ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

Τμήμα Μηχανικών Η/Υ και Πληροφορικής

Εισαγωγή στη Βιοπληροφορική - Προπτυχιακό (CEID1047)

Δεύτερο Σύνολο Ασκήσεων 2021-2022

Στεφανίδης Μάριος — 1067458
Μητροπούλου Αικατερίνα — 1067409

Πάτρα, Ιούνιος 2022

Περιεχόμενα

1 Εισαγωγή

Η παρούσα αναφορά συντάσσεται με αφορμή την εργασία στο μάθημα "Εισαγωγή στην Βιοπληροφορική" που διδάσκεται στο Πανεπιστήμιο Πατρών και παρακολουθείται από φοιτητές των τμημάτων Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών (ΗΜΤΥ) και Μηχανικών Υπολογιστών και Πληροφορικής (CEID). Ο κύριος στόχος είναι η σύντομη παρουσίαση τεσσάρων ερευνητικών κειμένων (papers) που αφορούν στον κλάδο της Βιοπληροφορικής.

2 Πρώτο Ερευνητικό Κείμενο

Στο συγκεκριμένο κεφάλαιο θα αναλυθεί η ερευνητική εργασία με τίτλο **A semi-supervised deep learning approach for predicting the functional effects of genomic non-coding variations** και συγγραφείς τους Hao Jia, Sung-Joon Park και Kenta Nakai.

2.1 Στόχος και Ερευνητικό Ενδιαφέρον

Είναι ευρέως γνωστό πως παραπάνω από το 95% του ανθρώπινου γονιδιώματος αποτελείται από μη-κωδικές αλληλουχίες DNA, οι οποίες δεν μεταφράζονται σε πρωτεΐνες. Πρόσφατες όμως μελέτες απέδειξαν πως οι παραπάνω περιοχές διαδραματίζουν σημαντικό ρόλο στην επιστήμη της Βιολογίας. Για παράδειγμα, μελέτες σε επίπεδο γονιδιώματος έχουν ανακαλύψει πως η πλειονότητα των διάφορων θέσεων στα χρωμοσώματα (variant loci) που σχετίζονται με ανθρώπινες ασθένειες, εντοπίζονται σε μη-κωδικές περιοχές και ρυθμίζουν το γονιδίωμα σε επίπεδο ιστών ή κυττάρων. Μερικές από τις παραπάνω μη-κωδικές περιοχές που έχουν υποστεί μεταλλάξεις επηρεάζουν σε σημαντικό βαθμό τις θέσεις δέσμευσης της μεταγραφής και τις επιγονιδιωματικές τροποποιήσεις που μελετήθηκαν από μεγάλα έργα - όπως είναι το ENCODE [4] και ο οδικός χάρτης Epigenomics [2] - και συνυπάρχουν με μη-κωδικές παραλλαγές που σχετίζονται με ασθένειες και ανθρώπινα χαρακτηριστικά.

Για την καλύτερη κατανόηση των λειτουργικών συνεπειών που επιφέρουν οι μη κωδικές γενετικές παραλλαγές, πολλοί ερευνητές έχουν χρησιμοποιήσει διάφορα υπολογιστικά εργαλεία, όπως είναι:

- Το FUN_LDA [1], ένα μοντέλο μη-εποπτευόμενης λανθάνουσας κατανομής Dirichlet
- Το GenoSkyline [12], ένα μοντέλο το οποίο εκπαιδεύεται από ένα μείγμα πιθανοτικών συστατικών

Οι παραπάνω προσεγγίσεις υπολογίζουν τις βαθμολογίες πρόβλεψης (prediction scores) χρησιμοποιώντας διάφορες τροποποιήσεις ιστόνης και την DNase I υπερευαισθησία. Ωστόσο, υπάρχουν και τα εξής Μοντέλα:

- Το Eigen [7] εφαρμόζει μια μέθοδο μη-εποπτευόμενης φασματικής εκμάθησης
- Το deltaSVMs [10] που είναι μια μηχανή διανυσμάτων υποστήριξης (SVM - Support Vector Machine) που προέρχεται από τον ταξινομητή gkm-SVM και αφορά την αποτελεσματική πρόβλεψη διάφορων ρυθμιστικών παραλλαγών
- Το CADD [9], ένας αλγόριθμος SVM γραμμικού πυρήνα
- Το DANN [18], ένα μοντέλο βαθιάς μάθησης
- Το DeepSEA [21], ένα μοντέλο βαθιάς μάθησης, το οποίο μαθαίνει από μοτίβα αλληλουχίας σε μη-κωδικές περιοχές προκειμένου να προβλέψει ένα ειδικό αλληλόμορφο προφίλ χρωματίνης

Τα τελευταία χρόνια, οι μέθοδοι μη-εποπτευόμενης μάθησης και βαθιάς μάθησης που αναφέρονται παραπάνω έχουν εφαρμοστεί με επιτυχία. Ωστόσο, οι παραπάνω προσεγγίσεις βασίζονται αποκλειστικά στο σύνολο δεδομένων εισόδου ενώ έχουν απρόβλεπτη συμπεριφορά όσον αφορά στα δεδομένα μεγάλης κλίμακας. Σε αυτό το πλαίσιο, πολλές έρευνες έχουν αγνοήσει το γεγονός πως έχει επικυρωθεί η λειτουργία ενός πολύ μικρού μέρους των μη-κωδικών γενετικών περιοχών που έχουν υποστεί μεταλλάξεις από τις εκατομμύρια υπάρχουσες.

Στην συγκεκριμένη μελέτη λοιπόν προτείνεται μια νέα μέθοδος, η οποία χρησιμοποιεί ένα ημί-εποπτευόμενο μοντέλο βαθιάς μάθησης με ψευδό-ετικέτες. Προκειμένου να ξεπεραστεί η σπανιότητα των διαθέσιμων δεδομένων, η συγκεκριμένη μέθοδος εκμεταλλεύεται τα πλεονεκτήματα που προκύπτουν από την εκμάθηση

τόσο από δεδομένα με ετικέτα όσο και από δεδομένα χωρίς ετικέτα. Επιπλέον, αξιοποιούνται διάφορες επιγενετικές παρατηρήσεις και χαρακτηριστικά ακολουθιών που εντοπίζονται σε περιοχές μη-κωδικών παραλλαγών για να συναχθούν οι σημαντικοί παράγοντες του φαινομένου υπό εξέταση.

2.2 Αποτελέσματα Έρευνας

2.2.1 Γενική Δομή του Μοντέλου

Η ημί-εποπτευόμενη μάθηση (SSL) έχει μελετηθεί εκτενώς και έχει γίνει ιδιαίτερα δημοφιλής σε διάφορα ερευνητικά πεδία. Συγκεκριμένα, το SSL παρέχει υψηλής ποιότητας ψευδό-ετικέτες σε μεγάλης κλίμακας δεδομένα που δεν φέρουν ετικέτες κατά τη διάρκεια της εκπαίδευσης, με αποτέλεσμα να επιτρέπει στα νευρωνικά δίκτυα να κάνουν πιο σίγουρες προβλέψεις. Οι ερευνητές εκμεταλλευόμενοι το παραπάνω, ανέπτυξαν ένα SSL μοντέλο για την ανάλυση γενετικών και επιγενετικών υπογραφών που εντοπίζονται σε μη-κωδικές γονιδιωματικές περιοχές μήκους 150 ζευγών βάσεων (150-bp region) όπου έχουν εμφανιστεί μεταλλάξεις.

Ως είσοδος στο νευρωνικό δίκτυο δίνονται τα νουκλεοτίδια μιας περιοχής 150-bp - με κέντρο οποιαδήποτε περιοχή χρωμοσώματος μη-κωδικών παραλλαγών - υπό τη μορφή κωδικοποίησης "one-hot". Ταυτόχρονα, οι περιοχές των 150-bp χαρακτηρίζονται από τρεις συναρτήσεις βαθμολόγησης - Peak, Max και Sum - μετρώντας 10 εμπλουτισμούς ιστόνης και DNase ευαισθησίας, ενώ μετρώνται και 10 διαφορετικοί τύποι συνθέσεων νουκλεοτιδίων. Τα παραπάνω χαρακτηριστικά επιγενετικής και νουκλεοτιδικής σύνθεσης συνδέονται με την έξοδο μιας συνάρτησης max-pooling στη δομή του νευρωνικού δικτύου.

2.2.2 Επίδοση του Μοντέλου και Συμπεράσματα

Για να ελεγχθεί η σκοπιμότητα της συγκεκριμένης προσέγγισης, οι ερευνητές χρησιμοποίησαν παραλλαγές μη-κωδικών περιοχών που εντοπίζονται σε γνωστές ανθρώπινες κυτταρικές σειρές - GM12878, HepG2 και K562 [5]. Δεδομένου ότι οι παραπάνω σειρές έχουν δοκιμαστεί εκτενώς στο ENCODE, θα μπορούσαν να έχουν πρόσβαση σε μεγάλης κλίμακας γονιδιωματικά και επιγονιδιωματικά δεδομένα, τα οποία μπορούν να χρησιμοποιηθούν για τον χαρακτηρισμό διάφορων θέσεων στα χρωμοσώματα σε γονιδιωματική κλίμακα.

Όσον αφορά στην κυτταρική σειρά K562, οι βαθμολογίες Max και Sum για επιγενετικά σημάδια επέδειξαν ευρείες περιοχές κατανομής ενώ τα μοτίβα τους ήταν παρόμοια μεταξύ τους. Οι παραπάνω βαθμολογίες ήταν παρόμοιες με αυτές της κατανομής των παραλλαγών σε μη-κωδικές περιοχές, ενώ δεν παρατηρήθηκε το ίδιο στα χαρακτηριστικά σύνθεσης νουκλεοτιδίων. Ιδιαίτερο ενδιαφέρον αποτέλεσε το γεγονός πως το χαρακτηριστικό της ευαισθησίας DNase συσχετίστηκε ισχυρά με τις παραλλαγές μη-κωδικών περιοχών σε όλες τις περιπτώσεις. Τέλος, με τη χρήση ενός διαφορετικού συνόλου δεδομένων, το μοντέλο αν και έφτασε σε απόδοση 0.75 σε AUC στο GM12878, δεν επέδειξε δραστικές διαφορές μεταξύ των κυτταρικών σειρών.

Προκειμένου να εξετάσουν ποια επιγενετικά χαρακτηριστικά συνέβαλαν περισσότερο στην απόδοση, ομαδοποίησαν τα παραπάνω σε 6 λειτουργικές κατηγορίες - I.Ενισχυτής, II.Προαγωγέας, III.Δομικά Σημάδια, IV.Ετεροχρωματίνη, V.Μεταγραφόμενο γονίδιο-σώμα και VI.Προσβασιμότητα. Τα μοντέλα που βασίζονται στο Max-score έδειξαν υψηλότερες τιμές AUC στις κατηγορίες I, II και VI με την τελευταία να συμβάλει σημαντικά στην απόδοση. Αντιθέτως, χαρακτηριστικά που βασίζονται σε νουκλεοτίδια ήταν λιγότερο αποτελεσματικά.

Συμπερασματικά, η προσβασιμότητα στο DNA είναι κυρίως αυτή που εξηγεί την παρουσία λειτουργικών παραλλαγών σε μη-κωδικές περιοχές όσον αφορά στις παραπάνω κυτταρικές σειρές.

2.2.3 Σε Σύγκριση με τα Υπόλοιπα Μοντέλα

Οι ερευνητές προκειμένου να επιδείξουν την καλύτερη απόδοση του μοντέλου τους (SSL_dnn), συνέκριναν το τελευταίο με επτά υπάρχοντα μη-εποπτευόμενα μοντέλα που αναφέρθηκαν και παραπάνω, χρησιμοποιώντας τις βαθμολογίες πρόβλεψης και τα δεδομένα από προηγούμενες μελέτες. Εφαρμόζοντας λοιπόν στο μοντέλο το ίδιο σύνολο δεδομένων επικύρωσης και σχεδιάζοντας τις AUC καμπύλες, το SSL_dnn εμφάνισε υψηλότερες τιμές AUC και πιο συγκεκριμένα 0,75 στο GM12878, 0,71 στο HepG2 και 0,69 στο K562.

