ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ

ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΝΑΓΝΩΡΙΣΗ ΤΗΣ ΛΕΙΤΟΥΡΓΙΑΣ MICRORNA ΜΕ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΙΩΑΝΝΗΣ ΜΑΡΙΟΣ ΠΑΠΑΣΤΑΜΑΤΗΣ

Επιβλέποντες καθηγητές: Καθ. Άρτεμις Χατζηγεωργίου

Αν. Καθ. Γεράσιμος Ποταμιάνος

Βόλος, 2014

Περίληψη

Τα microRNA είναι μικρά ενδογενή μόρια ριβονουκλεϊκού οξέος (RNA), τα οποία αποτελούνται κατά μέσο όρο από 22 νουκλεοτίδια. Ο ρόλος τους είναι ιδιαίτερα σημαντικός, αφού αποτελούν μετα-μεταγραφικούς ρυθμιστές που ελέγχουν την γονιδιακή έκφραση. Αυτή η διαδικασία έχει αποτέλεσμα την μεταγραφική καταστολή (translational repression) ή / και σίγηση των γονιδίων (gene silencing). Είναι πολύ σημαντικό να εντοπιστούν οι στόχοι των miRNA. Γι' αυτό το λόγο έχουν αναπτυχθεί αρκετοί αλγόριθμοι πρόβλεψης γονιδίων στόχων miRNA.

Σκοπός της παρούσας διπλωματικής εργασίας είναι η παρουσίαση καινούργιων αλγορίθμων πρόβλεψης γονιδίων στόχων miRNA, καθώς και η σύγκρισή τους με τους παλαιότερους ab initio αλγορίθμους και η εξαγωγή συμπερασμάτων. Τα τελευταία χρόνια με την ανάπτυξη των πειραματικών δεδομένων εμφανίζονται όλο και περισσότεροι αλγόριθμοι μηχανικής μάθησης και αρκετοί από τους νέους αλγορίθμους που παρουσιάζονται στην εργασία αυτή εμπίπτουν σε αυτήν την κατηγορία.

Abstract

MicroRNAs are small endogenous ribonucleic acids (RNA) which consist of 22 nucleotides on average. Their role is particularly important in that they are post-transcriptional regulators which control the gene expression. This process/procedure results in the translational repression and / or the gene silencing. The identification of miRNA targets is crucial. This is the reason for the development of a large number of miRNA target prediction algorithms.

The present dissertation aims to present new miRNA target prediction algorithms, compare them with the older ab initio algorithms and reach/draw conclusions. Due to the recent development of experimental data, an increasing number of machine learning algorithms are produced; many of the new algorithms hereby presented belong to this category.

Περιεχόμενα

Κατάλογος Σχημάτων

1. Εισαγωγή

1.1 MicroRNA

Τα microRna είναι μικρά ενδογενή μη κωδικοποιητικά μόρια ριβονουκλεϊκού οξέως (RNA), τα οποία ρυθμίζουν την γονιδιακή έκφραση μέσω της μεταγραφικής καταστολής (translational repression) ή της σίγησης του αγγελιοφόρου RNA (mRNA). Κάθε miRNA αποτελείται από περίπου 19-24 νουκλεοτίδια και έχει προέλθει από την επεξεργασία ενός μεγαλύτερου transcript, το οποίο αναφέρεται ως primary transcript (pri-miRNA). Τα primary transcripts επεξεργάζονται στον πυρήνα του κυττάρου και μετατρέπονται στα pre-miRNA, τα οποία αποτελούνται από 70 νουκλεοτίδια. Όσον αφορά τα ζώα αυτή η επεξεργασία γίνεται με την χρήση ενός πρωτεϊνικού συμπλέγματος, που περιλαμβάνει τα ένζυμα Drosha και Pasha. Έπειτα, τα pre-miRNA μετατρέπονται σε ώριμα miRNA στο κυτταρόπλασμα, με την βοήθεια της ενδονουκλεάσης Dicer (Alexiou et al.,2009). Στη συνέχεια, με τη βοήθεια του συμπλόκου RISC (RNA-induced silencing complex) το ώριμο miRNA οδηγείται στο mRNA, όπου θα ρυθμίσει την γονιδιακή έκφραση.

Αν και τα miRNA ανακαλύφθηκαν το 1993 (Lee et al., 1993), μόλις το 2001 προτάθηκε ότι είναι ευρύτερα διαδεδομένα και άφθονα (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee και Ambros, 2001). Οι ιδιότητες των microRNAs παρουσιάζουν μεγάλες διαφορές ανάμεσα στους φυτικούς και ζωικούς οργανισμούς. Συγκεκριμένα τα φυτικά miRNA έχουν σχεδόν τέλειο ταίριασμα με τους mRNA στόχους. Αντίθετα, στα ζώα η συμπληρωματικότητα δεν είναι τέλεια. Τα ζωικά miRNA είναι ικανά να αναγνωρίζουν τους mRNA στόχους, χρησιμοποιώντας μια περιοχή 6-8 νουκλεοτιδίων (seed περιοχή) στο 5΄ τέλος του miRNA. Επιπλέον, μια ακόμα διαφορά έγκειται στην περιοχή όπου βρίσκονται οι περιοχές στόχοι στα mRNA. Όσον αφορά τους ζωικούς οργανισμούς, οι στόχοι βρίσκονται συνήθως στην 3΄ αμετάφραστη περιοχή (3΄UTR). Αντίθετα, στα φυτά η συγκεκριμένη περιοχή μπορεί να βρίσκεται στην 3΄UTR, αλλά συνήθως βρίσκεται στην κωδική περιοχή (CDS).

Περίπου 2200 miRNA γονίδια έχουν αναφερθεί ότι υπάρχουν στο γονιδίωμα των θηλαστικών, από τα οποία τα 1000 περίπου ανήκουν στο ανθρώπινο γονιδίωμα. Πολλές βιολογικές διεργασίες όπως η κυτταρική ανάπτυξη, η διαφοροποίηση και ο

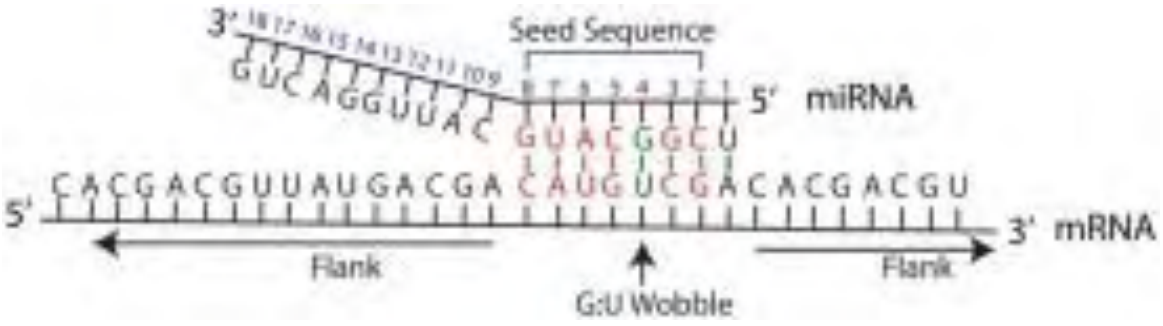
μεταβολισμός ρυθμίζονται από τα miRNA. Τα microRNA διαδραματίζουν ένα σημαντικό ρόλο σε αρκετές ανθρώπινες ασθένειες, από το έμφραγμα του μυοκαρδίου μέχρι αυτοάνοσα νοσήματα. Οι συνεχείς ανακαλύψεις και η πρόοδος σχετικά με τον ρόλο των miRNA στις ανθρώπινες ασθένειες, δίνει ελπίδα ότι θα διαδραματίσουν ένα σπουδαίο ρόλο στην διάγνωση και την θεραπεία διάφορων ασθενειών (Ardekani και Naeini, 2010).

Ο τεράστιος ρόλος των miRNA στην ανάπτυξη του καρκίνου δεν έχει ακόμα εκτιμηθεί στο βαθμό που θα έπρεπε. Έχει πλέον τεκμηριωθεί ότι η αυξορρύθμιση (upregulation) ή η μειορρύθμιση (down-regulation) ενός miRNA μπορεί να οδηγήσει σε διαφόρων ειδών καρκίνους (Naeini et al., 2009). Επίσης, τα επίπεδα έκφρασης ενός miRNA μπορούν να χρησιμεύσουν, ώστε να γίνει πρόγνωση για κάποιο είδος καρκίνου. Παραδείγματος χάρη, ψηλά επίπεδα miR-185 ή χαμηλά miR-133b μπορεί να συσχετίζονται με μετάσταση στον καρκίνο του παχέος εντέρου (Akçakaya et al.,2011).

Παρ' όλη την πολύ σημαντική συνεισφορά των high throughput τεχνικών στην επέκταση των επικυρωμένων βάσεων δεδομένων, που περιέχουν miRNA στόχους, η πλειονότητα των στόχων, ακόμα και για οργανισμούς που έχουν εξεταστεί εντατικά, όπως ο Homo Sapiens και ο Mus Musculus, είναι ακόμα άγνωστη. Διάφορες υπολογιστικοί μέθοδοι έχουν εμφανιστεί ώστε να αναγνωρίζουν πιθανούς στόχους. Οι υπολογιστικές μέθοδοι είναι η πιο πρακτική και αποτελεσματική επιλογή για να αναγνωριστούν άγνωστες αλληλεπιδράσεις miRNA-γονιδίων και να επιλεχθούν διακεκριμένοι υποψήφιοι για wet lab πειράματα. Βέβαια, πρέπει να έχουμε στο μυαλό μας ότι ακόμα και οι καλύτεροι αλγόριθμοι πρόβλεψης γονιδίων στόχων miRNA αποτυγχάνουν να αναγνωρίσουν ένα πολύ μεγάλο αριθμό αλληλεπιδράσεων γονιδίων-miRNA (Reczko et al.,2012). Επιπλέον, οι στόχοι που βρίσκονται δεν πρέπει να χρησιμοποιηθούν άμεσα, αλλά πρώτα να επικυρωθούν πειραματικά, καθώς υπάρχουν αρκετά λανθασμένα θετικά στα αποτελέσματα των αλγορίθμων (Vlachos και Hatzigeorgiou, 2013).

Τα πιο συνήθη χαρακτηριστικά που χρησιμοποιούν οι αλγόριθμοι πρόβλεψης γονιδίων στόχων miRNA για να αναγνωρίσουν τους πιθανούς στόχους είναι τέσσερα: το ταίριασμα στην κεντρική περιοχή (seed match), η εξελικτική διατήρηση

(conservation), η ελεύθερη ενέργεια του Gibbs (free energy) και η προσβασιμότητα (site accessibility). Σαν seed περιοχή ορίζονται τα πρώτα δύο με οχτώ νουκλεοτίδια αρχίζοντας από το 5΄ τέλος και μετρώντας προς το 3΄ τέλος. Όσον αφορά τα περισσότερα εργαλεία το seed ταίριασμα είναι ένα Watson-Crick ταίριασμα (δηλαδή αδενίνη με θυμίνη ή γουανίνη με κυτοσίνη) μεταξύ του miRNA και του στόχου του σε μια seed περιοχή. Υπάρχουν διάφοροι τύποι από seed ταιριάσματα που μπορούν να ληφθούν υπ' όψιν ανάλογα με τον κάθε αλγόριθμο. Στην εικόνα 1.1 παρατίθεται ένα διάγραμμα που δείχνει την seed περιοχή ενός miRNA, καθώς και την αλληλεπίδραση του με τον mRNA στόχο (W-C) ταιριάσματα, καθώς και ένα G-U.



Εικόνα 1.1: Αλληλεπίδραση miRNA με τον στόχο mRNA (Πηγή: Peterson et al., 2014)

Η εξελικτική διατήρηση (conservation) αναφέρεται στη συντήρηση ενός στόχου μεταξύ διαφόρων ειδών. Αφού οι θέσεις που δένουν τα microRNAs διατηρούνται μεταξύ διαφόρων ειδών, είναι πιθανότερο να είναι βιολογικά λειτουργικές. Η πρόβλεψη με κριτήριο τις συντηρημένες ακολουθίες μειώνει σημαντικά το ποσοστό λανθασμένα θετικών αποτελεσμάτων. Βέβαια, οι ορισμοί για το conservation διαφέρουν σε διαφορετικούς αλγόριθμους πρόβλεψης γονιδίων στόχων miRNA.

Η ελεύθερη ενέργεια (free energy ή Gibbs free energy) μπορεί να χρησιμοποιηθεί σαν μετρική για την σταθερότητα ενός βιολογικού συστήματος. Αν το δέσιμο ενός miRNA με έναν υποψήφιο στόχο mRNA υπολογίζεται να είναι σταθερό, τότε θεωρείται πιο πιθανό να είναι πραγματικός στόχος του miRNA (Peterson et al.,2014).

Η προσβασιμότητα (site accessibility) είναι μια μετρική σχετικά με την ευκολία που ένα miRNA εντοπίζει, τοποθετείται και διασταυρώνεται με τον στόχο mRNA.

2. Νέοι αλγόριθμοι πρόβλεψης στόχων miRNA

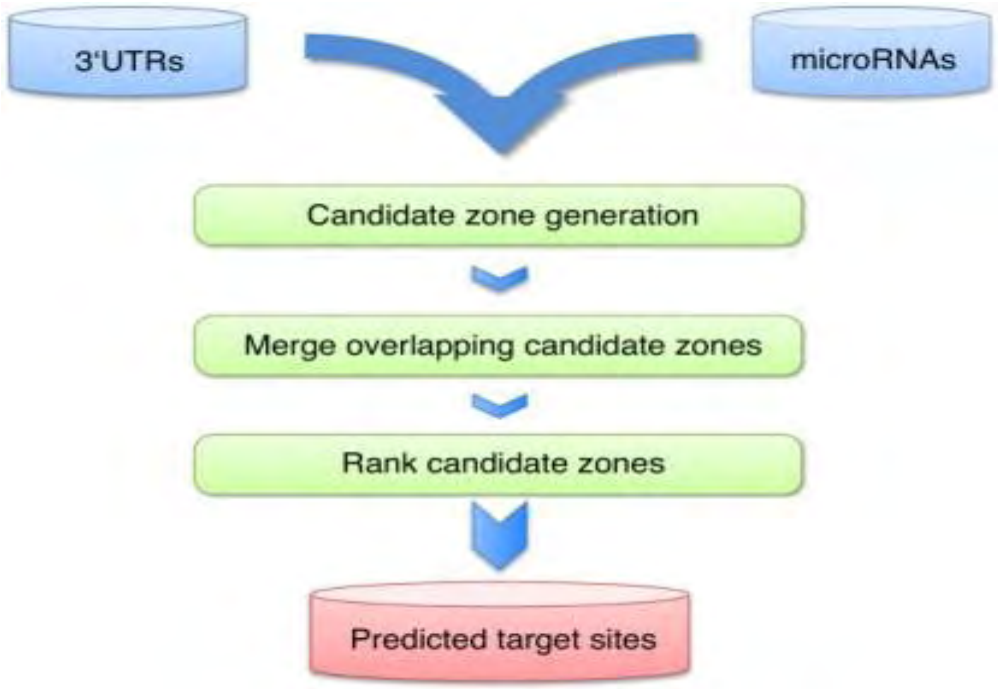
2.1 TargetSpy

Ο TargetSpy (Sturm et al., 2010) προβλέπει στόχους microRNA, ανεξάρτητα από την ύπαρξη κεντρικής περιοχής (seed region). Ο TargetSpy είναι ο πρώτος αλγόριθμος του οποίου η εκπαίδευση (training) πραγματοποιήθηκε πάνω σε ποντίκι. Το training βασίστηκε σε μία δημοσίευση που παρουσίασε ένα σύνολο από argonaute-mRNA θέσεις πρόσβασης (binding sites) για τα 20 πιο πολυάριθμα miRNA στον εγκέφαλο του ποντικιού P13. Οι πρωτεΐνες argonaute είναι τα καταλυτικά στοιχεία του RNA-induced silencing complex (RISC), το σύμπλεγμα πρωτεϊνών που είναι υπεύθυνο για την καταστολή της έκφρασης του στόχου mRNA. Συνολικά, χρησιμοποιήθηκαν 3872 θετικά και 4540 αρνητικά παραδείγματα.

Όπως προαναφέρθηκε, οι περισσότεροι αλγόριθμοι πρόβλεψης γονιδίων στόχων για miRNA βασίζονται κυρίως στις εξής παραμέτρους: α. την ύπαρξη seed περιοχής και β. την εξελικτική διατήρηση (conservation) του στόχου μεταξύ διάφορων ειδών γ. την ελεύθερη ενέργεια (free energy ή Gibbs free energy) και δ. την προσβασιμότητα (site accessibility). Αυτές οι παράμετροι οδηγούν σε πιο αξιόπιστα αποτελέσματα, όμως μειώνουν την ικανότητα του αλγορίθμου στο να βρίσκει στόχους. Έτσι, ένα μεγάλο μέρος των στόχων απορρίπτεται. Ο TargetSpy σε αντίθεση με τους περισσότερους αλγορίθμους δεν λαμβάνει υπ' όψιν το conservation ως κριτήριο και δεν επιβάλλει αυστηρή απαίτηση για seed ταίριασμα. Παρ' όλα αυτά στο paper παρουσιάζονται υποσύνολα των προβλέψεων που ικανοποιούν τις δύο παραμέτρους, ώστε να πραγματοποιηθεί καλύτερα η σύγκριση του TargetSpy με άλλους αλγορίθμους. Έτσι, παρατίθενται αποτελέσματα του TargetSpy no-seed που δεν πληροί καμιά από τις δύο παραμέτρους, του TargetSpy seed που ικανοποιεί την πρώτη παράμετρο και του TargetSpy cons. seed που ικανοποιεί και τις δύο.

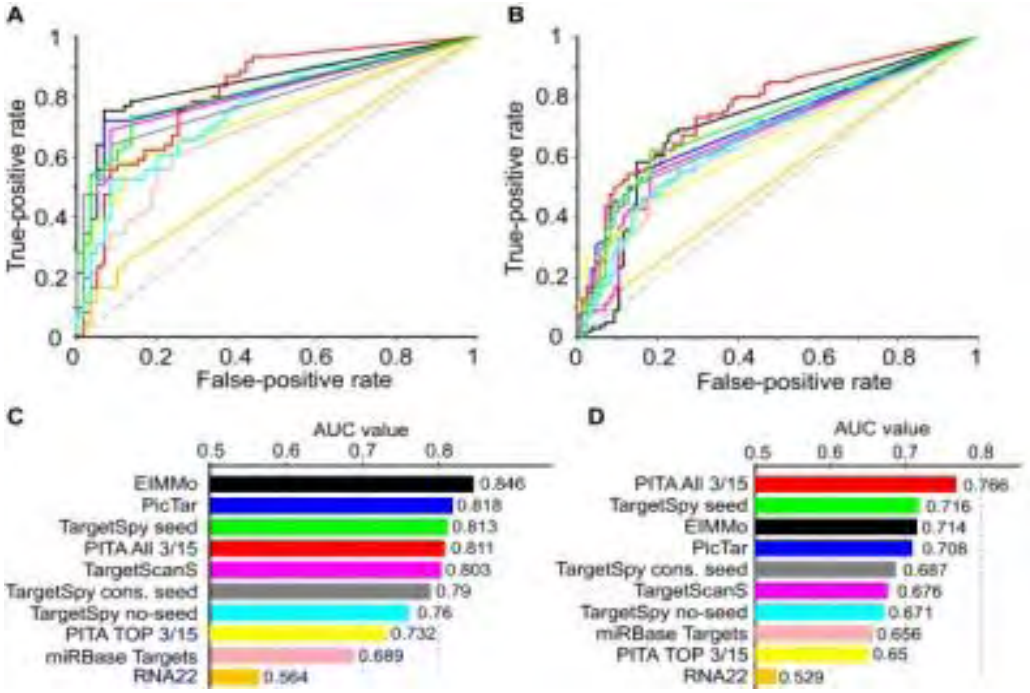
Για κάθε εισερχόμενο miRNA ο TargetSpy αναγνωρίζει ζώνες υποψηφίων (candidate zones) κατά μήκος όλων των σειρών 3΄ UTR. Στη συνέχεια υπολογίζει το σκορ για τον αντιπρόσωπο κάθε υποψήφιας ζώνης, συγχωνεύει τις επικαλυπτόμενες ζώνες υποψηφίων και ταξινομεί τις προβλέψεις ανάλογα με το σκορ τους. Το σκορ υπολογίζεται με βάση εφτά χαρακτηριστικά στα οποία κατέληξαν μετά από αξιολόγηση της απόδοσης 45 χαρακτηριστικών. Τα συγκεκριμένα εφτά

χαρακτηριστικά είχαν από κοινού την καλύτερη απόδοση. Το pipeline της διαδικασίας πρόβλεψης του αλγορίθμου TargetSpy φαίνεται στην εικόνα 2.1.



Εικόνα 2.1: Pipeline λειτουργίας του αλγορίθμου TargetSpy (Πηγή: Sturm et al., 2010)

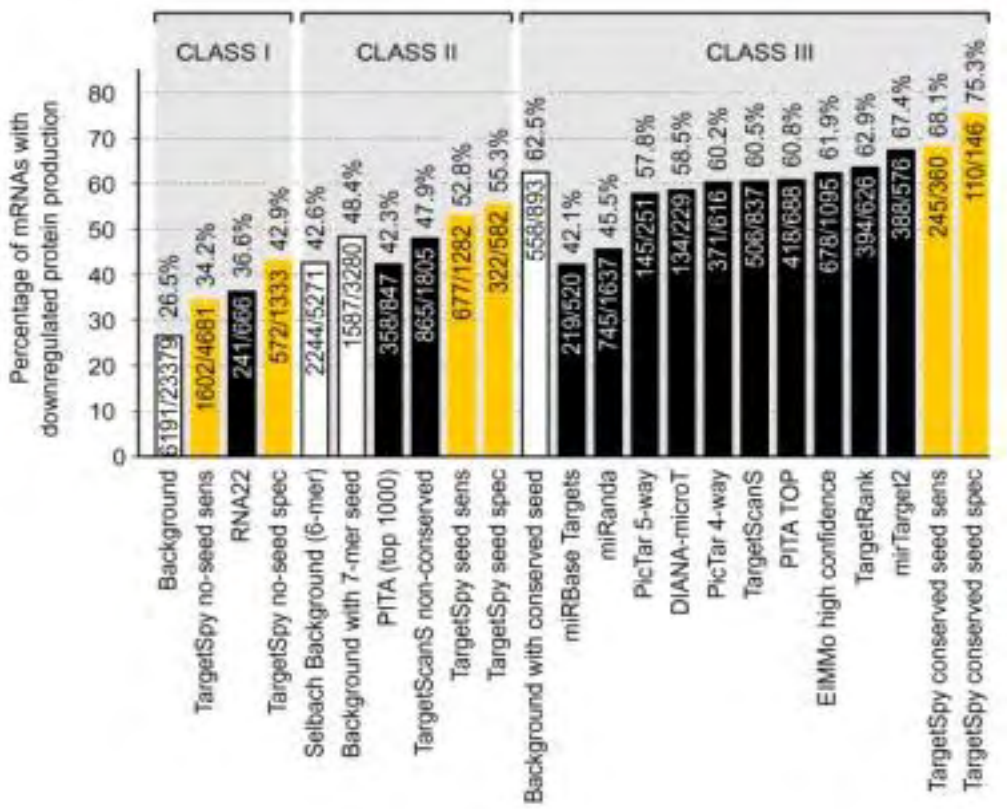
Για να συγκριθεί ο TargetSpy με διάφορους γνωστούς αλγορίθμους χρησιμοποιήθηκε ένα σετ δεδομένων που εξήχθησαν από μια σύγκριση που πραγματοποιήθηκε από τους Stark et al. (2005). Το συγκεκριμένο σετ δεδομένων αποτελείται από 120 αλληλεπιδράσεις miRNA-gene. Αργότερα, επεκτάθηκε από τους Kertesz et al. (2007), περιλαμβάνοντας 190 αλληλεπιδράσεις. Χρησιμοποιούνται και τα δύο σετ δεδομένων, ώστε να πραγματοποιηθεί καλύτερα η σύγκριση και παρατίθενται δύο διαγράμματα για κάθε σετ δεδομένων, η ROC καμπύλη και οι τιμές AUC (εικόνα 2.2). Η τιμή AUC αντιπροσωπεύει την πιθανότητα ότι ένα τυχαίο θετικό παράδειγμα ταξινομείται καλύτερα από ένα τυχαίο αρνητικό. Φαίνεται ότι ο TargetSpy έχει παρόμοια απόδοση με τους περισσότερους αλγορίθμους όταν χρησιμοποιεί το κριτήριο seed ταίριασμα. Όταν δεν το χρησιμοποιεί έχει καλύτερη απόδοση από τον RNA22, τον μοναδικό διαφορετικό αλγόριθμο που δεν προϋποθέτει το συγκεκριμένο κριτήριο. Στο σετ δεδομένων των Stark ο EIMMo πετυχαίνει την μεγαλύτερη τιμή AUC, ενώ στο σετ δεδομένων των Kertesz et al. ο PITA ALL 3/15.



