ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΓΙΑ ΤΗΝ ΥΠΟΛΟΓΙΣΤΙΚΗ ΠΡΟΒΛΕΨΗ MICRORNA ΣΤΟΧΩΝ

FEATURE SELECTION FOR COMPUTATIONAL MICRORNA TARGET PREDICTION

Μεταπτυχιακή διατριβή

από

Κοσμίδου Μαρία

Επιβλέποντες καθηγητές:

Χατζηγεωργίου Άρτεμις | Καθηγήτρια

Ποταμιάνος Γεράσιμος | Αναπληρωτής καθηγητής

Σταμούλης Γεώργιος | Καθηγητής

ΟΚΤΩΒΡΙΟΣ 2017,

ΒΟΛΟΣ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ Η/Υ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

UNIVERSITY OF THESSALY

Κοσμίδου Μαρία | MSc ΗΜΜΥ
Πανεπιστήμιο Θεσσαλίας

"*Bioinformatics is now the motor of the innovation.*
It not only answers the data inquiries, but also, more importantly,
determines what questions need to be asked in the first place, freeway."

— Wim Van Criekinge, bioinformatician

ΕΥΧΑΡΙΣΤΙΕΣ

Με το τέλος της μεταπτυχιακής μου διατριβής, θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια, κα. Άρτεμις Χατζηγεωργίου, καθηγήτρια του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ του πανεπιστημίου Θεσσαλίας στον Βόλο, για την ευκαιρία που μου έδωσε να γνωρίσω καλύτερα έναν τόσο ενδιαφέρον κλάδο, τον κλάδο της Βιοπληροφορικής καθώς και για την καθοδήγησή της. Επιπρόσθετα, θα ήθελα να ευχαριστήσω την διδακτορικό του πανεπιστημίου Θεσσαλίας Δήμητρα Καραγκούνη για την πολύτιμη βοήθειά της καθ' όλη τη διάρκεια της εργασίας. Τέλος, χωρίς την συγκατάθεση της τριμελούς επιτροπής Χατζηγεωργίου Άρτεμις, Ποταμιάνος Γεράσιμος, αναπληρωτής καθηγητής του Πανεπιστημίου Θεσσαλίας, Σταμούλης Γεώργιος, καθηγητής του Πανεπιστημίου Θεσσαλίας δε θα είχα την δυνατότητα να ξεκινήσω αυτό το ενδιαφέρον ταξίδι και τους ευχαριστώ γ' αυτό.

ΠΕΡΙΛΗΨΗ

Τα microRNAs (miRNAs) εμπλέκονται σε πολλές διαφορετικές βιολογικές διαδικασίες και μπορούν δυνητικά να ρυθμίσουν τις λειτουργίες χιλιάδων γονιδίων. Ωστόσο, ένα σημαντικό θέμα στις μελέτες miRNA είναι η έλλειψη προγραμμάτων βιοπληροφορικής για την ακριβή, γρήγορη και αποδοτική πρόβλεψη στόχων miRNA. Από την αρχική ανακάλυψη του miRNA lin-4 και μετά του miRNA let-7, το οποίο είναι εξαιρετικά συντηρημένο, οι συνδυασμένες πειραματικές και υπολογιστικές προσεγγίσεις είχαν ως αποτέλεσμα την ταυτοποίηση εκατοντάδων miRNAs σε ζωικά γονιδιώματα. Τα miRNAs των ζώων όμως έχουν περιορισμένη συμπληρωματικότητα αλληλουχίας με τους γονιδιακούς τους στόχους, πράγμα που καθιστά δύσκολη την κατασκευή μοντέλων πρόβλεψης στόχων με υψηλή εξειδίκευση.

Επιπρόσθετα, έρευνες έχουν αποδείξει ότι τα microRNAs είναι πολύ δραστικά και αποτελούν ρυθμιστές πολλών «ανθρώπινων» γονιδίων. Δρώντας στο μετα-μεταγραφικό επίπεδο, αυτά τα μόρια μπορούν να συντονίσουν την έκφραση όσο το 30% όλων των γονιδίων που κωδικοποιούν πρωτεΐνες θηλαστικών.

Για όλους αυτούς τους λόγους η Βιοπληροφορική καλείται λοιπόν, να λύσει απ' την πλευρά της τεχνολογίας αυτού του είδους τα προβλήματα με βέλτιστες μεθόδους όπως αυτές τις μηχανικής μάθησης και των νευρωνικών δικτύων καθώς και με άλλους έξυπνους τρόπους αναδιατύπωσης του προβλήματος. Πιο συγκεκριμένα, κάποια από τα αντικείμενα που καλείται να επιλύσει είναι η συλλογή, μελέτη, ομαδοποίηση ή διάσπαση ήδη εφευρημένων χαρακτηριστικών θέσεων στόχων miRNA σε ανθρώπινα γονιδιώματα.

Στα πλαίσια της μεταπτυχιακής διατριβής, έγινε η μελέτη και ανάλυση 6 αλγορίθμων που σκοπό έχουν να προβλέψουν θέσεις στόχους miRNA και συνεπώς καλούνται να αντιμετωπίσουν διάφορα απ' τα παραπάνω προβλήματα. Δύο απ' τους αλγορίθμους βασίζονται στην τεχνολογία της μηχανικής μάθησης. Κατά την ανάλυση συλλέχθηκαν κάποια βασικά χαρακτηριστικά για τον εκάστοτε αλγόριθμο αλλά και για τα datasets που χρησιμοποιεί ο καθένας.

Ως δεύτερο στάδιο της μεταπτυχιακής εργασίας είναι η εκτέλεση αυτών των αλγορίθμων και η συλλογή των αποτελεσμάτων. Μόνο για τον αλγόριθμο EIMMO2 δε βρέθηκαν αρκετές πληροφορίες και δεν εκτελέστηκε. Σύμφωνα, με τα αποτελέσματα των πέντε υπόλοιπων αλγορίθμων ακολούθησαν 2 βασικές προσεγγίσεις για την αξιολόγησή τους. Η 1η αφορά την αξιολόγηση της απόδοσης. Το ποσοστό των έγκυρων αλληλεπιδράσεων miRNAs-genes και το ποσοστό των συνολικών προβλέψεων από κάθε πρόγραμμα. Η διαδικασία αυτή έγινε έχοντας ως δεδομένο δύο σύνολα από επαληθευμένους στόχους miRNA. Σύμφωνα με την εύρεση αυτών των συνόλων υπολογίστηκαν κάποιες βασικές μετρικές απόδοσης προγραμμάτων, Sensitivity, Specificity, Precision. Η 2η αφορά τη συσχέτιση – ομοιότητα των αλγορίθμων, δημιουργώντας έναν πίνακα ομοιότητας Jaccard με βάση τις κοινές αλληλεπιδράσεις miRNA-γονιδίων και διαγράμματα Venn.

Λέξεις κλειδιά: miRNA, θέσεις, στόχοι, βιοπληροφορική, νευρωνικά δίκτυα, μηχανική μάθηση, χαρακτηριστικά, απόδοση, συσχέτιση

ABSTRACT

MicroRNAs (miRNAs) are involved in many biological processes and regulate the functions of thousands of genes. However, due to their recent discovery there is a lack of Bioinformatics algorithms with accurate, robust and efficient prediction of miRNA targets.

Since the initial discovery of miRNA lin-4 and afterwards the miRNA let-7, which persist gene evolution, combined experimental and computational studies have resulted in the discovery of hundreds of miRNAs in animal genes. Animal miRNAs have a restricted sequence with genomic targets, making it difficult to highlight any specific pattern.

Additional studies, have shown that microRNAs are highly reactive and regulate many of our human genes. MicroRNAs are a newly discovered class of uncoded RNAs that play key role in regulating gene expression. By tracing the post-transcriptional level, these fascinating molecules can coordinate gene expression as much as 30% of all genes in mammalian proteins.

To help further in collecting, studying, clustering or disrupting already introduced miRNA target site locations in human genomes, new Bioinformatics algorithms are starting to show up that employ cutting edge tools in analyzing data such as machine learning, neural networks as well as other sophisticated models and techniques.

In the context of postgraduate thesis, 6 algorithms were studied and analyzed in order to predict miRNA target sites. They are required to solve several of the above problems. Two of the algorithms are based on machine learning technology. During the analysis some key features were gathered for each given algorithm but also for the data sets used by each.

The second stage of postgraduate work is the execution of these algorithms and the collection of results. Not enough information was found about the EIMMO2 algorithm and was not executed. According to the results of the five remaining algorithms followed 2 key approaches to their evaluation. The first concerns the performance evaluation. The percentage of valid interactions of miRNAs-genes and the percentage of total program predictions. This process was done by obtaining two sets of verified miRNA targets. According to the estimation of these sets, some key metric performance programs, sensitivity, specificity, precision were calculated. The second relates to the correlation of the algorithms, creating a matrix correlation based on the common interactions of the miRNA genes and the Venn diagrams per four of the five algorithms.

Keywords: miRNA, sites, targets, bioinformatics, neural networks, machine learning, featuers, performance, correlation

Πίνακας περιεχομένων

Πινακας Εικονων

Πινακας Πινακων

ΠΙΝΑΚΑΣ ΓΡΑΦΗΜΑΤΩΝ

1.Κεφαλαιο - ΕΙΣΑΓΩΓΗ

1.1 Σκοπος

Τα τελευταία χρόνια η Βιοπληροφορική έχει αποδείξει σε όλους τους υπόλοιπους τομείς της επιστήμης την σπουδαιότητά της. Η φύση του ανθρώπου να προσπαθεί να εξηγήσει οτιδήποτε συμβαίνει γύρω του και πόσο μάλλον στον ίδιο του τον οργανισμό, καθώς και η ατελείωτη προσπάθειά του να λύνει προβλήματα, να εφεύρει θεραπείες και να εξελίσσει κατά κάποιον τρόπο το είδος του είναι η κινητήριος δύναμη που έχει φέρει τη βιοπληροφορική σε τόσο υψηλά επίπεδα.

Όσο εκείνη εξελίσσεται λοιπόν τόσες χιλιάδες microRNAs (miRNAs) έχουν εντοπιστεί. Αυτά τα miRNAs εμπλέκονται σε πολλές διαφορετικές βιολογικές διαδικασίες, όπως η ανάπτυξη, η διαφοροποίηση, η ιογενής λοίμωξη. Τα miRNAs λειτουργούν κυρίως μέσω κατασταλτικής ρύθμισης του επιπέδου έκφρασης των γονιδιακών στόχων τους. Τόσο οι υπολογιστικές όσο και οι πειραματικές μελέτες έχουν δείξει ότι χιλιάδες ανθρώπινα γονίδια πιθανόν να ρυθμίζονται από miRNAs. Λόγω των κρίσιμων ρόλων τους στη ρύθμιση της γονιδιακής έκφρασης, ο λειτουργικός χαρακτηρισμός των miRNAs έχει γίνει ένα από τα πιο ενεργά ερευνητικά πεδία στη βιολογία τα τελευταία χρόνια.

Ωστόσο, ένα σημαντικό θέμα που αντιμετωπίζει η έρευνα του miRNA είναι η έλλειψη υπολογιστικών εργαλείων για ακριβή, αποτελεσματική και γρήγορη πρόβλεψη στόχων. Αν και έχουν προταθεί πρόσφατα πολλαπλές υπολογιστικές προσεγγίσεις, αυτό παραμένει μια σημαντική πρόκληση για τους βιοπληροφορικούς λόγω της πολύ περιορισμένης αλληλεπίδρασης αλληλουχίας μεταξύ των miRNAs και των στόχων τους, καθώς και της έλλειψης πειραματικά επικυρωμένων γονιδιακών στόχων.

Μια στρατηγική για την πρόβλεψη στόχου είναι να χρησιμοποιηθούν προσεγγίσεις μηχανικής μάθησης. Οι μέθοδοι εκμάθησης μηχανών, όπως οι (SVMs), επιχειρούν να εξαγάγουν σχετικές πληροφορίες από δεδομένα, αυτόματα χρησιμοποιώντας υπολογιστικές και στατιστικές μεθόδους. Η μηχανική μάθηση έχει εφαρμοστεί σε πολλούς διαφορετικούς τομείς, συμπεριλαμβανομένης της βιολογικής έρευνας, αλλά δεν έχει εφαρμοστεί στην πρόβλεψη στόχων miRNA με μεγάλη επιτυχία μέχρι σήμερα. Ένα σημαντικό εμπόδιο είναι η έλλειψη δεδομένων υψηλής ποιότητας για την εκπαίδευση ισχυρών προγνωστικών μοντέλων. Υπάρχει μόνο περιορισμένος αριθμός επικυρωμένων στόχων miRNA από τη βιβλιογραφία. Επιπλέον, οι περισσότεροι από αυτούς τους στόχους επικυρώθηκαν επειδή είχαν προβλεφθεί στόχοι miRNA από υπάρχοντα προγράμματα. Ως αποτέλεσμα, τα δεδομένα επικύρωσης είναι «προκατειλημμένα» προς αυτούς τους αλγορίθμους και είναι λιγότερο χρήσιμα για την εξέλιξη και ανάπτυξη νέων αλγορίθμων πρόβλεψης στόχων.

Σκοπός αυτής της μεταπτυχιακής διατριβής λοιπόν είναι να αναφερθεί το θέμα των microRNAs και να επισημανθεί αφενός η σημασία τους και αφετέρου η αναγκαιότητα της εφεύρεσης βέλτιστων μεθόδων πρόβλεψης στόχων microRNA.

1.2 Διαρθρωση

Αυτή η εργασία διαρθρώνεται ως εξής, στο κεφάλαιο 2 που ακολουθεί ακριβώς μετά θα γίνει μια σχετικά μικρή αναφορά για τον ορισμό των microRNAs, τα πλεονεκτήματα και τα μειονεκτήματά τους καθώς και ποια είναι η σχέση τους με τον ανθρώπινο οργανισμό. Επίσης, στο ίδιο κεφάλαιο θα περιγραφεί και η έννοια της μηχανικής μάθησης μιας και σε αυτή την τεχνολογία στηρίζονται αρκετές μέθοδοι πρόβλεψης στόχων microRNA. Στη συνέχεια στο κεφάλαιο 3, θα γίνει η βασική ανάλυση έξι διαφορετικών

αλγορίθμων που σκοπό έχουν να προβλέψουν στόχους microRNA. Αφού ολοκληρωθεί η μελέτη και ανάλυση αυτών των αλγορίθμων, στο κεφάλαιο 4 θα αναφερθεί και η πειραματική ανάλυσή τους με σκοπό να ακολουθήσει στο κεφάλαιο 5 μια γενική συζήτηση, συγκρίνοντας τους αλγορίθμους αυτούς και βγάζοντας κάποια βασικά συμπεράσματα για την τρέχουσα κατάσταση καθώς και για την μελλοντική. Στο κεφάλαιο 6 υπάρχει όλο το επιπρόσθετο υλικό που χρειάστηκε για να εκελεστούν τα πειράματα, format και λεπτομέρειες των συνόλων δεδομένων που χρησιμοποιήθηκαν. Τέλος, στο κεφάλαιο 7 βρίσκεται η αντίστοιχη βιβλιογραφία της εργασίας.

2. Κεφαλαιο - ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 MicroRNAs

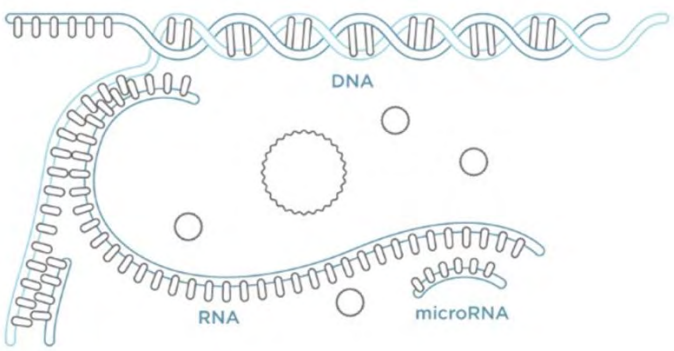
Το microRNA (miRNA) ανακαλύφθηκε αρχικά στο Caenorhabditis elegans από το εργαστήριο του Victor Ambros το 1993 ενώ μελέτησε το γονίδιο lin-14. Την ίδια στιγμή, ο Gary Ravkun ταυτοποίησε το πρώτο γονίδιο στόχο miRNA. Αυτές οι δύο πρωτοποριακές ανακαλύψεις προσδιόρισαν έναν νέο μηχανισμό ρύθμισης μετα-μεταγραφικού γονιδίου.

Εντούτοις, η σημασία του miRNA πραγματοποιήθηκε επτά χρόνια αργότερα, όταν τα εργαστήρια Ravukon και Horvitz ταυτοποίησαν ένα δεύτερο miRNA στο ίδιο μοντέλο νηματωδών ειδών (που ονομάστηκε let-7). Μόνο τότε έγινε φανερό ότι το μικρό μη κωδικοποιημένο μόριο RNA το οποίο προσδιορίστηκε το 1993 ήταν μέρος ενός πολύ μεγαλύτερου φαινομένου.

Έκτοτε, ένας αυξανόμενος αριθμός miRNAs έχει αναγνωριστεί στα θηλαστικά. Μόνο σε ανθρώπους έχουν ταυτοποιηθεί πάνω από 700 miRNAs και έχει προσδιοριστεί πλήρως η αλληλουχία τους και ο εκτιμώμενος αριθμός των miRNA γονιδίων σε ένα ανθρώπινο γονιδίωμα είναι μεγαλύτερο από 1000. Με βάση μοντέλα υπολογιστών, τα miRNAs στον άνθρωπο έχουν άμεση επίδραση στο τουλάχιστον 30% των γονιδίων σε όλο το γονιδίωμα.

Τα miRNAs αντιπροσωπεύουν μικρά μόρια RNA που κωδικοποιούνται στα γονιδιώματα φυτών και ζώων. Πρόκειται για μικρά (~ 22nt) μη-κωδικοποιημένα RNAs που καθοδηγούν το σύμπλεγμα που προκαλείται από το RNA (RISC) για να καταστείλει μετά την μεταγραφική έκφραση γονιδίων που κωδικοποιούν πρωτεΐνες με δέσμευση σε στοχευμένα mRNAs.

Αυτές οι εξαιρετικά διατηρημένες αλληλουχίες RNA ρυθμίζουν την έκφραση γονιδίων με δέσμευση στις 3'-αμετάφραστες περιοχές (3'-UTR) ειδικών mRNAs. Ένα αυξανόμενο σύνολο στοιχείων δείχνει ότι τα miRNAs είναι ένας από τους βασικούς παράγοντες στην κυτταρική διαφοροποίηση και ανάπτυξη, κινητικότητα και απόπτωση (προγραμματισμένος κυτταρικός θάνατος).



Εικόνα 1 Το microRNA

Μεταγραφή και επεξεργασία του microRNA

Τα γονίδια MicroRNA μεταγράφονται από την RNA πολυμεράση II ως μεγάλα πρωτογενή μεταγραφήματα (pri-microRNA) που υποβάλλονται σε επεξεργασία με ένα σύμπλεγμα πρωτεϊνών που περιέχει το ένζυμο Drosha της RNase III, για να σχηματιστεί ένα πρόδρομο μικροκλωνικό microRNA (pre-microRNA) περίπου 70 νουκλεοτιδίων. Αυτός ο πρόδρομος στη συνέχεια μεταφέρεται στο κυτταρόπλασμα όπου υποβάλλεται σε επεξεργασία με ένα δεύτερο ένζυμο RNase III, DICER, για να σχηματίσει ένα ώριμο microRNA (mature microRNA) περίπου 22 νουκλεοτιδίων . Το ώριμο microRNA στη συνέχεια ενσωματώνεται σε ένα ριβονουκλεϊκό σωματίδιο για να σχηματίσει το σύμπλεγμα σίγασης που προκαλείται από RNA, το RISC, το οποίο μεσολαβεί στη “σιωπή” του γονιδίου.

Τα miRNA ρυθμίζουν ποικίλες πτυχές της ανάπτυξης και της φυσιολογίας, κατανοώντας έτσι τον βιολογικό ρόλο τους που αποδεικνύεται όλο και πιο σημαντικός. Η ανάλυση της έκφρασης miRNA μπορεί να παρέχει πολύτιμες πληροφορίες, καθώς η δυσλειτουργία της λειτουργίας της μπορεί να οδηγήσει σε ανθρώπινες ασθένειες όπως ο καρκίνος, οι καρδιαγγειακές και μεταβολικές ασθένειες, οι καταστάσεις του ήπατος και η ανοσολογική δυσλειτουργία.

2.2 ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

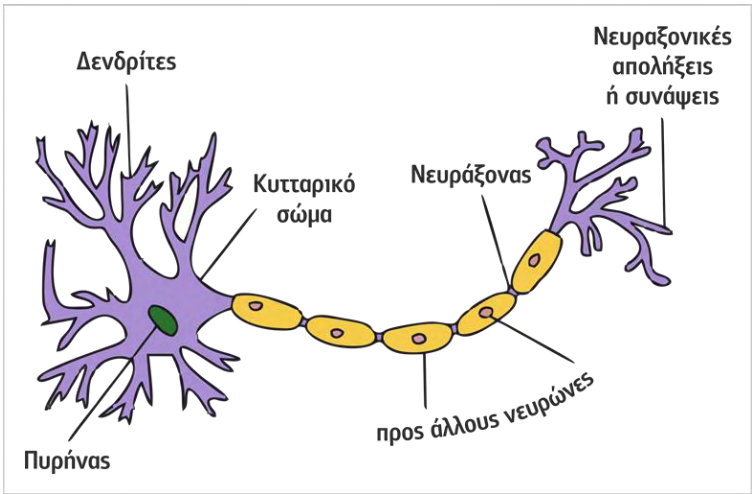
Ένα νευρωνικό δίκτυο είναι ένα σύνολο κόμβων ή νευρώνων που ο καθένας ενώνεται με όλους τους άλλους. Ουσιαστικά πρόκειται για ένα σύστημα επεξεργασίας πληροφοριών, σχεδιασμένο με βάση τη λογική συνδεσμολογίας των νευρώνων του ανθρώπινου εγκεφάλου. Το σημείο-κλειδί του συστήματος είναι η δικτυακή δομή του συστήματος επεξεργασίας πληροφορίας.

Αποτελείται από ένα μεγάλο αριθμό πολλαπλά διασυνδεδεμένων επεξεργαστών (νευρώνες), που δουλεύουν σε πλήρη συμφωνία μεταξύ τους. Τα νευρωνικά δίκτυα, όπως και ο άνθρωπος, έχουν τη δυνατότητα μάθησης μέσα από παραδείγματα. Στα βιολογικά συστήματα αυτό επιτυγχάνεται με την τροποποίηση των συνοπτικών συνδέσεων των νευρώνων. Η ίδια διαδικασία εφαρμόζεται και στα νευρωνικά δίκτυα. Ένα "εκπαιδευμένο" νευρωνικό δίκτυο είναι δυνατόν να θεωρηθεί "ειδήμονας" στην κατηγορία των πληροφοριών που του δίνονται για ανάλυση.

Ο όρος "Τεχνητά" για τα Νευρωνικά Δίκτυα χρησιμοποιείται για τον διαχωρισμό που γίνεται από τα Νευρωνικά Δίκτυα του ανθρώπινου εγκεφάλου. Τα Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ) είναι κατάλληλες απλουστεύσεις των νευρωνικών δικτύων του εγκεφάλου, δηλαδή εμπνέονται από τον τρόπο με τον οποίο ο εγκέφαλος παράγει πληροφορίες. Ένα τυπικό μοντέλο τεχνητού νευρωνικού δικτύου αποτελείται από διάφορα επίπεδα μονάδων επεξεργασίας. Η μονάδα επεξεργασίας μπορεί να θεωρηθεί ως ένας νευρώνας ή ομάδα νευρώνων. Ένας τεχνητός νευρώνας αθροίζει πληροφορίες από άλλους

νευρώνες, εκτελεί έναν απλό υπολογισμό σε αυτό το άθροισμα, π.χ. αποφασίζει αν είναι μεγαλύτερο από μια τιμή κατωφλίου και περνά το αποτέλεσμα σε άλλους νευρώνες. Ένα ΤΝΔ διαμορφώνεται για μια συγκεκριμένη εφαρμογή (όπως η αναγνώριση προτύπων ή η ταξινόμηση δεδομένων, η ανίχνευση ανωμαλιών σε καρδιογραφήματα και η αναγνώριση ύποπτων αντικειμένων στα αεροδρόμια) μέσω μιας διαδικασίας εκμάθησης.

Έτσι λοιπόν το κίνητρο για την ανάπτυξη της τεχνολογίας των νευρωνικών δικτύων προήλθε από την επιθυμία να αναπτυχθεί ένα τεχνητό σύστημα που θα μπορούσε να εκτελέσει τις "ευφυείς" στοιχειώδεις αυτές εργασίες, παρόμοιες με εκείνες που εκτελούνται από τον ανθρώπινο εγκέφαλο αφού τα νευρωνικά δίκτυα μοιάζουν με τον ανθρώπινο εγκέφαλο. Ένα νευρωνικό δίκτυο αποκτά τη γνώση μέσω της εκμάθησης. Η γνώση ενός νευρωνικού δικτύου καταχωρείται μέσα στις δια-νευρωνικές δυνάμεις, σύνδεσης γνωστές ως συνοπτικά βάρη.

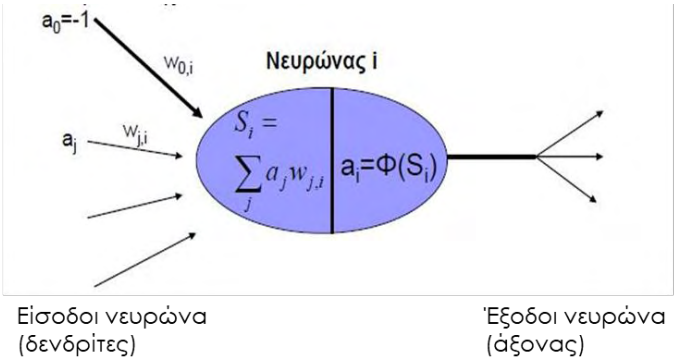


Εικόνα 2 Το νευρωνικό δίκτυο

2.3 Η ΔΟΜΗ ΤΟΥΣ

Ένας νευρώνας αποτελείται από το κυρίως σώμα, τον άξονα και τους δενδρίτες.

- Σώμα (Soma): Είναι ο κυρίως κορμός του νευρώνα μέσα στο οποίο βρίσκεται ο πυρήνας. Στον πυρήνα βρίσκεται όλο το γενετικό υλικό του οργανισμού, όπου γίνεται η πιο έντονη χημική δράση του κυττάρου για την σύνθεση των ενζύμων, πρωτεϊνών, και άλλων μορίων που είναι απαραίτητα για τη ζωή.
- Ο άξονας (Axon): Παράγει τη σύνδεση με άλλα νευρώνια του κεντρικού νευρικού συστήματος. Κάθε νευρώνας έχει μόνο έναν άξονα ο οποίος μεταδίδει τα σήματα σε άλλους νευρώνες.
- Δενδρίτης (Dendrite): Επεκτάσεις που χρησιμεύουν στη λήψη ερεθισμού. Ορισμένα νευρώνια διαθέτουν αρκετά πολύπλοκα τέτοια δέντρα ενώ άλλα είναι εξοπλισμένα μόνον με λίγες τέτοιες επεκτάσεις των νευρωνίων.
- Σύναψη (Synapse): Είναι το σημείο σύνδεσης ενός άκρου του άξονα με ένα άλλο κύτταρο.



Εικόνα 3 Η δομή του νευρώνα

Τα 4 στοιχεία (δενδρίτης, σύναψης, άξονας, σώμα) σχηματίζουν την ελάχιστη δομή που μπορεί να ληφθεί από τα βιολογικά μοντέλα, έτσι ώστε να σχηματισθούν τα τεχνητά νευρώνια.

Όπως ο ανθρώπινος εγκέφαλος, έτσι και ένα νευρωνικό δίκτυο αποτελείται από τους νευρώνες και τις συνδέσεις μεταξύ τους. Οι νευρώνες μεταφέρουν τις εισερχόμενες πληροφορίες στις εξερχόμενες συνδέσεις τους και κατόπιν σε άλλους νευρώνες. Στα νευρωνικά δίκτυα αυτές οι συνδέσεις καλούνται βάρη.

Είδαμε ότι οι αριθμοί των βιολογικών νευρώνων και των συνδέσεων τους είναι πράγματι πολλοί μεγάλοι. Η τάξη μεγέθους τους είναι πολύ μεγαλύτερη από τον αριθμό μονάδων που μπορεί να χειριστεί ένας ηλεκτρονικός υπολογιστής. Τα τεχνητά νευρωνικά δίκτυα (ΤΝΔ) δεν μπορούν να κάνουν πράγματα που ο εγκέφαλος ακόμη και ενός μικρού παιδιού επιτελεί με μεγάλη ευκολία. Ένα τεχνητό νευρωνικό δίκτυο μπορεί να έχει εκατοντάδες ή χιλιάδες νευρώνες αλλά όχι την τάξη μεγέθους που έχει ο εγκέφαλος. Οι συνάψεις στα βιολογικά είναι πολύ πιο περίπλοκες από ότι στα τεχνητά. Αυτή ίσως είναι η βασικότερη διαφορά τους. Η ταχύτητα στους υπολογιστές, είναι χιλιάδες φορές μεγαλύτερη από την ταχύτητα διάδοσης των βιολογικών δικτύων κάτι το οποίο έγκειται κατά μεγάλο βαθμό στη διαφορά πολυπλοκότητας. Ο εγκέφαλος μπορεί να κάνει σύγχρονη ή ασύγχρονη ενημέρωση των μονάδων του δηλ. και σε συνεχή χρόνο, ενώ τα ΤΝΔ κάνουν σύγχρονη ενημέρωση δηλ. σε διακριτό χρόνο.

Η διαδικασία εκτέλεσης ΤΝΔ

Οι πληροφορίες (είσοδος) στέλνονται στο νευρώνα στα εισερχόμενα βάρη του. Στη συνέχεια οι τιμές όλων προστίθενται των εισερχόμενων βαρών. Η προκύπτουσα αξία συγκρίνεται με μια συγκεκριμένη αξία ευαισθησίας. Εάν η είσοδος υπερβεί την αξία ευαισθησίας, ο νευρώνας θα ενεργοποιηθεί, διαφορετικά θα εμποδιστεί. Εάν ενεργοποιηθεί, ο νευρώνας στέλνει μια έξοδο στα εξερχόμενα βάρη του σε όλους τους συνδεδεμένους νευρώνες κ.ο.κ.

2.4 ΤΑ ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΤΟΥΣ

- Τη δυνατότητά τους να αντιπροσωπεύουν γραμμικές και μη γραμμικές σχέσεις και τη δυνατότητά τους να μάθουν αυτές τις σχέσεις άμεσα από τα στοιχεία που διαμορφώνονται.
- Δυνατότητα προσαρμοστικής εκμάθησης. Το δίκτυο είναι σε θέση να μάθει να εκτελεί διαδικασίες βασισμένες στα δεδομένα που του δίνονται.
- Αυτό-οργάνωση. Ένα νευρωνικό δίκτυο είναι σε θέση να οργανώνεται με την πάροδο του χρόνου και με βάση τα δεδομένα που δέχεται από τις διάφορες εισόδους του.

- Επεξεργασία σε συνθήκες πραγματικού χρόνου. Οι υπολογισμοί ενός νευρωνικού δικτύου γίνονται ιδιαίτερα γρήγορα, αφού η σχεδίασή τους είναι βασισμένη σε παράλληλη αρχιτεκτονική των επεξεργαστών του συστήματος.
- Δυνατότητα διόρθωσης λαθών. Ανοχή σφάλματος μέσο της περιττής κωδικοποίησης πληροφοριών. Αυτό σημαίνει ότι, εάν ένα νευρωνικό δίκτυο έχει εκπαιδευτεί για ένα συγκεκριμένο πρόβλημα, θα είναι σε θέση να ανακαλέσει τα σωστά αποτελέσματα, ακόμα κι αν το πρόβλημα προς λύση δεν είναι ακριβώς το ίδιο με αυτό στο οποίο έχει γίνει η εκμάθηση. Παραδείγματος χάριν, ας υποθέσουμε ότι ένα νευρωνικό δίκτυο έχει εκπαιδευτεί για να αναγνωρίζει την ανθρώπινη ομιλία. Κατά τη διάρκεια της διαδικασίας εκμάθησης, ένα συγκεκριμένο άτομο πρέπει να προφέρει μερικές λέξεις, τις οποίες μαθαίνει το δίκτυο. Εάν το νευρωνικό δίκτυο εκπαιδευτεί σωστά, πρέπει να είναι σε θέση να αναγνωρίσει εκείνες τις λέξεις και όταν αυτές λεχθούν από ένα άλλο πρόσωπο.

Τέλος, τα νευρωνικά δίκτυα κατασκευάζονται για να λύσουν τα προβλήματα που δεν μπορούν να λυθούν χρησιμοποιώντας τους συμβατικούς αλγορίθμους. Τέτοια προβλήματα είναι συνήθως προβλήματα βελτιστοποίησης ή ταξινόμησης όπως και το πρόβλημα της πρόβλεψης στόχων miRNA κατ' επέκταση της συλλογής χαρακτηριστικών διαφόρων θέσεων στόχων σ' ένα γονίδιο.

3. ΚΕΦΑΛΑΙΟ - ΑΝΑΛΥΣΗ ΑΛΓΟΡΙΘΜΩΝ

3.1 MIRTARGET2

3.1.1 ΕΙΣΑΓΩΓΗ

MirTarget2, πρόκειται για ένα πρόγραμμα πρόβλεψης στόχου miRNA που βασίζεται σε Support Vector Machines (SVMs) και σε ένα σύνολο δεδομένων μάθησης από microarrays (training dataset). Με τη συστηματική ανάλυση των δημόσιων δεδομένων microarrays, εντοπίστηκαν χαρακτηριστικά, τα οποία είναι σημαντικά για την πρόβλεψη downregulated στόχων (υπεύθυνα για την ρύθμιση προς τα κάτω μιας γονιδιακής έκφρασης). Με μη γραμμικό τρόπο ενσωματώνονται όλα τα χαρακτηριστικά που είναι απαραίτητα για την πρόβλεψη στόχων μέσω του μοντέλου μηχανικής μάθησης. Ο αλγόριθμος αυτός έχει επικυρωθεί με ανεξάρτητα πειραματικά δεδομένα για τη βελτιωμένη απόδοσή του.

3.1.2 ΣΥΝΟΨΗ

1) Γενική περιγραφή δεδομένων για εκπαίδευση
2) Πηγή προέλευσης
3) Το πλήθος τους
4) 1,2,3 για τα δεδομένα αξιολόγησης
5) Χαρακτηριστικά
6) Μέθοδοι μηχανικής μάθησης

7) Περιοχή γονιδιώματος που γίνεται η πρόβλεψη

8) Τύπος πρόσδεσης των miRNAs

9) Dependencies

---

1) Microarrays μεταγραφικού προφίλ από διάφορες αλληλουχίες γονιδίων ανθρώπινων κατά βάση αλλά και άλλων οργανισμών – αρουραίων, ποντικιών, σκύλων, κοτόπουλων – για την αξιοποίηση ενός χαρακτηριστικού, αυτού της συντηρησιμότητας. Με βάση 2 διαφορετικές μελέτες,

a. **Linsley**: let-7c, miR-103, miR-106b, miR-141, miR-15a και miR-215 επιλέχθηκαν για εκπαίδευση και δοκιμή μοντέλων

b. **Wang**: Overexpression του miR-124a

2) Από βιβλιογραφίες, από άλλους δημοσιευμένους αλγορίθμους - TargetScan, PicTar, miRanda, mirTarget1- Οι αλληλουχίες mRNA και τα αρχεία σχολιασμού λήφθηκαν από τις βάσεις δεδομένων NCBI.

3) Υπήρχαν 1.401 downregulated γονίδια (θετικά) και 16.761 φυσιολογικά γονίδια (αρνητικά) που ταυτοποιήθηκαν. Αναλύθηκαν 454 downregulated γονίδια (θετικά δείγματα) και 1017 φυσιολογικά γονίδια (αρνητικά).

4) Συσχέτιση του Spearman (Rs), καμπύλες ROC, καμπύλες Precision – Recall, στατιστικές Wald, σύγκριση με άλλους αλγορίθμους, σύγκριση με ανεξάρτητα δεδομένα που δεν είχαν λάβει μέρος στο σύνολο των δεδομένων για εκπαίδευση, υπολογισμός score για κάθε υποψήφια περιοχή 3' UTR. Υπήρχαν 151.087 προβλεπόμενοι στόχοι miRNA από 15.059 μοναδικά γονίδια στον άνθρωπο.

5) Seed conservation

Other seed types

Base Composition

Secondary structure

Location in 3'UTR

6) SVMs, LIBSVM με 10-fold cross validation, RNAfold (ΔG), DREES σύμφωνα με μέθοδο logistic regression πολλαπλών μεταβλητών, ανεξάρτητο t test ή χ^2 test για εύρεση P-value,

7) 3' UTR

8) οι θέσεις 1-8 (που ορίζονται ως seed 8)

οι θέσεις 1-7 (ως seed 7a)

οι θέσεις 2-8 (ως seed 7b)

οι θέσεις 2-7 (ως seed 6)

9) Οι αλληλουχίες μεταγραφής 3' UTR από άνθρωπο, ποντίκι, αρουραίο, σκύλο και κοτόπουλο αναλύθηκαν από τα αρχεία GenBank με BioPerl (http://www.bioperl.org). Οι προβλεπόμενοι στόχοι μεταγραφής εισήχθησαν έπειτα στο miRDB. Πολλαπλές μεταγραφές από τα ίδια γονίδια

χαρτογραφήθηκαν χρησιμοποιώντας αρχεία δείκτη γονιδίου NCBI και η μεταγραφή με το υψηλότερο σκορ πρόβλεψης στόχου παρουσιάστηκε στην ιστοσελίδα http://mirdb.org.

3.1.3 Δομη Αλγοριθμου

Η βασική δομή του αλγορίθμου mirTarget2, βασίζεται σε 3 κύριες μεθόδους:

1. Η ανάκτηση των δεδομένων
2. Η ταυτοποίηση μεταξύ downregulated γονιδίων και overexpression miRNAs
3. Η υπολογιστική ανάλυση

Ανάκτηση Δεδομένων (1),(2),(9)**

- Όλες οι αλληλουχίες mRNA και τα mapping index files για κάθε γονίδιο λήφθηκαν από το site NCBI.
- Οι αλληλουχίες 3'UTR αναλύθηκαν με BioPerl χρησιμοποιώντας τις παρατηρήσεις του GenBank.
- Οι σχέσεις ορθολογικών γονιδίων προβλέπονταν βάσει του NCBI HomoloGene.
- Αλληλουχίες 3' UTR από μετάγραφα, σε ποντίκια, αρουραίους, σκύλους και κοτόπουλα παρασκευάστηκαν και συμπεριλήφθηκαν στην υπολογιστική ανάλυση.
- Τα δεδομένα microarray λήφθηκαν από τη βάση δεδομένων NCBI GEO για δύο δημοσιευμένες μελέτες αντίστοιχα:
  i. Στη μελέτη Linsley, πολλαπλά miRNAs επίμολύνθηκαν σε δύο κυτταρικές σειρές και η γενική επίδραση της υπέρ-έκφρασης (overexpression) που δημιουργήθηκε εξετάστηκε με microarrays.
  ii. Στη μελέτη Wang, το miR-124a επιμολύνθηκε σε κύτταρα HepG2 και οι μεταβολές στα προφίλ γενικής γονιδιακής έκφρασης αξιολογήθηκαν με microarrays σε διαφορετικά χρονικά σημεία.
- Οι προβλεπόμενοι στόχοι miRNA από διάφορους δημοσιευμένους ήδη υπάρχοντες αλγορίθμους ανακτήθηκαν από τους δημόσιους ιστοτόπους (TargetScan, PicTar, miRanda, mirTarget1).
- Τα ids των προβλεπόμενων στόχων αυτών συσχετίσθηκαν με αυτά των γονιδίων που πάρθηκαν από το NCBI (με τα NCBI Gene IDs).
- Για την ανάλυση των δεδομένων, χρησιμοποιώντας:
  - Το σύνολο δεδομένων Linsley, οι αλληλουχίες εκπαίδευσης προεπιλέχτηκαν για τις θέσεις που γίνεται το αντίστοιχο match (seed matching sites). Έτσι, οι προβλεπόμενοι στόχοι χωρίς θέσεις αντιστοίχισης δε συμπεριλήφθηκαν στην ανάλυση.
  - Με τη χρήση του συνόλου δεδομένων Wang, όλοι οι προβλεπόμενοι στόχοι συμπεριλήφθηκαν στην ανάλυση.

** Αντιστοιχούν στα νούμερα από την Σύνοψη.

Ταυτοποίηση Downregulated Γονιδίων Από Υπέρ-εκφράσεις miRNA

- **Μελέτη Linsley:**

Δύο κυτταρικές γραμμές (HCT116 Dicerex5 και DLD-1 Dicerex5) συμπεριλήφθηκαν στη μελέτη. Τα περισσότερα από τα miRNAs επιμολύνθηκαν και στις δύο κυτταρικές αλληλουχίες για να αξιολογηθούν οι μεταβολές στα προφίλ γενικής γονιδιακής έκφρασης.

Έξι miRNAs από το σύνολο δεδομένων Linsley, let-7c, miR-103, miR-106b, miR-141, miR-15a και miR -215 επιλέχθηκαν για εκπαίδευση και δοκιμή μοντέλων. Όλα εκτός από ένα miRNA επιμολύνθηκαν και στα δύο κύτταρα HCT116 Dicerex5 και DLD-Dicerex5 και τα υποψήφια μετάγραφα επιλέχθηκαν με ανάλυση των δεδομένων microarray και από τις δύο κυτταρικές σειρές.

