ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΒΙΟΧΗΜΕΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΣΤΑΥΡΟΥΛΑ ΜΑΚΡΗ

ΕΦΑΡΜΟΓΕΣ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ

ΜΟΡΙΑΚΗ ΓΕΝΕΤΙΚΗ

ΔΙΑΓΝΩΣΤΙΚΟΙ ΔΕΙΚΤΕΣ

ΛΑΡΙΣΑ 2022

ΠΡΟΒΛΕΨΗ ΜΕΤΑΜΕΤΑΦΡΑΣΤΙΚΩΝ ΤΡΟΠΟΠΟΙΗΣΕΩΝ ΠΡΩΤΕΪΝΩΝ ΜΕ ΜΕΘΟΔΟΥΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

PREDICTING POST-TRANSLATIONAL MODIFICATIONS OF PROTEINS USING MACHINE LEARNING METHODS

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ

Αμούτζιας Γρηγόριος, Αναπληρωτής Καθηγητής Βιοπληροφορικής με έμφαση στη Μικροβιολογία, ΤΒΒ, Π.Θ.
Μόσιαλος Δημήτριος, Αναπληρωτής Καθηγητής Βιοτεχνολογίας Μικροβίων, ΤΒΒ, Π.Θ.
Ιωάννης Ηλιόπουλος, Αναπληρωτής Καθηγητής Μοριακής Βιολογίας και Βιοπληροφορικής, Τμήμα Ιατρικής, Πανεπιστήμιο Κρήτης

Ευχαριστίες

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στο εργαστήριο Βιοπληροφορικής του τμήματος Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας, με επιβλέποντα καθηγητή τον κ. Αμούτζια Γρηγόριο, τον οποίο και θα επιθυμούσα να ευχαριστήσω για την εμπιστοσύνη που μου επέδειξε αναθέτοντάς μου το συγκεκριμένο θέμα. Θα ήθελα να τον ευχαριστήσω για την καθοδήγηση και την ελευθερία που μου έδωσε, η οποία με παρότρυνε να παίρνω πρωτοβουλίες, να αναλαμβάνω ευθύνες και να είμαι παραγωγική στη δουλειά μου. Επίσης, ευχαριστώ πολύ και τα υπόλοιπα δύο μέλης της τριμελούς επιτροπής, τους καθηγητές κ. Μόσιαλο Δημήτριο, Αναπληρωτή Καθηγητή Βιοτεχνολογία Μικροβίων του τμήματος μας αλλά και τον κ. Ιωάννη Ηλιόπουλο, Αναπληρωτή Καθηγητή Μοριακής Βιολογίας και Βιοπληροφορικής από το Τμήμα Ιατρικής του Πανεπιστήμιο Κρήτης.

Θα ήθελα να ευχαριστήσω, ιδιαιτέρως, τον υποψήφιο Διδάκτορα Νικολαΐδη Μάριο, που με βοήθησε καθ΄ όλη τη διάρκεια της εργασίας. Είναι σίγουρο πως η συμβολή του ήταν καθοριστική και χωρίς εκείνον η παρούσα εργασία δεν θα μπορούσε να είναι η ίδια. Τέλος, οφείλω ένα μεγάλο ευχαριστώ στην οικογένειά μου και περισσότερο στους γονείς μου Ελευθερία και Φώτιο για την ψυχολογική τους υποστήριξη κατά τη διάρκεια της πειραματικής διαδικασίας αλλά και της συγγραφής της παρούσας εργασίας.

ΠΕΡΙΛΗΨΗ

Η πρόσφατη πρόοδος στην Πρωτεωμική μεγάλης κλίμακας αποκάλυψε τον κρίσιμο και γενικό ρόλο της πρωτεϊνικής μεθυλίωσης σε πολλές κυτταρικές διεργασίες καθώς και στον καρκίνο και σε άλλες ασθένειες. Ωστόσο, μέχρι στιγμής έχει βρεθεί μόνο το 20-40% του συνολικού μεθυλ-πρωτεώματος στον άνθρωπο. Οι πειραματικές τεχνικές, εμφανίζουν δυσκολίες στον εντοπισμό πιθανών θέσεων μεθυλίωσης τόσο στη λυσίνη όσο και στην αργινίνη, ενώ παράλληλα είναι αρκετά δαπανηρές. Οι υπολογιστικές προβλέψεις των θέσεων μεθυλίωσης με μεθόδους μηχανικής μάθησης έχουν βοηθήσει στην αντιμετώπιση αυτών των περιορισμών. Ωστόσο, τίθεται το ζήτημα της διερεύνησης της ποιότητας των δεδομένων που χρησιμοποιούνται στις υπολογιστικές προβλέψεις και αφορούν πεπτίδια με μεθυλιωμένες λυσίνες ή αργινίνες, καθώς αυτά είναι πιθανό να προέρχονται από πειράματα με χαμηλό ποσοστό ειδικότητας εμπλουτισμού και λανθασμένη εκτίμηση του FDR. Έτσι, υπάρχει ανάγκη δημιουργίας συνόλων δεδομένων υψηλής ποιότητας, ώστε να χρησιμοποιηθούν από τα εργαλεία πρόβλεψης θέσεων μεθυλίωσης και να αποδίδουν με μεγαλύτερη ακρίβεια και να απομακρύνουν αποτελεσματικά το «θόρυβο». Χρησιμοποιώντας τα πιο πρόσφατα δεδομένα του ανθρώπου και εφαρμόζοντας αυστηρά κριτήρια φιλτραρίσματος δημιουργήθηκαν σύνολα δεδομένων υψηλής ποιότητας αλλά και σύνολα δεδομένων χαμηλής ποιότητας όπου το φιλτράρισμα δεν ήταν αυστηρό. Βρέθηκε ότι τα σύνολα δεδομένων υψηλής ποιότητας παρουσίασαν υψηλότερη απόδοση σε σχέση με αυτά χαμηλής ποιότητας. Επίσης βρέθηκε ότι το μήκος των πεπτιδίων δεν επηρεάζει την ικανότητα πρόβλεψης, μετά από ένα ελάχιστο μήκος.

Λέξεις κλειδιά :
Μεθυλ-πρωτεωμική, μεθυλίωση αργινίνης, μεθυλίωση λυσίνης, πρόβλεψη μεθυλίωση

ABSTRACT

The latest advances in high-throughput Proteomics have revealed the crucial and proteome-wide role of protein methylation in many cellular processes as well as in cancer and other diseases. Nevertheless, only 20-40% of the total methyl-proteome has been identified in humans so far. The experimental approaches suffer difficulties in identification of lysine and arginine methylation sites. Computational prediction of methylation sites using machine learning methods have helped handle these limitations. However, the data used in computational prediction of peptides with methylated lysines or arginines raise doubts and it needs investigation, as these are likely to derive from experiments with low enrichment specificity and incorrect FDR estimation. Thus, there is a need for high-quality datasets which can be used by methylation site prediction tools, for more accurate performance and more effective noise reduction. By utilizing up to the latest human HTP datasets and applying stringent filtering criteria, high-quality datasets were generated as well as low-quality datasets in which were not applied stringent filtering criteria. It was found that the high-quality datasets have more accurate performance than the low-quality datasets. Also, the length of the peptides does not affect the prediction accuracy, after a minimum length.

Keywords:

Methyl-proteomics, arginine methylation, lysine methylation, methylation prediction

Περιεχόμενα

ΠΕΡΙΕΧΟΜΕΝΑ ΕΙΚΟΝΩΝ

ΠΕΡΙΕΧΟΜΕΝΑ ΠΙΝΑΚΩΝ

1. ΕΙΣΑΓΩΓΗ

1.1 Μετα-μεταφραστικές τροποποιήσεις

Η μετα-μεταφραστική τροποποίηση (Post-Translational Modification - PTM) των πρωτεϊνών, που είναι ένα από τα μεταγενέστερα στάδια της βιοσύνθεσης τους, αφορά τις αναστρέψιμες ή μη αναστρέψιμες χημικές αλλαγές που μπορεί να υποστούν οι πρωτεΐνες μετά τη μετάφραση. Με άλλα λόγια, τα PTM είναι χημικές τροποποιήσεις μιας πολυπεπτιδικής αλυσίδας που συμβαίνουν αφού το DNA μεταγραφεί σε RNA και μεταφραστεί σε πρωτεΐνη και αυξάνουν την πολυπλοκότητα της. Αυτές οι χημικές τροποποιήσεις κυμαίνονται από ενζυματική διάσπαση πεπτιδικών δεσμών έως ομοιοπολικές προσθήκες συγκεκριμένων χημικών ομάδων, λιπιδίων, υδατανθράκων ή ακόμα και ολόκληρων ολιγοπεπτιδίων σε πλευρικές αλυσίδες αμινοξέων. Η προσθήκη λειτουργικών ομάδων μπορεί να συμβεί σε πλούσιες σε ηλεκτρόνια ομάδες που δρουν ως πυρηνόφιλα, όπως υδροξυομάδες τυροσίνης, θρεονίνης, σερίνης, αμινομάδες αργινίνης και λυσίνης, καρβοξυλική ομάδα γλουταμικού και ασπαρτικού ή στα C ή N άκρα της πρωτεΐνης (Virág et al., 2020). Αυτές οι χημικές τροποποιήσεις μιας πολυπεπτιδικής αλυσίδας μετά τη βιοσύνθεσή της επεκτείνουν το εύρος της δομής και των ιδιοτήτων των αμινοξέων, και κατά συνέπεια, διαφοροποιούν τη δομή και τις λειτουργίες των πρωτεϊνών. Αν και το DNA τυπικά κωδικοποιεί 20 αμινοξέα, οι πρωτεΐνες περιέχουν περισσότερα από 140 διαφορετικά κατάλοιπα, λόγω διαφόρων PTMs. Η ικανότητα των πρωτεϊνών να υποβάλλονται σε PTM θεωρείται ως ένας από τους δύο γενικούς μηχανισμούς που επεκτείνουν την ικανότητα κωδικοποίησης ενός γονιδιώματος και να δημιουργεί εξαιρετικά διαφοροποιημένα πρωτεώματα. Αν και ορισμένες PTM μπορούν να βρεθούν σε προκαρυώτες, αυτές οι τροποποιήσεις είναι πολύ πιο συχνές σε ευκαρυωτικά κύτταρα, τα οποία τυπικά χαρακτηρίζονται από ένα ευρύ φάσμα τύπων PTM. Ορισμένες PTMs είναι άμεσα αναστρέψιμες με τη δράση συγκεκριμένων ενζύμων. Η αλληλεπίδραση μεταξύ τροποποιητικών και αποτροποποιητικών ενζύμων επιτρέπει τον γρήγορο και οικονομικό έλεγχο της πρωτεϊνικής λειτουργίας. Ένας παρόμοιος έλεγχος με αποικοδόμηση πρωτεϊνών και *de novo* σύνθεση θα έπαιρνε πολύ περισσότερο χρόνο και θα κόστιζε πολύ περισσότερη βιοενέργεια. (Uversky, 2013).

Τα PTM μπορεί να συμβούν σε οποιοδήποτε στάδιο της ζωής της πρωτεΐνης. Ορισμένες πρωτεΐνες τροποποιούνται λίγο μετά την ολοκλήρωση της μετάφρασής τους και πριν από τα τελικά στάδια της αναδίπλωσής τους. Αυτά τα πρώιμα PTMs μπορεί να επηρεάσουν την αποτελεσματικότητα της αναδίπλωσης πρωτεϊνών και τη σταθερότητα της πρωτεϊνικής διαμόρφωσης, ακόμη και να καθορίσουν τη μοίρα της πρωτεΐνης μέσω της μεταφοράς της σε διαφορετικά κυτταρικά διαμερίσματα. Άλλες πρωτεΐνες τροποποιούνται αφού ολοκληρωθεί η αναδίπλωση και ο εντοπισμός τους. Σε αυτή τη περίπτωση, τα PTM μπορούν να ενεργοποιήσουν ή να αδρανοποιήσουν καταλυτικές λειτουργίες ή να επηρεάσουν με άλλο τρόπο τη βιολογική δραστηριότητα της πρωτεΐνης (Uversky, 2013).

Υπάρχει μεγάλος αριθμός τύπων PTMs και καταλύονται κυρίως από ειδικά ένζυμα που αναγνωρίζουν συγκεκριμένες αλληλουχίες-στόχους σε συγκεκριμένες πρωτεΐνες. Σε ανώτερους ευκαρυώτες, έως και 5% των γονιδιωμάτων αναμένεται να κωδικοποιούν ένζυμα που σχετίζονται με τις μετα-μεταφραστικές τροποποιήσεις των πρωτεωμάτων. Συνολικά, έως και 300 μεταμεταφραστικές τροποποιήσεις πρωτεϊνών είναι γνωστό ότι συμβαίνουν φυσιολογικά. Οι πιο κοινές PTM είναι η ειδική διάσπαση πρόδρομων πρωτεϊνών, ο σχηματισμός δισουλφιδικών δεσμών, η ομοιοπολική προσθήκη ή αφαίρεση ομάδων χαμηλού μοριακού βάρους, το οποίο οδηγεί σε τροποποιήσεις όπως ακετυλίωση, αμίδωση, βιοτινυλίωση, κυστεϊνυλίωση, αποαμίδωση, φαρνεσυλίωση, φορμυλίωση, γερανυλγερανυλίωση, γλουταθειονυλίωση, γλυκοζυλίωση, υδροξυλίωση, μεθυλίωση, μονο-ADP-ριβοζυλίωση, μυριστουλίωση, οξείδωση, παλμιτοϋλίωση, φωσφορυλίωση, στεαροϋλίωση. Όλες οι πλευρικές αλυσίδες αμινοξέων είναι γνωστό ότι υφίστανται χημική διαφοροποίηση λόγω διαφόρων PTM. Ωστόσο, πιο συχνά, οι πρωτεϊνικές PTM συμβαίνουν σε πλευρικές αλυσίδες που μπορούν να δράσουν είτε ως ισχυρά (C, M, S, T, Y, K, H, R, D, E) ή ασθενή (N, Q) πυρηνόφιλα, ενώ τα υπόλοιπα κατάλοιπα (P, G, L, I, V, A, W, ΣΤ) σπάνια εμπλέκονται σε ομοιοπολικές τροποποιήσεις των πλευρικών αλυσίδων τους.(Uversky, 2013).

Μια πρωτεΐνη μπορεί να τροποποιηθεί από περισσότερους του ενός τύπους PTM ή να τροποποιηθεί πολλές φορές απο το ίδιο PTM σε διαφορετικά κατάλοιπα. Επιπλέον, τα PTMs μπορούν να διαχωριστούν χρονικά και μπορούν να συμβούν διαδοχικά σε μία μόνο θέση μιας πρωτεΐνης (Deribe et al., 2010). Η παρουσία διαφορετικών PTMs, που δρουν διαδοχικά ή/και συνεργατικά, φαίνεται να είναι ένας εξελικτικά διατηρημένος μηχανισμός για την έναρξη, τον τερματισμό ή τον ακριβή συντονισμό του αποτελέσματος των μονοπατιών σηματοδότησης. Αυτός ο μωσαϊκός καταρράκτης PTM δείχνει πώς μπορεί να δημιουργηθεί ένα σύνθετο δίκτυο σηματοδότησης. Είναι ιδιαίτερα σημαντικό ότι ένα εξελιγμένο γονιδιακό ρυθμιστικό σύστημα επιτρέπει την δημιουργία νέων μονοπατιών σηματοδότησης και κυτταρικών λειτουργιών (Deribe et al., 2010).

1.2 Πρωτεϊνική μεθυλίωση

Η μεταφορά της γενετικής πληροφορίας έχει περιγραφεί ως μια προς τα εμπρός ροή από το DNA στο RNA προς τις πρωτεΐνες. Ωστόσο, αυτός ο κλασικός ορισμός δεν καλύπτει τη βιολογική πολυπλοκότητα, ιδίως τον τρόπο με τον οποίο τα προϊόντα όπως οι πρωτεΐνες και οι μεταβολίτες δρουν στο DNA, το RNA και τις πρωτεΐνες με κληρονομικό τρόπο. Η μεθυλίωση βιολογικών προϊόντων όπως το DNA και οι πρωτεΐνες είναι αναμφισβήτητα μια πολύ σημαντική βιοχημική αντίδραση (Luo, 2015)

Η πρωτεϊνική μεθυλίωση είναι μια σημαντική μετα-μεταφραστική τροποποίηση (PTM) και εμπλέκεται σε όλες σχεδόν τις βασικές βιολογικές διεργασίες του κυττάρου (Murn & Shi, 2017). Συνήθως συμμετέχει στη μεταφορά πληροφοριών σε κυτταρικές οδούς μεταγωγής σήματος. Οι πρωτεΐνες μπορούν να μεθυλιωθούν τόσο στο N- και στο καρβοξυτελικό άκρο τους καθώς και σε πλευρικές αλυσίδες οκτώ καταλοίπων αμινοξέων. Αυτά είναι η λυσίνη (Lys ή K), η αργινίνη (Arg ή R), το ασπαρτικό (Asp ή D), το γλουταμικό (Glu ή E), η ιστιδίνη (His ή H), η ασπαραγίνη (Asn ή N), η γλουταμίνη (Gln ή Q) και η κυστεΐνη (Cys ή C) (Luo, 2018). Μεταξύ αυτών, η αργινίνη και η λυσίνη είναι τα πιο συχνά μεθυλιωμένα κατάλοιπα. Επιπλέον, διαφορετικές μορφές μεθυλίωσης θα μπορούσαν να συμβούν σε ένα κατάλοιπο (Wang et al., 2021).

Η πρωτεϊνική μεθυλίωση περιλαμβάνει την προσθήκη μιας μεθυλομάδας στην πολυπεπτιδική αλυσίδα. Τα ένζυμα που καταλύουν αυτές τις αντιδράσεις ονομάζονται πρωτεϊνικές μεθυλοτρανσφεράσες (Protein Methyltransferases, PMTs). Αυτά τα ένζυμα χρησιμοποιούν S-αδενοσυλ-L-μεθειονίνη (SAM ή AdoMet) ως δότη μεθυλομάδας και μπορούν να τροποποιήσουν μια ποικιλία πυρηνόφιλων ατόμων οξυγόνου, αζώτου και θείου στην πολυπεπτιδική αλυσίδα. Αυτές οι αντιδράσεις οδηγούν στη δημιουργία μεθυλεστέρων, μεθυλαμινών, μεθυλ-αμιδίων και σε άλλα παράγωγα στις πλευρικές αλυσίδες των καταλοίπων που μεθυλιώνουν, καθώς και στα κατάλοιπα που υπάρχουν στο N- ή καρβοξυλικό άκρο (Clarke, 1993).

