Πολυτεχνική Σχολή

Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Μάθημα: Εισαγωγή στην Βιοπληροφορική

Εφαρμογές Τεχνητών Νευρωνικών Δικτύων στην Βιοπληροφορική

Applications of Artificial Neural Networks in Bioinformatics

Αγγελόπουλος Γιώργος Α.Μ. 1067435

Κρεμανταλά Θεοδώρα Α.Μ. 1067445

Πάτρα, Ιούνιος 2023

Περιεχόμενα

1.1 *Μία κριτική ανασκόπηση για την εφαρμογή του Τεχνητού Νευρωνικού Δικτύου στην Βιοπληροφορική*

Η βιοπληροφορική είναι μια τεχνική που χρησιμοποιεί υπολογιστές για την κατασκευή μαθηματικού μοντέλου χρησιμοποιώντας διαφορετικά εργαλεία βιοπληροφορικής για την απόκτηση, διαχείριση και οπτικοποίηση βιολογικών δεδομένων. Για να συμπεράνει τη σχέση μεταξύ των στοιχείων ενός πολύπλοκου βιολογικού συστήματος, αναπτύσσει εργαλεία και μεθόδους λογισμικού για την κατανόηση και ανάλυση κλινικών δεδομένων και βιολογικών δεδομένων. Αυτό απαιτεί τη συλλογή και την ερμηνεία μεγάλων τμημάτων βιολογικών δεδομένων, έτσι ώστε να ληφθούν πληροφορίες σχετικά με ορισμένες βιολογικές διεργασίες.

Διαφορετικοί Τομείς Εφαρμογής της Βιοπληροφορικής:

Στον πραγματικό κόσμο, η βιοπληροφορική βρίσκει χρήση στους εξής τομείς:

- Τομέας υγείας
- Ανάπτυξη Φαρμάκων
- Μοριακή Ιατρική
- Προληπτική Ιατρική
- Γονιδιακή θεραπεία
- Καθαρισμός απορριμμάτων
- Μελέτη της Κλιματικής αλλαγής
- Βιοτεχνολογία
- Βελτίωση της καλλιέργειας
- Αντοχή στα έντομα
- Ανάπτυξη ποικιλιών ανθεκτικών στην ξηρασία
- Συγκριτικές σπουδές

Υπάρχει ένα ευρύ φάσμα εφαρμογών της βιοπληροφορικής στον τομέα της διάγνωσης, της ιατρικής, της γεωργίας, της βιοτεχνολογίας. Η μελέτη και η χρήση διαφορετικών εργαλείων βιοπληροφορικής θα επιτρέψει στους ερευνητές να επεκτείνουν τη γνώση πολύ πιο αποτελεσματικά και αποτελεσματικά μέσω ανάλυσης δεδομένων και πειραμάτων. Αυτό θα επιβεβαιώσει με μεγαλύτερη ακρίβεια τις σημαντικές ανακαλύψεις.

Προβλήματα βιοπληροφορικής:

Οι εφαρμογές βιοπληροφορικής έρχονται με πολλές προκλήσεις όταν σχετίζονται με ορισμένα ζητήματα που οφείλονται στα δεδομένα ή στις συσκευές που χρησιμοποιούνται για τη συλλογή ή την ανάλυσή τους. Επομένως, η αντιμετώπιση και η ανάλυση αυτών των θεμάτων απαιτείται για σωστή εκτέλεση και αποτελεσματικά συμπεράσματα.

Θέματα που σχετίζονται με τη δομή

Η μελέτη του DNA και της πρωτεΐνης περιλαμβάνει προβλήματα όπως η πρόβλεψη της δομής των πρωτεϊνών καθώς αναπαριστώνται σε τρισδιάστατα δεδομένα, επομένως η πρόβλεψη δομής, η ευθυγράμμιση και η ανάλυση γίνονται δύσκολη υπόθεση. Η πρόβλεψη της τρισδιάστατης δομής της πρωτεΐνης από την αλληλουχία μπορεί να λυθεί με την εφαρμογή ANN.

Ανάλυση Αλληλουχίας

Η ταξινόμηση του RNA, της Πρωτεϊνικής Αλληλουχίας και του DNA γίνεται πρόκληση λόγω της διαφοράς και της ομοιότητας πολλών οργανισμών. Το γονιδίωμα υποδηλώνει το πλήρες σύνολο των χρωμοσωμάτων ενός οργανισμού που αποτελείται από DNA. Η αλληλουχία γονιδιώματος είναι ένας τρόπος χαρτογράφησης DNA ή παραγγελίας DNA για οργάνωση, επεξεργασία και ερμηνεία των αλληλουχιών, κάτι που απαιτεί και πάλι βελτιώσεις στις στρατηγικές αλληλούχισης. Κάθε αλληλουχία DNA αντιμετωπίζει προκλήσεις στην αναζήτηση του σχεδίου αλληλουχίας, στο σχεδιασμό, την ανάλυση και την ερμηνεία των δεδομένων.

Η ανάλυση της αλληλουχίας ή της αλληλουχίας DNA είναι μια σημαντική εργασία, επειδή βοηθά στον εντοπισμό μεμονωμένων γονιδίων που σχετίζονται με μια ασθένεια. Οι παραδοσιακές μέθοδοι ανίχνευσης γονιδίων βασίστηκαν στη μέθοδο δοκιμής και σφάλματος, όμως πλέον πολλοί αλγόριθμοι μηχανικής μάθησης χρησιμοποιούνται για την ταξινόμηση των φυσιολογικών και των μη φυσιολογικών γονιδίων με μεγάλη ακρίβεια.

Βιολογικά σύνολα δεδομένων:

Η βιοπληροφορική ασχολείται με διάφορα βιολογικά σύνολα δεδομένων που συλλέγονται σε διαφορετικά επίπεδα ωμικών δεδομένων. Με βάση τον τύπο των δεδομένων, η βιολογική βάση δεδομένων μπορεί να χωριστεί σε δύο κατηγορίες:

Πρωτογενής Βάση Δεδομένων: Αυτού του είδους οι βάσεις δεδομένων έχουν αρχειακό χαρακτήρα, επειδή αυτές οι βάσεις δεδομένων δημιουργούνται από τα πειραματικά αποτελέσματα που υποβάλλονται απευθείας από τους ερευνητές. Αυτές οι βάσεις δεδομένων είναι γεμάτες με αλληλουχία πρωτεϊνών, αλληλουχία νουκλεοτιδίων ή μακρομοριακή δομή κ.λπ.

Δευτερεύουσα βάση δεδομένων: Αυτές οι βάσεις δεδομένων είτε δημιουργούνται χειροκίνητα είτε εξάγονται από την ανάλυση αποτελεσμάτων της κύριας βάσης δεδομένων για τη δημιουργία πιο δομημένων εγγραφών για εύκολη ανάκτηση δεδομένων.

Δόμηση Υπολογιστικού Μοντέλου:

Σε αυτή την ενότητα θα συζητήσουμε ορισμένες προϋποθέσεις που απαιτούνται για την κατασκευή του υπολογιστικού μοντέλου.

Προεπεξεργασία δεδομένων και η αναγκαιότητά της:

Προκειμένου να βελτιωθεί το αποτέλεσμα της ταξινόμησης, ξεκινά ένα βήμα προεπεξεργασίας ως βασικό βήμα πριν από την εξόρυξη των δεδομένων. Η τεχνική προεπεξεργασίας δεδομένων βελτιώνει σημαντικά την ποιότητα των δεδομένων, την απόδοση του μοντέλου ταξινόμησης και ελαχιστοποιεί τον χρόνο που απαιτείται για την πραγματική εξόρυξη.

Θα αντιμετωπίσουμε ορισμένα από τα προβλήματα που πρέπει να επιλυθούν για να επιτευχθεί καλύτερο αποτέλεσμα ταξινόμησης. Περιλαμβάνει τον καθαρισμό θορυβωδών δεδομένων, δεδομένων που λείπουν, διπλότυπων δεδομένων κ.λπ. από τη βάση δεδομένων για την ομαλή διεξαγωγή της διαδικασίας ταξινόμησης. Ένα άλλο μεγαλύτερο πρόβλημα στα βιολογικά δεδομένα είναι η απουσία τιμών. Σε πολύπλοκα βιολογικά σύνολα δεδομένων, αυτό το ζήτημα επηρεάζει σε μεγάλο βαθμό την απόδοση της ακρίβειας του μοντέλου. Τέλος, η επικάλυψη δεδομένων είναι συνεχές πρόβλημα ποιότητας δεδομένων που μαρτυρείται σε διάφορους τομείς, συμπεριλαμβανομένης της υγειονομικής περίθαλψης, των επιχειρήσεων και της μοριακής βιολογίας, κ.λπ. Αυτό το ζήτημα μπορεί να αντιμετωπιστεί με τον εντοπισμό και την εξάλειψη διπλότυπων τιμών.

Αυτά τα δεδομένα συχνά περιέχουν μεγάλο όγκο άσχετων δεδομένων που επηρεάζουν την ακρίβεια ταξινόμησης και την αποτελεσματικότητα της μηχανικής μάθησης . Η τεχνική μείωσης διαστάσεων εστιάζει στη μείωση του αριθμού των χαρακτηριστικών εισόδου που βοηθά στη μείωση του χρόνου υπολογισμού και των περιττών δεδομένων.

Ταξινόμηση βιολογικών δεδομένων:

Η ταξινόμηση είναι μια διαδικασία με την οποία τα δεδομένα οργανώνονται σε διαφορετικές κατηγορίες με τον καθορισμό μιας κλάσης για ένα στοιχείο στη βάση δεδομένων. Τα δεδομένα ομαδοποιούνται σε διαφορετικές κατηγορίες με βάση το σύνολο δεδομένων εκπαίδευσης. Η διαθεσιμότητα μεγάλου όγκου δεδομένων μικροσυστοιχιών έχει δημιουργήσει νέα πεδία στις μεθόδους ταξινόμησης. Όπως και η ταξινόμηση των μικροσυστοιχιών DNA συμβάλλει σημαντικά στη διάγνωση και την πρόγνωση σε πολλές κλινικές πρακτικές. Επίσης, η ταξινόμηση των δεδομένων γονιδιακής έκφρασης αντιμετωπίζει το θεμελιώδες πρόβλημα πολλών ασθενειών.

ML στη Βιοπληροφορική:

Η μηχανική μάθηση (ML) είναι μια τεχνική για την ανάπτυξη προγράμματος υπολογιστή για πρόσβαση σε δεδομένα και για αυτόματη εκμάθηση γνώσης από την εμπειρία χωρίς ανθρώπινη παρέμβαση και βοήθεια. . Η τεχνική μηχανικής

μάθησης χρησιμοποιεί δύο διαφορετικές μεθόδους για την εκπαίδευση του μοντέλου: την εποπτευόμενη μάθηση και τη μέθοδο μάθησης χωρίς επίβλεψη.