Στη συνέχεια, συνέκριναν το SSL_dnn με ένα εποπτευόμενο βαθύ νευρωνικό δίκτυο χωρίς ψευδοετικέτες. Αν και οι δύο ταξινομητές έδειξαν παρόμοιες τάσεις ανάπτυξης για τις τιμές AUC στην αρχή, το SSL_dnn απέκτησε σταδιακά καλύτερη απόδοση καθώς η εποχή αυξήθηκε στην κυτταρική σειρά K562.

Ακόμα, όταν η συνάρτηση απώλειας διασταυρούμενης εντροπίας του SSL_dnn άρχισε να ενσωματώνει την απώλεια των δεδομένων που δεν φέρουν ετικέτα, η απόδοση του μοντέλου ήταν εξαιρετική, γεγονός που υποδηλώνει πόσο καίρια είναι η συμβολή των ψευδοετικετών.

2.3 Μεθοδολογία

2.3.1 Σύνολα Δεδομένων

Χρησιμοποιήθηκαν παραλλαγές μη-κωδικών περιοχών και οι αντίστοιχες ετικέτες τους στις κυτταρικές περιοχές GM12878, HepG2 και K562. Πιο συγκεκριμένα δόθηκε η ετικέτα 1 για θετικούς τόπους (loci) που επηρεάζουν τη γονιδιακή ρύθμιση και ετικέτα 0 για αρνητικούς τόπους που δεν έχουν καμία σχέση με την γονιδιακή έκφραση. Επιπλέον, χρησιμοποιήθηκαν και επεξεργασμένα σύνολα δεδομένων των τροποποιήσεων ιστόνης και της ευαισθησίας DNase I από το ENCODE.

2.3.2 Δημιουργία Χαρακτηριστικών Διανυσμάτων (feature vectors)

Οι ερευνητές αφού έλεγξαν την επικάλυψη των μη-κωδικών περιοχών με τα επιγενετικά σημάδια, δημιούργησαν τρία χαρακτηριστικά διανύσματα:

- Peak, 1 για παραλλαγές σε μη-κωδικές περιοχές τοποθετημένες σε μια περιοχή κορυφής ενός επιγενετικού σημείου, διαφορετικά 0
- Max, η βαθμολογία του μέγιστου εμπλουτισμού σε μια περιοχή 150-bp που επικεντρώνεται από μια παραλλαγή σε μη-κωδική περιοχή
- Sum, το άθροισμα της βαθμολογίας εμπλουτισμού για την περιοχή των 150 bp

Έπειτα, υπολόγισαν τις νουκλεοτιδικές συνθέσεις της περιοχής των 150-bp - αριθμός μονονουκλεοτιδίων, αριθμός δινουκλεοτιδίων, αριθμός GC, αριθμός GT, αριθμός GA και λοξότητα - ενώ κωδικοποίησαν κάθε βάση στην παραπάνω περιοχή υιοθετώντας την τεχνική "one-hot".

2.3.3 Δημιουργία ενός Μοντέλου Βαθιάς Μάθησης

Το μοντέλο βαθιάς μάθησης αποτελείται από δύο συνελικτικά νευρωνικά στρώματα που αφορούν την μήτρα κώδικα με σχήμα 150 (μήκος ακολουθίας) $\times$ 4 (μέγεθος της κωδικοποίησης "one-hot"). Τα μεγέθη καναλιών εξόδου στα συνελικτικά στρώματα είναι 2 και 4, αντίστοιχα. Το πρώτο συνελικτικό νευρικό στρώμα χρησιμοποιεί ένα συνελικτικό φίλτρο (1×4) χωρίς επένδυση για την εξαγωγή πληροφοριών από λεξιλόγια νουκλεοτιδίων, ενώ το δεύτερο εφαρμόζει ένα φίλτρο (2×4) και ένα βήμα (2×1) διασκελισμού. Εφαρμόστηκε ακόμη μια συνάρτηση εγκατάλειψης ως τρίτο επίπεδο. Αυτή η λειτουργία εκχωρεί τυχαία μηδενικά για ορισμένες κρυφές μονάδες, με αποτέλεσμα να παραλείπονται κατά τη διάρκεια της εκπαίδευσης, γεγονός που συμβάλλει στην ελαχιστοποίηση της υπερπροσαρμογής. Επιπλέον, γίνεται χρήση ενός στρώματος max-pooling με μέγεθος πυρήνα (2×2), διατηρώντας τις μέγιστες τιμές στα παράθυρα και αφήνοντας ένα πυκνό χαρακτηριστικό χάρτη με μέγεθος ($4 \times 1 \times 72$) στο επόμενο επίπεδο. Τέλος, χρησιμοποιήθηκε η συνάρτηση ReLU (RectifieLinear Units) ως μέθοδος ενεργοποίησης για κάθε νευρωνική μονάδα.

Το μοντέλο περιελάμβανε τρία πλήρως συνδεδεμένα στρώματα (FC), τα οποία είναι επίσης γνωστά ως πυκνά στρώματα, με μεγέθη 40, 10 και 2, αντίστοιχα. Η είσοδος στο πρώτο επίπεδο FC δημιουργείται με τη σύνδεση της εξόδου της συνάρτησης max pooling με τον πρόσθετο χάρτη των χαρακτηριστικών επιγενετικής και νουκλεοτιδικής σύνθεσης. Πρόσθεσαν ακόμα τη συνάρτηση εγκατάλειψης και τη συνάρτηση κανονικοποίησης παρτίδας στο πρώτο και το δεύτερο επίπεδο FC, κάνοντας μη γραμμικούς μετασχηματισμούς για τα εισερχόμενα δεδομένα. Μετά την τρίτη στρώση FC εφαρμόστηκε η λειτουργία ενεργοποίησης ReLU. Το τελικό επίπεδο εξόδου αποτελείται από δύο νευρωνικές μονάδες που αντιστοιχούν στην πιθανότητα δύο ταξινομήσεων.

2.3.4 Δημιουργία ενός Μοντέλου Ημί-Εποπτευόμενης Μάθησης με τη Χρήση "Ψευδό-Ετικετών"

Η έννοια του μοντέλου εκπαίδευσης με ψευδό-ετικέτες για πραγματικά δεδομένα μεγάλης κλίμακας και χωρίς ετικέτες έχει αποδειχθεί. Πιο συγκεκριμένα, αποτελείται από μια συνάρτηση που αντιστοιχίζει απευθείας το χώρο εισόδου σε βαθμολογίες "εμπιστοσύνης" (confidence scores) και την έξοδο, η οποία είναι

ένα δισδιάστατο διάνυσμα για κάθε χαρακτηριστικό χάρτη εισόδου. Το δίκτυο εκπαιδεύεται ελαχιστοποιώντας την απώλεια της διασταυρούμενης εντροπίας.

Για την εκπαίδευση του νευρωνικού δικτύου βαθιάς μάθησης, το σύνολο δεδομένων μιας κυτταρικής γραμμής χωρίστηκε σε τρία μέρη: Ι.επισημασμένα σύνολα δεδομένων για εκπαίδευση, ΙΙ. μη-επισημασμένα σύνολα δεδομένων για εκπαίδευση και ΙΙΙ.ένα σύνολο δεδομένων επικύρωσης για δοκιμή. Προκειμένου να εξισορροπηθεί το σύνολο Ι και ΙΙΙ, οι εναπομείναντες θετικοί τόποι (loci) είναι πολύ λιγότεροι από τους αρνητικούς τόπους στο σύνολο ΙΙ. Χρησιμοποιώντας τα σύνολα δεδομένων εκπαίδευσης, πραγματοποιήθηκε η επαναληπτική διαδικασία εκπαίδευσης με την αρχικοποίηση τυχαίων παραμέτρων. Η διαδικασία με το σύνολο Ι παρακολουθήθηκε από έναν εποπτευόμενο όρο απώλειας και στη συνέχεια το σύνολο ΙΙ προβλέφθηκε με το εκπαιδευόμενο μοντέλο. Η κλάση που είχε τη μέγιστη προβλεπόμενη πιθανότητα στο δισδιάστατο διάνυσμα εξόδου επιλέχθηκε ως η "πραγματική" ετικέτα για την εκπαίδευση του μοντέλου. Έπειτα, υπολογίστηκε η απώλεια διασταυρούμενης εντροπίας για τη βελτιστοποίηση του μοντέλου. Σημειώνεται ακόμα πως ο αριθμός των μη επισημασμένων δεδομένων μειώνεται κατά τη διάρκεια της επανάληψης καθώς τα δεδομένα χωρίς ετικέτα με τις πιο σίγουρες ψευδοετικέτες προστίθενται στα επισημασμένα σύνολα δεδομένων που θα χρησιμοποιηθούν στην επόμενη εποχή.

Η αρχικοποίηση των παραμέτρων πραγματοποιήθηκε με μια συνάρτηση στοχαστικής κλίσης κατάβασης, η οποία ενημέρωνε τις παραμέτρους με ρυθμό εκμάθησης 0.03. Ορίστηκαν ακόμα τα μεγέθη mini-batch για τα επισημασμένα και μη επισημασμένα σύνολα δεδομένων εκπαίδευσης και για το σύνολο δεδομένων επικύρωσης σε 16, 32 και 20, αντίστοιχα. Το όριο για την επιλογή των εμπιστευτικών ψευδό-ετικέτες ήταν 0,95. Χρησιμοποιήθηκαν $T_1 = 100$, $T_2 = 600$.

2.4 Συμπεράσματα

Είναι προφανές λοιπόν πως η αξιοποίηση των ψευδό-ετικετών συνέβαλλε αφενός στην μέγιστη απόδοση του μοντέλου βαθιάς μάθησης και αφετέρου στη μελέτη διάφορων βιολογικών φαινομένων με τη χρήση μικρότερου αριθμού πειραματικά επιβεβαιωμένων δεδομένων και μεγάλου αριθμού δεδομένων που δεν φέρουν ετικέτα. Ακόμα, αφού οι ερευνητές εξασφάλισαν όσο το δυνατόν γίνεται πιο δίκαιες συγκρίσεις και σε συνδυασμό με τα παραπάνω, αποδείχθηκε πως η προσέγγισή τους ήταν καλύτερη σε σύγκριση με ήδη υπάρχοντα μοντέλα.

Όσον αφορά στα χαρακτηριστικά που επηρέασαν άμεσα την πρόβλεψη του μοντέλου, διαπιστώθηκε πως η προσβασιμότητα στο DNA που αντανακλά την κατάσταση ανοιχτής χρωματίνης [3] είναι το πιο σημαντικό από τα χαρακτηριστικά, μιας και παρουσίασε υψηλότερη συσχέτιση με την κατανομή των λειτουργικών παραλλαγών σε μη-κωδικές περιοχές. Αντίθετα, τα χαρακτηριστικά που βασίζονται στις νουκλεοτιδικές συνθέσεις ήταν λιγότερο αποτελεσματικά. Αξίζει να σημειωθεί πως το αναφερθέν μοντέλο έχει εκπαιδευτεί με το σύνολο δεδομένων μιας συγκεκριμένης κυτταρικής γραμμής, επομένως είναι απίθανο να πετύχει προβλέψεις σε παραλλαγές άλλων κυτταρικών σειρών. Αυτό άλλωστε υποδηλώνει πως οι ειδικοί για τον τύπο κυττάρου επιγενετικοί παράγοντες που σχετίζονται με τη διαμόρφωση ανοιχτής χρωματίνης αλληλεπιδρούν με τις λειτουργικές παραλλαγές μη-κωδικών περιοχών.

