ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΒΙΟΧΗΜΕΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

Κυριακίδου Πελαγία

ΕΦΑΡΜΟΓΕΣ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ

ΜΟΡΙΑΚΗ ΓΕΝΕΤΙΚΗ

ΔΙΑΓΝΩΣΤΙΚΟΙ ΔΕΙΚΤΕΣ

ΛΑΡΙΣΑ 2016

Εξόρυξη δεδομένων φωσφοπρωτεωμικής από τη βιβλιογραφία και βιοπληροφορική ανάλυση

Literature mining of phospho-proteomic data and bioinformatics analysis

Τίτλος πτυχιακής εργασίας: Εξόρυξη δεδομένων φωσφοπρωτεωμικής από τη βιβλιογραφία και βιοπληροφορική ανάλυση.

Φοιτήτρια: Κυριακίδου Πελαγία

Επιβλέπων Καθηγητής: Αμούτζιας Γρηγόριος

Εργαστήριο Βιοπληροφορικής

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ

Αμούτζιας Γρηγόρης (επιβλέπων): Επίκουρος Καθηγητής Βιοπληροφορικής στη Γενωμική, του Τμήματος Βιοχημείας & Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας

Μόσιαλος Δημήτριος: Επίκουρος Καθηγητής Βιοτεχνολογίας Μικροβίων, του Τμήματος Βιοχημείας & Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας

Στρατίκος Ευστράτιος: Ερευνητής Α' στο ΕΚΕΦΕ Δημόκριτος

Ευχαριστίες:

Η παρούσα διπλωματική εργασία εκπονήθηκε στο εργαστήριο Βιοπληροφορικής, του Τμήματος Βιοχημείας & Βιοτεχνολογίας (ΤΒΒ) του Πανεπιστημίου Θεσσαλίας στο πλαίσιο του μεταπτυχιακού προγράμματος: «Εφαρμογές Μοριακής Βιολογίας - Μοριακή Γενετική - Διαγνωστικοί Δείκτες». Πραγματοποιήθηκε υπό την επίβλεψη του Επίκουρου Καθηγητή κ. Αμούτζια Γρηγόρη τον οποίο θα ήθελα να ευχαριστήσω ιδιαίτερα που μου εμπιστεύτηκε την εκτέλεση αυτής της εργασίας και για το χρόνο που διέθεσε για την ολοκλήρωση αυτής. Επίσης, θα ήθελα να ευχαριστήσω τον Επίκουρο Καθηγητή Βιοτεχνολογίας Μικροβίων, κ. Μόσιαλο Δημήτριο καθώς και τον Δρ. Στρατίκο Ευστράτιο, Ερευνητή Α' στο ΕΚΕΦΕ Δημόκριτος, για την συμμετοχή στην τριμελή συμβουλευτική επιτροπή. Τέλος, ευχαριστώ θερμά όλα τα μέλη του εργαστηρίου για τις χρήσιμες συμβουλές και το ενδιαφέρον που έδειξαν καθ’ όλη τη διάρκεια εκπόνησης της εργασίας.

Η εργασία αυτή χρηματοδοτήθηκε από το ερευνητικό έργο FAB-PHOS “Φιλτράρισμα, λειτουργικός σχολιασμός και βιοπληροφορική ανάλυση δεδομένων από πειράματα φωσφοπρωτεωμικής μεγάλης κλίμακας” της Δράσης Αριστεία ΙΙ, που συντονίστηκε από τον Δρ. Αμούτζια.

Περίληψη

Εξόρυξη δεδομένων φωσφοπρωτεωμικής από τη βιβλιογραφία και βιοπληροφορική ανάλυση.

Η φωσφορυλίωση είναι η πιο συχνή μετα-μεταφραστική τροποποίηση και είναι ικανή να ρυθμίζει (ως μοριακός διακόπτης) τη δραστικότητα των ενζύμων, το σχηματισμό μοριακών συμπλόκων και τον υποκυτταρικό εντοπισμό ή την καταστροφή των πρωτεϊνών. Η φωσφορυλίωση/απο-φωσφορυλίωση είναι επίσης ένα βασικό συστατικό της μεταγωγής σήματος. Επομένως, είναι υψίστης σημασίας να γνωρίζουμε ποιες πρωτεΐνες είναι φωσφορυλιωμένες και σε ποια αμινοξέα τους.
Ο τομέας της πρωτεωμικής είναι ένα σχετικά καινούργιο και γρήγορα αναπτυσσόμενο ερευνητικό πεδίο. Η εφαρμογή της φασματομετρίας μάζας σε συνδυασμό με τις τεχνικές εμπλουτισμού (enrichment methods) των φωσφοπεπτιδίων έχει προωθήσει αρκετά τον τομέα των phosphoproteomics, έτσι ώστε εκατοντάδες ή ακόμα και χιλιάδες θέσεις φωσφορυλίωσης (p-sites) μπορούν να εντοπιστούν σε ένα μόνο πείραμα. Ωστόσο, όπως και με οποιαδήποτε τεχνική μεγάλης κλίμακας (High-Throughput - HTP), υπάρχουν ανησυχίες σχετικά με την ποιότητα των δεδομένων, με συνέπεια τα παραγόμενα δεδομένα να χρήζουν περαιτέρω φιλτραρίσματος και ανάλυσης. Επομένως, υπάρχει η ανάγκη για μια αυστηρή αξιολόγηση των δεδομένων ώστε να φιλτραριστούν πιθανώς ψευδείς ταυτοποιήσεις πεπτιδίων και των p-sites τους πριν καταλήξουμε σε συμπεράσματα σχετικά με τη δομή και τις ιδιότητες ενός φωσφοπρωτεώματος (phosphoproteome) (Bodenmiller et al., 2007).
Σε μια προσπάθεια να παρέχουμε μία αξιόπιστη συλλογή του φωσφοπρωτεώματος διαφόρων οργανισμών, ελέγχθηκαν πάνω από 1000 άρθρα και έγινε εξόρυξη δεδομένων και φιλτράρισμα από 205, για πέντε καλά μελετημένα είδη: του ανθρώπου, ποντικού, αρουραίου, *Arabidopsis* και ζύμης. Τα φιλτραρισμένα δεδομένα της ανάλυσης αυτής για όλους τους οργανισμούς που μελετήθηκαν είναι επίσης διαθέσιμα μέσω μίας Βάσης Δεδομένων, της PPProj.

Abstract

Literature- mining of phospho-proteomics data and bioinformatics analysis.

Phosphorylation is the most frequent post-translational modification made to proteins and may regulate (as a molecular switch) enzyme activity, complex formation, subcellular localization, or degradation. Phosphorylation/de-phosphorylation is also a key component of signal transduction. Thus, it is crucial to know which proteins are phosphorylated and on which of their amino acids.
Proteomics is a relatively new and rapidly developing field of research. The technology of mass spectrometry coupled with enrichment methods has been advanced considerably in the phosphoproteomics field so that hundreds or even thousands of phosphorylation sites (p-sites) can be detected in a single experiment. However, the quality of the data is sometimes questionable, as in every High-throughput Technology (HTP). Thus, there is a need for a step of rigorous assessment of the data to filter out potentially false peptide identifications and p-site localizations before drawing any conclusions about the properties of a phosphoproteome (Bodenmiller et al., 2007). In an effort to provide a credible collection of the phosphoproteome of various organisms, we mined over 1000 articles, from which we gathered and filtered 205 publicly available phosphoproteomic datasets/articles for five well-studied species (human, mouse, rat, Arabidopsis and yeast). The results of the data analysis for all organisms that were studied are stored in a Database, named “PPProj”.

Περιεχόμενα

Εισαγωγή

Σκοπός εργασίας

Η παρούσα εργασία εκπονήθηκε στο εργαστήριο Βιοπληροφορικής, του Τμήματος Βιοχημείας & Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας. Αποτελεί μέρος της ευρύτερης ερευνητικής δραστηριότητας του εργαστηρίου στον τομέα της Βιοπληροφορικής Ανάλυσης Βιολογικών Δεδομένων. Συγκεκριμένα, χρηματοδοτήθηκε από το ερευνητικό έργο FAB-PHOS "Φιλτράρισμα, λειτουργικός σχολιασμός και βιοπληροφορική ανάλυση δεδομένων από πειράματα φωσφοπρωτεωμικής μεγάλης κλίμακας" της Δράσης Αριστεία ΙΙ που συντονίστηκε από τον. Δρ. Αμούτζια. Το ερευνητικό αυτό έργο αποτελεί μία προσπάθεια συνέχισης και εμπλουτισμού μιας παλαιότερης έρευνας του εργαστηρίου με τίτλο: «Evaluation and properties of the budding yeast phosphoproteome» (Amoutzias *et al.*, 2012).
Στα πλαίσια της παρούσας διπλωματικής εργασίας ασχοληθήκαμε με την επεξεργασία βιολογικών, και πιο συγκεκριμένα φωσφο-πρωτεωμικών δεδομένων, που προκύπτουν από πειράματα μεγάλης κλίμακας (High Throughput Experiments - High Throughput Proteomics) (Lesley, 2001; Shen et al., 2001).
Ο βασικός στόχος αυτής της διπλωματικής εργασίας ήταν η δημιουργία ενός συνόλου δεδομένων φωσφοπρωτεωμικής υψηλής ποιότητας που θα χρησιμοποιούνταν για μελλοντικές αναλύσεις. Τα δεδομένα αυτά όντως χρησιμοποιήθηκαν σε μια ερευνητική εργασία που έχει υποβληθεί για δημοσίευση στο περιοδικό GigaScience, με τίτλο: "Estimating the total number of phosphoproteins and phosphorylation sites in eukaryotic proteomes" με συγγραφείς τους:" Panayotis Vlastaridis, Pelagia Kyriakidou, Anargyros Chaliotis, Yves Van de Peer, Stephen G. Oliver, Grigoris D. Amoutzias*".

Εισαγωγικές έννοιες

Σε αυτό το σημείο θα αναφερθούμε αναλυτικά στις μεθόδους που εφαρμόστηκαν σε αυτήν εδώ την εργασία. Πιο συγκεκριμένα, θα σταθούμε αρχικά στην αποσαφήνιση γενικότερων εννοιών όπως είναι η «εξόρυξη δεδομένων» και η «απόκτηση γνώσης». Κατόπιν, θα αναλύσουμε τα γνωρίσματα των δεδομένων που χρησιμοποιήθηκαν και ποιοι ήταν οι λόγοι επιλογής τους.

Σε προηγούμενα χρόνια, η έρευνα τόσο στον τομέα της Γενετικής όσο και της Πρωτεωμικής μπορούσε να εστιάσει σε ένα μόνο γονίδιο ή μία πρωτεΐνη τη φορά. Παρόλα αυτά, η εξέλιξη των τεχνικών στις βιοεπιστήμες έχει αρχίσει να παρέχει ένα εκθετικά αυξανόμενο όγκο δεδομένων. Εξαιτίας της ραγδαίας αύξησης των δεδομένων που είναι ελεύθερα διαθέσιμα στις Βάσεις Δεδομένων μέσω διαδικτύου σε συνδυασμό με το χαμηλό έως ανύπαρκτο έλεγχο της ορθότητας αυτών παρατηρείται μία μετάβαση στην έρευνα από μελέτες ωθούμενες από υπόθεση (hypothesis-driven) προς μελέτες ωθούμενες από δεδομένα (data-driven). Η μετατόπιση αυτή φέρνει αντιμέτωπους τους επιστήμονες με νέες προκλήσεις στην ανάλυση δεδομένων.

Εξόρυξη γνώσης από Βάσεις Δεδομένων

Η εξόρυξη δεδομένων αποτελεί φυσική απόρροια της εξέλιξης της τεχνολογίας της πληροφορίας η οποία περιλαμβάνει τη συλλογή των δεδομένων και την τοποθέτηση αυτών σε Βάσεις Δεδομένων ώστε να είναι ευρέως διαθέσιμα στους ερευνητές να τα επεξεργαστούν και να τα

ερμηνεύσουν (Han et al., 2011).
Τα δεδομένα έχουν συγκεντρωθεί για διάφορους λόγους, αλλά συνήθως δεν είναι οργανωμένα με τρόπο που να εξυπηρετεί τις διαδικασίες της μάθησης. Έτσι η ανακάλυψη γνώσης από Βάσεις Δεδομένων είναι μια σύνθετη διαδικασία αναγνώρισης έγκυρων, νέων, ενδεχομένως χρήσιμων και απόλυτα κατανοητών προτύπων και σχέσεων στα δεδομένα (Βλαχάβας et al, 2002).

Ανακάλυψη Γνώσης σε Βιολογικά Δεδομένα

Πριν περάσουμε στην παρουσίαση και την ανάλυση των μεθόδων που χρησιμοποιήθηκαν για την εξόρυξη δεδομένων στην συγκεκριμένη εργασία, θα ήταν απαραίτητη μια σύντομη εισαγωγική αναφορά στις βασικές έννοιες και αρχές που διέπουν τα δεδομένα που χρησιμοποιήθηκαν.
Τα τελευταία χρόνια η ολοένα αυξανόμενη υπολογιστική δύναμη και η κατακόρυφη αύξηση του χώρου αποθήκευσης δεδομένων έχει βοηθήσει στην επίλυση προβλημάτων σε διάφορους επιστημονικούς κλάδους. Ένας από τους επιστημονικούς κλάδους που πήρε ιδιαίτερη ώθηση ήταν αυτός της Βιολογίας. Ωστόσο, δημιούργησε και περαιτέρω προβλήματα εφόσον αυτοί οι τεράστιοι όγκοι δεδομένων διαθέτουν θόρυβο και κρύβουν σημαντικές πληροφορίες οι οποίες είναι αδύνατον να ανιχνευθούν χωρίς επιπρόσθετη επεξεργασία και φιλτράρισμα. Επομένως, σε τέτοιου είδους προβλήματα η Πληροφορική καθώς και άλλες εφαρμοσμένες επιστήμες, όπως αυτή των Μαθηματικών και της Στατιστικής, παίζουν κυρίαρχο ρόλο στην επίλυσή τους.
Η μεγάλη συσσώρευση των βιολογικών δεδομένων επιβεβαιώνεται από την εκθετική αύξηση του μεγέθους των διαφόρων Βάσεων Δεδομένων.

Η επιστήμη της Βιοπληροφορικής

Η Βιοπληροφορική ασχολείται με την ανάπτυξη και την εφαρμογή εργαλείων για τη διαχείριση Βιολογικών Δεδομένων. Ο όρος διαχείριση περιλαμβάνει την αποθήκευση, οργάνωση, ανάλυση ακόμα και την οπτικοποίηση των δεδομένων αυτών. Εκτός από την αποτελεσματική διαχείριση του όγκου της βιολογικής πληροφορίας, η Βιοπληροφορική συνεισφέρει και στην κατανόηση αυτής της πληροφορίας παρέχοντας ειδικές μεθοδολογίες και εργαλεία λογισμικού (Παπαϊωάννου, 2007).
Ο τεράστιος όγκος Βιολογικών δεδομένων που έχει προκύψει από τις διάφορες ερευνητικές δραστηριότητες με επίκεντρο το γονιδίωμα, το πρωτέωμα κ.α. εσωκλείει χρήσιμες πληροφορίες και ένα από τα μεγαλύτερα προβλήματα που αντιμετωπίζει η Βιολογία σήμερα είναι η εξόρυξη, η κατανόηση και η χρήση αυτών των πληροφοριών. Αυτά τα προβλήματα προσπαθεί να λύσει και η Βιοπληροφορική σήμερα.

Πιο συγκεκριμένα οι κύριοι στόχοι της Βιοπληροφορικής είναι:

1. Η οργάνωση των δεδομένων αυτών με κατάλληλο τρόπο σε Βάσεις Δεδομένων ώστε να είναι ελεύθερα προσβάσιμα από τους ερευνητές

2. Η ανάπτυξη υπολογιστικών εργαλείων και εφαρμογή ειδικών αλγορίθμων για την επεξεργασία και το φιλτράρισμα των δεδομένων αυτών καθώς και την ερμηνεία των αποτελεσμάτων που θα προκύψουν από την ανάλυση αυτή ώστε να υπάρξει βιολογικά σημαντική γνώση (Luscombe et al., 2001; Mount, 2004).

Τα εργαλεία Βιοπληροφορικής είναι ειδικά προγράμματα λογισμικού σχεδιασμένα για την εξαγωγή γνώσης από τη πληθώρα των βιολογικών δεδομένων που είναι κατατεθειμένα στις Βάσεις Δεδομένων. Τα πρώτα προγράμματα λογισμικού Βιοπληροφορικής ήταν γραμμένα σε C ή C++ και οι γλώσσες προγραμματισμού όπως η Python ή η Perl χρησιμοποιούνταν ως μέσο αλληλεπίδρασης των προγραμμάτων με τις Βάσεις Δεδομένων. Σήμερα χρησιμοποιούνται και άλλες γλώσσες, όπως η JAVA, για την ανάπτυξη λογισμικού. Πολλοί προγραμματιστές έχουν δημιουργήσει προγράμματα ανοιχτού κώδικα από τα οποία τα πιο ευρέως χρησιμοποιούμενα είναι τα: BioPerl, BioPython και BioJava.

Πρωτεωμική

Η πρωτεωμική είναι ο επιστημονικός κλάδος που αφορά την ανάλυση και μελέτη σύνθετων πρωτεϊνικών μιγμάτων, εστιάζοντας στη δομή και τη λειτουργία των πρωτεϊνών (Banks et al., 2000; Cash, 2002). Πιο συγκεκριμένα, είναι η μελέτη της σύνθεσης, της δομής, της λειτουργίας και των αλληλεπιδράσεων όλων των πρωτεϊνών μέσα σε ένα κύτταρο (Liebler, 2001). Υπάρχει μία άρρηκτη σύνδεση μεταξύ Γονιδιωματικής και Πρωτεωμικής δεδομένου ότι η έκφραση του γονιδιώματος του κυττάρου αποτελεί το πρωτέωμα.
Πριν από τα μέσα του 1994, η λέξη «πρωτέωμα» δεν υπήρχε. Ο Marc Wilkins, ένας φοιτητής στο Πανεπιστήμιο Macquarie της Αυστραλίας, προσπαθούσε να βρει τις σωστές λέξεις όταν τελείωνε τη διδακτορική του διατριβή για να περιγράψει το σύνολο των πρωτεϊνών. Στην εργασία του χρησιμοποιούσε συνεχώς τη φράση: «όλες οι πρωτεΐνες που εκφράζονται από ένα γονιδίωμα, κύτταρο ή ιστό», και επειδή θεωρούσε αυτή την επανάληψη κουραστική και άκομψη προσπάθησε να βρει μία μόνο λέξη που να εμπεριέχει το νόημα όλης της φράσης.
Έτσι, άρχισε να παίζει με τις λέξεις ώστε να δημιουργήσει μία καινούργια που θα εμπεριέχει την έννοια του «πρωτεϊνικού ισοδύναμου του γονιδιώματος σε μια χρονική στιγμή». Μετά την απόρριψη «proteinome» και «protome», κατέληξε στο «proteome» (πρωτέωμα) που αποτελεί μια σύντμηση των λέξεων «protein» και «genome» (PROTein complement of the genOME). Τον Σεπτέμβριο του 1994, ο Wilkins χρησιμοποίησε τη λέξη «πρωτέωμα» σε επιστημονικό συνέδριο στην Ιταλία, και η λέξη έμεινε (Huber, 2003; Wasinger et al., 1995).
Είναι σημαντικό να αναφέρεται το «σε μια χρονική στιγμή» επειδή το γονιδίωμα ενός οργανισμού είναι μοναδικό αλλά παράγει πολλές και διαφορετικές πρωτεΐνες ανάλογα με τον κυτταρικό τύπο, τις περιβαλλοντικές συνθήκες και τα ερεθίσματα που λαμβάνει το κύτταρο και τη θέση που βρίσκεται επομένως και ανάλογα με τη χρονική στιγμή που μελετάται.

Δυναμική φύση του πρωτεώματος

Παρόλο που το γονίδιο αποτελεί τη θέση αφετηρίας της παραγωγής μιας πρωτεΐνης, η σταθερότητα των παραγόμενων βιομορίων (mRNA κ.α.), η αποτελεσματικότητα της διαδικασίας της μετάφρασης, οι μετα-μεταφραστικές τροποποιήσεις και οι αλλαγές στο περιβάλλον του κυττάρου είναι σημαντικοί παράγοντες στην ρύθμιση.
Κυτταρικές και περιβαλλοντικές αλλαγές όπως το pH, οι συνθήκες υποξίας (έλλειψης οξυγόνου) και η χορήγηση φαρμάκων, είναι δυνατόν να μεταβάλλουν τη σύσταση του πρωτεώματος. Αυτό έχει ως αποτέλεσμα την αδυναμία πρόβλεψης των πρωτεϊνικών μορίων ακόμη και όταν είναι γνωστή η αλληλουχία του γονιδίου από το οποίο προέρχεται. Το εναλλακτικό μάτισμα αυξάνει τη γενετική πολυπλοκότητα οδηγώντας σε παραγωγή, από ένα γονίδιο, ενός συνόλου πρωτεϊνών με διαφορετικές ιδιότητες. Ο αριθμός των πρωτεϊνών που κωδικοποιούνται από τα γονίδια είναι άγνωστος αλλά εκτιμάται πως το κάθε γονίδιο μπορεί να κωδικοποιήσει από 10 μέχρι 20 διαφορετικούς τύπους πρωτεϊνών (Banks et al., 2000).
Εκτός από το εναλλακτικό μάτισμα, ένας άλλος λόγος δημιουργίας πρωτεϊνικής ποικιλομορφίας στο κύτταρο είναι οι μετα-μεταφραστικές τροποποιήσεις (PTM: post-translational modifications). Έχουν βρεθεί πάνω από 100 τέτοιου είδους τροποποιήσεις οι οποίες παράγουν πολλαπλά

ισόμορφα μίας πρωτεΐνης και δύναται να εμφανίζουν διαφορετικές βιολογικές λειτουργίες.

