

Ιόνιο Πανεπιστήμιο
Τμήμα Πληροφορικής



Αλγόριθμοι εύρεσης παθογενούς πρωτεϊνικής αναδίπλωσης

Πετσάβα Παναγιώτα
Επιβλέπων : Βλάμος Παναγιώτης

Επιβλέπων

Βλάμος Παναγιώτης, Καθηγητής,

Ιόνιο πανεπιστήμιο

Τριμελής Επιτροπή

Βλάμος Παναγιώτης, Καθηγητής,

Ιόνιο πανεπιστήμιο

Θεμιστοκλής Έξαρχος, Επίκουρος Καθηγητής,

Ιόνιο πανεπιστήμιο

Ανδρόνικος Θεόδωρος, Αναπληρωτής Καθηγητής,

Ιόνιο πανεπιστήμιο

Περίληψη

Η εύρεση παθογενούς πρωτεΐνης με αλγόριθμους είναι ένα από τα σημαντικότερα κομμάτια που καλείται ο τομέας της βιοπληροφορικής να επιλύσει. Τα τελευταία χρόνια έχουν γίνει πολλές έρευνες στο κομμάτι της εύρεσης παθογενούς πρωτεΐνης και με χρήση μαθηματικών μοντέλων αλλά και χρήση υπολογιστικών τεχνικών για την πρόγνωση.

Στην εργασία αυτή θα γίνει αναφορά σε βασικούς όρους που σχετίζονται με την ανάλυση και εύρεση παθογενούς πρωτεΐνης και σε βασικές τεχνικές που χρησιμοποιούνται στην ανάλυση των πρωτεϊνών με χρήση μαθηματικών και υπολογιστικών μοντέλων. Επιπλέον θα γίνει εκτενής αναφορά σε πρόσφατες εφαρμογές αλγορίθμων που έχουν επιτευχθεί για την εύρεση παθογενούς πρωτεΐνης.

Πίνακας Περιεχομένων

Περίληψη	3
Πίνακας περιεχομένων	4
Κατάλογος Σχημάτων	5
1.Εισαγωγή	7
1.1 Ορισμοί	7
1.2 DNA	7
1.3 RNA.....	9
1.4 Πρωτεΐνη	12
1.5 Πρωτεϊνική αναδίπλωση.....	13
1.6 Αλγόριθμος	14
1.7 Γενετικοί Αλγόριθμοι	15
2. Διαδικασία Εύρεσης	16
2.1 Δεδομένα	16
2.2 Ευριστικοί Αλγόριθμοι	17
2.3 Κωδικοποίηση	19
2.4 Νευρωνικά Δίκτυα	20
3. Σύγχρονοι Αλγόριθμοι	23
3.1 Unsupervised feature selection algorithm	24
3.2 GECKO algorithm.....	26
3.3 Developing parallel ant colonies filtered	27
3.4 Prediction Of Protein Secondary Structure	29
3.5 DeepCLIP	33
3.6 SCISSOR	35
3.7 Genetic algorithm-based personalized models.....	37
3.8 Predicting dynamic cellular protein	39
3.9 DeepLPI	41
3.10 An efficient three-Level parallel ABC algorithm	43
3.11 RNAsamba	45
3.12 MFEA	47
3.13 SAAFEC-SEQ.....	50
3.14 Systematic evaluation of machine learning methods	51
3.15 AnnapuRNA	54
4.Συμπεράσματα	57
5.Βιβλιογραφία	58

Κατάλογος Σχημάτων

1 Δομή του DNA.....	8
2 Δομή του RNA	10
3 Διαδικασία Μετάδοση Πληροφορίας	12
4 Κατηγορίες Δομών Πρωτεΐνης	13
5 Πρωτεϊνική Αναδίπλωση	14
6 Απεικόνιση Έννοιας Αλγορίθμου	15
7 Απεικόνιση έννοιας Γενετικού Αλγόριθμου	16
8 Απεικόνιση Αλγόριθμου FASTA	18
10 Βιολογικός Νεύρωνας	19
11 Σύγκριση Βιολογικού Νεύρωνα και Τεχνητού	21
12 Απλό Τεχνητό Νευρωνικό Δίκτυο	22
13 Διαγραμματική Απεικόνιση του Γενετικού Αλγόριθμου	24
14 Χωρισμός υπο-πινάκων	24
15 Ένωση 2 σημείων για την δημιουργία νέου	25
16 Χρήση Extreme Learning Machine	25
17 Σχηματική Απεικόνιση του Gecko Algorithm	27
18 Σχηματική Απεικόνιση του DracoRNA	28
19 Συνολική σχηματική απεικόνιση του Enhanced Fuzzy Random Forest	29
20 Συνολική διαδικασία του αλγόριθμου	30
21 Αλγόριθμος υπολογισμού Feature Vector	31
22 Αλγόριθμος εκπαίδευσης	32
23 Διαδικασία ταξινόμησης του DeepClip	33
24 Διαδικασία SCISSOR για την ανίχνευση αλλαγών σε ένα μόνο γονίδιο	36
25 Συγκριτικό του SCISSOR , JRBA και ELBA	37
26 α) Διάγραμμα του γενετικού αλγορίθμου και β) CAGE-based εξατομίκευση	38
27 Διαδικασία μοντελοποίησης και πρόβλεψης συνδέσεων RBP	39
28 Αρχιτεκτονική Νευρωνικού Δικτύου	40
29 Διάγραμμα της Αρχιτεκτονικής DeepLPI	41
30 Διάγραμμα αλγόριθμου 3L-PABCfold	43
31 Διαδικασία του αλγόριθμου DABC	45
32 Διαγραμματική Απεικόνιση RNAsamba	46
33 Προεπεξεργασία Δεδομένων σε αριθμητικές πληροφορίες	47
34 Τεχνική Μετάλλαξης για την δημιουργία πληθυσμού	48
35 Ο ψευδοκώδικας του αλγόριθμου MFEA	48
36 Αλγόριθμος υπολογισμού συνολικής απόστασης	49

37 Διαγραμματική Απεικόνιση SAAFEC-SEQ	50
38 Εφαρμογή Μηχανικής Μάθησης και Βασισμένο σε Πρότυπα	51
39 Διαγραμματική Απεικόνιση Εφαρμογής κατά την αξιολόγηση	53
40 Μεθοδολογία AnnapuRNA δημιουργία και εκτέλεση	55

Εισαγωγή

Η εύρεση παθογενούς πρωτεΐνης με αλγόριθμους είναι ένα από τα σημαντικότερα κομμάτια που καλείται ο τομέας της βιο-πληροφορικής να επιλύσει. Σε αυτό το κεφάλαιο θα γίνει εκτενής αναφορά στους όρους που θα αναφερθούν και τη χρησιμότητά τους.

Ορισμοί

Πριν γίνει εκτενής αναφορά στην εύρεση παθογενούς πρωτεΐνης θα πρέπει να γίνει αναφορά στους βασικούς ορισμούς:

- DNA
- RNA
- Πρωτεΐνη
- Πρωτεϊνική αναδίπλωση
- Αλγόριθμος
- Γενετικός Αλγόριθμος

αλλά και στην συσχέτιση των παραπάνω όρων για την μετάδοση της πληροφορίας.

DNA

Το DNA (ή αλλιώς το δεοξυριβονουκλεϊκό οξύ) είναι ένα χημικό μακρομόριο και συγκεκριμένα νουκλεϊκό οξύ, που αποτελεί το γενετικό υλικό όλων των ζωντανών οργανισμών καθώς και των περισσότερων ιών. Μέσα στο DNA περιέχονται οι πληροφορίες που καθορίζουν όλα τα χαρακτηριστικά ενός οργανισμού και οι οποίες οργανώνονται σε λειτουργικές μονάδες, τα γονίδια.

Στην αντιγραφή του DNA διατηρείται και μεταβιβάζεται η γενετική πληροφορία από κύτταρο σε κύτταρο και από έναν οργανισμό στους απογόνους του. Επίσης το DNA επιτυγχάνει την έκφραση των γενετικών πληροφοριών ελέγχοντας την σύνθεση των πρωτεϊνών.

Πρόκειται για μεγαλομοριακή ένωση που συγκροτείται από αζωτούχες-πρωτεϊνικές βάσεις, φωσφορικές ρίζες και ένα σάκχαρο με πέντε άτομα άνθρακα (πεντόζη), την δεοξυριβόζη. Στα ευκαρυωτικά κύτταρα ανιχνεύεται κυρίως μέσα στον πυρήνα του κυττάρου αλλά και σε μερικά άλλα οργανίδια, όπως τα μιτοχόνδρια και τα πλαστίδια, επιτρέποντάς τους να αναπαράγονται αυτόνομα (ημιαυτόνομα οργανίδια).

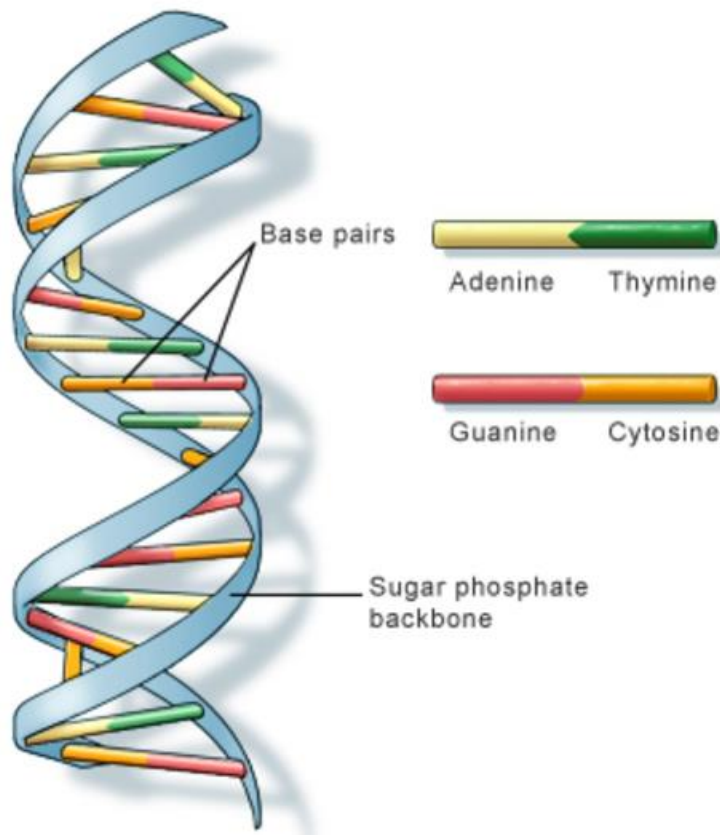
Το σύνολο των μορίων DNA που υπάρχουν σε ένα κύτταρο αποτελούν το γενετικό υλικό του. Το DNA είναι ο φορέας των γενετικών πληροφοριών του κυττάρου, όχι μόνο με την έννοια της μεταβίβασης χαρακτηριστικών, αναλλοίωτων από γενεά σε

γενεά, αλλά και της ρύθμισης της φυσιολογικής εξειδίκευσης κάθε κυττάρου για την επιτέλεση των ιδιαίτερων λειτουργιών του. Τέλος, το DNA επιτρέπει τη δημιουργία γενετικής ποικιλότητας, υφιστάμενο μεταλλάξεις.

Η διαμόρφωση των μεγάλων μορίων του DNA στο χώρο έχει τη μορφή δύο επιμήκων αλύσεων, οι οποίες συστρέφονται ελικοειδώς μεταξύ τους. Οι αζωτούχες βάσεις στο DNA είναι τέσσερις:

- κυτοσίνη C
- γουανίνη G
- θυμίνη T
- αδενίνη A

Οι αζωτούχες βάσεις, ανάλογα με την σειρά αλληλουχίας τους σε τριάδες, κωδικοποιούν το μήνυμα για τη σύνθεση των αμινοξέων του κυττάρου στα ριβοσώματα. Εκεί τα αμινοξέα συνδυάζονται, με τη σειρά κατά την οποία μεταφέρθηκαν στο ριβόσωμα και συντίθενται έτσι οι διαφορετικές πρωτεΐνες.



Εικόνα 1 Δομή του DNA

RNA

Το RNA (ή αλλιώς ριβονουκλεϊκό οξύ) είναι μία από τις δύο κατηγορίες των πολυμερών νουκλεϊκών οξέων στο κύτταρο.

Αποτελείται από μονομερή νουκλεοτίδια που παίζουν σημαντικό ρόλο στη διαδικασία της μετάφρασης του γενετικού κώδικα από την έτερη κατηγορία νουκλεϊκού οξέος, το δεοξυριβονουκλεϊκό οξύ (DNA).

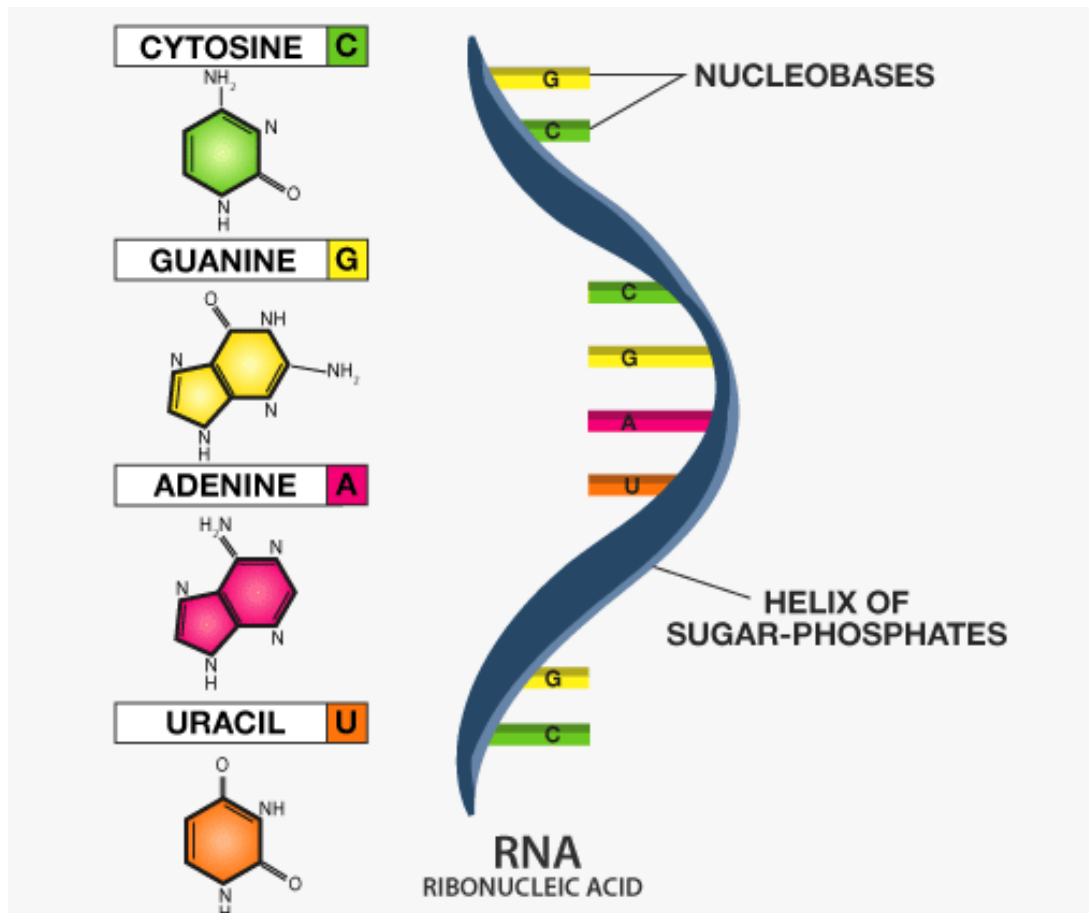
Το RNA χαρακτηρίζεται ως ο «αγγελιοφόρος» μεταξύ του DNA και των πρωτεϊνικών συμπλεγμάτων που είναι γνωστά σαν ριβοσώματα στο κυτταρόπλασμα του κυττάρου (αγγελιαφόρο RNA, mRNA). Έτσι το RNA μαζί με το DNA αποτελούν το γενετικό υλικό των οργανισμών.

Απο χημικής άποψης το RNA είναι όμοιο με το DNA. Και οι δύο αυτές κατηγορίες νουκλεϊκών οξέων είναι μακρομοριακές ενώσεις μεγάλου μοριακού βάρους.

Το μακρομόριο του RNA αποτελείται από επαναλαμβανόμενες δομικές μονάδες τα νουκλεοτίδια. Το μόριο του RNA περιλαμβάνει (όπως και του DNA), τέσσερις τύπους νουκλεοτιδίων που συνδέονται μεταξύ τους με 3'-5' φωσφοδιεστερικούς δεσμούς. Ωστόσο κύρια διαφορά του RNA από το DNA είναι ότι το μόριό του είναι μονόκλωνο έναντι του δίκλωνου του DNA, αποτελείται δηλαδή από μια μόνο αλυσίδα, ανάλογη της μιας εκ των δύο εκείνων της διπλής έλικας του DNA.

Βασική επίσης διαφορά είναι ότι το σάκχαρο στα νουκλεοτίδιά του είναι η ριβόζη, εξ ου και η ονομασία τους ριβονουκλεοτίδια, αντί της δεοξυριβόζης στο DNA, και ότι περιέχει την πυριμιδίνη ουρακίλη αντί της θυμίνης (που υπάρχει στο μόριο του DNA), χωρίς να είναι γνωστός ο λόγος της τελευταίας αυτής διαφοράς. Η μακρομοριακή πολυνουκλεοτιδική αλυσίδα του RNA εμφανίζεται από ελικοειδής μέχρι ευθύγραμμη.

Ένας άλλος τύπος RNA πέραν του mRNA, το μεταφορικό RNA (tRNA), αποτελεί δομικό συστατικό των ριβοσωμάτων και επιτελεί σημαντικές μεταφορές αμινοξέων που προορίζονται για την πρωτεϊνοσύνθεση. Το mRNA μεταγράφει τον γενετικό κώδικα από το DNA με τη βοήθεια κυρίως ενός ενζύμου που ονομάζεται RNA πολυμεράση, ενώ στη συνέχεια υφίσταται επεξεργασία από έναν αριθμό άλλων, δευτερευόντων ενζύμων. Ακολούθως το τελικό mRNA χρησιμοποιείται ως βάση για τη μετάφραση των γονιδίων σε πρωτεΐνες, με τη βοήθεια του tRNA το οποίο μεταφέρει αμινοξέα στα ριβοσώματα για να δημιουργηθούν πρωτεΐνες (πάντα χάρη στην αρχή της συμπληρωματικότητας των βάσεων).



Εικόνα 2 Δομή του RNA

Υπάρχουν διαφορετικά είδη RNA που εκτελούν διαφορετικές λειτουργίες όπως είναι:

- Αγγελιοφόρο RNA mRNA (Messenger RNA)
- Μεταφορικό RNA tRNA (Transfer RNA)
- Ριβοσωμικό RNA rRNA (Ribosomal RNA)
- Μη κωδικοποιητικό RNA (Non-coding RNA)

Το αγγελιοφόρο RNA (mRNA) είναι ένα μονόκλωνο μόριο RNA που αντιστοιχεί στη γενετική αλληλουχία ενός γονιδίου και διαβάζεται από ένα ριβόσωμα στη διαδικασία σύνθεσης μιας πρωτεΐνης.

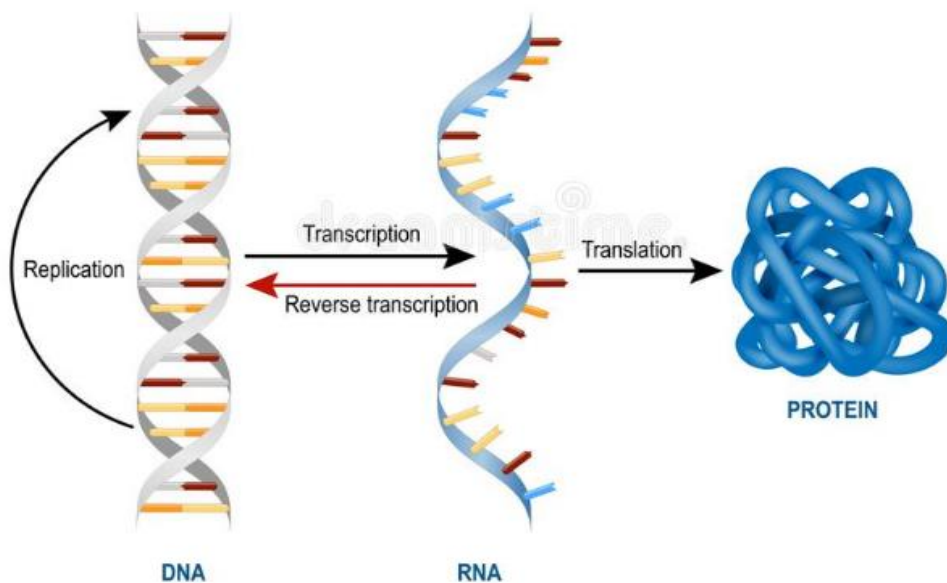
Το μεταφορικό RNA (tRNA) είναι μικρή αλυσίδα RNA, με μήκος 74-95 νουκλεοτιδίων, που μεταφέρει ειδικά αμινοξέα σε μια επεκτεινόμενη πολυπεπτιδική αλυσίδα στα ριβοσώματα του κυττάρου, με βάση τις οδηγίες του αγγελιαφόρου RNA. Έτσι γίνεται η πρωτεϊνοσύνθεση κατά τη διάρκεια της μετάφρασης στο κύτταρο. Είναι ένας τύπος μη κωδικοποιητικού RNA.

Το tRNA διαθέτει ειδικούς υποδοχείς για την πρόσδεση αμινοξέων καθώς και μια περιοχή αντικωδονίου, για την αναγνώριση του τρέχοντος κωδικονίου στο

αγγελιαφόρο RNA. Η αναγνώριση αυτή γίνεται με βάση τη συμπληρωματικότητα των βάσεων. Ένας τύπος tRNA αντιστοιχεί σε πολλά κωδικόνια αλλά μόνο σε ένα αμινοξύ.

