Université de Montréal

Identification de nouveaux substrats des kinases Erk1/2 par une approche bio-informatique, pharmacologique et phosphoprotéomique

par
Mathieu Courcelles

Département de Biochimie
Faculté de Médecine

Thèse présentée à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de Doctorat
en Bio-informatique

Décembre, 2011

Université de Montréal
Faculté des études supérieures et postdoctorales

Cette thèse intitulée:

Identification de nouveaux substrats des kinases Erk1/2 par une approche bio-informatique, pharmacologique et phosphoprotéomique

Présenté par:
Mathieu Courcelles

a été évaluée par un jury composé des personnes suivantes:

Nicolas Lartillot, président-rapporteur
Pierre Thibault, directeur de recherche
Sébastien Lemieux, co-directeur
Vincent Archambault, membre du jury
Anne-Claude Gingras, examinateur externe
Isabelle Royal, représentant du doyen de la FES

Résumé

La phosphorylation est une modification post-traductionnelle omniprésente des protéines Cette modification est ajoutée et enlevée par l'activité enzymatique respective des protéines kinases et phosphatases. Les kinases Erk1/2 sont au cœur d'une voie de signalisation importante qui régule l'activité de protéines impliquées dans la traduction, le cycle cellulaire, le réarrangement du cytosquelette et la transcription. Ces kinases sont aussi impliquées dans le développement de l'organisme, le métabolisme du glucose, la réponse immunitaire et la mémoire. Différentes pathologies humaines comme le diabète, les maladies cardiovasculaires et principalement le cancer, sont associées à une perturbation de la phosphorylation sur les différents acteurs de cette voie. Considérant l'importance biologique et clinique de ces deux kinases, connaître l'étendue de leur activité enzymatique pourrait mener au développement de nouvelles thérapies pharmacologiques.

Dans ce contexte, l'objectif principal de cette thèse était de mesurer l'influence de cette voie sur le phosphoprotéome et de découvrir de nouveaux substrats des kinases Erk1/2. Une étude phosphoprotéomique de cinétique d'inhibition pharmacologique de la voie de signalisation Erk1/2 a alors été entreprise. Le succès de cette étude était basé sur trois technologies clés, soit l'enrichissement des phosphopeptides avec le dioxyde de titane, la spectrométrie de masse haut débit et haute résolution, et le développement d'une plateforme bio-informatique nommée ProteoConnections. Cette plateforme permet d'organiser les données de protéomique, évaluer leur qualité, indiquer les changements d'abondance et accélérer l'interprétation des données. Une fonctionnalité distinctive de ProteoConnections est l'annotation des sites phosphorylés identifiés (kinases, domaines, structures, conservation, interactions protéiques phospho-dépendantes). Ces informations ont été essentielles à l'analyse des 9615 sites phosphorylés sur les 2108 protéines identifiées dans cette étude, soit le plus large ensemble rapporté chez le rat jusqu'à ce jour. L'analyse des domaines protéiques a révélé que les domaines impliqués dans les interactions avec les protéines, les acides nucléiques et les autres molécules sont les plus

fréquemment phosphorylés et que les sites sont stratégiquement localisés pour affecter les interactions.

Un algorithme a été implémenté pour trouver les substrats potentiels des kinases Erk1/2 à partir des sites identifiés selon leur motif de phosphorylation, leur cinétique de stimulation au sérum et l'inhibition pharmacologique de Mek1/2. Une liste de 157 substrats potentiels des kinases Erk1/2 a ainsi été obtenue. Parmi les substrats identifiés, douze ont déjà été rapportés et plusieurs autres ont des fonctions associées aux substrats déjà connus. Six substrats (Ddx47, Hmg20a, Junb, Map2k2, Numa1, Rras2) ont été confirmés par un essai kinase *in vitro* avec Erk1. Nos expériences d'immunofluorescence ont démontré que la phosphorylation de Hmg20a sur la sérine 105 par Erk1/2 affecte la localisation nucléocytoplasmique de cette protéine.

Finalement, les phosphopeptides isomériques positionnels, soit des peptides avec la même séquence d'acides aminés mais phosphorylés à différentes positions, ont été étudiés avec deux nouveaux algorithmes. Cette étude a permis de déterminer leur fréquence dans un extrait enrichi en phosphopeptides et d'évaluer leur séparation par chromatographie liquide en phase inverse. Une stratégie analytique employant un des algorithmes a été développée pour réaliser une analyse de spectrométrie de masse ciblée afin de découvrir les isomères ayant été manqués par la méthode d'analyse conventionnelle.

Mots-clés: Base de données biologiques, Bio-informatique, Erk, Kinase, Phosphoprotéomique, Phosphorylation, Protéomique quantitative, Signalisation cellulaire, Spectrométrie de masse

Abstract

Phosphorylation is an omnipresent post-translational modification of proteins that regulates numerous cellular processes. This modification is controlled by the enzymatic activity of protein kinases and phosphatases. Erk1/2 kinases are central to an important signaling pathway that modulates translation, cell cycle, cytoskeleton rearrangement and transcription. They are also implicated in organism development, glucose metabolism, immune response and memory. Different human pathologies such as diabetes, cardiovascular diseases, and most importantly cancer, are associated with misregulation or mutations in members of this pathway. Considering the biological and clinical importance of those two kinases, discovering the extent of their enzymatic activity could favor the development of new pharmacological therapies.

In this context, the principal objective of this thesis was to measure the influence of this pathway on the phosphoproteome and to discover new substrates of the Erk1/2 kinases. A phosphoproteomics study on the pharmacological inhibition kinetics of the Erk1/2 signaling pathway was initiated. The success of this study was based on three key technologies such as phosphopeptides enrichment with titanium dioxide, high-throughput and high-resolution mass spectrometry, and the development of ProteoConnections, a bioinformatics analysis platform. This platform is dedicated to organize proteomics data, evaluate data quality, report changes of abundance and accelerate data interpretation. A distinctive functionality of ProteoConnections is the annotation of phosphorylated sites (kinases, domains, structures, conservation, phospho-dependant protein interactions, etc.). This information was essential for the dataset analysis of 9615 phosphorylated sites identified on 2108 proteins during the study, which is, until now, the largest one reported for rat. Protein domain analysis revealed that domains implicated in proteins, nucleic acids and other molecules binding were the most frequently phosphorylated and that these sites are strategically located to affect the interactions.

An algorithm was implemented to find Erk1/2 kinases potential substrates of identified sites using their phosphorylation motif, serum stimulation and Mek1/2 inhibition kinetic profile. A list of 157 potential Erk1/2 substrates was obtained. Twelve of them were previously reported and many more have functions associated to known substrates. Six substrates (Ddx47, Hmg20a, Junb, Map2k2, Numa1, and Rras2) were confirmed by *in vitro* kinase assays with Erk1. Our immunofluorescence experiments demonstrated that the phosphorylation of Hmg20a on serine 105 by Erk1/2 affects the nucleocytoplasmic localization of this protein.

Finally, phosphopeptides positional isomers, peptides with the same amino acids sequence but phosphorylated at different positions, were studied with two new algorithms. This study allowed us to determine their frequency in an enriched phosphopeptide extract and to evaluate their separation by reverse-phase liquid chromatography. An analytical strategy that uses one of the algorithms was developed to do a targeted mass spectrometry analysis to discover the isomers that had been missed by the conventional method.

Keywords: Bioinformatics, Biological database, Erk, Kinase, Mass spectrometry, Phosphoproteomics, Phosphorylation, Quantitative proteomics, Signaling pathway

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

ACN	Acétonitrile
ADN	Acide désoxyribonucléique
AGC	*Automatic gain control*
ARN	Acide ribonucléique
ARNm	Acide ribonucléique messager
ATP	Adénosine triphosphate
CID	*Collision induced dissociation*
CSV	*Comma-separated values*
Da	Dalton
DBMS	*Database management system*
DDA	*Data-dependent acquisition*
DTT	Dithiothréitol
Erk	*Extracellular signal-regulated kinase*
ETD	*Electron transfer dissociation*
FA	*Formic acid*
FBS	*Fetal bovine serum*
FDR	*False discovery rate*
GST	Glutathione S-transférase
GO	*Gene Ontology*
GTP	Guanosine triphosphate
HCD	*Higher-energy C-trap dissociation*
HLB	*Hydrophilic-lipophilic balanced*
HPLC	*High performance liquid chromatography*
HTML	*HyperText Markup Language*
i.d.	*Internal diameter*
IEC-6	*Intestinal epithelial cell line, No. 6*

IMAC	*Immobilized metal ion affinity chromatography*
IPI	*International Protein Index*
LC-MS/MS	Chromatographie liquide couplée à la spectrométrie de masse en tandem
LTQ	*Linear trap quadrupole*
MAP kinase	*Mitogen-activated protein kinase*
Mek	*Map kinase/Erk kinase*
MGF	*Mascot generic file*
MS	Spectrométrie de masse
MS/MS	Spectrométrie de masse en tandem
m/z	Ratio masse sur charge
NES	*Nuclear export signal*
NLS	*Nuclear localization signal*
PAC	*Phosphoramidate chemistry*
PDB	*Protein DataBank*
PHP	*PHP: Hypertext Preprocessor*
pS	Phosphosérine
PTM	*Post-translational modification*
pT	Phosphothréonine
pY	Phosphotyrosine
RP-HPLC	*Reversed phase high-performance liquid chromatography*
rRNA	*Ribosomal ribonucleic acid*
SCX	*Strong cation exchange*
SQL	*Structured query language*
TFA	*Trifluoroacetic acid*
Tr	Temps de rétention
XML	*Extensible markup language*

À mes parents

Remerciements

L'aboutissement de cette thèse a été possible grâce aux nombreuses personnes avec lesquelles j'ai appris, discuté et collaboré tout au long de mon long parcours. Je voudrais d'abord remercier mon directeur de recherche Pierre Thibault pour l'opportunité d'entreprendre mon doctorat dans son laboratoire. Je lui suis très reconnaissant pour la liberté qu'il m'a accordée dans la poursuite de mes recherches sur le plan bio-informatique et pour la confiance qu'il m'a accordé au niveau du travail en laboratoire. Cela a été une occasion unique d'apprendre plusieurs aspects de la protéomique.

Je remercie mon co-directeur Sébastien Lemieux qui m'a éclairé sur plusieurs questions algorithmiques, bio-informatiques et statistiques. Pour l'aspect informatique, j'aimerais noter la contribution des programmeurs Gagandeep Jaitly, Kevin Eng et Olivier Caron-Lizotte qui ont développé le logiciel pour l'analyse quantitative de ce projet. Merci aussi à Dannis Jolicoeur et Patrick Gendron pour leur assistance technique.

Au niveau du travail en laboratoire, un merci spécial à Maria Marcantonio et Matthias Trost pour m'avoir enseigné les rudiments de la phosphoprotéomique et Gaëlle Bridon pour nos nombreuses discussions sur ce sujet et notre collaboration pour une partie de cette thèse. Je remercie aussi les employés Christelle Pomiès, pour son assistance technique dans la préparation d'échantillons, et Éric Bonneil pour son aide inestimable afin de régler les pépins de chromatographie et de spectrométrie de masse. Merci aussi à tous les autres membres du laboratoire Thibault pour leurs nombreuses suggestions constructives sur mon travail et ainsi que pour leur agréable compagnie tout au long de cette aventure. Un merci particulier à mes collaborateurs Laure Voisin, Christophe Frémin et Catherine Julien du laboratoire de Sylvain Meloche pour leur apport à nos découvertes. Finalement, je voudrais remercier ma conjointe, ma famille et mes amis pour leurs précieux encouragements et soutient tout au long de ma thèse. À tous ces gens MERCI.

CHAPITRE 1: Introduction

1.1 La phosphoprotéomique

La protéomique est un large champ d'étude qui s'intéresse à l'ensemble des protéines (voir Annexe 1 pour la théorie élémentaire sur les protéines) contenues dans les cellules. L'ensemble des protéines présentes dans une cellule se nomme le protéome. Sa taille se situe entre 20 000 et 40 000 protéines chez l'humain selon les plus récentes estimations (selon les cadres ouverts de lecture prédits), mais est en réalité beaucoup plus vaste dû aux isoformes formés par l'épissage alternatif, les clivages enzymatiques et les modifications post-traductionnelles. La protéomique s'intéresse à l'identification des protéines exprimées, comprendre leur fonction, déterminer leur structure, l'interaction de celles-ci avec d'autres molécules (métabolites, ADN, ARN, protéines), leur localisation dans les compartiments cellulaires ainsi qu'à leurs niveaux d'expression dans divers tissus. Ceci est étudié à divers moments du développement de l'organisme ou dans différents environnements cellulaires. La protéomique s'intéresse aussi aux diverses modifications chimiques post-traductionnelles (acétylation, méthylation, glycosylation, phosphorylation, ubiquitination, etc.) qui régulent l'activité des protéines. La phosphoprotéomique est orientée vers l'étude de la phosphorylation. L'intérêt particulier pour ce domaine de recherche provient de l'importance biologique et clinique de cette modification (perturbation fréquente de la phosphorylation dans plusieurs maladies humaines dont le cancer). La phosphoprotéomique a engendré le développement de plusieurs méthodes analytiques spécifiques pour détecter la phosphorylation.

1.1.1 La phosphorylation

La phosphorylation est une modification covalente post-traductionnelle des sérines, thréonines et tyrosines qui consiste en l'ajout d'un groupement phosphate sur l'hydroxyle de la chaîne latérale (Figure 1.1). Elle peut également survenir sur les cystéines, les arginines, les lysines, les histidines, les acides aspartiques et glutamiques, mais ces phosphorylations sont présentement considérées comme moins fréquentes et moins importantes fonctionnellement chez les eucaryotes [1]. La phosphorylation est d'origine ancienne puisqu'on peut la retrouver chez les procaryotes. Sa présence est toutefois beaucoup plus marquée au niveau des eucaryotes puisqu'elle y joue un rôle régulateur très important. On estime qu'environ 30 à 50% du protéome est phosphorylé chez les eucaryotes [2]. La phosphorylation est réversible et étroitement régulée par des centaines de protéines kinases et phosphatases qui reconnaissent des motifs spécifiques d'acides aminés sur les protéines. En réponse à des stimuli extra/intra cellulaires, ces familles de protéines vont respectivement phosphoryler et déphosphoryler leurs substrats en quelques minutes ou secondes.

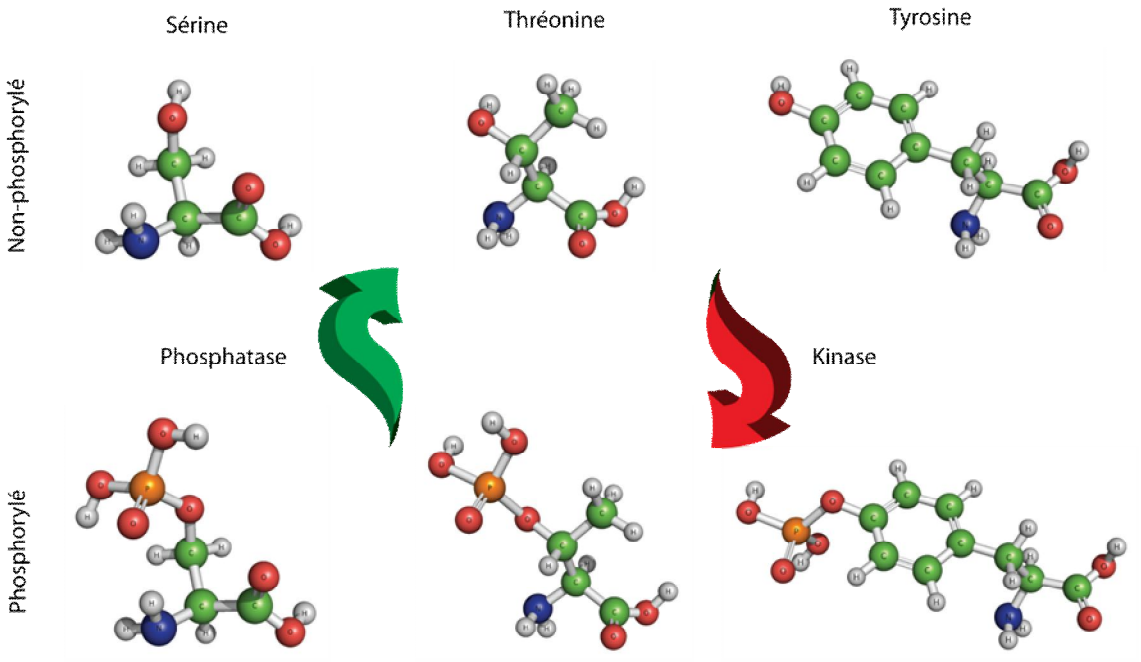


Figure 1.1: Structures des acides aminés phosphorylés et les familles de protéines impliquées dans le cycle de la phosphorylation.

La phosphorylation est une modification post-traductionnelle des sérines, thréonines et tyrosines. Cette ajout est catalysé par les protéines kinases et enlevé par les protéines phosphatases.

La phosphorylation est omniprésente dans plusieurs processus biologiques tels que la régulation du cycle cellulaire, la prolifération, la différentiation, la dégradation des protéines, etc. La régulation de processus biologiques par la phosphorylation peut nécessiter plusieurs évènements de phosphorylation successifs. Cette série d'interactions kinase-substrat forme une voie de signalisation. Au niveau moléculaire, la phosphorylation peut causé un changement conformationnel de la protéine en provoquant un encombrement stérique et en lui conférant une charge négative locale. Ces changements physicochimiques influencent l'activité de la protéine (Figure 1.2). Dans certain cas, ces changements activent ou désactivent totalement l'activité catalytique de la protéine. Ils régulent aussi les interactions protéine-protéine de façon positive en servant de point d'ancrage pour les domaines protéiques qui reconnaissent les résidus phosphorylés, ou négativement en

encombrant les interfaces d'interaction. Les interactions avec d'autres types de molécules comme les acides nucléiques sont aussi modulées par la phosphorylation. Dans une même région, plusieurs sites de phosphorylation peuvent agir coopérativement pour moduler l'activité ou les interactions des protéines. L'activité peut être régulée selon un mécanisme de type interrupteur (actif / inactif) ou graduel en réponse aux phosphorylations successives [3].

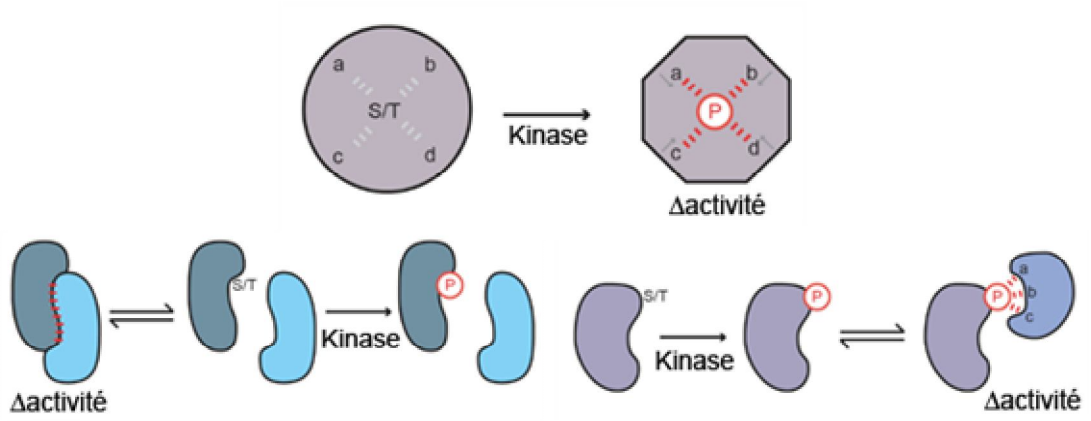


Figure 1.2: Régulation de l'activité des protéines par divers mécanismes dépendant de la phosphorylation.

Le groupement phosphate ajouté par la kinase peut interagir par liaisons ioniques avec les résidus voisins (a, b, c, d) dans la structure tertiaire de la protéine. Ces interactions peuvent engendrer des changements conformationnels qui influence l'activité des protéines. La phosphorylation module aussi les interactions protéine-protéine. Le groupement phosphate peut soit empêcher l'interaction par encombrement stérique ou la favoriser en formant des liaisons ioniques avec les résidus chargés de l'autre protéine. Adaptation de [4] avec la permission d'AASS: Science, copyright 2009

Tous les sites de phosphorylation détectés ne sont pas nécessairement fonctionnels. Autrement dit, certains pourraient n'avoir aucune influence notable sur l'activité de la protéine et sont le résultat d'une activité kinase hors cible. Cette hypothèse a été énoncée suite à l'observation de nombreux gains et pertes de phosphorylation au cours de

l'évolution [5]. Le phosphoprotéome évolue à un taux équivalent aux résidus non-phosphorylés (environ la même fréquence de substitutions du même acide aminé). La proportion de sites fonctionnels et non-fonctionnels est présentement inconnue étant donné le travail exigé pour comprendre l'effet de chaque phosphorylation. Toutefois, les sites fonctionnels connus semblent être plus conservés au cours de l'évolution. Il est possible que certaines pertes de phosphorylation soient compensées par le gain de nouveaux sites qui préservent la fonction. La conservation du phosphoprotéome est abordée dans la section 2.5.4.

1.1.1.1 Les protéines kinases

Les kinases forment une famille de protéines dont le rôle est d'effectuer la phosphorylation. Elles ont donc une activité phosphotransférase. Plus précisément, leur tâche est de transférer le groupement phosphate-γ provenant de l'ATP au groupement hydroxyle des sérines, thréonines et tyrosines d'une protéine appelée substrat. Les kinases peuvent aussi s'autophosphoryler. Structurellement, les kinases sont composées d'un domaine liant l'ATP riche en glycines qui contient une lysine, ainsi qu'un domaine catalytique caractérisé par un acide aspartique très conservé. Le domaine catalytique reconnait spécifiquement sur le substrat une région de quelques acides aminés avec le résidu à phosphoryler qui peut être acide (kinases CKI et CKII), basique (kinases PKA, PKC) ou riche en prolines (kinases MAPKs, CDKs). La région reconnue par une kinase est nommée séquence consensus de phosphorylation. D'autres domaines sur les kinases peuvent aussi contrôler les protéines qui seront des substrats. Les kinases représentent environ 2% des gènes dans les génomes eucaryotes. On estime à 518 le nombre de kinases chez l'homme, 540 chez la souris (dont 510 sont des orthologues humains), 251 chez la drosophile et 123 chez la levure [6]. Des transcrits épissés alternativement existent aussi pour 75% des kinases chez l'homme. Les kinases ont une spécificité pour les sérines/thréonines (352 chez l'humain dont 240 cytosoliques et 12 liées à un récepteur) et pour les tyrosines (90 dont 32 cytosoliques et 58

liées à un récepteur). Certaines possèdent une double spécificité. Les kinases ont été classifiées en 10 groupes: TK (les tyrosines kinases qui incluent les kinases Egfr, Pdgfr, Trk, Src, Abl et Eph), CMGC (pour Cdk1, Mapk, Gsk3 et Clk), CamK (« calcium/calmodulin-dependent protein kinase » - CaMK, Mark, Ampk et Rsk), TKL (« tyrosine kinase-like » - Irak, Mos, Ksr et Raf), STE (« homologs of Sterile kinases » - Mekk, Mek et Pak), CkI (« casein kinase I » - CkI, Myt et Bubl), AGC (pour protéines kinases A, G et C - Pka, Pkg, Pkc, S6k et Akt), RGC (« receptor guanylate cyclase ») et les kinases atypiques (40 membres). Le génome humain contient aussi 106 pseudogènes (gènes inactifs) et 50 kinases inactives qui ne possèdent pas le site de liaison de l'ATP ou encore qui sont déficientes en résidus catalytiques. Plus de 100 kinases avec un polymorphisme ont été identifiées comme responsables de plusieurs maladies et cancers humains. La fréquence et l'importance de ces polymorphismes ont soulevé un grand intérêt dans la caractérisation des kinases et de leurs substrats dans le but de développer des thérapies et médicaments. En 2009, dix inhibiteurs de kinases ont été approuvés dans le traitement du cancer chez l'homme et plusieurs centaines de molécules chimiques étaient à l'étude [7].

1.1.1.2 Les protéines phosphatases

Les phosphatases ont pour rôle d'enlever le groupement phosphate des acides aminés phosphorylés et de contrebalancer l'activité des kinases. Le génome humain contient 150 phosphatases. Elles forment quatre groupes distincts qui se différencient par leur domaine catalytique et leur préférence pour les substrats [1]. En plus des variations dans le domaine catalytique, les groupes de phosphatases ont différents domaines et sous-unités protéiques régulateurs pour contrôler la reconnaissance des substrats. Les trois premiers groupes effectuent la déphosphorylation des phosphosérines/thréonines. Le premier groupe est formé par les PPP (phosphoprotéine phosphatase) et inclut les protéines PP (protéine phosphatase) 1 à 7. Le deuxième groupe est constitué par les PPM (protéine phosphatase métallo-dépendante au Mg^{2+} ou Mn^{2+}; inclut Pp2c). Le troisième groupe, les phosphatases

aspartiques avec la signature du domaine catalytique DXDX[T/V], inclut les membres de la famille FCP/SCP (« TFIIF (transcription initiation factor IIF)-associating component of CTD phosphatase/small CTD phosphatase») et HAD (« haloacid dehalogenase »). Le quatrième groupe, spécifique aux tyrosines, les PTP (protéine tyrosine phosphatase) avec la signature du domaine catalytique CX_5R, est un large groupe (107 membres) qui est aussi reconnu pour déphosphoryler d'autres molécules comme les glucides, les ARNm et les phosphoinositides. Ce groupe inclut aussi les DUSP (« dual-specificity phosphatase») qui déphosphorylent les phosphosérines/thréonines en plus des phosphotyrosines. Les études génomiques [1] ont démontré que la famille des PPP est très ancienne (retrouvée chez les eucaryotes, bactéries et archaebactéries) tandis que la famille des PTP s'est étendue au cours de l'évolution des métazoaires. Cette observation suggère que le développement du système de signalisation phosphotyrosine est un facteur important de l'évolution des organismes pluricellulaires. Finalement, les protéines phosphatases sont considérées comme des gènes suppresseurs de tumeurs puisqu'ils agissent à l'opposé des kinases et sont donc aussi des cibles thérapeutiques intéressantes.

1.1.1.3 Domaines liant les acides aminés phosphorylés

Les sites phosphorylés peuvent servir de point d'ancrage aux protéines ayant un domaine les reconnaissant spécifiquement. Ces protéines servent de modulateurs dans la propagation du signal dans diverses voies de signalisation [8]. Il existe 10 domaines connus qui lient les sites phosphorylés: 8 spécifiques aux phosphosérines/thréonines et 2 aux phosphotyrosines. Les 14-3-3 sont une famille de 7 protéines chez les mammifères qui reconnaissent le motif RXXp[S/T]XP ou RX[Y/F/S]Xp[S/T]XP. Elles sont impliquées dans plusieurs processus tels que la rétention cytosolique, le contrôle du cycle cellulaire, l'apoptose, la transcription, etc. Le domaine FHA (« forkhead-associated ») se lie au motif pTXX[D/I/F/S] et est présent dans certaines protéines impliquées au niveau des points de contrôle de dommages de l'ADN. Quatre domaines sont constitués de répétition en tandem: BRCT, FF, WD40 et

LRR («leucine-rich repeat »). La séquence phosphorylée reconnue par ces domaines est variable d'une version du domaine à l'autre. Des variantes des domaines WD40 et LRR sont aussi connues pour lier les méthyl-lysines. Le domaine WW reconnait les sites avec le consensus minimal p[S/T]P. Les cibles de ce domaine sont impliquées dans le contrôle du cycle cellulaire. Le domaine FF est retrouvé sur la région CTD de l'ARN polymérase II et est fréquemment accompagné du domaine WW. Le domaine MH2 reconnait une séquence doublement phosphorylée, pSXpS. Le domaine POLO-box est présent seulement sur les kinases POLO. La séquence Sp[S/T] reconnue sert à contrôler la localisation de la kinase et possède aussi un rôle d'auto-inhibition. Les domaines PTB (« Phosphotyrosine binding ») et SH2 (« Src-homology 2 ») ont une spécificité pour les tyrosines phosphorylées. Le domaine PTB reconnait le motif NPXpY. Il est intéressant de noter que chaque domaine présente une structure distincte malgré qu'ils aient tous pour rôle de lier les acides aminés phosphorylés. Considérant la nature dynamique de la phosphorylation, les domaines liant les résidus phosphorylés sont des éléments importants dans le recrutement et l'assemblage de complexes protéiques de signalisation.

1.1.2 Méthodes analytiques et de séparations

Le premier objectif de la phosphoprotéomique est de détecter les protéines phosphorylées. À cette fin, ce domaine utilise plusieurs méthodes analytiques et de séparation pour détecter la phosphorylation. Une stratégie, ancienne mais toujours courante, a recours aux méthodes de la chimie des protéines et de l'enzymologie pour marquer les phosphoprotéines avec un isotope radioactif du phosphore (^{32}P) [9]. Pour incorporer cet isotope lors de la phosphorylation, l'orthophosphate radioactif ($H_3{}^{32}PO_4$) est ajouté aux cultures cellulaire *in vivo* (et sera incorporé à l'ATP par la cellule) tandis que pour les essais kinases *in vitro*, l'[γ-^{32}P]ATP sera plutôt utilisé dans le milieu réactionnel. Une fois marquées, les protéines sont généralement séparées par électrophorèse sur gel de polyacrylamide (séparation dans un champ électrique en fonction du poids moléculaire). Ensuite, un film photographique est

exposé au gel pour révéler les protéines phosphorylées (autoradiographie). L'électrophorèse bidimensionnelle sur gel (séparation par poids moléculaire et selon le point isoélectrique) est employée pour l'étude d'échantillons plus complexes. Cette méthode est très sensible et la limite de détection se situe dans l'ordre du femtomole. Un désavantage de cette méthode est le danger associé à la manipulation du matériel radioactif et aussi le temps nécessaire pour que le ^{32}P remplace le phosphore non-radioactif dans les cellules. Outre le marquage radioactif, un immunobuvardage de type « western » avec des anticorps spécifiques aux phosphosérines/thréonines/tyrosines est aussi utilisé pour détecter les protéines phosphorylées. Selon les anticorps, une sensibilité similaire à la méthode précédente peut-être obtenue. La sélection des anticorps doit être faite avec vigilance puisque la spécificité de certains anticorps peut être variable et certains peuvent se lier à des protéines non-phosphorylées. Récemment, un agent de coloration fluorescent nommé « Pro-Q® Diamond » (Molecular Probes®, Invitrogen) a été démontré comme suffisamment sensible (limite de détection dans l'ordre du picomole) pour devenir une alternative [10]. Cette méthode a cependant subi des critiques au niveau de la spécificité de la coloration. Une autre approche originale pour déterminer *in vitro* à l'échelle du protéome quelles protéines sont phosphorylées par une kinase consiste en l'emploi d'une micropuce à protéines. L'incubation de la micropuce de protéines avec de l'[γ-^{32}P]ATP et une kinase permet d'identifier les substrats phosphorylés. Cette méthode a permis l'identification de 4200 sites phosphorylés sur 1325 protéines phosphorylées par 87 kinases dans la levure [11]. Ces résultats ont mis en évidence la spécificité des différentes kinases et permis d'assembler un réseau potentiel de kinase-substrat chez la levure. Les reproches faits à cette méthode sont relatifs au contexte *in vitro* de l'expérience qui peut mener à des faux positifs ou négatifs. Dans ce contexte, il y a absence de cofacteurs, de pré-phosphorylation, de compartiments cellulaires limitant l'accès aux kinases et les protéines n'ont pas nécessairement leur conformation native sur la micropuce.

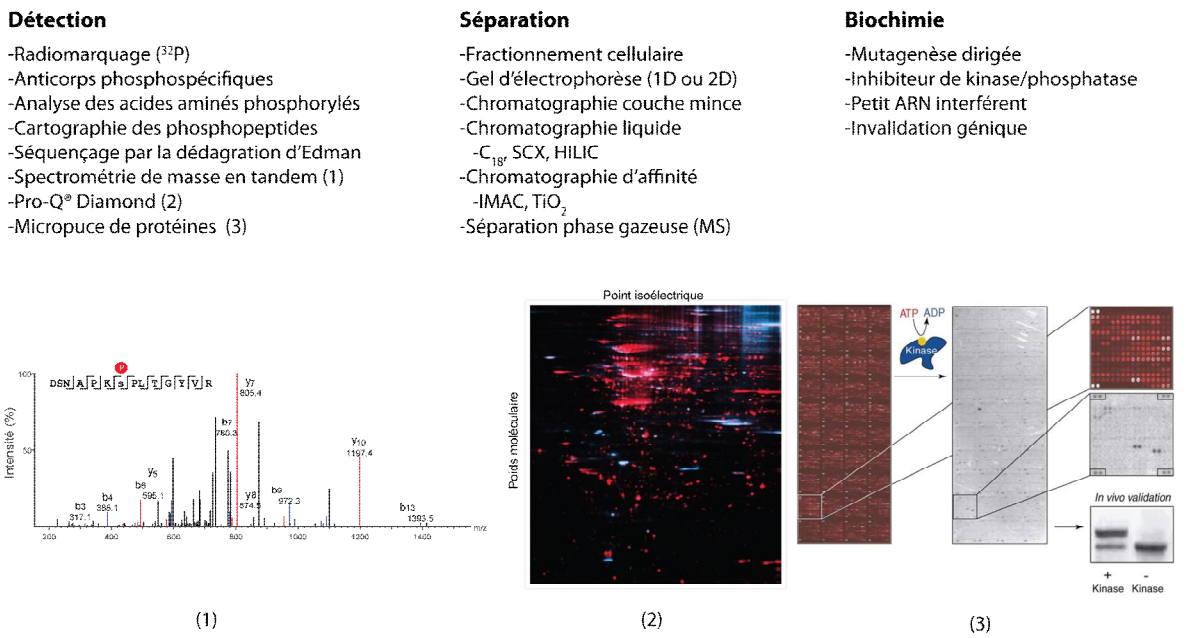


Figure 1.3 : Boîte à outils phosphoprotéomiques

Plusieurs méthodes de détection et séparation sont utilisées pour l'étude de la phosphorylation. Différentes approches biochimiques sont appliqués ensuite pour la validation (confirmer la localisation du site, identifier la kinase). Adaptation de la brochure de Molecular Probes®, Invitrogen et [12] avec la permission d'Elsevier: Cell, copyright 2006

Le deuxième objectif de la phosphoprotéomique est de localiser précisément la position des phosphorylations sur la protéine. Pour déterminer quel type d'acide aminé est phosphorylé, l'hydrolysat acide d'une protéine purifiée (ou d'un peptide suite à une digestion enzymatique) peut être séparé par chromatographie sur couche mince. La migration de l'hydrolysat peut être comparée à un standard suite à une autoradiographie ou à une coloration à la ninhydrine pour déterminer le type de résidus phosphorylés. Lorsque cette méthode est appliquée à un peptide connu on peut localiser la région de la protéine qui est phosphorylée si la séquence du peptide ne contient qu'un seul site phosphorylable.

Sinon pour localiser la position exacte du site phosphorylé, le séquençage des peptides, par la dégradation d'Edman, peut être effectué. Cette méthode laborieuse perd de son efficacité avec la longueur du peptide. Actuellement, le séquençage par la dégradation d'Edman a été remplacé dans la majorité des cas par la spectrométrie de masse en tandem couplée à la chromatographie liquide (LC-MS/MS) qui permet d'identifier simultanément des milliers de protéines phosphorylées tout en localisant la position des sites. Étant donné l'importance de cette méthodologie en phosphoprotéomique et de son usage prédominant dans les résultats présentés dans cette thèse, la section 1.2 en décrit les principes élémentaires. Dans le domaine de la biologie moléculaire, la mutagénèse dirigée est souvent employée en conjonction avec les méthodes de détection énumérées précédemment pour vérifier la localisation d'un site phosphorylé. Cette méthode consiste à modifier la séquence du gène afin d'obtenir un acide aminé différent à la position phosphorylée. S'il y a baisse ou abolition du signal de phosphorylation suite à la mutation, on pourra confirmer la position. Pour l'étude fonctionnelle de la phosphorylation, la mutagénèse dirigée est aussi employée pour créer des phosphomimétiques où la position phosphorylée est remplacée par un acide aminé chargé négativement dans le but d'obtenir la même fonction mais de façon constante.

La phosphoprotéomique cherche à répertorier l'ensemble complet des sites phosphorylés de tout le protéome. Pour atteindre ce but, il faut prendre en considération la nature dynamique du phosphoprotéome. Lorsqu'un échantillon est analysé, les résultats obtenus représentent le phosphoprotéome à un temps donné sous certaines conditions en fonction de l'activité des diverses kinases et phosphatases. Le niveau de phosphorylation de chaque site est différent d'une condition biologique à une autre et donc l'ensemble de sites détectés sera aussi partiellement différent. Pour obtenir un phosphoprotéome complet, divers états biologiques (conditions, tissus) devront être explorés. Au niveau de la préparation, un point technique impératif pour éviter la dégradation de l'échantillon est l'emploi d'un mélange d'inhibiteurs de protéases et de phosphatases lors de la lyse cellulaire [13]. D'autres facteurs à considérer pour maximiser la proportion du phosphoprotéome identifiée sont l'abondance des phosphoprotéines et la complexité du

mélange. Pour la détection de molécules de faible abondance, des méthodes de détection sensibles et/ou une augmentation de la quantité de protéines extraites sont nécessaires. Un échantillon contenant un mélange complexe de protéines pose aussi problème puisque les protéines abondantes masquent le signal des autres protéines. Les méthodes de séparation des protéines sont donc essentielles en protéomique pour en détecter le plus grand nombre. Outre l'électrophorèse sur gel, la chromatographie liquide haute performance (HPLC) est une méthode de séparation, applicable aux protéines et peptides, qui est largement utilisée puisqu'elle est automatisée et couplée à la spectrométrie de masse. Cette méthode sépare les molécules en fonction de leur affinité de liaison pour une colonne chromatographique (la phase stationnaire) lorsqu'un solvant (la phase mobile) y circule. En variant progressivement la composition du solvant au cours de l'expérience, l'affinité de certains composés pour la colonne est diminuée et ils se décrochent progressivement. On obtient donc une séparation. On compte plusieurs variantes de cette méthode en fonction du type de séparation désirée. La chromatographie liquide en phase inverse est une méthode courante pour la séparation des peptides en fonction de leur hydrophobicité. La phase stationnaire est composée de billes de silice sur lesquelles sont greffées des chaînes d'alkyles avec un certain nombre d'atomes de carbone (généralement 18 pour les peptides). La transition d'un solvant aqueux à un solvant organique (acétonitrile) selon un certain gradient permet d'éluer les composés de la colonne. La chromatographie par échange d'ions permet de séparer les molécules selon leur charge électrique nette en fonction du pH de la solution. Un échangeur de cations (SCX, « strong cation exchange ») utilise une phase stationnaire chargée négativement qui est composée de groupements d'acide sulfonique. Le solvant d'élution est constitué d'un gradient de sel de concentration croissante pour perturber les interactions ioniques.

Par chromatographie liquide, il est possible que des peptides de masse similaire coéluent et rendent difficile ou impossible la détection de l'un des peptides. Un cas particulier est la séparation d'isomères positionnels de phosphopeptides, des peptides avec la même séquence mais phosphorylés à différentes positions. Étant donné leurs propriétés

physico-chimiques semblables, la séparation peut être problématique et nécessiter d'autres moyens de séparations [14-17]. De plus, en présence d'un échantillon complexe où le nombre de peptides est important, il est avantageux de combiner diverses méthodes chromatographiques pour permettre un échantillonnage plus complet. Comme la chromatographie liquide en phase inverse et l'échangeur de cations séparent les peptides selon des propriétés physicochimiques orthogonales, la combinaison des deux méthodes permet d'obtenir une excellente séparation [18]. Finalement, la méthode de séparation qui a apportée le plus à la phosphoprotéomique est la stratégie d'enrichissement des phosphopeptides.

1.1.3 Méthodes d'enrichissement des phosphopeptides et phosphoprotéines

Une analyse typique LC-MS/MS d'un extrait total de protéines n'identifiera guère qu'un pourcent de phosphopeptides en général. Ceci s'explique par plusieurs facteurs [13]. D'abord les protéines sont généralement faiblement phosphorylées (environ 3 sites par protéine) et donc suite à une digestion trypsique, seulement une très faible proportion de la population de peptides aura un groupement phosphate. De plus, la présence de peptides non-phosphorylés entraîne une suppression du signal des phosphopeptides. Deuxièmement, les sites sont phosphorylés à une faible stœchiométrie et donc le peptide non-phosphorylé est présent en plus grande abondance que le même peptide phosphorylé. Troisièmement, l'ajout de la charge négative du ou des phosphates entraîne une baisse d'efficacité d'ionisation des phosphopeptides ce qui diminue l'intensité du signal mesuré. Tous ces facteurs combinés ne laissent qu'une mince probabilité aux phosphopeptides d'être échantillonnés. Afin de concentrer les efforts de détection uniquement sur les phosphopeptides, plusieurs méthodes d'enrichissement des phosphoprotéines ou phosphopeptides ont été développées pour augmenter leur proportion et la sensibilité de détection. Parmi les méthodes élaborées, on compte trois méthodologies (Figure 1.4):

l'immunoprécipitation, la chromatographie d'affinité avec ions métalliques immobilisés et la dérivatisation chimique.

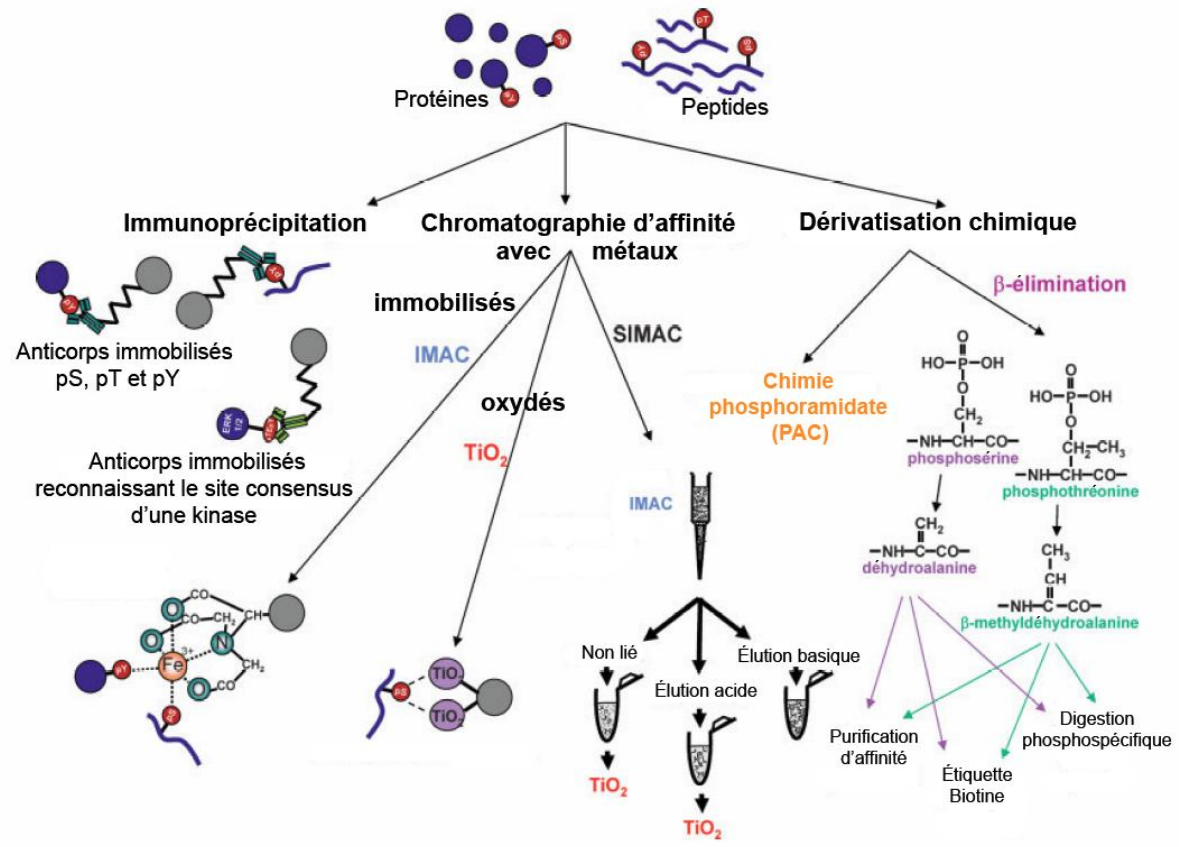


Figure 1.4: Méthodes d'enrichissement des phosphoprotéines et phosphopeptides.

Trois principales stratégies utilisées pour l'enrichissement des phosphopeptides et phosphoprotéines. L'immunoprécipitation et la chromatographie d'affinité sont les plus couramment utilisées. Adaptation de [13] avec la permission de Wiley-VCH Verlag GmbH & Co. KGaA: Proteomics, copyright 2009

1.1.3.1 Immunoprécipitation avec des anticorps phospho-spécifiques

L'immunoprécipitation est une méthode qui permet de précipiter une protéine en solution avec un anticorps qui reconnait une région spécifique de la protéine (épitope). Ainsi, il est possible d'isoler et de concentrer une protéine d'intérêt. Cette méthode est applicable pour l'enrichissement des phosphoprotéines et phosphopeptides. Des anticorps immobilisés sur billes qui reconnaissent les épitopes d'une sérine, thréonine ou tyrosine phosphorylée sont utilisés pour isoler les phosphopeptides. Le coût de production d'anticorps pour l'immunoprécipitation des phosphopeptides rend cette méthode plus onéreuse que les autres alternatives. En dépit du coût, la spécificité des anticorps procure un avantage pour certains phosphopeptides. Par exemple, l'immunoprécipitation est la méthode de choix pour l'étude des phosphotyrosines étant donné leur faible ratio dans le phosphoprotéome (environ 2% pY, 12% pT, 86% pS) [19]. La disponibilité d'anticorps hautement spécifiques pour les phosphotyrosines a permis d'étudier avec succès la signalisation cellulaire où la phosphorylation des tyrosines est très importante [20]. L'utilisation des anti-pS et anti-pT a été plus limitée en phosphoprotéomique dû à leur faible spécificité et affinité (causé par la faible immunogénicité des chaînes latérales) [21]. Une autre application où la spécificité des anticorps est avantageuse, est l'isolation des phosphopeptides ayant un site consensus d'une kinase particulière. Ceci a été utilisé par exemple pour identifier les substrats des kinases Erk1/2 [22].

1.1.3.2 Chromatographie d'affinité avec ions métalliques immobilisés et métaux oxydés

La chromatographie d'affinité avec ions métalliques immobilisés connue sous le nom d'IMAC (« Immobilized metal ion affinity chromatography ») a été adaptée pour la capture des peptides phosphorylés [23]. En condition acide, les ions positivement chargés de fer (Fe^{3+}), de gallénium (Ga^{3+}), d'aluminium (Al^{3+}) ou de cobalt (Co^{2+}) se comportent comme des acides de Lewis et ont les propriétés d'un échangeur d'ions. Les phosphopeptides

négativement chargés peuvent donc s'y lier en condition acide et être élués en condition basique. IMAC souffre d'un problème de liaison non-spécifique avec des peptides contenant plusieurs résidus acides. Pour pallier à ce problème, la méthyl-estérification des groupements carboxyliques préférablement effectuée [24]. Néanmoins, ce procédé complexifie l'échantillon si la réaction n'est pas complète et peut aussi mener à des réactions secondaires non désirées. Plus récemment, les dioxydes de titane (TiO_2) [25] et de zirconium (ZrO_2) ont été utilisés avec succès pour l'enrichissement. Les oxydes de niobium (Nb_2O_5) et d'hafnium (HfO_2) ont aussi été proposés, mais ils demeurent peu utilisés car leurs capacités et spécificités de rétention des phosphopeptides sont peu connues. Suite à l'optimisation du protocole d'enrichissement au dioxyde de titane [26], cette méthode s'est répandue car elle a un faible taux de liaisons non-spécifiques et est tolérante à différents tampons contenant des sels et détergents fréquemment utilisés lors de la préparation des échantillons. Étant donné les différentes affinités des diverses méthodes (biais dans l'enrichissement), elles sont parfois utilisées simultanément ou consécutivement pour maximiser le nombre d'identifications. La méthode SIMAC (« Sequential elution from IMAC ») combine les deux méthodes les plus efficaces (en nombre de phosphopeptides identifiés) soit IMAC (Fe^{3+}) et le dioxyde de titane [13]. SIMAC bénéficie du biais d'enrichissement de chaque méthode, soit des peptides multiplement phosphorylés pour IMAC et simplement pour TiO_2) pour enrichir des ensembles de phosphopeptides complémentaires et donc augmenter l'ensemble des phosphopeptides identifiés.

1.1.3.3 Dérivatisation chimique

La dérivatisation chimique procède au remplacement du groupement phosphate par un autre groupement fonctionnel dans le but d'enrichir le contenu de l'échantillon en phosphopeptides et de palier aux problèmes de détection (perte neutre de H_3PO_4 et mauvaise ionisation – voir section 1.2.2.1). Une première méthode, applicable seulement aux phosphosérines et phosphothréonines, consiste en une réaction de β-élimination suivie

d'une addition de Michael [27]. L'usage d'une base forte élimine le groupement phosphate et forme la déhydroalanine pour la sérine et la β-méthylhydroalanine pour la thréonine (Figure 1.5). Cette réaction n'est pas possible pour la tyrosine puisqu'il n'est pas énergétiquement favorable d'enlever le proton β du groupement aromatique et encore moins d'y former une liaison π supplémentaire. Ensuite, une molécule avec un groupe sulfhydrile peut réagir avec les composés précédemment formés par une réaction d'addition de Michael formant ainsi un di-thiol. Ce di-thiol peut servir à trois fonctions. Premièrement, une chromatographie d'affinité peut être effectuée directement afin de purifier les peptides. Deuxièmement, il peut être converti pour une digestion par une protéase spécifique à une lysine afin d'obtenir un clivage spécifique au site phosphorylé. Troisièmement, il peut être utilisé comme agent de réticulation pour y attacher une molécule de biotine afin de purifier aussi par chromatographie d'affinité. Un peptide marqué à la biotine a l'avantage d'être facilement détectable par un ion rapporteur dans le spectre MS/MS (*m/z* 446.3). Toutefois, des réactions non-désirées et des modifications post-traductionnelles favorables à la β-élimination (comme la O-glycosylation) peuvent causer une fausse interprétation des résultats.

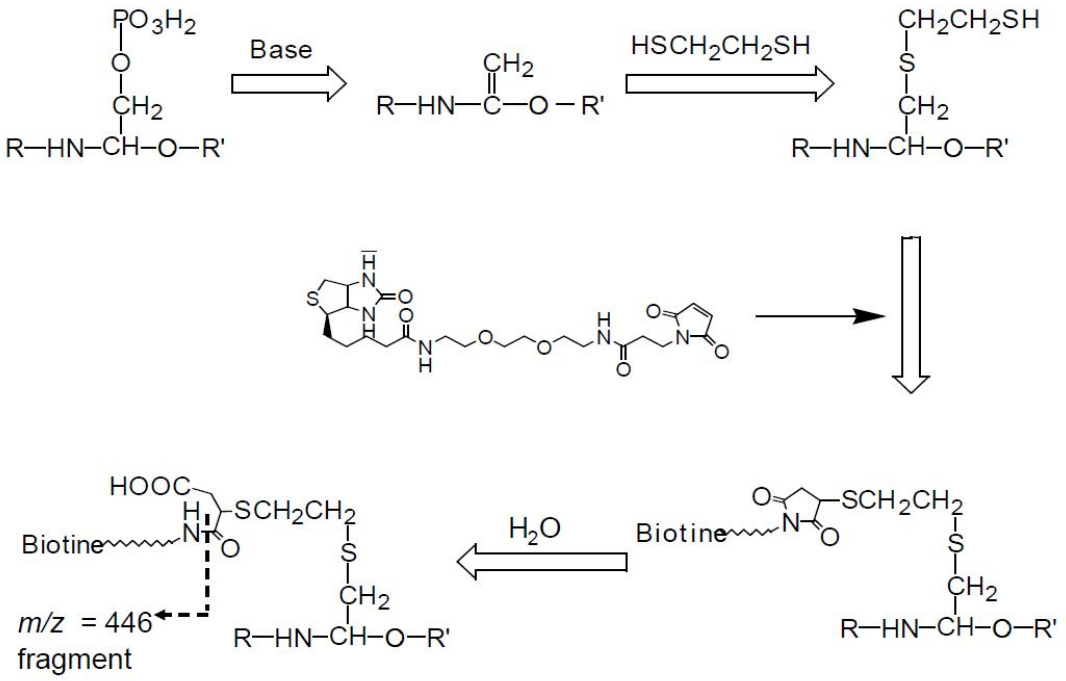


Figure 1.5: Conversion chimique d'une phosphosérine en résidu biotinylé pour l'enrichissement des phosphopeptides.

Schéma des 4 étapes réactionnelles nécessaires pour l'enrichissement des phosphopeptides biotinylés. La première étape est une β-élimination avec une base forte pour former un résidu déhydroalanyl réactif. L'éthanedithiol, un nucléophile, est ajouté par une addition de Michael. La biotine, le marqueur utilisé pour l'enrichissement des phosphopeptides, est ensuite greffé. Adaptation de [27] avec la permission de Macmillan Publishers Ltd: Nature Biotechnology, copyright 2001

Alternativement, la chimie de phosphoramidate (PAC) peut être employée pour fixer les phosphopeptides sur un support solide [28]. Contrairement à la β-élimination, cette approche est aussi valide pour les phosphotyrosines. Le protocole requiert six étapes (illustrées à la Figure 1.6) qui imposent une charge de travail considérable. Les nombreuses étapes de cette méthode, comme pour la β-élimination, induisent une perte d'échantillon dû au rendement limité de chaque étape réactionnelle. Les deux méthodes de dérivatisation

chimique présentées ici ont un biais différent d'enrichissement. La β-élimination est limitée aux phosphosérines/thréoniques tandis que la chimie des phosphoramidate permet aussi l'analyse des phosphotyrosines.

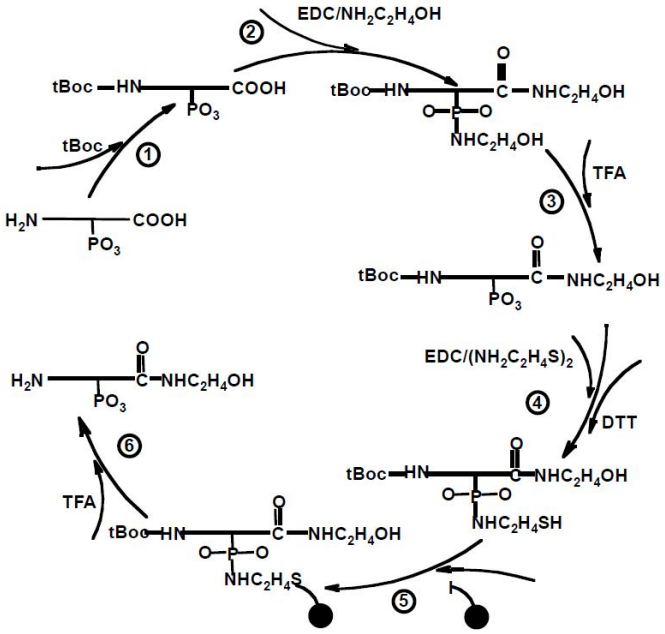


Figure 1.6: Stratégie d'isolation des phosphopeptides par la chimie de phosphoramidate (PAC).

Schéma des 6 étapes d'isolation des phosphopeptides. Dans cette méthode, les phosphopeptides sont fixés de façon covalente sur un support solide pour être enrichis. (1) Protection des groupements aminés avec le t-butyl-dicarbonate (tBOC) pour éviter des réactions de condensation inter ou intra moléculaires. (2) Réaction de condensation ave le N,N'-dimethylaminopropyle éthyle carbodiimide HCl (EDC) catalysée avec le carbodiimide pour former un amide et phosphoramidate avec le groupement carboxylate et phosphate respectivement. (3) Régénération du phosphate avec l'acide trifluoroacétique (TFA). (4) Condensation pour attacher une cystamine sur le groupement phosphate et réduction pour obtenir un groupement sulfhydrile libre. (5) Capture sur la phase solide par la réaction des sulfhydriles avec des groupements iodoacétyles couplés sur des billes de verre. (6) Enrichissement des phosphopeptides et clivage du lien phosphoramidate avec le TFA à une concentration qui enlève aussi le groupement protecteur tBoc. Les groupements

carboxyliques demeurent bloqués à la fin de cette procédure. Reproduction de [28] avec la permission de Macmillan Publishers Ltd: Nature Biotechnology, copyright 2001

Le développement des méthodes d'enrichissement a propulsé le champ d'étude de la phosphoprotéomique et permet actuellement d'identifier quelques milliers de sites phosphorylés par expérience. Toutefois, aucune méthode discutée ici n'est capable d'enrichir le phosphoprotéome en entier. Chaque méthode est spécifique envers un certain sous-ensemble de phosphopeptides et les identifications obtenues par les différentes méthodes ne se recoupent que partiellement. Il est donc avantageux, mais plus laborieux, de combiner les méthodes. L'optimisation et le développement de nouvelles méthodes d'enrichissement est toujours d'actualité. Il a été démontré que ces méthodes sont assez robustes et reproductibles pour les associer à la protéomique quantitative [29].

1.2 Spectrométrie de masse

Depuis le début des années 1990, diverses innovations technologiques ont permis à la spectrométrie de masse de s'imposer comme méthode de choix pour l'exploration du protéome. Cette technologie permet, en une seule expérience, d'identifier quelques milliers de protéines ainsi que leurs modifications chimiques. La spectrométrie de masse est donc un puissant outil en biologie moléculaire pour étudier les systèmes biologiques de la cellule.

1.2.1 Spectromètre de masse

Un spectromètre de masse mesure le rapport masse sur charge (*m/z*) des molécules ionisées. L'instrument est composé de trois principales parties: la source d'ionisation, l'analyseur de masse et le système de détection. Chaque partie peut être constituée par diverses technologies qui comportent différentes caractéristiques. La source d'ionisation sert à faire passer les analytes de la phase solide ou liquide en phase gazeuse pour les insérer dans l'instrument. Le processus crée des ions positifs ou négatifs qui sont ensuite accélérés dans le spectromètre de masse. Parmi les nombreuses sources d'ionisation existantes, deux se sont distinguées du lot en protéomique dû à leur capacité d'ioniser les protéines et peptides sans les dégrader. La méthode d'ionisation par désorption laser avec matrice [30] (MALDI, « Matrix assisted desorption ionization ») ionise par impulsion laser les molécules dissoutes dans une matrice sur une plaque. Cette méthode est pratique pour analyser rapidement plusieurs échantillons de faible complexité. L'ionisation par électronébulisation [31] (ESI, « Electrospray ionization ») consiste à appliquer une différence de potentiel élevée à un capillaire dans lequel une solution circule à faible débit afin de vaporiser celle-ci. Les gouttelettes formées s'évaporent progressivement et libèrent les analytes ionisés qui se dirigeront vers l'orifice de l'instrument sous l'effet d'un champ électrique. Cette méthode est la plus utilisée en protéomique car elle a l'avantage de pouvoir être couplée directement à la chromatographie liquide pour la séparation des peptides.

L'analyseur de masse sert à séparer les ions peptidiques selon leur rapport masse sur charge (*m/z*). Il existe plusieurs analyseurs de masse qui fonctionnent selon des principes différents. Une première catégorie d'analyseur se sert d'un champ électrique pour filtrer et/ou capturer les ions: le quadripôle (Q), la trappe ionique linéaire (LIT, « Linear ion trap » ou LTQ, « Linear trap quadrupole ») ou tridimensionnelle (IT, « Ion trap »). Les trappes ioniques ont l'avantage de pouvoir capturer les molécules, isoler la population d'une molécule d'intérêt, la fragmenter et analyser les fragments pour obtenir plus d'informations sur la molécule. La fragmentation des peptides survient par collision avec un gaz neutre (azote) qui est injecté dans la trappe. Cette approche est appelée la spectrométrie de masse

en tandem (MS/MS, voir section 1.2.2). Les analyseurs de cette première catégorie sont de basse résolution et ont une faible précision de masse (précision = $m_{théorique}$ - $m_{mesurée}$). La résolution est une mesure de l'étroitesse du pic mesuré d'un composé (R = $m/\Delta m$ où m représente la masse mesurée et Δm la largeur du pic à 50% de sa hauteur). La résolution indique la limite de l'instrument à distinguer le signal de deux molécules avec un *m/z* semblable (Figure 1.7). Un instrument de haute résolution peut distinguer le profil isotopique de grosses molécules multiplement chargées. Le profil isotopique représente la probabilité d'inclure un ou plusieurs isotopes lourds (principalement le ^{13}C) pour la population d'une molécule donnée. La différence de masse entre chaque isotope est de 1 Da. Le profil isotopique s'observe sur le spectre MS par une série de pics séparés d'une valeur de *m/z* de *1/z* en fonction de la charge du peptide. La taille et la forme du profil varient en fonction de la taille de la molécule.

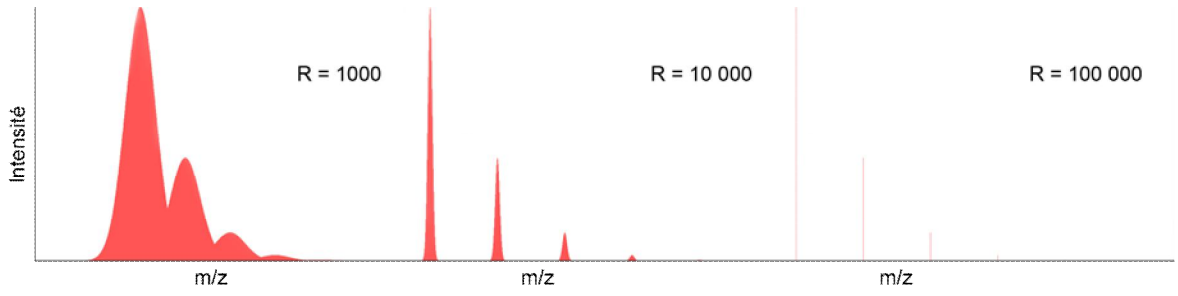


Figure 1.7: Profil isotopique d'un peptide à différentes résolutions.

Exemple d'un profil isotopique d'un peptide à trois résolutions croissantes. Une résolution (R) supérieure permet de mieux distinguer le signal de deux molécules avec des masses semblables. Les trois pics illustrés sont des isotopes d'un même peptide doublement chargé qui diffèrent de 0.5 *m/z*, soit 1 Da.

Dans la catégorie des instruments de haute résolution, on retrouve un analyseur qui mesure les *m/z* des ions en fonction du temps d'envol (TOF, « Time of flight ») et deux analyseurs qui mesurent la fréquence d'oscillation des ions dans un champ magnétique, la résonance cyclotronique ionique à transformée de Fourier (FTICR), et un champ électrique,

l'Orbitrap. Finalement, un spectromètre de masse contient aussi un système de détection, selon le type d'analyseur, qui rapporte l'abondance des espèces ioniques sous forme de signal électrique à un ordinateur.

Les spectromètres de masse utilisés en protéomique sont souvent de types hybrides. Ce type d'instrument est formé de plus d'un d'analyseur de masse afin de combiner différents avantages et plus de flexibilité dans les procédures d'analyses. Le triple quadripôle, le Q-TOF et le LTQ-Orbitrap sont des combinaisons intéressantes. Le LTQ-Orbitrap (Thermo Fisher Scientific), l'instrument employé pour cette thèse, est un instrument qui combine une trappe linéaire et l'Orbitrap (Figure 1.8) [32]. L'Orbitrap, inventé par Alexander Markarov, est un instrument composé de deux électrodes (une externe en forme de tonneau et une interne en forme de fuseau) qui génèrent un champ électrostatique quadro-logarithmique (une combinaison d'un champ de type quadripolaire d'une trappe ionique et d'un champ logarithmique d'un condensateur cylindrique; la géométrie est illustrée à la Figure 1.8). Lorsqu'on injecte des ions perpendiculairement à l'électrode dans l'appareil, les ions piégés oscillent selon un mouvement de va et vient le long de l'électrode interne avec la fréquence d'oscillation axiale ω (en radian/seconde) suivante:

$$\omega = \sqrt{(q/m)\,k} \qquad \textit{Équation 1}$$

où q est la charge de la molécule, m sa masse et k la courbure du champs. L'équation 1 et une transformée de Fourier appliquée sur le signal mesuré permettent de retrouver le rapport *m/z* des molécules présentes dans l'Orbitrap au moment de l'analyse. L'Orbitrap permet de déterminer avec haute résolution (>100 000) et précision (1-2 ppm) la masse d'une molécule. La précision de la masse est un facteur critique pour identifier une molécule sans ambiguïté et est définie comme le ratio de la différence entre la masse observée et la masse théorique sur la masse théorique. L'autre analyseur de masse de cet instrument est la trappe linéaire. Celle-ci est avantageuse car elle est capable de produire

très rapidement des spectres MS/MS de plusieurs peptides. Cette trappe a de plus une haute sensibilité car elle peut accumuler une grande quantité d'ions peptidiques. La combinaison de ces deux caractéristiques a permis à cet instrument d'être capable d'identifier avec une haute confiance plus d'un millier de peptides en une seule analyse.