Μετα-μεταφραστικές τροποποιήσεις

Οι μετα-μεταφραστικές τροποποιήσεις (ΜΜΤ) (Post-Translational modifications - PTMs) των πρωτεϊνών είναι μια φυσιολογική διεργασία, η οποία επιτελείται είτε ενζυμικά είτε αυθόρμητα. Αναφέρεται σε ομοιοπολικές τροποποιήσεις πλευρικών ομάδων αμινοξικών καταλοίπων των πρωτεϊνών με προσθήκη συμπληρωματικών ομάδων (λιπιδίων, υδατανθράκων, φωσφορυλιώσεων κ.τ.λ.). Οι ενζυμικές μετα-μεταφραστικές τροποποιήσεις καταλύονται από εξειδικευμένα ένζυμα, τα οποία αναγνωρίζουν ορισμένα μοτίβα στην αλληλουχία των πρωτεϊνών και εντοπίζονται σε συγκεκριμένα κυτταρικά διαμερίσματα. Τα PTMs παίζουν σημαντικό ρόλο σε όλες σχεδόν τις βιολογικές διαδικασίες. Κάποιες από αυτές τις τροποποιήσεις είναι ιδιαίτερα σημαντικές για τη ρύθμιση της λειτουργίας των πρωτεϊνών, τη θέση τους στο κύτταρο ή ακόμη και για την αποικοδόμησή τους. Η εύρεση και ανάλυση αυτών των τροποποιήσεων θα βοηθούσε στην κατανόηση των μηχανισμών που κρύβονται πίσω από λειτουργίες όπως η κυτταρική επικοινωνία, η κυτταρική διαμερισματοποίηση, αλλαγές στην πρωτεϊνική δραστηριότητα και σταθερότητα κ.τ.λ.. Πιο συγκεκριμένα, οι ακετυλιώσεις συμβάλλουν στο ξεδίπλωμα του DNA προκειμένου να δεσμευτούν πάνω σε αυτό οι πρωτεϊνικοί μεταγραφικοί παράγοντες ενώ οι φωσφορυλιώσεις συμβάλλουν στη επικοινωνία μεταξύ των βιομορίων του κυττάρου.

Φωσφορυλίωση πρωτεϊνών

Η φωσφορυλίωση αποτελεί την πιο συνηθισμένη και άφθονη μετα- μεταφραστική τροποποίηση που καταλύεται ενζυμικά. Η τροποποίηση αυτή είναι αντιστρεπτή, καταλύεται από τα ένζυμα κινάσες και αφορά την προσθήκη μίας φωσφορικής ομάδας σε αμινοξέα τα οποία συνήθως είναι σερίνες, θρεονίνες ή τυροσίνες. Η προσθήκη μιας φωσφορικής ομάδας σε μια πρωτεΐνη μπορεί να επηρεάσει πολλές ιδιότητές της, συμπεριλαμβανομένων της ικανότητας αναδίπλωσης, της λειτουργίας, της αλληλεπίδρασης με άλλες πρωτεΐνες, του κυτταρικού εντοπισμού και της αποικοδόμησής της με αποτέλεσμα να παίζει ουσιαστικό ρόλο στην ρύθμιση σχεδόν όλων των βιολογικών λειτουργιών, περιλαμβανομένου του πολλαπλασιασμού, της διαφοροποίησης, της απόπτωσης και της διακυτταρικής επικοινωνίας (Hunter, 2000).
Σε μια δεδομένη χρονική στιγμή, όλα τα αντίγραφα μιας δεδομένης πρωτεΐνης δεν είναι στη φωσφορυλιωμένη κατάσταση. Ως εκ τούτου, απαιτούνται εξαιρετικά ευαίσθητες τεχνικές για την απομόνωση, ανίχνευση και τον ποσοτικό προσδιορισμό των χαμηλών σε συγκέντρωση θέσεων φωσφορυλίωσης.

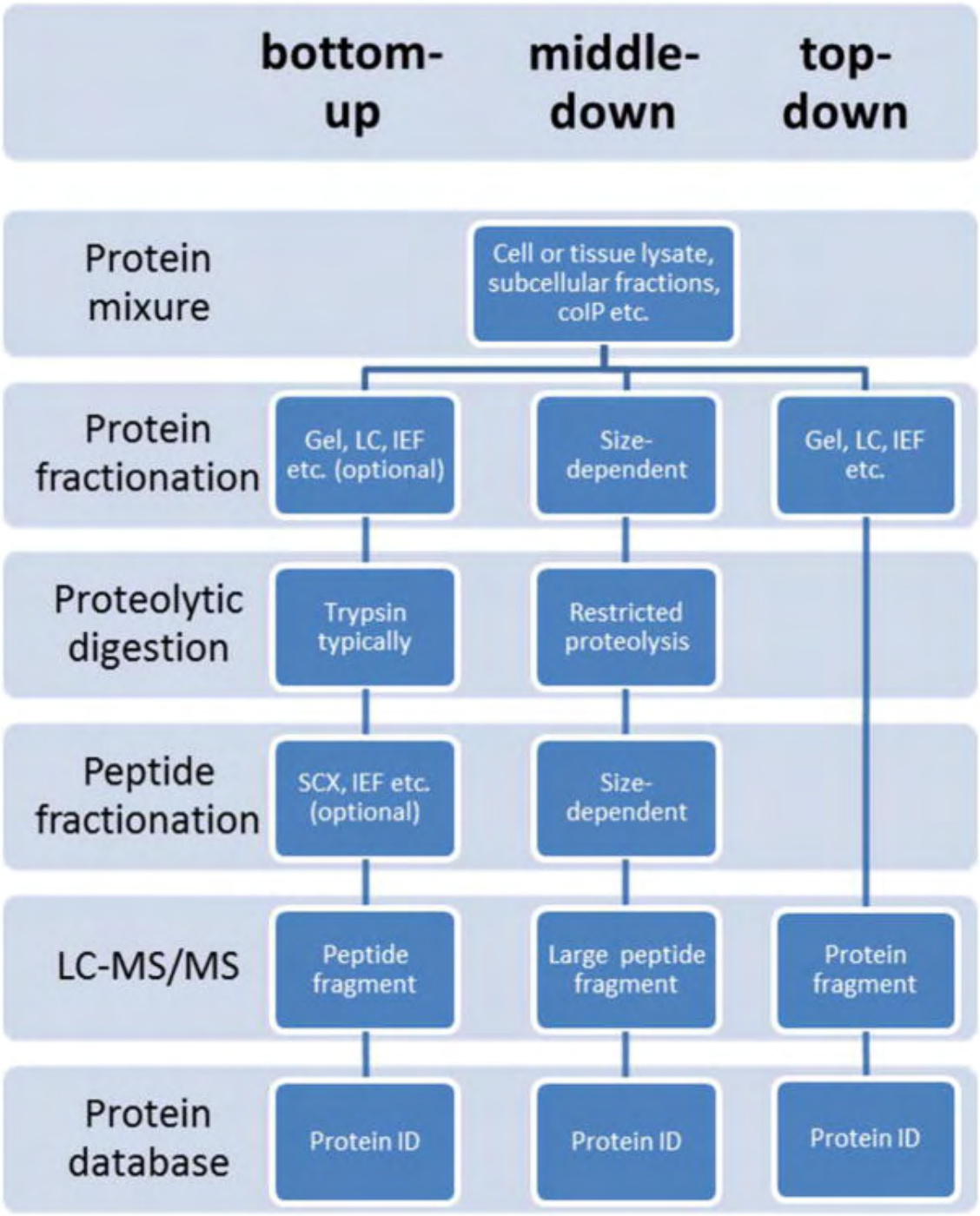
Τεχνικές Πρωτεωμικής

Ο πρωταρχικός στόχος κάθε πρωτεωμικής ανάλυσης είναι η ταυτοποίηση των πρωτεϊνών που βρίσκονται στο εξεταζόμενο δείγμα. Τα πρώτα χρόνια της ανάπτυξης των Proteomics, η ανάλυση και η μελέτη των πρωτεϊνικών μιγμάτων πραγματοποιούνταν αποκλειστικά με χρήση της πειραματικής τεχνικής two-dimensional polyacrylamide gel electrophoresis (2D PAGE – 2DE), η οποία στη συνέχεια συνδυάστηκε με τη χρήση τεχνικών Φασματομετρίας Μάζας (Mass Spectrometry – MS) (Hanash, 2003). Οι τεχνικές αυτές, σε συνδυασμό με εργαλεία Βιοπληροφορικής, έχουν καθιερωθεί πια ως οι βασικές συνιστώσες της κλασσικής πρωτεωμικής μεθοδολογίας (“the classical proteomics methodology”), με αποτέλεσμα να παίζουν καθοριστικό ρόλο στην πρωτεωμική ανάλυση (Beranova-Giorgianni,2003).

Η φασματομετρία μάζας (mass spectrometry-MS) έχει βελτιωθεί σημαντικά την τελευταία δεκαετία. Είναι μία τεχνική υψηλής ευαισθησίας αναγνώρισης και ποσοτικοποίησης βιομορίων (κυρίως πρωτεϊνών και πεπτιδίων). Πιο συγκεκριμένα, βασίζεται στον ιονισμό ατόμων ή μορίων, το διαχωρισμό των παραγόμενων ιόντων-ιονισμένων θραυσμάτων σύμφωνα με τη καταγραφή της σχετικής έντασης του ιοντικού ρεύματος που αντιστοιχεί σε κάθε λόγο μάζας προς φορτίο (m/z) και την καταγραφής της σχετικής αφθονίας τους.

Κατηγοριοποίηση τεχνικών Πρωτεωμικής

Ένας τρόπος κατηγοριοποίησης των διάφορων τεχνικών πρωτεωμικής ανάλυσης με σκοπό την ταυτοποίηση των πρωτεϊνών με τη χρήση φασματομετρία μάζας είναι η διάκριση τους σε «bottom-up» και «top-down» proteomics (βλέπε εικόνα 1).
Η «bottom-up» ανάλυση πρωτεϊνών αναφέρεται στον χαρακτηρισμό των πρωτεϊνών με ανάλυση των πεπτιδίων που δημιουργούνται μετά από πέψη της πρωτεΐνης με χρήση ειδικών πρωτεολυτικών ενζύμων. Οι απομονωμένες πρωτεΐνες ή τα μίγματα πρωτεϊνών πρώτα πέπτονται πλήρως σε πεπτίδια και στη συνέχεια τα πεπτίδια αναλύονται με φασματομετρία μάζας πολλών διαστάσεων (MS_n). Η ταυτότητα των πεπτιδίων συνήθως εξακριβώνεται μετά από σύγκριση θεωρητικών φασμάτων μάζας που δημιουργούνται μετά από εικονική πέψη κατάλληλης Βάσης Δεδομένων με τα πειραματικά φάσματα μάζας του δείγματος (Link et al., 1999; Wolters et al., 2001; Yates, 2004).
Η ανάλυση των πρωτεϊνών μέσω των πεπτιδίων που δημιουργούνται μετά από ενζυμική πρωτεόλυση παρακάμπτει μερικές από τις προκλήσεις που συνδέονται με τον διαχωρισμό και τον ιονισμό των άθικτων πρωτεϊνών στη φασματομετρία μάζας. Πιο συγκεκριμένα, το κυτταρικό πρωτεϊνικό μίγμα είναι ένα εξαιρετικά ετερογενές με ποικίλες φυσικοχημικές ιδιότητες. Επομένως μία υποθετική σκόπιμη αύξηση της πολυπλοκότητας του μίγματος πριν από την ανάλυση φαίνεται ότι μειώνει την αποτελεσματικότητα της διαδικασίας. Ωστόσο, η επιλεκτική πέψη με κάποια πρωτεάση φαίνεται να ομαλοποιεί και να διαχωρίζει τη βιοχημική ετερογένεια των πρωτεϊνών και μπορεί, στην πραγματικότητα, να δημιουργήσει ένα λιγότερο ετερογενές μίγμα όταν υπάρχουν πολλές ισομορφές πρωτεϊνών και μετα-μεταφραστικές τροποποιήσεις. Τέλος, οι μέθοδοι φασματομετρίας μάζας πεπτιδίων είναι τόσο θεωρητικά όσο και πειραματικά πιο απλές απ’ ότι σε επίπεδο πρωτεΐνης, καθιστώντας τις μεθόδους για την ανάλυση των πρωτεϊνών σε επίπεδο πεπτιδίου ερευνητικά και εργαστηριακά πιο προσιτές (Zhang et al., 2013).

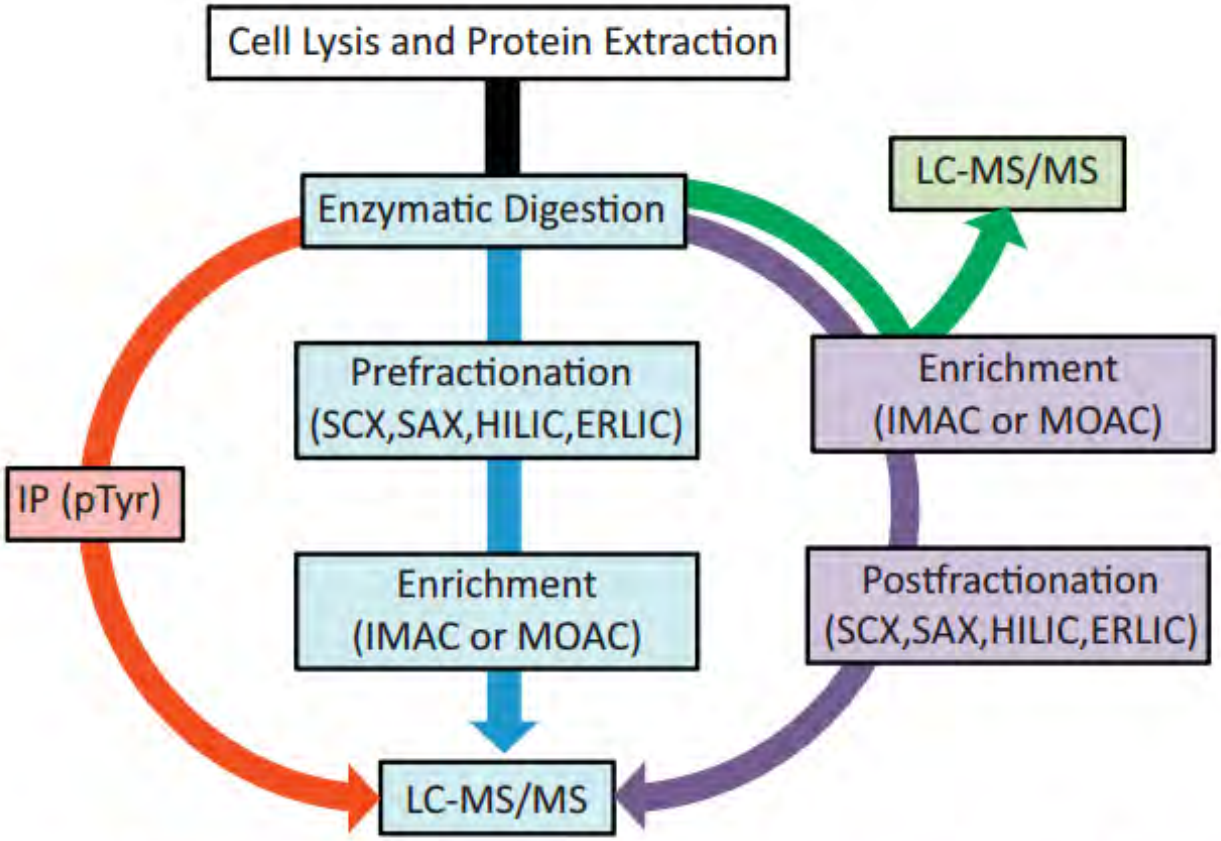


Εικόνα 1:Στρατηγικές Πρωτεωμικής.
(Zhang et al., 2013)

Η «bottom-up» ανάλυση πρωτεϊνών είναι αυτή που χρησιμοποιείται κυρίως στις διάφορες επιστημονικές έρευνες και βασίζεται στην δυνατότητα να μετατρέπεται η πρωτεΐνη σε πεπτίδια, να υπολογίζεται η ακολουθία των πεπτιδίων και στη συνέχεια να ταυτοποιείται η πρωτεΐνη. Όταν η ανάλυση «bottom-up» εκτελείται σε ένα μίγμα πρωτεϊνών στο οποίο γίνεται πέψη και τα παραγόμενα πεπτίδια διαχωρίζονται με υγρή χρωματογραφία πριν την εισαγωγή τους στο φασματογράφο μάζας τότε ονομάζεται «shotgun proteomics».

Τεχνικές ανάλυσης φωσφοπρωτεώματος

Οι τρέχουσες στρατηγικές για τις αναλύσεις φωσφοπρωτεωμικής συνήθως περιλαμβάνουν τέσσερα σημαντικά βήματα (βλέπε Εικόνα 2) τα οποία είναι τη λύση του κυττάρου και η εκχύλιση των πρωτεϊνών, ο διαχωρισμός των πρωτεϊνών και πεπτιδίων, ο εμπλουτισμός φωσφοπεπτιδίων και οι αναλύσεις φασματομετρίας μάζας. Αυτά τα στάδια μπορεί να προσαρμοστούν σύμφωνα με το πειραματικό πρωτόκολλο της κάθε μελέτης για να αποκτήσουν πρόσθετες πληροφορίες σχετικά με τις πρωτεΐνες όπως για παράδειγμα τις υποκυτταρικές μετατοπίσεις των πρωτεϊνών ή το υποσύνολο φωσφοπεπτιδίων που περιλαμβάνει ένα συγκεκριμένο μοτίβο αλληλουχίας ή συγκεκριμένα φωσφορυλιωμένα αμινοξέα (π.χ. φωσφοτυροσίνης).



Workflow	Specificity	Protein, mg	Number of IDs	Refs.
I	pY only	> 10 mg	+	[64,65]
II	pS/T/Y	5-10 mg	++++	[19,95-97]
III	pS/T/Y	0.25-0.5 mg	++	[75]
IV	pS/T/Y	2-4 mg	+++	[100]

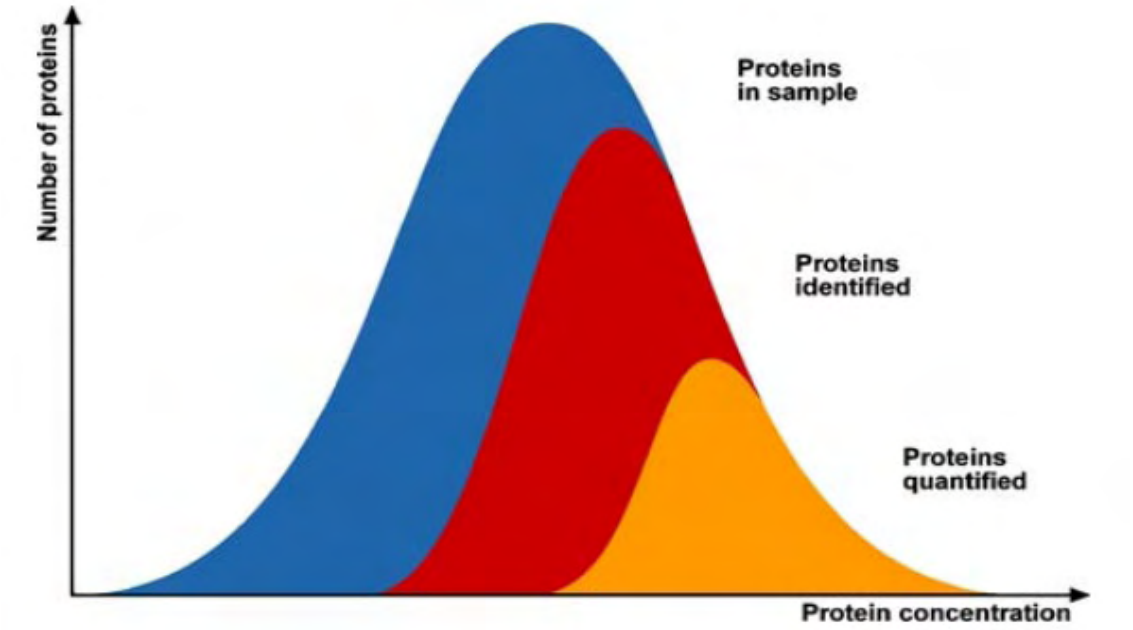
Εικόνα 2:Ροής εργασίας φωσφοπρωτεωμικής.

Τα εκχυλίσματα των κυττάρων υφίστανται πέψη, και τα αντίστοιχα πρωτεολυτικά πεπτίδια μπορούν να διαχωριστούν με LC ή / και με χρωματογραφία συγγένειας ώστε να ομαδοποιηθούν σε μικρότερες και πιο εύκολες να αναλυθούν συλλογές πεπτιδίων (Kanshin et al., 2012).

Ταυτοποίηση πεπτιδίων και των θέσεων φωσφορυλίωσης

Σήμερα, η συντριπτική πλειοψηφία των πρωτεϊνικών δεδομένων παράγονται με φασματομετρία μάζας. Τα μηχανήματα αυτά και οι διαφορετικές ροές εργασίας που ακολουθούνται έχουν το κοινό χαρακτηριστικό ότι δημιουργούν εκατοντάδες ως και δεκάδες χιλιάδες ιοντικά φάσματα θραυσμάτων πεπτιδίων. Η ανάθεση αυτού του τεράστιου όγκου φασμάτων μάζας σε πεπτιδικές ακολουθίες και στη συνέχεια σε πρωτεΐνες τις οποίες αντιπροσωπεύουν και από τις οποίες προέρχονται παρουσιάζει πολύπλοκες υπολογιστικές και στατιστικές προκλήσεις. Είναι απαραίτητο για τον τομέα της Πρωτεωμικής να αναπτύξει και να εφαρμόσει γενικά εργαλεία και λύσεις σε αυτά τα προβλήματα ώστε να παρέχουν ακριβή και επαναλήψιμα αποτελέσματα.

Αξίζει να σημειωθεί πως παρά την πρωτοφανή εξέλιξη των εργαλείων αλληλούχισης των πεπτιδίων-πρωτεϊνών (φασματόμετρα μάζας) και των διάφορων ερευνητικών πρωτοκόλλων που χρησιμοποιούνται, ο εντοπισμός και η ποσοτικοποίηση όλων των πρωτεϊνών σε ένα βιολογικό σύστημα εξακολουθεί να είναι αποτελεί μια πρόκληση.