Εικόνα 2.2: Σύγκριση της απόδοσης του TargetSpy χρησιμοποιώντας το σετ δεδομένων που συντάχθηκε από Stark (2005) –A και C- και από τον Kertesz (2007) –B και D- (Πηγή: Sturm et al., 2010)

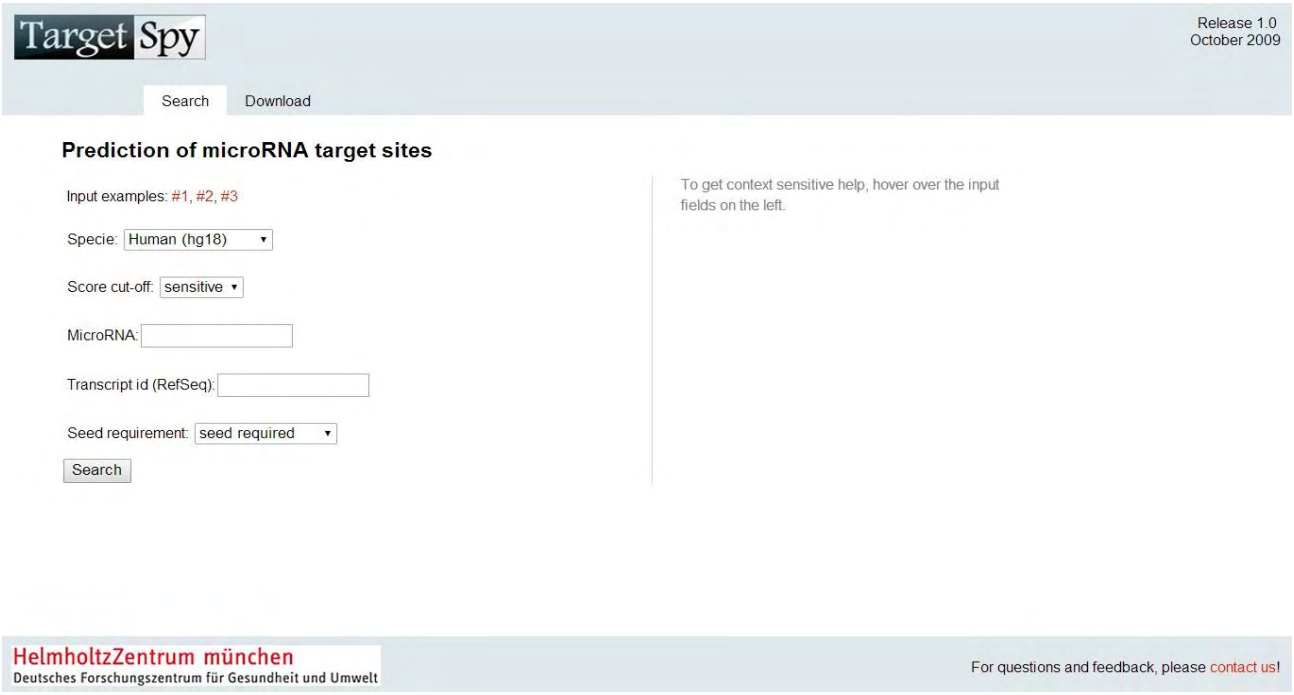
Οι Selbach et al. (2008) πραγματοποίησαν μια σύγκριση των πιο γνωστών αλγορίθμων μετρώντας το κομμάτι των γονιδίων που είχαν προβλεφθεί, το οποίο ήταν συνδεδεμένο με πρωτεΐνες που έχουν γίνει down-regulated περισσότερο από -0.1 $\log_2$ fold change. Τα microRNAs, πάνω στα οποία πραγματοποιήθηκε αυτή η μέτρηση είναι τα εξής: miR-1, miR-16, miR-155, miR-30a-5p και let-7b. Στην εικόνα 2.3, που παρουσιάζεται παρακάτω η πρώτη τιμή σε κάθε ράβδο αντιπροσωπεύει τον αριθμό των αλληλεπιδράσεων miRNA- στόχων που έχουν προβλεφθεί από τον κάθε αλγόριθμο και είναι συνδεδεμένοι με πρωτεΐνες που έχουν γίνει down-regulated, ενώ η δεύτερη τιμή τον συνολικό αριθμό των αλληλεπιδράσεων που έχουν προβλεφθεί από το pSILAC σετ δεδομένων. Το ποσοστό που φαίνεται αντιπροσωπεύει την ακρίβεια (accuracy) κάθε αλγορίθμου. Οι κλάσεις αντιπροσωπεύουν την ύπαρξη των κριτηρίων conservation και seed ταίριασμα (κλάση 3), την ύπαρξη του κριτηρίου seed ταίριασμα (κλάση 2) και τέλος την απουσία και των δύο κριτηρίων (κλάση 1). Παρουσιάζονται επίσης δύο εκδοχές του TargetSpy σε κάθε κλάση, η μία με ρυθμίσεις που ευνοούν το sensitivity (sens), ενώ η άλλη με ρυθμίσεις που ευνοούν το specificity (spec). Ο TargetSpy έχει την καλύτερη ακρίβεια σε κάθε κλάση. Στην τρίτη κλάση παρουσιάζει αρκετά χαμηλό sensitivity σε σχέση με τους υπόλοιπους

αλγορίθμους της ίδιας κλάσης, κάτι που δεν συμβαίνει στις δύο προηγούμενες κλάσεις.



Εικόνα 2.3: Σύγκριση της απόδοσης του TargetSpy βασιζόμενη στο σετ δεδομένων pSILAC (Πηγή: Sturm et al., 2010)

Το interface του TargetSpy παρουσιάζεται στην επόμενη εικόνα:

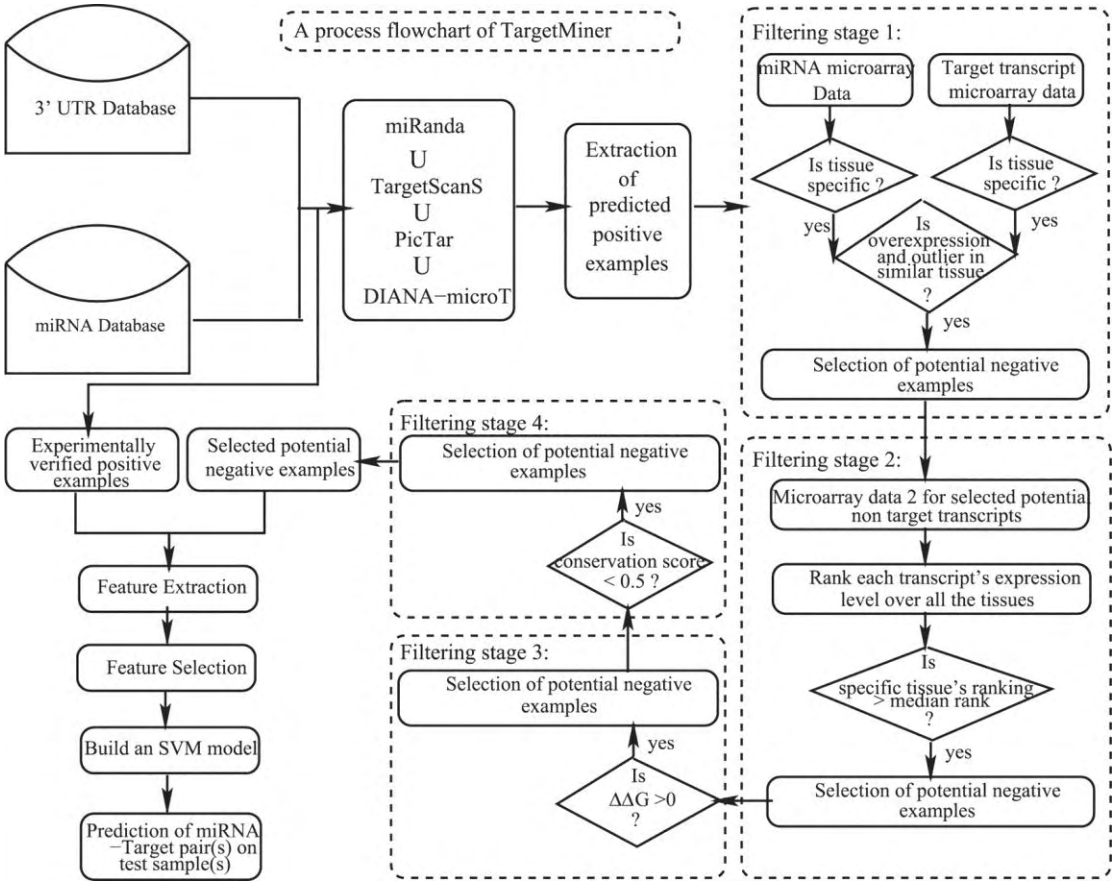


Εικόνα 2.4: TargetSpy (http://www.targetspy.org/)

2.2 TargetMiner

Ένα από τα βασικότερα προβλήματα των αλγόριθμων μηχανικής μάθησης (machine learning) είναι η έλλειψη αρνητικών παραδειγμάτων. Υπάρχουν πολύ λίγα αρνητικά παραδείγματα για να χτιστεί ένας αποτελεσματικός ταξινομητής. Η συστηματική αναγνώριση mRNAs που δεν αποτελούν στόχους δεν αντιμετωπίζεται σωστά και γι' αυτό οι περισσότεροι αλγόριθμοι μηχανικής μάθησης βασίζονται σε τεχνητά παραγόμενα αρνητικά παραδείγματα. Ως εκ τούτου, οι περισσότεροι έχουν λανθασμένα θετικά αποτελέσματα ή λανθασμένα αρνητικά αποτελέσματα σε μεγάλο βαθμό.

Ο TargetMiner (Bandyopadhyay και Mitra, 2009) είναι ένας αλγόριθμος πρόβλεψης στόχων miRNA, που βασίζεται στην συστηματική αναγνώριση ιστοειδικών (tissue-specific) αρνητικών παραδειγμάτων. Ένα σύνολο από 289 βιολογικά επικυρωμένα miRNA (θετικά παραδείγματα) εξήχθη από την βάση δεδομένων miRecords (Xiao et al., 2009). Για να βρουν πιθανά αρνητικά παραδείγματα κάνουν την εξής διαδικασία. Αρχικά, συγκεντρώνουν όλους τους προβλεφθέντες στόχους που βρέθηκαν από άλλους γνωστούς αλγορίθμους πρόβλεψης γονιδίων στόχων για miRNA (miRanda, TargetScanS, PicTar -4way και 5way- και DIANA-microT). Πολλά γονίδια θεωρούνται λανθασμένα στόχοι, επειδή οι αλγόριθμοι βασίζονται σε δομικές (structural) αλληλεπιδράσεις ή αλληλεπιδράσεις σειρών (sequence). Όμως, οι στόχοι ενός miRNA είναι ιστοειδικοί ή cell-type-specific. Έτσι βρίσκουν μεγάλο αριθμό από λανθασμένα θετικά αποτελέσματα για κάθε ιστοειδικό miRNA. Από αυτά κρατάνε τα τελικά αρνητικά παραδείγματα με κατάλληλο φιλτράρισμα. Υπάρχουν τέσσερα στάδια. Στα πρώτα δύο γίνεται ένα φιλτράρισμα του αριθμού των πιθανών στόχων. Στο τρίτο αυτά που έχουν παραμείνει ελέγχονται για θερμοδυναμική ευστάθεια (thermodynamic stability) του διπλότυπου (duplex) miRNA-mRNA και αν έχουν εφικτή αγνοούνται. Τέλος, στο τέταρτο ένα μέρος των στόχων mRNAs εξαιρούνται με βάση το conservation σκορ στην seed περιοχή. Η διαδικασία φαίνεται αναλυτικά στην εικόνα 2.5.

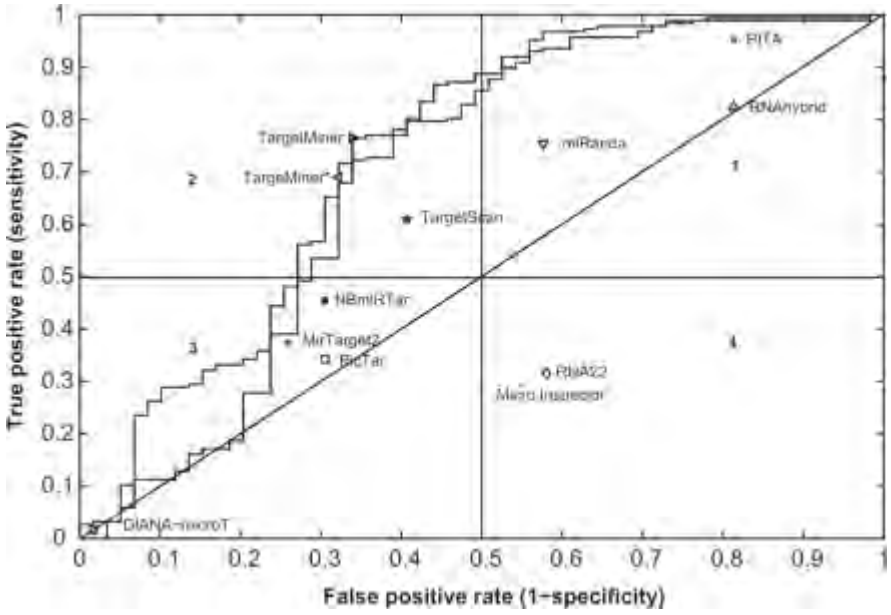


Εικόνα 2.5: Διάγραμμα αναγνώρισης των αρνητικών παραδειγμάτων και πρόβλεψης miRNA στόχων του αλγορίθμου TargetMiner (Πηγή: Bandyopadhyay και Mitra, 2009)

Με βάση τα αρνητικά παραδείγματα χτίζεται ένας ταξινομητής support vector machine (SVM) χρησιμοποιώντας 30 χαρακτηριστικά. Ο SVM ταξινομητής εκπαιδεύτηκε χρησιμοποιώντας την kernel radial basis function (RBF). Επίσης, παραθέτουν τα αποτελέσματα για έναν ακόμα ταξινομητή, που φτιάχνεται χρησιμοποιώντας 90 χαρακτηριστικά, αναφερόμενος ως TargetMiner*.

Σε ένα ανεξάρτητο σετ δεδομένων πραγματοποιείται σύγκριση του TargetMiner με 10 άλλους αλγόριθμους πρόβλεψης στόχων miRNA.Το συγκεκριμένο σετ δεδομένων απαρτίζεται από 246 αλληλεπιδράσεις, από τις οποίες οι 187 ήταν θετικές. Ένα σετ από 181 θετικά παραδείγματα πήραν από την βάση δεδομένων miRecords, ενώ τα υπόλοιπα θετικά από διάφορα πειράματα. Επίσης, χρησιμοποίησαν ένα σετ δεδομένων από 59 αρνητικά παραδείγματα από τους οργανισμούς Drosophila, ποντίκι και άνθρωπο, τα οποία εξήχθησαν από την βάση δεδομένων TarBase. Η απόδοση των αλγορίθμων αξιολογείται με βάση τις παραμέτρους sensitivity, specificity, Matthew's correlation coefficient (MCC) και average classwise accuracy (ACA). Οι

δύο εκδοχές του TargetMiner παρουσιάζουν τα καλύτερα αποτελέσματα όσον αφορά το MCC και το ACA. Παρατίθεται επίσης το διάγραμμα διασποράς (scatter plot) μεταξύ των σωστά θετικών (sensitivity) έναντι λανθασμένων θετικών αποτελεσμάτων (1- specificity). Η περιοχή του διαγράμματος χωρίζεται σε τέσσερα τεταρτημόρια. Το πρώτο τεταρτημόριο υποδηλώνει χαμηλό specificity και υψηλό sensitivity. Το τρίτο τεταρτημόριο υποδηλώνει υψηλό specificity και χαμηλό sensitivity. Το τέταρτο τεταρτημόριο υποδηλώνει χαμηλό sensitivity και χαμηλό specificity. Το καλύτερο τεταρτημόριο είναι το δεύτερο, το οποίο υποδηλώνει μεγάλο sensitivity και specificity. Εκεί βρίσκονται οι δύο εκδοχές του TargetMiner και ο TargetScan, με τον τελευταίο όμως να έχει μικρότερο MCC και να βρίσκεται περισσότερο κοντά στην διαγώνιο. Το διάγραμμα φαίνεται στην εικόνα 2.6.



Εικόνα 2.6 : Διάγραμμα διασποράς του sensitivity έναντι του (1-specificity) (Πηγή: Bandyopadhyay και Mitra, 2009)

Σε ένα σετ δεδομένων αποτελούμενο από mRNAs που είτε έχει καταστολεί η μετάφραση τους (translationally repressed) είτε έχουν αποικοδομηθεί (mRNA cleavage) το οποίο εξήχθη από την TarBase (Papadopoulos et al., 2009) μετρήθηκε το sensitivity και το specificity σε σχέση με πέντε ακόμα αλγορίθμους. Όσον αφορά το cleavage σετ δεδομένων τα sensitivities των TargetMiner, TargetScan, PicTar, miRanda, MirTarget2 και NBmiRTar είναι 0.816, 0.790, 0.684, 0.658, 0.658 και 0.526 αντίστοιχα. Όσον αφορά το translationally repressed σετ δεδομένων τα

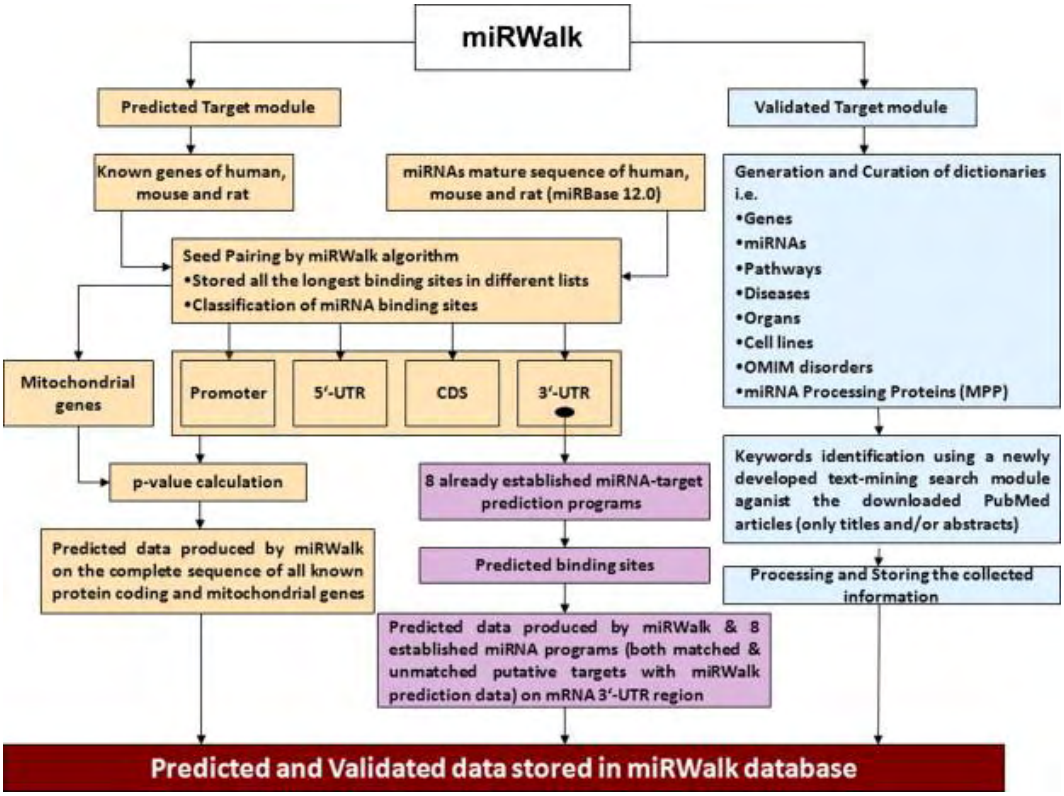
sensitivities των TargetMiner, TargetScan, NBmiRTar, miRanda, PicTar and MirTarget2 είναι 0.723, 0.661, 0.661, 0.569, 0.508 and 0.456 αντίστοιχα. Φαίνεται ότι και στις δύο περιπτώσεις ο TargetMiner πετυχαίνει το μεγαλύτερο sensitivity.

Το interface του TargetMiner παρουσιάζεται στην εικόνα 2.7.

Εικόνα 2.7: TargetMiner (http://www.isical.ac.in/~bioinfo_miu/targetminer20.htm**)**

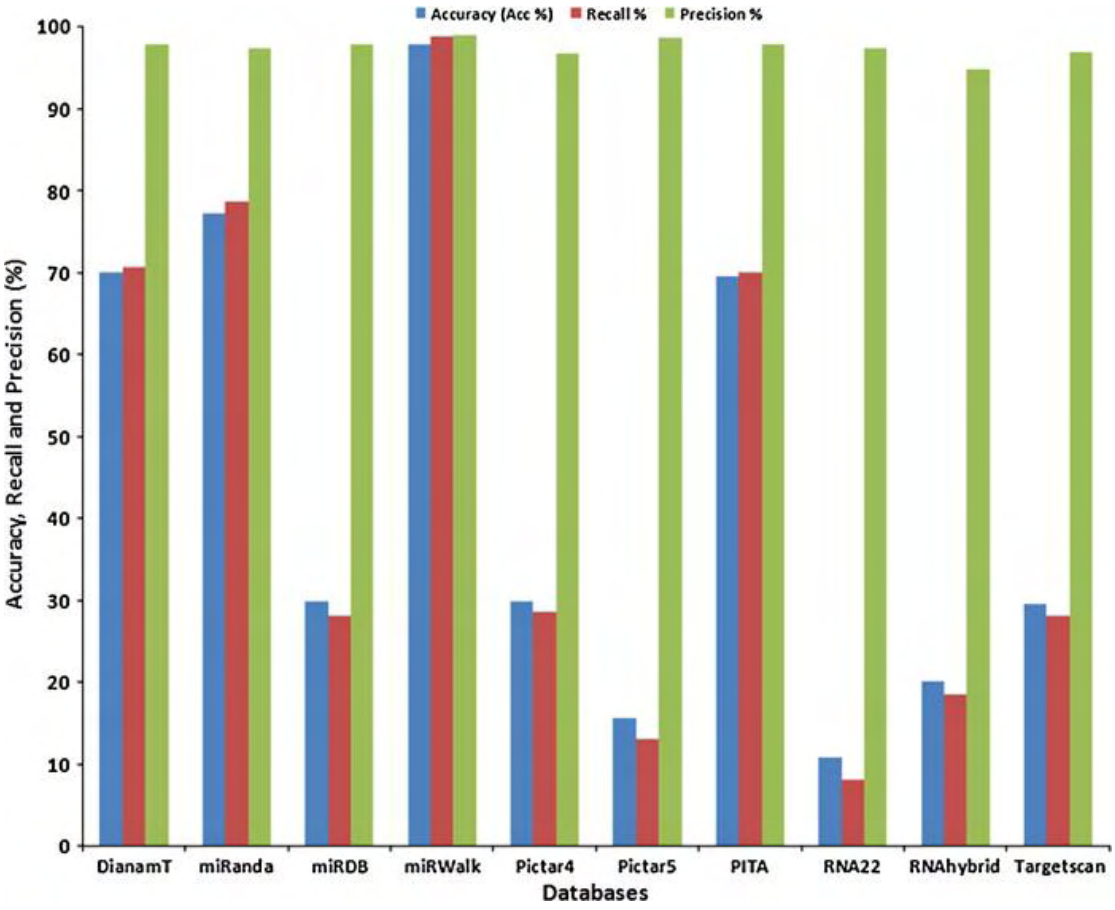
2.3 miRWalk

Ο miRWalk (Dweep et al., 2010) είναι μια βάση δεδομένων που παρέχει πληροφορίες για προβλεφθέντα και πιστοποιημένα binding sites για τον άνθρωπο, το ποντίκι και τον αρουραίο. Για να προβλέψει στόχους miRNA χρησιμοποιεί τον αλγόριθμο miRWalk. Ο συγκεκριμένος αλγόριθμος προσπαθεί να αναγνωρίσει πολλαπλές συνεχόμενες υποακολουθίες (subsequences) μεταξύ του miRNA και σειρών γονιδίων. Ο miRWalk ψάχνει για seed περιοχές, βασιζόμενος στην συμπληρωματικότητα Watson-Crick. Όταν βρίσκει ένα επταμερές (heptamer) που κάνει τέλειο base pairing αυξάνει το μήκος της seed περιοχής του miRNA μέχρι να βρει μια αστοχία. Ύστερα, επιστρέφει όλες τις πιθανές ευστοχίες με εφτά η περισσότερα ταιριάσματα. Τα binding sites της αναλυόμενης ακολουθίας χωρίζονται με βάση τις αναγνωρισμένες τοποθεσίες (θέση έναρξης, θέση τέλους και περιοχή). Έπειτα, χωρίζει τα αποτελέσματα σε πέντε κατηγορίες (promoter region, 5′-UTR, coding sequence (CDS), και 3′-UTR και mitochondrial genes). Ακολουθεί ο υπολογισμός της κατανομής πιθανότητας τυχαίων ταιριασμάτων σε μια υποακολουθία στην αναλυόμενη ακολουθία. Η διαδικασία φαίνεται αναλυτικά στην εικόνα 2.8. Όσον αφορά τους πιστοποιημένους στόχους ο miRWalk πραγματοποιεί μια αυτόματη αναζήτηση χρησιμοποιώντας εξόρυξη κειμένου (text mining) στους τίτλους και τις περιλήψεις των άρθρων του PubMed. Τα αποτελέσματα αποθηκεύονται στη βάση δεδομένων miRWalk.