Το ριβοσωμικό RNA (rRNA) είναι ένας τύπος RNA των ριβοσωμάτων που καταλύει την πρωτεϊνοσύνθεση στο κύτταρο. Τα ευκαρυωτικά ριβοσώματα αποτελούν το εργοστάσιο πρωτεϊνοσύνθεσης του κυττάρου και περιέχουν τέσσερα διαφορετικά μόρια rRNA: τα 18S, 5.8S, 28S, και 5S rRNA. Τρία από αυτά τα μόρια rRNA συντίθενται στον πυρήνισκο του κυττάρου. Τα διαφορετικά είδη rRNA είναι πολυάριθμα στο κύτταρο και αποτελούν το 80% των ολικών RNA σε ένα τυπικό ευκαρυωτικό κύτταρο. Στο κυτταρόπλασμα το ριβοσωμικό RNA συνδυάζεται με ειδικές πρωτεΐνες του πλάσματος και συνθέτουν ένα νουκλεοπρωτεϊνικό σύμπλεγμα που ονομάζεται ριβόσωμα. Το ριβόσωμα συνδέεται με το mRNA και είναι υπεύθυνο για την πρωτεϊνική σύνθεση, όπου και λαμβάνει χώρα. Ορισμένα ριβοσώματα μπορούν να είναι συνδεδεμένα με μία μονή αλυσίδα mRNA σε όλη τη διάρκεια της ζωής τους.

Το μη κωδικοποιητικό RNA (ncRNA) είναι ένα μόριο RNA που δεν μεταφράζεται σε πρωτεΐνη, όπως το αγγελιαφόρο RNA (mRNA), αλλά έχει κάποιον άλλο λειτουργικό ρόλο στο κύτταρο. Τα πιο αντιπροσωπευτικά είδη μη κωδικοποιητικού RNA είναι το μεταφορικό RNA (tRNA) και το ριβοσωμικό RNA (rRNA), που παίζουν σημαντικό υποστηρικτικό ρόλο στην μετάφραση. Η περιοχή του γενετικού κώδικα όπου κωδικοποιείται ένα μη κωδικοποιητικό RNA (αντί για μία πρωτεΐνη), ονομάζεται «γονίδιο RNA».



Εικόνα 3 Διαδικασία Μετάδοση Πληροφορίας

Πρωτεΐνη

Οι πρωτεΐνες αποτελούν τα πιο διαδεδομένα και πολυδιάστατα, τόσο στη μορφή όσο και στη λειτουργία τους, μακρομόρια. Ακόμη και σ' ένα απλό κύτταρο των βακτηρίων εντοπίζονται εκατοντάδες διαφορετικές πρωτεΐνες με την καθεμία εξ αυτών να έχει ιδιαίτερο ρόλο. Οι πρωτεΐνες αποτελούν είτε δομικά συστατικά των μεμβρανών του κυττάρου, είτε συνεργούν σε κάποια συγκεκριμένη λειτουργία, όπως η δημιουργία πρωτεϊνικών συμπλόκων.

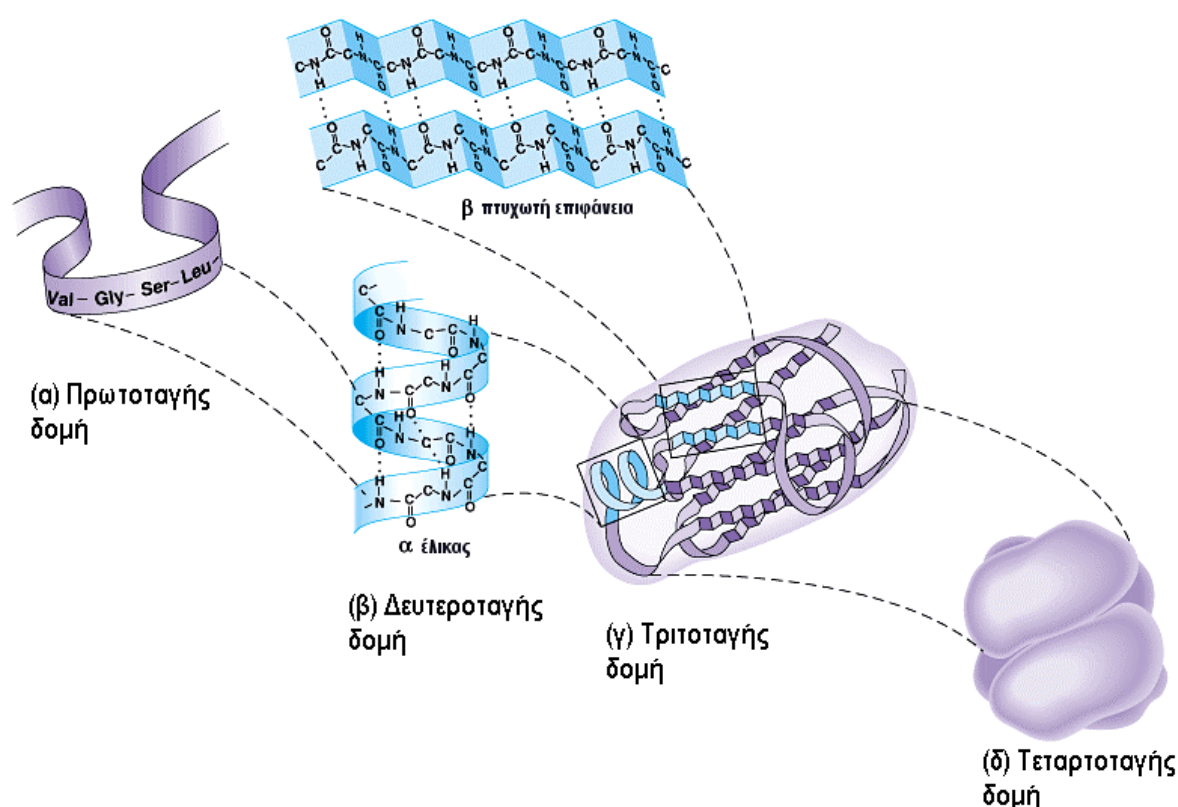
Είναι μεγάλα σύνθετα βιομόρια, με μοριακό βάρος από 10.000 μέχρι πάνω από 1 εκατομμύριο, αποτελούμενα από αμινοξέα, τα οποία ενώνονται μεταξύ τους με πεπτιδικούς δεσμούς σχηματίζοντας μια γραμμική αλυσίδα, καλούμενη αλυσίδα πολυπεπτιδίων. Όλες οι πρωτεΐνες περιέχουν άνθρακα, οξυγόνο και άζωτο και οι περισσότερες εξ αυτών και θείο και υδρογόνο.

Οι πρωτεΐνες επιτελούν σημαντικές βιολογικές λειτουργίες στους οργανισμούς. Για να συμβεί αυτό θα πρέπει να έχουν την κατάλληλη διαμόρφωση, δομή. Η δομή των πρωτεϊνών περιγράφεται από τέσσερα διαφορετικά επίπεδα δομών: την πρωτοταγή δομή, τη δευτεροταγή δομή, την τριτοταγή δομή και την τεταρτοταγή δομή. Οι δευτεροταγείς, τριτοταγείς και τεταρτοταγείς δομές ονομάζονται ανώτερες δομές της πρωτεΐνης.

Η πρωτοταγής δομή είναι η γραμμική αλληλουχία αμινοξέων σε ένα μόριο πρωτεΐνης. Η δευτεροταγής δομή είναι τα αμινοξέα, τα οποία είναι δίπλα το ένα στο

άλλο στην αλυσίδα, αλληλεπιδρούν (αλληλεπιδράσεις δεσμών υδρογόνου) με πολλαπλούς τρόπους: αυτό έχει σαν συνέπεια το ένα μέρος του γραμμικού πολυπεπτιδίου να γίνει ελικοειδές, σχηματίζοντας μια α-έλικα (α-helix) και το άλλο μέρος να σχηματίσει ένα πλέγμα με γραμμές που έχουν μία κατεύθυνση (β-sheet).

Η τριτοταγής δομή είναι η ολική δίπλωση μιας πρωτεϊνικής αλληλουχίας, η οποία σχηματίζεται από το πακετάρισμα των δευτεροταγών (δισουλφιδικοί θειούχοι δεσμοί ανάμεσα σε αμινοξέα που περιέχουν θείο σταθεροποιούν το μόριο στην τρισδιάστατη δομή του). Η τεταρτοταγής δομή είναι η διάταξη διαφορετικών πολυπεπτιδικών αλυσίδων (υποομάδες-subunits) σε ένα μόριο πρωτεΐνης.



Εικόνα 4 Κατηγορίες Δομών Πρωτεΐνης

Πρωτεϊνική αναδίπλωση

Πρωτεϊνική αναδίπλωση είναι η διεργασία κατά την οποία μια πολυπεπτιδική αλυσίδα αναδιπλώνεται στη σωστή τριτοταγή δομή, έτσι ώστε να καταλήξει στη βιολογικά ενεργή, φυσική κατάσταση (native state). Οι πρωτεΐνες αποτελούν θεμελιώδη μόρια για τους οργανισμούς και ο ρόλος τους είναι είτε δομικός είτε λειτουργικός.

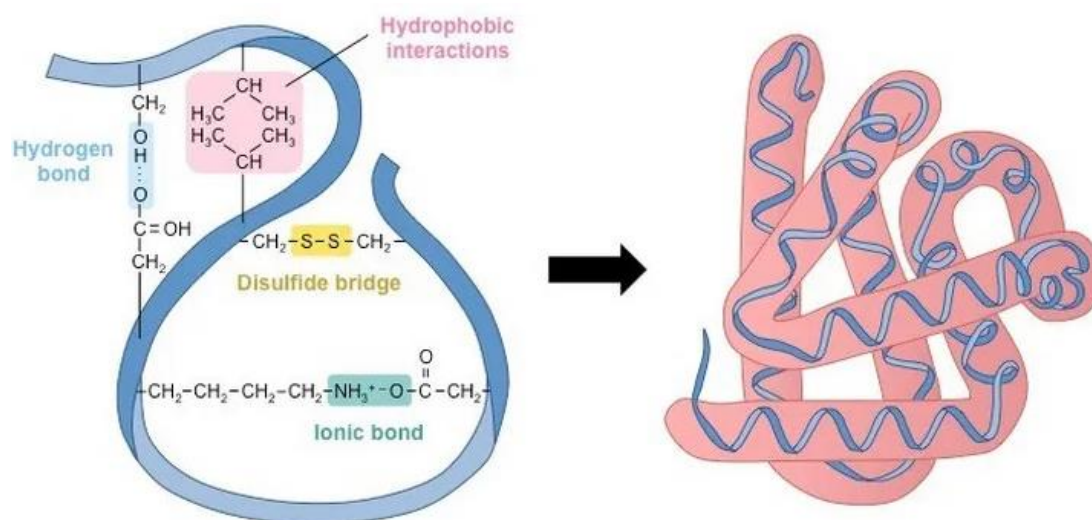
Το πολυπεπτίδιο αυτό δεν έχει ανεπτυγμένη τρισδιάστατη δομή. Παρόλα αυτά κάθε αμινοξύ της αλυσίδας θεωρείται ότι έχει συγκεκριμένα χημικά χαρακτηριστικά. Αυτά αλληλεπιδρούν μεταξύ τους καθώς και με τον περιβάλλοντα χώρο τους ώστε

να σχηματιστεί ένα καλώς ορισμένο τρισδιάστατο σχήμα, η αναδιπλωμένη πρωτεΐνη δηλαδή τελική στερεοδιάταξη.

Οι πρωτεΐνες αποτελούνται από αμινοξέα τα οποία ενώνονται με πεπτιδικούς δεσμούς, σχηματίζοντας κατ' αυτόν τον τρόπο μια γραμμική αλυσίδα. Η σωστή αλληλουχία των αμινοξέων αλλά και η σωστή τριτοταγής δομή απαιτούνται για να είναι μια πρωτεΐνη λειτουργική.

Η πρωτεΐνη θα αναδιπλωθεί αυθόρμητα κατά την διάρκεια ή μετά την σύνθεση. Ο μηχανισμός αναδίπλωσης εξαρτάται ισοδύναμα από χαρακτηριστικά όπως η φύση του διαλύματος, η συγκέντρωση των αλάτων, η θερμοκρασία και άλλα.

Η αναδίπλωση είναι μια αυθόρμητη διαδικασία ανεξάρτητη από ενεργειακές εισόδους. Οι δυνάμεις που οδηγούν στην αναδίπλωση είναι οι μη ομοιοπολικές αλληλεπιδράσεις μεταξύ ατόμων της πολυπεπτιδικής αλυσίδας. Ιδιαίτερη σημασία έχουν οι αλληλεπιδράσεις μεταξύ ατόμων που ανήκουν σε αμινοξέα που βρίσκονται μακριά το ένα από το άλλο κατά μήκος της αλυσίδας. Χάρη σ' αυτές, επιτυγχάνεται το δίπλωμα σε συμπαγή και σταθερή στερεοδομή.



Εικόνα 5 Πρωτεϊνική Αναδίπλωση

Αλγόριθμος

Με τον όρο αλγόριθμο ονομάζουμε μία σειρά από εντολές που έχουν αρχή και τέλος, είναι σαφείς και έχουν ως σκοπό την επίλυση κάποιου προβλήματος.

Οι αλγόριθμοι είναι σημαντικοί γιατί σχετίζονται άμεσα με τον τρόπο με τον οποίο οι υπολογιστές επεξεργάζονται δεδομένα και παράγουν πληροφορίες. Ένα πρόγραμμα υπολογιστών είναι ουσιαστικά ένας αλγόριθμος που λέει στον υπολογιστή ποια συγκεκριμένα βήματα να εκτελέσει (σε ποια συγκεκριμένη σειρά)

προκειμένου να επιτευχθεί ένας συγκεκριμένος στόχος.



Εικόνα 6 Απεικόνιση Έννοιας Αλγορίθμου

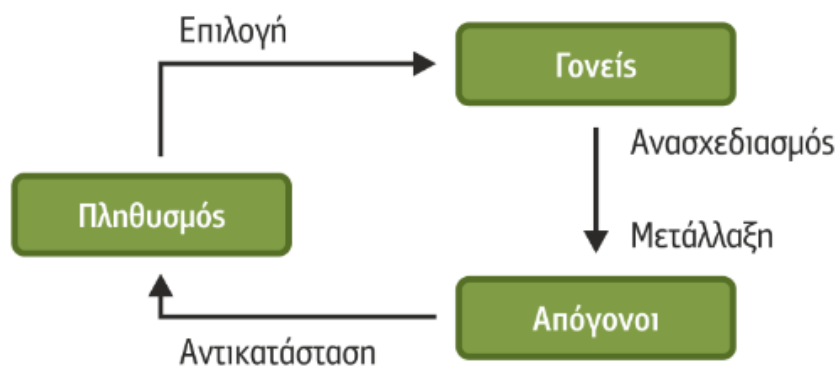
Όμως στο κομμάτι της πληροφορικής που ασχολείται με την μοριακή βιολογία εφαρμόζεται διαφορετικό είδος αλγορίθμων. Η διαδικασία που πρέπει να ακολουθούν είναι η ίδια όπως κάθε είδος αλγόριθμου. Οι αλγόριθμοι αυτοί ονομάζονται γενετικοί αλγόριθμοι.

Γενετικοί Αλγόριθμοι

Οι γενετικοί αλγόριθμοι είναι στοχαστικοί αλγόριθμοι με αναζήτηση γενικού σκοπού που καταφέρνουν μια αξιοσημείωτη ισορροπία μεταξύ της εξερεύνησης βέλτιστης λύσης και την στοχευμένη αναζήτηση στον ευρύτερο χώρο των εφικτών λύσεων του προβλήματος.

Οι αλγόριθμοι αυτοί χρησιμοποιούν την ιδέα της εξέλιξης μέσω γενετικής μετάλλαξης, φυσικής επιλογής και διασταύρωσης.

Οι Γενετικοί Αλγόριθμοι είναι αρκετά απλοί στην υλοποίησή τους. Οι τιμές για τις παραμέτρους του συστήματος πρέπει να κωδικοποιούνται με τρόπο ώστε να αναπαρασταθούν από μια μεταβλητή που περιέχει σειρά χαρακτήρων ή δυαδικών ψηφίων. Αυτή η μεταβλητή μιμείται το γενετικό κώδικα που υπάρχει στους ζωντανούς οργανισμούς.



Εικόνα 7 Απεικόνιση έννοιας Γενετικού Αλγόριθμου

Οι Γενετικοί Αλγόριθμοι διατηρούν έναν πληθυσμό πιθανών λύσεων, του προβλήματος που μας ενδιαφέρει, πάνω στον οποίο δουλεύουν, σε αντίθεση με άλλες μεθόδους αναζήτησης που επεξεργάζονται ένα μόνο σημείο του διαστήματος αναζήτησης.

Έτσι ένας Γενετικός Αλγόριθμος πραγματοποιεί αναζήτηση σε πολλές κατευθύνσεις και υποστηρίζει καταγραφή και ανταλλαγή πληροφοριών μεταξύ αυτών των κατευθύνσεων.

Διαδικασία Εύρεσης

Σε αυτό το κεφάλαιο θα γίνει εκτενής αναφορά στους διαθέσιμους τρόπους εύρεσης-πρόγνωσης πρωτεϊνικής αναδίπλωσης με υπολογιστικούς αλγόριθμους αλλά και στα δεδομένα τα οποία χρειάζονται για την πρόγνωση.

Δεδομένα

Τα δεδομένα που χρησιμοποιούνται για εύρεση-πρόγνωση πρωτεϊνικής αναδίπλωσης είναι αλληλουχίες. Για την εύρεση της αλληλουχίας μπορεί να γίνει και επιβάλλεται η χρήση των βιολογικών βάσεων δεδομένων. Αυτές οι βάσεις διαθέτουν την πληροφορία και τη δομή της αλληλουχίας ταξινομημένη ώστε να είναι οργανωμένη κατά την εξαγωγή των συμπερασμάτων σε βιολογικό επίπεδο. Οι βιολογικές βάσεις δεδομένων που ασχολούνται με τις πρωτεΐνες είναι:

- Βιολογικές βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών
- Βιολογικές βάσεις δεδομένων πρωτεϊνικών αλληλουχιών
- Βιολογικές βάσεις δεδομένων τρισδιάστατων βιολογικών δομών

Οι βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών έχουν να κάνουν με τις αλληλουχίες κυρίως του DNA αλλά και του RNA. Οι βάσεις δεδομένων πρωτεϊνικών αλληλουχιών έχουν να κάνουν με την δομή των πρωτεϊνών που παρουσιάζουν

μεγάλη ποικιλομορφία αλλά και στην λειτουργία τους. Οι βάσεις δεδομένων τρισδιάστατων βιολογικών δομών περιέχουν δεδομένα που έχουν να κάνουν με την τρισδιάστατη δομή βιολογικών μακρομορίων και αποτελούν το τελικό στάδιο της διαδικασίας που έγινε η χρήση μοριακών τεχνικών έχοντας σκοπό στην υπολογιστική επίλυση και ανάλυση των δομών των πρωτεϊνών.

Το μεγαλύτερο ποσοστό των βάσεων δεδομένων βρίσκεται στο Εθνικό Κέντρο Πληροφοριών Βιοτεχνολογίας (<https://www.ncbi.nlm.nih.gov/>).

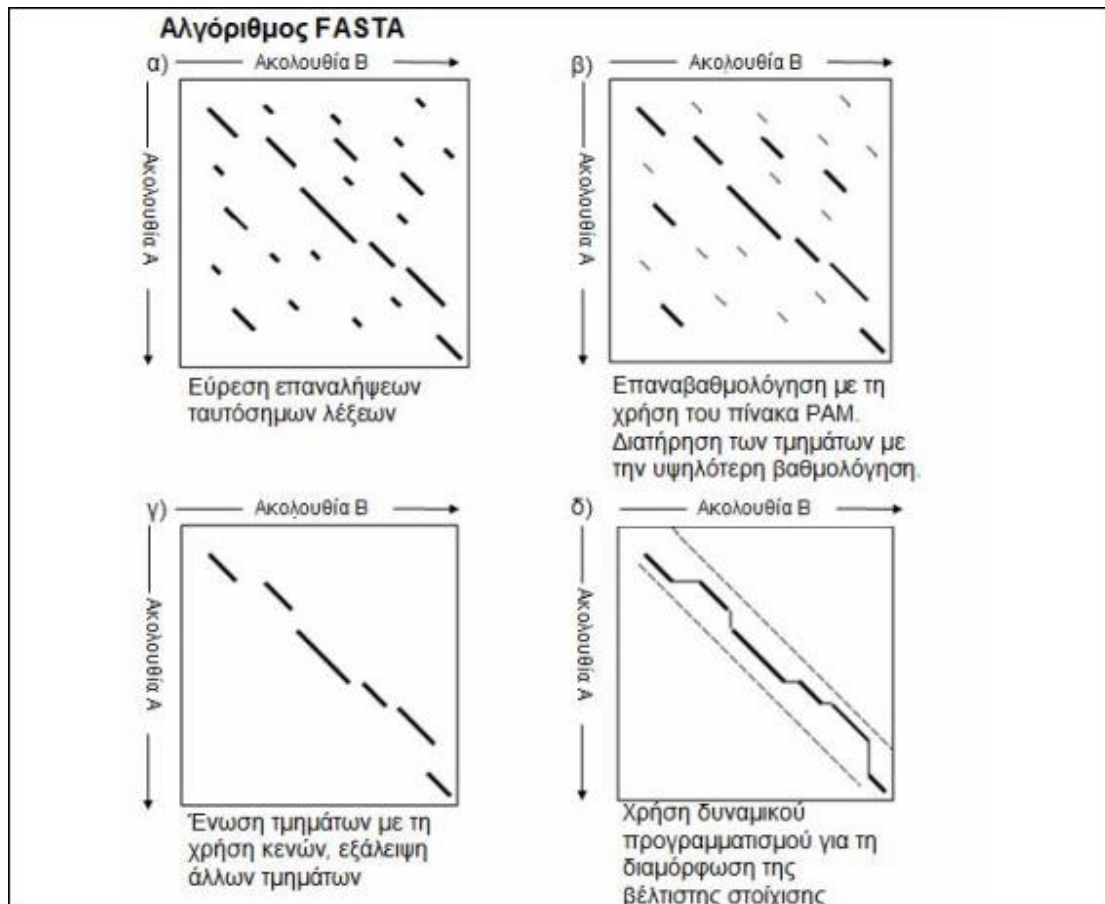
Ευριστικοί Αλγόριθμοι

Όμως για να γίνει η αναζήτηση της σωστής αλληλουχίας μέσω των βιολογικών βάσεων δεδομένων θα πρέπει να γίνει χρήση ευριστικών γενετικών αλγορίθμων που αποτελούν βασικό εργαλείο στην υπολογιστική ανάλυση. Η βασική απαίτηση που πρέπει να υπάρχει σε έναν ευριστικό αλγόριθμο είναι να αποδίδει το ίδιο καλά με έναν μαθηματικό αλγόριθμο αλλά και να εκτελεί πολλές φορές και με γρήγορο τρόπο τις αναλύσεις του. Υπάρχουν 2 τέτοιοι αλγόριθμοι που ασχολούνται με αναζήτηση σε βιολογικές βάσεις δεδομένων για την επίτευξη της σύγκρισης των αλληλουχιών. Οι 2 αλγόριθμοι είναι :

- FASTA
- BLAST

Ο αλγόριθμος FASTA εντοπίζει κατά προσέγγιση τη διαγώνιο γύρω από την οποία βρίσκεται η στοίχιση με σκοπό των περιορισμό του εύρους της αναζήτησης. Η διαδικασία που χρησιμοποιεί ο αλγόριθμος είναι η εξής:

1. Στην αρχή γίνεται η δημιουργία ενός ευρετηρίου με τις θέσεις όλων των λέξεων με μέγεθος k και το τυπικό μήκος για αμινοξικές αλληλουχίες είναι το 1 ή το 2 που υπάρχουν ταυτόχρονα στις 2 αλληλουχίες
2. Από τη διαφορά των θέσεων τους στις 2 αλληλουχίες εντοπίζεται η διαγώνιος στην οποία βρίσκονται
3. Ακολούθως , αυτές οι περιοχές ταύτισης συνενώνονται επιτρέποντας την εισαγωγή κενών με τον υπολογισμό της αντίστοιχης ποινής
4. Τέλος πραγματοποιείται η διαδικασία πλήρους δυναμικού προγραμματισμού περιορισμένου όμως μόνο σε ζώνη γύρω από τις συγκεκριμένες διαγώνιους.