Υπάρχουν πολλές τεχνικές μηχανικής μάθησης μεταξύ των οποίων το Τεχνητό Νευρωνικό Δίκτυο είναι μια αποτελεσματική τεχνική για την αναγνώριση, επιλογή, ταξινόμηση και πρόβλεψη του γονιδίου στις Αλληλουχίες DNA.

Εισαγωγή στο ANN:

Το ANN είναι ένα υπολογιστικό σύστημα που αποτελείται από ένα εξαιρετικά διασυνδεδεμένο δίκτυο μονάδων επεξεργασίας που ονομάζονται νευρώνες. Έχει τη δυνατότητα να χειρίζεται πολύπλοκα χαρακτηριστικά μέσα σε δεδομένα προκειμένου να επεξεργαστεί τις πληροφορίες.

Μία απλή ροή εργασίας ενός τεχνητού νευρωνικού δικτύου είναι με αρχιτεκτονική perceptron, το οποίο αποτελείται από ένα στρώμα εισόδου με λίγες μονάδες εισόδου που αντιπροσωπεύουν πολλαπλά χαρακτηριστικά που υπάρχουν στη βάση δεδομένων, ανάλογα με τον στόχο που ορίζεται. Κάθε είσοδος που συλλέγεται από το σύνολο δεδομένων πολλαπλασιάζεται με βάρη και τροφοδοτείται σε μια συνάρτηση που ονομάζεται συνάρτηση ενεργοποίησης για να παραχθεί η πραγματική έξοδος.

Ανάλογα με τη διαφορά μεταξύ της επιθυμητής εξόδου και της πραγματικής εξόδου τροποποιούνται τα βάρη σε κάθε επίπεδο σύνδεσης και τελικά προβλέπεται η έξοδος. Τα βάρη είναι τιμές που μαθαίνονται από μηχανή από νευρωνικά δίκτυα. Το νευρωνικό δίκτυο μαθαίνει μέσω μιας διαδικασίας ανάδρασης που ονομάζεται οπίσθια διάδοση, της οποίας ο αλγόριθμος βοηθά στη μείωση της συνολικής υπολογιστικής απώλειας του δικτύου κατά την εκμάθηση του δικτύου.

Τα πιο συχνά χρησιμοποιούμενα ANN γενικά ακολουθούν την αρχιτεκτονική τριών επιπέδων που έχουν ένα στρώμα εισόδου, ένα ή περισσότερα κρυφά επίπεδα και ένα επίπεδο εξόδου.

Υπάρχουν 2 μοντέλα ANN που χρησιμοποιούνται ευρέως για εποπτευόμενη μάθηση:

Δίκτυο Perceptron: Το μοντέλο Perceptron είναι ένας δυαδικός ταξινομητής που σημαίνει ότι διαχωρίζει τα δεδομένα εισόδου σε δύο κατηγορίες. Είναι το απλούστερο και παλαιότερο μοντέλο νευρωνικών δικτύων που μπορεί να εφαρμόσει γραμμικά διαχωρίσιμα προβλήματα όπως AND, OR, NOT gate αλλά δεν λειτουργεί για μη γραμμικά προβλήματα όπως η πύλη XOR.

Multilayer Perceptron (MLP) Network: Το MLP υποστηρίζει ταξινόμηση πολλαπλών κλάσεων. Αποτελείται από τουλάχιστον ένα κρυφό στρώμα μαζί με ένα στρώμα εισόδου και ένα στρώμα εξόδου. Σε αυτό το δίκτυο κάθε μεμονωμένος κόμβος συνδέεται με όλους τους άλλους κόμβους στο επόμενο επίπεδο συνδέοντας βάρη για την ανάπτυξη ενός πλήρως συνδεδεμένου νευρωνικού δικτύου. Το MLP

εφαρμόζει μη γραμμική λειτουργία ενεργοποίησης για την πρόβλεψη των μονάδων εξόδου.

Δεν θα ασχοληθούμε με την μη εποπτευόμενη μάθηση σε αυτό κεφάλαιο.

Έπειτα στο κείμενο ακολουθούν αναφορές σε εφαρμογές των ANN στην βιοπληροφορική, διάφορες τεχνικές που χρησιμοποιούνται για εποπτευόμενες τεχνικές μάθησης και μία εκτενή αναφορά σε βιβλιογραφικές αναφορές για συγκεκριμένα παραδείγματα χρήσης τους ανά τα χρόνια από μελετητές.

Στην συνέχεια γίνεται παράθεση ενός πίνακα που αναγράφει ανα στήλη σε ποιό άρθρο από τα προηγούμενα αναφέρεται η τρέχουσα γραμμή, ποιο σετ δεδομένων και μοντέλο χρησιμοποιήθηκε, ποιος ήταν ο σκοπός και ποια ήταν τα ευρήματα.

Κριτική Ανάλυση:

Σε αυτή τη μελέτη παρατηρήσαμε ότι ο ταξινομητής ANN ξεπέρασε όλους τους άλλους ταξινομητές με αποτέλεσμα λογικής ακρίβειας. Το ANN είναι το πιο ισχυρό εργαλείο ταξινόμησης και πρόβλεψης. Η απόδοση του αλγορίθμου ANN εξαρτάται από διάφορους παράγοντες όπως η προεπεξεργασία δεδομένων, η λειτουργία ενεργοποίησης που θα χρησιμοποιηθεί, η επιλογή του αριθμού εποχών και νευρώνων.

Στην περίπτωση του μοντέλου SVM φαίνεται ότι όταν ο αριθμός των χαρακτηριστικών υπερβαίνει τον αριθμό των δειγμάτων, το μοντέλο τείνει να αποδίδει αργά. Όμως, όταν υπάρχει σημαντική ποσότητα δεδομένων αλληλουχίας DNA για ταξινόμηση ασθενειών δύο τάξεων, μπορούμε να πούμε ότι το SVM είναι ένα εξαιρετικό μοντέλο ταξινόμησης. Παρατηρήσαμε επίσης ότι όταν οι είσοδοι είναι θορυβώδεις ή ελλιπείς, τα νευρωνικά δίκτυα εξακολουθούν να είναι σε θέση να παράγουν λογικά αποτελέσματα. Έτσι, η σωστή χρήση της τεχνικής προεπεξεργασίας δεδομένων θα μπορούσε να βελτιώσει την απόδοση του μοντέλου ταξινόμησης.

Ένας άλλος παράγοντας που επηρεάζει την απόδοση του μοντέλου είναι η σωστή επιλογή της συνάρτησης ενεργοποίησης για την ταξινόμηση γραμμικών και μη γραμμικών δεδομένων. Μία από τις ταχύτερες λειτουργίες ενεργοποίησης εκμάθησης είναι η λειτουργία ReLU που δίνει πιο ακριβή αποτελέσματα επειδή είναι εύκολο να βελτιστοποιηθεί με βαθμιδωτή κάθοδο και να οδηγήσει σε συνολική βέλτιστη λύση.

Παρατηρείται ότι όταν ο αριθμός των κρυφών επιπέδων αυξάνεται, το μοντέλο δίνει σχετικά υψηλή ακρίβεια. Όμως πρέπει να προσέξουμε στον αριθμό των κρυφών επιπέδων ανάλογα την εφαρμογή γιατί υπερβολικά πολλά επίπεδα υπονομεύουν συχνά την απόδοση σε ταχύτητα της ταξινόμησης αλλά πολύ λίγα την ποιότητα των αποτελεσμάτων. Η ενσωμάτωση του αλγορίθμου ANN με διαφορετικούς αλγόριθμους βελτιστοποίησης ελαχιστοποιεί το ποσοστό

σφάλματος που παράγεται από το μοντέλο ταξινόμησης, το οποίο ως αποτέλεσμα βελτιώνει την απόδοση του μοντέλου.

1.2 Paper 2: DNN (βαθέων νευρωνικών δικτύων) πρόβλεψη υπογραφών μεταγραφής σε όλο το γονιδίωμα – πέρα από το Μαύρο κουτί

Η έκφραση του μεταγραφόμενου mRNA αποτελεί ένα από τα πιο σημαντικά κομμάτια του ρυθμιστικού μηχανισμού των κυττάρων αλλά και των λειτουργιών των ιστών και των οργάνων. Πιο συγκεκριμένα, η ανάλυση της έκφρασης του MRNA μπορεί να συντελέσει τόσο στην μελέτη των ασθενειών όσο και στην διατήρηση της ταυτότητας των κυττάρων.

Επιπρόσθετα, έχουν δημιουργηθεί μεγάλες ποσότητες δεδομένων έκφρασης RNA από ανθρώπους τα οποία έχουν ομαδοποιηθεί με σκοπό την πιο εύκολη υπόθεση γονιδίων που εμπλέκονται σε ασθένειες. Οι παράγοντες της μεταγραφής, τα λεγόμενα TFs, είναι κρίσιμοι για τον ρυθμιστικό έλεγχο των γονίδιων και μεγάλος όγκος εργαλείων της Βιοπληροφορικής στοχεύουν στην πρόβλεψη των τοποθεσιών δέσμευσης των TF.

Μια άλλη συμπληρωματική στρατηγική για να πετύχουμε τον παραπάνω σκοπό είναι η επιστήμη του δικτύου η οποία αναφέρεται στην μελέτη των αλληλεπιδράσεων μεταξύ οντοτήτων, γονοτύπων και φαινοτύπων. Συγκεκριμένα, τα γονίδια που αλληλεπιδρούν με απορυθμισμένα γονίδια χωρίς να εκφράζονται διαφοροποιημένα τα ίδια συχνά παραβλέπονται σε μελέτες διαφορικής έκφρασης. Κατά συνέπεια, αυτά τα δίκτυα είναι δύσκολο να εξαχθούν από δεδομένα και έτσι η στρατηγική αποδεικνύεται χρήσιμη αλλά όχι ικανοποιητική.

Για αυτό τον λόγο λοιπόν, χάρις στην πρόσφατη πρόοδο της μηχανικής μάθησης υπάρχει το ερώτημα αν τέτοιες μέθοδοι μπορούν να διευκολύνουν την ανάλυση βιολογικών δικτύων. Οι εφαρμογές των βαθέων νευρωνικών δικτύων (DNN) στην γονιδιωματική σχετίζονται με την πρόβλεψη των θέσεων δέσμευσης TF και τα αποτελέσματα των μη-κωδικοποιημένων γενετικών παραλλαγών. Τα DNN περιλαμβάνουν την ικανότητα αποτύπωσης μη γραμμικών σχέσεων και επιπλέον απαιτούν σημαντικές ποσότητες δεδομένων. Τα βαθιά νευρωνικά δίκτυα έχουν επίσης εφαρμοστεί και για την κατανόηση της ρύθμισης της έκφρασης του mRNA. Αν και ήταν το πρώτο σημαντικό βήμα προς την πρόβλεψη των επιπέδων mRNA, οι ρυθμιστικοί παράγοντες μεταγραφής (TFs) δεν διαχωρίστηκαν από το υπόλοιπο μεταγραφικό στοιχείο, μετατρέποντας την βιολογική ερμηνεία και μετάφραση των ασθενειών δύσκολη.