1.2.1 Μεθυλίωση της αργινίνης

Η μεθυλίωση αργινίνης είναι μια μετα-μεταφραστική τροποποίηση (PTM) που διατηρείται σε όλους τους ευκαρυωτικούς οργανισμούς, από τους ζυμομύκητες έως τους ανθρώπους (Fulton et al., 2018). Αν και η μεθυλίωση της αργινίνης δεν μεταβάλλει το ηλεκτρικό φορτίο, αυξάνει τον όγκο των αμινοξέων και την υδροφοβικότητα των πρωτεϊνών, ρυθμίζοντας έτσι τον τρόπο με τον οποίο οι πρωτεΐνες αλληλεπιδρούν μεταξύ τους. Παράλληλα, συμμετέχει σε διάφορες κυτταρικές διεργασίες, συμπεριλαμβανομένου της επιδιόρθωσης βλαβών του DNA, της μεταγραφικής ρύθμισης και του μεταβολισμού του RNA. Έτσι, η μεθυλίωση της αργινίνης έχει μια βαθιά επίδραση σε ανθρώπινες ασθένειες όπως ο καρκίνος (Wei et al., 2021).

Η μεθυλίωση της αργινίνης καταλύεται από πρωτεϊνικές μεθυλτρανσφεράσες αργινίνης (PRMTs). Η μεθυλίωση της αργινίνης από τις PRMTs είναι μια εξαιρετικά ενεργοβόρα διαδικασία (12 μόρια ATP για κάθε μεθυλομάδα που προστίθεται) και συμβαίνει σε περισσότερο από το 10% όλων των ανθρώπινων πρωτεϊνών (Wei et al., 2021). Οι PRMTs καταλύουν τη μεταφορά μεθυλομάδων από το συν-υπόστρωμα S-αδενοσυλ-L-μεθειονίνη (AdoMet, SAM) στο άζωτο γουανιδίνης (ω-NG) ενός καταλοίπου πεπτιδυλ-αργινίνης, με αποτέλεσμα τον σχηματισμό μεθυλαργινίνης και S-αδενοσυλ- L-ομοκυστεΐνη (AdoHcy, SAH). Κατά την αντίδραση, τα PRMT χρησιμοποιούν ένα διαδοχικό σύνθετο μηχανισμό για να καταλύουν τη μεθυλίωση της αργινίνης, στον οποίο τόσο το SAM όσο και το υπόστρωμα αργινίνης απαιτείται να συνδεθούν στο ενεργό κέντρο των PRMT για να σχηματίσουν ένα τριμοριακό σύμπλοκο. Κατά τη δέσμευση, η γουανίνη στο κατάλοιπο αργινίνης βρίσκεται σε χωρική εγγύτητα με την αντιδρούσα μεθυλική ομάδα του SAM (Fulton et al., 2018).

Τα κύρια μέλη PRMT (δηλ. PRMT1 και PRMT5) εκφράζονται παντού στα κύτταρα και μέχρι στιγμής δεν έχει αναφερθεί γενικός μηχανισμός για τη ρύθμιση της δραστηριότητας PRMT (di Lorenzo & Bedford, 2011). Ως εκ τούτου, η ρύθμιση σε επίπεδο υποστρώματος μέσω αλληλεπιδράσεων PTM μπορεί να είναι ένας σημαντικός μηχανισμός για τη ρύθμιση των επιπέδων μεθυλίωσης αργινίνης (Fulton et al., 2018).

Υπάρχουν τρείς τύποι μεθυλαργινίνης που μπορούν να σχηματιστούν. Αυτοί είναι ω-N^G-μονομεθυλαργινίνη (MMA, Rme1), ω-N^G,N^G –ασύμμετρη-διμεθυλαργινίνη (ADMA, Rme2a) ω-N^G,N'^G-συμμετρική-διμεθυλαργίνινη (SDMA, Rme2s) (Tewary et al., 2019) Η μορφή της διμεθυλίωσης καθορίζεται από τη προσθήκη των δύο μεθυλομάδων. Αν οι δύο μεθυλομάδες προστίθενται στο ίδιο άτομο αζώτου στο τέλος της πλευρικής αλυσίδας της αργινίνης τότε έχουμε ασύμμετρη διμεθυλίωση. Στη περίπτωση που προστίθεται μία μεθυλομάδα σε κάθε ένα από τα δύο τελικά άτομα αζώτου τότε έχουμε συμμετρική διμεθυλίωση (Y.-H. Lee & Stallcup, 2009).

Στα σπονδυλωτά εμφανίζονται εννέα PRMTs ομαδοποιημένες σε τρεις τύπους σύμφωνα με την τελική μορφή των προϊόντων μεθυλαργινίνης. Τα PRMT τύπου I (PRMT-1, 2, 3, 4, 6 και 8) καταλύουν το σχηματισμό Rme1 και Rme2a. Τα PRMT τύπου II (PRMT-5 και 9) καταλύουν το σχηματισμό των Rme1 και Rme2s. Τέλος, ο τύπου III PRMT (PRMT7) μπορεί να δημιουργήσει μόνο Rme1. Επιπλέον, έχει περιγραφεί ένα σπάνια εμφανιζόμενο PRMT τύπου IV που οδηγεί σε μονομεθυλίωση του εσωτερικού δ-αζώτου της γουανιδίνης (Maron et al., 2021; Tewary et al., 2019). Το PRMT1, το πιο κυρίαρχο PRMT τύπου I στα κύτταρα θηλαστικών, αντιπροσωπεύει το 85% της κυτταρικής δραστηριότητας των PRMT και εμπλέκεται σε πολλές βιολογικές λειτουργίες (Tsai et al., 2011). Τόσο το PRMT1 όσο και το PRMT5, που θεωρούνται τα κύρια ένζυμα για τα Rme2a και Rme2s, αντίστοιχα, είναι απαραίτητα για τη βιωσιμότητα και την ανάπτυξη των κυττάρων.

Τα διαφορετικά PRMTs έχουν υψηλό βαθμό επικάλυψης στα υποστρώματά τους και υψηλές ομοιότητες μεταξύ των υποτιθέμενων μοτίβων μεθυλίωσης τους (H. H. Wei et al., 2021). Υπάρχουν κάποια μοτίβα αμινοξέων που αποτελούν υποστρώματα πολλαπλών PRMTs. Ίσως από τα πιο συχνά μοτίβα είναι τα RGG/RG, πλούσια σε γλυκίνη και αργινίνη, και αποτελεί στόχο των PRMT1, PRMT3, PRMT5, PRMT6 και PRMT8 (Blanc & Richard, 2017). Επιπλέον, κάποιοι προτιμούν να μεθυλιώνουν μοτίβα GAR, πλούσια σε γλυκίνη και αργινίνη, και άλλοι προτιμούν μοτίβα PGM, πλούσια σε προλίνη, γλυκίνη και μεθειονίνη (Tewary et al., 2019). Επίσης, νέα μοτίβα πλούσια σε SR, πλούσια σε DR, και πλούσια σε ER εντοπίστηκαν ως υποστρώματα PRMT (Cheng et al., 2007). Αυτό υποδηλώνει ότι και άλλες αλληλουχίες εκτός από το μοτίβο GR θα μπορούσαν να αναγνωριστούν ως θέσεις μεθυλίωσης αργινίνης και ότι οι διαφορετικές PRMTs έχουν διαφορετικές προτιμήσεις για το υπόστρωμά τους (Wei et al., 2021).

Η πλειονότητα των μοτίβων για τη μεθυλίωση αργινίνης είναι μικρά θραύσματα με χαμηλή πολυπλοκότητα αλληλουχίας, συμπεριλαμβανομένων των γνωστών πλούσιων σε GR μοτίβων και των πρόσφατα αναγνωρισμένων πλούσιων σε SR και ER μοτίβων. Επειδή οι περιοχές χαμηλής πολυπλοκότητας (π.χ., περιοχές πλούσιες σε GR και SR σε πρωτεΐνες δέσμευσης RNA) συνήθως σχηματίζουν μια μη δομική περιοχή, η αναγνώριση από τις PRMT πιθανότατα συμβαίνει στις μη δομημένες περιοχές των πρωτεϊνών. Αυτή η δέσμευση θα μπορούσε να εξηγήσει τις υψηλές επικαλύψεις μεταξύ των μορίων που δεσμεύονται στα διαφορετικά PRMTs, υποδηλώνοντας έναν λειτουργικό πλεονασμό μεταξύ των PRMTs (Wei et al., 2021). Ένα άλλο επαναλαμβανόμενο υπόστρωμα PRMT είναι το μοτίβο RXR δηλαδή δύο υπολείμματα αργινίνης που διαχωρίζονται από οποιοδήποτε αμινοξύ. Εκτός από αυτά τα κοινά μοτίβα, η μεθυλίωση αργινίνης βρίσκεται σε αλληλουχίες, όπως AKTRSS (ιστόνη H2AR17), VLRDNI (H4R23) και SVYRQQ (Mediator subunit MED12 R206). Επιπλέον, η μεθυλαργινίνη εμφανίζεται σε τμήματα πρωτεϊνών με χαμηλή δομική πολυπλοκότητα (Low Complexity) και σε περιοχές εμπλουτισμένες με θετικά φορτισμένα κατάλοιπα (Lorton & Shechter, 2019).

1.2.2 Μεθυλίωση Λυσίνης

Η μεθυλίωση της λυσίνης είναι μια δυναμική, αναστρέψιμη και ευρέως διαδεδομένη μετα-μεταφραστική τροποποίηση (PTM), που συμβάλλει στη ρύθμιση πολλών βιολογικών λειτουργιών (Han et al., 2019). Η μεθυλίωση της λυσίνης έχει παρατηρηθεί τόσο σε πυρηνικές όσο και σε κυτταροπλασματικές πρωτεΐνες και θεωρείται πλέον μια διαδεδομένη τροποποίηση σε ευκαρυώτες, προκαρυώτες και αρχαία (Lanouette et al., 2014). Η απορρύθμιση της μεθυλίωσης της λυσίνης σχετίζεται με διάφορες ανθρώπινες διαταραχές, ιδιαίτερα την ογκογένεση. Η ε-αμινομάδα της πρωτεϊνικης λυσίνης μπορεί να δεχτεί έως και τρεις μεθυλομάδες, με αποτέλεσμα τα κατάλοιπα λυσίνης να τροποποιούνται σε μία από τις τρείς μορφές μεθυλίωσης και να εμφανίζουν διαφορετικές λειτουργίες. Πιο συγκεκριμένα, μπορεί να προκύψει είτε μονομεθυλολυσίνη, διμεθυλολυσίνη και τριμεθυλολυσίνη (me1, me2 ή me3) (Eldjarn & Broughton, 1985). Κάθε τροποποίηση της λυσίνης καταλύεται ειδικά από πρωτεϊνικές μεθυλτρανσφεράσες λυσίνης (PKMTs) με τρόπο που εξαρτάται από τη SAM, η οποία αποτελεί τον δότη μεθυλομάδων, και μπορεί να αντιστραφεί από τις απομεθυλάσες λυσίνης (KDMs) μέσω μιας αντίδραση οξειδωτικής απομεθυλίωσης (Thinnes et al., 2014). Μέχρι σήμερα, έχουν αναφερθεί περισσότερες από 50 PKMTs και 20 KDMs (Han et al., 2019).

Υπάρχουν δύο κύριες ομάδες PKMTs, όπου και οι δύο χρησιμοποιούν SAM ως δότη μεθυλομάδας. Η πρώτη ομάδα PKMTs ομαδοποιεί τα ένζυμα που περιέχουν μια καταλυτική περιοχή SET (PMTs τάξης V). Τα ένζυμα της ομάδας αυτής κυρίως στοχεύουν στις λυσίνες των ιστονών (Han et al., 2019). Ο τομέας SET, χαρακτηρίζεται από τρεις περιοχές αναδιπλωμένες σε μια δομή που μοιάζει με αναδίπλωση β φύλλου και σχηματίζει την ενεργή θέση που αποτελείται από τα τέσσερα διατηρημένα μοτίβα GXG, YXG, NHXCXPN και ELXFDY (X. Cheng & Zhang, 2007; Qian & Zhou, 2006). Η σύνδεση του SAM και του υποστρώματος λαμβάνει χώρα σε κάθε πλευρά ενός καναλιού μεταφοράς μεθυλομάδας που σχηματίζεται από την παραπάνω δομή (Wilson et al., 2002). Ένα δίκτυο αρωματικών κατάλοιπων και δεσμών υδρογόνου σε αυτό το κανάλι μεταφοράς περιορίζει τους πιθανούς προσανατολισμούς του υποστρώματος λυσίνης, ελέγχοντας την ικανότητα των πρωτεϊνών της περιοχής SET να μεταφέρουν έναν συγκεκριμένο αριθμό μεθυλομάδων στο υπόστρωμα (Couture et al., 2008).

Με βάση τις ομοιότητες της αλληλουχίας και της οργάνωσης του τομέα SET, οι πρωτεΐνες που περιέχουν το τομέα SET μπορούν να χωριστούν περαιτέρω σε επτά οικογένειες, οι οποίες είναι οι SUV3/9, SET1, SET2, SMYD, EZ, SUV4-20 και RIZ. Μέλη του SUV3/9, SETDB1, SET1, SET2, SMYD και EZ μεθυλιώνουν υποστρώματα ιστονικών και μη πρωτεϊνών, ενώ τα υποστρώματα για τις οικογένειες SUV4-20 και RIZ περιορίζονται σε ιστονικές πρωτεΐνες (Lanouette et al., 2014).

1.4 Μεθυλ-πρωτεωμική

Μέσα στα ζωντανά κύτταρα, ο βαθμός μεθυλίωσης Lys/Arg ρυθμίζεται από πολλούς παράγοντες, συμπεριλαμβανομένων των ενζυματικών δραστηριοτήτων και των τοπικών συγκεντρώσεων συγκεκριμένων PMTs, της προσιτότητας των απομεθυλασών για την απομάκρυνση των στοιχείων της μεθυλίωσης, καθώς και της σχετικής διάρκειας ζωής των υποστρωμάτων PMT έναντι των προϊόντων μεθυλίωσης τους. Η διατάραξη αυτών των παραγόντων αναμένεται να μεταβάλει άμεσα τα επίπεδα των σχετικών γεγονότων μεθυλίωσης. Ωστόσο, μια τέτοια διατάραξη μπορεί επίσης να χρησιμεύσει ως ένα σήμα upstream που επηρεάζει όχι μόνο τα επίπεδα μεθυλίωσης των άμεσων downstream στόχων αλλά και το σχετικό μεθύλωμα. Μεμονωμένα γεγονότα μεθυλίωσης μπορούν να ανιχνευθούν με αντισώματα αKme/Rme, φασματομετρία μάζας (Mass Spectrometry - MS) ή αυτοραδιογραφία με SAM ή S-[methyl-14C/3H]-επισημασμένο ραδιενεργό SAM ως συμπαράγοντες δότες μεθυλομάδας. Ωστόσο, είναι δύσκολο να διαμορφωθεί το προφίλ του μεθυλώματος κάτω από συγκεκριμένες κυτταρικές συνθήκες αποκλειστικά ανάλογα με τα αντισώματα για Western blot ή MS για τον ποσοτικό προσδιορισμό μεθυλιωμένων πεπτιδίων, δεδομένης της κακής ποιότητας πολλών αντισωμάτων pan-αKme/Rme και ο περιορισμός της MS στην ανάλυση μεγάλου αριθμού πεπτιδικών ιόντων που έχουν μειωμένη αφθονία. Η άμεση χρήση του ραδιενεργού S-[methyl-14C/3H]-SAM ως συμπαράγοντα δότη μεθυλομάδας εφαρμόζεται συχνά σε καλά καθορισμένες *in vitro* συνθήκες λόγω της κακής διαπερατότητας της μεμβράνης, του υψηλού κόστους και της πιθανής περιβαλλοντικής μόλυνσης αυτών των ραδιενεργών αντιδραστηρίων. Οι συμβατικές μέθοδοι που χρησιμοποιούν μόνο αντισώματα, MS ή αυτοραδιογραφία είναι επομένως πιο κατάλληλες για την ανίχνευση μεμονωμένων γεγονότων μεθυλίωσης παρά μεθυλώματος, εκτός εάν μπορούν να συνδυαστούν με προηγμένες στρατηγικές όπως αναφέρεται παρακάτω (Luo, 2015).

1.4.1 Δυσκολίες στη μελέτη του μεθυλώματος

Η μελέτη του μεθυλώματος απαιτεί αποτελεσματικές στρατηγικές τόσο στον εμπλουτισμό οσο και στη βέβαιότητα ταυτοποιήσης των τροποποιήσεων. Ωστόσο, αντιμετωπίζει δυσκολίες και στους δύο τομείς. Είναι πολύ δύσκολο να αναπτυχθεί μια υψηλά αποτελεσματική στρατηγική για τον εμπλουτισμό λόγω της αμελητέας μεταβολής στις φυσικοχημικές ιδιότητες της πρωτεΐνης ή του πεπτιδίου που προκαλείται από τη μεθυλίωση. Ωστόσο, η μεθυλίωση της Arg/Lys δεν εξουδετερώνει το θετικό τους φορτίο και δεν αλλάζει σημαντικά την υδροφοβικότητά τους. Επίσης δύσκολο είναι να αναπτυχθούν χρωματογραφικές μέθοδοι για τον εμπλουτισμό μεθυλιωμένων πεπτιδίων. Εξάλλου, η μεθυλομάδα είναι πολύ μικρή για να εντοπιστεί και να συνδεθεί από συγκεκριμένα αντισώματα. Προς το παρόν, το υψηλότερο ποσοστό ειδικότητας του εμπλουτισμού των μεθυλοπεπτιδίων είναι μόνο 11%, η οποία είναι πολύ μικρότερη από αυτή στον εμπλουτισμό φωσφοπεπτιδίων όπου πάνω από 90% θα μπορούσαν να δεσμευθούν εύκολα (Wang et al., 2017).

