of my thesis. It has been based on a review that was prepared by me, under the supervision of the three members of my PhD advisory committee, Dr. Amoutzias, Prof. Oliver and Prof. Van de Peer. It was published as a peer-reviewed paper at the *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* in 2016. This publication was a conference peer-reviewed paper of the Computational Intelligence methods for Bioinformatics and Biostatistics conference, that was held in Naples in 2015. The conference was organized by the Bioinformatics Italian Society. The published review discusses how high-throughput phosphoproteomics have changed the field of post-translational regulation and what are the challenges of these technologies, in terms of biological noise, technical noise, incompleteness of the datasets. The focus is on phosphorylation, the most abundant post-translational modification. The review further explores the challenges of predicting phosphorylation sites with machine-learning methods and finally discusses the biological properties of a eukaryotic phosphoproteome, as it has been revealed by the bioinformatics analyses of high-throughput data. Although this publication is not available in Pubmed, it is visible within Scopus and the manuscript has obtained **DOI:** 10.1007/978-3-319-44332-4_15. Due to copyright issues with the publisher, the review is not included in the thesis and has to be accessed by the reader.

## The importance of post-translational regulation

Although a unicellular eukaryote like yeast needs only ~6,000 protein coding genes (Oliver *et al.* 1992; Goffeau *et al.* 1996), a very complex creature, like *Homo sapiens* only needs ~20,000 protein coding genes (Lander *et al.* 2001;

Venter *et al.* 2001). A plant genome may even harbor more genes (i.e. 45,000) than humans (Tuskan *et al.* 2006). This intriguing finding has been called the N-value paradox (Claverie 2001). In addition, approximately 1-2% of the human genome accounts for protein-coding sequences (Levine and Tjian 2003). Obviously, the complexity of an organism is not directly related to the number of protein coding genes, that constitute the building blocks of life. Comparative analyses in the genomic era revealed what many researchers previously suspected, that the greater complexity of life must be a result of more sophisticated gene regulation. Indeed, the ratio and the absolute number of transcription factors per genome increases as organisms become more complex (Levine and Tjian 2003; van Nimwegen 2003; Ranea *et al.* 2005) (van Nimwegen, 2003; Levine & Tjian, 2003; Ranea *et al*., 2005). However, gene regulation is not restricted at the level of transcription, but goes beyond that, at the post-transcriptional, translational and post-translational levels. Of particular interest is post-translational regulation, because it allows for fast and energy efficient control. More specifically, post-translational regulation very frequently includes enzymatic modifications of amino acids, that are termed post-translational modifications. It also includes cellular localization, as well as protein-protein interactions.

In this new era of high-throughput mass-spectrometry proteomics, especially during the last 10 years, it has become feasible to detect most of these post-translational modifications in a single experiment. Phosphorylation, methylation, acetylation and glycosylation are the most frequently detected modifications (Khoury *et al.* 2011), with thousands or even tens of thousands of them being detected in a singe experiment. These amino acid modifications may function as molecular switches or even as molecular rheostats, that regulate one or more functions of a protein, like enzyme activity, subcellular localization, complex formation or degradation (Vlastaridis *et al.* 2016). In addition, there may be interactions among neighboring modifications, that modify each other's effect. Thus, very complex higher-order molecular switches and rheostats may emerge,

where the combinatorics become explosive. It is conceivable that the complexity at the post-translational level is higher than any other level of gene regulation. Furthermore, post-translational modifications are very attractive targets for synthetic biology. A mutation of only one modified amino acid may have dramatic effects on the function of that protein, or at the pathway/s that the protein is involved and finally, at the phenotype of the cell. For example, a single mutation of phosphorylated amino acid (S42->A) in the yeast Cdc28 protein causes the cell-size to decrease, whereas the mutation of another phosphorylated amino acid may become lethal (Zhang *et al.* 2005; Vlastaridis *et al.* 2016).

However, the new high-throughput proteomic technologies also pose many challenges, related to the biological noise, technical noise, biases, and the incompleteness of the datasets (Lienhard 2008; Landry *et al.* 2009, 2014; Amoutzias *et al.* 2012). For example, the raw data need to be properly filtered by applying stringent filters related to the correct detection of modified peptides and the exact localization of modified amino acids within these peptides. A detailed review on this specific topic has been published as a conference paper at the beginning of my thesis, in (Vlastaridis *et al.* 2016). In order to avoid intellectual property problems, this review will not be included within this thesis, however, it is accessible over the internet, at the publisher's website.

Given the technical problems and biases that still afflict the high-throughput proteomic technologies, and the incompleteness of the datasets, a very important question arises, concerning the feasibility to use and analyze these data, in order to extract biological knowledge and understand the general properties of this particular level of gene regulation. Fortunately, a previous analysis on protein phosphorylation that compiled a compendium of high quality phosphorylation sites from yeast revealed that other previous analyses, based on single experiments revealed properties that were also present in the compendium (Amoutzias *et al.* 2012). Thus, even these incomplete samples of the eukaryotic

post-translational modifications are capable of revealing the general and major properties of this level of gene regulation.

## Organization of the Thesis

This thesis is organized in 7 chapters, with 4 of them already being published as a review (parts of this Introduction chapter) and research articles (chapters 3, 4, 7) in peer-reviewed journals.

Chapter 2 describes an annotation computational tool that was prepared by me, in order to help the team of annotators of the FAB-PHOS project to store and organize all the literature and supplementary data related to post-translational modifications, detected by high-throughput Mass Spectrometry technologies. The tool is already available for downloading and installing in a local computer by any research team and the goal is to publish it as a short technical note in a peer-reviewed journal in the near future. It is accessible (together with helping videos) at:
http://bioinf.bio.uth.gr/ptm-at.html

Chapter 3 describes how the FAB-PHOS team of annotators mined the literature for available high-throughput Mass Spectrometry phosphorylation data and compiled a compendium of phosphorylation sites from model eukaryotic organisms. This compendium was later analyzed by me, with several methods in order to estimate the total number of phosphoproteins and phosphorylation sites in the selected organisms. This work was included in a peer-reviewed research article in *Gigascience,* in 2017 with Pubmed ID: 28327990 (Vlastaridis *et al.* 2017a).

Chapter 4 describes how I analyzed the phosphorylation data that were previously compiled by our FAB-PHOS team, in order to understand the role of protein phosphorylation in the control of the yeast central metabolism and its

biotechnological implications. Data handling and statistical analyses were performed by me and were included in a publication in *Genes, Genomes, Genetics*, in 2017, with Pubmed ID: 28250014 (Vlastaridis *et al.* 2017b). The evolutionary analyses of the published paper were performed by Mr Chaliotis, while the computational structural analyses of the published paper were performed by Dr. Stratikos and Dr. Papakyriakou.

Chapter 5 describes how our FAB-PHOS team together with other colleagues, further mined methyl-proteomic data, integrated them with phosphorylation data and developed a neural network web server that predicts phosphorylation sites, methylation sites, as well as their meth-phos switches and clusters. In this chapter, other people contributed to the mining of the methyl-proteomic data and the development of the methylation Neural Network whereas I developed/trained the phosphorylation neural network and the web-server that integrates the results from the phosphorylation and methylation neural networks and visualizes them. This chapter has been prepared in the form of a research article for submission in a peer-reviewed journal, like *Bioinformatics*, in the near future. However, some of the data that we are planning to include as supplementary material in the submitted manuscript have been integrated within this chapter, due to relaxed limitations on word count. The server is accessible at:
http://bioinf.bio.uth.gr/meth-phos-prometheus/

Chapter 6 describes how I developed a database to store all the relevant information concerning phosphorylation sites. The structural data of the database have been prepared and provided by Dr. Stratikos and Dr. Papakyriakou, whereas I have developed the database schema, the server, the visualization and have integrated all the available data. The database is accessible at:
http://bioinf.bio.uth.gr/phospho-prometheus-db/

Chapter 7 is a brief description of my contributions to another three related peer-reviewed publications of Dr. Amoutzias research team. My contribution in these

publications was the development of the database schemas and web-servers. In all these three publications I was the second author. The software are accessible at:

http://bioinf.bio.uth.gr/lcr/

The publication of this software in *Nucleic Acids Research*, in 2019 has Pubmed ID: 31504783 (Ntountoumi *et al.* 2019)

http://bioinf.bio.uth.gr/nat-ncs2/

The publication of this software in *Gigascience*, in 2018 has Pubmed ID: 30418564 (Chaliotis *et al.* 2018)

http://bioinf.bio.uth.gr/aars/

The publication of this software in *Nucleic Acids Research*, in 2017 has Pubmed ID: 28180287 (Chaliotis *et al.* 2017).

## Bibliography

Amoutzias, G. D., Y. He, K. S. Lilley, Y. Van de Peer, and S. G. Oliver, 2012 Evaluation and properties of the budding yeast phosphoproteome. Mol. Cell Proteomics 11: M111.009555.

Chaliotis, A., P. Vlastaridis, D. Mossialos, M. Ibba, H. D. Becker *et al.*, 2017 The complex evolutionary history of aminoacyl-tRNA synthetases. Nucleic Acids Res. 45: 1059–1068.

Chaliotis, A., P. Vlastaridis, C. Ntountoumi, M. Botou, V. Yalelis *et al.*, 2018 NAT/NCS2-hound: a webserver for the detection and evolutionary classification of prokaryotic and eukaryotic nucleobase-cation symporters of the NAT/NCS2 family. Gigascience 7:.

Claverie, J. M., 2001 Gene number. What if there are only 30,000 human genes? Science 291: 1255–1257.

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon *et al.*, 1996 Life with 6000 genes. Science 274: 546, 563–567.

Khoury, G. A., R. C. Baliban, and C. A. Floudas, 2011 Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. Sci Rep 1:.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature 409: 860–921.

Landry, C. R., L. Freschi, T. Zarin, and A. M. Moses, 2014 Turnover of protein phosphorylation evolving under stabilizing selection. Front Genet 5: 245.

Landry, C. R., E. D. Levy, and S. W. Michnick, 2009 Weak functional constraints on phosphoproteomes. Trends Genet. 25: 193–197.

Levine, M., and R. Tjian, 2003 Transcription regulation and animal diversity. Nature 424: 147–151.

Lienhard, G. E., 2008 Non-functional phosphorylations? Trends Biochem. Sci. 33: 351–352.

van Nimwegen, E., 2003 Scaling laws in the functional content of genomes. Trends Genet. 19: 479–484.

Ntountoumi, C., P. Vlastaridis, D. Mossialos, C. Stathopoulos, I. Iliopoulos *et al.*, 2019 Low complexity regions in the proteins of prokaryotes perform important functional roles and are highly conserved. Nucleic Acids Res. 47: 9998–10009.

Oliver, S. G., Q. J. van der Aart, M. L. Agostoni-Carbone, M. Aigle, L. Alberghina *et al.*, 1992 The complete DNA sequence of yeast chromosome III. Nature 357: 38–46.

Ranea, J. A. G., A. Grant, J. M. Thornton, and C. A. Orengo, 2005 Microeconomic principles explain an optimal genome size in bacteria. Trends Genet. 21: 21–25.

Tuskan, G. A., S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev *et al.*, 2006 The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313: 1596–1604.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural *et al.*, 2001 The sequence of the human genome. Science 291: 1304–1351.

Vlastaridis, P., P. Kyriakidou, A. Chaliotis, Y. Van de Peer, S. G. Oliver *et al.*, 2017a Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes. Gigascience 6: 1–11.

Vlastaridis, P., S. G. Oliver, Y. Van de Peer, and G. D. Amoutzias, 2016 The Challenges of Interpreting Phosphoproteomics Data: A Critical View Through the Bioinformatics Lens, pp. 196–204 in *Computational Intelligence Methods for Bioinformatics and Biostatistics*, edited by C. Angelini, P. M. Rancoita, and S. Rovetta. Springer International Publishing, Cham.

Vlastaridis, P., A. Papakyriakou, A. Chaliotis, E. Stratikos, S. G. Oliver *et al.*, 2017b The Pivotal Role of Protein Phosphorylation in the Control of Yeast Central Metabolism. G3 (Bethesda) 7: 1239–1249.

Zhang, K., W. Lin, J. A. Latham, G. M. Riefler, J. M. Schumacher *et al.*, 2005 The Set1 methyltransferase opposes Ipl1 aurora kinase functions in chromosome segregation. Cell 122: 723–734.

# Chapter 2

# Development of a computational annotation tool for mining of the proteomic literature

## Abstract

An annotation tool was designed to help our team of annotators of the FAB-PHOS project to collect, store and organize experimental data on post-translational modifications of proteins (phosphorylations, methylations) from scientific publications in various journals. With this custom-designed tool, the annotator can store and categorize the publications found with various tags such as what type of experiments were performed, what organism and what type of tissue the experimental data came from and others. They can also archive the publication manuscript and the supplementary files on the server, for later review by themselves or their colleagues. In addition, a Solr search engine was integrated for more efficient searching in the huge amount of data stored from the annotators. The tool may be downloaded from the Bioinformatics laboratory website, at:

http://bioinf.bio.uth.gr/ptm-at.html

## Introduction

Within the framework of the FAB-PHOS project, a team of annotators investigated the phosphoproteomic literature, manually, in order to identify publications that contained high quality phosphoproteomic data (Vlastaridis *et al.* 2017). The goal was to gather these data, identify what software were used for the determination of the phosphopeptides and the exact location of the phosphorylation sites (p-sites) and apply the correct filters/cutoffs in order to retain phosphopeptides and p-sites that were identified with very high

probability (<1% false positives). This phase originally started with a simple storing of the data and metadata in directories and excel spreadsheets. However, the relevant literature was over 1000 publications in Pubmed and very soon, several problems arose concerning storing, coordination of the annotators and retrieval of the data in later times. Thus, it was realized that a web-based annotation tool was needed in order to store all this information, add all the relevant metadata, apply the correct filters, store the filter data and allow the annotation team to coordinate their efforts without duplicating the work. The annotation tool would need to show to anyone logged into it what publications were being handled by which annotator. Thus, I started developing this tool, in a simple form that was eventually evolved into a more complex structure, as it was used and trouble-shooted by the annotation team.

## Materials and Methods

For the annotation tool, a web-based application was developed that uses a MySQL database (“MySQL”), to store all the relevant information of metadata and filtered data. This needs to be installed separately by the user and is responsible for holding all data entered by the user. It also holds the information used for the authentication of the users such as their permissions (administrators and users with login), passwords, logins etc.

The web-application consists of three parts all included in source code provided by us. These are the i) Tomcat webserver (“Apache Tomcat® - Apache Tomcat 8 Software Downloads”), ii) the Java (“Java | Oracle”) and the Spring framework (“Spring Framework”), iii) the front-end part of our code written in Javascript language (“Free JavaScript training, resources and examples for the community”) and the Angular 8 Framework (“Angular”). More specifically, the Tomcat webserver is implemented to serve the web application and reply to all web requests in the network (or in the web if an external and static IP is set for the host of our web-application). The web-

requests include commands like save, delete or update a publication, show publications with various filters and pagination, add publications to Solr Search engine. The logic for all web-server requests and handling, including the authentication are written in Java and the Spring framework. Spring framework provides useful java libraries for dependency injection, REST API gateway programming, authentication and authorization. The front-end part of our code is written in Javascript language and the Angular 8 Framework. This is the code that is running on each client's browser when visiting our website. It is used for designing the user interface (UI) and programming a friendly and quickly responsive user experience (UX). The UI is designed with Bootstrap 4 css and javascript libraries and the UX with the Angular 8 - Typescript framework. Using Angular for the UI and following the Single Page Application Architecture a much-improved experience is offered to the user. The application feels faster because less bandwidth is being used, and no full-page refreshes are occurring as the user navigates through the application.

I also integrated the Solr search engine into the annotation tool. Solr is an open-source enterprise-search platform, written in Java, from the Apache Lucene project. Its major features include full-text search, hit highlighting, faceted search, real-time indexing and rich document (e.g., Word, PDF) handling. Solr runs as a standalone full-text search server. It uses the Lucene Java search library at its core for full-text indexing - searching and has REST-like HTTP/XML and JSON APIs that make it usable from most popular programming languages. Solr's external configuration allows it to be tailored to many types of application without Java coding, and it has a plugin architecture to support more advanced customization. Apache Lucene and Apache Solr are both produced by the same Apache Software Foundation development team.

# Results and Discussion

## Usage and capabilities

The users of the annotation tool are administered by an administrator. After the administrator approves the registration of a user then, the user can login to the system using a username and password in order to gain access to the data archived in the annotation tool.

After running the application, the user may use any browser (Chrome, Firefox, Safari or Edge) in order to view the publications archived in our annotation tool. Visiting the url: http://localhost:8080/#/publication or from top bar menu “Annotations” - > ”Publication List” will bring us to the table where all publications can be shown.

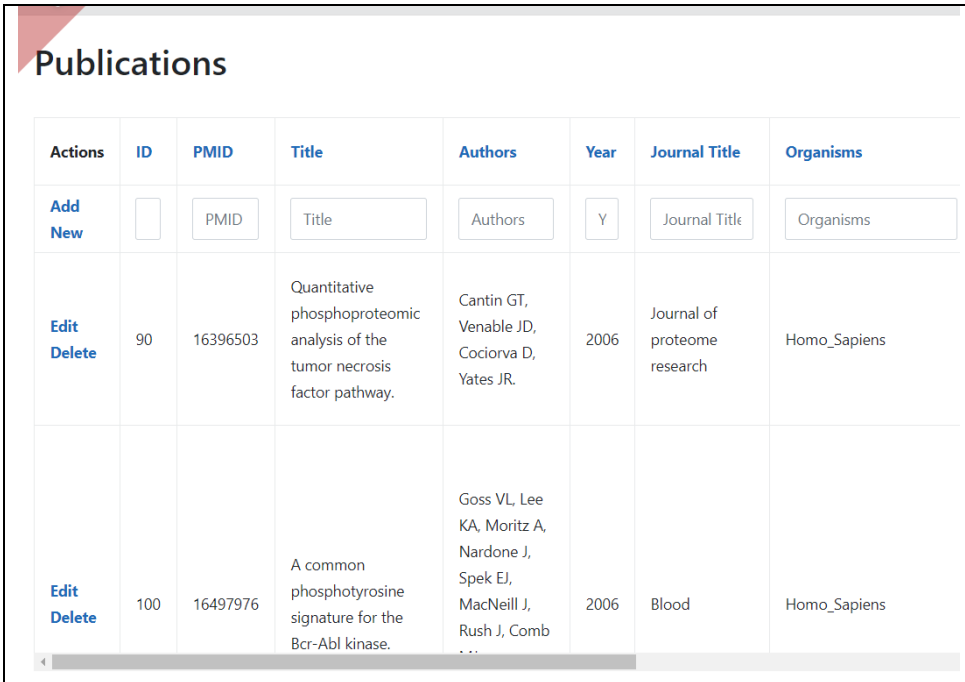



Figure 1. Partial view of the annotation tool that shows some of the metadata stored in the system.

On top of the table, various filters can be used, for the various categories and tags used on each publication so only publications of a certain organism for example can be listed. Also sorting order can be used by pressing on the title of each column (see figure 1). Adding, editing or deleting a publication is done by following the links on the left of this list. A form is opened that the annotator may insert the metadata of the publication (see figure 2).

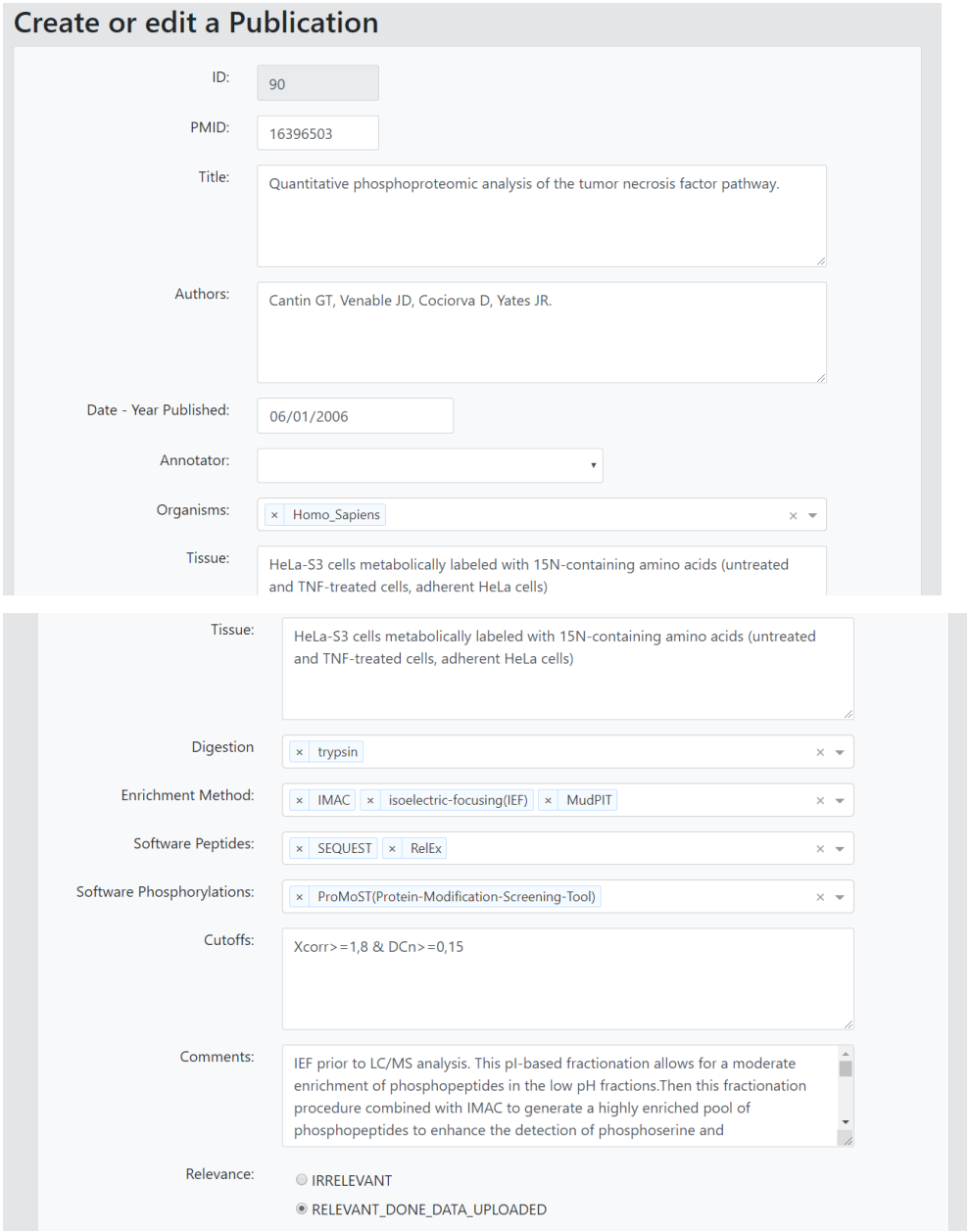



Figure 2. View of the annotation tool, concerning the metadata form that is completed by the annotator.

The annotator may also opt to add the pdf of the manuscript to the Solr search engine, for indexing as well for better searching capabilities within the full manuscript. In the Solr search page, the user may search the full pdf manuscripts with specific terms. The most commonly found terms are

proposed when you start typing. On the left it shows with descending order in which journals the most publications have a match for the terms you are searching for. Then, pressing on a Journal Title it shows only publications that appear in this Journal (facet searching). Also highlighting is being used and search yields portions of the text found in the manuscript file where the user-entered search terms were found including the terms highlighted.

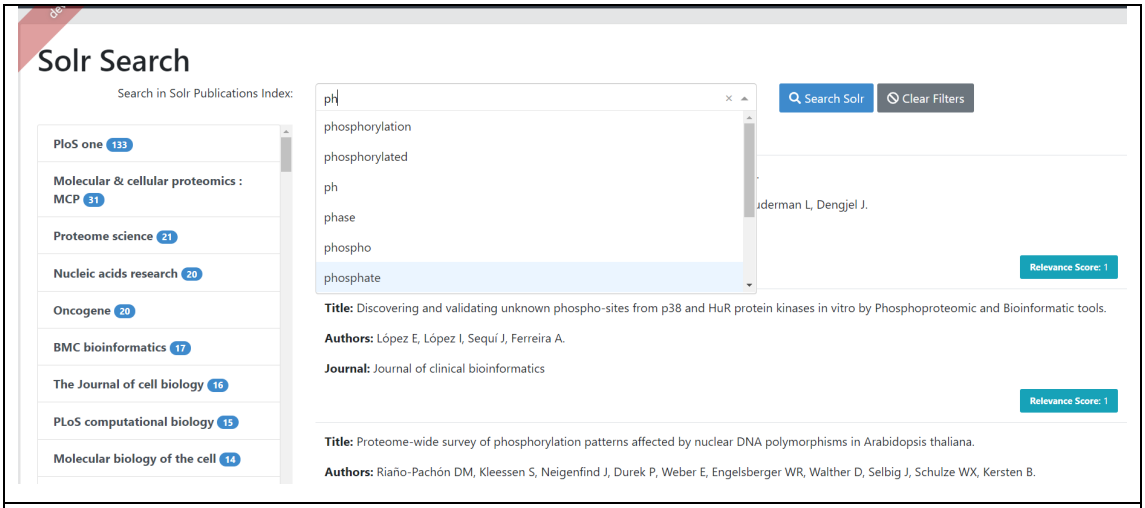



Figure 3. The user may search with keywords, within the stored full manuscripts, using the Solr search engine.

**The relation model of the MySQL database**

The publications are stored in a MySQL relational database management system The following enhanced entity relationship (EER) diagram of figure 4 shows the main table with its columns as well the relationship with two (not all of them) many-to-many relationships.

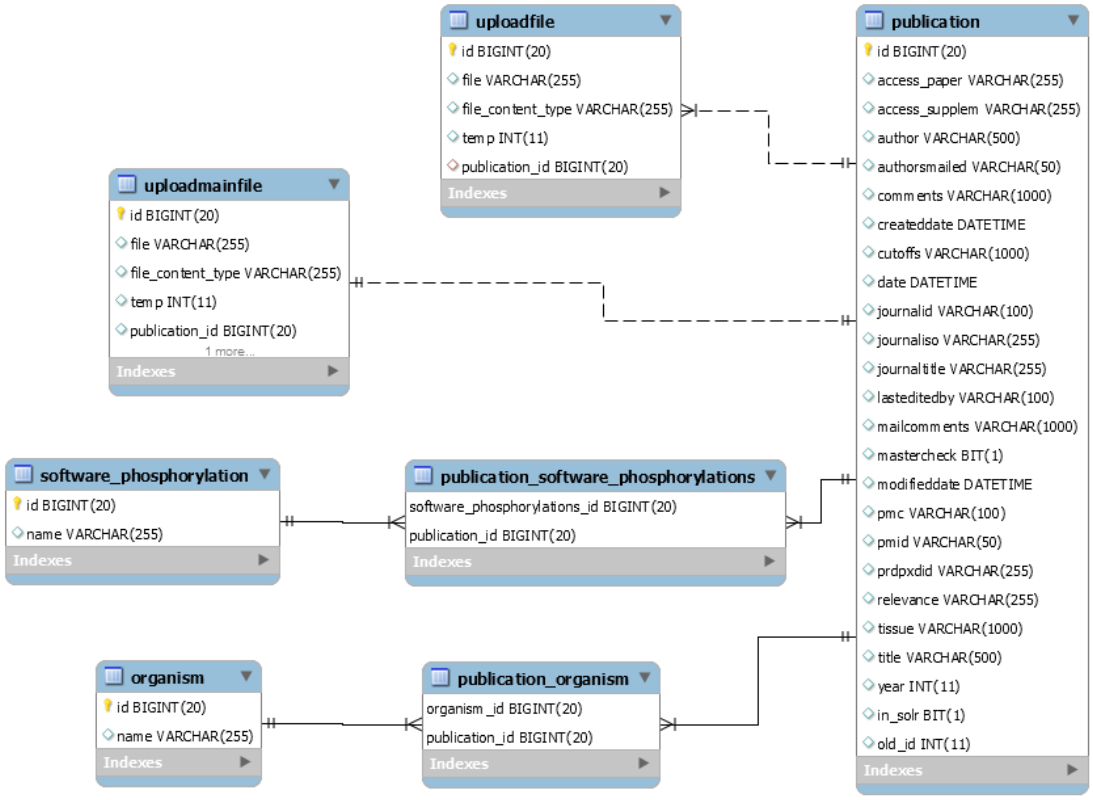



Figure 4. Enhanced entity relationship (EER) diagram of the annotator tool's underlying MySQL database.

Hence with this arrangement we can show in the user interface a drop-down list which is filtered by the typing characters entered by the user (autocomplete). User is also allowed to create new terms after typing a new term and pressing enter.

**Installation on Ubuntu 18.04**

The following section contains information on how to locally install the annotation tool to a Linux Ubuntu 18.04 operational system.

Install MySQL Server Version 5.7 Install the mysql-server package,

```
$ sudo apt install mysql-server
```

Then run the included security script

$ sudo mysql_secure_installation

Running mysql_secure_installation security script, the user can make some changes to the MySQL installation's security options. The first prompt will ask whether the user would like to set up the Validate Password Plugin, which can be used to test the strength of the MySQL password. Regardless of the choice, the next prompt will be to set a password for the MySQL root user. Enter and then confirm a secure password of your choice.

From there, the user can press Y and then ENTER to accept the defaults for all the subsequent questions. This will remove some anonymous users and the test database, disable remote root logins, and load these new rules so that MySQL immediately respects the changes that have been made.

Open up the MySQL prompt from the terminal:

$ sudo mysql

Then create the database:

mysql> CREATE SCHEMA annot_tool DEFAULT CHARACTER SET utf8;

Create user for this database with the following commands:

mysql> CREATE USER atdbuser@localhost IDENTIFIED BY 'nn0tT1';

mysql> GRANT ALL PRIVILEGES ON annot_tool.* TO atdbuser@localhost;

mysql> FLUSH PRIVILEGES;

2) Install Java 8:

sudo apt install openjdk-8-jdk

3) Install Solr:

Now download the required Solr version from its official site or mirrors. Or simply use the following command to download Apache Solr 7.7.2.

cd /opt

sudo wget http://www-eu.apache.org/dist/lucene/solr/7.7.2/solr-7.7.2.tgz

Now extract Apache Solr service installer shell script from the downloaded Solr archive file and run the installer using the following commands.

sudo tar xzf solr-7.7.2.tgz solr-7.7.2/bin/install_solr_service.sh --strip-components=2

sudo bash ./install_solr_service.sh solr-7.7.2.tgz

After successful installation of Solr on the system, the user may create the first collection in Apache Solr using the following command:

sudo su - solr -c "/opt/solr/bin/solr create -c techproducts -n data_driven_schema_configs"

4) Enable CORS in Apache SOLR

Open the file /opt/solr-7.7.3/server/solr-webapp/webapp/WEB-INF/web.xml and add the following XML before the existing filter section:

```
<filter>
   <filter-name>cross-origin</filter-name>
   <filter-class>org.eclipse.jetty.servlets.CrossOriginFilter</filter-class>
   <init-param>
      <param-name>allowedOrigins</param-name>
      <param-value>http://localhost*</param-value>
   </init-param>
    <init-param>
      <param-name>allowedMethods</param-name>
```

```
        <param-value>GET,POST,DELETE,PUT,HEAD,OPTIONS</param-
value>
    </init-param>
    <init-param>
        <param-name>allowedHeaders</param-name>
        <param-value>origin, content-type, cache-control, accept, options,
authorization, x-requested-with</param-value>
    </init-param>
   <init-param>
       <param-name>supportsCredentials</param-name>
       <param-value>true</param-value>
   </init-param>
   <init-param>
     <param-name>chainPreflight</param-name>
     <param-value>false</param-value>
   </init-param>
</filter>

<filter-mapping>
  <filter-name>cross-origin</filter-name>
  <url-pattern>/*</url-pattern>
</filter-mapping>
```

6) Import SOLR Data

Our project files can be downloaded from http://bioinf.bio.uth.gr/ptm-at.html. Please extract them in a location such as ~/share/ and then import date to SOLR by giving:

sudo cp -a ~/share/techproducts/. /var/solr/data/techproducts/

sudo service solr start

7) Copy folder and start project

$ mkdir ~/pmat

$ sudo cp -a ~/share/project/. ~/pmat/

$ cd ~/pmat

$ java -jar *.jar

8) Additional Settings for SOLR.

In case of large pdf files, with not very commonly used fonts, the files need to be indexed by SOLR. Follow the instructions below:

$ sudo apt-get install libpdfbox-java

After this modification, the user need to change solr start parameter in ./bin/solr from

SOLR_JAVA_STACK_SIZE='-Xss256kb'

to

SOLR_JAVA_STACK_SIZE='-Xss256M'

$ sudo service solr restart

At the accompanying download website of the Bioinformatics laboratory, a help page with videos exists on how the tool works and how it should be installed.

## Conclusions

A computational tool was designed to help annotators collect, store and organize experimental data on post-translational modifications of proteins. The tool is web-based and it uses a MySQL database schema and a Solr search engine. The web-application consists of three parts i) the Tomcat webserver, ii) the Java and the Spring framework, iii) the front-end part written in Javascript language and the Angular 8 Framework. The tool allows many annotators to coordinate their efforts and observe the status of the literature that is being annotated by the team, in real time. The tool has been developed to be available to other Bioinformatics teams as well, for local installation and usage. The tool may be downloaded from the Bioinformatics laboratory website, at:

http://bioinf.bio.uth.gr/ptm-at.html

Finally, a help page with videos exists at the accompanying website on how the tool works and how it should be installed.

# Bibliography

https://angular.io/

https://tomcat.apache.org/download-80.cgi

https://www.javascript.com/

https://www.java.com/en/ MySQL.

https://spring.io/

Vlastaridis, P., P. Kyriakidou, A. Chaliotis, Y. Van de Peer, S. G. Oliver *et al.*, 2017 Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes. Gigascience 6: 1–11.

# Chapter 3

## Preface

Chapter 3 describes how I analyzed a compendium of p-sites, with several methods in order to estimate the total number of phosphoproteins and phosphorylation sites in the selected organisms. The FAB-PHOS team of annotators mined the literature for available high-throughput Mass Spectrometry phosphorylation data and compiled the compendium of phosphorylation sites from model eukaryotic organisms. The work of this chapter was included within a peer-reviewed research article in *Gigascience,* in 2017 with Pubmed ID: 28327990.

## Title: Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes.

## Abstract

Despite the cornucopia of high-throughput (HTP) phosphoproteomic data in the last decade, it remains unclear how many proteins are phosphorylated and how many phosphorylation sites (p-sites) exist within a eukaryotic proteome. This chapter and its accompanying publication provides the first reliable estimates of the total number of phosphoproteins and phosphorylation sites (p-sites), for four eukaryotes (human, mouse, *Arabidopsis,* and yeast). It is based on 187 HTP phosphoproteomic datasets that were filtered, compiled and studied along with two low-throughput (LTP) compendia. Estimates of the number of phosphoproteins and p-sites were inferred by Capture-Recapture, and fitting the saturation curve of cumulative redundant vs. cumulative non-redundant phosphoproteins/p-sites. Estimates were also controled for various confounding factors. Thus, it is estimated that, 13,000, 11,000 and 3,000 phosphoproteins and 230,000, 156,000 and 40,000 p-sites exist in human, mouse and yeast, respectively, whereas estimates for Arabidopsis were not reliable enough.

# INTRODUCTION

It is of paramount importance to know which proteins are phosphorylated and on which of their amino acids. Nevrtheless, it is still unclear how many (in total) proteins are phosphorylated and how many phosphorylation sites (p-sites) can exist within a proteome. Experts still speculate on this. It has been suggested that between 1/3 and 2/3 of an organism's proteome could be phosphorylated [1–4]. For the human proteome, 57,000, 500,000, 700,000, or even 1,000,000 p-sites have been suggested to exist [5–8]. Sharma *et al.* performed a deep phosphoproteome analysis on HeLa cells and estimated that at least 75% of the proteome expressed in those cells can be phosphorylated, and this number may well rise to 90%, if phosphoproteomic experiments are performed at higher coverage [9]. In an effort to provide a statistically robust estimate based on current knowledge, over 1000 articles from the literature were investigated by our team. From them, 187 publicly available HTP phosphoproteomic datasets from four well-studied species were of use. By implementing two independent statistical methods - the Capture-Recapture method, and Curve-Fitting on the saturation curve of redundant phosphoproteins/p-sites vs non-redundant phosphoproteins/p-sites – I obtained estimates for humans and three other model eukaryotes.

# MATERIALS AND METHODS

Pubmed was searched with the keywords “phosphoproteomic OR phosphoproteomics”. Thus, over 1000 relevant article-hits were manually inspected for available raw data in human, mouse, *Arabidopsis,* and yeast. A cut-off criterion of 99% correct peptide identification and 99% correct p-site localization was implemented to filter phosphopeptides. Finally, phosphopeptides that exactly matched two or more genes/proteins were removed. Thus, 97, 42, 28 and 20 HTP datasets were retained for human, mouse, *Arabidopsis*, and budding yeast respectively. For every protein-encoding gene, only the longest peptide was retained. The Phosphosite plus database was used to retrieve human and mouse p-sites that were identified by LTP technologies [10], whereas the PhosphoGrid2 database was used to retrieve yeast LTP p-sites [4,11]. For *Arabidopsis,* no LTP compendium was available at the time.

The Capture-Recapture method is widely used in epidemiology and ecology for estimating unknown population sizes. It has been implemented as the Rcapture module within the R software package [12]. Here, the investigated population is sampled several times. Next, the prediction/estimate of the total population size is based on the observed pairwise overlap among the various samples. The Chao Mth model was selected among other models as the most appropriate, based on the Akaike information criterion test [13]. The user loads the matrix input file in R, where each row represents a protein or p-site and each column represents an experimental dataset. Zero is used for absence and 1 for presence.

the Capture-recapture method is run by executing the “closedp(matrix)” function. Due to the limitations of the software, estimates for each species were based on the 15 largest datasets.

