Πανεπιστήμιο Θεσσαλίας

Σχολή Επιστημών Υγείας

Τμήμα Βιοχημείας και Βιοτεχνολογίας

Πρόβλεψη θέσεων μεθυλίωσης σε ευκαρυωτικές πρωτεΐνες με τη βοήθεια αλγορίθμων μηχανικής μάθησης

Διπλωματική εργασία φοιτήτριας: Βασιλικής Φλιάτουρα
Επιβλέπων καθηγητής: Δρ. Γρηγόριος Αμούτζιας

University of Thessaly

School of Health Sciences

Dep. of Biochemistry and Biotechnology

Prediction of methylation sites in eukaryotic proteins with machine learning algorithms

Diploma Thesis of the Student: Vasiliki Fliatoura
Supervisor Professor: Dr. Grigorios Amoutzias

Η παρούσα διπλωματική εργασία εκπονήθηκε στο εργαστήριο Βιοπληροφορικής, του τμήματος Βιοχημείας και Βιοτεχνολογίας (ΤΒΒ), της Σχολής Επιστημών Υγείας του Πανεπιστημίου Θεσσαλίας (Π.Θ.)

Υπεύθυνος καθηγητής

Αμούτζιας Γρηγόριος, Επίκουρος Καθηγητής Βιοπληροφορικής στη Γενωμική, ΤΒΒ, Π.Θ.

Τριμελής επιτροπή

Αμούτζιας Γρηγόριος, Επίκουρος Καθηγητής Βιοπληροφορικής στη Γενωμική, ΤΒΒ, Π.Θ.

Νικόλαος Παπανικολάου, Επίκουρος Καθηγητής Βιοχημείας, Τμήμα Ιατρικής, Α.Π.Θ.

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

Περιεχόμενα

Περιεχόμενα εικόνων

Περιεχόμενα πινάκων

Ευχαριστίες

Θα ήθελα πρώτα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου, κ. Αμούτζια Γρηγόριο, Επίκουρο Καθηγητή Βιοπληροφορικής στη Γενωμική, ΤΒΒ, Π.Θ., για την ευκαιρία που μου έδωσε να εκπονήσω τη διπλωματική μου εργασία στο εργαστήριο του, δείχνοντας εμπιστοσύνη ως προς το πρόσωπό μου. Είμαι ευγνώμων για τις γνώσεις που μου προσέφερε καθώς και για την έμπειρη καθοδήγησή του καθ' όλη τη διάρκεια εκπόνησης της διπλωματικής μου. Επίσης, ευχαριστώ πολύ τους καθηγητές κ. Παπανικολάου Νικόλαο, Επίκουρο Καθηγήτη Βιοχημείας του Τμήματος Ιατρικής, Α.Π.Θ και κ. Ηλιόπουλο Ιωάννη, Επίκουρο Καθηγητή Μοριακής Βιολογίας-Γονιδιωματικής Βιοπληροφορικής του Τμήματος Ιατρικής του Πανεπιστημίου Κρήτης που συμμετείχαν στην τριμελή επιτροπή της εργασίας μου.

Στην συνέχεια θα ήθελα να ευχαριστήσω ιδιαίτερα την προπτυχιακή φοιτήτρια και καλή μου φίλη Ντουντούμη Χρύσα καθώς και τον υποψήφιο διδάκτορα Πάνο Βλασταρίδη για την αμέριστη βοήθειά τους, χωρίς την οποία δεν θα ήταν δυνατή η διεκπεραίωση αυτής της εργασίας. Ευχαριστώ και τα υπόλοιπα παιδιά του εργαστηρίου, τους προπτυχιακούς φοιτητές Νικολαΐδη Μάριο και Διακογεωργίου Αλεξάνδρα, για την υποστήριξη και τις ωραίες στιγμές.

Επίσης θα ήθελα να ευχαριστήσω πολύ την οικογένεια μου για την ηθική, ψυχολογική αλλά και οικονομική στήριξη που μου παρείχαν όλα αυτά τα χρόνια που φοιτώ στο τμήμα Βιοχημείας και Βιοτεχνολογίας του Π.Θ.

Λένε πως οι φίλοι είναι η οικογένεια που διαλέγουμε. 'Ετσι, δε θα μπορούσα να μην αναφέρω τους ανθρώπους που όλα αυτά τα χρόνια ήταν «εκεί» για μένα και μοιραστήκαμε ανησυχίες και όμορφες στιγμές. Ευχαριστώ, λοιπόν, μέσα απ' την καρδιά μου τον Τσιρνόβα Θάνο, την Παπαζλατάνη Χριστίνα, την Τσιλιπουνιδάκη Κατερίνα, την Τρικαλινού Κωνσταντίνα και ιδιαίτερα την Τσιάνου Βούλα, τη Χριστίδου Ουρανία, τη Σουλτσιώτη Μαρία, τη Μιχαλοπούλου Μαρία και την Ντουντούμη Χρύσα, για τη στήριξη και την αγάπη τους.

Περίληψη

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. Παρ' όλα αυτά, μέχρι στιγμής έχει βρεθεί μόνο το 20-40% του συνολικού μεθυλ-πρωτεώματος στον άνθρωπο, ενώ για πολλούς άλλους οργανισμούς-μοντέλα, τα δεδομένα είναι ακόμα πολύ λίγα. Έτσι, υπάρχει ανάγκη για ένα εργαλείο βιοπληροφορικής που να μπορεί να προβλέψει με ακρίβεια και ταχύτητα αυτές τις θέσεις μεθυλίωσης σε ολόκληρο το πρωτέωμα, ώστε να βοηθήσει στην καθοδήγηση μελλοντικών πειραμάτων σε αυτόν τον τομέα. Χρησιμοποιώντας τα πιο πρόσφατα δεδομένα του ανθρώπου και εφαρμόζοντας αυστηρά κριτήρια φιλτραρίσματος, αναπτύχθηκε ένας server-νευρωνικό δίκτυο πρόβλεψης που ονομάζεται Methyl-Prometheus, ο οποίος προβλέπει θέσεις μεθυλίωσης Λυσίνης και Αργινίνης σε πρωτεϊνικές αλληλουχίες. Αυτός ο server έχει ακρίβεια της τάξης του 81% στους ανθρώπους και 79% στον ποντικό. Επίσης, εμφανίζει υψηλή ακρίβεια της τάξης του 86% και 89% σε δύο σχετικά περιορισμένα πειραματικά δεδομένα από τον ζυμομύκητα *S. cerevisiae* και το Apicomplexan *Toxoplasma gondii*, που μοιράστηκαν έναν κοινό πρόγονο με τους ανθρώπους πριν από περισσότερα από ένα δισεκατομμύριο χρόνια. Έτσι, ο Methyl-Prometheus αναμένεται να προβλέπει πρωτεϊνικές θέσεις μεθυλίωσης με υψηλή ακρίβεια στη συντριπτική πλειοψηφία των ευκαρυωτικών πρωτεωμάτων. Ο server είναι ελεύθερα προσβάσιμος στη διεύθυνση: http://bioinf.bio.uth.gr/methyl-prometheus/

Abstract

The latest advances in high-throughput (HTP) 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, whereas for many other model species, data are still rare. Thus, there is a need for a bioinformatics tool that may accurately and rapidly predict these methylation sites in a whole proteome, so as to help guide future experiments in this field. By utilizing up to the latest human HTP datasets and applying stringent filtering criteria, a neural network prediction server named Methyl-Prometheus has been developed, that predicts Lysine and Arginine methylation sites in protein sequences. This server has an overall accuracy of 81% in humans and 79% in mouse, two rather distantly related mammals that diverged 90 million years ago. It also displays a high accuracy of 86% and 89% in two rather limited experimental datasets from and the budding yeast *S. cerevisiae*, and the apicomplexan *Toxoplasma gondii* that shared a common ancestor with humans more than a billion years ago. Thus, Methyl-Prometheus is expected to predict protein methylation sites with high accuracy in the vast majority of eukaryotic proteomes. The server is freely available at: http://bioinf.bio.uth.gr/methyl-prometheus/

1. Εισαγωγή

1.1 Μετα-μεταφραστικές τροποποιήσεις

Είναι γνωστό, ότι σε πολλές από τις πρωτεΐνες είναι δυνατόν να συμβούν διάφορες μετα-μεταφραστικές τροποποιήσεις, οι οποίες αυξάνουν την πολυπλοκότητα τους. Ορισμένες από αυτές τις ομοιοπολικές τροποποιήσεις μπορούν να συμβούν είτε με πρωτεολυτική διάσπαση της πρωτεΐνης είτε με την προσθήκη μιας λειτουργικής ομάδας, π.χ. φωσφορικής ομάδας ή μεθυλομάδας, σε ένα ή περισσότερα αμινοξέα (Mann and Jensen 2003). Συνεπώς, οι μετα-μεταφραστικές τροποποιήσεις, με ένα σχετικά χαμηλό ενεργειακό κόστος, μεταβάλλουν τις λειτουργικές ιδιότητες των πρωτεϊνών, παίζοντας σημαντικό ρόλο στη ρύθμιση της λειτουργίας, της ενδοκυτταρικής κατανομής και των αλληλεπιδράσεων των πρωτεϊνών (Murn and Shi 2017; Mann and Jensen 2003).

1.2 Πρωτεϊνική μεθυλίωση

1.2.1 Η ανακάλυψη της μεθυλίωσης σε ζωντανά κύτταρα

Η πρώτη αναφορά για την ύπαρξη πρωτεϊνικής μεθυλίωσης σε ζωντανά κύτταρα έγινε το 1959 από τους ερευνητές (Ambler and Rees 1959), οι οποίοι, κατά την ανάλυση πρωτεϊνών από μαστίγια βακτηρίων, παρατήρησαν την ύπαρξη ενός άγνωστου αμινοξέος, το οποίο τελικά ταυτοποιήθηκε ως μεθυλιωμένη λυσίνη. Αργότερα, οι ερευνητές πρότειναν ότι η μεθυλίωση λαμβάνει χώρα μετα-μεταφραστικά και υπέθεσαν την ύπαρξη ένος ενζύμου που ίσως να καταλύει αυτή την τροποποίηση. Η πρωτεϊνική μεθυλίωση, αρχικά, συγκέντρωσε μεγάλο ερευνητικό ενδιαφέρον. Όμως, παρά τα ενδιαφέροντα ευρήματα δεν ευδοκίμησε σε εκείνη την αρχική φάση, λόγω της έλλειψης γνώσεων σχετικά με τη βιολογική της δράση. Ωστόσο, η πρόοδος στη μοριακή βιολογία και οι πολυάριθμες ανακαλύψεις που έλαβαν χώρα στο τέλος του 20^{ου} και στις αρχές του 21^{ου} αιώνα, αποκάλυψαν τον σημαντικό της ρόλο, συμβάλλοντας στην άνθηση του πεδίου της πρωτεϊνικής μεθυλίωσης (Murn and Shi 2017).