Εικόνα 3:Σχηματική αναπαράσταση του κλάσματος ενός πρωτεώματος που μπορεί να προσδιοριστεί-αλληλουχηθεί ή να ποσοτικοποιηθεί με χρήση φασματόμετρου μάζας.
(Bantscheff et al., 2007)

Όπως φαίνεται και στην Εικόνα 3, οι πρωτεΐνες ενός κυττάρου καλύπτουν ένα ευρύ φάσμα έκφρασης και οι σημερινές φασματομετρικές τεχνικές μάζας μπορούν να καλύψουν-αλληλουχήσουν μόνο ένα μέρος αυτού. Επίσης, λόγω της περιορισμένης ποιότητας των δεδομένων, μόνο ένα κλάσμα όλων των ταυτοποιημένων πρωτεϊνών μπορεί να ποσοτικοποιηθεί με αξιοπιστία (Bantscheff et al., 2007).
Επομένως, η φασματομετρία μάζας πρωτεϊνών με πέψη αυτών για την ταυτοποίησή τους εξαρτάται από τον καλό πειραματικό σχεδιασμό, αλλά έχει να αντιμετωπίσει και μια σειρά από προκλήσεις που παρουσιάζει η ανάλυση των δεδομένων για την αποφυγή κυρίως ψευδών θετικών αποτελεσμάτων και τη μείωση των ψευδώς αρνητικών.
Σε αντίθετη περίπτωση, εισάγονται και διαδίδονται σφάλματα στη βιβλιογραφία, γεγονός που καθιστά δύσκολη την αξιολόγηση των συμπερασμάτων των ερευνών αυτών από τους κριτικούς των περιοδικών και των αναγνωστών με αποτέλεσμα να καθιστά ουσιαστικά άχρηστες τις Βάσεις Δεδομένων που περιέχουν τέτοιου είδους αμφισβητήσιμα δεδομένα (Carr et al., 2004).

Συστήματα αξιολόγησης πεπτιδικών και πρωτεϊνικών ταυτοποιήσεων

Το κομμάτι της πληροφορικής στη ροή εργασίας των πειραμάτων πρωτεωμικής με χρήση φασματόμετρου μάζας είναι πολύ σημαντικό για τη σωστή ανάλυση των δεδομένων, και μια ευρεία ποικιλία υπολογιστικών εργαλείων έχουν προκύψει για το σκοπό αυτό (Nesvizhskii, 2010).
Η τυπική ροή εργασιών πληροφορικής στα συγκεκριμένα πειράματα μπορεί να συνοψιστεί σε λίγα βήματα:

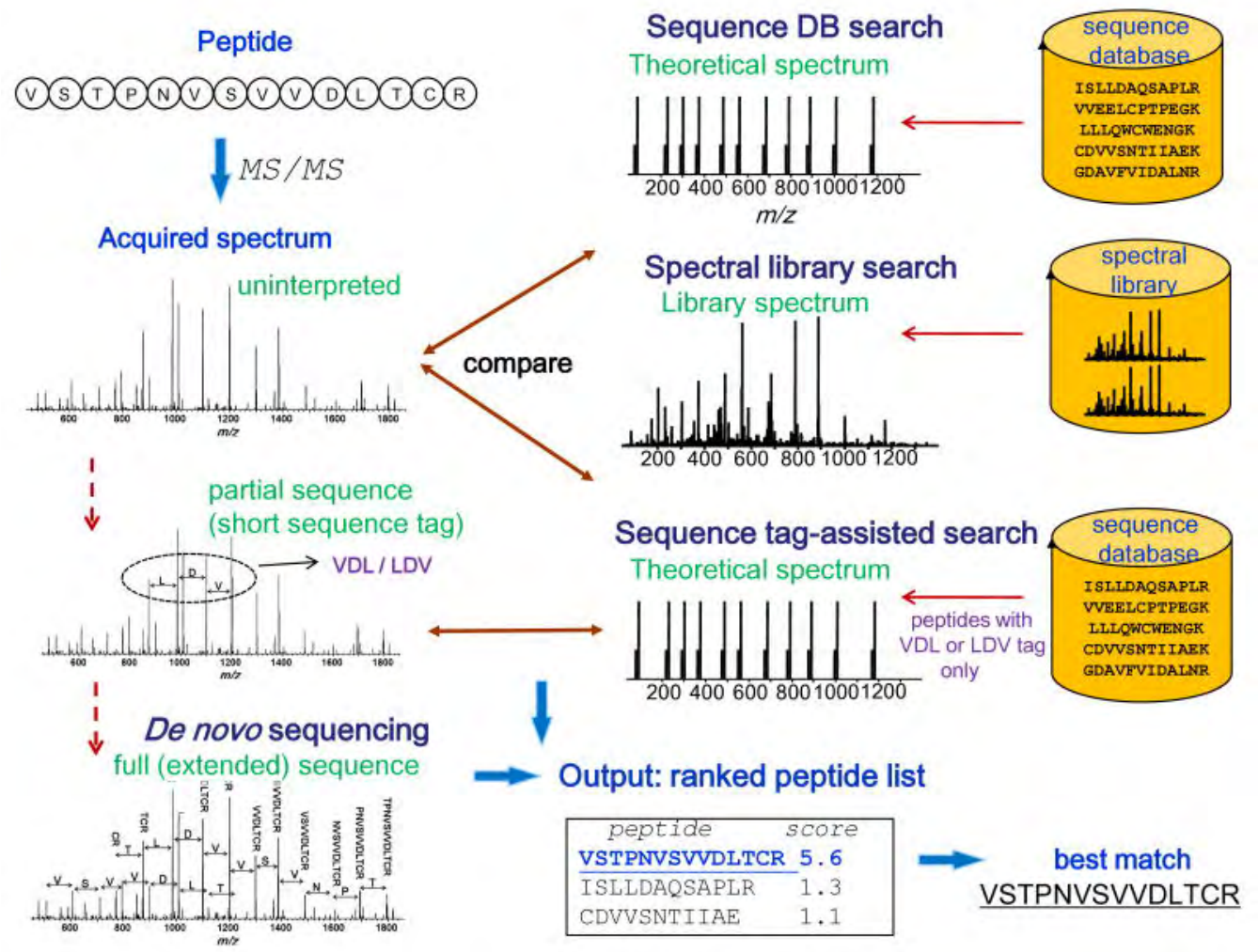
- μετατροπή της μορφής των αποτελεσμάτων της φασματομετρίας μάζας από ιδιόκτητο λογισμικό σε λογισμικό Ανοιχτού Κώδικα
- υψηλής απόδοσης ερμηνεία των φασμάτων MS / MS με χρήση μηχανής αναζήτησης
- στατιστική επικύρωση των αποτελεσμάτων μέσω του False Discovery Rate (FDR) με ένα επιλεγμένο κατώφλι
- ανακάλυψη των πρωτεϊνών που υπάρχουν στο δείγμα, και η αφθονία τους στο δείγμα αν είναι κάτι που αναζητείται, με βάση τις ταυτοποιήσεις των πεπτιδίων.

Οι πληροφορίες για τα πειράματα πρωτεωμικής με διαδοχική φασματομετρία μαζών (MS / MS) βρίσκονται με τη μορφή των φασμάτων θραυσμάτων ιόντων (φάσμα MS / MS) πεπτιδίων που έχουν παραχθεί κατά τον κατακερματισμό των πρωτεϊνών. Η σωστή σύνδεση-ταίριασμα ενός τέτοιου φάσματος σε μία πεπτιδική αλληλουχία ονομάζεται Peptide Spectrum Match ή PSM και αποτελεί ένα κεντρικό βήμα στην επεξεργασία αυτών των δεδομένων.
Ένα από τα πιο σημαντικά και δύσκολα καθήκοντα που αντιμετωπίζουν οι ερευνητές είναι η σωστή επιλογή αυτών των PSMs που παράγονται απομακρύνοντας τις ψευδώς θετικές αντιστοιχίσεις που προσθέτουν θόρυβο στα αποτελέσματα.
Ένας μεγάλος αριθμός υπολογιστικών προσεγγίσεων και εργαλείων λογισμικού έχουν δημιουργηθεί για να γίνεται αυτοματοποιημένη ανάθεση και σύνδεση των πεπτιδικών αλληλουχιών με τα αντίστοιχα φάσματα θραυσμάτων ιόντων. Τα υπολογιστικά αυτά εργαλεία βρίσκονται υπό την μορφή μηχανών αναζήτησης, δηλαδή αλγόριθμων που προσπαθούν να ερμηνεύσουν τα MS / MS φάσματα.
Υπάρχουν πολλές και διαφορετικές μηχανές αναζήτησης, που είναι εμπορικά διαθέσιμες και η καθεμία χρησιμοποιεί μία μοναδική τεχνική-αλγόριθμο για την αναγνώριση του και μία διαφορετική μορφή αρχείου για τα δεδομένα εξόδου (Deutsch, 2012).
Υπάρχουν τρεις κύριοι τύποι μηχανών αναζήτησης (βλέπε Εικόνα 4):

- Οι μηχανές αναζήτησης ακολουθίας (sequence search engines) όπως η X!Tandem (Craig and Beavis, 2004), η Mascot της Matrix Science Ltd που χρησιμοποιεί έναν αποκλειστικό αλγόριθμο βαθμολόγησης ο οποίος βασίζεται στον αλγόριθμο MOWSE (Perkins et al., 1999), η SEQUEST (Eng et al., 1994), το MyriMatch (Tabb et al., 2007), το MS-GFDB (Kim et al., 2010), και η OMSSA (Open Mass Spectrometry Search Algorithm) που είναι μια αποτελεσματική μηχανή αναζήτησης που προσδιορίζει MS-MS φάσματα πεπτιδίων με βάση στατιστικών μεθόδων (Barsnes et al., 2009). Αυτές οι μηχανές αναζήτησης προσπαθούν να ταιριάξουν τα φάσματα που αποκτήθηκαν με θεωρητικά φάσματα που παράγονται από πιθανές αλληλουχίες πεπτιδίων που περιέχονται σε πρωτεϊνικές Βάσεις Δεδομένων.
- Οι μηχανές αναζήτησης βιβλιοθηκών φασμάτων (spectral library search engines) όπως το SpectraST (Lam et al., 2007), το X!Hunter (Craig et al., 2006), και το BiblioSpec (Frewen et al., 2006), οι οποίες προσπαθούν να ταιριάξουν τα παρατηρούμενα φάσματα με μια βιβλιοθήκη που διαθέτει φάσματα που έχουν προηγουμένως παρατηρηθεί και προσδιορισθεί.
- Οι μηχανές αναζήτησης de novo όπως η PEAKS (Ma et al., 2003), η PepNovo (Frank and Pevzner, 2005), και η Lutefisk (Taylor and Johnson, 2001), οι οποίες προσπαθούν να ταυτοποιήσουν πεπτίδια βασισμένες αποκλειστικά στα μοτίβα κορυφών των ιόντων που εμφανίζονται

στα φάσματα MS / MS, χωρίς τη χρήση αλληλουχιών αναφοράς ή φασμάτων που έχουν ταυτοποιηθεί στο παρελθόν (Pevtsov et al., 2006).

- Οι μηχανές αναζήτησης υβριδικών προσεγγίσεων, όπου χρησιμοποιούν στοιχεία της de novo αναζήτησης με εξαγωγή μικρών αλληλουχιών-ετικετών, 3-5 υπολειμμάτων σε μήκος (Dasari et al., 2010), και στη συνέχεια πραγματοποιείται έρευνα σε Βάσεις Δεδομένων στις οποίες ο κατάλογος των υποψήφιων πεπτιδίων περιορίζεται σε εκείνα τα πεπτίδια μόνο που περιέχουν μία από αυτές τις ετικέτες αλληλουχίας (Creasy and Cottrell, 2002; Mann and Wilm, 1994). Μερικά παραδείγματα τέτοιων μηχανών αναζήτησης είναι η InSpecT (Tanner et al., 2005) και η PEAKS-DB (Zhang et al., 2012).



Εικόνα 4:Στρατηγικές ταυτοποίησης πεπτιδίων.
(Nesvizhskii, 2010)

Υποθέτοντας ότι ένα δεδομένο φάσμα MS / MS περιέχει επαρκείς πληροφορίες, δηλαδή έναν εύλογο αριθμό κορυφών θραύσματος ιόντος σε ικανοποιητικό λόγο σήματος προς θόρυβο (signal to noise ratio), οι πιθανοί λόγοι αποτυχίας ταιριάσματος είναι:

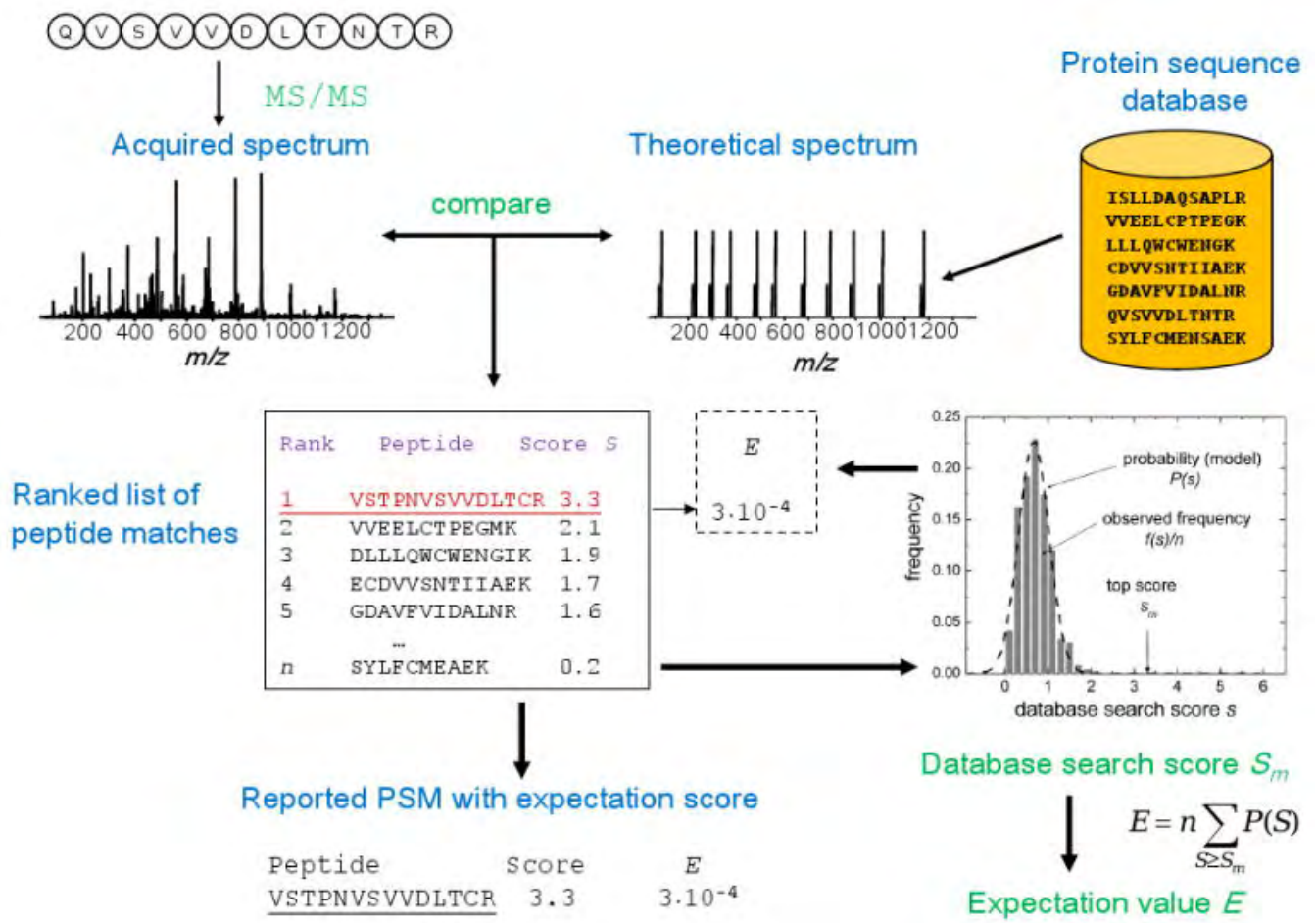
- Υποτιμημένο σφάλμα μέτρησης της μάζας
- Εσφαλμένος προσδιορισμός του φορτίου του προδρόμου μορίου
- Μη εξειδίκευση του ενζύμου
- Ύπαρξη χημικών και μετα-μεταφραστικών τροποποιήσεων που δεν υπολογίστηκαν
- Όταν η πεπτιδική αλληλουχία, επομένως και η αντίστοιχη πρωτεΐνη, δεν υπάρχει στη Βάση Δεδομένων

Γενικά, για τις μελέτες Πρωτεωμικής μεγάλης κλίμακας η αναζήτηση σε Βάσεις Δεδομένων παραμένει η πιο συχνά χρησιμοποιούμενη μέθοδος αναγνώρισης πεπτιδίου. Ωστόσο, οι άλλες στρατηγικές παρέχουν ελκυστικές εναλλακτικές λύσεις σε συγκεκριμένες καταστάσεις (Nesvizhskii et al., 2007) όπως για παράδειγμα στην περίπτωση που δεν βρεθεί αντιστοιχία μεταξύ παρατηρούμενου και προβλεπόμενου φάσματος, διότι η πρωτεΐνη δεν είναι παρούσα στη Βάση Δεδομένων. Τότε η «de novo» ταυτοποίηση των πεπτιδίων εξακολουθεί να είναι μια πολύτιμη μέθοδος, που βασίζεται σε γνωστούς κανόνες που διέπουν τον κατακερματισμό πεπτιδίων. Ένας άλλος τρόπος ταυτοποίησης αυτών των πεπτιδίων είναι η σύγκρισή τους με πεπτίδια που παρουσιάζουν ομολογία με αυτά και προέρχονται από πρωτεΐνες της ίδιας Βάσης Δεδομένων ή από Βάσεις Δεδομένων συγγενικών οργανισμών (Perkins et al., 1999).
Ωστόσο, όλες οι μηχανές αναζήτησης για την ταυτοποίηση των πεπτιδίων ακολουθούν μία κοινή λογική. Οι αλγόριθμοι αυτοί επιτρέπουν στο χρήστη να επιλέξει την κατάλληλη για τη μελέτη Βάση Δεδομένων, το είδος του οργανισμού απ' όπου προέρχεται το δείγμα ώστε να περιοριστεί ο όγκος δεδομένων από όπου θα γίνει η αναζήτηση, το πρωτεολυτικό ένζυμο που χρησιμοποιήθηκε και τις μετα-μεταφραστικές τροποποιήσεις που αναζητάει και αλλάζουν το μοριακό βάρος των πεπτιδίων. Στη συνέχεια, οι αλγόριθμοι αυτοί μετρούν την ομοιότητα μεταξύ των πειραματικών και των θεωρητικών φασμάτων μάζας. Τα θεωρητικά φάσματα μάζας έχουν δημιουργηθεί μετά από εικονική πέψη των πρωτεϊνών της Βάσης Δεδομένων του οργανισμού εκείνου που ερευνάται με το ίδιο πρωτεολυτικό ένζυμο που χρησιμοποιήθηκε στο εκάστοτε πείραμα.
Η παραπάνω διαδικασία είναι γνωστή ως Peptide Mass Fingerprinting και στηρίζεται στο γεγονός ότι αν μια πρωτεΐνη κοπεί με προβλέψιμο τρόπο, τα μεγέθη των πεπτιδίων που σχηματίζονται πρέπει να δημιουργούν ένα «δακτυλικό αποτύπωμα» για την εν λόγω πρωτεΐνη. Μία πρωτεΐνη μπορεί να οριστεί ως ένα σύνολο από αμινοξέα που είναι διευθετημένα με μια συγκεκριμένη ακολουθία, η οποία καθορίζει τις ιδιότητες και τη λειτουργία της. Παρά το γεγονός ότι ορισμένες πρωτεΐνες μπορεί να έχουν υψηλό βαθμό ομοιότητας με άλλες πρωτεΐνες, μερικά τμήματα αλληλουχίας τους θα είναι μοναδικά. Επομένως, αν κάθε πρωτεϊνική αλληλουχία σε μια Βάση Δεδομένων μπορεί να κοπεί με τον ίδιο τρόπο που έχει κοπεί το πειραματικό δείγμα, το δακτυλικό αποτύπωμα θα χρησιμεύσει για τον προσδιορισμό των πρωτεϊνών του δείγματος.
Όταν η μάζα των πεπτιδίων που έχουν δημιουργηθεί *in silico* από την επιλεγμένη Βάση Δεδομένων ταιριάζει με τις μετρούμενες πειραματικές μάζες, τότε τα συγκεκριμένα πεπτίδια ταξινομούνται με βάση τη μάζα τους και αυτά που ανήκουν στο επιλεγμένο όριο εμπιστοσύνης αποτελούν τα «hits». Στη συνέχεια για κάθε «hit» υπολογίζεται μια τιμή (score) η οποία χρησιμοποιείται για την αξιολόγηση της πρωτεΐνης που ταυτοποιείται επειδή τα πεπτίδια μπορεί είτε να έχουν ανατεθεί σε μία μοναδική πρωτεΐνη (1 match) είτε σε περισσότερες από μία (>1 matches).
Η βαθμολογία-σκορ αναζήτησης μετρά το βαθμό ομοιότητας μεταξύ του πειραματικού φάσματος και των διαφόρων θεωρητικών φασμάτων, και ως εκ τούτου λειτουργεί ως κύρια παράμετρος διάκρισης για το διαχωρισμό σωστών από λανθασμένων ταυτίσεων. Συνήθως, μόνο το πεπτίδιο με την καλύτερη αντιστοιχία βαθμολόγησης χρησιμοποιείται στο επόμενο στάδιο της στατιστικής ανάλυσης.

Αν δεν υπάρχουν «hits» για βαθμολογίες πάνω από το όριο εμπιστοσύνης, τότε η πρωτεΐνη από την οποία προέρχεται το πεπτίδιο παραμένει άγνωστη (McHugh and Arthur, 2008). Βέβαια, πρέπει να σημειωθεί ότι ένα χαμηλό σκορ, δηλαδή μια μικρή πιθανότητα εντοπισμού, δεν δηλώνει την απουσία του συγκεκριμένου πεπτιδίου αλλά απλώς την έλλειψη αποδεικτικών στοιχείων για την παρουσία του στο δείγμα.