Συμπερασματικά, το προτεινόμενο ημί-εποπτευόμενο μοντέλο βαθιάς μάθησης σε συνδυασμό με τη χρήση ψευδό-ετικετών είναι ιδιαίτερα χρήσιμο στη μελέτη βιολογικών φαινομένων με περιορισμένα σύνολα δεδομένων. Το παραπάνω μοντέλο σε συνδυασμό με την μελέτη των ερευνητών παρουσίασε μια αποτελεσματική προσέγγιση για την εύρεση μεταλλάξεων σε μη-κωδικές περιοχές που πιθανώς σχετίζονται με διάφορες ασθένειες.

3 Δεύτερο Ερευνητικό Κείμενο

Στο συγκεκριμένο κεφάλαιο θα αναλυθεί η ερευνητική εργασία με τίτλο **DEEPLY ESSENTIAL: a deep neural network for predicting essential genes in microbes** και συγγραφείς τους Md Abid Hasan και Stefano Lonardi.

3.1 Στόχος και Ερευνητικό Ενδιαφέρον

Βασικά γονίδια ονομάζονται εκείνα τα γονίδια που είναι κρίσιμα για την επιβίωση και την αναπαραγωγή ενός οργανισμού ενώ ταυτόχρονα αποτελούν ακρογωνιαίο λίθο για την εξέταση της προέλευσης και της εξέλιξης των οργανισμών. Συνεπώς, η μελέτη τους μπορεί να συμβάλλει στη δημιουργία αφενός

νέων αντιμικροβιακών/αντιβιοτικών φαρμάκων και αφετέρου τεχνητών αυτοσυντηρούμενων ζωντανών κυττάρων με ελάχιστο γονιδίωμα.

Η αναγνώριση βασικών γονιδίων μέσω πειραμάτων σε υγρό εργαστήριο είναι εντατική, δαπανηρή και χρονοβόρα, αφού αποτελείται από μια σειρά δύσκολων διαδικασιών, οι οποίες πολλές φορές παράγουν αντιφατικά αποτελέσματα. Με τις πρόσφατες εξελίξεις στην τεχνολογία προσδιορισμού αλληλουχίας υψηλής απόδοσης, έχουν προκύψει διάφορες υπολογιστικές μέθοδοι, όπως είναι η χαρτογράφηση ομολογίας (homology mapping) [14], που διευκολύνουν τις παραπάνω διαδικασίες. Με την εισαγωγή βάσεων δεδομένων γονιδίων - όπως είναι οι DEG, CEG και OGEE [13]-, οι ερευνητές σχεδίασαν πιο σύνθετα μοντέλα πρόβλεψης, χρησιμοποιώντας ένα ευρύτερο σύνολο χαρακτηριστικών, το οποίο είναι:

1. Χαρακτηριστικά αλληλουχίας, π.χ. συχνότητα κωδικονίων, μήκος γονιδίου
2. Τοπολογικά χαρακτηριστικά, δηλ. κεντρικός βαθμός, συντελεστής συστάδας
3. Λειτουργικά χαρακτηριστικά, π.χ. ομολογία και άλλες μοριακές ιδιότητες

Αξίζει να σημειωθεί πως σύμφωνα με μελέτες η τρισδιάστατη δομή των πρωτεϊνών μπορεί να ενσωματωθεί στο σύνολο των τοπολογικών χαρακτηριστικών.

Τα χαρακτηριστικά που ανήκουν στην πρώτη κατηγορία μπορούν να ληφθούν απευθείας από την πρωτογενή αλληλουχία DNA ενός γονιδίου και την αντίστοιχη αλληλουχία πρωτεΐνης του. Τα χαρακτηριστικά της τρίτης κατηγορίας απαιτούν γνώση του δικτύου αλληλεπίδρασης πρωτεΐνης-πρωτεΐνης - STRING [20] και HumanNET [6]. Η γονιδιακή έκφραση και οι πληροφορίες του λειτουργικού τομέα μπορούν να προέρχονται από βάσεις δεδομένων όπως είναι οι PROSITE και PFAM.

Έχουν δημοσιευτεί αρκετές μελέτες και υπολογιστικά εργαλεία σχετικά με το πρόβλημα της πρόβλεψης βασικών γονιδίων από τις αλληλουχίες τους. Αναφέρονται παρακάτω ενδεικτικά:

- Μελέτη [19] που ανέπτυξε το εργαλείο ZUPLS, το οποίο χρησιμοποιεί μια καμπύλη Z που προέρχεται από την αλληλουχία, χαρτογράφηση ομολογίας και βαθμολογία εμπλουτισμού τομέα για την πρόβλεψη βασικών γονιδίων σε δώδεκα προκαρυωτές μετά την εκπαίδευση του μοντέλου σε δύο βακτήρια. Αν και το ZUPLS λειτούργησε καλά στην πρόβλεψη διασταυρούμενων οργανισμών, ο περιορισμένος αριθμός βακτηριακών ειδών που χρησιμοποιήθηκαν στο σετ εκπαίδευσης, δημιούργησε αμφιβολίες για την ικανότητα του να γενικευτεί σε διαφορετικά βακτηριακά είδη
- Μελέτη [11] που πρότεινε τη χρήση του PCA σε χαρακτηριστικά που προέρχονται από τη γονιδιακή αλληλουχία, πρωτεϊνικούς τομείς, ομόλογα και τοπολογικές πληροφορίες
- Μελέτη [16], στην οποία χρησιμοποιήθηκαν νουκλεοτίδιο, δινουκλεοτίδιο, κωδικόνιο, συχνότητες αμινοξέων και ανάλυση χρήσης κωδικονίων για την πρόβλεψη της ουσιαστικότητας σε δεκαέξι βακτηριακά είδη. Οι συγγραφείς χρησιμοποίησαν CD-HIT για την ανίχνευση ομολογίας τόσο σε βασικά όσο και σε μη βασικά γονίδια
- Μελέτη [17], στην οποία χρησιμοποίησαν αρκετά γονιδιωματικά, φυσικοχημικά και υποκυτταρικά χαρακτηριστικά εντοπισμού για την πρόβλεψη της γονιδιακής ουσιαστικότητας σε δεκατέσσερα βακτηριακά είδη. Οι συγγραφείς αντιμετώπισαν πρόβλημα με τον πλεονασμό στο σύνολο δεδομένων (δηλαδή, κοινά ομόλογα γονίδια από πολλαπλά βακτηριακά γονιδιώματα) με ομαδοποίηση γονιδίων με βάση τις ομοιότητες της αλληλουχίας τους.
- Μελέτη [15], στην οποία οι συγγραφείς αναγνώρισαν βασικά γονίδια σε δεκαπέντε βακτηριακά είδη χρησιμοποιώντας θεωρητικά χαρακτηριστικά πληροφοριών, όπως είναι η απόκλιση Kullback-Leibler μεταξύ της κατανομής των k-mers (για k = 1, 2, 3), η αμοιβαία πληροφόρηση υπό όρους και η εντροπία. Αν και το έργο τους έδειξε πολλά υποσχόμενα αποτελέσματα για προβλέψεις ενδοοργανισμών και διασταυρούμενων οργανισμών, το μοντέλο είχε μάλλον κακή απόδοση όταν εκπαιδεύτηκε στο πλήρες βακτηριακό σύνολο δεδομένων.

Η πιο πρόσφατη εργασία για την πρόβλεψη ουσιαστικότητας γονιδίων χρησιμοποιεί χαρακτηριστικά που βασίζονται σε δίκτυο, Lasso για επιλογή χαρακτηριστικών και τον αλγόριθμο Random Forest ως ταξινομητή. Οι συγγραφείς χρησιμοποίησαν μια αναδρομική τεχνική εξαγωγής χαρακτηριστικών για να υπολογίσουν 267 χαρακτηριστικά σε τρεις διαφορετικές κατηγορίες, δηλαδή τοπικά χαρακτηριστικά όπως η κατανομή βαθμών, τα χαρακτηριστικά egonet που αναφέρονται στον κόμβο και το υπαγόμενο υπογράφημα που σχηματίζεται από όλους τους γείτονες του και τοπικά χαρακτηριστικά που αποτελούν ένα συνδυασμό

τοπικών και egonet χαρακτηριστικών. Χρησιμοποίησαν επίσης δεκατέσσερα μέτρα κεντρικότητας δικτύου ως ξεχωριστό σύνολο χαρακτηριστικών για την πρόβλεψη ουσιαστικότητας. Τέλος, συνδύασαν τα χαρακτηριστικά τους που βασίζονται στο δίκτυο με τα χαρακτηριστικά που βασίζονται σε ακολουθία για το μοντέλο πρόβλεψής τους. Για τη δημιουργία των μοντέλων, οι συγγραφείς πραγματοποίησαν δειγματοληψία σε μη βασικά γονίδια για την εξισορρόπηση του συνόλου εκπαίδευσης, αλλά δεν συνειδητοποίησαν ότι το σύνολο δεδομένων τους περιείχε πολλαπλά αντίγραφα ομόλογων γονιδίων που δημιούργησαν ένα ζήτημα «διαρροής δεδομένων» με αποτέλεσμα να επηρεάσει αρνητικά τα αποτελέσματά τους.

Στην αναφερθέν ερευνητική εργασία λοιπόν προτείνεται ένα βαθύ νευρωνικό δίκτυο τροφοδοσίας (DNN) που ονομάζεται DEEPLY ESSENTIAL που χρησιμοποιεί χαρακτηριστικά που προέρχονται αποκλειστικά από την κύρια γονιδιακή αλληλουχία, μεγιστοποιώντας έτσι την πρακτική εφαρμογή του.

3.2 Εργαλεία και Μεθοδολογίες

3.2.1 Σύνολο Δεδομένων και Επιλογή Χαρακτηριστικών

Τα γονιδιωματικά δεδομένα για τριάντα βακτηριακά είδη ελήφθησαν από τη βάση δεδομένων DEG, η οποία είναι μια ολοκληρωμένη αποθήκη πειραματικά καθορισμένων βακτηριακών και αρχαιοειδών βασικών γονιδίων. Μεταξύ των τριάντα βακτηριακών ειδών, τα εννέα είναι θετικά κατά Gram (GP) και τα είκοσι ένα είναι Gram-αρνητικά (GN). Το DEG παρέχει την πρωτογενή αλληλουχία DNA και την αντίστοιχη πρωτεϊνική αλληλουχία τόσο για βασικά όσο και για μη βασικά γονίδια, καθώς και γονιδιακές λειτουργικές παρατηρήσεις. Επιλέχθηκαν μόνο γονίδια που κωδικοποιούν πρωτεΐνες, συνεπώς αποκλείστηκαν γονίδια RNA, ψευδογονίδια και άλλα μη-κωδικά γονίδια. Παρατηρήθηκε, ωστόσο, πως το σύνολο δεδομένων είναι εξαιρετικά μη-ισορροπημένο, με αποτέλεσμα οι ερευνητές να μειώσουν τη δειγματοληψία σε μη-ουσιαστικής σημασίας γονίδια προκειμένου να βελτιώσουν την απόδοση του ταξινομητή.

Όσον αφορά στα χαρακτηριστικά, το DEEPLY ESSENTIAL χρησιμοποιεί τη συχνότητα κωδικονίων, τη μέγιστη σχετική χρήση συνώνυμου κωδικονίου, το δείκτη προσαρμογής κωδικονίων, το μήκος γονιδίου και το περιεχόμενο GC. Εκτός βέβαια από τα χαρακτηριστικά που προέρχονται από το DNA, χρησιμοποιούνται και οι συχνότητες αμινοξέων και το μήκος της πρωτεϊνικής αλληλουχίας.