1.2.2 Ο μηχανισμός της μεθυλίωσης πρωτεϊνών

Η μεθυλίωση των πρωτεϊνών είναι μια αντιστρέψιμη μετα-μεταφραστική τροποποίηση, η οποία περιλαμβάνει την ομοιοπολική προσθήκη ενός, δύο ή τριών μεθυλομάδων στην πλευρική αλυσίδα αμινοξέων. Ανάμεσα στα αμινοξέα που έχει βρεθεί ότι μεθυλιώνονται βρίσκονται η λυσίνη, η αργινίνη, η ιστιδίνη, η ασπαραγίνη και η γλουταμίνη (Sylvestersen et al. 2014). Ωστόσο, τα αμινοξέα που μεθυλιώνονται κατά κύριο λόγο είναι η λυσίνη και η αργινίνη (Bremang et al. 2013). Η αντίδραση της μεθυλίωσης διαμεσολαβείται από τις πρωτεϊνικές μεθυλοτρανφεράσες (Protein Methyltransferases,

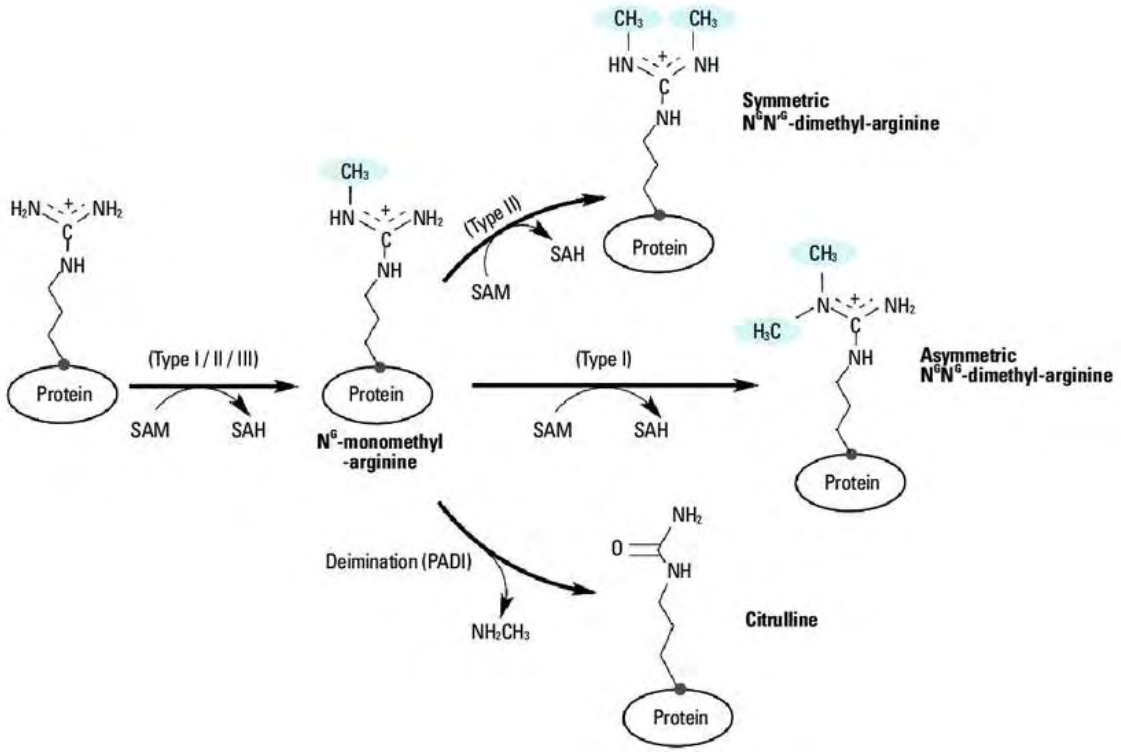
PMTs). Αυτά τα ένζυμα καταλύουν τη μεταφορά μίας μεθυλομάδας από την S-αδενοσυλ-L-μεθειονίνη στα προς μεθυλίωση πρωτεϊνικά υποστρώματα (Grillo and Colombatto 2005). Η αντίστροφη αντίδραση, δηλαδή η απομάκρυνση των μεθυλομάδων από τις μεθυλιωμένες πρωτεΐνες, καταλύεται από τις πρωτεϊνικές απομεθυλάσες (Protein Demethylases, PDMs) (Trojer and Reinberg 2006).

1.2.3 Η μεθυλίωση της αργινίνης

Η αργινίνη μπορεί να υποστεί μονομεθυλίωση (ω-N^G-μονομεθυλαργινίνη, MMA) ή διμεθυλίωση. Η διμεθυλίωση μπορεί να είναι ασύμμετρη, όταν και οι δύο μεθυλομάδες προστίθενται στο ίδιο άτομο αζώτου στο τέλος της πλευρικής αλυσίδας της αργινίνης (ω-N^G,N^G –ασύμμετρη-διμεθυλαργινίνη, ADMA) ή συμμετρική, όταν προστίθεται μία μεθυλομάδα σε κάθε ένα από τα δύο τελικά άτομα αζώτου (ω-N^G,N'^G-συμμετρική-διμεθυλαργίνινη, SDMA) (Blanc and Richard 2017; Lee and Stallcup 2009). Η μεθυλίωση της αργινίνης πραγματοποιείται από τις πρωτεϊνικές μεθυλοτρανσφεράσες της αργινίνης (Protein Arginine Methyltransferases, PRMTs). Όπως φαίνεται και στην Εικόνα 1, οι PRMTs διακρίνονται σε τρεις τύπους ανάλογα με το είδος της μεθυλίωσης που καταλύουν (Nicholson et al. 2015):

- Και οι τρεις τύποι ενζύμων μπορούν να οδηγήσουν στην παραγωγή μονομεθυλαργινίνης. Οι PRMTs τύπου III, όπως είναι η PRMT7, πραγματοποιεί μόνο μονομεθυλιώσεις.
- Οι PRMTs τύπου I, όπως είναι η PRMT4/CARM1, καταλύουν τη δημιουργία ασύμμετρων διμεθυλιώσεων της αργινίνης.
- Οι PRMTs τύπου II, όπως η PRMT5, πραγματοποιούν συμμετρικές διμεθυλιώσεις αργινίνης.

Έχει βρεθεί ότι οι PRMTs δείχνουν ιδιαίτερη «προτίμηση» και μεθυλιώνουν αργινίνες οι οποίες βρίσκονται εντός ή κοντά σε συγκεκριμένα μοτίβα αλληλουχιών. Σε αυτές τις αλληλουχίες συγκαταλέγονται τα μοτίβα που είναι πλούσια σε αργινίνη και γλυκίνη, τα οποία ονομάζονται μοτίβα RGG/RG ή περιοχές GAR (Glycine-Arginine-Rich). Αυτά τα μοτίβα παίζουν ρόλο τόσο στην πρόσδεση νουκλεϊκών οξέων όσο και στις αλληλεπιδράσεις μεταξύ των διάφορων πρωτεϊνών (Thandapani et al. 2013). Παρόλα αυτά, δεν μεθυλιώνουν όλες οι PRMTs μοτίβα RGG/RG. Ορισμένες μεθυλοτρανσφεράσες της αργινίνης, όπως είναι η PRMT4/CARM1, επιλέγουν αργινίνες που βρίσκονται κοντά σε μοτίβα πλούσια σε PGM (προλίνη, γλυκίνη και μεθειονίνη) (Yang and Bedford 2013). Τέλος, κάποιες άλλες, όπως η PRMT7, προτιμούν μοτίβα RxR που περιβάλλονται από αλληλουχίες πλούσιες σε λυσίνη (Feng et al. 2013).

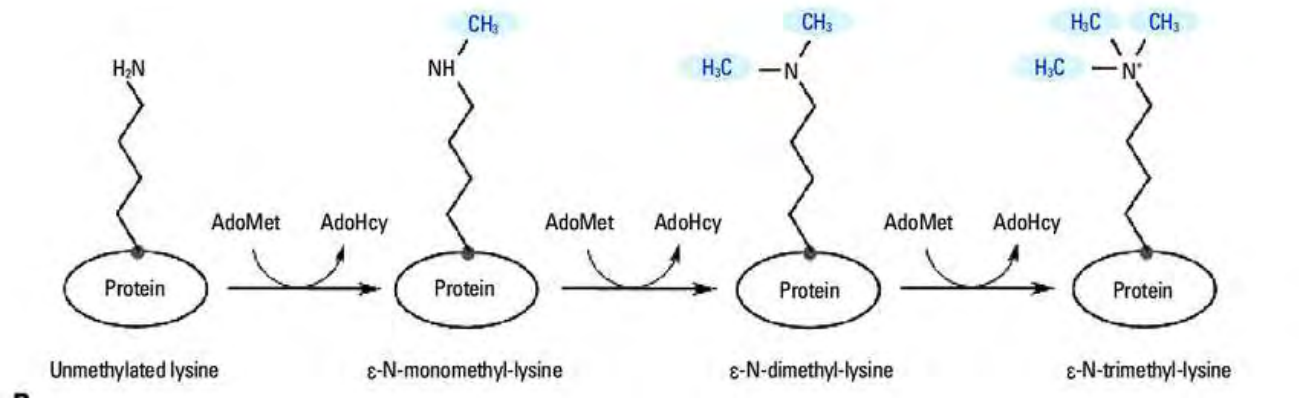


Εικόνα 1 *Οι τρεις τύποι μεθυλίωσης της αργινίνης και τα ένζυμα που καταλύουν κάθε αντίδραση* **(Paik, Kim, and Lim 2014)**.

1.2.4 Η μεθυλίωση της λυσίνης

Η μεθυλίωση της λυσίνης περιλαμβάνει την προσθήκη μίας, δύο ή τριών μεθυλομάδων στην ε-αμινομάδα της λυσίνης. Αυτή η αντίδραση καταλύεται από τις πρωτεϊνικές μεθυλοτρανσφεράσες της λυσίνης (Protein Lysine Methyltransferases, PKMTs) (Lee and Stallcup 2009; Guo et al. 2014). Οι ανθρώπινες PKMTs χωρίζονται σε δύο ομάδες, τάξη V και τάξη I. Οι PKMTs της τάξης V αποτελούν ένζυμα που περιέχουν την καταλυτική περιοχή SET. Η περιοχή SET πήρε το όνομα της από τις πρωτεΐνες SU(var), Enhancer of Zeste και Trithorax της Drosophila, στις οποίες είχε εντοπιστεί αρχικά (Tschiersch et al. 1994). Σε αυτή την περιοχή λαμβάνει χώρα η πρόσδεση του συμπαράγοντα SAM και του προς μεθυλίωση υποστρώματος (Min et al. 2002). Οι πρωτεΐνες που περιέχουν την περιοχή SET διακρίνονται σε επτά οικογένειες (Dillon et al. 2005): i) SUV3/9, ii) SET1, iii) SET2, iv) SMYD, v) EZ, vi) SUV4-20 και vii) RIZ.

Οι PKMTs της τάξης I ανήκουν σε μία υπεροικογένεια μεθυλοτρανσφερασών που έχουν βρεθεί σε ευκαρυώτες, προκαρυώτες αλλά και αρχαία. Μέλη αυτής της οικογένειας έχει βρεθεί ότι καταλύουν τη μεθυλίωση DNA, RNA ή αμινοξέων (Lanouette et al. 2014; Schubert et al. 2003).



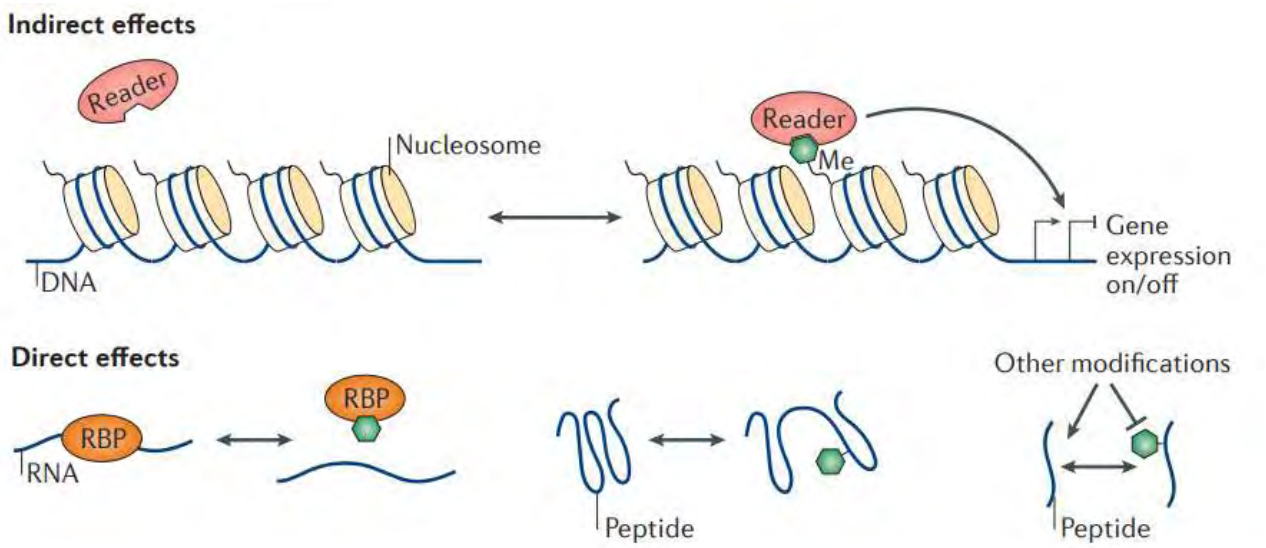
Εικόνα 2 *Οι τύποι μεθυλίωσης της λυσίνης* **(Paik, Kim, and Lim 2014)**.

