Πανεπιστήμιο Θεσσαλίας
Σχολή Επιστημών Υγείας
Τμήμα Βιοχημείας και Βιοτεχνολογίας

Πρόβλεψη θέσεων Ν-γλυκοζυλίωσης με μεθόδους μηχανικής εκμάθησης

Πατεράκη Γεωργία

Επιβλέπων καθηγητής: Αμούτζιας Γρηγόριος

Λάρισα 2019

University of Thessaly
School of Health Sciences
Department of Biochemistry and Biotechnology

Prediction of N-glycosylation sites with machine learning methods

Pateraki Georgia

Supervisor: Amoutzias Grigorios

Larissa 2019

Η παρούσα πτυχιακή εργασία με θέμα «Πρόβλεψη θέσεων N-γλυκοζυλίωσης με μεθόδους μηχανικής εκμάθησης» εκπονήθηκε στο εργαστήριο Βιοπληροφορικής, του τμήματος Βιοχημείας και Βιοτεχνολογίας, του Πανεπιστημίου Θεσσαλίας.

Τριμελής Επιτροπή:

Επιβλέπων καθηγητής: Αμούτζιας Γρηγόριος, Αναπληρωτής Καθηγητής Βιοπληροφορικής με έμφαση στη Μικροβιολογία, Τμήμα Βιοχημείας και Βιοτεχνολογίας, Πανεπιστήμιο Θεσσαλίας

Μόσιαλος Δημήτριος, Επίκουρος Καθηγητής Βιοτεχνολογίας Μικροβίων, Τμήμα Βιοχημείας και Βιοτεχνολογίας, Πανεπιστήμιο Θεσσαλίας

Ηλιόπουλος Ιωάννης, Επίκουρος Καθηγητής Μοριακής Βιολογίας-Γονιδιωματικής Βιοπληροφορικής, Τμήμα Ιατρικής, Πανεπιστήμιο Κρήτης

Περιεχόμενα

Περιεχόμενα εικόνων

Περιεχόμενα πινάκων

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, Αμούτζια Γρηγόριο, Αναπληρωτή Καθηγητή Βιοπληροφορικής με έμφαση στη Μικροβιολογία, του Τμήματος Βιοχημείας Βιοτεχνολογίας, του Πανεπιστημίου Θεσσαλίας, για την πολύτιμη βοήθεια και καθοδήγησή του καθ΄όλη τη διάρκεια της εκπόνησης της πτυχιακής μου εργασίας. Ευχαριστώ επίσης τα μέλη της τριμελούς επιτροπής κκ. Μόσιαλο και Ηλιόπουλο για την βοήθεια τους.

Επιπλέον, θα ήθελα να ευχαριστήσω τον προπτυχιακό συμφοιτητή μου Νικολαίδη Μάριο για την αμέριστη βοήθεια του.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου και τους φίλους μου για την στήριξή τους όλα αυτά τα χρόνια.

Περίληψη

Η γλυκοζυλίωση είναι μία συμμεταφραστική και μετα-μεταφραστική τροποποίηση η οποία ρυθμίζει την πρωτεϊνική αναδίπλωση, στόχευση και αλληλεπιδράσεις της με άλλα βιολογικά μόρια. Ως εκ τούτου, συμμετέχει σε ποικίλλες βιολογικές διεργασίες μεταξύ των οποίων η κυτταρική προσκόλληση και η κυτταρική επικοινωνία. Τα τελευταία χρόνια με την πρωτεωμική μεγάλης κλίμακας έχουν εντοπιστεί πολλά πεπτιδικά κατάλοιπα τα οποία υπόκεινται στη συγκεκριμένη τροποποίηση. Ωστόσο, απέχουμε ακόμη πολύ από την κάλυψη ολόκληρων των πρωτεωμάτων των ευκαρυωτικών ή μη οργανισμών. Οι υπολογιστικές προσεγγίσεις παρέχουν έναν γρήγορο και ακριβή τρόπο πρόβλεψης τέτοιων τροποποιήσεων. Στην παρούσα εργασία με τη συλλογή N-γλυκοζυλιωμένων πεπτιδίων και την εφαρμογή αυστηρών κριτηρίων φιλτραρίσματος επιτυγχάνεται η κατασκευή μοντέλων πρόβλεψης θέσεων N-γλυκοζυλίωσης, μέσα στα πλαίσια της σουίτας ελεύθερου λογισμικού Weka, με ποσοστό επιτυχούς πρόβλεψης μέχρι και 92.018%.

Abstract

Glycosylation is a co-translational and posttranslational modification that regulates protein folding, targeting and interaction. As such, it is related to various biological processes including cellular proliferation and communication. In recent years, with large-scale proteomics many peptide residues that are subject to this modification, have been identified. However, we are still far from covering entire proteomes of eukaryotes. Computational approaches provide a fast and accurate way of predicting such modifications. In the present work, by collecting glycosylated peptides and applying strict filtering criteria, the prediction of glycosylation site prediction models within the machine learning software, Weka, is achieved, with a n accuracy of up to 92.018%.

1 Εισαγωγή

1.1 Γονιδιακή ρύθμιση

Γονιδιακή ρύθμιση ονομάζεται η διαδικασία κατά την οποία πραγματοποιείται έλεγχος της έκφρασης των γονιδίων καθώς επίσης και των επιπέδων και λειτουργίας των προϊόντων τους, ως απόκριση σε εσωτερικά και εξωτερικά ερεθίσματα. Η ρύθμιση της γονιδιακής έκφρασης περιλαμβάνει ένα ευρύ φάσμα μηχανισμών σε πολλά επίπεδα.

1.1.1 Μεταγραφική ρύθμιση γονιδιακής έκφρασης

Σε μεταγραφικό επίπεδο περιλαμβάνονται ρυθμιστικά συστήματα που ελέγχουν την έναρξη και την ολοκλήρωση της μεταγραφής.

- ‘Εναρξη της μεταγραφής:
  Αναδιαμόρφωση χρωματίνης: η βασική μονάδα οργάνωσης της χρωματίνης είναι το νουκλεόσωμα το οποίο αποτελείται από DNA και πρωτεΐνες (ιστόνες). Η περιέλιξη του γενετικού υλικού με τη δομή νουκλεοσωμάτων έχει κατασταλτική επίδραση στη γονιδιακή έκφραση γιατί παρεμποδίζει την πρόσδεση μεταγραφικών παραγόντων. Ωστόσο, υπάρχουν ομάδες ενζύμων που τροποποιούν ομοιοπολικά τις ιστόνες με αποτέλεσμα την διευκόλυνση πρόσβασης των μεταγραφικών παραγόντων. Παράδειγμα αυτού του είδους των τροποποιήσεων αποτελεί η ακετυλίωση, μία τροποποίηση που διεκπεραιώνεται από τις ακετυλοτρανσφεράσες.
  Ρυθμιστικά στοιχεία: Η έναρξη της μεταγραφής διεκπεραιώνεται με την προσέλκυση του συμπλόκου έναρξης της μεταγραφής. Αυτό ελέγχεται από την αλληλεπίδραση των ρυθμιστικών στοιχείων των γονιδίων, δηλαδή του υποκινητή και των ενισχυτών με μία ειδική ομάδα πρωτεϊνών, τους μεταγραφικούς παράγοντες. Οι μεταγραφικοί παράγοντες προσδένονται στα παραπάνω ρυθμιστικά στοιχεία και δρουν είτε θετικά (επαγωγή έκφρασης γονιδίου) προσελκύοντας το σύμπλοκο έναρξης της μεταγραφής, είτε αρνητικά (καταστολή έκφρασης γονιδίου), παρεμποδίζοντας την πρόσδεση του συμπλόκου έναρξης της μεταγραφής στον υποκινητή του γονιδίου.
- Τερματισμός μεταγραφής:
  Η μεταγραφή ολοκληρώνεται μέσω ειδικών αλληλουχιών (terminators) που οδηγούν στην απελευθέρωση της πολυμεράσης.

1.1.2 Μετα-μεταγραφική ρύθμιση της γονιδιακής έκφρασης

- Επεξεργασία πρόδρομου mRNA (pre-mRNA):
  Προσθήκη μεθυλιωμένης κεφαλής και πολυαδενυλιωμένης ουράς: Το RNA που παράγεται από τη διαδικασία της μεταγραφής είναι ασταθές γεγονός που οδηγεί στην προσθήκη μιας καλύπτρας 7-μεθυλογουανοσίνης στο 5’-άκρο και την προσθήκη μιας πολυαδενυλιωμένης ουρά στο 3’-άκρο. Η προσθήκη της καλύπτρας χρησιμεύει, μεταξύ άλλων, ως προστασία έναντι στη δράση των 5’ 3’

εξωνουκλεασών (Ramanathan et al., 2016), ενώ η προσθήκη της πολυαδενυλιωμένης ουράς παρεμποδίζει την 3'→ 5' εξωνουκλεολυτική αποικοδόμηση (Ford et al., 1997).

Συρραφή: Για την παραγωγή του ώριμου mRNA από το πρόδρομο mRNA (pre-mRNA) απαιτείται η συρραφή, δηλαδή η απομάκρυνση των ιντρονίων και η συνένωση των εξωνίων. Με τη διαδικασία του εναλλακτικού ματίσματος ενδέχεται να προκύψουν περισσότερα του ενός, μόρια mRNA από ένα πρόδρομο RNA (εναλλακτικό μάτισμα).

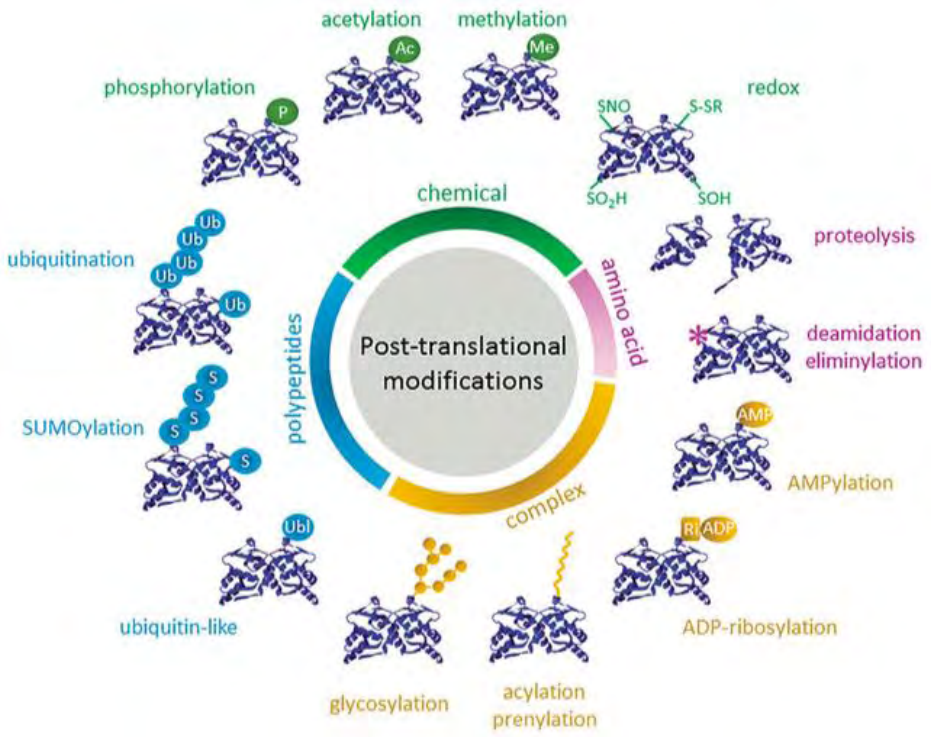
Μη κωδικοποιητικά RNAs (noncoding RNAs, ncRNAs): είναι τα RNAs τα οποία δεν κωδικοποιούν για κάποια πρωτεΐνη και διακρίνονται σε μικρά μη κωδικοποιητικά RNAs (small non-coding RNAs, sncRNAs) και μακρά μη κωδικοποιητικά RNAs (long non-coding RNAs, lncRNAs). Στα sncRNAs, η πιο καλά μελετημένη κατηγορία είναι τα micro-RNAs (miRNAs) τα οποία συμμετέχουν στη γονιδιακή ρύθμιση μέσω αποικοδόμησης των mRNAs στα οποία υβριδίζουν, ή της καταστολής της μετάφρασης τους (Catalanotto et al., 2016). Οι δράσεις των lncRNAs σε μετα-μεταγραφικό επίπεδο ποικίλλουν και περιλαμβάνουν είτε την επαγωγή, είτε την καταστολή της συρραφής καθώς επίσης και της μετάφρασης των mRNA στόχων (Yoon et al., 2013).

- Αποικοδόμηση:
  Ορισμένα από τα παραγόμενα, ώριμα mRNAs οδηγούνται στην αποικοδόμηση. Η ρύθμιση της αποικοδόμησης των mRNAs καθιστά ικανό το κύτταρο να ελέγχει τα επίπεδα σύνθεσης των πρωτεϊνών ανάλογα με τις ανάγκες του (McManus et al., 2015). Πρόκειται για μία διαδικασία η οποία ξεκινά με τη μείωση της poly-A ουράς και έπειτα είτε την αποικοδόμηση του μορίου με 3' → 5' κατεύθυνση, είτε την απομάκρυνση της μεθυλιωμένης κεφαλής και την αποικοδόμηση του μορίου με 5'→3' κατεύθυνση (Parker and Song, 2004).

1.1.3 Μετα-μεταφραστική ρύθμιση της γονιδιακής έκφρασης

Η γονιδιακή ρύθμιση σε μετα-μεταφραστικό επίπεδο περιλαμβάνει ένα ευρύ φάσμα διεργασιών που καθορίζουν την τύχη του πρωτεϊνικού μορίου. Οι διαδικασίες αυτές ελέγχουν τόσο τη διάρκεια ζωής μια πρωτεΐνης, όσο και τη δράση της:

- Υποκυτταρικός εντοπισμός:
  Η μεταφορά των πρωτεϊνών σε κάποιο υποκυτταρικό οργανίδιο ρυθμίζεται από αλληλουχίες των αμινοτελικών τους άκρων καθορίζοντας έτσι και την περιοχή δράσης τους.
- Μετα-μεταφραστικές τροποποιήσεις:
  Είναι οι τροποποιήσεις που πραγματοποιούνται στις πρωτεΐνες στο στάδιο μετά τη σύνθεσή τους και οδηγούν στην αλλαγή της δομής και της στερεοδιάταξής τους. Είναι ζωτικής σημασίας καθώς επηρεάζουν τη σταθερότητα των πρωτεϊνών, τη δραστηριότητα τους και τις αλληλεπιδράσεις τους με άλλα μόρια. Περιλαμβάνουν ένα ευρύ φάσμα τροποποιήσεων μεταξύ των οποίων η ομοιοπολική προσθήκη μιας χημικής ομάδας (φωσφορυλίωση, ακετυλίωση, μεθυλίωση), πολυπεπτιδίων (ουβικουϊτίνωση), ή πολυπλοκότερων μορίων (γλυκοζυλίωση, πρενοϋλίωση) (**Εικόνα 1**).