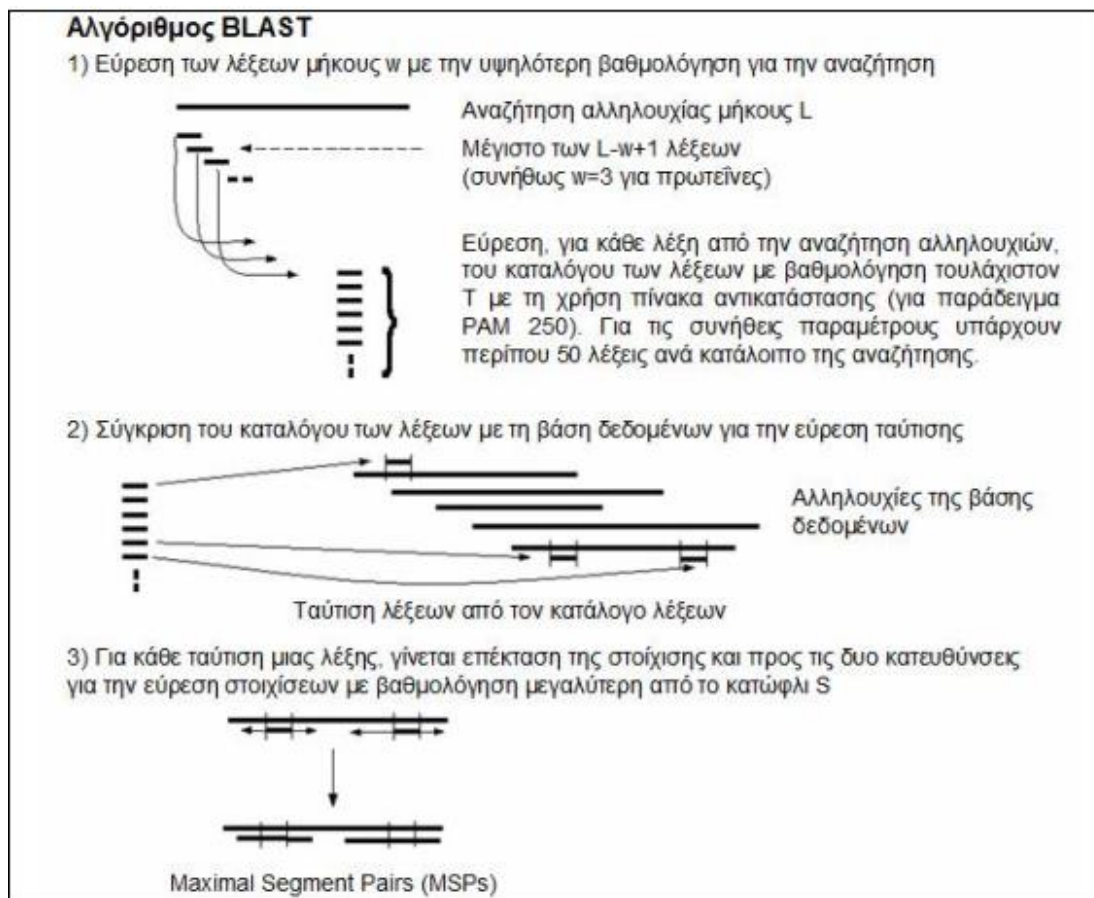


Εικόνα 8 Απεικόνιση Αλγόριθμου FASTA

Ο αλγόριθμος BLAST μοιάζει στα αρχικά στάδια με αυτή του FASTA αλλά είναι ακόμα πιο γρήγορος καθώς οι περισσότερες παραμέτρους είναι ήδη προϋπολογισμένες, έτσι ώστε να γίνει η αποφυγή της στοίχισης αλληλουχίας με αυτή της βάσης δεδομένων που ο αλγόριθμος κρίνει ότι δεν έχουν σημαντική ομοιότητα:

1. Η διαδικασία σύγκρισης ξεκινά από την κατασκευή ενός καταλόγου όλων των λέξεων που θα γινόταν το ταίριασμα με κάποια λέξη της άγνωστης αλληλουχίας και ξεπερνούν την τιμή κατωφλίου(με προκαθορισμένη τιμή στις πρωτεϊνικές αλληλουχίες)
2. Στη συνέχεια ο αλγόριθμος αναζητά αυτές τις λέξεις στις αλληλουχίες της βάσης δεδομένων και κάθε φορά που εντοπίζει κάποια, ξεκινάει η επέκταση του ευρήματος προς τις 2 κατευθύνσεις, όσο η βαθμολογία συνεχίζει και αυξάνει
3. Οι περιοχές μέγιστης βαθμολογίας που εντοπίζονται σε αυτό το στάδιο είναι οι υποψήφιες περιοχές ομοιότητας(high scoring pairs)
4. Απ' όλα τα high scoring pairs αναφέρονται στα αποτελέσματα εκείνες οι περιοχές στις οποίες η βαθμολογία υπερβαίνει μια δεύτερη τιμή κατωφλίου

5. Τέλος γίνεται η επιλογή μόνο των τοπικών ομοιοτήτων οι οποίες εμφανίζουν υψηλή στατιστική σημαντικότητα.



Εικόνα 9 Απεικόνιση αλγόριθμου BLAST

Κωδικοποίηση

Μετά την εκτέλεση των ευρεστικών αλγορίθμων για τη σωστή αλληλουχία, πρέπει να γίνει η κωδικοποίηση των αλληλουχιών και ανάλογα με την κωδικοποίηση μπορεί να γίνει και η πρόγνωση.

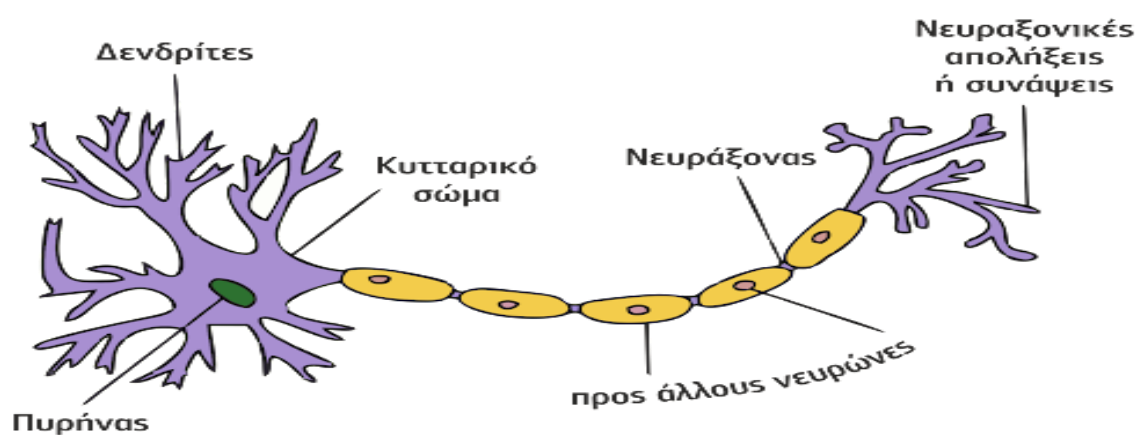
Για την κωδικοποίηση των αλληλουχιών υπάρχουν 2 τρόποι που μπορούν να πραγματοποιηθούν. Στην πρώτη κατηγορία κωδικοποίησης ενδιαφερόμαστε για τοπική πρόγνωση κατά το μήκος της αλληλουχίας. Στην δεύτερη κατηγορία ενδιαφερόμαστε για ταξινόμηση σε κατηγορίες των αλληλουχιών εύρεση-πρόγνωση της πρωτεϊνικής αναδίπλωσης. Και με τις 2 μεθόδους θα πρέπει να κωδικοποιηθούν σε επίπεδο υπολογιστικής ανάλυσης.

Αφού γίνει η κωδικοποίηση των αλληλουχιών έρχεται το κομμάτι της εύρεσης-πρόγνωσης. Στο κομμάτι της εύρεσης-πρόγνωσης θα πρέπει να γίνει χρήση τεχνικών μηχανικής μάθησης και συγκεκριμένα με νευρωνικά δίκτυα.

Νευρωνικά Δίκτυα

Τα Νευρωνικά Δίκτυα ή ΝΔ (Neural Networks ή NN), χωρίζονται σε Βιολογικά Νευρωνικά Δίκτυα (ΒΝΔ) ή Biological Neural Networks (BNN) και σε Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ) ή Artificial Neural Networks (ANN).

Τα Βιολογικά Νευρωνικά Δίκτυα αποτελούν μέρος του κεντρικού νευρικού συστήματος βιολογικών συστημάτων, παραδείγματος χάριν του ανθρώπου, ενώ συνίστανται από βιολογικό ιστό, χημικές ουσίες και ηλεκτρικά σήματα τα οποία τα διαχωρίζουν από τα τεχνητά νευρωνικά δίκτυα που προσπαθούν να μιμηθούν μέσα από ένα σύνολο ηλεκτρονικών και μηχανικών συστημάτων συνοδευόμενα από ευφυείς αλγορίθμους.



Εικόνα 10 Βιολογικός Νεύρωνας

Κάθε βιολογικός νευρώνας αποτελείται από 3 κύρια τμήματα:

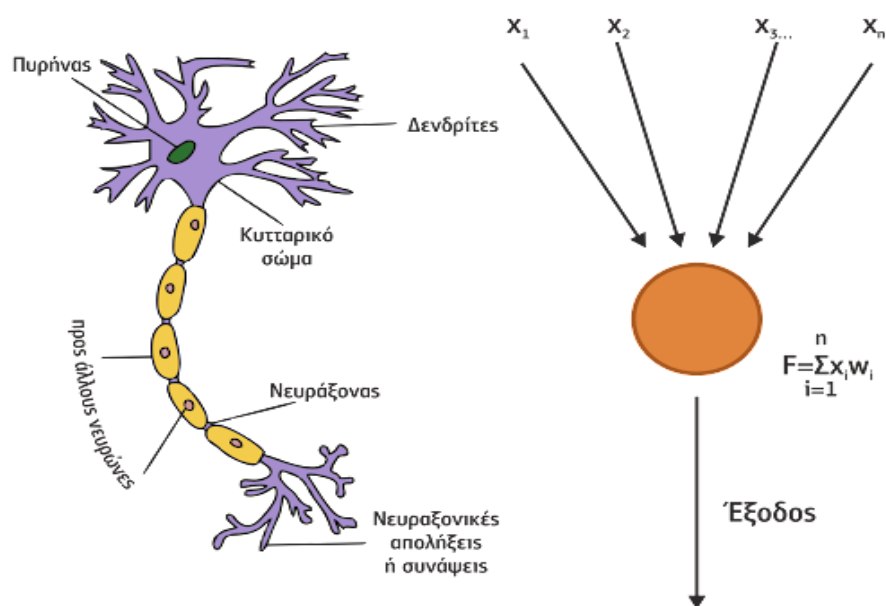
- τους δενδρίτες (dendrites), οι οποίοι λειτουργούν ως κανάλια εισόδου για το νευρώνα,
- το κυρίως κυτταρικό σώμα (cell body),
- τον άξονα του κυττάρου-νευράξονα (axon), που συνδέει ένα νευρώνα με άλλους νευρώνες.

Ο άξονας του ενός νευρώνα μεταφέρει σήματα στους δενδρίτες γειτονικών νευρώνων μέσω του σημείου ένωσης που ονομάζεται νευροαξονική απόληξη ή σύναψη (synapse). Ένας νευρώνας μπορεί να λάβει σήματα από ένα σύνολο γειτονικών νευρώνων μέσω των δενδριτών, να τα επεξεργαστεί και να τροφοδοτήσει την έξοδό του μέσω του άξονα προς ένα άλλο σύνολο γειτονικών νευρώνων. Τα σήματα που έρχονται μέσω των δενδριτών «ζυγίζονται» και τα αποτελέσματα αθροίζονται. Όταν το άθροισμα ξεπεράσει το οριακό επίπεδο (τιμή κατωφλίου), ο νευρώνας δημιουργεί μια έξοδο στον άξονά του, η οποία εν συνεχεία μέσω των συνάψεων θα μεταφερθεί στους γειτονικούς νευρώνες.

Τα σήματα που λαμβάνονται από ένα νευρώνα μεταβάλλονται από τα ηλεκτρικά χαρακτηριστικά των επαφών των συνάψεων, ώστε να εμποδίζονται μερικά και να επιτρέπεται σε άλλα να διαδοθούν. Τα ηλεκτρικά χαρακτηριστικά των συνάψεων αποτελούν κάποιο είδος πληροφορίας μοναδικής σε κάθε νευρώνα. Με αυτό τον τρόπο οι πληροφορίες που κρατούνται από ένα δίκτυο κατανέμονται στους νευρώνες του.

Η μεταβίβαση πληροφορίας γίνεται με βάση το δυναμικό ενέργειας που καθορίζεται όχι από τον τύπο του σήματος, αλλά από την οδό του εγκεφάλου μέσα από διακριτούς επικοινωνούντες νευρώνες από τους οποίους περνάει το σήμα.

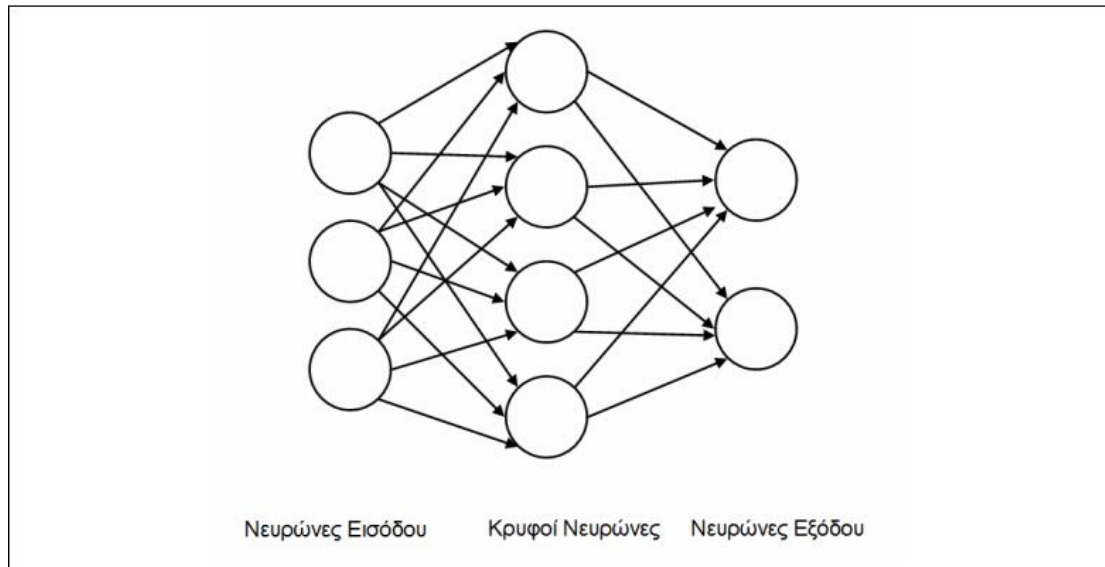
Τα τεχνητά νευρωνικά δίκτυα είναι μαθηματικές υπολογιστικές μηχανές που χρησιμοποιούνται για την επίτευξη γρηγορότερων λειτουργιών και προγνώσεων στο κομμάτι της μηχανικής μάθησης. Ο κάθε νευρώνας είναι απλά μια συνάρτηση που δέχεται μεταβλητές από άλλους νευρώνες και δίνει τελικά μεταβλητές (με βάση τη συνάρτηση αυτή) σε άλλους νευρώνες.



Εικόνα 11 Σύγκριση Βιολογικού Νεύρωνα και Τεχνητού

Συνήθως, τα δίκτυα αυτά αναπαριστούνται με ένα γράφο, με τα βέλη να αντιστοιχούν στις συνδέσεις (συνάψεις) μεταξύ των νευρώνων.

Πρακτικά, οι νευρώνες διαφοροποιούνται σε νευρώνες εισόδου στους οποίους κωδικοποιούνται οι μεταβλητές εισόδου, σε κρυφούς νευρώνες οι οποίοι δέχονται τα ερεθίσματα από τους νευρώνες εισόδου και στους νευρώνες εξόδου οι οποίοι δέχονται τα ερεθίσματα από τους κρυφούς νευρώνες και τελικά παράγουν το αποτέλεσμα του δικτύου.



Εικόνα 12 Απλό Τεχνητό Νευρωνικό Δίκτυο

Στην περίπτωση της πρόγνωσης της πρωτεϊνικής αναδίπλωσης το τεχνητό νευρωνικό δίκτυο δέχεται ως νευρώνες εισόδου μεταβλητές που έχουν προκύψει από μια κατάλληλη κωδικοποίηση μιας βιολογικής αλληλουχίας.

Βασικά στοιχεία της αρχιτεκτονικής των ΤΝΔ που πρέπει να καθοριστούν κατά τη δημιουργία τους είναι:

- Ο αριθμός των ενδιάμεσων κρυφών επιπέδων,
- Ο αριθμός των μονάδων (ή κόμβων) ανά επίπεδο,
- Ο τρόπος σύνδεσης των μονάδων μεταξύ τους,
- Η τιμή ενεργοποίησης (τιμή κατωφλίου),
- Η μορφή της συνάρτησης μετάβασης,
- Οι τιμές των αρχικών βαρών μεταξύ των μονάδων,
- Οι αλγόριθμοι (κανόνες εκπαίδευσης) που χρησιμοποιούνται, για να ενισχυθούν οι σύνδεσμοι μεταξύ των μονάδων κατά τη διαδικασία της εκπαίδευσης.

Αφού καθοριστεί η αρχιτεκτονική ενός ΤΝΔ, όσον αφορά τον τύπο τροφοδότησης και την εσωτερική του δομή, ακολουθούν δυο στάδια κρίσιμα για την ολοκλήρωσή του: εκπαίδευση (training) και ανάκληση (recall).

Κατά το στάδιο της εκπαίδευσης, τα συνδεσμικά βάρη του δικτύου «μαθαίνουν» προσαρμόζοντας τα βάρη τους.

Μετά την εκπαίδευση, ακολουθεί η φάση της ανάκλησης (δηλαδή του ελέγχου του δικτύου) με τη βοήθεια ενός νέου συνόλου δειγμάτων τα οποία δεν είχαν πάρει μέρος στη διαδικασία εκπαίδευσης.

Η εκπαίδευσή τους γίνεται βάσει αλγορίθμων που μπορεί να υιοθετούν μοντέλα μάθησης με επίβλεψη, χωρίς επίβλεψη ή ενισχυτικής μάθησης.

- Εκπαίδευση με επίβλεψη, όπου το νευρωνικό δίκτυο μαθαίνει να απεικονίζει δεδομένες εισόδους σε εξόδους εκ των προτέρων γνωστές (σύνολο εκπαίδευσης), με απώτερο στόχο τη γενίκευση της αναγνώρισης αυτής και για παρεμφερείς εισόδους στο μέλλον.
- Εκπαίδευση χωρίς επίβλεψη, όπου το νευρωνικό δίκτυο κατασκευάζει απεικονίσεις από μια αναπαράσταση σε μια άλλη.
- Ενισχυτική Εκπαίδευση, όπου ο αλγόριθμος μαθαίνει μια στρατηγική ενεργειών για δεδομένου τύπου παρατηρήσεις.

Η εκπαίδευση έχει ως βασικό στόχο να βρεθεί ένας τρόπος αλλαγής των συνδεσμικών βαρών που θα έχει ως αποτέλεσμα την αλλαγή της γενικής συμπεριφοράς του δικτύου με την αύξηση της ικανότητας του δικτύου να παρέχει στο μέλλον μία επιθυμητή έξοδο μετά από μία δεδομένη είσοδο. Όταν η επιθυμητή έξοδος είναι εκ των προτέρων γνωστή λέμε ότι το δίκτυο μαθαίνει με επίβλεψη (supervised learning), αλλιώς μαθαίνει χωρίς επίβλεψη (unsupervised learning).

Βασικό στοιχείο της αρχιτεκτονικής ενός ΤΝΔ είναι ο τρόπος ελέγχου της αλλαγής των βαρών κατά την εκπαίδευση, δηλαδή ο αλγόριθμος εκπαίδευσης (training algorithm) που υλοποιείται αποκλειστικά από το ίδιο το δίκτυο χωρίς εξωτερική επέμβαση.