1.3 Ο βιολογικός ρόλος της πρωτεϊνικής μεθυλίωσης

Η βιολογική δράση της πρωτεϊνικής μεθυλίωσης ασκείται, κατά κύριο λόγο, έμμεσα, μέσω της δημιούργιας επιφανειών πρόσδεσης πρωτεϊνικών επικρατειών, όπως είναι η περιοχή Tudor (Gayatri and Bedford 2014; Chen et al. 2011). Με αυτόν τον τρόπο, οι θέσεις μεθυλίωσης αναγνωρίζονται από πρωτεΐνες-τελεστές, οι οποίες αναφέρονται και ως «readers», οδηγώντας σε ποικίλες βιολογικές δράσεις. Για παράδειγμα, τέτοιες πρωτεΐνες αναγνωρίζουν θέσεις μεθυλίωσης στις ιστόνες και δρουν πραγματοποιώντας μεταγραφικές αλλαγές ή στρατολογώντας άλλες πρωτεΐνες. Ωστόσο, υπάρχουν ενδείξεις που υποστηρίζουν ότι η μεθυλίωση έχει και άμεσες δράσεις (Bedford and Richard 2005). Έχει αποδειχθεί ότι η μεθυλίωση της λυσίνης και της αργινίνης αυξάνει την υδροφοβικότητα των πλευρικών τους αλυσίδων, διευκολύνοντας την αλληλεπίδραση με νουκλεοτιδικές βάσεις του DNA, του RNA ή ακόμη με αρωματικά αμινοξέα (Evich et al. 2016) (

Εικόνα 3). Ωστόσο, μπορεί να έχει και αρνητική επίδραση στις αλληλεπιδράσεις πρωτεΐνης-πρωτεΐνης, όπως στην περίπτωση των ενζύμων που καταλύουν άλλες μετα-μεταφραστικές τροποποιήσεις (Beltran-Alvarez et al. 2015). Έτσι, η μεθυλίωση αυξάνει την δομική ποικιλομορφία και ρυθμίζει τη λειτουργία των πρωτεϊνών, παίζοντας σημαντικό ρόλο στις αλληλεπιδράσεις των πρωτεϊνών με άλλες πρωτεΐνες αλλά και με νουκλεϊκά οξέα.

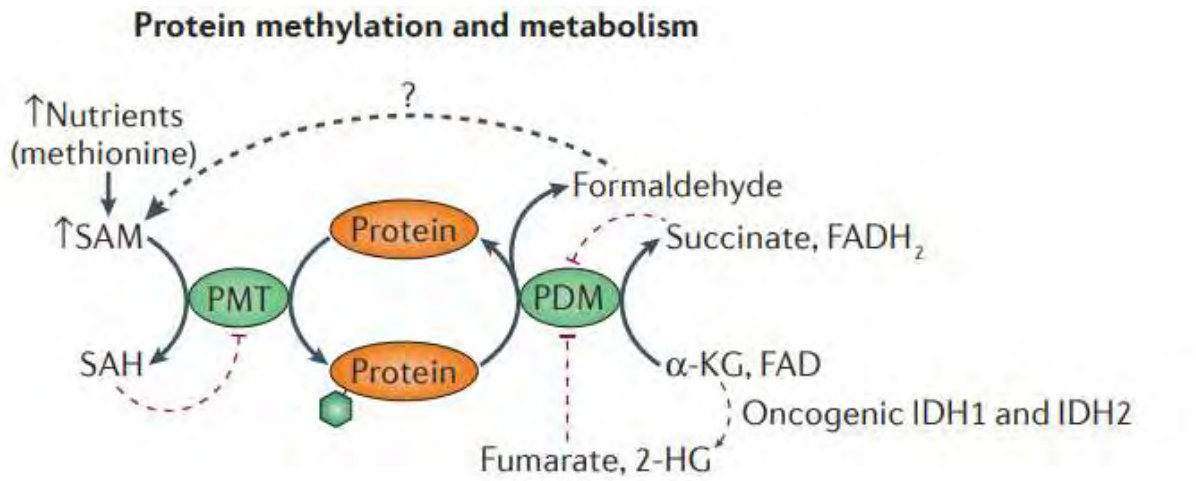
Η μεθυλίωση ρυθμίζει μία πληθώρα κυττατικών διεργασιών όπως την αναδιάταξη της χρωματίνης (Kouzarides 2007; Levy et al. 2011) και τη γονιδιακή μεταγραφή (Kontaki and Talianidis 2010) , δρώντας ως επιγενετικός ρυθμιστής, το μήκος και τη σταθερότητα των τελομερών (Mitchell et al. 2009), το μεταβολισμό του RNA (Kwak et al. 2003), τον κυτταρικό κύκλο (Carr et al. 2011), μεταγωγή σήματος (Biggar and Li 2015) και την πρωτεόλυση (Kontaki and Talianidis 2010).



Εικόνα 3 ***Βιολογική δράση πρωτεϊνικής μεθυλίωσης:*** *Έμμεση επίδραση, μέσω αναγνώρισης των θέσεων μεθυλίωσης από πρωτεΐνες-τελεστές. Στις ιστόνες τέτοιες πρωτεΐνες δρουν πραγματοποιώντας μεταγραφικές αλλαγές ή στρατολογώντας άλλες πρωτεΐνες για να το κάνουν. Άμεση δράση, μέσω διαμόρφωσης της αλληλεπίδρασης των πρωτεϊνών με άλλα κυτταρικά υποστρώματα, όπως είναι τα νουκλεϊκά οξέα* **(Murn and Shi 2017)**.

1.4 Αλληλεπίδραση πρωτεϊνικής μεθυλίωσης και μεταβολισμού.

Ιδιαίτερο ενδιαφέρον φαίνεται να έχει η στενή σχέση της πρωτεϊνικής μεθυλίωσης με βασικούς μεταβολίτες του κυττάρου. Σχεδόν όλες οι πρωτεϊνικές μεθυλοτρανσφεράσες απαιτούν την παρουσία ορισμένων συμπαραγόντων, όπως είναι η S-αδενοσυλο-L-μεθειονίνη (SAM), το δινουκλεοτίδιο φλαβίνης-αδενίνης (FAD) και το α-κετογλουταρικό οξύ (a-KG) , για να μπορέσουν να μεθυλιώσουν τους στόχους τους. Ο μεταβολισμός μπορεί να επηρεάσει τη μεθυλίωση μέσω αλλαγών στη διαθεσιμότητα αυτών των συμπαραγόντων ή μέσω της παραγωγής προϊόντων αντίδρασης, όπως η S-αδενοσυλο-ομοκυστεΐνη (SAH) και το $FADH_2$ (Εικόνα 4) (Murn and Shi 2017).



Εικόνα 4 *Σχηματική απεικόνιση της επίδρασης βασικών μεταβολιτών του κυττάρου, όπως είναι το FAD και το SAM, στην πρωτεΐνική μεθυλίωση* **(Murn and Shi 2017)**.

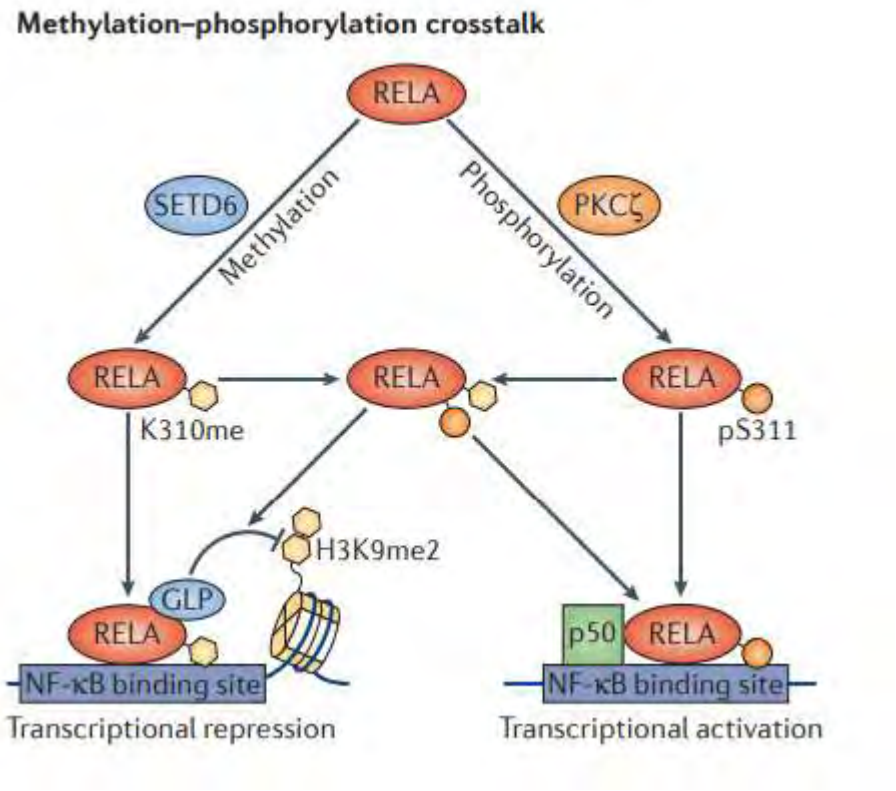
1.5 Αλληλεπίδραση μετα-μεταφραστικών τροποποιήσεων

Ορισμένες πρωτεΐνες υποβάλλονται σε πολυάριθμες μετα-μεταφραστικές τροποποιήσεις (PTMs), οι οποίες πρέπει να ενεργούν συντονισμένα για να εξασφαλίσουν το επιθυμητό βιολογικό αποτέλεσμα. Έτσι, δεν αποτελεί έκπληξη το γεγονός ότι υπάρχει αλληλεπίδραση μεταξύ των διαφορετικών PTMs, όπως είναι η μεθυλίωση, η φωσφορυλίωση, η ακετυλίωση και η ουβικουιτίνωση (Gayatri and Bedford 2014; Zhao et al. 2014). Συγκεκριμένα, οι διάφορες θέσεις μεθυλίωσης μπορούν να επηρεάσουν, είτε θετικά είτε αρνητικά, τη δημιουργία μεθυλίωσης ή άλλων τροποποιήσεων σε γειτονικά αμινοξέα, λειτουργώντας ως «μοριακοί διακόπτες» (Biggar and Li 2015). Επιπλέον, η μεθυλίωση ενός συγκεκριμένου αμινοξέος μπορεί να εμποδίσει άλλες τροποποιήσεις στο ίδιο αμινοξύ. Αυτό ισχύει ειδικά για τη λυσίνη, η οποία είναι το περισσότερο μετα-μεταφραστικά τροποποιημένο αμινοξύ (Lanouette et al. 2014). Στις ιστόνες, αυτή η αλληλεπίδραση μεταξύ των διαφορετικών μετα-μεταφραστικών τροποποιήσεων είναι γνωστή ως «κώδικας των ιστονών» (Strahl and Allis 2000).