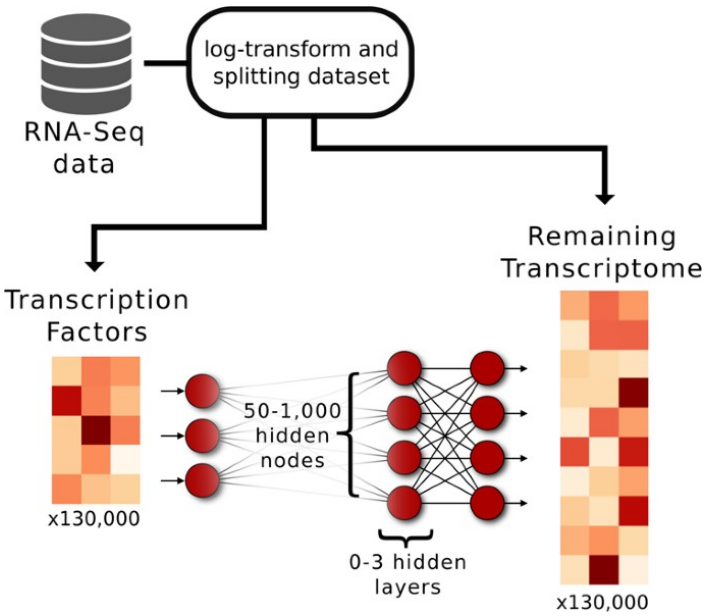
Στο παρόν κείμενο, αναπτύσσεται μια μεθοδολογία που υπερβαίνει την παραγωγή λιστών διαφορικά εκφραζόμενων γονιδίων αλλά δεν φτάνει στην δυσεπίλυτη ανακατασκευή ενός πλήρους ρυθμιστικού δικτύου γονιδίων.

Αντιθέτως, η μεθοδολογία στοχεύει στη ρύθμιση που ασκείται από τους παράγοντες μεταγραφής TFs και εξετάζεται εάν η εκπαίδευση ενός DNN σε δεδομένα γονιδιακής έκφρασης θα μπορούσε να προβλέψει ένα προγνωστικό δίκτυο Τ. Πρόκειται για μια μεθοδολογία πέρα από τα μοντέλα μηχανικής μάθησης μαύρου κουτιού, το οποίο με τη σειρά του είναι ένα πρώτο βήμα προς αυτό που θα μπορούσε να αναφερθεί ως πλήρως ερμηνεύσιμα μοντέλα λευκού κουτιού.

Διαπιστώνουμε ότι τέτοια μοντέλα μπορούν πράγματι να προβλέψουν την έκφραση γονιδίων που βασίζονται σε TF και ότι οι προβλεπόμενες σχέσεις μεταξύ των TF και των γονιδίων-στόχων τους επικαλύπτονται σε μεγάλο βαθμό με γνωστές δεσμεύσεις TF.

Πιο συγκεκριμένα, για την ακριβή και ισχυρή πρόβλεψη του επιπέδου έκφρασης των γονιδίων-στόχων με δίκτυα βαθιών νευρικών παραγόντων μεταγραφής χρησιμοποιήθηκαν οι TF ως είσοδοι και οι μη TF ως γονίδια – στόχοι εξόδου. Έγινε η εκπαίδευση των μοντέλων με χρήση της βάσης δεδομένων ARCHS4 και η απόδοση τους αξιολογήθηκε με τον ειδικό συντελεστή προσδιορισμού ($1\text{-}R^2$) στα δεδομένα δοκιμής.

Παρατηρήθηκε ότι τα DNN (βαθιά νευρωνικά δίκτυα είχαν καλύτερη απόδοση από τα ρηχά (shallow) μοντέλα και επίσης διαπιστώθηκε μια αύξηση στην ικανότητα πρόβλεψης της γονιδιακής έκφρασης σε σύγκριση με το 80% της επεξηγημένης διακύμανσης κατά την πρόβλεψη της αφθονίας του mRNA αποκλειστικά από την αλληλουχία DNA.



Στην συνέχεια, περιγράφεται στο κείμενο ότι έγινε εφαρμογή των εκπαιδευόμενων μοντέλων για πρόβλεψη έκφρασης γονιδίων σε κυτταρικές σειρές ανθρώπινου

όγκου. Μετά τα αποτελέσματα της παραπάνω μελέτης, καταλήξαμε στο συμπέρασμα ότι τα DNN μπορούσαν να προβλέψουν πιστά την πλειονότητα του ανθρώπινου μεταγραφώματος με δεδομένα τα επίπεδα έκφρασης, τόσο σε υγιείς όσο και σε προσβεβλημένες από ασθένειες καταστάσεις.

Παρακάτω στο κείμενο αναφέρεται ότι τεχνική φωτισμού κόμβου (node light-up technique) αποκάλυψε τον εμπλουτισμό επικυρωμένων συσχετίσεων στόχων TF μέσα στα δίκτυα πρόβλεψης. Αναλυτικότερα, γεννιέται το ερώτημα αν τα εκπαιδευμένα δίκτυα είναι ερμηνεύσιμα, δηλαδή εάν οι συσχετισμοί στόχων TF που έχουν μάθει είναι βιολογικά σχετικοί ή όχι. Είναι γνωστό ότι σε ένα DNN είναι ενσωματωμένες μη γραμμικές εξαρτήσεις. Με στόχο την προσέγγιση αυτών των εξαρτήσεων έγινε η υπόθεση ότι στο προβλεπόμενο δίκτυο οι διαταραχές έκφρασης των TFs διαδίδονται πιο αποτελεσματικά στα σχετικά γονίδια-στόχους. Μια τέτοια δηλαδή προσέγγιση καταγράφει τις αποτελεσματικές γονιδιακές εξαρτήσεις, γραμμικές ή μη και για την ανάλυση χρησιμοποιήθηκε η τεχνική φωτισμού κόμβου.

Επιπρόσθετα, χρησιμοποίησαν 4 πηγές αλληλεπιδράσεων TF-στόχου και λήφθηκαν δεδομένα από τις 3 βάσεις δεδομένων DoRothEA,TRRUST και RegNetWork. Ακολούθως έγιναν συγκρίσεις μεταξύ των βάσεων δεδομένων όσον αφορά τις κατατάξεις εμπλουτισμού. Τα αποτελέσματα που προέκυψαν ήταν ότι η υψηλότερη κατάταξη εμπλουτισμού λήφθηκε από το TRUUST ενώ η DoRothEA έδωσε τη δεύτερη υψηλότερη και το RegNetwork τη χαμηλότερη.

Δεδομένου ότι τα TFs μπορούν να δρουν τόσο ως αναστολείς όσο και ως εκκινητές της μεταγραφής, ήλθε στην επιφάνεια το ερώτημα για τον βαθμό στον οποίο οι αναλύσεις φωτισμού DNN κατέλαβαν επίσης την κατευθυντικότητα και το σημάδι των αλληλεπιδράσεων TF-στόχου. Γενικά, διαπιστώθηκε ότι ο αριθμός των κρυφών επιπέδων ή μονάδων για τα DNN έχει περιορισμένο αντίκτυπο στην απόδοση.

Όλα τα μοντέλα DNN έδειξαν συγκρίσιμη απόδοση σε αξιολογικές αναλύσεις, προβλέψεις έκφρασης πειραμάτων από τη βάση δεδομένων ARCHS4, προβλέψεις έκφρασης καρκίνου και σύγκριση φωτισμού με βάσεις δεδομένων TF-στόχου. Ας σημειωθεί ότι το ρηχό NN δεν έφτασε ποτέ σε ικανοποιητική απόδοση. Έτσι, το μεγαλύτερο κέρδος σε επεξηγηματική ισχύ και επικάλυψη με υπάρχουσες βάσεις δεδομένων ήρθε από την προσθήκη τουλάχιστον ενός ενδιάμεσου επιπέδου, επιτρέποντας έτσι μη γραμμικούς μετασχηματισμούς. Τελικά, οι μη γραμμικοί μετασχηματισμοί δεν αποτελέσαν εμπόδιο για την επιθεώρηση του προβλεπόμενου δικτύου και την εξαγωγή επικυρωμένων βιολογικών γνώσεων χρησιμοποιώντας την τεχνική κόμβου φωτισμού.

Συνεχίζοντας, στο κείμενο αναφέρεται η αλγοριθμική εξαγωγή του βασικού συνόλου επικυρωμένων ρυθμιστών TFs από DNN. Ειδικότερα, πραγματοποιήθηκε σταδιακή αφαίρεση TFs από το επίπεδο εισόδου με βάση την επεξηγηματική τους

ισχύ. Παρατηρήθηκε ότι επεξηγηματική ισχύς μειώνεται σταδιακά κατά την αφαίρεση των προγνωστικών, εδώ TF (predictors). Ακόμα, διαπιστώθηκε ότι το μοντέλο ανακάλυψε κυρίως γνωστές αλληλεπιδράσεις TF-στόχου ανεξάρτητα από το μέγεθος εισόδου του μοντέλου. Έτσι, παρά το γεγονός ότι είχε άφθονες μεταβλητές εισόδου, το DNN ανακάλυψε γνωστούς πυρήνες TF που σχετίζονται με κεντρικά και καλά σχολιασμένα μονοπάτια.

Ο λανθάνοντας χώρος DNN δείχνει εμπλουτισμό λειτουργικά σχετιζόμενων και σχετιζόμενων με νόσο γονιδίων. Αναλυτικότερα, εκτός από την αξιολόγηση της βιολογικής συνάφειας των συγκεκριμένων αλληλεπιδράσεων που ανακαλύφθηκαν από το προγνωστικό DNN, μπορεί κανείς να αναρωτηθεί εάν ο προγνωστικός παράγοντας θα μπορούσε να είναι ενημερωτικός σε ένα πλαίσιο ασθένειας. Πέρα από την ανάλυση των modules, μονάδων που μπορούν να χρησιμοποιηθούν στην μελέτη ασθενειών, περιγράφεται η χρήση αλληλεπιδράσεων γονιδίου στόχου TF εντός των δύο 250 μεταβλητών που μετρούν τα ενδιάμεσα λανθάνοντα στρώματα.