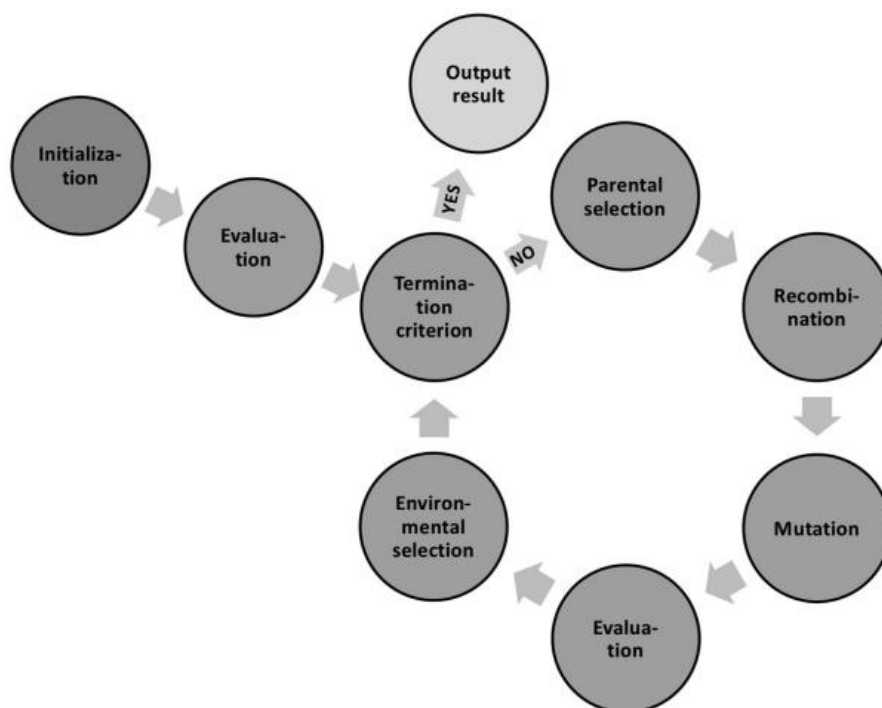
Στο επόμενο κεφάλαιο γίνεται χρήση τεχνητών νευρωνικών δικτύων με σκοπό την ανάλυση της αλληλουχίας για την πρόγνωση της πρωτεϊνικής αναδίπλωσης.

Σύγχρονοι Αλγόριθμοι

Σε αυτό το κεφάλαιο θα γίνει αναλυτική αναφορά σε 15 αλγόριθμους-εφαρμογές που έχουν δημοσιευτεί τα τελευταία χρόνια στο κομμάτι της εύρεσης παθογενούς πρωτεΐνης και μελέτης αλλαγής μορφής του RNA με χρήση μηχανικής μάθησης ή στοχαστικών μοντέλων ή γενετικών αλγορίθμων με χρήση δεδομένων αλληλουχίας.

Unsupervised feature selection algorithm

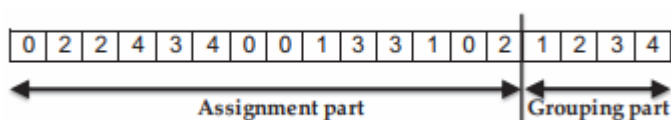
Το αντικείμενο της έρευνας του García-Díaz et al[26] είναι η χρήση ενός γενετικού αλγόριθμου για ομαδοποίηση με σκοπό την κατηγοριοποίηση των δεδομένων αλληλουχίας.



Εικόνα 13 Διαγραμματική Απεικόνιση του Γενετικού Αλγόριθμου

Ο συγκεκριμένος αλγόριθμος χρησιμοποιεί ειδικές στρατηγικές κωδικοποίησης με σκοπό την υψηλή απόδοση σε συγκεκριμένες μετρήσεις με ανάλογα προβλήματα.

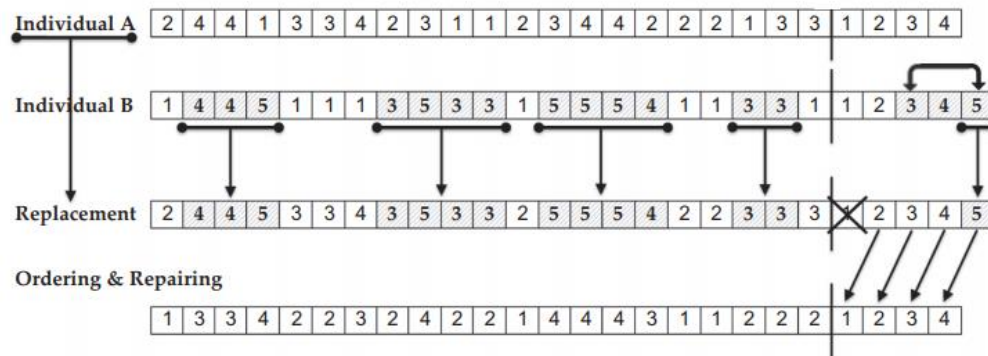
Αρχικά γίνεται η κωδικοποίηση σε πίνακα με στοιχεία τα οποία χωρίζονται σε 2 υπο-πίνακες. Το μήκος του τμήματος ανάθεσης συμπίπτει με τον αριθμό που πρέπει να ταξινομηθεί ενώ το μήκος του τμήματος ομαδοποίησης είναι ο αριθμός που έχει δημιουργηθεί για την κωδικοποιημένη λύση.



Εικόνα 14 Χωρισμός υπο-πινάκων

Στη συνέχεια ο αλγόριθμος εκτελεί την ίδια διαδικασία όπως ενός κανονικού γενετικού αλγορίθμου με την διαφορά ότι οι πράξεις του ανασυνδυασμού και της μετάλλαξης αποτελούν διαφορετικά χαρακτηριστικά σε σχέση με τον απλό γενετικό αλγόριθμο. Στο κομμάτι της μετάλλαξης ενός ατόμου η συγκεκριμένη λειτουργία μπορεί να θεωρηθεί και ως ανασυνδυασμός με άλλο άτομο. Σε αυτήν την περίπτωση

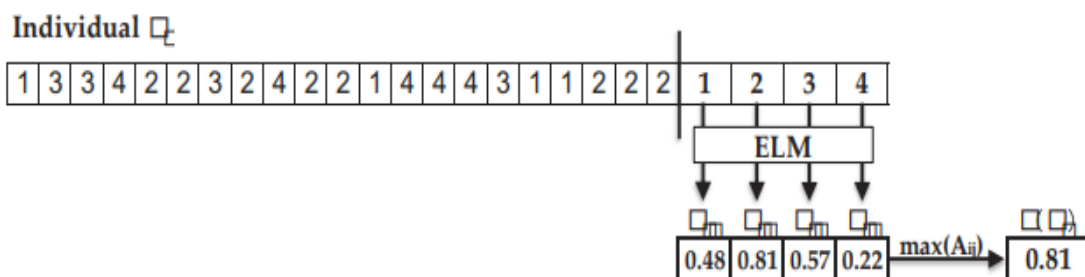
γίνεται η χρήση της crossover πράξης όπου γίνεται ένωση 2 σημείων-ατόμων με σκοπό την δημιουργία ενός νέου σημείου-ατόμου.



Εικόνα 15 Ένωση 2 σημείων για την δημιουργία νέου

Το επόμενο βήμα του αλγορίθμου είναι η ελαχιστοποίηση του σφάλματος εξόδου(output error). Εδώ γίνεται εμπλοκή του Extreme Learning Machine με σκοπό να υπολογιστεί η ακρίβεια(accuracy) ανάλογα με την επιλογή των δεδομένων.

Η ελαχιστοποίηση του σφάλματος θα έχει ως αποτέλεσμα την αύξηση της ακρίβειας. Για την αποτελεσματικότητά του θα πρέπει η μετρική της παλινδρόμησης να είναι αρκετά ακριβής και η διαδικασία της αξιολόγησης να είναι γρήγορη.



Εικόνα 16 Χρήση Extreme Learning Machine

Ο αλγόριθμος λαμβάνει ως δεδομένα αλληλουχίες από γονιδιακές εκφράσεις για ταξινόμηση του καρκίνου όπου αποτελούνταν από 5 διαφορετικούς τύπους όγκων στο σύνολο των δεδομένων:

1. BRCA (breast)
2. KIRC (kidney)
3. COAD (colon)
4. LUAD (lung)
5. PRAD (prostate)

Το μέγεθος των δεδομένων είναι 801 δείγματα και είναι σε 20.531 χαρακτηριστικά ή αλλιώς γονίδια.

Για την εκτέλεση του αλγορίθμου έπρεπε να γίνει ο διαχωρισμός των δεδομένων σε:

- 80% training data size
- 10% testing data size
- 10% validation data size

και ο ορισμός των παρακάτω παραμέτρων:

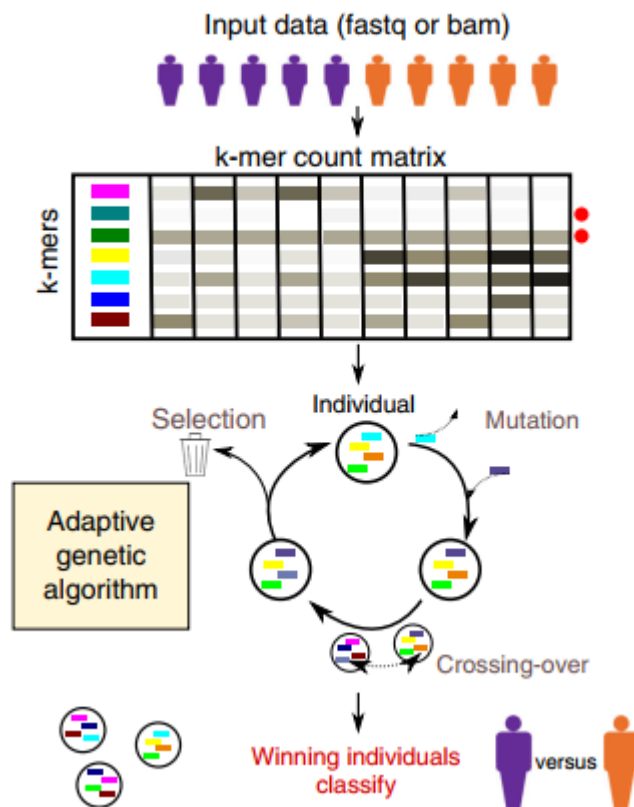
- Number of generations(G_{max}) = 60
- Population size (N_{ind}) = 50
- Offspring size = 25
- Mutation probability = 0.1
- Number of Neurons = 5
- Number of simulations = 3

με αποτέλεσμα ο μέσος όρος της ακρίβειας να είναι μεγαλύτερος από 90%.

GECKO algorithm

Το αντικείμενο της έρευνας του Thomas, Aubin, et al.[20] είναι η παρουσίαση ενός γενετικού αλγορίθμου και επιλογή χαρακτηριστικών για διεξοδική διερεύνηση μεγάλου όγκου αλληλουχίας δεδομένων για ταξινόμηση και ανακάλυψη νέων αλληλουχιών χωρίς εξειδικευμένων pipelines βιοπληροφορικής.

Ο συγκεκριμένος αλγόριθμος ασχολείται με την γενετική ταξινόμηση χρησιμοποιώντας k-mer βελτιστοποίηση διότι είναι σχεδιασμένος για υψηλής απόδοσης δεδομένα αλληλουχιών. Βασίζεται στην αποσύνθεση του βελτιστοποιητή k-mer σε συνδυασμό με έναν γενετικό αλγόριθμο που ασχολείται με υψηλής απόδοσης δεδομένα αλληλουχιών από 2 ή περισσότερες συνθήκες εισαγωγής και ταξινομεί τις κατηγορίες εισαγωγής με υψηλή ακρίβεια. Ο αλγόριθμος δεν βασίζεται σε έτοιμα γονίδια.



Εικόνα 17 Σχηματική Απεικόνιση του Gecko Algorithm

Έχει γίνει εφαρμογή του αλγόριθμου σε διάφορα προβλήματα και σε διαφορετικές αλληλουχίες δεδομένων :

1. Ανάλυση αλληλουχίας miRNA για την ταξινόμηση φυσιολογικών αιμοσφαιρίων
2. Ανάλυση αλληλουχίας mRNA για την ταξινόμηση δεδομένων καρκίνου του μαστού και την πρόβλεψη της απόκρισης με χημειοθεραπεία
3. Ανάλυση αλληλουχίας BS-seq δειγμάτων έναντι χρόνιας λεμφοκυτταρικής λευχαιμίας (CLL)

Ανεξάρτητα με τον τύπο δεδομένων ο αλγόριθμος χρησιμοποιεί ακριβείς υπογραφές για την ταξινόμηση των δειγμάτων και μπορούν να χρησιμοποιηθούν ως διαγνωστικοί και προγνωστικοί δείκτες.

Developing parallel ant colonies filtered

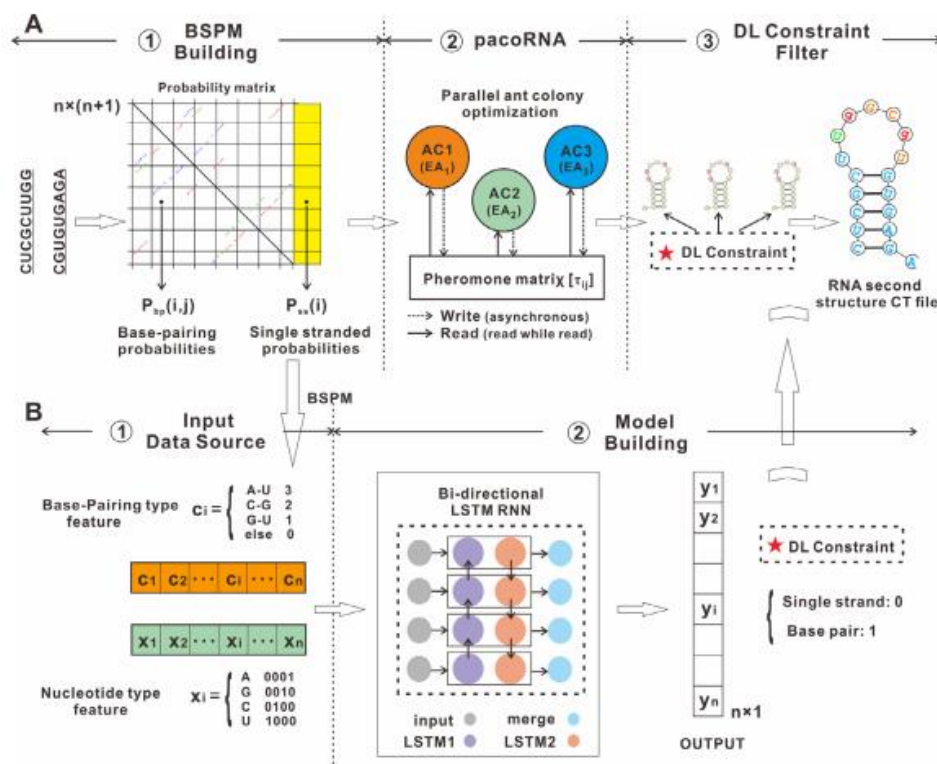
Το αντικείμενο της έρευνας Quan, Lijun, et al.[33] είναι η παρουσίαση ενός μοντέλου πρόβλεψης DracoRNA με σκοπό την βελτίωση της ακρίβειας στην πρόβλεψη της δευτερεύουσας δομής RNA με pseudo-knots , η οποία αποτελείται από 2 χαρακτηριστικά :

- Παράλληλο αλγόριθμο αναζήτησης
- Μάθηση δομικού περιορισμού με χρήση βαθιάς μάθησης

Ο συγκεκριμένος αλγόριθμος είναι παραλλαγή του αλγόριθμου PACO με χρήση τεχνικών βαθιάς μάθησης.

Ο αλγόριθμος αποτελείται από 3 στάδια :

1. Δημιουργείται ένας πίνακας πιθανοτήτων(BSPM) συναρμολογώντας έναν πίνακα πιθανότητας ζευγαρώματος βάσης και μια λίστα πιθανοτήτων μονόκλωνων
2. Σχεδιασμός του αλγορίθμου PACO για την αναζήτηση της καλύτερης δευτερεύοντος δομή RNA όπου το εκπαιδευμένο BSPM θα χρησιμοποιηθεί ως σημαντικό σύνολο αποτελεσμάτων
3. Το περιορισμένο ζεύγος βάσεων βαθιάς μάθησης θα χρησιμοποιηθεί για την τροποποίηση της δευτερεύουσας δομής που λαμβάνει ο αλγόριθμος PACO



Εικόνα 18 Σχηματική Απεικόνιση του DracoRNA

Για την εκτέλεση του αλγόριθμου έπρεπε να χρησιμοποιηθούν 386.585 RNA ζεύγη βάσεων όπου 3.171 δευτερεύουσες δομές RNA που ανήκαν στο κομμάτι του training, validating και testing sets όπου οι 1.039 δευτερεύουσες δομές RNA περιείχαν pseudo-knots.

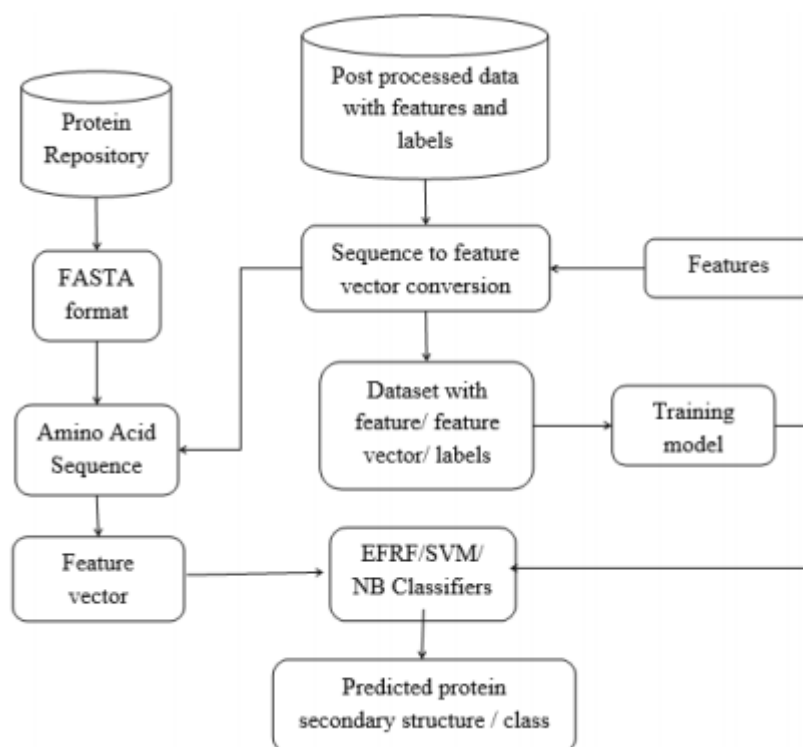
Prediction Of Protein Secondary Structure

Το αντικείμενο της έρευνας Sekhar, Sivagnanam R. Mani et al[15] είναι η παρουσίαση ενός αλγορίθμου ταξινόμησης Enhanced Fuzzy Random Forest που λαμβάνει πληροφορίες από διάφορες ιστοσελίδες για την εξαγωγή συνόλων δεδομένων μη επιβλεπόμενης μάθησης.

Χρησιμοποιώντας feature vectors δημιουργούνται δομές αλληλουχίας πρωτεϊνών τα οποία χρησιμοποιούνται ως είσοδο στον αλγόριθμο για την πρόβλεψη της δευτερεύουσας δομής RNA.

Τα βήματα της έρευνας που ακολούθησαν ήταν:

1. Παράλληλη συλλογή και ανάλυση των δεδομένων
2. Σχεδιασμός και δημιουργία feature vectors
3. Πρόβλεψη της πρωτεΐνης της δευτερεύουσας δομής με χρήση του αλγορίθμου Enhanced Fuzzy Random Forest



Εικόνα 19 Συνολική σχηματική απεικόνιση του Enhanced Fuzzy Random Forest

Για την εκτέλεση του αλγόριθμου έπρεπε να δημιουργηθούν 2 ξεχωριστοί αλγόριθμοι που θα εστίαζαν σε διαφορετικό κομμάτι υπολογισμού :

- την ταξινόμηση με την χρήση του random forest
- την δημιουργία αλληλουχίας πρωτεϊνών με feature vector

Algorithm 1: Proposed EFRF with Feature Vector Scoring

Input: Primary Protein Sequence

Output: Protein secondary structure

```
1. begin
2.     Compute feature vector score
3.         Apply Algorithm 2
4.     return vector score
5.     Generate protein secondary structure
6.         Apply Algorithm 3
7.     return protein class
8. end
```

Εικόνα 20 Συνολική διαδικασία του αλγόριθμου

Ο αλγόριθμος για το feature vector παίρνει την αλληλουχία ως είσοδο και υπολογίζει την σύνδεση, μετάβαση και διανομή τιμής. Αφού επιτευχθεί υπολογίζει τη συχνότητα εμφάνισης της ακολουθίας σε κάθε θέση του πίνακα και τέλος υπολογίζει την διανυσματική βαθμολογία για τη δεδομένη ακολουθία από τον πίνακα.

Algorithm 2: Proposed Feature Vector Scoring process

Input: Primary Protein Sequence

Output: A vector score for the Sequence

1. Procedure find Feature Vector (sequenceSQ)
2. **begin**
3. **for** each property **do**
4. divideAA --> 3Groups
5. calcComposition(SQ) --> return compArr
6. calcTransition(SQ) --> return transArr
7. calcDistribution(SQ) --> return distArr
8. **for** each AA **do**
10. calcf requencyof AAinSQ
11. **return** PFArr
12. **for** each in array **do**
13. fv[]+ = array
14. **return** fv
15. **end**

Εικόνα 21 Αλγόριθμος υπολογισμού Feature Vector

Ο αλγόριθμος για την παραγωγή πρωτεϊνικής δευτερεύουσας δομής ξεκινάει με την τοποθέτηση του εκπαιδευτικού μοντέλου(Enhanced Fuzzy Random Forest) για την κατηγορία των πρωτεϊνών. Στη συνέχεια διαλέγει τυχαία χαρακτηριστικά και αποθηκεύει το αποτέλεσμα του διαχωρισμού. Μετά γίνεται η δημιουργία ενός random forest με ασαφής ιδέα για την πρόβλεψη της κατηγορίας χρησιμοποιώντας το μοντέλο εκπαίδευσης.

Το αποτέλεσμα της κατηγορίας αποθηκεύεται σε συγκεκριμένη μεταβλητή και κάθε μεταβλητή υπολογίζεται ξεχωριστά και τέλος επιστρέφει την μεγαλύτερη τιμή της πρωτεϊνικής δευτερεύουσας δομής στον αλγόριθμο για να γίνει η πρόβλεψη με αποτέλεσμα ακρίβειας 96%.