1.5.1 Αλληλεπίδραση μεθυλίωσης-φωσφορυλίωσης

Η αλληλεπίδραση μεταξύ γειτονικών θέσεων μεθυλίωσης και φωσφορυλίωσης είναι πολύ συχνή (Biggar and Li 2015; Larsen et al. 2016). Χαρακτηριστικό παράδειγμα αποτελεί ο πυρηνικός παράγοντας-kB (NF-kB), ο οποίος εμπλέκεται στην παραγωγή κυτοκινίνης και στην κυτταρική επιβίωση. Η υπομονάδα RELA μπορεί να μεθυλιωθεί στη λυσίνη 310 από την SETD6. Αυτό έχει ως αποτέλεσμα τη δέσμευση της GLP στην ιστόνη H3 και τη διμεθυλίωση της λυσίνης 9, καταστέλλοντας τη μεταγραφή των γονιδίων-στόχων του NF-kB (Levy et al. 2011; Chang et al. 2011). Ωστόσο, η φωσφορυλίωση της RELA στη σερίνη 311 αποτρέπει τη δέσμευση της GLP στην RELA, αίροντας τη μεταγραφική καταστολή (Duran, et al. 2003). Με αυτόν τον τρόπο, η αλληλεπίδραση μεθυλίωσης-φωσφορυλίωσης στα δύο

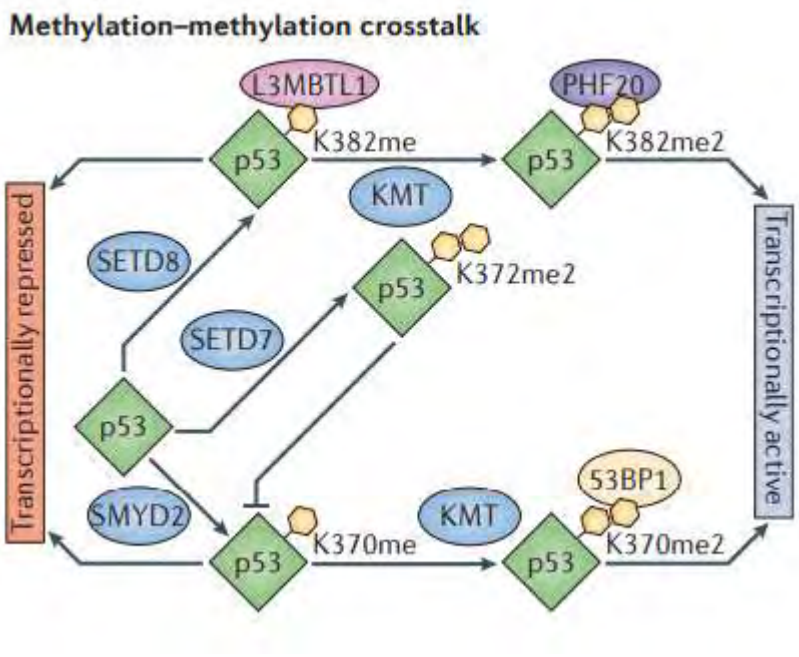
γειτονικά αμινοξέα της υπομονάδας RELA ελέγχει τη δομή της χρωματίνης και, κατ' επέκταση, την έκφραση των γονιδίων-στόχων του NF-kB (Levy et al. 2011).



Εικόνα 5 *Σχηματική απεικόνιση της αλληλεπίδρασης γειτονικών θέσεων μεθυλίωσης-φωσφορυλίωσης στην περιοχή RELA του πυρηνικού παράγοντα NF-kB , η οποία ρυθμίζει τη μεταγραφική του ικανότητα* **(Biggar and Li 2015)**.

1.5.2 Αλληλεπίδραση μεθυλίωσης-μεθυλίωσης

Η δράση του μεταγραφικού παράγοντα p53 μπορεί να ρυθμιστεί από τη μεθυλίωση ορισμένων λυσινών. Η μονομεθυλίωση του p53 στη λυσίνη 370 από τις μεθυλοτρανσφεράσες SET και MYND καταστέλλει τη μεταγραφική δραστηριότητα του p53. Ωστόσο, η διμεθυλίωση της λυσίνης 370 καθιστά δυνατή τη δέσμευση της πρωτεΐνης 53BP1 μέσω της TUDOR περιοχής της, με αποτέλεσμα την ενεργοποίηση της p53–εξαρτώμενης μεταγραφής. Επιπροσθέτως, η διμεθυλίωση της λυσίνης 372 από την SETD7 καταστέλλει τη μονομεθυλίωση της λυσίνης 370 από την SMYD2, προάγοντας την έκφραση των γονιδίων-στόχων του p53.



Εικόνα 6 *Ρύθμιση της μεταγραφικής δραστηριότητας του p53 από τη μεθυλίωση γειτονικών λυσινών* **(Biggar and Li 2015)**.

1.6 Μοριακή και βιοϊατρική σημασία μεθυλίωσης

Ορισμένες πρωτεϊνικές μεθυλοτρανσφεράσες και απομεθυλάσες έχει βρεθεί ότι εμπλέκονται σε ανθρώπινες ασθένειες. Συγκεκριμένα, οι περισσότερες από τις πρωτεϊνικές μεθυλοτρανσφεράσες απορρυθμίζονται σε πολλούς διαφορετικούς καρκίνους, όπως τον καρκίνο του οισοφάγου (Komatsu et al. 2009), της ουροδόχου κύστης (Cho et al. 2012), του μαστού (Frietze et al. 2008), του προστάτη (Majumder et al. 2006) και του εντέρου (Mathioudaki et al. 2008). Η απορρύθμιση τους μπορεί, επίσης, να συμβάλλει στη μεταστατική δυνατότητα του καρκίνου του πνεύμονα και του προστάτη (Larsen et al. 2016). Γι' αυτό τον λόγο αποτελούν θεραπευτικούς στόχους και έχουν αρχίσει να αναπτύσσονται εκλεκτικοί αναστολείς για αυτά τα ένζυμα (Kaniskan and Jin 2017). Επί του παρόντος, ήδη τρεις από αυτούς τους αναστολείς δοκιμάζονται σε κλινικές μελέτες στον άνθρωπο (Kaniskan and Jin 2017). Επίσης, αναστολείς απομεθυλασών (π.χ. για την LSD1) δοκιμάζονται σε κλινικές μελέτες για τη θεραπεία αιματολογικών κακοηθειών (Morera, Lübbert, and Jung 2016). Επιπλέον, μεθυλοτρανσφεράσες, όπως η PRMT5 έχουν προγνωστική αξία σε κακοήθιες, όπως το πολλαπλό μυέλωμα (Gullà et al. 2018). Πλέον, στοχευμένες και σε βάθος βιοχημικές μελέτες εκμεταλλεύονται τα ευρήματα μελετών σε επίπεδο πρωτεώματος για την επιλογή στόχων υψηλού θεραπευτικού και βιοτεχνολογικού ενδιαφέροντος (Larsen et al. 2016).

Ενδιαφέρον αποτελεί το γεγονός ότι η μεθυλίωση παίζει ρόλο στην αλληλεπίδραση παθογόνων βακτηρίων και ξενιστών. Συνεπώς, η κατανόηση αυτής της αλληλεπίδρασης θα οδηγήσει στην παραγωγή νέων και πιο αποτελεσματικών εμβόλιων για την αντιμετώπιση μολυσματικών ασθενειών, όπως ο τύφος (Chao et al. 2008; Lanouette et al. 2014).

1.7 Μεθυλ-πρωτεωμική

1.7.1 Μέθοδος ανίχνευσης θέσεων μεθυλίωσης

Το 2004 έκανε την εμφάνιση της μια επαναστατική μέθοδος βασισμένη στη φασματομετρία μάζας για την ποσοτική ταυτοποίηση/αναγνώριση θέσεων μεθυλίωσης *in vivo*. Αυτή η μέθοδος είναι γνωστή ως heavy-methyl SILAC και αποτελεί μια παραλλαγή της σταθερής ισοτοπικής σήμανσης με αμινοξέα σε κυτταρική καλλιέργεια. Κατά τη μέθοδο της heavy-methyl SILAC, τα κύτταρα αναπτύσσονται σε θρεπτικά μέσα που περιέχουν είτε light (12CH3)- είτε heavy (13CD3)- μεθειονίνη (Met-0 και Met-4, αντίστοιχα). Στη συνέχεια, η μεθειονίνη μετατρέπεται μεταβολικά σε σημασμένη SAM. Κατά την κυτταρική ανάπτυξη, η Met-4 ενσωματώνεται στις νεοσυντιθέμενες πρωτεΐνες, ενώ η heavy-μεθυλομάδα μεταφέρεται από τη SAM σε όλα τα μεθυλιωμένα υποστρώματα από τις αντίστοιχες μεθυλοτρανσφεράσες (Ong, Mittler, and Mann 2004) . Αυτή η προσέγγιση σε συνδυασμό με τον εμπλουτισμό των μεθυλιωμένων πρωτεϊνών με εξειδικευμένα αντισώματα αποτέλεσε την απαρχή της εποχής της μεθυλο-πρωτεωμικής. Αυτές οι προσπάθειες οδήγησαν στον εντοπισμό πάνω από 16000 μοναδικών θέσεων μεθυλίωσης σε περισσότερες από 5500 ανθρώπινες πρωτεΐνες (Murn and Shi 2017).

1.7.2 Το σημαντικό πρόβλημα του βιολογικού και τεχνικού θυρύβου στη μεθυλ-πρωτεωμική

Μία μεγάλη πρόκληση για το πεδίο της μεθυλο-πρωτεωμικής είναι η ποιότητα των παραγόμενων δεδομένων. Όπως συμβαίνει με κάθε νέα τεχνολογία μεγάλης κλίμακας, τα παραγόμενα δεδομένα μεθυλ-πρωτεωμικής επηρεάζονται από πειραματικό θόρυβο. Επιπλέον, μια πολύ πρόσφατη μελέτη κατέδειξε ότι η ταυτοποίηση των θέσεων μεθυλίωσης σε αμινοξέα είναι ιδιαίτερα επιρρεπής σε λάθη. Αυτό συμβαίνει λόγω των πολλών διαφορετικών συνδυασμών αμινοξέων που μπορούν να παράγουν πεπτίδια ισοβαρή με μεθυλιωμένα πεπτίδια διαφορετικής αλληλουχίας. Συνεπώς, η μέθοδος που προτιμάται για τον περιορισμό αυτού του προβλήματος είναι η heavy-methyl-SILAC (Hart-Smith et al. 2016).

1.8 Υπάρχοντα υπολογιστικά εργαλεία πρόβλεψης μεθυλίωσης των πρωτεϊνών

Η μηχανική μάθηση αποτελεί κλάδο της Τεχνητής Νοημοσύνης και ασχολείται με τη μελέτη αλγορίθμων, οι οποίοι βελτιώνουν προοδευτικά την απόδοση τους εξετάζοντας δεδομένα. Βρίσκει πολυάριθμες εφαρμογές από την οικονομία μέχρι την ιατρική και την βιοπληροφορική. Όσο αφορά τη μεθυλίωση, υπάρχουν διάφορα υπολογιστικά εργαλεία, βασισμένα σε τέτοιους αλγορίθμους που προβλέπουν θέσεις μεθυλίωσης σε πρωτεΐνες από την πρωτοταγή δομή τους. Η ύπαρξη τους συμβάλλει στο να ξεπεραστεί η έλλειψη πειραματικών δεδομένων, βοηθώντας τους επιστήμονες να καθοδηγήσουν τα πειράματά τους. Η πλειοψηφία αυτών των εργαλείων χρησιμοποιούν Support Vector Machines καθώς και Random Forests.