Η βεβαιότητα ταυτοποίησης των μεθυλ-πεπτιδίων είναι επίσης μεγάλη πρόκληση. Οι μετατοπίσεις μάζας που προκαλούνται από τη μεθυλίωση είναι ταυτόσημες με τις διαφορές μάζας μεταξύ πολλών αμινοξέων, επομένως, πολλά μη τροποποιημένα πεπτίδια μπορούν να αναγνωριστούν ως ψευδώς θετικά μεθυλιωμένα πεπτίδια. Αναφέρθηκε πρόσφατα ότι τα πραγματικά false discovery rates (FDR) ταυτοποιήσεων μεθυλ-πεπτιδίων ήταν συνήθως μεγαλύτερα από 80% ακόμη και όταν τα FDR εκτιμήθηκαν ότι ήταν <1% χρησιμοποιώντας τη συμβατική target-decoy προσέγγιση. Σαφώς ο έλεγχος εμπιστοσύνης δεν είναι ασήμαντο ζήτημα για ανάλυση μεθυλ-πρωτεώματος (Q. Wang et al., 2017).

1.4.2 Οι στρατηγικές για τον εμπλουτισμό του μεθυλ-πρωτεώματος

Ωστόσο, έχει επιτευχθεί αξιοσημείωτη πρόοδος στον τομέα της ανάλυσης μεθυλ-πρωτεώματος κατά την τελευταία δεκαετία. Διάφορες προσεγγίσεις εμπλουτισμού έχουν αναπτυχθεί. Από αυτά, ο εμπλουτισμός συγγένειας με χρήση αντισωμάτων είναι επί του παρόντος η κυρίαρχη προσέγγιση στην ανάλυση μεθυλ-πρωτεώματος. Μια προσέγγιση που βασίζεται στη χρωματογραφία είναι άλλη μια στρατηγική για τον εμπλουτισμό των μεθυλ-πεπτιδίων. Συχνά χρησιμοποιούνται οι στρατηγικές Strong cation exchange chromatography (SCX), isoelectric focusing (IEF) και hydrophilic interaction liquid chromatography (HILIC). Αξίζει να σημειωθεί ότι η συνδυαστική χρήση τεχνικών εμπλουτισμού συγγένειας και αυτών που βασίζονται στη χρωματογραφία θα μπορούσε να οδηγήσει σε υψηλότερη αποτελεσματικότητα αναγνώρισης. Η heavy-methyl SILAC που αποτελεί μια παραλλαγή της σταθερής ισοτοπικής σήμανσης με αμινοξέα σε κυτταρική καλλιέργεια (hM-SILAC) έχει γίνει η πιο ευρέως διαδεδομένη στρατηγική επικύρωσης μεθυλπεπτιδίου (Wang et al., 2017).

1.4.2.1 Εμπλουτισμός με χρήση αντισωμάτων

Εμπλουτισμός με βάση αντισώματα, είναι μια δημοφιλής προσέγγιση εμπλουτισμού για την ανάλυση της μεθυλιωμένης πρωτεΐνης. Ο εμπλουτισμός των μεθυλιωμένων πεπτιδίων αργινίνης μπορεί να επιτευχθεί με τη χρήση εξειδικευμένων αντισωμάτων, ενώ δεν είναι διαθέσιμο μέχρι στιγμής ένα υψηλής απόδοσης pan-specific αντίσωμα για τη μεθυλίωση της λυσίνης (Wang et al., 2017).

Ο ειδικός εμπλουτισμός των μεθυλπεπτιδίων είναι επίσης πολύ δύσκολος, επειδή η μεταβολή των φυσικοχημικών ιδιοτήτων τους που προκαλείται από τη μεθυλίωση είναι πολύ μικρή. Ωστόσο, η ανάλυση της μεθυλίωσης της λυσίνης δεν είναι τόσο επιτυχής λόγω της έλλειψης αντισωμάτων pan-K-methyl υψηλής απόδοσης. Για να αποκαλυφθεί η πλήρης κατάσταση του μεθυλ-πρωτεώματος, απαιτούνται ολοκληρωμένες αναλύσεις όλων των μορφών μεθυλίωσης στα αμινοξέα Arg/Lys, κάτι που απαιτεί τη χρήση πολλαπλών αντισωμάτων. Συνολικά, αυτή η μέθοδος κοστίζει και η αναπαραγωγιμότητα της επηρεάζεται από τα διαθέσιμα αντισώματα (Han et al., 2019).

1.4.2.2 Εμπλουτισμός με χρήση περιοχών αναγνώρισης μεθυλομάδας

Ορισμένα PTM μπορούν επίσης να αναγνωριστούν από συγκεκριμένα domains πρωτεϊνών. Τέτοια domains αναγνώρισης PTM είναι πιθανά αντιδραστήρια συγγένειας για τον εμπλουτισμό πρωτεϊνών/πεπτιδίων που φέρουν PTM. Από πολλά domains συμπεριλαμβανομένου του CD (chromodomain) του Tudor domain(s) και του MBT (malignant brain tumor domain) αναγνωρίζονται οι μεθυλιωμένες λυσίνες. Οι Liu et al. παρουσίασαν μια νέα στρατηγική συνδυάζοντας καθαρισμό συγγένειας που βασίζεται στην αναγνώριση περιοχής μεθυλυσίνης (MRD) με arrays πεπτιδίων, βιοπληροφορική και φασματομετρία μάζας για τον εντοπισμό θέσεων μεθυλίωσης. Ωστόσο, με την παραπάνω στρατηγική εντοπίστηκε μόνο ένας περιορισμένος αριθμός θέσεων μεθυλίωσης. Αυτό είναι ένα εγγενές μειονέκτημα του εμπλουτισμού σε επίπεδο πρωτεΐνης καθώς ένας μεγάλος αριθμός μη τροποποιημένων πεπτιδίων παρεμβαίνει σοβαρά στην ανίχνευση μεθυλιωμένων πεπτιδίων. Επιπλέον, αυτές οι περιοχές σύνδεσης απαιτούν την παρουσία πλευρικών καταλοίπων για πιο αποτελεσματική δέσμευση, επομένως ο εμπλουτισμός μεθυλιωμένων πεπτιδίων από μια πρωτεΐνη που έχει υποστεί πέψη είναι δύσκολος. Η ισχύς δέσμευσης μεταξύ ενός φυσικού domain αναγνώρισης PTM και ενός τροποποιημένου καταλοίπου είναι συνήθως ασθενής για να εμπλουτίσει τροποποιημένα πεπτίδια (Wang et al., 2017).

1.4.2.3 Μέθοδοι εμπλουτισμού με βάση τη χρωματογραφία

Η μεθυλίωση συμβαίνει συχνά σε μη δομικές περιοχές, και αυτές οι περιοχές είναι πάντα εκτεθειμένες σε νερό, επομένως τα μεθυλιωμένα πεπτίδια είναι συνήθως υψηλής υδροφιλικότητας. Επιπλέον, σε σύγκριση με τα μη τροποποιημένα πεπτίδια, τα μεθυλιωμένα πεπτίδια που έχουν υποστεί πέψη με τρυψίνη συνήθως φέρουν ≥3 θετικά φορτία. Το 2012, οι Uhlmann et al. διερεύνησαν την απόδοση τριών χρωματογραφικών μεθόδων, δηλαδή των HILIC, SCX και IEF, για τον εμπλουτισμό μεθυλιωμένων πεπτιδίων. Βρέθηκε ότι το HILIC έδωσε τα καλύτερα αποτελέσματα. Η κακή απόδοση του SCX οφείλεται κυρίως στην παρεμβολή από τα πεπτίδια που περιέχουν ιστιδίνη και τα μη τροποποιημένα πεπτίδια.

Θεωρητικά, το IEF μπορεί επίσης να χρησιμοποιηθεί για τον διαχωρισμό των μεθυλιωμένων πεπτιδίων λόγω των υψηλών ισοηλεκτρικών σημείων τους εξαιτίας της παρουσίας επιπλέον φορτίων. Η σπάνια παρουσία όξινων καταλοίπων που πλαισιώνουν τα κατάλοιπα μεθυλιωμένης αργινίνης, προκαλούν τα ισοηλεκτρικά σημεία αυτών των πεπτιδίων να είναι συνήθως υψηλότερα από 9. Λόγω της υψηλής ανάλυσης και της εύκολης πρόσβασης στο εργαστήριο, η μέθοδος που βασίζεται στο SCX είναι μια πολλά υποσχόμενη προσέγγιση για τον εμπλουτισμό μεθυλιωμένων πεπτιδίων.

Ωστόσο, η προσέγγισή αυτή δεν μπορεί να εξαλείψει την παρεμβολή από πεπτίδια που περιέχουν ιστιδίνη. Για να μειωθεί η παρεμβολή από αυτά τα πεπτίδια, δημιουργήθηκε μια στρατηγική διαχωρισμού SCX υψηλού pH (Q. Wang et al., 2017). Ο συμβατικός διαχωρισμός SCX συνήθως διεξάγεται σε χαμηλό pH, και σε αυτές τις συνθήκες, τα κατάλοιπα ιστιδίνης θα φέρουν επίσης θετικά φορτία. Ωστόσο, τα κατάλοιπα ιστιδίνης θα ήταν ως επί το πλείστον αφόρτιστα όταν το pH του διαλύματος είναι υψηλότερο από 9. Επομένως, ο διαχωρισμός SCX που πραγματοποιείται υπό αυτήν την κατάσταση μπορεί να μετριάσει την παρεμβολή από πεπτίδια που περιέχουν ιστιδίνη (Wang et al., 2017). Έτσι, ο διαχωρισμός SCX με υψηλό pH είναι ένα ισχυρό εργαλείο για την επίτευξη εμπλουτισμού μεθυλιωμένων πεπτιδίων χωρίς αντισώματα. Αν και η προσέγγιση εμπλουτισμού με βάση τη χρωματογραφία είναι κατώτερη από αυτή του εμπλουτισμού συγγένειας όσον αφορά την εξειδίκευση, το χαμηλότερο κόστος και η καλύτερη αναπαραγωγιμότητα της την καθιστούν μια ισχυρή προσέγγιση στην ολοκληρωμένη ανάλυση του μεθυλ-προτεώματος (Wang et al., 2017).

Εκτός από τον επιλεκτικό εμπλουτισμό μεθυλιωμένων πεπτιδίων, ο εξαντλητικός διαχωρισμός είναι μια άλλη στρατηγική για τη μείωση της πολυπλοκότητας του δείγματος και τη βελτίωση της ευαισθησίας ανίχνευσης των μεθυλιωμένων πεπτιδίων. Μια πολυδιάστατη μέθοδος διαχωρισμού, είναι το SCX-RPLC (Fisk et al., 2013). To Reversed phase liquid chromatography (RPLC) διαχωρίζει μόρια με βάση την υδροφοβικότητα της επιφάνειας και είναι η πιο συχνά χρησιμοποιούμενη και ευρέως εφαρμόσιμη τεχνική LC. Το RPLC χρησιμοποιείται συνήθως ως η τελευταία φάση του διαχωρισμού (Boone & Adamec, 2016).

Άλλες μέθοδοι διαχωρισμού είναι το CID καθώς και το ETD. To Collision-induced dissociation (CID), είναι μια τεχνική φασματομετρίας μάζας για την πρόκληση κατακερματισμού επιλεγμένων ιόντων στην αέρια φάση. Τα επιλεγμένα ιόντα επιταχύνονται εφαρμόζοντας ένα ηλεκτρικό δυναμικό για την αύξηση της κινητικής ενέργειας των ιόντων και στη συνέχεια συγκρούονται με ουδέτερα μόρια. Κατά τη σύγκρουση μέρος της κινητικής ενέργειας μετατρέπεται σε εσωτερική ενέργεια που έχει ως αποτέλεσμα τη θραύση του δεσμού και τον κατακερματισμό του μοριακού ιόντος σε μικρότερα θραύσματα. Αυτά τα ιόντα θραυσμάτων μπορούν στη συνέχεια να αναλυθούν με διαδοχική φασματομετρία μάζας (MS/MS) (Wells & McLuckey, 2005). Από την άλλη, το Electron-transfer dissociation (ETD) είναι μια μέθοδος κατακερματισμού πολλαπλών φορτισμένων αερίων μακρομορίων σε ένα φασματόμετρο μάζας μεταξύ των σταδίων της MS/MS. Η ETD επάγει τον κατακερματισμό μεγάλων, πολλαπλά φορτισμένων κατιόντων μεταφέροντας ηλεκτρόνια σε αυτά (Hart-Smith, 2014). Σύμφωνα με μια μελέτη του 2009 όπου συγκρίθηκαν η CID και η ETD φάνηκε ότι η δεύτερη είναι πιο αποτελεσματική στην ταυτοποίηση πεπτιδίων με μεθυλιωμένη αργινίνη, επειδή αυτά τα πεπτίδια συνήθως φέρουν περισσότερα θετικά φορτία και έχουν πολύ καλύτερο κατακερματισμό κατά τη λειτουργία ETD (H. Wang et al., 2009).

1.4.3 Στρατηγικές για αποτελεσματικότερη ταυτοποίηση των θέσεων μεθυλίωσης

Στην πρωτεομική ανάλυση της μεθυλίωσης των πρωτεϊνών, μπορεί να εμφανιστούν ψευδώς θετικές ταυτοποιήσεις λόγω διαφόρων παραγόντων όπως αναφέρθηκαν προηγουμένως. Έτσι, ο έλεγχος της ταυτοποίησης είναι σημαντικός στην μελέτη του μεθυλ-πρωτεώματος. Κυρίως, τρεις προσεγγίσεις χρησιμοποιούνται συνήθως για τον έλεγχο της βεβαιότητας ταυτοποίησης και παρουσιάζονται παρακάτω. Εκτός από τις πειραματικές προσεγγίσεις, εργαλεία βιοπληροφορικής έχουν επίσης αναπτυχθεί για την πρόβλεψη των θέσεων μεθυλίωσης (Q. Wang et al., 2017).

1.4.3.1 Προσέγγιση με αναζήτηση βάσεων δεδομένων target-decoy

Η προσέγγιση αυτή είναι ένα golden standard control για τον έλεγχο της βεβαιότητας των πεπτιδίων που έχουν ταυτοποιηθεί στην πρωτεϊνική ανάλυση. Αυτή η προσέγγιση εφαρμόστηκε επίσης για τον έλεγχο της εμπιστοσύνης των ταυτοποιήσεων μεθυλ-πεπτιδίων σε πολλές μελέτες μεθυλ-πρωτεώματος, συμπεριλαμβανομένων μερικών μελετών μεγάλης κλίμακας. Ωστόσο, οι Hart-Smith et al. βρήκαν ότι τα μεθυλπεπτίδια που προσδιορίστηκαν από αυτή την προσέγγιση δεν ήταν υψηλής ποιότητας. Οι πραγματικές FDR των spectra matches για τα μεθυλιωμένα πεπτίδια θα μπορούσαν να φτάσουν το 80% ακόμη και όταν οι FDR εκτιμήθηκαν ότι είναι <1% χρησιμοποιώντας αυτή τη μέθοδο. Ξεχωριστή αξιολόγηση FDR για μια υποομάδα τροποποιημένων πεπτιδίων έχει προταθεί για τη βελτίωση της βεβαιότητας ταυτοποίησης (Marx et al., 2013). Οι Hart-Smith et al. ερεύνησαν επίσης αυτή τη στρατηγική για ταυτοποιήσεις μεθυλπεπτιδίων και διαπίστωσαν ότι ούτε αυτή ήταν αξιόπιστη. Θεώρησαν ότι αυτό ήταν πιθανό να οφείλεται στον υψηλό αριθμό συνδυασμών αμινοξέων ικανών να παράγουν πεπτιδικές αλληλουχίες που είναι ισοβαρείς προς μεθυλιωμένα πεπτίδια διαφορετικής αλληλουχίας. Με άλλα λόγια, ένα ψευδώς θετικό ταυτοποιημένο μεθυλοπεπτίδιο με υψηλό score αντιστοίχισης θα μπορούσε να είναι ένα μη τροποποιημένο πεπτίδιο με την ίδια μάζα αλλά με ελαφρώς διαφορετική αλληλουχία, δηλαδή ένα πεπτίδιο που περιέχει ένα κατάλοιπο γλουταμίνης και ένα κατάλοιπο λυσίνης θα μπορούσε να αναγνωριστεί λανθασμένα ως ένα πεπτίδιο που περιέχει κατάλοιπο ασπαραγίνης και μονομεθυλιωμένη λυσίνη. Αυτό επιβεβαιώθηκε από το γεγονός ότι ένα σημαντικό μέρος ψευδώς θετικών ταυτοποιημένων μεθυλοπεπτιδίων είχε σχετικά υψηλά scores αντιστοίχισης και δεν μπορούσαν να αφαιρεθούν αποτελεσματικά με τη χρήση της στρατηγικής target-decoy ακόμη και όταν εφαρμόστηκε ξεχωριστή αξιολόγηση FDR (Hart-Smith et al., 2016). Επιπλέον, οι τεχνητές ή *ex vivo* τροποποιήσεις και οι εσφαλμένες θέσεις μεθυλίωσης μπορεί επίσης να συμβάλλουν στο υψηλό FDR. Καθώς οι περισσότερες μελέτες επικεντρώθηκαν στη διερεύνηση της μεθυλίωσης σε κατάλοιπα Arg και Lys, η μεθυλίωση που συμβαίνει σε άλλα κατάλοιπα μπορεί να αποδοθεί λανθασμένα στα κατάλοιπα Arg/Lys όταν οι μεταβλητές τροποποιήσεις σε αυτά δεν ορίστηκαν κατά την έρευνα της βάσης δεδομένων.

Ωστόσο, λαμβάνοντας υπόψη τη χαμηλή συχνότητα των μη μεθυλιωμένων Arg/Lys σε οργανισμούς, ο λανθασμένος εντοπισμός τέτοιων θέσεων μπορεί να μην αποτελεί μείζονα ανησυχία. Αντίθετα, ο εσφαλμένος εντοπισμός μεθυλίωσης μεταξύ διαφορετικών καταλοίπων λυσίνης ή αργινίνης μπορεί να είναι πιο συνηθισμένος. Έλεγχος με το μάτι μπορεί να είναι χρήσιμος για τη μείωση της πιθανότητας ψευδούς θετικής ταυτοποίησης μόνο όταν λαμβάνονται υπόψη τα μοναδικά χαρακτηριστικά των φασμάτων για τα μεθυλιωμένα πεπτίδια. Για παράδειγμα, η παρουσία μοναδικών θραυσμάτων ιόντων που σχετίζονται με την μεθυλομάδα, η οποία τεχνική συζητείται στην επόμενη ενότητα (Q. Wang et al., 2017).

