



ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΘΕΜΑ:

Διερευνητική ανάλυση (exploratory analysis) δεδομένων μεγάλου όγκου από σύγχρονες τεχνικές της Μοριακής Βιολογίας.

Εισηγήςτης: Ιωάννης-Δημήτριος Στανίτσας

Επιβλέπων Καθηγητής: Αριστείδης Βραχάτης

Τριμελής Επιτροπή:

- Αριστείδης Βραχάτης, Επίκουρος καθηγητής, Ιόνιο Πανεπιστήμιο**
- Θεμιστοκλής Έξαρχος, Αναπληρωτής καθηγητής, Ιόνιο Πανεπιστήμιο**
- Λήδα Κερμανίδου, Αναπληρώτρια καθηγήτρια, Ιόνιο Πανεπιστήμιο**

Περιεχόμενα

1	Εισαγωγή.....	6
1.1	Εισαγωγή στη Μοριακή Βιολογία.....	6
1.2	Σημασία της διερευνητικής ανάλυσης δεδομένων στη Μοριακή Βιολογία.....	11
1.3	Σκοπός και δομή της εργασίας.....	12
2	Ανασκόπηση.....	13
2.1	Εξέλιξη της Μοριακής Βιολογίας: Σημαντικά επιτεύγματα και τεχνολογικές προόδους.....	13
2.2	Σύγχρονες τεχνικές στη Μοριακή Βιολογία.....	14
2.2.1	Single-cell RNA-seq sequencing.....	16
2.2.2	ChIP-seq (Chromatin Immunoprecipitation Sequencing).....	18
2.2.3	Ανάλυση της μεθυλίωσης του DNA - DNA methylation analysis..	19
2.2.4	Μεταβολομική ανάλυση – Metabolomics.....	20
2.3	Εφαρμογές των σύγχρονων τεχνικών στην έρευνα και την κλινική πράξη	21
2.4	Προκλήσεις και προοπτικές της Μοριακής Βιολογίας: Ο ρόλος της διερευνητικής ανάλυσης δεδομένων.....	22
2.4.1	Υπολογιστικές Προκλήσεις στη Μοριακή Βιολογία.....	24
3	Διερευνητική Ανάλυση Δεδομένων: Μέθοδοι, Τεχνικές και Εφαρμογές....	26

3.1	Συλλογή δεδομένων.....	27
3.2	Προετοιμασία δεδομένων.....	29
3.3	Τεχνικές Διερευνητικής Ανάλυσης.....	30
3.3.1	Ανάλυση Κυρίων Συνιστωσών (PCA).....	31
3.3.2	Κατανομή Δεδομένων: Κυριότερες μέθοδοι.....	36
3.3.3	Συσταδοποίηση Δεδομένων.....	38
3.3.4	Άλλες σύγχρονες τεχνικές.....	40
3.4	Εργαλεία Οπτικοποίησης στη Διερευνητική Ανάλυση Δεδομένων.....	42
3.5	Εφαρμογή τεχνικών στα δεδομένα της Μοριακής Βιολογίας: Παραδείγματα εφαρμογών.....	44
3.6	Προβληματισμοί.....	46
4	Συζήτηση.....	48
4.1	Ταξινόμηση βιολογικών δεδομένων.....	48
4.2	Συσταδοποίηση βιολογικών δεδομένων.....	49
4.3	Προοπτικές και προκλήσεις για μελλοντική έρευνα.....	51
4.3.1	Εξόρυξη βιολογικών δεδομένων στο νέφος.....	51
4.3.2	Συμπεράσματα βιολογικού δικτύου με βάση το σύννεφο.....	55
4.3.3	Αλγόριθμος εξαγωγής συμπερασμάτων Γονιδιακού Ρυθμιστικού Δικτύου (GRN) με χρήση MI και GO.....	57
4.4	Συνεισφορά στον τομέα της Μοριακής Βιολογίας και ευρύτερα.....	59
4.4.1	Μια μελέτη περίπτωσης στην εξόρυξη δεδομένων δευτερογενούς δομής RNA.....	59
4.4.2	Συμπέρασμα.....	63

4.4.3	Μια μελέτη περίπτωσης στην εξόρυξη δεδομένων τριτογενών δομών RNA.....	65
4.4.4	Συμπέρασμα.....	66
5	Συμπεράσματα.....	69
5.1	Συνολική ανασκόπηση των ευρημάτων.....	69
5.2	Συνεισφορά της διερευνητικής ανάλυσης δεδομένων στη Μοριακή Βιολογία.....	70
5.3	Συνοπτική επισκόπηση των προτεινόμενων κατευθύνσεων για μελλοντική έρευνα.....	72
	Βιβλιογραφικές αναφορές.....	74

Κατάλογος Εικόνων

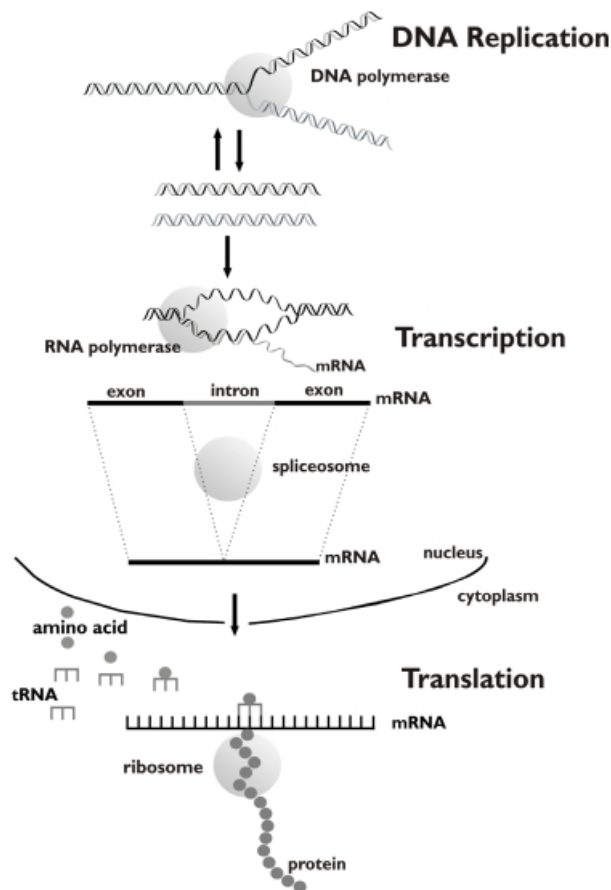
Εικόνα 1: Το κεντρικό δόγμα της μοριακής βιολογίας (Πηγή: Alberts, 2017).....	7
Εικόνα 2: Σχηματική απεικόνιση που δείχνει τη διαδικασία του scRNA-seq σε γοναδικούς ιστούς.....	17
Εικόνα 3: Διαδικασία του ChIP-seq.....	18
Εικόνα 4: Διαδικασία μείωσης διαστάσεων που εκτελείται με PCA.....	32
Εικόνα 5: t-SNE (t-Distributed Stochastic Neighbor Embedding).....	34
Εικόνα 6: UMAP (Uniform Manifold Approximation and Projection).....	35
Εικόνα 7: ISOMAP.....	36

1 Εισαγωγή

1.1 Εισαγωγή στη Μοριακή Βιολογία

Οι ζωντανοί οργανισμοί διέπονται από ένα σύνολο κληρονομικών οδηγιών που κωδικοποιούνται από το αλφάβητο των τεσσάρων γραμμάτων A, G, C και T. Τα "γράμματα" παίρνουν φυσική μορφή με την έννοια των τεσσάρων διαφορετικών νουκλεοτιδίων που αποτελούν τη βασική επαναλαμβανόμενη μονάδα των μορίων του δεοξυριβονουκλεϊκού οξέος (DNA). Κάθε νουκλεοτίδιο αποτελείται από ένα σάκχαρο 5 ατόμων άνθρακα με μια βάση αζώτου ομοιοπολικά συνδεδεμένη με το άτομο άνθρακα 1' και μια φωσφορική ομάδα ομοιοπολικά συνδεδεμένη με το άτομο άνθρακα 3' ή 5'. Ένα μόριο DNA είναι μια επαναλαμβανόμενη αλυσίδα νουκλεοτιδίων όπου κάθε φωσφορική ομάδα συνδέει το άτομο άνθρακα 3' του σακχάρου σε ένα νουκλεοτίδιο με το άτομο άνθρακα 5' του σακχάρου στο γειτονικό νουκλεοτίδιο. Υπάρχουν τέσσερις τύποι αζωτούχων βάσεων που καθορίζουν τα τέσσερα διαφορετικά νουκλεοτίδια στο DNA (αδενίνη (A), γουανίνη (G), κυτοσίνη (C) και θυμίνη (T)) και, ως εκ τούτου, κάθε μόριο DNA αντιπροσωπεύει μια μοναδική αλληλουχία αυτών των τεσσάρων χημικών "γραμμάτων" (Alberts, 2017).

Τα μόρια του DNA είναι επιπλέον δομικά οργανωμένα σε διπλά σύμπλοκα που αποτελούνται από δύο ελικοειδή μόρια DNA που συσπειρώνονται γύρω από έναν κοινό άξονα σχηματίζοντας μια διπλή έλικα. Οι δύο αλυσίδες της διπλής έλικας έχουν αντίθετες κατευθύνσεις για τη σύνδεση των ατόμων άνθρακα 3' με τα άτομα άνθρακα 5' (δηλαδή είναι αντι-παράλληλες) και συγκρατούνται μεταξύ τους με δεσμούς υδρογόνου μεταξύ αντίθετων βάσεων στις δύο αλυσίδες. Μια σημαντική ιδιότητα της διπλής έλικας είναι ότι οι δεσμοί υδρογόνου εμφανίζονται μόνο μεταξύ δύο συγκεκριμένων ζευγών βάσεων. Η A συνδέεται μόνο με την T και η C μόνο με την G. Αυτό σημαίνει ότι οι δύο αλυσίδες είναι συμπληρωματικές σε σχέση με την αλληλουχία που κωδικοποιούν, διευκολύνοντας έτσι βολικά σημαντικές διαδικασίες όπως η αντιγραφή και η μεταγραφή. Στους ευκαρυώτες, τα μόρια του DNA (το γονιδίωμα) είναι συστηματικά συσκευασμένα σε έναν αριθμό χρωμοσωμάτων που βρίσκονται στους πυρήνες κάθε κυττάρου (στα ζωικά κύτταρα ένα μικρό μέρος του DNA βρίσκεται στα μιτοχόνδρια). Ο πραγματικός αριθμός και το περιεχόμενο των χρωμοσωμάτων ποικίλλει από είδος σε είδος (Alberts, 2017).



Εικόνα 1: Το κεντρικό δόγμα της μοριακής βιολογίας (Πηγή: Alberts, 2017).

Το κεντρικό δόγμα της μοριακής βιολογίας αναφέρει ότι η γενετική πληροφορία που είναι σκληρά συνδεδεμένη στο DNA μεταγράφεται σε φορητά μόρια αγγελιοφόρου ριβονουκλεϊκού οξέος (mRNA), τα οποία στη συνέχεια μεταφράζονται σε πρωτεΐνες. Εκτός από την ουρακίλη (U) που αντικαθιστά τη θυμίνη (T) στην αλληλουχία του mRNA, ένα μόριο mRNA είναι ακριβές αντίγραφο ενός τμήματος μιας αλυσίδας του DNA και μεταφέρει τις απαραίτητες πληροφορίες για τη σύνθεση μιας ή ενός μικρού αριθμού πρωτεϊνών. Ενώ το DNA μπορεί να θεωρηθεί ως συσκευή αποθήκευσης γενετικών οδηγιών, οι πρωτεΐνες στην πραγματικότητα εκτελούν αυτές τις οδηγίες ως ένζυμα, υποδοχείς, πρωτεΐνες αποθήκευσης, πρωτεΐνες μεταφοράς, παράγοντες μεταγραφής, μόρια σηματοδότησης, ορμόνες κ.λπ. Εξάιρεση αποτελούν ορισμένα RNA που δεν μεταφράζονται σε πρωτεΐνες και τα οποία εκτελούν λειτουργίες άμεσα (tRNA, rRNA και snRNA).

Τα τμήματα του DNA που κωδικοποιούν RNA ονομάζονται γονίδια. Η μεταγραφή των γονιδίων σε RNA πραγματοποιείται από τα ένζυμα της RNA πολυμεράσης χρησιμοποιώντας ως πρότυπο μία από τις αλυσίδες του DNA. Το δίκλωνο DNA ξετυλίγεται κατά τη διάρκεια της μεταγραφής, έτσι ώστε η αλυσίδα που λειτουργεί ως πρότυπο για τη σύνθεση του RNA να μπορεί να σχηματίσει υβρίδιο με το νέο, αναπτυσσόμενο RNA. Κατά συνέπεια, το μεταγραφόμενο RNA είναι μια μονόκλωνη αλληλουχία συμπληρωματική προς την πρότυπη αλυσίδα και πανομοιότυπη προς την αλυσίδα DNA που δεν λειτουργεί ως πρότυπη (εκτός από το ότι το U αντικαθιστά το T) (Lander et al., 2001).

Τα γονίδια λέγεται ότι εκφράζονται σε ένα κύτταρο εάν μεταγράφονται. Η ικανότητα διαφορετικής έκφρασης των γονιδίων σε διαφορετικούς κυτταρικούς τύπους, στάδια του κυτταρικού κύκλου και κάτω από διάφορες αλλαγές στο περιβάλλον αποτελούν ένα σημαντικό επίπεδο δυναμικής στους κυτταρικούς οργανισμούς (ένα άλλο επίπεδο μοριακής δυναμικής είναι αυτό των πρωτεϊνών και των αλληλεπιδράσεών τους μεταξύ τους και με άλλα μόρια). Ένας αριθμός παραγόντων είναι σημαντικός για τη διαφορετική έκφραση ενός συγκεκριμένου γονιδίου σε διαφορετικά κύτταρα, συμπεριλαμβανομένου του ρυθμού μεταγραφής, του ρυθμού μετάφρασης και της σταθερότητας της πρωτεΐνης. Ωστόσο, ο σημαντικότερος παράγοντας είναι η έναρξη της πραγματικής μεταγραφής. Στους ευκαρυώτες, η μεταγραφή δεν εκκινείται από την πολυμεράση RNA, αλλά από έναν αριθμό ρυθμιστικών πρωτεϊνών που ονομάζονται μεταγραφικοί παράγοντες, οι οποίες συνδέονται με το DNA και τόσο ενεργοποιούν όσο και καθοδηγούν την πολυμεράση. Η ικανότητα αυτών των μεταγραφικών παραγόντων να αναγνωρίζουν επιλεκτικά συγκεκριμένα στοιχεία βραχείας αλληλουχίας στο DNA είναι επομένως σημαντική για τη ρύθμιση της γονιδιακής έκφρασης (δηλαδή τη γονιδιακή ρύθμιση). Πολλά από αυτά τα ρυθμιστικά στοιχεία ή θέσεις πρόσδεσης βρίσκονται σε μια περιοχή που ονομάζεται υποκινητής και βρίσκεται ανάντη της κωδικοποιητικής αλληλουχίας (τα ανάντη και τα κατάντη αναφέρονται στις αλληλουχίες που πλαισιώνουν ένα συγκεκριμένο γονίδιο στα 5' και 3' άκρα, αντίστοιχα).

Τα περισσότερα ευκαρυωτικά μεταγράμματα RNA περνούν από διάφορα στάδια προεπεξεργασίας, συμπεριλαμβανομένης της αφαίρεσης ορισμένων τμημάτων εντός του γονιδίου και της συγχώνευσης των υπόλοιπων τμημάτων (ματίσματος RNA). Αυτό

οφείλεται στην εσωτερική δομή των γονιδίων η οποία αποτελείται από κωδικοποιητικά τμήματα που ονομάζονται εξόνια και διαχωρίζονται από μη κωδικοποιητικές περιοχές που ονομάζονται ιντρόνια. Παρόλο που και τα δύο τμήματα μεταγράφονται, τα ιντρόνια αφαιρούνται αργότερα από ένα μεγάλο σύμπλοκο (το συγκολλητόσωμα) που αποτελείται από πέντε τύπους μικρών πυρηνικών RNAs (snRNAs) και πρωτεΐνες. Νεότερες μελέτες δείχνουν ότι τα εξόνια σε πολύπλοκους οργανισμούς όπως ο άνθρωπος ματίζονται με διαφορετικούς τρόπους, σχηματίζοντας διαφορετικές παραλλαγές ματίσματος και συνεπώς διαφορετικά πρωτεϊνικά προϊόντα από το ίδιο γονίδιο (Lander et al., 2001).

Η σύνθεση των πρωτεϊνών από το mRNA λαμβάνει χώρα στα ριβοσώματα που λειτουργούν ως δομικά πλαίσια για τη μετάφραση. Τα ριβοσώματα είναι μεγάλα σύμπλοκα RNA-πρωτεΐνης που αποτελούνται από έναν αριθμό ριβοσωμικών RNAs (rRNAs) και πρωτεΐνες. Τα βασικά δομικά στοιχεία των πρωτεϊνών είναι τα αμινοξέα. Υπάρχουν 20 αμινοξέα, τα οποία αποτελούνται όλα από ένα άτομο άνθρακα (C_α) συνδεδεμένο με μια αμινομάδα (NH_2), μια καρβοξυλομάδα ($COOH$), ένα άτομο υδρογόνου (H) και μια μεταβλητή ομάδα που καθορίζει τα 20 διαφορετικά αμινοξέα (τις πλευρικές αλυσίδες). Οι πρωτεΐνες είναι απλώς γραμμικές, μη διακλαδισμένες αλυσίδες αμινοξέων όπου η αμινομάδα ενός αμινοξέος σχηματίζει πεπτιδικό δεσμό με την καρβοξυλομάδα του γειτονικού αμινοξέος. Η επαναλαμβανόμενη αλυσίδα χωρίς τις μεταβλητές πλευρικές αλυσίδες ονομάζεται κύρια αλυσίδα ή ραχοκοκαλιά του πρωτεϊνικού μορίου. Οι πρωτεΐνες κωδικοποιούνται απευθείας στην αλληλουχία του mRNA με όρους διαδοχικών ομάδων τριών νουκλεοτιδίων (κωδικόνια). Δεδομένου ότι υπάρχουν τέσσερις διαφορετικές βάσεις στο RNA (και στο DNA) και τρεις θέσεις βάσεων σε ένα κωδικόνιο, υπάρχουν $4^3=64$ πιθανοί συνδυασμοί για την κωδικοποίηση 20 αμινοξέων. Ως εκ τούτου, κάθε αμινοξύ προσδιορίζεται κατά μέσο όρο από περίπου τρία διαφορετικά κωδικόνια (ο γενετικός κώδικας λέγεται εκφυλισμένος). Τα mRNA μεταφράζονται σε αλυσίδα αμινοξέων με τη βοήθεια των μεταφορικών RNA (tRNA). Υπάρχει ένα tRNA ανά αμινοξύ, ικανό να δεσμεύει και να μεταφέρει το συγκεκριμένο αμινοξύ. Κάθε tRNA περιλαμβάνει επίσης μια ειδική αλληλουχία (αντικωδικόνιο) που αναγνωρίζει το σχετικό κωδικόνιο στην αλληλουχία του mRNA, ώστε το σωστό αμινοξύ να μπορεί να εισαχθεί στην αναπτυσσόμενη αλυσίδα αμινοξέων (Leviton & Kaczmarek, 2015).

Μια σημαντική αρχή στη μοριακή βιολογία είναι ότι η αλληλουχία αμινοξέων μιας πρωτεΐνης καθορίζει το τρισδιάστατο σχήμα της (δηλαδή τη δομή της) και επιπλέον ότι η δομή μιας πρωτεΐνης καθορίζει τη λειτουργία της. Εφόσον η αλληλουχία των αμινοξέων κωδικοποιείται στο DNA, προκύπτει ότι οι μηχανισμοί της εξέλιξης (δηλαδή η μετάλλαξη και η διασταύρωση) συμβάλλουν σχεδόν άμεσα στην αλλαγή της λειτουργίας των πρωτεϊνών. Για να προσαρμόζονται σε διαφορετικές τρισδιάστατες διαμορφώσεις, τα 20 αμινοξέα ποικίλλουν ως προς το σχήμα, το φορτίο, την υδροφοβικότητα και την αντιδραστικότητα. Για παράδειγμα, τα υδρόφοβα αμινοξέα τείνουν να είναι θαμμένα στο εσωτερικό της πρωτεΐνης (όπου προστατεύονται από το νερό που περιβάλλει την πρωτεΐνη), ενώ τα υδρόφιλα αμινοξέα τείνουν να βρίσκονται στην επιφάνεια της πρωτεΐνης.

Η δομή των πρωτεϊνών είναι πιο πολύπλοκη από τη διπλή έλικα του DNA (βλ. Σχήμα 2 για παράδειγμα) και μπορεί να οργανωθεί σε τέσσερα επίπεδα. Η ίδια η αλληλουχία των αμινοξέων αναφέρεται ως πρωτογενής δομή. Όταν είναι σταθερή, η κύρια αλυσίδα της πρωτεΐνης αναδιπλώνεται είτε σε μια έλικα α (δηλ. μια σπειροειδή δομή), είτε σε ένα φύλλο β (δηλ. μια επίπεδη δομή περισσότερων του ενός κλώνων β) είτε σε μια σπείρα (δηλ. μια τυχαία δομή) (βλ. Εικόνα 2α). Αυτές οι επιβεβαιώσεις αποτελούν τη δευτερογενή δομή των πρωτεϊνών. Επιπλέον, τα στοιχεία της δευτεροταγούς δομής (φύλλα και έλικες) τείνουν να σχηματίζουν απλά μοτίβα που συνδέονται με σύντομες στροφές σχήματος U ή βρόχους που συχνά βρίσκονται στην επιφάνεια της πρωτεΐνης (π.χ. το κοινό μοτίβο φουρκέτας β που αποτελείται από δύο γειτονικά σκέλη β που ενώνονται με βρόχο). Αρκετά μοτίβα σχηματίζουν συμπαγείς σφαιρικές περιοχές που αναφέρονται ως τριτοταγής δομή των πρωτεϊνών. Ενώ η δευτεροταγής δομή σταθεροποιείται από δεσμούς υδρογόνου μεταξύ ορισμένων πλευρικών αλυσίδων, η τριτοταγής δομή σταθεροποιείται κυρίως από υδρόφοβες αλληλεπιδράσεις. Τέλος, ορισμένες πρωτεΐνες αποτελούνται από πολλές αλυσίδες αμινοξέων (που ονομάζονται επίσης υπομονάδες) και οι διατάξεις τους αναφέρονται ως τεταρτοταγής δομή των πρωτεϊνών. Όπως είδαμε ήδη με το συγκολλητόσωμα και το ριβόσωμα, οι πρωτεΐνες συχνά λειτουργούν σε μεγάλα σύμπλοκα στα οποία συμμετέχουν πολλές πρωτεΐνες και ενδεχομένως άλλα μακρομόρια (Levitan & Kaczmarek, 2015).

1.2 Σημασία της διερευνητικής ανάλυσης δεδομένων στη Μοριακή Βιολογία

Η εξερευνητική ανάλυση δεδομένων αναδεικνύει την ουσιαστική της σημασία στη Μοριακή Βιολογία, λειτουργώντας ως κρίσιμο εργαλείο για την ανακάλυψη νέων πτυχών και την εμβάθυνση στην κατανόηση της λειτουργίας των βιολογικών μορίων. Συγκεκριμένα, η Μοριακή Βιολογία συχνά αντιμετωπίζει τεράστια όγκο δεδομένων που προέρχονται από γονίδια, πρωτεΐνες, μεταβολίτες και άλλα βιολογικά μόρια. Η εξερευνητική ανάλυση δεδομένων σε αυτό το πλαίσιο μπορεί να αναδείξει μοτίβα και τάσεις που κρύβονται σε αυτά τα δεδομένα, επιτρέποντας την ανάπτυξη νέων υποθέσεων και την ανακάλυψη νέων πληροφοριών για τη λειτουργία και την αλληλεπίδραση των βιολογικών μορίων (Wilson et al., 2018).

Επιπλέον, η ανάλυση δεδομένων συμβάλλει στην κατανόηση παθολογικών διαδικασιών. Ενδεικτικά, μπορεί να αναγνωρίσει μοτίβα και παραμέτρους που σχετίζονται με διάφορες παθολογίες, προσφέροντας έτσι τη δυνατότητα ανακάλυψης νέων βιολογικών μαρκέρ και την προώθηση της κατανόησης της παθοφυσιολογίας των νοσημάτων (Wilson et al., 2018).

Επιπλέον, η εξερευνητική ανάλυση δεδομένων επιτρέπει την επαλήθευση ή ακόμα και την αντικατάσταση υποθέσεων που έχουν διατυπωθεί από επιστήμονες, με βάση την ανάλυση των πραγματικών δεδομένων που διατίθενται. Αυτό οδηγεί σε μια πιο ακριβή κατανόηση των βιολογικών διεργασιών και φαινομένων, βοηθώντας στην αποφυγή λανθασμένων συμπερασμάτων και προάγοντας την πρόοδο της επιστήμης (Wilson et al., 2018).

Τέλος, η ανάλυση δεδομένων καθοδηγεί τη στρατηγική σχεδίαση πειραμάτων, παρέχοντας πολύτιμες οδηγίες για την επιλογή κατάλληλων μεθόδων και προσεγγίσεων. Αυτό ενισχύει την αποτελεσματικότητα της έρευνας και βελτιώνει τις πιθανότητες επιτυχίας στην ανακάλυψη και την επεξήγηση των βιολογικών φαινομένων που μελετώνται (Wilson et al., 2018).

Συνολικά, η εξερευνητική ανάλυση δεδομένων αναδεικνύεται ως κρίσιμο εργαλείο για τη Μοριακή Βιολογία, προωθώντας την κατανόηση, την ανακάλυψη νέων γνώσεων και τη δημιουργία υποθέσεων για μελλοντική έρευνα στον τομέα (Wilson et al., 2018).

1.3 Σκοπός και δομή της εργασίας

Η Μοριακή Βιολογία βρίσκεται στην πρωτοπορία της επιστήμης, επιτρέποντας την κατανόηση των βιολογικών διεργασιών σε μοριακό επίπεδο. Η εξέλιξη στις τεχνολογίες συλλογής δεδομένων έχει οδηγήσει σε μια άνευ προηγουμένου ποσότητα πληροφοριών, απαιτώντας σύγχρονες τεχνικές ανάλυσης για την εξαγωγή σημαντικών ερευνητικών συμπερασμάτων. Η παρούσα μελέτη επικεντρώνεται στην εφαρμογή της διερευνητικής ανάλυσης δεδομένων μεγάλου όγκου μέσω σύγχρονων τεχνικών της Μοριακής Βιολογίας.