2. Σκοπός

Στόχος αυτής της εργασίας ήταν να αναπτύξουμε ένα νέο και πιο αποτελεσματικό υπολογιστικό εργαλείο πρόβλεψης θέσεων μεθυλίωσης σε λυσίνες και αργινίνες, χρησιμοποιώντας δημοσιευμένα δεδομένα μεθυλοπρωτεωμικής. Τα δεδομένα αυτά, αφού μαζέυτηκαν, φιλτραρίστηκαν και χρησιμοποιήθηκαν για την εκπαίδευση νευρωνικών δικτύων. Η ακρίβεια αυτού του εργαλείου συγκρίθηκε με την ακρίβεια πρόσφατα δημοσιευμένων εργαλείων και αποδείχθηκε ανώτερη. Τέλος, δημιουργήθηκε ένας server ελεύθερα διαθέσιμος στο διαδίκτυο για την ανάλυση πρωτεωμάτων και την πρόβλεψη μεθυλίωσης των λυσινών και αργινινών.

3. Υλικά και Μέθοδοι

3.1 Λήψη πρωτεωμάτων

Για την πραγματοποίηση της ανάλυσης λήφθηκαν τα πρωτεώματα του ανθρώπου (*Homo sapiens*), του ποντικού (*Mus musculus*), του σακχαρομύκητα (*S. cerevisiae*) και του πρωτόζωου *Toxoplasma gondii* (strain ATCC 50611/Me49). Πρέπει να επισημανθεί ότι τα πρωτεώματα του ανθρώπου και του ποντικού περιέχουν παραπάνω από μία ισοφορμές για κάθε γονίδιο. Γι' αυτόν το λόγο πραγματοποιήθηκε φιλτράρισμα των πρωτεωμάτων τους, λαμβάνοντας μόνο τη μεγαλύτερη σε μήκος πρωτεΐνη για κάθε γονίδιο. Στον Πίνακας 1 δίνεται για κάθε οργανισμό: το όνομα του αρχείου που περιέχει το πρωτέωμα σε FASTA format, η πηγή λήψης του πρωτεώματος και ο αριθμός των πρωτεϊνών που περιέχονται σε κάθε αρχείο.

Οργανισμός	Όνομα αρχείου	Πηγή	Αριθμός των πρωτεϊνών μετά το φιλτράρισμα
Homo sapiens	Homo_sapiens.GRCh38.pep.all.fa	Ensembl	23031
Mus musculus	Mus_musculus.GRCm38.pep.all.fa	Ensembl	22778
S.cerevisiae	orf_trans_all.fasta.gz	SGD	6425
T. gondii	uniprot-proteome%3AUP000001529.fasta	Uniprot	8315

Πίνακας 1 *Πληροφορίες για τα πρωτεώματα ανά οργανισμό*

3.2 Λήψη δεδομένων μεθυλ-πρωτεωμικής

Στη συνέχεια, ανατρέχοντας σε διάφορες εργασίες, ανακτήθηκαν δεδομένα μεθυλ-πρωτεωμικής για κάθε οργανισμό. Τα δεδομένα αποτελούν, ουσιαστικά, μεθυλιωμένα πεπτίδια ποικίλου μήκους. Τα μεθυλιωμένα αμινοξέα είναι είτε αργινίνη (R) είτε λυσίνη (K). Πρέπει να τονιστεί ότι επιλέχθηκαν δεδομένα μόνο από εργασίες που χρησιμοποίησαν τη μέθοδο heavy-methyl-SILAC ή/και εμπλουτισμό των μεθυλ-πεπτιδίων με εκλεκτικά αντισώματα για την ανίχνευση των θέσεων μεθυλίωσης. Συνεπώς, αυτά τα δεδομένα είναι υψηλής πιστότητας. Ωστόσο, έλαβε χώρα κι ένα περαιτέρω ιδιαίτερα αυστηρό φιλτράρισμα των δεδομένων, ώστε να αποφευχθεί η λήψη ψευδώς-θετικών θέσεων μεθυλίωσης. Στον Πίνακας 2 δίνονται οι εργασίες τις οποίες χρησιμοποιήσαμε. Για κάθε εργασία αναγράφεται: το όνομα των συγγραφέων, το έτος δημοσίευσης, το PMID, ο οργανισμός πάνω στον οποίο έγινε η έρευνα καθώς και τα κριτήρια με το οποία έγινε το φιλτράρισμα των δεδομένων.

Συγγραφείς	Έτος δημοσίευσης	PMID	Οργανισμός	Κριτήρια
Cao et al.	2013	23644510	*Homo sapiens*	• 1% FDR • pfind score <= 0.01
Bremang et al.	2013	23748837	*Homo sapiens*	Μόνο Class A πεπτίδια
Geoghegan et al.	2015	25849564	*Homo sapiens*	• IP probability >= 0.99 • q-value <= 0.01
Guo et al.	2014	24129315	• *Homo sapiens* • *Mus musculus*	• 1% FDR • a-score > 13
Larsen et al.	2016	27577262	*Homo sapiens*	• Localization probability >= 0.99 • PEP >= 0.01
Olsen et al.	2016	26750096	*Homo sapiens*	Peptide score <= 0.01
Onwuli et al.	2016	27600370	*Homo sapiens*	• Localization probability >= 0.99 • PEP >= 0.01
Sylvestersen et al.	2014	24563534	*Homo sapiens*	• Andromeda score >= 100 • Localization Probability >= 0.99

Wu et al.	2015	25505155	*Homo sapiens*	• MASCOT score >= 30 • p-value <= 0.01
Hornbeck et al. (PhosphositePlus)	2014	25514926	• *Homo sapiens* • *Mus musculus*	Μόνο πεπτίδια από Low throughput πειράματα
Hart-Smith et al.	2016	26699799	*S. cerevisiae*	• MASCOT score >= 30 • p-value <= 0.01
Plank et al.	2015	26046779	*S. cerevisiae*	MASCOT score >= 30
Yagoub et al.	2015	26081071	*S. cerevisiae*	MASCOT score >= 30
Yakubu et al.	2017	28143887	*T. gondii*	Localization probability >= 0.99

Πίνακας 2 *Εργασίες, των οποίων τα δεδομένα, χρησιμοποιήθηκαν στην ανάλυσή μας.*

3.3 Εύρεση των θέσεων μεθυλίωσης στο πρωτέωμα

Τα δεδομένα της μεθυλ-πρωτεωμικής που ανακτήσαμε χρησιμοποιήθηκαν για την εύρεση των θέσεων μεθυλίωσης στο πρωτέωμα. Αυτό πραγματοποιήθηκε με τη βοήθεια ενός perl script, που αντιστοιχεί τα μεθυλιωμένα πεπτίδια από κάθε οργανισμό στο ανάλογο πρωτέωμα, ανευρίσκοντας τις θέσεις μεθυλίωσης. Στον Πίνακας 3 δίνεται ο αριθμός των θέσεων μεθυλίωσης που συγκεντρώθηκαν για κάθε οργανισμό.

Οργανισμός	**Αριθμός θέσεων μεθυλίωσης**		
	R και K	**R**	**K**
Homo sapiens	8838	6966	1872
Mus musculus	1666	1635	31
S. cerevisiae	67	55	12
T. gondii	559	559	-

Πίνακας 3 *Αριθμός των θέσεων μεθυλίωσης που βρέθηκαν ανά οργανισμό.*

3.4 Δημιουργία ομάδων δεδομένων για τον άνθρωπο

Στη συνέχεια, όσον αφορά τον άνθρωπο, συγκεντρώθηκαν 8838 θέσεις μεθυλίωσης (6966 για αργινίνη και 1872 για λυσίνη). Στη συνέχεια δημιουργήσαμε μία ομάδα δεδομένων, το dataset HQ (High Quality) που περιέχει κάθε θέση που βρέθηκε σε low throughput πείραμα ή/και που βρέθηκε σε τουλάχιστον 2 από τα high-throughput πειράματα που φιλτράραμε. Το dataset HQ περιέχει συνολικά 1754 αργινίνες και 270 λυσίνες και θα χρησιμοποιηθεί για την κατασκευή του νευρωνικού δικτύου.

3.5 Προετοιμασία κατασκευής Τεχνητού Νευρωνικού Δικτύου

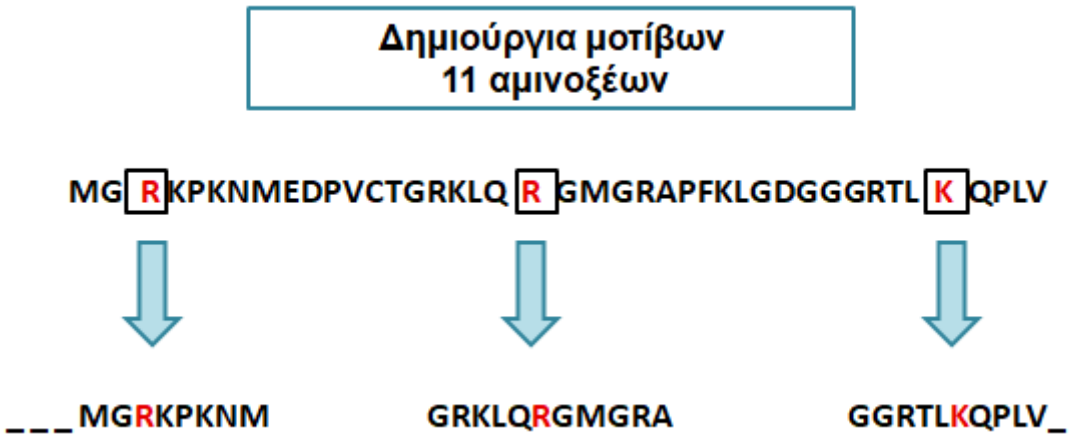
3.5.1 Δεδομένα για εκπαίδευση και αξιολόγηση

Για την κατασκευή του Τεχνητού Νευρωνικού Δικτύου είναι απαραίτητη η δημιουργία datasets, ένα για την εκπαίδευση (Training Dataset, TD – 70% των δεδομένων) και τουλάχιστον ένα για την αξιολόγηση του δικτύου (Evaluation Dataset, EVD – 30% των δεδομένων). Κάθε ένα από αυτά τα dataset περιέχει, εκτός από τις θετικές θέσεις (μεθυλιωμένες), ίδιο αριθμό τυχαία επιλεγμένων αρνητικών (μη-μεθυλιωμένων) αργινινών και λυσινών από πρωτεΐνες που δεν υπάρχει ένδειξη ότι μεθυλιώνονται.

Για την εκπαίδευση του δικτύου χρησιμοποιήθηκαν δεδομένα από τον άνθρωπο. Συγκεκριμένα, χρησιμοποιήθηκε το HQ dataset. Επιπλέον, για την αξιολόγηση χρησιμοποιήθηκαν και τα πειραματικά δεδομένα από τον ποντικό, τον σακχαρομύκητα και το *Toxoplasma gondii*. Οι οργανισμοί αυτοί επιλέχθηκαν ώστε να αξιολογήσουμε σε τι βαθμό είναι ικανό το υπολογιστικό εργαλείο μας να κάνει σωστή πρόβλεψη σε άλλους μακρινά συγγενικούς ευκαρυώτες.

3.5.2 Δημιουργία μοτίβων 11 αμινοξέων

Έπειτα, για κάθε θέση μεθυλίωσης δημιουργήθηκε ένα μοτίβο μήκους 11 αμινοξέων. Σε αυτό το μοτίβο, το μεθυλιωμένο αμινοξύ βρίσκεται στο κέντρο, ενώ δεξιά κι αριστερά πλαισιώνεται από 5 και 5 αμινοξέα, αντίστοιχα. Σε περίπτωση που το μεθυλιωμένο αμινοξύ βρίσκεται κοντά στην αρχή ή το τέλος της πρωτεΐνης, το μοτίβο δημιουργείται κανονικά, χρησιμοποιώντας το σύμβολο (_) για τα αμινοξέα που δεν υπάρχουν, όπως φαίνεται στην Εικόνα 7. Η ίδια διαδικασία ακολουθήθηκε και για τις μη-μεθυλιωμένες θέσεις.