Μέσα από την ανάλυση, διαπιστώθηκε ότι γονίδια νόσου φάνηκε να συνυπάρχουν σε φωτισμούς κρυφού κόμβου στο DNN. Το αποτέλεσμα αυτό αποδεικνύει ότι τα υπόλοιπα γονίδια που βρέθηκαν σε τέτοιες ενότητες ασθένειας θα μπορούσαν να είναι σχετικά με την ανάλυση και την ερμηνεία βιοδεικτών και μηχανισμών που σχετίζονται με ασθένειες.

Η ανάλυση με DNN δίνει πληροφορίες για μηχανισμούς ανθρώπινων ασθενειών που εμπλέκονται σε γονιδιακή ρύθμιση. Ειδικότερα, στο σημείο αυτό του κειμένου αναφέρονται οι αναλύσεις και τα πειράματα που έγιναν ώστε να διαπιστωθεί ότι οι μηχανισμοί ασθένειας απορρύθμισης από TFs θα μπορούσαν να διαδοθούν πιστά στο επίπεδο στόχο.

Τέθηκε το ερώτημα αν τα TF έπρεπε να εκφραστούν διαφοροποιημένα για να φέρουν προγνωστική ισχύ στη ρύθμιση γονιδίου στόχου που επηρεάζεται από ασθένεια. Αυτή η ερώτηση είναι ιδιαίτερα σημαντική, καθώς οι αιτιώδεις αλλαγές που σχετίζονται με την ασθένεια δεν εκδηλώνονται απαραίτητα μέσω διαρρυθμίσεων που είναι αρκετά μεγάλες ώστε να ανιχνευθούν σε διορθωμένα με πολλαπλές δοκιμές στατιστικά τεστ αλλαγών έκφρασης.

Με άλλα λόγια, η αφαίρεση της διαφορικής έκφρασης των TFs θα μπορούσε να προβλέψει TF που σχετίζονται με την ασθένεια, ακόμη και αν η αλλαγή στα επίπεδα mRNA ήταν μέτρια. Αυτό υποδηλώνει ότι η προσέγγισή είναι γενικά εφαρμόσιμη για την εύρεση στοιχείων που προκαλούν νόσο σε επίπεδο TF, πέρα από αυτό που ανιχνεύει μια πιο συμβατική ανάλυση RNA-seq της έκφρασης γονιδίων.

Η μέθοδος που παρουσιάστηκε στο παρόν κείμενο είναι μια βιολογικά ερμηνεύσιμη, γενική μέθοδος μηχανικής μάθησης για την πρόβλεψη μεταγραφικών υπογραφών, συμπεριλαμβανομένων των υπογραφών ασθενειών. Τα εκπαιδευμένα

μοντέλα προβλέπουν την έκφραση των γονιδίων από την έκφραση των μεταγραφικών παραγόντων (TFs). Οι προβλεπόμενες σχέσεις μεταξύ των TF και των γονιδίων-στόχων τους επικαλύπτονται σε μεγάλο βαθμό με γνωστές δεσμεύσεις TF.

Ως εκ τούτου, η μέθοδος DNN μας υπερβαίνει τις κλασικές περιγραφικές βιοπληροφορικές τεχνικές όπως η ομαδοποίηση και η ανάλυση εμπλουτισμού. Είναι σημαντικό ότι δεν αντιμετωπίζουμε το ακόμη δυσεπίλυτο πρόβλημα της πλήρους αποσυνέλιξης ολόκληρης της κυτταρικής αλληλεπίδρασης. Αντίθετα, η μέθοδός μας εξάγει ένα βασικό στοιχείο TF από μια τέτοια περίπλοκη ρυθμιστική αρχιτεκτονική.

Όπως σε διαφορετικούς τομείς μηχανικής μάθησης, αυτά τα συστήματα είναι χρήσιμοι προγνωστικοί παράγοντες, αλλά στην πράξη λειτουργούν ως συστήματα μαύρου κουτιού. Ένα μοντέλο μαύρου κουτιού δεν προσφέρεται για ερμηνεύσιμες και ουσιαστικές αναπαραστάσεις, καθιστώντας δυνητικά το μοντέλο πιο επιρρεπές σε επιθέσεις αντιπάλου.

Τα βαθιά νευρωνικά δίκτυα (DNN) έχουν την δυνατότητα να εντοπίζουν α εντοπίζουν βιολογικά σημαντικές μοριακές αναπαραστάσεις απευθείας από δεδομένα και να φέρουν επανάσταση στην ιατρική, επομένως είναι κρίσιμης σημασίας για το πεδίο να αναπτύξουν τεχνικές που υποστηρίζουν τη βιολογική ερμηνεία και τις ιδέες από τέτοια μοντέλα πρόβλεψης.

Ο κύριος στόχος είναι να σχεδιαστεί μια περιορισμένη προσέγγιση μηχανικής μάθησης, έτσι ώστε ο προγνωστικός παράγοντας να είναι ερμηνεύσιμος από βιολογική άποψη. Τα TF ήταν στην πρώτη γραμμή στην ανάλυση κυτταρικού επαναπρογραμματισμού και μετατροπής τύπων κυττάρων και το γεγονός αυτό τα καθιστά κατάλληλα αντικείμενα εστίασης για την παρούσα προσέγγιση.

Το πρόβλημα που παρουσιάζεται στην πρώτη γραμμή στη βιολογία συστημάτων από την αλληλουχία του ανθρώπινου γονιδιώματος είναι η λειτουργία των ρυθμιστικών δίκτυών να ελέγχουν την ταυτότητα των κυττάρων και τα αποτελέσματα φιλτραρίσματος των γενετικών παραλλαγών. Ωστόσο, εάν ισχυρές μέθοδοι, όπως η προτεινόμενη τεχνική DNN, μπορούσαν να διαλευκάνουν το τμήμα TF ενός τέτοιου δικτύου, θα μπορούσαμε ενδεχομένως να προσεγγίσουμε το συγκεκριμένο πρόβλημα με έναν σταδιακό τρόπο.

Στην συνέχεια του κειμένου, αναφέρονται τα υλικά και οι μέθοδοι που χρησιμοποιήθηκαν για την προσέγγιση μηχανικής μάθησης με χρήση DNN που παρουσιάστηκε για την πρόβλεψη υπογραφών μεταγραφής σε όλο το γονιδίωμα του ανθρώπου. Συγκεκριμένα, αναλύονται η επεξεργασία των δεδομένων, η σχεδίαση του μοντέλου και την ανάλυση light-up. Ακόμα, οι συγγραφείς παραθέτουν τον αλγόριθμο της αντίστροφης επιλογής που χρησιμοποίησαν για να προσδιορίσουν το βασικό σύνολο των TF, δηλαδή το ελάχιστο σύνολο TF που θα μπορούσε να προβλέψει τα γονίδια-στόχους.

Τέλος, περιγράφεται ο τρόπος που αναλυθήκαν οι ασθένειες και ακολούθως δίνονται οι αντίστοιχοι σύνδεσμοι για τα δεδομένα και για τον κώδικα που χρησιμοποιήθηκαν για την παραγωγή των αποτελεσμάτων.

1.3 Paper 3: Βιολογική ερμηνεία DNN (βαθέων νευρωνικών δικτύων) για πρόβλεψη φαινοτύπου με βάση τη γονιδιακή έκφραση

Η ιατρική ακριβείας συνίσταται στη χρήση γενετικών χαρακτηριστικών ασθενών με σκοπό την καθοδήγηση και τη βελτίωση της λήψης κλινικών αποφάσεων, όπως διάγνωση, πρόγνωση, επιλογή της καταλληλότερης θεραπείας κ.λπ. Έχει τη δυνατότητα να αλλάξει ριζικά τις ιατρικές πρακτικές. Η διαφοροποίηση της γονιδιακής έκφρασης επιτρέπει τη μελέτη πολύπλοκων παθολογιών. Η χρήση ταξινομητών, που κατασκευάζονται από προφίλ γονιδιακής έκφρασης στην κλινική έρευνα για να βοηθήσουν στη λήψη αποφάσεων, γίνεται όλο και πιο σημαντική.

Μεταξύ των διαφόρων προσεγγίσεων μηχανικής μάθησης, η βαθιά μάθηση έχει γίνει μια από τις πιο ισχυρές μεθόδους. Ο κύριος τομέας εφαρμογής του είναι η αναγνώριση εικόνας και η αναγνώριση ομιλίας, όπου έχει ξεπεράσει άλλα ρεκόρ τεχνικών μηχανικής εκμάθησης. Οι αλγόριθμοι βαθιάς μάθησης είναι πολλά υποσχόμενοι σε πολλούς άλλους τομείς της επιστήμης, ειδικά στην ιατρική ακριβείας και την ανάλυση δεδομένων γονιδιωματικής, καθώς είναι πολύ καλοί στην ανακάλυψη περίπλοκων δομών σε δεδομένα υψηλών διαστάσεων. Σε αντίθεση με τις εικόνες ή τα δεδομένα κειμένου, τα δεδομένα γονιδιακής έκφρασης δεν έχουν δομή που να μπορεί να αξιοποιηθεί στην αρχιτεκτονική ενός νευρωνικού δικτύου. Η αρχιτεκτονική που χρησιμοποιείται για την πρόβλεψη από δεδομένα γονιδιακής έκφρασης είναι επομένως το πολυστρωματικό perceptron. Ο αυτόματος κωδικοποιητής είναι μια άλλη αρχιτεκτονική που χρησιμοποιείται συνήθως για τη μείωση της διάστασης των δεδομένων γονιδιακής έκφρασης, όπως οι αυτοκωδικοποιητές αποθορυβοποίησης ή οι αυτοκωδικοποιητές παραλλαγών.

Ένα από τα κύρια προβλήματα της βαθιάς μάθησης στις ιατρικές εφαρμογές είναι η έλλειψη ερμηνευσιμότητας. Τα νευρωνικά δίκτυα μπορούν να θεωρηθούν ως μαύρα κουτιά, όπου το προφίλ γονιδιακής έκφρασης ενός ασθενούς τοποθετείται στο στρώμα εισόδου του και λαμβάνεται μια πρόβλεψη από το στρώμα εξόδου του χωρίς να παρέχεται καμία εξήγηση για τη διαδικασία λήψης απόφασης. Η ανάγκη να γίνουν τα βαθιά νευρωνικά δίκτυα πιο ερμηνεύσιμα αυξάνεται επομένως, ειδικά στον ιατρικό τομέα για δύο κυρίως λόγους. Πρώτον, είναι σημαντικό να διασφαλιστεί ότι ένα νευρωνικό δίκτυο βασίζει τις προβλέψεις του σε αξιόπιστες αναπαραστάσεις και δεν εστιάζει σε ένα τεχνούργημα των δεδομένων. Δεύτερον, ένα νευρωνικό δίκτυο με υψηλές επιδόσεις πρόβλεψης μπορεί να έχει εντοπίσει μοτίβα στη γονιδιακή έκφραση που θα μπορούσαν να οδηγήσουν σε νέες βιολογικές υποθέσεις. Για να διερευνήσουμε αυτά τα μοτίβα είναι σημαντικό να κατανοήσουμε ποια είναι η βιολογική σημασία των κρυφών στρωμάτων του δικτύου.