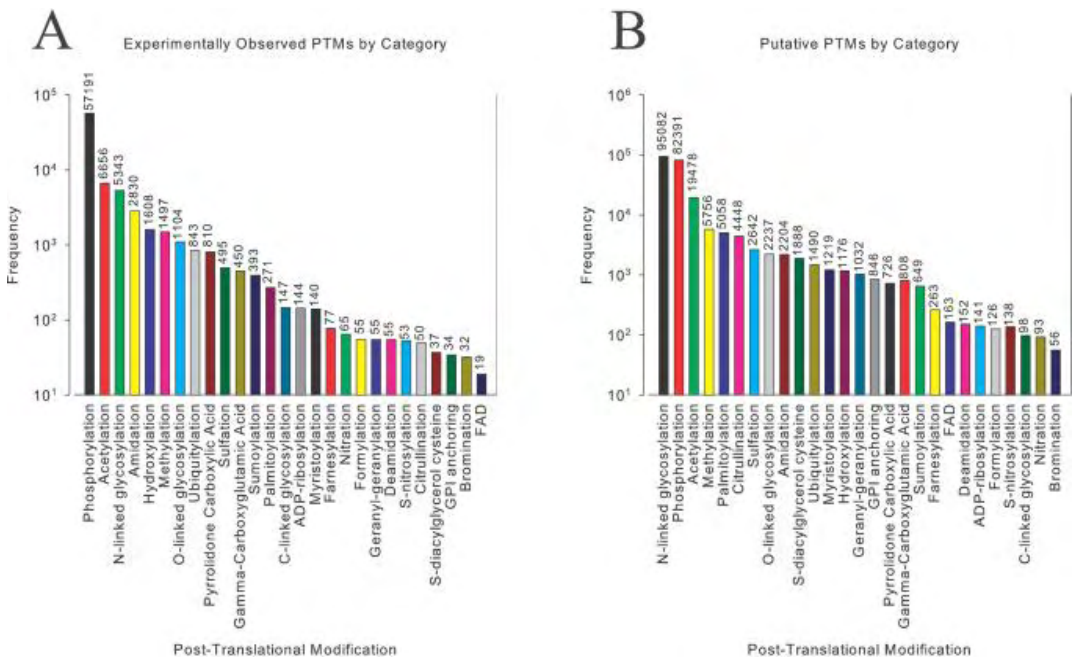
Εικόνα 1 *Ενα υποσύνολο μετα-μεταφραστικών τροποποιήσεων διαχωρισμένες με βάση το είδος της τροποποίησης. Οι χημικές τροποποιήσεις είναι αναστρέψιμες και περιλαμβάνουν τη φωσφορυλίωση (P), την ακετυλίωση (Ac), τη μεθυλίωση (Me) και τις τροποποιήσεις βασιζόμενες στη οξειδοαναγωγή (SNO, S-S, SOH, SO_2H). Οι τροποποιήσεις που περιλαμβάνουν την προσθήκη πολυπεπτιδίων είναι επίσης ενζυμικά αναστρέψιμες και περιλαμβάνουν την ουβικουιτίνωση, τη σουμοϋλίωση (S). Η τροποποίηση με την προσθήκη πιο πολύπλοκων ομάδων είναι ενζυμικά αναστρέψιμες και περιλαμβάνουν τη γλυκοζυλίωση, την προσθήκη λιπιδίων (ακυλίωση, πρενοϋλίωση), την ADP-ριβοσυλίωση (Ri-ADP), την αδενυλίωση (AMP). Τέλος, κάποιες τροποποιήσεις των αμινοξέων (αστερίσκος) ή του πολυπεπτιδικού σκελετού είναι μη αναστρέψιμες και περιλαμβάνουν την απαμίνωση* **(Spoel, 2018)**

1.2 Γλυκοζυλίωση

Η γλυκοζυλίωση αποτελεί μία από τις συνηθέστερες μετα-μεταφραστικές τροποποιήσεις των πρωτεϊνών (Khoury et al., 2011) (**Εικόνα *2***).

Με τον όρο γλυκοζυλίωση περιγράφεται η πρόσδεση μεγάλων υδατανθρακικών ομάδων σε πολυπεπτίδια. Υπάρχουν 2 γενικοί τύποι γλυκοζυλίωσης, η Ν-συνδεδεμένη γλυκοζυλίωση και η Ο-συνδεδεμένη γλυκοζυλίωση.

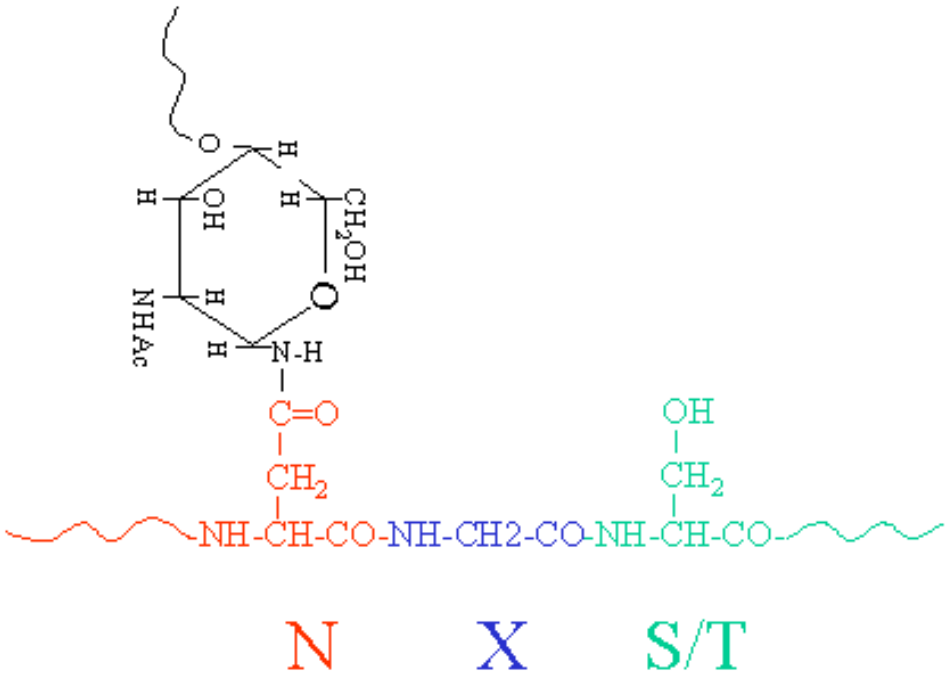
- Ο-συνδεδεμένη γλυκοζυλίωση: πρόσδεση μιας αλυσίδας σακχάρου μέσω της υδροξυλομάδας μιας σερίνης ή θρεονίνης
- Ν-συνδεδεμένη γλυκοζυλίωση: πρόσδεση μιας αλυσίδας σακχάρου μέσω της αμινομάδας της πλευρικής αλυσίδας της ασπαραγίνης.



Εικόνα 2 *Οι συχνότερες μετα-μεταφραστικές τροποποιήσεις που έχουν εντοπιστεί πειραματικά (A) και υποθετικά (B) με δεδομένα από τη Swiss-Prot* **(Khoury et al., 2011)**

1.2.1 N-συνδεδεμένη γλυκοζυλίωση

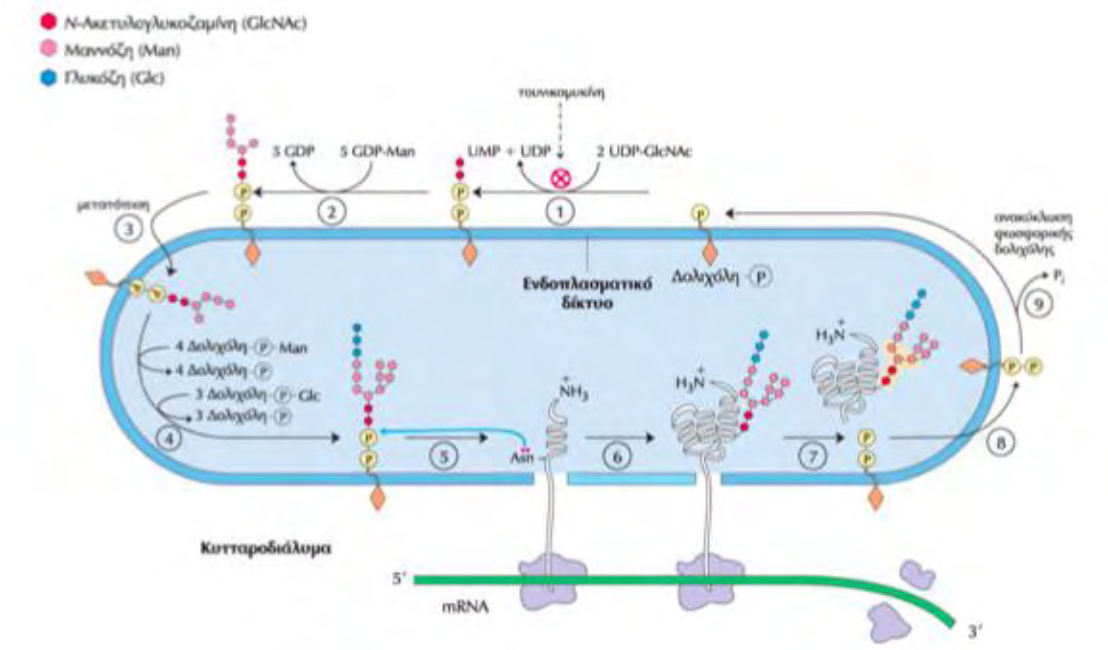
Στη N-συνδεδεμένη γλυκοζυλίωση η προσθήκη των υδατανθρακικών ομάδων γίνεται στο αμιδικό άτομο αζώτου του αμινοξέος ασπαραγίνη (N), όταν αυτό αποτελεί μέρος της ακολουθίας NXS/T (όπου X οποιοδήποτε αμινοξύ εκτός της προλίνης) (**Εικόνα 3**).



Εικόνα 3 *Απεικόνιση της πρόσδεσης του αμινοσακχάρου N-ακετυλογλυκοζαμίνη στο αμιδικό άτομο αζώτου της ασπαραγίνης, όταν αυτή αποτελεί μέρος της ακολουθίας NXS/T* (πηγή: https://www.ionsource.com/Card/carbo/nolink.htm)

Η διαδικασία της Ν-συνδεδεμένης γλυκοζυλίωσης αποτελεί μία συμμεταφραστική τροποποίηση καθώς ξεκινάει στην κυτταροπλασματική πλευρά του Ε.Δ., με την προσθήκη 2 καταλοίπων Ν-ακετυλογλυκοζαμίνης (GlcNAc) και 3 καταλοίπων μαννόζης, σε μία δολιχόλη της λιπιδικής διπλοστοιβάδας. Η νεοσχηματιζόμενη γλυκάνη μετακινείται κατά μήκος της μεμβράνης και γίνεται μία αλλαγή προσανατολισμού με αποτέλεσμα να μην είναι προσβάσιμη από τα ένζυμα του κυτταροπλάσματος, καθώς έχει πλέον στραφεί προς το εσωτερικό του Ε.Δ. Επειτα γίνεται περαιτέρω προσθήκη σακχαρικών καταλοίπων και ο κεντρικός ολιγοσακχαρίτης μεταφέρεται μέσω του ενζύμου ολιγοσακχαρυλο-τρανσφεράση, σε ένα κατάλοιπο Asn (**Εικόνα *4***). Ακολουθεί επεξεργασία του ολιγοσακχαρίτη, αφαίρεση και προσθήκη σακχαρικών καταλοίπων από ειδικά ένζυμα και μεταφορά στο Golgi όπου υφίστανται περαιτέρω επεξεργασία.

Με την προσθήκη των γλυκανών αφ'ενός υποβοηθείται άμεσα το πακετάρισμα των πρωτεινών καθώς επάγει τη σταθεροποίηση των πολυπεπτιδίων, αφ'ετέρου βοηθάει και έμμεσα εφ'όσον οι γλυκάνες χρησιμεύουν ως «ετικέτες» αναγνώρισης που τους επιτρέπουν να αλληλεπιδρούν με διάφορα ένζυμα όπως οι λεκτίνες, οι γλυκοσιδάσες και οι γλυκοζυλοτρανσφεράσες (Price et al., 2012). Ορισμένες από αυτές, όπως οι γλυκοσιδάσες Ι και ΙΙ, η καλνεξίνη και η καλρετικουλίνη παίζουν κεντρικό ρόλο στο πακετάρισμα, ενώ άλλες, όπως οι α-μαννοσιδάσες, εξυπηρετούν στην αποδόμηση σχετιζόμενη με το Ε.Δ. (Helenius, Aebi, 2004).

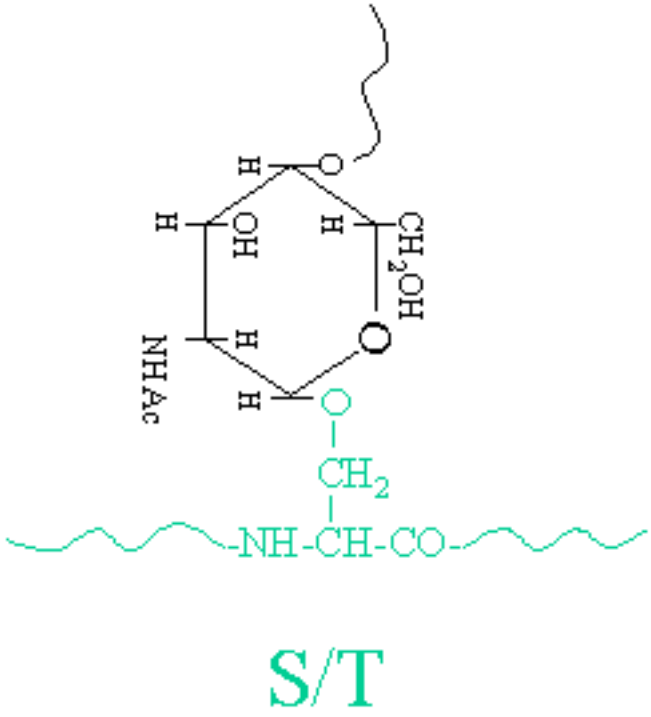


Εικόνα 4 *Σύνθεση του κεντρικού ολιγοσακχαρίτη των πρωτεϊνών 1,2: Τα αρχικά βήματα πραγματοποιούνται στην πλευρά του ενδοπλασματικού δικτύου προς το κυτταρόπλασμα 3: Ο ημιτελής ολιγοσακχαρίτης μετατοπίζεται διαμέσου της μεμβράνης και 4: η σύνθεση του κεντρικού ολιγοσακχαρίτη ολοκληρώνεται εντός του αυλού του ενδοπλασματικού δικτύου. Οι πρόδρομοι που συνεισφέρουν επιπρόσθετα κατάλοιπα μαννόζης και γλυκόζης στον αυξανόμενο ολιγοσακχαρίτη, στον αυλό του ενδοπλασματικού δικτύου, είναι παράγωγα της φωσφορικής δολιχόλης. Κατά την εκκίνηση της παρασκευής της Ν-συνδεδεμένης ολιγοκχαριτικής αλυσίδας 5,6: πραγματοποιείται μεταφορά του κεντρικού ολιγοσακχαρίτη από τη φωσφορική δολιχόλη σ'ένα κατάλοιπο ασπαραγίνης (Asn) της πρωτεΐνης μέσα στον αυλό*

του ενδοπλασματικού δικτύου. Εν συνεχεία, ο κεντρικός ολιγοσκχαρίτης τροποποιείται περαιτέρω στο ενδοπλασματικό δίκτυο και στη συσκευή Golgi σε οδούς οι οποίες διαφέρουν ανάλογα με την πρωτεΐνη. Τα πέντε υδατανθρακικά κατάλοιπα που περιλαμβάνονται στο μπεζ πλαίσιο (μετά το βήμα 7) διατηρούνται στην οριστική δομή όλων των N-συνδεδεμένων ολιγοσακχαριτών. 8: Η εκλυόμενη πυροφωσφορική δολιχόλη μετατοπίζεται ξανά έτσι ώστε η πυροφωσφορική ομάδα να βρίσκεται στην κυτταροπλασματική πλευρά της μεμβράνης του ενδοπλασματικού δικτύου. 9: Τέλος, η φωσφορική ομάδα αφαιρείται με υδρόλυση, αναγεννώντας τη φωσφορική δολιχόλη **(Cox and Nelson, 2000)**

1.2.2 Ο-συνδεδεμένη γλυκοζυλίωση

Στην Ο-συνδεδεμένη γλυκοζυλίωση η προσθήκη των υδατανθρακικών ομάδων γίνεται στο οξυγόνο της υδροξυλομάδας των αμινοξέων σερίνη (S) και θρεονίνη (T) (**Εικόνα *5***).



Εικόνα 5 *Απεικόνιση της πρόσδεσης του αμινοσακχάρου N-ακετυλογλυκοζαμίνη στο οξυγόνο της υδροξυλομάδας των αμινοξέων σερίνη ή θρεονίνη* (πηγή: **https://www.ionsource.com/Card/carbo/nolink.htm**)

Η Ο-συνδεδεμένη γλυκοζυλίωση, λαμβάνει χώρα αφού η πρωτεΐνη φτάσει στο σύμπλεγμα Golgi.
Υπάρχουν διάφορες κατηγορίες γλυκανών που συναντώνται στις Ο-γλυκοζυλιωμένες πρωτείνες:

- Σάκχαρα τύπου μουκινών (mucin-type O-glycans)

Κατά την εκκίνηση του σχηματισμού της γλυκάνης πραγματοποιείται προσθήκη, συνήθως, ενός μορίου N-ακετυλογαλακτοζαμίνης (GalNAc) στην πλευρική αλυσίδα ενός καταλοίπου αμινοξέος Ser ή Thr από το ένζυμο N-ακετυλογαλακτοζαμινοτρανσφεράση (GalNActransferase). Ακολουθεί περαιτέρω προσθήκη σακχαρικών καταλοίπων μέσω των ενζύμων γλυκοζυλοτρανσφεράσες (Steen et al., 1998).