Υπήρχαν 1.401 downregulated γονίδια (θετικά) και 16.761 φυσιολογικά γονίδια (αρνητικά) που ταυτοποιήθηκαν κατ' αυτόν τον τρόπο. **(3)**

Ένα γονίδιο ορίστηκε ως προς τα κάτω ρυθμιζόμενο (downregulated) αν σε σύγκριση με την ψευδή επιμόλυνση, το επίπεδο έκφρασής του μειώθηκε κατά τουλάχιστον 40% με τιμή $P < 0.001$ σε οποιαδήποτε κυτταρική γραμμή.

Ένα γονίδιο ορίστηκε ως φυσιολογικό (normal) εάν το επίπεδο γονιδιακής έκφρασης ήταν τουλάχιστον 95%, αλλά όχι περισσότερο από 120% με τιμή $P > 0.3$ και στις δύο κυτταρικές σειρές.

- **Μελέτη Wang:**

Αυτή ήταν μια χρονική μελέτη που αξιολόγησε την επίδραση της υπέρ-έκφρασης του miR-124a σε γενικά μεταγραφικά προφίλ. Τα downregulated γονίδια σε κάθε χρονικό σημείο ταυτοποιήθηκαν με τα ίδια κριτήρια που περιεγράφηκαν προηγουμένως.

Υπολογιστική Ανάλυση (6)

- Το πακέτο SVM LIBSVM χρησιμοποιήθηκε για την κατασκευή μοντέλων πρόβλεψης στόχων miRNA. Οι παράμετροι του μοντέλου εκπαίδευσης βελτιστοποιήθηκαν με πολλαπλούς κύκλους διασταυρούμενης επικύρωσης για να ελαχιστοποιηθεί ο κίνδυνος υπερβολικής εκπαίδευσης.
- Η διαδικασία κατασκευής μοντέλου αποτελείται από δύο βήματα: **(1)** επιλογή του μοντέλου **(2)** εκτίμηση των παραμέτρων του μοντέλου. Στην περίπτωσή αυτή, επελέγη 10 φορές η διασταυρούμενη επικύρωση για την επιλογή μοντέλου σε ανάλυση σταδιακής παλινδρόμησης.
- Η βασική ιδέα είναι να μεγιστοποιηθεί ο διαχωρισμός μεταξύ δύο ομάδων δεδομένων (θετικών και αρνητικών) σε ένα μη γραμμικό χώρο χαρακτηριστικών.
- Η σταθερότητα της δευτεροταγούς δομής RNA, που αντιπροσωπεύεται από την τιμή ΔG, υπολογίστηκε με RNAfold. Οι δευτερεύουσες δομές που προβλέπονται από το RNAfold αναλύθηκαν για να αναγνωριστούν τα νουκλεοτίδια που συνδυάστηκαν με αυτές τις δομές. Ο στατιστικός υπολογισμός πραγματοποιήθηκε με MATLAB και το πακέτο R.
- Η στατιστική σπουδαιότητα (P-value) για τα χαρακτηριστικά εκπαίδευσης υπολογίστηκε με ανεξάρτητο t test ή χ^2 test.
- Χρησιμοποιήθηκε το πακέτο DREES το οποίο είναι κατασκευασμένο με MATLAB, για την εξαγωγή των κορυφαίων σχετικών χαρακτηριστικών εκπαίδευσης.
- Το λογισμικό αυτό εφαρμόζει σταδιακή ανάλυση παλινδρόμησης (logistic regression) πολλαπλών μεταβλητών σε συνδυασμό με μεθόδους ανάδειγματοληψίας για τον εντοπισμό των πλέον σχετικών χαρακτηριστικών.
- Χρησιμοποιήθηκε επίσης για την αξιολόγηση μοντέλων πρόβλεψης με διαφορετικά σύνολα χαρακτηριστικών.

3.1.4 ΕΠΙΛΟΓΗ ΑΛΛΗΛΟΥΧΙΩΝ ΓΙΑ ΕΚΠΑΙΔΕΥΣΗ - SEEDS

Υπάρχουν τέσσερις κύριοι τύποι seed sequence: **(8)**

1. οι θέσεις 1-8 (που ορίζονται ως seed 8)
2. οι θέσεις 1-7 (ως seed 7a)
3. οι θέσεις 2-8 (ως seed 7b)
4. οι θέσεις 2-7 (ως seed 6)

Τα δεδομένα εκπαίδευσης αναλύθηκαν για να προσδιοριστεί ποιος τύπος ήταν πιο σχετικός με την ταυτοποίηση γονιδίων - downregulated. Το αποτέλεσμα συνοψίζεται στον Πίνακα 1.

Το seed 8 εμπλουτίστηκε 11,9 φορές στα downregulated γονίδια σε σύγκριση με τα φυσιολογικά γονίδια και αυτό ήταν στατιστικά πολύ σημαντικό (P=9,4E - 135). Ωστόσο, μόνο ένα μικρό ποσοστό (16,5%) των downregulated γονιδίων έχει σε seed 8 match στην 3' UTR.

Ο λόγος εμπλουτισμού (enrichment ratio) ορίστηκε ως το κλάσμα του ποσοστού των UTRs με ταίριασμα στα downregulated γονίδια προς το ποσοστό στα κανονικά γονίδια Πίνακας 1.

Downregulated genes/Normal genes

Από την άλλη μεριά, οι 6 αντίστοιχες θέσεις ήταν παρούσες στο 67,4% των downregulated γονιδίων. Ωστόσο, αυτός ο τύπος seed δεν εξετάστηκε περαιτέρω επειδή ένας μεγάλος αριθμός φυσιολογικών γονιδίων είχε επίσης σε seed 6 και συνεπώς ο λόγος εμπλουτισμού ήταν χαμηλός.

Seeds 7a και 7b εξετάστηκαν με παρόμοιο τρόπο. Οι θέσεις ζευγαρώματος των seed 7a εμπλουτίστηκαν 2,2 φορές σε 3' UTR περιοχές χωρίς καμία αντιστοιχία seed 7b. Σε σύγκριση, οι θέσεις ζευγαρώματος seed 7b εμπλουτίστηκαν σε ένα πολύ υψηλότερο επίπεδο (4,4 φορές) σε 3' UTR περιοχές χωρίς να ταιριάζει με seed 7a. Εάν η περιοχή τύπου seed 7b θεωρηθεί μόνη, ταυτοποιήθηκε το 43,5% των downregulated γονιδίων και ο λόγος εμπλουτισμού ήταν υψηλότερος από εκείνον από τη χρήση και των δύο seeds 7a και 7b, που υποδηλώνουν καλύτερο λόγο σήματος ως προς το θόρυβο. Έτσι, εξετάστηκε πρωτίστως η περιοχή τύπου seed 7b στην πρόβλεψη στόχων.

Τέλος, εφόσον μια θέση ζευγαρώματος seed 8 ταιριάζει επίσης με seed 7b (seed 7b συν ένα αντίστοιχο τερματικό βάσης είναι ισοδύναμο με seed 8), όλα τα ταιριάσματα seed 8 αξιολογήθηκαν στο πλαίσιο του seed 7b. Οι περισσότερες θέσεις ζευγαρώματος seed 7a (82,4%) εξετάστηκαν επίσης στην ανάλυση αυτή, επειδή μόνο ένα μικρό ποσοστό τύπου seed 7a αντιστοιχούσε σε 3' UTR χωρίς θέση seed 7b.

Στη μελέτη, επιλέχθηκαν 30 αλληλουχίες UTR με μόνο μεμονωμένες θέσεις ζευγαρώματος seeds. Με τον τρόπο αυτό, συμπεριλήφθηκαν 454 downregulated γονίδια και 1017 φυσιολογικά γονίδια στο σύνολο δεδομένων εκπαίδευσης για τις μοναδικές τους τοποθεσίες αντιστοίχισης seed 7b.

Seed type	Downregulated genes (%)	Normal genes (%)	Enrichment ratio
seed 8	16.5	1.4	11.9
seed 6	67.4	22.9	2.9
seed 7a; no seed 7b	9.3	4.3	2.2
seed 7b; no seed 7a	22.8	5.2	4.4
seed 7a or 7b	52.8	11.4	4.6
seed 7b	**43.5**	**7.1**	**6.1**

Πίνακας 1 Τύποι πρόσδεσης των miRNAs, seeds

από τη δημοσίευση: Prediction of both conserved and nonconserved microRNA targets in animals, Xiaowei Wang and Issam M. El Naqa [Αναφ. 8].

3.1.5 ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Εντοπίστηκαν νέα χαρακτηριστικά σημαντικά για το διαχωρισμό μεταξύ θετικών και αρνητικών δειγμάτων, τα οποία κατηγοριοποιούνται σε 5 κατηγορίες: **(5)**

- Seed conservation – Διατηρησιμότητα / Συντηρησιμότητα των seeds.
    - Η περιοχή που γίνεται το matching μεταξύ γονιδίου και miRNA (seed) σε 3' UTR διατηρείται συχνά σε πολλά είδη. Αυτό είναι ένα πρωτεύον φίλτρο επιλογής στους περισσότερους υπάρχοντες αλγορίθμους πρόβλεψης στόχων. Μια υποψήφια περιοχή συνήθως απορρίπτεται αν δεν πληρεί ένα επίπεδο κατωφλίου. Μετάγραφα σε 3' UTR από ορθολογικό ανθρώπινο γονίδιο σε ποντίκια, αρουραίους, σκύλους και κοτόπουλα αναλύθηκαν για τον εντοπισμό περιοχών ταιριάσματος miRNA και καταγράφηκε το επίπεδο συντήρησης αυτών των περιοχών. Οι υποψήφιες θέσεις από τα downregulated γονίδια ήταν σημαντικά πιο συντηρημένες από εκείνες των φυσιολογικών γονιδίων (P = 7,4E - 29). Αυτό το αποτέλεσμα συμφωνεί με πολλές προηγούμενες μελέτες που δείχνουν τη σημασία της διατήρησης των σπόρων στην πρόβλεψη στόχων.
- Other seed types – Άλλοι τύποι seeds
    - Μία θέση seed 7b είναι επίσης μία θέση seed 8 εάν η τερματική βάση αποτελεί matching, δηλαδή είναι ίδια.
    - 3' UTRs στα downregulated γονίδια εμπλουτίστηκαν πολύ σε θέσεις seed 8. Επομένως,
        - η αντιστοιχία τερματικής βάσης καταγράφηκε ως χαρακτηριστικό εκπαίδευσης. Η παρουσία του seed 7a μπορεί επίσης να είναι σημαντική για την αναγνώριση στόχου miRNA και συνεπώς
        - η συνολική καταμέτρηση των θέσεων seed 7a σε μία περιοχή 3' UTR (εκτός από τη θέση seed 7b) καταγράφηκε ως ένα άλλο χαρακτηριστικό εκπαίδευσης.
- Base composition – Σύνθεση βάσεων
    - Τοπικές 3' UTR περιοχές δίπλα στις θέσεις ζευγαρώματος seeds αναλύθηκαν για τη σύνθεση βάσης τους. Σε σύγκριση με τις υποψήφιες θέσεις στα φυσιολογικά γονίδια, οι

θέσεις στα downregulated γονίδια είχαν σημαντικά χαμηλότερη περιεκτικότητα σε GC γενικά (P=3.7E-46) και αυτό συμφωνούσε με προηγούμενες μελέτες. Και οι τέσσερις μετρήσεις βάσεων ήταν σημαντικά διαφορετικές μεταξύ των υποψήφιων θέσεων στα downregulated και κανονικά γονίδια, με την C ως την πιο υπό- εκπροσωπούμενη βάση στις θέσεις ελέγχου των downregulated γονιδίων (P=2.4E-18). Εκτός από τον αριθμό μονονουκλεοτιδίων, προσδιορίστηκαν οι συχνότητες και των 16 δινουκλεοτιδίων. Πολλοί από τους αριθμούς δινουκλεοτιδίων ήταν στατιστικά σημαντικοί, με τις UU, GC, AA και CC να κατατάσσονται στην κορυφή της λίστας (Πίνακας 3).

- Ωστόσο, υπήρξαν επίσης μερικές ενδιαφέρουσες παρατηρήσεις που μπορεί να σχετίζονται με συγκεκριμένες λειτουργικές απαιτήσεις. Για παράδειγμα, το C ήταν σημαντικά απών στις θέσεις -4 έως -1. Αυτό δεν μπορεί να εξηγηθεί από την απαίτηση χαμηλού περιεχομένου GC, καθώς η διαφορά G ήταν πολύ μικρότερη στις ίδιες θέσεις. Ένα τυπικό αμφίδρομο δέσμευσης miRNA / στόχου έχει δομή βρόχου ή διόγκωσης στην περιοχή αυτή και η απουσία των Cs μπορεί να σχετίζεται με αυτή την απαίτηση. Ένα άλλο παράδειγμα είναι η μέτρηση A στις θέσεις -5 και 7 (αμέσως μετά τη θέση δέσμευσης seed). Αν και δεν υπήρχε διαφορά στις μετρήσεις T σε αυτές τις θέσεις, οι A μετρήσεις ήταν σημαντικά υψηλότερες στις υποψήφιες θέσεις από τα downregulated γονίδια. Αυτές οι μετρήσεις ειδικών θέσεων βάσεων (20 βάσεις που περιβάλλουν τη θέση δέσμευσης seed) καταγράφηκαν ως χαρακτηριστικά εκπαίδευσης SVM.

- Secondary structure – Δευτεροταγής δομή
  - Μια θέση-στόχος δεν είναι πιθανόν να είναι λειτουργική εάν είναι απρόσιτη για τη δέσμευση του miRNA. Η πρόβλεψη δευτερογενούς δομής RNA εξακολουθεί να είναι δύσκολο έργο μέχρι σήμερα και γενικά η ακρίβεια της πρόβλεψης μειώνεται δραματικά καθώς το μήκος της αλληλουχίας αυξάνεται. Στη μελέτη αυτή, ένα σύντομο μήκος 25 nts σε κάθε πλευρά της τοποθεσίας αντιστοίχισης seeds συμπεριλήφθηκε στον υπολογισμό. Επίσης δοκιμάστηκαν και άλλα διάφορα μικρά μήκη αλληλουχίας και τα αποτελέσματα ήταν παρόμοια με αυτά που παρουσιάζονται εδώ. Η τοπική δευτερογενής δομή μιας υποψήφιας περιοχής υπολογίστηκε με RNAfold. Η τάση για σχηματισμό δευτεροταγούς δομής (μετρούμενη με το ΔG) ήταν σημαντικά υψηλότερη στις υποψήφιες θέσεις από τα φυσιολογικά γονίδια από εκείνες από τα downregulated. Εκτός από τη συνολική προσβασιμότητα μιας υποψήφιας θέσης, αξιολογήθηκε ο συνδυασμός βάσεων μεμονωμένων νουκλεοτιδίων και συνδυάστηκαν σημαντικές θέσεις (ο δείκτης προσβασιμότητας) για να ληφθεί το πιθανό αποτέλεσμα της συγκεκριμένης θέσης.
  - Οι πιθανές δομές υβριδοποίησης miRNA / στόχου υπολογίστηκαν για να προσδιορίσουν το επίπεδο της σύζευξης βάσεων νουκλεοτιδίων σε κάθε θέση. Τα νουκλεοτίδια 13-17 σε ένα miRNA ήταν πιο πιθανό να είναι ζευγαρωμένα σε βάσεις προς τις υποψήφιες θέσεις στα downregulated γονίδια παρά σε εκείνα στα κανονικά γονίδια. Αυτή η παρατήρηση ήταν σύμφωνη με προηγούμενες μελέτες που έδειξαν ότι η περιορισμένη ζεύξη βάσεων στην περιοχή 3' UTR μιας περιοχής στόχου συμβάλλει στην αναγνώριση στόχων miRNA. Από την άλλη πλευρά, η θέση 10 ήταν περισσότερο μη δομημένη στα γονίδια downregulated. Ένα τυπικό δίκλωνο που δεσμεύει το miRNA έχει μια διόγκωση / βρόχο μετά την περιοχή δέσμευσης seed, η οποία πιθανώς οδηγεί σε μία εκτεθειμένη βάση στη θέση 10. Αυτές οι σημαντικές θέσεις βάσης συνδυάστηκαν (ο δείκτης υβριδοποίησης) για να εκτιμηθεί ο γενικός συνδυασμός βάσης ειδικής θέσης του διπλού δεσμού. Οι συνολικές ευθυγραμμίσεις αλληλουχίας μεταξύ των miRNAs και των δυνητικών θέσεων στόχων τους προσδιορίστηκαν επίσης χρησιμοποιώντας τον

αλγόριθμο Smith-Waterman, αλλά δεν υπήρξε σημαντικός εμπλουτισμός στην ευθυγράμμιση με τα downregulated γονίδια.

- Location in 3' UTR – Τοποθεσία σε 3' UTR περιοχή **(7)**
    - Προηγούμενες μελέτες έχουν δείξει ότι η λειτουργία μιας θέσης δέσμευσης στόχου σχετίζεται με τη θέση της σε 3' UTR και μια θέση στη μέση ενός μακρού UTR είναι λιγότερο πιθανό να είναι λειτουργική. Στην ανάλυση αυτή, οι υποψήφιες θέσεις από τα downregulated γονίδια και τα φυσιολογικά γονίδια εξετάστηκαν για να προσδιοριστεί αν ήταν τουλάχιστον 600 ή 900 νουκλεοτίδια μακριά από τα δύο άκρα των UTRs. Σημαντική διαφορά παρατηρήθηκε μεταξύ των θέσεων στα downregulated γονίδια και στα φυσιολογικά γονίδια (Πίνακας 2).

Feature name	Fold change	*P*-value
Seed match conservation	2.47	7.4*E*−29
Terminal base match	1.98	3.2*E*−13
Seed 7a matching site	1.65	2.7*E*−03
Location >600 bases	0.52	3.1*E*−10
Location >900 bases	0.43	1.6*E*−07
Duplex hybridization ΔG	0.95	2.0*E*−05
Duplex hybridization index	1.15	6.7*E*−05
Features related to the target site		
ΔG	0.70	1.9*E*−31
GC content	0.80	3.7*E*−46
Accessibility index	0.82	2.4*E*−09
A count	1.22	5.5*E*−14

U count	1.19	4.0*E*−13
C count	0.75	2.4*E*−18
G count	0.82	3.3*E*−10
Dinucleotide counts		See Table 3
Position-specific base counts		See Table S1

Πίνακας 2 Χαρακτηριστικά mirTarget2

από τη δημοσίευση: Prediction of both conserved and nonconserved microRNA targets in animals, Xiaowei Wang and Issam M. El Naqa [Αναφ. 8].

Dinucleotide	Fold change	*P*-value
UU	1.55	1.8*E*−14
GC	0.63	4.1*E*−14
AA	1.53	7.4*E*−13
CC	0.63	3.8*E*−11
UA	1.42	1.2*E*−10
AU	1.34	1.3*E*−09
GG	0.64	8.5*E*−09
CA	0.82	2.5*E*−05
UC	0.81	2.9*E*−05
CG	0.59	5.0*E*−05
GA	1.20	1.1*E*−03
CU	0.88	2.4*E*−03

Πίνακας 3 Σπουδαιότητα της σύνθεσης των δινουκλεοτιδίων

από τη δημοσίευση: Prediction of both conserved and nonconserved microRNA targets in animals, Xiaowei Wang and Issam M. El Naqa [Αναφ. 8].

3.2 miRmap

3.2.1 Εισαγωγη

MiRmap, πρόκειται για μια βιβλιοθήκη λογισμικού ανοιχτού κώδικα (open source), η οποία για πρώτη φορά καλύπτει συνολικά τις τέσσερις προσεγγίσεις θερμοδυναμική, εξελικτική, πιθανοτική και ακολουθιακή χρησιμοποιώντας 11 χαρακτηριστικά πρόβλεψης, 3 από τα οποία είναι νέα. Αυτό επέτρεψε να εξετάσουνε συσχετισμούς χαρακτηριστικών και να συγκρίνουνε την προγνωστική τους δύναμη με έναν αμερόληπτο τρόπο χρησιμοποιώντας πειραματικά δεδομένα υψηλής απόδοσης από πειράματα ανοσοκαθαρισμού - immunopurification, μεταγραφικοποίησης - transcriptomics, πρωτεϊνωμάτωσης – proteomics και κλασματοποίησης πολυσώματος - polysome fractionation. Η προσβασιμότητα στην τοποθεσία στόχος φαίνεται να είναι το πιο προγνωστικό χαρακτηριστικό. Το νέο χαρακτηριστικό που βασίζεται στο PhyloP, το οποίο αξιολογεί τη σημασία της αρνητικής επιλογής, είναι ο καλύτερος προγνωστικός παράγοντας στην εξελικτική κατηγορία. Συνδυάσανε όλα τα χαρακτηριστικά σε ένα ολοκληρωμένο μοντέλο που σχεδόν διπλασιάζει την προβλεπτική ισχύ του TargetScan. Το miRmap διατίθεται δωρεάν από http://cegg.unige.ch/mirmap . Το miRmap δε βασίζεται σε τεχνολογία Machine Learning.

3.2.2 Συνοψη

1) Γενική περιγραφή δεδομένων εκπαίδευσης - εισόδου
2) Πηγή προέλευσης
3) Το πλήθος τους
4) 1,2,3 για τα δεδομένα αξιολόγησης
5) Χαρακτηριστικά
6) Μέθοδοι μηχανικής μάθησης
7) Περιοχή γονιδιώματος που γίνεται η πρόβλεψη
8) Τύπος πρόσδεσης των miRNAs - seeds
9) Dependencies

---

1) **Experimental data:**
    a) Microarrays από miRNAs επιμολυσμένων με κύτταρα HeLa από Grimson για τα miRNAs 122a, 128a, 132, 133a, 142, 148b, 181a, 7 και 9. Χρησιμοποιήσανε δεδομένα έκφρασης στις 24 ώρες μετά τη διαμόλυνση και επιλέχθηκαν μόνο ανιχνευτές με εντάσεις σήματος πάνω από το μέσο στα πειράματα επιμόλυνσης ελέγχου για να διατηρήσουν μόνο τα μετάγραφα που εκφράστηκαν αρκετά ώστε να παρατηρήσουν τη σιγή του miRNA.
    b) Microarrays από Linsley, τα πειράματα GSM156522, GSM156523, GSM156524, GSM156545, GSM156546, GSM156547, GSM156548, GSM156553, GSM156557, GSM156559, GSM156576, GSM156577, GSM156578, GSM156579 και GSM156581, τα οποία μετρήθηκαν στις 24 ώρες με τις ίδιες πειραματικές συνθήκες
    c) Proteomics overexpression από Selbach, συμπεριλαμβάνοντας fold-changes εκφράσεις που μετρήθηκαν στις 32 ώρες για miR-1, miR-155 και miR-16 και όχι τα let-7b και miR-30a καθώς αυτά τα miRNA ασκούν αρνητική επίδραση στο RNA silencing pathway
    d) δεδομένα HITS-CLIP από Chi, χρησιμοποίησαν το peak height ως μέτρο στόχευσης miRNA για τα 20 πιο άφθονα και διαθέσιμα miRNAs και κατέγραψαν τη συσχέτιση των κορυφών χρησιμοποιώντας μια βιολογική πολυπλοκότητα (BC, ένα μέτρο αναπαραγωγιμότητας μεταξύ βιολογικών αντιγράφων)
    e) από Hendrickson, επιμολύνθηκε το miR-124 σε κύτταρα HEK293T και μετρήθηκε (i) η σχέση miR-RISC με Ago IP, (ii) έκφραση μεταγράφου με microarrays και (iii) δραστικότητα μετάφρασης με κλασματοποίηση πολυσωμάτων. Χρησιμοποιήσαν το

σύνολο δεδομένων 5 από τις συμπληρωματικές πληροφορίες που περιλαμβάνει όλες τις μετρήσεις για κάθε μετάγραφο.

Sequence data:

RefSeq 47 χαρτογραφήθηκε σε ανθρώπινα (hg19) και μη (mm9) γονίδια (ποντικιών) και χρησιμοποιήθηκαν για τον ορισμό των σημείων mRNA, που περιορίστηκαν σε «NM_» μετάγραφα. Το miRBase 18 χρησιμοποιήθηκε για miRNA παρατηρήσεις & σχολιασμούς.

2) **Experimental data: (a)** λήφθηκαν από το GEO (GSE8501) **(b)** από την GEO (GSE6838) **(c)** αντίστοιχη ιστοσελίδα που αναφέρεται στο άρθρο του Selbach "Widespread changes in protein synthesis induced by microRNAs" (7) **(d)** απ' την αντίστοιχη ιστοσελίδα **(e)** απ' την αντίστοιχη ιστοσελίδα που αναφέρεται το άρθρο "Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA (25)

   Sequence data: από UCSC

3) –
4) Για την αξιολόγηση κάθε χαρακτηριστικού πήραν δεδομένα από 7 διαφορετικά ανεξάρτητα πειράματα:
   (i) Chi (9) διεξήγαγε ένα πείραμα διασταυρούμενης σύνδεσης Ago-RNA ακολουθούμενο από IP και προσδιορισμό αλληλουχίας από την οποία προσδιορίστηκαν οι θέσεις πρόσδεσης miRNA ***από*** IPcross.Chi, τύπου HITS-CLIP, Chi . (9)
   (ii) Hendrickson (25) διεξήγαγε ένα Ago-IP με μη διασταυρούμενη σύνδεση για να υπογραμμίσουν την επίδραση του σταδίου σταυρωτής σύνδεσης. Για τη μέτρηση της επίδρασης στα επίπεδα mRNA, χρησιμοποίησαν μελέτες που βασίζονται σε μεταγγίσεις miRNA ακολουθούμενες από μετρήσεις μικροσυστοιχίας ***από*** IP.Hendrickson, τύπου Immunopurification, Hendrickson . (25)
   (iii) Grimson (6), ***από*** Trans.Grimson, τύπου Microarray, Grimson . (6)
   (iv) Linsley (24) ***από*** Trans.Linsley, τύπου Microarray, Linsley . (24) και
   (v) Hendrickson (25) ***από*** Trans.Hendrickson, τύπου Microarray, Hendrickson . (25)
   (vi) Για να εκτιμηθεί η επίδραση του miRNA στην μετάφραση, επωφελήθηκαν από τα πειράματα κλασματοποίησης πολυσώματος από Hendrickson (25), ***από*** RibN.Hendrickson, τύπου Polysome fractionation, Hendrickson . (25)
   (vii) και από τα πειράματα πρωτεϊνωμάτωσης από Selbach (7) με βάση την τεχνολογία pSILAC για να αποκτήσει την τελική έξοδο μετάφρασης, ***από*** Prot.Selbach, τύπου pSILAC, Selbach . (7)

5) **Thermodynamic**:
   1. ΔG duplex,MFE with RNAcofold
   2. ΔG binding,Binding energy based on ensemble free energy
   3. ΔG open, mRNA opening free energy—Accessibility
   4. ΔG total, ΔG Duplex + ΔG open Other seed types

   Probabilistic:
   5. P.over binomial, site over-representation prob. (binomial dist.)
   6. P.over exact, site over-representation prob. (exact dist.)

   Conservation:
   7. BLS, branch length score on 3'-UTR fitted tree
   8. PhyloP, SPH test from PhyloP

   Sequence:

9. AU content, AU nucleotide composition around the seed
10. UTR position, Distance from the nearest 3'-UTR end
11. 3'-pairing, 3'-compensatory pairing

6) Ο αλγόριθμος αυτός δε βασίζεται στην τεχνολογία του Machine Learning
7) 3' UTR
8) 7-mer seed canonical περιοχές, αλληλουχίες μήκους 7 νουκλεοτιδίων και όχι 6, διότι η ισχυρότερη καταστολή του mRNA συνδέεται με τα μακρύτερα seeds. Επιτρέπεται μόνο κανονική αντιστοίχιση, χωρίς G:U ταλαντεύσεις και αναντιστοιχίες (non Watson Crick pairing).
9) miRmap web εφαρμογή: http://cegg.unige.ch/mirmap
   Ο ιστότοπος miRmap περιλαμβάνει υπηρεσίες τόσο για προϋπολογισμένες προβλέψεις στόχων miRNA για περιήγηση, όσο και για ηλεκτρονικό υπολογισμό στόχων miRNA σε ακολουθίες που υποβάλλονται από τον χρήστη.
   Το πρώτο βήμα είναι να επιλέξετε ένα είδος. Ο χρήστης μπορεί τότε να επιλέξει 'Select' ή να εισάγει την ακολουθία 'Sequence' ενός miRNA και / ή ενός γονιδίου κωδικοποίησης πρωτεΐνης. Είναι σημαντικό να σημειωθεί ότι για την περιήγηση των προυπολογισμένων προβλέψεων οι χρήστες μπορούν να επιλέξουν ένα miRNA ή ένα γονίδιο ή και τα δύο με δυνατότητα αυτόματης συμπλήρωσης για να διευκολύνουν τις γρήγορες επιλογές. Η επιλογή γονιδίου και οι επιλογές εισαγωγής αλληλουχίας μπορούν να συνδυαστούν έτσι ώστε ο χρήστης να μπορεί να εισάγει μια ακολουθία miRNA και να επιλέξει ένα γονίδιο/αντίγραφο με το όνομα ή το αντίστροφο. Όταν εισάγεται όνομα γονιδίου ή αναγνωριστικό, επιλέγεται το κανονικό αντίγραφο αυτού του γονιδίου. Είναι επίσης δυνατή η επιλογή ενός συγκεκριμένου μεταγράφου εισάγοντας απευθείας το αναγνωριστικό μεταγραφής. Όταν η φόρμα είναι έγκυρη, ο χρήστης μπορεί να υποβάλει το αίτημά του κάνοντας κλικ στο κουμπί "Get targets".
   Το miRmap περιλαμβάνει προϋπολογισμένες προβλέψεις για τα σχολιασμένα miRNA [miRBase 18] και τα γονίδια [Ensembl 83] οκτώ ειδών: Άνθρωπος, Χιμπατζής, Ποντίκι, Αρουραίος, Αγελάδα, Ψάρι-ζέβρα και Opossum.
   Υπάρχει και ένας πιο προγραμματιστικός τρόπος χρησιμοποιώντας την υπηρεσία REST. Χρησιμοποιήσανε το ευρέως χρησιμοποιούμενο εργαλείο ExtJS Javascript toolkit (sencha.com). Ένα λεπτομερές documentation της υπηρεσίας REST, καθώς και παραδείγματα, περιλαμβάνεται στο τμήμα βοήθειας του ιστότοπου miRmap http://mirmap.ezlab.org/help/rest.html .

3.2.3 Δομη Αλγοριθμου

Η βασική δομή του αλγορίθμου miRmap, βασίζεται σε 4 κύριες προσεγγίσεις, με βάση τις οποίες κατηγοριοποιούνται τα χαρακτηριστικά που λαμβάνει υπόψη ο αλγόριθμος:

1. Θερμοδυναμική - Thermodynamic
2. Πιθανοτική - Probabilistic
3. Εξελικτική - Evolutionary
4. Ακολουθιακή – Sequence

	Thermodynamic	Evolutionary	Probabilistic	Sequence-based	References
miRmap	✓	✓	✓	✓	
TargetScan		✓[a]		✓	Grimson *et al.* (6)
PITA	✓				Kertesz *et al.* (12)
PicTar	✓	✓	✓		Krek *et al.* (13)
miRanda	✓				John *et al.* (14)
RNAhybrid	✓				Rehmsmeier *et al.* (15)
DIANA-microT	✓	✓			Kiriakidou *et al.* (16)
ElMMo		✓	✓		Gaidatzis *et al.* (17)
PACMIT	✓		✓		Marín and Vanícek (18)

Πίνακας 4 Άλλες προσεγγίσεις που χρησιμοποιούνται από τα εργαλεία λογισμικού πρόβλεψης του miRNA.

από τη δημοσίευση: miRmap: Comprehensive prediction of microRNA target repression strength, Charles E. Vejnar and Evgeny M. Zdobnov [Αναφ. 19].

3.2.4 ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

(1) Thermodynamics of miRNA–mRNA interactions

1. ΔG duplex. Το ζεύγος miRNA-mRNA σχηματίζει ένα διπλό RNA. Χρησιμοποιώντας τη βιβλιοθήκη Δευτεροταγούς Δομής RNA της Βιέννης, υπολογίσανε την ελάχιστη ελεύθερη ενέργεια δίπλωσης (MFE - Minimum Free folding Energy) αυτού του διπλού RNA (με τη συνάρτηση 'cofold').

Χρησιμοποιήσανε τη συνάρτηση 'cofold' για τον υπολογισμό του ΔG duplex, καθώς αυτή η συνάρτηση της βιβλιοθήκης δευτερογενών δομών RNA της Βιέννης είναι πιο κατάλληλη από την τροποποιημένη λειτουργία 'duplexfold' που χρησιμοποιείται στο PITA για να υπολογίσει αυτό το χαρακτηριστικό. Η συνάρτηση 'duplexfold' γράφτηκε για γρήγορη σάρωση για πιθανές θέσεις υβριδισμού, ενώ η συνάρτηση 'cofold', σχεδιάστηκε ειδικά για να υπολογίσει την ελεύθερη ενέργεια του 2πλού RNA λαμβάνοντας υπόψη ενδομοριακά και δια μοριακά ζεύγη.

2. ΔG binding. Ενώ η δομή με τη χαμηλότερη προβλεπόμενη ενέργεια, MFE είναι η πιο σταθερή δομή, αρκετά σύνολα των RNA υιοθετούν διαφορετικές υπό-βέλτιστες δομές. Υπολογίσανε έτσι την ελεύθερη ενέργεια του συνόλου της δέσμευσης (με τη συνάρτηση 'co_pf_fold').

3. ΔG open. Υπολογίσανε την ενέργεια που απαιτείται για να ξεδιπλωθεί η περιοχή 3'-UTR της τοποθεσίας στόχου όπως και το PITA, με τη συνάρτηση 'pf_fold' της βιβλιοθήκης της Βιέννης. Ο υπολογισμός του 'ΔG open' απαιτεί δύο ενεργειακούς υπολογισμούς. Η ελεύθερη ενέργεια του mRNA που περιορίζεται στη διατήρηση της θέσης στόχου μονής έλικας αφαιρείται από την ελεύθερη ενέργεια του ίδιου μη περιορισμένου mRNA. Ο περιορισμός μονής έλικας τοποθετήθηκε σε ένα τμήμα 70 νουκλεοτιδίων επικεντρωμένο στη θέση στόχου.

4. ΔG total. Τέλος, το ΔG open αθροίζεται με 'ΔG duplex' ή 'ΔG binding' δίνοντας τη συνολική ενέργεια του συστήματος - αντίστοιχο ΔΔG στο PITA.

(2) Probability of the motif occurrence

Σχεδιάσανε την ακολουθία 3'-UTR ως διαδικασία Markov και προσδιορίστηκε η αναμενόμενη πιθανότητα να ανιχνευθούν τουλάχιστον n εμφανίσεις του μοτίβου που ορίστηκε ως ακριβής αντιστοίχιση seed ή την πλήρη θέση δέσμευσης miRNA, χρησιμοποιώντας δύο διαφορετικές μεθόδους. Στην πρώτη μέθοδο, η κατανομή πιθανότητας προσεγγίστηκε με μια διωνυμική κατανομή, όπως στο

Marin και Vanicek **5. P. over binomial,** ενώ στη δεύτερη μέθοδο υπολογίσανε την ακριβή κατανομή πιθανοτήτων **6. P. over exact** με βάση το θεωρητικό έργο Nuel (Αναφ. 31).

(3) Conservation of the target site

Χρησιμοποιώντας τις ευθυγραμμίσεις πολλαπλών γονιδιακών ακολουθιών UCSC (Αναφ. 27) MultiZ (hg19, MultiZ 46-way, mm9, MultiZ 30-way), αναζητήσανε διατηρούμενες θέσεις στόχου miRNA στα μπλοκ ευθυγράμμισης που καθορίστηκαν από τα 3' UTRs του είδους αναφοράς του ανθρώπου ή του ποντικιού για τα δεδομένα HITS-CLIP.

Από ένα δέντρο ειδών θηλαστικών (UCSC (Αναφ. 27)), καταρρίψανε πρώτα όλα τα είδη που δεν περιείχαν την τοποθεσία στόχο. Στη συνέχεια αθροίσανε τα μήκη των υπολοίπων κλάδων για να λάβουν τη βαθμολογία μήκους διακλάδωσης (**7. BLS** - Branch Length Score). Όπως εφαρμόζεται από τους Friedman (Αναφ. 19), αθροίσανε τα μήκη των κλάδων της τοπολογίας των ειδών για κάθε ευθυγράμμιση 3'-UTR με το μοντέλο REV χρησιμοποιώντας το πρόγραμμα PhyloFit από PHAST (Αναφ. 33). Οι χειρισμοί των δένδρων έγιναν με τη βιβλιοθήκη DendroPy.

Για να ελεγχθεί η ύπαρξη αρνητικών επιλογών που επηρεάζουν τις περιοχές στόχους miRNA, χρησιμοποιήσανε τις δοκιμασίες Siepel, Pollard και Haussler (SPH) που εφαρμόστηκαν στο πρόγραμμα **8. PhyloP**. Αυτή η δοκιμή αξιολογεί εάν τα μήκη των κλαδιών του δέντρου που είναι κατασκευασμένα από τις θέσεις στόχους είναι σημαντικά μικρότερα (λιγότερο αποκλίνουσες λόγω αρνητικής επιλογής) από το background (η 3' UTR).Οι αναφερόμενες τιμές στο κείμενο είναι το τεστ -log (P-value). Χρησιμοποιήθηκαν δεδομένα PhastCons 46-way από την UCSC (Αναφ. 27) για τον υπολογισμό της μέσης πιθανότητας αντιστοίχισης seeds ως ένα διατηρημένο στοιχείο. Τα αποτελέσματα των PhastCons κάθε βάσης στα seeds υπολογίστηκαν κατά μέσο όρο για να ληφθεί η βαθμολογία seed – seed score (Αναφ. 23,35).

(4) Sequence features

Εφαρμόσανε τα τρία χαρακτηριστικά αλληλουχίας του TargetScan (Αναφ. 6): **9. AU content.** Αναλογία νουκλεοτιδίων A και U πάνω από G και C, γύρω από την περιοχή seed, το χαρακτηριστικό **10. 3'-pairing** και **11. UTR position**. Μεταξύ της τοποθεσίας στόχου και του πλησιέστερου άκρου 3'-UTR.

Σύνολο 11 χαρακτηριστικά χρησιμοποιεί ο συγκεκριμένος αλγόριθμος.

Σχετική σημασία των χαρακτηριστικών

Επιπρόσθετα, υπολογίσανε τη σχετική σημασία των χαρακτηριστικών στα πολλαπλά γραμμικά μοντέλα με τη μέθοδο CAR (Αναφ. 36), η οποία αποσυνθέτει το ποσοστό της διακύμανσης που εξηγείται από κάθε μεταβλητή ενός μοντέλου λαμβάνοντας υπόψη τις συσχετίσεις μεταξύ των μεταβλητών.

3.3 TargetSpy

3.3.1 Εισαγωγη

Σχεδόν όλοι οι τρέχοντες αλγόριθμοι πρόβλεψης θέσης στόχου microRNA απαιτούν την παρουσία ενός (διατηρημένου) ταιριάσματος seed στο 5' άκρο του microRNA. Πρόσφατα όμως, έχει αποδειχθεί ότι η απαίτηση αυτή μπορεί να είναι υπερβολικά αυστηρή, πράγμα που οδηγεί σε σημαντικό αριθμό παραληφθέντων τοποθεσιών-στόχων. **TargetSpy**, μια νέα υπολογιστική προσέγγιση για την πρόβλεψη

τοποθεσιών στόχων miRNA ανεξάρτητα από την παρουσία ενός ταιριάσματος seed. Βασίζεται στη μηχανική μάθηση και την αυτόματη επιλογή χαρακτηριστικών χρησιμοποιώντας ένα ευρύ φάσμα χαρακτηριστικών σύνθεσης, δομών και ζευγών βάσεων που καλύπτουν τις τρέχουσες βιολογικές γνώσεις. Το μοντέλο δεν βασίζεται στην εξελικτική διατήρηση, που επιτρέπει την ανίχνευση ειδικών – μερικών αλληλεπιδράσεων για το είδος. Το targetSpy θεωρείται κατάλληλο για την ανάλυση μη διατηρημένων γονιδιακών ακολουθιών. Προκειμένου να καταστεί δυνατή η αμερόληπτη σύγκριση του targetSpy με άλλες μεθόδους, ταξινομήσανε όλους τους αλγόριθμους σε τρεις ομάδες:

I) μη απαίτηση αντιστοίχισης seed
II) απαίτηση αντιστοίχισης seed και
III) απαίτηση διατηρημένης αντιστοίχισης seed.

Σε μια ομάδα δεδομένων για τον άνθρωπο που αποκαλύπτει fold change στην παραγωγή πρωτεϊνών για πέντε επιλεγμένα microRNA, η μέθοδος δείχνει καλύτερη απόδοση σε όλες τις κατηγορίες. Στην Drosophila melanogaster, οι προβλέψεις της κατηγορίας II και III είναι ισοδύναμες με άλλους αλγορίθμους και οι προβλέψεις κατηγορίας I είναι απλώς οριακά λιγότερο ακριβείς. Εκτιμάται ότι το TargetSpy προβλέπει μεταξύ 26 και 112 λειτουργικών τοποθεσιών στόχων χωρίς αντιστοιχία seeds ανά microRNA που λείπουν από όλους τους άλλους διαθέσιμους επί του παρόντος αλγορίθμους.