1.4.3.2 Διάκριση μορφών μεθυλίωσης με χαρακτηριστικά ιόντα

Διαφορετικοί τύποι PTM συχνά δημιουργούν χαρακτηριστικά θραύσματα ιόντων σε διαδοχικά MS. Με βάση αυτό το γεγονός, η λειτουργία του MS για τη σάρωση πρόδρομων ιόντων χρησιμοποιήθηκε για την επιλεκτική ταυτοποίηση συγκεκριμένων τροποποιημένων πεπτιδίων (Petersson et al., 2001). Για τη μεθυλίωση, τα χαρακτηριστικά θραύσματα ιόντων χρησιμοποιούνταν συχνά για τη διάκριση της ασύμμετρης διμεθυλίωσης (ADMA) και της συμμετρικής διμεθυλίωσης (SDMA). Αν και δεν υπάρχει διαφορά μάζας μεταξύ των μεθυλοπεπτιδίων αυτών των δύο μορφών μεθυλίωσης, τα χαρακτηριστικά θραύσματα ιόντων είναι διαφορετικά. Το peak στα m/z 71,06 (ιόν διμεθυλ-καρβοδιιμιδίου) μπορεί να χρησιμοποιηθεί για την ταυτοποίηση διμεθυλιών (DMAs), αλλά το peak στα m/z 46,06 (ιόν διμεθυλ-αμμωνίου) εμφανίζεται μόνο στα φάσματα των ADMA υπό συνθήκες CID (Rappsilber et al., 2003). Εκτός από τη μεθυλίωση της αργινίνης, υπάρχουν επίσης μερικά χαρακτηριστικά ιόντα μεθυλίωσης της λυσίνης. Τόσο το Kme1 όσο και το Kme2 μπορούν να παράγουν ένα χαρακτηριστικό ιόν m/z 98. Ωστόσο, το peak στα m/z 112 εμφανίζεται μόνο στα spectra Kme2 (Couttas et al., 2008). Τα χαρακτηριστικά ιόντα πρέπει να είναι μοναδικά μεταξύ όλων των πεπτιδίων του πρωτεώματος, ώστε να επιτρέπεται η χρήση τους για την αξιολόγηση της ταυτοποίησης του μεθυλ-πεπτιδίου σε επίπεδο πρωτεώματος. Δυστυχώς, αυτό δεν ισχύει, τουλάχιστον, για ορισμένα ιόντα. Για παράδειγμα, ιόντα στα m/z 71 και 46, τα οποία συνήθως χρησιμοποιούνται για την ταυτοποίηση DMA και ADMA, αντίστοιχα, υπάρχουν επίσης στα φάσματα πεπτιδίου που περιέχει αργινίνη, προλίνη ή Kme2. Το χαρακτηριστικό θραύσμα-ιόν του MMA, το μονομεθυλ-γουανιδίνιο έχει παρόμοιο m/z με ένα βαρύ ισότοπο του ιόντος αμμωνίου της θρεονίνης. Εκτός από τη μη εξειδίκευση των χαρακτηριστικών ιόντων, υπάρχουν δύο πιθανά ζητήματα. Πρώτον, η αλληλουχία των πεπτιδίων επηρεάζει την ένταση του σήματος των χαρακτηριστικών ιόντων (Gehrig et al., 2004). Δεύτερον, η ανίχνευση χαρακτηριστικών ιόντων χαμηλής μάζας μπορεί να είναι δύσκολη για ορισμένα συστήματα MS λόγω των υψηλών τιμών που χρησιμοποιούνται ως κατώφλι (Tian et al., 2016).

1.4.3.3 Μέθοδος που βασίζεται στη μεταβολική σήμανση

Σε μια μελέτη μεθυλίωσης, οι ψευδώς θετικές ταυτοποιήσεις δεν μπορούν να φιλτραριστούν αποτελεσματικά από την παραδοσιακή target-decoy προσέγγιση. Για να βελτιωθεί η βεβαιότητα των ταυτοποιήσεων μεθυλίωσης, αναπτύχθηκε η μέθοδος hM-SILAC (heavy methyl stable isotope labeling by amino acids in cell culture) (Ong et al., 2004). Σε αυτή την προσέγγιση τα κύτταρα αναπτύσσονται σε ένα μέσο που περιέχει μόνο ‘heavy’ methionine, [13CD3]-Met, το οποίο μετατρέπεται μεταβολικά σε ‘heavy’ S-adenosyl methionine (AdoMet), με αποτέλεσμα την ενσωμάτωση σημασμένων heavy methyl groups (13CD3) σε όλες τις θέσεις μεθυλίωσης. Αυτό θα προκαλέσει μετατοπίσεις μάζας n × 18 Da. Στην πράξη, δημιουργείται επίσης ένα δείγμα ελέγχου που παρασκευάζεται από αναπτυσσόμενα κύτταρα σε μέσο που περιέχει ‘light’ methionine για να επιτρέπεται η εμφάνιση ενός ζεύγους φασμάτων (Ong et al., 2004). Η συνολική διαδικασία έχει ως εξής. Δύο δείγματα κυττάρων καλλιεργούνται σε μέσο που περιέχει [13CD3]-μεθειονίνη και μέσο που περιέχει [12CH3]-μεθειονίνη, αντίστοιχα. Όταν η αποτελεσματικότητα σήμανσης ενός σταθερού αμινοξέος που περιέχει ισότοπο ήταν αρκετά υψηλή (>97% μετά από πέντε διπλασιασμούς κυττάρων), οι πρωτεΐνες από δύο δείγματα κυττάρων εκχυλίζονται και αναμιγνύονται σε αναλογία μάζας 1 : 1 (Ong et al., 2004; Ong & Mann, 2006). Μετά τον εμπλουτισμό των μεθυλ-πρωτεϊνών, τη πέψη με τρυψίνη και την ανάλυση LC-MS/MS, χρησιμοποιήθηκαν δύο αυστηρά κριτήρια αξιολόγησης. Πρώτον, τα « light» και τα «heavy» μεθυλπεπτίδια θα πρέπει να ταυτοποιούνται ταυτόχρονα δηλαδή στο ίδιο πείραμα. Δεύτερον, η αναλογία έντασης των κορυφών των «light» και «heavy» πεπτιδίων θα πρέπει να είναι σχεδόν 1:1 στο χρωματογράφημα εκχυλισμένων ιόντων (XIC). Σε σύγκριση με την παραδοσιακή προσέγγιση target-decoy, το hM-SILAC καθιστά την ταυτοποίηση μεθυλίωσης πιο αξιόπιστη (Q. Wang et al., 2017).

1.5 Μηχανική μάθηση

Η Μηχανική Μάθηση (ΜΜ) είναι ένα πεδίο της Τεχνητής Νοημοσύνης και έχει ως στόχο τη δημιουργία αλγορίθμων με ικανότητα μάθησης. Πιο συγκεκριμένα, χρησιμοποιώντας ένα σύνολο δεδομένων, κατασκευάζονται μοντέλα. Για τη δημιουργία ενός μοντέλου κατασκευάζονται αλγόριθμοι, οι οποίοι βασίζονται σε ένα σύνολο εκπαίδευσης, προκειμένου τα δεδομένα να ταξινομηθούν σε κατηγορίες ή να βρεθούν τυχόν τυποποιημένες μορφές σε αυτά ή να γίνει πρόβλεψη για κάποια συμπεριφορά αυτών (Likas et al., 2014).

Έχουν αναπτυχθεί τρία είδη μάθησης. Το πρώτο είδος ονομάζεται Επιβλεπόμενη Μάθηση (Supervised learning) όπου ο αλγόριθμος εκπαιδεύεται λαμβάνοντας ως εισόδους ένα σύνολο δεδομένων εκπαίδευσης με γνωστές εξόδους. Στόχος της είναι η γενίκευση της συνάρτησης προκειμένου να λαμβάνει και να απεικονίζει στιγμιότυπα στα οποία η έξοδος είναι άγνωστη. Εφαρμόζεται σε προβλήματα ταξινόμησης (classification) και πρόβλεψης (prediction). Το δεύτερο είδος μάθησης είναι η Μη επιβλεπόμενη Μάθηση (Unsupervised Learning) όπου ο αλγόριθμος προσπαθεί να ανακαλύψει τυχόν συσχετίσεις μεταξύ των δεδομένων εισόδου με άγνωστη έξοδο προκειμένου να βρεθεί ένας τρόπος συσχέτισης ή ομαδοποίησης αυτών. Τέλος, στη Ενισχυτική Μάθηση (Reinforcement Learning) ο αλγόριθμος προσπαθεί να μάθει μια στρατηγική ενεργειών μέσα από την παρακολούθηση του τρόπου λειτουργίας ενός δυναμικού περιβάλλοντος δεδομένων (Witten et al., 2017).

1.5.1 Υπολογιστικά εργαλεία πρόβλεψης θέσεων μεθυλίωσης των πρωτεϊνών

Η ακριβής πρόβλεψη των θέσεων μεθυλίωσης πρωτεΐνης είναι σημαντική για την ανακάλυψη των μοριακών μηχανισμών στους οποίους συμμετέχει η μεθυλίωση. Όπως αναφέρθηκε παραπάνω, οι πειραματικές τεχνικές εμφανίζουν δυσκολίες στον εντοπισμό πιθανών θέσεων μεθυλίωσης ενώ παράλληλα είναι αρκετά δαπανηρές. Οι υπολογιστικές προβλέψεις των θέσεων μεθυλίωσης έχουν βοηθήσει στην αντιμετώπιση αυτών των περιορισμών καθώς επιτρέπουν την μείωση του αριθμού των πειραμάτων που απαιτούνται για τον προσδιορισμό των θέσεων μεθυλίωσης της πρωτεΐνης (Audagnotto & Dal Peraro, 2017). Τα τελευταία χρόνια, η υπολογιστική πρόβλεψη που βασίζεται σε αλγόριθμους μηχανικής μάθησης έχει αναδειχθεί ως μια ισχυρή προσέγγιση για τον εντοπισμό θέσεων μεθυλίωσης και έχει σημειωθεί μεγάλη πρόοδος στη βελτίωση της προγνωστικής απόδοσης (L. Wei et al., n.d) .

Υπάρχουν αρκετά προγράμματα στο διαδίκτυο τα οποία είναι διαθέσιμα για την πρόβλεψη θέσεων μεθυλίωσης. Η δημιουργία τέτοιων προγραμμάτων έχει ξεκινήσει εδώ και πολλά χρόνια και συνεχώς εξελίσσεται. Το MeMo είναι ένα από τα πρώτα εργαλεία που έγιναν διαθέσιμα. Χρησιμοποιεί το support vector machine (SVM) ως αλγόριθμο πρόβλεψης. Το σύνολο δεδομένων του προέρχονται από προσεκτική επιλογή των μεθυλιωμένων καταλοίπων που υπάρχουν στο SWISS-PROT, το οποίο αποτελείται από 264 μεμονωμένα πιστοποιημένες μεθυλ-λυσίνες και 107 μεθυλ-αργινίνες που προέκυψαν από την μελέτη περίπου 1700 επιστημονικών άρθρων. Το MeMo φαίνεται να είναι ένα ισχυρό εργαλείο για την πρόβλεψη θέσεων μεθυλιωμένης αργινίνης σε σύγκριση με τις θέσεις μεθυλιωμένης λυσίνης. Ωστόσο, η ακρίβειά του επηρεάζεται από την έλλειψη διαθέσιμων δεδομένων εκπαίδευσης τη στιγμή που δημιουργήθηκε (Chen et al., 2006). Στη συνέχεια, η αξιοπιστία της πρόβλεψης βελτιώθηκε από το BPB-PPMS, όπου χρησιμοποίησε μια Bi-Profile Bayesian στρατηγική για τον διαχωρισμό μεθυλιωμένων και μη μεθυλιωμένων θέσεων με βάση γνωστά πειραματικά δεδομένα (Shao et al., 2009). Το σύνολο δεδομένων αυξήθηκε σε 363 μεθυλιωμένες αργινίνες και 977 μεθυλιωμένες λυσίνες. Ο συνδυασμός των χαρακτηριστικών Bayesian Bi-Profile με ένα μεγαλύτερο σύνολο δεδομένων βελτίωσε την ακρίβεια πρόβλεψης μεθυλίωσης έως και 92% για τις πρωτεΐνες με μεθυλιωμένες λυσίνες και 88% για τις πρωτεΐνες με μεθυλιωμένες αργινίνες (Shao et al., 2009). Στη συνέχεια το MASA, ένα νέο πρόγραμμα, χρησιμοποίησε το Solvent Accessible Surface Are (SASA) όπου στηρίχθηκε στη δευτεροταγή δομή της πρωτεΐνης για την πρόβλεψη μεθυλιωμένων θέσεων (Shien et al., 2009). Το συγκεκριμένο πρόγραμμα επιτρέπει την πρόβλεψη όχι μόνο μεθυλιωμένων λυσινών και μεθυλιωμένων αργινινών, αλλά και μεθυλ-γλουταμινικών.

Ωστόσο, οι περισσότερες από αυτές τις μεθόδους χρησιμοποιούν μόνο πληροφορίες της πρωτεϊνικής αλληλουχίας χωρίς να λαμβάνεται υπόψη οι φυσικοχημικές ιδιότητες των αμινοξέων. Προκειμένου να βελτιωθεί η αποτελεσματικότητα της πρόβλεψης, δημιουργήθηκε μια νέα προσέγγιση που ονομάζεται PMes, η οποία εξετάζει τις φυσικοχημικές ιδιότητες των αμινοξέων που περιβάλλουν τις θέσεις μεθυλίωσης (Shi et al., 2012). Προτάθηκε επίσης ένα ειδικό εργαλείο πρόβλεψης μεθυλίωσης της λυσίνης για ιστονικές πρωτεΐνες, το METhK, το οποίο χρησιμοποιεί σαν χαρακτηριστικά την σύνθεση αμινοξέων, το SASA, την σύνθεση ζεύγους αμινοξέων κ.α. (T.-Y. Lee et al., 2014). Επίσης, ένα άλλο εργαλείο που εισήχθη για τον *in vivo* ή *in vitro* προσδιορισμό θέσεων μεθυλίωσης για συγκεκριμένα είδη είναι το PSSMe (Wen et al., 2016). Το συγκεκριμένο πρόγραμμα δοκιμάστηκε σε μεγάλης κλίμακας πειραματικό σύνολο δεδομένων με θέσεις μεθυλίωσης που προέρχονται από διαφορετικά είδη, αποκαλύπτοντας ότι τα πρότυπα μεθυλίωσης είναι πράγματι εξαρτώμενα από το είδος (Audagnotto & Dal Peraro, 2017).

Ωστόσο, η προγνωστική απόδοση των παραπάνω μεθόδων δεν είναι ικανοποιητική όσον αφορά τη συνολική ακρίβεια. Τα πλέον διαδεδομένα και πιο αποτελεσματικά είναι το MePred-RF, το PRmePRed και το GPS-MSP. Το MePred-RF βασίζεται στον αλγόριθμο Random forest, με ενσωμάτωση πολλών διακριτικών περιγραφικών χαρακτηριστικών που βασίζονται στην ακολουθία και βελτίωση της ικανότητας επιλογής χαρακτηριστικών χρησιμοποιώντας μια ισχυρή τεχνική επιλογής. Συγκριτικές μελέτες σε σύνολα δεδομένων αναφοράς δείχνουν ότι η προτεινόμενη μέθοδος MePred-RF ξεπερνά τα άλλα διαθέσιμα προγνωστικά εργαλεία κατά μέσο όρο 4,5 τοις εκατό όσον αφορά τη συνολική ακρίβεια (L. Wei et al., n.d.). Το PRmePRed είναι ένα εργαλείο πρόβλεψης βασισμένο στο SVM για την πρόβλεψη θέσεων μεθυλίωσης αργινίνης στις πρωτεΐνες. Για την πρόβλεψη των θέσεων μεθυλίωσης χρησιμοποιούνται η αλληλουχία και η δομή των πρωτεϊνών. Το PRmePRed αποδίδει αρκετά καλά με ακρίβεια 84,10%, ευαισθησία 82,38%, ειδικότητα 83,77% και συντελεστή συσχέτισης Matthew 0,7 σε 10-fold cross validation. Τέλος, το GPS-MSP (Methyl-group Specific Predictor) χρησιμοποίησε τον αλγόριθμο GPS 3.0 για την πρόβλεψη γενικών ή τυποειδικών καταλοίπων μεθυλ-αργινίνης σε πρωτεΐνες. Το πρόγραμμα αυτό διαθέτει πάνω από 28 predictors. Ένα από αυτά μπορεί να προβλέψει τύπους μονο-, δι- και τρι-μεθυλίωσης για συγκεκριμένες λυσίνες και μονο-, συμμετρικούς, δι- συμμετρικούς και ασύμμετρους τύπους διμεθυλίωσης για συγκεκριμένες αργινίνες.

1.6 ΣΚΟΠΟΣ

Στόχος αυτής της εργασίας ήταν η ανάπτυξη ενός νέου και πιο αποτελεσματικού υπολογιστικού εργαλείου πρόβλεψης θέσεων μεθυλίωσης σε λυσίνες και αργινίνες, χρησιμοποιώντας περισσότερα και πιο πρόσφατα δημοσιευμένα δεδομένα μεθυλοπρωτεωμικής. Παράλληλα, διερευνώνται οι παράγοντες που επηρεάζουν την απόδοση πρόβλεψης, οι οποίοι αφορούν τη μορφή των δεδομένων αλλά και την ποιότητα τους. Αρχικά, τα δεδομένα που αφορούν πεπτίδια με μεθυλιωμένες θέσεις, συγκεντρώθηκαν, φιλτραρίστηκαν και χρησιμοποιήθηκαν για την εκπαίδευση αλγορίθμων κατηγοριοποίησης. Για την εκπαίδευση των αλγορίθμων ως χαρακτηριστικά των παραδειγμάτων στη πρώτη ανάλυση χρησιμοποιήθηκε η πρωτοταγή δομή των πεπτιδίων, ενώ στη δεύτερη ανάλυση οι φυσικοχημικές ιδιότητες των αμινοξέων των πεπτιδίων. Τέλος, για κάθε ένα από τα σύνολα δεδομένων που δημιουργήθηκαν επιλέχθηκαν οι αλγόριθμοι που εμφάνισαν την υψηλότερη απόδοση.

2. ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

2.1 Λήψη δεδομένων μεθυλ-πρωτεωμικής

Για την πραγματοποίηση της ανάλυσης ήταν απαραίτητη η αναζήτηση και λήψη δεδομένων μεθυλ-πρωτεωμικής του ανθρώπου, πιο συγκεκριμένα για πεπτίδια στα οποία έχουν ταυτοποιηθεί μεθυλιωμένες αργινίνες και λύσινες. Έτσι, έγινε έρευνα σχετικών εργασιών σε ένα μεγάλο αριθμό περιοδικών, κυρίως στο Pubmed. Οι εργασίες που επιλέχθηκαν πληρούσαν ορισμένα κριτήρια. Θα έπρεπε οι ερευνητές να έχουν ακολουθήσει μια αποτελεσματική στρατηγική για τον εμπλουτισμό του μεθυλ-πρωτεώματος, όπως αυτές που αναφέρθηκαν παραπάνω, και να έχει γίνει έλεγχος της identification confidence με αξιόπιστες μεθόδους όπως η heavy-methyl-SILAC. Ένα ακόμη σημαντικό κριτήριο για την επιλογή των εργασιών που θα επιλέγονταν ήταν η μορφή των αρχείων όπου οι ερευνητές έχουν αποθηκεύσει τα δεδομένα μεθυλ-πρωτεωμικής. Τα αρχεία έπρεπε να περιέχουν δεδομένα της αλληλουχίας των πεπτιδίων, της θέσης των μεθυλιωμένων αργινίνων (R) και μεθυλιωμένων λυσίνων (K) στο πεπτίδιο, καθώς και συγκεκριμένα scores που αφορούν την ορθή ταυτοποίηση του κάθε τροποποιημένου πεπτιδίου αλλά και της θέσης μεθυλίωσης.

2.2 Φιλτράρισμα των δεδομένων που ανακτήθηκαν

Η χρήση των scores είναι ύψιστης σημασίας για αυστηρό φιλτράρισμα των δεδομένων, ώστε να αποφευχθεί η λήψη ψευδώς-θετικών θέσεων μεθυλίωσης. Πιο ειδικά, μετά την ολοκλήρωση της ανάλυσης της Φασματομετρίας μάζας προκύπτουν από αυτή τα raw files που αναλύονται με την βοήθεια υπολογιστικών εργαλείων βιοπληροφορικής. Στις εργασίες που επιλέχθηκαν οι ερευνητές χρησιμοποίησαν κυρίως το MaxQuant, αλλά και το Sequest HT, τα οποία διαθέτουν διάφορα scores που αφορούν τη αξιόπιστη ταυτοποίηση των μεθυλιωμένων πεπτιδίων αλλά και των μεθυλιωμένων θέσεων. Κάθε εργαλείο διαθέτει τα δικά του scores. Επομένως, για κάθε εργασία επιλέχτηκαν διαφορετικά scores, σύμφωνα με το software που χρησιμοποιήθηκε, προσφέροντας όσο το δυνατόν αυστηρότερο φιλτράρισμα. Στο πίνακα 1 παρουσιάζονται οι εργασίες που επιλέχθηκαν.

Πίνακας 1. *Εργασίες, των οποίων τα δεδομένα, χρησιμοποιήθηκαν στην ανάλυσή καθώς και τα scores και softwares που χρησιμοποιήθηκαν στη κάθε μια.*

Συγγραφείς	**Έτος δημοσίευσης**	**PMID**	**Κριτήρια**	**SOFTWARE**
Kapell et al.	2021	34046594	Localization Probability >= 0.99	MaxQuant
Wang et al.	2021	33682419	Localization Probability >= 0.99 PEP <= 0.01	MaxQuant
Marsden et al.	2021	34181092	Localization Probability >= 0.99 PEP <= 0.01	MaxQuant
Hartel et al.	2020	32429667	Localization Probability >= 0.99 Andromeda >= 100	MaxQuant
Wei et al.	2020	33344439	Localization Probability >= 0.99 Andromeda >= 100	MaxQuant

Kundinger et al.	2020	31994892	Localization Probability >= 0.99 Andromeda >= 100	MaxQuant
Lim et al.	2020	32467672	Modscore > 20	Sequest
Lim et al.	2020	32468700	Modscore > 20	Sequest
Spadotto et al.	2020	31777917	Localization Probability >= 0.99 Andromeda >= 100	MaxQuant-hmSEEKER
Hartel et al.	2019	31451547	Localization Probability >= 0.99 PEP <= 0.01	MaxQuant
Wang et al.	2019	31072472	Localization Probability >= 0.99 PEP <= 0.01	MaxQuant
Musiani et al.	2019	30940768	Localization Probability >= 0.99 PEP <= 0.01 Andromeda >= 100	MaxQuant-hmSEEKER
Fong et al.	2019	31408619	Localization Probability >= 0.99 Andromeda >= 100	MaxQuant-hmSEEKER
Gu et al.	2016	26635363	Modscore > 20	Sequest

Όπως φαίνεται στο πίνακα 1 στις εργασίες όπου οι ερευνητές χρησιμοποίησαν το software MaxQuant επιλέχθηκαν τα κριτήρια localization probability >= 0.99, PEP <= 0.01 και Andromeda >= 100. Το localization probability είναι η πιθανότητα ενός συγκεκριμένου αμινοξέος ενός πεπτιδίου να είναι μεθυλιωμένο..Το PEP (Posterior Error Probability) είναι η πιθανότητα μεταγενέστερου σφάλματος του καλύτερα ταυτοποιημένου τροποποιημένου πεπτιδίου που φέρει αυτή τη θέση μεθυλίωσης. Τέλος, το Andromeda score υπολογίζεται ως ο δεκαπλάσιος λογάριθμος της πιθανότητας να ταιριάζουν τουλάχιστον k από τις n θεωρητικές μάζες κατά τύχη (Tyanova et al., 2016).

Ο αλγόριθμος A-Score υπολογίζει το score πιθανοτήτων των μεθυλιωμένων θέσεων με βάση την αντιστοίχιση πιθανών “site-determining ” ιόντων b και y˙ Οι Lim et al. χρησιμοποίησαν τον όρο modscore για να χαρακτηρίσουν το αποτέλεσμα του A-Score αλγορίθμου. Συνήθως, από τους ερευνητές επιλέγεται το κατώφλι $p < 0,05$, AScore > 13. Ωστόσο, στη συγκεκριμένη ανάλυση επιλέχθηκαν τα μεθυλιωμένα πεπτίδια με modscore > 20 ώστε να επιτευχθεί καλύτερο φιλτράρισμα.

Το σύνολο των μεθυλιωμένων πεπτιδίων που αποκτήθηκαν από τις παραπάνω εργασίες είναι 84.998. Ωστόσο, σε αυτό το σύνολο υπάρχει μεγάλος αριθμός επαναλαμβανόμενων πεπτιδίων με ίδιες θέσεις μεθυλίωσης. Επίσης, τα περισσότερα πεπτίδια περιέχουν παραπάνω από μία θέσεις μεθυλίωσης. Έχοντας επιλέξει τα κατάλληλα scores καθώς και τα threshold αυτών δημιουργήθηκαν scripts για κάθε εργασία ξεχωριστά προκειμένου να γίνει εξόρυξη των μεθυλιωμένων πεπτιδίων που πληρούσαν τα κριτήρια. Για τη δημιουργία των scripts χρησιμοποιήθηκαν οι γλώσσες προγραμματισμού Python και Perl. Έτσι, δημιουργήθηκε το πρώτο σύνολο δεδομένων που αποτελούνταν από πεπτίδια όλων των εργασιών μετά το φιλτράρισμα και κάθε πεπτίδιο έφερε μοναδική θέση μεθυλίωσης R ή K. Σε αυτό το σύνολο δεδομένων υπήρχε και μια δεύτερη στήλη που προσδιόριζε το PMID της εργασίας και το πείραμα από το οποίο προήλθε το πεπτίδιο. Ο αριθμός των μεθυλιωμένων πεπτιδίων που περιέχονται σε αυτό το σύνολο δεδομένων είναι 17.061.

2.3 Αντιστοίχιση πρωτεϊνών και πεπτιδίων

Εν συνεχεία, δημιουργήθηκε ένα Python script προκειμένου να βρεθεί η πρωτεΐνη του ανθρώπινου πρωτεώματος από την οποία προήλθε το κάθε πεπτίδιο καθώς και την ακριβή θέση του μεθυλιωμένου αμινοξέος πάνω σε αυτή. Σκοπός αυτού ήταν να ελεγχθεί η πιστότητα των πεπτιδίων που προήλθαν από τη ανάλυση της φασματομετρίας μάζας, καθώς υπάρχει ο κίνδυνος να προκύψουν δεδομένα πεπτιδίων με λανθασμένες αλληλουχίες αμινοξέων. Αυτό μπορεί να οφείλεται στις συνθήκες κάτω από τις οποίες πραγματοποιήθηκε η ανάλυση καθώς και από πιθανές επιμολύνσεις του υπό εξέταση δείγματος. Παράλληλα, ένας ακόμη λόγος της αναζήτησης της θέσης των μεθυλιωμένων αμινοξέων στις πρωτεΐνες αφορούσε τη δημιουργία πεπτιδίων όπου το μεθυλιωμένο αμινοξύ θα βρισκόταν στο κέντρο του πεπτιδίου, και θα πλαισιωνόταν αριστερά και δεξιά από ένα συγκεκριμένο αριθμό αμινοξέων. Ωστόσο, αυτό θα αναλυθεί παρακάτω.

Σύμφωνα λοιπόν με την διαδικασία που ακολουθήθηκε, αρχικά ανακτήθηκε ολόκληρο το πρωτέωμα του ανθρώπου από την Ensembl με όνομα Homo_sapiens.GRCh38.pep.all.fa. Πρέπει να αναφερθεί ότι το πρωτέωμα του ανθρώπου περιέχει παραπάνω από μία ισομορφές για κάθε γονίδιο. Έτσι, δημιουργήθηκε ένα Python script το οποίο φίλτραρε το πρωτέωμα λαμβάνοντας μόνο τη μεγαλύτερη σε μήκος πρωτεΐνη για κάθε γονίδιο. Το ίδιο script στη συνέχεια διάβαζε το κάθε πεπτίδιο από το αρχικό σύνολο δεδομένων και αναζητούσε την πρωτεΐνη με την οποία υπάρχει πλήρη επικάλυψη. Στη περίπτωση που το πεπτίδιο ταίριαζε με περισσότερες από μία πρωτεΐνες το συγκεκριμένο πεπτίδιο δεν λαμβάνονταν υπόψη στη μετέπειτα ανάλυση. Αυτό αποτελεί ένα δεύτερο φιλτράρισμα του συνόλου δεδομένων, το οποίο τώρα περιέχει 3.622 πεπτίδια με μοναδικές θέσεις μεθυλίωσης R και K.

2.4 Δημιουργία μοτίβων 15, 29 και 41 αμινοξέων

Η αντιστοίχιση πεπτιδίων με τις πρωτεΐνες από τις οποίες προήλθαν εξυπηρετούσε και στη δημιουργία πεπτιδίων όπου η μεθυλιωμένη θέση θα βρισκόταν στο κέντρο του πεπτιδίου. Αυτή η μορφή είναι απαραίτητη για να δωθεί ως δεδομένα εισόδου στον αλγόριθμο μηχανικής μάθησης. Το ίδιο script λοιπόν έχοντας εντοπίσει τη θέση του μεθυλιωμένου αμινοξέος πάνω στη πρωτεΐνη λάμβανε ένα συγκεκριμένο αριθμό αμινοξέων αριστερά και δεξιά αυτής. Σε περίπτωση που το μεθυλιωμένο αμινοξύ βρισκόταν κοντά στην αρχή ή το τέλος της πρωτεΐνης, το μοτίβο δημιουργούνταν κανονικά, χρησιμοποιώντας το σύμβολο (_) για τα αμινοξέα που δεν υπήρχαν (βλ. εικόνα 1). Έτσι, προέκυψαν 3 σύνολα δεδομένων με τον ίδιο αριθμό πεπτιδίων, όπου στο πρώτο τα πεπτίδια είχαν μήκος 15 αμινοξέων, στο δεύτερο 29 και στο τρίτο 41· τα σύνολα δεδομένων συμβολίζονται προς το παρόν ως -7_AA_+7, -14_AA_+14 και -20_AA_+20 αντίστοιχα.

Εικόνα 1. *Σχηματική απεικόνιση της δημιουργίας των μοτίβων διαφορετικών μηκών. Παρουσιάζεται η διαδικασία δημιουργίας μοτίβων για μεθυλιωμένα αμινοξέα που βρίσκονται κοντά στη αρχή, στο κέντρο ή στο τέλος μιας πρωτεΐνης.*

2.5 Δημιουργία HQ και LQ συνόλων δεδομένων

Όπως αναφέρθηκε παραπάνω, κάθε πεπτίδιο των συνόλων δεδομένων συνοδεύεται από το PMID της εργασίας και του πειράματος από το οποίο προήλθε. Έτσι, για καθένα από τα τρία σύνολα δεδομένων δημιουργήθηκαν τα High Quality (HQ) και Low Quality (LQ). Πιο συγκεκριμένα, τα HQ σύνολα δεδομένων περιείχαν πεπτίδια όπου οι θέσεις μεθυλίωσης βρέθηκαν σε τουλάχιστον δύο high-throughput πειράματα. Από την άλλη, τα LQ σύνολα αποτελούνταν από όλα τα πεπτίδια που προέκυψαν μετά τα φιλτραρίσματα. Έτσι, προκύπτουν 6 νέα σύνολα δεδομένων.
Επιπλεόν, δημιουργήθηκαν αντίστοιχα αρνητικά σετ από λυσίνες και αργινίνες για τις οποίες δεν υπήρχαν ενδείξεις μεθυλίωσης.

2.6 Αλγόριθμοι μηχανικής μάθησης

Προκειμένου να δημιουργεί ένα αποτελεσματικό εργαλείο πρόβλεψης θέσεων μεθυλίωσης οι ομάδες δεδομένων που προέκυψαν χρησιμοποιήθηκαν για την εκπαίδευση αλγορίθμων που υπάρχουν στο πρόγραμμα Weka.

2.6.1 Αλγόριθμοι κατηγοριοποίησης του Weka

Το Weka είναι ένα ολοκληρωμένο λογισμικό, μια συλλογή αλγορίθμων μηχανικής μάθησης. Περιλαμβάνει εργαλεία για data preparation, classification, regression, clustering, association rules mining, and visualization.

2.6.2 Δημιουργία κλάσεων στα σύνολα δεδομένων

Ενας τρόπος πρόβλεψης θέσεων μεθυλίωσης στις πρωτεΐνες είναι με την χρήση αλγορίθμων κατηγοριοποίησης. Στη μηχανική μάθηση, κατηγοριοποίηση είναι μια supervised learning μέθοδος που βασικά κατηγοριοποιεί ένα σύνολο δεδομένων σε κλάσεις. Επομένως, καθένα από τα 6 σύνολα δεδομένων που δημιουργήθηκαν προηγουμένως έπρεπε να περιέχουν δεδομένα που να κατηγοριοποιούνται σε 2 κλάσεις. Η πρώτη κλάση αφορά πεπτίδια που περιέχουν μεθυλιωμένες θέσεις, τα οποία ονομάστηκαν positives, και είναι αυτά που υπήρχαν ήδη στο σύνολο δεδομένων. Αντίθετα, η δεύτερη κλάση αφορά πεπτίδια που περιέχουν μη μεθυλιωμένες θέσεις από πρωτεΐνες που δεν υπάρχει ένδειξη ότι μεθυλιώνονται, και ονομάστηκαν negatives. Για την δημιουργία των πεπτιδίων της δεύτερης κλάσης χρησιμοποιήθηκε το πρωτέωμα του ανθρώπου. Πιο συγκεκριμένα, δημιουργήθηκε ένα Python script το όποιο αρχικά εντόπισε όλες τις λυσίνες και αργινίνες των πρωτεϊνών και δημιουργούσε τα αντίστοιχα πεπτίδια μήκους 15, 21 και 41 όπου και πάλι το κεντρικό αμινοξύ ήταν λυσίνη ή αργινίνη. Από το σύνολο των πεπτιδίων που προέκυψαν αφαιρέθηκαν αυτά που υπήρχαν στη πρώτη κλάση. Έτσι, σε κάθε ένα από αυτά τα 6 σύνολα δεδομένων, εκτός από τα positives, προστέθηκε ίδιος αριθμός τυχαία επιλεγμένων negatives πεπτιδίων.

2.6.3 Εφαρμογή του προγράμματος Weka

Τα 6 σύνολα δεδομένων που δημιουργήθηκαν αποτελούνταν από πεπτίδια με μεθυλιωμένες λυσίνες και αργινίνες. Ωστόσο, η συγκεκριμένη μελέτη στοχεύει στη πρόβλεψη μεθυλιωμένων θέσεων μεμονωμένα για λυσίνες και αργινίνες. Επομένως, τα 6 σύνολα δεδομένων χωρίζονται επιμέρους για λυσίνες και αργινίνες και προκύπτουν 12 σύνολα (βλ. πίνακα 2).

Πίνακας 2. Συμβολισμοί συνόλων δεδομένων και ο αριθμός των πεπτιδίων που περιέχουν.