- Ο-συνδεδεμένη N-ακετυλογλυκοζαμίνη (O-GlcNAc)

  Συναντώνται μόνο σε θηλαστικά και κατά κύριο λόγο σε κυτταροπλασματικές και πυρηνικές πρωτείνες. Την προσθήκη και την

αφαίρεση του ολιγοσακχαρίτη καταλύουν τα ένζυμα O-GlcNAc transferase (OGT) και O- GlcNAcase (OGA), αντίστοιχα (Woo et al., 2018). Η παραπάνω τροποποίηση έχει πιθανολογηθεί ότι παίζει ρόλο σε σηματοδοτικούς «καταρράκτες» (Wells, Vosseller, Hart, 2001), ενώ έχει επιπλέον συσχετιστεί με την ενεργοποίηση των Τ-κυττάρων (Woo et al., 2018).

- Αλλες γλυκάνες όχι τύπου μουκινών περιλαμβάνουν τις O-φουκόζη, O-ξυλόζη, O-μαννόζη και O-γαλακτόζη.

1.2.3 Παράγωγα γλυκοζυλίωσης και βιολογικός ρόλος

Μία ομάδα υδατανθράκων μπορεί να προσδεθεί ομοιοπολικά σε μία πρωτεΐνη για να σχηματίσει μια γλυκοπρωτείνη. Υπάρχουν 3 τάξεις γλυκοπρωτεϊνών, οι απλές γλυκοπρωτεΐνες, οι πρωτεογλυκάνες και οι βλεννίνες (ή βλεννοπρωτεΐνες).

1.2.3.1 Γλυκοπρωτεΐνες

Η πρωτεϊνική συνιστώσα είναι το μεγαλύτερο συστατικό κατά βάρος. Αυτή η πολυδύναμη τάξη έχει διάφορους βιοχημικούς ρόλους.
Πολλές γλυκοπρωτείνες αποτελούν συστατικά των κυτταρικών μεμβρανών, όπου λαμβάνουν μέρος σε διεργασίες όπως η κυτταρική προσκόλληση και η πρόσδεση του σπερματοζωαρίου στο ωάριο. Αλλες γλυκοπρωτεΐνες σχηματίζονται με τη σύνδεση υδατανθράκων σε διαλυτές πρωτεΐνες. Ιδιαίτερα, πολλές από τις πρωτεΐνες που εκκρίνονται από τα κύτταρα είναι γλυκοζυλιωμένες ή τροποποιημένες από την πρόσδεση υδατανθράκων.

1.2.3.2 Πρωτεογλυκάνες

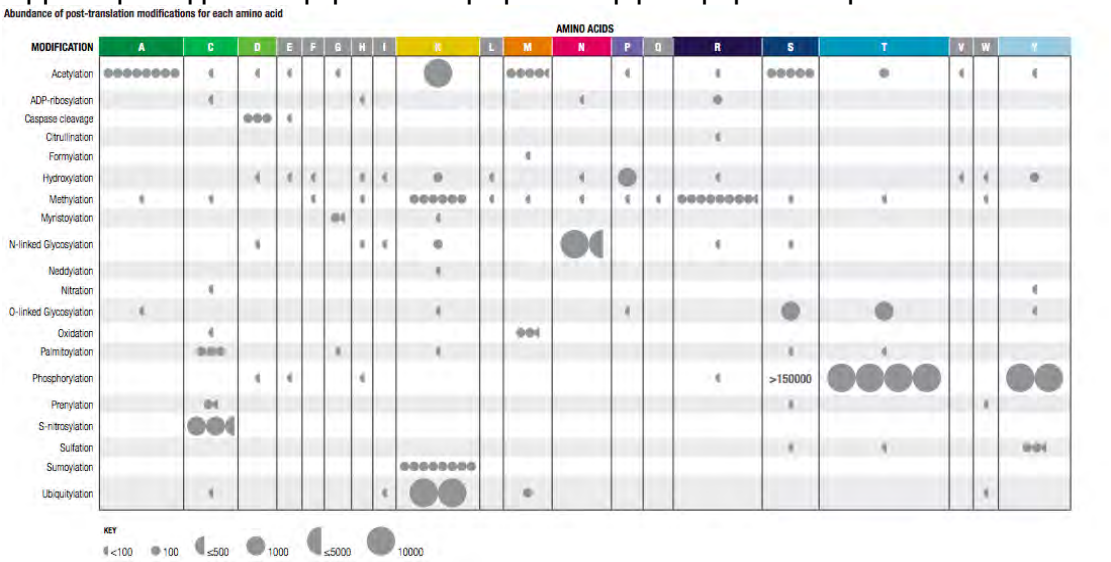
Το πρωτεϊνικό συστατικό των πρωτεογλυκανών είναι συνδεδεμένο με έναν ειδικό τύπο πολυσακχαρίτη που ονομάζεται γλυκοζαμινογλυκάνη (GAG). Οι υδατάνθρακες αποτελούν το μεγαλύτερο ποσοστό του βάρους των πρωτεογλυκανών. Οι πρωτεογλυκάνες αποτελούν ένα από τα σημαντικότερα συστατικά της εξωκυττάριας ουσίας και συμμετέχουν στην κυτταρική προσκόλληση, μετανάστευση και πολλαπλασιασμό (Wight et al., 1992).

1.2.3.3 Βλεννίνες

Οι βλεννίνες ή βλεννοπρωτείνες αποτελούνται, όπως οι πρωτεογλυκάνες, κυρίως από υδατάνθρακες. Η N-ακετυλογαλακτοζαμίνη (GalcNAc) είναι συνήθως η υδατανθρακική ομάδα που είναι προσδεδεμένη στην πρωτεΐνη στις βλεννίνες. Ενα ποσοστό ίσο περίπου με 80%, της σύστασης, είναι υδατάνθρακες και το υπόλοιπο 20% πρωτεΐνη. Διακρίνονται σε δύο υποκατηγορίες, τις εκκρινόμενες μουκίνες και τις διαμεμβρανικές μουκίνες, εκ των οποίων οι πρώτες συμμετέχουν στην άμυνα του οργανισμού έναντι σε παθογόνα και οι τελευταίες σε μονοπάτια μεταγωγής σήματος (Dhanisha et al., 2018).

1.2.4 Πιθανές και υπάρχουσες αλληλεπιδράσεις γλυκοζυλίωσης με άλλες μετα-μεταφραστικές τροποποιήσεις

Μέσα από διάφορες μελέτες έχει διαπιστωθεί ότι κάθε αμινοξικό κατάλοιπο υπόκειται σε διάφορες μετα-μεταφραστικές τροποποιήσεις, το είδος και η έκταση των οποίων, διαφέρει ανάλογα με το αμινοξύ (**Εικόνα 6**). Η τροποποίηση ενός αμινοξικού καταλοίπου ενδέχεται να έχει θετική ή αρνητική επίδραση στην τροποποίηση ενός άλλου αλλά και στο ίδιο. Αλληλεπιδράσεις αναμένουμε να εμφανιστούν και στην τροποποίηση της γλυκοζυλίωσης και πολύ έντονη αλληλεπίδραση έχει παρατηρηθεί ανάμεσα σε αυτού του είδους τη μετα-μεταφραστική τροποποίηση και στη φωσφορυλίωση.



Εικόνα 6 *Οι συχνότερες μετα-μεταφραστικές τροποποιήσεις για κάθε αμινοξύ* (πηγή: https://media.cellsignal.com/www/pdfs/content-fragments/gu-nt2-amino-acid-poster.pdf)

1.2.4.1 Αλληλεπιδράσεις γλυκοζυλίωσης-φωσφορυλίωσης

Εκτεταμένες αλληλεπιδράσεις έχουν παρατηρηθεί ανάμεσα στη φωσφορυλίωση και στην Ο-συνδεδεμένη γλυκοζυλίωση καθώς στοχεύουν κατά κύριο λόγο στα ίδια αμινοξικά κατάλοιπα (Ser Thr). Από πολύ νωρίς παρατηρήθηκε η πιθανή αλληλεπίδρασή τους όταν εντοπίστηκε η γλυκοζυλιωμένη Thr58 στην πρωτεΐνη c-myc, σε περιοχή όπου ήταν ήδη γνωστό οτι υφίσταται φωσφορυλίωση (Chou et al., 1995).

Εντοπίστηκαν περιοχές Ο-συνδεδεμένης γλυκοζυλίωσης σε πρωτεΐνες υπεύθυνες για τη συναρμολόγηση της μιτωτικής ατράκτου και την κυτταροκίνηση εκ των οποίων κάποιες είναι ίδιες ή βρίσκονται κοντά σε θέσεις φωσφορυλίωσης. Η Ο-συνδεδεμένη γλυκοζυλίωση είχε επιπτώσεις στη φωσφορυλίωση πρωτεϊνών που σχετίζονται με τη μιτωτική άτρακτο. Υπερέκφραση της OGT οδήγησε σε αύξηση της ανασταλτικής φωσφορυλίωσης της κυκλινο-εξαρτώμενης κινάσης 1 (CDK-1) και μείωσε τη φωσφορυλίωση των πρωτεϊνών-στόχων της (Wang et al., 2010).

Σε άλλες μελέτες, έχει παρατηρηθεί με αναστολή της δράσης της GSK3-β κινάσης, αύξηση της γλυκοζυλίωσης (O-GlcNAcylation) σε ορισμένες πρωτείνες, καθώς επίσης και μείωση σε άλλες. Το αντίστροφο φαινόμενο έχει παρατηρηθεί, επίσης, και μέσω αναστολής της O-GlcNAcase η οποία οδήγησε σε τριπλάσια αύξηση της γλυκοζυλίωσης και είχε ως αποτέλεσμα

σημαντική μείωση καθώς και σημαντική αύξηση της φωσφορυλίωσης διαφόρων πρωτεινών (Hart et al., 2011).

1.2.5 Βιοϊατρική σημασία γλυκοζυλίωσης

Οι συγγενείς διαταραχές γλυκοζυλίωσης αποτελούν μία ομάδα διαταραχών που περιλαμβάνουν περισσότερες από 100 ασθένειες, που προκύπτουν λόγω ανωμαλιών στη διαδικασία της γλυκοζυλίωσης και προάγουν διάφορες δυσμορφικές καταστάσεις όπως καθυστερημένη ανάπτυξη, ηπατοπάθεια, διαταραχές στην πήξη του αίματος και νευρολογικές ανωμαλίες (Chang, He, Lam, 2018).
Μεταλλαγές σε γονίδιο που κωδικοποιεί για μία υπομονάδα της ολιγοσακχαρυλοτρανσφεράσης έχουν συσχετιστεί με την εμφάνιση διανοητικής καθυστέρησης (Molinari et al., 2008).
Αλλοιώσεις σχετιζόμενες με αθρώπινους όγκους έχουν, επίσης, παρατηρηθεί στη Ν-συνδεδεμένη γλυκοζυλίωση μεταξύ των οποίων ο πρόωρος τερματισμός της γλυκάνης οδηγώντας σε ενδοκυτταρική συσσώρευση γλυκανών υψηλής περιεκτικότητας σε μαννόζη και αυξημένη διακλάδωση λόγω υψηλής έκφρασης του MGAT5 που επάγει τη δημιουργία περίπλοκων γλυκανών (Oliveira-Ferrer, Legler, & Milde-Langosch, 2017).
Η διαδικασία της Ο-γλυκοζυλίωσης έχει επίσης, συσχετιστεί με την ασθένεια του καρκίνου (Kudelka et al., 2015).

1.2.6 Εργαλεία πρόβλεψης θέσεων γλυκοζυλίωσης

Η συχνή εμφάνιση της γλυκοζυλίωσης και η σπουδαιότητα των απορροιών της στις λειτουργίες των πρωτεϊνών, καθιστούν αναγκαίο τον προσδιορισμό όσο το δυνατόν περισσότερων θέσεων εμφάνισης της. Τα τελευταία χρόνια με την ανάπτυξη καινούριων τεχνολογιών έχει γίνει εφικτή η ανίχνευση χιλιάδων θέσεων με ένα μόνο πείραμα. Αλλά λόγω της χαμηλής ευαισθησίας τους σε πρωτεΐνες με χαμηλά επίπεδα έκφρασης απέχουμε ακόμη πολύ από την κάλυψη ολόκληρου του πρωτεώματος του ανθρώπου, όσο και άλλων ευκαρυωτικών οργανισμών. Για τον λόγο αυτό, καθίσταται αναγκαία η ανάπτυξη υπολογιστικών μεθόδων πρόβλεψης, οι οποίες παρέχουν ακρίβεια, ταχύτητα και άνεση.
Για τον σκοπό αυτό έχουν δημιουργηθεί διάφορα εργαλεία πρόβλεψης θέσεων γλυκοζυλίωσης με μεθόδους μηχανικής μάθησης μεταξύ των οποίων τα NetNGlyc και NetOGlyc για πρόβλεψη θέσεων Ν-συνδεδεμένης γλυκοζυλίωσης και Ο-συνδεδεμένης γλυκοζυλίωσης, αντίστοιχα (Blom et al., 2004). Δυνατότητα ανίχνευσης N-,O- και C-συνδεδεμένης γλυκοζυλίωσης παρέχουν οι EnsemblGly (Caragea et al., 2007), GPP (Hamby and Hirst, 2008), GlycoEP (Chauhan et al., 2013). Τέλος, το GlycoPP είναι το μοναδικό εργαλείο που παρέχει τη δυνατότητα ανίχνευσης θέσεων γλυκοζυλίωσης σε προκαρυώτες (Chauhan et al., 2012).

1.3 Μηχανική Μάθηση

Μηχανική μάθηση είναι ο επιστημονικός κλάδος ο οποίος επικεντρώνεται στη μελέτη και κατασκευή αλγορίθμων από ένα σύνολο δεδομένων με την ικανότητα εξαγωγής συμπερασμάτων σχετικά με αυτά. Αποτελεί υποπεδίο της

επιστήμης των υπολογιστών και η διεργασία αυτή διεκπεραιώνεται από υπολογιστικά συστήματα, χωρίς αυτά να χρειαστεί να προγραμματιστούν εκ νέου.

1.3.1 Μη επιτηρούμενη μάθηση

Στη μη επιτηρούμενη μάθηση (Unsupervised Learning), τα δεδομένα εκπαίδευσης αποτελούνται μόνο από τις εισόδους (input) και ο αλγόριθμος προσπαθεί να ανακαλύψει κρυμμένα μοτίβα μέσα στο σύνολο των δεδομένων. Χαρακτηριστικό παράδειγμα εφαρμογής της προσέγγισης αυτής, αποτελούν οι μικροσυστοιχίες και η ανάλυση της γονιδιακής έκφρασης (Rana, Vijayeeta, Kar, Das, & Mishra, 2016).

1.3.2 Επιτηρούμενη μάθηση

Στην επιτηρούμενη μάθηση (Supervised Learning) απαιτείται η γνωστοποίηση των τιμών εξόδου στα δεδομένα εκπαίδευσης. Κάθε τιμή εισόδου αντιστοιχείται σε μία τιμή εξόδου και σκοπός, σε αυτού του είδους την προσέγγιση, είναι η χρήση αλγορίθμων για την κατασκευή μιας συνάρτησης με δυνατότητα αντιστοίχισης τιμών εισόδου σε δεδομένες τιμές εξόδου. Τη δημιουργία της συνάρτησης ακολουθεί η γενίκευση της, έτσι ώστε να καταστεί δυνατή η εξαγωγή συμπερασμάτων σε εισόδους με εξόδους που δεν έχουν ακόμη προσδιοριστεί (Kotsiantis, 2007).
Η μέθοδος αυτή, χρήζει τεράστιου βιολογικού και μη, ενδιαφέροντος και έχει εφαρμοστεί μεταξύ άλλων για την πρόβλεψη μεμβρανικών πρωτεινών (Gromiha & Yabuki, 2008), υποκυτταρικού εντοπισμού (Shen & Burger, 2007) καθώς επίσης και για την πρόβλεψη θέσεων ουβικουιτίνωσης (Tung & Ho, 2008). Παρακάτω ακολουθεί μία σειρά από τους κυριότερους κατηγοριοποιητές που χρησιμοποιούνται σε αυτού του είδους τη μάθηση.