The second estimation method is based on graphing in a scatter plot, the cumulative number of non-redundant (unique) phosphoproteins/p-sites (y-axis of a given scatter plot) identified as relevant experiments accumulated over time against the cumulative number of redundant p-sites/proteins (x-axis of a given scatter plot). Thus, the saturation level of the experiments is graphically visualized. Here, the cumulative number of non-redundant units (phosphoproteins or p-sites) rises steeply at the beginning and slows down later. The cumulative number of units should reach a plateau value that approximates the total number of units in that proteome. This process is very well modeled by an exponential recovery curve, shown by equation 1:

$$y=a*(1 - e^{(-x/b)})$$

In the above equation, **x** is the cumulative number of redundant units (p-sites/phosphoproteins), y is the cumulative number of non-redundant units that have been identified up to that point, **a** is a constant that reveals the maximum value of y (that is actually the estimated total number of non-redundant units) and **b** is a constant that defines the steepness of the curve and is the total number of redundant units needed to be detected in order to identify 63.2% of total non-redundant units. Microsoft Excel was used to do the curve-fitting, by optimizing the **a** and **b** parameters with the GRG non-linear

solving method. This was achieved by minimizing the sum of squared errors (SSE) between the observed and theoretical values. The curve-fitting process is explained in detail in supplementary file S6, which is a screencasting mp4 video in the accompanying publication.

It was also necessary to understand the effect of noise on the estimates. Towards this, three basic assumptions were made: i) noise has a stochastic nature; ii) the pool of noise (potential false-positive p-sites and phosphoproteins) is large; iii) the level of noise within a given experiment is relatively low, somewhere in the range of 1-10%. The above three assumptions are reasonably valid, therefore, the overlap of false-positive p-sites/phosphoproteins among the various experiments is expected to be minimal. Thus, 1%, 5%, and 10% more noise was added to all the datasets and the estimates were obtained again. Based on the results from this artificial increase, an appropriate downward adjustment of the original estimates was made, for each particular level of noise.

The effect of noise can also be investigated in the curve-fitting approach. Due to an expected minimal overlap among false positives of the various experiments, the number of total unique false-positives within the compendium will increase for some time in a linear fashion. Thus, while the number of experiments continues to increase, the number of true-positives will plateau, whereas noise will cause false-positives to continue to accumulate in a linear fashion, as shown in equation 2:

$$\mathbf{y=a*(1-e^{(-x/b)}) + c*x},$$

where: **c** is now the average noise level within the experiments.

It is also possible that the order in which the experiments were performed (or, at least, published) may affect the curve-fitting estimates. To control for this, the order of the experiments was changed in two ways. The largest experiment was placed either first or last in the temporal order, and the parameters of the curve re-calculated. In addition, the curve-fitting parameters were recalculated, but only for the earlier half of the experiments on each species.

# RESULTS AND DISCUSSION

## Estimation of the total number of phosphoproteins and p-sites in yeast, human and mouse

The best-studied unicellular eukaryote is the budding yeast, *S. cerevisiae* and has only ~6,000 proteins [14,15]. Eighteen published phosphoproteomic papers for this organism provided twenty HTP phosphoproteomic datasets, that were generated under a relatively wide range of conditions. Furthermore, within a single MS/MS experiment, almost 70% of its total proteome can be detected [16,17]. In addition, the PhosphoGrid2 database has compiled a very comprehensive compendium of LTP, but high quality, p-sites [4]. Therefore, yeast is the best organism to try to estimate the total number of phosphoproteins and p-sites.

To date, more than 2,500 phosphoproteins and more than 13,000 p-sites have been discovered, probably with some or even many of them as false-positives. Based on the HTP data, Figure 1A shows the saturation level of the yeast phosphoproteins. However, Figure 1B shows the different estimates of the total number of yeast phosphoproteins, using different methods and data treatments. As can be seen from Figure 1A, the detection of phosphoproteins with HTP methods has already approached saturation. The Curve-Fitting method estimates ~2,400 true-positive phosphoproteins, whereas the Capture-

Recapture method estimates ~2,800, assuming 1% noise in each experiment. Interestingly, Beltrao *et al*., also suggested that HTP phosphoproteomic studies have revealed about 80-90% of all *S. cerevisiae* phosphoproteins [18]. In addition, curve-fitting estimates were also obtained, based on highly confident phosphoproteins that have been detected in three or more experiments (this criterion is based on a previous analysis [1] - designated as 3X). This analysis by using only highly confident phosphoproteins suggests a total estimate of ~2,300 phosphoproteins. Therefore, based solely on the current HTP technologies, a gross estimate where ~40-50% of the proteome is phosphorylated, seems as a reasonable one. These conclusions appear robust, even if the order of the largest experiment is perturbed and even if only half of the experiments are used in Curve-Fitting (see Figure 1B).

Concerning the saturation level of p-sites, it is evident, especially from Figure 1C, that their detection is approaching saturation, although this trend is less intense than it is for the total number of phosphoproteins. For 1% expected noise in each experiment, the Curve-Fitting method estimates ~15,000 true positive p-sites, whereas the Capture-Recapture method raises this estimate to ~21,000. However, analysis of highly confident p-sites (detected in 3 or more experiments) with curve-fitting failed to provide a reasonable estimate.

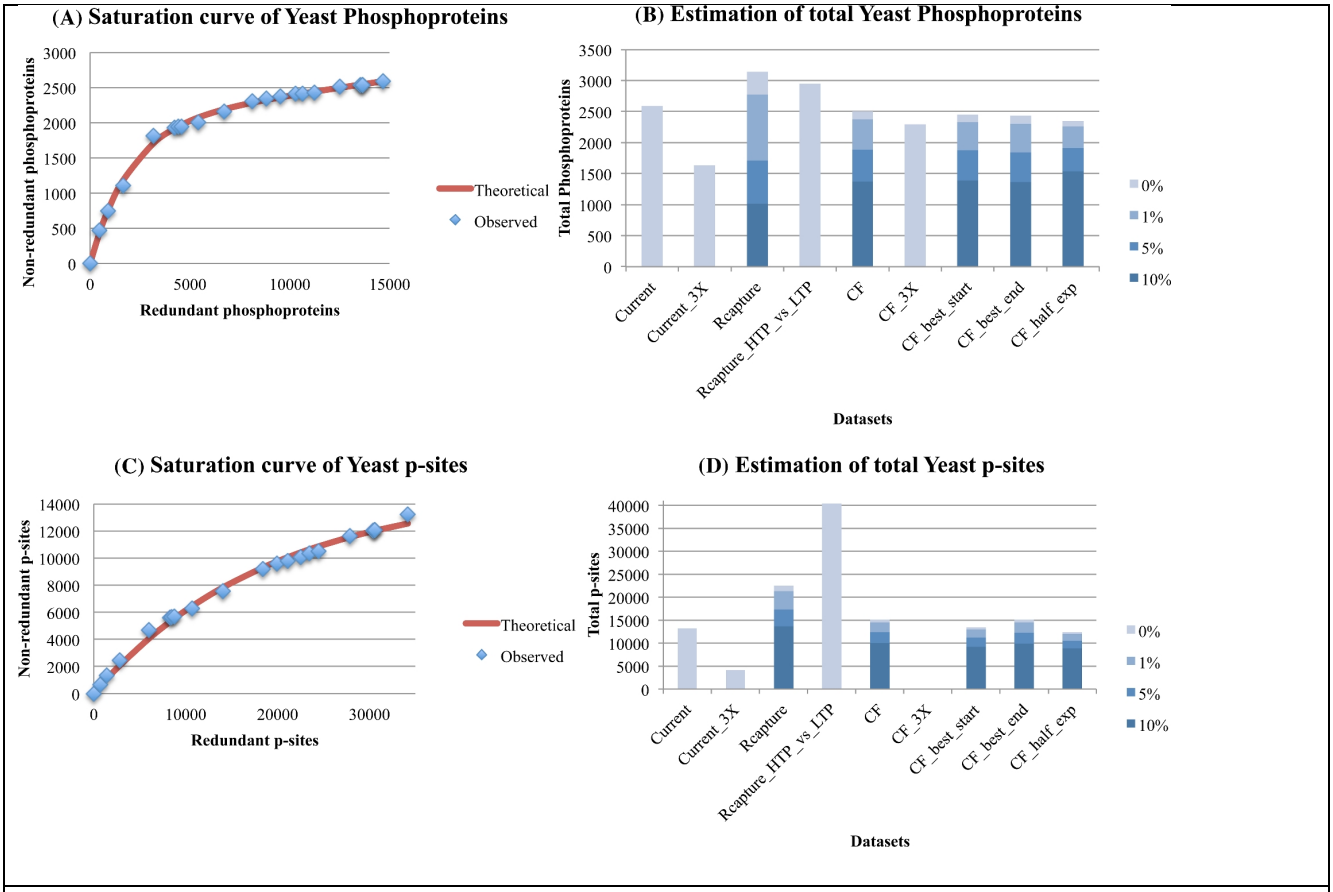



Figure 1. Estimation of the total number of phosphoproteins (1A, 1B) and p-sites (1C, 1D) for yeast, with the curve-fitting (assuming 1% noise) and Capture-Recapture methods, also correcting for 3 levels of noise (1%, 5%, 10%). In figures 1A and 1C, the x-axis is the cumulative number of redundant phosphoproteins/p-sites, whereas the y-axis is the cumulative number of non-redundant phosphoproteins/p-sites. The red curve is fitted for 1% noise. In figures 1B and 1D: Current is the total number of phosphoproteins/p-sites detected so far (by applying our filtering criteria). Current_3X is the total number of phosphoproteins/p-sites detected so far in at least 3 experiments. Rcapture is the estimation of maximum number of phosphoproteins/p-sites based on the Rcapture method (using the 15 largest datasets). Rcapture_HTP_vs_LTP is the estimation of maximum number of phosphoproteins/p-sites based on the Rcapture method, but this time using only

two datasets, where one of them is the compendium of all HTP experiments and the second is the compendium of all LTP experiments from PhosphoGrid2. CF is the estimation of maximum number of phosphoproteins/p-sites based on the curve-fitting method of the saturation curve from all experiments. CF_3X is the estimation of maximum number of phosphoproteins/p-sites identified in at least 3 experiments, based on the curve-fitting method (in this case, a reasonable estimate was not possible). CF_best_start is the estimation of maximum number of phosphoproteins/p-sites based on the curve-fitting method of the saturation curve from all experiments, but this time, the largest experiment is used as first in the series. CF_best_end is the estimation of maximum number of phosphoproteins/p-sites based on the curve-fitting method of the saturation curve from all experiments, but this time, the largest experiment is used as last in the series. CF_half_exp is the estimation of maximum number of phosphoproteins/p-sites based on the curve-fitting method of the saturation curve from the first half experiments. This figure is taken from the accompanying publication in *Gigascience*.

The above estimates are based only on 20 HTP experiments. Nevertheless, experimental and computational studies by others have highlighted a serious problem, where HTP phosphoproteomic experiments may fail to capture many known p-sites, depending on various parameters and protocols [1,11,19–23]. To control for this factor, the LTP (high confidence) data from PhosphoGrid2 were employed as well and were merged into one non-redundant LTP dataset. Similarly, all HTP experiments were merged into one non-redundant HTP dataset. Next, the Capture-Recapture method was implemented by using as input two datasets, the merged HTP one and the PhosphoGrid2 LTP one. This time, the

estimate significantly increased from 21,000 to 40,000 p-sites. On the contrary, the equivalent analysis for phosphoproteins estimated 2,951 total phosphoproteins, which is very close to the one generated by the Capture-Recapture method (2,772) that used the 15 largest HTP datasets individually. Most probably, the analysis that incorporates the LTP data provides a more realistic total estimate than an analysis based solely on HTP data. Therefore, the current HTP technologies seem to be capable of detecting the vast majority (94%) of the yeast phosphoproteome, but only half of the total p-sites.

Similar analyses to those performed on the *S. cerevisiae* proteome were also executed with three other species. The results are presented in Figures 2 (*Homo sapiens*), 3 (*Mus musculus*), and 4 (*Arabidopisis thaliana*), and Table 1 compares the outcomes of the analyses of all four proteomes. In the Table, the most reliable estimates, obtained by incorporating both the HTP and LTP non-redundant datasets are highlighted in bold.

Table 1. **Estimates on the total number of phosphoproteins and p-sites for the various species, based on different analyses. This table is taken from the accompanying published paper in *Gigascience*.**

| | | **Human** | **Mouse** | **Arabidopsis** | **Yeast** |
|---|---|---|---|---|---|
| PROTEINS | current | 10456 | 6512 | 4930 | 2587 |
| | current_3X | 6683 | 3827 | 1815 | 1630 |
| | **Rcapture_HTP_vs_LTP** | **12844** | **11190** | **NA** | **2951** |
| | Rcapture_1%_noise | 10239 | 8346 | 6531 | 2772 |
| | CF_1%_noise | 9160 | 7213 | 4292 | 2373 |

| | | | | | |
|---|---|---|---|---|---|
| | CF_3X | 7582 | 6789 | NA | 2297 |
| | CF_best_start_1%_noise | 8803 | 7167 | 4558 | 2328 |
| | CF_best_end_1%_noise | 8775 | 7099 | 4292 | 2304 |
| | CF_half_exp_1%_noise | 7885 | 6329 | 2373 | 2257 |
| P-SITES | current | 86181 | 36438 | 14796 | 13244 |
| | current_3X | 27110 | 10384 | 3078 | 4156 |
| | **Rcapture_HTP_vs_LTP** | **229616** | **155668** | **NA** | **40350** |
| | Rcapture_1%_noise | 124985 | 71456 | 27815 | 21343 |
| | CF_1%_noise | 94670 | 54031 | 23531 | 14533 |
| | CF_3X | 91500 | NA | 34457 | NA |
| | CF_best_start_1%_noise | 82092 | 45797 | 15122 | 12962 |
| | CF_best_end_1%_noise | 86723 | 49122 | 23531 | 14496 |
| | CF_half_exp_1%_noise | 89639 | 36615 | 6016 | 11980 |

Second column denotes the analysis and datasets: current: experimentally identified; current_3X: experimentally identified in three or more experiments; Rcapture_HTP_vs_LTP: The Capture-Recapture analysis that used the HTP compendium and the LTP compendium (shown in **bold** as the most reliable estimate); Rcapture_1%_noise: The Capture-Recapture analysis assuming 1% noise in each dataset; CF_1%_noise: The Curve-Fitting analysis assuming 1% noise; CF_3X: The Curve-Fitting analysis based on the datasets that have been identified in three or more experiments. CF_best_start_1%_noise: The Curve-Fitting analysis assuming 1% noise and changing the order of the largest experiment as first; CF_best_end_1%_noise: The Curve-Fitting analysis assuming 1% noise and changing the order of the largest experiment as last;

CF_half_exp_1%_noise: The Curve-Fitting analysis assuming 1% noise and using only the first half of experiments.

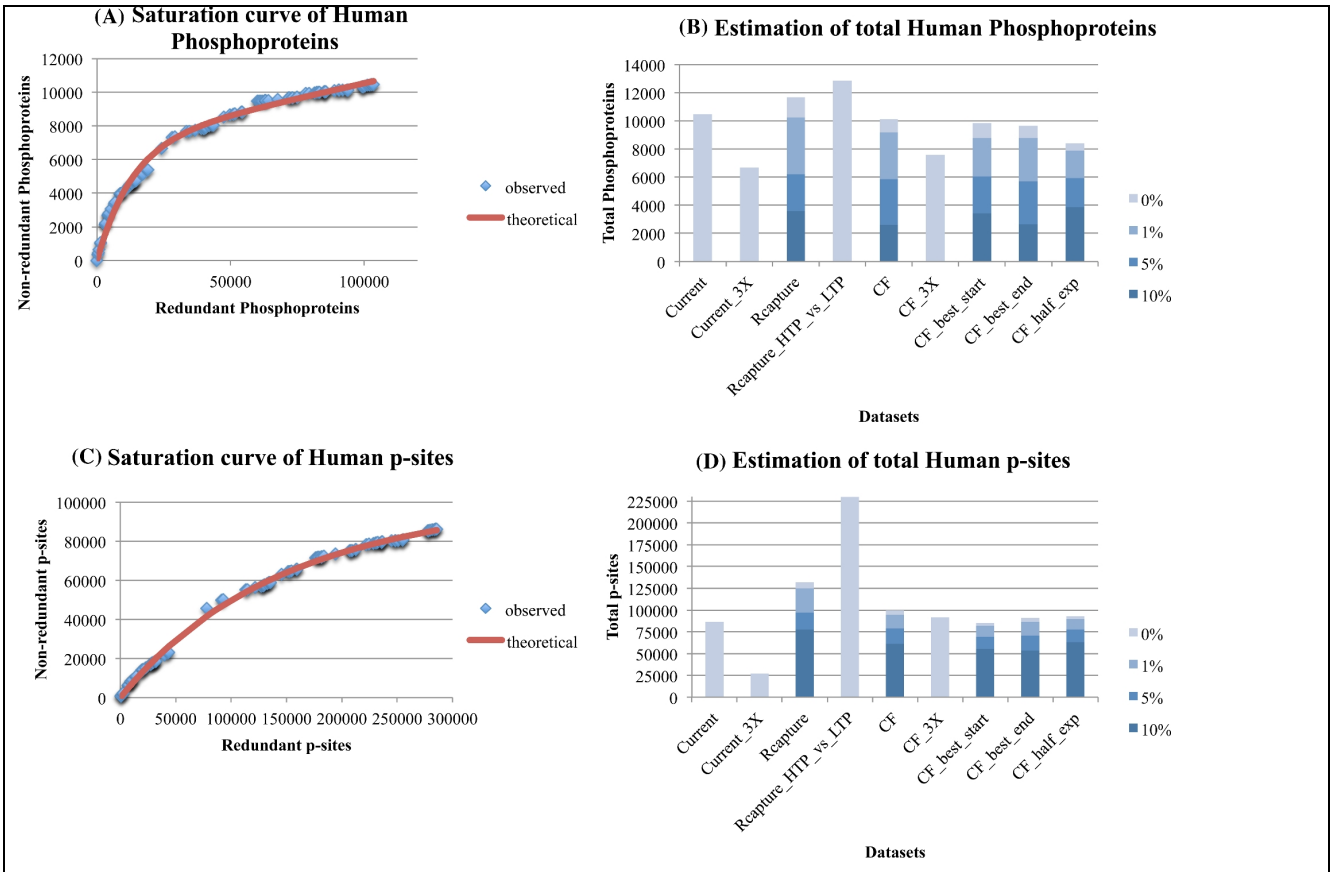



Figure 2. Estimation of the number of phosphoproteins (2A, 2B) and p-sites (2C, 2D) for human, with the Curve-Fitting (assuming 1% noise) and Capture-Recapture methods, also correcting for various levels of noise (1%, 5%, 10%). See legend of Figure 1 for explanations. This figure is taken from the accompanying publication in *Gigascience*.

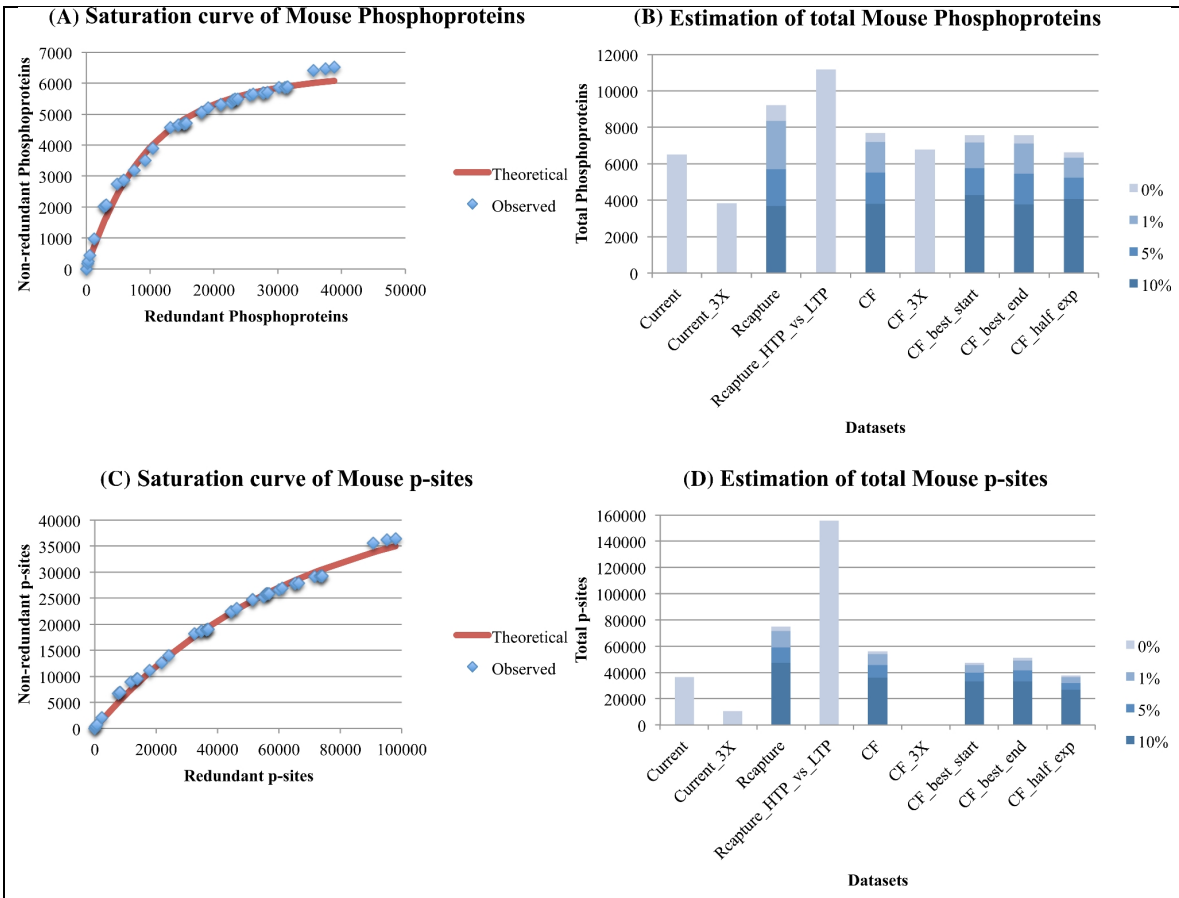



Figure 3. Estimation of the number of phosphoproteins (3A, 3B) and p-sites (3C, 3D) for mouse, with the Curve-Fitting (assuming 1% noise) and Capture-Recapture methods, also correcting for 3 levels of noise (1%, 5%, 10%). See legend of figure 1 for explanations. Estimates on figures 3B and 3D are obtained for a Vega annotated proteome of 16,000 protein-coding genes, where all estimates have been readjusted 25% upwards. This figure is taken from the accompanying publication in *Gigascience*.

The above estimates for human and mouse are based solely on 97 and 42 HTP experiments respectively. A compendium of LTP phosphoproteins/p-sites from Phosphosite plus was used to control again for the fact that HTP technologies may not be able to detect

the whole phosphoproteome. In addition, all HTP experiments were merged into one non-redundant HTP dataset for each species (human, mouse) separately. This time, the Capture-Recapture method was implemented in each species separately by using, as input, two datasets, the merged HTP one and the Phosphosite LTP one. For human, the maximum estimate of total p-sites significantly increased from 125,000 to 230,000. For mouse, the maximum estimate of total p-sites significantly increased from 71,000 to 156,000. In contrast, for human, the maximum estimate for phosphoproteins was increased from 10,200 to 12,800. For mouse, this number increased from 8,300 to 11,200 phosphoproteins. A reasonable interpretation is that the Capture-Recapture estimates that employ the LTP data are more realistic and that the current HTP technologies alone have the potential to capture the majority of the human (80%) and mouse (74%) phosphoproteome, but only half of their total p-sites. The estimates of the number of mouse phosphoproteins and p-sites are about 13% and 32% lower than those of the human phosphoproteins and p-sites respectively (See Table 1 for details).

**Estimation of *Arabidopsis* phosphoproteins and p-sites**

*Arabidopsis thaliana* is a model flowering plant (a eu-dicot) with ~28,000 protein-encoding genes [24] and multiple tissues and cell-types. Twenty eight HTP experimental datasets were collected for this organism, providing 14,796 p-sites in 4,930 phosphoproteins. The saturation level of the *Arabidopsis* phosphoproteins is depicted in Figure 4A, while the estimates on their total number, based on the different methods and data treatments, are depicted in Figure 4B. It is evident, especially from the final data point in Figure 4A, that the

detection of phosphoproteins has not approached saturation. This last experiment detected a lot of new phosphoproteins. In addition, Curve-Fitting estimates based on highly confident phosphoproteins (detected in 3 or more experiments) failed to provide a reasonable estimate. Even worse, Curve-Fitting based on half of the experiments provided an unrealistically low number. Apparently, the publicly available data have not yet reached saturation and most probably they are not sufficient enough to provide a reliable estimate of the total number of phosphoproteins. As a consequence, any attempt to estimate the total number of p-sites in *Arabidopsis* is even more problematic as it is evident from Figure 4C and 4D. Therefore, a gross estimate of 24,000-35,000 p-sites in Arabidopsis is currently suggested by the data, but should be considered of very low confidence.

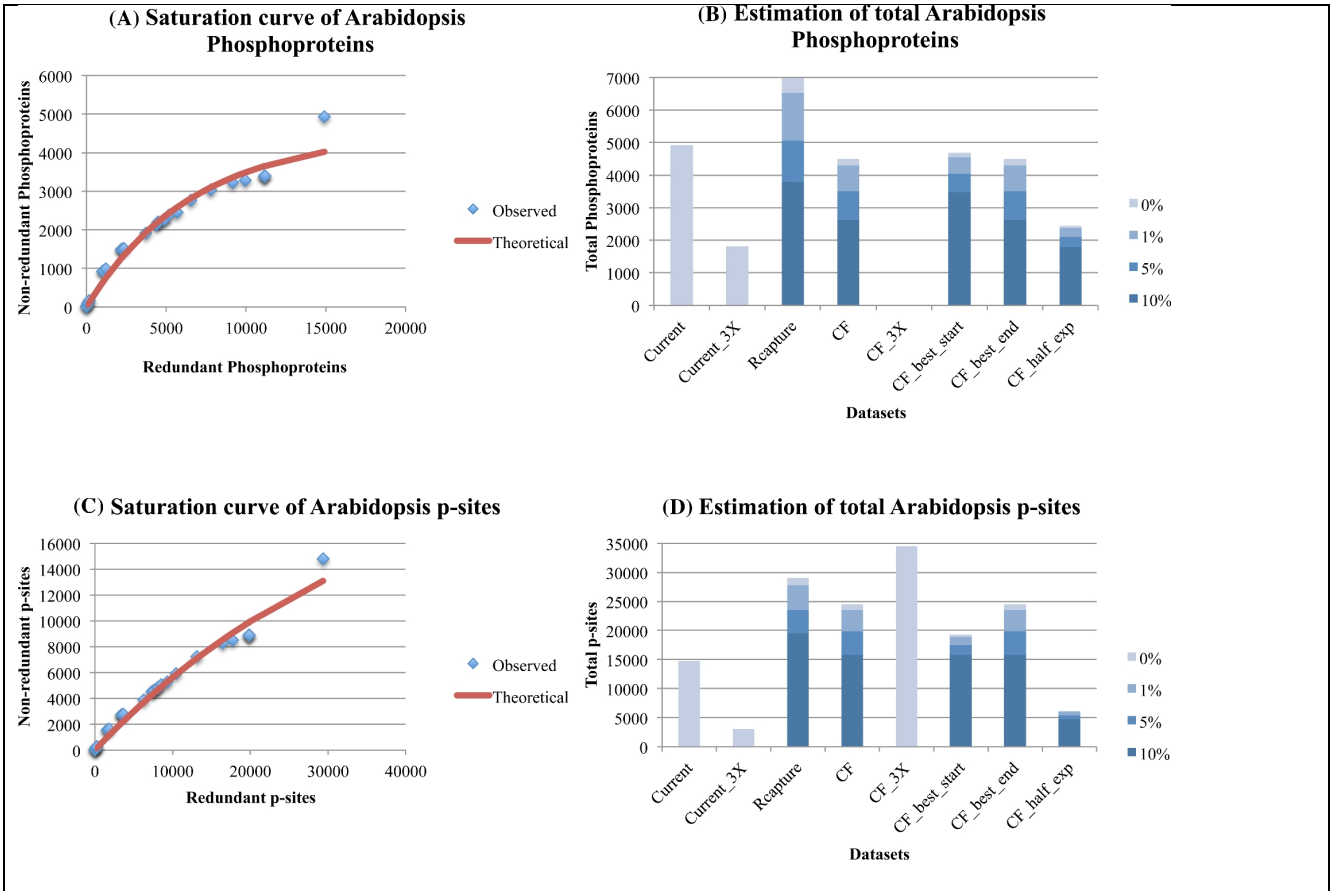



Figure 4. Estimation of the number of phosphoproteins (4A, 4B) and p-sites (4C, 4D) for

*Arabidopsis*, with the Curve-Fitting (assuming 1% noise) and Capture-Recapture methods, also correcting for 3 levels of noise (1%, 5%, 10%). See legend of Figure 1 for explanations. This figure is taken from the accompanying publication in *Gigascience*.

Obviously, the field of phosphoproteomics still faces significant experimental and computational challenges [25]. Several studies have reported that HTP phosphoproteomic experiments alone may fail to capture many known p-sites, depending on various parameters and protocols [1,11,19–23]. For example, consecutive proteolytic digestion by two or more enzymes increased phosphoprotein and p-site detection by 40-70%, compared to an experiment that used only one proteolytic enzyme [20,21,23]. Accordingly, a previous Proteomics analysis on yeast showed that the use of additional proteases, apart from the standard Trypsin resulted in a significant increase of proteomics coverage from 21% to 35% of total Serines, Threonines, Tyrosines [26]. Thus, the proteomics community is exploring the consecutive use of many more than one proteolytic enzymes [27]. P-sites are not evenly distributed across the proteome but tend to cluster, especially at disordered regions [1,28–30]. This increases the probability of missing many neighboring p-sites due to problematic enzymatic digestion of that peptide region. Also, the vast majority of the phosphoproteomic datasets are generated by three enrichment methods (IMAC, $TiO_2$ and p-Tyr pull down) that are well known to exhibit relatively low overlap among them [31,32]. Therefore, it is conceivable that a significant fraction of phosphopeptides are still undetectable from the current imperfect HTP protocols. Furthermore, several replicates may be needed to capture a certain phosphoproteome in a certain condition, as revealed by [5].

HTP technologies will eventually mature to a level that allows the discovery of the total number of p-sites within a proteome. Until then, experts in the field need to determine which p-sites are noisy and which ones have a functional effect on phenotype [33–35]. Considering the large number of p-sites estimated in this analysis, it is likely that such a task will need to use bioinformatics together with experimental processes that assess the phenotype of mutants in a high-throughput manner, like a robot scientist [36,37].

# REFERENCES

1. Amoutzias GD, He Y, Lilley KS, Van de Peer Y, Oliver SG. Evaluation and properties of the budding yeast phosphoproteome. Mol Cell Proteomics. 2012;11:M111.009555.

2. Cohen P. The origins of protein phosphorylation. Nat Cell Biol. 2002;4:E127-130.

3. Pinna LA, Ruzzene M. How do protein kinases recognize their substrates? Biochim Biophys Acta. 1996;1314:191–225.

4. Sadowski I, Breitkreutz B-J, Stark C, Su T-C, Dahabieh M, Raithatha S, et al. The PhosphoGRID Saccharomyces cerevisiae protein phosphorylation site database: version 2.0 update. Database (Oxford). 2013;2013:bat026.

5. Boekhorst J, Boersema PJ, Tops BBJ, van Breukelen B, Heck AJR, Snel B. Evaluating experimental bias and completeness in comparative phosphoproteomics analysis. PLoS ONE. 2011;6:e23276.

6. Boersema PJ, Foong LY, Ding VMY, Lemeer S, van Breukelen B, Philp R, et al. In-depth qualitative and quantitative profiling of tyrosine phosphorylation using a combination of phosphopeptide immunoaffinity purification and stable isotope dimethyl labeling. Mol Cell Proteomics. 2010;9:84–99.

7. Lemeer S, Heck AJR. The phosphoproteomics data explosion. Curr Opin Chem Biol. 2009;13:414–20.

8. Ubersax JA, Ferrell JE. Mechanisms of specificity in protein phosphorylation. Nat Rev Mol Cell Biol. 2007;8:530–41.

9. Sharma K, D'Souza RCJ, Tyanova S, Schaab C, Wiśniewski JR, Cox J, et al. Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. Cell Rep. 2014;8:1583–94.

10. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. 2015;43:D512-520.

11. Stark C, Su T-C, Breitkreutz A, Lourenco P, Dahabieh M, Breitkreutz B-J, et al. PhosphoGRID: a database of experimentally verified in vivo protein phosphorylation sites from the budding yeast Saccharomyces cerevisiae. Database (Oxford). 2010;2010:bap026.

12. Baillargeon S, Rivest L-P. The Rcapture Package: Loglinear Models for Capture-Recapture in R. Journal of Statistical Software [Internet]. 2007 [cited 2016 Mar 13];19. Available from: http://www.jstatsoft.org/v19/i05/

13. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974;19:716–23.

14. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science. 1996;274:546, 563–7.

15. Oliver SG, van der Aart QJ, Agostoni-Carbone ML, Aigle M, Alberghina L, Alexandraki D, et al. The complete DNA sequence of yeast chromosome III. Nature. 1992;357:38–46.

16. de Godoy LMF, Olsen JV, Cox J, Nielsen ML, Hubner NC, Fröhlich F, et al. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature. 2008;455:1251–4.

17. Wu R, Dephoure N, Haas W, Huttlin EL, Zhai B, Sowa ME, et al. Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes. Mol Cell Proteomics. 2011;10:M111.009654.

18. Beltrao P, Trinidad JC, Fiedler D, Roguev A, Lim WA, Shokat KM, et al. Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. PLoS Biol. 2009;7:e1000134.

19. Albuquerque CP, Smolka MB, Payne SH, Bafna V, Eng J, Zhou H. A multidimensional chromatography technology for in-depth phosphoproteome analysis. Mol Cell Proteomics. 2008;7:1389–96.

20. Choudhary G, Wu S-L, Shieh P, Hancock WS. Multiple enzymatic digestion for enhanced sequence coverage of proteins in complex proteomic mixtures using capillary LC with ion trap MS/MS. J Proteome Res. 2003;2:59–67.

21. Gauci S, Helbig AO, Slijper M, Krijgsveld J, Heck AJR, Mohammed S. Lys-N and trypsin cover complementary parts of the phosphoproteome in a refined SCX-based approach. Anal Chem. 2009;81:4493–501.

22. Lee DCH, Jones AR, Hubbard SJ. Computational phosphoproteomics: from identification to localization. Proteomics. 2015;15:950–63.

23. Wiśniewski JR, Mann M. Consecutive proteolytic digestion in an enzyme reactor increases depth of proteomic and phosphoproteomic analysis. Anal Chem. 2012;84:2631–7.

24. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Research. 2012;40:D1202–10.

25. Vlastaridis P, Oliver SG, Van de Peer Y, Amoutzias GD. The Challenges of Interpreting Phosphoproteomics Data: A Critical View Through the Bioinformatics Lens. In: Angelini C, Rancoita PM, Rovetta S, editors. Computational Intelligence Methods for Bioinformatics

and Biostatistics [Internet]. Cham: Springer International Publishing; 2016 [cited 2016 Nov 29]. p. 196–204. Available from: http://link.springer.com/10.1007/978-3-319-44332-4_15

26. Swaney DL, Wenger CD, Coon JJ. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. J Proteome Res. 2010;9:1323–9.

27. Giansanti P, Tsiatsiani L, Low TY, Heck AJR. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. Nat Protoc. 2016;11:993–1006.

28. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, et al. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. 2004;32:1037–49.

29. Moses AM, Hériché J-K, Durbin R. Clustering of phosphorylation site recognition motifs can be exploited to predict the targets of cyclin-dependent kinase. Genome Biol. 2007;8:R23.

30. Schweiger R, Linial M. Cooperativity within proximal phosphorylation sites is revealed from large-scale proteomics data. Biol Direct. 2010;5:6.

31. Bodenmiller B, Mueller LN, Mueller M, Domon B, Aebersold R. Reproducible isolation of distinct, overlapping segments of the phosphoproteome. Nat Methods. 2007;4:231–7.

32. Fíla J, Honys D. Enrichment techniques employed in phosphoproteomics. Amino Acids. 2012;43:1025–47.

33. Lienhard GE. Non-functional phosphorylations? Trends Biochem Sci. 2008;33:351–2.

34. Landry CR, Freschi L, Zarin T, Moses AM. Turnover of protein phosphorylation evolving under stabilizing selection. Front Genet. 2014;5:245.

35. Landry CR, Levy ED, Michnick SW. Weak functional constraints on phosphoproteomes. Trends Genet. 2009;25:193–7.

36. King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, et al. The automation of science. Science. 2009;324:85–9.

37. King RD, Whelan KE, Jones FM, Reiser PGK, Bryant CH, Muggleton SH, et al. Functional genomic hypothesis generation and experimentation by a robot scientist. Nature. 2004;427:247–52.

# Chapter 4

# Title: The pivotal role of protein phosphorylation in the control of yeast central metabolism

## Preface

Chapter 4 describes how the phosphorylation data that were previously compiled by our FAB-PHOS team, were further analyzed by me in order to understand the role of protein phosphorylation in the control of the yeast central metabolism and its biotechnological implications. This work was included within a peer-reviewed article that was published in *Genes, Genomes, Genetics*, in 2017, with Pubmed ID: 28250014 with me as first author**.** The evolutionary analyses in that publication were performed by Mr Chaliotis, while the computational structural analyses (within the publication) were performed by Dr. Stratikos and Dr. Papakyriakou. Data handling and statistical analyses were performed by me.

# ABSTRACT

Protein phosphorylation is the most frequent eukaryotic post-translational modification (PTM) and acts as a molecular switch or molecular rheostat for proteins. The manipulation of this PTM may regulate specific functions with high precision. My goal was to assess the significance of phosphorylation on the eukaryotic central metabolism, and thus its potential for biotechnological and medical applications. Therefore, a compendium of confident protein phosphorylation sites (p-sites) for the model organism *Saccharomyces cerevisiae* has been analyzed. This analysis highlights the global properties of the regulation of yeast central metabolism by protein phosphorylation, where almost half of the enzymes involved are subject to this sort of post-translational modification. These phosphorylated enzymes, compared to the non-phosphorylated enzymes are more abundant, have more protein-protein interactions and a higher fraction of them are ubiquitinated. All this integrated information together with other evolutionary and structural analyses has allowed to prioritize thousands of p-sites in terms of their potential phenotypic impact. Thus, guided future high-throughput mutation studies by wet labs may allow to identify key molecular switches/rheostats for the manipulation, of not only the metabolism of yeast, but also that of many other biotechnologically and medically important fungi and eukaryotes.

