Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων
Και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Προηγμένες υπηρεσίες αναζήτησης και διαχείρισης δεδομένων στις Βιοεπιστήμες

Διδακτορική Διατριβή

του

Αθανασίου Βεργούλη

Διπλωματούχου Μηχανικού Υπολογιστών &
Πληροφορικής Παν. Πατρών (2007)

Αθήνα, Αύγουστος 2014

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων Και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Προηγμένες υπηρεσίες αναζήτησης και διαχείρισης δεδομένων στις Βιοεπιστήμες

Διδακτορική Διατριβή

του

Αθανασίου Βεργούλη

Διπλωματούχου Μηχανικού Υπολογιστών &
Πληροφορικής Παν. Πατρών (2007)

Συμβουλευτική Επιτροπή: Τ. Σελλής
Ι. Βασιλείου
Θ. Δαλαμάγκας

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 7η Αυγούστου 2014.

Τ. Σελλής Καθ. ΕΜΠ	Ι. Βασιλείου Καθ. ΕΜΠ	Θ. Δαλαμάγκας Ερευνητής Β΄ Ερ. Κέντρου ΑΘΗΝΑ
Δ. Γουνόπουλος Καθ. ΕΚΠΑ	Δ. Φωτάκης Αναπ. Καθ. ΕΜΠ	Ν. Κοζύρης Καθ. ΕΜΠ
	Α. Δεληγιαννάκης Επ. Καθ. Πολ. Κρήτης	

Αθήνα, Αύγουστος 2014

...

Αθανάσιος Βεργούλης
Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

ΠΡΟΛΟΓΟΣ

Η παρούσα διατριβή εκπληρώνει τις απαιτήσεις για την απόκτηση διπλώματος στο βαθμό του Διδάκτορα στη Σχολή Ηλεκτρολόγων και Μηχανικών Υπολογιστών στο Εθνικό Μετσόβειο Πολυτεχνείο (ΕΜΠ). Η εργασία που παρουσιάζεται περιγράφει μεθόδους για διαχείριση μεγάλου όγκου δεδομένων από τις Βιοεπιστήμες και πραγματοποιήθηκε κατά τη διάρκεια των τελευταίων έξι χρόνων στο Εργαστήριο Βάσεων Γνώσεων και Δεδομένων του ΕΜΠ και στο Ινστιτούτο Πληροφοριακών Συστημάτων (ΙΠΣΥ) του Ερευνητικού Κέντρου Άθηνά.

Είμαι ευγνώμων στον Καθ. Τιμολέοντα Σελλή, τον Δρ. Θοδωρή Δαλαμάγκα και τον Δρ. Δημήτρη Σαχαρίδη για την καθοδήγηση και τις πολύτιμες συμβουλές που μου προσέφεραν. Η εμπειρία τους στο πεδίο της διαχείρισης δεδομένων ήταν μια πολύτιμη πηγή ιδεών. Επιπλέον, θα ήθελα να ευχαριστήσω την Καθ. Άρτεμη Γ. Χατζηγεωργίου και την ομάδα της για το γεγονός ότι με εισήγαγαν στο πεδίο της έρευνας σχετικά με τα μόρια miRNA και για την ευγενική τους απόκριση σε κάθε σχετική απορία που είχα. Επίσης θα ήθελα να αναγνωρίσω τη συνεισφορά του Νίκου Κωστούλα, του Ηλία Κανέλλου και της Ροδοθέας -Μυρσίνης Τσουπίδη που εργάστηκαν σε θέματα που σχετίζονται με την παρούσα διατριβή.

Τέλος, θα ήθελα να ευχαριστήσω θερμά όλους του συναδέλφους μου στο Εργαστήριο Βάσεων Γνώσεων και Δεδομένων και στο Ινστιτούτο Πληροφοριακών Συστημάτων για τη βοήθειά τους, τη συνεργασία που είχαμε και τις υπέροχες στιγμές που μοιραστήκαμε κατά τη διάρκεια των τελευταίων έξι ετών.

Θανάσης Βεργούλης
Αθήνα, Αύγουστος 2014

Στους γονείς μου, τον αδελφό μου και, φυσικά, στη Γεωργία.

ΠΕΡΙΛΗΨΗ

Η ανάγκη για προσεγγίσεις διαχείρισης και επεξεργασίας δεδομένων στις *βιοεπιστήμες* γίνεται εντονότερη λόγω των συνεχών τεχνολογικών εξελίξεων στις μηχανές που παράγουν δεδομένα από βιολογικά δείγματα. Στη σημερινή εποχή, αυτές οι μηχανές παράγουν τεράστιους όγκους δεδομένων τα οποία οφείλουν να επεξεργαστούν. Η πλειοψηφία αυτών των δεδομένων αναπαρίστανται ως *ακολουθίες* και η επεξεργασία τους συνίσταται, κυρίως, στην εφαρμογή αλγορίθμων στοίχισης ακολουθιών πάνω σε αυτές. Οι αλγόριθμοι αιχμής για στοίχιση ακολουθιών αποτυγχάνουν να αποδώσουν καλά για τόσο *μεγάλα δεδομένα*, έτσι, η εισαγωγή νέων προσεγγίσεων είναι απαραίτητη. Η κατάσταση γίνεται ακόμα δυσκολότερη καθώς νέα ευρήματα μερικές φορές δημιουργούν καινούριες ανάγκες επεξεργασίας που δεν μπορούν να ικανοποιηθούν μετασχηματίζοντας τις ήδη υπάρχουσες προσεγγίσεις. Και πάλι νέες μέθοδοι απαιτούνται. Τέλος, νέα ραγδαία εξελισσόμενα πεδία στις βιοεπιστήμες, όπως αυτό των μορίων miRNA παρουσιάζουν έλλειψη από κεντρικές υπηρεσίες πληροφόρησης. Η γνώση σε αυτά τα πεδία είναι διασκορπισμένη σε ένα μεγάλο πλήθος επιστημονικών δημοσιεύσεων επιβραδύνοντας την εργασία των ερευνητών.

ABSTRACT

The need for data management and processing approaches in *life sciences* is becoming more intense due to the continuous technological advances in the machines that produce data from biological samples. In today's era, these machines produce vast amount of data that need to be processed. Most of these data are represented as *sequences* and their processing consists, mainly, of applying sequence alignment algorithms on them. State-of-the-art sequence alignment algorithms fail to perform efficiently for such *big data*, thus, the introduction of novel approaches is apparent. To make the condition worse, novel findings sometimes raise novel processing needs that cannot be fulfilled by adapting already existent approaches. Again, new methods are required. Finally, new rapidly evolving fields in life sciences, like that of miRNA research, lack centralised information resources. The knowledge in such fields is scattered in a multitude of scientific publications slowing down the work of researchers.

Κεφάλαιο 1

Εισαγωγή

Οι *βιοεπιστήμες* περιλαμβάνουν τους τομείς της επιστήμης που αφορούν την μελέτη των ζωντανών οργανισμών και των μηχανισμών της ζωής γενικότερα. Η μελέτη των μηχανισμών της ζωής πραγματοποιείται κυρίως με τη διεξαγωγή βιοχημικών πειραμάτων τα οποία εξετάζουν τις πιθανές αλληλεπιδράσεις που προκύπτουν μεταξύ των χημικών ενώσεων των κυττάρων[1]. Ωστόσο, αυτά τα πειράματα κοστίζουν σε χρόνο και σε χρήματα. Οπότε, η εξέταση της πιθανής αλληλεπίδρασης μεταξύ δυο χημικών ουσιών οι οποίες δε συνδέονται μεταξύ τους αποτελεί σημαντική σπατάλη χρημάτων και χρόνου. Αυτή είναι η αιτία για την οποία η εξεύρεση κατάλληλων ψηφιακών αναπαραστάσεων των κυτταρικών ενώσεων και η εισαγωγή αποδοτικής υπολογιστικής ανάλυσής τους για την αποκάλυψη των πιο πιθανών αλληλεπιδράσεων αποτέλεσαν το αντικείμενο ενός νέου, ευρέους τομέα στις βιοεπιστήμες, που λέγεται *Βιοπληροφορική*.

Η ανάγκη για διαχείριση και επεξεργασία δεδομένων στις βιοεπιστήμες γίνεται πιο επιτακτική λόγω της συνεχούς τεχνολογικής ανάπτυξης του εργαστηριακού εξοπλισμού που αναλύει βιολογικά δείγματα ώστε να παράγει αναπαραστάσεις των κυτταρικών ενώσεων που μπορεί να διαβάσει ο υπολογιστής. Καθώς το πιο βολικό είναι να αναπαριστούμε αυτές τις ενώσεις ως ακολουθίες, ο προαναφερθείς εξοπλισμός λέγεται *μηχανή ακολουθιοποίησης* (sequencing machine). Στη σημερινή εποχή, οι μηχανές ακολουθιοποίησης παράγουν δεδομένα με συνεχώς αυξανόμενο όγκο. Συνεπώς, έχουμε προσεγγίσει το σημείο όπου το κόστος παραγωγής των δεδομένων είναι λιγότερο από εκείνο που απαιτείται για την ανάλυσή τους. Οπότε, η ανάγκη για αποδοτική διαχείριση αυτών των *μεγάλων δεδομένων* (big data) από τις βιοεπιστήμες γίνεται επιτακτική.

Προσεγγίσεις για διαχείριση και επεξεργασία δεδομένων ακολουθιών έχουν προταθεί από τις αρχές της δεκαετίας του ΄60. Ωστόσο, τα τελευταία χρόνια, οι ακολουθίες από τις βιοεπιστήμες δημιούργησαν σοβαρά ζητήματα τα οποία δεν μπορούσαν να επιλυθούν από τις τεχνολογίες αιχμής. Το πρώτο ζήτημα σχετίζεται με την προαναφερθείσα έκρηξη στην παραγωγή δεδομένων προς ανάλυση, που προκλήθηκε από την πρόσφατη πρόοδο στην τεχνολογία των μηχανών ακολουθιοποίησης. Μια σύγχρονη μηχανή ακολουθιοποίησης μπορεί να παράγει αρκετά GB δεδομένων ανά ημέρα και ένας τυπικός κύκλος εκτέλεσης μπορεί να κρατήσει αρκετές ημέρες. Οι υπάρχουσες προσεγγίσεις έχουν αποδειχτεί ανεπαρκείς, ως προς τις επιδόσεις ανάλυσης, για αυτόν το μεγάλο όγκο δεδομένων.

Ένα άλλο ζήτημα προκύπτει από την εμφάνιση μιας νέα γενιάς μηχανών ακολουθιοποίησης. Συνήθως, οι βιολογικές ακολουθίες έχουν μεγάλο μήκος (για παράδειγμα, ένα σύνηθες χρωμόσωμα DNA μπορεί να περιλαμβάνει περισσότερα από 100 εκ. σύμ-

[1]Τα κύτταρα είναι οι δομικές μονάδες όλων των ζωντανών οργανισμών.

βολα), ωστόσο, οι μηχανές ακολουθιοποίησης τα παράγουν σε τμήματα συγκεκριμένου μεγέθους, που λέγονται *αναγνώσματα* (reads). Καθώς οι πρώιμες μηχανές ακολουθιοποίησης παρήγαγαν μικρά αναγνώσματα τα οποία περιείχαν το πολύ 100 σύμβολα, η πρώτη γενιά προσεγγίσεων επεξεργασίας ακολουθιών ήταν βελτιστοποιημένη για ακολουθίες αυτού του μεγέθους. Ωστόσο, τα μικρά μεγέθη των αναγνωσμάτων μπορούν να οδηγήσουν σε προβλήματα κατά την επεξεργασία των ακολουθιών. Για παράδειγμα, ο πιο κοινός τύπος επεξεργασίας είναι η στοίχιση αναγνωσμάτων DNA μέσα σε μια δεδομένη ακολουθία γονιδιώματος. Η στοίχιση μικρών αναγνωσμάτων μπορεί να επηρεαστεί σημαντικά από την παρουσία δομικών μεταβολών ή από πολυμορφισμούς ενός νουκλεοτιδίου (SNP) μέσα σε αυτά. Κάτι τέτοιο έδωσε το κίνητρο για μια νέα γενιά μηχανών ακολουθιοποίησης που μπορούν να παράγουν πολύ μακρύτερα αναγνώσματα. Ωστόσο, για να παρέχει αυτό το προνόμιο, αυτή η νέα γενιά μηχανών ακολουθιοποίησης θυσιάζει την ακρίβεια, που σημαίνει ότι τα αναγνώσματά τους περιέχουν αυξημένο αριθμό λανθασμένα τοποθετημένων συμβόλων. Οι υπάρχουσες τεχνικές επεξεργασίας ακολουθιών έχουν αποδειχτεί ανεπαρκείς στην περίπτωση αναγνωσμάτων μεγάλου μήκους και μικρότερης ακρίβειας. Οπότε, η ανάγκη για προσεγγίσεις οι οποίες λειτουργούν καλά κάτω από το ανωτέρω σενάριο γίνεται προφανής.

Επιπλέον, νέα πορίσματα για τους μηχανισμούς της ζωής δημιουργούν μερικές φορές νέες ανάγκες για επεξεργασία δεδομένων. Σε αυτές τις περιπτώσεις, δεν υπάρχουν τεχνολογίες αιχμής και η προσαρμογή υπαρχουσών μεθόδων μπορεί να οδηγήσει σε χαμηλές επιδόσεις. Οπότε, πρέπει να αναπτυχθούν εξολοκλήρου νέες προσεγγίσεις επεξεργασίας. Για παράδειγμα, στις αρχές της δεκαετίας του 2000, η ανακάλυψη των μορίων miRNA και του ρόλου τους στην απενεργοποίηση γονιδίων ώθησε την ανάπτυξη υπολογιστικών μεθόδων που προσπαθούν να προβλέψουν τα γονίδια-στόχους κάθε miRNA. Ο χημικός δεσμός μεταξύ ενός miRNA και του μεταγράφου ενός γονιδίου που οδηγεί στην απενεροποίηση του γονιδίου ώθησε τη χρήση πολύπλοκων κριτηρίων στοίχισης ακολουθιών από τους πιο ακριβείς από αυτούς τους αλγορίθμους πρόβλεψης (όπως το DIANA microT). Ωστόσο, η προσαρμογή υπαρχουσών προσεγγίσεων στοίχισης ακολουθιών για την αναζήτηση τοποθεσιών γονιδίων που πληρούν τα ανωτέρω κριτήρια οδηγεί σε χαμηλές επιδόσεις. Έτσι, απαιτούνται πιο εξελιγμένες προσεγγίσεις.

Τέλος, ένα σημαντικό ζήτημα στις βιοεπιστήμες είναι ότι αν και υπάρχει πληθώρα αποθετηρίων που συλλέγουν και διανέμουν ενδιαφέρουσες πληροφορίες γενικού ενδιαφέροντος (όπως εκείνες που φιλοξενούνται από την Ensembl και το NCBI), υπάρχει απουσία παρόμοιων πόρων για πιο εξειδικευμένους τομείς. Για παράδειγμα, αυτό ισχύει στον πολύ σημαντικό τομέα της έρευνας των miRNA. Αν και υπάρχει μια κομβική βάση δεδομένων που συλλέγει κάποιες ενδιαφέρουσες πληροφορίες σχετικά με κάθε αναγνωρισμένο μόριο miRNA[2], η σχέση των miRNA με τα γονίδια, ο ρόλος τους στις μεταβολικές οδούς, το προφίλ έκφρασής τους, και πολλές άλλες πληροφορίες σχετικά με αυτά είτε είναι διασκορπισμένες σε σχετικές επιστημονικές δημοσιεύσεις είτε δεν υπάρχουν καθόλου (καθώς θα πρέπει να γίνει επιπλέον ανάλυση για την αποκάλυψή τους). Ωστόσο, τα miRNA συμμετέχουν σε πολύ σημαντικές λειτουργίες της ζωής, καθώς ευθύνονται για την απενεργοποίηση σημαντικών γονιδίων. Η απόκτηση γνώσης για τα miRNA θα μπορούσε να βοηθήσει την κατανόηση και θεραπεία σοβαρών ασθενειών, όπως είναι αρκετοί τύποι καρκίνου, η νόσος Αλτσχάιμερ κλπ. Για το λόγο αυτό, διαδικτυακά αποθετήρια και εργαλεία που να καλύπτουν το κενό των προαναφερθεισών πληροφοριών θα ήταν πολύτιμα.

[2] `http://www.mirbase.org`

1.1 Συνεισφορά

Η παρούσα διατριβή παρουσιάζει ποικίλες μεθόδους διαχείρισης δεδομένων από τις βιοεπιστήμες. Εστιάζουμε σε δυο σημαντικούς τομείς, *την πρόβλεψη στόχων miRNA* και *τη στοίχιση αναγνωσμάτων DNA*. Ο πρώτος τομέας περιλαμβάνει υπολογιστικές μεθόδους που προσπαθούν να αποκαλύψουν αλληλεπιδράσεις μεταξύ μορίων miRNA και γονιδίων, ενώ ο δεύτερος αφορά στη στοίχιση μικρών ακολουθιών DNA σε γονιδιωματικές ακολουθίες αναφοράς, κάτι που αποτελεί χρήσιμο βήμα προεπεξεργασίας σχεδόν για κάθε σημαντική υπολογιστική ανάλυση για τα βιομόρια. Επιπλέον, μελετούμε τρόπους υποβοήθησης της έρευνας σχετικά με τα μόρια miRNA μέσω της διάδοσης χρήσιμης γνώσης σχετικής με αυτά. Οι συνεισφορές μας περιλαμβάνουν τα ακόλουθα.

1. Ασχολούμαστε με το πρόβλημα της παροχής ακριβών προβλέψεων στόχων miRNA σε σχεδόν πραγματικό χρόνο. Επιλέγουμε να εστιάσουμε στη μέθοδο DIANA microT , καθώς συγκαταλέγεται μεταξύ των πιο αξιόπιστων και δημοφιλών μεθόδων. Μελετάμε την εμπλεκόμενη διαδικασία στοίχισης ακολουθιών επειδή είναι υπολογιστικά απαιτητική. Αυτή η διαδικασία περιλαμβάνει ένα νέο τύπο στοίχισης ακολουθιών, έτσι, μοντελοποιούμε μαθηματικά αυτό τον τύπο ερωτήματος εισάγοντας το *πρόβλημα Κατά Προσέγγιση Ταιριάσματος Περιοχών Ακολουθίας* (Approximate Regional Sequence Matching - ARSM). Επιπλέον, καθώς οι αλγόριθμοι αιχμής αποτυγχάνουν να αποδόσουν καλά για τα προαναφερθέντα ερωτήματα, προτείνουμε ένα νέο αλγόριθμο, που λέγεται *PS-ARSM*, ο οποίος εκμεταλλεύεται ειδικά χαρακτηριστικά αυτών των ερωτημάτων ώστε να αποφύγει πλεονάζοντες υπολογισμούς. Οι μέθοδοι που συζητήθηκαν και τα αποτελέσματα που παρατηρήθηκαν αναφέρονται στο [98].

2. Καθώς οι μέθοδοι πρόβλεψης στόχων miRNA περιλαμβάνουν επίσης μερικές άλλες υπολογιστικά απαιτητικές διαδικασίες, εκτός από τη στοίχιση ακολουθιών, μελετάμε το ενδεχόμενο να τις κατανείμουμε στους κόμβους μιας δομής Νέφους. Οπότε, σχεδιάσαμε δυο συστήματα πρόβλεψης στόχων βασισμένων στο Νέφος, που λέγονται *TarCloud* και *MR-microT*. Το πρώτο αναπτύχθηκε με τη χρήση του προγραμματιστικού πλαισίου Azure της Microsoft, ενώ το δεύτερο είναι μια εφαρμογή MapReduce με τη χρήση του προγραμματιστικού πλαισίου Hadoop. Οι μετρήσεις μας δείχνουν ότι και τα δυο συστήματα επιταχύνουν τη διαδικασία πρόβλεψης, με το MR-microT να είναι ανώτερο από πολλές απόψεις. Οι μέθοδοι που συζητήθηκαν και τα τα αποτελέσματα που παρατηρήθηκαν αναφέρονται στο [97] και στο [40].

3. Όσον αφορά στη διάδοση των πληροφοριών σχετικά με τα μόρια miRNA, κάναμε ουσιαστική δουλειά για να παρέχουμε αξιόλογα εργαλεία στους ερευνητές του κλάδου. Πιο συγκεκριμένα, σε συνεργασία με την ομάδα της Καθ. Άρτεμης Χατζηγεωργίου στο ΕΚΕΒΕ 'Αλ. Φλέμινγκ', συλλέξαμε δεδομένα που ήταν διασκορπισμένα σε πολλές επιστημονικές δημοσιεύσεις και βάσεις δεδομένων, τα συνδυάσαμε και τα επεξεργαστήκαμε για να εξάγουμε χρήσιμη γνώση. Τα αποτελέσματα διανέμονται στην επιστημονική κοινότητα μέσω μιας πληθώρας ισχυρών εργαλείων που έχουν διαισθητικές διεπαφές Παγκόσμιου Ιστού. Πιο συγκεκριμένα, αναπτύξαμε (α) το DIANA microT, το οποίο παρέχει στους βιοεπιστήμονες προβλέψεις για τα γονίδια που είναι στόχοι όλων των γνωστών miRNA, (β) το DIANA miRGen, το οποίο ενημερώνει τους χρήστες του για τις γονιδιωματικές

τοποθεσίες όλων των μεταγράφων miRNA και τη συμπεριφοράς έκφρασής τους, (γ) το DIANA TarBase, το οποίο παρέχει πειραματικά επιβεβαιωμένους στόχους miRNA, (δ) το DIANA mirPath, το οποίο ερευνά το ρόλο των miRNA στις γνωστές μεταβολικές οδούς, και (ε) το DIANA mirPub, ένα εργαλείο το οποίο βοηθά τους βιοεπιστήμονες στην έρευνα βιβλιογραφίας σχετικής με τα miRNA. Τα προαναφερθέντα συστήματα και οι μέθοδοι που συζητήθηκαν αναφέρονται στα [59, 60, 76], στο [3], στο [99], και στο [100].

4. Κατά την ανάπτυξη του DIANA TarBase αναγνωρίσαμε τις δυσκολίες που αντιμετωπίζουν οι επιμελητές επιστημονικών βάσεων δεδομένων όταν χρειάζεται να ταυτοποιήσουν αλληλεπιδράσεις μεταξύ miRNA και γονιδίων που καταγράφονται στο κείμενο σχετικών δημοσιεύσεων. Αυτό αποτέλεσε το κίνητρο για να μελετήσουμε τις ευκαιρίες στην αυτόματη αναγνώριση αλληλεπιδράσεων μεταξύ miRNA και γονιδίων. Τα αποτελέσματα της προκαταρκτικής αξιολόγησης, τα οποία παρουσιάστηκαν στο [95], δημιουργούν αισιοδοξία για την παροχή ανάλογων προτάσεων στους επιμελητές του DIANA TarBase.

5. Τέλος, παρουσιάσαμε το *Hitmap*, μια τεχνική ευρετηρίασης που υποστηρίζει αποδοτική στοίχιση για μεγάλα μήκη αναγνωσμάτων DNA και σχετικά μεγάλα κατώφλια σφαλμάτων. Το *Hitmap* καλύπτει το κενό στις τεχνικές στοίχισης αναγνωσμάτων DNA καθώς υπερνικά την τεχνολογία αιχμής στην περίπτωση μεγάλου μήκους αναγνωσμάτων DNA, ενώ η απόδοσή του για τη στοίχιση μικρού μήκους αναγνωσμάτων παραμένει παρόμοια με εκείνη των καλύτερων αλγορίθμων για τη στοίχιση μικρού μήκους αναγνωσμάτων DNA.

1.2 Σύνοψη

Το υπόλοιπο της παρούσας διατριβής δομείται ως εξής.

Το Κεφάλαιο 2 διαμορφώνει το απαραίτητο υπόβαθρο για την εισαγωγή της προτεινόμενης μεθοδολογίας μας. Πιο συγκεκριμένα, αναφέρει τα πιο δημοφιλή προβλήματα στη διαχείριση δεδομένων για βάσεις δεδομένων ακολουθιών και παρουσιάζει τις λύσεις αιχμής. Επιπλέον, εισάγει τον αναγνώστη σε κάποιες απαραίτητες έννοιες της βιολογίας και αναφέρει τις ανάγκες διαχείρισης και επεξεργασίας δεδομένων που προκύπτουν στις βιοεπιστήμες.

Στο Κεφάλαιο 3 αναφέρουμε τις προσπάθειές μας να ωθήσουμε την πρόβλεψη στόχων miRNA. Πρώτον, μελετάμε τη διαδικασία ταιριάσματος ακολουθιών που αποτελεί το πρώτο βήμα του DIANA microT και προτείνουμε μια τεχνική για την επιτάχυνση της διαδικασίας. Καθώς οι μέθοδοι πρόβλεψης στόχων περιλαμβάνουν επίσης και άλλες υπολογιστικά απαιτητικές διαδικασίες, πέρα από το ταίριασμα ακολουθιών, αναπτύσσουμε δυο τεχνικές βασισμένες στο Νέφος οι οποίες κατανέμουν αυτές τις διαδικασίες σε πολλούς υπολογιστικούς κόμβους. Τέλος, στην Ενότητα 3.3 συνοψίζουμε τη δουλειά που έχουμε κάνει σε αυτό τον τομέα και αναφέρουμε τη συνεισφορά μας.

Στο Κεφάλαιο 4, παρουσιάζουμε ένα σύνολο υποδομών που αναπτύχθηκαν για την υποστήριξη της έρευνας σχετικά με τα miRNA. Μέχρι πρόσφατα, οι σημαντικές πληροφορίες σχετικά με τη λειτουργία και τη ρύθμιση κάθε miRNA ήταν διασκορπισμένες σε πολλές βάσεις δεδομένων ή δεν ήταν καν διαθέσιμες. Αυτό αποτέλεσε σημαντικό εμπόδιο για τους ερευνητές στις βιοεπιστήμες, οι οποίοι προσπαθούσαν να κατανοήσουν το ρόλο των μορίων miRNA σε πολλά βιολογικά μονοπάτια, γνώση που θα μπορούσε

να βοηθήσει προς την ανακάλυψη θεραπειών για συγκεκριμένες ασθένειες. Παρουσιάζουμε λεπτομερώς ένα προς ένα όλα τα εργαλεία που αναπτύξαμε για να βοηθήσουμε στην έρευνα των miRNA, μιλώντας για το κίνητρο, τη λειτουργικότητά τους και τη συμβολή τους.

Στο Κεφάλαιο 5, παρουσιάζουμε το Hitmap, μια δομή ευρετηρίου που υποστηρίζει τη στοίχιση για μεγάλα μήκη αναγνωσμάτων και κατώφλια συντακτικής απόστασης.

Το Κεφάλαιο 6 παρουσιάζει τα συνολικά συμπεράσματα της διατριβής, συνοψίζοντας τη συμβολή της. Τέλος, καταγράφουμε πιθανές επεκτάσεις και προτείνουμε ιδέες για μελλοντικές εργασίες.

Κεφάλαιο 2

Προαπαιτούμενα

Στο παρόν κεφάλαιο παρέχουμε το απαραίτητο υπόβαθρο για την κατανόηση των προβλημάτων και των μεθόδων που παρουσιάζονται στα επόμενα κεφάλαια. Πιο συγκεκριμένα, το Κεφάλαιο 2.1 διατυπώνει κάποια προαπαιτούμενα σχετικά με τη διαχείριση ακολουθιών και παρουσιάζει τα πιο σημαντικά προβλήματα ταιριάσματος ακολουθιών καθώς και τις λύσεις αιχμής για αυτά. Το Κεφάλαιο 2.2 εισάγει τον αναγνώστη σε ορισμένες απαραίτητες έννοιες της βιολογίας και αναφέρει τις ανάγκες διαχείρισης και επεξεργασίας δεδομένων που προκύπτουν στις βιοεπιστήμες.

2.1 Διαχείριση βάσεων δεδομένων ακολουθιών

Σε πολλούς τομείς (όπως η υπολογιστική βιολογία, η επεξεργασία σήματος, η ανάκτηση κειμένου κλπ) οι ερευνητές συλλέγουν δεδομένα που αναπαριστώνται ως διατεταγμένα σύνολα συμβόλων, που λέγονται *ακολουθίες*, για να εξάγουν από αυτές χρήσιμες πληροφορίες εκτελώντας διάφορους τύπους ανάλυσης (π.χ. αναζήτηση για κοινά πρότυπα που εμφανίζονται σε αυτές). Τα σύμβολα σε αυτές τις ακολουθίες ανήκουν σε ένα πεπερασμένο σύνολο, που λέγεται *αλφάβητο*. Το μέγεθος του αλφαβήτου και τα περιεχόμενά του εξαρτώνται από την εκάστοτε εφαρμογή. Για παράδειγμα, ένας βιοεπιστήμονας μπορεί να χρησιμοποιεί μια μηχανή ακολουθιοποίησης για να ταυτοποιήσει την ακολουθία νουκλεοτιδίων η οποία ελέγχει την παραγωγή μιας συγκεκριμένης πρωτεΐνης σε ένα κύτταρο. Αυτές οι ακολουθίες νουκλεοτιδίων λέγονται γονίδια και το αλφάβητό τους περιλαμβάνει 4 σύμβολα (καθώς υπάρχουν 4 διακριτοί τύποι νουκλεοτιδίων).

Για το υπόλοιπο της παρούσας διατριβής, χρησιμοποιούμε κεφαλαία λατινικά γράμματα, όπως το S, για να αναπαραστήσουμε ακολουθίες. Χρησιμοποιούμε το γράμμα Σ για να αναπαραστήσουμε το αλφάβητο των ακολουθιών μας, δηλαδή $\forall S \in \Sigma$. Το $|S|$ δηλώνει το μήκος της S, δηλαδή, τον αριθμό συμβόλων που αυτή περιλαμβάνει. Για κάθε $i, j \in [1, |S|]$, το $S_{[i]}$ αντιστοιχεί στο i-οστό σύμβολο στην S, ενώ το $S_{[i,j]}$ στην υπακολουθία της S που ξεκινά στο i-οστό και τελειώνει στο j-οστό σύμβολο. Χρησιμοποιούμε το συμβολισμό $S_{[i,j]} \sqsubset S$ για να δηλώσουμε ότι το $S_{[i,j]}$ είναι υπακολουθία της S.

Μια *βάση δεδομένων ακολουθιών* είναι μια βάση δεδομένων που αποθηκεύει μια πληθώρα ακολουθιών καθώς και κάποια μεταδεδομένα που τις αφορούν. Τα πιο κοινά ερωτήματα στις βάσεις δεδομένων ακολουθιών είναι ερωτήματα επιλογής βάσει της ομοιότητας μιας δεδομένης ακολουθίας-ερωτήματος με οποιαδήποτε από τις ακολουθίες δεδομένων ή με οποιοδήποτε μέρος τους. Αυτή η αποτίμηση τέτοιων ερωτημάτων ανα-

φέρεται συχνά ως *ταίριασμα ακολουθιών ή στοίχιση ακολουθιών* ή, απλώς, *αναζήτηση ακολουθιών*. Στο Κεφάλαιο 2.1.1 περιγράφουμε τα πιο δημοφιλή προβλήματα ταιριάσματος ακολουθιών, ενώ στο Κεφάλαιο 2.1.1.3 αναφέρουμε τους αλγορίθμους και τις δομές ευρετηρίου αιχμής που ασχολούνται με κάθε ένα από αυτά.

2.1.1 Προβλήματα ταιριάσματος ακολουθιών

2.1.1.1 Ακριβές ταίριασμα ακολουθιών

Το πιο απλό ερώτημα επιλογής που μπορεί να τεθεί σε μια βάση δεδομένων ακολουθιών είναι εκείνο που απαιτεί την ανάκτηση όλων των εγγραφών της βάσης δεδομένων οι οποίες περιέχουν την ακριβή εμφάνιση μιας δεδομένης ακολουθίας-ερωτήματος. Το πρόβλημα αυτό είναι γνωστό ως *πρόβλημα ακριβούς ταιριάσματος ακολουθιών (exact sequence matching problem - ESM)* και τα ερωτήματα επιλογής αυτού του τύπου λέγονται *ερωτήματα ESM*. Για παράδειγμα, θεωρούμε μια βάση δεδομένων ακολουθιών {'girogea','roge','trvtt'} και ένα ερώτημα ESM 'roge'. Η απάντηση σε αυτό το ερώτημα είναι το σύνολο των πρώτων δύο εγγραφών της βάσης δεδομένων.

Υπάρχουν πολλοί γνωστοί αποδοτικοί αλγόριθμοι που μπορούν να χρησιμοποιηθούν για την αποτίμηση των ερωτημάτων ESM. Ο πιο δημοφιλής είναι ο αλγόριθμος Boyer-Moore [11]. Ο αλγόριθμος προεπεξεργάζεται την ακολουθία-ερώτημα και χρησιμοποιεί πληροφορίες που συλλέγονται κατά την προεπεξεργασία για να αποφύγει περιοχές των ακολουθιών της βάσης δεδομένων. Ο αλγόριθμος έχει χρόνο εκτέλεσης χειρότερης περίπτωσης $O(n+m)$, όπου m είναι το μήκος της ακολουθίας-ερωτήματος και n το άθροισμα των μηκών των ακολουθιών της βάσης δεδομένων. Να σημειωθεί ότι ο αλγόριθμος Boyer-Moore, σε συνδυασμό με κάποιες άλλες τεχνικές, βρίσκεται πίσω από το δημοφιλές εργαλείο GREP το οποίο παρέχεται από συστήματα βασισμένα σε Unix για την αναζήτηση εντός αρχείων.

Επιπλέον, υπάρχουν πολλές δομές ευρετηρίων που μπορούν να χρησιμοποιηθούν για να δώσουν ώθηση στην αποτίμηση ερωτημάτων ESM. Οι πιο δημοφιλείς είναι το *δέντρο επιθεμάτων* [102] και ο *πίνακας επιθεμάτων* [57]. Και οι δυο βασίζονται στην ίδια ιδέα και ανήκουν στην οικογένεια των *ευρετηρίων επιθεμάτων*. Έστω ότι δίνεται μια ακολουθία S, τότε το δέντρο επιθεμάτων είναι ένα δέντρο που περιλαμβάνει ένα φύλλο για κάθε επίθεμα της S. Κάθε ακμή του δέντρου έχει ετικέτα με μια υπακολουθία της S. Οποιοδήποτε μονοπάτι από τη ρίζα μέχρι ένα φύλλο κωδικοποιεί ένα επίθεμα της S (το επίθεμα του μονοπατιού μπορεί να βρεθεί συρράπτοντας τις ετικέτες των ακμών κατά τη σειρά της περιήγησης). Στην περίπτωση μιας βάσης δεδομένων ακολουθιών, μπορεί να φτιαχτεί ένα δέντρο που περιλαμβάνει τα επιθέματα όλων των ακολουθιών της βάσης δεδομένων. Τέτοια δέντρα είναι γνωστά ως *γενικευμένα δέντρα επιθεμάτων*.

Μετά την κατασκευή ενός δέντρου επιθεμάτων σε μια ακολουθία S, ερωτήματα ESM για ακολουθίες μήκους m μπορούν να απαντηθούν σε μέσο χρόνο $O(m)$. Αυτό γίνεται ξεκινώντας μια διάσχιση από τη ρίζα του δέντρου, διαβάζοντας σύμβολα της ακολουθίας-ερωτήματος (από τα αριστερά προς τα δεξιά). Κατά τη διάρκεια αυτής διαδικασίας, η επόμενη ακμή του δέντρου προς διάσχιση επιλέγεται βάσει των επόμενων συμβόλων που πρόκειται να διαβαστούν από την ακολουθία-ερώτημα. Πιο συγκεκριμένα, μεταξύ των εξερχόμενων ακμών του τρέχοντος κόμβου, εκείνη που έχει πρόθεμα το οποίο ταιριάζει ακριβώς με το επόμενο σύμβολο της ακολουθίας-ερωτήματος επιλέγεται για τη διάσχιση. Εάν υπάρχουν αταίριαστα σύμβολα στο τέλος του επιλεγμένου κόμβου ή εάν δεν υπάρχει τέτοια ακμή, η διάσχιση σταματάει. Η απάντηση στο ερώτημα ESM μπορεί να βρεθεί με επεξεργασία των φύλλων που περιέχονται στο υποδέντρο το

οποίο ορίζεται από τη θέση στο δέντρο όπου σταμάτησε η διάσχιση.

Ένα βασικό μειονέκτημα του δέντρου επιθεμάτων είναι ότι έχει πολύ μεγάλο μέγεθος (είναι μερικές φορές μεγαλύτερο από την αρχική ακολουθία). Ένας πίνακας επιθεμάτων είναι συμπιεσμένη μορφή του δέντρου επιθεμάτων. Κάθε αλγόριθμος που χρησιμοποιεί ένα δέντρο επιθεμάτων μπορεί να προσαρμοστεί ώστε να δουλεύει με πίνακες επιθεμάτων με κάποιο επιπλέον κόστος στο χρόνο εκτέλεσης. Για παράδειγμα, η αξιολόγηση των ερωτημάτων ESM μπορεί να εκτελεστεί σε χρόνο $O(m + logn)$ με ένα πίνακα επιθεμάτων. Ενισχύοντας τον πίνακα επιθεμάτων με μερικές επιπλέον πληροφορίες, μπορεί να επιτευχθεί η ίδια πολυπλοκότητα χρόνου εκτέλεσης με εκείνη των δέντρων επιθεμάτων [1].

Γενικά, το ESM είναι ένα καλά μελετημένο πρόβλημα και υπάρχουν πολλές ικανοποιητικές λύσεις σχεδόν για κάθε παραλλαγή του. Πιο απαιτητικά είναι τα ερωτήματα επιλογής που ζητούν να ανακτηθούν εγγραφές βάσεων δεδομένων οι οποίες περιέχουν κατά προσέγγιση εμφανίσεις μιας δεδομένης ακολουθίας-ερωτήματος. Το ενδιαφέρον μας επικεντρώνεται σε τέτοια ερωτήματα, έτσι, στο Κεφάλαιο 2.1.1.3 δεν πρόκειται να κάνουμε λόγο για λύσεις ESM. Ωστόσο, ο αναγνώστης μπορεί να βρει περισσότερες πληροφορίες για τους αλγορίθμους ESM στο [26].

2.1.1.2 Κατά προσέγγιση ταίριασμα ακολουθιών

Στην καρδιά των περισσότερων προβλημάτων ταιριάσματος ακολουθιών υπάρχει ένα ερώτημα επιλογής σε μια βάση δεδομένων ακολουθιών, όπου το κριτήριο ταιριάσματος βασίζεται στην ομοιότητα των ακολουθιών. Αυτή η ευρεία οικογένεια προβλημάτων είναι γνωστή ως προβλήματα *Κατά προσέγγιση ταιριάσματος ακολουθιών* (Approximate Sequence Matching - ASM) ή *στοίχισης ακολουθιών* (sequence alignment). Μια άριστη επιθεώρηση αυτών των προβλημάτων μπορεί να βρει κανείς στο [27].

Υπάρχουν πολλά μέτρα ομοιότητας ακολουθιών. Το πιο δημοφιλές και ισχυρό από αυτά είναι η *συντακτική απόσταση* [50, 51]. Δεδομένων δύο ακολουθιών S_1 και S_2, η συντακτική τους απόσταση $ed(S_1, S_2)$ ορίζεται ως ο ελάχιστος αριθμός των εισαγωγών, διαγραφών και αντικαταστάσεων συμβόλων που μπορούν να πραγματοποιηθούν για να μετατρέψουν τη μία ακολουθία στην άλλη. Για παράδειγμα, ισχύει ότι $ed(TOP, TAP) = 1$, καθώς μπορούμε να μετατρέψουμε το TOP σε TAP αντικαθιστώντας το σύμβολο O με ένα A. Στην πράξη, η συντακτική απόσταση ανιχνεύει πόσο ανόμοιες είναι δυο ακολουθίες μεταξύ τους, έτσι, μικρή συντακτική απόσταση μεταφράζεται σε μεγάλη ομοιότητα. Να σημειωθεί ότι η συντακτική απόσταση έχει τις ιδιότητες μιας μετρικής (metric).

Ένα άλλο ευρέως χρησιμοποιούμενο μέτρο ομοιότητας είναι η *απόσταση Hamming*. Δεδομένων των ακολουθιών S_1 και S_2, η απόσταση Hamming $hd(S_1, S_2)$ μεταξύ τους ορίζεται ως ο ελάχιστος αριθμός των αντικαταστάσεων συμβόλων που μπορεί να πραγματοποιηθούν για να μετατραπεί η μια ακολουθία στην άλλη. Η απόσταση Hamming είναι επίσης μετρική. Στην ουσία, αποτελεί ειδική περίπτωση της συντακτικής απόστασης όπου επιτρέπονται μόνο αντικαταστάσεις συμβόλων.

Υπάρχουν πολλά άλλα μέτρα ομοιότητας (όπως η ομοιότητα Smith-Waterman [89]), ωστόσο, στο εξής, θεωρούμε ότι η συντακτική απόσταση χρησιμοποιείται σε κάθε περίπτωση, εκτός εάν δηλώνεται κάτι διαφορετικό.

Υπάρχουν δυο κύριες κατηγορίες προβλημάτων κατά προσέγγιση ταιριάσματος ακολουθιών. Η μία αφορά στην ανάκτηση των k πιο όμοιων ακολουθιών βάσεων δεδομένων σε ένα δεδομένο ερώτημα (*κορυφαίες-k ακολουθίες*). Η άλλη έγκειται στην ανάκτηση όλων των ακολουθιών βάσεων δεδομένων που έχουν συντακτική απόσταση

από το ερώτημα το πολύ k, όπου το k είναι ένα κατώφλι που δίνεται από το χρήστη. Και στις δυο κατηγορίες, οι ανακτημένες ακολουθίες, που λέγονται *ταιριάσματα ή στοιχίσεις* του ερωτήματος, μπορούν να είναι είτε ολόκληρες εγγραφές της βάσης δεδομένων ακολουθιών (*ολική στοίχιση* [71, 84]) ή υπακολουθίες αυτών (*τοπική στοίχιση*), ανάλογα με το στόχο του προβλήματος. Καθώς η μελέτη των κορυφαίων-k ερωτημάτων δεν εντάσσεται στο πεδίο της παρούσας διατριβής, ο ενδιαφερόμενος αναγνώστης θα μπορούσε να αναφερθεί στο [27].

Σχετικά με τα ερωτήματα τοπικής στοίχισης, υπάρχει η επιλογή αναζήτησης στοιχίσεων τμημάτων του ερωτήματος [90]. Σε αυτή την περίπτωση, για κάθε εγγραφή της βάσης δεδομένων, στόχος είναι η εύρεση του ζεύγους ερωτήμα-υπακολουθία εγγραφής το οποίο επιτυγχάνει τη μεγαλύτερη ομοιότητα και στη συνέχεια η εξέταση του ενδεχομένου η ομοιότητα αυτή να ικανοποιεί το δεδομένο κατώφλι k. Αναφερόμαστε σε αυτά τα ερωτήματα ως ερωτήματα *διπλά τοπικής στοίχισης*.

2.1.1.3 Αλγόριθμοι και δομές ευρετηρίων για ASM

Οι προσεγγίσεις *δυναμικού προγραμματισμού*, όπως εκείνη που παρουσιάζεται στο [84], συνήθως χρησιμοποιούνται για την αξιολόγηση ερωτημάτων ολικής στοίχισης. Η προσέγγιση του βασικού δυναμικού προγραμματισμού απαιτεί $O(s_1 \cdot s_2)$ χρόνο για τον υπολογισμό της συντακτικής απόστασης δυο ακολουθιών S_1 και S_2, όπου s_1 και s_2, αντίστοιχα, είναι τα μήκη τους.

Σχετικά με τα ερωτήματα τοπικής στοίχισης, υπάρχουν αρκετές εργασίες που μελετούν σχετικούς αλγορίθμους [36, 67]. Η βασική λύση για την εύρεση των στοιχίσεων ενός δεδομένου ερωτήματος Q μέσα σε μια εγγραφή της βάσης δεδομένων D, που λέγεται 'ακολουθία δεδομένων', είναι επίσης μια προσέγγιση δυναμικού προγραμματισμού [85]. Ο χρόνος εκτέλεσής της είναι $O(q \cdot d)$, όπου q και d είναι τα μήκη των Q και D, αντίστοιχα. Έχουν προταθεί αρκετές βελτιώσεις για αυτή τη βασική προσέγγιση δυναμικού προγραμματισμού (για λεπτομέρειες δείτε το [67]). Μια τέτοια βελτιστοποίηση, που επιτυγχάνει χρόνο εκτέλεσης $O(k \cdot d)$, όπου k είναι το κατώφλι της συντακτικής απόστασης, είναι η *ευριστική αποκοπής* [96]. Η βασική της ιδέα έγκειται στην αποφυγή του υπολογισμού τμημάτων του πίνακα δυναμικού προγραμματισμού βάσει μιας ευριστικής που αναγνωρίζει κελιά του πίνακα για τα οποία είναι σίγουρο ότι η αξία τους θα είναι μεγαλύτερη από k.

Ωστόσο, οι πιο αποδοτικοί αλγόριθμοι για ερωτήματα τοπικής στοίχισης ανήκουν στην ομάδα των *αλγορίθμων φιλτραρίσματος*. Πιο συγκεκριμένα, οι αλγόριθμοι των Chang [15] και Fredriksson [22] έχουν τη βέλτιστη πολυπλοκότητα μέσης περίπτωσης. Αυτοί οι αλγόριθμοι πρώτα συγκρίνουν το ερώτημα με κάθε πιθανή ακολουθία ενός μήκους ℓ το οποίο έχει οριστεί εκ των προτέρων, που λέγεται ℓ-gram, και στη συνέχεια χρησιμοποιούν αυτές τις πληροφορίες για να φιλτράρουν τις περιοχές της ακολουθίας της βάσης δεδομένων οι οποίες δεν μπορούν να περιέχουν καμία στοίχιση. Οι υπόλοιπες περιοχές υφίστανται επεξεργασία με τη χρήση ενός συμβατικού αλγορίθμου τοπικής στοίχισης.

Προκύπτουν ενδιαφέροντα ζητήματα όταν πρέπει να αξιολογηθεί μεγάλος αριθμός ερωτημάτων στην ίδια βάση δεδομένων. Εάν ο αριθμός ερωτημάτων είναι μεγάλος, τότε είναι σίγουρο ότι υπάρχουν επικαλύψεις μεταξύ τους. Η ανεξάρτητη αξιολόγηση ερωτημάτων που έχουν επικαλύψεις οδηγεί σε φτωχή απόδοση λόγω επαναλαμβανόμενων υπολογισμών. Αυτό ήταν το κίνητρο για την εισαγωγή των αλγορίθμων *πολλαπλής τοπικής στοίχισης* (ή, γενικά, αλγορίθμων *πολλαπλού κατά προσέγγιση ταιριάσματος ακολουθιών*). Υπάρχουν αρκετές μέθοδοι (π.χ., οι [66, 9, 34, 22]), ωστόσο, ο αλγό-

ριθμος του Fredriksson [22] έχει αποδειχτεί ο βέλτιστος [70]. Ο αλγόριθμος αυτός διατρέχει κάθε ακολουθία δεδομένων χρησιμοποιώντας ένα κυλιόμενο παράθυρο. Για κάθε θέση παραθύρου, διαβάζει ανάποδα (δηλ. από τα δεξιά στα αριστερά) διαδοχικά, μη επικαλυπτόμενα ℓ-gram. Όταν η συνολική απόκλιση των αναγνωσμένων ℓ-gram υπερβαίνει ένα κατώφλι, ο αλγόριθμος προσπερνά το τρέχον παράθυρο και το σέρνει προς τα δεξιά. Διαφορετικά, πρέπει να εξετάσει το παράθυρο για ενδεχόμενα αποτελέσματα. Να σημειωθεί ότι αυτές οι μέθοδοι είναι σχεδιασμένες για ερωτήματα παρόμοιου μήκους, διαφορετικά δεν είναι πιθανό να βρεθεί μια σωστή τιμή ℓ που να επιτυγχάνει καλή απόδοση για όλα.

Όσον αφορά τα ερωτήματα διπλά τοπικής στοίχισης, ο δυναμικός προγραμματισμός των Smith-Waterman [90] εγγυάται να λύσει αυτό το πρόβλημα σε χρόνο $O(q \cdot d)$, για μια σημαντική τάξη μέτρων ομοιότητας. Όταν αναζητάμε υπακολουθίες υψηλής ομοιότητας, το υψηλό υπολογιστικό κόστος των αλγορίθμων δυναμικού προγραμματισμού καθιστά τις κατά προσέγγιση λύσεις (π.χ.,[56, 77, 4, 5]) πιο ελκυστικές. Αυτές οι μέθοδοι χρησιμοποιούν ευριστικές για την αποφυγή αναζήτησης τμημάτων των ακολουθιών που δεν είναι πιθανό να περιέχουν διπλά τοπικές στοιχίσεις. Σαν παράπλευρο αποτέλεσμα, μπορεί να χάσουν αποτελέσματα. Η πιο γνωστή προσεγγιστική λύση είναι το BLAST [4] και οι παραλλαγές του [5, 107, 45, 42].

Διάφορες δομές ευρετηρίων μπορούν να εφαρμοστούν για την επιτάχυνση των προβλημάτων ASM. Πρώτον, υπάρχουν αλγόριθμοι που χρησιμοποιούν ευρετήρια ESM, όπως τα δέντρα και οι πίνακες επιθεμάτων, για την αξιολόγηση των ερωτημάτων ASM [69]. Αυτό είναι πιθανό καθώς, συνήθως, ένα ερώτημα ASM μπορεί να μεταφραστεί σε ένα σύνολο ερωτημάτων ESM. Αυτή η κεντρική ιδέα βρίσκεται επίσης στον πυρήνα της πιο δημοφιλούς οικογένειας δομών ευρετηρίων ASM, των *ανεστραμμένων ευρετηρίων βασισμένων σε gram*. Αυτές οι δομές είναι απλώς λεξικά βασισμένα σε πίνακες κατακερματισμού τα οποία περιέχουν μια εγγραφή για κάθε διακριτή υπακολουθία των δεδομένων που έχουν ένα προκαθορισμένο μήκος. Κάθε εγγραφή έχει σαν κλειδί την ακολουθία και σαν τιμή τη λίστα όλων των θέσεων όπου εμφανίζεται η ακολουθία μέσα στα δεδομένα. Η αξιολόγηση των ερωτημάτων τοπικής στοίχισης, κάτω από ένα κατώφλι συντακτικής απόστασης k, μπορεί να εκτελεστεί με τη χρήση ενός ανεστραμμένου ευρετηρίου βασισμένου σε gram βάσει ενός θεωρήματος γνωστού ως η *αρχή του περιστερώνα*.

Η αυθεντική μορφή της αρχής του περιστερώνα είχε διατυπωθεί από τον P.G.L. Dirichlet στο πλαίσιο των διακριτών μαθηματικών. Η προσαρμογή της που χρησιμοποιείται από ανεστραμμένα ευρετήρια βασισμένα σε gram δηλώνει ότι εάν χωρίσουμε ένα ερώτημα Q σε $k + x$ διαδοχικά, μη επικαλυπτόμενα θραύσματα[1], τότε κάθε στοίχιση του Q στην S θα περιέχει ακριβείς εμφανίσεις τουλάχιστον x θραυσμάτων. Εάν η συνθήκη αυτής της αρχής δεν ισχύει για μια δεδομένη υπακολουθία δεδομένων, τότε δεν είναι πιθανό αυτή η υπακολουθία να περιέχει μια στοίχιση του Q.

Βάσει των ανωτέρω, θεωρούμε ένα ερώτημα Q, ένα κατώφλι συντακτικής απόστασης k και ένα ανεστραμμένο ευρετήριο βασισμένο σε gram χτισμένο πάνω στην ακολουθία δεδομένων D με τη χρήση ℓ-gram, όπου $\ell = \lfloor \frac{q}{k+x} \rfloor$. Τότε, το Q μπορεί να χωριστεί σε $k + x$ διαδοχικά, μη επικαλυπτόμενα θραύσματα και οι ανεστραμμένες λίστες όλων αυτών των θραυσμάτων μπορούν να ανακτηθούν από το ευρετήριο. Οι μόνες θέσεις της D που μπορεί να περιέχουν στοιχίσεις του Q είναι εκείνες που εμφανίζονται τουλάχιστον x φορές στις ανακτημένες λίστες. Για να τις βρούμε, πρέπει να

[1]Να σημειωθεί ότι μπορεί να υπάρχουν σύμβολα στο τέλος του Q που δεν ανήκουν σε κανένα θραύσμα.

εφαρμοστεί μια ειδική συνένωση των λιστών. Η προαναφερθείσα μέθοδος έχει διατυπωθεί στο [68], ενώ κάποιες βελτιστοποιήσεις σχετικά με το τελικό βήμα συνένωσης των λιστών έχουν προταθεί στο [52].

Καθώς το πρόβλημα τοπικής στοίχισης έγινε ζωτικής σημασίας στη Βιοπληροφορική λόγω της σχέσης του με τη διαδικασία στοίχισης αναγνωσμάτων DNA (βλ. Κεφάλαιο 2.2.3) πολλές προσεγγίσεις βασισμένες σε πίνακες κατακερματισμού, περιλαμβανομένων των Maq [29], SOAP [80], GSNAP [103] προτάθηκαν από ερευνητές του κλάδου. Αυτές οι προσεγγίσεις ελαχιστοποιούν το πρόβλημα της τοπικής στοίχισης σε ESM των θραυσμάτων των αναγνωσμάτων, ωστόσο, οι περισσότερες χρησιμοποιούν αυθαίρετο κατακερματισμό ερωτημάτων, όχι κατακερματισμό βάσει της αρχής του περιστερώνα. Μέθοδοι που υπερνικούν τις προηγούμενες, χρησιμοποιώντας συμπιεσμένους πίνακες επιθεμάτων συναντώνται επίσης στη Βιοπληροφορική (π.χ., οι BWA [53], Bowtie [10] και SOAP2 [81]). Αυτές οι μέθοδοι χρησιμοποιούν την *τεχνική μετασχηματισμού Burrows-Wheeler (BWT)* [13] για να συμπιέσουν τους πίνακες και να επιτύχουν πολύ μικρό αποτύπωμα μνήμης.

Ωστόσο, οι μέθοδοι συμπιεσμένων πινάκων επιθεμάτων βελτιστοποιούνται για πολύ μικρά μήκη ερωτημάτων (π.χ. για ερωτήματα με $30 - 50$ σύμβολα). Επιπλέον, λειτουργούν καλά όταν χρησιμοποιούνται για ταιριάσματα με μικρό αριθμό σφαλμάτων [54]. Ωστόσο, οι τεχνολογικές πρόοδοι στον εξοπλισμό που παράγει αναγνώσματα DNA εντείνουν την ανάγκη για προσεγγίσεις στοίχισης που υποστηρίζουν την τοπική στοίχιση για πιο μεγάλα αναγνώσματα και που είναι πιο ανεκτικές σε σφάλματα στοίχισης. Επιπλέον, καθώς το κόστος υπολογιστικής μνήμης εξακολουθεί να μειώνεται, δεν υπάρχει ανάγκη διατήρησης μικρού αποτυπώματος μνήμης. Τα ανωτέρω αποτέλεσαν το κίνητρο για μια νέα προσέγγιση ανεστραμμένου ευρετηρίου βασισμένου σε gram, που λέγεταιWHAM [54].

Η κεντρική ιδέα του WHAM είναι η ίδια με εκείνη του βασικού ανεστραμμένου ευρετηρίου βασισμένου σε gram [68]. Ωστόσο, κάθε εγγραφή του WHAM, αντί να καταγράφει τις εμφανίσεις κάθε πιθανού θραύσματος του ερωτήματος στην ακολουθία δεδομένων, καταγράφει τις συνδυασμένες εμφανίσεις των πιθανών θραυσμάτων. Ο αριθμός των θραυσμάτων σε κάθε συνδυασμό βασίζεται στην αρχή του περιστερώνα. Πιο συγκεκριμένα, εάν ένα ερώτημα Q πρέπει να στοιχιθεί με συντακτική απόσταση ίση το πολύ με k και επιλέξουμε να χωρίσουμε το ερώτημα σε $k + x$ θραύσματα, τότε κάθε εγγραφή του WHAM καταγράφει κάθε συνδυασμένη εμφάνιση x θραυσμάτων στη D. Με αυτό τον τρόπο, ένα ερώτημα τοπικής στοίχισης μπορεί να απαντηθεί απλώς παίρνοντας την ένωση των ανεστραμμένων λιστών όλων των συνδυασμών των θραυσμάτων των ερωτημάτων. Το μόνο ζήτημα που θέλει προσοχή είναι να ληφθούν υπόψη πιθανές εισαγωγές και διαγραφές σε κάθε συνδυασμό θραυσμάτων ολισθαίνοντας το θραύσμα με κάθε πιθανό τρόπο. Ας σημειωθεί ότι το WHAM γίνεται πανομοιότυπο με το βασικό ανεστραμμένο ευρετήριο βασισμένο σε gram στην περίπτωση όπου $x = 1$.

Τέλος, πέρα από τις λύσεις τις βασισμένες σε gram και σε πίνακες επιθεμάτων, υπάρχει επίσης το RBSA [75], μια προσέγγιση βασισμένη σε ακολουθία αναφοράς. Το RBSA βασίζεται στην ιδιότητα μετρικής της συντακτικής απόστασης. Πιο συγκεκριμένα, για κάθε υπακολουθία δεδομένων υπολογίζεται η συντακτική της απόσταση σε ένα σύνολο ακολουθιών αναφοράς. Κατά την αξιολόγηση των ερωτημάτων, υπολογίζεται επίσης η συντακτική απόσταση των ερωτημάτων στις ίδιες ακολουθίες αναφοράς. Έστω ότι έχουμε μια τυχαία υπακολουθία A των δεδομένων, Q το ερώτημα και k το κατώφλι της συντακτικής απόστασης. Εάν υπάρχει μια ακολουθία αναφοράς της A, που λέγεται R, για την οποία $|ed(A, R) - ed(Q, R) > k$, τότε, το $ed(A, Q)$ δεν μπορεί

να είναι μικρότερο ή ίσο με k. Η εξέταση της ανωτέρω συνθήκης για όλες τις ακολουθίες αναφοράς κάθε υπακολουθίας της D είναι λιγότερο υπολογιστικά απαιτητική από τον υπολογισμό της συντακτικής απόστασης του Q στην υπακολουθία. Οπότε, το RBSA παρέχει ένα τρόπο φιλτραρίσματος περιοχών των δεδομένων.

2.2 Διαχείριση και επεξεργασία δεδομένων στις Βιοεπιστήμες

Οι βιοεπιστήμες περιλαμβάνουν τους τομείς της επιστήμης που αφορούν στην επιστημονική μελέτη των ζωντανών οργανισμών και των μηχανισμών της ζωής, γενικότερα. Δυο τύποι βιολογικών μορίων παίζουν τον πιο σημαντικό ρόλο σε αυτούς τους μηχανισμούς: τα νουκλεϊκά οξέα και οι πρωτεΐνες. Τα νουκλεϊκά οξέα είναι αλυσίδες μονομερών που λέγονται νουκλεοτίδια, ενώ οι πρωτεΐνες είναι αλυσίδες αμινοξέων. Πολλοί μηχανισμοί της ζωής περιλαμβάνουν αλληλεπιδράσεις μεταξύ των προαναφερθέντων βιολογικών μορίων βάσει της αλληλουχίας των συστατικών στις αλυσίδες τους. Για το λόγο αυτό η αναπαράσταση των νουκλεϊκών οξέων και των πρωτεϊνών σαν ακολουθίες και η χρήση αλγορίθμων ταιριάσματος ακολουθιών για την ανάλυσή τους πυροδότησε την ανάπτυξη νέων κλάδων, όπως η *βιοπληροφορική* και η *υπολογιστική βιολογία*. Η έρευνα σε αυτούς τους κλάδους αφορά την εύρεση αποδοτικών τρόπων διαχείρισης και επεξεργασίας των βιολογικών ακολουθιών.

Στο Κεφάλαιο 2.2.1 θα εισάγουμε τον αναγνώστη σε κάποιες απαραίτητες έννοιες της βιολογίας. Στα Κεφάλαια 2.2.2 και 2.2.3 θα αναλύσουμε την πρόβλεψη στόχων miRNA και τη στοίχιση αναγνωσμάτων DNA, αντίστοιχα, τα οποία αποτελούν τα δυο προβλήματα από τον κλάδο της βιοπληροφορικής που αποτέλεσαν κίνητρο για την εργασία μας.

2.2.1 Έννοιες της βιολογίας

Τα *Νουκλεϊκά οξέα* είναι μεγάλα βιολογικά μόρια (ή μακρομόρια), ουσιώδους σημασίας για όλες τις γνωστές μορφές ζωής. Φτιάχνονται από μονομερή που λέγονται νουκλεοτίδια. Κάθε νουκλεοτίδιο έχει τρία συστατικά: ένα σάκχαρο με 5 άτομα άνθρακα, μια φωσφορική ομάδα, και μια αζωτούχα βάση. Εάν το σάκχαρο είναι δεσοξυριβόζη, το πολυμερές είναι DNA. Εάν το σάκχαρο είναι ριβόζη, το πολυμερές είναι RNA.

Τα δυο βασικά νουκλεϊκά οξέα είναι το *DNA* (δεσοξυριβονουκλεϊκό οξύ) και το *RNA* (ριβονουκλεϊκό οξύ). Αυτά τα νουκλεϊκά οξέα, μαζί με τις πρωτεΐνες, είναι τα πιο σημαντικά βιολογικά μακρομόρια και μπορούν να συναντηθούν σε αφθονία σε όλους τους ζωντανούς οργανισμούς. Η λειτουργία τους είναι να κωδικοποιούν, να εκπέμπουν, και να εκφράζουν τη γενετική πληροφορία. Παρακάτω, αναλύουμε τις λειτουργίες των πιο σημαντικών τύπων αυτών των βιομορίων.

2.2.1.1 DNA, γονιδίωμα, και γονίδια

Το DNA είναι ένα μόριο που κωδικοποιεί τις γενετικές πληροφορίες που χρησιμοποιούνται στην ανάπτυξη και λειτουργία όλων των γνωστών ζωντανών οργανισμών και πολλών ιών. Τα περισσότερα μόρια DNA αποτελούνται από δύο κλώνους βιοπολυμερών που περιστρέφονται το ένα γύρω από το άλλο ώστε να σχηματίσουν μια διπλή έλικα. Οι δυο κλώνοι DNA είναι γνωστοί ως πολυνουκλεοτίδια καθώς συντίθενται από πιο απλές μονάδες που λέγονται νουκλεοτίδια.

Κάθε νουκλεοτίδιο συντίθεται από μια νουκλεϊκή βάση που περιέχει άζωτο καθώς και από ένα μονοσακχαρίτη που λέγεται δεσοξυριβόζη και μια φωσφορική ομάδα. Υπάρχουν τέσσερις τύποι νουκλεϊκών βάσεων: η γουανίνη (G), η αδενίνη (A), η θυμίνη (T) και η κυτοσίνη (C). Τα νουκλεοτίδια συνδέονται μεταξύ τους σε μια αλυσίδα από ομοιπολικούς δεσμούς ανάμεσα στο σάκχαρο ενός νουκλεοτιδίου και το φωσφορικό άλας του επόμενου, οδηγώντας σε έναν εναλλασσόμενο αυχένα σακχάρου-φωσφορικού άλατος. Σύμφωνα με τους κανόνες ταιριάσματος των αζωτούχων βάσεων (A με T και C με G), οι δεσμοί υδρογόνου δεσμεύουν τις αζωτούχες βάσεις των δυο ξεχωριστών κλώνων πολυνουκλεοτιδίων για να φτιάξουν δίκλωνο DNA.

Μέσα στα κύτταρα, το DNA οργανώνεται σε μεγάλες δομές που λέγονται *χρωμοσώματα*. Κατά την κυτταρική διαίρεση αυτά τα χρωμοσώματα διπλασιάζονται στη διαδικασία αντιγραφής του DNA, παρέχοντας σε κάθε κύτταρο το δικό του πλήρες σύνολο χρωμοσωμάτων. Οι ευκαρυωτικοί οργανισμοί (όπως τα ζώα και τα φυτά) αποθηκεύουν το περισσότερο DNA τους μέσα στον πυρήνα του κυττάρου και ένα μέρος του DNA τους σε οργανίδια, όπως τα μιτοχόνδρια ή οι χλωροπλάστες. Αντίθετα, οι προκαρυωτικοί οργανισμοί (βακτήρια και αρχαία) αποθηκεύουν το DNA τους μόνο στο κυτταρόπλασμα. Στο εξής, θα επικεντρωθούμε στους ευκαρυωτικούς οργανισμούς.

Η ομάδα όλων των μορίων DNA που περιέχονται σε κάθε κύτταρο ενός ευκαρυωτικού οργανισμού λέγεται *γονιδίωμα*. Το γονιδίωμα είναι ακριβώς ίδιο για όλα τα κύτταρα ενός ατόμου[2]. Βρίσκεται μέσα σε μια προστατευμένη περιοχή του κυττάρου, που λέγεται *πυρήνας*. Όπως προαναφέρθηκε, το γονιδίωμα οργανώνεται σε χρωμοσώματα. Μέσα στις αλυσίδες DNA των χρωμοσωμάτων υπάρχουν κάποιες περιοχές, που λέγονται *γονίδια*, κάθε ένα από τα οποία κωδικοποιεί την ακολουθία των αμινοξέων που σχηματίζει μια *πρωτεΐνη*. Οι πρωτεΐνες είναι τα πιο σημαντικά μακρομόρια του κυττάρου, καθώς εκτελούν μια μεγάλη σειρά σημαντικών λειτουργιών, συμπεριλαμβανομένων των μεταβολικών αντιδράσεων, της αντιγραφής του DNA, της ανταπόκρισης σε ερεθίσματα, και της μεταφοράς μορίων από μια περιοχή σε μια άλλη. Για το λόγο αυτό, τα γονίδια κέντρισαν σε μεγάλο βαθμό την προσοχή των ερευνητών.

Το υπόλοιπο γονιδίωμα είναι γνωστό ως *αφανές γονιδίωμα* και για δεκαετίες οι επιστήμονες πίστευαν ότι είναι απλώς ένα παραπροϊόν της εξέλιξης. Ωστόσο, κατά τις τελευταίες δεκαετίες οι βιοεπιστήμονες βρήκαν ότι αυτές οι περιοχές είναι επίσης λειτουργικές (βλ. επίσης Κεφάλαιο 2.2.1.3).

2.2.1.2 Σύνθεση πρωτεϊνών και αγγελιαφόρο RNA

Οι πρωτεΐνες συντίθενται από αμινοξέα με τη χρήση πληροφοριών που είναι κωδικοποιημένες στα γονίδια. Κάθε πρωτεΐνη έχει τη δική της μοναδική ακολουθία αμινοξέων η οποία καθορίζεται από τη ακολουθία νουκλεοτιδίων του γονιδίου που κωδικοποιεί αυτή την πρωτεΐνη. Ο γενετικός κώδικας είναι ένα σύνολο ομάδων τριών νουκλεοτιδίων που λέγονται *κωδικόνια* και κάθε κωδικόνιο κωδικοποιεί ένα αμινοξύ, για παράδειγμα το AUG είναι το κωδικόνιο για τη μεθειονίνη. Επειδή το DNA περιέχει τέσσερα διαφορετικά νουκλεοτίδια, ο συνολικός αριθμός των πιθανών κωδικονίων είναι 64. Οπότε, υπάρχει κάποιος πλεονασμός στο γενετικό κώδικα, με κάποια αμινοξέα να προσδιορίζονται από περισσότερα του ενός κωδικόνια.

Τα γονίδια αντιγράφονται πρώτα σε *προ-αγγελιαφόρο RNA* (pre-mRNA) από πρωτεΐνες, όπως η RNA πολυμεράση. Οι περισσότεροι οργανισμοί στη συνέχεια επεξεργάζονται το προ-mRNA (γνωστό επίσης ως ένα *βασικό μετάγραφο*) για να σχηματίσουν

[2]Υπάρχουν κάποιες εξαιρέσεις, όπως οι γαμέτες.

το *ώριμο mRNA ή ώριμο μετάγραφο*, το οποίο στην πορεία χρησιμοποιείται ως πρότυπο για τη σύνθεση πρωτεϊνών. Οι ευκαρυωτικοί οργανισμοί φτιάχνουν mRNA στον πυρήνα των κυττάρων (όπου βρίσκεται το γονιδίωμα) και μετά το μεταφέρουν από την πυρηνική μεμβράνη στο κυτταρόπλασμα, όπου στη συνέχεια πραγματοποιείται η σύνθεση των πρωτεϊνών.

Η διαδικασία σύνθεσης μια πρωτεΐνης από ένα πρότυπο mRNA είναι γνωστή ως *μετάφραση*. Το mRNA φορτώνεται σε μια μεγάλη και πολύπλοκη μοριακή μηχανή, γνωστή ως ριβόσωμα. Μετά, διαβάζεται τρία νουκλεοτίδια τη φορά ταιριάζοντας κάθε κωδικόνιο με το αντικωδικόνιο που ταιριάζει ως προς τις αζωτούχες βάσεις του και το οποίο βρίσκεται πάνω σε ένα μεταφορικό μόριο RNA, το οποίο μεταφέρει το αμινοξύ που αντιστοιχεί στο κωδικόνιο που αναγνωρίζει. Κατά τη διαδικασία, τα αμινοξέα τοποθετούνται μαζί για να σχηματίσουν μεγάλες αλυσίδες, τις πρωτεΐνες (ή μέρη των πρωτεϊνών).

Καθώς τα γονίδια περιέχουν όλες τις πληροφορίες που χρησιμοποιούνται από το κύτταρο για τη σύνθεση μιας πρωτεΐνης, όταν παράγεται μια πρωτεΐνη που κωδικοποιείται από ένα γονίδιο, λέμε ότι αυτό το γονίδιο *εκφράζεται*.

2.2.1.3 Μόρια microRNA

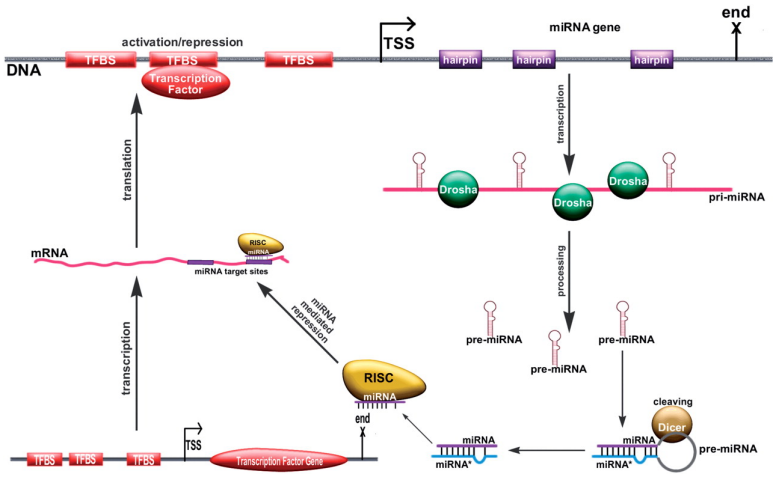
Τα microRNA (miRNA) είναι μονόκλωνα μόρια RNA που αποτελούνται περίπου από $20-30$ νουκλεοτίδια τα οποία δεν κωδικοποιούν καμία πρωτεΐνη. Αντ΄ αυτού, λειτουργούν ως ρυθμιστές της έκφρασης των γονιδίων προσδενόμενα σε μόρια αγγελιαφόρου RNA (mRNA) και αποσταθεροποιώντας τα ή εμποδίζοντας τη μετάφρασή τους. Τα miRNA εμπλέκονται σε μια μεγάλη γκάμα φυσιολογικών μοριακών διαδικασιών και η απορρύθμισή τους οδηγεί σε διάφορες ασθένειες [23, 21, 49].

Τα miRNA βρίσκονται σε διαγονιδιακές περιοχές ή στα εσώνια των γονιδίων που κωδικοποιούν πρωτεΐνες. Μεταγράφονται από την RNA Πολυμεράση II ως ανεξάρτητα μετάγραφα ή ως μέρος του μεταγράφου ενός γονιδίου ξενιστή. Μόνο μια μικρή ομάδα miRNA που βρίσκεται μέσα σε επαναλαμβανόμενα στοιχεία ALU μεταγράφεται από την RNA Πολυμεράση III. Ένα μετάγραφο miRNA μπορεί να φιλοξενήσει πάνω από ένα miRNA και μπορεί να έχει μήκος πολλών χιλιάδων νουκλεοτιδίων συμπεριλαμβανομένων των εσωνίων.

Μια περιοχή υποκινητής βρίσκεται γύρω από το σημείο εκκίνησης μεταγραφής (transcription start site - TSS) ενός μεταγράφου και ρυθμίζεται από πρωτεΐνες που προσδένονται σε αυτή την περιοχή. Τα ως τώρα στοιχεία δείχνουν ότι τα σημεία πρόσδεσης για μεταγραφικούς παράγοντες είναι παρόμοια κατανεμημένα μέσα στους υποκινητές τόσο των γονιδίων που κωδικοποιούν πρωτεΐνες όσο και των μεταγράφων miRNA [62]. Τα πρωτογενή μετάγραφα των miRNA (*pri-miRNA*) υφίστανται επεξεργασία στον πυρήνα για να σχηματίσουν *pre-miRNA*, δομές στελέχους-βρόχου που αποτελούνται περίπου από 70 νουκλεοτίδια και λέγονται επίσης *φουρκέτες (hairpins)* miRNA. Αυτά μετατρέπονται στη συνέχεια σε ώριμα miRNA στο κυτταρόπλασμα μέσω αλληλεπίδρασης με την ενδονουκλεάση Dicer, η οποία εκκινεί επίσης το σχηματισμό του RNA επαγόμενου συμπλόκου αποσιώπησης (RNA-induced silencing complex-RISC). Καθώς τα πρωτογενή μετάγραφα είναι βραχύβια και βρίσκονται μόνο μέσα στον πυρήνα, είναι δύσκολο να τα αναγνωρίσουμε με τις συνήθεις μοριακές τεχνικές.

Αφού το ένζυμο Dicer κόψει το στέλεχος-βρόχο του pre-miRNA, σχηματίζονται δύο συμπληρωματικά μικρά μόρια RNA, αλλά μόνο το ένα από αυτά, ο κλώνος-οδηγός, εντάσσεται κυρίως στο σύμπλοκο RISC. Ο άλλος κλώνος, γνωστός ως miRNA*, κλώνος αντι-οδηγός ή κλώνος-επιβάτης, συνήθως υποβαθμίζεται. Ωστόσο, η αναλογία

ένταξης κάθε κλώνου ποικίλει ανάλογα με τα είδη miRNA, με κάποια miRNA να έχουν σχεδόν ίδια αφθονία καθενός από τους δυο κλώνους που έχουν ενταχθεί στο RISC. Μια άλλη κοινή ονομασία για τους συμπληρωματικούς κλώνους miRNA είναι η σύμβαση ονοματολογίας '-3p' και '-5p'. Τα ονόματα αυτά δεν υπονοούν ποιο miRNA εντάσσεται πιο συχνά στο σύμπλοκο RISC. Οι ονοματολογίες miRNA-miRNA* και miRNA-3p-miRNA-5p χρησιμοποιούνται εξίσου ευρέως στην κοινότητα, συχνά για να υποδηλώσουν το ίδιο συμπληρωματικό ζεύγος miRNA. Τα ώριμα μόρια miRNA δεσμεύονται από το σύμπλοκο RISC, οδηγούνται σε συγκεκριμένα μοτίβα μέσα στο 3′-UTR ή στα mRNA που κωδικοποιούν πρωτεΐνες, και εμποδίζουν αυτά τα mRNA να μεταφραστούν σε πρωτεΐνες. Η βιογένεση των miRNA και η ρύθμισή τους από τους μεταγραφικούς παράγοντες φαίνεται στο Σχήμα 2.1.



Σχήμα 2.1: Λειτουργίες των miRNA.

2.2.2 Υπολογιστικές μέθοδοι πρόβλεψης στόχων microRNA

Η γνώση των miRNA που στοχεύουν ένα συγκεκριμένο γονίδιο βοηθά στην κατανόηση των αιτίων των ανθρώπινων ασθενειών και την αναζήτηση θεραπειών. Ωστόσο, τα βιοχημικά πειράματα που αποκαλύπτουν τους στόχους των miRNA είναι ακριβά και χρονοβόρα. Οπότε, έχει προταθεί πληθώρα υπολογιστικών μεθόδων για την πρόβλεψη των στόχων των miRNA, που λέγονται μέθοδοι πρόβλεψης στόχων miRNA. Οι πρώτες τέτοιες μέθοδοι αναπτύχθηκαν ήδη το 2003. Κατά την τελευταία δεκαετία, πάνω από δώδεκα τέτοιες μέθοδοι έχουν προταθεί, καθιστώντας τον κλάδο της πρόβλεψης στόχων miRNA έναν από τους πιο δραστήριους στη βιοπληροφορική. Μια άριστη επισκόπηση για τον κλάδο μπορεί να βρει κανείς στα [87, 2].

Μια από τις πιο ακριβείς και δημοφιλείς μεθόδους πρόβλεψης στόχων miRNA είναι το ΔΙΑΝΑ μιcroT. Καθώς στην παρούσα διατριβή παρουσιάζουμε πολλούς αποδοτικούς αλγορίθμους και συστήματα βασισμένα σε αυτή τη μέθοδο (βλ. Κεφάλαιο 3), στα επόμενα κεφάλαια συζητάμε τις πιο πρόσφατες εκδόσεις της μεθόδου αυτής.