Εικόνα 7 *Σχηματική απεικόνιση της δημιουργία μοτίβων 11 αμινοξέων. Με τη σειρά φαίνεται η διαδικασία δημιουργίας των μοτίβων για αμινοξέα που βρίσκονται κοντά στην αρχή, στο κέντρο και κοντά στο τέλος της πρωτεΐνης.*

3.5.3 Μετατροπή μοτίβων σε one-hot-encoding

Η κατασκευή των Τεχνητών Νευρωνικών Δικτύων (ΤΝΔ) απαιτεί την μετατροπή των δεδομένων που θέλουμε να επεξεργαστεί σε μία μορφή που «αναγνωρίζουν» οι αλγόριθμοι. Αυτό μπορεί να γίνει με την κωδικοποίηση one-hot-encoding, κατά την οποία κάθε λέξη ή χαρακτήρας «κωδικοποιείται» με τη χρήση των ψηφίων “0” και “1”. Για παράδειγμα, ένα αμινοξύ μπορεί να κωδικοποιηθεί με τον εξής τρόπο: Ως γνωστόν, τα αμινοξέα είναι στο σύνολο 20. Έτσι, όπως βλέπουμε στον Πίνακας 4, χρησιμοποιώντας τα ψηφία “0” και “1” , μπορούμε να δημιουργήσουμε μια σειρά από 20 ψηφία. Αυτή η σειρά, ανάλογα με τη θέση που βρίσκεται το ψηφίο “1”, θα είναι μοναδική και θα αντιστοιχεί σε ένα συγκεκριμένο αμινοξύ.

Αμινοξέα	**One-hot-encoding**																			
Αλανίνη (A)	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Κυστεΐνη (C)	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ασπαρτικό (D)	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Γλουταμικό (E)	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Φαινυλαλανίνη (F)	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Γλυκίνη (G)	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ιστίδινη (H)	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Ισολευκίνη (I)	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Λυσίνη (K)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Λευκίνη (L)	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Μεθειονίνη (M)	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Ασπαραγίνη (N)	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Προλίνη (P)	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Γλουταμίνη (Q)	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
Αργινίνη (R)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Σερίνη (S)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Θρεονίνη (T)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Βαλίνη (V)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Τρυπτοφάνη (W)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Τυροσίνη (Y)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Πίνακας 4 *One-hot encoding για τα 20 αμινοξέα.*

Χρησιμοποιώντας αυτή τη λογική, πραγματοποιήθηκε η μετατροπή των μοτίβων 11 αμινοξέων σε one-hot-encoding. Κάθε αμινοξύ αναπαριστάται από 20 ψηφία στη σειρά, επομένως κάθε μοτίβο αποτελείται από 220 ψηφία.

3.6 Κατασκευή Τεχνητού Νευρωνικού Δικτύου

Στη συνέχεια, προβήκαμε στην κατασκευή ενός Τεχνητού Νευρωνικού Δικτύου πρόβλεψης θέσεων μεθυλίωσης. Για τη δημιουργία και την εκπαίδευση του νευρωνικού δικτύου χρησιμοποιήθηκαν οι αλγόριθμοι Keras/Tensorflow.

Μετά από 243 διαφορετικούς συνδυασμούς, οι παράμετροι που χρησιμοποιήθηκαν για τη βέλτιστη λειτουργία του δικτύου είναι:

- ✓ Κόμβοι ή Nodes: 13
- ✓ Dropout: 0.6
- ✓ Epochs: 30
- ✓ Batch size: 160

Η εκπαίδευση κι η αξιολόγηση πραγματοποιήθηκαν με τις ομάδες δεδομένων που αναφέρθηκαν προηγουμένως.

3.7 Δημιουργία server

Έπειτα, αναπτύχθηκε ένας webserver με το όνομα Methyl-Prometheus από τον υποψήφιο διδάκτορα Πάνο Βλασταρίδη, που βασίζεται στο Jhipster Application Framework. Το Jhipster Application Framework χρησιμοποιεί τη γλώσσα προγραμματισμού Java και το Spring Framework για το back-end και το Angular Javascript Framework για το front-end. Ο Methyl-Prometheus είναι ελεύθερα προσβάσιμος στο διαδίκτυο στη διεύθυνση: http://bioinf.bio.uth.gr/methyl-prometheus/. Για όλες τις αξιολογήσεις του Methyl-Prometheus, χρησιμοποιήθηκε ένα score >= 0.5 ως κατώφλι (threshold) για την πρόβλεψη μιας θέσης ως μεθυλιωμένης.

3.8 Άλλοι αλγόριθμοι

Η ομάδα δεδομένων του ανθρώπου που χρησιμοποιήθηκε για την εκπαίδευση του νευρωνικού δικτύου χρησιμοποιήθηκε και για την εκπαίδευση 21 αλγορίθμων μηχανικής μάθησης, όπως είναι τα Random Forests. Αυτό πραγματοποιήθηκε με το πρόγραμμα WEKA, χρησιμοποιώντας τις προκαθορισμένες παραμέτρους του προγράμματος και 5-fold-cross-validation.

3.9 Άλλα προγράμματα πρόβλεψης θέσεων μεθυλίωσης

Τα δεδομένα αξιολόγησης από τον άνθρωπο, τον ποντικό, τον σαχκαρομύκητα και το *Toxoplasma gondii* χρησιμοποιήθηκαν για σύγκριση και σε άλλα δημοσιευμένα προγράμματα πρόβλεψης των θέσεων μεθυλίωσης όπως είναι το MePred-RF, PRmePRed και GPS-MSP (Wei et al. 2018; Kumar et al. 2017; Deng et al. 2017)

4. Αποτελέσματα-Συζήτηση

4.1 Εκτίμηση της λειτουργίας του Τεχνητού Νευρωνικού Δικτύου

Το δίκτυο που δημιουργήσαμε εκπαιδεύτηκε με το TD του ανθρώπου και κατόπιν έγινε εκτίμηση της λειτουργίας του με το EVD του ανθρώπου. Όπως φαίνεται στον Πίνακας 5, η ακρίβεια (accuracy) του δικτύου, όσον αφορά την πρόβλεψη μεθυλίωσης σε αργινίνη και λυσίνη (R&K), έφτασε το 81%. Όσον αφορά την ευαισθησία (sensitivity) και την ειδικότητα (specificity), υπήρξε ισορροπία μεταξύ των τιμών τους, οι οποίες ήταν 79% και 84%, αντίστοιχα. Επιπλέον, ο Συντελεστής Συσχέτισης Matthews (MCC) και η περιοχή κάτω από την καμπύλη (AUC) είχαν τιμές 0.62 και 0.894. Έπειτα, πραγματοποιήθηκε 5-fold cross-validation σε όλο το HQ dataset, πετυχαίνοντας ακρίβεια 81.5%, κατά μέσο όρο.

	Sensitivity	Specificity	Accuracy	MCC
R&K	0,79	0,84	0,81	0,62
R	0,82	0,85	0,83	0,67
K	0,61	0,73	0,67	0,34

Πίνακας 5 *Αποτελέσματα αξιολόγησης του ΤΝΔ.*

4.2 Ικανότητα ανίχνευσης θέσεων μεθυλίωσης σε άλλους ευκαρυωτικούς οργανισμούς

Στη συνέχεια, θέλαμε να διερευνήσουμε την ικανότητα του δικτύου να ανιχνεύει θέσεις μεθυλίωσης και σε άλλες μεγάλες ταξινομικές ομάδες των ευκαρυωτών. Για αυτόν τον σκοπό, χρησιμοποιήσαμε πειραματικά δεδομένα από τον ποντικό, τον σακχαρομύκητα και το *Toxoplasma gondii*. Όσον αφορά τον ποντικό, το δίκτυο έκανε πρόβλεψη θέσεων μεθυλίωσης με ακρίβεια 79%, ενώ υπήρχε μια σχετική ισορροπία μεταξύ ευαισθησίας (73%) και την ειδικότητας (86%), όπως και στον άνθρωπο (Πίνακας 6). Συνεπώς, αυτός ο αλγόριθμος πρέπει να είναι εξίσου ακριβής για τα περισσότερα από τα θηλαστικά. Επιπλέον, η απόδοση του δικτύου στο μικρό dataset του *S. cerevisiae* (91% ευαισθησία και 81% ειδικότητα), καθώς και στο dataset του *T. gondii* (93% ευαισθησία και 85% ειδικότητα) δείχνει ότι ο αλγόριθμος αναμένεται να έχει πολύ μεγάλη ακρίβεια στις περισσότερες από τις άλλες μεγάλες ευκαρυωτικές ταξινομικές ομάδες (Πίνακας 6).

Οργανισμός	Ευαισθησία	Ειδικότητα	Ακρίβεια	MCC
Mus musculus	0,73	0,86	0,79	0,59
S.cerevisiae	0,91	0,81	0,86	0,72
T. gondii	0,93	0,85	0,89	0,78

Πίνακας 6 *Αποτελέσματα αξιολόγησης του αλγορίθμου που αναπτύξαμε με τα EVD του ποντικού, του S.cerevisiae και του T.gondii.*

4.3 Σύγκριση με άλλα εργαλεία πρόβλεψης θέσεων μεθυλίωσης

Τα δεδομένα αξιολόγησης από τον άνθρωπο, τον ποντικό, τον σακχαρομύκητα και το *Toxoplasma gondii* χρησιμοποιήθηκαν και σε άλλα προγράμματα πρόβλεψης θέσεων μεθυλίωσης όπως είναι το MePred-RF, PRmePRed και GPS-MSP (Wei et al. 2018; Kumar et al. 2017; Deng et al. 2017). Το πιο πρόσφατο μεταξύ των προγραμμάτων είναι το PRmePRed, το οποίο αξιολογήθηκε σε σχέση με αρκετούς άλλους αλγορίθμους και εμφάνισε βελτιωμένη απόδοση σε σύγκριση με αυτούς. Παρόλα αυτά, ο αλγόριθμος μας εμφάνισε καλύτερη απόδοση σε σχέση με το PRmePRed αλλά και με άλλους δημοσιευμένους αλγορίθμους, πιθανότατα εξαιτίας του αυστηρού φιλτραρίσματος που

εφαρμόσαμε για την απομάκρυνση των ψευδώς-θετικών θέσεων. Στον άνθρωπο, τον ποντικό και το *Toxoplasma gondii,* ο αλγόριθμος που αναπτύξαμε ξεπέρασε τους υπόλοιπους αλγορίθμους από άποψη ακρίβειας και MCC, όπως φαίνεται στους Πίνακες Πίνακας 7, Πίνακας 8 καιΠίνακας 10. Μόνο στον *S. cerevisiae,* το πρόγραμμα MePred-RF ξεπέρασε τον αλγόριθμο μας στην ακρίβεια κατά 4% (Πίνακας 9).