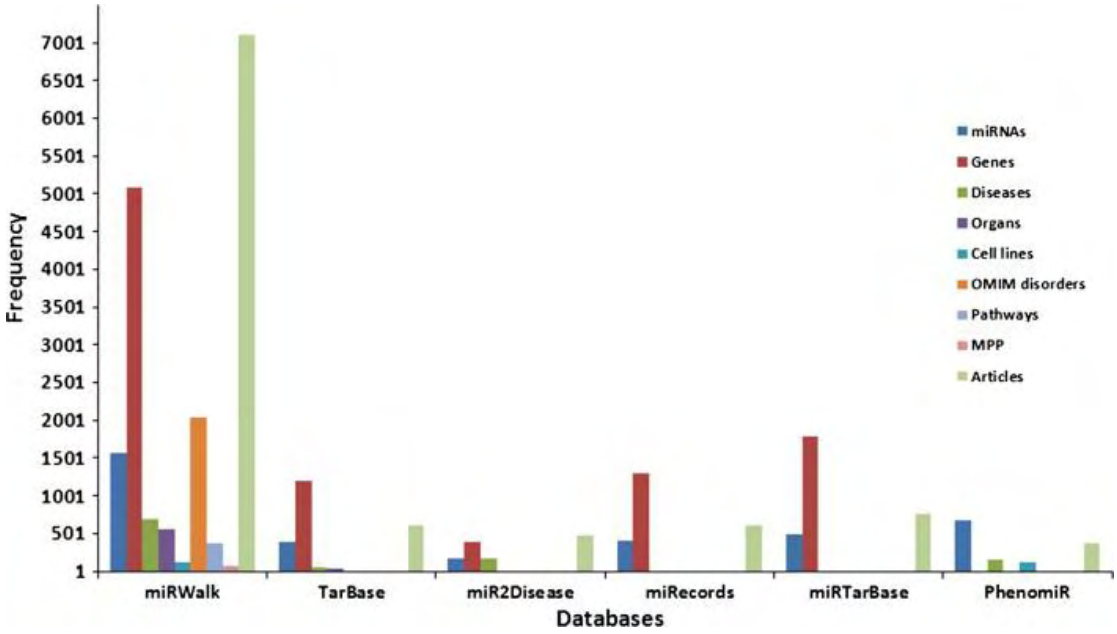
Εικόνα 2.8: Ροή εργασίας του αλγορίθμου miRWalk (Πηγή: Dweep et al., 2010)

Πραγματοποιήθηκε μια σύγκριση του miRWalk με οχτώ ακόμη αλγορίθμους πρόβλεψης στόχων miRNA, χρησιμοποιώντας 1870 θετικά και 61 αρνητικά ζευγάρια miRNA – στόχων, εξαγόμενα από την TarBase (Papadopoulos et al., 2009) , την miRecords (Xiao et al., 2009) και την miRTarBase (Hsu et al., 2011) (Εικόνα 2.9). Το θετικό και το αρνητικό σετ δεδομένων δόθηκαν σαν είσοδοι σε καθένα από τα προγράμματα και από τα αποτελέσματα υπολόγισαν τα Accuracy, Recall και Precision. Ο αλγόριθμος miRWalk πέτυχε Accuracy 97.93%, Recall 98.88% και Precision 98.98%. Οι DIANA microT, miRanda και PITA πέτυχαν επίσης σχετικά καλή απόδοση.



Εικόνα 2.9: Σύγκριση της απόδοσης του miRWalk με βάση το accuracy, το recall και το precision (Πηγή: Dweep et al., 2010)

Όσον αφορά τα πιστοποιημένα δεδομένα, πραγματοποιήθηκε μια ακόμα σύγκριση του miRWalk με πέντε γνωστές βάσεις δεδομένων, συγκεκριμένα τις εξής: TarBase, miR2Disease, miRecords, PhenomiR και miRTarBase (Εικόνα 2.10).



Εικόνα 2.10: Σύγκριση της απόδοσης του miRWalk με άλλες βάσεις δεδομένων (Πηγή: Dweep et al., 2010)

Το interface του miRWalk παρουσιάζεται στην εικόνα 2.11.

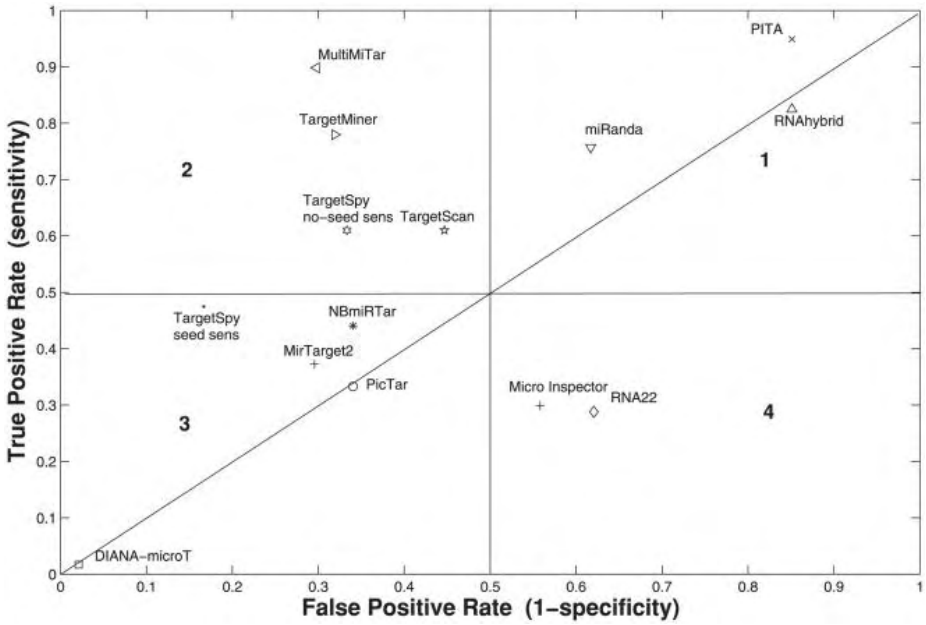
Εικόνα 2.11: miRWalk (http://www.umm.uni-heidelberg.de/apps/zmf/mirwalk/)

2.4 MultiMiTar

Ο MultiMiTar (Mitra και Bandyopadhyay, 2011) είναι ένας αλγόριθμος πρόβλεψης στόχων miRNA, ο οποίος δημιουργήθηκε από την ίδια επιστημονική ομάδα που δημιούργησε τον TargetMiner. Ο MultiMiTar χρησιμοποιεί τα ίδια σετ δεδομένων με τον TargetMiner, ώστε να εκπαιδευτεί ο ταξινομητής. Χρησιμοποιείται δηλαδή ένα σετ δεδομένων από 289 βιολογικά επικυρωμένα θετικά αποτελέσματα και 289 αρνητικά ιστοειδικά παραδείγματα. Η διαφορά τους έγκειται στον τρόπο με τον οποίο επιλέγονται τα χαρακτηριστικά, ώστε να χτισθεί ο ταξινομητής.

Όσον αφορά τον TargetMiner, χρησιμοποιήθηκε ένα μοναδικό κριτήριο, ώστε να επιλεχθούν τα τελικά χαρακτηριστικά. Συγκεκριμένα, χρησιμοποιήθηκε ένα τύπος ώστε να υπολογισθεί το σκορ και στην συνέχεια κρατήθηκε το 1/3 των χαρακτηριστικών που πετύχαιναν το μεγαλύτερο σκορ, καταλήγοντας με αυτόν τον τρόπο στα τελικά 30 χαρακτηριστικά. Η προσέγγιση αυτή (F-score) δεν πετυχαίνει βέλτιστα αποτελέσματα και γι' αυτό το πρόβλημα της επιλογής χαρακτηριστικών είναι προτιμότερο να αντιμετωπισθεί σαν multi-objective-optimization (MOO). Μια μέθοδος βελτιστοποίησης που έχουν δημοσιεύσει στο παρελθόν η Archived Multi-Objective Simulated Annealing (AMOSA) ενσωματώνεται με το Support Vector Machine (SVM) για να κτισθεί το AMOSA- SVM για επιλογή χαρακτηριστικών και ταξινόμηση. Το AMOSA υπερισχύει άλλων δημοφιλών MOO τεχνικών και ενισχύει την προγνωστική ικανότητα του MultiMiTar.

Σε ένα ανεξάρτητο σετ δεδομένων παρατίθεται η σύγκριση του MultiMiTar με 13 ακόμα αλγορίθμους πρόβλεψης στόχων miRNA. Περιέχει και τις τρεις εκδοχές του TargetSpy, για τις οποίες έχουν δοθεί αποτελέσματα, για τις οποίες μιλήσαμε προηγουμένως. Παρατίθεται το διάγραμμα διασποράς μεταξύ των σωστά θετικών (sensitivity) έναντι λανθασμένα αρνητικών αποτελεσμάτων (1-specificity), παρόμοιο με το διάγραμμα της εικόνας 2.6. Ο MultiMiTar πετυχαίνει τα καλύτερα αποτελέσματα, αφού βρίσκεται στο δεύτερο τεταρτημόριο (μεγάλο sensitivity και specificity), ενώ παράλληλα έχει μεγαλύτερο sensitivity και specificity από τους υπόλοιπους αλγορίθμους που βρίσκονται στο δεύτερο τεταρτημόριο (TargetMiner, TargetSpy και Targetscan). Το διάγραμμα φαίνεται στην εικόνα 2.12.

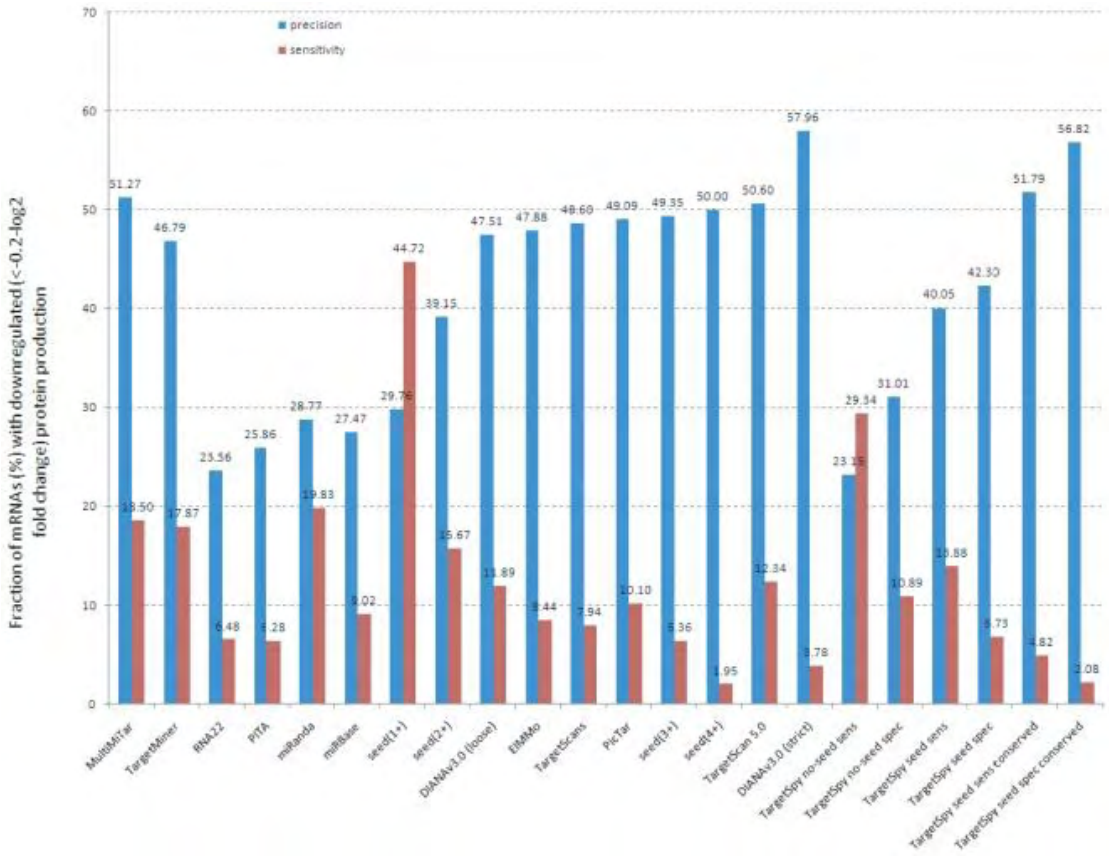


Εικόνα 2.12: Διάγραμμα διασποράς του sensitivity έναντι του (1- specificity) (Πηγή: Mitra και Bandyopadhyay, 2011)

Για να πραγματοποιηθεί με μεγαλύτερη αξιοπιστία η σύγκριση με τους συγκεκριμένους αλγορίθμους που ανήκουν επίσης στο δεύτερο τεταρτημόριο υπολογίζεται το Matthew's correlation coefficient (MCC) και το average classwise accuracy (ACA). Ο MultiMiTar πετυχαίνει τις καλύτερες τιμές (MCC=0.583 και ACA=0.8), ενώ ο TargetMiner (MCC = 0.403 και ACA = 0.73), ο TargetScan (MCC = 0.135 και ACA = 0.582) και τέλος ο TargetSpy no-seed sens (MCC = 0.209 και ACA = 0.56). Για την σύγκριση του MultiMiTar με τον TargetMiner υπολογίστηκε και η Area under the curve (AUC), με τον MultiMiTar να πετυχαίνει 0.7464 και τον TargetMiner 0.7085, δείχνοντας ότι η εξαγωγή χαρακτηριστικών με βάση το AMOSA στον MultiMiTar ευθύνεται για την βελτιωμένη συμπεριφορά.

Πραγματοποιούν μια ακόμα σύγκριση βασιζόμενοι στο σετ δεδομένων pSILAC. Χρησιμοποιούν μια πιο αυστηρή προσέγγιση από αυτή που περιγράφηκε στην σύγκριση του TargetSpy προηγουμένως (Selbach et al.,2008). Πρωτεΐνες που έχουν γίνει downregulated περισσότερο από -0.2log2 fold change, θεωρούνται στόχοι (Alexiou et al.,2009). Οι υπόλοιπες αλληλεπιδράσεις δεν θεωρούνται στόχοι. Στα αποτελέσματα της σύγκρισης που παρατίθεται στο Alexiou et al. (2009), πρόσθεσαν

αποτελέσματα για τον TargetMiner, τον MultiMiTar και τον TargetSpy όσον αφορά τις παραμέτρους precision και recall. Ενώ οι TargetScan 5.0, DIANA v3.0(strict) και κάποιες παραλλαγές του TargetSpy πετυχαίνουν το μεγαλύτερο precision, έχουν χαμηλές τιμές στο recall. Ο miRanda πετυχαίνει κάλο sensitivity, αλλά έχει πολύ χαμηλό precision. Από την άλλη ο MultiMiTar πετυχαίνει precision 51.27% και recall 18.50%, έχοντας μια ισορροπημένη απόδοση σε σχέση με τα υπόλοιπα προγράμματα που υστερούν είτε στο precision είτε στο recall. Το διάγραμμα με τα παρακάτω αποτελέσματα φαίνεται στην εικόνα 2.13.



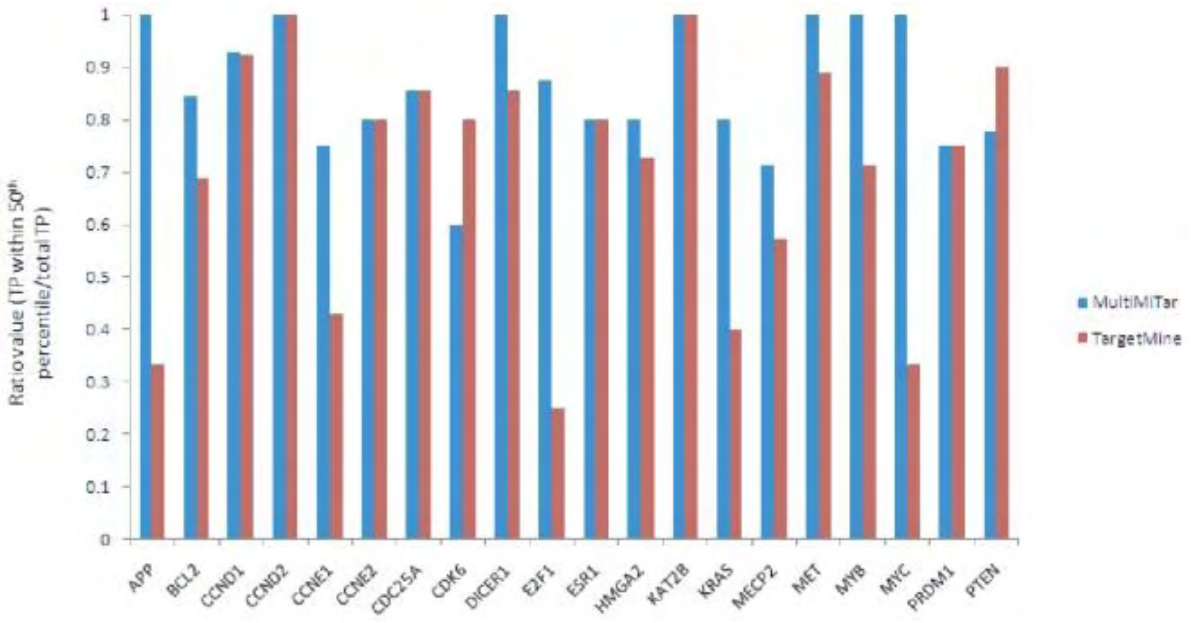
Εικόνα 2.13: Σύγκριση της απόδοσης του MultiMiTar βασιζόμενη στο σετ δεδομένων pSILAC (Πηγή: Mitra και Bandyopadhyay, 2011)

Το γονίδιο p21Cip1/Waf1, επίσης γνωστό ως Cyclin-dependent kinase inhibitor 1A (CDKN1A) μπορεί να γίνει στόχος από 28 miRNA (Wu et al., 2010). Το συγκεκριμένο γονίδιο είναι το μοναδικό που εξακριβώθηκε πειραματικά να ρυθμίζεται από ένα τόσο μεγάλο αριθμό από miRNA. Πάνω στο σετ δεδομένων της αλληλεπίδρασης του CDKN1A με τα miRNA πραγματοποιείται μια σύγκριση του MultiMiTar με τους υπόλοιπους αλγορίθμους πρόβλεψης στόχων miRNA, με

κριτήριο το sensitivity και την σειρά κατάταξης των miRNA. Ο MultiMiTar πετυχαίνει την καλύτερη απόδοση έχοντας sensitivity 67.86%. Οι μοναδικοί αλγόριθμοι που έχουν μεγαλύτερο sensitivity είναι οι PITA και RNAhybrid, κάτι που οφείλεται όμως στο ότι αναγνωρίζουν σχεδόν όλα τα παραδείγματα σαν θετικά, όπως είδαμε προηγουμένως.

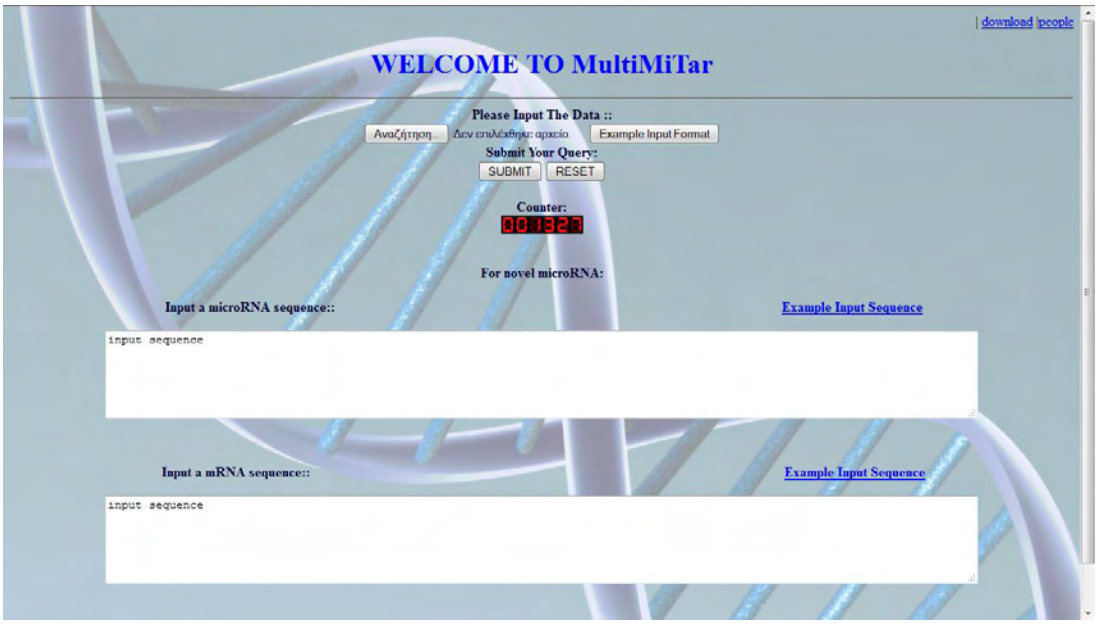
Στην συνέχεια πραγματοποιήθηκε μια σύγκριση με βάση το αν τα αληθινά θετικά αποτελέσματα είναι ομοιόμορφα κατανεμημένα κατά μήκος της λίστας των θετικών αποτελεσμάτων κάθε αλγορίθμου ή είναι κατανεμημένα στην αρχή της λίστας, όπως είναι προτιμότερο. Στην συγκεκριμένη περίπτωση η σύγκριση έγινε με τους έξι αλγορίθμους που πετυχαίνουν μεγαλύτερο sensitivity, δηλαδή με τους miRanda, NBmiRTar, PITA, RNAhybrid, DIANA-microT 3.0 και TargetMiner. Βρέθηκε ότι ο MultiMiTar κατανέμει το 78.95% των σωστά θετικών προβλέψεων στο πρώτο 20% της λίστας των αποτελεσμάτων του. Πετυχαίνει την καλύτερη ταξινόμηση (ranking) από τους υπόλοιπους αλγορίθμους καθώς οι TargetMiner, DIANA-microT 3.0, RNAhybrid, PITA, NBmiRTar και miRanda περιέχουν στο πρώτο 20% της λίστας των αποτελεσμάτων τους το 55%, 36.36%, 25%,25.93%,30.77% και 17.65% αντίστοιχα.