Η ερμηνεία των αλγορίθμων μηχανικής μάθησης εξακολουθεί να είναι ένα αναδυόμενο πεδίο έρευνας ειδικά για μοντέλα βαθιάς μάθησης. Μπορούν να εντοπιστούν δύο τύποι ερμηνείας: ερμηνεία πρόβλεψης και ερμηνεία μοντέλου. Η ερμηνεία πρόβλεψης συνίσταται στην εξήγηση της πρόβλεψης μιας συγκεκριμένης εισροής, ενώ η ερμηνεία του μοντέλου εξηγεί τη λογική πίσω από το μοντέλο κατά την πρόβλεψη των διαφορετικών εκροών σε ολόκληρο τον πληθυσμό. Και τα δύο είναι σημαντικά για ιατρικές εφαρμογές. Η ερμηνεία των νευρωνικών δικτύων που δημιουργούνται από την έκφραση γονιδίων δεν έχει μελετηθεί διεξοδικά. Στόχος όλων αυτών των μελετών είναι να εντοπιστούν δυνητικά ενδιαφέροντα γονίδια που σχετίζονται με την ασθένεια που μας ενδιαφέρει. Στόχος όλων αυτών των μελετών είναι να εντοπιστούν δυνητικά ενδιαφέροντα γονίδια που σχετίζονται με την ασθένεια που μας ενδιαφέρει. Ωστόσο, δεν εξηγούν τι κάνει το δίκτυο, ή τι αντιπροσωπεύει έναν νευρώνα, ή ποια αναπαράσταση του ασθενούς κατασκευάζεται στα κρυφά στρώματα. Πολύ λίγες εργασίες προσπάθησαν να ερμηνεύσουν τους κρυμμένους νευρώνες και σχεδόν όλες βασίζονται στην ανάλυση των τιμών ή στην κατανομή των βαρών σύνδεσης του μαθημένου νευρωνικού δικτύου.

Πρόσφατες εργασίες στην κοινότητα μηχανικής μάθησης δείχνουν ότι η χρήση μεθόδων gradient παράγει καλύτερες ερμηνείες ενός νευρωνικού δικτύου από την ανάλυση των βαρών τους. Η αρχή των μεθόδων κλίσης είναι να διαδίδεται η ενεργοποίηση του νευρώνα εξόδου μέσω του δικτύου και να εκτιμάται για κάθε επίπεδο η επίδραση των νευρώνων και των συνδέσεων στην έξοδο. Από ό,τι γνωρίζουμε, μόνο ένα έγγραφο χρησιμοποίησε τη μέθοδο της ενσωματωμένης κλίσης για να αναγνωρίσει τα πιο σημαντικά γονίδια που σχετίζονται με έναν χώρο αναπαράστασης χαμηλών διαστάσεων (LDR) που εκμάθησε χρησιμοποιώντας έναν αυτόματο κωδικοποιητή μεταβλητής. Ο κύριος στόχος της εργασίας μας είναι να ανοίξουμε το μαύρο κουτί ενός βαθιού νευρωνικού δικτύου που έχει δημιουργηθεί από δεδομένα έκφρασης γονιδίων συνδέοντας τους νευρώνες με τη βιολογική γνώση. Η προσέγγισή μας προσαρμόζει προσεγγίσεις ερμηνείας νευρωνικών δικτύων που βασίζονται σε κλίση προκειμένου να εντοπίσει τους σημαντικούς νευρώνες. Αν και ο κύριος σκοπός αυτής της εργασίας δεν είναι η απόδοση ταξινόμησης, δείχνουμε ότι για μεγάλα σετ εκπαίδευσης, η βαθιά μάθηση υπερτερεί των κλασικών μεθόδων μηχανικής μάθησης.

Αποτελέσματα

Σύνολο γονιδιακής έκφρασης

Εφαρμόσαμε τη μέθοδό μας σε ένα πρόβλημα διάγνωσης καρκίνου που εξήχθη από δεδομένα μικροσυστοιχιών. Ο συνδυασμός διαφορετικών συνόλων δεδομένων έκφρασης δίνει έναν παγκόσμιο χάρτη γονιδιακής έκφρασης που περιέχει μεταβλητότητα που σχετίζεται με τον τύπο των ιστών και τα πειραματικά πρωτόκολλα. Αυτό μας επιτρέπει να αντιμετωπίσουμε νέα ερωτήματα και να κάνουμε πρωτότυπες μελέτες που μπορεί να οδηγήσουν σε νέες βιολογικές ανακαλύψεις. Μετά τον ποιοτικό έλεγχο και την κανονικοποίηση, το σύνολο δεδομένων περιέχει την έκφραση 54675 ανιχνευτών από 27887 ιστούς από τους οποίους 9450 είναι υγιείς και 18437 καρκίνοι.

Μοντέλο νευρωνικών δικτύων

Κατασκευάσαμε ένα βαθύ πολυστρωματικό perceptron με ένα στρώμα εισόδου 54675 νευρώνων, τρία κρυφά στρώματα 500, 200, 50 νευρώνων αντίστοιχα και ένα στρώμα εξόδου δύο νευρώνων που αντιστοιχούν στις κατηγορίες μη καρκίνου και καρκίνου. Το δίκτυο μαθαίνεται χρησιμοποιώντας τον βελτιστοποιητή adam και το τέλος της προπόνησης ελέγχεται με διαδικασία έγκαιρης διακοπής με μέγιστο 500 εποχές. Οι υπερπαράμετροι βελτιστοποιούνται σε ένα σύνολο επικύρωσης που περιέχει το 10% του συνόλου εκπαίδευσης.

Οι επιδόσεις του νευρωνικού δικτύου (NN) συγκρίνονται με τις σύγχρονες μεθόδους εποπτευόμενης μάθησης. Δείχνουμε ότι για μικρό μέγεθος σετ προπόνησης οι αλγόριθμοι τελευταίας τεχνολογίας παρέχουν υψηλότερη ακρίβεια από το NN, για μέτριο μέγεθος σετ προπόνησης, η ακρίβεια του NN είναι παρόμοια με την ακρίβεια των άλλων μεθόδων, για μεγάλο μέγεθος συνόλου εκπαίδευσης το NN ξεπερνά σημαντικά τις άλλες μεθόδους. Η ακρίβεια του NN εξακολουθεί να αυξάνεται για πιο μεγάλο όγκο δεδομένων, άρα θα μπορούσαμε εύλογα να υποθέσουμε ότι η ακρίβειά του θα συνεχίσει να βελτιώνεται εάν υπάρχουν διαθέσιμα περισσότερα παραδείγματα εκπαίδευσης.

Σε αυτή την εργασία, εστιάζουμε στην ερμηνεία των μοντέλων βαθιάς μάθησης, προσδιορίζοντας τα σχετικά στοιχεία εισόδου και τα χαρακτηριστικά υψηλού επιπέδου που μαθαίνει το μοντέλο.

Ανάλυση βαθμολογιών συνάφειας

Προκειμένου να ερμηνευτεί το μοντέλο που μαθαίνεται, υπολογίζεται το διάνυσμα συνάφειας κάθε παραδείγματος στο σύνολο δοκιμής. Το διάνυσμα συνάφειας ενός παραδείγματος περιέχει τη βαθμολογία συνάφειας, που υπολογίζεται από το LRP, όλων των νευρώνων του δικτύου. Σημειώστε ότι το LRP εφαρμόζεται από τον νευρώνα εξόδου που αντιστοιχεί στην προβλεπόμενη κλάση. Επομένως, ένα διάνυσμα συνάφειας αντιπροσωπεύει ποιο τμήμα του δικτύου είναι το πιο υπεύθυνο για την πρόβλεψη ενός δεδομένου παραδείγματος. Μια ανάλυση αυτών των μεμονωμένων διανυσμάτων συνάφειας δείχνει ότι, για σχεδόν όλα τα παραδείγματα, μόνο ένα μικρό σύνολο νευρώνων είναι σημαντικό, δηλαδή έχει υψηλή βαθμολογία συνάφειας. Ωστόσο, οι σημαντικοί νευρώνες που σχετίζονται με δύο παραδείγματα μπορεί να είναι πολύ διαφορετικοί, ακόμα κι αν το μοντέλο τους εκχωρεί την ίδια κλάση. Για κάθε τάξη, μπορούμε να προσδιορίσουμε διαφορετικές ομάδες διανυσμάτων συνάφειας, που σημαίνει ότι διαφορετικά εξειδικεύονται στο να προβλέπουν από διαφορετικά γκρουπς παραδειγμάτων, ακόμα κι αν αυτά βρίσκονται σε διαφορετικές τάξεις.

Μια ενδιαφέρουσα παρατήρηση είναι ότι τα σφάλματα πρόβλεψης τείνουν να ομαδοποιούνται σε ορισμένες ομάδες. Αυτό σημαίνει ότι το λάθος των προβλέψεων προέρχεται συχνά από το ίδιο σύνολο νευρώνων. Αυτό σημαίνει ότι ο τρόπος με τον οποίο ένα παράδειγμα διαδίδεται μέσω του δικτύου δεν εξαρτάται από τον ιστό. Δύο εξηγήσεις είναι δυνατές. Το πρώτο είναι ότι το δίκτυο έχει ανακαλύψει μια γενική υπογραφή καρκίνου για κάθε τύπο ιστού. Η δεύτερη

εξήγηση είναι ότι το δίκτυο βρήκε διαφορετικές υπογραφές για τους διαφορετικούς ιστούς, αλλά αυτές οι υπογραφές συγχωνεύονται στο ίδιο σύνολο νευρώνων. Σε αυτή την περίπτωση, θα μπορούσε να είναι ενδιαφέρον να τροποποιήσουμε την αρχιτεκτονική του δικτύου προσθέτοντας βοηθητικές εξόδους που προβλέπουν τον τύπο ιστού από κρυφά στρώματα, έτσι ώστε οι διαφορετικοί ιστοί να χρησιμοποιούν διαφορετικούς νευρώνες στο κρυφό στρώμα.

Σύγκριση με ερμηνεία WM

Στην πλειονότητα των εργασιών ερμηνείας NN για γονιδιακή έκφραση, η αξιολόγηση της επίδρασης ενός νευρώνα (εισόδου ή κρυφού) βασίζεται στον μέσο όρο βάρους των συνδέσεων εξόδου τους (WM). Ωστόσο, σε πολλές περιπτώσεις, η βαθμολογία WM δεν αντιπροσωπεύει την πραγματική συμβολή μιας εισόδου ή νευρώνα στην πρόβλεψη.