# INTRODUCTION

Since the advent of the functional genomic era, there has been a continuous effort by the community to understand and model the metabolic network of the yeast, *Saccharomyces cerevisiae* (Herrgård *et al.* 2008). The final goal is to simulate yeast metabolism *in silico* and provide reliable predictions of the phenotype, after gene deletions (Szappanos *et al.* 2011) or gene additions (Szczebara *et al.* 2003; Galanie *et al.* 2015; Nielsen 2015). Genome-scale stoichiometric models of the yeast metabolic network that allow the computation of the steady-state distribution of metabolic fluxes (Flux Balance Analysis) have proved useful in this regard (Dobson *et al.* 2010; Orth *et al.* 2010).

However, there is a need to improve these models by incorporating post-translational modification of enzyme molecules, e.g. by phosphorylation. They are likely to play an important role in metabolic adaptations because they have rapid kinetics (Oliveira *et al.* 2012; Oliveira and Sauer 2012; Schulz *et al.* 2014; Tripodi *et al.* 2015; Chen and Nielsen 2016). Intriguingly, the energetic cost of protein synthesis is nine times higher than that of transcription (Schwanhäusser *et al.* 2011), therefore post-translational regulation via amino acid modifications seems as a very rapid and energy efficient level of regulation.

The identification of crucial phosphorylation sites (p-sites) in key proteins offers synthetic biologists the prospect of manipulating molecular pathways or organismal phenotypes with greater precision than can be achieved by either the deletion or under/over-expression of complete genes (Oliveira *et al.* 2012; Oliveira and Sauer

2012). However, these p-sites need to be stringently filtered before being investigated experimentally (Landry *et al.* 2009, 2014).

The goal of this study was to use a compendium of stringently filtered phosphoproteomic data from the best-studied model eukaryote, *S. cerevisiae,* together with functional genomic, and phenotypic data so as to: i) reveal the impact of protein phosphorylation on central metabolism, and ii) prioritize the metabolism-related yeast p-sites in terms of biological significance and assess their potential as targets of future mutation studies with a focus on biotechnological and medical applications. Therefore, if we identify crucial phosphorylation switches that regulate yeast metabolism, it should be possible, with a minimal effort, to significantly improve the predictive accuracy of metabolic flux balance analyses.

# MATERIALS AND METHODS

For *S. cerevisiae*, a high quality compendium of p-sites has been used from another computational analysis of our group (Vlastaridis *et al.* 2017). This compendium was generated from 20 HTP phosphoproteomic experiments found in 18 publications (Gruhler *et al.* 2005; Chi *et al.* 2007; Li *et al.* 2007; Albuquerque *et al.* 2008; Bodenmiller *et al.* 2008, 2010; Beltrao *et al.* 2009; Huber *et al.* 2009; Holt *et al.* 2009; Gnad *et al.* 2009; Soufi *et al.* 2009; Aguiar *et al.* 2010; Saleem *et al.* 2010; Wu *et al.* 2011; Oliveira *et al.* 2012; Mascaraque *et al.* 2013; Lee *et al.* 2013; Weinert *et al.* 2014). Very stringent criteria were applied, such as 99% correct phosphopeptide identification and 99% correct p-site localization (see supplementary file S1; spreadsheet: yeast p-sites). This compendium was an update of a previous yeast compendium from 12 HTP datasets (Amoutzias *et al.* 2012). In addition, the PhosphoGrid 2 dataset of manually curated low-throughput (LTP) p-sites (serving as a ‘gold standard) (Sadowski *et al.* 2013) was integrated into the compendium.

For the functional and statistical analyses, many publicly available functional genomics datasets were integrated, such as three protein abundance datasets from two publications (Ghaemmaghami *et al.* 2003; Newman *et al.* 2006), two protein half-lives datasets (Belle *et al.* 2006; Christiano *et al.* 2014), one list of highly confident essential genes (Giaever *et al.* 2002; Steinmetz *et al.* 2002; Pache *et al.* 2009), one protein ubiquitination dataset (Peng *et al.* 2003), one dataset of highly confident genetic interactions (Costanzo *et al.* 2010), one compendium of highly confident protein-protein interactions (Batada *et al.* 2006), a list of genes and the metabolic reactions that they

are involved in, included in the updated version 7.6 of the yeast metabolic model (Dobson *et al.* 2010) and a dataset of biotechnologically important genes that have been annotated as such in the Saccharomyces Genome Database (SGD) (Cherry *et al.* 2012). The integrated functional data are stored in the Excel spreadsheets “yeast p-sites” & “functional_information” of supplementary files S1 and S2 of the publication. Many of the above properties/measurements may be context-dependent or change significantly from one physiological condition to another.

A negative phosphoproteome of 2167 ORFs was also defined, that had no evidence of phosphorylation, even with less stringent filtering criteria.

Data integration was performed with the PERL programming language and statistical analyses with the R programming language (https://www.R-project.org/) (R Core Team 2015). Mapping of the yeast phospho-regulated enzymes to the KEGG metabolic map was performed with the KEGG mapper computational tool (Kanehisa *et al.* 2012), using the Uniprot identifiers of the yeast phosphorylated proteins.

To control for protein abundance as a potential confounding factor (Levy *et al.* 2012) in the comparison between the phosphoproteome and the negative phosphoproteome, relevant abundance measurements (based on the most thorough dataset of Ghaemmaghami *et al*., (2003) were converted to log10 values and binned in 8-10 groups. Equal numbers of phosphoproteins and non-phosphoproteins were randomly selected from each bin, thus generating a Protein-Abundance Controlled (PAC) phosphoproteome and negative phosphoproteome. The same procedure was followed for the metabolic phosphoproteome and the metabolic negative phosphoproteome.

# RESULTS AND DISCUSSION

### The updated yeast p-site compendium

The new *S. cerevisiae* compendium consists of 14339 p-sites in 2633 ORFs (see Table 1 and Excel spreadsheet “yeast p-sites” of supplementary file S1 in the publication) and constitutes a significant increase of 47% (for p-sites) over a previous compendium of 12 publicly available high-throughput phosphoproteomic datasets (Amoutzias *et al.* 2012). It is designated as 21UHQ, where 21 stands for the number of datasets, U stands for phosphopeptides uniquely matched to only one protein and HQ stands for high-quality phosphopeptides, based predominantly on 99% correct peptide identification and 99% correct p-site localization. Compared to the original yeast p-site compendium, the new one has been expanded by 8 more HTP datasets and also includes the latest version of the PhosphoGRID 2 (PG2) subset (Sadowski *et al.* 2013), which is based on manually curated low-throughput p-sites. PhosphoGrid is considered the gold standard of yeast p-sites.

Due to concerns about technical and biological noise in phosphoproteomic data (Lienhard 2008; Landry *et al.* 2009), a highly confident subset was constructed, consisting of 5519 p-sites in 1557 ORFs that includes p-sites identified in 3 or more HTP experiments and/or any of the PG2 Low ThroughPut (LTP) data (see Table 1). The criterion for 3 or more experiments was based on (Amoutzias *et al.* 2012). The corresponding highly confident subset is now designated as 21UHQ_HC, where HC stands for High Confidence.

| | Total p-sites | Total p-sites found in PFAM domains | Highly confident p-sites | Highly confident p-sites found in PFAM domains | Phospho-proteins | Phospho-proteins with highly confident p-sites |
|---|---|---|---|---|---|---|
| 12UHQ | 9783 | 2059 | 2566 (26%) | 431 | 2374 | 1112 (47%) |
| 20UHQ (only HTP) | 13244 | 2625 | 4156 (31%) | 698 | 2587 | 1421 (55%) |
| 21UHQ (including PG2) | 14339 | 3036 | 5519 (38%) | 1175 | 2633 | 1557 (59%) |
| 21UHQ metabolism (including PG2) | 1668 | 527 | 499 | 99 | 412 | 197 |
| 21UHQ metabolism essential proteins (including PG2) | 339 | 153 | 79 | 34 | 71 | 36 |

**Table 1.** The number of unique p-sites and phosphoproteins identified in the various phosphorylation compendiums and subsets. P-sites identified in three or more experiments are designated as **Highly Confident**. **12UHQ** refers to the Amoutzias et al., 2012 dataset. **20UHQ (only HTP)** refers to the p-sites identified by 20 HTP experiments in Vlastaridis et al., 2017. **21UHQ (including PG2)** refers to the p-sites identified by 20 HTP experiments and by PhosphoGrid2 in Vlastaridis et al., 2017. **21UHQ metabolism (including PG2)** refers to the p-sites identified by 20 HTP experiments and by PhosphoGrid2 in Vlastaridis et al., 2017 that were in metabolic proteins. **21UHQ metabolism essential proteins (including PG2)** refers to p-sites identified by 20 HTP experiments and by PhosphoGrid2 in Vlastaridis et al., 2017 that were in essential metabolic proteins.

**A substantial part of the yeast central metabolism is regulated by phosphorylation**

The Yeast 7.6 genome-scale metabolic model is manually curated by experts and contains 2302 reactions that have been assigned one or more from the 909 (15%) protein-coding genes to be catalyzing these specific reactions. Based on the stringency criteria to define a p-site (designated as ALL for all p-sites and HC for the subset of High Confidence p-sites) 412 (45%) or 197 (22%) of the metabolic proteins are phosphorylated and may control 1176 or 656 reactions correspondingly. Thus, protein phosphorylation is likely to exert significant control over the yeast central metabolism (see Figure 1). A previous analysis on older and less filtered datasets also identified half of the metabolic proteins as being phosphorylated (Oliveira and Sauer 2012). Similarly, a review focused on yeast carbon metabolism reported more than half of the relevant enzymes to be targets of post-translational modifications (Tripodi *et al.* 2015). Furthermore, genetic perturbations of the yeast kinome revealed significant changes in concentrations of hundreds of intracellular metabolites (Schulz *et al.* 2014). Although the current phosphoproteomic data are incomplete in terms of individual p-site detection, an analysis by our group has revealed that most of the phosphoproteins have already been detected (Vlastaridis *et al*., 2017), thus these conclusions appear robust.

A significant proportion of metabolic proteins are phosphorylated and yet there does not seem to be any major enrichment or depletion for phosphorylation in metabolic enzymes compared to the rest of the proteome (45% and 27% for ALL and HC respectively). Twelve percent (1668/14339) of ALL and 9% (499/5519) of HC p-sites are found in metabolic proteins (designated as phosphometabolic proteins). On average, phosphometabolic proteins have 4 and 2.5 p-sites (ALL and HC), whereas the rest of

the phosphoproteome have 5.7 and 3.7 p-sites respectively, a statistically significant difference (Wilcoxon p-value<0.006). 31% of ALL metabolic and 20% of HC metabolic p-sites are found within PFAM domains, indicating a potentially significant impact on structure, and probably on function. In contrast, 21% of ALL and 21% of HC p-sites are found within PFAM domains (see Table 1). Nevertheless, the next section shows that important enzymes tend to be regulated by phosphorylation.

**The general properties of yeast central metabolism likely to be regulated by phosphorylation**

The general properties of the phosphorylated metabolic proteins (designated as phosphometabolic), compared to the negative phosphometabolic proteins, are summarized in Figure 2 and in more detail in supplementary file S2, excel spreadsheet: stats in the publication. All subsequent reported differences are statistically significant (p-value < 0.05) and were performed with the appropriate Wilcoxon or Chi-squared test. Phosphometabolic proteins are i) significantly more abundant (305-540% higher), ii) have more kinase-target interactions (1-1.4 vs 0.3-0.4; 185-327% higher) iii) have longer total length (602-682 vs 369-388 aa; 55-76% higher), iv) longer intrinsically disordered regions (159-204 vs 71-74 aa; 114-175% higher), v) more protein-protein interactions (1-1.5 vs 0.5; 75-194% higher) and vi) regulate more reactions (4-5 vs 3.7-3.8; 3-36% higher). Furthermore, a higher fraction of them are ubiquitinated (37-53% vs 19-23%; 64-176% higher). It seems that there exists some synergism between protein phosphorylation and ubiquitination in the proteins of the yeast metabolic network (Tripodi *et al.* 2015). All the above conclusions hold even when controlling for protein abundance as a confounding factor. GOSlim analysis with Bingo (Maere *et al.* 2005) revealed an enrichment for the GO term “Vacuole”, when phosphometabolic proteins

were compared to the background (all metabolic proteins). In general, phosphometabolic proteins retain many of the general properties of the whole phosphoproteome (see figure 2), except the higher number of genetic interactions, the shorter protein half-lives (only for the Belle et al., dataset; conflicting results for the Christiano et al., dataset) and the higher fraction of essential genes (when controlling for protein abundance).

**Identification of p-sites in proteins that have a biotechnologically interesting phenotype related to metabolism and molecule production.**

The Saccharomyces Genome Database has mined and stored phenotypes caused by various gene perturbations, such as gene over/under-expression or even gene deletion. I manually inspected the phenotypes and focused on the ones that in my opinion are biotechnologically interesting. These phenotype terms mapped to 850 proteins, of which 408 were phosphoproteins, harboring 2363 p-sites. These phosphoproteins were not all annotated as participating in metabolism. By applying a stringent criterion of HC p-sites situated within conserved domains, I identified 180 of them in 73 phosphoproteins. These findings are summarized in Table 2. Obviously, there exist a significant number of very good candidate p-sites that may regulate biotechnologically important phenotypes, especially those related to increased chemical compound excretion and increased respiratory growth. These candidates should be the initial targets of future studies, e.g. to examine the phenotypic impact of deleting specific p-sites. Due to the inherent technical and biological noise of phosphorylation data, prioritization of p-sites for detailed study is an important task (Beltrao *et al.* 2012; Xiao *et al.* 2016). Readers can

perform their own customized prioritization on these data using supplementary file S1 of the publication.

| Phenotype_terms | P-sites/ proteins (ALL) | P-sites within domains / proteins (ALL) | P-sites/ proteins (HC) | P-sites within domains / proteins (HC) |
|---|---|---|---|---|
| chemical compound excretion: increased | 1497/248 | 284/ 189 | 564/147 | 109/43 |
| fermentative growth: increased | 7/3 | 1/1 | 2/1 | 0/0 |
| fermentative metabolism: increased | 85/10 | 10/6 | 38/10 | 3/3 |
| growth rate in exponential phase: increased | 73/8 | 14/5 | 38/6 | 9/2 |
| nutrient uptake/utilization: increased | 124/20 | 40/8 | 37/13 | 13/5 |
| respiratory growth: increased | 416/75 | 116/41 | 170/46 | 43/18 |
| respiratory metabolism: increased | 331/61 | 70/24 | 121/38 | 31/11 |
| utilization of carbon source: increased | 36/8 | 9/5 | 16/4 | 5/2 |
| vegetative growth: increased | 8/5 | 4/2 | 0/0 | 0/0 |
| viability: increased | 67/17 | 16/9 | 24/9 | 2/2 |
| ALL_RELATED_Phenotypes | 2363/408 | 496/183 | 887/247 | 180/73 |

Table 2. Number of p-sites that regulate proteins with a biotechnologically interesting phenotype.

# CONCLUSIONS

In summary, the integration of high-throughput data from various genomic, proteomic, and other functional sources has highlighted the pivotal role of protein phosphorylation in the control of yeast central metabolism, where almost half of the enzymes involved are phosphorylated. These phosphorylated enzymes, compared to the non-phosphorylated ones are more abundant, have more protein-protein interactions, regulate more reactions and a higher fraction of them are ubiquitinated. This analysis has also successfully identified and prioritized potential high-confidence p-sites that are likely to have a major impact on enzyme function and which should be targets of biotechnological and medical importance. The crucial question in this new era of high-throughput and integrative science is whether the numerous top-priority targets identified *in silico* will be investigated by low-throughput validation studies or by highly automated robotic procedures (King *et al.* 2004, 2009).

# FIGURES

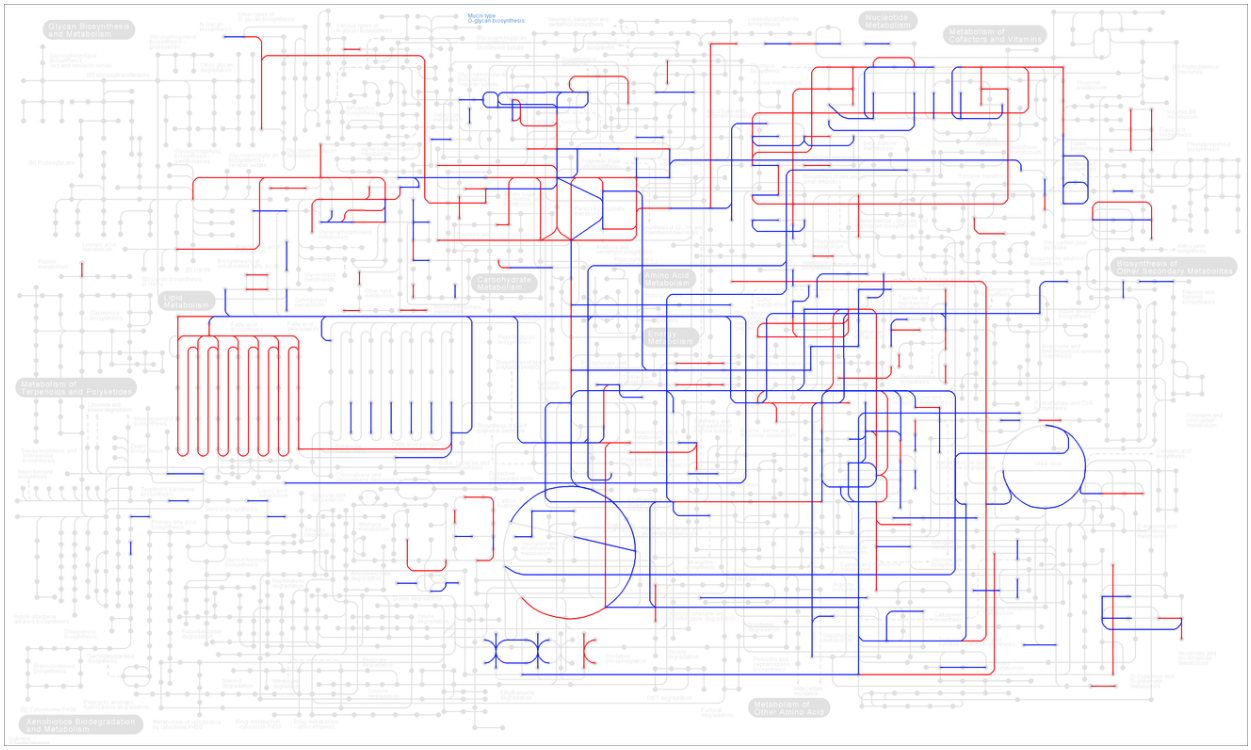


Figure 1. Protein phosphorylation is likely to exert significant control *over S. cerevisiae* central metabolism. Nodes represent metabolites and lines represent reactions in the KEGG metabolic map. Blue color is for reactions that are controlled by at least one enzyme that undergoes phosphorylation. Red color is for reactions that are controlled by at least one enzyme that contains High Confidence (HC) p-site/s. Mapping was performed with the KEGG mapper tool (Kanehisa *et al.* 2012), using the Uniprot identifiers of the yeast phosphorylated enzymes.

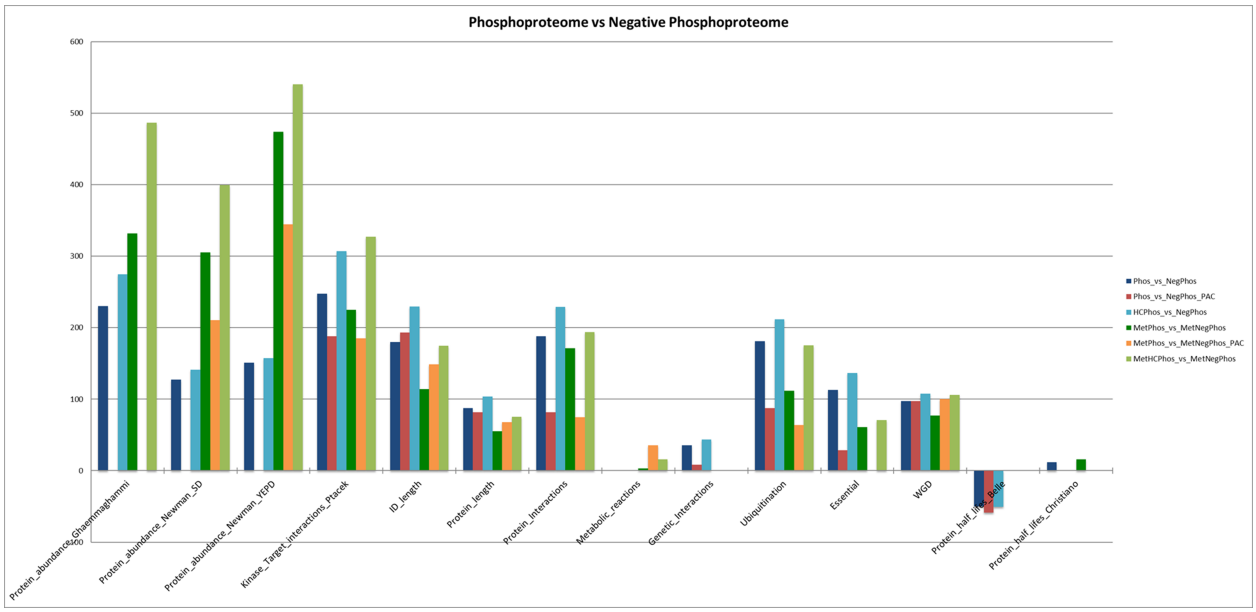



Figure 2. The general properties of the phosphoproteome, compared to the negative phosphoproteome. The bars show which properties of the phopshoproteome are higher/lower (% difference), compared to the negative phosphoproteome. Only statistically significant differences are shown. This is estimated for various datasets. Phos_vs_NegPhos: Phosphoproteome vs Negative Phosphoproteome; PAC stands for Protein Abundance Controlled dataset; HC stands for High Confidence subset of the Phosphoproteome. MetPhos_vs_MetNegPhos: Metabolic proteins of the Phosphoproteome vs Metabolic proteins of the Negative Phosphoproteome set.

# REFERENCES

Aguiar, M., W. Haas, S. A. Beausoleil, J. Rush, and S. P. Gygi, 2010 Gas-phase rearrangements do not affect site localization reliability in phosphoproteomics data sets. J. Proteome Res. 9: 3103–3107.

Albuquerque, C. P., M. B. Smolka, S. H. Payne, V. Bafna, J. Eng *et al.*, 2008 A multidimensional chromatography technology for in-depth phosphoproteome analysis. Mol. Cell Proteomics 7: 1389–1396.

Amoutzias, G. D., Y. He, K. S. Lilley, Y. Van de Peer, and S. G. Oliver, 2012 Evaluation and properties of the budding yeast phosphoproteome. Mol. Cell Proteomics 11: M111.009555.

Batada, N. N., T. Reguly, A. Breitkreutz, L. Boucher, B.-J. Breitkreutz *et al.*, 2006 Stratus not altocumulus: a new view of the yeast protein interaction network. PLoS Biol. 4: e317.

Belle, A., A. Tanay, L. Bitincka, R. Shamir, and E. K. O'Shea, 2006 Quantification of protein half-lives in the budding yeast proteome. Proc. Natl. Acad. Sci. U.S.A. 103: 13004–13009.

Beltrao, P., V. Albanèse, L. R. Kenner, D. L. Swaney, A. Burlingame *et al.*, 2012 Systematic functional prioritization of protein posttranslational modifications. Cell 150: 413–425.

Beltrao, P., J. C. Trinidad, D. Fiedler, A. Roguev, W. A. Lim *et al.*, 2009 Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. PLoS Biol. 7: e1000134.

Bodenmiller, B., D. Campbell, B. Gerrits, H. Lam, M. Jovanovic *et al.*, 2008 PhosphoPep--a database of protein phosphorylation sites in model organisms. Nat. Biotechnol. 26: 1339–1340.

Bodenmiller, B., S. Wanka, C. Kraft, J. Urban, D. Campbell *et al.*, 2010 Phosphoproteomic analysis reveals interconnected system-wide responses to perturbations of kinases and phosphatases in yeast. Sci Signal 3: rs4.

Chen, Y., and J. Nielsen, 2016 Flux Control through Protein Phosphorylation in Yeast. FEMS Yeast Research fow096.

Cherry, J. M., E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley *et al.*, 2012 Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. 40: D700-705.

Chi, A., C. Huttenhower, L. Y. Geer, J. J. Coon, J. E. P. Syka *et al.*, 2007 Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. Proc. Natl. Acad. Sci. U.S.A. 104: 2193–2198.

Christiano, R., N. Nagaraj, F. Fröhlich, and T. C. Walther, 2014 Global proteome turnover analyses of the Yeasts S. cerevisiae and S. pombe. Cell Rep 9: 1959–1965.

Costanzo, M., A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear *et al.*, 2010 The genetic landscape of a cell. Science 327: 425–431.

Dobson, P. D., K. Smallbone, D. Jameson, E. Simeonidis, K. Lanthaler *et al.*, 2010 Further developments towards a genome-scale metabolic model of yeast. BMC Syst Biol 4: 145.

Galanie, S., K. Thodey, I. J. Trenchard, M. Filsinger Interrante, and C. D. Smolke, 2015 Complete biosynthesis of opioids in yeast. Science 349: 1095–1100.

Ghaemmaghami, S., W.-K. Huh, K. Bower, R. W. Howson, A. Belle *et al.*, 2003 Global analysis of protein expression in yeast. Nature 425: 737–741.

Giaever, G., A. M. Chu, L. Ni, C. Connelly, L. Riles *et al.*, 2002 Functional profiling of the Saccharomyces cerevisiae genome. Nature 418: 387–391.

Gnad, F., L. M. F. de Godoy, J. Cox, N. Neuhauser, S. Ren *et al.*, 2009 High-accuracy identification and bioinformatic analysis of in vivo protein phosphorylation sites in yeast. Proteomics 9: 4642–4652.

Gruhler, A., J. V. Olsen, S. Mohammed, P. Mortensen, N. J. Faergeman *et al.*, 2005 Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. Mol. Cell Proteomics 4: 310–327.

Herrgård, M. J., N. Swainston, P. Dobson, W. B. Dunn, K. Y. Arga *et al.*, 2008 A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat. Biotechnol. 26: 1155–1160.

Holt, L. J., B. B. Tuch, J. Villén, A. D. Johnson, S. P. Gygi *et al.*, 2009 Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. Science 325: 1682–1686.

Huber, A., B. Bodenmiller, A. Uotila, M. Stahl, S. Wanka *et al.*, 2009 Characterization of the rapamycin-sensitive phosphoproteome reveals that Sch9 is a central coordinator of protein synthesis. Genes Dev. 23: 1929–1943.

Kanehisa, M., S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, 2012 KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40: D109-114.

King, R. D., J. Rowland, S. G. Oliver, M. Young, W. Aubrey *et al.*, 2009 The automation of science. Science 324: 85–89.

King, R. D., K. E. Whelan, F. M. Jones, P. G. K. Reiser, C. H. Bryant *et al.*, 2004 Functional genomic hypothesis generation and experimentation by a robot scientist. Nature 427: 247–252.

Landry, C. R., L. Freschi, T. Zarin, and A. M. Moses, 2014 Turnover of protein phosphorylation evolving under stabilizing selection. Front Genet 5: 245.

Landry, C. R., E. D. Levy, and S. W. Michnick, 2009 Weak functional constraints on phosphoproteomes. Trends Genet. 25: 193–197.

Lee, J., W. Reiter, I. Dohnal, C. Gregori, S. Beese-Sims *et al.*, 2013 MAPK Hog1 closes the S. cerevisiae glycerol channel Fps1 by phosphorylating and displacing its positive regulators. Genes Dev. 27: 2590–2601.

Levy, E. D., S. W. Michnick, and C. R. Landry, 2012 Protein abundance is key to distinguish promiscuous from functional phosphorylation based on evolutionary information. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 367: 2594–2606.

Li, X., S. A. Gerber, A. D. Rudner, S. A. Beausoleil, W. Haas *et al.*, 2007 Large-scale phosphorylation analysis of alpha-factor-arrested Saccharomyces cerevisiae. J. Proteome Res. 6: 1190–1197.

Lienhard, G. E., 2008 Non-functional phosphorylations? Trends Biochem. Sci. 33: 351–352.

Maere, S., K. Heymans, and M. Kuiper, 2005 BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21: 3448–3449.

Mascaraque, V., M. L. Hernáez, M. Jiménez-Sánchez, R. Hansen, C. Gil *et al.*, 2013 Phosphoproteomic analysis of protein kinase C signaling in Saccharomyces cerevisiae reveals Slt2 mitogen-activated protein kinase (MAPK)-dependent phosphorylation of eisosome core components. Mol. Cell Proteomics 12: 557–574.

Newman, J. R. S., S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble *et al.*, 2006 Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature 441: 840–846.

Nielsen, J., 2015 BIOENGINEERING. Yeast cell factories on the horizon. Science 349: 1050–1051.

Oliveira, A. P., C. Ludwig, P. Picotti, M. Kogadeeva, R. Aebersold *et al.*, 2012 Regulation of yeast central metabolism by enzyme phosphorylation. Mol. Syst. Biol. 8: 623.

Oliveira, A. P., and U. Sauer, 2012 The importance of post-translational modifications in regulating Saccharomyces cerevisiae metabolism. FEMS Yeast Res. 12: 104–117.

Orth, J. D., I. Thiele, and B. Ø. Palsson, 2010 What is flux balance analysis? Nat. Biotechnol. 28: 245–248.

Pache, R. A., M. M. Babu, and P. Aloy, 2009 Exploiting gene deletion fitness effects in yeast to understand the modular architecture of protein complexes under different growth conditions. BMC Syst Biol 3: 74.

Peng, J., D. Schwartz, J. E. Elias, C. C. Thoreen, D. Cheng *et al.*, 2003 A proteomics approach to understanding protein ubiquitination. Nat. Biotechnol. 21: 921–926.

R Core Team, 2015 *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sadowski, I., B.-J. Breitkreutz, C. Stark, T.-C. Su, M. Dahabieh *et al.*, 2013 The PhosphoGRID Saccharomyces cerevisiae protein phosphorylation site database: version 2.0 update. Database (Oxford) 2013: bat026.

Saleem, R. A., R. S. Rogers, A. V. Ratushny, D. J. Dilworth, P. T. Shannon *et al.*, 2010 Integrated phosphoproteomics analysis of a signaling network governing nutrient response and peroxisome induction. Mol. Cell Proteomics 9: 2076–2088.

Schulz, J. C., M. Zampieri, S. Wanka, C. von Mering, and U. Sauer, 2014 Large-scale functional analysis of the roles of phosphorylation in yeast metabolic pathways. Sci Signal 7: rs6.

Schwanhäusser, B., D. Busse, N. Li, G. Dittmar, J. Schuchhardt *et al.*, 2011 Global quantification of mammalian gene expression control. Nature 473: 337–342.

Soufi, B., C. D. Kelstrup, G. Stoehr, F. Fröhlich, T. C. Walther *et al.*, 2009 Global analysis of the yeast osmotic stress response by quantitative proteomics. Mol Biosyst 5: 1337–1346.

Steinmetz, L. M., C. Scharfe, A. M. Deutschbauer, D. Mokranjac, Z. S. Herman *et al.*, 2002 Systematic screen for human disease genes in yeast. Nat. Genet. 31: 400–404.

Szappanos, B., K. Kovács, B. Szamecz, F. Honti, M. Costanzo *et al.*, 2011 An integrated approach to characterize genetic interaction networks in yeast metabolism. Nat. Genet. 43: 656–662.

Szczebara, F. M., C. Chandelier, C. Villeret, A. Masurel, S. Bourot *et al.*, 2003 Total biosynthesis of hydrocortisone from a simple carbon source in yeast. Nat. Biotechnol. 21: 143–149.

Tripodi, F., R. Nicastro, V. Reghellin, and P. Coccetti, 2015 Post-translational modifications on yeast carbon metabolism: Regulatory mechanisms beyond transcriptional control. Biochim. Biophys. Acta 1850: 620–627.

Vlastaridis, P., P. Kyriakidou, A. Chaliotis, Y. Van de Peer, S. G. Oliver *et al.*, 2017 Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes. Gigascience.

Weinert, B. T., V. Iesmantavicius, T. Moustafa, C. Schölz, S. A. Wagner *et al.*, 2014 Acetylation dynamics and stoichiometry in Saccharomyces cerevisiae. Mol. Syst. Biol. 10: 716.

Wu, R., N. Dephoure, W. Haas, E. L. Huttlin, B. Zhai *et al.*, 2011 Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes. Mol. Cell Proteomics 10: M111.009654.

Xiao, Q., B. Miao, J. Bi, Z. Wang, and Y. Li, 2016 Prioritizing functional phosphorylation sites based on multiple feature integration. Sci Rep 6: 24735.

# Chapter 5

# Title: Meth-Phos-Prometheus: A webserver for the prediction of protein methylation sites, protein phosphorylation sites, their clusters and combinatorial switches.

## Preface

This chapter describes how our FAB-PHOS team further mined methyl-proteomic data, integrated them with phosphorylation data and developed a neural network web server that predicts phosphorylation sites, methylation sites, as well as their meth-phos switches and clusters. In this chapter, other people contributed to the mining of the methyl-proteomic data and the training of the methylation neural network whereas I trained the phosphorylation neural network and developed the web-server that integrates the results from the phosphorylation and methylation neural networks and visualizes them. This chapter has been prepared in the form of a research article for submission in a peer-reviewed journal, like *Bioinformatics*, in the near future. However, some of the data that we are planning to include as supplementary material in the submitted manuscript have been integrated within this chapter, due to relaxed limitations on word count. The server is accessible at:
http://bioinf.bio.uth.gr/meth-phos-prometheus/

## Abstract

**Summary:** We have developed Meth-Phos-Prometheus, a Neural Network (NN) webserver that accurately predicts in eukaryotes i) protein methylation sites, ii)

protein phosphorylation sites iii) the meth/phos switches they form iv) the clusters they form and v) further displays them graphically on the protein. The protein methylation NN has an accuracy of 86% (MCC: 0.72; sensitivity: 85%; specificity: 87%; AUC: 0.925; Precision: 86%), whereas the protein phosphorylation NN has an accuracy of 84% (MCC: 0.68; sensitivity: 86%; specificity: 83%; AUC: 0.91; Precision: 83%), thus significantly outperforming many other published prediction tools.

**Availability and implementation:** Freely available on the web at http://bioinf.bio.uth.gr/meth-phos-prometheus/. Website implemented in Python, Keras/Tensorflow, Java, Jhipster Application Framework, Angular Javascript Framework and Apache, with all major browsers supported.

# Introduction

Protein phosphorylation is the most frequently detected post-translational modification (PTM), with a very well established role in various diseases, cell physiology, signal transduction, (Vlastaridis *et al.*, 2016; Needham *et al.*, 2019) and with great potential for synthetic biology and biotechnology (Vlastaridis, Papakyriakou, *et al.*, 2017). More than 200,000 phosphorylation sites are expected to be found in more than half of the human proteins (Vlastaridis, Kyriakidou, Chaliotis, Van de Peer, Stephen G. Oliver, and Amoutzias, 2017). Advances in high-throughput proteomics have recently revealed the widespread incidence of protein methylation, an emerging post-translational modification (PTM), in eukaryotes, prokaryotes, and viruses, and as a promising therapeutic target in cancer, metabolic, neurodegenerative, muscular disorders and in stem-cell reprogramming (Lanouette *et al.*, 2014; Larsen *et al.*, 2016a; Blanc and Richard, 2017; Murn and Shi, 2017). Protein methylation involves the enzymatic addition of one, two, or even three methyl groups onto the side chain of amino acids and occurs most frequently on arginine and lysine residues. Despite recent advances in this field, technical challenges still hinder the detection of the complete set of methylation sites in human and most model organisms. So far,

high and low throughput proteomic studies have revealed 16000 unique lysine and arginine methylation events in 5500 human proteins, (Hornbeck *et al.*, 2015; Murn and Shi, 2017), but some of them may be of questionable quality (Hart-Smith *et al.*, 2016). By applying two independent methods on the current data, Capture-Recapture and Curve-fitting of the saturation curve, which we have previously applied to the identification of phosphorylation sites (Vlastaridis, Kyriakidou, Chaliotis, Van de Peer, Stephen G Oliver, and Amoutzias, 2017), we estimate that the total human methyl-proteome is between 23,000-44,000 methylation sites in 6,500-8,400 proteins.

Protein methylation may affect (block or promote) or be affected by other types of modifications in neighbouring residues, particularly phosphorylation (Larsen *et al.*, 2016a). Such neighboring PTMs interact with each other and may create higher order molecular AND/OR/NOT switches that in turn create signaling networks with properties like bistability, robustness and adaptability (Schreiber and Bernstein, 2002). The histone code (Strahl and Allis, 2000; Kouzarides, 2007; Sims and Reinberg, 2008) constitutes the best known case of functional methylation/phosphorylation (meth/phos) switches (Fischle *et al.*, 2005, 2003). Furthermore, this particular co-localization of methylation and phosphorylation sites is very frequent at the whole proteome level (Larsen *et al.*, 2016a), although their actual functional role still remains untested, for the vast majority of them. Targeted wet-lab experiments have revealed the important functional role of such meth/phos switches for protein-protein interactions, signal transduction, mitosis, the progression of colorectal and gastric cancers, for stem cells and differentiation, for DNA damage repair and apoptosis (Hong *et al.*, 2018; Noh *et al.*, 2015; Varier *et al.*, 2010; Fang *et al.*, 2014; Andrews *et al.*, 2016; Estève *et al.*, 2011; Li *et al.*, 2019; Song *et al.*, 2018; Poulard *et al.*, 2017; Yamagata *et al.*, 2008; Hsu *et al.*, 2011).