Αρχικά, θα γίνει συλλογή δεδομένων από διάφορες πηγές, συμπεριλαμβανομένων γονιδιωματικών ακολουθιών, πρωτεϊνικών δομών και μεταβολομικών προφίλ. Στη συνέχεια, θα χρησιμοποιηθούν σύγχρονες τεχνικές ανάλυσης όπως η ανάλυση PCA (Ανάλυση Κύριων Συνιστωσών) και η ιεραρχική συσταδοποίηση για την εξαγωγή μοτίβων και την ανάδειξη δομικών συνδέσεων μεταξύ των δεδομένων.

Η ανάλυση των δεδομένων αυτών θα συμβάλλει στην προσδιορισμό νέων υποψήφιων μορίων-στόχων για πειραματική επαλήθευση και την κατανόηση πιθανών βιολογικών μηχανισμών.

Η διερευνητική ανάλυση δεδομένων αποτελεί ισχυρό εργαλείο για την αντιμετώπιση της πληθώρας των πληροφοριών που παρέχονται από τις σύγχρονες τεχνολογίες της Μοριακής Βιολογίας. Η συνεχής εξέλιξη σε αυτόν τον τομέα απαιτεί σταθερή ενημέρωση και προσαρμογή των μεθόδων ανάλυσης για την αποτελεσματική εξαγωγή γνώσης.

2 Ανασκόπηση

2.1 Εξέλιξη της Μοριακής Βιολογίας: Σημαντικά επιτεύγματα και τεχνολογικές προόδους

Η εξέλιξη της Μοριακής Βιολογίας έχει σημαντικές διαστάσεις, συμπεριλαμβανομένων των επιτευγμάτων και των τεχνολογικών προόδων που έχουν σημειωθεί στον τομέα. Αυτή η εξέλιξη έχει επηρεάσει τον τρόπο που κατανοούμε και αντιμετωπίζουμε τα βιολογικά συστήματα, καθώς και τον τρόπο που εφαρμόζουμε αυτές τις γνώσεις σε διάφορους τομείς όπως η ιατρική, η γεωργία, η προστασία του περιβάλλοντος και η βιομηχανία (Baxevanis et al., 2020).

Μερικά σημαντικά επιτεύγματα και τεχνολογικές προόδους περιλαμβάνουν (Baxevanis et al., 2020):

1. **Αποκωδικοποίηση του ανθρώπινου γονιδιώματος:** Το έργο του ανθρώπινου γονιδιώματος (Human Genome Project) ήταν μια σημαντική προσπάθεια που οδήγησε στην αποκωδικοποίηση του γενετικού υλικού του ανθρώπου. Αυτή η προσπάθεια έδωσε μια σημαντική πηγή δεδομένων για την κατανόηση της ανθρώπινης γενετικής ποικιλότητας και της σχέσης του γονιδίου με την ανάπτυξη της νόσου.
2. **Ανάπτυξη νέων τεχνικών ανάλυσης DNA:** Η εξέλιξη των τεχνολογιών ανάλυσης DNA, όπως η υψηλής θέσης ακολουθιακής, έχει επιταχύνει τη δυνατότητά μας να αναλύουμε γονίδια και γενετικές αλληλουχίες με μεγάλη ακρίβεια και ταχύτητα.
3. **Πρόοδος στην ανάπτυξη της γενετικής επεξεργασίας:** Η τεχνική των CRISPR/Cas9 έχει ανοίξει νέους ορίζοντες στη γενετική επεξεργασία, επιτρέποντας την επεξεργασία του γονιδιώματος με μεγάλη ακρίβεια και αποτελεσματικότητα.
4. **Ανάπτυξη της βιοπληροφορικής:** Η βιοπληροφορική έχει παίξει καθοριστικό ρόλο στην ανάλυση και την ερμηνεία των γενετικών δεδομένων, επιτρέποντας την ανάπτυξη νέων αλγορίθμων και εργαλείων για την ανάλυση των ακολουθιών DNA και την πρόβλεψη της δομής και λειτουργίας των πρωτεϊνών.

Αυτά τα επιτεύγματα αποτελούν μόνο μερικά παραδείγματα του πώς η Μοριακή Βιολογία έχει προχωρήσει μπροστά και πώς οι τεχνολογικές πρόοδοι συμβάλλουν στην κατανόηση της ζωής σε μοριακό επίπεδο.

2.2 Σύγχρονες τεχνικές στη Μοριακή Βιολογία

Η ανάπτυξη της γονιδιωματικής και των πειραματικών τεχνολογιών υψηλής απόδοσης δημιούργησε την ανάγκη για υπολογιστές να αποθηκεύουν και να αναλύουν μεγάλες ποσότητες δεδομένων. Όπως συνέβη στην περίπτωση της γονιδιωματικής, η βιοπληροφορική αναπτύχθηκε από έναν κλάδο που σχετίζεται κυρίως με βάσεις δεδομένων αλληλουχιών και ανάλυση ακολουθιών σε μια υπολογιστική επιστήμη που χρησιμοποιεί βιολογικά δεδομένα για να κάνει π.χ. λειτουργική γονιδιωματική. Αν και υπάρχουν διαφορετικοί ορισμοί και απόψεις για τη βιοπληροφορική, οι περισσότεροι ερευνητές χρησιμοποιούν πλέον τη βιοπληροφορική ως γενικό όρο τόσο για την αποθήκευση και συντήρηση βιολογικών δεδομένων όσο και για τη χρήση μεθόδων και αλγορίθμων υπολογιστικής ανάλυσης δεδομένων σε μελέτες λειτουργικής γονιδιωματικής (Kanehisa & Bork, 2003).

Η βιοπληροφορική περιλαμβάνει επομένως μια σειρά από επιστημονικά πεδία, συμπεριλαμβανομένων των μαθηματικών, της στατιστικής, της πληροφορικής, της φυσικής, της χημείας, της βιολογίας και της ιατρικής.

Μια ευρέως χρησιμοποιούμενη μεθοδολογία στη βιοπληροφορική και στη λειτουργική γονιδιωματική είναι αυτή της μηχανικής μάθησης. Η μηχανική μάθηση αντιμετωπίζει το πρόβλημα της χρήσης υπολογιστών για την εκμάθηση γενικών εννοιών από παρατηρήσεις και γνώσεις και έχει αναπτυχθεί παραδοσιακά σε δύο διαφορετικά σχολεία. Οι στατιστικολόγοι αναπτύσσουν μεθόδους μάθησης που βασίζονται στα μαθηματικά πλαίσια της θεωρίας πιθανοτήτων και της στατιστικής. Οι επιστήμονες υπολογιστών συχνά αναπτύσσουν μεθόδους που βασίζονται σε μοντέλα ευφών συστημάτων (π.χ. μέθοδοι εμπνευσμένες από τη βιολογία, όπως γενετικοί αλγόριθμοι και νευρωνικά δίκτυα, ή μέθοδοι που βασίζονται στη λογική, όπως η εκμάθηση κανόνων) (Kanehisa & Bork, 2003).

Οι διαφορές οφείλονται κυρίως στο γεγονός ότι οι στατιστικολόγοι ενδιαφέρθηκαν κυρίως για την καθαρή ανάλυση δεδομένων, ενώ οι επιστήμονες υπολογιστών ενδιαφέρθηκαν επίσης για την κατασκευή ευφών συστημάτων (π.χ. ρομπότ με τεχνητή νοημοσύνη). Ωστόσο, αυτές οι διαφορετικές απόψεις συγκλίνουν κάπως, σχηματίζοντας υβρίδια χρησιμοποιώντας στοιχεία τόσο από τη στατιστική όσο και από την επιστήμη των υπολογιστών (π.χ. αναγνώριση προτύπων).

Η επαγωγή αναφέρεται στη γενίκευση από τις παρατηρήσεις σε ευρείες έννοιες και διαφέρει από την εξαγωγή που αναφέρεται στη χρήση γενικών εννοιών (ή θεωριών) για να συναχθούν συγκεκριμένες υποθέσεις. Στη μοριακή βιολογία, η επαγωγή είναι ιδιαίτερα σημαντική αφού οι γενικές θεωρίες δεν έχουν ακόμη επεξεργαστεί. Για παράδειγμα, γνωρίζουμε ότι υπάρχει μια σχέση μεταξύ αλληλουχίας και δομής, αλλά αυτή η σχέση δεν είναι καλά κατανοητή από την άποψη των θεωριών που μπορούν να χρησιμοποιηθούν για την εξαγωγή καλών δομικών μοντέλων για μια συγκεκριμένη αλληλουχία πρωτεΐνης. Ωστόσο, έχουμε παραδείγματα αυτής της σχέσης όσον αφορά τις δομές πρωτεΐνης που επιλύονται πειραματικά. Και οι μέθοδοι μηχανικής μάθησης έχουν σχεδιαστεί για να προκαλούν μοντέλα που βασίζονται σε παραδείγματα, τα οποία περιγράφουν εν μέρει την υποτιθέμενη υποκείμενη λειτουργική σχέση μεταξύ, σε αυτήν την περίπτωση, της ακολουθίας και της δομής. Η πιο κοινή εφαρμογή τέτοιων μοντέλων είναι αυτή της πρόβλεψης. Ωστόσο, δεδομένου ενός μοντέλου που μπορεί να προβλέψει αξιόπιστα τη δομή της πρωτεΐνης από την αλληλουχία (ιδίως για πρωτεΐνες που δεν εμφανίζονται, δηλ. πρωτεΐνες που δεν ήταν διαθέσιμες όταν προκλήθηκε το μοντέλο), αυτό το μοντέλο περιλαμβάνει προφανώς γενικές έννοιες που μπορούν επίσης να χρησιμοποιηθούν για την κατανόηση της σχέσης. Και αυτή η κατανόηση μπορεί με τον καιρό να οδηγήσει σε γενικές θεωρίες. Κατά συνέπεια, η μηχανική μάθηση μπορεί να χρησιμοποιηθεί τόσο για προγνωστικούς όσο και για περιγραφικούς σκοπούς (Voet et al., 2016).

Στη μοριακή βιολογία, και ιδιαίτερα στη λειτουργική γονιδιωματική, θα δούμε ότι μια σειρά προβλημάτων μπορούν να αντιμετωπιστούν χρησιμοποιώντας τις έννοιες των παραδειγμάτων και της μηχανικής μάθησης. Και η επιτυχής εφαρμογή τέτοιων μεθόδων θα μπορούσε να οδηγήσει σε καταστάσεις όπου τα βιολογικά πειράματα χρησιμοποιούνται για τη λήψη πληροφοριών για ένα (αντιπροσωπευτικό) σύνολο περιπτώσεων, τα μοντέλα δημιουργούνται αυτόματα από αυτά τα παραδείγματα και

τελικά χρησιμοποιούνται για τη συμπλήρωση της γνώσης που λείπει για τις υπόλοιπες περιπτώσεις. Αυτή είναι η φιλοσοφία της δομικής γονιδιωματικής που αναφέρθηκε προηγουμένως: να λυθεί πειραματικά η δομή τουλάχιστον μίας πρωτεΐνης από κάθε δομική τάξη (π.χ. διπλώστε) και να προβλέψετε τη δομή των υπόλοιπων πρωτεϊνών χρησιμοποιώντας ομοιότητα αλληλουχίας με πρωτεΐνες με λυμένες δομές.

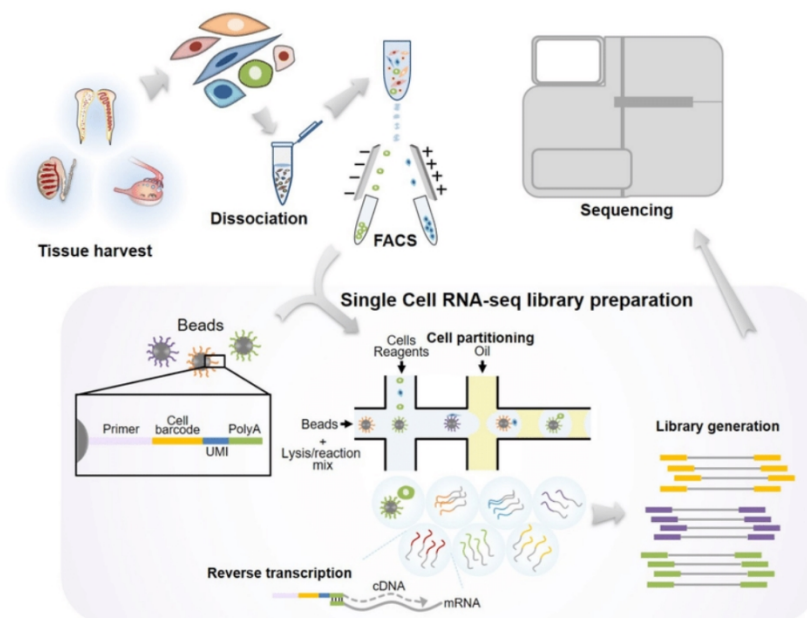
Ένα από τα σημαντικότερα εμπόδια για την αποτελεσματική χρήση της μηχανικής μάθησης στη λειτουργική γονιδιωματική ήταν η έλλειψη δομής στην υπάρχουσα βιολογική γνώση όσον αφορά τις αναγνώσιμες βάσεις δεδομένων και τους σχολιασμούς από υπολογιστή. Η εξόρυξη κειμένου και η αυτόματη εξαγωγή συμπερασμάτων από το ελεύθερο κείμενο ήταν επομένως ένα σημαντικό μέρος της βιοπληροφορικής και θα συνεχίσει να είναι (Voet et al., 2016).

2.2.1 Single-cell RNA-seq sequencing

Το Single-cell RNA sequencing (scRNA-seq) είναι μια προηγμένη τεχνική στη μοριακή βιολογία που επιτρέπει την ανάλυση του μεταγραφικού προφίλ σε επίπεδο μονάδας κυττάρων. Σε αντίθεση με τις παραδοσιακές μεθόδους RNA sequencing, η scRNA-seq προσφέρει τη δυνατότητα ανάλυσης της εκφραστικής ποικιλομορφίας σε ατομικό κυτταρικό επίπεδο, αναδεικνύοντας έτσι την ετερογένεια μέσα σε κυτταρικούς πληθυσμούς και αποκαλύπτοντας νέους υποτύπους κυττάρων ή κατανοώντας τις αλληλεπιδράσεις μεταξύ κυττάρων σε πολυκυτταρικά συστήματα (Chen, Ning & Shi, 2019).

Η διαδικασία διερεύνησης των δεδομένων scRNA-seq περιλαμβάνει αρχικά τον καθορισμό των δεδομένων, όπου τα scRNA-seq δεδομένα ελέγχονται για ποιότητα και καθαρίζονται. Στο στάδιο αυτό, τα RNA από μονοκύτταρα υποβάλλονται στην απαραίτητη προετοιμασία, η οποία περιλαμβάνει την εκχύλιση του RNA και την καταμέτρηση των κυττάρων. Στη συνέχεια, πραγματοποιείται η εξόρυξη χαρακτηριστικών, όπου αναλύονται τα δεδομένα για την ανίχνευση της εκφραστικής δραστηριότητας του RNA σε επίπεδο μονάδας κυττάρων. Αυτή η ανάλυση περιλαμβάνει την καταμέτρηση των μορίων RNA και τη χρήση μεθόδων όπως η Principal Component Analysis (PCA) ή η t-distributed Stochastic Neighbor Embedding

(t-SNE) για τη μείωση της διάστασης των δεδομένων και τον εντοπισμό πιθανών υποτύπων κυττάρων.



Εικόνα 2: Σχηματική απεικόνιση που δείχνει τη διαδικασία του scRNA-seq σε γοναδικούς ιστούς.

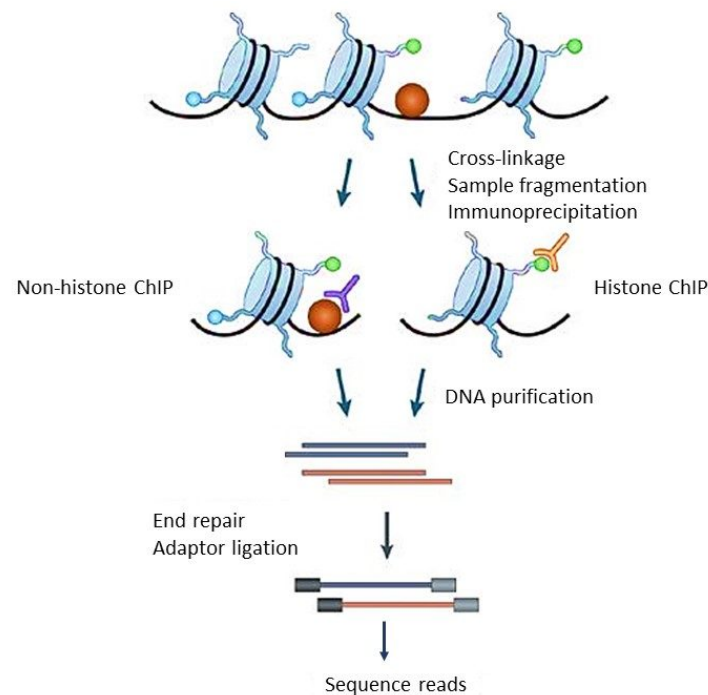
Ακολουθεί η ανάλυση διαφορικής έκφρασης γονιδίων, όπου συγκρίνονται τα επίπεδα έκφρασης μεταξύ διαφορετικών υποτύπων κυττάρων για να αποκαλυφθούν τα γονίδια που είναι διαφορεικά εκφρασμένα μεταξύ αυτών των υποτύπων, και συνεπώς πιθανώς υπεύθυνα για τη διαφοροποίησή τους. Τέλος, η κατηγοριοποίηση των κυττάρων επιτρέπει την ταξινόμηση των εντοπισμένων υποτύπων με βάση τις λειτουργικές τους ιδιότητες, την εξέλιξή τους ή το περιβάλλον τους.

Η scRNA-seq αναδεικνύει την ετερογένεια και την πολυπλοκότητα εντός των κυττάρων και έχει εφαρμογές σε πολλούς τομείς της βιολογίας, όπως η ανάπτυξη, η νευροβιολογία, η καρκινογένεση και η ανοσολογία (Haque et al., 2017).

2.2.2 ChIP-seq (Chromatin Immunoprecipitation Sequencing)

Το ChIP-seq (Chromatin Immunoprecipitation Sequencing) είναι μια προηγμένη τεχνική στη μοριακή βιολογία που χρησιμοποιείται για τη μελέτη των

αλληλεπιδράσεων μεταξύ πρωτεϊνών και DNA σε επίπεδο γονιδίων. Αυτή η τεχνική επιτρέπει στους ερευνητές να ανιχνεύσουν τις περιοχές του γονιδιώματος στις οποίες συγκεκριμένες πρωτεΐνες συνδέονται, προσφέροντας σημαντικές πληροφορίες για τη ρύθμιση της γονιδιακής έκφρασης (Nakato & Sakata, 2021).



Εικόνα 3: Διαδικασία του ChIP-seq.

Η διαδικασία ChIP-seq ξεκινά με την εκχύλιση των πυρήνων από κύτταρα που έχουν υποστεί επεξεργασία, διατηρώντας έτσι τις πρωτεΐνες συνδεδεμένες με το DNA. Στη συνέχεια, οι πρωτεΐνες που πιθανώς συνδέονται με το DNA σταθεροποιούνται με τη χρήση αντισωμάτων ή άλλων ανιχνευτικών μεθόδων. Το αντίσωμα συνδέεται με τις πρωτεΐνες που είναι δεσμευμένες στο DNA, σχηματίζοντας ένα σύμπλοκο πρωτεΐνης-αντισώματος-γονιδίου. Ακολούθως, το DNA εξάγεται από αυτό το σύμπλοκο. Το εξαγόμενο DNA στη συνέχεια αλληλουχείται με τη χρήση τεχνολογιών αλληλούχισης επόμενης γενιάς (NGS). Τα δεδομένα που προκύπτουν αναλύονται υπολογιστικά για να εντοπιστούν οι περιοχές του γονιδιώματος όπου συνδέονται οι πρωτεΐνες.

Το ChIP-seq έχει εφαρμογές σε πολλούς τομείς της μοριακής βιολογίας. Για παράδειγμα, μπορεί να χρησιμοποιηθεί για τον εντοπισμό των γονιδιακών περιοχών που είναι υπεύθυνες για τη ρύθμιση της γονιδιακής έκφρασης, τον εντοπισμό σημάτων

μετάδοσης που σχετίζονται με την ενεργοποίηση ή την απενεργοποίηση γονιδίων, καθώς και για την κατανόηση της δομής και της λειτουργίας του χρωματίνικου DNA (Nakato & Shirahige, 2017).

2.2.3 Ανάλυση της μεθυλίωσης του DNA - DNA methylation analysis

Η ανάλυση της μεθυλίωσης του DNA είναι μια τεχνική που χρησιμοποιείται για να μετρηθεί η μεθυλίωση των νουκλεοτιδίων στο γονιδίωμα ενός οργανισμού. Αυτή η διαδικασία είναι σημαντική γιατί η μεθυλίωση του DNA μπορεί να επηρεάσει την έκφραση των γονιδίων και να διαδραματίσει κρίσιμο ρόλο σε πολλές βιολογικές διαδικασίες (Kurdyukov & Bullock, 2016).

Η διαδικασία ανάλυσης της μεθυλίωσης του DNA ξεκινά με την εξαγωγή του DNA από τα δείγματα βιολογικού υλικού, όπως το αίμα, οι ιστοί ή τα κύτταρα. Στη συνέχεια, το DNA υποβάλλεται σε χημικές αντιδράσεις που αφαιρούν τις μεθυλικές ομάδες από τα νουκλεοτίδια. Ακολουθεί η ανάλυση της μεθυλίωσης, η οποία γίνεται συνήθως με τη χρήση διαφόρων τεχνικών, όπως το bisulfite sequencing ή η μαζική ακολουθιασμολογία (massively parallel sequencing). Τέλος, τα δεδομένα που προκύπτουν αναλύονται στατιστικά και υπολογιστικά για να ανακαλυφθούν οι μεθυλιωμένες περιοχές του γονιδιώματος.

Η ανάλυση της μεθυλίωσης του DNA έχει πολλές εφαρμογές στη βιολογία και την ιατρική. Για παράδειγμα, στη μελέτη του καρκίνου, η ανάλυση μπορεί να αναδείξει περιοχές του γονιδιώματος που συμβάλλουν στην κακοήθεια και την πρόοδο του καρκίνου. Στον τομέα των νευροεπιστημών, η μεθυλίωση του DNA παίζει σημαντικό ρόλο στην ανάπτυξη και λειτουργία του εγκεφάλου, καθώς και σε νευρολογικές διαταραχές. Επιπλέον, στην αναπαραγωγική υγεία, η μεθυλίωση μπορεί να επηρεάσει την έκφραση γονιδίων που σχετίζονται με την αναπαραγωγή και την υγεία των απογόνων.

Γενικά, η ανάλυση της μεθυλίωσης του DNA παρέχει πολύτιμες πληροφορίες για την κατανόηση της λειτουργίας του γονιδιώματος, την ανάπτυξη ασθενειών και την εξέλιξη νέων θεραπευτικών και διαγνωστικών μεθόδων (Feng & Lou, 2019).

2.2.4 Μεταβολομική ανάλυση – Metabolomics

Η μεταβολομική ανάλυση, γνωστή και ως μεταβολομική προφίλωση, είναι μια τεχνική που μετρά και αναλύει τους μεταβολίτες ενός οργανισμού. Οι μεταβολίτες, οι οποίοι είναι μικροί οργανικοί πυρήνες, εμπλέκονται σε βιοχημικές αντιδράσεις και παρέχουν πολύτιμες πληροφορίες για την κατάσταση της υγείας ενός οργανισμού, το μεταβολικό του προφίλ και την αντίδρασή του σε εξωτερικούς παράγοντες (Nalbantoglu, 2019).

Η διαδικασία της μεταβολομικής ανάλυσης ξεκινά με τη δειγματοληψία, όπου συλλέγονται δείγματα βιολογικού υλικού, όπως αίμα, ούρα, ιστοί και κύτταρα, από τον οργανισμό που εξετάζεται. Στη συνέχεια, τα δείγματα υποβάλλονται σε διάφορες διαδικασίες επεξεργασίας, όπως εκχύλιση, απόσταξη ή διάλυση, ώστε να προετοιμαστούν για ανάλυση. Οι μεταβολίτες στη συνέχεια μετρούνται και αναλύονται μέσω διαφόρων τεχνικών, όπως υγρή χρωματογραφία, μαζική φασματομετρία (mass spectrometry) ή ανάλυση πυρηνικού μαγνητικού συντονισμού (nuclear magnetic resonance, NMR). Τα δεδομένα που προκύπτουν από την ανάλυση επεξεργάζονται στατιστικά και υπολογιστικά για να εντοπιστούν μεταβολομικά πρότυπα και συσχετίσεις.

Η μεταβολομική ανάλυση έχει πολλές εφαρμογές στη βιολογία και την ιατρική. Για παράδειγμα, στον τομέα της ιατρικής προσαρμοστικής βιολογίας, μπορεί να βοηθήσει στη διάγνωση και την παρακολούθηση ασθενειών, καθώς και στην επιλογή και προσαρμογή θεραπειών. Στη γενετική και γονιδιωματική έρευνα, αποκαλύπτει πώς γονίδια και γενετικές διαφορές επηρεάζουν το μεταβολισμό. Στην αγροτική βιολογία, η μεταβολομική ανάλυση συμβάλλει στη βελτίωση της παραγωγής τροφίμων και στην επιλογή φυτών με επιθυμητές μεταβολομικές ιδιότητες. Οι εφαρμογές της μεταβολομικής ανάλυσης είναι ευρείες και συνεχώς εξελίσσονται, προσφέροντας βαθύτερη κατανόηση της λειτουργίας των ζωντανών οργανισμών (Nalbantoglu, 2019).

2.3 Εφαρμογές των σύγχρονων τεχνικών στην έρευνα και την κλινική πράξη

Οι σύγχρονες τεχνικές στη Μοριακή Βιολογία έχουν εφαρμογές τόσο στην έρευνα όσο και στην κλινική πράξη, προσφέροντας καινοτόμες λύσεις σε προβλήματα υγείας και προαγωγής της καλύτερης κατανόησης της ανθρώπινης φυσιολογίας. Ορισμένες από αυτές τις εφαρμογές περιλαμβάνουν (Milone & O'Doherty, 2018):

1. **Διάγνωση και θεραπεία νόσων:** Η γνώση της γενετικής ποικιλομορφίας και η εφαρμογή τεχνικών όπως η γονομική ανάλυση μπορεί να βοηθήσει στη διάγνωση και την πρόβλεψη των γενετικών νόσων, καθώς και στην εξατομικευμένη αγωγή.
2. **Ανακάλυψη νέων φαρμάκων:** Η ανάλυση των γενετικών δεδομένων και η κατανόηση των μοριακών μηχανισμών νόσων μπορεί να οδηγήσει στην ανακάλυψη νέων φαρμάκων και θεραπειών που είναι πιο αποτελεσματικά και επιδραστικά.
3. **Επιτάχυνση της έρευνας:** Οι τεχνικές όπως η μεταγραφομική ανάλυση και η πρωτεομική ανάλυση επιτρέπουν την ταχεία ανακάλυψη νέων γονιδίων και πρωτεϊνών που σχετίζονται με διάφορες παθήσεις και καταστάσεις.
4. **Προσαρμοσμένη θεραπεία:** Η γνώση των γενετικών δεδομένων ενός ασθενούς μπορεί να χρησιμοποιηθεί για την προσαρμοσμένη θεραπεία, όπως η επιλογή του κατάλληλου φαρμάκου ή η πρόβλεψη της ανταπόκρισης σε μια θεραπεία.
5. **Πρόληψη ασθενειών:** Η γνώση της γενετικής προδιάθεσης για συγκεκριμένες παθήσεις μπορεί να βοηθήσει στην πρόληψη των ασθενειών μέσω προληπτικών μέτρων και της προσαρμογής του τρόπου ζωής.