2.2.2.1 DIANA microT v.3

Στη βάση της μεθόδου πρόβλεψης DIANA microT v.3 [59] βρίσκεται ένα σύνολο βιολογικών παρατηρήσεων. Ένα τυπικό miRNA έχει μήκος περίπου $20-30$ νουκλεοτίδια,

αλλά τα νουκλεοτίδια που βρίσκονται κοντά στο 5′-άκρο είναι πολύ μεγάλης σημασίας για την αναγνώριση μιας ακολουθίας-στόχου και την πρόσδεση σε αυτή. Συνήθως, απαιτείται ισχυρή πρόσδεση (δηλαδή, τουλάχιστον 7 διαδοχικά ταιριάσματα βάσεων τύπου Watson-Crick) μεταξύ των πρώτων 9 νουκλεοτιδίων από το 5′-άκρο της ακολουθίας miRNA (που θα λέμε *ακολουθία miRNA οδηγός*) και του γονιδίου-στόχος για επαρκή καταστολή της παραγωγής πρωτεϊνών. Ωστόσο, υπάρχουν πειραματικά στοιχεία [12] ότι μια πιο αδύναμη πρόσδεση, που αφορά μόνο 6 νουκλεοτίδια που έχουν ταιριάξει διαδοχικά ή περιλαμβάνει ζεύγη wobble G:U, μπορεί επίσης να καταστείλει την παραγωγή πρωτεϊνών εάν υπάρχει επιπλέον πρόσδεση μεταξύ του 3′-άκρο του miRNA και του γονιδίου-στόχου.

Η μέθοδος DIANA microT v.3 θεωρεί υποψήφιες περιοχές πρόσδεσης, εκείνες τις περιοχές UTR που έχουν διαδοχικά ταιριάσματα βάσεων τύπου Watson-Crick των 7, 8, ή 9 νουκλεοτιδίων με το miRNA, ξεκινώντας από την πρώτη ή τη δεύτερη θέση από το 5′-άκρο του miRNA. Για θέσεις με επιπλέον ταίριασμα βάσεων που αφορούν το 3′-άκρο του miRNA, επιτρέπονται επίσης ένα μονό ζεύγος wobble G:U ή πρόσδεση μόνο 6 διαδοχικών νουκλεοτιδίων στην ακολουθία-οδηγό. Χρησιμοποιώντας σαν χαρακτηριστικά τον τύπο πρόσδεσης και το προφίλ διατήρησης των υποψήφιων θέσεων πρόσδεσης, υπολογίζεται μια βαθμολογία για κάθε ένα από αυτά μέσω μιας συγκριτικής ανάλυσης προς ένα σύνολο υποψήφιων θέσεων πρόσδεσης που έχει αναγνωριστεί με βάση πλαστές ακολουθίες miRNA. Η συνολική βαθμολογία πρόβλεψης στόχων miRNA υπολογίζεται ως το σταθμισμένο άθροισμα των βαθμολογιών όλων των αναγνωρισμένων υποψήφιων θέσεων πρόσδεσης στο 3′-UTR. Η μέθοδος χρησιμοποιεί περίπου 27 είδη για την αξιολόγηση των προφίλ διατήρησης των υποψήφιων θέσεων πρόσδεσης, λαμβάνοντας υπόψη τόσο τις διατηρημένες όσο και τις μη διατηρημένες υποψήφιες θέσεις πρόσδεσης για την εκτίμηση της τελικής βαθμολογίας πρόβλεψης στόχων miRNA.

Για την αξιολόγηση των προβλεφθησών αλληλεπιδράσεων κάθε miRNA, η μέθοδος τις συγκρίνει με εκείνες που έχουν προβλεφθεί για ένα σύνολο πλαστών miRNA. Τα πλαστά miRNA δημιουργούνται ανεξάρτητα για κάθε πραγματικό miRNA και σχεδιάζονται έτσι ώστε να έχουν περίπου τον ίδιο αριθμό προβλεπόμενων στόχων με το πραγματικό miRNA. Αυτό επιτρέπει τον υπολογισμό του λόγου του σήματος προς το θόρυβο για συγκεκριμένο miRNA σε διαφορετικά κατώφλια αποκοπής βαθμολογίας πρόβλεψης στόχων miRNA καθώς και την εκτίμηση της βαθμολογίας ακρίβειας που παρέχει μια ένδειξη του λόγου ψευδών θετικών αποτελεσμάτων για μια συγκεκριμένη πρόβλεψη αλληλεπίδρασης miRNA.

2.2.2.2 DIANA microT v.4

Η μέθοδος πρόβλεψης στόχων DIANA microT v.4 [60] είναι σχεδόν ίδια με τη μέθοδο DIANA microT v.3. Η μόνη βελτίωση έγκειται στο ότι, σε αντίθεση με την προηγούμενη έκδοση που βασιζόταν σε χαρακτηριστικά που διαχώριζαν τα πραγματικά και τα πλαστά (ανακατεμένα) miRNA (βλ. Κεφάλαιο 2.2.2.1, η DIANA microT v.4 χρησιμοποιεί πειραματικά δεδομένα υψηλής ρυθμαπόδοσης για τον ίδιο σκοπό. Αυτά τα δεδομένα αποκτήθηκαν από τα πειράματα των Selbach et al. στο [83].

2.2.2.3 DIANA microT v.5

Η μέθοδος DIANA microT v.5 [78] βασίζεται στις προηγούμενες εκδόσεις του DIANA microT και σε χαρακτηριστικά που εξάγονται από υπάρχοντα δεδομένα ανοσοκαθίζησης και ακολουθιοποίησης υψηλής ρυθμαπόδοσης από θηλαστικά. Η ανάλυση

εκτελείται ανεξάρτητα για τις κωδικές περιοχές και τις περιοχές 3′-UTR των γονιδίων και αποκαλύπτει διαφορετικά σύνολα χαρακτηριστικών και μοντέλα για τις δυο περιοχές. Τα δυο μοντέλα συνδυάζονται σε ένα νέο υπολογιστικό μοντέλο για τα γονίδια-στόχους των miRNA, το οποίο πετυχαίνει μεγαλύτερη ευαισθησία συγκριτικά με άλλα δημοφιλή προγράμματα και τις προηγούμενες εκδόσεις του DIANA microT . Περαιτέρω ανάλυση δείχνει ότι γονίδια με πιο κοντά 3′-UTR προτιμούνται ως στόχοι στις κωδικές περιοχές, δείχνοντας ότι η εξελικτική επιλογή ίσως ευνοεί επιπλέον θέσεις στις κωδικές περιοχές σε περιπτώσεις όπου υπάρχει περιορισμένος χώρος στο 3′-UTR. Να σημειωθεί ότι, το μοντέλο 3′-UTR είναι το ίδιο με τις προηγούμενες μεθόδους DIANA microT .

2.2.3 Στοίχιση αναγνωσμάτων DNA

Η μελέτη των μηχανισμών της ζωής πραγματοποιείται κυρίως με την εκτέλεση βιοχημικών πειραμάτων που ερευνούν τις πιθανές αλληλεπιδράσεις που προκύπτουν μεταξύ βιομορίων (όπως το DNA, το RNA και οι πρωτεΐνες). Ωστόσο, αυτά τα πειράματα είναι ακριβά και χρονοβόρα. Οπότε, η έρευνα της πιθανής αλληλεπίδρασης μεταξύ δυο μορίων τα οποία δε συσχετίζονται είναι απλώς σπατάλη χρόνου και χρημάτων. Αυτός είναι ο λόγος που οι ερευνητές αναπαριστούν τα μόρια αυτά ως ακολουθίες συμβόλων και εφαρμόζουν προσεγγίσεις υπολογιστικής επεξεργασίας για να αποκαλύψουν πιθανές αλληλεπιδράσεις.

Οι υπολογιστικές αναπαραστάσεις βιομορίων παράγονται με τη χρήση μηχανών ειδικού σκοπού, που λέγονται *μηχανές ακολουθιοποίησης*. Οι μηχανές αυτές λαμβάνουν βιολογικά δείγματα που περιέχουν κύτταρα από συγκεκριμένους ιστούς και παράγουν μικρές ακολουθίες, κάθε μια από τις οποίες αναπαριστά ένα τμήμα του DNA που περιέχεται στο δείγμα. Οι προαναφερθείσες μικρές ακολουθίες λέγονται *αναγνώσματα DNA*, ή απλώς *αναγνώσματα*. Πολλαπλά αναγνώσματα που εξάγονται από δείγματα του ίδιου ζωντανού οργανισμού μπορούν να στοιχιστούν μεταξύ τους για να παράγουν ολόκληρη τη γονιδιακή ακολουθία αυτού του οργανισμού. Επιπλέον, πολλοί άλλοι σημαντικοί τύποι ανάλυσης DNA περιλαμβάνουν στοίχιση αναγνωσμάτων DNA σε μια ήδη γνωστή γονιδιακή ακολουθία. Ο όρος *στοίχιση αναγνωσμάτων DNA* χρησιμοποιείται για την περιγραφή όλων των ανωτέρω τύπων ανάλυσης.

Στον πυρήνα του, το πρόβλημα στοίχισης αναγνωσμάτων μοιάζει με το σύνηθες πρόβλημα στοίχισης ακολουθιών στην επεξεργασία δεδομένων (βλ. Κεφάλαιο 2.1.1.2). Οπότε, όλες οι μέθοδοι που περιγράφονται στο Κεφάλαιο 2.1.1.3 μπορούν επίσης να εφαρμοστούν για την παραγωγή των στοιχίσεων αναγνωσμάτων DNA.

Η πρώτη αυτοματοποιημένη μηχανή ακολουθιοποίησης εισήχθη από την Applied Biosystems το 1987 και επέτρεψε την ολοκλήρωση του έργου αποκωδικοποίησης του ανθρώπινου γονιδιώματος το 2001. Έκανε χρήση της μεθόδου ακολουθιοποίησης Sanger, μιας τεχνολογίας που χρησιμοποιούν όλες οι μηχανές ακολουθιοποίησης πρώτης γενιάς.

Το έργο αποκωδικοποίησης του ανθρώπινου γονιδιώματος κινητοποίησε την ανάπτυξη μιας νέας γενιάς πιο φθηνών, υψηλής ρυθμαπόδοσης μηχανών ακολουθιοποίησης. Αυτή η γενιά, γνωστή ως *μηχανές ακολουθιοποίησης νέας γενιάς*, περιλαμβάνει τις πλατφόρμες ακολουθιοποίησης 454, SoLiD και Illumina. Οι μηχανές ακολουθιοποίησης νέας γενιάς ώθησαν το ρυθμό της ακολουθιοποίησης DNA σε σύγκριση με τις προηγούμενες μεθόδους που είναι βασισμένες στην τεχνική Sanger. Τα αναγνώσματα DNA που παράγονται από αυτές τις μηχανές είναι πολύ μικρά και αποτελούνται από

$30 - 50$ σύμβολα. Μπορούν επίσης να περιέχουν σφάλματα που προκαλούνται από την τεχνική ακολουθιοποίησης του DNA ή από μια επεξεργασία που πραγματοποιείται μετά από αυτή, την ενίσχυση PCR. Τις τελευταίες δεκαετίες, έχουν προταθεί πολλές μέθοδοι στοίχισης ακολουθιών και έχει αποδειχτεί ότι δουλεύουν ικανοποιητικά για αναγνώσματα DNA που έχουν τα ανωτέρω χαρακτηριστικά (βλ. Κεφάλαιο 3.1.4.2).

Ωστόσο, τα μικρά μεγέθη αναγνωσμάτων μπορούν να οδηγήσουν σε προβλήματα κατά την επεξεργασία της ακολουθίας. Για παράδειγμα, η στοίχιση μικρών αναγνωσμάτων DNA μπορεί να επηρεαστεί δραματικά από την παρουσία δομικών παραλλαγών ή πολυμορφισμών ενός νουκλεοτιδίου [single-nucleotide polymorphisms (SNPs)] μέσα τους. Κάτι τέτοιο έδωσε ώθηση σε μια νέα γενιά μηχανών ακολουθιοποίησης που μπορούν να παράγουν πολύ μεγαλύτερα αναγνώσματα. Για παράδειγμα, η μηχανή Ion Torrent PGM παράγει αναγνώσματα $200-400$ συμβόλων, η 454 GS FLX αναγνώσματα 700 συμβόλων, και η PacBio αναγνώσματα που περιέχουν αρκετές χιλιάδες συμβόλων. Ένα παράπλευρο αποτέλεσμα είναι ότι, προκειμένου να παρέχουν αυτό το προνόμιο, αυτές οι μηχανές συνήθως θυσιάζουν την ακρίβεια, κάτι που σημαίνει ότι τα αναγνώσματά τους περιέχουν αυξημένο πλήθος λανθασμένα τοποθετημένων συμβόλων. Αυτό δημιουργεί μια πρόκληση για τη στοίχιση ακολουθιών καθώς οι υπάρχουσες προσεγγίσεις αιχμής είναι βέλτιστες για μικρά αναγνώσματα που περιλαμβάνουν μικρό πλήθος λαθών και οι επιδόσεις τους είναι φτωχές για μεγάλα και μη ακριβή αναγνώσματα.

Κεφάλαιο 3

Αποδοτική πρόβλεψη στόχων **miRNA**

Για πολλά χρόνια, οι ερευνητές των βιοεπιστημών πίστευαν ότι μόνο οι περιοχές του γονιδιώματος που μεταφράζονται σε πρωτεΐνες είναι σημαντικές για τη ζωή. Αυτό έχει αλλάξει δραματικά μετά την ανακάλυψη, στα τέλη της δεκαετίας του ′90, περιοχών στο 'μη μεταφράσιμο' γονιδίωμα που παίζουν σημαντικό ρόλο σε αρκετές λειτουργίες της ζωής. Μεταξύ αυτών των λειτουργιών, μια από τις πιο σημαντικές είναι η απενεργοποίηση ('αδρανοποίηση') των γονιδίων από μικρά μόρια RNA, που λέγονται microRNA (*miRNA*). Εν συντομία, κάθε miRNA *στοχεύει* συγκεκριμένα γονίδια, καταστρέφοντας τα μετάγραφά τους και, συνεπώς, εμποδίζοντας την παραγωγή της κωδικοποιούμενης πρωτεΐνης (βλ. Κεφάλαιο 2.2.1).

Η γνώση των miRNA που στοχεύουν ένα συγκεκριμένο γονίδιο βοηθά στην κατανόηση των αιτίων των ανθρώπινων ασθενειών και στην ανάπτυξη θεραπειών για αυτές τις ασθένειες. Ωστόσο, τα βιοχημικά πειράματα που μπορούν να προσδιορίσουν τους στόχους είναι ακριβά και χρονοβόρα. Οπότε, έχουν προταθεί υπολογιστικές μέθοδοι για την πρόβλεψη των στόχων των miRNA (βλ. Κεφάλαιο 2.2.2). Αν και αυτές οι μέθοδοι αποτελούν μεγάλη ώθηση σε σχέση με τα βιοχημικά πειράματα, περιλαμβάνουν απαιτητικούς υπολογισμούς και, έτσι, απαιτούν σημαντικό χρόνο για την παραγωγή των στόχων όλων των γνωστών ακολουθιών miRNA.

Επιλέξαμε να μελετήσουμε και να δώσουμε ώθηση στο DIANA microT (Κεφάλαιο 2.2.2.3), μια από τις πιο δημοφιλείς και αξιόπιστες μεθόδους πρόβλεψης στόχων miRNA. Στο παρόν κεφάλαιο, μελετάμε τη δουλειά μας προς αυτή την κατεύθυνση. Πρώτα, στο Κεφάλαιο 3.1 μελετάμε τη διαδικασία ταιριάσματος ακολουθιών η οποία αποτελεί το πρώτο βήμα του DIANA microT και προτείνουμε μια προσέγγιση για την επιτάχυνση της διαδικασίας. Καθώς οι μέθοδοι πρόβλεψης στόχων ενσωματώνουν και άλλες υπολογιστικά απαιτητικές διαδικασίες, πέρα από το ταίριασμα ακολουθιών, στο Κεφάλαιο 3.2 αναπτύσσουμε δυο προσεγγίσεις βασισμένες στο Νέφος οι οποίες κατανέμουν αυτές τις διαδικασίες σε πολλούς υπολογιστικούς κόμβους. Τέλος, στο Κεφάλαιο 3.3 ανακεφαλαιώνουμε την εργασία μας σε αυτό τον κλάδο και αναφέρουμε τη συνεισφορά μας.

3.1 Αποδοτικό ταίριασμα ακολουθιών για πρόβλεψη στόχων

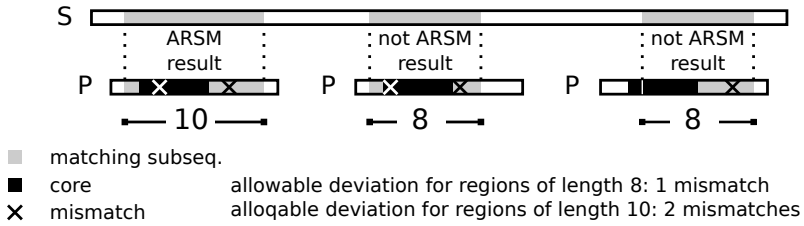
3.1.1 Κίνητρο και συνεισφορά

Τα προβλήματα ταιριάσματος ακολουθιών (π.χ., ακριβής/κατά προσέγγιση, ολική/ τοπική/ διπλά-τοπική στοίχιση) έχουν μελετηθεί εκτεταμένα, και έχουν προταθεί αρκετοί αλγόριθμοι, που εξετάζονται στα [82, 26] (βλ. Κεφάλαιο 2.1.1.3). Αυτά τα προβλήματα είναι πολύ δημοφιλή καθώς προκύπτουν φυσικά στον πυρήνα πολλών διαφορετικών εφαρμογών. Για παράδειγμα, στις βιολογικές βάσεις δεδομένων, που περιέχουν μεγάλες ακολουθίες συμβόλων (όπως τα νουκλεοτίδια, τα αμινοξέα, κλπ.), οι αλγόριθμοι ταιριάσματος ακολουθιών βοηθούν στον εντοπισμό ομόλογων (δηλαδή, παρόμοιας λειτουργικότητας) βιολογικών οντοτήτων, όπως τα γονίδια, οι πρωτεΐνες κλπ (βλ. επίσης Κεφάλαιο 2.2).

Συχνά, η πρόοδος σε συγκεκριμένους ερευνητικούς κλάδους δημιουργεί την ανάγκη για πολύπλοκα κριτήρια ταιριάσματος που ωθεί την ανάπτυξη *νέων προβλημάτων ταιριάσματος ακολουθιών*. Για παράδειγμα, έχει παρατηρηθεί ότι μια χημική συσχέτιση, γνωστή ως πρόσδεση, μιας μη κωδικοποιούμενης ακολουθίας RNA (π.χ. κάποιο miRNA ή κάποιο μικρό παρεμβαλλόμενο RNA κλπ.), που λέγεται *πρότυπο*, με μια μεγαλύτερη (π.χ. ένα γονίδιο), που λέγεται τα *δεδομένα*, συνήθως συμβαίνει γύρω από μια τοποθεσία-κλειδί του προτύπου, που λέγεται *πυρήνας* (π.χ. τα νουκλεοτίδια κοντά στην αρχή του miRNA [19]). Καθώς τα εργαστηριακά πειράματα κοστίζουν και είναι χρονοβόρα, έχουν προταθεί υπολογιστικές μέθοδοι για την πρόβλεψη προσδέσεων βάσει της προηγούμενης παρατήρησης (βλ. Κεφάλαιο 2.2.2). Για παράδειγμα, ο αλγόριθμος πρόβλεψης DIANA microT (βλ. Κεφάλαιο 2.2.2.1) χρησιμοποιεί ένα συμβατικό αλγόριθμο ταιριάσματος ακολουθιών για να ελέγξει εάν κάποια υπερακολουθία του πυρήνα, που λέγεται *περιοχή*, ταιριάζει *κατά προσέγγιση* (δηλαδή, μερικά ξεχωριστά σύμβολα μπορεί να μην ταιριάζουν) με μια υπακολουθία των δεδομένων. Αυτή η διαδικασία επαναλαμβάνεται για κάθε περιοχή, και ο μέγιστος επιτρεπτός αριθμός αστοχιών (μη ταιριασμάτων) ορίζεται εμπειρικά βάσει του μήκους της περιοχής. Όσο μεγαλύτερη είναι η περιοχή και όσο καλύτερα ταιριάζει με τα δεδομένα, τόσο πιο πιθανή είναι μια πρόσδεση.

Κινητοποιημένοι από αυτή την πραγματική εφαρμογή, γενικεύσαμε τα ανωτέρω κριτήρια ταιριάσματος και εισάγαμε το πρόβλημα του *Κατά Προσέγγιση Ταιριάσματος Περιοχών Ακολουθίας* (Approximate Regional Sequence Matching - ARSM) [98] (βλ. επίσης Κεφάλαιο 3.1.2. Έστω ότι υπάρχει μια ακολουθία *δεδομένων* S, μια ακολουθία *πρότυπο* P, και ένας *πυρήνας* (δηλαδή, μια υπακολουθία) του P. Εν συντομία, ένα *αποτέλεσμα του ARSM* είναι μια υπακολουθία της S που ταιριάζει κατά προσέγγιση με κάποια υπακολουθία του P υπό τις ακόλουθες προϋποθέσεις: (α) η υπακολουθία του P είναι μια *περιοχή*, δηλαδή, περικλείει τον πυρήνα, και (β) η επιτρεπτή απόκλιση, που αφορά τον αριθμό συμβόλων που δεν ταιριάζουν, ανάμεσα στην υπακολουθία της S και την περιοχή P αυξάνεται με το μήκος της τελευταίας.

Το Σχήμα 3.1 παρουσιάζει ένα παράδειγμα προβλήματος ARSM. Το πάνω μέρος του σχήματος απεικονίζει την ακολουθία δεδομένων S, ενώ το κάτω μέρος δείχνει τρία αντίγραφα της ακολουθίας πρότυπο P στοιχισμένα σε διαφορετικές περιοχές κάτω από την S. Το σκούρο σκιασμένο τμήμα κάθε αντιγράφου του P αντιστοιχεί στην περιοχή του πυρήνα. Από την άλλη πλευρά, το ανοιχτόχρωμο σκιασμένο τμήμα απεικονίζει μια υπακολουθία του P (διαφορετική σε κάθε αντίγραφο) που ταιριάζει με την αντίστοιχη



Σχήμα 3.1: Ένα παράδειγμα προβλήματος ARSM.

ανοιχτόχρωμη σκιασμένη υπακολουθία της S. Σε κάθε υπακολουθία του P, ο αριθμός κάτω από αυτή υποδηλώνει το μήκος της, ενώ ένας σταυρός δείχνει ένα σύμβολο που δεν ταιριάζει σε σχέση με την S. Επιπλέον, ο επιτρεπτός αριθμός συμβόλων που δεν ταιριάζουν είναι 1 (αντίστοιχα 2) για περιοχές μήκους 8 (αντίστοιχα 10).

Παρατηρούμε ότι η υπακολουθία της S που ταιριάζει με τη δεύτερη υπακολουθία του P δεν είναι αποτέλεσμα ARSM επειδή αυτή η υπακολουθία του P έχει μήκος 8 και περιέχει περισσότερες αστοχίες από το επιτρεπτό. Επιπλέον, ούτε η υπακολουθία της S που αντιστοιχεί στην τρίτη υπακολουθία του P είναι αποτέλεσμα ARSM, καθώς η υπακολουθία του P δεν περικλείει τον πυρήνα, δηλαδή, δεν είναι περιοχή. Από την άλλη πλευρά, η υπακολουθία της S που αντιστοιχεί στην πρώτη υπακολουθία του P *είναι* αποτέλεσμα ARSM καθώς ικανοποιεί και τις δυο προϋποθέσεις.

Το διακριτό χαρακτηριστικό του ARSM, σε σύγκριση με άλλα προβλήματα κατά προσέγγιση ταιριάσματος ακολουθιών, είναι ότι *πολλαπλές* ακολουθίες, οι περιοχές, εξετάζονται για ταιριάσματα κάτω από *ποικίλες* επιτρεπτές τιμές απόκλισης. Σημειώστε ότι είναι δυνατό να επεκτείνουμε υπάρχουσες μεθόδους για την επίλυση του προβλήματος ARSM. Η αφελής προσέγγιση είναι η εφαρμογή ενός αλγορίθμου αιχμής για κατά προσέγγιση ταίριασμα ακολουθιών (ASM) (π.χ. [67]) για κάθε πιθανή περιοχή. Προφανώς, αυτή η μέθοδος ωμής βίας δεν είναι αποδοτική, καθώς δεν κάνει προσπάθεια να μοιράστει υπολογισμούς μεταξύ των περιοχών που είναι επικαλυπτόμενες.

Μια καλύτερη εναλλακτική είναι να εφαρμόσουμε έναν αλγόριθμο πολλαπλού ASM (MASM) (π.χ., τα [66, 22]) που μπορεί να επεξεργάζεται πολλαπλά πρότυπα ταυτόχρονα και να εκμεταλλεύεται τις επικαλύψεις τους. Καθώς οι αλγόριθμοι MASM έχουν σχεδιαστεί για να λειτουργούν σε ένα σύνολο προτύπων παρόμοιου μήκους (βλ. επίσης Κεφάλαιο 2.1.1.3), μια προσέγγιση βασισμένη σε MASM πρέπει πρώτα να ομαδοποιήσει τις περιοχές σύμφωνα με το μήκος τους, και να εκτελέσει MASM μια φορά ανά ομάδα. Ωστόσο, αυτή η μέθοδος δεν μπορεί να εκμεταλλευτεί τις επικαλύψεις που εμφανίζονται σε περιοχές που ανήκουν σε διαφορετικές ομάδες. Επιπλέον, για γονιδιακές βάσεις δεδομένων, όπου έχουμε μικρό μέγεθος αλφαβήτου (4 σύμβολα), μικρά πρότυπα (λίγες δεκάδες σύμβολα) και μεγάλες επιτρεπτές αποκλίσεις (περίπου 20% του μήκους του προτύπου), οι αλγόριθμοι MASM είναι γνωστό ότι υστερούν [22].

Να σημειωθεί ότι οι αλγόριθμοι τοπικής στοίχισης (π.χ. ο αλγόριθμος των Smith-Waterman [90]), που ψάχνουν ταιριάσματα όλων των πιθανών υπακολουθιών προτύπων (και έτσι και των περιοχών), δεν μπορούν να προσαρμοστούν στο πρόβλημα ARSM για τρεις λόγους. Πρώτον, απαιτούν η επιτρεπτή απόκλιση να είναι σταθερή και ανεξάρτητη από το μήκος της υπακολουθίας. Δεύτερον, οι δημοφιλείς ευριστικοί αλγόριθμοι αιχμής (όπως το BLAST [4]) δεν αναγνωρίζουν όλα τα ταιριάσματα. Τρίτον, και πιο σημαντικό, ακόμα και αν χρησιμοποιηθούν ακριβείς αλγόριθμοι, κάποιες απαντήσεις ARSM μπορεί και πάλι να χαθούν. Έστω, για παράδειγμα, η ακολουθία δεδομένων $S = \cdots \texttt{GTTGA} \cdots$, και η περιοχή $R = \texttt{GCCGA}$. Προφανώς, υπάρχει εμφάνιση της R στην S με κόστος (συντακτική απόσταση) 2. Ωστόσο, ο αλγόριθμος SW θα αποτύγχανε

να την αναγνωρίσει[1]. Ο λόγος είναι ο εξής. Στον πίνακα δυναμικού προγραμματισμού, το κελί που αντιστοιχεί στα δυο `A` στις ακολουθίες έχει την υψηλότερη τιμή 2, κάτι που σημαίνει ότι υπάρχουν δύο υπακολουθίες που λήγουν σε `A` οι οποίες έχουν σκορ ομοιότητας 2. Ωστόσο, καθώς αυτό το σκορ αντιστοιχεί στις υπακολουθίες `GA`, δεν υπάρχει τρόπος για ανάστροφη ιχνηλάτηση και αναγνώριση των υπακολουθιών `GCCGA` και `GTTGA`.

Για να ξεπεράσουμε τους ανωτέρω περιορισμούς, προτείνουμε τη μέθοδο `PS-ARSM` [98], η οποία εκμεταλλεύεται τις επικαλύψεις προθέματος και επιθέματος μεταξύ των περιοχών (βλ. επίσης Κεφάλαιο 3.1.4). Εν συντομία, η μέθοδός μας πρώτα ορίζει τις υπακολουθίες δεδομένων όπου η πιο μικρή περιοχή (ο πυρήνας) ταιριάζει κάτω από τη μεγαλύτερη δυνατή επιτρεπτή απόκλιση. Έπειτα, βάσει ενός συνόλου κανόνων επέκτασης που είναι ορθοί και πλήρεις, ο αλγόριθμος σταδιακά επεκτείνει τις υπακολουθίες των δεδομένων για να αντλήσει όλα τα αποτελέσματα του ARSM.

Στα επόμενα κεφάλαια εισάγουμε το πρόβλημα ARSM, περιγράφουμε τη μέθοδο `PS-ARSM` και αξιολογούμε την αποδοτικότητά της συγκριτικά με μεθόδους βασισμένες σε ASM και MASM αλγορίθμους εκτελώντας πειράματα σε γονιδιακές βάσεις δεδομένων.

3.1.2 Το πρόβλημα ARSM

Έστω ένα αλφάβητο Σ. Στο υπόλοιπο κομμάτι αυτού του κεφαλαίου, κάθε ακολουθία $S \in \Sigma^*$. Το $|S|$ υποδηλώνει το μήκος της ακολουθίας S. Η $S_{[i]}$ αντιστοιχεί στο i-οστό σύμβολο στην S, ενώ η $S_{[i,j]}$ στην υπακολουθία της S που ξεκινά στο i-οστό και καταλήγει στο j-οστό σύμβολο. Χρησιμοποιούμε το συμβολισμό $S_{[i,j]} \sqsubset S$ για να δείξουμε ότι η $S_{[i,j]}$ είναι υπακολουθία της S.

Έχοντας δυο ακολουθίες, αποκαλούμε *συντακτικό μετάγραφο* (ή απλώς *μετάγραφο*) κάθε διατεταγμένο σύνολο τ από *συντακτικές λειτουργίες* που μετατρέπουν τη μία ακολουθία στην άλλη. Χαρακτηριστικά, οι ακόλουθες συντακτικές λειτουργίες επιτρέπονται για μια ακολουθία S:

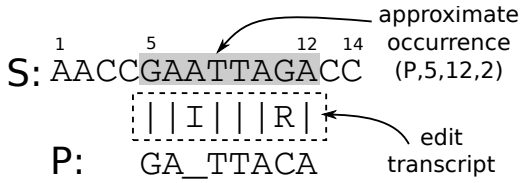
- Εισάγουμε (`I`) ένα σύμβολο στην S,
- Διαγράφουμε (`D`)ένα σύμβολο από την S,
- Αντικαθιστούμε (`R`) ένα σύμβολο της S με ένα άλλο,
- Ταιριάζουμε (`M`), δηλαδή, διατηρούμε, ένα σύμβολο της S.

Το *κόστος* $c(\tau)$ ενός μεταγράφου είναι ο αριθμός των λειτουργιών `I`, `D`, και `R` που περιέχει.

Έστω δυο ακολουθίες S και P, που λέγονται ακολουθία δεδομένων και ακολουθία πρότυπο, αντίστοιχα. Μπορεί κανείς πάντα να βρει ένα μετάγραφο που μετατρέπει την P σε οποιαδήποτε υπακολουθία $S_{[i,j]}$ της ακολουθίας δεδομένων. Λέμε ότι το πρότυπο P *εμφανίζεται* στα δεδομένα S σε τοποθεσία $[i, j]$ με κόστος μεταγράφου ϵ, εάν υπάρχει ένα μετάγραφο τ που μετατρέπει την P σε $S_{[i,j]}$ με κόστος $c(\tau) = \epsilon$. Χρησιμοποιούμε το συμβολισμό (P, i, j, ϵ) για να δείξουμε αυτή την *εμφάνιση*.

[1] Υποθέτουμε ότι ένα ταίριασμα έχει σκορ 1, ενώ μια αστοχία, διαγραφή ή εισαγωγή έχει σκορ -1. Σημειώστε ότι υπάρχουν παρόμοια αντιπαραδείγματα και για άλλα σκορ.

Το Σχήμα 3.2 δείχνει μια εμφάνιση $(P, 5, 12, 2)$ του προτύπου $P =$ `GATTACA` ιν $S =$ `AACCGAATTAGACC` στην τοποθεσία $[5, 12]$ με κόστος μεταγράφου $\epsilon = 2$. Πράγματι, σύμφωνα με το μετάγραφο $\tau =$ `MMIMMMRM`, η P μπορεί να μετατραπεί σε $S_{[5,12]}$ εισάγοντας (`I`) το σύμβολο `A` μεταξύ της $P_{[2]}$ και της $P_{[3]}$, και αντικαθιστώντας (`R`) την $P_{[6]} =$ `C` με την `G`. Όλα τα άλλα σύμβολα μένουν τα ίδια (`M`).



Σχήμα 3.2: Μια κατά προσέγγιση εμφάνιση της `GATTACA` στην `AACCGAATTACACC`.

Σημειώστε ότι μπορούν να υπάρχουν περισσότερα από ένα μετάγραφα για να μετατρέψουμε την P σε $S_{[i,j]}$ με το ίδιο κόστος ϵ. Για παράδειγμα, στο παράδειγμα του Σχήματος 3.2, το μετάγραφο $\tau' =$ `MIMMMMRM` έχει το ίδιο κόστος με το εικονιζόμενο.

Ένα πρότυπο μπορεί να εμφανιστεί σε μια συγκεκριμένη τοποθεσία στα δεδομένα με ποικίλα κόστη μεταγράφων. Μια εμφάνιση (P, i, j, ϵ) λέγεται *ελάχιστη* εάν έχει το χαμηλότερο δυνατό κόστος, δηλαδή, εάν δεν υπάρχει άλλη εμφάνιση (P, i, j, ϵ') ούτως ώστε $\epsilon' < \epsilon$. Για παράδειγμα, η εμφάνιση $(P, 5, 12, 2)$ στο Σχήμα 3.2 είναι ελάχιστη.

Με δεδομένη μια ακολουθία πρότυπο P και μια τοποθεσία $[a, b]$ στην P που λέγεται *πυρήνας* ($1 \leq a \leq b \leq |P|$), μια υπακολουθία $R = P_{[i,j]}$ της P λέγεται *περιοχή* εάν $i \leq a$ και $j \geq b$. Δυο ειδικές περιοχές είναι η *περιοχή του πυρήνα* $P_{[a,b]}$ και η *περιοχή του προτύπου* $P_{[1,|P|]}$. Στη συνέχεια εισάγουμε το πρόβλημα ARSM.

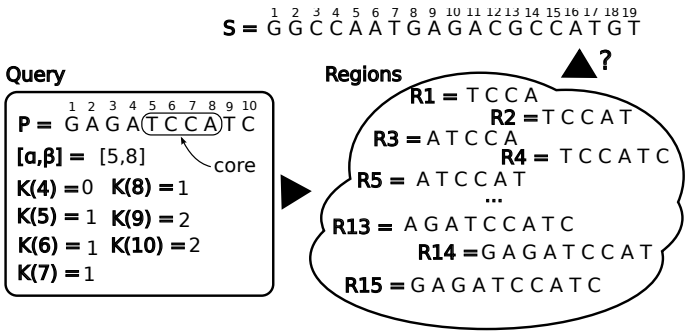
Ορισμός 3.1 (Πρόβλημα ARSM)**.** *Δεδομένης μιας ακολουθίας δεδομένων S, ενός προτύπου P, του πυρήνα του $[a, b]$, και μιας μονοτονικά αυξανόμενης συνάρτησης κατωφλίου $\mathcal{K} : \mathbb{N} \to \mathbb{N}$, το* πρόβλημα ARSM *αφορά την ανάκτηση των ελάχιστων εμφανίσεων (R, i, j, ϵ) κάθε περιοχής R του P, ούτως ώστε: (1) $\epsilon \leq \mathcal{K}(|R|)$, και (2) καμιά άλλη ελάχιστη εμφάνιση (R', i, j, ϵ'), όπου $R \sqsubset R'$, δεν έχει $\epsilon' \leq \mathcal{K}(|R'|)$.* □

Ο πρώτος περιορισμός ορίζει ότι το επιτρεπόμενο κόστος για μια εμφάνιση περιοχής εξαρτάται από το μέγεθος της περιοχής και δίνεται από τη συνάρτηση κατωφλίου. Μεγαλύτερες περιοχές επιτρέπεται να έχουν μεγαλύτερο κόστος. Ο δεύτερος περιορισμός υπονοεί ότι εάν δυο διαφορετικές περιοχές R, R' εμφανιστούν στην ίδια τοποθεσία $[i, j]$ της ακολουθίας δεδομένων S, τότε μόνο η εμφάνιση της μεγαλύτερης περιοχής επιστρέφεται. Αποκαλούμε όλες αυτές τις ανακτηθείσες ελάχιστες εμφανίσεις *αποτελέσματα ARSM*.

Το Σχήμα 3.3 απεικονίζει ένα στιγμιότυπο του προβλήματος ARSM στο οποίο η ακολουθία δεδομένων S έχει 19 σύμβολα και το πρότυπο P έχει 10. Ο πυρήνας είναι η τοποθεσία $[5, 8]$ του προτύπου. Σε αυτό το στιγμιότυπο, υπάρχουν 15 πιθανές περιοχές που ονομάζονται $R1$ έως $R15$. Το Σχήμα 3.3 απεικονίζει επίσης τις τιμές της συνάρτησης κατωφλίου $\mathcal{K}$ για όλα τα πιθανά μήκη περιοχών (από μήκος 4 έως 10).

3.1.3 Χαρακτηριστικά του ARSM

Το Κεφάλαιο 3.1.3.1 παρουσιάζει κάποιες βασικές παρατηρήσεις σχετικά με εμφανίσεις επικαλυπτόμενων περιοχών. Υπενθυμίζουμε ότι οι περιοχές είναι σε μεγάλο βαθμό επικαλυπτόμενες (βλ. Σχήμα 3.3), καθώς κάθε μια είναι υπακολουθία του προτύπου και υπερακολουθία του πυρήνα. Οπότε, το Κεφάλαιο 3.1.3.2 εκμεταλλεύεται αυτές τις



Σχήμα 3.3: Παράδειγμα ARSM

παρατηρήσεις για να εισάγει ένα σύνολο κανόνων επέκτασης που κατασκευάζουν το σύνολο των ελάχιστων εμφανίσεων μιας περιοχής από εκείνες μιας μικρότερης.

3.1.3.1 Επικαλυπτόμενες εμφανίσεις

Έστω μια ακολουθία δεδομένων S και ένα πρότυπο P. Υποθέτοντας ότι το P εμφανίζεται στην τοποθεσία $[i, j]$ της S, τα επόμενα δύο λήμματα δείχνουν πώς αυτή η εμφάνιση συνδέεται με μια εμφάνιση του P στις τοποθεσίες $[i, j + 1]$ και $[i - 1, j]$. Διαισθητικά, μια συντακτική λειτουργία `I` μπορεί να προσαρτηθεί στο μετάγραφο για να βρει χώρο για το επιπλέον σύμβολο της ακολουθίας δεδομένων της τοποθεσίας $j+1$ ή $i - 1$.

Λήμμα 3.1. *Εάν το (P, i, j, ϵ) είναι εμφάνιση του P στην ακολουθία δεδομένων S και τ είναι ένα από τα μετάγραφά του, τότε το $(P, i, j + 1, \epsilon + 1)$ είναι και αυτό εμφάνιση του P στην S και τ`I` είναι ένα από τα μετάγραφά του.*

Απόδειξη. Το μετάγραφο τ`I` περιέχει τις ίδιες συντακτικές λειτουργίες με το τ, και μια επιπλέον `I` λειτουργία. Οπότε, $c(\tau\texttt{I}) = c(\tau) + 1 = \epsilon + 1$. Καθώς το τ μετατρέπει το P σε $S_{[i,j]}$, το τελευταίο `I` στο τ`I` εισάγει $S_{[j+1]}$ στο τέλος του P. Οπότε, το τ`I`, με κόστος μεταγράφου $\epsilon + 1$, μετατρέπει το P σε $S_{[i,j+1]}$, δηλαδή το $(P, i, j + 1, \epsilon + 1)$ είναι εμφάνιση του P στην S και το τ`I` είναι ένα από τα μετάγραφά του. □

Λήμμα 3.2. *Εάν το (P, i, j, ϵ) είναι εμφάνιση του P στην ακολουθία δεδομένων S και το τ είναι ένα από τα μετάγραφά του, τότε το $(P, i - 1, j, \epsilon + 1)$ είναι και αυτό εμφάνιση του P στην S και το $I\tau$ είναι ένα από τα μετάγραφά του.*

Απόδειξη. Παρόμοια με εκείνη του Λήμματος 3.1. □

Έπειτα, έστω δυο ακολουθίες πρότυπα, P και $P\gamma$, όπου η τελευταία δημιουργείται με την προσάρτηση ενός συμβόλου $\gamma \in \Sigma$ στο τέλος της πρώτης. Αυτό αντιπροσωπεύει την περίπτωση όπου περιοχές μοιράζονται το ίδιο πρόθεμα και διαφέρουν κατά ένα σύμβολο, π.χ. οι περιοχές $R2$ και $R4$ στο Σχήμα 3.3. Υποθέτοντας ότι το P εμφανίζεται στην τοποθεσία $[i, j]$ στην S, το επόμενο λήμμα δείχνει πώς αυτή η εμφάνιση σχετίζεται με εμφανίσεις του $P\gamma$ στις τοποθεσίες $[i, j]$ και $[i, j + 1]$. Διαισθητικά, μια συντακτική λειτουργία (`D`, `R`, ή `M`) μπορεί να προσαρτηθεί στο μετάγραφο για να βρει χώρο για το επιπλέον σύμβολο γ της ακολουθίας πρότυπο.

Λήμμα 3.3. *Εάν το (P, i, j, ϵ) είναι εμφάνιση του P στην ακολουθία δεδομένων S και το τ είναι ένα από τα μετάγραφά του, τότε:*

1. *το* $(P\gamma, i, j, \epsilon + 1)$ *είναι εμφάνιση του* $P\gamma$ *στην* S *και το* τ`D` *είναι ένα από τα μετάγραφά του,*

2. *το* $(P\gamma, i, j + 1, \epsilon)$ *είναι εμφάνιση του* $P\gamma$ *στην* S *και το* τ`M` *είναι ένα από τα μετάγραφά του, εάν* $S_{[j+1]} = \gamma$*,*

3. *το* $(P\gamma, i, j + 1, \epsilon + 1)$ *είναι εμφάνιση του* $P\gamma$ *στην* S *και το* τ`R` *είναι ένα από τα μετάγραφά του, εάν* $S_{[j+1]} \neq \gamma$*.*

Απόδειξη. Αποδεικνύουμε την περίπτωση 1· οι περιπτώσεις 2 και 3 μπορούν να αποδειχτούν με παρόμοιο τρόπο. Το μετάγραφο τ`D` περιέχει τις ίδιες συνταχτικές λειτουργίες με το τ, και μια επιπλέον `D` λειτουργία. Οπότε, $c(\tau\texttt{D}) = c(\tau) + 1 = \epsilon + 1$. Καθώς το τ μετατρέπει το P σε $S_{[i,j]}$ και το τελευταίο `D` σε τ`D` απλώς διαγράφει το γ από το $P\gamma$, έπεται ότι το τ`D` μετατρέπει το $P\gamma$ σε $S_{[i,j]}$. Οπότε, το $(P\gamma, i, j, \epsilon + 1)$ είναι εμφάνιση του $P\gamma$ στην S και το τ`D` είναι ένα από τα μετάγραφά του. □

Τέλος, έστω δυο ακολουθίες πρότυπα, P και θP, όπου η τελευταία δημιουργείται με την προσάρτηση ενός συμβόλου $\theta \in \Sigma$ στην αρχή της πρώτης. Αυτό αντιπροσωπεύει την περίπτωση όπου περιοχές μοιράζονται το ίδιο επίθεμα και διαφέρουν κατά ένα σύμβολο, π.χ. οι περιοχές $R1$ και $R3$ στο Σχήμα 3.3. Υποθέτοντας ότι το P συμβαίνει στην τοποθεσία $[i, j]$ στην S, το επόμενο λήμμα δείχνει πώς αυτή η εμφάνιση σχετίζεται με εμφανίσεις του θP στις τοποθεσίες $[i, j]$ και $[i - 1, j]$. Όπως προηγουμένως, μια συνταχτική λειτουργία (`D`, `R`, ή `M`) μπορεί να προσαρτηθεί στο μετάγραφο για να βρει χώρο για το επιπλέον σύμβολο θ της ακολουθίας πρότυπο.

Λήμμα 3.4. *Εάν το* (P, i, j, ϵ) *είναι εμφάνιση του* P *στην ακολουθία δεδομένων* S *και το* τ *είναι ένα από τα μετάγραφά του, τότε:*

1. *το* $(\theta P, i, j, \epsilon + 1)$ *είναι εμφάνιση του* θP *στην* S *και το* `D`τ *είναι ένα από τα μετάγραφά του,*

2. *το* $(\theta P, i - 1, j, \epsilon)$ *είναι εμφάνιση του* θP *στην* S *και το* `M`τ *είναι ένα από τα μετάγραφά του, εάν* $S_{[i-1]} = \theta$*,*

3. *το* $(\theta P, i - 1, j, \epsilon + 1)$ *είναι εμφάνιση του* θP *στην* S *και το* `R`τ *είναι ένα από τα μετάγραφά του, εάν* $S_{[i-1]} \neq \theta$*.*

Απόδειξη. Παρόμοια με εκείνη του Λήμματος 3.3. □

3.1.3.2 Επεκτάσεις Προθέματος και Επιθέματος

Βάσει των λημμάτων της προηγούμενης ενότητας, δείχνουμε στη συνέχεια ότι είναι δυνατό να κατασκευάσουμε το σύνολο των ελάχιστων εμφανίσεων μιας περιοχής από εκείνες μιας μικρότερης. Έστω μια ακολουθία δεδομένων S, ένα πρότυπο P, και ένα κόστος μεταγράφου k. Έστω ότι $\mathcal{O}$ είναι το σύνολο των ελάχιστων εμφανίσεων του P στην S με κόστος μεταγράφου που δεν ξεπερνά το k. Στη συνέχεια, περιγράφουμε ένα σύνολο κανόνων επέκτασης που, όταν εφαρμόζονται στο $\mathcal{O}$, παράγουν τις ελάχιστες εμφανίσεις των προτύπων $P\gamma$ και θP, όπου $\gamma, \theta \in \Sigma$. Περιγράφουμε δυο σύνολα κανόνων επέκτασης: κανόνες επέκτασης προθέματος και επιθέματος.

Πρώτα παρουσιάζουμε τους κανόνες επέκτασης προθέματος. Θεωρούμε την περίπτωση του προτύπου $P\gamma$, που έχει το P ως πρόθεμα.

Ορισμός 3.2 (Επέκταση Προθέματος)**.** *Η επέκταση προθέματος του $\mathcal{O}$ με το σύμβολο $\gamma \in \Sigma$, που συμβολίζεται ως $\mathcal{O}^\gamma$, είναι ένα σύνολο εμφανίσεων του προτύπου $P\gamma$ στην S, με κόστος που δεν ξεπερνά το k, και προκύπτει σύμφωνα με τους ακόλουθους κανόνες επέκτασης.*

Για κάθε $(P, i, j, \epsilon) \in \mathcal{O}$:

1. *Εάν $\epsilon + 1 \leq k$, εισήγαγε στο $\mathcal{O}^\gamma$ τις εμφανίσεις $(P\gamma, i, j + x, \epsilon + x + 1)$ για κάθε $0 \leq x \leq k - \epsilon - 1$.*

2. *Εάν $\epsilon \leq k$ και $S_{[j+1]} = \gamma$, εισήγαγε στο $\mathcal{O}^\gamma$ τις εμφανίσεις $(P\gamma, i, j + x + 1, \epsilon + x)$ για κάθε $0 \leq x \leq k - \epsilon$.*

3. *Εάν $\epsilon + 1 \leq k$ και $S_{[j+1]} \neq \gamma$, εισήγαγε στο $\mathcal{O}^\gamma$ τις εμφανίσεις $(P\gamma, i, j + x + 1, \epsilon + x + 1)$ για κάθε $0 \leq x \leq k - \epsilon - 1$.*

Κατά την εισαγωγή μιας εμφάνισης, εάν κάποια άλλη στο $\mathcal{O}^\gamma$ συμβαίνει στην ίδια τοποθεσία, κρατάμε εκείνη με το μικρότερο κόστος μεταγράφου. □

Διαισθητικά, αυτοί οι κανόνες εφαρμόζουν τις περιπτώσεις 1, 2, ή 3 ου Λήμματος 3.3, αντίστοιχα, στην εμφάνιση (P, i, j, ϵ) και, μετά, εφαρμόζουν το Λήμμα 3.1 επαναλαμβανόμενα (μια φορά ανά τιμή x έτσι ώστε να μην ξεπερνάται το κατώφλι λάθους k) σε κάθε μια επιστρεφόμενη εμφάνιση του $P\gamma$.

Σαν παράδειγμα, θεωρούμε την ακολουθία δεδομένων S του Σχήματος 3.3 και έστω ότι $P = \texttt{GCCA}$, $\gamma = \texttt{T}$. Το $(P, 13, 16, 0)$ είναι μια ελάχιστη εμφάνιση του P στην S με κόστος που δεν ξεπερνά το $k = 1$. Ο κανόνας επέκτασης 1 στην εμφάνιση $(P, 13, 16, 0)$ παράγει το $(P\gamma, 13, 16, 1)$, ενώ ο κανόνας 2 παράγει το $(P\gamma, 13, 17, 0)$ και το $(P\gamma, 13, 18, 1)$. Σημειώστε ότι ο κανόνας επέκτασης 3 δεν εφαρμόζεται, καθώς $S_{[17]} = \gamma = \texttt{T}$.

Το επόμενο θεώρημα δείχνει ότι οι κανόνες επέκτασης προθέματος είναι *ορθοί*, δηλαδή, παράγουν εμφανίσεις του $P\gamma$ που είναι ελάχιστες, και *πλήρεις*, δηλαδή, παράγουν όλες τις ελάχιστες εμφανίσεις του $P\gamma$.

Θεώρημα 3.1. *Εάν το $\mathcal{O}$ είναι το σύνολο όλων των ελάχιστων εμφανίσεων του P στην S με κόστος μεταγράφου που δεν ξεπερνά το k, τότε η επέκταση προθέματός του $\mathcal{O}^\gamma$ είναι το σύνολο όλων των ελάχιστων εμφανίσεων του $P\gamma$ στην S με κόστος μεταγράφου που δεν ξεπερνά το k.*

Απόδειξη. Έστω ότι το $\mathcal{O}'$ είναι το σύνολο όλων των ελάχιστων εμφανίσεων του $P\gamma$ στην S με κόστος μεταγράφου που δεν ξεπερνά το k. Αποδεικνύουμε πρώτα ότι $\mathcal{O}' \subseteq \mathcal{O}^\gamma$ μέσω απαγωγής σε άτοπον.

Υποθέτουμε ότι υπάρχει μια ελάχιστη εμφάνιση $(P\gamma, i, j, \epsilon)$ του $P\gamma$ με $\epsilon \leq k$ που δεν εμφανίζεται στο $\mathcal{O}^\gamma$. Έστω ότι το τ είναι ένα οποιοδήποτε από τα συντακτικά μετάγραφα αυτής της εμφάνισης. Υπάρχουν τέσσερις περιπτώσεις βάσει της τελευταίας λειτουργίας στο τ. Έστω ότι το τ' υποδηλώνει το μετάγραφο που προκύπτει παραλείποντας αυτή την τελευταία λειτουργία.

Υποθέτουμε ότι η τελευταία λειτουργία είναι D, δηλαδή, $\tau = \tau'\texttt{D}$. Είναι εύκολο να δούμε ότι $c(\tau') = \epsilon - 1$ και ότι το $(P, i, j, \epsilon - 1)$ είναι εμφάνιση του P. Δείχνουμε ότι αυτή η εμφάνιση είναι ελάχιστη.

Έστω ότι υποθέτουμε το αντίθετο. Τότε, υπάρχει ένα μετάγραφο $\tau*$ που αντιστοιχεί σε μια εμφάνιση $(P, i, j, c(\tau^*))$, όπου $c(\tau^*) < c(\tau') = \epsilon - 1$. Σύμφωνα με την περίπτωση 1 του Λήμματος 3.3, το $(P\gamma, i, j, c(\tau^*) + 1)$ είναι εμφάνιση του $P\gamma$ με

μετάγραφο τ^*D και κόστος $c(\tau^*) + 1 < \epsilon$. Ωστόσο, αυτό δεν είναι δυνατό, καθώς η ελάχιστη εμφάνιση του $P\gamma$ στην τοποθεσία $[i, j]$ έχει κόστος ϵ, όπως υποθέσαμε. Συνεπώς, το $(P, i, j, \epsilon - 1)$ είναι ελάχιστη εμφάνιση του P.

Κατά συνέπεια, σύμφωνα με τον πρώτο κανόνα (για $x = 0$), το $(P, i, j, \epsilon - 1)$ επεκτείνεται σε εμφάνιση $(P\gamma, i, j, \epsilon)$. Όμως, η τελευταία υποτέθηκε ότι δεν εμφανίζεται στο $\mathcal{O}^\gamma$, κάτι που αποτελεί αντίφαση (άτοπο).

Χρησιμοποιώντας παρόμοιους ισχυρισμούς, μπορεί κάποιος να δείξει ότι αυτή η αντίφαση εμφανίζεται για τις περιπτώσεις όπου η τελευταία λειτουργία είναι R και M, χρησιμοποιώντας τις περιπτώσεις 3 και 2 του Λήμματος 3.3, αντίστοιχα.

Μένει να δείξουμε μια αντίφαση για την τελευταία περίπτωση, όπου η τελευταία λειτουργία στο μετάγραφο τ είναι I. Έστω ότι το y είναι ο μεγαλύτερος ακέραιος έτσι ώστε οι τελευταίες λειτουργίες y στο τ να είναι όλες εισαγωγές. Επιπλέον, έστω ότι το τ_1 είναι το μετάγραφο που αποκτάται από το τ παραλείποντας εκείνες τις τελευταίες λειτουργίες y. Παρατηρούμε ότι το $(P\gamma, i, j - y, c(\tau_1))$ πρέπει να είναι ελάχιστη εμφάνιση του $P\gamma$ στην τοποθεσία $[i, j-y]$ με κόστος μεταγράφου $c(\tau_1) = \epsilon - y$. Διαφορετικά, μπορεί κανείς να κατασκευάσει μια εμφάνιση του $P\gamma$ στην τοποθεσία $[i, j]$ με κόστος χαμηλότερο από το ελάχιστο ϵ.

Ανάλογα με την τελευταία λειτουργία στο τ_1 (που δεν μπορεί να είναι I), μπορεί κάποιος να κατασκευάσει μια ελάχιστη εμφάνιση του P από το $(P\gamma, i, j - y, c(\tau_1))$ με τρόπο παρόμοιο με τις τρεις περιπτώσεις που εξετάσαμε ανωτέρω. Έπειτα, εφαρμόζοντας την αντίστοιχη ρύθμιση κανόνα επέκτασης $x = y$, προκύπτει ότι το $(P\gamma, i, j, \epsilon)$ εμφανίζεται στο $\mathcal{O}^\gamma$, δηλαδή, αυτό είναι αντίφαση. Οπότε, έχουμε $\mathcal{O}' \subseteq \mathcal{O}^\gamma$.

Τέλος, δείχνουμε ότι $\mathcal{O}' \supseteq \mathcal{O}^\gamma$. Ας υποθέσουμε διαφορετικά, δηλαδή, ότι υπάρχει μια εμφάνιση $(P\gamma, i, j, \epsilon)$ οφ $\mathcal{O}^\gamma$ που δε βρίσκεται στο $\mathcal{O}'$. Καθώς αυτή η εμφάνιση δεν είναι ελάχιστη, πρέπει να υπάρχει μια άλλη, μια $(P\gamma, i, j, \epsilon') \in \mathcal{O}'$, με $\epsilon' < \epsilon$. Έχουμε αποδείξει ήδη ότι $\mathcal{O}' \subseteq \mathcal{O}^\gamma$, κάτι που συνεπάγεται ότι το $(P\gamma, i, j, \epsilon')$ βρίσκεται επίσης στο $\mathcal{O}^\gamma$. Ως αποτέλεσμα, τόσο το $(P\gamma, i, j, \epsilon)$ όσο και το $(P\gamma, i, j, \epsilon')$ βρίσκονται στο $\mathcal{O}^\gamma$. Αυτό αποτελεί αντίφαση επειδή οι κανόνες επέκτασης ορίζουν ότι μόνο η εμφάνιση με το μικρότερο κόστος μεταξύ εκείνων που εμφανίζονται στην ίδια τοποθεσία επιτρέπεται στο $\mathcal{O}^\gamma$. □

Τέλος, θεωρούμε την περίπτωση του προτύπου θP, που έχει επίθεμα P.

Ορισμός 3.3 (Επέκταση Επιθέματος)**.** *Η επέκταση επιθέματος του $\mathcal{O}$ με σύμβολο $\theta \in \Sigma$, που συμβολίζεται ως ${}^\theta\mathcal{O}$, περιέχει ένα σύνολο εμφανίσεων του προτύπου θP στην S με κόστος που δεν ξεπερνά το k, και προκύπτει σύμφωνα με τους ακόλουθους κανόνες επέκτασης.*

Για κάθε $(P, i, j, \epsilon) \in \mathcal{O}$:

1. *Εάν το $\epsilon + 1 \leq k$, εισήγαγε στο ${}^\theta\mathcal{O}$ τις εμφανίσεις $(\theta P, i - x, j, \epsilon + x + 1)$ για κάθε $0 \leq x \leq k - \epsilon - 1$.*

2. *Εάν τα $\epsilon \leq k$ και $S_{[i-1]} = \theta$, εισήγαγε στο ${}^\theta\mathcal{O}$ τις εμφανίσεις $(\theta P, i-x-1, j, \epsilon+x)$ για κάθε $0 \leq x \leq k - \epsilon$.*

3. *Εάν τα $\epsilon + 1 \leq k$ και $S_{[i-1]} \neq \theta$, εισήγαγε στο ${}^\theta\mathcal{O}$ τις εμφανίσεις $(\theta P, i - x - 1, j, \epsilon + x + 1)$ για κάθε $0 \leq x \leq k - \epsilon - 1$.*

Κατά την εισαγωγή εμφανίσεων, εάν κάποια άλλη στο ${}^\theta\mathcal{O}$ εμφανιστεί στην ίδια τοποθεσία, κρατάμε εκείνη με το χαμηλότερο κόστος μεταγράφου. □

Διαισθητικά, αυτοί οι κανόνες εφαρμόζουν τις περιπτώσεις 1, 2, ή 3 του Λήμματος 3.4, αντίστοιχα, στην εμφάνιση (P, i, j, ϵ) και, έπειτα, εφαρμόζουν το Λήμμα 3.2 επαναλαμβανόμενα (μια φορά ανά τιμή x έτσι ώστε να μην ξεπερνάται το κατώφλι λάθους k) σε κάθε μια δημιουγηθείσα εμφάνιση του θP. Το επόμενο θεώρημα δείχνει την ορθότητα και την πληρότητα των κανόνων επέκτασης επιθέματος.

Θεώρημα 3.2. *Εάν το $\mathcal{O}$ είναι το σύνολο όλων των ελάχιστων εμφανίσεων του P στην S με κόστος μεταγράφου που δεν ξεπερνά το k, τότε η επέκταση επιθέματός του ${}^{\theta}\mathcal{O}$ είναι το σύνολο όλων των ελάχιστων εμφανίσεων του θP στην S με κόστος μεταγράφου που δεν ξεπερνά το k.*

Απόδειξη. Παρόμοια με εκείνη του Θεωρήματος 3.1. □

Να σημειωθεί ότι συνεχείς εφαρμογές των κανόνων επέκτασης προθέματος και επιθέματος μπορούν να παράγουν τις ελάχιστες εμφανίσεις ενός προτύπου από τις ελάχιστες εμφανίσεις μιας από τις υπακολουθίες του, όπως φαίνεται στο ακόλουθο θεώρημα.

Θεώρημα 3.3. *Εάν δίνονται δυο πρότυπα P, P' έτσι ώστε $P \sqsubset P'$, και το σύνολο $\mathcal{O}$ όλων των ελάχιστων εμφανίσεων του P στην S με κόστος μεταγράφου που δεν ξεπερνά το k, είναι δυνατό να κατασκευάσουμε το σύνολο όλων των ελάχιστων εμφανίσεων του P' στην S με κόστος μεταγράφου που δεν ξεπερνά το k', για κάθε $k' \leq k$.*

Απόδειξη. Παρατηρούμε ότι για κάθε πρότυπο το σύνολο των εμφανίσεών του με κόστος που δεν ξεπερνά το k είναι ένα υπερσύνολο του συνόλου των εμφανίσεών του με κόστος που δεν ξεπερνά το k', όπου $k' \leq k$. Οπότε, χρειάζεται μόνο να αποδείξουμε ότι το σύνολο των ελάχιστων εμφανίσεων του P' με κόστος μεταγράφου που δεν ξεπερνά το k μπορεί να αποκτηθεί από το $\mathcal{O}$.

Καθώς ισχύει ότι $P \sqsubset P'$, υπάρχει μια ακολουθία προτύπων $P_1, \ldots, P_n$, όπως τα $P_1 = P$, $P_n = P'$, και είτε ισχύει το $P_{i+1} = P_i\gamma$ ή το $P_{i+1} = \theta P_i$, όπου $\gamma, \theta \in \Sigma$. Λόγω των Θεωρημάτων 3.1 και 3.2, μια εφαρμογή των κατά προσέγγιση (πρόθεμα εάν $P_{i+1} = P_i\gamma$, διαφορετικά επίθεμα) κανόνων επέκτασης στο σύνολο των ελάχιστων εμφανίσεων του P_i κατασκευάζει το σύνολο όλων των ελάχιστων εμφανίσεων του P_{i+1}. Μετά από διαδοχικές εφαρμογές, μπορεί να κατασκευαστεί το ζητούμενο σύνολο. □

3.1.4 Η μέθοδος PS-ARSM

Η μέθοδος *Prefix-Suffix ARSM* (`PS-ARSM`) εκμεταλλεύεται τις επικαλύψεις μεταξύ των περιοχών και εφαρμόζει τους κανόνες επέκτασης του Κεφαλαίου 3.1.3.2 για να παράγει αποδοτικά όλα τα αποτελέσματα ARSM. Η κεντρική ιδέα αφορά αρχικά τον καθορισμό των ελάχιστων εμφανίσεων για την πιο μικρή δυνατή περιοχή, τον πυρήνα, και έπειτα την σταδιακή επέκτασή τους για την κατασκευή των ελάχιστων εμφανίσεων για όλες τις περιοχές. Απαιτείται ειδική προσοχή ώστε οι παραγόμενες εμφανίσεις να υπακούν στις δυο προϋποθέσεις που τίθενται στον Ορισμό 3.1. Σημειώστε ότι καθώς όλες οι εμφανίσεις που παράγονται από τη μέθοδό μας είναι ελάχιστες, παραλείπουμε το χαρακτηρισμό ελάχιστες στο υπόλοιπο της παρούσας εργασίας.

Αρχικά, εισάγουμε κάποιες σημαντικές έννοιες στο Κεφάλαιο 3.1.4.1. Έπειτα, περιγράφουμε τον αλγόριθμο `PS-ARSM` στο Κεφάλαιο 3.1.4.2, και επεξηγούμε λεπτομερώς την εκτέλεσή του στο Κεφάλαιο 3.1.4.3. Τέλος, παρουσιάζουμε μια ανάλυση κόστους του `PS-ARSM` και προτείνουμε μια βελτιστοποίηση στο Κεφάλαιο 3.1.5. Για λόγους επίδειξης, περιλαμβάνουμε τα πιο κοινά σύμβολα και τους ορισμούς τους στον Πίνακα 3.1.

Πίνακας 3.1: Κοινοί συμβολισμοί στο ARSM.

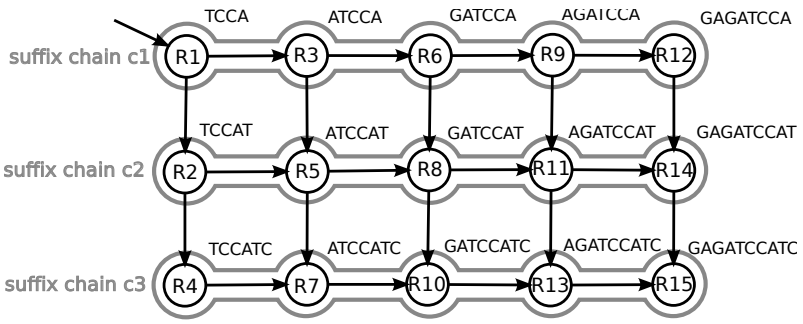
Σύμβολο	**Ορισμός**
Σ	Αλφάβητο
S	Ακολουθία δεδομένων
P	Πρότυπο
R	Μια περιοχή του P
C	Πυρήνας του P
a	Αρχική τοποθεσία του C στο P
b	Τελική τοποθεσία του C στο P
$\mathcal{K}()$	Συνάρτηση κατωφλίου
c	Μια αλυσίδα επιθεμάτων στο δικτυωτό (μερικώς διατεταγμένο σύνολο περιοχών - lattice)
sc_num	Αριθμός αλυσίδων επιθεμάτων
sc_length	Μήκος των αλυσίδων επιθεμάτων

3.1.4.1 Δικτυωτό (Lattice)

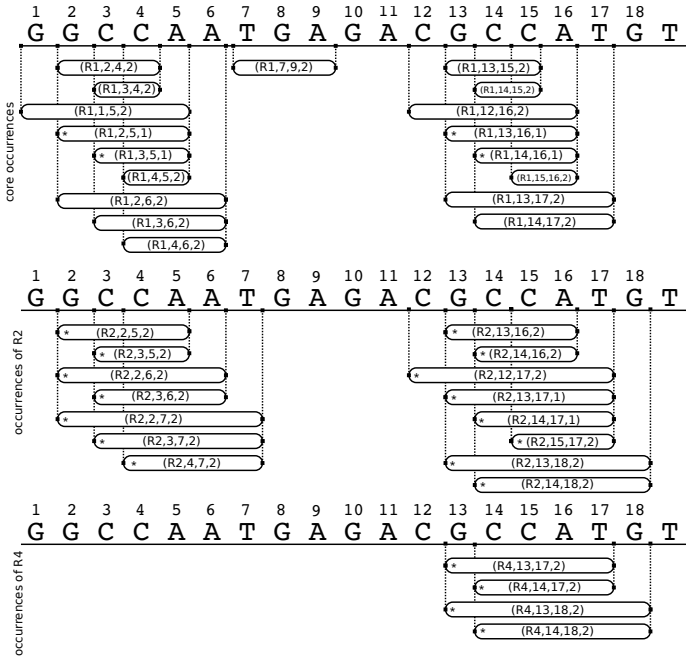
Το PS-ARSM εργάζεται πάνω στο *δικτυωτό (lattice)* που προκύπτει από τη σχέση υπακολουθίας $\sqsubset$ μεταξύ των περιοχών του προτύπου. Το Σχήμα 3.4 παρουσιάζει το δικτυωτό για το παράδειγμα του Σχήματος 3.3. Η περιοχή $R1$ πάνω αριστερά αντιστοιχεί στον πυρήνα, ενώ η περιοχή $R15$ κάτω δεξιά στο πρότυπο. Ένα οριζόντιο (αντίστ. κατακόρυφο) βέλος από μια περιοχή σε μια άλλη υπονοεί ότι η πρώτη είναι επίθεμα (αντίστ. πρόθεμα) της δεύτερης.

Μια *κεφαλή* είναι μια περιοχή τέτοια ώστε κανένα από τα επιθέματά της, εκτός από την ίδια, δεν είναι περιοχή. *Ουρά* είναι μια περιοχή τέτοια που δεν είναι το επίθεμα καμιάς άλλης περιοχής. Στο Σχήμα 3.4, υπάρχουν τρεις κεφαλές, το $R1$, το $R2$ και το $R4$, και τρεις ουρές, το $R12$, το $R14$ και το $R15$. Οι κεφαλές και οι ουρές ενός δικτυωτού διατάσσονται πλήρως (totally ordered) βάσει της σχέσης $\sqsubset$. Η πρώτη κεφαλή είναι ο πυρήνας, ενώ η τελευταία ουρά είναι το πρότυπο.

Μια *αλυσίδα επιθεμάτων* είναι ένα πλήρως διατεταγμένο σύνολο περιοχών που περιέχει μια ουρά και όλα τα επιθέματά της. Οι αλυσίδες επιθεμάτων ενός δικτυωτού διατάσσονται σύμφωνα με την διάταξη της ουράς τους. Υπάρχουν τρεις αλυσίδες επιθεμάτων στο Σχήμα 3.4, που ονομάζονται $c1$, $c2$, $c3$, και κάθε μια αντιστοιχεί σε μια σειρά του δικτυωτού. Παρατηρούμε ότι η πιο μικρή περιοχή σε μια αλυσίδα επιθεμάτων είναι μια κεφαλή και η μεγαλύτερη είναι μια ουρά. Π.χ. στην αλυσίδα $c1 = \{R1, R3, R6, R9, R12\}$, τα $R1$ και $R12$ είναι η κεφαλή και η ουρά της, αντίστοιχα.



Σχήμα 3.4: Το δικτυωτό των περιοχών για το παράδειγμα προβλήματος ARSM που χρησιμοποιούμε.



Σχήμα 3.5: Εμφανίσεις του πυρήνα, εμφανίσεις κεφαλών και σπόροι για τις αλυσίδες επιθεμάτων c_1, c_2, και c_3.

3.1.4.2 Περιγραφή του Αλγορίθμου

Το PS-ARSM αποτελείται από τρεις φάσεις εκτέλεσης. Στη συνέχεια περιγράφουμε αυτές τις φάσεις λεπτομερώς.

Φάση 1. Σε αυτή τη φάση, το PS-ARSM καθορίζει τις εμφανίσεις του πυρήνα με κόστος μεταγράφου που δεν ξεπερνά το $\mathcal{K}(|P|)$. Να σημειωθεί ότι το κατώφλι κόστους $\mathcal{K}(|P|)$ επιλέγεται έτσι ώστε κανένα αποτέλεσμα του ARSM, που παράγεται από επεκτάσεις των εμφανίσεων του πυρήνα, να μη χάνεται, όπως εξηγείται παρακάτω.

Υπενθυμίζουμε ότι μια εμφάνιση μιας περιοχής R μπορεί να είναι αποτέλεσμα ARSM μόνο εάν το κόστος της δεν ξεπερνά το $\mathcal{K}(|R|)$ (πρώτη απαίτηση στον Ορισμό 3.1). Καθώς το πρότυπο P είναι η μεγαλύτερη περιοχή και η $\mathcal{K}$ αυξάνεται μονοτονικά, το $\mathcal{K}(|P|)$ είναι το υψηλότερο κόστος που επιτρέπεται να έχει κάθε αποτέλεσμα ARSM. Από το Θεώρημα 3.3, προκύπτει ότι κάθε αποτέλεσμα ARSM πρέπει να βρίσκεται μεταξύ των επεκτάσεων των εμφανίσεων του πυρήνα με το πιο χαλαρό κατώφλι κόστους που μπορεί να υπάρξει, δηλαδή, το $\mathcal{K}(|P|)$.

Για το παράδειγμα ARSM που φαίνεται στο Σχήμα 3.3 και το αντίστοιχο δικτυωτό στο Σχήμα 3.4, η πρώτη φάση του PS-ARSM υπολογίζει τις εμφανίσεις του πυρήνα, δηλαδή, εκείνες της περιοχής $R1 = \texttt{TCCA}$, με κόστος το πολύ $\mathcal{K}(|P|) = \mathcal{K}(10) = 2$. Όλες αυτές οι εμφανίσεις απεικονίζονται στην κορυφή του Σχήματος 3.5 σαν ωοειδή κουτιά που στοιχίζονται ως προς την ακολουθία δεδομένων S. Για παράδειγμα, το $(R1, 2, 4, 2)$ αντιστοιχεί σε μια εμφάνιση του πυρήνα $R1$ στην τοποθεσία $[2, 4]$ στην S.

Φάση 2. Σε αυτή τη φάση, το PS-ARSM εφαρμόζει πρώτα τους κανόνες επέκτασης προθέματος στις εμφανίσεις του πυρήνα για να παράγει τις εμφανίσεις όλων των κεφαλών. Π.χ. στο δικτυωτό του Σχήματος 3.4, το PS-ARSM επεκτείνει με πρόθεμα τις εμφανίσεις του πυρήνα για να κατασκευάσει τις εμφανίσεις της περιοχής $R2 = \texttt{TCCAT}$, όπως φαίνεται στο μέσο του Σχήματος 3.5. Οι εμφανίσεις κεφαλών που προκύπτουν επεκτείνονται έπειτα για την απόκτηση των εμφανίσεων του $R4 = \texttt{TCCATC}$, όπως φαίνεται στο κάτω μέρος του Σχήματος 3.5.

Έπειτα, το PS-ARSM φιλτράρει τις εμφανίσεις κεφαλών κάθε αλυσίδας επιθεμάτων για να παρέχει την κατάλληλη είσοδο στη φάση 3. Θεωρούμε μια αλυσίδα c, και έστω ότι το R_1^c και το R_n^c είναι η κεφαλή και η ουρά της, αντίστοιχα. Εν συντομία, στόχος της φάσης 3 είναι η παραγωγή των εμφανίσεων κάθε περιοχής στο c με επέκταση των εμφανίσεων κεφαλής του. Να σημειωθεί ότι όλες οι παραγόμενες εμφανίσεις έχουν κόστος το πολύ $\mathcal{K}(|P|)$, καθώς είναι επεκτάσεις των εμφανίσεων του πυρήνα. Ωστόσο, παρατηρούμε ότι το $\mathcal{K}(|R_n^c|)$, που δεν ξεπερνά το $\mathcal{K}(|P|)$, είναι το υψηλότερο κόστος που επιτρέπεται να έχει κάθε εμφάνιση μιας περιοχής στο c (βλ. Ορισμό 3.1). Οπότε, από το Θεώρημα 3.3, προκύπτει ότι μόνο οι εμφανίσεις κεφαλών με κόστος που δεν ξεπερνά το $\mathcal{K}(|R_n^c|)$ θα πρέπει να επεκταθούν με επίθεμα στη φάση 3. Αναφερόμαστε σε αυτές τις εμφανίσεις του R_1^c ως τους *σπόρους* της αλυσίδας επιθεμάτων c. Για παράδειγμα, οι εμφανίσεις κεφαλών στη Σχήμα 3.5 που είναι επίσης σπόροι, μαρκάρονται με αστερίσκο, π.χ. το $(R2, 2, 5, 2)$ είναι σπόρος της αλυσίδας επιθεμάτων $c2$. Έτσι, η φάση 2 απομακρύνει τις εμφανίσεις κεφαλών που δεν είναι σπόροι για να παρέχει την είσοδο της φάσης 3.

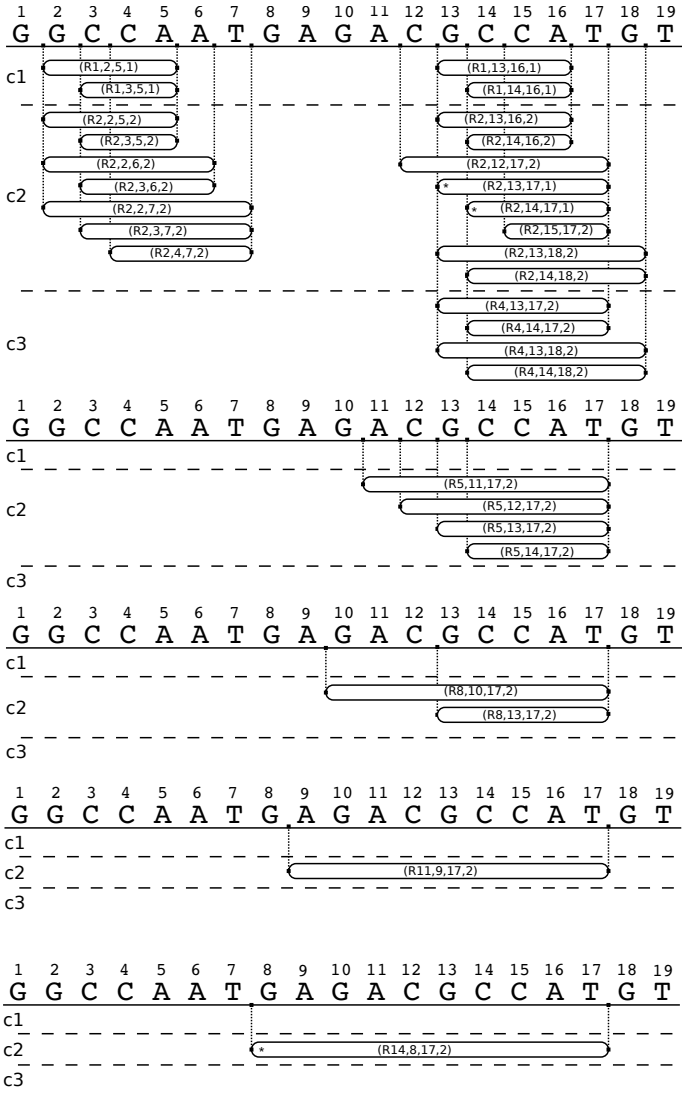
Φάση 3. Σε αυτή τη φάση, το PS-ARSM παράγει τα αποτελέσματα ARSM. Ο αλγόριθμος λειτουργεί ολιστικά σε όλες τις αλυσίδες, αλλά για λόγους κατανόησης περιγράφουμε μόνο τη διαδικασία για μια αλυσίδα c. Για κάθε περιοχή κατά μήκος της αλυσίδας, ξεκινώντας από την κεφαλή R_1^c και τελειώνοντας στην ουρά R_n^c, το PS-ARSM εκτελεί τις ακόλουθες εργασίες.