Η ανάπτυξη των συστημάτων βαθμολόγησης για ταυτοποίηση πρωτεϊνών και πεπτιδίων ξεκίνησε με την προσαρμογή των καλά ανεπτυγμένων γενικών στατιστικών μεθόδων που χρησιμοποιούνται σε πολλούς τομείς της επιστήμης και της τεχνολογίας, όπως το θεώρημα της αλληλοσυσχέτισης (ετεροσυσχέτισης, cross-correlation), την πιθανότητα κατά Bayes, τον αλγόριθμο μεγιστοποίησης προσδοκίας (Expectation Maximization) και την μηχανική μάθηση (machine learning). Σταδιακά όλο και πιο εξελιγμένα και βελτιωμένα συστήματα βαθμολόγησης κατασκευάζονται με την εισαγωγή νέων στατιστικών μεθόδων (McHugh and Arthur, 2008).

Επίσης, για το κάθε καλύτερο ταίριασμα ανάθεσης πεπτιδίου σε ένα φάσμα MS / MS, η κατανομή πιθανοτήτων συχνά εκτιμάται από την κατασκευή του ιστογράμματος των scores όλων των πεπτιδίων που προήλθαν από τη Βάση Δεδομένων και σκόραραν εναντίον του συγκεκριμένου φάσματος (εκτός από το πεπτίδιο με την καλύτερη βαθμολογία ώστε να μπορεί να αξιολογηθεί). Όσο πιο μακριά βρίσκεται η παρατηρηθείσα-καλύτερη βαθμολογία από τον πυρήνα του ιστογράμματος τόσο πιο σημαντικό και λιγότερο τυχαίο είναι το ταίριασμα πεπτιδίου με το παρατηρούμενο φάσμα (δηλαδή, έχει περισσότερες πιθανότητες να είναι το σωστό πεπτίδιο).

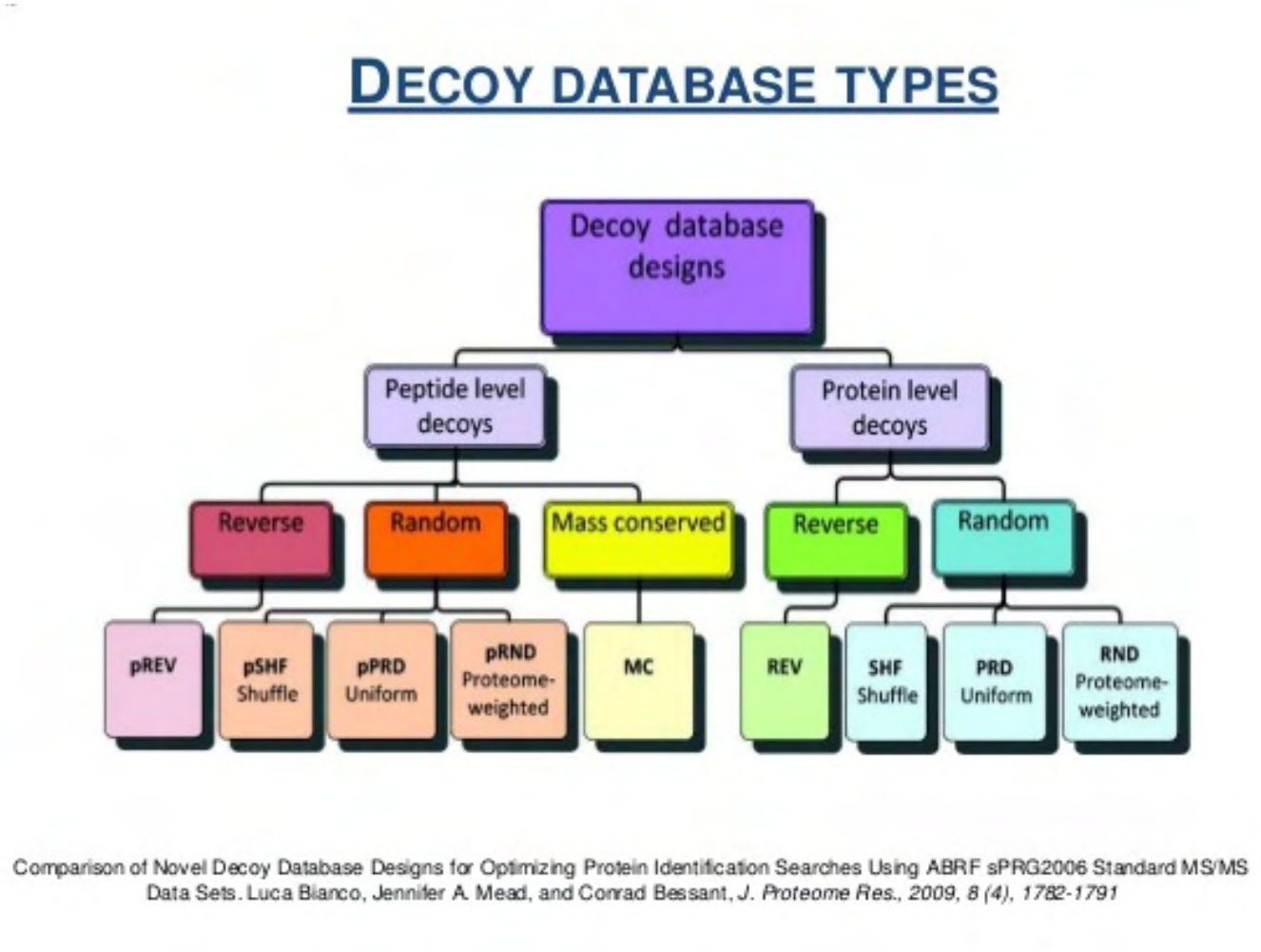


Εικόνα 5:Βαθμολογίες εμπιστοσύνης για μοναδιαία φάσματα

Ένα φάσμα MS / MS που αποκτήθηκε από ένα πείραμα φασματομετρίας μάζας συγκρίνεται με τα θεωρητικά φάσματα που κατασκευάστηκαν για κάθε υποψήφιο πεπτίδιο που δημιουργήθηκε από τη Βάση Δεδομένων. Τα υποψήφια πεπτίδια κατατάσσονται σύμφωνα με το score που ορίζεται από τον αλγόριθμο βαθμολόγησης που χρησιμοποιήθηκε. Η αλληλουχία του πεπτιδίου με την υψηλότερη βαθμολογία (το καλύτερο “match”) επιλέγεται για περαιτέρω ανάλυση. Κατασκευάζεται ένα ιστόγραμμα της συχνότητας της εμφάνισης των βαθμολογιών “s” μεταξύ όλων των συγκρίσεων που πραγματοποιήθηκαν (στο συγκεκριμένο παράδειγμα χρησιμοποιείται το σκορ «Xcorr» από τη μηχανή αναζήτησης SEQUEST). Στη συνέχεια γίνεται κανονικοποίηση προς το συνολικό αριθμό των υποψηφίων πεπτιδίων της Βάσης Δεδομένων “n”, και προσαρμόζεται χρησιμοποιώντας ένα μοντέλο κατανομής “P(s)” (κατανομή Gaussian, διακεκομμένη γραμμή). Η περιοχή κάτω από την P(s) στα δεξιά που εκτείνεται πέρα από την κορυφαία βαθμολογία “Sm” υπολογίζεται, και στη συνέχεια μετατρέπεται σε τιμή E-value. Η τιμή E-value χρησιμοποιείται στη θέση έναντι της αρχικής βαθμολογίας αναζήτησης στη Βάση Δεδομένων για όλες τις μετέπειτα αναλύσεις και για το φιλτράρισμα των δεδομένων (Nesvizhskii, 2010).

False Discovery Rate

Οι περισσότεροι ερευνητές, στις αναλύσεις τους χρησιμοποιούν μια διόρθωση των στατιστικών τιμών με μια μια μέθοδο γνωστή ως εκτίμηση του FDR (False Discovery Rate) που συνήθως βασίζεται στη δημιουργία μία Βάση Δεδομένων-δόλωμα.
Η χρήση FDR έχει αποδειχθεί ότι είναι ένα σημαντικό και χρήσιμο εργαλείο στη σύγχρονη πρωτεωμική, δίνοντας εμπειρική στατιστική επικύρωση των πεπτιδικών και πρωτεϊνικών ταυτοποιήσεων. Έτσι, ο αριθμός των ταιριασμάτων από τη Βάση Δεδομένων-δόλωμα αποτελεί μια εξαιρετική εκτίμηση του αριθμού των ψευδώς θετικών ταιριασμάτων που προκύπτουν από την πραγματική ή Βάση Δεδομένων-στόχο (Elias et al., 2005; Keller et al., 2002). Ωστόσο, όσο μεγαλύτερο είναι το μέγεθος της Βάσης Δεδομένων-στόχου, τόσο μεγαλύτερο είναι και το μέγεθος της Βάσης Δεδομένων-δολώματος με αποτέλεσμα η πιθανότητα για τυχαίες αντιστοιχήσεις να αυξάνεται.
Διάφορες μέθοδοι έχουν δημοσιευθεί για τη δημιουργία αυτών των Βάσεων Δεδομένων-δόλωμα, αλλά δεν υπάρχει μία η οποία έχει αποδειχθεί καλύτερη και πιο αποτελεσματική από τις υπόλοιπες (Bianco et al., 2009).



Comparison of Novel Decoy Database Designs for Optimizing Protein Identification Searches Using ABRF sPRG2006 Standard MS/MS Data Sets. Luca Bianco, Jennifer A. Mead, and Conrad Bessant, *J. Proteome Res., 2009, 8 (4), 1782-1791*

Εικόνα 6:Σχεδιασμός decoy-Βάσεων Δεδομένων.
(Bianco et al., 2009)

Επομένως, ακόμη και οι «καλύτερες» και πιο ξεκάθαρες συσχετίσεις-αντιστοιχήσεις (PSMs) μπορεί να περιέχουν ένα σημαντικό αριθμό ψευδών θετικών ταυτοποιήσεων αν δεν εφαρμόζονται στην ανάλυση των αποτελεσμάτων τα κατάλληλα κατώτερα όρια των μέτρων της στατιστικής σημαντικότητας (p-values, E-values, και/ή false-discovery rates (FDR).

Για τα υψηλής ποιότητας φάσματα, δηλαδή όταν η αναλογία σήματος προς θόρυβο είναι υψηλή και τα φάσματα περιέχουν σαφώς καθορισμένη σειρά ιόντων, οι περισσότεροι αλγόριθμοι αποδίδουν αρκετά καλά. Ωστόσο, όταν ο κατακερματισμός του πεπτιδίου στο πείραμα δεν είναι ο βέλτιστος και το φάσμα περιέχει πολύ λίγες διακριτές κορυφές, ή όταν η αντίστοιχη πεπτιδική αλληλουχία αμινοξέων δεν υπάρχει στη Βάση Δεδομένων τότε οδηγούμαστε σε λάθος ταυτοποιήσεις πεπτιδίων. Όταν το φασματόμετρο μάζας παρέχει υψηλή ακρίβεια μάζας τότε περιορίζεται ο αριθμός των υποψηφίων πεπτιδικών αλληλουχιών με αποτέλεσμα να μειωθούν σημαντικά οι ψευδείς ταυτοποιήσεις.
Αν συνεκτιμηθεί και το εύρος λάθους του φασματόμετρου που χρησιμοποιήθηκε τότε ο αριθμός των matches αυξάνεται. Επίσης, όσο μεγαλύτερη είναι η πρωτεΐνη τόσο περισσότερα πεπτίδια δημιουργούνται με αποτέλεσμα να αυξάνεται και η πιθανότητα των false-positives. Επομένως, η ανάγκη δημιουργίας αυτών των αλγορίθμων ήταν επιτακτική (Liebler, 2001).
Επιπλέον, η εξαντλητική αναζήτηση μετα-μεταφραστικών τροποποιήσεων (PTMs) και παραλλαγών αλληλουχίας, όπως της κωδικοποίησης πολυμορφισμών μονού νουκλεοτιδίου (SNPs) και του εναλλακτικού ματίσματος, μπορεί να οδηγήσει σε στατιστικά σημαντικές ψευδώς θετικές ταυτοποιήσεις.
Από τα αποτελέσματα της φασματομετρίας μάζας συνήθως υπάρχουν χιλιάδες πειραματικά φάσματα τα οποία είτε μπορούν να μελετηθούν ώστε να ταυτοποιηθούν τα φωσφοπεπτίδια από κάποιον έμπειρο ερευνητή είτε να γίνει χρήση κάποιου ειδικού λογισμικού. Αν και η επικύρωση των θέσεων φωσφορυλίωσης από τον ίδιο τον ερευνητή είναι πιο ακριβής και σίγουρη ωστόσο είναι μία διαδικασία χρονοβόρα, πολύ κουραστική και είναι αδύνατο να εφαρμοστεί σε πολύπλοκα δείγματα. Αντίθετα, τα διάφορα είδη λογισμικού χρησιμοποιούν κάποιο ευριστικό ή πιθανολογικό μοντέλο-αλγόριθμο που είναι συνήθως αρκετά γρήγορος.
Γι’ αυτό έχουν δημιουργηθεί διάφοροι αλγόριθμοι, όπως SEQUEST και Mascot, που μπορούν να χρησιμοποιηθούν τόσο για την ταυτοποίηση της πεπτιδικής ακολουθίας όσο και για την ταυτοποίηση των θέσεων φωσφορυλίωσης.
Ωστόσο, με τη χρήση αυτών των αλγορίθμων δημιουργείται ένα μικρό επίπεδο ασάφειας στην τοποθέτηση των φωσφορυλιωμένων θέσεων σε ένα πεπτίδιο είτε επειδή τα ιόντα δεν καθορίζουν επαρκώς τις θέσεις αυτές είτε επειδή είναι αρκετά δύσκολο όταν ένα πεπτίδιο φωσφορυλιώνεται σε περισσότερα από ένα αμινοξικά κατάλοιπα.

Συστήματα αξιολόγησης ταυτοποίησης των θέσεων φωσφορυλίωσης

Γενικά, είναι συχνά δύσκολο να γίνει διάκριση μεταξύ των πιθανών θέσεων φωσφορυλίωσης σε ένα πεπτίδιο. Η δυσκολία αυτή έχει να κάνει με την παρουσία περισσότερων της μίας πιθανής θέσης φωσφορυλίωσης (Ser, Thr, Tyr) σε ένα πεπτίδιο και με τη χαμηλή συνολική ένταση των κορυφών που δημιουργούνται σε ένα φάσμα μάζας από τα ιόντα που παράγονται από φωσφορυλιωμένα πεπτίδια.
Γενικά οι αναζητήσεις φωσφοπεπτιδίων είναι εγγενώς πιο απαιτητικές από εκείνες των μη τροποποιημένων πεπτιδίων λόγω της αύξησης του μεγέθους της Βάσης Δεδομένων γεγονός που οφείλεται στις δυναμικές τροποποιήσεις (δηλαδή, στη δυνατότητα φωσφορυλίωσης σε κάθε σερίνη, θρεονίνη ή τυροσίνη).
Η βασική δυσκολία είναι, συνεπώς, ότι επειδή τα πεπτίδια ταυτοποιούνται μέσω σύγκρισης των πειραματικών φασμάτων με εκείνων των αναμενόμενων από τη Βάση Δεδομένων, πρέπει και η ανάλυση αυτή να τροποποιηθεί ώστε να λάβει υπόψη τις διαφορές της μάζας των πεπτιδίων λόγω των μετα-μεταφραστικών τροποποιήσεων. Μια απλή λύση φαίνεται να είναι το να προστεθούν τα φάσματα για τα τροποποιημένα πεπτίδια στη Βάση Δεδομένων. Όταν θεωρητικά υπάρχει μία μόνο τροποποίηση, ειδικά αν είναι καθολική (π.χ., όλες οι κυστεΐνες αλκυλιώνονται), η στρατηγική αυτή λειτουργεί καλά. Ωστόσο, επειδή στην πραγματικότητα υπάρχουν >200 μετα-μεταφραστικές τροποποιήσεις που συναντώνται στη φύση (Gooley and Packer, 1997; Walsh, 2006), καθεμία από τις οποίες μπορεί να συμβεί σε συνδυασμό με άλλες, δεν μπορεί κανείς να αντιπροσωπεύσει όλους τους δυνατούς συνδυασμούς των τροποποιήσεων χωρίς η Βάση

Δεδομένων να γίνει γρήγορα δύσχρηστη. Για παράδειγμα, επιτρέποντας οποιεσδήποτε δύο τροποποιήσεις ανά πεπτίδιο (από τις >200 πιθανές) έχει ως αποτέλεσμα τη δημιουργία μίας Βάσης Δεδομένων που είναι περίπου 40000, φορές μεγαλύτερη. Και, όπως ο αριθμός των ψευδώς θετικών μεγαλώνει με το μέγεθος της Βάσης Δεδομένων, η απόδοση κάθε αναζήτησης μειώνεται με κάθε επιπλέον τροποποίηση που προστίθεται (Marcotte, 2007).
Οι περισσότεροι αλγόριθμοι αναζήτησης σε βάσεις δεδομένων δεν αξιολογούν τις διαφορετικές ισομορφές ενός πεπτιδίου με ειδική εξέταση, με αποτέλεσμα το κορυφαίο ταίριασμα πειραματικού και θεωρητικού φάσματος μάζας, ανεξάρτητα από το σκορ ταυτοποίησης του πεπτιδίου και την ακρίβεια μάζας, μπορεί να αντιστοιχεί στο σωστό πεπτίδιο αλλά σε λανθασμένο ισομερές θέσης.
Ως εκ τούτου, αλγόριθμοι για την ταυτοποίηση της θέσης φωσφορυλίωσης είναι αναγκαίοι.
Υπάρχουν δύο κατηγορίες αλγορίθμων εντοπισμού των θέσεων φωσφορυλίωσης:

- οι αλγόριθμοι που στηρίζονται σε πιθανότητες (probability-based localizers-PBLs)
- οι αλγόριθμοι που στηρίζονται σε search engine difference-SED scores

(Chalkley and Clauser, 2012)

Probability-based localizers

Πολλά εργαλεία PBL προέρχονται από αλγόριθμους που αρχικά είχαν σχεδιαστεί για να επεξεργάζονται φάσματα μάζας MS3 και στη συνέχεια εφαρμόστηκαν σε προβλήματα ανακάλυψης μετα-μεταφραστικών τροποποιήσεων (Olsen and Mann, 2004).
Ένα χαρακτηριστικό παράδειγμα τέτοιου αλγόριθμου που χρησιμοποιείται ευρέως είναι αυτό που σχεδιάστηκε από τους Olsen και Mann και ονομάζεται PTM score algorithm. Ο συγκεκριμένος αλγόριθμος προσπαθεί να λύσει το πρόβλημα ταυτοποίησης της θέσης τροποποίησης μέσω υπολογισμού διωνυμικής πιθανότητας, υπολογίζοντας μία πιθανότητα για κάθε υποψήφιο p-site.
Ο αλγόριθμος Ascore έχει αναπτύξει μια παρόμοια πιθανολογική προσέγγιση με το PTM score και είναι αναμφισβήτητα ο πιο γνωστός και ευρέως χρησιμοποιούμενος αλγόριθμος στα πειράματα φωσφοπρωτεωμικής (Beausoleil et al., 2006).
Άλλα παραδείγματα PBL αλγορίθμων είναι το PhosphoRS, το SLoMo όπως επίσης και η μηχανή αναζήτησης Andromeda που υπάρχει στην εφαρμογή της MaxQuant και χρησιμοποιεί μία δικιά της μορφή «PTM score».

Search Engine Difference scores

Οι SED βαθμολογίες υπολογίζονται όταν για μία δεδομένη τροποποίηση υπάρχουν πολλαπλές δυνατές θέσεις δημιουργώντας πολλά πιθανά PTM ισομερή. Αυτές οι προσεγγίσεις έχουν αποδειχθεί αρκετά δημοφιλείς λόγω της απλότητας τους και γιατί μπορούν να εφαρμοστούν σε οποιαδήποτε μέθοδο βαθμολόγησης. Παραδείγματα SED scores για τον εντοπισμό της θέσης τροποποίησης είναι το Mascot Delta score στην μηχανή αναζήτησης Mascot (Savitski et al., 2011) και το SLIP score (Site Localization In Peptides) στην μηχανή αναζήτησης ProteinProspector (Chalkley and Clauser, 2012).

Υλικά και μέθοδοι

Στην ενότητα αυτή γίνεται περιγραφή της στρατηγικής που ακολουθήσαμε για να εντοπιστούν τα δεδομένα. Πιο συγκεκριμένα, γίνεται μια παρουσίαση των βασικών βημάτων που ακολουθήθηκαν και συζητιούνται διάφορα προβλήματα και πώς αντιμετωπίστηκαν.

Όπως είναι γνωστό, ένα μεγάλο πρόβλημα είναι η εύρεση κατάλληλων δεδομένων. Στις επόμενες παραγράφους θα περιγράψουμε οτιδήποτε σχετίζεται με τα δεδομένα που χρησιμοποιήθηκαν, από την αναζήτησή τους μέχρι και τον μετασχηματισμό τους στην επιθυμητή μορφή και την οργάνωση τους σε μία καινούργια Βάση Δεδομένων.

Στρατηγική Αναζήτησης Δεδομένων από τη Βιβλιογραφία και από Βάσεις Δεδομένων

Για την υλοποίηση της παρούσας εργασίας ήταν απαραίτητο να βρεθούν τα κατάλληλα δεδομένα φωσφοπρωτεωμικής διαφόρων οργανισμών αλλά με ιδιαίτερη έμφαση και προτεραιότητα στον άνθρωπο (*Homo sapiens*), στο ποντίκι (*Mus_musculus*), στον αρουραίο (*Rattus norvegicus*), στην *Arabidopsis thaliana* και στο σακχαρομύκητα (*Saccharomyces cerevisiae*). Το γεγονός ότι υπάρχουν πολλά δεδομένα πρωτεωμικής γι' αυτούς τους οργανισμούς σε σχέση με άλλους οδήγησε στην επιλογή τους.
Η εύρεση των κατάλληλων δεδομένων αποτελεί ένα από τα πιο κρίσιμα στάδια καθώς η απόφαση για την πηγή προέλευσης των δεδομένων είναι καθοριστική για όλα τα επόμενα βήματα της συγκεκριμένης εργασίας. Όσο περισσότερα άρθρα συλλεχθούν με αξιόπιστα δεδομένα τόσο καλύτερης ποιότητας θα είναι και τα συμπεράσματα στα οποία θα καταλήξουμε μετά από την επεξεργασία τους. Αντίθετα, περιορισμένα ή κακής ποιότητας δεδομένα είθισται να οδηγούν σε μη αξιόπιστα συμπεράσματα. Ωστόσο η συγκεκριμένη διαδικασία είναι αρκετά χρονοβόρα και δύσκολη. Ένας από τους λόγους που δικαιολογεί τον αυξημένο βαθμό δυσκολίας στην εύρεση των κατάλληλων δημοσιευμένων άρθρων για τη συγκεκριμένη εργασία είναι ότι αναζητούσαμε ερευνητικά άρθρα που στόχευαν σε δεδομένα πρωτεωμικής με συγκεκριμένη μετα-μεταφραστική τροποποίηση (φωσφορυλίωση) και χρησιμοποιούσαν οπωσδήποτε μεθόδους αποφυγής ψευδώς θετικών αποτελεσμάτων και θορύβου τόσο στην ταυτοποίηση των πεπτιδικών αλληλουχιών όσο και στην τοποθέτηση των φωσφορυλιώσεων.
Τα δεδομένα μπορούν να ληφθούν από διαφορετικές και ετερογενείς πηγές. Στη συγκεκριμένη μελέτη χρησιμοποιήθηκαν διάφορες Βάσεις Δεδομένων.