3.2.2 Σπουδαιότητα Γονιδίων

Η συχνότητα κωδικονίων έχει αναγνωριστεί ως ένα σημαντικό χαρακτηριστικό για την πρόβλεψη της σπουδαιότητας του γονιδίου. Δεδομένης της πρωτογενούς αλληλουχίας DNA ενός γονιδίου, η συχνότητα κωδικονίων του υπολογίζεται ολισθαίνοντας ένα παράθυρο τριών νουκλεοτιδίων κατά μήκος του γονιδίου.

Άλλα διακριτικά χαρακτηριστικά που επιδεικνύουν την σπουδαιότητα των γονιδίων είναι το μήκος γονιδίου και το περιεχόμενο GC. Όσον αφορά στο μήκος γονιδίου, παρατηρήθηκε ότι τα βασικά γονίδια στο σύνολο δεδομένων GP (Gram-Positive) είναι κατά μέσο όρο μεγαλύτερα από τα μη-βασικά γονίδια. Σχετικά με το περιεχόμενο GC, είναι άλλο ένα χαρακτηριστικό που παρέχει πληροφορίες για την πρόβλεψη της σπουδαιότητας γονιδίου. Παρατηρήθηκε πως τα μη-βασικά γονίδια έχουν υψηλότερη περιεκτικότητα σε GC σε σύγκριση με τα βασικά γονίδια.

Ακόμη ένα ενημερωτικό σύνολο χαρακτηριστικών που χρησιμοποιείται για την πρόβλεψη της σπουδαιότητας των γονιδίων είναι αυτά που προέρχονται από τις αντίστοιχες πρωτεϊνικές αλληλουχίες. Συνεπώς, το DEEPLY ESSENTIAL χρησιμοποιεί τα χαρακτηριστικά συχνότητες αμινοξέων και μήκη πρωτεϊνικών αλληλουχιών.

Τα παραπάνω χαρακτηριστικά τα οποία σε σύνολο είναι 89 συνδυάζονται και υπολογίζονται σύμφωνα με μια πρωτογενή αλληλουχία DNA ενός γονιδίου αλλά και της πρωτεϊνικής του αλληλουχίας, προκειμένου να δημιουργηθεί το ζητούμενο μοντέλο.

3.2.3 Χρήση Συνώνυμων Κωδικωνίων και Δείκτης Προσαρμογής Κωδικονίων

Η μη ισορροπημένη χρήση συνώνυμων κωδικονίων συναντάται τόσο στους προκαρυωτικούς όσο και στους ευκαρυωτικούς οργανισμούς. Ο βαθμός μεροληψίας ποικίλλει μεταξύ γονιδίων όχι μόνο σε διαφορετικά είδη αλλά και μεταξύ γονιδίων του ίδιου είδους. Οι διαφορές στη χρήση κωδικονίων σε ένα γονίδιο σε σύγκριση με τα γύρω γονίδια μπορεί να συνεπάγονται την ξένη προέλευσή του, διαφορετικούς λειτουργικούς περιορισμούς ή διαφορετική τοπική μετάλλαξη. Ως αποτέλεσμα, η εξέταση της χρήσης κωδικονίων βοηθά στον εντοπισμό αλλαγών στις μεταλλάξεις μεταξύ των γονιδιωμάτων. Για τον υπολογισμό της σχετικής χρήσης συνώνυμων κωδικονίων συγκρίνεται ο παρατηρούμενος αριθμός εμφάνισης κάθε κωδικονίου με τον αναμενόμενο αριθμό εμφανίσεων.

Ο δείκτης προσαρμογής κωδικονίων (CAI) εκτιμά την προκατάληψη προς ορισμένα κωδικόνια που είναι πιο κοινά σε γονίδια υψηλής έκφρασης. Το CAI ορίζεται ως ο γεωμετρικός μέσος όρος των στατιστικών σχετικής προσαρμοστικότητας. Η σχετική προσαρμοστικότητα για ένα κωδικόνιο ορίζεται ως η σχετική συχνότητα του κωδικονίου σε ένα ειδικό για το είδος σύνολο αναφοράς γονιδίων υψηλής έκφρασης. Το CAI βασίζεται στον αριθμό των κωδικονίων στο γονίδιο εξαιρουμένης της μεθειονίνης, της τρυπτοφάνης και του κωδικονίου τερματισμού. Το εύρος του CAI είναι (0, 1] όπου οι υψηλότερες τιμές υποδεικνύουν υψηλότερη αναλογία στα πιο άφθονα κωδικόνια.

3.2.4 Perceptron Πολλαπλών Στρωμάτων

Η αρχιτεκτονική του DEEPLY ESSENTIAL υλοποιεί ένα πολυστρωματικό perceptron (MLP). Πιο συγκεκριμένα, αποτελείται από ένα επίπεδο εισόδου, πολλαπλά κρυφά επίπεδα και ένα επίπεδο εξόδου. Το τελευταίο κωδικοποιεί την πιθανότητα ένα γονίδιο να είναι απαραίτητο. Ακόμα, η προσθήκη ενός στρώματος εγκατάλειψης καθιστά το δίκτυο λιγότερο ευαίσθητο στο θόρυβο κατά τη διάρκεια της εκπαίδευσης και αυξάνει την ικανότητα του να γενικεύει. Αυτό το επίπεδο εκχωρεί τυχαία μηδενικά βάρη σε ένα μικρό μέρος των νευρώνων στο δίκτυο.

Σε ένα γενικό πλαίσιο, η έξοδος εξαρτάται από την είσοδο του προηγούμενου επιπέδου ενώ συσχετίζεται με την συνάρτηση ενεργοποίησης, την πόλωση και τη μήτρα βάρους που αφορά στις ακμές του δικτύου. Κατά τη φάση της εκπαίδευσης, το δίκτυο μαθαίνει τα βάρη και την προκατάληψη. Ακόμα, το DEEPLY ESSENTIAL χρησιμοποιεί μια γραμμική μονάδα (ReLU) σε κάθε νευρώνα των κρυφών στρωμάτων, η οποία περιορίζει όλες τις αρνητικές τιμές στο μηδέν. Στο επίπεδο εξόδου χρησιμοποιείται ένα σιγμοειδές ως συνάρτηση ενεργοποίησης για την εκτέλεση της διακριτής ταξινόμησης ενώ η συνάρτηση απώλειας είναι μια δυαδική διασταυρούμενη εντροπία που εξαρτάται από τον αριθμό των κλάσεων (2 στην συγκεκριμένη περίπτωση), έναν δυαδικό δείκτη σε περίπτωση που η ετικέτα κλάσης c είναι η σωστή ταξινόμηση για την παρατήρηση o και την p, μια προβλεπόμενη πιθανότητα πως η παρατήρηση o ανήκει στην κατηγορία c.

3.3 Επεξήγηση Παραμέτρων και Τρόπος Αξιολόγησης

3.3.1 Υπερπαράμετροι Μοντέλου

Ο αριθμός των κρυφών επιπέδων, ο αριθμός των κόμβων σε κάθε ένα από τα κρυφά επίπεδα, το μέγεθος παρτίδας, το ποσοστό εγκατάλειψης και ο τύπος του βελτιστοποιητή επιλέχθηκαν με στόχο την βελτιστοποίηση της απόδοσης του ταξινομητή κατά τη διασταυρούμενη επικύρωση. Το τελικό πλήρως συνδεδεμένο στρώμα μειώνει το διάνυσμα διαστάσεων 1024 σε ένα δισδιάστατο διάνυσμα που αντιστοιχεί στις δύο κατηγορίες πρόβλεψης (ουσιώδες/μη ουσιώδες). Η σιγμοειδής συνάρτηση ενεργοποίησης αναγκάζει την έξοδο των δύο νευρώνων στο στρώμα εξόδου να αθροιστεί σε ένα. Έτσι, η τιμή εξόδου αντιπροσωπεύει την πιθανότητα κάθε κλάσης. Ως βελτιστοποιητής επιλέχθηκε ο adadelta, μιας και δεν έχει παραμέτρους, επομένως δεν χρειάστηκε να οριστεί κάποιο ποσοστό εκμάθησης. Η εκπαίδευση διεξήχθη για 100 εποχές με κριτήρια πρόωρης διακοπής.

Το DEEPLY ESSENTIAL εκπαιδεύτηκε σε τρία σύνολα δεδομένων - GP, GN και GP+GN. Για κάθε σύνολο δεδομένων, 80% δεδομένα χρησιμοποιήθηκαν για εκπαίδευση, 10% δεδομένα για επικύρωση και 10% δεδομένα για δοκιμές. Η τυχαία επιλογή επαναλήφθηκε δέκα φορές, δηλαδή έγινε δεκαπλάσια διασταυρούμενη επικύρωση για να προκύψει το συμπέρασμα.

3.3.2 Μέτρα Αξιολόγησης

Η απόδοση του DEEPLY ESSENTIAL αξιολογήθηκε χρησιμοποιώντας την AUC περιοχή της χαρακτηριστικής καμπύλης λειτουργίας του δέκτη (ROC). Το διάγραμμα ROC αντιπροσωπεύει την αντιστάθμιση μεταξύ ευαισθησίας και ειδικότητας για όλα τα πιθανά κάτω όρια ενώ υποδεικνύει επίσης τη σχέση μεταξύ του αριθμού των δειγμάτων εκπαίδευσης και της σταθερότητας της απόδοσης της πρόβλεψης. Έχουν ακόμη βεβαία χρησιμοποιηθεί και κάποια πρόσθετα μέτρα απόδοσης που αντιπροσωπεύουν τον αριθμό των αληθινών θετικών, αληθινών αρνητικών, ψευδών θετικών και ψευδών αρνητικών τιμών.

3.4 Αποτελέσματα Έρευνας

3.4.1 Πρόβλεψη Σπουδαιότητας Γονιδίων

Το DEEPLY ESSENTIAL απέδωσε μια περιοχή κάτω από την καμπύλη 0,838, 0,829 και 0,842 για GP, GN και GP+GN κατά μέσο όρο, αντίστοιχα. Γίνεται προφανώς ότι η απόδοση του DEEPLY ESSENTIAL ήταν πιο σταθερή στο σύνολο δεδομένων GP+GN από το σύνολο δεδομένων GP (το οποίο περιέχει τον μικρότερο αριθμό δειγμάτων). Οι καμπύλες ανάκλησης ακρίβειας δείχνουν την ικανότητα του μοντέλου να αποδίδει σταθερά χαμηλό ποσοστό ψευδώς θετικών τιμών και χαμηλό ποσοστό ψευδώς αρνητικών ποσοστών σε όλα τα σύνολα δεδομένων.

3.4.2 Σύγκριση με Μεθόδους Μείωσης Δειγματοληψίας

Όπως αναφέρθηκε παραπάνω το σύνολο δεδομένων της σπουδαιότητας των γονιδίων ήταν εξαιρετικά ανισόρροπο, γεγονός που μπορεί να επηρεάσει αρνητικά την απόδοση ενός ταξινομητή. Προκειμένου να γίνει αντιληπτό με ποιον τρόπο η ανισορροπία της κλάσης επηρεάζει την απόδοση του παραπάνω, το DEEPLY ESSENTIAL εκπαιδεύτηκε σε ένα πλήρες μη-ισορροπημένο σύνολο δεδομένων. Παρατηρήθηκε λοιπόν πως η ευαισθησία και η θετική προγνωστική τιμή (PPV) του ταξινομητή είναι πολύ χειρότερες από το ισορροπημένο σύνολο δεδομένων. Για το λόγο αυτό χρησιμοποιήθηκε η δειγματοληψία προς τα κάτω, η οποία είχε ως αποτέλεσμα το DEEPLY ESSENTIAL να επιτύχει καλύτερο AUC, ευαισθησία και PPV σε σύγκριση με μεθόδους που χρησιμοποίησαν καθολική δειγματοληψία.