Furthermore, protein phosphorylation sites tend to cluster together (Amoutzias *et al.*, 2012; Moses *et al.*, 2007; Schweiger and Linial, 2010) and form molecular

rheostats (with an analog effect) or molecular barcodes/clusters, where different combinations lead to different outcomes (Landry *et al.*, 2014). Given the emerging role of protein methylation as well, it is reasonable to assume that such clusters must be present for this type of PTM as well.

Many machine-learning computational tools have been developed to predict either protein methylation (Chen *et al.*, 2016; Deng *et al.*, 2017; Ju *et al.*, 2015; Kumar *et al.*, 2017; Lee *et al.*, 2014; Qiu *et al.*, 2016, 2014; Shao *et al.*, 2009; Shi *et al.*, 2015; Shien *et al.*, 2009; Wei *et al.*, 2017; Wen *et al.*, 2016) or phosphorylation sites (Trost and Kusalik, 2011, 2013; Blom *et al.*, 1999; Xue *et al.*, 2008; Luo *et al.*, 2019; Wang *et al.*, 2017), in order to overcome the lack of experimental data and guide the experiments of wet-lab scientists. However, no available tool combines these two predictions in order to identify molecular meth/phos switches and/or clusters that could potentially function as rheostats at the protein or proteome level. In addition, many of the published tools are found to display an imbalanced sensitivity/specificity ratio, to be incapable of analyzing more than a few proteins or to be unavailable as user-friendly webservers. Their performance is heavily dependent upon abundant and high-quality training datasets; yet, many recent experiments that detect thousands of new methylation and phosphorylation sites have not been utilized, even by the most recent algorithms. Also, several reports emphasize the large number of false positive or non-functional p-sites and m-sites that are detected by high-throughput proteomics (Hart-Smith *et al.*, 2016; Landry *et al.*, 2009; Lienhard, 2008). We thus focused on training two NNs (one for methylation and one phosphorylation) with recent, abundant and highly filtered quality data, instead of training them with a large number of dubious-quality data.

# Materials and Methods

## The eukaryotic phosphoproteomic compendium

Our eukaryotic phosphoproteomic compendium consisted of 53,585 high quality phosphorylation sites (p-sites) (41,863 serines, 7,204 threonines, 4,518 tyrosines) and was compiled by manually mining and filtering 216 published datasets in human, mouse, yeast and *Arabidopsis thaliana* (see supplementary tables 1-4). We applied stringent filtering criteria, including 99% correct phosphorylated peptide identification and 99% correct phosphorylation site localization, whenever applicable. In addition, we retrieved human and mouse phosphorylation sites from low-throughput studies documented in Phosphosite+ (Hornbeck *et al.*, 2019) and yeast phosphorylation sites from low-throughput studies documented in PhosphoGrid2 (Sadowski *et al.*, 2013). We only included p-sites that had been identified in any low-throughput study (availalble in Phosphosite+ or PhosphoGrid2) and/or identified in at least three of the high-throughput published datasets that we filtered. This way, we tried to enrich our dataset for true positive p-sites with a functional role. We have already shown that p-sites identified in several experiments have a higher probability of being functional (Amoutzias *et al.*, 2012).

More specifically, the very stringent human set contained 24099, 4559 and 4005 phosphorylated serines, threonines, tyrosines, respectively, that had been identified in any low-throughput study (availalble in Phosphosite+) and/or identified in at least three of the 106 high-throughput published datasets that we filtered. Of note, the publication of (Bian *et al.*, 2016) provided 7 different datasets. The human phosphoproteomic datasets that we used are shown in Table **1**. (Alayev *et al.*, 2014; Aslanian *et al.*, 2014; Bai *et al.*, 2012; Beck *et al.*, 2014; Beltran *et al.*, 2012; Bian *et al.*, 2016, 2014, 2012, 2013; Breitkopf *et al.*,

2010; Brill *et al.*, 2009; Cantin *et al.*, 2006; Casado *et al.*, 2013, 2014; Chen *et al.*, 2011; Christensen *et al.*, 2010; Chylek *et al.*, 2014; Dammer *et al.*, 2015; Daub *et al.*, 2008; Di Palma *et al.*, 2013; Everley and Dillman, 2010; Ficarro *et al.*, 2011a; Fleitz *et al.*, 2013; Francavilla *et al.*, 2013; Franchin *et al.*, 2014, 2015; Franz-Wachtel *et al.*, 2012; Galan *et al.*, 2014; Ge *et al.*, 2010; Gerarduzzi *et al.*, 2014; Giansanti *et al.*, 2014, 2013; Glowinski *et al.*, 2014; Goss *et al.*, 2006; Grosstessner-Hain *et al.*, 2011; Hammond *et al.*, 2010; Han *et al.*, 2008; Harder *et al.*, 2014; Helm *et al.*, 2014; Helou *et al.*, 2013; Herskowitz *et al.*, 2010; Højlund *et al.*, 2009; Hornbeck *et al.*, 2019; Iliuk *et al.*, 2012; Joughin *et al.*, 2009; Jouy *et al.*, 2015; Kettenbach and Gerber, 2011; Kettenbach *et al.*, 2011, 2012; Klammer *et al.*, 2014; Lai *et al.*, 2012; Luerman *et al.*, 2014; Ly *et al.*, 2014; Mäusbacher *et al.*, 2010; McNulty and Annan, 2008; Melo-Braga *et al.*, 2014; Narumi *et al.*, 2012; Nguyen *et al.*, 2009; Old *et al.*, 2009; Osinalde *et al.*, 2015; Ozlü *et al.*, 2010; Palmisano *et al.*, 2012a; Park and Maudsley, 2011; Ruperez *et al.*, 2012; Ruse *et al.*, 2008; Šalovská *et al.*, 2014; Santamaria *et al.*, 2011; Schweppe *et al.*, 2013; Sharma *et al.*, 2014, 2012; Shevchuk *et al.*, 2014; Shiromizu *et al.*, 2013; Soderblom *et al.*, 2013; Song *et al.*, 2012; Stokes *et al.*, 2012a; Sui *et al.*, 2008; Tan *et al.*, 2015; Taus *et al.*, 2011; Tong *et al.*, 2017; van den Biggelaar *et al.*, 2014; Van Hoof *et al.*, 2009; C. Wang *et al.*, 2013; Weber *et al.*, 2012; Wiese *et al.*, 2015; Wojcechowskyj *et al.*, 2011; F. Wu *et al.*, 2010; J. Wu *et al.*, 2010; Xia *et al.*, 2008; Xiao *et al.*, 2010; Xie *et al.*, 2010; Xue *et al.*, 2012; Yan *et al.*, 2011; Yang *et al.*, 2007, 2010; X.-L. Yang *et al.*, 2013; Yao *et al.*, 2011; Ye and Li, 2014; H. Zhang *et al.*, 2013; L. Zhang *et al.*, 2013; Zhang *et al.*, 2017; Zheng *et al.*, 2013).

| Publication Year | Title | Publication Title | PMID |
|---|---|---|---|
| 2006 | Quantitative phosphoproteomic analysis of the tumor necrosis factor pathway | Journal of Proteome Research | 16396503 |
| 2006 | A common phosphotyrosine signature for the Bcr-Abl kinase | Blood | 16497976 |
| 2007 | Applying a targeted label-free approach using LC-MS AMT tags to | Journal of Proteome Research | 17929957 |

| | | | |
|---|---|---|---|
| | evaluate changes in protein phosphorylation following phosphatase inhibition | | |
| 2008 | Hydrophilic interaction chromatography reduces the complexity of the phosphoproteome and improves global phosphopeptide isolation and detection | Molecular & cellular proteomics: MCP | 18212344 |
| 2008 | Large-scale phosphoproteome analysis of human liver tissue by enrichment and fractionation of phosphopeptides with strong anion exchange chromatography | Proteomics | 18318008 |
| 2008 | Motif-specific sampling of phosphoproteomes | Journal of Proteome Research | 18452278 |
| 2008 | Phosphoproteome analysis of the human Chang liver cells using SCX and a complementary mass spectrometric strategy | Proteomics | 18491316 |
| 2008 | Phosphoproteomic analysis of human brain by calcium phosphate precipitation and mass spectrometry | Journal of Proteome Research | 18510355 |
| 2008 | Kinase-selective enrichment enables quantitative phosphoproteomics of the kinome across the cell cycle | Molecular Cell | 18691976 |
| 2009 | An integrated comparative phosphoproteomic and bioinformatic approach reveals a novel class of MPM-2 motifs upregulated in EGFRvIII-expressing glioblastoma cells | Molecular bioSystems | 19081932 |
| 2009 | Functional proteomics identifies targets of phosphorylation by B-Raf signaling in melanoma | Molecular Cell | 19362540 |
| 2009 | A new approach for quantitative phosphoproteomic dissection of signaling pathways applied to T cell receptor activation | Molecular & cellular proteomics: MCP | 19605366 |
| 2009 | Phosphoproteomic analysis of human embryonic stem cells | Cell Stem Cell | 19664994 |
| 2009 | Phosphorylation dynamics during early differentiation of human embryonic stem cells | Cell Stem Cell | 19664995 |
| 2009 | In vivo phosphoproteome of human skeletal muscle revealed by phosphopeptide enrichment and HPLC-ESI-MS/MS | Journal of Proteome Research | 19764811 |
| 2010 | Binding partner switching on microtubules and aurora-B in the mitosis to cytokinesis transition | Molecular & cellular proteomics: MCP | 19786723 |
| 2010 | A large-scale quantitative proteomic approach to identifying sulfur mustard-induced protein | Chemical Research in Toxicology | 19845377 |

| | | | |
|---|---|---|---|
| | phosphorylation cascades | | |
| 2010 | Quantitative analysis of HGF and EGF-dependent phosphotyrosine signaling networks | Journal of Proteome Research | 20222723 |
| 2010 | Phosphoproteomic analysis of primary human multiple myeloma cells | Journal of Proteomics | 20230923 |
| 2010 | Integrating titania enrichment, iTRAQ labeling, and Orbitrap CID-HCD for global identification and quantitative analysis of phosphopeptides | Proteomics | 20340162 |
| 2010 | Quantitative phosphoproteomics dissection of seven-transmembrane receptor signaling using full and biased agonists | Molecular & cellular proteomics: MCP | 20363803 |
| 2010 | Studies of phosphoproteomic changes induced by nucleophosmin-anaplastic lymphoma kinase (ALK) highlight deregulation of tumor necrosis factor (TNF)/Fas/TNF-related apoptosis-induced ligand signaling pathway in ALK-positive anaplastic large cell lymphoma | Molecular & cellular proteomics: MCP | 20393185 |
| 2010 | A comparative phosphoproteomic analysis of a human tumor metastasis model using a label-free quantitative approach | Electrophoresis | 20446291 |
| 2010 | Glycoprotein capture and quantitative phosphoproteomics indicate coordinated regulation of cell migration upon lysophosphatidic acid stimulation | Molecular & cellular proteomics: MCP | 20639409 |
| 2010 | Global phosphorylation analysis of beta-arrestin-mediated signaling downstream of a seven transmembrane receptor (7TMR) | Proceedings of the National Academy of Sciences of the United States of America | 20686112 |
| 2011 | Characterization of phosphoproteins in gastric cancer secretome | Omics: A Journal of Integrative Biology | 20726782 |
| 2011 | The Plk1-dependent phosphoproteome of the early mitotic spindle | Molecular & cellular proteomics: MCP | 20860994 |
| 2010 | Proteomics analysis of cellular imatinib targets and their candidate downstream effectors | Journal of Proteome Research | 20866107 |
| 2010 | Phosphoproteomic analysis reveals site-specific changes in GFAP and NDRG2 phosphorylation in frontotemporal lobar degeneration | Journal of Proteome Research | 20886841 |
| 2011 | Discontinuous pH gradient-mediated separation of TiO2-enriched phosphopeptides | Analytical Biochemistry | 20946866 |
| 2011 | Increasing phosphoproteome coverage and identification of | Journal of Chromatography. B, | 21130716 |

| | | | |
|---|---|---|---|
| | phosphorylation motifs through combination of different HPLC fractionation methods | Analytical Technologies in the Biomedical and Life Sciences | |
| 2010 | Phosphoproteomics profiling of human skin fibroblast cells reveals pathways and proteins affected by low doses of ionizing radiation | PloS One | 21152398 |
| 2011 | SUMOylation-regulated protein phosphorylation, evidence from quantitative phosphoproteomics analyses | The Journal of Biological Chemistry | 21685386 |
| 2011 | Quantitative phosphoproteomics identifies substrates and functional modules of Aurora and Polo-like kinase activities in mitotic cells | Science Signaling | 21712546 |
| 2011 | Online nanoflow multidimensional fractionation for high efficiency phosphopeptide analysis | Molecular & cellular proteomics: MCP | 21788404 |
| 2011 | Quantitative phospho-proteomics to investigate the polo-like kinase 1-dependent phospho-proteome | Molecular & cellular proteomics: MCP | 21857030 |
| 2011 | Rapid and reproducible single-stage phosphopeptide enrichment of complex peptide mixtures: application to general and phosphotyrosine-specific phosphoproteomics experiments | Analytical Chemistry | 21899308 |
| 2011 | Quantitative phosphoproteomics of CXCL12 (SDF-1) signaling | PloS One | 21949786 |
| 2011 | Universal and confident phosphorylation site localization using phosphoRS | Journal of Proteome Research | 22073976 |
| 2012 | Dual phosphoproteomics and chemical proteomics analysis of erlotinib and gefitinib interference in acute myeloid leukemia cells | Journal of Proteomics | 22115753 |
| 2012 | Quantitative proteomics reveals that Hsp90 inhibition preferentially targets kinases and the DNA damage response | Molecular & cellular proteomics: MCP | 22167270 |
| 2012 | PTMScan direct: identification and quantification of peptides from critical signaling proteins by immunoaffinity enrichment coupled with LC-MS/MS | Molecular & cellular proteomics: MCP | 22322096 |
| 2012 | Sensitive kinase assay linked with phosphoproteomics for identifying direct kinase substrates | Proceedings of the National Academy of Sciences of the United States of America | 22451900 |
| 2012 | Phosphoproteomics identifies driver tyrosine kinases in sarcoma cell lines and tumors | Cancer Research | 22461510 |
| 2012 | Improve the coverage for the analysis of phosphoproteome of HeLa cells by | Journal of Proteome Research | 22468782 |

| | | | |
|---|---|---|---|
| | a tandem digestion approach | | |
| 2012 | Global detection of protein kinase D-dependent phosphorylation events in nocodazole-treated human cells | Molecular & cellular proteomics: MCP | 22496350 |
| 2012 | Quantitative phosphoproteomic analysis reveals a role for serine and threonine kinases in the cytoskeletal reorganization in early T cell receptor activation in human primary T cells | Molecular & cellular proteomics: MCP | 22499768 |
| 2012 | Chemical visualization of phosphoproteomes on membrane | Molecular & cellular proteomics: MCP | 22593177 |
| 2012 | Rapid determination of multiple linear kinase substrate motifs by mass spectrometry | Chemistry & Biology | 22633412 |
| 2012 | Systematic analysis of protein phosphorylation networks from phosphoproteomic data | Molecular & cellular proteomics: MCP | 22798277 |
| 2012 | A novel method for the simultaneous enrichment, identification, and quantification of phosphopeptides and sialylated glycopeptides applied to a temporal profile of mouse brain development | Molecular & cellular proteomics: MCP | 22843994 |
| 2012 | Complementary Fe(3+)- and Ti(4+)-immobilized metal ion affinity chromatography for purification of acidic and basic phosphopeptides | Rapid communications in mass spectrometry: RCM | 22886815 |
| 2012 | A strategy for large-scale phosphoproteomics and SRM-based validation of human breast cancer tissue samples | Journal of Proteome Research | 22985185 |
| 2012 | Global profiling of protein kinase activities in cancer cells by mass spectrometry | Journal of Proteomics | 23041048 |
| 2013 | Proteomic analysis of ERK1/2-mediated human sickle red blood cell membrane protein phosphorylation | Clinical Proteomics | 23286773 |
| 2013 | Identification of missing proteins in the neXtProt database and unregistered phosphopeptides in the PhosphoSitePlus database as part of the Chromosome-centric Human Proteome Project | Journal of Proteome Research | 23312004 |
| 2013 | Identification of complex relationship between protein kinases and substrates during the cell cycle of HeLa cells by phosphoproteomic analysis | Proteomics | 23322592 |
| 2013 | Phosphoproteomics data classify hematological cancer cell lines according to tumor type and sensitivity to kinase inhibitors | Genome Biology | 23628362 |

| | | | |
|---|---|---|---|
| 2013 | Characterization of the novel broad-spectrum kinase inhibitor CTx-0294885 as an affinity reagent for mass spectrometry-based kinome profiling | Journal of Proteome Research | 23692254 |
| 2013 | Determination of CK2 specificity and substrates by proteome-derived peptide libraries | Journal of Proteome Research | 23808766 |
| 2013 | ERK positive feedback regulates a widespread network of tyrosine phosphorylation sites across canonical T cell signaling and actin cytoskeletal proteins in Jurkat T cells | PloS One | 23874979 |
| 2013 | Interrogating cAMP-dependent kinase signaling in Jurkat T cells via a protein kinase A targeted immune-precipitation phosphoproteomics approach | Molecular & cellular proteomics: MCP | 23882029 |
| 2013 | Enhanced detection of multiply phosphorylated peptides and identification of their sites of modification | Analytical Chemistry | 23889490 |
| 2013 | Quantitative phosphoproteomic profiling of human non-small cell lung cancer tumors | Journal of Proteomics | 23911959 |
| 2013 | Finding the same needles in the haystack? A comparison of phosphotyrosine peptides enriched by immuno-affinity precipitation and metal-based affinity chromatography | Journal of Proteomics | 23917254 |
| 2013 | Functional proteomics defines the molecular switch underlying FGF receptor trafficking and cellular outputs | Molecular Cell | 24011590 |
| 2014 | Phosphoproteomic evaluation of pharmacological inhibition of leucine-rich repeat kinase 2 reveals significant off-target effects of LRRK-2-IN-1 | Journal of Neurochemistry | 24117733 |
| 2013 | SILAC-based phosphoproteomics reveals an inhibitory role of KSR1 in p53 transcriptional activity via modulation of DBC1 | British Journal of Cancer | 24129246 |
| 2014 | Comprehensive quantitative comparison of the membrane proteome, phosphoproteome, and sialiome of human embryonic and neural stem cells | Molecular & cellular proteomics: MCP | 24173317 |
| 2013 | Urinary proteomic and non-prefractionation quantitative phosphoproteomic analysis during pregnancy and non-pregnancy | BMC genomics | 24215720 |
| 2014 | An enzyme assisted RP-RPLC | Journal of Proteomics | 24275569 |

| | | | |
|---|---|---|---|
| | approach for in-depth analysis of human liver phosphoproteome | | |
| 2014 | Inducing autophagy: a comparative phosphoproteomic study of the cellular response to ammonia and rapamycin | Autophagy | 24300666 |
| 2013 | Global screening of CK2 kinase substrates by an integrated phosphoproteomics workflow | Scientific Reports | 24322422 |
| 2014 | Time-resolved characterization of cAMP/PKA-dependent signaling reveals that platelet inhibition is a concerted process involving multiple signaling pathways | Blood | 24324209 |
| 2014 | Environmental stress affects the activity of metabolic and growth factor signaling networks and induces autophagy markers in MCF7 breast cancer cells | Molecular & cellular proteomics: MCP | 24425749 |
| 2014 | Mass spectrometry-based quantification of the cellular response to methyl methanesulfonate treatment in human cells | DNA repair | 24461736 |
| 2014 | Quantitative phosphoproteomics unveils temporal dynamics of thrombin signaling in human endothelial cells | Blood | 24501219 |
| 2014 | A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells | eLife | 24596151 |
| 2014 | HOPE-fixation of lung tissue allows retrospective proteome and phosphoproteome studies | Journal of Proteome Research | 24702127 |
| 2014 | Evaluating the promiscuous nature of tyrosine kinase inhibitors assessed in A431 epidermoid carcinoma cells by both chemical- and phosphoproteomics | ACS chemical biology | 24804581 |
| 2014 | Macroporous reversed-phase separation of proteins combined with reversed-phase separation of phosphopeptides and tandem mass spectrometry for profiling the phosphoproteome of MDA-MB-231 cells | Electrophoresis | 24888630 |
| 2014 | Phosphoproteomic analysis identifies the tumor suppressor PDCD4 as a RSK substrate negatively regulated by 14-3-3 | Proceedings of the National Academy of Sciences of the United States of America | 25002506 |
| 2014 | Radiosensitization of human leukemic HL-60 cells by ATR kinase inhibitor (VE-821): phosphoproteomic analysis | International Journal of Molecular Sciences | 25003641 |
| 2014 | Analysis of T4SS-induced signaling | Frontiers in Microbiology | 25101063 |

| | | | |
|---|---|---|---|
| | by H. pylori using quantitative phosphoproteomics | | |
| 2014 | Ion mobility tandem mass spectrometry enhances performance of bottom-up proteomics | Molecular & cellular proteomics: MCP | 25106551 |
| 2014 | Identification of significant features by the Global Mean Rank test | PloS One | 25119995 |
| 2015 | Simultaneous dissection and comparison of IL-2 and IL-15 signaling pathways by global quantitative phosphoproteomics | Proteomics | 25142963 |
| 2014 | Phosphorylation site dynamics of early T-cell receptor signaling | PloS One | 25147952 |
| 2014 | Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling | Cell Reports | 25159151 |
| 2015 | Quantitative phosphoproteomics reveals the protein tyrosine kinase Pyk2 as a central effector of olfactory receptor signaling in prostate cancer cells | Biochimica Et Biophysica Acta | 25219547 |
| 2014 | Quantitative phosphoproteomic analysis of signaling downstream of the prostaglandin e2/g-protein coupled receptor in human synovial fibroblasts: potential antifibrotic networks | Journal of Proteome Research | 25223752 |
| 2015 | Quantitative analysis of a phosphoproteome readily altered by the protein kinase CK2 inhibitor quinalizarin in HEK-293T cells | Biochimica Et Biophysica Acta | 25278378 |
| 2015 | Refined phosphopeptide enrichment by phosphate additive and the analysis of human brain phosphoproteome | Proteomics | 25307156 |
| 2014 | Phosphoproteomics reveals resveratrol-dependent inhibition of Akt/mTORC1/S6K1 signaling | Journal of Proteome Research | 25311616 |
| 2015 | Quantitative phosphoproteomics of Alzheimer's disease reveals cross-talk between kinases and small heat shock proteins | Proteomics | 25332170 |
| 2014 | Identification of the PLK2-dependent phosphopeptidome by quantitative proteomics [corrected] | PloS One | 25338102 |
| 2015 | Integration of conventional quantitative and phospho-proteomics reveals new elements in activated Jurkat T-cell receptor pathway maintenance | Proteomics | 25348772 |
| 2016 | Ultra-deep tyrosine phosphoproteomics enabled by a phosphotyrosine superbinder | Nature Chemical Biology | 27642862 |

| | | | |
|---|---|---|---|
| 2017 | Protein-phosphotyrosine proteome profiling by superbinder-SH2 domain affinity purification mass spectrometry, sSH2-AP-MS | Proteomics | 27880036 |
| 2017 | Quantitative Tyrosine Phosphoproteomics of Epidermal Growth Factor Receptor (EGFR) Tyrosine Kinase Inhibitor-treated Lung Adenocarcinoma Cells Reveals Potential Novel Biomarkers of Therapeutic Response | Molecular & cellular proteomics: MCP | 28331001 |
| 2019 | 15 years of PhosphoSitePlus®: integrating post-translationally modified sites, disease variants and isoforms | Nucleic Acids Research | 30445427 |

Table 1. Published human phosphoproteomic datasets.

The very stringent mouse set contained 9703, 1365 and 464 phosphorylated serines, threonines, tyrosines, respectively, that had been identified in any low-throughput study (availalble in Phosphosite+) and/or identified in at least three of the 42 high-throughput published datasets that we filtered. The mouse phosphoproteomic datasets that we used are shown in Table 2 (Azimifar *et al.*, 2012; Ballif *et al.*, 2008; Batth *et al.*, 2014; Caruthers *et al.*, 2014; Chang *et al.*, 2013; Chevrier *et al.*, 2011; Choudhary *et al.*, 2009; Corradini *et al.*, 2014; Dong *et al.*, 2012; Doubleday and Ballif, 2014; Edwards *et al.*, 2014; Ferrando *et al.*, 2012; Ficarro *et al.*, 2011b; Gnad *et al.*, 2010; Goswami *et al.*, 2012; Griaud *et al.*, 2012; Gündisch *et al.*, 2013; Han *et al.*, 2015; Herring *et al.*, 2015; Hornbeck *et al.*, 2019; Iwai *et al.*, 2010; Jedrychowski *et al.*, 2011; Lu *et al.*, 2013; Matic *et al.*, 2014; Nakayasu *et al.*, 2013; Ostasiewicz *et al.*, 2010; Palmisano *et al.*, 2012b; Pan *et al.*, 2013; Pines *et al.*, 2011; Pinto *et al.*, 2015; Scholten *et al.*, 2013; Sharma *et al.*, 2010; Stokes *et al.*, 2012b; Sweet *et al.*, 2009; Weintz *et al.*, 2010; Wilson-Grady *et al.*, 2013; Wu *et al.*, 2012, 2009; Xu *et al.*, 2008; Yu *et al.*, 2011; Zhong *et al.*, 2014, 2012, 2015).

| Publication Year | Title | Publication Title | PMID |
|---|---|---|---|
| 2008 | Large-scale identification and evolution indexing of tyrosine phosphorylation sites | Journal of Proteome Research | 18034455 |

| | | | |
|---|---|---|---|
| | from murine brain | | |
| 2008 | Capture of phosphopeptides using alpha-zirconium phosphate nanoplatelets | Analytical Chemistry | 18522436 |
| 2009 | Large scale localization of protein phosphorylation by use of electron capture dissociation mass spectrometry | Molecular & cellular proteomics: MCP | 19131326 |
| 2009 | Concurrent quantification of proteome and phosphoproteome to reveal system-wide association of protein phosphorylation and gene expression | Molecular & cellular proteomics: MCP | 19674963 |
| 2009 | Mislocalized activation of oncogenic RTKs switches downstream signaling outcomes | Molecular Cell | 19854140 |
| 2010 | Quantitative analysis of kinase-proximal signaling in lipopolysaccharide-induced innate immune response | Journal of Proteome Research | 20222745 |
| 2010 | Quantitative phosphoproteomic analysis of T cell receptor signaling in diabetes prone and resistant mice | Journal of Proteome Research | 20438120 |
| 2010 | Proteome, phosphoproteome, and N-glycoproteome are quantitatively preserved in formalin-fixed paraffin-embedded tissue and analyzable by high-resolution mass spectrometry | Journal of Proteome Research | 20469934 |
| 2010 | The phosphoproteome of toll-like receptor-activated macrophages | Molecular Systems Biology | 20531401 |
| 2010 | Evolutionary constraints of phosphorylation in eukaryotes, prokaryotes, and mitochondria | Molecular & cellular proteomics: MCP | 20688971 |
| 2011 | Phosphoproteomic analysis identifies Grb10 as an mTORC1 substrate that negatively regulates insulin signaling | Science (New York, N.Y.) | 21659605 |
| 2011 | Online nanoflow multidimensional fractionation for high efficiency phosphopeptide analysis | Molecular & cellular proteomics: MCP | 21788404 |
| 2011 | Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics | Molecular & cellular proteomics: MCP | 21917720 |
| 2011 | Global phosphoproteome profiling reveals unanticipated networks responsive to cisplatin treatment of embryonic stem cells | Molecular and Cellular Biology | 22006019 |
| 2011 | Systematic discovery of TLR signaling components delineates viral-sensing circuits | Cell | 22078882 |
| 2012 | PTMScan direct: identification and quantification of peptides from critical signaling proteins by immunoaffinity enrichment coupled with LC-MS/MS | Molecular & cellular proteomics: MCP | 22322096 |
| 2012 | TSLP signaling network revealed by SILAC-based phosphoproteomics | Molecular & cellular proteomics: MCP | 22345495 |
| 2012 | Induction of membrane circular dorsal ruffles requires co-signalling of integrin-ILK-complex and EGF receptor | Journal of Cell Science | 22357970 |

| | | | |
|---|---|---|---|
| 2012 | Identification of targets of c-Src tyrosine kinase by chemical complementation and phosphoproteomics | Molecular & cellular proteomics: MCP | 22499769 |
| 2012 | BCR/ABL modulates protein phosphorylation associated with the etoposide-induced DNA damage response | Journal of Proteomics | 22705319 |
| 2012 | Comparative phosphoproteomic analysis of neonatal and adult murine brain | Proteomics | 22807455 |
| 2012 | A novel method for the simultaneous enrichment, identification, and quantification of phosphopeptides and sialylated glycopeptides applied to a temporal profile of mouse brain development | Molecular & cellular proteomics: MCP | 22843994 |
| 2012 | Depletion of acidic phosphopeptides by SAX to improve the coverage for the detection of basophilic kinase substrates | Journal of Proteome Research | 22871156 |
| 2012 | Investigation of receptor interacting protein (RIP3)-dependent protein phosphorylation by quantitative phosphoproteomics | Molecular & cellular proteomics: MCP | 22942356 |
| 2013 | Global protein phosphorylation dynamics during deoxynivalenol-induced ribotoxic stress response in the macrophage | Toxicology and Applied Pharmacology | 23352502 |
| 2013 | Quantitative comparison of the fasted and re-fed mouse liver phosphoproteomes using lower pH reductive dimethylation | Methods (San Diego, Calif.) | 23567750 |
| 2013 | Facile synthesis of Fe3O4@mesoporous TiO2 microspheres for selective enrichment of phosphopeptides for phosphoproteomics analysis | Talanta | 23597982 |
| 2013 | Quantitative phosphoproteomic study of pressure-overloaded mouse heart reveals dynamin-related protein 1 as a modulator of cardiac hypertrophy | Molecular & cellular proteomics: MCP | 23882026 |
| 2013 | Phosphoproteomics study based on in vivo inhibition reveals sites of calmodulin-dependent protein kinase II regulation in the heart | Journal of the American Heart Association | 23926118 |
| 2013 | Comparative phosphoproteomics reveals components of host cell invasion and post-transcriptional regulation during Francisella infection | Molecular & cellular proteomics: MCP | 23970565 |
| 2013 | Delayed times to tissue fixation result in unpredictable global phosphoproteome changes | Journal of Proteome Research | 23984901 |
| 2014 | Mercury alters B-cell protein phosphorylation profiles | Journal of Proteome Research | 24224561 |
| 2014 | Quantitative phosphoproteomic analysis of RIP3-dependent protein phosphorylation in the course of TNF-induced necroptosis | Proteomics | 24453211 |
| 2014 | Neuronal process structure and growth proteins are targets of heavy PTM regulation during brain development | Journal of Proteomics | 24560892 |
| 2014 | Alterations in the cerebellar | Molecular & cellular | 24925903 |

| | | | |
|---|---|---|---|
| | (Phospho)proteome of a cyclic guanosine monophosphate (cGMP)-dependent protein kinase knockout mouse | proteomics: MCP | |
| 2015 | Integrated approach using multistep enzyme digestion, TiO2 enrichment, and database search for in-depth phosphoproteomic profiling | Proteomics | 25159016 |
| 2014 | Quantitative phosphoproteomics of murine Fmr1-KO cell lines provides new insights into FMRP-dependent signal transduction mechanisms | Journal of Proteome Research | 25168779 |
| 2014 | Developmentally-Dynamic Murine Brain Proteomes and Phosphoproteomes Revealed by Quantitative Proteomics | Proteomes | 25177544 |
| 2015 | Quantitative phosphoproteomics reveals crosstalk between phosphorylation and O-GlcNAc in the DNA damage response pathway | Proteomics | 25263469 |
| 2014 | Off-line high-pH reversed-phase fractionation for in-depth phosphoproteomics | Journal of Proteome Research | 25338131 |
| 2015 | Quantitative phosphoproteomic analysis of IL-33-mediated signaling | Proteomics | 25367039 |
| 2015 | Development of a tandem affinity phosphoproteomic method with motif selectivity and its application in analysis of signal transduction networks | Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences | 25777480 |
| 2019 | 15 years of PhosphoSitePlus®: integrating post-translationally modified sites, disease variants and isoforms | Nucleic Acids Research | 30445427 |

Table 2. Published mouse phosphoproteomic datasets.

The very stringent yeast set contained 4767, 947 and 29 phosphorylated serines, threonines, tyrosines, respectively, that had been identified in any low-throughput study (availalble in PhosphoGrid2) and/or identified in at least three of the 21 high-throughput published datasets that we filtered. The yeast phosphoproteomic datasets that we used are shown in Table 3 (Aguiar *et al.*, 2010; Albuquerque *et al.*, 2008; Beltrao *et al.*, 2009; Bodenmiller *et al.*, 2008, 2010; Chi *et al.*, 2007; Gnad *et al.*, 2009; Gruhler *et al.*, 2005; Holt *et al.*, 2009; Huber *et al.*, 2009; Lee *et al.*, 2013; Li *et al.*, 2007; Mascaraque *et al.*, 2013; Oliveira *et al.*, 2012; Sadowski *et al.*, 2013; Saleem *et al.*, 2010; Soufi *et al.*, 2009; Studer *et al.*, 2016;

Weinert *et al.*, 2014; Wu *et al.*, 2011). Of note, the publication of (Holt *et al.*, 2009) provided 3 different datasets.

| Publication Year | Title | Publication Title | PMID |
|---|---|---|---|
| 2005 | Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway | Molecular & cellular proteomics: MCP | 15665377 |
| 2007 | Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry | Proceedings of the National Academy of Sciences of the United States of America | 17287358 |
| 2007 | Large-scale phosphorylation analysis of alpha-factor-arrested Saccharomyces cerevisiae | Journal of Proteome Research | 17330950 |
| 2008 | A multidimensional chromatography technology for in-depth phosphoproteome analysis | Molecular & cellular proteomics: MCP | 18407956 |
| 2008 | PhosphoPep--a database of protein phosphorylation sites in model organisms | Nature Biotechnology | 19060867 |
| 2009 | Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species | PLoS biology | 19547744 |
| 2009 | Characterization of the rapamycin-sensitive phosphoproteome reveals that Sch9 is a central coordinator of protein synthesis | Genes & Development | 19684113 |
| 2009 | Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution | Science (New York, N.Y.) | 19779198 |
| 2009 | High-accuracy identification and bioinformatic analysis of in vivo protein phosphorylation sites in yeast | Proteomics | 19795423 |
| 2009 | Global analysis of the yeast osmotic stress response by quantitative proteomics | Molecular bioSystems | 19823750 |
| 2010 | Gas-phase rearrangements do not affect site localization reliability in phosphoproteomics data sets | Journal of Proteome Research | 20377248 |
| 2010 | Integrated phosphoproteomics analysis of a signaling network governing nutrient response and peroxisome induction | Molecular & cellular proteomics: MCP | 20395639 |
| 2010 | Phosphoproteomic analysis reveals interconnected system-wide responses to perturbations of kinases and phosphatases in yeast | Science Signaling | 21177495 |
| 2011 | Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes | Molecular & cellular proteomics: MCP | 21551504 |
| 2012 | Regulation of yeast central metabolism by enzyme phosphorylation | Molecular Systems Biology | 23149688 |

| | | | |
|---|---|---|---|
| 2013 | Phosphoproteomic analysis of protein kinase C signaling in Saccharomyces cerevisiae reveals Slt2 mitogen-activated protein kinase (MAPK)-dependent phosphorylation of eisosome core components | Molecular & cellular proteomics: MCP | 23221999 |
| 2013 | The PhosphoGRID Saccharomyces cerevisiae protein phosphorylation site database: version 2.0 update | Database: The Journal of Biological Databases and Curation | 23674503 |
| 2013 | MAPK Hog1 closes the S. cerevisiae glycerol channel Fps1 by phosphorylating and displacing its positive regulators | Genes & Development | 24298058 |
| 2014 | Acetylation dynamics and stoichiometry in Saccharomyces cerevisiae | Molecular Systems Biology | 24489116 |
| 2016 | Evolution of protein phosphorylation across 18 fungal species | Science (New York, N.Y.) | 27738172 |

Table 3. Published yeast phosphoproteomic datasets.