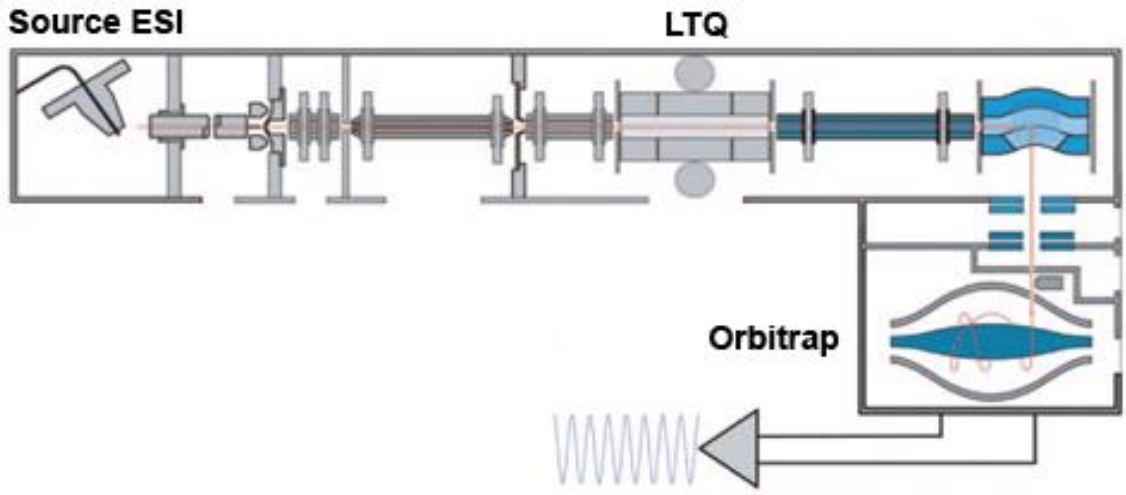


Figure 1.8: Schéma technique du spectromètre de masse hybride LTQ-Orbitrap.

Ce schéma illustre les composants du LTQ-Orbitrap. Les ions peptidiques sont introduits dans l'instrument par une source d'ionisation par électronébulisation (ESI). Ceux-ci sont ensuite détectés par l'analyseur de masse Orbitrap et un spectre MS haute résolution est alors obtenu. La trappe linéaire (LTQ), le second analyseur de masse, permet d'isoler un peptide détecté et de le fragmenter pour obtenir un spectre MS/MS qui indique la séquence de celui-ci. Adaptation du schéma de Thermo Fisher Scientific.

1.2.2 Spectrométrie de masse en tandem

En protéomique, la spectrométrie de masse en tandem permet d'identifier les protéines d'un échantillon. Il existe deux stratégies principales pour identifier les protéines: l'approche descendante (« top-down ») et l'approche ascendante (« bottom-up »). L'approche descendante identifie les protéines en les injectant dans leur forme entière dans le spectromètre de masse et en les fragmentant ensuite dans la phase gazeuse pour déterminer

leur identité. Cette méthode a l'avantage de pouvoir distinguer les différents isoformes et les combinaisons des différentes modifications post-traductionnelles sur les protéines. Toutefois cette approche est le plus souvent restreinte aux protéines de petit poids moléculaire car les analyseurs de masse sont limités à une gamme de faible masse. Cette barrière se réduit cependant à chaque nouvelle génération d'instruments. L'approche ascendante consiste à morceler d'abord les protéines en peptides par une digestion enzymatique. La digestion à la trypsine, une enzyme qui clive les protéines après les lysines et arginines non suivies d'une proline, est fréquemment employée car elle génère des peptides d'une taille et d'une charge favorable pour la détection avec la spectrométrie de masse en tandem. Les peptides sont ensuite séparés par chromatographie liquide en phase inverse. Au fur et à mesure que les peptides trypsiques éluent dans le spectromètre de masse, un cycle d'analyse se produit afin de les identifier (Figure 1.9).

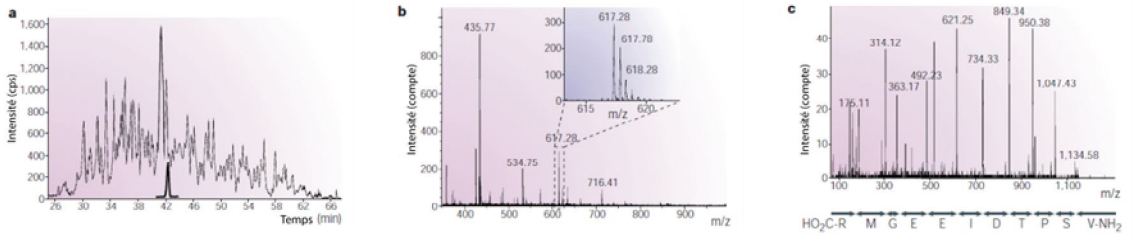


Figure 1.9: Cycle d'analyse MS/MS.

a) Chromatogramme total des ions peptidiques. L'intensité est en comptes par seconde (cps). b) Spectre MS des peptides présents à 42 minutes. L'agrandissement illustre le profil isotopique du précurseur peptidique avec un *m/z* de 617.28. c) Spectre MS/MS du peptide isolé et fragmenté avec la séquence correspondante assignée. Adaptation de [33] avec la permission de Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology, copyright 2004.

La première étape de ce cycle est d'effectuer un balayage des ions peptidiques présents. Le rapport *m/z* de chaque ion détecté et leur abondance sont enregistrées sous la forme d'un spectre MS. À partir des données du spectre MS, un algorithme choisit l'ion en plus forte abondance (non choisi préalablement) et donne l'instruction à l'instrument de l'isoler et de le fragmenter. L'ion peptidique isolé est fragmenté par collision avec un gaz neutre comme l'azote (CID, « collision induced dissociation »). Par cette méthode, on obtient généralement une fragmentation au niveau du lien peptidique. Selon la nomenclature de Biemann (Figure 1.10), on aura un ion *b* si la charge est retenue sur un fragment N-terminal et un ion *y* si elle est sur la portion C-terminale du peptide.

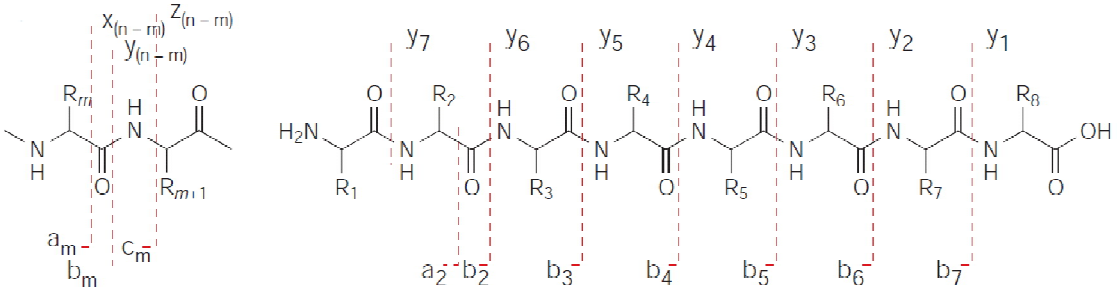


Figure 1.10: Nomenclature de Biemann des fragments peptidiques.

Selon la méthode de fragmentation employée lors de la spectrométrie de masse en tandem, le squelette peptidique fragmente à des endroits spécifiques nommés selon la nomenclature de Biemann. Adaptation de [33] avec la permission de Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology, copyright 2004.

Les fragments détectés sont rapportés par un spectre MS/MS. Grâce à la masse intact du peptide et celles de ses fragments, il est possible de déduire la séquence du peptide. L'espacement de masse entre deux fragments permet de déterminer les acides aminés qui composent la séquence du peptide (Figure 1.9C). L'interprétation automatisée des spectres MS/MS est expliquée en détails à la section 1.3.3. Une fois la séquence du peptide déterminée, on peut finalement associer celui-ci à une protéine si on connait au

préalable la séquence de celle-ci. Plus on obtient de peptides pour une protéine, plus la confiance en l'identification est élevée. L'approche ascendante est, jusqu'à maintenant, la méthode la plus efficace pour identifier les protéines présentes dans un échantillon. Cette façon de faire possède par contre certains aspects négatifs. Tout d'abord, en pratique, la couverture peptidique (le pourcentage de la séquence identifiée) d'une protéine est généralement faible. Plusieurs peptides sont trop courts ou trop longs pour être détectés et d'autres n'ionisent ou ne fragmentent pas bien. Ceci est problématique lorsque l'on cherche à savoir quelles régions d'une protéine portent des modifications. De plus, ces données fragmentaires empêchent de savoir si ces modifications sont présentes simultanément sur la protéine ou de déterminer de quels isoformes protéiques originent les peptides identifiés.

1.2.2.1 Analyse MS/MS des phosphopeptides

La spectrométrie de masse en tandem permet de détecter et localiser la phosphorylation sur les peptides. Pour déceler la phosphorylation, on doit chercher dans les fragments obtenus une sérine, une thréonine ou une tyrosine avec l'ajout de HPO_3. Cet ajout correspond à une augmentation de masse de 80 Da. Il faut une haute précision de masse pour distinguer la phosphorylation (79.96633 Da) de la sulfonation (79.95681 Da) qui diffèrent de seulement 9.5 mDa. La sulfonation est une modification post-traductionnelle [34] mais est aussi un artéfact chimique causé par la préparation de l'échantillon [35]. Ces deux modifications peuvent être distinguées en MS/MS par leurs signatures ioniques respectives. Pour la phosphorylation, on observe la présence d'un fragment qui porte la masse du précurseur peptidique moins la perte neutre de –98 Da ($–H_3PO_4$) pour les phosphosérines/thréonines et –80 Da ($–HPO_3$) pour les phosphotyrosines. Une perte de –80 Da ($−SO_3$) est détectable au niveau des sulfosérines/thréonines. Cet ion caractéristique est observé car le lien phosphodiester est très labile. Ce bris est problématique pour identifier les phosphopeptides car il diminue l'efficacité de fragmentation du peptide et la rétention de la phosphorylation sur les fragments appropriés. Ceci rend difficile la localisation de la

phosphorylation. Pour certains phosphopeptides, ceci est problématique au point où le squelette peptidique ne fragmente plus. Le spectre MS/MS contient alors un fragment très intense mais ne fournit pas d'information sur la séquence peptidique. La perte neutre de –98 Da ainsi que l'ion immonium de la phosphotyrosine (*m/z* 216) sont des signatures des phosphopeptides qui peuvent servir pour les cibler, optimiser leur détection et confirmer leur présence. Sur les instruments avec une trappe ionique, il est possible d'isoler le fragment avec la perte neutre et de le fragmenter pour déterminer sa séquence (méthode MS^3 [36]). Il est aussi possible d'ignorer l'étape d'isolation et de faire une seconde étape d'activation pour fragmenter le peptide (méthode d'activation multi-étapes ou pseudo-MS^3 [37]). Récemment, il a été démontré que la méthode de fragmentation ETD (« Electron transfer dissociation »), disponible commercialement depuis 2008 sur le LTQ-Orbitrap, pouvait être complémentaire dans l'identification des phosphopeptides [38]. Cette méthode induit la fragmentation des peptides ou protéines par le transfert d'un électron provenant d'un anion radicalaire du fluoranthène et induit la formation d'ions *c* et *z*. L'intense perte de neutre présente en CID n'est pas observée avec ETD permettant ainsi de déterminer la séquence peptidique et de localiser la position du groupement phosphate. La méthode ETD n'est cependant pas efficace pour tous les phosphopeptides et le peptide précurseur fragmente dans une faible proportion. Étant donné ce problème et la complémentarité des mécanismes de fragmentation CID et ETD, ces deux méthodes ont été combinées avec un algorithme d'arbre de décisions, inclus dans le logiciel d'acquisition, qui déclenche la méthode la plus efficace selon le *m/z* et la charge du phosphopeptide [39]. Finalement, un autre problème rencontré qui limite l'analyse des phosphopeptides est leur mauvaise ionisation. La forme phosphorylée ionise moins bien que la forme non-phosphorylée dû à la charge négative supplémentaire. Malgré les difficultés d'ionisation et de fragmentation, il est actuellement possible, avec la spectrométrie de masse en tandem combinée aux méthodes d'enrichissement des phosphopeptides, d'identifier environ 10 000 sites phosphorylés de l'ensemble du phosphoprotéome.

1.2.3 Méthodes quantitatives

La spectrométrie de masse permet de quantifier l'abondance des protéines et peptides dans un échantillon. La quantification en phosphoprotéomique est primordiale pour suivre les variations de phosphorylation qui contrôlent les voies de signalisation cellulaires. On distingue deux types de quantification : absolue et relative. La quantification absolue consiste à déterminer la quantité exacte en mole d'un composé. Pour quantifier un peptide A, un peptide identique synthétique Å marqué avec des isotopes lourds (^{13}C, ^{15}N) sera ajouté à une quantité connue dans l'échantillon. La comparaison de l'intensité du signal du peptide A avec celui du peptide Å permet de déterminer la quantité du peptide A. Une courbe d'étalonnage doit être tracée pour plus de précision et assurer une réponse linéaire. Cette méthode se limite habituellement à un ou quelques peptides puisque la synthèse de peptides lourds est relativement coûteuse. Afin d'obtenir une haute sélectivité et sensibilité de détection, l'analyse de ces peptides est souvent faite avec les procédures SRM (« Selected reaction monitoring ») ou MRM (« Multiple reaction monitoring ») [40], la version multiplexe de SRM, qui sont effectuées sur un instrument de type triple quadripôle. La procédure consiste à filtrer avec le premier quadripôle pour ne garder qu'une trajectoire stable pour le m/z d'un précurseur peptidique désiré. Le peptide isolé est fragmenté dans le deuxième quadripôle qui sert de cellule de collisions. Le troisième quadripôle est configuré pour transmettre jusqu'au détecteur un ou quelques fragments prédéterminés du peptide. Cette procédure est hautement sélective car l'instrument filtre à deux reprises les ions attendus et est hautement sensible car seulement les ions désirés se rendent au détecteur évitant ainsi sa saturation par d'autres ions présents en forte abondance (meilleur signal sur bruit). Ce type d'approche requiert toutefois une connaissance préalable de la masse du peptide et de ses ions fragments (transitions). Pour la quantification d'un phosphopeptide, les transitions doivent être sélectionnées en considérant les fragments permettant la localisation spécifique de la phosphorylation [41]. Jusqu'à maintenant, la quantification SRM n'a pas été appliquée en phase découverte des études phosphoprotéomiques étant donné les connaissances préalables nécessaires de la fragmentation des peptides pour

réaliser l'expérience avec succès. Suite à l'accumulation des données phosphoprotéomiques et le développement de méthodes SRM planifiées, la méthode peut analyser plus de peptides par analyse MS et semble très prometteuse pour la quantification d'un ensemble peptidique correspondant à une voie de signalisation prédéterminée.

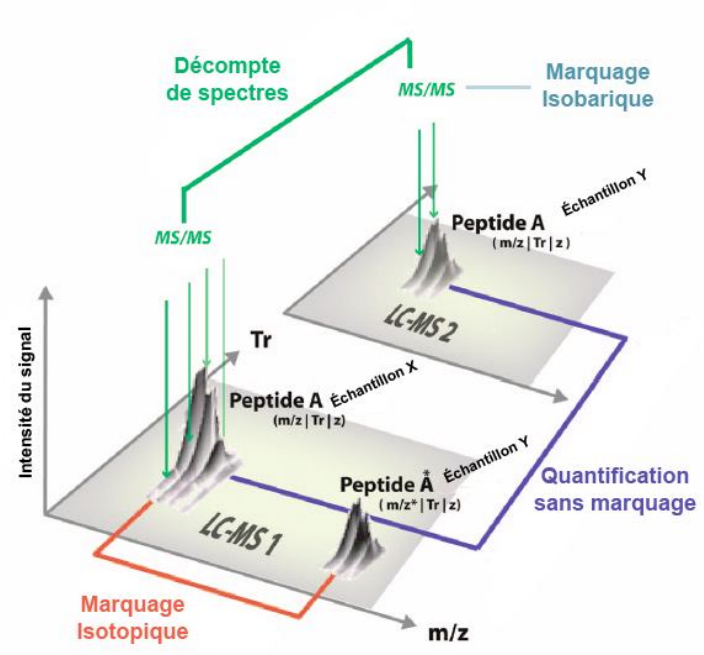


Figure 1.11: Méthodes de quantification des peptides en spectrométrie de masse.

Un éventail de méthodes quantitatives existe en protéomique pour mesurer l'abondance des peptides. La quantification peut nécessiter un marquage des peptides, s'effectuée par MS ou MS/MS, ou être multiplexée pour l'analyse simultanée de plusieurs échantillons. Adaptation de [42] avec la permission de l'American Chemical Society: Journal of proteome research, copyright 2008.

Pour quantifier plusieurs peptides, on a plutôt recours à la quantification relative. La quantification relative indique le changement en abondance d'un composé sous forme de ratio lorsque que l'on compare deux conditions (Figure 1.11). Le marquage isotopique est aussi compatible avec ce mode de quantification. La méthode SILAC (« Stable isotope labelling with amino acids in cell culture ») [43] consiste à marquer les protéines en

incorporant des acides aminés lourds dans le milieu de culture. L'arginine et la lysine lourde sont les plus fréquemment ajoutées afin que chaque peptide soit marqué une seule fois lors d'une digestion trypsique de l'extrait protéique. SILAC est utilisé, par exemple, pour comparer le protéome de cellules d'un groupe contrôle X à celles traitées par une drogue d'un groupe Y. Les protéines des cellules du groupe contrôle ne seront pas marquées tandis que celles du groupe traité le seront. Les deux échantillons seront combinés pour une seule analyse LC-MS/MS. Pour déterminer la variation en abondance d'un peptide, il suffit de calculer le ratio d'intensité du signal du peptide léger A de l'échantillon X et du peptide lourd Å de l'échantillon Y. Le peptide lourd Å élut au même temps de rétention (Tr) mais à une masse de *m* Da supérieur selon les acides aminés lourds choisis. La méthode SILAC est avantageuse puisqu'elle permet de faire la comparaison en une seule analyse mais a le désavantage de complexifier l'échantillon car le signal des peptides est dupliqué. Dans certaines lignées cellulaires, une conversion enzymatique de l'arginine lourde en proline est aussi problématique. De plus, la nécessité d'incorporer des acides aminés lourds dans les protéines ne permet pas l'utilisation de SILAC pour des patients humains lors de l'étude de maladies. Cette limitation a été récemment contournée par « Spike-in SILAC », une nouvelle méthode qui emploie un échantillon protéique marqué comme standard [44]. Alternativement, la méthode ICAT (« Isotope-coded affinity tags ») [45] utilise aussi un marquage isotopique mais celui-ci est effectué sur l'extrait protéique. Des réactifs chimiques sont employés comme étiquette légère ou lourde pour marquer les peptides au niveau des cystéines. La détection se fait ensuite selon le même principe que SILAC en cherchant une paire de peptides dans le spectre MS avec une différence de masse de *m* Da. Les étiquettes ICAT sont compatibles avec la chromatographie d'affinité ce qui permet de réduire la complexité de l'échantillon à analyser. Toutefois comme ICAT marque seulement les peptides avec cystéines, on a une couverture incomplète du protéome car seulement 80 à 90% des protéines en contiennent. SILAC et ICAT ne sont que deux exemples parmi plusieurs méthodes quantitatives qui emploient un marquage aux isotopes lourds [46].

Une autre méthode de quantification fait appel à des marqueurs isobariques. Dans ce cas, les peptides de chaque échantillon sont marqués et ensuite combinés. Les étiquettes ajoutées sont isobariques, c'est-à-dire qu'elles ont toutes la même masse, et n'augmentent pas la complexité de l'échantillon en principe si la réaction est complète. Un peptide marqué avec différentes étiquettes est donc détecté comme un seul pic sur un spectre MS. C'est en MS/MS que l'abondance du peptide dans chaque échantillon est révélée. Les étiquettes sont structurellement conçues pour générer des ions rapporteurs de masses différentes lorsque le peptide est fragmenté. Les ions rapporteurs sont de faibles masses et sont détectés dans une région du spectre MS/MS où il y a peu de fragments provenant du peptide. Le ratio d'intensité de ces ions rapporteurs est utilisé pour la quantification. On retrouve commercialement deux types de marqueurs isobariques: iTRAQ (« Isobaric tags for relative and absolute quantitation ») [47] et TMT (« Tandem mass tags ») [48]. Une particularité intéressante du marquage isobarique est la possibilité de faire une analyse multiplexe. Il est possible d'analyser simultanément jusqu'à 6 échantillons pour TMT et 8 pour iTRAQ. Les marqueurs isobariques peuvent aussi servir à effectuer une quantification absolue si on marque des peptides synthétiques dont la concentration est connue. La quantification des phosphopeptides avec les marqueurs isobariques ne semble pas appropriée car ceux-ci nuisent au succès d'identification en élevant la charge nette [49]. Malgré ceci, iTRAQ est occasionnellement utilisé en phosphoprotéomique et permet l'identification de plusieurs milliers de phosphopeptides [50].

Finalement, il y a aussi les méthodes quantitatives sans marquage. Ici chaque échantillon est analysé indépendamment par LC-MS/MS et l'abondance des peptides est comparée entre les analyses. Une première méthode simpliste et semi-quantitative consiste à faire le décompte des spectres MS/MS acquis pour chaque protéine [51]. Cette méthode est basée sur la prémisse que plus une protéine est abondante, plus les signaux de ses peptides seront élevés et susceptibles d'être échantillonnés à répétition. La fréquence d'échantillonnage des peptides d'une protéine est ensuite transformée en index d'abondance. Cet index d'abondance est dépendant du nombre de peptides, de la qualité et

de l'exactitude des identifications. Cet index peut être biaisé par la taille des protéines et l'observabilité de ses peptides en fonction de leurs propriétés physicochimiques qui influence leur rétention chromatographique, ionisation et fragmentation. Cette approche est appliquée sur les instruments de basse résolution puisqu'elle est simple et rapide à mettre en place. Sur les instruments de haute résolution, on procède souvent par comparaison directe de l'intensité du signal des peptides [52]. La haute précision de masse des spectromètres de masse, la reproductibilité de la chromatographie et la consistance dans la préparation d'échantillons sont critiques pour trouver les peptides identiques entre les conditions/ réplicats et limiter la variabilité. La corrélation automatisée d'abondance des peptides entre les échantillons est aussi sujette aux erreurs s'il y a variation du temps de rétention et si la complexité (nombre d'ions peptidiques) augmente. Cette méthode est en principe plus précise et non influencée par l'échantillonnage des peptides. Un désavantage majeur de l'approche sans marquage est l'impossibilité de faire une analyse multiplexe pour économiser du temps d'analyse. Finalement, pour toutes les méthodes quantitatives présentées dans cette section, le traitement informatisé de données MS est essentiel pour extraire les variations d'abondances pour chaque peptide. Ce processus s'avère plus compliqué pour la quantification sans marquage puisque la dimension temporelle doit être considérée en plus car les échantillons ne sont pas analysés en même temps. Étant donné l'usage de cette méthode quantitative dans cette thèse, le traitement des données quantitatives sans marquage est expliqué en détails à la section 1.3.4.

1.3 Protéomique computationnelle

L'informatique est impliqué en protéomique de l'acquisition des données avec le spectromètre de masse jusqu'à l'interprétation des données. L'identification et la caractérisation des protéines à partir des données MS/MS est un aspect important de ce processus. La comparaison et la quantification sont aussi des tâches importantes dans la compréhension des différences entre les diverses conditions biologiques ou expérimentales.

1.3.1 Acquisition des données

L'acquisition des données sur le spectromètre de masse est contrôlée par divers algorithmes selon la stratégie la plus appropriée pour identifier ou quantifier les peptides d'intérêts. La méthode dominante de découverte et à la base de l'approche ascendante est l'acquisition dépendante des données (DDA, « Data-dependent acquisition»). Cette méthode contrôle le cycle d'analyse où un balayage des peptides présents est effectué en premier lieu (spectre MS). Ce balayage est ensuite suivi de l'acquisition des spectres MS/MS sur un nombre déterminé de peptides les plus abondants. Une autre stratégie d'acquisition fréquemment utilisée est l'analyse ciblée. Il est possible de générer, à partir de données précédentes, des listes d'inclusion des ions peptidiques d'intérêts et d'ordonner ensuite à l'instrument de les analyser.

Les instruments de la génération actuelle sont capables d'effectuer plusieurs milliers de cycles au cours d'une seule heure et génèrent un fichier brut de sortie de quelques centaines de mégaoctets. Ceci engendre le déploiement d'une logistique importante dans la gestion et le traitement des données. Les données brutes de l'instrument transmises à l'ordinateur contiennent les spectres MS et MS/MS. Les spectres MS rapportent les ions peptidiques présents et sont utilisés pour la quantification des peptides. Les spectres MS/MS peuvent être utiles aussi à la quantification mais ils servent principalement à l'identification des peptides. Les spectres MS/MS bruts sont d'abord prétraités avant d'effectuer leur interprétation.

1.3.2 Prétraitement des spectres MS/MS

Le prétraitement des spectres MS/MS sert à générer une liste de pics (de fragments) pour identifier les peptides [53]. Un spectre de masse rapporte l'intensité des ions détectés par le spectromètre de masse à un certain rapport masse sur charge (m/z) sous forme de pic. Le pic mesuré d'un composé est défini de façon continue par plusieurs valeurs d'intensités

dans un court intervalle de *m/z* en fonction de la résolution de l'instrument. Cette structure de données est trop complexe pour permettre une interprétation efficace et rapide du spectre MS/MS. Alors, seulement les valeurs *m/z* et d'intensité au centroïde du pic sont retenues pour la génération de la liste de pics. Cette étape est généralement faite simultanément avec l'acquisition des données pour économiser de l'espace disque. Le prétraitement doit de plus distinguer les pics provenant d'un peptide de ceux du bruit afin d'éliminer les données non informatives. Cette étape peut être complexe si l'abondance du peptide est faible ou si certains fragments sont générés en faibles proportions. Il y a un risque que cette procédure enlève de l'information pertinente contenue dans le spectre MS/MS. Ceci peut réduire la confiance ou empêcher l'identification du peptide. Dans le cas de données MS/MS de haute résolution où les isotopes d'un fragment sont résolus, la série de pics de l'enveloppe isotopique d'un fragment est convertie en un seul pic avec la *m/z* le plus faible. Un fragment présent à plusieurs états de charge peut être réduit en un seul pic. La masse du précurseur peptidique associée au spectre MS/MS nécessite parfois un ajustement lorsque le pic monoisotopique n'a pas été sélectionné lors de l'isolation. La combinaison de spectres pour les mêmes précurseurs est aussi faite pour obtenir un meilleur signal sur bruit. La qualité du spectre est parfois évaluée pour déterminer s'il y a suffisamment de signal pour l'interpréter ou s'il vaut mieux l'éliminer. Bref, le prétraitement réduit la complexité des données et aussi la taille du fichier de données afin d'interpréter efficacement et plus rapidement les spectres. Ces procédures, qui éliminent le bruit et la redondance présents dans les spectres, augmentent le taux de succès d'identification des peptides en éliminant les ambiguïtés et en réduisant le taux de faux positifs.

1.3.3 Interprétation des spectres MS/MS et identification des protéines

La protéomique par l'approche ascendante procède à l'identification des protéines par l'identification de peptides. Étant donné qu'une analyse LC-MS/MS génère plusieurs milliers de spectres MS/MS par analyse, l'identité des peptides est déduite à partir de

méthodes automatisées. Plusieurs algorithmes, qui effectuent cette tâche, procèdent selon l'une des deux stratégies suivantes: la recherche d'ions MS/MS ou le séquençage *de novo*.

1.3.3.1 Recherche d'ions MS/MS

Cette méthode d'interprétation de spectres MS/MS a connue son essor en protéomique suite au séquençage complet du génome d'organismes modèles puisqu'elle requiert une base de données de séquences de protéines. Grâce à cette base de données, il est possible d'identifier les peptides provenant de diverses protéines. Deux éléments sont utilisés pour identifier un peptide: sa masse et la masse de ses fragments. Pour identifier efficacement un peptide à l'aide de la base de données, une digestion *in silico* des protéines selon le procédé expérimental est effectuée au préalable pour générer un index de masse des peptides. Pour chaque MS/MS, l'algorithme de recherche débute par l'interrogation de cet index avec la masse du peptide observé. Cette étape produit une liste de peptides candidats qui ont une masse théorique qui ne s'écarte pas de plus d'une certaine marge d'erreur en fonction de la précision de l'instrument. Chaque candidat est ensuite considéré en vérifiant la corrélation de masse entre les fragments observés et les fragments attendus en considérant aussi une marge d'erreur. Un pointage est attribué à chaque peptide candidat et le meilleur est assigné au spectre MS/MS. Un pointage minimum est appliqué pour filtrer les candidats de faible confiance. L'algorithme retourne à la fin une liste de peptides identifiés avec leur protéine correspondante. L'assignation des peptides aux protéines est parfois ambiguë dans le cas où la séquence de celui-ci est partagée par plusieurs protéines ou isoformes.

Plusieurs implémentations de cet algorithme existent. Celles-ci diffèrent fondamentalement au niveau du modèle de calcul du pointage des spectres MS/MS. Mascot (Matrix Science) emploie une approche probabiliste qui est basée sur la probabilité p d'observer par hasard un peptide et ses fragments en fonction de la tolérance de masse et du nombre de fragments observés [54]. Les détails de ces calculs n'ont jamais été publiés mais les p obtenus corrèlent bien avec une distribution de probabilité binominale. Le score final

de Mascot est calculé par $S = -10 \times \log_{10}(p)$. SEQUEST (Thermo Fisher Scientific) utilise un algorithme de corrélation croisée pour calculer le pointage [55]. Cet algorithme consiste à additionner l'intensité relative des fragments observés s'ils correspondent aux fragments du modèle théorique. Pour pondérer la contribution du bruit, l'auto-corrélation est calculée de la même façon, mais en faisant une translation de masse des fragments du modèle théorique. Le pointage XCorr, retourné par SEQUEST, consiste en la division de la valeur de corrélation croisée par la valeur moyenne de l'auto-corrélation sur un intervalle de 150 Da. Le XCorr sera donc élevé si on a plusieurs fragments avec un signal sur bruit élevé. X!Tandem [56], un outil plus récent, calcule un pointage nommé l' « Hyperscore » qui est calculé ainsi:

$$HyperScore = \left(\sum_{i=0}^{n} I_i \times P_i \right) \times N_b! \times N_y!$$

Équation 2

où I_i est l'intensité d'un ion fragment, P_i est égale à 1 si i est un ion fragment théorique ou 0 dans le cas opposé, N_b est le nombre d'ions b assignés et N_y le nombre d'ions y assignés. Ce score est donc basé sur l'intensité des fragments et le nombre observé. Une valeur d'espérance (« E-value ») est aussi calculée empiriquement pour indiquer la probabilité que l'identification soit incorrecte relativement aux autres N candidats peptidiques. Plusieurs autre algorithmes de recherche existent: OMSSA (NCBI) [57], Protein Prospector [58], Spectrum Mill (Agilent Technologies), etc.

Les résultats d'une recherche d'ions MS/MS sont susceptibles de contenir un certain taux de faux positifs. Pour estimer cette valeur, il est possible d'effectuer la recherche en ajoutant à la base de données de protéines des séquences leurres [59]. Les séquences leurres sont générées en randomisant ou inversant l'ordre des acides aminés des protéines. Les identifications issues des peptides leurres sont utilisées pour estimer le taux de faux positifs. Le taux de faux positifs est calculé en doublant le nombre de peptides faux positifs

(car on considère qu'il y a autant de chance qu'il y ait des faux positifs dans les séquences non-randomisées) et en divisant ensuite ce nombre par le nombre total d'identifications. En pratique, un taux de faux positifs inférieur à un pourcent est toléré dans les données protéomiques et une limite de pointage des peptides est appliquée en conséquence.

Ce type d'interprétation obtient en moyenne un taux de succès d'identification avoisinant les 20-25% (pour un LTQ-Orbitrap) pour l'ensemble des spectres MS/MS acquis. Les sources de ce faible taux sont multiples. De façon générale et indépendante de la stratégie algorithmique sélectionnée. on retrouve les problèmes suivants: l'ion sélectionné n'est pas un peptide, la piètre qualité des spectres MS/MS (faible abondance du précurseur ou mauvaise fragmentation du peptide), des peptides avec une modification chimique non-explorée (augmentation exponentielle de l'espace de recherche pour chaque modification considéré), des peptides avec une extrémité non conforme à la protéase employée ou une digestion incomplète. La recherche d'ions MS/MS est limitée par la base de données de séquences protéiques utilisée. Il est donc impossible de trouver un peptide qui ne s'y trouve pas comme dans les cas suivants: polymorphisme d'un acide aminé, isoforme inconnu, séquence codante non-prédite par les algorithmes de prédiction de régions codantes pour des protéines et des contaminants protéiques provenant d'une autre espèce.

1.3.3.2 Séquençage *de novo*

Le séquençage *de novo* identifie la séquence peptidique en employant seulement la masse du peptide et les fragments contenus dans les spectres MS/MS sans aucune connaissance préalable de la séquence. Aucune base de données de protéines n'est utilisée pour déterminer la séquence des peptides, ce qui comble la problématique principale de la recherche d'ions MS/MS. Cette méthode a par contre un coût de calcul plus élevé que l'autre méthode car elle doit évaluer toutes les possibilités de séquences en fonction des fragments observés. Les algorithmes de séquençage *de novo* procèdent à la construction

d'un graphe pour évaluer ces possibilités [53]. Les sommets du graphe correspondent aux fragments observés et les arcs à un acide aminé. Le problème est de trouver la série de fragments *b* et *y* pour reconstituer la séquence du peptide. Pour résoudre le problème, les sommets sont d'abord ordonnés de façon croissante selon leur masse. Le graphe est construit en évaluant si la différence de masse entre deux pics correspond à la masse d'un acide aminé. Dans un tel cas, un arc étiqueté de l'acide aminé correspondant est ajouté entre deux sommets. Les arcs vont toujours d'un pic de petite masse vers un de plus grande. En théorie, avec le graphe, on devrait retrouver la séquence complète du peptide avec la série *b* et pour la série *y*, la séquence inversée. En pratique toutefois, on s'écarte largement du cas idéal puisque la fragmentation n'est pas uniformément distribuée sur la séquence du peptide. On n'observe seulement qu'une partie des fragments attendus. Dans ce cas, des arcs correspondant à plusieurs acides aminés sont alors ajoutés au graphe en acceptant le coût d'avoir une incertitude sur l'ordre de la séquence. Un élément critique du succès de cette méthode est de reconnaître les fragments peptiques du bruit contenu dans le spectre MS/MS. La présence de sous-fragments, les masses isobariques des acides aminés (I = L) ou les combinaisons (comme W = E + G) compliquent l'interprétation car ils causent des ambiguïtés. Il est très important d'avoir des spectres MS/MS de haute résolution pour éviter de confondre les fragments de masses similaires. Le graphe est donc, en pratique, partiel et il contient plusieurs possibilités. Un pointage intervient alors pour trouver le meilleur chemin dans le graphe pour déterminer la séquence du peptide. L'identification des protéines est finalement faite en cherchant la séquence exacte ou similaire dans une base de données de séquences protéiques. La base de données peut servir à résoudre certaines ambiguïtés de séquence et indiquer la présence de polymorphisme. Parmi les outils qui implémentent cette stratégie, on retrouve Lutefisk [60], PEAKS [61], PepNovo [62], etc.

Il existe aussi une méthode intermédiaire entre le séquençage de *novo* et la recherche d'ions MS/MS nommée « étiquette de séquence » [63]. Le spectre MS/MS est d'abord inspecté manuellement et les masses de quelques fragments sont employées pour

déduire une séquence partielle du peptide. Cette séquence ainsi que la masse du précurseur peptidique sont utilisés pour trouver les peptides potentiels correspondants.

1.3.3.3 Comparaison de spectres

Une dernière procédure pour déterminer le peptide correspondant à un spectre est la comparaison de spectres. Cette méthode est dépendante d'au moins une des deux précédentes méthodes présentées puisqu'elle nécessite la construction d'une base de données de spectres MS/MS-peptide. Un algorithme compare ensuite chaque spectre acquit avec les spectres contenus dans la base de données pour obtenir la séquence du peptide. La corrélation des fragments entre un spectre inconnu et le spectre de référence est calculée et l'identité du peptide est rapportée si la valeur dépasse un seuil prédéterminé par l'utilisateur. Puisque la comparaison de spectres est très rapide (1000 fois plus), cette méthode est employée comme première étape de recherche et les spectres non identifiés sont ensuite soumis à l'une des deux autres méthodes. Les implémentations de l'algorithme de comparaison de spectres sont X!Hunter [64], BiblioSpec [65], Bonanza [66], SpectraST [67], etc. Quelques groupes de recherche rendent disponible publiquement leur librairie de spectre MS/MS annotés (GPM [68], PeptideAtlas [69]).

1.3.3.4 Particularité pour les phosphopeptides

Considérer les modifications post-traductionnelles lors de l'interprétation des spectres MS/MS n'implique pas de changement important dans les algorithmes précédents. Pour la phosphorylation, il suffit d'ajouter en surplus dans le modèle théorique les acides aminés sérine, thréonine et tyrosine en version phosphorylée, donc avec +80 Da supplémentaire. Ensuite, pour chaque peptide contenant un ou plusieurs de ces acides aminés, l'algorithme doit itérer sur les différentes combinaisons possibles. Comme la perte de neutre de H_3PO_4

est très importante dans le cas des phosphopeptides, certain des algorithmes précédents considèrent aussi la perte de 98 Da pour améliorer le pointage accordé aux spectres.

Toutefois, une limitation avec les outils d'interprétation des spectres MS/MS pour la recherche de phosphopeptides est l'absence d'indicateur révélant que le site a été localisé avec confiance. La localisation du site phosphorylé peut être ambiguë si le peptide est composé de plusieurs sérines/thréonines/tyrosines et que le patron de fragmentation du peptide ne fournit pas suffisamment d'indices. Ceci est particulièrement fréquent avec les phosphopeptides puisque le lien phosphoester est plus labile que le lien peptidique. Il est aussi hautement probable que des isomères positionnels du phosphopeptide (la même séquence mais phosphorylée à des positions différentes) coéluent créant ainsi un spectre MS/MS avec une mixture des formes. Ce spectre mixte est donc difficile à interpréter. Des algorithmes ont été développés pour déterminer la probabilité de localisation. Toutes les méthodes proposées cherchent la présence d'ions fragments spécifiques dans les spectres MS/MS et calculent une probabilité avec une distribution binomiale cumulative:

$$P = \sum_{k=n}^{N} \binom{N}{k} p^k (1 - p)^{N-k}$$

Équation 3

où N est le nombre d'ions fragments potentiels du peptide, n est le nombre d'ions fragments attribués à un pic dans le spectre MS/MS et p la probabilité d'assigner un ion à un pic. Ascore [70] a été le premier outil à implémenter cette méthode pour localiser la phosphorylation. MaxQuant [71], PhosphoScan [72] et PhosCalc [73] intègrent en plus l'information des ions fragments provenant de spectres MS^3 lorsque disponible et supportent différent formats de données. SLoMo [74], l'outil le plus récent, considère en plus les ions *c* et *z* issus de la fragmentation ETD. Une autre stratégie est aussi utilisée pour assigner la confiance de localisation. Celle-ci consiste à utiliser la liste de candidats peptidiques phosphorylés proposés par les outils de recherche (différentes positions phosphorylées suggérées) et leur différence de pointage pour calculer la confiance de localisation [19, 75]. La distinction entre le signal et le bruit dans un spectre ainsi que

l'assignation erronée d'un pic (hautement probable avec les spectres MS/MS de basse résolution couramment utilisés) peut mener à de fausses interprétations avec ses algorithmes. Finalement, une combinaison des outils précédemment énumérés est couramment employée pour traiter les données de phosphoprotéomique afin d'obtenir une liste de sites phosphorylés avec un faible taux de faux positifs et une haute confiance de localisation.

1.3.4 Protéomique quantitative sans marquage

La quantification relative sans marquage requiert l'emploi de logiciels pour extraire l'abondance des pics des données MS. Deux étapes sont nécessaires pour obtenir ces valeurs d'abondance: la détection des ions peptidiques et l'alignement des peptides entre les conditions (Figure 1.12). La première étape consiste à utiliser les données brutes des spectres MS pour détecter les ions peptidiques. Pour chaque peptide, il faut obtenir sa valeur m/z, son temps de rétention et son abondance. Ces valeurs servent de coordonnées pour comparer les peptides entre les différentes conditions. MassSense, un logiciel développé à notre laboratoire qui est basé sur l'algorithme THRASH [76], est employé dans ce travail pour effectuer cette tâche. L'algorithme recherche dans un groupe de pics d'ions le centroïde bidimensionnel (*m/z* et temps) qui définit un peptide. L'algorithme comprend quatre étapes: 1) la détection des pics dans chaque spectre de masse pour éliminer le bruit, 2) la construction de pistes d'élutions en groupant les pics qui sont détectés dans des spectres consécutifs, 3) le groupement des pistes d'élutions qui forment un profil isotopique, et 4) la sélection des candidats peptidiques par comparaison avec un modèle isotopique basé sur l'averagine (acide aminé *in silico* avec la masse moyenne des 20 acides aminés) et la méthode des moindres carrés. Une fois les peptides détectés, l'algorithme retient la valeur *m/z* du premier pic de la série, soit le pic monoisotopique. La charge du peptide est déterminée avec le profil isotopique (espacement *1/z*). La valeur d'abondance est obtenue soit en intégrant l'aire sous la courbe du pic ou en cherchant le sommet

d'intensité maximal du peptide au cours de son élution. Le sommet d'abondance correspond au temps de rétention du peptide. Les valeurs de chaque peptide détecté dans un échantillon sont rapportées sous forme de carte peptidique.

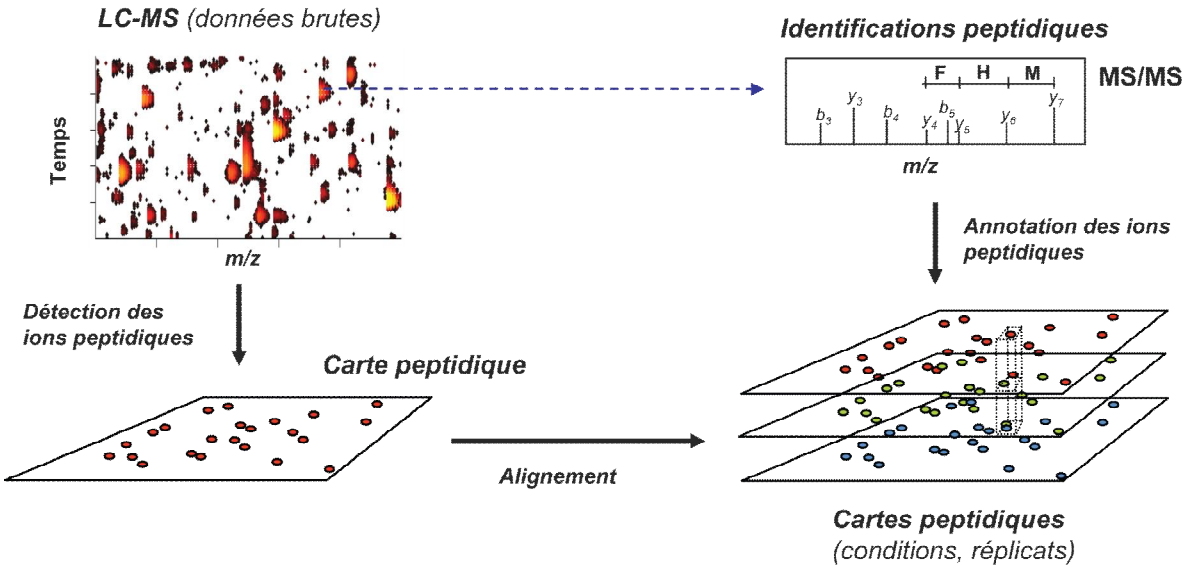


Figure 1.12: Traitement des données pour la quantification MS sans marquage.

Ce schéma illustre les différentes étapes algorithmiques nécessaires pour passer des données brutes de spectrométrie de masse à une liste où chaque peptide identifié est quantifié.

Les cartes peptidiques des conditions ainsi que celles des réplicats sont alignées par notre logiciel ProteoProfile les unes aux autres pour comparer l'abondance des peptides. L'algorithme d'alignement recherche itérativement pour tous les peptides identifiés leur présence dans chaque carte peptidique. La recherche d'un peptide candidat dans chaque carte est effectuée dans une fenêtre de valeurs *m/z* et de temps de rétention prédéfinis par l'utilisateur. Finalement, un vecteur des valeurs d'abondance pour chaque condition est retourné pour chaque peptide. Des données de haute résolution et une chromatographie reproductible diminuent le risque d'erreur lors de l'alignement. Lorsque les données sont complexes (bruit élevé, nombre important de peptides, peptides isobariques,

chevauchement du signal des peptides), il est plus risqué que le signal corrélé entre les conditions ne provienne pas du même peptide. La validation manuelle est nécessaire pour éviter les erreurs dans cette situation. D'autres données sont aussi employées pour augmenter la confiance de l'alignement comme la charge du peptide, son environnement ou la différence d'abondance (pour les réplicats). Suite à l'alignement, l'abondance des peptides est normalisée afin de compenser pour les variations techniques. La normalisation modifie les valeurs afin que la moyenne d'abondance de la population des peptides de chaque échantillon soit identique. Cette procédure est applicable seulement s'il y a peu de différences d'expression entre les échantillons. Les cartes peptidiques alignées sont finalement annotées avec les peptides identifiés par MS/MS. À partir de ce point, les données quantitatives sont prêtes afin d'effectuer les calculs pour trouver les variations d'abondance. Étant donné que l'échantillonnage MS/MS des peptides est incomplet ou parfois infructueux, plusieurs ions peptidiques détectés restent inconnus. Les cartes peptidiques peuvent être utilisées pour générer des listes d'inclusions pour une analyse MS/MS subséquente pour identifier les ions inconnus.

On peut déterminer les changements d'abondances des peptides en calculant des ratios d'abondance moyenne entre les conditions. Un test-t de Student est ensuite employé pour déterminer si les variations observées sont significatives entre une condition contrôle *c* et une condition traitée *t* (l'hypothèse nulle est $\mu_c = \mu_t$). La statistique *t* est calculée comme suit:

$$t = (\mu_c - \mu_t) \Big/ \sqrt{\frac{s_c^2}{n_c} + \frac{s_t^2}{n_t}}$$

Équation 4

où μ, *s* et *n* correspondent respectivement à la moyenne, la variance et le nombre de réplicats. La distribution t de Student permet de déterminer la probabilité d'observer la valeur *t* obtenue et de calculer la valeur *p*, soit la probabilité d'obtenir la même valeur ou une valeur plus extrême.

Les données de quantification relative sans marquage obtenues au cours de cette thèse ont été traitées avec l'implémentation de cette stratégie de notre laboratoire, soit MassSense et ProteoProfile. Plusieurs outils de quantification sans marquage en source libre existent: SpecArray [77], MsInspect [78], MSight [79], TOPP [80], PEPPeR [81] et SuperHirn [82]. Des versions commerciales sont aussi disponibles: QuantLynx (Waters), SIEVE (Thermo Fischer Scientific), Elucidator (Rosetta Biosoftware) et Expressionist (GeneData). Ces logiciels se différencient selon les algorithmes quantitatifs implémentés, les formats de fichier supportés, l'intégration à des plateformes d'analyses et l'analyse statistique effectuée [83] (voir cette référence pour une comparaison des fonctionnalités).

1.4 Bio-informatique appliquée à la phosphoprotéomique

L'accumulation des données de phosphoprotéomique a stimulé le développement d'outils informatiques pour comprendre les divers aspects du phosphoprotéome. L'accumulation de sites phosphorylés dans les bases de données telles Swissprot [84], PhosphoSite [85], Phospho.ELM [86], PHOSIDA [87], PhosphoPep [88], a engendré le développement de plusieurs outils de prédiction de sites de phosphorylation. Ces différentes méthodes prédisent si un site peut être phosphorylé et certaines classifient en plus les sites selon les kinases potentiellement impliquées. Parmi ces méthodes, PROSITE [89] et Scansite [90] utilisent les patrons et profils de position, NetPhos [91] un réseau de neurones, KinasePhos [92] un modèle de Markov caché, PredPhospho [93] les machines à vecteurs de support (SVM) et PPSP [94] la théorie de décision Bayésienne. La spécificité et la sensibilité des outils de prédictions sont souvent critiquées puisque ces algorithmes sont biaisés par divers facteurs. Entre autre, l'ensemble d'exemples négatifs employés pour l'entraînement contient des S/T/Y pouvant être phosphorylés *in vivo* mais non rapportés pour le moment. Aussi, certains outils sont entraînés exclusivement avec des données d'expériences *in vitro* ou encore un nombre insuffisant d'exemples pour certaines kinases. De plus, ces outils ne considèrent pas l'accessibilité des sites prédits qui peuvent être enfouis à l'intérieur de la

structure protéique ou la possibilité que la kinase et le substrat ne partagent pas le même compartiment cellulaire. Seul NetworKIN [95] utilise les données d'interactions protéine-protéine afin d'augmenter la spécificité de prédiction. Il est probable qu'une description tridimensionnelle des motifs donnerait une spécificité supplémentaire mais le nombre de structures disponibles est insuffisant pour entraîner un classificateur.

Des algorithmes de recherche de motifs de phosphorylation sur-représentés dans les études de phosphoprotéomique ont aussi été développés afin de découvrir la spécificité de kinases. Autrement dit, le problème est de choisir un ensemble de motifs $\mathcal{M}$ qui distinguent l'ensemble des sites identifiés de ceux non-phosphorylés dans le protéome. À cette fin, Motif-X [96] emploie un algorithme glouton itératif et MoDL [97] utilise le principe de la «description de longueur minimum » tiré de la théorie de l'information. Pratiquement, le premier outil détermine de nombreux motifs avec de la redondance et à l'opposé, le dernier n'en retourne que quelques uns (< 5). La recherche avec ces outils de motifs enrichis dans les données de phosphoprotéomique a été faite pour plusieurs espèces dont la levure, la mouche, la souris et l'humain. Ces outils ont servi à la découverte de nouveaux motifs de phosphorylation dont le motif S/T-Q associé aux substrats des kinases ATM/ATR [98]. Les nouveaux motifs découverts par ces algorithmes pourront aussi servir à la prédiction future de sites de phosphorylation.

D'autres outils et ressources bio-informatiques, non conçus spécifiquement pour l'étude de la phosphorylation, sont aussi utiles pour comprendre le rôle et le mécanisme d'action des évènements de phosphorylation. Les répertoires (SwissProt [84], PhosphositePlus [85]) et outils de prédictions d'O-glycosylation (NetOGlyc [99], YingOYang [100] and OGlcNAcScan [101]) peuvent indiquer une compétition entre la glycosylation et la phosphorylation pour les sérines et les thréonines. Les bases de données pour d'autres modifications sont aussi pertinentes pour évaluer leur association à la phosphorylation pour le contrôle de processus biologique. Au niveau fonctionnel, les annotations Interpro [102] des domaines protéiques peuvent indiquer quelle fonction est modulée par la phosphorylation. Au niveau moléculaire, le rôle de la phosphorylation peut

être suggéré à partir des structures protéiques tridimensionnelles répertoriées par la base de donnée PDB [103]. Celles-ci permettent d'observer les interactions intra ou intermoléculaires entre le site phosphorylé et les autres résidus via des interactions électrostatiques ou un encombrement stérique. Évaluer la conservation des sites au cours de l'évolution avec des alignements multiples de séquences protéiques provenant de différents organismes peut révéler l'importance de la phosphorylation. Cette information est aussi pertinente pour déterminer si l'activité d'un site phosphorylé est transposable chez l'homme et peut être une cible thérapeutique potentielle s'il y a une association avec une maladie. Bref, l'intégration des diverses ressources et outils bio-informatiques aux résultats de phosphoprotéomique est primordiale pour émettre de nouvelles hypothèses fonctionnelles et comprendre le rôle individuel de chaque site.

Toutefois, malgré la disponibilité des ressources et des outils bio-informatiques énumérés ici, il est ardu, dans un contexte de phosphoprotéomique, de les utiliser car chaque outil doit être exécuté manuellement pour chaque liste de protéines identifiées ou voire même pour chaque protéine. De plus, intégrer toutes ces ressources en un seul document pour une étude approfondie des données est aussi une tâche complexe. Une plateforme d'analyse bio-informatique dédiée à la phosphoprotéomique était donc nécessaire pour simplifier et automatiser ces tâches afin d'aider à l'interprétation des données (Chapitre 2). Au cours de cette thèse, deux plateformes, PTMScout [104] et SysPTM [105], ont été développées parallèlement à la nôtre. Elles intègrent plusieurs outils et bases de données dédiés à la phosphoprotéomique. Ceux-ci ne traitent par contre pas les données d'identifications brutes des peptides et des protéines obtenus suite à l'interprétation automatisée des spectres MS/MS. Somme toute, le développement de ce type de plateforme d'analyse bio-informatique est crucial pour gérer la complexité des données de phosphoprotéomique, pour accélérer la recherche en phosphoprotéomique et en extraire connaissances biologiques.

1.5 Implication de la voie de signalisation Erk1/2 dans le cancer

1.5.1 Qu'est-ce que le cancer?

Le cancer est la première cause de mortalité au Canada. Approximativement 45% des Canadiens recevront ce diagnostique au cours de leur vie. Selon les estimés de la Société canadienne du cancer, 2011 sera marqué par 177 800 nouveaux patients diagnostiqués et 75 000 décès. Pathologiquement, le cancer est caractérisé par une croissance cellulaire rapide, invasive et non contrôlée qui peut survenir dans tous les tissus cellulaires ou organes. Après plusieurs cycles de division, les cellules forment une masse cellulaire appelée tumeur (sauf pour les leucémies) qui envahie les tissus avoisinants. Le déplacement des cellules cancéreuses dans l'organisme peut entraîner la formation de nouveaux foyers cancéreux, les métastases. Les cellules cancéreuses sont insensibles aux signaux qui empêchent les cellules normales de grandir et évitent aussi les mécanismes de mort cellulaire programmés (apoptose). Cette maladie se produit par l'accumulation au fil du temps de mutations sur divers gènes et éléments régulateurs. Plusieurs rondes de sélection naturelle sélectionnent et amplifient les cellules qui en possèdent [106]. Ces mutations sont majoritairement causées par des facteurs environnementaux (tabac, diète, forme physique, infection, pollution, vieillesse) et plus faiblement d'origine génétique (mutations héréditaires, erreurs dans la réplication de l'ADN). La combinaison des mutations correspond à la signature moléculaire du cancer. Chaque cancer a un ensemble particulier de mutations et ceci complique le traitement. Toutefois, certaines mutations semblent être présentes à une fréquence plus élevée dans certains types de cancer. Les kinases et phosphatases sont deux familles de protéines où l'on retrouve fréquemment des mutations dans les cancers humains [107].

1.5.2 La voie de signalisation Erk1/2

La régulation de processus biologiques par la phosphorylation nécessite plusieurs évènements de phosphorylation successifs qui constituent un modèle que l'on nomme voie de signalisation. La voie de signalisation Erk1/2 (« extracellular signal-regulated kinase ») ou Map kinase (« mitogen-activated protein ») régule diverses fonctions moléculaires comme la traduction, la progression du cycle cellulaire, le réarrangement du cytosquelette, l'activation de la transcription de certains gènes, etc. Au niveau des processus biologiques, elle est impliquée dans le développement, le métabolisme du glucose, la réponse immunitaire et la mémoire. Cette voie est caractérisée par une cascade de phosphorylations où une kinase en phosphoryle une autre jusqu'à la phosphorylation des protéines effectrices en aval (voir schéma Figure 1.13). La cascade sert à propager une information extracellulaire qui est perçue par un récepteur au niveau de la membrane cellulaire et transmise jusqu'aux protéines ciblées.

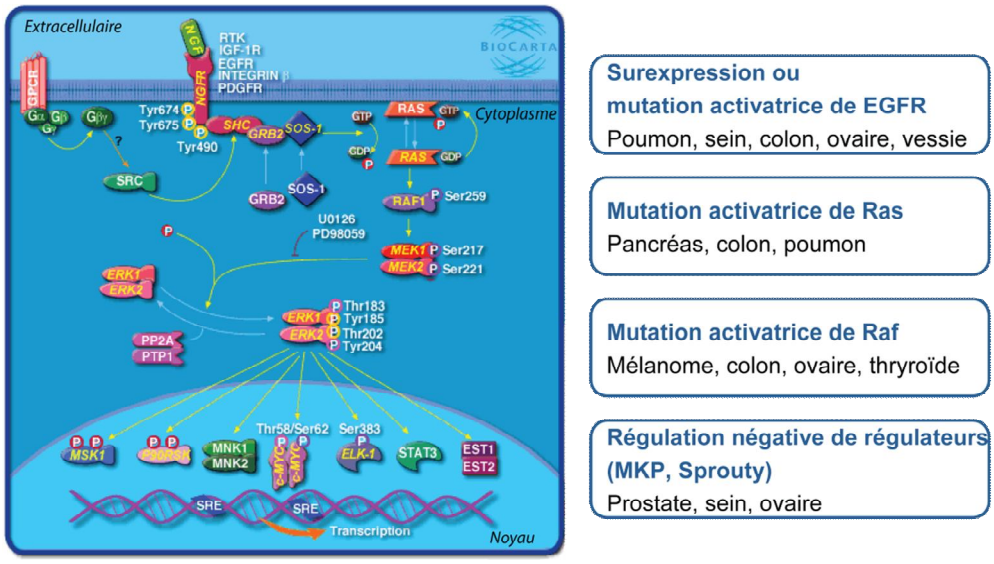


Figure 1.13: Voie de signalisation Erk1/2 et problèmes moléculaires dans le cancer.

Schéma illustrant la cascade de phosphorylations de la voie de signalisation Erk1/2 qui survient lorsque le récepteur membranaire est activé. Adaptation de « Erk1/Erk2 Mapk Signaling pathway » (Michael Shih, Ph.D) avec la permission de BioCarta Inc., 2010

Plus en détails, la cascade de phosphorylation est activée par la liaison d'un ligand à un récepteur (Rtk, Igf-1r, Egfr, intégrine, Pdgfr, Ngfr) situé à la membrane plasmique. Les récepteurs sont activés par des stimuli comme des facteurs de croissances (Egf, Pdgf), des esters de phorbol (PMA, TPA), des cytokines (Tgf), des hormones (insulines), des mitogènes, les ligands des récepteurs couplés aux protéines G et le stress osmotique. L'activation des récepteurs tyrosines kinases par un ligand entraîne une autophosphorylation et le recrutement du complexe de couplage Shc/Grb2/Sos via un domaine SH2. Ce complexe recrute ensuite la petite GTPase Ras qui échange alors son GDP pour un GTP. À son tour Ras actif recrute Raf, une Map3k. Il y a trois isoformes de Raf (A, B, C). Chez la souris, l'inactivation des gènes B-Raf ou C-Raf est létal et celle de A-Raf entraîne des problèmes de développement neuronal et gastro-intestinal. B-Raf est la forme majoritaire qui phosphoryle et active les kinases Mek1/Mek2 (« Map kinase/Erk kinase 1/2 » ou Mapkk). Mek1 est considérée comme la kinase majoritairement active malgré que les deux isoformes partagent 80% d'identité de séquence. $Mek1^{-/-}$ est létal tandis que $Mek2^{-/-}$ semble n'avoir aucun phénotype particulier chez la souris [108]. Les kinases Mek1/2 sont des kinases à double spécificité, c'est-à-dire qu'elles phosphorylent autant les sérines/thréonines que les tyrosines. Cette spécificité leur permet d'activer les kinases Erk1/2 (Mapk) en phosphorylant doublement le motif activateur TGY. Cette double phosphorylation induit un changement conformationnel, un réarrangement des liens hydrogènes au site de liaison du substrat, qui active les kinases. Erk1/2 sont les kinases principales de cette voie de signalisation. Elles partagent 84% d'identité de séquence et sont conservées de la levure à l'homme. Erk1 est absent chez le poulet et une seule forme est présente chez les invertébrés. Les mutants $Erk1^{-/-}$ chez la souris sont viables, fertiles mais de tailles réduites (prolifération réduite et entrée lente en phase S du cycle cellulaire). $Erk2^{-}$

$^{/-}$ est létal, car elles engendrent des défauts au niveau cardiaque, placentaire et du mésoderme. Ces kinases sont désactivées par l'action des phosphatases Pp2a, Ptp1 et Dusp.

Les kinases Erk1/2 phosphorylent les sérines et thréonines suivies d'une proline (p[ST]P) et avec une plus haute efficacité lorsque qu'une proline est présente à la position -2 du site phosphorylé (PXp[ST]P). Outre le site de liaison du substrat, des domaines d'interactions additionnels participent à la sélectivité des substrats. Le motif CD (« common docking »), séquence longue de 8 acides aminés sur les Erks, interagit avec le motif de liaison D (ou DEJL) qui a comme séquence consensus Arg/Lys_2-Xaa_{2-6}-Φaa-Xaa-Φaa (où Xaa est n'importe quel acide aminé et Φaa est un des résidus hydrophobes suivant: Leu, Ile et Val). Le motif D est localisé généralement à moins de 20 acides aminés du site phosphorylé. Le motif CD forme un site d'ancrage permettant de former des complexes kinase/substrat avant même que la kinase soit active. Ce pré-assemblage augmente la vitesse de la cinétique de phosphorylation des substrats. Un second motif de liaison connu est le DEF (« Docking site for Erk »). Ce court motif, localisé environ à 10 acides aminés du côté C-terminal du site phosphorylé, est formé par Phe-Xaa-Phe-Pro et se lie dans une poche hydrophobique sur Erk1/2 près du site catalytique. Cette région est seulement accessible lorsque les kinases Erk1/2 sont actives. Ces deux motifs de liaisons ne sont pas présents sur tous les substrats connus. Leur présence augmente par contre l'efficacité de la phosphorylation.

Présentement, Erk1/2 sont connues pour phosphoryler environ 160 substrats [109]. Il n'a pas été démontré si Erk1 et Erk2 régulent chacun un ensemble commun ou distinct de sites phosphorylés. Les protéines phosphorylées par Erk1/2 peuvent varier selon les conditions cellulaires, les différents compartiments, le type de cellules et les espèces. Environ la moitié des substrats ont été trouvés dans le noyau, et le reste est partagé entre le cytosol, la membrane plasmique et d'autres organelles. Suite à l'activation des kinases Erk1/2 par différents stimuli, la phosphorylation des divers substrats peut s'effectuer selon une dynamique temporelle rapide où le maximum de phosphorylation est atteint en quelques minutes, ou tardive, maximum en plus d'une heure. Les substrats modulés par les

kinases Erk1/2 sont associés à plusieurs classes de protéines impliquées dans différents processus biologiques. Les facteurs de transcription forment, avec leurs cinquante substrats connus, la classe de protéines comportant le plus grand nombre de substrats. La phosphorylation module ici l'expression des gènes soit en augmentant l'affinité de liaison du facteur pour l'ADN (par exemple: Elk1), soit en recrutant des co-activateurs via des interactions protéine-protéine dépendant de la phosphorylation ou soit en stabilisant la protéine pour éviter sa dégradation (c-Fos). L'expression de ces gènes est importante pour le contrôle de la prolifération, de la différentiation et une régulation inadéquate peut mener à des transformations oncogéniques. Vingt-deux kinases (dont les familles Rsk et Mapkapk) et phosphatases (dont la famille Dusp) sont phosphorylées par les kinases Erk1/2. La modulation de celles-ci prolonge la cascade de phosphorylation qui modifie conséquemment une partie du phosphoprotéome. La famille Rsk est un relai de signalisation important pour la prévention de l'apoptose et l'induction du cycle cellulaire. Les kinases Erk1/2 phosphorylent aussi vingt protéines du cytosquelette (par exemple la paxilline). Ces protéines sont importantes pour la prolifération, l'adhésion, la morphologie et la motilité cellulaire. La phosphorylation de ces protéines sert principalement au recrutement de protéines pour la signalisation intra ou extracellulaire. Pour le reste des substrats connus des kinases Erk1/2, on trouve 24 protéines impliquées dans la signalisation, 9 dans l'apoptose et 21 avec diverses autres fonctions. L'ensemble complet des substrats d'Erk1/2 et leurs sites associés ne sont pas connus pour le moment et d'autres études sont nécessaires pour comprendre l'étendue des effets de l'activité de ces kinases.