3.4.3 Τι σημαίνει «διαρροή δεδομένων» και πως επηρεάζει την πρόβλεψη της γονιδιακής σπουδαιότητας

Στα βακτηριακά είδη, ένα σημαντικό μέρος των γονιδίων τους διατηρείται αφού αφενός εκτελούν παρόμοιες θεμελιώδεις βιολογικές λειτουργίες και αφετέρου είναι αρκετά παρόμοια σε επίπεδο αλληλουχίας. Όπως έχει αναφερθεί και παραπάνω, αυτά τα γονίδια έχουν επισημανθεί ως απαραίτητα ή μη. Ας θεωρηθούν x και y δύο ομόλογα γονίδια, δηλαδή γονίδια που έχουν πολύ παρόμοια πρωτογενή αλληλουχία DNA. Αν το x χρησιμοποιείται στην διαδικασία της εκπαίδευσης και το y χρησιμοποιείται για δοκιμή, το παραπάνω εισάγει ενός είδους μεροληψίας ή διαφορετικά μια "διαρροή δεδομένων".

Προκειμένου να ποσοτικοποιηθεί η επίδραση του ζητήματος διαρροής δεδομένων, το σύνολο όλων των γονιδίων των τριάντα βακτηριακών ειδών ομαδοποιήθηκε χρησιμοποιώντας το OrthoMCL. Το τελευταίο είναι μια δημοφιλής μέθοδος για τη ομαδοποίηση ορθολογικών, ομόλογων και παραλογικών πρωτεϊνών που χρησιμοποιούν αφενός αμοιβαία ευθυγράμμιση βέλτιστου χτυπήματος για την ανίχνευση πιθανού ζεύγους ενδοπαραλόγου/πρόσφατου παραλόγου και αφετέρου αμοιβαίες ευθυγραμμίσεις βέλτιστου χτυπήματος μεταξύ δύο γονιδιωμάτων για τον εντοπισμό πιθανών ορθολογικών ζευγών. Στη συνέχεια, δημιουργήθηκε ένα γράφημα ομοιότητας με βάση τις πρωτεΐνες που συνδέονται μεταξύ τους. Για τον διαχωρισμό μεγάλων συστάδων, χρησιμοποιήθηκε ο αλγόριθμος ομαδοποίησης Markov (MCL). Μέσα στις συστάδες MCL, τα βάρη μεταξύ κάθε ζεύγους πρωτεϊνών κανονικοποιούνται, ώστε να διορθωθούν οι εξελικτικές διαφορές.

Για να γίνει αντιληπτή λοιπόν η επίδραση της ομοιότητας αλληλουχίας γονιδίων στην απόδοση πρόβλεψης, δημιουργήθηκε ένα σύνολο δεδομένων, στο οποίο κανένα γονίδιο από μία μόνο συστάδα δεν μπορεί να εκχωρηθεί τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο δοκιμών. Η πρόβλεψη επαναλήφθηκε δέκα φορές και πρέκυψαν τα παρακάτω αποτελέσματα. Η τιμή AUC μειώθηκε περισσότερο από 7%, ενώ η ακρίβεια μειώθηκε κατά 6.9%. Βέβαια, ενώ οι τιμές AUC παρέμειναν σταθερές σε όλα τα πειράματα, η ευαισθησία, η ειδικότητα και το PPV διέφεραν σε μεγάλο βαθμό μεταξύ των πειραμάτων που αφορούσαν στο ομαδοποιημένο σύνολο δεδομένων.

Σε ορισμένες άλλες δημοσιευμένες μελέτες το ζήτημα της διαρροής δεδομένων αντιμετωπίστηκε με διάφορες τεχνικές όπως είναι η απόκλιση Kullback-Leibler που μετράει την απόσταση μεταξύ της κατανομής K-MER που προέκυψε από ακολουθίες και το CD-HIT που αφαιρεί τους πλεονασμούς στα δεδομένα εκπαίδευσης και βελτιώνει την ικανότητα γενίκευσης του μοντέλου.

3.4.4 Η Σπουδαιότητα των Χαρακτηριστικών

Η αρχιτεκτονική DNN, σε σχέση με άλλα μοντέλα, δεν παρέχει εύκολα εικόνα σχετικά με το σύνολο των χαρακτηριστικών που συνέβαλαν στο μέγιστο βαθμό όσον αφορά στην απόδοση πρόβλεψης. Προκειμένου λοιπόν να κατανοήσουν οι ερευνητές το αντίκτυπο ενός χαρακτηριστικού, πραγματοποίησαν μια μελέτη κατάλυσης, η οποία ουσιαστικά αφαιρεί ένα ή περισσότερα χαρακτηριστικά -που δεν συσχετίζονται σε

μεγάλο βαθμό - από την είσοδο και καθορίζει τη διαφορά απόδοσης. Σε αυτή την περίπτωση όμως, η αφαίρεση ενός χαρακτηριστικού αντισταθμίζεται από το εξαιρετικά συσχετιζόμενο χαρακτηριστικό του. Για να αντιμετωπιστεί το παραπάνω λοιπόν υπολογίστηκε η Pearson συσχέτιση ανά ζεύγη όλων των χαρακτηριστικών εισόδου. Στον χάρτη θερμότητας που δημιουργήθηκε λοιπόν υπήρξαν δεκαεννέα ζεύγη χαρακτηριστικών που έδειξαν υψηλότερη συσχέτιση από 0.9 (σε απόλυτη τιμή).

Δοκιμάστηκαν ακόμα οι αλλαγές απόδοσης στο σύνολο δεδομένων GP+GN χρησιμοποιώντας πενταπλάσια διασταυρούμενη επικύρωση. Πιο συγκεκριμένα, μετρήθηκε η διαφορά στις τιμές AUC και λήφθηκαν υπόψιν τα χαρακτηριστικά με βάση τον αντίκτυπό τους στη μείωση της προγνωστικής απόδοσης. Κατέληξαν στο συμπέρασμα πως το μήκος του γονιδίου και της πρωτεΐνης είναι εξαιρετικά κατατοπιστικά χαρακτηριστικά για την πρόβλεψη σπουδαιότητας ενός γονιδίου. Επιπλέον, είναι ευρέως γνωστό ότι σε βασικά γονίδια εντός της λειτουργικής κατηγορίας που σχετίζεται με την αποθήκευση και τη διεργασία πληροφοριών, τα κωδικοποιημένα αμινοξέα K, L και οι υποκατηγορίες κωδικοποιημένων αμινοξέων C, G, E και F ταιριάζουν κατά προτίμηση στον κύριο κλώνο όπου αυτά είναι υπεύθυνα για την παραγωγή και τη μετατροπή ενέργειας, τη μεταφορά υδατανθράκων και άλλες βασικές μεταβολικές διεργασίες.

3.5 Συμπεράσματα

Έχει χρησιμοποιηθεί ένας μεγάλος αριθμός δομικών και λειτουργικών χαρακτηριστικών για την πρόβλεψη της γονιδιακής σπουδαιότητας. Ωστόσο, τα παραπάνω χαρακτηριστικά δεν μπορούν να ληφθούν από τις γονιδιακές ακολουθίες και συχνά δεν είναι διαθέσιμα για πολλά βακτηριακά είδη. Για να μεγιστοποιηθεί λοιπόν η πρακτική χρησιμότητα του DEEPLY ESSENTIAL χρησιμοποιούνται αποκλειστικά χαρακτηριστικά που προέρχονται απευθείας από τις παραπάνω ακολουθίες.

Για να αποδειχθεί η καλύτερη απόδοση του μοντέλου συγκρίθηκε με άλλες δύο προσεγγίσεις. Τα πειράματα απέδειξαν πως το DEEPLY ESSENTIAL έχει καλύτερη προγνωστική απόδοση τόσο σε σύνολα δεδομένων με κάτω-δειγματοληψία (down-sampling) όσο και σε ομαδοποιημένα σύνολα δεδομένων. Όσον αφορά στην πρώτη κατηγορία, το μοντέλο έδειξε βελτίωση 12,8% σε τιμές AUC ενώ παρήγαγε σημαντικά καλύτερη ευαισθησία και ακρίβεια σε σύγκριση με τις υπόλοιπες προσεγγίσεις, επιτυγχάνοντας 6.2% βελτιωμένη ευαισθησία και 137.4% βελτιωμένη ακρίβεια. Το DEEPLY ESSENTIAL πέτυχε επίσης καλύτερη απόδοση σε ομαδοποιημένα σύνολα δεδομένων, αφού έδωσε 7,9% και 29,2% βελτιωμένη AUC συγκρίνοντας το με άλλες δύο προσεγγίσεις.

Ως εναλλακτική προσέγγιση θα μπορούσε κάποιος να εξετάσει την εκπαίδευση ενός συνελικτικού νευρωνικού δικτύου (CNN) με χρήση κωδικοποίησης "one-hot" του DNA και της αλληλουχίας της πρωτεΐνης ως είσοδο. Ωστόσο, το περιορισμένο μέγεθος των διαθέσιμων δεδομένων εκπαίδευσης θα ήταν ανεπαρκές, για να επιτρέψει στο παραπάνω δίκτυο να εξάγει σχετικά χαρακτηριστικά. Αυτό σημαίνει πως οι ταξινομητές που βασίζονται στο CNN δεν είναι τόσο ακριβείς σε σύγκριση με την αρχιτεκτονική του DEEPLY ESSENTIAL.

Συμπερασματικά, το ερευνητικό κείμενο προτείνει ένα νευρωνικό δίκτυο "βαθιάς" αρχιτεκτονικής για την πρόβλεψη της σπουδαιότητας των γονιδίων στα μικρόβια που έχει ως απώτερο στόχο την ανακάλυψη φαρμάκων και την βελτιστοποίηση των πειραμάτων συνθετικής βιολογίας σε μικρόβια. Ως είσοδο δέχεται αλληλουχίες γονιδίου, μεγιστοποιώντας την πρακτική του εφαρμογή, σε σύγκριση με άλλες προσεγγίσεις που απαιτούν δομικά ή τοπολογικά χαρακτηριστικά που μπορεί να μην είναι άμεσα διαθέσιμα. Εκτεταμένα πειράματα έδειξαν πως το DEEPLY ESSENTIAL έχει καλύτερη προγνωστική απόδοση από τα υπάρχοντα εργαλεία πρόβλεψης. Ωστόσο, θα μπορούσε να βελτιωθεί ακόμη παραπάνω, εάν ήταν διαθέσιμα περισσότερα βακτηριακά δεδομένα.

4 Τρίτο Ερευνητικό Κείμενο

Στο συγκεκριμένο κεφάλαιο θα αναλυθεί η ερευνητική εργασία με τίτλο **MethylNet: an automated and modular deep learning approach for DNA methylation analysis** και συγγραφείς τους Joshua J. Levy, Alexander J. Titus, Curtis L. Petersen, Youdinghuan Chen, Lucas A. Salas and Christensen Brock C.

4.1 Στόχος και Ερευνητικό Ενδιαφέρον

Η μεθυλίωση αποτελεί μία απλή, ζωτικής όμως σημασίας, βιοχημική διαδικασία του σώματος, η οποία βοηθάει στη ρύθμιση της δραστηριότητας του καρδιοαγγειακού, νευρολογικού, αναπαραγωγικού και απο-

τοξινωτικού συστήματος. Κατά τη διάρκεια αυτής της διεργασίας ομάδες μεθυλίου προστίθενται στο μόριο του DNA. Η μεθυλίωση έχει την ικανότητα να αλλάξει τη δραστηριότητα ενός τμήματος DNA χωρίς να αλλάξει την αλληλουχία του. Συνήθως διεξάγεται μεταξύ των αλληλουχιών κυτοσίνης και γουανίνης και όταν βρίσκεται σε έναν προαγωγέα γονιδίου, δρα για να καταστείλει τη γονιδιακή μεταγραφή. Οι μη μεθυλιωμένες περιοχές του DNA, συνδέονται με ανοιχτές καταστάσεις χρωματίνης και επιτρεπτές στη γονιδιακή μεταγραφή. Τέλος, σημειώνεται ότι μία δυσλειτουργική - απορυθμισμένη διαδικασία μεθυλίωσης μπορεί να έχει ως αποτέλεσμα ασθένειες όπως ο καρκίνος, ενώ επηρεάζεται άμεσα από παράγοντες όπως η ηλικία και το κάπνισμα. Συνεπώς, καθίσταται σαφές, ότι η παρατήρηση της διεργασίας της μεθυλίωσης ενός οργανισμού θα βοηθήσει σημαντικά στην έγκαιρη διάγνωση σοβαρών προβλημάτων.