Τα άρθρα που συλλέχθηκαν ήταν αποτελέσματα αναζήτησης:

1) στην Pubmed με τις λέξεις-κλειδιά (keywords):
   - «phosphoproteomics», από την οποία βρέθηκαν 718 άρθρα
   - «phosphoproteome», από την οποία βρέθηκαν 843 άρθρα

2) στο Scopus, όπου έγινε αναζήτηση των αναφορών-άρθρων που δημοσίευσαν τις πιο δημοφιλείς υ-πολογιστικές μεθόδους εντοπισμού της θέσεως φωσφορυλίωσης σε ένα πεπτίδιο, από δεδομένα MS/MS και τις χρησιμοποιήσαμε ως λέξεις-δολώματα αφού αναζητούσαμε άρθρα μελετών που θα χρησιμοποιούσαν οπωσδήποτε κάποιο τέτοιου είδους πρόγραμμα. Συγκεκριμένα, αναζητήθηκαν οι αναφορές στις παρακάτω μεθόδους:
   - Ascore από την οποία βρέθηκαν 477 άρθρα
   - PhosphoRS από την οποία βρέθηκαν 54 άρθρα
   - PhosphoScore από την οποία βρέθηκαν 42 άρθρα
   - SloMo από την οποία βρέθηκαν 39 άρθρα
   - Mascot Delta Score από την οποία βρέθηκαν 47 άρθρα

- PTM Score από την οποία βρέθηκαν 1386 άρθρα
- InsPecT από την οποία βρέθηκαν 262 άρθρα
- PhosCalc από την οποία βρέθηκαν 19 άρθρα.

3) στη Βάση Δεδομένων PRIDE με τα εξής φίλτρα:
   - Species: Human, modification: phosphorylation (βρέθηκαν 61 αποτελέσματα)
   - Species: *Mus musculus*, modification: phosphorylation (βρέθηκαν 16 αποτελέσματα)
4) στη Βάση Δεδομένων ProteomeXchange με λέξεις κλειδιά:
   - *Homo sapiens*
   - Mouse
   - Rat
   - *Saccharomyces cerevisiae*

από την οποία βρέθηκαν 214, 78, 9 και 14 αποτελέσματα αντίστοιχα.

5) στη Βάση Δεδομένων PhosPhAt, η οποία περιέχει δεδομένα φωσφοπρωτεομικής μόνο για τον οργανισμό *Arabidopsis thaliana* από την οποία βρέθηκαν 35 άρθρα.

Μεταξύ των αποτελεσμάτων που εντοπίστηκαν από τις διαφορετικές στρατηγικές αναζήτησης υπήρξε όπως αναμενόταν αλληλοεπικάλυψη. Συνολικά ανακτήθηκαν 4186 άρθρα και από αυτά μελετήθηκαν τα 1029. Από τα 1029 άρθρα που μελετήθηκαν, τα 205 μόνο είχαν αξιοποιήσιμα δεδομένα. Αυτή η αρκετά μεγάλη διαφορά στον αριθμό μεταξύ των άρθρων που μελετήθηκαν και αυτών που τελικά αξιοποιήθηκαν οφείλεται σε διάφορους λόγους. Για παράδειγμα, η αναζήτηση άρθρων, με τους τρόπους που αναφέρθηκαν προηγουμένως, έδινε ως αποτελέσματα και άρθρα που δεν ήταν πρωτότυπες μελέτες. Αντίθετα, μπορεί να ήταν κριτικές ανασκοπήσεις άλλων άρθρων ή μετα-αναλύσεις ήδη δημοσιευμένων φωσφοπρωτεϊνικών αποτελεσμάτων.
Επιπλέον, πολλά από τα άρθρα, κυρίως τα πιο πρόσφατα, ήταν κλειδωμένα και ήταν αδύνατη η συλλογή πληροφοριών για τα προγράμματα που χρησιμοποιήθηκαν για την αξιολόγηση των δεδομένων με αποτέλεσμα να μην μπορούμε να κρατήσουμε τα δεδομένα αυτά για ανάλυση. Ωστόσο, μετά από επικοινωνία με τους ερευνητές μερικών από αυτών των κλειδωμένων άρθρων έγινε δυνατή η λήψη τους και τελικά η χρήση των δεδομένων.
Επίσης, το γεγονός ότι τα προγράμματα αξιολόγησης της ταυτοποίησης των πεπτιδίων και της τοποθέτησης της τροποποίησης χρησιμοποιούνται στις μελέτες όχι μόνο της φωσφορυλίωσης αλλά και όλων των υπόλοιπων μετα-μεταφραστικών τροποποιήσεων (ακετυλίωση κ.α.) αλλά και η μελέτη συνθετικών πεπτιδίων οδήγησε σε συλλογή άρθρων μη σχετικών με το θέμα της παρούσας εργασίας.
Πολλά από τα άρθρα που μελετήθηκαν δεν διέθεταν δεδομένα καλής ποιότητας με αποτέλεσμα την κατακόρυφη μείωση των τελικά αξιοποιήσιμων αποτελεσμάτων.
Τέλος, τα δεδομένα φωσφοπρωτεωμικής πολλών άρθρων βρισκόταν σε μη αξιοποιήσιμη μορφή, όπως για παράδειγμα σε μορφή .pdf και όχι .xls (excel files), με αποτέλεσμα να μην μπορούμε να τα συμπεριλάβουμε στην έρευνα γιατί ήταν αδύνατη η ανάλυση τους.

Οργάνωση των δημοσιευμένων εργασιών

Η μελέτη κάθε άρθρου συμπεριελάμβανε την αποθήκευση του ίδιου του άρθρου και του συμπληρωματικού υλικού του και προσεκτική ανάγνωση.
Για την αρχική οργάνωση και διαχείριση των εργασιών που εντοπίστηκαν, δημιουργήθηκε ένα αρχείο excel στο οποίο προσδιορίζουμε τα εξής, μετά από ανάγνωση του κάθε άρθρου:

- τίτλο άρθρου,
- συγγραφείς,
- ημερομηνία δημοσίευσης,
- PubmedID,
- οργανισμό,
- ιστό και κύτταρα,
- μέθοδο πέψης,
- μέθοδο εμπλουτισμού των φωσφοπεπτιδίων,
- software αναγνώρισης των πεπτιδίων,
- software προσδιορισμού της θέσης φωσφορυλίωσης,
- σχόλια για το κάθε άρθρο,
- αν είναι σχετικό με τη συγκεκριμένη έρευνα,
- τη στρατηγική αναζήτησης που το εντόπισε,
- αν είναι ελεύθερα προσβάσιμο ή όχι,
- αν ήταν προσβάσιμο το excel με τα αποτελέσματα
- οι τυχόν υπάρχοντες υπερσύνδεσμοι που οδηγούν στα δεδομένα των βάσεων PRIDE και ProteomeXchange,
- αν υπάρχει η πεπτιδική ακολουθία στο excel από το συμπληρωματικό υλικό του κάθε άρθρου,
- αν υπάρχει το localization score στο excel από το συμπληρωματικό υλικό του κάθε άρθρου.

Για την βέλτιστη οργάνωση αυτών των δεδομένων και την πιο εύκολη διαχείρισή τους στη συνέχεια δημιουργήθηκε και μία Βάση Δεδομένων από τον Π. Βλασταρίδη, διδακτορικό φοιτητή του εργαστηρίου, που συμπεριλάμβανε όλες τις λεπτομέρειες των άρθρων που αναλύθηκαν και τα αξιοποιήσιμα δεδομένα.
Ωστόσο, τα δεδομένα φωσφοπρωτεωμικής των άρθρων που μελετήθηκαν και διέθεταν αξιοποιήσιμο υλικό ήταν τις περισσότερες φορές οργανωμένα για άλλη χρήση και γι’ αυτό προκαλούσαν σύγχυση με αποτέλεσμα να μην μπορούν να χρησιμοποιηθούν εύκολα για εξαγωγή πληροφορίας. Επίσης, όπως αναφέρθηκε και στην εισαγωγή, τα πειραματικά δεδομένα περιέχουν θόρυβο καθώς και άλλου τύπου αποκλίσεις που δημιουργήθηκαν κατά τη διάρκεια της πειραματικής διαδικασίας και αφορούν την ταυτοποίηση της πεπτιδικής αλληλουχίας και τον εντοπισμό της θέσης φωσφορυλίωσης, με αποτέλεσμα να μην μπορούν να χρησιμοποιηθούν με αυτή τη μορφή στο επόμενο στάδιο ανάλυσης, αφού δεν είναι αρκετά αξιόπιστα και ακριβή. Επομένως, υπήρξε ένα βήμα προ επεξεργασίας των δεδομένων ώστε να μετασχηματιστούν σε απλούστερες μορφές που διευκολύνουν την εύρεση γνώσης και μπορούν με αυτό τον τρόπο να αναλυθούν πιο εύκολα.
Αρχικά, επειδή δεν υπάρχει ένας μόνο συγκεκριμένος τρόπος δήλωσης της ύπαρξης της φωσφορυλίωσης στις πεπτιδικές αλληλουχίες που ακολουθούν όλα τα ερευνητικά εργαστήρια η επεξεργασία των δεδομένων με ένα τυποποιημένο υπολογιστικό τρόπο προαπαιτούσε την επεξεργασία τους. Επομένως, στη παρούσα εργασία συμφωνήθηκε ότι το σημείο φωσφορυλίωσης σε κάθε πεπτιδική ακολουθία θα δηλωνόταν με το πεζό γράμμα «p» και ακολουθούσε πάντα το φωσφορυλιωμένο αμινοξύ που θα ήταν ένα από τα: σερίνη, θρεονίνη ή τυροσίνη.
Η διαδικασία αυτή έγινε ημι-αυτόματα, χρησιμοποιώντας διάφορα perl scripts και τρόπους ταξινόμησης και φιλτραρίσματος στην επιφάνεια εργασίας excel, για τον εντοπισμό των

προβληματικών δεδομένων υπό την εποπτεία του ανθρώπινου παράγοντα που θα αποφασίσει τελικά αν θα κρατήσει τα δεδομένα ή όχι.
Σε αρκετά από αυτά τα excel spreadsheets υπήρχαν τα φωσφοπεπτίδια καθώς επίσης και ένα peptide score (συνήθως Mascot score) ή ακόμα και ένα phosphosite localization score. Επομένως από κάθε αρχείο excel από το συμπληρωματικό υλικό κάθε δημοσιευμένου άρθρου που μελετήθηκε, απομονώθηκαν σε ξεχωριστό αρχείο excel (.xls) μόνο τα φωσφοπεπτίδια εκείνα που είχαν peptide score και phosphosite localization score που ξεπερνούσε το κατώφλι βεβαιότητας που ορίσαμε (1% πιθανότητα λάθους). Τα scores αυτά αντικατοπτρίζουν την βεβαιότητα με την οποία έχει εντοπιστεί το πεπτίδιο και η συγκεκριμένη θέση φωσφορυλίωσης αντίστοιχα. Αυτό βέβαια συνεπάγεται ότι για κάθε έναν αλγόριθμο πρέπει να έχει οριστεί μια τιμή κατώφλι. Το κάθε score υπολογίζεται διαφορετικά, το κοινό κριτήριο όμως ήταν βεβαιότητα 99% για εντοπισμό πεπτιδίου και βεβαιότητα 99% για εντοπισμό θέσης φωσφορυλίωσης. Σε κάθε μία από αυτές τις εργασίες εντοπίστηκαν τα μοναδικά φωσφοπεπτίδια που ικανοποιούσαν τα κριτήριά μας. Το αρχείο τύπου excel (.xls) από κάθε δημοσιευμένη εργασία μετασχηματίστηκε και σε αρχείο τύπου text (.txt) για να είναι εφικτή η ανάγνωση και ο χειρισμός των δεδομένων στη γλώσσα προγραμματισμού της Perl. Η δομή του αρχείου αυτού δημιουργήθηκε ικανοποιώντας τις απαιτήσεις των διάφορων perl scripts που χρησιμοποιήθηκαν.

Στη συνέχεια δημιουργήθηκαν προγράμματα στη γλώσσα προγραμματισμού Perl που:

- εντόπισαν για κάθε ένα φωσφοπεπτίδιο σε ποια πρωτεΐνη/πρωτεΐνες αντιστοιχεί
- εντόπισαν την ακριβή θέση φωσφορυλίωσης κάθε φωσφοπεπτιδίου πάνω στην πρωτεΐνη.

Για όλες τις πρωτεΐνες χρησιμοποιήσαμε τα δεδομένα από την Βάση Ensembl. Για κάθε γονίδιο χρησιμοποιήσαμε το μεγαλύτερο ισόμορφο.

Αποτελέσματα

Χρονολόγηση άρθρων

Όλα τα άρθρα που μελετήθηκαν χρονολογούνται από το 2003 έως το 2015 ενώ αυτά που έχουν αξιοποιήσιμα δεδομένα χρονολογούνται από το 2005 έως το 2015.

Άρθρα για κάθε είδος οργανισμού

Στον παρακάτω πίνακα φαίνεται πόσα άρθρα από αυτά που μελετήσαμε είχαν δεδομένα τα οποία μπορούσαμε να αξιοποιήσουμε και ο συνολικός αριθμός των φωσφοπεπτιδίων και φωσφοπρωτεϊνών που ανακτήθηκαν.

Πίνακας 1: Αριθμός άρθρων, φωσφοπεπτιδίων και φωσφοπρωτεϊνών ανά οργανισμό

	Homo sapiens	*Mus musculus*	*Rattus norvegicus*	*Arabidopsis thaliana*	*Saccharomyces cerevisiae*
#articles	97	42	17	28	21
#phosphosites	79424	45596	19598	14796	14339
#phosphosites found in >=3 experiments	23997	13087	2640	3078	4239
#phosphosites found in 1 experiment	43031	25151	13227	9567	7792
#phosphoproteins	10005	8129	5746	4930	2633
#phosphoproteins found in >=3 experiments	6275	4736	2018	1815	1660
#phosphoproteins found in 1 experiment	2598	2277	2326	2322	642

Πιο συγκεκριμένα για κάθε οργανισμό:

Homo sapiens

Πίνακας 2:Αριθμός φωσφοπεπτιδίων και φωσφοπρωτεϊνών για τον οργανισμό Homo sapiens

PMID	#phosphosites	#phosphosites incremental sum	#phosphoproteins	#phosphoproteins incremental sum
16396503	588	588	339	339
16497976	158	745	93	423
17929957	627	1310	268	624
18212344	903	2051	523	967
18318008	244	2220	147	1041
18452278	4294	5967	1519	2060
18491316	175	6101	129	2127
18510355	145	6208	66	2158
18691976	2094	7890	891	2614
19081932	60	7917	41	2625
19362540	247	8012	195	2665
19605366	117	8101	99	2710
19664994	1110	8722	818	2972
19664995	2722	10209	1167	3268
19764811	656	10725	235	3367
19786723	4933	13636	1702	3826
19845377	616	13815	311	3849
20222723	54	13854	38	3856
20230923	127	13865	112	3859
20340162	2417	14721	1279	4022
20363803	4170	16506	1572	4274
20393185	569	16845	389	4387
20446291	356	16895	131	4390
20639409	875	17193	477	4476
20686112	198	17243	155	4486
20726782	31	17250	24	4487
20860994	1505	17733	570	4531
20866107	937	17983	442	4562
20886841	171	18046	116	4588
20946866	450	18080	274	4598
21130716	7204	21190	2595	4995
21152398	2901	22334	1262	5239
21628590	1	22334	2	5239
21685386	947	22397	638	5247
21712546	33441	44349	4833	6494
21788404	12723	48611	3876	7103
21857030	1597	48854	1001	7128

21899308	19749	53793	4322	7465
21949786	1488	53860	939	7477
22073976	6701	54815	2123	7536
22115753	5401	55498	2160	7616
22167270	1488	55530	883	7619
22322096	421	55716	154	7629
22451900	490	55978	378	7652
22461510	620	56183	380	7679
22468782	2082	56726	1181	7771
22496350	1802	57171	1024	7804
22499768	334	57232	239	7810
22593177	651	57276	265	7813
22633412	9876	61357	3853	8275
22798277	5638	62625	2219	8426
22843994	1768	62759	1098	8445
22886815	1234	62977	677	8462
22985185	4522	63691	2314	8569
23041048	248	63722	198	8577
23286773	103	63746	62	8578
23312004	16457	69596	5607	9195
23322592	1056	69611	574	9195
23628362	807	69717	544	9222
23692254	1145	69753	727	9225
23808766	1868	69906	894	9234
23874979	157	69940	137	9236
23882029	202	70025	159	9239
23889490	1998	70321	1028	9259
23911959	9979	71394	3231	9323
23917254	12534	72945	4025	9392
24011590	445	73084	329	9398
24117733	1453	73143	962	9402
24129246	172	73145	152	9402
24173317	3049	73527	1573	9442
24215720	19	73531	18	9444
24275569	9272	76158	3146	9665
24300666	2354	76231	1312	9665
24322422	4031	76644	1988	9703
24324209	1689	76930	770	9717
24425749	1238	76983	662	9722
24461736	833	77069	358	9726
24501219	3155	77247	1775	9743
24596151	73	77248	60	9744
24702127	9	77252	9	9745
24804581	607	77406	375	9749
24888630	38	77431	26	9756
25002506	7975	77935	3115	9831

25003641	3538	78002	1726	9837
25101063	11	78002	11	9837
25106551	2260	78114	1258	9842
25119995	3869	78245	1719	9849
25142963	18	78252	17	9851
25147952	530	78494	401	9865
25219547	670	78500	350	9867
25223752	51	78503	47	9868
25278378	1230	78737	656	9915
25307156	1878	79129	954	9970
25311616	1829	79171	933	9972
25332170	1112	79385	594	10004
25338102	65	79416	50	10005
25348772	1060	79424	695	10005

Mus musculus

Πίνακας 3: Αριθμός φωσφοπεπτιδίων και φωσφοπρωτεϊνών για τον οργανισμό Mus musculus

PMID	#phosphosites	#phosphosites incremental sum	#phosphoproteins	#phosphoproteins incremental sum
18034455	232	232	181	181
18522436	179	410	89	263
19131326	462	828	324	533
19674963	1893	2553	863	1210
19854140	6819	8331	1974	2484
20222745	454	8490	316	2538
20438120	49	8522	36	2551
20469934	4327	11005	2150	3387
20531401	2693	11780	1246	3530
20688971	98	11803	56	3539
21659605	5040	13885	2018	3935
21788404	4795	15679	2111	4310
21917720	2927	17353	1452	4809
22006019	10598	22687	3464	5655
22078882	2572	23181	1432	5743
22322096	163	23299	72	5756
22345495	2037	23559	1080	5804
22357970	69	23586	36	5804
22499769	255	23727	160	5826
22705319	354	23771	228	5827
22807455	9151	27860	2917	6283
22843994	18	27863	15	6284
22871156	2243	28677	1285	6456
22942356	6320	30781	2466	6594
23352502	42	30787	35	6594
23567750	4507	31411	1955	6664
23597982	934	32135	350	6791
23882026	496	32231	301	6820
23926118	152	32255	119	6824
23970565	729	32299	564	6832
23984901	3922	33167	2310	7005
24224561	1263	33530	693	7039
24453211	5330	34712	1911	7087
24560892	184	34751	99	7090
24925903	1321	34841	680	7101
25159016	6393	36364	2280	7304
25168779	2209	36442	1099	7311
25177544	678	36490	417	7325
25263469	202	36513	147	7329

25338131	21111	44460	5041	8014
25367039	5488	45320	2325	8074
25777480	3598	45596	1736	8129

Rattus norvegicus

Πίνακας 4: Αριθμός φωσφοπεπτιδίων και φωσφοπρωτεϊνών για τον οργανισμό Rattus norvegicus

PMID	#phosphosites	#phosphosites incremental sum	#phosphoproteins	#phosphoproteins incremental sum
16396499	222	222	132	132
17683130	28	246	23	147
20028136	1031	1232	703	785
20568813	109	1285	87	817
20628157	483	1695	410	1092
21630457	2534	3712	1120	1723
21738781	1563	4603	967	2098
22276854	1974	5557	1132	2460
22345510	36	5568	30	2466
22609512	56	5613	34	2487
22673903	12785	15241	4302	4791
23800682	2184	15815	1161	4932
23984901	3051	17388	1921	5368
24214862	529	17470	357	5385
24467267	80	17503	64	5401
24945867	690	17598	367	5416
25403869	4437	19598	2037	5746

Arabidopsis thaliana

Πίνακας 5: Αριθμός φωσφοπεπτιδίων και φωσφοπρωτεϊνών για τον οργανισμό Arabidopsis thaliana

PMID	#phosphosites	#phosphosites incremental sum	#phosphoproteins	#phosphoproteins incremental sum
16807317	63	63	18	18
17317660	80	143	62	80
17586839	29	172	27	105
17651370	105	264	73	158
17934214	2	266	2	159
18463617	1339	1549	813	919
18686298	36	1572	33	936
19245862	210	1648	189	982
19376835	1565	2711	973	1494
19688752	47	2722	44	1498
19900291	12	2725	11	1498
20374526	149	2787	116	1522
20466843	2639	3889	1356	1917
21175636	1093	4504	659	2130
22060019	155	4613	120	2202
22325874	22	4613	15	2202
22438062	562	4834	356	2267
22616989	19	4840	15	2270
22631563	390	5073	284	2389
23094866	796	5269	536	2460
23111157	1110	5917	833	2766
23328941	2695	7244	1241	3045
23572148	3312	8311	1391	3224
23660473	1306	8520	781	3281
23776212	2092	8857	1207	3383
23820729	15	8872	9	3390
24299221	28	8900	14	3398
25561503	9480	14796	3736	4930