Μόνο μερικοί αλγόριθμοι μπορούν να προβλέψουν τις θέσεις-στόχους χωρίς να απαιτήσουν αντιστοίχιση seed και το targetSpy επιδεικνύει σημαντική βελτίωση στην ακρίβεια της πρόβλεψης σε αυτή την κατηγορία. Επιπλέον, όταν απαιτείται διατήρηση και παρουσία ενός ταιριάσματος seed, η απόδοση είναι συγκρίσιμη με τους πλέον σύγχρονους αλγορίθμους. Το targetSpy εκπαιδεύτηκε σε ποντίκι και αποδίδει καλά στον άνθρωπο και τη drosophila, υποδηλώνοντας ότι μπορεί να εφαρμοστεί σε ένα ευρύ φάσμα ειδών. Επιπλέον, έχουν αποδείξει ότι η εφαρμογή τεχνικών μηχανικής μάθησης σε συνδυασμό με τα επερχόμενα δεδομένα αλληλουχίας καταλήγει σε ένα ισχυρό εργαλείο πρόβλεψης ιστοτόπου microRNA http://www.targetspy.org

3.3.2 ΣΥΝΟΨΗ

1) Γενική περιγραφή δεδομένων εκπαίδευσης - εισόδου
2) Πηγή προέλευσης
3) Το πλήθος τους
4) 1,2,3 για τα δεδομένα αξιολόγησης
5) Χαρακτηριστικά
6) Μέθοδοι μηχανικής μάθησης
7) Περιοχή γονιδιώματος που γίνεται η πρόβλεψη
8) Τύπος πρόσδεσης των miRNAs - seeds
9) Dependencies

---

1) i) 3' UTR αλληλουχίες, human (hg18, March 2006), mouse (mm8, July 2007), rat (rn4, November 2004) και chicken (galGal2, May 2006) χρησιμοποιώντας το RefSeq Genes Track, για fly (dme, April 2006) χρησιμοποίησαν παρατηρήσεις από το FlyBase.
   Για τη δημιουργία προβλέψεων θέσεων στόχων σχετικά με διατηρήσιμες περιοχές seeds χρησιμοποίησαν το Galaxy (Αναφ. 28) για να εξάγουν 3'UTR alignments για άνθρωπο, χιμπατζή, ποντίκι, αρουραίο και σκύλο από 17-way ολόκληρο ανθρώπινο γονιδίωμα και D. melanogaster, D. yakuba, D. ananassae, και D. pseudoobscura από το 15-way D. melanogaster ολόκληρο γονιδίωμα.

ii) microRNA αλληλουχίες
iii) σετ Ago - mRNA (τα 20 πιο συχνα microRNAs που παρουσιάζονται στο P13 του εγκεφάλου του ποντικιού). Αφαιρέθηκαν οι περιοχές που δεν βρίσκονταν σε 3'UTR ή που δεν είχαν αριθμό πρόσβασης RefSeq συνδεδεμένο.

2) i) UCSC Genome Database [Αναφ. 27], χρησιμοποιώντας το UCSC Πίνακας Browser
ii) miRBase, release 12
iii) http://ago.rockefeller.edu/
3) 692 microRNAs για άνθρωπο, 513 για ποντίκι, 443 για κοτόπουλο and 147 για μύγα.
4) **Προηγούμενες επαληθευμένες μέθοδοι**:
i) PicTar. Κατέβηκαν προβλέψεις στόχων από τον αλγόριθμο PicTar [Αναφ. 22] από το hg17 στο hg18 'μετανάστευσαν' εφαρμόζοντας την εντολή liftover. Χρησιμοποίησαν τις προβλέψεις διατηρήσιμες σε άνθρωπο, ποντίκι, αρουραίο, χιμπατζή και σκύλο (4-way), όπως και για το κοτόπουλο (5-way). Για τη μύγα κατέβασαν το σύνολο S1 από το PicTar το οποίο απαρτίζεται από προβλέψεις συντηρήσιμες σε D. melanogaster, D. yakuba, D. ananassae. and D. Pseudoobscura
ii) miRanda. Μόνο προβλέψεις για μετάγραφα που περιλαμβάνονται στη βάση δεδομένων RefSeq εξετάστηκαν. Ανθρώπων και της μύγας προβλέψεις από miRBase Targets[Αναφ. 7], version 5, λήφθηκαν από http://microrna.sanger.ac.uk/targets/v5/ .
iii) Προβλέψεις RNA22 [Αναφ. 23] για ανθρώπινες αλληλουχίες 3'UTR. Μέχρι να γίνουν οι προβλέψεις με τη χρήση Ensembl transcripts, χαρτογραφήσανε τις προβλέψεις στα γονίδια RefSeq εφαρμόζοντας πίνακες αντιστοίχισης που παρέχει η Ensembl και η UCSC.
iv) PITA. Χρησιμοποίησαν τα TOP και ALL σετς με 3/15 flankings.
v) Οι TargetScanS [Αναφ. 13] προβλέψεις και ο αντίστοιχος microRNA family mapping πίνακας χρησιμοποιηθήκανε.
vi) Οι προβλέψεις Gaidatzis. [20] λήφθηκαν για τον αλγόριθμο EIMMo.
vii) Στόχοι που προβλέπονται από το mirTarget2 (Έκδοση 3) [Αναφ. 21]
viii) DIANAmicroTv3.0. Για κατώτα όρια - χαλαρά (score 7.3) και αυστηρά (score 19)
ix) TargetRank. Στόχους για τον άνθρωπο.
Πειραματικά:
Χρησιμοποιήθηκαν δύο ομάδες πειραματικά επαληθευμένων θέσεων στόχου για την σύγκριση των αλγορίθμων πρόβλεψης στόχων. Για αξιολόγηση σε Drosophila melanogaster, χρησιμοποιήσανε τις 120 πειραματικά δοκιμασμένες αλληλεπιδράσεις microRNA - γονιδίου που συντάχθηκαν από Stark [Αναφ. 17] και τις 190 αλληλεπιδράσεις που δημοσιεύονται από Kertesz [Αναφ. 9]. Οι κατάλληλες ακολουθίες 3'UTR προήλθαν από τις παρατηρήσεις FlyBase που παρείχε η UCSC. Οι μεταγραφές για τις οποίες δεν ήταν διαθέσιμη η 3 'UTR περιοχή απορρίφθηκαν. Για την αξιολόγηση στον άνθρωπο, χρησιμοποιήσανε ένα σετ που βασίζεται στην τεχνική pSILAC και αποκαλύπτει τις fold changes της παραγωγής πρωτεϊνών που προκαλούνται από πέντε επιλεγμένα microRNA [Αναφ. 18], τα οποία λήφθηκαν από τη διεύθυνση http://psilac.mdcberlin.de .
Origin:
i) UCSC database, χρησιμοποιώντας το Πίνακας Browser.
ii) http://microRNA.org [29]
iii) http://cbcsrv.watson.ibm.com/rna22.html
iv)http://genie.weizmann.ac.il/pubs/mir07/mir07_data.html
v)http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_50
vi) http://www.mirz.unibas.ch/
vii)http://mirdb.org/miRDB

viii) http://diana.cslab.ece.ntua.gr/microT/
ix)http://hollywood.mit.edu/targetrank/

Size:
εκτεταμένο σύνολο 190 πειραματικά επαληθευμένων αλληλεπιδράσεων microRNA-στόχου

Positive & Negative:
Το παλιό σύνολο αποτελείται από 61 λειτουργικές(positive) και 59 μη λειτουργικές (negative) αλληλεπιδράσεις. Το τελευταίο σύνολο αποτελείται από 102 λειτουργικές και 88 μη λειτουργικές αλληλεπιδράσεις.

5) Compactness
Αναλογία περιεχομένου G+C σε microRNA και θέση στόχο
Μεγαλύτερο μήκος διαδοχικών ζεύξεων βάσεων οπουδήποτε στο υβρίδιο
Ασύμμετρη δέσμευση
Αναλογία περιεχομένου G+C της θέσης στόχου
Αριθμός ζευγών βάσεων του miRNA, 8mer seed
Η τοποθεσία του στόχου στο 3' UTR
6) Ο αλγόριθμος αυτός βασίζεται στην τεχνολογία του Machine Learning, πρόκειται για μία supervised machine learning προσέγγιση. Πραγματοποιήθηκε με τη βοήθεια του data mining λογισμικού WEKA(version 3.5.3).
7) 3' UTR
8) Δεν απαιτείται τέλεια αντιστοίχιση, δεν εξαρτάται από seeds
9) targetSpy εφαρμογή: http://www.targetspy.org
Ένα stand-alone java πρόγραμμα υλοποιημένο σε java virtual machine version 1.5 με το πακέτο Vienna για RNA secondary structure prediction(version 1.6.1)

3.3.3 Δομη Αλγοριθμου

Ταξινόμηση προσεγγίσεων πρόβλεψης

Τα τρέχοντα εργαλεία για την πρόβλεψη των θέσεων στόχου microRNA μπορούν να ομαδοποιηθούν σε τρεις ξεχωριστές κατηγορίες (Πίνακας 5). Ο βασικός λόγος είναι για να υπάρξει στη συνέχεια μια δίκαιη σύγκριση με τους άλλους αλγορίθμους που λαμβάνουν υπόψη τη διατηρησιμότητα & την αντιστοίχιση των seeds.

Η κατηγορία I αποτελείται από εκείνες τις προσεγγίσεις που δεν χρησιμοποιούν ούτε την απαίτηση αντιστοίχισης seeds ούτε τη συντήρηση. Η κατηγορία II περιέχει όλες τις προσεγγίσεις που απαιτούν ταίριασμα seeds, αλλά δεν κάνουν χρήση της διατήρησης. Τέλος, η τάξη III είναι για τις προβλέψεις όπου και οι δύο απαιτούν ένα ταίριασμα seeds και βασίζονται στη διατήρηση.

	Απαίτηση seeds	**Απαίτηση διατηρησιμότητας**
Κατηγορία I	Όχι	Όχι
Κατηγορία II	Ναι	Όχι
Κατηγορία III	Ναι	Ναι

Πίνακας 5 Ταξινόμηση των διαφόρων αλγορίθμων

Organism	Seed match not required	Seed match required	Seed match required and conservation considered
Human	RNA22 [23] TargetSpy no-seed	PITA All 3/15 [9] TargetScanS non-conserved TargetSpy seed	EIMMo [21] MiRBase Targets [24] MiRanda [6] PicTar [22] DIANA-microT [30] TargetScanS [13] PITA TOP [9] MirTarget2 [21] TargetRank [26] TargetSpy cons. seed
Fly	RNA22 [23] TargetSpy no-seed	PITA All 3/15 [9] TargetSpy seed	EIMMo [21] PicTar [22] MiRBase Targets [24] TargetSpy cons. seed TargetScanS [13]

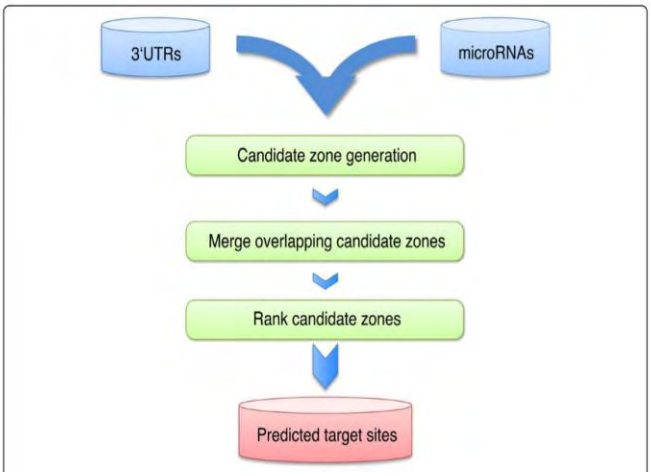
Εικόνα 4 Κατηγοριοποίηση των προγραμμάτων

από τη δημοσίευση: TargetSpy: a supervised machine learning approach for microRNA target prediction, Martin Sturm, Michael Hackenberg, David Langenberger, Dmitrij Frishman [Αναφ. 28]

Ορισμένες μέθοδοι δεν μπορούν να προσαρμοστούν τέλεια σε αυτό το σχήμα. Για παράδειγμα, ενώ ο αλγόριθμος miRanda [Αναφ. 6] δεν απαιτεί τέλεια αντιστοίχιση seed, το βάρος της περιοχής των seeds είναι τόσο υψηλό που κατά μέσο όρο περίπου το 7% όλων των προβλεπόμενων θέσεων στόχων εμφανίζουν αναντιστοιχίες σε 7mer seeds (νουκλεοτίδια microRNA 1-7 ή 2 -8). Ο αλγόριθμος miRBase [Αναφ. 7] χρησιμοποιεί το miRanda και επιτρέπει μια ενιαία αναντιστοιχία στην περιοχή των seeds. Καθώς και οι δύο προσεγγίσεις απαιτούν επιπρόσθετα τη συντήρηση του τόπου στόχου, θεωρούνται ως μέλη της κλάσης ΙΙΙ. Παρ' όλα αυτά, το TargetSpy ανήκει γενικά στην κλάση Ι, αφού το μοντέλο δεν επιβάλλει αυστηρή απαίτηση αντιστοίχισης seed και δεν βασίζεται στη διατήρηση των τοποθεσιών στόχων.

Η βασική διαδικασία του αλγορίθμου

Η βασική δομή του αλγορίθμου targetSpy, βασίζεται στο pipeline που φαίνεται στην Εικόνα 5.



Εικόνα 5 Η δομή του targetSpy

από τη δημοσίευση: TargetSpy: a supervised machine learning approach for microRNA target prediction, Martin Sturm, Michael Hackenberg, David Langenberger, Dmitrij Frishman [Αναφ. 28]

Συνοπτικά ισχύει,

- Οι σημειωμένες αλληλουχίες 3 'UTR και όλα τα γνωστά microRNA από ένα δεδομένο είδος χρησιμεύουν ως είσοδος.
- Τα MicroRNAs ταιριάζουν με τα 3'UTRs για να δημιουργήσουν πιθανές υποψήφιες ζώνες.
- Οι προκύπτουσες υποψήφιες ζώνες ταξινομούνται και ταξινομούνται σύμφωνα με τη βαθμολογία τους,
- ενώ οι επικαλυπτόμενες ζώνες συγχωνεύονται.

Στόχος εδώ είναι να χτίσουν έναν αγωγό για την πρόβλεψη τοποθεσιών στόχου microRNA με βάση πολλαπλά χαρακτηριστικά που περιγράφονται παρακάτω στην ενότητα ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ. Σημειώνεται ότι η προσέγγιση αυτή είναι απαλλαγμένη από την απαίτηση της αντιστοίχισης seed και της διατήρησης αυτής της περιοχής σε πολλά είδη.

A. Το TargetSpy παίρνει ως είσοδο δύο πολλαπλά αρχεία FASTA. Ένα με τις αλληλουχίες 3 'UTR και ένα με τις ώριμες ακολουθίες microRNA. Σημειώστε ότι δεν πρέπει να παρέχονται άλλες εξωγενείς πληροφορίες, όπως η εξελικτική διατήρηση.
B. Για κάθε είσοδο microRNA, το TargetSpy αναγνωρίζει τις υποψήφιες ζώνες (τμήματα αλληλουχίας RNA που ενδεχομένως φιλοξενούν μια θέση στόχου) σε όλες τις αλληλουχίες 3'UTR.
C. Υπολογίζει τη βαθμολογία για την εκάστοτε υποψήφια ζώνη,
D. συγχωνεύει επικαλυπτόμενες υποψήφιες ζώνες και
E. κατατάσσει τις προβλέψεις σύμφωνα με τις βαθμολογίες τους.

Χρησιμοποιώντας αυτό το μοντέλο βρέθηκαν θέσεις στόχοι για Homo sapiens, Mus musculus, Rattus norvegicus, Gallus gallus και Drosophila melanogaster (Πίνακας 6).

				Number of predicted target sites			
	Number of 3'UTRs	Average 3'UTR length	Number of microRNAs	TargetSpy no-seed sens	TargetSpy no-seed spec	TargetSpy seed sens	TargetSpy seed spec
Human	26161	1210	692	4837 k	1023 k	829 k	339 k
Mouse	18694	1082	513	1906 k	407 k	340 k	137 k
Rat	11859	760	292	535 k	113 k	91 k	36 k
Chicken	3676	927	443	372 k	80 k	59 k	24 k
Fly	15884	471	147	247 k	54 k	50 k	20 k

Πίνακας 6 Προβλεπόμενοι στόχοι για κάθε είδος από διάφορες εκδοχές του targetSpy

από τη δημοσίευση: TargetSpy: a supervised machine learning approach for microRNA target prediction, Martin Sturm, Michael Hackenberg, David Langenberger, Dmitrij Frishman [Αναφ. 28]

Υποψήφιες θέσεις στόχοι

Ένα συνηθισμένο σημείο εκκίνησης μιας ροής εργασιών πρόβλεψης είναι η αναζήτηση τέλειων αντιστοιχιών seeds στο 3'UTR των ενδιαφερόμενων μεταγραφών. Δεδομένου ότι ο στόχος είναι να αναπτυχθεί ένα μοντέλο που δεν βασίζεται στην παρουσία ενός ταιριάσματος seeds, έπρεπε να επαναπροσδιοριστούν οι κανόνες για την επιλογή των αρχικών υποψήφιων θέσεων-στόχων.

Σύμφωνα με τη συλλογιστική ότι ένας λειτουργικός χώρος προσελκύεται περισσότερο από το σύμπλεγμα RISC σε σχέση με την περιβάλλουσα περιοχή, αναζητήθηκαν περιοχές όπου η προβλεπόμενη ελεύθερη ενέργεια Gibbs του διπλότυπου στόχου microRNA είναι κάτω από ένα συγκεκριμένο για το microRNA κατώτατο όριο ενέργειας. Για να διασφαλιστεί υψηλή κάλυψη των λειτουργικών σημείων πρόσδεσης, επιλέχθηκε μια συντηρητική αποκοπή και εντοπίσανε περίπου 150 εκατομμύρια υποψήφιους στόχους για όλα τα microRNA στον άνθρωπο.

Training set

Προκειμένου να αναπτυχθεί ένας ταξινομητής υψηλής ποιότητας, είναι απαραίτητο να δημιουργηθεί ένα σετ εκπαίδευσης που να είναι πραγματικά αντιπροσωπευτικό τόσο για τις θετικές (πραγματικές τοποθεσίες στόχων) όσο και για τις αρνητικές (όχι θέσεις στόχους). Δημοσιεύθηκε ένα σύνολο θέσεων δέσμευσης mRNA αργοναύτη (Ago), ταυτοποιημένο με μια νέα τεχνική που απομονώνει RNA με διασταυρούμενη immunoprecipitation σε πειράματα υψηλής απόδοσης (HITS-CLIP), για τα 20 πλέον άφθονα microRNA που υπάρχουν στον εγκέφαλο ποντικού P13 [Αναφ. 19].

Πρόκειται για το πρώτο σύνολο πειραματικών δεδομένων που αναφέρει απευθείας τις θέσεις στόχους microRNA σε μεγάλη κλίμακα και ο TargetSpy είναι ο πρώτος αλγόριθμος που το χρησιμοποιεί για την εκπαίδευση. Ανέκτησαν τα δεδομένα από http://ago.Rockefeller.edu/

και αφαιρέθηκαν όλες οι τοποθεσίες που δεν χαρτογραφήθηκαν σε 3'UTR ή δεν είχαν κανένα αριθμό πρόσβασης RefSeq.

Δεδομένου ότι μόνο η οικογένεια microRNA προσδιορίζεται σε αυτή τη δημοσίευση, εντοπίσανε τους υποψήφιους στόχους για όλα τα microRNA που ανήκουν στην οικογένεια αυτή.

Θετικές & Αρνητικές περιπτώσεις

Αυτοί οι υποψήφιοι στόχοι που επικαλύπτονταν τις πειραματικά προερχόμενες θέσεις διατηρήθηκαν ως θετικές περιπτώσεις. Σε περιπτώσεις όπου αρκετοί υποψήφιοι της ίδιας οικογένειας microRNA αλληλοεπικαλύπτονταν, επέλεγαν τον πιο ενεργητικό. Οι υποψήφιοι στοχευόμενων θέσεων που δεν έχουν κανένα ισοδύναμο στο σύνολο πειραματικών προερχόμενων θέσεων δέσμευσης Ago είναι απίθανο να είναι βιολογικώς σχετικοί. Επομένως, προσδιορίσανε τον ενεργητικό πιο σταθερό υποψήφιο για μια αναφερθείσα αλληλεπίδραση Ago-mRNA που δεν επικαλύπτει την επικυρωμένη θέση σύνδεσης Ago. Αυτοί οι υποψήφιοι χρησίμευσαν ως αρνητικές περιπτώσεις. Συνολικά λάβανε 3.872 θετικές και 4.540 αρνητικές περιπτώσεις.

3.3.4 ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Οι αργοναύτες είναι πρωτεΐνες οι οποίες κατά την συσχέτιση με τα microRNAs σχηματίζουν το σύμπλεγμα σίγασης που προκαλείται από RNA (RISC), το οποίο είναι υπεύθυνο για την καταστολή της έκφρασης του mRNA στόχου.

Αξιολογήσανε ένα ευρύ φάσμα χαρακτηριστικών των θέσεων - στόχου, εφαρμόζοντας την τεχνική ReliefF [Αναφ. 8] (Πίνακας 7). Ορισμένα χαρακτηριστικά που γενικά θεωρούνται ότι έχουν μεγάλη σημασία για την αναγνώριση της θέσης-στόχου από το microRNA, όπως ο αριθμός των βασικών γονιδίων στην περιοχή seed, έχουν εκτελεστεί πολύ καλά. Από την άλλη πλευρά, η προσβασιμότητα των χαρακτηριστικών με 3 upstream νουκλεοτίδια και 15 downstream πλευρικώς, που αναφέρθηκε στο [9], εκτιμήθηκε ως

ανεπαρκής απόδοση. Για να αναλύσουνε αν αυτό οφείλεται στις επιλεγμένες πλευρικές αλληλουχίες, εξετάσανε άλλες πλευρικές αλληλουχίες και βρήκαν 30 upstream νουκλεοτίδια και 30 downstream για να υπάρξει ελαφρώς καλύτερη απόδοση από τη ρύθμιση 3/15 nt (τα δεδομένα δεν παρουσιάζονται).

Είναι ενδιαφέρον το γεγονός ότι το χαρακτηριστικό compactness (που συνδυάζει το μήκος της θέσης στόχου και τον αριθμό των νουκλεοτιδίων που δεσμεύονται στο microRNA), παρουσιάζει μια από τις καλύτερες αποδόσεις.

Rank	Features	Score
1	Number of base parings to the microRNA 8-mer seed	0.03175
2	G+C content of target site	0.01263
3	Number of base pairings to the first 8 nucleotides of the microRNA 3' end	0.01038
4	Number of consecutive base-pairings to the microRNA 3' end with two allowed non-pairing positions	0.00995
5	Occurrence of CpG in target site	0.00799
6	G+C content ratio between the microRNA and the target site	0.00642
7	Compactness	0.00619
8	T9 anchor	0.00556
9	Longest stretch of consecutive base-pairings in the hybrid	0.00513
10	Number of bulges in the microRNA of size three	0.00498
11	T1 S/W anchor	0.00491
12	Total number of base-pairings	0.00475
13	Number of bulges on the target site of size seven or greater	0.00442
14	T1 anchor	0.00434
15	Number of bulges in the microRNA of size two	0.00433
16	Occurrence of CpG in the upstream flanking area	0.00383
17	Number of bulges in the target site of size one	0.00374
18	Total bulge length of the target site	0.00362
19	Length of the target site	0.00336
20	Total bulge length of the microRNA	0.00334
21	Target site position within the 3'UTR	0.00333
22	Number of symmetric bulges	0.00290
23	G + C content upstream of the target site	0.00287
24	Number of bulges on the target site	0.00286
25	Length of the second largest bulge on the target site	0.00268
26	Mean length of bulges on the target site	0.00263
27	T9 S/W anchor	0.00261
28	Binding asymmetry	0.00255
29	Number of bulges in the target site of size two	0.00240
30	Total number of G:U wobble base pairs	0.00227
31	Local RISC accessibility 30/30	0.00220
32	Local RISC accessibility 3/15	0.00215
33	Number of bulges in the target site of size four	0.00210
34	Difference in G+C content between the first and the last nt of the target site	0.00201
35	Occurrence of CpG in downstream flanking area	0.00179

36	Number of bulges in the microRNA of size one	0.00179
37	Length of the second largest bulge on the microRNA	0.00174
38	Number of bulges on the microRNA	0.00153
39	Number of bulges in the microRNA of size five	0.00128
40	Number of bulges in the target site of size three	0.00113
41	Difference in G + C content between the target site and the 20 nt upstream and downstream flanking region	0.00112
42	Number of bulges in the target site of size five	0.00100
43	Number of bulges in the microRNA of size four	0.00084
44	G+C content downstream of the target site	0.00084
45	Number of bulges in the target site of size six	0.00021

Πίνακας 7 Ταξινομημένη λίστα χαρακτηριστικών targetSpy

από τη δημοσίευση: TargetSpy: a supervised machine learning approach for microRNA target prediction, Martin Sturm, Michael Hackenberg, David Langenberger, Dmitrij Frishman [Αναφ. 28]

Αξιολόγηση απόδοσης ταξινομητή

Στη συνέχεια αξιολογήσανε την απόδοση του ταξινομητή σε σχέση με τα χαρακτηριστικά που χρησιμοποιούνται για την εκπαίδευση. Ξεκινήσανε με την καλύτερη δυνατή λειτουργία και προσθέσανε σταδιακά χαρακτηριστικά από τον Πίνακα 7 μία προς μία, σύμφωνα με την κατάταξη. Κάθε ταξινομητής αξιολογήθηκε στην εκπαιδευτική ομάδα με μια τυπική 10-fold cross validation διαδικασία, όπως εφαρμόζεται στο WEKA [10]. Η Εικόνα 6 δείχνει τον αριθμό των χαρακτηριστικών που χρησιμοποιούνται για την κατασκευή του ταξινομητή και της αντίστοιχης περιοχής κάτω από τις τιμές καμπύλης (AUC). Όπως φαίνεται, η τιμή AUC αυξάνεται μέχρι το 14ο προστιθέμενο χαρακτηριστικό. Με την προσθήκη άλλων χαρακτηριστικών η απόδοση αρχίζει να κυμαίνεται γύρω από την τιμή AUC 0,79 και δεν βελτιώνεται περαιτέρω. Δεδομένου ότι η ReliefF λαμβάνει υπόψη μόνο ένα χαρακτηριστικό κάθε φορά και δεν λαμβάνει υπόψη τη συσχέτιση μεταξύ των χαρακτηριστικών, εφαρμόσανε επιπρόσθετα τη συνάρτηση επιλογής βασισμένη στη συσχέτιση CFS (Correlation-based Feature Selection) [11] για τον προσδιορισμό του καλύτερου υποσυνόλου χαρακτηριστικών. Αυτή η προσέγγιση επέστρεψε μια τιμή AUC 0,79 (Εικόνα 6, κόκκινη γραμμή) με ένα σύνολο επτά χαρακτηριστικών, δηλαδή

1. compactness
2. G+C content ratio between microRNA and target site
3. length of the longest stretch of consecutive base-pairings anywhere in the hybrid
4. binding asymmetry
5. G+C content of the target site
6. number of base-pairings to the microRNA 8-mer seed και
7. the position of the target site in the 3'UTR.

Τα πρώτα τέσσερα χαρακτηριστικά χρησιμοποιούνται εδώ για πρώτη φορά, ενώ τα τελευταία τρία έχουν προταθεί ξανά από τις [Αναφ. 12-15].

Είναι ενδιαφέρον ότι πολλά από αυτά τα χαρακτηριστικά που αξιολογούνται ξεχωριστά, κατατάσσονται αρκετά χαμηλά, ενώ σε συνδυασμό ο ταξινομητής εκμεταλλεύεται τις συνεργατικές επιδράσεις μεταξύ των χαρακτηριστικών, καθιστώντας σε ένα σύνολο από πιο αδύναμα χαρακτηριστικά μια αρκετά καλή & συγκρίσιμη απόδοση.

Επιλέξανε το σύνολο χαρακτηριστικών που δημιουργήθηκε από την CFS για την τεχνική μηχανικής μάθησης.

Compactness:

Υποστήριξαν ότι τα υβρίδια που είναι πιο συμπαγή, δηλ. έχουν μόνο λίγα μη ζευγαρωμένα νουκλεοτίδια τόσο στο microRNA όσο και στη θέση στόχο είναι πιθανότερο να είναι βιολογικά λειτουργικά από άλλα. Ως εκ τούτου ο σκοπός εδώ ήταν να ενοποιήσουν αυτά τα δύο χαρακτηριστικά σε ένα μόνο μέτρο. Ορίζοντας το compactness ενός υβριδίου ως τη μέση τιμή των ακόλουθων αναλογιών: αριθμός ζευγών βάσεων / μήκος microRNA και αριθμός ζευγών βάσεων / μήκους θέσης στόχου. Συνεπώς, οι τιμές compactness είναι μεταξύ 0 και 1, με την τελευταία τιμή να αντιστοιχεί στην τέλεια συμπληρωματικότητα. Εάν ο τόπος στόχος είναι μικρότερος από το microRNA, εισάγεται ποινή, καθώς η περίπτωση αυτή δεν λαμβάνεται υπόψη με το μέσο όρο των λόγων που αναφέρθηκαν παραπάνω:

Τα κατώφλια αναγνώρισης ορίστηκαν κατά τρόπο ώστε οι θέσεις-στόχοι με ψευδώς θετικό ρυθμό μικρότερο από 5% (όπως εκτιμήθηκε σε 10-fold διασταυρούμενη επικύρωση) να αντιστοιχούν στο ευαίσθητο υποσύνολο και εκείνοι με ψευδώς θετικό ρυθμό 1 % ή λιγότερο στο σύνολο εξειδίκευσης (Πίνακας 8).

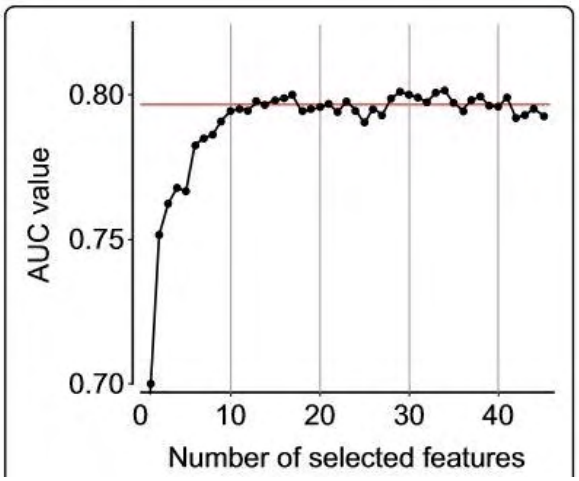
Σύνολο ευαισθησίας: False-positive < 5%

Σύνολο εξειδίκευσης: False-

$$\text{comp} = \frac{1}{2}\left(\frac{\#\text{pair}}{\text{miRLen}} + \frac{\#\text{pair}}{\text{tsLen}} \times f\left(\frac{\text{tsLen}}{\text{miLen}}\right)\right)$$

$$f(x) = \begin{cases} \sqrt{\cos\left(\pi \times \frac{(1-x)}{2}\right)} & x < 1 \\ 1 & else \end{cases},$$

Στην εξίσωση πάνω για το compactness, **#pair** είναι ο αριθμός ζευγών βάσεων, **tsLen** το μήκος της θέσης στόχου και **miRLen** το μήκος του microRNA.



Εικόνα 6 Αξιολόγηση ταξινομητή

από τη δημοσίευση: TargetSpy: a supervised machine learning approach for microRNA target prediction, Martin Sturm, Michael Hackenberg, David Langenberger, Dmitrij Frishman [Αναφ. 28]

Ο ταξινομητής αξιολογήθηκε σε μια επαναληπτική διαδικασία όπου ένα χαρακτηριστικό προστέθηκε κάθε φορά. Τα χαρακτηριστικά επιλέχθηκαν σύμφωνα με τον κατάλογο χαρακτηριστικών (Πίνακας 7), ξεκινώντας με την καλύτερη δυνατότητα. Στην εικόνα 6, σε μαύρο εμφανίζονται οι τιμές AUC (άξονας y) για το αντίστοιχο μέγεθος συνόλου χαρακτηριστικών (άξονας x). Η κόκκινη γραμμή υποδεικνύει την τιμή AUC του συνόλου χαρακτηριστικών που επιτεύχθηκε με την προσέγγιση επιλογής υποσύνολου χαρακτηριστικών.

Από τη στιγμή που το targetSpy προσπαθεί στο πρώτο βήμα να εντοπίσει όσο το δυνατόν περισσότερους δυνητικούς στόχους και στη συνέχεια ταξινομεί αυτούς σύμφωνα με το βαθμό ταξινόμησης, παράγονται τεράστιες ποσότητες τοποθεσιών στόχου από τις οποίες μόνο ένα κομμάτι, οι κορυφαίες προβλέψεις, παρουσιάζουν ενδιαφέρον. Για να καταστεί η εφαρμογή και η συγκριτική αξιολόγηση του targetSpy πιο διαφανή, δημιουργήσανε ένα υποσύνολο προβλέψεων με υψηλή ευαισθησία - sensitivity και υψηλή εξειδίκευση - specificity.

Prediction dataset name	Seed match required	Conservation considered	False-positive rate threshold
TargetSpy no-seed sens	No	No	0.05
TargetSpy no-seed spec	No	No	0.01
TargetSpy seed sens	Yes	No	0.05
TargetSpy seed spec	Yes	No	0.01
TargetSpy cons. seed sens	Yes	Yes	0.05
TargetSpy cons. seed spec	Yes	Yes	0.01

Πίνακας 8 Υποσύνολα πρόβλεψης targetSpy και threshold

από τη δημοσίευση: TargetSpy: a supervised machine learning approach for microRNA target prediction, Martin Sturm, Michael Hackenberg, David Langenberger, Dmitrij Frishman [Αναφ. 28]

3.4 PITA

3.4.1 Εισαγωγη

Τα microRNAs είναι βασικοί ρυθμιστές της γονιδιακής έκφρασης, αλλά οι ακριβείς μηχανισμοί στους οποίους βασίζεται η αλληλεπίδρασή τους με τους στόχους του mRNA παραμένουν ελάχιστα κατανοητοί. Εδώ, διερευνάται συστηματικά ο ρόλος της προσβασιμότητας του τόπου στόχου, όπως προσδιορίζεται από τις αλληλεπιδράσεις ζευγαρώματος βάσεων εντός του mRNA, στην αναγνώριση στόχου microRNA. Δείχνονται πειραματικά ότι οι μεταλλάξεις που μειώνουν την προσβασιμότητα στο στόχο μειώνουν ουσιαστικά την μεταγραφική καταστολή με τη μεσολάβηση του microRNA, έχοντας αποτελέσματα συγκρίσιμα με εκείνα των μεταλλάξεων που διαταράσσουν τη συμπληρωματικότητα της αλληλουχίας.

PITA, πρόκειται λοιπόν, για ένα μοντέλο χωρίς παραμέτρους για την αλληλεπίδραση microRNA-στόχου που υπολογίζει τη διαφορά μεταξύ της ελεύθερης ενέργειας που αποκτάται από το σχηματισμό του διπλότυπου στόχου microRNA και του ενεργειακού κόστους της μη ανάμειξης του στόχου για να καταστεί προσβάσιμο στο microRNA. Αυτό το μοντέλο προβλέπει τους επικυρωμένους στόχους με μεγαλύτερη ακρίβεια από τους υπάρχοντες αλγόριθμους και δείχνει ότι τα γονιδιώματα προσαρμόζουν την προσβασιμότητα του τόπου, τοποθετώντας κατά προτίμηση τους στόχους σε περιοχές με μεγάλη προσιτότητα. **Η μελέτη αυτή δείχνει ότι η προσβασιμότητα στο στόχο είναι ένας κρίσιμος παράγοντας στη λειτουργία του microRNA**.

3.4.2 ΣΥΝΟΨΗ

1) Γενική περιγραφή δεδομένων εκπαίδευσης - εισόδου
2) Πηγή προέλευσης
3) Το πλήθος τους
4) 1,2,3 για τα δεδομένα αξιολόγησης
5) Χαρακτηριστικά
6) Μέθοδοι μηχανικής μάθησης
7) Περιοχή γονιδιώματος που γίνεται η πρόβλεψη
8) Τύπος πρόσδεσης των miRNAs - seeds
9) Dependencies

---

1) Γονιδιακές αλληλουχίες μύγας (dm2), ποντικού (mm8), ανθρώπου (hg17) και σκουληκιών (ce4) μαζί με σχόλια και διαδρομές διατήρησης. Για τα γονίδια χωρίς σχολιασμό 3’ UTR, χρησιμοποίησαν προβλεπόμενες τιμές 3’ UTRs, με οριακή απόκλιση 500 bp(μύγα), 800 bp (ανθρώπου και ποντικού), 300 bp (σκουλήκι) downstream του άκρου της κωδικοποιημένης αλληλουχίας . Για τους καταλόγους PITA του γονιδιώματος, που μπορούν να ληφθούν από τον ιστότοπο, χρησιμοποίησαν μια χαλαρότερη αποκοπή 1 s.d. πάνω από το μέσο μήκος του 3’ UTR.
2) Από το Πανεπιστήμιο Καλιφόρνιας (Santa Clus, UCSC) Genome Bioinformatics site (http://genome.ucsc.edu) μαζί με σχόλια γονιδίων και διαδρομές διατήρησης.

3) Τα UTRs αντιπροσωπεύουν το 21% (3.914 γονίδια) σε μύγα, 5% (843 γονίδια) στον άνθρωπο, 10% (1.858 γονίδια) σε ποντίκια και 54% (10.821 γονίδια) σε σκουλήκι.
4) Συλλογή πειραματικά δοκιμασμένων ζευγών Drosophila microRNA-mRNA (Supplementary Πίνακας2) βασισμένη σε **1.** προηγούμενη δημοσιευμένη λίστα, στην οποία προστέθηκαν στόχοι αναφερόμενοι σε TarBase και στους στόχους **2.** miR-2 και **3.** miR-184 από τα πειράματα τα δικά τους. Μια συλλογή ιδίου πλήθους τοποθεσιών με την ίδια κατανομή περιεχομένου GC με αυτή των πραγματικών seeds microRNA, επιλέχθηκε τυχαία ως σύνολο ελέγχου. **4.** Για τη δοκιμή περιορισμένης διατήρησης περιοχής seed κατέβασαν διαδρομές συντήρησης βασισμένες σε ένα φυλογενετικό κρυφό μοντέλο Markov (phastCons) (σύγκριση 12 ειδών Drosophila, κουνουπιών, μέλισσας και κόκκινου σκαθαριού για το track μύγας και 17 σπονδυλωτά, συμπεριλαμβανομένων θηλαστικών, αμφιβίων , είδη πτηνών και ψαριών για τα tracks του ανθρώπου και του ποντικιού).
   Origin:
   1. Αναφ. 49.
   2. Αναφ. 25.
   3. Αναφ. 30.
   4. UCSC
   Size: 190 συνολικά ζεύγη

Positive & Negative: 102 αναφέρθηκαν ως λειτουργικοί στόχοι microRNA(positive) και 88 αναφέρθηκαν ως μη λειτουργικοί(negative).

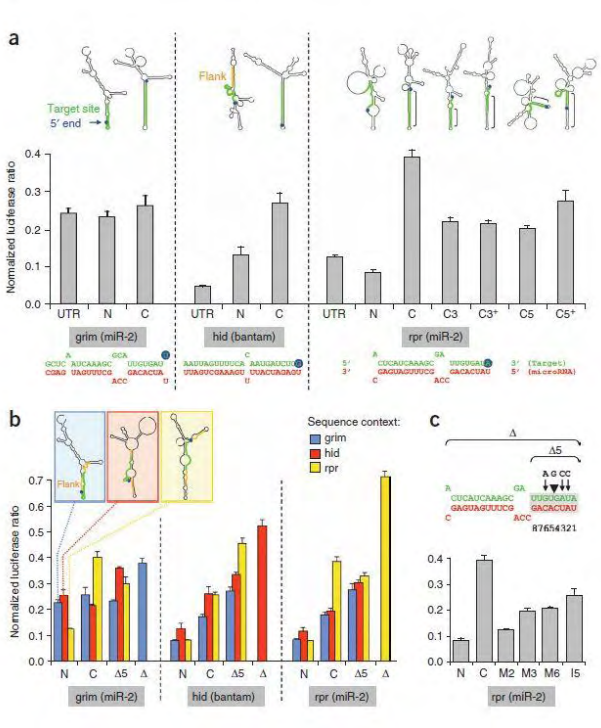
5) ΔΔG = ΔGduplex - ΔGopen
   ΔΔG: ενεργειακή βαθμολογία
   Δgduplex: ενέργεια που αποκτάται από τη δέσμευση του microRNA στο στόχο
   Δgopen: ενέργεια που απαιτείται για να καταστεί η περιοχή-στόχος προσβάσιμη για τη δέσμευση microRNA
   Δε διαθέτει παραμέτρους ο αλγόριθμος παρά μόνο ό,τι χρειάζεται για να αξιοποιηθεί σωστά το χαρακτηριστικό της προσβασιμότητας της υποψήφιας περιοχής-στόχος.
6) -
7) 3' UTR
8) seeds μήκους 6-8 βάσεων, αρχίζοντας από τη θέση 2 του microRNA
   Δεν επιτρέπονται αναντιστοιχίες ή βρόχοι, αλλά επιτρέπεται μία μόνο ταλάντωση G: U σε 7 ή 8 mers.
   Για κάθε οργανισμό που δοκίμασαν, σαρώσανε τα 3' UTRs για τέλεια matches seeds στόχων microRNA, τουλάχιστον επτά βάσεων μήκους, αποκλείοντας τη G: U ταλάντωση, αναντιστοιχίες ή βρόχους.
9) PITA εφαρμογή: http://genie.weizmann.ac.il/pubs/mir07
   Για τα δεδομένα, τα αποτελέσματα, ένα ηλεκτρονικό εργαλείο για την πρόβλεψη των αλληλεπιδράσεων microRNA-στόχου και το εκτελέσιμο αρχείο PITA μπορείτε να δείτε τον αντίστοιχο ιστότοπο.