Les découvertes des substrats des kinases Erk1/2 ont été faites, pour la plupart, par des études individuelles de protéines [109]. Quelques études ont employé des approches à large échelle par spectrométrie de masse pour les identifier. Les substrats potentiels ont été ciblés soit par l'usage d'anticorps phosphospécifique [110], en mesurant par quantification sans marquage [111], SILAC [110], par l'électrophorèse 2D différentielle en fluorescence [112], ou par l'inhibition de la phosphorylation en utilisant un inhibiteur de Mek1/2 (U0126). Récemment, une kinase synthétique Erk2 sensible à un analogue de l'ATP a aussi

été utilisée *in vitro* pour marquer et isoler les substrats de cette kinase [113]. Les substrats potentiels obtenus par ces différentes études démontrent un faible taux de recouvrement entre elles et aussi avec la liste des substrats connus (voir Figure 3.4). Les différents types cellulaires, les conditions biologiques et les différentes procédures analytiques employées ont possiblement mené à un échantillonnage différent du phosphoprotéome. Étant donné qu'il est long et ardu d'effectuer une validation systématique de chaque substrat candidat, seulement quelques nouveaux substrats de ces études s'ont venu s'ajouter à la liste des substrats connus des kinases Erk1/2. Des efforts supplémentaires de recherche sont donc nécessaires pour compléter cette liste.

1.5.3 Intérêt clinique

La dérégulation et la présence de mutations sur les membres de la voie Erk1/2 sont fréquemment associées à différentes pathologies humaines comme le diabète, les maladies cardiovasculaires et principalement le cancer (voir Figure 1.13). Les deux tiers des tumeurs solides de cellules épithéliales sur-expriment le récepteur, un tiers des cancers possèdent une mutation de Ras et deux tiers une mutation dans la protéine B-Raf pour les mélanomes [114]. Ces problèmes engendrent une sur-activation de la voie Erk1/2 dans les tumeurs. Ces observations ont stimulé le développement d'inhibiteurs pharmacologiques comme agents anticancéreux [107]. Des effets antiprolifératifs, antimétastatiques et antiangiogéniques sont attendus de ces agents selon les observations précédentes. Des petites molécules ciblant et inhibant les kinases EGFR (Tarceva®OSI-Pharmaceuticals, Iressa®AstraZeneca), Raf (ISIS5312®Isis, L-779®Merck, BAY 43-9006® Bayer) et MEK (CI-1040®Pfizer) ont été développées. Plusieurs molécules sont en essais cliniques tandis que quelques unes ont été approuvées jusqu'à maintenant.

1.6 Défis futurs en phosphoprotéomique

Pour conclure cette introduction, quelques questions ouvertes et problématiques en phosphoprotéomique seront abordées en traitant l'aspect analytique, informatique et biologique. Pour étudier le phosphoprotéome et repousser les limites de notre compréhension, il faut d'abord être capable d'identifier et quantifier les sites de phosphorylation. Ceci requiert une variété de moyens techniques. Les récentes percées de la spectrométrie de masse ont permis un énorme élargissement de notre connaissance du phosphoprotéome qui a permis de cataloguer des milliers de sites de phosphorylation. Malgré ces chiffres impressionnants, les expériences actuelles n'identifient qu'une fraction de l'ensemble du phosphoprotéome [2] dû à plusieurs limitations techniques (sensibilité, gamme dynamique du détecteur, rapidité d'échantillonnage de l'instrument insuffisante, exploration restreinte de l'espace de recherche pour l'interprétation automatique de spectres MS/MS, taille incompatible de certains peptides générés par la digestion enzymatique, biais des méthodes d'enrichissement des phosphopeptides, etc.). Donc, l'ensemble des sites de phosphorylation présentement détecté limite la compréhension de la régulation d'une protéine par plusieurs sites phosphorylés ou encore l'interprétation des voies de signalisation. Sur le plan analytique, la phosphoprotéomique devra être plus sensible pour détecter l'ensemble des phosphorylations présentes ou sinon pouvoir cibler les protéines d'intérêts dans le but de comprendre les voies de signalisation.

Jusqu'à présent, le phosphoprotéome a principalement été étudié de façon statique. La phosphorylation est toutefois un processus dynamique. Elle régule l'activité des protéines qui sont organisées sous forme de voies de signalisation qui permettent d'adapter rapidement le système cellulaire en fonction de différentes conditions. Les méthodes quantitatives présentées plus tôt permettent désormais de quantifier la variation d'abondance de phosphorylation pour plusieurs milliers de sites simultanément. Ce type de données relativement récent ouvre donc la porte à la modélisation informatique pour améliorer notre compréhension des réseaux de signalisation cellulaire. Les modèles

informatiques devraient aussi intégrer les autres modifications post-traductionnelles puisqu'elles sont toutes aussi importantes dans la logique de la signalisation. Récemment, il a été rapporté que la phosphorylation peut agir en synergie avec d'autres types de modifications post-traductionnelles pour réguler l'action des protéines [115]. Par exemple, il y a compétition entre la phosphorylation et la glycosylation pour la même sérine/thréonine. Considéré dans un contexte de modélisation des voies signalisations, cette compétition pourrait agir comme une porte logique « OU exclusif». La présence de modifications supplémentaires (phosphorylation, glycosylation, ubiquitination, acétylation, etc.) sur d'autres résidus pourrait correspondre à une porte logique « ET » ou « OU ». La présence de condition « SI » est nécessaire pour l'activité de kinase de Gsk3. La phosphorylation par la kinase Gsk3 est conditionnelle à la présence d'une pré-phosphorylation du substrat [110]. Le concept de modélisation de la signalisation considérant plusieurs modifications émerge tranquillement en parallèle avec le développement des méthodes analytiques. Elles sont limitées actuellement par l'accès aux données nécessaires pour explorer cet aspect de façon globale.

Une autre difficulté rencontrée pour modéliser adéquatement les voies de signalisation est le peu d'information disponible sur l'activité des kinases et phosphatases. Il est difficile de déterminer expérimentalement et informatiquement quelles kinases et phosphatases catalysent l'ajout et le retrait du site *in vivo*. Les outils informatiques de prédiction de kinases ne sont pas suffisamment spécifiques en pratique et identifient plusieurs kinases potentielles pour un même site. Ils sont aussi inefficaces pour les kinases non caractérisées et les espèces éloignées de celles utilisées pour l'entraînement du classificateur. À l'opposé, aucun outil n'existe présentement pour prédire les substrats potentiels des phosphatases puisque très peu de motifs reconnus par les phosphatases sont répertoriés [116].

Finalement, bien que la liste des sites de phosphorylation soit incomplète, sa taille est suffisamment importante pour rendre ardue la tâche de faire ressortir les sites pertinents pour une condition cellulaire particulière. Pour la majorité des sites identifiés, leur fonction

biologique est inconnue et le demeure tant qu'ils ne sont pas investigués un par un avec les méthodes laborieuses traditionnelles de biochimie et de biologie moléculaire. Il n'est pas clair si les sites identifiés ont tous une fonction ou s'ils sont seulement présents dû à l'activité des kinases qui phosphorylent toutes les régions qu'elles reconnaissent et qui sont accessibles. La conservation évolutive de la phosphorylation pourrait contribuer à identifier quels sites sont fonctionnellement importants [5]. Ultimement, la combinaison des connaissances du rôle spécifique de chaque phosphorylation, de l'état du phosphoprotéome sous une condition particulière et des modèles informatiques de signalisation pourrait permette de prédire la réponse cellulaire. Ce type d'application pourrait être fort utile pour comprendre les effets bénéfiques et néfastes d'un médicament en développement.

1.7 Objectifs de cette thèse

L'objectif de cette thèse était d'identifier de nouveaux substrats des kinases Erk1/2. Pour y parvenir, l'analyse phosphoprotéomique quantitative de la cinétique du phosphoprotéome a été effectuée suite à la stimulation de cellules intestinales épithéliales de rat et l'inhibition pharmacologique de la voie Erk1/2. Considérant l'importance biologique et clinique de cette voie, il est primordial de déterminer l'étendue de l'activité enzymatique des kinases Erk1/2. Ce projet de grande envergure a été divisé en deux objectifs principaux.

Le premier objectif de cette thèse était le développement d'une plateforme d'analyse bio-informatique conçue pour faciliter et accélérer l'étude du phosphoprotéome. Cette plateforme devait d'abord gérer le volume de données générées par cette expérience phosphoprotéomique par spectrométrie de masse. Ensuite, celle-ci devait intégrer les outils nécessaires pour résoudre les 2 problèmes suivants. Le premier problème rencontré était de définir les critères nécessaires pour extraire des données de spectrométrie de masse l'ensemble de sites phosphorylés et de s'assurer qu'il n'y a aucune ambigüité sur la localisation des sites de phosphorylation. Ce problème nécessitait de déterminer l'ensemble

des phosphopeptides identifiés avec un taux de faux positifs acceptable et aussi d'implémenter un algorithme pour calculer la probabilité de localisation des sites de phosphorylation adapté à l'outil d'interprétation des spectres MS/MS utilisé. Deuxièmement, l'étude du phosphoprotéome se déroule dans un contexte exploratoire et peu de connaissances sont disponibles sur la plupart des sites de phosphorylation identifiés. Cette situation limite conséquemment l'interprétation des données de phosphoprotéomique. Pour gérer ce problème, des informations supplémentaires provenant de diverses bases de données et outils bio-informatiques ont été incorporées aux sites de phosphorylation identifiés dans le but d'émettre de nouvelles hypothèses sur la fonction de chacun pour réguler l'activité des protéines. Ces annotations couvrent les kinases potentielles, le contexte structural, la proximité d'un domaine protéique, la conservation au cours de l'évolution, la médiation potentielle d'une interaction protéique phospho-dépendante et la compétition avec d'autres modifications. L'accomplissement de ce premier objectif a fourni les outils nécessaires pour l'étude ponctuelle de la phosphorylation sur une protéine ainsi que pour l'étude globale du phosphoprotéome.

Le deuxième et principal objectif était d'effectuer une expérience de cinétique du phosphoprotéome lors de la stimulation et l'inhibition pharmacologique de la voie Erk1/2 afin d'identifier les substrats des kinases Erk1/2. L'atteinte de cet objectif nécessitait donc le développement d'un algorithme basé sur les motifs et les profils cinétiques d'abondance de phosphorylation pour sélectionner les substrats potentiels. Le problème à résoudre était de cibler, avec les profils cinétiques obtenus par la phosphoprotéomique quantitative, les sites de phosphorylation avec le motif consensus des kinases Erk1/2 dont l'abondance est régulée positivement suite à la stimulation et négativement suite à l'inhibition pharmacologique de la voie Erk1/2. Il était donc nécessaire de déterminer comment optimiser les paramètres de l'algorithme effectuant la détection des peptides pour la quantification sans marquage, normaliser les valeurs d'abondances pour corriger les variations expérimentales, contrôler la qualité des données et déterminer les variations de phosphorylation significatives.

Finalement, un autre sujet de recherche a été abordé lors de cette thèse suite à l'analyse des données issues de l'expérience de phosphoprotéomique sur le rat. Il a été observé qu'un certain nombre de phosphopeptides sont présents sous divers isomères positionnels, ou autrement dit, des peptides avec la même séquence mais phosphorylés à différentes positions. La fréquence d'occurrence de ces isomères dans un échantillon et l'efficacité des méthodes de séparations utilisées en protéomique sont inconnues. L'objectif fixé pour cet aspect a été de développer des algorithmes pour faciliter la détection des phosphopeptides isomériques positionnels avec leur profil d'élution LC-MS, pour ceux séparés par chromatographie liquide, ou avec leur spectre MS/MS, pour ceux qui ne le sont pas. Ces algorithmes permettront donc de mieux caractériser ces phosphopeptides isomériques peu étudiés en phosphoprotéomique.

1.8 Organisation des chapitres

Le **CHAPITRE 1: Introduction** décrit la théorie élémentaire de la phosphoprotéomique, les méthodes analytiques et de séparations, la spectrométrie de masse, la protéomique computationnelle, la bio-informatique appliquée à la phosphoprotéomique, la voie de signalisation Erk1/2 et son implication dans le cancer, et les défis futurs de la phosphoprotéomique. Cette section se termine par les objectifs de cette thèse.

Le **CHAPITRE 2: ProteoConnections: a bioinformatics platform to facilitate proteome and phosphoproteome analyses** présente l'article publié dans **Proteomics** [117]. ProteoConnections est une nouvelle plateforme d'analyse bio-informatique dédiée à l'interprétation des données issues d'expériences de phosphoprotéomique générées par spectrométrie de masse. La plateforme sert d'abord de base de données et de support pour évaluer la qualité des données. Le point fort de cette plateforme est l'intégration d'une multitude d'outils bio-informatiques pour faciliter l'exploration du phosphoprotéome afin de proposer de nouvelles hypothèses biologiques fonctionnelles. L'utilité de la plateforme

est démontrée grâce à une analyse phosphoprotéomique chez le rat. L'ensemble des sites phosphorylés identifiés est étudié sous divers aspects comme les interactions modulées par la phosphorylation, l'environnement structurel, leur présence dans les domaines des protéines, la compétition avec la glycosylation et la conservation au cours de l'évolution.

Le **CHAPITRE 3: Phosphoproteome dynamics reveal novel Erk1/2 MAP kinase substrates in epithelial cells** présente l'étude phosphoprotéomique effectuée pour identifier de nouveaux substrats des kinases Erk1/2. Une approche bio-informatique, pharmacologique et phosphoprotéomique a été élaborée pour les découvrir. D'abord, la cinétique de phosphorylation de 7936 sites de phosphorylation a été mesurée à quatre temps (0, 5, 15, 60 minutes) dans les cellules épithéliales de rat suite à la stimulation au sérum et à l'inhibition pharmacologique de la voie Erk1/2 avec l'inhibiteur des kinases Mek1/2 (PD184352). L'abondance de la phosphorylation a été mesurée avec une procédure phosphoprotéomique quantitative combinant l'enrichissement des phosphopeptides avec le dioxyde de titane, une séparation chromatographique bidimensionnelle en ligne avec un échangeur de cations suivi d'une phase inverse et une quantification sans marquage obtenue avec le spectromètre de masse LTQ-Orbitrap XL. À partir des données obtenues, 157 substrats potentiels des kinases Erk1/2 ont été trouvés avec un nouvel algorithme exploitant le motif de phosphorylation reconnu par ces kinases et les profils cinétiques. Les candidats trouvés ont des fonctions similaires aux substrats déjà connus à l'exception d'un groupe de protéines impliqué dans le métabolisme des acides nucléiques et l'épissage alternatif, suggérant ainsi une nouvelle implication biologique des kinases Erk1/2. Six substrats (Ddx47, Hmg20a, Junb, Map2k2, Numa1, Rras2) ont été validés par essai kinase *in vitro* avec Erk1 et par abolition de la phosphorylation avec la mutagénèse dirigée. Selon les résultats de nos expériences d'immunofluorescence, la localisation nucléocytoplasmique de Hmg20a est modulée par la phosphorylation de la sérine 105 par la kinase Erk1/2.

Le **CHAPITRE 4: Algorithms to detect phosphopeptide positional isomers** relate nos travaux pour détecter la présence d'isomères positionnels des phosphopeptides, autrement dit, des peptides qui partagent la même séquence d'acides aminés mais qui sont

phosphorylés sur différents résidus. La fréquence d'occurrence de ces isomères dans un échantillon enrichi en phosphopeptides n'a jamais été rapportée et la littérature scientifique n'indique pas clairement si la séparation de ces isomères est problématique avec les méthodes de séparations utilisées en protéomique. Deux algorithmes ont donc été développés dans ce projet pour mettre en évidence la présence de ces isomères positionnels dans un extrait enrichi en phosphopeptides. Le premier algorithme cherche les isomères séparés dans les profils d'élution LC-MS. Cet algorithme produit une liste d'inclusion pouvant être utilisée pour une analyse ciblée LC-MS/MS. Le second algorithme détecte les isomères coéluants par une recherche de fragments caractéristiques dans les spectres MS/MS qui indiquent la présence de plus d'un phosphopeptide. Les résultats obtenus avec ces algorithmes indiquent que la fréquence de ces phosphopeptides isomériques positionnels dans un échantillon enrichi en phosphopeptides est d'environ un pourcent.

CHAPITRE 2: ProteoConnections: a bioinformatics platform to facilitate proteome and phosphoproteome analyses

Mathieu Courcelles[1,2], Sébastien Lemieux[1,4], Laure Voisin[1], Sylvain Meloche[1,5], Pierre Thibault[1,2,3]

Article publié dans le journal ***Proteomics,*** 11, 2654-2671 [117]

Inclut avec la permission de John Wiley and Sons

[1]IRIC, Institut de recherche en immunologie et cancérologie, [2]Département de Biochimie, [3]Département de Chimie, [4]Département d'informatique et de recherche opérationnelle, [5]Département de Pharmacologie, Université de Montréal, Montréal, Canada

2.1 Contribution des auteurs

Mathieu Courcelles et Pierre Thibault ont écrit le manuscrit. Sébastien Lemieux, Laure Voisin et Sylvain Meloche ont révisé et commenté le manuscrit. **Mathieu Courcelles** a conçu et implémenté la plateforme d'analyse ProteoConnections. Laure Voisin a effectué la culture cellulaire. **Mathieu Courcelles** a procédé à la préparation des échantillons et l'analyse de spectrométrie de masse. **Mathieu Courcelles** a fait l'analyse bio-informatique des données du phosphoprotéome du rat. Pierre Thibault et Sébastien Lemieux ont supervisé cette analyse.

2.2 Abstract

Novel and improved computational tools are required to transform large-scale proteomics data into valuable information of biological relevance. To this end, we developed ProteoConnections, a bioinformatics platform tailored to address the pressing needs of proteomics analyses. The primary focus of this platform is to organize peptide and protein identifications, evaluate the quality of the acquired dataset, profile abundance changes and accelerate data interpretation. Peptide and protein identifications are stored into a relational database to facilitate data mining and to evaluate the quality of datasets using graphical reports. We integrated databases of known post-translational modifications and other bioinformatics tools to facilitate the analysis of phosphoproteomics datasets and to provide insights for subsequent biological validation experiments. Phosphorylation sites are also annotated according to kinase consensus motifs, contextual environment, protein domains, binding motifs and evolutionary conservation across different species. The practical application of ProteoConnections is further demonstrated for the analysis of the phosphoproteomics datasets from rat intestinal IEC-6 cells where we identified 9615 phosphorylation sites on 2108 phosphoproteins. Combined proteomics and bioinformatics analyses revealed valuable biological insights on the regulation of phosphoprotein functions via the introduction of new binding sites on scaffold proteins or the modulation of protein–protein, protein-DNA or protein-RNA interactions. Quantitative proteomics data can be integrated into ProteoConnections to determine changes in protein phosphorylation under different cell stimulation conditions or kinase inhibitors, as demonstrated here for the Mek1/2 inhibitor PD184352.

ProteoConnections is available at http://www.thibault.iric.ca/proteoconnections.

2.3 Introduction

The availability of sensitive mass spectrometers with high duty cycle has paved the way to high throughput proteomics experiments where several hundred proteins and modifications thereof can be identified at an unprecedented speed in a single experiment. The relatively large datasets obtained from these experiments have prompted the development of several computational tools to facilitate data interpretation. More specifically, different proteomics groups have developed customized analysis pipeline including Trans-Proteomic Pipeline [118], TOPP [80], VEMS [119], Prequips [120], HTAPP [121]. PEDRO [122], PRISM [123], myProMS [124], MASPECTRAS [125], PrestOMIC [126], PeptideDepot [127], 2DDB [128], Qupe [129] to process data according to their respective experimental setups. Most of these pipelines use relational database management systems (DBMS) to consolidate and archive raw mass spectrometry (MS) spectra and the corresponding peptide and protein identifications. However, they differ on the basis of data analysis tools available for MS/MS data preprocessing (peak picking, MS/MS quality and file format conversion), data exchange via open extensible markup language (XML) file formats, support of multiple search engines, evaluation of peptide assignments and quantitative analysis. These data analysis platforms have been supported by large data repositories such as OPD [130], PRIDE [131], GPM [132], PeptideAtlas [133], Human Proteinpedia [134], NCBI Peptidome [135], TRANCHE [136] to favour data exchange (raw data and peptide/protein identifications) or to provide comprehensive repertoire of organism-specific proteomes. These repositories are not only useful to record specific experimental details but also to compare datasets with different methods and algorithms.

More specialized database resources, including PhosphositePlus [85], Phospho.ELM [86], PHOSIDA [87], PhosphoPep [88] and LymPHOS [137] have emerged recently in support to the large amount of phosphorylation data available through large-scale phosphoproteomics studies. Furthermore, some resources such as Phospho.ELM [86] PHOSIDA [87], Scansite [90], KinasePhos [138], and PPSP [94] also provide prediction

and/or annotation tools to identify potential kinases associated to the phosphorylation of specific sites. Recently, two database resources, SysPTM [105] and PTMScout [104], have been developed to study post-translational modifications (PTMs) that regulate cell signaling processes. Both of them contain PTM datasets from public databases and peer-reviewed MS papers to allow the reporting of conditions under which a PTM was observed. Other tools such as protein domains (Pfam [139]), motif searches (Scansite [90]), Gene Ontology, pathways (KEGG [140]), predictors, evolutive conservation and search for multiple PTM clustered in a protein region have been included to gain biological insight from each dataset. Comparison of published datasets can be performed with simple tools although comparison is limited to the filtering parameters used by the original author. Dataset management and filtering of raw peptide and protein identifications are not available from those tools.

In this context, we have developed ProteoConnections, a bioinformatics analysis platform that facilitates the exploration of proteome and phosphoproteome datasets. ProteoConnections distinguishes itself from other proteomics data processing pipelines by its strong focus on phosphoproteomics analyses to facilitate the interpretation of the data and derive insightful biological information. We integrated different bioinformatics tools to annotate phosphoproteomics data in a global or a targeted fashion (consensus motifs, contextual environment, protein domains, binding motifs and evolutionary conservation across different species). New analysis tools were developed to search putative protein-protein interactions mediated by phosphorylation, enrichment of phosphorylation in protein domains and conservation of phosphorylation sites across species for specific molecular function. Furthermore, the rapid pace of changes in databases requires frequent updates of annotated datasets and maintenance of third party information. To this end, external database resources used by ProteoConnections are updated by web services and automated scripts. Similarly to existing pipelines, ProteoConnections also manages dataset of protein identification from MS/MS experiments by integrating tools for data filtering, quality control of identification (graphical displays describing datasets at

peptide/protein/modification levels) and quantitative analyses. The versatility of this platform makes it possible to share proteomics information and biological inference amongst groups of users, an advantage that can be of practical utility to a large number of proteomics core facilities.

We demonstrate the application of ProteoConnections for the analysis of the phosphoproteomics datasets from IEC-6 rat intestinal epithelial cells. These analyses provided an unprecedented number of identifications with 9615 phosphorylation sites (15 700 phosphopeptides) on 2108 phosphoproteins, of which 80% of the sites were not previously reported in the rat and 43% represent novel identification yet unreported in orthologs from other species. We explored this relatively large dataset with different bioinformatics tools to profile the distribution of phosphorylation sites across different protein domains (e.g. protein binding, nucleic acid binding, nuclear localization and export sequences, etc.) and identified those where phosphorylation represented a statistically significant enrichment. More than 50% of phosphorylation sites identified in this study are recognized by specialized phospho-binding domains, and construction of protein-protein interaction network in combination with consensus motifs enabled the identification of known and potential interactors such as those represented by classical and non-classical 14-3-3 binding motifs. Furthermore, detailed structural analysis of 292 sites found in proteins where three dimensional structures were available revealed potentially important protein-protein and protein-RNA interactions mediated by protein phosphorylation. For example, phosphorylated residues of proteins such as Histone H4 (Ser48) or 60S ribosomal protein L15 (Ser97) can interact with neighboring basic amino acids to prevent efficient binding to DNA or RNA. The availability of convenient bioinformatics tools is explored in the larger context of phosphoproteomics studies to provide multiple biological hypotheses for subsequent validation.

2.4 Materials and methods

2.4.1 Cell cultures, protein extraction and sample preparation

Rat intestinal epithelial (IEC-6) cells were made quiescent by serum starvation for 24 h and then stimulated with 10% fetal bovine serum for 0, 5, 15 and 60 minutes in presence or absence of 2 μM PD184352 (selective Mek1/2 inhibitor). Three biological replicates were prepared for each condition. Cells were collected by scraping, washed twice with ice cold PBS (HyClone), lysed with buffer B (10 mM Tris pH 8.4 (VWR), 140 mM NaCl, 1.5 mM $MgCl_2$, 0.5% NP-40 (Calbiochem), 1 mM dithiothreitol (DTT), protease and phosphatase inhibitor (1:100 added freshly), and spun at 1,000xg for 3 min at 4°C. All chemicals were purchased from Sigma-Aldrich unless otherwise indicated. Supernatant was transferred to another tube (cytoplasmic fraction). Pellet was resuspend in lysis buffer B plus 1/10 of detergent stock (3.3% w:v sodium deoxycholate, 6.6% Tween 40), vortexed at slow speed, incubated on ice for 5 min, spun at 1,000xg for 3 min at 4°C and the supernatant was discarded. The pellet was rinsed with lysis buffer B, and the nuclei extract was lysed with extraction buffer B (20 mM HEPES pH 7.9, 1.5 mM MgCl2, 0.42 M NaCl, 0.2 mM EDTA, 25% glycerol) and sonicated. Benzonase nuclease HC (Novagen) was added to digest nucleic acids of the nuclear fraction. Proteins were precipitated overnight with cold acetone (-20°C) (EMD Chemicals Inc., Gibbstown) and resuspended the next day with a solution of 1% SDS and 50 mM ammonium bicarbonate.

Cytosolic and nuclear proteins extracts were reduced for 20 min at 37°C with 0.5 mM tris(2-carboxyethyl)phosphine (Pierce) and then alkylated with 50 mM iodoacetamide for 20 min at 37°C. The excess of iodoacetamide was neutralized by adding 50 mM DTT. Proteins were quantified using microBCA (Pierce), diluted 10 times with 50 mM ammonium bicarbonate, digested with sequencing grade trypsin (1:100) (Promega)

overnight at 37°C with high agitation, acidified with trifluoro acetic acid (TFA) and dried in a SpeedVac (Thermo Fisher Scientific, San Jose, CA).

Phosphopeptides, 1 mg/replicate, were enriched as previously described [141] on home-made TiO_2 affinity columns (1.25 mg Titansphere, 5 µm, GL Sciences), using 250mM lactic acid (Fluka) and eluted with 30 µL of 1% ammonium hydroxide. Samples were acidified with 1 µL of TFA, desalted using 30 mg HLB cartridge (Waters Corporation, Milford, MA), dried and resuspended in 2% acetonitrile (ACN) (Thermo Fisher Scientific), 0.2% formic acid (FA)(EMD Chemicals Inc.) prior to analysis.

2.4.2 Mass spectrometry analyses

Enriched phosphopeptides extracts from biological triplicates of cytosol and nucleus were analyzed separately using online 2D-nanoLC-MS/MS on an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific) coupled to an Eksigent 2D nanoLC system. SCX separation was obtained using an Opti-Guard 1 mm cation column (Optimize Technologies) and eluted with five different ammonium acetate salt fractions (0, 0.25, 0.5, 1.0 and 2.0 M), pH 3.0 in 2% ACN (0.2% FA). Eluted salt fractions were loaded on a self-packed reverse phase trap column (4 mm length, 360 µm i.d.) and injected on a reverse phase analytical column (10 cm length, 150 µm i.d.) (Jupiter C_{18}, 3 µm, 300 Å, Phenomenex). Peptides were eluted using a gradient from 2 to 33% ACN over 53 min followed by a gradient from 33 to 60% ACN for the next 10 min using a flow rate of 600 nL/min. Detailed procedure is described in [142].

2.4.3 Protein identification and bioinformatics analyses

MS/MS spectra were preprocessed using Mascot Distiller v2.1.1 (Matrix Science). The centroided MS/MS data were merged into single peak-list file and searched with the Mascot search engine v2.2 (Matrix Science) against the combined forward and reversed IPI rat database v3.54 containing 39,928 forward protein sequences. The following parameters were used: parent and fragment mass tolerance of 0.02 Da and 0.5 Da respectively, trypsin with 2 missed cleavages and the following modifications: carbamidomethyl (C), deamidation (NQ), oxidation (M), phosphorylation (STY). False discovery rate (FDR) was obtained using a decoy database and applying the following filtering criteria: ± 10 ppm peptide mass precursor, peptides assigned to proteins with a $p < 0.05$ significance threshold and provided a FDR below 1%. To remove the redundancy in the peptides list, only the identification with the highest score was retained. Dataset is available online via ProteoConnections at http://www.thibault.iric.ca/proteoconnections and in Supplemental Table A2.II.

2.4.4 ProteoConnections architecture

ProteoConnections is constructed with a three-tier architecture that comprises the data, the logical tools and the presentation browser (Figure 2.1). The data tier is based on a data warehouse implementation model that uses MySQL as a database management system (DBMS) to consolidate data in a single location, provide fast and reliable access for data mining, and integrate analysis tools. The centralized database system facilitates the comparison of experiments, the re-analysis of previously acquired data and combination of data from species-specific proteomes. The current database model consists of 37 tables (Figure A2.1). The logic tier comprises different PHP and Perl scripts that are used to extract and format the requested information from the database. The presentation tier uses

an Apache web server to facilitate data access via a web browser. A series of views (not SQL) are available to consult the datasets in HTML format reports.

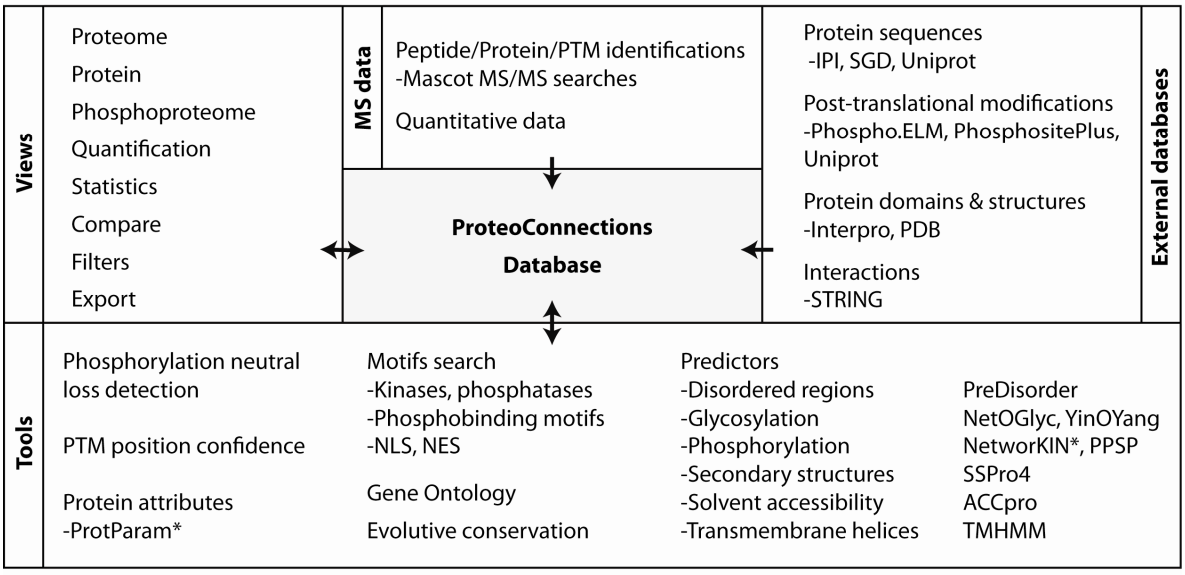


Figure 2.1: Overview of available features in ProteoConnections.

Identifications of peptides/proteins from mass spectrometry are stored into MySQL relational database management system. External databases are included to annotate identified proteins. ProteoConnections offers different views to consult inserted data and facilitate their analysis using direct access to bioinformatics tools installed locally or on web services (marked with *).

Programming interfaces to bioinformatics tools and databases were written for web services and stand alone applications installed locally (Table A2.I). ProteoConnections must be installed under the Linux operating system in order to use all bioinformatics tools. ProteoConnections was tested on a desktop computer (Dual Intel Xeon 3GHz cpu, 2 Gb of RAM) but is used routinely on a server (8 Intel Xeon 3GHz cpu, 32 Gb of RAM) by a group of 15 users. To control access and facilitate data sharing, ProteoConnections has been designed with user/group permissions according to three user types: collaborators with read-only access to data, regular users that can enter and review data into the database and admin users that have access to automated maintenance tools to update the external

resources. The multi-user environment of ProteoConnections makes it an inexpensive solution for core facilities.

2.4.5 Data organization, filters and searches

At present, ProteoConnections support the entry of database searches from Mascot. In order to benefit from the tools integrated in ProteoConnections, data obtained from other search engines must be imported into a specific CSV format recognized by ProteoConnections. We planed to add support for mzIdentML (exchange format for peptides and proteins identified from mass spectra), a format now supported by major search engines (Mascot, Sequest, X!Tandem, Omssa, Scaffold). Search results can also be entered automatically into ProteoConnections using Mascot Daemon or by manually selecting a specific job number. Data is organized in a hierarchical fashion with projects (1st level), experiments (2nd level) and Mascot searches from each sample (3rd level). This structure is useful to create different flexible views for subsequent comparison. The explorer mode provides a simple panel where projects, experiments and searches from specific users are displayed. Entries can be filtered by peptide score and modifications to generate selected list of peptides. Additional filters such as peptide min/max length, mass tolerance in ppm or dalton, peptides with quantitation data and assignments with highest score can be used to remove redundant identifications. Admin users can also create filtered views to consolidate data from specie-specific proteome that have been collected over time in various projects. The combination of all datasets can be valuable for organism proteogenomics annotation projects. This feature is also useful to re-analyze data from previous experiments and investigate other hypotheses.

Filtered views can be visualized by a detailed graphical overview for the distribution of unique peptides/proteins, false discovery rate, charge, mass, length, retention time, mass deviation and score (Figure A2.2 & Figure A2.3). These graphical displays can

provide a practical visual cue to identify instrumental biases and to improve analytical methods. The false discovery rate enables the user to select an appropriate peptide cut-off score based on the distribution of decoy identification. The distribution of mass deviation can reveal potential mass calibration shift during the analysis, a bias that can be corrected by post-acquisition recalibration. In addition to filtering steps, a search page can be used to find specific protein by accession identifier, protein name, gene name, by peptide sequence or subsequence (Figure A2.4). The user can define a short list of protein identifiers to limit the result output, and reduce the manual sorting of protein candidates of interest.

ProteoConnections also comprise a module to compare identification from different datasets in order to determine common/distinct protein and peptides. This comparison can be performed for multiple datasets, including data representing different projects experiments or sample searches. Identification can be sorted according to unique peptide sequences (…ANS…), a sequence with chemical modification (…AN(deamidate)S…), a sequence with PTM (…ANS(phospho)…), or a sequence with a specific charge state (…ANS…, 2+). The selection of appropriate criteria avoids confusion in the interpretation and accounting of relevant unique identification. ProteoConnections allows all of the previous selection to be reported independently.

2.4.6 Proteome view

The proteome view reports a list of identified proteins (Figure A2.5) according to three possible selection criteria: Mascot, all and unique. By selecting Mascot, ProteoConnections will only report protein assignments as obtained from the Mascot search engine to facilitate comparison between Mascot reports. A disadvantage of this report feature is that peptides can be assigned to different protein entries depending on the set of identified peptides. This results in an over estimation of protein identifications and ambiguities when comparing different lists of proteins to determine common assignments. Furthermore, protein

descriptors selected by the search engine may not be the most relevant or easily recognized by users. To alleviate this difficulty, the all mode correlates all identified peptides to all entries in the protein sequences database to determine whether or not a peptide is shared by multiple proteins. This allows for comparing changes in protein abundances when peptides are shared with multiple proteins. This peptide-protein mapping method can also be used to facilitate the correlation of identification between different versions of sequence databases, especially when identifiers and peptide start-stop positions change between releases.

In the unique mode, protein assignments are performed by combining Mascot search results using the following greedy algorithm: i) count the number of distinct peptides matching to each protein, ii) select the protein with the highest number of peptides and the highest priority (see below), iii) add the top ranking protein candidate to the list and remove the corresponding assigned peptides from the peptide list, iv) repeat from step i until the peptide list is empty. In situations where proteins have the same number of peptides, a priority level is assigned to yield the most meaningful protein identification. The algorithm prioritizes the selection of protein names with specifics keywords (Isoform > Fragment > similar > Uncharacterized protein = hypothetical = kDa protein). For datasets searched using International Protein Index (IPI) database [143], a routine is used to select the best annotated protein based on the source database. IPI is assembled from several sources using a hierarchical order of decreasing level of high quality annotations: Swiss-Prot [84] > RefSeq [144] > TrEMBL > Ensembl. Since IPI is now deprecated, this code will benefit to people that are still using it while it is still maintained or for past datasets.

The list of proteins can be filtered according to a minimum number of identified peptide and protein score cut-off value. The report output also displays the distribution of protein molecular weight, protein score, peptide count and sequence coverage (Figure A2.6). Additional information such as the protein sequence, physicochemical properties with ProtParam [145], signal peptide with SignalP [146], transmembrane helices prediction with TMHMM [147], Interpro domains [102] and PDB protein structures identifiers can also be included to the report.

Detailed information on identified peptides and modifications of thereof, including those reported from the literature (e.g. acetylation, glycosylation, methylation, phosphorylation, sumoylation and ubiquitination as reported in Phospho.ELM [86], PhosphositePlus [85] and Uniprot [84]), are available for each protein in the protein and peptide evidence view (Figure A2.7 & Figure A2.8). User can view the MS/MS spectrum and peptide fragment assignment made by Mascot using the peptide evidence view (Figure A2.8), provided that appropriate privilege was granted to access the Mascot server or spectral data from open access. When available, changes in protein and peptide abundances from quantitative proteomics experiments are also displayed (see section 2.4.8). Multiple sequences can be aligned using Muscle [148] to determine the level of conservation across species, and the extent to which a modified site is conserved. Alignments can be exported in ClustalW, Fasta, MSF and Phylip formats for further use.

2.4.7 Phosphorylation sites view

ProteoConnections integrates different bioinformatics tools for the analysis of large-scale phosphoproteomics datasets in a single platform (Figure A2.9). The neutral loss of H_3PO_4 (98 Da) characteristic of phosphopeptides is queried in all MS/MS spectra from the mascot generic file (mgf) prior to database search. We also probe for the presence of 80 Da neutral losses in the corresponding MS/MS spectra, the occurrence of which can be indicative of sulfated peptides or artifact arising from silver stained gel [35]. To enhance the confidence in the location of phosphorylation sites, we integrated a script that determines the probability of a specific assignment based on the method by Olsen et al.[19]. This method retrieves first the list of candidate positions of phosphorylated site on the phosphopeptides with their corresponding peptide score. For each candidate, the peptide score is transformed in $1/p$ knowing that peptideScore = $-10 \log_{10}(p)$ and p is the probability of the peptide assignment. Then the sum of $1/p$ for all candidates is done and all the $1/p$ are transformed in proportional probability (P-site) by dividing by the sum. The site localization probability is

obtained for each position by summing the P-site value for each candidate that has this site specific position. Phosphopeptide candidates can be selected based on a minimal probability threshold. Phosphorylation sites are compared against other databases such as Swissprot v15.53 [84], Phospho.ELM v8.2 [86] and PhosphositePlus v2.0 [85] to determine if sites have been previously reported. This correlation is performed by mapping the identified sites on sequences from these databases to avoid ambiguities from entries having insertions/deletions errors, inconsistency in the presence of initiating methionine, or variable accession number between releases.

All known kinase/phosphatase and binding phosphomotifs (specific amino acid sequence at the phosphorylated site, e.g. PXp[ST]P) found in HPRD [116] and PepCyber [149] have been integrated in ProteoConnections, and can be use to scan regular expression and annotate the identified phosphorylation sites. Motif scan can be performed for a single kinase by selecting the motif description for a dropdown menu or for all possible kinases motifs. Similarly, all sites are searched against phosphorylation sites predictors PPSP [94] and NetworKIN [95] to determine potential kinases. The list of phosphorylated peptides can be exported in a compatible format to find over-represented phosphorylation motifs using Motif-X [96] or MoDL [97] (these tools are not included). These tools can be used to determine activation of a kinase in a particular condition or to discover new phosphorylation motifs for uncharacterized kinases. Similarly, MAP kinases docking sequences can be searched to increase the specificity of motif analyses. Additional tools are also available to determine solvent accessibility, secondary structures and disordered regions using ACCpro, SSPro4 [150] and PreDisorder 1.0 [151]. We also integrated Interpro protein signature database [102] to identify potential regulatory sites, and NLSdb [152] and NES consensus motifs [153] to determine if identified phosphorylation sites are proximal to nuclear localization signal (NLS) or nuclear export signal (NES), respectively. To determine the interplay between phosphorylation and O-glycosylation at specific site, we included known glycosylation sites from SwissProt [84] and PhosphositePlus [85], and

integrated NetOGlyc [99], YingOYang [100] and OGlcNAcScan [101] as site prediction tools for O-glycosylation.

Finally, phosphorylation sites can be aligned with 10 different species (*M.musculus, H.sapiens, B.taurus, G.gallus, D.rerio, D.melanogaster, C.elegans, A.thaliana, S.cerevisiae and E.coli*) to determine the extent to which sites are evolutionary conserved to support functional significance [5, 154]. Orthologs are first identified from BLAST [155] searches of UniProt protein sequences corresponding to each species using an e-value threshold of 10^{-10}. A Smith-Waterman alignment [156] is then performed using the SSearch tool which is included in the Fasta package [157], and the conservation of phosphorylation sites is evaluated in a pairwise fashion between the query sequence and those from other species. We also evaluated sequence alignment from InParanoid7 [158] to reduce the number of BLAST candidates that are not true orthologs and could lead to false positive identification of conserved sites. However, close inspection of the data revealed that isoforms are missing from the database and correlation between identifiers is problematic, leading to an underestimation of the number of conserved sites. Accordingly, the BLAST alignment strategy was thus selected in the present platform. For each ortholog, ProteoConnections reports protein identifier, gene name, site position and whether or not the site has been identified before. For non-conserved sites, the amino acid substitution and the sequence alignment are shown. This module also report statistics on the level of conservation for each species including the number of conserved sites, the frequency of specific mutations, and the proportion of sites conserved across species. Different graphical views display the distribution of S/T/Y sites, the number of phosphorylated residues per peptide and proteins, the proportion of previously reported sites, and the distribution of sites assigned with different confidence levels.

2.4.8 Quantification view

Quantitative analysis from label free, SILAC, and iTRAQ experiments can be imported into ProteoConnections once data are properly formatted in comma-separated value (CSV) file. Normalization of peptide intensities and validation of quantification values should be carried prior insertion. Peptide abundances are displayed in a scatter plot to determine the reproducibility and variability from experimental replicates. In proteome view, changes in proteins abundances across conditions can be filtered by fold-change at the peptide level. Identified peptides are displayed with their fold change, p-value from the two-tailed Student t-test and manual validation status. A volcano plot of $-\log_{10}$(p-value) versus $\log_2$ (fold change) of all identified peptides is also displayed to facilitate the identification of peptides showing statistically significant changes in abundance. A plot of peptide abundance versus time is also generated for quantitative proteomics experiments involving different replicates and time points.

2.4.9 Network, domain and motif analyses of a rat phosphoproteomics dataset

The analysis of rat phosphoproteomics datasets was performed using functionalities available in ProteoConnections. For these analyses, only phosphorylation sites with high localization confidence were used (greater than 75%).

To evaluate the enrichment or depletion of phosphorylation in protein domains, the $\log_2$-odds ratio was used as a measure of effect size (i.e. the strength of the relationship between two variables in a sample). The odds-ratio is determined from the following ratio: (phosphorylated sites in domain/ phosphorylated sites not in domain)/ (non-phosphorylated sites in domain/non-phosphorylated sites not in domain). Non-phosphorylated sites are serine, threonine and tyrosine residues not identified as phosphorylated in our study. A

Fisher`s exact test was used to test for a dependence between the phosphorylation state and the category of interest, since it is applicable to samples of small and large sizes. $Log_2$-odds ratio and the Fisher`s exact test were calculated using R (function: fisher.test()), a statistical computing software. Additional details are provided in supplementary material. A Perl script is available in ProteoConnections under the “tools” section to run this analysis.

Putative phospho-interactions and kinase/phosphatase-substrate interactions were searched using a Perl script integrated into ProteoConnections. Using the STRING interaction database, high confidence interactions (STRING score > 0.9) from databases and experiments were extracted. The depth of the network was set to 1. Biomart is queried to retrieve the list of proteins with phospho-binding domains (14-3-3, BRCT, C2, FHA, MH2, PBD, PTB, SH2, WD40 and WW) using Interpro annotations. The nodes of the network that comprised non-phosphoproteins or proteins without a phospho-binding domain were removed. To check if the interactions are modulated by phosphorylation, protein interactors with phospho-binding domains were searched for phosphorylation sites with the motif recognized by the corresponding phospho-binding domains. ProteoConnections filters and annotates the phosphorylated sites recognized by phospho-binding domains using the motif description retrieved from HPRD [116] and PepCyber [149]. Finally, interactions with match between phospho-binding domains and sites with the corresponding motif are retained. A similar strategy is done for kinase/phosphatase using GO annotations and phosphorylation motifs of kinases.

ProteoConnections was used to determine the number of conserved phosphorylation sites in 10 other species (*M.musculus, H.sapiens, B.taurus, G.gallus, D.rerio, D.melanogaster, C.elegans, A.thaliana, S.cerevisiae and E.coli*). A Perl script calculated the number of conserved non-phosphorylated sites for phosphoproteins with an ortholog in the compared species. A Fisher`s exact test was used to determine the significance of the proportion of (phosphorylated sites conserved/ phosphorylated sites not conserved)/ (non-phosphorylated sites conserved/non-phosphorylated sites not conserved). A Perl script is available in ProteoConnections under the “tools” section to run this analysis.

2.4.10 Miscellaneous features

Other specific features have been integrated into ProteoConnections in support to in-house projects. These include the creation of smaller sequence databases that can be customized to add or remove protein sequences without editing Mascot database configuration in order to facilitate rapid and targeted protein identification. Other features include the development of specific filters for the selection of peptides binding to the major histocompatibility complex I (MHC) [110]. The algorithm is able to recognize the specific signature of MHC I peptides for mouse and human. To obtain a global system overview of the identified proteins, Gene Ontology (GO) analysis [159] and mapping to protein-protein interaction networks from STRING [160] can be performed. These two analyses can be performed on identified proteins reported in proteome and phosphosites views. The GO terms enrichment/depletion analysis (for biological process, molecular function or cellular compartment) is determined using our in house tool. Odds ratios for all terms (no slim available) are calculated by comparing against the whole proteome (number of proteins with the GO term in the dataset / number of proteins in the dataset) / (number of proteins with the GO term in the proteome / number of proteins in the proteome) and p-values, calculated with Fisher's exact test, report their significance. In the report, the GO terms with related functions based on the GO hierarchy are clustered together with a list of identified proteins for each term.

All views described above can be exported into a CSV file format for subsequent processing with dedicated applications. Provision are presently made to ProteoConnections to import and export data into a mzIdentML file format to facilitate data sharing and use with other softwares developed by the proteomics community.

2.5 Results and discussion

2.5.1 Analysis of a rat phosphoproteomics dataset

ProteoConnections was used to analyze a rat phosphoproteomics dataset from IEC-6 intestinal epithelial cells stimulated with FBS. We identified a total of 9615 distinct phosphorylation sites on 15 700 phosphopeptides assigned to 2108 unique proteins (Figure A2.2 for more details on the identified peptide population). Twenty percent of the identified sites were also identified in previous studies and could be correlated with available phosphorylation databases, while 43 % of all identified sites represent novel identification unreported in orthologs from other species. Overall, 6419 (66 %) sites were identified with a high degree of confidence (>75%, Figure 2.2A) and only those were retained for subsequent analyses. The proportion of S/T/Y residues was 80:18:2 and 81% of peptides were identified as singly phosphorylated, 17% were doubly phosphorylated and 2% showed evidences for more than 2 phosphorylation sites (Figure 2.2B). The extent of protein phosphorylation ranged from one to 169 sites (Figure 2.2C), whereas the molecular weight of the corresponding protein substrates displayed a normal distribution (Figure 2.2D).

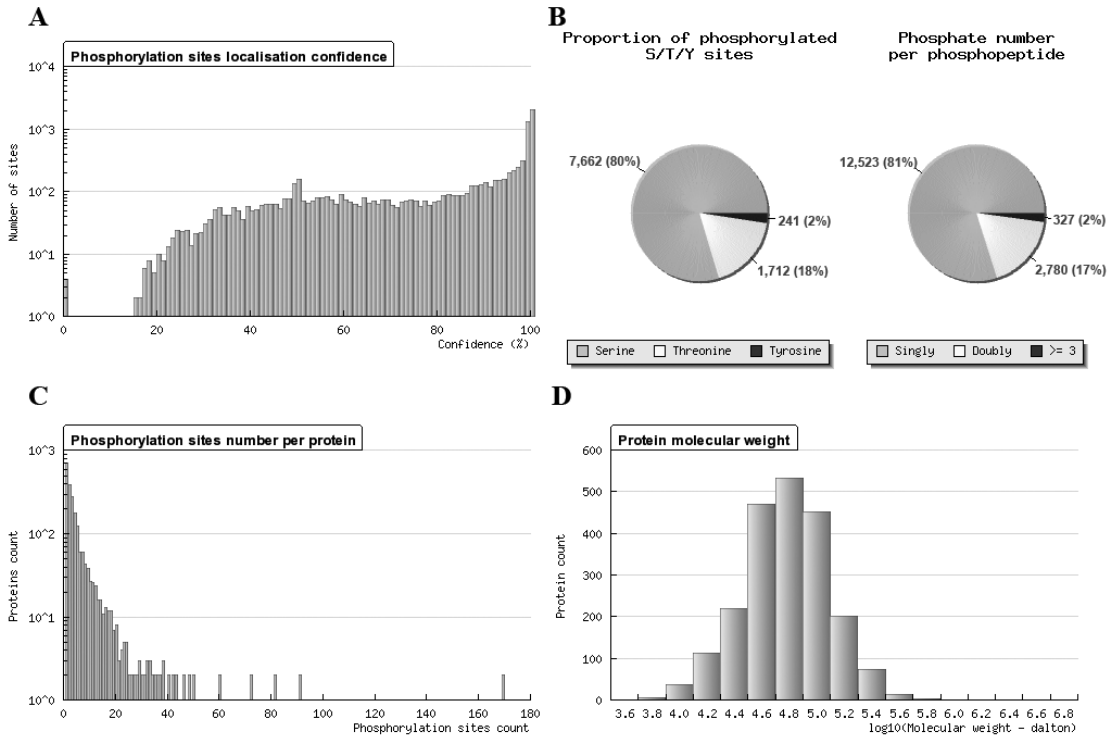


Figure 2.2: Statistical overview of rat phosphoproteome dataset.

In the phosphosite view, ProteoConnections provides detailed graphical display to represent phosphoproteome analyses based on the confidence in the location of the phosphorylation site (A), the proportion of phosphorylated residues (B), the number of phosphorylated sites per protein (C), and the distribution of phosphoproteins molecular weight (D). Additional statistical features are shown in Figure A2.2.

We used PreDisorder 1.0 [151] to determine the disordered regions for each protein and SSpro4 and ACCpro [150] to predict secondary structures and residue accessibility. Our analyses revealed that the fraction of phosphorylation sites observed in disordered regions (70%) is slightly lower than that reported previously for human (84%) and yeast (79%) [5]. However, these differences could be partly attributed to different prediction

algorithms used to identify disordered regions. Consistent with previous studies, identified sites are mostly located in loops (92%) and are also accessible (77 %) [5].

Kinases and phosphatases involved in phosphorylation are particularly important to understand the regulation of cellular signaling. A kinase motifs search found matches for 5875 (92%) of the sites. The phosphorylation prediction algorithms NetworKIN and PPSP (medium stringency) suggested putative kinases for 3595 (56%) and 6419 (100%) sites, respectively. These three methods report multiple kinase suggestions for individual sites, and differences noted in the number of reported sites reflect the variation in sensitivity and specificity of the different methods. Motif descriptors can be either strict or loose which result in missed prediction (poor sensitivity) or wrong/uncertain kinase prediction (poor specificity). Loose descriptors can give rise to overlapping prediction for different kinase families. Prediction tools have similar caveats but accuracy can vary between kinase families depending on the training sets used. NetworKIN predictions are more conservative than those of PPSP since it uses additional data from protein interactions network. Motifs search and predictions are useful for selecting a subset of phosphorylated sites that can regulate a signaling pathway of interest.

Several tools have been integrated in ProteoConnections to facilitate the assignment of potential regulatory roles that can be inferred from identified phosphorylation sites. First, we determine the location of modification sites within different protein domains. Most domains have a defined function (e.g. catalytic activity, protein-protein interaction, or protein-nucleic acid interaction), and phosphorylation site located within these domains may possibly modulate protein activity. A sizable proportion of identified sites (865 out of 6419, 15%) are located within a domain annotated by Interpro (Figure 2.3A). Approximately half of these sites are associated with catalytic activities including protein kinases (66 sites, IPR011009 & IPR002290), dehydrogenases (7 sites, IPR001017 & IPR020829), translation initiation factors (6 sites, IPR015760), acyl transferases (4 sites, IPR016035) and phosphodiesterases (2 sites, IPR017946). Other sites are located in domains associated with protein binding (242 sites, 4%), nucleic acid binding (137 sites,

2%) and other domains binding to ATP, GTP, metal ions, steroid, calmodulin, phospholipid, NAD(P) and acyl-CoA (28 sites, 0.4%). To determine if phosphorylation sites were significantly enriched in specific domains, we performed Fisher`s exact tests on log-odds ratio of identified sites versus all non-phosphorylated sites from the associated domains (see Figure 2.3B). These analyses indicated that protein domains are generally depleted in phosphorylation sites (log odds-ratio -1.5, p-value < 5.5 x 10^{-223}). These results suggest that protein domain activity is regulated by a few phosphorylated sites in situ. In view of this result and the low frequency of sites located within domain, it is possible that sites outside of domain, though structurally proximal, may contribute to regulating domain activity. It is however difficult to locate computationally distant regulatory site without protein structure information.

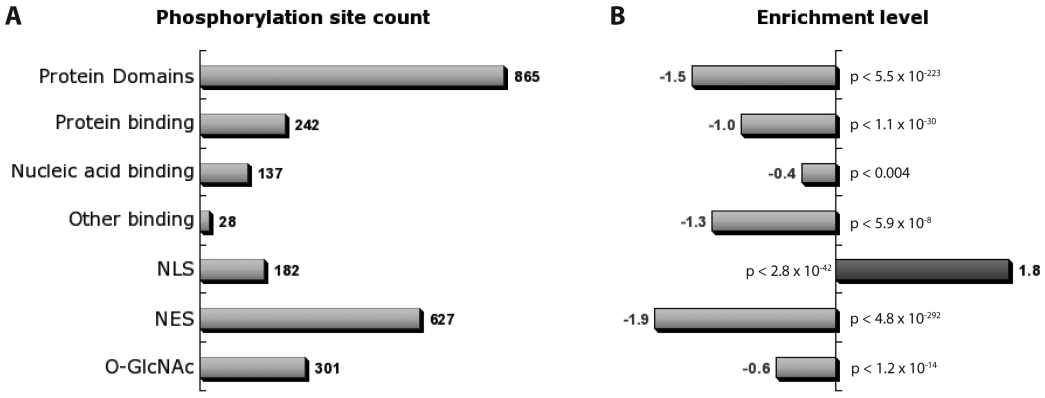


Figure 2.3: Distribution and enrichment of phosphorylation in different protein domains.

A) Approximately 15% of all identified phosphorylation sites (865/6419 sites) are located within an annotated Interpro protein domain. B) Log-odds ratio of phosphorylated sites versus non-phosphorylated sites for different domains based on a Fisher`s exact test. Protein domains that show a higher level of structure are generally depleted in phosphorylation sites. A higher proportion of phosphorylation sites was found in nucleic acid binding domains and sites proximal (10 amino acids) to NLS (based on NLSdb).

Domains implicated in binding of small molecules are depleted by a log-odds ratio of 1.3 (p-value < 5.9×10^{-8}), suggesting that phosphorylation is not highly represented although these domains contain 19 % of phosphorylatable sites from all protein domains. Interestingly, we found that nucleic acid binding domains have a larger representation of phosphorylation sites compare to other protein domains. Among the 137 sites found in nucleic acid binding domains, different proteins showed a higher enrichment level. These included the intermediate filament DNA-binding region (IPR006821) that is enriched by a log-odds ratio of 2.6 (p-value < 1.4×10^{-4}), eukaryotic transcription factor Skn-1-like (IPR008917) by 3.8 (p-value < 6.9×10^{-5}) and splicing factor 3B subunit 1 (IPR015016) by 3.6 (p-value < 1.2×10^{-6}). The high number of phosphorylated residues found in DNA-binding region suggests that protein phosphorylation plays an important regulatory role in mediating protein-nucleic acid interactions as previously reported by Hyland et al. [161]. Similarly, the presence of phosphorylation sites within transcription factor binding regions is known to affect DNA binding by preventing their association with recognized DNA elements [162]. Furthermore, protein phosphorylation plays an essential role in functional spliceosomes although no molecular association with the pre-mRNA was revealed yet [163].

We performed similar analyses to determine the extent to which phosphorylation sites were found in close proximity to sequence motifs such as nuclear localization signal (NLS) and nuclear export signal (NES) that are associated with nucleocytoplasmic protein translocation. NLS consists of one or more short sequences of lysine or arginine residues that are recognized by nuclear transport receptors and target protein to the nucleus. We searched the NLSdb [152] and found 182 occurrences on 89 proteins that showed a phosphorylation site within 10 amino acids of a NLS (Figure 2.3A). Interestingly, putative NLS domains were enriched in phosphorylation sites by a log-odds ratio of 1.8 (p-value < 2.8×10^{-42}). Among proteins having phosphorylation sites proximal to NLS, we identified Mcm3 and RanBP3, two known examples where import is blocked cooperatively by multiple phosphorylations [164, 165]. For nine proteins, we observed that multiple sites

were located within 10 amino acids from the NLS. We also identified 35 phosphorylation sites near NLS with a 14-3-3 consensus motif (Ctr9, Scaf1, Sfrs8, Smarcad1, Sptbn1, Ranbp3) and 76 that are recognized by other phosphorylation binding domains. It is noteworthy that 14-3-3 proteins act as molecular scaffolds and their binding to phosphorylation sites near NLS could interfere with the nuclear import machinery [166]. The family of 14-3-3 proteins also contains a NES domain that could trigger nuclear export of bound phosphorylated proteins. Similar analyses performed on phosphorylation sites near NES identified 627 occurrences on 406 proteins with a log-odds ratio of -1.9 (p-value $< 4.8 \times 10^{-292}$).

We next investigated the extent to which phosphorylation and O-glycosylation sites can compete for the same serine or threonine residues, a reciprocity often referred to as "yin-yang" sites [167]. Glycosylation and phosphorylation can have opposite biological effects. In the case of PEST regions, glycosylation can protect from proteolytic degradation while phosphorylation favors protein degradation [168]. Current data on O-glycosylation is limited to 241 sites for the rat, 318 for the mouse and 607 for human. Unfortunately, none of the phosphorylation sites identified in this study where reported to be O-glycosylated. Consequently, we relied on O-glycosylation predictors to obtain a more comprehensive list of potential modification sites. We predicted 613 (10%) O-GalNac sites with NetOGlyc [99] while 1,259 (20%) and 301 (5%) O-GlcNAc sites were obtained with YinOYang [100] and OGlcNAcScan [101], respectively. The overlap of these two O-GlcNAc prediction algorithms yielded only 114 sites. These largely different results reflect the limitations of current prediction models that are trained with small dataset of identified O-GlcNAc sites. OGlcNAcScan, which is the most recent prediction algorithm, was trained with the largest dataset and provide a more conservative number of glycosylated residues compared to other approaches. Using this narrower list of O-GlcNAc sites we performed a Fisher`s exact test on log-odd ratio and obtained a value of -0.6 (p-value $< 1.2 \times 10^{-14}$) suggesting that O-GlcNAc compete less favorably on sites occupied by phosphorylated residues. Gene Ontology annotations of phosphoproteins having predicted glycosylated sites indicated

diverse biological processes that matched known glycosylated proteins [169] including proteins involved in nucleotide metabolism (21 proteins), transport (14 proteins), transcription (9 proteins), carbohydrate metabolic process (5 proteins), cell proliferation (6 proteins) and amino acid metabolic process (1 proteins). No functional enrichment was observed when we compared phosphoproteins having predicted glycosylated sites with all identified phosphoproteins, suggesting that competition between these two modifications is not specific to a biological function. A more accurate definition of the phosphorylation and O-GlcNAc interplay will be gained with more comprehensive identification of the O-GlcNAcome.

2.5.2 Protein interactions modulated by phosphorylation

Protein phosphorylation can also provide a scaffold for the recruitment of other binding partners and mediate protein-protein interactions. Specialized protein domains such as 14-3-3, BRCT, C2, FHA, MH2, PBD, PTB, SH2, WD40 and WW recognize phosphorylated amino acids within specific sequence motifs. A regular expression search with all the phospho-binding motifs found in HPRD [116] and PepCyber [149] identified 3936 sites (61%) that can be recognized by at least one phospho-binding domain. To determine sites that mediate protein-protein interactions, we combined this list with a protein interactions network of identified phosphoproteins and proteins with phospho-binding domains. We then searched interactions with match between phospho-binding domains and sites of the corresponding motif. This network is shown in Figure 2.4A (and zoomable version is available in Supporting Information Figure A2.12).

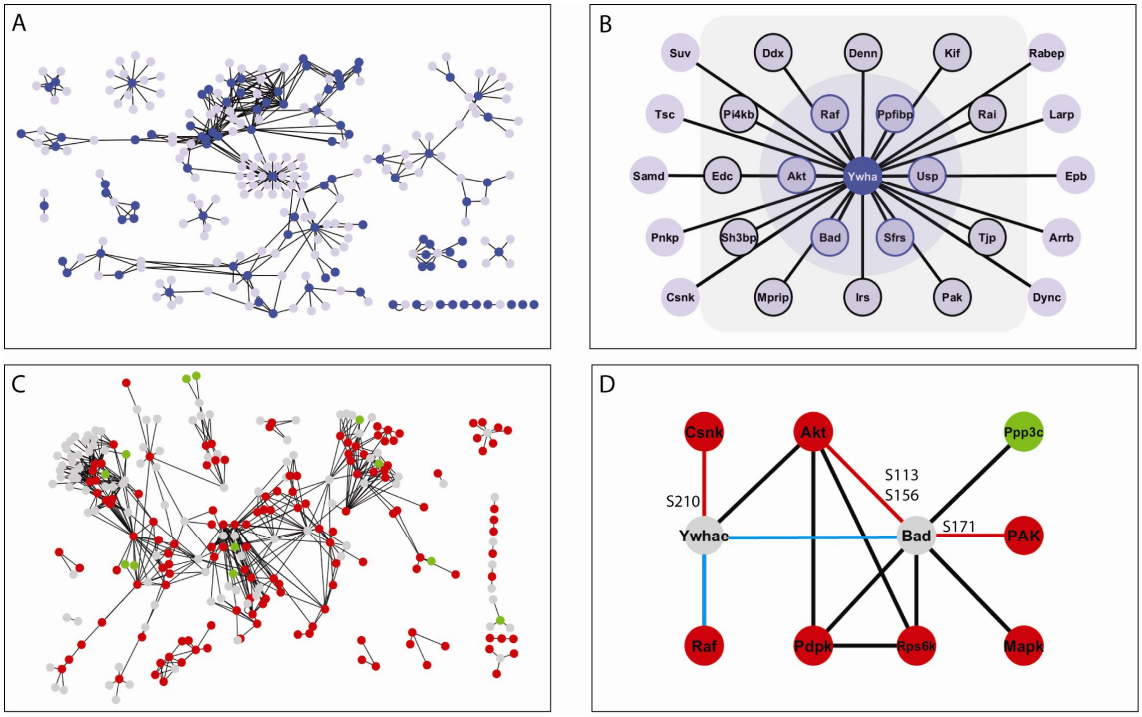


Figure 2.4: Protein interactions mediated by phosphorylation.