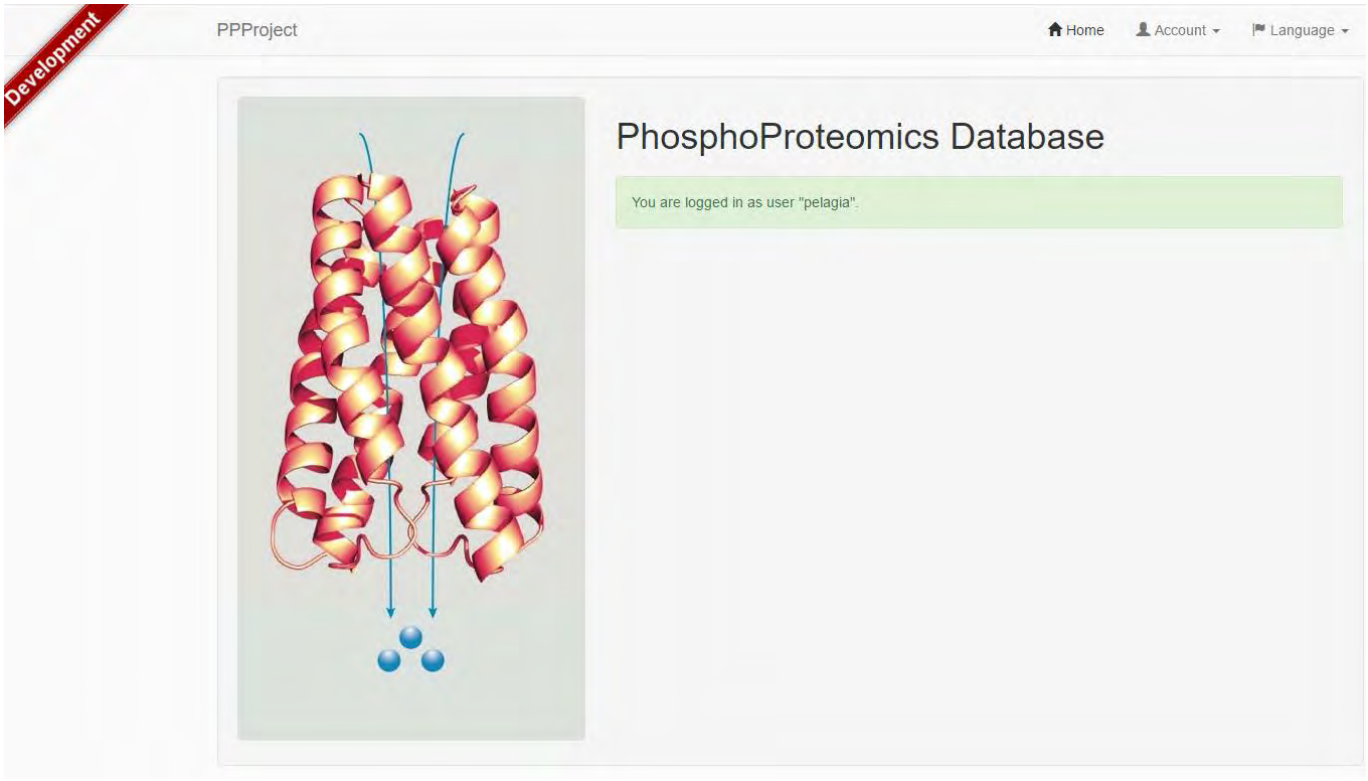
Saccharomyces cerevisiae

Πίνακας 6: Αριθμός φωσφοπεπτιδίων και φωσφοπρωτεϊνών για τον οργανισμό Saccharomyces cerevisiae

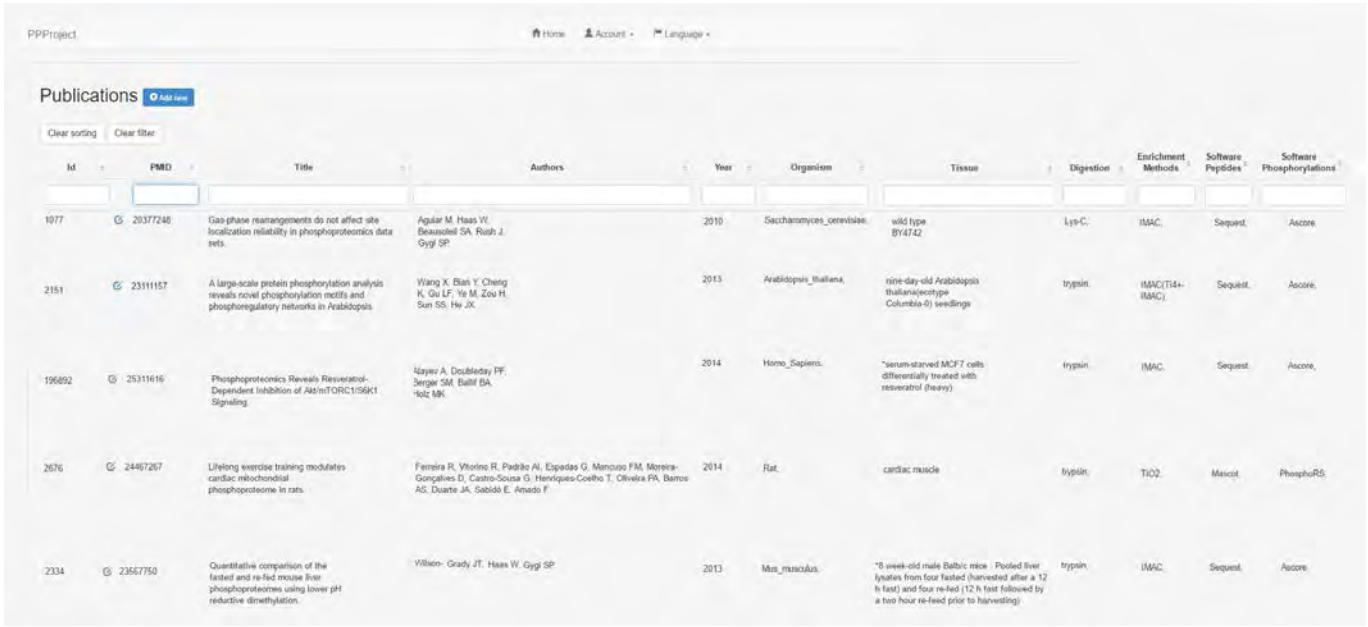
Dataset	#phosphosites	#phosphosites incremental sum	#phosphoproteins	#phosphoproteins incremental sum
Gruhler	676	676	470	470
Chi	724	1349	422	744
Li	1433	2455	755	1105
Albuquerque	3155	4696	1513	1811
Bodenmiller_1	2274	5588	1071	1933
Beltrao	201	5655	177	1947
Huber	311	5668	160	1947
Holt_1P	1939	6265	857	2009
Holt_2P	3348	7554	1286	2158
Holt3P	4321	9216	1400	2309
Gnad	1546	9580	727	2348
Soufi	1155	9783	683	2374
Aguiar	1419	10065	772	2410
Saleem	960	10342	352	2412
Bodenmiller_2	973	10509	588	2429
Wu	3428	11632	1271	2512
Oliviera	2538	12040	1031	2539
Mascaraque	186	12078	109	2540
Lee	12	12083	2	2540
Weinert	3576	13242	1029	2587
PG	1633	14339	374	2633

Βάση Δεδομένων PPProject

Δεδομένου του μεγάλου όγκου των βιολογικών δεδομένων που ανακτήθηκε από τη συγκεκριμένη μελέτη κρίθηκε απαραίτητη η ανάπτυξη/χρήση κάποιας βάσης δεδομένων η οποία θα επέτρεπε την αποθήκευση και οργάνωση όλων αυτών των πληροφοριών αλλά και θα παρείχε την δυνατότητα άμεσης προσπέλασης από τον χρήστη. Για τον λόγο αυτό εγκαταστάθηκε τοπικά, από τον διδακτορικό φοιτητή Π. Βλασταρίδη, μία Βάση Δεδομένων με το όνομα “PPProj” – Phospho Proteomics Project Database στην οποία βρίσκονται οργανωμένα τα δεδομένα που προέκυψαν από τη συγκεκριμένη εργασία και στην οποία ο χρήστης έχει τη δυνατότητα να την ανανεώνει χειροκίνητα, όταν χρειάζεται. Η PPProj οργανώθηκε με τρόπο ανάλογο των μεγάλων Βάσεων Δεδομένων. Αποτελείται από πίνακες καθένας από τους οποίους περιέχει στήλες και γραμμές στα κελιά των οποίων αποθηκεύονται όλα τα δεδομένα και πληροφορίες των δημοσιευμένων εργασιών που μελετήθηκαν.



Εικόνα 7 : Η Βάση Δεδομένων PPProj



Εικόνα 8: Η Βάση Δεδομένων PPProj

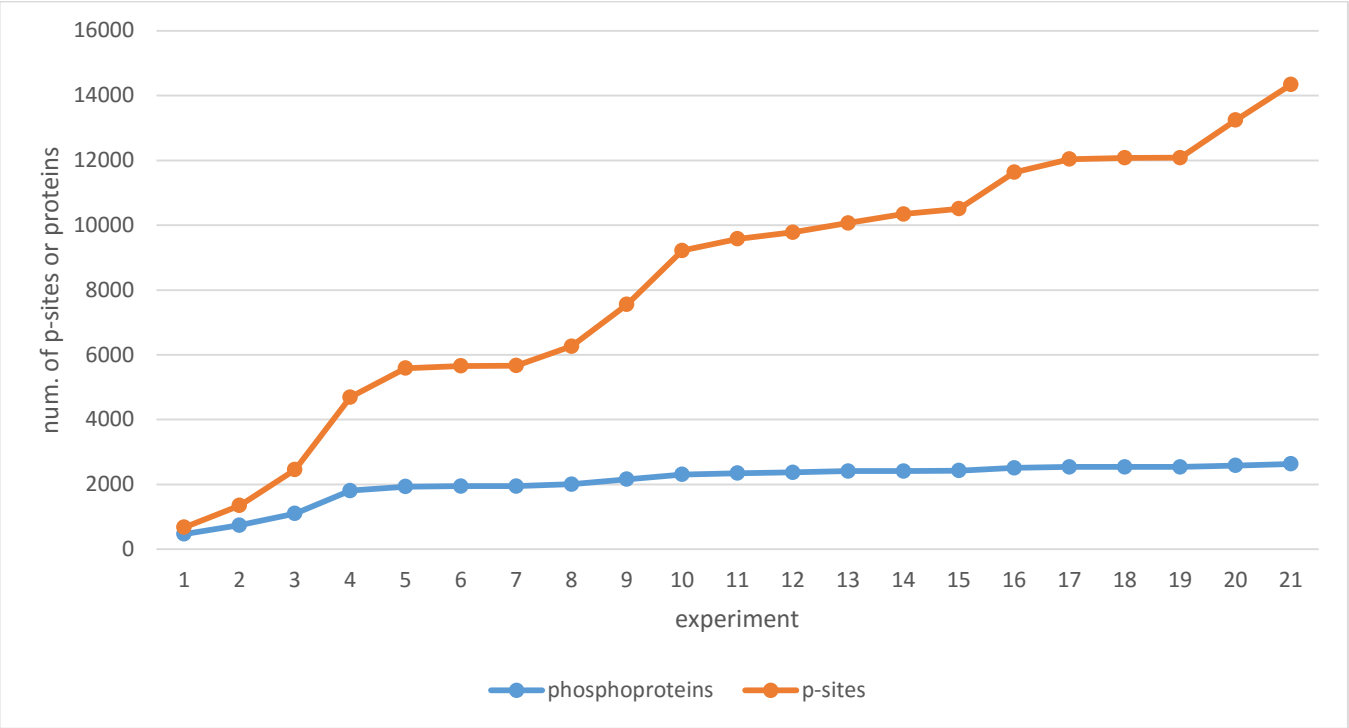
Ο κορεσμός της ανακάλυψης των δεδομένων φωσφορυλίωσης υψηλής-ποιότητας

Στο σημείο αυτό μελετήθηκε η πιθανότητα να ισχύει ότι το σύνολο των δεδομένων φωσφοπρωτεωμικής που συλλέξαμε και φιλτράραμε για κάθε οργανισμό συνθέτει το σύνολο των φωσφοπεπτιδίων και φωσφοπρωτεϊνών που όντως εκφράζονται.
Το πλήρες «phosphoproteome» ενός οργανισμού αναφέρεται στην απογραφή όλων των αμινοξέων που φωσφορυλιώνονται με μία ή περισσότερες συνθήκες. Ωστόσο, λόγω των δύσκολων πειραματικών ροών εργασίας που χαρακτηρίζουν τα πειράματα Πρωτεωμικής και οδηγούν σε χαρακτηρισμό/προσδιορισμό ενός ποσοστού μόνο των φωσφοπεπτιδίων κάθε φορά σε συνδυασμό με την αδυναμία διεξαγωγής πειραμάτων υπό κάθε πιθανή συνθήκη το συνολικό μέγεθος του φωσφοπρωτεώματος κάθε οργανισμού είναι δύσκολο να αποτιμηθεί. Πιο συγκεκριμένα, όπως έχει ήδη αναφερθεί, αν και παρατηρείται μια ραγδαία αύξηση της απόδοσης των τεχνικών και στρατηγικών που χρησιμοποιούνται σε πειράματα φωσφοπρωτεωμικής (Grimsrud et al., 2010) ωστόσο ακόμη και με την πιο προηγμένη προσέγγιση ποτέ δεν αποκαλύπτονται όλα τα φωσφοπεπτίδια από ένα βιολογικό δείγμα (Solari et al., 2015). Το γεγονός αυτό αποδεικνύεται από το ότι η επανάληψη ενός πειράματος πάντα οδηγεί σε ανακάλυψη νέων φωσφοπεπτιδίων.
Εμείς εδώ εκμεταλλευόμαστε το φαινόμενο κορεσμού για να εκτιμηθεί η πληρότητα των συνόλων δεδομένων φωσφοπρωτεωμικής με τη γραφική αναπαράσταση των μοναδικών φωσφοπεπτιδίων που έχουν βρεθεί σε σχέση με τον αύξοντα αριθμό των πειραμάτων.

Saccharomyces cerevisiae:

Ο οργανισμός *S. cerevisiae* (budding yeast) είναι ο πιο καλά μελετημένος μονοκυτταρικός ευκαριωτικός οργανισμός και εκφράζει μόνο ~ 6.000 πρωτεΐνες (Oliver et al., 1992). Περισσότερο από 70% του πρωτεώματός του είναι ανιχνεύσιμο με ένα μόνο πείραμα MS/MS (de Godoy et al., 2008; Wu et al., 2011). Εικοσιένα σύνολα δεδομένων έχουν συλλεχθεί από αυτόν τον οργανισμό για την παρούσα διπλωματική εργασία, κάτω από ένα αρκετά ευρύ φάσμα συνθηκών. Επομένως, ο ζυμομύκητας φαίνεται να αποτελεί το ιδανικό σύστημα με το οποίο μπορεί να εκτιμηθεί η πληρότητα του συνόλου των φωσφοπρωτεϊνών και p-sites που έχουν βρεθεί μέχρι τώρα.

Η Εικόνα 9 δείχνει την αύξηση του συνολικού αριθμού των μοναδικών p-sites και των φωσφοπρωτεϊνών στον οργανισμό *Saccharomyces cerevisiae* που ανακαλύφθηκαν/προσδιορίστηκαν μετά από κάθε πείραμα.

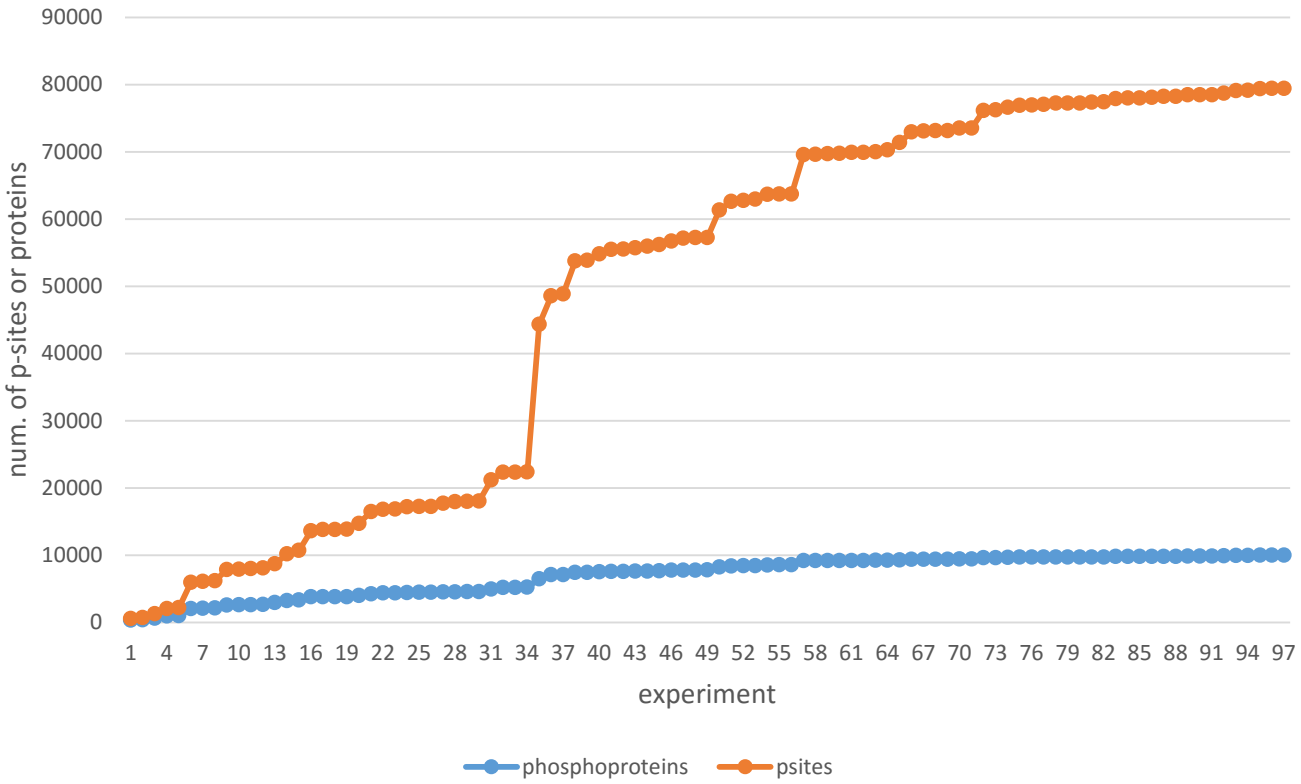


Εικόνα 9:Χρονολογική προσαύξηση των μοναδικών p-sites και φωσφοπρωτεϊνών σε Saccharomyces cerevisiae

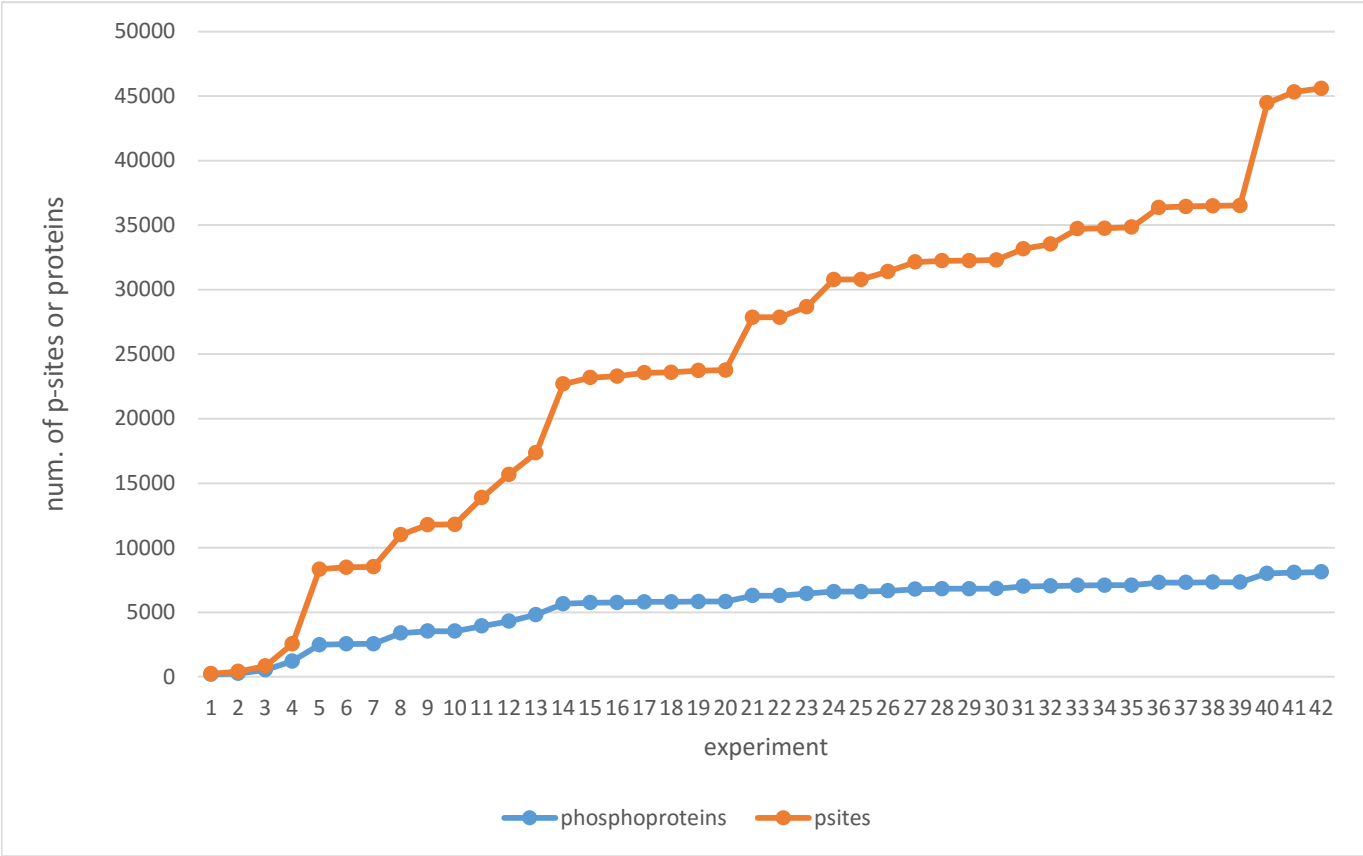
Μια πρόσφατη ανάλυση ενός συνόλου δεδομένων μεγάλης εμπιστοσύνης από 12 ανεξάρτητες μελέτες φωσφοπρωτεωμικής HTP, το οποίο ονομάζεται 12 HQ data set, βρήκε ότι το φωσφοπρωτέωμα της μαγιάς (*Saccharomyces cerevisiae*) σταδιακά πλησιάζει τον κορεσμό, επειδή η μέση επικάλυψη των φωσφοπεπτιδίων μεταξύ οποιωνδήποτε δύο πειραμάτων είναι 12%, ενώ η επικάλυψη μεταξύ φωσφοπρωτεϊνών είναι 28% (Amoutzias et al., 2012). Ομοίως, από τις συγκρίσεις επικάλυψης μεταξύ διαφορετικών μελετών HTP, εκτιμήθηκε ότι η ανίχνευση του ολοκληρωμένου φωσφοπρωτεώματος της ζύμης έχει φτάσει 80-90% κάλυψη (Beltrao et al., 2009).
Στο συμπέρασμα ότι οι φωσφοπρωτεΐνες που έχουν προσδιοριστεί μέχρι τώρα στις διάφορες έρευνες φωσφοπρωτεωμικής υψηλής ποιότητας φτάνουν σταδιακά σε κορεσμό και αποτελούν κατά ένα μεγάλο ποσοστό το συνολικό φωσφοπρωτέωμα της ζύμης καταλήγουν και τα δεδομένα στη παρούσα μελέτη.
Η έλλειψη κορεσμού όσον αφορά τα p-sites θα μπορούσε να αποδοθεί στον ανεπαρκή αριθμό των περιβαλλοντικών συνθηκών που έχουν δοκιμαστεί σε πειράματα, ή οφείλεται στις αδυναμίες των τεχνολογιών και πρωτοκόλλων φωσφοπρωτεωμικής που χρησιμοποιούνται.
Όντως 7792 από τα συνολικά 14339 μοναδικά p-sites (54%) που συλλέχθηκαν για τον οργανισμό *Saccharomyces cerevisiae* έχουν εντοπιστεί μία μόνο φορά, γεγονός που επιβεβαιώνει την ανάγκη για την περαιτέρω ανάπτυξη και βελτίωση των τεχνολογιών που χρησιμοποιούνται σε πειράματα φωσφοπρωτεωμικής. Αντίθετα τα 4239 από τα 14339 μοναδικά p-sites (30%) εντοπίστηκαν σε 3 ή περισσότερα πειράματα, τα οποία μπορούν να ομαδοποιηθούν σε μία υποκατηγορία υψηλής σημαντικότητας.