1.3.3 Μέθοδοι κατηγοριοποίησης

1.3.3.1 Κατηγοριοποιητές Κανόνων (Rule-Based Classifiers)

Οι κατηγοριοποιητές κανόνων είναι μία τεχνική η οποία βασίζεται στην χρήση μιας συλλογής κανόνων για την διεκπεραίωση της κατηγοριοποίησης.

1.3.3.2 Μπαϋεσιανοί κατηγοριοποιητές (Bayesian Classifiers)

Οι Μπαϋεσιανοί κατηγοριοποιητές αποτελούν σημαντικά εργαλεία για τη μοντελοποίηση πιθανολογούμενων σχέσεων μεταξύ των χαρακτηριστικών και των μεταβλητών κατηγορίας. Το θεωρητικό του υπόβαθρο είναι το θεώρημα του Bayes, μέσω του οποίου υπολογίζεται η υπό συνθήκη πιθανότητα P(H|E), δηλαδή την πιθανότητα να ισχύει η υπόθεση H, με δεδομένο ότι ισχύει το γεγονός E.
Τέτοιου είδους κατηγοριοποιητές είναι οι Αφελείς Μπαϋεσιανοί Κατηγοριοποιητές (Naïve Bayesian) και τα Μπαϋεσιανά Δίκτυα (Bayesian Networks).

1.3.3.3 Δένδρα Απόφασης (Desicion Trees)

Αυτού του είδους οι κατηγοριοποιητές κατασκευάζουν ένα δένδρο όπου κάθε κόμβος αντιπροσωπεύει ένα χαρακτηριστικό και κάθε εξερχόμενο κλαδί αντιπροσωπεύει μία πιθανή τιμή που μπορεί να λάβει αυτό το χαρακτηριστικό. Η διαδικασία της κατηγοριοποίησης ξεκινάει από τον κόμβο-ρίζα του δέντρου και συνεχίζεται στον επόμενο κόμβο ανάλογα με την εκάστοτε τιμή του χαρακτηριστικού. Στην **Εικόνα *7*** παρατίθεται ένα παράδειγμα όπου απεικονίζεται η κατασκευή ενός δένδρου απόφασης βασιζόμενο στα δεδομένα του πίνακα που δίνεται από κάτω.

Ο πιο γνωστός αλγόριθμος για την κατασκευή δένδρων αποφάσεων είναι ο C4.5 , μία παραλλαγή του ID3 του Quinlann.

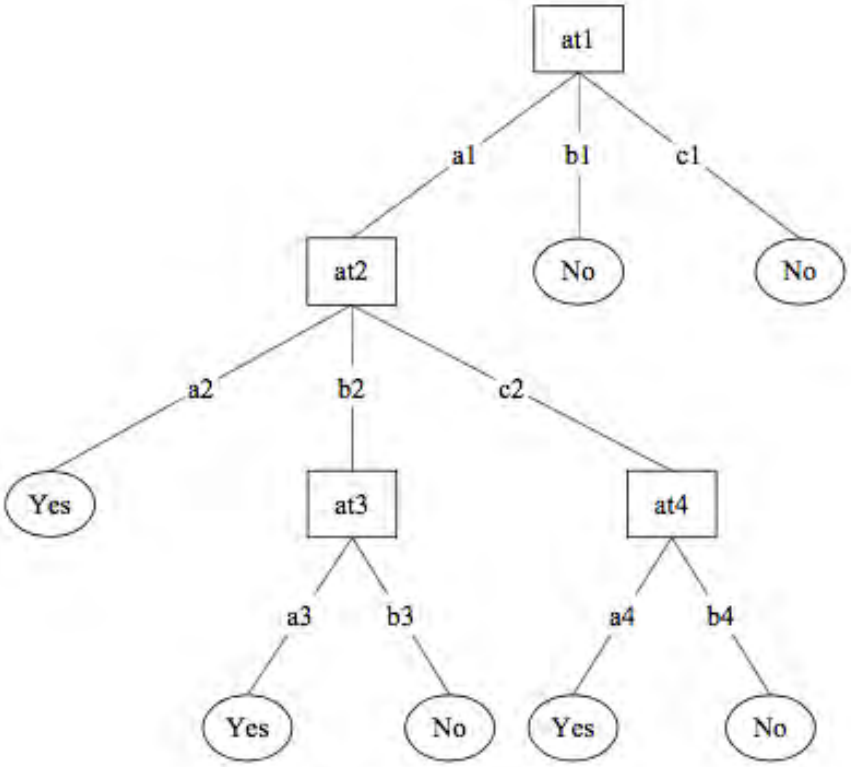


Figure 2. A decision tree

at1	at2	at3	at4	Class
a1	a2	a3	a4	Yes
a1	a2	a3	b4	Yes
a1	b2	a3	a4	Yes
a1	b2	b3	b4	No
a1	c2	a3	a4	Yes
a1	c2	a3	b4	No
b1	b2	b3	b4	No
c1	b2	b3	b4	No

Table 2. Training Set

Εικόνα 7 *Παράδειγμα κατασκευής δένδρου απόφασης (επάνω), από τα στοιχεία του πίνακα (κάτω)* **(Kotsiantis, 2007)**

1.3.3.4 Μέθοδοι ομάδας (Ensemble Learning)

Σε αυτή την κατηγορία, δημιουργούνται περισσότερα του ενός, μοντέλα πρόβλεψης και η ετικέτα κατηγορίας που αποδίδεται σε μία εγγραφή εισόδου είναι το αποτέλεσμα της συνάθροισης των προβλέψεων πολλών κατηγοριοποιητών. Χαρακτηριστικά παραδείγματα μεθόδων αυτής της κατηγορίας αποτελούν η εμφωλίαση (bagging), η ενίσχυση (boosting) και το τυχαίο δάσος (random forest).

1.3.4 Εκτίμηση απόδοσης

Για την εκτίμηση της απόδοσης ενός κατηγοριοποιητή, απαιτείται πρώτα η δημιουργία ενός συνόλου ελέγχου. Υπάρχουν διάφορες μέθοδοι, εκ των οποίων οι πιο βασικές είναι η Εκτίμηση μέσω συνόλου ελέγχου (Holdout Method) και η Διασταυρώμενη Επικύρωση (Cross-Validation).

1.3.4.1 Εκτίμηση μέσω συνόλου-ελέγχου (Holdout method)

Ο συνολικός αριθμός δεδομένων διαρείται σε 2 υποσύνολα, το σύνολο εκπαίδευσης που απαρτίζεται από τα δεδομένα που θα χρησιμοποιηθούν για την εκπαίδευση του αλγορίθμου, και το σύνολο ελέγχου στο οποίο περιέχονται τα δεδομένα που θα χρησιμοποιηθούν για την εκτίμηση της απόδοσης του μοντέλου. Η αναλογία των 2 υποσυνόλων είναι στην κρίση του αναλυτή.

1.3.4.2 Διασταυρωμένη Επικύρωση (Cross-Validation)

Σε αυτού του είδους την προσέγγιση το σύνολο των δεδομένων χωρίζεται τυχαία σε k υποσύνολα, εκ των οποίων το ένα χρησιμοποιείται για την εκτίμηση της απόδοσης του κατηγοριοποιητή και τα υπόλοιπα για την εκπαίδευση του. Η διαδικασία αυτή επαναλαμβάνεται k φορές με διαφορετικό σύνολο ελέγχου κάθε φορά και το συνολικό σφάλμα υπολογίζεται από τα σφάλματα των k εκτελέσεων.

1.3.4.3 Μέτρα αξιολόγησης της απόδοσης

1.3.4.3.1 Πίνακας σύγχυσης (Confusion Matrix)

Απεικονίζεται πόσες εγγραφές κατηγοριοποιήθηκαν και σε ποια τιμή εξόδου. Τα στοιχεία δείχνουν τον αριθμό των εγγραφών των οποίων η πραγματική τάξη είναι η σειρά και η προβλεφθείσα η στήλη (**Πίνακας *1***).

Οι θετικές εγγραφές οι οποίες κατηγοριοποιήθηκαν ως θετικές, χαρακτηρίζονται ως Αληθώς Θετικές (True Positive, TP), ενώ εκείνες οι οποίες κατηγοριοποιήθηκαν ως αρνητικές χαρακτηρίζονται ως Ψευδώς Αρνητικές (False Negative, FN). Ανάλογα, οι αρνητικές εγγραφές οι οποίες αναγνωρίστηκαν ως αρνητικές ονομάζονται Αληθώς Αρνητικές (True Negatve, TN), ενώ εκείνες που κατηγοριοποιήθηκαν ως θετικές καλούνται Ψευδώς Θετικές (False Positive, FP).

		Προβλεπόμενη	
Πραγματική τάξη		Θετικά	Αρνητικά
	Θετικά	ΑΘ	ΨΑ
	Αρνητικά	ΨΘ	ΑΑ

Πίνακας 1 *Πίνακας σύγχυσης. Οριζόντια εμφανίζεται η πραγματική τάξη, ενώ κάθετα η τάξη που προβλέπει ο κατηγοριοποιητής (ΑΘ: Αληθώς θετικά, ΨΘ: Ψευδώς θετικά, ΑΑ: Αληθώς αρνητικά, ΨΑ: Ψευδώς αρνητικά)*

1.3.4.3.2 Ορθότητα (Accuracy)

Με τον όρο ορθότητα, εκφράζεται ο αριθμός των εγγραφών εισόδου που κατηγοριοποιείται σωστά και υπολογίζεται από την εξής σχέση:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

1.3.4.3.3 Ευαισθησία (Sensitivity)

Η ευαισθησία ή αληθής θετικός βαθμός (True Positive Rate, TPR) ορίζεται ως η αναλογία των θετικών δειγμάτων, που έχουν προβλεφθεί σωστά από το μοντέλο, δηλαδή:

$$TPR = \frac{TP}{TP+FN}$$

1.3.4.3.4 Ειδικότητα (Specificity)

Η ειδικότητα ή αλλιώς ο αληθής αρνητικός βαθμός (True Negative Rate, TNR) ορίζεται ως η αναλογία των αρνητικών δειγμάτων που έχουν προβλεφθεί σωστά από το μοντέλο, δηλαδή:

$$TNR = \frac{TN}{TN+FP}$$

1.3.4.3.5 Ακρίβεια (Precision)

Προσδιορίζει το πλήθος των εγγραφών οι οποίες αποδεικνύεται στην πραγματικότητα ότι είναι θετικές στην ομάδα που ο κατηγοριοποιητής έχει δηλώσει ως θετική κατηγορία. Οσο πιο υψηλή είναι η ακρίβεια, τόσο πιο μικρό είναι το πλήθος των ψευδώς θετικών σφαλμάτων που κάνει ο κατηγοριοποιητής.

$$Precision = \frac{TP}{TP+FP}$$

1.3.4.3.6 F-measure

Εξαρτάται από την ακρίβεια και την ανάκληση και προκύπτει από την παρακάτω σχέση:

$$\text{F-measure} = 2 \text{ x } \frac{Precision\ x\ Recall}{Precision+Recall}$$

1.3.4.3.7 Matthews Correlation Coefficient (MCC)

Ο συντελεστής Matthews Correlation Coefficient (MCC) αποτελεί ακόμη ένα μέτρο αξιολόγησης και προκύπτει απο την εξής σχέση:

MCC = (TP x TN – FP x FN) / $\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}$

1.3.4.3.8 ROC Area

Η καμπύλη ROC (Receiver Operating Characteristic Curve), είναι ένας γραφικός τρόπος αναπαράστασης της σχέσης ανάμεσα στον αληθή θετικό βαθμό (TPR) και τον ψευδή θετικό βαθμό ενός κατηγοριοποιητή (False Positive Rate, FPR). Ο αληθής θετικός βαθμός σχεδιάζεται κατά μήκος του άξονα ψ, ενώ ο ψευδής θετικός κατά μήκος του άξονα χ. Δημιουργείται σημειώνοντας σε διάφορα κατώφλια την αναλογία από τις τιμές των αληθούς θετικού βαθμού και ψευδούς θετικού βαθμού.

Η περιοχή κάτω από την καμπύλη ROC (ROC Area), παρέχει ακόμη μία προσέγγιση για την εκτίμηση του καλύτερου, κατά μέσο όρο, μοντέλου. Αν το μοντέλο είναι τέλειο τότε η περιοχή κάτω από την καμπύλη του ROC θα είναι ίση με το 1. Αν το μοντέλο κάνει απλώς τυχαία πρόβλεψη, τότε η περιοχή κάτω από την καμπύλη του ROC θα είναι ίση με 0.5. Ένα μοντέλο που είναι αυστηρά καλύτερο από ένα άλλο θα πρέπει να έχει μεγαλύτερη περιοχή κάτω από την καμπύλη ROC.

1.3.4.3.9 PRC Area

Η PRC καμπύλη (Precision Recall Curve) κατασκευάζεται με ανάλογο, της ROC καμπύλης, τρόπο με τη διαφορά ότι στον ψ άξονα μπαίνει η ανάκληση και στον χ άξονα, η ακρίβεια.

Η περιοχή κάτω από την καμπύλη (PRC Area), αποτελεί ακόμη μία σημαντική μέθοδος εκτίμησης της απόδοσης ενός κατηγοριοποιητή.

1.4 Weka

Το Weka (Waikato Enviroment for Knowledge Analysis) αποτελεί μία δημοφιλή σουίτα λογισμικού για μηχανική μάθηση και εξόρυξη δεδομένων, γραμμένη σε Java, η οποία αναπτύχθηκε από ερευνητές του πανεπιστημίου του Waikato στη Νέα Ζηλανδία.

Με την εκκίνηση του διατίθενται οι εξής τρεις κύριες εφαρμογές:

- Explorer, το πιο εύχρηστο γραφικό περιβάλλον χρήστη (Graphical User Interface). Μέσω αυτού, ο χρήστης έχει τη δύνατοτητα να πραγματοποιήσει ένα μεγάλο εύρος διεργασιών μεταξύ των οποίων η προ-επεξεργασία των δεδομένων, η απαλοιφή θορύβου, η κατηγοριοποίηση, η συσταδοποίηση κ.ά.
- Experimenter, είναι ένα περιβάλλον που παρέχει τη δυνατότητα διεξαγωγής, τροποποίησης και ανάλυσης πειραμάτων με περισσότερο ευέλικτους χειρισμούς.
- Knowledge Flow, στο οποίο πραγματοποιούνται οι ίδιες δράσεις με το Explorer διαθέτει όμως διαφορετική διεπαφή (interface).

1.5 Σκοπός

Σκοπός της παρούσας εργασίας ήτανε η χρήση δεδομένων γλυκοζυλοπρωτεωμικής για την πρόβλεψη θέσεων Ν-συνδεδεμένης γλυκοζυλίωσης με τη χρήση αλγορίθμων, στα πλαίσια της σουίτας λογισμικού Weka.

2 Υλικά-Μέθοδοι

2.1 Λήψη δεδομένων γλυκοζυλοπρωτεωμικής

Ανατρέχοντας σε διάφορες μελέτες γλυκοζυλίωσης, συλλέξαμε έναν μεγάλο αριθμό Ν-γλυκοζυλιωμένων πεπτιδίων.
Έπειτα εφαρμόστηκαν αυστηρά κριτήρια φιλτραρίσματος, αποσκοπώντας στην αύξηση της πιστότητας των αποτελεσμάτων.
Στον **Πίνακας *2*** παρατίθενται τα άρθρα στα οποία βασιστήκαμε για την συλλογή των πεπτιδίων, οι συγγραφείς, τα PubMed ID τους, ο οργανισμός από τον οποίο προήλθαν καθώς επίσης και τα κριτήρια που χρησιμοποιήθηκαν για το φιλτράρισμα των πεπτιδίων.