A) Protein interactions mediated by phosphorylation in binding domains such as 14-3-3, BRCT, C2, FHA, MH2, PBD, PTB, SH2, WD40 and WW. Dark blue nodes represent protein with a binding domain while the light blue ones are phosphorylated proteins. For an enlarged version, see Supporting Information Figure A2.12. B) Subnetwork showing phosphorylated interactors of Ywha (14-3-3) with a classic motif RSXp[ST]XP (inner circle) and a non classic motif (outer rectangle box) recognized by 14-3-3. C) Kinase and phosphatase substrates interactions network. Red nodes represent kinases, green phosphatases and grey are phosphorylated proteins. For an enlarged version, see Supporting Information file Figure A2.13. D) Subnetwork of kinases (red) and phosphatases (green) interacting with Ywae and Bad substrates. Modified sites from protein substrates are shown. Red edges indicated kinases phosphorylating Ywhae and Bad while blue edges represent interactions mediated by 14-3-3 phospho-binding domains. The networks were generated from ProteoConnections using the STRING database and the figure made with Cytoscape.

The network contains 173 phosphoproteins and 97 proteins with phospho-binding domains, and 79 out of 412 protein interactions correspond to phosphoproteins with sites specifically recognized by these domains. Using this method, we successfully identified known phospho-interactors of Ywha (14-3-3 protein eta) such as Bad, Raf, Irs (Figure

2.4B). Ywha is an adapter protein involved in numerous signaling pathways and binds to its partners via the recognition of specific phosphoserine or phosphothreonine motifs. We also identified 27 binding partners of Ywha, including 6 proteins with a phosphorylation site with the 14-3-3 consensus motif (RSXp[ST]XP), 11 with the non-classical 14-3-3 consensus while the remaining sites have no similarity with known 14-3-3 motifs. Among the 11 proteins, the protein Tsc is known to have a phospho-dependant interaction with 14-3-3, the Arrb protein has no site matching 14-3-3 motif and the remaining 9 have potential 14-3-3 phosphorylation sites not yet identified by MS. The low stoichiometry of phosphorylation and detection limits of MS methods might explain why these sites were not detected in the present study. Protein network provides a convenient tool to identify interacting proteins, including kinases and phosphatases that regulate the phosphorylation of their protein substrates. We used the list of identified phosphoproteins (143 proteins) that interact with a kinase (141 proteins) or a phosphatase (13 proteins) to construct a network of interactors (Figure 2.4C and a zoomable version in Supporting Information Figure A2.13). Only seven percent of the phosphoproteins have known interactions with a kinase or a phosphatase showing the transient nature of the interaction enzyme/substrate. Only 67 of the 543 interactions corresponded to phosphoproteins with phosphorylation sites recognized by a specific kinase. This is further illustrated in Figure 2.4D where we selected two identified phosphorylated proteins, Ywhae (14-3-3 epsilon) and Bad, that are known to interact with 3 and 5 kinases, respectively. Regular expression searches of kinase phosphorylation consensus motifs often yield multiple possibilities for a given phosphorylated site. Using information from protein interaction network and the specificity of the kinase consensus motifs, we could attribute the phosphorylated sites to a specific kinase (Csnk for Ywhae-S10, Akt for Bad-S113,S156 and PAK for Bad-S171). The combination of motif analysis and interaction network provides a valuable approach to define potential kinases for subsequent validation experiments. Obviously, the success of this approach depends on the quality and comprehensiveness of the interacting protein network.

2.5.3 Molecular definition of interacting residues from structural studies

Protein phosphorylation can affect the activity of substrates by conferring a new scaffold platform for protein interactions. However, the changes imparted by this modification and its impact on protein-protein or protein-DNA interactions require a level of structural definition that can only be obtained from X-ray crystallography or NMR studies. Determination of interactions between critical residues of interacting partners can be helpful in highlighting a possible function. Among all phosphorylation sites identified in the present study, we obtained relevant 3D structures for only 59 proteins (2.8% of all sites) from the Protein DataBank (PDB). In spite of the relatively few structures available, we identified several potential interactions involving phosphorylation sites with proximal residues within protein substrates or their binding partners. For example, the phosphorylation of Ser48 from Histone H4 is located in the histone core domain (IPR007125) implicated in nucleic acid binding. Close examination of the protein structure wrapped around the DNA showed that the unmodified serine residue is 3.37 Å from the phosphate group of the DNA backbone (Figure 2.5A, PDB: 1AOI). Phosphorylation of Ser48 presumably interferes with DNA binding due to electrostatic repulsion between the two negatively charged groups. In yeast, this phosphorylated site is associated with a phenotype where chromatin is less accessible to the transcription machinery [110]. Similarly, protein-RNA interactions can also be disrupted by phosphorylation. A case in point is the 60S ribosomal protein L15e (rpL15e) that interacts with the 28S ribosomal RNA (rRNA) (Figure 2.5B, PDB: 2ZKR). Phosphorylation of Ser97 from rpL15e introduces a negative group only 6.35 Å away from the phosphate group of the adenine nucleotide, thus interfering directly with RNA binding. Once this protein is assembled on the ribosome, the phosphorylated site is no longer accessible to kinases. This protein dedicates 58.7% of its surface for contact with all domains of 28S rRNA (except domain IV) and is surrounded to a high degree by rRNA [170]. Among all ribosomal proteins, rpL15e is the second protein that buries the largest surface of rRNA area. It has been proposed that rpL15e is incorporated early in the ribosome assembly, due to the large

protein extensions penetrating into the rRNA core and its interaction with the first transcribed domain. Since this site is not accessible once rpL15e is assembled in the ribosome, we hypothesize that its phosphorylation may interfere with ribosome assembly. The phosphorylated site is located between the globular domain and the extension loop, a region implicated in early interaction between rpL15e and the 28S rRNA in the assembly and folding pathway proposed by Klein et al [170].

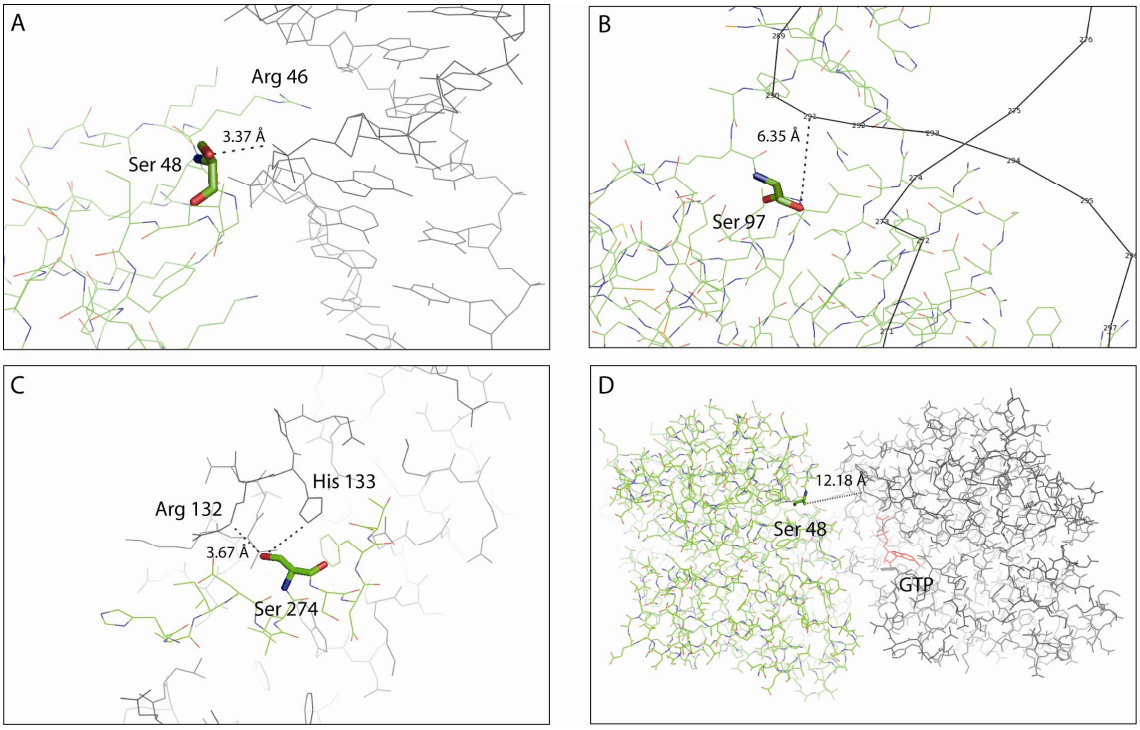


Figure 2.5: Interactions regulated by protein phosphorylation.

A) Protein-DNA interaction: Phosphorylation on Histone H4 on S48 in close proximity to phosphate of DNA (PDB: 1AOI). B) Protein-RNA interaction: 60S ribosomal protein L15e (rpL15e) phosphorylated at S97 is interacting with the 28S ribosomal RNA (PDB: 2ZKR). Only the phosphorus atoms of the rRNA are shown in the PDB file. C) Protein-protein interaction: Phosphorylation of S274 from calpastatin and its interacting residues on Calpain small subunit 1 (Capns1). (PDB: 3DF0). D) Phosphorylation on tubulin alpha-chain of S48 could modulate the interaction with tubulin beta and Stathmin-4 (PDB: 3HKC). All four sites are evolutionary conserved supporting their important regulatory role. Unphosphorylated residues are shown. ProteoConnections was used to retrieve PDB identifiers for phosphorylated proteins and PyMOL to visualize them.

Phosphorylation can also interfere with protein-protein interactions. An example of this was found for calpastatin, a protein involved in the specific inhibition of calcium-dependent cysteine proteases (calpains). The phosphorylation of Ser274 from calpastatin can form ionic interactions with neighboring Arg132 and His133 residues from Calpain small subunit 1, Capns1 (Figure 2.5C, PDB: 3DF0). The interaction between the regulatory subunit of calpain and the region C of the inhibitory domain 1 of calpastatin serves as an anchor point to potentiate the inhibition of the calpain. Previous studies suggested that phosphorylation by Pka controls intracellular localization of calpastatin and that Pkc phosphorylation decreases its inhibitory activity [171]. Ser274 of calpastatin harbors a casein kinase I (CkI), phosphorylation motif and this modified residue might have a different inhibition mechanism to what has been reported previously. The ion-pair interactions between the phosphorylated Ser274 residue of Calpastatin and neighboring Arg132 and His133 from Capns1 would favor an increase in its inhibitory activity.

Another example of protein-protein interaction is the phosphorylation of Ser48 from tubulin alpha, a modification that could influence its interaction with tubulin beta (Figure 2.5D, PDB: 3HKC) and Stathmin-4 (not-shown). In this case, the phosphorylation site is located inside the Tubulin FtsZ GTPase domain (IPR003008) where it likely affects the formation of microtubules. Interestingly, the four sites presented here showed a high degree of conservation across different species including mouse, human, bovine, chicken, zebra fish, fly, worm, plant and yeast. The first one is conserved in all species, the second is absent only in fly, the third is conserved in all vertebrates, and the fourth is missing in plant and yeast. This level of conservation across such a long evolutionary distance likely indicates the significance of these residues in regulating protein activity.

2.5.4 Conservation of phosphorylation sites

The impact of protein kinases and phosphatases on the regulation of protein phosphorylation make them prime candidates for evolutionary studies. Using large-scale phosphoproteomics data, previous reports have estimated the evolutionary rates of change of phosphorylation between different yeast species to determine the functional significance of site conservation [172, 173]. Recently, a comparison of phagosomal phosphoproteomes from evolutionary distant organisms has provided a first detailed bioinformatics analysis on how existing orthologs have been remodeled by the modification of core constituents to transform the phagosome from a phagotrophic compartment to a fully competent organelle for antigen presentation [174]. These analyses have highlighted the varying degree of plasticity of protein phosphorylation through evolution, and the constraints under which certain regions display a high conservation level due to their functional importance [5]. To evaluate the site conservation, we integrated into ProteoConnections a tool that aligns and compares phosphorylated sites across 10 different species (see experimental section for additional details). The relevance of this tool was evaluated in the context of the present phosphoproteomics dataset that comprises more than 6419 phosphorylation sites, where we considered both site identity and non-phosphorylated sites. A pair-wise comparison of site identity and non-phosphorylated residues between rat and all ten other species revealed a linear relationship where conservation increased according to evolution proximity (Figure A2.10). Most phosphorylation sites follow a similar trend whereby closely related species display a high degree of site conservation that progressively decreased with evolutionary distance. The increased evolutionary distance makes it more challenging to find true orthologs and obtain correct alignments. In species closely related to rat, site identity is approximately 4% higher than those from non-phosphorylated residues (Fisher's exact test: mouse p-value $< 4.2\times10^{-22}$, human p-value $< 2.0\times10^{-31}$ and bovine 7.0×10^{-18}, see Supporting Information data). We next compared site conservation on 905 phosphoproteins (43% of all identifications) for which GO terms could be obtained. We determined the difference between the proportion of conserved phosphorylation sites and that from

conserved phosphorylatable sites, and evaluated the significance with the Fisher`s exact test between these two distributions (Figure A2.11). These analyses revealed no systematic or significant site conservation for any of the molecular function categories examined except for phosphosites associated to DNA binding and SH3 domains. For these two categories, we noted that phosphorylation sites were consistently more conserved from human to Xenopus, and that the proportion of phosphorylated vs. phosphorylatable sites was higher by 5.5 % (p-value $< 4\times10^{-4}$) for DNA binding and by 17% for SH3 binding domain (p-value $< 8\times10^{-3}$). The overall lack of phosphosite conservation across species observed here is not entirely unexpected, and is consistent with previous bioinformatics analyses that indicated an overrepresentation of phosphosites in disordered regions in which a higher rate of evolution is typically observed [5]. It should be noted however that phosphosites are not expected to experience stronger evolutionary constraints than phosphorylatable residues in disordered regions. The evolution rate of phosphosites in disordered regions could suggest that other compensatory mechanisms are operative to maintain function.

2.5.5 Quantitative analysis

ProteoConnections also integrates data from quantitative proteomics experiments. Profiling changes in protein phosphorylation across conditions can provide valuable insights on key residues modulated by specific cell stimuli within signaling pathways. Abundance profiles from label-free or isotopically labeled peptides can be displayed as plots for different time points, conditions and replicates. This is examplified in Figure 2.6 for Gap junction alpha 1 (Gja1), a protein from the connexin family that forms channels between cells to facilitate the diffusion of small molecules. The sequence of Gja1 is shown in Figure 2.6A along with the location of phosphorylation sites identified in this study or from the literature. The abundance of the native phosphopeptides was monitored over the first 60 minutes following incubation of IEC-6 cells with the MEK inhibitor PD184352.

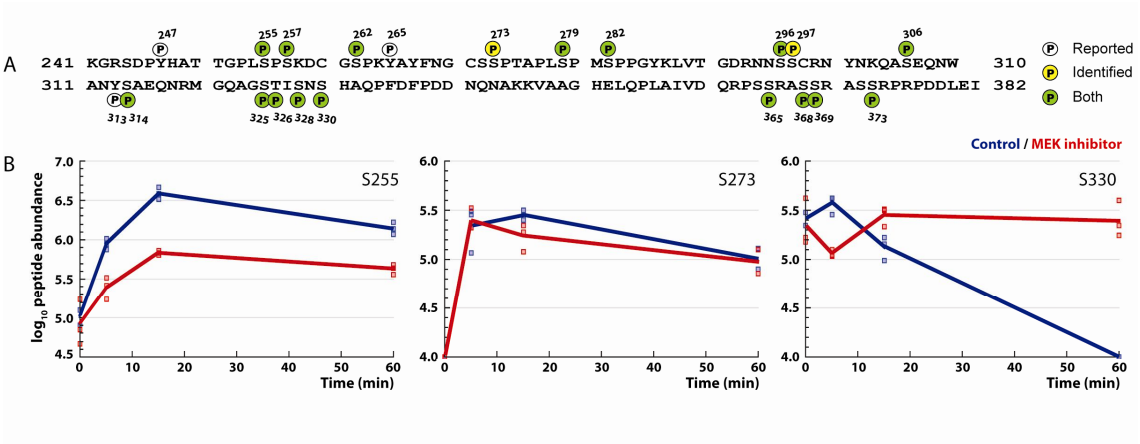


Figure 2.6: Profiling kinetic changes in protein phosphorylation of the Gap junction alpha 1 (Gja1) protein.

A) Gja1 sequence showing twenty one phosphorylation sites identified in this study or from previous reports. B) Changes in protein phosphorylation for three different sites of Gja1 following bovine serum stimulation and inhibition with MEK inhibitor (PD184352) using label-free quantitative proteomics. Graphs represent actual output from ProteoConnection for quantitative kinetic experiments.

We identified 18 phosphorylated sites of which 2 sites were not reported previously. Gja1 protein is composed of four transmembrane segments, and all identified sites are located exclusively in the cytoplasmic C-terminal region of the protein. Interestingly, the kinetic changes in phosphopeptide abundance were found to be site specific. For example, Figure 2.6B shows three distinct phosphorylation kinetic profiles for Gja1. The first profile shows the abundance of the mono-phosphorylated peptide SDPYHATTGPLpSPSK identified with a high level of confidence (Mascot score 70.4). The site is located on serine 255 and bears the PXSP consensus motif recognized by the Mitogen-activated protein (MAP) kinase family. This residue was previously reported as a known phosphorylation site of Erk1/2 kinases [175], and was found to be partly inhibited by PD184252. The profile obtained is consistent for the behavior expected for an Erk1/2 kinase substrate where phosphorylation first increases following serum stimulation (control) and subsequently decreases in presence of the inhibitor. This type of profile combined with motif search can

be used to report potential substrates of Erk1/2 MAP kinases. The second profile shows the abundance of the phosphopeptide YAYFNGCSpSPTAPLSPMSPPGYK (S273 from Gja1) identified by Mascot with a score of 62.5. This site exhibits no significant change after the inhibition, although it harbors a S/TP consensus motif recognized by MAP kinases. The third profile corresponds to the phosphopeptide MGQAGSTISNpSHAQPFDFPDDNQNAKK (S330 from Gja1) and was identified with a Mascot score of 73.9. This site was reported to be phosphorylated in vitro and in vivo by CkI [176]. In contrast to S273, this site showed an increase in phosphorylation following incubation with PD184252, suggesting that either CkI is activated or that the corresponding phosphatase is inactivated in presence of the inhibitor. These results highlight the significance of obtaining simultaneous kinetic profiles on protein substrates to precisely define sites of inhibition and determine differential regulation by kinases and phosphatases. The last example shows the signal propagation of the inhibitory effect to other pathways found in the phosphorylation network. For kinases with similar phosphorylation consensus (e.g. S/TP for Mapk and Cdk), it is however challenging to determine if observed changes originate from direct or secondary effect of kinase inhibition. Phosphorylation kinetic profiles are important to dissect phosphorylation events implicated in a particular phenotype or in a specific signaling cascade. A complete analysis of the phosphorylation kinetics will be presented separately.

2.5.6 Comparison with other processing pipelines

Many pipelines currently exist to identify peptides and proteins from the large collections of acquired MS/MS spectra. Each pipeline has its own strength and weaknesses depending on the algorithms and data formats chosen. In comparison with other identification platform, ProteoConnections does not include raw MS data file management (PrestOMIC [126], HTAPP [121]), MS/MS preprocessing step (TOPP [80]) and automated searches submission (Trans-Proteomic Pipeline [118], HTAPP [121]). ProteoConnections imports

identification from search engines and stores them in a relational database. The database structure allows quick and automated remapping of identifications to different protein sequence databases, a useful feature that enables the comparison of search outputs while avoiding the need to repeat database searches directly from raw data. This also facilitates the correlation of identifications with frequently updated version of databases. The database structure also enables efficient retrieval of peptides with post-translational modifications. ProteoConnections supports protein grouping, a feature also available in pipelines such as GPM [132], VEMS [119], and MASPECTRAS [125]. To determine false discovery rate, we implemented the target-decoy analysis while other pipelines such as TPP [118] and MASPECTRAS [125] use PeptideProphet [177]. Both methods are widely accepted but PeptideProphet provides the advantage of combining data from different search engines. ProteoConnections can also filter data based on annotations (protein domain, structure) or phosphorylation features. In contrast to other platforms such as TPP [118], HTAPP [121], or TOPP [80] that use dedicated quantification software, ProteoConnections import abundance data in a text file format to accommodate different quantitative proteomics approaches (e.g. label-free, iTRAQ, or SILAC). ProteoConnections quantification view is not meant to be a full quantification package but a view to report abundance change of peptides.

ProteoConnections provides an easy and rapid access to bioinformatics tools to annotate datasets, an important step toward the generation of biological hypotheses. Similarly to TPP [118], Prequips [120], and HTAPP [121], we integrated protein interaction network mapping tools to establish connectivity of identifications at the proteome level. ProteoConnections distinguish itself from other existing proteomics data processing pipelines by the integration of relevant bioinformatics tools for phosphoproteomics analyses in a single platform (e.g. PTM localization confidence, motifs search, kinases prediction, PDB structure, protein domain, known PTMs, Gene Ontology and conservation of phosphorylation with other species). In addition, this platform includes searches for NLS/NES proximal to phosphorylation sites, docking sequences for the MAP

kinase, known O-glycosylation sites, tools to determine modifications competing with phosphorylation, and specific filters for the selection of peptides binding to the major histocompatibility complex I (MHC). Finally, ProteoConnections uses a combination of web services and automated import scripts to update databases from external sources.

2.5.7 Completeness of phosphoproteomics dataset

The dataset used in the present study represents only a fraction of the whole rat phosphoproteome. Current phosphoproteomics experiments are biased toward abundant phosphopeptides, and are subject to variability of phosphoproteome coverage depending on the enrichment methods used. Furthermore, specific conditions used to stimulate cells will give rise to phosphorylation events defined by a subset of the entire phosphoproteome. We identified a total of 9615 distinct phosphorylation sites on 15 700 phosphopeptides, of which only 1935 sites (20 %) were previously reported in the rat UniProt database. For mouse, a closely related species that has been more extensively studied, we identified 3866 orthologous sites from 26 829 known phosphorylated sites (UniProt). The extent of the phosphoproteome repertoire is largely unknown for most species, and is estimated to exceed 100 000 phosphorylation sites in human alone, a number that is far beyond what can be identified in a single phosphoproteomics experiment. Observations made from a given experiment are likely to differ from the real sample distribution of the phosphoproteome. Correspondingly, odds-ratios will be affected by this sampling bias, a consideration that should be taken into account when interpreting results from phosphoproteomics experiments.

2.6 Concluding remarks

ProteoConnections was developed to facilitate data management, quality control and analysis of complex data sets obtained from large-scale MS-based proteomics experiments. This data analysis platform also facilitates the comparison and sharing of data sets between user groups, a feature that can improve productivity of proteomics core platforms or lab with limited bioinformatics resources. The database structure of ProteoConnections is flexible, robust and handled more than 196 projects with 3474 Mascot search jobs and a total of 8 830 083 peptide identifications (1 058 888 unique peptides). Several tools were integrated in ProteoConnections to reduce manual analysis and enable users to derive valuable biological insights for subsequent validation experiments or structural studies. A particular emphasis was placed on bioinformatics tools and programs tailored for phosphoproteome analysis that leverage phosphopeptide identifications to provide consensus motifs, potential kinases, and site conservation across different species. Using these tools we demonstrated the application of ProteoConnections on a unique phosphoproteome data set from rat intestinal IEC-6 cells comprising 9615 phosphorylation sites. The combination of site localization and three dimensional protein structures enabled the identification of important residues modulating interactions with their binding partners such as Ser28 of Histone H4 that interacts with DNA, or Ser274 of Calpastatin inhibitor that favors contacts with basic residues of Calpain small subunit 1. ProteoConnections can also incorporate quantitative phosphoproteomics datasets to profile changes in protein phosphorylation in response to cell stimulation or following incubation of specific kinase inhibitor. This capability was demonstrated for Gap-junction alpha 1 (Gja1), where specific sites are differentially phosphorylated following incubation with the Mek1/2 inhibitor PD184352.

ProteoConnections can be accessed at http://www.thibault.iric.ca/proteoconnections and the source code for this application is available from SourceForge at http://sourceforge.net/projects/proteoconnect.

2.7 Acknowledgements

We thank all proteomics lab members for valuable comments during the development, testing and implementation of ProteoConnections. Alexandre Bramoullé implemented the interface for GO. MC acknowledges the Canadian Institute for Health Research (CIHR) BiT program and the Fonds de recherche sur la nature et les technologies du Québec (FQRNT) for a graduate scholarship. IRIC receives infrastructure support funds from the Fonds de la Recherche en Santé du Québec (FRSQ) and from a Canadian Institutes for Health Research (CIHR) multi-resource grant. This work was carried out with the financial support of operating grants from the National Science and Engineering Research Council (NSERC) to PT, the Canadian Cancer Society Research Institute to SM, and from the Canada Research Chair program to SM and PT.

CHAPITRE 3: Phosphoproteome dynamics reveal novel Erk1/2 MAP kinase substrates in epithelial cells

Mathieu Courcelles[1,2], Christophe Frémin[1], Laure Voisin[1], Sébastien Lemieux[1,4], Sylvain Meloche[1,5], Pierre Thibault[1,2,3]

[1]IRIC, Institut de recherche en immunologie et cancérologie, [2]Département de Biochimie, [3]Département de Chimie, [4]Département d'informatique et de recherche opérationnelle, [5]Département de Pharmacologie, Université de Montréal, Montréal, Canada

3.1 Contribution des auteurs

Mathieu Courcelles et Pierre Thibault ont écrit le manuscrit. Tous ont révisé et commenté le manuscrit. Laure Voisin a effectué la culture cellulaire. **Mathieu Courcelles** a procédé à la préparation des échantillons et les analyses de spectrométrie de masse. **Mathieu Courcelles** a fait l'analyse bio-informatique des données du phosphoprotéome du rat et généré la liste des substrats potentiels des kinases Erk1/2. Laure Voisin et Christophe Frémin ont fait la validation des substrats des kinases Erk1/2 sélectionnés. Pierre Thibault, Sylvain Meloche et Sébastien Lemieux ont supervisé ce projet.

3.2 Abstract

The Ras-dependent Erk1/2 mitogen-activated protein (MAP) kinase pathway is a major regulator of cell proliferation, differentiation and survival. Aberrant activation of receptor tyrosine kinases or gain-of-function mutations in RAS or RAF genes are frequent hallmarks of human cancer and lead to misregulated Erk1/2 MAP kinase activities. To date, approximately 160 Erk1/2 substrates were reported, but only a limited number of them are validated. In this study, we performed quantitative phosphoproteomics and bioinformatics analyses to identify putative Erk1/2 substrates from their phosphorylation signature and kinetic profiles in rat intestine epithelial cells in response to serum cell stimulation and selective Mek1/2 inhibition with PD184352. We identified a total of 7936 phosphorylation sites within 1861 proteins, of which 145 are new putative Erk1/2 substrates. Six substrates (Ddx47, Hmg20a, Junb, Map2k2, Numa1, and Rras2) were confirmed by *in vitro* kinase assays, and provided additional evidences for the expanding roles of Erk1/2 in different cellular pathways. Immunofluorescence experiments demonstrated that the phosphorylation of Hmg20a on serine 105 by Erk1/2 affects the nucleocytoplasmic localization of this protein.

3.3 Introduction

Extracellular signal-regulated kinases 1 and 2 (Erk1/2), of the mitogen-activated protein (MAP) kinases family, are evolutionarily conserved signaling enzymes that regulate numerous cellular functions such as proliferation, differentiation, survival, and migration [178, 179]. The Erk1/2 kinases modulate the activities of transcription factors, kinases, phosphatases, cytoskeletal proteins, apoptotic proteins and proteinases [109]. Their misregulation is frequently associated with many diseases including cancer, diabetes, inflammatory disorders and neuro-cardio-facial-cutaneous syndromes [110]. Erk1/2 kinases are activated by the dual specificity Mek1/2 kinases in response to mitogenic factors, growth factors, differentiation stimuli and cytokines acting through receptor tyrosine kinases, cytokine receptors, integrins and G protein-coupled receptors. Aberrant activation of receptor tyrosine kinases or gain-of-function mutations in RAS or RAF genes often lead to hyperactivation of the Ras-dependent Erk1/2 pathway resulting in a condition that favors cancer initiation and progression. The frequent observation of these molecular anomalies in patient prompted the development of small-molecule inhibitors of Mek1/2 for targeted cancer therapy [107, 180].

The advent of affinity purification media [141], high sensitivity mass spectrometry (MS) and bioinformatics tools have facilitated the identification of thousands of phosphorylation sites from large-scale proteomics experiments, and expanded the repertoire of proteins regulated by phosphorylation. Quantitative phosphoproteomics using SILAC [181], iTRAQ [48] or label-free [182], enable the profiling of phosphopeptide abundance across different experimental paradigms to dissect cell signaling pathways and identify regulated sites. Large-scale phosphoproteomics studies enabled the profiling and identification of phosphorylation sites in hundreds of protein substrates in response to EGF stimulation [19, 20, 183, 184]. Quantitative phosphoproteomics using SILAC [110], label-

free [111], 2D-DIGE [112] or ATP analog sensitive Erk2 kinase [113] were used to identify Erk1/2 kinase substrates, but yielded limited information on the dynamic changes in protein phosphorylation of the corresponding substrates.

In the present study, we capitalized on advances in MS instrumentation, quantitative phosphoproteomics and bioinformatics to identify novel Erk1/2 substrates by correlating temporal changes in phosphorylation in response to cell stimulation and pharmacological inhibition of Mek1/2 kinases. To select putative Erk1/2 substrates, we developed a bioinformatics approach that searched for specific phosphorylation motifs and dynamic profiles from our quantitative phosphoproteomics dataset. This approach enabled the identification of 157 candidates including 12 known substrates from the dynamic profiles of 7936 identified phosphorylation sites. Six putative substrates (Ddx47, Hmg20a, Junb, Map2k2, Numa1, and Rras2) were confirmed by *in vitro* Erk1 kinase assays. Several protein substrates are involved in nucleic acid metabolic processes and alternative splicing, thus highlighting potentially new functions associated with Erk1/2 kinases.

3.4 Materials and methods

Additional details on experimental procedures can be found in Supplemental Methods (Annex 3).

3.4.1 Cell culture

Intestine epithelial rat cells (IEC-6) were grown to confluence in 150 mm Petri dishes, made quiescent by serum starvation for 24 h, and treated with DMSO (control) or with 2 μM PD184352 (Pfizer), a selective Mek1/2 inhibitor, for 1 h prior to stimulation with 10% fetal bovine serum (FBS) for different time periods (0, 5, 15, 60 minutes).

3.4.2 Cell fractionation and protein extraction

A total of 5 x 10^8 cells per biological replicate were washed twice with ice cold PBS (HyClone), collected by scrapping, lysed with lysis buffer containing protease and phosphatase inhibitors and centrifuged at 1000g for 3 min at 4°C. Supernatants were transferred to separate tubes (cytoplasmic fraction) and pellets were resuspended in lysis buffer, spun at 1000g for 3 min at 4°C, extracted and sonicated (nuclei fraction). Proteins from all fractions were precipitated overnight with cold acetone (-20°C) and resuspended in 1% SDS and 50 mM ammonium bicarbonate prior to reduction and alkylation. Proteins were quantified with microBCA protein assay (Pierce), digested with sequencing grade trypsin (Promega) overnight at 37°C. Tryptic digests were acidified and dried in a SpeedVac (Thermo).

3.4.3 Phosphopeptides enrichment and mass spectrometry

Phosphopeptides were enriched (1 mg tryptic digest/replicate) using home-made TiO_2 affinity media (Titansphere, 5 µm, GL Sciences), as described previously [141, 182]. Phosphopeptide extracts were analyzed in an interleaved fashion using a 2D-nanoLC system (Eksigent) coupled to a LTQ-Orbitrap XL mass spectrometer (Thermo Fisher Scientific). Samples were injected on SCX/C_{18} trap columns prior to their separation on a C_{18} analytical column using a linear gradient of 2-33% ACN in 53 min. The mass spectrometer was operated in a data-dependent acquisition mode with a 1 s survey scan at 60 000 resolution using a lock mass, followed by MS/MS fragmentation on the three most intense ions in the linear ion trap. All MS/MS spectra were searched with Mascot 2.1 (Matrix Science) against a concatenated target/decoy IPI rat database v3.54 (39 928 protein sequences) [143] to obtain a FDR < 1% using the following parameters: peptide mass tolerance ± 10 ppm, fragment mass tolerance ± 0.5 Da, trypsin with 2 missed cleavages, variable modifications: carbamidomethyl (C), deamidation (NQ), oxidation (M), phosphorylation (STY). All search results and identified phosphorylation sites are available in ProteoConnections [117]. Label-free proteomics quantitation was performed on all identified and detected peptides following time and *m/z* alignment [142, 182]. A median normalization procedure was applied to minimize variability of peptide ion abundances, and only reproducibly detected peptides were considered for candidate selection.

3.4.4 Selection of candidate Erk1/2 substrates

An algorithm, implemented in Perl, was written to select putative Erk1/2 substrates by selecting high confidence sites (≥ 75%) with the minimal Erk consensus [pS/T]P motifs,

and showing increase in phosphorylation after serum stimulation. Phosphopeptides were filtered with a cut-off of $\Sigma\log_{10}$(stimulated $t_{5\text{-}60}$/control t_0) ≥ 0.3, a value above most fold change found across replicates. A cut-off of $\Sigma\log_{10}$(PD184352 treated/control) ≤ -0.7 was used to select down regulated phosphopeptides upon inhibitor treatment. A two-tailed Student t-test (p-value ≤ 0.05) was performed on down-regulated phosphopeptides and manual inspection of abundance profiles was performed on all potential candidates.

3.4.5 Kinase assays and immunofluorescence microscopy analysis

Recombinant GST-tagged candidates were produced in *E.coli* and purified by adsorption to glutathione-Sepharose beads. Recombinant proteins were incubated in kinase buffer (20 mM Tris-HCl, pH 7.4, 20 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT) supplemented with 50 µM ATP and 5 µCi [γ-^{32}P]ATP for 20 minutes at 30°C in presence of 30 ng recombinant Erk1 protein. The reaction products were analyzed by SDS-PAGE, autoradiography and MS (no [γ-^{32}P]ATP). Immunofluorescence microscopy analysis was done with HA-tagged Hmg20a (pcDNA3 plasmid with CMV promoter) transiently transfected with polyethyleneimine in Madin-Darby canine kidney (MDCK) cells and stained with anti-HA antibody conjugated with Alexa fluor 488. DAPI was used to stained nucleus. At least 150 cells were scored for each coverslip.

3.5 Results

3.5.1 Phosphoproteome analyses and determination of temporal profiles

Measurement of site-specific phosphorylation dynamics was achieved using the following quantitative phosphoproteomics approach (Figure 3.1A). To profile signaling events associated with the Erk1/2 pathway, biological triplicate from two populations of intestinal epithelial rat cells (IEC-6) were stimulated for 0, 5, 15 or 60 minutes with fetal bovine serum in presence or absence of the selective Mek1/2 inhibitor PD184352. A low concentration of PD184352 (2 μM) was used to avoid simultaneous inhibition of the Mek5 kinase [110]. Proteins from cytosolic and nuclear fractions were digested with trypsin and phosphopeptides were enriched on TiO_2 micro-columns. Phosphopeptides were separated into 5 different SCX fractions and analyzed by LC-MS/MS on a hybrid linear ion trap/Orbitrap (LTQ-Orbitrap) mass spectrometer. Label-free quantitation was used to profile the abundance of identified phosphopeptides and checked manually for accuracy [142, 182]. To select putative Erk1/2 substrates, additional filtering was performed on identified phosphopeptides to keep sites with Erk1/2 consensus motifs that display specific abundance profiles upon cell stimulation and kinase inhibition (Figure 3.1B).

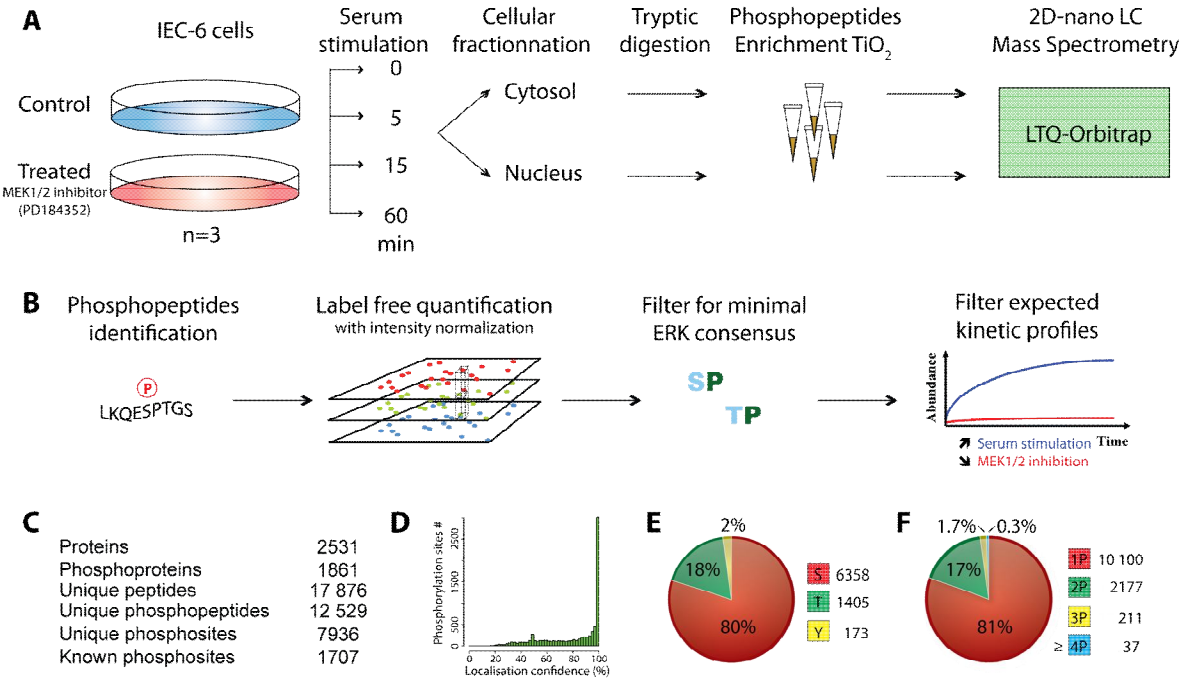


Figure 3.1: Experimental workflow and data processing for Erk1/2 substrates discovery.

A) Experimental workflow used for sample processing and MS analyses. B) Data analysis steps for the identification of Erk1/2 substrates using consensus motifs dynamic profiles of quantifiable phosphopeptides upon cell stimulation and kinase inhibition. Statistics on identified phosphopeptides (C), distribution of confidence in site localization (D), modified residues (E) and number of phosphorylation site per peptide (F).

We identified 12 529 non-redundant phosphopeptides from a total of 17 876 peptides assigned to 2551 proteins (Figure 3.1C). Phosphopeptides were typically sequenced several times in different forms containing for instance oxidized methionine or missed tryptic cleavage sites. For convenience, all identifications are available online in ProteoConnections [117]. This analysis enabled the identification of 7936 unique phosphorylation sites on 1861 proteins, of which two third represent high confidence assignment with a localization probability of at least 0.75 (Figure 3.1D and Table A3.I). A comparison of these sites with those reported in large-scale phosphoproteomics databases revealed that only 1707 sites (21%) are already identified in rat while 4660 sites (59%) are found in orthologues. Also, we identified 4055 and 5088 phosphorylation sites in the

cytosolic and nuclear fraction, respectively. The cell fractionation procedure provided an increase in phosphoproteome coverage since we obtained an overlap of only 1127 phosphorylation sites between cellular fractions. The distribution of pS, pT, and pY sites is 80%, 18%, 2% (Figure 3.1E-F), and peptides were predominantly singly and doubly phosphorylated, consistent with previous reports [185].

We obtained the dynamic profiles of 3015 and 5222 phosphopeptides from the cytosol and nucleus fractions respectively (Table A3.II). The reproducibility of abundance measurements from our label-free quantitative procedure was evaluated to determine a fold change cutoff above technical and biological variability. The standard deviation of these measurements was 37% and 95% of phosphopeptides showed less than 2-fold change across biological replicates (Figure A3.1). A cutoff of 2-fold change in either direction and a p-value less than 0.05 (two-tailed Student t-test) over kinetic profiles indicated that 2510 phosphopeptides displayed a significant change in abundance in at least one time point.

We next summed the abundance changes of the four time points using the $\Sigma\log_{10}$(PD184352 treated/control) to determine the overall inhibitory effect on all quantifiable substrates. A normal distribution centered on zero was obtained for both cytosol and nucleus extracts, and indicated that Mek1/2 inhibition resulted in equal distribution of up- and down-regulation of protein phosphorylation (Figure A3.2). It is noteworthy that this distribution can also reflect the activity of other kinases (CkI, Cdk, Gsk3, Jnk1, Sapk and other MAP kinases). However, many phosphopeptides with the minimal Erk1/2 consensus were down-regulated in both cytosol and nucleus indicating that a sizable proportion of them could be direct Erk1/2 substrates.

3.5.2 Phosphoproteome dynamics identify potential Erk1/2 substrates

Cell stimulation with serum activates the Erk1/2 pathway and gives rise to a signaling cascade that mediates phosphorylation of a large number of protein substrates. Changes in phosphorylation were determined for several kinases and protein substrates upstream of Erk1/2, including Pak1, Raf1, Map3k1, Mek1/2, Ksr1, Rras2 and Ywhae, (Figure 3.2A). Phosphorylated sites on Raf1-S621, Mek1-S222 and Mek2-S226 showed the expected hyperactivation upon incubation with PD184352 resulting from negative feedback loops [186, 187]. A decrease in phosphorylated Ywhae-S210 was observed upon kinase inhibition, consistent with its negative regulation for effective binding to activated Raf1-S621 [188]. Interestingly, phosphorylated sites from the same protein were observed to be regulated differently in presence of the Mek1/2 inhibitor. Our analyses also identified known Erk1/2 substrates including Fosb-S116, Gja1-S255, Pxn-S83 and Stmn1-S25 that all exhibited down-regulation of phosphorylation upon Mek1/2 inhibition (Figure 3.2B). These examples confirmed previous findings reported on known Erk1/2 substrates and validated the data mining approaches used in the present study.

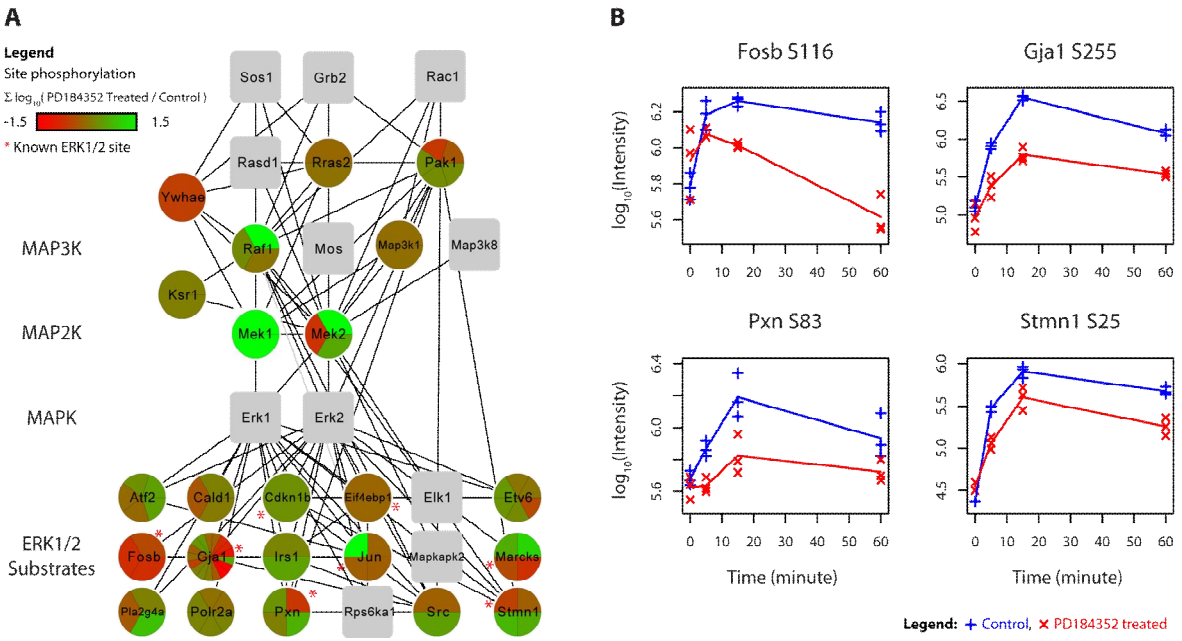


Figure 3.2: Identification of interacting proteins of the Ras-Raf-Mek-Erk1/2 MAP kinase pathway.

A) Identified phosphoproteins mapped on the interaction network from the STRING database. A color gradient is used to represent the modulation of each phosphorylation site following PD184352 treatment. B) Site-specific changes of phosphorylation for four known Erk1/2 substrates in response to cell stimulation (control) and Mek1/2 inhibition.

To determine potential Erk1/2 substrates, we analyzed phosphorylation motifs, cell stimulation and inhibition profiles (Figure 3.1B). The Erk1/2 kinases phosphorylate the minimal consensus motif [pS/T]P and more efficiently at the optimal motif PX[pS/T]P [189]. We identified 2296 high confidence phosphosites on 987 proteins (1092 sites in cytosol and 1816 in nucleus) that contained the minimal motif. We then selected phosphopeptides clusters in both nuclear and cytosol extracts that are positively regulated in abundance upon serum stimulation and negatively following Mek1/2 inhibition. We obtained 157 putative Erk1/2 substrates (233 sites), including 52 sites (22%) that contained the optimal Erk1/2 phosphorylation consensus PX[ST]P motif (Figure 3.3 and Table A3.III).

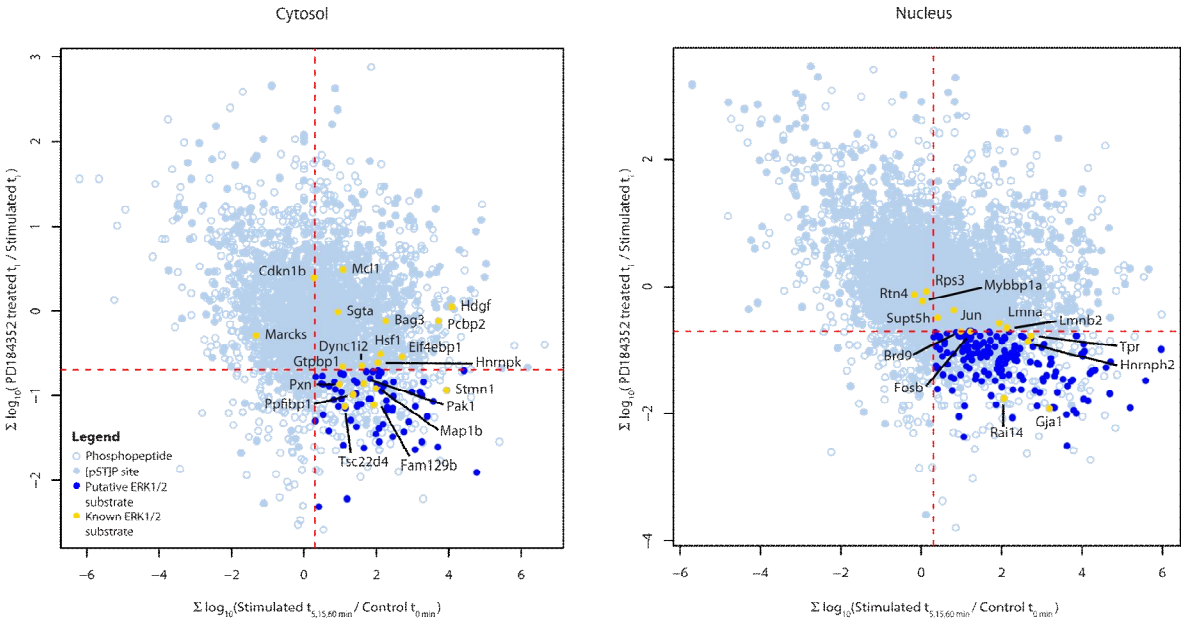


Figure 3.3: Dynamic changes of protein phosphorylation identify Erk1/2 substrates from cytosol and nuclear extracts.

Bi-dimensional representation of changes observed in protein phosphorylation following cell stimulation and Mek1/2 inhibition where each point represents a different phosphopeptide. Putative Erk1/2 substrates are shown as dark blue circles in the bottom right quadrant for sites displaying increase in phosphorylation upon serum stimulation and decrease following PD184352 treatment (two-tailed t-test). Yellow circles correspond to known Erk1/2 substrates.

We also analyzed the time profiles of regulated phosphopeptides using fuzzy c-means clustering and obtained 6 groups that displayed representative stimulation and inhibition trends from our substrates dataset (Figure A3.3). Bioinformatics analyses revealed that 94% of these sites and their following proline are conserved in either human or mouse. Additional specificity of Erk1/2 kinases is conferred by binding to docking sites such as the DEF domain or D domain [190]. Our analyses indicated that 22 sites and 1 site were located within D or DEF domains respectively.

We compared our list of potential Erk1/2 substrates with a review that regroups almost 160 members identified from different experimental approaches and cell lines [109]. Potential candidates were also compared with recent reports using alternate Mek1/2 inhibitor, or an ATP analog sensitive Erk2 kinase (Figure 3.4 and Table A3.V). A total of 32 Erk1/2 substrates proteins including Atf2, Cald1, Etv6, Fosb, Gja1, Jun, Marcks, Pla2g4a, Pxn, Src, and Stmn1, overlapped with these studies. The observed overlap between substrates in all the studies was limited and often the identified site on substrates differed.

3.5.3 Comparison of ERK1/2 substrates discovered by phosphoproteomics studies

We then compared our list with other MS phosphoproteomic studies (Figure 3.4). The use of U0126, an alternate MEK1/2 inhibitor, or an ATP analog sensitive ERK1 kinase were central to substrates targeting but analytical strategies deployed were different in each. The analytical strategies deployed in each study were different. The work done by Hattori's group used mouse fibroblasts (NIH3T3), IMAC, 2D-DIGE, phosphomotif-specific antibodies and MALDI-TOF MS (peptide mass fingerprint) to discover ERK1/2 dependent changes [191]. They detected 37 spots with more than 1.5-fold change and found 24 new candidates ERK1/2 targets. Their candidates were selected with antibody reactivity to ERK1/2 consensus. Thirteen of them were validated by in vitro kinase assays. Unfortunately, the mass spectrometry methodology used did not provide precise localization of phosphorylated sites. Thus, our comparison could only be done at the protein level. An overlap of seven proteins was found with our list and seven with the reference dataset. In another study by Ahn's group, a human melanoma cell line (WM115) stimulated with EGF was used with a negative ionization on a QTrap mass spectrometer to

detect phosphopeptides based on the fragment ion signature PO3- [111]. Label-free quantification reported 90 sites with changes (≥ 1.7-fold). Only 28 sites on 15 proteins were [pS/T]P sites and decreased in abundance. Five proteins are in our list but only two had the same phosphorylated sites and six are in the reference dataset. In work by Mann's group, EGF stimulated HeLa cells were used, phosphopeptides enriched by TiO2, separated by SCX and quantified by SILAC on LTQ-Orbitrap [192]. They reported that 85% of phosphopeptides were not influenced by inhibitors or growth factor (< 2-fold change) and that about 6.6% were potential substrates ERK1/2 (35 proteins), showing both EGF-up regulation and U0126 inhibition. Ten proteins were common between the two studies but 6 have sites located at the same positions and four proteins are common to the reference dataset. In a recent study, ERK2 substrates in mouse fibroblasts (NIH 3T3-L1) were discovered using an ATP analog-sensitive ERK2 kinase that tag substrates in vitro [113]. Substrates were isolated using the thiophospho tag and further enriched by IMAC. Their approach identified 80 ERK2 substrates of which 13 were known. Five substrates have the exact sites found in our study and thirteen more are common but at an alternate position. Overall, the overlap is small between MS studies and with the reference dataset (Figure 3.4). Finally, our analysis strategy was more comprehensive and produced the longest list ERK1/2 candidates list of all phosphoproteomics studies.

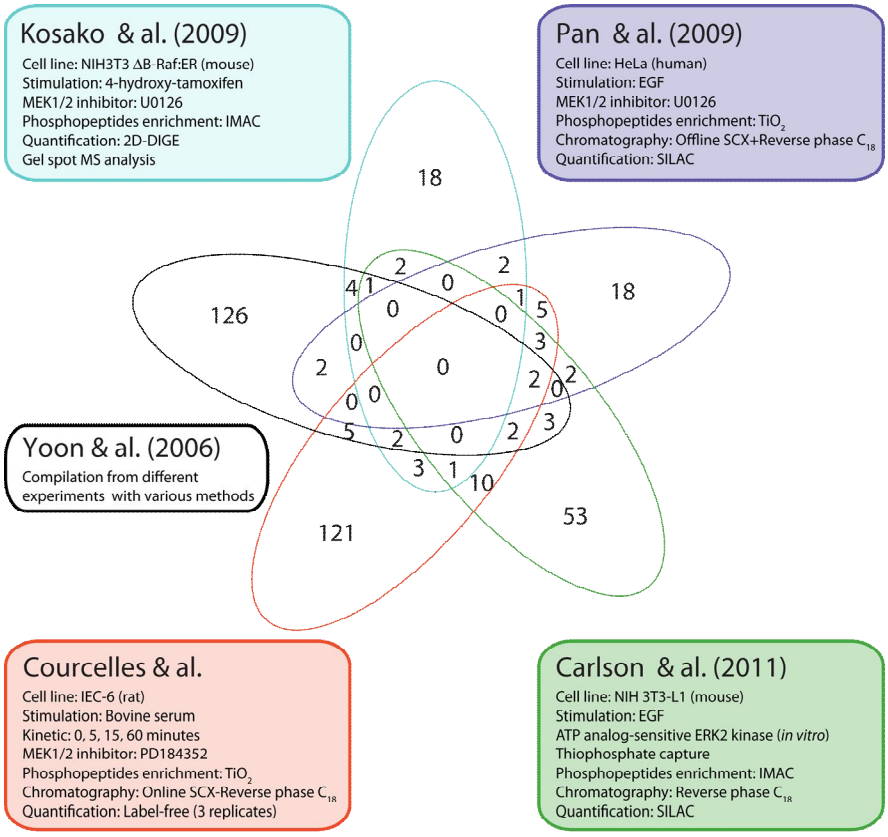


Figure 3.4: Comparison of Erk1/2 substrates between phosphoproteomics studies

A compilation dataset by Yoon & al, three phosphoproteomics experiments and our substrates list were compared all against all to find overlap and differences between reported Erk1/2 substrates. The comparison was done at protein level since the dataset of Kosako & al. did not report the position of phosphorylated sites.

Using Gene Ontology (GO) annotations, we observed that potential Erk1/2 substrates share common functions with known Erk1/2 substrates (Table A3.IV). Several proteins were found with related GO terms: transcription regulation (21 proteins, GO:0006350), cytoskeleton (15, GO:0005856), signaling (13, GO:0007242), kinase (6, GO:0016301), apoptosis (4, GO:0006915) and phosphatase (1, GO:0004721). Further GO analyses also revealed that the putative Erk1/2 substrates are enriched by 5.4 fold for the term cell junction (10 proteins, GO:0030054, p-value $\leq 8.63 \times 10^{-6}$). Other categories such

as nucleic acid binding (51 proteins, GO:0003676) and nucleic acid metabolic process (29 proteins, GO:0006139) are represented with high frequency. Interestingly, we noted significant enrichment (> 10-fold) for two related categories associated with nuclear speck (4 proteins, GO:0016607, p-value $\leq 1.25 \times 10^{-5}$) and spliceosomes (3 proteins, GO:0005681, p-value ≤ 0.002). Closer inspection of these data revealed that these two categories comprise members of the SR protein family associated with RNA splicing (e.g. Sfrs 1, 2, 6, 11). SR proteins are highly regulated by the activity of multiple kinases, namely Akt, Gsk3, Srpk, Clk, that affect their nuclear import and export, and their recruitment to nascent transcripts [193]. Our data suggest that several SR proteins are potential Erk1/2 substrates, a novel finding that extend the role of these kinases to the regulation of protein-splicing factors.

3.5.4 Erk1/2 substrates identified by quantitative phosphoproteomics show site-specific phosphorylation by Erk1 *in vitro*

From the 157 potential substrates, we next selected potential Erk1/2 substrates for further validation, using *in vitro* kinase assays, based on the biological interest, functional diversity, site with functional hypothesis that can be easily tested, the availability of reagents and the feasibility of site-directed mutagenesis experiments. Six candidates with different function were selected for validation: Ddx47 (RNA helicase or hydrolase), Hmg20a (chromatin regulator), Junb (transcription factor), Mek2 (kinase), Numa1 (structural component of nucleus) and Rras2 (signaling) (Table 3.I). Hmg20a and Junb were selected because their site is proximal to DNA binding and NLS motifs (effect on cellular localization can be easily determined experimentally by immunofluorescence). Mek2 and Rras2 were two interesting proteins because they are members the Erk1/2 pathway. Ddx47 and Numa1 were selected only for substrate confirmation purpose since GST-protein constructs were available. All sites except Mek2 had the optimal PXSP

consensus but none had a proximal D or DEF motif. Differences were observed in their respective stimulation and inhibition profiles (Figure 3.5A).

Table 3.I: Putative Erk1/2 substrates confirmed by *in vitro* kinase assay

Gene Symbol	Description	Molecular function	Position	Cellular fraction	$\Sigma \log_{10}$ (Stimulated t_i / Control t_0 min)	$\Sigma \log_{10}$ (PD184352 treated t_i / stimulated t_i)
Ddx47	DEAD (Asp-Glu-Ala-Asp) box polypeptide 47	Probable RNA helicase or hydrolase	S9	Nuclear	0.46	-0.81
Hmg20a	High mobility group 20A	Chromatin regulator	S105	Nuclear	3.86	-2.00
Junb	Transcription factor jun-B	Transcription factor	S256	Nuclear	1.20	-1.29
Map2k2	Dual specificity mitogen-activated protein kinase kinase 2	Kinase	S295	Cytosol	2.15	-0.95
Numa1	Nuclear mitotic apparatus protein 1 isoform 2	Probable structural component of nucleus	T1994	Nuclear	0.47	-0.99
Rras2	Related RAS viral (R-ras) oncogene homolog 2	Signaling	S186	Cytosol	2.98	-0.96

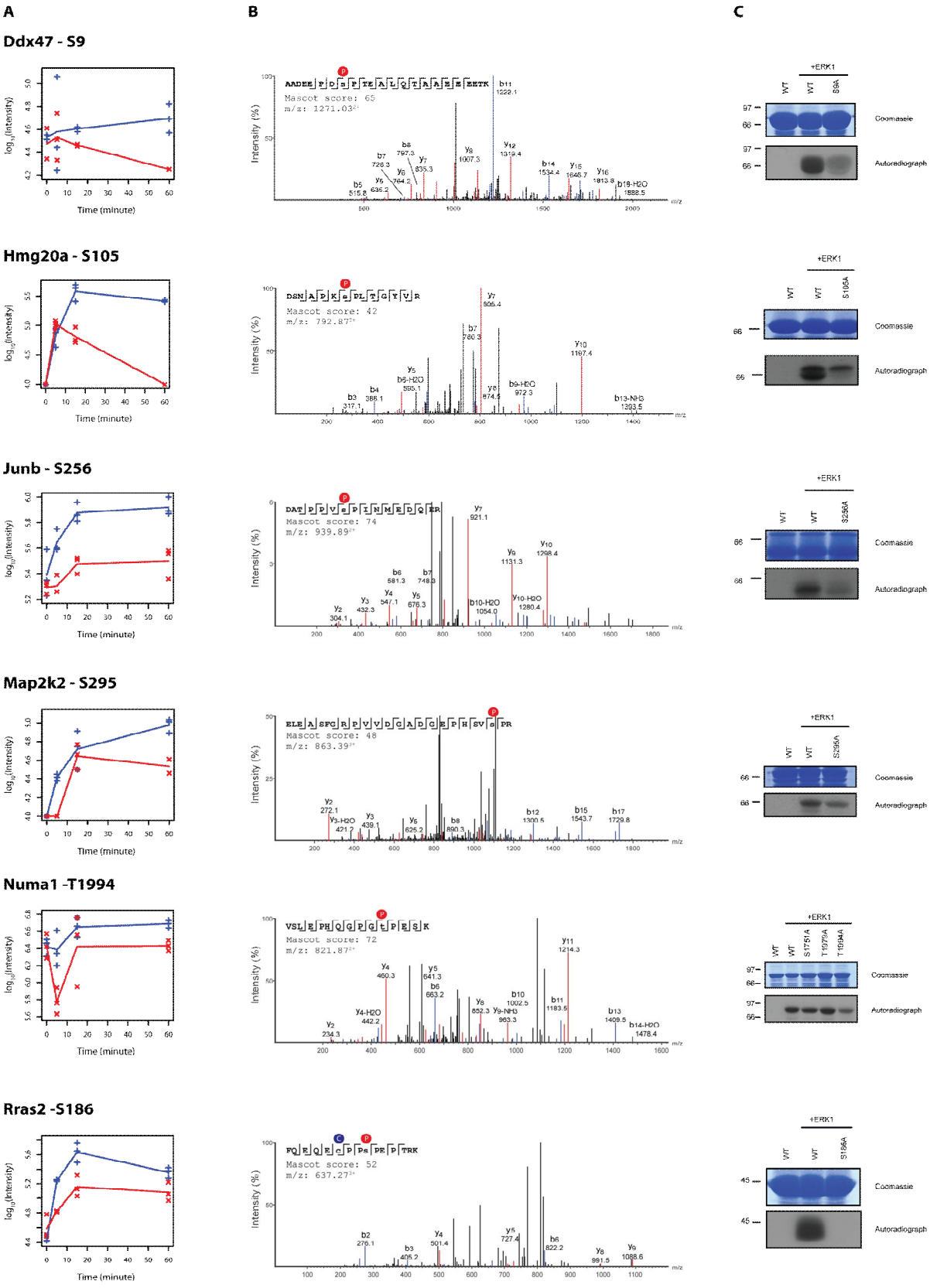
Time point t_i : 0, 5, 15, 60 minutes

Substrates were expressed as GST-fusion proteins to perform *in vitro* kinase assay with active Erk1. In the case of Mek2, the inactive GST-Mek2-K101 mutant was used to avoid potential kinase autophosphorylation. Alanine mutants to each identified phosphorylation site were obtained to confirm enzyme specificity. These included GST-Ddx47 (S9A), GST-Hmg20a (S105A), GST-Junb (S256A), GST-Mek2-K101 (S295A), GST-Numa1 (T1994A) and GST-Rras2 (S186A). We performed *in vitro* kinase assays on all 6 substrates and confirmed the kinase specificity for sites identified *in vivo* from our quantitative phosphoproteomics analyses (Figure 3.5C). All sites had a phosphorylation decrease for the alanine mutant. Except Rras2, residual phosphorylation is observed after mutagenesis indicating the presence of alternate phosphorylated site in the *in vitro* kinase assay. Figure 3.5A-B shows the dynamic changes of phosphorylated S105 from Hmg20a following Mek1/2 inhibition, and the MS/MS spectrum confirming the location of the modified residue. The autoradiograph obtained from wild type (WT) and S105A mutant Hmg20a for the *in vitro* kinase assay indicated that S105 is directly phosphorylated by Erk1

(Figure 3.5C). The higher band on this gel is probably a protein contaminant from the bacterial expression system that gets phosphorylated by Erk1.