Σημείωση: Συμπληρωματικές πληροφορίες διατίθενται στον ιστότοπο Nature Genetics.

3.4.3 Δομη Αλγοριθμου

Η διαδικασία

Αρχικά, εστίασαν σε επικυρωμένες θέσεις με σχεδόν τέλεια συμπληρωματικότητα αλληλουχίας με το microRNA σε 3 'UTRs των τριών γονιδίων hid (στοχευμένο από bantam), rpr (miR-2) και grim (miR-2). Και οι τρεις στόχοι διαμεσολάβησαν για τη σημαντική καταστολή στην δοκιμασία αυτή, τόσο όταν χρησιμοποιήσανε UTRs πλήρους μήκους όσο και όταν χρησιμοποιήσανε τα ζευγάρια βάσεων 200 (bp) επικεντρωμένα στο κέντρο (Εικόνα 7α).



Εικόνα 7 Ο ρόλος της προσβασιμότητας της θέσης στόχου microRNA στην καταστολή με τη μεσολάβηση microRNA

από τη δημοσίευση: The role of site accessibility in microRNA target recognition, Michael Kertesz, Nicola Iovino, Ulrich Unnerstall2, Ulrike Gaul & Eran Segal [Αναφ. 33].

Για να ελεγχθεί η επίδραση της προσβασιμότητας του τόπου στη δύναμη της καταστολής των microRNA, ώθησαν στόχους σε δομές εξαιρετικά ζευγαρωμένων δομών με μηχανικές μεταλλάξεις εισάγοντας το ανάστροφο συμπλήρωμά τους κοντά στον στόχο. Αυτές οι κλειστές δομές μειώνουν σημαντικά την καταστολή που προκαλείται από microRNA για τις θέσεις hid και rpr. Ωστόσο, παρατηρήθηκε μόνο μια μικρή διαφορά για το grim σημείο. Η διαφορά στη συμπεριφορά μπορεί να εξηγηθεί από την προσβασιμότητα στο στόχο: η αρχική grid θέση προβλέπεται ότι είναι μέρος μιας κλειστής δομής, σε αντίθεση με τις αρχικές θέσεις rpr και hid, οι οποίες βρίσκονται σε δομές ανοιχτού βρόχου (Εικόνα 7a).

Έπειτα, εξετάσανε αν η σημασία της προσβασιμότητας της τοποθεσίας είναι ασύμμετρη σε σχέση με το 5’ ή 3’ άκρο του microRNA, όπως συμβαίνει στην περίπτωση της συμπληρωματικής αλληλουχίας. Διαπιστώσανε ότι οι δομές των στελεχών που καλύπτουν την περιοχή 5’-seed ή την 3’ συμπληρωματική περιοχή της rpr οδηγούν σε συγκρίσιμη ενδιάμεση καταστολή, υποδηλώνοντας ότι για μια θέση με σημαντική 3’ pairing, η προσβασιμότητα και των δύο τελικών άκρων είναι εξίσου σημαντική (Εικόνα 7a).

Τέλος, διερευνήσανε τις επιδράσεις του ευρύτερου πλαισίου αλληλουχίας στην αποτελεσματικότητα της καταστολής του microRNA με εναλλαγή κάθε θέσης στην άλλη ~200-bp UTR. Τόσο η hid όσο και η rpr θέση λειτουργούσαν εξίσου καλά σε όλα τα πλαίσια αλληλουχίας, σύμφωνα με την προβλεπόμενη ανοικτή δομή τους σε όλες τις UTRs. Αντίθετα, η grim περιοχή λειτούργησε ανεπαρκώς σε grim και hid UTRs, όπου η δομή του είναι κλειστή, αλλά σαφώς καλύτερη σε rpr UTR, όπου η δομή του είναι πιο ανοικτή (Εικόνα 7β).

3.4.4 ΜΕΘΟΔΟΙ

Κατασκευή reporter και δοκιμασία λουσιφεράσης.

Η δοκιμασία χρησιμοποίησε ένα σύστημα διπλής λουσιφεράσης στο οποίο δύο ένζυμα λουσιφεράσης, ένα (από Renilla reniformis) που περιέχει την πειραματική αλληλουχία στόχου και ένα άλλο (από firefly) που περιέχει τον έλεγχο, εκφράζονται από ένα απλό πλασμίδιο.

Και τα δύο γονίδια λουσιφεράσης ελέγχθηκαν με ετερόλογους προαγωγούς (SV40, HSVTK, Promega) και προέκυψε ότι σε κύτταρα S2 αυξήθηκαν δέκα φορές τα επίπεδα μεταγραφής απ' ότι σε ενδογενή επίπεδα (δοκιμάστηκαν για Gli, ttk, τα δεδομένα δεν δείχνονται). Συγκεκριμένα, ακόμη και όταν εισήχθησαν ισχυρές αλληλουχίες στόχου microRNA στη λουσιφεράση της Renilla, η αφθονία των αντιγράφων τους δεν έδειξε σημαντική αλλαγή, υποδεικνύοντας ότι οποιαδήποτε παρατηρούμενη μείωση στην δραστικότητα λουσιφεράσης οφειλόταν σε μεταφραστική καταστολή και όχι σε αποικοδόμηση του μετάγραφου.

Για τη διευκόλυνση της ανάλυσης μεγαλύτερης κλίμακας των αλληλουχιών 3' UTR, δημιουργήσανε μια έκδοση (Invitrogen) του διπλού φορέα λουσιφεράσης psicheck-2 (Promega) συνδέοντας μία κασέτα με αμβλύ άκρο που περιέχει θέσεις attR που φέρουν το γονίδιο ccdB και το γονίδιο αντίστασης σε χλωραμφενικόλη στη θέση PmeI εντός του πολυσυνδετήρα psicheck-2.

Οι αλληλουχίες UTR πλήρους μήκους και οι περικομμένες ενισχύθηκαν με PCR από γονιδιωματικό ή πλασμιδιακό RNA, κλωνοποιήθηκαν σε φορέα pEntr και στη συνέχεια ανασυνδιάστηκαν στον φορέα προορισμού psicheck-2 χρησιμοποιώντας το kit ανασυνδυασμού LR (Invitrogen).

Για τα hid, grim και rpr 3 'UTRs, κατασκευάσανε ~200 bp περικομμένες εκδόσεις χρησιμοποιώντας συνθετικά λίγα νουκλεοτίδια μήκους 50 bp έτσι ώστε η κεντρική θέση στόχου microRNA να φέρεται από θέσεις περιορισμού για τη διευκόλυνση των μεταλλαξιογόνων αντικαταστάσεων (5'-NotI-target site-XhoI-3'). Τα λίγα νουκλεοτίδια καθαρίστηκαν με PAGE, πυρακτώθηκαν, συνδέθηκαν και κλωνοποιήθηκαν σε psicheck2. Επιβεβαιώσανε όλους τους κλώνους με προσδιορισμό αλληλουχίας DNA (Genewiz). Οι αλληλουχίες όλων των κατασκευών που χρησιμοποιήθηκαν σε αυτή τη μελέτη καταρτίζονται στον συμπληρωματικό πίνακα 4. Επιμολύνουνε κύτταρα 10^6 S2 με πλασμίδιο αναφοράς (1 μg) χρησιμοποιώντας Cellfectin (Invitrogen) και μετά από 20 ώρες λύσανε τα κύτταρα και ελέγχανε για τη δράση της λουσιφεράσης χρησιμοποιώντας το kit προσδιορισμού διπλής λουσιφεράσης (Promega). Οι αναλογίες Renilla / firefly λουσιφεράσης κανονικοποιήθηκαν έναντι του κενού φορέα psicheck-2 και καταμετρήθηκαν κατά μέσο όρο σε 4-8 αντίγραφα. Αξιολογήσανε τη στατιστική σημασία χρησιμοποιώντας ANOVA one-factor και two-factor με τη δοκιμασία post-hoc Student-Newman-Keul. Σημειώστε ότι καμία από τις δοκιμαστικές κατασκευές, ακόμη και εκείνες στις οποίες διαγράφηκε ολόκληρη η τοποθεσία, δεν επανέρχεται πλήρως στο επίπεδο έκφρασης του κεντρικού φορέα, υποδηλώνοντας ότι οποιαδήποτε σημαντική προσθήκη αλληλουχίας 3 'UTR (>150 βάσεις) μειώνει την αποτελεσματικότητα της μετάφρασης.

UTRs για προβλέψεις γονιδιωματικών στόχων microRNA.

Οι γονιδιωματικές αλληλουχίες μύγας (dm2), ποντικού (mm8), ανθρώπου (hg17) και σκουληκιών (ce4) λήφθηκαν από το Πανεπιστήμιο Καλιφόρνιας (Santa Clus, UCSC) Genome Bioinformatics Site (http://genome.ucsc.edu/) μαζί με σχόλια γονιδίων και διαδρομές διατήρησης. Για τα γονίδια που δεν έφεραν σχολιασμό 3' UTR, χρησιμοποιήσανε προβλεπόμενες τιμές 3' UTRs, με οριακή απόκλιση 500 bp, 800 bp (ανθρώπου και ποντικού) ή 300 bp (σκουλήκι) downstream του άκρου της κωδικοποιημένης αλληλουχίας . Για τους καταλόγους PITA, που μπορούν να ληφθούν από τον αντίστοιχο ιστότοπο, χρησιμοποιήσανε μια χαλαρότερη αποκοπή 1 s.d. πάνω από το μέσο μήκος του 3' UTR. Τα προβλεπόμενα UTRs αντιπροσωπεύουν το 21% (3.914 γονίδια) σε μύγα, 5% (843 γονίδια) στον άνθρωπο, 10% (1.858 γονίδια) σε ποντίκια και 54% (10.821 γονίδια) σε σκουλήκι.

3.4.5 ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Θερμοδυναμικό μοντέλο για αλληλεπιδράσεις microRNA-στόχου.

Το μοντέλο αξιολογεί τις αλληλεπιδράσεις microRNA-στόχου με μια ενεργειακή βαθμολογία, **ΔΔG**. Το **ΔG duplex** είναι η ελεύθερη ενέργεια δέσμευσης της διπλής δομής microRNA στόχου στην οποία το microRNA και ο στόχος συνδυάζονται σύμφωνα με τους περιορισμούς ζευγαρώματος που επιβάλλονται από τα seeds. Για να υπολογίσουνε αυτή την τιμή, τροποποιήσανε το RNAduplex έτσι ώστε επιπλέον των αλληλουχιών microRNA και στόχου να δίδονται σαφείς συνδυασμοί seeds πάνω στους οποίους επιλέχθηκε ο στόχος και θεωρεί μόνο δομές που συμμορφώνονται με αυτούς τους περιορισμούς. Ο κώδικας υπολογίζει έπειτα την δεσμευτική ελεύθερη ενέργεια κάθε δομής που συμμορφώνεται (ένας υπολογισμός που μπορεί να γίνει αποτελεσματικά με δυναμικό προγραμματισμό) και επιλέγει την ελάχιστη δομή ελεύθερης ενέργειας ως ΔG duplex. Το **ΔG open** υπολογίζεται ως η διαφορά μεταξύ της ελεύθερης ενέργειας του συνόλου όλων των δευτερογενών δομών της περιοχής στόχου και της ελεύθερης ενέργειας όλων των δομών περιοχής-στόχου στις οποίες τα νουκλεοτίδια απαιτείται να είναι μη ζευγαρωμένα. Οι ελεύθερες ενέργειες αυτών των δύο συνόλων υπολογίζονται χρησιμοποιώντας το RNAFold, μεταβάλλοντας όλες τις πιθανές δομές (για το δεύτερο όρο, όλες τις δομές που υπόκεινται στους παραπάνω ανυπότακτους περιορισμούς) και αθροίζοντας κατάλληλα τις αντίστοιχες ελεύθερες ενέργειές τους. Αυτοί οι υπολογισμοί μπορούν να γίνουν αποτελεσματικά χρησιμοποιώντας δυναμικό προγραμματισμό. Η περιοχή του στόχου που δίνεται στο RNAFold για δίπλωση αποτελείται από την περιοχή στόχου (και την περιοχή φθοράς όταν εφαρμόζεται) και 70 επιπλέον νουκλεοτιδίων upstream και downstream. Η τιμή των 70 νουκλεοτιδίων επιλέχθηκε για τη μείωση της χρονικής πολυπλοκότητας των παραπάνω υπολογισμών και βασίζεται στο γεγονός ότι υπάρχει μικρή πιθανότητα αλληλεπιδράσεων ζευγαρώματος βάσης δευτερογενούς δομής μεταξύ νουκλεοτιδίων που διαχωρίζονται από περισσότερα από 70 νουκλεοτίδια (τα δεδομένα δεν παρουσιάζονται).

Τέλος, **η συνολική βαθμολογία αλληλεπίδρασης, ΔΔG, είναι ίση με τη διαφορά μεταξύ ΔG duplex και ΔG open.**

Για την ενσωμάτωση πολλαπλών θέσεων με αποτελέσματα ΔΔG s_1, … , s_n, για ένα microRNA στο ίδιο UTR σε μια συνολική βαθμολογία αλληλεπίδρασης microRNA-UTR, υπολογίζεται το στατιστικό βάρος όλων των παραμέτρων στις οποίες ακριβώς μία από τις θέσεις δεσμεύεται από microRNA σύμφωνα με το $\mathbf{T = \log \sum_{i=1}^{n} e^{si}}$. Επιλέξανε αυτή την απλή μέθοδο ενσωμάτωσης πολλαπλών θέσεων πάνω από τον υπολογισμό της πραγματικής πιθανότητας δέσμευσης microRNA και πάνω από υπολογισμούς που περιλαμβάνουν διαμορφώσεις στις οποίες δύο θέσεις μπορούν να συνδεθούν ταυτόχρονα, επειδή αυτοί οι υπολογισμοί θα απαιτούσαν γνώση μιας επιπλέον (άγνωστης) παραμέτρου ελεύθερης συγκέντρωσης microRNA.

ΔΔG ισούται με τη διαφορά μεταξύ της ενέργειας που αποκτάται από τη δέσμευση του microRNA στο στόχο, ΔG duplex και την ενέργεια που απαιτείται για να καταστεί η περιοχή στόχος προσβάσιμη για τη δέσμευση microRNA, ΔG open.

3.5 EIMMO2

3.5.1 ΕΙΣΑΓΩΓΗ

Τα microRNAs έχουν αναδειχθεί ως σημαντικά ρυθμιστικά γονίδια σε μια ποικιλία κυτταρικών διεργασιών και, τα τελευταία χρόνια, εκατοντάδες τέτοιων γονιδίων έχουν ανακαλυφθεί σε ζώα. Αντίθετα, οι λειτουργικοί σχολιασμοί είναι διαθέσιμοι μόνο για ένα πολύ μικρό κλάσμα αυτών των miRNAs, και ακόμη και σε αυτές τις περιπτώσεις μόνο εν μέρει.

Αναπτύχθηκε μια γενική Bayesian μέθοδο για τη συμπερίληψη των θέσεων στόχων miRNA, όπου, για κάθε miRNA, μοντελοποιήσανε ρητά την εξέλιξη των ορθολογικών θέσεων στόχων σε ένα σύνολο συναφών ειδών. Χρησιμοποιώντας αυτή τη μέθοδο προβλέπουμε τοποθεσίες-στόχους για όλα τα γνωστά miRNAs σε μύγες, σκουλήκια, ψάρια και θηλαστικά. Συγκρίνοντας τις προβλέψεις στη μύγα με ένα σύνολο αναφοράς πειραματικά δοκιμασμένων αλληλεπιδράσεων, δείχνεται ότι η γενική αυτή μέθοδος αποδίδει τουλάχιστον το ίδιο όσο και οι πιο ακριβείς διαθέσιμες μέχρι σήμερα μέθοδοι, συμπεριλαμβανομένων εκείνων που είναι ειδικά προσαρμοσμένες για την πρόβλεψη στόχου στη μύγα. Ένα σημαντικό νέο χαρακτηριστικό του μοντέλου είναι ότι ξεκινάει ρητά τη φυλογενετική κατανομή των λειτουργικών θέσεων στόχων, ανεξάρτητα για κάθε miRNA. Δείχνεται επίσης ότι, σε μακριές ανθρώπινες 3 'UTRs, οι περιοχές στόχοι miRNA υπάρχουν κατά προτίμηση κοντά στην αρχή και κοντά στο τέλος του 3' UTR.

Για να χαρακτηριστεί η λειτουργία miRNA πέρα από τις προβλεπόμενες λίστες στόχων, παρουσιάζεται περαιτέρω μια μέθοδο για να συναγάγουμε σημαντικές συσχετίσεις μεταξύ των συνόλων στόχων που προβλέπονται για μεμονωμένα miRNAs και συγκεκριμένες βιοχημικές οδούς, ιδιαίτερα εκείνες της βάσης δεδομένων για το μονοπάτι KEGG. Δείχνεται ότι αυτή η προσέγγιση ανακτά αρκετές γνωστές λειτουργικές ενώσεις miRNA-mRNA και προβλέπει νέες λειτουργίες για γνωστά miRNAs στην κυτταρική ανάπτυξη.

Παρουσιάζεται λοιπόν, ένας αλγόριθμος πρόβλεψης Bayesian στόχου χωρίς οποιεσδήποτε εύρυθμες παραμέτρους, που μπορούν να εφαρμοστούν σε αλληλουχίες από οποιαδήποτε ομάδα ειδών. Ο αλγόριθμος καταρτίζει αυτόματα τη φυλογενετική κατανομή των λειτουργικών θέσεων για κάθε miRNA και αναθέτει μια πιθανότητα posterior σε κάθε πιθανή τοποθεσία στόχο.

Τα αποτελέσματα που παρουσιάζονται εδώ δείχνουν ότι η γενική μέθοδος επιτυγχάνει πολύ καλές επιδόσεις στην πρόβλεψη θέσεων στόχων miRNA, παρέχοντας ταυτόχρονα πληροφορίες για την εξέλιξη των θέσεων στόχων για μεμονωμένα miRNAs. Επιπλέον, συνδυάζοντας τις προβλέψεις με την ανάλυση των οδών, προτείνονται λειτουργίες συγκεκριμένων miRNA στην ανάπτυξη νευρικού συστήματος, την ενδοκυτταρική επικοινωνία και την κυτταρική ανάπτυξη. Οι πλήρεις προβλέψεις βρίσκονται στον αντίστοιχο ιστότοπο **ElMMo**.

3.5.2 ΣΥΝΟΨΗ

1) Γενική περιγραφή δεδομένων εκπαίδευσης - εισόδου
2) Πηγή προέλευσης
3) Το πλήθος τους
4) 1,2,3 για τα δεδομένα αξιολόγησης
5) Χαρακτηριστικά
6) Μέθοδοι μηχανικής μάθησης
7) Περιοχή γονιδιώματος που γίνεται η πρόβλεψη
8) Τύπος πρόσδεσης των miRNAs - seeds
9) Dependencies

---

1) i)Διαθέσιμα μετάγραφα ανθρώπου, μύγας, ψαριών και σκουληκιών RefSeq
   ii) Στα hg17 (ανθρώπινο), dm2 (μύγα), ceWB05 (σκουλήκι) και danRer3 (ψάρια) γονιδιώματα χαρτογράφησαν όλες τις προβλέψεις

iii) Ζευγαρωμένες ευθυγραμμίσεις αρκετών γονιδιωμάτων με το γονιδίωμα των ειδών αναφοράς, ως εξής: για τον άνθρωπο κατεβάσανε hg17-προς-panTro1, hg17-προς-rheMac2, hg17-tocanFam2, hg17- προς-bosTau2, hg17- Έως-mm7, hg17-προς-rn3, hg17-προς-monDom1 και hg17-προς-galGal2. Για μύγα χρησιμοποιήσανε dm2-προς-droSim1, dm2-προς-droYak1, dm2-προς-droAna1, dm2-προς-dp3, dm2-προς-droMoj1, dm2-προς-droVir1. Για τα ψάρια χρησιμοποιήσανε τα danRer3-to-fr1 και danRer3-to-tetNig1.
iv)Για το σκουλήκι χρησιμοποιήσανε το λογισμικό Threading Blocks igner (TBA) [67] για την ευθυγράμμιση των C. briggsae και C. remanei με τον C. elegans.

2) i)Υπάρχουν στη 17η έκδοση της βάσης δεδομένων Refseq.
ii)Παρέχονται από το Genome Bioinformatics, Ομάδα στο Πανεπιστήμιο της Καλιφόρνιας, Σάντα Κρουζ, Μέσω του προγράμματος cDNA spa
iii) Παρέχονται από το Genome Bioinformatics, Ομάδα στο Πανεπιστήμιο της Καλιφόρνιας, Σάντα Κρουζ
3) –
4) –
5)
   - «πρότυπο διατήρησης» $c_i = 1$, εάν η τοποθεσία διατηρείται στο είδος i και $c_i = 0$ εάν η τοποθεσία δεν είναι διατηρημένη.
   - «πρότυπο υποβάθρου - background» b_t παρατηρεί το πρότυπο διατήρησης c "τυχαία" για ένα seed τύπου t, δηλ. ένα συγκεκριμένο 7-μερές ή 8-μερές με την "τυχαία διατήρηση" εννοούμε ότι δεν υπάρχει ειδική επιλογή για τη διατήρηση της συμπληρωματικότητας της εν λόγω περιοχής στο 5' άκρο του miRNA.
   - «πρότυπο επιλογής» $s_i = 1$, εάν ο χώρος είναι λειτουργικός (υπό την επιλογή) στο είδος i, και $s_i = 0$ διαφορετικά.
   - Φυλογένεια, posterrior πιθανότητες (βλέπε παρακάτω).
   - «conservation fold enrichment» f_t του τύπου seed t ορίζεται ως ο λόγος των παρατηρούμενων προτύπων διατήρησης και των ρυθμών διατήρησης του background: $f_t = c_t / b_t$.
6) Bayesian μοντέλο
7) 3’ UTR
8)
   - Τέλεια συμπληρωματικότητα με τις αλληλεπιδράσεις Watson-Crick μεταξύ των θέσεων 1-8 του miRNA και της θέσης στόχου mRNA.
   - Τέλεια συμπληρωματικότητα στις θέσεις 1-7, αλλά όχι θέση 8.
   - Τέλεια συμπληρωματικότητα στις θέσεις 2-8 αλλά όχι στην θέση 1.
   - Τέλεια συμπληρωματικότητα σε 2-7 αλλά όχι σε 1 και8.
   - Συμπληρωματικότητα στις θέσεις 1-8 με ένα μοναδικό G-U pair που συμβαίνει με το U στο miRNA (GUM).
   - Συμπληρωματικότητα στις θέσεις 1-8 με ένα μόνο G-U pair που συμβαίνει με το U στο στόχο (GUT).
   - Συμπληρωματικότητα με ένα μονόπλευρο νουκλεοτίδιο στην πλευρά miRNA (BM).
   - Συμπληρωματικότητα με ένα απλό διογκωμένο νουκλεοτίδιο στην πλευρά στόχου (BT).
   - Συμπληρωματικότητα με ένα μόνο εσωτερικό βρόχο που περιλαμβάνει νουκλεοτίδιο τόσο στο miRNA όσο και στο στόχο (LP).

Ο αλγόριθμος επικεντρώνεται στους 3 πρώτους τύπους seeds που παρουσιάζουν ισχυρές ενδείξεις εμπλουτισμού διατήρησης σε όλα τα clades, εκείνα με τέλεια συμπληρωματικότητα Watson-Crick με θέσεις 1-7, 2-8 ή 1-8 του miRNA.

9) EIMMO2 εφαρμογή: www.mirz.unibas.ch/ElMMo2/
Οι πλήρεις προβλέψεις για τον στόχο καθώς και οι ενώσεις miRNA / μονοπατιού είναι διαθέσιμες στον εξυπηρετητή ιστού ElMMo.

3.5.3 Δομη Αλγοριθμου

Η διαδικασία

Bayesian Φυλογενετικό μοντέλο

Πρόκειται για ένα Bayesian πιθανοτικό μοντέλο για την εκχώρηση, σε κάθε πιθανή "θέση" σε ένα 3' UTR που είναι συμπληρωματικό ως προς ένα seed miRNA, μια posterrior πιθανότητα ότι η θέση αυτή είναι μια λειτουργική περιοχή στόχος για το miRNA. Τα κύρια συστατικά του μοντέλου αυτού είναι τα εξής:

A. Για κάθε miRNA και κάθε τύπο seed t συλλέγουν όλες τις πιθανές θέσεις στα 3 'UTRs του είδους αναφοράς της εν λόγω κατηγορίας, δηλ. τα 3' UTR τμήματα που είναι συμπληρωματικά προς το δεδομένο seed miRNA.
B. Για κάθε μία από αυτές τις υποτιθέμενες θέσεις προσδιορίζουν τότε το πρότυπο διατήρησης c, που ορίζεται ως ένας δυαδικός φορέας με $c_i = 1$ εάν η θέση διατηρείται στα είδη i και $c_i = 0$ αν δεν συντηρείται στα είδη i.
C. Στη συνέχεια υπολογίζουν τον αριθμό των χρόνων n (c, t) που το πρότυπο συντήρησης c παρατηρείται για πιθανές θέσεις-στόχους τύπου seed t.
D. Για να υπολογίσουν τις πιθανότητες posterior για μεμονωμένους τόπους, το μοντέλο στη συνέχεια συγκρίνει αυτούς τους αριθμούς n (c, t) με εκείνους που θα αναμένονταν δεδομένων των "background "συχνοτήτων ρ (c | t, bg) με τις οποίες επιλέχθηκαν τυχαία τμήματα αλληλουχίας 3 'UTR του ίδιου μήκους με το seed miRNA που παρουσιάζει το πρότυπο συντήρησης c.

Το πρότυπο διατήρησης μιας δεδομένης θέσης είναι τυπικά το αποτέλεσμα της επιλογής της διατήρησης της θέσης σε ορισμένα από τα είδη σε συνδυασμό με την πιθανή διατήρηση της θέσης σε άλλα είδη, ιδιαίτερα εκείνα που είναι εξελικτικά κοντά. Το μοντέλο EIMMO2 το λαμβάνει υπόψη αυτό και θεωρεί όλα τα πιθανά "μοτίβα επιλογής" s, τα οποία είναι και δυαδικοί φορείς, με $s_i = 1$ αν ο χώρος βρίσκεται υπό επιλογή στο είδος i, και $s_i = 0$ αν δεν είναι.

E. Για κάθε miRNA προσδιορίζονται οι συχνότητες p (s) διαφορετικών προτύπων επιλογής που μεγιστοποιούν το σύνολο πιθανοτήτων των παρατηρούμενων χρόνων n (c, t). Δηλαδή, καθορίζουν την κατανομή των μοτίβων επιλογής p (s) που εξηγεί καλύτερα τις παρατηρούμενες μετρήσεις n (c, t) των προτύπων διατήρησης για αυτό το miRNA.
F. Χρησιμοποιώντας τις εκτιμώμενες συχνότητες p (s) μπορούν στη συνέχεια να προσδιορίσουν, για κάθε υποθετική περιοχή στόχου, την posterrior πιθανότητα ότι η περιοχή είναι λειτουργική δεδομένου του μοντέλου διατήρησής c.
G. Τέλος, για να προσδιοριστεί η συνολική πιθανότητα ότι ένα δεδομένο 3 'UTR στοχεύεται από ένα δεδομένο miRNA, συνδυάζουν τις πιθανές πιθανότητες όλων των θέσεων για το miRNA που εμφανίζεται στο 3' UTR.

Η σημασία των διαφορετικών τύπων seeds

Διάφορες ενδείξεις [Αναφ. 22, 27-29, 37] υποδηλώνουν ότι η συμπληρωματικότητα της θέσης στόχου με τις πρώτες 8 βάσεις στο 5 'άκρο του miRNA είναι καθοριστικής σημασίας για την αναγνώριση της θέσης στόχου. Lewis [Αναφ. 22] έχουν διερευνήσει τη σημασία του "seed" miRNA, που ορίζεται ως οι θέσεις 2-7 του miRNA, συγκρίνοντας στατιστικά στοιχεία διατήρησης τμημάτων mRNA που είναι συμπληρωματικά προς τους miRNA seeds με αυτά των τυχαίων συνόλων ελέγχου. **Καταλήγουν στο συμπέρασμα ότι οι διατηρούμενες περιοχές 3 'UTR που προβλέπεται να υβριδοποιηθούν τέλεια με τις θέσεις 2-8 του miRNA ή με τις θέσεις 2-7 του miRNA, αλλά έχοντας ένα νουκλεοτίδιο A που πλαισιώνει την αντιστοίχιση των seeds στο 3' άκρο, είναι πιθανό να είναι στόχος miRNA.** Από αυτές τις μεθόδους αποφασίσανε να επανεξετάσουνε τα στατιστικά στοιχεία διατήρησης διαφορετικών "τύπων seeds" σε διαφορετικά σύνολα οργανισμών. Σε όλες τις περιπτώσεις, επικεντρώσανε σε μόνο τις πρώτες 8 θέσεις του miRNA και αναλύσανε τους ακόλουθους 9 τύπους seeds:

1. Τέλεια συμπληρωματικότητα με τις αλληλεπιδράσεις Watson-Crick μεταξύ των θέσεων 1-8 του miRNA και της θέσης στόχου mRNA.

2. Τέλεια συμπληρωματικότητα στις θέσεις 1-7, αλλά όχι θέση 8.

3. Τέλεια συμπληρωματικότητα στις θέσεις 2-8 αλλά όχι στην θέση 1.

4. Τέλεια συμπληρωματικότητα σε 2-7 αλλά όχι σε 1 και8.

5. Συμπληρωματικότητα στις θέσεις 1-8 με ένα μοναδικό G-U pair που συμβαίνει με το U στο miRNA (GUM).

6. Συμπληρωματικότητα στις θέσεις 1-8 με ένα μόνο G-U pair που συμβαίνει με το U στο στόχο (GUT).

7. Συμπληρωματικότητα με ένα μονόπλευρο νουκλεοτίδιο στην πλευρά miRNA (BM).

8. Συμπληρωματικότητα με ένα απλό διογκωμένο νουκλεοτίδιο στην πλευρά στόχου (BT).
9. Συμπληρωματικότητα με ένα μόνο εσωτερικό βρόχο που περιλαμβάνει νουκλεοτίδιο τόσο στο miRNA όσο και στο στόχο (LP).

Για κάθε ένα από αυτούς τους 9 τύπους seed, t και κάθε ένα από τα τέσσερα clades (θηλαστικά συν κοτόπουλο, ψάρια, μύγες και σκουλήκια), προσδιορίσανε **1.** το κλάσμα f_t των πιθανών θέσεων στόχων που διατηρούνται τέλεια σε όλα τα είδη του clade. Θεωρήσανε μόνο miRNAs τα οποία διατηρήθηκαν σε όλα τα είδη του clade. Προσδιορίσανε επίσης **2.** το κλάσμα συντήρησης "**background",** τυχαίων επιλεγμένων τμημάτων αλληλουχίας 3 'UTR του ίδιου μήκους με τους αντίστοιχους τύπους seeds που διατηρούνται σε όλα τα είδη της ομάδας.

Η αναλογία αυτών των δύο κλασμάτων, ονομάστηκε **"conservation fold enrichment"**. Όπως αναμένεται, οι 8mer τοποθεσίες έχουν τα περισσότερα στοιχεία λειτουργικότητας.

Το κλάσμα αυτό μειώνεται δραματικά καθώς μειώνεται η έκταση της συμπληρωματικότητας που απαιτείται μεταξύ του miRNA και του υποθετικού στόχου. Συγκεκριμένα, οι θέσεις στις οποίες μόνο τα νουκλεοτίδια 2-7 του miRNA προβλέπεται ότι σχηματίζουν ζεύγη βάσεων με το mRNA, καθώς και θέσεις που προβλέπουν ότι σχηματίζουν ζεύγη βάσεων G-U ή περιέχουν εσωτερικούς βρόχους παρουσιάζουν σχετικές ελάχιστες ενδείξεις εμπλουτισμού διατήρησης. Αυτό δεν σημαίνει ότι τέτοιες τοποθεσίες δεν λειτουργούν ποτέ.

Ωστόσο, για τη μέθοδο πρόβλεψης στόχων αποφασίσανε να επικεντρωθούν στους τρεις τύπους seeds που παρουσιάζουν ισχυρές ενδείξεις εμπλουτισμού διατήρησης σε όλα τα clades: εκείνα με τέλεια συμπληρωματικότητα Watson-Crick με θέσεις 1-7, 2-8 ή 1-8 του miRNA.

3.5.4 Βασικες Αναλυσεις

Bayesian φυλογενετικός αλγόριθμος ταυτοποίησης miRNA στόχου

Για κάθε miRNA και για κάθε έναν από τους τρεις τύπους seeds προσδιορίζουνε χωριστά τις θέσεις υποψήφιου στόχου και εκχωρούνε μια πιθανότητα posterior σε κάθε τοποθεσία στόχο ως εξής.

- Αρχικά βρίσκουνε όλες τις "τοποθεσίες" που είναι συμπληρωματικές του seed στα 3' UTR του είδους αναφοράς.
- Χρησιμοποιώντας ζευγαρωτές ευθυγραμμίσεις μεταξύ των ειδών αναφοράς και των άλλων ειδών προσδιορίζουνε, για κάθε υποθετική περιοχή, ποια άλλα είδη έχουν διατηρήσει τη θέση. Μια μεμονωμένη περιοχή θεωρήθηκε συντηρημένη εάν όλα τα ζεύγη βάσεων που προβλεπόταν να σχηματιστούν μεταξύ του miRNA και αυτής της θέσης στο είδος αναφοράς θα μπορούσαν επίσης να σχηματιστούν με τις αντίστοιχες θέσεις, που εξάγονται από τις ευθυγραμμίσεις του γονιδιώματος στα άλλα είδη.

  Αυτό το ορίζει ένα «πρότυπο διατήρησης» για κάθε θέση, το οποίο είναι ένας δυαδικός c με $c_i = 1$ εάν η τοποθεσία διατηρείται στα είδη i και $c_i = 0$ εάν η τοποθεσία δεν είναι διατηρημένη. Για παράδειγμα, για την τριπλέτα των C. elegans, C. briggsae και C. remanei, χρησιμοποιώντας τον C. elegans ως είδος αναφοράς, ο φορέας c = (1, 1) υποδεικνύει μια θέση C. elegans που διατηρείται και στα δύο άλλα σκουλήκια, ο φορέας c = (1, 0) μια θέση που διατηρείται μόνο στο C. briggsae, ο φορέας c = (0, 1) μια θέση που διατηρείται μόνο στο C. Remanei και ο φορέας c = (0, 0) ένας τόπος συντηρημένος σε κανένα από τα άλλα δύο σκουλήκια.

Το γεγονός ότι μια πιθανή τοποθεσία στόχος διατηρείται, δεν σημαίνει απαραίτητα ότι η περιοχή είναι λειτουργική. Για παράδειγμα, τα 7-μερή και 8-μερή των miRNA seeds, μπορούν εύκολα να συντηρηθούν τυχαία.

Αυτή η εξελικτική εξάρτηση μεταξύ των ορθολογικών περιοχών μπορεί να ληφθεί υπόψη με διάφορους τρόπους. Για παράδειγμα, στο RNAhybrid τα pvalues για τις ορθολογικές θέσεις-στόχους συνδυάζονται προσαρμόζοντας έναν "αποτελεσματικό" αριθμό ορθολογικών αλληλουχιών στην παρατηρούμενη κατανομή ρ-τιμών για τυχαία παραγόμενα miRNAs.

Εδώ στοχεύουν στην ενσωμάτωση των στατιστικών διατήρησης σε ένα Bayesian πλαίσιο που λαμβάνει ρητά υπόψη τη φυλογένεια του είδους και αναγνωρίζει ότι μια διατηρημένη περιοχή μπορεί να βρίσκεται υπό επιλογή σε οποιαδήποτε από τις υποομάδες των ειδών στις οποίες διατηρείται η τοποθεσία.

- Για το σκοπό αυτό καθορίζουνε πρώτα ένα "background" που δίνει τις πιθανότητες p (c | t, bg) να παρατηρήσει το πρότυπο διατήρησης c "τυχαία" για ένα seed του τύπου t, δηλ. ένα συγκεκριμένο 7-μερές ή 8-μερές. Με την "τυχαία συντήρηση" εννοούμε ότι δεν υπάρχει ειδική επιλογή για τη διατήρηση της συμπληρωματικότητας της εν λόγω περιοχής στο 5 'άκρο του miRNA. Ωστόσο, δεν χρησιμοποιήσανε ένα μοντέλο background που αντικατοπτρίζει απλώς τις πιθανότητες να παρατηρηθούν διαφορετικά πρότυπα συντήρησης υπό ουδέτερη εξέλιξη. Οποιοσδήποτε συγκεκριμένος υποψήφιος χώρος στόχος μπορεί να επικαλύπτεται ή να αποτελεί μέρος ενός τόπου που είναι λειτουργικός για κάποιο άλλο λόγο και μπορεί συνεπώς να είναι πιο συντηρημένος από ότι αναμένεται μόνο υπό την ουδέτερη εξέλιξη.
- Επομένως, για να υπολογίσουνε τις πιθανότητες του background p (c | t, bg) υπολογίσανε τη συνολική συχνότητα με τους οποίους όλα τα πρότυπα συντήρησης c εμφανίζονται στις ευθυγραμμίσεις, υπολογίζονται κατά μέσον όρο για όλες τις 8 μετρήσεις για τον τύπο seed 1-8 και υπολογίζονται κατά μέσο όρο για όλες τις 7 μετρήσεις για τους τύπους seed 1-7 και 2-8. Σε προηγούμενες εργασίες, άλλοι [Αναφ. 22] έχουν υπολογίσει τις συχνότητες background των διατηρημένων matches seed ανεξάρτητα για τα seeds που έχουν διαφορετικές απόλυτες συχνότητες στα 3 'UTRs του είδους αναφοράς. Αντίθετα, απαιτούν μόνο τις σχετικές συχνότητες διαφορετικών προτύπων διατήρησης .