Ας υποθέσουμε ότι το R_i^c είναι η τρέχουσα περιοχή. Το PS-ARSM πρώτα επεκτείνει με επίθεμα τις εμφανίσεις της προηγούμενης περιοχής R_{i-1}^c στην αλυσίδα για να παράγει τις εμφανίσεις του R_i^c. Αυτές οι επεκτεταμένες εμφανίσεις λέγονται *υποψήφιες* και οι υποψήφιες του R_1^c είναι οι σπόροι. Τότε, το PS-ARSM επιβάλλει τις δυο απαιτήσεις του Ορισμού 3.1. Για την πρώτη, αποκλείει υποψήφιες με κόστος πάνω από $\mathcal{K}(|R_i^c|)$. Οι υπόλοιπες εισάγονται στο σύνολο του αποτελέσματος. Ωστόσο, κατά την εισαγωγή, το PS-ARSM απομακρύνει κάθε εμφάνιση (είτε μια υποψήφια, είτε μια που υπάρχει ήδη στο σύνολο του αποτελέσματος) που παραβιάζει τη δεύτερη απαίτηση. Είναι σημαντικό να σημειωθεί ότι όλες οι υποψήφιες για το R_i^c (δηλαδή, όχι μόνο εκείνες που έχουν εισαχθεί στο σύνολο του αποτελέσματος) είναι απαραίτητες για την απόκτηση των υποψηφίων για την επόμενη περιοχή R_{i+1}^c στην αλυσίδα.

Το Σχήμα 3.6 απεικονίζει την τρίτη φάση για όλες τις αλυσίδες επιθεμάτων του δικτυωτού που φαίνεται στο Σχήμα 3.4. Θεωρούμε μια αλυσίδα επιθεμάτων c_2. Η κεφαλή της είναι το $R_1^{c2} = R2$ και η ουρά είναι το $R_n^{c2} = R14$. Οι σπόροι αυτής της αλυσίδας, δηλαδή, οι εμφανίσεις του $R2$ με κόστος το πολύ $\mathcal{K}(|R14|) = \mathcal{K}(9) = 2$, έχουν προσδιοριστεί στη δεύτερη φάση. Υποθέτουμε ότι η εξεταζόμενη περιοχή είναι η $R5$. Οι τέσσερις υποψήφιές της (βλ. το δεύτερο τμήμα του Σχήματος 3.6) παράγονται από την επέκταση επιθέματος των υποψηφίων του $R2$ (στο μέσο του πρώτου τμήματος του Σχήματος 3.6). Τότε, το PS-ARSM λαμβάνει υπόψη μόνο εκείνες τις υποψήφιες που έχουν κόστος 1 (καθώς $\mathcal{K}(|R5|) = \mathcal{K}(6) = 1$) για εισαγωγή στο σύνολο του αποτελέσματος. Στο παράδειγμά μας, αυτό είναι ένα κενό σύνολο.

Ψευδοκώδικας. Το Σχήμα 3.7 παρουσιάζει τον ψευδοκώδικα του PS-ARSM . Ο αλγόριθμος παίρνει ως είσοδο την ακολουθία δεδομένων S, το πρότυπο P, τον πυρήνα $[a, b]$ και τη συνάρτηση κατωφλίου $\mathcal{K}$, και εξάγει τα αποτελέσματα ARSM. Στην πρώτη φάση, το PS-ARSM εφαρμόζει έναν αλγόριθμο ASM για τον υπολογισμό των εμφανίσεων του πυρήνα (γραμμή 1). Λεπτομέρειες δίνονται στο Κεφάλαιο 3.1.4.3.

Στη δεύτερη φάση (γραμμές 2–6), το PS-ARSM κατασκευάζει τις εμφανίσεις των κεφαλών. Οι εμφανίσεις της πρώτης κεφαλής είναι εκείνες του πυρήνα (γραμμή 2). Οι



Σχήμα 3.6: Σπόροι, υποψήφιες και χρονικά αποτελέσματα.

εμφανίσεις κάθε άλλης κεφαλής αποκτώνται από εκείνες της προηγούμενης κεφαλής (γραμμή 4). Η μέθοδος `pExp` υλοποιεί τους κανόνες επέκτασης προθέματος. Τότε, το `PS-ARSM` αναγνωρίζει τους σπόρους για την αντίστοιχη αλυσίδα επιθεμάτων, επικαλούμενο τη μέθοδο `getSeeds` (γραμμή 5). Και οι δυο μέθοδοι συζητώνται στο Κεφάλαιο 3.1.4.3.

Στην τρίτη φάση (γραμμές 7–14), το `PS-ARSM` υπολογίζει τα αποτελέσματα ARSM. Για λόγους κατανόησης, ο ψευδοκώδικας παρουσιάζει τα απαιτούμενα βήματα ανεξάρτητα για κάθε αλυσίδα επιθεμάτων c_j. Ωστόσο, αυτά τα βήματα εκτελούνται ολιστικά για τους σπόρους όλων των αλυσίδων επιθεμάτων (βλ. Κεφάλαιο 3.1.4.3 για λεπτομέρειες). Πίσω στον ψευδοκώδικα, το `PS-ARSM` κατασκευάζει πρώτα τις υποψήφιες, δηλαδή, τις εμφανίσεις κάθε περιοχής στην αλυσίδα. Να σημειωθείω ότι οι υποψήφιες της κεφαλής είναι οι σπόροι της αλυσίδας (γραμμή 8). Οι υποψήφιες για κάθε άλλη περιοχή αποκτώνται από εκείνες της προηγούμενης περιοχής (γραμμή 10). Πιο συγκεκριμένα, η μέθοδος `sExp` (που παρουσιάζεται στο Κεφάλαιο 3.1.4.3) εφαρμόζει τους κανόνες επέκτασης επιθέματος. Έπειτα, μεταξύ των υποψηφίων της περιοχής $R_i^{c_j}$, εκείνες με κόστος που δεν ξεπερνά το $\mathcal{K}(|R_i^{c_j}|)$ αναγνωρίζονται από τη μέθοδο `sieve` (γραμμή 11 και Κεφάλαιο 3.1.4.3). Τέλος, οι υπόλοιπες εμφανίσεις εισάγονται στο σύνολο του

PS-ARSM()
Input: S, P, $[a,b]$, $\mathcal{K}$
Output: *results*
begin
 # Phase 1
01. $core\_occs \leftarrow$ `ASMfull`$(S,\ P_{[a,b]},\ \mathcal{K}(|P|))$
 # Phase 2
02. $head\_occs_{c_1} \leftarrow core\_occs$
03. **foreach** suffix chain c_j
04. **if** $(j>1)$ **then** $head\_occs_{c_j} \leftarrow$ `pExp`$(head\_occs_{c_{j-1}},\ R_1^{c_j})$
05. $seeds_{c_j} \leftarrow$ `getSeeds`$(head\_occs_{c_j},\ \mathcal{K}(|c_j.tail|))$
06. **end**
 # Phase 3
07. **foreach** suffix chain c_j
08. $cands_1 \leftarrow seeds_{c_j}$
09. **foreach** region $R_i^{c_j}$ of c_j
10. **if** $(i>1)$ **then** $cands_i \leftarrow$ `sExp`$(cands_{i-1},\ R_i^{c_j})$
11. $tmp \leftarrow$ `sieve`$(cands_i,\ \mathcal{K}(|R_i^{c_j}|))$
12. `insert`$(results,\ tmp)$
13. **end**
14. **end**
end.

Σχήμα 3.7: Ο αλγόριθμος `PS-ARSM` .

αποτελέσματος, ενισχύοντας τη δεύτερη απαίτηση Ορισμού 3.1 (γραμμή 12).

Ορθότητα. Το `PS-ARSM` εφαρμόζει το Θεώρημα 3.3, δηλαδή, εκτελεί κανόνες προθέματος και επιθέματος, που είναι ορθοί και πλήρεις όπως έχουμε αποδείξει. Τα προηγούμενα εγγυώνται ότι το `PS-ARSM` θα επιστρέψει όλα τα αποτελέσματα ARSM.

3.1.4.3 Υλοποίηση

Μια σημαντική παρατήρηση για το `PS-ARSM` είναι ότι η επέκταση προθέματος (αντίστ. επιθέματος) ενός συνόλου εμφανίσεων που λήγουν (αντίστ. ξεκινούν) στην ίδια θέση στην ακολουθία δεδομένων, απαιτεί πολλούς ίδιους υπολογισμούς. Για να γίνει αυτό κατανοητό, θεωρούμε το Σχήμα 3.5 και παρατηρούμε τις εμφανίσεις πυρήνα $(R1, 1, 5, 2)$ και $(R1, 2, 5, 1)$, οι οποίες λήγουν και οι δύο στη θέση 5. Στη συνέχεια, περιγράφουμε τις απαραίτητες λειτουργίες για την επέκταση προθέματος αυτών των εμφανίσεων και την παραγωγή των εμφανίσεων κεφαλής της περιοχής $R2$. Πρώτον, παρατηρούμε ότι $R2 = R1\,\mathtt{T}$ και ότι το επόμενο σύμβολο στην ακολουθία δεδομένων, δηλαδή αυτό στη θέση 6, είναι το `A`. Έπειτα, εφαρμόζοντας το Λήμμα 3.3 για $P = R1$, $i \in \{1,2\}$ $j = 5$, $\epsilon \in \{2,1\}$, και καθώς $S[j+1] = \mathtt{A} \neq \mathtt{T} = \gamma$, προκύπτει ότι κάθε εμφάνιση του πυρήνα έχει ως αποτέλεσμα δυο εμφανίσεις κεφαλής: $(R2, 1, 5, 3)$, $(R2, 1, 6, 3)$ από την πρώτη και $(R2, 2, 5, 2)$, $(R2, 2, 6, 2)$ από τη δεύτερη. Παρατηρήστε ότι αυτές μπορούν να αναπαρασταθούν εν συντομία ως οι δυο εμφανίσεις κεφαλής $(R2, i, 5{+}\underline{0}, \epsilon{+}\underline{1})$ και $(R2, i, 5{+}\underline{1}, \epsilon{+}\underline{1})$, για $i \in \{1,2\}$, $\epsilon \in \{2,1\}$. Είναι σημαντικό να σημειώσουμε εδώ ότι οι πληροφορίες που έχουμε υπογραμμίσει στην προηγούμενη αναπαράσταση (δηλαδή, $+0$, $+1$ για την τελευταία θέση και $+1$, $+1$ για το κόστος) περιγράφουν πλήρως το πώς κατασκευάζονται οι εμφανίσεις κεφαλής από τις εμφανίσεις του πυρήνα.

Βάσει αυτής της παρατήρησης, το `PS-ARSM` επιστρατεύει δομές δεδομένων που εξυπηρετούν δύο σκοπούς: (1) να αναπαριστούν τις εμφανίσεις περιοχών με συμπαγή τρόπο και (2) να διευκολύνουν στις επεκτάσεις επιθέματος και προθέματος των εμφανίσεων περιοχών αποφεύγοντας πλεονάζοντες υπολογισμούς.

Πίνακας κατακερματισμού `head_occs` **.** Ο πίνακας κατακερματισμού `head_occs` χρησιμοποιείται για τον υπολογισμό και την αποθήκευση των εμφανίσεων κεφαλής κατά τη δεύτερη φάση του `PS-ARSM` . Αρχικά, περιέχει τις εμφανίσεις πυρήνα, δηλαδή, τις εμφανίσεις κεφαλής της πρώτης αλυσίδας επιθεμάτων. Στη συνέχεια, ανανεώνονται οι καταχωρήσεις του $sc\_num - 1$ φορές, όπου sc_num είναι ο αριθμός των αλυσίδων επιθεμάτων στο δικτυωτό των περιοχών. Μετά από την $(i-1)$-οστή ανανέωση, το `head_occs` περιέχει τις εμφανίσεις κεφαλής για την i-οστή αλυσίδα επιθεμάτων.

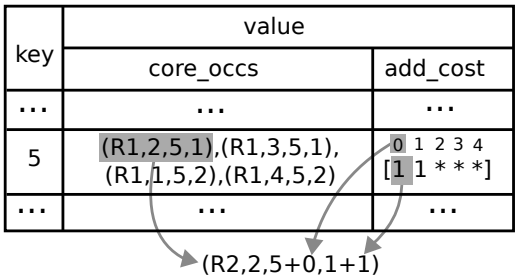
Στη συνέχεια, περιγράφουμε τα περιεχόμενα του `head_occs` για την αλυσίδα επιθεμάτων c_i. Η διαδικασία ανανέωσης εξηγείται αργότερα. Το `head_occs` συνίσταται από καταχωρήσεις κλειδιού-τιμής. Ένα *κλειδί* αντιστοιχεί στην θέση τερματισμού μιας εμφάνισης του πυρήνα στην ακολουθία δεδομένων. Η *τιμή* για το κλειδί j, που συμβολίζεται ως `head_occs[j]`, είναι μια σύνθετη τιμή που περιέχει πληροφορίες για τις εμφανίσεις της κεφαλής της c_i που παράγονται από την επέκταση επιθέματος των εμφανίσεων του πυρήνα που λήγουν στη θέση j. Πιο συγκεκριμένα, αυτή η σύνθετη τιμή συνίσταται από:

- `core_occs`: τη λίστα εμφανίσεων του πυρήνα που λήγουν στη θέση j, και ταξινομούνται ανάλογα με το κόστος μεταγράφου τους.
- `add_cost`: έναν πίνακα στον οποίο η x-οστή καταχώρηση υποδηλώνει το επιπλέον κόστος μεταγράφου που απαιτείται για την επέκταση κάθε εμφάνισης του πυρήνα που υπάρχει στο `core_occs` σε μία εμφάνιση κεφαλής του c_i που λήγει στη θέση $j + x$.

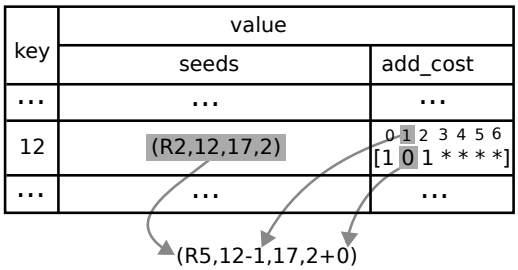
Να σημειωθεί ότι το μέγεθος του πίνακα `add_cost` είναι $sc\_num + \mathcal{K}(|P|)$, καθώς αυτή η τιμή είναι το περισσότερο που μπορεί να επεκταθεί μια εμφάνιση του πυρήνα προς τα δεξιά ενώ το κόστος μεταγράφου της θα εξακολουθεί να μην ξεπερνά το $\mathcal{K}(|P|)$.

Παρατηρούμε ότι το `head_occs` περιγράφει έμμεσα τις εμφανίσεις κεφαλής της αλυσίδας επιθεμάτων c_i. Δηλαδή, αποθηκεύει τις εμφανίσεις του πυρήνα (`core_occs`), και το πώς αυτές επεκτείνονται (`add_cost`) για να παραχθούν οι εμφανίσεις κεφαλής. Επιδεικνύουμε τα ανωτέρω χρησιμοποιώντας ένα παράδειγμα. Έστω ότι έχουμε το δικτυωτό του Σχήματος 3.4 και έχουμε το `head_occs` μετά την πρώτη ανανέωση, δηλαδή όταν περιλαμβάνει τις εμφανίσεις κεφαλής της αλυσίδας επιθεμάτων c_2. Το Σχήμα 3.8 απεικονίζει τα περιεχόμενα της καταχώρησης `head_occs[5]`. Το `core_occs` περιέχει όλες τις εμφανίσεις του πυρήνα που λήγουν στη θέση 5 στην ακολουθία δεδομένων S. Να σημειωθεί ότι ένας πίνακας `add_cost` αναπαρίσταται σε δύο σειρές: η χαμηλότερη δείχνει τα περιεχόμενα, ενώ η πιο πάνω παρουσιάζει τη δεικτοδότηση αυτού του πίνακα η οποία ξεκινά από το μηδέν. Ο εικονιζόμενος `add_cost` μας ενημερώνει ότι οι εμφανίσεις κεφαλής του c_2 (οι οποίες παράγονται από οποιαδήποτε εμφάνιση του πυρήνα στο `core_occs`) που λήγουν στις θέσεις $5+0 = 5$ και $5+1 = 6$, έχουν επιπλέον κόστος 1. Επιπλέον, κάθε εμφάνιση κεφαλής που λήγει σε επόμενες θέσεις, π.χ., η $5+2 = 7$, έχει επιπλέον κόστος μεγαλύτερο από $\mathcal{K}(|P|)$, που συμβολίζεται ως $*$ στο Σχήμα 3.8. Οπότε, οι εμφανίσεις κεφαλής στο `head_occs[5]` είναι $(R2, 2, 5, 2)$, $(R2, 2, 6, 2)$, $(R2, 3, 5, 2)$ και $(R2, 3, 6, 2)$ και είναι όλες εμφανίσεις κεφαλής του c_2 οι οποίες είναι επεκτάσεις εμφανίσεων πυρήνα που λήγουν στη θέση 5.

Πίνακας κατακερματισμού `cands` **.** Ο πίνακας κατακερματισμού `cands` χρησιμοποιείται για τον υπολογισμό και την αποθήκευση των υποψηφίων κατά την τρίτη φάση του `PS-ARSM` . Αρχικά, περιέχει τους σπόρους όλων των αλυσίδων επιθεμάτων, δηλαδή, τις υποψήφιες για όλες τις κεφαλές. Μετά, οι καταχωρήσεις του ανανεώνονται $sc\_length - 1$ φορές, όπου sc_length είναι το μήκος κάθε αλυσίδας επιθεμάτων



Σχήμα 3.8: Ένα απόσπασμα από τον πίνακα κατακερματισμού `head_occs` που περιέχει τις εμφανίσεις κεφαλής της αλυσίδας επιθεμάτων c_2.



Σχήμα 3.9: Ένα απόσπασμα από τον πίνακα κατακερματισμού `cands` , που περιέχει τις υποψήφιες για τις περιοχές $R3$, $R5$, και $R7$.

στο δικτυωτό της περιοχής. Μετά την $(i+1)$-οστή ανανέωση, ο `cands` περιέχει τις υποψήφιες για τις i-οστές περιοχές όλων των αλυσίδων επιθεμάτων.

Στη συνέχεια, περιγράφουμε τα περιεχόμενα του `cands` για τις i-οστές περιοχές όλων των αλυσίδων επιθεμάτων. Η διαδικασία ανανέωσης εξηγείται αργότερα. Ο `cands` αποτελείται από ζεύγη κλειδιών-τιμών. Ένα *κλειδί* αντιστοιχεί στην αρχική θέση ενός σπόρου στην ακολουθία δεδομένων. Η *τιμή* για το κλειδί j, που συμβολίζεται ως `cands[j]`, είναι μια σύνθετη τιμή με πληροφορίες για τις υποψήφιες που παράγονται από την επέκταση επιθέματος των σπόρων που ξεκινούν στη θέση j. Πιο συγκεκριμένα, αυτή η σύνθετη τιμή αποτελείται από:

- `seeds`: τη λίστα σπόρων που ξεκινούν στη θέση j και ταξινομούνται σύμφωνα με το κόστος μεταγράφου τους.

- `add_cost`: έναν πίνακα στον οποίο η x-οστή καταχώρηση δηλώνει το επιπλέον κόστος μεταγράφου που απαιτείται για την επέκταση κάθε σπόρου των `seeds` σε μια υποψήφια που ξεκινά στη θέση $j - x$.

Να σημειωθεί ότι το μήκος του πίνακα `add_cost` είναι $sc\_length + \mathcal{K}(|P|)$, καθώς αυτή η τιμή είναι όσο πιο μακριά μπορεί να επεκταθεί ένας σπόρος προς τα αριστερά ενώ το κόστος μεταγράφου του εξακολουθεί να μην ξεπερνά το $\mathcal{K}(|P|)$.

Παρόμοια με τον `head_occs` , ο `cands` περιγράφει έμμεσα τις υποψήφιες. Δηλαδή, αποθηκεύει τους σπόρους (`seeds`), και το πώς τους επεκτείνουμε (`add_cost`) ώστε να παράγουμε τις υποψήφιες. Επιδεικνύουμε το ανωτέρω με τη χρήση του δικτυωτού που υπάρχει ως παράδειγμα στο Σχήμα 3.4. Υποθέτουμε ότι ο `cands` μετά την πρώτη ανανέωση, περιέχει τις υποψήφιες για τις περιοχές $R3$, $R5$, και $R7$. Το Σχήμα 3.9 απεικονίζει την καταχώρηση `cands[12]`. Παρατηρούμε ότι ο σπόρος $(R2, 12, 17, 2)$ παράγει τη μόνη υποψήφια $(R5, 11, 17, 2)$, όπως εξηγείται στο σχήμα, εφόσον όλες οι άλλες εμφανίσεις ξεπερνούν το κατώφλι κόστους.

Στόχος της φάσης 1 είναι να παράγει όλες τις εμφανίσεις του πυρήνα με κόστος το πολύ $\mathcal{K}(|P|)$. Να σημειωθεί ότι ένας έτοιμος αλγόριθμος ASM (όπως εκείνοι που αναφέρονται στο Κεφάλαιο 2.1.1.3) *δεν μπορεί* να παράγει όλες τις εμφανίσεις του πυρήνα. Αυτό συμβαίνει επειδή, λόγω σχεδιασμού, όλοι οι αλγόριθμοι ASM αγνοούν κάποιες από τις επικαλυπτόμενες εμφανίσεις. Για παράδειγμα, εάν δυο εμφανίσεις έχουν την ίδια τελική θέση αλλά διαφορετικά κόστη μεταγράφου, αναφέρεται μόνο εκείνη με το χαμηλότερο κόστος. Για να το δείξουμε αυτό, έστω ότι έχουμε τον κλασικό αλγόριθμο δυναμικού προγραμματισμού που περιγράφεται στο [26]. Θεωρούμε την εμφάνιση του πυρήνα $(R1, 12, 16, 2)$ στο Σχήμα 3.5. Η συγκεκριμένη εμφάνιση δεν θα περιλαμβανόταν μεταξύ των αποτελεσμάτων, επειδή η $(R1, 13, 16, 1)$ και η $(R1, 14, 16, 1)$ λήγουν στην ίδια θέση και έχουν καλύτερο κόστος μεταγράφου. Ωστόσο, η $(R1, 12, 16, 2)$ δε θα έπρεπε να απορριφθεί, καθώς παράγει (μέσω επεκτάσεων στις δυο τελευταίες φάσεις) την $(R14, 8, 17, 2)$, η οποία είναι αποτέλεσμα ASM.

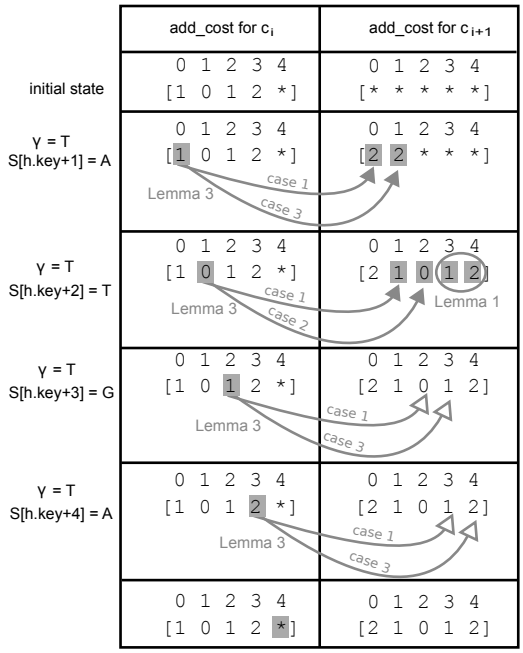
Προκειμένου να παράγουμε όλες τις εμφανίσεις του πυρήνα, ακολουθούμε τρία βήματα: (1) εκτελούμε ένα συμβατικό αλγόριθμο ASM για να ανακαλύψουμε τα καταληκτικά σημεία των εμφανίσεων, (2) κατασκευάζουμε ασύζευκτα παράθυρα γύρω από αυτά τα καταληκτικά σημεία, ώστε κάθε εμφάνιση να τοποθετείται πλήρως μέσα σε ένα παράθυρο, και (3) σε κάθε παράθυρο, εκτελούμε μια παραλλαγή του αλγορίθμου δυναμικού προγραμματισμού (που περιγράφεται παρακάτω) για να παράγουμε αποτελεσματικά όλες τις εμφανίσεις.

Να σημειωθεί ότι κάθε αλγόριθμος ASM, συμπεριλαμβανομένων των λύσεων που βασίζονται σε ευρετήρια, μπορεί να χρησιμοποιηθεί στο βήμα 1. Στη συνέχεια, στο βήμα 2, κατασκευάζουμε το παράθυρο $[i - |C| - \mathcal{K}(|P|) + 1, i]$ για κάθε καταληκτικό σημείο i που βρήκαμε στο προηγούμενο βήμα. Αυτό συμβαίνει επειδή είναι αδύνατο για μια εμφάνιση να ξεκινά πριν την $(i - |C| - \mathcal{K}(|P|) + 1)$-οστή θέση. Για την αποφυγή περιττών υπολογισμών, συγχωνεύουμε τα επικαλυπτόμενα παράθυρα.

Τέλος, στο βήμα 3, εκτελούμε μια παραλλαγή του αλγορίθμου δυναμικού προγραμματισμού του [96] που προσαρμόζεται ειδικά για να επιστρέφει όλες τις εμφανίσεις. Υπενθυμίζουμε ότι στους συμβατικούς αλγορίθμους, κάθε κελί δυναμικού προγραμματισμού $DP[i, j]$ περιέχει το κόστος και την εναρκτήρια θέση της καλύτερης εμφάνισης του προθέματος προτύπου $P_{[1,i]}$ που λήγει στη θέση j στην ακολουθία δεδομένων S. Από την άλλη μεριά, στη δική μας παραλλαγή, κρατάμε σε κάθε κελί $DP[i, j]$ έναν πίνακα που περιέχει τα κόστη των καλύτερων εμφανίσεων του $P_{[1,i]}$ που λήγει στη θέση j *για κάθε δυνατή εναρκτήρια θέση* στην S. Τέλος, καθώς οι συμβατικοί αλγόριθμοι δεν μπορούν να ανακτήσουν εμφανίσεις τα μετάγραφα των οποίων ξεκινούν με λειτουργίες Εισαγωγής, πρέπει να προσέξουμε ιδιαίτερα ώστε να μην τις χάσουμε.

Η δεύτερη φάση του `PS-ARSM` παράγει τους σπόρους όλων των αλυσίδων επιθεμάτων, επεκτείνοντας σταδιακά με πρόθεμα τις εμφανίσεις κεφαλής. Πρώτα, ο `head_occs` ξεκινάει με τις εμφανίσεις του πυρήνα που προέκυψαν από τη Φάση 1. Μετά, το `PS-ARSM` προχωράει σε $sc\_num - 1$ επαναλήψεις. Σε κάθε επανάληψη, το `PS-ARSM` εκτελεί δυο εργασίες: (1) επικαλείται το `pExp` για να ανανεώσει τον `head_occs` ώστε να περιέχει τις εμφανίσεις κεφαλής της επόμενης αλυσίδας επιθεμάτων, και (2) επικαλείται το `getSeeds` για να αναγνωρίσει τους σπόρους και να αρχικοποιήσει τον πίνακα κατακερματισμού `cands` , που απαιτείται για την Φάση 3.

Θα περιγράψουμε τώρα τη μέθοδο `pExp` . Έστω ότι ο `head_occs` περιέχει τις εμφανίσεις κεφαλής της αλυσίδας επιθεμάτων c_i. Έπειτα το `pExp` ανανεώνει τον `head_occs` (τροποποιώντας τους πίνακες `add_cost`) έτσι ώστε να περιέχει τις εμφανίσεις κεφαλής του c_{i+1}. Συγκεκριμένα, το `pExp` επισκέπτεται μια καταχώρηση του



Σχήμα 3.10: Η μέθοδος `pExp` ανανεώνει τον πίνακα `add_cost` της καταχώρησης h του `head_occs` .

πίνακα κατακερματισμού `head_occs` και εφαρμόζει τους κανόνες επέκτασης προθέματος για κάθε πιθανή καταληκτική θέση μιας εμφάνισης. Φτιάχνουμε μια καταχώρηση h του `head_occs` με κλειδί $h.key$, και έστω ότι γ είναι το τελευταίο σύμβολο στην κεφαλή (πρώτη περιοχή) της αλυσίδας επιθεμάτων c_{i+1}. Έστω ότι το `add_cost`(c_i) (αντίστ. το `add_cost`(c_{i+1})) δηλώνει τον πίνακα για την αλυσίδα επιθεμάτων c_i (αντίστ. c_{i+1}). Αρχικά το `add_cost`(c_{i+1}) συμπληρώνεται με τιμές $*$. Έπειτα, το `pExp` σκανάρει το `add_cost`(c_i) και εφαρμόζει την ακόλουθη διαδικασία για κάθε καταχώρηση μέχρι να συναντηθεί μια τιμή $*$. Θεωρούμε την j-οστή καταχώρηση στο `add_cost`(c_i). Υπενθυμίζουμε ότι αυτή αντιπροσωπεύει τις εμφανίσεις κεφαλής του c_i που λήγουν στη θέση $h.key + j$. Πρώτα, η διαδικασία ανανέωσης εφαρμόζει το Λήμμα 3.3 (περιπτώσεις 1, 2 εάν τα σύμβολα γ και $S_{[h.key+j]}$ ταιριάζουν, ή περιπτώσεις 1, 3 διαφορετικά) για να υπολογίσει τα επιπλέον κόστη των εμφανίσεων κεφαλής του c_{i+1} που λήγουν στις θέσεις $h.key + j$ και $h.key + j + 1$. Εάν οι αντίστοιχες καταχωρήσεις στο `add_cost`(c_{i+1}) έχουν υψηλότερα επιπλέον κόστη, ανανεώνονται. Τέλος, το `pExp` εφαρμόζει το Λήμμα 3.1 για κάθε καταχώρηση του `add_cost`(c_{i+1}). Όπως προηγουμένως, ανανεώνει μια καταχώρηση μόνο όταν το υπολογιζόμενο επιπλέον κόστος είναι μικρότερο από την υπάρχουσα τιμή της καταχώρησης.

Το Σχήμα 3.10 απεικονίζει μια εφαρμογή του `pExp` , υποθέτοντας ότι το μέγιστο επιτρεπτό κόστος είναι 2. Η αριστερή στήλη δείχνει τον πίνακα `add_cost` για την αλυσίδα επιθεμάτων c_i, ενώ η δεξιά στήλη δείχνει πώς ο πίνακας `add_cost` ανανεώνεται για την επόμενη αλυσίδα επιθεμάτων c_{i+1}. Η πρώτη σειρά δείχνει ότι το `add_cost`(c_{i+1}) έχει αρχικά όλες τις τιμές του ίσες με $*$. Έπειτα, το `pExp` εξετάζει την καταχώρηση `add_cost`$(c_i)[0]$ και εφαρμόζει τις περιπτώσεις 1 και 3 του Λήμματος 3.3 ($S_{[h.key+1]} =$ `A` $\neq$ `T` $= \gamma$). Αυτό υπονοεί ότι οι καταχωρήσεις 0 και 1 του `add_cost`(c_{i+1}) έχουν κόστος 1 περισσότερο από εκείνο στο `add_cost`$(c_i)[0]$. Να σημειωθεί ότι το Λήμμα 3.1 δεν ανανεώνει καμία καταχώρηση του `add_cost`(c_{i+1}) καθώς θα αποκτούσε επιπλέον κόστος μεγαλύτερο από το μέγιστο επιτρεπτό.

Μετά, το `pExp` εξετάζει την καταχώρηση `add_cost`$(c_i)[1]$ (βλ. την τρίτη σειρά του

Σχήματος 3.10). Αυτή τη φορά, εφαρμόζονται οι περιπτώσεις 1 και 2 του Λήμματος 3.3. Οπότε, η καταχώρηση 1 του `add_cost`(c_{i+1}) ανανεώνεται με επιπλέον κόστος $0+1=1$ (περίπτωση 1), καθώς είναι χαμηλότερο από την τρέχουσα τιμή της. Από την άλλη μεριά, η καταχώρηση 2 του `add_cost`(c_{i+1}) έχει επιπλέον κόστος $0+0=0$ (περίπτωση 2). Το Λήμμα 3.1 συμπληρώνει κάθε υπολειπόμενη καταχώρηση με επιπλέον κόστη 1 περισσότερο συγκριτικά με την προηγούμενη καταχώρηση.

Οι περιπτώσεις 1 και 3 του Λήμματος 3.3 εφαρμόζονται για το `add_cost`(c_i)[2]. Ωστόσο, καθώς παράγουν εμφανίσεις με κόστη χειρότερα από εκείνα που παράγονται στο προηγούμενο βήμα, το `add_cost`(c_{i+1}) δεν ανανεώνεται (βλ. την τέταρτη σειρά του Σχήματος 3.10). Η μέθοδος `pExp` συνεχίζει με τις επόμενες καταχωρήσεις και σταματάει όταν φτάσει το $*$ στην τελική καταχώρηση.

Τέλος, περιγράφουμε τη μέθοδο `getSeeds` . Στόχος της είναι η αναγνώριση των σπόρων ανάμεσα στις εμφανίσεις κεφαλής της τρέχουσας αλυσίδας επιθεμάτων. Χρησιμοποιώντας τις πληροφορίες στα πεδία $core\_occs$ και add_cost, το `getSeeds` επιλέγει εκείνες τις εμφανίσεις κεφαλής με κόστος που δεν ξεπερνά το επιτρεπόμενο για την τρέχουσα αλυσίδα επιθεμάτων. Οι επιλεγμένες εμφανίσεις είναι οι σπόροι, και εισάγονται στον πίνακα κατακερματισμού `cands` .

Η τρίτη φάση του `PS-ARSM` παράγει τα πραγματικά αποτελέσματα ARSM, επεκτείνοντας σταδιακά με επίθεμα τους σπόρους όλων των αλυσίδων επιθεμάτων. Να σημειωθεί ότι η Φάση 2 έχει αρχικοποιήσει τον πίνακα κατακερματισμού `cands` με όλους τους σπόρους. Τότε, το `PS-ARSM` προχωράει σε sc_length επαναλήψεις. Σε κάθε επανάληψη, το `PS-ARSM` εκτελεί δυο εργασίες: (1) επικαλείται το `sExp` (εκτός από την πρώτη επανάληψη) για να ανανεώσει το `cands` έτσι ώστε να περιέχει τις υποψήφιες της επόμενης περιοχής όλων των αλυσίδων επιθεμάτων, και (2) επικαλείται το `sieve` και το `insert` για να παράγει τα πραγματικά αποτελέσματα ARSM.

Η μέθοδος `sExp` είναι παρόμοια με το `pExp` , αλλά έχει τις ακόλουθες διαφορές: (1) λειτουργεί στον πίνακα κατακερματισμού `cands`, (2) εφαρμόζει τα Λήμματα 3.4 και 3.2 και (3) εξετάζει την ακολουθία δεδομένων προς τα πίσω.

Η μέθοδος `sieve` προσδιορίζει τα αποτελέσματα ARSM μεταξύ των τρεχουσών υποψηφίων στον πίνακα κατακερματισμού `cands` . Τέλος, το `insert` προσθέτει αυτές τις εμφανίσεις στο σύνολο των αποτελεσμάτων ARSM, φροντίζοντας να μην παραβιάζεται η δεύτερη απαίτηση του Ορισμού 3.1.

3.1.5 Ανάλυση Κόστους και Βελτιστοποίηση

Κόστος Φάσης 1. Η πρώτη φάση περιλαμβάνει τρία βήματα. Στα δυο πρώτα, χρησιμοποιείται ένας συμβατικός αλγόριθμος ASM για να μαρκάρει τα παράθυρα της ακολουθίας δεδομένων S τα οποία περιέχουν εμφανίσεις του πυρήνα με κόστος που δεν ξεπερνά το $\mathcal{K}(|P|)$. Υποθέτουμε ότι εφαρμόζεται ο αλγόριθμος δυναμικού προγραμματισμού με ευριστική αποκοπής [96]. Μετά, σύμφωνα με το [8], ο μέσος χρόνος επεξεργασίας των πρώτων δυο βημάτων είναι το πολύ $\left(\frac{\mathcal{K}(|P|)}{1-e/\sqrt{|\Sigma|}}+O(1)\right)\cdot|S|\cdot T_{DP}=O\left(\mathcal{K}(|P|)\cdot|S|\right)$, όπου το T_{DP} είναι ο απαιτούμενος χρόνος για τον υπολογισμό κάθε κελιού δυναμικού προγραμματισμού (σταθερά για μια δεδομένη παραμετροποίηση του συστήματος).

Στο τρίτο βήμα της Φάσης 1, η δική μας παραλλαγή δυναμικού προγραμματισμού εκτελείται για κάθε παράθυρο που έχει σχηματιστεί στα προηγούμενα βήματα. Στην ανάλυσή μας, υποθέτουμε το χειρότερο σενάριο, όπου ολόκληρη η ακολουθία δεδομένων μαρκάρεται ως ένα μονό παράθυρο. Ακολουθώντας την ανάλυση στο [8], αυτό

το βήμα απαιτεί κατά μέσο όρο $O(\mathcal{K}(|P|)^2 \cdot |S|)$ χρόνο, επειδή κάθε κελί δυναμικού προγραμματισμού περιέχει τιμές $2 \cdot \mathcal{K}(|P|) + 1$.

Συνοψίζοντας, η Φάση 1 έχει συνολικά χρόνο επεξεργασίας $O(\mathcal{K}(|P|)^2 \cdot |S|)$.

Κόστος των Φάσεων 2 και 3. Για να καθορίσουμε το κόστος των δυο τελευταίων φάσεων, παρατηρούμε ότι κάθε εμφάνιση του πυρήνα, που υπολογίζεται στην πρώτη φάση, θα υποστεί τον ίδιο αριθμό επεκτάσεων (προθέματος και επιθέματος). Συγκεκριμένα, ο αριθμός των επεκτάσεων προθέματος είναι κατά ένα μικρότερος από τον αριθμό των αλυσίδων επιθεμάτων, δηλαδή, $|P| - \beta$. Από την άλλη πλευρά, ο αριθμός των επεκτάσεων επιθέματος είναι κατά ένα μικρότερος από το μήκος μιας αλυσίδας επιθεμάτων, δηλαδή, $\alpha - 1$. Συνολικά, μια εμφάνιση του πυρήνα θα υποστεί $|P| - \beta + \alpha - 1 = |P| - |C|$ επεκτάσεις. Επιπλέον, ο συνολικός αριθμός των εμφανίσεων του πυρήνα που παράγονται στη Φάση 1 δίνεται από το $|S| \cdot f(|C|, \mathcal{K}(|P|))$, όπου $f(|C|, \mathcal{K}(|P|))$ είναι η πιθανότητα μιας τυχαίας ακολουθίας μήκους $|C|$ να ταιριάξει σε μια δεδομένη θέση της ακολουθίας δεδομένων με κόστος μεταγράφου που δεν ξεπερνά το $\mathcal{K}(|P|)$.

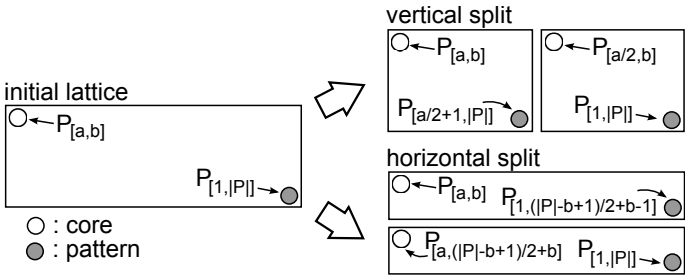
Συνοψίζοντας, και υποθέτοντας ότι κάθε επέκταση απαιτεί χρόνο T_{EX} (σταθερά για μια δεδομένη παραμετροποίηση του συστήματος), ο χρόνος επεξεργασίας των Φάσεων 2 και 3 είναι $|S| \cdot f(|C|, \mathcal{K}(|P|)) \cdot (|P| - |C|) \cdot T_{EX} = O(|S| \cdot f(|C|, \mathcal{K}(|P|)) \cdot (|P| - |C|))$.

Να σημειωθεί ότι ο υπολογισμός ενός κλειστού τύπου για τη συνάρτηση $f()$ είναι δύσκολη εργασία [8]. Ωστόσο, είναι δυνατό να προκύψει το ακόλουθο ανώτατο όριο για το $f(|C|, \mathcal{K}(|P|))$ βάσει της ανάλυσης που περιλαμβάνεται στο παράρτημα του [8]:

$$\sum_{i=|C|-\mathcal{K}(|P|)}^{|C|} \frac{1}{|\Sigma|^{|C|-\mathcal{K}(|P|)}} \binom{|C|}{|C| - \mathcal{K}(|P|)} \binom{i}{|C| - \mathcal{K}(|P|)}$$
$$+ \sum_{i=|C|+1}^{|C|+\mathcal{K}(|P|)} \frac{1}{|\Sigma|^{i-\mathcal{K}(|P|)}} \binom{|C|}{i - \mathcal{K}(|P|)} \binom{i}{i - \mathcal{K}(|P|)}.$$

Βελτιστοποίηση. Σε κάποιες περιπτώσεις του ARSM (π.χ. όταν ο λόγος $\mathcal{K}(|P|)/|R|$ είναι μεγάλος), είναι δυνατό το πλήθος εμφανίσεων του πυρήνα να είναι τόσο υψηλός ώστε ο συνολικός χρόνος εκτέλεσης του `PS-ARSM` να κυριαρχείται από τις Φάσεις 2 και 3. Για τέτοιες περιπτώσεις, προτιμάται συχνά να διχοτομείται το πρόβλημα σε δυο υπο-προβλήματα ARSM, όπου κάθε ένα έχει πολύ λιγότερες εμφανίσεις πυρήνα από το αρχικό και τα δύο προβλήματα να λύνονται ανεξάρτητα. Αυτό μπορεί να επιτευχθεί με το χωρισμό του δικτυωτού στα δύο και με τον κατάλληλο ορισμό της ακολουθίας πρότυπο και του πυρήνα για κάθε υπο-πρόβλημα — η συνάρτηση κατωφλίου $\mathcal{K}$ είναι κοινή. Το Σχήμα 3.11 δείχνει τους δύο τρόπους (κατακόρυφη και οριζόντια διάτμηση) με τους οποίους μπορούμε να θεωρήσουμε ένα πρόβλημα ARSM ως δυο ανεξάρτητα υπο-προβλήματα. Φαίνονται επίσης οι περιοχές του προτύπου και του πυρήνα για κάθε υπο-πρόβλημα. Σημειώστε ότι στο τέλος, οι εμφανίσεις από το ένα υπο-πρόβλημα πρέπει να ελεγχθούν σε σχέση με εκείνες από το άλλο, για να απομακρυνθούν οι εμφανίσεις που παραβιάζουν τη δεύτερη απαίτηση του Ορισμού 3.1.

Το μόνο ερώτημα που παραμένει είναι πότε και πώς πρέπει να εκτελείται η διχοτόμηση σε υπο-προβλήματα. Βάσει της ανάλυσης των προηγούμενων παραγράφων, εκτιμάμε το συνολικό χρόνο εκτέλεσης τριών σεναρίων: (α) της επίλυσης του αρχικού προβλήματος, (β) της επίλυσης των δυο υπο-προβλημάτων που προκύπτουν από κατακόρυφη διχοτόμηση του δικτυωτού, και (γ) της επίλυσης των δυο υπο-προβλημάτων που προκύπτουν από οριζόντια διχοτόμηση του δικτυωτού. Το σενάριο με το λιγότερο



Σχήμα 3.11: Παράδειγμα κατακόρυφης και οριζόντιας διχοτόμησης του δικτυωτού.

εκτιμώμενο κόστος είναι αυτό που επιλέγεται.

3.1.6 Πειραματική αξιολόγηση

Εκτελούμε ένα ολοκληρωμένο σύνολο πειραμάτων για την αξιολόγηση των επιδόσεων της μεθόδου μας, του ΠΣ-ΑΡΣΜ, τόσο σε συνθετικά όσο και σε πραγματικά σύνολα δεδομένων. Το Κεφάλαιο 3.1.6.1 περιγράφει τη διάταξη των πειραμάτων και το Κεφάλαιο 3.1.6.2 παρουσιάζει τα ευρήματά μας.

3.1.6.1 Πειραματική διάταξη

Αλγόριθμοι. Η αξιολόγηση περιλαμβάνει τρεις ακριβείς αλγορίθμους, δηλαδή αλγορίθμους που επιστρέφουν σωστά όλα τα αποτελέσματα ARSM.

- `PS-ARSM` , η λύση που προτείνουμε για το ARSM. Ο δυναμικός προγραμματισμός με τον αλγόριθμο ευριστικής αποκοπής [96] χρησιμοποιείται στη Φάση 1.
- `N-ARSM`, η αφελής προσέγγιση που εκτελεί έναν αλγόριθμο δυναμικού προγραμματισμού ASM για κάθε περιοχή. Η ευριστική αποκοπής [96] χρησιμοποιείται για τη βελτίωση της αποτελεσματικότητας του δυναμικού προγραμματισμού.
- `M-ARSM`, που εκτελεί το MASM [22] για κάθε ομάδα περιοχών που έχουν το ίδιο μήκος. Πειράματα έχουν δείξει ότι αποτελεί τον πιο αποτελεσματικό αλγόριθμο βασισμένο στο MASM (βλ. επίσης Ενότητα 2.1.1.3).

Υλοποιήσαμε όλους τους αλγορίθμους σε C++, και τρέξαμε όλα τα πειράματα σε ένα Linux PC Intel Core 2 Duo CPU, E8400, στα 3.00GHz. Στο σύστημά μας, $T_{DP} = 3.14 \cdot 10^{-5} msec$ και $T_{EX} = 2.11 \cdot 10^{-5} msec$.

Παράμετροι. Μετράμε την απόδοση σχετικά με το συνολικό χρόνο που απαιτείται για την παραγωγή αποτελεσμάτων ARSM, ενώ διαφοροποιούμε το ακόλουθο σύνολο

Πίνακας 3.2: Πειραματικές παράμετροι ARSM.

Παράμετρος	**Εύρος τιμών**	**Προεπιλογή**
$\|P\|$	10, 20, 30, 40, 50	30
$\|S\|$	1M, 5M, 10M, 50M, 100M	10M
$\|C\|/\|P\|$	$0,2$, $0,3$, $0,4$, $0,5$, $0,6$, $0,7$, $0,8$	$0,5$
$cPos$	left, middle, right	middle
α	$0,1$, $0,15$, $0,2$, $0,25$, $0,3$	$0,2$
$\|\Sigma\|$	4, 20, 94	4

παραμέτρων: το μήκος της ακολουθίας πρότυπο $|P|$ (μετρημένο σε αριθμό συμβόλων), το μήκος της ακολουθίας δεδομένων $|S|$, το λόγο του μήκους του πυρήνα προς το μήκος του προτύπου $|core|/|P|$, τη θέση του πυρήνα $cPos$, το λόγο του επιτρεπόμενου κόστους μεταγράφου για κάθε περιοχή προς το μήκος της $\alpha = \mathcal{K}(|R|)/|R|$ (θεωρούμε γραμμικές συναρτήσεις κατωφλίου)· το μέγεθος του αλφαβήτου $|\Sigma|$. Ο Πίνακας 3.2 περιέχει όλες τις παραμέτρους και το εύρος των εξεταζόμενων τιμών. Σε κάθε πείραμα, διαφοροποιούμε μια παράμετρο και θέτουμε τις υπόλοιπες στις προεπιλεγμένες τους τιμές όπως φαίνεται στον πίνακα.

Σύνολα δεδομένων. Χρησιμοποιούμε συνθετικά και πραγματικά σύνολα δεδομένων. Για τα συνθετικά (D1), χρησιμοποιούμε τυχαίες ακολουθίες που ακολουθούν το ομοιόμορφο μοντέλο Bernoulli, δηλαδή, κάθε σύμβολο έχει πιθανότητα $1/|\Sigma|$ να εμφανιστεί και επιλέγεται ανεξάρτητα από τα άλλα. Δημιουργούμε 20 ζεύγη ακολουθιών δεδομένων και προτύπων και, για κάθε ζεύγος, εκτελούμε τους αλγορίθμους 5 φορές. Οπότε, κάθε αναφερόμενη τιμή χρόνου είναι ο μέσος όρος 100 εκτελέσεων.

Παίρνουμε επίσης πραγματικά σύνολα δεδομένων, που τα παίρνουμε από τη βάση δεδομένων Ensembl[2]. Το σύνολο δεδομένων 3′UTR (D2) είναι μια ακολουθία 44 εκατομμυρίων νουκλεοτιδίων για την 3′ μη μεταφρασμένη περιοχή των μεταγράφων των ανθρώπινων γονιδίων. Το σύνολο δεδομένων CDS (D3) είναι μια ακολουθία 74 εκατομμυρίων νουκλεοτιδίων για την κωδική περιοχή των μεταγράφων των ανθρώπινων γονιδίων. Να σημειωθεί ότι αυτά είναι γονιδιακά σύνολα δεδομένων και μπορούν να χρησιμοποιηθούν μόνο σε πειράματα με μέγεθος αλφαβήτου $\Sigma = 4$. Εξάγουμε 20 τυχαίες υπακολουθίες από αυτά τα σύνολα δεδομένων για να χρησιμεύσουν ως τα πρότυπα και για κάθε μια από αυτές εκτελούμε τους αλγορίθμους 5 φορές. Ως αποτέλεσμα, κάθε αναφερόμενη τιμή χρόνου είναι ο μέσος όρος 100 εκτελέσεων.

3.1.6.2 Αποτελέσματα

Ανάλυση χρόνου εκτέλεσης του `PS-ARSM` . Ερευνάμε τις επιδόσεις χρόνου εκτέλεσης του `PS-ARSM` για την προεπιλεγμένη πειραματική ρύθμιση. Να σημειωθεί ότι, για αυτές τις τιμές παραμέτρων, το `PS-ARSM` επιλέγει να χωρίσει το δικτυωτό περιοχών οριζόντια.

Ο Πίνακας 3.3 παρουσιάζει τη μνήμη που καταλαμβάνουν οι δυο πίνακες κατακερματισμού του `PS-ARSM` . Οι δυο εκτελέσεις που φαίνονται σε αυτόν αντιστοιχούν στις δυο εκτελέσεις του `PS-ARSM` , μια για κάθε μισό του αρχικού δικτυωτού. Να σημειωθεί ότι το μέγεθος του `cands` είναι σημαντικά μικρότερο από εκείνο του `head_occs` . Ο λόγος είναι ότι πολύ λιγότερες εμφανίσεις επιβιώνουν μετά τη Φάση 2. Για παράδειγμα, όπως δείχνει ο πίνακας για την πρώτη εκτέλεση στο D1, υπάρχουν 128.577 διακριτές θέσεις όπου λήγουν εμφανίσεις στη Φάση 2, αλλά μόνο 2.006 θέσεις όπου εμφανίσεις ξεκινούν στη Φάση 3.

Ο Πίνακας 3.4 παρουσιάζει το σχετικό χρόνο που ξοδεύτηκε σε κάθε φάση του `PS-ARSM` · οι τιμές βασίζονται στο συνολικό χρόνο εκτέλεσης και για τις δυο εκτελέσεις. Η Φάση 1 είναι μακράν η πιο ακριβή καθώς καταναλώνει περίπου 97% του συνολικού χρόνου εκτέλεσης. Μεταξύ των άλλω δυο φάσεων, η Φάση 3 απαιτεί περισσότερο χρόνο καθώς επεκτείνει λιγότερες εμφανίσεις από τη Φάση 2 (βλ. Πίνακα 3.3).

Διαφοροποιώντας το μήκος της ακολουθίας πρότυπο. Το Σχήμα 3.12 παρουσιάζει τους χρόνους εκτέλεσης (σε λογαριθμική κλίμακα) όλων των αλγορίθμων καθώς διαφοροποιείται το μήκος της ακολουθίας πρότυπο $|P|$. Τα ευρήματα είναι

[2]http://www.ensembl.org/biomart/martview/

Πίνακας 3.3: Κατανάλωση μνήμης των πινάκων κατακερματισμού του PS-ARSM .

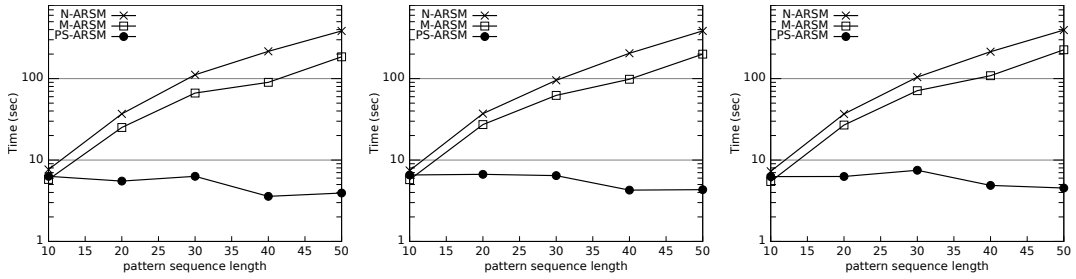
Σύνολο δεδομένων	**μέγεθος του** head_occs	**μέγεθος του** cands
D1 run 1	7,69MB (128.577 καταχωρήσεις)	0,17MB (2.006 καταχωρήσεις)
D1 run 2	1,61MB (27.001 καταχωρήσεις)	0,03MB (255 καταχωρήσεις)
D2 run 1	7,82MB (126.852 καταχωρήσεις)	0,37MB (3.221 καταχωρήσεις)
D2 run 2	2,11MB (33.802 καταχωρήσεις)	0,18MB (944 καταχωρήσεις)
D3 run 1	9,51MB (155.912 καταχωρήσεις)	0,33MB (3.898 καταχωρήσεις)
D3 run 2	2,52MB (41.282 καταχωρήσεις)	0,07MB (698 καταχωρήσεις)

Πίνακας 3.4: Κατάτμηση του χρόνου εκτέλεσης για τις φάσεις του PS-ARSM .

Σύνολο δεδομένων	**Φάση 1 (%)**	**Φάση 2 (%)**	**Φάση 3 (%)**
D1	97,26	2,57	0,17
D2	96,70	2,88	0,42
D3	96,58	3,20	0,22

παρόμοια για όλα τα σύνολα δεδομένων. Όσο μεγαλώνει το $|P|$, τόσο αυξάνονται ο αριθμός των περιοχών και το μέσο μήκος τους. Αυτό εξηγεί γιατί ο χρόνος εκτέλεσης του M-ARSM και του N-ARSM αυξάνεται.

Από την άλλη πλευρά, ο χρόνος εκτέλεσης του PS-ARSM παραμένει κάτω από 10 δευτερόλεπτα, μη επηρεαζόμενος από το $|P|$. Να σημειωθεί ότι όσο το $|P|$ μεγαλώνει και οι άλλες παράμετροι μένουν σταθερές, τόσο το μήκος του πυρήνα $|C|$ όσο και το κατώφλι κόστους $\mathcal{K}(|P|)$ μεγαλώνουν επίσης. Συγκεκριμένα, η διαφορά $|C| - \mathcal{K}(|P|) = 0.5 \cdot |P| - 0.2 \cdot |P| = 0.3 \cdot |P|$ (για τις προεπιλεγμένες τιμές του Πίνακα 3.2) μεγαλώνει γραμμικά με το μήκος του προτύπου. Αυτή η διαφορά παίζει σημαντικό ρόλο στον αριθμό των εμφανίσεων του πυρήνα που προκύπτουν κατά την πρώτη φάση του PS-ARSM . Όπως δείχνει η ανάλυση του Κεφαλαίου 3.1.5, ο αριθμός των εμφανίσεων του πυρήνα μειώνεται γρήγορα καθώς το $|C| - \mathcal{K}(|P|)$ αυξάνεται. Οπότε, αν και ο χρόνος εκτέλεσης της Φάσης 1 αυξάνεται, εκείνος των Φάσεων 2 και 3 μειώνεται με το $|P|$. Ως αποτέλεσμα, για μεγάλες τιμές $|P|$, το PS-ARSM είναι σχεδόν δυο τάξεις μεγέθους πιο γρήγορο.

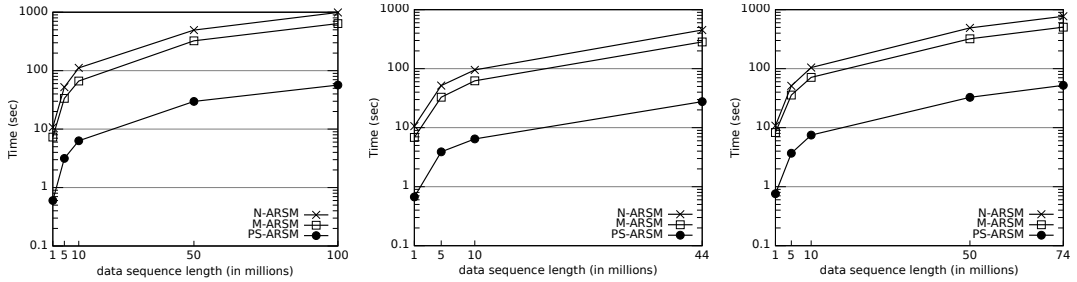


(α′) Συνθ. σύνολο δεδομένων, D1 (β′) 3'UTR σύνολο δεδομένων, D2 (γ′) CDS σύνολο δεδομένων, D3

Σχήμα 3.12: Διαφοροποιώντας το μήκος της ακολουθίας πρότυπο $|P|$.

Διαφοροποιώντας το μήκος της ακολουθίας δεδομένων. Το Σχήμα 3.13 δείχνει το χρόνο εκτέλεσης σε ακολουθίες δεδομένων διαφορετικών μηκών. Να σημειωθεί ότι για να κατασκευάσουμε μια ακολουθία με μήκος $|S|$ από τα πραγματικά σύνολα δεδομένων, εξάγουμε τα πρώτα σύμβολα $|S|$, και ότι το μέγιστο δυνατό μήκος είναι 44M για το Δ2 και 74M για το Δ3. Ο χρόνος εκτέλεσης όλων των μεθόδων αυξάνεται γραμμικά, καθώς αυξάνεται το μέγεθος της ακολουθίας δεδομένων. Οπότε, η βελτίωση απόδοσης του PS-ARSM επί του M-ARSM και του N-ARSM, σε όλα τα σύνολα

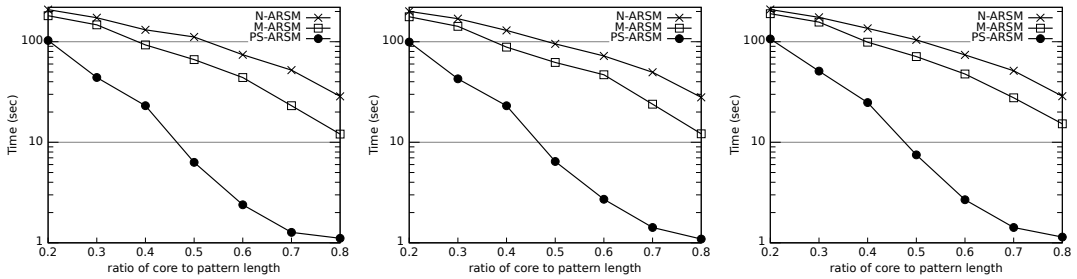
δεδομένων και τις τιμές $|S|$, είναι είναι πάνω από μια τάξη μεγέθους.



(α′) Συνθ. σύνολο δεδομένων, D1 (β′) 3'UTR σύνολο δεδομένων, D2 (γ′) CDS σύνολο δεδομένων, D3

Σχήμα 3.13: Διαφοροποιώντας το μήκος της ακολουθίας δεδομένων $|S|$.

Διαφοροποιώντας το λόγο μήκους πυρήνα/μήκος προτύπου. Το Σχήμα 3.14 απεικονίζει το χρόνο εκτέλεσης για αρκετές τιμές του λόγου $|C|/|P|$. Τα αποτελέσματα είναι παρόμοια για όλα τα σύνολα δεδομένων. Καθώς αυξάνεται ο λόγος $|C|/|P|$, ο χρόνος εκτέλεσης όλων των μεθόδων μειώνεται. Αυτό οφείλεται στο ότι ο αριθμός των περιοχών μειώνεται και το μέσο μήκος περιοχής αυξάνεται. Να σημειωθεί ότι το πλεονέκτημα του `PS-ARSM` επί των ανταγωνιστών του αυξάνεται με το λόγο καθώς, επιπροσθέτως, ο αριθμός εμφανίσεων του πυρήνα που προκύπτουν στην πρώτη φάση του μειώνεται. Ο λόγος είναι ότι το $|C|$, και επομένως το $|C| - \mathcal{K}(|P|)$, αυξάνεται (βλ. Ενότητα 3.1.5).

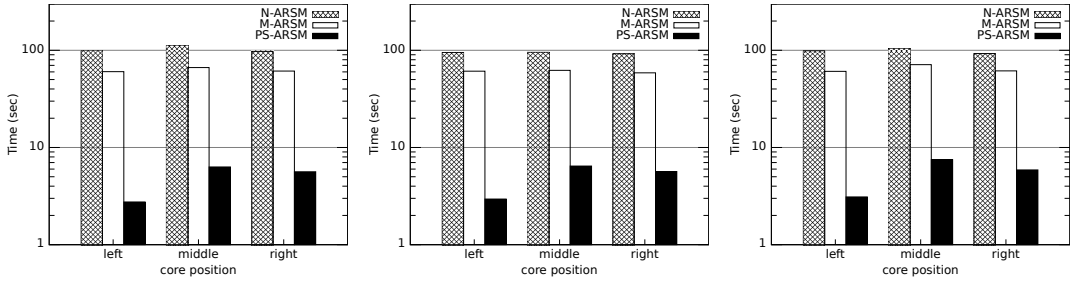


(α′) Συνθ. σύνολο δεδομένων, D1 (β′) 3'UTR σύνολο δεδομένων, D2 (γ′) CDS σύνολο δεδομένων, D3

Σχήμα 3.14: Διαφοροποιώντας το λόγο μήκους πυρήνα προς μήκος προτύπου $|C|/|P|$.

Διαφοροποιώντας τη θέση του πυρήνα. Το Σχήμα 3.15 παρουσιάζει το χρόνο εκτέλεσης καθώς αλλάζει η τοποθεσία του πυρήνα στο πρότυπο. Καθώς μετακινούμε τον πυρήνα στην αρχή ή στο τέλος του προτύπου (ενώ οι άλλες παράμετροι παραμένουν σταθερές) ο αριθμός περιοχών μειώνεται λίγο. Οπότε, όλοι οι αλγόριθμοι τρέχουν γρηγορότερα σε αυτή την περίπτωση. Ωστόσο, η μείωση στο χρόνο εκτέλεσης του `PS-ARSM` είναι πιο έντονη. Εν συντομία, ο λόγος είναι ότι, επιπροσθέτως, ο αριθμός των εμφανίσεων του πυρήνα είναι μικρότερος όταν ο πυρήνας βρίσκεται κοντά στα άκρα του προτύπου. Ακολουθεί μια λεπτομερής εξήγηση για την περίπτωση όπου ο πυρήνας βρίσκεται αριστερά (η εξήγηση για την άλλη περίπτωση είναι παρόμοια).

Θεωρούμε την περίπτωση όπου ο πυρήνας είναι στο κέτρο του $C_m = [a_m, b_m]$ και εκείνη όπου βρίσκεται στα αριστερά του προτύπου $C_l = [a_l, b_l] = [a_m - c, b_m - c]$, όπου $c > 0$. Και στις δυο περιπτώσεις, το `PS-ARSM` επιλέγει να χωρίσει το δικτυωτό οριζόντια. Έστω C_m^{top} και P_m^{top} (αντίστ. C_m^{bot} και P_m^{bot}) δηλώνουν τον πυρήνα και το πρότυπο για το πάνω (αντίστ. κάτω) μισό δικτυωτό όπου ο πυρήνας βρίσκεται στο μέσο. Ο αντίστοιχος συμβολισμός για τον πυρήνα στα αριστερά προκύπτει αντικαθιστώντας το m με l. Μετά, για το πάνω μισό του δικτυωτού, ισχύει ότι $|C_m^{top}| = |C_l^{top}|$

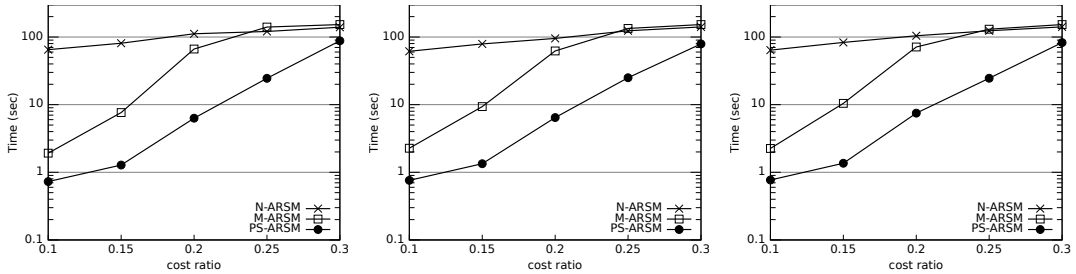


(α′) Συνθ. σύνολο δεδομένων, D1 (β′) 3'UTR σύνολο δεδομένων, D2 (γ′) CDS σύνολο δεδομένων, D3

Σχήμα 3.15: Διαφοροποιώντας τη θέση του πυρήνα $cPos$.

και $|P_l^{top}| = (|P_l| + b_l - 1)/2 = |P_m^{top}| - c/2 < |P_m^{top}|$. Με άλλα λόγια, όταν ο πυρήνας είναι αριστερά, το πρότυπο του πάνω μισού είναι πιο μικρό, και έτσι το επιτρεπόμενο κόστος είναι επίσης μικρότερο, κάτι που οδηγεί σε λιγότερες εμφανίσεις του πυρήνα. Από την άλλη πλευρά, για το κάτω μισό δικτυωτό, ισχύει ότι $|P_m^{bot}| = |P_l^{bot}|$, ανδ $|C_l^{bot}| = (|P_l| + b_l + 1)/2 - a_l + 1 = |C_m^{bot}| + c/2 > |C_m^{bot}|$. Εδώ, το πρότυπο του κάτω μισού, και επομένως το επιτρεπόμενο κόστος, είναι το ίδιο, αλλά ο πυρήνας του κάτω μισού είναι μεγαλύτερος, κάτι που σημαίνει πάλι λιγότερες εμφανίσεις όταν ο πυρήνας είναι στα αριστερά. Συνολικά, ο αριθμός των εμφανίσεων του πυρήνα και στα δύο δικτυωτών είναι μικρότερος όταν ο πυρήνας είναι στα αριστερά.

Διαφοροποιώντας το λόγο κόστους. Το Σχήμα 3.16 δείχνει το χρόνο εκτέλεσης καθώς αλλάζει ο λόγος κατωφλίου κόστους. Υψηλότερες τιμές α απαιτούν περισσότερη προσπάθεια από όλες τις μεθόδους. Ωστόσο, η απόδοση του `M-ARSM` και του `PS-ARSM` επιδεινώνεται γρηγορότερα και προσεγγίζει εκείνη της αφελούς μεθόδου. Διαισθητικά, ο λόγος είναι ο ακόλουθος. Αυτές οι δυο μέθοδοι προσπαθούν να ξεσκαρτάρουν ορισμένες περιοχές της ακολουθίας δεδομένων που δεν περιέχουν αποτελέσματα ARSM. Όταν αυξάνεται το επιτρεπόμενο κόστος μεταγράφου, λιγότερες και μικρότερες περιοχές αποκλείονται, και έτσι το προνόμιο φιλτραρίσματός τους μειώνεται και συμπεριφέρονται παρόμοια με τη μέθοδο ωμής βίας του `N-ARSM`.

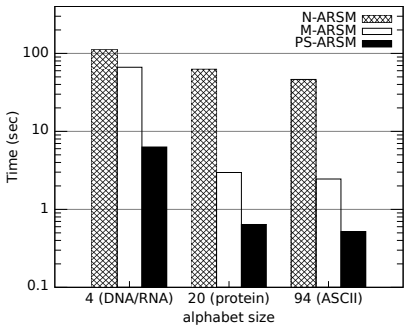


(α′) Συνθ. σύνολο δεδομένων, D1 (β′) 3'UTR σύνολο δεδομένων, D2 (γ′) CDS σύνολο δεδομένων, D3

Σχήμα 3.16: Διαφοροποιώντας το λόγο κόστους α.

Διαφοροποιώντας το μέγεθος του αλφαβήτου. Αν και το ενδιαφέρον της παρούσας εργασίας εστιάζεται στα γονιδιακά σύνολα δεδομένων, όπου $|\Sigma| = 4$, εξετάζουμε επίσης τη συμπεριφορά όλων των μεθόδων σε βάσεις δεδομένων με διαφορετικό μέγεθος αλφαβήτου. Συγκεκριμένα, θεωρούμε τις πρωτεϊνικές ακολουθίες, όπου $|\Sigma| = 20$, και ακολουθίες κειμένου ASCII, όπου $|\Sigma| = 94$. Σε αυτό το πείραμα, χρησιμοποιούμε μόνο συνθετικά δεδομένα. Το Σχήμα 3.17 δείχνει τα αποτελέσματα. Ο χρόνος εκτέλεσης όλων των μεθόδων μειώνεται καθώς αυξάνεται το $|\Sigma|$, επειδή υπάρχουν λιγότερες πιθανές εμφανίσεις (βλ. Ενότητα 3.1.5). Σημειώστε ότι τόσο το `M-ARSM` όσο και το `PS-ARSM` γίνονται σημαντικά πιο αποτελεσματικά από την αφελή

μέθοδο, ενώ το προνόμιο του PS-ARSM επί του M-ARSM παραμένει κοντά στη μια τάξη μεγέθους.



Σχήμα 3.17: Διαφοροποιώντας το μέγεθος του αλφαβήτου Σ.

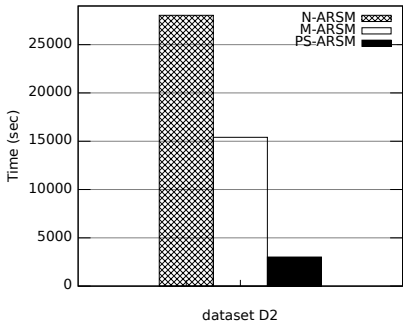
Χρησιμοποιώντας πραγματικά πρότυπα. Στο τελικό πείραμα, ερευνούμε την απόδοση του PS-ARSM για το πραγματικό πρόβλημα πρόβλεψης προσδέσεων miRNA. Βάσει αρκετών βιολογικών παρατηρήσεων (π.χ. [19]), επιλέγουμε τις ακόλουθες παραμέτρους. Χρησιμοποιούμε το σύνολο δεδομένων 3′UTR (D2) σαν βάση δεδομένων, καθώς οι πιο γνωστές προσδέσεις miRNA βρίσκονται σε αυτή την περιοχή γονιδίων. Επιπλέον, επιλέγουμε τυχαία 100 ακολουθίες miRNA από το MirBase[3] σαν πρότυπα. Καθώς τα πιο σημαντικά σύμβολα για την πρόσδεση βρίσκονται στα αριστερά των ακολουθιών miRNA, επιλέγουμε $|C| = 10$, $cPos = left$, και ορίζουμε $\alpha = 0.2$ για όλες τις περιοχές, εκτός από τον πυρήνα όπου το επιτρεπόμενο κόστος μεταγράφου περιορίζεται σε 1.

Το Σχήμα 3.18 παρουσιάζει τα αποτελέσματα. Να σημειωθεί ότι αναφέρουμε το συνολικό χρόνο εκτέλεσης για όλα τα 100 πρότυπα miRNA. Τα ευρήματα είναι παρόμοια με την περίπτωση συνθετικών προτύπων. Το PS-ARSM είναι περίπου μια τάξη μεγέθους πιο γρήγορο από το N-ARSM, και πάνω από πέντε φορές πιο γρήγορο από το M-ARSM.

3.2 Συστήματα πρόβλεψης στόχων βασισμένα στο Νέφος

Η επιτάχυνση του βήματος ταιριάσματος ακολουθιών των μεθόδων πρόβλεψης στόχων δεν αρκεί για την επίτευξη επιδόσεων σχεδόν πραγματικού χρόνου, καθώς αυτές οι μέθοδοι περιλαμβάνουν επίσης κάποιες άλλες υπολογιστικά απαιτητικές διεργασίες. Η κατανομή των εμπλεκόμενων υπολογισμών στους κόμβους μιας υποδομής Νέφους μπορεί να βοηθήσει στην επίλυση του ζητήματος αυτού. Οπότε, σχεδιάσαμε και αναπτύξαμε δυο συστήματα πρόβλεψης στόχων miRNA βασισμένα στο Νέφος, το TarCloud [97] και το MR-microT [40]. Το πρώτο, που αναπτύχθηκε χρησιμοποιώντας το προγραμματιστικό πλαίσιο Azure της Microsoft, μπορούσε να παρέχει τους στόχους ενός δεδομένου miRNA σε περίπου 5 λεπτά, ωστόσο δεν ήταν πολύ κλιμακώσιμο για μεγάλο αριθμό ταυτόχρονων αιτημάτων πρόβλεψης. Το δεύτερο, που βασίστηκε στο MapReduce, παρέχει τις προβλέψεις κάθε δεδομένου miRNA σε λιγότερο από 2 λεπτά, και κλιμακώνεται καλά όσο μεγαλώνει ο αριθμός των ταυτόχρονων προβλέψεων

[3]http://www.mirbase.org/, βάση δεδομένων όπου είναι καταγεγραμμένες όλες οι ακολουθίες miRNA.



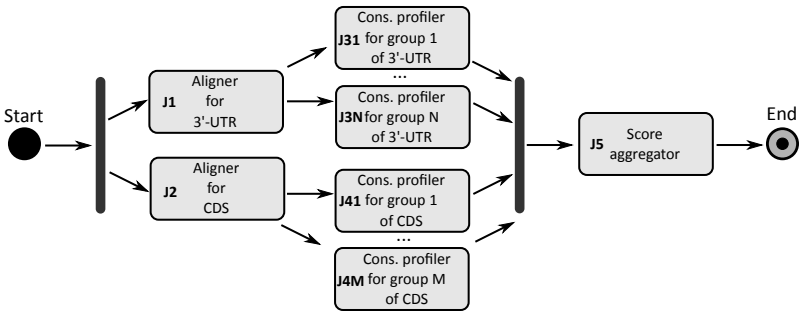
Σχήμα 3.18: Πείραμα με πραγματικά πρότυπα.

(με την προϋπόθεση ότι υπάρχουν αρκετοί πόροι). Το Κεφάλαιο 3.2.1 περιγράφει το TarCloud, ενώ το Κεφάλαιο 3.2.2 εισάγει το MR-microT.

3.2.1 TarCloud: πρόβλεψη στόχων στο MS Azure

3.2.1.1 Τεχνολογίες και αρχιτεκτονική συστήματος

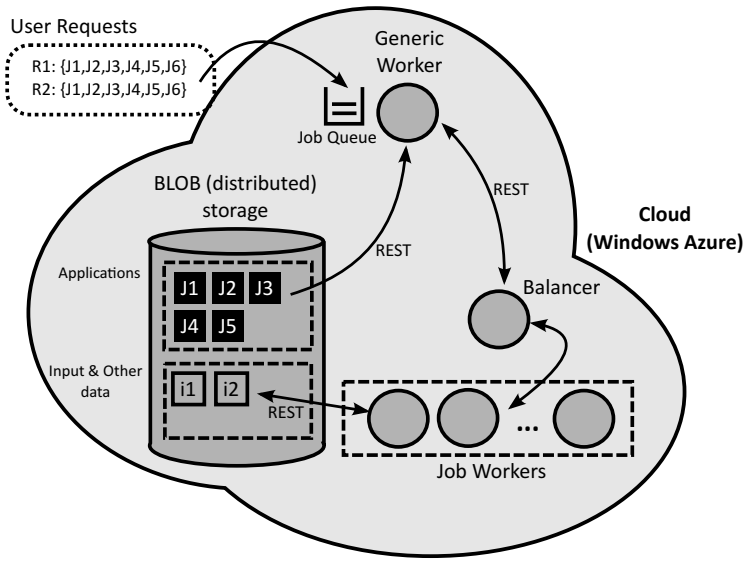
Το TarCloud [97] είναι ένα σύστημα πρόβλεψης στόχων miRNA, βασισμένο στη μέθοδο DIANA microT v.5 [78] (βλ. Κεφάλαιο 2.2.2) και υλοποιημένο χρησιμοποιώντας τη πλατφόρμα Νέφους *Microsoft Azure*[4]. Η ανωτέρω πλατφόρμα αποτελείται από διάφορες υπηρεσίες εμπορευματοποιημένες μέσω τριών προϊόντων. Αυτά είναι: (α) το *Windows Azure*, ένα λειτουργικό σύστημα που παρέχει κλιμακώσιμο εξοπλισμό υπολογισμού και αποθήκευσης, (β) το *SQL Azure*, μια εκδοχή του SQL Server βασισμένη στο Νέφος, και (γ) το *Windows Azure AppFabric*, μια συλλογή υπηρεσιών που υποστηρίζουν εφαρμογές Νέφους. Η πλατφόρμα παρέχει ένα API χτισμένο σε REST, HTTP και XML, το οποίο επιτρέπει σε έναν προγραμματιστή να αλληλεπιδρά με τις υπηρεσίες Azure.



Σχήμα 3.19: Η ροή εργασιών του TarCloud

Το TarCloud ακολουθεί την προσέγγιση του DIANA microT v.5 για την πρόβλεψη πιθανών στόχων για μια δεδομένη ακολουθία miRNA (για λεπτομέρειες σχετικά με το DIANA microT v.5 βλ. Κεφάλαιο 2.2.2). Βάσει αυτού, μια διεργασία του TarCloud αποτελείται από ένα σύνολο διακριτών εργασιών. Το Σχήμα 3.19 παρουσιάζει τη ροή εργασιών δείνχοντας τις εξαρτήσεις μεταξύ αυτών των εργασιών. Κάθε ορθογώνιο αναπαριστά μια ανεξάρτητη εργασία του TarCloud. Τα βέλη δείχνουν τη ροή δεδομένων μεταξύ των εργασιών. Οι γκρίζες κατακόρυφες μπάρες αναπαριστούν μια διακλάδωση ή μια συνένωση. Μια διακλάδωση πυροδοτεί παράλληλες δραστηριότητες, ενώ μια

[4] `http://www.windowsazure.com`



Σχήμα 3.20: Η αρχιτεκτονική του TarCloud

συνένωση συγχωνεύει το προϊόν των παράλληλων δραστηριοτήτων. Ακολουθεί μια σύντομη περιγραφή κάθε μιας από της εικονιζόμενες εργασίες.