Figure 3.5: Phosphorylation profiles, MS/MS spectrum and *in vitro* kinase assay experiments for validated Erk1/2 substrates

A) Observed phosphorylation kinetic profiles of Erk1/2 substrates after serum stimulation (blue) and with the Mek1/2 inhibitor PD184352 (red). B) Confirmation of the phosphorylated position in kinase assays or *in vivo* by CID MS/MS analysis. C) Erk1 *in vitro* kinase assays with wild type and alanine mutant candidate substrates. GST-tagged protein (10 μg) incubated with and without active Erk1 for 30 min at 30°C in presence of [γ-^{32}P]ATP. Coomassie-stained gel (top) and autoradiograph (bottom) of phosphorylation reaction products.


A
B
C
Ddx47 - S9
Hmg20a - S105
Junb - S256
Map2k2 - S295
Numa1 -T1994
Rras2 -S186
Time (minute)
Intensity (%)
m/z
+ERK1
WT
S9A
S105A
S256A
S295A
S1751A
T1979A
T1994A
S186A
Coomassie
Autoradiograph

Interestingly, serine 105 from Hmg20a displays the optimal Erk phosphorylation consensus PX[ST]P motif, and is located within 10 residues of a putative nuclear localization signal (NLS) (Figure 3.6A). Serine 105 and the NLS also show a high degree of conservation among mammalian orthologs (not shown). These observations prompted us to examine if the phosphorylation of this site could regulate the nuclear localization of Hmg20a. Immunofluorescence experiments were performed on MDCK cells transiently transfected with HA-tagged Hmg20a WT or S105A mutant (SA). Immunostaining with anti-HA antibody revealed the preferential nuclear localization of the Hmg20a WT in exponentially proliferating cells (Figure 3.6B). Independent measurements of the Hmg20a distribution in the nucleocytoplasm (N/C), or nucleus only (N) on at least 150 cells confirmed that WT Hmg20a containing a phosphorylatable Ser105 residue show 16% increase in nuclear localization compared to its corresponding SA mutant (Figure 3.6C). This preliminary observation suggests that phosphorylated Ser105 residue affects Hmg20a nucleocytoplasmic distribution.

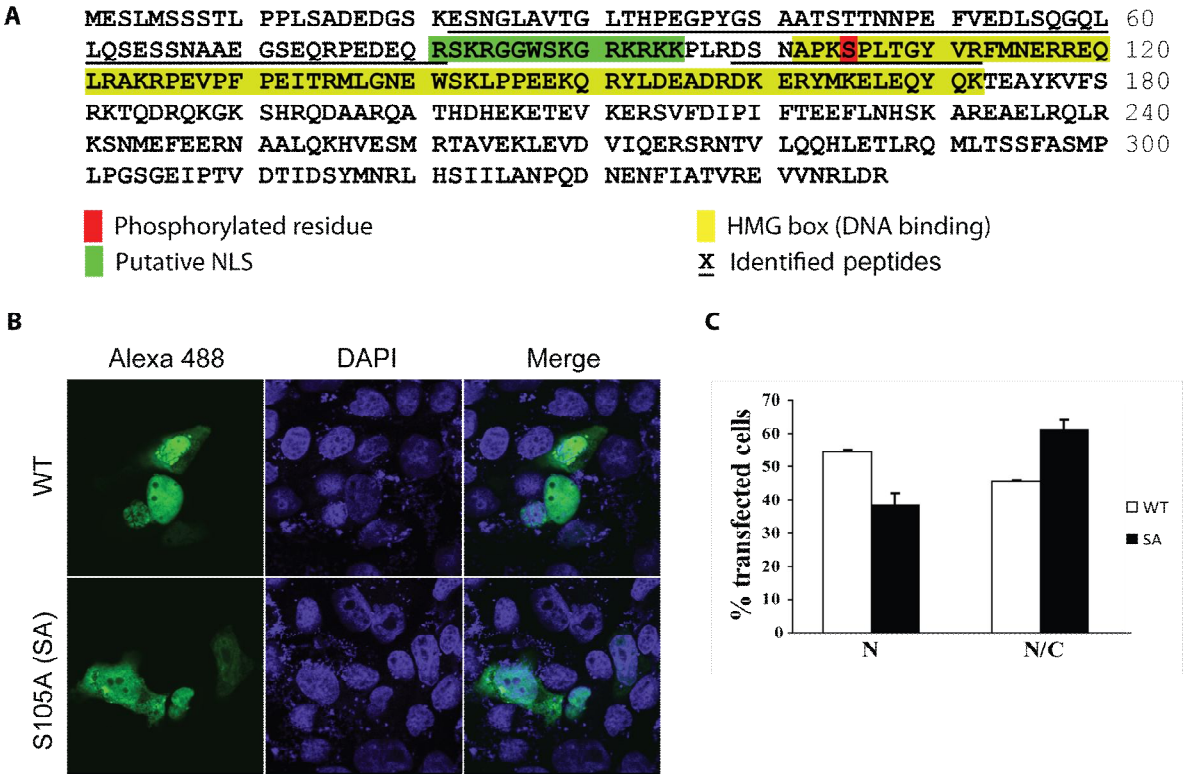


Figure 3.6: Phosphorylation of Ser105 Hmg20a influences its nucleocytoplasmic distribution.

A) Amino acids sequence of Hmg20a indicating the location of the identified phosphorylation site and the proximal NLS. B) Immunofluorescence microscopy analysis of the subcellular localization of a HA-tagged Hmg20a transiently transfected in MDCK cells. Transfected Hmg20a is stained with anti-HA antibody conjugated with Alexa fluor 488 and nucleus with DAPI. C) The localization of ectopic Hmg20a in exponentially proliferating cells was quantified by fluorescence microscopy. The cellular localization of Hmg20a was scored in two categories: distribution in the nucleus and in the cytoplasm (N/C), only nuclear (N). The bar graph represents the mean ± S.E. of two independent experiments.

3.6 Discussion

This report makes use of quantitative phosphoproteomics and specific kinase inhibitor to identify putative Erk1/2 substrates in complex cell extracts. Dynamic changes of protein phosphorylation were systematically assessed on 7936 different phosphorylation sites to determine patterns consistent with specific Erk1/2 kinase activities. The underlying premise is that substrates with characteristic consensus motifs would be distinguished by an increase in phosphorylation in response to Erk1/2 activation, and a decrease upon specific Mek1/2 kinase inhibition. Based on a previous report, protein expression should remain unchanged for the 60 min activation and inhibition period, and thus abundance changes observed can be attributed to rapid modulation of protein kinases and phosphatases activities [111]. The utility of this approach was evaluated here with an attempt to identify candidate substrates for the protein kinase Erk1/2 in IEC-6 epithelial cells.

As indicated earlier, other MS studies have searched for Erk1/2 substrates but using the Mek1/2 inhibitor U0126, a different inhibitor than the one used in this study, or an ATP analog-sensitive Erk2 kinase [109, 112, 113, 192]. Overall, the overlap is ratter small between all the phosphoproteomics studies and with the Erk1/2 substrates list compiled by Seger and co-workers. An important disparity in biomaterials and protocols used by the different labs could explain this difference. The overlap with our list and other studies could have been greater if some filtering parameters were relaxed. In some cases, localization confidence of phosphorylation and amplitude of inhibition were not sufficiently high to pass the defined filter criteria. Furthermore, it is possible that some potential substrates reported in the other studies are in fact Erk5 substrates. Indeed in other studies, U0126 was used at a concentration of at least 10 uM which is sufficient to inhibit Erk5 activity. Our protocol with PD184352 was used with an optimized concentration to avoid inhibition of Erk5 [194]. Our analysis suggested the highest number of putative substrates

and was the one with the highest number of identified phosphorylation sites. The cellular fractionation provided more identifications and the kinetic analysis permitted the detection of short and late phosphorylation of substrates. Finally, the poor overlap between studies is raising a concern about false positive hits. While each study validated a few substrates, the level of false positive in the reported lists is unknown. Until systematic evaluation of putative substrates is done by new and faster methods, it is impossible to determine the best approach to detect substrates in term of sensitivity and specificity.

Quantitative proteomics and phosphorylation dynamics enabled the identification of 157 putative Erk1/2 substrates (233 sites), thus representing the largest pool of Erk1/2 substrates identified in a single study (Figure 3.4). While the regulation of protein phosphorylation was observed for sites having the minimal Erk consensus motif, it is possible that our list of putative Erk1/2 substrates contains non-direct targets as recently reported in a large-scale inhibitory screen for 97 kinases [195]. This consensus motif is not exclusive to Erk1/2 kinases and is also shared with CkI, Cdk, Gsk3, Jnk1, Sapk and other MAP kinases [116]. However, only three sites in our list were annotated with conflicting kinases (Cdk5, Pkca, p38-delta). Interestingly, the majority of putative substrates share related functions with known Erk1/2 substrates and several proteins are associated with nucleic acid metabolism and splicing. Several lines of evidences suggested that Erk1/2 activation can modify alternative splicing of CD44 pre-mRNA transcript and can phosphorylate Sam68, a protein present in the splicing complex [110]. Dasatinib inhibition of Bcr-Abl, a kinase acting upstream of Erk1/2, also lead to decrease in the phosphorylation of proteins implicated in RNA splicing [110]. This suggests that the Erk1/2 signaling pathway could be involved in these processes never reported for known Erk1/2 substrates.

The validation and characterization of bona fide Erk1/2 substrates require different types of experiments including *in vitro* kinase assay, site localization with MS/MS, site mutagesis, and kinase knockout. As part of this study, we validated 6 putative substrates using *in vitro* kinase assays on WT and alanine-mutant proteins, and showed site-specific phosphorylation by Erk1. Phosphorylation of residues near NLS are known to modulate the

nuclear translocation of protein substrates [196]. It is noteworthy that log-odds ratio of phosphorylated sites vs. non phosphorylated sites for different protein domains of our entire phosphoproteome data set indicated an almost two-fold enrichment ($p < 2.8 \times 10^{-42}$) in sites proximal to NLS [117]. Two validated Erk1/2 substrates (Hmg20a and Junb) and three other putative substrates (Marcks, Sh3pxd2a, and Zfp148) are phosphorylated at residues in close proximity to NLS. This observation raises the possibility that Erk1/2 phosphorylation on those 5 five substrates mediates nuclear translocation. This hypothesis was verified for Hmg20a protein. Our preliminary results indicated that phosphorylation of Hmg20a at Ser105 favors its accumulation in the nucleus. Hmg proteins bind DNA, modulate chromatin structure, and can control the expression of different genes [197]. Hmg20a Ser105 is located at the beginning of the HMG box which is a DNA binding module. Further experiments should be done to determine if the nuclear translocation is modulated either by the nuclear import machinery via NLS or by nuclear retention via HMG box DNA binding.

In summary, the present contribution presents the first phosphoproteome dynamic study of the Erk1/2 MAP kinase pathway using a specific Mek1/2 inhibitor. Taken together, these data provide one of the most comprehensive phosphoproteomics study to identify Erk1/2 substrates, and provide a unique resource for the development of pharmacological inhibitors in the treatment of cancers that harbor activating mutations in the MAP kinase signaling pathway.

3.7 Acknowledgments

We thank Éric Bonneil for the technical assistance during the MS analyses and Ivan Topisirovic for the cellular fractionation protocol. MC acknowledges the Canadian Institute for Health Research (CIHR) BiT program and the Fonds de recherche sur la nature et les technologies du Québec (FQRNT) for a graduate scholarship. IRIC receives infrastructure

support funds from the Fonds de la Recherche en Santé du Québec (FRSQ) and from a Canadian Institutes for Health Research (CIHR) multi-resource grant. This work was carried out with the financial support of operating grants from the National Science and Engineering Research Council (NSERC) to PT, the Canadian Cancer Society Research Institute to SM, and from the Canada Research Chair program to SM and PT.

CHAPITRE 4: Algorithms to detect phosphopeptide positional isomers

Mathieu Courcelles[1,2], Gaëlle Bridon[1,2], Sébastien Lemieux[4], Pierre Thibault[1,2,3]

[1]IRIC, Institut de recherche en immunologie et cancérologie, [2]Département de Biochimie, [3]Département de Chimie, [4]Département d'informatique et de recherche opérationnelle, Université de Montréal, Montréal, Canada

4.1 Contribution des auteurs

Mathieu Courcelles est le principal auteur du manuscrit. Gaëlle Bridon a contribué à l'écriture et aux figures. Les autres auteurs ont révisé et édité le manuscrit. Gaëlle Bridon a effectué la culture cellulaire et la préparation des échantillons. Gaëlle Bridon et **Mathieu Courcelles** ont effectué la spectrométrie de masse et l'analyse des données. **Mathieu Courcelles** a développé les algorithmes pour détecter les isomères positionnels des phosphopeptides. Pierre Thibault et Sébastien Lemieux ont supervisé cette analyse.

4.2 Abstract

The last decade has seen remarkable advances in phosphoproteomics that facilitated the detection of thousands of phosphorylation sites in a single experiment. The development of phosphopeptide enrichment protocols, combined with the enhancement of mass spectrometer sensitivity and bioinformatics algorithms have contributed to this success. However, some phosphopeptides are difficult to identify and some sites remain elusive. Also, the precise localization of phosphorylation site can be challenging when several phosphorylatable residues are present within the same proteolytic peptide. This difficulty is further exacerbated by the possibility of having combinatorial distribution of phosphorylation sites giving rise to different phosphopeptide positional isomers. These peptides have similar physicochemical properties and their separation by RP-HPLC has been reported to be problematic in some cases. Relatively few studies have described the frequency and distribution of phosphoisomers in large-scale phosphoproteomics experiments and no convenient informatics tools currently exist to facilitate their detection. To address this analytical challenge, we developed two algorithms to detect separated phosphopeptide isomers from LC-MS experiments and recognize co-eluting isomers from mixed MS/MS spectra. Using these algorithms, we determined that the proportion of phosphopeptide positional isomers present in large-scale phosphoproteomics studies from mouse, rat and fly cell extracts correspond to approximately one percent of all identified phosphopeptides. While conventional analysis can identify chromatographically separated phosphopeptides, a targeted LC-MS/MS analysis using inclusion list generated complementary identification.

4.3 Introduction

Phosphorylation is a frequent and reversible protein post-translational modification imparting structural changes on the hydroxyl group of serine, threonine and tyrosine residues. Phosphorylation modulates protein activity and mediates interactions with other molecules. This protein modification regulates different biological processes and its misregulation is often associated with many human diseases, including cancer. The necessity to identify and profile this modification in thousands of different proteins simultaneously has prompted the development of large-scale phosphoproteomics studies. Over the past decade, the development of phosphopeptide enrichment procedures, together with improvement in mass spectrometer sensitivity and bioinformatics algorithms to facilitate phosphopeptide detection have literally revolutionized the fields of cell biology and cell signaling [13].

Phosphoproteomics is now routinely performed in many labs as part of discovery studies, though several difficulties need to be resolved to achieve comprehensiveness of analyses. For example, current mass spectrometry tools enabling the sequencing of phosphopeptides do not consistently provide unambiguous localization of the modified residue within the peptide backbone, and about one third of identified phosphorylation sites have ambiguous assignments [117]. This problem can arise from two situations. First, the labile nature of the phosphoester bound and the poor quality of phosphopeptide fragmentation can account for a sizeable proportion of ambiguous identifications. It is well recognized that under CID fragmentation, some phosphopeptides undergo a neutral loss of phosphoric acid (H_3PO_4 , -98 Da), a situation that can prevent site localization due to low abundance of fragment ions [13]. Precursor ion activation via ETD typically leaves phosphoester bonds intact and favors consecutive cleavages of backbone C-N bonds [110]. Different algorithms were designed to evaluate the probability that the position assigned by

database search engines is correct. These include Ascore [70], PhosphoScore [198], PhosphoScan [72], PhosCalc [73] for CID MS/MS, and, SLoMo [74], Phosphinator [38] that can be used for both CID/ETD MS/MS spectra. The presence of co-eluting phosphopeptide positional isomers corresponds to a second case where localization of phosphorylation can be ambiguous (Figure 4.1). Those peptides have the same amino acid sequence but are phosphorylated at different serine, threonine or tyrosine residues. In this situation, the MS/MS analysis results in a mixed spectrum that lead to ambiguous interpretation due to confounding fragment ions from both phosphoisomers. Software tools mentioned above were not designed to distinguish positional isomers and may not use sufficient spectral information to distinguish different possibilities. The probabilities calculated by these algorithms will favor candidates with the highest number of correlated fragment ions or report ambiguous localization when different candidates are possible. These softwares do not typically report if uncertain localization is due to lack of structurally relevant fragment ions or conflicting fragments originating from different isomers.

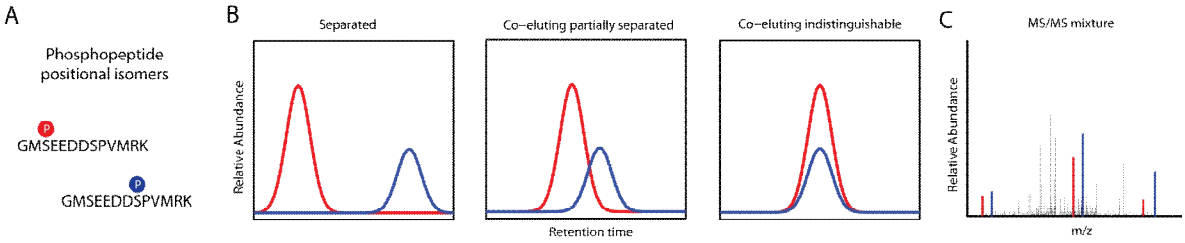


Figure 4.1: Phosphopeptide positional isomers separation and detection.

A) Phosphopeptide positional isomers are peptides with the same amino acid sequence but phosphorylated at different position. B) Under liquid chromatography, phosphopeptide isomers can be detected if they are completely or partially separated. C) The mixture of two co-eluting phosphopeptide positional isomers can be detected from a MS/MS spectrum by two distinct sets of fragments.

Separation of phosphopeptide positional isomers is challenging because they share similar physiochemical properties. Those were first reported for the separation of synthetic

phosphopeptide isomers from NEFM and microtubule-associated protein τ using reverse-phase liquid chromatography (RP-HPLC) [199, 200]. Other examples of phosphoisomers were encountered in large scale phosphoproteomics including peptides from: Fus3p [201], Erk2 [202], Egfr [203], Dok1 [204], and Sprouty2 [205]. Different chromatography media such as monolithic column [14] and HILIC with aminopropyl stationary phase [15] were reported to facilitate the separation of phosphopeptide isomers. More recently, capillary zone electrophoresis [16] and FAIMS [17] were used as alternative techniques to separate phosphoisomers. However, the frequency and difficulties associated with phosphoisomers from large-scale phosphoproteomics studies still remain unknown.

Furthermore, the detection of isomeric phosphopeptides by mass spectrometry can be problematic depending on the MS/MS acquisition method used. For example, the dynamic exclusion time windows (typically 1-5 minutes) can be too long to efficiently identify closely eluting phosphoisomers. In those cases, the instrument will trigger an MS/MS for the first isomer while the second form will be excluded. Also, the MS/MS scan grouping function often applied at the data processing can sum two or more spectra together making it impossible for the search engine to differentiate potential isomers. Therefore, proper tailoring of acquisition methods and data processing are required to identify potential isomers.

The primary goal of the present study is to detect and target phosphopeptide positional isomers in order to determine their occurrence in large-scale phosphoproteomics experiments. To achieve this goal, we developed two algorithms to detect separated isomers from the LC-MS elution profiles and identify co-eluting isomers from mixed MS/MS spectra. An inclusion list of potential phosphopeptides isomers is then generated for subsequent targeted LC-MS/MS experiments. These new algorithms facilitate the identification of biological relevant phosphorylation events that could have been missed in typical phosphoproteomics analysis. This approach is complementary to data-dependant acquisition and provided additional identification that expand the scope of large-scale phosphoproteomics studies.

4.4 Materials and methods

4.4.1 Materials

Modified porcine sequencing grade trypsin was obtained from Promega (Madison, WI, USA). Acetonitrile (ACN) and HPLC grade water were purchased from Fisher Scientific (Whitby, ON, Canada). Ammonium bicarbonate and formic acid were obtained from EM Science (Mississauga, ON, Canada). Ammonium hydroxide, trifluoroacetic acid (TFA), DTT (DL-dithiothreitol), iodoacetamide, sucrose, $MgCl_2$, $CaCl_2$, Hepes, NaCl, glycerol, protease inhibitor cocktail (4-(2-aminoethyl)benzenesulfonyl fluoride [AEBSF], pepstatinA, E-64, bestatin, leupeptin, and aprotinin), phosphatase inhibitor cocktail (sodium vanadate, sodium molybdate, sodium tartrate, and imidazole), serum-free and protein-free insect medium were purchased from Sigma-Aldrich (Oakville, ON, Canada). Bradford protein assay was obtained from Bio-Rad (Mississauga, ON, Canada). Lactic acid was purchased from Fluka (Saint Louis, MO). Titanium dioxide bulk (5 μm, 500 mg) was obtained from Canadian Life Science (Peterborough, ON). Bond breaker TCEP (tris[2-carboxyethyl] phosphine) was purchased from Pierce Biotechnology Inc. (Rockford, IL). Phosphate buffered saline (PBS) and ethylene-diamine-tetra-acetic acid (EDTA) were obtained from HyClone (Thermo Scientific, Logan, UT). Oasis HLB cartridges (1 cc, 30 mg) were purchased from Waters (Milford, MA). Capillary columns for LC-MS were packed in-house using Jupiter C_{18} (3 μm, 300 Å) particles from Phenomenex (Torrance, CA), and fused silica tubing from Polymicro Technologies (Phoenix, AZ). Synthetic phosphopeptides were purchased as Spike Tides from JPT Peptides technologies (Berlin, Germany).

4.4.2 Cell culture and protein extraction

Drosophila melanogaster Schneider S2 cells were cultured in serum-free and protein-free insect medium at 28°C. Cells were plated in 100 mm Petri dishes at a density of 18 million cells/dish. Total cell lysates were obtained using a detergent-free cell lysis. Cells were washed with PBS and cell lysis was performed in an extraction buffer (20 mM Hepes, 1.5 mM $MgCl_2$, 0.42 M NaCl, 0.2 mM EDTA, 25% glycerol) after sonication to extract cytoplasmic and nuclear proteins. After centrifugation at 13 000 g, the supernatant was isolated and the membrane pellet was discarded. During buffer preparation, 1mM DTT, proteases and phosphatases inhibitor cocktails were added at inhibitor/proteins ratio of 1:100. Total protein amount was quantified using Bradford protein assay.

4.4.3 Trypsin digestion

Proteins were reduced in 50 mM ammonium bicarbonate containing 0.5 mM TCEP and 1% SDS for 20 min at 37°C, and then alkylated in 5 mM iodoacetamide for 20 min at 37°C. Excess iodoacetamide was neutralized using 5 mM DTT. After dilution to 0.1% SDS with 50 mM ammonium bicarbonate, proteins were digested overnight with sequencing grade modified trypsin (enzyme/protein ratio 1:50) at 37°C with high agitation. The digest was acidified with TFA, and then evaporated to dryness in a SpeedVac.

4.4.4 TiO_2 phosphopeptides enrichment

Phosphopeptides from *D. melanogaster* S2 cells were enriched on TiO_2 micro-columns as described previously [206]. Briefly, tryptic peptides were resuspended in 250 mM lactic acid (3% TFA/70% ACN) before loading on micro-columns pre-equilibrated with 3% TFA/70% ACN. A total of 750 µg of protein digest was loaded on micro-columns each

containing 6 mg of TiO_2 affinity media. Each micro-column was washed with lactic acid solution followed by 3% TFA/70% ACN to remove non-specific binding peptides. Phosphopeptides were eluted with 1% NH_4OH pH 10 in water and then acidified with TFA before desalting on Oasis HLB cartridges. Samples were evaporated to dryness on a SpeedVac and the equivalent of 250 µg of starting material was resuspended in 20 µL of 0.2% formic acid/5% ACN for LC-MS/MS experiments.

4.4.5 Mass spectrometry

Nano-LC separation was performed using an Eksigent 2D nano-LC (Dublin, CA) system coupled to a LTQ-Orbitrap XL or Velos mass spectrometers (Thermo Fisher Scientific, San Jose, CA). Custom reversed-phase trap (4 mm length, 360 µm i.d.) and analytical column (15 cm length, 150 µm id) were packed with Jupiter C_{18} stationary phase (3 µm particle, 300 Å pore size). Chromatographic separations of enriched phosphopeptides extracts were typically achieved using a flow rate of 600 nL/min and a linear gradient from 5-30% ACN over of 70 or 120 min. Online 2D-LC separations were performed using an Opti-Guard SCX column (1 mm length, 350 µm id, Optimize Technologies) and eluted with 0, 0.25, 0.5, 0.75, 1.0, and 2.0 M ammonium acetate salt fractions, pH 3 (2% ACN, 0.2% FA).

The survey scans were acquired at a resolution of 60 000 (AGC: 10^6, injection time: 500 ms) over the acquisition range of *m/z* 300-2000. MS/MS spectrum with multistage activation were acquired in CID mode using data-dependent acquisition for multiply charged ions exceeding a threshold of 10 000 counts (isolation width: 2 Da, normalized collision energy: 35, activation q: 0.25, activation time: 30 ms, dynamic exclusion: 30 s). MS/MS spectra acquired using ETD mode were obtained using 100 ms activation time with supplemental activation. HCD MS/MS spectra were obtained using 0.1 ms activation time. In each cycle, 12 MS/MS scans on the most abundant peptides were performed in the LTQ (AGC: 5×10^4, maximum injection time: 300 ms) for CID/ETD and 6 MS/MS for HCD.

Targeted analyses were achieved using an inclusion list generated from the algorithms described in section 4.4.7. MS/MS were triggered on selected parent masses with ± 0.02 Da tolerance. To facilitate the identification of MS/MS spectra of phosphoisomers, repeated precursor selection was performed every 30 or 5 seconds for survey and targeted analysis, respectively.

4.4.6 MS/MS data processing for peptide and protein identifications

Mascot Distiller 2.3.2 was used to generate MS/MS peak-list files from raw data. For the detection of phosphoisomers, MS/MS spectra from the same precursor *m/z* were not combined together. All MS/MS spectra were searched with Mascot 2.3 (Matrix Science) against a concatenated target/decoy database of *D. melanogaster* protein sequences built from Uniprot (v11.0, 29 447 proteins) using the following parameters: peptide mass tolerance ± 15 ppm, fragment mass tolerance ± 0.5 Da for CID/ETD and 0.02 Da for HCD, trypsin with up to 4 missed cleavages, and carbamidomethyl (C), deamidation (NQ), oxidation (M), phosphorylation (STY) as variable modifications. Finally, ProteoConnections was used to constrain the peptide false discovery rate (FDR) to less than 1% and to assign the phosphorylation site localization confidence [117].

4.4.7 Algorithms for detecting phosphopeptide positional isomers from LC-MS/MS analysis

Two algorithms were developed to detect potential phosphoisomers from LC-MS elution profiles and from mixed MS/MS spectra. The algorithms were implemented in C# and run under Windows operating system.

The detection of phosphopeptide isomers from LC-MS elution profile comprises four steps (Figure 4.2A): *i*) peptides map generation, *ii*) identified phosphopeptides selection, *iii*) isobaric peptides search and i*v*) output list of candidates to file. Step *i* and *ii* are pre-requisite for step *iii*. Step *i* detect peptide features from raw LC-MS data files. Peptide detection is performed using ProteoProfile, that implements an in-house version of the THRASH algorithm [76]. It creates a peptides map with the following peptide features: *m/z*, retention time (minimum, peak top, maximum) and peak intensities for all elution tracks of each peptide's isotope. Step *ii* consists of selecting phosphopeptides that can have more than one possible phosphorylatable residue. Identified phosphopeptides attributes (*m/z*, retention time, peptide sequence and modifications) are loaded from ProteoConnections output CSV file. The step *iii* uses both inputs to find isobaric peptides. For each identified phosphopeptide, the algorithm searches the peptides map for neighboring phosphoisomers within a retention time window *rtWin* of 4 minutes and an *m/z* tolerance *mztol* of 10 ppm with identical charge state *z*. Next, the algorithm compares the isotopic profiles of candidates to detect phosphoisomers separated by a minimum valley minRtValley of 0.1 min. Multiple replicate LC-MS analyses can be used to report only reproducibly detected phosphoisomer candidates and to reduce false positive identifications arising from neighboring peptides of similar masses but of distinct sequences. Finally in step i*v*, phosphoisomer candidates are reported in an inclusion list (target *m/z*, minimum and maximum retention times). This format is compatible with the acquisition method interface of Thermo Xcalibur software. The inclusion list can be separated into multiple files with appropriate number of precursor ions when the list of candidates exceeds the physical limit of data acquisition of the instrument. A text report is also generated for review and editing before the launch of the targeted analysis.

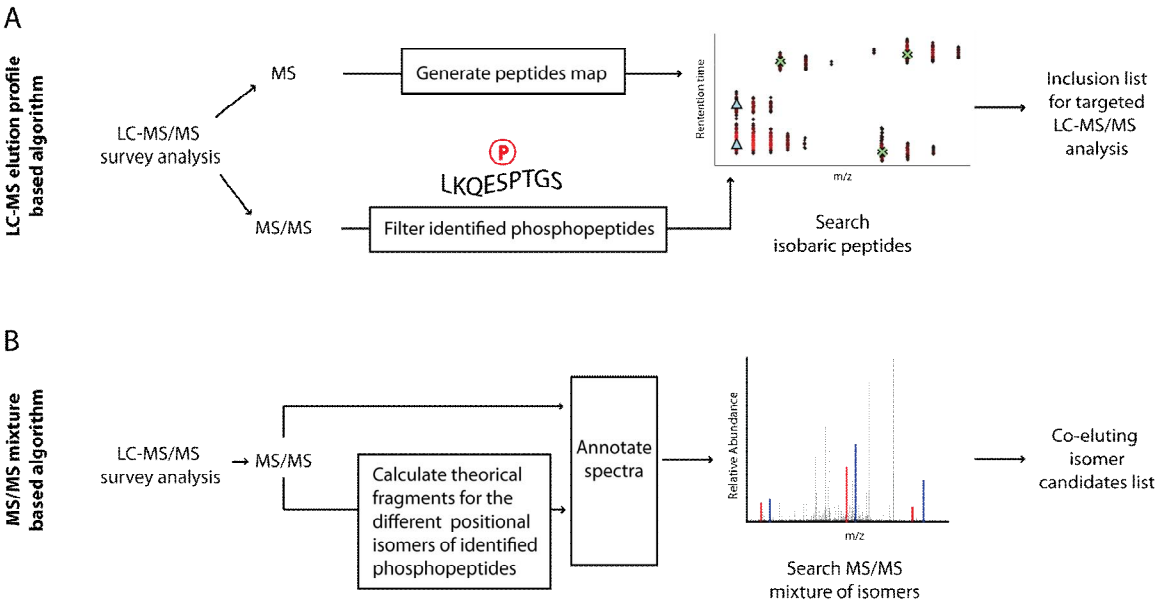


Figure 4.2 : Algorithm workflows for detecting phosphopeptide positional isomers.

Two algorithms were implemented to search for phosphopeptide positional isomers. (A) Search for partially or fully separated isobaric peptides from LC-MS elution profiles. (B) Detection of phosphoisomers from mixed MS/MS spectra.

The algorithm that enables the detection of phosphopeptides isomers from mixed MS/MS spectra searches for site specific fragment ions (Figure 4.2B), and comprises 4 steps: *i*) calculation of phosphopeptide fragments for each positional isomer, *ii*) fragment ion annotation, *iii*) search for site specific fragment ions for each positional isomer, and *iv*) report the potential phosphoisomer candidates. In step *i*, all phosphorylatable residues (serine, threonine and tyrosine amino acids) are identified to generate all possible combination of phosphopeptide sequences. Fragment ion patterns are generated for all sequences using Peptide Fragmentation Modeller software [207] to create a the list of site-specific fragment ions. In step *ii*, each peak p_i from a spectrum s is annotated using the theoretical fragments of each positional isomer based on user's specified fragment mass tolerance Δm. The presence of multiple isomers is detected in step *iii* by searching pairs of fragments suggesting the occurrence of more than one positional isomer. These represent

pairs of unique fragments to each isomer, partially shared fragments (indicate that at least two forms are present but cannot be distinguished), or a shift of 80 Da for a specific fragment (see Table 4.I for example). Finally at step *iv*, a list of co-eluting isomer candidates is generated.

Table 4.I: Types of fragment ions that reveal the presence of phosphopeptide positional isomers.

Fragment m/z	Fragment type	RKpSEEPTAPSGNK	RKSEEPpTAPSGNK	RKSEEPTAPpSGNK
302.36	–	c2	c2	c2
388.32	Partialy shared	z4	z4	
389.23	Shifted		c3	c3
469.28	Shifted	c3		
598.33	Unique	c4		
657.44	Unique	z7		
824.43	Unique	c6		
925.48	Partialy shared	c7	c7	
963.41	Partialy shared		z9	z9
1013.36	Shifted			c9
1092.41	Partialy shared		z10	z10
1093.53	Shifted	c9	c9	
1179.43	–	z11	z11	z11
1180.45	–	c10	c10	c10

Unique: Peak assigned to a fragment specific to one phosphopositional isomer
Shifted: Pair of peaks separated by 80 Da assigned to the same fragment that indicates the presence of two phosphopositional isomers.
Partialy shared: Peak assigned to a fragment that is shared by few phosphopositional isomers. Compared with unique or other partialy shared fragments and if the is no overlap, it can reveal if more than one isoform is truly present.
– : Not informative to distinguish isomers.

The application also contains an automated validation software tool to review the isobaric phosphopeptide candidates of a targeted LC-MS/MS analysis. For each candidate, a text report indicates all acquired MS/MS spectra and the associated peptide identification when successfully assigned by the search engine. Phosphopeptides sharing the same

sequence but differing in the site of the phosphorylated residues are highlighted as phosphoisomers using a simple binary classifier.

4.4.8 Conformation prediction of phosphopeptide positional isomers

PEP-FOLD, a *de novo* 3D peptide conformation modeling algorithm was used to predict peptide conformation [208]. This software uses a greedy-OPEP forcefield algorithm and a hidden Markov model with a structural alphabet. The lowest energy conformations were used in this study. The solvent accessible surfaces of the whole peptide and the charged residues were computed using the *get_area* function in PyMOL (version 0.99rc6).

4.4.9 Datasets availability

All datasets generated or used in this study are available online via ProteoConnections at http://www.thibault.iric.ca/proteoconnections. For the rat phosphoproteomics dataset, description of the experimental and analytical procedures are available in Courcelles & al [117].

4.5 Results

4.5.1 Phosphopeptide positional isomers occurrence in large-scale phosphoproteomics studies

The presence of phosphopeptide positional isomer has been reported only in few papers for specific proteins [110]. However, there is limited information available on their occurrence in complex phosphopeptide samples and how to detect phosphopeptides isomers in a comprehensive fashion from LC-MS datasets. In an attempt to answer these questions, we searched two previously acquired datasets for the presence of these phosphoisomers. The first phosphoproteomics dataset was acquired from mouse macrophage cell line J774 (unpublished) and the second, from rat intestinal epithelial cell line IEC-6 datasets [117]. In both cases, phosphopeptides were enriched using TiO_2 media and then analyzed by online 2D-LC-MS/MS on the LTQ-Orbitrap XL. We identified 4798 phosphopeptides (1257 proteins) and 15 700 phosphopeptides (2108 proteins) in the J774 and IEC-6 samples, respectively.

To detect chromatographically separated phosphopeptide isomers, we developed a new algorithm that used peptides maps and MS/MS spectra (Figure 4.2A). The algorithm uses phosphopeptide identifications from MS/MS searches, and lists candidate isomers that are separated by at least 0.1 min. Only peptides with a Mascot score higher than 20 and site localization confidence greater than 0.75 were retained. For J774, the algorithm detected 598 phosphopeptides with two chromatographically separated peaks. The analysis of MS/MS spectra indicated that 53 phosphopeptides had the same sequence but differed in the localization of the phosphorylated residues (Table 4.II). Further examination of MS/MS spectra and peptides maps revealed that 19 phosphopeptides were true isomers, while 34

had ambiguous identification due to insufficient time separation and/or uncertain site localization. In the IEC-6 dataset, we detected 1298 phosphopeptide isomer candidates where 48 out of the 145 acquired MS/MS supported their assignment as true positional isomers. The survey of these large-scale phosphoproteome datasets indicated that distinguishable phosphopeptide isomers represent approximately 0.4% of all identified phosphopeptides. However, this relative proportion could be underestimated in view of the number of ambiguous cases, the bias in MS/MS sampling toward most abundant peptide ions and the potential of co-eluting species.

Table 4.II : Relative distribution of phosphopeptide positional isomers from large-scale phosphoproteomics studies.

Sample	Peptides	Phosphopeptides			
		All	Separated in LC elution profile	Separated in LC but unclear in MS/MS ids	Separated in LC and MS/MS ids
J774	5372	4798	598	34	19
IEC-6	23902	15700	1298	97	48

Chromatographically separated phosphopeptide isomers from both datasets were combined into a unique list of 64 isomers to study their intrinsic physicochemical properties (Supplementary file Table A4.II). The length of separated phosphoisomers ranged from 7 to 35 amino acids (average of 17), and the distance between phosphorylated sites spanned from 1 to 23 (average of 15) (Figure 4.3A-B). These results are consistent with those reported in the literature where phosphopeptide length varied between 13 and 26 amino acids, while phosphorylated sites were separated by 1 to 16 amino acids [110]. The measured shifts in retention time caused by the alternate modification site varied from 0.5 min to 6 min (outlier at 22.3 min) with an average of 2.3 min (Figure 4.3C) and LC peak resolution ranged from 1 to 12.8 with an average of 4.6 (Figure 4.3D). To rationalize the differences in retention times between isomeric phosphopeptides, we calculated the

corresponding changes in local hydrophobicity and electrostatic interactions. The difference in local hydrophobicity scale was calculated using a windows of 3 amino acids centered on the phosphosite and the amino acids scale values from Kyte J., Doolittle R.F. [209] (Figure A4.2A). Electrostatic interaction effects were considered by calculating the difference in distance between the phosphosite and the charged residues (Figure A4.2B-C). Change in these physicochemical properties should affect the interaction with the reverse phase column. However, we could not obtain any meaningful correlation between changes in these properties and variation of retention times, suggesting that subtle conformational changes could not be properly predicted based on primary peptide sequence alone.

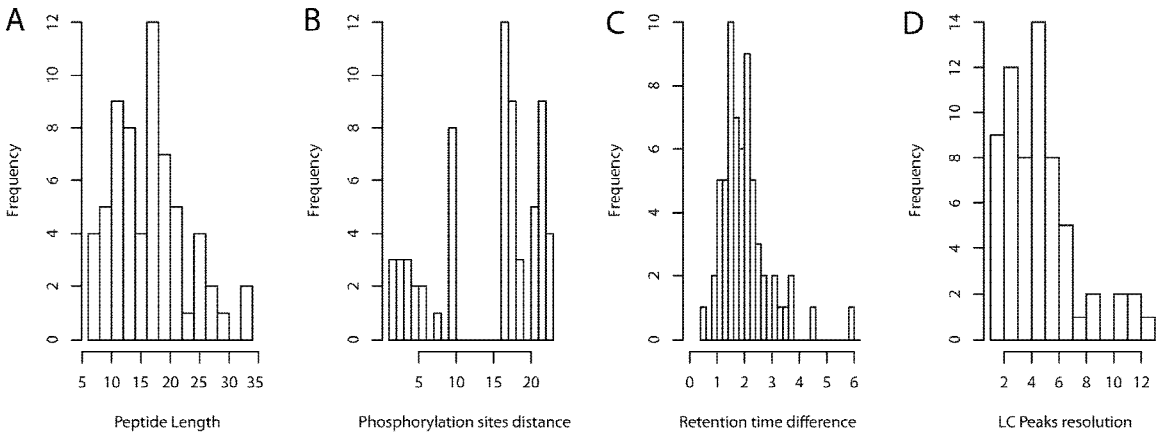


Figure 4.3 : Properties of phosphopeptide isomers separated by RP-HPLC.

Properties of 64 phosphopeptide isomers separated by RP-HPLC identified in IEC-6 and J774 cells using the LC-MS elution profile detection algorithm and MS/MS identifications. (A) Histograms for the distribution of the isomers peptide length. (B) Distance in amino acids between phosphorylated site positions. (C) Retention time difference between the peak top of phosphopeptide isomers (an outlier with 22.3 min of separation is not displayed for clarity). (D) Chromatographic resolution of isomeric peaks.

We next investigated conformational changes of positional phosphopeptide isomers using PEP-FOLD, a *de novo* 3D peptide conformation modeling algorithm [208]. Since this software cannot directly predict the 3D structure of phosphopeptides, we used glutamic acid as a phosphomimetic residue. As an example, we investigated the conformational changes imparted by the localization of the phosphorylated residues in the following two phosphoisomers VGGpSSVDLHR and VGGSpSVDLHR. Although these phosphopeptides only differ by the phosphorylation of Ser4 and Ser5 residues, they were chromatographically separated by 1.2 min (Figure 4.4A). In both cases, the phosphorylated site was assigned with a confidence level of 1.0 (Figure 4.4B and C). The conformation of each isomer was predicted using PEP-FOLD, and suggested strickingly different 3D structures (Figure A4.3). For example, the change in phosphorylated residues Ser4 to Ser5 resulted in a transition from a random coil (Figure A4.3B) to an helix (Figure A4.3D) conformation. The formation of an electrostatic interaction and N-H···O=C hydrogen bond between the N-terminal residue and the phosphorylated serine 5 stabilize the helix secondary structure. The two conformations have an approximate difference of 33 $Å^2$ in solvent accessible surface. The Ser4 isomer which has the largest solvent accessible surface (1133 $Å^2$) is more retained on the reverse phase column. In addition, the charged residues surface of the Ser4 isomer is 5 $Å^2$ lower than its Ser5 isomer counterpart. Increase in hydrophobic surface calculated between these two isomers could explain the difference observed in their retention times.

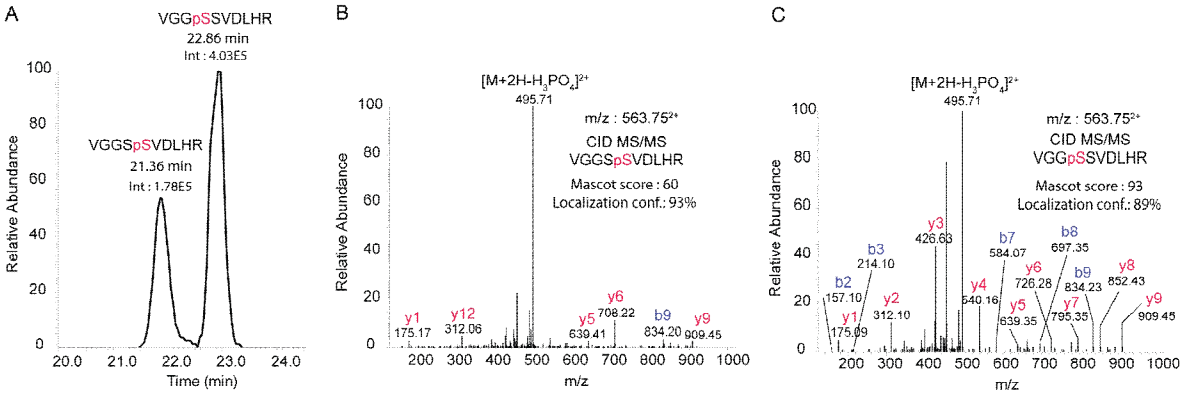


Figure 4.4 : RP-HPLC separation of two isomeric phosphopeptides.

A) Extracted ion chromatogram for the doubly-protonated phosphopeptides VGGSpSVDLHR and VGGpSSVDLHR at *m/z* 563.75, from a TiO_2-enriched digest of IEC-6 cell extract showing a 1.2 min separation between the two phosphoisomers. MS/MS spectra of peaks eluting at 21.4 (B) and 22.9 min (C) shown in (A).

4.5.2 Analysis of synthetic phosphopeptide positional isomers

We next evaluated the ability of our algorithm to detect chromatographically separated isomers using a set of nine synthetic phosphopeptides. These phosphopeptides were selected from candidates previously identified in a large-scale phosphoproteomics study of *D. melanogaster* S2 cells, and a list of these peptides is presented in Figure 4.5. We obtained monophosphorylated peptides for all possible phosphorylatable residues of each peptide sequence. Three peptides can be phosphorylated at two positions, four at three, and two at four positions. Phosphopeptides were mixed together in order to analyze each pair of positional isomers from each peptide sequence. We performed LC-MS/MS experiments on the LTQ-Orbitrap XL using an inclusion list to trigger the acquisition of CID and ETD MS/MS spectra of target synthetic phosphopeptides every 15 seconds.

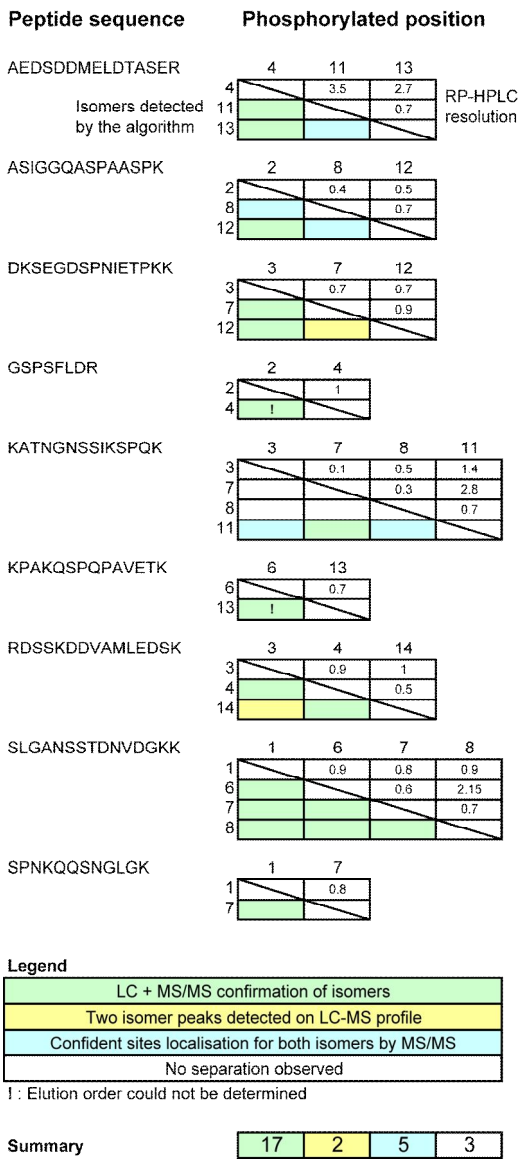


Figure 4.5 : Detection of synthetic phosphopeptide isomers with the algorithm based on LC-MS elution profile.

Positional isomers of 9 synthetic phosphopeptides analyzed in pairs by LC-MS/MS on the LTQ-Orbitrap XL using CID and ETD fragmentation activation modes. The separation resolution of isomers is reported for phosphopeptides detected above an intensity value of 10 000 counts and the efficiency of the algorithm to detect phosphopeptide isomers is indicated by the color code.

Most isomers were separated with different degree of overlap and only four were completely separated with a resolution greater than 1.5 (Figure 4.5 and Supplemental Table A4.I). We then applied our algorithm to generate the inclusion lists of potential isomer based on the LC-MS elution profiles. Our algorithm could detect meaningful separation for most of the phosphopeptide isomers, except for 8 specific cases where poor resolution was obtained. The algorithm also found more phosphopeptide peaks than expected, and close examination of the corresponding MS/MS spectra revealed almost superimposable fragmentation patterns, suggesting the presence of separated phosphopeptide conformers as previously reported for synthetic phosphopeptides [210]. These conformers are wrongly assigned as potential isomer candidates and cannot be distinguished from their LC-MS elution profile alone. The presence of phosphopeptide conformers could lead to ambiguous isomer assignment if their occurrence is frequently observed in digests of cell extracts. This topic is further discussed in section 4.5.3.

We then applied our validation algorithm to confirm phosphopeptide isomers from CID and ETD MS/MS spectra. For five isomers pairs, the phosphorylation localization could not be confidently determined irrespective of the activation mode used, while CID and ETD MS/MS spectra provided complementary identification for 20 isomers. Five phosphopeptide isomers were identified using both fragmentation methods. Isomers that are partially chromatographically separated can be distinguished by MS/MS if the acquisition has been triggered in a region with limited overlap. Finally, 18 out of 27 isomeric pairs were confirmed with distinct retention times and MS/MS spectra. The elution order of two pairs could not be determined by the MS/MS identifications obtained. These experiments indicated that the algorithm can detect isomers showing partial separation, although some assignments could be misrepresented if phosphorylated sites are ambiguously localized.

4.5.3 Targeted analysis of phosphopeptide positional isomers

We evaluated the analytical potential of our algorithm to detect phosphopeptide isomers using TiO_2-enriched extracts of *Drosophila melanogaster* S2 cells. All experiments were performed using 2D-LC-MS/MS on the LTQ-Orbitrap Velos. To identify a large number of phosphopeptides, triplicate analyses were obtained using CID, ETD and HCD activation modes with 30 s dynamic exclusion time. We identified a total of 10 110 phosphopeptides (unique *m/z*) of which 1304 phosphopeptide ions were detected with 2 potential isomers separated by at least 0.1 min (Table 4.III and Supplementary Table A4.III). Comparison of acquired MS/MS spectra confirmed 77 pairs of phosphopeptide positional isomers and 199 ambiguous pairs with uncertain localization. The proportion of true isomers is consistent with the observation made in the large-scale phosphoproteomics analyses of J774 and IEC-6 cell extracts.

Table 4.III : Targeted analysis of RP-HPLC separated phosphopeptide positional isomers from the fly.

Survey analysis		Targeted analysis		True isomers			
Ambiguous	True isomers	Ambiguous	True isomers	Common	Survey	Targeted	Total
199	77	216	86	46	31	40	117

The targeted analysis was conducted in two 2D-LC-MS/MS analyses to evenly distribute the high number isomer candidates. The list of isomer candidates were thus split in two parts for each SCX fraction and added to distinct instrument acquisition methods. Each precursor ion was fragmented sequentially by CID, ETD, HCD every 5 seconds to obtain good quality MS/MS spectra and confirm the site localization of both isomers with high confidence. We identified a total of 86 true isomers together with 216 ambiguous phosphopeptide isomers. A subset of 46 true isomers was also identified in the survey analysis while 40 were unique to the targeted analysis. Altogether, these analyses identified

a total of 117 phosphopeptide positional isomers corresponding to 1.2% of all identified phosphopeptides. The targeted analysis yielded complementary results to the survey analysis with approximately 50% overlap of identification (Figure A4.4). A comparison of phosphoisomer physicochemical properties of S2 cells (Figure A4.5) with those identified in J774 and IEC-6 cells (Figure 4.3) revealed similar trends, except for the occurrence of the relatively long peptides (> 35 a.a.) observed in the S2 extracts.

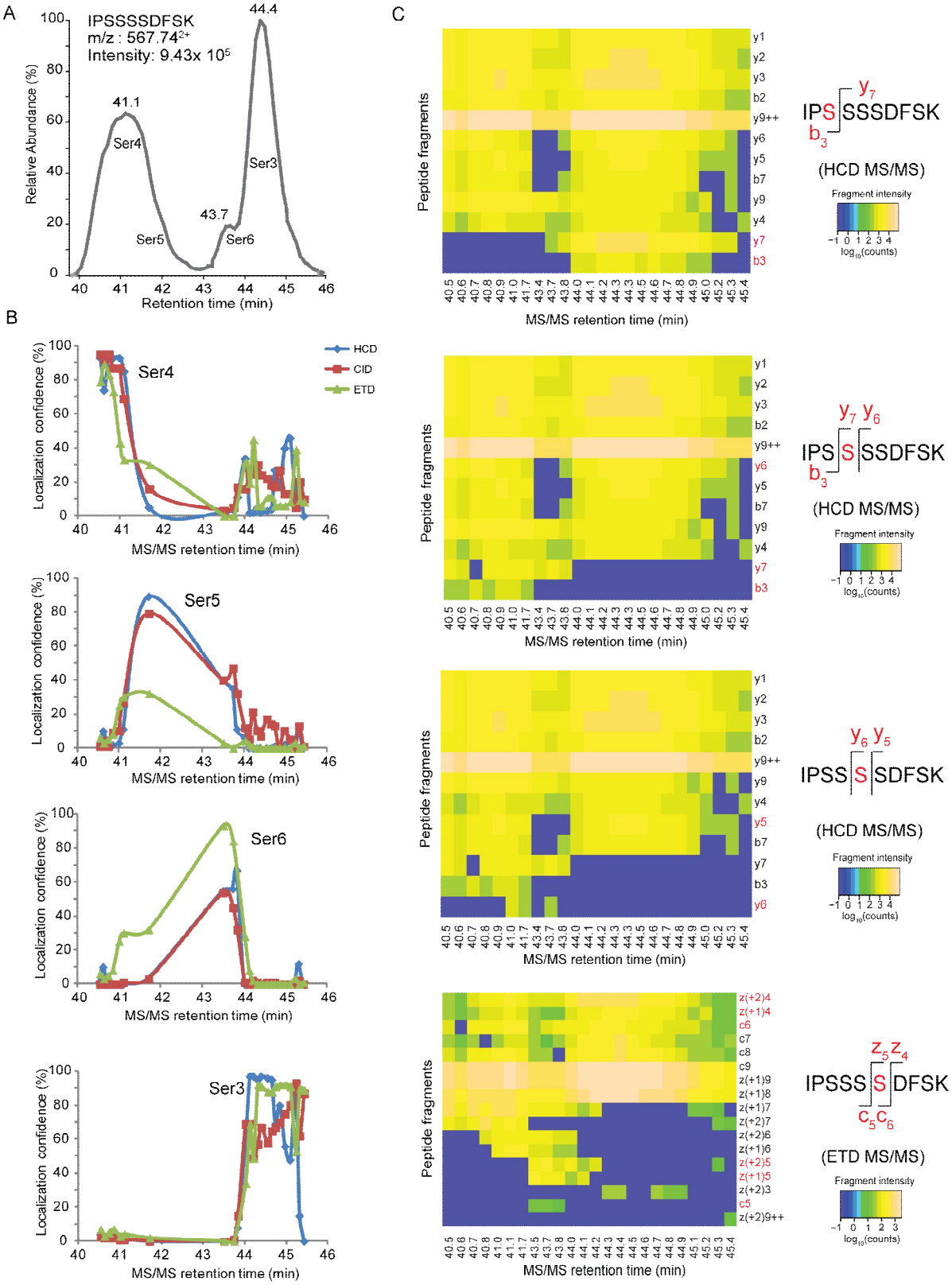
Artifacts that can be confounded with isomers were also observed (Figure A4.1). Those artifacts represent 7% of included peptide ions. There were 8 cases with an alternative tryptic cleavage, 56 with an alternative peptide sequence, 2 with position shift of an alternative modification and 20 with an alternative conformer. Alternative tryptic cleavage arises in presence of a missed cleavage and when two arginines or lysines are found at both peptide extremities (e.g. XR_KPEPTIDER_KX ↔ XRK_PEPTIDERK_X, where trypsin cleavages occur at _). Phosphopeptide conformers are present at a low frequency in protein extracts and not only in synthetic phosphopeptides. The separation of the phosphopeptide GIMEEIEMRpSPLSDR (625.6^{3+}) conformers is shown as an example (Figure A4.6). Both conformers are well resolved by RP-HPLC and have a common HCD fragmentation pattern that localizes the phosphorylated site to the same residue. The second eluting form has an extra fragment ($y9^{++}$) that is highly abundant and that might be favorized by the specific conformation.

In addition, we identified 7 phosphopeptides with three positional isomers and one phosphopeptide with four isomers (Figure 4.6). The sequence of the latter peptide IPSSSSDFSK from a protein of unknown function contains 5 phosphorylatable residues, 4 of which were identified in our analysis. Interestingly, the localization of the phosphorylation site had a significant influence on its hydrophobicity as shown for phosphorylated Ser3 and Ser4 residues that are separated by 3.3 min (Figure 4.6A). For the two other phosphorylated isomers, Ser5 and Ser6, both of them co-eluted with the two predominant forms modified at Ser3 and Ser4 residues. Only Ser6 isomer could be distinguished by a small bump in the elution profile. We next compared the confidence in

the site localization for all MS/MS spectra acquired over this time period (Figure 4.6B and Figure A4.7). While phosphorylation at Ser3 and Ser4 residues was localized confidently using all three fragmentation methods, the localization of Ser5 was confirmed only by HCD/CID and S6 by ETD. This result indicates that multiple fragmentation methods can be required to identify all positional isomers of a peptide. This peptide can also be phosphorylated on S9 but none of fragmentation methods supported this identification (data not shown). Repeated MS/MS acquisitions were found to be useful to confirm the presence of each phosphorylated isomer. For each phosphorylated form, a consistent fragmentation pattern across multiple MS/MS scans can be observed for fragment ions with site-specific modification (Figure 4.6C). The use of repeated MS/MS acquisition enables the profiling of fragment ion patterns during the peptide elution and facilitates the unambiguous localization of phosphorylation sites for isomeric phosphopeptides.

Figure 4.6: Identification of four phosphopeptide positional isomers of IPSSSSDFSK

A) LC-MS elution profile of the four monophosphorylated positional isomers of IPSSSSDFSK identified by targeted LC-MS/MS analysis of TiO_2-enriched phosphopeptides from *D. melanogaster* S2 cells. B) Localization confidence (%) for all MS/MS acquired in HCD, CID and ETD fragmentation modes. C) Heatmap representing each phosphorylated position according to the intensity of fragment ions. Fragments enabling site-specific localization are highlighted in red.


A
IPSSSSDFSK
m/z : 567.74²⁺
Intensity: 9.43x 10⁵
41.1
Ser4
43.7
Ser5
Ser6
44.4
Ser3
Relative Abundance (%)
Retention time (min)
B
Ser4
Ser5
Ser6
Ser3
HCD
CID
ETD
Localization confidence (%)
MS/MS retention time (min)
C
Peptide fragments
MS/MS retention time (min)
y7 y6
IPS S SSDFSK
b3
y6 y5
IPSS S SDFSK
z5 z4
IPSSS S DFSK
c5 c6
(HCD MS/MS)
(ETD MS/MS)
Fragment intensity
log10(counts)

4.5.4 Detection of co-eluting phosphopeptide positional isomers

The analysis of chromatographically separated phosphopeptide isomers revealed that the difference in retention times between isomeric pairs extends from 0.5-7.0 min and more than 56% were observed within 2 min of each other. The close proximity of retention times between isomeric pairs raises the possibility that a sizable proportion of phosphoisomers could co-elute together. To evaluate this possibility, we extended the functionality of the present algorithm to detect isomers from fragment ion features of the MS/MS spectra. Enriched phosphopeptides from *Drosophila melanogaster* S2 cell line were analyzed by LC-MS/MS on the LTQ-Orbitrap Velos using CID, ETD and HCD activation modes. A list of potential phosphoisomers was generated from a preliminary LC-MS/MS analysis obtained under data-dependent acquisition (DDA). A dynamic exclusion of 10s was used to acquire multiple MS/MS spectra across the peak elution and identify co-eluting isomers partially separated by LC. A total of 2901 unique phosphopeptides were identified, of which 1544, 1251 and 861 phosphopeptides were assigned from their corresponding CID, ETD and HCD MS/MS spectra.

The algorithm annotates each MS/MS spectrum based on theoretical fragment ions calculated for each possible isomer. A search is then initiated to identify fragment ion pairs that distinguish two or more phosphopeptide isomers (Table 4.I). The distributions of different types of fragment ions observed between phosphopeptide isomers is presented in Figure 4.7, and indicates that the number of observed fragment ions enabling isomer detection depends on the fragmentation method. For an arbitrary cutoff of at least two fragment pair, the proportion of candidate isomers co-eluting was 65% (1011/1544), 9.8% (123/1251), 3.6% (31/861) for CID, ETD and HCD, respectively. MS/MS spectra acquired with HCD generate a lower proportion of isomer candidates than other fragmentation modes. The false positive rate is certainly lower due to enhanced mass accuracy. Visual inspection of elution profile of ETD and HCD isomer candidates confirmed that 23 pairs of phosphopeptide isomers were partly separated. This result suggests that co-eluting

phosphopeptide positional isomers represent at least 0.8% of all identified phosphopeptides.

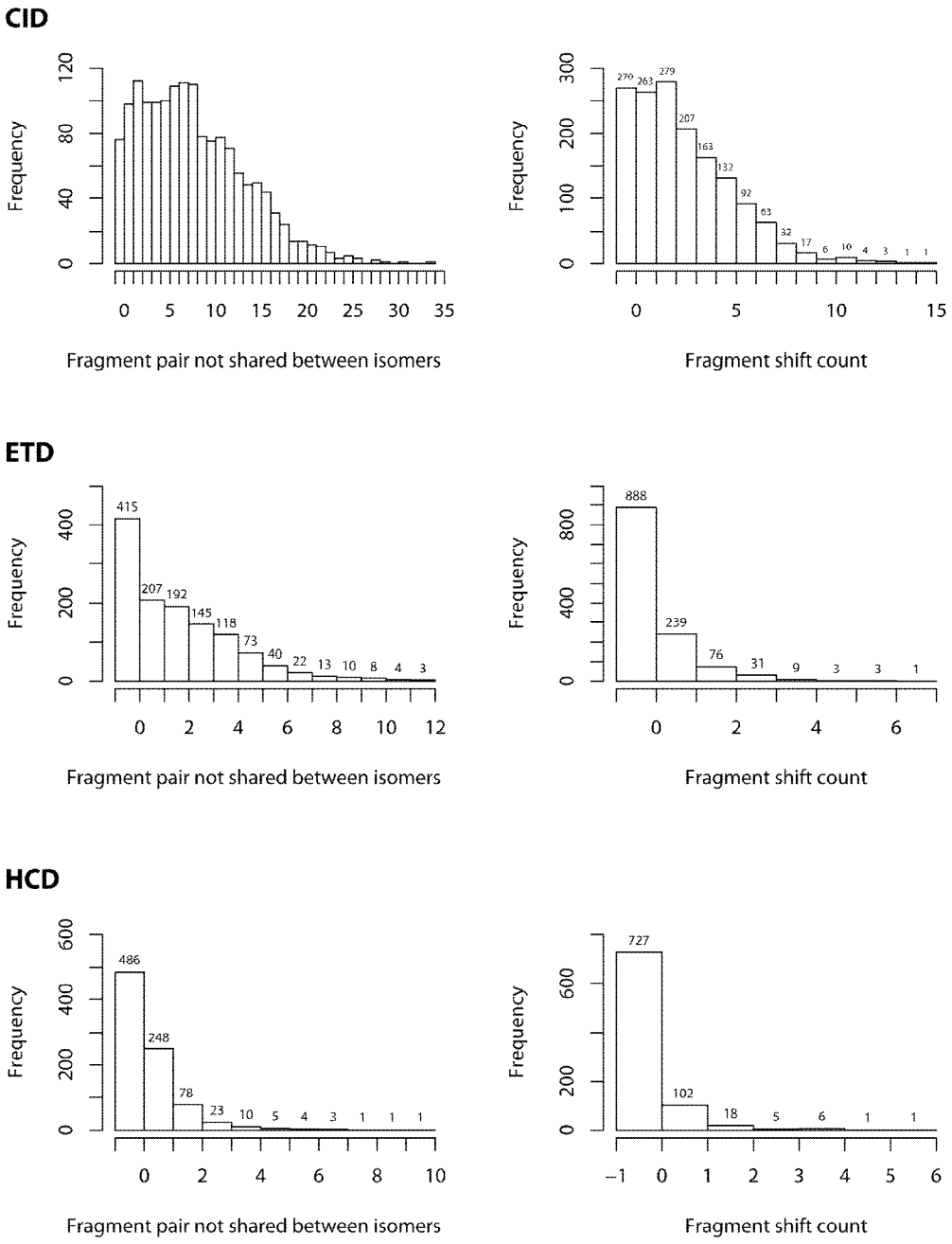


Figure 4.7 : Detection of co-eluting phosphopeptide isomers using distinctive fragment ion features for MS/MS spectra acquired with different activation modes.

The MS/MS based algorithm was tested using three different fragmentation methods (CID, ETD, and HCD) to detect co-eluting phosphopeptide positional isomers. Histograms show specific characteristics used to detect isomer from MS/MS spectra for each fragmentation method.