Παρόμοιες αναλύσεις με εκείνες που πραγματοποιούνται στο πρωτέωμα του οργανισμού *S. cerevisiae* έγιναν και για τα άλλα τέσσερα είδη. Τα αποτελέσματα παρουσιάζονται στην Εικόνα 10 (*Homo sapiens*), στην Εικόνα 11 (*Mus musculus*), στην Εικόνα 12 (*Rattus norvegicus*) και στην

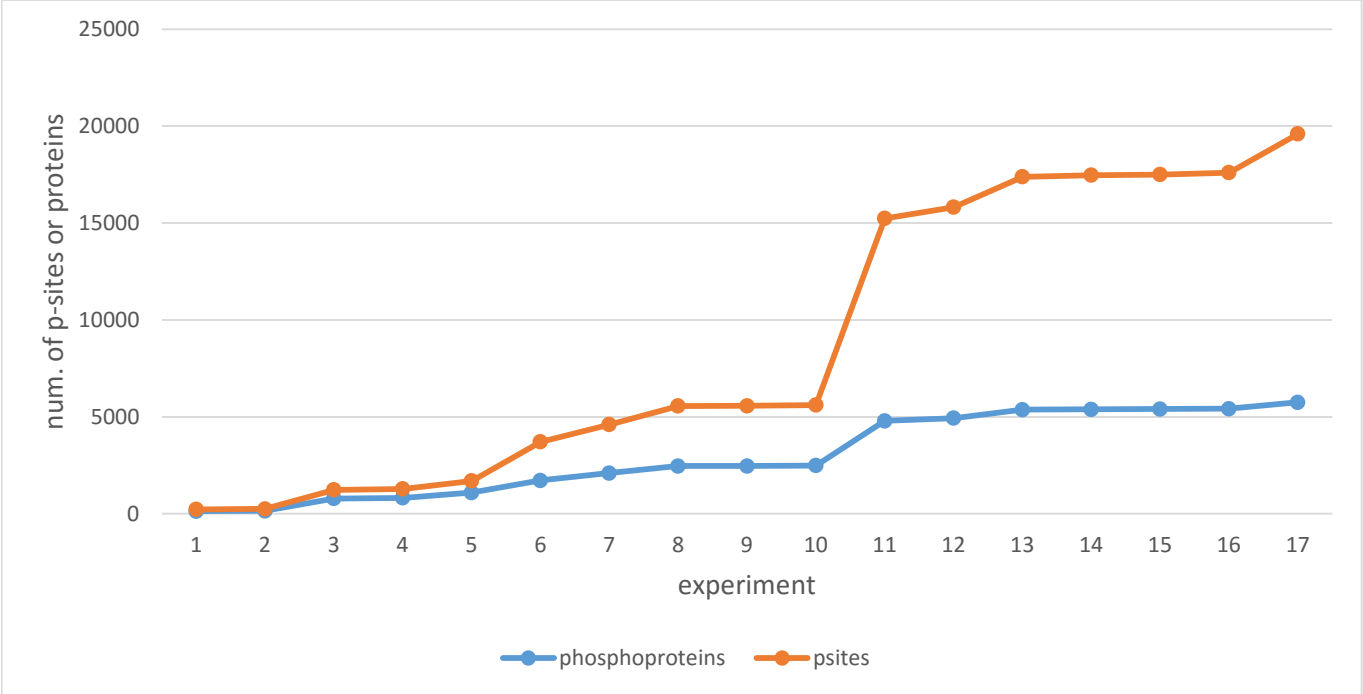
Εικόνα 13 (*Arabidopisis thaliana*). Διαπιστώνεται ότι στον άνθρωπο, ποντικό και αρουραίο ισχύουν τα ίδια βασικά συμπεράσματα που βρέθηκαν για την μαγιά, δηλαδή ότι το φωσφοπρωτέωμα πλησιάζει σε κορεσμό, ενώ ο αριθμός των θέσεων φωσφορυλίωσης όχι. Στην *Arabidopsis thaliana*, ούτε το φωσφοπρωτέωμα, ούτε ο αριθμός των θέσεων φωσφορυλίωσης πλησιάζει σε κορεσμό. Αυτό πιθανώς οφείλεται στον σχετικά μικρό αριθμό και μέγεθος των επιμέρους δεδομένων, όπως επίσης και στην πολυπλοκότητα αυτού του πολυκύτταρου οργανισμού.



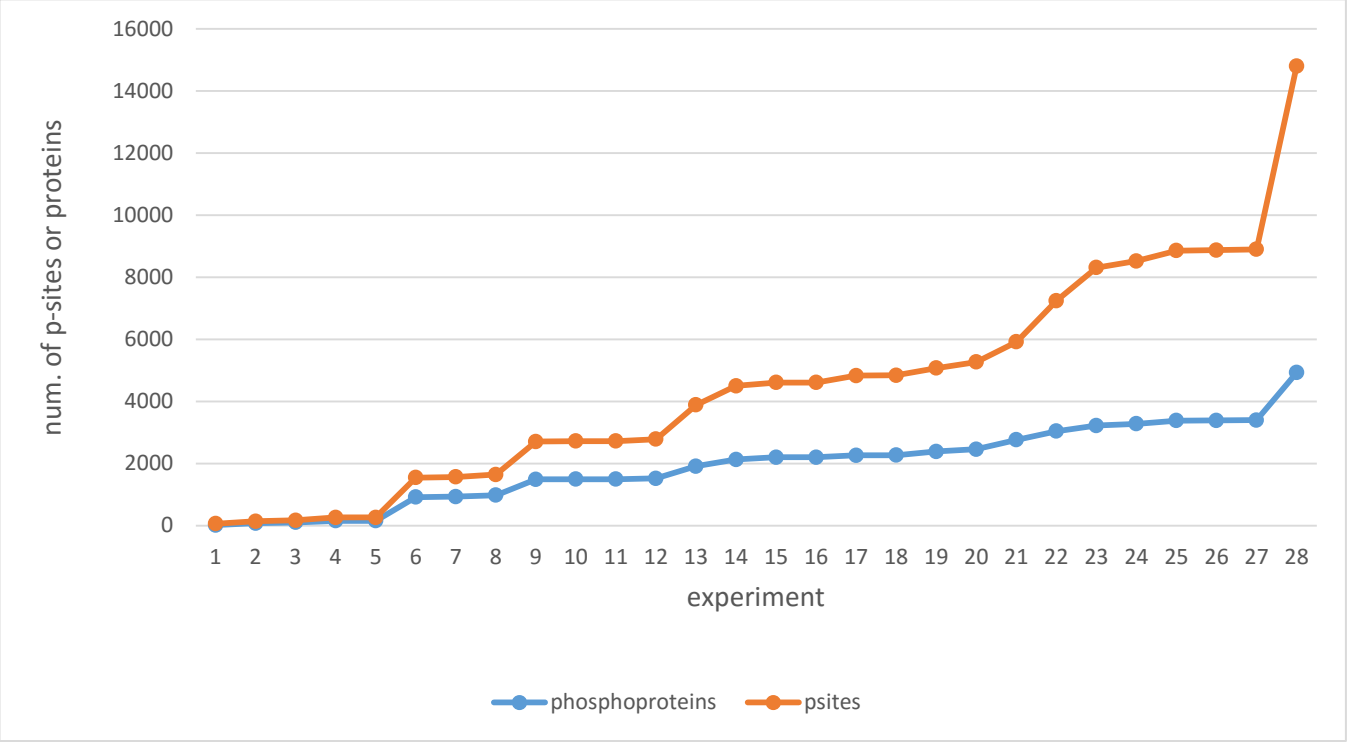
Εικόνα 10:Χρονολογική αύξηση των μοναδικών p-sites και φωσφοπρωτεϊνών σε Homo sapiens



Εικόνα 11:Χρονολογική αύξηση των μοναδικών p-sites και φωσφοπρωτεϊνών σε Mus musculus



Εικόνα 12:Χρονολογική αύξηση των μοναδικών p-sites και φωσφοπρωτεϊνών σε Rattus norvegicus



Εικόνα 13:Χρονολογική αύξηση των μοναδικών p-sites και φωσφοπρωτεϊνών σε Arabidopsis thaliana

Ανάλυση εμπλουτισμού γονιδιακών οντολογιών (Gene ontology enrichment).

Η αυξανόμενη πολυπλοκότητα των λειτουργικών δεδομένων γονιδιωματικής, οδήγησε επίσης στην ανάπτυξη μεθόδων και εργαλείων για την ενοποίηση των δεδομένων και την απεικόνιση τους. Το Cytoscape (Shannon et al., 2003) είναι μια πλατφόρμα λογισμικού ανοικτού κώδικα για την οπτικοποίηση των δικτύων μοριακής αλληλεπίδρασης και την ενσωμάτωση αυτών των αλληλεπιδράσεων με τα προφίλ της γονιδιακής έκφρασης και άλλα δεδομένα της λειτουργικής γονιδιωματικής. Η πλατφόρμα Cytoscape στηρίζει ενεργά την ανάπτυξη των εργαλείων plugin που επεκτείνει τη λειτουργικότητα του. Ένα τέτοιο παράδειγμα plugin αποτελεί το εργαλείο οντολογία γονιδίων βιολογικών δικτύων (Biological Networks Gene Ontology-Bingo). Το Bingo αξιολογεί την υπερεκπροσώπηση/εμπλουτισμό των κατηγοριών GO ενός βιολογικού δικτύου, ή οποιαδήποτε άλλου συνόλου γονιδίων (Maere et al., 2005).
Με τη χρήση του Cytoscape για το σύνολο των δεδομένων που ανακτήσαμε παρατηρήσαμε ποιες κατηγορίες GO υπερεκπροσωπούνται σε κάθε έναν οργανισμό ξεχωριστά.

Πίνακας 7:Υπερεκπροσώπηση κατηγοριών GO ανά οργανισμό

Homo sapiens	*Mus musculus*	*R. norvegicus*	*A. thaliana*	*S. cerevisiae*
intracellular	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	protein binding	plasma membrane	nucleus
intracellular part	nitrogen compound metabolic process	intracellular part	intracellular	site of polarized growth
protein binding	regulation of metabolic process	intracellular	membrane	transcription
binding	nucleotide binding	Binding	Cytoplasm	cellular bud
intracellular organelle	regulation of cellular metabolic process	cytoplasm	Cytosol	cell cycle
organelle	biological regulation	membrane-bounded organelle	Binding	signal transduction
nucleus	nucleic acid metabolic process	intracellular membrane-bounded organelle	nucleotide binding	cell cortex
cytoplasm	regulation of primary metabolic process	intracellular organelle	Nucleus	response to stress
cellular process	regulation of biological process	organelle	kinase activity	cytoskeleton organization
cellular macromolecule metabolic process	organelle part	cellular process	Cell	protein kinase activity

Συζήτηση

Το βασικό ερώτημα που θέλουμε να απαντήσουμε όταν βλέπουμε τα αποτελέσματα οποιασδήποτε επιστημονικής μελέτης είναι «Μπορώ να εμπιστευτώ τα δεδομένα της συγκεκριμένης ανάλυσης;». Οι συντάκτες και οι κριτές (peer reviewers) των ιατρικών και επιστημονικών περιοδικών δεν έχουν συνήθως το χρόνο για να εκτελέσουν μια σωστή αξιολόγηση των δεδομένων. Το πρόβλημα αυτό επιδεινώνεται από το γεγονός ότι τα σύνολα δεδομένων και οι αναλύσεις αυτών γίνονται όλο και πιο περίπλοκες, ο ρυθμός υποβολής μελετών σε επιστημονικά περιοδικά συνεχίζει να αυξάνεται και οι απαιτήσεις από άποψη στατιστικής αυξάνονται κατακόρυφα. Οι πιέσεις αυτές έχουν μειώσει την αποτελεσματικότητα της αξιολόγησης των ερευνητικών μελετών από τα περιοδικά. Αυτό οδηγεί σε ένα ντόμινο παραγωγής πιθανώς λανθασμένων συμπερασμάτων, εφόσον τα δημοσιευμένα σε περιοδικά αποτελέσματα χρησιμοποιούνται για μετα-ανάλυση από άλλες ερευνητικές ομάδες χωρίς να αναρωτηθούν για την αξιοπιστία τους.
Ο καλύτερος τρόπος για να αποφευχθεί η κακή ποιότητα των δεδομένων στην επιστημονική βιβλιογραφία είναι με την αύξηση του αριθμού των εκπαιδευμένων αναλυτών δεδομένων στην επιστημονική κοινότητα και με τη δημιουργία όλο και περισσότερων και πιο αποτελεσματικών λογισμικών στατιστικής που παράγουν κάποιου είδους σκορ που οδηγεί σε συγκεκριμένο βαθμό βεβαιότητας των αποτελεσμάτων (Leek and Peng, 2015).
Η δημιουργία τεχνικών και εργαλείων λογισμικού για την ερμηνεία και αξιολόγηση των δεδομένων αποτελεί ένα πεδίο εμπειρογνωμόνων, ανθρώπων που είναι γνώστες, όχι μόνο για τα οφέλη των μεθόδων αυτών αλλά και των πραγματικών και δυνητικών τους αδυναμιών. Τώρα, καθώς οι τεχνικές φασματομετρίας μάζας και τα εργαλεία Πρωτεωμικής γίνονται όλο και πιο διαθέσιμα και προσβάσιμα, ένα πολύ ευρύτερο φάσμα ερευνητών εφαρμόζει αυτές τις μεθοδολογίες, συχνά με σημαντικά λιγότερη κατανόηση των περιορισμών που επηρεάζουν την αξιοπιστία και τη σημασία των αποτελεσμάτων.
Ιδανικά, η ερευνητική κοινότητα που ασχολείται με την Πρωτεωμική πρέπει να θεσπίσει κριτήρια για την ταυτοποίηση των πρωτεϊνών με φασματομετρία μάζας που πρέπει να εφαρμόζονται από όλους τους ερευνητές. Καθώς όμως η Πρωτεωμική παραμένει ένας ταχέως αναπτυσσόμενος τομέας με πολλές διαφορετικές πειραματικές προσεγγίσεις και διαφορετικούς τρόπους αναζήτησης και ερμηνείας των δεδομένων, είναι δύσκολο να διαδοθούν αυστηροί και άμεσοι κανόνες με καθολική εφαρμογή.
Επομένως, είναι απαραίτητο για τον τομέα της Πρωτεωμικής να αναπτυχθούν και να εφαρμοστούν γενικά εργαλεία και λύσεις σε αυτά τα προβλήματα ώστε να οδηγούν σε ακριβή και αναπαραγωγίσιμα αποτελέσματα. Σε αντίθετη περίπτωση, εισάγονται και διαδίδονται λάθη στη βιβλιογραφία, καθιστώντας δύσκολο για τους κριτές και τους αναγνώστες να αξιολογήσουν τα συμπεράσματα μιας έρευνας ή να συγκρίνουν τα αποτελέσματα μεταξύ διαφόρων μελετών.
Στο εγγύς μέλλον, οι πειραματικές διαδικασίες στον τομέα της Πρωτεωμικής θα γίνουν ρουτίνα και αναπόφευκτα θα οδηγήσουν σε ακόμη μεγαλύτερες και ποικίλες ποσότητες δεδομένων. Ως εκ τούτου, η επιστημονική κοινότητα πρέπει να συμφωνήσει και να δεσμευτεί σε μια κοινή μορφή για τη δημοσίευση το σχεδιασμό και την ανάλυση αυτών των μελετών, που θα διασφαλίσει τη συμβατότητα, επαναληψιμότητα και την επαναχρησιμοποίηση των δεδομένων που προκύπτουν.
Η επαναχρησιμοποίηση των δεδομένων δημοσιευμένων μελετών απαιτεί ακριβή και ολοκληρωμένη αναφορά των στοιχείων που περιγράφουν τον πειραματικό σχεδιασμό, την απόκτηση και την προετοιμασία του δείγματος, τα πρωτόκολλα του οργάνου που χρησιμοποιήθηκε και τα στάδια επεξεργασίας των δεδομένων.
Ένα τέτοιο παράδειγμα επαναχρησιμοποίησης των δεδομένων πραγματοποιήθηκε στην παρούσα διπλωματική που είχε ως σκοπό τη συλλογή ενός συνόλου δεδομένων Φωσφοπρωτεωμικής από τη Βιβλιογραφία με φιλτράρισμα του θορύβου και της χαμηλής