Στοιχιστές (J1 και J2). Ο *στοιχιστής* είναι υπεύθυνος για (α) τον εντοπισμό των MRE[5] της ακολουθίας miRNA, (β) τον υπολογισμό του σκορ στοίχισης ακολουθιών του σπόρου miRNA στο MRE, και (ς) τον υπολογισμό της ενέργειας πρόσδεσης του miRNA στο γονίδιο (με χρήση RNAhybrid [79]). Ο στοιχιστής εκτελεί την εργασία J1 για τις περιοχές 3′-UTR όλων των γονιδίων του επιλεγμένου γονιδιώματος, και τη J2 για τις περιοχές CDS. Να σημειωθεί ότι αυτές οι δυο εργασίες εκτελούνται παράλληλα.

Αναλυτής διατηρησιμότητας (J31 ... J3N και J41 ... J4M). Ο *αναλυτής διατηρησιμότητας* είναι υπεύθυνος για τον έλεγχο του τρόπου με τον οποίο κάθε MRE διατηρείται στα γονιδιώματα διαφόρων ειδών. Επί του παρόντος, το TarCloud αξιοποιεί περίπου 27 είδη για την αξιολόγηση του προφίλ διατηρησιμότητας των MRE, λαμβάνοντας υπόψη τόσο διατηρημένα όσο και μη διατηρημένα MRE για την εκτίμηση του τελικού σκορ. Η ανάλυση διατηρησιμότητας για όλα τα MRE που ανήκουν στην ίδια περιοχή (3′-UTR/CDS) του ίδιου χρωμοσώματος πραγματοποιείται σαν δεσμίδα. Πολλές δεσμίδες υπολογισμών ανάλυσης διατηρησιμότητας μπορούν να εκτελεστούν παράλληλα. Στο Σχήμα 3.19 οι εργασίες ανάλυσης διατηρησιμότητας σε περιοχές 3′-UTR έχουν το πρόθεμα 'J3', ενώ οι υπόλοιπες έχουν το πρόθεμα 'J4'.

Συναθροιστής σκορ (J5). Ο *συναθροιστής σκορ* υπολογίζει, για κάθε γονίδιο, το συνολικό σκορ όλων των MRE του. Η συνάθροιση είναι ένα σταθμισμένο άθροισμα που λαμβάνει υπόψη του επίσης τη μοριακή αναδίπλωση (δηλαδή, την τρισδιάστατη δομή των μορίων). Ο συναθροιστής σκορ εκτελεί την εργασία J5.

Η αρχιτεκτονική του TarCloud, που απεικονίζεται στο Σχήμα 3.20, εμπλέκει Εικονικές Μηχανές (VM), που λέγονται *Εργάτες*, και είναι πλήρως ικανές να εκτελέσουν οποιαδήποτε από τις ανωτέρω εργασίες. Κάθε Εργάτης ακούει αιτήματα HTTP (REST) και εκτελεί τις εργασίες του TarCloud που περιγράφουν αυτά τα αιτήματα. Κάθε δεδομένο εισόδου και εξόδου αποθηκεύεται κυρίως σαν BLOB στον κατανεμημένο αποθηκευτικό χώρο στο Νέφος του Azure της Microsoft. Αναφερόμαστε σε αυτό τον αποθηκευτικό χώρο σαν *τον αποθηκευτικό χώρο BLOB*. Ο αριθμός των μηχανημάτων

[5]Τα MRE είναι τοποθεσίες στα γονίδια για τις οποίες είναι πολύ πιθανό, βάσει στοίχισης ακολουθιών, να είναι σημεία πρόσδεσης miRNA. Βλ. Κεφάλαιο 2.2.2 για περισσότερες λεπτομέρειες.

Εργατών που απασχολούνται για μια συγκεκριμένη παραμετροποίηση του TarCloud, είναι μια παράμετρος σχεδιασμού ουσιώδης για την αποδοτικότητα του συστήματος. Για την κατανομή των αιτημάτων HTTP στα μηχανήματα των Εργατών, χρησιμοποιούμε επίσης μια VM, που λέγεται *Εξισορροπητής*. Να σημειωθεί ότι δεν είναι ουσιώδες ο Εξισορροπητής να βρίσκεται σε ξεχωριστή VM. Ένας από τους Εργάτες μπορεί να φιλοξενεί επίσης τον Εξισορροπητή.

Για κάθε μια από τις εργασίες τουTarCloud, υλοποιείται μια ξεχωριστή εφαρμογή του Azure[6] και αποθηκεύεται στον αποθηκευτικό χώρο BLOB. Αυτές οι εφαρμογές είναι απλώς ελαφρά τερματικά (thin clients) που δημιουργούν αιτήματα HTTP (REST). Τα αιτήματα λένε στον Εξισορροπητή να απασχολήσει κάποιους Εργάτες για να εκτελέσουν μια συγκεκριμένη εργασία του TarCloud. Μια άλλη VM, που λέγεται *Γενικός Εργάτης (General Worker - GW)*, είναι υπεύθυνη να φορτώσει τις ανωτέρω εφαρμογές από τον αποθηκευτικό χώρο BLOB, και να τις εκτελέσει με τη σειρά που καθορίζει η ροή εργασιών του Σχήματος 3.19, μετά από αίτημα κάποιου χρήστη. Τα αιτήματα των χρηστών συγκεντρώνονται σε μια *ουρά εργασιών*. Για την επίτευξη συγχρονισμού, ο GW ελέγχει απλώς εάν η είσοδος για κάθε εργασία είναι έτοιμη. Εν συντομία, για κάθε εργασία, ο GW γνωρίζει τα URI των αρχείων εισόδου της (στον αποθηκευτικό χώρο BLOB) και, πριν την εκτέλεση της εργασίας, εκτελεί ανίχνευση (polling) για να ελέγξει εάν τα αρχεία εισόδου είναι πλήρη. Η εκτέλεση αρχίζει μόνο όταν τα αρχεία εισόδου είναι έτοιμα. Για παράδειγμα, για τη ροή εργασιών του Σχήματος 3.19, ο GW δε θα εκτελέσει την J31, μέχρι τα αρχεία εξόδου από την J1 (τα οποία είναι είσοδος για την J31) να είναι έτοιμα.

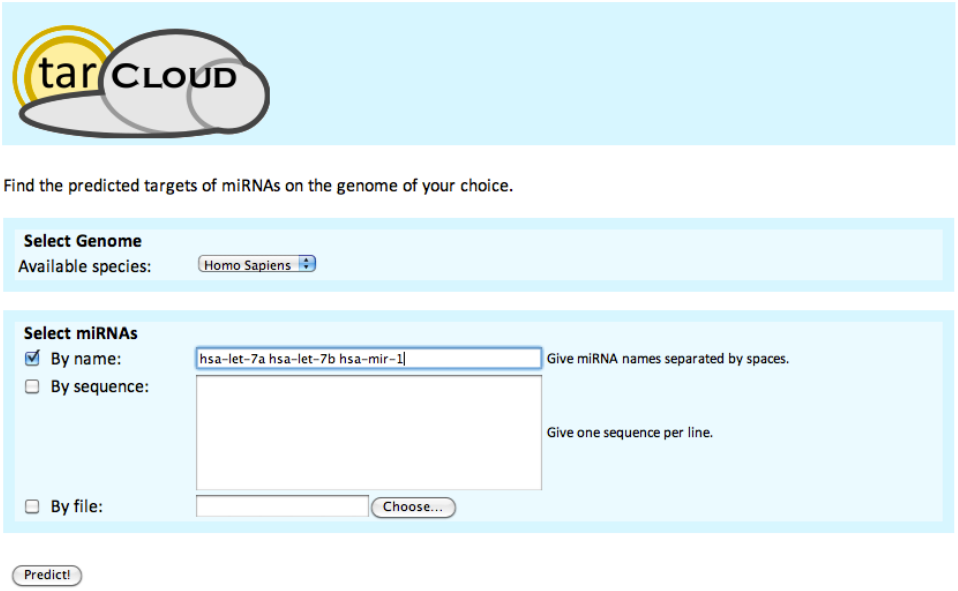
3.2.1.2 Διεπαφή χρήστη και αξιολόγηση

Υλοποιήσαμε επίσης μια διεπαφή Ιστού, για να δώσουμε στους χρήστες του TarCloud έναν εύκολο και διαισθητικό τρόπο να ξεκινούν εργασίες πρόβλεψης στόχων miRNA. Το Σχήμα 3.21 απεικονίζει ένα στιγμιότυπο αυτής της διεπαφής. Ο χρήστης απλώς επιλέγει το επιθυμητό είδος από μια αναπτυσσόμενη λίστα. Έπειτα, επιλέγει μια ή περισσότερες ακολουθίες miRNA. Έχει 3 επιλογές: (α) να επιλέξει ήδη γνωστές ακολουθίες miRNA εισάγοντας τα ονόματά τους στο πεδίο 'Select miRNA by name' (αποθηκεύουμε όλα τα miRNA από την τελευταία έκδοση της miRBase), (β) να επιλέξει άγνωστα miRNA καθορίζοντας τις ακολουθίες τους στο πεδίο 'Select miRNA by sequence', ή (γ) να επιλέξει ένα σύνολο άγνωστων miRNA ανεβάζοντας ένα αρχείο που να περιέχει ένα όνομα και μια ακολουθία για κάθε ένα από αυτά. Όταν ο χρήστης επιλέξει το κουμπί 'Predict!', αποστέλλεται ένα αίτημα για την προαναφερθείσα εργασία στο Γενικό Εργάτη.

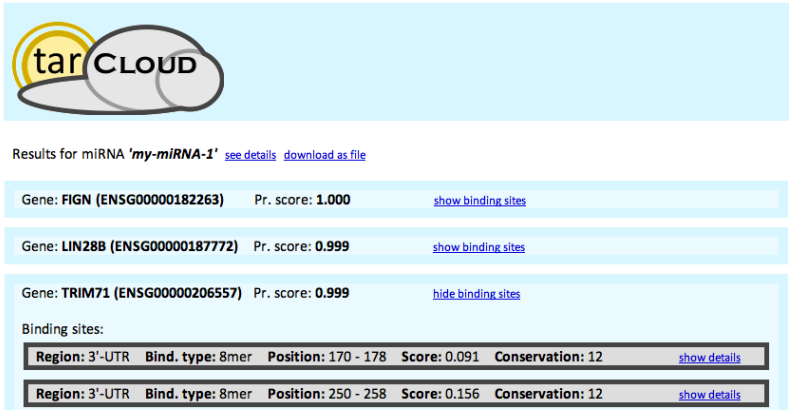
Όταν το προϊόν εξόδου είναι έτοιμο και αποθηκευμένο στον αποθηκευτικό χώρο BLOB, η διεπαφή Ιστού του TarCloud δημιουργεί μια ιστοσελίδα που περιέχει το προϊόν εξόδου. Ο χρήστης έχει δυο επιλογές: (α) να ψάξει τους προβλεπόμενους στόχους χρησιμοποιώντας το δικό μας ενσωματωμένο απεικονιστή αποτελεσμάτων (βλ. Σχήμα 3.22), ή (β) να κατεβάσει τα αρχεία κειμένου του προϊόντος εξόδου στον υπολογιστή του.

Πραγματοποιήσαμε επίσης πειράματα για την αξιολόγηση της απόδοσης του συστήματός μας ζητώντας προβλέψεις για 65 ακολουθίες miRNA. Αυτές οι ακολουθίες ήταν ακολουθίες miRNA ανθρώπου ή ποντικού, επιλεγμένες τυχαία από τη miRBase (τη βάση δεδομένων όλων των γνωστών miRNA). Η υποδομή μας αποτελούνταν από

[6]Οι εφαρμογές του Azure είναι εκτελέσιμα 32-bit των Windows.



Σχήμα 3.21: Ένα στιγμιότυπο της διεπαφής αναζήτησης του TARCLOUD



Σχήμα 3.22: Ένα στιγμιότυπο του απεικονιστή αποτελεσμάτων του TARCLOUD για το miRNA "my-miRNA-1"

3 μηχανήματα Εργατών (μηχανήματα Azure μεσαίου μεγέθους) και 1 μηχάνημα Γενικού Εργάτη. Τα αποτελέσματα επιβεβαιώθηκαν με σύγκριση με το προϊόν εξόδου της αυθεντικής υλοποίησης του DIANA microT για τις ίδιες ακολουθίες miRNA και φαίνονται στον Πίνακα 3.5.

Σύστημα	Μέσος χρόνος εκτέλεσης
Αρχική υλοποίηση	37 μιν
TarCloud	5 μιν

Πίνακας 3.5: Μετρήσεις για την απόδοση του TarCloud.

3.2.2 MR-microT: πρόβλεψη στόχων με χρήση MapReduce

3.2.2.1 Κίνητρο

Το TarCloud εισάγει μόνο τετριμμένη παραλληλοποίηση στην πρόβλεψη των στόχων για ένα δεδομένο miRNA. Συγκεκριμένα, το βήμα στοίχισης ακολουθιών χωρίζεται σε δυο μη συνδεόμενα σύνολα υπολογισμών ενώ η ανάλυση διατηρησιμότητας χωρίζεται μόνο σε λίγα μη συνδεόμενα σύνολα υπολογισμών (βάσει του αριθμού των χρωμοσωμάτων). Ως αποτέλεσμα, ο χρόνος εκτέλεσης για ένα *μόνο* miRNA δεν μπορεί να επιταχυνθεί σημαντικά και έτσι δεν μπορεί να κλιμακωθεί με τη χρήση περισσότερων κόμβων επεξεργασίας. Επιπλέον, η υλοποίηση του TarCloud εξαρτάται από την πλατφόρμα Azure, καθιστώντας αδύνατη τη μεταφορά του συστήματος σε έναν άλλο πάροχο Νέφους ή σε μια ιδιωτική συστάδα, αν χρειαστεί.

Για τους παραπάνω λόγους, δημιουργήσαμε το σύστημα *MR-microT* [40], που είναι μια προσαρμογή της μεθόδου microT βασισμένη στο MapReduce. Το βασικό χαρακτηριστικό του MR-microT είναι ο παραλληλισμός της διαδικασίας πρόβλεψης, ώστε ο χρόνος εκτέλεσης για ένα *μόνο* miRNA να μπορεί να επιταχυνθεί όπως είναι επιθυμητό, χρησιμοποιώντας περισσότερους πόρους. Αυτό επιτυγχάνεται με τον προσεκτικό χωρισμό των δεδομένων εισόδου (γονιδίωμα και πληροφορίες διατηρισιμότητας σε διάφορα είδη) μεταξύ διαφόρων κόμβων επεξεργασίας. Αυτή η πρόσέγγιση επιτυγχάνει επίσης την επιτάχυνση της εκτέλεσης για *πολλά* miRNA. Τέλος, το MR-microT έρχεται με μια διαισθητική και ισχυρή διεπαφή Ιστού[7] που μπορεί να χρησιμοποιηθεί από ερευνητές για την παραγωγή προβλέψεων στόχων για τυχαίες ακολουθίες miRNA. Ως αποτέλεσμα, το MR-microT είναι η πρώτη υλοποίηση microT που μπορεί να προβλέψει σε σχεδόν πραγματικό χρόνο ad hoc στόχους miRNA. Καθώς αρκετές επιλογές σχεδιασμού στο MR-microT θα μπορούσαν να εφαρμοστούν σε άλλες παρόμοιες υπολογιστικά απαιτητικές μεθόδους της βιοπληροφορικής, η φιλοδοξία μας για το MR-microT είναι να χρησιμεύσει σαν ένα υποδειγματικό σύστημα στον κλάδο.

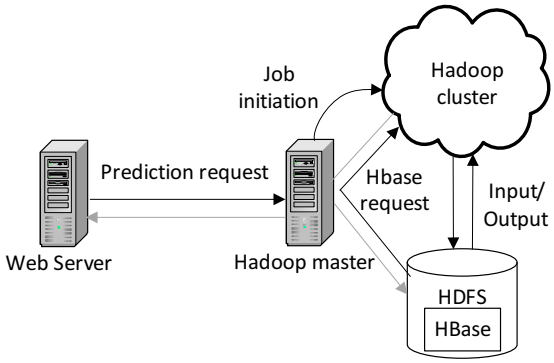
3.2.2.2 Τεχνολογίες και αρχιτεκτονική συστήματος

Το MR-microT είναι μια κατανεμημένη υλοποίηση της μεθόδου πρόβλεψης στόχων microT, σχεδιασμένη να τρέχει σε υπολογιστικές συστάδες αυθαίρετων μεγεθών. Χρησιμοποιεί το πλαίσιο Hadoop[8] για την κατανομή των υπολογισμών του microT στους κόμβους της συστάδας. Το Hadoop είναι ένα πλαίσιο ανοιχτού κώδικα που υλοποιεί το κατανεμημένο προγραμματιστικό παράδειγμα του *MapReduce* [18] και μπορεί να εγκατασταθεί εύκολα σε σχεδόν κάθε προσωρινή υπολογιστική συστάδα. Η εκτέλεση ενός προγράμματος MapReduce αποτελείται από τις φάσεις *Map* και *Reduce*. Κατά τη φάση Map, η είσοδος χωρίζεται σε μη συνδεόμενα κομμάτια και κάθε κομμάτι υφίσταται ξεχωριστή επεξεργασίας από ένα διαφορετικό κόμβο της συστάδας για την παραγωγή ενός συνόλου ζευγών κλειδιού-τιμής. Η φάση Reduce ομαδοποιεί αυτά τα ζεύγη κλειδιού-τιμής και επεξεργάζεται τις τιμές τους μαζί για να παράγει το προϊόν εξόδου του προγράμματος. Χρησιμοποιούμε το εργαλείο Hadoop streaming[9] για να εκμεταλλευτούμε μεγάλες ποσότητες υπαρχόντων κωδίκων Perl γραμμένων από βιοπληροφορικούς.

[7] `http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=mrmicrot`

[8] `http://hadoop.apache.org/`

[9] `http://hadoop.apache.org/docs/stable1/streaming.html`



Σχήμα 3.23: Αρχιτεκτονική συστήματος του MR-microT

Τα περισσότερα από τα δεδομένα που απαιτούνται από την εκτέλεση του microT αποθηκεύονται στο Κατανεμημένο Σύστημα Αρχείων του Hadoop (Hadoop Distributed File System - HDFS). Αυτό το κατανεμημένο σύστημα αρχείων βρίσκεται στους σκληρούς δίσκους των κόμβων της συστάδας και χρησιμοποιείται για την αποθήκευση της εισόδου και της εξόδου κάθε κώδικα Hadoop. Χρησιμοποιούμε το HBase[10], μια μη σχεσιακή κατανεμημένη βάση δεδομένων εμπνευσμένη από το BigTable [14] της Google, για την αποθήκευση δεδομένων που πρέπει να ανακτώνται ad hoc κατά την εκτέλεση.

Η συστάδα που φιλοξενεί το MR-microT είναι ένα Εικονικό Δίκτυο που αποτελείται από 19 Εικονικές Μηχανές Ubuntu, που παρέχονται από την υπηρεσία Νέφους του ∼okeanos[11] [46]. Το ∼okeanos είναι μια πλατφόρμα IaaS που παρέχει πόρους στην Ελληνική ερευνητική και ακαδημαϊκή κοινότητα.

Η αρχιτεκτονική του MR-microT απεικονίζεται στο Σχήμα 3.23. Συγκεκριμένα, το MR-microT αποτελείται από (α) έναν εξυπηρετητή Ιστού, που συλλέγει τα αιτήματα των χρηστών για πρόβλεψη στόχων ακολουθιών miRNA, (β) μια συστάδα Hadoop, στους κόμβους της οποίας εκτελείται ο κώδικάς μας, και (γ) ένα Hadoop master, το οποίο είναι μια Εικονική Μηχανή υπεύθυνη για τη διαχείριση της συστάδας Hadoop και των πόρων αποθήκευσης των HDFS και HBase.

Ο εξυπηρετητής ιστού είναι μια Εικονική Μηχανή Ubuntu που φέρει ένα μηχάνημα του εξυπηρετητή Apache και έχει τη διεπαφή του χρήστη του MR-microT (βλ. Κεφάλαιο 3.2.2.3). Αυτό το front-end είναι γραμμένο σε PHP και συλλέγει αιτήματα HTTP του χρήστη στο σύστημα του MR-microT για δεδομένες ακολουθίες miRNA. Ο εξυπηρετητής Ιστού αποστέλλει αυτά τα αιτήματα στο Hadoop master, που κατανέμει τους υπολογισμούς που απαιτούν τα αιτήματα σε κόμβους της συστάδας Hadoop. Κάθε κόμβος διαβάζει αρχεία εισόδου από το HDFS κατά την αρχική φάση και ζητάει καταχωρήσεις HBase όταν χρειαστεί. Να σημειωθεί ότι ο κύριος κόμβος είναι επίσης ο πληρεξούσιος του HDFS και ένας από τους εξυπηρετητές των αιτημάτων HBase.

Καθώς οι κόμβοι της συστάδας εκτελούν τις εργασίες τους, ο κύριος κόμβος παρακολουθεί την πρόοδό τους. Ο εξυπηρετητής Ιστού απευθύνεται στον κύριο κόμβο για αναφορές προόδου και τις παρουσιάζει στο χρήστη σε μορφή φιλική προς τον άνθρωπο. Η τελευταία αναφορά προόδου του κυρίου κόμβου ενημερώνει τον εξυπηρετητή Ιστού ότι η διαδικασία έχει ολοκληρωθεί, και, έπειτα, τα δεδομένα εξόδου μεταφέρονται από το HDFS στον εξυπηρετητή Ιστού προκειμένου να παρουσιαστούν στο χρήστη.

Υπάρχουν τρία βασικά βήματα εκτέλεσης στη μέθοδο microT για την εξεύρεση των στόχων ενός μορίου miRNA για ένα συγκεκριμένο είδος. Πρώτον, παράγεται ένα

[10] `https://hbase.apache.org/`
[11] `https://okeanos.grnet.gr/`

σύνολο υποψήφιων στόχων στοιχίζοντας τα πρώτα 9 ζεύγη βάσεων της ακολουθίας miRNA εντός της περιοχής που κωδικοποιεί την πρωτεΐνη και την περιοχή 3′-UTR κάθε ακολουθίας γονιδίων αυτού του είδους. Ένα γονίδιο λέγεται υποψήφιος στόχος εάν περιέχει τουλάχιστον μια τοποθεσία στην περιοχή που κωδικοποιεί την πρωτεΐνη του ή στις περιοχές 3′-UTR όπου αυτά τα 9 ζεύγη βάσεων μπορούν να στοιχιστούν. Αυτές οι τοποθεσίες λέγονται *στοιχεία αναγνώρισης miRNA (miRNA recognition elements - MREs)* και είναι υποψήφια σημεία πρόσδεσης για το μόριο miRNA μέσα στο στόχο. Για κάθε MRE, καταγράφεται η ποιότητα στοίχισης ως το *σκορ στοίχισής* του. Στο δεύτερο βήμα, υπολογίζεται ένα *σκορ διατηρησιμότητας* για κάθε MRE βάσει του αριθμού ειδών στα οποία διατηρείται το MRE. Τέλος, στο τρίτο βήμα, υπολογίζεται ένα αθροιστικό σκορ για κάθε γονίδιο του είδους, λαμβάνοντας υπόψη τα υπολογισμένα σκορ για όλα τα MRE του, καθώς και κάποιους άλλους παράγοντες, όπως η αναδίπλωση των εμπλεκόμενων μορίων.

Ο σχεδιασμός MapReduce που έχουμε κάνει βασίζεται στην ακόλουθη παρατήρηση. Είναι πιθανό να εκτελέσουμε τους υπολογισμούς που εμπλέκονται στα δυο πρώτα βήματα της μεθόδου microT ανεξάρτητα για κάθε γονίδιο. Οπότε, επιλέγουμε να αναθέσουμε αυτές τις ανεξάρτητες εργασίες σε διαφορετικούς κόμβους της συστάδας (βέβαια, εάν ο συνολικός αριθμός εργασιών είναι μεγαλύτερος από τον αριθμό των διαθέσιμων κόμβων, τότε κάποιοι κόμβοι μπορεί να χρειαστεί να εκτελέσουν διαδοχικά πολλαπλές εργασίες). Μια τέτοια κατανομή εργασιών μπορεί να πραγματοποιηθεί υλοποιώντας τα δυο πρώτα βήματα της μεθόδου microT σαν την φάση Map ενός κώδικα Hadoop. Η είσοδος που απαιτείται σε αυτή τη φάση Map αποθηκεύεται στο HDFS σαν ένα σύνολο από καταχωρήσεις, όπου κάθε μια περιέχει δεδομένα σχετικά με μια περιοχή (που κωδικοποιεί την πρωτεΐνη ή περιοχή 3′-UTR) ενός συγκεκριμένου γονιδίου. Η έξοδος από τη φάση Map θα ήταν ένα σύνολο καταχωρήσεων επίσης, όπου κάθε καταχώρηση κωδικοποιεί την τοποθεσία ενός MRE, τα σκορ που έχουν υπολογιστεί για αυτό, και το αναγνωριστικό του στόχου, δηλαδή, το γονίδιο, στο οποίο βρίσκεται.

Το τρίτο βήμα του microT συνίσταται στον υπολογισμό ενός τελικού σκορ για κάθε στόχο συνδυάζοντας τα σκορ στοίχισης και διατηρησιμότητας όλων των MRE που βρίσκονται σε αυτό. Οπότε, στο MR-microT, αυτό το βήμα υλοποιείται σαν φάση Reduce, καταναλώνοντας την έξοδο της φάσης Map που περιγράφεται ανωτέρω. Για όλες τις καταγραφές που έχουν το ίδιο αναγνωριστικό στόχου, η φάση Reduce υπολογίζει το τελικό σκορ λαμβάνοντας υπόψη επίσης τη μοριακή αναδίπλωση. Περισσότερες πληροφορίες για τη φάση Map όσο και για τη φάση Reduce του MR-microT ακολουθούν.

Η φάση Map παίρνει την περιοχή που κωδικοποιεί την πρωτεΐνη ή την περιοχή 3′-UTR μιας ακολουθίας γονιδίων μαζί με κάποιες πληροφορίες σχετικά με τη διατήρησή της σε ένα προκαθορισμένο σύνολο ειδών, και παράγει τα MRE εκείνης της ακολουθίας μαζί με τα σκορ στοίχισης και διατηρησιμότητάς της. Η είσοδος οργανώνεται σε αρχεία αποθηκευμένα στο HDFS. Κάθε σειρά αυτών των αρχείων αντιστοιχεί σε μια περιοχή (που κωδικοποιεί την πρωτεΐνη ή περιοχή 3′-UTR) ενός γονιδίου και δομείται ως ζεύγος κλειδιού-τιμής. Το κλειδί αυτού του ζεύγους είναι η συγκέντρωση του του αναγνωριστικού του γονιδίου από την Ensembl με ένα αλφαριθμητικό που δηλώνει τον τύπο της περιοχής (που κωδικοποιεί την πρωτεΐνη ή περιοχή 3′-UTR). Η τιμή του ζεύγους είναι μια δομή που περιέχει την ακολουθία της περιοχής του γονιδίου μαζί με κάποιες βασικές πληροφορίες για το γονίδιο και τη διατηρησιμότητά του σε ένα πλήθος ειδών.

Για κάθε ζεύγος κλειδιού-τιμής, ο κώδικας Map εκτελεί πρώτα έναν αλγόριθμο

στοίχισης ακολουθιών για να βρει όλες τις στοιχίσεις των πρώτων 9 ζευγών βάσεων της ακολουθίας miRNA στην ακολουθία της περιοχής του γονιδίου που περιέχεται στην τιμή του ζεύγους. Κάθε στοίχιση αυτών των ζευγών βάσεων που βρίσκεται είναι ένα MRE και υπολογίζεται ένα σκορ στοίχισης για αυτό βάσει της ποιότητας της στοίχισης και της ισχύος των δεσμών που πρόκειται να δημιουργηθούν σε περίπτωση μοριακής πρόσδεσης.

Τότε, οι πληροφορίες διατηρησιμότητας από την ακολουθία του γονιδίου λαμβάνονται υπόψη και υπολογίζεται ένα σκορ διατηρησιμότητας. Όσο περισσότερα είναι τα είδη που διατηρούν την ακολουθία γονιδίων απαράλλακτη, τόσο μεγαλύτερο είναι το σκορ διατηρησιμότητας που προκύπτει. Σημειώστε ότι, ο υπολογισμός αυτού του σκορ εξαρτάται από προϋπολογισμένα βάρη για κάθε πιθανό ζεύγος των 3-gram (ένα ζεύγος αποτελείται από ένα gram από το είδος αναφοράς και ένα gram από ένα άλλο είδος). Αυτά τα βάρη απαιτούνται ad hoc κατά την εκτέλεση της φάσης Map και αποθηκεύονται έτσι στο HBase.

Τέλος, ένα σύνολο ζευγών κλειδιού-τιμής, ένα για κάθε ευρεθέν MRE, παράγεται σαν έξοδος από τη φάση Map. Το κλειδί κάθε ζεύγους είναι το αναγνωριστικό της Ensembl για το γονίδιο στο οποίο βρίσκεται το MRE και η τιμή κωδικοποιεί την τοποθεσία ενός MRE μαζί με τα υπολογισμένα σκορ στοίχισης και διατηρησιμότητάς του.

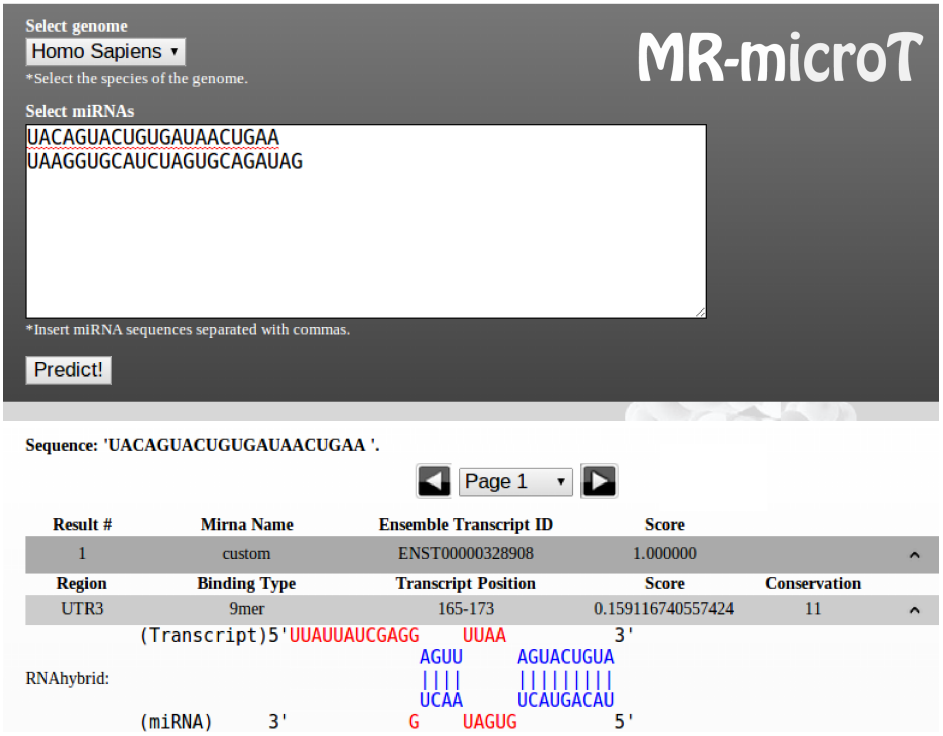
Η φάση Reduce καταναλώνει την έξοδο από τη φάση Map. Το πλαίσιο Hadoop εξασφαλίζει ότι όλα τα ζεύγη κλειδιού-τιμής που έχουν το ίδιο κλειδί (δηλαδή το ίδιο αναγνωριστικό γονιδίου) θα υποστούν επεξεργασία από τον ίδιο κόμβο της συστάδας. Αυτός ο κόμβος αθροίζει τα σκορ στοίχισης και διατηρησιμότητας όλων των MRE που έχουν βρεθεί για αυτό το γονίδιο, και βγάζει ένα σκορ πρόβλεψης για το ίδιο το γονίδιο. Κατά τον υπολογισμό αυτό, η φάση Reduce λαμβάνει υπόψη της επίσης την αναδίπλωση των εμπλεκόμενων μορίων (αυτό είναι σημαντικό καθώς η αναδίπλωση θα μπορούσε στην ουσία να καταστρέψει μια πιθανή πρόσδεση). Σημειώστε ότι οι πληροφορίες αναδίπλωσης αποθηκεύονται στο HBase και απαιτούνται ad hoc κατά την εκτέλεση. Η έξοδος της φάσης Reduce αποθηκεύεται στο HDFS και ο κύριος κόμβος του Hadoop ενημερώνεται για την τοποθεσία αποθήκευσης.

3.2.2.3 Διεπαφή χρήστη και αξιολόγηση

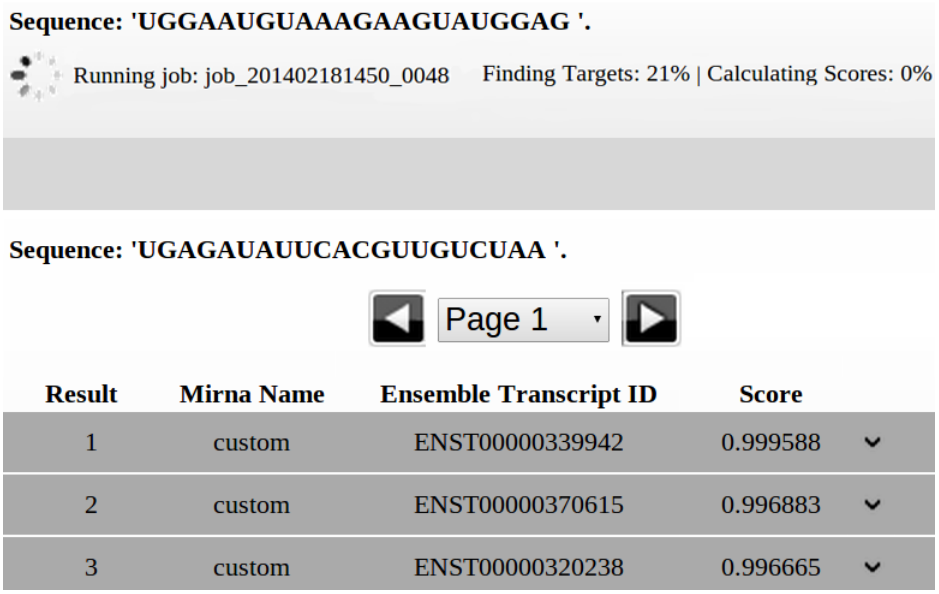
Οι υπηρεσίες του MR-microT είναι προσβάσιμες μέσω μιας ισχυρής και διαισθητικής διεπαφής Ιστού. Ένα στιγμιότυπο της διεπαφής παρουσιάζεται στο Σχήμα 3.24. Πρώτα, ο χρήστης επιλέγει το είδος στο οποίο θέλει να βρει στόχους από μια αναπτυσσόμενη λίστα. Αυτή τη στιγμή, υπάρχουν δυο διαθέσιμες επιλογές: Homo Sapiens (δηλαδή ο άνθρωπος) και Mus Musculis (δηλαδή το ποντίκι). Μετά, εισάγει μια ή περισσότερες ακολουθίες miRNA σε ένα πλαίσιο κειμένου.

Όταν καθοριστούν τα είδη και οι ακολουθίες miRNA, ο χρήστης μπορεί να στείλει αυτό το αίτημα πρόβλεψης στο σύστημα. Αυτό γίνεται πατώντας το κουμπί που έχει τίτλο 'Predict!'. Μετά από αυτό, ο χρήστης έχει την ευκαιρία να παρακολουθήσει τη διαδικασία πρόβλεψης στόχων για *κάθε* μια από τις ακολουθίες που έχει θέσει. Το σύστημα εμφανίζει μια χωριστή *θέαση προόδου* για κάθε ακολουθία miRNA μέσα στην οποία απεικονίζεται η πρόοδος της πρόβλεψης στόχων. Ένα παράδειγμα θέασης προόδου φαίνεται στο πάνω μέρος του Σχήματος 3.25.

Όταν ολοκληρωθεί η μέθοδος πρόβλεψης στόχων για μια ακολουθία miRNA, η θέαση προόδου εμπλουτίζεται με τη θέαση των αποτελεσμάτων, που φαίνεται στο κάτω



Σχήμα 3.24: Στιγμιότυπο της διεπαφής χρήστη του MR-microT



Σχήμα 3.25: Θέαση προόδου του MR-microT, που δείχνει την τρέχουσα κατάσταση της πρόβλεψης στόχων ενός μόνο miRNA, και θέαση αποτελεσμάτων που παρουσιάζει τη λίστα των προβλεπόμενων στόχων

μέρος του Σχήματος 3.25, η οποία περιέχει πληροφορίες για τους υπολογισμένους στόχους. Καθώς ο αριθμός στόχων είναι συνήθως μεγάλος, οργανώνονται σε σελίδες.

Κάθε σελίδα δείχνει πληροφορίες σχετικά με έναν αριθμό στόχων. Οι πληροφορίες για το στόχο βρίσκονται μέσα στα γκρίζα πλαίσια που φαίνονται στο Σχήμα ;;. Κάθε πλαίσιο πληροφοριών περιέχει το αναγνωριστικό γονιδίου και το τελικό σκορ πρόβλεψης στόχων που υπολογίζεται για αυτό το γονίδιο. Λεπτομέρειες σχετικά με τη λίστα των προβλεπόμενων σημείων πρόσδεσης του στόχου μπορούν να βρεθούν πατώντας στο

βελάκι που βρίσκεται στα δεξιά του γκρίζου πλαισίου.

Εκτελέσαμε κάποια πειράματα για να αξιολογήσουμε την απόδοση του MR-microT. Ο Πίνακας 3.6 συνοψίζει το μέσο χρόνο εκτέλεσης που απαιτείται από το MR-microT για να παράγει τους στόχους ενός δεδομένου miRNA στο γονιδίωμα του ανθρώπου και του ποντικιού. Και για τις δυο μετρήσεις, έχουμε δοκιμάσει τρία διαφορετικά μεγέθη συστάδων, ένα με 19 κόμβους, ένα άλλο με 10 κόμβους, και ένα τελευταίο με ένα κόμβο. Κάθε κόμβος είναι μια VM που παρέχεται από την υπηρεσία ~okeanos και περιέχει δυο πυρήνες.

Γονιδίωμα	1 **κόμβος**	10 **κόμβοι**	19 **κόμβοι**
Άνθρωπος	48λ 4δ	4λ 3δ	1λ 58δ
Ποντίκι	44λ 30δ	3λ 55δ	1λ 44δ

Πίνακας 3.6: Χρόνος εκτέλεσης του MR-microT για διάφορα μεγέθη συστάδων για το γονιδίωμα του ανθρώπου και του ποντικιού.

Είναι προφανές ότι ο χρόνος εκτέλεσης ανά miRNA μειώνεται γραμμικά καθώς αυξάνεται το μέγεθος της συστάδας. Αυτή είναι η βασική διαφορά από το TarCloud, που δεν μπορεί να μειώσει το χρόνο εκτέλεσης ανά miRNA κάτω από τα 5 λεπτά προσθέτοντας περισσότερους υπολογιστικούς κόμβους. Αντίθετα, το MR-microT θα μπορούσε να πετύχει ακόμα μικρότερο χρόνο εκτέλεσης εάν χρησιμοποιούσε περισσότερους πόρους. Υπάρχει φυσικά ένα κατώτατο όριο, επειδή δεν μπορούμε να χωρίσουμε την είσοδο κάθε διαδικασίας microT σε μικρότερα κομμάτια από το μέγεθος των γονιδίων, ωστόσο, είναι δυνατό να παράγουμε προβλέψεις στόχων για ένα δεδομένο miRNA σε λιγότερο από ένα λεπτό εκτελώντας το MR-microT σε μια συστάδα κατάλληλου μεγέθους.

Σχετικά με το σενάριο των πολλαπλών, ταυτόχρονων αιτημάτων χρηστών για την πρόβλεψη πολλών miRNA, το MR-microT μπορεί να εξυπηρετήσει όλα αυτά τα αιτήματα χωρίς παρατηρήσιμες καθυστερήσεις, με την προϋπόθεση οι πόροι που διαθέτει (δηλαδή, οι υπολογιστικοί πυρήνες) να μην έχουν εξαντληθεί. Διαφορετικά, θα παρατηρηθούν καθυστερήσεις, ωστόσο, εν τέλει, το σύστημα θα παρέχει αποτέλεσμα για όλα. Σημειώστε ότι χρησιμοποιώντας υπηρεσίες για ελαστική χρήση πόρων, είναι πιθανό να αποφευχθούν καθυστερήσεις όπως αναφέρθηκε ανωτέρω.

3.3 Συμπεράσματα

Χωρίς αποδοτικές και αξιόπιστες υπολογιστικές μεθόδους για πρόβλεψη στόχων miRNA, ο απαιτούμενος χρόνος και το απαιτούμενο κόστος για εκτέλεση βιοχημικών πειραμάτων τα οποία αποκαλύπτουν το ρόλο συγκεκριμένων miRNA στην ανάπτυξη και την θεραπεία σημαντικών ασθενειών θα ήταν τεράστια. Επιλέξαμε να επιταχύνουμε το DIANA microT μία από τις πιο δημοφιλείς και ακριβείς μεθόδους πρόβλεψης στόχων miRNA. Ο σκοπός μας ήταν να επιτύχουμε παραγωγή στόχων miRNA σε σχεδόν πραγματικό χρόνο.

Προς αυτή την κατεύθυνση, πρώτα μελετήσαμε τη διεργασία ταιριάσματος ακολουθιών που αποτελεί το πρώτο βήμα της μεθόδου. Βρήκαμε ότι εμπλέκει ένα νέο είδος ερωτήματος ταιριάσματος ακολουθιών. Μοντελοποιήσαμε μαθηματικά το προηγούμενο τύπο ερωτήματος εισάγοντας το πρόβλημα ARSM. Ο στόχος του προβλήματος αυτού είναι να ανακτηθούν όλες οι εμφανίσεις περιοχής ενός προτύπου μέσα σε μία ακολουθία

δεδομένων. Οι περιοχές του προτύπου που ταιριάζουν οφείλουν να περιέχουν μια προκαθορισμένη υπακολουθία του προτύπου, τον πυρήνα. Επιπλέον, η επιτρεπτή απόκλιση από την ακολουθία δεδομένων είναι πιο αυστηρή για μικρότερες και πιο χαλαρή για μεγαλύτερες περιοχές. Για να αντιμετωπίσουμε το προηγούμενο πρόβλημα προτείναμε τη μέθοδο PS-ARSM. Η μέθοδός μας εκμεταλλεύεται τις επικαλύψεις προθέματος και επιθέματος μεταξύ των περιοχών αποφεύγοντας επαναλαμβανόμενους υπολογισμούς. Μια αναλυτική πειραματική αξιολόγηση έδειξε ότι το PS-ARSM είναι περίπου δύο τάξεις μεγέθους γρηγορότερο από τις υπάρχουσες τεχνικές που μπορούν να μετασχηματιστούν στο πρόβλημα ARSM.

Παρόλα αυτά, η επιτάχυνση του βήματος ταιριάσματος ακολουθιών των μεθόδων πρόβλεψης στόχων δεν είναι αρκετό για να επιτευχθούν επιδόσεις σχεδόν πραγματικού χρόνου. Ο λόγς είναι επειδή αυτές οι μέθοδοι επίσης εμπλέκουν κάποιες υπολογιστικά έντονες διεργασίες (πχ τον υπολογισμό της αναδίπλωσης των εμπλεκόμενων μορίων, τον τρόπο με τον οποίο οι προβλεπόμενοι τόποι πρόσδεσης διατηρούνται σε διάφορα είδη κτλ). Ακολουθήσαμε την προσέγγιση να κατανείμουμε αυτούς τους υπολογισμούς στους κόμβους μιας υποδομής Νέφους. Προς αυτή την κατεύθυνση, σχεδιάσαμε δύο συστήματα πρόβλεψης στόχων βασισμένα στο Νέφος, το TarCloud και το MR-microT. Το πρώτο αναπτύχθηκε χρησιμοποιώντας το πλαίσιο Microsoft Azure, ενώ το δεύτερο χρησιμοποιώντας το πλαίσιο Hadoop. Με βάση τις μετρήσεις μας και τα δύο συστήματα επιταχύνουν τη διαδικασία πρόβλεψης, όμως το MR-microT είναι ανώτερο γιατί είναι ανεξάρτητο της πλατφόρμας (μπορεί να εγκατασταθεί σε οποιαδήποτε σύγχρονη συστάδα υπολογιστών), παρέχει βελτιωμένη παραλληλοποίηση των εμπλεκόμενων εργασιών και έχει σχεδιαστεί για να υποστηρίζει χωρίς προβλήματα μεγάλα πλήθη από αιτήματα πρόβλεψης.

Κεφάλαιο 4

Υποδομές για έρευνα πάνω στα miRNA

Στο παρόν κεφάλαιο, παρουσιάζουμε ένα σύνολο υποδομών που αναπτύχθηκαν για να στηρίξουν την έρευνα πάνω στα miRNA. Μέχρι πρόσφατα, σημαντικές πληροφορίες για τη λειτουργία και τη ρύθμιση κάθε miRNA ήταν διασκορπισμένες σε πολλές βάσεις δεδομένων ή δεν ήταν καν διαθέσιμες. Αυτό αποτέλεσε σημαντικό εμπόδιο για τους ερευνητές στις βιοεπιστήμες, οι οποίοι προσπαθούσαν να κατανοήσουν το ρόλο των miRNA σε πολλά βιολογικά μονοπάτια, γνώση που θα μπορούσε να βοηθήσει προς την ανακάλυψη θεραπειών για συγκεκριμένες ασθένειες. Στο Κεφάλαιο 4.1 αναφέρουμε κάποια ενδιαφέροντα δεδομένα και εργαλεία τα οποία θα μπορούσαν να είναι πολύτιμα για τους βιοεπιστήμονες στον τομέα της έρευνας πάνω στα miRNA. Η λίστα δημιουργήθηκε ρωτώντας ειδικούς του κλάδου.

Βάσει των καταγεγραμμένων αναγκών, και σε συνεργασία με μια ερευνητική ομάδα από το ΕΚΕΒΕ Αλ. Φλέμινγκ, αναπτύξαμε τα *εργαλεία DIANA*[1], ένα μεγάλο σύνολο δημόσια διαθέσιμων βάσεων δεδομένων και ηλεκτρονικών εργαλείων προοριζόμενων για την διευκόλυνση της έρευνας στις βιοεπιστήμες σχετικά με τα miRNA. Στο υπόλοιπο αυτού του κεφαλαίου, παρουσιάζουμε λεπτομερώς ένα προς ένα όλα αυτά τα εργαλεία, αναλύοντας το κίνητρο που ώθησε στη δημιουργία τους, τη λειτουργικότητά τους και τη συμβολή τους.

4.1 Ανάγκες για εργαλεία και δεδομένα σχετικά με τα miRNA

Υπάρχει πληθώρα αποθετηρίων που συλλέγουν και διανέμουν ενδιαφέρουσες πληροφορίες γενικού ενδιαφέροντος στις βιοεπιστήμες. Ένα ενδεικτικό παράδειγμα είναι το σύνολο πόρων που παρέχονται από την Ensembl και το NCBI. Ωστόσο, υπάρχει απουσία παρόμοιων πόρων στον κλάδο της έρευνας πάνω στα miRNA. Αν και το miRBase[2] είναι ένα κέντρο που συλλέγει ενδιαφέρουσες πληροφορίες σχετικά με κάθε υπάρχον μόριο miRNA, πολλά χρήσιμα δεδομένα σχετικά με τα miRNA είτε είναι διασκορπισμένα στις επιστημονικές δημοσιεύσεις είτε δεν υπάρχουν καν (καθώς επιπλέον ανάλυση θα έπρεπε να γίνει για να αποκαλυφθούν). Σε αυτό το κεφάλαιο, βασισμένοι σε ερωτήσεις προς ειδικούς του κλάδου, συλλέγουμε και καταγράφουμε κάποια δεδομένα

[1] `http://http://diana.imis.athena-innovation.gr`

[2] `http://www.mirbase.org`

σχετικά με τα miRNA τα οποία ανήκουν στην παραπάνω κατηγορία.

Οι πιο σημαντικές πληροφορίες σχετικά με ένα μόριο miRNA είναι η λίστα των γονιδίων που επηρεάζει η παρουσία του, δηλαδή, η λίστα των γονιδίων στόχων του. Ένα γονίδιο στόχος μπορεί να αποκαλυφθεί με βιοχημικά πειράματα. Τα αποτελέσματα αυτών των πειραμάτων δημοσιεύονται σε επιστημονικά περιοδικά. Οπότε, ένας ερευνητής που ενδιαφέρεται για τους στόχους ενός συγκεκριμένου miRNA θα έπρεπε να ψάξει σε όλη τη διαθέσιμη βιβλιογραφία για να βρει τα πειράματα που τον ενδιαφέρουν. Κάποιες επιμελημένες βάσεις δεδομένων που συλλέγουν τις αλληλεπιδράσεις miRNA-γονιδίου υπάρχουν, ωστόσο μπορούν να περιέχουν πολύ περιορισμένο αριθμό αλληλεπιδράσεων και συνήθως δεν ανανεώνονται σε τακτική βάση.

Πιστεύεται ότι υπάρχουν εκατομμύρια αλληλεπιδράσεις miRNA-γονιδίων. Ωστόσο, τα ήδη εκτελεσμένα βιοχημικά πειράματα έχουν αποκαλύψει μόνο ένα μικρό μέρος τους. Αυτό οφείλεται στο γεγονός ότι ένα σύνολο πειραμάτων σαν τα ανωτέρω απαιτεί σημαντικό χρόνο για να ολοκληρωθεί. Για το λόγο αυτό, υπολογιστικές μέθοδοι που προβλέπουν γρήγορα αλληλεπιδράσεις miRNA-γονιδίων είναι πολύτιμες καθώς παρέχουν γνώσεις για πιθανές σχέσεις μεταξύ miRNA και γονιδίων. Υπάρχει πληθώρα μεθόδων πρόβλεψης που έχουν προταθεί. Κάποιες από αυτές παρέχουν μια διεπαφή Ιστού που παρουσιάζει τους προβλεπόμενους στόχους όλων των γνωστών miRNA. Το DIANA microT είναι μια από τις πιο ακριβείς μεθόδους πρόβλεψης στόχων, ωστόσο, υστερεί στη διασπορά των προβλέψεών του καθώς δεν παρέχει καμία διεπαφή Ιστού για να τις παρουσιάσει.

Μια άλλη σημαντική πληροφορία για τα miRNA είναι η ακριβής γονιδιακή τοποθεσία των μεταγράφων τους και τα προφίλ έκφρασής τους. Πάλι αυτή η πληροφορία είναι διασκορπισμένη σε σχετικές επιστημονικές δημοσιεύσεις. Η αλληλεπίδραση των μεταγράφων miRNA με συγκεκριμένους μεταγραφικούς παράγοντες, η λίστα των ιστών μέσα στους οποίους εκφράζονται αυτά τα μετάγραφα, και τα μεταβολικά μονοπάτια στα οποία συμμετέχουν θα μπορούσαν να είναι επίσης ισχυρά κομμάτια πληροφοριών για τους βιοεπιστήμονες στον κλάδο της έρευνας πάνω στα miRNA.

Τέλος, καθώς υπάρχει τεράστιος αριθμός δημοσιεύσεων σχετικά με τα miRNA, ο οποίος συνεχώς αυξάνεται, προκύπτει η ανάγκη αναζήτησης για όλες τις δημοσιεύσεις που είναι σχετικές με ένα συγκεκριμένο μόριο. Αυτό δεν είναι απλή εργασία καθώς η ονομασία των miRNA χρησιμοποιείται ασυνεπώς στη βιβλιογραφία και ακόμα και τα επίσημα ονόματα miRNA μπορεί να αλλάξουν κατά καιρούς.

Στόχος μας είναι να αναπτύξουμε ένα σύνολο ισχυρών εργαλείων που θα καλύψουν το κενό πληροφόρησης που υπάρχει στους ανωτέρω τομείς. Στο Κεφάλαιο 4.2 παρουσιάζουμε διάφορες εκδόσεις του DIANA microT, ενός εξυπηρετητή Ιστού που παρέχει προβλέψεις για τα γονίδια που στοχεύονται από όλα τα γνωστά miRNA, βάσει της πολύ ακριβούς μεθόδου πρόβλεψης DIANA microT. Στο Κεφάλαιο 4.3 περιγράφουμε το DIANA miRGen, ένα σύστημα που ενημερώνει τους χρήστες του για τις γονιδιακές τοποθεσίες όλων των μεταγράφων miRNA και τη συμπεριφορά έκφρασής τους. Επιπλέον, στο Κεφάλαιο 4.4 εισάγουμε το DIANA TarBase, μια βάση δεδομένων που συλλέγει πειραματικά επικυρωμένους στόχους miRNA. Στο Κεφάλαιο 4.5, παρουσιάζουμε το DIANA mirPath, ένα εργαλείο που ερευνά το ρόλο των miRNA στα γνωστά μεταβολικά μονοπάτια. Τέλος, στο Κεφάλαιο 4.6, συζητάμε για το DIANA mirPub, ένα ηλεκτρονικό εργαλείο που βοηθά τους βιοεπιστήμονες στην αναζήτηση βιβλιογραφίας σχετικά με τα miRNA. Πιστεύουμε ότι τα παραπάνω εργαλεία είναι πολύτιμη προσθήκη στην εργαλειοθήκη οποιουδήποτε βιοεπιστήμονα στον κλάδο της έρευνας πάνω στα miRNA.

4.2 Εξυπηρετητής Ιστού DIANA microT: Αναζήτηση προβλεπόμενων στόχων miRNA

Στη δεκαετία του 2000, εισήχθη μεγάλος αριθμός μεθόδων πρόβλεψης στόχων miRNA [87, 2]. Οι περισσότερες βασίζονταν κυρίως στη στοίχιση ακολουθιών της περιοχής σπόρων miRNA (δηλαδή, νουκλεοτίδια $2-7$ από το 5′-άκρο του μορίου) με την περιοχή 3′-UTR των υποψήφιων γονιδίων στόχων για την αναγνώριση θεωρούμενων σημείων πρόσδεσης. Επιπλέον, για τη βελτίωση της ακρίβειάς τους, αυτές οι μέθοδοι συνήθιζαν να εκμεταλλεύονται κάποια επιπλέον χαρακτηριστικά όπως την εξελικτική διατηρησιμότητα των υποψήφιων σημείων πρόσδεσης, τη δομική διαθεσιμότητα των εμπλεκόμενων μορίων, τη σύνθεση των νουκλεοτιδίων κλπ. Για περισσότερες πληροφορίες σχετικά με τις τεχνικές που χρησιμοποιούν οι μέθοδοι πρόβλεψης στόχων miRNA βλ. επίσης Κεφάλαιο 2.2.2.

Τότε εισήχθη η πρώτη έκδοση της μεθόδου *microt* [43]. Το 2009, η ερευνητική ομάδα που ήταν υπεύθυνη για την ανάπτυξη του DIANA microT εργαζόταν στην τρίτη έκδοση της μεθόδου. Ξεκίνησαν συνεργασία με την ομάδα μας και, σε αυτό το πλαίσιο, ξεκινήσαμε να εργαζόμαστε πάνω σε (α) τεχνικές για βελτίωση της αποδοτικότητας της μεθόδου τους και (β) υποδομές έρευνας για την ώθηση των αποτελεσμάτων της. Το πρώτο αποτέλεσμα αυτής της συνεργασίας ήταν ο εξυπηρετητής Ιστού DIANA microT v.3 [59], μια υποδομή έρευνας που συλλέγει τους προβλεπόμενους στόχους όλων των γνωστών miRNA στα γονιδιώματα του ανθρώπου και του ποντικιού βάσει της μεθόδου DIANA microT. Αυτός ο εξυπηρετητής έγινε πολύ δημοφιλής (αριθμώντας περισσότερους από 150 μοναδικούς επισκέπτες ανά ημέρα) λόγω της ακρίβειας της μεθόδου DIANA microT v.3 και της μεγάλης ποικιλίας των χαρακτηριστικών που παρέχονται μέσω της καλά σχεδιασμένης διεπαφής του στον Ιστό. Κατά τα επόμενα χρόνια, δυο επιπλέον εκδόσεις του εξυπηρετητή DIANA microT δημοσιεύτηκαν [60, 76] κάνοντάς το ακόμα πιο δημοφιλή (πάνω από 500 μοναδικούς επισκέπτες ανά ημέρα). Στις ακόλουθες παραγράφους συζητάμε τη συμβολή των ανωτέρω εκδόσεων του εξυπηρετητή DIANA microT.

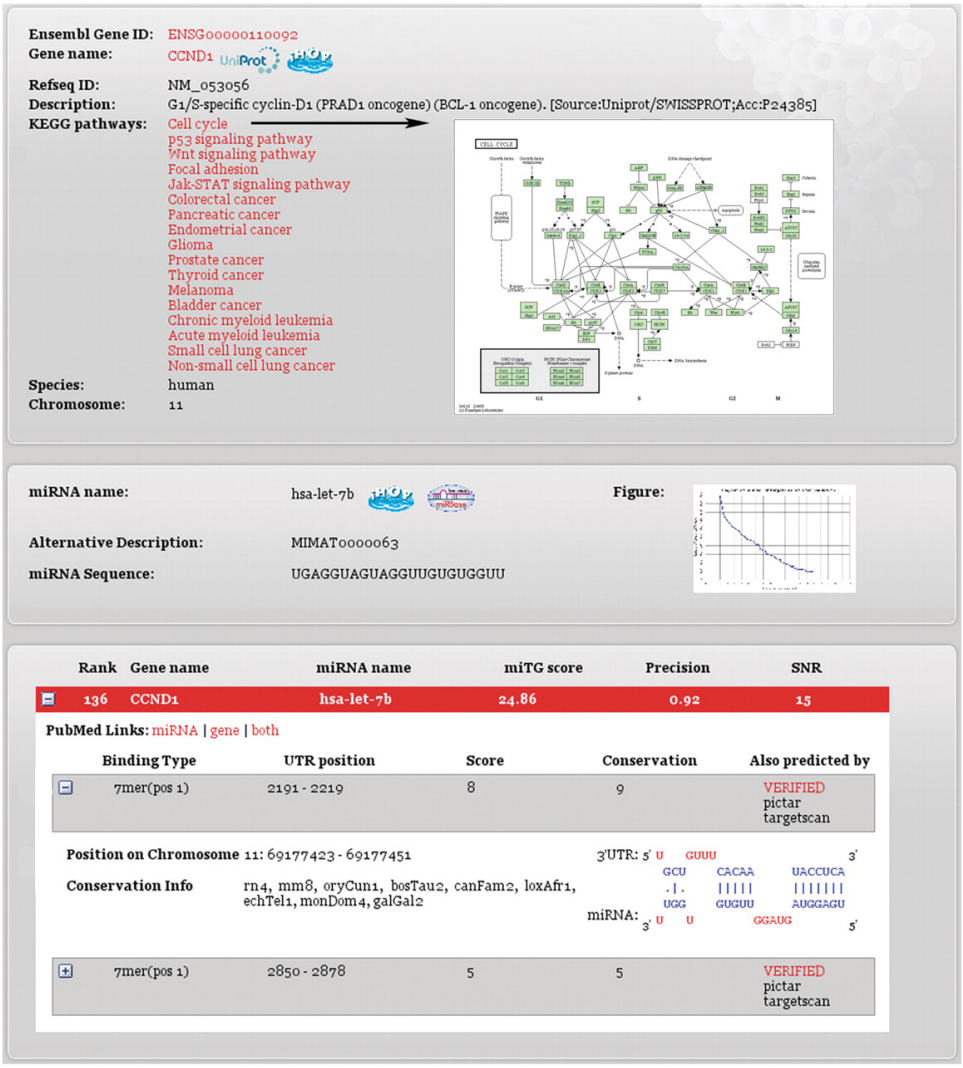
4.2.1 DIANA microT v.3

Η μέθοδος DIANA microT v.3 αναγνωρίζει τοποθεσίες στην περιοχή 3′-UTR του γονιδιώματος που πιθανόν αντιστοιχούν σε σημεία πρόσδεσης miRNA βάσει της στοίχισης ακολουθιών. Μετά, για κάθε ένα από αυτά τα προβλεπόμενα σημεία, υπολογίζεται ένα σκορ βάσει της ποιότητας της στοίχισης, της διατηρησιμότητας του σημείου σε πολλά είδη, και της αναδίπλωσης των εμπλεκόμενων μορίων. Τα σκορ όλων των προβλεπόμενων σημείων κάθε γονιδίου αθροίζονται για να παράγουν το σκορ $miTG$, ένα δείκτη της πιθανότητας το miRNA να στοχεύει όντως το γονίδιο. Για περισσότερες λεπτομέρειες σχετικά με τη μέθοδο βλ. επίσης Κεφάλαιο 2.2.2.1.

Ο εξυπηρετητής DIANA microT v.3 [58, 59] αναπτύχθηκε για να συλλέγει τα προβλεπόμενα σημεία στόχους όλων των γνωστών miRNA στα γονίδια του ανθρώπου και του ποντικιού και, μετά, να διαδώσει αυτή τη γνώση σε ερευνητές των βιοεπιστημών, μέσω μιας ισχυρής και διαισθητικής διεπαφής Ιστού. Ο χρήστης μπορεί να ψάξει για προβλέψεις σχετικά με ένα συγκεκριμένο miRNA, γονίδιο ή βιολογικό μονοπάτι[3] εισάγοντας λέξεις κλειδιά που περιγράφουν αυτές τις οντότητες σε πλαίσια αναζήτησης.

[3]Τα μονοπάτια της KEGG (Kyoto Encyclopedia of Genes and Genomes) λήφθησαν υπόψη [37].

Κάθε προβλεπόμενη αλληλεπίδραση miRNA-γονιδίου πέρα από το σκορ miTG, έχει μαζί ένα σκορ του λόγου σήματος προς θόρυβο (SNP) και ένα σκορ ακριβείας. Αυτά τα σκορ μπορούν να χρησιμοποιηθούν ως δείκτες για την ορθότητα κάθε πρόβλεψης. Επιπλέον, κάθε προβλεπόμενη αλληλεπίδραση παρουσιάζεται μαζί με συνδέσμους σε εξωτερικούς βιολογικούς πόρους και σε σχετικές δημοσιεύσεις. Αλληλεπιδράσεις που προβλέπονται από άλλες μεθόδους πρόβλεψης ή αλληλεπιδράσεις επιβεβαιωμένες από βιοχημικά πειράματα μαρκάρονται ανάλογα στη διεπαφή. Τέλος, ο εξυπηρετητής βρίσκεται στην ίδια πλατφόρμα με άλλα δυο λειτουργικά εργαλεία ανάλυσης miRNA: το mirPath, ένα εργαλείο ανάλυσης μονοπατιών προβλεπόμενων στόχων και το mirExTra, ένα εργαλείο ανάλυσης miRNA βάσει διαφορικά εκφρασμένων προφίλ mRNA. Το Σχήμα 4.1 παρουσιάζει ένα στιγμιότυπο της σελίδας αποτελεσμάτων του DIANA microT v.3 για τη συνδυαστική αναζήτηση ενός miRNA ("hsa-let-7a") και ενός γονιδίου ("CCND1").



Σχήμα 4.1: Στιγμιότυπο της διεπαφής του DIANA microT v.3 καθώς δείχνει τους προβλεπόμενους στόχους του miRNA "hsa-let-7a" στο γονίδιο "CCND1".

Το DIANA microT v.3 υποστηρίζει επίσης αιτήματα πρόβλεψης για ακολουθίες miRNA καθορισμένες από το χρήστη. Τα αποτελέσματα των de novo προβλέψεων αποθηκεύονται σε μια βάση δεδομένων από την οποία μπορούν να ανακτηθούν αργότερα

και να παρουσιαστούν στο χρήστη ο οποίος λαμβάνει έναν κωδικό μέσω ειδοποίησης στο ηλεκτρονικό ταχυδρομείο. Αυτό το σχήμα ήταν απαραίτητο επειδή, τότε, η de novo πρόβλεψη για μια ακολουθία miRNA ήταν υπολογιστικά απαιτητική και χρειαζόταν πάνω από 20 λεπτά για να ολοκληρωθεί σε μια συστάδα των 32 κόμβων (256 πυρήνων) που βρίσκεται στο Εθνικό Μετσόβιο Πολυτεχνείο (ΕΜΠ). Υπενθυμίζουμε ότι προσεγγίσεις Νέφους χρησιμοποιήθηκαν για να επιτύχουν de novo προβλέψεις σε σχεδόν πραγματικό χρόνο (βλ. Κεφάλαια 3.2.1 και 3.2.2).

Να σημειωθεί ότι όλες οι διεπαφές χρήστη του εξυπηρετητή Ιστού του DIANA microT v.3 αναπτύχθηκαν με χρήση PHP, ενώ όλα τα απαραίτητα δεδομένα αποθηκεύτηκαν σε μια σχεσιακή βάση δεδομένων (χρησιμοποιήθηκε η MySQL). Το ΕΜΠ φιλοξένησε το εργαλείο Ιστού στο `http://diana.cslab.ece.ntua.gr/microT/`.

4.2.2 DIANA microT v.4

Ο εξυπηρετητής Ιστού DIANA microT v.4 [60] είναι μια εκτεταμένη ανανέωση του DIANA microT v.3 με πολλές σημαντικές βελτιώσεις:

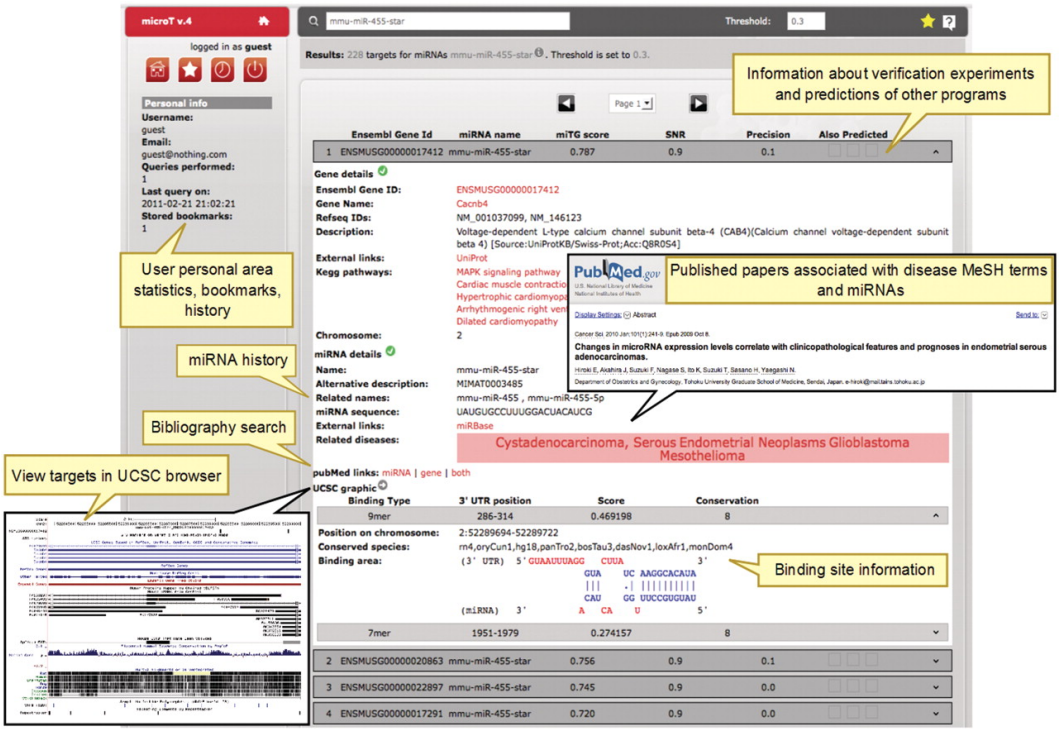
- Συσχέτιση των miRNA με ασθένειες βάσει προχωρημένης βιβλιογραφικής ανάλυσης
- Υποστήριξη προβλέψεων για δυο επιπλέον είδη (Drosophila melanogaster και C. elegans)
- Γραφική αναπαράσταση με όλες τις σχετικές λειτουργικές πληροφορίες από το UCSC genome browser,
- Υποστήριξη στα παλιά ονόματα των miRNA μέσω παρακολούθησης των αλλαγών στην ονοματοδοσία των miRNA
- Εξατομικευμένες συνεδρίες χρήστη που επιτρέπουν προσωπικό ιστορικό ερωτημάτων και σελιδοδείκτες

Ο εξυπηρετητής Ιστού DIANA microT v.4 παρέχει λειτουργική ανάλυση των miRNA που ξεπερνά τα όρια της απλής καταγραφής των στόχων των miRNA. Αυτό επιτυγχάνεται μέσω της ενοποίησης της γνώσης που εξάγεται τόσο από την επιστημονική βιβλιογραφία όσο και από τις πληροφορίες που περιλαμβάνονται σε γνωστά βιολογικά μονοπάτια. Προς αυτή την κατεύθυνση, ο εξυπηρετητής Ιστού παρέχει συνδέσεις των miRNA με ασθένειες, βάσει κοινής ανάλυσης κειμένου που πραγματοποιείται σε τίτλους και περιλήψεις δημοσιεύσεων του PubMed. Συγκεκριμένα, ένα miRNA θεωρούνταν ότι σχετίζεται με μια ασθένεια εάν υπάρχει τουλάχιστον μια δημοσίευση που (α) περιέχει το όνομα του miRNA ή το όνομα της οικογένειάς του στον τίτλο ή την περίληψη και (β) είναι σχολιασμένη με έναν όρο MeSH[4] που περιγράφει την ασθένεια. Όλες οι ασθένειες MeSH που εντοπίζεται ότι σχετίζονται με ένα miRNA οπτικοποιούνται μέσω ενός νέφους ετικετών που βοηθά το χρήστη να κατανοήσει τη γνώση που περιγράφεται στη βιβλιογραφία. Οι όροι MeSH στο νέφος ετικετών χρησιμεύουν επίσης ως υπερσυνδέσεις στις σχετικές δημοσιεύσεις.

Ένα άλλο σημαντικό νέο χαρακτηριστικό του εξυπηρετητή Ιστού DIANA microT v.4 είναι το γεγονός ότι συλλέγει προβλεπόμενους στόχους για δυο ακόμα είδη. Συγκεκριμένα, η πρώτη έκδοση του εξυπηρετητή σχεδιάστηκε για να υποστηρίζει τη

[4]Medical Subject Headings που παρέχονται από το National Library of Medicine.

λειτουργική ανάλυση των miRNA του ανθρώπου και του ποντικιού. Ο εξυπηρετητής DIANA microT v.4 έχει ανανεωθεί για να περιέχει επίσης προβλέψεις για τα Drosophila melanogaster και C. elegans. Επιπλέον, δεδομένα από πιο πρόσφατες εκδόσεις του miRBase (miRBase 13) και της Ensembl (Ensembl 54) συμπεριλήφθησαν στη βάση δεδομένων του νέου εξυπηρετητή Ιστού. Συνολικά, προστέθηκαν προβλέψεις για 723 καινούρια miRNA, τα 147 εκ των οποίων αντιστοιχούν στο Drosophila melanogaster και τα 154 στο C. elegans (τα υπόλοιπα είναι νέα miRNA του Homo sapiens και του Mus musculus). Αυτό οδηγεί σε περίπου διπλάσιο αριθμό προβλεπόμενων στόχων σε σύγκριση με το DIANA microT v.3, αριθμώντας περισσότερα από έξι εκατομμύρια προβλεπόμενα γονίδια στόχους. Τέλος, ενώ το DIANA microT v.3 βασίζεται σε χαρακτηριστικά που διακρίνουν τα αλληθινά από τα ψευδή (τυχαία) miRNA (βλ. επίσης Κεφάλαιο 2.2.2.1) το DIANA microT v.4 χρησιμοποιεί πειραματικά δεδομένα υψηλής απόδοσης για τον ίδιο σκοπό.

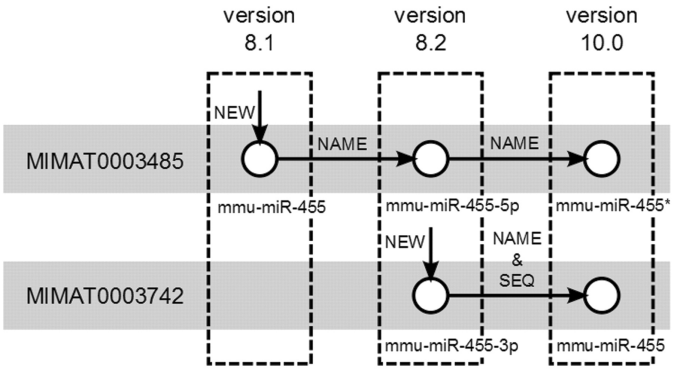


Σχήμα 4.2: Ένα στιγμιότυπο της διεπαφής DIANA microT v.4 καθώς δείχνει τους προβλεπόμενους στόχους του miRNA "mmu-miR-455-star".

Η διεπαφή του εξυπηρετητή Ιστού DIANA microT ανασχεδιάστηκε πλήρως και έγινε πιο διαισθητική (βλ. Σχήμα 4.2). Η πιο σημαντική αλλαγή ήταν ότι αντί να παρέχει ανεξάρτητα πλαίσια αναζήτησης για την αναζήτηση miRNA, γονιδίων και μονοπατιών, ο εξυπηρετητής Ιστού παρέχει μόνο ένα πλαίσιο αναζήτησης που εξυπηρετεί όλους τους τύπους αναζήτησης. Επιπλέον, οι θέσεις των σημείων πρόσδεσης στο μετάγραφο του γονιδίου στόχοσν αναπαριστώνται γραφικά μέσω του UCSC genome browser. Αυτή η αυτόματη φόρτωση μπορεί να χρησιμοποιηθεί για την παροχή πληροφοριών σε σύγκριση με άλλα ευρήματα ενδιαφέροντος όπως πολυμορφισμούς ενός νουκλεοτιδίου (SNP), επαναλαμβανόμενα στοιχεία, και εναλλακτικές μορφές σύνδεσης 3′-UTR. Μια άλλη σημαντική βελτίωση είναι ότι το DIANA microT v.4 παρέχει έναν ενσωματωμένο προσωπικό χώρο χρήστη στον οποίο οι χρήστες μπορούν να αποθηκεύ-

ουν εύκολα σημαντικές αναζητήσεις και αποτελέσματα που επιθυμούν να κρατήσουν για μελλοντική ανάλυση. Συγκεκριμένα, το σύστημα κρατά τις πιο πρόσφατες αναζητήσεις του χρήστη, παρέχοντας την ευκαιρία να επαναλάβει αναζητήσεις. Ένας μηχανισμός δημιουργίας σελιδοδεικτών παρέχει τη δυνατότητα αποθήκευσης ενδιαφέροντων αποτελεσμάτων μαζί με τα σχόλια του χρήστη. Ο προσωπικός χώρος παρέχει στατιστικά χρήσης σχετικά με τις πιο πρόσφατες αναζητήσεις, δίνοντάς τους τη δυνατότητα έτσι να κρατούν επαφή με τα πιο πρόσφατα ευρήματά τους.

Καθώς η βιολογία των miRNA είναι ακόμα ένας αναπτυσσόμενος κλάδος, μπορεί ένα miRNA να αλλάξει όνομα μεταξύ δυο διαδοχικών εκδόσεων της miRBase. Λόγω τέτοιων αλλαγών, οι ερευνητές μπορεί να χάσουν επαφή με το πλήρες ιστορικού ενός miRNA και οι σχετικές βιβλιογραφικές αναζητήσεις να μείνουν μισές. Για την αντιμετώπιση αυτού του ζητήματος, εκτελείται μια εκτεταμένη ανάλυση σε 13 εκδόσεις της miRBase (εκδόσεις 7.1. έως 14), και εξάγεται το ιστορικό ονοματοδωσίας κάθε miRNA. Η ανάλυση χρησιμοποιεί την έκδοση 13 του miRBase σαν βάση δεδομένων αναφοράς. Αυτή η έκδοση περιλαμβάνει 1.884 ώριμα miRNA για τα τέσσερα είδη που υποστηρίζονται στον εξυπηρετητή Ιστού. Κάθε miRNA αποκτά ένα μοναδικό αριθμό αναγνώρισης, που λέγεται `MIMAT id΄ και ένα συγκεκριμένο συνδεδεμένο όνομα miRNA. Μεταξύ των εκδόσεων, συναντώνται αλλαγές σε 77 MIMAT id (38 στον άνθρωπο, 37 στο ποντίκι και 2 στο Drosophila) και 151 ονόματα miRNA (76 στον άνθρωπο, 71 στο ποντίκι και 4 στο Drosophila). Αυτό υποδηλώνει ότι οι αλλαγές ονομάτων είναι πιο συχνές από τις αλλαγές των MIMAT id. Για την παρακολούθηση αυτών των αλλαγών, ένα ευρετήριο ιστορικού ενσωματώθηκε στον εξυπηρετητή Ιστού. Οι πληροφορίες αυτού του ευρετηρίου έγιναν διαθέσιμες στο χρήστη μέσω ενός συγκεκριμένου χαρακτηριστικού που λέγεται `miRNA history΄ το οποίο χρησιμοποιήθηκε επίσης για τις βιβλιογραφικές αναζητήσεις σχετικά με τα miRNA (βλ. Σχήμα 4.2). Για παράδειγμα, το miRNA "mmu-miR-455" εμφανίστηκε πρώτη φορά στο miRBase v8.1. Το όνομά του άλλαξε αργότερα σε "mmu-miR-455-5p" στην έκδοση 8.2 και αργότερα εμφανίστηκε ως "mmu-miR-455*" στην έκδοση 10.0 (βλ. Σχήμα 4.3).



Σχήμα 4.3: Το ιστορικό εξέλιξης δεδομένων σχετικά με το miRNA με όνομα "mmu-miR-455*".