4.6 Discussion

The objective of this work was to study the occurrence and separation of phosphopeptide positional isomers. For this purpose two algorithms were created to detect the isomers from the LC-MS elution profile and MS/MS spectrum. Using the first algorithm, we found out that about 1% of identified phosphopeptide are positional isomers separated by RP-HPLC. This proportion was similar for mouse, rat and fly species. The estimated proportion of co-eluting isomers in fly was about the same as separated isomers. Co-eluting isomers need to be confirmed by further experimentations and the appropriate method required to separate them must be determined.

Detecting phosphopeptide positional isomers in a complex phosphopeptides sample was harder than expected. First, an important issue for the discovery of phosphopeptide positional isomers is the ability to localize confidently the phosphorylated sites. In our targeted analysis, 25% of the targeted isomers were ambiguous in site localization. The poor fragmentation of phosphopeptides prevents the site localization to a specific residue. Even the use of different fragmentation methods (CID, ETD, and HCD) was not sufficient to localize the phosphorylation site for some phosphopeptides and therefore impede the confirmation of the isomers presence. Additionally, it was found, from an inclusion list experiment with repeated MS/MS acquisitions, that the timing for triggering MS/MS acquisition is important to obtain a good fragmentation that allows site localization and not only the peptide identification.

The second issue in detecting phosphopeptide positional isomers in a complex phosphopeptide sample is the possibility that different peptide artifacts are mistaken for isomers (Figure A4.1). Artifacts are appearing as two separated isobaric peaks and are considered as isomer candidates by the algorithm. Those artifacts corresponded to 7% of all included phosphopeptide ions and four types of artifacts were observed. The most frequent one is peptides with equal masses but different amino acid sequences that have a small difference in retention time. Another recurrent class of artifacts is phosphopeptide

conformers that are separated by reverse-phase liquid chromatography. For those, two peaks on the LC chromatogram are clearly separated and MS/MS spectra of each peak share a common fragmentation pattern and their interpretation yields the same peptide sequence and site localization. Enzymatic digestion with trypsin can also create isobaric species with alternative cleavage position. While this artifact class can be anticipated by looking in the protein sequence, only 15% of the potential cases had this alternate missed cleavage observed. Finally, other modifications (e.g. carbamidomethylation, oxidation) present on a phosphopeptide were also found with alternative position.

The two algorithms created to detect phosphopeptide positional isomers have several limitations. Those can either affect detection of isomers or report false positive hits. The algorithm based on LC-MS profile should preferably be used with high resolution MS data to avoid artifact hits from peptides with close masses. Isomers search is limited to phosphopeptides identified in the LC-MS/MS survey analysis to avoid wasting acquisition time on non phosphorylated peptides or phosphopeptides with no alternative phosphorylable position. Poor sampling of peptide ions in the survey step can thus limit the candidate list. Additionally, only phosphopeptides with clear isotopic and elution profiles above the user-defined intensity threshold are considered. Only phosphopeptide isomers separated by at least 0.1 min are considered as candidate to avoid too many false positive hits caused by missing MS signal data points.

The successful use of the inclusion list generated by this algorithm depends on the acquisition strategy. The targeted acquisition method used here has several pitfalls and not all targeted ions were acquired. The inclusion rate, proportion of peptide ions with MS/MS triggered, was 81% (1058/1304) and the identification rate after database search was 78% (1014/1304). The amount of MS/MS acquired for each phosphopeptide is unequal because the inclusion strategy favors the most abundant peptide first. Some phosphopeptides got many MS/MS since repeats were allowed every 5 seconds and others were not acquired on both peaks in crowded elution windows. This dynamic exclusion period is therefore too short. Different dynamic exclusion times were selected based on trials and errors across the

experiments presented in this paper. However, it was not systematically optimized to select the optimal value. More optimization would be necessary to determine the maximum overlapping inclusion windows tolerated by the instrument. The optimization should consider the number of repeated MS/MS acquisition across the elution peak in order to get a confident identification of the phosphopeptide and site localization. This optimization would help to determine how much time and acquisition run are required to successfully analyze all targets.

The included list of phosphopeptide isomer candidates generated by the algorithm has an important size that represents 13% of all identified phosphopeptides and only 9% of them were true isomers. The discovery rate of isomers with this method is thus low. After considering true isomers, ambiguous cases, artifacts and non acquired phosphopeptides, the remaining 40% of the phosphopeptides on the list show no sign to be isomers. Those might be false positive cases either caused by the peptide detection algorithm or unstable elution profiles. To reduce the number of false positive candidates, the algorithm was used post-acquisition on acquired data and set to keep reproducibly separated phosphopeptides in the three replicates. The inclusion list size dropped from 1304 (13%) to 418 (4%), about three times smaller, but the number of true isomers also decreased by 28% (117 to 84) which is 1.4 times smaller. As this filter has a cost on discovery of isomers, it could be useful to prioritize a subset of candidates for inclusion.

Detection of isomers with the MS/MS based algorithm has also its set of challenges. The low abundance or low fragmentation efficiency of some phosphopeptides leads to low intensity fragments that can be confused with noise peaks. Setting a low peak intensity threshold will cause false positive isomer hits while a high one will miss true isomers. To increase the specificity of match made by our algorithm, the minimum number of fragments supporting the presence of two isomers can be adjusted to decrease the probability of hazardous match. Resolution of acquired MS/MS spectra is also important to limit false positives. For MS/MS acquired in the LTQ analyzer, the low mass accuracy of observed fragments can lead to false peak annotation. An additional problem with CID fragmentation

in the LTQ is the possible low level of gas phase phosphate rearrangement [211]. This phenomenon generates unnatural isomers that will be picked-up by this algorithm. For a proper use of this algorithm, high resolution MS/MS with HCD fragmentation is recommended to avoid both problem enumerated here. Two more points should also be considered for detection of co-eluting isomers. In cases where co-eluting isomers are slightly separated in time, the MS/MS acquisition must be triggered when there is elution overlap. Repeated MS/MS acquisition over the elution time interval is therefore necessary to generate an MS/MS spectrum at the right time. Stoichiometry of co-eluting isomers should also not exceed the dynamic range of the mass analyzer as this will impede the detection of one isomer.

To understand the effect of the shift in the phosphorylated position on the retention time of isomers on a reverse phase column, several properties were calculated from the primary peptide sequence. While differences were observed in local hydrophobicity scale and proximity to charged residues, those values could not be correlated with the amplitude of separation and thus were not useful for predicting drift in retention time between two isomers. A quick investigation of tridimensional conformation of isomers was done in this study using the PEP-FOLD software. From the predicted conformation of the two isomers, the difference in solvent accessible surface for the whole phosphopeptide and the charged regions were calculated. These two values supported the elution order of two isomers of VGGSSVDLHR phosphorylated at Ser4 or Ser5. For that reason, tridimensional structures looked promising for the prediction of retention time difference of phosphopeptide positional isomers on a reverse phase column. PEP-FOLD predictions were extended to all isomers detected in S2 cells. From those, we noted for 90% of the cases that the phosphate is interacting with a different functional group on the peptide between isomers. Qualitatively, this affected the secondary structure of about 38% of them. However poor correlation was observed between the retention time shifts observed based on the surface area. Besides solvent accessible surface, additional structural factors such as the shape or the distribution of charged patch could influence the interaction with the reverse phase

column. The results obtained from this approach could also be biased by the accuracy of the structure prediction. PEP-FOLD algorithm was designed for prediction in aqueous solution, and therefore, might not reflect the conformation dynamic in the course of a separation with a gradient of aqueous/organic solvent. The phosphorylation was also simulated by a glutamic acid in those predictions. This might not perfectly reproduces the space and charge of the phosphate group. Supplementary work is therefore required to understand the separation of those isomers.

In this work, algorithms were developed specifically to detect phosphopeptide positional isomers in order to study them. Other than this specific purpose, these algorithms can be useful in different research scenarios. In case of large scale phosphoproteomics, those algorithms can provide additional sites that would be missed by standard analytical procedure. The gain in number of sites will however be low considering isomers occurrence frequency. The biological importance of those sites should be considered versus the time and efforts required for the targeted analysis to identify the isomers. A second scenario where those algorithms can be helpful is in situations where positional isomers are problematic for phosphopeptides quantification. Co-eluting forms are quantified by the software as a single peptide and not the combination of the abundance of two peptides. Isomers separated by LC can also be wrongly quantified in label-free quantification. If isomers are poorly resolved and retention time shifted slightly between runs, the algorithm that searches for the peptide with the minimal *m/z* and retention time difference can select the alternate isomer and again reports the wrong abundance value. For both quantification problems, our two algorithms can be used to highlight problematic phosphopeptides to be manually reviewed by the MS expert. Finally, the two algorithms can also be used beyond phosphoproteomics and be employed to study the occurrence and separation of different post-translational modification isomers.

4.7 Conclusion

This investigation presents the first detailed study on phosphopeptide isomers and describes two algorithms to facilitate their detection in complex cell extracts. This led to two unexpected findings about the occurrence and types of phosphopeptide isomers present in large-scale phosphoproteomics studies. First, the proportion of phosphopeptide isomers is relatively small and represents typically 1% of all phosphopeptides identified from TiO_2-enriched digests of mouse, rat or fly protein extracts. These phosphorylated sites closely located could act as cooperative, independent or mutually exclusive signaling events to modulate proteins activity. Second, the location of the phosphorylated residue within the same peptide sequence can significantly affect its physicochemical properties leading to conformational changes and variable retention times on reverse phase columns. The distribution of shift in retention times between more than 110 isomeric pairs indicated a variation of 0.5-7.0 min, where more than 56% were observed within 2 min of each other. Closely eluting phosphopeptide isomers are often difficult to detect in complex biological extracts due to ambiguity in the site localization and the stochastic nature of data dependant MS/MS acquisition. Targeted LC-MS/MS using inclusion list yielded complementary information on phosphopeptide isomers, although site localization could not be obtained for 25% of targeted precursor ions irrespective of the activation mode used (e.g. HCD, ETD or CID). The use of inclusion list with repeated MS/MS scans facilitated the resolution of phosphopeptide isomers that are either missed or incorrectly assigned due to non optimal triggering of MS/MS spectra when using data dependant acquisition. This approach can also unravel confounding non phosphorylated peptides with closely related mass, and phosphopeptides with alternate conformations or cleavage sites that typically represented 7% of suspected phosphopeptide isomers. Importantly, the data mining approach developed here for the identification of phosphopeptides isomers extend beyond a simple cataloguing exercise, and can be of practical application for the confirmation of phosphorylation sites for subsequent mutagenesis studies, and for the identification of other types of peptide isomers or conjugates.

4.8 Acknowledgments

We thank Éric Bonneil for the technical assistance and Dev Sriranganadane for comments on this manuscript. MC acknowledges scholarships from the Fonds de recherche sur la nature et les technologies du Québec (FQRNT) and the Faculté des études supérieures et postdoctorales (FESP). IRIC is supported in part by the Canadian Center of Excellence in Commercialization and Research, the Canada Foundation for Innovation, and the FRSQ. This work was carried out with the financial support of operating grants from the National Science and Engineering Research Council (NSERC) and from the Canada Research Chair program.

Conclusion

La phosphorylation est une modification post-traductionnelle omniprésente qui régule divers processus cellulaires. Cette modification est contrôlée spécifiquement par l'activité enzymatique des familles des protéines kinases et phosphatases. Les kinases Erk1/2, de la famille des « mitogen-activated protein kinases », sont au cœur d'une voie de signalisation importante qui module la traduction des protéines, la progression du cycle cellulaire, le réarrangement du cytosquelette, l'activation de la transcription de certains gènes, etc. Au niveau des processus biologiques, elles sont impliquées dans le développement de l'organisme, le métabolisme du glucose et la réponse immunitaire. Différentes pathologies humaines comme le diabète, les maladies cardiovasculaires et principalement le cancer, ont fréquemment été associées au dérèglement d'expression des protéines et à la présence de mutations sur les membres de la voie Erk1/2. Compte tenu de l'importance biologique et clinique de ces deux kinases, il est pertinent de connaître l'étendue de leur activité enzymatique afin de développer de nouvelles thérapies pharmacologiques.

Dans ce contexte, l'objectif principal de cette thèse a été de mesurer l'influence de cette voie sur l'ensemble du phosphoprotéome pour découvrir de nouveaux substrats des kinases Erk1/2. Afin d'atteindre cet objectif, nous avons entrepris une étude phosphoprotéomique de l'inhibition pharmacologique de la voie de signalisation Erk1/2. Le succès de cette étude est basé sur trois technologies clés, soit l'enrichissement des phosphopeptides avec le dioxyde de titane, la spectrométrie de masse à haut débit et haute résolution, et le développement d'une plateforme bio-informatique pour l'analyse des données.

ProteoConnections, présenté au Chapitre 2, est une nouvelle plateforme d'analyse bio-informatique dédiée à l'organisation des données de protéomique, à l'évaluation de leur qualité, à indiquer les changements d'abondance et à accélérer l'interprétation des données. Une base de données relationnelle est employée pour organiser logiquement les milliers de peptides et protéines identifiés. La création du schéma de la base de données a été axée sur la présence de modifications post-traductionnelles des peptides (Figure A2.1). La structure

du schéma permet donc de faire de simples requêtes SQL pour obtenir la liste des sites de phosphorylation et le niveau de confiance correspondant à la localisation de ces sites. Pour accélérer l'interprétation, la liste de sites phosphorylés obtenue est annotée par ProteoConnections grâce à des interfaces de programmation qui accèdent à diverses ressources dédiées à la phosphoprotéomique et à l'analyse des protéines. Les informations additionnelles sur les kinases potentielles, le contexte structural, la proximité d'un domaine protéique, la conservation au cours de l'évolution, la médiation potentielle d'une interaction protéique phospho-dépendante et la compétition avec d'autres modifications pour une position précise sur la protéine sont annexées à chaque site identifié si disponible. Ces informations supplémentaires peuvent fournir un point de départ pour élucider la fonction d'un site particulier.

Ces annotations peuvent aussi être utilisées à l'échelle du phosphoprotéome pour déterminer les processus biologiques ou les fonctions moléculaires fréquemment régulés par la phosphorylation. À cette fin, nous avons utilisé l'ensemble des sites phosphorylés identifiés chez les cellules épithéliales intestinales de rat (IEC-6) lors de notre expérience phosphoprotéomique de l'inhibition pharmacologique de la voie de signalisation Erk1/2. Cet ensemble de 9615 sites de phosphorylation sur 2108 protéines est, jusqu'à ce jour, le plus important en nombre à être répertorié pour le phosphoprotéome de rat. À partir des 6419 sites localisés avec haute confiance, la fréquence de phosphorylation des domaines protéiques a été calculée avec un nouvel outil qui a été développé et intégré à ProteoConnections. Seulement 15% des sites, soit 865, sont situés à l'intérieur des limites d'un domaine protéique (Figure 2.3). Cette faible proportion indique que peu de sites modulent localement l'activité d'un domaine protéique. Cette valeur est toutefois liée à la connaissance actuelle des domaines protéiques. Les domaines kinases, avec 66 sites identifiés, forment la classe de domaines catalytiques la plus fréquemment phosphorylée. De plus, les domaines associés aux interactions sont les plus fréquemment phosphorylés parmi tous les domaines. On trouve 242 sites dans les domaines d'interactions avec les protéines, 137 sites avec les acides nucléiques et 28 sites associés à des domaines liant

d'autres types de molécules. Au niveau moléculaire, les sites de phosphorylation peuvent être situés stratégiquement à l'interface d'interaction tant pour les protéines et que les acides nucléiques (ADN et ARN) (Figure 2.5). Le groupement phosphate peut autant stabiliser que perturber un complexe d'interaction.

Les interactions protéiques phospho-dépendantes, c'est-à-dire celles intervenant entre un site phosphorylé et un domaine liant celui-ci comme 14-3-3, BRCT, C2, FHA, MH2, PBD, PTD, SH2, WD40, ont été recherchées avec un nouvel outil intégré à ProteoConnections. L'algorithme de cet outil superpose d'abord les identifications des sites phosphorylés sur le réseau d'interactions protéine-protéine STRING. Il scrute ensuite itérativement les interactions du réseau pour trouver celles faisant intervenir une protéine ayant un domaine liant les sites phosphorylés et une protéine avec un site phosphorylé dont le motif correspond au motif reconnu par le domaine. Les cas trouvés sont considérés comme des interactions potentiellement phospho-dépendantes. Pour les phosphoprotéines identifiées chez le rat, nous avons identifié 79 sur 412 interactions potentiellement phospho-dépendantes. Dix-sept d'entre elles correspondent à des interactions phospho-dépendantes connues avec les protéines de la famille 14-3-3. Le reste des interactions potentielles devront faire l'objet d'une validation expérimentale utilisant la mutagénèse dirigée afin de confirmer l'implication directe du site de phosphorylation et de son rôle fonctionnel.

Les séquences d'import/export au noyau (NLS et NES respectivement) sont connues pour être modulées par la phosphorylation. Dans le phosphoprotéome du rat, nous avons trouvé 182 et 627 sites de phosphorylation localisées à proximité des régions NLS et NES. Selon nos résultats, les domaines NLS sont enrichis en phosphorylation. Un autre aspect qui a été étudié, avec les outils de prédiction intégrés à ProteoConnections, est la compétition entre la phosphorylation et la glycosylation survenant sur les mêmes acides aminés. Seulement 301 des 6419 sites de phosphorylation identifiés avec haute confiance sont également glycosylés sur une sérine ou thréonine par le monosaccharide GlcNAc. Ces prédictions indiquent donc une proportion de compétition d'environ 5% entre ces deux

modifications. Les valeurs présentées ici dépendent des paramètres des outils employés qui contrôlent la sensibilité et la spécificité des prédictions. Cette valeur pourra possiblement être évaluée dans un avenir rapproché puisque des développements technologiques récents en glycoprotéomique permettront d'augmenter le nombre de sites de glycosylation répertoriés [212]. Le résumé de cette analyse de données démontre comment ProteoConnections est utile pour explorer divers aspects du phosphoprotéome.

ProteoConnections n'est pas la seule plateforme d'analyse existante et se distingue des autres plateformes par un schéma de base de données axé sur les modifications post-traductionnelles et aussi par l'intégration de ressources et d'outils bio-informatiques. Récemment, SysPTM [105] et PTMScout [104] qui sont deux plateformes similaires à ProteoConnections ont été publiées. Pour l'instant, ProteoConnections est la plateforme avec le plus d'outils dédiés à la phosphoprotéomique dont certains ont été développés pour l'analyse des domaines fréquemment phosphorylés et la recherche d'interactions phospho-dépendantes. Finalement, ProteoConnections a contribué à l'analyse de données de phosphoprotéomique pour divers projets menés dans notre laboratoire comme la découverte d'un artéfact de sulfatation pouvant être confondu avec la phosphorylation [35], l'analyse du phosphoprotéome du phagosome de macrophages activés par l'interféron gamma [182] et le phosphoprotéome de *L. kluyveri* [213]. Pour cette thèse, ProteoConnections a été un outil essentiel pour l'analyse phosphoprotéomique de l'inhibition pharmacologique de la voie de signalisation Erk1/2.

Afin de comprendre l'étendue de l'activité enzymatique des kinases Erk1/2 et leurs répercussions biologiques, il est indispensable de connaître l'ensemble des substrats de ces kinases. Une étude phosphoprotéomique de cinétique d'inhibition pharmacologique de la voie de signalisation Erk1/2 a alors été entreprise pour découvrir de nouveaux substrats (Chapitre 3). Plusieurs aspects distinguent cette étude des études phosphoprotéomiques précédentes sur le sujet [111-113, 192]. D'abord, une cinétique de phosphorylation avec quatre mesures temporelles (0, 5, 15, 60 minutes) a été effectuée afin de détecter les substrats phosphorylés à différents temps. Deuxièmement, nous avons eu recours à

l'inhibiteur Mek1/2 PD184352 (CI-1040, Pfizer) de seconde génération à une concentration limitée pour éviter l'inhibition de Mek5, plutôt que d'utiliser le U0126 de moindre spécificité. Cette précaution permettra, contrairement à certaines études précédentes, d'éviter de confondre les substrats de la kinase Erk5 pour ceux des kinases Erk1/2. Troisièmement, une analyse quantitative sans marquage a été effectuée avec 3 réplicats biologiques pour éviter l'incertitude sur les mesures d'abondance de phosphorylation. Les réplicats ne sont utilisés que dans très peu d'études de phosphoprotéomique quantitative. Cette analyse phosphoprotéomique a bénéficié d'un échantillonnage plus large du phosphoprotéome grâce à un fractionnement cellulaire (cytosol, noyau) et à trois étapes chromatographiques comprenant l'enrichissement des phosphopeptides au dioxyde de titane et deux autres étapes orthogonales de séparation utilisant l'échange cationique et la chromatographie à phase inverse. Ces analyses ont été effectuées sur un LTQ-Orbitrap XL, un spectromètre de masse avec un haut débit d'acquisition MS/MS. Au terme de cette analyse, les profils cinétiques d'abondance de phosphorylation de 7936 sites sur 1861 protéines ont été obtenus grâce à la quantification MS sans marquage et avec la reconstruction automatisée des profils avec le logiciel ProteoProfile.

Pour extraire une liste de substrats potentiels des kinases Erk1/2, un algorithme a été implémenté pour filtrer les sites identifiés selon certains critères prédéterminés. L'algorithme interroge ProteoConnections pour récupérer la liste des sites localisés avec haute confiance et portant le motif consensus minimal de phosphorylation [pS/T]P reconnu par les kinases Erk1/2. Ensuite, il sélectionne les sites dont l'abondance de la phosphorylation augmente suite à la stimulation au sérum et diminue suite à l'inhibition pharmacologique avec l'inhibiteur de kinases Mek1/2. Cet algorithme a généré une liste de 157 substrats potentiels des kinases Erk1/2 comprenant 233 sites de phosphorylation. Cette liste est actuellement la plus longue rapportée par une étude de phosphoprotéomique pour ces deux kinases.

Cette liste a été d'abord comparée à une compilation récente de 160 substrats confirmés et potentiels des kinases Erk1/2 [109] et à 4 études phosphoprotéomiques à large

échelle [111-113, 192] (Table A3.V). La comparaison a indiqué 32 substrats en commun entre notre liste et les 5 autres. Toutefois, seulement 12 ont exactement le même site phosphorylé. Parmi toutes nos identifications, 31 sites rapportés dans les autres listes n'ont pas été retenus n'étant pas localisés avec une confiance suffisante ou ne montrant pas un changement d'abondance de phosphorylation significatif et reproductible. La Figure 3.3 indique que plus de la moitié des substrats connus n'ont pas les changements d'abondances attendus. Pour la fraction nucléaire, les substrats connus manqués n'ont pas un changement significatif ou suffisamment important dans la direction attendue. Étrangement pour le cytosol, il y a plus de variabilité sur les cas de figures obtenus: 5 ont un changement faible dans la direction attendue, 4 sont stimulés par le sérum mais ne répondent pas à l'inhibiteur et 3 cas ont un changement modéré dans la direction opposée. Cette observation soulève quelques questions et critiques. D'abord pourquoi il y a-t-il plus de variabilité dans le cytosol que dans le noyau au niveau de la quantification des substrats connus? Est-ce que la condition cytosolique est plus bruitée et donc la quantification donne de faux négatifs? Les coefficients de variation moyens des peptides quantifiés n'indiquaient pas de différence majeure. De plus étant donné que l'abondance des peptides des substrats connus a été manuellement validée, les erreurs d'alignements qui donnent des résultats erronés sont aussi exclues. Ensuite, on note que la variation d'abondance des peptides de certains substrats connus est sur les limites définies pour être classé comme substrat potentiel. Il y a donc un compromis entre la sélectivité et la sensibilité dans notre approche. Finalement, il est aussi possible que les substrats des kinases Erk1/2 rapportés dans la littérature soit des faux positifs. Pour les études de cas par cas, quelques candidats n'ont peut-être pas été validés nécessairement avec rigueur et pour les autres études à grande échelle, elles sont aussi susceptibles que la notre de contenir des faux positifs. Certains cas pourraient même être le substrat d'une autre Map kinase. Par exemple pour les 4 cas où l'on note une stimulation au sérum mais pas d'inhibition, on pourrait suspecter qu'ils seraient les substrats de la kinase Erk5 étant donné qu'ils sont trouvés seulement dans les études utilisant le U0126. Dans ces études, contrairement à la notre, la concentration utilisée de cet inhibiteur était suffisante pour inhiber aussi Mek5.

La comparaison des études phosphoprotéomiques montre un faible recouvrement entre les listes des substrats des kinases Erk1/2 trouvés (Figure 3.4). Selon cette observation, il semble qu'aucune méthode ne soit actuellement capable de détecter l'ensemble des substrats des kinases Erk1/2 et que le développement de nouvelles méthodes soit toujours nécessaire. Plusieurs facteurs peuvent expliquer ce faible recouvrement: les types cellulaires, la stimulation/inhibition des cellules, les méthodes d'enrichissement des phosphopeptides, les méthodes de quantification et les différents spectromètres de masse. Le problème majeur qui ressort de notre analyse lors de la comparaison de ces listes est la difficulté d'identifier les protéines phosphorylées qui devraient être présentes universellement dans tous les types cellulaires. Il est évident que les différentes procédures utilisées ont un degré différent de sensibilité qui permettait l'identification d'un ensemble restreint de protéines. Au niveau du traitement des données, certaines études (dont la notre) ont peut-être rejeté des peptides avec un score faible ou des sites avec une faible confiance de localisation. L'assignation des peptides aux mêmes protéines et la comparaison des identifiants entre les espèces peuvent aussi compliquer la comparaison. L'enrichissement des phosphopeptides est aussi probablement fortement responsable du faible recouvrement. Une étude comparative des méthodes d'enrichissement a démontré qu'il y a environ 75% de recoupement détecté par LC-MS (peptides détectés en MS) entre deux isolats de phosphopeptides indépendamment de la méthode et qu'il y a seulement 35% d'identifications communes (MS et MS/MS) entre IMAC et TiO_2 [29]. Au niveau quantitatif, il est possible que des faux positifs ou négatifs sortent dans toutes les études phosphoprotéomiques. Les études faites avec SILAC sont susceptibles, malgré la robustesse de la méthode, de contenir des faux positifs puisque la reproductibilité des résultats n'a pas été évaluée (un seul réplicat). Une autre limite des analyses quantitatives est que certains peptides identifiés n'ont pas de valeur quantitative s'ils ont une faible abondance. La complexité des échantillons analysés peut aussi confondre les algorithmes dans les pics sélectionnés pour la quantification. Comme indiqué à la figure Figure 3.3, nous avons observé des variations d'abondance contradictoires pour certains substrats connus. Outre la variabilité technique, des facteurs biologiques peuvent expliquer les

différences entre les études. Il est possible pour certaines protéines que la variation d'abondance observée ne soit pas due à la phosphorylation mais bien au niveau d'expression de la protéine. Aucune étude n'a effectué une validation systématique de cet aspect. Les espèces et les types cellulaires sont deux autres facteurs de différences. Entre les espèces, on note qu'une fraction des sites phosphorylés n'est pas conservé (différence de 7% rat/souris et 13% rat/humain pour tous les sites identifiés chez le rat) et l'activité des kinases/phosphatases est peut-être aussi régulée légèrement de façon différente. Au niveau du type cellulaire, les études ont employé des cellules épithéliales de colons (IEC-6), des fibroblastes (NIH-3T3) et des cellules cancéreuses épithéliales du col de l'utérus (HeLa). Une étude récente a démontré à l'échelle du phosphoprotéome pour 9 différents tissus de souris que 50% des sites sont identifiable dans un seul tissu [214]. Finalement, la stimulation et l'inhibition des cellules peuvent aussi introduire des différences. Les différents agents de stimulation (EGF, sérum et 4-hydroxy-tamoxifen) activent différentes ramification des voies de signalisation. Pour les inhibiteurs de Mek1/2 (U0126 et PD184352), des effets sur d'autres kinases (comme Mek5) ou autres protéines (non-documenté) sont possible. Bref, la différence de recouvrement entre les études est multifactorielle.

Ayant constaté le faible recouvrement entre les études, il est logique de s'interroger sur la proportion de protéines de ces listes qui sont réellement des substrats des kinases Erk1/2 et ceux qui sont des faux positifs. Les effets pléiotropiques directs ou indirects de l'inhibiteur des kinases Mek1/2 sur les diverses voies de signalisation affectent l'ensemble du phosphoprotéome. L'activité des phosphatases et des kinases telles CkI, Cdk, Gsk3, Jnk1, Sapk et d'autres Mapk qui phosphorylent aussi le consensus minimal [pS/T]P pourraient être responsables de la variation d'abondance observée. Les substrats potentiels doivent donc être validés par une méthode alternative, tel que des essais kinases *in vitro*, pour démontrer que les kinases Erk1/2 sont bien responsables. Toutefois, l'ampleur du travail et le temps nécessaire pour tous les valider est trop important pour un seul

laboratoire. C'est donc pourquoi chaque étude, incluant la notre, a validé quelques substrats seulement.

Parmi la liste des 157 substrats potentiels trouvés, six substrats ont été sélectionnés pour validation en fonction de leur intérêt biologique et l'accessibilité à des réactifs pour essais *in vitro* ou *ex vivo*. Les substrats suivants ont été confirmés par un essai kinase *in vitro* avec Erk1: Ddx47 S9 (probablement une hélicase d'ARN ou hydrolase), Hmg20a S105 (régulateur de la chromatine), Junb S256 (facteur de transcription), Map2k2 S295 (kinase), Numa1 T1994 (probablement un constituent structural du noyau), Rras2 S186 (signalisation). Selon les résultats préliminaires de nos expériences d'immunofluorescence, la localisation nucléocytoplasmique de Hmg20a est modulée par la phosphorylation de la sérine 105 par les kinases Erk1/2. La forme sauvage de Hmg20a montre 16% plus de localisation nucléaire que la forme mutante (Ser105Ala). Étant donné que le site phosphorylé est près d'un NLS et d'une boîte HMG liant l'ADN, ce site pourrait soit affecter l'import/export de la protéine au noyau ou encore influencer la capacité de la protéine à lier l'ADN.

Une analyse ontologique de gènes sur notre liste de substrats potentiels des kinases Erk1/2 a permis d'observer des liens fonctionnels avec la liste de substrats connus compilée par Yoon et al. [109]. Plus du tiers des protéines sont annotées avec une des catégories suivantes associées aux substrats connus : régulation de la transcription, cytosquelette, signalisation, apoptose, kinase et phosphatase. Outre les fonctions déjà associées, il a été intéressant d'observer que deux catégories fonctionnelles, « liaison aux acides nucléiques » et « métabolisme des acides nucléiques », sont représentées avec une haute fréquence dans notre liste de substrats. Cette catégorie inclue plusieurs facteurs d'épissage alternatif dont des protéines de la famille SR. Ces résultats suggèrent une implication nouvelle des kinases Erk1/2 dans l'épissage alternatif des transcrits.

Le Chapitre 4 présente l'étude des phosphopeptides isomériques positionnels dont la phosphorylation peut survenir sur différent sites de la même séquence peptidique. Notre

attention a été portée sur ces peptides suite à l'observation de leur présence lors de la validation des données quantitatives des substrats potentiels des kinases Erk1/2. Trois substrats potentiels démontraient la présence d'isomères positionnels : Ahnak phosphorylé à la sérine 5397 (ASLGSLEGEAEAETSSPK en S16, alternativement en S2 et S5), Srrm2 en S1345 (SSSELSPEIVEK en S6 et alternativement en S3) et Rsf1 en S1366 (VGSPLDYSLVDLPSTNGQSPGK en S19, alternativement en S3). La présence de ces isomères est problématique lors de la quantification puisqu'ils causent une ambigüité potentielle pour sélectionner le précurseur correspondant lorsque plusieurs pics au même *m/z* sont présents dans un court intervalle de temps. Une co-élution de ces isomères peut aussi rendre ambigüe la localisation du site et certains peuvent être manqués lors de l'analyse si la séparation n'est pas suffisante. Devant ces problèmes, il nous est apparu important de déterminer la fréquence des phosphopeptides avec des isomères positionnels. Les études antérieures sur le sujet ne traitent que de quelques exemples qui se limitent à rapporter la séparation observée ou à résoudre la séparation problématique de cas biologiquement importants avec la chromatographie, l'électrophorèse capillaire ou la mobilité ionique couplée à la spectrométrie de masse.

Pour déterminer la proportion de phosphopeptides isomériques positionnels dans les analyses phosphoprotéomiques, un algorithme a été implémenté pour détecter ces isomères à partir de leur profil d'élution LC-MS et des identifications par MS/MS. L'analyse de données phosphoprotéomiques provenant des cellules IEC-6 et J774 avec cet algorithme indique que les isomères séparés par chromatographie en phase inverse représentent moins d'un pourcent de tous les phosphopeptides identifiés. Cependant, ce nombre est possiblement sous-estimé compte tenu du nombre de MS/MS ne permettant pas une identification définitive, les cas d'isomères n'ayant pas été échantillonnés par MS/MS et les co-élutions potentielles. La combinaison des deux analyses a permis d'obtenir une liste de 64 paires de phosphopeptides isomériques. Différentes propriétés ont été calculées à partir de la séquence primaire des peptides pour tenter de comprendre les différences de temps de rétention observées sur la colonne avec phase inverse C_{18}. Aucune corrélation n'a été

observée entre la différence de temps et les propriétés suivantes : longueur du peptide, différence de position du site phosphorylé, l'hydrophobicité locale et la proximité du site phosphorylé avec les régions chargées du peptide. Selon les études précédentes, les différentes conformations tridimensionnelles des isomères expliqueraient la différence de séparation observée [16, 17]_ENREF_19. Les prédictions de structures des isomères positionnels, où la phosphorylation a été simulée par un acide glutamique, indiquent que certains cas peuvent être expliqués par un changement de conformation important comme, par exemple, la transition d'une structure désordonnée à une hélice (Figure A4.3). La surface hydrophobe et chargée exposée au solvant pourrait également expliquer la rétention sur la phase inverse C_{18} et expliquer l'ordre d'élution. Toutefois, cette approche n'a pas été approfondie étant donné qu'elle est limitée par l'exactitude des prédictions. Celle-ci ne considère pas la dynamique conformationnelle se produisant lors d'un gradient eau-acétonitrile utilisé lors de la chromatographie. De plus, il faudrait aussi considérer la surface effective qui interagit avec la phase stationnaire de la colonne.

L'algorithme de détection d'isomères positionnels basé sur le profil d'élution LC-MS a aussi été conçu pour générer des listes d'inclusions afin de faire une analyse LC-MS/MS ciblée sur les isomères qui n'ont pas été identifiés par la méthode conventionnelle. Cet algorithme a d'abord été évalué avec un ensemble de 9 phosphopeptides isomériques positionnels. Ce test a révélé que l'algorithme détecte les isomères bien séparés mais aussi la présence d'artéfacts comme des conformères qui peuvent être séparés par chromatographie. La difficulté la plus importante pour détecter et confirmer la présence d'isomères est la qualité des spectres MS/MS tant en mode CID et ETD qui est parfois insuffisante pour localiser le site avec confiance. Théoriquement, il serait bénéfique de combiner les fragments complémentaires des spectres CID et ETD (*b/y* et *c/z* respectivement) avant d'effectuer la recherche pour augmenter la confiance de localisation des sites. Les résultats d'une telle approche pourraient toutefois être mitigés tel qu'observé dans le cadre d'une étude de phosphoprotéomique [215]. Cette étude a observé une augmentation marginale du pointage associé aux phosphopeptides identifiés. Certains

spectres ont obtenu un pointage plus élevé et d'autres plus faible. Il a été suggéré que l'impossibilité de distinguer les ions fragments *b/y* des ions *c/z*, et l'augmentation du nombre de fragments augmentent la probabilité d'une assignation erronée. Pour bénéficier de l'information complémentaire des spectres CID et ETD, il serait d'abord pertinent de revoir les algorithmes de recherche ainsi que les méthodes du calcul de pointage. Des spectres MS/MS de haute résolution pourraient être bénéfiques pour cette application.

Une expérience supplémentaire pour évaluer l'algorithme aurait été d'ajouter les peptides synthétiques des isomères positionnels dans un extrait enrichi de phosphopeptides. La complexité de cet échantillon serait plus représentative de l'usage anticipé de l'algorithme et permettrait de mieux estimer le taux de fausses découvertes. De plus, cette expérience permettrait d'optimiser la méthode d'acquisition (temps d'exclusion dynamique et nombre de précurseurs pour l'inclusion) afin de s'assurer que tous les isomères positionnels attendus soient identifiés.

Ensuite, une analyse ciblée a été effectuée avec l'algorithme sur un extrait de phosphopeptides enrichis provenant de la lignée cellulaire S2 de *Drosophila melanogaster*. Cette analyse nous a permis d'identifier 117 phosphopeptides isomériques, soit 1.2% de tous les phosphopeptides identifiés. L'algorithme est complémentaire à la méthode conventionnelle non-ciblée. La méthode d'acquisition employant la liste d'inclusion devra toutefois être optimisée afin d'obtenir des MS/MS sur l'ensemble des ions peptidiques inclus. Une difficulté supplémentaire observée lors de cette analyse est la présence d'artéfacts isobariques (différentes séquences peptidiques, digestions enzymatiques alternatives, modifications alternatives, conformères) qui sont confondus pour des isomères positionnels phosphorylés. Un second algorithme a été employé pour détecter les isomères coéluant à partir des spectres MS/MS. Selon les résultats obtenus, il a été estimé qu'au moins 0.8% des phosphopeptides pourraient être des isomères positionnels coéluant. Des expériences supplémentaires seront nécessaires pour confirmer l'ensemble des cas trouvés et déterminer quelles méthodes permettent leur séparation. Tel qu'indiqué dans l'introduction du chapitre 4, un éventail de méthodes alternatives à la chromatographie

liquide en phase inverse ont été investiguées pour séparer les cas problématiques d'isomères positionnels coéluants. Les approches chromatographiques alternatives (HILIC, colonne monolithique) sont compatibles avec les algorithmes développés dans ce projet. Elles seraient relativement faciles et économiques à déployer puisqu'elles ne nécessiteraient que le changement de la colonne et des solvants sur le système actuel chromatographique. L'électrophorèse capillaire en zone serait aussi théoriquement compatible avec notre algorithme. Cette technologie n'est toutefois pas aussi répandue dans les laboratoires de protéomique que la chromatographie liquide et nécessiterait pour plusieurs l'achat d'un système d'électrophorèse. La troisième alternative est la mobilité ionique. Cette approche nécessiterait l'adaptation notre algorithme pour la recherche des isomères positionnels. Cette technologie n'est pas standard sur les spectromètres de masse. Les instruments SYNAPT de Waters intègrent une cellule de mobilité ionique tandis que les instruments de type Orbitrap de Thermo Fisher Scientific nécessite l'ajout d'un appareil nommé FAIMS pour obtenir cette capacité. Une comparaison systématique de ces différentes méthodologies serait nécessaire pour évaluer lesquelles sont complémentaires et le plus simple à utiliser dans le contexte de la découverte d'isomères positionnels.

Ce quatrième chapitre a donc présenté la première étude sur la population de phosphopeptides isomériques positionnels et celle-ci a été possible grâce au développement de deux nouveaux algorithmes.

Perspectives

Les travaux de cette thèse ont permis d'approfondir nos connaissances en phosphoprotéomique et en bioinformatique, tout en permettant d'identifier de nouveaux substrats des kinases Erk1/2 d'intérêts biologiques. D'abord sur le plan bio-informatique, ProteoConnections, une plateforme d'analyse bio-informatique dédiée à l'analyse des données phosphoprotéomiques, a été développée. Cette plateforme a été utile au cours de ce projet ainsi que pour quelques études phosphoprotéomiques de notre laboratoire. Le code source de ProteoConnections a été rendu publiquement disponible afin que d'autres laboratoires puissent en bénéficier. La conception modulaire du code de cette plateforme permettrait aussi d'augmenter les fonctionnalités disponibles pour d'autres types d'expériences de protéomique. Il est possible d'envisager le développement d'outils permettant l'exploitation des identifications qui s'accumuleront dans la base de données au fil du temps. Ces données pourraient, par exemple, être utilisées pour déterminer sous quelles conditions une protéine ou un site de phosphorylation est observé par MS ou encore déterminer les peptides protéo-typiques qui pourraient être utilisés pour des études quantitatives. De plus, avec l'avènement de techniques d'enrichissement pour d'autres modifications post-traductionnelles, ProteoConnections pourrait déterminer s'il existe une interrelation entre celles-ci. Les phosphodégrons sont un exemple où la relation entre la phosphorylation et l'ubiquitination module la dégradation des protéines.

Deuxièmement, deux nouveaux algorithmes ont été conçus pour détecter et mettre en évidence la présence des phosphopeptides isomériques positionnels dans les extraits enrichis de phosphopeptides. Ces algorithmes pourront servir à la détection des sites phosphorylés biologiquement importants qui sont manqués par les méthodes conventionnelles et aussi à résoudre les ambiguïtés associées à leur présence au niveau de la localisation et de la quantification. L'usage de ces algorithmes sera donc pertinent avant d'entreprendre la validation fonctionnelle de certains sites de phosphorylation d'intérêt. Alternativement, ces algorithmes pourront être utilisés pour d'autres modifications post-traductionnelles qui forment des peptides isomériques. Pour la poursuite de ce projet, il faudra revoir ces algorithmes pour réduire le taux de faux positifs ou déterminer comment

prioriser les candidats dont la séparation est plus évidente. Les méthodes d'acquisitions de l'instrument devront aussi être optimisées afin de s'assurer que chaque candidat est inspecté ainsi que le plus grand nombre dans une seule analyse.

Troisièmement, un autre algorithme a été conçu pour détecter les substrats potentiels des kinases Erk1/2 à partir de leur motif consensus de phosphorylation et des profils cinétiques d'abondance. La procédure de cet algorithme pourra servir à d'autres études de découverte de substrats pour des kinases alternatives ou encore pour compléter l'identification des substrats des kinases Erk1/2 sous différentes conditions expérimentales.

L'étude phosphoprotéomique entreprise dans cette thèse a permis d'identifier plusieurs milliers de sites de phosphorylation chez les protéines du rat. Cette liste a été rendue publique et pourra servir de ressource pour l'étude fonctionnelle des protéines. De plus, pour la majorité des sites identifiés, un profil cinétique d'abondance a été obtenu suite à la stimulation des cellules au sérum et en présence de l'inhibiteur de Mek1/2. Donc contrairement à plusieurs études phosphoprotéomiques, ces données ne sont pas une liste statique de sites de phosphorylation et pourront potentiellement servir à d'autres chercheurs pour comprendre la modulation des voies de signalisation non explorées dans cette thèse. L'étude de la phosphorylation des protéines impliquées dans l'angiogenèse, la motilité et la prolifération serait un autre aspect intéressant qui pourrait être étudié avec ces données. Ces processus biologiques sont importants pour le développement et le traitement du cancer. Nos données montrent que la phosphorylation des protéines impliquées dans ces processus est affectée par l'inhibiteur de Mek1/2 (résultats non présentés). Des expériences pour comprendre le rôle de ces phosphorylations pourraient améliorer notre compréhension du mécanisme d'action des inhibiteurs de Mek1/2.

Enfin, une liste de 157 substrats potentiels des kinases Erk1/2 a été obtenue. Au terme de cette thèse, seulement 6 substrats ont été validés et plusieurs candidats restent à être confirmés comme substrats authentiques des kinases Erk1/2. Il sera donc important de déterminer la nature fonctionnelle de chacun des sites des substrats. Nos collaborateurs s'affèrent présentement à étudier le rôle de la phosphorylation de la sérine 105 de Hmg20a.

Ces nouvelles connaissances sur l'étendue enzymatique de la voie Erk1/2 et leurs effets sur l'ensemble du phosphoprotéome permettront peut-être un jour de développer de nouvelles thérapies pharmacologiques pour le traitement du cancer et d'autres maladies. Pour conclure, la comparaison approfondie avec les précédentes études phosphoprotéomiques a indiqué que chaque étude apporte de nouveaux substrats mais que celles-ci ne se recoupent que très peu. Il semble donc que la découverte de substrats des kinases Erk1/2 reste un défi majeur à résoudre et que de nouvelles méthodes analytiques plus sensibles et polyvalentes soient nécessaires pour y parvenir.

Bibliographie

[1] Moorhead, G. B., De Wever, V., Templeton, G., Kerk, D., Evolution of protein phosphatases in plants and animals. *Biochem J* 2009, *417*, 401-409.

[2] Piggee, C., Phosphoproteomics: miles to go before it's routine. *Anal Chem* 2009, *81*, 2418-2420.

[3] Serber, Z., Ferrell, J. E., Jr., Tuning bulk electrostatics to regulate protein function. *Cell* 2007, *128*, 441-444.

[4] Holt, L. J., Tuch, B. B., Villen, J., Johnson, A. D.*, et al.*, Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science* 2009, *325*, 1682-1686.

[5] Landry, C. R., Levy, E. D., Michnick, S. W., Weak functional constraints on phosphoproteomes. *Trends Genet* 2009, *25*, 193-197.

[6] Manning, G., Whyte, D. B., Martinez, R., Hunter, T., Sudarsanam, S., The protein kinase complement of the human genome. *Science* 2002, *298*, 1912-1934.

[7] Janne, P. A., Gray, N., Settleman, J., Factors underlying sensitivity of cancers to small-molecule kinase inhibitors. *Nat Rev Drug Discov* 2009, *8*, 709-723.

[8] Yaffe, M. B., Elia, A. E., Phosphoserine/threonine-binding domains. *Curr Opin Cell Biol* 2001, *13*, 131-138.

[9] Hjerrild, M., Gammeltoft, S., Phosphoproteomics toolbox: computational biology, protein chemistry and mass spectrometry. *FEBS Lett* 2006, *580*, 4764-4770.

[10] Stasyk, T., Morandell, S., Bakry, R., Feuerstein, I.*, et al.*, Quantitative detection of phosphoproteins by combination of two-dimensional difference gel electrophoresis and phosphospecific fluorescent staining. *Electrophoresis* 2005, *26*, 2850-2854.

[11] Ptacek, J., Devgan, G., Michaud, G., Zhu, H.*, et al.*, Global analysis of protein phosphorylation in yeast. *Nature* 2005, *438*, 679-684.

[12] Ptacek, J., Snyder, M., Charging it up: global analysis of protein phosphorylation. *Trends Genet* 2006, *22*, 545-554.

[13] Thingholm, T. E., Jensen, O. N., Larsen, M. R., Analytical strategies for phosphoproteomics. *Proteomics* 2009, *9*, 1451-1468.

[14] Tholey, A., Toll, H., Huber, C. G., Separation and detection of phosphorylated and nonphosphorylated peptides in liquid chromatography-mass spectrometry using monolithic columns and acidic or alkaline mobile phases. *Anal Chem* 2005, *77*, 4618-4625.

[15] Singer, D., Kuhlmann, J., Muschket, M., Hoffmann, R., Separation of multiphosphorylated peptide isomers by hydrophilic interaction chromatography on an aminopropyl phase. *Anal Chem* 2010, *82*, 6409-6414.

[16] Muetzelburg, M. V., Hoffmann, R., Separation of multiphosphorylated peptide isomers by CZE. *Electrophoresis* 2008, *29*, 4381-4385.

[17] Xuan, Y., Creese, A. J., Horner, J. A., Cooper, H. J., High-field asymmetric waveform ion mobility spectrometry (FAIMS) coupled with high-resolution electron transfer dissociation mass spectrometry for the analysis of isobaric phosphopeptides. *Rapid Commun Mass Spectrom* 2009, *23*, 1963-1969.

[18] Washburn, M. P., Wolters, D., Yates, J. R., 3rd, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* 2001, *19*, 242-247.

[19] Olsen, J. V., Blagoev, B., Gnad, F., Macek, B.*, et al.*, Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* 2006, *127*, 635-648.

[20] Zhang, Y., Wolf-Yadlin, A., Ross, P. L., Pappin, D. J.*, et al.*, Time-resolved mass spectrometry of tyrosine phosphorylation sites in the epidermal growth factor receptor signaling network reveals dynamic modules. *Molecular & cellular proteomics : MCP* 2005, *4*, 1240-1250.

[21] Sopko, R., Andrews, B. J., Linking the kinome and phosphorylome--a comprehensive review of approaches to find kinase targets. *Molecular bioSystems* 2008, *4*, 920-933.

[22] Edbauer, D., Cheng, D., Batterton, M. N., Wang, C. F.*, et al.*, Identification and characterization of neuronal mitogen-activated protein kinase substrates using a specific phosphomotif antibody. *Molecular & cellular proteomics : MCP* 2009, *8*, 681-695.

[23] Andersson, L., Porath, J., Isolation of phosphoproteins by immobilized metal (Fe3+) affinity chromatography. *Anal Biochem* 1986, *154*, 250-254.

[24] Ficarro, S. B., McCleland, M. L., Stukenberg, P. T., Burke, D. J.*, et al.*, Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nat Biotechnol* 2002, *20*, 301-305.

[25] Pinkse, M. W., Uitto, P. M., Hilhorst, M. J., Ooms, B., Heck, A. J., Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns. *Anal Chem* 2004, *76*, 3935-3943.

[26] Larsen, M. R., Thingholm, T. E., Jensen, O. N., Roepstorff, P., Jorgensen, T. J., Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns. *Molecular & cellular proteomics : MCP* 2005, *4*, 873-886.

[27] Oda, Y., Nagasu, T., Chait, B. T., Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. *Nat Biotechnol* 2001, *19*, 379-382.

[28] Zhou, H., Watts, J. D., Aebersold, R., A systematic approach to the analysis of protein phosphorylation. *Nat Biotechnol* 2001, *19*, 375-378.

[29] Bodenmiller, B., Mueller, L. N., Mueller, M., Domon, B., Aebersold, R., Reproducible isolation of distinct, overlapping segments of the phosphoproteome. *Nat Methods* 2007, *4*, 231-237.

[30] Karas, M., Hillenkamp, F., Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* 1988, *60*, 2299-2301.

[31] Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., Whitehouse, C. M., Electrospray ionization for mass spectrometry of large biomolecules. *Science* 1989, *246*, 64-71.

[32] Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., *et al.*, Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal Chem* 2006, *78*, 2113-2120.

[33] Steen, H., Mann, M., The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* 2004, *5*, 699-711.

[34] Medzihradszky, K. F., Darula, Z., Perlson, E., Fainzilber, M., *et al.*, O-sulfonation of serine and threonine: mass spectrometric detection and characterization of a new posttranslational modification in diverse proteins throughout the eukaryotes. *Molecular & cellular proteomics : MCP* 2004, *3*, 429-440.

[35] Gharib, M., Marcantonio, M., Lehmann, S. G., Courcelles, M., *et al.*, Artifactual sulfation of silver-stained proteins: implications for the assignment of phosphorylation and sulfation sites. *Molecular & cellular proteomics : MCP* 2009, *8*, 506-518.

[36] Beausoleil, S. A., Jedrychowski, M., Schwartz, D., Elias, J. E., *et al.*, Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc Natl Acad Sci U S A* 2004, *101*, 12130-12135.

[37] Schroeder, M. J., Shabanowitz, J., Schwartz, J. C., Hunt, D. F., Coon, J. J., A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry. *Anal Chem* 2004, *76*, 3590-3598.

[38] Swaney, D. L., Wenger, C. D., Thomson, J. A., Coon, J. J., Human embryonic stem cell phosphoproteome revealed by electron transfer dissociation tandem mass spectrometry. *Proc Natl Acad Sci U S A* 2009, *106*, 995-1000.

[39] Swaney, D. L., McAlister, G. C., Coon, J. J., Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat Methods* 2008, *5*, 959-964.

[40] Elschenbroich, S., Kislinger, T., Targeted proteomics by selected reaction monitoring mass spectrometry: applications to systems biology and biomarker discovery. *Molecular bioSystems* 2010, *7*, 292-303.

[41] St-Denis, N., Gingras, A. C., Mass spectrometric tools for systematic analysis of protein phosphorylation. *Progress in molecular biology and translational science* 2012, *106*, 3-32.

[42] Mueller, L. N., Brusniak, M. Y., Mani, D. R., Aebersold, R., An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. *J Proteome Res* 2008, *7*, 51-61.

[43] Ong, S. E., Kratchmarova, I., Mann, M., Properties of 13C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC). *J Proteome Res* 2003, *2*, 173-181.

[44] Geiger, T., Wisniewski, J. R., Cox, J., Zanivan, S., *et al.*, Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics. *Nature protocols* 2011, *6*, 147-157.

[45] Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., *et al.*, Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 1999, *17*, 994-999.

[46] Gevaert, K., Impens, F., Ghesquiere, B., Van Damme, P., *et al.*, Stable isotopic labeling in proteomics. *Proteomics* 2008, *8*, 4873-4885.

[47] Thompson, A., Schafer, J., Kuhn, K., Kienle, S., *et al.*, Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* 2003, *75*, 1895-1904.

[48] Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., *et al.*, Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics : MCP* 2004, *3*, 1154-1169.

[49] Thingholm, T. E., Palmisano, G., Kjeldsen, F., Larsen, M. R., Undesirable charge-enhancement of isobaric tagged phosphopeptides leads to reduced identification efficiency. *J Proteome Res* 2010, *9*, 4045-4052.

[50] Hoffert, J. D., Pisitkun, T., Saeed, F., Song, J. H., *et al.*, Dynamics of the G protein-coupled vasopressin V2 receptor signaling network revealed by quantitative phosphoproteomics. *Molecular & cellular proteomics : MCP* 2012, *11*, M111 014613.

[51] Liu, H., Sadygov, R. G., Yates, J. R., 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* 2004, *76*, 4193-4201.

[52] Bondarenko, P. V., Chelius, D., Shaler, T. A., Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal Chem* 2002, *74*, 4741-4749.

[53] Eidhammer, I., *Computational methods for mass spectrometry proteomics*, John Wiley & Sons, Chichester, England ; Hoboken, NJ 2007.

[54] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999, *20*, 3551-3567.

[55] Yates, J. R., 3rd, Eng, J. K., McCormack, A. L., Schieltz, D., Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* 1995, *67*, 1426-1436.

[56] Craig, R., Beavis, R. C., TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004, *20*, 1466-1467.

[57] Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., *et al.*, Open mass spectrometry search algorithm. *J Proteome Res* 2004, *3*, 958-964.

[58] Clauser, K. R., Baker, P., Burlingame, A. L., Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal Chem* 1999, *71*, 2871-2882.

[59] Kall, L., Storey, J. D., MacCoss, M. J., Noble, W. S., Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J Proteome Res* 2008, *7*, 29-34.

[60] Taylor, J. A., Johnson, R. S., Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal Chem* 2001, *73*, 2594-2604.

[61] Ma, B., Zhang, K., Hendrie, C., Liang, C.*, et al.*, PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2003, *17*, 2337-2342.

[62] Frank, A., Pevzner, P., PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* 2005, *77*, 964-973.

[63] Mann, M., Wilm, M., Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem* 1994, *66*, 4390-4399.

[64] Yen, C. Y., Meyer-Arendt, K., Eichelberger, B., Sun, S.*, et al.*, A simulated MS/MS library for spectrum-to-spectrum searching in large scale identification of proteins. *Molecular & cellular proteomics : MCP* 2009, *8*, 857-869.

[65] Frewen, B., MacCoss, M. J., Using BiblioSpec for creating and searching tandem MS peptide libraries. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* 2007, *Chapter 13*, Unit 13 17.

[66] Falkner, J. A., Falkner, J. W., Yocum, A. K., Andrews, P. C., A spectral clustering approach to MS/MS identification of post-translational modifications. *J Proteome Res* 2008, *7*, 4614-4622.

[67] Lam, H., Deutsch, E. W., Eddes, J. S., Eng, J. K.*, et al.*, Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* 2007, *7*, 655-667.

[68] Fenyo, D., Eriksson, J., Beavis, R., Mass spectrometric protein identification using the global proteome machine. *Methods in molecular biology* 2010, *673*, 189-202.

[69] Deutsch, E. W., The PeptideAtlas Project. *Methods in molecular biology* 2010, *604*, 285-296.

[70] Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J., Gygi, S. P., A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol* 2006, *24*, 1285-1292.

[71] Olsen, J. V., Mann, M., Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc Natl Acad Sci U S A* 2004, *101*, 13417-13422.

[72] Wan, Y., Cripps, D., Thomas, S., Campbell, P., *et al.*, PhosphoScan: a probability-based method for phosphorylation site prediction using MS2/MS3 pair information. *J Proteome Res* 2008, *7*, 2803-2811.

[73] MacLean, D., Burrell, M. A., Studholme, D. J., Jones, A. M., PhosCalc: a tool for evaluating the sites of peptide phosphorylation from mass spectrometer data. *BMC Res Notes* 2008, *1*, 30.

[74] Bailey, C. M., Sweet, S. M., Cunningham, D. L., Zeller, M., *et al.*, SLoMo: automated site localization of modifications from ETD/ECD mass spectra. *J Proteome Res* 2009, *8*, 1965-1971.

[75] Savitski, M. M., Lemeer, S., Boesche, M., Lang, M., *et al.*, Confident phosphorylation site localization using the Mascot Delta Score. *Molecular & cellular proteomics : MCP* 2010, M110.003830.

[76] Horn, D. M., Zubarev, R. A., McLafferty, F. W., Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J Am Soc Mass Spectrom* 2000, *11*, 320-332.

[77] Li, X. J., Yi, E. C., Kemp, C. J., Zhang, H., Aebersold, R., A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Molecular & cellular proteomics : MCP* 2005, *4*, 1328-1340.

[78] May, D., Fitzgibbon, M., Liu, Y., Holzman, T., *et al.*, A platform for accurate mass and time analyses of mass spectrometry data. *J Proteome Res* 2007, *6*, 2685-2694.

[79] Palagi, P. M., Walther, D., Quadroni, M., Catherinet, S., *et al.*, MSight: an image analysis software for liquid chromatography-mass spectrometry. *Proteomics* 2005, *5*, 2381-2384.

[80] Kohlbacher, O., Reinert, K., Gropl, C., Lange, E., *et al.*, TOPP--the OpenMS proteomics pipeline. *Bioinformatics* 2007, *23*, e191-197.

[81] Jaffe, J. D., Mani, D. R., Leptos, K. C., Church, G. M., *et al.*, PEPPeR, a platform for experimental proteomic pattern recognition. *Molecular & cellular proteomics : MCP* 2006, *5*, 1927-1941.

[82] Mueller, L. N., Rinner, O., Schmidt, A., Letarte, S., *et al.*, SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* 2007, *7*, 3470-3480.

[83] Neilson, K. A., Ali, N. A., Muralidharan, S., Mirzaei, M., *et al.*, Less label, more free: approaches in label-free quantitative mass spectrometry. *Proteomics* 2011, *11*, 535-553.

[84] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., *et al.*, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research* 2003, *31*, 365-370.

[85] Hornbeck, P. V., Chabra, I., Kornhauser, J. M., Skrzypek, E., Zhang, B., PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* 2004, *4*, 1551-1561.

[86] Diella, F., Gould, C. M., Chica, C., Via, A., Gibson, T. J., Phospho.ELM: a database of phosphorylation sites--update 2008. *Nucleic acids research* 2008, *36*, D240-244.

[87] Gnad, F., Ren, S., Cox, J., Olsen, J. V.*, et al.*, PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol* 2007, *8*, R250.

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

[89] Bairoch, A., The PROSITE dictionary of sites and patterns in proteins, its current status. *Nucleic acids research* 1993, *21*, 3097-3103.

[90] Obenauer, J. C., Cantley, L. C., Yaffe, M. B., Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic acids research* 2003, *31*, 3635-3641.

[91] Blom, N., Gammeltoft, S., Brunak, S., Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol* 1999, *294*, 1351-1362.

[92] Huang, H. D., Lee, T. Y., Tzeng, S. W., Horng, J. T., KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic acids research* 2005, *33*, W226-229.

[93] Kim, J. H., Lee, J., Oh, B., Kimm, K., Koh, I., Prediction of phosphorylation sites using SVMs. *Bioinformatics* 2004, *20*, 3179-3184.

[94] Xue, Y., Li, A., Wang, L., Feng, H., Yao, X., PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC bioinformatics [electronic resource]* 2006, *7*, 163.

[95] Linding, R., Jensen, L. J., Pasculescu, A., Olhovsky, M.*, et al.*, NetworKIN: a resource for exploring cellular phosphorylation networks. *Nucleic acids research* 2008, *36*, D695-699.

[96] Schwartz, D., Gygi, S. P., An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat Biotechnol* 2005, *23*, 1391-1398.

[97] Ritz, A., Shakhnarovich, G., Salomon, A. R., Raphael, B. J., Discovery of phosphorylation motif mixtures in phosphoproteomics data. *Bioinformatics* 2009, *25*, 14-21.

[98] Ren, J., Gao, X., Liu, Z., Cao, J.*, et al.*, Computational analysis of phosphoproteomics: progresses and perspectives. *Current protein & peptide science* 2011, *12*, 591-601.

[99] Julenius, K., Molgaard, A., Gupta, R., Brunak, S., Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology* 2005, *15*, 153-164.

[100] Gupta, R., Brunak, S., Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac Symp Biocomput* 2002, 310-322.

[101] Hu, Z.-Z., *dbOGAP: A Bioinformatics Resource for the O-GlcNAcylated Proteins and Sites, 6th US HUPO Annual Meeting*, Denver, CO 2010.

[102] Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A.*, et al.*, InterPro: the integrative protein signature database. *Nucleic acids research* 2009, *37*, D211-215.

[103] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G.*, et al.*, The Protein Data Bank. *Nucleic acids research* 2000, *28*, 235-242.

[104] Naegle, K. M., Gymrek, M., Joughin, B. A., Wagner, J. P.*, et al.*, PTMScout, a Web resource for analysis of high throughput post-translational proteomics studies. *Molecular & cellular proteomics : MCP* 2010, *9*, 2558-2570.

[105] Li, H., Xing, X., Ding, G., Li, Q.*, et al.*, SysPTM: a systematic resource for proteomic research on post-translational modifications. *Molecular & cellular proteomics : MCP* 2009, *8*, 1839-1849.

[106] Alberts, B., *Molecular biology of the cell*, Garland Science, New York 2002.

[107] Kohno, M., Pouyssegur, J., Targeting the ERK signaling pathway in cancer therapy. *Ann Med* 2006, *38*, 200-211.

[108] LeFloch, R., *Les isoformes ERK1 et ERK2 ont-elles les mêmes fonctions cellulaires*, Universite de Nice-Sophia Antipolis - UFR Sciences, Ecole doctorale des sciences de la vie et de la sante, Nice 2007.

[109] Yoon, S., Seger, R., The extracellular signal-regulated kinase: multiple substrates regulate diverse cellular functions. *Growth Factors* 2006, *24*, 21-44.

[110] !!! INVALID CITATION !!!

[111] Old, W. M., Shabb, J. B., Houel, S., Wang, H.*, et al.*, Functional proteomics identifies targets of phosphorylation by B-Raf signaling in melanoma. *Mol Cell* 2009, *34*, 115-131.

[112] Kosako, H., Yamaguchi, N., Aranami, C., Ushiyama, M.*, et al.*, Phosphoproteomics reveals new ERK MAP kinase targets and links ERK to nucleoporin-mediated nuclear transport. *Nat Struct Mol Biol* 2009, *16*, 1026-1035.

[113] Carlson, S. M., Chouinard, C. R., Labadorf, A., Lam, C. J.*, et al.*, Large-Scale Discovery of ERK2 Substrates Identifies ERK-Mediated Transcriptional Regulation by ETV3. *Sci Signal* 2011, *4*, rs11.

[114] Olive, D. M., Quantitative methods for the analysis of protein phosphorylation in drug development. *Expert Rev Proteomics* 2004, *1*, 327-341.

[115] Hunter, T., The age of crosstalk: phosphorylation, ubiquitination, and beyond. *Mol Cell* 2007, *28*, 730-738.

[116] Amanchy, R., Periaswamy, B., Mathivanan, S., Reddy, R.*, et al.*, A curated compendium of phosphorylation motifs. *Nat Biotechnol* 2007, *25*, 285-286.

[117] Courcelles, M., Lemieux, S., Voisin, L., Meloche, S., Thibault, P., ProteoConnections: a bioinformatics platform to facilitate proteome and phosphoproteome analyses. *Proteomics* 2011, *11*, 2654-2671.

[118] Keller, A., Eng, J., Zhang, N., Li, X. J., Aebersold, R., A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* 2005, *1*, 2005 0017.

[119] Matthiesen, R., Virtual Expert Mass Spectrometrist v3.0: an integrated tool for proteome analysis. *Methods in molecular biology* 2007, *367*, 121-138.

[120] Gehlenborg, N., Yan, W., Lee, I. Y., Yoo, H.*, et al.*, Prequips--an extensible software platform for integration, visualization and analysis of LC-MS/MS proteomics data. *Bioinformatics* 2009, *25*, 682-683.