Algorithm 3: Enhanced Fuzzy Random Forest

Input: Feature Vector of size 1 x 105

Output: Predicted Protein Structure

```
1. TrainingModel ()
2. begin
3.   k=RandomSelectFeatures() (From m features, k
   << m)
4.   d = FuzzyBestSplit(k)
5.   i = 0
6.   while ((i < n) && (n! =1)) do
7.     d = FuzzyBestSplit(d)
8.     BuildFuzzyForest(d, i)
9.     PredictionFromTrainedModel()
10.    targetOutcome[n] =
    RandomDecisionTree(k) {AlphaHelix, RandomCoil,
        EtendedStrand}
11.    votes[m] ==>
    CalculateVotes(targetOutcome[i]) {i = 0, 1, n =size} {m <<
    n}
12.    max = 0
13.    i = 0
14.    while (i < n) do
15.      if (max < votes[j] {j = 0, 1, ..m})
16.        max = votes[j]
17.    return max
18. end
```

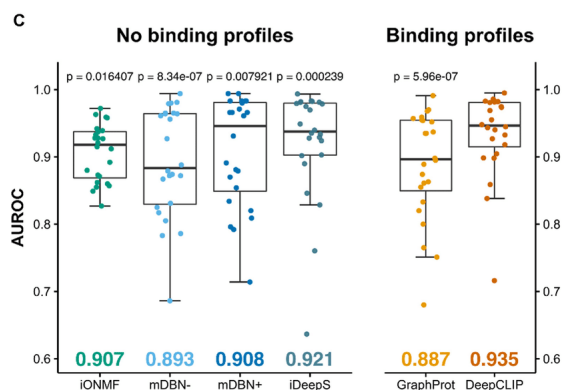
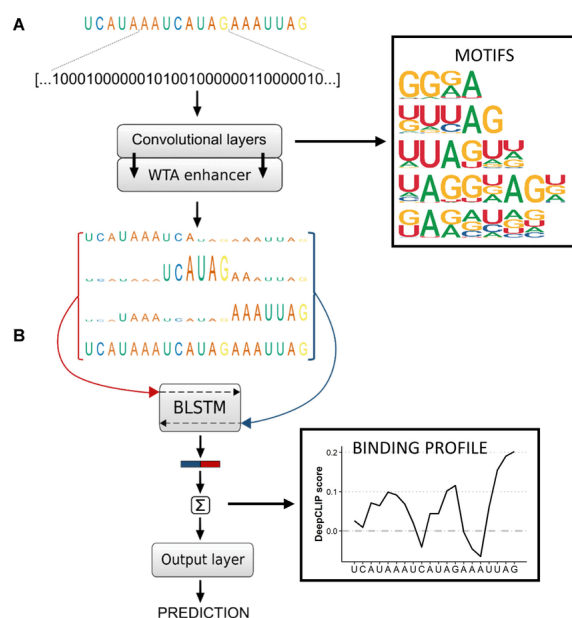
Εικόνα 22 Αλγόριθμος εκπαίδευσης

DeepCLIP

Το αντικείμενο της έρευνας Alexander Gulliver, et al[28] είναι η παρουσίαση ενός μοντέλου βαθιάς μάθησης για την πρόβλεψη και μοντελοποίηση πρωτεϊνικής σύνδεσης με νουκλεϊκά οξέα RNA με χρήση δεδομένων αλληλουχίας(sequencing data).

Το συγκεκριμένο μοντέλο είναι νευρωνικό δίκτυο βαθιάς μάθησης που χρησιμοποιεί:

- συνελκτικά στρώματα(convolutional layers) μονοδιάστατου μεγέθους για εύρεση και ενίσχυση χαρακτηριστικών ενός συνόλου ακολουθιών
- αμφίδρομη μακροπρόθεσμη μνήμη(bidirectional long short term memory) για την εξαγωγή χαρακτηριστικών και πληροφορίας με βάση τα συμφραζόμενα των ακολουθιών για την εύρεση των αλληλουχιών-RNA που σχετίζονται με RBP σύνδεση



Εικόνα 23 Διαδικασία ταξινόμησης του DeepClip

Το μοντέλο λαμβάνει μια ή περισσότερες μικρές αλληλουχίες RNA ως είσοδο και προβλεπόμενη πιθανότητα σύνδεσης και υπολογίζει ένα προφίλ σύνδεσης.

Ο κύριος σκοπός του είναι ο εντοπισμός θέσεων δέσμευσης πρωτεϊνών σε νέες δοκιμασμένες ακολουθίες χρησιμοποιώντας εκπαιδευμένα μοντέλα που έχουν γίνει εξαγωγή δεσμευτικής πληροφορίας ιστού η οποία παρέχεται από ήδη υπάρχοντα δεδομένα, να προβλέπει την επίδραση των παραλλαγών ακολουθίας στη σύνδεση και για τον προσδιορισμό της σημασίας των μεμονωμένων νουκλεοτιδίων να γίνει η δέσμευση της πρωτεΐνης.

Στο κομμάτι του διαχωρισμού των δεδομένων γίνεται ο διαχωρισμός στους στα εξής:

- Το 80% των δεδομένων ως εισόδου χρησιμοποιείται για training
- Το 10% των δεδομένων για validating
- Το υπόλοιπο 10% για testing

Τα δεδομένα που χρησιμοποιούνται για training και validating εκπαιδεύονται στο μοντέλο πάρα πολλές φορές. Στην διάρκεια της εκπαίδευσής του η απόδοση μετριέται με το validating που έχει τεθεί σε κάθε επανάληψη και το μοντέλο με τα καλύτερα αποτελέσματα αποθηκεύεται. Μπορεί να γίνει η εφαρμογή πρόωρης διακοπής για την αποφυγή της μέγιστης απόδοσης.

Η τελική απόδοση του μοντέλου χρησιμοποιείται στο 10% των δεδομένων που αποτελούνται για testing ώστε να μην υπάρχουν δεδομένα τα οποία έχουν εισήχθη στο μοντέλο.

Στο κομμάτι της αξιολόγησης του δεδομένου χρησιμοποιείται η μέθοδος 10-fold cross validation ίσου μεγέθους.

Τα συμπεράσματα στα οποία κατέληξαν ήταν :

- Το μοντέλο τους ξεπερνά τα δομικά και μη δομικά μοντέλα με βάση των δεδομένων αλληλουχίας μόνο
- Το μοντέλο μπορεί να γίνει η πρόβλεψη σε δεσμευτικά μοτίβα
- Προβλέπει και αναλύει τις μεταλλάξεις των συνδέσεων RNA
- Τα αποτελέσματα του μοντέλου με εξώνια hnRNP A1 ΚΑΙ SRSF1 συσχετίζονται με τα επίπεδα συμπερίληψης εξωνίου ενός γνωστού εξαρτώμενου από SRSF1 εξωνίου.

SCISSOR

Το αντικείμενο της έρευνας Choi, Hyo Young, et al[10] είναι η παρουσίαση μιας εφαρμογής για την ανακάλυψη παρεκκλίνουσας δομής μεταγραφής RNA με την χρήση δεδομένων αλληλουχίας σύντομης ανάγνωσης με βάση τις αλλαγές του σχήματος ενός προφίλ ανακάλυψης RNA αλληλουχίας.

Η εφαρμογή παρουσιάζει μια διαφορετική προσέγγιση για την συστηματική ανακάλυψη αλλαγών έκφρασης mRNA. Επίσης χρησιμοποιεί βασική ανάλυση δεδομένων κάλυψης που παρέχουν σημαντικές πληροφορίες σε ότι αφορά τις τιμές έκφρασης γονιδίου/ εξωνίου ή συνδέσμων συναρμογής.

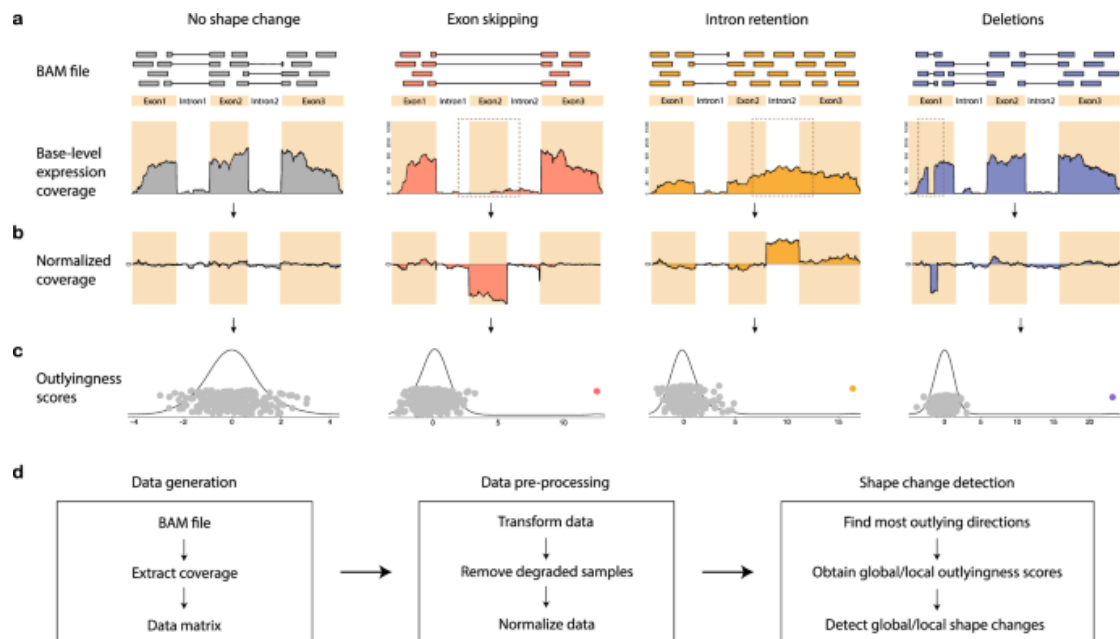
Η μέθοδος αυτή στοχεύει στον εντοπισμό διαρθρωτικών παραλλαγών ή διαφορικών προτύπων κάλυψης σε όλες τις ομάδες των αλληλουχιών RNA χωρίς καμία υποκείμενη υπόθεση του μηχανισμού. Αυτό επιτρέπει την μείωση της εξάρτησης από γνωστά γονιδιακά μοντέλα ώστε να γίνει η αύξηση της δυνατότητας προσδιορισμού με διαφορετικές συγκαλυμμένες γενετικές εκτροπές.

Πριν γίνει η αλλαγή του σχήματος πρέπει να γίνει :

- Διαδικασία ανίχνευσης
- Ομαλοποίηση
- Σταθεροποίηση διακύμανσης

για την δημιουργία περισσότερων συμμετρικών κατανομών ώστε να γίνει η εκτέλεση του SCISSOR αλλά και η διευκόλυνση της ερμηνείας των αποτελεσμάτων σε όλα τα γονίδια. Επιπροσθέτως πραγματοποιήθηκε πρόσθετη προ-επεξεργασία που μπορεί να θεωρηθεί ότι καθορίζει την υπολειμματική έκφραση prebase για κάθε δείγμα μετά την αφαίρεση ενός μοντέλου που αντιπροσωπεύει.

Επίσης γίνεται αναφορά και σε πιθανή ανώμαλη υπολειμματική δομή ως λανθασμένη γονιδιακή έκφραση που δεν παρατηρήθηκε άμεσα κατά την αφαίρεση του μοντέλου από τα κανονικοποιημένα δεδομένα.



Εικόνα 24 Διαδικασία SCISSOR για την ανίχνευση αλλαγών σε ένα μόνο γονίδιο

Για την εκτέλεση της μεθόδου-εφαρμογής χρησιμοποιήθηκαν 522 πλακώδους TCGA κεφαλής και λαιμού καρκινώματος(HNSC) για τον εντοπισμό γνωστών ανωμαλιών σε καρκινικά γονίδια.

Στη συνέχεια εντοπίστηκαν νέες εκτροπές που δεν είχαν αναφερθεί σε παρόμοιες εφαρμογές ενώ είχε γίνει χρήση των ίδιων δεδομένων συμπεριλαμβανομένης της εναλλακτικής σύνδεσης ισομορφών , συντήξεις και ενδογενείς διαγραφές. Τέλος έγινε εφαρμογή σε ολόκληρο το γονιδίωμα για την αναγνώριση ενός νέου συνόλου γονιδίων με ισχυρές ενδείξεις παρεκκλίνουσας δομής στο καρκίνωμα (HNSC)

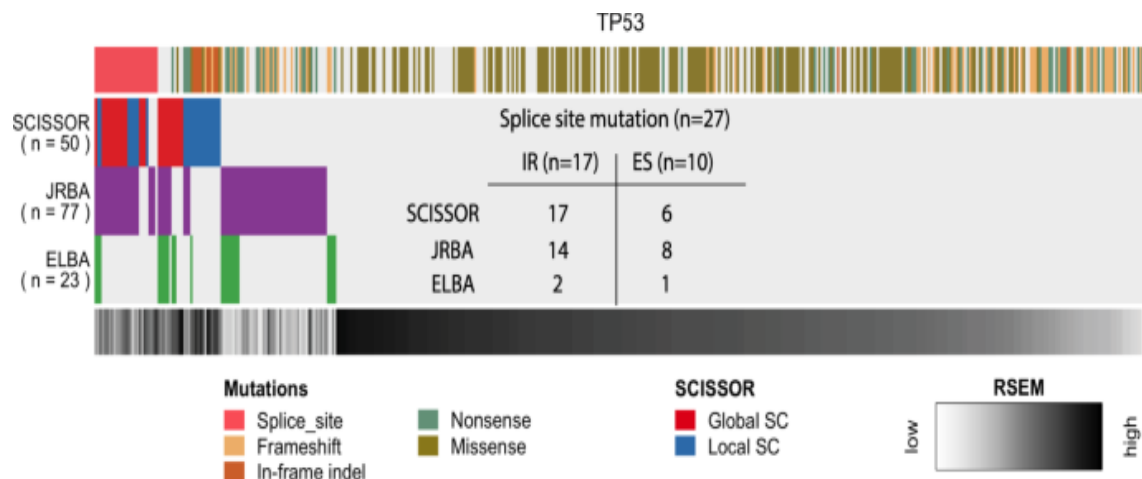
Στο κομμάτι της εφαρμογής έγινε και η επικύρωση της λειτουργίας του μοντέλου. Για την επικύρωση του μοντέλου έγινε χρήση δημόσιων δεδομένων σε μεταλλάξεις από TCGA καθώς και σε TCGA RNA αλληλουχία δεδομένων. Αν και τα γονίδια μεταλλάσσονται συχνά , η μεμονωμένη δομική αλλαγή θα είναι επαναλαμβανόμενη ή αλληλεπικαλυπτόμενη.

Σε αυτό το κομμάτι η μέθοδος ανιχνεύει μη επαναλαμβανόμενα εξωτερικά γεγονότα που επεκτείνουμε στα απλά γονίδια και σπάνια στα μεταλλαγμένα γονίδια.

Υπάρχουν προκλήσεις στη σύγκριση του SCISSOR με άλλες πολλές τεχνικές ανάλυσης μεταγραφής στις οποίες επικεντρώνονται σε μεγάλο βαθμό στην εποπτευόμενη ανάλυση με στόχο την ανίχνευση διαφορεικής ισομορφής μεταξύ ομάδων δειγμάτων.

Παρόλα αυτά, έγινε προσπάθεια σύγκρισης για τυχόν πιθανά οφέλη από την υψηλή διάσταση του μετασχηματισμού του RNA που προσφέρει το SCISSOR σε προσεγγίσεις που βασίζονται σε περισσότερα δεδομένα με χαμηλότερες διαστάσεις, όπως δεδομένα περίληψης εξωνίου ή εναλλακτική σύνδεση 13,25.

Έγινε και η προσαρμογή μεθόδων που είναι αντιπροσωπευτικές στην ανάλυση βασισμένων σε ανάγνωση (JRBA) και βασισμένες σε έκφραση επιπέδου εξωνίου (ELBA) και έγινε σύγκριση της απόδοσή τους στον εντοπισμό αλλοιώσεων του TP53 καθώς και σε προσομοιωμένα δεδομένα.



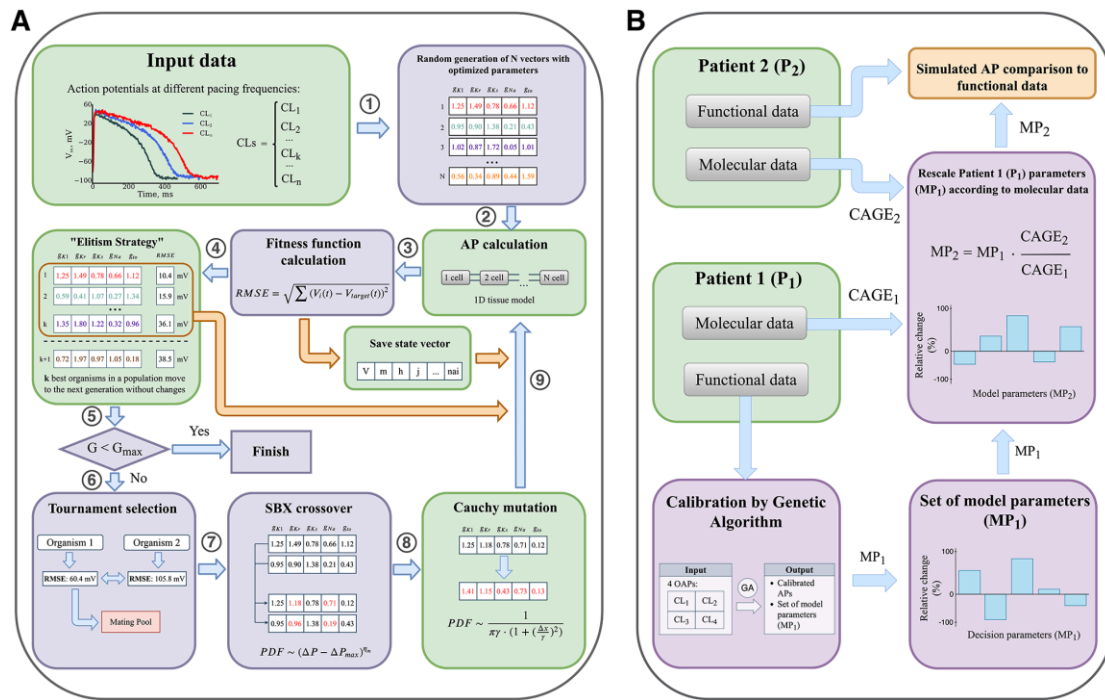
Εικόνα 25 Συγκριτικό του SCISSOR, JRBA και ELBA

Με βάση την παραπάνω σύγκριση, η μέθοδος SCISSOR παρήγαγε αμερόληπτα αποτελέσματα, λιγότερο εξαρτημένα από το γονίδιο σε επίπεδα έκφρασης, χάρη στη διαδικασία ομαλοποίησης που βοήθησε στην ελαχιστοποίηση ασήμαντων και αναξιόπιστων σημάτων από τη χαμηλή κάλυψη.

Genetic algorithm-based personalized models

Το αντικείμενο της έρευνας Smirnov, Dmitrii, et al.[23] είναι η παρουσίαση ενός τροποποιημένου γενετικού αλγόριθμου που καθορίζει συγκεκριμένες παραμέτρους βασισμένο στο μοντέλο ηλεκτροφυσιολογίας καρδιομυοκυττάρων σύμφωνα με σύνολο πειραματικών ανθρώπων action potential(AP) που καταγράφεται σε διαφορετικούς καρδιακούς ρυθμούς.

Για την βέλτιστη λύση ο αλγόριθμος εκτελεί ταυτόχρονη αναζήτηση σε παραμετρικούς και αργούς μεταβλητούς χώρους. Ο αλγόριθμος για να εκτελεστεί θα πρέπει να γίνουν μερικές τροποποιήσεις για την αποτελεσματική σύγκλιση.



Εικόνα 26 α) Διάγραμμα του γενετικού αλγορίθμου και β) CAGE-based εξατομίκευση

Οι τροποποιήσεις που έγιναν αφορούν στα παρακάτω:

1. Δεδομένα εισόδου όπου τα δεδομένα έγιναν renormalized κατά το μήκος του άξονα του χρόνου προκειμένου να επιβάλει συγκριτικές κυματομορφές. Συγκεκριμένα, η μισή μέγιστη από-πόλωση των κυματομορφών χρησιμοποιήθηκε ως σημείο αναφοράς.
2. Υπολογισμός AP όπου έγινε προσομοίωση σε 1d ιστό αντί για ένα μεμονωμένο κύτταρο όταν εφαρμοζόταν ο αλγόριθμος σε πειραματικά δεδομένα που κατέγραφαν ιστό
3. Λειτουργία φυσικής κατάστασης όπου χρησιμοποιήθηκε η μετρική RMSE για την αξιολόγηση του πόσο κοντά είναι ένα δεδομένο οργανισμού AP σε ένα PCL δεδομένο
4. Αποθήκευση φορέα κατάστασης όπου έγινε το υπολογιστικό κόστος βηματοδότησης κάθε οργανισμού σε κάθε γενιά έως την επίτευξη της σταθερής κατάστασης κατά την διάρκεια μιας εκτέλεσης του αλγόριθμου
5. «Στρατηγική ελιτισμού» όπου όταν γίνεται χρήση ενός μεγάλου αριθμού ελίτ οργανισμού τότε γίνεται αποτελεσματικά η εκτέλεση του αλγόριθμου.

6. Μετάλλαξη Cauchy όπου οδηγεί σε καλύτερη σύγκλιση για λειτουργίες με πολλά τοπικά ελάχιστα. Επομένως, έγινε η τροποποίηση του τελεστή μετάλλαξης ώστε να γίνει χρήση της διανομής Cauchy:

$$f_x(x) = \frac{c}{\pi} \left[\frac{\gamma}{(x - x_0)^2 + \gamma^2} \right]$$

Η απόδοση του αλγόριθμου επιτεύχθηκε σε πειραματικά δεδομένα καρδιακής οπτικής χαρτογράφησης και σε προφίλ έκφρασης mRNA.

Στο αποτέλεσμα που κατέληξαν ήταν στην ακρίβεια των παραμέτρων εξόδου όπου τα μοντέλα που βασίζονται σε mRNA προβλέπουν την κυματομορφή AP και την αποκατάσταση των ασθενών.