PubMed ID	**Ετος Δημοσίευσης**	**Συγγραφείς**	**Οργανισμός**	**Επιλεγμένοι παράμετροι**
19159218	2009	Chen et al.	Homo sapiens	• dCn ≥ 0.15 • Xcorr ≥ 1.8
20510933	2010	Zielinska et al.	Mus musculus	Confidence ≥ 0.99
21441315	2011	Parker et al.	Rattus Norvegicus	IonScore ≥ 45
2823882	2012	Kaji et al.	Mus musculus	P-value < 0.05
22633491	2012	Zielinska et al.	• Arabidopsis Thaliana • Schizosaccharomyces Pombe • Saccharomyces cerevisae • Caenorabdhitis elegans • Drosophila melanogaster • Danio rerio	Confidence ≥ 0.99
24090084	2013	Parker et al.	Rattus Norvegicus	hcd ≥ 20

25182382	2014	Cheng et al.	Homo sapiens	Score ≥ 45
24495048	2014	Zhu et al.	Homo sapiens	• PepVelos Score ≥ 90 • Pep5600 Score ≥ 90

Πίνακας 2 *Τα PubMed IDs (1η στήλη), έτη δημοσίευσης (2η στήλη) και συγγραφείς (3η στήλη) των εργασιών απ'όπου συλλέχθηκαν τα γλυκοζυλιωμένα πεπτίδια μαζί με τους οργανισμούς από τους οποίους προήλθαν (4η στήλη) και τις παραμέτρους που χρησιμοποιήθηκαν για το φιλτράρισμά τους (5η στήλη)*

2.2 Λήψη Πρωτεωμάτων

Εν συνεχεία, πραγματοποιήθηκε λήψη των πρωτεωμάτων του ανθρώπου (*Homo sapiens*), του ποντικού (*Mus musculus*), του αρουραίου (*Rattus norvegicus*), του ψαριού ζέβρα (*Danio rerio*), της μύγας (*Drosophila melanogaster*), του *Caenorhabditis elegans*, της αραβιδόψης (*Arabidopsis thaliana*), του σακχαρομύκητα (*Saccharomyces cerevisae*) και του σχιζοσακχαρομύτα (*Schizosaccharomyces pombe*). Ολα λήφθηκαν από την Enseml με εξαίρεση το πρωτέωμα του *Schizosaccharomyces pombe*, το οποίο λήφθηκε από τη βάση δεδομένων PomBase (pombase.org)
Αναφορικά με τις πρωτεΐνες οι οποίες διαθέτουν περισσότερες από μία ισομορφές δημιουργήθηκε perl script για την συλλογή των μεγαλύτερων σε μήκος.

2.3 Αντιστοίχιση πεπτιδίων στα πρωτεώματα

Βασιζόμενοι στο protein id του κάθε πεπτιδίου έγινε αντιστοίχιση του τελευταίου πάνω στο πρωτέωμα του αντίστοιχου οργανισμού. Επειτα έγινε προσθήκη ή αφαίρεση αμινοξικών καταλοίπων έτσι ώστε κάθε πεπτιδική αλληλουχία να αποτελείται από 29 αμινοξικά κατάλοιπα και το τροποποιημένο αμινοξύ (Ασπαραγίνη) να βρίσκεται πάντα στη μέση της αλληλουχίας, δηλαδή στη θέση 15.
Με το πέρας των παραπάνω διαδικασιών καταφέραμε να ολοκληρώσουμε τη δημιουργία του θετικού υποσυνόλου εκπαίδευσης.

2.4 Συλλογή αρνητικών δεδομένων εκπαίδευσης

Για τη δημιουργία ενός ολοκληρωμένου συνόλου δεδομένων εκπαίδευσης, απαιτείται, πέρα από το θετικό υποσύνολο, και το αρνητικό, πεπτίδια δηλαδή τα οποία δεν υπόκεινται σε Ν-συνδεδεμένη γλυκοζυλίωση.
Για την πραγματοποίηση του παραπάνω χρησιμοποιήθηκαν perl scripts, μέσω των οποίων έγινε τυχαία επιλογή πεπτιδίων από τα πρωτεώματα των οργανισμών, τα οποία δεν έχουν ήδη συμπεριληφθεί στο θετικό υποσύνολο.

2.5 Προεπεξεργασία δεδομένων

Πριν την εισαγωγή των δεδομένων εκπαίδευσης, στη σουίτα λογισμικού Weka, απαιτείται η μετατροπή του αρχείου σε συγκεκριμένα format, μεταξύ των οποίων το CSV (Comma Separated Values) το οποίο χαρακτηρίζεται από την επέκταση .csv. Σε αυτού του είδους το format, οι τιμές εντός μιας εγγραφής πρέπει να διαχωρίζονται με κόμμα. Στην **Εικόνα *8*** απεικονίζεται ένα τμήμα του αρχείου όπου φαίνονται οι τιμές κάθε εγγραφής χωρισμένες με κόμμα.

```
D,L,R,V,Q,A,R,A,Q,P,G,T,M,S,N,G,T,E,T,R,G,T,G,L,T,A,V,A,V,YES
W,D,Q,L,P,V,D,V,Q,N,G,F,I,R,N,Y,T,I,F,Y,R,T,I,I,G,N,E,T,A,YES
F,I,R,N,Y,T,I,F,Y,R,T,I,I,G,N,E,T,A,V,N,V,D,S,S,H,T,E,Y,T,YES
V,Q,E,G,N,S,D,V,V,E,V,R,L,A,N,R,T,G,G,L,E,V,L,L,N,Q,E,V,L,YES
G,L,L,G,D,P,F,R,P,L,P,Q,Q,V,N,L,T,D,G,R,W,H,R,V,A,V,S,I,D,YES
E,N,F,S,L,L,I,T,L,R,G,Q,P,A,N,Q,S,V,L,L,S,I,Y,D,E,R,G,A,R,YES
L,V,D,S,S,D,I,F,A,H,I,S,G,A,N,S,F,K,C,N,Y,P,I,Q,S,W,I,_,_,NO
Q,V,C,S,G,F,I,V,G,T,N,H,G,F,N,V,Y,S,C,K,P,M,I,K,K,S,I,S,R,NO
L,A,V,L,K,G,K,S,V,I,T,S,N,T,N,L,G,V,Y,G,Q,G,M,L,T,L,S,G,P,NO
M,G,D,W,L,N,K,Q,L,R,L,P,V,S,N,S,L,N,K,S,D,V,I,E,I,D,L,G,P,NO
E,A,D,N,E,E,E,F,W,N,I,K,K,R,N,D,H,E,G,V,L,D,T,S,S,G,N,G,N,NO
```

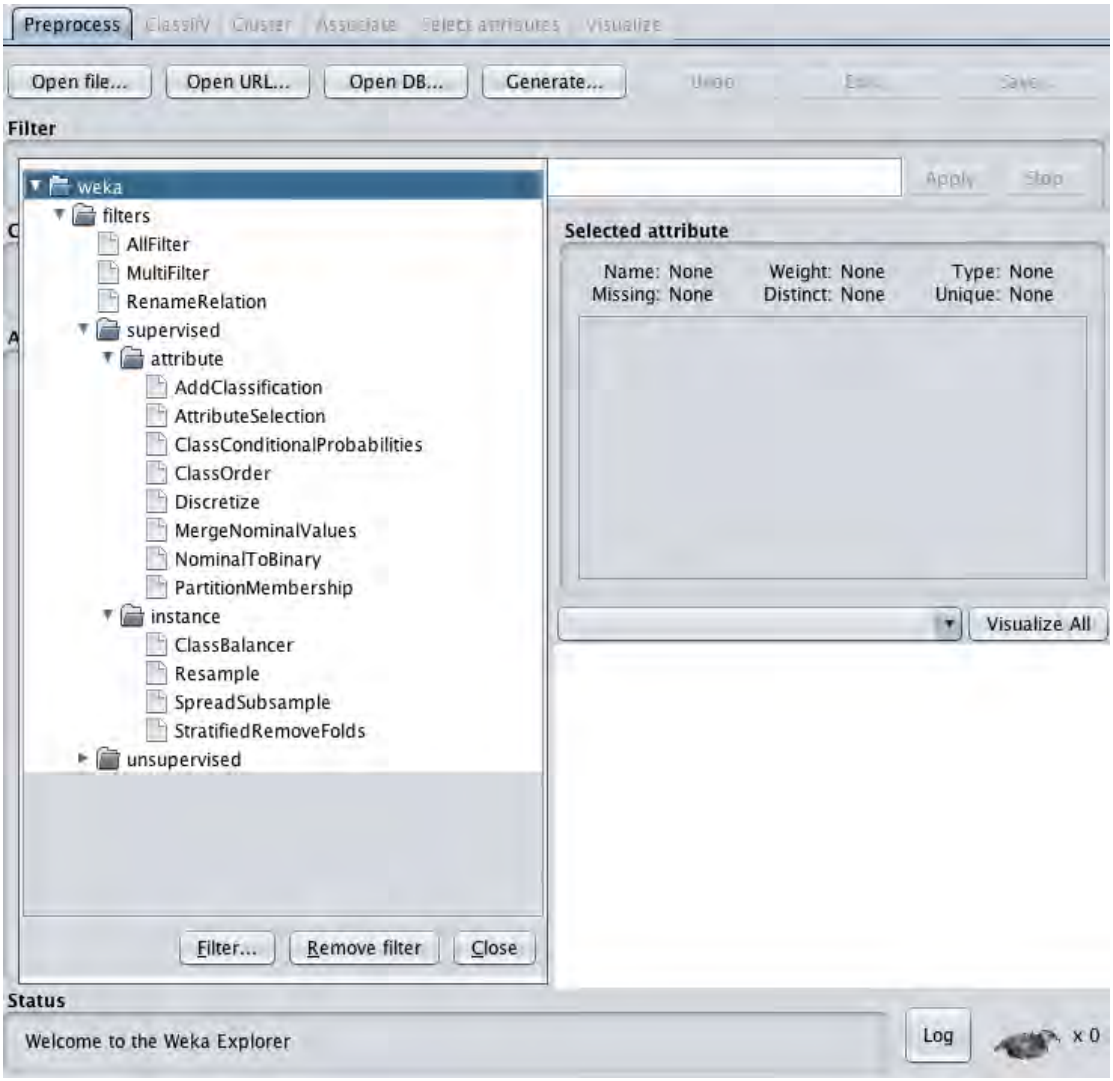
Εικόνα 8 *Μορφή αρχείου σε CSV fomat. Οι τιμές εντός των εγγραφών διαχωρίζονται με κόμμα*

2.6 Ανάλυση μέσω Weka

Το επόμενο στάδιο, ήτανε η εισαγωγή του αρχείου με τα δεδομένα γλυκοζυλίωσης στο περιβάλλον του Weka. Αφού επιλέξαμε από την αρχική οθόνη το γραφικό περιβάλλον χρήστη με το όνομα Explorer, φορτώσαμε το αρχείο μας στο πρώτο tab που εμφανίζεται (Preprocess). Εδώ, γίνεται η εισαγωγή των δεδομένων, τα οποία μπορούν να μεταφερθούν μέσω ενός αρχείου (με το κατάλληλο format), μέσω ενός URL ή μέσω μιας SQL βάσης δεδομένων.

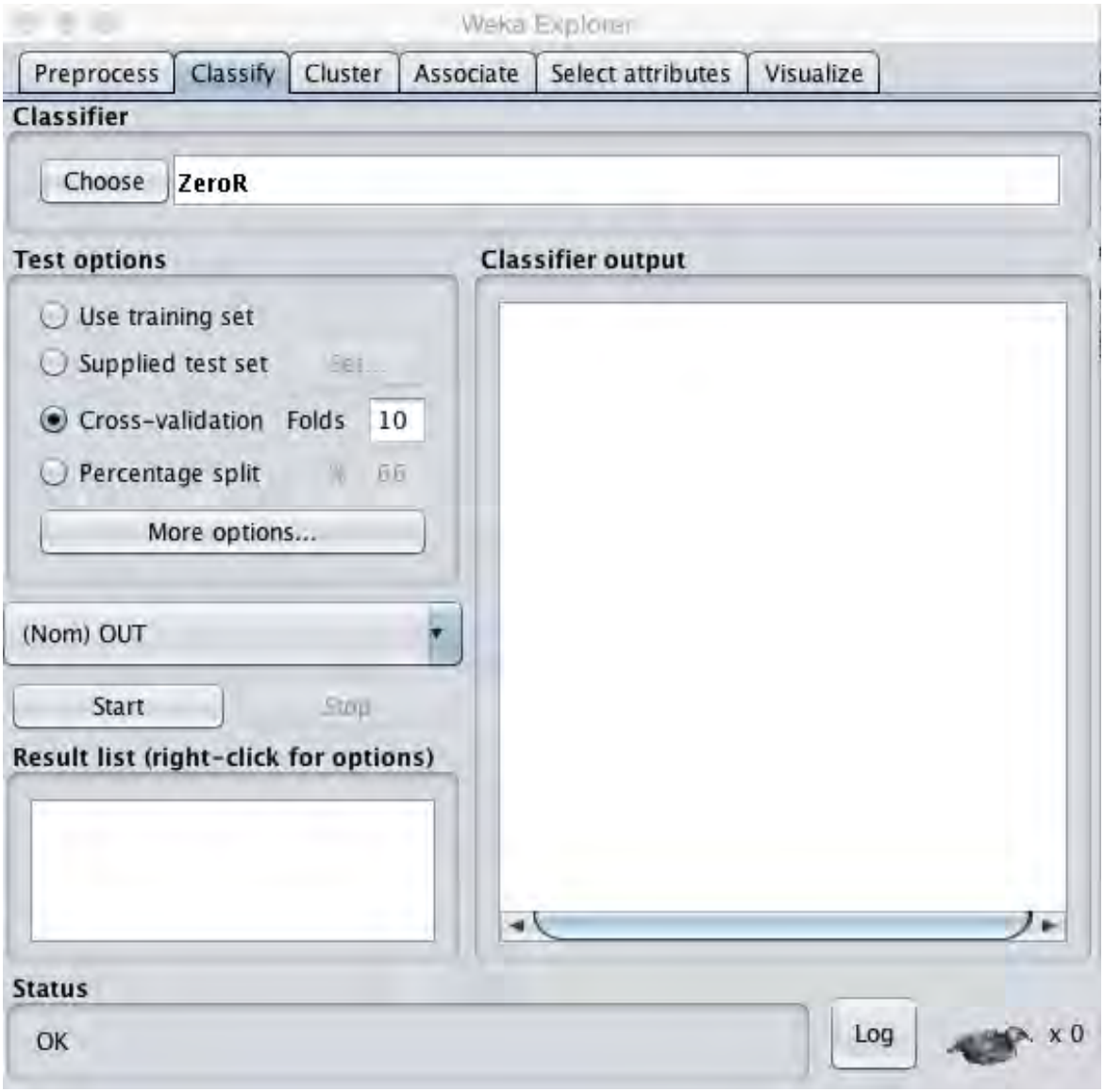
Επίσης, στο σημείο αυτό παρέχεται η δυνατότητα επεξεργασίας των χαρακτηριστικών (attributes) οι οποίες μπορούν είτε να απομακρυνθούν (remove), είτε να τροποποιηθούν (edit).

Επιπλέον παρέχεται και μία ποικιλία φίλτρων με τη χρήση των οποίων μας δίνεται η δυνατότητα να τροποποιήσουμε τις εισόδους ή τα χαρακτηριστικά μας (***Εικόνα 9***).



Εικόνα 9 Το tab "Preprocess" του Weka explorer. Φαίνονται τα εργαλεία που παρέχουν τη δυνατότητα τροποποίησης των εισόδων

Επειτα, στο παράθυρο Classify, ακολούθησε η διαδικασία κατασκευής του μοντέλου και πρόβλεψης των τιμών εξόδου, με διάφορους αλγόριθμους που παρέχει αυτή η σουίτα λογισμικού συμπεριλαμβανομένων των οποίων οι Κατηγοριοποιητές Κανόνων (ZeroR, OneR), Μπαϋεσιανοί (NaiveBayes) και Δένδρα Αποφάσεων (RandromTree, J48).
Για την αξιολόγηση της απόδοσης του μοντέλου επιλέξαμε το 10-fold Cross Validation. Μετά την επιλογή τους πατάμε «Start» και στο πεδίο με τίτλο «Classifier Output» εμφανίζονται τα αποτελέσματα της κατηγοριοποίησης, όπου περιλαμβάνονται το ποσοστό επιτυχούς πρόβλεψης του κατηγοριοποιητή και ορισμένα μέτρα αξιολόγησης της απόδοσης του (**Εικόνα *10***).