	Αργινίνες	Πεπτίδια	Λυσίνες	Πεπτίδια
HQ	7_R_7	2566	7_K_7	332
	14_R_14	2566	14_K_14	332
	20_R_20	2566	20_K_20	332
LQ	7_R_7	5612	7_K_7	1630
	14_R_14	5612	14_K_14	1630
	20_R_20	5612	20_K_20	1630

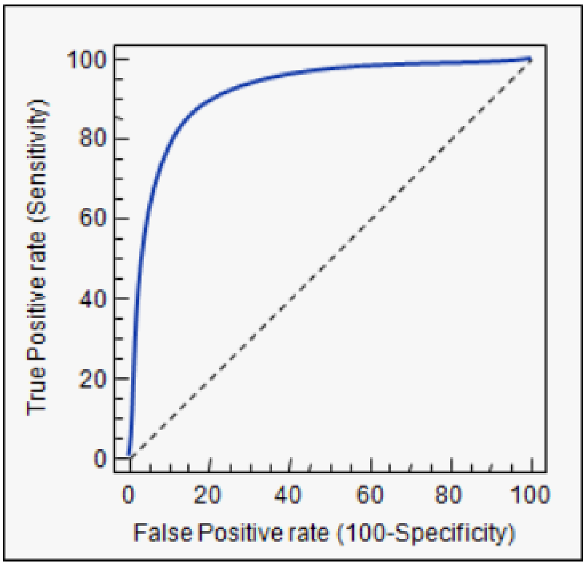
Για τον έλεγχο εγκυρότητας ενός μοντέλου, το Weka παρέχει την τεχνική Cross-validation που αξιολογεί τον τρόπο με τον οποίο τα αποτελέσματα μιας στατιστικής ανάλυσης θα μπορούν να γενικευτούν σε ένα ανεξάρτητο σύνολο δεδομένων. Χρησιμοποιείται κυρίως σε αναλύσεις όπου ο στόχος είναι η πρόβλεψη και πρέπει να εκτιμηθεί η ακρίβεια απόδοσης του μοντέλου πρόβλεψης στην πράξη. Επομένως, το Cross-validation πρέπει να δοκιμάσει τη ικανότητα του μοντέλου να προβλέπει νέα δεδομένα που δεν χρησιμοποιήθηκαν για την εκπαίδευση του, προκειμένου να επισημανθούν προβλήματα όπως η υπερπροσαρμογή ή η μεροληψία επιλογής (Witten et al., 2017). Όσον αφορά το τρόπο λειτουργίας της μεθόδου, αυτή χρησιμοποιεί διαφορετικά τμήματα των δεδομένων για να δοκιμάσει και να εκπαιδεύσει ένα μοντέλο σε διαφορετικές επαναλήψεις. Έτσι, καθορίζεται ένας σταθερός αριθμός folds. Στη ανάλυση χρησιμοποιήθηκαν 5-folds. Τα δεδομένα χωρίζονται τυχαία σε πέντε περίπου ίσα τμήματα στα οποία η κάθε κλάση αναπαρίσταται σε περίπου ίδιες αναλογίες όπως στο πλήρες σύνολο δεδομένων · το καθένα με τη σειρά του χρησιμοποιεί ένα σύνολο για testing και το υπόλοιπο χρησιμοποιείται για training. Δηλαδή, χρησιμοποιούνται τα 4/5 για training και το 1/5 για testing, το οποίο επαναλαμβάνεται πέντε φορές. Κάθε τμήμα εκτελείται με τη σειρά του και το πρόγραμμα εκμάθησης εκπαιδεύεται στα υπόλοιπα τέσσερα πέμπτα· τότε το ποσοστό σφάλματός του υπολογίζεται στο τμήμα που δεν χρησιμοποιείται εκείνη τη στιγμή (holdout set). Έτσι η διαδικασία εκμάθησης εκτελείται συνολικά 5 φορές σε διαφορετικά σετ εκπαίδευσης. Τέλος, οι 5 εκτιμήσεις σφαλμάτων υπολογίζονται κατά μέσο όρο για να προκύψει μια συνολική εκτίμηση σφάλματος της πρόβλεψης (Witten et al., 2017).

Για την αξιολόγηση της απόδοσης πρόβλεψης των αλγορίθμων, το πρόγραμμα Weka παρέχει ένα σύνολο metrics τα οποία υπολογίζονται κατά την εφαρμογή των αλγορίθμων. Τα metrics που επιλέχθηκαν για να ελεγχθεί η απόδοση πρόβλεψης των μοντέλων ήταν το Matthews correlation coefficient (MCC), το F-measure και ROC-AREA.

Ο MCC χρησιμοποιείται στη μηχανική μάθηση ως μέτρο της ποιότητας της κατηγοριοποίησης με πολλές κλάσεις. Λαμβάνει υπόψη τα αληθινά και τα ψευδή θετικά και αρνητικά και γενικά θεωρείται ως ένα ισορροπημένο μέτρο που μπορεί να χρησιμοποιηθεί ακόμα κι αν οι κλάσεις είναι πολύ διαφορετικών μεγεθών. Το MCC είναι στην ουσία μια τιμή συντελεστή συσχέτισης μεταξύ -1 και +1. Ένας συντελεστής +1 αντιπροσωπεύει μια τέλεια πρόβλεψη, 0 μια μέση τυχαία πρόβλεψη και -1 μια αντίστροφη πρόβλεψη.

Το F-Measure παρέχει έναν τρόπο συνδυασμού της ακρίβειας και της recall σε ένα ενιαίο μέτρο που καταγράφει και τις δύο ιδιότητες. Η ακρίβεια είναι μια μέτρηση που ποσοτικοποιεί τον αριθμό των αληθινών θετικών προβλέψεων. Υπολογίζεται ως το άθροισμα των αληθινών θετικών σε όλες τις κλάσεις διαιρούμενο με το άθροισμα των αληθινών θετικών και των ψευδώς θετικών σε όλες τις κλάσεις. Από την άλλη, η recall είναι ένα μέτρο που ποσοτικοποιεί τον αριθμό των αληθινών θετικών προβλέψεων που έγιναν από όλες τις θετικές προβλέψεις που θα μπορούσαν να είχαν συμβεί. Υπολογίζεται ως το άθροισμα των αληθινών θετικών σε όλες τις κλάσεις διαιρούμενο με το άθροισμα των αληθινών θετικών και των ψευδών αρνητικών σε όλες τις κλάσεις. Από μόνη της, ούτε η ακρίβεια ούτε η recall είναι αρκετή ως metrics για να ελεγχθεί η απόδοση πρόβλεψης ενός μοντέλου. Μπορούμε να έχουμε ταυτόχρονα εξαιρετική ακρίβεια με κακή recall, ή εναλλακτικά, κακή ακρίβεια με εξαιρετική recall. Το F-measure παρέχει έναν τρόπο έκφρασης και των δύο με ένα μόνο σκορ. Μόλις υπολογιστεί η ακρίβεια και η recall για ένα πρόβλημα κατηγοριοποίησης με πολλαπλές κλάσεις οι δύο τιμές μπορούν να συνδυαστούν στον υπολογισμό του F-Measure (Hripcsak, 2005).

Τέλος, σημαντικό μέτρο αξιολόγησης είναι το εμβαδόν κάτω από την καμπύλη ROC (Receiver Operating Characteristic). Για το σχηματισμό της καμπύλης ROC, χρησιμοποιούνται διάφορες τιμές κατωφλίου και σημειώνονται κάθε φορά τα ποσοστά True Positive Rate (TPR=TP/P) δηλαδή η ευαισθησία και False Positive Rate (FPR=FP/N) το οποίο είναι εξειδίκευση. Αυτά τα ζεύγη τιμών προστίθενται σε ένα γράφημα όπου ο άξονας y αντιστοιχεί στα TPR και ο άξονας x στα FPR (βλ. εικόνα 2). Τα πλεονεκτήματα του μέτρου είναι ότι αποκτάται πληροφορία για την ποιότητα της πρόβλεψης του μοντέλου για διάφορες τιμές του κατωφλίου και επίσης είναι ανεξάρτητη από την ανισορροπία των κλάσεων στα δεδομένα (Faraggi & Reiser, 2002)



Εικόνα 2. *Διάγραμμα απεικόνισης της καμπύλης ROC. Στον άξονα x απεικονίζεται το true positive rate, ενώ στον άξονα y το false positive rate (Zweig & Campbell, 1993).*

Οι διακεκομμένες τελείες στο διάγραμμα απεικονίζουν ένα μοντέλο ταξινόμησης το οποίο προβλέπει όχι καλύτερα από κάποιο μοντέλο το οποίο έχει επιλεχθεί τυχαία. Τα πολύ αποτελεσματικά πλησιάζουν την πάνω αριστερή γωνία. Η περιοχή κάτω από την καμπύλη ROC (Area Under the Curve, AUC) είναι ένα ποσοτικό μέτρο αξιολόγησης της διακριτικής/προβλεπτικής ικανότητας ενός μοντέλου, το οποίο βασίζεται στην καμπύλη ROC. Όσο μεγαλύτερο και πιο κοντά στην μονάδα είναι το AUC, τόσο πιο αποτελεσματικό είναι το μοντέλο, και στη περίπτωση που είναι μονάδα το μοντέλο είναι ιδανικό για τη πρόβλεψη (Zweig & Campbell, 1993)

3.6.4 Χρήση Αλγορίθμων με τη μέθοδο Κατηγοριοποίησης με μοναδική ετικέτα

Προηγουμένως αναφέρθηκε η χρησιμοποίηση αλγορίθμων κατηγοριοποίησης. Ωστόσο, για μεγαλύτερη ακρίβεια ο τρόπος της συγκεκριμένης μάθησης με επίβλεψη είναι η κατηγοριοποίηση με μοναδική ετικέτα, που στην ουσία κατηγοριοποιεί ένα παράδειγμα σε μία από τις δύο κλάσεις. Συγκεκριμένα, ονομάζεται single label classification, όπου στη διαδικασία της μάθησης σε ένα σύνολο παραδειγμάτων συσχετίζεται κάθε παράδειγμα του με μία μοναδική ετικέτα. Αν ο αριθμός των ετικετών είναι 2, όπως συμβαίνει και στη δικιά μας ανάλυση, το πρόβλημα μάθησης ονομάζεται δυαδικό (binary) πρόβλημα κατηγοριοποίησης. Στα σύνολα δεδομένων που έχουν δημιουργηθεί, το κάθε πεπτίδιο αποτελεί ένα παράδειγμα το οποίο φέρει μία ετικέτα, η οποία εξαρτάται από το αν το πεπτίδιο φέρει θέση μεθυλίωσης ή όχι, δηλαδή αυτό που αναφέρθηκε προηγουμένως ως positive και negative πεπτίδιο.

Σε μια ανάλυση που στοχεύει στη πρόβλεψη μέσω μάθησης με κατηγοριοποίηση μοναδικής ετικέτας, τα σύνολα δεδομένων πρέπει να αποτελούνται από παραδείγματα τα οποία φέρουν μοναδική ετικέτα και διάφορα χαρακτηριστικά που διακρίνουν το κάθε παράδειγμα. Στη προκειμένη ανάλυση τα παραδείγματα είναι τα πεπτίδια που προέκυψαν μετά τα φιλτραρίσματα, ενώ οι ετικέτες είναι δύο και χαρακτηρίζουν το πεπτίδιο ως μεθυλιωμένο ή όχι. Όσον αφορά τα χαρακτηριστικά των παραδειγμάτων, επιλέχθηκαν δύο τρόποι προσδιορισμού και εκπροσώπησης τους. Αρχικά, στην πρώτη εφαρμογή των αλγορίθμων η πρόβλεψη στηρίζεται στη πρωτοταγή δομή των πεπτιδίων, καθώς ως χαρακτηριστικά των παραδειγμάτων χρησιμοποιήθηκε η αλληλουχία των αμινοξέων των πεπτιδίων. Πιο συγκεκριμένα, στα σύνολα δεδομένων όπου τα πεπτίδια αποτελούνταν από 15 αμινοξέα, κάθε αμινοξύ θεωρήθηκε ως ένα ξεχωριστό χαρακτηριστικό, με αποτέλεσμα να δημιουργείται ένα νέο σύνολο δεδομένων προσαρμοσμένο στο πρόγραμμα Weka αποτελούμενο από 15 στήλες χαρακτηριστικών που κάθε μία έφερε ένα μοναδικό αμινοξύ. Ύπηρχε και μια δέκατη έκτη στήλη που αφορούσε την ετικέτα του παραδείγματος (Εικόνα 3. Α). Με τον ίδιο τρόπο προσαρμόστηκαν και τα υπόλοιπα σύνολα δεδομένων στο πρόγραμμα Weka, όπου σε κάθε περίπτωση ο αριθμός των αμινοξέων είναι ταυτόσημος με τον αριθμό των χαρακτηριστικών/στήλων των παραδειγμάτων.

A

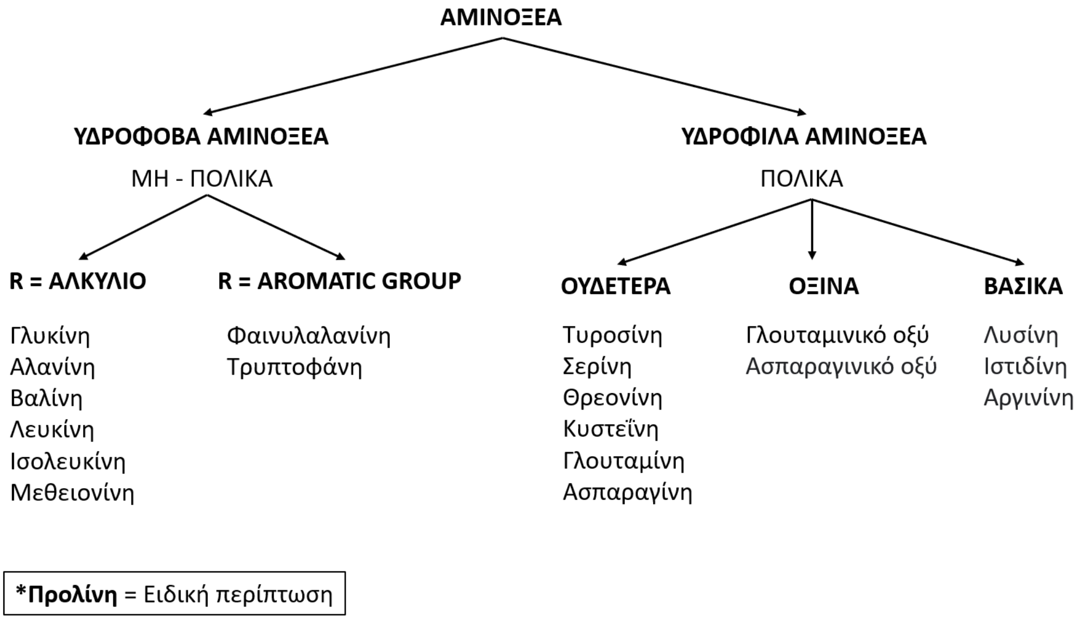
P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	Outcome
K	I	M	N	H	K	E	R	F	Q	F	P	A	Q	V	n
V	S	A	A	L	Q	S	R	Q	Q	A	A	P	D	A	n
L	K	R	F	L	L	A	R	R	S	R	R	G	L	F	n
S	C	D	A	G	M	C	R	A	Y	F	H	V	T	C	n
P	D	P	A	D	P	K	R	E	P	L	P	S	R	P	n
A	V	S	G	T	V	R	R	L	Q	G	V	L	G	G	n
R	T	L	V	A	V	E	R	P	L	D	D	I	I	A	n

B

P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	Outcome
PB	NPALK	NPALK	PNEU	PB	PB	PA	PB	NPAROM	PNEU	NPAROM	SC	NPALK	PNEU	NPALK	n
NPALK	PNEU	NPALK	NPALK	NPALK	PNEU	PNEU	PB	PNEU	PNEU	NPALK	NPALK	SC	PA	NPALK	n
NPALK	PB	PB	NPAROM	NPALK	NPALK	NPALK	PB	PB	PNEU	PB	PB	NPALK	NPALK	NPAROM	n
PNEU	PNEU	PA	NPALK	NPALK	NPALK	PNEU	PB	NPALK	PNEU	NPAROM	PB	NPALK	PNEU	PNEU	n
SC	PA	SC	NPALK	PA	SC	PB	PB	PA	SC	NPALK	SC	PNEU	PB	SC	n
NPALK	NPALK	PNEU	NPALK	PNEU	NPALK	PB	PB	NPALK	PNEU	NPALK	NPALK	NPALK	NPALK	NPALK	n
PB	PNEU	NPALK	NPALK	NPALK	NPALK	PA	PB	SC	NPALK	PA	PA	NPALK	NPALK	NPALK	n

Εικόνα 3. *Απεικόνιση των συνόλων δεδομένων που χρησιμοποιήθηκαν στο πρόγραμμα Weka. A) Αποτελεί το σύνολο δεδομένων όπου τα χαρακτηριστικά των παραδειγμάτων είναι τα αμινοξέα. Στο τέλος υπάρχει η στήλη που αφορά την ετικέτα του παραδείγματος. B) Αποτελεί το σύνολο δεδομένων όπου τα χαρακτηριστικά των παραδειγμάτων αφορούν τις φυσικοχημικές ιδιότητες των αμινοξέων. Σε κάθε στήλη αντί να αντιπροσωπεύεται το αμινοξύ, παρουσιάζεται η ομάδα που ανήκει. Το PB συμβολίζει την ομάδα polar basic, το NPALK την ομάδα non polar alkyl, το PNEU την ομάδα polar neutral, το PA την ομάδα polar acidic, το NPAROM την ομάδα non polar aromatic και το SC special case που αφορά μόνο την προλίνη.*

Ο δεύτερος τρόπος προσδιορισμού των χαρακτηριστικών των παραδειγμάτων αφορά τις φυσικοχημικές ιδιότητες των αμινοξέων. Τα αμινοξέα μπορούν να ομαδοποιηθούν με βάση τις ιδιότητες της πλευρικής ομάδας του κάθε αμινοξέος. Μπορεί να είναι πολικά ή μη πολικά. Στα πολικά αμινοξέα οι ομάδες "R" είναι υδρόφιλες, ενώ στα μη πολικά αμινοξέα είναι υδρόφοβες (Hubbard & Kamran Haider, 2010). Η ομαδοποίηση των αμινοξέων φαίνεται στη εικόνα 4. Έτσι, σε αυτή τη περίπτωση ο αριθμός των στήλων παραμένει ο ίδιος, αλλάζει μόνο ο συμβολισμός των αμινοξέων. Το κάθε αμινοξύ συμβολίζεται με βάση την ομάδα που ανήκει, σε αντίθεση με την πρώτη περίπτωση που αναφέρεται το αμινοξύ αυτό καθ' αυτό (βλ. εικόνα 3.Β). Οι πέντε ομάδες αμινοξέων είναι μη πολικά-alkyl αμινοξέα, μη πολικά- αρωματικά αμινοξέα, πολικά ουδέτερα αμινοξέα, πολικά όξινα αμινοξέα και πολικά βασικά αμινοξέα. Υπάρχει και μία έκτη ομάδα που αφορά την προλίνη καθώς θεωρείται ως μια ιδιαίτερη ομάδα αμινοξέος. Αυτό συμβαίνει γιατί η προλίνη είναι μοναδική μεταξύ των τυπικών αμινοξέων στο ότι δεν έχει ελεύθερες αμινο- και ελεύθερες καρβοξυλικές ομάδες. Αντίθετα, οι πλευρικές του αλυσίδες σχηματίζουν μια κυκλική δομή καθώς το άτομο αζώτου συνδέεται με δύο άτομα άνθρακα (Allen et al., 2004).