Σημειώστε ότι όλες οι διεπαφές χρήστη του εξυπηρετητή Ιστού DIANA microT v.4 υλοποιήθηκαν σε PHP χρησιμοποιώντας καλά καθιερωμένα πρότυπα σχεδιάσης για προγραμματισμό Ιστού (όπως τα MVC model, Active Records, κλπ). Όλα τα απαραίτητα δεδομένα αποθηκεύτηκαν σε μια σχεσιακή βάση δεδομένων (χρησιμοποιήθηκε η MySQL). Η υπηρεσία ViMa που παρέχει η GRNET χρησιμοποιήθηκε για τη φιλοξενία του ηλεκτρονικού εργαλείου στο `http://diana.imis.athena-innovation.gr/`

DianaTools/index.php?r=microtv4.

4.2.3 DIANA microT v.5

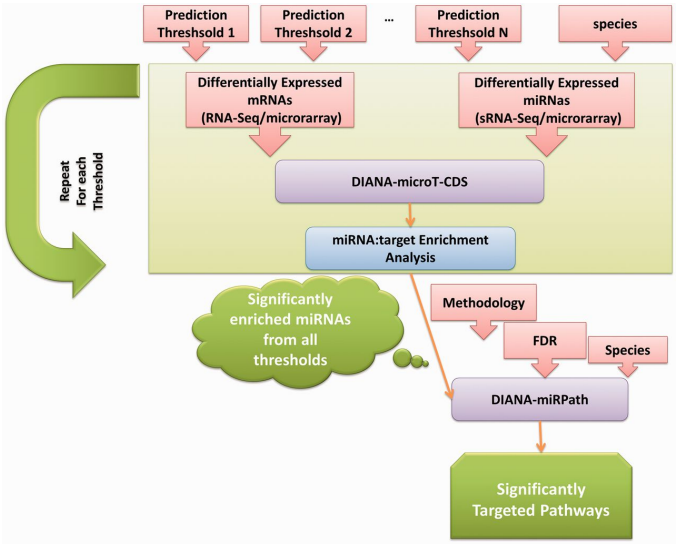
Ο εξυπηρετητής Ιστού DIANA microT v.5 [76] είναι μια σημαντικά ανανεωμένη έκδοση του εξυπηρετητή Ιστού DIANA microT v.4. Χρησιμοποιεί τη μέθοδο αιχμής για πρόβλεψη στόχων DIANA microT-CDS [78], η οποία είναι ειδικά σχεδιασμένη για να αναγνωρίζει τους στόχους των miRNA τόσο στην περιοχή 3′-UTR και στις κωδικές ακολουθίες γονιδίων (για λεπτομέρειες βλ. επίσης Κεφάλαιο 2.2.2.3). Αυτό είναι πολύ σημαντικό καθώς, αν και η αρχική έρευνα έδειχνε ότι τα miRNA προσδένονται μόνο στις περιοχές 3′-UTR των γονιδίων, αθροισμένα πειραματικά στοιχεία έχουν αποκαλύψει ότι σημεία πρόσδεσης miRNA εντός κωδικών ακολουθιών είναι επίσης λειτουργικά στο να ελέγχουν την έκφραση των γονιδίων [92]. Επιπλέον, η μέθοδος DIANA microT CDS παρέχει αυξημένη ακρίβεια και την υψηλότερη ευαισθησία σε κάθε επίπεδο εξειδίκευσης πάνω σε άλλες διαθέσιμες υλοποιήσεις αιχμής, όταν ελέγχεται για πρωτεϊνικά σύνολα δεδομένων pSILAC [83].

Ο εξυπηρετητής DIANA microT v.5 φιλοξενεί στόχους miRNA σε ακολουθίες γονιδίων των Homo sapiens, Mus musculus, Drosophila melanogaster και C. elegans. Συγκεκριμένα, η βάση δεδομένων του, κατά την ημερομηνία της πρώτης του δημοσίευσης, περιείχε $7,3 \times 10^6$ αλληλεπιδράσεις για το H.sapiens, $3,5 \times 10^6$ για το M. musculus, $4,4 \times 10^5$ για το D. melanogaster και $2,5 \times 10^5$ για το C.elegans. Επιπλέον, το DIANA microT v.5 φιλοξενεί, σε σύγκριση με το DIANA microT v.4, ανανεωμένες πληροφορίες για τα miRNA (miRBase v.18), τα γονίδια (Ensembl v.69), και τα μονοπάτια KEGG [38]. Τέλος, σημειώσεις έκφρασης γονιδίων και miRNA έχουν επίσης ενσωματωθεί στον εξυπηρετητή Ιστού, δίνοντας στο χρήστη τη δυνατότητα να εκτελέσει προηγμένο φιλτράρισμα αποτελεσμάτων βάσει της έκφρασης των ιστών.

Πέρα από τη βασική λειτουργικότητα αναζήτησης, η πέμπτη έκδοση του εξυπηρετητή DIANA microT παρέχει επίσης υποστήριξη για προηγμένες διοχετεύσεις (pipelines) για αναλύσεις υψηλής ρυθμαπόδοσης. Συγκεκριμένα, ο εξυπηρετητής Ιστού DIANA microT v.5 φιλοξενεί πολλές ενσωματωμένες αναλύσεις στη μορφή έτοιμων προηγμένων διοχετεύσεων, που καλύπτουν μια ευρεία γκάμα ερωτημάτων σχετικά με τις προβλεπόμενες ή επιβεβαιωμένες αλληλεπιδράσεις miRNA-γονιδίου και τον αντίκτυπό τους στα μεταβολικά μονοπάτια και μονοπάτια σηματοδοσίας. Αυτές οι διοχετεύσεις μπορεί να χρησιμοποιηθούν για την ανάλυση δεδομένων χρηστών που προέρχονται από πειράματα μικρής κλίμακας και υψηλής ρυθμαπόδοσης απευθείας από τη διεπαφή του εξυπηρετητή Ιστού DIANA-microT, χωρίς την αναγκαιότητα εγκατάστασης ή υλοποίησης κανενός είδους λογισμικού.

Για παράδειγμα, μία από τις διαθέσιμες ροές εργασιών (Σχήμα 4.4) μπορεί να αναλύσει δεδομένα έκφρασης mRNA και miRNA (αλλαγή έκφρασης και αναδίπλωσης). Καταπιεσμένα γονίδια ταιριάζουν αυτόματα με υπερεκφρασμένα miRNA (και αντίστροφα). Η ροή εργασιών εκτελεί ανάλυση εμπλουτισμού πειραματικά επιβεβαιωμένων στόχων που προέρχονται από το DIANA TarBase v.6 (βλ. Κεφάλαιο 4.4) ή/και προβλεπόμενων αλληλεπιδράσεων από το microT-CDS. Αυτό το βήμα θεωρείται ουσιώδες για την αναγνώριση των miRNA που καθορίζουν σημαντικά τα διαφορικά εκφρασμένα γονίδια.

Το κατώφλι σκορ πρόβλεψης μπορεί να επηρεάσει σημαντικά τα βήματα ανάλυσης που ακολουθούν. Στην περίπτωση των προβλεπόμενων αλληλεπιδράσεων, η διοχέτευση μπορεί να βελτιστοποιηθεί μέσω αυτόματων επαναλήψεων διαφόρων κατοφλίων πρόβλεψης (από ευαίσθητα σε πιο αυστηρά), για την ελαχιστοποίηση της επίδρασης



Σχήμα 4.4: Διάγραμμα ροής που απεικονίζει μια διοχέτευση ανάλυσης άμεσα διαθέσιμη από τη διεπαφή του εξυπηρετητή Ιστού.

των επιλεγμένων ρυθμίσεων στο αποτέλεσμα. Με χρήση στατιστικών μετα-ανάλυσης, ο εξυπηρετητής συνδυάζει τις τιμές P από κάθε επανάληψη σε μια συνολική τιμή P για κάθε miRNA, συμβολίζοντας την επίδραση στα επιλεγμένα γονίδια για όλα τα χρησιμοποιημένα κατώφλια. Στο τελευταίο βήμα της διοχέτευσης, τα αναγνωρισμένα miRNA υπόκεινται σε λειτουργική ανάλυση, όπου εντοπίζονται μονοπάτια που ελέγχονται από τη συνδυασμένη δράση αυτών των miRNA με τη χρήση του DIANA mirPath v.2.1 (βλ. Ενότητα 4.5).

Άλλες διαθέσιμες διοχετεύσεις μπορούν να χειριστούν λίστες miRNA και γονιδίων, για να εκτελέσουν την ανάλυση εμπλουτισμού, ή ακόμα να επιλέξουν τον τύπο των χρησιμοποιούμενων αλληλεπιδράσεων (προβλεπόμενων ή επικυρωμένων πειραματικά). Στην τελευταία ροή εργασιών, ο αλγόριθμος ἐξατομικεύει΄ την ενότητα αναγνώρισης στόχων για κάθε miRNA. Αναγνωρίζει αρχικά τον αριθμό των διαθέσιμων αλληλεπιδράσεων στο DIANA TarBase και το DIANA microT-CDS (επιβεβαιωμένες έναντι προβλεπόμενων) και επιλέγει αυτόματα να χρησιμοποιήσει επιβεβαιωμένους στόχους μόνο στις περιπτώσεις των καλά σημειωμένων miRNA. Οι υπολογιστικά αναγνωρισμένες αλληλεπιδράσεις χρησιμοποιούνται διαφορετικά.

Ο νέος εξυπηρετητής Ιστού DIANA microT δίνει στους χρήστες τη δυνατότητα να εκτελέσουν τέτοιες αναλύσεις απευθείας από τη διαδικτυακή διεπαφή χρήστη, ή να δημιουργήσουν πιο εκτεταμένες διοχετεύσεις προγραμματιστικά ή χρησιμοποιώντας εποπτικά εργαλεία (Taverna WMS [33]). Προς αυτό το σκοπό, ο εξυπηρετητής Ιστού DIANA microT v5.0 παρέχει πλήρη ενσωμάτωση με το Taverna WMS, χρησιμοποιώντας το DIANA-Taverna Plug-in που αναπτύχθηκε. Το DIANA-Taverna Plug-in δίνει στο χρήστη τη δυνατότητα να έχει άμεση πρόσβαση στον εξυπηρετητή πρόβλεψης στόχων (microT-CDS) από τη γραφική διεπαφή του Taverna και να ενσωματώσει προηγμένες λειτουργικότητες ανάλυσης miRNA σε ειδικές διοχετεύσεις. Επιπλέον, το plug-in δίνει τη δυνατότητα επέκτασης τέτοιων διοχετεύσεων μέσω της χρήσης άλλων εργαλείων DIANA και βάσεων δεδομένων, παρέχοντας πρόσβαση στην πιο εκτεταμένη συλλογή επικυρωμένων στόχων miRNA (DIANA TarBase v.6) και στο DIANA mirPath v.2.1, ένα εργαλείο σχεδιασμένο για την αναγνώριση στοχευμέ-

νων μονοπατιών miRNA. Επιπλέον, ο εξυπηρετητής Ιστού υποστηρίζει επίσης άμεση προγραμματιστική πρόσβαση σε όλες τις ανωτέρω παροχές στη μορφή υπηρεσιών, για να διευκολύνει τους χρήστες να έχουν ήδη υλοποιημένες διοχετεύσεις με τη χρήση γνωσσών σεναρίων ή γλωσσών προγραμματισμού.

Να σημειωθεί ότι όλες οι διεπαφές χρήστη του εξυπηρετητή Ιστού DIANA microT v.5 υλοποιήθηκαν σε PHP με χρήση καλά καθιερωμένων προτύπων σχεδιασμού για προγραμματισμό Ιστού (όπως τα MVC model, Active Records, κλπ). Όλα τα απαραίτητα δεδομένα αποθηκεύτηκαν σε μια σχεσιακή βάση δεδομένων (χρησιμοποιήθηκε η MySQL). Η υπηρεσία ViMa[5] παρασχέθηκε από την GRNET[6] και χρησιμοποιήθηκε για τη φιλοξενία του ηλεκτρονικού εργαλείου στο `http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=microtv5`.

4.3 DIANA miRGen: Αποκαλύπτοντας πληροφορίες για μετάγραφα miRNA

Όπως άλλα μόρια RNA, τα ώριμα microRNA παράγονται από μετάγραφα. Κάθε ένα από αυτά τα μετάγραφα μπορεί να κωδικοποιήσει περισσότερα από ένα ώριμαmiRNA (βλ. επίσης Κεφάλαιο 2.2.1.3). Η γνώση της γονιδιακής τοποθεσίας αυτών των μεταγράφων και του πότε εκφράζονται είναι πολύ χρήσιμες πληροφορίες για κάθε ερευνητή που μελετά μόρια miRNA. Η πρώτη έκδοση του DIANA miRGen [63] ήταν μια βάση δεδομένων που συνέλεγε αυτές τις πληροφορίες. Ωστόσο, οι συστάδες των miRNA που εκφράζονται μαζί αναγνωρίζονταν βάσει της σχετικής τους απόστασης και των γονιδιακών χαρακτηριστικών που τα περιέβαλαν. Αυτή η μεθοδολογία οδηγεί σε πολλά λάθη, έτσι, απαιτούνταν μια πιο ακριβής βάση δεδομένων.

4.3.1 DIANA miRGen v.2

Το DIANA miRGen v.2 [3] είναι μια βάση δεδομένων που στοχεύει στην παροχή πλήρων πληροφοριών για τη θέση των μεταγράφων που κωδικοποιούν microRNA στον άνθρωπο και στο ποντίκι και τη ρύθμισή τους από μεταγραφικούς παράγοντες, συμπεριλαμβάνοντας μια μοναδική συλλογή τόσο προβλεπόμενων όσο και πειραματικά υποστηριζόμενων δεδομένων.

Κατά την ανάπτυξη του DIANA miRGen v.2:

- πρωτεύοντα μετάγραφα σε γονιδιώματα θηλαστικών (συγκεκριμένα, στο γονιδίωμα του ανθρώπου και του ποντικιού) αναγνωρίστηκαν εξορύσσοντας σημαντικές πηγές βιβλιογραφίας.

- σημεία πρόσδεσης μεταγραφικών παραγόντων (Transcription Factor Binding Sites - TFBS) χαρτογραφήθηκαν εντός των περιοχών ανάντη των σημείων εκκίνησης της μεταγραφής (TSS) των ανωτέρω πρωτευόντων μεταγράφων miRNA

- ενσωματώθηκαν προφίλ έκφρασης miRNA σε αρκετούς ιστούς, η χαρτογράφηση των SNP εντός γονιδιακών τοποθεσιών φουρκέτων miRNA, και η χαρτογράφηση των SNP εντός των TFBS που βρέθηκαν ανάντη των γονιδίων miRNA.

[5]http://vima.grnet.gr
[6]https://www.grnet.gr

Η αλληλεπίδραση αυτών των διαφόρων πηγών πληροφορίας σχετικά με γονιδιακά χαρακτηριστικά που συνδέονται με γονίδια miRNA και τα επίπεδα έκφρασής τους μπορούν να χρησιμοποιηθούν για τη μελέτη της λειτουργίας των miRNA και της απορρύθμισής τους στην ασθένεια. Για παράδειγμα, ένας χρήστης που ενδιαφέρεται για ένα συγκεκριμένο μεταγραφικό παράγοντα μπορεί να βρει γονίδια miRNA που να συνδέονται με αυτόν, να βρει τα επίπεδα έκφρασης αυτών των miRNA σε ένα πιθανό ιστό ενδιαφέροντος, να βρει ίσως κάποια SNP στα TFBS ή τις τοποθεσίες miRNA στο γονιδίωμα που συνδέονται με μια πιθανή ασθένεια ενδιαφέροντος και, τέλος, να βρει προβλεπόμενους στόχους των miRNA που συνδέονται με το μεταγραφικό παράγοντα που τον ενδιαφέρει, και μοριακά μονοπάτια στα οποία οι στόχοι κάθε ενός από αυτά τα miRNA εμπλέκονται ξεχωριστά ή μαζί.

Τα μετάγραφα miRNA στον άνθρωπο και το ποντίκι αναγνωρίστηκαν από τέσσερις βιβλιογραφικές πηγές:

- Οι Corcoran et al. [17] χρησιμοποίησαν δεδομένα ανοσοκαθίζησης PolII και ChIP-chip σε επιθηλιακά πνευμονικά κύτταρα για να αναγνωρίσουν τα μετάγραφα miRNA και τις περιοχές-υποκινητές τους.

- Οι Landgraf et al. [48] ακολουθιοποίησαν 250 μικρές βιβλιοθήκες RNA που αντιστοιχούν σε 26 διαφορετικά οργανικά συστήματα και τύπους κυττάρων του ανθρώπου και του ποντικιού, με περίπου 1.000 κλώνους miRNA ανά βιβλιοθήκη και αναγνωρισμένα γονίδια που κωδικοποιούν miRNA. Σε αυτή την εργασία όλα τα μετάγραφα των γονιδίων που κωδικοποιούν miRNA αναγνωρίστηκαν, καθώς και τα γονίδια που κωδικοποιούν πρωτεΐνες και περιέχουν miRNA.

- Οι Oszolak et al. [72] προέβλεψαν την τοποθεσία των εγγείων υποκινητών των ανθρώπινων miRNA συνδυάζοντας χαρτογράφηση νουκλεοσωμάτων με υπογραφές χρωματίνης των υποκινητών σε κύτταρα MALME, HeLa και UACC62. Αν και αναγνωρίστηκε σε αυτή τη μελέτη το TSS των γονιδίων miRNA, δεν παρασχέθηκε το τέλος του μεταγράφου. Παρείχαμε το τέλος του τελευταίου miRNA που είναι μέλος ενός γονιδίου ως προσεγγιστική αποτίμηση του τέλους του μεταγράφου.

- Οι Marson et al. [61] χρησιμοποίησαν δεδομένα ChIP-seq για την αναγνώριση υποκινητών γονιδίων miRNA σε εμβρυϊκά βλαστοκύτταρα. Αναγνώρισαν υποκινητές και συν-ρυθμισμένα miRNA, αλλά η ακριβής θέση του TSS δεν αναγνωρίστηκε. Για το λόγο αυτό, χρησιμοποιήσαμε την αρχή του πρώτου miRNA κάθε συστάδας ως πεωρούμενο TSS. Επιπροσθέτως, οι συντεταγμένες που είχαν δώσει οι Marson et al. έπρεπε να αρθούν χρησιμοποιώντας το 'UCSC lift over tool' στο πιο πρόσφατο, τότε, genome build (hg18, mm9). Σε περιπτώσεις όπου χρησιμοποιούνται θεωρούμενες και όχι πειραματικά επιβεβαιωμένες θέσεις, δηλώνεται στη γραφική διεπαφή ως ὑπολογιστικό TSS΄.

Συνολικά, αναγνωρίστηκαν 812 μετάγραφα που κωδικοποιούν ανθρώπινα miRNA και 386 μετάγραφα που κωδικοποιούν miRNA ποντικιού. Από αυτά, τα 423 εμφανίζονταν σε αντίστοιχες εργασίες να συνδέονται με γονίδια που κωδικοποιούν πρωτεΐνες (ενδογονιδιακά μετάγραφα miRNA). Πάνω από μία από τις παραπάνω δημοσιεύσεις συνήθως αναγνωρίζουν μετάγραφα που αντιστοιχούν σε ένα miRNA. Όταν συμβαίνει αυτό, μετάγραφα από όλες τις μεθόδοτς επιστρέφονται στο χρήστη.

Καθώς δημοσιεύτηκαν αυτές οι μελέτες, αναγνωρίστηκαν επιπλέον miRNA. Όταν καινούρια miRNA βρίσκονται εντός των συντεταγμένων των συστάδων που δίνονται

από οποιαδήποτε από αυτές τις δημοσιεύσεις, αυτό το miRNA προστίθεται στη συστάδα. Για ονόματα που άλλαξαν ή αποδόθηκαν διαφορετικά από τον τρέχοντα κανόνα, χρησιμοποιήθηκε χειροκίνητη φύλαξη με αναφορά στο mirBase για την αναγνώριση και αντικατάσταση αυτών των ονομάτων σύμφωνα με τον τρέχοντα κανόνα. Για όλους αυτούς τους λόγους, είναι δυνατό ο αριθμός γονιδίων που χρησιμοποιούνται στο miRGen να μην αντιστοιχεί τέλεια στον αριθμό που δηλώνεται στις αντίστοιχες δημοσιεύσεις.

Προκειμένου να προσδιορίσουμε θεωρούμενα TFBS κοντά στο TSS των πρωτεύοντων μεταγράφων miRNA, χρησιμοποιήσαμε το ελεύθερα διαθέσιμο εργαλείο Match™ [41]. Το Match™ χρησιμοποιεί τη δημόσια βιβλιοθήκη των μητρώων ζύγισης θέσης από το Transfac 6.0. Ταιριάξαμε όλα τα μητρώα μεταγραφικών παραγόντων των σπονδυλωτών στις περιοχές που εκτείνονται από τα 5 kb ανάντη του κάθε TSS στο 1 kb κατάντη του TSS. Σαν κριτήριο για τον προσδιορισμό των τιμών αποκοπής, επιλέγουμε την ελαχιστοποίηση των ψευδοθετικών αποτελεσμάτων προκειμένου να παράγουμε ένα αυστηρό σύνολο προβλέψεων χωρίς πολλά εσφαλμένως προβλεπόμενα TFBS. Υπολογίζονται δυο σκορ για κάθε πεωρούμενο TFBS. Το σκορ ομοιότητας μητρώων περιγράφει την ποιότητα ταιριάσματος μεταξύ ενός ολόκληρου μητρώου και ενός αυθαίρετου τμήματος των ακολουθιών εισόδου. Ανάλογα, το σκορ ομοιότητας πυρήνων υποδηλώνει την ποιότητα του ταιριάσματος ανάμεσα στην ακολουθία του πυρήνα ενός μητρώου (δηλαδή τις πέντε πιο διατηρημένες θέσεις μέσα στο μητρώο) και σε ένα τμήμα της ακολουθίας εισόδου.

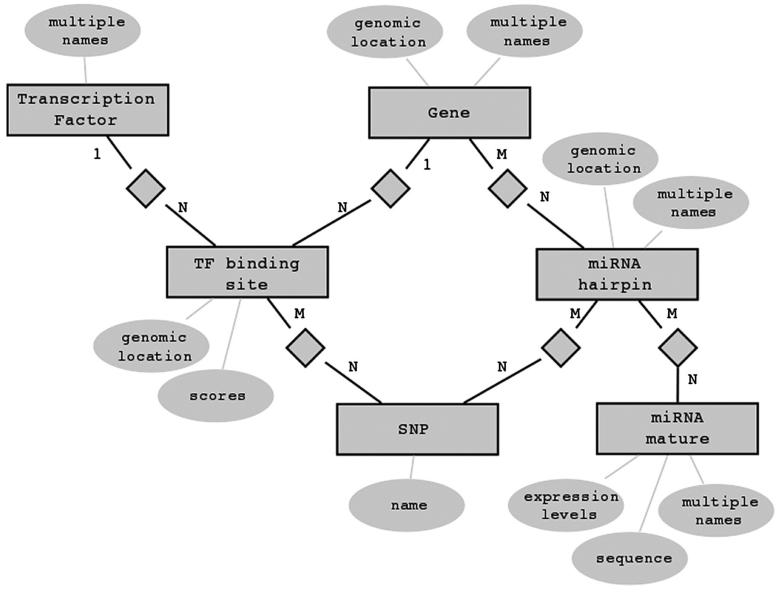
Τα προφίλ έκφρασης miRNA αναγνωρίστηκαν από τον άτλαντα έκφρασης miRNA των θηλαστικών [48]. Πληροφορίες για τα προφίλ έκφρασης των 548 ανθρώπινων miRNA και των 451 miRNA από ποντίκια σε 172 μικρές βιβλιοθήκες RNA για τον άνθρωπο 68 για τα ποντίκια προέκυψαν από κυτταρικές σειρές και ιστούς.

Τα SNP που βρίσκονται μέσα στις γονιδιακές θέσεις των φουρκετών miRNA και τα αντίστοιχα TFBS συλλέχθηκαν από το UCSC table browser. Για τον άνθρωπο, δεδομένα Πολυμορφισμού από τη βάση δεδομένων dbSnp [88] ή γονοτυπικοί πίνακες SNP130 χρησιμοποιήθηκαν με 18.833.531 αναγνωρισμένα SNP. Για το ποντίκι, το SNP128 χρησιμοποιήθηκε με 14.893.502 αναγνωρισμένα SNP.

Το αποθετήριο miRGen έχει υλοποιηθεί με χρήση τεχνολογίας σχεσιακής βάσης δεδομένων. Όλα τα δεδομένα αποθηκεύονται σε ένα σύστημα διαχείρισης σχεσιακής βάσης δεδομένων MySQL. Το Σχήμα 4.5 απεικονίζει μέρος του μοντέλου οντότητας-σχέσης της εφαρμογής μας. Για περισσότερες πληροφορίες σχετικά με τις βιολογικές οντότητες που περιγράφονται σε αυτό το μοντέλο και τις σχέσεις μεταξύ τους βλ. Ενότητα 2.2.1.

Όλα τα αποτελέσματα είναι διαθέσιμα μέσω μιας διεπαφής φιλικής προς το χρήστη, η οποία επιτρέπει αναζητήσεις για miRNA και για μεταγραφικούς παράγοντες ενδιαφέροντος. Για τα ώριμα miRNA, είναι πιθανό να παρουσιάσουμε στόχους που προβλέπονται από τη μέθοδο DIANA microT και για τα miRNA που βρίσκονται στο ίδιο μετάγραφο, ο χρήστης μπορεί να δει μια λειτουργική σημείωση των στόχων τους σε μοριακά μονοπάτια μέσω της εφαρμογής DIANA mirPath (βλ. Ενότητα 4.5). Το Σχήμα 4.6 δείχνει μια επισκόπηση της διεπαφής και τονίζει τους συνδέσμους σε εξωτερικές βάσεις δεδομένων-τα UCSC genome browser, iHop, dbSNP, και mirBase.

Όταν εκτελείται αναζήτηση ενός miRNA (Σχήμα 4.6-A), όλες οι διακριτές τοποθεσίες στο γονιδίωμα (φουρκέτες) που θα μπορούσαν να κωδικοποιήσουν για αυτό το miRNA επιστρέφονται, και ο χρήστης μπορεί να δει λεπτομέρειες για κάθε ένα από τα πιθανά επικαλυπτόμενα μετάγραφα που αναγνωρίζονται για κάθε τοποθεσία, και προβλέπονται συνήθως από διαφορετικές εργασίες. Κάθε καρτέλα μεταγράφου περιέχει



Σχήμα 4.5: Μέρος του μοντέλου οντότηταση-σχέσης της βάσης δεδομένων miRGen

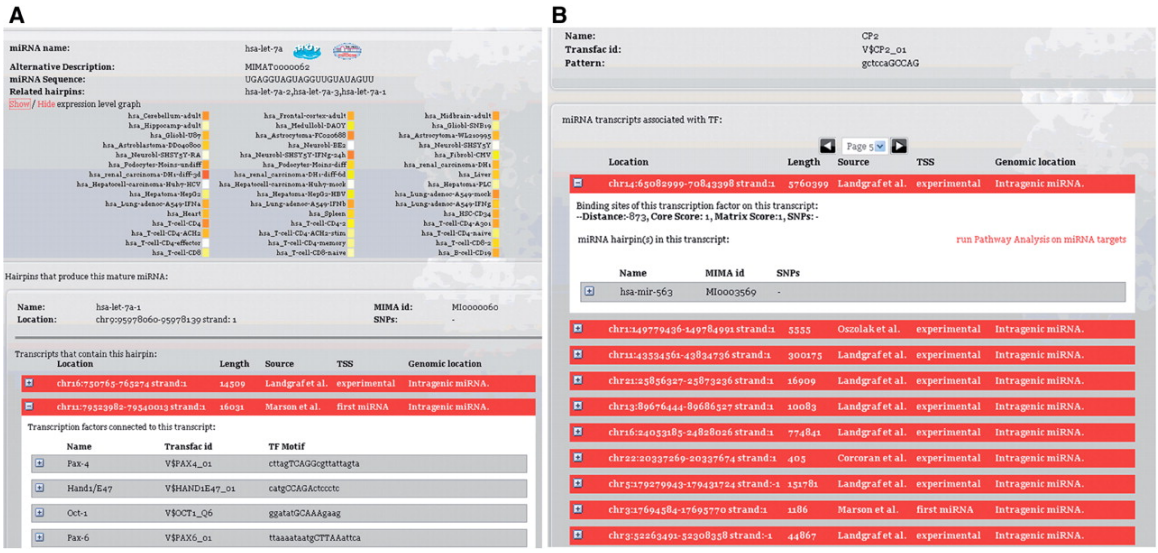
πληροφορίες για τα TFBS που βρίσκονται από τα 5 kb ανάντη έως το 1 kb κατάντη της αρχής του μεταγράφου. Επιπροσθέτως, πληροφορίες για τα επίπεδα έκφρασης του ώριμου miRNA παρουσιάζονται ως χάρτης θερμότητας.

Η αναζήτηση ενός μεταγραφικού παράγοντα που μας ενδιαφέρει (Σχήμα 4.6-Β) επιστρέφει όλα τα γονίδια που κωδικοποιούν miRNA για τα οποία βρίσκεται τουλάχιστον ένα σημείο πρόσδεσης για αυτό το μεταγραφικό παράγοντα. Πληροφορίες για το γονίδιο, τα TFBS, και τα ώριμα miRNA που κωδικοποιούνται από το γονίδιο μπορούν να φανούν σε καρτέλες. Όλες οι περιπτώσεις TFBS και φουρκετών miRNA συνδέονται με την αντίστοιχη χαρτογράφηση των SNP στις γονιδιακές τους τοποθεσίες. Για όλα τα μετάγραφα, φαίνεται η βιβλιογραφική πηγή του γονιδίου, η αναγνώριση του TSS (πειραματικό εάν το TSS αναγνωριζόταν στην εργασία, υπολογιστικό εάν υπολογιζόταν με υπολογιστικά μέσα και πρώτο miRNA εάν η αρχή του πρώτου miRNA λειτουργεί ως υποκατάστατο για ένα άγνωστο TSS), και εάν το γονίδιο είναι ενδογονιδιακό ή συν-εκφράζεται με ένα γονίδιο που κωδικοποιεί πρωτεΐνη.

Σημειώστε ότι όλες οι διεπαφές χρήστη DIANA miRGen v.2 υλοποιήθηκαν σε PHP με χρήση καλά καθιερωμένων προτύπων σχεδιασμού για προγραμματισμό (όπως τα like MVC model, Active Records, κλπ). Όλα τα απαραίτητα δεδομένα αποθηκεύτηκαν σε μια σχεσιακή βάση δεδομένων (χρησιμοποιήθηκε η MySQL). Η υπηρεσία ViMa παρασχέθηκε από την GRNET και χρησιμοποιήθηκε για τη φιλοξενία του ηλεκτρονικού εργαλείου στο `http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=mirgen`.

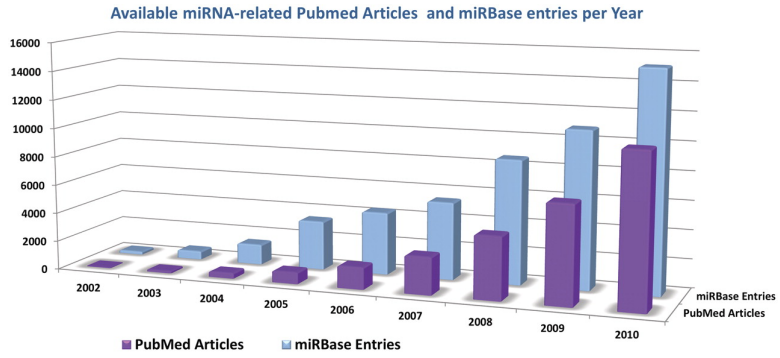
4.4 DIANA TarBase: Αναζητώντας πειραματικά επιβεβαιωμένους στόχους miRNA

Κατά την προηγούμενη δεκαετία έντονες ερευνητικές προσπάθειες είχαν ως αποτέλεσμα την παραγωγή σημαντικής ποσότητας δεδομένων που σχετίζονται με τη βιογένεση και τη λειτουργία των miRNA. Αυτή η διεργασία αντανακλάται στην υπεργραμμική αύξηση



Σχήμα 4.6: Δυο αποτυπώσεις της διεπαφής miRGen. Μία για μια αναζήτηση miRNA (A) και μια άλλη για αναζήτηση του μεταγραφικού παράγοντα (B).

του πλήθους εγγραφών της miRBase και των άρθρων του PubMed που σχετίζονται με miRNA (βλ. Σχήμα 4.7). Κάτω από αυτές τις συνθήκες η αναγνώριση στόχων miRNA αποτελεί μία κουραστική εργασία για τους ερευνητές των βιοεπιστημών.



Σχήμα 4.7: Η ετήσια αύξηση των δημοσιεύσεων που αφορούν σε miRNA στο PubMed και το πλήθος των καταχωρήσεων στη βάση δεδομένων miRBase.

Ένα μεγάλο πλήθος από μεθόδους πρόβλεψης στόχων, οι οποίες προσπαθούν να προσεγγίσουν το πρόβλημα υπολογιστικά, είναι διαθέσιμες (βλ. Κεφάλαιο 2.2.2). Κάποιοι στόχοι μπορούν να προβλεφθούν με μεγάλη βεβαιότητα από τις παρούσες τεχνικές, όμως, η ακρίβεια και το ποσοστό ανάκτησης των αλγορίθμων αιχμής εκτιμάται ως περίπου 50% και 12%, αντίστοιχα, όταν δοκιμάζονται έναντι στόχων miRNA που υποστηρίζονται πρωτεομικά [2], δίνοντας έμφαση στην ανάγκη για μαζική πειραματική επιβεβαίωση στόχων miRNA [93].

Οι στόχοι miRNA μπορούν να επιβεβαιωθούν πειραματικά με ειδικές τεχνικές για κάθε γονίδιο (gene-specific), όπως επίσης και με τεχνικές υψηλής ρυθμαπόδοσης. Οι ειδικές τεχνικές περιλαμβάνουν τις δοκιμασίες γονιδίων ανταπόκρισης, την αξιολόγηση συν-έκφρασης μορίων miRNA και στόχων (π.χ. Northern blotting ή qPCR) και την εκτίμηση της επίδρασης του miRNA πάνω σε πρωτεΐνη στόχο (π.χ. ELISA, Western

blotting, ανοσοιστοχημεία). Οι τεχνικές υψηλής ρυθμαπόδοσης μπορούν να είναι μια απλή επέκταση κάποιας ειδικής τεχνικής σε περιβάλλον υψηλής ρυθμαπόδοσης, π.χ. η εκμετάλλευση microarray screening αντί για qPCR. Μπορούν επίσης να εμπλέκουν καινοφανείς σχετικές μεθοδολογίες, όπως είναι οι RNA-Seq, ανοσοκαθίζηση RISC component, ανάλυση HITS-CLIP, PAR-CLIP και διάφορες πρωτεωμικές προσεγγίσεις όπως η SILAC.

Καθώς η σχετική βιβλιογραφία και το πλήθος των πειραμάτων αυξάνονται με υπερ γραμμικό ρυθμό, βάσεις δεδομένων που επιμελούνται και συγκεντρώνουν πειραματικά επιβεβαιωμένους στόχους εμφανίστηκαν σταδιακά. Αυτές οι βάσεις δεδομένων προσπαθούν να παρέχουν πρόσβαση σε αυτή την ποικιλία πειραματικών δεδομένων, τα οποία είναι διασκορπισμένα σε χιλιάδες δημοσιευμένες εργασίες. Το DIANA TarBase v.1 [86], η πρώτη βάση δεδομένων για στόχους miRNA που είναι πειραματικά επιβεβαιωμένοι δημιουργήθηκε για να βοηθά το σχεδιασμό των πιο συμπαγών μεθόδων πρόβλεψης. Καθώς τα συνολικά δεδομένα αυξάνονταν, το DIANA TarBase και άλλες σχετικές βάσεις δεδομένων μετατράπηκαν σε ανεκτίμητα εργαλεία για όλες τις πτυχές της έρευνας που σχετίζεται με τα miRNA. Στο υπόλοιπο αυτού του κεφαλαίου πρώτα κάνουμε μία επισκόπηση των βάσεων δεδομένων που προαναφέρθηκαν και στη συνέχεια περιγράφουμε το DIANA TarBase v.6 το οποίο αναπτύξαμε.

4.4.1 Σχετικές εργασίες

Μια σύντομη επισκόπηση των διαθέσιμων βάσεων δεδομένων παρουσιάζεται παρακάτω σε αλφαβητική σειρά:

Το miR2Disease [35] έγινε διαθέσιμο για πρώτη φορά το 2008. Είναι μια επιμελημένη βάση δεδομένων που έχει σκοπό να παρέχει πληροφορίες σχετικά με παθολογίες που αφορούν miRNA. Το miR2Disease επιμελείται 809 αλληλεπιδράσεις miRNA-γονιδίου για τον Homo sapiens συνδεδεμένες με πληροφορίες σχετικών ασθενειών σύμφωνα με τη σχετική βιβλιογραφία. Οι $3,273$ καταχωρήσεις για ασθένειες miRNA αποτελούν το ισχυρότερο πλεονέκτημα της βάσης δεδομένων. Ο χρήστης μπορεί να πραγματοποιήσει αναζήτηση είτε για miRNA, είτε για στοχευόμενο γονίδιο ή όνομα ασθένειας. Περισσότερες λεπτομέρειες παρέχονται όπως η μέθοδος επιβεβαίωσης, οι επιβεβαιωμένοι στόχοι με βάση μία παλιά έκδοση του DIANA TarBase , πληροφορίες για την επιστημονική εργασία και σύνδεσμοι προς σελίδες αλγορίθμων πρόβλεψης στόχων.

Το MirnaMAP [31] έγινε διαθέσιμο για πρώτη φορά το 2006. Περιέχει δεδομένα που προέρχονται από μια ξεπερασμένη έκδοση του DIANA TarBase (η οποία αριθμούσε 346 στόχους) και από επιμέλεια των συγγραφέων (29 στόχοι). Το MirnaMAP δεν ανανεωνόταν για περισσότερα από 4 χρόνια και περιέχει ένα περιορισμένο πλήθος από πειραματικά επιβεβαιωμένους στόχους για τον H. sapiens. Οι περισσότερες από τις καταχωρήσεις του mirRNAMAP βασίζονται σε προβλέψεις αλληλεπιδράσεων για $2,464$ miRNA σε 12 είδη. Το MirnaMAP παρέχει μία πληθώρα από διαθέσιμα δεδομένα για κάθε καταχώρηση της βάσης δεδομένων, τα οποία περιλαμβάνουν πληροφορίες για το miRNA και το γονίδιο, πληροφορία για το προφίλ έκφρασης του miRNA σε ιστούς, προβλέψεις για τα στοχευόμενα γονίδια, όπως επίσης και σχετική βιβλιογραφία.

Το MiRecords [104] έγινε διαθέσιμο για πρώτη φορά το 2008. Περιέχει επιμελημένους στόχους miRNA και προβλέψεις στόχων. Το υποσύστημα με τους επιβεβαιωμένους στόχους της βάσης δεδομένων περιέχει $2,286$ αλληλεπιδράσεις μεταξύ 548 miRNA και $1,579$ γονιδίων-στόχων σε εννιά είδη (τα συγκεκριμένα νούμερα αναφέρονται στην έκδοση που ήταν διαθέσιμη στις 25 Νοεμβρίου του 2010). Το μεγαλύτερο

μέρος από αυτές τις αλληληεπιδράσεις προέρχονται από ειδικά για κάθε γονίδιο (gene-specific) πειράματα. Η βάση δεδομένων παρέχει πληροφορίες για τα miRNA, τα γονίδια και τις τοποθεσίες των στόχων, όπως επίσης και συνδέσμους προς το miRBase και τη RefSeq. Οι αλληλεπιδράσεις miRNA-γονιδίου υποστηρίζονται από δεδομένα που αφορούν την επιστημονική δημοσίευση, την πειραματική μέθοδο που χρησιμοποιήθηκε για την επιβεβαίωση όπως επίσης και ένα επιλεγμένο εδάφιο του κειμένου της δημοσίευσης το οποίο διατυπώνει το πειραματικό αποτέλεσμα. Παρόλα αυτά ο χρήστης δεν μπορεί να φιλτράρει τα αποτελέσματα με βάση τα υπάρχοντα πεδία. Η διεπαφή του miRecords επιτρέπει επίσης στο χρήστη να εισάγει νέες αλληλεπιδράσεις miRNA-γονιδίου.

Η βάση δεδομένων miRSel [65] έγινε διαθέσιμη για πρώτη φορά το 2010. Περιέχει δεδομένα αλληλεπιδράσεων miRNA τα οποία προέρχονται αποκλειστικά από εξόρυξη κειμένου σε περιλήψεις από τη MedLine. Ο αλγόριθμος εξόρυξης κειμένου επιτυγχάνει να βρει αλληλεπιδράσεις miRNA-γονιδίου με ακρίβεια 65%, ποσοστό ανάκτησης 90%, με βάση μία δοκιμή που εφαρμόστηκε σε 89 επιλεγμένες προτάσεις που προέρχονταν από 50 περιλήψεις από το PubMed. Το MiRSel περιέχει 3,690 αλληλεπιδράσεις miRNA-γονιδίου που έχουν προέλθει από εξόρυξη κειμένου. Εφαρμόζοντας λιγότερο αυστηρά κριτήρια, ο χρήστης μπορεί να έχει πρόσβαση σε περίπου 8,000 ζεύγη τα οποία χαρακτηρίζονται ως λιγότερο αξιόπιστα από τους δημιουργούς. Στο miRSel ο χρήστης μπορεί επίσης να αναζητήσει miRNA που σχετίζονται με συγκεκριμένα άρθρα Medline τα οποία περιέχουν ένα υποσύνολο επιθυμητών όρων ή που είναι σχετικά με όρους της Gene Ontology. Σύνδεσμοι σε εξωτερικές βάσεις δεδομένων όπως είναι η miRBase και η Entrez Gene παρέχονται για κάθε καταχώρηση. Πληροφορίες σχετικά με την πειραματική μέθοδο που χρησιμοποιήθηκε για την επιβεβαίωση των στόχων miRNA δεν είναι διαθέσιμοι. Δεδομένα τα οποία προέρχονται από άλλες επιμελημένες βάσεις δεδομένων που συγκεντρώνουν αλληλεπιδράσεις miRNA, όπως είναι η DIANA TarBase v.5, η miR2Disease και η miRecords έχουν επίσης ενσωματωθεί.

Η miRTarBase [32] έγινε για πρώτη φορά διαθέσιμη το 2010. Περιέχει επιμελημένα δεδομένα για 3,969 πειραματικά επιβεβαιωμένες αλληλεπιδράσεις miRNA-γονιδίου για 14 είδη (τα νούμερα αναφέρονται στην έκδοση που ήταν διαθέσιμη στις 15 Απριλίου 2011). Παρέχει πληροφορίες σχετικές με το miRNA, το στοχευόμενο γονίδιο και τον τόπο στόχευσης. Σε πολλές περιπτώσεις, όπου τα άρθρα δεν παρουσιάζουν ξεκάθαρα πληροφορία για τον τόπο στόχευσης, η miRTarBase μπορεί να παρέχει προβλεπόμενες περιοχές χρησιμοποιώντας έναν αλγόριθμο πρόβλεψης στόχων. Πληροφορίες σχετικά με τα διαθέσιμα πειραματικά ευρήματα που υποστηρίζουν την αλληλεπίδραση περιλαμβάνονται επίσης. Η διεπαφή χρήστη προσφέρει συνδέσμους προς εξωτερικές πηγές δεδομένων όπως είναι η NCBI Entrez, ο UCSC Genome Browser, η miRBase, η BioGPS, η iHOP και η HGNC. Προαιρετικά ο χρήστης μπορεί να υποβάλλει δεδομένα για μη καταχωρημένες αλληλεπιδράσεις.

Το miRWalk [20] έγινε για πρώτη φορά διαθέσιμο το 2010. Παρέχει πειραματικά υποστηριζόμενους miRNA στόχους που αναγνωρίζονται αποκλειστικά από εξόρυξη δεδομένων σε περιλήψεις διαθέσιμες από τη MedLine. Το υποσύστημα επιβεβαιωμένων στόχων του miRWalk φιλοξενεί αλληλεπιδράσεις προερχόμενες από εξόρυξη κειμένου για 1,572 miRNA τα οποία αλληλεπιδρούν με 5,080 γονίδια από τρία είδη (άνθρωπος, ποντίκι και αρουραίος). Μια άμεση εκτίμηση της ακρίβειας του συστήματος κατά την εξαγωγή αλληλεπιδράσεων δεν παρέχεται. Η προσέγγιση εξόρυξης κειμένου των συγγραφέων τους επέτρεψε επίσης να συγκεντρώσουν δεδομένα για ασθένειες-στόχους, όργανα, κυτταρικές σειρές και μονοπάτια.

Η StarBase [106] έγινε για πρώτη φορά διαθέσιμη το 2010. Είναι μία πλατφόρ-

μα που επικεντρώνεται στην ανάλυση CLIP-Seq υψηλής ρυθμαπόδοσης (HITS-CLIP και PAR-CLIP) και δεδομένα Degradome-Seq και PARE για έξι οργανισμούς. Τα δεδομένα της StarBase που σχετίζονται με miRNA προέρχονται από οκτώ διαφορετικές μελέτες. Οι προγραμματιστές χρησιμοποιήσαν πέντε προγράμματα πρόβλεψης για να εντοπίσουν θεωρούμενους στόχους, και στην συνέχεια πήραν την τομή τους με τα δεδομένα υψηλής ρυθμαπόδοσης που είχαν προηγουμένως αναλυθεί, έχοντας ως αποτέλεσμα ένα υψηλό πλήθος από θεωρούμενους στόχους (περίπου $500,000$). Ο χρήστης μπορεί να μειώσει το πλήθος των εσφαλμένα αναφερόμενων επιλέγοντας τα αποτελέσματα μόνο ενός αλγορίθμου πρόβλεψης. Η StarBase παρέχει τον εσωτερικής ανάπτυξης deepView Genome Browser, το οποίο παρέχει πρόσβαση στα αναγνώσματα που έχουν ταιριάξει, στους προβλεπόμενους και επιβεβαιωμένους στόχους miRNA, στα ncRNA, στα γονίδια κτλ. Οι παρεχόμενες πληροφορίες περιλαμβάνουν δεδομένα για τα miRNA και τα γονίδια, όρους της GO και μονοπάτια KEGG που σχετίζονται με κάθε γονίδιο-στόχο.

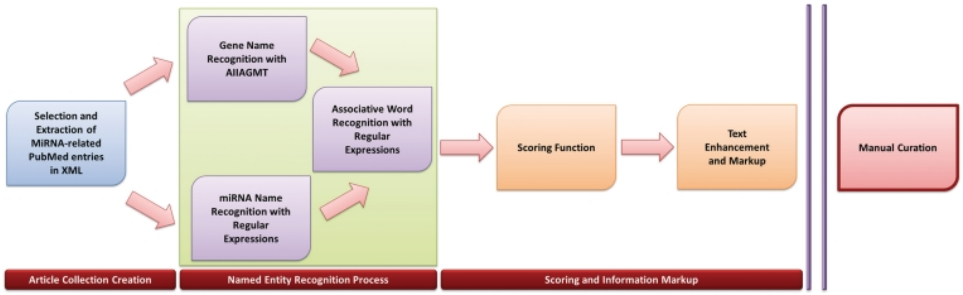
Το DIANA TarBase v.5 [74] έγινε για πρώτη φορά διαθέσιμο το 2005 και η πέμπτη του έκδοση περιλαμβάνει $1,300$ πειραματικά υποστηριζόμενους στόχους από οκτώ είδη τα οποία επιμελούνται από τη σχετική βιβλιογραφία. Τα παρεχόμενα δεδομένα για κάθε αλληλεπίδραση περιλαμβάνουν πληροφορίες για το miRNA, γενικές πληροφορίες για τα στοχευόμενα γονίδια, τη φύση της πειραματικής επιβεβαίωσης, τις σειρές κυττάρων, το υποστηρικτικό άρθρο κτλ. Η βάση δεδομένων επίσης περιλαμβάνει συνδέσμους σε εξωτερικές πηγές δεδομένων για κάθε καταχώρηση, όπως προς την Ensembl, τη HUGO, την UVSV και τη Swiss-Prot.

4.4.2 DIANA TarBase v.6

Καθώς η έρευνα για τα miRNA εξελίσσεται, οι απαιτήσεις για σχετικές βάσεις δεδομένων αυξάνουν σημαντικά. Είναι προφανές ότι υπάρχει αναγκαιότητα για βάσεις δεδομένων που να επιμελούνται μεγάλα πλήθη από πειραματικά επιβεβαιωμένους στόχους miRNA από ειδικές για κάθε γονίδιο (gene-specific) και υψηλής ρυθμαπόδοσης τεχνικές. Βάσεις δεδομένων που περιέχουν και τα δύο είδη από πειραματικά δεδομένα μέχρι τον καιρό της δημοσίευσης περιείχαν λιγότερους από $5,000$ επιβεβαιωμένους στόχους, με ένα σημαντικό μέρος αυτών των στόχων να προέρχονται από λίγα πειράματα υψηλής ρυθμαπόδοσης. Η συμπερίληψη ενός μεγάλου πλήθους από υποστηρικτικές έρευνες αυξάνει την εγκυρότητα των δεδομένων και επιτρέπει σε τέτοιες βάσεις δεδομένων να μετατρέπονται σε συνοπτικά σύνολα δεδομένων για εκπαίδευση και έλεγχο συμπαγών αλγορίθμων πρόβλεψης στόχων miRNA. Ένα άλλο σημαντικό θέμα το οποίο σύντομα έγινε εμφανές ήταν ότι οι ερευνητές απαιτούν προχωρημένες δυνατότητες αναζήτησης και φιλτραρίσματος αποτελεσμάτων προκειμένου να ανακαλύψουν miRNA ή γονίδια ειδικού ενδιαφέροντος. Τέλος, ο εμπλουτισμός των συνόλων δεδομένων με επιπρόσθετη πληροφορία από εξωτερικές πηγές (όπως μονοπάτια ή όροι της οντολογίας GO) και με μεταδεδομένα μπορεί να επιτρέψει αποδοτική εξόρυξη δεδομένων των υπαρχόντων πειραματικά επιβεβαιωμένων αποτελεσμάτων, οδηγώντας σε χρήσιμες νέες παρατηρήσεις.

Ο στόχος του DIANA TarBase v.6 [99] είναι να αντιμετωπίσει τις προαναφερθείας σύγχρονες προκλήσεις και να εγκαινιάσει την επόμενη γενιά βάσεων δεδομένων για επιβεβαιωμένους στόχους miRNA παρέχοντας μια σημαντική αύξηση για διαθέσιμους στόχους που προέρχονται από διαθέσιμες πειραματικές τεχνικές, ενώ περιλαμβάνει ένα ισχυρό σύνολο εργαλείων με μια φιλική προς το χρήστη διεπαφή.

Η επιμέλεια βιβλιογραφίας σχετικής με miRNA είναι μια χρονικά απαιτητική διεργασία που απαιτεί πολύ καλά εκπαιδευμένο προσωπικό. Υλοποιήθηκε μια διοχέτευση (pipeline) ανθρώπινης επιμέλειας με υποβοήθηση από εξόρυξη κειμένου προκειμένου να μειωθεί ο απαιτούμενος χρόνος για επιμέλεια άρθρων. Καθώς η ακριβής αυτόματη εξαγωγή αλληλεπιδράσεων από βιοϊατρική βιβλιογραφία είναι ακόμα ένα ανοιχτό πρόβλημα, η εφαρμογή εξόρυξης κειμένου σχεδιάστηκε σε αυτό το σημείο μόνο για να βοηθά τους επιμελητές στην εργασία τους. Μία απλοϊκή διοχέτευση σχεδιάστηκε η οποία περιλάμβανε αναγνώριση επώνυμων οντοτήτων (named entity recognition - NER), αναγνώριση αλληλεπιδράσεων miRNA-στόχου, βαθμολόγηση και ενισχυμένη παρουσίαση κειμένου (Σχήμα 4.8). Το σχεδιασμένο υποσύστημα παρέχει: επισημείωση των γονιδίων, των miRNA και λέξεων συσχέτισης με διαφορετικούς συμβολισμούς. Στη συνέχεια πραγματοποιείται βαθμολόγηση και ταξινόμηση των περιλήψεων με βάση ένα σκορ πιθανότητας ύπαρξης στόχου. Τέλος προστέθηκε ειδική επισημείωση για ενότητες που έχουν αυξημένο ενδιαφέρον (πχ υπογράμμιση προτάσεων που φαίνόταν να έχουν μεγάλη πιθανότητα να περιέχουν μία συσχέτιση miRNA-στόχου.



Σχήμα 4.8: Η διοχέτευση υποστηριζόμενης επιμέλειας που υιοθετήθηκε.

Αρχικά, όλες οι σχετικές περιλήψεις που περιείχαν όρους σχετικούς με miRNA (όπως microRNA ή micro-RNA ή miRNA ή "micro RNA" στον τίτλο τους, το κείμενο, τις λέξεις-κλειδιά ή τους όρους MeSH συλλέχθηκαν από τη MedLine σε μορφή XML κειμένων. Στη συνέχεια η διεργασία αναγνώρισης επώνυμων οντοτήτων εκτελέστηκε σε δύο διακριτά βήματα: Στο πρώτο βήμα, αναγνωρίστηκαν γονίδια με τη χρήση του AIIAGMT του εργαστηρίου AIIA, ενός από τα καλύτερα συστήματα αναγνώρισης επώνυμων όρων με βάση τις προκλήσεις BioCreative [91]. Το AIIAGMT προσφέρει μια διεπαφή XML-RPC μέσω ενός Perl module, που αναπτύχθηκε από τους συγγραφείς το οποίο ενσωματώθηκε μέσα στη διοχέτευσή μας. Ο κώδικάς μας σημείωσε συγκεκριμένα όλες τις αναφορές γονιδίων που εντοπίστηκαν από το AIIAGMT. Η αναγνώριση των miRNA είναι πολύ πιο εύκολη αφού ακολουθούν πολύ πιο συντηρητική ονοματολογία και δεν έχουν ένα τεράστιο πλήθος συνωνύμων, δύο ζητήματα που είναι κοινός τόπος στο συμβολισμό των γονιδίων. Έτσι η αναγνώριση miRNA πραγματοποιήθηκε μέσω κανονικών εκφράσεων (regular expressions) που υλοποιήθηκαν σε Perl. Μετά από την NER, μια λίστα από stems 16 λέξεων χρησιμοποιήθηκε ως βάση για την αναγνώριση λέξεων συσχέτισης μέσα στο κείμενο. Αυτά τα stems συλλέχθηκαν αφαιρώντας τα επιθέματα (από τα ρήματα ή τα ουσιαστικά) από ένα σύνολο λέξεων που συμβολίζουν συσχετίσεις miRNA-στόχων (π.χ. target-s, target-ing, target-ed κτλ). Τέλος μια συνάρτηση βαθμολόγησης εφαρμόστηκε η οποία προωθούσε τις περιλήψεις που περιείχαν ένα μεγάλο πλήθος προτάσεων που περιείχαν συσχετίσεις miRNA-στόχων. Αυτό το εργαλείο αύξησε σημαντικά το αποτέλεσμα της ημερήσιας επιμέλειας και μας

επέτρεψε να ενσωματώσουμε μεγάλο πλήθος από επιμελημένους στόχους μέσα στην DIANA TarBase.

Ένα νέο σχήμα σχεσιακής βάσης δεδομένων σχεδιάστηκε και υλοποιήθηκε για να υποδέχεται τα τρέχοντα δεδομένα και τις μελλοντικές ανανεώσεις του DIANA TarBase. Το DIANA TarBase v.6 φιλοξενεί ένα σημαντικό εύρος από πληροφορίες για κάθε αλληλεπίδραση miRNA-γονιδίου οι οποίες περιλαμβάνουν από πληροφορίες για miRNA και γονίδια μέχρι και πληροφορίες για την αλληλεπίδρασή τους, τις μεθοδολογίες πειραματικής επιβεβαίωσης και τα επακόλουθα αποτελέσματά τους. Όλες οι καταχωρήσεις της βάσης δεδομένων είναι εμπλουτισμένες με ένα σημαντικό ποσό από δεδομένα σχετικά με τις λειτουργίες όπως και γενικές πληροφορίες οι οποίες προέρχονται από εξωτερικές βάσεις δεδομένων. Οι καταχωρήσεις της βάσης δεδομένων αντιστοιχίζονται σε εξωτερικές πηγές όπως η UniProt, η Ensembl, η RefSeq και άλλες, προκειμένου να παρέχεται ανεμπόδιστη ενσωμάτωση με άλλες υπηρεσίες. Τα σκορ του DIANA microT v.4 [60] και σύνδεσμοι προς τις σχετικές καταχωρήσεις του microT επίσης προστέθηκαν στις αλληλεπιδράσεις. Το νέο, διευρυμμένο σχήμα βάσης δεδομένων σχεδιάστηκε για να υποδέχεται δεδομένα για αλληλεπιδράσεις miRNA-γονιδίων σε μεγάλη λεπτομέρεια και για να είναι αποδοτικό κατά τη διάρκεια της αποτίμησης ερωτημάτων. Η βάση δεδομένων υποστηρίζεται από ένα μεγάλο πλήθος ευρετήρια και υλοποιημένες όψεις (matterialized views) για βελτίωση των επιδόσεων.

Για κάθε καταχώρηση στην DIANA TarBase v.6 οι επιμελητές καταγράφουν το miRNA, το σχετικό στοχευόμενο γονίδιο όπως επίσης και πληροφορία που αφορά στο πείραμα, όπως είναι η σειρά κυττάρων ή ο ιστός που χρησιμοποιήθηκε. Η μεθοδολογία που χρησιμοποιήθηκε (ειδική τεχνική ή τεχνική υψηλής ρυθμαπόδοσης) αναφέρεται συγκεκριμένα. Οι υποστηριζόμενες μεθοδολογίες είναι 'Reporter genes', 'qPCR', 'Western blotting', 'MicroArray', 'Proteomics' (όπως το pSILAC), 'Sequencing' (πχ RNA-Seq, HITS-CLIP, PAR-CLIP), 'Degradome-Seq' και 'Other' (πχ ELISA, RACE, immunohistochemistry, κτλ). Για κάθε καταχώρηση, το πειραματικό αποτέλεσμα (θετικό ή αρνητικό), ο τύπος της συσχέτισης (άμεση ή έμμεση) ο τύπος της ρύθμισης (πάνω ή κάτω), ο τόπος πρόσδεσης, όπως επίσης και το αποτέλεσμα κάθε μεθοδολογίας εισήχθη στη βάση δεδομένων από τους επιμελητές. Ένα μικρό εδάφιο από το άρθρο επίσης προστέθηκε το οποίο κρίθηκε ότι περιέχει σημαντική πληροφορία σχετικά με το πείραμα.

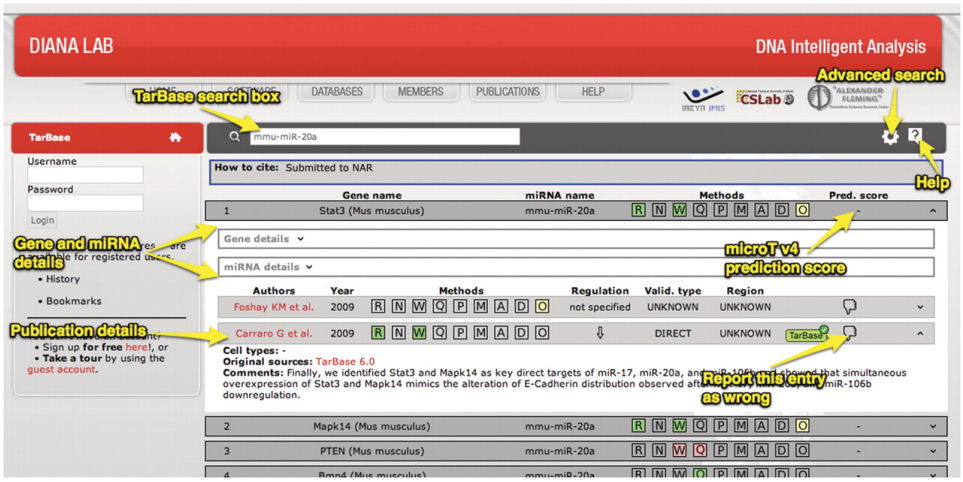
Μη επεξεργασμένα σύνολα δεδομένων από πειράματα υψηλής ρυθμαπόδοσης τα οποία είναι αποθηκευμένα σε σχετικά αποθετήρια ή παρέχονται ως βοηθητικό υλικό από επτά δημοσιεύσεις αναλύθηκαν (τέσσερα microarray [24, 101, 55, 25], δύο HITS-CLIP [16, 108], ένα PAR-CLIP [28]). Τα δεδομένα PAR-CLIP και HITS-CLIP αποτελούνται από γενομικές συντεταγμένες που καθορίζουν του πιθανούς τόπους πρόσδεσης των miRNA. Κάθε θέση σαρώνεται για συμπληρωματικές ακόλουθίες σε σχέση με τους σπόρους όλων των γνωστών miRNA και αν τέτοια εμφάνιση βρεθεί, τότε το γονίδιο σημειώνεται ως στοχευόμενο από το αντίστοιχο miRNA. Για τα δεδομένα PAR-CLIP η αναζήτηση των σπόρων διυλίζεται με τον περιορισμό των περιοχών προς σάρωση εντός 10 νουκλεοτιδίων τριγύρω από την μετάλλαξη του T σε C. Η μετάλλαξη βρέθηκε [28] να προκύπτει κοντά στο σημείο διασταύρωσης (crosslink) της πρωτεΐνης Ago στο miRNA και επομένως είναι καλύτερης ακρίβειας ένδειξη της θέσης πρόσδεσης του τόπου πρόσδεσης του miRNA.

Η νέα βάση δεδομένων περιλαμβάνει 65,814 πειραματικά επιβεβαιωμένες αλληλεπιδράσεις miRNA-γονιδίου οι οποίες εξάγονται από σχετική βιβλιογραφία από τους επιμελητές του DIANA TarBase. Αυτό είναι μία αύξηση των καταχωρήσεων σε σχέση

με την προηγούμενη έκδοση του DIANA TarBase κατά 50 φορές και μία αύξηση από τις άλλες επιμελημένες βάσεις δεδομένων κατά 16.5 με 175 φορές. Το DIANA TarBase v.6 υποδέχεται ένα σημαντικό πλήθος αποτελεσμάτων που παράγονται από τις υψηλής ρυθμαπόδοσης μελέτες αιχμής. Σημαντικό στοιχείο αποτελεί το ότι το DIANA TarBase φιλοξενεί δεδομένα που προέρχονται από 3 μελέτες CLIP-Seq και 12 μελέτες Degradome-Seq, γεγονός που αποτελεί μεγάλη αύξηση συγκριτικά με τις οκτώ μελέτες που υποστηρίζει το StarBase, μια βάση δεδομένων αφιερωμένη στη συλλογή δεδομένων από αυτές τις μεθοδολογίες.

Η νέα διεπαφή του TarBase επιχειρεί να εξισορροπήσει την ευκολία χρήσης με τη λειτουργικότητα. Οι χρήστες μπορούν να φυλλομετρήσουν ένα μεγάλο πλήθος από στόχους miRNA και να διερευνήσουν τα αποτελέσματα από εκατοντάδες πειραματικές μελέτες με απλό και διαισθητικό τρόπο. Παρόλα αυτά, η απλότητα δεν θα έπρεπε να αποτρέπει τους χρήστες από την πραγματοποίηση πολύπλοκων λειτουργιών, όπως την επεξεργασία των αποτελεσμάτων χρησιμοποιώντας διάφορους τύπους φίλτρων ή την εκτέλεση συνδυαστικών αναζητήσεων. Τέτοιες λειτουργίες είναι συνήθως απούσες από τις περισσότερες προηγούμενες βάσεις δεδομένων για στόχους miRNA.

Μέσω του πλαισίου αναζήτησης του DIANA TarBase (Σχήμα 4.9) οι χρήστες μπορούν να αναζητήσουν στόχους εισάγοντας αναγνωριστικά miRNA ή γονιδίων ή ένα συνδυασμό τους (συνδυαστική αναζήτηση). Για τη διευκόλυνση των χρηστών παρέχεται εκτεταμένη υποστήριξη σε ένα μεγάλο πλήθος από ονοματολογίες για miRNA και γονίδια , όπως είναι τα ονόματα miRNA ή τα αναγνωριστικά MIMAT για τα γονίδια και το όνομα, το αναγνωριστικό Ensembl ή το αναγνωριστικό RefSeq για τα γονίδια. Σε περιπτώσεις τυπογραφικών λαθών ή άλλων ασαφειών στην είσοδο του χρήστη το σύστημα παρέχει αυτόματες προτάσεις. Σε αυτές τις περιπτώσεις ο χρήστης μπορεί να επιλέξει οποιαδήποτε από τις επιλογές που του δίνονται για να ολοκληρώσει το ερώτημα στη βάση δεδομένων.



Σχήμα 4.9: Ένα στιγμιότυπο της διεπαφής του DIANA TarBase v.6.

Έπειτα από την υποβολή ενός ερωτήματος χρήστη, το σύστημα απεικονίζει τους σχετικούς στόχους και όλες τις διαθέσιμες σχετικές πληροφορίες σε μορφή μιας επεκτάσιμης λίστας κάτω από το πλαίσιο αναζήτησης του DIANA TarBase . Κάθε γραμμή στη λίστα αντιστοιχεί σε έναν πειραματικά επιβεβαιωμένο στόχο miRNA. Κάθε καταχώρηση συνοδεύεται από ένα σημαντικό πλήθος πληροφοριών που σχετίζοντια με το γονίδιο, το miRNA και την αλληλεπίδρασή τους. Υποστηρικτικά πειραματικά δεδο-

μένα παρέχονται επίσης, όπως είναι οι μέθοδοι που χρησιμοποιήθηκαν, το αποτέλεσμα κάθε μεθόδου και το σκορ πρόβλεψης με βάση τον αλγόριθμο DIANA microT v.4. Το τελευταίο σκορ είναι επίσης υπερσύνδεση προς την αντίστοιχη σελίδα του DIANA microT για να παρέχει λεπτομερή πληροφορία για την πρόβλεψη του στόχου.

Κάθε καταχώρηση στη νέα διεπαφή αποτελεσμάτων είναι επεκτάσιμη, εμφανίζοντας περισσότερες πληροφορίες με έναν εύκολο και διαισθητικό τρόπο. Ο χρήστης μπορεί να εμφανίσει ή να αποκρύψει αυτή την πληροφορία απλώς πατώντας επάνω στα βέλη που εμφανίζονται δίπλα σε κάθε καταχώρηση. Η επιπλέον πληροφορία που παρουσιάζεται περιλαμβάνει (α) λεπτομερή περιγραφή του εμπλεκόμενου γονιδίου και miRNA (όνομα, περιγραφή, σχετικοί όροι MeSH, μετάγραφα, κτλ) και (β) λίστα δημοσιεύσεων που υποστηρίζουν την αλληλεπίδραση μαζί με λεπτομέρειες σχετικά με τις πειραματικές μεθόδους που χρησιμοποιήθηκαν και τα αποτελέσματά τους. Οι πληροφορίες του γονιδίου και του miRNA εμπλουτίζονται με σχετικά μονοπάτια KEGG, συνδέσμους σε εξωτερικές πηγές δεδομένων όπως η UniProt και η Ensembl και ταιριάζονται με κωδικούς αναγνώρισης σε σχετικές εξωτερικές βάσεις δεδομένων, όπως είναι η RefSeq. Τα δεδομένα των αλληλεπιδράσεων υποστηρίζονται από συμβολισμούς για άμεση/έμμεση αλληλεπίδραση, πάνω/κάτω ρύθμιση ενός στόχου, τύπο κυττάρων και ιστών που χρησιμοποιήθηκαν κατά το πείραμα και την περιοχή πρόσδεσης του miRNA στο στόχο. Επιπλέον, προσφέρεται ένας σύνδεσμος στο PubMed προς την υποστηρίζουσα δημοσίευση μαζί με ένα εδάφιο από το κείμενο του άρθρου που καταγράφει χρήσιμη πληροφορία σχετικά με κάθε αλληλεπίδραση.

Η λειτουργία της Προχωρημένης Αναζήτησης (Advanced Search) της διεπαφής του DIANA TarBase βοηθά τους χρήστες να παραμετροποιήσουν τη λίστα των αποτελεσμάτων με βάση τις ιδιαίτερες ανάγκες τους. Οι χρήστες μπορούν να ενεργοποιήσουν αυτή τη λειτουργία πατώντας επάνω στο εικονίδιο γραναζιού στα δεξιά του πλαισίου αναζήτησης (Σχήμα 4.9) για να ενεργοποιήσουν και να εμφανίσουν ένα εκτεταμμένο σύνολο από επιλογές φιλτραρίσματος. Οι χρήστες μπορούν επίσης να παραμετροποιήσουν τις λίστες αποτελεσμάτων επιλέγοντας την επιθυμητή τιμή για κάθε επιλογή. Το σύνολο των παρεχόμενων επιλογών περιλαμβάνει (α) είδη, (β) πειραματική μέθοδο επιβεβαίωσης (δηλ. qPCR, microarray, κτλ), (γ) τύπο ρύθμισης (δηλ. πάνω/κάτω), (δ) τύπο αλληλεπίδρασης (δηλ. άμεσος/έμμεσος), (ε) αποτέλεσμα επιβεβαίωσης (δηλ θετικό/αρνητικό), (στ) χρονιά δημοσίευσης, (ζ) σκορ πρόβλεψης DIANA microT .

Επιπλέον, το DIANA TarBase v.6 παρέχει την επιλογή στο χρήστη να αποκτήσει πρόσβαση σε στόχους που έχουν καταχωρηθεί σε όλες τις άλλες επιμελημένες βάσεις δεδομένων. Συγκεκριμένα, το DIANA TarBase ενσωματώνει καταχωρήσεις από τη miRecords, τη miRTarBase και τη miR2Disease (βλ. επίσης Κεφάλαιο 4.4.1). Αυτή η λειτουργικότητα είναι ανενεργή αρχικά όμως ο χρήστης μπορεί εύκολα να ενεργοποιήσει το υποσύστημα ενσωμάτωσης μέσω του μενού προχωρημένης αναζήτησης. Έπειτα από την ενεργοποίηση, οι εγγραφές από τις εξωτερικές βάσεις δεδομένων ενσωματώνονται στη λίστα αποτελεσμάτων. Κάθε καταχώρηση συνοδεύεται από μία σαφή αναφορά στη βάση δεδομένων από την οποία προέρχεται. Όλα τα στατιστικά που αναφέρθηκαν προηγουμένως σχετικά με τη βάση δεδομένων DIANA TarBase δεν εμπλέκουν τους ενσωματωμένους στόχους ή τις εξωτερικές πηγές αλληλεπιδράσεων miRNA.

Τέλος, ο χρήστης του DIANA TarBase v.6 έχει τη δυνατότητα να αναφέρει λανθασμένες εγγραφές με το πάτημα ενός πλήκτρου ή να υποβάλλει εγγραφές από ένα πείραμα χρησιμοποιώντας εύκολες στη χρήση ενεργές φόρμες. Να σημειωθεί ότι όλες οι διεπαφές χρήστη του DIANA TarBase v.6 υλοποιήθηκαν σε PHP χρησιμοποιώντας καθιερωμένα για χρόνια πρότυπα σχεδιασμού (design pattern) για τον προγραμματισμό

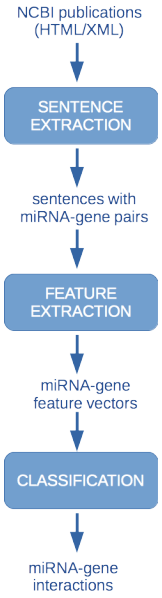
Ιστού (όπως το μοντέλο MVC, οι Active Records, κτλ). Όλα τα απαιτούμενα δεδομένα αποθηκεύτηκαν σε μια σχεσιακή βάση δεδομένων (η MySQL χρησιμοποιήθηκε). Η υπηρεσία ViMa που παρέχεται από την ΕΔΕΤ χρησιμοποιήθηκε για να φιλοξενήσει το εργαλείο Ιστού στη διεύθυνση `http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase`.

4.4.3 Αυτόματη εξαγωγή στόχων miRNA: στην κατεύθυνση του DIANA TarBase v.7

Ένα σημαντικό ζήτημα που αντιμετωπίζουν όλες οι επιμελημένες βάσεις δεδομένων είναι ότι το κόστος προκειμένου να διατηρούνται τα δεδομένα τους ανανεωμένα είναι μεγάλο. Στην περίπτωση του DIANA TarBase, οι ανανεώσεις είναι πολύ ακριβές γιατί οι επιμελητές οφείλουν να ανακαλύπτουν συνεχώς δημοσιεύσεις που μπορεί να περιέχουν πληροφορία σχετικά με στόχους μορίων miRNA και στη συνέχεια να διαβάζουν το περιεχόμενό τους προκειμένου να ανακαλύψουν πιθανές επιβεβαιωμένες αλληλεπιδράσεις miRNA-γονιδίων. Παρόλα αυτά, συνήθως, τα εργαστήρια που διαχειρίζονται τις επιμελημένες βάσεις δεδομένων για αλληλεπιδράσεις miRNA-γονιδίου δεν διαθέτουν τα απαραίτητα κεφάλαια για να προσλάβουν ειδικούς αφοσιωμένους στο σκοπό της συνεχούς επιμέλειας. Έτσι, τελικά, οι περισσότερες από αυτές τις βάσεις δεδομένων είναι καταδικασμένες να περιέχουν παλιά δεδομένα για μεγάλες χρονικές περιόδους.

Προκειμένου να βοηθήσουμε τόσο την DIANA TarBase όσο και άλλες παρόμοιες βάσεις δεδομένων να διατηρήσουν τα δεδομένα τους ανανεωμένα χωρίς να απαιτούνται σημαντικά πρόσθετα έξοδα, αναπτύξαμε ένα σύστημα που διατρέχει ένα δεδομένο σύνολο δημοσιεύσεων και ανακαλύπτει μέσα τους προτάσεις που μπορεί να κωδικοποιούν κάποια αλληλεπίδραση μεταξύ γονιδίου και miRNA. Μια πρότυπη υλοποίηση αυτού τους συστήματος και πρώιμη πειραματική αξιολόγησή του με βάση τις αλληλεπιδράσεις που περιέχονται στην DIANA TarBase v.6 πραγματοποιήθηκε στο [95]. Σκοπός μας είναι να αναπτύξουμε μια πιο σταθερή έκδοση του συστήματος, να πραγματοποιήσουμε ενδελεχή πειράματα για τις επιδόσεις του και, στη συνέχεια, να το καταστήσουμε διαθέσιμο ως έργο ανοιχτού κώδικα.

Το Σχήμα 4.10 απεικονίζει τη διεργασία που ακολουθείται από το σύστημα προκειμένου να αναγνωρίσει τις πραγματικές αλληλεπιδράσεις miRNA-γονιδίου που περιέχονται σε ένα σύνολο από δημοσιεύσεις. Η είσοδος είναι ένα σύνολο δημοσιεύσεων σε μορφή HTML/XML. Τέτοια αρχεία μπορούν να συγκεντρωθούν από τη βάση δεδομένων PMC της NCBI. Το σύστημα διαβάζει τα αρχεία των δημοσιεύσεων και εξάγει εκείνες τις προτάσεις που περιέχουν τουλάχιστον έναν όρο miRNA μαζί με ένα όρο γονιδίου. Τα miRNA αναγνωρίζονται χρησιμοποιώντας μια σχετικά απλή γραμματική αφού υπάρχει επίσημη ονοματολογία που χρησιμοποιείται με κάποιες καθιερωμένες παραλλαγές (για λεπτομέρειες βλ. [95]). Για την αναγνώριση των γονιδίων ειδικά λεξικά χρησιμοποιούνται (πχ τα Entrez Gene και HGNC για τα ανθρώπινα γονίδια, το MGI για τα γονίδια του ποντικιού, κτλ). Οι προτάσεις που περιέχουν αναφορές σε άλλες δημοσιεύσεις δεν λαμβάνονται υπόψη εφόσον θεωρείται ότι το περιεχόμενό τους αναφέρεται στα ευρήματα της δημοσίευσης που αναφέρεται. Επιπλέον, προτάσεις που περιέχουν ζεύγη που δεν εμφανίζονται σε καμία άλλη πρόταση της ίδια δημοσίευσης επίσης δεν λαμβάνονται υπόψη. Αυτό γιατί υπάρχουν κάποιες τυχαίες εμφανίζεις ζευγών όρων miRNA και γονιδίου. Παρατηρήσαμε ότι αν μια μελέτη υποστηρίζει ότι υπάρχει στόχος ενός μορίου miRNA σε ένα συγκεκριμένο γονίδιο, τότε, συνήθως, το miRNA και το γονίδιο εμφανίζονται μαζί σε πολλές προτάσεις του κειμένου.



Σχήμα 4.10: Η διεργασία αναγνώρισης αλληλεπιδράσεων miRNA-γονιδίων στη βιβλιογραφία.

Στις προτάσεις που απομένουν πραγματοποιείται επεξεργασία φυσικής γλώσσας (NLP) προκειμένου να εξαχθεί πληροφορία που απαιτείται για να κατασκευαστούν διανύσματα χαρακτηριστικών για κάθε ζεύγος miRNA-γονιδίου το οποίο θα χρησιμοποιηθεί στη συνέχεια από έναν κατηγοριοποιητή που αποφασίζει αν ένα ζεύγος αντιστοιχεί σε πραγματική αλληλεπίδραση miRNA-γονιδίου ή όχι. Κατά τη διάρκεια του NLP σε κάθε λέξη της πρότασης εφαρμόζεται αποκοπή καταλήξεων (χρησιμοποιώντας τον Morpha Stemmer [64]) και αναγνωρίζεται το μέρος του λόγου της (χρησιμοποιώντας τον Stanford Tagger[7] [94]). Επιπλέον, παράγονται οι εξαρτήσεις μεταξύ των λέξεων κάθε πρότασης χρησιμοποιώντας τον Dependency Parser[8] [44] του πακέτου Stanford NLP. Τέλος, κάθε πρόταση διαχωρίζεται σε φράσεις από ένα δικής μας υλοποίησης λογισμικό διαχωρισμού (chunking).