[121] Yu, K., Salomon, A. R., HTAPP: High-throughput autonomous proteomic pipeline. *Proteomics* 2010, *10*, 2113-2122.

[122] Garwood, K., McLaughlin, T., Garwood, C., Joens, S.*, et al.*, PEDRo: a database for storing, searching and disseminating experimental proteomics data. *BMC Genomics* 2004, *5*, 68.

[123] Kiebel, G. R., Auberry, K. J., Jaitly, N., Clark, D. A.*, et al.*, PRISM: a data management system for high-throughput proteomics. *Proteomics* 2006, *6*, 1783-1790.

[124] Poullet, P., Carpentier, S., Barillot, E., myProMS, a web server for management and validation of mass spectrometry-based proteomic data. *Proteomics* 2007, *7*, 2553-2556.

[125] Hartler, J., Thallinger, G. G., Stocker, G., Sturn, A.*, et al.*, MASPECTRAS: a platform for management and analysis of proteomics LC-MS/MS data. *BMC bioinformatics [electronic resource]* 2007, *8*, 197.

[126] Howes, C. G., Foster, L. J., PrestOMIC, an open source application for dissemination of proteomic datasets by individual laboratories. *Proteome Sci* 2007, *5*, 8.

[127] Yu, K., Salomon, A. R., PeptideDepot: flexible relational database for visual analysis of quantitative proteomic data and integration of existing protein information. *Proteomics* 2009, *9*, 5350-5358.

[128] Malmstrom, L., Marko-Varga, G., Westergren-Thorsson, G., Laurell, T., Malmstrom, J., 2DDB - a bioinformatics solution for analysis of quantitative proteomics data. *BMC bioinformatics [electronic resource]* 2006, *7*, 158.

[129] Albaum, S. P., Neuweger, H., Franzel, B., Lange, S.*, et al.*, Qupe--a Rich Internet Application to take a step forward in the analysis of mass spectrometry-based quantitative proteomics experiments. *Bioinformatics* 2009, *25*, 3128-3134.

[130] Prince, J. T., Carlson, M. W., Wang, R., Lu, P., Marcotte, E. M., The need for a public proteomics repository. *Nat Biotechnol* 2004, *22*, 471-472.

[131] Jones, P., Cote, R. G., Cho, S. Y., Klie, S.*, et al.*, PRIDE: new developments and new datasets. *Nucleic acids research* 2008, *36*, D878-883.

[132] Craig, R., Cortens, J. P., Beavis, R. C., Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* 2004, *3*, 1234-1242.

[133] Deutsch, E. W., Lam, H., Aebersold, R., PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* 2008, *9*, 429-434.

[134] Mathivanan, S., Ahmed, M., Ahn, N. G., Alexandre, H.*, et al.*, Human Proteinpedia enables sharing of human protein data. *Nat Biotechnol* 2008, *26*, 164-167.

[135] Slotta, D. J., Barrett, T., Edgar, R., NCBI Peptidome: a new public repository for mass spectrometry peptide identifications. *Nat Biotechnol* 2009, *27*, 600-601.

[136] Falkner JA, A. P., Tranche: secure decentralized data storage for the proteomics community. *J Bio Tech 18: 3* 2007.

[137] Ovelleiro, D., Carrascal, M., Casas, V., Abian, J., LymPHOS: design of a phosphosite database of primary human T cells. *Proteomics* 2009, *9*, 3741-3751.

[138] Wong, Y. H., Lee, T. Y., Liang, H. K., Huang, C. M.*, et al.*, KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic acids research* 2007, *35*, W588-594.

[139] Finn, R. D., Tate, J., Mistry, J., Coggill, P. C.*, et al.*, The Pfam protein families database. *Nucleic acids research* 2008, *36*, D281-288.

[140] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M., The KEGG resource for deciphering the genome. *Nucleic acids research* 2004, *32*, D277-280.

[141] Thingholm, T. E., Jorgensen, T. J., Jensen, O. N., Larsen, M. R., Highly selective enrichment of phosphorylated peptides using titanium dioxide. *Nature protocols* 2006, *1*, 1929-1935.

[142] Marcantonio, M., Trost, M., Courcelles, M., Desjardins, M., Thibault, P., Combined enzymatic and data mining approaches for comprehensive phosphoproteome analyses: application to cell signaling events of interferon-gamma-stimulated macrophages. *Molecular & cellular proteomics : MCP* 2008, *7*, 645-660.

[143] Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y.*, et al.*, The International Protein Index: an integrated database for proteomics experiments. *Proteomics* 2004, *4*, 1985-1988.

[144] Pruitt, K. D., Tatusova, T., Klimke, W., Maglott, D. R., NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic acids research* 2009, *37*, D32-36.

[145] Wilkins, M. R., Gasteiger, E., Bairoch, A., Sanchez, J. C.*, et al.*, Protein identification and analysis tools in the ExPASy server. *Methods in molecular biology* 1999, *112*, 531-552.

[146] Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H., Locating proteins in the cell using TargetP, SignalP and related tools. *Nature protocols* 2007, *2*, 953-971.

[147] Sonnhammer, E. L., von Heijne, G., Krogh, A., A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* 1998, *6*, 175-182.

[148] Edgar, R. C., MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 2004, *32*, 1792-1797.

[149] Gong, W., Zhou, D., Ren, Y., Wang, Y.*, et al.*, PepCyber:P~PEP: a database of human protein protein interactions mediated by phosphoprotein-binding domains. *Nucleic acids research* 2008, *36*, D679-683.

[150] Cheng, J., Randall, A. Z., Sweredoski, M. J., Baldi, P., SCRATCH: a protein structure and structural feature prediction server. *Nucleic acids research* 2005, *33*, W72-76.

[151] Deng, X., Eickholt, J., Cheng, J., PreDisorder: ab initio sequence-based prediction of protein disordered regions. *BMC bioinformatics [electronic resource]* 2009, *10*, 436.

[152] Nair, R., Carter, P., Rost, B., NLSdb: database of nuclear localization signals. *Nucleic acids research* 2003, *31*, 397-399.

[153] Kosugi, S., Hasebe, M., Tomita, M., Yanagawa, H., Nuclear export signal consensus sequences defined using a localization-based yeast selection system. *Traffic* 2008, *9*, 2053-2062.

[154] Yachie, N., Saito, R., Sugahara, J., Tomita, M., Ishihama, Y., In silico analysis of phosphoproteome data suggests a rich-get-richer process of phosphosite accumulation over evolution. *Molecular & cellular proteomics : MCP* 2009, *8*, 1061-1071.

[155] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., Basic local alignment search tool. *J Mol Biol* 1990, *215*, 403-410.

[156] Smith, T. F., Waterman, M. S., Identification of common molecular subsequences. *J Mol Biol* 1981, *147*, 195-197.

[157] Pearson, W. R., Lipman, D. J., Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* 1988, *85*, 2444-2448.

[158] Berglund, A. C., Sjolund, E., Ostlund, G., Sonnhammer, E. L., InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic acids research* 2008, *36*, D263-266.

[159] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D.*, et al.*, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, *25*, 25-29.

[160] Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S.*, et al.*, STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research* 2009, *37*, D412-416.

[161] Hyland, E. M., Cosgrove, M. S., Molina, H., Wang, D.*, et al.*, Insights into the role of histone H3 and histone H4 core modifiable residues in Saccharomyces cerevisiae. *Mol Cell Biol* 2005, *25*, 10060-10070.

[162] Karin, M., The regulation of AP-1 activity by mitogen-activated protein kinases. *J Biol Chem* 1995, *270*, 16483-16486.

[163] Wang, C., Chua, K., Seghezzi, W., Lees, E.*, et al.*, Phosphorylation of spliceosomal protein SAP 155 coupled with splicing catalysis. *Genes Dev* 1998, *12*, 1409-1414.

[164] Liku, M. E., Nguyen, V. Q., Rosales, A. W., Irie, K., Li, J. J., CDK phosphorylation of a novel NLS-NES module distributed between two subunits of the Mcm2-7 complex prevents chromosomal rereplication. *Mol Biol Cell* 2005, *16*, 5026-5039.

[165] Kodiha, M., Tran, D., Morogan, A., Qian, C., Stochaj, U., Dissecting the signaling events that impact classical nuclear import and target nuclear transport factors. *PLoS One* 2009, *4*, e8420.

[166] Obsilova, V., Silhan, J., Boura, E., Teisinger, J., Obsil, T., 14-3-3 proteins: a family of versatile molecular regulators. *Physiol Res* 2008, *57 Suppl 3*, S11-21.

[167] Hart, G. W., Greis, K. D., Dong, L. Y., Blomberg, M. A.*, et al.*, O-linked N-acetylglucosamine: the "yin-yang" of Ser/Thr phosphorylation? Nuclear and cytoplasmic glycosylation. *Adv Exp Med Biol* 1995, *376*, 115-123.

[168] Rechsteiner, M., Rogers, S. W., PEST sequences and regulation by proteolysis. *Trends Biochem Sci* 1996, *21*, 267-271.

[169] Nandi, A., Sprung, R., Barma, D. K., Zhao, Y.*, et al.*, Global identification of O-GlcNAc-modified proteins. *Anal Chem* 2006, *78*, 452-458.

[170] Klein, D. J., Moore, P. B., Steitz, T. A., The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J Mol Biol* 2004, *340*, 141-177.

[171] Averna, M., de Tullio, R., Passalacqua, M., Salamino, F.*, et al.*, Changes in intracellular calpastatin localization are mediated by reversible phosphorylation. *Biochem J* 2001, *354*, 25-30.

[172] Ba, A. N., Moses, A. M., Evolution of characterized phosphorylation sites in budding yeast. *Mol Biol Evol* 2010, *27*, 2027-2037.

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

[174] Boulais, J., Trost, M., Landry, C. R., Dieckmann, R.*, et al.*, Molecular characterization of the evolution of phagosomes. *Mol Syst Biol* 2010, *6*, 423.

[175] Warn-Cramer, B. J., Cottrell, G. T., Burt, J. M., Lau, A. F., Regulation of connexin-43 gap junctional intercellular communication by mitogen-activated protein kinase. *J Biol Chem* 1998, *273*, 9188-9196.

[176] Cooper, C. D., Lampe, P. D., Casein kinase 1 regulates connexin-43 gap junction assembly. *J Biol Chem* 2002, *277*, 44962-44968.

[177] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 2002, *74*, 5383-5392.

[178] Meloche, S., Mitogen-activated protein kinases. *In Encyclopedia of Signaling Molecules. S. Choi, editor. New York: Springer. in press.* 2011.

[179] Pearson, G., Robinson, F., Beers Gibson, T., Xu, B. E.*, et al.*, Mitogen-activated protein (MAP) kinase pathways: regulation and physiological functions. *Endocr Rev* 2001, *22*, 153-183.

[180] Fremin, C., Meloche, S., From basic research to clinical development of MEK1/2 inhibitors for cancer therapy. *Journal of hematology & oncology* 2010, *3*, 8.

[181] Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B.*, et al.*, Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & cellular proteomics : MCP* 2002, *1*, 376-386.

[182] Trost, M., English, L., Lemieux, S., Courcelles, M.*, et al.*, The phagosomal proteome in interferon-gamma-activated macrophages. *Immunity* 2009, *30*, 143-154.

[183] Blagoev, B., Ong, S. E., Kratchmarova, I., Mann, M., Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nat Biotechnol* 2004, *22*, 1139-1145.

[184] Wolf-Yadlin, A., Kumar, N., Zhang, Y., Hautaniemi, S.*, et al.*, Effects of HER2 overexpression on cell signaling networks governing proliferation and migration. *Mol Syst Biol* 2006, *2*, 54.

[185] Pan, C., Gnad, F., Olsen, J. V., Mann, M., Quantitative phosphoproteome analysis of a mouse liver cell line reveals specificity of phosphatase inhibitors. *Proteomics* 2008, *8*, 4534-4546.

[186] Delaney, A. M., Printen, J. A., Chen, H., Fauman, E. B., Dudley, D. T., Identification of a novel mitogen-activated protein kinase kinase activation domain recognized by the inhibitor PD 184352. *Mol Cell Biol* 2002, *22*, 7593-7602.

[187] McCubrey, J. A., Steelman, L. S., Abrams, S. L., Lee, J. T.*, et al.*, Roles of the RAF/MEK/ERK and PI3K/PTEN/AKT pathways in malignant transformation and drug resistance. *Advances in enzyme regulation* 2006, *46*, 249-279.

[188] Dubois, T., Rommel, C., Howell, S., Steinhussen, U.*, et al.*, 14-3-3 is phosphorylated by casein kinase I on residue 233. Phosphorylation at this site in vivo regulates Raf/14-3-3 interaction. *J Biol Chem* 1997, *272*, 28882-28888.

[189] Sheridan, D. L., Kong, Y., Parker, S. A., Dalby, K. N., Turk, B. E., Substrate discrimination among mitogen-activated protein kinases through distinct docking sequence motifs. *J Biol Chem* 2008, *283*, 19511-19520.

[190] Fernandes, N., Allbritton, N. L., Effect of the DEF motif on phosphorylation of peptide substrates by ERK. *Biochemical and biophysical research communications* 2009, *387*, 414-418.

[191] Kosako, H., Yamaguchi, N., Aranami, C., Ushiyama, M.*, et al.*, Phosphoproteomics reveals new ERK MAP kinase targets and links ERK to nucleoporin-mediated nuclear transport. *Nat Struct Mol Biol* 2009.

[192] Pan, C., Olsen, J. V., Daub, H., Mann, M., Global effects of kinase inhibitors on signaling networks revealed by quantitative phosphoproteomics. *Molecular & cellular proteomics : MCP* 2009, *8*, 2796-2808.

[193] Yeakley, J. M., Tronchere, H., Olesen, J., Dyck, J. A.*, et al.*, Phosphorylation regulates in vivo interaction and molecular targeting of serine/arginine-rich pre-mRNA splicing factors. *J Cell Biol* 1999, *145*, 447-455.

[194] Kamakura, S., Moriguchi, T., Nishida, E., Activation of the protein kinase ERK5/BMK1 by receptor tyrosine kinases. Identification and characterization of a signaling pathway to the nucleus. *J Biol Chem* 1999, *274*, 26563-26571.

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

[196] Nardozzi, J. D., Lott, K., Cingolani, G., Phosphorylation meets nuclear import: a review. *Cell communication and signaling : CCS* 2010, *8*, 32.

[197] Bianchi, M. E., Agresti, A., HMG proteins: dynamic players in gene regulation and differentiation. *Current opinion in genetics & development* 2005, *15*, 496-506.

[198] Ruttenberg, B. E., Pisitkun, T., Knepper, M. A., Hoffert, J. D., PhosphoScore: an open-source phosphorylation site assignment tool for MSn data. *J Proteome Res* 2008, *7*, 3054-3059.

[199] Otvos, L., Jr., Tangoren, I. A., Wroblewski, K., Hollosi, M., Lee, V. M., Reversed-phase high-performance liquid chromatographic separation of synthetic phosphopeptide isomers. *J Chromatogr* 1990, *512*, 265-272.

[200] Hoffmann, R., Segal, M., Otvos, L., Separation of sets of mono- and diphosphorylated peptides by reversed-phase high performance liquid chromatography. *Analytica Chimica Acta* 1997, *352*, 327-333.

[201] Gruhler, A., Olsen, J. V., Mohammed, S., Mortensen, P.*, et al.*, Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Molecular & cellular proteomics : MCP* 2005, *4*, 310-327.

[202] Ballif, B. A., Roux, P. P., Gerber, S. A., MacKeigan, J. P.*, et al.*, Quantitative phosphorylation profiling of the ERK/p90 ribosomal S6 kinase-signaling cassette and its targets, the tuberous sclerosis tumor suppressors. *Proc Natl Acad Sci U S A* 2005, *102*, 667-672.

[203] Boeri Erba, E., Matthiesen, R., Bunkenborg, J., Schulze, W. X.*, et al.*, Quantitation of multisite EGF receptor phosphorylation using mass spectrometry and a novel normalization approach. *J Proteome Res* 2007, *6*, 2768-2785.

[204] Cunningham, D. L., Sweet, S. M., Cooper, H. J., Heath, J. K., Differential phosphoproteomics of fibroblast growth factor signaling: identification of Src family kinase-mediated phosphorylation events. *J Proteome Res* 2010, *9*, 2317-2328.

[205] Sweet, S. M., Mardakheh, F. K., Ryan, K. J., Langton, A. J.*, et al.*, Targeted online liquid chromatography electron capture dissociation mass spectrometry for the localization of sites of in vivo phosphorylation in human Sprouty2. *Anal Chem* 2008, *80*, 6650-6657.

[206] Bridon, G., Bonneil, E., Muratore-Schroeder, T., Caron-Lizotte, O., Thibault, P., Improvement of phosphoproteome analyses using FAIMS and decision tree fragmentation; Application to the insulin signalling pathway in Drosophila melanogaster S2 cells. *J Proteome Res* 2011.

[207] Monroe, M., Pacific Northwest National Laboratory, Richland, WA 2011.

[208] Maupetit, J., Derreumaux, P., Tuffery, P., PEP-FOLD: an online resource for de novo peptide structure prediction. *Nucleic acids research* 2009, *37*, W498-503.

[209] Kyte, J., Doolittle, R. F., A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982, *157*, 105-132.

[210] Winter, D., Pipkorn, R., Lehmann, W. D., Separation of peptide isomers and conformers by ultra performance liquid chromatography. *J Sep Sci* 2009, *32*, 1111-1119.

[211] Kelstrup, C. D., Hekmat, O., Francavilla, C., Olsen, J. V., Pinpointing phosphorylation sites: Quantitative filtering and a novel site-specific x-ion fragment. *J Proteome Res* 2011, *10*, 2937-2948.

[212] Pan, S., Chen, R., Aebersold, R., Brentnall, T. A., Mass spectrometry based glycoproteomics--from a proteomics perspective. *Molecular & cellular proteomics : MCP* 2011, *10*, R110 003251.

[213] Freschi, L., Courcelles, M., Thibault, P., Michnick, S. W., Landry, C. R., Phosphorylation network rewiring by gene duplication. *Mol Syst Biol* 2011, *7*, 504.

[214] Huttlin, E. L., Jedrychowski, M. P., Elias, J. E., Goswami, T.*, et al.*, A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* 2010, *143*, 1174-1189.

[215] Kim, M. S., Zhong, J., Kandasamy, K., Delanghe, B., Pandey, A., Systematic evaluation of alternating CID and ETD fragmentation for phosphorylated peptides. *Proteomics* 2011, *11*, 2568-2572.

Annexe 1 Protéines et acides aminés

Les protéines sont des biopolymères qui assument un rôle prédominant dans l'accomplissement des divers processus biologiques. Ils assurent aussi un soutient structurel pour la cellule. Les protéines sont issues des gènes encodés dans le génome qui est constitué d'ADN. Les gènes sont transcrits en ARNm qui sert de guide pour la synthèse des protéines (la traduction) par les ribosomes. Les protéines sont constituées d'une sélection de 20 acides aminés (Figure A1.1) qui diffèrent par leur chaîne latérale au niveau du carbone α. Les acides aminés sont attachés les uns aux autres par la formation d'un lien covalent, le lien peptidique (Figure A1.2), et forment une chaîne linéaire. Une fois liés entre eux, les acides aminés sont nommés résidus. La taille de ce polymère peut varier de quelques acides aminés à plusieurs milliers. Une protéine de moins de 50 résidus est appelée peptide. L'ordre d'assemblage des différents acides aminés influence l'activité et le repliement de la structure protéique. La structure des protéines est définie selon quatre niveaux (Figure A1.2). La structure primaire correspond à la séquence ordonnée d'acides aminés de l'extrémité N-terminale à l'extrémité C-terminale. La structure secondaire correspond à de simples structures tridimensionnelles comme l'hélice α et le feuillet β qui sont stabilisées par des liaisons hydrogènes. Les régions non-ordonnées entre ces structures sont reliées par des boucles flexibles. La structure tertiaire est la structure tridimensionnelle totale de la protéine. Des ponts disulfures peuvent se créer entre les cystéines pour stabiliser la structure. La structure quaternaire correspond à l'assemblage de façon non-covalente de deux ou plusieurs chaînes polypeptidiques qui peuvent être identiques ou différentes. Ce type d'assemblage est appelé complexe protéique. La structure d'une protéine peut être modifiée par des réactions chimiques ou l'activité enzymatique de d'autres protéines. Des molécules de petite ou de très grande taille peuvent être conjuguées aux divers acides aminés des protéines. Ces modifications affectent les propriétés physico-chimiques des protéines ainsi que leur activité. La structure des protéines n'est pas complètement rigide et peut adopter différentes conformations pour interagir avec diverses molécules et protéines. Ces diverses conformations sont essentielles pour l'activité des protéines. Les protéines

sont composées de domaines qui ont divers rôles fonctionnels très spécifiques: activité catalytique, reconnaissance du substrat, interaction protéine-protéine ou protéine-ADN, etc.

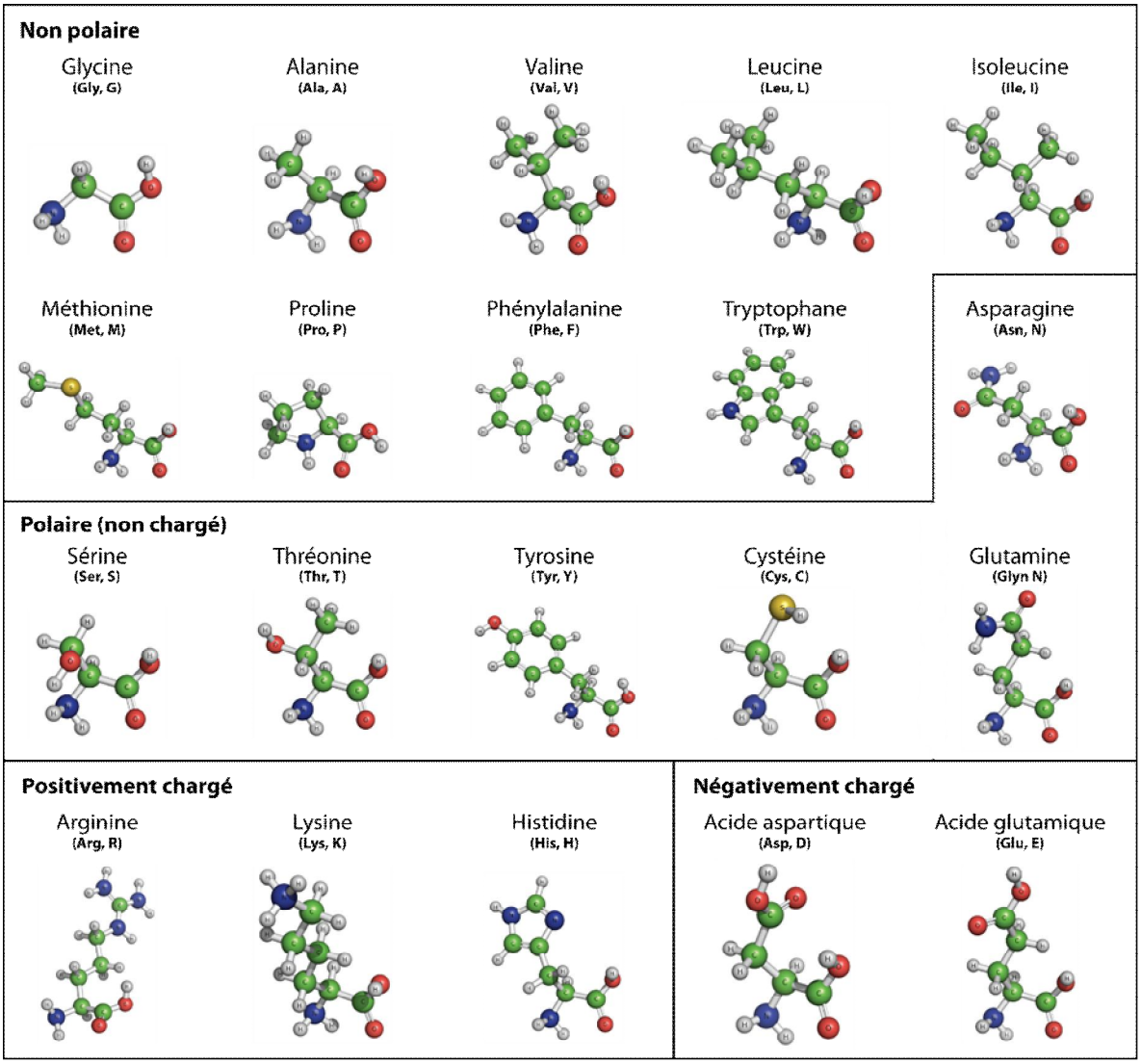


Figure A1.1: Structures tridimensionnelles des 20 acides aminés standards.

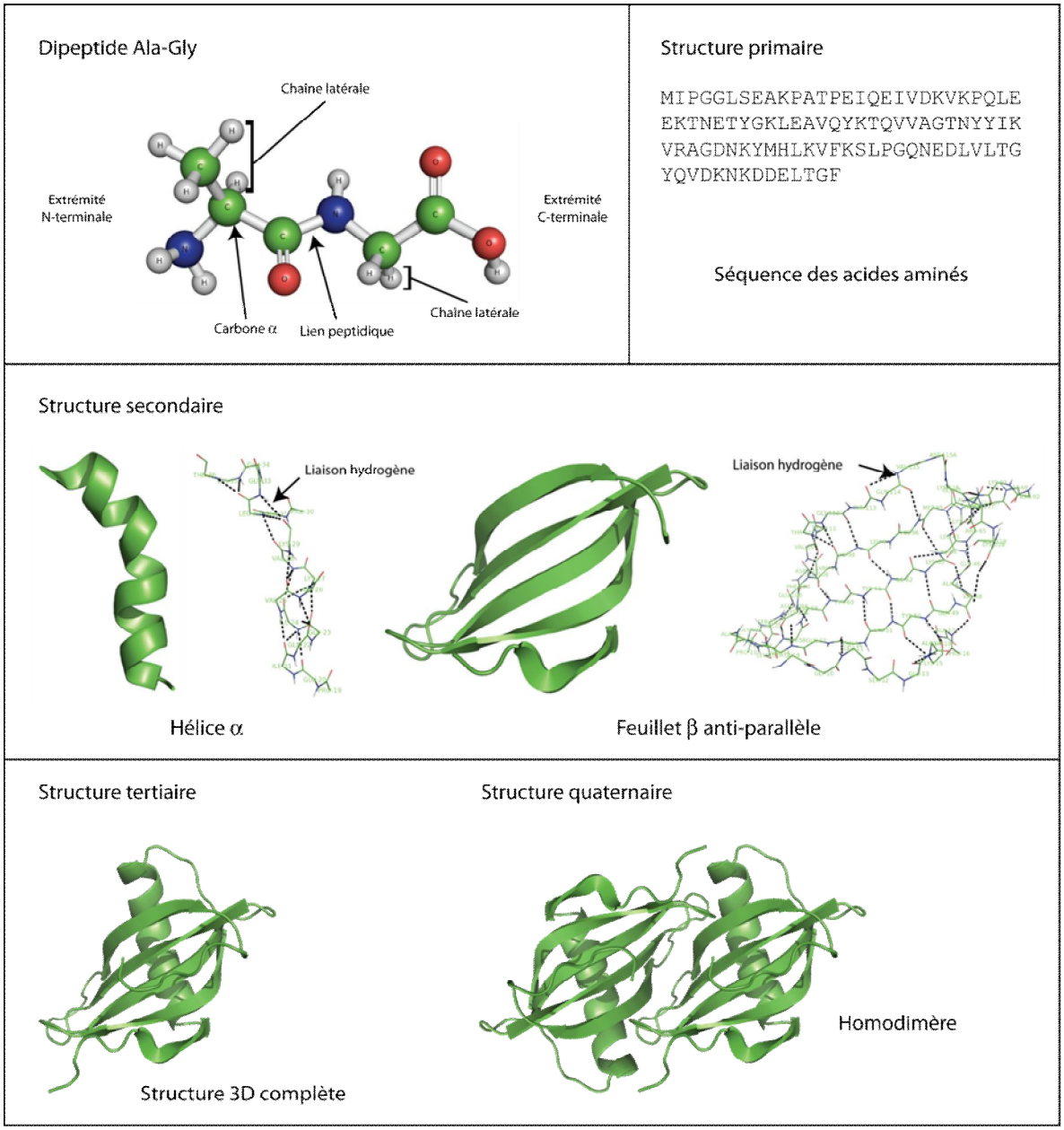


Figure A1.2: Niveaux d'organisation des structures des protéines.

Annexe 2 Figures et tableaux supplémentaires du chapitre 2

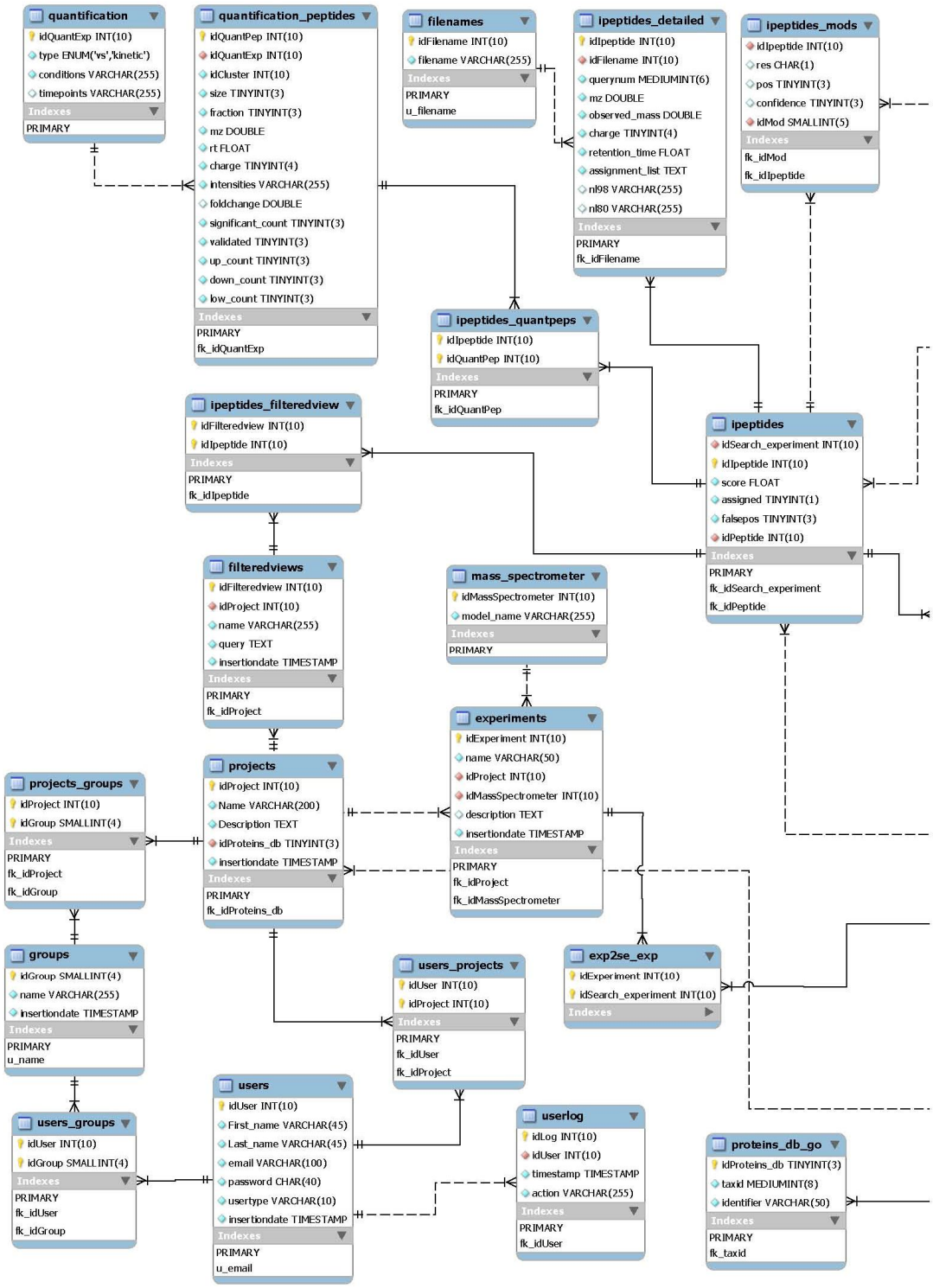


Figure A2.1: Entity-relationship schema of ProteoConnections database.

This schema represents the 37 tables that hold data for the analysis platform.

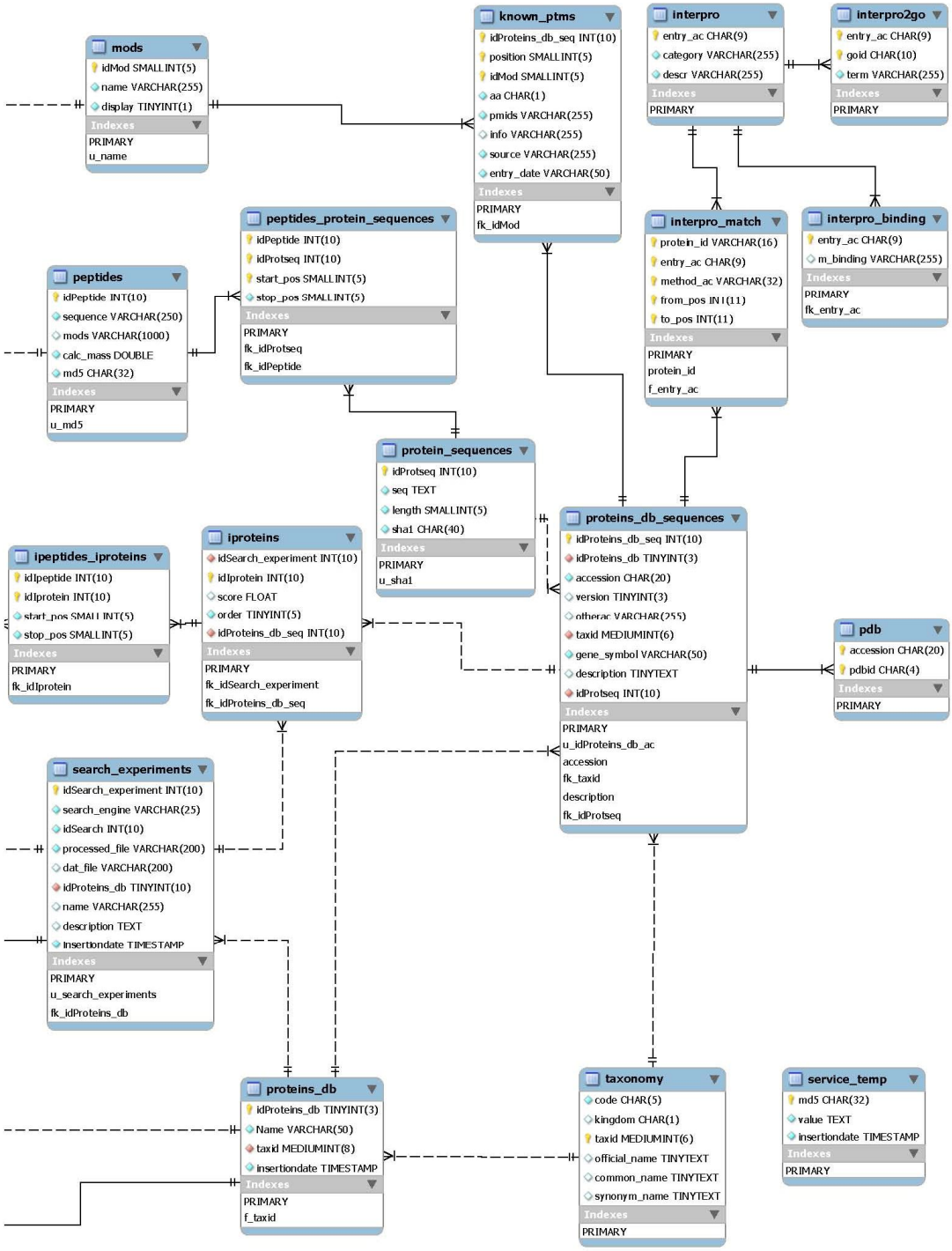

mods
idMod SMALLINT(5)
name VARCHAR(255)
display TINYINT(1)
Indexes
PRIMARY
u_name
known_ptms
idProteins_db_seq INT(10)
position SMALLINT(5)
idMod SMALLINT(5)
aa CHAR(1)
pmids VARCHAR(255)
info VARCHAR(255)
source VARCHAR(255)
entry_date VARCHAR(50)
Indexes
PRIMARY
fk_idMod
interpro
entry_ac CHAR(9)
category VARCHAR(255)
descr VARCHAR(255)
Indexes
PRIMARY
interpro2go
entry_ac CHAR(9)
goid CHAR(10)
term VARCHAR(255)
Indexes
PRIMARY
peptides_protein_sequences
idPeptide INT(10)
idProtseq INT(10)
start_pos SMALLINT(5)
stop_pos SMALLINT(5)
Indexes
PRIMARY
fk_idProtseq
fk_idPeptide
interpro_match
protein_id VARCHAR(16)
entry_ac CHAR(9)
method_ac VARCHAR(32)
from_pos INT(11)
to_pos INT(11)
Indexes
PRIMARY
protein_id
f_entry_ac
interpro_binding
entry_ac CHAR(9)
m_binding VARCHAR(255)
Indexes
PRIMARY
fk_entry_ac
peptides
idPeptide INT(10)
sequence VARCHAR(250)
mods VARCHAR(1000)
calc_mass DOUBLE
md5 CHAR(32)
Indexes
PRIMARY
u_md5
protein_sequences
idProtseq INT(10)
seq TEXT
length SMALLINT(5)
sha1 CHAR(40)
Indexes
PRIMARY
u_sha1
proteins_db_sequences
idProteins_db_seq INT(10)
idProteins_db TINYINT(3)
accession CHAR(20)
version TINYINT(3)
otheracVARCHAR(255)
taxid MEDIUMINT(6)
gene_symbol VARCHAR(50)
description TINYTEXT
idProtseq INT(10)
Indexes
PRIMARY
u_idProteins_db_ac
accession
fk_taxid
description
fk_idProtseq
ipeptides_iproteins
idIpeptide INT(10)
idIprotein INT(10)
start_pos SMALLINT(5)
stop_pos SMALLINT(5)
Indexes
PRIMARY
fk_idIprotein
iproteins
idSearch_experiment INT(10)
idIprotein INT(10)
score FLOAT
order TINYINT(5)
idProteins_db_seq INT(10)
Indexes
PRIMARY
fk_idSearch_experiment
fk_idProteins_db_seq
pdb
accession CHAR(20)
pdbid CHAR(4)
Indexes
PRIMARY
search_experiments
idSearch_experiment INT(10)
search_engine VARCHAR(25)
idSearch INT(10)
processed_file VARCHAR(200)
dat_file VARCHAR(200)
idProteins_db TINYINT(10)
name VARCHAR(255)
description TEXT
insertiondate TIMESTAMP
Indexes
PRIMARY
u_search_experiments
fk_idProteins_db
proteins_db
idProteins_db TINYINT(3)
Name VARCHAR(50)
taxid MEDIUMINT(8)
insertiondate TIMESTAMP
Indexes
PRIMARY
f_taxid
taxonomy
code CHAR(5)
kingdom CHAR(1)
taxid MEDIUMINT(6)
official_name TINYTEXT
common_name TINYTEXT
synonym_name TINYTEXT
Indexes
PRIMARY
service_temp
md5 CHAR(32)
value TEXT
insertiondate TIMESTAMP
Indexes
PRIMARY

Table A2.I: Programming interfaces for bioinformatics resources integrated to ProteoConnections.

Bioinformatics tool or database	Ressource description	Interface	Url
ACCpro	Residues accessiblity	Local install	
Biomart	Protein domains (Interpro), PDB	Web service	http://www.biomart.org/
EBI IPI	Protein sequences, annotations	Database*	http://www.ebi.ac.uk/IPI
EBI PICR	Accession conversion, PDB	Web service	http://www.ebi.ac.uk/Tools/picr/
HPRD	Phosphorylation binding, kinases & phosphatase motifs	Database	http://www.hprd.org
MoDL	Phosphorylation motif discovery	Data formated for manual use.	http://cs.brown.edu/people/braphael/software.html
Motif-X	Phosphorylation motif discovery	Data formated for manual use.	http:/motif-x.med.harvard.edu
Muscle	Multiple sequences aligment	Local install	http:/www.drive5.com/muscle/
NetOglyc-3.1d	Glycosylation predictor	Local install	http://www.cbs.dtu.dk/services/NetOGlyc/
NetworKIN	Phosphorylation predictor	Web service	http://networkin.info
NLSdb	Nuclear localization signal	Database	http://rostlab.org/services/nlsdb/
OGlcNAcScan	Glycosylation predictor	Web service	http://cbsb.lombardi.georgetown.edu/hulab/OGAP.html
PepCyber	Phosphorylation binding, kinases & phosphatase motifs	Database	http://pepcyber.biolead.org/PPEP/
Phospho.ELM	Post-translastional modifications	Database*	http://phospho.elm.eu.org/
PhosphoSitePlus	Post-translastional modifications	Database*	http://www.phosphosite.org
Predisorder1.0	Disordered region predictor	Local install	http://casp.rnet.missouri.edu/download/predisorder1.0.tar.gz
PPSP	Phosphorylation predictor	Local install	http://ppsp.biocuckoo.org/
ProtParam	Protein physiochemical	Web service	http://ca.expasy.org/tools/protparam.html
SGD	Protein sequences, annotations	Database*	http://www.yeastgenome.org/
SignalP 3	Signal peptide predictor	Local install	http://www.cbs.dtu.dk/services/SignalP/
SSpro4	Secondary structure predictor	Local install	http://download.igb.uci.edu/sspro4.html
STRING	Protein-protein interaction network	Database*	http://string-db.org/
TMHMM2.0c	Transmembrane helice predictor	Local install	http://www.cbs.dtu.dk/services/TMHMM/
Uniprot	Protein sequences, ptms, annotations	Database*	http://www.uniprot.org/
YinOyang-1.2	Glycosylation predictor	Local install	http://www.cbs.dtu.dk/services/YinOYang/

*Database are imported into ProteoConnections database and can be updated automaticaly or semi-automaticaly via the administrator interface.
Local install: Software must be installed by the user on the server.

Table A2.II: Identified phosphorylation sites list.

List of identified phosphorylation sites in a EXCEL file (Supplemental file on CD-ROM).

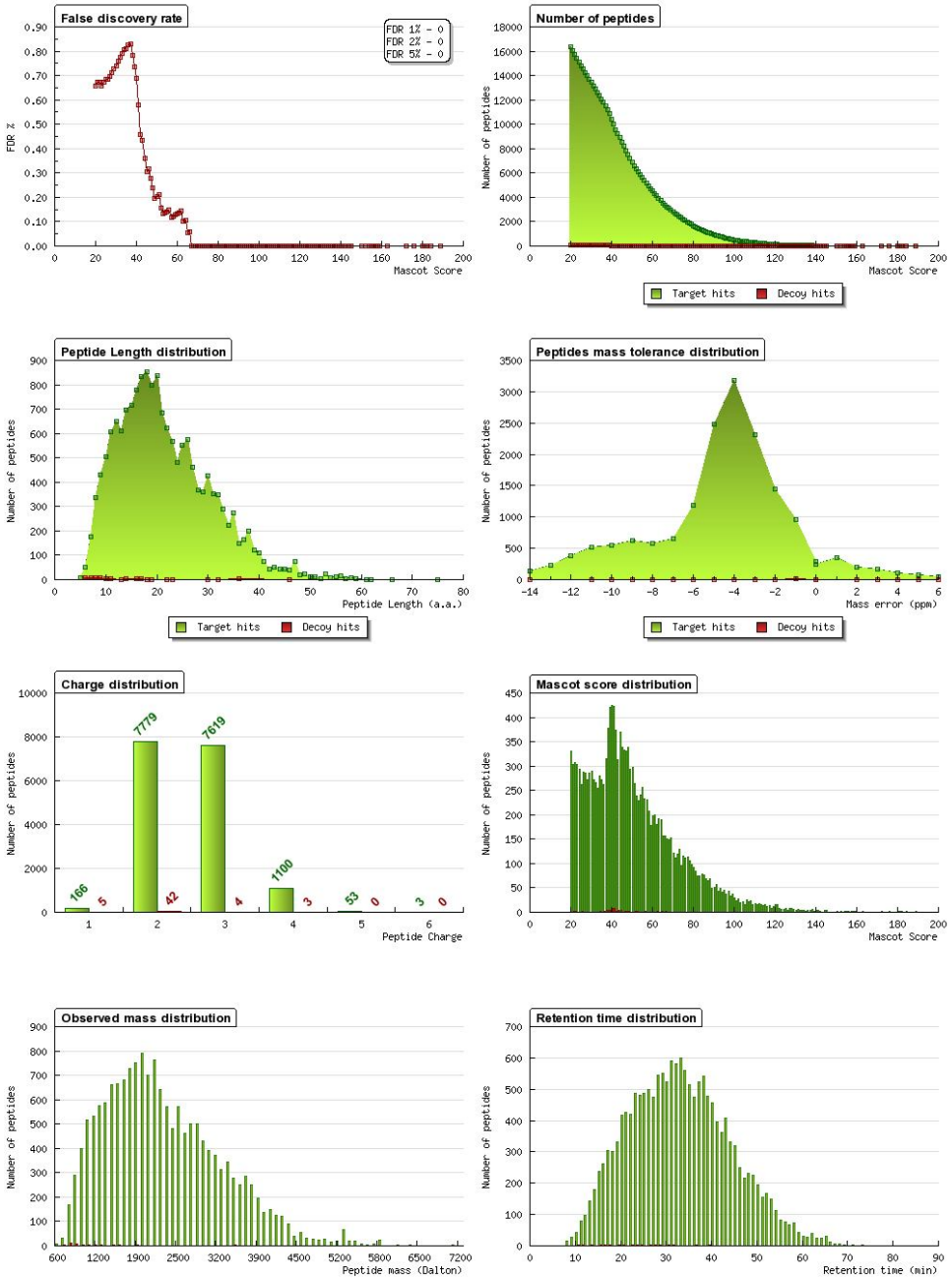


Figure A2.2: Graphical overview of the identified phosphopeptides population.

This figure has been generated by ProteoConnections using uniquely identified phosphopeptides. MS/MS spectra were searched with Mascot v2.2 using IPI rat databases v3.54. False discovery rate below 1% was obtained by applying the following filtering criteria: ± 10 ppm peptide mass precursor, peptides assigned to proteins with a $p < 0.05$ significance threshold.

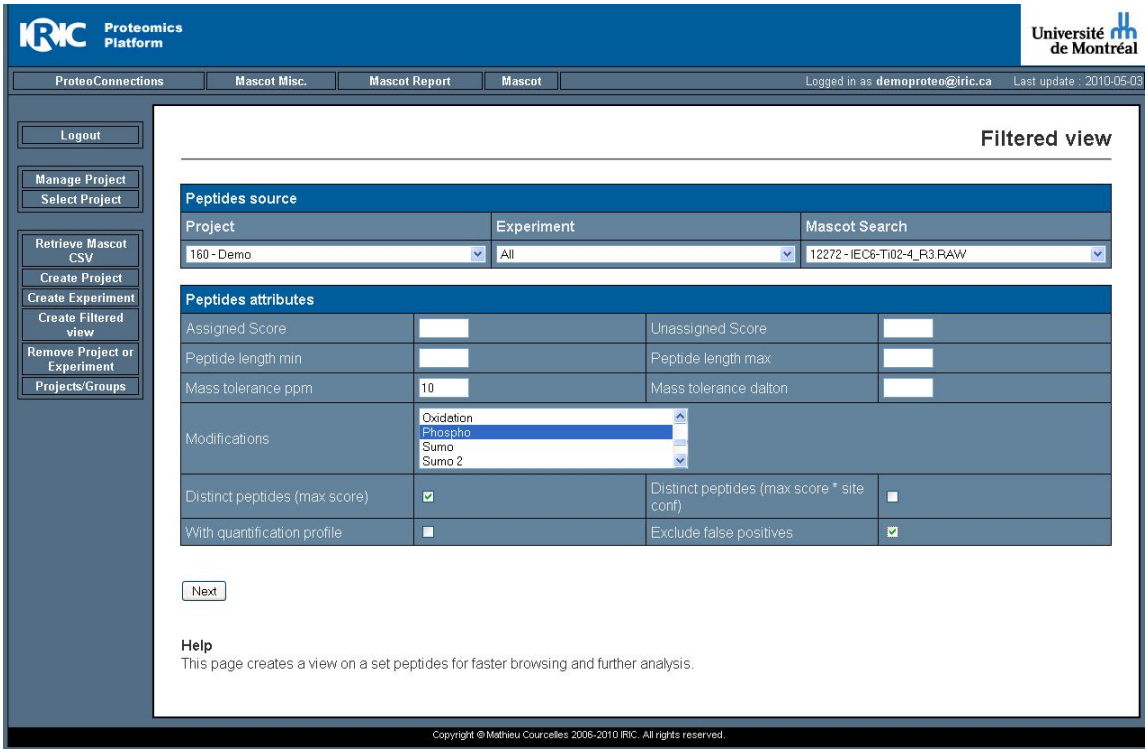


Figure A2.3: Filtered view screenshot.

Identified peptides can be filtered by mascot score, length, mass tolerance and modifications to create the restraint view needed by the user. Redundancy can also be removed to accelerate the browsing of the data.

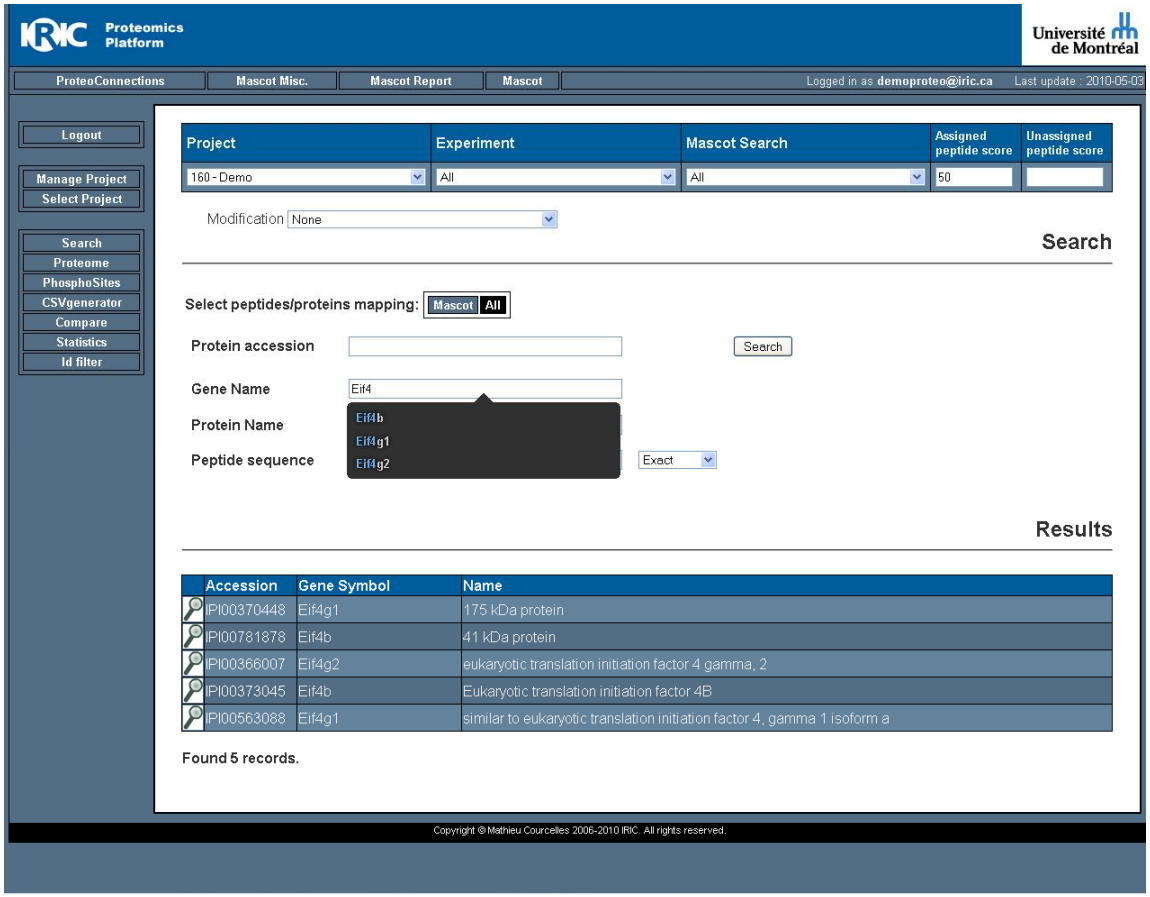


Figure A2.4: Search view screenshot.

For quicker access to data, user can search in the dataset by the protein identifier, gene name, protein name and exact or partial peptide sequence.

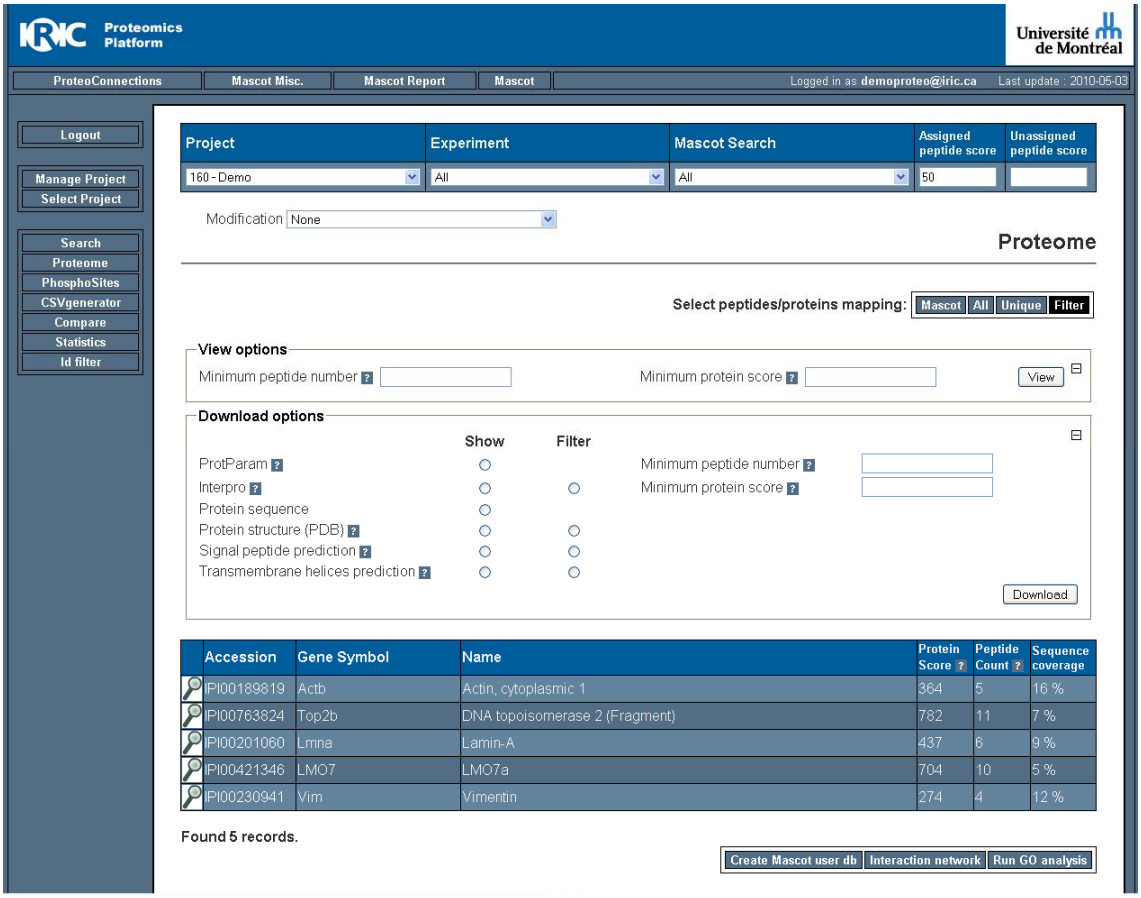


Figure A2.5: Proteome view screenshot.

This view displays the list of identified proteins with details on protein score, peptide count and sequence coverage for each protein. Various annotations can be added to the protein lists in the download mode.

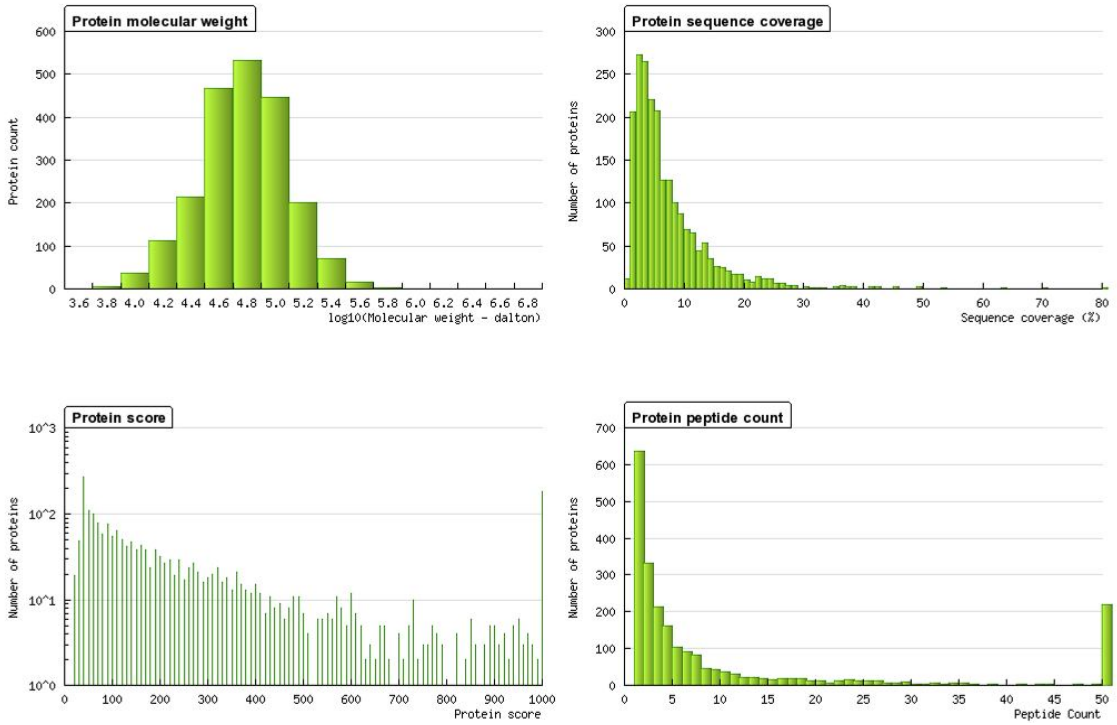


Figure A2.6: Graphical overview of the identified protein population.

This figure has been generated by ProteoConnections using identified phosphopeptides from rat samples. MS/MS spectra were searched with Mascot v2.2 using IPI rat database v3.54. Peptides were filtered with a ±10 ppm mass tolerance and a Mascot score for a FDR < 1% using a target-reverse database.

XXX

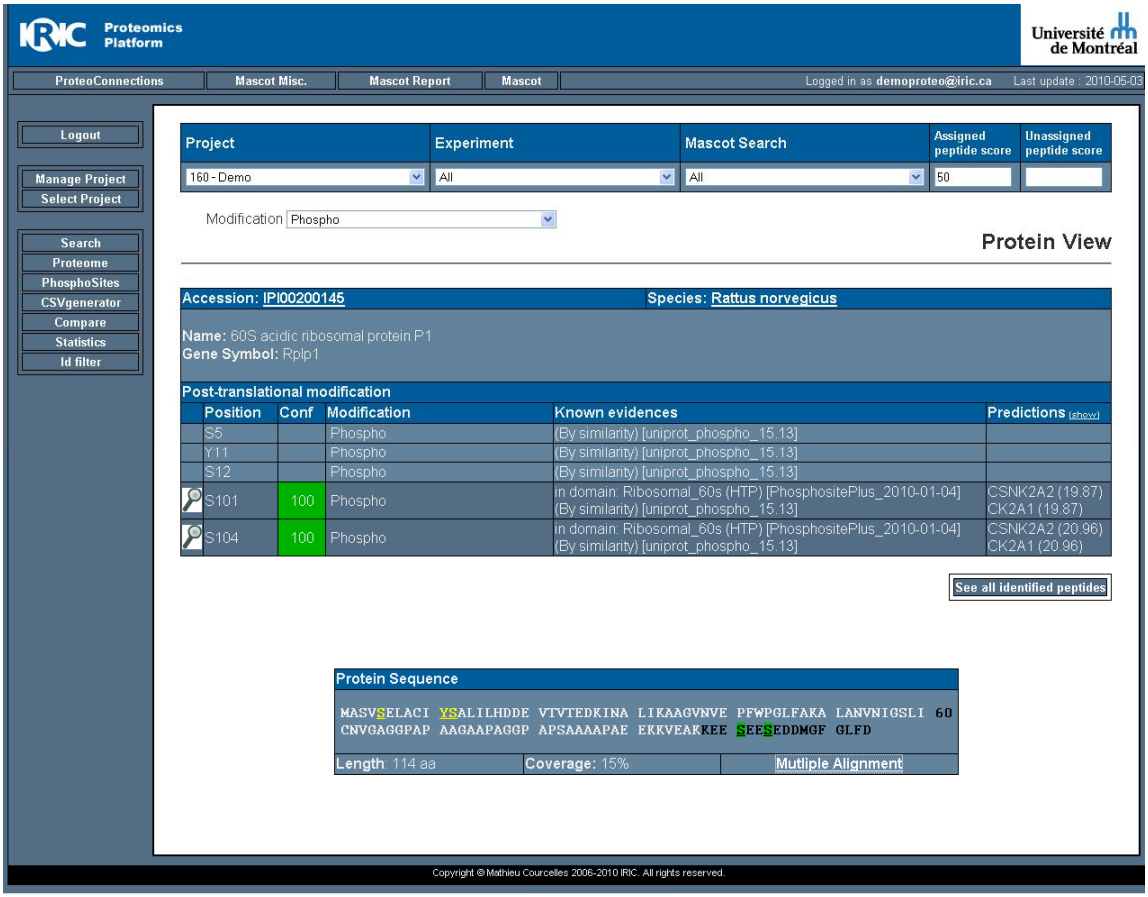


Figure A2.7: Protein view screenshot.

The protein view reports the identified peptides for a specific protein. Emphasis has been put on the presentation of identified post-translational modifications. Localization confidence, known modification and kinase predictions (in the case of phosphorylated residue) are reported. Multiple sequences alignment can be generated by one click to verify if the modified site is conserved in other species.

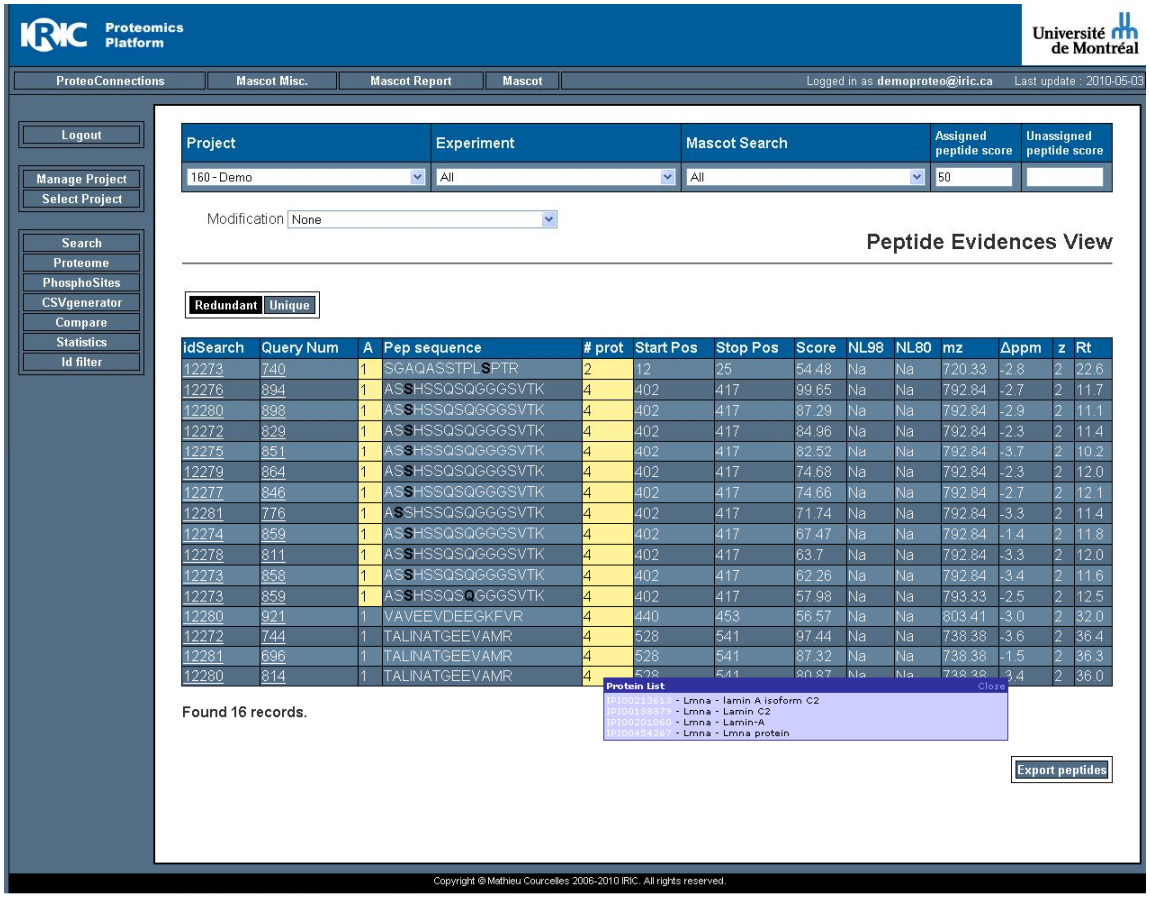


Figure A2.8: Peptide evidences view screenshot.