Για την δεύτερη έρευνα χρησιμοποιήθηκαν 20 mRNAs κάθε ένα από τα οποία αποτελεί στόχο 6 ή και περισσοτέρων miRNA, τα οποία εξήχθησαν από την miRTarBase. Στην συνέχεια υπολογίζεται για κάθε mRNA πόσα αληθινά θετικά αποτελέσματα υπάρχουν στο πρώτα μισά των αποτελεσμάτων του MultiMiTar. Αυτό το ποσό διαιρείται με το σύνολο των αληθινά θετικών προβλέψεων. Η συγκεκριμένη διαδικασία πραγματοποιήθηκε και για τον TargetMiner. Και σε αυτή τη περίπτωση ο MultiMiTar φαίνεται να αποδίδει καλύτερα. Στα περισσότερα γονίδια ξεπερνάει τον TargetMiner Το διάγραμμα με τα αποτελέσματα φαίνεται στην εικόνα 2.14.



Εικόνα 2.14: Σύγκριση του MultiMiTar με τον TargetMiner με κριτήριο την ταξινόμηση (ranking) (Πηγή: Mitra και Bandyopadhyay, 2011)

Το interface του MultiMiTar φαίνεται στην εικόνα 2.15.



Εικόνα 2.15: MultiMiTar (http://www.isical.ac.in/~bioinfo_miu/multimitar.htm)

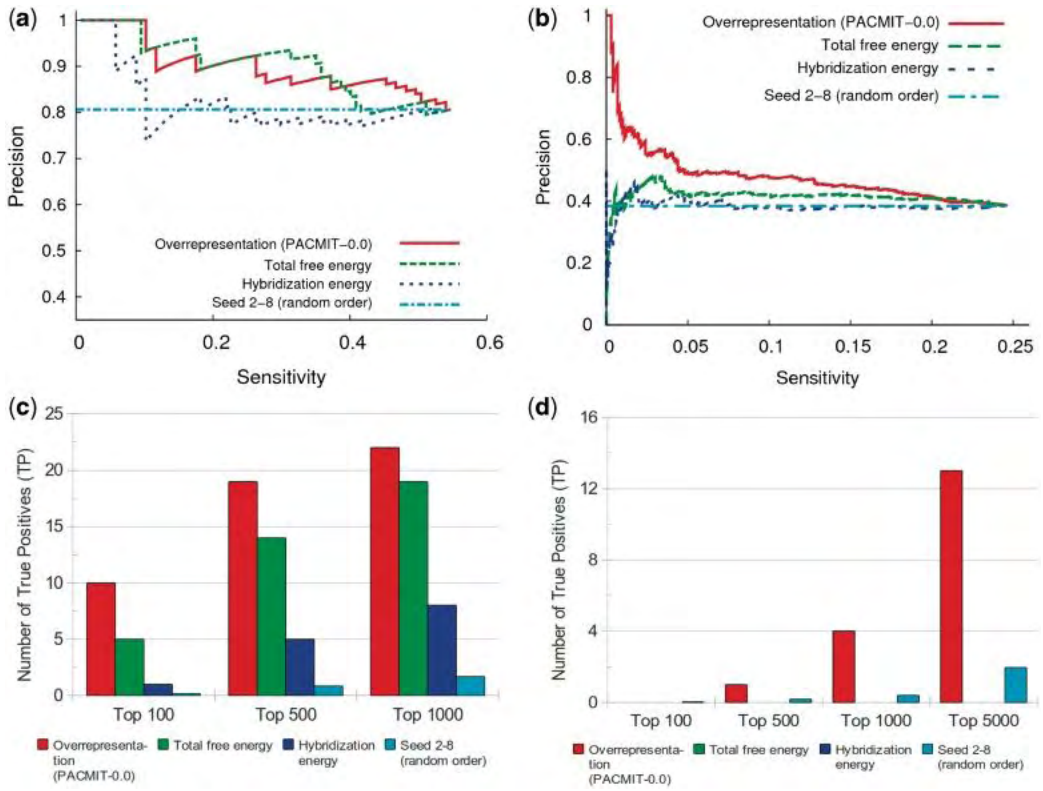
2.5 PACMIT

Ο PACMIT (Marin και Vanicek, 2011) είναι ένας αλγόριθμος πρόβλεψης στόχων microRNA, βασισμένος στην προσβασιμότητα (accessibility). Ταξινομεί τις προβλέψεις χρησιμοποιώντας σαν κριτήριο την υπερβολικά μεγάλου βαθμού αναπαράσταση (over-representation) των προσβάσιμων συμπληρωματικών χώρων.

Η βασική υπόθεση που στηρίζεται ο συγκεκριμένος αλγόριθμος είναι ότι οι βιολογικώς λειτουργικές αλληλεπιδράσεις miRNA-mRNA προκύπτουν από την συνεξέλιξη (coevolution) του miRNA με τον στόχο του. Συνεπώς, τα συμπληρωματικά sites σε πραγματικούς στόχους αντιστοιχούν σε over-represented ολιγομερή. Επιπρόσθετα, εφόσον τα λειτουργικά ζευγάρια miRNA-3΄UTR μπορούν να προκύψουν είτε από απλό δυνατό binding site είτε από πολλαπλά αδύναμα binding sites ολόκληρο το 3΄UTR θεωρείται σαν πιθανός στόχος. Για να συγκριθούν οι δύο πιθανότητες δίκαια ο αλγόριθμος αξιολογεί το επίπεδο του over-representation ενός η περισσοτέρων συμπληρωματικών sites υπολογίζοντας μια τιμή P-value (Psh) για κάθε ζευγάρι miRNA- 3΄UTR. Όσο μικρότερη είναι η πιθανότητα Psh, τόσο πιθανότερο το 3΄UTR να είναι λειτουργικός στόχος. Η γενική διαδικασία είναι να υπολογισθεί το Psh για όλα τα πιθανά ζευγάρια miRNA-3΄UTR για να βγουν οι τελικές προβλέψεις ταξινομημένες με το κριτήριο Psh.

Κάνουν μια σύγκριση για να ελέγξουν την απόδοση του over-representation ως κριτηρίου ταξινόμησης. Για να το πετύχουν αυτό, συγκρίνουν την ακρίβεια που πέτυχαν, όταν όλα τα ζευγάρια miRNA-3΄UTR με τουλάχιστον ένα τέλειο ταίριασμα seed 2-8 ταξινομήθηκαν με τέσσερα διαφορετικά κριτήρια: over-representation, hybridization, total free energy και random order. Και στον άνθρωπο και στην Drosophila Melanogaster το over-representation έχει καλύτερη απόδοση από τα υπόλοιπα κριτήρια, ενώ δεν υπάρχει μεγάλη διαφορά μεταξύ hybridization και random order. Το κριτήριο total free energy πετυχαίνει καλύτερη ακρίβεια από το hybridization και το random order, αλλά στον άνθρωπο το over-representation πετυχαίνει αρκετά καλύτερα αποτελέσματα. Επίσης, στο διάγραμμα που παρατίθεται παρακάτω φαίνεται ότι η ταξινόμηση με κριτήριο το over-representation τοποθετεί πολλά σωστά θετικά στις πρώτες (top) προβλέψεις, σε σχέση με την ταξινόμηση με βάση τα υπόλοιπα κριτήρια. Από την σύγκριση φαίνεται ότι το over-representation είναι πολύ κριτήριο για να ταξινομηθούν οι προβλέψεις. Παραθέτουν ένα διάγραμμα

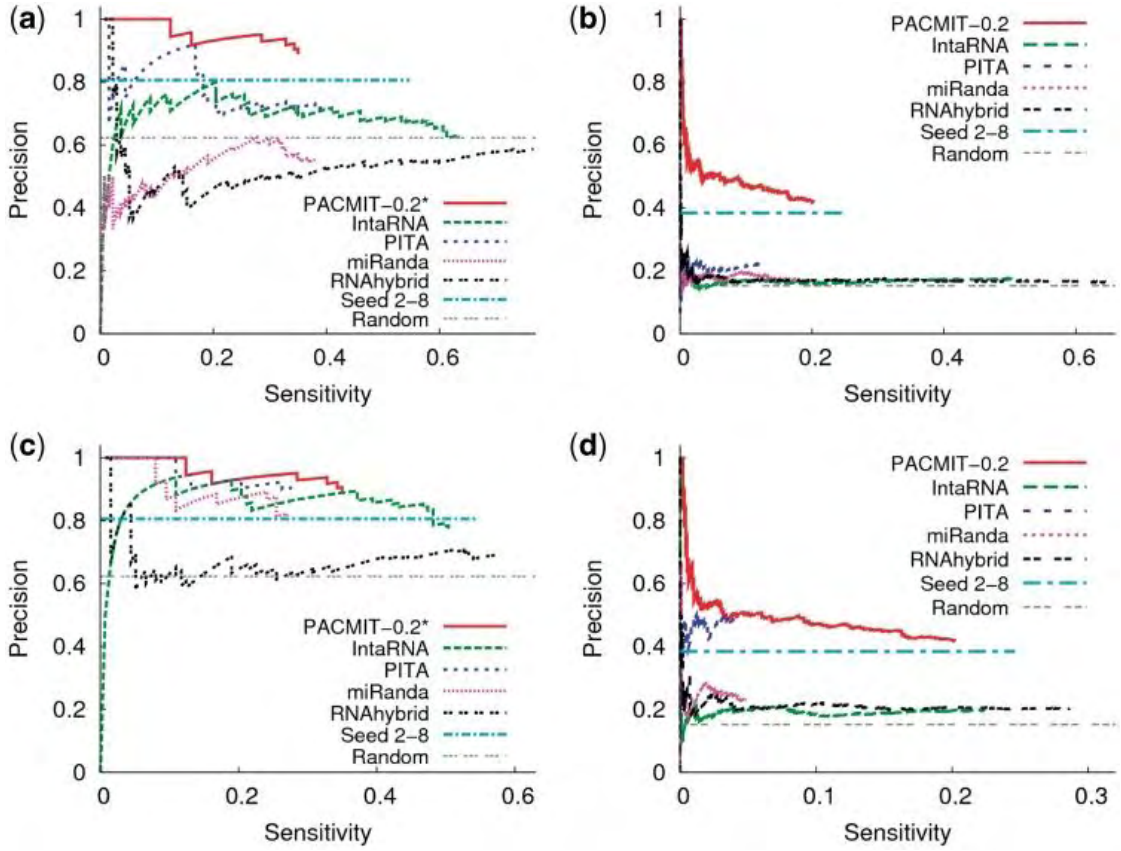
με τέσσερις γραφικές παραστάσεις, η πρώτη εκ των οποίων δείχνει το precision έναντι sensitivity στην Drosophila Melanogaster για τα τέσσερα κριτήρια ταξινόμησης και η δεύτερη το ίδιο στον άνθρωπο. Η τρίτη και η τέταρτη γραφική παράσταση δείχνουν τον αριθμό των σωστά θετικών στις top προβλέψεις για Drosophila Melanogaster και άνθρωπο αντίστοιχα. Τα παραπάνω φαίνονται στην εικόνα 2.16.



Εικόνα 2.16: Σύγκριση των τεσσάρων κριτηρίων ταξινόμησης στην Drosophila Melanogaster και στον άνθρωπο (Πηγή: Marin και Vanicek, 2011)

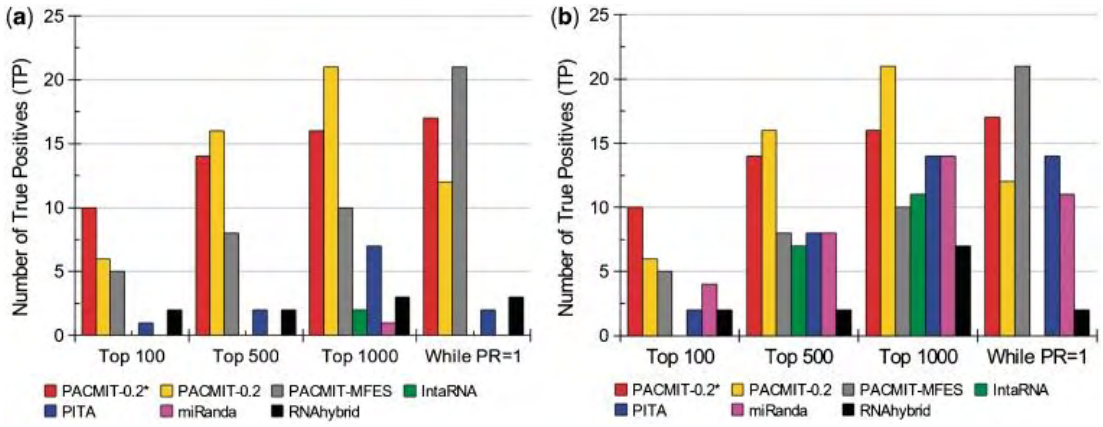
Έπειτα συγκρίνουν τον PACMIT με δύο αλγορίθμους που χρησιμοποιούν την προσβασιμότητα (accessibility) έναντι του conservation για να αυξήσουν την ακρίβεια των προβλέψεων τους, συγκεκριμένα με τους PITA και IntaRNA. Επίσης, με δύο ακόμα αλγορίθμους που χρησιμοποιούν hybridization energy, τον miRanda και τον RNAhybrid. Συμπεριέλαβαν επίσης το seed 2-8 σύμφωνα με το οποίο όλα τα 3΄UTR με ένα τουλάχιστον τέλειο ταίριασμα θεωρούνται στόχοι, καθώς και το τυχαίο (Random) το οποίο υποδηλώνει την απουσία οποιουδήποτε αλγορίθμου. Έκαναν δύο συγκρίσεις την πρώτη με default παραμέτρους, ενώ την δεύτερη με τις βέλτιστες παραμέτρους που βρήκαν οι ίδιοι ελέγχοντας τους άλλους αλγορίθμους. Αυτές τις

παραμέτρους τις ονομάζουν optimized. Στην εικόνα 2.17 φαίνονται τα αποτελέσματα. Η πρώτη και η δεύτερη γραφική παράσταση δείχνουν το precision έναντι sensitivity στην Drosophila Melanogaster και στον άνθρωπο αντίστοιχα, με τις default παραμέτρους, ενώ η τρίτη και η τέταρτη με τις optimized παραμέτρους. Και στους δύο οργανισμούς ο PACMIT πετυχαίνει τα καλύτερα αποτελέσματα. Ειδικά, όσον αφορά τον άνθρωπο το precision είναι αρκετά μεγαλύτερο από το precision διαφορετικών μεθόδων. Με optimized παραμέτρους ο PACMIT συνεχίζει να έχει καλύτερα αποτελέσματα, όμως με μικρότερη διαφορά.



Εικόνα 2.17: Σύγκριση του PACMIT με άλλους αλγορίθμους έχοντας σαν κριτήριο το precision και το sensitivity (Πηγή: Marin και Vanicek, 2011)

Πραγματοποιούν μια ακόμα σύγκριση που αφορά τον αριθμό των σωστά θετικών ανάμεσα στις top προβλέψεις στην Drosophila Melanogaster. Παρατίθενται δύο γραφήματα το ένα με τις default παραμέτρους, ενώ το δεύτερο με τις optimized. Ο PACMIT πετυχαίνει τα καλύτερα αποτελέσματα και σε αυτήν την περίπτωση. Το διάγραμμα φαίνεται στην εικόνα 2.18.

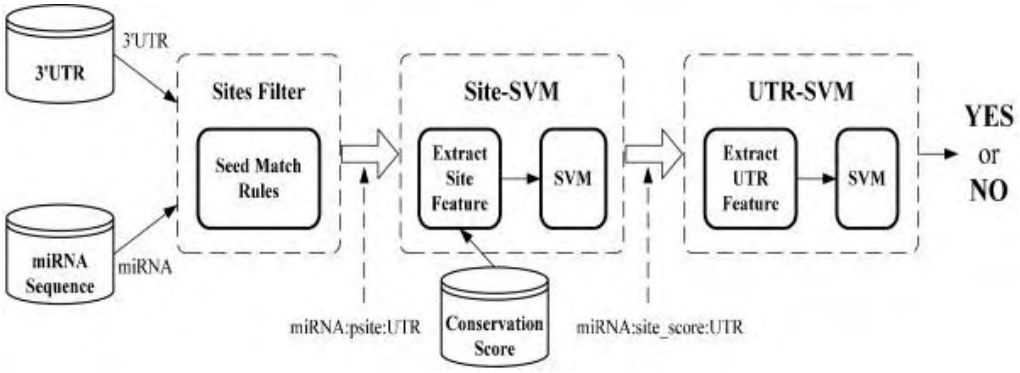


Εικόνα 2.18: Σύγκριση του PACMIT με άλλους αλγορίθμους έχοντας σαν κριτήριο τον αριθμό των σωστά θετικών στις top προβλέψεις (Πηγή: Marin και Vanicek, 2011)

Τέλος, παρατήρησαν ότι ο PACMIT πετυχαίνει την μικρότερη κατανάλωση CPU, συγκρίνοντας την κατανάλωση CPU των διάφορων αλγορίθμων στον οργανισμό Drosophila Melanogaster. Επομένως ο PACMIT είναι και υπολογιστικά πιο αποτελεσματικός από τους υπόλοιπους αλγορίθμους.

2.6 SVMicrO

Ο SVMicrO (Liu et al., 2010) είναι ένας αλγόριθμος πρόβλεψης στόχων miRNA βασισμένος σε SVM (Support Vector Machine) δύο σταδίων. Επιδιώκει να προβλέψει στόχους θηλαστικών miRNA. Έχει τρία στάδια. Πρώτα, εφαρμόζεται ένα site filter που χρησιμοποιεί την ακολουθία miRNA και ανιχνεύει την 3'UTR ακολουθία, ώστε να βρεθούν πιθανά binding sites. Στόχος είναι να επιλεχθούν sites με υψηλή ευαισθησία, αφού από αυτό το στάδιο εξαρτάται η ευαισθησία του αλγορίθμου. Ύστερα, τα πιθανά sites που έχουν φιλτραριστεί υποβάλλονται στο Site-SVM, το οποίο εξάγει features για κάθε site και βάζει ένα σκορ για να δείξει την πιθανότητα ένα site να είναι πραγματικό. Τέλος, τα site scores μαζί με UTR χαρακτηριστικά υποβάλλονται στο UTR-SVM, για να παραχθεί η τελική πρόβλεψη του UTR σαν στόχος. Αυτή η διαδικασία φαίνεται στην εικόνα 2.19.



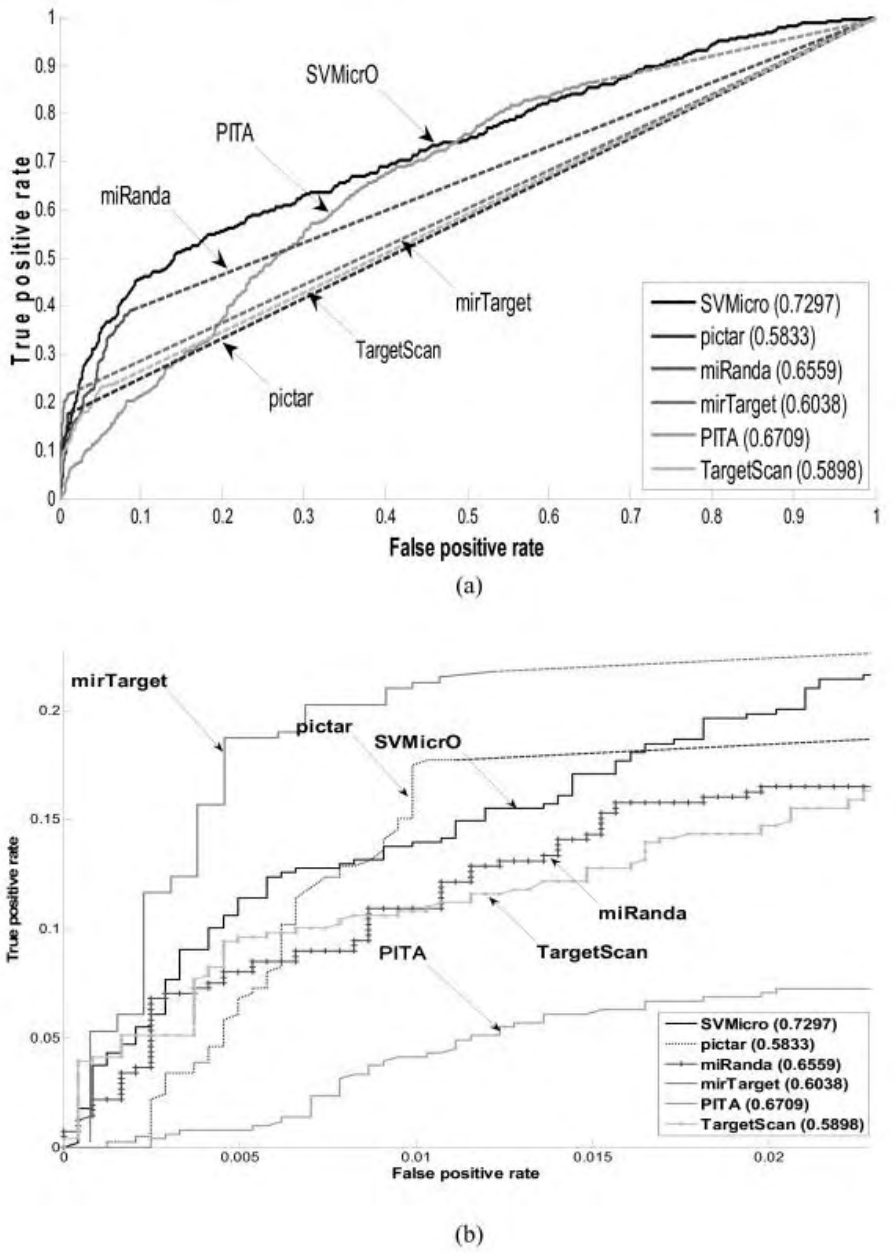
Εικόνα 2.19: Λειτουργία του αλγορίθμου SVMicrO (Πηγή: Liu et al., 2010)

Τα θετικά δεδομένα που χρησιμοποίησαν για training τα πήραν από την βάση δεδομένων miRecords. Για τα αρνητικά δεδομένα μάζεψαν 20 miRNA over-expression microarray δεδομένα από NCBI Gene Expression Omnibus. Από αυτά διάλεξαν τα πιο σίγουρα up-regulated γονίδια περιορίζοντας την διαφορική expression p-value, την fold change και την συνοχή των samples σταδιακά. Ύστερα τα αρνητικά ζευγάρια περνάνε από το site filter. Στο τέλος βρίσκουν 3542 αρνητικά ζευγάρια miRNA-mRNA.

Ο SVMicrO κάνει προβλέψεις βασισμένος σε 21 βέλτιστα site χαρακτηριστικά και 18 βέλτιστα UTR χαρακτηριστικά. Αυτά προέκυψαν από μια μεγάλη συλλογή 113 site

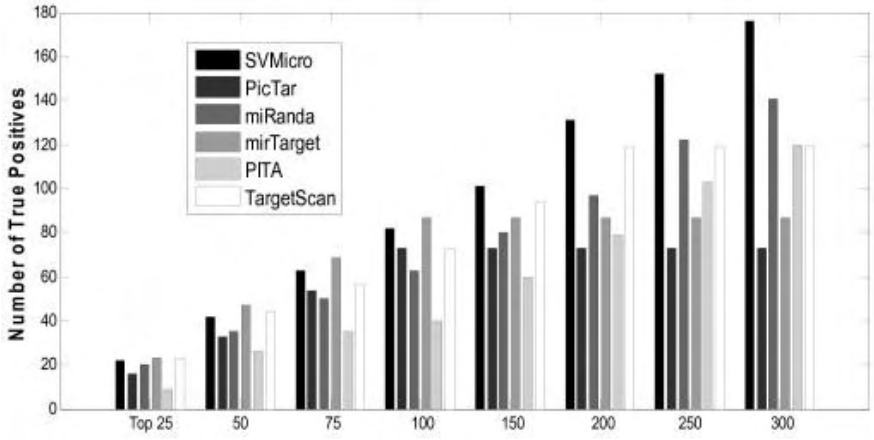
και 30 UTR χαρακτηριστικών. Για κάθε site και UTR-SVM το RBF διαλέγεται σαν kernel λειτουργία. Γίνεται 5-fold cross validation για να γίνουν training οι παράμετροι και να διαλεχτούν τα βέλτιστα χαρακτηριστικά και για τα δύο SVM. Σε κάθε γύρο του cross validation, ένας sequential forward search αλγόριθμος εφαρμόζεται για επιλογή χαρακτηριστικών βασισμένος σε ταξινομημένα χαρακτηριστικά που παίρνει σαν έξοδο από τον redundancy maximal relevance (mRMR) αλγόριθμο. Ο mRMR αλγόριθμος έχει σχεδιαστεί για να διαλέξει ένα υποσύνολο χαρακτηριστικών. Τελικά, διαλέγονται 21 βέλτιστα site χαρακτηριστικά και 18 βέλτιστα UTR χαρακτηριστικά.