Όλα τα δεδομένα που παράγονται από την NLP διεργασία χρησιμοποιούνται για να διαμορφωθούν διανύσματα χαρακτηριστικών για κάθε ζεύγος miRNA-γονιδίου. Αυτά τα διανύσματα χαρακτηριστικών ανιχνεύουν κοινά πρότυπα κειμένου που εμφανίζονται σε προτάσεις που περιγράφουν πειραματικά επιβεβαιωμένες αλληλεπιδράσεις miRNA-γονιδίων. Κάθε διάνυσμα χαρακτηριστικών που παράγεται από κάθε δημοσίευση προωθείται σε έναν κατηγοριοποιητή Δυαδικής Μέγιστης Εντροπίας (Binary Maximum Entropy). Η έξοδος αυτού του κατηγοριοποιητή για το διάνυσμα χαρακτηριστικών ενός ζεύγους miRNA-γονιδίου που περιέχεται σε μια δεδομένη δημοσίευση αντιστοιχεί στην πιθανότητα ότι η δημοσίευση περιέχει πειραματικά αποτελέσματα που επιβεβαιώνουν ότι το συγκεκριμένο miRNA στοχεύει το συγκεκριμένο γονίδιο. Αν αυτή η πιθανότητα είναι μεγαλύτερη από 0.5 θεωρούμε ότι η αναφερόμενη αλληλεπίδραση είναι πραγματική. Να σημειωθεί ότι ο κατηγοριοποιητής εκπαιδεύτηκε χρησιμοποιώντας δεδομένα από την DIANA TarBase v.6. Για περισσότερες λεπτομέρειες βλ. [95].

Ο Πίνακας 4.1 παρουσιάζει τις επιδόσεις του συστήματος που περιγράφηκε πιο πάνω με βάσει κάποια πρώιμα αποτελέσματα που παρουσιάστηκαν στο [95]. Κατά τη διάρ-

[7]http://nlp.stanford.edu/software/tagger.shtml
[8]http://nlp.stanford.edu/software/lex-parser.shtml

κεια αυτών των πειραμάτων παρατηρήσαμε ότι πολλές επιβεβαιωμένες αλληλεπιδράσεις μεταξύ miRNA και γονιδίων δεν παρατίθενται μέσω του κειμένου της δημοσίευσης αλλά μέσω υποστηρικτικών σχημάτων, φύλλων Excel ή πινάκων. Το να ληφθούν υπόψη εκείνες οι αλληλεπιδράσεις που δεν κωδικοποιούνται στο κείμενο των δημοσιεύσεων ήταν έξω από τη στόχευση του [95] και μπορεί να αποτελέσει τμήμα μελλοντικής εργασίας. Προκειμένου να αξιολογήσουμε τις επιδόσεις του υλοποιημένου προτύπου συστήματος έπρεπε να απομονώσουμε από τα πειράματά μας τις περισσότερες δημοσιεύσεις που περιέχουν πολλές αλληλεπιδράσεις στο βοηθητικό υλικό. Θεωρήσαμε ότι είναι απίθανο μία δημοσίευση που διατυπώνει περισσότερες από 10 αλληλεπιδράσεις να κωδικοποιεί τη συγκεκριμένη πληροφορία μέσα στο κείμενο. Έτσι μετρήσαμε την ακρίβεια, το ποσοστό ανάκτησης και το σκορ-F1 για το σύστημά μας χρησιμοποιώντας όλες τις δημοσιεύσεις του συνόλου δεδομένων μας που περιείχαν το πολύ 2, 3, 5, ή 10 αλληλεπιδράσεις. Να σημειωθεί ότι το σύνολο δεδομένων μας περιέχει 1,167 δημοσιεύσεις που συλλέχθηκαν από την NCBI PMC και τις αλληλεπιδράσεις miRNA-γονιδίου από την DIANA TarBase v.6. Αυτό το σύνολο δεδομένων διαχωρίστηκε σε σύνολο εκπαίδευσης (85% των δημοσιεύσεων) και σύνολο δοκιμών (το υπόλοιπο 15%). Οι μετρήσεις στον Πίνακα 4.1 αναφέρονται στις επιδόσεις του υλοποιημένου προτύπου συστήματος πάνω στο σύνολο δοκιμών.

Μεγ. #αλληλ. ανά δημ.	**Ακρίβεια (%)**	**Ανάκτηση (%)**	**Σκορ-F1 (%)**
2	76.9	69.1	72.6
3	79.9	66.2	72.3
5	76.7	60.5	67.6
10	71.8	55.9	62.8

Πίνακας 4.1: Πρώιμη αξιολόγηση της αυτόματης εξαγωγής αλληλεπιδράσεων miRNA-γονιδίων.

Είναι σαφές ότι το πρωτότυπο σύστημα είναι αξιόπιστο σε ικανοποιητικό βαθμό για να παρέχει προτάσεις στους επιμελητές της DIANA TarBase. Αν ληφθούν επίσης υπόψη οι αλληλεπιδράσεις που αναφέρονται στο βοηθητικό υλικό των δημοσιεύσεων, το σύστημα θα είναι ακόμα πιο πρακτικό.

4.5 DIANA mirPath: Ανακαλύπτοντας το ρόλο των miRNA στα μεταβολικά μονοπάτια

Στη βιοχημεία, τα μεταβολικά μονοπάτια είναι σειρές χημικών αντιδράσεων που συμβαίνουν εντός ενός κυττάρου. Σε κάθε μονοπάτι, ένα βασικό χημικό στοιχείο μετατρέπεται από μια σειρά χημικών αντιδράσεων. Τα ένζυμα καταλύουν αυτές τις αντιδράσεις, και συχνά απαιτούν διατροφικά μεταλλικά στοιχεία, βιταμίνες, και άλλους συμπαράγοντες προκειμένου να λειτουργήσουν σωστά. Ενώ αυτή η συμμετοχή των γονιδίων στα μονοπάτια έχει μελετηθεί σχετικά καλά, δεν ισχύει το ίδιο για τα miRNA.

Το miTalos [47] είναι ένα λογισμικό που μπορεί να χρησιμοποιηθεί για την ανάλυση ενός υποσυνόλου ανθρώπινων μονοπατιών σήμανσης. Αναγνωρίζει στόχους χρησιμοποιώντας πέντε διαφορετικούς εξωτερικούς αλγορίθμους πρόβλεψης στόχων miRNA ενώ λαμβάνει επίσης υπόψη τα δεδομένα έκφρασης. Επιπλέον, το miRTar [30] μπορεί να χρησιμοποιηθεί για την αναζήτηση εναλλακτικώς κατακερματισμένων στόχων

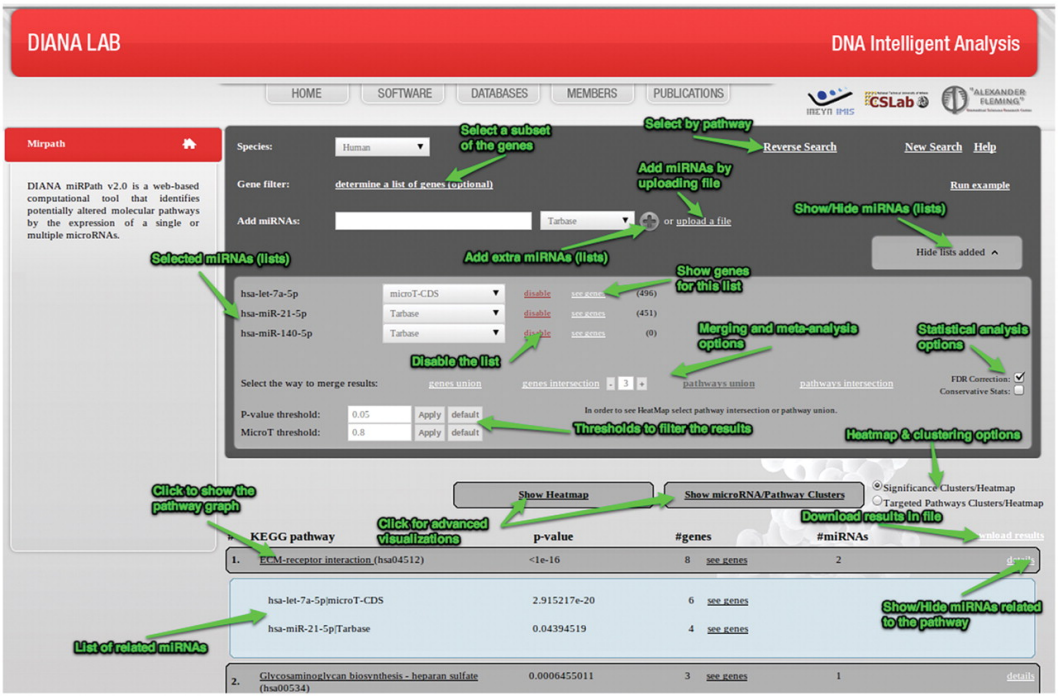
miRNA, που αναγνωρίζονται ενσωματώνοντας εξωτερικούς αλγορίθμους πρόβλεψης. Το miRTar μπορεί επίσης να κάνει ανάλυση εμπλουτισμού συνόλων γονιδίων για την αναγνώριση των μονοπατιών στόχων miRNA. Το GeneTrail [6] είναι εξυπηρετητής Ιστού που φιλοξενεί υπηρεσίες εμπλουτισμού συνόλων γονιδίων και έχει ικανότητες υπερ-αναπαράστασης έναντι σε διάφορα σύνολα δεδομένων, όπως τα GO και KEGG. Το GeneTrail έχει επεκταθεί με ένα εργαλείο, το οποίο αναζητά αναγνωριστικά miRNA έναντι στη βάση δεδομένων MicroCosm Targets για θεωρούμενους στόχους γονιδίων [7]. Παρέχει πολλές επιλογές κατά τη διαδικασία ανάλυσης εμπλουτισμού γονιδίων, όπως τα κατώφλια τιμών P και διόρθωση σημαντικότητας μέσω πολλαπλών ελέγχων (multiple testing significance correction) αλλά δεν έχει ειδικές λειτουργίες για τα miRNA, όπως πληροφορίες σχετικά με θέσεις πρόσδεσης και τύπους πρόσδεσης. Τέλος, το DIANA mirPath v.1 [73] ήταν μια από τις πρώτες διαθέσιμες εφαρμογές που επικεντρώθηκαν στην ανάλυση εμπλουτισμού προβλεπόμενων γονιδίων στόχων, ικανή να εντοπίσει μονοπάτια που στοχεύονται από ένα ή πολλά miRNA.

Ο εξυπηρετητής Ιστού DIANA mirPath v.2 [100] είναι ένας εξ ολοκλήρου ανασχεδιασμένος εξυπηρετητής Ιστού με πολλά νέα χαρακτηριστικά. Στοχεύει στην σημαντική αύξηση της ακρίβειας των χρησιμοποιούμενων αλγορίθμων και στατιστικών, καθώς και στην ενίσχυση της υπολογιστικής ταχύτητας, συγκριτικά με την προηγούμενη έκδοση του DIANA mirPath. Κάτι πολύ σημαντικό, το DIANA mirPath v.2 προσφέρει για πρώτη φορά μια σειρά εργαλείων ειδικά επικεντρωμένων στην ανάλυση μονοπατιών που στοχεύονται από τα miRNA. Ο χρήστης του DIANA mirPath v.2 μπορεί να χρησιμοποιήσει προβλεπόμενους ή πειραματικά επικυρομένους στόχους, να συνδυάσει τα αποτελέσματα με αλγορίθμους συγχώνευσης και μετα-ανάλυσης, να εκτελέσει ιεραρχική συσταδοποίηση των miRNA και των μονοπατιών βάσει των επιπέδων αλληλεπίδρασής τους, καθώς και να εξηγήσει αναλυτικά εξεζητημένες οπτικοποιήσεις, όπως δεντρογράμματα ή χάρτες θερμότητας αλληλεπίδρασης miRNA/μονοπατιού, από μια διαισθητική και εύχρηστη διεπαφή Ιστού. Ο νέος εξυπηρετητής παρέχει επιπλέον πληροφορίες σχετικά με παθογόνα SNP στα προβλεπόμενα σημεία στόχους των miRNA. Επιπλέον, το υποσύστημα αντίστροφης ανάλυσης σημειώνει όλα τα προβλεπόμενα ή πειραματικά επικυρωμένα miRNA που στοχεύουν ένα επιλεγμένο μοριακό μονοπάτι.

Το DIANA mirPath v.2 βασίζεται σε ένα νέο σχεσιακό σχήμα, ειδικά σχεδιασμένο για να το φιλοξενεί, αυτό καθώς και μελλοντικές ανανεώσεις. Πληροφορίες σχετικά με τα miRNA και τα μονοπάτια αποκτήθηκαν από το miRBase 18 και την Kyoto Encyclopedia of Genes and Genomes (KEGG) v58.1 [39], αντίστοιχα.

Η διεπαφή DIANA mirPath v.2 (Σχήμα 4.11) έχει σχεδιαστεί να είναι υψηλά προσαρμόσιμη σε διαφορετικά σενάρια χρήσης και να παρέχει αποτελέσματα σε πραγματικό χρόνο. Για να πραγματοποιήσει την ανάλυση, ο χρήστης μπορεί να επιλέξει ένα ή περισσότερα miRNA και την πηγή των γονιδίων στόχων για κάθε miRNA. Προαιρετικά, μπορεί να φορτωθεί μια λίστα εκφραζόμενων γονιδίων. Συνεπώς, ο εξυπηρετητής παρουσιάζει τα σημαντικά εμπλουτισμένα μονοπάτια, τα γονίδια στόχους σε κάθε μονοπάτι και τον αριθμό των miRNA με θετικά αναγνωρισμένους στόχους για κάθε μονοπάτι στη μορφή διαδραστικού πίνακα.

Στην περίπτωση προβλεπόμενων αλληλεπιδράσεων miRNA-γονιδίων, ο εξυπηρετητής παρέχει ένα σύνδεσμο στις καταχωρήσεις του σχετικού εξυπηρετητή DIANA microT v.5 (Κεφάλαιο 4.2). Εκεί, ο χρήστης μπορεί να ερευνήσει περαιτέρω την προβλεπόμενη αλληλεπίδραση miRNA-γονιδίου. Τέτοιες αλληλεπιδράσεις περιλαμβάνουν την περιοχή, τη θέση και τον τύπο πρόσδεσης. Εάν μια αλληλεπίδραση miRNA-γονιδίου

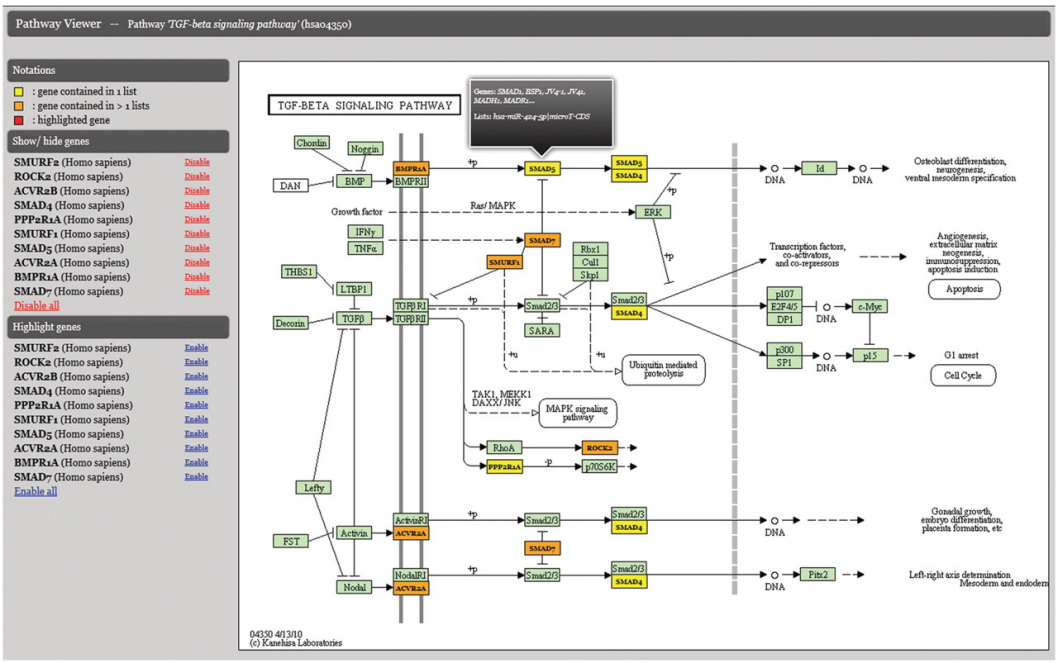


Σχήμα 4.11: Μια αποτύπωση της διεπαφής DIANA mirPath .

είναι πειραματικά επικυρωμένη, ο εξυπηρετητής παρέχει ένα σύνδεσμο προς τη συγκεκριμένη ενότητα της ιστοσελίδας του DIANA TarBase v.6 (Κεφάλαιο 4.4). Η σχετική καταχώρηση παρέχει πληροφορίες σχετικά με την υλοποιημένη πειραματική μέθοδο που χρησιμοποιήθηκε για την επικύρωση και την υποστηρικτική βιβλιογραφία. Το DIANA mirPath v.2 προσφέρει εμπλουτισμένες οπτικοποιήσεις μονοπατιών KEGG, όπου τα γονίδια στόχοι είναι ειδικά μαρκαρισμένα για καλύτερο εντοπισμό (Σχήμα 4.12).

Το νέο υποσύστημα αντίστροφης αναζήτησης μπορεί να χρησιμοποιηθεί για την αναγνώριση όλων των miRNA που προβλέπονται ή έχουν επικυρωθεί πειραματικά να στοχεύουν ένα συγκεκριμένο μονοπάτι KEGG. Το υποσύστημα παίρνει ως είσοδο ένα όνομα ή αναγνωριστικό μονοπατιού KEGG και την πηγή των στόχων miRNA. Έπειτα αναγνωρίζει όλα τα miRNA που στοχεύουν το επιλεγμένο μονοπάτι. Το νέο υποσύστημα μπορεί να γίνει ισχυρό εργαλείο για τους επιστήμονες που μελετούν συγκεκριμένα μονοπάτια. Μπορεί να βοηθήσει την εξέταση επικυρωμένων σχέσεων μεταξύ μονοπατιών και miRNA που εκφράζονται στη διαθέσιμη βιβλιογραφία (στόχοι DIANA TarBase) ή τη μελέτη νέων αλληλεπιδράσεων miRNA-μονοπατιών (στόχοι DIANA microT). Εάν η ανάλυση εκτελείται in silico, ο χρήστης μπορεί να προσδιορίσει τα επιθυμητά επίπεδα για ευαισθησία και ακρίβεια εφαρμόζοντας ένα κατώφλι σκορ DIANA microT v.5.

Το DIANA mirPath v.2 παρέχει επίσης προηγμένα χαρακτηριστικά, στατιστικά και βοηθήματα οπτικοποίησης, τα οποία αυξάνουν σημαντικά το βάθος της ανάλυσης και μεγιστοποιούν την επίδραση του χρήστη στον υπολογισμό και την παρουσίαση των αποτελεσμάτων. Επιπλέον, είναι ικανό να εκτελεί επαναληπτικές αναλύσεις, όπως ιεραρχική συσταδοποίηση στοχευμένων μονοπατιών και miRNA. Το DIANA mirPath v.2 πραγματοποιεί συσταδοποίηση των επιλεγμένων miRNA βάσει της επιρροής τους στα μοριακά μονοπάτια. Παρέχει συσταδοποίηση μονοπατιών βάσει του υποσυνόλου των miRNA που στοχεύει κάθε μονοπάτι και του επιπέδου σημαντικότητας της αλληλεπί-



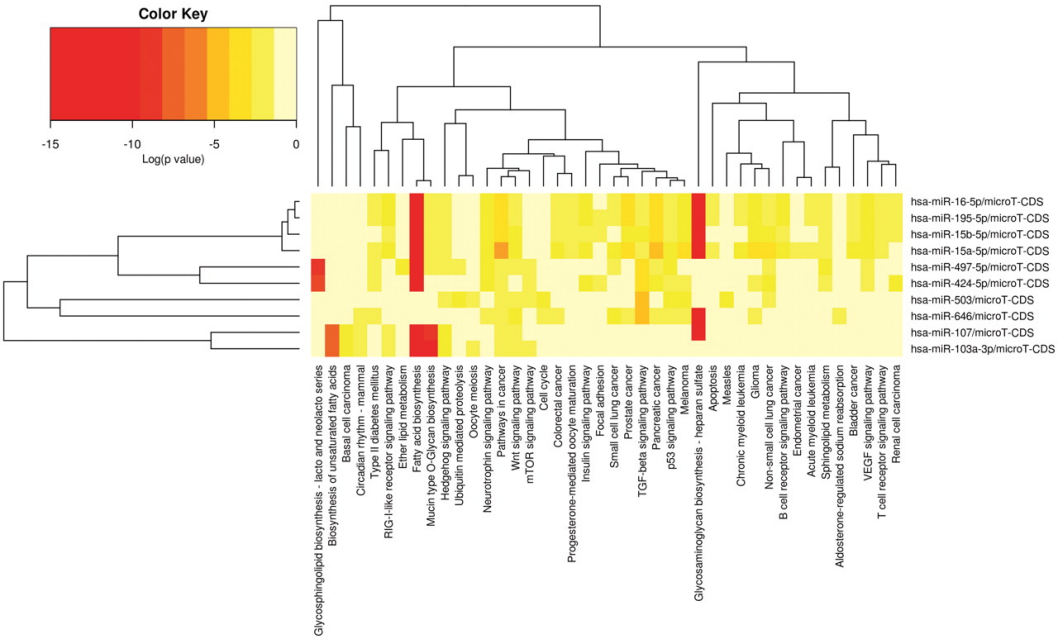
Σχήμα 4.12: Μια αποτύπωση της οπτικοποίησης ενός μονοπατιού KEGG.

δρασής τους. Ο εξυπηρετητής εκτελεί την ιεραρχική ανάλυση συστάδων βάσει μιας πλήρους μεθόδου συσταδοποίησης linkage, όπου τετραγωνικές Ευκλείδιες αποστάσεις υπολογίζονται ως μέτρα απόστασης. Ο εξυπηρετητής Ιστού μπορεί να χρησιμοποιήσει απόλυτες τιμές P σε όλους τους υπολογισμούς (επιλογή: 'Significance Clusters') ή δυαδικές τιμές (0=μη στοχευμένο, 1=στοχευμένο), εάν προτιμηθεί η επιλογή 'Targeted Pathways Clusters'. Χρησιμοποιώντας αυτές τις επιλογές, ο αλγόριθμος μπορεί να συσταδοποιήσει microRNA που στοχεύουν παρόμοιες λίστες μονοπατιών, καθώς και μονοπάτια, που στοχεύονται από παρόμοιες λίστες microRNA (Targeted Pathways Clusters), μπορεί επίσης να λάβει υπόψη τα επίπεδα σημαντικότητας των αλληλεπιδράσεων (Significance Clusters) κατά τη διαδικασία συσταδοποίησης.

Αυτά τα προηγμένα χαρακτηριστικά μπορούν να βοηθήσουν το χρήστη να αναγνωρίσει σχέσεις μεταξύ miRNA ή μονοπατιών ανάλογα με το ενεργό μέγεθος των αλληλεπιδράσεων miRNA-μονοπατιών. Ο εξυπηρετητής Ιστού παρέχει οπτικοποιήσεις της ιεραρχικής συσταδοποίησης στη μορφή δεντρογραμμάτων miRNA και μονοπατιών.

Τέλος, ο νέος εξυπηρετητής DIANA mirPath δίνει στο χρήστη τη δυνατότητα να δημιουργήσει προηγμένες οπτικοποιήσεις όπως χάρτες θερμότητας miRNA έναντι μονοπατιών (Σχήμα 4.13). Οι χάρτες θερμότητας είναι γραφικές αναπαραστάσεις των δεδομένων όπου οι τιμές σε ένα μητρώο αναπαριστώνται σαν χρώματα. Αυτές οι διαισθητικές οπτικοποιήσεις έχουν αποδειχτεί χρήσιμες σε διάφορους κλάδους, καθώς δίνουν στους χρήστες τη δυνατότητα να αναγνωρίσουν μονοπάτια στα δεδομένα, που δεν ήταν εύκολα ορατά όταν εξετάζονταν οι παράμετροι ανεξάρτητα. Επιπλέον, καθιστούν ικανή την οπτικοποίηση ενός πολύ μεγάλου αριθμού μεταβλητών, των μεταξύ τους σχέσεων και των επιπέδων αλληλεπίδρασής τους. Ο εξυπηρετητής Ιστού χρησιμοποιεί τα αποτελέσματα ιεραρχικής συσταδοποίησης και στους δυο άξονες (μονοπάτια και miRNA), προκειμένου να κατασκευάσει την οπτικοποίηση του χάρτη θερμότητας. Όπως στην περίπτωση της ανάλυσης συστάδων, ο εξυπηρετητής παρέχει δυο επιλογές για υπολογισμό χάρτη θερμότητας: 'Significance Heat Maps' και 'Targeted Pathways

Heat Maps′. Η πρώτη αφορά τη χρήση απόλυτων τιμών P σε όλους τους υπολογισμούς, ενώ η δεύτερη αντικαθιστά όλες τις τιμές P που είναι χαμηλότερες από το κατώφλι που έχει ορίσει ο χρήστης με 0, και 1 διαφορετικά. Με τη χρήση αυτών των προηγμένων εργαλείων, η χρήστης μπορεί να εξετάσει πολλές σχέσεις miRNA-miRNA, miRNA-μονοπατιού και μονοπατιού-μονοπατιού. Τέτοιες αναπαραστάσεις μπορούν να βοηθήσουν τους ερευνητές να ανακαλύψουν πρότυπα και σχέσεις που είναι κρυμμένα μέσα στα δεδομένα. Όλα τα διαγράμματα σχεδιάζονται σε υψηλή ανάλυση.



Σχήμα 4.13: Ένα παράδειγμα χάρτη θερμότητας από τη διεπαφή DIANA mirPath .

Όλες οι διεπαφές χρήστη DIANA mirPath v.2 υλοποιήθηκαν σε PHP με χρήση καλά καθιερωμένων προτύπων σχεδιασμού για Προγραμματισμό (όπως τα MVC model, Active Records, κλπ). Όλα τα απαραίτητα δεδομένα αποθηκεύτηκαν σε μια σχεσιακή βάση δεδομένων (χρησιμοποιήθηκε η MySQL). Σενάρια που εκτελούν την εμπλεκόμενη στατιστική ανάλυση αναπτύχθηκαν στην R. Τέλος, η υπηρεσία ViMa παρασχέθηκε από την GRNET και χρησιμοποιήθηκε για τη φιλοξενία του ηλεκτρονικού εργαλείου στο `http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=mirpath`.

4.6 DIANA mirPub: Αναζητώντας δημοσιεύσεις σχετικές με miRNA

Η αναγνώριση, ανάμεσα στις εκατομμύρια δημοσιεύσεις που είναι διαθέσιμες στην MEDLINE, αυτών που είναι σχετικές με ένα συγκεκριμένο miRNA ενδιαφέροντος με βάση αναζήτηση με λέξεις-κλειδιά αντιμετωπίζει σημαντικά προβλήματα. Το βασικό πρόβλημα σχετίζεται με την ονοματολογία των miRNA. Συχνά χρησιμοποιείται ασυνεπώς έχοντας ως αποτέλεσμα να είναι είναι πιθανό, για παράδειγμα, το ίδιο miRNA να αναφέρεται με δύο διαφορετικά ονόματα σε δύο διαφορετικά άρθρα. Ακόμα χειρότερα, σε ορισμένες περιπτώσεις η ίδια η ονοματολογία εξελίσσεται. Για παράδειγμα, μια δη-

μοσίευση μπορεί να αναφέρεται σε ονόματα miRNA τα οποία δε χρησιμοποιούνται πια στις μέρες μας.

Παρουσιάζουμε εδώ τη DIANA mirPub, μια βάση δεδομένων με μια διαισθητική διεπαφή που παρέχει ένα ισχυρό εργαλείο για την αποτελεσματική αναζήτηση δημοσιεύσεων που σχετίζονται με miRNA. Η μηχανή αναζήτησης της DIANA mirPub λαμβάνει υπόψη της όχι μόνο παραλλαγές ονομάτων miRNA, αλλά επίσης και αλλαγές που μπορεί να συμβούν στα ονόματα και στις ακολουθίες miRNA (με βάση όλες τις διαθέσιμες εκδόσεις της miRBase). Για να ανακαλύψει συσχετίσεις miRNA με δημοσιεύσεις εκμεταλλεύεται τόσο την εξόρυξη κειμένου πάνω στη MEDLINE όσο και δεδομένα που προέρχονται από επιμελημένες βάσεις δεδομένων. Επιπλέον, η DIANA mirPub ακολουθεί μια προσέγγιση crowdsourcing: οι χρήστες της μπορούν να υποβάλλουν τα δικά τους δεδομένα για να βοηθήσουν την επιμέλεια των σχετικών με miRNA δημοσιεύσεων. Μια άλλη σημαντική λειτουργία της mirPub είναι μια διαδραστική υπηρεσία οπτικοποίησης που απεικονίζει διαισθητικά την εξέλιξη των δεδομένων miRNA. Άλλες προσφερόμενες λειτουργίες περιλαμβάνουν νέφη ετικετών που συνοψίζουν τη σχέση δημοσιεύσεων με συγκεκριμένες ασθένειες, κυτταρικές σειρές και ιστούς και πρόσβαση σε δεδομένα του DIANA TarBase προκειμένου ο χρήστης να μπορεί να κάνει επισκόπηση των γονιδίων που σχετίζονται με τις δημοσιεύσεις miRNA.

4.6.1 Περιγραφή του DIANA mirPub

Σε αυτό το κεφάλαιο, περιγράφουμε το σχεδιασμό και την ανάπτυξη της βάσης δεδομένων και της εφαρμογής Ιστού mirPub και επιδεικνύουμε τη λειτουργικότητα της mirPub και τις υπηρεσίες της. Συγκεκριμένα, στο Κεφάλαιο 4.6.1.1 περιγράφουμε πώς συλλέγουμε συσχετίσεις miRNA με δημοσιεύσεις. Στο Κεφάλαιο 4.6.1.2, παρουσιάζουμε μεθόδους που καταγράφουν την εξέλιξη δεδομένων των miRNA μέσω της επεξεργασίας πολλαπλών εκδόσεων της miRBase. Τέλος, στο Κεφάλαιο 4.6.1.3 περιγράφουμε τη διεπαφή του mirPub και τα βασικά του χαρακτηριστικά.

4.6.1.1 Συσχετίσεις miRNA-δημοσιεύσεων

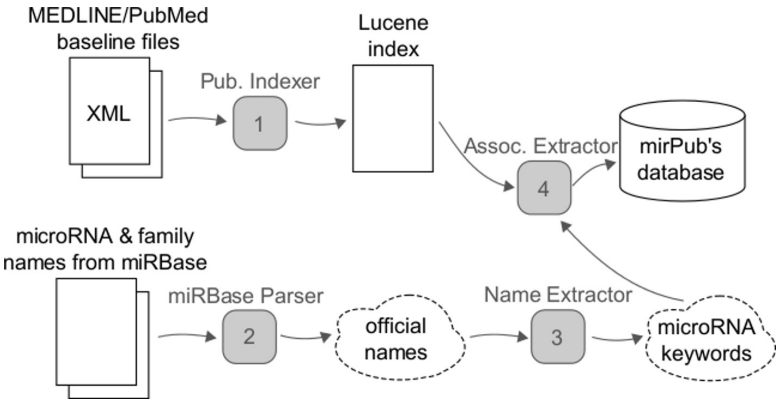
Το mirPub παρέχει συσχετίσεις miRNA με δημοσιεύσεις χρησιμοποιώντας τις παρακάτω τρεις πηγές:

1. Μια επιμελημένη βάση δεδομένων που βασίζεται σε συσχετίσεις miRNA με δημοσιεύσεις που έχουν ήδη καταγραφεί στη miRBase, την DIANA TarBase και την mir2disease.

2. Συσχετίσεις που προέρχονται από την εφαρμογή τεχνικών εξόρυξης κειμένου σε τίτλους και περιλήψεις που προέρχονται από δημοσιεύσεις της MEDLINE/PubMed.

3. Δεδομένα που προέρχονται από τους χρήστες, αφού οι χρήστες του mirPub μπορούν να συνεισφέρουν στο περιεχόμενό του με το να αναφέρουν προβληματικές εγγραφές ή με το να προτείνουν συσχετίσεις miRNA-δημοσιεύσεων που δεν υπάρχουν στη βάση δεδομένων.

Να σημειωθεί ότι μια σημαντική συνεισφορά εδώ είναι ότι για το ζητούμενο της εξόρυξης λαμβάνουμε υπόψη την εξέλιξη της ονοματολογίας των miRNA και τις παραλλαγές στα ονόματα των miRNA. Στη συνέχεια παρουσιάζουμε τη μέθοδο εξόρυξης λεπτομερώς.

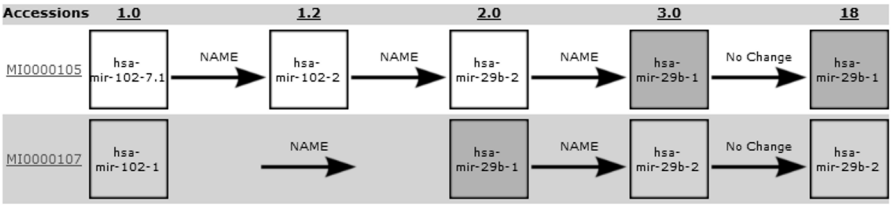


Σχήμα 4.14: Η ροή εργασιών της εξόρυξης συσχετίσεων miRNA-δημοσιεύσεων.

Η βασική ιδέα της μεθόδου εξόρυξης κειμένου πάνω στα δεδομένα της MEDLINE/PubMed απεικονίζεται στο Σχήμα 4.14. Εν συντομία, πραγματοποιούμε αναζήτηση μέσα στους τίτλους, τις περιλήψεις και τα πλήρη κείμενα (όποτε αυτά είναι διαθέσιμα) σε όλα τα άρθρα προκειμένου να εντοπίσουμε τις εμφανίσεις λέξεων-κλειδιών που περιγράφουν miRNA, αφού αυτές οι εμφανίσεις υπονοούν συσχετίσεις miRNA-δημοσιεύσεων. Για να υποστηρίξουμε αποδοτική αναζήτηση, οργανώνουμε τα δεδομένα των δημοσιεύσεων σε βολικές μορφές αναπαράστασης και δομές δεδομένων. Συνεπώς, στο πρώτο βήμα, επιλέγουμε το Lucene για να χτίσουμε ένα ανεστραμμένο ευρετήριο πάνω στα βασικά αρχεία του MEDLINE/Pubmed (υποσύστημα *publication indexer*). Το ευρετήριο παρέχει αποδοτική πρόσβαση σε περιλήψεις και τίτλους που περιέχουν μια λέξη-κλειδί και παρέχει επίσης άλλα χρήσιμα μεταδεδομένα (πχ τον τίτλο της δημοσίευσης, τους σχετικούς όρους MeSH κτλ).

Μετά από την κατασκευή του ευρετηρίου, ο *αναλυτής miRBase* εκτελείται για να παράγει όλα τα ονόματα miRNA και οικογενειών miRNA που καταγράφονται στη miRBase. Συγεκριμένα, όλα τα .dat αρχεία από τις εκδόσεις 1.0 έως 18 της miRBase χρησιμοποιούνται προκειμένου να συλλεχθούν τα ονόματα των φουρκετών και ώριμων miRNA και το αρχείο .fam της έκδοσης miRBase 18 χρησιμοποιείται για να συλλεχθούν όλα τα ονόματα οικογενειών. Αναφερόμαστε στην ένωση των προηγούμενων ονομάτων ως το σύνολο των *επίσημων ονομάτων*. Στη συνέχεια επεκτείνουμε αυτό το σύνολο προκειμένου να περιέχει επίσης και τις παραλλαγές τους, επειδή τα άρθρα δεν περιέχουν πάντα ακριβείς εμφανίσεις των επίσημων ονομάτων. Αυτή είναι η εργασία του υποσυστήματος *εξαγωγέας ονομάτων*.

Ο εξαγωγέας ονομάτων χρησιμοποιεί ένα σύνολο προκαθορισμένων κανόνων για να παράγει μεταβλητές των επίσημων ονομάτων. Μια πρώτη τάξη μεταβλητών περιέχει ονόματα από τα οποία παραλείπεται το πρόθεμα του είδους (στην ουσία, αυτή είναι μια πολύ συχνή χρήση ονόματος). Μια άλλη τάξη περιέχει τις μεταβλητές που παράγονται αντικαθιστώντας μερικές λεκτικές μονάδες (token) ενός όρου με άλλες. Για παράδειγμα, η λεκτική μονάδα "mir" στις λέξεις κλειδιά miRNA συχνά αντικαθίσταται από τις λεκτικές μονάδες "mirna" ή "microrna", η λεκτική μονάδα "let" αντικαθίσταται από τα "mir-let", "mirna-let", ή "microrna-let". Να σημειωθεί ότι παρόμοιοι κανόνες έχουν χρησιμοποιηθεί επίσης στο ([105]) για την αναγνώριση των ονομάτων miRNA. Η έξοδος του εξαγωγέα ονομάτων είναι η ένωση όλων των επίσημων ονομάτων και όλων των παραγόμενων μεταβλητών τους (αναφερόμαστε σε αυτή την ένωση ως το σύνολο των λέξεων-κλειδιών των *miRNA*).



Σχήμα 4.15: Το χρονοδιάγραμμα εξέλιξης του όρου "hsa-mir-29b-1" (που εμπλέκει τα miRNA-φουρκέτες MI0000105 και MI0000107).

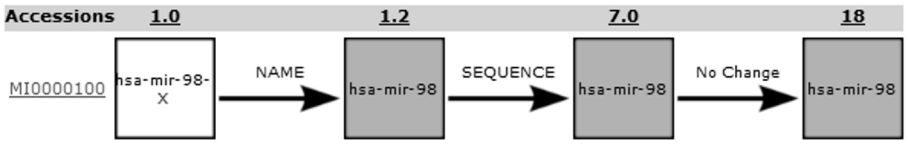
Το τελικό βήμα εκτελείται από τον *εξαγωγέα συσχετίσεων*. Σκοπός του είναι να χρησιμοποιήσει όλες τις λέξεισ-κλειδιά για τα miRNA για να διερευνήσει το ευρετήριο Lucene. Το αποτέλεσμα είναι μια λίστα εγγραφών $< keyword,\ PubMed\ ID >$ που ορίζουν στην ουσία τις συσχετίσεις miRNA-δημοσιεύσεων. Αποθηκεύουμε αυτές τις εγγραφές στη σχεσιακή βάση δεδομένων του mirPub.

Η μέθοδος εξόρυξης κειμένου μπορεί να αποτύχει να ανακαλύψει κάποιες από τις συσχετίσεις miRNA-δημοσιεύσεων. Για παράδειγμα, δεν λαμβάνουμε υπόψη μας δημοσιεύσεις που δεν περιέχουν λέξεις-κλειδιά miRNA ούτε στον τίτλο ούτε στην περίληψή τους. Η εξόρυξη κειμένου στο πλήρες κείμενο μπορεί να επιστρέψει πολλές από τις δημοσιεύσεις που χάνουμε, παρόλα αυτά, αυτός ο τύπος επεξεργασίας μπορεί να εφαρμοστεί μόνο σε άρθρα που είναι ανοιχτής πρόσβασης, τα οποία είναι απλώς ένα κομμάτι της υπάρχουσας βιβλιογραφίας. Η χρήση δεδομένων από επιμελημένες βάσεις δεδομένων (όπως τη miRBase, την TarBase κτλ) μπορεί να βοηθήσει προς αυτή την κατεύθυνση. Παρόλα αυτά, ακόμα και αυτή η λύση δεν είναι ικανοποιητική αφού τέτοιες βάσεις δεδομένων συνήθως δεν παρέχουν ανανεώσεις σε τακτική βάση. Ακόμα και αν το κάνουν, ο χρόνος που περνά ανάμεσα σε δύο συνεχόμενες ανανεώσεις μπορεί να είναι μεγάλος. Για το λόγο αυτό, στη mirPub ενθαρρύνουμε τους χρήστες να αναφέρουν νέες συσχετίσεις miRNA-δημοσιεύσεων, όπως επίσης και λάθη. Ένα πρωτόκολλο έχει οριστεί για να εγγυάται ότι κάθε νέα αναφερόμενη συσχέτιση θα εξετάζεται από έναν επιμελητή πριν την τελική αποδοχή του μέσα στη βάση δεδομένων.

4.6.1.2 Καταγράφοντας την εξέλιξη δεδομένων miRNA

Αφού η έρευνα για τα miRNA είναι συνεχώς σε εξέλιξη, νέες δημοσιεύσεις τροποποιούν τα αποτελέσματα προηγούνενων εργασιών και έτσι το περιεχόμενο των σχετικών βάσεων δεδομένων, όπως της miRBase, πρέπει να μετατρέπεται για να συμφωνεί με αυτές τις αλλαγές. Εφόσον πολλές από αυτές τις αλλαγές περιλαμβάνουν τα ονόματα των miRNA, η καταγραφή της ιστορίας της ονοματολογίας κάθε μορίου miRNA είναι σημαντική για κάθε ερευνητή που αναζητά δημοσιεύσεις που συνδέονται με ένα συγκεκριμένο μόριο. Αυτό συμβαίνει γιατί η γνώση παλαιότερων ονομάτων ενός μορίου ή άλλων μορίων που είχαν το ίδιο νόημα στο παρελθόν μπορεί να βοηθήσει τους ερευνητές τόσο να επεκτείνουν την αναζήτησή τους ώστε να περιλαμβάνει πιο πολλά αποτελέσματα όπως επίσης και να αφαιρέσουν άσχετα άρθρα από τη λίστα αποτελεσμάτων. Στο mirPub επεξεργαστήκαμε όλες τις διαθέσιμες εκδόσεις της miRBase για να εξάγουμε χρήσιμη πληροφορία για την εξέλιξη των miRNA δεδομένων. Αναφερόμαστε λεπτομερώς στο συγκεκριμένο θέμα στις επόμενες παραγράφους.

Στη miRBase κάθε ώριμο miRNA και κάθε miRNA-φουρκέτα έχει ένα όνομα και μία ακολουθία που μπορεί να είναι διαφορετικά από έκδοση σε έκδοση, όπως επίσης και ένα αναγνωριστικό πρόσβασης το οποίο παραμένει πάντα σταθερό και αναγνωρί-



Σχήμα 4.16: Το χρονοδιάγραμμα εξέλιξης του όρου "hsa-mir-98" (που εμπλέκει το miRNA-φουρκέτα MI0000100).

ζει το μόριο μοναδικά σε όλες τις εκδόσεις miRBase. Όλα αυτά τα δεδομένα μαζί με μεταδεδομένα αποθηκεύονται σε αρχεία που ανανεώνονται ανά έκδοση και τα οποία έχουν κατάληξη .dat και θα τα ονομάζουμε, από εδώ και πέρα, *dat αρχεία* της miRBase. Τα αρχεία dat ακολουθούν μια δομή παρόμοια με αυτή των αρχείων EMBL και έχουν μία εγγραφή για κάθε ξεχωριστό αναγνωριστικό miRNA-φουρκέτας. Κάθε εγγραφή περιέχει ένα όνομα, μία ακολουθία, τα παραγόμενα ώριμα miRNA, τις σχετικές δημοσιεύσεις κτλ, μιας συγκεκριμένης φουρκέτας. Τα δεδομένα των ώριμων miRNA αποθηκεύονται ως υπο-εγγραφές στις εγγραφές των φουρκετών που τα παράγουν.

Με τη σύγκριση του αρχείου dat μίας έκδοσης προς το αρχείο dat της επόμενης έκδοσης όλες οι αλλαγές για ώριμα miRNA και miRNA-φουρκέτες μπορούν να παραχθούν. Η miRBase παρέχει επίσης και ένα *αρχείο διαφορών (diff file)* για κάθε έκδοση, το οποίο μπορεί να επεξεργαστεί για να εξαχθεί η ίδια πληροφορία. Όμως διαθέσιμα αρχεία διαφορών υπάρχουν μόνο για εκδόσεις της miRBase που είναι νεότερες από την 3.1. Συνεπώς, όλα τα δεδομένα εξέλιξης δεδομένων miRNA παράγονται συγκρίνοντας τα dat αρχεία των εκδόσεων miRBase μεταξύ τους (από την έκδοση 1.0 έως 18). Χρησιμοποιήσαμε τα παρεχόμενα αρχεία διαφορών μόνο για να επιβεβαιώσουμε τα αποτελέσματα από την προηγούμενη ανάλυση για τις εκδόσεις που είναι νεότερες της 3.1. Να σημειωθεί ότι η miRBase δεν παρέχει αναγνωριστικά πρόσβασης για ώριμα miRNA σε εκδόσεις που είναι παλαιότερες από την 6.0. Γι΄ αυτό το λόγο καταγράφουμε τις αλλαγές που αφορούν στα ώριμα miRNA μόνο για εκδόσεις της miRBase που είναι νεότερες από την 5.1.

Έπειτα από την ανάλυση όλων των αρχείων miRBase από την έκδοση 1.0 έως την 18, αναγνωρίσαμε τα παρακάτω είδη αλλαγών:

- Για ώριμα miRNA και για miRNA-φουρκέτες:
  1. NEW: ένα νέο miRNA εισήχθη στην τρέχουσα έκδοση της miRBase
  2. NAME: ένα όνομα miRNA μεταβλήθηκε
  3. SEQUENCE: μια ακολουθία miRNA μεταβλήθηκε
  4. NAME-SEQUENCE: τόσο το όνομα όσο και η ακολουθία ενός miRNA μεταβλήθηκαν
  5. DELETE: ένα miRNA αφαιρέθηκε από τη miRBase και το αναγνωριστικό του έγινε παρωχημένο
- Μόνο για φουρκέτες:
  1. FORWARD: το αναγνωριστικό ενός miRNA αντικαταστάθηκε από ένα άλλο
- Για ζεύγη ώριμων-φουρκετών:

1. ADD MATURE-HAIRPIN ASSOC: ένα ώριμο-miRNA βρέθηκε ότι παράγεται από ένα συγκεκριμένο miRNA-φουρκέτα
2. REMOVE MATURE-HAIRPIN ASSOC: ένα ώριμο-miRNA βρέθηκε ότι δεν παράγεται από ένα συγκεκριμένο miRNA-φουρκέτα

Επεξεργαστήκαμε όλα τα αρχεία της miRBase, συλλέξαμε δεδομένα που σχετίζονται με αλλαγές που προκύπτουν στα miRNA και αποθηκεύσαμε αυτά τα δεδομένα σε μία σχεσιακή βάση δεδομένων.

4.6.1.3 Λειτουργίες και διεπαφή

Οι υπηρεσίες του mirPub παρέχονται μέσω μιας διεπαφής Ιστού ελεύθερης πρόσβασης. Στα επόμενα κεφάλαια, περιγράφουμε λεπτομερώς ζητήματα που σχετίζονται με τη λειτουργικότητα της διεπαφής του mirPub δείχνοντας, για παράδειγμα, πώς εκτελείται η αναζήτηση δημοσιεύσεων και πώς συγκεκριμενοποιούνται τα αποτελέσματα αναζήτησης με τη χρήση φίλτρων, πώς γίνεται πλοήγηση στα χρονοδιαγράμματα απεικόνισης της εξέλιξης των miRNA δεδομένων και πώς οι χρήστες μπορούν να συνεισφέρουν στο περιεχόμενο της mirPub.

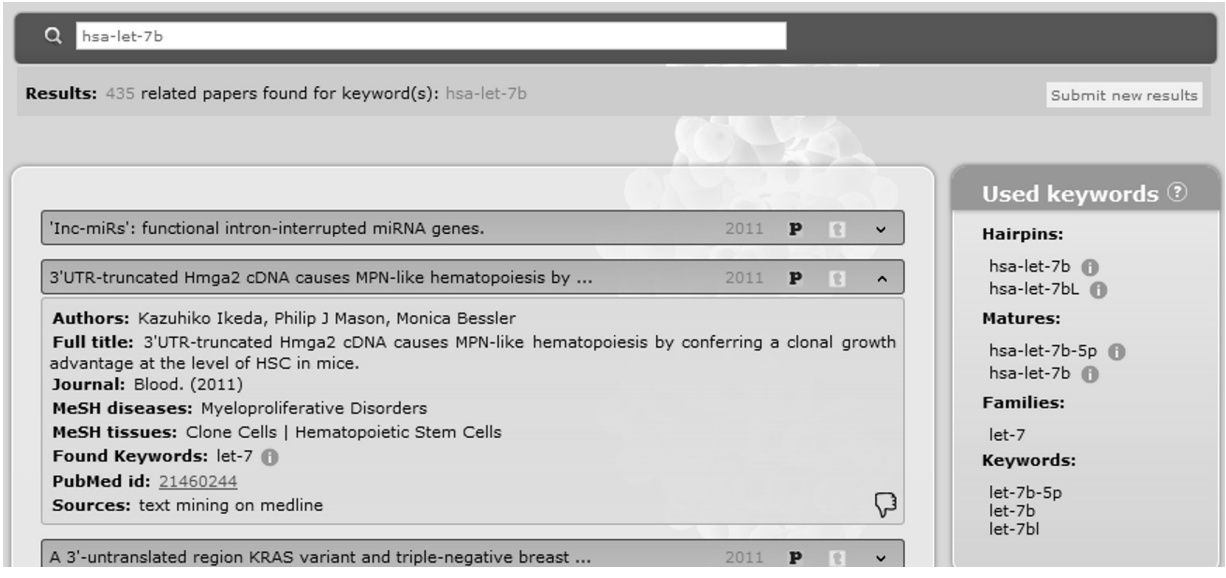
4.6.1.3.1 Αναζήτηση miRNA δημοσιεύσεων Ο χρήστης μπορεί να αναζητήσει δημοσιεύσεις που σχετίζονται με συγκεκριμένα miRNA εισάγοντας λέξεις-κλειδιά που περιγράφουν αυτά τα miRNA σε ένα πλαίσιο αναζήτησης της διεπαφής του mirPub (βλ. επίσης το Σχήμα 4.17). Η mirPub αντιστοιχίζει αυτές τις *λέξεις-κλειδιά* στις αποθηκευμένες λέξεις-κλειδιά miRNA (βλ. επίσης Κεφάλαιο 4.6.1.1). Αν δεν είναι εφικτό να βρεθεί ένα ακριβές ταίριασμα για μία λέξη-κλειδί χρήστη, τότε η mirPub εντοπίζει τις πιο όμοιες σε αυτή λέξεις-κλειδιά και τις προτείνει στο χρήστη. Για να μετρήσουμε την ομοιότητα λέξεων-κλειδιών υιοθετούμε μια παραλλαγή της συντακτικής απόστασης που θεωρεί ότι οι διαγραφές έχουν μηδενικό κόστος. Επιπλέον, εάν μία λέξη-κλειδί του χρήστη μπορεί να αντιστοιχιστεί σε περισσότερες από μία αποθηκευμένες λέξεις-κλειδιά, τότε η mirPub παρουσιάζει όλες τις διαθέσιμες επιλογές στο χρήστη και του ζητά να επιλέξει ποια από αυτές να συμπεριλάβει στην αναζήτησή του. Σε οποιαδήποτε άλλη περίπτωση, η mirPub παράγει τη λίστα δημοσιεύσεων που βρέθηκαν να είναι σχετικές με τις αντιστοιχισμένες λέξεις-κλειδιά (βλ. Σχήμα 4.17). Αυτή η λίστα διατάσσεται με βάση την ημερομηνία δημοσίευσης των άρθρων (τα πιο πρόσφατα άρθρα εμφανίζονται πρώτα). Κάθε αντικείμενο της λίστας είναι ένα πτυσσόμενο ορθογώνιο που απεικονίζει (στη συμπιεσμένη μορφή του) τον τίτλο και την ημερομηνία της δημοσίευσης. Εάν υπάρχουν, εμφανίζονται επίσης δύο σύνδεσμοι, ένας προς το PubMed και ένας προς το TarBase. Όταν ο χρήστης επιλέγει να επεκτείνει το ορθογώνιο, περισσότερη πληροφορία για τη δημοσίευση γίνεται διαθέσιμη: οι πηγές που χρησιμοποιήθηκαν για να συσχετίσουν τη δημοσίευση με το ερώτημα του χρήστη και οι MeSH ασθένειες, ιστοί και σειρές κυττάρων που σχετίζονται με τη δημοσίευση.

Να σημειωθεί ότι όταν ο χρήστης υποβάλλει ένα σύνολο από λέξεις-κλειδιά μέσω του πλαισίου αναζήτησης, η mirPub επεκτείνει το σύνολο αυτό προσθέτοντας και άλλες σχετικές λέξεις-κλειδιά miRNA. Συγκεκριμένα, για κάθε αναγνωρισμένο miRNA η mirPub προσθέτει στο σύνολο των λέξεων-κλειδιών αναζήτησης το σχετικό όνομα οικογένειας, σχετικά ονόματα με βάση την ιστορία και όλες τις γνωστές παραλλαγές ονομάτων (με βάση κανόνες που παρουσιάζονται στο Κεφάλαιο 4.6.1.1). Όλες οι λέξεις-κλειδιά miRNA που χρησιμοποιήθηκαν για να παραχθούν τα αποτελέσματα

εμφανίζονται ως υπερσύνδεσμοι στο πλαίσιο "used keywords" που εμφανίζεται στα δεξιά της λίστας αποτελεσμάτων (βλ. Σχήμα 4.17). Μετακινώντας το δείκτη του ποντικιού πάνω από οποιαδήποτε από αυτές τις λέξεις-κλειδιά, οι σχετικές με αυτή δημοσιεύσεις υπογραμμίζονται. Ο χρήστης μπορεί να επιλέξει να κρατήσει μόνο τα αποτελέσματα που σχετίζονται με κάποιες από τις λέξεις-κλειδιά απλώς πατώντας μία φορά πάνω στους αντίστοιχους υπερσυνδέσμους. Πατώντας και πάλι πάνω σε μία επιλεγμένη λέξη-κλειδί, η επιλογή εξαφανίζεται. Να σημειωθεί ότι οι χρησιμοποιούμενες λέξεις-κλειδιά οργανώνονται σε τέσσερις ξεχωριστές κατηγορίες: (α) φουρκέτες, (β) ώριμα miRNA, (γ) οικογένειες και (δ) λέξεις-κλειδιά (βλ. τις αντίστοιχες ετικέτες στο Σχήμα 4.17). Η κατηγορία "hairpins" (αντίστοιχα "matures") περιέχει ακριβή ονόματα φουρκετών (αντίστοιχα ώριμων miRNA) από οποιαδήποτε έκδοση της miRBase. Αυτά τα ονόματα ακολουθούνται από ένα πλήκτρο πληροφοριών το οποίο ενεργοποιεί ένα αναδυόμενο παράθυρο που απεικονίζει την ιστορία των miRNA που σχετίζονται με αυτή τη λέξη-κλειδί (βλ. επίσης Κεφάλαιο 4.6.1.3.2). Η κατηγορία "families" περιέχει ονόματα οικογενειών από την τελευταία έκδοση της miRBase. Τέλος, η κατηγορία "keywords" περιέχει παραλλαγές των ονομάτων miRNA.

Πέρα από την παρουσίαση των σχετικών ασθενειών, ιστών και σειρών κυττάρων MeSH σε κάθε καταχώρηση της λίστας αποτελεσμάτων, η mirPub συναθροίζει αυτή την πληροφορία για όλες τις καταχωρήσεις σε δύο χρήσιμα σύννεφα ετικετών (tag cloud): ένα που συγκεντρώνει τις σχετικές στις απεικονιζόμενες δημοσιεύσεις ασθένειες MeSH και ένα δεύτερο που συγκεντρώνει τους σχετικούς ιστούς και σειρές κυττάρων MeSH. Το μέγεθος γραμματοσειράς κάθε όρου MeSH εξαρτάται από το πλήθος των σχετικών άρθρων στο σύνολο αποτελεσμάτων.

4.6.1.3.2 Οπτικοποιώντας την εξέλιξη miRNA δεδομένων Οι χρήστες της mirPub μπορούν να εξερευνήσουν τις αλλαγές σε ονόματα, ακολουθίες miRNA και άλλες αλλαγές στα δεδομένα των miRNA με ένα εργαλείο που οπτικοποιεί το χρονοδιάγραμμα αυτών των αλλαγών. Αυτό το εργαλείο είναι προσβάσιμο μέσω του πλήκτρου πληροφοριών που συνοδεύει κάθε αναφορά σε ένα όνομα miRNA μέσα στη διεπαφή χρήστη της mirPub (πχ. υπάρχουν τέσσερα τέτοια πλήκτρα στο πλαίσιο "used keywords" του Σχήματος 4.17). Πατώντας πάνω σε αυτό το πλήκτρο εμφανίζεται ένα



Σχήμα 4.17: Στιγμιότυπο από τη διεπαφή της mirPub.

αναδυόμενο παράθυρο το οποίο περιέχει την ιστορία όλων των ώριμων miRNA και miRNA-φουρκετών που έχουν συσχετιστεί με αυτό το όνομα για τουλάχιστον μία έκδοση της miRBase στο παρελθόν.

Το Σχήμα 4.15 παρουσιάζει το χρονοδιάγραμμα που παράγεται για το "hsa-mir-29b-1" και το Σχήμα 4.16 το χρονοδιάγραμμα που παράγεται για το "hsa-mir-98". Το πρώτο όνομα ήταν συνδεδεμένο με τη φουρκέτα MI0000105 για τις εκδόσεις 2.0 ως 2.2 της miRBase και με τη φουρκέτα MI0000107 από την έκδοση 3.0 μέχρι και σήμερα. Συνεπώς, ένα χρονοδιάγραμμα για καθεμία από αυτές τις φουρκέτες εμφανίζεται (βλ. Σχήμα 4.15). Αυτά τα χρονοδιαγράμματα δείχνουν ότι το "hsa-mir-29b-1" αρχικά είχε ανατεθεί στο MI0000107 μετά από μία αλλαγή τύπου NAME κατά τη διάρκεια της ανανέωσης της miRBase στην έκδοση 2.0 (το προηγούμενο όνομα της συγκεκριμένης φουρκέτας ήταν το "hsa-mir-102-1"). Έπειτα, στην έκδοση 3.0, το ίδιο όνομα ανατέθηκε στο MI0000105, ενώ το MI0000107 μετονομάστηκε σε "hsa-mir-29b-2". Από την άλλη, το όνομα "hsa-mir-98" συνδέεται με τη φουρκέτα MI0000100 για όλες τις εκδόσεις που είναι νεότερες από την 1.2. Το χρονοδιάγραμμα αλλαγών για αυτή τη φουρκέτα (που απεικονίζεται στο Σχήμα 4.16) δείχνει ότι το όνομα "hsa-mir-90" ανατέθηκε σε αυτή μετά από μια αλλαγή τύπου NAME. Στη συνέχεια, στην έκδοση 7.0, η miRBase μετέβαλλε την ακολουθία αυτής της φουρκέτας. Να σημειωθεί ότι το πορτοκαλί χρώμα χρησιμοποιείται σε κάθε χρονοδιάγραμμα για να υπογραμμίσει τις καταστάσεις των φουρκετών που αφορούν το όνομα που χρησιμοποιήθηκε για να παραχθούν τα χρονοδιαγράμματα. (πχ "hsa-mir-29b-1" στο Σχήμα 4.15 και "hsa-mir-98" στο Σχήμα 4.16). Επιπλέον, να σημειωθεί ότι για λόγους παρουσίασης η τελευταία κατάσταση μιας φουρκέτας ανήκει στην τελευταία έκδοση miRBase πάντα, έτσι μια ψευδοαλλαγή που συνδέεται σε αυτή την τελευταία κατάσταση (με τον τίτλο "No Change") απαιτείται στις περισσότερες περιπτώσεις.

4.6.1.3.3 Δεδομένα συνεισφερόμενα από τους χρήστες Όπως αναφέρθηκε στο Κεφάλαιο 4.6.1.1, το mirPub παρέχει στους χρήστες του τη δυνατότητα να υποβάλλουν δεδομένα έτσι ώστε να αυξήσει την ακρίβειά του και ως μια επιπλέον επιλογή για να παραμένει ανανεωμένο. Συγκεκριμένα, υπάρχουν δύο διαθέσιμες λειτουργίες: (α) αιτήματα για νέες συσχετίσεις miRNA-δημοσιεύσεων μπορούν να υποβληθούν και (β) μια υπάρχουσα συσχέτιση miRNA-δημοσίευσης μπορεί να αναφερθεί ως λανθασμένη. Το πρώτο γίνεται πατώντας στο πλήκτρο "Submit new results" που βρίσκεται κάτω από το πλαίσιο αναζήτησης (βλ. στα δεξιά του Σχήματος 4.17), ενώ το δεύτερο γίνεται πατώντας στο κουμπί "thumbs down" που βρίσκεται στην λεπτομερή όψη του κάθε αποτελέσματος (βλ. επίσης Σχήμα 4.17). Και οι δύο ενέργειες έχουν ως αποτέλεσμα την απεικόνιση μιας νέας σελίδας που περιέχει μια φόρμα που συγκεντρώνει δεδομένα από το χρήστη. Τα δεδομένα που συλλέγονται γίνονται μέρος της βάσης δεδομένων μόνο αφού περάσουν από την έγκριση ενός ειδικού επιμελητή. Το τελικό βήμα είναι απαραίτητο για να διατηρηθεί ένα υψηλό επίπεδο ποιότητας δεδομένων.

4.6.2 Πειραματική αξιολόγηση

Σε αυτό το κεφάλαιο, πρώτα αναλύουμε τη βάση δεδομένων της mirPub και παρουσιάζουμε ενδιαφέροντα στατιστικά σχετικά με τις δημοσιεύσεις που είναι σχετικές με miRNA και με την εξέλιξη miRNA δεδομένων. Μετά αξιολογούμε την αποδοτικότητα της mirPub κατά τη διάρκεια ανάκτησης βιβλιογραφίας miRNA.

4.6.2.1 Δημοσιεύσεις miRNA και στατιστικά για την εξέλιξη των miRNA δεδομένων

Αυτή τη στιγμή η βάση δεδομένων mirPub περιέχει περισσότερα από 57,436 διακριτά ζεύγη λέξης-κλειδιού miRNA και PubMed_id που εμπλέκουν περισσότερες από 7,471 διακριτές δημοσιεύσεις. Ο Πίνακας 4.2 συγκρίνει τη mirPub με τις πιο γνωστές βάσεις δεδομένων που περιέχουν δημοσιεύσεις που είναι σχετικές με miRNA. Η σύγκριση γίνεται με βάση το πλήθος των άρθρων που βρέθηκαν να συνδέονται με τουλάχιστον ένα miRNA. Είναι έκδηλο ότι η mirPub περιέχει το μεγαλύτερο πλήθος από δημοσιεύσεις που σχετίζονται με miRNA σε σύγκριση με τις άλλες βάσεις δεδομένων.

Πίνακας 4.2: Σύγκριση διαφόρων βάσεων δεδομένων που περιέχουν δημοσιεύσεις που σχετίζονται με miRNA.

	#άρθρων
mirPub	7,471
miRBase	407
TarBase	1,392
mir2disease	519
miRCancer	573

Ο Πίνακας 4.3 συνοψίζει μερικά ενδιαφέροντα στατιστικά για τις δημοσιεύσεις που είναι αποθηκευμένες στη βάση δεδομένων της mirPub. Συγκεκριμένα παρουσιάζονται το μέγιστο και το μέσο πλήθος λέξεων-κλειδιών miRNA, MeSH ασθενειών, MeSH ιστών και σειρών κυττάρων ανά δημοσίευση. Μία δημοσίευση περιέχει 2,510 ξεχωριστές λέξεις-κλειδιά miRNA (από ένα σύνολο 43,908). Αυτό συμβαίνει επειδή αυτή η δημοσίευση είναι μια μεγάλη έρευνα έκφρασης miRNA [48] η οποία πραγματοποιήθηκε ακολουθιοποιώντας πάνω από 250 βιβλιοθήκες από μικρά RNA 26 διαφορετικών συστημάτων οργάνων και σειρών κυττάρων. Έτσι περιέχει ένα μεγάλο πλήθος από σχετικές λέξεις-κλειδιά miRNA. Ο Πίνακας 4.3 παρουσιάζει επίσης το μέγιστο και μέσο πλήθος από πηγές που χρησιμοποιήθηκαν για να ανακτήσουν καθένα από τα άρθρα. Να σημειωθεί επίσης ότι η mirPub χρησιμοποιεί 5 πηγές δεδομένων: τη MEDLINE/PubMed, τη miRBase, την TarBase, τη mir2disease και δεδομένα συνεισφοράς των χρηστών.

Πίνακας 4.3: Μερικά ενδιαφέροντα στατιστικά για τις δημοσιεύσεις της mirPub.

	Μεγ.	**Μες.**
miRNA kwds/paper	2,510	9.71
diseases/paper	8	1.86
tissues & cells/paper	8	1.58
sources/paper	4	1.32

Περισσότερα στατιστικά μπορούν να βρεθούν στον Πίνακα 4.4, όπου παρουσιάζονται τα πιο δημοφιλή επιστημονικά περιοδικά και όροι MeSH που συνδέονται με άρθρα miRNA.

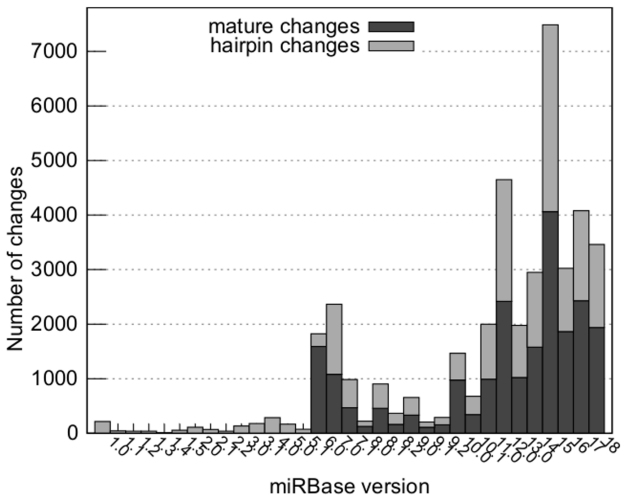
Τα Σχήματα 4.18 και 4.19, απεικονίζουν το πλήθος αλλαγών που εισάγονται σε κάθε έκδοση της miRBase για φουρκέτες-miRNA και ώριμα miRNA. Συγκεκριμένα, το πρώτο παρουσιάζει το πλήθος των αλλαγών τύπου NEW και DELETE ενώ το δεύτερο το πλήθος των αλλαγών τύπου NAME, SEQUENCE και NAME-SEQUENCE για

Πίνακας 4.4: Τα πιο σημαντικά περιοδικά που περιέχουν δημοσιεύσεις σχετικές με miRNA και οι πιο σημαντικές MeSH ασθένειες που σχετιζόνται με miRNA με βάση τις δημοσιεύσεις.

	Περιοδικό	**#άρθρων (ποσ.)**
1	**PloS one**	380 (5.09%)
2	**Proc. of the Nat. Acad. of Sciences of the USA**	241 (3.23%)
3	**Cancer research**	210 (2.81%)
4	**The Journal of biological chemistry**	208 (2.78%)
5	**Biochemical & biophysical res. comm.**	150 (2.01%)
6	**Blood**	132 (1.77%)
7	**Nucl. Acids Research**	122 (1.63%)
8	**Oncogene**	115 (1.54%)
9	**Cell cycle (Georgetown, Tex.)**	88 (1.18%)
10	**Journal of virology**	86 (1.15%)

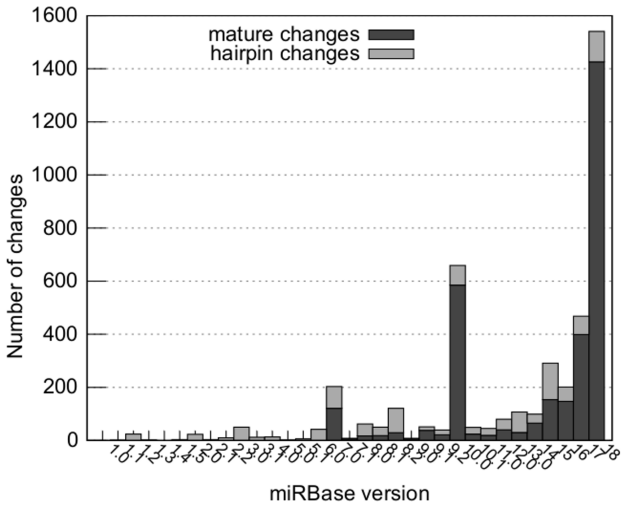
	Ασθένειες MeSH	**#άρθρων**
1	**Neoplasm Invasiveness**	317
2	**Breast Neoplasms**	305
3	**Neoplasms**	239
4	**Lung Neoplasms**	237
5	**Liver Neoplasms**	218
6	**Carcinoma, Hepatocellular**	198
7	**Disease Models, Animal**	194
8	**Cell Transform., Neoplastic**	168
9	**Neoplasm Metastasis**	156
10	**Carcinoma, Squamous Cell**	150

κάθε έκδοση miRBase. Στα ιστογράμματα οι ανοιχτές γκρι μπάρες αναπαριστούν αλλαγές φουρκετών ενώ οι σκούρες γκρι μπάρες αναπαριστούν αλλαγές ώριμων miRNA. Η γενική τάση είναι ότι κάθε νέα έκδοση της miRBase εισάγει περισσότερες αλλαγές από τις προηγούμενες εκδόσεις.



Σχήμα 4.18: Το πλήθος των αλλαγών τύπου NEW & DELETE που εισάγονται σε κάθε έκδοση της miRBase.

Τέλος, μια ενδιαφέρουσα παρατήρηση είναι ότι μια αλλαγή miRNA μπορεί συχνά να ενεργοποιεί μια άλλη αλλαγή. Για παράδειγμα, η εισαγωγή μιας νέας φουρκέτας στη miRBase συνήθως ακολουθείται από μια εισαγωγή τουλάχιστον ενός ώριμου miRNA. Στην πραγματικότητα, η ανάλυσή μας δείχνει ότι το 86.29% των αλλαγών τύπου NEW για ώριμα miRNA ακολουθούν μία αλλαγή τύπου NEW για φουρκέτες.



Σχήμα 4.19: Το πλήθος των αλλαγών τύπου NAME, SEQUENCE και NAME-SEQUENCE που εισάγονται σε κάθε έκδοση της miRBase.

4.6.2.2 Αξιολογώντας πειραματικά την ανάκτηση βιβλιογραφίας

Η MirPub περιέχει ένα υποσύνολο των δημοσιεύσεων του PubMed. Παρόλα αυτά, είναι πιο αποδοτική όταν χρησιμοποιείται για την αναζήτηση δημοσιεύσεων που σχετίζονται με συγκεκριμένα miRNA. Η δύναμη της mirPub βασίζεται στο μεγάλο σύνολο από συσχετίσεις miRNA με δημοσιεύσεις που περιέχει. Οι πιο πολλές από αυτές τις συσχετίσεις δεν μπορούν να ανακτηθούν από την ίδια την PubMed γιατί δεν περιλαμβάνει όλα τα επιμελημένα δεδομένα που περιέχει η mirPub και επειδή δεν μπορεί να εκμεταλλευτεί τις παραλλαγές των miRNA ονομάτων (βλ. το Κεφάλαιο 4.6.1.1) και την ιστορία τους (βλ. Κεφάλαιο 4.6.1.2).

Για να αξιολογήσουμε την ικανότητα της mirPub έναντι της PubMed κατά την ανάκτηση βιβλιογραφίας σχετικής με miRNA εξετάσαμε τα αποτελέσματα που επιστρέφονται δοθέντος ενός συνόλου ερωτημάτων από 50 τυχαία επιλεγμένα ονόματα miRNA. Παραμετροποιήσαμε τις μηχανές αναζήτησης τόσο της mirPub όσο και της PubMed ώστε να ανακτούν για την ίδια χρονική περίοδο μόνο αποτελέσματα που περιέχουν ακριβείς εμφανίσεις των όρων αναζήτησης. Στη mirPub αυτό μπορεί να γίνει εύκολα επιλέγοντας τις κατάλληλες λέξεις-κλειδιά από το φίλτρο "Used keywords". Να σημειωθεί ότι για τη PubMed εκμεταλλευόμαστε επιπλέον τις συσχετίσεις miRNA-άρθρων που περιλαμβάνονται στη βάση δεδομένων NCBI gene αφού αυτά τα δεδομένα είναι επίσης προσβάσιμα μέσω της σελίδας αποτελεσμάτων της PubMed.

Τα αποτελέσματα παρουσιάζονται στον Πίνακα 4.5. Να σημειωθεί ότι, κάτω από το σενάριο που εξετάζουμε, τόσο η mirPub όσο και η PubMed θεωρείται ότι έχουν ακρίβεια 100%. Αυτό σημαίνει ότι όλα τα ανακτημένα άρθρα συνδέονται με την αναζήτηση του χρήστη. Η ακριβής τιμή του ποσοστού ανάκτησης δεν μπορεί να ποσοτικοποιηθεί, αφού το συνολικό πλήθος των σχετικών δημοσιεύσεων είναι άγνωστο. Παρόλα αυτά, το γεγονός ότι η mirPub επιστρέφει περισσότερα αποτελέσματα από τη PubMed μας επιτρέπει να συμπεράνουμε ότι η mirPub έχει υψηλότερο ποσοστό ανάκλησης. Συγκεριμένα, με βάση το πλήθος των ανακτημένων άρθρων, η mirPub αναμένεται να έχει ποσοστό ανάκτησης που είναι περισσότερες από δύο φορές το ποσοστό ανάκτησης της PubMed.

Επίσης αξιολογήσαμε πειραματικά τον τρόπο με τον οποίο η γνώση για την εξέλιξη των δεδομένων miRNA βελτιώνει την ικανότητα ανάκτησης βιβλιογραφίας. Συγκε-

Πίνακας 4.5: Πειραματική αξιολόγηση της ικανότητας της mirPub στην ανάκτηση βιβλιογραφίας miRNA συγκριτικά με την PubMed.

	#άρθρων
mirPub	134
PubMed	54

κριμένα, ζητήσαμε από έναν ειδικό να χρησιμοποιήσει τη mirPub για να ψάξει για 25 miRNA. Για τις ανάγκες του πειράματος αυτού επιλέξαμε miRNA που έχουν σημαντική ιστορία. Πρώτα ζητήσαμε από τον ειδικό να παραμετροποιήσει τα φίλτρα της mirPub με τον τρόπο που πίστευε ότι θα του επιστρέψει τα πιο σχετικά αποτελέσματα σε σχέση με αυτό που έψαχνε. Μετά του ζητήσαμε να εξετάσει την ιστορία όλων των ώριμων miRNA και των miRNA φουρκετών στο πλαίσιο "Used keywords" και στη συνέχεια να παραμετροποιήσει και πάλι τα φίλτρα. Στο τέλος έπρεπε να δώσει κρίσεις για την σχετικότητα των ανακτημένων αποτελεσμάτων. Ο πίνακας 4.6 συγκεντρώνει τα αποτελέσματα του πειράματος. Είναι εμφανές ότι η γνώση για την ιστορία των miRNA μπορεί να φανεί χρήσιμη για την αναζήτηση βιβλιογραφίας. Συγκεκριμένα, πραγματοποιώντας αναζητήσεις χωρίς τη γνώση της ιστορίας των miRNA αποτυγχάνει να επιστρέψει το 30.23% των σχετικών δημοσιεύσεων που επιστρέφονται όταν αυτή η πληροφορία είναι διαθέσιμη. Επιπλέον, η γνώση για την εξέλιξη αυτών των δεδομένων βοηθά το χρήστη να αποφύγει λάθη κατά την αναζήτηση (πχ τη χρήση παλιών λέξεων-κλειδιών που αναφέρονται σε αλλαγμένες ακολουθίες) παρουσιάζοντας υψηλά επίπεδα ακρίβειας (περίπου $90,2\%$ για το πείραμά μας).

Πίνακας 4.6: Αξιολογώντας πώς η γνώση της εξέλιξης των δεδομένων miRNA βελτιώνει την αναζήτηση δημοσιεύσεων.

	#άρθρα
ανακτημένα χωρίς γνώση εξελ.	90
ανακτημένα με γνώση εξελ.	143
σχετικά	129

4.7 Συμπεράσματα

Πραγματοποιήσαμε σημαντική δουλειά για την παροχή χρήσιμων εργαλείων για τη διευκόλυνση των επιστημόνων που εργάζονται στον κλάδο της έρευνας πάνω στα miRNA. Για το σκοπό αυτό, συλλέξαμε δεδομένα που ήταν σκορπισμένα σε πολλές επιστημονικές βάσεις δεδομένων και δημοσιεύσεις, τα συνδυάσαμε και τα επεξεργαστήκαμε για να εξάγουμε γνώση σχετικά με το ρόλο των μορίων miRNA σε πολλούς μηχανισμούς της ζωής. Τα αποτελέσματα διανέμονται στην ερευνητική κοινότητα μέσω μιας πληθώρας ισχυρών εργαλείων με διαισθητικές διεπαφές Ιστού. Χρησιμοποιώντας τα, οι βιοεπιστήμονες μπορούν τόσο να φυλλομετρήσουν τη συγκεντρωμένη γνώση του κλάδου που έχει καταγραφεί στις βάσεις δεδομένων μας όσο και να εκτελέσουν πολλούς τύπους ανάλυσης στα αποθηκευμένα δεδομένα.