Οι τεχνολογίες που χρησιμοποιούνται ως τώρα στον τομέα των αναλύσεων της βαθιάς μάθησης σχετικά με τη μεθυλίωση δεν έχουν καταφέρει να δημιουργήσουν ένα πλαίσιο φιλικό προς το χρήστη για μοντέλα εκτέλεσης, εκπαίδευσης και ερμηνείας. Ωστόσο, μέσω του άρθρου παρουσιάζεται μία DNAm μεθοδολογία βαθιάς μάθησης, η οποία μπορεί να κατασκευάσει ενσωματώσεις, να κάνει προβλέψεις, να δημιουργήσει νέα δεδομένα και να αποκαλύψει άγνωστη ετερογένεια με ελάχιστη επίβλεψη από τη πλευρά του χρήστη. Πιο συγκεκριμένα, αξιοποιούνται εργασίες βαθιάς μάθησης λανθάνουσας παλινδρόμησης και ταξινόμησης χώρου μέσω της ανάπτυξης ενός αρθρωτού πλαισίου που είναι εξαιρετικά προσιτό στους επιγενετικούς ερευνητές. Το **MethylNet**, όπως αποκαλείται, φιλοδοξεί να προσφέρει εργασίες δημιουργίας και ομαδοποίησης χωρίς επίβλεψη, αποσυνέλιξη τύπου κυττάρου, ταξινόμηση υπόπτου παν-καρκίνου, παλινδρόμηση ηλικίας και ταξινόμηση κατάστασης καπνίσματος.

4.2 Αποτελέσματα

Από την πειραματική διαδικασία προέκυψε ότι ο αλγόριθμος του **MethylNet** κρίνεται ιδιαιτέρως αποτελεσματικός για την κωδικοποίηση δεδομένων DNAm (δεδομένα μεθυλίωσης μορίου DNA), εντοπίζοντας τα λανθάνοντα χαρακτηριστικά που έχουν υψηλή πιστότητα [1] στο αρχικό σύνολο δεδομένων. Λαμβάνοντας υπόψιν τις πολύ καλές επιδόσεις του **MethylNet** στη διαδικασία της κατηγοριοποίησης, η ομάδα αποφάσισε να δοκιμάσει την εγκυρότητά του και ως προβλεπτικό μοντέλο, για εκτίμηση ηλικίας, κυτταρική αναλογία και κατηγοριοποίηση ασθενειών, για DNAm δεδομένα.

4.2.1 Εργαλεία και Μεθοδολογία

Για την εκμάθηση του νευρωνικού δικτύου χρησιμοποιήθηκαν έξι δημόσια DNAm datasets, ενώ ακόμη κατασκευάστηκαν use cases, με σκοπό την επίδειξη της ικανότητας επιτυχούς αποτύπωσης των χαρακτηριστικών, που σχετίζονται με την ηλικία, τα κύτταρα, τη γενεαλογία, τις καταστάσεις των ασθενειών και τις εκθέσεις σε ασθένειες. Τα σύνολα δεδομένων χρησιμοποιήθηκαν για την κατηγοριοποίηση των παραπάνω χαρακτηριστικών από τον αλγόριθμο, ενώ πιο συγκεκριμένα όσον αφορά στις ασθένειες πραγματοποιήθηκε εκτενής μελέτη για διάφορους τύπους καρκίνου, ενώ ακόμη συγκρίθηκαν δείγματα από καπνιστές και μη για τη μελέτη της ρευματοειδούς αρθρίτιδας.

Τα σύνολα δεδομένων αναλύθηκαν και στη συνέχεια χωρίστηκαν και χρησιμοποιήθηκαν σε τρεις κατηγορίες: εκπαίδευση νευρωνικού δικτύου, δοκιμές και έλεγχο εγκυρότητας.

4.2.2 Ηλικιακά Αποτελέσματα

Τα αποτελέσματα για την εκτίμηση της ηλικίας έδειξαν μία αρκετά ακριβή εκτίμηση της ηλικίας, χωρίς βέβαια να παρέχουν αρκετές πληροφορίες για το πως γίνεται η αντιστοίχηση των χαρακτηριστικών που οδηγούν στο συμπέρασμα, αλλά ούτε και για τα ίδια τα χαρακτηριστικά.

Τελευταία, παρατηρείται ενδιαφέρον για την περαιτέρω εξερεύνηση της διαφοράς μεταξύ της χρονολογικής ηλικίας και της ηλικίας μεθυλίωσης. Οι όροι αυτοί έχουν οριστεί στην επιστήμη της βιολογίας και με βάση αυτούς τους ορισμούς η ομάδα φιλοδοξεί ότι ο αλγόριθμος τους θα μπορέσει να χρησιμοποιηθεί για την εξερεύνηση του συγκεκριμένου τομέα.

4.2.3 Αποτελέσματα Αποσυνέλιξης Κυτταρικού Τύπου

Η διαδικασία της κυτταρικής αποσυνέλιξης σχετίζεται με τις υπολογιστικές τεχνικές που χρησιμοποιούνται στην προσπάθεια να εκτιμηθούν οι ποσότητες των διαφορετικών τύπων κυττάρων σε ένα

[1] Αναφερόμενοι στον όρο "υψηλή πιστότητα", εννοούμε την όσο το δυνατό καλύτερη και ακριβέστερη διαδικασία μεταφοράς δεδομένων από τον κόμβο X στον κόμβο Y, δηλαδή κατά τη μεταφορά αυτή τα δεδομένα να μείνουν αμετάβλητα

δείγμα. ενώ υπάρχουν αρκετά αξιόπιστα συστήματα σήμερα για τη συγκεκριμένη διεργασία, η ομάδα αποφάσισε να δοκιμάσει τις δυνατότητες του **MethylNet** στον τομέα αυτό, με την ελπίδα ότι θα μπορέσει να λειτουργήσει ως αρωγή στο μέλλον σε μη επιβλεπόμενες διαδικασίες.

Κατά τις πρώτες δοκιμές το **MethylNet** χρησιμοποιώντας τα σύνολα δεδομένων που χρησιμοποιήθηκαν και για την ηλικιακή έρευνα, έπρεπε να αναγνωρίσει έξι διαφορετικούς τύπους ανοσοποιητικών κυττάρων. Η χωρίς επίβλεψη παραγωγή έξι λανθάνοντων συστάδων χρησιμοποιώντας ενσωματώσεις VAE (τεχνική βαθιάς μάθησης για την αναγνώριση λανθανουσών αναπαραστάσεων) επέδειξε διαχωρισμό των κυτταρικών αναλογιών, χωρίς, πρώτα, να έχει εκπαιδευτεί σε ένα σύνολο αναφοράς κυτταρικών αναλογιών για προφίλ DNAm. Οι SHAP τιμές (οι τιμές πρόβλεψης, το πως δηλαδή εκτιμούμε ένα σύστημα, όπως για παράδειγμα ένα νευρωνικό δίκτυο, πρόκειται να συμπεριφερθεί) των κυτταρικών τύπων, ομαδοποιήθηκαν ιεραρχικά και το αποτέλεσμα δείχνει ότι οι τιμές που εκτιμήθηκαν συμπίπτουν με αυτές που έχουν γνωστοποιηθεί από την κυτταρική γενεολογία, ενώ ακόμα και τύποι που εμφανίζουν βελτιωμένες μετρήσεις συμφωνίας σε σχέση με άλλους τύπους συνεχίζουν να έχουν παρόμοια απόλυτα σφάλματα.

4.2.4 Αποτελέσματα Πρόβλεψης Όλων των Τύπων Καρκίνου

Το **MethylNet** συμμετείχε σε παν-καρκινική κλινική έρευνα που διεξήχθη και επέδειξε εκπληκτικά αποτελέσματα, καθώς ήταν σε θέση να αναγνωρίσει τριάντα δύο διαφορετικούς τύπους καρκίνου με πολλή μεγάλη ακρίβεια, ξεπερνώντας κατά πολύ άλλα πολύ αξιόπιστα συστήματα. Ακόμη, και στα λανθάνοντα προφίλ παν-καρκινικών υποτύπων χάρη στην εκπαίδευσή του κατόρθωσε να παράγει ομαδοποιήσεις σε υψηλή συμφωνία με διάφορους τύπους καρκίνου, καταφέρνοντας να δημιουργήσει με τους υποτύπους ομάδες που συμφωνούν με τα χαρακτηριστικά που έχουν τεθεί για τον καρκίνο από την επιστήμη της βιολογίας. Με αυτόν τον τρόπο, το σύστημα αποδεικνύει ότι όχι μόνο είναι σε θέση να κάνει ακριβείς προβλεπτικές κατηγοριοποιήσεις, αλλά και ότι είναι σε θέση να ξεχωρίσει τα λανθάνοντα χαρακτηριστικά με βάση τη βιολογία και τους γνωστούς καρκινικούς τύπους.

Βέβαια σημειώνεται ότι υπήρχαν και αρκετές περιπτώσεις λανθασμένης κατηγοριοποίησης για τους υποτύπους. Πιο συγκεκριμένα, ενώ, οι υπερομάδες επέδειξαν θετικά αποτελέσματα, όπως προαναφέρθηκε, παρατηρείται ότι οι υπό-ομάδες μέσα στις υπερκλάσεις παρουσιάζουν λάθη, κυρίως μεταξύ συγκεκριμένων ζευγαριών πλησίον περιοχών, συγχέοντας μεταξύ τους παρόμοιους τύπους καρκίνων.

4.2.5 Εφαρμογή EWAS, Προκαταρκτική Υποτυποποίηση και Εξωτερική Επικύρωση

EWAS - Epigenome-Wide Association Study: πρόκειται για μία εξέταση ενός συνόλου μετρήσιμων επιγενετικών σημάτων σε όλο το γονιδίωμα, όπως η μεθυλίωση του DNA, σε διαφορετικά άτομα για να προσδιοριστούν συσχετίσεις μεταξύ των επιγενετικών ποικιλιών και ένας συγκεκριμένος αναγνωρίσιμος φαινότυπος/χαρακτηριστικό.

Εξαιτίας της επιτυχίας του **MethylNet** να ομαδοποιήσει μη γραμμικά στοιχεία και όλα όσα αναφέρθηκαν νωρίτερα αποφασίστηκε να δοκιμαστεί και σε δεδομένα με σκοπό την ανάλυση του καπνίσματος, με δείγματα καπνιστών αλλά και ανθρώπων που δεν έχουν καπνίσει ποτέ. Παρά τη βραχυχρόνια εκπαίδευση του στο συγκεκριμένο κομμάτι κατάφερε να αποδώσει αρκετά ικανοποιητικά αποτελέσματα πρόβλεψης κατάστασης καπνίσματος. Στη συνέχεια, δοκιμάστηκε και σε ένα μεγαλύτερο σύνολο δεδομένων και απέδειξε ότι τα αποτελέσματα που είναι σε θέση να εξάγει έχουν υψηλή συσχέτιση με αυτά των κλασσικών EWAS τεχνολογιών [8], παρά τις διαφορές του σε σχέση με αυτές.