Για να συγκρίνουν τον SVMicrO με άλλους γνωστούς αλγορίθμους πρόβλεψης στόχων miRNA παραθέτουν μια καμπύλη ROC βασισμένη στα training δεδομένα, που δείχνει την απόδοση των SVMicrO , PITA, TargetScan, miRanda, MirTarget και PicTar. Εκτός από το PITA για το οποίο οι προβλέψεις αποκτήθηκαν από τον παρεχόμενο αλγόριθμο, στα υπόλοιπα το TPR και FPR υπολογίστηκε βασιζόμενο σε προβλέψεις δημοσιευμένες στα website τους. Συνολικά, ο SVMicrO έχει το μεγαλύτερο AUC (area under the curve) και την καλύτερη ROC. Ο PITA που έχει την δεύτερη καλύτερη AUC έχει την χειρότερη απόδοση σε FPR. Στην εικόνα 2.20 παρατίθενται δύο διαγράμματα, το δεύτερο είναι μια μεγέθυνση του πρώτου.



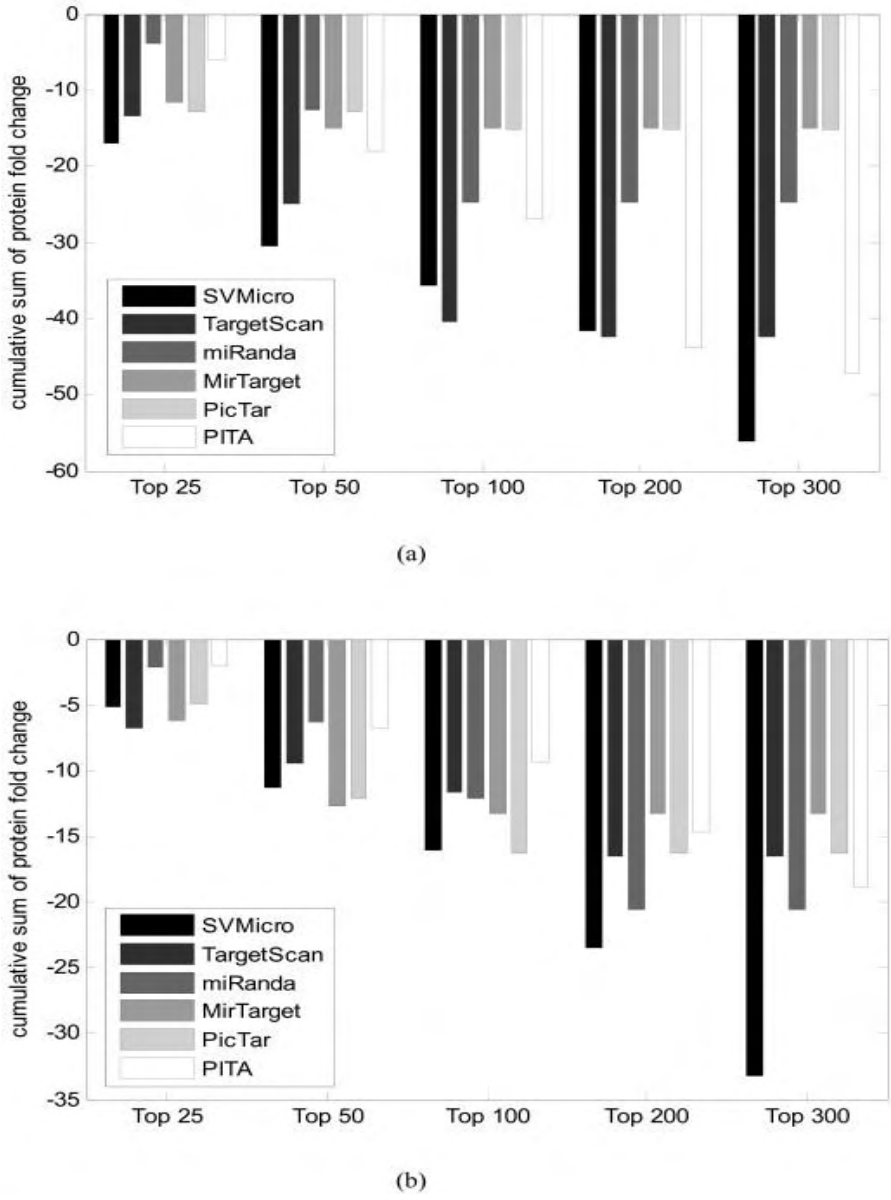
Εικόνα 2.20: Διάγραμμα ROC για την σύγκριση του SVMicrO με λοιπούς αλγορίθμους (Πηγή: Liu et al., 2010)

Στην συνέχεια μετράνε την ακρίβεια (τον αριθμό των σωστά θετικών στο σύνολο των προβλέψεων) κάθε αλγορίθμου. Στις 100 top προβλέψεις ο SVMicrO και ο mirTarget έχουν περίπου ίδια αποτελέσματα, ενώ όταν πηγαίνουμε πάνω από 150 προβλέψεις ο SVMicrO έχει αρκετά περισσότερα σωστά θετικά από τους υπόλοιπους αλγορίθμους. Τα αποτελέσματα φαίνονται στην εικόνα 2.21.



Εικόνα 2.21: Σύγκριση του SVMicrO με λοιπούς αλγορίθμους με βάση τον αριθμό των σωστά θετικών προβλέψεων (Πηγή: Liu et al., 2010)

Για να πραγματοποιήσουν μια ακόμα σύγκριση βασίστηκαν σε high throughput proteomics δεδομένα (Baek et al.,2008 και Selbach et al.,2008). Σε αυτά τα δύο papers η fold αλλαγή της πρωτεΐνης εξαιτίας του over-expression κάποιου συγκεκριμένου miRNA υπολογίσθηκε από stable-isotope-labelling-of-amino-acids-in culture (SILAC). Κανένα γονίδιο δεν θεωρείται επακριβώς στόχος στα paper, όμως είναι λογικό να θεωρηθεί ότι όσο μεγαλύτερο το down-fold μιας πρωτεΐνης, τόσο πιθανότερο το αντίστοιχο γονίδιο να είναι πραγματικός στόχος. Στην εικόνα 2.22 φαίνεται η Cumulative Fold Change (CFC) για τα γονίδια miR-124 (a) και miR-1 (b) στις πρώτες 300 προβλέψεις. Όσον αφορά το miR-124, ο SVMicrO και ο TargetScan έχουν την καλύτερη απόδοση. Όσον αφορά το miR-1, μέχρι τις πρώτες 200 προβλέψεις αρκετοί αλγόριθμοι έχουν παρόμοια αποτελέσματα, όμως πάνω από τις 200 ο SVMicrO φαίνεται να πετυχαίνει καλύτερη απόδοση.



Εικόνα 2.22: Σύγκριση του SVMicrO με βάση την Cumulative Fold Change στις πρώτες 300 προβλέψεις (Πηγή: Liu et al., 2010)

3. Ab Initio αλγόριθμοι πρόβλεψης στόχων miRNA

3.1 miRanda

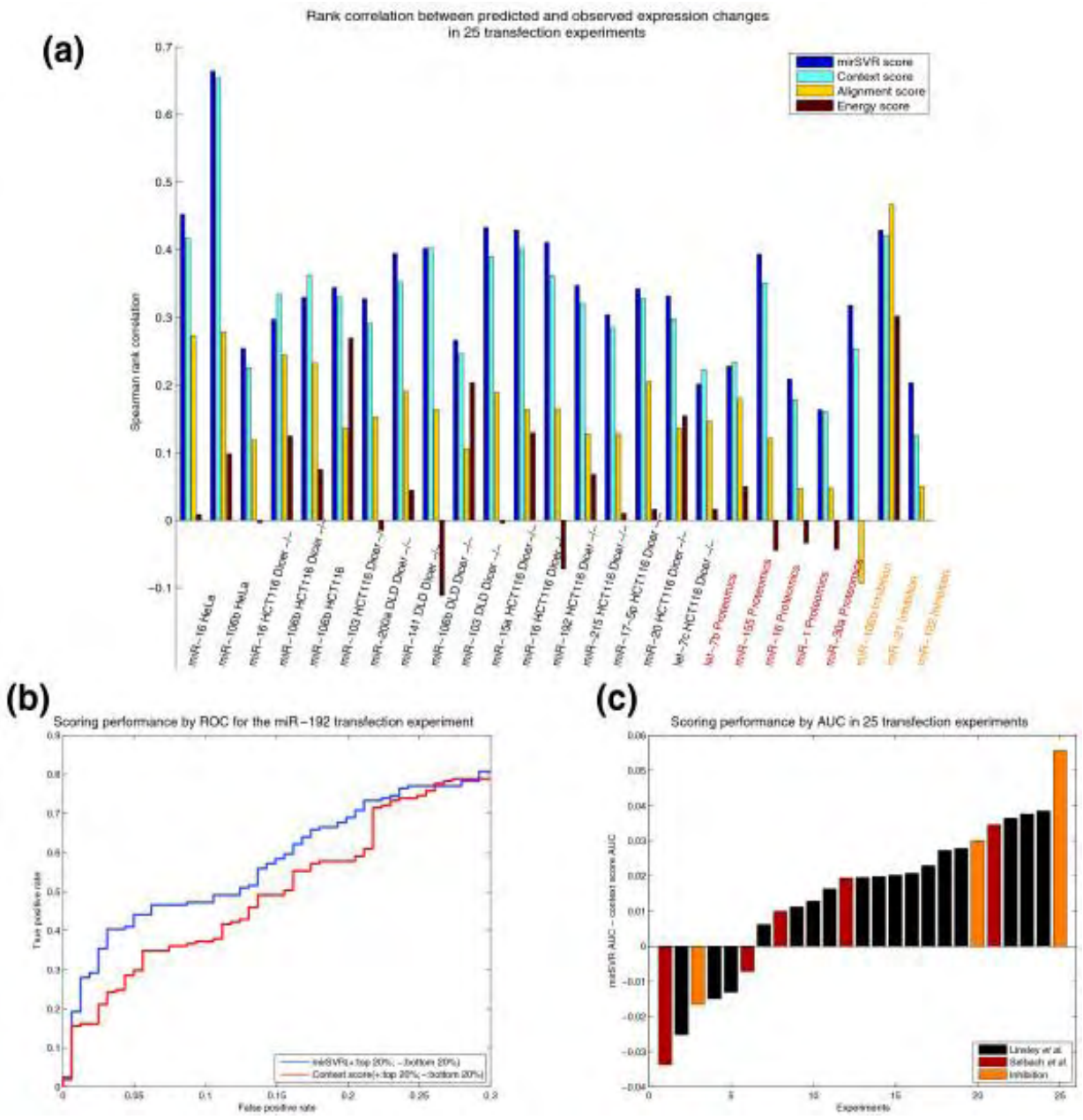
Ο miRanda (Enright et al.,2003) είναι ένας από τους παλιότερους και πιο σημαντικούς αλγόριθμους πρόβλεψης στόχων miRNA. Αρχικά, δημιουργήθηκε με σκοπό να βρίσκει στόχους στην Drosophila Melanogaster (Enright et al.,2003), στην συνέχεια όμως ανανεώθηκε και για τον άνθρωπο (John et al.,2004). Αν και ο miRanda είναι διαθέσιμος online σαν μέρος του miRanda-mirSVR, για να χρησιμοποιηθεί πλέον μόνος του πρέπει να γίνει download.

Ο miRanda βασίζεται στον υπολογισμό της ιδανικής ακολουθιακής συμπληρωματικότητας (sequence complementarity), μεταξύ ενός συνόλου από ώριμα microRNA και ενός δοθέντος mRNA χρησιμοποιώντας έναν weighted αλγόριθμο δυναμικού προγραμματισμού. Η ενέργεια δεσίματος (binding energy) του duplex structure, το εξελικτικό conservation ολόκληρου του site του στόχου και η θέση του μέσα στο 3΄UTR υπολογίζονται και λογαριάζονται για τα τελικά αποτελέσματα (alignment score) που είναι ένα σταθμισμένο (weighted) σύνολο από σκορ ταιριασμάτων και όχι ταιριασμάτων και ποινών για κενά (gap penalties). Σαν ένα δεύτερο φίλτρο χρησιμοποιούν μια εκτίμηση της ελεύθερης ενέργειας (free energy) της διάταξης του microRNA:mRNA duplex χρησιμοποιώντας το Vienna package. Σαν τελευταίο φίλτρο χρησιμοποιείται το conservation. Για να φιλτράρουν τους λιγότερο conserved προβλεφθέντες στόχους χρησιμοποίησαν το PhastCons conservation score, που υπολογίζει το εξελικτικό conservation ακολουθιακών κομματιών χρησιμοποιώντας ένα φυλογενετικό hidden Markov model.

Πρόσφατα, έγινε μια σημαντική βελτίωση στον miRanda με την δημιουργία ενός αλγορίθμου βασισμένου σε support vector regression (SVR), του mirSVR, ο οποίος κάνει ανάθεση σκορ και ταξινόμηση στις αλληλεπιδράσεις miRNA-στόχων που προέρχονται από τον αλγόριθμο miRanda. Είναι μια μέθοδος μηχανικής μάθησης και ταξινομεί τους στόχους των microRNA με ένα downregulation σκορ. Ένα από τα κυριότερα πλεονεκτήματά του είναι ότι αναγνωρίζει στόχους που δεν είναι conserved και δεν είναι canonical, (δηλαδή sites με τέλεια seed συμπληρωματικότητα). Ο mirSVR εκπαιδεύτηκε με 9 miR transfection πειράματα που έγιναν σε κύτταρα Hela και ενσωματώνει ένα σύνολο από χαρακτηριστικά που βρήκαν σχετικά όπως site

accessibility, AU flanking content, τα base pairings στην περιοχή seed και το μήκος UTR. Πραγματοποίησαν δύο διαφορετικούς τρόπους εκπαίδευσης για το mirSVR, το πρώτο εκ των οποίων κάνει training μόνο σε γονίδια που περιέχουν ένα απλό canonical site στο 3΄UTR και λέγεται μοντέλο canonical-only. Στο δεύτερο γίνεται training σε γονίδια που περιέχουν ένα απλό canonical ή μη canonical site στο 3΄UTR, όπου επιτρέπονται μη canonical sites με ακριβώς ένα G:U wobble η mitsmatch σε ένα εξαμερές στην περιοχή seed και λέγεται μοντέλο all-sites. Το πρώτο παράγει ένα μοντέλο που είναι εύκολα συγκρίσιμο με τους περισσότερους υπάρχοντες αλγορίθμους πρόβλεψης στόχων miRNA, ενώ το δεύτερο επιτρέπει να δούμε αν είναι δυνατό να επιτευχθούν αξιόλογα αποτελέσματα σε μη canonical sites.

Πραγματοποιούν μια σύγκριση χρησιμοποιώντας το μοντέλο canonical-only. Σύγκριναν τον mirSVR με άλλους αντιπροσωπευτικούς αλγορίθμους πρόβλεψης στόχων miRNA διαφορετικών μεθοδολογιών: το context score του TargetScan, το alignment score του miRanda και το energy score του PITA. Για να αξιολογήσουν την απόδοση, υπολόγισαν το Spearman rank correlation μεταξύ της παρατηρούμενης log expression αλλαγής και του σκορ πρόβλεψης, που δίνει ένα γενικό μέτρο της απόδοσης ταξινόμησης κάθε αλγορίθμου. Στην συγκεκριμένη έρευνα δεν φιλτράρανε τα πιθανά target sites με κριτήριο το conservation, λόγω του ότι δεν θέλουν να πραγματοποιήσουν μια τυπική σύγκριση των υπαρχόντων προγραμμάτων, όπως υπάρχουν στους web servers, αλλά θέλουν να αξιολογηθεί η αξία διαφορετικών μεθόδων scoring να υπολογίσουν την έκταση του microRNA regulation. Οι miRanda και PITA δεν έχουν εκπαιδευτεί σε genome-wide expression δεδομένα και συνεπώς είναι λογικό ο mirSVR να πετυχαίνει καλύτερη απόδοση. Στην σύγκριση όμως με το context σκορ του TargetScan, ο mirSVR έχει καλύτερη απόδοση στα 21 από τα 25 σετ δεδομένων. Επίσης, μέτρησαν την απόδοση του mirSVR και του context σκορ με μια ROC ανάλυση ταξινομώντας το πρώτο 20% των περισσότερο downregulated στόχων (το οποίο όρισαν true positives) και το πρώτο 20% των λιγότερο downregulated στόχων (το οποίο όρισαν true negatives) για το miR-192. Τα αποτελέσματα φαίνονται στην εικόνα 3.1, όπου το (a) δείχνει το Spearman rank correlation, το (b) το ROC διάγραμμα για το miR-192 και (c) μια σύνοψη της ROC ανάλυσης και για τα 25 σετ δεδομένων υπολογίζοντας το AUC για το mirSVR και το context score. Το mirSVR έχει μεγαλύτερη AUC σε 19 από τα 25 σετ δεδομένων.



Εικόνα 3.1: Σύγκριση του mirSVR με το alignment score του miRanda, το context score του TargetScan και το energy score του PITA (Πηγή: Betel et al., 2010)

Το interface του miRanda-mirSVR φαίνεται στην εικόνα 3.2.

Εικόνα 3.2: Interface του αλγορίθμου miRanda-mirSVR (http://www.microrna.org/microrna/home.do)

3.2 TargetScan

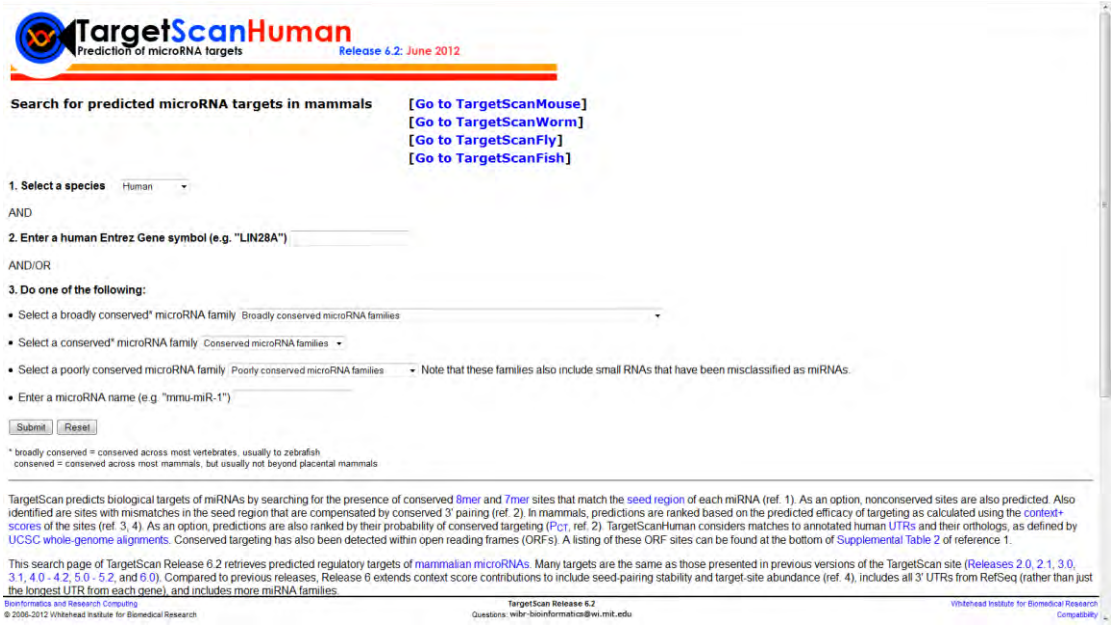
Ο TargetScan είναι από τους παλαιότερους και σημαντικότερους αλγόριθμους πρόβλεψης στόχων miRNA και ήδη έχει συμπεριληφθεί σχεδόν σε όλες τις έρευνες που παρουσιάστηκαν προηγουμένως. Αποτελεί ένα online λογισμικό που παρέχεται από το πανεπιστήμιο του MIT. Ο TargetScan δίνει την δυνατότητα στον χρήστη να ψάξει με βάση το miR όνομα, το όνομα του γονιδίου ή από broadly conserved, conserved ή poorly conserved οικογένειες miR μεταξύ διάφορων ειδών. Broadly conserved miRNA families θεωρούνται αυτές που είναι conserved στα περισσότερα σπονδυλωτά, ενώ conserved miRNA families αυτές που είναι conserved στα περισσότερα θηλαστικά, αλλά συνήθως όχι στα πλακούντια. Σαν poorly conserved families θεωρούνται όλα τα υπόλοιπα (Friedman et al.,2009) Ο TargetScan έχει εκπαιδευτεί να ανιχνεύει στόχους στην 3΄UTR περιοχή. Αν και τα περισσότερα miRNA Recognition Elements (MREs) βρίσκονται σε αυτήν την περιοχή, κάποιοι λειτουργικοί στόχοι έχουν βρεθεί και στην coding sequence (CDS).

Η ταξινόμηση των στόχων γίνεται πλέον με βάση το context score + είτε την πιθανότητα conserved targeting (Pct). Τα context+ scores χρησιμοποιήθηκαν το 2011 στην θέση του προηγούμενου κριτηρίου που ονομάζεται context scores. Τα context scores χρησιμοποιήθηκαν για πρώτη φορά το 2007 στην έκδοση TargetScan 4.0 για να υπάρχει καλύτερο specificity στην πρόβλεψη στόχων. Τα context scores υπολογίζονται με βάση τα εξής τέσσερα χαρακτηριστικά: site type συνεισφορά η οποία εκφράζει το είδος του seed ταιριάσματος (8-mer, 7-mer, 7mer-1A), 3' pairing συνεισφορά, η οποία εκφράζει την miRNA-target συμπληρωματικότητα έξω από την seed περιοχή, την τοπική AU συνεισφορά, η οποία εκφράζει την μεταγραφή AU upstream και downstream του προβλεπόμενου site και τέλος συνεισφορά θέσης που εκφράζει την απόσταση από το κοντινότερο τέλος ενός annotated UTR του στόχου. Το context score είναι το άθροισμα αυτών των τεσσάρων χαρακτηριστικών και το context score percentile είναι η ποσοστιαία ταξινόμηση κάθε site, συγκρινόμενη με όλα τα sites για την συγκεκριμένη miRNA οικογένεια. Όσο πιο αρνητικό είναι το σκορ, τόσο ευνοϊκότερο το site. Σε ένα γονίδιο με πολλά sites για μια miRNA family, το συνολικό context score υπολογίζεται σαν το άθροισμα των context scores των πιο ευνοϊκών (αρνητικών) miRNA σε αυτή την οικογένεια. Αν το context score για κάποιο από αυτά τα sites είναι θετικό, η συνεισφορά στο τελικό αποτέλεσμα είναι

0. Το representative miRNA είναι το miRNA με το ευνοϊκότερο συνολικό context σκορ.