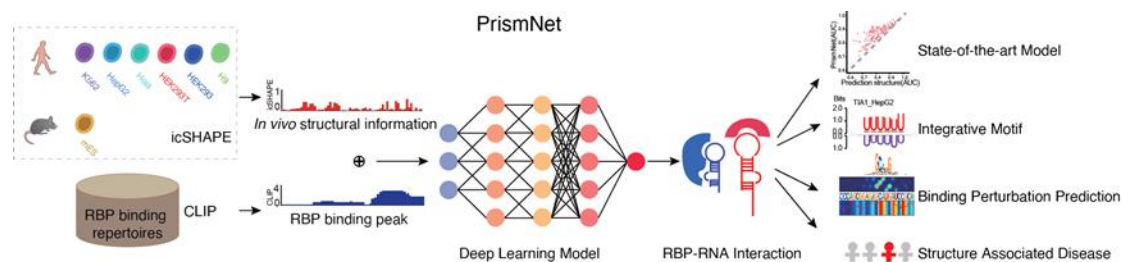
Ουσιαστικά ο συνδυασμός ενός γενετικού αλγόριθμου με τις μετρήσεις έκφρασης mRNA παρέχει μια νέα τεχνική για το μοντέλο εξατομίκευσης.

Predicting dynamic cellular protein

Το αντικείμενο της έρευνας Sun, Lei, et al.[35] είναι η παρουσίαση ενός νευρωνικού δικτύου βαθιάς μάθησης το οποίο προβλέπει με ακρίβεια δεδομένα πρωτεϊνών που δεσμεύουν RNA.

Το συγκεκριμένο νευρωνικό δίκτυο ενσωματώνει πειραματικά δεδομένα νίνο RNA και δεδομένα δέσμευσης πρωτεϊνών RNA(ή αλλιώς RBP) για αντιστοιχισμένα κελιά με σκοπό την δυνατότερη πρόβλεψη RBP σε διάφορες κυτταρικές συνθήκες.

Το νευρωνικό δίκτυο μπορεί να χρησιμοποιηθεί για την ανατομή και την πρόβλεψη διαταραχών των γενετικών παραλλαγών που σχετίζονται με όσους νοσούν.



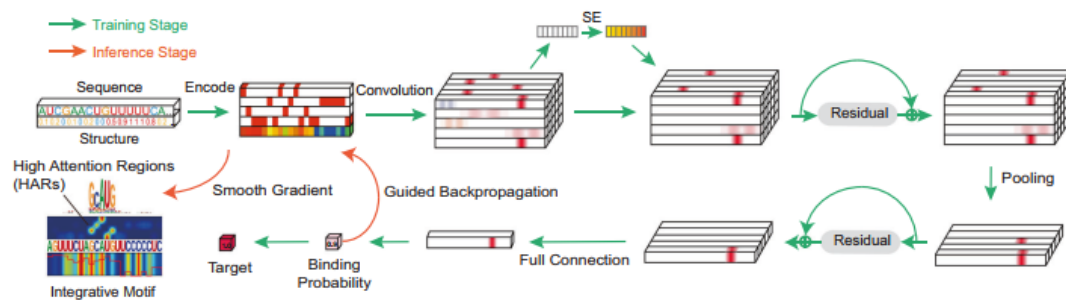
Εικόνα 27 Διαδικασία μοντελοποίησης και πρόβλεψης συνδέσεων RBP

Για την εισαγωγή των δεδομένων γίνεται κωδικοποίηση σε 4d διαστάσεων μονόχρωμη αλλά και η χρήση icSHAPE για την δομική βαθμολόγηση ως 5d διάσταση.

Το νευρωνικό δίκτυο αποτελείται από:

- Πολλαπλά συνελκτικά στρώματα(Convolutional layers)

- Δίκτυα συμπίεσης και διέγερσης(SE networks)
- Υπολειμματικά μπλοκ για να συλλάβουν τους προσδιοριστές αλληλουχίας και δομικής σύνδεσης RBP



Εικόνα 28 Αρχιτεκτονική Νευρωνικού Δικτύου

Για την αποτελεσματικότερη πρόβλεψη και εκτέλεση του νευρωνικού δικτύου δημιουργήθηκε ένα μεγάλο σύνολο δεδομένων δομικών προφίλ RNA σε διαφορετικό επίπεδο ανάλογα με τις κυτταρικές γραμμές θηλαστικών.

Τα συγκεκριμένα δεδομένα δείχνουν ότι η δομή του RNA εμφανίζει σημαντική ποσότητα εξαρτώμενη από την μεταβλητή της κυτταρικής γραμμής δηλώνοντας έντονα την πιθανή ισχυρή επιρροή του στην ακρίβεια και συνάφεια με βάση την πρόβλεψη των μοντέλων σε ανάλογα δεδομένα.

Με τα παραπάνω δεδομένα κατέληξαν στο συμπέρασμα ότι το μοντέλο αφού μαθαίνει την αλληλεπίδραση πρωτεΐνης-RNA και κατά συνέπεια ήταν και σε θέση να προβλέψει με ακρίβεια το RBP σε διαφορετικές κυτταρικές συνθήκες.

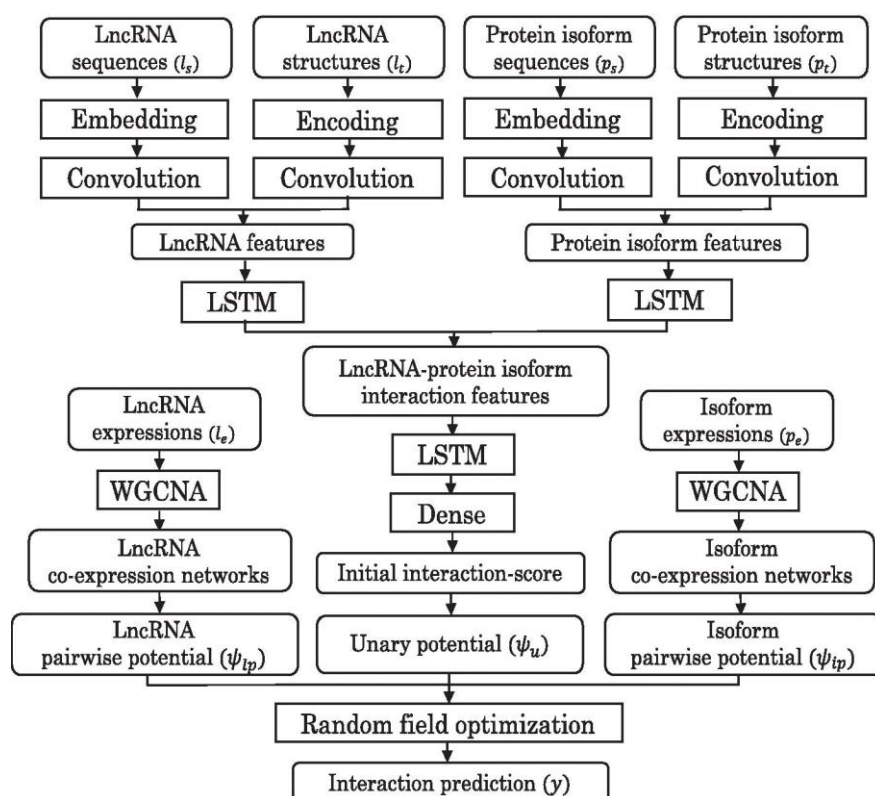
DeepLPI

Το αντικείμενο της έρευνας Shaw, Dipan, et al.[3] ήταν η παρουσίαση μιας μεθόδου που προβλέπει την αλληλεπίδραση μεταξύ lncRNAs και πρωτεϊνικών ισομορφών.

Η συγκεκριμένη μέθοδος χρησιμοποιεί ως δεδομένα ακολουθία και δομή για την εξαγωγή εγγενών χαρακτηριστικών και δεδομένα έκφρασης για εξαγωγή τοπολογικών χαρακτηριστικών. Για την ένωση των διαφορετικών δεδομένων έγινε η δημιουργία ενός υβριδικού πλαισίου που θα είναι ενσωματωμένο σε νευρωνικό δίκτυο βαθιάς μάθησης και σε τυχαίο πεδίο με όρους(CRF).

Το νευρωνικό δίκτυο βαθιάς μάθησης εξάγει και συγχωνεύει τα χαρακτηριστικά ,με χρήση convolutional layers, από την ακολουθία και δομή δεδομένων των lncRNA και των ισομορφών πρωτεϊνών και υπολογίζει τις αλληλεπιδράσεις μεταξύ lncRNA και των ισομορφών πρωτεΐνης.

Για τον υπολογισμό έγινε χρήση τεχνικής ενσωμάτωσης για την χαρτογράφηση των ακολουθιών lncRNA και των πρωτεϊνών σε ένα διανυσματικό πίνακα 64 διαστάσεων.



Εικόνα 29 Διάγραμμα της Αρχιτεκτονικής DeepLPI

Αφού γίνει η εκπαίδευση , τα embedding layers είναι σε θέση να μάθουν τις κατάλληλες αντιστοιχίσεις δηλαδή χαρτογραφημένα πυκνά διανύσματα θα μπορούσαν να αναζητήσουν ομοιότητες των ακολουθιών. Στη συνέχεια , οι dense

διανυσματικοί πίνακες που αντιπροσωπεύουν τις ακολουθίες και τους πίνακες που κωδικοποιούν το προβλεπόμενο lncRNA και τις πρωτεΐνες διέρχονται από μονοδιάστατες convolutional layers με 4 φίλτρα για να ληφθούν τα τοπικά χαρακτηριστικά της αλληλουχίας και των δομών δεδομένων.

Μετά απ' αυτό, χρησιμοποιήθηκαν τα max pooling layers για την μείωση του δείγματος της εξόδου των convolutional layers με σκοπό την μείωση του χρόνου εκμάθησης των επόμενων στρωμάτων. Βασισμένα στα μειωμένα χαρακτηριστικά, τα long short-term memory layers(LSTM) χρησιμοποιούνται για την εκμάθηση των χαρακτηριστικών που αντιπροσωπεύουν το μικρό εύρος και τις ιδιότητες μακράς εμβέλειας των αλληλουχιών και των δομών.

Αυτά τα χαρακτηριστικά εξάγονται από αλληλουχίες και δομές lncRNA, αλληλουχίες ισομορφών πρωτεΐνης και οι δομές συγχωνεύονται μαζί ως είσοδο ενός LSTM. Το στρώμα LSTM και το dense layer προορίζονται για να μάθουν τα μοτίβα αλληλεπίδρασης μεταξύ lncRNA και πρωτεΐνης.

Τέλος, η έξοδος του dense layer τροφοδοτείται σε ένα επίπεδο παλινδρόμησης logistics για να υπολογίσει την αρχική βαθμολογία αλληλεπίδρασης.

Αφού επιτευχθεί η εκπαίδευση του νευρωνικού, το μοντέλο CRF εκτελεί την τελική πρόβλεψη που βασίζεται τόσο στις αρχικές βαθμολογίες που υπολογίζονται από το νευρωνικό δίκτυο όσο και στα δεδομένα έκφρασης των mRNA και lncRNA. Για να ξεπεραστεί η έλλειψη βασικής αλήθειας αλληλεπιδράσεων μεταξύ lncRNA και πρωτεϊνών, αναπτύχθηκε ένας αλγόριθμος μη επιβλεπόμενης μάθησης για την εκπαίδευση του νευρωνικού και του CRF μαζί επαναληπτικά.

Για τη μη δημιουργία ελλείψεων γνωστικών αλληλεπιδράσεων μεταξύ lncRNA και πρωτεϊνικών ισομορφών εφαρμόστηκε μια πολλαπλή προσέγγιση εκμάθησης(MIL).

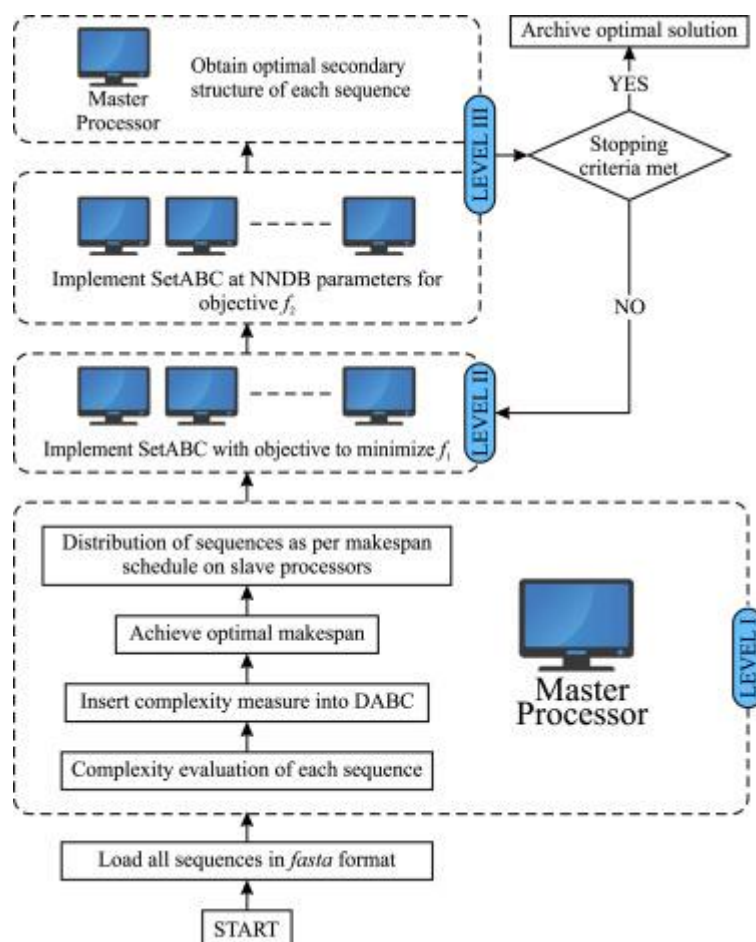
Το DeepLPI αρχικά εκκινεί όλες τις περιπτώσεις θετικές μεταβλητές-περιπτώσεις και τις υπόλοιπες αρνητικές περιπτώσεις. Στη συνέχεια, οι παράμετροι του μοντέλου βελτιστοποιούνται με τις αρχικές ετικέτες ακολουθώντας τον τυπικό τρόπο επιβλεπόμενης μάθησης.

An efficient three-Level parallel ABC algorithm

Το αντικείμενο της έρευνας Lalwani et.al.[19] ήταν η παρουσίαση ενός παράλληλου τεχνητού αλγόριθμου(3L-PABCfold) για την αποτελεσματική πρόβλεψη δευτερογενούς δομής πολύπλοκων ακολουθιών RNA με χρήση παράλληλων πολλαπλών επεξεργαστών χρήση υπολογιστή.

Ο συγκεκριμένος αλγόριθμος αποτελείται από 3 υπό-αλγόριθμους :

1. Παράλληλος αλγόριθμος ABC τριών επιπέδων
2. Καθορισμένος αλγόριθμος ABC
3. Διακριτός ABC



Εικόνα 30 Διάγραμμα αλγόριθμου 3L-PABCfold

Ο αλγόριθμος χρησιμοποιεί 1.000.000 δεδομένα RNA αλληλουχιών. Για την λειτουργία του αλγορίθμου γίνονται τα παρακάτω βήματα:

- Τοποθετούνται τα δεδομένα των αλληλουχιών σε μορφή fasta
- Για το πρώτο επίπεδο:
 1. Γίνεται αξιολόγηση πολυπλοκότητας κάθε αλληλουχίας

2. Εισαγωγή του μέτρου πολυπλοκότητας στον DABC αλγόριθμο για την βέλτιστη διάρκεια ζωής της αλληλουχίας
 3. Διανομή αλληλουχιών σύμφωνα με το πρόγραμμα λειτουργίας slave processors
- Για το δεύτερο επίπεδο γίνεται η εκτέλεση των slave processors με την βοήθεια του καθορισμένου αλγόριθμου ABC όπου εκτελεί τα ακόλουθα βήματα:
 1. Φάση Αρχικοποίησης
 2. Εφαρμογή του μετρητή επανάληψης μέχρι τον μέγιστο αριθμό επιτυχούς επανάληψης
 3. Φάση εκτέλεσης
 4. Αξιολόγηση των τιμών όπου αν είναι επιτυχής η αξιολόγηση για το τρίτο επίπεδο τότε αντικατέστησε τις τιμές f_1, f_2
 5. Εφαρμογή greedy διαδικασίας επιλογής
 6. Αξιολόγηση της πιθανότητας
 7. Φάση Παρατήρησης μέχρις ότου ο αριθμός των αναβαθμίσεων να είναι μικρότερος από τους αριθμούς των αποτελεσμάτων
 8. Φάση scout όπου προσδιορίζεται η εγκαταλειμμένη λύση. Αν υπάρχει τέτοια τιμή γίνεται αντικατάστασή της με τυχαία παραγόμενη νέα λύση στο χώρο αναζήτησης προβλημάτων
 9. Αποθήκευση της καλύτερης λύσης μέχρι τώρα.
 - Για το τρίτο επίπεδο γίνεται εκτέλεση στον master server χρησιμοποιώντας τα αποτελέσματα των λύσεων από το δεύτερο επίπεδο όπου γίνεται η μέτρηση της μέγιστης αντιστοιχίας με σκοπό την αξιολόγηση των ζευγών για την πλήρη δευτερεύουσα δομή στις παραμέτρους του NNDB χρησιμοποιώντας τον καθορισμένο αλγόριθμο ABC.

Algorithm 1 Process flow of DABC

Initialize the population of solutions x_i , $i=1,2,...,n$; Set the iteration counter t and maximum number of iterations T .
Evaluate $obj=f(x)$ for each solution.

While $t < T$

Employed bee phase:

Produce a new solution v_i in the neighbourhood of x_i using:

$$v_i = LPT \left(x_i \oplus (rand \otimes (x_i \odot x_r)) \right) \quad (3)$$

here x_r is a randomly selected solution. Evaluate the fitness value of x_i by:

$$fit_i(x_i) = \begin{cases} \frac{1}{1+f(x_i)} \\ 1+f(x_i) \end{cases} \quad (4)$$

Apply greedy selection process between v_i and x_i .

Evaluate probability P_i :

$$P_i = \frac{fit_i(x_i)}{\sum_{i=1}^n fit_i(x_i)} \quad (5)$$

Onlooker bee phase:

While Number of update < Number of Solutions

Select a solution x_i , based on P_i .

Produce a new solution v_i as shown in Eq. (2).

Compute fitness fit_i of each solution.

Apply greedy selection process between v_i and x_i .

endWhile

Scout bee phase:

Determine the abandoned solution x_i .

Replace it with a randomly produced new solution within the problem search space.

Save the best solution so far.

endWhile

Εικόνα 31 Διαδικασία του αλγόριθμου DABC

RNAsamba

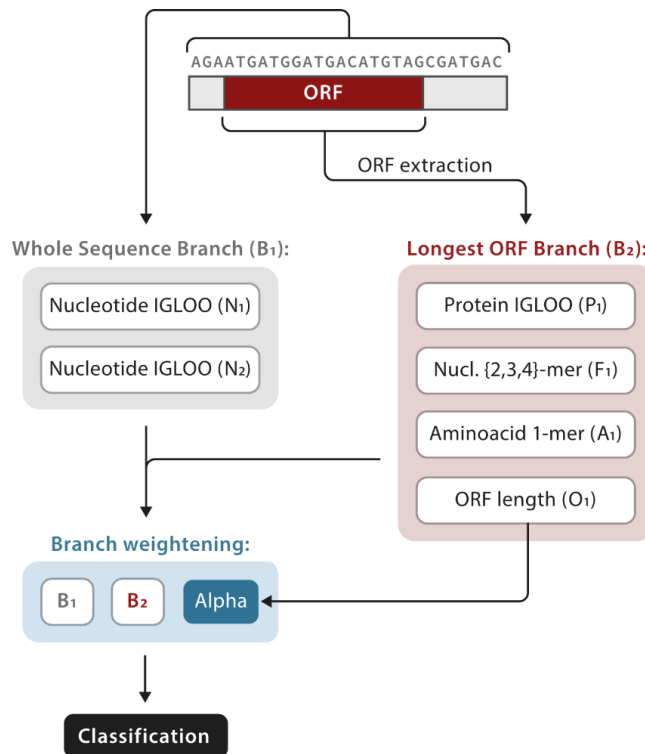
Το αντικείμενο της έρευνας Camargo, Antonio P., et al. [24] ήταν η περιγραφή του αλγόριθμου RNAsamba για την πρόβλεψη της δυνατότητας κωδικοποίησης των μορίων RNA από πληροφορίες αλληλουχιών με χρήση νευρωνικού δικτύου που μοντελοποιεί ολόκληρη την ακολουθία και την χρήση του ORF για τον εντοπισμό μοτίβων που διακρίνουν την κωδικοποίηση από τις μεταγραφές που δεν κωδικοποιούν.

Ξεκινώντας από την αρχική αλληλουχία ο RNAsamba υπολογίζει τη κωδικοποίηση μιας δεδομένης μεταγραφής συνδυάζοντας πληροφορίες που προέρχονται από 2 διαφορετικές πηγές:

- Ολόκληρο κλάδο ακολουθίας(B1)
- Μεγαλύτερος ORF Branch(B2)

Ο B1 περιέχει την μεταγραφή ακολουθιών και μπορεί να συλλάβει την κωδικοποίηση των πρωτεϊνών ανεξάρτητα από την ταυτότητα του ORF. Αντίθετα, ο B2 μεταγράφει τις πληροφορίες που αντλήθηκαν από το μεγαλύτερο και έχουν ταυτοποιηθεί στον ORF.

Λαμβάνοντας υπόψη αυτές τις 2 πηγές πληροφοριών αλληλουχίας, ο RNAsamba δημιουργεί μια λεπτομερή αναπαράσταση του αντίγραφου , βελτιώνοντας την απόδοση της ταξινόμησης του αλγόριθμου.