ποιότητας δεδομένων. Πιο συγκεκριμένα, σκοπός της είναι η εύρεση των έγκυρων σημείων φωσφορυλίωσης σε πρωτεϊνικές αλληλουχίες που έχουν ήδη δημοσιευθεί.
Επειδή οι τρέχουσες ερευνητικές δραστηριότητες στον τομέα της πρωτεωμικής παράγουν πλούτο δεδομένων για ένα ευρύ φάσμα ιστών και για διάφορες συνθήκες, πρέπει τα αποτελέσματα που παράγονται να οργανώνονται και να παρουσιάζονται σε μορφή που να τα καθιστά όσο το δυνατόν πιο εύκολα προσβάσιμα για περαιτέρω ανάλυση. Για το σκοπό αυτό, έχουν δημιουργηθεί πολλές Βάσεις Δεδομένων, που περιέχουν τόσο οπτικοποιημένα δεδομένα όσο και πληροφορίες με εκτεταμένους υπερ-συνδέσμους σε άλλες Βάσεις. Πολλές μελέτες που ασχολούνται είτε με το συνολικό πρωτέωμα ενός οργανισμού είτε με το λειτουργικό πρωτέωμα (που ορίζεται ως η πρωτεϊνική έκφραση υπό συγκεκριμένες συνθήκες περιβάλλοντος) έχουν ως σημείο αφετηρίας αυτές τις Βάσεις Δεδομένων.
Είναι πολύ σημαντικό ότι δημόσιες Βάσεις Δεδομένων με πληροφορίες μετα-μεταφραστικών τροποποιήσεων (PTM) μπορεί ακόμα να περιέχουν εντοπισμούς θέσεων στις οποίες η ύπαρξη μετα-μεταφραστικής τροποποίησης είτε δεν έχει επικυρωθεί με ειδικούς αλγόριθμους ή δεν έχει αναφερθεί κάποια πιθανότητα, και ως εκ τούτου περιλαμβάνουν ένα σημαντικό ποσοστό λανθασμένων ή χωρίς βεβαιότητα θέσεων φωσφορυλίωσης. Με την εισαγωγή των κατευθυντήριων γραμμών για τις συγκεκριμένες συνθήκες που πρέπει να ακολουθούν οι ερευνητικές ομάδες για την ετοιμασία του δείγματος και τη φασματομετρία μάζας και των ορίων που πρέπει να έχουν τα αποτελέσματά τους σχετικά με τη βεβαιότητα της πεπτιδικής ταυτοποίησης και τοποθέτησης της φωσφορυλίωσης θα μειωθούν τα λανθασμένα αποτελέσματα στο μέλλον. Σε συνδυασμό με τη δημιουργία αυτοματοποιημένου ελέγχου και μετα-ανάλυσης της ποιότητας των ήδη δημοσιευμένων δεδομένων η ποιότητα αυτών σε δημόσιες βάσεις δεδομένων θα βελτιωθεί σημαντικά.
Η μετα-ανάλυση επιτρέπει τη δημιουργία πιο ολοκληρωμένων ομάδων δεδομένων, την εύρεση των λανθασμένων λόγω μειωμένης βεβαιότητας αποτελεσμάτων, την εξαγωγή των πιο ισχυρών και σωστών δεδομένων, και την ανακάλυψη νέων προτύπων και σχέσεων.
Η εύρεση των θέσεων φωσφορυλίωσης των πρωτεϊνών έγινε ένας σημαντικός στόχος στην εφαρμοσμένη έρευνα, επειδή προσφέρει ένα μοναδικό εργαλείο για να διαλευκανθούν τα διάφορα μονοπάτια σηματοδότησης και οι κρίσιμοι κόμβοι που μπορεί να μεταβληθούν κατά τη διάρκεια ανάπτυξης μιας ασθένειας.
Στρατηγικές εύρεσης των φωσφοπεπτιδίων που δεν στηρίζονται σε φασματομετρία μάζας έχουν χρησιμοποιηθεί στο παρελθόν για να προσδιοριστούν υποστρώματα κινασών χωρίς όμως να έχουν τη δυνατότητα να προσδιορίσουν επακριβώς τη θέση φωσφορυλίωσης, να ανακαλύψουν νέες θέσεις, ή να διακρίνουν διαφορετικές θέσεις φωσφορυλίωσης εντός της ίδιας πρωτεΐνης.
Οι λόγοι αυτοί σε συνδυασμό με την ερευνητική ανάγκη για μελέτες υψηλής απόδοσης οδήγησαν στην πρωτοκαθεδρία των στρατηγικών που βασίζονται σε MS στο τομέα της φωσφοπρωτεωμικής.
Παρά το γεγονός ότι οι στρατηγικές που στηρίζονται στη φασματομετρία μάζας προσφέρουν το καλύτερο εργαλείο για να προσδιοριστούν και να χαρτογραφηθούν με ακρίβεια τα p-sites, ωστόσο υπάρχουν πολλές λεπτομέρειες που οδηγούν σε λάθος αποτελέσματα και γενικά περιπλέκουν την ανίχνευση του φωσφορυλιώσεων.
Η επιτυχής ταυτοποίηση φωσφοπεπτιδίων με MS εξαρτάται σε πρώτο στάδιο από την χρήση του κατάλληλου πρωτοκόλλου για την προετοιμασία του δείγματος για την παροχή επαρκούς εμπλουτισμού των φωσφοπεπτιδίων για αξιόπιστη ανίχνευση τους. Έτσι, ειδικές προφυλάξεις πρέπει να λαμβάνονται για να διατηρηθεί η ακεραιότητα του δείγματος κατά τη διάρκεια της διαδικασίας απομόνωσης των φωσφοπεπτιδίων.
Αρχικά η χαμηλή στοιχειομετρική αφθονία των φωσφορυλιώσεων στο πρωτέωμα και η πιθανότητα απώλειας της φωσφορικής ομάδας κατά τον κατακερματισμό στην φασματομετρία μάζας οδηγεί σε μείωση των πραγματικών φωσφορυλιώσεων στα αποτελέσματα. Σε μια δεδομένη στιγμή, μόνο ένα μικρό ποσοστό των πρωτεϊνών που υπάρχουν σε ένα κύτταρο είναι φωσφορυλιωμένο και η κατάσταση φωσφορυλίωσης της ίδιας πρωτεΐνης μπορεί να ποικίλει. Ως εκ τούτου, το στάδιο εμπλουτισμού για την απομόνωση των φωσφοπρωτεϊνών ή

φωσφοπεπτιδίων είναι απαραίτητο πριν την ανάλυση φασματομετρίας μάζας. Επιπλέον, ο εντοπισμός και χαρτογράφηση p-sites σε ένα δείγμα απαιτεί ιδιαίτερη προσοχή στη επιλογή της κατάλληλης μεθοδολογίας MS-MS. Η προσέγγιση CID δεν είναι απαραίτητα η καλύτερη επιλογή γιατί συχνά χάνεται πληροφορία μετά την κατάτμηση των πεπτιδίων που διαθέτουν φωσφορυλιωμένη σερίνη ή θρεονίνη. Πολλές διαφορετικές προσεγγίσεις κατακερματισμού έχουν δοκιμαστεί και προταθεί τα τελευταία χρόνια για την αντιμετώπιση αυτού του ζητήματος, συμπεριλαμβανομένων των MSA, HCD και ETD, αλλά μια πλήρης συναίνεση δεν έχει επιτευχθεί για το ποια από αυτές είναι η πιο αποτελεσματική. Φαίνεται ότι η επιτυχία μιας προσέγγισης έναντι κάποιας άλλης εξαρτάται από την πολυπλοκότητα του δείγματος, τη διάταξη LC και συγκεκριμένων ρυθμίσεων MS. Επίσης, πεπτίδια φωσφοτυροσίνης είναι ακόμη πιο δύσκολο να εντοπιστούν, τόσο λόγω του χαμηλότερου επιπέδου της φωσφορυλίωσης τυροσίνης σε σύγκριση με τη σερίνη και θρεονίνη και λόγω της δυναμικής της φύσης (Lombardi et al., 2015).

Δεν μπορούμε να μετρήσουμε με ακρίβεια ό, τι χρειαζόμαστε ή θέλουμε. Οι πρωτεΐνες και άλλα μικρά μόρια δεν μπορούν να ταυτοποιηθούν και να ποσοτικοποιηθούν με ακρίβεια με τα φασματόμετρα μάζας όπως αυτά έχουν εξελιχθεί μέχρι σήμερα. Γι’ αυτό και οι γνώσεις μας για την πρωτεωμική δεν έχουν ωριμάσει αρκετά.

Ωστόσο όταν στο μέλλον τα ερευνητικά μηχανήματα και εργαλεία είναι εξελιγμένα σε τέτοιο βαθμό ώστε να έχουμε με μεγάλη αξιοπιστία όλα τα δεδομένα πρωτεωμικής, η πιο έξυπνη μηχανή στον πλανήτη, ο ανθρώπινος εγκέφαλος, θα είναι ανίκανος να τα κατανοήσει όλα. Είναι δεδομένα μεγάλων διαστάσεων (high-dimensional data) που ως άνθρωποι απλά δεν μπορούμε να τα καταλάβουμε και να επεξεργαστούμε.

Ο ερευνητής Jun Wang υποστηρίζει ότι η τεχνητή νοημοσύνη είναι το μέλλον της γονιδιωματικής και πρωτεωμικής και ότι το μόνο πράγμα που μπορεί να καταλάβει πραγματικά την πολυπλοκότητα ενός ζωντανού κυττάρου, ή ενός ζωντανού οργανισμού, είναι μια νοημοσύνη πολύ μεγαλύτερη από την ανθρώπινη (Cyranoski, 2015).

Παρά το γεγονός ότι οι τεχνολογίες HTP αργά ή γρήγορα θα εξελιχθούν και θα ωριμάσουν σε επίπεδο που επιτρέπει την ανακάλυψη του συνολικού αριθμού των p-sites και των φωσφοπρωτεϊνών, η πραγματική πρόκληση που βρίσκεται μπροστά μας είναι να προσδιοριστεί ποια από αυτά έχουν λειτουργική επίδραση στο φαινότυπο (Landry et al., 2009, 2014; Lienhard, 2008). Μια τέτοια πρόκληση μπορεί να αντιμετωπιστεί μόνο με ένα συνδυασμό βιοπληροφορικής ανάλυσης και εξαιρετικά αυτοματοποιημένων πειραματικών διαδικασιών για την φαινοτυπική αξιολόγηση μεταλλάξεων πάνω στις θέσεις φωσφορυλίωσης (King et al., 2004, 2009).

Βιβλιογραφία

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

Banks, R.E., Dunn, M.J., Hochstrasser, D.F., Sanchez, J.C., Blackstock, W., Pappin, D.J., and Selby, P.J. (2000). Proteomics: new perspectives, new biomedical opportunities. Lancet Lond. Engl. *356*, 1749–1756.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007). Quantitative mass spectrometry in proteomics: a critical review. Anal. Bioanal. Chem. *389*, 1017–1031.

Barsnes, H., Huber, S., Sickmann, A., Eidhammer, I., and Martens, L. (2009). OMSSA Parser: an open-source library to parse and extract data from OMSSA MS/MS search results. Proteomics *9*, 3772–3774.

Beausoleil, S.A., Villén, J., Gerber, S.A., Rush, J., and Gygi, S.P. (2006). A probability-based approach for high-throughput protein phosphorylation analysis and site localization. Nat. Biotechnol. *24*, 1285–1292.

Beltrao, P., Trinidad, J.C., Fiedler, D., Roguev, A., Lim, W.A., Shokat, K.M., Burlingame, A.L., and Krogan, N.J. (2009). Evolution of Phosphoregulation: Comparison of Phosphorylation Patterns across Yeast Species. PLoS Biol *7*, e1000134.

Bianco, L., Mead, J.A., and Bessant, C. (2009). Comparison of Novel Decoy Database Designs for Optimizing Protein Identification Searches Using ABRF sPRG2006 Standard MS/MS Data Sets. J. Proteome Res. *8*, 1782–1791.

Bodenmiller, B., Mueller, L.N., Mueller, M., Domon, B., and Aebersold, R. (2007). Reproducible isolation of distinct, overlapping segments of the phosphoproteome. Nat. Methods *4*, 231–237.

Bunge, J., and Fitzpatrick, M. (1993). Estimating the Number of Species: A Review. J. Am. Stat. Assoc. *88*, 364–373.

Burnham, K.P., and Overton, W.S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika *65*, 625–633.

Carr, S., Aebersold, R., Baldwin, M., Burlingame, A., Clauser, K., and Nesvizhskii, A. (2004). The Need for Guidelines in Publication of Peptide and Protein Identification Data Working Group On Publication Guidelines For Peptide And Protein Identification Data. Mol. Cell. Proteomics *3*, 531–533.

Cash, P. (2002). Proteomics: the protein revolution. Biol. Lond. Engl. *49*, 58–62.

Chalkley, R.J., and Clauser, K.R. (2012). Modification site localization scoring: strategies and performance. Mol. Cell. Proteomics MCP *11*, 3–14.

Chao, A. (2001). An overview of closed capture-recapture models. J. Agric. Biol. Environ. Stat. *6*, 158–175.

Craig, R., and Beavis, R.C. (2004). TANDEM: matching proteins with tandem mass spectra. Bioinforma. Oxf. Engl. *20*, 1466–1467.

Craig, R., Cortens, J.C., Fenyo, D., and Beavis, R.C. (2006). Using annotated peptide mass spectrum libraries for protein identification. J. Proteome Res. *5*, 1843–1849.

Creasy, D.M., and Cottrell, J.S. (2002). Error tolerant searching of uninterpreted tandem mass spectrometry data. Proteomics *2*, 1426–1434.

Cyranoski, D. (2015). Exclusive: Genomics pioneer Jun Wang on his new AI venture. Nature.

Dasari, S., Chambers, M.C., Slebos, R.J., Zimmerman, L.J., Ham, A.-J.L., and Tabb, D.L. (2010). TagRecon: high-throughput mutation identification through sequence tagging. J. Proteome Res. *9*, 1716–1726.

Deutsch, E.W. (2012). File formats commonly used in mass spectrometry proteomics. Mol. Cell. Proteomics MCP *11*, 1612–1621.

Diz, A.P., Carvajal-Rodríguez, A., and Skibinski, D.O.F. (2011). Multiple Hypothesis Testing in Proteomics: A Strategy for Experimental Work. Mol. Cell. Proteomics MCP *10*.

Elias, J.E., Haas, W., Faherty, B.K., and Gygi, S.P. (2005). Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. Nat. Methods *2*, 667–675.

Eng, J.K., McCormack, A.L., and Yates, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. *5*, 976–989.

Fenyö, D., and Beavis, R.C. (2003). A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes. Anal. Chem. *75*, 768–774.

Frank, A., and Pevzner, P. (2005). PepNovo: de novo peptide sequencing via probabilistic network modeling. Anal. Chem. *77*, 964–973.

Frewen, B.E., Merrihew, G.E., Wu, C.C., Noble, W.S., and MacCoss, M.J. (2006). Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. Anal. Chem. *78*, 5678–5684.

Geer, L.Y., Markey, S.P., Kowalak, J.A., Wagner, L., Xu, M., Maynard, D.M., Yang, X., Shi, W., and Bryant, S.H. (2004). Open mass spectrometry search algorithm. J. Proteome Res. *3*, 958–964.

de Godoy, L.M.F., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Fröhlich, F., Walther, T.C., and Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature *455*, 1251–1254.

Gooley, A.A., and Packer, N.H. (1997). The Importance of Protein Co- and Post-Translational Modifications in Proteome Projects. In Proteome Research: New Frontiers in Functional Genomics, D.M.R. Wilkins, P.K.L. Williams, D.R.D. Appel, and P.D.F. Hochstrasser, eds. (Springer Berlin Heidelberg), pp. 65–91.

Griffiths, J. (2008). A Brief History of Mass Spectrometry. Anal. Chem. *80*, 5678–5683.

Grimsrud, P.A., Swaney, D.L., Wenger, C.D., Beauchene, N.A., and Coon, J.J. (2010). Phosphoproteomics for the masses. ACS Chem. Biol. *5*, 105–119.

Han, J., Kamber, M., and Pei, J. (2011). Data Mining: Concepts and Techniques, Third Edition (Burlington, MA: Morgan Kaufmann).

Hanash, S. (2003). Disease proteomics. Nature *422*, 226–232.

Huber, L.A. (2003). Is proteomics heading in the wrong direction? Nat. Rev. Mol. Cell Biol. *4*, 74–80.

Hunter, T. (2000). Signaling--2000 and beyond. Cell *100*, 113–127.

Käll, L., Storey, J.D., MacCoss, M.J., and Noble, W.S. (2008). Posterior error probabilities and false discovery rates: two sides of the same coin. J. Proteome Res. *7*, 40–44.

Kanshin, E., Michnick, S., and Thibault, P. (2012). Sample preparation and analytical strategies for large-scale phosphoproteomics experiments. Semin. Cell Dev. Biol. *23*, 843–853.

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

Kicman, A.T., Parkin, M.C., and Iles, R.K. (2007). An introduction to mass spectrometry based proteomics-detection and characterization of gonadotropins and related molecules. Mol. Cell. Endocrinol. *260-262*, 212–227.

Kim, S., Mischerikow, N., Bandeira, N., Navarro, J.D., Wich, L., Mohammed, S., Heck, A.J.R., and Pevzner, P.A. (2010). The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. Mol. Cell. Proteomics MCP *9*, 2840–2852.

King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G.K., Bryant, C.H., Muggleton, S.H., Kell, D.B., and Oliver, S.G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. Nature *427*, 247–252.

King, R.D., Rowland, J., Oliver, S.G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L.N., et al. (2009). The automation of science. Science *324*, 85–89.

Koziol, J.A., Feng, A.C., and Schnitzer, J.E. (2006). Application of Capture−Recapture Models to Estimation of Protein Count in MudPIT Experiments. Anal. Chem. *78*, 3203–3207.

Lam, H., Deutsch, E.W., Eddes, J.S., Eng, J.K., King, N., Stein, S.E., and Aebersold, R. (2007). Development and validation of a spectral library searching method for peptide identification from MS/MS. Proteomics *7*, 655–667.

Lander, E.S. (2011). Initial impact of the sequencing of the human genome. Nature *470*, 187–197.

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

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

Leek, J.T., and Peng, R.D. (2015). Opinion: Reproducible research can still be wrong: Adopting a prevention approach. Proc. Natl. Acad. Sci. *112*, 1645–1646.

Lesley, S.A. (2001). High-throughput proteomics: protein expression and purification in the postgenomic world. Protein Expr. Purif. *22*, 159–164.

Liebler, D. (2001). Introduction to Proteomics: Tools for the New Biology (Totowa, NJ: Humana Press).

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

Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M., and Yates, J.R. (1999). Direct analysis of protein complexes using mass spectrometry. Nat. Biotechnol. *17*, 676–682.

Lombardi, B., Rendell, N., Edwards, M., Katan, M., and Zimmermann, J.G. (2015). Evaluation of phosphopeptide enrichment strategies for quantitative TMT analysis of complex network dynamics in cancer-associated cell signalling. EuPA Open Proteomics *6*, 10–15.

Luscombe, N.M., Greenbaum, D., and Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. Methods Inf. Med. *40*, 346–358.

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., and Lajoie, G. (2003). PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. RCM *17*, 2337–2342.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics *21*, 3448–3449.

Mann, M., and Wilm, M. (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal. Chem. *66*, 4390–4399.

Marcotte, E.M. (2007). How do shotgun proteomics algorithms identify proteins? Nat. Biotechnol. *25*, 755–757.

McHugh, L., and Arthur, J.W. (2008). Computational methods for protein identification from mass spectrometry data. PLoS Comput. Biol. *4*, e12.

Mount, D. (2004). Bioinformatics: Sequence and Genome Analysis (Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press).

Nesvizhskii, A.I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J. Proteomics *73*, 2092–2123.

Nesvizhskii, A.I., Vitek, O., and Aebersold, R. (2007). Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat. Methods *4*, 787–797.

Oliver, S.G., van der Aart, Q.J., Agostoni-Carbone, M.L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J.P., and Benit, P. (1992). The complete DNA sequence of yeast chromosome III. Nature *357*, 38–46.

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

Olsen, J.V., Ong, S.-E., and Mann, M. (2004). Trypsin cleaves exclusively C-terminal to arginine and lysine residues. Mol. Cell. Proteomics MCP *3*, 608–614.

Perkins, D.N., Pappin, D.J.C., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. ELECTROPHORESIS *20*, 3551–3567.

Pevtsov, S., Fedulova, I., Mirzaei, H., Buck, C., and Zhang, X. (2006). Performance evaluation of existing de novo sequencing algorithms. J. Proteome Res. *5*, 3018–3028.

Sadygov, R.G., and Yates, J.R. (2003). A Hypergeometric Probability Model for Protein Identification and Validation Using Tandem Mass Spectral Data and Protein Sequence Databases. Anal. Chem. *75*, 3792–3798.

Savitski, M.M., Lemeer, S., Boesche, M., Lang, M., Mathieson, T., Bantscheff, M., and Kuster, B. (2011). Confident phosphorylation site localization using the Mascot Delta Score. Mol. Cell. Proteomics MCP *10*, M110.003830.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. *13*, 2498–2504.

Shen, Y., Tolić, N., Zhao, R., Pasa-Tolić, L., Li, L., Berger, S.J., Harkewicz, R., Anderson, G.A., Belov, M.E., and Smith, R.D. (2001). High-throughput proteomics using high-efficiency multiple-capillary liquid chromatography with on-line high-performance ESI FTICR mass spectrometry. Anal. Chem. *73*, 3011–3021.

Shokouhi, M., Zobel, J., Scholer, F., and Tahaghoghi, S.M.M. (2006). Capturing Collection Size for Distributed Non-cooperative Retrieval. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (New York, NY, USA: ACM), pp. 316–323.

Solari, F.A., Dell'Aica, M., Sickmann, A., and Zahedi, R.P. (2015). Why phosphoproteomics is still a challenge. Mol. Biosyst. *11*, 1487–1493.

Tabb, D.L., Fernando, C.G., and Chambers, M.C. (2007). MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J. Proteome Res. *6*, 654–661.

Tanner, S., Shu, H., Frank, A., Wang, L.-C., Zandi, E., Mumby, M., Pevzner, P.A., and Bafna, V. (2005). InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal. Chem. *77*, 4626–4639.

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

Tran, B.Q., Hernandez, C., Waridel, P., Potts, A., Barblan, J., Lisacek, F., and Quadroni, M. (2011). Addressing Trypsin Bias in Large Scale (Phospho)proteome Analysis by Size Exclusion Chromatography and Secondary Digestion of Large Post-Trypsin Peptides. J. Proteome Res. *10*, 800–811.

Walsh, C. (2006). Posttranslational Modification of Proteins: Expanding Nature's Inventory (Roberts and Company Publishers).

Wasinger, V.C., Cordwell, S.J., Cerpa-Poljak, A., Yan, J.X., Gooley, A.A., Wilkins, M.R., Duncan, M.W., Harris, R., Williams, K.L., and Humphery-Smith, I. (1995). Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. Electrophoresis *16*, 1090–1094.

Wolters, D.A., Washburn, M.P., and Yates, J.R. (2001). An automated multidimensional protein identification technology for shotgun proteomics. Anal. Chem. *73*, 5683–5690.

Wu, C., Tran, J.C., Zamdborg, L., Durbin, K.R., Li, M., Ahlf, D.R., Early, B.P., Thomas, P.M., Sweedler, J.V., and Kelleher, N.L. (2012). A Protease for Middle Down Proteomics. Nat. Methods *9*, 822–824.

Wu, R., Dephoure, N., Haas, W., Huttlin, E.L., Zhai, B., Sowa, M.E., and Gygi, S.P. (2011). Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes. Mol. Cell. Proteomics MCP *10*, M111.009654.

Yates, J.R. (2004). Mass spectral analysis in proteomics. Annu. Rev. Biophys. Biomol. Struct. *33*, 297–316.

Zhang, J., Xin, L., Shan, B., Chen, W., Xie, M., Yuen, D., Zhang, W., Zhang, Z., Lajoie, G.A., and Ma, B. (2012). PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. Mol. Cell. Proteomics MCP *11*, M111.010587.

Zhang, W., Li, F., and Nie, L. (2010). Integrating multiple "omics" analysis for microbial biology: application and methodologies. Microbiol. Read. Engl. *156*, 287–301.

Zhang, Y., Fonslow, B.R., Shan, B., Baek, M.-C., and Yates, J.R. (2013). Protein analysis by shotgun/bottom-up proteomics. Chem. Rev. *113*, 2343–2394.

(2002). Current Topics in Computational Molecular Biology (Cambridge, Mass: A Bradford Book).

Βλαχάβας Ι, Κεφαλάς Π, Βασιλειάδης Ν, Ρεφανίδης Ι, Κόκκορας Φ και Σακελλαρίου Η(2002). Τεχνητή Νοημοσύνη, Εκδόσεις Γαρταγάνη

Παπαϊωάννου Στέλλα-Συλβάνα (2007). Βιοπληροφορική-Βάσεις Δεδομένων με γενετικά στοιχεία από μελέτες ψαριών – Αλεξάνδρειο Τεχνολογικό εκπαιδευτικό ίδρυμα Θεσσαλονίκης.