4.2.6 Τι είναι το MethylNet

Το **MethylNet** αποτελεί μία δομή βαθιάς μάθησης, η οποία χρησιμοποιεί το interface μίας αντικειμενοστρεφούς εφαρμογής και περιέχει μία ενσωματωμένη λειτουργικότητα για την εύκολη εναλλαγή μεταξύ αναλύσεων σε σχέση με εργασίες ενσωμάτωσης, δημιουργίας, ταξινόμησης και παλινδρόμησης. Μεγίστης σημασίας είναι η προσπάθεια για την ανάδειξη της ικανότητάς του να καταγράφει χαρακτηριστικά που αποσπάστηκαν από τα αρχικά DNAm δεδομένα ενώ ταυτόχρονα δημιουργήθηκαν ακριβείς προβλέψεις που συμφωνούν με τις αναμενόμενες που προκύπτουν από τη βιολογία. Σε αντίθεση με προηγούμενες εφαρμογές και προσπάθειες το **MethylNet** προσφέρει την τελειοποίηση του συστήματος εξαγωγής των παραπάνω χαρακτηριστικών, ενώ προσθέτει επιπλέον στρώματα ανάλυσης στις εργασίες πρόβλεψης. Ακόμη, το σύστημα του επιτρέπει το μοντέλο που χρησιμοποιεί για την ανάλυση των δεδομένων να γενικοποιηθεί και για δεδομένα πάνω στα οποία δεν έχει εκπαιδευτεί. Τα παραπάνω χαρακτηριστικά του σημαίνουν την αρχή μιας εποχής όπου είναι δυνατή η καλύτερη κατανόηση της βιολογίας μέσω της βαθιάς μάθησης.

4.2.7 Δυνατά σημεία, περιορισμοί και μελλοντικές κατευθύνσεις

Ανάμεσα σε όλα το **MethylNet** έρχεται αντιμέτωπο με την ανάλυση πολυδιάστατων δεδομένων, κάτι που ακόμη αποτελεί ζήτημα σημαντικής δυσκολίας, καθώς κατά το στάδιο της προσαρμογής των συντελεστών των προβλέψεων, τα αποτελέσματα τείνουν να μην είναι ευκόλως ερμηνεύσιμα. Παρόλα αυτά, ένα από τα πιο δυνατά του σημεία είναι όπως αναφέρθηκαν παραπάνω οι παν-καρκινικές του προβλέψεις, όπου σε σχέση με άλλες μεθόδους, οι οποίες μένουν απλά στην εξερεύνηση μερικών δεσμών θυμίνης-γουανίνης, το **MethylNet** παρέχει μία πιο ολοκληρωμένη εικόνα κι εφόσον είναι σε θέση να καταλάβει και να αναλύσει τη βιολογία ανάμεσα σε παρόμοιες περιπτώσεις παρέχει πιθανότατα τη δυνατότητα εξερεύνησης ασθενειών και λύσεων σε ασθένειες, είτε αυτές έχουν να κάνουν με τον καρκίνο είτε όχι.

Σημαντικό είναι να σημειωθεί, επίσης, ότι το **MethylNet** σε αντίθεση με άλλες μεθόδους είναι σε θέση να αναλύσει πολύ μεγαλύτερο όγκο δεδομένων , οπότε σε επόμενη φάση κρίνεται απαραίτητο, να αντιμετωπιστούν τα θέματα που προαναφέρθηκαν τα οποία προκαλούν σύγχυση στην ανάγνωση των αποτελεσμάτων μέσω της επιλογής των χαρακτηριστικών, ενώ ακόμη να δοθεί μεγαλύτερο βάρος στις βιολογικές ερμηνείες και στις πληροφοριακές μεθόδους που εφαρμόζονται για την ανάλυση των δεσμών θυμίνης και γουανίνης.

Τέλος, για την ανάπτυξη του κομματιού που αφορά στη βαθιά μάθηση, δεν αποκλείεται το ενδεχόμενο να χρησιμοποιηθεί Common Workflow Language (CWL), ενώ ακόμη πιθανή είναι η χρήση ενός Μπεϋζιανού δικτύου για την αυτοματοποίηση της κατασκευής της αρχιτεκτονικής ενός ιδανικού νευρωνικού δικτύου.

4.2.8 Μέθοδοι

Το **MethylNet** έχει σχεδιαστεί για να λειτουργεί μέσω μερικών απλών εντολών (σε περιβάλλον UNIX/Linux), όλες εκ των οποίων μπορούν να εφαρμοστούν σε οποιοδήποτε προβλεπτικό μοντέλο. Πώς έχουν σχεδιαστεί όμως αυτά τα μοντέλα; Πρώτο στάδιο αποτελεί η προ-εκπαίδευση των μοντέλων βαθιάς μάθησης χρησιμοποιώντας πληθώρα από αυτόματους κωδικοποιητές, ενώ τα επίπεδα των κωδικοποιητών αυτών είναι που χρησιμοποιούνται για την εξαγωγή των βιολογικά σημαντικών χαρακτηριστικών. Τα στρώματα των νευρωνικών δικτύων χρησιμοποιούνται για την ενσωμάτωση των δεδομένων και την εξαγωγή χαρακτηριστικών που θα χρησιμεύσουν στην ομαδοποίηση ενός συνόλου κατά τη διάρκεια μίας μη-επιβλεπόμενης υλοποίησης, δημιουργώντας κατά αυτόν τον τρόπο νέα δεδομένα που παρουσιάζουν υψηλή πιστότητα σε σχέση με αυτά της αρχικής πηγής. Σε δεύτερο στάδιο, τα στρώματα πρόβλεψης συμπεριλαμβάνονται στον κωδικοποιητή, ρυθμίζοντας λεπτομερώς την πρόβλεψη του μοντέλου και παρουσιάζοντας την εξόρυξη χαρακτηριστικών κατά τις διαδικασίες της παλινδρόμησης και της ταξινόμησης πολλαπλών εξόδων. Η διεργασία της προ-εκπαίδευσης βελτιστοποιούν το νευρωνικό δίκτυο στο κομμάτι των προβλέψεων. Τρίτον, η πραγματοποίηση αυτόνομων σαρώσεων υπερπαραμέτρων βοηθάει στη βελτιστοποίηση των παραμέτρων του μοντέλου για την πρώτη και τη δεύτερη διεργασία, ενώ παράλληλα δημιουργεί πλούσιες οπτικοποιήσεις των δεδομένων. Τέλος, οι συνδυασμοί θυμίνης-γουανίνης σε κάθε πρόβλεψη προσδιορίζονται μέσω μεθόδων απόδοσης χαρακτηριστικών Shapley.

5 Τέταρτο Ερευνητικό Κείμενο

Στο συγκεκριμένο κεφάλαιο θα αναλυθεί η ερευνητική εργασία με τίτλο **DeepGRN: prediction of transcription factor binding site across cell-types using attention-based deep neural networks** και συγγραφείς τους Chen Chen, Jie Hou, Xiaowen Shi, Hua Yang, James Birchler and Jianlin Cheng.

5.1 Στόχος και Ερευνητικό Εδνιαφέρον

Οι μεταγραφικοί παράγοντες αποτελούν πρωτεΐνες που προσδένονται σε συγκεκριμένες γονιδιωματικές αλληλουχίες και επηρεάζουν πολυάριθμες κυτταρικές διεργασίες. Ρυθμίζουν τους ρυθμούς των μεταγραφικών δραστηριοτήτων των μεταγενέστερων γονιδίων μέσω τέτοιων γεγονότων πρόσδεσης, ώστε να δρουν ως ενεργοποιητές ή καταστολείς στα γονιδιακά ρυθμιστικά δίκτυα ελέγχοντας το επίπεδο έκφρασης και την αφθονία πρωτεϊνών των γονιδίων-στόχων τους. Η ChIPSeq (ακολουθία χρωματίνης) αποτελεί το μοντέλο πρότυπο για τον προσδιορισμό των αλληλεπιδράσεων ενός μεταγραφικού παράγοντα και όλων των πιθανών περιοχών πρόσδεσης σε γονιδιωματικές αλληλουχίες. Ωστόσο, τα πειράματα που απαιτούνται

για τη μελέτη της συγκεκριμένης ακολουθίας χρειάζονται αντιδραστήρια και υλικά, τα οποία είναι ανέφικτο να αποκτηθούν. Μία λύση στο συγκεκριμένο ζήτημα αποτελεί η πρόβλεψη των πιθανών περιοχών πρόσδεσης μέσω υπολογιστικών μεθόδων και μέχρι σήμερα έχει αναπτυχθεί μία πληθώρα αλγορίθμων για την επίτευξη αυτού ακριβώς του σκοπού, οι οποίοι όμως βασίζονται σε προηγούμενη γνώση για την παροχή των αποτελεσμάτων και ίσως να ήταν λιγότερο αξιόπιστα στο ενδεχόμενο που χρησιμοποιούνται σε "περιοχές", όπου προηγούμενη γνώση δεν είναι διαθέσιμη.

Στο συγκεκριμένο ζήτημα η λύση φαίνεται να παρέχεται από τα νευρωνικά δίκτυα, τα οποία μπορούν να διαχειριστούν πολύ καλύτερα δεδομένα μεγάλου όγκου και παρέχουν καλύτερα αποτελέσματα όταν τους παρέχεται λίγη ή και καθόλου προηγούμενη γνώση (στη συγκεκριμένη περίπτωση γνώση για πιθανές περιοχές πρόσδεσης). Έρευνες έχουν καταλήξει στο συμπέρασμα ότι για τον προσδιορισμό των πιθανών περιοχών πρόσδεσης χρησιμεύει ο συνδυασμός των συνελικτικών και επαναλαμβανόμενων νευρωνικών δικτύων, όπου το συνελικτικό στρώμα εξάγει τοπικά χαρακτηριστικά από γονιδιωματικά σήματα και περιοχές, ενώ το επαναλαμβανόμενο αξιοποιεί τις χρήσιμες πληροφορίες σε ολόκληρη την αλληλουχία δεδομένων. Πρόσφατα, ο μηχανισμός της προσοχής έχει σημειώσει σημαντική επιτυχία στη νευρωνική μηχανικής μετάφρασης και στη συναισθηματική ανάλυση. Παρά τις επιτυχίες των μηχανισμών αυτών, παρατηρείται ένα μειονέκτημα, το οποίο είναι η μεγάλη δυσκολία της ερμηνείας των βαρών ενός νευρωνικού, εξαιτίας του πλεονασμού τους και της μη γραμμικής τους σχέσης με την έξοδο.

Στόχος της έρευνας του συγκεκριμένου συγγράμματος είναι η ανάπτυξη ενός μοντέλου πρόβλεψης περιοχών πρόσδεσης των μεταγραφικών παραγόντων το οποίο στηρίζεται στη βαθιά μάθηση με μηχανισμό προσοχής, ενώ αποδεικνύεται ότι η αξιοποίηση πληροφοριακών μοτίβων, τόσο DNase-seq όσο και στις αλληλουχίες DNA είναι σημαντική για την παροχή μίας ακριβούς πρόβλεψης.

5.2 Διαδικασία

Για τους μεταφραστικούς παράγοντες και τους κυτταρικούς τύπους που παρέχονται στα σύνολα δεδομένων πρόσδεσης, η ετικέτα της δέσμευσης κατάστασης των μεταγραφικών παραγόντων παράγεται από πειράματα CHIP-sEQ και χρησιμοποιείται ως βασική αλήθεια. Για την εκπαίδευση του μοντέλου ως χαρακτηριστικά εισόδου χρησιμοποιούνται πληροφορίες προσβασιμότητας της χρωματίνης (δεδομένα DNA-Seq) και δεδομένα RNA-Seq, ενώ ακόμα ακολουθούνται οι κανόνες και οι περιορισμοί της πρόκλησης DREAM [2]. Αρχικά, τα μοντέλα εκπαιδεύονται σε όλα τα χρωμοσώματα εκτός από τα 1, 8 και 21 και το χρωμόσωμα 11 χρησιμοποιείται ως επικύρωση Στη συνέχεια, το μοντέλο με την καλύτερη απόδοση στα δεδομένα επικύρωσης χρησιμοποιείται για την τελική πρόβλεψη εάν δεν υπάρχει ήδη κάποιο "leaderboard" δοσμένο από την πρόκληση (DREAM). Στην πορεία, τα δεδομένα του "leaderboard" είναι διαθέσιμα για συγκριτική αξιολόγηση με ορισμένους μεταφραστικούς παράγοντες και κάθε συμμετέχων μπορεί να δοκιμάσει την απόδοση με αυτούς τους μεταφραστικούς παράγοντες πραγματοποιώντας έως και δέκα υποβολές. Κατά αυτόν τον τρόπο, επιλέγονται τα δέκα καλύτερα μοντέλα ως προαιρετικό βήμα επιλογής μοντέλου.