Εικόνα 32 Διαγραμματική Απεικόνιση RNAsamba

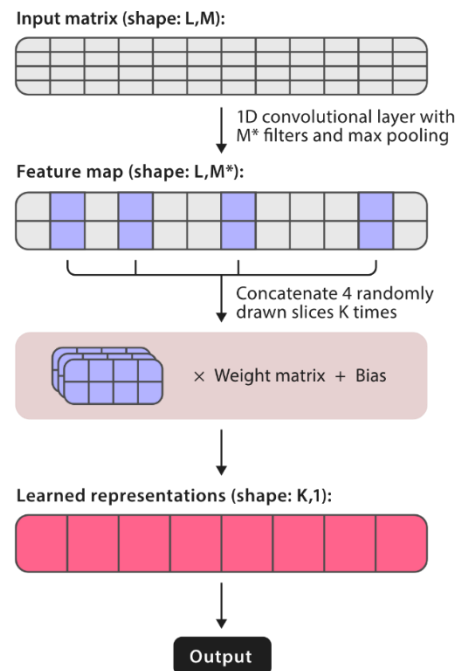
Για την εξαγωγή των ORF αλληλουχιών πληροφοριών για την τοποθέτηση του στο B2 ο RNAsamba σαρώνει καθένα από τα τρία πλαίσια ανάγνωσης αναζητώντας θράσματα που ξεκινούν με κωδικόνιο έναρξης (ATG) και τελειώνουν είτε με κωδικόνιο στάσης (TAG, TAA ή TGA) ή στο τέλος της μεταγραφής. Το μεγαλύτερο κομμάτι απ' αυτά που βρέθηκαν σε όλα τα πλαίσια ανάγνωσης εξάγονται στη συνέχεια ανεξάρτητα με το κωδικόνιο. Στην αντίθετη περίπτωση το B2 δεν θα φέρει καμία πληροφορία και η ταξινόμηση θα έχει αποκλειστικά από το B1.

Στο κομμάτι της προεπεξεργασίας ο RNAsamba δημιουργεί αναπαραστάσεις υψηλού επιπέδου και των 2 αλληλουχιών νουκλεοτιδίων και αμινοξέων χρησιμοποιώντας μονάδες IGLOO.

Καθώς αυτές οι μονάδες απαιτούν ακολουθίες σταθερού μήκους εισόδου, οι αλληλουχίες μεταγραφής και πρωτεΐνης περικόπτονται στο μέγιστο με μήκος 3000 νουκλεοτιδίων και 1000 αμινοξέων. Παρόλο που υπάρχουν τα όρια, ο αλγόριθμος παρουσιάζει γρηγορότερους χρόνους εκπαίδευσης και μπορεί να συλλάβει αρκετές πληροφορίες για την σωστή ταξινόμηση πολύ μεγάλων μεταγραφών.

Κατά την εκπαίδευση του αλγορίθμου, οι αλληλουχίες υποβάλλονται σε προεπεξεργασία σε αριθμητικές πληροφορίες και διαδίδονται μέσω του νευρωνικού δικτύου σε κομμάτια προσαρμόζοντας σταδιακά τα βάρη μάθησης για την βελτιστοποίηση της απόδοσης ταξινόμησης του μοντέλου. Οι μεταγραφές εισόδου περνούν από την ίδια προεπεξεργασία αλλά αντί να χρησιμοποιηθούν για

την ενημέρωση του δικτύου, χρησιμοποιούνται για τον υπολογισμό της δυνατότητας κωδικοποίησής τους.



Εικόνα 33 Προεπεξεργασία Δεδομένων σε αριθμητικές πληροφορίες

Τα αποτελέσματα στα οποία κατέληξαν ήταν ότι ο αλγόριθμος RNAamba:

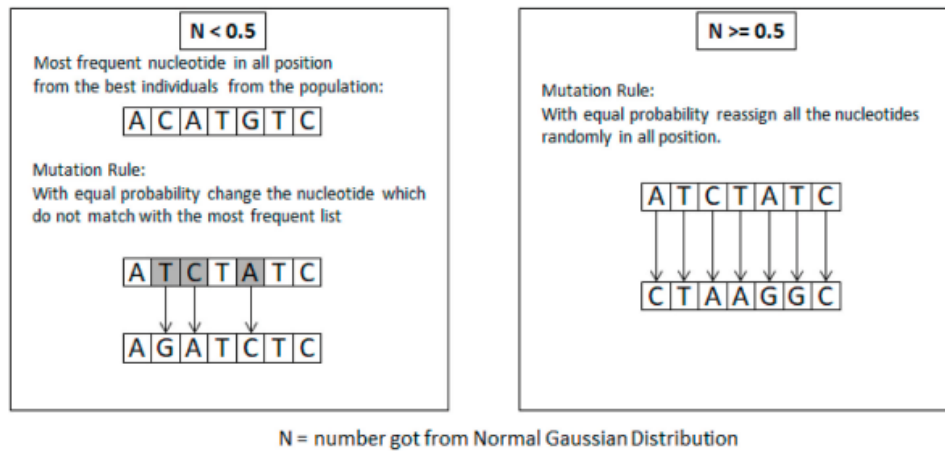
1. Διακρίνει με ακρίβεια τα mRNA από τα lncRNA σε πολλά σύνολα δεδομένων
2. Γενικεύεται σε διαφορετικά είδη δεδομένων
3. Μπορεί να προσδιορίσει περικομμένες ακολουθίες κωδικοποίησης
4. Μπορεί να ανιχνεύσει μια αλληλουχία που σχετίζεται με την μετάφραση εκτός του ORF
5. Μπορεί να χρησιμοποιηθεί για τον εντοπισμό mRNA που κωδικοποιεί μικροπεπτίδια
6. Είναι ταχύτερο από τις εναλλακτικές λύσεις που βασίζονται σε νευρωνικά δίκτυα

MFEA

Το αντικείμενο της έρευνας Ashraf, Faisal Bin et.al[16] η δημιουργία ενός εξελικτικού αλγόριθμου που ελαχιστοποιεί την περιττή εξερεύνηση στον motif space υποψήφιο και οδηγεί στην πιο ακριβή αναζήτηση ζώνης στο χώρο διαλύματος ακολουθώντας μερικές τεχνικές μετάλλαξης στις υπο-ακολουθίες των αλληλουχιών DNA.

Ο αλγόριθμος δημιουργεί έναν αρχικό πληθυσμό υποψηφίων μοτίβων (άτομα του πληθυσμού) και στην συνέχεια μετράει την καταλληλότητα όλων των ατόμων. Στη συνέχεια, γίνεται αναγέννηση του πληθυσμού κρατώντας τα καλύτερα γονίδια και εφαρμόζοντας διασταύρωση και μετάλλαξη μεταξύ τους. Σε κάθε επανάληψη,

ενημερώνεται η καλύτερη λύση και γίνεται κράτηση της επανάληψης μέχρι να ξεκινήσει η καλύτερη λύση για μεγάλο χρονικό διάστημα.



Εικόνα 34 Τεχνική Μετάλλαξης για την δημιουργία πληθυσμού

Ο αλγόριθμος εκτελεί αυτήν την προσέγγιση αρκετές φορές με διαφορετικό αρχικό πληθυσμό ώστε να διασφαλιστεί η παγκόσμια βέλτιστη λύση

Algorithm for Motif Finding
<pre> listOfBest ← {} μ ← number of parents selected λ ← number of children generated by parents l ← length of motif we want to find for l times do P ← buildInitialPopulation() Best ← {} repeat for each candidate $P_i \in P$ do AssessFitness(P_i) if Best={} or Fitness(P_i) > Fitness(Best) then Best ← P_i end if end for Q ← μ individuals in P which fitness are greatest P ← Q for each candidate motif $Q_j \in Q$ do for $\frac{\lambda}{\mu}$ times do $P \leftarrow P \cup \{Mutate(Copy(Q_j))\}$ end for end for until Best is ideal solution listOfBest ← listOfBest ∪ Best end for MOTIF ← Consensus(listOfBest) return MOTIF </pre>

Εικόνα 35 Ο ψευδοκώδικας του αλγόριθμου MFEA

Για την δημιουργία του αρχικού πληθυσμού ο αλγόριθμος επιλέγει τυχαία υπο-ακολουθία μήκους 1 από τις αλληλουχίες DNA και θα γεμίσουν με τον αρχικό πληθυσμό ως υποψήφια μοτίβα. Η επιλογή του αρχικού πληθυσμού από το σύνολο δεδομένων βρίσκεται κοντά στην βέλτιστη λύση ως υπάρχον σύνολο δεδομένων.

Αφού γίνει η δημιουργία του πληθυσμού, αξιολογούμε την καταλληλότητα των ατόμων. Με την εύρεση της ελάχιστης απόστασης του υποψήφιου μοτίβου με τις ακολουθίες του συνόλου δεδομένων, θα υπολογιστεί η συνολική απόσταση από ολόκληρο το σύνολο δεδομένων που θα είναι η φυσική κατάσταση. Η χαμηλότερη τιμή θα αντιπροσωπεύει την υψηλότερη προσαρμοσμένη υποψήφια λύση για το σύνολο των δεδομένων.

Algorithm for *AssessFitness*(P_i)

```
fitness  $\leftarrow$  0
l  $\leftarrow$  length of motif we want to find
for each sequence in data set do
    tempMin  $\leftarrow$   $\infty$ 
    for each sub-sequence of length l in the sequence do
        d = Hamming Distance between  $P_i$  and sub-sequence
        if d < tempMin then
            tempMin  $\leftarrow$  d
        end if
    end for
    fitness  $\leftarrow$  fitness + tempMin
end for
return fitness
```

Εικόνα 36 Αλγόριθμος υπολογισμού συνολικής απόστασης

Το κομμάτι της μετάλλαξης παίζει σημαντικό ρόλο στον συγκεκριμένο αλγόριθμο, διότι διαχειρίζεται την ανταλλαγή μεταξύ της εξερεύνησης και της εκμετάλλευσής του στο χώρο της αναζήτησης. Σε αυτή τη μέθοδο, βρίσκεται η υψηλότερη συχνότητα νουκλεοτιδίου σε κάθε θέση από τα καλύτερα υποψήφια μοτίβα που βρέθηκαν στην επανάληψη.

Η λίστα των νουκλεοτιδίων μας δίνει τον καλύτερο δυνατό συνδυασμό τους με βάση τη φυσική κατάσταση που έχει αξιολογηθεί. Με την χρήση και εισαγωγή της Gaussian Κατανομή διατηρήθηκε ο συμβιβασμός μεταξύ εξερεύνησης και εκμετάλλευσης. Η κανονική κατανομή δίνει χαμηλότερους αριθμούς μεταξύ 0 και 1 και σπάνια θα δίνει μεγαλύτερους αριθμούς.

Χρησιμοποιήθηκε αυτή η τεχνική ώστε να μπορούμε πάντα να βρισκόμαστε στο καλύτερο μέρος του χώρου λύσεων και σπάνια να υπάρξει μεγάλη αύξηση τιμών για την μετάβαση σε διαφορετικό μέρος του χώρου λύσεων εάν υπάρχει καλύτερη λύση.

Από την στιγμή που παραμένουν στον ίδιο χώρο λύσεων, τείνουν να μεταβούν σε καλύτερη λύση αντικαθιστώντας μόνο τα νουκλεοτίδια που δεν είναι πιο συχνά με την βοήθεια μιας λίστας όπου η μεγαλύτερη συχνότητα νουκλεοτιδίων αποθηκεύεται σε διαφορετικές θέσεις. Σε περίπτωση μετακίνησης, γίνεται αλλαγή

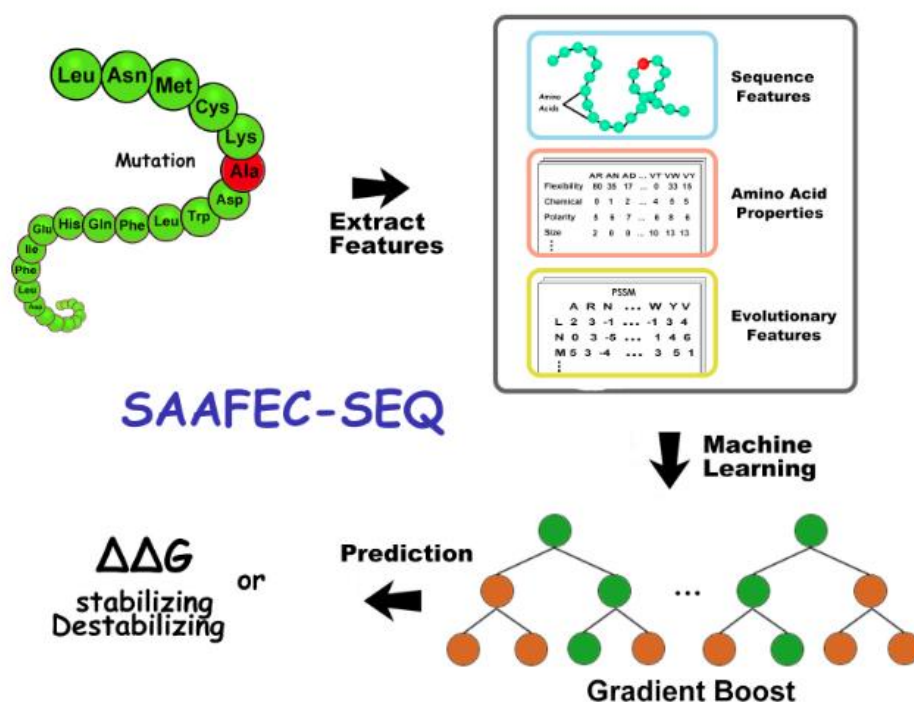
με τυχαίο τρόπο των νουκλεοτιδίων σε διαφορετική θέση με αποτέλεσμα να δημιουργηθεί διαφορετικό μοτίβο υποψηφίου.

Το τελικό κομμάτι του αλγορίθμου είναι να υπολογίζει τη συχνότητα του καθενός νουκλεοτιδίου σε κάθε θέση και επιστρέφει την μήτρα που περιέχει την πιθανότητα της κάθε θέσης νουκλεοτιδίου.

SAAFEC-SEQ

Το αντικείμενο της έρευνας Li, Gen et al.[12] ήταν η ενίσχυση του αλγορίθμου SAAFEC με χρήση μηχανικής μάθησης, δέντρα απόφασης, για την πρόβλεψη της αλλαγής της αναδιπλούμενης ελεύθερης ενέργειας που προκαλείται από τις αντικαταστάσεις των αμινοξέων.

Ο αλγόριθμος δεν χρειάζεται τρισδιάστατη δομή της αντίστοιχης πρωτεΐνης αλλά μόνο την αλληλουχία του και μπορεί να εφαρμοστεί σε έρευνες κλίμακας γονιδίων. Ο αλγόριθμος SAAFEC-SEQ χρησιμοποιεί φυσικοχημικές ιδιότητες, χαρακτηριστικά αλληλουχίας και εξελικτικές πληροφορίες για την πρόβλεψη του αποτελέσματος.



Εικόνα 37 Διαγραμματική Απεικόνιση SAAFEC-SEQ

Έγινε χρήση αλγορίθμου μηχανικής μάθησης με εξαγόμενα χαρακτηριστικά τις φυσικοχημικές ιδιότητες των θέσεων μετάλλαξης, τα χαρακτηριστικά ακολουθίας και τις εξελικτικές πληροφορίες για την πρόβλεψη της αναδιπλούμενης ελεύθερης ενέργειας που προκύπτει από μια μετάλλαξη αμινοξέων.

Για την εκπαίδευση του αλγορίθμου χρησιμοποιήθηκε γνωστό σύνολο δεδομένων που περιέχει πειραματικά DDG2648 μεταλλάξεων απλού σημείου από 131 διαφορετικές πρωτεΐνες.

Ο αλγόριθμος χρησιμοποιεί 4 διαφορετικά χαρακτηριστικά ομάδας για να διακρίνει τις αλλαγές σταθερότητας που προκαλούνται από λανθασμένες μεταλλάξεις ενός σημείου.

Για την αξιολόγηση της σημασίας του χαρακτηριστικού αλλά και για την εκπαίδευση του μοντέλου χρησιμοποιήθηκε δέντρο απόφασης. Η σημασία του χαρακτηριστικού υπολογίζεται με την χρήση της αύξησης της κλίσης οι οποία περιλαμβάνει τον υπολογισμό του ποσού βελτίωσης του μέτρου απόδοσης λόγω κάθε σημείου διαίρεσης χαρακτηριστικών σταθμισμένο με τον αριθμό των παρατηρήσεων.

Στη συνέχεια υπολογίζεται ο μέσος όρος για τη σημασία των χαρακτηριστικών σε όλα τα δέντρα αποφάσεων για κάθε χαρακτηριστικό. Αυτά τα 2 χαρακτηριστικά αποτυπώνουν την εξελικτική διατήρηση ενός δεδομένου αμινοξέος στη θέση της μετάλλαξης καθώς και του περιβάλλοντος της θέσης της μετάλλαξης αλλά και την αλλαγή του κατά την μετάλλαξη.

Το τρίτο σημαντικότερο χαρακτηριστικό της ομάδας είναι η φυσικοχημική ιδιότητα. Το τέταρτο σημαντικό χαρακτηριστικό είναι η ακολουθία γείτονα. Το συγκεκριμένο χαρακτηριστικό αποτυπώνει την επιρροή του γειτονικού υπολείμματος αμινοξέων στη θέση μετάλλαξης.

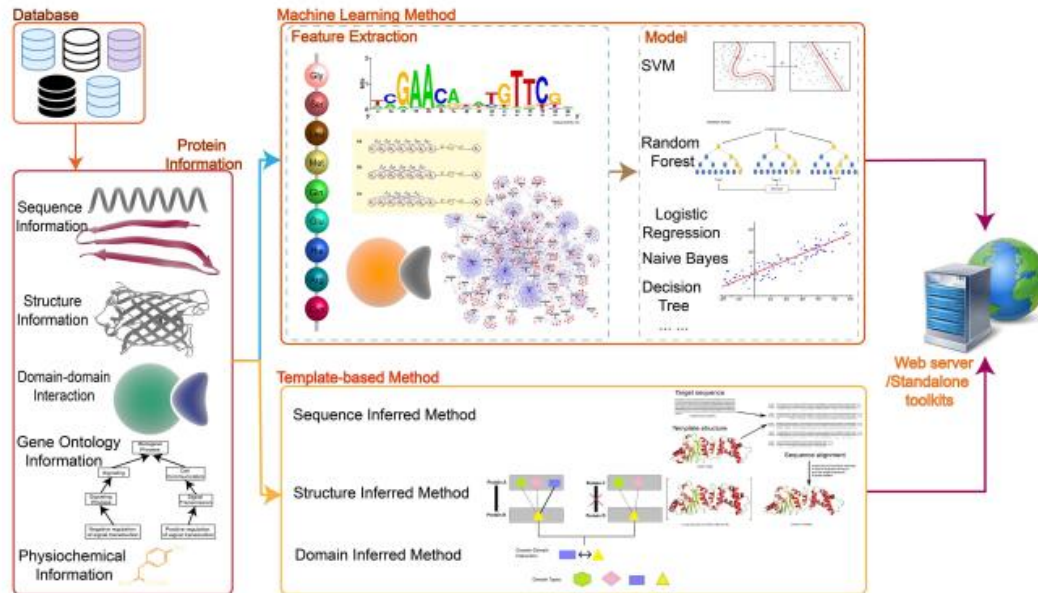
Για το κομμάτι της αξιολόγησης του αλγορίθμου χρησιμοποιήθηκε ο 5-fold cross validation για 100 εκτελέσεις. Ο αλγόριθμος αξιολογήθηκε σε 350 μεταλλάξεις που επιλέχθηκαν τυχαία από το σύνολο δεδομένων S2648. Επιτεύχθηκε δοκιμή σε 3 ανεξάρτητα σύνολα δεδομένων(S276,p53,CAGI15) για την καλύτερη απόδοση του αλγορίθμου και σύγκρισή του με άλλες μεθόδους που βασίζονται σε αλληλουχία.

Systematic evaluation of machine learning methods

Το αντικείμενο της έρευνας Chen, Huaming, et al.[29] ήταν η παρουσίαση μιας συστηματικής αξιολόγησης των υπολογιστικών μεθόδων που βασίζονται στην μηχανική μάθηση για τις αλληλεπιδράσεις ανθρώπου(πρωτεΐνης)-βακτηρίου(πρωτεΐνης) (HB-PPIs). Η συστηματική αξιολόγηση επιτεύχθηκε από 2 πλευρές. Η πρώτη εξέτασε την εφαρμογή των αλγορίθμων αναπαράστασης χαρακτηριστικών στα πρωτεϊνικά δεδομένα, ενώ η δεύτερη σχετίζεται με μοντέλα μηχανικής μάθησης.

Το μοντέλο βασισμένο σε πρότυπο χρησιμοποιεί διαφορετικούς τύπους πρωτεϊνών πληροφοριών για τη δημιουργία πρόβλεψης, συμπεριλαμβανομένων των ακολουθιών πληροφοριών, πληροφοριών δομών και τομέων. Τέτοια μοντέλα

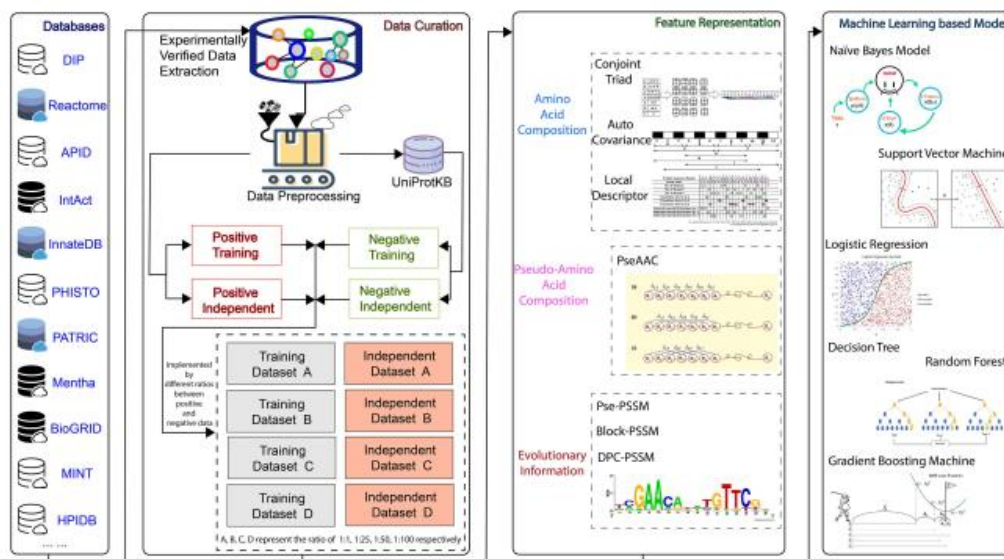
χρησιμοποιούν διαφορετικές πρωτεϊνικές πληροφορίες για την ανίχνευση υψηλής βαθμολογίας ομολογίας που θα μπορούσε να αποφέρει παρόμοιες λειτουργίες. Όμως τα μοντέλα αυτά ενδεχομένως να αποτύχουν ασχέτως εάν αλληλοεπιδρά η απομακρυσμένη ομολογία με γνωστές πρωτεΐνες.