Εικόνα 4. *Σχηματική απεικόνιση της κατηγοριοποίησης των αμινοξέων όσον αφορά τις φυσικοχημικές τους ιδιότητες.*

Στη συνέχεια, στο κάθε σύνολο δεδομένων ξεχωριστά εφαρμόστηκαν όλοι οι αλγόριθμοι κατηγοριοποίησης του προγράμματος Weka.

3.ΑΠΟΤΕΛΕΣΜΑΤΑ-ΣΥΖΗΤΗΣΗ

3.1 Απόδοση πρόβλεψης των μοντέλων κατηγοριοποίησης με βάση τη πρωτοταγή δομή των πεπτιδίων

Αφού εφαρμόστηκαν όλοι οι αλγόριθμοι κατηγοριοποίησης που παρείχε το πρόγραμμα Weka επιλέχθηκαν για κάθε σύνολο δεδομένων αυτοί που πέτυχαν την καλύτερη απόδοση πρόβλεψης θέσεων μεθυλίωσης R και K όπως φαίνεται στους παρακάτω πίνακες.

3.1.1 Πρόβλεψη θέσεων Μεθυλίωσης Αργινίνης

Πίνακας 3. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων HQ 7_R_7.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,81	0,62	0,868	80,98%
NaiveBayes	0,809	0,619	0,868	80,94%
NaiveBayesUpdateable	0,809	0,619	0,868	80,94%
SimpleLogistic	0,798	0,596	0,858	79,77%
LMT	0,798	0,596	0,858	79,77%
RandomForest	0,791	0,582	0,858	79,11%

Πίνακας 4. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων HQ 14_R_14.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,824	0,648	0,884	82,36%
NaiveBayes	0,824	0,648	0,884	82,36%
NaiveBayesUpdateable	0,824	0,648	0,884	82,36%
SimpleLogistic	0,782	0,565	0,85	78,23%
LMT	0,782	0,565	0,85	78,23%
RandomForest	0,814	0,629	0,881	81,43%

Πίνακας 5. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων HQ 20_R_20.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,83	0,661	0,882	83,00%
NaiveBayes	0,83	0,661	0,882	83,00%
NaiveBayesUpdateable	0,83	0,661	0,882	83,00%
SimpleLogistic	0,766	0,533	0,842	76,61%
LWL	0,661	0,428	0,877	68,37%
RandomForest	0,819	0,638	0,881	81,87%

Πίνακας 6. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων LQ 7_R_7.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,707	0,417	0,771	70,74%
NaiveBayes	0,705	0,415	0,754	70,60%
NaiveBayesUpdateable	0,705	0,415	0,754	70,60%
LWL	0,611	0,351	0,752	64,33%
classificationViaRegression	0,693	0,388	0,752	69,33%
RandomForest	0,707	0,417	0,771	70,74%

Πίνακας 7. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων LQ 14_R_14.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,72	0,447	0,761	72,14%
NaiveBayes	0,721	0,448	0,761	72,19%
NaiveBayesUpdateable	0,721	0,448	0,761	72,19%
LWL	0,584	0,317	0,754	62,40%
LMT	0,688	0,378	0,751	68,88%
RandomForest	0,712	0,426	0,77	71,25%

Πίνακας 8. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων LQ 20_R_20.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,727	0,462	0,764	72,84%
NaiveBayes	0,726	0,461	0,764	72,79%
NaiveBayesUpdateable	0,726	0,461	0,764	72,79%
SimpleLogistic	0,687	0,376	0,756	68,77%
LWL	0,584	0,308	0,757	62,26%
RandomForest	0,715	0,432	0,773	71,54%

Σύμφωνα με τους παραπάνω πίνακες τα μοντέλα κατηγοριοποίησης για την πρόβλεψη θέσεων μεθυλίωσης R αποδίδουν καλύτερα στα HQ σύνολα δεδομένων συγκριτικά με την εφαρμογή αυτών στα LQ σύνολα δεδομένων. Πιο συγκεκριμένα, το F- measure, το MCC και το ROC AREA εμφανίζουν υψηλότερες τιμές στα HQ σύνολα δεδομένων. Οι τιμές του ROC Area στα HQ σύνολα δεδομένων είναι μεγαλύτερες του 0,86 το οποίο αντικατοπτρίζει μία εξαιρετικά ικανοποιητική απόδοση των μοντέλων. Αντίθετα, οι τιμές του ROC Area στα LQ σύνολα δεδομένων δεν ξεπερνούν το 0,77, γεγονός που υποδηλώνει μια σαφώς χαμηλότερη απόδοση. Το ίδιο συμπέρασμα προκύπτει από τα αποτελέσματα του F – measure αλλά και του MCC. Στα HQ σύνολα δεδομένων οι τιμές του F - measure ξεπερνούν το 0,81 ενώ στα LQ σύνολα δεδομένων η μεγαλύτερη τιμή του φτάνει στο 0,72. Οι τιμές του MCC στα HQ σύνολα δεδομένων φτάνει στο 0,66 ενώ στα LQ σύνολα δεδομένων φτάνει στο 0,44.

Από την άλλη, το μήκος των μεθυλιωμένων πεπτιδίων φαίνεται πως δεν επηρεάζει σημαντικά την πρόβλεψη θέσεων μεθυλίωσης R, καθώς στα διάφορα μήκη οι τιμές των μέτρων αξιολόγησης της απόδοσης της πρόβλεψης τόσο μεταξύ των HQ συνόλων δεδομένων όσο και των LQ συνόλων δεδομένων διαφέρουν ελάχιστα. Η καλύτερη απόδοση παρουσιάζεται στο σύνολο δεδομένων HQ 14_R_14 με την εφαρμογή του αλγόριθμου NaiveBayes. Φαίνεται να υπάρχει μια πολύ μικρή αύξηση της απόδοσης σε αυτό το μήκος πεπτιδίου αλλά αυτή η διαφορά δεν μπορεί να θεωρηθεί σημαντική. Πρέπει να αναφερθεί πως στα HQ σύνολα δεδομένων όταν εφαρμόζεται ο αλγόριθμος NaiveBayes η απόδοση είναι η υψηλότερη. Ωστόσο, στα LQ σύνολα δεδομένων αυτό συμβαίνει με τον αλγόριθμο Random Forest.

3.1.2 Πρόβλεψη θέσεων Μεθυλίωσης Λυσίνης

Πίνακας 9. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων HQ 7_K_7.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
bagging	0,545	0,09	0,581	54,52%
BayesNet	0,548	0,096	0,567	54,82%
classificationViaRegression	0,551	0,103	0,588	55,12%
LMT	0,57	0,146	0,587	57,23%
RandomSubSpace	0,563	0,127	0,588	56,33%
SimpleLogistic	0,57	0,146	0,587	57,23%

Πίνακας 10. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων HQ 14_K_14.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,601	0,203	0,667	60,12%
NaiveBayes	0,619	0,239	0,674	61,90%
NaiveBayesUpdateable	0,619	0,239	0,674	61,90%
MultiClassClassifierUpdateable	0,61	0,22	0,61	61,01%
hoeffdingTree	0,619	0,239	0,674	61,90%
RandomForest	0,577	0,155	0,637	57,74%

Πίνακας 11. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων HQ 20_K_20.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,607	0,215	0,656	60,71%
NaiveBayes	0,61	0,22	0,66	61,01%
NaiveBayesUpdateable	0,607	0,215	0,656	60,71%
Logistic	0,616	0,232	0,67	61,61%
MultiClassClassifier	0,616	0,232	0,67	61,61%
hoeffdingTree	0,61	0,22	0,661	61,01%

Πίνακας 12. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων LQ 7_K_7.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,604	0,208	0,651	60,41%
hoeffdingTree	0,604	0,208	0,651	60,41%
NaiveBayes	0,604	0,208	0,651	60,41%
NaiveBayesUpdateable	0,604	0,208	0,651	60,41%
RandomForest	0,6	0,2	0,647	59,98%
kStar	0,588	0,181	0,642	58,99%

Πίνακας 13. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων LQ 14_K_14.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,597	0,199	0,643	59,85%
NaiveBayes	0,598	0,201	0,643	59,98%
NaiveBayesUpdateable	0,598	0,201	0,643	59,98%
RandomCommittee	0,559	0,118	0,608	55,88%
hoeffdingTree	0,598	0,201	0,643	59,98%
RandomForest	0,588	0,178	0,634	58,87%

Πίνακας 14. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων LQ 20_K_20.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,612	0,23	0,669	61,38%
NaiveBayes	0,613	0,231	0,669	61,44%
NaiveBayesUpdateable	0,613	0,231	0,669	61,44%
HoeffdingTree	0,613	0,231	0,669	61,44%
RandomForest	0,584	0,169	0,636	58,45%
RandomCommittee	0,581	0,162	0,624	58,08%

Η απόδοση των μοντέλων ταξινόμησης για τη πρόβλεψη θέσεων μεθυλίωσης της Κ διαφέρει αρκετά με αυτή της R. Πιο συγκεκριμένα, από τους παραπάνω πίνακες φαίνεται ότι η απόδοση πρόβλεψης θέσεων μεθυλίωσης Κ δεν είναι ικανοποιητική, το οποίο υποστηρίζεται από τις χαμηλές τιμές των μέτρων ROC AREA, MCC και F – Measure. Επομένως, δεν μπορούν να διεξαχθούν αξιόπιστα συμπεράσματα για την πρόβλεψη μεθυλιωμένων θέσεων Κ. Αυτό αφορά τόσο τα HQ σύνολα δεδομένων όσο και τα LQ σύνολα δεδομένων. Πιο συγκεκριμένα, η μέγιστη τιμή του ROC Area φτάνει το 0,69, η μέγιστη τιμή του MCC φτάνει το 0,23 και η μέγιστη τιμή του F-Measure φτάνει το 0,61. Οι μέγιστες τιμές αυτών των μέτρων παρατηρούνται στο σύνολο δεδομένων LQ 20_K_20 με την εφαρμογή του αλγορίθμου NaiveBayes. Πρέπει να αναφερθεί πως τα σύνολα δεδομένων που παρουσιάζονται στους παραπάνω πίνακες εμφανίζουν καλύτερη απόδοση όταν εφαρμόζεται σε αυτά ο αλγόριθμος NaiveBayes. Δεν παρατηρείται στατιστικά σημαντική διαφορά στην απόδοση μεταξύ των HQ και LQ συνόλων δεδομένων. Αυτό ίσως οφείλεται στο μικρό αριθμό δεδομένων που περιέχονται στα σύνολα δεδομένων της Κ, όπως φαίνεται και στο πίνακα 2, γεγονός που επηρεάζει σημαντικά την ικανότητα πρόβλεψης.

Παράλληλα, ούτε στη περίπτωση της Κ δεν παρατηρείται σημαντική διαφορά της απόδοσης πρόβλεψης που να οφείλεται στο μήκος των πεπτιδίων. Παρατηρείται μια αύξηση της απόδοσης στα σύνολα δεδομένων LQ 20_K_20 και HQ 14_K_14. Ωστόσο, και σε αυτή τη περίπτωση η αύξηση αυτή δεν μπορεί να θεωρηθεί στατιστικά σημαντική.

3.2 Απόδοση πρόβλεψης των μοντέλων κατηγοριοποίησης με βάση το φυσικοχημικό χαρακτήρα των αμινοξέων των πεπτιδίων

Όπως αναφέρθηκε παραπάνω στα σύνολα δεδομένων έγινε τροποποίηση των χαρακτηριστικών των παραδειγμάτων που αφορούσε το συμβολισμό των αμινοξέων σύμφωνα με την ομάδα στην οποία ανήκουν. Αφού εφαρμόστηκαν εκ νέου όλοι οι αλγόριθμοι κατηγοριοποίησης σε όλα τα σύνολα δεδομένων επιλέχθηκαν και πάλι αυτοί που πέτυχαν υψηλότερη απόδοση στη πρόβλεψη της μεθυλίωσης των Κ και R όπως φαίνεται στους παρακάτω πίνακες.

3.2.1 Πρόβλεψη θέσεων Μεθυλίωσης Αργινίνης

Πίνακας 15. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων HQ 7_R_7 όπου τα χαρακτηριστικά των παραδειγμάτων αφορούν το φυσικοχημικό χαρακτήρα.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,744	0,487	0,809	74,36%
NaiveBayes	0,743	0,487	0,809	74,32%
NaiveBayesUpdateable	0,743	0,487	0,809	74,32%
classificationViaRegression	0,745	0,49	0,804	74,47%
HoeffdingTree	0,743	0,487	0,809	74,32%
RandomForest	0,734	0,47	0,803	73,46%

Πίνακας 16. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων HQ 14_R_14 όπου τα χαρακτηριστικά των παραδειγμάτων αφορούν το φυσικοχημικό χαρακτήρα.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,759	0,518	0,816	75,90%
NaiveBayes	0,759	0,519	0,817	75,93%
NaiveBayesUpdateable	0,759	0,519	0,817	75,93%
SimpleLogistic	0,742	0,484	0,81	74,18%
HoeffdingTree	0,759	0,519	0,817	75,93%
RandomForest	0,763	0,526	0,829	76,29%

Πίνακας 17. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων HQ 20_R_20 όπου τα χαρακτηριστικά των παραδειγμάτων αφορούν το φυσικοχημικό χαρακτήρα.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,746	0,493	0,805	74,59%
NaiveBayes	0,746	0,494	0,806	74,63%
NaiveBayesUpdateable	0,746	0,494	0,806	74,63%
SimpleLogistic	0,729	0,458	0,8	72,88%
HoeffdingTree	0,746	0,494	0,806	74,63%
RandomForest	0,74	0,479	0,814	73,97%

Πίνακας 18. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων LQ 7_R_7 όπου τα χαρακτηριστικά των παραδειγμάτων αφορούν το φυσικοχημικό χαρακτήρα.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,644	0,289	0,68	64,43%
NaiveBayes	0,644	0,289	0,68	64,42%
NaiveBayesUpdateable	0,644	0,289	0,68	64,42%
SimpleLogistic	0,636	0,273	0,679	63,65%
RandomSubSpace	0,629	0,258	0,682	62,90%
RandomForest	0,653	0,306	0,706	65,31%

Πίνακας 19. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων LQ 14_R_14 όπου τα χαρακτηριστικά των παραδειγμάτων αφορούν το φυσικοχημικό χαρακτήρα..*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,652	0,309	0,69	65,30%
NaiveBayes	0,651	0,308	0,69	65,28%
NaiveBayesUpdateable	0,651	0,308	0,69	65,28%
Logistic	0,643	0,286	0,693	64,29%
classificationViaRegression	0,644	0,289	0,694	64,43%
RandomForest	0,655	0,312	0,715	65,57%

Πίνακας 20. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων LQ 20_R_20 όπου τα χαρακτηριστικά των παραδειγμάτων αφορούν το φυσικοχημικό χαρακτήρα..*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,656	0,325	0,696	65,92%
NaiveBayes	0,656	0,325	0,696	65,92%
NaiveBayesUpdateable	0,656	0,325	0,696	65,92%
Logistic	0,655	0,311	0,706	65,53%
MultiClassClassifier	0,655	0,311	0,706	65,53%
RandomForest	0,663	0,327	0,722	66,35%

Σύμφωνα με τους παραπάνω πίνακες τα HQ σύνολα δεδομένων αποδίδουν καλύτερα σε σχέση με τα LQ σύνολα δεδομένων. Και σε αυτή τη περίπτωση βλέπουμε μια σαφώς ικανοποιητική απόδοση των HQ σύνολα δεδομένων, αλλά συνολικά χαμηλότερη από την απόδοση που προέκυψε από την πρώτη εφαρμογή των αλγορίθμων όπου τα χαρακτηριστικά των παραδειγμάτων αντιπροσώπευαν την πρωτοταγή δομή των πεπτιδίων.

Οι τιμές του ROC Area στα HQ σύνολα δεδομένων φτάνουν το 0,83 το οποίο δείχνει μία καλή απόδοση των μοντέλων. Αντίθετα, οι τιμές του ROC Area στα LQ σύνολα δεδομένων δεν ξεπερνούν το 0,72, γεγονός που υποδηλώνει μια σαφώς χαμηλότερη απόδοση. Το ίδιο συμπέρασμα προκύπτει από τα αποτελέσματα του F – measure αλλά και του MCC. Στα HQ σύνολα δεδομένων οι τιμές του F - measure φτάνουν το 0,74 ενώ στα LQ σύνολα δεδομένων η μεγαλύτερη τιμή του φτάνει στο 0,66. Οι τιμές του MCC στα HQ σύνολα δεδομένων φτάνει στο 0,53 ενώ στα LQ σύνολα δεδομένων φτάνει στο 0,33.

Ούτε σε αυτή την ανάλυση φαίνεται να επηρεάζεται η απόδοση πρόβλεψης της μεθυλίωσης των R απο το μήκος των πεπτιδίων. Ωστόσο πρέπει να αναφερθεί πως η καλύτερη απόδοση αφορά το σύνολο δεδομένων HQ 14_R_14 και τον αλγόριθμο Random Forest. Υπάρχει μια μικρή αύξηση της απόδοσης πρόβλεψης στα υψηλότερα μήκη όμως και πάλι αυτές οι διαφορές δεν μπορούν να θεωρηθούν στατιστικά σημαντικές. Είναι αξιοσημείωτο ότι με αυτή τη αλλαγή της δομής των χαρακτηριστικών των παραδειγμάτων ο αλγόριθμος που δίνει τα καλύτερα αποτελέσματα πρόβλεψης δεν είναι ο NaiveBayes αλλά ο Random Forest, αυτό συμβαίνει σε όλα τα παραπάνω σύνολα δεδομένων εκτός από το HQ 7_R_7 όπου είναι και πάλι ο NaiveBayes.