- Στη συνέχεια υπολογίσανε πόσο πιθανό είναι να παρατηρήσουνε διαφορετικά πρότυπα διατήρησης c δεδομένου ότι η υποτιθέμενη περιοχή στόχος είναι λειτουργική σε τουλάχιστον ένα από τα είδη.
- Για να μπορέσουν να υπολογίσουν τις σημαντικές πιθανότητες παρατηρώντας διαφορετικά πρότυπα διατήρησης c για λειτουργικούς τόπους, κάνουν τις ακόλουθες απλουστευτικές υποθέσεις. Πρώτον, υποθέτουν ότι δεδομένου ενός συνόλου διατηρημένων υποθετικών θέσεων στόχων, κάθε μία από τις συντηρημένες τοποθεσίες μπορεί να είναι είτε "λειτουργική" είτε "μη λειτουργική". Στο πλαίσιο αυτό, το "λειτουργικό" σημαίνει ότι η επιλογή έχει ενεργήσει για να εξασφαλίσει ότι ο χώρος στόχος παραμένει συντηρημένος και "μη λειτουργικός" σημαίνει ότι ο τόπος στόχος έχει εξελιχθεί σύμφωνα με το background. Για να πάρουμε το παράδειγμα του σκουληκιού, εάν μια λειτουργική περιοχή του C. elegans είναι λειτουργική και στα δύο άλλα σκουλήκια, τότε η περιοχή θα διατηρηθεί αναγκαστικά και στα δύο, δηλαδή θα έχουμε c = (1, 1). Εάν η περιοχή είναι λειτουργική στο C. Briggsae μόνο τότε μπορούμε να παρατηρήσουμε είτε το c = (1, 0) ή το c = (1, 1), επειδή η περιοχή είναι αναγκαστικά συντηρημένη στο C. briggsae και μπορεί να παραμείνει μόνιμη κατά τύχη στην C. remanei.
- Έτσι, σε γενικές γραμμές, εξετάζουμε όλα τα πιθανά "μοτίβα επιλογής" για τον τόπο στα διάφορα είδη. Όπως και το conservation pattern,το μοτίβο επιλογής s είναι ένας δυαδικός φορέας με $s_i = 1$ αν ο χώρος είναι λειτουργικός (υπό την επιλογή) στο είδος i, και $s_i = 0$ διαφορετικά. Υπολογίζουμε τις πιθανότητες P (c | t, s) για να παρατηρήσουμε το πρότυπο διατήρησης c που δίνεται επιλέγοντας (και τον τύπο seed t) ως εξής:
  Έστω C (s) υποδηλώνει το σύνολο όλων των προτύπων διατήρησης c που είναι συνυφασμένο με το μοτίβο επιλογής s. Για να είναι συνεπής με το πρότυπο επιλογής, η τοποθεσία πρέπει να διατηρείται σε όλα τα είδη στα οποία υποτίθεται ότι βρίσκεται υπό επιλογή, δηλ. για όλα τα c στο C (s) έχουμε ότι $c_i = 1$ για όλα τα i για τα οποία $s_i=1$. Η πιθανότητα p (c | t, s) δίνεται στη συνέχεια από

$$p(\vec{c} \mid t, \vec{s}) = \frac{p(\vec{c} \mid t, \mathrm{bg})}{\sum_{\vec{c}' \in C(\vec{s})} p(\vec{c}' \mid t, \mathrm{bg})}. \qquad (1)$$

- Σημειώστε ότι το p (c | t, s) είναι απλώς η πιθανότητα ότι η θέση είναι συντηρημένη τυχαία σε εκείνα τα είδη που έχουν $c_i = 1$ αλλά δεν είναι υπό επιλογή, δηλ. $s_i = 0$.
- Τέλος, πρέπει να προσδιορίσουμε ποσοτικά το πόσο πιθανό είναι εκ των προτέρων η επιλογή ενός συγκεκριμένου τόπου στο είδος αναφοράς σε ένα συγκεκριμένο υποσύνολο των άλλων ειδών. Δηλαδή, χρειαζόμαστε μια προηγούμενη κατανομή πιθανότητας p (s) που δίνει την πιθανότητα ότι μια περιοχή miRNA θα είναι υπό επιλογή σε όλα τα είδη i για τα οποία $s_i = 1$.
- Ένα από τα βασικά νέα χαρακτηριστικά του μοντέλου είναι ότι επιτρέπεται αυτήν η προηγούμενη κατανομή p (s) να μεταβάλλεται μεταξύ διαφορετικών miRNAs. Λαμβάνεται συνεπώς υπόψη η ειδική διατήρηση των λειτουργικών στόχων ως προς το είδος ή το clade, δηλαδή ότι τα είδη αναφοράς μπορούν να μοιράζονται λειτουργικές θέσεις στόχους με διαφορετικά υποσύνολα ειδών για διαφορετικά miRNAs.
- Για κάθε miRNA πρέπει να υπολογίσουμε τα προηγούμενα p (s) όλων των πιθανών προτύπων επιλογής. Για να γίνει αυτό, μπορούμε πρώτα να χρησιμοποιήσουμε τη διατήρηση του γονιδίου miRNA. Δηλαδή, αν το γονίδιο miRNA δεν είναι διατηρημένο σε ένα δεδομένο είδος i, τότε θα υποθέσουμε ότι οι θέσεις αυτού του miRNA δεν μπορούν να βρίσκονται υπό επιλογή στο είδος i. Έτσι, για κάθε miRNA στο είδος αναφοράς ελέγχεται ποιο από τα άλλα είδη περιέχει ένα miRNA με το ίδιο seed. Στη συνέχεια, ορίστε p (s) = 0 για όλους τους φορείς s στα οποία η περιοχή υποτίθεται ότι έχει υποβληθεί σε επιλογή σε ένα είδος που δεν περιέχει το miRNA. Σημειώστε ότι είναι πιθανό με τη σύνθεση γονιδιώματος να υπάρχει πιθανότητα να χάσουμε το ορθόλογο ενός

συγκεκριμένου γονιδίου miRNA. Αυτό θα έχει ως αποτέλεσμα να αγνοηθούν οι πληροφορίες διατήρησης αυτού του είδους για το συγκεκριμένο miRNA.

- Η πιο γενική προσέγγιση για την εκτίμηση του p (s) θα ήταν να βρούμε απλώς το p (s) διανομής που έχει συνολική μέγιστη πιθανότητα δεδομένων. Επίσημα, η πιθανότητα p (c, t) για να παρατηρήσουμε το πρότυπο συντήρησης c για μια δεδομένη πιθανή θέση στόχου του τύπου seed t δίνεται από την εξίσωση (2)

$$p(\vec{c},t) = \sum_{\vec{s}\in S} p(\vec{c} \mid t,\vec{s})p(\vec{s}), \qquad (2)$$

- Έστω ότι το n (c, t) υποδηλώνει τον αριθμό των περιπτώσεων υποθετικών θέσεων στόχων του τύπου seed t που έχουν πρότυπο διατήρησης c. Η πιθανότητα L που δίνεται στα δεδομένα, δηλ. οι παρατηρούμενες μετρήσεις n (c, t), δίνονται στη συνέχεια από τον τύπο

$$L = \prod_{\vec{c},t} p(\vec{c},t)^{n(\vec{c},t)}. \qquad (3)$$

Δεδομένου ότι υπάρχουν επαρκή δεδομένα, δηλαδή n (c, t) >> 0 για όλα τα c_i θα μπορούσαμε να υπολογίζουμε το p (s) μεγιστοποιώντας το L σε σχέση με το p (s). Ωστόσο, η ποσότητα των δεδομένων είναι περιορισμένη και η διανομή p (s) γενικά έχει μεγάλο αριθμό ανεξάρτητων συστατικών (2^g για είδη g). Επειδή πιστεύουμε ότι δεν είναι πιθανό να χωρέσει σωστά ολόκληρη τη διανομή p (s) χωρίς σημαντικό κίνδυνο υπερβολικής τοποθέτησης, στόχος ήταν να μετριάσουν λογικές διανομές p (s) χρησιμοποιώντας ένα πολύ μικρότερο σύνολο παραμέτρων, δηλ. g αντί των παραμέτρων 2^g. Μια δεύτερη πληροφορία που μπορεί να βοηθήσει είναι ότι το p (s) αποτελείται από τις φυλογενετικές σχέσεις μεταξύ των ειδών. Δηλαδή, θα αναμενόταν γενικά ότι οι λειτουργικοί στόχοι των ειδών αναφοράς είναι συχνότερα λειτουργικοί και σε στενά συγγενή είδη από ό, τι σε μακρινά. Είναι επομένως φυσικό να μοντελοποιήσουν την εξέλιξη των μοτίβων επιλογής κατά μήκος των κλάδων του φυλογενετικού δένδρου της ομάδας. Σε αναλογία με εξελικτικά μοντέλα για την εξέλιξη των αλληλουχιών γονιδίων θα μπορούσε κανείς να εξετάσει τα μοντέλα στα οποία η επιλογή για μια περιοχή μπορεί να «μεταλλαχθεί» από το «on» σε «off» κατά μήκος κάθε κλαδιού του δέντρου, με πιθανότητα «μετάλλαξης» ανάλογη με το μήκος του κλάδου.

- Για την παραμετροποίηση του p (s) θα έπρεπε λοιπόν να τοποθετήσουμε ανεξάρτητα ποσοστά απώλειας και κέρδους επιλογής κατά μήκος κάθε κλάδου του δέντρου για το κάθε miRNA. Επιπλέον, για κάθε μοτίβο επιλογής θα πρέπει να εξετάσουμε όλες τις εξελικτικές ιστορίες της απώλειας και κέρδους επιλογής που είναι συνεπείς με το προκύπτον μοτίβο επιλογής στα φύλλα του δέντρου.

3.6 PicTar2

3.6.1 Εισαγωγη

Τα microRNAs είναι μικρά μη κωδικοποιημένα RNA που αναγνωρίζουν και δεσμεύονται σε μερικώς συμπληρωματικές θέσεις στις μη μεταφραζόμενες περιοχές των γονιδίων-στόχων σε ζώα και, με άγνωστους μηχανισμούς, ρυθμίζουν την παραγωγή πρωτεϊνών του μεταγραφικού στόχου [1-3]. Διαφορετικοί συνδυασμοί microRNAs εκφράζονται σε διαφορετικούς κυτταρικούς τύπους και μπορούν να συντονίζουν τα γονίδια-στόχους των κυττάρων. Εδώ παρουσιάζεται ο **PicTar2, μια υπολογιστική μέθοδος για τον εντοπισμό κοινών στόχων των microRNAs.** Στατιστικά τεστ που χρησιμοποιούν γονιδιωματικές ευθυγραμμίσεις σε οκτώ γονιδιώματα σπονδυλωτών, η ικανότητα του PicTar να

ανακάμψει ειδικά τους δημοσιευμένους στόχους microRNA και η πειραματική επικύρωση επτά προβλεπόμενων στόχων υποδηλώνουν ότι το PicTar έχει καλό ποσοστό επιτυχίας στην πρόβλεψη στόχων για μεμονωμένα microRNAs και για συνδυασμούς microRNAs. Συγκεκριμένα, επικυρώνει πειραματικά την κοινή ρύθμιση του Mtpn με miR-375, miR-124 και let-7b και έτσι παρέχουν αποδείξεις για τον έλεγχο του συντονισμού microRNA σε θηλαστικά.

3.6.2 ΣΥΝΟΨΗ

1) Γενική περιγραφή δεδομένων εκπαίδευσης - εισόδου
2) Πηγή προέλευσης
3) Το πλήθος τους
4) 1,2,3 για τα δεδομένα αξιολόγησης
5) Χαρακτηριστικά
6) Μέθοδοι μηχανικής μάθησης
7) Περιοχή γονιδιώματος που γίνεται η πρόβλεψη
8) Τύπος πρόσδεσης των miRNAs - seeds
9) Dependencies

---

1) I) πολλαπλές ευθυγραμμίσεις γονιδιώματος οκτώ σπονδυλωτών - Ανθρώπου, Μάιος 2004 (hg17). Χιμπατζή, Νοέμβριος 2003 (panTro1). Ποντίκι, Μάιος 2004 (mm5). Αρουραίος, Ιούνιος 2003 (rn3); Σκύλος, Ιούλιος 2004 (canFam1); Κοτόπουλο, Φεβρουάριος 2004 (galGal2). Puffer fish, Αύγουστος 2002 (fr1) και Ζέβρα, Νοέμβριος 2003 (danRer1).
   Χρησιμοποιήσανε τις απεικονίσεις UCSC των δεδομένων mRNA του ανθρώπινου RefSeq(Έκδοση 6) στο ανθρώπινο γονιδίωμα για να καθορίσουν πολλαπλές ευθυγραμμίσεις των 3' UTRs.
   II) σταθερό σετ αναζήτησης συνεκφρασμένων microRNAs απεικονίσεις UCSC ανθρώπινων αλληλουχιών mRNA RefSeq, 500 bp upstream των θέσεων έναρξης μεταγραφής («promoters»).
   Για να εξαιρέσουνε πιθανές επικαλύψεις με τις 3' UTRs, δεν συμπεριλάβανε αλληλουχίες που επικαλύπτονταν με οποιοδήποτε μετάγραφο, φτάνοντας σε συνολικά 17.883 ανθρώπινες αλληλουχίες.
2) I) Βάση Δεδομένων Genomic UCSC
   II) RefSeq
3) 19.971 αλληλουχίες για τον άνθρωπο και τον χιμπατζή.
   19.289 για τον άνθρωπο, τον χιμπαντζή και τον ποντικό.
   18.717 για ανθρώπους, χιμπαντζήδες, ποντικούς και αρουραίους.
   18.567 για ανθρώπους, χιμπαντζήδες, ποντικούς, αρουραίους και σκύλους.
   11.190 για ανθρώπους, χιμπαντζήδες, ποντικούς, αρουραίους, σκύλους και κοτόπουλο.
   6.136 για ανθρώπινο, χιμπαντζή, ποντίκι, αρουραίο, σκύλο, κοτόπουλο και pufferfish.
   4.355 για ανθρώπους, χιμπαντζήδες, ποντικούς, αρουραίους, σκύλους, κοτόπουλο, pufferfish και ζέβρα.
4) **i)** Για την εκτίμηση ψευδώς θετικών ποσοστών για τις προβλέψεις σπονδυλωτών στόχων microRNA, παρήγαγαν ώριμες αλληλουχίες microRNA και προσθέσανε εννέα microRNAs. Εκπόνησαν ένα υποσύνολο miRNA που διατηρήθηκε μεταξύ ανθρώπου, χιμπαντζή, ποντικού, αρουραίου, σκύλου και κοτόπουλου χρησιμοποιώντας σημειώσεις Rfam ώριμων miRNAs σπονδυλωτών ομόλογων σε ανθρώπινο microRNA. Όποτε δεν υπήρχε σχολιασμός, χρησιμοποίησαν αυστηρά κριτήρια. Κατασκευάσανε ένα σύνολο μοναδικών microRNAs συγκεντρώνοντας μαζί τα microRNAs με πανομοιότυπες βάσεις στις θέσεις 1-7 ή 2-8 (ξεκινώντας από το 5' end).

Για να δοκιμάσουν το PicTar, το εφαρμόσανε για να αναζητήσουνε στο σύνολο του γονιδιώματος C. elegans και Caenorhabditis briggsae 3' UTR αλληλουχίες για στόχους lin-4 ή let-7. Οι γνωστοί στόχοι lin-14, hbl-1, daf-12 και lin-28 κατατάχθηκαν ως εξής πρώτος, δεύτερος, τέταρτος και έβδομος αντίστοιχα και μόνο ένα γνωστό γονίδιο στόχος (lin-41) δεν ανακτήθηκε, υποδηλώνοντας ότι το PicTar έχει καλό specificity και sensitivity.
ii) Εξετάσανε περαιτέρω το PicTar υπολογίζοντας προβλέψεις για κάθε microRNA ξεχωριστά σε όλα τα C. elegans 3 'UTRs χωρίς συγκρίσεις μεταξύ των ειδών. Οι δοκιμές τυχαίας επιλογής έδειξαν ότι ένα κλάσμα εξαιρετικά σημαντικών (>10 s.d.) προβλεπόμενων θέσεων διατηρείται εξελικτικά, ενισχύοντας την εμπιστοσύνη στο PicTar.
Origin:
i) από Rfam (Release 5.0)
ii) βάση δεδομένων του Πανεπιστημίου της Καλιφόρνια στη Σάντα Κρουζ (UCSC)
Size:

Για προβλέψεις στόχων σε σπονδυλωτά, κατασκευάσανε πολλαπλές ευθυγραμμίσεις 20.254 σχολιασμένων ανθρώπινων 3' UTRs σε γονιδιακές αλληλουχίες από επτά άλλα σπονδυλωτά, χιμπαντζή, ποντίκι, αρουραίο, σκύλο, κοτόπουλο, pufferfish και ζέβρα.

Το αποτέλεσμα ήταν 58 μοναδικά microRNA που διατηρούνται σε ανθρώπους, χιμπατζήδες, ποντικούς, αρουραίους, σκύλους και κοτόπουλο.

5)
- η τέλεια αντιστοιχία seeds
- η αναζήτηση συνδυασμών θέσεων δέσμευσης microRNA για σύνολα συν-εκφρασμένων miRNAs.
- score για κάθε υποψήφια θέση - βέλτιστη εκτίμηση ελεύθερης ενέργειας των RNA:RNA duplexes.

6) –
7) 3’ UTR
8) «τέλειο seed» ως μια τέλεια αντιστοίχιση βάσης Watson-Crick με ζεύγη 7 nt ξεκινώντας από την πρώτη ή τη δεύτερη βάση του microRNA (υπολογιζόμενη από το άκρο 5' end).
Εισαγωγές ή μεταλλάξεις στην αλληλουχία mRNA ενός τέλειου seed επιτρέπονται εφ' όσον η ελεύθερη ενέργεια της δέσμευσης, που προσδιορίζεται από πρότυπο λογισμικό πρόβλεψης δευτερογενούς δομής του RNA, δεν αυξάνεται και δεν περιέχει G:U ζεύξεις βάσεων. Αυτοί οι μεταλλαγμένοι πυρήνες-seeds ονομάζονται ατελείς πυρήνες.
Τέλεια seeds cutoff: 33%
Ατελή seeds cutoff: 66%
Η πιθανότητα να είναι ένα τέλειο seed όντως μια θέση πρόσδεσης για το miRNA είναι p (p ~0.8), ενώ για ατελείς seeds είναι 1-p.
9) PicTar2 εφαρμογή: http://pictar.mdc-berlin.de/
Ο αλγόριθμος ακολουθεί τη γενική λογική του Ahab, έναν επικυρωμένο πιθανοτικό αλγόριθμο για την αναγνώριση συνδυασμών θέσεων πρόσδεσης.
Τα αποτελέσματα του PicTar είναι διαθέσιμα στη διεύθυνση http://pictar.bio.nyu.edu (δε λειτουργεί)
Ο UCSC Genome Browser διατίθεται στη διεύθυνση http://www.genome.ucsc.edu/
Η βάση στην οποία υπάρχουν και τα αποτελέσματα απ' τον Pictar2 http://dorina.mdc-berlin.de/

3.6.3 Δομη Αλγοριθμου

Η βασική διαδικασία του αλγορίθμου

Η προδιάθεση έκφρασης των microRNAs με κλωνοποίηση και προσδιορισμό αλληλουχίας, northern blotting ή microarrays [Αναφ. 4] έδειξε ότι ένας μικρός αριθμός microRNAs (τυπικά ένας έως δέκα) εκφράζονται συχνά σε συγκεκριμένους ιστούς και αναπτυξιακά στάδια. Πολλά γνωστά γονίδια-στόχοι των microRNAs περιέχουν αρκετές θέσεις δέσμευσης microRNA και ο βαθμός της μεταφραστικής καταστολής μπορεί να αυξηθεί εκθετικά με τον αριθμό των θέσεων δέσμευσης του microRNA στην μη μεταφραζόμενη περιοχή 3 (UTR). Έτσι, όπως και στην μεταγραφική ρύθμιση, οι συγκεντρώσεις των διεκπεραιωμένων microRNAs σε ένα κύτταρο μπορούν να αναγνωσθούν από ρυθμιστικές θέσεις και να χρησιμοποιηθούν για να καθοριστεί η γονιδιακή έκφραση.

Συνεπώς, μπορεί να **είναι σημαντικό να αναζητηθούν συνδυασμοί θέσεων δέσμευσης microRNA για σύνολα συν-εκφρασμένων microRNAs.**

Οι προηγουμένως αναπτυγμένοι υπολογιστικοί αλγόριθμοι μπορούν να εντοπίσουν στόχους για μεμονωμένα microRNAs [Αναφ. 7-14] αλλά μέχρι στιγμής δεν έχουν χρησιμοποιηθεί για να επιτύχουν κοινούς στόχους διαφόρων microRNAs. Περαιτέρω, τυπικά έχουν σχετικά υψηλά ψευδώς θετικά ποσοστά όταν ο αριθμός των θέσεων δέσμευσης για ένα δεδομένο microRNA σε ένα 3 'UTR είναι μικρός.

Η μέθοδος αυτή, **η πιθανολογική ταυτοποίηση συνδυασμών θέσεων στόχων (PicTar), υπερνικά αυτά τα προβλήματα με γενίκευση προηγούμενων μεθόδων και επιτρέπει τον εντοπισμό στόχων τόσο για τα απλά microRNAs όσο και για τους συνδυασμούς των microRNAs.**

- Η είσοδος στο PicTar (Εικόνα 8a) είναι ένα σταθερό σετ αναζήτησης microRNAs και πολλαπλές ευθυγραμμίσεις των ορθολογικών αλληλουχιών νουκλεοτιδίων (3’ UTRs). Τα αποτελέσματα είναι αποτελέσματα που ταξινομούν τα γονίδια με την πιθανότητα να είναι ένας κοινός στόχος των μελών (υποσύνολα) του συνόλου αναζήτησης και πιθανότητες για τις προβλεπόμενες θέσεις δέσμευσης σε κάθε UTR. Ο χώρος αλληλουχίας κάθε είδους στην ευθυγράμμιση δίνεται από τον αντίστοιχο αριθμό νουκλεοτιδίων. Οι πολλαπλές ευθυγραμμίσεις καλύπτουν τον άνθρωπο, τον χιμπαντζή, τον ποντικό, τον αρουραίο και τον σκύλο για το 90% όλων των ανθρώπινων νουκλεοτιδίων αλληλουχίας 3 'UTR. Οι αλληλουχίες των οκτώ ειδών ευθυγραμμίζονται για το 21% όλων των ανθρώπινων 3' UTRs. Η κάλυψη για ανθρώπους, χιμπατζήδες, ποντίκια, αρουραίους, σκύλους και κοτόπουλο (55%) είναι σύμφωνος με τον εκτιμώμενο αριθμό ορθολογικών γονιδίων ανθρώπινου - κοτόπουλου [Αναφ. 24]. Για την παραγωγή των στατιστικών, δημιουργήσανε ένα σύνολο δεδομένων 3’ UTRs περιορίζοντας το ανθρώπινο 3’ UTR σε μοναδικές ακολουθίες και καλύπτοντας επαναλήψεις χρησιμοποιώντας τις μάσκες επανάληψης UCSC.
- Ο αλγόριθμος ακολουθεί τη γενική λογική του Ahab, έναν επικυρωμένο πιθανοτικό αλγόριθμο για την αναγνώριση συνδυασμών θέσεων πρόσδεσης μεταγραφών [Αναφ. 15,16]. Το PicTar αντιστοιχεί σε όλες τις κατατμήσεις μιας ακολουθίας σε θέσεις δέσμευσης και στο background και υπολογίζει τη βαθμολογία μέγιστης πιθανότητας ότι η ακολουθία δεσμεύεται από συνδυασμούς microRNAs. (Εικόνα 8b)
- Το μοντέλο αποδίδει συνεργατικές επιδράσεις πολλαπλών θέσεων σύνδεσης ενός microRNA ή διαφόρων microRNAs που δρουν μαζί, καθώς και για το κατάλληλο score επικαλυπτόμενων περιοχών. Οι πιθανότητες που αποδίδονται σε μία μοναδική τοποθεσία διαμορφώθηκαν σύμφωνα με τα πειραματικά [Αναφ. 7,8,12,17] και τα υπολογιστικά [Αναφ. 7-14] αποτελέσματα.
- Οι συγκρίσεις μεταξύ των ειδών είναι ζωτικής σημασίας για την αποτύπωση ψευδών θετικών: τα υποψήφια γονίδια-στόχοι ορίζονται ως UTRs με έναν ελάχιστο (καθορισμένο από τον χρήστη) αριθμό εξελικτικά διατηρημένων υποθετικών δεσμευτικών θέσεων. Το PicTar τότε βαθμολογεί τις υποψήφιες αλληλουχίες για κάθε είδος ξεχωριστά. Οι βαθμολογίες που προκύπτουν

συνδυάζονται για να ληφθεί η τελική βαθμολογία PicTar για ένα γονίδιο. Οι μελλοντικές γνώσεις σχετικά με την αναγνώριση της θέσης στόχου microRNA και την αποτελεσματικότητα της καταστολής μπορούν εύκολα να ενσωματωθούν στο μοντέλο. Στην Caenorhabditis elegans, η διαδοχική εξειδίκευση της έκφρασης των microRNAs lin-4 και let-7 συντονίζει την ανάπτυξη του χρονισμού[Αναφ. 18].

3.6.4 ΒΑΣΙΚΕΣ ΑΝΑΛΥΣΕΙΣ

Αλγόριθμος PicTar: αναγνώριση μοναδικών θέσεων στόχου microRNA.

Ο «πυρήνας» (ή ο «seed»), τυπικά ένα άριστα συνδεδεμένο Watson-Crick-ζευγαρωμένο τμήμα ~7 nt στο διπλό miRNA - mRNA, διαδραματίζει βασικό ρόλο τόσο στην αναγνώριση της θέσης στόχου όσο και στην καταστολή του μεταγραφικού στόχου. Ο πυρήνας βρίσκεται συνήθως στο 5' άκρο του microRNA ξεκινώντας από την πρώτη ή τη δεύτερη θέση [Αναφ. 3]. Η ελεύθερη ενέργεια της συσχετίζεται με την ικανότητα του διπλού miRNA mRNA να καταστείλει τη μετάφραση του στοχευόμενου μεταγραφήματος [Αναφ. 17]. Χρησιμοποιήσανε αυτά και άλλα πειραματικά αποτελέσματα σε πιθανότητες για μια αλληλουχία mRNA να είναι μια θέση δέσμευσης για ένα δεδομένο microRNA. Πιο συγκεκριμένα, ορίσανε έναν «τέλειο πυρήνα» ως μια τέλεια ζεύξη βάσης Watson-Crick με ζεύγη 7 nt ξεκινώντας από την πρώτη ή τη δεύτερη βάση του microRNA (υπολογιζόμενη από το 5’ άκρο). Εισαγωγές ή μεταλλάξεις στην αλληλουχία mRNA ενός τέλειου πυρήνα επιτρέπονται εφ' όσον η ελεύθερη ενέργεια της δέσμευσης, που προσδιορίζεται από πρότυπο λογισμικό πρόβλεψης δευτερογενούς δομής του RNA, δεν αυξάνεται και δεν περιέχει G:U ζεύξεις βάσεων. Αυτοί οι μεταλλαγμένοι πυρήνες ονομάζονται ατελείς πυρήνες. Σύμφωνα με προηγούμενες μελέτες [Αναφ. 7-14], απαιτούν επίσης ότι η ελεύθερη ενέργεια ολόκληρου του duplex miRNA - mRNA να είναι κάτω από μια τιμή αποκοπής. Για τις θέσεις με τέλειους πυρήνες, η τιμή αυτή ρυθμίζεται στο 33% της βέλτιστης ελεύθερης ενέργειας ολόκληρης της ώριμης δέσμευσης του microRNA σε μια τέλεια συμπληρωματική θέση στόχου. Αυτό το φίλτρο απορρίπτει, κατά μέσο όρο, μόνο το 5% όλων των τέλειων πυρήνων, αλλά αυξάνει τον λόγο σήματος προς θόρυβο. Προς το παρόν χρησιμοποιούν πολύ πιο αυστηρό φίλτρο (66% της βέλτιστης ελεύθερης ενέργειας) για τοποθεσίες με ατελείς πυρήνες για να προστατεύσουν από ψευδώς θετικά.

Ένας τέλειος πυρήνας που επιβιώνει από το φιλτράρισμα αποδίδει μια πιθανότητα ρ να είναι μια θέση πρόσδεσης για το microRNA. Η πιθανότητα για ατελείς πυρήνες είναι 1 - p διαιρούμενη από τον συνολικό αριθμό ατελών πυρήνων (συνήθως στην περιοχή 2-20). Δουλέψαν με ένα υψηλό p (p E 0.8) επειδή οι περισσότερες από τις γνωστές περιοχές στόχου δεν έχουν ατελείς πυρήνες, αλλά έλεγξαν ότι οι βαθμολογίες των UTR με βαθμολογίες PicTar δεν ήταν ευαίσθητες σε συγκεκριμένες θέσεις λογικά υψηλών τιμών p. Πιο εξελιγμένοι τρόποι για την εκχώρηση πιθανοτήτων στις θέσεις δέσμευσης microRNA θα είναι δυνατές όταν επικυρωθούν περισσότεροι στόχοι.

Αλγόριθμος PicTar: βαθμολόγηση συνδυασμών τοποθεσιών στόχων.

Το PicTar υπολογίζει μια βαθμολογία μέγιστης πιθανότητας ότι μια δεδομένη αλληλουχία RNA (τυπικά ένα 3’ UTR) στοχεύεται από ένα καθορισμένο σύνολο microRNAs. Μόλις οι πιθανότητες για κάθε αλληλουχία της αλληλουχίας RNA να είναι μια θέση δέσμευσης για ένα microRNA είναι σταθερές, η βαθμολόγηση του PicTar είναι παρόμοια με τον αλγόριθμο Ahab όπως περιγράφεται [Αναφ. 15] με τις ακόλουθες λεπτομέρειες εφαρμογής.

Πρώτον, το PicTar θέτει το μήκος των υποτιθέμενων θέσεων δέσμευσης microRNA στο μήκος των αντίστοιχων πυρήνων. Αυτό συλλαμβάνει το πειραματικό αποτέλεσμα ότι οι αλληλεπικαλυπτόμενες θέσεις σύνδεσης φαίνεται να δρουν ανεξάρτητα εφόσον οι πυρήνες τους δεν αλληλεπικαλύπτονται [Αναφ. 17]. **Δεύτερον**, ένα σύντομο 3’ UTR (<300 bp) δεν μπορεί να χρησιμοποιηθεί για αξιόπιστη

εκτίμηση των δικών του νουκλεοτιδικών συχνοτήτων. Σε αυτές τις περιπτώσεις, λαμβάνεται ο γραμμικός συνδυασμός των υπό-μονάδων νουκλεοτιδίων υποβάθρου - background που υπολογίζονται από την UTR και τις συχνότητες υποβάθρου που υπολογίζονται από όλα τα UTR για το ίδιο είδος στο σύνολο δεδομένων. **Τρίτον**, χρησιμοποιούν τον αλγόριθμο Baum-Welch [Αναφ. 25] για να υπολογίσουν τις μέγιστες πιθανότητες. Η σύγκλιση του λογαρίθμου του διαχωριστικού αθροίσματος ελέγχεται μέχρι ακρίβειας 0,0005. **Τέταρτον**, χρησιμοποιούν τη βελτιστοποιημένη προηγούμενη για background όταν υπολογίζουν το ποσό διαίρεσης μόνο για background. **Πέμπτον**, η σειρά του μοντέλου για την αλληλουχία background ορίζεται στο 0.

Αλγόριθμος PicTar: εκτελέσεις PicTar σε επίπεδο γονιδιώματος και συγκρίσεις μεταξύ ειδών.
Καταρχήν, προ-ρυθμίστηκαν οι θέσεις όλων των πιθανών πυρήνων microRNA σε όλες τις αλληλουχίες UTR με το πρόγραμμα nuclMap. Ελέγξαν αν πυρήνες για το ίδιο microRNA πέφτουν σε αλληλεπικαλυπτόμενες θέσεις ευθυγράμμισης για όλα τα είδη κάτω από ένα cutoff.

Εάν οι πυρήνες συντηρηθούν με αυτά τα κριτήρια, ελέγξαν εάν η βέλτιστη ελεύθερη ενέργεια των προβλεπόμενων miRNA mRNA πέρασε τα κριτήρια φιλτραρίσματος. Τέλειοι πυρήνες που επιβίωσαν αυτά τα βήματα ονομάζονται άγκυρες.

Ο αριθμός των αγκυρών σε ένα UTR καθορίζει εάν θα γίνει βαθμολογία από το PicTar. Εάν ναι, η βέλτιστη ελεύθερη ενέργεια όλων των θέσεων με τέλειους ή ατελείς πυρήνες σε κάθε ακολουθία UTR χρησιμοποιείται για τον εντοπισμό απίθανων θέσεων στόχων. Οι υπόλοιποι τόποι για κάθε UTR εισάγονται στο PicTar για να υπολογίσουν μια βαθμολογία για κάθε UTR στην πολλαπλή ευθυγράμμιση. Για να ληφθεί μια τελική βαθμολογία που αντανακλά την πιθανότητα να ρυθμιστεί το UTR από το δεδομένο σύνολο microRNAs, υπολογίσανε κατά μέσον όρο τις βαθμολογίες για όλα τα είδη που χρησιμοποιήθηκαν για τις θέσεις πρόσδεσης. Ο μέσος όρος αυτός θα πρέπει να αντανακλά τις διαφορετικές εξελικτικές αποστάσεις μεταξύ των ειδών. Καταμετρήσανε κατά μέσον όρο τα αποτελέσματα των ανθρώπων και των χιμπατζήδων και τα αποτελέσματα των ποντικών και των αρουραίων ανεξάρτητα για να λάβουν ένα σκορ πρωτεύοντος και ένα σκορ του τρωκτικού. Αυτές οι βαθμολογίες στη συνέχεια υπολογίστηκαν κατά μέσο όρο με το σκορ σκύλου για να ληφθεί μια βαθμολογία που αντικατοπτρίζει τη διατήρηση σε όλα τα θηλαστικά. Παρομοίως, καταμετρήσανε κατά μέσο όρο αυτό το σκορ θηλαστικού, το σκορ κοτόπουλου και τις μέσες βαθμολογίες fish για ένα συνολικό σκορ, ανάλογα με την περίπτωση. Η εκτέλεση μιας ολόκληρης ανάλυσης σε έναν τυποποιημένο υπολογιστή με 2 GB μνήμης διαρκούσε 15 λεπτά κατά την αναζήτηση στόχων από ένα έως έξι microRNAs.

Βέλτιστες εκτιμήσεις της ελεύθερης ενέργειας των RNA: RNA duplexes.

Υπολογίσανε τις ελεύθερες ενέργειες των διπλών RNA: RNA χρησιμοποιώντας RNAhybrid14 με επιλογές -s3utr_human για ακολουθίες σπονδυλωτών, -s3utr_worm για τις ακολουθίες νηματωδών και τις προεπιλεγμένες ρυθμίσεις αλλιώς.

Σύνολα δεδομένων γνωστών και τυχαίων ώριμων αλληλουχιών microRNA.
Παρήγαγαν ώριμες αλληλουχίες microRNA από Rfam26 (Release 5.0) και προσθέσανε εννέα microRNAs. Εκπονήσανε ένα υποσύνολο microRNA που διατηρήθηκε μεταξύ ανθρώπου, χιμπαντζή, ποντικού, αρουραίου, σκύλου και κοτόπουλου χρησιμοποιώντας σημειώσεις Rfam ώριμων microRNAs σπονδυλωτών ομόλογων σε ανθρώπινο microRNA. Όποτε δεν υπήρχε σχολιασμός, χρησιμοποιήσανε αυστηρά κριτήρια για να ελέγξουν τη διατήρηση. Κατασκευάσανε ένα σύνολο μοναδικών microRNAs συγκεντρώνοντας μαζί τα microRNAs με πανομοιότυπες βάσεις στις θέσεις 1-7 ή 2-8 (ξεκινώντας από το 5' end). Λάβανε 58 μοναδικά microRNA που διατηρούνται σε ανθρώπους, χιμπατζήδες, ποντικούς,

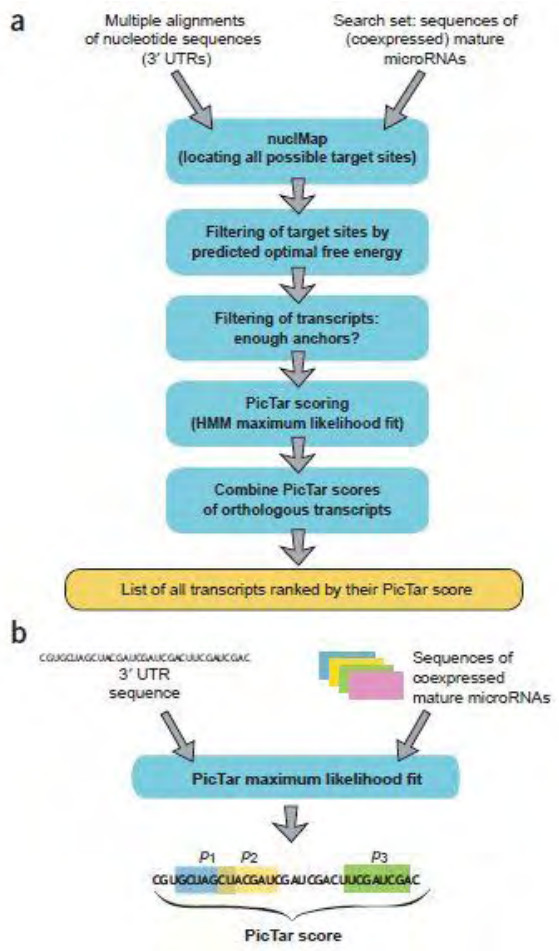
αρουραίους, σκύλους και κοτόπουλο. Παρόμοια με μια προηγούμενη μέθοδο [Αναφ. 11], δημιουργήσανε ομάδες μοναδικών τυχαιοποιημένων microRNAs με εκχύλιση 8-μερών με περίπου την ίδια αφθονία (+-15%) του 7μερούς που αρχίζει στις θέσεις 1 και 2 και το αντίστοιχο 7-μερές του θεωρούμενου microRNA σε όλα τα ανθρώπινα 3‘ UTRs. Ο πειραματισμός με πολλά άλλα συστήματα τυχαιοποίησης οδήγησε σε συγκρίσιμες αναλογίες σήματος προς θόρυβο. Συνδέσανε το άκρο 3 του κάθε microRNA στο αντίστοιχο τυχαίο 8-μερές.

Δοκιμασία δραστηριότητας λουσιφεράσης.

Εξαιρέσανε την άγριου τύπου μυοτροπίνη 3' UTR από τον κλώνο IMAGE 6839739 και τον υποκλωνοποιήσανε προς τα κάτω του κωδικονίου τερματισμού στο pRL-TK (Promega). Επιμολύνανε κύτταρα N2A με 0,1 mg φορέα αναφοράς pRL-TK που κωδικοποιεί Rr-luc και 0,1 mg φορέα ελέγχου pGL3 που κωδικοποιεί Pp-luc (Promega). Διαμολύνανε κύτταρα με 200 ng siRNAs. Σε περιπτώσεις όπου οι ομάδες συγκρίθηκαν με μεμονωμένα siRNAs, η διαφορά δημιουργήθηκε χρησιμοποιώντας si-GFP. Συλλέξαν και ανιχνεύσαν τα κύτταρα 30-36 ώρες μετά τη διαμόλυνση.

SiRNAs.

Συνθετικά microRNAs και siRNA συντέθηκαν από την Dharmacon Research, Inc. Επιμολύνθηκαν κύτταρα N2A με φορείς και siRNA χρησιμοποιώντας Lipofectamine 2000 (Invitrogen) σύμφωνα με τις οδηγίες του κατασκευαστή.



Εικόνα 8 PicTar2 αλγόριθμος

από τη δημοσίευση: Combinatorial microRNA target predictions, Azra Krek, Dominic Grun, Matthew N Poy, Rachel Wolf, Lauren Rosenberg, Eric J Epstein, Philip MacMenamin, Isabelle da Piedade, Kristin C Gunsalus, Markus Stoffel & Nikolaus Rajewsky [Αναφ. 47].

Στην εικόνα 8 παρουσιάζεται η δομή του αλγορίθμου συνοπτικά,

- (a) Η είσοδος στο PicTar αποτελείται από πολλαπλές ευθυγραμμίσεις των αλληλουχιών RNA (τυπικά 3 'UTRs) και ένα σύνολο αναζήτησης ώριμων (συνεκφραζόμενων) αλληλουχιών microRNA. Το πρόγραμμα nuclMap εντοπίζει όλους τους τέλειους πυρήνες (μήκος 7, ξεκινώντας από τη θέση 1 ή 2 του άκρου 5 'του microRNA) και ατελείς πυρήνες σε ακολουθίες 3' UTR.
- Πυρήνες που επιβιώνουν από τον βέλτιστο ελεύθερο ενεργειακό φίλτρο και πέφτουν σε αλληλεπικαλυπτόμενες θέσεις στις ευθυγραμμίσεις για όλα τα υπό εξέταση είδη καλούνται άγκυρες.
- Εάν μια πολλαπλή ευθυγράμμιση 3 'UTR έχει έναν ελάχιστο (προσδιορισμένο από τον χρήστη) αριθμό αγκυρών, κάθε UTR στην ευθυγράμμιση θα βαθμολογηθεί με την κεντρική PicTar διαδικασία μέγιστης πιθανότητας (b).
- Οι βαθμολογίες για μεμονωμένα UTR σε μια ευθυγράμμιση συνδυάζονται για να ληφθεί η τελική βαθμολογία PicTar, η οποία μπορεί να χρησιμοποιηθεί για την απόκτηση μιας κατάταξης λίστας όλων των συνόλων ορθολογικών μεταγραφών.
- (b) Βαθμολόγηση PicTar μιας μοναδικής ακολουθίας 3 'UTR. Το PicTar αντιστοιχεί σε όλες τις κατατμήσεις της αλληλουχίας RNA (3 'UTR) σε θέσεις δέσμευσης και αλληλουχίες υποβάθρου. Το PicTar υπολογίζει τη βαθμολογία μέγιστης πιθανότητας (βαθμολογία PicTar) ότι η αλληλουχία RNA στοχεύεται από συνδυασμούς μικροRNAs από το σύνολο αναζήτησης σε σύγκριση με το υπόβαθρο και την ατομική πιθανότητα pi για κάθε υποαλληλουχία της αλληλουχίας RNA που δεσμεύεται από ένα microRNA (μόνο οι πυρήνες για τις θέσεις δέσμευσης που απεικονίζονται). Αυτές οι posterior πιθανότητες είναι διαφορετικές από την πιθανότητα ότι μια μόνο υπο-αλληλουχία είναι μια θέση πρόσδεσης του microRNA.

4. ΚΕΦΑΛΑΙΟ – ΠΕΙΡΑΜΑΤΑ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 ΑΞΙΟΛΟΓΗΣΗ MIRTARGET2

Ένας αλγόριθμος πρόβλεψης στόχου miRNA με βάση την τεχνολογία SVM αναπτύχθηκε με την ενσωμάτωση 131 ετερογενών χαρακτηριστικών πρόβλεψης που περιεγράφηκαν παραπάνω. Συνολικά, αναλύθηκαν 454 downregulated γονίδια (θετικά δείγματα) και 1017 φυσιολογικά γονίδια (αρνητικά δείγματα). Το πακέτο LIBSVM χρησιμοποιήθηκε για την κατασκευή ταξινομητών στόχων miRNA. Οι παράμετροι του μοντέλου εκπαίδευσης βελτιστοποιήθηκαν με πολλαπλούς κύκλους διασταυρούμενης επικύρωσης για να ελαχιστοποιηθεί ο κίνδυνος υπερβολικής προπόνησης.

Η συσχέτιση με τις αληθινές ετικέτες κλάσης αξιολογήθηκε χρησιμοποιώντας τη συσχέτιση του Spearman (Rs).