Homo sapiens				
K				
Αλγόριθμος	**Ευαισθησία**	**Ειδικότητα**	**Ακρίβεια**	**MCC**
PRmePRed	Διαθέσιμο μόνο για αργινίνες!			
MePred-RF	0,49	0,71	0,60	0,21
GPS-MSP	0,31	0,99	0,65	0,41
Methyl-Prometheus	0,61	0,73	0,67	0,34
R				
Αλγόριθμος	**Ευαισθησία**	**Ειδικότητα**	**Ακρίβεια**	**MCC**
PRmePRed	0,96	0,56	0,76	0,56
MePred-RF	0,53	0,96	0,75	0,55
GPS-MSP	0,14	0,98	0,56	0,23
Methyl-Prometheus	0,82	0,85	0,83	0,67
R&K				
Αλγόριθμος	**Ευαισθησία**	**Ειδικότητα**	**Ακρίβεια**	**MCC**
PRmePRed	-	-	-	-
MePred-RF	0,53	0,93	0,73	0,50
GPS-MSP	0,16	0,99	0,57	0,26
Methyl-Prometheus	0,79	0,84	0,81	0,62

Πίνακας 7 *Αποτελέσματα αξιολόγησης διάφορων αλγορίθμων με το EVD dataset του ανθρώπου.*

Mus musculus				
K				
Αλγόριθμος	**Ευαισθησία**	**Ειδικότητα**	**Ακρίβεια**	**MCC**
PRmePRed	Διαθέσιμο μόνο για αργινίνες!			
MePred-RF	0,68	0,71	0,69	0,38
GPS-MSP	0,23	1	0,61	0,36
Methyl-Prometheus	0,61	0,74	0,68	0,36
R				
Αλγόριθμος	**Ευαισθησία**	**Ειδικότητα**	**Ακρίβεια**	**MCC**
PRmePRed	0,92	0,52	0,72	0,47
MePred-RF	0,49	0,94	0,71	0,48
GPS-MSP	0,21	0,99	0,60	0,31
Methyl-Prometheus	0,73	0,86	0,79	0,59
R&K				
Αλγόριθμος	**Ευαισθησία**	**Ειδικότητα**	**Ακρίβεια**	**MCC**
PRmePRed	-	-	-	-
MePred-RF	0,49	0,93	0,71	0,47
GPS-MSP	0,21	0,99	0,60	0,31
Methyl-Prometheus	0,73	0,86	0,79	0,59

Πίνακας 8 *Αποτελέσματα αξιολόγησης διάφορων αλγορίθμων με το EVD dataset του ποντικού.*

S. cerevisiae				
K				
Αλγόριθμος	**Ευαισθησία**	**Ειδικότητα**	**Ακρίβεια**	**MCC**
PRmePRed	Διαθέσιμο μόνο για αργινίνες!			
MePred-RF	0,67	0,67	0,67	0,33
GPS-MSP	0,83	1	0,92	0,85
Methyl-Prometheus	0,58	0,58	0,58	0,17
R				
Αλγόριθμος	**Ευαισθησία**	**Ειδικότητα**	**Ακρίβεια**	**MCC**
PRmePRed	0,98	0,65	0,82	0,67
MePred-RF	0,93	0,96	0,95	0,89
GPS-MSP	0,40	1	0,70	0,50
Methyl-Prometheus	0,98	0,85	0,92	0,84
R&K				
Αλγόριθμος	**Ευαισθησία**	**Ειδικότητα**	**Ακρίβεια**	**MCC**
PRmePRed				
MePred-RF	0,88	0,91	0,90	0,79
GPS-MSP	0,48	1	0,74	0,56
Methyl-Prometheus	0,91	0,81	0,86	0,72

Πίνακας 9 *Αποτελέσματα αξιολόγησης διάφορων αλγορίθμων με το EVD dataset του S. cereviasiae.*

Toxoplasma gondii				
R				
Αλγόριθμος	**Ευαισθησία**	**Ειδικότητα**	**Ακρίβεια**	**MCC**
PRmePRed	0,85	0,91	0,88	0,75
MePred-RF	0,99	0,28	0,64	0,39
GPS-MSP	0,20	1	0,60	0,33
Methyl-Prometheus	0,93	0,85	0,89	0,78

Πίνακας 10 *Αποτελέσματα αξιολόγησης διάφορων αλγορίθμων με το EVD dataset του T. gondii.*

4.4 Εκπαίδευση άλλων αλγορίθμων

Επιπλέον, το HQ dataset του ανθρώπου χρησιμοποιήθηκε για την εκπαίδευση 21 διαφορετικών αλγορίθμων μηχανικής μάθησης (όπως είναι οι SVM και οι Random Forests), μέσω του προγράμματος WEKA. Έγινε χρήση των προκαθορισμένων παραμέτρων του WEKA και 5-fold cross-validation. Όπως φαίνεται στον Πίνακας 11, οι 4 πρώτοι σε απόδοση αλγόριθμοι πέτυχαν ακρίβεια της τάξης του 80,5-81%, συμπεραίνοντας ότι δεν ξεπέρασαν το νευρωνικό δίκτυο που εκπαιδεύσαμε με το Keras/Tensorflow.

Αλγόριθμος	**Ακρίβεια (%)**
LWL	71,94
OneR	72,31
AdaBoostM1	73,54
Decision Table	73,86
J48	73,91
IBk	74,36
JRip	74,56
PART	74,88
Kstar	75,74
REPTree	76,24

MultilayerPerceptor	79,32
LibSVM	79,37
RandomForest	79,40
ClassificationViaRegression	79,40
NaiveBayes	79,79
NaiveBayesUpdateable	79,79
BayesNet	79,84
SGD	80,53
Logistic	80,61
SimpleLogistic	80,93
LMT	80,93

Πίνακας 11 *Αποτελέσματα αξιολόγησης 21 διαφορετικών αλγορίθμων μηχανικής μάθησης τα οποία εκπαιδεύτηκαν με το HQ dataset του ανθρώπου.*

5. Συμπεράσματα

Εν κατακλείδι, χρησιμοποιώντας τα πιο πρόσφατα δεδομένα μεθυλ-πρωτεωμικής, στα οποία εφαρμόσαμε πολύ αυστηρά κριτήρια για να απομακρύνουμε το «θόρυβο», εκπαιδεύσαμε ένα νευρωνικό δίκτυο με τη χρήση του Keras/Tensorflow. Το νευρωνικό δίκτυο είναι ελεύθερα προσβάσιμο ως webserver με το όνομα Methyl-Prometheus. Η απόδοση του ξεπερνάει εκείνη των άλλων πρόσφατα δημοσιευμένων αλγορίθμων και εμφανίζει μία ισορροπημένη ευαισθησία και ειδικότητα. Με βάση την ανάλυση μας, αυτός ο αλγόριθμος επιτυγχάνει ακρίβεια της τάξης του 79-81% στα θηλαστικά. Υψηλή ακρίβεια επιτυγχάνει και σε άλλους ευκαρυώτες, όπως είναι οι μύκητες και τα Apicomplexa. Ο αλγόριθμος είναι, επίσης, ικανός να σαρώσει το ανθρώπινο πρωτέωμα (~23.000 πρωτεΐνες) μέσα σε 10 λεπτά. Δεδομένου ότι γίνονται διαθέσιμα όλο και περισσότερα υψηλής ποιότητας δεδομένα από τον άνθρωπο αλλά κι από άλλους οργανισμούς, η απόδοση του αλγορίθμου μας θα συνεχίσει να αυξάνεται, υιοθετώντας, εκτός από το τοπικό πλαίσιο πρωτοταγούς αλληλουχίας, κι άλλους τύπους λειτουργικών πληροφοριών, όπως μερικά εργαλεία πρόβλεψης πρωτεϊνικής φωσφορυλίωσης (Fan et al. 2014; Wang, Wang, and Li 2017; Xu et al. 2014).

Βιβλιογραφία

1. Ambler, R. P., and M. W. Rees. 1959. “Ε-N-Methyl-Lysine in Bacterial Flagellar Protein.” *Nature* 184 (4679): 56–57. https://doi.org/10.1038/184056b0.

2. Bedford, Mark T., and Stéphane Richard. 2005. “Arginine Methylation: An Emerging Regulatorof Protein Function.” *Molecular Cell* 18 (3): 263–72. https://doi.org/10.1016/j.molcel.2005.04.003.

3. Beltran-Alvarez, Pedro, Ferran Feixas, Sílvia Osuna, Rubí Díaz-Hernández, Ramon Brugada, and Sara Pagans. 2015. “Interplay between R513 Methylation and S516 Phosphorylation of the Cardiac Voltage-Gated Sodium Channel.” *Amino Acids* 47 (2): 429–34. https://doi.org/10.1007/s00726-014-1890-0.

4. Biggar, Kyle K., and Shawn S.-C. Li. 2015. “Non-Histone Protein Methylation as a Regulator of Cellular Signalling and Function.” *Nature Reviews Molecular Cell Biology* 16 (1): 5–17. https://doi.org/10.1038/nrm3915.

5. Blanc, Roméo S., and Stéphane Richard. 2017. “Arginine Methylation: The Coming of Age.” *Molecular Cell* 65 (1): 8–24. https://doi.org/10.1016/j.molcel.2016.11.003.

6. Bremang, Michael, Alessandro Cuomo, Anna Maria Agresta, Magdalena Stugiewicz, Valeria Spadotto, and Tiziana Bonaldi. 2013. “Mass Spectrometry-Based Identification and Characterisation of Lysine and Arginine Methylation in the Human Proteome.” *Molecular BioSystems* 9 (9): 2231–47. https://doi.org/10.1039/C3MB00009E.

7. Carr, Simon M, Shonagh Munro, Benedikt Kessler, Udo Oppermann, and Nicholas B La Thangue. 2011. “Interplay between Lysine Methylation and Cdk Phosphorylation in Growth Control by the Retinoblastoma Protein.” *The EMBO Journal* 30 (2): 317–27. https://doi.org/10.1038/emboj.2010.311.

8. Chang, Yanqi, Dan Levy, John R. Horton, Junmin Peng, Xing Zhang, Or Gozani, and Xiaodong Cheng. 2011. “Structural Basis of SETD6-Mediated Regulation of the NF-KB Network via Methyl-Lysine Signaling.” *Nucleic Acids Research* 39 (15): 6380–89. https://doi.org/10.1093/nar/gkr256.

9. Chao, Chien-Chung, Zhiwen Zhang, Hui Wang, Abdulnaser Alkhalil, and Wei-Mei Ching. 2008. “Serological Reactivity and Biochemical Characterization of Methylated and Unmethylated Forms of a Recombinant Protein Fragment Derived from Outer Membrane Protein B of Rickettsia Typhi.” *Clinical and Vaccine Immunology : CVI* 15 (4): 684–90. https://doi.org/10.1128/CVI.00281-07.

10. Chen, Chen, Timothy J. Nott, Jing Jin, and Tony Pawson. 2011. “Deciphering Arginine Methylation: Tudor Tells the Tale.” *Nature Reviews Molecular Cell Biology* 12 (10): 629–42. https://doi.org/10.1038/nrm3185.

11. Cho, Hyun-Soo, Shinya Hayami, Gouji Toyokawa, Kazuhiro Maejima, Yuka Yamane, Takehiro Suzuki, Naoshi Dohmae, et al. 2012. “RB1 Methylation by SMYD2 Enhances Cell Cycle Progression through an Increase of RB1 Phosphorylation.” *Neoplasia (New York, N.Y.)* 14 (6): 476–86.

12. Deng, Wankun, Yongbo Wang, Lili Ma, Ying Zhang, Shahid Ullah, and Yu Xue. 2017. “Computational Prediction of Methylation Types of Covalently Modified Lysine and Arginine Residues in Proteins.” *Briefings in Bioinformatics* 18 (4): 647–58. https://doi.org/10.1093/bib/bbw041.

13. Dillon, Shane C, Xing Zhang, Raymond C Trievel, and Xiaodong Cheng. 2005. “The SET-Domain Protein Superfamily: Protein Lysine Methyltransferases.” *Genome Biology* 6 (8): 227. https://doi.org/10.1186/gb-2005-6-8-227.

14. Duran, Angeles, María T. Diaz-Meco, and Jorge Moscat. 2003. “Essential Role of RelA Ser311 Phosphorylation by ZPKC in NF-KB Transcriptional Activation.” *The EMBO Journal* 22 (15): 3910–18. https://doi.org/10.1093/emboj/cdg370.

15. Evich, Marina, Ekaterina Stroeva, Yujun George Zheng, and Markus W. Germann. 2016. “Effect of Methylation on the Side- chain PK a Value of Arginine.” *Protein Science : A Publication of the Protein Society* 25 (2): 479–86. https://doi.org/10.1002/pro.2838.

16. Fan, Wenwen, Xiaoyi Xu, Yi Shen, Huanqing Feng, Ao Li, and Minghui Wang. 2014. “Prediction of Protein Kinase-Specific Phosphorylation Sites in Hierarchical Structure Using Functional Information and Random Forest.” *Amino Acids* 46 (4): 1069–78. https://doi.org/10.1007/s00726-014-1669-3.