Εικόνα 38 Εφαρμογή Μηχανικής Μάθησης και Βασισμένο σε Πρότυπα

Για την κωδικοποίηση πρωτεϊνών ως φορέων χαρακτηριστικών, πολλά διαφορετικά χαρακτηριστικά έχουν συμπεριληφθεί για την πρόβλεψη των PPI όπως είναι τα εξής:

- Αμινοξέα πρωτεΐνη
- Ψευδοαμινοξέα πρωτεΐνη
- Εξελικτική πληροφόρησης πρωτεΐνης



Εικόνα 39 Διαγραμματική Απεικόνιση Εφαρμογής κατά την αξιολόγηση

Ένας άλλος τρόπος υπολογιστικού μοντέλου είναι τα μοντέλα μηχανικής μάθησης όπου οι πληροφορίες για τις πρωτεΐνες είναι οι πρώτες που εισάγονται για να μάθουν τις εγγενείς σχέσεις τους αυτόματα, τα οποία χρησιμοποιούνται έτσι για την δημιουργία του μοντέλου και την πρόβλεψη των αλληλεπιδράσεων. Συγκεκριμένα για τα PPI, οι πρωτεϊνικές πληροφορίες μπορεί να είναι πληροφορίες αλληλουχίας, γονιδιακής οντολογίας, πληροφορίες γονιδιακής έκφρασης και πληροφορίες δικτύου αλληλεπίδρασης. Οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται είναι:

- Support vector machine(SVM) όπου εφαρμόζεται σε πολλές εργασίες ταξινόμησης και παλινδρόμησης στην υπολογιστική βιολογία. Περιέχει πυρήνα συνάρτησης ακτινικής βάσης η οποία ανατίθεται για την ταξινόμηση των ζευγών HP-PPI
- Decision Tree(DT) όπου εφαρμόζεται ως μη παραμετρικό μοντέλο μη επιβλεπόμενης μάθησης. Χρησιμοποιεί γράφημα που μοιάζει με δέντρο για την πρόβλεψη εισερχόμενης περίπτωσης βασισμένη στους κανόνες απόφασης από δείγματα δεδομένων και χαρακτηριστικών. Είναι κατανοητά και ερμηνεύσιμα και ικανά να χειριστούν και αριθμητικά και κατηγορηματικά δεδομένα
- Random Forest(RF) χρησιμοποιεί τυχαία μέθοδο εκμάθησης για την κατασκευή συνδυασμού decision tree. Έχει ανώτερες επιδόσεις στο κομμάτι της ταξινόμησης και παλινδρόμησης σε σύγκριση με υπόλοιπους αλγόριθμους μηχανικής μάθησης. Τεχνικά , είναι ένα μοντέλο εκμάθησης συνόλου που βασίζεται σε μεθόδους αποθήκευσης δέντρων για την δημιουργία δέσμης τυχαίων decision tree με σκοπό την αποφυγή λανθάνοντος προβλήματος που μπορεί να προκληθεί από δυνητικά προκατειλημμένα δεδομένα

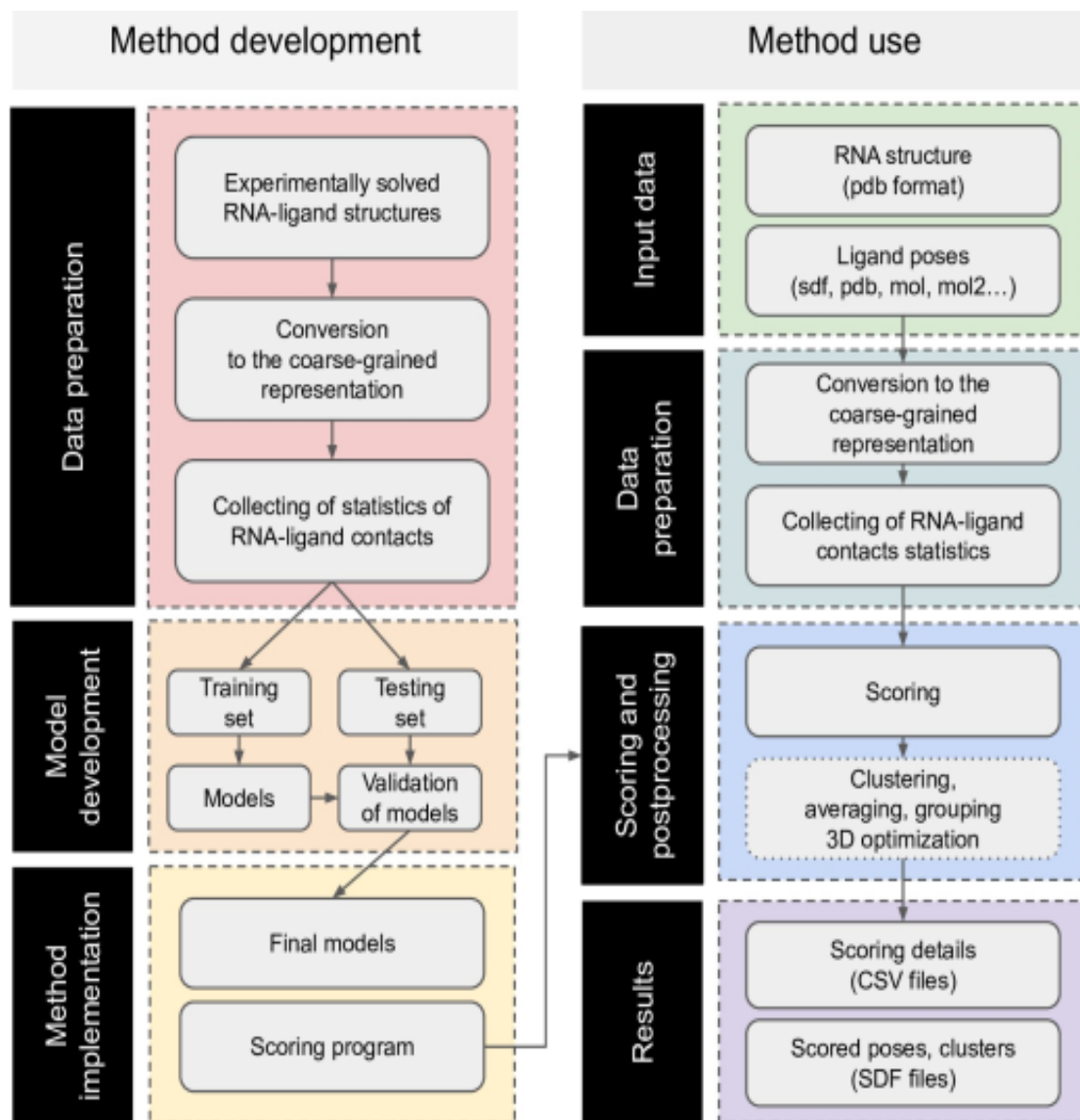
- Gaussian Naïve bayes (GNB) είναι πιθανολογικός ταξινομητής που θεωρεί τα χαρακτηριστικά ως ανεξάρτητες μεταβλητές όταν είναι η ετικέτα της κλάσης δεδομένου. Το συγκεκριμένο μοντέλο χρησιμοποιείται στην αντιμετώπιση των συνεχών δεδομένων που παράγονται από διάφορους αλγόριθμους αναπαράστασης χαρακτηριστικών
- Gradient boosting machine(GBM) είναι greedy μοντέλο βελτιστοποίησης τόσο για παλινδρόμηση όσο και για ταξινόμηση. Το μοντέλο αυτό είναι ενσωματωμένο σε μοντέλο decision tree

AnnapuRNA

Το αντικείμενο της έρευνας Stefaniak, Filip, et.al[14] ήταν η δημιουργία του AppuRNA, μιας λειτουργίας βαθμολόγησης βασισμένη στη γνώση που έχει σχεδιαστεί για την αξιολόγηση σύνθετων δομών RNA που δημιουργείται από οποιαδήποτε υπολογιστική μέθοδο σύνδεσης. Έγινε αξιολόγηση τριών κύριων παραγόντων που μπορεί να επηρεάσει την πρόβλεψη της δομής , δηλαδή τον αρχικό συμμορφωτή ενός συνδετήρα, πρόγραμμα σύνδεσης και το πως χρησιμοποιείται η λειτουργία βαθμολόγησης.

Ο αλγόριθμος αποτελεί στατιστική λειτουργία βαθμολόγησης μηχανικής μάθησης που αναπτύχθηκε για την πρόβλεψη των συμπλεγμάτων δομών RNA με υψηλή ακρίβεια. Χρησιμοποιεί εσωτερικά μια χονδροειδή αναπαράσταση και για τους 2 αλληλοεπιδρώντες RNA- συνδετήρων μικρών μορίων.

Χρησιμοποιώντας την συγκεκριμένη αναπαράσταση έχει αποδειχθεί επιτυχής προσομοίωση διαφόρων βιομοριακών συστημάτων, συμπεριλαμβανομένων της αναδίπλωσης του μορίου RNA από το πρόγραμμα SimRNA. Τα μοντέλα μηχανικής μάθησης που εκφράζουν την πιθανότητα αλληλεπιδράσεων RNA-συνδέτη, με βάση τα δεδομένα αλληλεπίδρασης που συλλέγονται από πειραματικές βάσεις δεδομένων.



Εικόνα 40 Μεθοδολογία Αναπαράστασης RNA δημιουργία και εκτέλεση

Χρησιμοποιήθηκαν 2 διαφορετικά δεδομένα για την εκτέλεση της μεθοδολογίας. Τα δεδομένα χωρίστηκαν σε 2 σετ δεδομένων. Το training set περιείχε 87 και 131 δομές RNA ενώ το test 33 δομές κοινές και στα 2 σετ δεδομένων.

Για να αποκτηθεί ένα γενικό μοντέλο αλληλεπιδράσεων από το τρέχον σύνολο σύνθετων δομών μικρού μορίου RNA χρησιμοποιήθηκε η αλληλεπίδραση RNA-συνδέτη.

Για τα μόρια RNA χρησιμοποιήθηκαν χονδροειδή αναπαράσταση βάση της μεθόδου προσομοίωσης SimRNA, με 5 σφαιρίδια ανά υπόλειμμα ριβονουκλεοτιδίου με θέσεις που αντιστοιχούν σε πραγματικά άτομα.

Για να αντληθούν τα μοντέλα βαθμολόγησης από τα στατιστικά στοιχεία πολυδιάστατων αλληλεπιδράσεων, χρησιμοποιήθηκαν αλγόριθμοι επιβλεπόμενης μάθησης:

- Deep Learning-Τεχνητό νευρωνικό δίκτυο τροφοδοσίας πολλαπλών στρωμάτων(DL)
- Gaussian Naïve Bayes(GNB)
- K Nearest Neighbors(kNN)
- Random Forests(RF)
- Support Vector Machine(SVM)-RBF kernel

Για την δημιουργία του AnnapuRNA με τα μοντέλα μηχανικής μάθησης χρησιμοποιήθηκε το training set δεδομένων και ως αξιολόγηση χρησιμοποιήθηκε 5-fold cross validation.

Για το βελτιστοποιημένο σύνολο παραμέτρων και τον ανεπτυγμένο αγωγό επεξεργασίας δημιουργήθηκαν 4 εκδόσεις προγνωστικών μοντέλων για 2 επιλεγμένους αλγόριθμους μηχανικής μάθησης(kNN και DL) χρησιμοποιώντας τα ίδια σύνολα δεδομένων.

Δεδομένου ότι το AnnapuRNA είναι μια μέθοδος knowledge based και εξαρτάται από το περιεχόμενο και την ποιότητα των βάσεων δεδομένων, ο κύριος περιορισμός του είναι ο αριθμός των πειραματικών προσδιορισμένων αλληλεπιδράσεων RNA-συνδέτη. Ωστόσο αυτός ο αριθμός αυξάνεται σταθερά και θα μπορούν να επιτευχθούν αναβαθμίσεις.

Συμπεράσματα

Οι παραπάνω εφαρμογές-αλγόριθμοι έδωσαν ως αποτέλεσμα την εύρεση της σύνδεσης RNA-πρωτεΐνη με αποτέλεσμα την επιτυχή εύρεση παθογενούς πρωτεΐνης. Η παρουσίαση των αλγορίθμων-εφαρμογών δίνει την δυνατότητα να γίνει σωστά η επεξεργασία των αλληλουχιών πρωτεϊνών ανεξαρτήτως δομής.

Για κάθε εκτέλεση αλγόριθμου-εφαρμογής χρησιμοποιήθηκαν δεδομένα είτε έτοιμες αλληλουχίες είτε αλληλουχίες που έπρεπε να γίνει προ-επεξεργασία. Ο κάθε αλγόριθμος-εφαρμογή χρησιμοποίησε τα δεδομένα για πρόγνωση δευτεροταγούς ή τριτοταγούς δομής με σκοπό την εύρεση παθογενούς πρωτεΐνης.

Η εύρεση παθογενούς πρωτεΐνης ήταν ένα απ' τα πιο δύσκολα κομμάτια στον τομέα της βιο-πληροφορικής με αποτέλεσμα να μην γινόταν σωστά η εύρεση της αναδίπλωσής τους. Χωρίς την αποτελεσματική εύρεση της σύνδεσης RNA-πρωτεΐνη δεν θα μπορούν να αναλύσουν και να αναζητήσουν την παθογενούς πρωτεΐνη.

Η μελέτη του μηχανισμού της αναδίπλωσης και γενικά η κατανόηση της προέλευσης ασθενειών που σχετίζονται με την μη σωστή αναδίπλωση των πρωτεϊνών στο χώρο είναι σε γενικές γραμμές ένα ανοιχτό ερευνητικό θέμα.

Μέσω της χρήσης των παραπάνω εφαρμογών-αλγορίθμων μπορεί να επιτευχθεί ως αποτέλεσμα με τον ίδιο ακριβώς τρόπο η εύρεση της παθογενούς πρωτεϊνικής αναδίπλωσης στην τριτοταγούς δομή.

Βιβλιογραφία

- [1] Acharya, Debarun, and Tapan K. Dutta. "Elucidating the network features and evolutionary attributes of intra-and interspecific protein–protein interactions between human and pathogenic bacteria." *Scientific Reports* 11.1 (2021): 1-11.
- [2] Richards, Alicia L., Manon Eckhardt, and Nevan J. Krogan. "Mass spectrometry-based protein–protein interaction networks for the study of human diseases." *Molecular Systems Biology* 17.1 (2021).
- [3] Shaw, Dipan, et al. "DeepLPI: a multimodal deep learning method for predicting the interactions between lncRNAs and protein isoforms." *BMC bioinformatics* 22.1 (2021): 1-22.
- [4] Jena, N. R., Suyash Pant, and Hemant Kumar Srivastava. "Artificially expanded genetic information systems (AEGISs) as potent inhibitors of the RNA-dependent RNA polymerase of the SARS-CoV-2." *Journal of Biomolecular Structure and Dynamics* (2021): 1-17.
- [5] Pontecorvi, Paola, et al. "Protein–protein interaction network analysis applied to DNA copy number profiling suggests new perspectives on the aetiology of Mayer–Rokitansky–Küster–Hauser syndrome." *Scientific reports* 11.1 (2021): 1-11.
- [6] Donepudi, Ajay C., et al. "Multidrug resistance-associated protein 4 (Mrp4) is a novel genetic factor in the pathogenesis of obesity and diabetes." *The FASEB Journal* 35.2 (2021): e21304.
- [7] Riffelmacher, Thomas, et al. "Metabolic activation and colitis pathogenesis is prevented by lymphotoxin β receptor expression in neutrophils." *Mucosal Immunology* (2021): 1-12.
- [8] Lu, Wei, et al. "Expression profiling and bioinformatics analysis of exosomal long noncoding RNAs in patients with myasthenia gravis by RNA sequencing." *Journal of Clinical Laboratory Analysis* (2021): e23722.
- [9] Eameh, Reza Zolfaghari, et al. "Identification and characterization of a silent mutation in RNA binding domain of N protein coding gene from SARS-CoV-2." *BMC Research Notes* 14.1 (2021): 1-6.
- [10] Choi, Hyo Young, et al. "SCISSOR: a framework for identifying structural changes in RNA transcripts." *Nature communications* 12.1 (2021): 1-12.
- [11] Bai, Shuheng, et al. "Construct a circRNA/miRNA/mRNA regulatory network to explore potential pathogenesis and therapy options of clear cell renal cell carcinoma." *Scientific Reports* 10.1 (2020): 1-15.

- [12] Li, Gen, Shailesh Kumar Panday, and Emil Alexov. "SAAFEC-SEQ: A Sequence-Based Method for Predicting the Effect of Single Point Mutations on Protein Thermodynamic Stability." *International Journal of Molecular Sciences* 22.2 (2021): 606.
- [13] Gupta, Shruti, Ajay Kumar Verma, and Shandar Ahmad. "Feature Selection for Topological Proximity Prediction of Single-Cell Transcriptomic Profiles in *Drosophila* Embryo Using Genetic Algorithm." *Genes* 12.1 (2021): 28.
- [14] Stefaniak, Filip, and Janusz M. Bujnicki. "AnnapuRNA: A scoring function for predicting RNA-small molecule binding poses." *PLoS computational biology* 17.2 (2021): e1008309.
- [15] Sekhar, Sivagnanam R. Mani, Siddesh G. Matt, and Sunilkumar S. Manvi. "Prediction of Protein Secondary Structure from Amino Acid Sequences by Integrating Fuzzy, Random Forest and Feature Vector Methodologies." *Jordan Journal of Biological Sciences* 13.4 (2020).
- [16] Ashraf, Faisal Bin, and Md Shafiur Raihan Shafi. "MFEA: An evolutionary approach for motif finding in DNA sequences." *Informatics in Medicine Unlocked* 21 (2020): 100466.
- [17] Cascarina, Sean M., and Eric D. Ross. "Natural and pathogenic protein sequence variation affecting prion-like domains within and across human proteomes." *BMC genomics* 21.1 (2020): 1-18.
- [18] Oliver, Gavin Robert, Garrett Jenkinson, and Eric W. Klee. "Computational detection of known pathogenic gene fusions in a Normal tissue database and implications for genetic disease research." *Frontiers in genetics* 11 (2020): 173.
- [19] Lalwani, Soniya, and Rajesh Kumar. "An efficient three-Level parallel ABC algorithm for secondary structure prediction of complex RNA sequences." *Applied Soft Computing* 99 (2021): 106848.
- [20] Thomas, Aubin, et al. "GECKO is a genetic algorithm to classify and explore high throughput sequencing data." *Communications biology* 2.1 (2019): 1-8.
- [21] Chen, Kunqi, et al. "RMDisease: a database of genetic variants that affect RNA modifications, with implications for epitranscriptome pathogenesis." *Nucleic acids research* 49.D1 (2021): D1396-D1404.
- [22] Luo, Xiaotong, et al. "RMVar: an updated database of functional variants involved in RNA modifications." *Nucleic Acids Research* 49.D1 (2021): D1405-D1412.
- [23] Smirnov, Dmitrii, et al. "Genetic algorithm-based personalized models of human cardiac action potential." *PloS one* 15.5 (2020): e0231695.

- [24] Camargo, Antonio P., et al. "RNAsamba: neural network-based assessment of the protein-coding potential of RNA sequences." *NAR Genomics and Bioinformatics* 2.1 (2020): lqz024.
- [25] Li, Wei, et al. "Integrated analysis of the functions and prognostic values of RNA binding proteins in lung squamous cell carcinoma." *Frontiers in genetics* 11 (2020): 185.
- [26] García-Díaz, Pilar, et al. "Unsupervised feature selection algorithm for multiclass cancer classification of gene expression RNA-Seq data." *Genomics* 112.2 (2020): 1916-1925.
- [27] Zhao, Qijie, et al. "m6A RNA modification modulates PI3K/Akt/mTOR signal pathway in Gastrointestinal Cancer." *Theranostics* 10.21 (2020): 9528.
- [28] Grønning, Alexander Gulliver Bjørnholt, et al. "DeepCLIP: predicting the effect of mutations on protein–RNA binding with deep learning." *Nucleic acids research* 48.13 (2020): 7099-7118.
- [29] Chen, Huaming, et al. "Systematic evaluation of machine learning methods for identifying human–pathogen protein–protein interactions." *Briefings in Bioinformatics* (2020).
- [30] Miladi, Milad, et al. "MutaRNA: analysis and visualization of mutation-induced changes in RNA structure." *Nucleic acids research* 48.W1 (2020): W287-W291.
- [31] Deschamps-Francoeur, Gabrielle, Joël Simoneau, and Michelle S. Scott. "Handling multi-mapped reads in RNA-seq." *Computational and Structural Biotechnology Journal* (2020).
- [32] Kuksa, Pavel P., et al. "HiPR: High-throughput probabilistic RNA structure inference." *Computational and structural biotechnology journal* 18 (2020): 1539-1547.
- [33] Quan, Lijun, et al. "Developing parallel ant colonies filtered by deep learned constrains for predicting RNA secondary structure with pseudo-knots." *Neurocomputing* 384 (2020): 104-114.
- [34] Barman, Bahnisikha, and Suvendra N. Bhattacharyya. "mRNA targeting to endoplasmic reticulum precedes ago protein interaction and microRNA (miRNA)-mediated translation repression in mammalian cells." *Journal of Biological Chemistry* 290.41 (2015): 24650-24656.
- [35] Sun, Lei, et al. "Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures." *Cell research* (2021): 1-22.

[36] Vemuri, Sasank, et al. "Slicelt: A genome-wide resource and visualization tool to design CRISPR/Cas9 screens for editing protein-RNA interaction sites in the human genome." *Methods* 178 (2020): 104-113.

[37] Pevsner, Jonathan. *Bioinformatics and functional genomics*. John Wiley & Sons, 2015.

[38] Lesk, Arthur. *Introduction to bioinformatics*. Oxford university press, 2019.

[39] Baxevanis, Andreas D., and BF Francis Ouellette. "A Practical Guide to the Analysis of Genes and Proteins."

[40] Μπάγκος, Παντελεήμων. "Βιοπληροφορική." (2015).

[41] Γεωργούλη, Α., 2015. Τεχνητή νοημοσύνη. [ηλεκτρ. βιβλ.] Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Διαθέσιμο στο: <http://hdl.handle.net/11419/3381>