Συγκεκριμένα, το DIANA microT παρέχει στους βιοεπιστήμονες προβλέψεις για τα γονίδια που στοχεύονται από όλα τα γνωστά miRNA, βάσει της πολύ ακριβούς μεθόδου DIANA microT . Το DIANA miRGen ενημερώνει τους χρήστες του για τις γονιδιακές τοποθεσίες όλων των μεταγράφων miRNA και τη συμπεριφορά έκφρασής

τους. Το DIANA TarBase παρέχει το πληρέστερο σύνολο πειραματικά επιβεβαιωμένων στόχων miRNA. Το DIANA mirPath ερευνά το ρόλο των miRNA στα γνωστά μεταβολικά μονοπάτια. Τέλος, το DIANA mirPub βοηθά τους βιοεπιστήμονες στην αναζήτηση της βιβλιογραφίας σχετικά με τα miRNA.

Ο αντίκτυπος αυτών των εργαλείων έχει αξιολογηθεί μέσω της χρήσης του κατά τα προηγούμενα χρόνια. Περίπου 500 διακριτοί ερευνητές τα χρησιμοποιούν καθημερινά, πάνω από 100 από αυτούς είναι εγγεγραμμένα μέλη που επωφελούνται από τα εξατομικευμένα χαρακτηριστικά που παρέχουν τα εργαλεία μας.

Κεφάλαιο 5

Στοίχιση ακολουθιών για αναγνώσματα DNA μεγάλου μήκους

Η στοίχιση αναγνωσμάτων DNA είναι μία βασική ανάλυση για πολλές εφαρμογές στις βιοεπιστήμες (βλ. επίσης Κεφάλαιο 2.2.3). Συνίσταται στη στοίχιση αναγνωσμάτων DNA σε γονιδιακές ακολουθίες αναφοράς. Τα αναγνώσματα είναι μικρές ακολουθίες που εξάγονται από βιολογικά δείγματα χρησιμοποιώντας εξειδικευμένο εξοπλισμό που ονομάζεται μηχανή ακολουθιοποίησης.

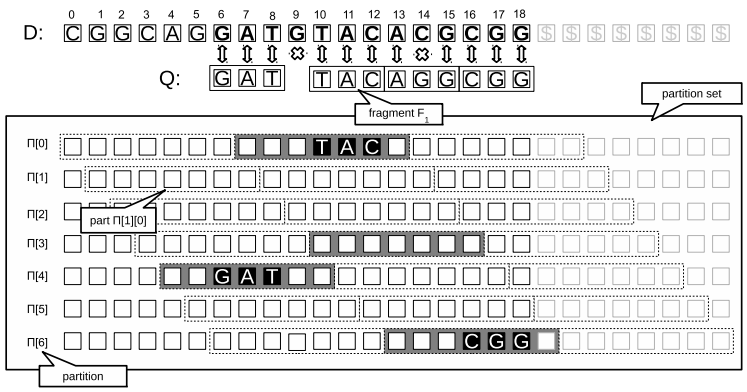
Παρόλο που πολλές δομές ευρετηρίου και αλγόριθμοι που επιταχύνουν τη στοίχιση αναγνωσμάτων DNA έχουν προταθεί στο παρελθόν, οι περισσότεροι από αυτούς είναι βελτιστοποιημένοι για αναγνώσματα που εξάγονται από παλιάς γενιάς μηχανές ακολουθιοποίησης και δεν συμπεριφέρονται καλά για αναγνώσματα που παράγονται από πρόσφατες μηχανές. Συγκεκριμένα, οι μηχανές ακολουθιοποίησης νέας γενιάς παράγουν αναγνώσματα DNA μεγαλύτερου μήκους (αποτελούμενα από περισσότερα των 200 συμβόλων) και η ακρίβεια ανάγνωσής τους μειώνεται, καθιστώντας τις τρέχουσες προσεγγίσεις φιλτραρίσματος για στοίχιση ακολουθιών μη ικανοποιητικές.

Σε αυτό το κεφάλαιο, παρουσιάζουμε τον *Hitmap*, μια προσέγγιση ευρετηρίασης που υποστηρίζει αποδοτική στοίχιση για αναγνώσματα μεγάλου μήκους και για μεγάλα κατώφλια συντακτικής απόστασης. Στο Κεφάλαιο 5.1 εξηγούμε κάποιες εισαγωγικές έννοιες, αναφέρουμε τα κίνητρά μας και παρουσιάζουμε τους αλγορίθμους αιχμής για στοίχιση αναγνωσμάτων DNA. Στο Κεφάλαιο 5.2 εισάγουμε τη δομή ευρετηρίου *Hitmap* και στο Κεφάλαιο 5.3 περιγράφουμε έναν αλγόριθμο που εκμεταλλεύεται αυτή τη δομή για να βρει με αποδοτικό τρόπο όλα τα σημεία στοίχισης μιας δεδομένης ακολουθίας-ερωτήματος μέσα σε μία ακολουθία δεδομένων. Στο Κεφάλαιο 5.4 παρουσιάζουμε πειράματα που αξιολογούν την αποδοτικότητα της προσέγγισής μας σε σύγκριση με τις προσεγγίσεις αιχμής. Τέλος, στο Κεφάλαιο 5.5 συγκεντρώνουμε τη συνεισφορά μας.

5.1 Υπόβαθρο

5.1.1 Χρήσιμοι συμβολισμοί και ορισμοί

Μόνο κατά τη διάρκεια του παρόντος κεφαλαίου και για λόγους παρουσίασης, δοθείσης μιας ακολουθίας S, συμβολίζουμε ως $S^{x:y}$ την υπακολουθία του S που ξεκινά από



Σχήμα 5.1: Ένα δείγμα ακολουθίας δεδομένων και ερωτήματος μαζί με τις αντίστοιχες διατμήσεις.

το x-στο σύμβολο και τελειώνει στο $(y-1)$-στο σύμβολο (προσοχή: $0 \leq x, y \leq |S|$). Επιπλέον, αναφερόμαστε στο x-στο της σύμβολο ως S^x. Τέλος, ακολουθούμε τη σύμβαση ότι ένα κεφαλαίο γράμμα, π.χ. το X, συμβολίζει ένα αντικείμενο (π.χ. μια ακολουθία), ενώ ένα μικρό γράμμα, π.χ. το x, αντιστοιχεί σε μια ποσότητα που σχετίζεται με το X (συνήθως το μήκος του).

Στο υπόλοιπο κείμενο, αναφερόμαστε σε μια ακολουθία δεδομένων (π.χ. μια γονιδιωματική ακολουθία αναφοράς) ως D και σε μία ακολουθία-ερώτημα η οποία πρέπει να στοιχιστεί μέσα της (π.χ. ένα ανάγνωσμα DNA) ως Q. Το Σχήμα 5.1 παρουσιάζει ένα απλουστευμένο παράδειγμα ακολουθιών δεδομένων και ερωτήματος με μήκη $d = 19$ και $q = 12$, αντίστοιχα. Να σημειωθεί ότι σε πραγματικά σενάρια το d ανέρχεται σε πλήθος συμβόλων της τάξης των δισεκατομμυρίων και το q σε πλήθος συμβόλων της τάξης των εκατοντάδων. Ακολουθούν μερικοί χρήσιμοι ορισμοί.

Στοίχιση (Alignment). Μια ακολουθία-ερώτημα Q *στοιχίζεται* στη θέση x του D αν το Q και το $D^{x:x+q+\psi}$ έχουν συντακτική απόσταση μικρότερη ή ίση από ϵ, για κάποιο $\psi \in [-\epsilon, \epsilon]$. Αναφερόμαστε στο ϵ ως το *κατώφλι στοίχισης*, και στην υπακολουθία $D^{x:x+q+\psi}$ ως *τόπος στοίχισης*.

Λόγου χάρη, στο παράδειγμα του Σχήματος 5.1 θεωρώντας ότι $\epsilon = 2$, υπάρχει μία στοίχιση του Q στο D στη θέση $x = 6$. Η στοίχιση εμπλέκει την αντικατάσταση του Q^7 από το σύμβολο `C` και την εισαγωγή του συμβόλου `G` στη θέση 3 του Q.

Θραύσμα (Fragment). Δοθέντος ενός ερωτήματος Q και ενός ακεραίου f, όπου $0 < f \leq q$, το Q διαχωρίζεται σε $\phi = \lfloor q/f \rfloor$ μη-επικαλυπτόμενες υπακολουθίες μήκους f, που ονομάζονται *θραύσματα*. Χρησιμοποιούμε το F_k για να αναπαραστήσουμε το θραύσμα $Q^{k \cdot f:(k+1) \cdot f}$, όπου $k \in [0, \phi)$. Να σημειωθεί ότι μερικά από τα σύμβολα στο τέλος του Q ενδέχεται να μην ανήκουν σε κανένα θραύσμα.

Για παράδειγμα, στο Σχήμα 5.1 θεωρώντας $f = 3$, το Q διαχωρίζεται σε $\phi = 4$ μη-επικαλυπτόμενα θραύσματα: $F_0 = \texttt{GAT}$, $F_1 = \texttt{TAC}$, $F_2 = \texttt{AGG}$ και $F_3 = \texttt{CGG}$.

f-gram. Ένα f-gram είναι απλώς μία ακολουθία f συμβόλων. Ως φυσικό επακόλουθο, ένα θραύσμα του ερωτήματος είναι ένα f-gram.

Εμφάνιση (Appearance). Ένα θραύσμα F_k (μιας ακολουθίας-ερωτήματος Q), όπου $k \in [0, \phi)$, *εμφανίζεται* μέσα σε μια υπακολουθία $D^{y:z}$ των δεδομένων, αν υπάρχει $x \in [y, z - f]$ τέτοιο ώστε τα F_k και $D^{x:x+f}$ να ταυτίζονται. Αναφερόμαστε στην υπακολουθία $D^{x:x+f}$ ως μια *εμφάνιση* του F_k στο x.

Θεωρούμε την υπακολουθία $D^{7:14}$ της ακολουθίας δεδομένων στο Σχήμα 5.1. Υπάρχει μια εμφάνιση του F_1 μέσα σε αυτή την υπακολουθία, αφού $F_1 = D^{10:13} = \texttt{TAC}$.

5.1.2 Το θεώρημα του περιστερώνα

Το θεώρημα του περιστερώνα είναι μια απλή παρατήρηση στην οποία βασίζονται πολλές τεχνικές φιλτραρίσματος ακολουθιών. Διατυπώνει ότι εάν περιστέρια τοποθετηθούν σε θέσεις ενός περιστερώνα και υπάρχουν περισσότερα περιστέρια από θέσεις, τότε υπάρχει τουλάχιστον μία θέση που περιέχει περισσότερα από ένα περιστέρια. Η εφαρμογή αυτής της αρχής στο πρόβλημα της στοίχισης ακολουθιών έχει ως αποτέλεσμα το επόμενο θεώρημα.

Θεώρημα 5.1. *Θεωρούμε ένα ερώτημα Q και μια ακολουθία δεδομένων D. Έστω ϵ είναι το κατώφλι στοίχισης, f είναι το μέγεθος των θραυσμάτων του ερωτήματος και $\phi = \lfloor q/f \rfloor$ το πλήθος τους.*
Εάν το Q στοιχίζεται στη θέση x του D τότε τουλάχιστον $\phi - \epsilon$ θραύσματα του Q εμφανίζονται μέσα στον αντίστοιχο τόπο στοίχισης.

Το επόμενο πόρισμα είναι απαραίτητη απόρροια του Θεωρήματος 5.1 και, κατά τη διάρκεια της στοίχισης μιας δεδομένης ακολουθίας Q μέσα σε μια ακολουθία δεδομένων D με κατώφλι ϵ, προσφέρει τα μέσα για να φιλτράρουμε περιοχές της D που δεν περιέχουν καμία στοίχιση του Q.

Πόρισμα 5.1. *Εάν λιγότερα από $\phi - \epsilon$ θραύσματα της Q εμφανίζονται μέσα σε μια υπακολουθία $D^{y:z}$ των δεδομένων, τότε η Q δεν μπορεί να στοιχιστεί σε οποιοδήποτε $x \in [y, z - q]$.*

5.1.3 Το κίνητρό μας

Στο Κεφάλαιο 2.2.3 συζητήσαμε λεπτομερώς την παραγωγή των αναγνωσμάτων DNA από μηχανές ακολουθιοποίησης και την ανάγκη για στοίχιση αυτών των αναγνωσμάτων σε γονιδιώματα αναφοράς. Χρησιμοποιώντας τους συμβολισμούς που εισήχθησαν στο Κεφάλαιο 5.1.1 μπορούμε να ορίσουμε το πρόβλημα ως ακολούθως.

Ορισμός 5.1 (Στοίχισης αναγνωσμάτων DNA)**.** *Θεωρούμε ένα γονιδίωμα αναφοράς D, ένα ανάγνωσμα DNA Q και ένα κατώφλι στοίχισης ϵ. Το ζητούμενο της στοίχισης αναγνωσμάτων DNA είναι να βρεθούν οι θέσεις όλων των στοιχίσεων του Q μέσα στην D.*

Από εδώ και στο εξής, θα χρησιμοποιούμε τους όρους 'γονιδίωμα αναφοράς' και 'δεδομένα' εναλλάξ. Το ίδιο ισχύει και για τους όρους 'ανάγνωσμα DNA' και 'ερώτημα'.

Στον πυρήνα το πρόβλημα της στοίχισης αναγνωσμάτων είναι παρόμοιο με το κοινό κατά προσέγγιση ταίριασμα ακολουθιών (βλ. Κεφάλαιο 2.1.1.2). Συνεπώς οι γνωστές τεχνικές ταιριάσματος ακολουθιών ήταν οι πρώτες που προσαρμόστηκαν για να χρησιμοποιηθούν σε αυτό το πρόβλημα. Στο παρελθόν, πολλές προσεγγίσεις βασισμένες σε πίνακες κατακερματισμού από grams ή συμπιεσμένους πίνακες επιθεμάτων προτάθηκαν για να λύσουν το πρόβλημα (βλ. Κεφάλαιο 2.2.3). Όλες αυτές οι μέθοδοι σχεδιάστηκαν για να συμπεριφέρονται καλά κατά την στοίχιση μικρών αναγνωσμάτων DNA, αφού η πρώτη γενιά μηχανών ακολουθιοποίησης ήταν ικανή να παράγει αναγνώσματα μήκους $30 - 50$ συμβόλων. Όμως, οι πιο πρόσφατες τεχνολογικές εξελίξεις είχαν ως αποτέλεσμα την εμφάνιση μηχανών που παράγουν σημαντικά μεγαλύτερα αναγνώσματα DNA θυσιάζοντας μέρος της ακρίβειας ανάγνωσης (βλ. Κεφάλαιο 2.2.3).

Κάτω από τις προαναφερθείσες συνθήκες, οι περισσότερες προσεγγίσεις αιχμής αποτυγχάνουν να επιτύχουν ικανοποιητική αποδοτικότητα. Το WHAM ήταν μία προσέγγιση που προτάθηκε για να καλύψει αυτό το κενό [54]. Πειράματα που συγκρίνουν

το WHAM με άλλες προσεγγίσεις αιχμής αποκάλυψαν ότι το WHAM υπερνικά τους ανταγωνιστές του τόσο για αναγνώσματα μικρού ($30 - 50$ σύμβολα) όσο και κάπως μεγάλου μήκους (100 σύμβολα). Επιπλέον, το WHAM βρέθηκε να συμπεριφέρεται καλά ακόμα και για σενάρια στοίχισης αναγνωσμάτων χρησιμοποιώντας σχετικά χαλαρά κατώφλια στοίχισης.

Από την άλλη, το WHAM υποφέρει από δύο σημαντικά μειονεκτήματα. Το πρώτο είναι ότι η αυθεντική του έκδοση επιτρέπει αναζήτηση για στοιχίσεις που ικανοποιούν είτε (α) ένα δοθέν κατώφλι απόστασης Χάμινγκ είτε (β) τρία δοθέντα κατώφλια, καθένα από τα οποία καθορίζει το μέγιστο πλήθος εισαγωγών, διαγραφών ή αστοχιών, αντίστοιχα. Όμως, η χρήση ενός κατωφλίου συντακτικής απόστασης είναι η δημοφιλέστερη, διαισθητικότερη και πιο ισχυρή επιλογή. Το δεύτερο μειονέκτημα του WHAM είναι ότι παρόλο που συμπεριφέρεται καλύτερα από τις υπόλοιπες προσεγγίσης στοίχισης για μικρά αναγνώσματα DNA, η αποδοτικότητά του κατά τη διάρκεια της στοίχισης αναγνωσμάτων μήκους μεγαλύτερου από 100 σύμβολα δεν είναι ικανοποιητική. Αυτό είναι ένα σημαντικό ζήτημα αφού κάποιες σύγχρονες μηχανές ακολουθιοποίησης παράγουν αναγνώσματα DNA που αριθμούν > 400 σύμβολα (κάποιες άλλες παρέχουν αναγνώσματα χιλιάδων συμβόλων) και αυτό είναι μία τάση που αναμένεται να συνεχιστεί (βλ. Κεφάλαιο 2.2.3).

Η πρόθεσή μας ήταν να προσφέρουμε μια νέα προσέγγιση για την αποδοτική στοίχιση αναγνωσμάτων που παράγονται από σύγχρονες μηχανές ακολουθιοποίησης. Συνεπώς, η προσέγγισή μας πρέπει να συμπεριφέρεται καλά για αναγνώσματα που περιέχουν > 100 σύμβολα και πρέπει να διατηρεί τα πλεονεκτήματά της ακόμα και σε περιπτώσεις που η ακρίβεια ανάγνωσης της μηχανής ακολουθιοποίησης μειώνεται δραματικά. Επιπλέον, για τη διευκόλυνση των βιοεπιστημόνων, η μέθοδός μας πρέπει να υποστηρίζει κατώφλια συντακτικής απόστασης.

Για να αξιολογήσουμε την αποδοτικότητα αυτής της μεθόδου προσαρμόσαμε το WHAM ώστε να υποστηρίζει τη συντακτική απόσταση προκειμένου αυτό να διαμορφωθεί στο βασικό ανταγωνιστή της μεθόδου μας. Ως έναν επιπλέον ανταγωνιστή χρησιμοποιήσσαμε το RBSA [75] μια μέθοδο αιχμής που είναι γνωστό ότι υποστηρίζει συντακτική απόσταση και που είναι ικανή για στοίχιση αναγνωσμάτων μεγάλου μήκους. Στις επόμενες παραγράφους περιγράφουμε λεπτομερώς την προσέγγιση που προτείνουμε και παρουσιάζουμε πειράματα που αξιολογούν τις επιδόσεις της έναντι των προαναφερθέντων ανταγωνιστών.

5.2 Το ευρετήριο Hitmap

Στο Κεφάλαιο 5.2.1, πρώτα παρουσιάζουμε κάποιες παρατηρήσεις και βασικές έννοιες πάνω στις οποίες βασίζεται το ευρετήριο *Hitmap*, όπως είναι τα τμήματα δεδομένων, τα καλύμματα θέσης και οι κλάσεις καλυμμάτων. Έπειτα, στο Κεφάλαιο 5.2.2, περιγράφουμε λεπτομερώς τη δομή ευρετηρίου *Hitmap* και τον τρόπο κατασκευής της πάνω σε μια δεδομένη ακολουθία δεδομένων (ακολουθία αναφοράς). Τέλος, στο Κεφάλαιο 5.2.3, παρουσιάζουμε μια μέθοδο συμπίεσης που χρησιμοποιείται για να μειωθεί το αποτύπωμα μνήμης του *Hitmap*.

5.2.1 Βασικές έννοιες

Θεωρούμε μια ακολουθία δεδομένων D, ένα ερώτημα Q, το κατώφλι στοίχισης ϵ και το επιλεγμένο μέγεθος θραυσμάτων f (με ϕ να είναι το πλήθος αυτών των θραυσμάτων).

Τμήματα (Parts). Η πιο σημαντική έννοια σχετικά με το ευρετήριο *Hitmap* είναι αυτή του *τμήματος*, το οποίο απλώς αντιστοιχεί σε μια υπακολουθία της D μήκους $\pi = f + 2 \cdot \epsilon$. Ένα τμήμα είναι μεγαλύτερο από ένα θραύσμα κατά ακριβώς δύο φορές το κατώφλι στοίχισης ϵ. Ο λόγος είναι για να μπορεί να υποδεχθεί εισαγωγές και διαγραφές με τέτοιο τρόπο που να εξασφαλίζει ότι θα υπάρχει πάντα τουλάχιστον ένα τμήμα των δεδομένων που να περικλύει οποιαδήποτε εμφάνιση θραύσματος του ερωτήματος, όπως θα γίνει πιο σαφές παρακάτω.

Το ευρετήριο *Hitmap* αποθηκεύει πληροφορία σχετικά με όλα τα πιθανά τμήματα της D (οι επικαλύψεις επιτρέπονται). Για ευκολία στην αναφορά στα τμήματα, τα οργανώνουμε σε μια δισδιάστατη δομή και χρησιμοποιούμε δύο συντεταγμένες για να αναφερθούμε σε οποιοδήποτε από αυτά με μοναδικό τρόπο. Συγκεκριμένα, ένα τμήμα συμβολίζεται ως $\Pi[i][j]$ και αντιστοιχεί στην υπακολουθία των δεδομένων $D^{i+j\cdot\pi:i+(j+1)\cdot\pi}$ μήκους π, όπου η συντεταγμένη i παίρνει τιμές μέσα στο $[0, \pi)$ και η συντεταγμένη j μέσα στο $[0, \lceil d/\pi \rceil)$. Να σημειωθεί ότι προκειμένου να εγγυηθούμε ότι τα τμήματα στη συντεταγμένη $j = \lceil d/\pi \rceil - 1$ μπορούν να οριστούν, επεκτείνουμε τη D προσθέτοντας $(\pi - 1) + \lceil d/\pi \rceil \cdot \pi - d$ ειδικά σύμβολα (τα οποία δεν ανήκουν στο αλφάβητο) στο τέλος.

Αναφερόμαστε στο σύνολο των τμημάτων που έχουν κοινή τη συντεταγμένη i (δηλαδή μια σειρά από τμήματα) ως την i-στη *διάτμηση (partition)* συμβολιζόμενη ως $\Pi[i]$. Υπάρχουν π διατμήσεις, ακριβώς τόσες όσο είναι το μήκος του κάθε τμήματος. Ο λόγος είναι γιατί τα τμήματα σε μια διάτμηση $\Pi[i+1]$ μπορούν να προκύψουν από εκείνα της διάτμησης $\Pi[i]$ μέσω της ολίσθησης των θέσεών τους κατά ένα σύμβολο. Αφού ένα τμήμα έχει π σύμβολα, υπάρχουν π πιθανές ολισθήσεις. Να σημειωθεί ότι κάθε διάτμηση περιέχει μη-επικαλυπτόμενα τμήματα.

Θεωρώντας και πάλι το παράδειγμα του Σχήματος 5.1, για $f = 3$ και $\epsilon = 2$ (η τιμή του ϵ είναι εσκεμμένα πολύ υψηλή για λόγους επίδειξης), θέτουμε το μήκος τμήματος ίσο με $\pi = 3 + 2 \cdot 2 = 7$. Έπειτα, υπάρχουν $\pi = 7$ διατμήσεις και κάθε διάτμηση περιέχει $\lceil d/\pi \rceil = 3$ τμήματα μήκους 7. Κάθε τμήμα αναπαρίσταται από ένα διάστικτο ορθογώνιο. Τα επιπλέον σύμβολα που προστίθονται στο τέλος της D συμβολίζονται ως \$.

Καλύμματα (Covers). Το κίνητρο για τον ορισμό τμημάτων είναι για τη διευκόλυνση του να καθοριστεί αν το ερώτημα μπορεί να στοιχιστεί σε μια συγκεκριμένη θέση των δεδομένων. Για να το επιτύχουμε αυτό, χρησιμοποιούμε το θεώρημα του περιστερώνα και ελέγχουμε αν ικανοποιητικό πλήθος θραυσμάτων του ερωτήματος εμφανίζονται μέσα σε προσεκτικά επιλεγμένα τμήματα. Αυτά τα τμήματα ανήκουν σε ένα σύνολο, που ονομάζεται κάλυμμα και το οποίο ορίζεται με βάση μια θέση των δεδομένων ως ακολούθως.

Δοθείσης μίας θέσης x στην D, ορίζουμε το *κάλυμμα* της x, συμβολιζόμενο από το $C^x = \{C^x_0, \ldots, C^x_{\phi-1}\}$, ως ένα διατεταγμένο σύνολο από ϕ τμήματα τέτοια ώστε $C^x_k = \Pi[i^x_k][j^x_k]$, όπου $i^x_0 = mod(x-\epsilon, \pi)$, $j^x_0 = \lfloor \frac{x-\epsilon}{\pi} \rfloor$, $i^x_k = mod(x-\epsilon+k\cdot f, \pi)$, και $j^x_k = \lfloor \frac{x-\epsilon+k\cdot f}{\pi} \rfloor$, για $k \in [0, \phi)$. Να σημειωθεί ότι τα τμήματα σε ένα κάλυμμα μπορούν να ανήκουν σε διαφορετικές διατμήσεις. Για παράδειγμα, στο Σχήμα 5.1, τα γκρι τμήματα διαμορφώνουν το κάλυμμα της θέσης $x = 6$, δηλ. $C^6 = \{\Pi[4][0], \Pi[0][1], \Pi[3][1], \Pi[6][1]\}$.

Κάθε τμήμα σε ένα κάλυμμα επιλέγεται ώστε να αιχμαλωτίζει την πιθανή εμφά-

νιση ενός συγκεκριμένου θραύσματος του ερωτήματος. Πιο συγκεκριμένα, το k-στο θραύσμα F_k συνδέεται με το k-στο τμήμα C_k^x στο κάλυμμα. Δοθείσης αυτής της αντιστοίχισης, το ακόλουθο αποτέλεσμα ισχύει λόγω του θεωρήματος του περιστερώνα.

Θεώρημα 5.2. *Αν το πλήθος των θραυσμάτων F_k, για $k \in [0, \phi)$, που εμφανίζονται στο C_k^x είναι μικρότερο από $\phi - \epsilon$, τότε δεν υπάρχει στοίχιση στη θέση x.*

Απόδειξη. Θα αποδείξουμε το θεώρημα μέσω απαγωγής σε άτοπο. Έστω ότι υπάρχει μία στοίχιση στη x. Τότε, με βάση το Θεώρημα 5.1, ισχύει ότι τουλάχιστον $\phi - \epsilon$ θραύσματα του Q εμφανίζονται μέσα στο $D^{x:x+\epsilon+\psi}$, όπου $\psi \in [-\epsilon, \epsilon]$.

Έστω F_k, το k-στο θραύσμα του Q, είναι ένα από τα εμφανιζόμενα θραύσματα, όπου $k \in [0, \phi)$. Εφόσον το Q στοιχίζεται στη x με το πολύ ϵ λάθη, η αρχή μίας εμφάνισης του F_k θα πρέπει να βρίσκεται μεταξύ των $x + k \cdot f - \epsilon$ (στην περίπτωση που ϵ διαγραφές εφαρμόζονται στο Q πριν από το F_k) και $x + k \cdot f + \epsilon$ (αν ϵ εισαγωγές εφαρμόζονται στο Q πριν από το F_k). Βασισμένοι στα προηγούμενα, κάθε εμφάνιση του F_k βρίσκεται στο $D^{x+k\cdot f-\epsilon:x+(k+1)\cdot f+\epsilon}$.

Θεωρούμε τώρα το C_k^x, το k-στο τμήμα του καλύμματος του x. Εξ ορισμού, $C_k^x = \Pi[i_k^x][j_k^x]$, όπου $i_k^x = mod(x - \epsilon + k \cdot f, \pi)$ και $j_k^x = \left\lfloor \frac{x-\epsilon+k\cdot f}{\pi} \right\rfloor$. Από τον ορισμό των τμημάτων, το προηγούμενο τμήμα αντιστοιχεί στην υπακολουθία των δεδομένων $D^{i_k^x+j_k^x\cdot\pi:i_k^x+(j_k^x+1)\cdot\pi}$. Ισχύει ότι $i_k^x + j_k^x \cdot \pi = mod(x - \epsilon + k \cdot f, \pi) + \left\lfloor \frac{x-\epsilon+k\cdot f}{\pi} \right\rfloor \cdot \pi = x - \epsilon + k \cdot f$, με βάση τον ορισμό του υπολοίπου (modulo): $mod(a, b) = a - \left\lfloor \frac{a}{b} \right\rfloor \cdot b$. Επιπλέον, $i_k^x + (j_k^x + 1) \cdot \pi = (i_k^x + j_k^x \cdot \pi) + \pi = x - \epsilon + k \cdot f + f + 2 \cdot \epsilon = x + \epsilon + (k+1) \cdot f$.

Συνεπώς, η υπακολουθία των δεδομένων που αντιστοιχεί στο C_k^x ταυτίζεται με την υπακολουθία μέσα στην οποία βρίσκεται κάθε εμφάνιση του F_k. Εφόσον επιλέξαμε το F_k με τυχαίο τρόπο από το σύνολο των θραυσμάτων του Q που εμφανίζονται μέσα στη στοίχιση στη x, αυτό σημαίνει ότι η προηγούμενη παρατήρηση ισχύει για καθένα από τα, τουλάχιστον, $\phi - \epsilon$ θραύσματα που πρέπει να εμφανίζονται μέσα στη στοίχιση στην x. Έτσι, υπάρχουν τουλάχιστον $\phi - \epsilon$ θραύσματα F_k που εμφανίζονται στο C_k^x, για $k \in [0, \phi)$ το οποίο οδηγεί σε άτοπο.

□

Δοθέντος ενός ερωτήματος Q, κατά τη διάρκεια της αναζήτησης για τις στοιχίσεις του, μπορούμε να χρησιμοποιήσουμε το Θεώρημα 5.2 προκειμένου να φιλτράρουμε περιοχές της D, για τις οποίες είναι εγγυημένο ότι δεν περιέχουν καμία στοιχίση. Αντίστοιχα, μπορούμε να βρούμε όλες τις περιοχές που έχουν την πιθανότητα να περιέχουν μία στοίχιση με βάση το επόμενο πόρισμα.

Πόρισμα 5.2. *Εάν το πλήθος των θραυσμάτων F_k, για $k \in [0, \phi)$, που εμφανίζονται στο C_k^x είναι τουλάχιστον $\phi - \epsilon$, τότε είναι δυνατό να υπάρχει μία στοίχιση στη x. Αναφερόμαστε στη x ως μία* υποψήφια στοίχιση (candidate alignment).

Λόγου χάρη, στο παράδειγμα του Σχήματος 5.1, υπάρχει μία στοίχιση που ξεκινά από τη θέση $x = 6$ και, πράγματι, υπάρχουν $3 > (\phi - \epsilon) = 2$ θραύσματα που εμφανίζονται μέσα στο C^6. Να σημειωθεί ότι παρόλο που στη συγκεκριμένη περίπτωση υπάρχει μία στοίχιση, σε άλλες περιπτώσεις θα μπορούσε να είναι απλώς ένα εσφαλμένα θετικό αποτέλεσμα.

Κλάσεις καλυμμάτων (Cover classes). Καλύμματα που ανήκουν σε διάφορες διακριτές θέσεις εμφανίζουν μια καλή συμμετρία. Έστω X είναι το σύνολο των d πιθανών θέσεων που υπάρχουν στην ακολουθία δεδομένων D. Δοθέντος ενός κατωφλίου στοίχισης ϵ και ενός μήκους τμήματος π, ορίζουμε π διακριτές κλάσεις ισοδυναμίας ως

ακολούθως: $[\alpha] = \{x \in X : mod(x - \epsilon, \pi) = \alpha\}$, $\forall \alpha \in [0, \pi)$. Να σημειωθεί ότι αν το x ανήκει στην κλάση ισοδυναμίας $[\alpha]$, τότε μπορεί να εκφραστεί ως $x = \alpha + \lambda \cdot \pi + \epsilon$, όπου $\lambda \in [0, \lceil d/\pi \rceil)$.

Ορίζουμε την *κλάση καλυμμάτων* $[\alpha]$ ως το σύνολο των καλυμμάτων για τις θέσεις $x \in [\alpha]$. Σε πολλές περιπτώσεις, για λόγους απλότητας, αντί να αναφέρουμε ότι το κάλυμμα μιας θέσης x ανήκει σε μια κλάση καλυμμάτων $[\alpha]$, αναφέρουμε ότι η x ανήκει στην $[\alpha]$.

Το επόμενο θεώρημα διατυπώνει διάφορες συσχετίσεις που ισχύουν για τις συντεταγμένες i_k^x, j_k^x ενός τμήματος στο κάλυμμα της x.

Θεώρημα 5.3. *Για κάθε $x_1, x_2 \in [\alpha]$, όπου $\alpha \in [0, \pi)$, και κάθε $k \in [0, \phi)$, οι ακόλουθες διατυπώσεις ισχύουν για τις συντεταγμένες των τμημάτων στα καλύμματα C^{x_1}, C^{x_2}:*

$$(1)\ i_k^{x_1} = i_k^{x_2} = i_k^{[\alpha]}$$

$$(2)\ j_k^{x_1} - j_0^{x_1} = j_k^{x_2} - j_0^{x_2} = \Delta j_k^{[\alpha]}$$

Επιπλέον, γράφοντας $x_1 = \alpha + \lambda^{x_1} \cdot \pi + \epsilon$ και $x_2 = \alpha + \lambda^{x_2} \cdot \pi + \epsilon$, όπου $x_1 \leq x_2$ και $\lambda^{x_1}, \lambda^{x_2} \in [0, \lceil d/\pi \rceil)$, τότε, για κάθε $k \in [0, \phi)$, ισχύει:

$$(3)\ j_k^{x_2} = j_k^{x_1} + \lambda^{x_2} - \lambda^{x_1}$$

Απόδειξη. Για τις ανάγκες της παρούσας απόδειξης θεωρούμε δύο θέσεις των δεδομένων $x_1 = \alpha + \lambda^{x_1} \cdot \pi + \epsilon$ και $x_2 = \alpha + \lambda^{x_2} \cdot \pi + \epsilon$, όπου $x_1 \leq x_2$ και $\lambda^{x_1}, \lambda^{x_2} \in [0, \lceil d/\pi \rceil)$.

Αρχικά αποδεικνύουμε τη διατύπωση (1). Ισχύει ότι $x_2 = \alpha + \lambda^{x_2} \cdot \pi + \epsilon = \alpha + (\lambda^{x_2} - \lambda^{x_1}) \cdot \pi + \lambda^{x_1} \cdot \pi + \epsilon = x_1 + (\lambda^{x_2} - \lambda^{x_1}) \cdot \pi$. Συνεπώς, $\forall k \in [0, \phi)$ ισχύει ότι $i_k^{x_2} = mod(x_2 - \epsilon + k \cdot f, \pi) = mod(x_1 + (\lambda^{x_2} - \lambda^{x_1}) \cdot \pi - \epsilon + k \cdot f, \pi)$ $= mod(x_1 - \epsilon + k \cdot f, \pi) = i_k^{x_1}$.

Έπειτα, αποδεικνύουμε τη διατύπωση (3). Για κάθε $k \in [0, \phi)$, ισχύει: $j_k^{x_2} = \left\lfloor \frac{x_2 - \epsilon + k \cdot f}{\pi} \right\rfloor = \left\lfloor \frac{x_1 + (\lambda^{x_2} - \lambda^{x_1}) \cdot \pi - \epsilon + k \cdot f}{\pi} \right\rfloor = \left\lfloor \frac{x_1 - \epsilon + k \cdot f}{\pi} \right\rfloor + \frac{(\lambda^{x_2} - \lambda x_1) \cdot \pi}{\pi} = j_k^{x_1} + \lambda^{x_2} - \lambda^{x_1}$.

Τέλος, αποδεικνύουμε τη διατύπωση (2). Για κάθε $k \in [0, \phi)$, ισχύει: $j_k^{x_2} - j_0^{x_2} = j_k^{x_1} + \lambda^{x_2} - \lambda^{x_1} - j_0^{x_1} - \lambda^{x_2} + \lambda^{x_1} = j_k^{x_1} - j_0^{x_1}$.

□

Ένα άμεσο πόρισμα του Θεωρήματος 5.3 είναι ότι δύο καλύμματα C^{x_1}, C^{x_2} μέσα σε μία κλάση περιέχουν τμήματα που κατανέμονται με παρόμοιο τρόπο στο σύνολο των διατμήσεων. Συγκεκριμένα, (1) το k-στο τμήμα στο C^{x_1} και στο C^{x_2} ανήκει στην ίδια διάτμηση, και (2) η διαφορά στη συντεταγμένη j δύο συνεχόμενων τμημάτων στα καλύμματα C^{x_1} και C^{x_2} είναι ίδια. Συνεπώς, μπορούμε να φανταστούμε το κάλυμμα της x_2 ως το κάλυμμα της x_1 ολισθημένο στα δεξιά κατά $\lambda^{x_2} - \lambda^{x_1}$ τμήματα, θεωρώντας $\lambda^{x_1} \leq \lambda^{x_2}$.

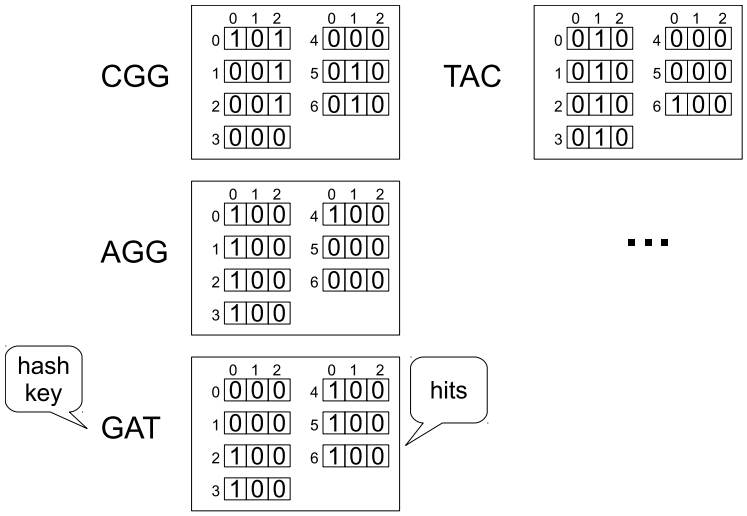
Ως παράδειγμα, θεωρούμε τις θέσεις $x_1 = 6$ και $x_2 = 13$ της ακολουθίας δεδομένων στο Σχήμα 5.1. Και οι δύο ανήκουν στην κλάση ισοδυναμίας [4], αφού $x_1 = 4 + 0 \cdot 7 + 2 = 6$ και $x_2 = 4 + 1 \cdot 7 + 2 = 13$. Ισχύει ότι $C^{x_1} = \{\Pi[4][0], \Pi[0][1], \Pi[3][1], \Pi[6][1]\}$ και $C^{x_2} = \{\Pi[4][1], \Pi[0][2], \Pi[3][2], \Pi[6][2]\}$, δηλ. το C^{x_2} είναι απλώς το C^{x_1} ολισθημένο στα δεξιά κατά ένα τμήμα.

5.2.2 Δομή ευρετηρίου

Περιγραφή της δομής. Υπάρχει ένα πεπερασμένο σύνολο $|\Sigma|^f$ διαφορετικών f-grams που μπορούν να είναι θραύσματα μιας ακολουθίας-ερωτήματος. Η δομή ευρετηρίου *Hitmap* $\mathcal{H}$ είναι ένας πίνακας κατακερματισμού που κωδικοποιεί τις εμφανίσεις τους μέσα στα τμήματα της ακολουθίας δεδομένων D.

Συγκεκριμένα, το $\mathcal{H}$ περιέχει μία καταχώρηση για κάθε f-gram το οποίο εμφανίζεται τουλάχιστον μία φορά μέσα στο D. Το κλειδί κ κάθε καταχώρησης είναι το ίδιο το f-gram και η τιμή $\mathcal{H}(\kappa)$, ή $\mathcal{H}_\kappa$ για απλότητα, αντιστοιχεί σε ένα πίνακα από δυαδικά σύνολα (bitsets) $\{\mathcal{H}_\kappa[0], \cdots, \mathcal{H}_\kappa[\pi-1]\}$ που κωδικοποιούν τις εμφανίσεις του κ μέσα στα τμήματα της D. Πιο συγκεκριμένα, το bitset $\mathcal{H}_\kappa[i]$ αντιστοιχεί στην $\Pi[i]$, δηλ. στην i-στη κατάτμηση της D και περιέχει ένα δυαδικό σύμβολο για καθένα από τα τμήματά της. Η τιμή του $\mathcal{H}_\kappa[i][j]$ είναι 1 αν το κ εμφανίζεται μέσα στην $\Pi[i][j]$, ή 0 αν όχι. Χρησιμοποιούμε τον όρο *χτύπημα (hit)* για να αναφερθούμε σε κάθε δυαδικό σύμβολο του οποίου η τιμή είναι ίση με 1.

Το Σχήμα 5.2 απεικονίζει κάποιες καταχωρήσεις του ευρετηρίου *Hitmap* που δημιουργείται στην ακολουθία δεδομένων του Σχήματος 5.1. Εφόσον η ακολουθία `TAC` περιέχεται στα τμήματα $\Pi[0][1], \Pi[1][1]$, $\Pi[2][1], \Pi[3][1]$ και $\Pi[6][0]$, ισχύει ότι $\mathcal{H}_{\texttt{TAC}}[0][1] = \mathcal{H}_{\texttt{TAC}}[1][1] = \mathcal{H}_{\texttt{TAC}}[2][1] = \mathcal{H}_{\texttt{TAC}}[3][1] = \mathcal{H}_{\texttt{TAC}}[6][0] = 1$. Όλα τα υπόλοιπα δυαδικά σύμβολα του $\mathcal{H}_{\texttt{TAC}}$ τίθενται στο 0.



Σχήμα 5.2: Τμήμα του ευρετηρίου Hitmap της ακολουθίας δεδομένων που παρουσιάζεται στο Σχήμα 5.1.

Κατασκευή ευρετηρίου. Δοθείσης μιας ακολουθίας δεδομένων D, το ευρετήριο *Hitmap* της $\mathcal{H}$ κατασκευάζεται ολισθαίνοντας ένα παράθυρο μεγέθους f από την πρώτη προς την τελευταία θέση της D, ενώ ταυτόχρονα καταγράφονται οι εμφανίσεις των f-grams μέσα στα τμήματα της D.

Έστω $W^x = D^{x:x+f}$ είναι η υπακολουθία της D που περικλείεται από το παράθυρο όταν η αρχή του παραθύρου είναι τοποθετημένη στη θέση x της D. Ένα f-gram που έχει την ίδια ακολουθία με το W^x θα εμφανίζεται σε όλα τα τμήματα που περιέχουν εξ ολοκλήρου το W^x, άρα τα αντίστοιχα δυαδικά σύμβολα στο $\mathcal{H}_{W^x}$ πρέπει να τεθούν στο 1. Να σημειωθεί ότι, για κάθε θέση x, υπάρχουν $\pi - f + 1 = 2 \cdot \epsilon + 1$ τμήματα που περιέχουν εξ ολοκλήρου το W^x.

Για παράδειγμα, θεωρούμε ότι το *Hitmap* $\mathcal{H}$ της ακολουθίας δεδομένων που πα-

ρουσιάζεται στο Σχήμα 5.1 είναι υπό κατασκευή και ότι το παράθυρο που ολισθαίνει τοποθετείται στη θέση $x = 10$. Τότε, $W^{10} = TAC$ και τα τμήματα Π[0][1], Π[1][1], Π[2][1], Π[3][1], και Π[6][0] περιέχουν εξ ολοκλήρου το W^{10}. Συνεπώς, τα δυαδικά σύμβολα $\mathcal{H}_{TAC}$[0][1], $\mathcal{H}_{TAC}$[1][1], $\mathcal{H}_{TAC}$[2][1], $\mathcal{H}_{TAC}$[3][1], και $\mathcal{H}_{TAC}$[6][0] είναι hits του TAC και πρέπει να τεθούν στην τιμή 1.

5.2.3 Συμπίεση του ευρετηρίου

Κάθε καταχώρηση του *Hitmap* περιέχει π bitset (ένα για κάθε διάτμηση) καθένα από τα οποία αποτελείται από $\lceil d/\pi \rceil$ δυαδικά σύμβολα (ένα για κάθε τμήμα). Ως αποτέλεσμα, μία καταχώρηση περιέχει συνολικά d δυαδικά σύμβολα, το οποίο είναι ένα πολύ μεγάλο πλήθος για μεγάλες ακολουθίες δεδομένων. Ευτυχώς, τα bitset στο *Hitmap* είναι πάρα πολύ αραιά για τα περισσότερα πραγματικά σενάρια και, συνεπώς, το ευρετήριο μπορεί να συμπιεστεί σε μεγάλο βαθμό επιτυγχάνοντας πολύ μεγάλους λόγους συμπίεσης χωρίς να θυσιάζεται η αποδοτικότητα των ερωτημάτων.

Η κεντρική ιδέα πίσω από το σχήμα συμπίεσης είναι να αποφεύγεται η αποθήκευση μεγάλων συνόλων από συνεχόμενα μηδενικά δυαδικά ψηφία. Επιπλέον, υπάρχει μία ακόμα παρατήρηση που μπορεί επίσης να βοηθήσει: Μέσα στην ίδια καταχώρηση του *Hitmap* , τα μη-μηδενικά δυαδικά ψηφία (hits) τείνουν να εμφανίζονται στην ίδια θέση στα διάφορα bitset που τα περιέχουν. Θεωρούμε για παράδειγμα τον πίνακα από bitsets μιας καταχώρησης $\mathcal{H}_{AGG}$ στο Σχήμα 5.2. Όλα τα hits της εμφανίζονται στο 0-στο δυαδικό ψηφίο του κάθε bitset. Το προαναφερθέν πρότυπο είναι πολύ κοινό γιατί οι ίδιες θέσεις διαφορετικών bitset στην ίδια καταχώρηση *Hitmap* αντιστοιχούν σε τμήματα της ακολουθίας δεδομένων που επικαλύπτονται σε πολύ μεγάλο βαθμό.

Ένα παρελκόμενο των προηγούμενων παρατηρήσεων είναι ότι, αν αναπαραστήσουμε τον πίνακα των bitset κάθε καταχώρησης ως ένα δυσδιάστατο μητρώο, όπου κάθε bitset αντιστοιχεί σε μία γραμμή, τότε οι περισσότερες από τις στήλες αυτού του μητρώου αναμένεται να είναι γεμάτες μηδενικά. Εξαιτίας αυτού, επιλέγουμε να αποθηκεύουμε τα δυαδικά ψηφία των καταχωρήσεων του *Hitmap* ανά στήλη και όχι ανά γραμμή και αποφεύγουμε να αποθηκεύουμε στήλες που είναι γεμάτες μηδενικά. Επιπλέον, εφόσον, στα περισσότερα πραγματικά σενάρια, οι στήλες είναι αραιές από μόνες τους, κατορθώσαμε να μειώσουμε ακόμα περισσότερο τον απαιτούμενο χώρο χτίζοντας μια μεγάλη δεξαμενή με όλα τα διαφορετικά bitset που υπάρχουν και αποθηκεύοντας στις καταχωρήσεις *Hitmap* μόνο δείκτες προς τις εγγραφές της δεξαμενής.

5.3 Αποτίμηση ερωτήματος με χρήση του ευρετηρίου Hitmap

Σε αυτό το κεφάλαιο, περιγράφουμε έναν αλγόριθμο που εκμεταλλεύεται το ευρετήριο *Hitmap* για να βρει όλες τις στοιχίσεις ενός δοθέντος ερωτήματος Q μέσα σε μία δοθείσα ακολουθία δεδομένων D, όπου το κατώφλι στοίχισης είναι ϵ.

5.3.1 Ο αλγόριθμος Hitmap

Θεωρούμε μια ακολουθία-ερώτημα Q και ένα κατώφλι στοίχισης ϵ. Είναι εφικτό να φιλτράρουμε περιοχές στη D οι οποίες δεν περιέχουν καμία στοίχιση του Q εκμεταλλευόμενοι το $\mathcal{H}$, το ευρετήριο *Hitmap* της D. Αυτή η διαδικασία φιλτραρίσματος

αποτελείται από (α) εξέταση του $\mathcal{H}$ για να ανακτήσουμε τους πίνακες των bitsets $\mathcal{H}_{F_0}, \ldots, \mathcal{H}_{F_{\phi-1}}$, όπου $F_0, \ldots, F_{\phi-1}$ είναι θραύσματα του Q και (β) εκτέλεση bitwise λειτουργιών πάνω στα ανακτημένα bitset για να εξετάσουμε τη συνθήκη του Πορίσματος 5.2 για όλες τις θέσεις της D. Αν η συνθήκη ισχύει για μία θέση x, τότε η περιοχή τριγύρω από αυτή τη θέση πρέπει να εξεταστεί για πιθανές στοιχίσεις. Διαφορετικά, η x φιλτράρεται.

Θεωρούμε τη θέση x της D και το k-στο της τμήμα καλύμματος $C_k^x = \Pi[i_k^x][j_k^x]$. Εξ ορισμού (βλ. Κεφάλαιο 5.2.1) αυτό το τμήμα είναι υπεύθυνο να καταγράψει την εμφάνιση του k-στου θραύσματος του Q (δηλ. του F_k) σε περίπτωση που υπάρχει μία στοίχιση του Q στη x. Το δυαδικό ψηφίο στο ευρετήριο *Hitmap* που αντιστοιχεί σε αυτό το τμήμα είναι το $\mathcal{H}_{F_k}[i_k^x][j_k^x]$. Από εδώ και στο εξής, θα αναφερόμαστε σε αυτό το δυαδικό ψηφίο ως το k-στο *δυαδικό ψηφίο καλύμματος (cover bit)* της x.

Είναι εμφανές ότι η συνθήκη του Πορίσματος 5.2 μπορεί να εξεταστεί για τη θέση x μετρώντας το πλήθος των δυαδικών ψηφίων καλύμματός του τα οποία είναι hit. Εάν τουλάχιστον $\phi - \epsilon$ από αυτά είναι hit, τότε υπάρχει μία υποψήφια στοίχιση στη x. Για παράδειγμα, στο Σχήμα 5.1, η θέση $x = 6$ είναι μία υποψήφια στοίχιση εφόσον μεταξύ των δυαδικών του ψηφίων καλύμματος $\mathcal{H}_{GAT}[4][0]$, $\mathcal{H}_{TAC}[0][1]$, $\mathcal{H}_{AGG}[3][1]$ και $\mathcal{H}_{CGG}[6][1]$, τρία από αυτά είναι hit (συγκεκριμένα, όλα εκτός από το τρίτο, βλ. Σχήμα 5.2).

Μια ενδιαφέρουσα παρατήρηση είναι η ακόλουθη: Έστω x_1 και x_2 είναι οι δύο θέσεις της D που ανήκουν στην ίδια κλάση καλυμμάτων $[\alpha]$. Από την πρώτη ιδιότητα του Θεωρήματος 5.3 προκύπτει ότι, για κάθε $k \in [0, \phi)$, το k-στο δυαδικό ψηφίο καλύμματος τόσο της x_1 όσο και της x_2 ανήκει στο ίδιο bitset $\mathcal{H}_{F_k}[i_k^{[\alpha]}]$. Αναφερόμαστε σε αυτά τα bitset $B^{[\alpha]}[0], \cdots, B^{[\alpha]}[\phi-1]$, όπου $B^{[\alpha]}[k] = \mathcal{H}_{F_k}[i_k^{[\alpha]}]$ για κάθε $k \in [0, \phi)$, ως τα *bitset καλύμματος (cover bitsets)* της κλάσης $[\alpha]$. Το επόμενο θεώρημα ισχύει για τα bitset καλύμματος.

Θεώρημα 5.4. *Θεωρούμε μία θέση $x = \alpha + \lambda \cdot \pi + \epsilon$ η οποία ανήκει στην κλάση καλυμμάτων $[\alpha]$, και θεωρούμε ότι ολισθαίνουμε καθένα από τα bitset καλύμματος $B^{[\alpha]}[k]$ κατά $\Delta j_k^{[\alpha]}$ θέσεις προς το λιγότερο σημαντικό δυαδικό ψηφίο (least significant bit - LSB). Τότε η x είναι μια υποψήφια θέση στοίχισης ανν μπορούμε να βρούμε τουλάχιστον $\phi - \epsilon$ άσσους στο λ-στο δυαδικό ψηφίο των ολισθημένων bitset καλύμματος.*

Απόδειξη. Με βάση το Πόρισμα 5.2 υπάρχει μία υποψήφια στοίχιση στη θέση x ανν το F_k εμφανίζεται στο C_k^x, για τουλάχιστον $\phi - \epsilon$ διακριτές τιμές του $k \in [0, \phi)$. Εφόσον $\mathcal{H}_{F_k}[i_k^x][j_k^x] = B^{[\alpha]}[k][j_k^x]$, δηλ. το k-στο δυαδικό ψηφίο καλύμματος της x, είναι αυτό που κωδικοποιεί τις πιθανές εμφανίσεις του F_k μέσα στο τμήμα C_k^x, αυτό σημαίνει ότι υπάρχει μία υποψήφια στοίχιση της x ανν $B^{[\alpha]}[k][j_k^x] = 1$ για τουλάχιστον $\phi - \epsilon$ διακριτές τιμές του k. Αν ολισθήσουμε κάθε $B^{[\alpha]}[k]$ κατά $\Delta j_k^{[\alpha]}$ προς το LSB, αυτό σημαίνει ότι το x είναι μία υποψήφια στοίχιση ανν $B^{[\alpha]}[k][j'] = 1$ για τουλάχιστον $\phi - \epsilon$ διακριτές τιμές του k, όπου $j' = j_k^x = \Delta j_k^{[\alpha]} = j_0^x = \left\lfloor \frac{x-\epsilon}{\pi} \right\rfloor = \left\lfloor \frac{\alpha+\lambda\cdot\pi}{\pi} \right\rfloor = \lambda$. Συνεπώς, ανν τουλάχιστον $\phi - \epsilon$ από τα λ-στα δυαδικά ψηφία των ολισθημένων bitset καλύμματος είναι άσσοι, τότε το x είναι μία υποψήφια στοίχιση. □

Και ένα άμεσο αποτέλεσμα του προηγούμενο θεωρήματος είναι το παρακάτω.

Πόρισμα 5.3. *Θεωρούμε την κλάση καλυμμάτων $[\alpha]$ και θεωρούμε ότι ολισθαίνουμε καθένα από τα bitset καλύμματος της $B^{[\alpha]}[k]$ κατά $\Delta j_k^{[\alpha]}$ θέσεις προς το λιγότερο σημαντικό δυαδικό ψηφίο. Τότε, εάν μπορούμε να βρούμε τουλάχιστον $\phi - \epsilon$ άσσους*

hitmap()
Input: $\mathcal{H}, Q, \epsilon, \phi$
Output: $\mathbb{A}$
begin
01. **foreach** α **in** $[0, \pi)$
02. $B^{[\alpha]} \leftarrow$ new array
03. **foreach** k **in** $[0, \phi)$
04. $F_k \leftarrow Q.getFrag(k, \phi)$
05. $B^{[\alpha]}[k] \leftarrow \mathcal{H}_{F_k}[i_k^{[\alpha]}]$
06. $B^{[\alpha]}[k] \leftarrow lsb\_shift(B^{[\alpha]}[k], \Delta j_k^{[\alpha]})$
07. **end**
08. $O_{B^{[\alpha]}}^{\phi-\epsilon} \leftarrow t\_occ\_bitset(B^{[\alpha]}, \phi - \epsilon)$
09. $\mathbb{A}.addCandidates(O_{B^{[\alpha]}}^{\phi-\epsilon}, [\alpha])$
10. **end**
end.

Σχήμα 5.3: Ο αλγόριθμος *Hitmap*.

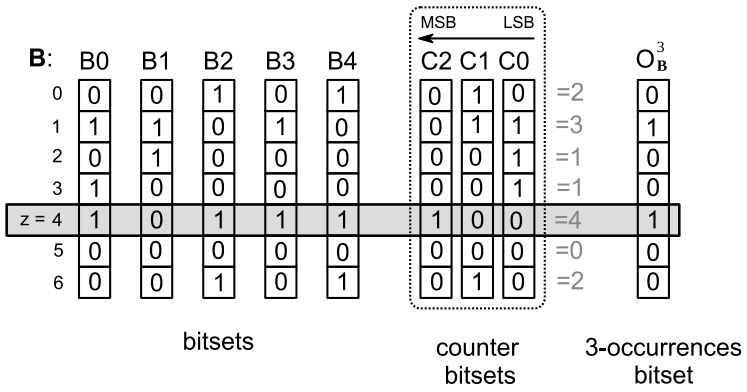
στο λ-στο δυαδικό ψηφίο των ολισθημένων bitset καλύμματος προκύπτει ότι η θέση $x = \alpha + \lambda \cdot \pi + \epsilon$ *είναι μία υποψήφια στοίχιση.*

Με βάση το Πόρισμα 5.3 μπορούμε να βρούμε τις υποψήφιες στοιχίσεις ενός ερωτήματος Q που βρίσκονται σε θέσεις που ανήκουν σε μια συγκεκριμένη κλάση καλυμμάτων. Συνεπώς, εφαρμόζοντας το Πόρισμα 5.3 για όλες τις κλάσεις καλυμμάτων, είναι εφικτό να εντοπιστούν υποψήφιες στοιχίσεις οπουδήποτε μέσα στη D. Ο ψευδοκώδικας του Σχήματος 5.3 παρουσιάζει την προαναφερθείσα προσέγγιση. Αναφερόμαστε σε αυτό τον αλγόριθμο ως ο *αλγόριθμος Hitmap*.

Η είσοδος του αλγορίθμου *Hitmap* αποτελείται από το ευρετήριο *Hitmap* $\mathcal{H}$ της D, το ερώτημα Q, το κατώφλι στοίχισης ϵ και το επιθυμητό πλήθος θραυσμάτων ερωτήματος ϕ (το τελευταίο καθορίζει επίσης και το μέγεθος των θραυσμάτων f). Η έξοδος του αλγορίθμου είναι το σύνολο όλων των υποψηφίων θέσεων στοίχισης $\mathbb{A}$.

Ο αλγόριθμος αποτελείται από ένα βρόχο που εκτελείται μία φορά για κάθε κλάση καλύμματος (γραμμές 1-10). Κατά τη διάρκεια κάθε επανάληψης, ο αλγόριθμος, πρώτα δεσμεύει χώρο για έναν πίνακα από bitset, τα οποίο προορίζονται (στο τέλος) να κρατούν τα bitset καλύμματος της κλάσης (γραμμή 2). Μετά, ο αλγόριθμος εξάγει ένα-ένα όλα τα θραύσματα του Q (γραμμή 4) και τα χρησιμοποιεί για να ανακτήσει τα bitset καλύμματος τους για την τρέχουσα κλάση $[\alpha]$ (γραμμή 5). Στη συνέχεια, κάθε bitset καλύμματος ολισθαίνει κατάλληλα σύμφωνα με το Πόρισμα 5.3 (γραμμή 6). Όταν όλα τα bitset καλύμματος είναι έτοιμα και αποθηκευμένα στον πίνακα bitset $B^{[\alpha]}$, ο αλγόριθμος επεξεργάζεται αυτόν τον πίνακα για να βρει εκείνες τις θέσεις του bitset καλύμματος $\lambda \in [0, \lceil d/\pi \rceil)$ για τις οποίες ισχύει ότι $B^{[\alpha]}[k][\lambda] = 1$ για τουλάχιστον $\phi - \epsilon$ διακριτές τιμές του k (γραμμή 8). Το τελευταίο επιτυγχάνεται από τη συνάρτηση $t\_occ\_bitset()$, η οποία εμπλέκει λειτουργίες bitwise (περιγράφονται στο Κεφάλαιο 5.3.1.1) για να δημιουργήσουν ένα bitset $O_{B^{[\alpha]}}^{\phi-\epsilon}$ για το οποίο $O_{B^{[\alpha]}}^{\phi-\epsilon}[\lambda] = 1$ ανν λ είναι μια θέση bitset για την οποία η συνθήκη που περιγράψαμε προηγουμένως ισχύει. Τέλος, το bitset $O_{B^{[\alpha]}}^{\phi-\epsilon}$ επεξεργάζεται για να παραχθούν οι υποψήφιες στοιχίσεις του με βάση το Πόρισμα 5.3 και οι υποψήφιες για όλες τις κλάσεις προστίθενται στο ίδιο σύνολο $\mathbb{A}$.

Στο επόμενο κεφάλαιο παρέχουμε λεπτομέρειες για τις λειτουργίες bitwise που χρησιμοποιούνται από τη συνάρτηση $t\_occ\_bitset()$ για να παραθχεί το bitset που κωδικοποιεί τις υποψήφιες στοιχίσεις μίας κλάσης.



Σχήμα 5.4: Ένα παράδειγμα συνόλου bitsets, το bitset t-παρουσιών του και τα bitset μετρητών που χρησιμοποιούνται για τον υπολογισμό του bitset t-παρουσιών ($t = 3$).

5.3.1.1 Αποδοτικές λειτουργίες bitwise για να βρεθούν οι υποψήφιες στοιχίσεις

Προτού προχωρήσουμε παρακάτω, να σημειωθεί ότι τα σύμβολα $\oplus$, $\wedge$ και $\vee$ χρησιμοποιούνται για να συμβολίσουν τις πράξεις bitwise-XOR, bitwise-AND και bitwise-OR, αντίστοιχα, ενώ τα σύμβολα $\oplus^b$, $\wedge^b$ και $\vee^b$ χρησιμοποιούνται για να συμβολίσουν τις δυαδικές πράξεις bit-XOR, bit-AND και bit-OR.

Θεωρούμε το σύνολο των bitset $\boldsymbol{B} = \{B_0, \cdots, B_n\}$, όπου κάθε B_i περιέχει m δυαδικά σύμβολα, $\forall i \in [0, n]$. Ορίζουμε το *bitset t-παρουσιών (t-occurrences bitset)* του $\boldsymbol{B}$ ως το bitset $O^t_{\boldsymbol{B}}$ για το οποίο ισχύει ότι:

$$O^t_{\boldsymbol{B}}[z] = \begin{cases} 1 & \text{ανν } B_i[z] = 1 \text{ για τουλάχιστον } t \text{ διακριτές τιμές του } i \text{ στο } [0, n] \\ 0 & \text{διαφορετικά} \end{cases}$$

Το Σχήμα 5.4 απεικονίζει ένα σύνολο από 5 bitset (στα αριστέρα) και το 3-παρουσιών bitset τους (στα δεξιά). Να σημειωθεί ότι $O^t_{\boldsymbol{B}}[1] = 1$ γιατί υπάρχουν 3 bitsets (B_0, B_1 και B_3) τα οποία έχουν άσσο στη θέση 1. Από την άλλη, $O^t_{\boldsymbol{B}}[2] = 0$ επειδή υπάρχει μόνο ένα bitset (B_1) το οποίο έχει άσσο στη θέση 2.

Το Σχήμα 5.5 παρουσιάζει έναν αλγόριθμο που υπολογίζει το bitset t-παρουσιών ενός δεδομένου συνόλου bitset. Η είσοδός του αποτελείται από το σύνολο bitset $\boldsymbol{B}$ και το κατώφλι t. Η έξοδός του είναι το bitset t-παρουσιών. Να σημειωθεί ότι η συνάρτηση $bin(\alpha, \beta)$ επιστρέφει ένα bitset που περιέχει τη δυαδική αναπαράσταση του α χρησιμοποιώντας β δυαδικά ψηφία.

Έπειτα από τις απαραίτητες αρχικοποιήσεις (γραμμές $1-3$), μετράμε τους άσσους για όλες τις δυνατές θέσεις bitset (γραμμές $4-11$) και έπειτα, με βάση το αποτέλεσμα, υπολογίζουμε το bitset t-παρουσιών (γραμμές $12-17$). Κατά τη διάρκεια της καταμέτρησης, χρησιμοποιούμε κάποια βοηθητικά bitset. Καταρχάς, χρησιμοποιούμε το $\boldsymbol{C} = \{C_0, \cdots, C_{|\boldsymbol{C}|}\}$, ένα σύνολο από $|\boldsymbol{C}| = \lceil log_2(n+1) \rceil$ bitset, που χρησιμοποιούνται για να κωδικοποιηθεί το πλήθος των άσσων για κάθε πιθανή θέση bitset. Αναφερόμαστε σε αυτά τα βοηθητικά bitset ως τα *bitset μετρητές (counter bitsets)*.

Τα bitset μετρητές κωδικοποιούν τα πλήθη δυαδικών ψηφίων που είναι άσσοι με τον ακόλουθο τρόπο. Θεωρούμε τα δυαδικά ψηφία $C_0[z]$, ..., $C_{|\boldsymbol{C}|}[z]$, για κάποιο $z \in [0, m]$. Αν θεωρήσουμε ότι αυτά τα δυαδικά ψηφία σχηματίζουν ένα 'εικονικό bitset' το οποίο έχει το $C_0[z]$ ως το λιγότερο σημαντικό του δυαδικό ψηφίο (LSB), το $C_1[z]$ το επόμενο σε σημασία δυαδικό του ψηφίο και ούτω καθεξής, τότε αυτό το εικονικό bitset περιέχει τη δυαδική αναπαράσταση του πλήθους των z-στων δυαδικών

t_occ_bitset()
Input: $\boldsymbol{B}$,t
Output: $O^t_{\boldsymbol{B}}$
begin
#Initialisations
01. **foreach** y **in** $[0, \lceil log_2(n+1) \rceil)$
02. $\quad C_y \leftarrow bin(0, m+1)$
03. **end**
#Count the set bits
04. **foreach** w in $[0, n]$
05. $\quad R_0 \leftarrow C_0 \wedge B_w$
06. $\quad C_0 \leftarrow C_0 \oplus B_w$
07. $\quad$ **foreach** y **in** $[1, \lceil log_2(n+1) \rceil)$
08. $\quad\quad R_y \leftarrow C_y \wedge R_{y-1}$
09. $\quad\quad C_y \leftarrow C_y \oplus R_{y-1}$
10. $\quad$ **end**
11. **end**
#Compute t-set flags
12. $tbin \leftarrow bin(t, \lceil log_2(n+1) \rceil)$
13. $O^t_{\boldsymbol{B}} \leftarrow \neg bin(0, m+1)$
14. **foreach** y **in** $[0, \lceil log_2(n+1) \rceil)$
15. $\quad$ **if** $tbin[y] == 0$ **then** $O^t_{\boldsymbol{B}} \leftarrow O^t_{\boldsymbol{B}} \vee C_y$
16. $\quad$ **else** $O^t_{\boldsymbol{B}} \leftarrow O^t_{\boldsymbol{B}} \wedge C_y$
17. **end**
end.

Σχήμα 5.5: Ψευδοκώδικας ενός αλγορίθμου για τον υπολογισμό του bitset t-παρουσιών ενός δοθέντος συνόλου bitset.

ψηφίων που είναι άσσοι. Αναφερόμαστε σε αυτό το εικονικό bitset ως τον *μετρητή δυαδικών ψηφίων της θέσης z (bit-counter of position z)*.

Για παράδειγμα, θεωρούμε $z = 4$ στο Σχήμα 5.4. Ισχύει ότι υπάρχουν 4 δυαδικά ψηφία που είναι άσσοι στη θέση 4, εφόσον $B_0[4] = B_2[4] = B_3[4] = B_4[4] = 1$. Μπορούμε να δούμε ότι ο μετρητής δυαδικών ψηφίων της θέσης 4 είναι $C_2[4]C_1[4]C_0[4] = 100$, το οποίο είναι η δυαδική αναπαράσταση του 4. Αναφερόμαστε στο C_0 ως το *LSB bitset μετρητή* (επειδή περιέχει τα λιγότερο σημαντικά δυαδικά ψηφία όλων των μετρητών δυαδικών ψηφίων) και στο $C_{|\boldsymbol{C}|}$ ως το *MSB bitset μετρητή*. Για παράδειγμα, στο Σχήμα 5.4, το C_0 είναι το LSB bitset μετρητής και το C_2 είναι το MSB bitset μετρητής.

Αρχικώς, τα bitset μετρητές περιέχουν μόνο μηδενικά (γραμμή 2 στο Σχήμα 5.5). Ανανεώνουμε τα δυαδικά ψηφία τους μία φορά για καθένα από τα bitset B_w στο $\boldsymbol{B}$, όπου $w \in [0, n]$ (γραμμές $4 - 11$). Αφού ανανεωθούν τα bitset μετρητές για το B_w, κωδικοποιούν το πλήθος από άσσους σε κάθε θέση για τα bitset $B_0, \ldots, B_w$, δηλ. για τα bitset που έχουμε εξετάσει μέχρι εκείνη τη στιγμή. Ως επακόλουθο, μετά την ανανέωση για το B_n, τα bitset μετρητές κωδικοποιούν το πλήθος των άσσων για όλες τις δοθείσες θέσεις δυαδικών ψηφίων.

Η ανανέωση των bitset μετρητών (γραμμές $4-11$) βασίζεται στη δυαδική πρόσθεση. Συγκεκριμένα, για να ανανεωθούν οι bitset μετρητές για το B_w, αρχικά προσθέτουμε δυαδικά καθένα από τα δυαδικά ψηφία B_w στο αντίστοιχό του δυαδικό ψηφίο του LSB bitset μετρητή και μετά προωθούμε τα κρατούμενα από όλες αυτές τις προσθέσεις ώστε να προστεθούν δυαδικά στα δυαδικά ψηφία του επόμενου bitset μετρητή. Μετά, τα νέα κρατούμενα προωθούνται για επιπλέον προσθέσεις στο επόμενο bitset μετρητή κοκ μέχρι τα δυαδικά ψηφία του MSB bitset μετρητή να ανανεωθούν. Να σημειωθεί

ότι χρησιμοποιούμε το $\boldsymbol{R} = \{R_0 \cdots R_{|\boldsymbol{R}|}\}$, ένα σύνολο από $|\boldsymbol{R}| = \lceil log_2(n+1) \rceil$ βοηθητικά bitset, στο οποίο αποθηκεύονται τα κρατούμενα όλων των δυαδικών προσθέσεων που πραγματοποιούνται (το R_w περιέχει τα κρατούμενα της δυαδικής πρόσθεσης που πραγματοποιείται στο bitset μετρητή C_w). Επιπλέον, αφού η δυαδική πρόσθεση δύο δυαδικών ψηφίων είναι το bit-XOR και το κρατούμενο αυτής της πρόσθεσης είναι το bit-AND, υλοποιούμε τις απαιτούμενες προσθέσεις χρησιμοποιώντας τις λειτουργίες bitwise-XOR και bitwise-AND (γραμμές $5-6$ και $8-9$).

Για παράδειγμα, θεωρούμε ξανά τα bitset του Σχήματος 5.4 και θεωρούμε ότι μόλις έχουμε ανανεώσει τα bitset μετρητές για το B_3 και ότι είμαστε έτοιμοι να κάνουμε το ίδιο για το B_4. Σε αυτό το βήμα, το περιεχόμενο του μετρητή δυαδικών ψηφίων της θέσης 4 είναι $C_2[4]C_1[4]C_0[4] = 011$, αφού τρία από τα bitset B_0, B_1, B_2 και B_3 (τα οποία έχουν ληφθεί υπόψη μέχρι αυτό το βήμα) έχουν άσσους στη θέση 4. Το επόμενο bitset B_4, έχει επίσης άσσο στη θέση 4. Προσθέτουμε το συγκεκριμένο δυαδικό ψηφίο στο $C_0[4]$ (το οποίο είναι το αντίστοιχο δυαδικό ψηφίο του LSB-bitset των bitset μετρητών) και παίρνουμε $C_0[4] \leftarrow C_0[4] \oplus^b B_4[4] = 0$ και $R_0[4] \leftarrow C_0[4] \wedge^b B_4[4] = 1$. Το παραγόμενο κρατούμενο $R_0[4]$ προστίθεται στη συνέχεια στο αντίστοιχο δυαδικό σύμβολο του bitset μετρητή επόμενης σημασίας και έτσι έχουμε: $C_1[4] \leftarrow C_1[4] \oplus^b R_0[4] = 0$ και $R_1[4] \leftarrow C_1[4] \wedge^b R_0[4] = 1$. Στο τελικό βήμα, παίρνουμε: $C_2[4] \leftarrow C_2[4] \oplus^b R_1[4]$, $R_2[4] \leftarrow C_2[4] \wedge^b R_1[4]$ και έτσι ο μετρητής δυαδικών ψηφίων της θέσης 4 γίνεται $C_2[4]C_1[4]C_0[4] = 100$. Να υπενθυμίσουμε ότι, στην πραγματικότητα, πραγματοποιούμε τις λειτουργίες δυαδικών ψηφίων ταυτόχρονα για όλες τις θέσεις χρησιμοποιώντας λειτουργίες bitwise (αναφέρουμε εδώ τις λειτουργίες δυαδικών ψηφίων μόνο για λόγους κατανόησης).

Αφού έχουμε υπολογίσει τα bitset μετρητές, πρέπει να τα χρησιμοποιήσουμε προκειμένου να υπολογίσουμε το bitset t-παρουσιών. Για να το κάνουμε αυτό, πρέπει να συγκρίνουμε τους μετρητές δυαδικών ψηφίων κάθε θέσης με το t. Αν ο μετρητής δυαδικών ψηφίων της θέσης $z \in [0, m]$ είναι μεγαλύτερος ή ίσος από το t, τότε θέτουμε $O^t_{\boldsymbol{B}}[z] = 1$. Διαφορετικά, θέτουμε $O^t_{\boldsymbol{B}}[z] = 0$. Προκειμένου να υπολογίσουμε όλα τα δυαδικά ψηφία των t-παρουσιών ταυτόχρονα, μετατρέπουμε το t στη δυαδική του αναπαράσταση $tbin$ (γραμμή 12 στο Σχήμα 5.5), αρχικοποιούμε τα δυαδικά ψηφία των t-παρουσιών να είναι όλα άσσοι (γραμμή 13), και μετά εφαρμόζουμε κάποιες λειτουργίες bitwise πάνω στους bitset μετρητές και το bitset t-παρουσιών, βασισμένοι στα δυαδικά ψηγία του $tbin$ (γραμμές $14-17$). Να σημειωθεί ότι το $tbin$ περιέχει ένα δυαδικό ψηφίο για κάθε bitset μετρητή.

Η βασική ιδέα πίσω από τις γραμμές $14-17$ είναι η ακόλουθη. Ανανεώνουμε το $O^t_{\boldsymbol{B}}$, μία φορά για κάθε bitset μετρητή, δηλ. μία φορά για κάθε δυαδικό ψηφίο στους μετρητές δυαδικών ψηφίων. Η $(y+1)$-στη ανανέωση του $O^t_{\boldsymbol{B}}$, όπου $y \in [0, \lceil log_2(n+1) \rceil)$, επιτυγχάνεται εφαρμόζοντας μια λειτουργία bitwise $\diamond$ μεταξύ του $O^t_{\boldsymbol{B}}$ και του bitset μετρητή C_y, δηλ. $O^t_{\boldsymbol{B}} \leftarrow O^t_{\boldsymbol{B}} \diamond C_y$. Για κάθε δυαδικό ψηφίο $O^t_{\boldsymbol{B}}[z]$ του ανανεωμένου $O^t_{\boldsymbol{B}}$, όπου $z \in [0, m]$, πρέπει να ισχύει ότι $O^t_{\boldsymbol{B}}[z] = 1$ ανν $C_0[z] \cdots C_y[z] \geq tbin[0] \cdots tbin[y]$, ή $O^t_{\boldsymbol{B}}[z] = 0$ σε διαφορετική περίπτωση. Συνεπώς, η bitwise λειτουργία $\diamond$ πρέπει να επιλέγεται για να εγγυηθεί η προηγούμενη συνθήκη. Την ώρα πρίν από την $(y+1)$-στη ανανέωση, το $O^t_{\boldsymbol{B}}[z]$ περιέχει δυαδικά ψηφία που καθορίζονται από τα πρώτα (δηλ. τα λιγότερο σημαντικά) y δυαδικά ψηφία του $tbin$ και τα πρώτα y bitsets μετρητές (ή, ισοδύναμα, τα πρώτα y δυαδικά ψηφία όλων των μετρητών δυαδικών ψηφίων). Κατά τη διάρκεια αυτής της ανανέωσης ασχολούμαστε με το $tbin[y]$ (δηλ. το $(y+1)$-στο δυαδικό ψηφίο του $tbin$) και το C_y (δηλ. το $(y+1)$-στο bitset μετρητή). Έστω ότι $tbin[y] = 0$. Για κάθε $z \in [0, m]$, υπάρχουν δύο περιπτώσεις: (α) αν $C_y[z] = 0 = tbin[y]$, συνε-

πώς δεν μεταβάλλουμε το $O^t_{\boldsymbol{B}}[z]$ (όλα τα εμπλεκόμενα δυαδικά νούμερα παραμένουν τα ίδια καθώς μηδενικά προστέθηκαν μετά από το προηγούμενο πιο σημαντικό δυαδικό ψηφίο τους) και (β) αν $C_y[z] = 1 > tbin[y]$, τότε $C_0[z] \cdots C_y[z]$ είναι μεγαλύτερο από το $tbin[0] \cdots tbin[y]$. Από τα προηγούμενα, προκύπτει ότι $O^t_{\boldsymbol{B}}[z] \diamond^b 0 = F[z]$ και $O^t_{\boldsymbol{B}}[z] \diamond^b 1 = 1$, επομένως το $\diamond^b$ πρέπει να είναι το bit-OR και, ως αποτέλεσμα, το $\diamond$ πρέπει να είναι το bitwise-OR. Ακολουθώντας παρόμοια λογική μπορούμε να δείξουμε ότι αν $tbin[y] = 1$, τότε το $\diamond$ πρέπει να είναι το bitwise-AND.

5.3.2 Βελτιστοποιώντας τις λειτουργίες bitwise

Οι λειτουργίες bitwise που εμπλέκονται στην εκτέλεση του αλγορίθμου *Hitmap* εκτελούνται σε μεγάλα bitset, συνεπώς, μπορεί να είναι απαιτητικές. Παρόλα αυτά, εφόσον τα εμπλεκόμενα bitset είναι αραία (περιέχουν πολύ περιορισμένο πλήθος από άσσους), υπάρχει η δυνατότητα για κάποιες βελτιστοποιήσεις.