17. Feng, You, Ranjan Maity, Julian P. Whitelegge, Andrea Hadjikyriacou, Ziwei Li, Cecilia Zurita-Lopez, Qais Al-Hadid, et al. 2013. “Mammalian Protein Arginine Methyltransferase 7 (PRMT7) Specifically Targets RXR Sites in Lysine- and Arginine-Rich Regions.” *The Journal of Biological Chemistry* 288 (52): 37010–25. https://doi.org/10.1074/jbc.M113.525345.

18. Frietze, Seth, Mathieu Lupien, Pamela A. Silver, and Myles Brown. 2008. “CARM1 Regulates Estrogen-Stimulated Breast Cancer Growth through Up-Regulation of E2F1.” *Cancer Research* 68 (1): 301–6. https://doi.org/10.1158/0008-5472.CAN-07-1983.

19. Gayatri, Sitaram, and Mark T. Bedford. 2014. “Readers of Histone Methylarginine Marks.” *Biochimica et Biophysica Acta* 1839 (8): 702–10. https://doi.org/10.1016/j.bbagrm.2014.02.015.

20. Grillo, M. A., and S. Colombatto. 2005. “S-Adenosylmethionine and Protein Methylation.” *Amino Acids* 28 (4): 357–62. https://doi.org/10.1007/s00726-005-0197-6.

21. Gullà, A., T. Hideshima, G. Bianchi, M. Fulciniti, M. Kemal Samur, J. Qi, Y.-T. Tai, et al. 2018. “Protein Arginine Methyltransferase 5 Has Prognostic Relevance and Is a

Druggable Target in Multiple Myeloma." *Leukemia* 32 (4): 996–1002. https://doi.org/10.1038/leu.2017.334.

22. Guo, Ailan, Hongbo Gu, Jing Zhou, Daniel Mulhern, Yi Wang, Kimberly A. Lee, Vicky Yang, et al. 2014. "Immunoaffinity Enrichment and Mass Spectrometry Analysis of Protein Methylation." *Molecular & Cellular Proteomics : MCP* 13 (1): 372–87. https://doi.org/10.1074/mcp.O113.027870.

23. Hart-Smith, Gene, Daniel Yagoub, Aidan P. Tay, Russell Pickford, and Marc R. Wilkins. 2016. "Large Scale Mass Spectrometry-Based Identifications of Enzyme-Mediated Protein Methylation Are Subject to High False Discovery Rates." *Molecular & Cellular Proteomics : MCP* 15 (3): 989–1006. https://doi.org/10.1074/mcp.M115.055384.

24. Kaniskan, H. Ümit, and Jian Jin. 2017. "Recent Progress in Developing Selective Inhibitors of Protein Methyltransferases." *Current Opinion in Chemical Biology*, Molecular Imaging Chemical Genetics and Epigenetics, 39 (August): 100–108. https://doi.org/10.1016/j.cbpa.2017.06.013.

25. Komatsu, Shuhei, Issei Imoto, Hitoshi Tsuda, Ken-ich Kozaki, Tomoki Muramatsu, Yutaka Shimada, Satoshi Aiko, et al. 2009. "Overexpression of SMYD2 Relates to Tumor Cell Proliferation and Malignant Outcome of Esophageal Squamous Cell Carcinoma." *Carcinogenesis* 30 (7): 1139–46. https://doi.org/10.1093/carcin/bgp116.

26. Kontaki, Haroula, and Iannis Talianidis. 2010. "Lysine Methylation Regulates E2F1-Induced Cell Death." *Molecular Cell* 39 (1): 152–60. https://doi.org/10.1016/j.molcel.2010.06.006.

27. Kouzarides, Tony. 2007. "Chromatin Modifications and Their Function." *Cell* 128 (4): 693–705. https://doi.org/10.1016/j.cell.2007.02.005.

28. Kumar, Pawan, Joseph Joy, Ashutosh Pandey, and Dinesh Gupta. 2017. "PRmePRed: A Protein Arginine Methylation Prediction Tool." *PLoS ONE* 12 (8). https://doi.org/10.1371/journal.pone.0183318.

29. Kwak, Youn Tae, Jun Guo, Shashi Prajapati, Kyu-Jin Park, Rama M. Surabhi, Brady Miller, Peter Gehrig, and Richard B. Gaynor. 2003. "Methylation of SPT5 Regulates Its Interaction with RNA Polymerase II and Transcriptional Elongation Properties." *Molecular Cell* 11 (4): 1055–66. https://doi.org/10.1016/S1097-2765(03)00101-1.

30. Lanouette, Sylvain, Vanessa Mongeon, Daniel Figeys, and Jean- François Couture. 2014. "The Functional Diversity of Protein Lysine Methylation." *Molecular Systems Biology* 10 (4). https://doi.org/10.1002/msb.134974.

31. Larsen, Sara, Kathrine B. Sylvestersen, Andreas Mund, David Lyon, Meeli Mullari, Maria V. Madsen, Jeremy Daniel, Lars J. Jensen, and Michael L. Nielsen. 2016. "Proteome-Wide Analysis of Arginine Monomethylation Reveals Widespread Occurrence in Human Cells." *Science Signaling* 9 (August): rs9–rs9. https://doi.org/10.1126/scisignal.aaf7329.

32. Lee, Young-Ho, and Michael R. Stallcup. 2009. “Minireview: Protein Arginine Methylation of Nonhistone Proteins in Transcriptional Regulation.” *Molecular Endocrinology* 23 (4): 425–33. https://doi.org/10.1210/me.2008-0380.

33. Levy, Dan, Alex J. Kuo, Yanqi Chang, Uwe Schaefer, Christopher Kitson, Peggie Cheung, Alexsandra Espejo, et al. 2011. “Lysine Methylation of the NF-KB Subunit RelA by SETD6 Couples Activity of the Histone Methyltransferase GLP at Chromatin to Tonic Repression of NF-KB Signaling.” *Nature Immunology* 12 (1): 29–36. https://doi.org/10.1038/ni.1968.

34. Majumder, Samarpan, Yuanbo Liu, O. Harris Ford, James L. Mohler, and Young E. Whang. 2006. “Involvement of Arginine Methyltransferase CARM1 in Androgen Receptor Function and Prostate Cancer Cell Viability.” *The Prostate* 66 (12): 1292–1301. https://doi.org/10.1002/pros.20438.

35. Mann, Matthias, and Ole N. Jensen. 2003. “Proteomic Analysis of Post-Translational Modifications.” *Nature Biotechnology* 21 (3): 255–61. https://doi.org/10.1038/nbt0303-255.

36. Mathioudaki, K, A Papadokostopoulou, A Scorilas, D Xynopoulos, N Agnanti, and M Talieri. 2008. “The PRMT1 Gene Expression Pattern in Colon Cancer.” *British Journal of Cancer* 99 (12): 2094–99. https://doi.org/10.1038/sj.bjc.6604807.

37. Min, Jinrong, Xing Zhang, Xiaodong Cheng, Shiv I. S. Grewal, and Rui-Ming Xu. 2002. “Structure of the SET Domain Histone Lysine Methyltransferase Clr4.” *Nature Structural & Molecular Biology* 9 (11): 828–32. https://doi.org/10.1038/nsb860.

38. Mitchell, Taylor R. H., Kimberly Glenfield, Kajaparan Jeyanthan, and Xu-Dong Zhu. 2009. “Arginine Methylation Regulates Telomere Length and Stability.” *Molecular and Cellular Biology* 29 (18): 4918–34. https://doi.org/10.1128/MCB.00009-09.

39. Morera, Ludovica, Michael Lübbert, and Manfred Jung. 2016. “Targeting Histone Methyltransferases and Demethylases in Clinical Trials for Cancer Therapy.” *Clinical Epigenetics* 8 (May). https://doi.org/10.1186/s13148-016-0223-4.

40. Murn, Jernej, and Yang Shi. 2017. “The Winding Path of Protein Methylation Research: Milestones and New Frontiers.” *Nature Reviews Molecular Cell Biology* 18 (8): 517–27. https://doi.org/10.1038/nrm.2017.35.

41. Nicholson, Thomas B., Nicolas Veland, and Taiping Chen. 2015. “Chapter 3 - Writers, Readers, and Erasers of Epigenetic Marks.” In *Epigenetic Cancer Therapy*, edited by Steven G. Gray, 31–66. Boston: Academic Press. https://doi.org/10.1016/B978-0-12-800206-3.00003-3.

42. Paik, Woon Ki, Sangduk Kim, and In Kyoung Lim. 2014. “Protein Methylation and Interaction with the Antiproliferative Gene, BTG2/TIS21/Pc3.” *Yonsei Medical Journal* 55 (2): 292–303. https://doi.org/10.3349/ymj.2014.55.2.292.

43. Schubert, Heidi L., Robert M. Blumenthal, and Xiaodong Cheng. 2003. “Many Paths to Methyltransfer: A Chronicle of Convergence.” *Trends in Biochemical Sciences* 28 (6): 329–35. https://doi.org/10.1016/S0968-0004(03)00090-2.

44. Strahl, Brian D., and C. David Allis. 2000. “The Language of Covalent Histone Modifications.” *Nature* 403 (6765): 41–45. https://doi.org/10.1038/47412.

45. Sylvestersen, Kathrine B., Heiko Horn, Stephanie Jungmichel, Lars J. Jensen, and Michael L. Nielsen. 2014. “Proteomic Analysis of Arginine Methylation Sites in Human Cells Reveals Dynamic Regulation During Transcriptional Arrest.” *Molecular & Cellular Proteomics : MCP* 13 (8): 2072–88. https://doi.org/10.1074/mcp.O113.032748.

46. Thandapani, Palaniraja, Timothy R. O’Connor, Timothy L. Bailey, and Stéphane Richard. 2013. “Defining the RGG/RG Motif.” *Molecular Cell* 50 (5): 613–23. https://doi.org/10.1016/j.molcel.2013.05.021.

47. Trojer, Patrick, and Danny Reinberg. 2006. “Histone Lysine Demethylases and Their Impact on Epigenetics.” *Cell* 125 (2): 213–17. https://doi.org/10.1016/j.cell.2006.04.003.

48. Tschiersch, B, A Hofmann, V Krauss, R Dorn, G Korge, and G Reuter. 1994. “The Protein Encoded by the Drosophila Position-Effect Variegation Suppressor Gene Su(Var)3-9 Combines Domains of Antagonistic Regulators of Homeotic Gene Complexes.” *The EMBO Journal* 13 (16): 3822–31.

49. Wang, BingHua, Minghui Wang, and Ao Li. 2017. “Prediction of Post-Translational Modification Sites Using Multiple Kernel Support Vector Machine.” *PeerJ* 5 (April): e3261. https://doi.org/10.7717/peerj.3261.

50. Wei, L., P. Xing, G. Shi, Z. L. Ji, and Q. Zou. 2018. “Fast Prediction of Protein Methylation Sites Using a Sequence-Based Feature Selection Technique.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1–1. https://doi.org/10.1109/TCBB.2017.2670558.

51. Xu, Xiaoyi, Ao Li, Liang Zou, Yi Shen, Wenwen Fan, and Minghui Wang. 2014. “Improving the Performance of Protein Kinase Identification via High Dimensional Protein–Protein Interactions and Substrate Structure Data.” *Molecular BioSystems* 10 (3): 694–702. https://doi.org/10.1039/C3MB70462A.

52. Yang, Yanzhong, and Mark T. Bedford. 2013. “Protein Arginine Methyltransferases and Cancer.” *Nature Reviews Cancer* 13 (1): 37–50. https://doi.org/10.1038/nrc3409.

53. Zhao, Yu, Joshua R. Brickner, Mona C. Majid, and Nima Mosammaparast. 2014. “Crosstalk between Ubiquitin and Other Post-Translational Modifications on Chromatin during DSB Repair.” *Trends in Cell Biology* 24 (7): 426–34. https://doi.org/10.1016/j.tcb.2014.01.005.