Αυτές οι εφαρμογές αποτελούν μόνο μερικά παραδείγματα του πώς οι σύγχρονες τεχνικές στη Μοριακή Βιολογία επηρεάζουν την έρευνα και την κλινική πράξη, προσφέροντας καινοτόμες λύσεις σε προβλήματα υγείας και βιολογικής κατανόησης (Milone & O'Doherty, 2018).

2.4 Προκλήσεις και προοπτικές της Μοριακής Βιολογίας: Ο ρόλος της διερευνητικής ανάλυσης δεδομένων

Ο ρόλος της διερευνητικής ανάλυσης δεδομένων στη Μοριακή Βιολογία είναι κρίσιμος για την αξιοποίηση και την ερμηνεία των δεδομένων που παράγονται από τις σύγχρονες τεχνικές. Αντιμετωπίζονται ορισμένες προκλήσεις και ανοίγουν νέες προοπτικές στον τομέα της Μοριακής Βιολογίας λόγω της αυξανόμενης πολυπλοκότητας των δεδομένων και των αναγκών για νέες μεθόδους ανάλυσης.

Ορισμένες από αυτές τις προκλήσεις και προοπτικές περιλαμβάνουν (Rehman, Naz & Razzak, 2022):

Αντιμετώπιση του Όγκου των Δεδομένων

Η σύγχρονη τεχνολογία έχει οδηγήσει σε σημαντική αύξηση του όγκου και της πολυπλοκότητας των δεδομένων στη Μοριακή Βιολογία. Τεχνικές όπως το RNA-seq και το ChIP-seq παράγουν δεδομένα υψηλής διαστασιμότητας και πολυπλοκότητας. Η διερευνητική ανάλυση δεδομένων βοηθά στη διαχείριση αυτού του μεγάλου όγκου δεδομένων και στην ανάγκη να εξαχθούν σημαντικά συμπεράσματα από αυτά, απαιτώντας εξελιγμένες μεθόδους ανάλυσης για να αναλυθούν αποτελεσματικά.

Ολοκληρωμένη Ανάλυση Δεδομένων

Η διερευνητική ανάλυση δεδομένων συμβάλλει στην ολοκληρωμένη ανάλυση διαφορετικών ειδών δεδομένων, όπως γονιδιακά δεδομένα, δεδομένα μεταγραφομικής και πρωτεομικής έκφρασης, καθώς και δεδομένα βιολογικών δρόμων. Τα δεδομένα στη μοριακή βιολογία προέρχονται από διάφορες πηγές και τεχνικές, όπως το RNA-seq, το ChIP-seq, και η μεταβολομική ανάλυση, που απαιτούν την ενοποίηση των δεδομένων παρά τις διαφορές τους. Αυτή η ολοκληρωμένη ανάλυση επιτρέπει την απόκτηση μιας πλήρους εικόνας των διαδικασιών σε κυτταρικό και μοριακό επίπεδο, απαιτώντας εξελιγμένους αλγόριθμους που μπορούν να διαχειριστούν αυτή την πολυπλοκότητα.

Ανάπτυξη Νέων Μεθόδων Ανάλυσης

Η ανάγκη για πιο αποτελεσματικές μεθόδους ανάλυσης και ερμηνείας των μοριακών δεδομένων οδηγεί στην ανάπτυξη νέων αλγορίθμων και εργαλείων διερευνητικής ανάλυσης δεδομένων. Οι περιορισμοί στο υπολογιστικό κόστος για την ανάλυση μεγάλων δεδομένων στη μοριακή βιολογία, όπως η ανάγκη για σημαντική υπολογιστική ισχύ και αποθηκευτικό χώρο, ωθούν στην ανάπτυξη νέων μεθόδων που μπορούν να επεξεργαστούν αποτελεσματικά τα δεδομένα. Οι υπολογισμοί που

απαιτούνται για την ανάλυση τέτοιων δεδομένων μπορούν να είναι ιδιαίτερα έντονοι, απαιτώντας εξειδικευμένο υλικό και λογισμικό.

Ενσωμάτωση Δεδομένων από Διαφορετικές Πηγές

Η διερευνητική ανάλυση δεδομένων βοηθά στην ενσωμάτωση δεδομένων από διαφορετικές πηγές, όπως γονοτύπους, δεδομένα κλίνης και πειραματικά δεδομένα, παρέχοντας μια πιο ολοκληρωμένη κατανόηση των βιολογικών φαινομένων. Αυτή η ετερογένεια καθιστά την ενοποίηση των δεδομένων δύσκολη, καθώς κάθε τεχνική μπορεί να έχει διαφορετικά χαρακτηριστικά και θορύβους. Η συνδυαστική ανάλυση τέτοιων ετερογενών δεδομένων απαιτεί εξελιγμένους αλγόριθμους που μπορούν να διαχειριστούν αυτή την πολυπλοκότητα.

Η διερευνητική ανάλυση δεδομένων αναδεικνύεται έτσι ως ένα εργαλείο που αντιμετωπίζει τις προκλήσεις αυτές και προσφέρει νέες προοπτικές για την προώθηση της Μοριακής Βιολογίας και της βιολογικής έρευνας. Η ανάγκη για διερευνητική ανάλυση καθίσταται επιτακτική λόγω της συνεχώς αυξανόμενης πολυπλοκότητας και του όγκου των δεδομένων που παράγονται από τις σύγχρονες τεχνικές μοριακής βιολογίας (Rehman, Naz & Razzak, 2022).

2.4.1 Υπολογιστικές Προκλήσεις στη Μοριακή Βιολογία

Στη μοριακή βιολογία, η διερευνητική ανάλυση δεδομένων αντιμετωπίζει πληθώρα υπολογιστικών προκλήσεων που καθιστούν την επεξεργασία και ερμηνεία των δεδομένων περίπλοκη και απαιτητική. Αυτές οι προκλήσεις οφείλονται κυρίως στην πολυπλοκότητα και τον όγκο των δεδομένων, την ετερογένεια των πηγών και την ανάγκη για σημαντική υπολογιστική ισχύ. Παρακάτω αναλύονται κάποιες από τις κύριες προκλήσεις (Rehman, Naz & Razzak, 2022):

Μεγάλη Πολυπλοκότητα

Οι τεχνικές που χρησιμοποιούνται στη μοριακή βιολογία, όπως το RNA-seq και το ChIP-seq, παράγουν δεδομένα υψηλής διαστασιμότητας και πολυπλοκότητας. Αυτά τα δεδομένα απαιτούν εξελιγμένες μεθόδους ανάλυσης για να αναλυθούν αποτελεσματικά. Η πολυπλοκότητα αυτή δυσχεραίνει την εξαγωγή σαφών και αξιόπιστων συμπερασμάτων, απαιτώντας την ανάπτυξη εξειδικευμένων υπολογιστικών εργαλείων.

Ετερογένεια Δεδομένων

Τα δεδομένα στη μοριακή βιολογία προέρχονται από διάφορες πηγές και τεχνικές, όπως το RNA-seq, το ChIP-seq, και η μεταβολομική ανάλυση. Αυτή η ετερογένεια καθιστά την ενοποίηση των δεδομένων δύσκολη, καθώς κάθε τεχνική μπορεί να έχει διαφορετικά χαρακτηριστικά και θορύβους. Η συνδυαστική ανάλυση τέτοιων ετερογενών δεδομένων απαιτεί εξελιγμένους αλγόριθμους που μπορούν να διαχειριστούν αυτή την πολυπλοκότητα.

Υπολογιστικό Κόστος

Η ανάλυση μεγάλων δεδομένων στη μοριακή βιολογία απαιτεί σημαντική υπολογιστική ισχύ και αποθηκευτικό χώρο. Οι υπολογισμοί που απαιτούνται για την ανάλυση τέτοιων δεδομένων μπορούν να είναι ιδιαίτερα έντονοι, απαιτώντας εξειδικευμένο υλικό και λογισμικό για την αποτελεσματική επεξεργασία τους. Οι περιορισμοί αυτοί μπορούν να επηρεάσουν την ταχύτητα και την αποδοτικότητα της έρευνας.

Ερμηνευσιμότητα Αποτελεσμάτων

Η πολυπλοκότητα των μοντέλων που χρησιμοποιούνται για την ανάλυση δεδομένων μοριακής βιολογίας μπορεί να δυσχεραίνει την ερμηνεία των αποτελεσμάτων. Τα εξελιγμένα μοντέλα μηχανικής μάθησης και στατιστικής ανάλυσης που

χρησιμοποιούνται για την επεξεργασία των δεδομένων μπορεί να είναι δύσκολο να ερμηνευτούν από τους ερευνητές, περιορίζοντας έτσι την κατανόηση των βιολογικών διαδικασιών που μελετώνται.

Καταλήγοντας, οι παραπάνω προκλήσεις υπογραμμίζουν την ανάγκη για διερευνητική ανάλυση δεδομένων στη μοριακή βιολογία. Η διερευνητική ανάλυση επιτρέπει την καλύτερη κατανόηση και ερμηνεία των δεδομένων, παρέχοντας στους ερευνητές τα εργαλεία που χρειάζονται για να αντλήσουν αξιόπιστα και ουσιαστικά συμπεράσματα από τα δεδομένα τους. Αυτή η ανάγκη καθίσταται επιτακτική λόγω της συνεχώς αυξανόμενης πολυπλοκότητας και του όγκου των δεδομένων που παράγονται από τις σύγχρονες τεχνικές μοριακής βιολογίας (Rehman, Naz & Razzak, 2022).

3 Διερευνητική Ανάλυση Δεδομένων: Μέθοδοι, Τεχνικές και Εφαρμογές

Η Διερευνητική Ανάλυση Δεδομένων (ΔΑΔ) (Exploratory Data Analysis - EDA) είναι μια διαδικασία που ασχολείται με τον εντοπισμό, την εξαγωγή και την ερμηνεία πληροφοριών από μεγάλα σύνολα δεδομένων. Στόχος της είναι να αποκαλύψει τα μοτίβα, τις σχέσεις και τις τάσεις που μπορεί να κρύβουν τα δεδομένα, προκειμένου να ληφθούν σημαντικές αποφάσεις ή να προβλεφθούν μελλοντικές εξελίξεις (Jebb et al., 2017).

Η ΔΑΔ χρησιμοποιεί μια ποικιλία τεχνικών και εργαλείων από διάφορους επιστημονικούς τομείς, συμπεριλαμβανομένων της στατιστικής, της μηχανικής μάθησης, της τεχνητής νοημοσύνης και της οπτικοποίησης δεδομένων. Οι βασικές φάσεις της διαδικασίας περιλαμβάνουν την προετοιμασία των δεδομένων, την εξόρυξη δεδομένων, την ανάλυση και την ερμηνεία των αποτελεσμάτων, καθώς και την αναφορά και την παρουσίαση των ευρημάτων.

Οι βασικές φάσεις της διαδικασίας της ανάλυσης δεδομένων συνήθως περιλαμβάνουν τα ακόλουθα στάδια (Jebb et al., 2017):

1. **Προετοιμασία των Δεδομένων:** Αυτό το στάδιο περιλαμβάνει τη συλλογή, την καθαριότητα και την προετοιμασία των δεδομένων για ανάλυση. Συχνά, τα δεδομένα πρέπει να επεξεργαστούν για να αφαιρεθούν ανωμαλίες, να αντιμετωπιστούν απουσίες και να μετατραπούν σε μορφές που μπορούν να χρησιμοποιηθούν από τους αλγόριθμους ανάλυσης.
2. **Εξόρυξη Δεδομένων:** Σε αυτό το στάδιο, οι αλγόριθμοι και οι τεχνικές εφαρμόζονται στα προετοιμασμένα δεδομένα προκειμένου να εξαχθούν πληροφορίες, μοτίβα ή τάσεις. Αυτό μπορεί να γίνει μέσω διαφόρων τεχνικών, όπως αναζήτηση προτύπων, ανάλυση συστάδων ή ταξινόμηση δεδομένων.
3. **Ανάλυση:** Κατά τη διάρκεια αυτής της φάσης, οι αναλυτές εφαρμόζουν μεθόδους και τεχνικές για να εξάγουν πληροφορίες από τα δεδομένα. Αυτό μπορεί να περιλαμβάνει στατιστική ανάλυση, μοντέλα πρόβλεψης ή αναγνώριση παρατηρήσεων που ξεπερνούν το καθορισμένο καθορισμένο όριο.

4. **Ερμηνεία των Αποτελεσμάτων:** Αυτή η φάση περιλαμβάνει την ερμηνεία των αποτελεσμάτων και την αναγνώριση των συμπερασμάτων ή των προτάσεων που προκύπτουν από την ανάλυση. Εδώ, οι αναλυτές προσπαθούν να κατανοήσουν το νόημα πίσω από τα δεδομένα και να εξηγήσουν τις ευρήματα σε μια κατανοητή μορφή.
5. **Αναφορά και Παρουσίαση των Ευρημάτων:** Τέλος, οι αναλυτές πρέπει να αναφέρουν τα αποτελέσματα της ανάλυσης σε μια σαφή και συνοπτική μορφή. Αυτό μπορεί να περιλαμβάνει τη σύνταξη αναφορών, τη δημοσίευση άρθρων ή την παρουσίαση των ευρημάτων σε επιστημονικές συναντήσεις.

Οι εφαρμογές της ΔΑΔ είναι ευρείες και καλύπτουν πολλούς τομείς, όπως η οικονομία, η υγεία, οι επιχειρήσεις, η εκπαίδευση, η επιστήμη και η κοινωνία γενικότερα. Παραδείγματα εφαρμογών περιλαμβάνουν την πρόβλεψη των τάσεων αγοράς, την ανίχνευση απάτης, τη βελτιστοποίηση διαδικασιών και αποφάσεων, καθώς και την ανακάλυψη νέων γνώσεων και συνδέσεων μεταξύ δεδομένων (Jebb et al., 2017).

Ωστόσο, η ΔΑΔ δεν είναι απλά μια τεχνική διαδικασία. Είναι ένα ευρύ φάσμα δραστηριοτήτων που περιλαμβάνει τη σωστή κατανόηση του περιβάλλοντος, την ανάπτυξη ερωτημάτων, την εργασία με ομάδες ειδικών και ενίοτε την ανασκόπηση και την αναθεώρηση των αποτελεσμάτων. Συνεπώς, η επιτυχής εφαρμογή της ΔΑΔ απαιτεί τόσο τεχνικές όσο και συναισθηματικές δεξιότητες.

Τέλος, είναι σημαντικό να τονιστεί ότι η ΔΑΔ πρέπει να γίνεται με σεβασμό προς την ιδιωτικότητα και την ασφάλεια των δεδομένων, καθώς και με επίγνωση των δεοντολογικών και νομικών περιορισμών που ενδέχεται να υπάρχουν (Jebb et al., 2017).

3.1 Συλλογή δεδομένων

Η συλλογή δεδομένων είναι ένα σημαντικό και βασικό βήμα στην επιστημονική έρευνα, στην επιχειρηματική ανάλυση, και σε πολλούς άλλους τομείς. Αποτελεί τη διαδικασία συγκέντρωσης πληροφοριών ή δεδομένων από διάφορες πηγές, με σκοπό

την ανάλυση, την εξαγωγή συμπερασμάτων, την κατανόηση ή τη λήψη αποφάσεων βασισμένων σε αυτά (Sahoo et al., 2019).

Σημασία της Συλλογής Δεδομένων:

- Τα δεδομένα αποτελούν τη βάση κάθε ανάλυσης και έρευνας, καθώς παρέχουν τις πληροφορίες πάνω στις οποίες θα γίνουν οι εκτιμήσεις, οι συμπεράσματα και οι αποφάσεις.
- Η ποιότητα και η εγκυρότητα των δεδομένων επηρεάζουν αποφασιστικά την αξιοπιστία των αποτελεσμάτων και των συμπερασμάτων που προκύπτουν.

Πηγές Δεδομένων: Οι πηγές από τις οποίες μπορούν να συλλεχθούν δεδομένα είναι πολλές και ποικίλες, περιλαμβάνοντας:

1. **Εσωτερικές πηγές:** Αυτές περιλαμβάνουν τα εσωτερικά συστήματα και βάσεις δεδομένων του οργανισμού ή της επιχείρησης.
2. **Εξωτερικές πηγές:** Αυτές περιλαμβάνουν δεδομένα από διάφορες δημόσιες ή ιδιωτικές πηγές, όπως δημόσιες βάσεις δεδομένων, αρχεία κειμένου, ιστοσελίδες, αισθητήρες, κ.ά.

Τεχνικές Συλλογής Δεδομένων: Η συλλογή δεδομένων μπορεί να γίνει με διάφορες τεχνικές, συμπεριλαμβανομένων:

1. **Παρατήρηση:** Η συλλογή δεδομένων γίνεται μέσω παρατηρήσεων.
2. **Ερωτηματολόγια:** Οι συμμετέχοντες απαντούν σε ερωτήσεις που τίθενται από τον ερευνητή.
3. **Συνεντεύξεις:** Ο ερευνητής επικοινωνεί απευθείας με τους συμμετέχοντες για τη συλλογή δεδομένων.
4. **Δοκιμές:** Οι συμμετέχοντες υπόκεινται σε δοκιμές ή πειράματα για τη συλλογή δεδομένων.

Συνολικά, η συλλογή δεδομένων είναι ένα σημαντικό στάδιο που απαιτεί προσεκτικό σχεδιασμό και εκτέλεση, καθώς επηρεάζει την ακρίβεια και την αξιοπιστία των

αποτελεσμάτων και των συμπερασμάτων που θα προκύψουν από την ανάλυση (Sahoo et al., 2019).

3.2 Προετοιμασία δεδομένων

Η προετοιμασία δεδομένων είναι ένα σημαντικό στάδιο στη διαδικασία ανάλυσης δεδομένων, καθώς επιτρέπει τη διόρθωση και την επεξεργασία των δεδομένων προκειμένου να είναι έτοιμα για ανάλυση. Αυτή η φάση περιλαμβάνει διάφορες διαδικασίες που στοχεύουν στην εξασφάλιση της ποιότητας, της ακρίβειας και της εγκυρότητας των δεδομένων πριν από την ανάλυσή τους (Milo & Somech, 2020).

Οι βασικές διαδικασίες προετοιμασίας δεδομένων περιλαμβάνουν τα ακόλουθα:

1. Καθαρισμός Δεδομένων:

- Ανίχνευση και διόρθωση λαθών και ανωμαλιών στα δεδομένα, όπως απουσιάζουσες τιμές, διπλότυπες εγγραφές ή αντιφάσεις.
- Αφαίρεση ή αντικατάσταση ανωμαλιών με λογικές τιμές ή μέσους όρους.

2. Διαχείριση Απουσιάζουσων Τιμών:

- Αντιμετώπιση απουσιάζουσων τιμών με την αντικατάστασή τους με μέσους όρους, μεταθέτοντας την πληροφορία από άλλες παρόμοιες εγγραφές, ή αφαιρώντας εγγραφές.

3. Διαχείριση Κατηγορικών Δεδομένων:

- Μετατροπή κατηγορικών δεδομένων σε δυαδικές μεταβλητές (one-hot encoding) ή σε αριθμητικές τιμές.
- Κωδικοποίηση κατηγορικών μεταβλητών με μέσο την κατάταξη (ordinal encoding).

4. Κανονικοποίηση Δεδομένων:

- Κλιμάκωση των δεδομένων σε συγκεκριμένο εύρος τιμών ή μετατροπή τους σε συγκεκριμένη κατανομή, ώστε να μειωθούν οι διαφορές κλίμακας μεταξύ μεταβλητών.

5. Δημιουργία Νέων Χαρακτηριστικών:

- Παραγωγή νέων μεταβλητών με βάση τις υπάρχουσες, όπως πολλαπλασιασμός, διαίρεση ή σύγκριση μεταξύ μεταβλητών.

6. Απόρριψη Μη Χρήσιμων Χαρακτηριστικών:

- Αφαίρεση μεταβλητών που δεν συμβάλλουν στην ανάλυση ή δεν παρέχουν πληροφορία.

Η προετοιμασία δεδομένων είναι κρίσιμη για την επιτυχία της ανάλυσης δεδομένων και την αξιοπιστία των αποτελεσμάτων. Η καλή πρακτική σε αυτό το στάδιο εξασφαλίζει την ακρίβεια, τη συνοχή και την αξιοπιστία των αποτελεσμάτων που προκύπτουν από την ανάλυση (Milo & Somech, 2020).

3.3 Τεχνικές Διερευνητικής Ανάλυσης

Η διερευνητική ανάλυση δεδομένων είναι ένα κρίσιμο εργαλείο στην επιστήμη των δεδομένων και την αναγνώριση προτύπων, επιτρέποντας την εξερεύνηση, κατανόηση και ανάδειξη κρυμμένων πληροφοριών από ένα σύνολο δεδομένων χωρίς την ανάγκη συγκεκριμένων υποθέσεων (Camacho et al., 2017).

Αυτές οι τεχνικές δίνουν στους ερευνητές και αναλυτές τη δυνατότητα να εξετάσουν τα χαρακτηριστικά, τα πρότυπα και τις σχέσεις μεταξύ των μεταβλητών χωρίς περιορισμούς από προκαταλήψεις ή προϋπάρχουσες θεωρίες. Ο κύριος στόχος είναι να αποκαλυφθούν νέες πληροφορίες και να διαμορφωθούν νέες ερωτήσεις για μελλοντική έρευνα.

Οι τεχνικές διερευνητικής ανάλυσης καλύπτουν μια ευρεία γκάμα διαδικασιών, συμπεριλαμβανομένων της οπτικοποίησης δεδομένων, της ανάλυσης κατανομής, της ανίχνευσης προτύπων και της συσταδοποίησης. Ο συνδυασμός αυτών των τεχνικών μπορεί να οδηγήσει σε πολλαπλές ερμηνείες και στην ανακάλυψη νέων γνώσεων (Camacho et al., 2017).

Η διερευνητική ανάλυση αποτελεί ένα αναγκαίο βήμα στη διαδικασία ανάλυσης δεδομένων, καθώς επιτρέπει την κατανόηση της φύσης και της δομής των δεδομένων πριν από την εφαρμογή πιο προηγμένων αλγορίθμων ή την εξαγωγή συμπερασμάτων.

Οι βασικές τεχνικές διερευνητικής ανάλυσης περιλαμβάνουν τις αναλύσεις κεντρικής τάσης και διασποράς, όπου υπολογίζονται και ερμηνεύονται μετρικές όπως ο μέσος όρος, η διάμεσος, το ποσοστό και οι τυπικές αποκλίσεις, βοηθώντας στην κατανόηση της κεντρικής τάσης και της διασποράς των δεδομένων. Ιστογράμματα και διαγράμματα σκαλίσματος απεικονίζουν την κατανομή των δεδομένων, βοηθώντας στην αναγνώριση πιθανών μοτίβων και τάσεων. Διαγράμματα διασποράς επιτρέπουν την οπτικοποίηση των σχέσεων μεταξύ δύο μεταβλητών, διευκολύνοντας την ανίχνευση πιθανών συσχετίσεων ή μοτίβων (Camacho et al., 2017).

Επιπλέον, διαγράμματα και γραφήματα πιθανότητας απεικονίζουν την κανονική κατανομή των δεδομένων, επιτρέποντας την εκτίμηση του κατά πόσο οι δεδομένοι ακολουθούν μια κανονική κατανομή. Κουτιογράμματα απεικονίζουν την κατανομή των δεδομένων και ανιχνεύουν ανωμαλίες ή εξαιρέσεις. Τέλος, η διακριτική ανάλυση εξετάζει την κατανομή των δεδομένων σε διαφορετικές ομάδες ή κατηγορίες, βοηθώντας στην αναγνώριση διαφορών ή προτύπων.

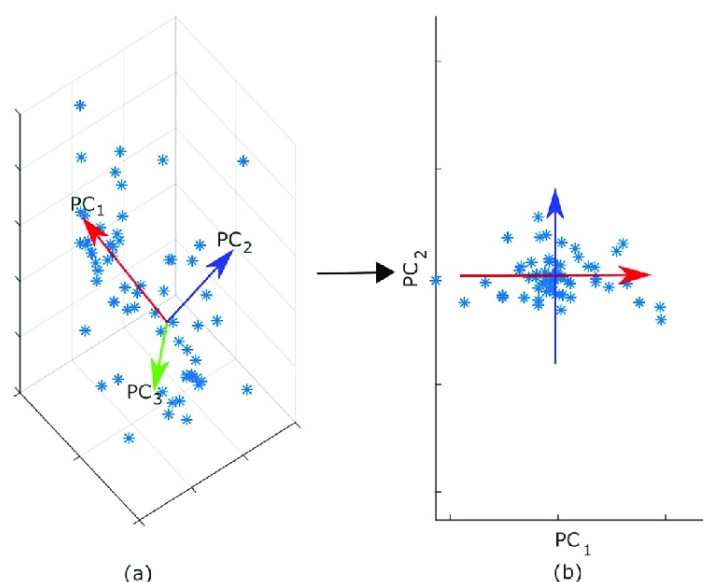
Συνολικά, οι τεχνικές διερευνητικής ανάλυσης παρέχουν μια αρχική εικόνα των δεδομένων και επιτρέπουν στους αναλυτές να προχωρήσουν στην εξερεύνηση πιο σύνθετων σχέσεων και προτύπων, αποτελώντας ένα απαραίτητο βήμα πριν από την εφαρμογή πιο προηγμένων αλγορίθμων ανάλυσης δεδομένων (Camacho et al., 2017).

3.3.1 Ανάλυση Κυρίων Συνιστωσών (PCA)

Η Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis - PCA) είναι μια ισχυρή και ευρέως χρησιμοποιούμενη τεχνική διαστατικής μείωσης δεδομένων και ανάλυσης πολυδιάστατων συνόλων δεδομένων. Στόχος της PCA είναι να μειώσει τη διαστατικότητα του συνόλου δεδομένων ενώ διατηρεί τη μεγαλύτερη δυνατή ποσότητα πληροφορίας (Kherif & Latypova, 2020).

Ο βασικός στόχος της PCA είναι η εύρεση των κυρίων συνιστωσών (principal components) των δεδομένων. Οι κυρίως συνιστώσες είναι νέες μεταβλητές που προκύπτουν από μια γραμμική συνδυαστική περίληψη των αρχικών μεταβλητών.