Τα context+ scores (Garcia et al.,2011) περιλαμβάνουν τα τέσσερα χαρακτηριστικά των context score που αναφέρθηκαν προηγουμένως και επίσης τα εξής: την αφθονία των target-site (TA) και την σταθερότητα seed-pairing. Τα δύο καινούργια χαρακτηριστικά συμπεριλήφθηκαν για να αντιμετωπιστεί το φαινόμενο των miRNA που παρουσιάζουν χαμηλή επίδοση. Αυτά τα miRNA επιδεικνύουν μειωμένη ικανότητα καταστολής (repressing), ακόμα και σε στόχους με seed ταίριασμα 7 ή 8 νουκλεοτιδίων. Το συγκεκριμένο υποσύνολο από miRNA έχει ασυνήθιστα AU-rich seed περιοχές, που μειώνουν την σταθερότητα των αλληλεπιδράσεων seed-pairing και επίσης επιδεικνύουν αυξημένη target-site αφθονία. Όσον αφορά την πιθανότητα conserved targeting (Pct), αρχικά υπολόγισαν την signal-to-background αναλογία (S/B) για κάθε site σε διαφορετικά τμήματά του. Έπειτα, μετέτρεψαν την (S/B) στην πιθανότητα conserved targeting με τον εξής τύπο: (S/B - 1)/(S/B). Αυτό το σκορ αντανακλά την Bayesian εκτίμηση της πιθανότητας ότι ένα site είναι conserved επιλεκτικά και όχι τυχαία ή για κάποιον άλλο λόγο. Τέλος, ο τύπος για το aggregate (συνολικό) Pct είναι ο εξής: $1 - ((1 - P_{CT})_{site1} \times (1 - P_{CT})_{site2} \times (1 - P_{CT})_{site3} \ldots)$.

Το interface του TargetScan φαίνεται στην εικόνα 3.3.



Εικόνα 3.3: Interface του αλγορίθμου TargetScan (http://www.targetscan.org/)

3.3 DIANA-microT

Ο αλγόριθμος DIANA-microT είναι ένας από τους σημαντικότερους αλγορίθμους πρόβλεψης στόχων miRNA. Αυτή τη στιγμή είναι ενεργές δύο εκδόσεις του. Ο DIANA-microT-CDS αποτελεί την πέμπτη έκδοση του DIANA-microT. Έχει εκπαιδευτεί σε ένα θετικό και ένα αρνητικό σετ από miRNA Recognition Elements (MREs), τα οποία έχουν εντοπιστεί και στην 3΄UTR και στην CDS περιοχή. Με την συγκεκριμένη έκδοση υπάρχει μια σημαντική αύξηση στο sensitivity σε σχέση με την προηγούμενη έκδοση του αλγορίθμου. Η DIANA-microT v4 είναι η τέταρτη έκδοση του microT αλγορίθμου. Ο server έχει προβλέψεις για τέσσερα είδη και υποστηρίζει το miRBase 18 και την Ensembl 69. Υπολογίζονται miRNAστόχοι στα εξής είδη: Homo sapiens, Mus musculus, Drosophila melanogaster και C.elegans.

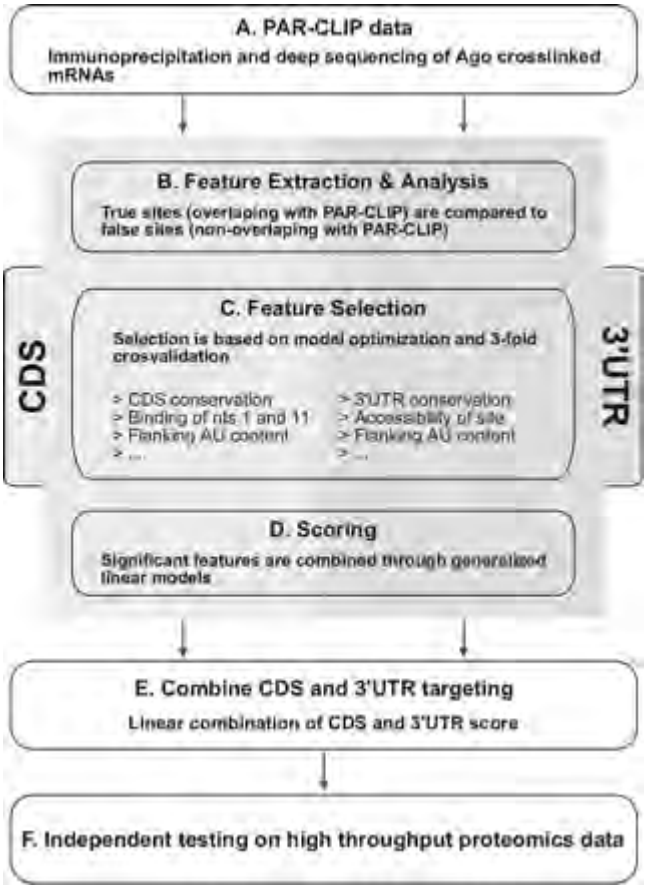
Ο αλγόριθμος DIANA-microT αναπτύχθηκε από το εργαστήριο DIANA. Το συγκεκριμένο εργαστήριο έχει στόχο να παρέχει αλγορίθμους, βάσεις δεδομένων και εργαλείων για την ερμηνεία και την αρχειοθέτηση γονιδιακών δεδομένων. Ορισμένα από αυτά είναι τα microT v4 και microT-CDS που αναφέρθηκαν προηγουμένως, τα TarBase και LncBase και το mirPath.

Ο αλγόριθμος αυτός είναι βασισμένος σε παραμέτρους που υπολογίζονται ξεχωριστά για κάθε miRNA και για κάθε miRNA recognition element (MRE), με βάση binding και conservation επίπεδα. Το συνολικό προβλεπόμενο σκορ ενός πιθανού στόχου είναι το σταθμισμένο άθροισμα των conserved και unconserved MREs ενός γονιδίου. Επίσης, υπολογίζεται μια αναλογία signal-to-noise (SNR) και ένα σκορ ακρίβειας για κάθε αλληλεπίδραση, το οποίο μπορεί να χρησιμοποιηθεί για να γίνει μια εικασία του false positive rate για κάθε miRNA targeted gene (miTG) (Maragkakis et al.,2009). Τα παραπάνω μπορούν να βρεθούν στον web server του DIANA-microT που τώρα βρίσκεται στην έκδοση 5.0.

Στην παρούσα στιγμή ο DIANA-microT-CDS είναι ένας από τους λίγους αλγορίθμους που υποστηρίζουν την αναζήτηση στόχων miRNA μαζί στην 3′ UTR και στην CDS περιοχή. Αυτό είναι πολύ σημαντικό, καθώς υπάρχουν όλο και περισσότερα στοιχεία ότι τα miRNA δένουν και στην CDS περιοχή (Forman et al.,2008, Huang et al.,2010, Qin et al.,2010, Ott et al.,2011), με αποτέλεσμα να υπάρχει μια μείωση στο sensitivity για τους αλγορίθμους που ψάχνουν μόνο την 3΄UTR περιοχή. Η εφαρμογή high throughput προσεγγίσεων για την απομόνωση

Argonaute-bound στόχων υποδεικνύει ότι τα CDS sites είναι εξίσου πολυάριθμα με αυτά που βρίσκονται στην 3΄UTR περιοχή, αν και με μεγαλύτερη πυκνότητα στην τελευταία (Chi et al.,2009, Hafner et al.,2010). High throughput proteomics πειράματα που μετράνε τις αλλαγές για χιλιάδες γονίδια δείχνουν ότι περίπου τα μισά γονίδια που η έκφραση (expression) τους μεγαλώνει ή μικραίνει λόγω miRNA (transfection) δεν έχουν ένα miRNA seed ταίριασμα στην 3΄UTR περιοχή. Ο DIANA-microT-CDS αναγνωρίζει 12% από αυτά τα downregulated γονίδια σαν επιπλέον στόχους miRNA, έχοντας τους στόχους στην κωδική περιοχή (CDS).

Ο αλγόριθμος αυτός έχει εκπαιδευτεί σε ένα θετικό και ένα αρνητικό σετ από MREs ορισμένο από PAR-CLIP δεδομένα του Hafner et al. (2010). Αρχικά, τα MREs που προσδιορίζονται από τα PAR-CLIP δεδομένα χωρίζονται σε δύο κατηγορίες ανάλογα με την περιοχή στην οποία ανήκουν (CDS και 3΄UTR). Για αυτά τα δύο σετ δεδομένων διάφορα χαρακτηριστικά εξάγονται και τα πιο ενημερωτικά από αυτά επιλέγονται, συγκρίνοντας θετικά με αρνητικά MREs. Κατόπιν, η επιλογή γίνεται χρησιμοποιώντας three-fold cross-validation. Για κάθε αναγνωρισμένο mRNA MRE τα επιλεγόμενα χαρακτηριστικά συνδυάζονται σε ένα MRE σκορ μέσα από γενικευμένα γραμμικά μοντέλα. Για κάθε γονίδιο το CDS σκορ και το 3΄UTR σκορ ορίζεται αθροίζοντας τα MRE σκορ που υπάρχουν στην CDS και 3΄UTR περιοχή αντίστοιχα. Αυτά τα δύο σκορ συνδυάζονται γραμμικά, ώστε να δώσουν ένα τελικό σκορ. Η διαδικασία φαίνεται αναλυτικά στην εικόνα 3.4.

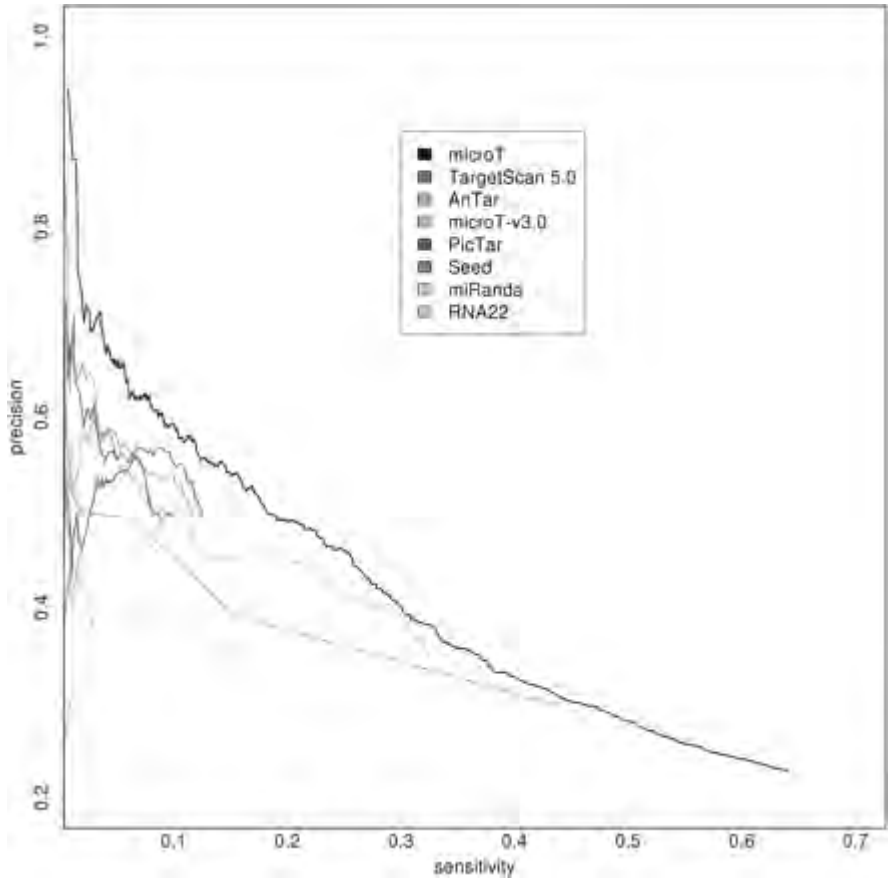


Εικόνα 3.4: Διάγραμμα ανάλυσης στα δεδομένα PAR-CLIP (Πηγή: Reczko et al., 2012)

Πραγματοποίησαν μια σύγκριση βασισμένη σε ένα μεγάλο ανεξάρτητο τεστ δεδομένων που παρέχεται από τον Selbach et al. (2008). Αυτό το σετ δεδομένων παρέχει τους πειραματικά επιβεβαιωμένους στόχους για 5 miRNA που αναγνωρίστηκαν με high-throughput μεθόδους. Περίπου τα μισά γονίδια που θεωρούνται στόχοι των συγκεκριμένων miRNA δεν έχουν ένα απλό ταίριασμα seed στις 3΄UTR sequences και γι' αυτό δεν αναγνωρίζονται από τα υπάρχοντα προγράμματα πρόβλεψης στόχων miRNA. Το DIANA-microT-CDS αυξάνει το sensitivity από 52% σε 65%, κρατώντας το specificity σε σταθερό επίπεδο στο 32%. Η προηγούμενη σύγκριση, η οποία είχε sensitivity 52%, είχε γίνει με παλαιότερη έκδοση του αλγορίθμου που έψαχνε μόνο στην 3΄UTR περιοχή.

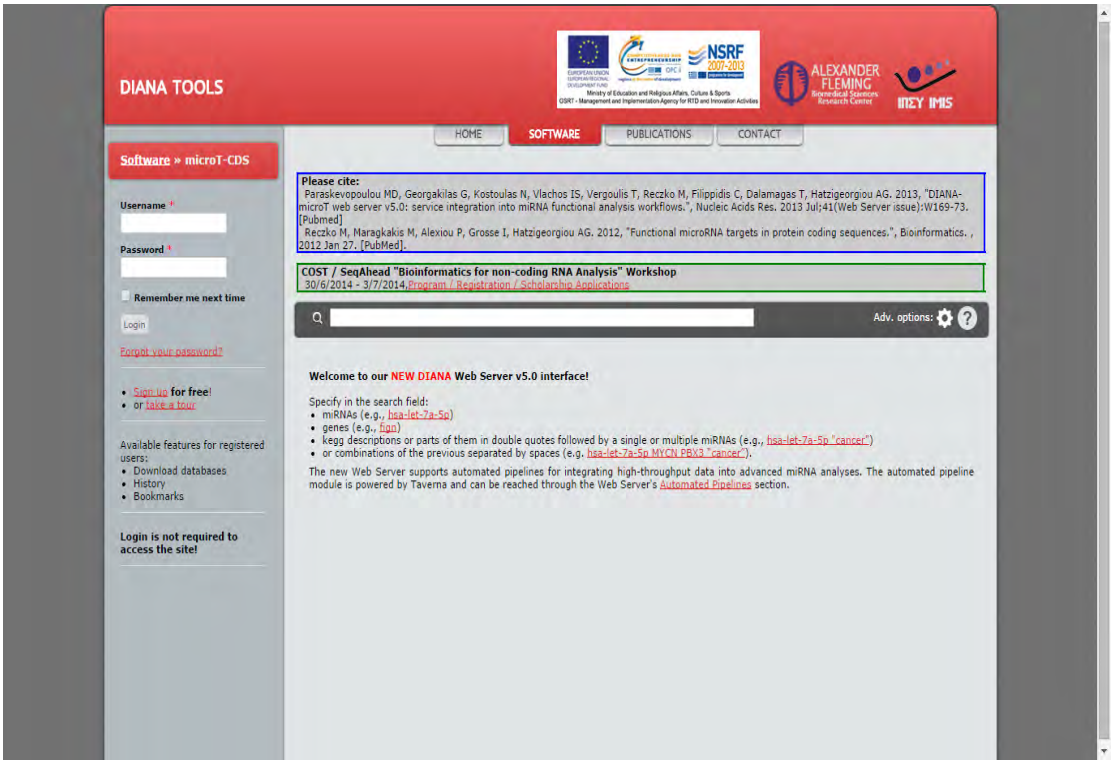
Σύγκριναν επίσης τον DIANA-microT-CDS με άλλους γνωστούς αλγορίθμους πρόβλεψης στόχων miRNA και συγκεκριμένα με τους TargetScan 5.0, PicTar, RNA22, AnTar, καθώς και μία απλή τεχνική seed, όπου το σκορ πρόβλεψης καθορίζεται από τον αριθμό των seed ταιριασμάτων στο 3΄UTR των γονιδίων. Το

sensitivity και το precision μετριέται σε διαφορετικά cutoffs. Ο DIANA-microT-CDS πετυχαίνει το μεγαλύτερο sensitivity σε οποιοδήποτε επίπεδο specificity σε σύγκριση με τους υπόλοιπους αλγορίθμους. Τέλος, παρατηρήθηκε μια μεγάλη αύξηση του sensitivity σε χαμηλότερες τιμές specificity ξεπερνώντας και την απλή τεχνική seed. Τα παραπάνω φαίνονται στην εικόνα 3.5.



Εικόνα 3.5: Διάγραμμα pROC (precision receiver operating curve) για την σύγκριση του DIANA-microT με λοιπούς αλγορίθμους (Πηγή: Reczko et al., 2012)

Το interface του DIANA-microT-CDS (Web Server v5.0) φαίνεται στην εικόνα 3.6.



Εικόνα 3.6: Interface του αλγορίθμου DIANA- microT-CDS (http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=microT_CDS/index)

4. Έρευνες

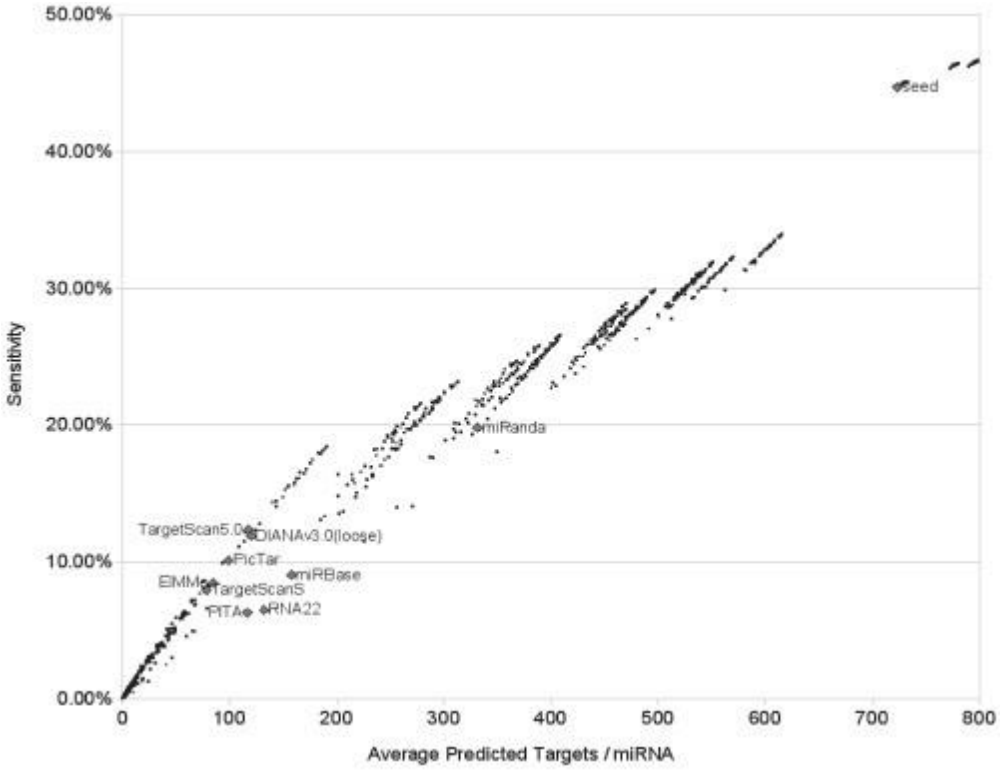
4.1. Συμπεράσματα από τις έρευνες σχετικά με την απόδοση των αλγορίθμων πρόβλεψης στόχων miRNA

Σε αυτή την εργασία παρουσιάστηκαν σε βάθος αρκετές έρευνες σχετικά με την αποδοτικότητα κάθε αλγορίθμου και την σύγκριση των αλγορίθμων μεταξύ τους. Οι αλγόριθμοι πρόβλεψης βάσεων δεδομένων είναι αρκετοί και τα πλεονεκτήματα κάποιου συγκεκριμένου αλγορίθμου δυσδιάκριτα. Αποτελεί σημαντικό ζήτημα για τους ερευνητές να βρεθεί ποιο είναι το καταλληλότερο εργαλείο, ώστε να το χρησιμοποιήσουν. Είναι αξιόπιστα τα αποτελέσματα των ερευνών που παρατίθενται; Οι περισσότερες έρευνες παρέχουν πειραματικά αποτελέσματα από διαφορετικά σετ δεδομένων και χρησιμοποιούν διαφορετικά κριτήρια αξιολόγησης. Υπάρχουν ελάχιστες ανεξάρτητες έρευνες που συγκρίνουν την απόδοση των αλγορίθμων και αυτές δεν είναι πρόσφατες, ώστε να εμπεριέχουν τους νέους αλγορίθμους που αναφέρθηκαν προηγουμένως, αλλά και τις καινούργιες βελτιωμένες εκδόσεις των παλαιότερων αλγορίθμων.

Μια επιλογή που γίνεται όλο και πιο διαδεδομένη είναι η χρήση πολλών αλγορίθμων και η χρήση των πιθανών ενώσεων ή διασταυρώσεων των συνδυασμών, ώστε να υπάρχουν πιο σίγουρα αποτελέσματα. Μάλιστα, αρκετά προγράμματα έχουν δημιουργηθεί αυτοματοποιώντας αυτή τη διαδικασία, όπως τα GOmir και miRecords. Άξιο να αναφερθεί είναι ότι το miRecords ενσωματώνει τους στόχους 11 διαφορετικών προγραμμάτων, των DIANA-microT, MicroInspector, miRanda, MirTarget2, miTarget, NBmiRTar, PicTar, PITA, RNA22, RNAhybrid και TargetScan/TargetScanS. Ενώ διαισθητικά μοιάζει λογικό οι συνδυασμοί αλγορίθμων πρόβλεψης γονιδίων στόχων miRNA να υπερέχουν των μεμονωμένων αλγορίθμων, μια έρευνα των Alexiou et al. (2009) αποδεικνύει το αντίθετο.

Στην συγκεκριμένη έρευνα υπολογίστηκαν όλοι οι πιθανοί συνδυασμοί ενώσεων και διασταυρώσεων μεταξύ των DIANA-microT v3.0, EIMMo, PITA, PicTar, RNA22, TargetScan 5.0, miRanda και miRBase. Η ένωση περιλαμβάνει όλους τους στόχους των προγραμμάτων που ανήκουν στην συγκεκριμένη ένωση, ενώ η διασταύρωση δύο ή και περισσοτέρων προγραμμάτων περιλαμβάνει στόχους που βρίσκουν όλα τα αντίστοιχα προγράμματα. Για να πραγματοποιήσουν την σύγκριση χρησιμοποίησαν

τα high-throughput δεδομένα, που παρέχονται για 5 miRNA από τους Selbach et al. (2008). Πολλοί συνδυασμοί λειτουργούν χειρότερα από μεμονωμένους αλγορίθμους. Τα αποτελέσματα φαίνονται στην εικόνα 4.1.