A very stringent *Arabidopsis* set contained 3294, 333 and 20 phosphorylated serines, threonines, tyrosines, respectively, that had been identified in at least three of the 47 high-throughput published datasets that we filtered. Of note, many of these *Arabidopsis* datasets were identified by initial investigation of the PhosphAT database (Durek *et al.*, 2010). The *Arabidopsis* phosphoproteomic datasets that we used are shown in Table 4 (Benschop *et al.*, 2007; Bhaskara *et al.*, 2017; Bigeard *et al.*, 2014; Carroll *et al.*, 2008; Chang *et al.*, 2012; Chen and Hoehenwarter, 2015; Chen *et al.*, 2010; Cho *et al.*, 2016; Choudhary *et al.*, 2015; de la Fuente van Bentem *et al.*, 2006; E Stecker *et al.*, 2014; Engelsberger and Schulze, 2012; Hoehenwarter *et al.*, 2013; Howden *et al.*, 2011; Hsu *et al.*, 2009; Ingelsson and Vener, 2012; Ito *et al.*, 2009; Jones *et al.*, 2009; Lan *et al.*, 2012; Lassowskat *et al.*, 2014, 2013; Li *et al.*, 2014; Lin *et al.*, 2015; Mattei *et al.*, 2016; Mayank *et al.*, 2012; Menz *et al.*, 2016; Mithoe *et al.*, 2012; Nakagami *et al.*, 2010; Niittylä *et al.*, 2007; Nühse *et al.*, 2007; Nukarinen *et al.*, 2017; Qing *et al.*, 2016; Rayapuram *et al.*, 2014; Reiland *et al.*, 2009; Roitinger *et al.*, 2015; Sugiyama *et al.*, 2008; Umezawa *et al.*, 2013; Vandenbogaert *et al.*, 2012; Vu *et al.*, 2016; P. Wang *et al.*, 2013; X. Wang *et al.*, 2013; Wang *et al.*, 2009; Whiteman *et al.*, 2008; Wu *et al.*, 2013; Xue *et al.*, 2013; Z. Yang *et al.*, 2013; Hongtao Zhang *et al.*, 2013).

| Publication Year | Title | Publication Title | PMID |
|---|---|---|---|
| 2006 | Phosphoproteomics reveals extensive in vivo phosphorylation of Arabidopsis proteins involved in RNA metabolism | Nucleic Acids Research | 16807317 |
| 2007 | Quantitative phosphoproteomics of early elicitor signaling in Arabidopsis | Molecular & cellular proteomics: MCP | 17317660 |
| 2007 | Temporal analysis of sucrose-induced phosphorylation changes in plasma membrane proteins of Arabidopsis | Molecular & cellular proteomics: MCP | 17586839 |
| 2007 | Quantitative phosphoproteomic analysis of plasma membrane proteins reveals regulatory mechanisms of plant innate immune responses | The Plant Journal: For Cell and Molecular Biology | 17651370 |
| 2008 | Analysis of the Arabidopsis cytosolic ribosome proteome provides detailed insights into its components and their post-translational modification | Molecular & cellular proteomics: MCP | 17934214 |
| 2008 | Large-scale phosphorylation mapping reveals the extent of tyrosine phosphorylation in Arabidopsis | Molecular Systems Biology | 18463617 |
| 2008 | Identification of novel proteins and phosphorylation sites in a tonoplast enriched membrane fraction of Arabidopsis thaliana | Proteomics | 18686298 |
| 2009 | Phosphoproteomic analysis of nuclei-enriched fractions from Arabidopsis thaliana | Journal of Proteomics | 19245862 |
| 2009 | A simple and effective method for detecting phosphopeptides for phosphoproteomic analysis | Journal of Proteomics | 19341826 |
| 2009 | Large-scale Arabidopsis phosphoproteome profiling reveals novel chloroplast kinase substrates and phosphorylation networks | Plant Physiology | 19376835 |
| 2009 | A survey of the Arabidopsis thaliana mitochondrial phosphoproteome | Proteomics | 19688752 |
| 2009 | Functional phosphoproteomic profiling of phosphorylation sites in membrane fractions of salt-stressed Arabidopsis thaliana | Proteome Science | 19900291 |
| 2010 | Comparative analysis of phytohormone-responsive phosphoproteins in Arabidopsis thaliana using TiO2-phosphopeptide enrichment and mass accuracy precursor alignment | The Plant Journal: For Cell and Molecular Biology | 20374526 |
| 2010 | Large-scale comparative phosphoproteomics identifies conserved phosphorylation sites in plants | Plant Physiology | 20466843 |
| 2011 | The phosphoproteome of Arabidopsis plants lacking the oxidative signal-inducible1 (OXI1) protein kinase | The New Phytologist | 21175636 |
| 2012 | Nitrate and ammonium lead to distinct global dynamic phosphorylation patterns when | The Plant Journal: For Cell and | 22060019 |

| | | | |
|---|---|---|---|
| | resupplied to nitrogen-starved Arabidopsis seedlings | Molecular Biology | |
| 2012 | Targeted quantitative phosphoproteomics approach for the detection of phospho-tyrosine signaling in plants | Journal of Proteome Research | 22074104 |
| 2012 | Comparative phosphoproteomic analysis of microsomal fractions of Arabidopsis thaliana and Oryza sativa subjected to high salinity | Plant Science: An International Journal of Experimental Plant Biology | 22325874 |
| 2012 | Quantitative phosphoproteome profiling of iron-deficient Arabidopsis roots | Plant Physiology | 22438062 |
| 2012 | Phosphoproteomics of Arabidopsis chloroplasts reveals involvement of the STN7 kinase in phosphorylation of nucleoid protein pTAC16 | FEBS letters | 22616989 |
| 2012 | Characterization of the phosphoproteome of mature Arabidopsis pollen | The Plant Journal: For Cell and Molecular Biology | 22631563 |
| 2012 | Automated phosphopeptide identification using multiple MS/MS fragmentation modes | Journal of Proteome Research | 23094866 |
| 2013 | A large-scale protein phosphorylation analysis reveals novel phosphorylation motifs and phosphoregulatory networks in Arabidopsis | Journal of Proteomics | 23111157 |
| 2013 | Identification of novel in vivo MAP kinase substrates in Arabidopsis thaliana through use of tandem metal oxide affinity chromatography | Molecular & cellular proteomics: MCP | 23172892 |
| 2013 | Quantitative phosphoproteomics after auxin-stimulated lateral root induction identifies an SNX1 protein phosphorylation site required for growth | Molecular & cellular proteomics: MCP | 23328941 |
| 2013 | Genetics and phosphoproteomics reveal a protein phosphorylation network in the abscisic acid signaling pathway in Arabidopsis thaliana | Science Signaling | 23572148 |
| 2013 | Quantitative measurement of phosphoproteome response to osmotic stress in arabidopsis based on Library-Assisted eXtracted Ion Chromatogram (LAXIC) | Molecular & cellular proteomics: MCP | 23660473 |
| 2013 | Quantitative phosphoproteomics identifies SnRK2 protein kinase substrates and reveals the effectors of abscisic acid action | Proceedings of the National Academy of Sciences of the United States of America | 23776212 |
| 2013 | Sucrose-induced receptor kinase SIRK1 regulates a plasma membrane aquaporin in Arabidopsis | Molecular & cellular proteomics: MCP | 23820729 |
| 2013 | Stable isotope metabolic labeling-based quantitative phosphoproteomic analysis of Arabidopsis mutants reveals ethylene-regulated time-dependent phosphoproteins and putative substrates of constitutive triple | Molecular & cellular proteomics: MCP | 24043427 |

| | | | |
|---|---|---|---|
| | response 1 kinase | | |
| 2014 | The protein phosphatase subunit PP2A-B'? is required to suppress day length-dependent pathogenesis responses triggered by intracellular oxidative stress | The New Phytologist | 24299221 |
| 2014 | Identification of novel PAMP-triggered phosphorylation and dephosphorylation events in Arabidopsis thaliana by quantitative phosphoproteomic analysis | Journal of Proteome Research | 24601666 |
| 2014 | Phosphoproteomic Analyses Reveal Early Signaling Events in the Osmotic Stress Response | Plant Physiology | 24808101 |
| 2014 | Proteomic and phosphoproteomic analyses of chromatin-associated proteins from Arabidopsis thaliana | Proteomics | 24889360 |
| 2014 | Sustained mitogen-activated protein kinase activation reprograms defense metabolism and phosphoprotein profile in Arabidopsis thaliana | Frontiers in Plant Science | 25368622 |
| 2015 | Quantitative phosphoproteomics of the ataxia telangiectasia-mutated (ATM) and ataxia telangiectasia-mutated and rad3-related (ATR) dependent DNA damage response in Arabidopsis thaliana | Molecular & cellular proteomics: MCP | 25561503 |
| 2015 | Quantitative Circadian Phosphoproteomic Analysis of Arabidopsis Reveals Extensive Clock Control of Key Components in Physiological, Metabolic, and Signaling Pathways | Molecular & cellular proteomics: MCP | 26091701 |
| 2015 | Integrating Phosphoproteomics and Bioinformatics to Study Brassinosteroid-Regulated Phosphorylation Dynamics in Arabidopsis | BMC genomics | 26187819 |
| 2015 | Changes in the Phosphoproteome and Metabolome Link Early Signaling Events to Rearrangement of Photosynthesis and Central Metabolism in Salinity and Oxidative Stress Response in Arabidopsis | Plant Physiology | 26471895 |
| 2016 | Quantitative and Functional Phosphoproteomic Analysis Reveals that Ethylene Regulates Water Transport via the C-Terminal Phosphorylation of Aquaporin PIP2;1 in Arabidopsis | Molecular Plant | 26476206 |
| 2016 | Quantitative phosphoproteomics of protein kinase SnRK1 regulated protein phosphorylation in Arabidopsis under submergence | Journal of Experimental Botany | 27029354 |
| 2016 | Early nitrogen-deprivation responses in Arabidopsis roots reveal distinct differences on transcriptome and (phospho-) proteome levels between nitrate and ammonium nutrition | The Plant Journal: For Cell and Molecular Biology | 27419465 |

| | | | |
|---|---|---|---|
| 2016 | Comprehensive Analysis of the Membrane Phosphoproteome Regulated by Oligogalacturonides in Arabidopsis thaliana | Frontiers in Plant Science | 27532006 |
| 2016 | Up-to-Date Workflow for Plant (Phospho)proteomics Identifies Differential Drought-Responsive Phosphorylation Events in Maize Leaves | Journal of Proteome Research | 27643528 |
| 2017 | Protein Phosphatase 2Cs and Microtubule-Associated Stress Protein 1 Control Microtubule Stability, Plant Growth, and Drought Response | The Plant Cell | 28011693 |
| 2013 | PAPE (Prefractionation-Assisted Phosphoprotein Enrichment): A Novel Approach for Phosphoproteomic Analysis of Green Tissues from Plants | Proteomes | 28250405 |
| 2017 | Protein sumoylation and phosphorylation intersect in Arabidopsis signaling | The Plant Journal: For Cell and Molecular Biology | 28419593 |

Table 4. Published *Arabidopsis thaliana* Phosphoproteomic datasets.

For each of these four species we also randomly included the same number of «negative» (i.e. non-phosphorylated) serines, threonines and tyrosines from proteins that were not phosphorylated. All positive and negative datasets from the four species were compiled together in a eukaryotic positive and a eukaryotic negative dataset. Again, each of these four species sets (positive and negative sites) was later randomly split into a training dataset (70%) and an evaluation dataset (30%).

**The eukaryotic methylproteomic dataset**

The eukaryotic methylproteomic compendium consisted of 4316 methylation sites (m-sites) (4003 arginines and 313 lysines) and was compiled by manually mining and filtering 14 published studies in human, mouse, yeast and *Toxoplasma gondii.*(see Table 5) (Bremang *et al.*, 2013; Cao *et al.*, 2010; Geoghegan *et al.*, 2015; Guo *et al.*, 2014; Hart-Smith *et al.*, 2016; Hornbeck *et al.*, 2019; Larsen *et al.*, 2016b; Olsen *et al.*, 2016; Onwuli *et al.*, 2017; Plank *et al.*, 2015; Sylvestersen *et al.*, 2014; Wu *et al.*, 2015; Yagoub *et al.*, 2015; Yakubu *et al.*, 2017).

We applied stringent filtering criteria, including 99% correct methylated peptide identification and 99% correct methylation site localization, whenever applicable. We included studies that used methyl-arginine/methyl-lysine antibody enrichment and/or heavy-methyl-SILAC in order to further filter out false-positives, as suggetsed by (Hart-Smith *et al.*, 2016). In addition, we retrieved methylation sites from low-throughput studies documented in Phosphosite+ (Hornbeck *et al.*, 2019). For human, we compiled a stringent set. It contained 270 and 1754 methylated lysines and arginines, respectively, that had been identified in any low-throughput study (availalble in Phosphosite+) and/or identified in at least two of the 9 high-throughput published studies that we filtered. The mouse dataset (1635 and 31 methylated arginines and lysines respectively) was created from a high-throughput experiment (Guo *et al.*, 2014) and by further adding mouse methylation sites from low-throughput studies in Phosphosite+. The yeast dataset (55 and 12 methylated arginines and lysines respectively) was assembled from three publications (Yagoub *et al.*, 2015; Plank *et al.*, 2015; Hart-Smith *et al.*, 2016) whereas the *Toxoplasma gondii* dataset came from one study on 559 methylated arginines (Yakubu *et al.*, 2017). For each of these four species we also randomly included the same number of «negative» (i.e. non-methylated) lysines and arginines from proteins that were not methylated. Again, each of these four species sets (positive and negative sites) was later randomly split into a training dataset (70%) and an evaluation dataset (30%).

| Publication Year | Title | Publication Title | PMID | Species |
|---|---|---|---|---|
| 2010 | High-coverage proteome analysis reveals the first insight of protein modification systems in the pathogenic spirochete Leptospira interrogans | Cell Research | 19918266 | Human |
| 2013 | Mass spectrometry-based identification and characterisation of lysine and arginine methylation in the human proteome | Molecular bioSystems | 23748837 | Human |
| 2014 | Immunoaffinity enrichment and mass spectrometry analysis of | Molecular & cellular | 24129315 | Human, Mouse |

| | | | | |
|---|---|---|---|---|
| | protein methylation | proteomics: MCP | | |
| 2014 | Proteomic analysis of arginine methylation sites in human cells reveals dynamic regulation during transcriptional arrest | Molecular & cellular proteomics: MCP | 24563534 | Human |
| 2015 | A chemical proteomics approach for global analysis of lysine monomethylome profiling | Molecular & cellular proteomics: MCP | 25505155 | Human |
| 2015 | Comprehensive identification of arginine methylation in primary T cells reveals regulatory roles in cell signalling | Nature Communicatio ns | 25849564 | Human |
| 2015 | Expanding the yeast protein arginine methylome | Proteomics | 26046779 | Yeast |
| 2015 | Yeast proteins Gar1p, Nop1p, Npl3p, Nsr1p, and Rps2p are natively methylated and are substrates of the arginine methyltransferase Hmt1p | Proteomics | 26081071 | Yeast |
| 2016 | Large Scale Mass Spectrometry-based Identifications of Enzyme-mediated Protein Methylation Are Subject to High False Discovery Rates | Molecular & cellular proteomics: MCP | 26699799 | Yeast |
| 2016 | Quantitative Profiling of the Activity of Protein Lysine Methyltransferase SMYD2 Using SILAC-Based Proteomics | Molecular & cellular proteomics: MCP | 26750096 | Human |
| 2016 | Proteome-wide analysis of arginine monomethylation reveals widespread occurrence in human cells | Science Signaling | 27577262 | Human |
| 2017 | Mapping arginine methylation in the human body and cardiac disease | Proteomics. Clinical Applications | 27600370 | Human |
| 2017 | Comparative Monomethylarginine Proteomics Suggests that Protein Arginine Methyltransferase 1 (PRMT1) is a Significant Contributor to Arginine Monomethylation in Toxoplasma gondii | Molecular & cellular proteomics: MCP | 28143887 | Toxoplasma |
| 2019 | 15 years of PhosphoSitePlus®: integrating post-translationally modified sites, disease variants and isoforms | Nucleic Acids Research | 30445427 | Phosphosite _plus |

Table 5. Published eukaryotic methylproteomic datasets.

**Training of Neural Networks with Keras/Tensorflow and development of the webserver**

Keras/Tensorflow was used to build and train a methylation and a phosphorylation Neural Network (NN) using, as parameters, one hidden layer of 13 & 200 nodes, batch size of 160 & 400, 30 & 50 epochs of training, a dropout value of 0.6 & 0.8 respectively, (loss: Binary crossentropy; optimizer: adam; activation: relu; all other parameters default). Each network used, as input, sequence motifs of 29 amino acids - the methylated residue and 14 amino acids either side of it (we tried different motif lengths ranging from 11 – 35 amino acids). These values were selected after testing more than 250 different combinations of the above parameters. The motifs were converted to one-hot encoding. In addition, sequence motif redundancy was removed at the early stages of compiling the compendia, by using a 100% aminoacid identity cutoff.

A webserver, named Meth-Phos-Prometheus was developed based on the Jhipster Application Framework, which utilises the Java language and Spring Framework for the back-end and Angular Javascript Framework for the front-end. For all evaluations using Meth-Phos-Prometheus, a threshold score of >= 0.5 was used to predict a site as methylated or phosphorylated.

## Results and Discussion

Our strategy was successful for both types of prediction, as can be seen in Table 6 and Table 7. Concerning protein methylation, Meth-Phos-Prometheus was very accurate (85.8% accuracy, AUC value of 0.925) and had a balanced sensitivity/specificity. It significantly outperformed three other recently published prediction algorithms, GPS-MSP (Deng *et al.*, 2017), MePred-RF (Wei *et al.*, 2017) and PRme-PRed (Kumar *et al.*, 2017) against the entire eukaryotic evaluation dataset. Of note, the *T. gondii* dataset consisted only of arginines. The most recently published algorithm, PrmePRed (Kumar *et al.*, 2017), was assessed by its developers against a number of other algorithms and displayed an enhanced performance compared to them. In addition, it can only predict methylated arginines. We found that Meth-Phos-Prometheus displayed a significantly better performance (accuracy) of more than 11%, compared to the three other published algorithms, most probably due to the stringent filtering of noisy methylation sites that we applied, despite the rather limited number of training data.

| EUKARYOTIC METHYLATION EVALUATION DATASET | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | TP | FP | FN | TN | Sensitivity | Specificity | Accuracy | MCC |
| Lysine | | | | | | | | |
| PRme-PRed | X | X | X | X | X | X | X | X |
| MePred-RF | 48 | 29 | 49 | 68 | 49.5% | 70.1% | 59.8% | 0.20 |
| GPS-MSP | 31 | 1 | 66 | 96 | 32.0% | 99.0% | 65.5% | 0.42 |
| MethPhos-Prometheus | 70 | 31 | 27 | 66 | 72.2% | 68.0% | 70.1% | 0.40 |
| Arginine | | | | | | | | |
| PRme-PRed | 1124 | 603 | 67 | 588 | 94.4% | 49.4% | 71.9% | 0.49 |
| MePred-RF | 681 | 62 | 510 | 1129 | 57.2% | 94.8% | 76.0% | 0.56 |
| GPS-MSP | 215 | 13 | 976 | 1178 | 18.1% | 98.9% | 58.5% | 0.29 |
| MethPhos-Prometheus | 1026 | 143 | 165 | 1048 | 86.1% | 88.0% | 87.1% | 0.74 |
| Lysine and Arginine | | | | | | | | |
| PRme-PRed | X | X | X | X | X | X | X | X |
| MePred-RF | 729 | 91 | 559 | 1197 | 56.6% | 92.9% | 74.8% | 0.53 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| GPS-MSP | 246 | 14 | 1042 | 1274 | 19.1% | 98.9% | 59.0% | 0.30 |
| MethPhos-Prometheus | 1096 | 174 | 192 | 1114 | 85.1% | 86.5% | 85.8% | 0.72 |

Table 6. Performance of the four predictors against the entire eukaryotic methylation evaluation dataset. TP: True Positive; FP: False Positive; FN: False Negative; TN: True Negative; MCC: Matthews Correlation Coefficient.

Concerning protein phosphorylation, Meth-Phos-Prometheus was again very accurate (84% accuracy, AUC value of 0.916) and had a balanced sensitivity/specificity. It significantly outperformed three of the four evaluated prediction algorithms, Netphos 3 (Ingrell *et al.*, 2007), GPS3 (Xue *et al.*, 2008) and Phosfer (Trost and Kusalik, 2013), against the entire eukaryotic evaluation dataset. The most recently published algorithm, DeepPhos (Luo *et al.*, 2019), based on convolutional neural networks was assessed by its developers against four other well-established or recently published algorithms (i.e. MusiteDeep) (Wang *et al.*, 2017) and displayed an enhanced performance compared to them. For a prediction score of 0.5, Meth-Phos-Prometheus also outperformed DeepPhos, but the AUC value of DeepPhos was 0.926 (AUC: 0.916 for Meth-Phos-Prometheus). Therefore, we consider the performance of these two algorithms as comparable.

| EUKARYOTIC PHOSPHORYLATION EVALUATION DATASET | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Algorithm | TP | FP | FN | TN | Sensitivity | Specificity | Accuracy | MCC |
| Serine | | | | | | | | |
| Netphos | 11624 | 8881 | 878 | 3621 | 93.0% | 29.0% | 61.0% | 0.29 |
| GPS3H | 12291 | 11903 | 211 | 599 | 98.3% | 4.8% | 51.6% | 0.09 |
| Phosfer | 10334 | 3382 | 2168 | 9120 | 82.7% | 72.9% | 77.8% | 0.56 |
| DeepPhos | 11955 | 4257 | 444 | 8109 | 96.4% | 65.6% | 81.0% | 0.65 |
| Meth-Phos Prometheus | 10833 | 2070 | 1669 | 10432 | 86.7% | 83.4% | 85.0% | 0.70 |
| Threonine | | | | | | | | |
| Netphos | 1776 | 1198 | 350 | 928 | 83.5% | 43.7% | 63.6% | 0.30 |
| GPS3H | 2102 | 2007 | 24 | 119 | 98.9% | 5.6% | 52.2% | 0.12 |
| Phosfer | 1776 | 690 | 350 | 1436 | 83.5% | 67.5% | 75.5% | 0.52 |
| DeepPhos | 1888 | 405 | 222 | 1689 | 89.5% | 80.7% | 85.1% | 0.70 |
| Meth-Phos Prometheus | 1800 | 352 | 326 | 1774 | 84.7% | 83.4% | 84.1% | 0.68 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Tyrosine | | | | | | | | |
| Netphos | 884 | 488 | 490 | 886 | 64.3% | 64.5% | 64.4% | 0.29 |
| GPS3H | 1332 | 1138 | 42 | 236 | 96.9% | 17.2% | 57.1% | 0.23 |
| Phosfer | 967 | 449 | 407 | 925 | 70.4% | 67.3% | 68.9% | 0.38 |
| DeepPhos | 1234 | 521 | 135 | 838 | 90.1% | 61.7% | 76.0% | 0.54 |
| Meth-Phos Prometheus | 1088 | 363 | 286 | 1011 | 79.2% | 73.6% | 76.4% | 0.53 |
| Serine/Threonine/Tyrosine | | | | | | | | |
| Netphos | 14284 | 10567 | 1718 | 5435 | 89.3% | 34.0% | 61.6% | 0.28 |
| GPS3H | 15725 | 15048 | 277 | 954 | 98.3% | 6.0% | 52.1% | 0.11 |
| Phosfer | 13077 | 4521 | 2925 | 11481 | 81.7% | 71.7% | 76.7% | 0.54 |
| DeepPhos | 15077 | 5183 | 801 | 10636 | 95.0% | 67.2% | 81.1% | 0.65 |
| Meth-Phos Prometheus | 13721 | 2785 | 2281 | 13217 | 85.7% | 82.6% | 84.2% | 0.68 |

Table 7. Performance of the five predictors against the entire eukaryotic phosphorylation evaluation dataset. TP: True Positive; FP: False Positive; FN: False Negative; TN: True Negative; MCC: Matthews Correlation Coefficient.

The Meth-Phos-Prometheus webserver takes as input protein sequences in FASTA format and requires a prediction score threshold (default value 0.5). However, the user may change the prediction score threshold for phosphorylation and methylation prediction, thus adjusting the sensitivity and specificity of the two NNs. Figure 1 and Figure 2 show how sensitivity and specificity change for the two NNs, as the prediction score threshold is adjusted by the user. In addition, the user may select the maximum distance (1-5 aa) between two neighbouring predicted amino acids (one methylated, one phosphorylated) in order to mark them as a predicted meth/phos switch. The same distance threshold is used for identifying clusters, where three or more predicted neighbouring p-sites or m-sites are found. For the clusters, the distance threshold is between two consecutive sites. For usage, see also the help page in the website, as well as an embedded video. The prediction results are available for downloading in tab-delimited format (see help page of website for detailed explanation). Furthermore, the results of the 10 sequences with the most predicted methylation/phosphorylation sites are available in a graphical representation, as can be seen in Figure **3** and at the website help-page. The

blue circles correspond to methylation sites and the red circles correspond to phosphorylation sites. The user may hover the mouse over a circle to see the prediction score of that site. The entire yeast proteome may be analyzed in 50 min.

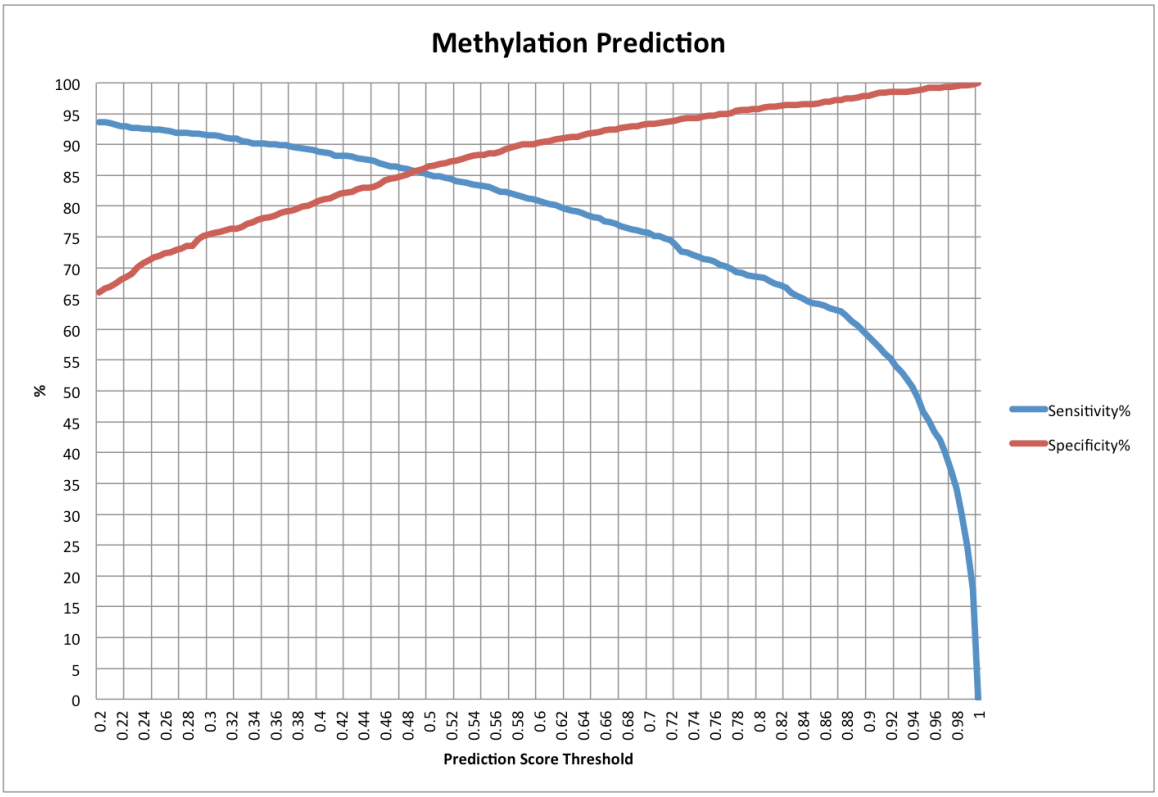



Figure 1. Sensitivity and Specificity of Meth-Phos Prometheus for methylation site prediction, with various prediction score thresholds.

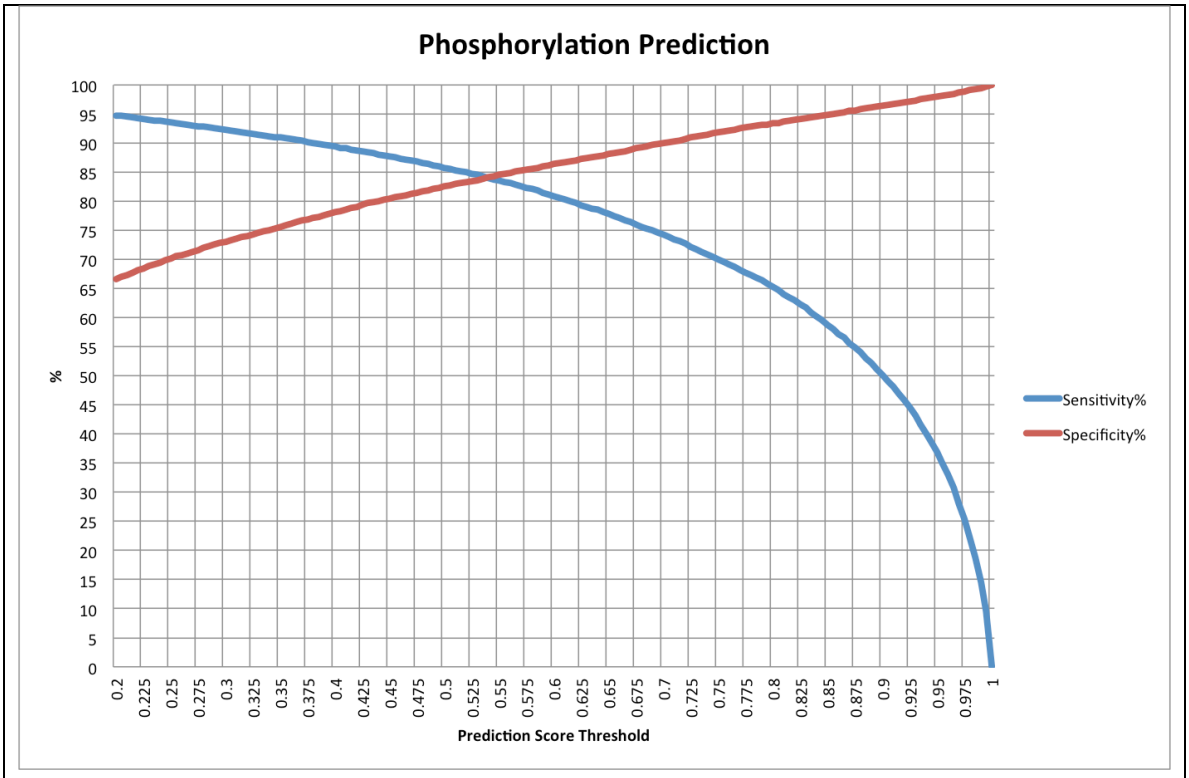
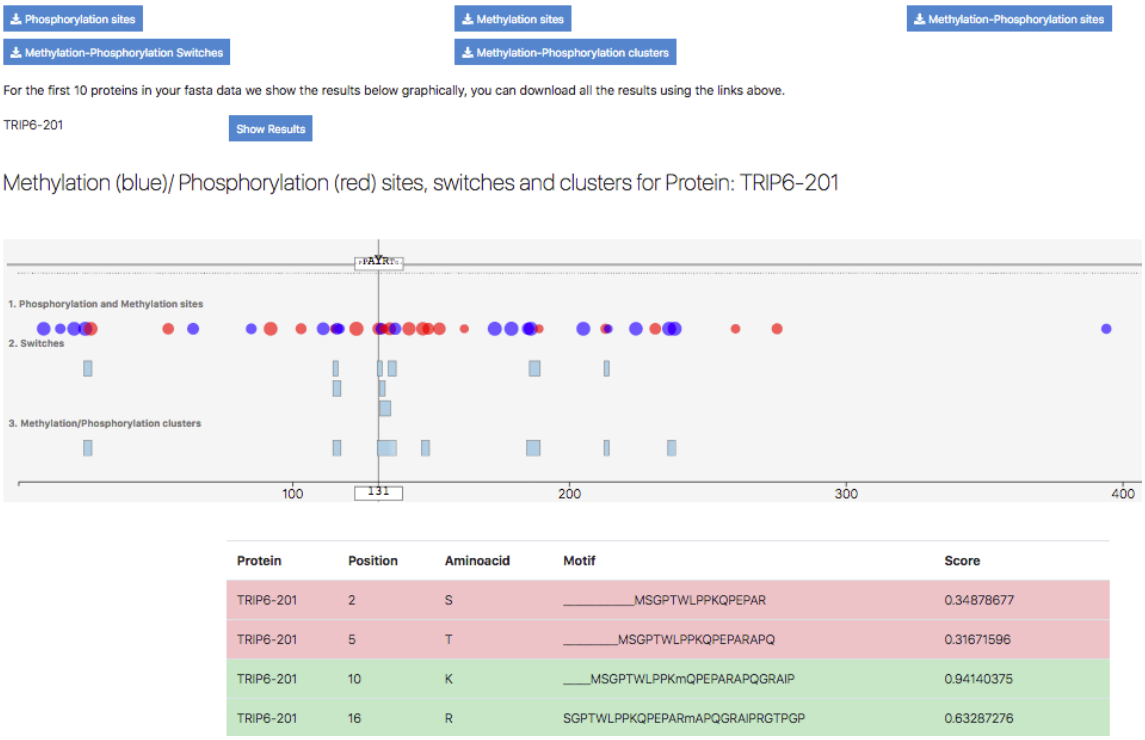



Figure 2. Sensitivity and Specificity of Meth-Phos Prometheus for phosphorylation site prediction, with various prediction score thresholds.




| Protein | Position | Aminoacid | Motif | Score |
|---|---|---|---|---|
| TRIP6-201 | 2 | S | _____MSGPTWLPPKQPEPAR | 0.34878677 |
| TRIP6-201 | 5 | T | _____MSGPTWLPPKQPEPARAPQ | 0.31671596 |
| TRIP6-201 | 10 | K | ____MSGPTWLPPKmQPEPARAPQGRAIP | 0.94140375 |
| TRIP6-201 | 16 | R | SGPTWLPPKQPEPARmAPQGRAIPRGTPGP | 0.63287276 |

Figure 3. Graphical view of predicted methylation and phosphorylation sites as well as the meth/phos switches and clusters/rheostats they form, in Meth-Phos-Prometheus webserver.

# Bibliography


Aguiar,M. *et al.* (2010) Gas-phase rearrangements do not affect site localization reliability in phosphoproteomics data sets. *J. Proteome Res.*, **9**, 3103–3107.

Alayev,A. *et al.* (2014) Phosphoproteomics reveals resveratrol-dependent inhibition of Akt/mTORC1/S6K1 signaling. *J. Proteome Res.*, **13**, 5734–5742.

Albuquerque,C.P. *et al.* (2008) A multidimensional chromatography technology for in-depth phosphoproteome analysis. *Mol. Cell Proteomics*, **7**, 1389–1396.

Amoutzias,G.D. *et al.* (2012) Evaluation and properties of the budding yeast phosphoproteome. *Mol. Cell Proteomics*, **11**, M111.009555.

Andrews,F.H. *et al.* (2016) Regulation of Methyllysine Readers through Phosphorylation. *ACS Chem. Biol.*, **11**, 547–553.

Aslanian,A. *et al.* (2014) Mass spectrometry-based quantification of the cellular response to methyl methanesulfonate treatment in human cells. *DNA Repair (Amst.)*, **15**, 29–38.

Azimifar,S.B. *et al.* (2012) Induction of membrane circular dorsal ruffles requires co-signalling of integrin-ILK-complex and EGF receptor. *J. Cell. Sci.*, **125**, 435–448.

Bai,Y. *et al.* (2012) Phosphoproteomics identifies driver tyrosine kinases in sarcoma cell lines and tumors. *Cancer Res.*, **72**, 2501–2511.

Ballif,B.A. *et al.* (2008) Large-scale identification and evolution indexing of tyrosine phosphorylation sites from murine brain. *J. Proteome Res.*, **7**, 311–318.

Batth,T.S. *et al.* (2014) Off-line high-pH reversed-phase fractionation for in-depth phosphoproteomics. *J. Proteome Res.*, **13**, 6176–6186.

Beck,F. *et al.* (2014) Time-resolved characterization of cAMP/PKA-dependent signaling reveals that platelet inhibition is a concerted process involving multiple signaling pathways. *Blood*, **123**, e1–e10.

Beltran,L. *et al.* (2012) Global profiling of protein kinase activities in cancer cells by mass spectrometry. *J Proteomics*, **77**, 492–503.

Beltrao,P. *et al.* (2009) Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. *PLoS Biol.*, **7**, e1000134.

Benschop,J.J. *et al.* (2007) Quantitative phosphoproteomics of early elicitor signaling in Arabidopsis. *Mol. Cell Proteomics*, **6**, 1198–1214.

Bhaskara,G.B. *et al.* (2017) Protein Phosphatase 2Cs and Microtubule-Associated Stress Protein 1 Control Microtubule Stability, Plant Growth, and Drought Response. *Plant Cell*, **29**, 169–191.

Bian,Y. *et al.* (2014) An enzyme assisted RP-RPLC approach for in-depth analysis of human liver phosphoproteome. *J Proteomics*, **96**, 253–262.

Bian,Y. *et al.* (2013) Global screening of CK2 kinase substrates by an integrated phosphoproteomics workflow. *Sci Rep*, **3**, 3460.

Bian,Y. *et al.* (2012) Improve the coverage for the analysis of phosphoproteome of HeLa cells by a tandem digestion approach. *J. Proteome Res.*, **11**, 2828–2837.

Bian,Y. *et al.* (2016) Ultra-deep tyrosine phosphoproteomics enabled by a phosphotyrosine superbinder. *Nat. Chem. Biol.*, **12**, 959–966.

Bigeard,J. *et al.* (2014) Proteomic and phosphoproteomic analyses of chromatin-associated proteins from Arabidopsis thaliana. *Proteomics*, **14**, 2141–2155.
van den Biggelaar,M. *et al.* (2014) Quantitative phosphoproteomics unveils temporal dynamics of thrombin signaling in human endothelial cells. *Blood*, **123**, e22-36.
Blanc,R.S. and Richard,S. (2017) Arginine Methylation: The Coming of Age. *Mol. Cell*, **65**, 8–24.
Blom,N. *et al.* (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **294**, 1351–1362.
Bodenmiller,B. *et al.* (2008) PhosphoPep--a database of protein phosphorylation sites in model organisms. *Nat. Biotechnol.*, **26**, 1339–1340.
Bodenmiller,B. *et al.* (2010) Phosphoproteomic analysis reveals interconnected system-wide responses to perturbations of kinases and phosphatases in yeast. *Sci Signal*, **3**, rs4.
Breitkopf,S.B. *et al.* (2010) Proteomics analysis of cellular imatinib targets and their candidate downstream effectors. *J. Proteome Res.*, **9**, 6033–6043.
Bremang,M. *et al.* (2013) Mass spectrometry-based identification and characterisation of lysine and arginine methylation in the human proteome. *Mol Biosyst*, **9**, 2231–2247.
Brill,L.M. *et al.* (2009) Phosphoproteomic analysis of human embryonic stem cells. *Cell Stem Cell*, **5**, 204–213.
Cantin,G.T. *et al.* (2006) Quantitative phosphoproteomic analysis of the tumor necrosis factor pathway. *J. Proteome Res.*, **5**, 127–134.
Cao,X.-J. *et al.* (2010) High-coverage proteome analysis reveals the first insight of protein modification systems in the pathogenic spirochete Leptospira interrogans. *Cell Res.*, **20**, 197–210.
Carroll,A.J. *et al.* (2008) Analysis of the Arabidopsis cytosolic ribosome proteome provides detailed insights into its components and their post-translational modification. *Mol. Cell Proteomics*, **7**, 347–369.
Caruthers,N.J. *et al.* (2014) Mercury alters B-cell protein phosphorylation profiles. *J. Proteome Res.*, **13**, 496–505.
Casado,P. *et al.* (2014) Environmental stress affects the activity of metabolic and growth factor signaling networks and induces autophagy markers in MCF7 breast cancer cells. *Mol. Cell Proteomics*, **13**, 836–848.
Casado,P. *et al.* (2013) Phosphoproteomics data classify hematological cancer cell lines according to tumor type and sensitivity to kinase inhibitors. *Genome Biol.*, **14**, R37.
Chang,I.-F. *et al.* (2012) Comparative phosphoproteomic analysis of microsomal fractions of Arabidopsis thaliana and Oryza sativa subjected to high salinity. *Plant Sci.*, **185–186**, 131–142.
Chang,Y.-W. *et al.* (2013) Quantitative phosphoproteomic study of pressure-overloaded mouse heart reveals dynamin-related protein 1 as a modulator of cardiac hypertrophy. *Mol. Cell Proteomics*, **12**, 3094–3107.
Chen,X. *et al.* (2016) A homology-based pipeline for global prediction of post-translational modification sites. *Sci Rep*, **6**, 25801.
Chen,X. *et al.* (2011) Increasing phosphoproteome coverage and identification of phosphorylation motifs through combination of different HPLC fractionation methods. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, **879**, 25–34.
Chen,Y. *et al.* (2010) Comparative analysis of phytohormone-responsive phosphoproteins in Arabidopsis thaliana using TiO2-phosphopeptide enrichment and mass accuracy precursor alignment. *Plant J.*, **63**, 1–17.