Αυτές οι νέες μεταβλητές είναι οργανωμένες με τέτοιο τρόπο ώστε η πρώτη κυρίως συνιστώσα να αποτελεί τον άξονα με τη μεγαλύτερη διακύμανση στα δεδομένα, η δεύτερη κυρίως συνιστώσα τον άξονα με τη δεύτερη μεγαλύτερη διακύμανση, και ούτω καθεξής.



Εικόνα 4: Διαδικασία μείωσης διαστάσεων που εκτελείται με PCA.

Η διαδικασία της PCA περιλαμβάνει τα ακόλουθα βήματα (Kherif & Latypova, 2020):

1. **Κέντρωση των δεδομένων:** Αφαίρεση του μέσου τιμήματος από κάθε μεταβλητή ώστε να κεντραριστούν τα δεδομένα γύρω από το σημείο (0,0).
2. **Υπολογισμός του Πίνακα Συνδιακύμανσης ή Πίνακα Διασποράς:** Υπολογισμός του πίνακα που περιέχει όλες τις συνδιακυμανσές μεταξύ των μεταβλητών.
3. **Υπολογισμός των Κυρίων Συνιστωσών:** Εφαρμογή μιας μεθόδου για τον υπολογισμό των κυρίων συνιστωσών, συνήθως με τη χρήση μεθόδων αυτοσυσχέτισης ή μεθόδων ανάλυσης ιδιοτιμών.
4. **Επιλογή Κυρίων Συνιστωσών:** Επιλέγονται οι πρώτες κυρίως συνιστώσες που εξηγούν το μεγαλύτερο ποσοστό της διακύμανσης των δεδομένων.

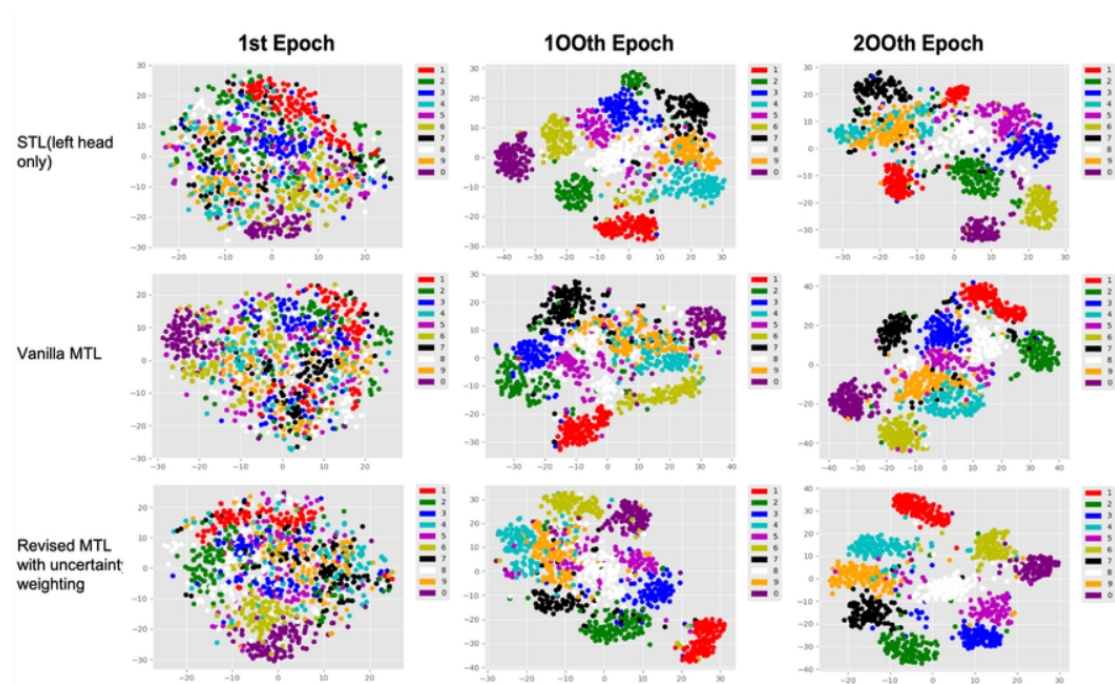
5. Μείωση της Διαστατικότητας: Εφαρμογή της μείωσης διαστατικότητας αφαιρώντας τις λιγότερο σημαντικές κυρίως συνιστώσες.

Η PCA χρησιμοποιείται ευρέως σε πολλούς τομείς, όπως η ανάλυση δεδομένων, η αναγνώριση προτύπων, η επεξεργασία εικόνας, η αναγνώριση προτύπων και η μείωση διαστατικότητας για την απλοποίηση των δεδομένων και την αποτελεσματικότερη ανάλυσή τους (Kherif & Latypova, 2020).

Μείωση Διαστάσεων με t-SNE, UMAP και ISOMAP

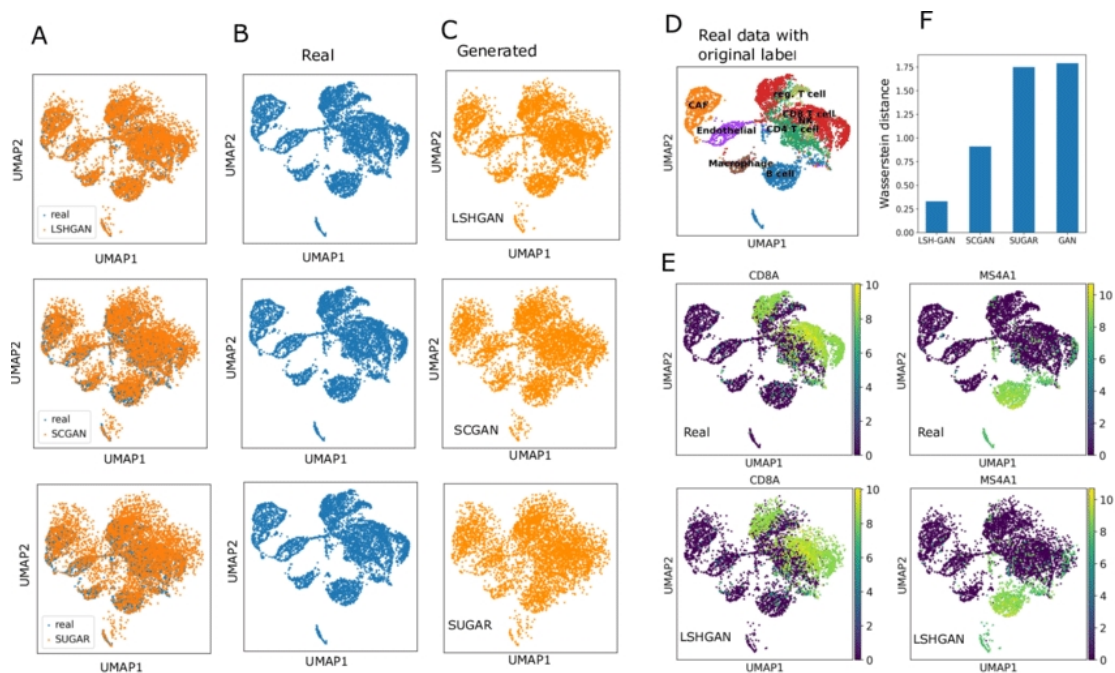
Η μείωση διάστασης είναι ένας σημαντικός τομέας στην ανάλυση δεδομένων, ιδιαίτερα όταν τα δεδομένα είναι υψηλής διάστασης και πολύπλοκα. Κάθε μια από αυτές τις τεχνικές έχει διαφορετικά πλεονεκτήματα και εφαρμογές, προσφέροντας διαφορετικές προσεγγίσεις στη διαχείριση των δεδομένων και την αποκάλυψη κρυφών δομών (Monsen & Monsen, 2018).

t-SNE (t-Distributed Stochastic Neighbor Embedding): Το t-SNE είναι μια τεχνική μείωσης διάστασης που έχει σχεδιαστεί για να χαρτογραφεί δεδομένα υψηλής διάστασης σε δύο ή τρεις διαστάσεις, διατηρώντας τη σχετική απόσταση μεταξύ των δεδομένων. Αυτό επιτυγχάνεται μέσω μιας στοχαστικής διαδικασίας που ελαχιστοποιεί τις διαφορές μεταξύ των αρχικών και των χαμηλής διάστασης αναπαραστάσεων των δεδομένων. Το t-SNE είναι ιδιαίτερα αποτελεσματικό στην αποκάλυψη σύνθετων μοτίβων και σχέσεων, όπως οι ομάδες (clusters) σε σύνολα δεδομένων. Ωστόσο, η υπολογιστική του πολυπλοκότητα μπορεί να είναι αυξημένη, γεγονός που το καθιστά λιγότερο αποδοτικό για πολύ μεγάλα σύνολα δεδομένων.



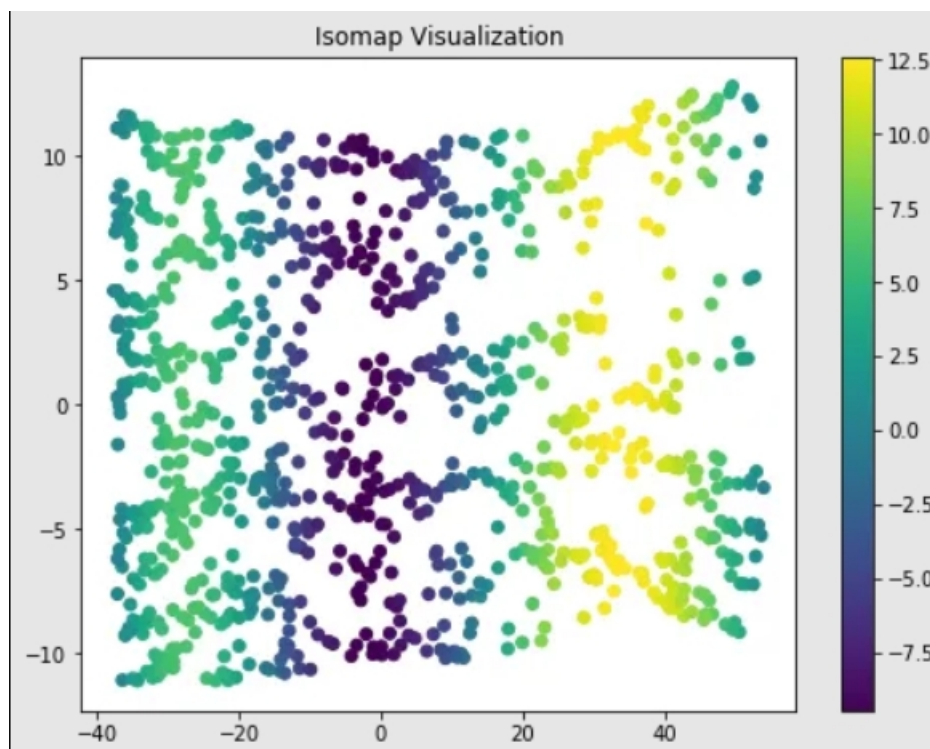
Εικόνα 5: *t*-SNE (*t*-Distributed Stochastic Neighbor Embedding).

UMAP (Uniform Manifold Approximation and Projection): Το UMAP είναι μια πιο πρόσφατη τεχνική που προσφέρει ταχύτερη και πιο αποδοτική μείωση διάστασης σε σύγκριση με το *t*-SNE. Το UMAP διατηρεί καλύτερα τις τοπικές και παγκόσμιες δομές των δεδομένων, κάνοντάς το ιδανικό για την ανάλυση σύνθετων πολυδιάστατων δεδομένων. Η τεχνική αυτή χρησιμοποιεί τη θεωρία των πολλαπλών (manifolds) για να προσεγγίσει τη δομή των δεδομένων και να τη διατηρήσει σε χαμηλότερες διαστάσεις. Λόγω της ταχύτητας και της ακρίβειάς του, το UMAP χρησιμοποιείται ευρέως σε εφαρμογές όπως η ανάλυση γονιδιωματικών δεδομένων και η ανάλυση εικόνας.



Εικόνα 6: UMAP (Uniform Manifold Approximation and Projection).

ISOMAP: Το ISOMAP είναι μια τεχνική μείωσης διάστασης που επεκτείνει την κλασική ανάλυση πολυδιάστατων κλιμάκων (MDS) λαμβάνοντας υπόψη τις γεωδαιτικές αποστάσεις στον πολλαπλόχωρο των δεδομένων. Αντί να χρησιμοποιεί τις ευκλείδειες αποστάσεις, το ISOMAP χρησιμοποιεί τις γεωδαιτικές αποστάσεις για να διατηρήσει την πραγματική μορφή και δομή των δεδομένων σε χαμηλότερες διαστάσεις. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη για δεδομένα που έχουν φυσική πολλαπλή δομή, όπως τα δεδομένα από βιολογικές ή κοινωνικές επιστήμες.



Εικόνα 7: ISOMAP.

Κάθε μία από αυτές τις τεχνικές μείωσης διάστασης προσφέρει μοναδικά πλεονεκτήματα και εφαρμογές. Η επιλογή της κατάλληλης τεχνικής εξαρτάται από τις ιδιαιτερότητες του συνόλου δεδομένων και τους στόχους της ανάλυσης. Εν κατακλείδι, η κατανόηση και η σωστή εφαρμογή αυτών των τεχνικών μπορεί να οδηγήσει σε σημαντικές ανακαλύψεις και να βελτιώσει την ικανότητα ανάλυσης των δεδομένων.

3.3.2 Κατανομή Δεδομένων: Κυριότερες μέθοδοι

Η κατανομή δεδομένων αποτελεί ένα κεντρικό στοιχείο στον χώρο της στατιστικής και της ανάλυσης δεδομένων, αναφερόμενη στον τρόπο με τον οποίο τα δεδομένα κατανέμονται σε ένα σύνολο. Αυτό περιλαμβάνει τον προσδιορισμό της συχνότητας εμφάνισης των διαφορετικών τιμών και τη μορφή της κατανομής τους (Monsen & Monsen, 2018).

Οι κατανομές δεδομένων μπορεί να ποικίλουν σημαντικά, ανάλογα με τη φύση των δεδομένων και τον τρόπο συλλογής τους. Η κανονική κατανομή, γνωστή και ως Γκαουσιανή κατανομή, είναι μία από τις πιο γνωστές και χρησιμοποιούμενες. Χαρακτηρίζεται από τη μορφή καμπάνας και έχει ένα μοναδικό μέσο όρο και

διακύμανση, με πολλά φυσικά και κοινωνικά φαινόμενα να την ακολουθούν. Η εκθετική κατανομή, από την άλλη, χαρακτηρίζεται από την τάση των δεδομένων να συγκεντρώνονται κοντά σε μία άκρη του εύρους και να μειώνονται εκθετικά καθώς απομακρύνονται από αυτήν.

Αντίθετα με τις συνεχείς κατανομές, οι διακριτικές κατανομές αντιπροσωπεύουν μετρήσεις που προέρχονται από μια πεπερασμένη σειρά πιθανών τιμών. Ένα παράδειγμα διακριτικής κατανομής είναι η κατανομή των χαρτονομισμάτων που πέφτουν με το πρόσωπο προς τα πάνω. Η βήτα κατανομή, η οποία χρησιμοποιείται συχνά σε πειραματικά δεδομένα, χαρακτηρίζεται από δύο θετικές παραμέτρους και μπορεί να πάρει ποικίλες μορφές ανάλογα με τις τιμές αυτών των παραμέτρων. Η γεωμετρική κατανομή, μία διακριτική κατανομή, χρησιμοποιείται για να περιγράψει τον αριθμό των δοκιμών που απαιτούνται πριν συμβεί ένα συγκεκριμένο γεγονός (Monsen & Monsen, 2018).

Η κατανόηση της κατανομής των δεδομένων είναι κρίσιμη για την αξιολόγηση των στατιστικών ιδιοτήτων τους και την προετοιμασία της ανάλυσης. Αυτές οι κατανομές αποτελούν θεμελιώδη εργαλεία στη στατιστική και την ανάλυση δεδομένων, παρέχοντας τη βάση για περαιτέρω έρευνα και ανάλυση.

Οι κυριότερες μέθοδοι κατανομής δεδομένων περιλαμβάνουν:

1. **Ιστογράμματα:** Τα ιστογράμματα απεικονίζουν τη συχνότητα εμφάνισης κάθε τιμής σε ένα σύνολο δεδομένων. Χρησιμοποιούνται ευρέως για να απεικονίσουν την κατανομή των δεδομένων και να εντοπίσουν πιθανά πρότυπα ή τάσεις.
2. **Πιθανοτικά Διαγράμματα:** Τα πιθανοτικά διαγράμματα είναι γραφήματα που χρησιμοποιούνται για την απεικόνιση της κατανομής των δεδομένων και τη σύγκρισή τους με μια θεωρητική κανονική κατανομή.
3. **Κουτιογράμματα (Box Plots):** Τα κουτιογράμματα ή Box plots παρουσιάζουν τη διακύμανση και τη διακριτική ανάλυση μιας μεταβλητής. Παρέχουν πληροφορίες για τη μέση τιμή, τη διάμεσο και τις ακραίες τιμές του συνόλου δεδομένων.

4. **Κανονικότητα Κατανομής (Normal Distribution):** Η κανονική κατανομή είναι μια σημαντική μέθοδος κατανομής, όπου τα δεδομένα κατανέμονται κανονικά γύρω από το μέσον τιμής με μια συγκεκριμένη διακύμανση.
5. **Διαγράμματα Σκαλίσματος (Scatter Plots):** Τα διαγράμματα σκαλίσματος απεικονίζουν τη σχέση μεταξύ δύο μεταβλητών και μπορούν να δείξουν το πλάτος και τη διασπορά της κατανομής των δεδομένων.

Αυτές οι μέθοδοι κατανομής δεδομένων είναι βασικές στη διαδικασία ανάλυσης δεδομένων και χρησιμοποιούνται ευρέως για την κατανόηση των χαρακτηριστικών και των προτύπων των δεδομένων. Η επιλογή της κατάλληλης μεθόδου εξαρτάται από τη φύση των δεδομένων και τους στόχους της ανάλυσης (Monsen & Monsen, 2018).

3.3.3 Συσταδοποίηση Δεδομένων

Η συσταδοποίηση δεδομένων είναι μια τεχνική ανάλυσης δεδομένων που αποσκοπεί στην ομαδοποίηση των δεδομένων σε ομάδες ή "συστάδες", ώστε τα μέλη της κάθε ομάδας να είναι πιο ομοιόμορφα μεταξύ τους από ό,τι με τα υπόλοιπα μέλη σε διαφορετικές ομάδες. Στόχος της συσταδοποίησης είναι η ανακάλυψη φυσικών ή κρυφών δομών στα δεδομένα, έτσι ώστε να μπορούν να γίνουν πιο εύκολα κατανοητά και ερμηνευτικά (Martinez, Martinez & Solka, 2017).

Οι βασικές έννοιες και η λειτουργία της συσταδοποίησης δεδομένων περιλαμβάνουν:

1. **Κριτήρια Ομοιότητας:** Η συσταδοποίηση βασίζεται στην ανάθεση των δεδομένων σε ομάδες με βάση κάποια μέτρα ομοιότητας. Τα κριτήρια ομοιότητας μπορεί να περιλαμβάνουν την ευκλείδεια απόσταση, τη συσχέτιση Pearson, τη συννημιτονική απόσταση, κλπ.
2. **Αλγόριθμοι Συσταδοποίησης:** Υπάρχουν πολλοί αλγόριθμοι συσταδοποίησης, όπως ο K-Means, ο Hierarchical Clustering, ο DBSCAN και ο Gaussian Mixture Model (GMM), με κάθε έναν να έχει τα πλεονεκτήματα και τους περιορισμούς του.

3. **Επιλογή Πλήθους Συστάδων:** Ένα σημαντικό βήμα είναι η επιλογή του κατάλληλου αριθμού συστάδων. Αυτό μπορεί να γίνει με τη χρήση μεθόδων όπως οι elbow method, silhouette score κλπ.
4. **Ερμηνευσιμότητα Συστάδων:** Αφού ολοκληρωθεί η συσταδοποίηση, είναι σημαντικό να ερμηνευτούν οι συστάδες που προέκυψαν, εξετάζοντας τα χαρακτηριστικά που τις προσδιορίζουν.

Η συσταδοποίηση δεδομένων χρησιμοποιείται ευρέως σε πολλούς τομείς, όπως η αναγνώριση προτύπων, η ανάλυση κοινωνικών δικτύων, η ανακάλυψη γνώσης από μεγάλα σύνολα δεδομένων κλπ. Η αποτελεσματική χρήση της απαιτεί κατανόηση των αλγορίθμων και την επιλογή της κατάλληλης μεθόδου συσταδοποίησης ανάλογα με τα χαρακτηριστικά των δεδομένων και τους στόχους της ανάλυσης (Martinez, Martinez & Solka, 2017).

Οι τεχνικές συσταδοποίησης δεδομένων είναι διάφορες μέθοδοι που χρησιμοποιούνται για να ομαδοποιήσουν μια συλλογή δεδομένων σε συστάδες με βάση την ομοιότητά τους. Κάθε τεχνική έχει τα δικά της πλεονεκτήματα και περιορισμούς, και η επιλογή της κατάλληλης τεχνικής εξαρτάται από τη φύση των δεδομένων και τους στόχους της ανάλυσης. Ορισμένες από τις κυριότερες τεχνικές συσταδοποίησης περιλαμβάνουν:

1. **K-Means:** Ο αλγόριθμος K-Means είναι ένας από τους πιο δημοφιλείς αλγορίθμους συσταδοποίησης. Ξεκινώντας με έναν αριθμό προκαθορισμένων συστάδων (k), ο αλγόριθμος εναλλάσσει τα κέντρα των συστάδων έως ότου βρει τα κέντρα που ελαχιστοποιούν την απόσταση από τα μέλη της συστάδας.
2. **Hierarchical Clustering:** Η ιεραρχική συσταδοποίηση είναι μια μέθοδος που δημιουργεί ένα δενδρόγραμμα των συστάδων, όπου κάθε κόμβος αντιστοιχεί σε μια συστάδα. Ο αλγόριθμος μπορεί να είναι άνοδος (αρχίζοντας με μια συστάδα και συνενώνοντας σταδιακά) ή καθοδικός (αρχίζοντας με κάθε στοιχείο ως ξεχωριστή συστάδα και διαχωρίζοντας σταδιακά).
3. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** Ο αλγόριθμος DBSCAN βασίζεται στην πυκνότητα των δεδομένων. Εντοπίζει περιοχές με υψηλή πυκνότητα και ομαδοποιεί τα δεδομένα που βρίσκονται

κοντά μεταξύ τους σε συστάδες, ενώ ετικετοποιεί τα εναπομείναντα δεδομένα ως θόρυβο.

4. **Gaussian Mixture Model (GMM):** Ο GMM αναπαριστά κάθε συστάδα με μια κατανομή Gauss, με παραμέτρους που υπολογίζονται με χρήση της μέθοδος μέγιστης πιθανοφάνειας.
5. **Spectral Clustering:** Η φασματική συσταδοποίηση μετατρέπει τα δεδομένα σε ένα γράφο και χρησιμοποιεί τις ιδιοτιμές του γράφου για να ομαδοποιήσει τα δεδομένα.

Καθένας από αυτούς τους αλγορίθμους έχει τα πλεονεκτήματά του και τις προϋποθέσεις του. Η επιλογή της κατάλληλης τεχνικής συσταδοποίησης εξαρτάται από τα δεδομένα, την εφαρμογή και τους στόχους της ανάλυσης (Martinez, Martinez & Solka, 2017).

3.3.4 Άλλες σύγχρονες τεχνικές

Πέρα από τις παραδοσιακές τεχνικές ανάλυσης δεδομένων, οι σύγχρονες τεχνικές διερευνητικής ανάλυσης έχουν αναπτυχθεί για την αντιμετώπιση συγκεκριμένων προκλήσεων και τη βελτίωση της απόδοσης των μεθόδων ανάλυσης. Ένα παράδειγμα είναι η βαθιά μάθηση (Deep Learning), όπου τα νευρωνικά δίκτυα βαθιάς μάθησης έχουν επαναπροσδιορίσει πολλούς τομείς της ανάλυσης δεδομένων, αναγνωρίζοντας πολύπλοκα πρότυπα και παράγοντας προβλέψεις από μεγάλα σύνολα δεδομένων, συμπεριλαμβανομένων εικόνων, ήχου και κειμένου (Pearson, 2018).

Ακόμη, τα αναδρομικά νευρωνικά δίκτυα επιτρέπουν την αναδρομική επεξεργασία δεδομένων, εξετάζοντας σχέσεις και μοτίβα σε δεδομένα που επιτρέπουν την αναφορά σε προηγούμενες καταστάσεις. Μία άλλη ισχυρή τεχνική είναι οι μηχανές υποστήριξης βαθμονόμησης (Support Vector Machines - SVMs), οι οποίες χρησιμοποιούνται για ταξινόμηση και προβλέψεις παλινδρόμησης, ιδιαίτερα όταν τα δεδομένα είναι μη γραμμικά διαχωρίσιμα ή έχουν υψηλή διάσταση.

Η αυτόματη μηχανική μάθηση (Automated Machine Learning - AutoML) αυτοματοποιεί τη διαδικασία επιλογής, εκπαίδευσης και βελτιστοποίησης μοντέλων μηχανικής μάθησης, μειώνοντας την ανάγκη για ανθρώπινη παρέμβαση και

επιτρέποντας στους αναλυτές να επικεντρωθούν στην ερμηνεία των αποτελεσμάτων. Παράλληλα, οι μέθοδοι εξόρυξης δεδομένων επικεντρώνονται στην ανακάλυψη προτύπων, τάσεων και συσχετίσεων σε μεγάλα σύνολα δεδομένων, και χρησιμοποιούνται σε εφαρμογές όπως η ανίχνευση απάτης και η πρόβλεψη των αναγκών των πελατών (Pearson, 2018).

Η μεταφορά μάθησης (Transfer Learning) αξιοποιεί τη γνώση που έχει αποκτηθεί από την εκπαίδευση ενός μοντέλου σε ένα πρόβλημα και την εφαρμόζει σε ένα διαφορετικό πρόβλημα, ιδιαίτερα χρήσιμη όταν τα δεδομένα εκπαίδευσης είναι περιορισμένα. Τα πιθανολογικά γραφικά μοντέλα επιτρέπουν την αναπαράσταση των δεδομένων σε μορφή γράφου, όπου οι κόμβοι αντιστοιχούν σε μεταβλητές και οι ακμές στις συσχετίσεις μεταξύ τους, χρήσιμα για την ανάλυση δεδομένων με σημαντικές σχέσεις μεταξύ μεταβλητών.

Η μη αρνητική παραγοντοποίηση μητρώου (Non-negative Matrix Factorization - NMF) αναλύει μη αρνητικά δεδομένα, παράγοντας παραγοντοποιημένους πίνακες που διατηρούν τη μη αρνητικότητα. Οι τεχνικές ανίχνευσης ανωμαλιών επιτρέπουν την αναγνώριση απρόβλεπτων παρατηρήσεων ή συμπεριφορών σε ένα σύνολο δεδομένων, κρίσιμες για την ανίχνευση απάτης και την ασφάλεια δεδομένων.