This view displays the attributes of the identified peptides for a specific protein. Ambiguity in peptide sequence or modification assignment is highlighted. Protein entries in the database sharing a peptide are reported.

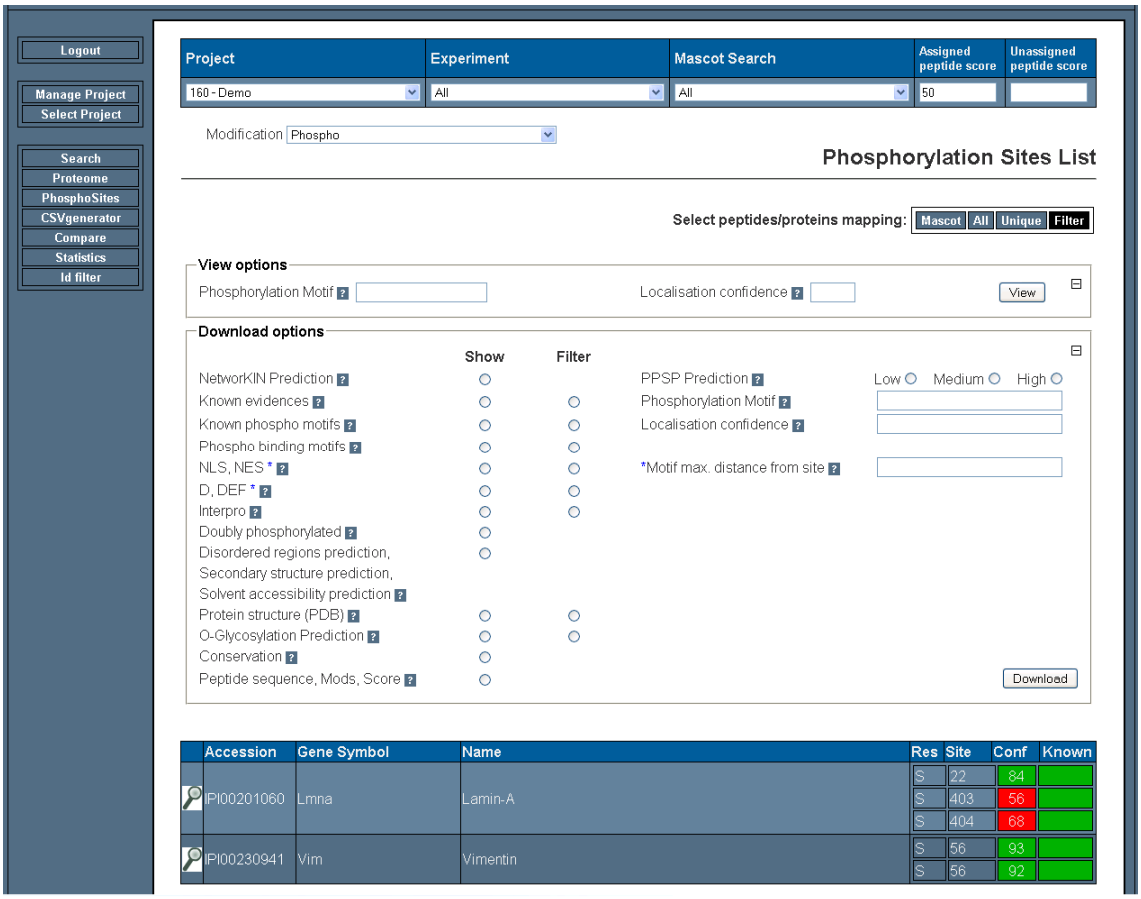


Figure A2.9: Phosphorylation sites list view screenshot.

This view generates a non-redundant phosphorylation sites list from the identified peptides. This list can be constrained to different phosphorylation motif, localization confidence, site located in domains, structure or predicted to be glycosylated.

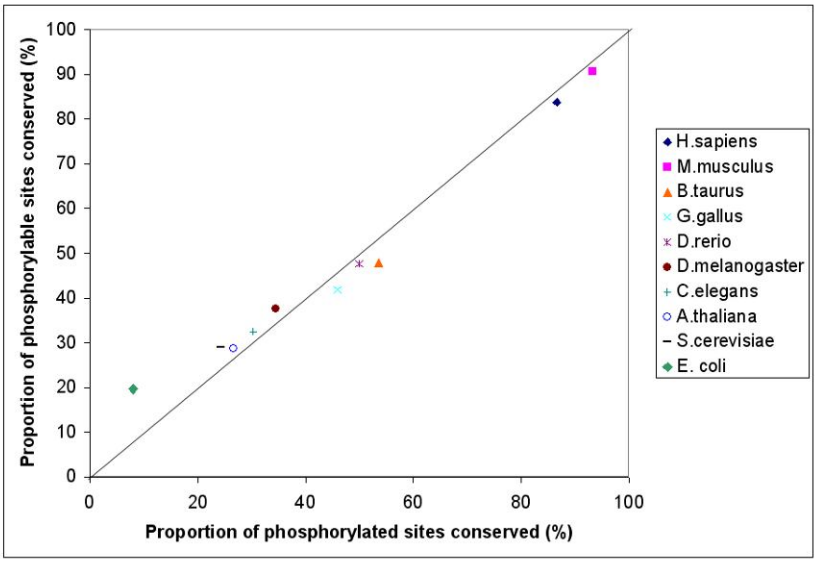


Figure A2.10: Conservation of phosphorylation sites between rat and other species.

This graph shows a higher conservation of phosphorylated sites in close species.

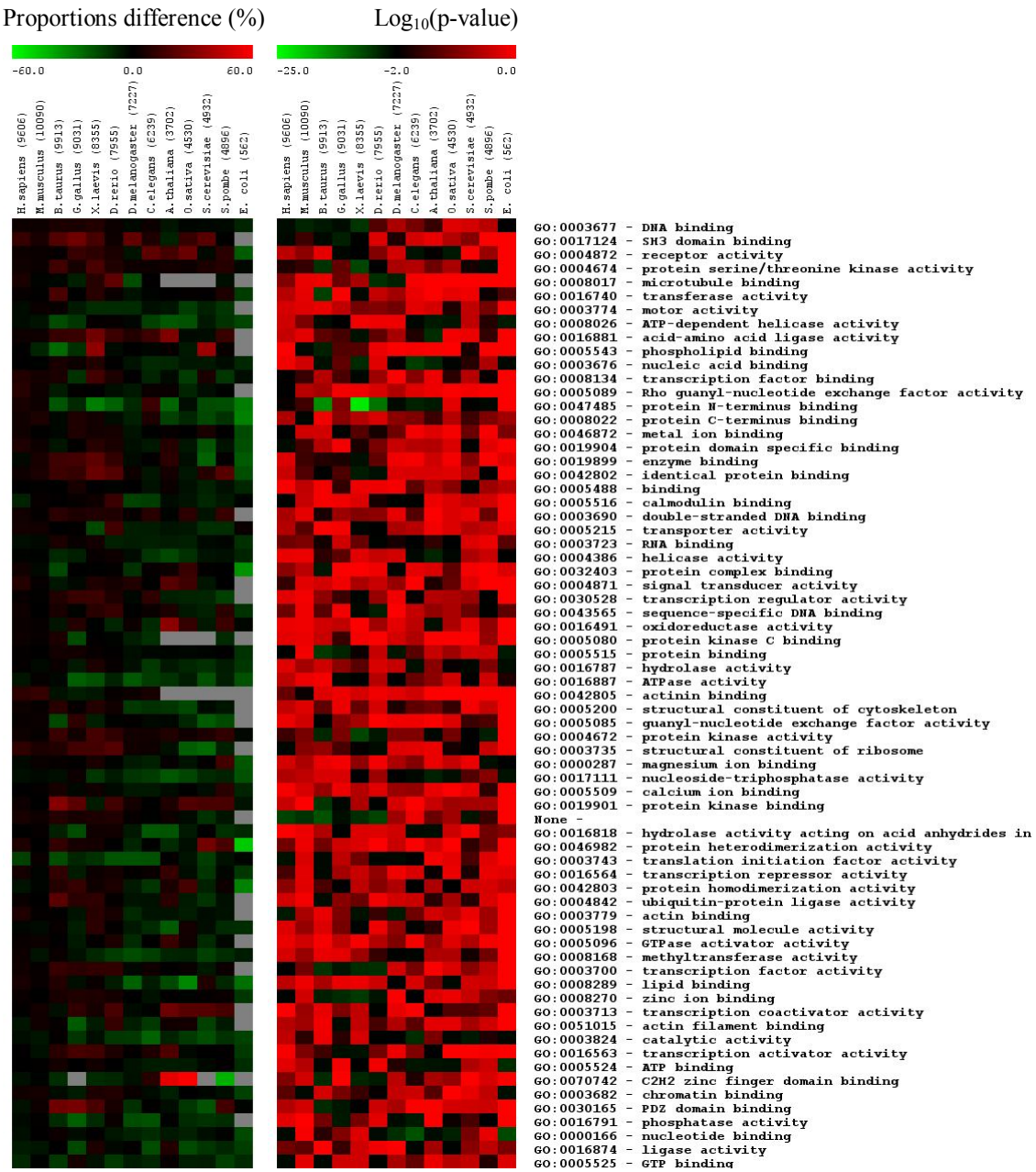


Figure A2.11: Rat phosphorylation sites conservation versus various species by molecular function.

This figure shows that no molecular function category has phosphorylation sites that are systematically and significantly more conserved. Heat maps showing the difference of the proportion of conserved phosphorylation sites and the proportion of conserved phosphorylable sites (left) and $\log_{10}$(p-value) obtained with Fisher exact test (right) for the same GO functional annotation.

Figure A2.12: Protein interactions potentially mediated by phosphorylation.

Dark blue nodes are proteins with phospho-binding domain and light blue nodes are identified phosphoproteins. Dark blue edges are interactions potentially mediated by phosphorylation (i.e. the phosphorylated protein has a phosphorylation site with a motif recognized by the phospho-binding domain) (Supplemental file on CD-ROM).

Figure A2.13: Kinases and phosphatases protein interactions with potential substrates.

Red nodes are proteins kinases, green are phosphatases and grey nodes are identified phosphoproteins. Red edges are potential interactions between a kinase and its substrates (i.e. the phosphorylated protein has a phosphorylation site with a motif recognized by the kinase) (Supplemental file on CD-ROM).

Log odds ratio & Fisher exact test

This example demonstrates how log odds ratio and Fisher exact test were calculated in Figure 3B using the library provided in R v2.6.2.

	Phosphorylated	**Phosphorylable***	**Total**
In domain	n_{11}	n_{10}	N_{1t}
Not in domain	n_{01}	n_{00}	N_{0t}
Total	N_{t1}	N_{t0}	N_{tt}

*but not phosphorylated

$$\text{Log odds ratio} = \log \frac{n_{11}/n_{01}}{n_{10}/n_{00}}$$

Équation 5

Fisher exact test calculates the table configuration probability using the hypergeometric distribution with this formula:

$$P = \frac{N_{1t}!N_{0t}!N_{t1}!N_{t0}!}{N_{tt}!\, n_{11}!n_{10}!n_{01}!n_{00}!}$$

Équation 6

Using fixed marginal totals (N_{xx}), the p-value is equal to the sum to of the table configuration probabilities that are less or equals then the observed table configuration probability.

Details of the algorithm implemented in R can be found in this publication:

Cyrus R. Mehta & Nitin R. Patel (1986). Algorithm 643. FEXACT: A Fortran subroutine for Fisher's exact test on unordered *r**c* contingency tables. *ACM Transactions on Mathematical Software*, **12**, 154–161.

Annexe 3 Figures et tableaux supplémentaires du chapitre 3

Supplemental methods

Cellular fractionation and protein extraction

Collected cells were split in three equal parts for triplicate analysis. Cells were washed twice with ice cold PBS (HyClone), collected by scrapping, lysed with lysis buffer (10 mM Tris pH 8.4 (VWR), 140 mM NaCl, 1.5 mM $MgCl_2$, 0.5% NP-40 (Calbiochem), 1 mM dithiothreitol (DTT), protease and phosphatase inhibitor 1:100 added freshly) and spun at 1000 g at 4°C for 3 min. Supernatants were transferred to another tube (cytoplasmic fraction). Pellet was resuspended in lysis buffer plus 1/10 of detergent stock (3.3% w/v sodium deoxycholate, 6.6% Tween 40), vortexed at slow speed, incubated on ice for 5 min, spun at 1000 g at 4°C for 3 min and the supernatant was discarded. The pellet was rinsed with lysis buffer, and the nuclei were lysed with extraction buffer (20 mM HEPES pH 7.9, 1.5 mM $MgCl_2$, 0.42 M NaCl, 0.2 mM EDTA, 25% glycerol) and sonicated. Benzonase nuclease HC (Novagen) was added to cut nucleic acids and to obtain the nuclear fraction. Proteins were precipitated overnight with cold acetone (-20°C) (EMD) and resuspended the next day with a solution of 1% SDS and 50 mM ammonium bicarbonate. Cytosolic and nuclear proteins extracts were reduced for 20 min at 37°C with 0.5 mM tris(2-carboxyethyl)phosphine (TCEP) (Pierce) and then alkylated with 50 mM iodoacetamide for the same time and temperature. The excess of iodoacetamide was neutralized by the addition of 50 mM DTT. Proteins were quantified with microBCA protein assay (Pierce), diluted 10 times with 50 mM ammonium bicarbonate, digested with sequencing grade trypsin (1:100) (Promega) overnight at 37°C with high agitation, acidified below pH 4 with trifluoroacetic acid (TFA) to inactivate trypsin and dried in SpeedVac apparatus (Thermo).

Phosphopeptides enrichment and mass spectrometry

Phosphopeptides (1 mg/replicate) were enriched as previously described [141] on home-made TiO_2 affinity columns (4 columns - 250 µg of protein digest/column), 1.25 mg Titansphere, 5 µm, GL Sciences), using 250mM lactic acid (Fluka) and eluted with 30 µL of 1% ammonium hydroxide. Sample was acidified with 1 µL of TFA, desalted using 30 mg HLB cartridge (Waters), dried and resuspended in 2% acetonitrile (ACN) (Fischer Scientific)/0.2% formic acid (FA) (EMD) prior to analysis. Phosphopeptides were separated by online 2D-nanoLC using Eksigent 2D nanoLC system. Opti-Guard 1mm cation SCX column (Optimize Technologies) was used with five ammonium acetate salt fractions (0, 0.25, 0.5, 1 & 2M) in 2% ACN pH 3. Fractions were loaded on a trap column (4 mm length, 360 µm i.d.) and separated on a reverse phase analytical column (10 cm length, 150 µm i.d.) (Jupiter C_{18}, 3µm, 300 Å, Phenomenex). Both columns were packed manually. A gradient from 2 to 33% ACN over 53 min followed by a gradient from 33 to 60% ACN over 10 min with a flow rate of 600 nL/min was used to elute the peptides to the MS system with the nanoelectrospray source voltage set to 1.7 kV. MS analysis was done on a LTQ-Orbitrap XL mass spectrometer (Thermo Fisher Scientific). MS spectra were acquired with a resolution of 60 000 in FTMS using lock mass. CID MS/MS spectra were acquired in data-dependent mode for the three most abundant multiply charged ions with intensity above 10 000 counts. A dynamic exclusion window was set to 90 seconds.

MS/MS processing for peptide and protein identifications

MS/MS spectra peak lists were extracted from Xcalibur raw data files (Thermo Fischer Scientific) and preprocessed using Mascot Distiller v2.1.1 (Matrix Science) using the configuration file for low resolution MS/MS for the LTQ-Orbitrap. MGF peak lists were searched with Mascot 2.1 on a concatenated target/decoy IPI rat database v3.54 (39 928 protein sequences)[143] using the following parameters: peptide mass tolerance ± 10 ppm,

fragment mass tolerance ± 0.5 Da, trypsin with 2 missed cleavages, variable modifications: carbamidomethyl (C), deamidation (NQ), oxidation (M), phosphorylation (STY). All search results were then transferred to ProteoConnections, our in-house bioinformatics platform dedicated for phosphoproteomics analysis [117]. Identifications and MS/MS spectra are available online in ProteoConnections (http://www.thibault.iric.ca/proteoconnections). From there, a 1% peptide identifications FDR cut-off was applied to peptides assigned to proteins with a $p < 0.05$ significance threshold. Phosphorylation site localization confidence was assigned by ProteoConnections as previously proposed [19].

Data processing for label free peptide quantification

Peptide detection for all the raw files was done using our in-house peptides detection software. It retrieves peak intensity value, mass to charge ratio, retention time and charge state for each peptide. An intensity threshold of 10 000 counts was selected to detect peptides above the noise level. Detected peptides were then aligned with an *m/z* tolerance of 15 ppm, retention time window of ±1 min and same charge state to get abundance values for all conditions and replicates. A median normalization procedure was applied to minimize experimental variability of the peptides population abundance. Since the peptide population size are not equals between experimental conditions, only reproducibly detected peptides were used to calculate the median of samples. Soft-clustering of phosphopeptide kinetic profiles was done using fuzzy c-means clustering of MFuzz R package (number of clusters centroid $c = 6$, parameter $m = 1.5$). The number of clusters was chosen arbitrarily to 6 to show diverse phosphorylation change trends. Grouping is done by minimizing the Euclidian distance between the phosphopeptide kinetic fold change profiles with a weighted square error function. Fuzzy c-means clustering is a soft-clustering algorithm that distinguishes itself from hard-clustering algorithm by providing a membership probability value to each member of the clusters. This information indicates how similar a profile is compared to the rest of the cluster members.

Selection of candidate Erk1/2 substrates

An algorithm, implemented in Perl, was written to select putative Erk1/2 substrates using the following criteria. First, high confidence phosphorylation sites (≥ 75%) with the minimal ERK consensus [pS/T]P motifs were selected using ProteoConnections. Second, phosphorylation level must increase after serum stimulation. Phosphopeptides were filtered with a cut-off of $\Sigma\log_{10}$(fold control t_x/control t_0) ≥ 0.3, a value above most peptide fold change found between replicates. Third, a decrease in the phosphorylation site abundance after treatment with the Mek1/2 inhibitor must be measured. To select decreasing kinetic profiles, a cut-off of $\Sigma\log_{10}$(PD184352 treated/control) ≤ -0.7 was used, a value 2.5 time above fold change found between replicates. At least one time point of the kinetics must indicate a significant decrease with two-tailed t-test (p-value ≤ 0.05). Finally, manual inspection of the abundance profiles was performed on potential candidates to fix potential peak selection errors of the label-free quantification software.

Bioinformatics analyses

Gene ontology analysis for the putative for Erk1/2 substrates was carried out using the following methods. Enrichment and depletion of categories were calculated using odds ratios. Ratios are calculated this way: (number of proteins with the GO term in the dataset / number of proteins in the dataset) / (number of proteins with the GO term in the proteome / number of proteins in the proteome). P-values, which report the statistical significance, were calculated with Fisher's exact test. Protein interaction network of the Erk1/2 pathway was generated from ProteoConnections using STRING interaction dataset (high confidence interactions with score > 0.9 from database and experiments). Identified phosphorylated sites were compared with Swissprot v15.53 [84], Phospho.ELM v8.2 [86] and PhosphositePlus v2.0 [85] databases to report the fraction of novel sites. ProteoConnections was used to indicate the presence of potential docking sequences DEF (motif FX[FY]P) or D (motif [KR]2-5X1-6[LIV]X[LIV]) in the protein sequence of putative Erk1/2 substrates.

Supplemental results

Table A3.I: Annotated phosphorylation sites

The list of 7936 identified phosphorylation sites were annotated using ProteoConnections. The following data fields are reported: accession number, gene symbol, description, species, residue, position, site localization confidence, peptide sequence, peptide modifications, Mascot peptide score, known phosphorylation site, kinase motifs and predictions, phosphorylation binding motif, NLS, NES, protein domains, disordered regions prediction, secondary structure prediction, solvent accessibility prediction, PDB structure, glycosylation predictions and site conservation in other species (Supplemental file on CD-ROM).

Table A3.II: Kinetic profiles of phosphorylated peptides

Kinetic profiles of identified phosphorylated peptides for cytosolic (3015 profiles) and nuclear (5222 profiles) fractions are reported with the following values: peptide *m/z*, retention time, charge, sequence, modifications, score, protein accession, gene symbol, name, peptide start & stop position, modification positions on protein, ProteoConnections protein & kinetic graph links, average peptide abundance for each time point, number of replicates where the peptide is found, fold change (treated/control) and p-value (two-tailed Student t-test) (Supplemental file on CD-ROM).

Table A3.III: Putative Erk1/2 substrates kinetic profiles

Kinetic profiles of putative Erk1/2 substrates are reported for cytosolic (74 profiles) and nuclear (171 profiles) fractions. Reported values are the same as in Table A3.II. In addition, annotations for each phosphorylation site of putative substrates are included in the file (reported values are the same as in Table A3.I with two additional columns to show the presence of potential DEF & D-domain on the protein). Validated substrates by the *in vitro* kinase assays are highlighted in green (Supplemental file on CD-ROM).

Table A3.IV: Gene Ontology enrichment analyses on putative Erk1/2 substrates

Gene Ontology enrichment analyses were performed with the tool included in ProteoConnections. The following fields are reported: GO term identifier, ontology, ontology name, definition, depth, p-value (Fisher's exact test), number in putative Erk1/2 substrates list, number in the rat proteome, ratio, enrichment and IPI protein identifiers (Supplemental file on CD-ROM).

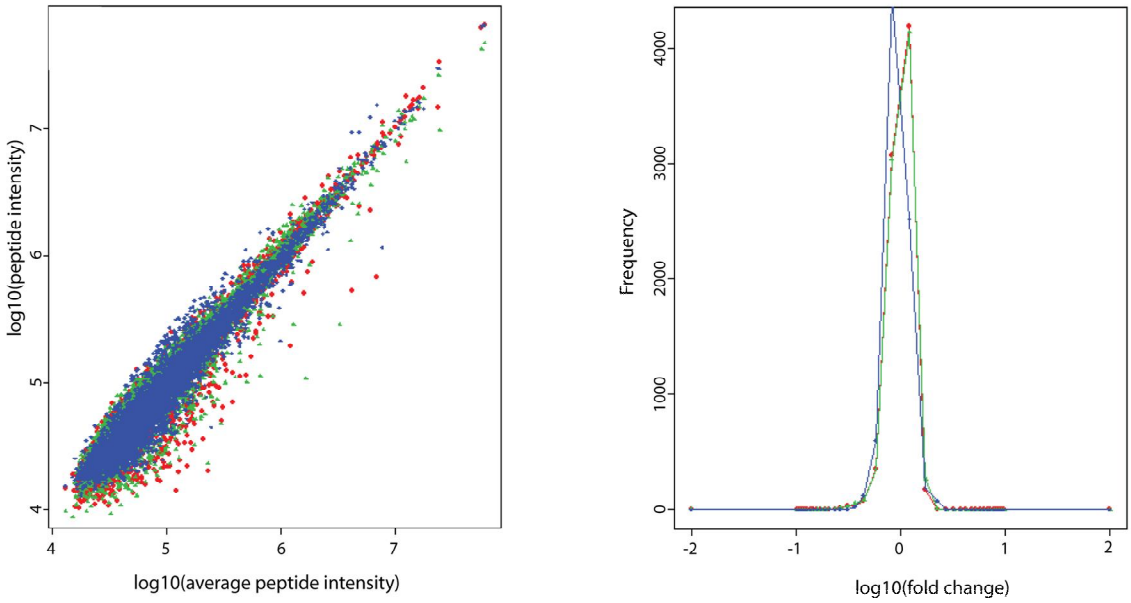


Figure A3.1: Reproducibility of label-free phosphopeptide quantification

TiO_2 enriched phosphopeptides were analyzed in triplicate by nanoLC-MS/MS on LTQ-Orbitrap and quantified by a label-free method. Phosphopeptides were detected from raw MS spectrum and aligned between replicates. Biological replicates of the Mek1/2 inhibitor treated cytosolic fraction at time 5 min are shown here to demonstrate the reproducibility of the experiment. 95% of phosphopeptides show less than two-fold change and the coefficient of variation (CV) of the measured peak intensities for the replicates of the same condition was on average 37%.

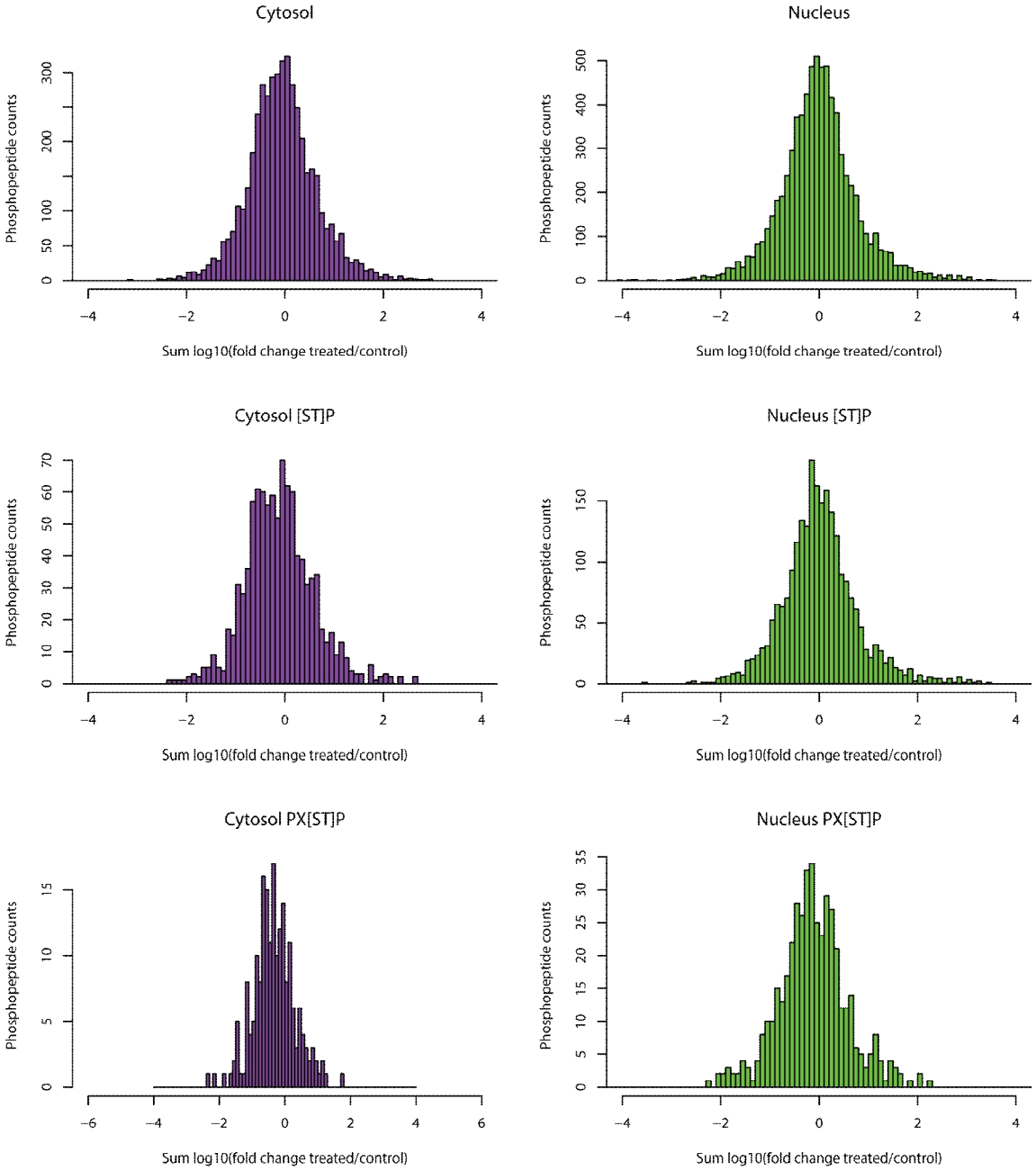


Figure A3.2: Distributions of phosphorylation abundance changes after Mek1/2 inhibition

These distributions show the measured $\Sigma\log_{10}$(PD184352 treated/control) for the four time points of all quantified phosphopeptides.

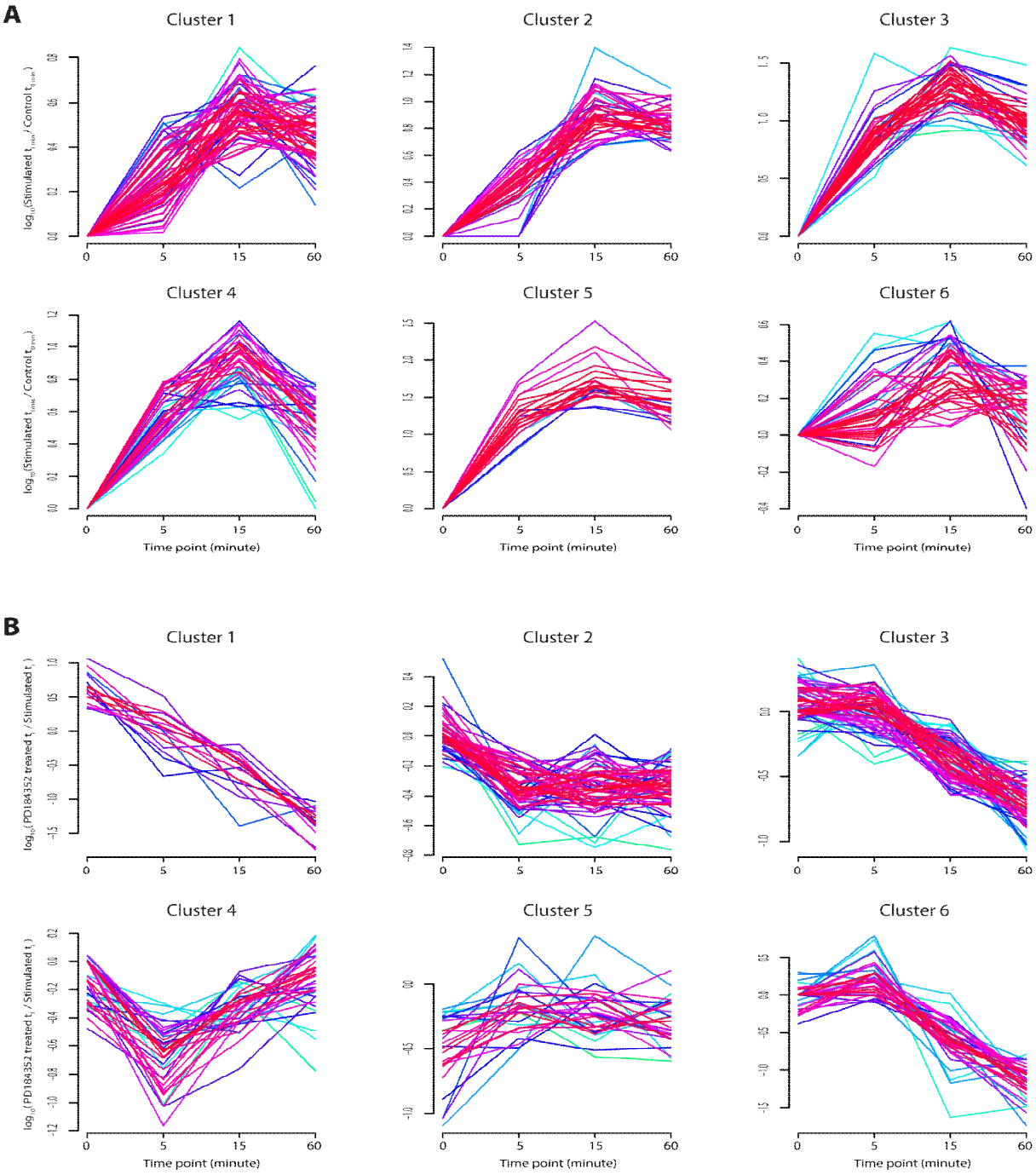


Figure A3.3: Kinetic profiles of putative Erk1/2 substrates

Stimulation (A) and inhibition (B) profiles of selected putative Erk1/2 substrates are displayed in this figure. Soft-clustering of kinetic profiles (fuzzy c-means clustering MFuzz R package, c=6, m=1.5) was done to show 6 groups (arbitrary chosen number) of phosphorylation change trends for both groups. High and low memberships are shown with a gradient from red and green. Members of clusters shown in A and B differ.

Table A3.V: Pairwise comparison of putative Erk1/2 substrates with known Erk1/2 substrates or candidates from phosphoproteomics experiments

Dataset	**Putative Erk1/2 substrates common to both dataset**	**No significant changes observed or low localization confidence**	**Different phosphorylated site observed**	**Proteins not identified**
Yoon & al. (reference dataset of known Erk1/2 substrates)	Cttn*, Gja1, Gtf2i*, Map1b, Map2k2, Marcks, Pak1, Pxn, Smarca5*, Stmn1, Tpr	Atf2, Canx, Cdkn1b, Dmd, Eif4ebp1, Hnrnpk, Hsf1, Jun, Lmnb2, Mcl1, Raf1	Cald1, Map4, Mycn, Pla2g4a, Polr2a, Top2a, Trim24, Wipf3	Adrbk1, Amph, Anxa11, Apbb1, Ar, Bad, Bcl2l11, Bcl6, Bdp1, Bmal1, Braf, Btg2, Capn2, Cdk2ap2, Crem, Cry1, Cryaa, Dlg4, Dusp1, Dusp16, Dusp4, Dusp6, Egfr, Elk1, Elk3, Elk4, Esr, Etv1, Fos, Fosl1, Gab1, Gab2, Gata1, Gorasp2, Grb10, H3, Hif1a, Ier3, Kcnd2, Lat, Lifr, Map2, Mapk1ip1, Mapkapk3, Mapkapk5, Mapt, Mitf, Mnk1, Mnk2, Myc, Mylk2, Nckipsd, Ncoa1, Nefh, Nefm, Nfatc4, Nr4a2, Pax6, Pgr, Plcb1, Plcg1, Pparg, Ptpn11, Rab4, Rgs19, Rps6ka2, Rps6ka4, Rps6ka5, Rps6kb1, Runx1, Runx2, Scnn1b, Scnn1g, Sh2b1, Shc1, Smad1, Smad2/3, Smad4, Sorbs3, Sos1, Sp1, Stat1/3, Stat5a, Syk, Syn1, Tal1, Tcf3, Tgif, Th, Tnfrsf1a, Tnip1, Tob, Tp53, Ubtf

Kosako & al.[%]	Cast, Cttn, Dpysl3, Dync1li1, Fam129b, Lima1, Map2k2	Cald1, Dync1i2, Eif4b, Hnrnpk, Lmna, Pla2g4a, Ppil4, Sgta, Smarcc1, Stip1, Tpd52l2, Trim28, Ubxn1	Ik , Map2k1	Dmn1, Eef1g , Eif4e, Fos, Lcp1, Mkiaa0776, Nup50, Prrc1, Snrnp70, Snx5, Sorbs3, Ugdh
Old & al.	Cttn*, Fam129b*, Marcks*, Stmn1, Tpr	Bag3, Gtpbp1, Hdgf, Hnrnpk, Pcbp2	Arhgap17, Rps6ka1, Smc4	Stat3, Stk10
Pan & al.	Ahnak*, Brd9, Fam129b, Nup153*, Ppfibp1, Rai14*, Rsf1*, Stmn1, Tpr, Tsc22d4	Bag3, Sgta	Cttnbp2nl, Dock5, Eps15, Irs1, Mkl1, Nup214, Plec1, Rps6ka1, Shroom3	Ablim3, Dennd1a, Etv3, Flj61655, Gigyf2, Huwe1, Kiaa1967, Mapk7, Mkl2, Mycbp2, Mylk, Nup50, Rbm12b
Carlson & al.	Ahctf1*, Ahnak*, Aim1*, Bat2*, Cdc42ep1*, Cdc42ep2*, Dync1li1*, Gja1, Hnrnph2, Map1b Nup153*, Phf2*, Rai1*, Rai14, Stmn1, Tnks1bp1*, Tpr*, Wiz*	Dync1i2, Lmna, Mybbp1a, Rps3, Rtn4, Supt5h	Arhgef17, Kab, Eif4enif1, Erf, Irs2, LOC500726, Ksr1, Larp1, Map1a, Mgea5, Myo9b, Ncor2, Pdxdc1, Phldb1, Raph1, Rell1, Rexo1, Tpx2, Ubap2, Xrn2	Ahnak2, Akap12, Atg2b, C21orf70, Dennd4c, Dlg5, Dlg7, Dnajc30, Egfr, Etv3, Fam165a, Fam195b, Fox2, Gigyf2, Gli2, Gtse1, Mapkapk2, Mast2, Nid1, Rbpms, Ripk3, Safb2, Sned1, Snx2, Sorbs3, Ssbp3, Stk10, Supv3l1, Tanc1, Tbc1d23, Tfpi, Tgfb1i1, Tp53bp2, Ubap2l, Udpgdh

[%] Comparison at protein level only, * Putative Erk1/2 substrates in both dataset but with different location of phosphorylation

Annexe 4 Figures et tableaux supplémentaires du chapitre 4

Table A4.I : Synthetic phosphopeptides analysis.

List of the 9 synthetic phosphopeptides used to test separation of isomers by RP-HPLC and the algorithm performance. Peptide sequence, modification, *m/z*, localisation confidence, score, retention time and resolution of LC separation is reported (Supplemental file on CD-ROM).

Table A4.II: Phosphopeptide positional isomers discovered in mouse and rat.

List of the 64 isomers detected from the mouse and rat datasets. In addition to the data fields reported in Table A4.I, the distance between phosphorylated amino acid, distance to the nearest acid or basic residues and local hydropathicity scale are reported (Supplemental file on CD-ROM).

Table A4.III : Phosphopeptide positional isomers discovered in fly.

List of the 117 isomers detected from the fly dataset and included peptide ions for the targeted analysis. In addition to the data fields reported in Table A4.II, inclusion *m/z*, inclusion time window, SCX fraction, inclusion & identification success, true isomers & ambiguous cases, and artifacts are reported for the survey and targeted analysis (Supplemental file on CD-ROM).

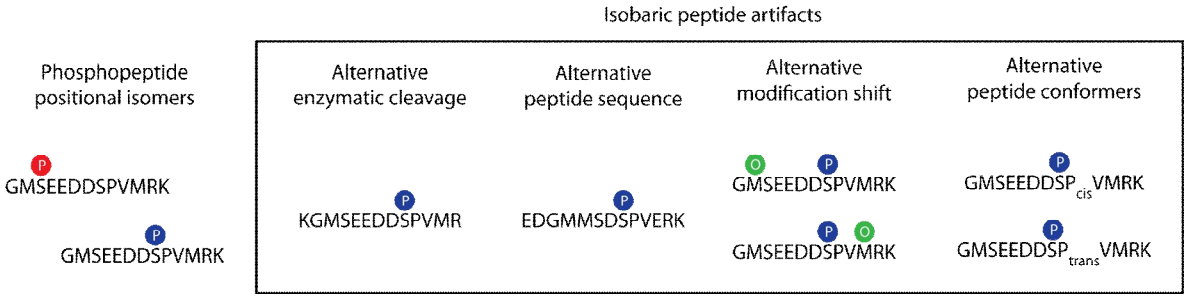


Figure A4.1: Isobaric peptide artifacts.

Isobaric peptide artifacts can be confounded to phosphopeptide positional isomers. Alternative enzymatic cleavage, peptide sequence, other modification position shift (e.g. alternate position of oxidation) and conformers are peptide species with the same mass that can be separated by liquid chromatography.

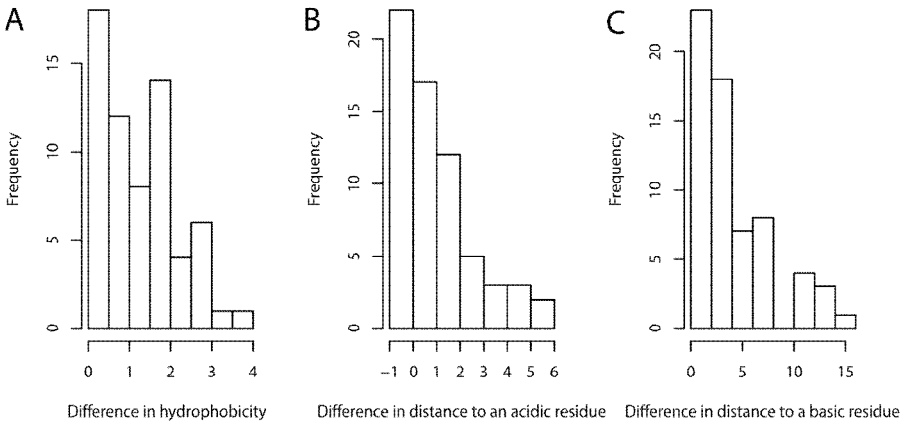


Figure A4.2: Physicochemical properties phosphopeptide positional isomers separated by RP-HPLC.

Difference in physicochemical properties (hydrophobicity and electrostatic forces) of 64 phosphopeptide isomers identified in IEC-6 & J774 cells that could influence retention time drift on a reverse-phase column. A) Difference in absolute local hydrophobicity scale between the region centered on the phosphorylated sites (windows of 3 amino acids, amino acids scale values from Kyte J., Doolittle R.F. [209]). Difference in proximity (in amino acids) of the phosphate group to an acidic residue (B) or a basic residue (C) that could affect electrostatic interactions that stabilize a peptide conformation are reported.

l

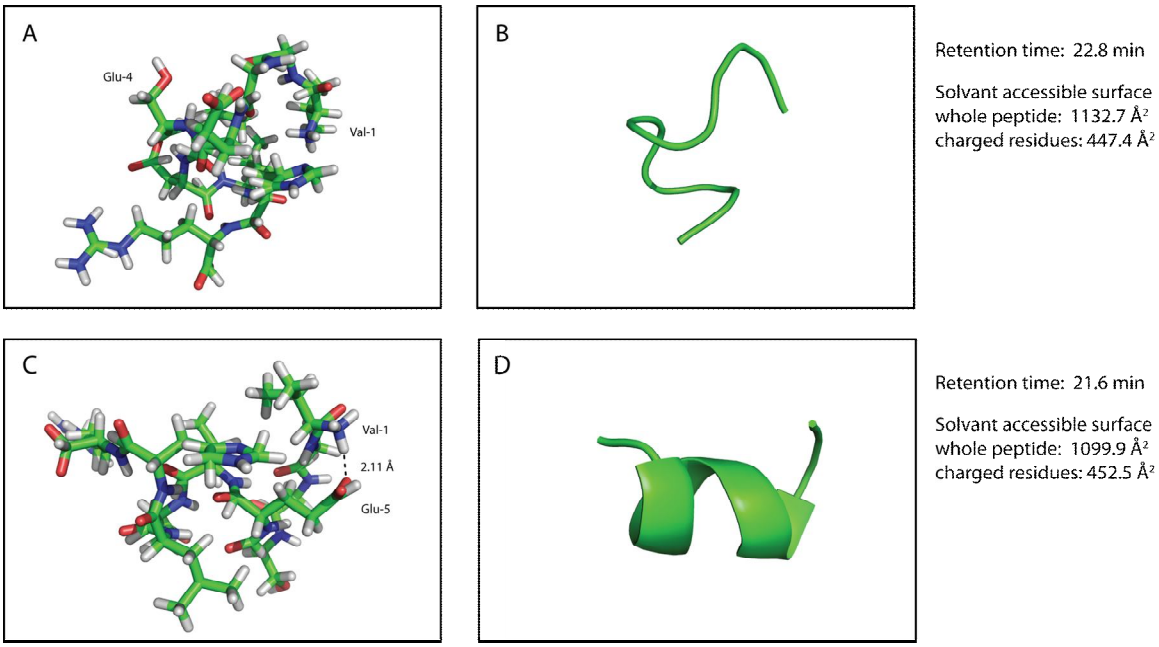


Figure A4.3: Predicted conformations of phosphopeptide positional isomers.

The positional isomers of VGGSSVDLHR phosphorylated at serine 4 (A,B) or at serine 5 (C, D) were predicted using PEP-FOLD [208]. Phosphorylated residues were mimic by a glutamic acid in the prediction stage. This prediction suggests that the change in the phosphorylated position cause a transition from a random coil (B) to an helix (D). The phosphorylated serine 5 stabilizes the helix secondary structure by an electrostatic interaction and the formation of an N-H···O=C hydrogen bond with the N-terminal residue. Physiochemical properties of these different conformations can explain the change in retention time on RP-HPLC between the two isomers species.

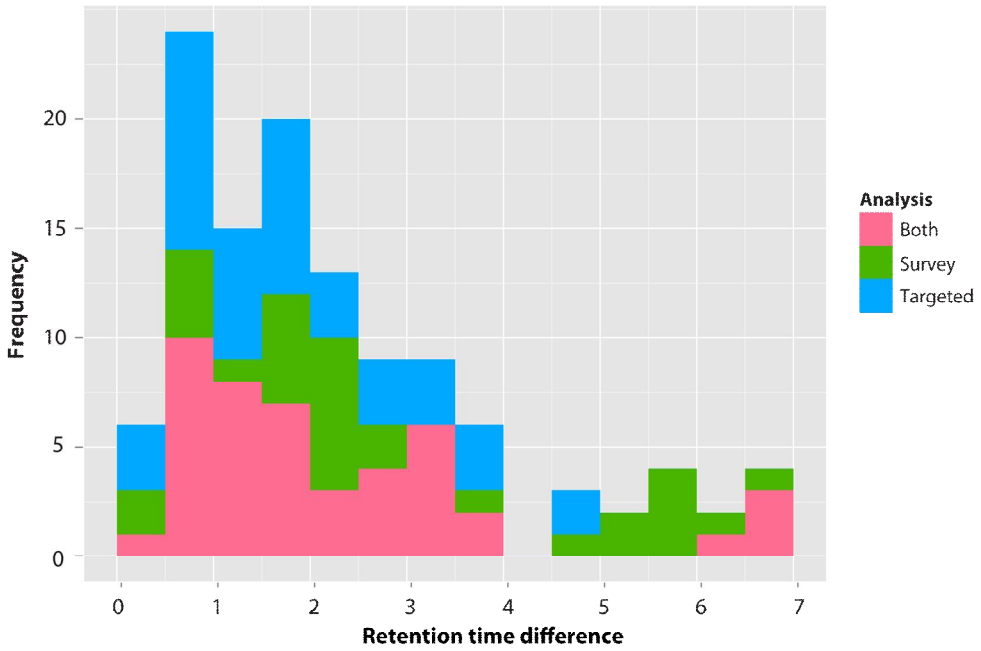


Figure A4.4: Retention time difference of isomers separated by RP-HPLC in the fly S2 sample found in the survey and targeted analysis.

Retention time difference of 117 phosphopeptide isomers separated by RP-HPLC identified in fly S2 cells using the LC-MS elution profile detection algorithm and supported by MS/MS identifications. The histogram shows the common and uniquely detected isomers from the survey and the targeted analysis.

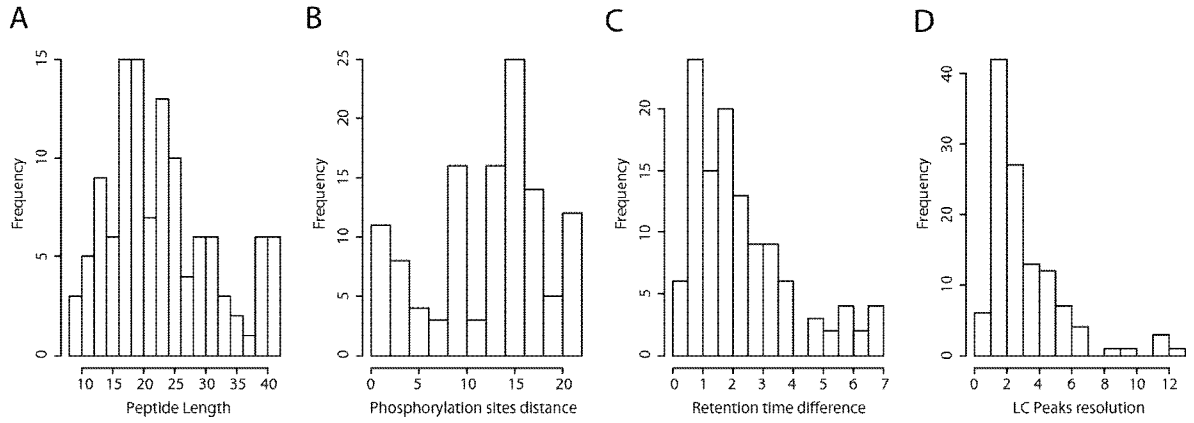


Figure A4.5: Properties of phosphopeptide isomers separated by RP-HPLC in the fly S2 sample.

Properties of 117 phosphopeptide isomers separated by RP-HPLC identified in fly S2 cells using the LC-MS elution profile detection algorithm and supported by MS/MS identifications. Histograms for the distribution of the isomers peptide length is presented in (A), distance in amino acids between phosphorylated site positions (B), retention time difference at peak top (C), and the chromatographic resolution of isomer peaks at intensity threshold of 10 000 counts (D).

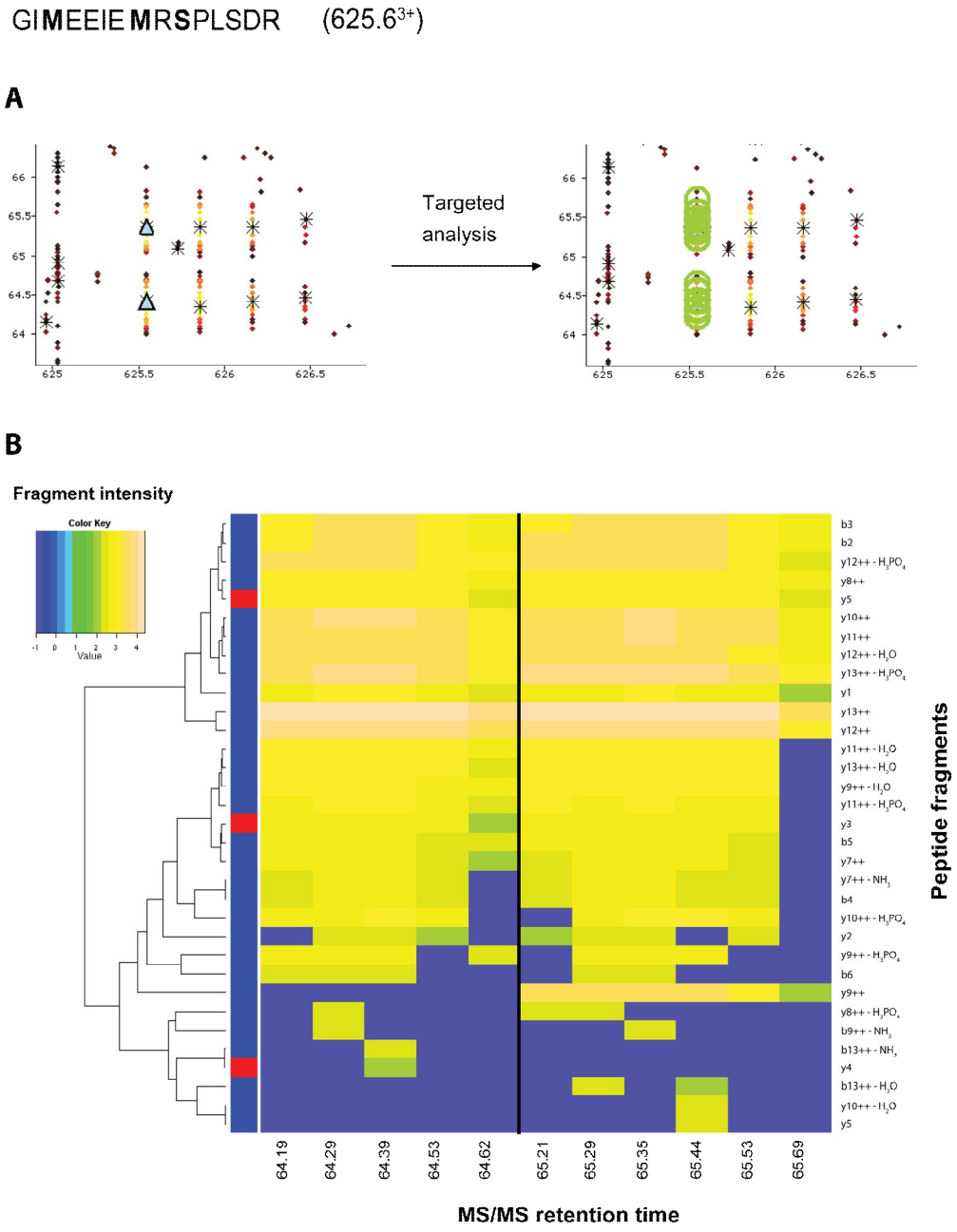


Figure A4.6: RP-HPLC separation of phosphopeptide conformers.

Example of separated phosphopeptide conformers: GIMEEIEMRpSPLSDR (625.63+). Both methionine of this peptide are oxidized. A) LC-MS profile showing the two separated conformers and the acquired HCD MS/MS spectra within the targeted analysis (green circles). B) Heatmap of the observed peptide fragments. Each row is a fragment and each column a MS/MS scan. Fragments marked with red are specific to localize the phosphorylated site. The vertical black line separates both conformers.

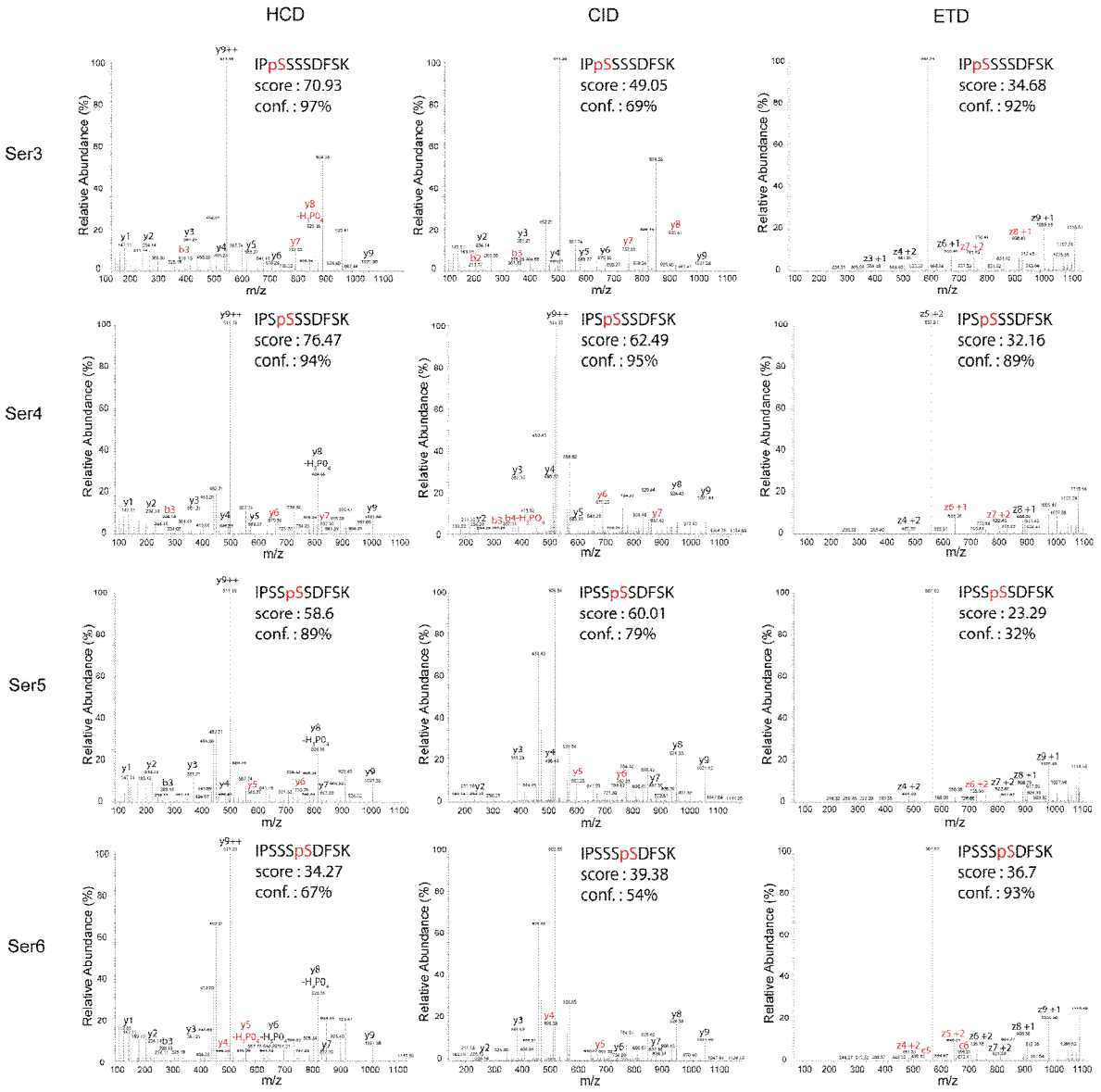


Figure A4.7: MS/MS spectra of four phosphopeptide isomers of IPSSSSDFSK

CID, ETD and HCD MS/MS spectra are shown for the four phosphopeptide positional isomers of IPSSSSDFSK. Mascot score and phosphorylation localization confidence are reported for each spectrum.

Annexe 5 Contributions scientifiques

Publications

2011

Courcelles M, Lemieux S, Voisin L, Meloche S, Thibault P
ProteoConnections: a bioinformatics platform to facilitate proteome and phosphoproteome analyses, Proteomics, 11, 2654-2671.

Freschi L, Courcelles M, Thibault P, Michnick SW, Landry CR
Phosphorylation network rewiring by gene duplication, Molecular Systems Biology, 7, 504.

2010

Galisson F, Mahrouche L, Courcelles M, Bonneil E, Meloche S, Chelbi-Alix M, Thibault P
A novel proteomics approach to identify SUMOylated proteins and their modification sites in human cells, Mol Cell Proteomics, 10, M110 004796.

2009

Trost M, English L, Lemieux S, Courcelles M, Desjardins M, Thibault P
The phagosomal proteome in interferon-gamma-activated macrophages, Immunity, 30, 143-154.

Gharib M, Marcantonio M, Lehmann S, Courcelles M, Meloche S, Verreault A, Thibault P
Artifactual sulfation of silver-stained proteins: implications for the assignment of phosphorylation and sulfation sites, Mol Cell Proteomics, 8, 506-518.

2008

Marcantonio M, Trost M, Courcelles M, Desjardins M, Thibault P

Combined enzymatic and data mining approaches for comprehensive phosphoproteome analyses: application to cell signaling events of interferon-gamma-stimulated macrophages, Mol Cell Proteomics, 7, 645-660.

Conférences et présentations

2011

Courcelles M, Voisin L, Fremin C, Lemieux S, Meloche S, Thibault P

Biological insights into the rat phosphoproteome

-HUPO 10th world congress, Palexpo, Genève, Genève, Suisse, 4-7 Septembre.

Courcelles M, Lemieux S, Voisin L, Meloche S, Thibault P

ProteoConnections: an analysis platform to accelerate proteomes and phosphoproteomes exploration

-Rapport de recherche de l'IRIC, Institut de recherche en immunologie et en cancérologie, Montréal, Québec, Canada, 21 Janvier.

2010

Courcelles M, Voisin L, Julien C, Lemieux S, Meloche S, Thibault P

Cinétique du phosphoprotéome des cellules épithéliales de rat suite à l'inhibition de la voie Erk

-Symposium Protéomique et biologie des systèmes au 78e congrès de l'Acfas, Université de Montréal, Montréal, Québec, Canada, 11 Mai.

Courcelles M, Lemieux S, Voisin L, Meloche S, Thibault P

ProteoConnections: an analysis platform to accelerate proteomes and phosphoproteomes exploration

-3e journée scientifique de l'IRIC, Institut de recherche en immunologie et en cancérologie, Montréal, Québec, Canada, 27 Novembre.

-Canadian National Proteomics Network, Fairmont The Queen Elizabeth, Montréal, Québec, Canada, 9-10 Mai.

-RECOMB Satellite Conference on Computational Proteomics, University of California San Diego, San Diego, Californie, États-Unis, 27-28 Mars.

2009

Courcelles M, Voisin L, Julien C, Lemieux S, Meloche S, Thibault P

Profiling global changes in the phosphoproteome of epithelial cells following the inhibition of Erk1/2 MAP kinase pathway

-Lake Louise 22st Workshop on Tandem Mass Spectrometry, Fairmont Chateau Lake Louise, Lake Louise, Alberta, Canada, 3-5 Décembre.

-MonBUG Bioinformatics Symposium, Institut de recherches cliniques de Montréal, Montréal, Québec, Canada, 3 Septembre.

-Systems Biology in cancer and immunology, Institut de recherche en immunologie et en cancérologie, Montréal, Québec, Canada, 14-15 Juillet.

-57th ASMS Conference on Mass Spectrometry and Allied Topics, Pennsylvania Convention Center, Philadelphie, Pennsylvanie, États-Unis, 31 Mai au 4 Juin.

2008

Courcelles M, Voisin L, Julien C, Lemieux S, Meloche S, Thibault P

Phosphoproteome kinetic profiling upon Erk pathway inhibition

-2e journée scientifique de l'IRIC, Institut de recherche en immunologie et en cancérologie, Montréal, Québec, Canada, 29 Novembre.

Courcelles M, Lemieux S, Voisin L, Meloche S, Thibault P

ProteoConnections: an analysis platform to accelerate proteomes and phosphoproteomes exploration

-5e Colloque bio-informatique Robert Cedergren, Université de Montréal, Montréal, Québec, Canada, 3-4 Novembre.

-16th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB), Metro Toronto Convention Centre, Toronto, Ontario, Canada, 19-23 Juillet.

-Proteomics, Bioinformatics and Systems biology (Canadian Proteome Society regional meeting), Université Laval, Québec, Québec, Canada, 9 Mai.

Gharib M, Courcelles M, Verreault A, Thibault P

Silver staining-induced sulfonation: an obstacle in the identification of genuine protein phosphorylation

-56th ASMS Conference on Mass Spectrometry and Allied Topics, Colorado Convention Center, Denver, Colorado, États-Unis, 1-5 Juin.

Fortier MH, Caron E, Courcelles M, Perreault C, Thibault P

The mTOR signaling pathway and its influence on the MHC Class I peptide repertoire

-56th ASMS Conference on Mass Spectrometry and Allied Topics, Colorado Convention Center, Denver, Colorado, États-Unis, 1-5 Juin.

2007

Courcelles M, Voisin L, Meloche S, Thibault P

MSdatabase: A novel proteomics and phosphoproteomics analysis platform

-4e Colloque bio-informatique Robert Cedergren, Université de Montréal, Montréal, Québec, Canada, 8-9 Novembre.

Trost M, Marcantonio M, Courcelles M, Desjardins M, Thibault P

System-biology analysis of Interferon-gamma activated mouse macrophages: from the cytosol to the phagosome

-55th ASMS Conference on Mass Spectrometry and Allied Topics, Indiana Convention Center, Indianapolis, Indiana, États-Unis, 3-7 Juin.