Chen,Y. and Hoehenwarter,W. (2015) Changes in the Phosphoproteome and Metabolome Link Early Signaling Events to Rearrangement of Photosynthesis and Central Metabolism in Salinity and Oxidative Stress Response in Arabidopsis. *Plant Physiol.*, **169**, 3021–3033.
Chevrier,N. *et al.* (2011) Systematic discovery of TLR signaling components delineates viral-sensing circuits. *Cell*, **147**, 853–867.
Chi,A. *et al.* (2007) Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 2193–2198.
Cho,H.-Y. *et al.* (2016) Quantitative phosphoproteomics of protein kinase SnRK1 regulated protein phosphorylation in Arabidopsis under submergence. *J. Exp. Bot.*, **67**, 2745–2760.
Choudhary,C. *et al.* (2009) Mislocalized activation of oncogenic RTKs switches downstream signaling outcomes. *Mol. Cell*, **36**, 326–339.
Choudhary,M.K. *et al.* (2015) Quantitative Circadian Phosphoproteomic Analysis of Arabidopsis Reveals Extensive Clock Control of Key Components in Physiological, Metabolic, and Signaling Pathways. *Mol. Cell Proteomics*, **14**, 2243–2260.
Christensen,G.L. *et al.* (2010) Quantitative phosphoproteomics dissection of seven-transmembrane receptor signaling using full and biased agonists. *Mol. Cell Proteomics*, **9**, 1540–1553.
Chylek,L.A. *et al.* (2014) Phosphorylation site dynamics of early T-cell receptor signaling. *PLoS ONE*, **9**, e104240.
Corradini,E. *et al.* (2014) Alterations in the cerebellar (Phospho)proteome of a cyclic guanosine monophosphate (cGMP)-dependent protein kinase knockout mouse. *Mol. Cell Proteomics*, **13**, 2004–2016.
Dammer,E.B. *et al.* (2015) Quantitative phosphoproteomics of Alzheimer's disease reveals cross-talk between kinases and small heat shock proteins. *Proteomics*, **15**, 508–519.
Daub,H. *et al.* (2008) Kinase-selective enrichment enables quantitative phosphoproteomics of the kinome across the cell cycle. *Mol. Cell*, **31**, 438–448.
Deng,W. *et al.* (2017) Computational prediction of methylation types of covalently modified lysine and arginine residues in proteins. *Brief. Bioinformatics*, **18**, 647–658.
Di Palma,S. *et al.* (2013) Finding the same needles in the haystack? A comparison of phosphotyrosine peptides enriched by immuno-affinity precipitation and metal-based affinity chromatography. *J Proteomics*, **91**, 331–337.
Dong,M. *et al.* (2012) Depletion of acidic phosphopeptides by SAX to improve the coverage for the detection of basophilic kinase substrates. *J. Proteome Res.*, **11**, 4673–4681.
Doubleday,P.F. and Ballif,B.A. (2014) Developmentally-Dynamic Murine Brain Proteomes and Phosphoproteomes Revealed by Quantitative Proteomics. *Proteomes*, **2**, 197–207.
Durek,P. *et al.* (2010) PhosPhAt: the Arabidopsis thaliana phosphorylation site database. An update. *Nucleic Acids Res.*, **38**, D828-834.
E Stecker,K. *et al.* (2014) Phosphoproteomic Analyses Reveal Early Signaling Events in the Osmotic Stress Response. *Plant Physiol.*, **165**, 1171–1187.
Edwards,A.V.G. *et al.* (2014) Neuronal process structure and growth proteins are targets of heavy PTM regulation during brain development. *J Proteomics*, **101**, 77–87.

Engelsberger,W.R. and Schulze,W.X. (2012) Nitrate and ammonium lead to distinct global dynamic phosphorylation patterns when resupplied to nitrogen-starved Arabidopsis seedlings. *Plant J.*, **69**, 978–995.
Estève,P.-O. *et al.* (2011) A methylation and phosphorylation switch between an adjacent lysine and serine determines human DNMT1 stability. *Nat. Struct. Mol. Biol.*, **18**, 42–48.
Everley,P.A. and Dillman,J.F. (2010) A large-scale quantitative proteomic approach to identifying sulfur mustard-induced protein phosphorylation cascades. *Chem. Res. Toxicol.*, **23**, 20–25.
Fang,L. *et al.* (2014) A methylation-phosphorylation switch determines Sox2 stability and function in ESC maintenance or differentiation. *Mol. Cell*, **55**, 537–551.
Ferrando,I.M. *et al.* (2012) Identification of targets of c-Src tyrosine kinase by chemical complementation and phosphoproteomics. *Mol. Cell Proteomics*, **11**, 355–369.
Ficarro,S.B. *et al.* (2011a) Online nanoflow multidimensional fractionation for high efficiency phosphopeptide analysis. *Mol. Cell Proteomics*, **10**, O111.011064.
Ficarro,S.B. *et al.* (2011b) Online nanoflow multidimensional fractionation for high efficiency phosphopeptide analysis. *Mol. Cell Proteomics*, **10**, O111.011064.
Fischle,W. *et al.* (2003) Binary switches and modification cassettes in histone biology and beyond. *Nature*, **425**, 475–479.
Fischle,W. *et al.* (2005) Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. *Nature*, **438**, 1116–1122.
Fleitz,A. *et al.* (2013) Enhanced detection of multiply phosphorylated peptides and identification of their sites of modification. *Anal. Chem.*, **85**, 8566–8576.
Francavilla,C. *et al.* (2013) Functional proteomics defines the molecular switch underlying FGF receptor trafficking and cellular outputs. *Mol. Cell*, **51**, 707–722.
Franchin,C. *et al.* (2014) Identification of the PLK2-dependent phosphopeptidome by quantitative proteomics [corrected]. *PLoS ONE*, **9**, e111018.
Franchin,C. *et al.* (2015) Quantitative analysis of a phosphoproteome readily altered by the protein kinase CK2 inhibitor quinalizarin in HEK-293T cells. *Biochim. Biophys. Acta*, **1854**, 609–623.
Franz-Wachtel,M. *et al.* (2012) Global detection of protein kinase D-dependent phosphorylation events in nocodazole-treated human cells. *Mol. Cell Proteomics*, **11**, 160–170.
de la Fuente van Bentem,S. *et al.* (2006) Phosphoproteomics reveals extensive in vivo phosphorylation of Arabidopsis proteins involved in RNA metabolism. *Nucleic Acids Res.*, **34**, 3267–3278.
Galan,J.A. *et al.* (2014) Phosphoproteomic analysis identifies the tumor suppressor PDCD4 as a RSK substrate negatively regulated by 14-3-3. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E2918-2927.
Ge,F. *et al.* (2010) Phosphoproteomic analysis of primary human multiple myeloma cells. *J Proteomics*, **73**, 1381–1390.
Geoghegan,V. *et al.* (2015) Comprehensive identification of arginine methylation in primary T cells reveals regulatory roles in cell signalling. *Nat Commun*, **6**, 6758.
Gerarduzzi,C. *et al.* (2014) Quantitative phosphoproteomic analysis of signaling downstream of the prostaglandin e2/g-protein coupled receptor in human

synovial fibroblasts: potential antifibrotic networks. *J. Proteome Res.*, **13**, 5262–5280.
Giansanti,P. *et al.* (2014) Evaluating the promiscuous nature of tyrosine kinase inhibitors assessed in A431 epidermoid carcinoma cells by both chemical- and phosphoproteomics. *ACS Chem. Biol.*, **9**, 1490–1498.
Giansanti,P. *et al.* (2013) Interrogating cAMP-dependent kinase signaling in Jurkat T cells via a protein kinase A targeted immune-precipitation phosphoproteomics approach. *Mol. Cell Proteomics*, **12**, 3350–3359.
Glowinski,F. *et al.* (2014) Analysis of T4SS-induced signaling by H. pylori using quantitative phosphoproteomics. *Front Microbiol*, **5**, 356.
Gnad,F. *et al.* (2010) Evolutionary constraints of phosphorylation in eukaryotes, prokaryotes, and mitochondria. *Mol. Cell Proteomics*, **9**, 2642–2653.
Gnad,F. *et al.* (2009) High-accuracy identification and bioinformatic analysis of in vivo protein phosphorylation sites in yeast. *Proteomics*, **9**, 4642–4652.
Goss,V.L. *et al.* (2006) A common phosphotyrosine signature for the Bcr-Abl kinase. *Blood*, **107**, 4888–4897.
Goswami,T. *et al.* (2012) Comparative phosphoproteomic analysis of neonatal and adult murine brain. *Proteomics*, **12**, 2185–2189.
Griaud,F. *et al.* (2012) BCR/ABL modulates protein phosphorylation associated with the etoposide-induced DNA damage response. *J Proteomics*, **77**, 14–26.
Grosstessner-Hain,K. *et al.* (2011) Quantitative phospho-proteomics to investigate the polo-like kinase 1-dependent phospho-proteome. *Mol. Cell Proteomics*, **10**, M111.008540.
Gruhler,A. *et al.* (2005) Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol. Cell Proteomics*, **4**, 310–327.
Gündisch,S. *et al.* (2013) Delayed times to tissue fixation result in unpredictable global phosphoproteome changes. *J. Proteome Res.*, **12**, 4424–4434.
Guo,A. *et al.* (2014) Immunoaffinity enrichment and mass spectrometry analysis of protein methylation. *Mol. Cell Proteomics*, **13**, 372–387.
Hammond,D.E. *et al.* (2010) Quantitative analysis of HGF and EGF-dependent phosphotyrosine signaling networks. *J. Proteome Res.*, **9**, 2734–2742.
Han,D. *et al.* (2015) Integrated approach using multistep enzyme digestion, TiO2 enrichment, and database search for in-depth phosphoproteomic profiling. *Proteomics*, **15**, 618–623.
Han,G. *et al.* (2008) Large-scale phosphoproteome analysis of human liver tissue by enrichment and fractionation of phosphopeptides with strong anion exchange chromatography. *Proteomics*, **8**, 1346–1361.
Harder,L.M. *et al.* (2014) Inducing autophagy: a comparative phosphoproteomic study of the cellular response to ammonia and rapamycin. *Autophagy*, **10**, 339–355.
Hart-Smith,G. *et al.* (2016) Large Scale Mass Spectrometry-based Identifications of Enzyme-mediated Protein Methylation Are Subject to High False Discovery Rates. *Mol. Cell Proteomics*, **15**, 989–1006.
Helm,D. *et al.* (2014) Ion mobility tandem mass spectrometry enhances performance of bottom-up proteomics. *Mol. Cell Proteomics*, **13**, 3709–3715.
Helou,Y.A. *et al.* (2013) ERK positive feedback regulates a widespread network of tyrosine phosphorylation sites across canonical T cell signaling and actin cytoskeletal proteins in Jurkat T cells. *PLoS ONE*, **8**, e69641.
Herring,L.E. *et al.* (2015) Development of a tandem affinity phosphoproteomic method with motif selectivity and its application in analysis of signal transduction networks. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, **988**, 166–174.

Herskowitz,J.H. *et al.* (2010) Phosphoproteomic analysis reveals site-specific changes in GFAP and NDRG2 phosphorylation in frontotemporal lobar degeneration. *J. Proteome Res.*, **9**, 6368–6379.
Hoehenwarter,W. *et al.* (2013) Identification of novel in vivo MAP kinase substrates in Arabidopsis thaliana through use of tandem metal oxide affinity chromatography. *Mol. Cell Proteomics*, **12**, 369–380.
Højlund,K. *et al.* (2009) In vivo phosphoproteome of human skeletal muscle revealed by phosphopeptide enrichment and HPLC-ESI-MS/MS. *J. Proteome Res.*, **8**, 4954–4965.
Holt,L.J. *et al.* (2009) Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science*, **325**, 1682–1686.
Hong,Xuehui *et al.* (2018) Targeting posttranslational modifications of RIOK1 inhibits the progression of colorectal and gastric cancers. *Elife*, **7**.
Hornbeck,P.V. *et al.* (2019) 15 years of PhosphoSitePlus®: integrating post-translationally modified sites, disease variants and isoforms. *Nucleic Acids Res.*, **47**, D433–D441.
Hornbeck,P.V. *et al.* (2015) PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.*, **43**, D512-520.
Howden,A.J.M. *et al.* (2011) The phosphoproteome of Arabidopsis plants lacking the oxidative signal-inducible1 (OXI1) protein kinase. *New Phytol.*, **190**, 49–56.
Hsu,J.-L. *et al.* (2009) Functional phosphoproteomic profiling of phosphorylation sites in membrane fractions of salt-stressed Arabidopsis thaliana. *Proteome Sci*, **7**, 42.
Hsu,J.-M. *et al.* (2011) Crosstalk between Arg 1175 methylation and Tyr 1173 phosphorylation negatively modulates EGFR-mediated ERK activation. *Nat. Cell Biol.*, **13**, 174–181.
Huber,A. *et al.* (2009) Characterization of the rapamycin-sensitive phosphoproteome reveals that Sch9 is a central coordinator of protein synthesis. *Genes Dev.*, **23**, 1929–1943.
Iliuk,A. *et al.* (2012) Chemical visualization of phosphoproteomes on membrane. *Mol. Cell Proteomics*, **11**, 629–639.
Ingelsson,B. and Vener,A.V. (2012) Phosphoproteomics of Arabidopsis chloroplasts reveals involvement of the STN7 kinase in phosphorylation of nucleoid protein pTAC16. *FEBS Lett.*, **586**, 1265–1271.
Ingrell,C.R. *et al.* (2007) NetPhosYeast: Prediction of Protein Phosphorylation Sites in Yeast. *Bioinformatics*, **23**, 895–897.
Ito,J. *et al.* (2009) A survey of the Arabidopsis thaliana mitochondrial phosphoproteome. *Proteomics*, **9**, 4229–4240.
Iwai,L.K. *et al.* (2010) Quantitative phosphoproteomic analysis of T cell receptor signaling in diabetes prone and resistant mice. *J. Proteome Res.*, **9**, 3135–3145.
Jedrychowski,M.P. *et al.* (2011) Evaluation of HCD- and CID-type fragmentation within their respective detection platforms for murine phosphoproteomics. *Mol. Cell Proteomics*, **10**, M111.009910.
Jones,A.M.E. *et al.* (2009) Phosphoproteomic analysis of nuclei-enriched fractions from Arabidopsis thaliana. *J Proteomics*, **72**, 439–451.
Joughin,B.A. *et al.* (2009) An integrated comparative phosphoproteomic and bioinformatic approach reveals a novel class of MPM-2 motifs upregulated in EGFRvIII-expressing glioblastoma cells. *Mol Biosyst*, **5**, 59–67.
Jouy,F. *et al.* (2015) Integration of conventional quantitative and phospho-proteomics reveals new elements in activated Jurkat T-cell receptor pathway maintenance. *Proteomics*, **15**, 25–33.

Ju,Z. *et al.* (2015) iLM-2L: A two-level predictor for identifying protein lysine methylation sites and their methylation degrees by incorporating K-gap amino acid pairs into Chou's general PseAAC. *J. Theor. Biol.*, **385**, 50–57.
Kettenbach,A.N. *et al.* (2011) Quantitative phosphoproteomics identifies substrates and functional modules of Aurora and Polo-like kinase activities in mitotic cells. *Sci Signal*, **4**, rs5.
Kettenbach,A.N. *et al.* (2012) Rapid determination of multiple linear kinase substrate motifs by mass spectrometry. *Chem. Biol.*, **19**, 608–618.
Kettenbach,A.N. and Gerber,S.A. (2011) Rapid and reproducible single-stage phosphopeptide enrichment of complex peptide mixtures: application to general and phosphotyrosine-specific phosphoproteomics experiments. *Anal. Chem.*, **83**, 7635–7644.
Klammer,M. *et al.* (2014) Identification of significant features by the Global Mean Rank test. *PLoS ONE*, **9**, e104504.
Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.
Kumar,P. *et al.* (2017) PRmePRed: A protein arginine methylation prediction tool. *PLoS ONE*, **12**, e0183318.
Lai,A.C.-Y. *et al.* (2012) Complementary Fe(3+)- and Ti(4+)-immobilized metal ion affinity chromatography for purification of acidic and basic phosphopeptides. *Rapid Commun. Mass Spectrom.*, **26**, 2186–2194.
Lan,P. *et al.* (2012) Quantitative phosphoproteome profiling of iron-deficient Arabidopsis roots. *Plant Physiol.*, **159**, 403–417.
Landry,C.R. *et al.* (2014) Turnover of protein phosphorylation evolving under stabilizing selection. *Front Genet*, **5**, 245.
Landry,C.R. *et al.* (2009) Weak functional constraints on phosphoproteomes. *Trends Genet.*, **25**, 193–197.
Lanouette,S. *et al.* (2014) The functional diversity of protein lysine methylation. *Mol. Syst. Biol.*, **10**, 724.
Larsen,S.C. *et al.* (2016a) Proteome-wide analysis of arginine monomethylation reveals widespread occurrence in human cells. *Sci Signal*, **9**, rs9.
Larsen,S.C. *et al.* (2016b) Proteome-wide analysis of arginine monomethylation reveals widespread occurrence in human cells. *Sci Signal*, **9**, rs9.
Lassowskat,I. *et al.* (2013) PAPE (Prefractionation-Assisted Phosphoprotein Enrichment): A Novel Approach for Phosphoproteomic Analysis of Green Tissues from Plants. *Proteomes*, **1**, 254–274.
Lassowskat,I. *et al.* (2014) Sustained mitogen-activated protein kinase activation reprograms defense metabolism and phosphoprotein profile in Arabidopsis thaliana. *Front Plant Sci*, **5**, 554.
Lee,J. *et al.* (2013) MAPK Hog1 closes the S. cerevisiae glycerol channel Fps1 by phosphorylating and displacing its positive regulators. *Genes Dev.*, **27**, 2590–2601.
Lee,T.-Y. *et al.* (2014) Identification and characterization of lysine-methylated sites on histones and non-histone proteins. *Computational Biology and Chemistry*, **50**, 11–18.
Li,S. *et al.* (2014) The protein phosphatase subunit PP2A-B'γ is required to suppress day length-dependent pathogenesis responses triggered by intracellular oxidative stress. *New Phytol.*, **202**, 145–160.
Li,W. *et al.* (2019) A methylation-phosphorylation switch determines Plk1 kinase activity and function in DNA damage repair. *Sci Adv*, **5**, eaau7566.
Li,X. *et al.* (2007) Large-scale phosphorylation analysis of alpha-factor-arrested Saccharomyces cerevisiae. *J. Proteome Res.*, **6**, 1190–1197.
Lienhard,G.E. (2008) Non-functional phosphorylations? *Trends Biochem. Sci.*, **33**, 351–352.

Lin,L.-L. *et al.* (2015) Integrating Phosphoproteomics and Bioinformatics to Study Brassinosteroid-Regulated Phosphorylation Dynamics in Arabidopsis. *BMC Genomics*, **16**, 533.
Lu,J. *et al.* (2013) Facile synthesis of Fe3O4@mesoporous TiO2 microspheres for selective enrichment of phosphopeptides for phosphoproteomics analysis. *Talanta*, **105**, 20–27.
Luerman,G.C. *et al.* (2014) Phosphoproteomic evaluation of pharmacological inhibition of leucine-rich repeat kinase 2 reveals significant off-target effects of LRRK-2-IN-1. *J. Neurochem.*, **128**, 561–576.
Luo,F. *et al.* (2019) DeepPhos: prediction of protein phosphorylation sites with deep learning. *Bioinformatics*.
Ly,T. *et al.* (2014) A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells. *Elife*, **3**, e01630.
Mascaraque,V. *et al.* (2013) Phosphoproteomic analysis of protein kinase C signaling in Saccharomyces cerevisiae reveals Slt2 mitogen-activated protein kinase (MAPK)-dependent phosphorylation of eisosome core components. *Mol. Cell Proteomics*, **12**, 557–574.
Matic,K. *et al.* (2014) Quantitative phosphoproteomics of murine Fmr1-KO cell lines provides new insights into FMRP-dependent signal transduction mechanisms. *J. Proteome Res.*, **13**, 4388–4397.
Mattei,B. *et al.* (2016) Comprehensive Analysis of the Membrane Phosphoproteome Regulated by Oligogalacturonides in Arabidopsis thaliana. *Front Plant Sci*, **7**, 1107.
Mäusbacher,N. *et al.* (2010) Glycoprotein capture and quantitative phosphoproteomics indicate coordinated regulation of cell migration upon lysophosphatidic acid stimulation. *Mol. Cell Proteomics*, **9**, 2337–2353.
Mayank,P. *et al.* (2012) Characterization of the phosphoproteome of mature Arabidopsis pollen. *Plant J.*, **72**, 89–101.
McNulty,D.E. and Annan,R.S. (2008) Hydrophilic interaction chromatography reduces the complexity of the phosphoproteome and improves global phosphopeptide isolation and detection. *Mol. Cell Proteomics*, **7**, 971–980.
Melo-Braga,M.N. *et al.* (2014) Comprehensive quantitative comparison of the membrane proteome, phosphoproteome, and sialiome of human embryonic and neural stem cells. *Mol. Cell Proteomics*, **13**, 311–328.
Menz,J. *et al.* (2016) Early nitrogen-deprivation responses in Arabidopsis roots reveal distinct differences on transcriptome and (phospho-) proteome levels between nitrate and ammonium nutrition. *Plant J.*, **88**, 717–734.
Mithoe,S.C. *et al.* (2012) Targeted quantitative phosphoproteomics approach for the detection of phospho-tyrosine signaling in plants. *J. Proteome Res.*, **11**, 438–448.
Moses,A.M. *et al.* (2007) Clustering of phosphorylation site recognition motifs can be exploited to predict the targets of cyclin-dependent kinase. *Genome Biol.*, **8**, R23.
Murn,J. and Shi,Y. (2017) The winding path of protein methylation research: milestones and new frontiers. *Nat. Rev. Mol. Cell Biol.*, **18**, 517–527.
Nakagami,H. *et al.* (2010) Large-scale comparative phosphoproteomics identifies conserved phosphorylation sites in plants. *Plant Physiol.*, **153**, 1161–1174.
Nakayasu,E.S. *et al.* (2013) Comparative phosphoproteomics reveals components of host cell invasion and post-transcriptional regulation during Francisella infection. *Mol. Cell Proteomics*, **12**, 3297–3309.
Narumi,R. *et al.* (2012) A strategy for large-scale phosphoproteomics and SRM-based validation of human breast cancer tissue samples. *J. Proteome Res.*, **11**, 5311–5322.

Needham,E.J. *et al.* (2019) Illuminating the dark phosphoproteome. *Sci Signal*, **12**.
Nguyen,V. *et al.* (2009) A new approach for quantitative phosphoproteomic dissection of signaling pathways applied to T cell receptor activation. *Mol. Cell Proteomics*, **8**, 2418–2431.
Niittylä,T. *et al.* (2007) Temporal analysis of sucrose-induced phosphorylation changes in plasma membrane proteins of Arabidopsis. *Mol. Cell Proteomics*, **6**, 1711–1726.
Noh,K.-M. *et al.* (2015) ATRX tolerates activity-dependent histone H3 methyl/phos switching to maintain repetitive element silencing in neurons. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 6820–6827.
Nühse,T.S. *et al.* (2007) Quantitative phosphoproteomic analysis of plasma membrane proteins reveals regulatory mechanisms of plant innate immune responses. *Plant J.*, **51**, 931–940.
Nukarinen,E. *et al.* (2017) Protein sumoylation and phosphorylation intersect in Arabidopsis signaling. *Plant J.*, **91**, 505–517.
Old,W.M. *et al.* (2009) Functional proteomics identifies targets of phosphorylation by B-Raf signaling in melanoma. *Mol. Cell*, **34**, 115–131.
Oliveira,A.P. *et al.* (2012) Regulation of yeast central metabolism by enzyme phosphorylation. *Mol. Syst. Biol.*, **8**, 623.
Olsen,J.B. *et al.* (2016) Quantitative Profiling of the Activity of Protein Lysine Methyltransferase SMYD2 Using SILAC-Based Proteomics. *Mol. Cell Proteomics*, **15**, 892–905.
Onwuli,D.O. *et al.* (2017) Mapping arginine methylation in the human body and cardiac disease. *Proteomics Clin Appl*, **11**.
Osinalde,N. *et al.* (2015) Simultaneous dissection and comparison of IL-2 and IL-15 signaling pathways by global quantitative phosphoproteomics. *Proteomics*, **15**, 520–531.
Ostasiewicz,P. *et al.* (2010) Proteome, phosphoproteome, and N-glycoproteome are quantitatively preserved in formalin-fixed paraffin-embedded tissue and analyzable by high-resolution mass spectrometry. *J. Proteome Res.*, **9**, 3688–3700.
Ozlü,N. *et al.* (2010) Binding partner switching on microtubules and aurora-B in the mitosis to cytokinesis transition. *Mol. Cell Proteomics*, **9**, 336–350.
Palmisano,G. *et al.* (2012a) A novel method for the simultaneous enrichment, identification, and quantification of phosphopeptides and sialylated glycopeptides applied to a temporal profile of mouse brain development. *Mol. Cell Proteomics*, **11**, 1191–1202.
Palmisano,G. *et al.* (2012b) A novel method for the simultaneous enrichment, identification, and quantification of phosphopeptides and sialylated glycopeptides applied to a temporal profile of mouse brain development. *Mol. Cell Proteomics*, **11**, 1191–1202.
Pan,X. *et al.* (2013) Global protein phosphorylation dynamics during deoxynivalenol-induced ribotoxic stress response in the macrophage. *Toxicol. Appl. Pharmacol.*, **268**, 201–211.
Park,S.-S. and Maudsley,S. (2011) Discontinuous pH gradient-mediated separation of TiO2-enriched phosphopeptides. *Anal. Biochem.*, **409**, 81–88.
Pines,A. *et al.* (2011) Global phosphoproteome profiling reveals unanticipated networks responsive to cisplatin treatment of embryonic stem cells. *Mol. Cell. Biol.*, **31**, 4964–4977.
Pinto,S.M. *et al.* (2015) Quantitative phosphoproteomic analysis of IL-33-mediated signaling. *Proteomics*, **15**, 532–544.

Plank,M. *et al.* (2015) Expanding the yeast protein arginine methylome. *Proteomics*, **15**, 3232–3243.
Poulard,C. *et al.* (2017) A post-translational modification switch controls coactivator function of histone methyltransferases G9a and GLP. *EMBO Rep.*, **18**, 1442–1459.
Qing,D. *et al.* (2016) Quantitative and Functional Phosphoproteomic Analysis Reveals that Ethylene Regulates Water Transport via the C-Terminal Phosphorylation of Aquaporin PIP2;1 in Arabidopsis. *Mol Plant*, **9**, 158–174.
Qiu,W.-R. *et al.* (2014) iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. *Biomed Res Int*, **2014**, 947416.
Qiu,W.-R. *et al.* (2016) iPTM-mLys: identifying multiple lysine PTM sites and their different types. *Bioinformatics*, **32**, 3116–3123.
Rayapuram,N. *et al.* (2014) Identification of novel PAMP-triggered phosphorylation and dephosphorylation events in Arabidopsis thaliana by quantitative phosphoproteomic analysis. *J. Proteome Res.*, **13**, 2137–2151.
Reiland,S. *et al.* (2009) Large-scale Arabidopsis phosphoproteome profiling reveals novel chloroplast kinase substrates and phosphorylation networks. *Plant Physiol.*, **150**, 889–903.
Roitinger,E. *et al.* (2015) Quantitative phosphoproteomics of the ataxia telangiectasia-mutated (ATM) and ataxia telangiectasia-mutated and rad3-related (ATR) dependent DNA damage response in Arabidopsis thaliana. *Mol. Cell Proteomics*, **14**, 556–571.
Ruperez,P. *et al.* (2012) Quantitative phosphoproteomic analysis reveals a role for serine and threonine kinases in the cytoskeletal reorganization in early T cell receptor activation in human primary T cells. *Mol. Cell Proteomics*, **11**, 171–186.
Ruse,C.I. *et al.* (2008) Motif-specific sampling of phosphoproteomes. *J. Proteome Res.*, **7**, 2140–2150.
Sadowski,I. *et al.* (2013) The PhosphoGRID Saccharomyces cerevisiae protein phosphorylation site database: version 2.0 update. *Database (Oxford)*, **2013**, bat026.
Saleem,R.A. *et al.* (2010) Integrated phosphoproteomics analysis of a signaling network governing nutrient response and peroxisome induction. *Mol. Cell Proteomics*, **9**, 2076–2088.
Šalovská,B. *et al.* (2014) Radiosensitization of human leukemic HL-60 cells by ATR kinase inhibitor (VE-821): phosphoproteomic analysis. *Int J Mol Sci*, **15**, 12007–12026.
Santamaria,A. *et al.* (2011) The Plk1-dependent phosphoproteome of the early mitotic spindle. *Mol. Cell Proteomics*, **10**, M110.004457.
Scholten,A. *et al.* (2013) Phosphoproteomics study based on in vivo inhibition reveals sites of calmodulin-dependent protein kinase II regulation in the heart. *J Am Heart Assoc*, **2**, e000318.
Schreiber,S.L. and Bernstein,B.E. (2002) Signaling network model of chromatin. *Cell*, **111**, 771–778.
Schweiger,R. and Linial,M. (2010) Cooperativity within proximal phosphorylation sites is revealed from large-scale proteomics data. *Biol. Direct*, **5**, 6.
Schweppe,D.K. *et al.* (2013) Quantitative phosphoproteomic profiling of human non-small cell lung cancer tumors. *J Proteomics*, **91**, 286–296.
Shao,J. *et al.* (2009) Computational Identification of Protein Methylation Sites through Bi-Profile Bayes Feature Extraction. *PLoS ONE*, **4**, e4920.

Sharma,K. *et al.* (2010) Quantitative analysis of kinase-proximal signaling in lipopolysaccharide-induced innate immune response. *J. Proteome Res.*, **9**, 2539–2549.
Sharma,K. *et al.* (2012) Quantitative proteomics reveals that Hsp90 inhibition preferentially targets kinases and the DNA damage response. *Mol. Cell Proteomics*, **11**, M111.014654.
Sharma,K. *et al.* (2014) Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep*, **8**, 1583–1594.
Shevchuk,O. *et al.* (2014) HOPE-fixation of lung tissue allows retrospective proteome and phosphoproteome studies. *J. Proteome Res.*, **13**, 5230–5239.
Shi,Y. *et al.* (2015) Position-specific prediction of methylation sites from sequence conservation based on information theory. *Sci Rep*, **5**, 12403.
Shien,D.-M. *et al.* (2009) Incorporating structural characteristics for identification of protein methylation sites. *J Comput Chem*, **30**, 1532–1543.
Shiromizu,T. *et al.* (2013) Identification of missing proteins in the neXtProt database and unregistered phosphopeptides in the PhosphoSitePlus database as part of the Chromosome-centric Human Proteome Project. *J. Proteome Res.*, **12**, 2414–2421.
Sims,R.J. and Reinberg,D. (2008) Is there a code embedded in proteins that is based on post-translational modifications? *Nat. Rev. Mol. Cell Biol.*, **9**, 815–820.
Soderblom,E.J. *et al.* (2013) Proteomic analysis of ERK1/2-mediated human sickle red blood cell membrane protein phosphorylation. *Clin Proteomics*, **10**, 1.
Song,C. *et al.* (2012) Systematic analysis of protein phosphorylation networks from phosphoproteomic data. *Mol. Cell Proteomics*, **11**, 1070–1083.
Song,H. *et al.* (2018) Crosstalk between lysine methylation and phosphorylation of ATG16L1 dictates the apoptosis of hypoxia/reoxygenation-induced cardiomyocytes. *Autophagy*, **14**, 825–844.
Soufi,B. *et al.* (2009) Global analysis of the yeast osmotic stress response by quantitative proteomics. *Mol Biosyst*, **5**, 1337–1346.
Stokes,M.P. *et al.* (2012a) PTMScan direct: identification and quantification of peptides from critical signaling proteins by immunoaffinity enrichment coupled with LC-MS/MS. *Mol. Cell Proteomics*, **11**, 187–201.
Stokes,M.P. *et al.* (2012b) PTMScan direct: identification and quantification of peptides from critical signaling proteins by immunoaffinity enrichment coupled with LC-MS/MS. *Mol. Cell Proteomics*, **11**, 187–201.
Strahl,B.D. and Allis,C.D. (2000) The language of covalent histone modifications. *Nature*, **403**, 41–45.
Studer,R.A. *et al.* (2016) Evolution of protein phosphorylation across 18 fungal species. *Science*, **354**, 229–232.
Sugiyama,N. *et al.* (2008) Large-scale phosphorylation mapping reveals the extent of tyrosine phosphorylation in Arabidopsis. *Mol. Syst. Biol.*, **4**, 193.
Sui,S. *et al.* (2008) Phosphoproteome analysis of the human Chang liver cells using SCX and a complementary mass spectrometric strategy. *Proteomics*, **8**, 2024–2034.
Sweet,S.M.M. *et al.* (2009) Large scale localization of protein phosphorylation by use of electron capture dissociation mass spectrometry. *Mol. Cell Proteomics*, **8**, 904–912.
Sylvestersen,K.B. *et al.* (2014) Proteomic analysis of arginine methylation sites in human cells reveals dynamic regulation during transcriptional arrest. *Mol. Cell Proteomics*, **13**, 2072–2088.

Tan,H. *et al.* (2015) Refined phosphopeptide enrichment by phosphate additive and the analysis of human brain phosphoproteome. *Proteomics*, **15**, 500–507.
Taus,T. *et al.* (2011) Universal and confident phosphorylation site localization using phosphoRS. *J. Proteome Res.*, **10**, 5354–5362.
Tong,J. *et al.* (2017) Protein-phosphotyrosine proteome profiling by superbinder-SH2 domain affinity purification mass spectrometry, sSH2-AP-MS. *Proteomics*, **17**.
Trost,B. and Kusalik,A. (2013) Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights. *Bioinformatics*, **29**, 686–694.
Trost,B. and Kusalik,A. (2011) Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics*, **27**, 2927–2935.
Umezawa,T. *et al.* (2013) Genetics and phosphoproteomics reveal a protein phosphorylation network in the abscisic acid signaling pathway in Arabidopsis thaliana. *Sci Signal*, **6**, rs8.
Van Hoof,D. *et al.* (2009) Phosphorylation dynamics during early differentiation of human embryonic stem cells. *Cell Stem Cell*, **5**, 214–226.
Vandenbogaert,M. *et al.* (2012) Automated phosphopeptide identification using multiple MS/MS fragmentation modes. *J. Proteome Res.*, **11**, 5695–5703.
Varier,R.A. *et al.* (2010) A phospho/methyl switch at histone H3 regulates TFIID association with mitotic chromosomes. *EMBO J.*, **29**, 3967–3978.
Vlastaridis,P., Kyriakidou,P., Chaliotis,A., Van de Peer,Y., Oliver,Stephen G., and Amoutzias,G.D. (2017) Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes. *Gigascience*, **6**, 1–11.
Vlastaridis,P., Kyriakidou,P., Chaliotis,A., Van de Peer,Y., Oliver,Stephen G, and Amoutzias,G.D. (2017) Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes. *Gigascience*.
Vlastaridis,P. *et al.* (2016) The Challenges of Interpreting Phosphoproteomics Data: A Critical View Through the Bioinformatics Lens. In, Angelini,C. *et al.* (eds), *Computational Intelligence Methods for Bioinformatics and Biostatistics*. Springer International Publishing, Cham, pp. 196–204.
Vlastaridis,P., Papakyriakou,A., *et al.* (2017) The Pivotal Role of Protein Phosphorylation in the Control of Yeast Central Metabolism. *G3 (Bethesda)*, **7**, 1239–1249.
Vu,L.D. *et al.* (2016) Up-to-Date Workflow for Plant (Phospho)proteomics Identifies Differential Drought-Responsive Phosphorylation Events in Maize Leaves. *J. Proteome Res.*, **15**, 4304–4317.
Wang,C. *et al.* (2013) Determination of CK2 specificity and substrates by proteome-derived peptide libraries. *J. Proteome Res.*, **12**, 3813–3821.
Wang,D. *et al.* (2017) MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics*, **33**, 3909–3916.
Wang,P. *et al.* (2013) Quantitative phosphoproteomics identifies SnRK2 protein kinase substrates and reveals the effectors of abscisic acid action. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 11205–11210.
Wang,X. *et al.* (2013) A large-scale protein phosphorylation analysis reveals novel phosphorylation motifs and phosphoregulatory networks in Arabidopsis. *J Proteomics*, **78**, 486–498.
Wang,Z. *et al.* (2009) A simple and effective method for detecting phosphopeptides for phosphoproteomic analysis. *J Proteomics*, **72**, 831–835.