Αυτές οι σύγχρονες τεχνικές ανάλυσης δεδομένων αντιπροσωπεύουν μόνο μερικές από τις πολλές προσεγγίσεις που χρησιμοποιούνται σήμερα για την κατανόηση, την εξόρυξη και την αξιοποίηση των πληροφοριών που περιέχονται στα δεδομένα, επιδεικνύοντας την πολυπλοκότητα και την καινοτομία στον τομέα της επιστήμης δεδομένων (Pearson, 2018).

3.4 Εργαλεία Οπτικοποίησης στη Διερευνητική Ανάλυση Δεδομένων

Η διερευνητική ανάλυση δεδομένων (EDA) στη μοριακή βιολογία χρησιμοποιεί μια ποικιλία εργαλείων οπτικοποίησης για να βοηθήσει τους ερευνητές να κατανοήσουν τα δεδομένα τους. Αυτά τα εργαλεία μπορούν να χωριστούν σε διάφορες κατηγορίες (Pearson, 2018):

1. Βασικά Γραφήματα Στατιστικής:

- **Ιστογράμματα:** Εμφανίζουν την κατανομή μιας μεταβλητής, αποκαλύπτοντας το σχήμα, το κέντρο και τη διασπορά των δεδομένων. Αυτό βοηθά στην αναγνώριση ακραίων τιμών, στην αξιολόγηση της κανονικότητας και στον εντοπισμό πιθανών μοτίβων.
- **Διαγράμματα Σκέδασης (Scatter plots):** Εμφανίζουν τη σχέση μεταξύ δύο μεταβλητών, βοηθώντας στον εντοπισμό συσχετίσεων, συστάδων ή τάσεων στα δεδομένα.
- **Box Plots:** Παρέχουν μια συνοπτική σύνοψη της κατανομής των δεδομένων, εμφανίζοντας τη διάμεσο, τα τεταρτημόρια και τυχόν ακραίες τιμές. Αυτό βοηθά στη σύγκριση της κατανομής μεταξύ διαφορετικών ομάδων ή συνθηκών.

2. Εργαλεία Οπτικοποίησης για Ανάλυση Εκφράσης Γονιδίων:

- **Heatmaps:** Εμφανίζουν τα επίπεδα έκφρασης γονιδίων σε διαφορετικά δείγματα ή συνθήκες χρησιμοποιώντας μια κλίμακα χρωμάτων. Αυτό βοηθά στον εντοπισμό μοτίβων γονιδιακής έκφρασης, διαφορικά εκφρασμένων γονιδίων και συστάδων γονιδίων με παρόμοια προφίλ έκφρασης.
- **Volcano Plots:** Οπτικοποιούν τα αποτελέσματα της διαφορικής ανάλυσης έκφρασης γονιδίων, απεικονίζοντας τη στατιστική σημαντικότητα έναντι του μεγέθους του αποτελέσματος. Βοηθούν στον εντοπισμό γονιδίων που είναι σημαντικά υπερεκφρασμένα ή υποεκφρασμένα.
- **MA Plots:** Εμφανίζουν την κατανομή του log fold change σε σχέση με τη μέση έκφραση για κάθε γονίδιο. Βοηθούν στην αξιολόγηση της ποιότητας των δεδομένων και στον εντοπισμό πιθανών προκαταλήψεων.

3. Εργαλεία Οπτικοποίησης για Ανάλυση Δικτύων:

- **Cytoscape:** Μια πλατφόρμα λογισμικού ανοιχτού κώδικα για την οπτικοποίηση και ανάλυση βιολογικών δικτύων. Επιτρέπει στους

ερευνητές να δημιουργούν διαδραστικές αναπαραστάσεις δικτύων γονιδίων, πρωτεϊνών ή μεταβολιτών και να εκτελούν διάφορες αναλύσεις δικτύου, όπως εντοπισμό συστάδων ή μονοπατιών.

- **STRING:** Μια βάση δεδομένων και ένα εργαλείο web που παρέχουν προβλέψεις για τις αλληλεπιδράσεις πρωτεϊνών. Οπτικοποιεί τις προβλεπόμενες αλληλεπιδράσεις ως δίκτυο, βοηθώντας τους ερευνητές να διερευνήσουν λειτουργικές σχέσεις μεταξύ πρωτεϊνών.
- **BioGRID:** Μια δημόσια αποθήκη αλληλεπιδράσεων και γενετικών σχέσεων. Παρέχει εργαλεία οπτικοποίησης για την εξερεύνηση αλληλεπιδράσεων πρωτεϊνών, γενετικών αλληλεπιδράσεων και φαινοτύπων.

4. Εργαλεία Οπτικοποίησης για Μείωση Διαστάσεων:

- **t-SNE plots:** Μια τεχνική μη γραμμικής μείωσης διαστάσεων που συχνά χρησιμοποιείται για την οπτικοποίηση κυτταρικών υποπληθυσμών σε δεδομένα αλληλουχίας RNA ενός κυττάρου (scRNA-seq). Βοηθά στον εντοπισμό μοτίβων και ομάδων στα δεδομένα που μπορεί να μην είναι εμφανή σε υψηλότερες διαστάσεις.
- **UMAP plots:** Μια άλλη τεχνική μείωσης διαστάσεων που μπορεί να χρησιμοποιηθεί για την οπτικοποίηση πολύπλοκων δεδομένων, όπως δεδομένα scRNA-seq. Συχνά συγκρίνεται με το t-SNE και προσφέρει παρόμοια οφέλη στην ανακάλυψη προτύπων και σχέσεων στα δεδομένα.

Αυτά είναι μόνο μερικά παραδείγματα εργαλείων οπτικοποίησης που χρησιμοποιούνται συνήθως στη διερευνητική ανάλυση δεδομένων στη μοριακή βιολογία. Η επιλογή των κατάλληλων εργαλείων εξαρτάται από τον συγκεκριμένο τύπο δεδομένων και τους ερευνητικούς στόχους. Συνολικά, αυτά τα εργαλεία επιτρέπουν στους ερευνητές να αποκτήσουν πολύτιμες γνώσεις από πολύπλοκα βιολογικά δεδομένα, οδηγώντας σε νέες ανακαλύψεις και προόδους στον τομέα (Pearson, 2018).

3.5 Εφαρμογή τεχνικών στα δεδομένα της Μοριακής Βιολογίας: Παραδείγματα εφαρμογών

Η εφαρμογή τεχνικών ανάλυσης δεδομένων στη μοριακή βιολογία ανοίγει νέους ορίζοντες για την κατανόηση των μοριακών μηχανισμών που διέπουν τη ζωή. Από την αποκωδικοποίηση του γονιδιώματος μέχρι την ανάλυση πρωτεϊνών και μεταβολιτών, οι τεχνικές ανάλυσης δεδομένων προσφέρουν πολλά στη μοριακή βιολογία (Abzalov & Abzalov, 2016).

Ένα κλασικό παράδειγμα είναι η ανάλυση γονιδιωματικής εκφραστικότητας. Αυτή περιλαμβάνει τη μέτρηση της εκφραστικότητας των γονιδίων σε διάφορες συνθήκες, όπως ασθένειες ή στάδια ανάπτυξης, χρησιμοποιώντας τεχνικές όπως η μικροσυστοιχία DNA ή η ακολουθιακή ανάλυση RNA. Αυτή η ανάλυση αποκαλύπτει γονίδια που εκφράζονται διαφορετικά σε διάφορες συνθήκες και μπορεί να αναδείξει σημαντικά μοριακά μοτίβα.

Η ανάλυση διαδραστικών δικτύων αφορά την ανίχνευση των αλληλεπιδράσεων μεταξύ πρωτεϊνών μέσω τεχνικών όπως η μαζική φασματομετρία ή η πρωτεομική. Αυτές οι τεχνικές αποκαλύπτουν το δίκτυο των πρωτεϊνών που αλληλεπιδρούν, αποκαλύπτοντας τη λειτουργία ολόκληρων βιολογικών συστημάτων (Abzalov & Abzalov, 2016).

Η ανάλυση δομής και λειτουργίας πρωτεϊνών περιλαμβάνει την πρόβλεψη της δομής και της λειτουργίας πρωτεϊνών μέσω υπολογιστικών μεθόδων, όπως η πυρηνική μαγνητική αντήχηση ή η προσομοίωση μοριακής δυναμικής. Αυτές οι τεχνικές επιτρέπουν την κατανόηση της λειτουργικής δομής των πρωτεϊνών και των αλληλεπιδράσεών τους με άλλα μόρια.

Η ανίχνευση ανωμαλιών στη γενετική αλληλουχία χρησιμοποιεί αλγόριθμους μηχανικής μάθησης για την ανίχνευση ανωμαλιών στις γενετικές αλληλουχίες. Αυτό είναι χρήσιμο για την ανίχνευση γενετικών παθήσεων ή την πρόβλεψη των αντιστοιχιών των γονιδίων (Abzalov & Abzalov, 2016).

Υπάρχουν πολλά παραδείγματα εφαρμογών τεχνικών ανάλυσης δεδομένων στη μοριακή βιολογία. Ένα από αυτά είναι η ανάλυση διασποράς γονιδιωμάτων σε διαφορετικά βιολογικά παραδείγματα. Με τη χρήση τεχνικών ανάλυσης γονιδιωμάτων, ερευνητές μπορούν να συγκρίνουν τη διασπορά των γονιδιωμάτων μεταξύ διαφορετικών ειδών ή μεταξύ υγιών και ασθενών ατόμων, οδηγώντας στην ανακάλυψη

γονιδιακών μοτίβων που συσχετίζονται με συγκεκριμένες βιολογικές καταστάσεις (Paliy & Shankar, 2016).

Η ανάλυση δικτύου πρωτεϊνών σε πάθηση επιτρέπει την ανακάλυψη των πρωτεϊνικών αλληλεπιδράσεων που συμβαίνουν σε συγκεκριμένες παθήσεις, αποκαλύπτοντας κλειδιά πρωτεϊνικά δίκτυα που είναι υπεύθυνα για την πάθηση και μπορεί να οδηγήσουν στην ανάπτυξη νέων θεραπειών.

Η πρόβλεψη δομής πρωτεϊνών και αλληλεπιδράσεων μέσω υπολογιστικών μεθόδων επιτρέπει την πρόβλεψη της τρισδιάστατης δομής των πρωτεϊνών και των αλληλεπιδράσεών τους, κρίσιμη για τη σχεδίαση φαρμάκων και την κατανόηση της λειτουργίας των πρωτεϊνών σε κυτταρικό επίπεδο (Paliy & Shankar, 2016).

Η ανίχνευση γονιδιακών μεταλλαγών μέσω της ανάλυσης γονιδίων και της σύγκρισης γενετικών αλληλουχιών μεταξύ υγιών και ασθενών ατόμων επιτρέπει την ανίχνευση γονιδιακών μεταλλάξεων που συσχετίζονται με παθήσεις όπως ο καρκίνος.

Η μελέτη εκφραστικών προφίλ γονιδίων αποκαλύπτει πώς αλληλοεπιδρούν τα γονίδια σε διάφορες συνθήκες, όπως η νόσος ή η αντίδραση σε θεραπεία, παρέχοντας σημαντική κατανόηση για τη βιολογία των παθήσεων και την ανάπτυξη φαρμάκων.

Αυτά τα παραδείγματα αποδεικνύουν τη σημαντική συμβολή των τεχνικών ανάλυσης δεδομένων στη μοριακή βιολογία και την κατανόηση των πολύπλοκων βιολογικών διεργασιών (Paliy & Shankar, 2016).

3.6 Προβληματισμοί

Η διερευνητική ανάλυση δεδομένων συχνά συνοδεύεται από προβληματισμούς και ερωτήσεις που αναδύονται κατά τη διάρκεια της διαδικασίας. Αυτοί οι προβληματισμοί μπορούν να αποτελέσουν κατευθυντήριες γραμμές για περαιτέρω έρευνα ή να υποδείξουν τυχόν αδυναμίες στα δεδομένα ή στη μεθοδολογία ανάλυσης. Ανάλογα με τον τύπο των δεδομένων και τον σκοπό της έρευνας, μπορούν να προκύψουν διάφοροι είδη προβληματισμών (Jebb, Parrigon & Woo, 2017).

Μερικοί από αυτού είναι οι εξής (Jebb, Parrigon & Woo, 2017):

1. **Αντιφατικά Αποτελέσματα:** Κάποιες φορές, οι αναλύσεις μπορεί να οδηγήσουν σε αντιφατικά αποτελέσματα ή σε ευρήματα που δεν είναι συμβατά

με τις αρχικές υποθέσεις. Αυτό μπορεί να απαιτήσει περαιτέρω έρευνα για να κατανοήσουμε την πηγή των αντιφάσεων.

2. **Ελλείπουσες Πληροφορίες ή Δεδομένα:** Η έλλειψη ορισμένων δεδομένων μπορεί να καθιστά δύσκολη την ερμηνεία ή να περιορίζει την εύρεση συνολικών προτύπων. Σε αυτήν την περίπτωση, οι ερευνητές μπορεί να πρέπει να εξετάσουν τις δυνατότητες επέκτασης του συνόλου δεδομένων ή τις εναλλακτικές μεθόδους αντιμετώπισης του προβλήματος.
3. **Μη Αναπαράσταση της Πραγματικότητας:** Οι αναλύσεις μπορεί να μην αντανακλούν πλήρως την πραγματικότητα λόγω περιορισμένης αναπαράστασης ή λόγω παραγόντων που δεν έχουν ληφθεί υπόψη. Αυτό μπορεί να απαιτήσει προσεκτική επανεξέταση των μεθόδων ανάλυσης ή τη χρήση εναλλακτικών τεχνικών.
4. **Πιθανή Στρέβλωση ή Ανεπάρκεια Δεδομένων:** Είναι σημαντικό να εξετάζονται πιθανές στρεβλώσεις ή ανεπάρκειες στα δεδομένα, όπως προκαταλήψεις στη συλλογή των δεδομένων ή μη αντιπροσωπευτικότητα του δείγματος. Αυτό μπορεί να απαιτήσει επιπλέον ανάλυση ή διορθώσεις στη μεθοδολογία.
5. **Σημασία των Αποτελεσμάτων:** Παρά την ανάλυση των δεδομένων, μπορεί να είναι δύσκολο να ερμηνευθεί η πραγματική σημασία των αποτελεσμάτων ή η εφαρμογή τους στο πραγματικό κόσμο. Αυτό μπορεί να ενισχυθεί από προβληματισμούς σχετικά με την εξωτερική επικύρωση των ευρημάτων.

Οι παραπάνω προβληματισμοί αναδεικνύουν τη σημασία της κριτικής ανάλυσης και του συνεχούς ερωτηματισμού κατά τη διαδικασία της διερευνητικής ανάλυσης δεδομένων. Η αντιμετώπιση αυτών των προβληματισμών μπορεί να οδηγήσει σε βελτιωμένη κατανόηση και ερμηνεία των δεδομένων, καθώς και σε πιο αξιόπιστα αποτελέσματα (Jebb, Parrigon & Woo, 2017).

4 Συζήτηση

4.1 Ταξινόμηση βιολογικών δεδομένων

Η ταξινόμηση αποδίδει τα δεδομένα σε προκαθορισμένες κατηγορικές ετικέτες κλάσεων. Μια ταξινόμηση είναι ένα χαρακτηριστικό ή μια ιδιότητα σε ένα σύνολο δεδομένων για την οποία ενδιαφέρεται περισσότερο ένας ερευνητής. Ορίζεται ως η εξαρτημένη μεταβλητή στη στατιστική. Για την ταξινόμηση δεδομένων, ένας αλγόριθμος ταξινόμησης δημιουργεί ένα μοντέλο ταξινόμησης που αποτελείται από κανόνες ταξινόμησης. Για παράδειγμα, οι ασφαλιστές αυτοκινήτων έχουν αναπτύξει μοντέλα ταξινόμησης για να κατηγοριοποιούν τις εφαρμογές των οδηγών ως «επικίνδυνες» ή «ασφαλείς». Στον ιατρικό τομέα, η ταξινόμηση μπορεί να χρησιμοποιηθεί για να βοηθήσει στον καθορισμό της ιατρικής διάγνωσης και πρόγνωσης με βάση τα συμπτώματα, το ιστορικό οικογενειακών ασθενειών και τις συνθήκες υγείας (Han, Pei & Tong, 2022).

Η ταξινόμηση είναι μια διαδικασία δύο βημάτων που αποτελείται από εκπαίδευση και δοκιμή. Στο πρώτο βήμα, κατασκευάζεται ένα μοντέλο εκπαίδευσης το οποίο αποτελείται από κανόνες ταξινόμησης. Το μοντέλο εκπαίδευσης «διδάσκεται» πρώτα τους κανόνες. Στη συνέχεια, όταν το μοντέλο εκπαίδευσης «δοκιμάζεται», δηλαδή όταν του παρουσιάζονται νέα δεδομένα που δεν προέρχονται από τα δεδομένα εκπαίδευσης, το μοντέλο εκπαίδευσης είναι σε θέση να λάβει μια απόφαση με βάση όσα έχει μάθει από τα δεδομένα εκπαίδευσης.

Ορισμένοι κανόνες ταξινόμησης χρησιμοποιούν έναν μαθηματικό τύπο για τον προσδιορισμό μιας ετικέτας κλάσης. Οι κανόνες ταξινόμησης δεν είναι απαραίτητα 100% αληθείς- γενικά, οι κανόνες με ακρίβεια 90-95% θεωρούνται σταθεροί κανόνες. Η ακρίβεια ενός ταξινομητή (ή μοντέλου ταξινόμησης) εξαρτάται από το βαθμό στον οποίο οι κανόνες ταξινόμησης είναι αληθείς (Hand, 2007).

Το δεύτερο βήμα, η δοκιμή, εξετάζει έναν ταξινομητή χρησιμοποιώντας δεδομένα δοκιμής για την ακρίβεια. Οι ετικέτες κλάσης για τα δεδομένα δοκιμής είναι γνωστές. Ο ταξινομητής αναμένεται να προβλέψει την ετικέτα κλάσης για κάθε περίπτωση δοκιμής με βάση τον τρόπο διδασκαλίας του. Γενικά, η διαδικασία δοκιμής είναι πολύ απλή και υπολογιστικά ανέξοδη σε σύγκριση με το βήμα εκπαίδευσης, το οποίο μπορεί να είναι πολύπλοκο και να απαιτεί σημαντικούς υπολογιστικούς πόρους.

Μια ενδιαφέρουσα τεχνική στην ταξινόμηση είναι η προσέγγιση ensemble. Η λογική πίσω από την προσέγγιση ensemble είναι ότι πολλαπλοί ταξινομητές (ή μοντέλα ταξινόμησης) που συνεργάζονται μπορούν να αποδώσουν καλύτερη ακρίβεια ταξινόμησης από τη χρήση ενός μόνο ταξινομητή. Ως απλό παράδειγμα, εάν οι ταξινομητές A, B και Γ προβλέπουν ότι ένας δύσκολα ταξινομήσιμος ασθενής (ασθενής 1) έχει μια ασθένεια και οι ταξινομητές Δ και Ε προβλέπουν ότι ο ασθενής 1 δεν έχει αυτή την ασθένεια, τότε, με τη χρήση μιας στρατηγικής ψηφοφορίας, ο ensemble προβλεπτής θα προέβλεπε ότι ο ασθενής 1 έχει την ασθένεια. Σε ορισμένες περιπτώσεις, σε κάθε ταξινομητή μπορεί να ανατίθενται διαφορετικά βάρη και η τελική πρόβλεψη του συνόλου θα είναι τότε ένας σταθμισμένος μέσος όρος των ψήφων των ταξινομητών. Η μελέτη περίπτωσης ταξινόμησης σε αυτό το κεφάλαιο διερευνά τη μέθοδο ensemble με περισσότερες λεπτομέρειες (Han, Pei & Tong, 2022).

4.2 Συσταδοποίηση βιολογικών δεδομένων

Η ομαδοποίηση ορίζεται ως μάθηση χωρίς επίβλεψη που πραγματοποιείται παρατηρώντας μόνο τις ανεξάρτητες μεταβλητές (σε αντίθεση με την επιβλεπόμενη μάθηση που αναλύει τόσο τις ανεξάρτητες μεταβλητές όσο και τις εξαρτημένες μεταβλητές). Με άλλα λόγια, σε αντίθεση με την ταξινόμηση, η ομαδοποίηση δεν χρησιμοποιεί «τάξη». Στην πραγματικότητα, αυτή είναι η κύρια διαφορά μεταξύ ταξινόμησης και ομαδοποίησης. Για τον λόγο αυτό, η ομαδοποίηση μπορεί να χρησιμοποιηθεί καλύτερα για μελέτες διερευνητικού χαρακτήρα, ιδίως εάν οι εν λόγω μελέτες περιλαμβάνουν μεγάλο όγκο δεδομένων, αλλά πολύ λίγα είναι γνωστά για τα δεδομένα (όπως η μάζα των δεδομένων που παράγονται συνήθως από την ανάλυση μικροσυστοιχιών) (Hand, 2007).

Η συσταδοποίηση χρησιμοποιείται για την ομαδοποίηση αντικειμένων σε συγκεκριμένο αριθμό συστάδων έτσι ώστε τα αντικείμενα εντός μιας συστάδας να έχουν πολύ υψηλή ομοιότητα και τα αντικείμενα από διαφορετικές συστάδες να έχουν πολύ χαμηλή ομοιότητα. Οι ομοιότητες μεταξύ δύο αντικειμένων μετρώνται χρησιμοποιώντας τις τιμές των χαρακτηριστικών τους. Μια πολύ πρώιμη εφαρμογή της συσταδοποίησης στη βιολογία ήταν η ομαδοποίηση παρόμοιων φυτών και ζώων για τη δημιουργία ταξινομιών με βάση τα χαρακτηριστικά τους (όπως ο αριθμός των πετάλων

και ο αριθμός των ποδιών). Τις τελευταίες δεκαετίες έχουν εισαχθεί και χρησιμοποιηθεί διάφοροι αλγόριθμοι συσταδοποίησης. Αυτοί οι αλγόριθμοι κατηγοριοποιούνται κυρίως σε ιεραρχικούς και μερικούς.

Οι αλγόριθμοι ιεραρχικής συσσωρευτικής ομαδοποίησης συγχωνεύουν διαδοχικά τις δύο πιο παρόμοιες ομάδες αντικειμένων με βάση τις αποστάσεις ανά ζεύγη μεταξύ δύο ομάδων αντικειμένων μέχρι να ισχύει μια συνθήκη τερματισμού, ώστε τα αντικείμενα να ομαδοποιηθούν ιεραρχικά. Για το λόγο αυτό, οι ίδιοι οι ιεραρχικοί αλγόριθμοι μπορούν να κατηγοριοποιηθούν αποτελεσματικά ανάλογα με τις αντίστοιχες μεθόδους υπολογισμού της ομοιότητας (ή της απόστασης) μεταξύ δύο ομάδων αντικειμένων. Με άλλα λόγια, η κατηγοριοποίηση αυτή βασίζεται στον τρόπο με τον οποίο επιλέγεται το αντιπροσωπευτικό αντικείμενο κάθε ομάδας για τον υπολογισμό της ομοιότητας (Hand, 2007).

Ενώ η ιεραρχική ομαδοποίηση είναι συσσωρευτική, δηλαδή ξεκινά με ατομικά στοιχεία και τα συγκεντρώνει σε ομάδες αυξανόμενου μεγέθους, η διαχωριστική ομαδοποίηση ξεκινά με το ένα πλήρες σύνολο δεδομένων και υποδιαίρει το σύνολο δεδομένων σε μικρότερα τμήματα. Ένας αλγόριθμος διααιρετικής συσταδοποίησης εκτελεί επαναληπτικά τα εξής δύο υποπροβλήματα (Savaresi et al., 2002):

1. αποφασίζει την καλύτερη συστάδα που πρέπει να διαχωριστεί
2. αποφασίζει τον καλύτερο τρόπο διαχωρισμού της επιλεγμένης συστάδας.

Σε αντίθεση με τους αλγορίθμους ιεραρχικής ομαδοποίησης, οι αλγόριθμοι μερικής ομαδοποίησης απαιτούν από τον χρήστη να εισάγει μια παράμετρο k , η οποία είναι ο αριθμός των ομάδων. Γενικά, οι αλγόριθμοι *partitional* μεταφέρουν τα αντικείμενα απευθείας σε k συστάδες. Οι *partitional* αλγόριθμοι κατηγοριοποιούνται ανάλογα με τον τρόπο με τον οποίο μετατοπίζουν τα αντικείμενα, τον τρόπο με τον οποίο επιλέγουν ένα κεντροειδές συστάδας (ή εκπρόσωπο) μεταξύ των αντικειμένων εντός μιας (ατελούς) συστάδας και τον τρόπο με τον οποίο μετρούν τις ομοιότητες μεταξύ των αντικειμένων και των κεντροειδών συστάδων. Για παράδειγμα, ο *kmeans*, ο πιο ευρέως χρησιμοποιούμενος *partitional* αλγόριθμος, επιλέγει πρώτα τυχαία k κεντροειδή (αντικείμενα) και στη συνέχεια αποσυνθέτει τα αντικείμενα σε k διαχωρισμένες ομάδες μετατοπίζοντας επαναληπτικά τα αντικείμενα με βάση την ομοιότητα μεταξύ των κεντροειδών και των αντικειμένων. Στον *k-means*, το κεντροειδές μιας συστάδας είναι

η μέση τιμή των αντικειμένων της συστάδας. Σε πολλές περιπτώσεις, τα κεντροειδή της συστάδας δεν είναι πραγματικά αντικείμενα της συστάδας. Σε αντίθεση με το k-means, το k-medoids επιλέγει το κοντινότερο αντικείμενο στη μέση τιμή των αντικειμένων σε μια συστάδα (Savaresi et al., 2002).

Το σημαντικότερο πλεονέκτημα των αλγορίθμων ομαδοποίησης κατά τμήματα έναντι άλλων μεθόδων, είναι η ανώτερη ακρίβεια ομαδοποίησης σε σύγκριση με τους αλγορίθμους ιεραρχικής ομαδοποίησης, η οποία είναι αποτέλεσμα της συνολικής στρατηγικής βελτιστοποίησής τους (δηλαδή των αναδρομικών μετατοπίσεων των αντικειμένων). Επιπλέον, οι partitional αλγόριθμοι μπορούν να χειριστούν μεγάλα σύνολα δεδομένων, τα οποία οι ιεραρχικοί αλγόριθμοι δεν μπορούν (δηλ. καλύτερη κλιμάκωση) και μπορούν να ομαδοποιήσουν τα δεδομένα πιο γρήγορα. Εν ολίγοις, οι partitional αλγόριθμοι είναι πιο αποτελεσματικοί και αποδοτικοί από τους ιεραρχικούς αλγορίθμους. Ένα σημαντικό μειονέκτημα στη χρήση των partitional αλγορίθμων είναι ότι τα αποτελέσματα της συσταδοποίησής τους εξαρτώνται σε κάποιο βαθμό από τα αρχικά κεντροειδή των συστάδων, επειδή τα κεντροειδή επιλέγονται τυχαία. Έτσι, τα αποτελέσματα ομαδοποίησης που λαμβάνονται είναι λίγο διαφορετικά κάθε φορά που εκτελείται ο partitional αλγόριθμος. Μια τέτοια διαδικασία είναι γνωστή ως μη ντετερμινιστική. Μια ντετερμινιστική διαδικασία, συγκριτικά, θα δίνει το ίδιο αποτέλεσμα κάθε φορά που εκτελείται (Hand, 2007).