Εικόνα 4.1: Σύγκριση των συνδυασμών προγραμμάτων πρόβλεψης γονιδίων στόχων miRNA (Πηγή: Alexiou et al., 2009)

Οι έρευνες που έχουν ξεχωρίσει και έχουν χρησιμοποιηθεί από πολλούς ερευνητές είναι τέσσερις: των Sethupathy et al. (2006), των Baek et al. (2008), των Selbach et al. (2008) και των Alexiou et al. (2009). Η πρώτη έρευνα που αφορά την απόδοση των αλγορίθμων πρόβλεψης γονιδίων στόχων miRNA έγινε το 2006, από τους Sethupathy et al. Το specificity και το sensitivity υπολογίστηκαν χρησιμοποιώντας ένα σετ δεδομένων πειραματικά επικυρωμένων στόχων θηλαστικών από την TarBase. Πέντε αλγόριθμοι εξετάστηκαν σε αυτήν την έρευνα, συγκεκριμένα οι TargetScanS, PicTar, miRanda, TargetScan και DIANA-microT. Οι TargetScanS, PicTar και miRanda μόνοι ή σαν ένωση (οι στόχοι υποδεικνύονται από τουλάχιστον έναν αλγόριθμο) είχαν την καλύτερη απόδοση σε sensitivity και specificity, ενώ οι TargetScan και DIANA-microT δεν πέτυχαν καλή απόδοση. Επιπλέον, μόνο η

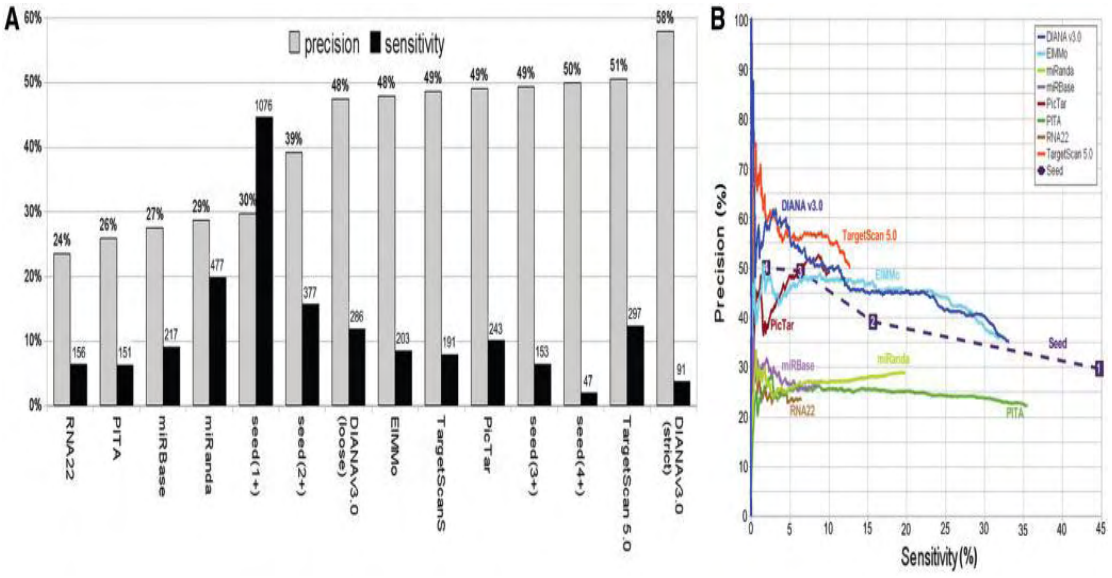
διασταύρωση μεταξύ TargetScanS και PicTar (οι στόχοι υποδεικνύονται και από τα δύο προγράμματα) είχε καλά αποτελέσματα, το οποίο μπορεί να εξηγηθεί από το ότι χρησιμοποιούν παρόμοια πολιτική για την πρόβλεψη στόχων miRNA. Ωστόσο, αυτή η έρευνα μπορεί να οδηγήσει σε λάθος συμπεράσματα. Το σετ δεδομένων από την TarBase δεν ήταν αρκετά μεγάλο και κυρίως είχε μόνο αλληλεπιδράσεις που ανακαλύφθηκαν με την χρήση προαποφασισμένων κανόνων για την πρόβλεψη στόχων miRNA. Επειδή ακόμα δεν έχουν οριστεί όλοι οι πιθανοί τύποι αλληλεπιδράσεων mRNA-miRNA, το σετ δεδομένων δεν μπορεί να θεωρηθεί ως αντιπροσωπευτικό (Witkos et al. 2011).

Οι Baek et al. (2008) εφάρμοσαν quantitative mass spectrometry χρησιμοποιώντας SILAC (stable isotope labeling with amino acids in cell culture) για να μετρήσουν την επίδραση του miR-223 στην παραγωγή των πρωτεϊνών στα neutrophils του ποντικιού. Η συγκεκριμένη έρευνα ασχολήθηκε με τους εξής αλγορίθμους: TargetScan (έκδοση 4.1), PicTar, miRanda (έκδοση Ιανουαρίου 2008), miRBase Targets (έκδοση 5), RNA22 και PITA. Η σύγκριση μεταξύ των in vivo και των in silico αποτελεσμάτων έδειξε ότι ο TargetScan και ο PicTar φαίνονται να είναι οι καλύτεροι και ότι μόνο το συνολικό context score του TargetScan συσχετίζεται με το downregulation των πρωτεϊνών. Αλγόριθμοι όπως ο RNA22 και μια πιο ανεκτική έκδοση του PITA που δεν χρησιμοποιούν site conservation δεν είχαν καλύτερη απόδοση από μια απλή αναζήτηση με 7μερή-8μερή seed ταιριάσματα.

Οι Selbach et al. (2008) σύγκριναν 7 διαφορετικά προγράμματα πρόβλεψης στόχων miRNA, συγκεκριμένα τα TargetScanS, PicTar, RNA22, PITA, miRBase, miRanda, και DIANA-microT 3.0, καθώς και δύο απλές τεχνικές, το background που κάνει μια τελείως τυχαία επιλογή και το with seed που έχει σαν μοναδικό κριτήριο τα seed sites. Τα αποτελέσματα αυτών των αλγορίθμων τα συνέκριναν με τα αποτελέσματα από την ανάλυση pSILAC που πραγματοποιήθηκε στο συγκεκριμένο paper. Μόνο τρεις αλγόριθμοι πετυχαίνουν καλύτερη απόδοση από την απλή seed τεχνική, οι TargetScanS , PicTar και DIANA-microT, οι οποίοι χρησιμοποιούν το εξελικτικό conservation των seed sites σαν ένα επιπλέον φίλτρο.

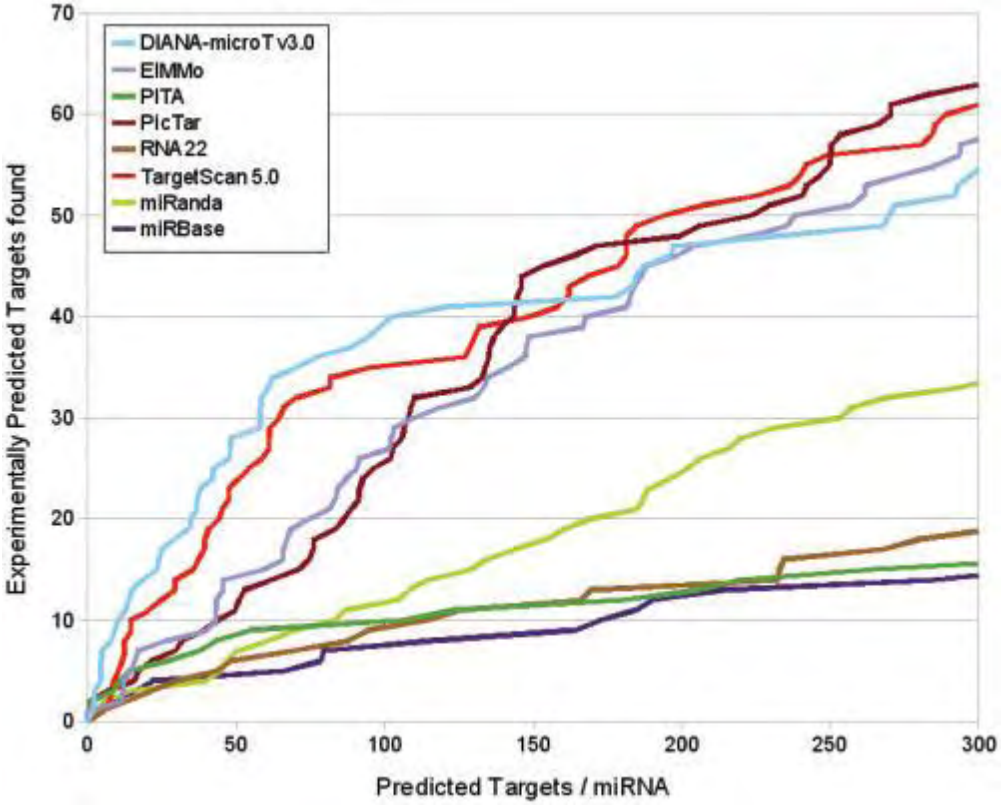
Οι Alexiou et al. (2009) έκαναν μια σύγκριση χρησιμοποιώντας τα γονίδια που προτάθηκαν ότι αποτελούν στόχο miRNA από τους Selbach et al. (2008). Η σύγκριση περιλάμβανε τους ακόλουθους αλγορίθμους: DIANA-microT 3.0, ElMMo, PITA,

PicTar, miRBase, miRanda, RNA22, TargetScan 5.0 και μια απλή τεχνική seed σύμφωνα με την οποία τα γονίδια αναγνωρίζονται σαν στόχοι και ταξινομούνται ανάλογα με το επίπεδο της συμπληρωματικότητάς τους με την seed περιοχή. Μέτρησαν το precision και το sensitivity και παρατήρησαν ότι πέντε από τα προγράμματα, συγκεκριμένα τα DIANA-microT 3.0, TargetScan 5.0, TargetScanS, PicTar και ElMMo πετυχαίνουν precision κοντά στο 50%, με το sensitivity να κυμαίνεται μεταξύ 6% και 12%. Τα συγκεκριμένα προγράμματα βασίζονται στο εξελικτικό conservation στην seed περιοχή (μερικά με κάποιες μικρές επεκτάσεις) και συνδυάζουν αυτή την πληροφορία με άλλα στοιχεία που χαρακτηρίζουν τα miTGs (miRNA targeted genes). Τα αποτελέσματα φαίνονται στην πρώτη γραφική παράσταση της εικόνας 4.2. Στην δεύτερη γραφική παράσταση της εικόνας 4.2 παραθέτουν ένα διάγραμμα pROC που δείχνει το precision έναντι του sensitivity.



Εικόνα 4.2: Σύγκριση εννέα αλγορίθμων πρόβλεψης γονιδίων στόχων miRNA με δεδομένα από το Selbach et al. (2008) (Πηγή: Alexiou et al., 2009)

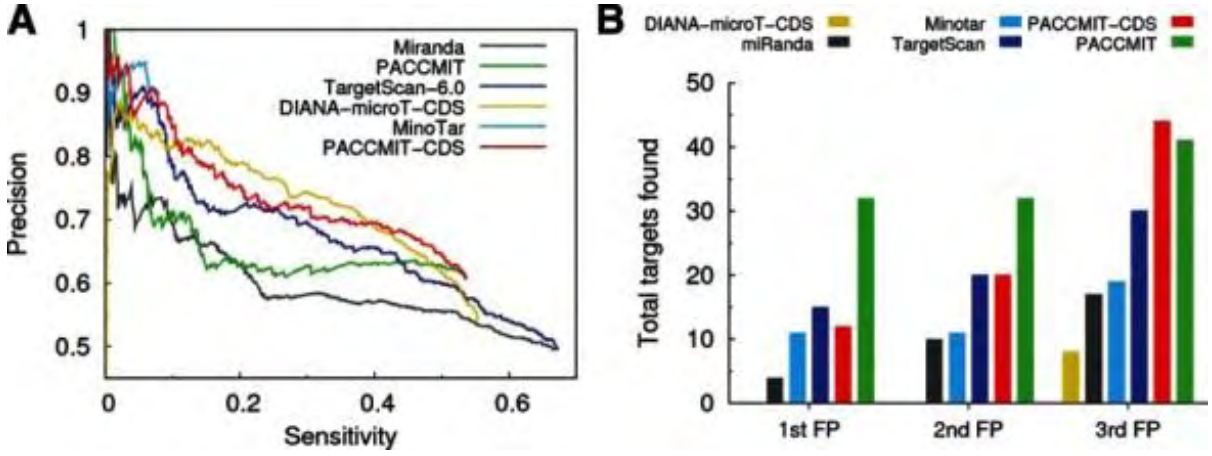
Επίσης διεξήγαγαν μια ακόμα έρευνα χρησιμοποιώντας ένα σετ δεδομένων από πειραματικούς στόχους οι οποίοι εξήχθησαν από την TarBase. Το συγκεκριμένο σετ δεδομένων περιλαμβάνει 150 στόχους από 61 διαφορετικά miRNA. Τα αποτελέσματα όπως φαίνονται στην εικόνα 4.3 επιβεβαιώνουν την προηγούμενη έρευνα ως προς την απόδοση των αλγορίθμων.



Εικόνα 4.3: Σύγκριση των προγραμμάτων πρόβλεψης γονιδίων στόχων miRNA βασισμένη στο σετ δεδομένων από την TarBase (Πηγή: Alexiou et al., 2009)

Μία ακόμη έρευνα των Marin et al. (2013) πραγματοποιεί μια σύγκριση αρκετών αλγορίθμων, αλλά αυτή τη φορά στην CDS περιοχή. Στην έρευνα αυτή συμπεριλαμβάνονται η DIANA-microT-CDS (Reczko et al.,2012) , ο miRanda (Enright et al.,2003), ο TargetScan-6.0 (Garcia et al.,2011), ο PACMIT (Marin και Vanicek,2012), ο PACCMIT-CDS και ο MinoTar ο οποίος είναι επέκταση του αλγορίθμου που παρουσιάστηκε από το Forman et al. (2008) και στον οποίο οι στόχοι αποφασίζονται από την παρουσία conserved 8μερών στην CDS περιοχή. Από τους αλγορίθμους πρόβλεψης γονιδίων στόχων miRNA, οι MinoTar, PACCMIT-CDS και DIANA-microT-CDS είναι από τους λίγους αλγορίθμους που έχουν ειδικευτεί ώστε να προβλέπουν στόχους στην CDS περιοχή και για αυτό συμπεριλήφθηκαν στην έρευνα. Επίσης, συμπεριλήφθηκαν οι υπόλοιποι αλγόριθμοι για να αξιολογηθεί αν οι αλγοριθμικές βελτιώσεις που απευθύνονται στην CDS περιοχή μειώνουν το ποσοστό των λανθασμένα θετικών. Για να αποφύγουν κάποια διαστρέβλωση λόγω της 3'UTR περιοχής, αξιολογήθηκε μόνο η πρόβλεψη στόχων στην CDS.

Στην εικόνα 4.4 φαίνονται τα αποτελέσματα της έρευνας. Ο PACCMIT-CDS και ο MinoTar έχουν καλύτερο precision για χαμηλότερο sensitivity, όμως η DIANA-microT-CDS έχει το καλύτερο precision, όταν αυξάνεται το sensitivity. Όσον αφορά το sensitivity η DIANA-microT-CDS και ο PACCMIT-CDS ξεπερνούν σε απόδοση τον MinoTar. Τέλος, φαίνεται ότι οι αλγόριθμοι που ειδικεύονται στην CDS περιοχή έχουν καλύτερη απόδοση από αυτούς που ειδικεύονται στην 3΄UTR περιοχή, επιβεβαιώνοντας ότι οι επιπλέον κανόνες που αφορούν την CDS περιοχή έχουν θετικό αντίκτυπο στο precision.



Εικόνα 4.4: Σύγκριση αλγορίθμων πρόβλεψης στόχων miRNA στην CDS περιοχή με κριτήριο το precision έναντι sensitivity (A) και ο αριθμός των σωστών θετικών πριν το πρώτο, δεύτερο και τρίτο αρνητικό θετικό (B) (Πηγή: Marin et al. , 2013)

Τα κριτήρια ώστε να συγκριθούν οι αλγόριθμοι διαφέρουν μεταξύ τους, με τα πιο συνηθισμένα να είναι το sensitivity, το specificity, το precision, το accuracy, το διάγραμμα ROC και η AUC του διαγράμματος ROC. Άλλα κριτήρια που συναντώνται στην παρούσα διπλωματική είναι τα Matthew's correlation coefficient (MCC), το Spearman rank correlation και το average classwise accuracy (ACA). Επίσης, ένας ακόμα τρόπος για να συγκριθούν οι αλγόριθμοι, είναι το κατά πόσο τα θετικά αποτελέσματα βρίσκονται στην αρχή της λίστας των αποτελεσμάτων του αλγορίθμου ή βρίσκονται ομοιόμορφα κατανεμημένα. Η πρώτη επιλογή είναι η προτιμότερη, αφού δείχνει ότι ο αλγόριθμος έχει και μεγάλη ικανότητα ταξινόμησης. Ένας παρόμοιος τρόπος που χρησιμοποιούν κάποιες έρευνες είναι η σύγκριση των σωστά θετικών αποτελεσμάτων στις πρώτες χ προβλέψεις, όπου ο αριθμός προβλέψεων μπορεί να είναι 100 ή 500 ή να παρατίθεται μια γραφική παράσταση που δείχνει τον αριθμό των σωστά θετικών σε διαφορετικές τιμές top προβλέψεων.

Επιπλέον, σε μια έρευνα των Liu et al. (2010), παρατίθεται η Cumulative Fold Change (CFC), στην οποία ένας αλγόριθμος με υψηλότερο precision και λιγότερα λανθασμένα θετικά αναμένεται να έχει γρηγορότερη μείωση. Τέλος, οι Marik και Vanicek (2010) παραθέτουν μια έρευνα που μετράει την κατανάλωση CPU, ώστε να δειχθεί ποιος αλγόριθμος είναι υπολογιστικά αποτελεσματικότερος.

Στην εικόνα 2.3, βλέπε σελίδα 15 οι Sturm et al. (2010) χρησιμοποίησαν την έρευνα των Selbach et al. (2008). Εκεί πρόσθεσαν κάποιους νέους αλγορίθμους όπως τις διάφορες εκδόσεις του TargetSpy, τον mirTarget2, τον EIMMo και τον TargetRank οι οποίοι πετυχαίνουν οριακά καλύτερα αποτελέσματα από τους τρεις αλγορίθμους που ξεχώρισαν στην έρευνα των Selbach et al., δηλαδή τους DIANA-microT, TargetScanS και PicTar. Βασισμένη επίσης στην έρευνα των Selbach et al., οι Alexiou et al. είχαν παραθέσει μια έρευνα που έδειχνε το precision και το sensitivity κάθε αλγορίθμου (εικόνα 4.2, σελίδα 56), σύμφωνα με την οποία τα DIANA-microT 3.0, TargetScan 5.0, TargetScanS, PicTar και ElMMo πέτυχαν precision κοντά στο 50% με το sensitivity να κυμαίνεται μεταξύ 6% και 12%.

Πολλές είναι οι έρευνες που εξετάζουν το sensitivity ή recall κάθε αλγορίθμου. Εκεί τα αποτελέσματα είναι διαφορετικά για τους ίδιους αλγόριθμους γεγονός που οφείλεται στην ελλιπή αντικειμενικότητα των ερευνών, ειδικά όταν δεν είναι ανεξάρτητες. Ένα ακόμα πρόβλημα βέβαια είναι η χρήση διαφορετικών εκδόσεων ενός αλγορίθμου, που μπορεί να επηρεάσει σε μεγάλο βαθμό την απόδοση. Παραδείγματος χάρη, ο TargetScan που εμπεριέχεται στις περισσότερες έρευνες, πετυχαίνει sensitivity γύρω στο 12% στους Alexiou et al., 66% σε μία έρευνα των Bandyopadhyay και Mitra και γύρω στο 26% σε μια έρευνα των Dweep et al. (2011). Τα αποτελέσματα δεν μπορούν να εξεταστούν ξεχωριστά από το σετ δεδομένων που χρησιμοποιείται.

Μετρικές όπως το sensitivity, το specificity κ.ο.κ. είναι προτιμότερο να εξετάζονται σε συνδυασμό, αφού το μεγάλο sensitivity πολλές φορές οδηγεί σε αρκετά χαμηλό specificity και το αντίθετο. Έτσι, χρησιμοποιώντας χαλαρά φίλτρα μπορεί ένας αλγόριθμος να πετυχαίνει μεγάλο precision ή sensitivity, παρουσιάζοντας φαινομενικά καλύτερη απόδοση από άλλους αλγορίθμους, όμως θυσιάζοντας την απόδοση του σε άλλους τομείς. Βέβαια, εξαρτάται από τον κάθε ερευνητή τι ακριβώς

θέλει να χρησιμοποιήσει και τι τον ενδιαφέρει περισσότερα στα αποτελέσματα που θα αποκομίσει (π.χ. προτιμάει να μην υπάρχουν λανθασμένα θετικά ή να μην υπάρχουν λανθασμένα αρνητικά). Σε αυτό το επίπεδο υπάρχει ένα tradeoff στους περισσότερους αλγορίθμους. Επίσης, έχει σημασία αν δουλεύει με ζώα ή φυτά.

Ένας τύπος έρευνας είναι το διάγραμμα ROC. Στα πλαίσια της παρούσας διπλωματικής παρουσιάστηκαν διάφορα διαγράμματα από διαφορετικές έρευνες. Τα τρία ήταν από δύο διαφορετικά papers των Bandyopadhyay και Mitra, καθώς και των Sturm et al. (2010), στα οποία την καλύτερη απόδοση την πετυχαίνουν οι νέοι αλγόριθμοι TargetMiner, MultiMiTar και TargetSpy. Ενδεικτικό είναι ότι στις έρευνες των Bandyopadhyay και Mitra αλγόριθμοι όπως ο DIANA-microT και ο PicTar πετυχαίνουν πολύ χαμηλό sensitivity και θεωρείται ότι έχουν χαμηλή υπολογιστική ικανότητα, ενώ ο PicTar στην έρευνα των Sturm et al. πετυχαίνει πολύ καλή απόδοση. Τέλος, ενώ η AUC του ROC διαγράμματος χρησιμοποιείται αρκετά συχνά, υπάρχουν τελευταία κάποιες ενστάσεις σχετικά την αξιοπιστία της. Μια έρευνα των Hanczar et al. (2010) έδειξε ότι η AUC είναι αρκετά θορυβώδης για μετρική ταξινόμησης, ενώ οι έρευνες των Lobo et al. (2008) και Hand και David (2009) ότι υπάρχουν και άλλα αρκετά σημαντικά προβλήματα στην σύγκριση μέσω της AUC. Παρ' όλα αυτά, η AUC σαν μέτρο συγκεντρωτικής ταξινόμησης έχει αθωωθεί στα πλαίσια της ομοιόμορφης κατανομής (Flach et al., 2011).

Αρκετές έρευνες παραθέτουν το precision σε συνδυασμό με το sensitivity των αλγορίθμων που συγκρίνουν. Σε αυτές τις έρευνες τα αποτελέσματα είναι σχετικά παρόμοια. Στην έρευνα των Dweep et al. (εικόνα 2.9, σελίδα 24) στην οποία ο miRWalk που παρουσιάζεται στο συγκεκριμένο paper πετυχαίνει τέλεια απόδοση, πολύ καλή απόδοση πετυχαίνουν και οι DIANA-microT PITA και miRanda, ενώ ο TargetScan πετυχαίνει πολύ χαμηλό sensitivity. Σε μια ακόμα έρευνα που εξετάζεται το precision κοινά με το sensitivity, των Bandyopadhyay και Mitra (εικόνα 2.13, σελίδα 28), οι TargetScan και DIANA-microT πετυχαίνουν πολύ καλό precision, αλλά χαμηλό sensitivity, οι MultiMitar και TargetMiner λίγο χειρότερο precision, όμως καλύτερο sensitivity, ενώ ο miRanda χαμηλό precision αλλά υψηλό sensitivity. Τέλος, στην έρευνα των Alexiou et al. ο DIANA-microT πετυχαίνει πολύ καλό precision με χαμηλό sensitivity (DIANA v3.0 strict) ή λίγο χαμηλότερο precision με

καλύτερο sensitivity (DIANA v3.0 loose), ο TargetScan καλό precision και καλό sensitivity, ενώ ο miRanda χαμηλό precision, αλλά πολύ καλό sensitivity.

Κατά την διάρκεια της παρούσας διπλωματικής παρατέθηκαν διάφορες έρευνες σχετικά με νέους αλγόριθμους πρόβλεψης microRNA, αρκετοί εκ των οποίων περιέχουν μεθόδους που εμπίπτουν στο πεδίο της μηχανικής μάθησης. Στην εικόνα 4.5 παρατίθεται μία συνοπτική παρουσίαση των σημαντικότερων σχετικών ερευνών που παρουσιάστηκαν, καθώς και το training set και ο αλγόριθμος ο οποίος χρησιμοποιήθηκε, ο αριθμός των χαρακτηριστικών και ο αριθμός των interactions miRNA-γονιδίου που χρησιμοποιήθηκαν κατά την διεξαγωγή της κάθε έρευνας.