Συγκεκριμένα, μπορούμε να αναπαραστήσουμε οποιοδήποτε bitset ως ένα σύνολο από μπλοκ συγκεκριμένου μήκους. Μετά οι λειτουργίες bitwise που εφαρμόζονται στα δύο bitset μετασχηματίζονται στην εφαρμογή των ίδιων λειτουργιών στα μπλοκ τους. Να σημειωθεί ότι εφόσον τα bitset είναι αραιά, τα περισσότερα από τα μπλοκ τους αναμένεται να είναι άδεια, δηλ. γεμάτα με μηδενικά. Αυτό είναι πολύ χρήσιμο επειδή οι λειτουργίες bitwise που εμπλέκουν άδεια bitset έχουν αναμενόμενα αποτελέσματα, επομένως, αυτές οι λειτουργίες bitwise είναι δυνατό να αποφευχθούν.

Θεωρούμε ένα μη-άδειο μπλοκ B και έστω $B_\varnothing$ ότι συμβολίζει ένα άδειο μπλοκ. Τότε, οι ακόλουθες ιδιότητες ισχύουν:

$$B \wedge B_\varnothing = B_\varnothing$$

$$B \vee B_\varnothing = B$$

$$B \oplus B_\varnothing = B$$

Χρησιμοποιώντας τις παραπάνω τρεις ιδιότητες, αποφεύγουμε την εφαρμογή πολλών λειτουργιών bitwise κάτι που έχει ως αποτέλεσμα βελτιωμένες επιδόσεις για το *Hitmap*. Να σημειωθεί ότι το μέγεθος του μπλοκ είναι μία παράμετρος του *Hitmap* που μπορεί να επηρεάσει σημαντικά το χρόνο εκτέλεσης. Τέλος, να σημειωθεί ότι ο τρόπος που τα bitset αποθηκεύονται στο ευρετήριο (και ο οποίος περιγράφεται στο Κεφάλαιο 5.2.3) δεν μεταβάλλεται. Τα συμπιεσμένα bitset μετατρέπονται στην αναπαράσταση τύπου μπλοκ ακριβώς πριν από την εφαρμογή των λειτουργιών bitwise πάνω τους.

5.4 Αποτίμηση

Σε αυτό το κεφάλαιο περιγράφουμε ένα σύνολο πειραμάτων που εκτελέσαμε για να αξιολογήσουμε τις επιδόσεις του ευρετηρίου και του αλγορίθμου *Hitmap* σε σύγκριση με τις προσεγγίσεις αιχμής στην περίπτωση της στοίχισης αναγνωσμάτων DNA μεγάλου μήκους. Οι βασικοί ανταγωνιστές μας είναι (α) το ευρετήριο *WHAM* [54], το οποίο είναι η λύση αιχμής για στοίχιση αναγνωσμάτων DNA μικρού μήκους και (β) το ευρετήριο *RBSA*[75], το οποίο έχει διαπιστωθεί να συμπεριφέρεται καλά για αναγνώσματα DNA μεγάλου μήκους στο παρελθόν.

Στο Κεφάλαιο 5.4.1 συζητούμε τη πειραματική διάταξη και τα σύνολα δεδομένων που χρησιμοποιήθηκαν κατά τη διάρκεια των πειραμάτων, ενώ στο Κεφάλαιο 5.4.2 αξιολογούμε τις επιδόσεις του *Hitmap* έναντι των προαναφερθέντων ανταγωνιστών.

5.4.1 Πειραματική διάταξη

Σύστημα. Υλοποιήσαμε όλους τους αλγορίθμους σε C++, και πραγματοποιήσαμε τα πειράματα σε έναν εξυπηρετητή Linux που έχει δύο Intel Xeon E5607 επεξεργαστές στα $2.27GHz$ και κύρια μνήμη μεγάθους $64GBs$.

Σύνολα δεδομένων. Για τα συγκριτικά πειράματα του Κεφαλαίου 5.4.2 που εμπλέκουν το RBSA χρησιμοποιήσαμε ένα μικρό συνθετικό σύνολο δεδομένων που περιείχε $100,000$ σύμβολα, όπου κάθε σύμβολο επιλέχθηκε με τυχαίο τρόπο με βάση το ομοιόμορφο μοντέλο Bernoulli (δηλ. κάθε σύμβολο έχει πιθανότητα $1/|\Sigma|$ να εμφανιστεί και επιλέγεται ανεξάρτητα από τα υπόλοιπα). Αυτή η επιλογή έγινε αφού η κατασκεύη του RBSA είναι υπολογιστικά απαιτητική (εμπλέκει υπολογισμό συντακτικής απόστασης μεταξύ κάθε ακολουθίας αναφοράς και κάθε πιθανής υπακολουθίας των δεδομένων). Ακόμα και στην περίπτωση αυτής της απλοϊκής ακολουθίας δεδομένων που χρησιμοποιήθηκε για τα πειράματά μας η κατασκευή του ευρετηρίου RBSA απαιτούσε αρκετές ώρες για κάποιες συγκεκριμένες παραμετροποιήσεις. Αναφερόμαστε στο προηγούμενο σύνολο δεδομένων ως SYNTH.

Για τα υπόλοιπα πειράματα χρησιμοποιήσαμε δεδομένα από το αφιλτράριστο (unmasked) ανθρώπινο γονιδίωμα το οποίο αντλήσαμε από το *Genome Reference Consortium* (συγκεκριμένα, η έκδοση GRCh38 χρησιμοποιήθηκε). Αυτό το σύνολο δεδομένων περιέχει ακολουθίες από τα 24 ανθρώπινα χρωμοσώματα και έχει συνολικό μέγεθος 3.14GB (αυτά τα δεδομένα περιλαμβάνουν τις ίδιες τις ακολουθίες μαζί με κάποια μεταδεδομένα). Κάθε ακολουθία χρωμοσώματος αποτελείται από $50-250$ εκατομμύρια σύμβολα.

Να σημειωθεί ότι σε όλες τις περιπτώσεις (δηλ. τόσο για τα συνθετικά όσο και για τα πραγματικά δεδομένα), οι ακολουθίες ερωτήματα ήταν τυχαία επιλεγμένες υπακολουθίες των χρησιμοποιούμενων συνόλων δεδομένων.

Παραμετροποιήσεις των ευρετηρίων. Πραγματοποιήσαμε ενδελεχείς ελέγχους προκειμένου να βεβαιωθούμε ότι κάθε ευρετήριο είναι παραμετροποιημένο βέλτιστα για κάθε πείραμα.

Βρήκαμε ότι για την περίπτωση των σχετικά μεγάλων ακολουθιών ερωτημάτων (> 50 σύμβολα), το WHAM αποδίδει καλύτερα όταν παραμετροποιείται ώστε να ταυτίζεται με το κλασσικό *ευρετήριο gram του Navarro* [68], δηλ. όταν το πλήθος των θραυσμάτων του ερωτήματος επιλέγεται να είναι ίσο με $\epsilon+1$ (βλ. επίσης Κεφάλαιο 2.1.1.3). Αυτό δεν ήταν έκπληξη εφόσον οι δημιουργοί του ακολουθούν την ίδια προσέγγιση στο [54]. Παρόλα αυτά, αυτή είναι η πρώτη φορά που το WHAM μετασχηματίζεται για να λειτουργεί με κατώφλια συντακτικής απόστασης (βλ. Κεφάλαιο 5.1.3), έτσι ήταν απαραίτητο να αποκαλυφθούν οι ιδανικές παραμετροποιήσεις για το συγκεκριμένο σενάριο. Να σημειωθεί ότι με την αύξηση του πλήθος των θραυσμάτων παρατηρήθηκε χειρότερη επίδοση σε όλες τις περιπτώσεις, ακόμα και σε αυτές που το φιλτράρισμα που επιτυγχανόταν ήταν καλύτερο. Ο λόγος είναι ότι σε αυτές τις περιπτώσεις το πλήθος των ελέγχων στο ευρετήριο WHAM αυξάνεται δραματικά.

Για το ευρετήριο RBSA, χρησιμοποιήσαμε ένα σύνολο από 500 τυχαία επιλεγμένες ακολουθίες αναφοράς. Κάθε θέση στα δεδομένα συνδέθηκε με 50 από τις ακολουθίες αναφοράς με βάση μια ελαφρώς παραλλαγμένη έκδοση της άπληστης προσέγγισης που περιγράφεται στο [75]. Η παραλλαγή μας ήταν ότι, αντί να επιλέγουμε τυχαία το δείγμα ερωτημάτων, χρησιμοποιήσαμε τις ακολουθίες ερωτημάτων για τη δουλειά αυτή. Φυσικά αυτή η μεταβολή δεν είναι δίκαιη για το WHAM και το *Hitmap*, αφού το ευρετήριο RBSA που κατσκευάζεται επιτυγχάνει βέλτιστο φιλτράρισμα για τα ερωτήματα των

πειραμάτων μας χρησιμοποιώντας μια εκ των προτέρων γνώση για αυτά. Παρόλα αυτά, χρησιμοποιήσαμε αυτό το μη-ρεαλιστικό σενάριο απλώς για να δείξουμε ότι ακόμα και υπό αυτές τις ιδανικές συνθήκες, το RBSA συμπεριφέρεται σημαντικά χειρότερα από το WHAM και το το *Hitmap*.

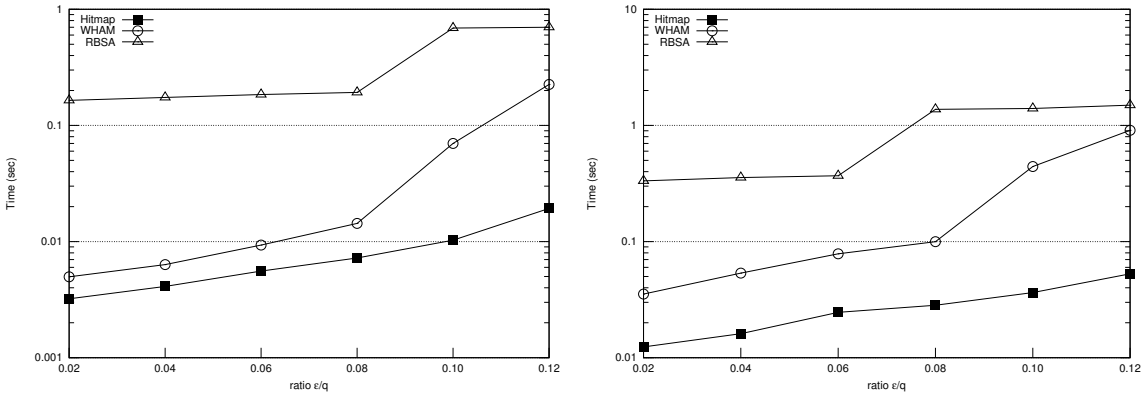
Το ευρετήριο *Hitmap* παραμετροποιήθηκε για να έχει $\phi = \epsilon + 2$ ή $\phi = \epsilon + 1$ σε περιπτώσεις που η τιμή ϵ/q ήταν μικρή. Για τις υπόλοιπες περιπτώσεις περισσότερα θραύσματα έπρεπε να χρησιμοποιηθούν προκειμένου να επιτευχθεί βελτιωμένο φιλτράρισμα. Το μέγεθος μπλοκ bitset τέθηκε στο 500 σε όλες τις περιπτώσεις.

5.4.2 Σύγκριση με WHAM και RBSA

Στο Κεφάλαιο 5.4.2.1 συγκρίνουμε τις επιδόσεις του *Hitmap* με αυτές του WHAM και του RBSA χρησιμοποιώντας το συνθετικό μας σύνολο δεδομένων. Εφόσον το *Hitmap* και το WHAM διαπιστώθηκε ότι υπερνικούν κατά κράτος το RBSA, επικεντρωνόμαστε στη συγκριτική τους αποτίμηση στο Κεφάλαιο 5.4.2.2.

5.4.2.1 Σύγκριση τριών αλγορίθμων

Το Σχήμα 5.6 απεικονίζει τις επιδόσεις της προσέγγισης *Hitmap* έναντι του WHAM και του RBSA στην περίπτωση των ερωτημάτων μεγάλου μήκους για το σύνολο δεδομένων SYNTH. Κάθε τμήμα του σχήματος αντιστοιχεί σε ένα διαφορετικό μήκος ερωτήματος q. Εκτελέσαμε πειράματα για $q = 500$ και $1,000$. Για κάθε τιμή του q μετρήσαμε το χρόνο εκτέλεσης (άξονας y, σε λογαριθμική κλίμακα) καθενός από τους τρεις αλγορίθμους για πολλά διαφορετικά κατώφλια στοίχισης ϵ που ποικίλουν από $0.02 \cdot q$ έως $0.12 \cdot q$ (άξονας x). Δεν εξετάζουμε τα κατώφλια στοίχισης που είναι μεγαλύτερα από 12% του μήκους του ερωτήματος γιατί δεν είναι πρακτικά.



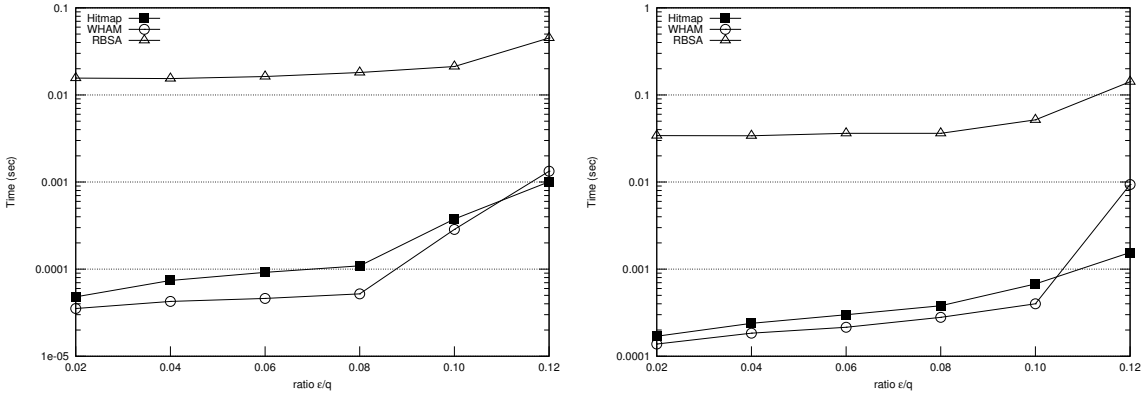
(α′) $q = 500$ (β′) $q = 1,000$

Σχήμα 5.6: Χρόνοι εκτέλεσης για το *Hitmap*, το WHAM και το RBSA για διάφορα κατώφλια στοίχισης στο σύνολο δεδομένων SYNTH, στην περίπτωση των αναγνωσμάτων μεγάλου μήκους.

Είναι εμφανές ότι το *Hitmap* υπερνικά τόσο το RBSA όσο και τοWHAM για όλα τα εξεταζόμενα κατώφλια στοίχισης. Συγκεκριμένα, σε όλες τις περιπτώσεις, το *Hitmap* είναι γρηγορότερο από το RBSA για περισσότερες από μία τάξεις μεγέθους. Επιπλέον το *Hitmap* υπερνικά το WHAM για μία τάξη μεγέθους ή και περισσότερο σε όλες τις περιπτώσεις όπου $\epsilon = 0.10 \cdot q$. Να σημειωθεί ότι για $\epsilon \geq 0.10 \cdot q$ και $q = 500$ και για $\epsilon \geq 0.08 \cdot q$ και $q = 1,000$, το RBSA αποτυγχάνει να φιλτράρει οποιοδήποτε

τμήμα των δεδομένων, έτσι, η πλήρης ακολουθία δεδομένων πρέπει να εξεταστεί για πιθανές στοιχίσεις του Q.

Ενώ το *Hitmap* είναι βελτιστοποιημένο για αναγνώσματα μεγάλου μήκους, επιτυγχάνει ικανοποιητικές επιδόσεις για αναγνώσματα μικρού μήκους. Στο Σχήμα 5.7 παρουσιάζουμε το χρόνο εκτέλεσης για το *Hitmap*, το WHAM και το RBSA για ερωτήματα μήκους $q = 50$ και 100. Ο άξονας x και πάλι αντιστοιχεί σε διαφορετικά κατώφλια στοίχισης. Να σημειωθεί ότι ο άξονας y παρουσιάζεται σε λογαριθμική κλίμακα.



(α′) $q = 50$ (β′) $q = 100$

Σχήμα 5.7: Χρόνοι εκτέλεσης για το *Hitmap*, το WHAM και το RBSA για διάφορα κατώφλια στοίχισης στο σύνολο δεδομένων SYNTH, στην περίπτωση των αναγνωσμάτων μικρού μήκους.

Το *Hitmap* υπερνικά ξεκάθαρα το RBSA για περισσότερες από μία τάξεις μεγέθους. Επιπλέον, παρόλο που συμπεριφέρεται χειρότερα από το WHAM για τις περισσότερες περιπτώσεις στα μικρά αναγνώσματα, οι επιδόσεις του είναι πάντα συγκρίσιμες με αυτές του WHAM. Να υπενθυμίσουμε ότι η στοίχιση μικρών αναγνωσμάτων είναι ένα καλά μελετημένο αντικείμενο και οι υπάρχουσες προσεγγίσεις όπως είναι το WHAM και η [68] επιτυγχάνουν σχεδόν βέλτιστες επιδόσεις. Αυτό συμβαίνει επειδή κάτω από τέτοια σενάρια ακόμα και χαλαρές συνθήκες φιλτραρίσματος (όπως αυτές που χρησιμοποιούνται από το WHAM και το [68]) επιτυγχάνουν την αποφυγεί περισσότερου από το 99% των δεδομένων. Από την άλλη, το*Hitmap* επικεντρώνεται στην παροχή σχεδόν βέλτιστου φιλτραρίσματος ακόμα και στις δύσκολες περιπτώσεις, όπου το φιλτράρισμα του WHAM και του RBSA αποτυγχάνουν να αποφύγουν μεγάλο μέρος των δεδομένων.

Στον Πίνακα 5.1 παρουσιάζουμε την τιμή του ϕ που χρησιμοποιήσουμε για το *Hitmap* σε καθένα από τα προηγούμενα πειράματα.

Για να συνοψίσουμε, το *Hitmap* αποδείχθηκε ότι συμπεριφέρεται καλύτερα από το WHAM και το RBSA για μεγάλα ερωτήματα και αποτελεί τη μόνη ικανοποιητική λύση φιλτραρίσματος καθώς τα λάθη που περιέχονται στο ερώτημα γίνονται περισσότερα. Επιπλέον, το RBSA υπερνικάται ξεκάθαρα από τις άλλες δύο μεθόδους και έτσι στο επόμενο κεφάλαιο θα πραγματοποιήσουμε πειράματα σε πραγματικά σύνολα δεδομένων συγκρίνοντας μόνο το *Hitmap* και το WHAM.

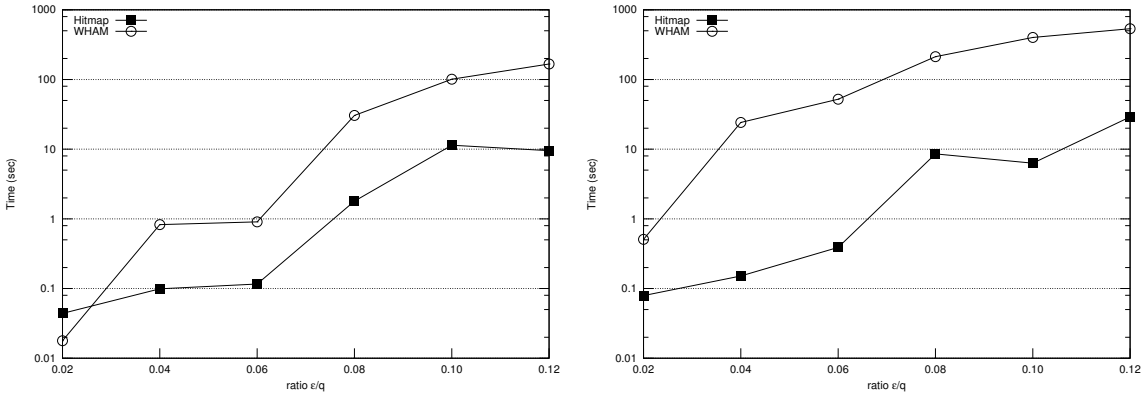
5.4.2.2 Hitmap εναντίον WHAM για μεγάλα ερωτήματα

Σε αυτό το κεφάλαιο, συγκρίνουμε τις επιδόσεις του *Hitmap* έναντι αυτών του WHAM για ερωτήματα μεγάλου μήκους σε ένα πραγματικό σύνολο δεδομένων. Συγκεκριμένα,

Πίνακας 5.1: Παραμετροποίηση του *Hitmap* για τα πειράματα στο σύνολο δεδομένων SYNTH.

q	ϵ/q	ϕ	q	ϵ/q	ϕ	q	ϵ/q	ϕ	q	ϵ/q	ϕ
50	0.02	$\epsilon+1$	100	0.02	$\epsilon+1$	500	0.02	$\epsilon+2$	1,000	0.02	$\epsilon+2$
50	0.04	$\epsilon+1$	100	0.04	$\epsilon+2$	500	0.04	$\epsilon+2$	1,000	0.04	$\epsilon+2$
50	0.06	$\epsilon+1$	100	0.06	$\epsilon+2$	500	0.06	$\epsilon+2$	1,000	0.06	$\epsilon+2$
50	0.08	$\epsilon+1$	100	0.08	$\epsilon+2$	500	0.08	$\epsilon+2$	1,000	0.08	$\epsilon+3$
50	0.10	$\epsilon+2$	100	0.10	$\epsilon+2$	500	0.10	$\epsilon+3$	1,000	0.10	$\epsilon+4$
50	0.12	$\epsilon+2$	100	0.12	$\epsilon+2$	500	0.12	$\epsilon+5$	1,000	0.12	$\epsilon+5$

πραγματοποιούμε τα πειράματά μας χρησιμοποιώντας ως δεδομένα το χρωμόσωμα 19 του Homo sapiens (μήκους 58,617,616 συμβόλων). Το Σχήμα 5.8 δείχνει τους χρόνους εκτέλεσης του *Hitmap* και του WHAM (ο άξονας y είναι σε λογαριθμική κλίμακα, ενώ ο άξονας x αντιστοιχεί σε διαφορετική επιλογή από κατώφλια στοίχισης).



(α′) $q = 500$ (β′) $q = 1,000$

Σχήμα 5.8: Χρόνοι εκτέλεσης για το *Hitmap* και το WHAM για διάφορα κατώφλια στοίχισης στο χρωμόσωμα 19 του Homo sapiens, στην περίπτωση των μεγάλων αναγνωσμάτων.

Για την περίπτωση $q = 500$, το *Hitmap* υπερνικά το WHAM για περίπου μία τάξη μεγέθους σχεδόν σε όλες τις περιπτώσεις. Εξαίρεση αποτελεί η περίπτωση $\epsilon = 0.02 \cdot q$, όπου το WHAM συμπεριφέρεται ελάχιστα καλύτερα. Ο λόγος είναι ότι σε αυτή την περίπτωση, περισσότερος χρόνος καταναλώνεται στις κλήσεις του ευρετηρίου παρά στην επαλήθευση υποψηφίων στοιχίσεων. Και οι δύο μέθοδοι επιτυγχάνουν ικανοποιητικό φιλτράρισμα.

Για την περίπτωση $q = 1,000$, οι επιδόσεις του *Hitmap* είναι περίπου δύο τάξεις μεγέθους καλύτερες από αυτές του WHAM για σχεδόν όλες τις περιπτώσεις. Η μόνη εξαίρεση είναι, και πάλι, η περίπτωση $\epsilon = 0.02 \cdot q$, για την οποία το *Hitmap* υπερνικά το WHAM για μία τάξη μεγέθους.

Στον Πίνακα 5.2 παρουσιάζουμε τις τιμές ϕ που χρησιμοποιήθηκαν για το *Hitmap* σε καθένα από τα προηγούμενα πειράματα.

Συνοψίζοντας, το *Hitmap* αποδείχθηκε να αποδίδει καλύτερα από το WHAM για τις περισσότερες περιπτώσεις στα μεγάλα ερωτήματα και αποτελεί τη μόνη ικανοποιητική λύση φιλτραρίσματος καθώς το πλήθος των λαθών που περιέχονται στο ερώτημα αυξάνονται.

Πίνακας 5.2: Παραμετροποίηση του *Hitmap* για τα πειράματα στο χρωμόσωμα 19 του Homo sapiens.

q	ϵ/q	ϕ	q	ϵ/q	ϕ
500	0.02	$\epsilon + 2$	1,000	0.02	$\epsilon + 3$
500	0.04	$\epsilon + 3$	1,000	0.04	$\epsilon + 3$
500	0.06	$\epsilon + 4$	1,000	0.06	$\epsilon + 6$
500	0.08	$\epsilon + 5$	1,000	0.08	$\epsilon + 10$
500	0.10	$\epsilon + 10$	1,000	0.10	$\epsilon + 11$
500	0.12	$\epsilon + 10$	1,000	0.12	$\epsilon + 13$

5.5 Συμπεράσματα

Παρουσιάσαμε το *Hitmap* μια δομή ευρετηρίου μαζί με έναν αλγόριθμο που υποστηρίζει αποδοτική στοίχιση για αναγνώσματα μεγάλου μήκους και για μεγάλο πλήθος λαθών. Το *Hitmap* καλύπτει το κενό που αφήνουν οι προσεγγίσεις στοίχισης αναγνωσμάτων καθώς υπερνικά τις προσεγγίσεις αιχμής στην περίπτωση των αναγνωσμάτων DNA μεγάλου μήκους, ενώ οι επιδόσεις του για στοίχιση μικρών αναγνωσμάτων παραμένει συγκρίσιμη με αυτή των καλύτερων αλγορίθμων για στοίχιση αναγνωσμάτων μικρού μήκους.

Κεφάλαιο 6

Συμπεράσματα και Μελλοντικές εργασίες

Αυτή η διατριβή παρουσίασε διάφορες μεθόδους για διαχείριση μεγάλου όγκου δεδομένων από τις βιοεπιστήμες. Το ενδιαφέρον μας επικεντρώθηκε σε δύο ενδιαφέροντα προβλήματα: (α) την αποδοτική πρόβλεψη στόχων miRNA και (β) την αποδοτική στοίχιση αναγνωσμάτων DNA μεγάλου μήκους. Για το πρώτο πρόβλημα κατορθώσαμε να κατασκευάσουμε μεθόδους και συστήματα που παράγουν τους στόχους των miRNA σε σχεδόν πραγματικό χρόνο. Αυτό έγινε δυνατό με τη συνδυαστική χρήση νέων μεθόδων κατά προσέγγιση ταιριάσματος ακολουθιών και την υιοθέτηση προσεγγίσεων βασισμένων στο Νέφος. Για το δεύτερο πρόβλημα, προτείναμε το *Hitmap* μια νέα δομή ευρετηρίου που επιτρέπει την αποδοτική στοίχιση αναγνωσμάτων DNA σε γονιδιώματα αναφοράς για κάθε μέγεθος αναγνώσματος, ακόμα και για αναγνώσματα που αποτελούνται από εκατοντάδες ή χιλιάδες σύμβολα. Τέλος, υλοποιήσαμε μια μεγάλη ποικιλία εργαλείων Ιστού για να διευκολύνουμε τους επιστήμονες του χώρου της έρευνας για τα miRNA. Τα εργαλεία μας καθιστούν διαθέσιμη και εύκολη στην αναζήτηση πληροφορία η οποία προηγουμένως είτε ήταν διασκορπισμένη είτε δεν υπήρχε καθόλου.

Στο υπόλοιπο αυτού του κεφαλαίου συζητάμε πιο λεπτομερειακά τις συνεισφορές μας και αναγνωρίζουμε ενδιαφέροντα θέματα τα οποία προτείνουμε για μελλοντική εργασία.

6.1 Σύνοψη

Αρχικά ασχοληθήκαμε με το πρόβλημα της πρόβλεψης στόχων miRNA. Ο στόχος μας ήταν να παρέχουμε στους βιοεπιστήμονες ακριβή πρόβλεψη για στόχους miRNA σε σχεδόν πραγματικό χρόνο χρησιμοποιώντας ως βάση τη μέθοδο αιχμής DIANA microT. Προς αυτή την κατεύθυνση, μελετήσαμε τη διεργασία ταιριάσματος ακολουθιών που αποτελεί το πρώτο βήμα της μεθόδου. Βρήκαμε ότι αυτή συνιστά ένα νέο είδος ερωτήματος ταιριάσματος ακολουθιών. Διατυπώσαμε μαθηματικά αυτό το ερώτημα εισάγοντας το πρόβλημα ARSM. Το ζητούμενο σε αυτό το πρόβλημα είναι να ανακτηθούν όλες οι εμφανίσεις περιοχής ενός προτύπου μέσα σε μία ακολουθία δεδομένων. Οι περιοχές του προτύπου που ταιριάζουν πρέπει να περιέχουν ένα προκαθορισμένο κομμάτι του προτύπου, τον πυρήνα. Επιπλέον, η επιτρεπόμενη διαφοροποίηση από την ακολουθία δεδομένων είναι πιο αυστηρή για μικρότερες και πιο χαλαρή για μεγαλύτερες περιοχές. Για να αντιμετωπίσουμε το προηγούμενο πρόβλημα προτείναμε τη μέθοδο `PS-ARSM`. Η μέθοδός μας εκμεταλλεύεται τις επικαλύψεις προθέματος και επιθέματος

των περιοχών για να αποφύγει περιττούς υπολογισμούς. Εκτενής πειραματική μελέτη έδειξε ότι η PS-ARSM είναι περίπου δύο τάξεις μεγέθους γρηγορότερη από τις υπάρχουσες τεχνικές όταν αυτές μετασχηματιστούν ώστε να απαντούν στο πρόβλημα ARSM.

Παρόλα αυτά, η επιτάχυνση του βήματος ταιριάσματος ακολουθιών των μεθόδων πρόβλεψης στόχων δεν είναι αρκετή για να επιτευχθεί απόδοση σχεδόν πραγματικού χρόνου. Αυτό συμβαίνει επειδή αυτές οι μέθοδοι εμπλέκουν και κάποιες άλλες υπολογιστικά απαιτητικές διεργασίες. Για να επιταχυνθεί η εκτέλεσή τους ακολουθήσαμε την προσέγγιση της κατανομής αυτών των διεργασιών στους κόμβους μιάς υποδομής Νέφους. Έτσι σχεδιάσαμε δύο συστήματα πρόβλεψης στόχων βασισμένα στο Νέφος, το TarCloud και το MR-microT. Το πρώτο αναπτύχθηκε χρησιμοποιώντας το πλαίσιο Microsoft Azure ενώ το δεύτερο ήταν μια υλοποίηση MapReduce που χρησιμοποιεί το πλαίσιο Hadoop. Με βάση τις μετρήσεις μας, και τα δύο συστήματα επιταχύνουν τη διεργασία πρόβλεψης, όμως το MR-microT είναι καλύτερο επειδή (α) είναι ανεξάρτητο πλατφόρμας (μπορεί να εγκατασταθεί σε οποιαδήποτε συστάδα σύγχρονων υπολογιστικών κόμβων, (β) παρέχει βελτιωμένο παραλληλισμό των εμπλεκόμενων εργασιών και (γ) είναι σχεδιασμένο για να υποστηρίζει ανεμπόδιστα αυξανόμενο πλήθος αιτήσεων πρόβλεψης.

Έπειτα, εργαστήκαμε συστηματικά για να προσφέρουμε πολύτιμα εργαλεία που διευκολύνουν τους ερευνητές που εργάζονται στο πεδίο της έρευνας για τα miRNA. Για αυτό το σκοπό, συλλέξαμε δεδομένα που ήταν διασκορπισμένα σε πολλές επιστημονικές δημοσιεύσεις και βάσεις δεδομένων, τα συνδυάσαμε και τα επεξεργαστήκαμε για να εξάγουμε γνώση σχετικά με το ρόλο των μορίων miRNA σε πολλούς μηχανισμούς της ζωής. Τα αποτελέσματα διανέμονται στην ερευνητική κοινότητα μέσω ενός πλήθους από πλούσια εργαλεία που διαθέτουν διαισθητικές διεπαφές Ιστού. Χρησιμοποιώντας τα οι βιοεπιστήμονες μπορούν τόσο να φυλλομετρούν τη γνώση του πεδίου που έχει καταγραφεί στις βάσες δεδομένων μας και να εκτελούν πολλούς τύπους αναλύσεων στα αποθηκευμένα δεδομένα. Συγκεκριμένα, αναπτύξαμε (α) το DIANA microT , που προσφέρει στους βιοεπιστήμονες προβλέψεις για τα γονίδια που στοχεύονται από όλα τα γνωστά miRNA, (β) το DIANA miRGen που ενημερώνει τους χρήστες σχετικά με τις γονιδιωματικές τοποθεσίες όλων των μεταγράφων miRNA και τη συμπεριφορά έκφρασής τους, (γ) το DIANA TarBase , το οποίο παρέχει πειραματικά επιβεβαιωμένους στόχους miRNA, (δ) το DIANA mirPath , που διερευνά το ρόλο των miRNA στα γνωστά μεταβολικά μονοπάτια και (ε) το DIANA mirPub , ένα εργαλείο που βοηθά τους βιοεπιστήμονες σε αναζητήσεις βιβλιογραφίας που είναι σχετική με miRNA. Κατά την ανάπτυξη του DIANA TarBase εντοπίσαμε τις δυσκολίες που αντιμετωπίζουν οι επιμελητές όταν πρέπει να αναγνωρίσουν αλληλεπιδράσεις miRNA με γονίδια οι οποίες καταγράφονται σε κείμενο σε σχετικές δημοσιεύσεις. Αυτό ήταν το κίνητρο για να διερευνήσουμε τις δυνατότητες που υπάρχουν για αυτόματη αναγνώριση αλληλεπιδράσεων miRNA με γονίδια στο κείμενο σχετικών δημοσιεύσεων. Τα αποτελέσματα πρώιμης πειραματικής αποτίμησης μας δημιουργούν αισιοδοξία σχετικά με την παροχή ικανοποιητικών προτάσεων προς του επιμελητές του DIANA TarBase. Να σημειωθεί ότι η αξία των προαναφερθέντων εργαλείων έχει αποδειχθεί από τη χρήση τους μέσα στα προηγούμενα χρόνια. Περίπου 500 διαφορετικοί ερευνητές τα χρησιμοποιούν καθημερινά, με περισσότερους από 100 από αυτούς να είναι εγγεγραμμένοι και να εκμεταλλεύονται τις εξατομικευμένες λειτουργίες που τους παρέχουμε.

Τέλος, παρουσιάσαμε τη *Hitmap* μια προσέγγιση ευρετηρίου που υποστηρίζει αποδοτική στοίχιση για αναγνώσματα DNA μεγάλου μήκους και για σχετικά μεγάλα κατώφλια λαθών. Η *Hitmap* καλύπτει το κενό στις προσεγγίσεις στοίχισης καθώς

υπερνικά τις προσεγγίσεις αιχμής στην περίπτωση των αναγνωσμάτων DNA μεγάλου μήκους ενώ οι επιδόσεις του για στοίχιση μικρών αναγνωσμάτων παραμένει παρόμοια με αυτή των καλύτερων αλγορίθμων στοίχισης για μικρά αναγνώσματα.

6.2 Μελλοντικές εργασίες

Κατά την εργασία στην παρούσα διατριβή, αναγνωρίσαμε τα ακόλουθα ενδιαφέροντα θέματα τα οποία προτείνουμε για μελλοντική εργασία.

- Σχετικά με τη στοίχιση αναγνωσμάτων DNA μεγάλου μήκους, πιστεύουμε ότι είναι εύκολο να μετασχηματίσουμε το ευρετήριο *Hitmap* προκειμένου να μπορεί να χρησιμοποιηθεί από συστήματα που βασίζονται στο Νέφος. Μια τέτοια εξέλιξη θα καταστήσει εφικτό να δημιουργηθούν εργαλεία Ιστού για στοίχιση αναγνωσμάτων DNA στο ανθρώπινο γονιδίωμα σε πραγματικό χρόνο. Πολλά εργαστήρια που έχουν ανάγκη για στοίχιση αλλά δεν έχουν την απαραίτητη χρηματοδότηση για να αγοράσουν ισχυρούς υπολογιστές θα μπορούσαν να επωφεληθούν από αυτά τα εργαλεία. Επιπλέον, αφού τα αναγνώσματα DNA παράγονται σε δεσμίδες θα ήταν ενδιαφέρον να εφαρμοστεί προεπεξεργασία πάνω σε αυτές τις δεσμίδες ούτως ώστε να γίνει εκμετάλλευση των επικαλύψεων που έχουν τα αναγνώσματα κατά τη διάρκεια της στοίχισής τους στο γονιδίωμα αναφοράς.

- Μια άλλη κατεύθυνση που σχετίζεται με το ταίριασμα ακολουθιών σε δεδομένα από τις βιοεπιστήμες είναι η αποδοτική αποθήκευση πολλαπλών γονιδιωματικών ακολουθιών με τρόπου που να εκμεταλλεύεται τις μεταξύ τους επικαλύψεις και να διευκολύνει τη γρήγορη στοίχιση αναγνωσμάτων DNA ταυτόχρονα σε όλες αυτές τις ακολουθίες γονιδιώματος. Υπάρχουν πολλά σενάρια (πχ εξελικτική ανάλυση ειδών, εξατομικευμένη ιατρική, κτλ) όπου τα αναγνώσματα DNA χρειάζεται να στοιχιστούν όχι μόνο σε ένα γονιδίωμα αλλά σε ένα σύνολο από γονιδιώματα. Αυτά τα γονιδιώματα μπορούν να είναι είτε γονιδιώματα αναφοράς από διαφορετικά είδη ή γονιδιώματα ατόμων του ίδιου είδους. Η αποθήκευση κάθε γονιδιώματος ξεχωριστά έχει ως αποτέλεσμα μη αποδοτική χρήση του δίσκου επειδή αυτές οι ακολουθίες περιέχουν πολλές επικαλύψεις. Έτσι, αποδοτικές τεχνικές αποθήκευσης που εκμεταλλεύονται τις προηγούμενες επικαλύψεις πρέπει να χρησιμοποιηθούν. Παρόλο που υπάρχουν κάποιες τεχνικές συμπίεσης γονιδιώματος οι περισσότερες από αυτές επικεντρώνονται στη βελτιστοποίηση της χρήσης του αποθηκευτικού χώρου. Όμως, ένα άλλο σημαντικό ζήτημα είναι η ευρετηρίαση των αποθηκευμένων γονιδιωμάτων προκειμένου να προσφέρεται αποδοτική στοίχιση ενός δοθέντος αναγνώσματος μέσα σε όλα αυτά τα γονιδιώματα. Η διερεύνηση αυτού του προβλήματος μπορεί να παρουσιάζει ενδιαφέρον.

- Κατά τη διάρκεια της ανάπτυξης του DIANA mirPub διαπιστώσαμε ότι οι υπάρχουσες τεχνικές για ιεράρχηση επιστημονικών δημοσιεύσεων εμφανίζουν πολλά προβλήματα. Για παράδειγμα, πολλές από αυτές δίνουν πλεονέκτημα σε παλιές δημοσιεύσεις επειδή υπάρχουν πολλές άλλες δημοσιεύσεις που αναφέρονται σε αυτές. Παρόλα αυτά, αυτό δεν είναι δίκαιο για τις νέες δημοσιεύσεις καθώς εκείνες από αυτές που είναι σημαντικές πρόκειται να προσελκύσουν τις περισσότερες από τις αναφορές τους στο μέλλον. Η πρόβλεψη της σημασίας των νέων δημοσιεύσεων είναι ένα ενδιαφέρον ανοικτό πρόβλημα που η επίλυσή του μπορεί να βοηθήσει προς αυτή την κατεύθυνση.

- Η δουλειά που πραγματοποιήθηκε στο πεδίο της αυτόματης εξαγωγής γνώσης από δημοσιεύσεις για miRNA (Κεφάλαιο 4.4.3) οφείλει να επεκταθεί και να αξιολογηθεί χρησιμοποιώντας περισσότερες κρίσεις από ειδικούς του πεδίου για τις προτεινόμενες αλληλεπιδράσεις miRNA-γονιδίων που προσφέρονται από την προσέγγισή μας. Επιπλέον, μια προσέγγιση που εκμεταλλεύεται τις αλληλεπιδράσεις που καταγράφονται σε υποστηρικτικούς πίνακες και σχήματα πρέπει να παρέχεται.

Συμπερασματικά, πιστεύουμε ότι υπάρχει μια πληθώρα από ενδιαφέρονται και νέα θέματα που σχετίζονται με την αποδοτική διαχείριση δεδομένων από τις βιοεπιστήμες. Ελπίζουμε η παρούσα διατριβή να αποτελέσει εφαλτήριο για επιπλέον έρευνα στο συγκεκριμένο πεδίο.

Bibliography

[1] M. I. Abouelhoda, S. Kurtz, and E. Ohlebusch. Replacing suffix trees with enhanced suffix arrays. *J. Discrete Algorithms*, 2(1):53–86, 2004.

[2] P. Alexiou, M. Maragkakis, G. L. Papadopoulos, M. Reczko, and A. G. Hatzigeorgiou. Lost in translation: an assessment and perspective for computational microrna target identification. *Bioinformatics*, 25(23):3049–3055, 2009.

[3] P. Alexiou, T. Vergoulis, M. Gleditzsch, G. Prekas, T. Dalamagas, M. Megraw, I. Grosse, T. Sellis, and A. G. Hatzigeorgiou. mirgen 2.0: a database of microrna genomic information and regulation. *Nucleic Acids Res*, 41(suppl 1):D137–D141, 2010.

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403 – 410, 1990.

[5] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, Sept. 1997.

[6] C. Backes, A. Keller, J. Kuentzer, B. Kneissl, N. Comtesse, Y. A. Elnakady, R. Muller, E. Meese, and H. P. Lenhof. GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Res.*, 35(Web Server issue):W186–192, Jul 2007.

[7] C. Backes, E. Meese, H. P. Lenhof, and A. Keller. A dictionary on microRNAs and their putative target pathways. *Nucleic Acids Res.*, 38(13):4476–4486, Jul 2010.

[8] R. A. Baeza-Yates and G. Navarro. Faster approximate string matching. *Algorithmica*, 23(2):127–158, 1999.

[9] R. A. Baeza-Yates and G. Navarro. New and faster filters for multiple approximate string matching. *Random Structure and Algorithms*, 20(1):23 – 49, 2002.

[10] M. P. Ben Langmead, Cole Trapnell and S. L. Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25.

[11] R. S. Boyer and J. S. Moore. A fast string searching algorithm. *Commun. ACM*, 20(10):762–772, 1977.

[12] J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen. Principles of microrna-target recognition. *PLOS Biology*, 3(3):e85, 2005.

[13] M. Burrows and D. J. Wheeler. A block-sorting lossless data compression algorithm. Technical report, 1994.

[14] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, T. C. Mike Burrows, A. Fikes, and R. E. Gruber. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems*, 26, 2008.

[15] W. I. Chang and T. G. Marr. Approximate string matching and local similarity. In *Combinatorial Pattern Matching (CPM)*, volume 807 of *Lecture Notes in Computer Science (LNCS)*, pages 259–273. Springer, 1994.

[16] S. W. Chi, J. B. Zang, A. Mele, and R. B. Darnell. Argonaute hits-clip decodes microrna-mrna interaction maps. *Nature*, 460(7254):479–486, 2009.

[17] D. L. Corcoran, K. V. Pandit, B. Gordon, A. Bhattacharjee, N. Kaminski, and P. V. Benos. Features of mammalian microrna promoters emerge from polymerase ii chromatin immunoprecipitation data. *PLoS One*, 4(4):e5279, 2009.

[18] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51:107–113, 2008.

[19] J. G. Doench and P. A. Sharp. Specificity of microrna target selection in translational repression. *Genes Dev*, 18(5):504–511, 2004.

[20] H. Dweep, C. Sticht, P. Pandey, and N. Gretz. mirwalk - database: Prediction of possible mirna binding sites by "walking" the genes of three genomes. *J Biomed Inform*, 44(5):839–847, 2011.

[21] M. Fabbri, C. M. Croce, and G. A. Calin. Micrornas in the ontogeny of leukemias and lymphomas. *Leukemia and Lymphoma*, 50(2):160–170, 2009.

[22] K. Fredriksson and G. Navarro. Average-optimal single and multiple approximate string matching. *ACM Journal of Experimental Algorithms*, 9, 2004.

[23] A. L. Gartel and E. S. Kandel. mirnas: Little known mediators of oncogenesis. *Seminars in Cancer Biology*, 18(2):103 – 110, 2008. Postgenetic Oncology - MicroRNA and Cell Proliferation.

[24] V. A. Gennarino, M. Sardiello, R. Avellino, N. Meola, V. Maselli, S. Anand, L. Cutillo, A. Ballabio, and S. Banfi. Microrna target prediction by expression analysis of host genes. *Genome Res*, 19:481–490, 2009.

[25] A. Grimson, K. K.-H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel. Microrna targeting specificity in mammals: Determinants beyond seed pairing. *Mol Cell*, 27(1):91–105, 2007.

[26] D. Gusfield. *Algorithms on strings, trees, and sequences*. Cambridge University Press, 1999.

[27] M. Hadjieleftheriou and D. Srivastava. Approximate string processing. *Foundations and Trends in Databases*, 2(4):267–402, 2011.

[28] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. A. Jr, A.-C. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, and T. Tuschl. Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip. *Cell*, 141(1):129–141, 2009.

[29] J. R. Heng Li and R. Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18:1851 – 1858, 2008.

[30] J. B. Hsu, C. M. Chiu, S. D. Hsu, W. Y. Huang, C. H. Chien, T. Y. Lee, and H. D. Huang. miRTar: an integrated system for identifying miRNA-target interactions in human. *BMC Bioinformatics*, 12:300, 2011.

[31] S.-D. Hsu, C.-H. Chu, A.-P. Tsou, S.-J. Chen, H.-C. Chen, P. W.-C. Hsu1, Y.-H. Wong, Y.-H. Chen, G.-H. Chen, and H.-D. Huang. mirnamap 2.0: genomic maps of microrna s in metazoan genomes. *Nucleic Acids Res*, 36(suppl 1):D165–D169, 2008.

[32] S.-D. Hsu, F.-M. Lin, W.-Y. Wu, C. Liang, W.-C. Huang, W.-L. Chan, W.-T. Tsai, G.-Z. Chen, C.-J. Lee, C.-M. Chiu1, C.-H. Chien, M.-C. Wu, C.-Y. Huang, A.-P. Tsou, and H.-D. Huang. mirtarbase: a database curates experimentally validated microrna-target interactions. *Nucleic Acids Res*, 39(suppl 1):D163–D169, 2011.

[33] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res*, 34(suppl 2):W729–W732, 2006.

[34] H. Hyyrö and G. Navarro. Faster bit-parallel approximate string matching. In *Combinatorial Pattern Matching (CPM)*, volume 2373 of *Lecture Notes in Computer Science (LNCS)*, pages 203–224. Springer, 2002.

[35] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang, and Y. Liu. mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic Acids Res*, 37(suppl 1):D98–D104, 2009.

[36] P. Jokinen, J. Tarhio, and E. Ukkonen. A comparison of approximate string matching algorithms. *Softw., Pract. Exper.*, 26(12):1439–1458, 1996.

[37] M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, 2000.

[38] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, 40(D1):D109–D114, 2012.

[39] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, 40(Database issue):D109–114, Jan 2012.

[40] I. Kanellos, T. Vergoulis, D. Sacharidis, T. Dalamagas, A. G. Hatzigeorgiou, S. Sartzetakis, and T. Sellis. Mr-microt: A mapreduce-based microrna target prediction method. In *Proceedings of the International Conference on Scientific and Statistical Database Management (SSDBM)*, 2014 (to appear).

[41] A. E. Kel, E. Gößling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender. Match^{TM}: a tool for searching transcription factor binding sites in dna sequences. *Nucleic Acids Res*, 31(13):3576–3579, 2003.

[42] Y. J. Kim, A. Boyd, B. D. Athey, and J. M. Patel. miblast: scalable evaluation of a batch of nucleotide sequence queries with blast. *Nucleic Acids Research*, 33(13):4335–44, 2005.

[43] M. Kiriakidou, P. T. Nelson, A. Kouranov, P. Fitziev, C. Bouyioukos, Z. Mourelatos, and A. G. Hatzigeorgiou. A combined computational-experimental approach predicts human microrna targets. *Genes Dev*, 18(10):1165–1178, 2004.

[44] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *ACL*, pages 423–430, 2003.

[45] I. Korf and W. Gish. Mpblast: Improved blast performance with multiplexed queries. *Bioinformatics*, 16(11):1052–1053, November 2000.

[46] E. Koukis and P. Louridas. ∼okeanos iaas. *Proceedings of Science*, 2012.

[47] A. Kowarsch, M. Preusse, C. Marr, and F. J. Theis. miTALOS: analyzing the tissue-specific regulation of signaling pathways by human and mouse microRNAs. *RNA*, 17(5):809–819, May 2011.

[48] P. Landgraf, M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, S. Pfeffer, A. Rice, A. O. Kamphorst, M. Landthaler, C. Lin, N. D. Socci, L. Hermida, V. Fulci, S. Chiaretti, R. Foa, J. Schliwka, U. Fuchs, A. Novosel, R.-U. Muller, B. Schermer, U. Bissels, J. Inman, Q. Phan, M. Chien, D. B. Weir, R. Choksi, G. D. Vita, D. Frezzetti, H.-I. Trompeter, V. Hornung, G. Teng, G. Hartmann, M. Palkovits, R. D. Lauro, P. Wernet, G. Macino, C. E. Rogler, J. W. Nagle, J. Ju, F. N. Papavasiliou, T. Benzing, P. Lichter, W. Tam, M. J. Brownstein, A. Bosio, A. Borkhardt, J. J. Russo, C. Sander, M. Zavolan, and T. Tuschl. A mammalian microrna expression atlas based on small rna library sequencing. *Cell*, 129(7):1401–1414, 2007.

[49] M. V. G. Latronico, D. Catalucci, and G. Condorelli. Microrna and cardiac pathologies. *Physiological Genomics*, 34(3):239–242, 2008.

[50] V. Levenshtein. Binary codes capable of correcting spurious insertions and deletions of ones. *Probl. Inf. Transmission*, 1:8 – 17, 1965.

[51] V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.*, 10(8):707 – 710, 1966. Original in Russian in Dokl. Akad. Nauk SSSR 163(4): 845-848, 1965.

[52] C. Li, J. Lu, and Y. Lu. Efficient merging and filtering algorithms for approximate string searches. In *ICDE*, pages 257–266, 2008.

[53] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[54] Y. Li, J. M. Patel, and A. Terrell. Wham: A high-throughput sequence alignment method. *ACM Trans. Database Syst.*, 37(4):28, 2012.

[55] P. S. Linsley, J. Schelter, J. Burchard, M. Kibukawa, M. M. Martin, S. R. Bartz, J. M. Johnson, J. M. Cummins, C. K. Raymond, H. Dai, N. Chau, M. Cleary, A. L. Jackson, M. Carleton, and L. Lim. Transcripts targeted by the microrna-16 family cooperatively regulate cell cycle progression. *Mol Cell Biol*, 27(6):2240–2252, 2007.

[56] D. J. Lipman and W. R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435 – 1441, March 1985.

[57] U. Manber and G. Myers. Suffix arrays: A new method for on-line string searches. In *SODA*, pages 319–327, 1990.

[58] M. Maragkakis, P. Alexiou, G. L. Papadopoulos, M. Reczko, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, V. A. Simossis, P. Sethupathy, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas, and A. G. Hatzigeorgiou. Accurate microrna target prediction correlates with protein repression levels. 10:295, 2009.

[59] M. Maragkakis, M. Reczko, V. A. Simossis, P. Alexiou, G. L. Papadopoulos, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas, and A. G. Hatzigeorgiou. Diana-microt web server: elucidating microrna functions through target prediction. *Nucleic Acids Res*, 37(suppl 2):W273–W276, 2009.

[60] M. Maragkakis, T. Vergoulis, P. Alexiou, M. Reczko, K. Plomaritou, M. Gousis, K. Kourtis, N. Koziris, T. Dalamagas, and A. G. Hatzigeorgiou. Diana-microt web server upgrade supports fly and worm mirna target prediction and bibliographic mirna to disease association. *Nucleic Acids Res*, 39(suppl 2):W145–W148, 2011.

[61] A. Marson, S. S. Levine, M. F. Cole, G. M. Frampton, T. Brambrink, S. Johnstone, M. G. Guenther, W. K. Johnston, M. Wernig, J. Newman, J. M. Calabrese, L. M. Dennis, T. L. Volkert, S. Gupta, J. Love, N. Hannett, P. A. Sharp, D. P. Bartel, R. Jaenisch, and R. A. Youngemail. Connecting microrna genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, 134(3):521–533, 2008.

[62] M. Megraw, V. Baev, V. Rusinov, S. Jensen, K. Kalantidis, and A. G. Hatzigeorgiou. Microrna promoter element discovery in arabidopsis. *RNA*, 12:1612–1619, 2006.

[63] M. Megraw, P. Sethupathy, B. Corda, and A. G. Hatzigeorgiou. mirgen: a database for the study of animal microrna genomic organization and function. *Nucleic Acids Res*, 35(suppl 1):D149–D155, 2006.

[64] G. Minnen, J. A. Carroll, and D. Pearce. Applied morphological processing of english. *Natural Language Engineering*, 7(3):207–223, 2001.

[65] miRSel: Automated extraction of associations between microRNAs and genes from the biomedical literature.

[66] R. Muth and U. Mamber. Approximate multiple string search. In *Combinatorial Pattern Matching (CPM)*, volume 1075 of *Lecture Notes in Computer Science (LNCS)*, pages 75–86. Springer, 1996.

[67] G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys (CSUR)*, 33(1):31 – 88, Mar. 2001.

[68] G. Navarro and R. A. Baeza-Yates. A practical q -gram index for text retrieval allowing errors. *CLEI Electron. J.*, 1(2), 1998.

[69] G. Navarro, R. A. Baeza-Yates, E. Sutinen, and J. Tarhio. Indexing methods for approximate string matching. *IEEE Data Engineering Bulletin (DEBU)*, 24(4):19 – 27, Dec. 2001.

[70] G. Navarro and K. Fredriksson. Average complexity of exact and approximate multiple string matching. *Theor. Comput. Sci.*, 321(2-3):283 – 290, 2004.

[71] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–53, Mar. 1970.

[72] F. Ozsolak, L. L. Poling, Z. Wang, H. Liu, X. S. Liu, R. G. Roeder, X. Zhang, J. S. Song, and D. E. Fisher. Chromatin structure analyses identify mirna promoters. *Genes Dev*, 22:3172–3183, 2008.

[73] G. L. Papadopoulos, P. Alexiou, M. Maragkakis, M. Reczko, and A. G. Hatzigeorgiou. DIANA-mirPath: Integrating human and mouse microRNAs in pathways. *Bioinformatics*, 25(15):1991–1993, Aug 2009.

[74] G. L. Papadopoulos, M. Reczko, V. A. Simossis, P. Sethupathy, and A. G. Hatzigeorgiou. The database of experimentally supported targets: a functional update of tarbase. *Nucleic Acids Res*, 37(suppl 1):D155–D158, 2009.

[75] P. Papapetrou, V. Athitsos, G. Kollios, and D. Gunopulos. Reference-based alignment in large sequence databases. *PVLDB*, 2(1):205–216, 2009.

[76] M. D. Paraskevopoulou, G. Georgakilas, N. Kostoulas, I. S. Vlachos, T. Vergoulis, M. Reczko, C. Filippidis, T. Dalamagas, and A. G. Hatzigeorgiou. Diana-microt web server v5.0: service integration into mirna functional analysis workflows. *Nucleic Acids Res*, 41(W1):W169–W173, 2013.

[77] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA*, 85(8):2444 – 2448, April 1988.

[78] M. Reczko, M. Maragkakis, P. Alexiou, I. Grosse, and A. G. Hatzigeorgiou. Functional microrna targets in protein coding sequences. *Bioinformatics*, 28(6):771–776, 2012.

[79] M. Rehmsmeier, P. Steffen, M. Höchsmann, and R. Giegerich. Fast and effective prediction of microrna/target duplexes. *RNA*, 10:1507–1517, 2004.

[80] K. K. Ruiqiang Li, Yingrui Li and J. Wang. Soap: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713 – 714.

[81] Y. L. T.-W. L. S.-M. Y. K. K. Ruiqiang Li, Chang Yu and J. Wang. Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967.

[82] D. Sankoff and J. Kruskal. *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison.* Addison-Wesley,Reading,MA, 1983.

[83] M. Selbach, B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky. Widespread changes in protein synthesis induced by microrna. *Nature*, 455(7209):58–63, 2008.

[84] P. H. Sellers. An algorithm for the distance between two finite sequences. *J. Combin. Theory Ser. A*, 16:253–258, 1974.

[85] P. H. Sellers. The theory and computation of evolutionary distances: Pattern recognition. *Journal of Algorithms*, 1(4):359 – 373, 1980.

[86] P. Sethupathy, B. Corda, and A. G. Hatzigeorgiou. Tarbase: A comprehensive database of experimentally supported animal microrna targets. *RNA*, 12:192–197, 2006.

[87] P. Sethupathy, M. Megraw, and A. G. Hatzigeorgiou. A guide through present computational approaches for the identification of mammalian microrna targets. *Nat Methods*, 3:881–886, 2006.

[88] E. M. Smigielski, K. Sirotkin, M. Ward, and S. T. Sherry. dbsnp: a database of single nucleotide polymorphisms. *Nucleic Acids Res*, 28(1):352–355, 2000.

[89] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–7, Mar. 1981.

[90] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–7, Mar. 1981.

[91] L. Smith1, L. K. Tanabe, R. J. nee Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. M. Friedrich, K. Ganchev, M. Torii, H. Liu, B. Haddow, C. A. Struble, R. J. Povinelli, A. Vlachos, W. A. Baumgartner, L. Hunter, B. Carpenter, R. T.-H. Tsai, H.-J. Dai, F. Liu, Y. Chen, C. Sun, S. Katrenko, P. Adriaans, C. Blaschke, R. Torres, M. Neves, P. Nakov, A. Divoli, M. Maña-López, J. Mata, and W. J. Wilbur. Overview of biocreative ii gene mention recognition. *Genome Biol*, 9(suppl 2):S2, 2008.

[92] Y. Tay, J. Zhang, A. M. Thomson, B. Lim, and I. Rigoutsos. Micrornas to nanog, oct4 and sox2 coding regions modulate embryonic stem cell differentiation. *Nature*, 455(7216):1124–1128, 2008.

[93] D. W. Thomson, C. P. Bracken, and G. J. Goodall. Experimental strategies for microrna target identification. *Nucleic Acids Res*, 39(16):6845–6853, 2011.

[94] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL*, 2003.

[95] R.-M. Tsoupidi. Automatic extraction of microrna-gene interactions from scientific publications in life sciences. *diploma thesis at NTUA*, 2014.

[96] E. Ukkonen. Finding approximate patterns in strings. *Journal of Algorithms*, 6:132–137, 1985.

[97] T. Vergoulis, M. Alexakis, T. Dalamagas, M. Maragkakis, A. G. Hatzigeorgiou, and T. Sellis. Tarcloud: A cloud-based platform to support mirna target prediction. In *Proceedings of the International Conference on Scientific and Statistical Database Management (SSDBM)*, pages 628–633, 2012.

[98] T. Vergoulis, T. Dalamagas, D. Sacharidis, and T. Sellis. Approximate regional sequence matching for genomic databases. *VLDB J.*, 21(6):779–795, 2012.

[99] T. Vergoulis, I. S. Vlachos, P. Alexiou, G. Georgakilas, M. Maragkakis, M. Reczko, S. Gerangelos, N. Koziris, T. Dalamagas, and A. G. Hatzigeorgiou. Tarbase 6.0: capturing the exponential growth of mirna targets with experimental support. *Nucleic Acids Res*, 40(D1):D222–D229, 2012.

[100] I. S. Vlachos, N. Kostoulas, T. Vergoulis, G. Georgakilas, M. Reczko, M. Maragkakis, M. D. Paraskevopoulou, K. Prionidis, T. Dalamagas, and A. G. Hatzigeorgiou. Diana mirpath v.2.0: investigating the combinatorial effect of microrna in pathways. *Nucleic Acids Res*, 40(W1):W498–W504, 2012.

[101] X. Wang and X. Wang. Systematic identification of microrna functions by combining target prediction and expression profiling. *Nucleic Acids Res*, 34(5):1646–1652, 2006.

[102] P. G. Weiner. Linear pattern matching algorithms. In *14th Annual IEEE Symposium on Switching and Automata Theory*, pages 1–11, 1973.

[103] T. D. Wu and S. Nacu. Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 24(5):873 – 881.

[104] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li. mirecords: an integrated resource for microrna-target interactions. *Nucleic Acids Res*, 37(suppl 1):D105–D110, 2009.

[105] B. Xie, Q. Ding, H. Han, and D. Wu. miRCancer: a microRNA-cancer association database constructed by text mining on literature. *Bioinformatics*, 29(5):638–644, Mar 2013.

[106] J.-H. Yang, J.-H. Li, P. Shao, H. Zhou, Y.-Q. Chen, and L.-H. Qu. starbase: a database for exploring microrna-mrna interaction maps from argonaute clip-seq and degradome-seq data. *Nucleic Acids Res*, 39(suppl 1):D202–D209, 2010.

[107] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller. A greedy algorithm for aligning dna sequences. *J. Comput. Biol.*, 7(1-2):203–214, 2000.

[108] D. G. Zisoulis, M. T. Lovci, M. L. Wilbert, K. R. Hutt, T. Y. Liang, A. E. Pasquinelli, and G. W. Yeo. Comprehensive discovery of endogenous argonaute binding sites in caenorhabditis elegans. *Nat Struct Mol Biol*, 17:173–179, 2010.

Παράρτημα Α΄

Γλωσσάρι

Ελληνικός όρος	**Αγγλικός όρος**
ακολουθία	sequence
ακολουθία δεδομένων	data sequence
ακολουθία-ερώτημα	query sequence
αλυσίδα επιθεμάτων	suffix chain
αλφάβητο	alphabet
αλφαριθμητικό	string
αναδυόμενο παράθυρο	pop-up window
αναλυτής διατηρησιμότητας	conservation profiler
αναλυτής	parser
ανεστραμμένη λίστα	inverted list
ανεστραμμένο ευρετήριο	inverted index
αποκοπή	cut-off
αποκοπή καταλήξεων	stemming
αποτύπωμα μνήμης	memory footprint
βάση δεδομένων ακολουθιών	sequence database
γενικευμένα δένδρα επιθεμάτων	generalised suffix trees
γονίδιο	gene
γονιδίωμα	genome
δεξαμενή	pool
δεσμίδα	batch
διάνυσμα χαρακτηριστικών	feature vector
διάτμηση	partition
διαφορικά εκφρασμένος	differential expressed
διεπαφή	interface
διοχέτευση	pipeline
διπλά-τοπική στοίχιση	double-local alignment
δοκιμασίες γονιδίων ανταπόκρισης	reporter gene assays
δυαδική μέγιστη εντροπία	binary maximum entropy
δυαδικό ψηφίο	bit
δυαδικό ψηφίο καλύμματος	cover bit
δυναμικός προγραμματισμός	dynamic programming
εγγραφή	record
ειδικά για κάθε γονίδιο πειράματα	gene specific experiments
εξισορροπητής	balancer

Ελληνικός όρος	**Αγγλικός όρος**
εξόρυξη κειμένου	text mining
επίθεμα	suffix
ευρετηρίαση	indexing
ευρετήριο	index
ευριστική αποκοπής	cur-off heuristic
ιεράρχηση	ranking
Ιστός	Web
κάλυμμα	cover
κατά προς. ταίριασμα περιοχών ακολουθίας	approx. regional sequence matching (ARSM)
καταπιεσμένο γονίδιο	suppressed gene
καταχώρηση	entry
κατηγοριοποιητής	classifier
κατώφλι	threshold
κατώφλι στοίχισης	alignment threshold
μερική διάταξη	partial order
μετρητής δυαδικός ψηφίων	bit-counter
μετρική	metric
μητρώο ζύγισης θέσης	position weight matrix
Νέφος	Cloud
ολική διάταξη	total order
ολική στοίχιση	global alignment
Παγκόσμιος Ιστός	World Wide Web
παρουσία	occurrence
πολυμορφισμός ενός νουκλεοτιδίου	single nucleotide polymorphism (SNP)
πρόβλεψη στόχων miRNA	miRNA target prediction
πρόθεμα	prefix
πρότυπο σχεδιασμού	design pattern
ρυθμαπόδοση	throughput
στόχος miRNA	miRNA target
σύμβολο	symbol
σύννεφο ετικετών	tag cloud
ταίριασμα ακολουθιών	string matching
τάξη μεγέθους	order of magnitude
τοπική στοίχιση	local alignment
τμήμα	part
υπακολουθία	subsequence
υπερακολουθία	supersequence
υπερεκφρασμένος	overexpressed
υπόλοιπο	modulo
φουρκέτα	hairpin
χτύπημα	hit
ώριμο miRNA	mature miRNA
bitset t-παρουσιών	t-occurrences bitset
miRNA φουρκέτα	hairpin miRNA

Παράρτημα Β′

Συνοπτικό Βιογραφικό Σημείωμα

Β′.1 Προσωπικά Στοιχεία

- Όνομα: Αθανάσιος Βεργούλης
- Ειδικότητα: Δρ. Μηχανικός Η/Υ & Πληροφορικής
- Τηλέφωνο: +30 2106875423
- Διεύθυνση: Ινστιτούτο Πληροφοριακών Συστημάτων, Ερευνητικό Κέντρο 'Αθηνά', Αρτέμιδος 6 & Επιδαύρου, Παράδεισος Αμαρουσίου 15125
- Διεύθυνση ηλ. ταχυδρομείου: vergoulis@imis.athena-innovation.gr
- Προσωπική ιστοσελίδα: http://www.dblab.ece.ntua.gr/ vergoulis

Β′.2 Σπουδές, Επαγγελματικές Άδειες και Πιστοποιήσεις

- 2007 - 2014: Διδακτορικό στην Επιστήμη Υπολογιστών ('Διαχείριση Δεδομένων στις Βιοεπιστήμες') στο Εθνικό Μετσόβιο Πολυτεχνείο. Επιβλέπων: Καθ. Τιμολέων Σελλής
- 2002 - 2007: Δίπλωμα από το τμήμα Μηχανικών Η/Υ και Πληροφορικής του Πανεπιστημίου Πατρών. Βαθμός: 8,22/10
- Άδεια άσκησης επαγγέλματος Μηχανικού Η/Υ και Πληροφορικής από το Τεχνικό Επιμελητήριο Ελλάδος (ΤΕΕ)
- Απολυτήριο Λυκείου από το 1ο Ενιαίο Λύκειο Ναυπάκτου
- Ξένες γλώσσες: Αγγλικά (πιστοποίηση επιπέδου C1 από το University of Cambridge)

Β′.3 Επαγγελματικές συνεργασίες

- 10/2012 - 8/2014: Διδακτορικός ερευνητικός συνεργάτης στο Ινστιτούτο Πληροφοριακών Συστημάτων του Ε.Κ. 'Αθηνά'.
  - Συμμετοχή σε ερευνητικά έργα 'ΜΙΚΡΟΡΝΑ', 'LODGOV' και 'ΜΕΔΑ'.
- 5/2008 - 8/2012: Διδακτορικός ερευνητικός υπότροφος στο Ινστιτούτο Πληροφοριακών Συστημάτων του Ε.Κ. 'Αθηνά'.
  - Χρηματοδότηση από τακτικό προϋπολογισμό ΙΠΣΥ και από συμμετοχή σε έργα (π.χ. 'ΜΙΚΡΟΡΝΑ')
- 9/2008 - 11/2008: Καθηγητής Πληροφορικής στα εκπαιδευτήρια TechnoKids-TechnoPlus.
- 7/2006 - 8/2006: Πρακτική άσκηση στο Τμήμα Παροχής Τεχνικών Υπηρεσιών της 'Ελληνικά Πετρέλαια Α.Ε.'.
- 7/2005 - 8/2005, 8/2006 - 9/2006 και 7/2007 - 8/2007: Πρακτική άσκηση στον Τομέα Μεγάλων Πελατών της 'ΔΕΗ Α.Ε'.

Β′.4 Ερευνητικά και αναπτυξιακά έργα

- 12/2013 - σήμερα, ΜΕΔΑ
  - Θέμα έργου: 'Μεγάλα Δεδομένα'
  - Χρηματοδότηση έργου: Συγχρηματοδοτούμενο (πλαίσιο χρηματοδότησης: ΕΣΠΑ 2007-2013, επιχειρησιακό πρόγραμμα: ΚΡΗΠΙΣ)
  - Ρόλος στο έργο: Κύριος ερευνητής από ΙΠΣΥ σε θέματα επιστημονικών βάσεων δεδομένων, συντονισμός εργασιών, ανάπτυξη κώδικα
- 11/2012 - 10/2013, LODGOV
  - Θέμα έργου: 'Διακυβέρνηση δεδομένων στην εποχή του Ιστού Δεδομένων: Δημιουργία, διαχείριση, διατηρησιμότητα, κοινοχρηστία και προστασία πόρων στον Ιστό Δεδομένων'
  - Χρηματοδότηση έργου: Συγχρηματοδοτούμενο (πλαίσιο χρηματοδότησης: ΕΣΠΑ 2007-2013, επιχειρησιακό πρόγραμμα: Ανταγωνιστικότητα & Επιχειρηματικότητα, δράση: ΑΡΙΣΤΕΙΑ, αριθμός πρότασης έργου: 315)
  - Ρόλος στο έργο: Βιβλιογραφική επισκόπηση, συμμετοχή στη συγγραφή εξαμηνιαίων αναφορών
- 10/2010 - 12/2012, ΜΙΚΡΟΡΝΑ
  - Θέμα έργου: 'Εξερευνώντας τον ρόλο των μικρών RNA στις ασθένειες: μια υπολογιστική πειραματική προσέγγιση'
  - Χρηματοδότηση έργου: Συγχρηματοδοτούμενο (πλαίσιο χρηματοδότησης: ΕΣΠΑ 2007-2013, επιχειρησιακό πρόγραμμα: Εκπαίδευση & Δια Βίου Μάθηση, δράση: Συνεργασία Ι, κωδικός έργου: 09ΣΥΝ-13-1055)

  - Ρόλος στο έργο: Κύριος ερευνητής από ΙΠΣΥ, συντονισμός εργασιών, ανάπτυξη κώδικα

- 6/2011 - 6/2012, VENUS-C
  - Θέμα έργου: 'Building a highly-scalable and flexible Cloud infrastructure'
  - Χρηματοδότηση έργου: Ευρωπαϊκό (πλαίσιο χρηματοδότησης: ICT-EC's 7th Framework Programme 2007-2013, κωδικός έργου: RI-261565)
  - Σχόλιο: Συμμετοχή του ΙΠΣΥ στο έργο ως 'subcontracted pilot', η επιλογή έγινε μετά από σχετικό διαγωνισμό που διενεργήθηκε από τους διαχειριστές του έργου
  - Ρόλος στο έργο: Κύριος ερευνητής από ΙΠΣΥ, συμμετοχή στη συγγραφή πρότασης που έγινε δεκτή από το διαγωνισμό, συντονισμός εργασιών, ανάπτυξη κώδικα

Β΄.5 Εκπαιδευτική εμπειρία

- 2010-2011: Επικουρικό εκπαιδευτικό έργο στο Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Ε.Μ.Π. για τα ακόλουθα μαθήματα:
  - 'Προχωρημένα Θέματα Βάσεων Δεδομένων'

- 2009-2010: Επικουρικό εκπαιδευτικό έργο στο Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Ε.Μ.Π. για τα ακόλουθα μαθήματα:
  - 'Προχωρημένα Θέματα Βάσεων Δεδομένων'

- 2008-2009: Επικουρικό εκπαιδευτικό έργο στο Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Ε.Μ.Π. για τα ακόλουθα μαθήματα:
  - 'Προγραμματισμός Ηλεκτρονικών Υπολογιστών (εργαστήριο)'
  - 'Βάσεις Δεδομένων'

- Συμμετοχή στην επίβλεψη των παρακάτω διπλωματικών εργασιών:
  - 'Σύστημα διαχείρισης διαχρονικών δεδομένων για γονίδια', Κώστας Ζαγγανάς, 2014
  - 'Σύστημα αυτόματης εξαγωγής αλληλεπιδράσεων μεταξύ μορίων microRNA και γονιδίων από επιστημονικές δημοσιεύσεις στις Βιοεπιστήμες', Ροδοθέα-Μυρσίνη Τσουπίδη, 2014
  - 'Αναζήτηση σε επιστημονικές βάσεις δεδομένων με βάση την ιστορική εξέλιξη των δεδομένων', Ηλίας Κανέλλος, 2012
  - 'Τεχνικές Σύστασης Όρων Σε Επιστημονικές Βάσεις Δεδομένων', Θεόφιλος Πέτσιος, 2011
  - 'DIANA 2.0: Προηγμένη εφαρμογή ιστού διαχείρισης δεδομένων Βιοεπιστημών', Γεωργία Φωτοπούλου, 2010
  - 'Μελέτη και Υλοποίηση Αλγορίθμων Βιολογικών Εφαρμογών στο MapReduce περιβάλλον', Δανάη Κούτρα, 2010

- 'Υλοποίηση κατασκευής δέντρου επιθεμάτων με Hadoop MapReduce', Αλέξανδρος Κωνσταντινάκησ-Κάρμης, 2010
- 'Προηγμένη Εφαρμογή Ιστού Διαχείρισης Δεδομένων Βιοεπιστημών', Γιώργος Πρέκας, 2010
- 'Τεχνικές κατασκευής δένδρων επιθεμάτων πολύ μεγάλου μεγέθους και χρήσης τους για γρήγορη αναζήτηση βιολογικών δεδομένων', Βασίλης Πολυχρονόπουλος, 2009

- Καθηγητής Πληροφορικής στα εκπαιδευτήρια TechnoKids-TechnoPlus κατά την περίοδο 15/9/2008 έως 15/11/2008.

Β΄.6 Συμμετοχή σε Επιτροπές

- Συμμετοχή στην Επιτροπή Προγράμματος του διεθνούς συνεδρίου CIKM 2011 (μέλος της επιτροπής προγράμματος αφισών).
- Εξωτερικός κριτής σε διεθνή περιοδικά και συνέδρια (BMC Bioinformatics, ACM SIGMOD, TODS, CIKM, KDD,ICDE,TPDL,VLDB)

Β΄.7 Ερευνητικά Ενδιαφέροντα

Επιστημονικές βάσεις δεδομένων, διαχείριση Δεδομένων για βιολογικά δεδομένα, αλγόριθμοι και ευρετήρια αναζήτησης ακολουθιακών δεδομένων, MapReduce και Υπολογισμός Νέφους (cloud computing).

Β΄.8 Τεχνικές Δεξιότητες

- Γλώσσες προγραμματισμού: C, C++, Java, Perl, Python
- Ανάπτυξη διαδικτυακών εφαρμογών: PHP, HTML, XML, Javascript, AJAX, Joomla
- Συστήματα βάσεων δεδομένων: MySQL, MS SQL Server, PostgreSQL
- Υπολογισμός Νέφους: Hadoop, Microsoft Azure

Β΄.9 Επιστημονικές Δημοσιεύσεις

Αναφορές: Με βάση το προφίλ στο Γοογλε Σχηολαρ 806

- I. Kanellos, T. Vergoulis, D. Sacharidis, T. Dalamagas, A. Hatzigeorgiou, S. Sartzetakis, T. Sellis. MR-microT: A MapReduce-based microRNA target prediction method. SSDBM 2014.
- Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, Filippidis C, Dalamagas T, Hatzigeorgiou AG. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W169-73.

- I. S. Vlachos, N. Kostoulas, T. Vergoulis, G. Georgakilas, M. Reczko, M. Maragkakis, M. D. Paraskevopoulou, K. Prionidis, T. Dalamagas, A. G. Hatzigeorgiou. DIANA miRPath v.2.0: investigating the combinatorial effect of microRNAs in pathways. Nucleic Acids Research 2012 (Web server issue)

- T. Vergoulis, M. Alexakis, T. Dalamagas, M. Maragkakis, A. G. Hatzigeorgiou, T. Sellis. TARCLOUD: A Cloud-based platform to support miRNA target prediction. SSDBM 2012

- T. Vergoulis, T. Dalamagas, D. Sacharidis, T. Sellis. Approximate Regional Sequence Matching for Genomic Databases. VLDB Journal 2012

- T. Vergoulis, I. Vlachos, P. Alexiou, G. Georgakilas, M. Maragkakis, M. Reczko, S. Gerangelos, N. Koziris, T. Dalamagas, AG Hatzigeorgiou. Tarbase 6.0: Capturing the Exponential Growth of miRNA Targets with Experimental Support. Nucleic Acids Research 2012 (Database issue)

- M. Maragkakis, T. Vergoulis, P. Alexiou, M. Reczko, K. Plomaritou, M. Gousis, K. Kourtis, N. Koziris, T. Dalamagas, AG. Hatzigeorgiou. DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and bibliographic miRNA to disease association. Nucleic Acids Research 2011 (Webserver issue)

- P. Alexiou, T. Vergoulis, M. Gleditzsch, G. Prekas, T. Dalamagas, M. Megraw, I. Grosse, T. Sellis, A.G. Hatzigeorgiou. miRGen 2.0: a database of microRNA genomic information and regulation. Nucleic Acids Research 2010, Vol. 38(Database issue)

- M. Maragkakis, P. Alexiou, G. Papadopoulos, M. Reczko, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis,V. Simossis, P. Sethupathy, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas, A. Hatzigeorgiou. Accurate microRNA target prediction correlates with protein repression levels. BMC Bioinformatics 2009, 10:295

- M. Maragkakis, M. Reczko, V. A. Simossis, P. Alexiou, G. L. Papadopoulos, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas, and A. G. Hatzigeorgiou. DIANA-microT web server: elucidating microRNA functions through target prediction. Nucleic Acids Research 2009, Vol. 37(Web Server issue)