4.3 Προοπτικές και προκλήσεις για μελλοντική έρευνα

4.3.1 Εξόρυξη βιολογικών δεδομένων στο νέφος

Οι εξελίξεις στα υπολογιστικά εργαλεία βιολογικής ανάλυσης πρέπει να συμβαδίζουν με τα βιολογικά Big Data, δηλαδή την εκθετική αύξηση των βιολογικών γονιδιωματικών δεδομένων. Το υπολογιστικό νέφος είναι «ένα μοντέλο για την παροχή πανταχού παρούσας, βολικής, δικτυακής πρόσβασης κατά παραγγελία σε μια κοινή δεξαμενή διαμορφώσιμων υπολογιστικών πόρων (π.χ. δίκτυα, διακομιστές, αποθήκευση, εφαρμογές και υπηρεσίες) που μπορούν να παρέχονται και να απελευθερώνονται γρήγορα με ελάχιστη προσπάθεια διαχείρισης ή αλληλεπίδραση με τον πάροχο υπηρεσιών» (Gunarathne et al., 2013).

Τα εργαλεία που βασίζονται στο νέφος και διαθέτουν δυνατότητες μαζικής παράλληλης επεξεργασίας μπορούν να χρησιμοποιηθούν για τη σύγχρονη υπολογιστική βιολογική έρευνα σε μεγάλα δεδομένα. Το μοντέλο υπολογιστικού νέφους προσφέρει μια εναλλακτική λύση στις ακριβές υπολογιστικές και αποθηκευτικές ανάγκες των ερευνητικών υπολογισμών. Η ευθυγράμμιση αλληλουχιών DNA ή RNA ανά ζεύγη, ο υπολογισμός αποστάσεων, η ομαδοποίηση, η πολυδιάστατη κλιμάκωση και η οπτικοποίηση γονιδιακών αλληλουχιών είναι υπολογιστικά εντατικές λειτουργίες. Ομοίως υπολογιστικά δαπανηροί είναι οι διαδραστικοί παράλληλοι υπολογισμοί που μπορούν να χρησιμοποιηθούν για τον προσδιορισμό των παραμέτρων διαφορικών εξισώσεων κατά την επίλυση προβλημάτων συμπερασμού γονιδιακών δικτύων.

Ένα πρόβλημα Μεγάλων Δεδομένων μπορεί να επιλυθεί με το σπάσιμό του σε μικρότερα μέρη, την ταυτόχρονη παράλληλη επίλυση των μικρότερων μερών και, στη συνέχεια, τη συγκέντρωση των επιμέρους αποτελεσμάτων. Ο χρόνος που εξοικονομείται με αυτή την προσέγγιση «παράλληλης επεξεργασίας» ονομάζεται επιτάχυνση. Ο νόμος του Amdahl (Εξ. 1) που ορίζει την «Επιτάχυνση» χρησιμοποιείται συχνά στον παράλληλο υπολογισμό για την πρόβλεψη της θεωρητικής μέγιστης επιτάχυνσης με τη χρήση πολλαπλών επεξεργαστών (Gunarathne et al., 2013).

$$Amdahl's Law : SpeedUp = \frac{1}{(1 - P) + \frac{P}{S}} \quad (1)$$

Στην Εξίσωση 1.1, P είναι το ποσοστό ενός προβλήματος που μπορεί να χωριστεί σε εργασίες παράλληλου υπολογισμού και 1 - P είναι το ποσοστό που δεν μπορεί να παραλληλιστεί, δηλαδή το «διαδοχικό» ποσοστό. Η μέγιστη δυνατότητα επιτάχυνσης βρίσκεται στο S, που σημαίνει ότι το P μπορεί να επιταχυνθεί S φορές με τη χρήση S παράλληλων επεξεργαστών. Η επιτάχυνση ενός προγράμματος που χρησιμοποιεί πολλαπλούς επεξεργαστές στον παράλληλο υπολογισμό περιορίζεται από το χρόνο που απαιτείται για το ακολουθιακό τμήμα του προγράμματος. Για παράδειγμα, υποθέστε ότι ένα πρόγραμμα χρειάζεται 20 ώρες χρησιμοποιώντας έναν πυρήνα ενός επεξεργαστή. Υποθέστε ότι ένα συγκεκριμένο τμήμα του προγράμματος (δηλαδή το 1 - P) που χρειάζεται μία ώρα (δηλαδή 5%) για να εκτελεστεί δεν μπορεί να παραλληλιστεί. Οι υπόλοιπες 19 ώρες (δηλ. 95%) του χρόνου εκτέλεσης μπορούν να παραλληλιστούν (δηλ. P). Σε αυτή την περίπτωση, ανεξάρτητα από το πόσοι

επεξεργαστές αφιερώνονται σε μια παραλληλοποιημένη εκτέλεση αυτού του προγράμματος, ο ελάχιστος χρόνος εκτέλεσης δεν μπορεί να είναι μικρότερος από μία ώρα. Ως εκ τούτου, η επιτάχυνση περιορίζεται το πολύ στο 20. Ο νόμος του Amdahl αφορά την επιτάχυνση που μπορεί να επιτευχθεί με την «παραλληλοποίηση» του ποσοστού P ενός υπολογισμού όπου η βελτίωση έχει επιτάχυνση S . Ο νόμος του Amdahl περιγράφει με σαφήνεια την αποτελεσματικότητα και τον περιορισμό του παράλληλου υπολογισμού και έχει υιοθετηθεί ευρέως.

Το MapReduce είναι ένα κατανεμημένο πλαίσιο ανάπτυξης λογισμικού επεξεργασίας δεδομένων που αναπτύχθηκε από την Google για την υποστήριξη της παράλληλης κατανεμημένης εκτέλεσης των εφαρμογών έντασης δεδομένων της. Η προσέγγιση MapReduce χρησιμοποιεί το «διαίρει και βασίλευε» για να επιταχύνει την επεξεργασία των μεγάλων δεδομένων. Με τη χρήση του MapReduce, μια λύση επεξεργασίας δεδομένων αποτελείται από βήματα map και reduce. Η Google χρησιμοποιεί αυτό το πλαίσιο εσωτερικά για την εκτέλεση χιλιάδων εφαρμογών MapReduce την ημέρα, επεξεργαζόμενη peta-bytes δεδομένων, και όλα αυτά σε commodity hardware. Μεταξύ των φάσεων map και reduce εκτελείται μια εσωτερική φάση shuffle για τον χειρισμό των ενδιάμεσων αποτελεσμάτων. Το πλαίσιο MapReduce εκτελεί αυτόματα αυτές τις λειτουργίες παράλληλα σε οποιονδήποτε αριθμό επεξεργαστών (Dean & Ghemawat, 2008).

Το πλαίσιο MapReduce περιέχει δύο κύριες φάσεις, map και reduce, οι οποίες ελέγχονται από ένα πρόγραμμα οδήγησης σε ένα καθορισμένο «κύριο» μηχάνημα. Στη φάση map, το πρόγραμμα οδήγησης προωθεί τμήματα δεδομένων εισόδου σε πολλαπλούς υπολογιστικούς κόμβους που ορίζονται ως «slave» μηχανές. Ο οδηγός MapReduce δίνει εντολή στους υπολογιστικούς κόμβους να εκτελέσουν υπολογισμούς σύμφωνα με ένα πρόγραμμα που έχει οριστεί από τον χρήστη για τη διαδικασία map. Τα αποτελέσματα του χάρτη αποθηκεύονται σε άμεσα αρχεία (Dean & Ghemawat, 2008).

Στη φάση μείωσης (αφού ολοκληρωθούν οι υπολογισμοί στους χαρτογράφους), ο οδηγός κατευθύνει τους υπολογιστικούς κόμβους να συγκεντρώσουν τα αποτελέσματα από τα ενδιάμεσα αρχεία. Αυτά τα ενδιάμεσα αποτελέσματα επεξεργάζονται χρησιμοποιώντας ένα πρόγραμμα που έχει οριστεί από τον χρήστη για το στάδιο reduce

(AKA reducer). Στη συνέχεια, οι μειωτές συνεισφέρουν τα αποτελέσματά τους για να σχηματίσουν την πλήρη έξοδο.

Το MapReduce παρέχει ένα εύχρηστο μοντέλο προγραμματισμού που διαθέτει ανοχή σε σφάλματα, αυτόματο παραλληλισμό, επεκτασιμότητα και βελτιστοποιήσεις βασισμένες στην εντοπιότητα δεδομένων. Πολλοί αλγόριθμοι επιστημονικών υπολογισμών που βασίζονται σε επαναληπτικούς υπολογισμούς μπορούν να υλοποιηθούν με έναν υπολογισμό MapReduce που καθορίζεται για κάθε επαναληπτικό βήμα (Ghemawat, Gobioff & Leung, 2003).

Ένας σημαντικός παράγοντας για την επιτυχία του MapReduce είναι το καινοτόμο κατακεντρωμένο σύστημα αρχείων που αναπτύχθηκε από την Google και ονομάζεται Google File System (GFS). Το GFS είναι εξαιρετικά ανεκτικό σε σφάλματα για να οφείλεται σε εκτεταμένη αναπαραγωγή δεδομένων σε φθηνό εξοπλισμό αποθήκευσης βασικών προϊόντων.

Το μοντέλο προγραμματισμού MapReduce παρέχει ένα εύκολα υλοποιήσιμο πλαίσιο με δυνατότητες ανοχής σφαλμάτων. Αυτό το μοντέλο έχει χρησιμοποιηθεί για την επιτυχή επίλυση πολλών επιστημονικών υπολογιστικών προβλημάτων μεγάλης κλίμακας, συμπεριλαμβανομένων προβλημάτων στις βιοεπιστήμες. Ο στόχος του MapReduce είναι η ανάπτυξη ενός μεγάλου όγκου εργασιών που απαιτούν χρόνο και μνήμη σε πολλούς υπολογιστικούς κόμβους που επεξεργάζονται τις εργασίες παράλληλα εκτελώντας αλγορίθμους που ορίζονται από τον χρήστη. Η ροή της διαδικασίας MapReduce περιλαμβάνει μια κύρια μηχανή και πολλές δευτερεύουσες μηχανές που εκτελούν εργασίες MapReduce σύμφωνα με τις οδηγίες της διεργασίας οδηγού στην κύρια μηχανή (Ghemawat, Gobioff & Leung, 2003).

Το Apache Hadoop είναι μια ευρέως χρησιμοποιούμενη εφαρμογή ανοιχτού κώδικα του κατακεντρωμένου πλαισίου επεξεργασίας δεδομένων Google MapReduce. Το Apache Hadoop χρησιμοποιεί το Hadoop Distributed File System (HDFS) για αποθήκευση δεδομένων, το οποίο αποθηκεύει τα δεδομένα στους τοπικούς δίσκους των κόμβων υπολογιστών ενώ παρουσιάζει μια ενιαία προβολή συστήματος αρχείων μέσω της διεπαφής προγράμματος εφαρμογής HDFS (API). Το HDFS είναι μια εφαρμογή ανοιχτού κώδικα του GFS. Όπως το GFS, το HDFS προορίζεται να εκτελείται σε εξοπλισμό αποθήκευσης εμπορευμάτων. Το HDFS παρέχει την

υποστήριξη καταναμημένου συστήματος αρχείων για τη διασφάλιση υψηλής απόδοσης και ανοχής σφαλμάτων. Υψηλά επίπεδα αξιοπιστίας επιτυγχάνονται μέσω της αναπαραγωγής δεδομένων. Η απόδοση διέλευσης βελτιστοποιείται με τον προγραμματισμό της μεταφοράς δεδομένων από το αντίγραφο δεδομένων που βρίσκεται πλησιέστερα στη θέση του κόμβου υπολογισμού. Το Apache Hadoop εκτελεί διπλότυπες εκτελέσεις πιο αργών εργασιών και χειρίζεται αποτυχίες εκτελώντας ξανά τις αποτυχημένες εργασίες χρησιμοποιώντας διαφορετικά μηχανήματα εργαζομένων (Hashem et al., 2016).

Η βιολογική έρευνα που βασίζεται στο σύννεφο είναι επί του παρόντος πολύ ενεργή. Το Stormbow, που αναφέρθηκε προηγουμένως, αναλύει δείγματα αλληλουχίας νουκλεοτιδίων RNA (δηλαδή RNA-Seq) χρησιμοποιώντας πόρους που βασίζονται σε σύννεφο. Το AWS ήταν η υποδομή που χρησιμοποιήθηκε για την ανάπτυξη του Stormbow. Ο Stormbow χρειάστηκε 6 έως 8 ώρες για να επεξεργαστεί ένα δείγμα RNA-Seq που περιείχε πάνω από 100 εκατομμύρια μεμονωμένα μόρια RNA.

Η βιοπληροφορική τίθεται υπό αμφισβήτηση από ολοένα και μεγαλύτερα σύνολα δεδομένων που οδηγούν σε υπολογιστικές εργασίες που απαιτούν απαράδεκτα μεγάλους χρόνους εάν πραγματοποιηθούν σε μικρό αριθμό μηχανών. Για αυτές τις περιπτώσεις, ο καταναμημένος υπολογισμός σε πολλαπλά συμπλέγματα σε διαφορετικές τοποθεσίες γίνεται μια ελκυστική, αν όχι απαραίτητη, προσέγγιση για την επίτευξη βραχυπρόθεσμων χρόνων. Το Galaxy, για παράδειγμα, είναι μια ισχυρή πλατφόρμα ανοιχτού συστήματος που βασίζεται στο διαδίκτυο για βιοϊατρική έρευνα εντατικής δεδομένων. Το Galaxy επιτρέπει στον χρήστη να εκτελεί, να αναπαράγει και να μοιράζεται πλήρεις αναλύσεις στον ελεύθερα διαθέσιμο διακομιστή ή σε μια παρουσία ιδιωτικού διακομιστή (Hashem et al., 2016).

4.3.2 Συμπεράσματα βιολογικού δικτύου με βάση το σύννεφο

Οι πρωτόγνωρες ποσότητες βιολογικών δεδομένων που παράγονται σε μια ποικιλία μορφών με εξίσου πρωτόγνωρες ταχύτητες παρουσιάζουν ένα μείγμα ευλογιών και προκλήσεων για τους ερευνητές της βιοπληροφορικής. Αυτές οι τρεις διαστάσεις των προκλήσεων, δηλαδή η ποικιλία, η ταχύτητα και ο όγκος, αποτελούν το τρέχον φαινόμενο «μεγάλων δεδομένων» στη βιοπληροφορική. Καινοτόμες μεθοδολογίες

στην υπολογιστική βιολογία, συμπεριλαμβανομένης της μείωσης διαστάσεων, της επιλογής χαρακτηριστικών, της παραλληλοποίησης και του υπολογιστικού νέφους, μετριάζουν τις πολυπλοκότητες υψηλού όγκου επιταχύνοντας τη διαδικασία μετατροπής μεγάλων δεδομένων σε χρήσιμη γνώση. Δεν υπάρχει αμφιβολία ότι η εξόρυξη βιολογικών μεγάλων δεδομένων και η αποκάλυψη λεπτομερειών σχετικά με τις ρυθμιστικές αλληλεπιδράσεις των γονιδίων θα οδηγήσει σε μεγαλύτερη κατανόηση των κυτταρικών λειτουργιών (Laney, 2001).

Τα προφίλ γονιδιακής έκφρασης χρονοσειρών αποκαλύπτουν το κυτταρικό μεταγράφημα σε τακτά χρονικά διαστήματα. Χαρτογραφώντας τις αναγνώσεις σε αυτά τα προφίλ στην αντίστοιχη αλληλουχία γονιδιώματος αναφοράς τους, μπορεί να αναγνωριστεί το γονίδιο πηγής για κάθε μεταγραφή. Με την εφαρμογή μεθοδολογιών αμοιβαίας πληροφόρησης (MI), μπορεί να κατασκευαστεί ένα κατευθυνόμενο ακυκλικό γράφημα (DAG) για την πρόβλεψη ενός δικτύου γονιδιακής ρύθμισης (GRN) για τον οργανισμό. Στο DAG, οι δύο κόμβοι κάθε ακμής θα αντιπροσώπευαν ένα ζεύγος γονιδίων έτσι ώστε ένα γονίδιο να ρυθμίζει την έκφραση του δεύτερου γονιδίου. Η έρευνα για το συμπέρασμα των GRN παράγει πολλά υποσχόμενα αποτελέσματα (Lee, Hsiao & Hwang, 2014).

Καμία έρευνα συμπερασμάτων GRN που χρησιμοποιεί μεθοδολογίες MI μέχρι σήμερα δεν έχει αξιοποιήσει την τεράστια ικανότητα παράλληλης επεξεργασίας του υπολογιστικού νέφους σε συνδυασμό με τη βελτίωση της γνώσης της Οντολογίας Γονιδίων (GO) για τη βελτίωση των αποτελεσμάτων πρόβλεψης. Θα εφαρμόσουμε το συμπέρασμα GRN χρησιμοποιώντας μεθοδολογίες MI και τελειοποίηση αποτελεσμάτων GO σε περιβάλλον 83 που βασίζεται σε σύννεφο για να προβλέψουμε ένα ρυθμιστικό δίκτυο γονιδίων (GRN) τόσο από πειραματικά σύνολα δεδομένων RNA-Seq όσο και από συνθετικά σύνολα δεδομένων χρονοσειράς. Επιπλέον, θα επιλέξουμε σύνολα δεδομένων για τα οποία υπάρχουν γνωστά ή προβλεπόμενα GRN διαθέσιμα για ανάλυση συγκριτικής αξιολόγησης. Θα δείξουμε ότι αυτή η μέθοδος που βασίζεται σε σύννεφο μπορεί εύκολα να κλιμακωθεί κατά πολλαπλές τάξεις μεγέθους με μικρό αντίκτυπο στην απόδοση.

Αυτή η νέα εφαρμογή που βασίζεται σε σύννεφο για να συμπεράνουμε ένα GRN θα είναι ένα σημαντικό εργαλείο βιοπληροφορικής. Οι πόροι του cloud γίνονται όλο και πιο ευέλικτοι και προσιτοί σε σύγκριση με τους τοπικούς παραδοσιακούς

υπολογιστικούς πόρους. Οι πόροι του cloud είναι διαθέσιμοι κατόπιν ζήτησης με ελάχιστη οικονομική ή χρονική δέσμευση. Τα πλεονεκτήματα του cloud computing στον τομέα της έρευνας στη βιοπληροφορική είναι γνωστά (Lee, Hsiao & Hwang, 2014).

4.3.3 Αλγόριθμος εξαγωγής συμπερασμάτων Γονιδιακού Ρυθμιστικού Δικτύου (GRN) με χρήση MI και GO

Αλγόριθμος εξαγωγής γονιδιακού ρυθμιστικού δικτύου με χρήση MI και GO, ενσωματώνοντας Map-Reduce σε πλατφόρμα Elastic Cloud Computing (EC2) του Amazon Web Services (AWS), με χρήση του ανοιχτού κώδικα Apache Hadoop, εφαρμόζοντας το πλαίσιο παράλληλης επεξεργασίας MapReduce της Google. Ο αλγόριθμος αυτός εφαρμόζεται σε σύνολα δεδομένων χρονοσειράς RNA-Seq.

Η διαδικασία ξεκινά με την αξιολόγηση των προφίλ έκφρασης γονιδίων για την εύρεση σημείου αλλαγής της έκφρασης, χρησιμοποιώντας προσέγγιση MI. Στη συνέχεια, δημιουργείται ένα γονιδιακό ρυθμιστικό δίκτυο (GRN) που απεικονίζει τις επιδράσεις μεταξύ των γονιδίων. Αυτό το δίκτυο κλαδεύεται με βάση τη βάση δεδομένων GO και τις τροποποιήσεις της ανισότητας επεξεργασίας δεδομένων.

Στη συνέχεια, ο κύριος υπολογιστικός κόμβος προετοιμάζει τα δεδομένα για τους κόμβους υπολογισμού Slave. Κάθε γονίδιο συγκρίνεται με όλα τα υπόλοιπα γονίδια για την εντοπισμό του σημείου αλλαγής στην έκφραση. Αν η αλλαγή υπερβαίνει ένα ορισμένο κατώφλι, καθορίζεται το αρχικό σημείο αλλαγής έκφρασης για το γονίδιο. Έπειτα, υπολογίζεται η επίδραση ενός γονιδίου σε κάθε άλλο, και αυτό γίνεται παράλληλα για πολλά ζεύγη γονιδίων. Η παραλληλοποίηση αυτή επιτρέπει την αποτελεσματική επεξεργασία μεγάλων συνόλων δεδομένων.

Με αυτόν τον τρόπο, ο αλγόριθμος εξάγει πληροφορίες για το γονιδιακό ρυθμιστικό δίκτυο από σύνολα δεδομένων χρονοσειράς RNA-Seq, χρησιμοποιώντας τη δύναμη του πλαισίου MapReduce για την επεξεργασία των δεδομένων σε παράλληλη βάση.

Σύμφωνα με το πλαίσιο MapReduce, στο Hadoop, μια συνάρτηση χάρτη ορίζεται για την επεξεργασία δεδομένων ως ένα ζεύγος στοιχείων. Για την έρευνά μας, το πρώτο

στοιχείο του ζεύγους είναι ένα μοναδικό αναγνωριστικό ζεύγους γονιδίων, όπου το πρώτο γονίδιο έχει τιμή IcE μικρότερη από το δεύτερο. Αυτό το αναγνωριστικό παραμένει ίδιο για όλα τα ζευγάρια γονιδίων εντός μιας εργασίας χάρτη. Το δεύτερο στοιχείο του ζεύγους είναι τα δύο πλήρη σύνολα εκφράσεων γονιδίων για κάθε γονίδιο στο ζεύγος.

Μια συνάρτηση Reduce, σύμφωνα με το ίδιο πλαίσιο, επεξεργάζεται δεδομένα ως ένα ζεύγος στοιχείων που προκύπτουν από τη συνάρτηση Map. Σε αυτήν την έρευνα, η Reduce ενοποιεί όλα τα ζευγάρια γονιδίων που παράγονται από το Mapper βάσει του πρώτου γονιδιακού στοιχείου. Επιπλέον, αξιολογεί τη βάση δεδομένων GO για να επιβεβαιώσει τη ρυθμιστική ικανότητα του πρώτου γονιδιακού στοιχείου στο ζεύγος. Οι κόμβοι Reducer παράγουν ένα ζεύγος κλειδιού-τιμής για κάθε μοναδικό πρώτο γονίδιο.

Το κλειδί κάθε ζεύγους περιλαμβάνει ένα μοναδικό αριθμητικό αναγνωριστικό γονιδίου, ενώ η τιμή περιέχει έναν πίνακα τριπλών δεδομένων. Κάθε τρίπλο δεδομένο περιλαμβάνει το δεύτερο γονίδιο του ζεύγους, την τιμή INFL και την τιμή πιστοποίησης GO που σχετίζεται με το συγκεκριμένο ζεύγος γονιδίων. Αυτή η δυαδική τιμή προσδιορίζει εάν το γονίδιο έχει πιστοποιηθεί ως έχον ρυθμιστική λειτουργία ή όχι. Τα δεδομένα που παράγονται από τη συνάρτηση Reduce αξιολογούνται από τον κύριο υπολογιστικό κόμβο.

Στο τέλος της διαδικασίας Map και Reduce, ο κύριος υπολογιστικός κόμβος εξετάζει τη συλλογή των τιμών πιστοποίησης INFL και GO για τα ζεύγη γονιδίων. Αυτές οι πληροφορίες χρησιμοποιούνται για τη δημιουργία ενός γραφήματος κατευθυνόμενων ακμών (DAG) που αντιπροσωπεύει τις προβλέψεις ενός γονιδιακού ρυθμιστικού δικτύου (GRN) για τον εκάστοτε οργανισμό. Αυτό το DAG περιλαμβάνει μια κατευθυνόμενη ακμή για κάθε ζεύγος γονιδίων στη συλλογή, ενώ τα άκρα του επισημαίνονται ως "πιστοποιημένο GO" εάν το γονίδιο έχει επιβεβαιωθεί ότι διαθέτει ρυθμιστική λειτουργία. Το DAG ελέγχεται για κατευθυνόμενες ακμές προς τις δύο κατευθύνσεις, και σε κάθε τέτοια περίπτωση, αφαιρούνται οι ακμές που δεν είναι "πιστοποιημένο GO" και έχουν μικρότερη τιμή INFL από το δεύτερο άκρο.

Το προκύπτον DAG αντιπροσωπεύει μια πρόβλεψη του GRN για τον εκάστοτε οργανισμό. Αυτή η πρόβλεψη μπορεί να συγκριθεί με παρόμοιες προβλέψεις που

παράχθηκαν για το ίδιο προφίλ γονιδιακής έκφρασης, προσφέροντας έτσι σημαντική πληροφορία για τη λειτουργία του GRN στον εκάστοτε οργανισμό.

Για την εμπλουτισμένη κατανόηση του GRN, προστίθενται δύο νέοι αλγόριθμοι. Ο πρώτος είναι ο αλγόριθμος αλληλεπίδρασης γονιδίων, ο οποίος χρησιμοποιεί τεχνικές μηχανικής μάθησης για να ανιχνεύσει κρυφές σχέσεις μεταξύ γονιδίων σε μεγάλα σύνολα δεδομένων. Αυτός ο αλγόριθμος εστιάζει στην αξιολόγηση της στατιστικής σημαντικότητας των σχέσεων που προκύπτουν, παρέχοντας μια επιπλέον διάσταση στην κατανόηση των γονιδιακών αλληλεπιδράσεων. Ο δεύτερος είναι ο αλγόριθμος ομοιογένειας γονιδίων, ο οποίος εφαρμόζει μεθόδους ομαδοποίησης για να εντοπίσει ομάδες γονιδίων με παρόμοια εκφραστικά προφίλ. Αυτή η μέθοδος μπορεί να προσφέρει πληροφορίες σχετικά με την πιθανή συνεργασία των γονιδίων μέσα σε συγκεκριμένα βιολογικά μονοπάτια, ενισχύοντας έτσι την ερμηνεία των δεδομένων.

Αυτοί οι αλγόριθμοι συμβάλλουν στην καλύτερη κατανόηση των γονιδιακών ρυθμιστικών δικτύων, διευρύνοντας τη δυνατότητα ανάλυσης και ερμηνείας των γονιδιακών αλληλεπιδράσεων και της λειτουργίας τους στο πλαίσιο των βιολογικών διεργασιών.