Έπειτα παρουσιάζεται εξήγηση στο θέμα μέσα από λεπτομερή παραδείγματα.

Βιολογική ερμηνεία του μοντέλου.

Σε αυτή την ενότητα, επεξηγούμε το ενδιαφέρον της μεθόδου μας παρέχοντας μια βιολογική ερμηνεία του νευρωνικού δικτύου που προβλέπει τον καρκίνο από την προηγούμενη ενότητα. Για κάθε τάξη, οι σημαντικοί νευρώνες κάθε στρώματος προσδιορίζονται από το μέσο διάνυσμα των βαθμολογιών συνάφειας. Εν συντομία, μπορούμε να πούμε ότι οι σημαντικοί νευρώνες που εντοπίσαμε, επικεντρώθηκαν κυρίως στον κυτταρικό κύκλο και στις οδούς που συνυπάρχουν μαζί του, όπως μεταβολικές διεργασίες και διεργασίες ματίσματος RNA. Το πρώτο στρώμα ήταν το πιο γενικό στρώμα με ελάχιστη εξειδίκευση. Καθώς προχωρούσαμε στα στρώματα, κάθε νευρώνας είχε την τάση να ειδικεύεται σε ένα στοιχείο του κυτταρικού κύκλου με το τελευταίο στρώμα να έχει έναν νευρώνα ειδικό για την επεξεργασία μη κωδικοποιητικού RNA.

Συζήτηση

Επισημαίνουμε ότι στόχος της ερμηνείας είναι να εξηγήσει πώς λειτουργεί το μοντέλο και όχι πώς λειτουργεί η βιολογία. Όταν μια βιολογική λειτουργία, που δεν σχετίζεται με τον φαινότυπο, προσδιορίζεται στους σημαντικούς νευρώνες, είναι πιθανό αυτή η λειτουργία είτε να έχει έμμεση συσχέτιση είτε να συνδέεται με μια άγνωστη σχέση αιτιότητας με τον φαινότυπο. Επιπλέον, ας μην ξεχνάμε ότι οι βιολογικές βάσεις δεδομένων που χρησιμοποιούνται δεν είναι μια τέλεια περιγραφή της βιολογίας, αλλά απλώς μια αναπαράσταση της τρέχουσας γνώσης της βιολογίας.

Εντοπίζουμε τρεις περιπτώσεις με βάση τα αποτελέσματα της ερμηνείας. Στην πρώτη περίπτωση, η πλειονότητα των στοιχείων που παρέχει η βιολογική ερμηνεία σχετίζονται με τον προβλεπόμενο φαινότυπο. Αυτό σημαίνει ότι το μοντέλο βασίζει τις προβλέψεις του σε στοιχεία που συνάδουν με τη βιολογική γνώση. Αυτό θα βελτιώσει την εμπιστοσύνη στο μοντέλο εκτός από την απόδοση πρόβλεψής του. Η δεύτερη περίπτωση είναι το αντίθετο. Τα περισσότερα μέρη των στοιχείων που παρέχονται από τη βιολογική ερμηνεία είναι γνωστό ότι δεν σχετίζονται με τον προβλεπόμενο φαινότυπο. Δεδομένου ότι οι προβλέψεις βασίζονται σε στοιχεία που δεν συνάδουν με την τρέχουσα βιολογική γνώση, η αξιοπιστία του μοντέλου πρέπει να αμφισβητηθεί. Το μοντέλο μπορεί να ταιριάζει υπερβολικά ή να

παραπλανηθεί από μια προκατάληψη στο σετ εκπαίδευσης. Στην τελευταία περίπτωση, η βιολογική ερμηνεία παρέχει κυρίως στοιχεία που μπορεί να σχετίζονται ή να μην σχετίζονται με τον προβλεπόμενο φαινότυπο αλλά μπορεί να οδηγήσει σε νέες βιολογικές υποθέσεις που θα διερευνηθούν από τους βιολόγους.

Συμπέρασμα

Σε αυτό το άρθρο, προτείνουμε μια πρωτότυπη προσέγγιση για τη βιολογική ερμηνεία των μοντέλων βαθιάς μάθησης για την πρόβλεψη φαινοτύπων από δεδομένα γονιδιακής έκφρασης. Ο κύριος στόχος μας είναι να εντοπίσουμε τους νευρώνες και τις εισόδους του ΝΝ που συμβάλλουν στις προβλέψεις και να τις συνδέσουμε με τη βιολογική γνώση. Το μοντέλο περιορίζεται σε ένα υποδίκτυο που περιέχει τις σχετικές συνδέσεις και τους νευρώνες που εμπλέκονται στην πρόβλεψη. Στη συνέχεια, αυτοί οι νευρώνες συνδέονται με μια λίστα γονιδίων και την αντίστοιχη βιολογική γνώση (GO, KEGG και DOLite). Τα πειράματά μας, που βασίζονται στην πρόβλεψη του καρκίνου, δείχνουν ότι (1) η προσέγγιση βαθιάς μάθησης υπερτερεί των κλασικών μεθόδων μηχανικής μάθησης σε μεγάλα σύνολα δεδομένων εκπαίδευσης, (2) η προσέγγισή μας παράγει ερμηνείες πιο συνεπείς με τη βιολογία από την τελευταία λέξη της τεχνολογίας με βάση την προσέγγιση WM , (3) μπορούμε να παρέχουμε μια περιεκτική εξήγηση των προβλέψεων για βιολόγους και γιατρούς. Οι συνεχιζόμενες εργασίες αφορούν πρόσθετη βιολογική ανάλυση, σύγκριση και επικύρωση που είναι απαραίτητες για να έχουμε μια ολοκληρωμένη εικόνα της λογικής πίσω από τις προβλέψεις των νευρωνικών δικτύων. Οι μελλοντικές εργασίες αφορούν την εισαγωγή βιολογικής γνώσης μέσα στο νευρωνικό δίκτυο προκειμένου να καθοδηγηθεί η φάση εκμάθησης του μοντέλου. Αυτό επιτρέπει την εκμάθηση γνωστών βιολογικών εννοιών και μπορεί να οδηγήσει σε βιολογικές ανακαλύψεις. Οι μελλοντικές εργασίες αφορούν την εισαγωγή βιολογικής γνώσης μέσα στο νευρωνικό δίκτυο προκειμένου να καθοδηγηθεί η φάση εκμάθησης του μοντέλου. Αυτό επιτρέπει την εκμάθηση γνωστών βιολογικών εννοιών και μπορεί να οδηγήσει σε βιολογικές ανακαλύψεις. Οι μελλοντικές εργασίες αφορούν την εισαγωγή βιολογικής γνώσης μέσα στο νευρωνικό δίκτυο προκειμένου να καθοδηγηθεί η φάση εκμάθησης του μοντέλου. Αυτό επιτρέπει την εκμάθηση γνωστών βιολογικών εννοιών και μπορεί να οδηγήσει σε βιολογικές ανακαλύψεις.

Μέθοδοι

Παρουσιάζουμε την αρχιτεκτονική του βαθιού νευρωνικού δικτύου που χρησιμοποιείται για τα δεδομένα γονιδιακής έκφρασης και την προσέγγιση βιολογικής ερμηνείας μας. Η μέθοδος gradient για την ερμηνεία του νευρωνικού δικτύου είναι η Layer-wise Relevance Propagation (LRP), η οποία είναι προσαρμοσμένη για να προσδιορίζει τους σημαντικότερους νευρώνες που οδηγούν στην πρόβλεψη καθώς και στην αναγνώριση του συνόλου των γονιδίων που ενεργοποιούν αυτούς τους σημαντικούς νευρώνες. Τέλος, οι σημαντικοί νευρώνες και γονίδια συνδέονται με την Gene Ontology (GO), την Kyoto Encyclopedia of Genes and Genomes (KEGG) και τη Disease Ontology Annotation List (DOLite) προκειμένου να προταθεί μια βιολογική ερμηνεία του μοντέλου νευρωνικών δικτύων.

Βαθιού πολυστρωματικού perceptron
Έπειτα ακολουθεί μια λεπτομερής περιγραφή του βαθιού πολυστρωματικού perceptron και του πως λειτουργεί.

Διάδοση συνάφειας βάση επιπέδου (LRP)

Η πρόβλεψη του φαινοτύπου ενός ασθενούς λαμβάνεται με τη διάδοση του προφίλ γονιδιακής έκφρασης του μέσω του δικτύου και την αξιολόγηση των νευρώνων στο πέρασμα τροφοδοσίας προς τα εμπρός. Οι μέθοδοι κλίσης υπολογίζουν την επίδραση κάθε μεταβλητής και νευρώνα του δικτύου για μια δεδομένη πρόβλεψη. Μεταξύ όλων των μεθόδων κλίσης που παρουσιάζονται στην ενότητα «Εισαγωγή», επιλέξαμε το LRP για δύο λόγους: πρώτον, το LRP παράγει αποτελέσματα καλά ευθυγραμμισμένα με την ανθρώπινη διαίσθηση, δεύτερον δεν χρειάζεται εισόδους αναφοράς.

Η ιδέα του LRP είναι να διαδώσει πίσω το σήμα του νευρώνα εξόδου που μας ενδιαφέρει μέσω του δικτύου. Αρχικά, το LRP αναπτύχθηκε για να ερμηνεύει την πρόβλεψη από εικόνες, δηλαδή να υπολογίζει τη συμβολή κάθε pixel στην πρόβλεψη της κατηγορίας μιας δεδομένης εικόνας. Σε αυτή την εργασία, προσαρμόζουμε και χρησιμοποιούμε το LRP στο πλαίσιο των δεδομένων γονιδιακής έκφρασης. Επιπλέον, η εργασία μας επικεντρώνεται στο πρόβλημα της ερμηνείας του μοντέλου παρά στην ερμηνεία πρόβλεψης. Επομένως, η ανάλυσή μας βασίζεται στον μέσο όρο συνάφειας που υπολογίζεται από ένα υποσύνολο των συνόλων δοκιμών και όχι σε μεμονωμένες βαθμολογίες συνάφειας. Το LRP μπορεί επίσης να χρησιμοποιηθεί για να εξηγήσει την μεμονωμένη πρόβλεψη, αλλά αυτή η ανάλυση δεν εμπίπτει στο πεδίο εφαρμογής αυτής της εργασίας.