Στόχο για γην πρόβλεψη αποτελούν τα δεδομένα πρόσδεσης μεταγραφικών παραγόντων των CHIP-Seq πειραμάτων. Για την επεξεργασία των δεδομένων αυτών το γονιδίωμα χωρίζεται σε bins [3] των 200 bp με μέγεθος βήματος ολίσθησης 50 bp, καθένα από τα οποία ανήκει σε μία κατηγορία: δεσμευμένο, μη-δεσμευμένο ή διφορούμενο, με βάση τα αποτελέσματα των πειραμάτων. Κατά τη διαδικασία της εκπαίδευσης ή της επικύρωσης δε χρησιμοποιούνται τα διφορούμενα bins.

Για την ανάλυση της πρωτογενούς αλληλουχίας DNA χρησιμοποιήθηκε ένα γονιδίωμα ως αναφορά και όπως συνηθίζεται στους αλγορίθμους εξαγωγής χαρακτηριστικών από προφίλ χρωματίνης επεκτείνεται κάθε bin κατά 400 bp προς τα πάνω αλλά και προς τα κάτω, παράγοντας έτσι μία περιοχή εισόδου 1000 bp. Μέσω της αξιολόγησης, της απόδοσης διαφορετικών περιοχών εισόδων αποδείχθηκε ότι ένα εύρος πάνω από 600 bp είναι επαρκές για την απόκτηση σταθερής πρόβλεψης. Η αλληλουχία αυτής της περιοχής αναπαρίσταται σε έναν πίνακα, όπου κάθε γραμμή αντιστοιχεί σε ένα νουκλεοτίδιο. Σημειώνεται ότι εξαιτίας του γεγονότος ότι τα χαμηλά επίπεδα αντιστοίχησης αλληλουχιών μπορεί να εισάγουν μεροληψία σε πειράματα παράλληλης αντιστοίχησης, η μοναδικότητα της αλληλουχίας συνδέεται στενά με την ποιότητα των δεδομένων αλληλουχίας και για το λόγο αυτό, επιλέγεται ως πρόσθετο χαρακτηριστικό το σκορ μοναδικότητας. Στη συγκεκριμένη περίπτωση παρατηρούνται βαθμολογίες δύο τιμών (0 και 1), όπου

[2] διαγωνισμοί (προκλήσεις) που αναζητούν απάντηση σε βασικά ερωτήματα που αφορούν τη συστημική βιολογία και τη μεταφραστική ιατρική μέσω ανεπτυγμένων υπολογιστικών μεθόδων

[3] ομαδοποίηση δεδομένων για την καλύτερη προετοιμασία των δεδομένων μέσω της χρήσης τους στη διαδικασία της βαθιάς μάθησης

το 1 χαρακτηρίζει τις αλληλουχίες που θεωρούνται μοναδικές, ενώ το 0 αντιστοιχεί σε αλληλουχίες που εμφανίζονται τουλάχιστον τέσσερις φορές. Το ENCODE Project Consortium παρείχε μία μαύρη λίστα γονιδιωματικών περιοχών. Επομένως, τα bins που βρίσκονται στις συγκεκριμένες περιοχές αποκλείονται ως είσοδοι δεδομένων εκπαίδευσης και το σκορ τους τίθεται αυτομάτως σε μηδέν, εάν βρίσκονται σε περιοχές-στόχους της πρόβλεψης.

Η ανάλυση των δεδομένων DNase-Seq χρησιμοποιείται για τη λήψη χαρτών κατά μήκος όλου του γονιδιώματος με σκοπό την εύρεση πληροφοριών για την προσβασιμότητα της χρωματίνης, καθώς η προσβασιμότητα της χρωματίνης συνδέεται άμεσα με την προσβασιμότητα των περιοχών ενός χρωμοσώματος και με τα γεγονότα πρόσδεσης μεταγραφικών παραγόντων.

Το χαρακτηριστικό που αντιστοιχεί στο σχολιασμό κάθε bin κωδικοποιείται ως δυαδικό διάνυσμα μήκους έξι, με κάθε τιμή να περιγράφει αν υπάρχει επικάλυψη μεταξύ του δυαδικού πεδίου εισόδου και κάθε ενός από τα έξι γονιδιωματικά χαρακτηριστικά. Στη συνέχεια, η Principal Component Analysis εκτελείται σε κανονικοποιημένες μετρήσεις μεταγραφών ανά εκατομμύριο από δεδομένα RNA-Seq που παρέχονται από το διαγωνισμό και οι οκτώ πρώτες κύριες συνιστώσες ενός κυτταρικού τύπου χρησιμοποιούνται ως βαθμολογίες έκφρασης για όλες τις εισόδους από το συγκεκριμένο κυτταρικό τύπο, δημιουργώντας ένα διάνυσμα μήκους οκτώ.

Τα κομμάτια διατήρησης του γονιδιώματος χρησιμοποιούνται ως χαρακτηριστικό γνώρισμα για πρόσθετα μοντέλα. Τα συγκεκριμένα συνοδεύονται από μία βαθμολογία PhastCons, η οποία με τη σειρά της αναπαριστά βαθμολογίες διατήρησης ανά βάση, δηλαδή αναπαριστά την πιθανότητα τα διατηρημένα στοιχεία να βρίσκονται σε συντηρημένες περιοχές του γονιδιώματος.

5.2.1 Κατασκευή Νευρωνικού

Το σχήμα κάθε διαδοχικής εισόδου καθορίζεται από ένα τύπο με βάση το μήκος L κάθε περιοχής. Οι διαδοχικές είσοδοι παράγονται τόσο για το εμπρόσθιο μέρος όσο και για την αντίστροφη συμπληρωματική αλυσίδα. Τα βάρη όλων των στρωμάτων του μοντέλου μοιράζονται τις δύο εισόδους ώστε να σχηματιστεί μια σιαμαία αρχιτεκτονική. Τα διανύσματα των μη διαδοχικών χαρακτηριστικών δεδομένων γονιδιακής έκφρασης και γονιδιωματικού σχολιασμού συγχωνεύονται στο μοντέλο και πιο συγκεκριμένα στο πρώτο πυκνό στρώμα. Το μοντέλο χωρίζεται σε δύο κύριες ενότητες, τη μεμονωμένη προσοχή και την προσοχή ανά ζεύγη, οι οποίες χρησιμοποιούν την ίδια είσοδο και αρχιτεκτονική πέρα από τον εσωτερικό μηχανισμό προσοχής, ενώ το τελικό αποτέλεσμα είναι ο μέσος όρος της εξόδου των δύο μονάδων.

5.3 Αποτελέσματα

Μελετώντας τα αποτελέσματα καθίσταται σαφές ότι το μοντέλο σημείωσε με μεγάλη διαφορά καλύτερη επίδοση σε σχέση με τα υπάρχοντα λογισμικά. Πιο συγκεκριμένα, το μοντέλο παρουσιάζει καλύτερη απόδοση για κάθε συνδυασμό μεταγραφικού παράγοντα και κυτταρικού τύπου παρέχοντας παράλληλα ένα πιο ευρύ φάσμα αποτελεσμάτων. Παράλληλα, παρουσιάζει το καλύτερο σκορ σε επτά από τους δεκατρείς στόχους και τον καλύτερο γενικό μέσο όρο επίδοσης, ενώ η αξιοπιστία των αποτελεσμάτων αξιολογήθηκε επίσης, παρατηρώντας τις επιδόσεις του μοντέλου στους στόχους του διαγωνισμού, όπου παρουσίασε εξίσου ικανοποιητικές επιδόσεις.

5.4 Συμπεράσματα

Το μοντέλο που παρουσιάζεται ενσωματώνει το μηχανισμό προσοχής με την αρχιτεκτονική που βασίζεται στα CNNs-RNNs, το οποίο τέθηκε σε ελέγχους και απέδειξε ότι είναι ικανό να ανταγωνιστεί τις τέσσερις κορυφαίες μεθόδους στον πίνακα κατάταξης του διαγωνισμού DREAM, ενώ ακόμη, παρουσιάζει ότι οι μονάδες προσοχής βοηθούν στην ερμηνεία του τρόπου αναγνώρισης κρίσιμων μοτίβων από διαφορετικούς τύπους χαρακτηριστικών εισόδου.

Βιβλιογραφία

[1] et al Backenroth D. *FUN_LDA: a latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications,.* Am J Hum Genet, 2018.

[2] et al Bernstein BE. *The NIH roadmap epigenomics mapping consortium,.* Nat Biotechnol, 2010.

[3] et al Boyle AP. *High-resolution mapping and characterization of open chromatin across the genome,*. Cell, 2008.

[4] et al Dunham I. *An integrated encyclopedia of DNA elements in the human genome,*. Nature, 2012.

[5] et al He Z. *A semi-supervised approach for predicting cell-type specific functional consequences of non-coding variation using MPRAs,*. Nat Commun, 2018.

[6] Kim CY Hwang S. *HumanNet v2: human gene networks for disease research.* Nucleic Acids Res, 2019.

[7] et al Ionita-Laza I. *A spectral approach integrating functional genomic annotations for coding and noncoding variants,*. Nat Genet., 2016.

[8] Marioni RE Joehanes R, Just AC. *Epigenetic signatures of cigarette smoking. Circ Cardiovasc Genet.* Circ Cardiovasc Genet, 2016.

[9] et al Kircher M. *A general framework for estimating the relative pathogenicity of human genetic variants,*. Nat Genet., 2014.

[10] et al Lee D. *A method to predict the impact of regulatory variants from DNA sequence,*. Nat Genet., 2015.

[11] Zhang F-Z Lin Y. *Identifying bacterial essential genes based on a feature-integrated method.* IEEE/ACM Trans Comput Biol Bioinform, 2017.

[12] et al Lu Q. *Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies,*. PLoS Genet, 2016.

[13] Gao F Luo H, Lin Y. *DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements,*. Nucleic Acids Res, 2014.

[14] Koonin EV Mushegian AR. *A minimal gene set for cellular life derived by comparison of complete bacterial genomes,*. Proc Natl Acad Sci USA, 1996.

[15] Sobetzko P Nigatu D. *Sequence-based information-theoretic features for gene essentiality prediction.* BMC Bioinformatics, 2017.

[16] Lin H Ning LW. *Predicting bacterial essential genes using only sequence composition information.* Genet Mol Res, 2014.

[17] Mukherjee S Palaniappan K. *Predicting "essential" genes across microbial genomes: A machine learning approach.* In: 2011 10th International Conference on Machine Learning and Applications and Workshops, vol. 2. ieeexplore.ieee.org, 2011.

[18] et al Quang D. *DANN: a deep learning approach for annotating the pathogenicity of genetic variants,*. Bioinformatics, 2015.

[19] Tong T Song K. *Predicting essential genes in prokaryotic genomes using a linear method: ZUPLS.* Integr Biol., 2014.

[20] Morris JH Szklarczyk D. *The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible.* Nucleic Acids Res, 2017.

[21] Troyanskaya OG Zhou J. *Predicting effects of noncoding variants with deep learning-based sequence model,*. Nat Methods, 2015.