Χρησιμοποιήθηκαν καμπύλες λειτουργικών χαρακτηριστικών δέκτη (ROC) για την αξιολόγηση της ευαισθησίας πρόβλεψης και της εξειδίκευσης. Η Εικόνα 9 δείχνει 3 διαφορετικά μοντέλα SVMs:

I. Συμπεριλαμβάνει όλα τα χαρακτηριστικά και εξάγει ένα 10-πλάσιο αποτέλεσμα πρόβλεψης διασταυρούμενης επικύρωσης, με περιοχή ROC 0,79. Η σταδιακή παλινδρόμηση πραγματοποιήθηκε για

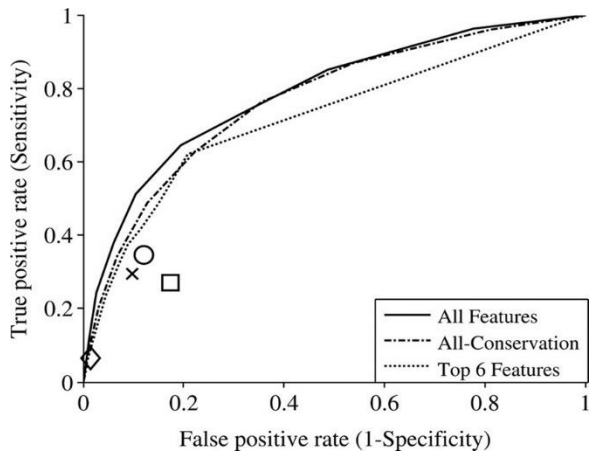
να εντοπιστούν τα χαρακτηριστικά που συνέβαλαν περισσότερο στο μοντέλο πρόβλεψης. Οι τιμές P εκτιμήθηκαν με στατιστικές Wald. Το χαρακτηριστικό συντήρησης των seeds σε διασταυρούμενα είδη κατατάχθηκε στην κορυφή της λίστας με τον μεγαλύτερο συντελεστή συσχέτισης (Rs = 0.28).

Precision = το κλάσμα των αληθινών θετικών (true positive) μεταξύ όλων των προβλεπόμενων θετικών

Recall = το κλάσμα των αληθινών θετικών (true positive) μεταξύ όλων των θετικών δειγμάτων

II. Συμπεριλαμβάνει όλα τα χαρακτηριστικά εκτός του χαρακτηριστικού της συντηρησιμότητας. Ως αποτέλεσμα, η περιοχή ROC μειώθηκε σε 0,77. Η αλγοριθμική βελτίωση από το χαρακτηριστικό συντήρησης ήταν πιο εμφανής όταν ο ψευδώς θετικός (false positive) ρυθμός ήταν σχετικά χαμηλός.

III. Ένα εναλλακτικό μοντέλο SVM χτίστηκε επίσης με βάση τα έξι κορυφαία χαρακτηριστικά, οδηγώντας σε μειωμένη περιοχή ROC 0,72.



Εικόνα 9 ROC καμπύλες και σύγκριση με άλλους αλγορίθμους

από τη δημοσίευση: Prediction of both conserved and nonconserved microRNA targets in animals, Xiaowei Wang and Issam M. El Naqa [Αναφ. 8].

Επίσης, έγινε σύγκριση της απόδοσης του μοντέλου πρόβλεψης με άλλους υπάρχοντες αλγορίθμους πρόβλεψης, TargetScan, PicTar, miRanda, mirTarget1. Όλοι αυτοί οι αλγόριθμοι έχουν τιμές αποκοπής (cutoffs) για την εκχώρηση βαθμολογίας πρόβλεψης και πάνω από το 90% όλων των ανθρώπινων γονιδίων δεν έλαβαν κάποιο score. Ως αποτέλεσμα, οι καμπύλες ROC δεν μπορούν να κατασκευαστούν πλήρως για αυτούς τους αλγορίθμους. Αντ' αυτού, η συνολική απόδοση πρόβλεψης αυτών των αλγορίθμων θα μπορούσε να αναπαρασταθεί μόνο από επιλεγμένα σημεία αποκοπής στο διάγραμμα ROC.

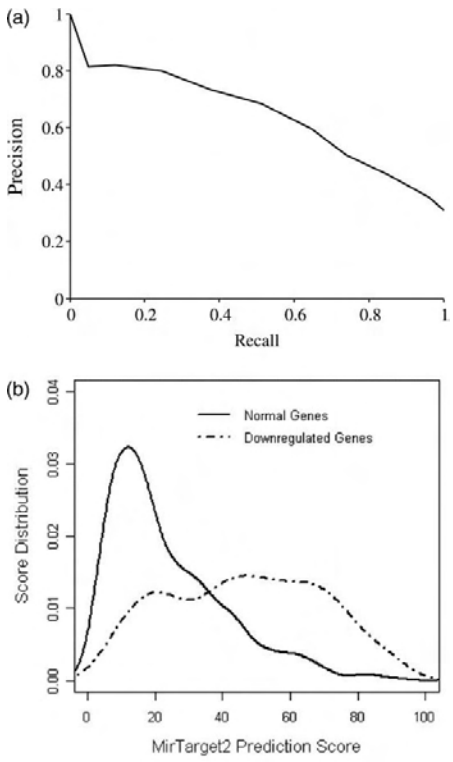
Όπως φαίνεται στην Εικόνα 9, τα μοντέλα SVMs είχαν πιο ισχυρή απόδοση από τους δημοσιευμένους αλγόριθμους στα επιλεγμένα σημεία αποκοπής.

Το mirTarget1 παρουσίασε μια πολύ μικρή ψευδώς θετική τιμή (0,01) με ευαισθησία 0,07 όταν συγκρίθηκε με τα μοντέλα SVMs, εν μέρει λόγω της πολύ συντηρητικής στρατηγικής του να προβλέψει

μόνο τις πιο έμπιστες υποψήφιες θέσεις (296 προβλεπόμενοι στόχοι σε σχέση με πάνω από 1.500 από άλλα δημοσιευμένα προγράμματα).

Οι καμπύλες precision - recall κατασκευάστηκαν για να αξιολογήσουν την ακρίβεια της πρόβλεψης. Οι καμπύλες ακριβείας - ανάκτησης χρησιμοποιούνται συνήθως στη μηχανική μάθηση για την αξιολόγηση της ακρίβειας της πρόβλεψης.

Όπως φαίνεται στο σχήμα 10a, ο ρυθμός ακριβείας ήταν <80% όταν ο ρυθμός ανάκλησης ήταν κάτω από 20%.



Εικόνα 10 Precision – Recall καμπύλες

από τη δημοσίευση: Prediction of both conserved and nonconserved microRNA targets in animals, Xiaowei Wang and Issam M. El Naqa [Αναφ. 8].

Δημιουργήθηκε ένα σύστημα βαθμολόγησης για τη βαθμολόγηση όλων των 3' UTR περιοχών με τοποθεσίες που αντιστοιχούν σε seeds. Ένα μικρό μέρος αυτών των περιοχών είχε πολλαπλές υποψήφιες τοποθεσίες και όλες οι θέσεις σε ένα UTR συνδυάστηκαν για να υπολογίσουν την βαθμολογία ως εξής:

$$S = 100 \times (1 - \prod_{i=1}^{n} P_i)$$

Όπου,

- n αντιπροσωπεύει τον αριθμό των υποψήφιων θέσεων στόχων σε ένα UTR
- Pi αντιπροσωπεύει την στατιστική σημασία P-τιμής για κάθε μία από αυτές τις υποψήφιες θέσεις όπως υπολογίζεται από την SVM.

Οι βαθμολογίες για UTRs μιας τοποθεσίας υπολογίστηκαν χρησιμοποιώντας την ίδια εξίσωση με n=1. Αυτές οι βαθμολογίες χρησιμοποιήθηκαν για την ανάθεση τάξεων για την αξιολόγηση της σχετικής σημασίας των προβλεπόμενων στόχων.

Οι κατανομές των βαθμολογιών για τα downregulated και τα κανονικά γονίδια συγκρίθηκαν για να αξιολογηθεί η προβλεπτική ισχύς του συστήματος βαθμολόγησης (Εικόνα 12b). Η κατανομή των βαθμολογιών για τα φυσιολογικά γονίδια διαχωρίστηκε σημαντικά από εκείνη των downregulated γονιδίων. Κατά μέσο όρο, οι βαθμολογίες για τις υποψήφιες θέσεις από τα φυσιολογικά γονίδια ήταν πολύ χαμηλότερες, με μια μέγιστη κορυφή κάτω από 20. Οι περιοχές με downregulated γονίδια είχαν μια πιο κατανεμημένη κατανομή σκορ, αντανακλώντας το γεγονός ότι το 50% των θέσεων είχε βαθμολογίες υψηλότερες από 50.

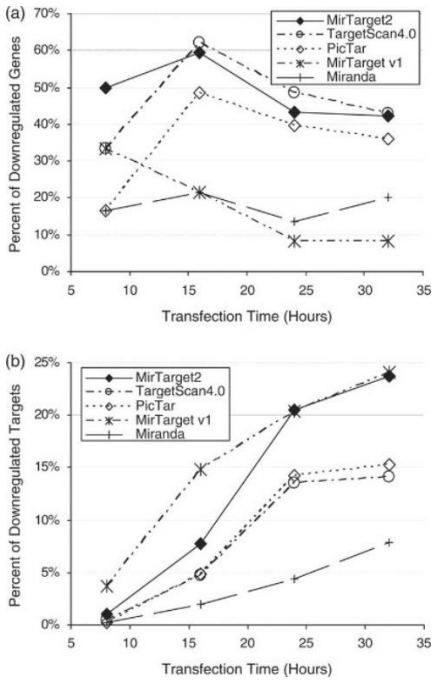
Ο αλγόριθμος πρόβλεψης υλοποιήθηκε ως πρόγραμμα Perl. Η απόδοση του mirTarget2 αξιολογήθηκε με ένα ανεξάρτητο σύνολο δεδομένων. Σε αυτό το πείραμα, το miR-124a υπέρ - εκφράστηκε σε κύτταρα HepG2 και οι μεταβολές στα προφίλ ολικής έκφρασης αξιολογήθηκαν με microarrays σε πολλαπλά χρονικά σημεία. Το miR-124a δεν είχε συμπεριληφθεί στη διαδικασία εκπαίδευσης SVM. Τα downregulated γονίδια ταυτοποιήθηκαν στις 8, 16, 24 και 32 ώρες μετά την υπέρ-έκφραση του miR-124a.

Το mirTarget2 συγκρίθηκε με άλλους τέσσερις αλγόριθμους, το TargetScan, το PicTar, το mirTarget1 και το miRanda για την πρόβλεψη στόχου miR-124a. Τα γονίδια 283, 482, 372, 54 και 408 προέβλεπαν τους στόχους miR-124a από mirTarget2, TargetScan, PicTar, mirTarget1 και miRanda αντίστοιχα. Δεν θα μπορούσαν να γίνουν καθόλου πλήρεις καμπύλες ROC για οποιονδήποτε από αυτούς τους αλγορίθμους (συμπεριλαμβανομένου του mirTarget2), επειδή αποδίδουν βαθμολογίες πρόβλεψης στο ~ 10% όλων των ανθρώπινων γονιδίων.

Στην ανάλυση αυτή, η απόδοση του αλγορίθμου συγκρίθηκε με την αξιολόγηση της γενικής εξειδίκευσης πρόβλεψης και της ευαισθησίας για downregulated γονίδια miRNA σε πολλαπλά χρονικά σημεία. Αρχικά υπολογίστηκε ο αριθμός των γονιδίων που είχαν ρυθμιστεί με ακρίβεια και είχαν προβλεφθεί επίσης στόχοι για το miR-124a. Συνολικά, το mirTarget2 και το TargetScan προέβλεπαν παρόμοια υψηλότερα ποσοστά γονιδίων (Εικόνα 11a).

Τα ποσοστά των προβλεπόμενων downregulated στόχων από το miR-124a καθορίστηκαν επίσης. Όπως φαίνεται στο Σχήμα 11b, τα ποσοστά για όλους τους αλγόριθμους ήταν σχετικά χαμηλά στα αρχικά χρονικά σημεία. Αυτό υποδηλώνει ότι η πρόβλεψη του downregulated στόχου από miRNA γενικά μπορεί να μην είναι μια ταχεία διαδικασία.

Μεταξύ αυτών των αλγορίθμων, το mirTarget1 ήταν πιο επιλεκτικό στην πρόβλεψη των downregulated στόχων γονιδίου. Το mirTarget2 εκτελέστηκε παρόμοια με το mirTarget1, ειδικά στις 24 και 32 ώρες. Και οι δύο αλγόριθμοι ήταν πιο επιλεκτικοί στην ταυτοποίηση των downregulated γονιδίων από τους άλλους αλγορίθμους που περιλαμβάνονται στην ανάλυση.



Εικόνα 11 Σύγκριση εκδόσεων MirTarget2

από τη δημοσίευση: Prediction of both conserved and nonconserved microRNA targets in animals, Xiaowei Wang and Issam M. El Naqa [Αναφ. 8].

4.2 ΑΞΙΟΛΟΓΗΣΗ MIRMAP

- **Συνδυασμός χαρακτηριστικών πρόβλεψης**

Τα χαρακτηριστικά συσχετίζονται γραμμικά με τα πειραματικά μετρούμενα επίπεδα καταστολής του miRNA. Συνδυάσανε 10 χαρακτηριστικά της βιβλιοθήκης miRmap (εκτός του «ΔG total», καθώς αυτό το χαρακτηριστικό είναι απλώς το άθροισμα των «ΔG duplex» και «ΔG open») με πολλαπλή γραμμική παλινδρόμηση στο 'Trans. Grimson' σύνολο. Αυτό το μοντέλο αποτελεί το 12,7% της διακύμανσης, κοντά σε μια διπλάσια αύξηση σε σχέση με το σκορ TargetScan, με τον ίδιο τύπο παλινδρόμησης, τα τρία χαρακτηριστικά του σκορ TargetScan ('AU content', '3-pairing ' «UTR position») αποτελούν μόνο το 7,49% της διακύμανσης.

Αυτή η βελτιωμένη απόδοση του μοντέλου επιβεβαιώνεται από τις υψηλότερες συσχετίσεις με τις πειραματικές μετρήσεις, υπολογιζόμενες με τον ίδιο τρόπο όπως οι συσχετίσεις μεμονωμένων χαρακτηριστικών. Η συνεισφορά κάθε χαρακτηριστικού (στο σύνολο δεδομένων του 'Trans.Grimson', Εικόνα 12Β) αντικατοπτρίζει γενικά τις ταξινομήσεις με βάση τις συσχετίσεις μεμονωμένων χαρακτηριστικών: Το «AU content» είναι το πιο επεξηγηματικό χαρακτηριστικό, αλλά το «P.over exact» συμβάλλει περισσότερο στο μοντέλο παλινδρόμησης. Είναι ενδιαφέρον ότι τα χαρακτηριστικά συντήρησης «PhyloP» και «BLS» συμβάλλουν το 14 και το 11% αντίστοιχα, παρά τη χρήση των ίδιων δεδομένων εισόδου (ευθυγράμμιση αλληλουχιών πολλαπλών γονιδιωμάτων) και τα δύο συμβάλλουν ουσιαστικά στη διακύμανση. Μεταξύ των χαρακτηριστικών προσβασιμότητας, το «ΔG open» συμβάλλει μόνο κατά το ήμισυ σε σχέση με το «AU content» (15 και 30%, αντίστοιχα). Από τα πέντε χαρακτηριστικά με τη μεγαλύτερη συνεισφορά στο μοντέλο με όλα τα χαρακτηριστικά (αντιπροσωπεύουν το 90,5% του πλήρους μοντέλου) εξακολουθεί να αντιστοιχεί στο 11,6% της διακύμανσης.

Αντί για την αξιολόγηση του μοντέλου απευθείας από την άποψη της επεξηγούμενης διακύμανσης, η ποιότητα της κατάταξης μπορεί να εκτιμηθεί με την ταξινόμηση των θέσεων στόχων με την

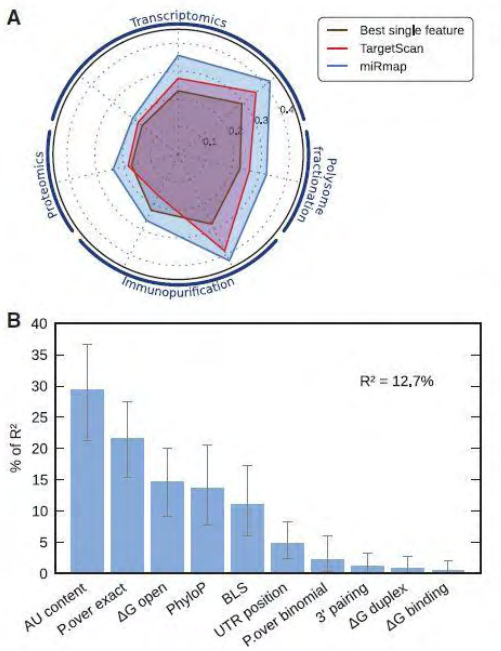
προβλεπόμενη δύναμη, τη συσσώρευση τους και τον υπολογισμό της μέσης έκφρασης fold change κάθε bin. Αυτή η προσέγγιση, που χρησιμοποιήθηκε επίσης για την αξιολόγηση της κατάταξης διαφορετικών εργαλείων για την πρόβλεψη της δύναμης της καταστολής του miRNA στη μετάφραση με δεδομένα πρωτεϊνωμάτωσης, εφαρμόστηκε σε 10 ποσοτικά μεγέθη των ταξινομημένων προβλέψεων (Συμπληρωματικό Εικόνα S2). Η συνολική κατανομή μετατοπίστηκε σε χαμηλότερες αναδιπλώσεις για το miRmap σε σύγκριση με το σκορ TargetScan, υποδεικνύοντας μια καλύτερη κατάταξη, καθώς η μείωση fold change αντιστοιχεί σε μεγαλύτερη καταστολή. Για το πρώτο ποσό, η μέση αναδιπλούμενη αλλαγή μειώθηκε από το 0,32 έως το 0,39 με το miRmap.

Οι πολλαπλές γραμμικές παλινδρομήσεις με τα άλλα σύνολα δεδομένων υποστηρίζουν περαιτέρω τα συμπεράσματα από τις αναλύσεις της επίδοσης των μεμονωμένων χαρακτηριστικών.

Επιβεβαίωσαν (i) τη σημασία του «PhyloP» για το σύνολο δεδομένων IPcross.Chi (64% του R2) σε 24% για το «AU content», (ii) την παρόμοια σημασία των «PhyloP» και «AU content» για πρωτεϊνικά σύνολα (31%, 39% αντίστοιχα) (iii) η συνάφεια του πειράματος κλασματοποίησης πολυσωμάτων (σύνολο δεδομένων RibN.Hendrickson) για τη μέτρηση της δύναμης καταστολής του miRNA σε σύγκριση με την πρωτεϊνωματική, καθώς το 10,6% της διακύμανσης εξηγείται από το μοντέλο (5,75% για την πρωτεϊνωματική).

Παρατηρήσανε επίσης ότι το μοντέλο που υπολογίζεται στο σύνολο δεδομένων Trans.Linsley εξηγεί μόνο το 4,36% της διακύμανσης, αν και αυτό το σύνολο δεδομένων είναι μεγαλύτερο και βασίζεται στις ίδιες τεχνικές με το σύνολο δεδομένων Trans.Grimson (R^2 = 12,7%). Τα μικρότερα seeds μπορούν επίσης να προωθήσουν την καταστολή του miRNA, αλλά συνήθως με χαμηλότερη απόδοση. Ως εκ τούτου, δοκιμάσανε την προσέγγισή σε κανονικά 6mer seeds με υπολογισμό ενός μοντέλου με αυτά τα matches των seeds στο σύνολο δεδομένων «Trans.Grimson». Ενώ η γενική σημασία κάθε χαρακτηριστικού παρέμεινε γενικά παρόμοια, με τα χαρακτηριστικά προσβασιμότητας να είναι τα πιο επεξηγηματικά, το R^2 έπεσε στο 8,31% της διακύμανσης (Συμπληρωματικό Εικόνα S4A), το οποίο εξακολουθεί να υπερβαίνει το σκορ TargetScan (R^2 = 4,70%). Είναι ενδιαφέρον ότι η σημασία του πιθανολογικού χαρακτηριστικού «P.over exact» μειώθηκε από 22 σε 7% - από τη δεύτερη θέση στη πέμπτη - όπως αναμενόταν με βραχύτερα seeds, όπου τα matches συμβαίνουν συχνότερα τυχαία και ως εκ τούτου είναι λιγότερο στατιστικά διακρινόμενα από το background .

Αξιολογήσανε επίσης το μοντέλο με υπολογισμό της κατανομής των fold changes. Όπως αναμενόταν, οι μέσες αναδιπλώσεις δεν ήταν τόσο χαμηλές όσο με τα 7mer seeds, ωστόσο επιβεβαίωσαν την καλύτερη κατάταξη που επιτεύχθηκε με το miRmap σε σύγκριση με το σκορ TargetScan, π.χ. Η μέση αναδιπλούμενη αλλαγή της πρώτης ποσότητας μειώθηκε από -0.16 σε -0.21. Αυτά τα αποτελέσματα υποστηρίχθηκαν περαιτέρω από την ανάλυση των άλλων συνόλων δεδομένων.



Εικόνα 12 (Α) Σύγκριση απόδοσης miRmap (Β) Σημαντικότητα χαρακτηριστικών

από τη δημοσίευση miRmap: Comprehensive prediction of microRNA target repression strength, Charles E. Vejnar and Evgeny M. Zdobnov [Αναφ. 19].

(Α) Συγκριτική απόδοση του χαρακτηριστικού με τις καλύτερες επιδόσεις (καφέ), του TargetScan (κόκκινο) και του miRmap (μπλε). (Β) Παρουσιάζει σχετική σημασία στο μοντέλο miRmap πολλαπλής γραμμικής παλινδρόμησης που προβλέπει την ισχύ καταστολής του miRNA. Το R^2 είναι το ποσοστό διακύμανσης που εξηγείται από το μοντέλο. Το περιεχόμενο 'AU' είναι η πιο επεξηγηματική μεταβλητή με το 29% του R^2.

Συνδυασμός πολλαπλών θέσεων στόχων

Κάθε mRNA μπορεί να περιέχει πολλές θέσεις στόχου miRNA. Αν και τα περισσότερα πειραματικά σύνολα δεδομένων επικεντρώνονται σε ένα απλό miRNA κάθε φορά (ή όλα τα miRNAs για το σύνολο δεδομένων IPcross.Chi), ένα πλαίσιο που μπορεί να συλλάβει την πολλαπλότητα αυτών των αλληλεπιδράσεων θα πρέπει να βελτιώσει την προβλεπτική ισχύ. Εξετάσανε τρεις απλές λειτουργίες για να συνδυάσουνε τα μεμονωμένα αποτελέσματα των τοποθεσιών στόχων σε μια γενική μετρική στο επίπεδο mRNA: το καλύτερο (το ελάχιστο ή το μέγιστο ανάλογα με το σήμα της συσχέτισης), το άθροισμα και το log του αθροίσματος των exponentials. Για αυτή την ανάλυση, επιλέξανε μετάγραφα από το σύνολο δεδομένων Trans.Grimson με ακριβώς δύο θέσεις στόχους, προκύπτοντας ένα δείγμα μεγέθους 370 mRNAs (μόνο τα 53 mRNAs έχουν ακριβώς τρεις θέσεις στόχους). Για αυτή τη μελέτη, μόνο τα χαρακτηριστικά που προβλέπουν διαφορετικές δυνάμεις για κάθε θέση στόχου σε ένα 3-UTR είναι κατάλληλα, καθώς θα έδειχναν διαφορετικούς συσχετισμούς για κάθε λειτουργία, επιτρέποντας έτσι τη σύγκριση λειτουργιών. Δεδομένου ότι τα πιθανοτικά χαρακτηριστικά υπολογίζουν την πιθανότητα ενός σταθερού αριθμού αντιστοιχιών seeds στο 3-UTR και καθώς η βαθμολογία «BLS» υπολογίζεται επίσης για ολόκληρο το 3-UTR, δεν θα μπορούσαν να χρησιμοποιηθούν.

4.3 ΑΞΙΟΛΟΓΗΣΗ TARGETSPY

- **Αξιολόγηση απόδοσης πρόβλεψης**

Για την αξιολόγηση της ποιότητας του ταξινομητή, χρησιμοποιήσανε τα ακόλουθα μέτρα απόδοσης: sensitivity, specificity και accuracy. Όπως σε κάθε διαδικασία ταξινόμησης, πρέπει να ληφθούν υπόψη τέσσερις διαφορετικές δυνατότητες: αληθή θετικά (TP), αληθή αρνητικά (TN), ψευδώς θετικά (FP) και ψευδή αρνητικά (FN). Για την αξιολόγηση του σετ εκπαίδευσης , αποκτήσανε αυτές τις τιμές με τη μορφή ενός confusion matrix, εκτελώντας μια τυπική 10-fold cross validation ακολουθούμενη από την απεικόνιση μιας καμπύλης (ROC). Από αυτό υπολογίσανε την περιοχή κάτω από τα στατιστικά στοιχεία καμπύλης (AUC), ένα μέτρο που εννοείται ως η πιθανότητα ο ταξινομητής να εκχωρήσει σε μια θετική περίπτωση υψηλότερη βαθμολογία από μια αρνητική περίπτωση κατά την επιλογή τυχαίας περίπτωσης από κάθε κλάση. Δεδομένου του confusion matrix, η ευαισθησία και η εξειδίκευση καθορίζονται από τις ακόλουθες εξισώσεις:

$$sensitivity = \frac{TP}{TP+FN}$$

$$specificity = \frac{TN}{FP+TN}$$

Η αξιολόγηση των δεδομένων pSILAC πραγματοποιήθηκε όπως στο Selbach 2008. Το μέτρο απόδοσης (accuracy) ορίστηκε ως το κλάσμα των προβλεπόμενων στόχων mRNA με μειωμένη παραγωγή πρωτεΐνης (log_2 fold change < -0.1), που ταιριάζει με τον ορισμό της θετικής προβλεπόμενης τιμής (PPV – Positive Predicted Value).

- **Αξιολόγηση σε πειραματικά επαληθευμένα δεδομένα**

Καθορίσανε **(i)** την αξιολόγηση της ποιότητας του συνόλου εκπαίδευσης που εφαρμόστηκε στο targetSpy σε πειραματικά επαληθευμένα δεδομένα και **(ii)** τη συγκριτική αξιολόγηση του targetSpy έναντι των μεθόδων που χρησιμοποιούνται συνήθως.

Το **(ii)** είναι δύσκολο καθώς οι δημοσιευμένες μέθοδοι βασίζονται σε διαφορετικές αρχές, πράγμα που καθιστά δύσκολο να τις συγκρίνουμε με δίκαιο τρόπο.

Πρόσφατα, η επίδραση της υπέρ-έκφρασης του microRNA και του knockdown αναλύθηκε σε πρωτεϊνικές μελέτες μεγάλης κλίμακας [16,18] που δεν έπασχαν από μεροληψία επιλογής που υπάρχει σε άλλα data sets. Ένα ακόμη πλεονέκτημα αυτού του συνόλου δεδομένων είναι ότι καμία από τις προσεγγίσεις πρόβλεψης δεν εκπαιδεύτηκε σε αυτά τα δεδομένα. Πέραν του μικρού αριθμού των μετρηθέντων microRNAs (πέντε) και του γεγονότος ότι η ακριβής θέση της θέσης στόχου στο αντίστοιχο μετάγραφο δεν προσδιορίζεται στο pSILAC, μέχρις ότου δεδομένα αλληλουχίας υψηλής ποιότητας όπως το HITSCLIP από το [19] , γίνουν διαθέσιμα σε μεγάλες ποσότητες. Τα δεδομένα αυτά αποτελούν το τρέχον «πρότυπο χρυσού».

Ως εκ τούτου, χρησιμοποιούνε το σύνολο δεδομένων [αναφ. 9,17] και το σύνολο δεδομένων για τον άνθρωπο pSILAC [18] για αξιολόγηση.

- **Σύγκριση απόδοσης σε Drosophila melanogaster**

Το 2005, Stark [17] διεξήγαγε μια ευρεία σύγκριση των ευρέως χρησιμοποιούμενων προσεγγίσεων πρόβλεψης στόχων. Μια σειρά από 133 πειραματικά δοκιμασμένες λειτουργικές και μη λειτουργικές αλληλεπιδράσεις γονιδίου microRNA συντάχθηκε, από τις οποίες 120 (λειτουργικές: 61, μη λειτουργικές: 59) χρησιμοποιήθηκαν για την πραγματική σύγκριση [17]. Αυτό το σύνολο δεδομένων χρησίμευσε ως πρότυπο της αλήθειας για την αξιολόγηση της εξελικτικής προσέγγισης στην πρόβλεψη στόχου microRNA που δημοσίευσε ο Stark [20] και αργότερα επεκτάθηκε σε 190 αλληλεπιδράσεις από Kertesz [9].

Για να αξιολογήσουνε την προβλεπτική ισχύ της μεθόδου και να την συγκρίνουνε με άλλες μεθόδους, την εφαρμόσανε πρώτα στην original σύνολο (Gaidatzis) και έπειτα στο επεκταμένο σύνολο στόχων (Kertesz).

Δεδομένου ότι εκχωρούνε κάθε υποψήφια ζώνη σε μια βαθμολογία, μπορούνε να ποσοτικοποιήσουμε την απόδοση με μια καμπύλη (ROC), καθιστώντας τη σύγκριση με άλλες προσεγγίσεις πιο διαφανή.

Κατά τη σύγκριση της απόδοσης των μεθόδων στο αρχικό σύνολο δεδομένων των Stark (Εικόνα 13A, C):

ο αλγόριθμος EIMMo [21] επιτυγχάνει τα καλύτερα αποτελέσματα, δείχνοντας ένα υψηλό πραγματικό θετικό ποσοστό σε συνδυασμό με ένα χαμηλό ψευδώς θετικό ποσοστό.

Παρά το πλεονέκτημα ότι είναι σε θέση να συγκρίνει όλες τις μεθόδους με μία μόνο τιμή, η εξέταση της εξέλιξης της καμπύλης ROC είναι ακόμη πιο διαφωτιστική, ειδικά για εκείνες τις μεθόδους που συσσωρεύονται στενά μεταξύ τους με την τιμή AUC.

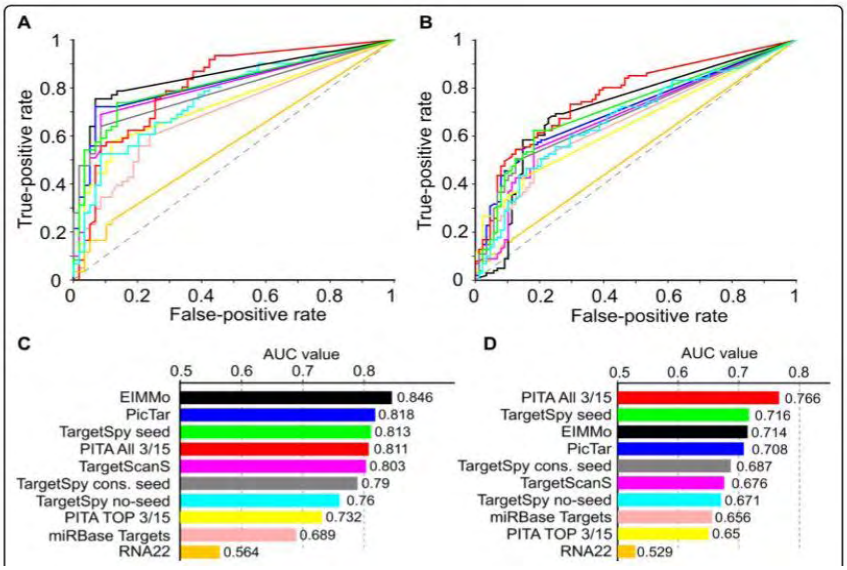
Ιδιαίτερα ενδιαφέρον είναι το χαρακτηριστικό των καμπυλών σε χαμηλά ψευδώς θετικά ποσοστά, όπως για πολλά πειράματα η ποσότητα των δειγμάτων μπορεί να είναι πολύ περιορισμένη. Η εκδοχή του TargetSpy με συντηρημένες περιοχές seed, παρουσιάζει τη χαμηλότερη τιμή ψευδώς θετικού ποσοστού (FPR) στη δοκιμή με ένα πραγματικό θετικό ποσοστό (TPR) 48%. Η εκδοχή του TargetSpy με seed, παρουσιάζει το δεύτερο χαμηλότερο FPR, αλλά προσφέρει ελαφρώς καλύτερη TPR, συγκρίσιμη με αυτή του PicTar. Ενώ, η εκδοχή του TargetSpy χωρίς seeds & διατηρησιμότητα, εμφανίζει απόδοση που πλησιάζει τις μεθόδους των κατηγοριών II και III, ειδικά για τις κορυφαίες προβλέψεις του που καλύπτουν περισσότερο από το 50% του TPR.

Η συγκριτική αξιολόγηση στο εκτεταμένο σύνολο 190 πειραματικά επαληθευμένων αλληλεπιδράσεων microRNA-στόχου παράγει αρκετές ενδιαφέρουσες παρατηρήσεις (Εικόνα 13B, D):

Πρώτον, οι τιμές AUC είναι γενικά χαμηλότερες σε σύγκριση με το αρχικό σύνολο. Δεύτερον, το PITA, ειδικά προσαρμοσμένο σε αυτό το σετ, βρίσκεται πολύ μπροστά από όλες τις άλλες προσεγγίσεις. Τρίτον, η κατάταξη των άλλων προσεγγίσεων δεν έχει αλλάξει εκτός από το ότι i) Η εκδοχή του TargetSpy με seeds, αποδίδει καλύτερα από τα PicTar και EIMMo, η εκδοχή του TargetSpy με seeds & διατηρησιμότητα υπερέχει των TargetScanS και miRBase Targets [24] και έχει καλύτερη απόδοση από την PITA TOP 3/15 και ii) Η εκδοχή του TargetSpy χωρις seeds & διατηρησιμότητα και το TargetScanS μειώνονται. Τέλος, το specificity του EIMMo και του TargetScanS, μειώθηκε ιδιαίτερα έντονα για τις κορυφαίες προβλέψεις τους.

Συνοπτικά, η αξιολόγηση των πειραματικών δεδομένων μύγας υποδηλώνει ότι το TargetSpy, το οποίο εκπαιδεύτηκε στα δεδομένα του ποντικιού, λειτουργεί τόσο καλά όσο οι σύγχρονοι αλγόριθμοι που λαμβάνουν υπόψη το κριτήριο της αντιστοίχισης seeds.

Επιπλέον, η πρόβλεψη για καμία αντιστοίχιση seed είναι σημαντικά καλύτερη από τον RNA22, τον άλλο δοκιμασμένο αλγόριθμο που δεν απαιτεί τέλεια αντιστοίχιση seeds.



Εικόνα 13 Σύγκριση απόδοσης διαφόρων προσεγγίσεων

από τη δημοσίευση: TargetSpy: a supervised machine learning approach for microRNA target prediction, Martin Sturm, Michael Hackenberg, David Langenberger, Dmitrij Frishman [Αναφ. 28]

- **Αξιολόγηση στα δεδομένα pSILAC**

Selbach [αναφ. 18] πραγματοποίησε μια σύγκριση των πιο ευρέως χρησιμοποιούμενων προσεγγίσεων μετρώντας το κλάσμα των προβλεπόμενων θέσεων στόχων που σχετίζονται με πρωτεΐνες οι οποίες είναι πιο έντονα ρυθμισμένες προς τα κάτω - downregulated από -0,1 $\log_2$ fold change.

Δημιούργησαν δύο σύνολα (τυχαία) background:

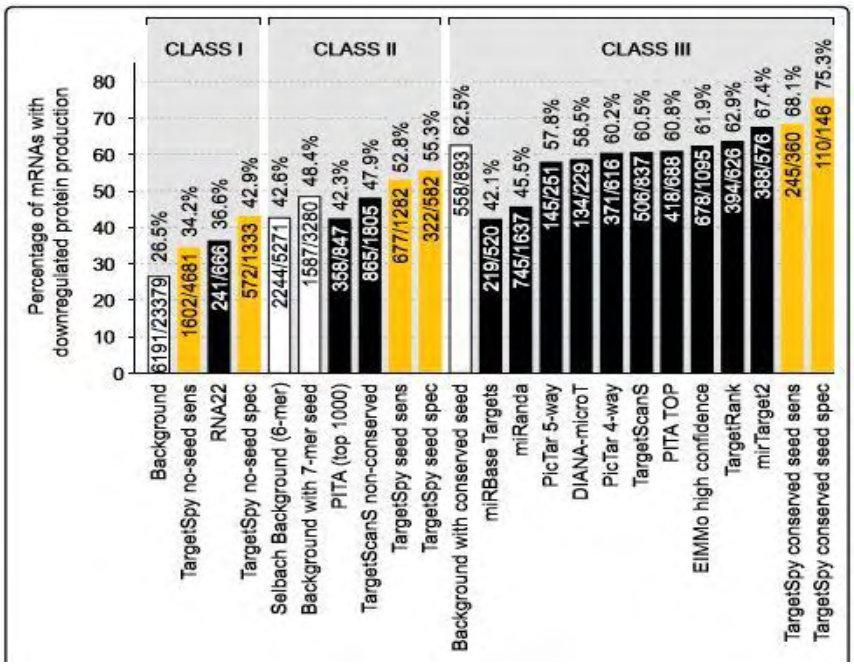
(i) ένα σύνολο όπου όλα τα mRNA που υπάρχουν θεωρούνται ως στόχοι και

(ii) ένα σύνολο όλων των mRNA που έχουν ένα ταίριασμα 6mer seed στην αλληλουχία τους, που αναφέρεται περαιτέρω ως "background Selbach".

Δημιουργήσανε επιπλέον background 7mer seeds για την κατηγορία II, καθώς το PicTar και το TargetSpy απαιτούν τέλειες αντιστοιχίες 7mer seeds.

Προκειμένου να συγκριθούν οι προγνωστικοί δείκτες κατηγορίας III με την τυχαία προσδοκία, εισήγαγαν επιπλέον ένα σύνολο δεδομένων background για συντηρημένες αντιστοιχίσεις 7mer seeds.

Συγκεκριμένα, ερευνήσανε αντιστοιχίσεις 6mer seeds (θέσεις 2-7) που διατηρούνται τέλεια σε αρουραίους, ανθρώπους, χιμπαντζήδες, ποντικούς και σκύλους, οι οποίες δείχνουν επιπρόσθετα ένα ταίριασμα είτε στη βάση 1 είτε στην 8 [25]. Όπως φαίνεται στην Εικόνα 14 **για την κατηγορία I** (χωρίς seed / χωρίς συντήρηση), μια εντελώς τυχαία επιλογή των θέσεων στόχων θα απέδιδε yield ~ 27% (ακρίβεια background) με ρυθμισμένες προς τα κάτω πρωτεΐνες στο pSILAC. Τόσο ο RNA22 όσο και ο no seed TargetSpy έχουν καλύτερη απόδοση από την τυχαία.



Εικόνα 14 Σύγκριση απόδοσης διαφόρων προσεγγίσεων με pSILAC δεδομένα

από τη δημοσίευση: TargetSpy: a supervised machine learning approach for microRNA target prediction, Martin Sturm, Michael Hackenberg, David Langenberger, Dmitrij Frishman [Αναφ. 28]

Ο no seed sens TargetSpy επιτυγχάνει την ακρίβεια 34,2%, σημαντική βελτίωση σε σύγκριση με τις τυχαίες προβλέψεις. Το RNA22 παρουσιάζει ακρίβεια 36,2%.

Ο no seed spec TargetSpy δεν περιέχει περισσότερες τοποθεσίες στόχους από το RNA22, αλλά επιτυγχάνει ακρίβεια 42,9% και είναι επομένως στο ίδιο επίπεδο με το background 6mer seeds της κατηγορίας II.

Η ακρίβεια υποβάθρου **για την κατηγορία II** είναι στο 42,6% όταν χρησιμοποιούνται 6mer seeds (background Selbach).

Δεδομένου ότι ο PITA καλύπτει όλους τους υποψήφιους στόχους με matches που ξεκινούν από το μέγεθος των 6 nt και στη συνέχεια τους κατατάσσει ανάλογα με την προσβασιμότητά τους, είναι απαραίτητο να ληφθούν υπόψη μόνο οι κορυφαίες προβλέψεις. Έτσι, πήραν τις κορυφαίες 1000 προβλέψεις ανά microRNA και βρήκαν μια αλληλεπικάλυψη 42,3% με μειωμένες πρωτεΐνες που αποδεικνύουν ακρίβεια κάτω από το επίπεδο background.

Στην κατηγορία III, ο miranda & ο mirBase είναι σαν να ανήκουν στην κλάση 2.