Προσέγγιση βιολογικής ερμηνείας
Ο στόχος είναι να εντοπιστούν οι βιολογικές λειτουργίες και τα μεταβολικά μονοπάτια που χρησιμοποιεί το νευρωνικό δίκτυο για να προβλέψει κάθε κατηγορία. Για κάθε τάξη, η προτεινόμενη ερμηνευτική προσέγγιση μπορεί να αποσυντεθεί σε τρία βήματα. Στο πρώτο βήμα, υπολογίζουμε τις βαθμολογίες συνάφειας μέσω του δικτύου και προσδιορίζουμε τους πιο σημαντικούς νευρώνες που επιτρέπουν την πρόβλεψη της τάξης. Στη συνέχεια, συσχετίζουμε με κάθε σημαντικό νευρώνα μια λίστα με τα σημαντικά γονίδια που επηρεάζουν την ενεργοποίηση του νευρώνα. Τέλος, βιολογικές λειτουργίες, μεταβολικές οδοί και ασθένειες συνδέονται με κάθε σημαντικό νευρώνα.

1.4. Paper 4: OptNCMiner: μια προσέγγιση βαθιάς μάθησης για την ανακάλυψη φυσικών ενώσεων που διαμορφώνουν πολλαπλούς στόχους για συγκεκριμένες ασθένειες

Τα φυσικά προϊόντα (NPs) ορίζονται ως οι ουσίες που παράγονται από ζωντανούς οργανισμούς και χρησιμοποιούνται για να καλύψουν τις φυσικές ενώσεις (NCs) και μείγματα αυτών που προέρχονται από φυσικές πηγές. Οι βιολογικές δραστηριότητες των NPs οφείλονται στις δραστηριότητες των NC που τα αποτελούν.

Ας αναλύσουμε τα χαρακτηριστικά των φυσικών ενώσεων και τους τομείς που μπορούν να χρησιμοποιηθούν. Πιο συγκεκριμένα, βρίσκουν εφαρμογή στην φαρμακολογία και στην ιατρική καθώς χρησιμοποιούνται ευρέως για την θεραπεία σοβαρών ασθενειών. Επιπλέον, η χρήση τους έχει επεκταθεί και σε βιομηχανίες τροφίμων αλλά και καλλυντικών.

Ένα επιπλέον χαρακτηριστικό των φυσικών ενώσεων είναι ότι έχουν την τάση να διαμορφώνουν πολλαπλές πρωτεΐνες – στόχους. Το συγκεκριμένο χαρακτηριστικό μπορεί να εφαρμοστεί για την δημιουργία φαρμάκων πολλαπλών στόχων για ασθένειες με πολύπλοκες αιτιολογίες ή προβλήματα αντοχής στα φάρμακα. Η παραπάνω εφαρμογή των NCs έχει αλλάξει από το ένας στόχος – μια ασθένεια και έχει ως στόχο να ενισχύσει την κλινική αποτελεσματικότητα και να βελτιώσει ζητήματα ασφάλειας.

Παρόλα αυτά όμως, η διαδικασία ανακάλυψης για NCs που διαμορφώνουν πολλαπλούς στόχους είναι κουραστική και δαπανηρή. Για αυτό τον λόγο, έχουν μέθοδοι τρισδιάστατης μοντελοποίησης για την προσομοίωση μιας πολύπλοκης μοριακής δομής και διαμορφωτικού χώρου των NCs, καθώς και των αλληλεπιδράσεών τους με τις πρωτεΐνες-στόχου. Τα εργαλεία αυτά μαζί με τις βάσεις δεδομένων αλληλεπίδρασης χημικών πρωτεϊνών, έχουν ρίξει φως στην ανάπτυξη μοντέλων μηχανικής μάθησης που προβλέπουν νέα NC και βελτιώνουν τη διαδικασία ανακάλυψης NC.

Ειδικότερα, τα DNN (Βαθιά Νευρωνικά Δίκτυα) έχουν εφαρμοστεί ευρέως στον τομέα ανακάλυψης ενεργών ενώσεων καθώς επιτρέπουν την αυτοματοποίηση της

διαδικασίας μηχανικής χαρακτηριστικών που συχνά γίνεται εμπόδιο στις συμβατικές μεθόδους μηχανικής μάθησης. Επιπλέον, το DNN δημιουργεί ή εξάγει σημαντικά κύρια χαρακτηριστικά από τα διανύσματα εισόδου των ενώσεων που είναι υπεύθυνες για την δραστηριότητα τους. Αξίζει να σημειωθεί όμως ότι η απόδοση του DNN εξαρτάται από την ποσότητα και την ποιότητα των δεδομένων και επιπλέον δεν έχουν όλες οι πρωτεΐνες επαρκή δεδομένα για την αξιόπιστη εκπαίδευση μοντέλων μηχανικής μάθησης.

Για την αντιμετώπιση των θεμάτων που αναφέρθηκαν προτάθηκε το SNN (Siamese Neural Network), ένα ισχυρό εργαλείο που αποτελείται από 2 πανομοιότυπα δίκτυα (κεφαλές και σώμα) και επιτρέπει τις συγκρίσεις ομοιότητας σύνθετων δεδομένων και μπορεί να εφαρμοστεί σε μάθηση μιας η πολλών λήψεων. Ακόμα, το SNN χρησιμοποιεί συναρτήσεις αντίθεσης απώλειας και είναι ικανό για few-shot learning, ένα είδος εκπαίδευσης για προβλέψεις σε μονό η μικρό αριθμό δειγμάτων. Είναι ένα συμβατό μοντέλο και ειδικό για εκμάθηση δεδομένων ενώσεων που αλληλεπιδρούν με πρωτεΐνη.

Στο παρόν κείμενο, παρουσιάζεται το μοντέλο μηχανικής μάθησης OptNCMiner που είναι κατάλληλο για πρόβλεψη των βέλτιστων NCs που διαμορφώνουν πολλαπλούς στόχους για συγκεκριμένες ασθένειες. Το συγκεκριμένο μοντέλο είναι χτισμένο σε δομή SNN και διατηρεί τα πλεονεκτήματα του DNN για αποτελέσματα εξαγωγής χαρακτηριστικών των φυσικών ενώσεων NCs που σχετίζονται με αλληλεπιδράσεις χημικής πρωτεΐνης. Το OptNCMiner δοκιμάστηκε με την ανακάλυψη φυσικών πηγών που περιέχουν NCs που ρυθμίζουν πρωτεΐνες-στόχους που σχετίζονται με επιπλοκές που σχετίζονται με τον σακχαρώδη διαβήτη τύπου 2 (T2DM).

Σχετικά με αυτή την δοκιμή, αναφέρεται ότι το μοντέλο OptNCMiner υπολογίζει τη βαθμολογία δραστηριότητας των NCs με κάθε πρωτεΐνη-στόχο στο πλαίσιο μιας βαθμολογίας ομοιότητας μεταξύ NC και γνωστών ενεργών ενώσεων πρωτεϊνών-στόχων. Επιπλέον, ύστερα από συγκρίσεις ανακαλύφθηκε ότι το μοντέλο προβλέπει επιτυχώς τόσο γνωστές όσο και άγνωστες αλληλεπιδράσεις χημικής πρωτεΐνης.

Στην συνέχεια, γίνεται γνωστή η συλλογή των δεδομένων από 3 διαφορετικά σύνολα που αποτελούνται από ενεργές και ανενεργές ενώσεις. Τα σύνολα δεδομένων αυτά είναι τα: base dataset, transfer learning dataset, few-shot learning dataset και αναλύονται το καθένα ξεχωριστά. Θα πρέπει ακόμα να αναφερθεί ότι το μοντέλο OptNCMiner είναι δίκτυο που δέχεται εισόδους με την μορφή ζευγών και υπολογίζει την ομοιότητα μεταξύ των δύο. Πιο συγκεκριμένα, αυτά τα ζεύγη ενώσεων θεωρούνται θετικά αν ταξινομηθούν και τα 2 ως ενεργά αλλιώς χαρακτηρίζονται ως αρνητικά.

Όπως είπαμε και νωρίτερα, το OptNCMiner έχει χτιστεί σε δομή SNN και τα ζεύγη εισόδων τροφοδοτούνται σε πανομοιότυπα perceptrons πολλαπλών επιπέδων που αποτελούν την head function για δημιουργία ενσωματωμένων διανυσμάτων. Η ομοιότητα για τα 2 ενσωματωμένα διανύσματα υπολογίζεται από το body function.

Στην συνέχεια, ορίζονται και άλλα μέρη του μοντέλου όπως πχ η συνάρτηση απώλειας και στο τέλος για την ταξινόμηση της δέσμευσης της ένωσης με την πρωτεΐνη στόχο, η υψηλότερη ομοιότητα για κάθε πρωτεΐνη συγκρίθηκε με ένα κατώφλι.

Ανάλογα με τα μεγέθη των συνόλων δεδομένων, χρησιμοποιήθηκαν και διαφορετικές προσεγγίσεις εκπαίδευσης: transfer learning και few-shot learning. Όσον αφορά το πρώτο, σύμφωνα με αυτό, το μοντέλο εκπαιδεύεται πρώτα σε ένα μεγαλύτερο σύνολο δεδομένων πριν εκπαιδευτεί περαιτέρω στο σύνολο δεδομένων του στόχου. Για δεδομένα πολύ μικρά για να μπορέσουν να εκπαιδευτούν μπορεί να χρησιμοποιηθεί η δεύτερη προσέγγιση εκμάθησης δηλαδή το few-shot learning, η εκμάθηση με λίγες λήψεις.

Επιπρόσθετα, μπορούμε να αναλύσουμε τις μετρήσεις αξιολόγησης του μοντέλου OptNCMiner. Συγκεκριμένα, είναι ένα μοντέλο ταξινόμησης και για αυτό το λόγο χρησιμοποιούνται τυπικές μετρικές για την αξιολόγηση της απόδοσης του. Κάποιες από τις μετρήσεις είναι η ανάκληση (recall) που αξιολογεί την εσφαλμένη ταξινόμηση των πραγματικών θετικών στοιχείων, η ακρίβεια (accuracy) που μετράει την γενική απόδοση του μοντέλου και η περιοχή κάτω από τα χαρακτηριστικά λειτουργίας του δέκτη (AUROC) που αντιπροσωπεύει τον βαθμό διαχωρισιμότητας κλάσεων του μοντέλου.

Από τα αποτελέσματα που προκύπτουν για τις ενώσεις μπορούμε να συμπεράνουμε ότι οι ενώσεις έχουν παρόμοιες φυσικοχημικές ιδιότητες αλλά είναι δομικά διαφορετικές. Εξαιτίας αυτού όμως δεν υπήρχαν προκαθορισμένες δομές ή φυσικοχημικές ιδιότητες για να προβλέψουμε διαφορετικές πρωτεΐνες στόχους. Για αυτό το λόγο λοιπόν, το OptNCMiner εκπαιδεύτηκε για τέτοια σύνθετα δεδομένα ώστε μάθει αυτά τα κρυφά χαρακτηριστικά για να διακρίνει την δεσμευτική φύση των ενώσεων με διαφορετικές πρωτεΐνες στόχους.