Εικόνα 10 *Το tab "Classify" του Weka Explorer. Γίνεται η επιλογή του επιθυμητού κατηγοριοποιητή και η επιλογή του τρόπου διαχωρισμού του συνόλου ελέγχου. Μετά την επιλογή τους, στο πεδίο με τίτλο «Classifier Output» εμφανίζονται τα αποτελέσματα της κατηγοριοποίησης, μεταξύ των οποίων το ποσοστό επιτυχούς πρόβλεψης του κατηγοριοποιητή και διάφορα μέτρα αξιολόγησης του.*

3 Αποτελέσματα-Συζήτηση

3.1 Συλλογή πεπτιδίων

Glycoproteomics Analysis of Human Liver Tissue by Combination of Multiple Enzyme Digestion and Hydrazide Chemistry (Chen et al., 2009)

Στην εργασία αυτή, οι συγγραφείς κατάφεραν να ανιχνεύσουν 939 περιοχές N-γλυκοζυλίωσης από ανθρώπινο ιστό ήπατος. Η διαδικασία της πέψης πραγματοποιήθηκε με τη χρήση 3 ενζύμων (θρυψίνη, πεψίνη και

θερμολυσίνη) και τα παραγόμενα πεπτίδια διαχωρίστηκαν μέσω υγρής χρωματογραφίας (Liquid Chromatography). Επειτα ακολούθησε ανάλυση με φασματομετρία μαζών σε σειρά (Tandem Mass Spectometry) και περαιτέρω ανάλυση με τη χρήση της μηχανής αναζήτησης SEQUEST.
Από το σύνολο των πεπτιδίων, συλλέξαμε μόνο αυτά με dCn ≥ 0.15 και Xcorr ≥ 1.8.

Precision Mapping of an In Vivo N-Glycoproteome Reveals Rigid Topological and Sequence Constraints

Οι (Zielinska et al., 2010) κατάφεραν να ανιχνεύσουν 6367 περιοχές N-γλυκοζυλίωσης σε 2352 πρωτεΐνες από το πλάσμα του αίματος και 4 διαφορετικούς ιστούς ποντικών (εγκέφαλος, καρδιά, νεφρό, ήπαρ). Η μέθοδος περιλαμβάνει την υποβοηθούμενη με φίλτρο προετοιμασία δείγματος (Filter-aided Sample Preparation) με τη χρήση λεκτινών στις οποίες προσδένονται οι γλυκοζυλιωμένες πρωτεΐνες. Ακολούθησε χρήση της γλυκοσιδάσης PNGase F (peptide-N-glycosidase F) για την απογλυκοζυλίωση των πρωτεϊνών και πέψη με 2 ενδοπρωτεϊνάσες (θρυψίνη και Glu-C). Η ανάλυση πραγματοποιήθηκε με το φασματόμετρο μάζας LTQ-Orbitrap Velos, η αντιστοίχιση των φασμάτων με το λογισμικό MaxQuant και η εύρεση των πρωτεϊνών με τη μηχανή αναζήτησης MASCOT.
Η συλλογή των πεπτιδίων υψηλής πιστότητας έγινε με εφαρμογή του κριτηρίου Confidence ≥ 0.99.

Quantitative N-linked Glycoproteomics of Myocardial Ischemia and Reperfusion Injury Reveals Early Remodeling in the Extracellular Environment

Στη συγκεκριμένη εργασία πραγματοποιήθηκε πολυενζυμική πέψη με τα ένζυμα θρυψίνη, Asp-N και θερμολυσίνη σε πρωτεΐνες που απομονώθηκαν από το μυοκάρδιο αρουραίων, ακινητοποίηση με τη χρήση υδραζιδίου (hydrazide capture) και απογλυκοζυλίωση των πεπτιδίων με PNGase F. Ακολούθησε ενίσχυση μέσω χρωματογραφίας σε στήλη ZIC-HILIC και φασματομετρία μάζας με το LTQ-Orbitrap XL. Τα αποτελέσματα αναλύθηκαν με το λογισμικό Mascot (Parker et al., 2011).
Από τα 5607 γλυκοπεπτίδια που προέκυψαν με αυτές τις μεθόδους, συλλέξαμε εκείνα με IonScore ≥ 45.

Large-scale Identification of N-Glycosylated Proteins of Mouse Tissues and Construction of a Glycoprotein Database, GlycoProtDB (Kaji et al., 2012)

Εγινε αναζήτηση γλυκοζυλιωμένων πρωτεϊνών σε διάφορους ιστούς ποντικών και κατάφεραν να ανιχνεύσουν 5060 N-γλυκοζυλιωμένα πεπτίδια από 2556 γλυκοπρωτείνες, το οποίο οδήγησε στην κατασκευή μια βάσης δεδομένων γλυκοπρωτεϊνών (GlycoProtBD). Αμφότερα πραγματοποιήθηκαν με μια διαδικασία η οποία περιλαμβάνει την ακινητοποίηση των γλυκοπρωτεϊνών σε στήλες λεκτινών και πέψη με θρυψίνη. Επειτα πραγματοποιήθηκε καθαρισμός με χρωματογραφία υδρόφιλης αλληλεπίδρασης, χημική απογλυκοζυλίωση των πρωτεϊνών με PNGase F και σήμανση με O_{18} μια διαδικασία που ονομάζεται Isotope-coded Glycoyslation site-specific Tagging (IGOT).

Ακολούθησε υγρή χρωματογραφία και φασματομετρία μάζας με το LTQ-Orbitrap-Velos και τέλος τα πεπτίδια αναλύθηκαν με το MASCOT.
Το κριτήριο για τη συλλογή των πεπτιδίων ήτανε P-value < 0.05.

Mapping N-Glycosylation Sites across Seven Evolutionarily Distant Species Reveals a Divergent Substrate Proteome Despite a Common Core Machinery (Zielinska et al., 2012)

Η Zielinska και οι συνεργάτες της, σε συνδυασμό με την προηγούμενη μελέτη τους (Zielinska et al., 2010), κατόρθωσαν να ανιχνεύσουν 15.771 περιοχές N-γλυκοζυλίωσης σε 7 εξελικτικά απομακρυσμένους οργανισμούς. Η μέθοδος περιλαμβάνει, παρομοίως με την προηγούμενη εργασία, προετοιμασία του δείγματος με τη N-Glyco-FASP μέθοδο, πέψη των πρωτεϊνών με θρυψίνη και Glu-C, ακινητοποίηση των γλυκοζυλιωμένων πρωτεϊνών με τη χρήση λεκτινών και απογλυκοζυλίωση με την PNGase F. Η φασματομετρία μάζας διεκπεραιώθηκε με το LTQ-Orbitrap Velos και τα φάσματα αναλύθηκαν με το λογισμικό MaxQuant.

Από το σύνολο των πεπτιδίων της παρούσας εργασίας, επιλέχθηκαν εκείνα με Confidence ≥ 0.99.

Site-Specific Glycan-Peptide Analysis for Determination of N- Glycoproteome Heterogeneity (Parker et al., 2013)

Αρχικά έγινε χαρακτηρισμός όλων των N-συνδεδεμένων γλυκανών με τη χρήση υγρής χρωματογραφίας και φασματομετρία μάζας σε σειρά και τα δεδομένα χρησιμοποιήθηκαν για την κατασκευή μίας βάσης δεδομένων N-γλυκανών. Επειτα έγινε πέψη των γλυκοπεπτιδίων με θρυψίνη, τα οποία εμπολουτίστηκαν με τη μέθοδο (Zwitterionic Hydrophilic Interaction Liquid Chromatography, Zic-HILIC) και μελέτη των παραγόμενων πεπτιδίων με υγρή χρωματογραφία αντίστροφης φάσης (Reverse Phase Liquid Chromatography, RPLC). Τα παραγόμενα μόρια μοιράστηκαν σε δύο υποσύνολα. Στο πρώτο υποσύνολο έγινε αφαίρεση των γλυκανών με τη γλυκοσιδάση PNGase F και τα απογλυκοζυλιωμένα πεπτίδια αναλύθηκαν με νανο-υγρή χρωματογραφία αντίστροφης φάσης (nano-Reverse Phase Liquid Chromatography, nano-RPLC) και τα φάσματα ταυτοποιήθηκαν με τη μηχανή αναζήτησης Mascot. Το δεύτερο υποσύνολο αναλύθηκε με nano-RPLC αφού πρώτα προηγήθηκε θραυσματοποίηση με τις μεθόδους CID (Collision Induced Dissociation) και HCD (Higher-Energy Collision Dissociation). Η αντιστοίχιση των γλυκανών στις πεπτιδικές αλληλουχίες πραγματοποιήθηκε μέσω αναζήτησης των φασμάτων που παρήγαγαν τα N-γλυκοζυλιωμένα πεπτίδια έναντι σε αυτά των απογλυκοζυλιωμένων πεπτιδίων, οδηγώντας στην ταυτοποίηση 862 N-γλυκοζυλιωμένων πεπτιδίων.

Στην παρούσα εργασία, με εφαρμογή του φίλτρου hcd ≥ 20 συλλέχθησαν 845 πεπτίδια.

Large-scale characterization of intact N-glycopeptides using an automated glycoproteomic method (Cheng et al., 2014)

Έγινε πέψη πρωτεϊνών από ανθρώπινο νεφρό και εμπλουτισμούς με χρωματογραφία υδρόφιλης αλληλεπίδρασης. Τα παραγόμενα γλυκοπεπτίδια χωρίστηκαν σε δύο επιμέρους ομάδες. Η μία ομάδα υπέστη απογλυκοζυλίωση με PNGase F, ακολουθούμενη από υγρή χρωματογραφία και φασματομετρία μάζας. Στη δεύτερη ομάδα πραγματοποιήθηκε απ'ευθείας υγρή χρωματογραφία και φασματομετρία μάζας. Τα αποτελέσματα αναλύθηκαν με το Mascot.
Με εφαρμογή του φίλτρου Score ≥ 45 συλλέξαμε 1169 πεπτίδια

Comprehensive Mapping of Protein N- Glycosylation in Human Liver by Combining Hydrophilic Interaction Chromatography and Hydrazide Chemistry (Zhu et al., 2014)

Επιτεύχθηκε η ανίχνευση 14498 γλυκοπεπτιδίων από ανθρώπινο ήπαρ. Η διαδικασία είχε ως εξής: οι πρωτεΐνες χωρίστηκαν σε τρεις ομάδες. Στην πρώτη ομάδα πραγματοποιήθηκε πέψη των πρωτεϊνών και καθαρισμός μέσω υγρής χρωματογραφίας υδρόφιλης αλληλεπίδρασης. Επειτα ακολούθησε απογλυκοζυλίωση με την PNGase F. Στη δεύτερη ομάδα ακολουθήθηκαν τα ίδια βήματα με τη διαφορά ότι ο καθαρισμός έγινε με διαφορετική μέθοδο (hydrazide chemistry). Η τρίτη ομάδα, που αποτελούνται απο N-γλυκοζυλιωμένες πρωτείνες καθαρίστηκαν με τη μέθοδο hydrazide chemistry, επειτα ακολούθησε πολυενζυμική πέψη και τέλος, απογλυκοζυλίωση από το ίδιο ένζυμο. Τα αποτελέσματα αναλύθηκαν με τα LTQ-Orbitral Velos και TripleTOF 5600. Η εύρεση των πεπτιδίων έγινε με το λογισμικό MaxQuant.

Η συλλογή των πεπτιδίων έγινε με εφαρμογή του φίλτρου Score ≥ 90.

3.2 Εκτίμηση απόδοσης

3.2.1 ZeroR

Ο αλγόριθμος ZeroR ανήκει στους κατηγοριοποιητές κανόνων και η αρχή λειτουργίας του είναι πολύ απλή: αναθέτει σε κάθε εγγραφή εισόδου την τάξη που χαρακτηρίζει την πλειονότητα των εγγραφών.

	Προβλεπόμενη τάξη		
Πραγματική τάξη		Θετικά	Αρνητικά
	Θετικά	13840	3462
	Αρνητικά	13842	3460

Πίνακας 3 *Πίνακας σύγχυσης για την κατασκευή του μοντέλου με βάση τον αλγόριθμο ZeroR*

Τάξη	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Θετικά	0.800	0.800	0.500	0.800	0.615	-0.000	0.500	0.500

Αρνητικά	0.200	0.200	0.500	0.200	0.286	-0.000	0.500	0.500
Μέση τιμή	0.500	0.500	0.500	0.500	0.450	-0.000	0.500	0.500

Πίνακας 4 *Μέτρα αξιολόγησης του μοντέλου, το οποίο κατασκευάστηκε με βάση τον αλγόριθμο ZeroR*

3.2.2 OneR

Ο αλγόριθμος OneR ανήκει στους κατηγοριοποιητές κανόνων και πραγματοποιεί την κατηγοριοποίηση με βάση μια ομάδα κανόνων που εξετάζει μόνο ένα χαρακτηριστικό.

	Προβλεπόμενη τάξη	
Πραγματική τάξη	Θετικά	Αρνητικά
Θετικά	16845	457
Αρνητικά	2654	14648

Πίνακας 5 *Πίνακας σύγχυσης για την κατασκευή του μοντέλου με βάση τον αλγόριθμο OneR*

Τάξη	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Θετικά	0.974	0.153	0.864	0.974	0.915	0.827	0.910	0.854
Αρνητικά	0.847	0.026	0.970	0.847	0.904	0.827	0.910	0.898
Μέση τιμή	0.910	0.090	0.917	0.910	0.910	0.827	0.910	0.876

Πίνακας 6 *Μέτρα αξιολόγησης του μοντέλου, το οποίο κατασκευάστηκε με βάση τον αλγόριθμο OneR*

3.2.3 J48

Ο αλγόριθμος J48 είναι η εκδοχή του C4.5, που συναντούμε στο Weka. Ανήκει στους κατηγοριοποιητές δένδρων απόφασης και η λειτουργία του βασίζεται σε μία στατιστική ιδιότητα που ονομάζεται κέρδος πληροφορίας (information gain).

	Προβλεπόμενη τάξη	
Πραγματική τάξη	Θετικά	Αρνητικά
Θετικά	16845	457
Αρνητικά	2654	14648

Πίνακας 7 *Πίνακας σύγχυσης για την κατασκευή του μοντέλου με βάση τον αλγόριθμο J48*

Τάξη	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Θετικά	0.974	0.153	0.864	0.974	0.915	0.827	0.926	0.883
Αρνητικά	0.847	0.026	0.970	0.847	0.904	0.827	0.926	0.939
Μέση τιμή	0.910	0.090	0.917	0.910	0.910	0.827	0.926	0.911

Πίνακας 8 *Μέτρα αξιολόγησης του μοντέλου, το οποίο κατασκευάστηκε με βάση τον αλγόριθμο J48*

3.2.4 HoeffdingTree

Ο αλγόριθμος Hoeffding Tree, ανήκει στους κατηγοριοποιητές δένδρων απόφασης και χρησιμοποιεί το όριο του Hoeffding για την κατασκευή και ανάλυση των δένδρων απόφασης.

	Προβλεπόμενη τάξη	
Πραγματική τάξη	Θετικά	Αρνητικά
Θετικά	16833	469
Αρνητικά	2515	14787

Πίνακας 9 *Πίνακας σύγχυσης για την κατασκευή του μοντέλου με βάση τον αλγόριθμο HoeffdingTree*

Τάξη	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Θετικά	0.973	0.145	0.870	0.973	0.919	0.833	0.934	0.897
Αρνητικά	0.855	0.027	0.969	0.855	0.908	0.833	0.934	0.942
Μέση τιμή	0.914	0.086	0.920	0.914	0.913	0.833	0.934	0.919

Πίνακας 10 *Μέτρα αξιολόγησης του μοντέλου, το οποίο κατασκευάστηκε με βάση τον αλγόριθμο HoeffdingTree*

3.2.5 NaiveBayes

Ο αλγόριθμος NaiveBayes , ανήκει στους μπαϋεσιανούς κατηγοριοποιητές και αποτελεί ευθεία εφαρμογή του θεωρήματος Bayes.