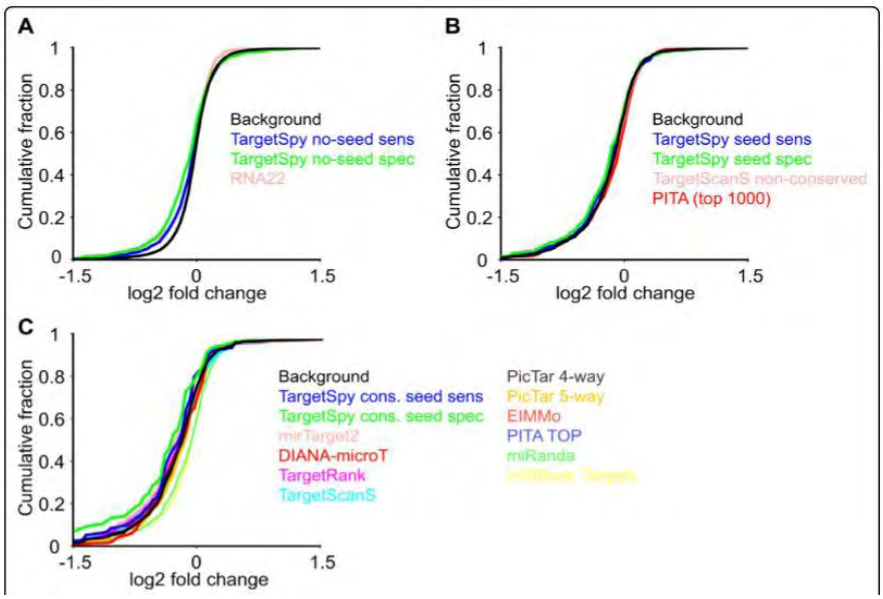
Επίσης, το DIANA-microT, το TargetScanS, το PITA TOP, ένα επίσημο υποσύνολο της PITA με διατηρημένα ταιριάσματα 8μερών seeds που απαιτούνται και το EIMMo με τη ρύθμιση υψηλής εμπιστοσύνης (score> = 0.5) βρίσκονται ελαφρώς κάτω απ' το background.

Η πρώτη προσέγγιση που βρίσκεται πάνω από το background της κατηγορίας III είναι η TargetRank [26] (62,9%), ακολουθούμενη από το mirTarget2 [21] (67,4%) και το conserved seed sens TargetSpy (68,1%).

Τελικά, το TargetSpy conserved seed spec επιτυγχάνει την υψηλότερη ακρίβεια όλων των μεθόδων (75,3%).

Πρέπει να σημειωθεί ότι παρόλο που το TargetSpy επιτυγχάνει ανώτερη απόδοση όσον αφορά την ακρίβεια και την ευαισθησία στις κατηγορίες I και II, η ευαισθησία του TargetSpy στην κατηγορία III είναι χαμηλότερη σε σύγκριση με άλλες προσεγγίσεις.

Για να αποκλειστεί η πιθανότητα ότι η εκτελούμενη αξιολόγηση ισχύει μόνο για το επιλεγμένο κατώφλι -0.1 $\log_2$ fold change, ερευνήσανε επίσης το σωρευτικό κλάσμα των προβλεπόμενων θέσεων στόχων ως συνάρτηση της πρωτεΐνης $\log_2$ fold change. Η Εικόνα 15 δείχνει τις κατανομές για καθεμία από τις τρεις κατηγορίες. Είναι προφανές ότι η σχετική απόδοση των υπολογιστικών προσεγγίσεων παραμένει ουσιαστικά αμετάβλητη για κάθε τιμή fold change σε κάθε κατηγορία. Εντούτοις, όπως φαίνεται στο Σχήμα 15c, η πρόοδος της TargetSpy cons. seed spec (πράσινη γραμμή) είναι ιδιαίτερα έντονη για χαμηλές τιμές fold change. Εφόσον οι χαμηλές τιμές fold change αντιστοιχούν σε ισχυρότερη μείωση των πρωτεϊνών, αυτό μπορεί να σημαίνει ότι η προσέγγιση αυτή αποδίδει ακόμα καλύτερα για πολύ αποδοτικές τοποθεσίες στόχους.



Εικόνα 15 Σύγκριση προσεγγίσεων σε διάφορα κατώφλια fold change

από τη δημοσίευση: TargetSpy: a supervised machine learning approach for microRNA target prediction, Martin Sturm, Michael Hackenberg, David Langenberger, Dmitrij Frishman [Αναφ. 28]

Γενικά τα αποτελέσματά για τα δεδομένα pSILAC υποδηλώνουν ότι το TargetSpy αποδίδει καλύτερα σε κάθε κατηγορία, επιδεικνύοντας επιπλέον ένα σταθερό κέρδος στην ακρίβεια από το sensitive στο specific threshold. Αυτό το εύρημα υποδηλώνει ότι η ποιότητα πρόβλεψης αυξάνεται με το σκορ και επομένως η κατάταξη των θέσεων στόχων που επιβάλλονται από το σκορ του μοντέλου TargetSpy φαίνεται ότι έχει βιολογική σημασία.

Υπό την παραδοχή ότι όλες οι αλληλεπιδράσεις pSILAC που δεν ρυθμίζονται προς τα κάτω είναι αληθή αρνητικά (true negatives), οι παράμετροι ROC επιβεβαιώνουν τις αξιολογήσεις απόδοσης με βάση την ακρίβεια.

Τέλος, προσδιορίσανε τον αριθμό των λειτουργικών θέσεων στόχων (ρυθμισμένες προς τα κάτω πρωτεΐνες) που δε διαθέτουν περιοχές seed, οι οποίες έχουν προβλεφθεί σωστά αποκλειστικά από το TargetSpy αλλά όχι από οποιονδήποτε άλλο αλγόριθμο. Προκειμένου να αποφευχθούν υποτιθέμενες λανθασμένες τοποθεσίες στόχων, αποκλείσανε από την εξέταση αυτών των αλληλεπιδράσεων γονιδίου-

microRNA που έδειξαν τουλάχιστον ένα τέλειο ταίριασμα seed. Μετά την αφαίρεση αυτών των αλληλεπιδράσεων γονιδίου-microRNA ελήφθησαν 564 μοναδικές θέσεις στόχου χωρίς περιοχή seed στο sensitivity σύνολο και 134 στο specificty σετ. Αυτό σημαίνει ότι το TargetSpy αναφέρει κατά μέσο όρο μεταξύ 26 λειτουργικών στόχων (spec) και 112 (sens), οι οποίοι δεν δείχνουν αντιστοιχία seed ανά microRNA που δεν μπορούσε να ανιχνευθεί από οποιοδήποτε άλλο εργαλείο.

4.4 ΑΞΙΟΛΟΓΗΣΗ PITA

Ζεύγη αλληλεπίδρασης microRNA-στόχου δοκιμάστηκαν πειραματικά στη βιβλιογραφία.

Η συλλογή πειραματικά δοκιμασμένων ζευγών Drosophila microRNA-mRNA βασίστηκε σε προηγουμένως δημοσιευμένο κατάλογο, στον οποίο προστέθηκαν στόχοι αναφερόμενοι σε TarBase [Αναφ. 49] και στόχων miR-2 και miR-184 από τα δικά τους πειράματα. Η προκύπτουσα μη περιττή λίστα αποτελείται από 190 συνολικά ζεύγη, εκ των οποίων 102 αναφέρθηκαν ως στόχοι λειτουργικού microRNA και 88 αναφέρθηκαν ως μη λειτουργικά.

Έλεγχος της επιλογής της τοποθέτησης microRNA seeds για περιοχές ανοικτής δευτερογενούς δομής.

Για κάθε οργανισμό, σαρώσανε τα 3’ UTRs για τέλεια seeds στόχους microRNA τουλάχιστον επτά βάσεων μήκους, αποκλείοντας τη G:U ταλάντωση, αναντιστοιχίες ή βρόχους. Επειδή το ΔG open επηρεάζεται από το περιεχόμενο GC, μια συλλογή ίσων ποσοτήτων τοποθεσιών με την ίδια κατανομή περιεχομένου GC με αυτή των πραγματικών seeds microRNA επιλέχθηκε τυχαία ως σύνολο ελέγχου.

Υπολογίσανε τη βαθμολογία προσβασιμότητας (ΔG open) για κάθε αληθινό και ελεγχόμενο seed (χωρίς flank) και συγκρίνανε τις κατανομές τους χρησιμοποιώντας το τεστ Kolmogorov-Smirnov. Για τη δοκιμή περιορισμένης διατήρησης ελήφθησαν διαδρομές συντήρησης βασισμένες σε ένα φυλογενετικό κρυφό μοντέλο Markov (phastCons) από τη θέση UCSC (σύγκριση 12 ειδών Drosophila, κουνουπιών, μέλισσας και κόκκινου σκαθαριού για την αλληλουχία μύγας και 17 σπονδυλωτά, συμπεριλαμβανομένων θηλαστικών, αμφιβίων , είδη πτηνών και ψαριών για τις αλληλουχίες του ανθρώπου και του ποντικιού). Μόνο seeds microRNA που έχουν μέση (κατά μήκος της seed περιοχής) βαθμολογία συντήρησης μεγαλύτερη από 0,9 εξετάστηκαν. Για αυτές τις συγκρίσεις, το σετ ελέγχου των τυχαίων seeds επιλέχθηκε επίσης από περιοχές του 3 'UTR με μια τιμή διατήρησης μεγαλύτερη από 0,9.

4.5 ΑΞΙΟΛΟΓΗΣΗ EIMMO2

Pathway enrichment ανάλυση

Χρησιμοποιήσανε τη βάση δεδομένων KEGG για να συναγάγουν μονοπάτια στοχευμένα από μεμονωμένα miRNAs. Η βάση δεδομένων KEGG (ftp.genome.jp) περιέχει αντιστοιχίσεις από αναγνωριστικά γονιδίων NCBI σε αναγνωριστικά μονοπατιών, ενώ η βάση δεδομένων Gene του NCBI αντιστοιχίσεις από αναγνωριστικά γονιδίων σε αναγνωριστικά του Refseq.

Διασυνδέοντας αυτά τα σύνολα δεδομένων, αποκτήσανε τις αντιστοιχίσεις από αναγνωριστικά του Refseq σε διαδρομές. Στη συνέχεια χρησιμοποιήσανε Bayesian μέθοδο για να προσδιορίσουνε τη σημασία της αλληλεπικάλυψης μεταξύ των στόχων κάθε σετ ισοδύναμων seeds miRNAs και κάθε συγκεκριμένης οδού.

Για ένα δεδομένο μονοπάτι και το miRNA αφήνουμε τα n01, n10, n00 και n11 να υποδηλώσουν αντίστοιχα τον αριθμό των προβλεπόμενων στόχων του miRNA που δεν είναι μέρος της οδού, ο αριθμός των γονιδίων στο μονοπάτι που δεν στοχεύουν το miRNA, ο αριθμός των γονιδίων που δεν είναι ούτε στόχοι του miRNA ούτε μέλη του μονοπατιού και ο αριθμός των γονιδίων στο μονοπάτι που προβλέπεται

ότι θα στοχεύουν το miRNA. Ενώ η συμμετοχή στο μονοπάτι είναι μια απλή boolean μεταβλητή (ένα γονίδιο είναι είτε μέλος ενός δεδομένου μονοπατιού είτε δεν είναι), μπορούμε να εκχωρήσουμε μόνο πιθανότητες για ένα δεδομένο γονίδιο να είναι στόχος miRNA. Ας υποθέσουμε ότι ένα δεδομένο γονίδιο έχει n υποθετικές θέσεις-στόχους για ένα δεδομένο miRNA και αφήνει το p_i να υποδηλώσει την posterrior πιθανότητα της i-θέσης. Η πιθανότητα ότι τουλάχιστον μία από τις θέσεις είναι λειτουργική δίνεται στη συνέχεια από

$$p_{\text{tar}} = 1 - \prod_{i=1}^{n} (1 - p_i)$$

Χρησιμοποιούνε το ptar ως την πιθανότητα να στοχεύσουμε το γονίδιο από το miRNA και να πάρουμε n01 και n11 με άθροισμα των ptar πάνω από όλα τα γονίδια που δεν βρίσκονται στο μονοπάτι και όλα τα γονίδια στο μονοπάτι αντίστοιχα.

Ομοίως αθροίζουμε (1 - ptar) πάνω από όλα τα γονίδια που δεν βρίσκονται στο μονοπάτι και όλα τα γονίδια στο μονοπάτι για να λάβουμε n00 και n10 αντίστοιχα.

Τέλος, υπολογίζουμε την πιθανότητα των παρατηρούμενων αριθμών n00, n10, n01 και n11 κάτω από ένα «ανεξάρτητο μοντέλο», στο οποίο η πιθανότητα να στοχεύεται από το miRNA είναι ανεξάρτητη από την ιδιότητα του μονοπατιού και ένα «εξαρτώμενο μοντέλο» στο οποίο η πιθανότητα της στόχευσης του miRNA εξαρτάται γενικά από την ένταξη στο μονοπάτι.

Η πιθανότητα από το ανεξάρτητο μοντέλο δίνεται από

$$\begin{aligned} L_{\text{indep}} &= \int_0^1 (pq)^{n_{11}} (p(1-q))^{n_{10}} \\ &\quad ((1-p)q)^{n_{01}} ((1-p)(1-q))^{n_{00}} \, dpdq \qquad (10) \\ &= \frac{\Gamma(n_{1.}+1)\Gamma(n_{0.}+1)\Gamma(n_{.0}+1)\Gamma(n_{.1}+1)}{\Gamma(n+2)\Gamma(n+2)}, \end{aligned}$$

Όπου Γ (x) είναι η συνάρτηση γάμμα, μια τελεία δηλώνει άθροιση πάνω στην εν λόγω μεταβλητή, δηλ. n1. = n10 + n11 και n είναι ο συνολικός αριθμός γονιδίων.

Για το εξαρτώμενο μοντέλο η πιθανότητα δίνεται από

$$\begin{aligned} L_{\text{dep}} &= \int p_{00}^{n_{00}} p_{10}^{n_{10}} p_{01}^{n_{01}} p_{11}^{n_{11}} \, dp_{00} dp_{01} dp_{10} dp_{11} \\ &= \frac{\Gamma(4)\Gamma(n_{11}+1)\Gamma(n_{10}+1)\Gamma(n_{01}+1)\Gamma(n_{00}+1)}{\Gamma(n+4)}, \qquad (11) \end{aligned}$$

Όπου το ολοκλήρωμα είναι πάνω από το simplex p00 + p10 + p01 + p11 = 1.

Η αναλογία πιθανότητας $\mathbf{L_{dep}/L_{indep}}$ ποσοτικοποιεί την ποσότητα αποδεικτικών στοιχείων για τη σχέση μεταξύ των στόχων miRNA και της οδού. Αυτή η συσχέτιση μπορεί είτε να είναι θετική (στόχοι miRNA εμπλουτίζονται στο μονοπάτι) είτε αρνητικοί (στόχοι miRNA εξαντλούνται στο μονοπάτι).

Καταγράψανε την ένδειξη της ποσότητας (n11n .. - n1.n.1) p_{dep}, όπου

$$p_{\text{dep}} = \frac{L_{\text{dep}}}{L_{\text{indep}} + L_{\text{dep}}}$$

είναι η posterior πιθανότητα απ' το εξαρτημένο μοντέλο (υποθέτοντας μια ομοιόμορφη prior).

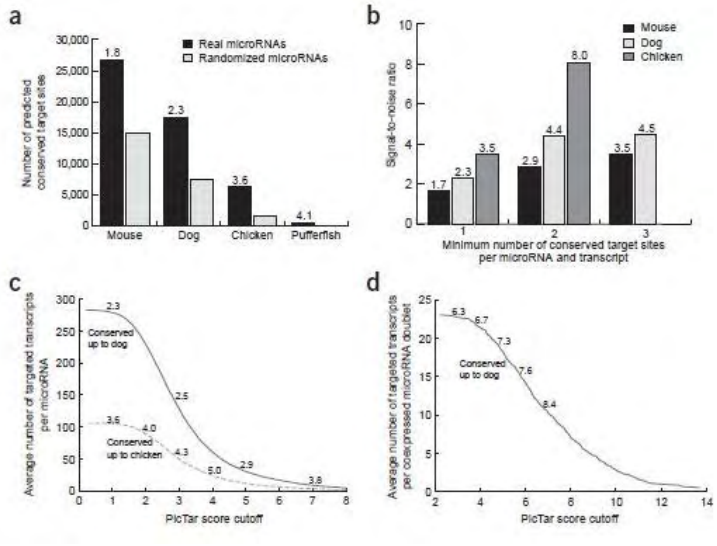
4.6 Αξιολογηση PicTar2

Για να δοκιμάσουν το PicTar, το εφάρμοσαν για να αναζητήσουν στο σύνολο του γονιδιώματος 10.607 C. elegans και Caenorhabditis briggsae 3' UTR αλληλουχίες (συμπληρωματικές μέθοδοι) για στόχους lin-4 ή let-7. Οι γνωστοί στόχοι lin-14, hbl-1, daf-12 και lin-28 κατατάχθηκαν πρώτος, δεύτερος, τέταρτος και έβδομος αντίστοιχα και μόνο ένα γνωστό γονίδιο στόχος (lin-41) δεν ανακτήθηκε, υποδηλώνοντας ότι το PicTar έχει καλό specificity και sensitivity. Σύμφωνα με προηγούμενες μελέτες [18,19], το PicTar προβλέπει ότι τα lin-14 και lin-28 θα πρέπει να στοχεύονται και από το lin-4 και το let-7. Συγκεκριμένα, το PicTar βρήκε μόνο μερικά γονίδια με τοποθεσίες τόσο για το lin-4 όσο και για το let-7, γεγονός που υποδηλώνει ότι ο αριθμός των κοινών στόχων lin-4-let-7 είναι σχετικά μικρός.

Εξετάσανε περαιτέρω το PicTar υπολογίζοντας προβλέψεις για κάθε microRNA ξεχωριστά σε όλα τα C. elegans 3' UTRs χωρίς συγκρίσεις μεταξύ των ειδών. Οι δοκιμές τυχαίας επιλογής (συμπληρωματικές μέθοδοι) έδειξαν ότι ένα κλάσμα εξαιρετικά σημαντικών (>10 s.d.) των προβλεπόμενων θέσεων διατηρείται εξελικτικά, ενισχύοντας την εμπιστοσύνη στο PicTar. Για προβλέψεις στόχων σε σπονδυλωτά, κατασκευάσανε πολλαπλές ευθυγραμμίσεις 20.254 σχολιασμένων ανθρώπινων 3 'UTRs σε γονιδιακές αλληλουχίες από επτά άλλα σπονδυλωτά, χιμπαντζή, ποντίκι, αρουραίο, σκύλο, κοτόπουλο, pufferfish και ζέβρα, χρησιμοποιώντας τη βάση δεδομένων του Πανεπιστημίου της Καλιφόρνια (UCSC). Από αυτές τις ευθυγραμμίσεις, το 92% κάλυψε όλα τα είδη θηλαστικών, το 55% περιλάμβανε τις ακολουθίες κοτόπουλου και το 21% κάλυψε και τα οκτώ σπονδυλωτά. Συγκρίνοντας τους συνδυασμούς αλληλουχιών ανθρώπου-ποντικιού των ευθυγραμμίσεων με ζευγαρώματα που ανεξάρτητα προσδιορίστηκαν μέσω ενός πίνακα ορθολογίας γονιδίου έδωσε χαμηλό ποσοστό σφάλματος 3%.

Για την εκτίμηση ψευδώς θετικών ποσοστών για τις προβλέψεις σπονδυλωτών στόχων microRNA, καταγράψανε τέλειες αντιστοιχίσεις διατηρημένου στόχου («anchors») για 58 μοναδικά ανθρώπινα microRNAs συντηρημένα σε ανθρώπινα, χιμπατζή, ποντίκι, αρουραίο, σκύλο και κοτόπουλο και για τυχαιοποιημένα microRNAs [11,13]. Το σχήμα 16a δείχνει την αναλογία του συνολικού αριθμού θέσεων πρόσδεσης για πραγματικά και τυχαιοποιημένα microRNA σε διαφορετικούς βαθμούς συντήρησης.

Η συμπερίληψη των γονιδιωμάτων σκύλου και κοτόπουλου βελτίωσε σημαντικά την αναλογία σήματος προς θόρυβο από 1,8 για τον άνθρωπο, τον χιμπαντζή και τον ποντικό σε 2,3 και 3,6, αντίστοιχα. **Συνολικά, τα αποτελέσματα υποδεικνύουν ότι, κατά μέσο όρο, κάθε microRNA στοχεύει μεταγραφές 200 πάνω από θόρυβο, παρόμοια με άλλες προβλέψεις [21].** Καταγράψανε λόγους σήματος προς θόρυβο για τον αριθμό των μεταγραφών με τουλάχιστον n συντηρημένες άγκυρες για κάθε microRNA χωριστά (Σχήμα 16b). Η πολλαπλότητα των τοποθεσιών σε ένα UTR, το οποίο βαθμολογείται από το PicTar, οδηγεί επίσης σε μια σημαντική αύξηση του σήματος-προς-θόρυβο [11]. Από αυτά τα αποτελέσματα, χρησιμοποιήσανε το PicTar για να κατατάξουνε τις προβλέψεις στόχων για όλα τα διαθέσιμα σήμερα, διατηρημένα microRNAs.



Εικόνα 16 Αναλογία σήματος προς θόρυβο για προβλέψεις θέσης στόχου microRNA για σπονδυλωτά

από τη δημοσίευση: Combinatorial microRNA target predictions, Azra Krek, Dominic Grun, Matthew N Poy, Rachel Wolf, Lauren Rosenberg, Eric J Epstein, Philip MacMenamin, Isabelle da Piedade, Kristin C Gunsalus, Markus Stoffel & Nikolaus Rajewsky [Αναφ. 47].

Στην εικόνα 16 φαίνεται συνοπτικά,

(α) Αναλογία θορύβου σήματος για προβλεπόμενους μοναδικούς τόπους στόχους. Ο αριθμός των προβλεπόμενων θέσεων στόχων (άγκυρες) για το σύνολο 58 μοναδικών συντηρημένων ανθρώπινων microRNAs έναντι του αντίστοιχου αριθμού για τυχαιοποιημένα microRNAs, απαιτώντας τη διατήρηση θέσεων αγκύρωσης μεταξύ ανθρώπου, χιμπαντζή, ποντικού (πρώτη στήλη), αρουραίου και σκύλου (δεύτερη στήλη) , κοτόπουλο (τρίτη στήλη) και φούσκα (τελευταία στήλη). Η συμπερίληψη των πιο απομακρυσμένων σχετικών ειδών ενισχύει σημαντικά την αναλογία σήματος προς θόρυβο (που υποδεικνύεται πάνω από τις μαύρες ράβδους). Για ανθρώπους, χιμπαντζήδες, ποντικούς, αρουραίους και σκύλους, προβλέπουμε 17.542 διατηρούμενες περιοχές στόχους με λόγο σήματος προς θόρυβο 2.3 και ως εκ τούτου 10.000 πραγματικές θέσεις στόχου. (β) Η πολλαπλότητα των θέσεων στόχων ενισχύει την αναλογία σήματος προς θόρυβο. Η αναλογία του αριθμού των μεταγραφών με τουλάχιστον θέσεις n άγκυρας ανά microRNA για πραγματικά έναντι τυχαίων microRNA παρέχει μια εκτίμηση της αναλογίας σήματος προς θόρυβο για θέσεις που διατηρούνται σε ανθρώπους, χιμπατζήδες, ποντικούς (μαύρες ράβδους), αρουραίους, σκύλους μπαρ) και κοτόπουλο (σκούρο γκρι μπάρες). Η πολλαπλότητα των τοποθεσιών στόχων (βαθμολογία από το PicTar) βοηθά στην αύξηση του λόγου σήματος προς θόρυβο. (γ) PicTar εξαρτώμενη από τη βαθμολογία ευαισθησία και συγκεκριμένη πρόβλεψη μοναδικής microRNA θέσης στόχου. Ο μέσος αριθμός προβλεπόμενων στόχων ενός μοναδικού microRNA με τουλάχιστον μία θέση άγκυρας ανά μεταγραφή σε άνθρωπο, χιμπαντζή, ποντικό, αρουραίο και σκύλο (ανώτερη καμπύλη) ή σε ανθρώπους, χιμπαντζήδες, ποντικούς, αρουραίους, σκύλους και κοτόπουλο (κατώτερη καμπύλη) είναι με γραφική παράσταση ως συνάρτηση μιας διακοπής βαθμολογίας PicTar (απόρριψη μεταγραφών με βαθμολογία κάτω από την αποκοπή). Οι αναλογίες σήματος προς θόρυβο υποδεικνύονται πάνω από κάθε καμπύλη. (δ) PicTar εξαρτώμενη από την βαθμολογία ευαισθησία και συγκεκριμένη πρόβλεψη θέσης στόχου για τέσσερα σύνολα συν-εκφρασμένων microRNA4 και αντίστοιχα σύνολα τυχαιοποιημένων microRNAs, απαιτώντας δύο θέσεις άγκυρας για διαφορετικά microRNAs στον άνθρωπο, τον χιμπαντζή, τον ποντικό, τον αρουραίο και τον σκύλο . Η γραφική παράσταση δείχνει τον μέσο αριθμό στόχων ανά ζεύγος μικροRNAs ως συνάρτηση της απόκλισης βαθμολογίας PicTar (λόγοι σήματος προς θόρυβο πάνω από την καμπύλη).

Το specificity και η ευαισθησία συσχετίζονται έντονα με τη βαθμολογία PicTar (Εικόνα 16c). Το specificity ως συνάρτηση του PicTar για σύνολα συνεκφρασμένων microRNAs που χρησιμοποιούνται για συγκεκριμένες προβλέψεις στόχου σε τέσσερις ιστούς θηλαστικών δείχνεται στο Σχήμα 16d.

4.7 Σύγκριση των Προγραμματων

Στα πλαίσια της μελέτης των έξι αλγορίθμων, εκτελέστηκαν όλα τα προγράμματα, το καθένα στο αντίστοιχο υπολογιστικό περιβάλλον που περιγράφηκε σε προηγούμενο κεφάλαιο, συλλέχθηκαν τα αποτελέσματά τους και πραγματοποιήθηκαν σειρές πειραμάτων με απώτερο σκοπό την σύγκριση των αλγορίθμων και τη συσχέτισή τους.

Πιο συγκεκριμένα, συλλέχθηκαν τα αποτελέσματα από κάθε πρόγραμμα, περισσότερες πληροφορίες των οποίων περιγράφονται στους παρακάτω πίνακες.

Χαρακτηριστικά προγραμμάτων

Program	Website	Organism	Reference
mirTarget2	http://mirdb.org/	Σπονδυλωτά	8
miRmap	http://mirmap.ezlab.org/	Σπονδυλωτά	19
targetSpy	http://webclu.bio.wzw.tum.de/targetspy/	Σπονδυλωτά & Μύγα	28
PITA	https://genie.weizmann.ac.il/pubs/mir07/	Όλα	33
EIMMo2	-	Όλα	38
PicTar2	http://pictar.mdc-berlin.de/	Όλα	48

Πίνακας 9 Βασικές πληροφορίες για τα προγράμματα

Features	mirTarget2	miRmap	targetSpy	PITA	EIMMo	PicTar2
Machine Learning	x	-	x	-	x	-
Sequence						
3'UTR Region	x	x	x	x	x	x
Perfect seed match	-	x	x[a]	-	x	-
Preference for perfect seed match	x	-	-	x	-	x
3/15 flank requirements	-	-	-	x	-	-
3' compensatory site	-	-	x	-	-	-

Thermodynamics	-	x	x	x	-	x
Probabilistic	-	x	-	-	x	-
Conservation	x	x	-	x[b]	x	x

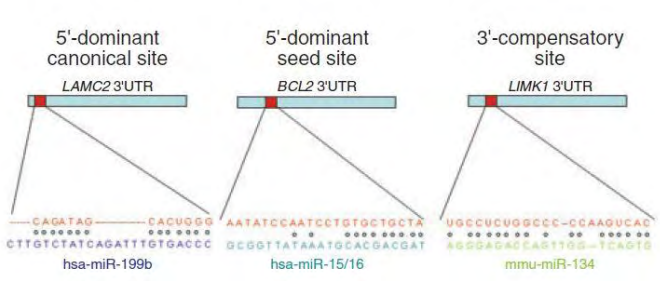
Πίνακας 10 Σύνοψη βασικών χαρακτηριστικών όλων των προγραμμάτων

X[a]: Ο αλγόριθμος targetSpy, κανονικά δεν απαιτεί καμία απαίτηση όσον αφορά την πρόσδεση σε seed type περιοχή, παρ' όλα αυτά για λόγους καλύτερης και πιο αντικειμενικής σύγκρισης με άλλους αλγορίθμους έχουν δημιουργηθεί κατάλληλα υποσύνολα λαμβάνοντας υπόψη την τέλεια αντιστοίχιση 7-mer σε seed type περιοχή. Αυτά τα υποσύνολα λήφθηκαν για τη μελέτη της μεταπτυχιακής διατριβής.

X[b]: Ο αλγόριθμος PITA δε λαμβάνει υπόψη καμία απαίτηση όσον αφορά τη συντηρισιμότητα σε άλλα ειδή για την εύρεση θέσεων στόχων miRNA. Απλώς για λόγους σύγκρισης έχουν δημιουργήσει διάφορα υποσύνολα που είναι διατηρημένα σε διάφορα είδη.

Στον παραπάνω πίνακα αναπαρίσταται μια γενική εικόνα για τα χαρακτηριστικά - απαιτήσεις που πρέπει να πληρούν οι αλγόριθμοι. Τα χαρακτηριστικά αυτά κατηγοριοποιήθηκαν σε 4 βασικές ομάδες,

1. **Sequence – Αλληλουχία:**
   Πρόκειται για χαρακτηριστικά που αφορούν τις αλληλουχίες των γονιδίων. Πιο συγκεκριμένα, αν γίνεται η πρόβλεψη σε 3' UTR περιοχές του γονιδιώματος, αν απαιτείται τέλεια πρόσδεση (1 προς 1 αντιστοίχιση) σε seed περιοχή ή αν προβλέπονται όλες οι προσδέσεις αλλά προτιμάται η τέλεια αντιστοίχιση σε seed type, αν λαμβάνονται υπόψη 3/15 flanks (αφορά περιοχές 3 upstream και 15 downstream νουκλεοτίδια), αν προβλέπουν και σε 3' compensatory type περιοχή. Για τους τύπους των περιοχών Βλέπε Εικόνα 17.
2. **Thermodynamics – Θερμοδυναμική:**
   Πρόκειται για χαρακτηριστικά που βασίζονται σε θερμοδυναμικά μοντέλα, υπολογίζοντας ενεργειακά σκορ για κάθε θέση στόχο miRNA.
3. **Probabilistic – Πιθανοτική:**
   Πρόκειται για χαρακτηριστικά που βασίζονται σε πιθανολογικά μοντέλα και μεθόδους όπως η Bayesian μέθοδος έτσι ώστε με βάση πιθανοτήτων να προσδιορίζονται οι θέσεις στόχων miRNA.
4. **Conservation – Συντηρισιμότητα:**
   Πρόκειται για χαρακτηριστικά που απαιτούν τη διατήρηση του στόχου miRNA σε διάφορα άλλα είδη οργανισμών. Ο mirTarget2 λαμβάνει υπόψη τη συντηρισιμότητα σε ποντίκι, αρουραίο, κοτόπουλο και σκύλο. Ο miRmap σε ποντίκι και άνθρωπο. Ο targetSpy δεν εξαρτάται από αυτό τον παράγοντα, ο PITA αν και δεν το έχει σαν απαίτηση ελέγχει και συγκρίνει τις προβλέψεις του αλγορίθμου σε διάφορα σύνολα που είναι συντηρημένα σε άνθρωπο, ποντίκι, μύγα και σκουλήκι. Ο EIMMo2 σε άνθρωπο, μύγα, σκουλήκι και zebrafish, ενώ τέλος ο PicTar2 σε άνθρωπο, ποντίκι, χιμπατζή, σκύλο, κοτόπουλο και pufferfish.
   Τέλος στον πίνακα δείχνεται και ποια προγράμματα βασίζονται στη μηχανική μάθηση.



Εικόνα 17 Τρεις κατηγορίες θέσεων στόχων microRNA. Canonical (αριστερά), seed (κέντρο) και 3′-compensatory (δεξιά) θηλαστικών miRNA θέσεων στόχων [Αναφ. 37]

Χαρακτηριστικά συνόλων δεδομένων

Format συνόλων δεδομένων:

Όλα τα ονόματα miRNA είναι από miRBase18 και τα ονόματα των γονιδίων από Ensembl 83. Για όσα χρειάστηκε έγινε η αντιστοίχιση από GenBank (NCBI) σε Ensembl. Όλα τα δεδομένα, miRNA, γονίδια αφορούν τον άνθρωπο [hsa-].

Η αντιστοίχιση από Gene ID Ensembl σε Gene association έγινε μέσω του archive Ensembl83 http://dec2015.archive.ensembl.org/ και του εργαλείου BioMart. Από τη βάση της Ensemble 83 για γονίδια ανθρώπου, μέσα από την επιλογή Filters ανεβάσαμε το αρχείο με τα αντίστοιχα γονίδια Gene IDs της Ensembl και από την επιλογή Attributes - External επιλέξαμε το gene association RefSeq [e.g. NM_001195597], τέλος, επιλέξαμε Results και Download. Εφόσον εξάγαμε ένα αρχείο με τις αντιστοιχίες, στη συνέχεια με αλγόριθμο σε Python κάναμε την αντιστοίχιση σωστά για κάθε dataset.

mirTarget2:

1η στήλη	miRNA name
2η στήλη	Gene association από GenBank NCBI (e.g. NM_)
3η στήλη	mirtarget2 score

miRmap:

1η στήλη	transcript_id
2η στήλη	gene_id (Ensembl83)
3η στήλη	gene_name
4η στήλη	transcript_chr
5η στήλη	transcript_strand
6η στήλη	mature_microRNA_name
7η στήλη	site_id
8η στήλη	site_end
9η στήλη	transcript2genome
10η στήλη	seed_length
11η στήλη	seed_mismatches_nogu
12η στήλη	seed_gu
13η στήλη	tgs_au
14η στήλη	tgs_position
15η στήλη	tgs_pairing3p
16η στήλη	dg_duplex
17η στήλη	dg_binding
18η στήλη	dg_open
19η στήλη	dg_total
20η στήλη	prob_exact
21η στήλη	prob_binomial
22η στήλη	cons_bls
23η στήλη	selec_phylop
24η στήλη	miRmap_score

targetSpy:

1η στήλη	miRNA name without species prefix (hsa-)
2η στήλη	Gene association από GenBank NCBI (e.g. NM_)
3η στήλη	Position in 3'UTR start
4η στήλη	Position in 3'UTR end
5η στήλη	Gene name
6η στήλη	Sequence
7η στήλη	Energy
8η στήλη	TargetSpy score

PITA:

1η στήλη	Gene association από GenBank NCBI (e.g. NM_)
2η στήλη	Gene name
3η στήλη	microRNA name
4η στήλη	Sites, number of sites
5η στήλη	PITA score

PicTar2:

1η στήλη	Chromosome

2η στήλη	Start
3η στήλη	End
4η στήλη	Gene association από GenBank NCBI (e.g. NM_)
5η στήλη	miRNA name
6η στήλη	Strand
7η στήλη	PicTar2 score

Test dataset 1:

1η στήλη	miRNA
2η στήλη	Ensembl_Gene_id
3η στήλη	Method
4η στήλη	MIMAT
5η στήλη	Gene_name

Test dataset 2:

Επιπρόσθετα με το προηγούμενο format συν :

6η στήλη	chr: Το χρωμόσωμα
7η στήλη	start: Η γονιδιωματική θέση που ξεκινάει η περιοχή πρόσδεσης του miRNA
8η στήλη	end: Η γονιδιωματική θέση που τελειώνει η περιοχή πρόσδεσης
9η στήλη	strand: Σε ποια αλυσίδα προσδένεται («+», «-»).

Όσον αφορά τα test dataset 1 & 2, πρόκειται για δύο σύνολα επαληθευμένων δεδομένων που προέρχονται από το Diana-TarBase[Αναφ. 49].

Το DIANA-TarBase κυκλοφόρησε αρχικά το 2006 και ήταν η πρώτη βάση δεδομένων με στόχο την καταγραφή δημοσιευμένων πειραματικά επικυρωμένων αλληλεπιδράσεων γονιδίου miRNA: γονιδίων. Το DIANA-TarBase v7.0 παρέχει για πρώτη φορά εκατοντάδες χιλιάδες υψηλής ποιότητας χειρωνακτικά επιλεγμένες και πειραματικά επαληθευμένες αλληλεπιδράσεις γονιδίου miRNA: γονιδίου, ενισχυμένες με λεπτομερή επιπρόσθετα δεδομένα. Το DIANA-TarBase v7.0 είναι διαθέσιμο εδώ : http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase/index/

***Μεγέθη συνόλων δεδομένων*:**

	mirTarget2	miRmap	targetSpy sens (a)	targetSpy spec (a)	PITA_ALL (b)	PITA_TOP (b)	PicTar2	Test dataset1	Test dataset2
Predictions	1.196.548	4.217.829	804.711	341.602	4.084.453	209.509	484.585	25.671	2.175

Πίνακας 11 Το πλήθος των ολικών προβλέψεων από κάθε πρόγραμμα

Στον Πίνακα 11 εμφανίζονται τα αρχικά μεγέθη όλων των προϋπολογισμένων συνόλων δεδομένων. Όταν αναφερόμαστε σε προβλέψεις εννοείται ο αριθμός των αλληλεπιδράσεων miRNA:gene που υπάρχει σε κάθε πρόγραμμα ξεχωριστά. Πιο συγκεκριμένα,

(a): Το πρόγραμμα targetSpy όπως έχει αναφερθεί και σε προηγούμενο κεφάλαιο δημιουργεί κάποιες εκδόσεις συνόλων δεδομένων για να είναι εφικτή η σωστή και αποτελεσματική σύγκριση του αλγορίθμου με τα υπόλοιπα προγράμματα. Ο διαχωρισμός γίνεται σύμφωνα με το αν τηρούν δύο απαιτήσεις, την απαίτηση της συντηρησιμότητας και την απαίτηση των seeds, βλέπε Πίνακα 5.

Επιπρόσθετα, για να περιοριστεί ο όγκος των δεδομένων όρισαν κάποια κατώφλια με βάση το false-positive ratio με αποτέλεσμα να δημιουργηθούν δύο υποσύνολα με τον εξής τρόπο:

οι θέσεις-στόχοι με ψευδώς θετικό ρυθμό μικρότερο από 5% (όπως εκτιμήθηκε σε 10-fold διασταυρούμενη επικύρωση) να αντιστοιχούν στο ευαίσθητο-sensitive υποσύνολο και εκείνοι με ψευδώς θετικό ρυθμό 1 % ή λιγότερο στο σύνολο εξειδίκευσης-specific.

Σύνολο sensitive: False-positive < 5%, 0.05

Σύνολο specific: False-positive < 1%, 0.01

Κατά αυτόν τον τρόπο δημιουργηθήκανε τα σύνολα targetSpy_sens και targetSpy_spec. Στην αξιολόγηση της μεταπτυχιακής διατριβής λήφθηκαν και τα δύο υποσύνολα αυτά πληρώντας την απαίτηση των seeds, για πιο δίκαιη σύγκριση με τους υπόλοιπους αλγορίθμους.

(b): Ο αλγόριθμος PITA, διαχωρίζει επίσης τα σύνολα δεδομένων του σε PITA_ALL όπου είναι όλες οι προβλέψεις του και σε PITA_TOP όπου είναι πιο σίγουρες προβλέψεις σύμφωνα με τον αλγόριθμο, όσον αφορά την λειτουργικότητά τους.

Το 2ο υποσύνολο αφορά seeds 7-8mer με καμία αναντιστοιχία και φίλτρα conservation, ενώ το 1ο αφορά και 6mer seeds με κανένα φίλτρο conservation κάνοντας τις περιοχές αυτές να μη θεωρούνται σίγουρα λειτουργικές.

Επιπρόσθετα, για κάθε τέτοιο υποσύνολο γίνεται ένας εξίσου διαχωρισμός σε no flank 0/0 και 3/15 flanks. Το 2ο αφορά περιοχές 3 upstream και 15 downstream nt, νουκλεοτίδια.

Στις συγκρίσεις της διατριβής συμπεριλήφθηκαν τα υποσύνολα PITA_ALL, PITA_0_0_TOP και PITA_3_15_TOP. Παρατηρώντας όμως, ότι μεταξύ των PITA_0_0_TOP και PITA_3_15_TOP εκδόσεων, δεν παρουσιάστηκε μεγάλη αλλαγή στα αποτελέσματα, συμπεριλήφθηκαν μόνο τα σύνολα PITA_ALL και PITA_TOP.

Το σκορ που αντιστοιχεί στις προβλέψεις των υποσυνόλων αυτών είναι ενεργειακό μιας και βασίζεται σε θερμοδυναμικό μοντέλο υπολογισμού του σκορ. Αυτό σημαίνει ότι όσο μικρότερο το ποσοστό τόσο το καλύτερο. Επομένως, αν η βαθμολογία είναι μεγαλύτερη του -10 θεωρείται λειτουργική θέση στόχος σύμφωνα με τον αλγόριθμο, ενώ όσο πλησιάζει στο +10 και μετά όχι τόσο.

Αφού συλλέχθηκαν τα δεδομένα από κάθε πρόγραμμα ξεχωριστά, πραγματοποιήθηκαν 2 βασικές προσεγγίσεις αξιολόγησης των προγραμμάτων με βάση τα δύο διαφορετικά σύνολα επικυρωμένων δεδομένων (test dataset 1, test dataset 2).

- ΑΠΟΔΟΣΗ. Η 1η αποτελείται από την εύρεση αρχικά των έγκυρων συνδυασμών miRNA:gene για κάθε πρόγραμμα – ***True Positive σύνολο*** και την εύρεση των συνολικών προβλέψεων – ***Total σύνολο***.
- ΣΥΣΧΕΤΙΣΗ - ΟΜΟΙΟΤΗΤΑ. Η 2η αποτελείται από την σύγκριση των προγραμμάτων όσον αφορά τη συσχέτιση των προβλέψεων τους, τις κοινές miRNA:gene αλληλεπιδράσεις.