Στην συνέχεια, αναφέρονται οι τρόποι με τους οποίους αξιολογήθηκε η απόδοση του OptNCMiner. Ένας από τους τρόπους είναι εξετάσαμε την ικανότητά του να μαθαίνει ταξινόμηση πολλαπλών ετικετών συγκρίνοντας την απόδοση του μοντέλου μετά την εκπαίδευση με δεδομένα μιας ετικέτας και πολλαπλών ετικετών.

Η απόδοση του OptNCMiner αξιολογήθηκε επίσης με τις ενώσεις που δεν χρησιμοποιήθηκαν στη δημιουργία ζεύγους εκπαίδευσης από το βασικό σύνολο δεδομένων και το σύνολο δεδομένων μεταφοράς. Τέλος, η απόδοση του μοντέλου συγκρίθηκε με 5 βασικά μοντέλα ικανά για ταξινόμηση πολλαπλών ετικετών: ομοιότητα, συνημίτονου, MB, LR, RF, MLP και η απόδοση του OptNCMiner ήταν η καλύτερη σε σχέση με τα υπόλοιπα μοντέλα καθώς πάνω από το 80% των γνωστών θετικών είχαν προβλεφθεί σωστά.

Επιπλέον, όλες οι μετρικές τιμές αξιολόγησης έχουν βελτιωθεί στο σύνολο δεδομένων εκμάθησης μεταφοράς (transfer learning) σε σύγκριση με το βασικό σύνολο δεδομένων. Η απόδοση του OptNCMiner επίσης βελτιώθηκε με την μέθοδο transfer learning ανεξάρτητα από τα χαρακτηριστικά των δεδομένων. Από την άλλη η μέτρηση της ανάκλησης ήταν καλύτερη τόσο για το βασικό όσο για το transfer learning σύνολο δεδομένων.

Ένα από τα πλεονεκτήματα του μοντέλου OptNCMiner που εντοπίζουμε είναι η ικανότητα να προβλέπει αλληλεπιδράσεις χημικής πρωτεΐνης για πρωτεΐνες με περιορισμένα δεδομένα εκμάθησης. Από την απόδοση με few-shot learning σύνολο δεδομένων επιβεβαιώνεται ότι το μοντέλο διαθέτει την ικανότητα να εντοπίζει δομικές ιδιότητες ενώσεων που επιτρέπουν συγκεκριμένες αλληλεπιδράσεις χημικής πρωτεΐνης από μικρό αριθμό δειγμάτων.

Το γεγονός ότι το OptNCMiner προβλέπει και άγνωστες πρωτεΐνες στόχους και αυτό μπορεί να οδηγήσει σε υψηλά ψευδώς θετικά ποσοστά. Για να επιβεβαιωθεί η ικανότητα αυτή θα πρέπει να γίνει επικύρωση των ψευδώς θετικών αποτελεσμάτων στο few-shot learning.

Το αποτέλεσμα του in silico docking δείχνει ότι τα ψευδώς θετικά που προβλέπονται από το OptNCMiner είναι πραγματικά άγνωστα θετικά με μεγάλη πιθανότητα.

Στην συνέχεια, παρουσιάζεται ένα use case scenario που σχετίζεται με τον εντοπισμό φυσικών ενώσεων NCs που υπάρχουν σε φυσικές πηγές που ρυθμίζουν πρωτεΐνες -στόχους που σχετίζονται με T2DM επιπλοκές. Πιο συγκεκριμένα, για αυτή την περίπτωση ανάλογα με τα 2 διαφορετικά εύρη μεγεθών των δεδομένων ένωσης που αλληλεπιδρούν εφαρμόστηκε transfer learning ή few- shot learning. Τα δεδομένα λήφθηκαν από μια βάση δεδομένων FooDB συστατικών τροφίμων και δόθηκαν σαν είσοδο στο OptNCMiner.

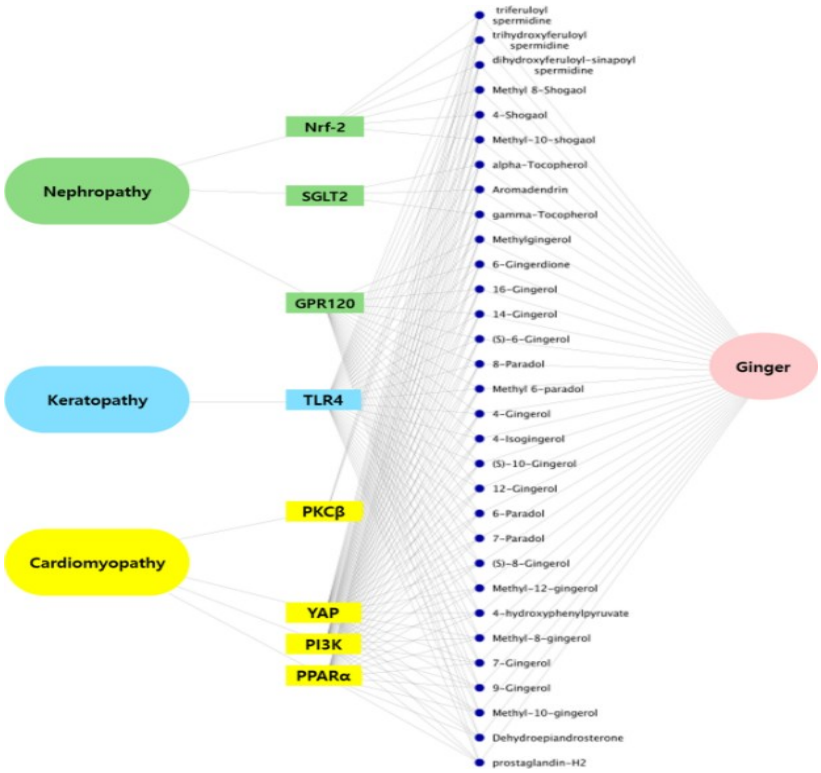
Έπειτα, βρέθηκαν NCs σε πολλαπλές πηγές τροφίμων με το φυτό τζιντζερ να αποτελεί παράδειγμα για την πηγή τροφής με το μεγαλύτερο επιλεγμένων NC (31).

Στην ρίζα του συγκεκριμένου φυτού υπάρχει τεράστια ποικιλία από NCs υπεύθυνα για βιολογική δραστηριότητα. Μεταξύ των 160 αναγνωρισμένων NCs που υπάρχουν στο τζίντζερ, οι φαινολικές και τερπενικές ενώσεις, όπως οι τζιντζερόλες και οι σογκαόλες, έχουν διερευνηθεί ευρέως για την έντονη αυστηρότητα, την αφθονία και τα πολλαπλά οφέλη για την υγεία.

Είναι ενδιαφέρον ότι το OptNCMiner προέβλεψε την 6-gingerol ως NC που βελτιώνει τις επιπλοκές του ΣΔ2, διαμορφώνοντας 5 διαφορετικές πρωτεΐνες: TLR4,

PI3K, YAP, PPARα και GPR120. Η καθεμιά από αυτές τις πρωτεΐνες σχετίζεται με επιπλοκές τουT2DM.

Σε όλες τις παραπάνω πρωτεΐνες που πρόκυψαν εμφανίζεται η 6-gingerol η όποια με κατάλληλες λειτουργίες προστατεύει και βοηθά στην βελτίωση των τριών επιπλοκών που φαίνονται και στην εικόνα του T2DM.



Το OptNCMiner έχει προβλέψει σωστά τους προηγουμένως γνωστούς στόχους της 6-gingerol καθώς και πιθανούς στόχους που δεν είναι ακόμη γνωστοί, αποδεικνύοντας την ικανότητά του να προβλέπει NC που σχετίζονται με συγκεκριμένες ασθένειες.

Συνεχίζοντας, αναφέρονται συστάσεις για τους χρήστες του μοντέλου OptNCMiner καθώς και περιθώρια για βελτίωση της απόδοσης του. Πιο συγκεκριμένα, το μοντέλο μπορεί να χρησιμοποιηθεί συνδυαστικά με άλλες μεθόδους ώστε να πετυχαίνονται καλυτέρα αποτελέσματα και σε περισσότερους τομείς όπως αλληλεπιδράσεις RNA – πρωτεϊνών , γονίδιων -ασθενειών κα.

Γενικά, υπάρχει περιθώριο για βελτίωση της απόδοσης του OptNCMiner. Ένας από τους παράγοντες που περιορίζει την απόδοση του προγράμματος είναι η πολυπλοκότητα της βιολογικής δραστηριότητας των φυσικών ενώσεων NCs στο ανθρώπινο σώμα.

Καταλήγοντας, μια σημαντική σύσταση για τους χρήστες του μοντέλου είναι η προσεκτική επιλογή πρωτεϊνών-στόχων για ανακάλυψη NC. Στο παράδειγμα των

επιπλοκών T2DM, μόνο οι πρωτεΐνες που βελτιώνουν τις τρεις ασθένειες θεωρήθηκαν πρωτεΐνες-στόχοι. Ωστόσο, για τον εντοπισμό NC που ρυθμίζουν μόνο την επιθυμητή πρωτεΐνη-στόχο, στην πράξη, πιθανές πρωτεΐνες εκτός στόχου μπορεί να επηρεαστούν με βάση τη γνώση του υποβάθρου.

Το OptNCMiner μπορεί επίσης να χρησιμοποιηθεί σε συνδυασμό με άλλα προγράμματα για την υποστήριξη της ολιστικής ανακάλυψης NC, όπως προγράμματα που προβλέπουν την απορρόφηση και την κατανομή των NC μετά την κατάποση.

2. Βιβλιογραφία

Jhalia, V. and Swarnkar, T. (2021). A Critical Review on the Application of Artificial Neural Network in Bioinformatics. In Data Analytics in Bioinformatics (eds R. Satpathy, T. Choudhury, S. Satpathy, S.N. Mohanty and X. Zhang).

https://doi.org/10.1002/9781119785620.ch3

Rasmus Magnusson, Jesper N. Tegnér & Mika Gustafsson (2022)

Deep neural network prediction of genome-wide transcriptome signatures – beyond the Black-box *Systems Biology and Applications (2022) 8:9*
https://doi.org/10.1038/s41540-022-00218-9

Blaise Hanczar , Farida Zehraoui, Tina Issa & Mathieu Arles (2020)

Biological interpretation of deep neural network for phenotype prediction based on gene expression *Hanczar et al. BMC Bioinformatics (2020) 21:501*
https://doi.org/10.1186/s12859-020-03836-4

Seo Hyun Shin, Seung Man Oh , Jung Han Yoon Park, Ki Won Lee & Hee Yang (2022)

OptNCMiner: a deep learning approach for the discovery of natural compounds modulating disease-specifc multi-targets Shin et al. BMC Bioinformatics (2022) 23:218
https://doi.org/10.1186/s12859-022-04752-5