	Προβλεπόμενη τάξη	
Πραγματική τάξη	Θετικά	Αρνητικά
Θετικά	16716	586
Αρνητικά	2373	14929

Πίνακας 11 *Πίνακας σύγχυσης για την κατασκευή του μοντέλου με βάση τον αλγόριθμο NaiveBayes*

Τάξη	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Θετικά	0.966	0.137	0.876	0.966	0.919	0.833	0.944	0.924
Αρνητικά	0.863	0.034	0.962	0.863	0.910	0.833	0.944	0.941
Μέση τιμή	0.914	0.086	0.919	0.914	0.914	0.833	0.944	0.933

Πίνακας 12 *Μέτρα αξιολόγησης του μοντέλου, το οποίο κατασκευάστηκε με βάση τον αλγόριθμο NaiveBayes*

3.2.6 BayesNet

Ο αλγόριθμος αυτός ανήκει στους μπαϋεσιανούς κατηγοριοποιητές και έχει την ιδιότητα να μαθαίνει από τα μπαυεσιανά δίκτυα (Bayesian networks).

	Προβλεπόμενη τάξη		
Πραγματική τάξη		Θετικά	Αρνητικά
	Θετικά	16715	587
	Αρνητικά	2373	14929

Πίνακας 13 *Πίνακας σύγχυσης για την κατασκευή του μοντέλου με βάση τον αλγόριθμο BayesNet*

Τάξη	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Θετικά	0.966	0.137	0.876	0.966	0.919	0.833	0.944	0.924
Αρνητικά	0.863	0.034	0.962	0.863	0.910	0.833	0.944	0.941
Μέση τιμή	0.914	0.086	0.919	0.914	0.914	0.833	0.944	0.933

Πίνακας 14 *Μέτρα αξιολόγησης του μοντέλου, το οποίο κατασκευάστηκε με βάση τον αλγόριθμο BayesNet*

3.2.7 Logistic

Η λογιστική παλινδρόμηση αποτελεί ένα μοντέλο ταξινόμησης με βάση τη θεωρία πιθανοτήτων.

	Προβλεπόμενη τάξη		
Πραγματική τάξη		Θετικά	Αρνητικά
	Θετικά	16738	564

	Αρνητικά	2332	14970

Πίνακας 15 *Πίνακας σύγχυσης για την κατασκευή του μοντέλου με βάση τον αλγόριθμο Logistic*

Τάξη	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Θετικά	0.967	0.135	0.878	0.967	0.920	0.837	0.948	0.923
Αρνητικά	0.865	0.033	0.964	0.865	0.912	0.837	0.948	0.957
Μέση τιμή	0.916	0.084	0.921	0.916	0.916	0.837	0.948	0.940

Πίνακας 16 *Μέτρα αξιολόγησης του μοντέλου, το οποίο κατασκευάστηκε με βάση τον αλγόριθμο Logistic*

3.2.8 RandomForest

Κάθε σύνολο εκπαίδευσης δημιουργείται με τυχαία επαναδειγματοληψία του αρχικού συνόλου δεδομένων. Κάθε σύνολο εκπαίδευσης χρησιμοποιείται για την εκπαίδευση ενός δένδρου απόφασης.

	Προβλεπόμενη τάξη		
Πραγματική τάξη		Θετικά	Αρνητικά
	Θετικά	16850	452
	Αρνητικά	2337	14965

Πίνακας 17 *Πίνακας σύγχυσης για την κατασκευή του μοντέλου με βάση τον αλγόριθμο RandomForest*

Τάξη	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Θετικά	0.974	0.135	0.878	0.974	0.924	0.844	0.964	0.960
Αρνητικά	0.865	0.026	0.971	0.865	0.915	0.844	0.964	0.967
Μέση τιμή	0.919	0.081	0.924	0.919	0.919	0.844	0.964	0.964

Πίνακας 18 *Μέτρα αξιολόγησης του μοντέλου, το οποίο κατασκευάστηκε με βάση τον αλγόριθμο RandomForest*

3.3 Περαιτέρω επεξεργασία με RandomForest

3.3.1 Μείωση sequence window

Ακολούθησε σταδιακή αφαίρεση των χαρακτηριστικών σε κάθε άκρο του sequence window με σκοπό την εύρεση εκείνου που θα εκπαιδεύσει τον

αλγόριθμο με την μεγαλύτερη ορθότητα πρόβλεψης. Μετά από διάφορες δοκιμές (**Πίνακας *19***) καταλήξαμε στο ότι το sequence window με το μέγιστο ποσοστό πρόβλεψης και δη 92.0761%, είναι εκείνο που διαθέτει 25 χαρακτηριστικά με το τροποποιημένο αμινοξύ (N) πάντα στη μέση, δηλαδή στη θέση 15. Τα καινούρια δεδομένα παρουσιάζονται στον **Πίνακας *21*** και στον **Πίνακας *21***.

Sequence window	Ποσοστό επιτυχούς πρόβλεψης (%)
29	91.9402
27	92.0327
25	92.0761
23	92.0067
21	92.0529
19	92.0616
17	91.9951
15	92.0327
13	92.0154
11	91.8709
9	91.8449
7	91.7697
5	90.163
3	59.4382

Πίνακας 19 *Μεταβολή ποσοστών επιτυχούς πρόβλεψης του μοντέλου με βάση το μήκος του sequence window*

	Προβλεπόμενη τάξη		
Πραγματική τάξη		Θετικά	Αρνητικά
	Θετικά	16878	424
	Αρνητικά	2318	14984

Πίνακας 20 *Πίνακας σύγχυσης του μοντέλου, το οποίο κατασκευάστηκε με βάση τον αλγόριθμο RandomForest και μοτίβου 25 αμινοξέων*

Τάξη	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Θετικά	0.975	0.134	0.879	0.975	0.925	0.847	0.964	0.961
Αρνητικά	0.866	0.025	0.972	0.866	0.916	0.847	0.964	0.966
Μέση τιμή	0.921	0.079	0.926	0.921	0.921	0.847	0.964	0.964

Πίνακας 21 *Μέτρα αξιολόγησης του μοντέλου, το οποίο κατασκευάστηκε με βάση τον αλγόριθμο RandomForest και μοτίβου 25 αμινοξέων*

3.3.2 Εκτίμηση της απόδοσης με τη μέθοδο holdout

Επιπλέον, δοκιμάστηκε η δημιουργία συνόλου ελέγχου μέσω της μεθόδου holdout, όπως περιγράφηκε στην ενότητα 1.3.4.1. Διαπιστώθηκε ότι με τον

διαχωρισμό των δεδομένων σε σύνολο εκπαίδευσης και σύνολο ελέγχου με αναλογία 70-30, πραγματοποιείται πρόβλεψη του κατηγοριοποιητή, η οποία λαμβάνει πλέον την τιμή 92.0817%. Τα αποτελέσματα απεικονίζονται στον ***Πίνακας 22*** και στον **Πίνακας *23***.

		Προβλεπόμενη τάξη	
Πραγματική τάξη		Θετικά	Αρνητικά
	Θετικά	5039	128
	Αρνητικά	694	4520

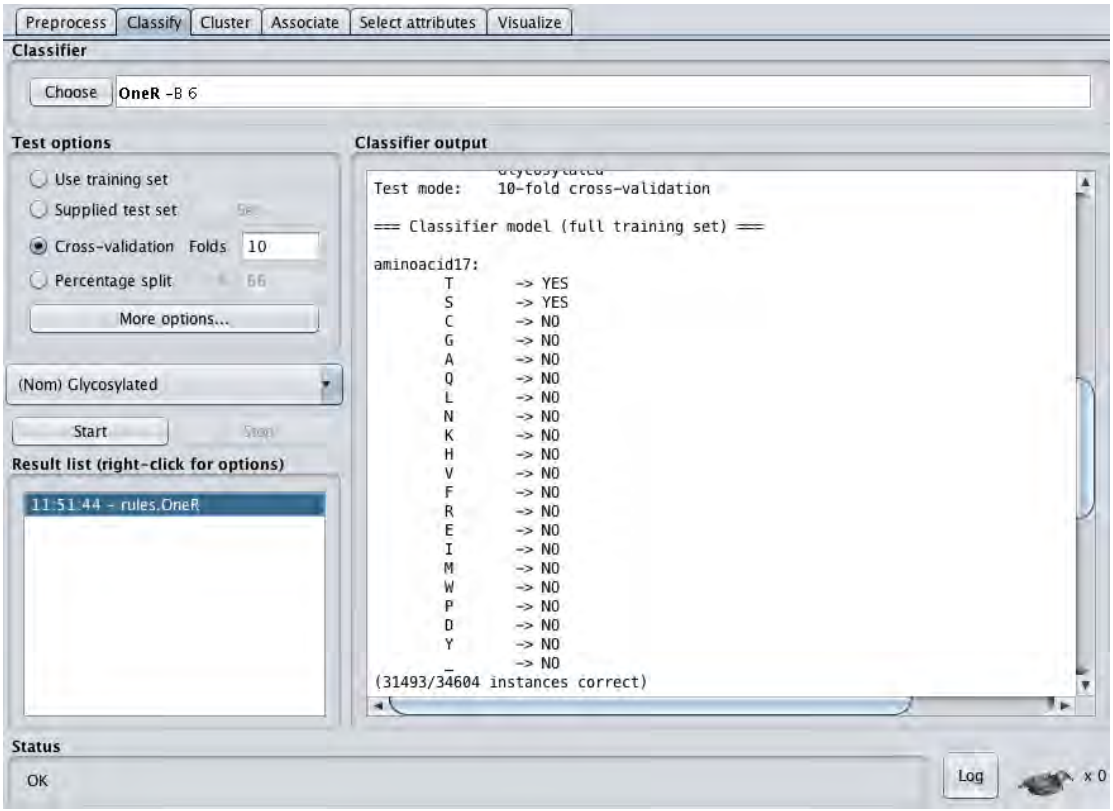
Πίνακας 22 *Πίνακας σύγχυσης του μοντέλου, το οποίο κατασκευάστηκε με βάση τον αλγόριθμο RandomForest, μοτίβου 25 αμινοξέων και holdout μεθόδου διαχωρισμού των δεδομένων*

Τάξη	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Θετικά	0.975	0.133	0.879	0.975	0.925	0.847	0.964	0.959
Αρνητικά	0.867	0.025	0.972	0.867	0.917	0.847	0.964	0.967
Μέση τιμή	0.921	0.079	0.926	0.921	0.921	0.847	0.964	0.963

Πίνακας 23 *Μέτρα αξιολόγησης του μοντέλου, το οποίο κατασκευάστηκε με βάση τον αλγόριθμο RandomForest, μοτίβου 25 αμινοξέων και holdout μεθόδου διαχωρισμού των δεδομένων*

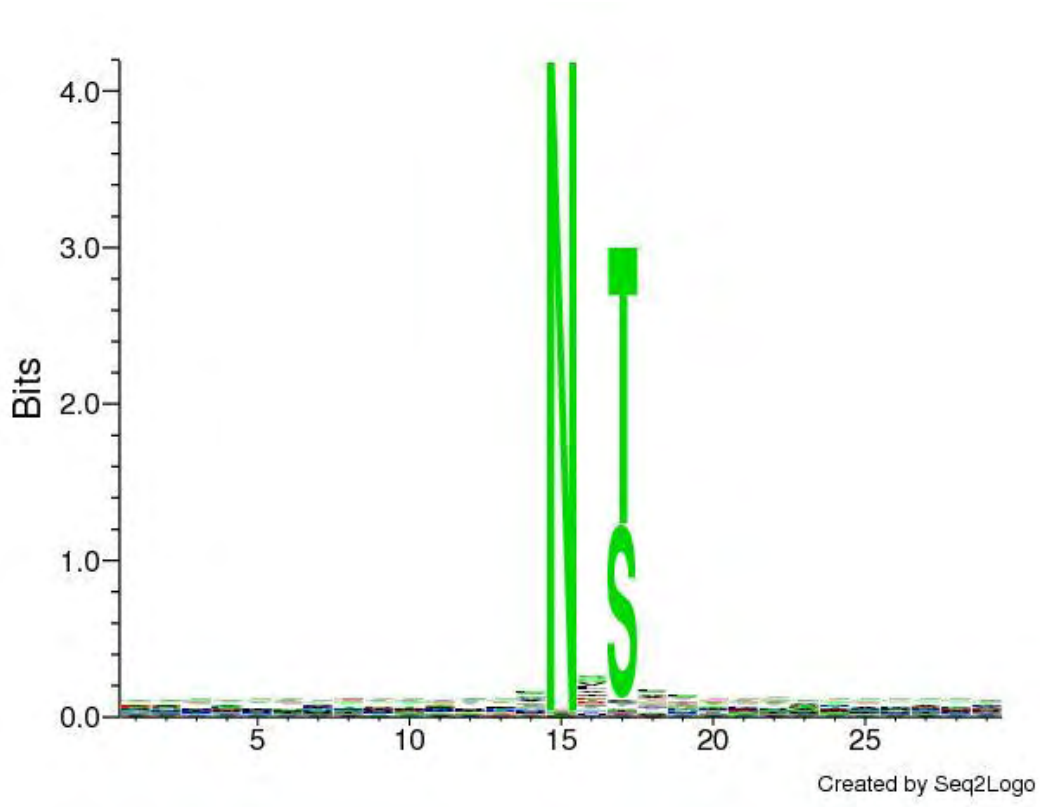
3.4 Προσδιορισμός σημαντικότερων χαρακτηριστικών

Η πολύ καλή απόδοση του κατηγοριοποιητή κανόνων OneR (3.2.2) μας οδήγησε στην ανίχνευση του πιο καθοριστικού χαρακτηριστικού. Σύμφωνα με τα αποτελέσματα του συγκεκριμένου μοντέλου, η κατηγοριοποίηση έγινε βασισμένη εξ'ολοκλήρου στο χαρακτηριστικό 17, δηλαδή δύο θέσεις μετά την Asn. Ειδικότερα, θετική κατηγορία αποδιδόταν στην εγγραφή όποτε αυτό ήταν Ser ή Thr, ενώ στις υπόλοιπες περιπτώσεις κατηγοριοποιούταν ως αρνητικό (**Εικόνα *11***).



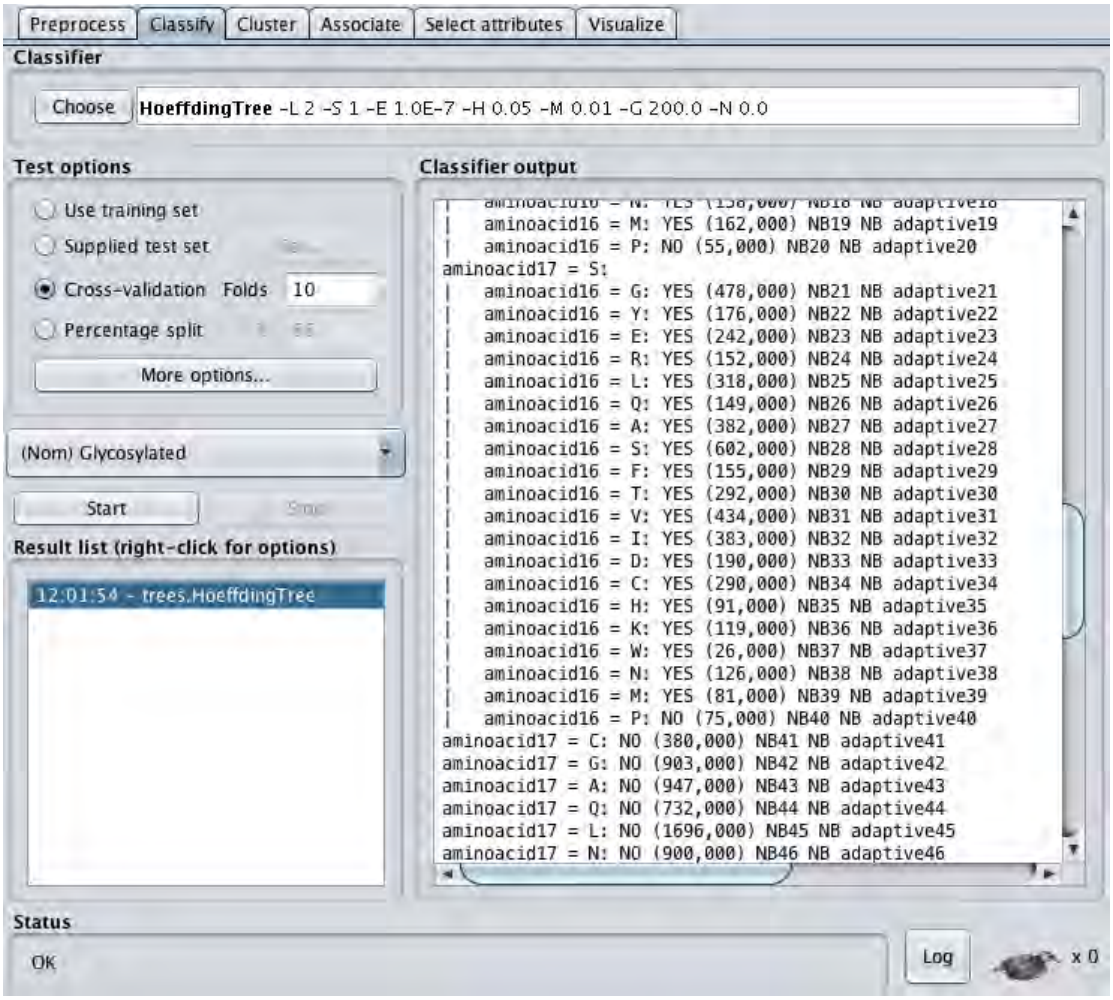
Εικόνα 11 *Τμήμα αποτελεσμάτων κατηγοριοποίησης με τον αλγόριθμο OneR. Απεικονίζεται το καθοριστικότερο χαρακτηριστικό για το αν μια αλληλουχία θα τροποποιηθεί με N-γλυκοζυλίωση*

Αυτό επιβεβαιώθηκε και από τη δημιουργία του Logo για το θετικό σύνολο δεδομένων, στο site http://www.cbs.dtu.dk/biotools/Seq2Logo/ επιλέγοντας για το logo type την επιλογή Shannon. Παρατηρούμε ότι δύο θέσεις μετά τη γλυκοζυλιωμένη ασπαραγίνη βρίσκεται πάντα είτε το αμινοξύ θρεονίνη, είτε το αμινοξύ σερίνη (**Εικόνα 12**).