4.4 Συνεισφορά στον τομέα της Μοριακής Βιολογίας και ευρύτερα

Δύο περιπτώσιολογικές μελέτες εξόρυξης βιολογικών δεδομένων σε όλο το γονιδίωμα παρουσιάζονται παρακάτω. Η πρώτη μελέτη περίπτωσης είναι ένα παράδειγμα ανακάλυψης βιολογικών προτύπων όπου οι δευτερογενείς δομές ενός RNA αξιολογούνται χρησιμοποιώντας ένα μοντέλο συνδιακύμανσης. Η δεύτερη μελέτη περίπτωσης είναι ένα παράδειγμα ταξινόμησης βιολογικών δεδομένων όπου οι τριτοταγείς δομές RNA αξιολογούνται χρησιμοποιώντας έναν τυχαίο ταξινομητή δασών.

4.4.1 Μια μελέτη περίπτωσης στην εξόρυξη δεδομένων δευτερογενούς δομής RNA

Παρουσιάζεται ένα παράδειγμα μελέτης περίπτωσης στον τομέα της εξόρυξης δεδομένων δευτερογενούς δομής RNA, με τον στόχο της ανάλυσης βιολογικών προτύπων. Συγκεκριμένα, η μελέτη επικεντρώνεται στην ανίχνευση και ταξινόμηση λειτουργικών περιοχών μη κωδικοποιητικού RNA σε είδη της μύγας των φρούτων (Byron et al., 2010).

Αναφέρεται πώς εξελικτικά διατηρημένες περιοχές RNA στο χρωμόσωμα X της μύγας των φρούτων (roX1) έχουν ταυτοποιηθεί σε οκτώ είδη του γένους *Drosophila*. Ενδιαφέρον παρουσιάζει η ανακάλυψη διατηρημένων πρωτογενών αλληλουχιών εντός των RNA roX1 αυτών των ειδών. Επιπλέον, αναφέρεται η χρήση ενός μοντέλου συνδιακύμανσης για την εντοπισμό των χαρακτηριστικών των λειτουργικών περιοχών roX1, με σκοπό την ταξινόμηση παρόμοιων παραδειγμάτων RNA σε άλλα είδη *Drosophila* (Khaladkar et al., 2008).

Παρά τις υψηλές δραστηριότητες γονιδιωματικής αλληλουχίας, ο σχολιασμός των ειδών *Drosophila* παραμένει ατελής, με πολλές πληροφορίες για τις συντεταγμένες των χρωμοσωμάτων να παραμένουν άγνωστες. Ως αποτέλεσμα, πραγματοποιήθηκε σάρωση ολόκληρων γονιδιωμάτων της *Drosophila* για την αναγνώριση διαθέσιμων σχολιασμένων περιοχών, όπως χρωμοσώματα ή ικριώματα (Byron et al., 2010).

Με τη χρήση γνωστών παραδειγμάτων roX1 για σύγκριση, προτείνεται ότι η πρόβλεψη νέων λειτουργικών περιοχών roX1 είναι εφικτή με ακρίβεια μόνο από πληροφορίες αλληλουχίας. Η ανάλυση του roX1 σε γονιδιωματική κλίμακα παρέχει πληροφορίες για τις εξελικτικές διαδικασίες μεταξύ διαφορετικών ειδών. Τα αποτελέσματα της μελέτης υποστηρίζουν την αποτελεσματικότητα μιας διαδικασίας αναζήτησης και ταξινόμησης μέσω μοντέλου συνδιακύμανσης στην εξόρυξη γονιδίων RNA roX1. Επιπλέον, λόγω της ευελιξίας της μεθοδολογίας αναζήτησης, η προσέγγιση αυτή είναι πιθανό να επιτύχει και σε παρόμοιες αναζητήσεις σε διάφορους άλλους οργανισμούς. Η μελέτη αυτή αναδεικνύει τη σημασία της εφαρμογής προηγμένων μεθόδων ανάλυσης δεδομένων στην ανακάλυψη και την κατανόηση βιολογικών προτύπων. Επιπλέον, ανοίγει τον δρόμο για περαιτέρω έρευνα σε παρόμοιους τομείς, προσφέροντας ένα πλαίσιο για την ανάπτυξη νέων μεθόδων εξόρυξης δεδομένων στον τομέα της

βιολογίας και της γενετικής. Η εφαρμογή τέτοιων μεθόδων μπορεί να συμβάλει στην κατανόηση της λειτουργίας και της εξέλιξης των οργανισμών, ενισχύοντας έτσι την επιστημονική κοινότητα στον τομέα της βιολογίας και της γενετικής (Eddy, 2001).

Το βασικό εύρημα της μελέτης ήταν η κατάταξη στοιχείων λειτουργικής δομής, δηλαδή του μη-κωδικοποιητικού RNA, σε γονίδια ειδών *Drosophila* ως πιθανά ομόλογα του roX1. Αυτό επιτεύχθηκε μέσω μιας προσέγγισης που προσαρμόστηκε στο πλαίσιο της μελέτης. Αρχικά, επιλέχθηκαν οκτώ αλληλουχίες μεταγραφών RNA roX1, και στη συνέχεια χρησιμοποιήθηκε η μέθοδος "slide and fold" για την κατασκευή δομών RNA, με βάση τις θερμοδυναμικές τους ιδιότητες (Eddy, 2001).

Συγκεκριμένα, αυτή η προσέγγιση περιλάμβανε την ανάπτυξη δομών RNA με διπλώσεις υποακολουθιών μήκους 20 νουκλεοτιδίων ή μικρότερου, με επικάλυψη 50 νουκλεοτιδίων μεταξύ τους. Οι δομές RNA προβλέφθηκαν με χρήση του πακέτου Vienna RNAfold, ενώ πραγματοποιήθηκαν συγκρίσεις ανά ζεύγη ειδών για την αξιολόγηση της ακρίβειας των προβλέψεων. Η διαδικασία αυτή επιτελέστηκε με χρήση ενός συστήματος παράλληλης επεξεργασίας, το οποίο επέτρεψε την ταχεία ολοκλήρωση περίπου 520.000 ευθυγραμμίσεων ανά ζεύγη σε λιγότερο από πέντε λεπτά (Eddy, 2001).

Στη συνέχεια, επιλέχθηκε μια ομάδα τριών δομών RNA από τα είδη *D. melanogaster*, *D. simulans* και *D. yakuba*, οι οποίες είχαν παρόμοια βαθμολογία ευθυγράμμισης και μήκος, για περαιτέρω ανάλυση. Αυτή η ομάδα δημιούργησε ένα μοντέλο συνδιακύμανσης (CM) από τις τρεις δομές, με χρήση του προγράμματος Infernal CMbuild. Τα πλήρη γονιδιώματα οκτώ ειδών *Drosophila*, στα οποία επιβεβαιώθηκε η παρουσία μεταγραφών roX1 ncRNA, χρησιμοποιήθηκαν ως στόχοι για τις αναζητήσεις CM (Vollmers et al., 2022).

Οι αναζητήσεις CM εντόπισαν τα γονίδια roX1 σε έξι είδη *Drosophila*, ενώ απέτυχαν να τα εντοπίσουν σε δύο άλλα είδη, τα *D. mojavensis* και *D. virilis*. Σε πέντε από τις έξι επιτυχημένες αναζητήσεις, η βαθμολογία ευθυγράμμισης υποδήλωνε μια αλληλουχία εντός του γνωστού εύρους γονιδιακών συντεταγμένων roX1 για το εκάστοτε είδος. Η έκτη επιτυχημένη αναζήτηση στο *D. pseudoobscura* παρήγαγε το τρίτο αποτέλεσμα αναζήτησης με τη μεγαλύτερη βαθμολογία, που αντιπροσώπευε μια αλληλουχία εντός

του γνωστού εύρους γονιδιακών συντεταγμένων roX1 για αυτό το είδος (Vollmers et al., 2022).

Για να διασφαλιστεί η ακρίβεια και η ταχύτητα των υπολογισμών, τα αρχεία γονιδιωμάτων διαχωρίστηκαν σε μικρότερα αρχεία ανά 2 εκατομμύρια ζεύγη βάσεων (Mbp) και εκτελέστηκαν πολλαπλές αναζητήσεις CM παράλληλα, χρησιμοποιώντας ένα σύμπλεγμα υπολογιστών υψηλής απόδοσης (HPC). Αυτή η προσέγγιση εξασφάλισε την ακρίβεια και την ταχύτητα των υπολογισμών, καθώς η αναζήτηση CM ενός ολόκληρου γονιδιώματος διήρκεσε περίπου 10 λεπτά (Storz, 2002).

Η μελέτη επικεντρώνεται επίσης στην αναγνώριση πιθανών στοιχείων λειτουργικής δομής roX1 σε γονιδιώματα τεσσάρων ειδών *Drosophila*, στα οποία δεν είχε ακόμη επιβεβαιωθεί η παρουσία μεταγραφών roX1. Αυτά τα ευρήματα υποδεικνύουν υψηλή πιθανότητα διατήρησης της λειτουργικότητας του roX1 μεταξύ των ειδών. Για παράδειγμα, η ανάλυση των αλληλουχιών γονιδιώματος roX1 ανάμεσα στα *D. melanogaster* και *D. sechellia* απέδειξε υψηλό ποσοστό ομοιότητας, φτάνοντας το 94%. Αυτό υποδεικνύει ότι η λειτουργικότητα του roX1 είναι πιθανό να διατηρείται μεταξύ αυτών των δύο ειδών (Storz, 2002).

Η μελέτη παρέχει σημαντικές πληροφορίες σχετικά με την παρουσία και τη διατήρηση του roX1 σε διάφορα είδη *Drosophila*, καθώς και την πιθανή ομολογία του μεταξύ αυτών των ειδών. Η χρήση προηγμένων υπολογιστικών μεθόδων, όπως η ευθυγράμμιση και η ανάλυση αλληλουχιών, παρείχε σημαντικές ενδείξεις για τη διατήρηση της λειτουργικότητας του γονιδίου roX1 σε διαφορετικά είδη (Wimberly et al., 2000).

Ωστόσο, απαιτούνται πειραματικές επαληθεύσεις για την επιβεβαίωση αυτών των αποτελεσμάτων. Η περαιτέρω έρευνα μπορεί να περιλαμβάνει πειράματα σε εργαστήριο για την επαλήθευση της λειτουργικότητας των περιοχών roX1 στα είδη *Drosophila* που μελετήθηκαν. Επίσης, η σύγκριση μεταξύ περισσότερων ειδών μπορεί να αποκαλύψει περαιτέρω πληροφορίες σχετικά με την εξέλιξη και τη διατήρηση του γονιδίου roX1 στο γένος *Drosophila*. Αυτές οι πληροφορίες θα μπορούσαν να είναι χρήσιμες για την κατανόηση της λειτουργίας και της εξέλιξης του γενετικού υλικού σε αυτά τα είδη (Wimberly et al., 2000).

Στη μελέτη περίπτωσης, αναπτύχθηκε μια συστηματική και υπολογιστικά αποτελεσματική προσέγγιση για την αναγνώριση και ταξινόμηση των δομικών στο. Επισημαίνεται ότι η προσέγγιση αυτή περιλαμβάνει τρία βασικά στάδια (Pavlopoulos et al., 2023):

1. Σύγκριση Δομών RNA: Αρχικά, πραγματοποιείται η σύγκριση των δομών RNA roX1 μεταξύ όλων των ειδών.
2. Επιλογή Ομάδων Δομής RNA: Στη συνέχεια, επιλέγονται ομάδες δομής RNA που σχετίζονται σημαντικά με άλλα είδη.
3. Χρήση Μεθοδολογίας Δομικής Αναζήτησης: Τέλος, χρησιμοποιείται μια μεθοδολογία δομικής αναζήτησης με μεγάλη ευαισθησία (όπως τα μοντέλα συνδιακύμανσης), η οποία είναι επίσης ευέλικτη.

Η εφαρμογή αυτής της προσέγγισης επιτρέπει τη βελτίωση της αναπαράστασης ενός συμπλέγματος δομών RNA, με την προσθήκη ή αφαίρεση δομών ανάλογα με τις ανάγκες. Επιπλέον, η παράλληλη επεξεργασία συμβάλλει στην αντιμετώπιση των μακρών χρόνων επεξεργασίας. Αυτή η μέθοδος εφαρμόζεται για την αναγνώριση μικρών δομών RNA, καθώς αυτές μπορούν να ταξινομηθούν με μεγαλύτερη ακρίβεια σε σύγκριση με τις μεθόδους που βασίζονται μόνο στη θερμοδυναμική ελαχιστοποίηση (Pavlopoulos et al., 2023).

Για να αξιολογηθεί η αποτελεσματικότητα διαφορετικών εργαλείων αναζήτησης, πραγματοποιήθηκε σύγκριση μεταξύ της αναζήτησης BLAST και της Infernal σε αναζήτηση διατηρημένων δομικών μοτίβων. Τα αποτελέσματα έδειξαν ότι η Infernal είχε καλύτερη απόδοση από το BLAST, καθώς αυτό τείνει να αποτύχει στην ανίχνευση ορισμένων δομών σε σχέση με το Infernal (Park et al., 2007).

Συνολικά, αυτή η μελέτη επιβεβαιώνει την παρουσία λειτουργικών περιοχών roX1 στο χρωμόσωμα X σε διάφορα είδη *Drosophila*, ενώ επιπλέον αναδεικνύει την πολυπλοκότητα που εμπεριέχεται στην ταξινόμηση των μοτίβων ncRNA. Η σύγκριση ομολογίας μεταξύ ειδών επιβεβαιώνει τη διατήρηση της λειτουργικότητας του γονιδίου σε διαφορετικά είδη, καθιστώντας δυνατή την κατανόηση της σημασίας του στο ευρύτερο πλαίσιο της εξέλιξης (Park et al., 2007).

4.4.2 Συμπέρασμα

Η μελέτη περίπτωσης εντάσσεται στο πλαίσιο της διερευνητικής ανάλυσης δεδομένων (EDA) καθώς εστιάζει στην ανακάλυψη μοτίβων και σχέσεων σε μεγάλα σύνολα δεδομένων, χωρίς την ύπαρξη προκαθορισμένων υποθέσεων. Στόχος της είναι η εξαγωγή χρήσιμων πληροφοριών και η διατύπωση νέων ερευνητικών ερωτημάτων. Στη συγκεκριμένη περίπτωση, η ανάλυση δευτερογενών δομών RNA αποτελεί ένα παράδειγμα εφαρμογής της EDA στη μοριακή βιολογία. Μέσω της χρήσης εργαλείων όπως το RSmatch και το Infernal, οι ερευνητές μπορούν να ανακαλύψουν μοτίβα που υποδεικνύουν λειτουργικές ή δομικές ομοιότητες μεταξύ διαφορετικών μορίων RNA (Pavlopoulos et al., 2023).

Η ανακάλυψη μοτίβων σε δευτερογενείς δομές RNA μπορεί να οδηγήσει σε νέες υποθέσεις σχετικά με τη λειτουργία αυτών των μορίων. Για παράδειγμα, η παρουσία συγκεκριμένων μοτίβων σε μη κωδικοποιητικά RNA (ncRNA) μπορεί να υποδηλώνει τη συμμετοχή τους σε σημαντικές κυτταρικές διεργασίες, όπως η ρύθμιση της γονιδιακής έκφρασης ή η σταθερότητα του mRNA. Επιπλέον, η σύγκριση μοτίβων μεταξύ διαφορετικών ειδών μπορεί να αποκαλύψει εξελικτικές σχέσεις και να οδηγήσει σε νέες υποθέσεις σχετικά με τη λειτουργία και την προέλευση αυτών των μορίων.

Τα αποτελέσματα της μελέτης περίπτωσης αναδεικνύουν την αποτελεσματικότητα της χρήσης μοντέλων συνδιακύμανσης (CM) και του εργαλείου CSminer για την ανίχνευση ομοαξονικών ελικοειδών στοιβάδων σε συνδέσμους RNA 3 κατευθύνσεων. Η επιτυχής αναγνώριση αυτών των μοτίβων σε διάφορα είδη *Drosophila* υποδηλώνει τη διατήρησή τους κατά την εξέλιξη, γεγονός που υποστηρίζει τη λειτουργική τους σημασία. Επιπλέον, η ανακάλυψη ενός πιθανού ομολόγου roX1 στο *D. sechellia*, το οποίο δεν είχε αναφερθεί προηγουμένως, ανοίγει νέους δρόμους για την κατανόηση της λειτουργίας του roX1 σε αυτό το είδος (Pavlopoulos et al., 2023).

Η σύγκριση μεταξύ της μεθόδου CMsearch του Infernal και του BLAST αποκάλυψε την ανωτερότητα της μεθόδου CMsearch στην ανίχνευση μοτίβων RNA. Αυτό οφείλεται στην ικανότητα των μοντέλων συνδιακύμανσης να λαμβάνουν υπόψη τις συσχετίσεις μεταξύ των βάσεων, οι οποίες είναι κρίσιμες για τη διατήρηση της δευτερογενούς δομής του RNA.

Τα ευρήματα της μελέτης περίπτωσης έχουν σημαντικές εφαρμογές στη μοριακή βιολογία. Η αναγνώριση μοτίβων σε δευτερογενείς δομές RNA μπορεί να βοηθήσει στην κατανόηση της λειτουργίας των ncRNA, τα οποία παίζουν κρίσιμο ρόλο σε πολλές κυτταρικές διεργασίες. Η ανακάλυψη νέων μοτίβων μπορεί να οδηγήσει στην ανάπτυξη νέων θεραπευτικών στόχων για ασθένειες που σχετίζονται με τη δυσλειτουργία των ncRNA. Επιπλέον, η κατανόηση των εξελικτικών σχέσεων μεταξύ των μοτίβων RNA μπορεί να βοηθήσει στην ανασύνθεση της ιστορίας της ζωής και στην κατανόηση των μηχανισμών της εξέλιξης (Vollmers et al., 2022).

Η μελέτη περίπτωσης αναδεικνύει επίσης τη σημασία της ανάπτυξης νέων υπολογιστικών εργαλείων και μεθόδων για την ανάλυση βιολογικών δεδομένων. Η χρήση προηγμένων τεχνικών, όπως η μηχανική μάθηση και η τεχνητή νοημοσύνη, μπορεί να βελτιώσει την ακρίβεια και την αποτελεσματικότητα της ανάλυσης δεδομένων, επιτρέποντας την ανακάλυψη νέων μοτίβων και σχέσεων που δεν ήταν προηγουμένως εφικτή (Vollmers et al., 2022).

4.4.3 Μια μελέτη περίπτωσης στην εξόρυξη δεδομένων τριτογενών δομών RNA

Σε αυτήν την έρευνα παρουσιάζεται η ανάπτυξη ενός υπολογιστικού εργαλείου για την ανίχνευση τρισδιάστατων μοτίβων RNA σε γονιδιωματικές αλληλουχίες, χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων. Το εργαλείο αυτό, γνωστό ως CSminer, χρησιμοποίησε δύο ευέλικτες μεθοδολογίες, τις τυχαίες δάσης και τα μοντέλα συνδιακύμανσης, για να προβλέψει τις θέσεις γονιδιωματικών συντεταγμένων που περιέχουν πιθανώς ομοαξονικές ελικοειδείς στοιβάδες σε συνδέσεις RNA 3 κατευθύνσεων. Η προσέγγιση αυτή επέτρεψε την ανίχνευση του ομοαξονικού ελικοειδούς τριτογενούς μοτίβου στοίβαξης, όπου δύο ξεχωριστά ελικοειδή στοιχεία ευθυγραμμίζονται σε έναν κοινό άξονα, παρέχοντας θερμοδυναμική σταθερότητα στο μόριο RNA. Η αποτελεσματικότητα της προσέγγισης επιβεβαιώθηκε μέσω πειραματικών αποτελεσμάτων (Laing et al., 2012).

Η ανάπτυξη εργαλείων ανίχνευσης προτύπων αποτελεί κύριο ενδιαφέρον για τους βιοπληροφορικούς, οι οποίοι χρησιμοποιούν όλο και πιο προηγμένες υπολογιστικές μεθόδους στη βιολογία. Το CSminer ανέπτυξε μια προσέγγιση για την πρόβλεψη της θέσης των ομοαξονικών ελικοειδών στοίβαξης σε γονίδια, επεκτείνοντας τη

λειτουργικότητα του εργαλείου Infernal. Αυτή η προσέγγιση επιτρέπει την ανίχνευση του τριτογενούς μοτίβου RNA με χρήση τυχαίου ταξινομητή δασών, ενώ κάθε αποτέλεσμα αναζήτησης εκτιμάται για την παρουσία ομοαξονικής στοίβαξης (Laing et al., 2012).

Πραγματοποιήθηκε μια σειρά πειραμάτων για να αξιολογηθεί η αποτελεσματικότητα της ανακάλυψης προτύπων στην προσέγγιση που παρουσιάστηκε σε αυτή τη μελέτη περίπτωσης. Στο πρώτο πείραμα, το CSminer δοκιμάστηκε εναντίον του πλήρους γονιδιώματος του *Deinococcus radiodurans*, με GenBank ID NC 001263.1, που ανακτήθηκε από τη βάση δεδομένων NCBI GenBank. Ανακαλύφθηκε ένα τριτογενές μοτίβο ncRNA, συγκεκριμένα η ομοαξονική ελικοειδής στοίβαξη H2H3, ανάμεσα στις θέσεις 251047 και 251117 στον συν κλώνο του γονιδιώματος. Το CSminer προέβλεψε σωστά αυτήν τη διασταύρωση RNA 3 κατευθύνσεων με ομοαξονική ελικοειδή στοίβαξη τύπου H2H3 (Ban et al., 2000).

Αυτό το αποτέλεσμα επικυρώθηκε με τη χρήση του NCBI BLAST και χειροκίνητων αναλύσεων. Επιπλέον, πραγματοποιήθηκαν πειράματα με τρία βακτηριακά γονιδιώματα που σχετίζονται στενά φυλογενετικά με είδη που αντιπροσωπεύονται στα μοντέλα συνδιακύμανσης που χρησιμοποιήθηκαν. Αυτά τα πειράματα δεν ανέδειξαν σημαντικά αποτελέσματα αναζήτησης σε άλλα βιολογικά βασίλεια. Για τα δύο γονιδιώματα του ιού, το γονίδιο του ζώου, το γονίδιο του μύκητα και το γονίδιο του protista, δεν προβλέφθηκε κανένα μοτίβο, υποδηλώνοντας ότι τα μοτίβα ενδιαφέροντος ήταν συγκεκριμένα για τα γονιδιώματα/είδη που αντιπροσωπεύονταν στα μοντέλα συνδιακύμανσης που χρησιμοποιήθηκαν από την CSminer (Ban et al., 2000).

Τέλος, δεν βρέθηκε κάποια αντιστοίχιση με γνωστά γονίδια RNA που είναι μέλη της οικογένειας RFAM, οπότε η ακρίβεια της ανίχνευσης δεν μπορεί να επιβεβαιωθεί σε αυτές τις περιπτώσεις (Ban et al., 2000).

4.4.4 Συμπέρασμα

Η μελέτη περίπτωσης εστιάζει στην ανίχνευση ομοαξονικών ελικοειδών στοιβάδων σε συνδέσμους RNA 3 κατευθύνσεων, χρησιμοποιώντας συνδυασμό μοντέλων συνδιακύμανσης (CMs) και τυχαίων δασών (Li & Chen, 2023).

Η μελέτη περίπτωσης αποτελεί ένα παράδειγμα εφαρμογής της EDA στη μοριακή βιολογία, καθώς εξερευνά τα δεδομένα για την ανακάλυψη μοτίβων και σχέσεων που δεν ήταν προηγουμένως γνωστά. Η χρήση εργαλείων όπως το CSminer επιτρέπει την ανίχνευση πολύπλοκων μοτίβων σε τριτογενείς δομές RNA, τα οποία μπορεί να μην είναι εμφανή με άλλες μεθόδους. Αυτά τα μοτίβα μπορούν να οδηγήσουν σε νέες υποθέσεις σχετικά με τη λειτουργία και τη δομή του RNA, καθώς και να αποκαλύψουν εξελικτικές σχέσεις μεταξύ διαφορετικών οργανισμών (Li & Chen, 2023).

Τα αποτελέσματα της μελέτης περίπτωσης δείχνουν ότι το CSminer μπορεί να προβλέψει με ακρίβεια την παρουσία ομοαξονικών ελικοειδών στοιβάδων σε γονιδιώματα βακτηρίων. Η επικύρωση των προβλέψεων με γνωστές δομές από τη βάση δεδομένων Protein Data Bank (PDB) ενισχύει την αξιοπιστία του εργαλείου. Επιπλέον, η σύγκριση με το NCBI BLAST αναδεικνύει την ανωτερότητα του CSminer στην ανίχνευση μοτίβων RNA, καθώς λαμβάνει υπόψη τις συσχετίσεις μεταξύ των βάσεων (Greenfield et al., 2010).

Η αναγνώριση ομοαξονικών ελικοειδών στοιβάδων σε συνδέσμους RNA 3 κατευθύνσεων έχει σημαντικές εφαρμογές στη μοριακή βιολογία. Αυτά τα μοτίβα παίζουν κρίσιμο ρόλο στη σταθερότητα και τη λειτουργία του RNA. Η κατανόηση της δομής και της λειτουργίας αυτών των μοτίβων μπορεί να οδηγήσει σε νέες θεραπευτικές προσεγγίσεις για ασθένειες που σχετίζονται με τη δυσλειτουργία του RNA. Επιπλέον, η ανίχνευση αυτών των μοτίβων σε διαφορετικά είδη μπορεί να βοηθήσει στην κατανόηση της εξέλιξης του RNA και των μηχανισμών που διέπουν τη λειτουργία του (Greenfield et al., 2010).

Η ανάπτυξη και εφαρμογή του εργαλείου CSminer αντιμετώπισε αρκετές προκλήσεις. Η κύρια δυσκολία έγκειται στην επιλογή κατάλληλων χαρακτηριστικών για την εκπαίδευση του ταξινομητή τυχαίων δασών. Τα χαρακτηριστικά αυτά έπρεπε να είναι επαρκώς διακριτικά για να επιτρέπουν την ακριβή πρόβλεψη της παρουσίας ομοαξονικών ελικοειδών στοιβάδων, αλλά και αρκετά γενικά ώστε να μπορούν να

εφαρμοστούν σε διαφορετικές αλληλουχίες RNA. Επιπλέον, ο μεγάλος όγκος των δεδομένων που απαιτούνται για την εκπαίδευση του ταξινομητή, καθώς και η πολυπλοκότητα των τριτογενών δομών RNA, αποτελούσαν πρόσθετες προκλήσεις (Berman et al., 2000).

Για την αντιμετώπιση αυτών των προκλήσεων, οι ερευνητές χρησιμοποίησαν μια ποικιλία τεχνικών. Η επιλογή των χαρακτηριστικών βασίστηκε σε προηγούμενη γνώση για τη δομή και τη λειτουργία του RNA, καθώς και σε στατιστική ανάλυση των δεδομένων. Επιπλέον, η χρήση αλγορίθμων μηχανικής μάθησης, όπως τα τυχαία δάση, επέτρεψε την αποτελεσματική επεξεργασία μεγάλου όγκου δεδομένων και την εξαγωγή αξιόπιστων προβλέψεων (Berman et al., 2000).