3.2.2 Πρόβλεψη θέσεων Μεθυλίωσης Λυσίνης

Πίνακας 21. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων HQ 7_K_7 όπου τα χαρακτηριστικά των παραδειγμάτων αφορούν το φυσικοχημικό χαρακτήρα..*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,584	0,169	0,615	58,43%
NaiveBayes	0,584	0,169	0,617	58,43%
NaiveBayesUpdateable	0,584	0,169	0,617	58,43%
classificationViaRegression	0,614	0,229	0,627	61,45%
RandomCommittee	0,584	0,169	0,623	58,43%
HoeffdingTree	0,584	0,169	0,617	58,43%

Πίνακας 22. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων HQ 14_K_14 όπου τα χαρακτηριστικά των παραδειγμάτων αφορούν το φυσικοχημικό χαρακτήρα.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,6	0,204	0,653	60,12%
NaiveBayes	0,6	0,204	0,654	60,12%
NaiveBayesUpdateable	0,6	0,204	0,654	60,12%
RandomSubSpace	0,542	0,083	0,553	54,17%
HoeffdingTree	0,6	0,204	0,654	60,12%
RandomForest	0,604	0,208	0,658	60,42%

Πίνακας 23. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων HQ 20_K_20 όπου τα χαρακτηριστικά των παραδειγμάτων αφορούν το φυσικοχημικό χαρακτήρα.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,565	0,131	0,602	56,55%
NaiveBayes	0,562	0,125	0,605	56,25%
NaiveBayesUpdateable	0,562	0,125	0,605	56,25%
SimpleLogistic	0,56	0,119	0,593	55,95%
HoeffdingTree	0,562	0,125	0,605	56,25%
RandomForest	0,568	0,137	0,621	56,85%

Πίνακας 24. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων LQ 7_K_7 όπου τα χαρακτηριστικά των παραδειγμάτων αφορούν το φυσικοχημικό χαρακτήρα.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,578	0,162	0,606	58,00%
NaiveBayes	0,578	0,161	0,606	58,00%
NaiveBayesUpdateable	0,578	0,161	0,606	58,00%
RandomSubSpace	0,566	0,132	0,604	56,59%
HoeffdingTree	0,578	0,161	0,606	58,00%
RandomForest	0,57	0,14	0,603	57,02%

Πίνακας 25. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων LQ 14_K_14 όπου τα χαρακτηριστικά των παραδειγμάτων αφορούν το φυσικοχημικό χαρακτήρα.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,601	0,225	0,645	60,83%
NaiveBayes	0,601	0,225	0,645	60,83%
NaiveBayesUpdateable	0,601	0,225	0,645	60,83%
Logistic	0,6	0,201	0,634	60,04%
HoeffdingTree	0,601	0,225	0,645	60,83%
RandomForest	0,591	0,186	0,646	59,24%

Πίνακας 26. *Αποτελέσματα αξιολόγησης των αλγορίθμων που εφαρμόστηκαν στο σύνολο δεδομένων LQ 2O_K_20 όπου τα χαρακτηριστικά των παραδειγμάτων αφορούν το φυσικοχημικό χαρακτήρα.*

Algorithm	F-Measure	MCC	ROC Area	Correctly Classified
BayesNet	0,591	0,201	0,645	59,73%
NaiveBayes	0,592	0,205	0,645	59,91%
NaiveBayesUpdateable	0,592	0,205	0,645	59,91%
SimpleLogistic	0,591	0,181	0,639	59,06%
HoeffdingTree	0,592	0,205	0,645	59,91%
RandomForest	0,613	0,228	0,655	61,38%

Η απόδοση των μοντέλων ταξινόμησης για τη πρόβλεψη θέσεων μεθυλίωσης της K διαφέρει και πάλι αρκετά με αυτή της R. Πιο συγκεκριμένα, από τους παραπάνω πίνακες φαίνεται ότι η απόδοση πρόβλεψης θέσεων μεθυλίωσης K δεν είναι ικανοποιητική, το οποίο υποστηρίζεται από τις χαμηλές τιμές των μέτρων ROC AREA, MCC και F – Measure. Επομένως, δεν μπορούν να διεξαχθούν αξιόπιστα συμπεράσματα για την πρόβλεψη μεθυλιωμένων θέσεων K. Αυτό αφορά τόσο τα HQ σύνολα δεδομένων όσο και τα LQ σύνολα δεδομένων. Πιο συγκεκριμένα, η μέγιστη τιμή του ROC Area φτάνει το 0,66, η μέγιστη τιμή του MCC φτάνει το 0,23 και η μέγιστη τιμή του F-Measure φτάνει το 0,61. Οι μέγιστες τιμές αυτών των μέτρων παρατηρούνται και πάλι στο σύνολο δεδομένων LQ 20_K_20 με την εφαρμογή του αλγορίθμου Random Forest. Πρέπει να αναφερθεί ότι στο ίδιο σύνολο δεδομένων με την προηγούμενη μορφή των χαρακτηριστικών των παραδειγμάτων η καλύτερη απόδοση προέκυψε με τον αλγόριθμο NaiveBayes. Επίσης, σε όλα τα σύνολα δεδομένων που παρουσιάζονται στους παραπάνω πίνακες εμφανίζουν καλύτερη απόδοση όταν εφαρμόζεται σε αυτά ο αλγόριθμος Random Forest, εκτός από το σύνολο δεδομένων HQ 7_K_7. Δεν παρατηρείται και πάλι σημαντική διαφορά στην απόδοση μεταξύ των HQ και LQ σύνολα δεδομένων.

Παράλληλα, δεν παρατηρείται σημαντική διαφορά της απόδοσης πρόβλεψης που να οφείλεται στο μήκος των πεπτιδίων. Παρατηρείται μια αύξηση της απόδοσης στα σύνολα δεδομένων LQ 20_K_20 και HQ 14_K_14. Ωστόσο, και σε αυτή τη περίπτωση η αύξηση αυτή δεν μπορεί να θεωρηθεί στατιστικά σημαντική.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Εν κατακλείδι, η μελέτη του μεθυλώματος αντιμετωπίζει δυσκολίες τόσο στις στρατηγικές εμπλουτισμού όσο και στη βεβαιότητα ταυτοποίησης των μεθυλιωμένων αργινίνων και λυσίνων, με αποτέλεσμα να προκύπτουν δεδομένα που προκαλούν «θόρυβο» στις υπολογιστικές προβλέψεις της μεθυλίωσης αυτών των αμινοξέων. Χρησιμοποιώντας τα πιο πρόσφατα δεδομένα μεθυλ-πρωτεωμικής, διερευνήθηκε η σημασία της ποιότητας των δεδομένων όταν αυτά χρησιμοποιούνται στις υπολογιστικές προβλέψεις με μηχανική μάθηση. Βρέθηκε ότι τα δεδομένα υψηλής ποιότητας παρουσιάζουν υψηλότερη απόδοση σε σχέση με αυτά χαμηλής ποιότητας, ακόμα και εάν είναι περισσότερα. Παράλληλα, μελετήθηκε η επίδραση της μορφής αυτών των δεδομένων στην απόδοση της πρόβλεψης. Σε σύνολα δεδομένων όπου τα χαρακτηριστικά των παραδειγμάτων τους αφορούσαν την πρωτοταγή δομή μεθυλιωμένων πεπτιδίων, και το κάθε σύνολο έφερε πεπτίδια με διαφορετικό αριθμό αμινοξέων, βρέθηκε ότι το μήκος των πεπτιδίων δεν επηρεάζει την ικανότητα πρόβλεψης, από ένα μήκος και μετά. Παράλληλα, ερευνήθηκε η απόδοση πρόβλεψης σε σύνολα δεδομένων όπου τα χαρακτηριστικά των παραδειγμάτων αφορούσαν τις φυσικοχημικές ιδιότητες των αμινοξέων αυτών των πεπτιδίων. Προκύπτει ,λοιπόν, πως η απόδοση είναι υψηλότερη όταν τα χαρακτηριστικά των συνόλων δεδομένων αφορούν την πρωτοταγή δομή των πεπτιδίων και όχι τις φυσικοχημικές ιδιότητες των αμινοξέων τους.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Audagnotto, M., & Dal Peraro, M. (2017). Protein post-translational modifications: In silico prediction tools and molecular modeling. In *Computational and Structural Biotechnology Journal* (Vol. 15, pp. 307–319). Elsevier B.V.

Beausoleil, S. A., Villén, J., Gerber, S. A., Rush, J., & Gygi, S. P. (2006). A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnology*, *24*(10), 1285–1292.

Blanc, R. S., & Richard, S. (2017). Arginine Methylation: The Coming of Age. *Molecular Cell*, *65*(1), 8–24.

Boone, C., & Adamec, J. (2016). Top-Down Proteomics. In *Proteomic Profiling and Analytical Chemistry: The Crossroads: Second Edition* (pp. 175–191). Elsevier Inc.

Chalkley, R. J., & Clauser, K. R. (2012). Modification Site Localization Scoring: Strategies and Performance. *Molecular & Cellular Proteomics*, *11*(5), 3–14.

Chen, H., Xue, Y., Huang, N., Yao, X., & Sun, Z. (2006). MeMo: a web tool for prediction of protein methylation modifications. *Nucleic Acids Research*, *34*(Web Server issue), W249-53.

Cheng, D., Côté, J., Shaaban, S., & Bedford, M. T. (2007). The arginine methyltransferase CARM1 regulates the coupling of transcription and mRNA processing. *Molecular Cell*, *25*(1), 71–83.

Couttas, T. A., Raftery, M. J., Bernardini, G., & Wilkins, M. R. (2008). Immonium Ion Scanning for the Discovery of Post-Translational Modifications and Its Application to Histones. *Journal of Proteome Research*, *7*(7), 2632–2641.

Couture, J.-F., Dirk, L. M. A., Brunzelle, J. S., Houtz, R. L., & Trievel, R. C. (2008). Structural origins for the product specificity of SET domain protein methyltransferases. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(52), 20659–20664.

di Lorenzo, A., & Bedford, M. T. (2011). Histone arginine methylation. *FEBS Letters*, *585*(13), 2024–2031.

Eldjarn, L., & Broughton, P. M. (1985). Methods of assigning accurate values to reference serum. Part 2. The use of definitive methods, reference laboratories, transferred values and consensus values. *Annals of Clinical Biochemistry*, *22 (Pt 6)*, 635–649.

Faraggi, D., & Reiser, B. (2002). Estimation of the area under the ROC curve. *Statistics in Medicine*, *21*(20), 3093–3106.

Fisk, J. C., Li, J., Wang, H., Aletta, J. M., Qu, J., & Read, L. K. (2013). Proteomic Analysis Reveals Diverse Classes of Arginine Methylproteins in Mitochondria of Trypanosomes. *Molecular & Cellular Proteomics*, *12*(2), 302–311.

Gehrig, P. M., Hunziker, P. E., Zahariev, S., & Pongor, S. (2004). Fragmentation pathways of N ^{G} -methylated and unmodified arginine residues in peptides studied by ESI-MS/MS and MALDI-MS. *Journal of the American Society for Mass Spectrometry*, *15*(2), 142–149.

Han, D., Huang, M., Wang, T., Li, Z., Chen, Y., Liu, C., Lei, Z., & Chu, X. (2019). Lysine methylation of transcription factors in cancer. In *Cell Death and Disease* (Vol. 10, Issue 4). Nature Publishing Group.

Hart-Smith, G. (2014). A review of electron-capture and electron-transfer dissociation tandem mass spectrometry in polymer chemistry. *Analytica Chimica Acta*, *808*, 44–55.

Hart-Smith, G., Yagoub, D., Tay, A. P., Pickford, R., & Wilkins, M. R. (2016). Large Scale Mass Spectrometry-based Identifications of Enzyme-mediated Protein Methylation Are Subject to High False Discovery Rates. *Molecular & Cellular Proteomics*, *15*(3), 989–1006.

Hripcsak, G. (2005). Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*, *12*(3), 296–298.

Hubbard, R. E., & Kamran Haider, M. (2010). Hydrogen Bonds in Proteins: Role and Strength. In *eLS*. Wiley.

Krüger, D. M., Neubacher, S., & Grossmann, T. N. (2018). Protein-RNA interactions: structural characteristics and hotspot amino acids. *RNA (New York, N.Y.)*, *24*(11), 1457–1465.

Lanouette, S., Mongeon, V., Figeys, D., & Couture, J. F. (2014). The functional diversity of protein lysine methylation. In *Molecular Systems Biology* (Vol. 10, Issue 4). Blackwell Publishing Ltd.

Lee, T.-Y., Chang, C.-W., Lu, C.-T., Cheng, T.-H., & Chang, T.-H. (2014). Identification and characterization of lysine-methylated sites on histones and non-histone proteins. *Computational Biology and Chemistry*, *50*, 11–18.

Lee, Y.-H., & Stallcup, M. R. (2009). Minireview: Protein Arginine Methylation of Nonhistone Proteins in Transcriptional Regulation. *Molecular Endocrinology*, *23*(4), 425–433.

Likas, A., Blekas, K., & Kalles, D. (Eds.). (2014). *Artificial Intelligence: Methods and Applications* (Vol. 8445). Springer International Publishing.

Lorton, B. M., & Shechter, D. (2019). Cellular consequences of arginine methylation. *Cellular and Molecular Life Sciences : CMLS*, *76*(15), 2933–2956.

Luo, M. (2015). Current Methods for Methylome Profiling. In *Epigenetic Technological Applications* (pp. 187–217). Elsevier Inc.

Luo, M. (2018). Chemical and Biochemical Perspectives of Protein Lysine Methylation. *Chemical Reviews*, *118*(14), 6656–6705.

Marx, H., Lemeer, S., Schliep, J. E., Matheron, L., Mohammed, S., Cox, J., Mann, M., Heck, A. J. R., & Kuster, B. (2013). A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nature Biotechnology*, *31*(6), 557–564.

Morera, L., Lübbert, M., & Jung, M. (2016). Targeting histone methyltransferases and demethylases in clinical trials for cancer therapy. *Clinical Epigenetics*, *8*, 57.

Murn, J., & Shi, Y. (2017). The winding path of protein methylation research: milestones and new frontiers. *Nature Reviews. Molecular Cell Biology*, *18*(8), 517–527.

Ong, S.-E., & Mann, M. (2006). A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nature Protocols*, *1*(6), 2650–2660.

Ong, S.-E., Mittler, G., & Mann, M. (2004). Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. *Nature Methods*, *1*(2), 119–126.

Petersson, A. S., Steen, H., Kalume, D. E., Caidahl, K., & Roepstorff, P. (2001). Investigation of tyrosine nitration in proteins by mass spectrometry. *Journal of Mass Spectrometry : JMS*, *36*(6), 616–625.

Petrossian, T. C., & Clarke, S. G. (2011). Uncovering the Human Methyltransferasome. *Molecular & Cellular Proteomics*, *10*(1), M110.000976.

Rappsilber, J., Friesen, W. J., Paushkin, S., Dreyfuss, G., & Mann, M. (2003). Detection of Arginine Dimethylated Peptides by Parallel Precursor Ion Scanning Mass Spectrometry in Positive Ion Mode. *Analytical Chemistry*, *75*(13), 3107–3114.

Shao, J., Xu, D., Tsai, S.-N., Wang, Y., & Ngai, S.-M. (2009). Computational Identification of Protein Methylation Sites through Bi-Profile Bayes Feature Extraction. *PLoS ONE*, *4*(3), e4920.

Shi, S.-P., Qiu, J.-D., Sun, X.-Y., Suo, S.-B., Huang, S.-Y., & Liang, R.-P. (2012). PMeS: Prediction of Methylation Sites Based on Enhanced Feature Encoding Scheme. *PLoS ONE*, *7*(6), e38772.

Sun, F., & Wu, R. (2019). *Systematic and site-specific analysis of N-glycoproteins on the cell surface by integrating bioorthogonal chemistry and MS-based proteomics* (pp. 223–247).

Thinnes, C. C., England, K. S., Kawamura, A., Chowdhury, R., Schofield, C. J., & Hopkinson, R. J. (2014). Targeting histone lysine demethylases - progress, challenges, and the future. *Biochimica et Biophysica Acta*, *1839*(12), 1416–1432.

Tian, H., Sun, Y., Liu, C., Duan, X., Tang, W., & Li, Z. (2016). Precise Quantitation of MicroRNA in a Single Cell with Droplet Digital PCR Based on Ligation Reaction. *Analytical Chemistry*, *88*(23), 11384–11389.

Tsai, Y.-J., Pan, H., Hung, C.-M., Hou, P.-T., Li, Y.-C., Lee, Y.-J., Shen, Y.-T., Wu, T.-T., & Li, C. (2011). The predominant protein arginine methyltransferase PRMT1 is critical for zebrafish convergence and extension during gastrulation. *The FEBS Journal*, *278*(6), 905–917.

Tyanova, S., Temu, T., & Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols*, *11*(12), 2301–2319.

Uversky, V. N. (2013). Posttranslational Modification. In *Brenner's Encyclopedia of Genetics: Second Edition* (pp. 425–430). Elsevier Inc.

Wang, H., Straubinger, R. M., Aletta, J. M., Cao, J., Duan, X., Yu, H., & Qu, J. (2009). Accurate localization and relative quantification of arginine methylation using nanoflow liquid chromatography coupled to electron transfer dissociation and drbitrap mass spectrometry. *Journal of the American Society for Mass Spectrometry*, *20*(3), 507–519.

Wang, Q., Wang, K., & Ye, M. (2017). Strategies for large-scale analysis of non-histone protein methylation by LC-MS/MS. In *Analyst* (Vol. 142, Issue 19, pp. 3536–3548). Royal Society of Chemistry.

Wei, H. H., Fan, X. J., Hu, Y., Tian, X. X., Guo, M., Mao, M. W., Fang, Z. Y., Wu, P., Gao, S. X., Peng, C., Yang, Y., & Wang, Z. (2021). A systematic survey of PRMT interactomes reveals the key roles of arginine methylation in the global control of RNA splicing and translation. *Science Bulletin*, *66*(13), 1342–1357.

Wei, L., Xing, P., Shi, G., Ji, Z., & Zou, Q. (n.d.). Fast Prediction of Protein Methylation Sites Using a Sequence-Based Feature Selection Technique. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *16*(4), 1264–1273.

Wells, J. M., & McLuckey, S. A. (2005). Collision-induced dissociation (CID) of peptides and proteins. *Methods in Enzymology*, *402*, 148–185.

Wen, P.-P., Shi, S.-P., Xu, H.-D., Wang, L.-N., & Qiu, J.-D. (2016). Accurate *in silico* prediction of species-specific methylation sites based on information gain feature optimization. *Bioinformatics*, *32*(20), 3107–3115.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data Mining Practical Machine Learning Tools and Techniques Fourth Edition*. https://www.elsevier.com

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*(4), 561–577.