Weber,C. *et al.* (2012) Dual phosphoproteomics and chemical proteomics analysis of erlotinib and gefitinib interference in acute myeloid leukemia cells. *J Proteomics*, **75**, 1343–1356.
Wei,L. *et al.* (2017) Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Trans Comput Biol Bioinform*.
Weinert,B.T. *et al.* (2014) Acetylation dynamics and stoichiometry in Saccharomyces cerevisiae. *Mol. Syst. Biol.*, **10**, 716.
Weintz,G. *et al.* (2010) The phosphoproteome of toll-like receptor-activated macrophages. *Mol. Syst. Biol.*, **6**, 371.
Wen,P.-P. *et al.* (2016) Accurate in silico prediction of species-specific methylation sites based on information gain feature optimization. *Bioinformatics*, **32**, 3107–3115.
Whiteman,S.-A. *et al.* (2008) Identification of novel proteins and phosphorylation sites in a tonoplast enriched membrane fraction of Arabidopsis thaliana. *Proteomics*, **8**, 3536–3547.
Wiese,H. *et al.* (2015) Quantitative phosphoproteomics reveals the protein tyrosine kinase Pyk2 as a central effector of olfactory receptor signaling in prostate cancer cells. *Biochim. Biophys. Acta*, **1854**, 632–640.
Wilson-Grady,J.T. *et al.* (2013) Quantitative comparison of the fasted and re-fed mouse liver phosphoproteomes using lower pH reductive dimethylation. *Methods*, **61**, 277–286.
Wojcechowskyj,J.A. *et al.* (2011) Quantitative phosphoproteomics of CXCL12 (SDF-1) signaling. *PLoS ONE*, **6**, e24918.
Wu,F. *et al.* (2010) Studies of phosphoproteomic changes induced by nucleophosmin-anaplastic lymphoma kinase (ALK) highlight deregulation of tumor necrosis factor (TNF)/Fas/TNF-related apoptosis-induced ligand signaling pathway in ALK-positive anaplastic large cell lymphoma. *Mol. Cell Proteomics*, **9**, 1616–1632.
Wu,J. *et al.* (2010) Integrating titania enrichment, iTRAQ labeling, and Orbitrap CID-HCD for global identification and quantitative analysis of phosphopeptides. *Proteomics*, **10**, 2224–2234.
Wu,R. *et al.* (2011) Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes. *Mol. Cell Proteomics*, **10**, M111.009654.
Wu,X. *et al.* (2012) Investigation of receptor interacting protein (RIP3)-dependent protein phosphorylation by quantitative phosphoproteomics. *Mol. Cell Proteomics*, **11**, 1640–1651.
Wu,X.N. *et al.* (2013) Sucrose-induced receptor kinase SIRK1 regulates a plasma membrane aquaporin in Arabidopsis. *Mol. Cell Proteomics*, **12**, 2856–2873.
Wu,Y.-B. *et al.* (2009) Concurrent quantification of proteome and phosphoproteome to reveal system-wide association of protein phosphorylation and gene expression. *Mol. Cell Proteomics*, **8**, 2809–2826.
Wu,Z. *et al.* (2015) A chemical proteomics approach for global analysis of lysine monomethylome profiling. *Mol. Cell Proteomics*, **14**, 329–339.
Xia,Q. *et al.* (2008) Phosphoproteomic analysis of human brain by calcium phosphate precipitation and mass spectrometry. *J. Proteome Res.*, **7**, 2845–2851.
Xiao,K. *et al.* (2010) Global phosphorylation analysis of beta-arrestin-mediated signaling downstream of a seven transmembrane receptor (7TMR). *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 15299–15304.

Xie,X. *et al.* (2010) A comparative phosphoproteomic analysis of a human tumor metastasis model using a label-free quantitative approach. *Electrophoresis*, **31**, 1842–1852.
Xu,S. *et al.* (2008) Capture of phosphopeptides using alpha-zirconium phosphate nanoplatelets. *Anal. Chem.*, **80**, 5542–5549.
Xue,L. *et al.* (2013) Quantitative measurement of phosphoproteome response to osmotic stress in arabidopsis based on Library-Assisted eXtracted Ion Chromatogram (LAXIC). *Mol. Cell Proteomics*, **12**, 2354–2369.
Xue,L. *et al.* (2012) Sensitive kinase assay linked with phosphoproteomics for identifying direct kinase substrates. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 5615–5620.
Xue,Y. *et al.* (2008) GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol. Cell Proteomics*, **7**, 1598–1608.
Yagoub,D. *et al.* (2015) Yeast proteins Gar1p, Nop1p, Npl3p, Nsr1p, and Rps2p are natively methylated and are substrates of the arginine methyltransferase Hmt1p. *Proteomics*, **15**, 3209–3218.
Yakubu,R.R. *et al.* (2017) Comparative Monomethylarginine Proteomics Suggests that Protein Arginine Methyltransferase 1 (PRMT1) is a Significant Contributor to Arginine Monomethylation in Toxoplasma gondii. *Mol. Cell Proteomics*, **16**, 567–580.
Yamagata,K. *et al.* (2008) Arginine methylation of FOXO transcription factors inhibits their phosphorylation by Akt. *Mol. Cell*, **32**, 221–231.
Yan,G.-R. *et al.* (2011) Characterization of phosphoproteins in gastric cancer secretome. *OMICS*, **15**, 83–90.
Yang,F. *et al.* (2007) Applying a targeted label-free approach using LC-MS AMT tags to evaluate changes in protein phosphorylation following phosphatase inhibition. *J. Proteome Res.*, **6**, 4489–4497.
Yang,F. *et al.* (2010) Phosphoproteomics profiling of human skin fibroblast cells reveals pathways and proteins affected by low doses of ionizing radiation. *PLoS ONE*, **5**, e14152.
Yang,X.-L. *et al.* (2013) Identification of complex relationship between protein kinases and substrates during the cell cycle of HeLa cells by phosphoproteomic analysis. *Proteomics*, **13**, 1233–1246.
Yang,Z. *et al.* (2013) Stable isotope metabolic labeling-based quantitative phosphoproteomic analysis of Arabidopsis mutants reveals ethylene-regulated time-dependent phosphoproteins and putative substrates of constitutive triple response 1 kinase. *Mol. Cell Proteomics*, **12**, 3559–3582.
Yao,Q. *et al.* (2011) SUMOylation-regulated protein phosphorylation, evidence from quantitative phosphoproteomics analyses. *J. Biol. Chem.*, **286**, 27342–27349.
Ye,X. and Li,L. (2014) Macroporous reversed-phase separation of proteins combined with reversed-phase separation of phosphopeptides and tandem mass spectrometry for profiling the phosphoproteome of MDA-MB-231 cells. *Electrophoresis*, **35**, 3479–3486.
Yu,Y. *et al.* (2011) Phosphoproteomic analysis identifies Grb10 as an mTORC1 substrate that negatively regulates insulin signaling. *Science*, **332**, 1322–1326.
Zhang,Hongtao *et al.* (2013) Quantitative phosphoproteomics after auxin-stimulated lateral root induction identifies an SNX1 protein phosphorylation site required for growth. *Mol. Cell Proteomics*, **12**, 1158–1169.

Zhang,H. *et al.* (2013) SILAC-based phosphoproteomics reveals an inhibitory role of KSR1 in p53 transcriptional activity via modulation of DBC1. *Br. J. Cancer*, **109**, 2675–2684.
Zhang,L. *et al.* (2013) Characterization of the novel broad-spectrum kinase inhibitor CTx-0294885 as an affinity reagent for mass spectrometry-based kinome profiling. *J. Proteome Res.*, **12**, 3104–3116.
Zhang,X. *et al.* (2017) Quantitative Tyrosine Phosphoproteomics of Epidermal Growth Factor Receptor (EGFR) Tyrosine Kinase Inhibitor-treated Lung Adenocarcinoma Cells Reveals Potential Novel Biomarkers of Therapeutic Response. *Mol. Cell Proteomics*, **16**, 891–910.
Zheng,J. *et al.* (2013) Urinary proteomic and non-prefractionation quantitative phosphoproteomic analysis during pregnancy and non-pregnancy. *BMC Genomics*, **14**, 777.
Zhong,C.-Q. *et al.* (2014) Quantitative phosphoproteomic analysis of RIP3-dependent protein phosphorylation in the course of TNF-induced necroptosis. *Proteomics*, **14**, 713–724.
Zhong,J. *et al.* (2015) Quantitative phosphoproteomics reveals crosstalk between phosphorylation and O-GlcNAc in the DNA damage response pathway. *Proteomics*, **15**, 591–607.
Zhong,J. *et al.* (2012) TSLP signaling network revealed by SILAC-based phosphoproteomics. *Mol. Cell Proteomics*, **11**, M112.017764.

# Chapter 6

# Title: Development of a graph database for storing and organizing protein phosphorylation data.

## Preface

The structural data of the database have been prepared and provided by Dr. Stratikos and Dr. Papakyriakou, whereas I have developed the database schema, the server, the visualization and have integrated all the available data.

## Abstract

This chapter describes the graph database that was developed to store and organize the filtered phosphorylation sites, together with other types of data, such as domains, SNPs, structural information. A description of the web application is given in order to allow the user to conduct searches/queries in the graph database. Furthermore, I provide explanations of how the results are interpreted. The database is accessible at:

http://bioinf.bio.uth.gr/phospho-prometheus-db/

## Introduction

For the needs of the FAB-PHOS project, more than 1000 publications related to phosphoproteomics had to be manually inspected by the annotators, in order to determine which specific publications had high quality phosphorylation sites (p-sites) and with what criteria the raw data had to be filtered. Once the annotators identified the right publications and applied the stringent filtering criteria that our team implemented, a compendium of p-sites from various organisms was compiled. However, in order to efficiently use these data for complex and in-depth bioinformatics analyses, other types

of relevant data had to be integrated. These other data included the functional annotation of the phosphoproteins, whether a p-site was detected within a conserved functional domain or not, whether Single Nucleotide Polymorphisms (SNPs) of clinical importance were found in the vicinity of the p-site, that could affect its phosphorylation by the cognate kinase. Furthermore, it was important to know how many publications supported the detection of a specific p-site, how well conserved was this p-site in other homologous proteins and whether this particular p-site was identified in a structurally important region of the protein. Thus, the goal was to integrate all these types of data in a newly developed database that would facilitate our bioinformatics analyses. In addition, the goal was to make this database publicly available and help other researchers in the field identify targets of interest, or verify some of their findings. A widely used and publicly available database that integrates post-translational modifications from various organisms is Phosphosite+ (Hornbeck *et al.* 2015). However, this database provides compendia of phosphorylation sites based on its own filtering criteria and does not integrate all the above mentioned information.

# Materials and Methods

## Database schema/organization

For our database management system, we chose the Neo4J graph database ("(Neo4j)-[") where data is stored on a graph data model (Angles and Gutierrez 2008; kumar Kaliyar 2015). Neo4J is an open – source project written in Java and has an enterprise Edition for commercial use. Neo4J is the most popular technology for graph databases as shown in (Angles and Gutierrez 2008) and in (Kolomičenko *et al.* 2013).

Graph databases have three main key advantages over traditional relational database management systems (RDBMS). One of them is performance. For intensive data relationship handling, graph databases improve performance by several orders of magnitude. The second key advantage is flexibility. With graph databases, developers and researchers have the advantage to alter their database schema without rebuilding the database from the beginning, because the structure and schema of a graph model flexes as applications and requirements change. Rather than exhaustively modeling a domain ahead of time, data teams can add to the existing graph structure without endangering current functionality. The third key advantage is agility. Developing with graph databases aligns perfectly with today's agile, test-driven development practices, allowing our graph database to evolve in step with the rest of the application and any changing research requirements. Modern graph databases are equipped for frictionless development and graceful systems maintenance.

A graph model is composed mainly of two elements: a node and a relationship. Each node represents an entity such as genes, proteins, phosphorylation sites, domains, experiments or other types of data), and each relationship represents how two nodes are associated. A node can have properties such as Ensemble identifier, organism

name, coordinates on a genome, length of a nucleotide sequence, a protein or a motif amino-acid sequence and others. Nodes can also have labels which are used to group nodes together, such as human genes etc.

**Webserver development**

The webserver consists of two parts. These are the i) the Java (“Java | Oracle”) and the Spring framework (“Spring Framework”) for the back-end, ii) the front-end part of our code written in Javascript language (“Free JavaScript training, resources and examples for the community”) and the AngularJS 1.4 Framework (“Angular”). More specifically, the webserver accepts requests to REST endpoints. The web-requests are translated into queries for the graph database and are served back to the front-end in JSON format. Spring framework provides useful java libraries for dependency injection, REST API gateway programming as well as drivers for connecting to Neo4J server. The front-end part of our code is written in Javascript language and the AngularJS Framework. This is the code that is running on each client’s browser when visiting our website. It is used for designing the user interface (UI) and programming a friendly and quickly responsive user experience (UX). The UI is designed with Bootstrap 4 CSS and javascript libraries and the UX with the AngularJS - Javascript framework. Using Angular for the UI and following the Single Page Application Architecture a much-improved experience is offered to the user. The application feels faster because less bandwidth is being used, and no full-page refreshes are occurring as the user navigates through the application.

# Results and Discussion

### The entities in the graph model and their relationships.

The main entity of our database model is the phosphorylation site. A phosphorylation site belongs to a protein, which in turn is a translation of a transcript and finally this transcript belongs to a gene. Genes can also have gene annotations assigning Gene Ontology terms to them according to the Gene Ontology Consortium. All the above information can be included in a graph model as shown in Figure 1.

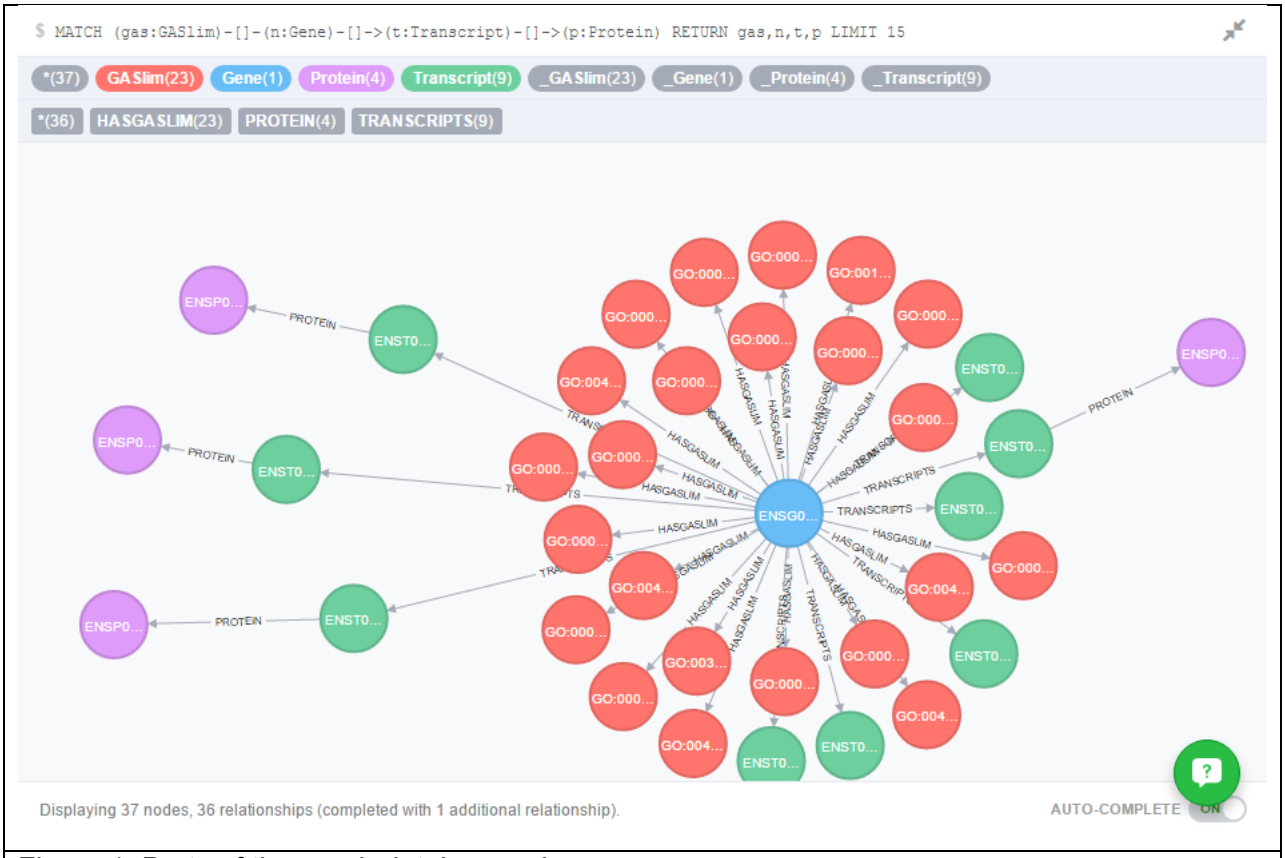



Figure 1. Parts of the graph database schema.

With blue, we have a gene entity which connects to gene annotation entities denoted by red, transcript nodes denoted by green. Transcript entities are connected to their translated protein entities denoted by color magenta.

Further down the graph, a protein may have phosphorylation sites, whereas proteins may also have domains. Phosphorylation sites are discovered in published papers by our annotators. Also phosphorylation sites may have homologs to other phosphorylation sites for mouse and rat. This can be modelled in our graph model as shown in Figure 2.

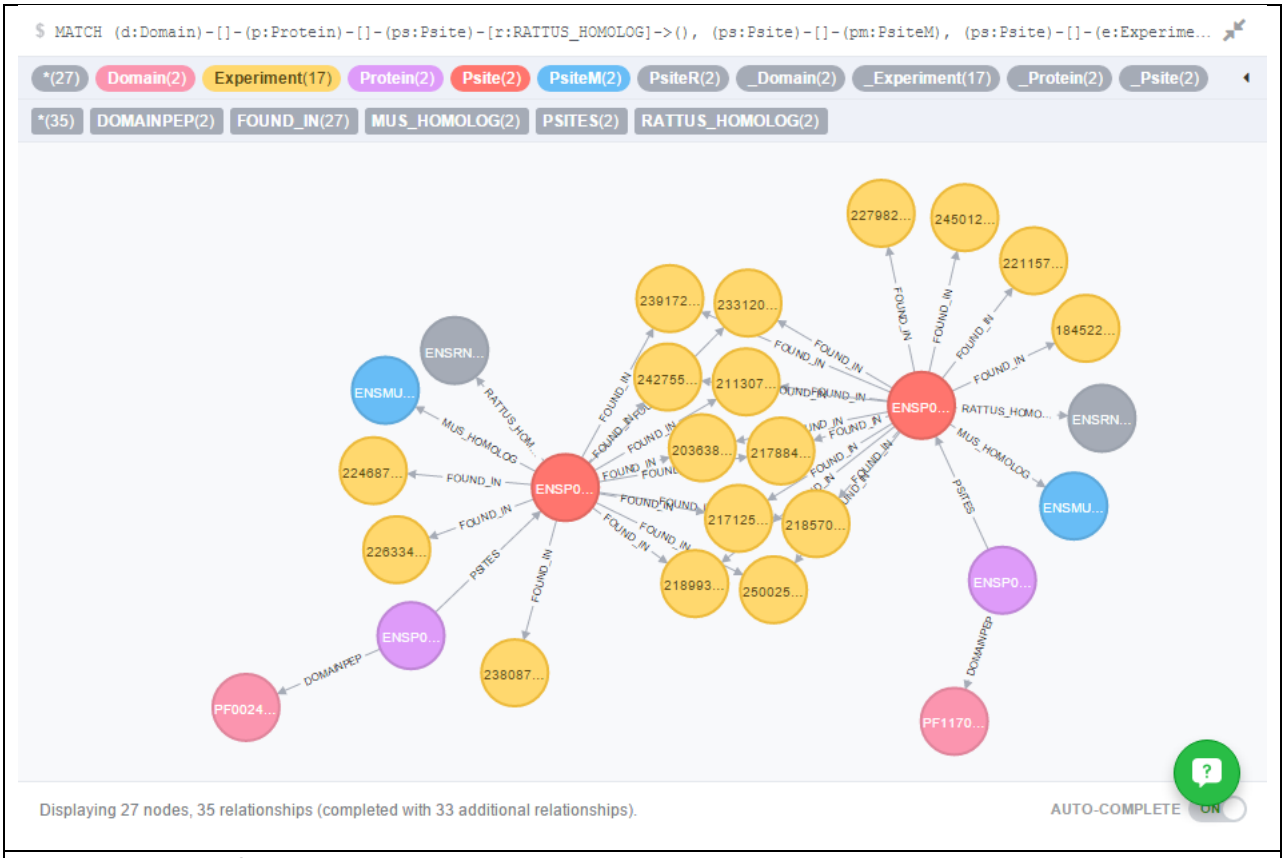



Figure 2. Parts of the graph database schema, related to proteins.

With orange we have phosphorylation site entities which connect to experiment entities denoted by yellow, mouse and rat homologs are denoted by blue and green respectively. Finally, domains denoted by red are connected to the phosphorylation sites' protein entities denoted by magenta.

In our dataset we have included 105,475 human phosphorylation sites, 51,151 mouse and 24,514 rat p-sites.

In order to perform data queries in Neo4J, I use Cypher, a declarative graph query language. Cypher is to graph databases what is SQL in relational data models.

**Web interface for searching the stored data.**

On the landing page of our web application, users can search in our graph database using our search form. Users must choose a species and a database entity (gene, domain, phosphorylation site, protein, GOA term or Experiment) as well as entering a text term for which the query is performed. Examples for the species *Homo sapiens* of text queries are shown in figure 3:

**Gene ensembl Identifier:** ENSG00000096384
**Gene Description:** heat shock protein 90kDa alpha
**Gene Name:** HSP90AA1
**Protein ensembl Identifier:** ENSP00000360609
**Phosphorylation Site ensembl protein identifier with position:** ENSP00000306010 or ENSP00000306010_58
**Domain PFAM Name:** PF09439.6 or PF09439
**Domain Acronym:** PF02290
**Domain Description:** Signal recognition particle
**Gene Annotation Term ID:** GO:0048646
**Gene Annotation Term Description:** anatomical structure
**Experiment Title: Functional proteomics**
**Experiment Author Surname: Francavilla**

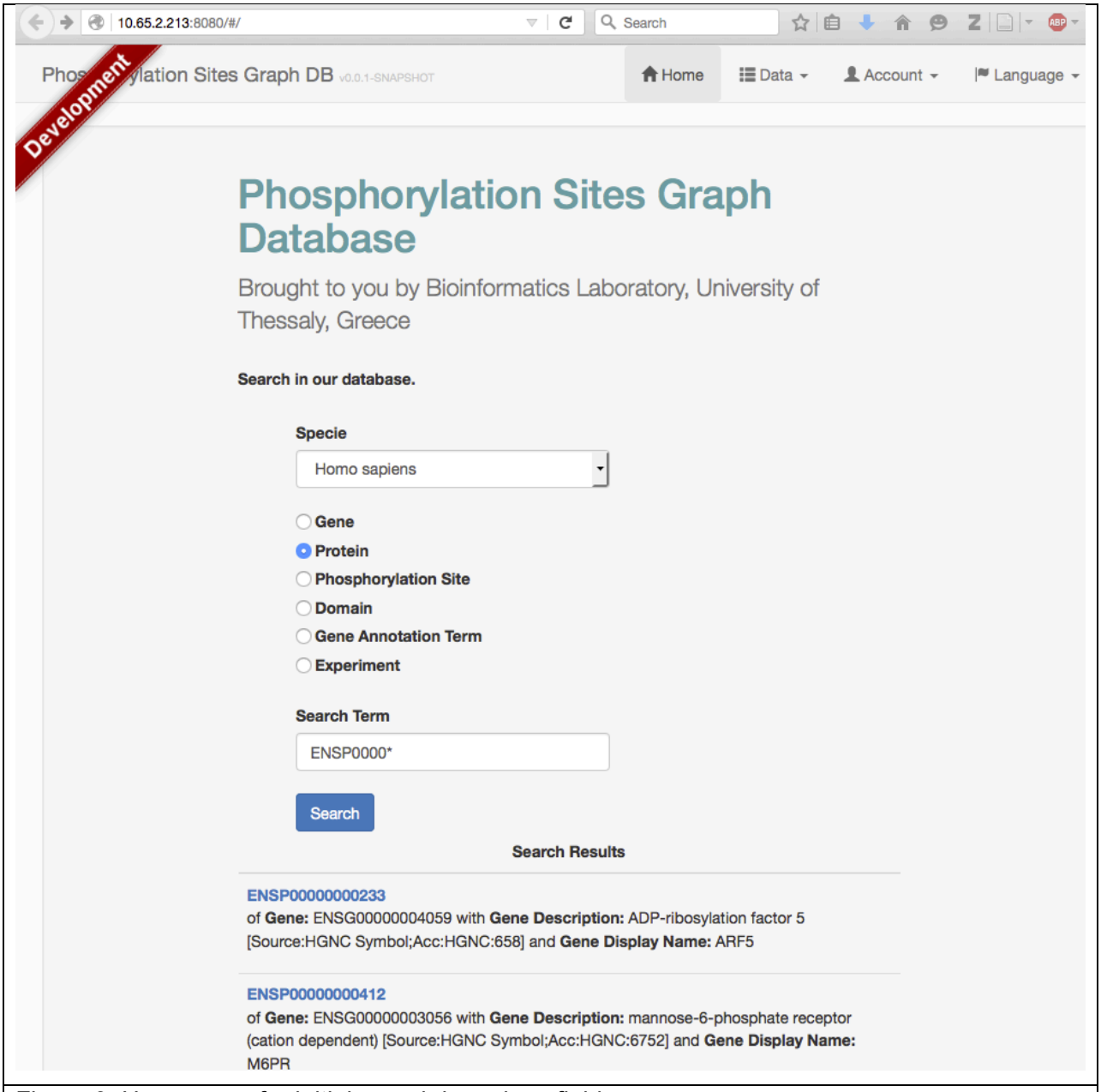



Figure 3. Home page for initial search in various fields.

On the gene page, information regarding a gene is shown, such as chromosome, coordinates, description etc. The transcript of the typical protein is denoted with red background. From the gene page, users can use a hyperlink to visit the corresponding protein page.

On the Protein page (see figure 4), the Transcript and Gene information are displayed, along with domains and phosphorylation sites of that protein. In addition, the coordinates of the domains on the protein are shown.

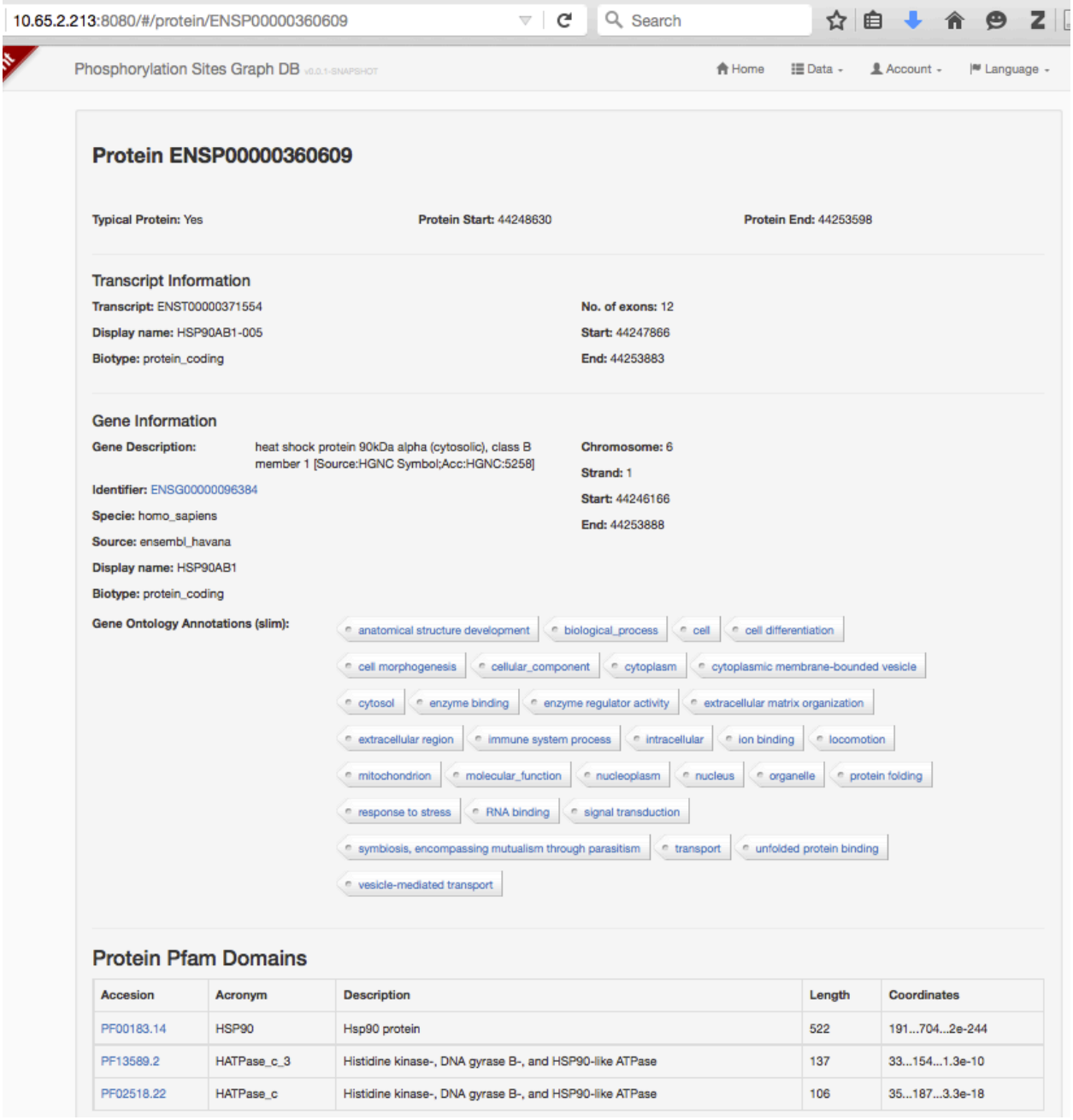
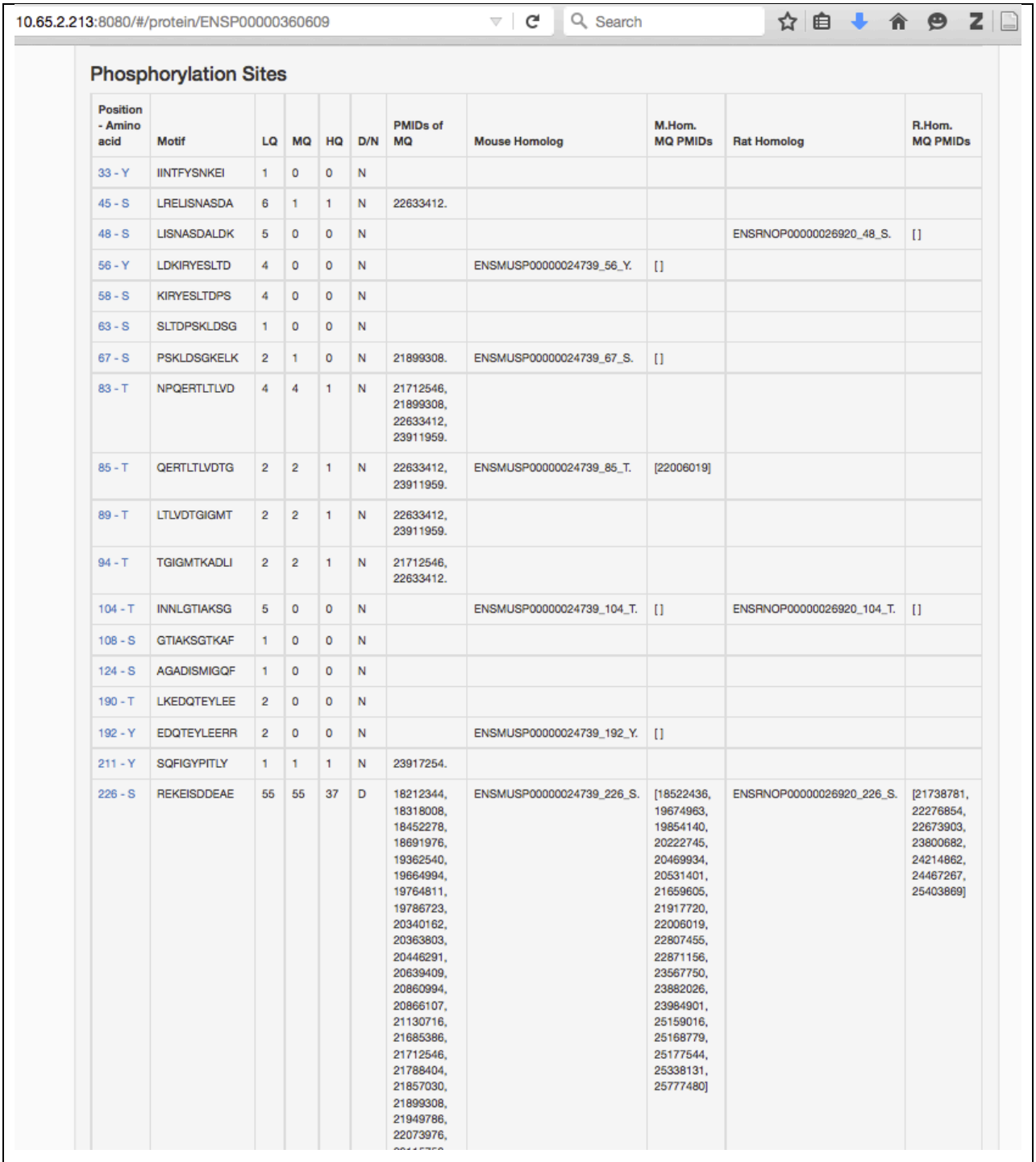



Figure 4. Search results for a certain protein. Gene ontology information as well as protein domain information is incorporated.

Within the protein page, there exists a phosphorylations table (see figure 5) that has the following information:

- LQ, MQ, HQ where a phosphorylation site is marked as low quality, medium quality or high quality depending on the techniques used in the experiments to identify them as phosphorylation sites
- D/N: denotes whether the site is located in a disordered region or not.
- PMIDs: Shows the PUBMED ids of the papers that detected this phosphorylation site.
- The other collumns show if the mouse of rat orthologs have a conserved amino acids and whether this conserved amino acid has been detected as phosphorylated, by literature.

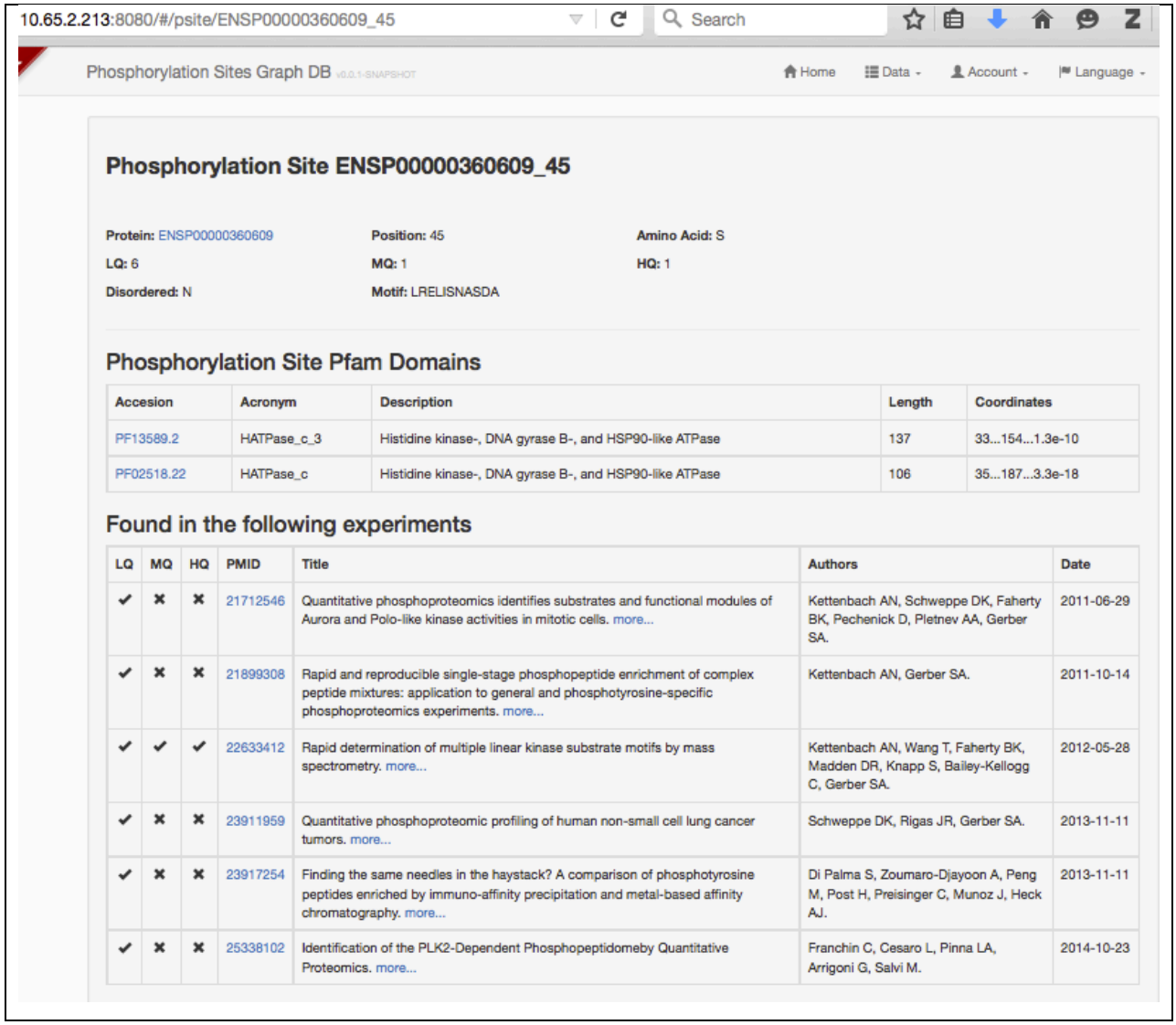



## Phosphorylation Sites

| Position - Amino acid | Motif | LQ | MQ | HQ | D/N | PMIDs of MQ | Mouse Homolog | M.Hom. MQ PMIDs | Rat Homolog | R.Hom. MQ PMIDs |
|---|---|---|---|---|---|---|---|---|---|---|
| 33 - Y | IINTFYSNKEI | 1 | 0 | 0 | N | | | | | |
| 45 - S | LRELISNASDA | 6 | 1 | 1 | N | 22633412. | | | | |
| 48 - S | LISNASDALDK | 5 | 0 | 0 | N | | | | ENSRNOP00000026920_48_S. | [] |
| 56 - Y | LDKIRYESLTD | 4 | 0 | 0 | N | | ENSMUSP00000024739_56_Y. | [] | | |
| 58 - S | KIRYESLTDPS | 4 | 0 | 0 | N | | | | | |
| 63 - S | SLTDPSKLDSG | 1 | 0 | 0 | N | | | | | |
| 67 - S | PSKLDSGKELK | 2 | 1 | 0 | N | 21899308. | ENSMUSP00000024739_67_S. | [] | | |
| 83 - T | NPQERTLTLVD | 4 | 4 | 1 | N | 21712546, 21899308, 22633412, 23911959. | | | | |
| 85 - T | QERTLTLVDTG | 2 | 2 | 1 | N | 22633412, 23911959. | ENSMUSP00000024739_85_T. | [22006019] | | |
| 89 - T | LTLVDTGIGMT | 2 | 2 | 1 | N | 22633412, 23911959. | | | | |
| 94 - T | TGIGMTKADLI | 2 | 2 | 1 | N | 21712546, 22633412. | | | | |
| 104 - T | INNLGTIAKSG | 5 | 0 | 0 | N | | ENSMUSP00000024739_104_T. | [] | ENSRNOP00000026920_104_T. | [] |
| 108 - S | GTIAKSGTKAF | 1 | 0 | 0 | N | | | | | |
| 124 - S | AGADISMIGQF | 1 | 0 | 0 | N | | | | | |
| 190 - T | LKEDQTEYLEE | 2 | 0 | 0 | N | | | | | |
| 192 - Y | EDQTEYLEERR | 2 | 0 | 0 | N | | ENSMUSP00000024739_192_Y. | [] | | |
| 211 - Y | SQFIGYPITLY | 1 | 1 | 1 | N | 23917254. | | | | |
| 226 - S | REKEISDDEAE | 55 | 55 | 37 | D | 18212344, 18318008, 18452278, 18691976, 19362540, 19664994, 19764811, 19786723, 20340162, 20363803, 20446291, 20639409, 20860994, 20866107, 21130716, 21685386, 21712546, 21788404, 21857030, 21899308, 21949786, 22073976, | ENSMUSP00000024739_226_S. | [18522436, 19674963, 19854140, 20222745, 20469934, 20531401, 21659605, 21917720, 22006019, 22807455, 22871156, 23567750, 23882026, 23984901, 25159016, 25168779, 25177544, 25338131, 25777480] | ENSRNOP00000026920_226_S. | [21738781, 22276854, 22673903, 23800682, 24214862, 24467267, 25403869] |

Figure 5. Here, the p-sites of the protein are displayed together with their experimental overlap and confidence, as well as evolutionary information regarding conservation of p-site in other species.