Algorithms	Training set	Training algorithm	Test set	Number of interactions	Number of Features
Targetspy	3872 positive and 4540 negative examples taken from 20 argonaute-mRNA binding sites for the 20 most abundant microRNAs present in the P13 mouse brain	Automatic feature selection	Stark et al., Kertesz et al.	120 190	7
Targetspy	3872 positive and 4540 negative examples taken from 20 argonaute-mRNA binding sites for the 20 most abundant microRNAs present in the P13 mouse brain	Automatic feature selection	Selbach et al.	23.806	7
TargetMiner	289 validated positive targets from TarBase 289 tissue-specific negative examples	SVM classifier (RBF kernel)	Independent	246	30
MultiMiTar	289 validated positive targets from TarBase 289 tissue-specific negative examples	SVM based classifier integrated with a multiobjective metaheuristic based feature selection technique	Independent	246	39
MultiMiTar	289 validated positive targets from TarBase 289 tissue-specific negative examples	SVM based classifier integrated with a multiobjective metaheuristic based feature selection technique	Selbach et al.	15.806	39
SVMicrO	896 positive examples from miRecords 3542 negative examples from 20 miRNA over-expression data taken from NCBI Gene Expression Omnibus	2-stage structure including a site-SVM followed by a UTR-SVM	Training Data	4438	39

Εικόνα 4.5: Παρουσίαση των σημαντικότερων ερευνών σχετικά με τους νέους αλγορίθμους μηχανικής μάθησης

Επίσης, στη εικόνα 4.6 παρατίθεται ένα ακόμα διάγραμμα που περιλαμβάνει τους νέους αλγόριθμους που παρουσιάστηκαν, τα websites από τα οποία μπορεί να αποκτήσει πρόσβαση στους αλγόριθμους αυτούς, τα άρθρα που τους περιγράφουν, καθώς και το αν παρέχουν κάποιο εκτελέσιμο αρχείο του αλγορίθμου.

Algorithm	Website	Reference	Executable
TargetSpy	http://www.targetspy.org	Sturm et al., 2010	Yes
MultiMiTar	http://www.isical.ac.in/~bioinfo_miu/multimitar.htm	Mitra and Bandyopadhyay, 2011	Yes
TargetMiner	http://www.isical.ac.in/~bioinfo_miu/targetminer20.htm	Bandyopadhyay and Mitra, 2009	Yes
miRWalk	http://www.umm.uni-heidelberg.de/apps/zmf/mirwalk/	Dweep et al., 2010	No
PACMIT	http://lcpt.epfl.ch/page-67906-en.html	Marin and Vanicek, 2011	No
SVMicrO	http://www.mybiosoftware.com/rna-analysis/12293	Liu et al., 2010	Yes

Εικόνα 4.6: Νέοι αλγόριθμοι πρόβλεψης στόχων miRNA

Οι αλγόριθμοι πρόβλεψης στόχων miRNA, που φαίνονται να ξεχωρίζουν, είναι οι DIANA-micro T, TargetScan και miRanda. Από τους συγκεκριμένους αλγορίθμους οι TargetScan και DIANA-microT πετυχαίνουν καλύτερο specificity και precision στις περισσότερες έρευνες, όμως ο miRanda έχει καλύτερο sensitivity. Όσον αφορά τους καινούργιους αλγορίθμους που παρουσιάστηκαν στην παρούσα διπλωματική, πρέπει να πραγματοποιηθούν περισσότερες ανεξάρτητες έρευνες για να εξεταστούν ενδελεχώς, αλλά τα αποτελέσματα που παρατίθενται μέχρι στιγμής είναι αρκετά ικανοποιητικά και οι μελλοντικές επεκτάσεις τους τα καθιστούν αρκετά σημαντικά για το μέλλον της πρόβλεψης γονιδίων στόχων miRNA.

4.2 Σύγκριση των ερευνών με την έρευνα των Alexiou et al. (2009)

Όπως προαναφέρθηκε, οι Alexiou et al. (2009) έκαναν μια σύγκριση (εικόνα 4.2, σελίδα 56) χρησιμοποιώντας τα proteomics δεδομένα από τους Selbach et al. (2008), διαθέσιμα στο http://psilac.mdc-berlin.de. Οι Selbach et al. εφάρμοσαν γενετική επιμόλυνση (transfection), ώστε να πραγματοποιηθεί το overexpression σε πέντε ανθρώπινα miRNA στα HeLa κύτταρα. Στην συνέχεια μέτρησαν την αλλαγή στην παραγωγή των πρωτεϊνών. Ορισμένα από τα συγκεκριμένα miRNA είναι ιστοειδικά και άφθονα στα HeLa κύτταρα (miR-1, miR-155), ενώ άλλα είναι εκφρασμένα σε πολλούς ιστούς (miR-16, miR-30a, let-7b), συμπεριλαμβανομένων των HeLa κυττάρων. Οι Alexiou et al. αναγνώρισαν συνολικά 15806 πιθανές αλληλεπιδράσεις miRNA-γονιδίου και από τα 5 σετ δεδομένων. Οι πρωτεΐνες που παρουσιάζουν αλλαγή fold μικρότερη του -0.2 θεωρήθηκαν ότι έχουν επηρεαστεί από τα miRNA. Στους Selbach et al., ο ορισμός δεν ήταν τόσο αυστηρός, αφού η αλλαγή fold χρειαζόταν να είναι μικρότερη του -0.1, ώστε οι πρωτεΐνες να θεωρηθούν απορρυθμισμένες.

Σε μια σύγκριση που διεξήγαγαν οι Sturm et al. (εικόνα 2.2, σελίδα 14), τα δεδομένα που χρησιμοποίησαν προέρχονται από δύο διαφορετικά σετ δεδομένων, το σετ δεδομένων των Stark et al. (2005) και το σετ δεδομένων των Kertesz et al. (2007). Το σετ δεδομένων των Stark et al. περιέχει 120 αλληλεπιδράσεις miRNA-γονιδίου του οργανισμού Drosophila Melanogaster. Το σετ δεδομένων των Kertesz et al. περιέχει 190 αλληλεπιδράσεις miRNA-γονιδίου του οργανισμού Drosophila Melanogaster και αποτελεί επέκταση του προηγούμενου σετ δεδομένων των Stark et al. Την καλύτερη απόδοση στο σετ δεδομένων των Stark et al. πέτυχε ο EIMMo, δείχνοντας ένα πολύ καλό δείκτη σωστά θετικών μαζί με ένα πολύ χαμηλό δείκτη λανθασμένα θετικών. Στην συνέχεια ακολουθούν οι PicTar, TargetSpy seed, PITA ALL 3/15, TargetScanS κ.ά., ενώ ο DIANA-microT δεν περιλαμβάνεται στην σύγκριση. Όσον αφορά το σετ δεδομένων των Kertesz et al. ο PITA ALL 3/15 πετυχαίνει την καλύτερη απόδοση, με την κατάταξη των υπόλοιπων αλγορίθμων να μην αλλάζει σημαντικά. Στην έρευνα των Alexiou et al., ο PITA είχε πολύ άσχημη απόδοση, όμως και οι δύο έρευνες συμφωνούν στην πολύ κακή απόδοση του RNA22.

Οι Sturm et al. (2009) πραγματοποίησαν επίσης μια σύγκριση (εικόνα 2.3, σελίδα 15) βασισμένη στα δεδομένα των Selbach et al. και στην προσέγγιση των Selbach et al.,

σύμφωνα με την οποία οι πρωτεΐνες θεωρούνται απορρυθμισμένες αν έχει υπάρξει αλλαγή fold μικρότερη του -0.1. Όπως αναφέρθηκε πριν, στην έρευνα των Alexiou et al. το κριτήριο ήταν αυστηρότερο με τις πρωτεΐνες να θεωρούνται απορρυθμισμένες αν έχει υπάρξει αλλαγή fold μικρότερη του -0.2. Οι Sturm et al. παραθέτουν το accuracy κάθε αλγορίθμου, με τις διάφορες εκδόσεις του TargetSpy να πετυχαίνουν τα καλύτερα αποτελέσματα και πολλούς γνωστούς αλγορίθμους να κυμαίνονται μεταξύ 55-65%. Εξαίρεση αποτελεί ο miRanda και ο RNA22 που πετυχαίνουν 45.5% και 36.6% αντίστοιχα. Όσον αφορά το sensitivity, τα αποτελέσματα συμπίπτουν σε μεγάλο βαθμό με τα αποτελέσματα των Alexiou et al.

Οι Bandyopadhyay και Mitra (2009) διεξήγαγαν μια σύγκριση χρησιμοποιώντας ένα σετ δεδομένων από mRNAs που είτε έχει κατασταλεί η μετάφρασή τους (translationally repressed) είτε έχουν αποικοδομηθεί (mRNA cleavage) το οποίο εξήχθη από την TarBase (Papadopoulos et al., 2009). Μέτρησαν το sensitivity και το specificity έξι αλγορίθμων, συγκεκριμένα των TargetMiner, TargetScan, PicTar, miRanda, MirTarget2 και NBmiRTar. Όσον αφορά το sensitivity, ο TargetMiner υπερέχει των υπολοίπων με δεύτερο τον TargetScan και τον miRanda τέταρτο, γεγονός που δεν συμφωνεί με την έρευνα των Alexiou et al., σύμφωνα με την οποία ο miRanda πετυχαίνει το καλύτερο sensitivity.

Οι Mitra και Bandyopadhyay (2011) διεξήγαγαν μια σύγκριση βασισμένη στα δεδομένα των Selbach et al., αλλά χρησιμοποιώντας την προσέγγιση των Alexiou et al. σύμφωνα με την οποία οι πρωτεΐνες θεωρούνται απορρυθμισμένες αν έχει υπάρξει αλλαγή fold μικρότερη του -0.2. Χρησιμοποιούν τις 15.806 αλληλεπιδράσεις miRNA-γονιδίου που παρατήρησαν οι Alexiou et al. με συνολικά 2.406 αλληλεπιδράσεις να θεωρούνται στόχοι. Τα αποτελέσματα για τους κοινούς αλγορίθμους συμπίπτουν με την έρευνα των Alexiou et al. τόσο στο sensitivity όσο και στο precision με τις όποιες αποκλίσεις να μην ξεπερνούν το 1%. Επιπρόσθετα, στην έρευνα των Mitra και Bandyopadhyay παρουσιάζεται και η απόδοση των νέων αλγορίθμων MultiMiTar, TargetMiner και TargetSpy με τους δύο πρώτους να πετυχαίνουν πολύ καλή απόδοση, συνδυάζοντας υψηλό sensitivity (μόνο το miRanda παρουσιάζει τόσο υψηλό) μαζί με precision σχεδόν ίσο με των γνωστών αλγορίθμων πρόβλεψης γονιδίων στόχων miRNA κοντά στο 50% (με εξαίρεση τον DIANA v3.0 strict που πετυχαίνει μεγαλύτερο precision). Ο TargetSpy επίσης πετυχαίνει καλή απόδοση στην συγκεκριμένη έρευνα.

Μια σύγκριση των Mitra και Bandyopadhyay (2011) βασίστηκε στο ανθρώπινο γονίδιο p21Cip1/Waf1, το οποίο είναι επίσης γνωστό ως Cyclin-dependent kinase inhibitor 1A (CDKN1A) και το οποίο μπορεί να γίνει στόχος από 28 miRNA (Wu et al., 2010). Μετρήθηκε το sensitivity κάθε αλγορίθμου, βλέποντας πόσα από τα 28 miRNA προβλέφθηκαν. Ο PITA βρήκε τα 27 από τα 28 miRNA, ενώ ο miRanda πέτυχε πολύ υψηλό sensitivity (60.71%). Ο DIANA-microT v3.0 πέτυχε sensitivity περίπου 39.28%, ενώ ο TargetScan περίπου 28% και ο PicTar περίπου 21%. Επίσης, ο MultiMiTar πέτυχε πολύ υψηλό sensitivity (67.86%). Η κατάταξη των αλγορίθμων όσον αφορά το sensitivity συμπίπτει σε μεγάλο βαθμό με την έρευνα των Alexiou et al.

Οι Marin και Vanicek (2011) διεξήγαγαν μια σύγκριση στην οποία συνέκριναν τους πιο γνωστούς αλγορίθμους πρόβλεψης γονιδίων στόχων miRNA στους οργανισμούς Drosophila Melanogaster και στον άνθρωπο. Όσον αφορά τον οργανισμό Drosophila Melanogaster χρησιμοποίησαν ένα σετ δεδομένων αποτελούμενο από 220 πειραματικά ελεγμένες αλληλεπιδράσεις miRNA-3΄UTR, αναφερόμενες ως λειτουργικές ή μη λειτουργικές. Αυτό το σετ δεδομένων είναι βασισμένο στο σετ δεδομένων των Kertesz et al. (2007) και συμπληρωμένο με επικυρωμένους στόχους από την βάση δεδομένων miRecords. Όσον αφορά τον άνθρωπο, χρησιμοποίησαν τα δεδομένα των Selbach et al. Από τα 15806 ζευγάρια miRNA-3΄UTR θεώρησαν λειτουργικά τα 2406, ενώ τα 13400 μη λειτουργικά (Alexiou et al.,2009). Στην σύγκριση που διεξήγαγαν περιλαμβάνονται οι αλγόριθμοι: PITA, IntaRNA, miRanda και RNAhybrid.

Οι Dweep et al. (2010) πραγματοποίησαν μια σύγκριση χρησιμοποιώντας ένα σύνολο από γονίδια (θετικά και αρνητικά σετ δεδομένων), τα οποία εξήχθησαν από την TarBase, την miRecords και την miRTarBase για τα οποία τα miRNA binding sites είναι επικυρωμένα και δημοσιευμένα στην βάση δεδομένων του PubMed. Συνολικά, χρησιμοποίησαν 1870 θετικές αλληλεπιδράσεις miRNA – στόχου και 61 αρνητικές. Όσον αφορά τα αποτελέσματα, το precision είναι σχεδόν το ίδιο σε όλους τους αλγορίθμους και κοντά στο 100%, γεγονός που δεν συμφωνεί με την έρευνα των Alexiou et al, όπου ο DIANA v3.0 (strict) υπερέχει σημαντικά των υπολοίπων και οι υπόλοιποι δεν παρουσιάζουν κοινή απόδοση. Όσον αφορά το sensitivity, την καλύτερη απόδοση πετυχαίνει ο mirWalk με τον miRanda δεύτερο με απόδοση κοντά στο 80%, τους DIANA-microT και PITA να πετυχαίνουν λίγο πάνω από 70%, ενώ ο

TargetScan έχει πολύ χαμηλό sensitivity, γύρω στο 30%. Στην έρευνα των Alexiou et al. ο miRanda είχε πετύχει το καλύτερο sensitivity, αλλά ο TargetScan είχε πετύχει το δεύτερο καλύτερο. Το PITA επίσης δεν παρουσίαζε τόσο καλά αποτελέσματα όσον αφορά το sensitivity στην έρευνα των Alexiou et al.

5. Επίλογος

Η σημασία των miRNA τόσο στον άνθρωπο όσο και σε άλλους οργανισμούς τα καθιστά κέντρο της προσοχής στην επιστήμη της βιοπληροφορικής. Οι υπολογιστικοί μέθοδοι αποτελούν την πρακτικότερη και αποτελεσματικότερη επιλογή για να αναγνωριστούν άγνωστες αλληλεπιδράσεις miRNA-γονιδίων και να επιλεχθούν διακεκριμένοι υποψήφιοι για wet lab πειράματα. Όμως, ένας πολύ μεγάλος αριθμός αλληλεπιδράσεων γονιδίων-miRNA δεν αναγνωρίζεται ακόμα και από τους καλύτερους αλγορίθμους πρόβλεψης γονιδίων στόχων miRNA. Επίσης, στα αποτελέσματα των συγκεκριμένων αλγορίθμων υπάρχουν πολλά λανθασμένα θετικά, γεγονός που καθιστά απαραίτητη και αναγκαία την πειραματική επικύρωσή τους, προτού χρησιμοποιηθούν. Τα ολοένα και περισσότερο αυξανόμενα δεδομένα ίσως βοηθήσουν στην αναγνώριση νέων κανόνων που διέπουν την λειτουργία των miRNA, καθώς και να στο χρησιμοποιηθούν για να εκπαιδευτούν οι εφαρμογές που χρησιμοποιούν μηχανική μάθηση.

Τα προγράμματα που εξετάστηκαν στην παρούσα διπλωματική αποτελούν ένα σημαντικό κομμάτι της πρόβλεψης γονιδίων στόχων miRNA. Στην παρούσα διπλωματική πραγματοποιήθηκε η παρουσίαση καινούργιων προγραμμάτων πρόβλεψης γονιδίων στόχων miRNA που πληρούν κάποια κριτήρια, καθώς και η σύγκρισή τους με βάση διάφορες έρευνες που παρατίθενται. Κατά την προσπάθεια να πραγματοποιηθεί η σύγκριση παρουσιάστηκαν αρκετές δυσκολίες. Οι έρευνες πολλές φορές χρησιμοποιούν διαφορετικά σετ δεδομένων ή και διαφορετικούς ορισμούς σχετικά με τα κριτήρια ώστε να χαρακτηριστεί ένα γονίδιο σαν στόχος. Επίσης, χρησιμοποιούνται διαφορετικά κριτήρια αξιολόγησης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Akçakaya P, Ekelund S, Kolosenko I, Caramuta S, Ozata DM, Xie H, Lindforss U, Olivecrona H, Lui WO, 2011. *miR-185 and miR-133b deregulation is associated with overall survival and metastasis in colorectal cancer.* In *Int J Oncol.* 39(2):311-8.

Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG, 2009. *Lost in translation: an assessment and perspective for computational microRNA target identification*. In *Bioinformatics*, 25(23):3049-55.

Ardekani AM, Naeini MM, 2010. *The Role of MicroRNAs in Human Diseases*. In *Avicenna J Med Biotechnol*., 2(4): 161–179.

Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP, 2008. *The impact of microRNAs on protein output.* In *Nature,* 455(7209):64–71.

Bandyopadhyay S, Mitra R, 2009. *TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples.* In *Bioinformatics,* 25(20):2625-31.

Betel D, Koppal A, Agius P, Sander C, Leslie C, 2010. *Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites.* In *Genome Biol.,* 11(8):R90.

Betel D, Wilson M, Gabow A, Marks DS, Sander C, 2007. *The microRNA.org resource: targets and expression.* In *Nucleic Acids Res,* 36(Database issue):D149-53.

Chi SW, Zang JB, Mele A, Darnell RB, 2009. *Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps.* In *Nature,* 460:479–486.

Dweep H, Gretz N, Sticht C, 2011. *miRWalk - database: prediction of possible miRNA binding sites by "walking" the genes of 3 genomes.* In *Journal of Biomedical Informatics,* 44: 839-7.

Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS, 2003. *MicroRNA targets in Drosophila.* In *Genome Biology,* 5:R1.

Flach PA, Hernandez-Orallo J, Ferri C, 2011. *A coherent interpretation of AUC as a measure of aggregated classification performance*. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11),* pp. 657–664.

Forman JJ, Legesse-Miller A, Coller HA, 2008. *A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence.* In *Proc Natl Acad Sci,* 105: 14879–14884.

Friedman RC, Farh KK, Burge CB, Bartel DP, 2009. *Most mammalian mRNAs are conserved targets of microRNAs.* In *Genome Res,* 19(1):92-105.

Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP, 2011. *Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs.* In *Nat Struct Mol Biol.,* 18(10):1139-46.

Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP., 2007. *MicroRNA targeting specificity in mammals: determinants beyond seed pairing.* In *Mol Cell.,* 27(1): 91-105.

Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T, 2010. *Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP*. In *Cell,* 141:129–141.

Hanczar B, Hua J, Sima C, Weinstein J, Bittner M, Dougherty ER, 2010. *Small-sample precision of ROC-related estimates.* In *Bioinformatics,* 26 (6): 822–830.

Hand DJ, 2009. *Measuring classifier performance: A coherent alternative to the area under the ROC curve.* In *Machine Learning,* 77: 103–12.

Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, Tsai WT, Chen GZ, Lee CJ, Chiu CM, Chien CH, Wu MC, Huang CY, Tsou AP, Huang HD, 2011. *MiRTarBase: a database curates experimentally validated microRNA-target interactions*. In *Nucleic Acids Res.,* 39: D163–D169.

Huang FWD, Qin J, Reidys CM, Stadler PF, 2010. *Target prediction and a statistical sampling algorithm for RNA-RNA interaction.* In *Bioinformatics,* 26: 175–181.

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS, 2004. *Human MicroRNA Targets.* In *PLoS Biol.,* 3(7):e264.

Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E, 2007. *The role of site accessibility in microRNA target recognition.* In *Nature genetics*, 39(10):1278–1284.

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T, 2001. *Identification of novel genes coding for small expressed RNAs.* In *Science*, 294(5543):853-8.

Lau NC, Lim LP, Weinstein EG, Bartel DP, 2001. *An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans*. In *Science*, 294(5543):858-62.

Lee RC, Ambros V, 2001. *An extensive class of small RNAs in Caenorhabditis elegans*. In *Science*, 294(5543):862-4.

Lee RC, Feinbaum RL, Ambros V, 1993. *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14*. In *Cell*, 75(5):843-54.

Lewis BP, Burge CB, Bartel DP, 2005. *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* In *Cell, 120(1):15-20.*

Liu H, Yue D, Chen Y, Gao SJ, Huang Y, 2010. *Improving performance of mammalian microRNA target prediction.* In *BMC Bioinformatics, 11:476.*

Lobo JM, Jiménez-Valverde A, Real R, 2008. *AUC: a misleading measure of the performance of predictive distribution models.* In *Global Ecology and Biogeography,* 17: 145–151.

Maragkakis M, Alexiou P, Papadopoulos GL, Reczko M, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K, Simossis VA, Sethupathy P, Vergoulis T, Koziris N, Sellis T, Tsanakas P, Hatzigeorgiou AG, 2009. *Accurate microRNA target prediction correlates with protein repression levels.* In *BMC Bioinformatics ,* 10:295.

Marín RM, Sulc M, Vanícek J, 2013. *Searching the coding region for microRNA targets.* In *RNA,* 19(4):467-74.

Marín RM, Vanícek J, 2011. *Efficient use of accessibility in microRNA target prediction.* In *Nucleic Acids Res.,* 39(1):19-29.

Mitra R, Bandyopadhyay S, 2011. *MultiMiTar: a novel multi objective optimization based miRNA-target prediction method.* In *PLoS One,* 6(9):e24583.

Naeini MM, Ardekani AM, 2009. *Noncoding RNAs and Cancer.* In *Avicenna J Med Biotech*., 1(2):55–70.

Ott CE, Grünhagen J, Jäger M, Horbelt D, Schwill S, Kallenbach K, Guo G, Manke T, Knaus P, Mundlos S, Robinson PN, 2011. *MicroRNAs differentially expressed in postnatal aortic development downregulate elastin via 3′ UTR and coding-sequence binding sites.* In *PLoS ONE,* 6:e16250.

Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, Hatzigeorgiou AG, 2009. *The database of experimentally supported targets: a functional update of tarbase.* In *Nucleic Acids Res,37:* D155-8.

Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, Filippidis C, Dalamagas T, Hatzigeorgiou AG, 2013. *DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows.* In *Nucleic Acids Research*, 41:W169-73.

Peterson SM, Thompson JA, Ufkin ML, Sathyanarayana P, Liaw L, Congdon CB, 2014. *Common features of microRNA target prediction tools.* In *Front Genet.,* 5: 23.

Qin W, Shi Y, Zhao B, Yao C, Jin L, Ma J, Jin Y, 2010. *miR-24 regulates apoptosis by targeting the open reading frame (ORF) region of FAF1 in cancer cells.* In *PLoS ONE,* 5: e9429.

Reczko M, Maragkakis M, Alexiou P, Grosse I, Hatzigeorgiou AG, 2012. *Functional microrna targets in protein coding sequences*. In *Bioinformatics*, 28:771-6.

Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N, 2008. *Widespread changes in protein synthesis induced by microRNAs*. In *Nature*, 455(7209):58–63.

Sethupathy P, Megraw M, Hatzigeorgiou AG, 2006. *A guide through present computational approaches for the identification of mammalian microRNA targets.* In *Nat Methods,* 3: 881-886.

Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM, 2005. *Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution*. In *Cell*, 123(6):1133–1146.

Sturm M, Hackenberg M, Langenberger D, Frishman D, 2010. *TargetSpy: a supervised machine learning approach for microRNA target prediction.* In *BMC Bioinformatics,* 11:292.

Vlachos IS, Hatzigeorgiou AG, 2013.*Online resources for miRNA analysis*. In *Clin Biochem.,* 46(10-11):879-900.

Witkos TM, Koscianska E, Krzyzosiak WJ, 2011. *Practical Aspects of microRNA Target Prediction.* In *Curr Mol Med.,* 11(2): 93–109.

Wu S, Huang S, Ding J, Zhao Y, Liang L, Liu T, Zhan R, He X, 2010. *Multiple microRNAs modulate p21Cip1/Waf1 expression by directly targeting its 30 untranslated region.* In *Oncogene,* 29:2302–2308.

Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T, 2009. *miRecords: an integrated resource for microRNA-target interactions.* In *Nucleid Acids Res.,* D105-10.