Εικόνα 12 *Δημιουργία Logo για το θετικό σύνολο δεδομένων, τύπου Shannon* (πηγή: **http://www.cbs.dtu.dk/biotools/Seq2Logo/**)

Περαιτέρω ανάλυση και με τα υπόλοιπα μοντέλα αποκάλυψε ακόμη ένα σημαντικό χαρακτηριστικό. Πιο συγκεκριμένα, στην κατηγοριοποίηση με τον αλγόριθμο του HoeffdingTree, παρ΄όλο που στις περισσότερες των περιπτώσεων που το αμινοξύ δύο θέσεις μετά την Asn ήταν Ser/Thr, η εγγραφή κατηγοριοποιούταν ως θετική, όταν το αμινοξύ μία θέση μετά την Asn ήταν Pro (Προλίνη), η εγγραφή κατηγοριοποιούταν ως αρνητική, άρα ότι δεν υφίσταται Ν-γλυκοζυλίωση (**Εικόνα 13**).



Εικόνα 13 *Τμήμα των αποτελεσμάτων κατηγοριοποίησης με τον αλγόριθμο HoeffdingTree. Φαίνεται η απόδοση αρνητικής κατηγορίας στην εγγραφή όταν το αμινοξύ στη θέση 16 είναι προλίνη*

Καταλήγουμε λοιπόν στο ότι, όταν η αλληλουχία είναι NP!S/T (όπου P! οποιοδήποτε αμινοξύ εκτός της προλίνης) η αλληλουχία αποτελεί μία πιθανή θέση Ν-συνδεδεμένης γλυκοζυλίωσης. Ωστόσο δεν μπορούμε να αποφανθούμε με σιγουριά αν αυτούς του είδους η μετα-μεταφραστική τροποποίηση θα λάβει σίγουρα χώρα καθώς υπάρχει ένα ευρύ φάσμα τροποποιήσεων που έχει ως στόχο το ίδιο αμινοξικό κατάλοιπο (Asn) ή τα γειτονικά της (Ser,Thr) (***Εικόνα 6***), με συχνότερη τη φωσφορυλίωση των αμινοξέων Ser, Thr ή την ακετυλίωση της Ser.

3.5 Σύγκριση απόδοσης με άλλα δημοσιευμένα εργαλεία πρόβλεψης θέσεων Ν-γλυκοζυλίωσης

Στον **Πίνακας *24*** φαίνονται οι τιμές των διαφόρων μεθόδων αξιολόγησης για δημοσιευμένα εργαλεία πρόβλεψης θέσεων Ν-συνδεδεμένης γλυκοζυλίωσης, τα οποία περιλαμβάνουν το SPRINT-Gly (Taherzadeh et al., 2019), το GPP (Hamby and Hirst, 2008), το Glycomine (Li et al., 2015), το NetNGlyc (Blom et al., 2004) και το GlycoPP (Chauhan et al., 2012). Την καλύτερη απόδοση εμφανίζει το πιο πρόσφατο δημοσιευμένο εργαλείο (SPRINT-Gly) με τιμές για

τον συντελεστή MCC 0.939, Roc Area 0.984, Accuracy 0.978, Sensitivity 0.978, Specificity 0.979 και Precision 0.927.
Η δική μας μέθοδος με τη χρήση του αλγορίθμου RandomForest και 10-fold Cross Validation παρουσίασε MCC 0.847, Roc Area 0.964, Accuracy 0.92, Sensitivity 0.975, Specificity 0.866 και Precision 0.879, δηλαδή καλύτερα από τα περισσότερα διαθέσιμα εργαλεία. Ωστόσο, τα αποτελέσματα αυτά δεν είναι άμεσα συγκρίσιμα καθώς εφαρμόστηκε διαφορετικό σύνολο ελέγχου για το μοντέλο στα πλαίσια του Weka.

Methods		MCC	AUC	Acc	Sn	Sp	Pr
N-linked Human	SPRINT-Gly	0.939	0.984	0.978	0.978	0.979	0.927
	GPP	0.534	0.823	0.758	0.936	0.711	0.465
	GlycoMine	0.235	0.763	0.793	0.076	0.999	0.946
	NetNGlyc	0.794	0.95	0.934	0.745	0.985	0.928
	GlycoPP	0.742	0.952	0.918	0.91	0.919	0.682

Πίνακας 24 *Τιμές μέτρων αξιολόγησης απόδοσης για δημοσιευμένα εργαλεία πρόβλεψης N-γλυκοζυλίωσης (AUC: ROC Area, Sn: Sensitivity, Sp: Specificity, Pr: Precision)* **(Taherzadeh et al., 2019)**

3.6 Συμπέρασμα

Από όλες τις μεθόδους που δοκιμάσαμε στη σουίτα λογισμικού Weka, το καλύτερο ήταν ο ensemble κατηγοριοποιητής που κατασκευάστηκε βάσει του αλγορίθμου RandomForest. Εξίσου καλά απέδωσε και ο OneR το οποίο μας οδήγησε στη διευκρίνιση του χαρακτηριστικού που παίζει τόσο μεγάλο ρόλο. Η δημιουργία του Logo από θετικά δεδομένα έδειξε ότι δύο θέσεις μετά τη γλυκοζυλιωμένη ασπαραγίνη βρίσκεται σχεδόν πάντοτε ένα από τα αμινοξέα σερίνη ή θρεονίνη. Το μοντέλο που κατασκευάστηκε από τον αλγόριθμο HoeffdingTree υπέδειξε ακόμη ένα σημαντικό χαρακτηριστικό, το αμινοξύ ακριβώς δίπλα από τη γλυκοζυλιωμένη ασπαραγίνη, το οποίο δεν είναι ποτέ προλίνη. Ως εκ τούτου, η τελικά αλληλουχία συναίνεσης είναι NP!S/T (όπου P! οποιοδήποτε αμινοξύ εκτός της προλίνης), πράγμα που συμφωνεί με τη βιβλιογραφία (Aebi, 2013). Αυτό πιθανότατα είναι απόρροια του ενζύμου που καταλύει τη N-συνδεδεμένη γλυκοζυλίωση, δηλαδή της ολιγοσακχαρυλοτρανσφεράσης και δη της θέσης εξειδίκευσής της. Ωστόσο έχουν βρεθεί και μοτίβα τα οποία δεν συμφωνούν με το προαναφερθέν (Lowenthal et al., 2016).
Παρά την ύπαρξη μεθόδων που επιτρέπουν την ταυτοποίηση τροποποιημένων πεπτιδίων με ένα μόνο πείραμα, απέχουμε ακόμη πολύ από την κάλυψη ολόκληρων πρωτεωμάτων. Ετσι, οι υπολογιστικές προσεγγίσεις αποτελούν μία καλή εναλλακτική, προσφέροντας παράλληλα ταχύτητα, ακρίβεια και άνεση. Επιπλέον, μας βοηθούν να κατανοήσουμε ποιά είναι τα χαρακτηριστικά που καθορίζουν αν ένα αμινοξύ θα γλυκοζυλιωθεί ή όχι.

4 Βιβλιογραφία

Aebi, M. (2013). N-linked protein glycosylation in the ER. Biochim. Biophys. Acta BBA - Mol. Cell Res. *1833*, 2430–2437.

Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics *4*, 1633–1649.

Caragea, C., Sinapov, J., Silvescu, A., Dobbs, D., and Honavar, V. (2007). Glycosylation site prediction using ensembles of Support Vector Machine classifiers. BMC Bioinformatics *8*, 438.

Catalanotto, C., Cogoni, C., and Zardo, G. (2016). MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions. Int. J. Mol. Sci. *17*.

Chauhan, J.S., Bhat, A.H., Raghava, G.P.S., and Rao, A. (2012). GlycoPP: A Webserver for Prediction of N- and O-Glycosites in Prokaryotic Protein Sequences. PLoS ONE *7*.

Chauhan, J.S., Rao, A., and Raghava, G.P.S. (2013). In silico Platform for Prediction of N-, O- and C-Glycosites in Eukaryotic Protein Sequences. PLOS ONE *8*, e67008.

Chen, R., Jiang, X., Sun, D., Han, G., Wang, F., Ye, M., Wang, L., and Zou, H. (2009). Glycoproteomics analysis of human liver tissue by combination of multiple enzyme digestion and hydrazide chemistry. J. Proteome Res. *8*, 651–661.

Cheng, K., Chen, R., Seebun, D., Ye, M., Figeys, D., and Zou, H. (2014). Large-scale characterization of intact N-glycopeptides using an automated glycoproteomic method. J. Proteomics *110*, 145–154.

Chou, T.Y., Hart, G.W., and Dang, C.V. (1995). c-Myc is glycosylated at threonine 58, a known phosphorylation site and a mutational hot spot in lymphomas. J. Biol. Chem. *270*, 18961–18965.

Cox, M., and Nelson, D. (2000). Lehninger Principles of Biochemistry.

Dhanisha, S.S., Guruvayoorappan, C., Drishya, S., and Abeesh, P. (2018). Mucins: Structural diversity, biosynthesis, its role in pathogenesis and as possible therapeutic targets. Crit. Rev. Oncol. Hematol. *122*, 98–122.

Ford, L.P., Bagga, P.S., and Wilusz, J. (1997). The poly(A) tail inhibits the assembly of a 3’-to-5’ exonuclease in an in vitro RNA stability system. Mol. Cell. Biol. *17*, 398–406.

Hamby, S.E., and Hirst, J.D. (2008). Prediction of glycosylation sites using random forests. BMC Bioinformatics *9*, 500.

Hart, G.W., Slawson, C., Ramirez-Correa, G., and Lagerlof, O. (2011). Cross Talk Between O-GlcNAcylation and Phosphorylation: Roles in Signaling, Transcription, and Chronic Disease. Annu. Rev. Biochem. *80*, 825–858.

Kaji, H., Shikanai, T., Sasaki-Sawa, A., Wen, H., Fujita, M., Suzuki, Y., Sugahara, D., Sawaki, H., Yamauchi, Y., Shinkawa, T., et al. (2012). Large-scale identification of N-glycosylated proteins of mouse tissues and construction of a glycoprotein database, GlycoProtDB. J. Proteome Res. *11*, 4553–4566.

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

Kotsiantis, S.B. (2007). Supervised Machine Learning: A Review of Classification Techniques. 20.

Kudelka, M.R., Ju, T., Heimburg-Molinaro, J., and Cummings, R.D. (2015). Simple sugars to complex disease--mucin-type O-glycans in cancer. Adv. Cancer Res. *126*, 53–135.

Li, F., Li, C., Wang, M., Webb, G.I., Zhang, Y., Whisstock, J.C., and Song, J. (2015). GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome. Bioinformatics *31*, 1411–1419.

Lowenthal, M.S., Davis, K.S., Formolo, T., Kilpatrick, L.E., and Phinney, K.W. (2016). Identification of novel N-glycosylation sites at non-canonical protein consensus motifs. J. Proteome Res. *15*, 2087–2101.

McManus, J., Cheng, Z., and Vogel, C. (2015). Next-generation analysis of gene expression regulation – comparing the roles of synthesis and degradation. Mol. Biosyst. *11*, 2680–2689.

Molinari, F., Foulquier, F., Tarpey, P.S., Morelle, W., Boissel, S., Teague, J., Edkins, S., Futreal, P.A., Stratton, M.R., Turner, G., et al. (2008). Oligosaccharyltransferase-Subunit Mutations in Nonsyndromic Mental Retardation. Am. J. Hum. Genet. *82*, 1150–1157.

Parker, R., and Song, H. (2004). The enzymes and control of eukaryotic mRNA turnover. Nat. Struct. Mol. Biol. *11*, 121–127.

Parker, B.L., Palmisano, G., Edwards, A.V.G., White, M.Y., Engholm-Keller, K., Lee, A., Scott, N.E., Kolarich, D., Hambly, B.D., Packer, N.H., et al. (2011). Quantitative N-linked glycoproteomics of myocardial ischemia and reperfusion injury reveals early remodeling in the extracellular environment. Mol. Cell. Proteomics MCP *10*, M110.006833.

Ramanathan, A., Robb, G.B., and Chan, S.-H. (2016). mRNA capping: biological functions and applications. Nucleic Acids Res. *44*, 7511–7526.

Spoel, S.H. (2018). Orchestrating the proteome with post-translational modifications. J. Exp. Bot. *69*, 4499–4503.

Taherzadeh, G., Dehzangi, A., Golchin, M., Zhou, Y., and Campbell, M.P. (2019). SPRINT-Gly: predicting N- and O-linked glycosylation sites of human and mouse proteins by using sequence and predicted structural properties. Bioinformatics.

Wang, Z., Udeshi, N.D., Slawson, C., Compton, P.D., Sakabe, K., Cheung, W.D., Shabanowitz, J., Hunt, D.F., and Hart, G.W. (2010). Extensive Crosstalk Between O-GlcNAcylation and Phosphorylation Regulates Cytokinesis. Sci. Signal. *3*, ra2.

Wight, T.N., Kinsella, M.G., and Qwarnström, E.E. (1992). The role of proteoglycans in cell adhesion, migration and proliferation. Curr. Opin. Cell Biol. *4*, 793–801.

Yoon, J.-H., Abdelmohsen, K., and Gorospe, M. (2013). Post-transcriptional gene regulation by long noncoding RNA. J. Mol. Biol. *425*, 3723–3730.

Zhu, J., Sun, Z., Cheng, K., Chen, R., Ye, M., Xu, B., Sun, D., Wang, L., Liu, J., Wang, F., et al. (2014). Comprehensive mapping of protein N-glycosylation in human liver by combining hydrophilic interaction chromatography and hydrazide chemistry. J. Proteome Res. *13*, 1713–1721.

Zielinska, D.F., Gnad, F., Wiśniewski, J.R., and Mann, M. (2010). Precision Mapping of an In Vivo N-Glycoproteome Reveals Rigid Topological and Sequence Constraints. Cell *141*, 897–907.

Zielinska, D.F., Gnad, F., Schropp, K., Wiśniewski, J.R., and Mann, M. (2012). Mapping N-glycosylation sites across seven evolutionarily distant species reveals a divergent substrate proteome despite a common core machinery. Mol. Cell *46*, 542–548.