Η έρευνα στον τομέα της εξόρυξης δεδομένων τριτογενών δομών RNA είναι ακόμα σε εξέλιξη. Μια μελλοντική κατεύθυνση θα μπορούσε να είναι η επέκταση του CSminer για την ανίχνευση και άλλων τύπων μοτίβων RNA, πέρα από τις ομοαξονικές ελικοειδείς στοιβάδες. Επιπλέον, η εφαρμογή του εργαλείου σε άλλους οργανισμούς, πέρα από τα βακτήρια, θα μπορούσε να αποκαλύψει νέα μοτίβα και να διευρύνει την κατανόησή μας για τη δομή και τη λειτουργία του RNA (Waleń et al., 2014).

Μια άλλη πολλά υποσχόμενη κατεύθυνση είναι η ενσωμάτωση πρόσθετων πηγών δεδομένων, όπως πειραματικά δεδομένα από κρυσταλλογραφία ακτίνων X ή φασματοσκοπία NMR, για την περαιτέρω βελτίωση της ακρίβειας των προβλέψεων. Επιπλέον, η ανάπτυξη νέων αλγορίθμων μηχανικής μάθησης, όπως οι αλγόριθμοι βαθιάς μάθησης, θα μπορούσε να οδηγήσει σε ακόμα πιο ισχυρά εργαλεία για την ανάλυση τριτογενών δομών RNA (Waleń et al., 2014).

5 Συμπεράσματα

5.1 Συνολική ανασκόπηση των ευρημάτων

Στην προσέγγιση μελέτης περίπτωσης που εξετάζεται στη μελέτη περίπτωσης στην εξόρυξη δεδομένων δευτερογενούς δομής RNA, τα εργαλεία RSmash και Infernal αποδείχθηκαν αποτελεσματικά στην ανακάλυψη νέων προτύπων ncRNAs. Αν και η αναζήτηση ομολογίας είναι μια διαδεδομένη πρακτική στη βιοπληροφορική, παλαιότερες μέθοδοι όπως το BLAST και το FASTA έχουν παρουσιάσει περιορισμένη ακρίβεια. Η αναζήτηση ομολογίας για το μη κωδικοποιητικό RNA είναι ακόμα πιο προβληματική λόγω των ενδομοριακών συμμεταβλητών ζευγών βάσεων που διατηρούνται σε υψηλό βαθμό (Nawrocki & Eddy, 2013).

Η εκτέλεση ενός ερωτήματος αναζήτησης με το Infernal απαιτεί μεγάλο χρονικό διάστημα, όπως αναφέρεται στη μελέτη. Παράλληλες προσεγγίσεις χρησιμοποιούνται για να βελτιωθεί η απόδοση, συμπεριλαμβανομένων των μεθόδων φιλτραρίσματος HMM και των ευρετικών βασιζόμενων σε ακολουθίες.

Στο ίδιο πλαίσιο, πραγματοποιήθηκε σάρωση ολόκληρων γονιδιωμάτων των 12 ειδών της *Drosophila* με αλληλουχία. Οι φυλογενετικές σχέσεις βασίστηκαν στην υπόθεση ότι πιο πρόσφατα εξελιγμένα είδη θα έχουν μεγαλύτερη γενετική ομοιότητα μεταξύ τους. Η μελέτη επιβεβαίωσε την ύπαρξη του roX1 ncRNA σε έξι είδη *Drosophila*, ενώ παρέχει νέα δεδομένα για την παρουσία του στο *D. Sechellia* (Nawrocki & Eddy, 2013).

Η μελέτη περίπτωσης στην εξόρυξη δεδομένων τριτογενών δομών RNA αναδεικνύει την αποτελεσματικότητα του CSminer, το οποίο συνδυάζει το Infernal με έναν ταξινομητή τυχαίων δασών βασισμένο σε σύνολο. Αυτή η προσέγγιση αποδείχθηκε αποτελεσματική στην εξαγωγή βιολογικών δεδομένων, με έμφαση στην ανακάλυψη μοτίβων RNA ομοαξονικής ελικοειδούς στοίβαξης σε όλο το γονιδίωμα. Αυτή η προσέγγιση είναι εφικτή και οικονομικά αποδοτική, παρέχοντας σημαντική συμβολή στην κατανόηση του μη κωδικοποιητικού RNA.

Συνοψίζοντας, οι μελέτες περίπτωσης παρουσιάζουν την αποτελεσματικότητα διαφόρων εργαλείων και μεθόδων στον τομέα της αναζήτησης και της εξόρυξης δεδομένων RNA. Αυτές οι προσεγγίσεις δεν μόνο βοηθούν στην ανακάλυψη νέων

μοτίβων και προτύπων αλλά και στην κατανόηση της λειτουργίας και της εξέλιξης του μη κωδικοποιητικού RNA σε επίπεδο γονιδιώματος (Nawrocki & Eddy, 2013).

5.2 Συνεισφορά της διερευνητικής ανάλυσης δεδομένων στη Μοριακή Βιολογία

Η Διερευνητική Ανάλυση Δεδομένων (EDA) έχει συνεισφέρει σημαντικά στον τομέα της Μοριακής Βιολογίας παρέχοντας πολύτιμες γνώσεις για πολύπλοκα βιολογικά συστήματα. Ακολουθούν μερικοί τρόποι με τους οποίους η EDA έχει επηρεάσει και έχει προχωρήσει τη Μοριακή Βιολογία (Pearson, 2018):

1. **Ανακάλυψη προτύπων:** Οι τεχνικές EDA βοηθούν στον εντοπισμό προτύπων και τάσεων σε μεγάλα βιολογικά σύνολα δεδομένων, όπως δεδομένα έκφρασης γονιδίων ή δίκτυα αλληλεπίδρασης πρωτεΐνης-πρωτεΐνης. Εξερευνώντας οπτικά τα δεδομένα, οι ερευνητές μπορούν να αποκαλύψουν κρυφά μοτίβα που μπορεί να υποδεικνύουν ρυθμιστικούς μηχανισμούς, βιοδείκτες για ασθένειες ή εξελικτικές σχέσεις μεταξύ των ειδών.
2. **Μείωση διαστάσεων:** Τα σύνολα δεδομένων μοριακής βιολογίας είναι συχνά υψηλών διαστάσεων, περιέχουν χιλιάδες μεταβλητές (π.χ. γονίδια ή πρωτεΐνες) που μετρώνται σε διαφορετικές συνθήκες ή δείγματα. Οι τεχνικές EDA, όπως η ανάλυση κύριου συστατικού (PCA) ή η ενσωμάτωση στοχαστικού γείτονα (t-SNE), επιτρέπουν στους ερευνητές να μειώσουν τη διάσταση των δεδομένων διατηρώντας παράλληλα τη βασική δομή τους. Αυτή η απλοποίηση διευκολύνει την οπτικοποίηση και την ερμηνεία πολύπλοκων βιολογικών φαινομένων.
3. **Ομαδοποίηση και ταξινόμηση:** Οι μέθοδοι EDA διευκολύνουν τη ομαδοποίηση βιολογικών δειγμάτων με βάση τις ομοιότητες στα μοριακά τους προφίλ. Οι αλγόριθμοι ομαδοποίησης όπως η ιεραρχική ομαδοποίηση ή η ομαδοποίηση k-means βοηθούν στον εντοπισμό ομάδων δειγμάτων με παρόμοια πρότυπα γονιδιακής έκφρασης, αλληλεπιδράσεις πρωτεϊνών ή μεταβολικά προφίλ. Ομοίως, τεχνικές ταξινόμησης όπως οι μηχανές φορέα υποστήριξης (SVM) ή τα τυχαία δάση μπορούν να προβλέψουν τις βιολογικές ιδιότητες των δειγμάτων με βάση τα μοριακά τους χαρακτηριστικά.
4. **Ανάλυση Δικτύου:** Η μοριακή βιολογία συχνά περιλαμβάνει τη μελέτη πολύπλοκων δικτύων αλληλεπιδράσεων μεταξύ γονιδίων, πρωτεϊνών και

άλλων βιομορίων. Οι προσεγγίσεις EDA επιτρέπουν την οπτικοποίηση και ανάλυση βιολογικών δικτύων, επιτρέποντας στους ερευνητές να αναγνωρίζουν βασικούς κόμβους (π.χ. γονίδια κεντρικού ή πρωτεΐνες) και να συμπεραίνουν λειτουργικές μονάδες ή μονοπάτια μέσα στο δίκτυο. Οι τεχνικές EDA που βασίζονται σε δίκτυο έχουν συμβάλει καθοριστικά στην κατανόηση της κυτταρικής σηματοδότησης, της γονιδιακής ρύθμισης και των μηχανισμών ασθενειών.

5. **Ενσωμάτωση δεδομένων Multi-Omics:** Με την έλευση των τεχνολογιών υψηλής απόδοσης, οι ερευνητές μπορούν τώρα να δημιουργήσουν σύνολα δεδομένων πολλαπλών ωμικών που περιλαμβάνουν δεδομένα γονιδιωματικής, μεταγραφομικής, πρωτεομικής και μεταβολομικής. Οι μέθοδοι EDA διαδραματίζουν κρίσιμο ρόλο στην ενοποίηση και ανάλυση αυτών των ετερογενών συνόλων δεδομένων, επιτρέποντας στους ερευνητές να αποκτήσουν μια ολοκληρωμένη κατανόηση των βιολογικών συστημάτων σε πολλαπλά μοριακά επίπεδα.
6. **Δημιουργία Υποθέσεων:** Η EDA είναι μια επαναληπτική διαδικασία που προωθεί τη δημιουργία υποθέσεων και τη βελτίωση στην έρευνα της μοριακής βιολογίας. Διερευνώντας διαδραστικά οπτικοποιήσεις δεδομένων και στατιστικές περιλήψεις, οι ερευνητές μπορούν να διατυπώσουν νέες υποθέσεις σχετικά με τις βιολογικές διεργασίες, να δοκιμάσουν υπάρχουσες θεωρίες και να σχεδιάσουν στοχευμένα πειράματα για να επικυρώσουν τα ευρήματά τους.
7. **Αξιολόγηση ποιότητας δεδομένων:** Οι τεχνικές EDA βοηθούν στην αξιολόγηση της ποιότητας και της αξιοπιστίας των δεδομένων μοριακής βιολογίας, εντοπίζοντας ακραίες τιμές, τιμές που λείπουν ή τεχνικά τεχνουργήματα που μπορεί να επηρεάσουν τις μεταγενέστερες αναλύσεις. Διασφαλίζοντας την ακεραιότητα και τη συνέπεια των δεδομένων, ο EDA συμβάλλει στην αναπαραγωγιμότητα και την ευρωστία των ευρημάτων της βιολογικής έρευνας.

Συνοπτικά, η διερευνητική ανάλυση δεδομένων διαδραματίζει κεντρικό ρόλο στη Μοριακή Βιολογία δίνοντας τη δυνατότητα στους ερευνητές να εξερευνήσουν, να οπτικοποιήσουν και να ερμηνεύσουν πολύπλοκα βιολογικά σύνολα δεδομένων. Αξιοποιώντας τις τεχνικές EDA, οι ερευνητές μπορούν να αποκτήσουν νέες γνώσεις για τα βιολογικά συστήματα, να ξεδιαλύνουν μοριακούς μηχανισμούς που

υποκρύπτουν ασθένειες και τελικά να συμβάλλουν στην πρόοδο της υγειονομικής περίθαλψης και της βιοτεχνολογίας (Sahoo et al., 2019).

5.3 Συνοπτική επισκόπηση των προτεινόμενων κατευθύνσεων για μελλοντική έρευνα

Η έρευνα δείχνει ότι, όταν η προηγούμενη γνώση είναι διαθέσιμη και εφαρμόζεται στο συμπέρασμα ρυθμιστικού δικτύου γονιδίων (GRN), τα αποτελέσματα της απόδοσης πρόβλεψης του δικτύου γενικά βελτιώνονται. Επτά πρόσφατα εργαλεία βιοπληροφορικής για συμπεράσματα GRN με βάση χρονοσειρές εξετάστηκαν και συγκρίθηκαν, χρησιμοποιώντας τόσο προσομοιωμένα όσο και πειραματικά σύνολα δεδομένων. Οι τυπικές μετρήσεις απόδοσης για αυτά τα εργαλεία συμπερασμάτων GRN είναι γενικά χαμηλές, υποδηλώνοντας ότι απαιτούνται περαιτέρω προσπάθειες για την ανάπτυξη πιο αξιόπιστων εργαλείων εξαγωγής συμπερασμάτων δικτύου. Η χρήση πολλαπλών εργαλείων μαζί μπορεί να βοηθήσει στον εντοπισμό αληθινών ρυθμιστικών αλληλεπιδράσεων μεταξύ γονιδίων, εύρημα που συνάδει με αυτά που αναφέρονται στη βιβλιογραφία. Η διερευνητική ανάλυση δεδομένων (EDA) στη μοριακή βιολογία βρίσκεται σε μια συνεχή εξέλιξη, καθώς οι τεχνολογίες αλληλούχισης και οι μέθοδοι ανάλυσης συνεχίζουν να βελτιώνονται. Η αυξανόμενη πολυπλοκότητα και ο όγκος των βιολογικών δεδομένων απαιτούν την ανάπτυξη νέων προσεγγίσεων και τεχνικών για την αποτελεσματική εξαγωγή γνώσης. Μερικές από τις πολλά υποσχόμενες κατευθύνσεις για μελλοντική έρευνα περιλαμβάνουν (Barquist, Burge & Gardner, 2016):

1. **Ανάπτυξη νέων αλγορίθμων για την ανάλυση δικτύων γονιδιακής ρύθμισης (GRNs):** Οι υπάρχοντες αλγόριθμοι για την εξαγωγή συμπερασμάτων GRN από δεδομένα χρονοσειρών γονιδιακής έκφρασης έχουν περιορισμούς στην απόδοση. Η εξερεύνηση νέων αλγορίθμων που ενσωματώνουν προηγούμενη γνώση ή χρησιμοποιούν προσεγγίσεις συχνής εξόρυξης μοτίβων μπορεί να βελτιώσει την ακρίβεια και την αξιοπιστία των συμπερασμάτων GRN.
2. **Ενσωμάτωση μηχανικής μάθησης και τεχνητής νοημοσύνης (AI):** Η εφαρμογή τεχνικών μηχανικής μάθησης και τεχνητής νοημοσύνης (AI) στην ανάλυση βιολογικών δεδομένων μπορεί να φέρει επανάσταση στον τομέα. Οι αλγόριθμοι

βαθιάς μάθησης έχουν τη δυνατότητα να εντοπίζουν πολύπλοκα μοτίβα και αλληλεπιδράσεις σε μεγάλα σύνολα δεδομένων, που μπορεί να μην είναι εμφανή με τις παραδοσιακές στατιστικές μεθόδους. Η ενσωμάτωση τεχνικών ΑΙ, όπως η επεξεργασία φυσικής γλώσσας, θα μπορούσε να βοηθήσει στην εξαγωγή γνώσης από τη βιβλιογραφία και την ενσωμάτωσή της στα μοντέλα EDA.

3. **Ανάπτυξη διεπαφών φιλικών προς τον χρήστη:** Ένα σημαντικό εμπόδιο στην ευρεία υιοθέτηση των εργαλείων EDA στη μοριακή βιολογία είναι η πολυπλοκότητά τους και η απαίτηση εξειδικευμένων γνώσεων βιοπληροφορικής. Η ανάπτυξη διαισθητικών και φιλικών προς το χρήστη διεπαφών θα μπορούσε να κάνει αυτά τα εργαλεία πιο προσιτά σε ένα ευρύτερο φάσμα ερευνητών, συμπεριλαμβανομένων των βιολόγων με περιορισμένη υπολογιστική εμπειρία.
4. **Βελτιστοποίηση της επεκτασιμότητας και της απόδοσης:** Καθώς τα βιολογικά δεδομένα συνεχίζουν να αυξάνονται σε μέγεθος και πολυπλοκότητα, η επεκτασιμότητα και η απόδοση των εργαλείων EDA καθίστανται όλο και πιο σημαντικές. Η εξερεύνηση νέων υπολογιστικών πλαισίων, όπως η υπολογιστική cloud και οι μέθοδοι παράλληλης επεξεργασίας, θα μπορούσε να επιτρέψει την ανάλυση τεράστιων συνόλων δεδομένων σε λογικά χρονικά πλαίσια.
5. **Επικύρωση και αναπαραγωγιμότητα:** Η επικύρωση των ευρημάτων της EDA με πειραματικά δεδομένα και η διασφάλιση της αναπαραγωγιμότητας των αναλύσεων είναι ζωτικής σημασίας για την οικοδόμηση εμπιστοσύνης στα αποτελέσματα. Η ανάπτυξη τυποποιημένων αγωγών EDA και η ενσωμάτωση ελέγχων ποιότητας θα μπορούσαν να βελτιώσουν την αξιοπιστία και τη μεταφραστικότητα της έρευνας στη μοριακή βιολογία.

Βιβλιογραφικές αναφορές

- [1]. Abzalov, M., & Abzalov, M. (2016). Exploratory data analysis. *Applied Mining Geology*, 207-219.
- [2]. Alberts, B. (2017). *Molecular biology of the cell*. Garland science.
- [3]. Amezquita, R. A., Lun, A. T., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., ... & Hicks, S. C. (2020). Orchestrating single-cell analysis with Bioconductor. *Nature methods*, 17(2), 137-145.
- [4]. Ban, N., Nissen, P., Hansen, J., Moore, P. B., & Steitz, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, 289(5481), 905-920.
- [5]. Barquist, L., Burge, S. W., & Gardner, P. P. (2016). Studying RNA homology and conservation with infernal: from single sequences to RNA families. *Current Protocols in Bioinformatics*, 54(1), 12-13.
- [6]. Baxevanis, A. D., Bader, G. D., & Wishart, D. S. (Eds.). (2020). *Bioinformatics*. John Wiley & Sons.
- [7]. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... & Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1), 235-242.
- [8]. Byron, K., Cervantes-Cervantes, M., Wang, J. T., Lin, W. C., & Park, Y. (2010). Mining roX1 RNA in *Drosophila* genomes using covariance models. *International journal of computational bioscience*, 1(1), 22-32.
- [9]. Camacho, J., Rodríguez-Gómez, R. A., & Saccenti, E. (2017). Group-wise principal component analysis for exploratory data analysis. *Journal of Computational and Graphical Statistics*, 26(3), 501-512.
- [10]. Chen, G., Ning, B., & Shi, T. (2019). Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in genetics*, 10, 441123.

- [11]. Cooper, G., & Adams, K. (2022). *The cell: a molecular approach*. Oxford University Press.
- [12]. Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1), 107-113.
- [13]. Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12), 919-929.
- [14]. Feng, L., & Lou, J. (2019). DNA methylation analysis. *Nanotoxicity: Methods and Protocols*, 181-227.
- [15]. Ghemawat, S., Gobioff, H., & Leung, S. T. (2003, October). The Google file system. In *Proceedings of the nineteenth ACM symposium on Operating systems principles* (pp. 29-43).
- [16]. Greenfield, A., Madar, A., Ostrer, H., & Bonneau, R. (2010). DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models. *PloS one*, 5(10), e13397.
- [17]. Gunarathne, T., Zhang, B., Wu, T. L., & Qiu, J. (2013). Scalable parallel computing on clouds using Twister4Azure iterative MapReduce. *Future Generation Computer Systems*, 29(4), 1035-1048.
- [18]. Han, J., Pei, J., & Tong, H. (2022). *Data mining: concepts and techniques*. Morgan kaufmann.
- [19]. Hand, D. J. (2007). Principles of data mining. *Drug Safety*, 30, 621-622.
- [20]. Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome medicine*, 9, 1-12.
- [21]. Hashem, I. A. T., Anuar, N. B., Gani, A., Yaqoob, I., Xia, F., & Khan, S. U. (2016). MapReduce: Review and open challenges. *Scientometrics*, 109, 389-422.

- [22]. Jakobsson, J. E., Spjuth, O., & Lagerström, M. C. (2021). scConnect: a method for exploratory analysis of cell–cell communication based on single-cell RNA-sequencing data. *Bioinformatics*, 37(20), 3501-3508.
- [23]. Jebb, A. T., Parrigon, S., & Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, 27(2), 265-276.
- [24]. Kanehisa, M., & Bork, P. (2003). Bioinformatics in the post-sequence era. *Nature genetics*, 33(3), 305-310.
- [25]. Khaladkar, M., Patel, V., Bellofatto, V., Wilusz, J., & Wang, J. T. (2008). Detecting conserved secondary structures in RNA molecules using constrained structural alignment. *Computational biology and chemistry*, 32(4), 264-272.
- [26]. Kherif, F., & Latypova, A. (2020). Principal component analysis. In *Machine learning* (pp. 209-225). Academic Press.
- [27]. Kurdyukov, S., & Bullock, M. (2016). DNA methylation analysis: choosing the right method. *Biology*, 5(1), 3.
- [28]. Laing, C., Wen, D., Wang, J. T., & Schlick, T. (2012). Predicting coaxial helical stacking in RNA junctions. *Nucleic acids research*, 40(2), 487-498.
- [29]. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., ... & Doyle, M. (2001). FitzHugh Wet al: Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860-921.
- [30]. Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, 6(70), 1.
- [31]. Lee, W. P., Hsiao, Y. T., & Hwang, W. C. (2014). Designing a parallel evolutionary algorithm for inferring gene networks on the cloud computing environment. *BMC systems biology*, 8, 1-19.
- [32]. Levitan, I. B., & Kaczmarek, L. K. (2015). *The neuron: cell and molecular biology*. Oxford University Press, USA.

- [33]. Li, J., & Chen, S. J. (2023). RNAJP: enhanced RNA 3D structure predictions with non-canonical interactions and global topology sampling. *Nucleic acids research*, 51(7), 3341-3356.
- [34]. Lössl, P., van de Waterbeemd, M., & Heck, A. J. (2016). The diverse and expanding role of mass spectrometry in structural and molecular biology. *The EMBO journal*, 35(24), 2634-2657.
- [35]. Mangan, N. M., Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Inferring biological networks by sparse identification of nonlinear dynamics. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2(1), 52-63.
- [36]. Martinez, W. L., Martinez, A. R., & Solka, J. (2017). *Exploratory data analysis with MATLAB*. Chapman and Hall/CRC.
- [37]. Milo, T., & Somech, A. (2020, June). Automating exploratory data analysis via machine learning: An overview. In *Proceedings of the 2020 ACM SIGMOD international conference on management of data* (pp. 2617-2622).
- [38]. Milone, M. C., & O'Doherty, U. (2018). Clinical use of lentiviral vectors. *Leukemia*, 32(7), 1529-1541.
- [39]. Monsen, K. A., & Monsen, K. A. (2018). Exploratory data analysis. *Intervention Effectiveness Research: Quality Improvement and Program Evaluation*, 77-85.
- [40]. Moses, L., Einarsson, P. H., Jackson, K., Luebbert, L., Boeshaghi, A. S., Antonsson, S., ... & Pachter, L. (2023). Voyager: exploratory single-cell genomics data analysis with geospatial statistics. *bioRxiv*.
- [41]. Nakato, R., & Sakata, T. (2021). Methods for ChIP-seq analysis: A practical workflow and advanced applications. *Methods*, 187, 44-53.

- [42]. Nakato, R., & Shirahige, K. (2017). Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Briefings in bioinformatics*, 18(2), 279-290.
- [43]. Nalbantoglu, S. (2019). Metabolomics: basic principles and strategies. *Molecular medicine*, 10(1).
- [44]. Nawrocki, E. P., & Eddy, S. R. (2013). Computational identification of functional RNA homologs in metagenomic data. *RNA biology*, 10(7), 1170-1179.
- [45]. Obermayer, B., Holtgrewe, M., Nieminen, M., Messerschmidt, C., & Beule, D. (2020). SCellVis: exploratory single cell data analysis on the desktop and in the cloud. *PeerJ*, 8, e8607.
- [46]. Paliy, O., & Shankar, V. (2016). Application of multivariate statistical techniques in microbial ecology. *Molecular ecology*, 25(5), 1032-1057.
- [47]. Park, S. W., Kang, Y. I., Sypula, J. G., Choi, J., Oh, H., & Park, Y. (2007). An evolutionarily conserved domain of roX2 RNA is sufficient for induction of H4-Lys16 acetylation on the Drosophila X chromosome. *Genetics*, 177(3), 1429-1437.
- [48]. Pavlopoulos, G. A., Baltoumas, F. A., Liu, S., Selvitopi, O., Camargo, A. P., Nayfach, S., ... & Kyrpides, N. C. (2023). Unraveling the functional dark matter through global metagenomics. *Nature*, 622(7983), 594-602.
- [49]. Pearson, R. K. (2018). *Exploratory data analysis using R*. Chapman and Hall/CRC.
- [50]. Rehman, A., Naz, S., & Razzak, I. (2022). Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. *Multimedia Systems*, 28(4), 1339-1371.
- [51]. Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory data analysis using Python. *International Journal of Innovative Technology and Exploring Engineering*, 8(12), 4727-4735.

- [52]. Savaresi, S. M., Boley, D. L., Bittanti, S., & Gazzaniga, G. (2002, April). Cluster selection in divisive clustering algorithms. In *Proceedings of the 2002 SIAM International Conference on Data Mining* (pp. 299-314). Society for Industrial and Applied Mathematics.
- [53]. Storz, G. (2002). An expanding universe of noncoding RNAs. *Science*, 296(5571), 1260-1263.
- [54]. Voet, D., Voet, J. G., & Pratt, C. W. (2016). *Fundamentals of biochemistry: life at the molecular level*. John Wiley & Sons.
- [55]. Vollmers, J., Wiegand, S., Lenk, F., & Kaster, A. K. (2022). How clear is our current view on microbial dark matter?(Re-) assessing public MAG & SAG datasets with MDMcleaner. *Nucleic Acids Research*, 50(13), e76-e76.
- [56]. Waleń, T., Chojnowski, G., Gierski, P., & Bujnicki, J. M. (2014). ClaRNA: a classifier of contacts in RNA 3D structures based on a comparative analysis of various classification schemes. *Nucleic acids research*, 42(19), e151-e151.
- [57]. Wilson, K., Hofmann, A., Walker, J. M., & Clokie, S. (Eds.). (2018). *Wilson and Walker's principles and techniques of biochemistry and molecular biology*. Cambridge university press.
- [58]. Wilson, K., Hofmann, A., Walker, J. M., & Clokie, S. (Eds.). (2018). *Wilson and Walker's principles and techniques of biochemistry and molecular biology*. Cambridge university press.
- [59]. Wimberly, B. T., Brodersen, D. E., Clemons Jr, W. M., Morgan-Warren, R. J., Carter, A. P., Vornrhein, C., ... & Ramakrishnan, V. (2000). Structure of the 30S ribosomal subunit. *Nature*, 407(6802), 327-339.
- [60]. Zhang, Y., Kim, M. S., Reichenberger, E. R., Stear, B., & Taylor, D. M. (2020). Scedar: A scalable Python package for single-cell RNA-seq exploratory data analysis. *PLoS computational biology*, 16(4), e1007794.