The user may select a specific p-site and view further information about it, as shown in figure 6.



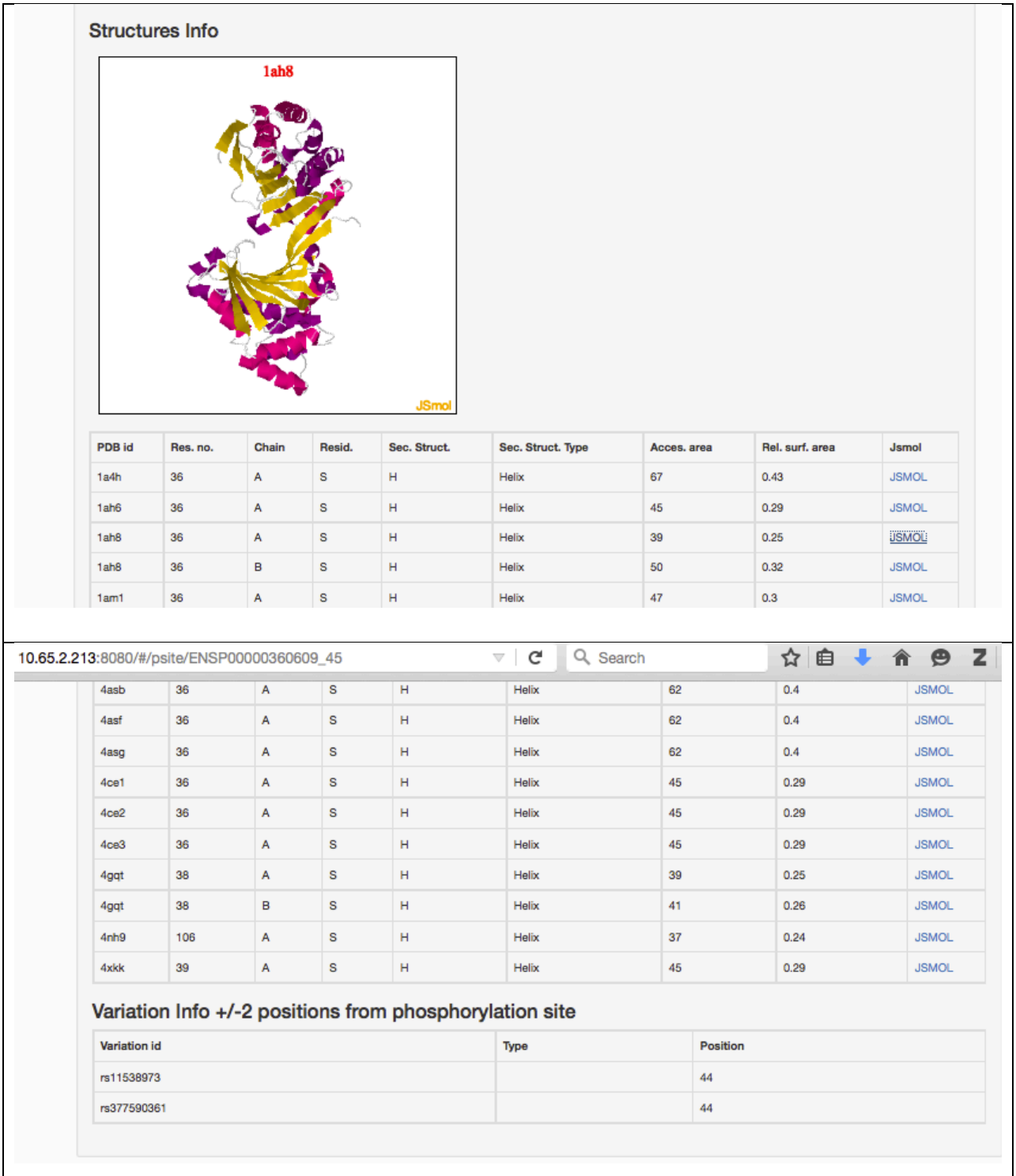



Figure 6. The information page, regarding a selected p-site of a phosphoprotein. Here, information is displayed regarding the experiments where the p-site was identified as well as the 3D crystal structures that are available for this protein and the SNPs that are found in the

vicinity.

Structural information is displayed in 3D using the JSmol application. JSmol is a javascript library used to enable viewing of JMOL java applets. JMOL is a Java open source application for displaying 3D chemical structures.

## Conclusions – Future perspectives

A neo4J database schema was developed to store all relevant information concerning phosphoproteins and phosphorylation sites, the expreriments that identified them, the 3D structures of the phosphoproteins and the SNPs that are found in the vicinity of the p-sites. This database was populated with raw data in 2015, for the needs of the FAB-PHOS project and is functional. Therefore, in order to make this tool useful to the community, there is a need to re-populate the database with recent data, concerning phosphoproteomic experiments, crystal structures and SNPs. In addition, the viewing tool of the crystal structures needs to be adjusted, so as to display the position of the phosphorylation sites. Finally, the database needs to be updated with other types of post-translational modifications, such as methylation sites. The above mentioned updates will allow this database to be published in a peer-reviewed journal in the near future. In its current form, the database is accessible at:

http://bioinf.bio.uth.gr/phospho-prometheus-db/

# Bibliography

Angles, R., and C. Gutierrez, 2008 Survey of graph database models. ACM Computing Surveys 40: 1–39.

Hornbeck, P. V., B. Zhang, B. Murray, J. M. Kornhauser, V. Latham *et al.*, 2015 PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. 43: D512-520.

kumar Kaliyar, R., 2015 Graph databases: A survey, pp. 785–790 in *International Conference on Computing, Communication & Automation*, IEEE, Greater Noida, India.

Kolomičenko, V., M. Svoboda, and I. H. Mlýnková, 2013 Experimental Comparison of Graph Databases, pp. 115–124 in *Proceedings of International Conference on Information Integration and Web-based Applications & Services - IIWAS '13*, ACM Press, Vienna, Austria.

https://neo4j.com/

https://spring.io/

https://angular.io/

https://www.javascript.com/

https://www.java.com/en/

# Chapter 7

## Contributions to other related Bioinformatics projects

This is a brief description of my contributions to another three relevant peer-reviewed publications of Dr. Amoutzias research team, while I was a Ph.D student in his laboratory.

**Title:** Low complexity regions in the proteins of prokaryotes perform important functional roles and are highly conserved.
Ntountoumi C, Vlastaridis P, Mossialos D, Stathopoulos C, Iliopoulos I, Promponas V, Oliver SG, Amoutzias GD.
Nucleic Acids Res. 2019 Nov 4;47(19):9998-10009. doi: 10.1093/nar/gkz730. PMID: 31504783

**Abstract of the publication:** “We provide the first high-throughput analysis of the properties and functional role of Low Complexity Regions (LCRs) in more than 1500 prokaryotic and phage proteomes. We observe that, contrary to a widespread belief based on older and sparse data, LCRs actually have a significant, persistent and highly conserved presence and role in many and diverse prokaryotes. Their specific amino acid content is linked to proteins with certain molecular functions, such as the binding of RNA, DNA, metal-ions and polysaccharides. In addition, LCRs have been repeatedly identified in very ancient, and usually highly expressed proteins of the translation machinery. At last, based on the amino acid content enriched in certain categories, we have developed a neural network web server to identify LCRs and accurately predict whether they can bind nucleic acids, metal-ions or are involved in chaperone functions. An evaluation of the tool showed that it is highly accurate for eukaryotic proteins as well.”

**Contribution:** My contribution in this publication (Ntountoumi *et al.* 2019) was the development of the web-server, named LCR hound, that accepts proteins in FASTA format, scans them for low-complexity regions and then graphically displays the results (see figure 1). The web server, was developed based on the Jhipster Application Framework, which utilizes the Java language and Spring Framework for the back-end and Angular Javascript Framework for the front-end. LCR hound is freely available at: http://bioinf.bio.uth.gr/lcr/

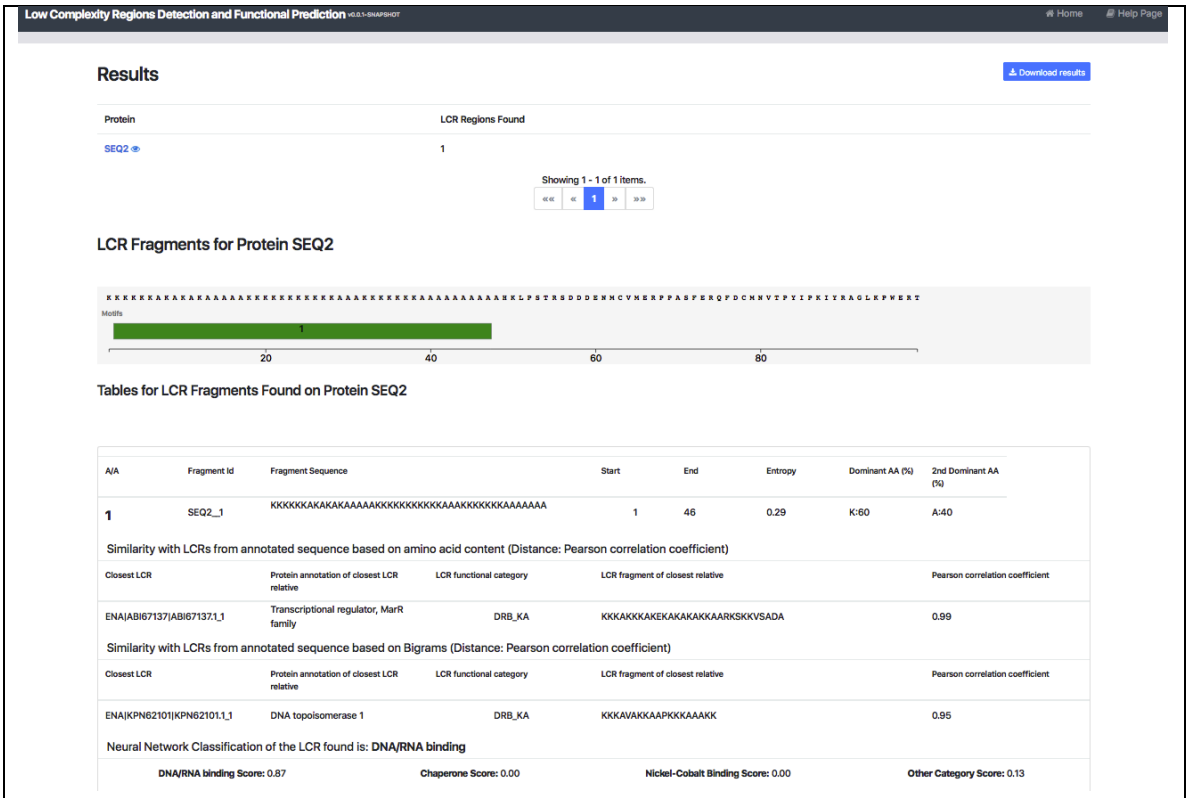



Figure 1. Graphical display of predicted low-complexity regions, by our published server (Ntountoumi *et al.* 2019).

**Title:** NAT/NCS2-hound: a webserver for the detection and evolutionary classification of prokaryotic and eukaryotic nucleobase-cation symporters of the NAT/NCS2 family.

Chaliotis A, Vlastaridis P, Ntountoumi C, Botou M, Yalelis V, Lazou P, Tatsaki E, Mossialos D, Frillingos S, Amoutzias GD.

Gigascience. 2018 Dec 1;7(12):giy133. doi: 10.1093/gigascience/giy133. PMID: 30418564

**Abstract of the publication: "**Nucleobase transporters are important for supplying the cell with purines and/or pyrimidines, for controlling the intracellular pool of nucleotides, and for obtaining exogenous nitrogen/carbon sources for metabolism. Nucleobase transporters are also evaluated as potential targets for antimicrobial therapies, since several pathogenic microorganisms rely on purine/pyrimidine salvage from their hosts. The majority of known nucleobase transporters belong to the evolutionarily conserved and ubiquitous nucleobase-ascorbate transporter/nucleobase-cation symporter-2 (NAT/NCS2) protein family. Based on a large-scale phylogenetic analysis that we performed on thousands of prokaryotic proteomes, we developed a webserver that can detect and distinguish this family of transporters from other homologous families that recognize different substrates. We can further categorize these transporters to certain evolutionary groups with distinct substrate preferences. The webserver scans whole proteomes and graphically displays which proteins are identified as NAT/NCS2, to which evolutionary groups and subgroups they belong to, and which conserved motifs they have. For key subgroups and motifs, the server displays annotated information from published crystal-structures and mutational studies pointing to key functional amino acids that may help experts assess the transport capability of the target sequences. The server is 100% accurate in detecting NAT/NCS2 family members. We also used the server to analyze 9,109 prokaryotic proteomes and identified Clostridia, Bacilli, β- and γ-Proteobacteria, Actinobacteria, and Fusobacteria as the taxa with the largest number of

NAT/NCS2 transporters per proteome. An analysis of 120 representative eukaryotic proteomes also demonstrates the server's capability of correctly analyzing this major lineage, with plants emerging as the group with the highest number of NAT/NCS2 members per proteome."

**Contribution:** My contribution in this publication (Chaliotis *et al.* 2018) was the development of a web-server, named NAT/NCS2-hound, that accepts proteins in FASTA format, scans them for NAT/NSC2 transporters, classifies them in pre-defined evolutionary groups and then graphically displays the results. Figure 2 shows how the server displays the results. All of the developed HMMs and MEME motifs were incorporated into the webserver. The webserver is based on the Jhipster Application Framework that utilizes Angular Javascript Framework for the front end and the Java language and Spring Framework for the back end. The server is freely available at:
http://bioinf.bio.uth.gr/nat-ncs2/
The server and instructions for local installation are found in the Supplementary Material "Server_for_local_installation" of the publication. Also, the server is registered at SciCrunch.org with RRID:SCR_016473.

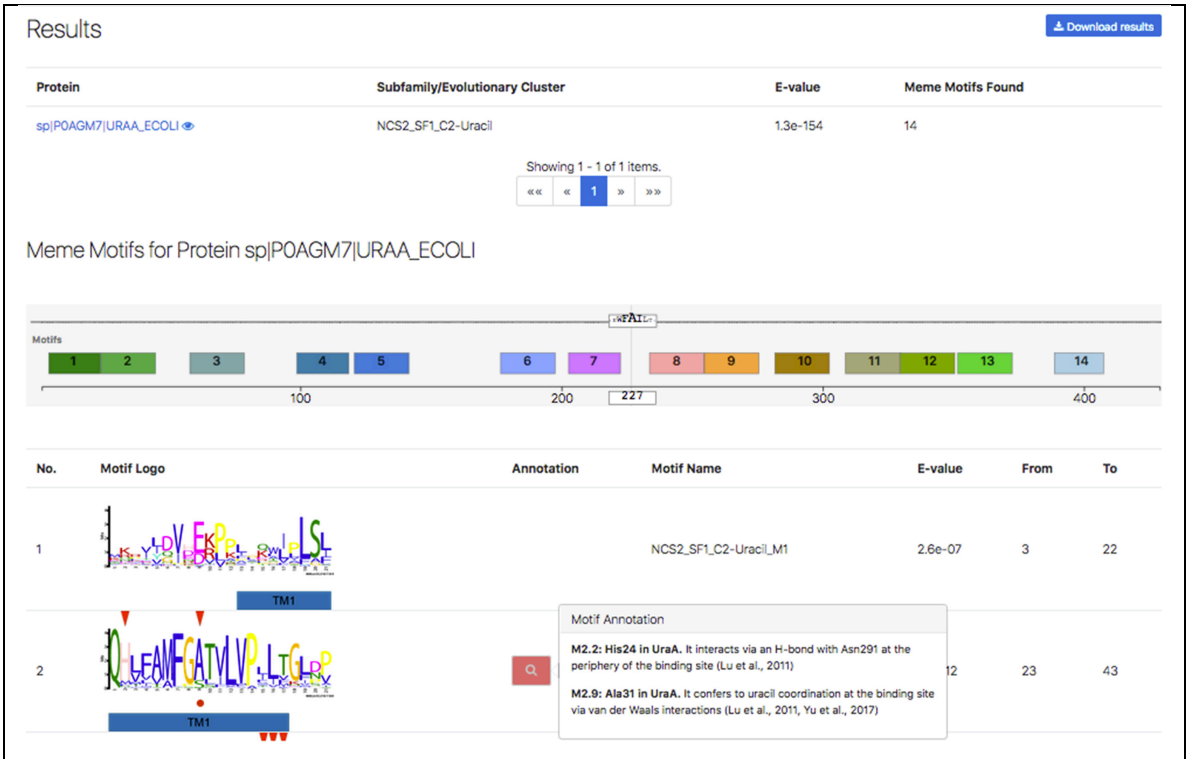



Figure 2. Display of results from the NAT/NCS2-hound server, including the best HMM that detects the protein sequence, the various MEME conserved motifs, and any available functional information/annotation for specific sites in certain motifs (Chaliotis *et al.* 2018).

**Title:** The complex evolutionary history of aminoacyl-tRNA synthetases.

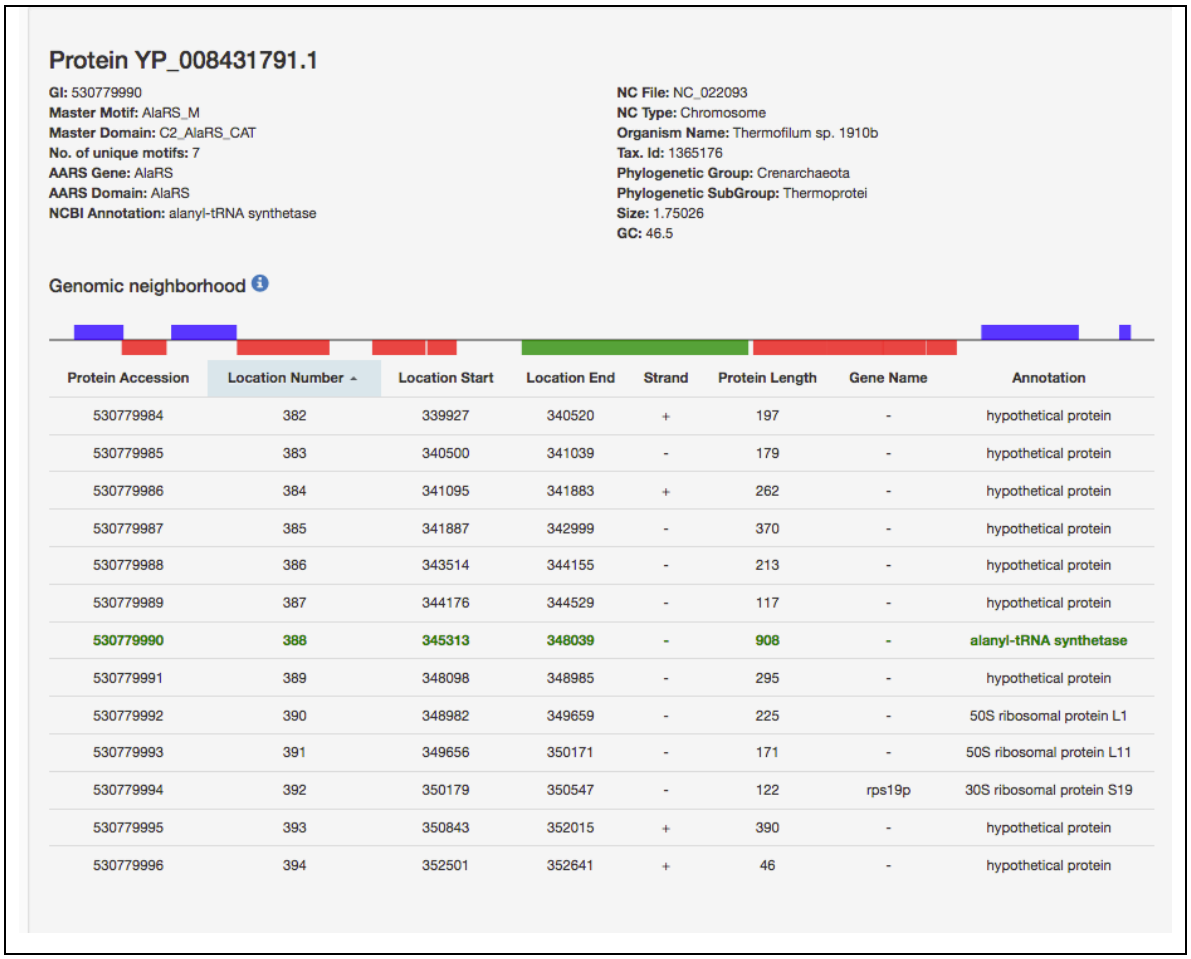
Chaliotis A, Vlastaridis P, Mossialos D, Ibba M, Becker HD, Stathopoulos C, Amoutzias GD.

Nucleic Acids Res. 2017 Feb 17;45(3):1059-1068. doi: 10.1093/nar/gkw1182. PMID: 28180287

**Abstract of the publication: "**Aminoacyl-tRNA synthetases (AARSs) are a superfamily of enzymes responsible for the faithful translation of the genetic code and have lately become a prominent target for synthetic biologists. Our large-scale analysis of >2500 prokaryotic genomes reveals the complex evolutionary history of these enzymes and their paralogs, in which horizontal gene transfer played an important role. These results show that a widespread belief in the evolutionary stability of this superfamily is misconceived. Although AlaRS, GlyRS, LeuRS, IleRS, ValRS are the most stable members of the family, GluRS, LysRS and CysRS often have paralogs, whereas AsnRS, GlnRS, PylRS and SepRS are often absent from many genomes. In the course of this analysis, highly conserved protein motifs and domains within each of the AARS loci were identified and used to build a web-based computational tool for the genome-wide detection of AARS coding sequences. This is based on hidden Markov models (HMMs) and is available together with a cognate database that may be used for specific analyses. The bioinformatics tools that we have developed may also help to identify new antibiotic agents and targets using these essential enzymes. These tools also may help to identify organisms with alternative pathways that are involved in maintaining the fidelity of the genetic code."

**Contribution:** My contribution in this publication (Chaliotis *et al.* 2017) was the development of the AARS database and accompanying web-server that accepts proteins in FASTA format, scans them for aminoacyl-tRNA synthetases and then graphically displays the results (see figure 3). Data organization and storage was implemented in a MySQL database. A web graphical interface was generated

with Java Language and Spring Framework for the back-end and Angular JS Framework for the front-end that is developed in a single-page application format. Front-end and back-end communication is established through an authenticated RESTful API. The database and web-server are available at: http://bioinf.bio.uth.gr/aars/



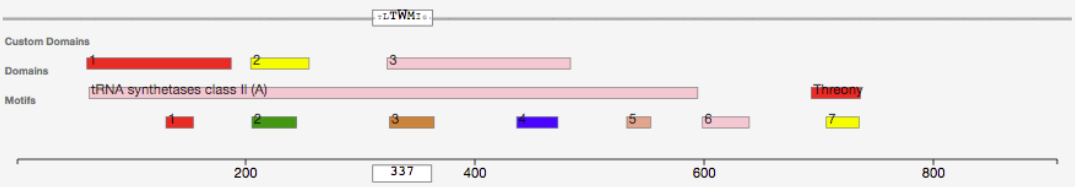
**Protein YP_008431791.1**

**GI:** 530779990
**Master Motif:** AlaRS_M
**Master Domain:** C2_AlaRS_CAT
**No. of unique motifs:** 7
**AARS Gene:** AlaRS
**AARS Domain:** AlaRS
**NCBI Annotation:** alanyl-tRNA synthetase

**NC File:** NC_022093
**NC Type:** Chromosome
**Organism Name:** Thermofilum sp. 1910b
**Tax. Id:** 1365176
**Phylogenetic Group:** Crenarchaeota
**Phylogenetic SubGroup:** Thermoprotei
**Size:** 1.75026
**GC:** 46.5

Genomic neighborhood

| Protein Accession | Location Number | Location Start | Location End | Strand | Protein Length | Gene Name | Annotation |
|---|---|---|---|---|---|---|---|
| 530779984 | 382 | 339927 | 340520 | + | 197 | - | hypothetical protein |
| 530779985 | 383 | 340500 | 341039 | - | 179 | - | hypothetical protein |
| 530779986 | 384 | 341095 | 341883 | + | 262 | - | hypothetical protein |
| 530779987 | 385 | 341887 | 342999 | - | 370 | - | hypothetical protein |
| 530779988 | 386 | 343514 | 344155 | - | 213 | - | hypothetical protein |
| 530779989 | 387 | 344176 | 344529 | - | 117 | - | hypothetical protein |
| **530779990** | **388** | **345313** | **348039** | - | **908** | - | **alanyl-tRNA synthetase** |
| 530779991 | 389 | 348098 | 348985 | - | 295 | - | hypothetical protein |
| 530779992 | 390 | 348982 | 349659 | - | 225 | - | 50S ribosomal protein L1 |
| 530779993 | 391 | 349656 | 350171 | - | 171 | - | 50S ribosomal protein L11 |
| 530779994 | 392 | 350179 | 350547 | - | 122 | rps19p | 30S ribosomal protein S19 |
| 530779995 | 393 | 350843 | 352015 | + | 390 | - | hypothetical protein |
| 530779996 | 394 | 352501 | 352641 | + | 46 | - | hypothetical protein |




**Motifs**

| Motif Name | Motif Number | Score | Hmm Start | Hmm End | Ali Start | Ali End | E-value |
|---|---|---|---|---|---|---|---|
| AlaRS_M2_cat | 1 | 47.7 | 1 | 24 | 130 | 153 | 1.8e-14 |
| AlaRS_M4_cat | 2 | 52.8 | 3 | 41 | 205 | 243 | 5.3e-16 |
| AlaRS_M5 | 3 | 65.9 | 2 | 40 | 325 | 363 | 3.9e-20 |
| AlaRS_M6 | 4 | 41 | 1 | 36 | 436 | 471 | 2.7e-12 |
| AlaRS_M7 | 5 | 34.3 | 1 | 21 | 532 | 552 | 1.8e-10 |
| AlaRS_M8 | 6 | 71.2 | 1 | 41 | 598 | 638 | 9e-22 |
| AlaRS_M10 | 7 | 46.3 | 1 | 29 | 706 | 734 | 6.7e-14 |

**Custom domains**

| Motif Name | Motif Number | Score | Hmm Start | Hmm End | Ali Start | Ali End | E-value |
|---|---|---|---|---|---|---|---|
| C2_AlaRS_CAT | 1 | 127 | 3 | 122 | 61 | 186 | 8.5e-54 |
| C2_AlaRS_CAT | 2 | 45 | 190 | 240 | 204 | 254 | 8.5e-54 |
| C2_AlaRS_ACB | 3 | 141.7 | 28 | 185 | 323 | 482 | 1.6e-44 |

**Pfam Domains**

| Model | Description | Model Accession | Model Length | Protein Length | Model Hit | E-value | Score | Hmm Start | Hmm End | Ali Start | Ali End |
|---|---|---|---|---|---|---|---|---|---|---|---|
| tRNA-synt_2c | tRNA synthetases class II (A) | PF01411.16 | 552 | 908 | 1 | 1.6e-130 | 436.1 | 1 | 551 | 63 | 593 |
| tRNA_SAD | Threonyl and Alanyl tRNA synthetase second additional domain | PF07973.11 | 44 | 908 | 1 | 8.8e-14 | 51.2 | 1 | 43 | 693 | 735 |

Figure 3. Graphical display of the results from a query in our AARS database and web-server (Chaliotis *et al.* 2017).

# Bibliography

Chaliotis, A., P. Vlastaridis, D. Mossialos, M. Ibba, H. D. Becker *et al.*, 2017 The complex evolutionary history of aminoacyl-tRNA synthetases. Nucleic Acids Res. 45: 1059–1068.

Chaliotis, A., P. Vlastaridis, C. Ntountoumi, M. Botou, V. Yalelis *et al.*, 2018 NAT/NCS2-hound: a webserver for the detection and evolutionary classification of prokaryotic and eukaryotic nucleobase-cation symporters of the NAT/NCS2 family. Gigascience 7:.

Ntountoumi, C., P. Vlastaridis, D. Mossialos, C. Stathopoulos, I. Iliopoulos *et al.*, 2019 Low complexity regions in the proteins of prokaryotes perform important functional roles and are highly conserved. Nucleic Acids Res. 47: 9998–10009.

# Conclusions – Future Perspectives

The goal of this doctoral thesis was to investigate the impact of the new high-throughput technologies and their generated data, on understanding the role of post-translational regulation. The focus was on protein phosphorylation, which is the most abundant post-translational modification.

At the beginning of the thesis, a review was prepared and published as a conference paper, that discussed the importance of protein phosphorylation, the challenges of the phosphoproteomic technologies and their underlying biases, how to filter these data and how to extract biological knowledge (Vlastaridis *et al.* 2016).

During this thesis, bioinformatics tools were developed to mine, organize and store the large volume of publicly available data. Next, these data were analyzed with a plethora of bioinformatics methods in order to extract biological knowledge. More specifically, a locally installed annotation tool was developed to help a team of annotators of the FAB-PHOS project to collect, store and organize experimental data on post-translational modifications of proteins (phosphorylations, methylations) from scientific publications in various journals. With this custom-designed tool, many annotators can simultaneously store and categorize the publications found with various tags, such as what type of experiments were performed, what organism and what type of tissue the experimental data came from and others. They can also archive the publication manuscript and the supplementary files on a local server, for later review by themselves or their colleagues. In addition, a Solr search engine was integrated for more efficient searching in the huge amount of data stored from the annotators. The tool may be downloaded from the Bioinformatics laboratory website, at:
http://bioinf.bio.uth.gr/ptm-at.html

This tool allowed our team of annotators to successfully mine more than 1000 publicly available research papers from Pubmed. The details of the development and characteristics of this computational tool are found in Chapter 2.

Next, thanks to this developed annotator tool, a compendium of 150,000 phosphorylation sites from human, mouse, yeast and *Arabidopsis* was compiled. This compendium was based on 187 filtered high-throughput phosphoproteomic datasets and on two low-throughput compendia from the PhosphoGrid and Phosphosite+ databases. Based on this compendium, two different methods were applied to estimate the total number of phosphoproteins and phosphorylation sites in each of the four model organisms, Capture-Recapture, and fitting the saturation curve of cumulative redundant vs. cumulative non-redundant phosphoproteins/p-sites. The estimates were also adjusted for different levels of noise in the underlying data and for other confounding factors. The estimates for the first three model organisms were considered reliable, whereas for *Arabidopsis*, they were not considered reliable, due to the limited underlying raw data for this species. Thus, 13,000, 11,000 and 3,000 phosphoproteins and 230,000, 156,000 and 40,000 p-sites are expected to exist in human, mouse and yeast, respectively. These analyses revealed that most of the phosphoproteins have already been discovered for human, mouse and yeast, whereas this is not the case for their phosphorylation sites. All the details of the mining, filtering and analysis of the data are described in chapter 3 and in (Vlastaridis *et al.* 2017a), in *Gigascience*, under the Creative Commons CC BY license, with Pubmed ID: 28327990.

Based on the yeast compendium that was compiled in Chapter 3, an in-depth bioinformatics analysis was conducted, in order to assess the role of protein phosphorylation in the yeast central metabolism. Towards this aim, many other types of omic data from yeast were integrated as well. The ultimate goal was to identify phosphorylation sites that may regulate important enzymes and pathways with biotechnological applications. Indeed, this bioinformatics study

clearly demonstrated the pivotal role of protein phosphorylation, since half of the enzymes of the central metabolism are phosphorylated. In addition, important enzymes that are more abundant, regulate more reactions, and have more protein-protein interactions tend to be regulated by phosphorylation more frequently than the other enzymes. Furthermore, these analyses helped to prioritize thousands of p-sites in terms of their potential phenotypic impact. Thus, this study constitutes a catalogue of p-sites that need to be explored experimentally, in the future, for identifying key molecular switches/rheostats with potential biotechnological and even medical applications. All the details of the study are described in chapter 4 and in (Vlastaridis *et al.* 2017b), in *Genes, Genomes, Genetics,* under the Creative Commons CC BY license, with Pubmed ID: 28250014.

Next, in chapter 5, the phosphorylation compendium that was compiled in chapter 3 was integrated with another protein methylation compendium that was compiled by other members of our FAB-PHOS team, in order to train two Neural Networks (NNs) that predict phosphorylation and methylation sites in proteins. I trained the phosphorylation NN. The results of the two NNs are further integrated in order to predict meth/phos switches and clusters as well. The protein methylation NN has an accuracy of 86%, whereas the protein phosphorylation NN has an accuracy of 84%, thus significantly outperforming many other published prediction tools. These tools have been integrated into a webserver named Meth-Phos-Prometheus available at http://bioinf.bio.uth.gr/meth-phos-prometheus/. The goal is to publish this webserver in a peer-reviewed journal, such as *Bioinformatics*.

Another goal of this thesis was to develop a graph database to store and organize the filtered phosphorylation sites, together with other types of data, such domains, SNPs, structural information. Thus these data would become available not only to other team members, but to the scientific community as well. This database was developed based in Neo4J. The details of the database schema,

the webserver, and how are the results organized and displayed are described in detail in chapter 6. Furthermore, the database is accessible at: http://bioinf.bio.uth.gr/phospho-prometheus-db/

Finally, the computational expertise that I developed during this thesis allowed me to develop a database and three webservers for three more relevant publications of the Bioinformatics group led by Dr. Amoutzias, concerning i) the evolution of t-RNA synthetases, (Chaliotis *et al.* 2017) in *Nucleic Acids Research*, ii) the evolution of NAT/NCS2 transporters (Chaliotis *et al.* 2018), in *Gigascience* and iii) the role of low-complexity regions in prokaryotes (Ntountoumi *et al.* 2019), in *Nucleic Acids Research*.

## Bibliography

Chaliotis, A., P. Vlastaridis, D. Mossialos, M. Ibba, H. D. Becker *et al.*, 2017 The complex evolutionary history of aminoacyl-tRNA synthetases. Nucleic Acids Res. 45: 1059–1068.

Chaliotis, A., P. Vlastaridis, C. Ntountoumi, M. Botou, V. Yalelis *et al.*, 2018 NAT/NCS2-hound: a webserver for the detection and evolutionary classification of prokaryotic and eukaryotic nucleobase-cation symporters of the NAT/NCS2 family. Gigascience 7:.

Ntountoumi, C., P. Vlastaridis, D. Mossialos, C. Stathopoulos, I. Iliopoulos *et al.*, 2019 Low complexity regions in the proteins of prokaryotes perform important functional roles and are highly conserved. Nucleic Acids Res. 47: 9998–10009.

Vlastaridis, P., P. Kyriakidou, A. Chaliotis, Y. Van de Peer, S. G. Oliver *et al.*, 2017a Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes. Gigascience 6: 1–11.

Vlastaridis, P., S. G. Oliver, Y. Van de Peer, and G. D. Amoutzias, 2016 The Challenges of Interpreting Phosphoproteomics Data: A Critical View Through the Bioinformatics Lens, pp. 196–204 in *Computational Intelligence Methods for Bioinformatics and Biostatistics*, edited by C. Angelini, P. M. Rancoita, and S. Rovetta. Springer International Publishing, Cham.

Vlastaridis, P., A. Papakyriakou, A. Chaliotis, E. Stratikos, S. G. Oliver *et al.*, 2017b The Pivotal Role of Protein Phosphorylation in the Control of Yeast Central Metabolism. G3 (Bethesda) 7: 1239–1249.