1ο στάδιο αξιολόγησης

Το 1ο στάδιο αξιολόγησης αποτελεί τη διαδικασία σύγκρισης των προγραμμάτων σύμφωνα με τα κατώφλια των σκορ στους εκάστοτε στόχους, που είναι ήδη ορισμένα απ' το κάθε πρόγραμμα. Ενώ στο 2ο στάδιο αξιολόγησης θα ορίσουμε εμείς πειραματικά εκ νεόυ τα κατώφλια στο σκορ των εκάστοτε στόχων για κάθε πρόγραμμα διαφορετικά.

1η προσέγγιση αξιολόγησης - ΑΠΟΔΟΣΗ:

(Α) Ορισμός κατωφλίων σε κάθε αλγόριθμο:

Αρχικά, ορίστηκαν κάποια βασικά thresholds στο σκορ της κάθε πρόβλεψης για τον κάθε αλγόριθμο ξεχωριστά. Τα thresholds, είναι αυτά που ορίστηκαν απ' τα προγράμματα και φαίνονται στον παρακάτω πίνακα.

	mirTarget2	miRmap	targetSpy sens	targetSpy spec	PITA_ALL	PITA_TOP	PicTar2
threshold	50	0.684467	0.99088	0.99930	32.5	17.65	6

Πίνακας 12 Thresholds για κάθε πρόγραμμα

Αναλυτικότερα, για το mirTarget2 ως αρχικό threshold, για να είναι μια θέση στόχος λειτουργική από το ίδιο το πρόγραμμα ορίστηκε το 50 σε κλίμακα [0-100] και τα σκορ στις προβλέψεις κυμαίνονται από [50 – 100].

Το miRmap υπολογίζει ένα ενεργειακό σκορ με αποτέλεσμα όσο πιο μικρό είναι το σκορ τόσο πιο λειτουργική θεωρείται η θέση στόχος. Ως ορισμένο threshold είναι 0.684467 και οι τιμές των σκορ κυμαίνονται από [-0.663517 έως 0.684467] με καλύτερη απόδοση το -0.663517.

Το targetSpy_sens έχει ορισμένο threshold στα 0.99088 με τα σκορ των προβλέψεων να κυμαίνονται από [0.990882764647 έως 0.999997131133], ενώ Το targetSpy_spec έχει ορισμένο threshold στα 0.99930 με τις τιμές των σκορ να κυμαίνονται από [0.999300086036 έως 0.999997131133].

Το PITA_TOP έχει ορισμένο threshold 17.65 και οι τιμές γενικά κυμαίνονται από [-34.35 έως 17.65] με καλύτερη απόδοση το -34.35 και το PITA ALL έχει ορισμένο threshold 32.5 και οι τιμές γενικά κυμαίνονται από [-41.04 έως 32.5] με καλύτερη απόδοση το -41.04.

Τέλος, στο PicTar2 πρόγραμμα οι τιμές των σκορ κυμαίνονται από [6, 203].

Εύρεση κοινών συνδυασμών miRNA:genes

Εφόσον, ορίσαμε τα κάτω όρια των σκορ για κάθε αλγόριθμο, ξεκινά η διαδικασία της εύρεσης κοινών προβλέψεων μεταξύ του Test_dataset1 και του εκάστοτε συνόλου δεδομένων που περιγράφηκαν πιο πάνω.

Πιο πριν βέβαια, έχει γίνει έλεγχος των datasets όσον αφορά το format τους, και το αν δίνουν ίδιες πληροφορίες τα δύο σύνολα δεδομένων μεταξύ τους (π.χ ονόματα απ' την ίδια έκδοση Ensembl για τα γονίδια ή απ' την ίδια έκδοση miRbase για τα miRNAs). Αν αυτό δεν ισχύει τότε πρέπει να γίνει η αντιστοίχιση των ονομάτων αν είναι εφικτό.

Αρχικά, πρέπει να οριστεί το σύνολο των συνολικών προβλέψεων. Αυτό γίνεται με δύο βασικά βήματα. Αναπτύσσοντας αλγόριθμο σε Python έχοντας ως είσοδο τα δύο datasets (test_dataset1- **1ο σύνολο** και εκάστοτε dataset αλγορίθμου–**2ο σύνολο**) φιλτράρουμε το 2ο σύνολο με βάση τα κοινά miRNAs μόνο, του 1ου συνόλου. Προκύπτει ένα αποτέλεσμα (.txt) που στην ουσία είναι ένα υποσύνολο**(A)** του 2ου συνόλου.

Ακολούθως, ξανά τρέχουμε τον αλγόριθμο αλλά αυτή τη φορά με είσοδο τα δύο datasets (test_dataset1- **1ο σύνολο** και το προηγούμενο αποτέλεσμα υποσύνολο(A) –**2ο σύνολο**) και φιλτράροντας τώρα το 2ο σύνολο με τα genes του 1ου συνόλου. Στο τέλος, θα έχουμε ως έξοδο ένα υποσύνολο**(B)** του υποσυνόλου(A). Αυτό αποτελεί και το σύνολο των total predictions – ***Total set.***

Η διαδικασία αυτή γίνεται, διότι τα αποτελέσματα που κατεβάσαμε ήταν προϋπολογισμένα. Αυτό σημαίνει ότι συμπεριελάμβαναν πολλή περισσότερη πληροφορία απ' όση θα χρειαζόμασταν, σύμφωνα πάντα με τα test datasets που έχουμε στη διάθεσή μας και η σύγκριση με αυτά δε θα ήταν καθόλου αντικειμενική.
Στη συνέχεια, αναπτύσσοντας σε Python έναν αλγόριθμο που λαμβάνει ως είσοδο το test dataset και τα Total predictions που βρήκαμε από πριν, διατρέχει ανά σειρά τα δεδομένα και βρίσκει τους κοινούς συνδυασμούς miRNA και genes μαζί.

Όταν ολοκληρωθεί ο αλγόριθμος, δίνει ως αποτέλεσμα ένα νέο αρχείο (.txt) στο οποίο βρίσκονται όλες οι κοινές προβλέψεις μεταξύ των ήδη επαληθευμένων δεδομένων απ' το test_dataset1 και του εκάστοτε dataset – Total predictions των αλγορίθμων. Το αποτέλεσμα αυτό είναι στην ουσία το σύνολο των Αληθή Θετικών προβλέψεων – ***True Positive set***.

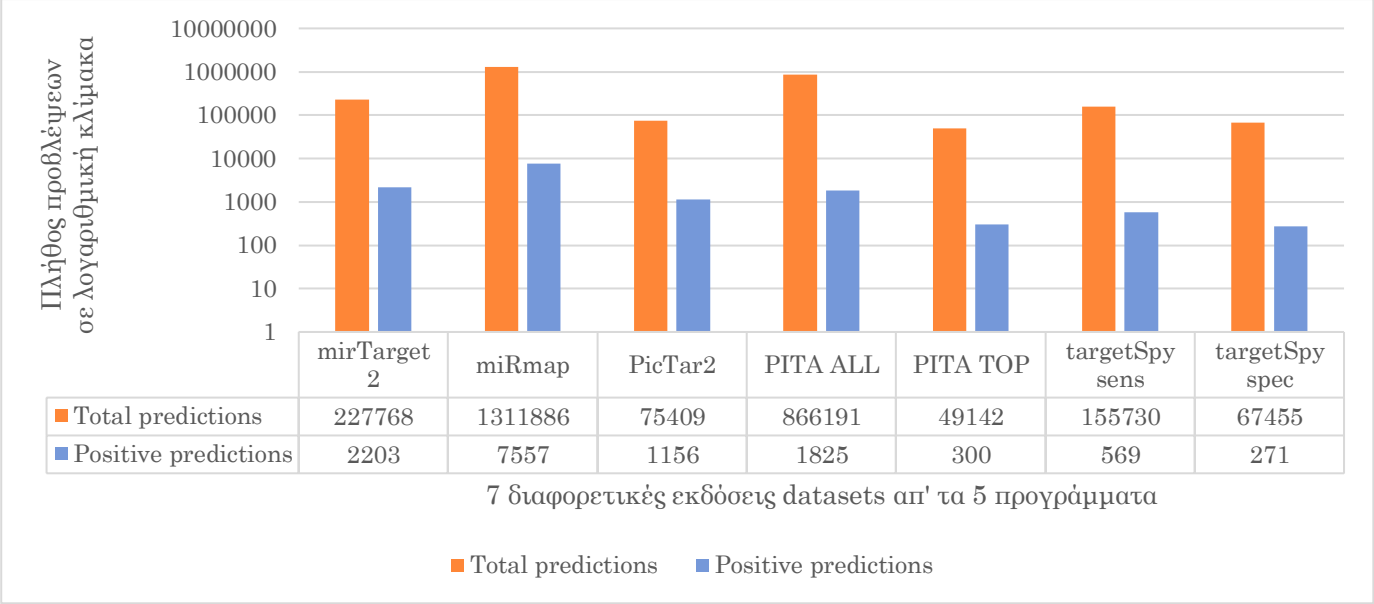
Με βάση τον Πίνακα 13 που συνοψίζει όλες τις πληροφορίες που χρειαζόμαστε για τα επικυρωμένα σύνολα δεδομένων, τα αποτελέσματα απεικονίζονται στα παρακάτω γραφήματα και πίνακες. Στον πίνακα 13, όταν αναφερόμαστε σε total predictions εννοείται ο αριθμός των αλληλεπιδράσεων miRNA:gene.

Test datasets	Total Predictions	Unique miRNAs	Unique genes	Source
Test dataset 1	25.671	886	10.436	Diana TarBase
Test dataset 2	2.175	257	1.585	Diana TarBase

Πίνακας 13 Βασικές πληροφορίες των test συνόλων δεδομένων

Πρόγραμμα	Συνολικές προβλέψεις (μεμονωμένες αλληλεπιδράσεις miRNA - γονιδίων) TOTAL SET	Αληθής θετικές προβλέψεις (κοινές μεμονωμένες αλληλεπιδράσεις miRNA-γονιδίων με το test dataset 1) TRUE POSITIVE SET
mirTarget2	227,768	2,203
miRmap	1,311,886	7,557
PicTar2	75,409	1,156
PITA_ALL	866,191	1,825
PITA_TOP	49,142	300
targetSpy_sens	155,730	569
targetSpy_spec	67,455	271

Πίνακας 14 Total set και True Positive set για κάθε πρόγραμμα, σύμφωνα με το test dataset 1

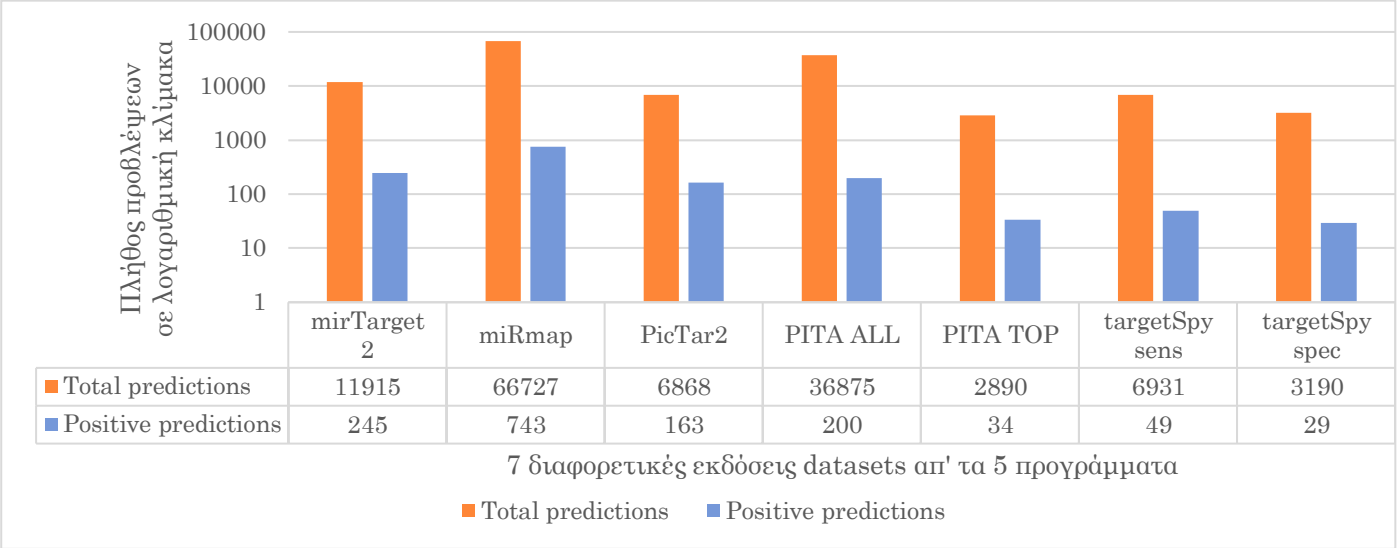


Γράφημα 1 Απεικόνιση Positive και Total predictions με βάση το Test dataset 1

Η ίδια διαδικασία ακριβώς γίνεται και για το Test_dataset 2. Δημιουργώντας έτσι ένα αντίστοιχο Γράφημα 2.

Πρόγραμμα	**Συνολικές προβλέψεις (μεμονωμένες αλληλεπιδράσεις miRNA - γονιδίων) TOTAL SET**	**Αληθής θετικές προβλέψεις (κοινές μεμονωμένες αλληλεπιδράσεις miRNA-γονιδίων με το test dataset 2) TRUE POSITIVE SET**
mirTarget2	11,915	245
miRmap	66,727	743
PicTar2	6,868	163
PITA_ALL	36,875	200
PITA_TOP	2,890	34
targetSpy_sens	6,931	49
targetSpy_spec	3,190	29

Πίνακας 15 Total set και True Positive set για κάθε πρόγραμμα, σύμφωνα με το test dataset 2



Γράφημα 2 Απεικόνιση Positive και Total predictions με βάση το Test dataset 2

Εύρεση sensitivity

Η απόδοση ενός υπολογιστικού προγράμματος αρχικά μετράται με τη μετρική Sensitivity ή Recall, δίνεται από τον παρακάτω τύπο:

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives} = \frac{True\ Positives}{All\ Positives}$$

Όπου ως True Positives ορίζεται ο αριθμός των επαληθευμένων δεδομένων miRNA:gene θέσεων στόχων που προβλέπονται από το εκάστοτε πρόγραμμα που θέλουμε να αξιολογήσουμε.

Ως False Negatives ορίζεται ο αριθμός των επαληθευμένων δεδομένων miRNA:gene θέσεων στόχων που δεν προβλέπονται από το εκάστοτε πρόγραμμα που θέλουμε να αξιολογήσουμε.

Χρησιμοποιώντας μόνο αυτή την εξίσωση, ωστόσο, ένα πρόγραμμα που προβλέπει κάθε γονίδιο ως στόχο για κάθε miRNA θα έχει την «καλύτερη» απόδοση σε σύγκριση με τα άλλα, διότι θα περιλαμβάνει όλες τις πειραματικά επαληθευμένες αλληλεπιδράσεις γονιδίου στόχου miRNA. Ωστόσο, φυσικά, θα περιλαμβάνει επίσης έναν τεράστιο αριθμό ψευδών προβλέψεων. Για το λόγο αυτό, είναι επίσης απαραίτητο να υπολογιστεί το specificity ή ο 'false positive rate'.

Ο ψευδώς θετικός ρυθμός ορίζεται ως το ποσοστό όλων των αρνητικών (πειραματικά αντικρουόμενων αλληλεπιδράσεων γονιδίου miRNA-στόχου) που προβλέπονται λανθασμένα.

Πρόγραμμα	**Ποσοστό πειραματικά επαληθευμένων προβλέψεων miRNA θέσεων στόχων (Sensitivity)**	**Total προβλέψεις miRNA θέσεων στόχων (Total set)**
1.mirTarget2	8.6%	227,768
2.miRmap	29%	1,311,886
3.PicTar2	4.5%	75,409
4.PITA_ALL	7.1%	866,191
5.PITA_TOP	1.17%	49,142

6.targetSpy_sens	2.2%	155,730
7.targetSpy_spec	1%	67,455

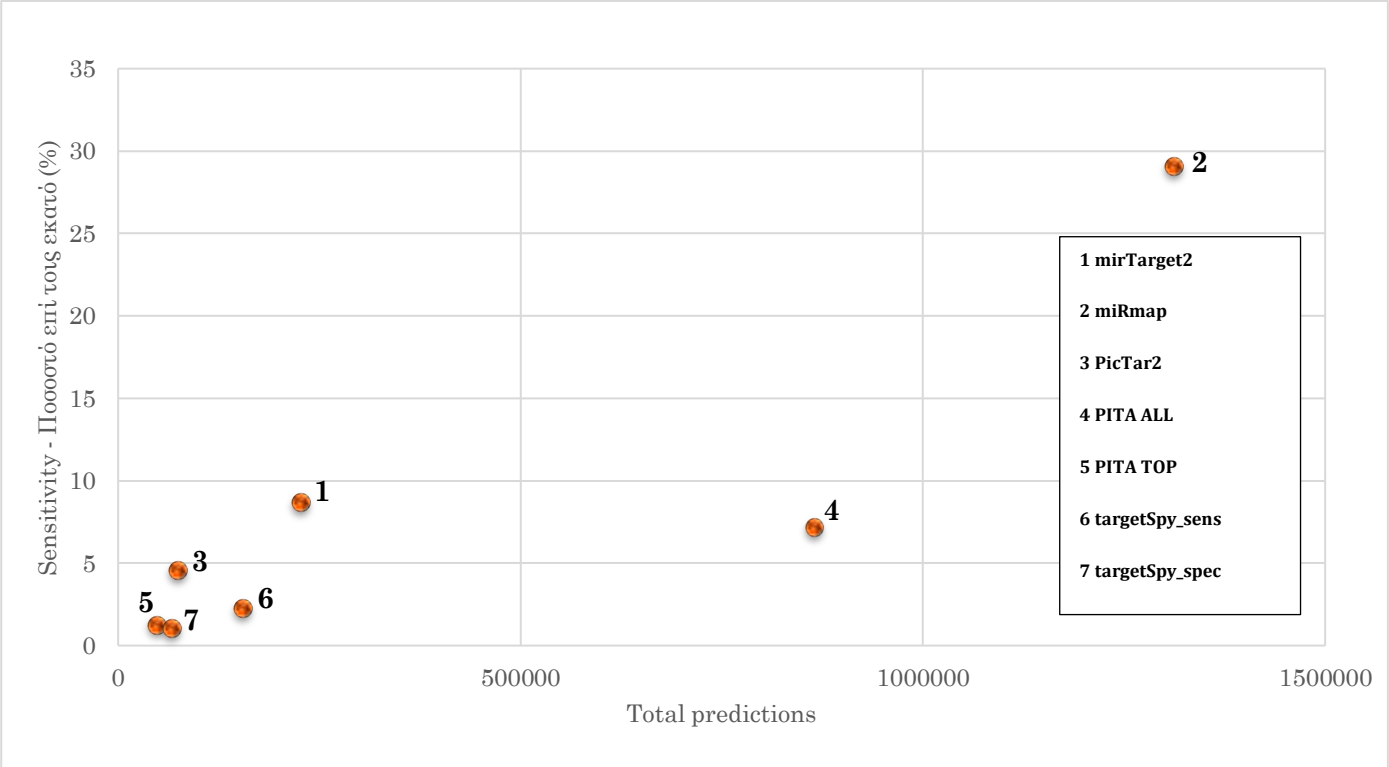
Πίνακας 16 Sensitivity και ο αριθμός των total predictions για κάθε πρόγραμμα, σύμφωνα με το test dataset 1

Πρόγραμμα	**Ποσοστό πειραματικά επαληθευμένων προβλέψεων miRNA θέσεων στόχων (Sensitivity)**	**Total προβλέψεις miRNA θέσεων στόχων (Total set)**
1.mirTarget2	11.3%	11,915
2.miRmap	34%	66,727
3.PicTar2	7.5%	6,868
4.PITA_ALL	9.2%	3,6875
5.PITA_TOP	1.6%	2,890
6.targetSpy_sens	2.3%	6,931
7.targetSpy_spec	1.3%	3,190

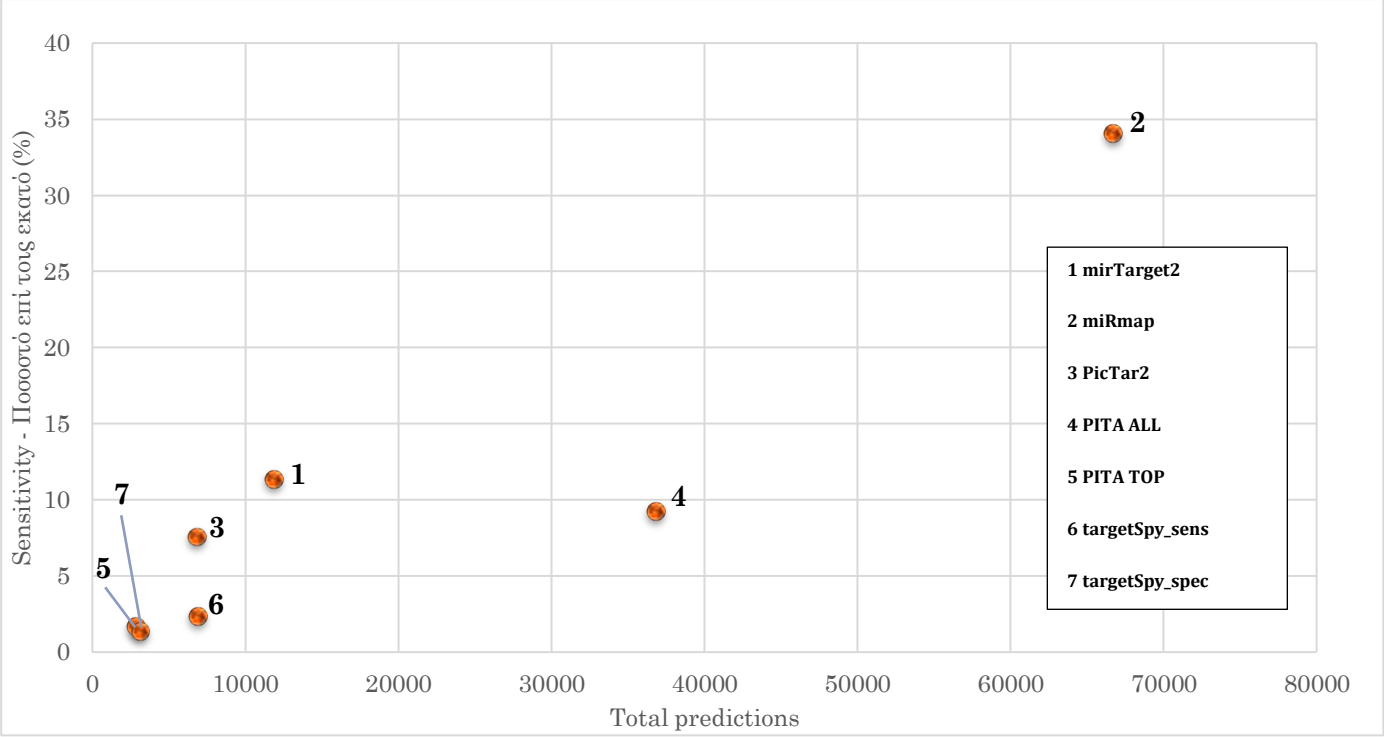
Πίνακας 17 Sensitivity και ο αριθμός των total predictions για κάθε πρόγραμμα, σύμφωνα με το test dataset 2

Κάθε γραμμή των παραπάνω πινάκων 16 & 17 αντιστοιχεί σε ένα σημείο στα Γραφήματα 3 & 4, αντίστοιχα.

Τα υπάρχοντα προγράμματα πρόβλεψης στόχων βρίσκονται σε ένα φάσμα απόδοσης το οποίο αποκαλύπτει την ανταπόκριση ευαισθησίας-εξειδίκευσης sensitivity-specificity του καθενός. Τα Γραφήματα 3 & 4 παρέχουν μια γραφική παράσταση αυτού του φάσματος απόδοσης.



Γράφημα 3 Το φάσμα απόδοσης των προγραμμάτων πρόβλεψης στόχων. Μια γραφική παράσταση Sensitivity σε σχέση με τον αριθμό των συνολικών προβλέψεων του γονιδίου στόχου miRNA, χρησιμοποιώντας το test dataset 1.



Γράφημα 4 Το φάσμα απόδοσης των προγραμμάτων πρόβλεψης στόχων. Μια γραφική παράσταση Sensitivity σε σχέση με τον αριθμό των συνολικών προβλέψεων του γονιδίου στόχου miRNA, χρησιμοποιώντας το test dataset 2.

Όσον αφορά τα Γραφήματα 3 και 4 παρατηρούμε ότι και για τα δύο test datasets ισχύει ότι ο miRmap(2) έχει το υψηλότερο sensitivity αλλά το μικρότερο specificity μιας και όσο μεγαλύτερος ο αριθμός των Total predictions σε σχέση με τα TP (true positive) τόσο μικραίνει το specificity.

Στη συνέχεια, 2ος έρχεται ο mirTarget2 (1) με το ακριβώς επόμενο μεγαλύτερο sensitivity και με αρκετά μεγαλύτερο specificity σε σχέση με τον miRmap, αλλά συνεχίζεται να θεωρείται ότι το specificity είναι μικρό.

Σίγουρα, ο mirTarget2 (1) είναι καλύτερος σε σύγκριση με τον PITA ALL (4), διότι ο (1) παρουσιάζει μεγαλύτερο specificity με σχεδόν ίδιο sensitivity.

Στους δύο αλγορίθμους miRmap (2) και PITA ALL(4), το γεγονός ότι έχουν μεγάλο sensitivity σε μεγάλο πλήθος προβλέψεων (μικρό specificity) σημαίνει ότι προβλέπουν αρκετά θετικά αλλά και ένα πολύ μεγαλύτερο ποσοστό αρνητικών ή ψευδών θετικών.

Για τα δύο test datasets, το μεγαλύτερο specificity το έχει ο PITA_TOP(5) με το μικρότερο sensitivity μαζί με το targetSpy_spec(7).

Τα αποτελέσματα επιβεβαιώνουν το γεγονός ότι τα προγράμματα με μεγάλο sensitivity δε σημαίνει ότι αποτελούν και τα αποδοτικότερα προγράμματα σε σύγκριση με τα υπόλοιπα και αυτό διότι μπορεί να προβλέπουν αρκετά Positives ή ακόμη και όλα σύμφωνα με κάποιο test σύνολο επαληθευμένων δεδομένων αλλά μέσα στις προβλέψεις τους να υπάρχει και μεγάλο ποσοστό ψευδών θετικών, γεγονός που φαίνεται με το πόσο υψηλό είναι το specificity.

Αντίστροφα, το γεγονός ότι τα προγράμματα με μεγάλο specificity δε σημαίνει ότι αποτελούν και τα αποδοτικότερα προγράμματα σε σύγκριση με τα υπόλοιπα και αυτό διότι μπορεί να προβλέπουν λίγες

σχετικά συνολικές προβλέψεις και να μειώνεται το ποσοστό των ψευδών θετικών αλλά να παραλείπουν αρκετά μεγάλο ποσοστό από τα Positives.

Το ζητούμενο είναι να έχουν και τις δύο μετρικές υψηλές όσο είναι δυνατόν, για να μπορέσουμε να αποφανθούμε στο αν ένα πρόγραμμα έχει όντως «καλή» απόδοση ή όχι.

Το mirTarget2 (1) βρίσκεται σε καλύτερη θέση σε σχέση με τα υπόλοιπα προγράμματα, εννοώντας ότι σε ένα σχετικά μέτριο σύνολο προβλέψεων προβλέπει και ένα σχετικά καλό ποσοστό θετικών, αλλά και αυτό από την πλευρά του έχει αρκετά μικρό specificity σε σχέση με το sensitivity, δηλαδή χάνει αρκετά από τα θετικά.

Το σίγουρο είναι, ότι :

- ο mirTarget2 (1) είναι καλύτερος από τον PITA ALL (4)
- ο PicTar2 (3) είναι καλύτερος από τους targetSpy_sens (6), targetSpy_spec (7), PITA_TOP (5) διότι παρουσιάζει καλύτερο sensitivity με βάση στα ίδια επίπεδα specificity, από τους υπόλοιπους.
- ο mirTarget2 (1) έχει καλύτερο sensitivity από τον PicTar2 (3) με λίγο μικρότερο specificity, όμως.
- ο miRmap (2) ενώ παρουσιάζει το μεγαλύτερο sensitivity, χάνει τη θέση του στην κατάταξη των προγραμμάτων εξαιτίας του μικρότερου specificity.

Ίσως, στο 2ο στάδιο της αξιολόγησης που θα ορίσουμε εκ νέου thresholds για το κάθε πρόγραμμα να δούμε διαφορές στις αποδόσεις των προγραμμάτων.

2η προσέγγιση αξιολόγησης – ΟΜΟΙΟΤΗΤΑ:

Στη συνέχεια, έγιναν οι κατάλληλες διαδικασίες για να μπορέσουμε να απεικονίσουμε τη συσχέτιση – ομοιότητα (similarity) των αλγορίθμων αυτών σύμφωνα με τις έγκυρες προβλέψεις τους (**TP set)** με κάποιο τον πίνακα συσχέτισης/ομοιότητας.

Οι ακόλουθοι πίνακες βασίζονται στις κοινές αλληλεπιδράσεις miRNA:genes μεταξύ των προγραμμάτων. Πιο συγκεκριμένα, συγκρίνοντας τα σύνολα True Positive των προγραμμάτων, το κάθε κελί στους πίνακες 18 & 19 αποτελεί το πλήθος των κοινών TP μεταξύ από κάθε ζεύγος αλγορίθμων.

	mirTarget2	**PicTar2**	**PITA ALL**	**PITA TOP**	**targetSpy sens**	**targetSpy spec**
mirTarget2	-	474	308	159	210	112
PicTar2	474	-	128	74	71	42
PITA ALL	308	128	-	300	457	220
PITA TOP	159	74	300	-	155	78
targetSpy sens	210	71	457	155	-	271
targetSpy spec	112	42	220	78	271	-

Πίνακας 18 Πίνακας συσχετίσεων 6 διαφορετικών εκδόσεων των αλγορίθμων σύμφωνα με το test dataset 1

	mirTarget2	PicTar2	PITA ALL	PITA TOP	targetSpy sens	targetSpy spec
mirTarget2	-	79	25	16	18	12
PicTar2	79	-	15	7	8	4
PITA ALL	25	15	-	34	43	27
PITA TOP	16	7	34	-	18	11
targetSpy sens	18	8	43	18	-	29
targetSpy spec	12	4	27	11	29	-

Πίνακας 19 Πίνακας συσχετίσεων 6 διαφορετικών εκδόσεων των αλγορίθμων σύμφωνα με το test dataset 2

Εύρεση Jaccard index

Ο δείκτης Jaccard, επίσης γνωστός ως ο συντελεστής ομοιότητας Jaccard, χρησιμοποιείται στατιστικά για τη σύγκριση της ομοιότητας και της ποικιλομορφίας των συνόλων δειγμάτων. Ο συντελεστής Jaccard μετρά την ομοιότητα μεταξύ των συνόλων πεπερασμένων δειγμάτων και ορίζεται ως το μέγεθος της τομής διαιρούμενο με το μέγεθος της ένωσης των συνόλων δειγμάτων:

$$J(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \qquad , 0 \leq J(A,B) \leq 1.$$

Αν A και B είναι κενά, τότε J(A,B) = 1.

Το ποσοστό ομοιότητας – similarity προκύπτει με τον πολλαπλασιασμό αυτού του δείκτη με το 100.

Σύμφωνα με τους πίνακες 18 & 19 και με την εξίσωση εύρεσης του Jaccard δείκτη καταλήγουμε στους παρακάτω πίνακες ξεχωριστά σύμφωνα με κάθε test dataset.

	mirTarget2	PicTar2	PITA ALL	PITA TOP	targetSpy sens	targetSpy spec
mirTarget2	-	474 **(16.4%)**	308 **(8.3%)**	159 **(6.8%)**	210 **(8.2%)**	112 **(4.7%)**
PicTar2	474 **(16.4%)**	-	128 **(4.5%)**	74 **(5.4%)**	71 **(4.3%)**	42 **(3%)**
PITA ALL	308 **(8.3%)**	128 **(4.5%)**	-	300 **(16.4%)**	457 **(23.6%)**	220 **(11.7%)**
PITA TOP	159 **(6.8%)**	74 **(5.4%)**	300 **(16.4%)**	-	155 **(21.7%)**	78 **(15.8%)**
targetSpy sens	210 **(8.2%)**	71 **(4.3%)**	457 **(23.6%)**	155 **(21.7%)**	-	271 **(47.6%)**
targetSpy spec	112 **(4.7%)**	42 **(3%)**	220 **(11.7%)**	78 **(15.8%)**	271 **(47.6%)**	-

Πίνακας 20 Πίνακας ομοιότητας των προγραμμάτων με βάση τον Jaccard δείκτη, σύμφωνα με το test dataset 1

Ακολουθεί ένα παράδειγμα εύρεσης του ποσοστού ομοιότητας Jaccard δύο προγραμμάτων,

Έστω ότι θέλουμε να βρούμε το ποσοστό ομοιότητας των αποτελεσμάτων μεταξύ του mirTarget2 και PicTar2.

— Ο mirTarget2 (1), υπολογίσαμε πριν ότι έχει 2.203 μεμονωμένες κοινές αλληλεπιδράσεις miRNA-γονιδίων με το test dataset 1 [πρόκειται για το True Positive set που υπολογίστηκε στον πίνακα 14].

— Ο PicTar2 (3) έχει 1.156 μεμονωμένες κοινές αλληλεπιδράσεις miRNA-γονιδίων με το test dataset 1.

Μεταξύ αυτών των δύο συνόλων, τα κοινά είναι 474. Επομένως, προκύπτει ότι ο δείκτης Jaccard ισούται με J((1),(3)) = 474 / (2203 + 1156 - 474) = 0.164. Το ποσοστό ομοιότητας Jaccard είναι 16.4%.

	mirTarget2	**PicTar2**	**PITA ALL**	**PITA TOP**	**targetSpy sens**	**targetSpy spec**
mirTarget2	-	79 **(24%)**	25 **(6%)**	16 **(6%)**	18 **(23.7%)**	12 **(4.6%)**
PicTar2	79 **(24%)**	-	15 **(4.3%)**	7 **(3.7%)**	8 **(4%)**	4 **(2.1%)**
PITA ALL	25 **(6%)**	15 **(4.3%)**	-	34 **(17%)**	43 **(20.9%)**	27 **(13.4%)**
PITA TOP	16 **(6%)**	7 **(3.7%)**	34 **(17%)**	-	18 **(27.7%)**	11 **(21.1%)**
targetSpy sens	18 **(23.7%)**	8 **(4%)**	43 **(20.9%)**	18 **(27.7%)**	-	29 **(59.1%)**
targetSpy spec	12 **(4.6%)**	4 **(2.1%)**	27 **(13.4%)**	11 **(21.1%)**	29 **(59.1%)**	-

Πίνακας 21 Πίνακας ομοιότητας των προγραμμάτων με βάση τον Jaccard δείκτη, σύμφωνα με το test dataset 2

Παρατηρείται ότι τη μεγαλύτερη ομοιότητα στα αποτελέσματά τους όσον αφορά τα True Positives, την έχουν ο targetSpy_spec με targetSpy_sens στα 47.6% σύμφωνα με το test dataset 1 και στα 59.1% σύμφωνα με το test dataset 2.

- Σύμφωνα με το test dataset 1, ακριβώς μετά, έρχονται ο targetSpy_sens με τον PITA ALL στα 23.6%.

Στα 21.7% ο targetSpy_sens με τον PITA TOP και στα 16.4% PicTar2 με mirTarget2 και PITA ALL με PITA TOP.

- Ενώ σύμφωνα με το test dataset 2, ακριβώς μετά, έρχονται ο targetSpy_spec με τον PITA TOP στα 27.7%.

Στα 24% ο PicTar2 με mirTarget2. Στα 23.7% ο targetSpy_sens με τον mirTarget 2 και στα 21.1% ο targetSpy_spec με PITA TOP καθώς και σε κοντινά ποσοστά με 20.9% ο PITA ALL με targetSpy_sens.

2ο στάδιο αξιολόγησης

Εφόσον, συγκρίθηκαν τα προγράμματα με βάση τα σύνολα δεδομένων με κατώφλι αυτό που ορίζουν και οι αντίστοιχες δημοσιεύσεις τους, ακολουθεί ο ορισμός των κατωφλίων όσον αφορά το σκορ των προβλέψεων για κάθε πρόγραμμα ξεχωριστά με σκοπό μια πιο αντικειμενική σύγκριση.

Ο ορισμός των κατωφλίων έγινε με πειραματικό τρόπο και παρατηρώντας αρχικά τι είναι αυτό που επιζητάμε. Στην ουσία, για να μπορεί να υπάρξει μια αντικειμενική σύγκριση των προγραμμάτων θα πρέπει να εξισορροπηθεί ο αριθμός των προβλέψεων μεταξύ τους, έτσι ώστε σύμφωνα με ένα παρόμοιο πλήθος προβλέψεων να αξιολογηθεί η απόδοσή τους.

Αυτό είναι αναγκαίο, διότι προηγουμένως μπορεί να εξήχθησαν κάποια αποτελέσματα και να βγήκαν κάποια συμπεράσματα όμως, δεν είναι πλήρως αντικειμενικά, διότι το πλήθος των ολικών προβλέψεων διαφέρει από κάθε πρόγραμμα, καθώς και τα αποτελέσματα των διαφορετικών προγραμμάτων σχετίζονται πολλές φορές με το πλήθος των ολικών προβλέψεων ή με το πλήθος των προβλέψεων ανά microRNA. Για παράδειγμα, μπορεί ο miRmap να φαίνεται ότι προβλέπει αρκετά θετικά και το sensitivity να αυξάνεται αντίστοιχα, αλλά σε σχέση με το μεγάλο πλήθος προβλέψεών του να χάνεται το specificity, ή αντίστροφα ο targetSpy να προβλέπει ένα μικρό πλήθος δεδομένων και να αυξάνεται το specificity, αλλά εξαιτίας αυτού του μικρού πλήθους δεδομένων να χάνει αρκετά θετικά.

Το γεγονός ότι τα προγράμματα PITA_TOP και targetSpy_spec προβλέπουν το πολύ έως 50.000 και 60.000 περίπου, αντίστοιχα, δε μας επιτρέπει να ορίσουμε νέα κατώφλια μιας και ήδη ο αριθμός των προβλέψεων τους είναι σχετικά μικρός.

Η πειραματική διαδικασία του ορισμού νέων κατωφλίων έγινε για τα υπόλοιπα προγράμματα και προέκυψαν οι πίνακες 23 – 27.

Βέβαια, πρέπει να αναφερθεί ότι για τον ορισμό των κατωφλίων παίζουν και άλλοι παράγοντες ρόλο αλλά σε αυτή την εργασία και επειδή κάποια απ' τα προγράμματα βασίζονται σε συγκεκριμένα προϋπολογισμένα σύνολα δεδομένων θα αρκεστούμε σε αυτή τη σκέψη.

Προσέγγιση αξιολόγησης – ΑΠΟΔΟΣΗ:

Στους παρακάτω πίνακες εμφανίζονται για κάθε πρόγραμμα ξεχωριστά, οι τιμές των διαφορετικών κατωφλίων που δοκιμάστηκαν, οι αντίστοιχες τιμές των ολικών προβλέψεων (Total set), ολικών προβλέψεων ανά miRNA, αληθή θετικών αποτελεσμάτων (True Positive set), και το ποσοστό sensitivity.

Με βάση τον πίνακα 22, ο οποίος δείχνει για κάθε πρόγραμμα το πλήθος των miRNAs για τα οποία έχει εξάγει αποτελέσματα, προκύπτει το πλήθος των ολικών προβλέψεων ανά miRNA.

Για παράδειγμα, για τον mirTarget2 για να βρούμε τις ολικές προβλέψεις ανά microRNA, διαιρέσαμε το total predictions του πίνακα 2 με το unique miRNAs του πίνακα 1. Για το 1ο κατώφλι 179,922 / 883 ≃203 ολικές προβλέψεις ανά microRNA, η ίδια διαδικασία και για τα υπόλοιπα κατώφλια.

Πρόγραμμα	Unique miRNAs
mirTarget2	883
miRmap	886
PicTar2	566
PITA_ALL	301
targetSpy_sens	307

Πίνακας 22 Πλήθος miRNAs που εξάγουν αποτελέσματα, για κάθε πρόγραμμα

mirTarget2 κατώφλια	Total Predictions	Total Predictions για κάθε miRNA	True Positives	Sensitivity
55	179922	203	1813	7%
60	141901	160	1488	5.8%
65	108072	122	1207	4.7%
70	80092	90	942	3.6%
75	57003	64	688	2.7%
80	40348	45	524	2%

Πίνακας 23 Έξι διαφορετικά πειράματα του mirTarget2 ανάλογα με διάφορες τιμές κατωφλίων

miRmap κατώφλια	Total Predictions	Total Predictions για κάθε miRNA	True Positives	Sensitivity
0.40888	1075235	1213	6560	≃25%

-0.053491	762563	860	5126	19.9%
-0.128015	454971	513	3510	13.6%
-0.20714	177590	200	1578	6.14%
-0.25246	80542	≃91	777	3%
-0.26791	58728	66	573	2.23%
-0.28505	40528	45	395	1.5%

Πίνακας 24 Εφτά διαφορετικά πειράματα του miRmap ανάλογα με διάφορες τιμές κατωφλίων

PicTar2 κατώφλια	Total Predictions	Total Predictions για κάθε miRNA	True Positives	Sensitivity
7	71155	125	1112	4.3%
8	65034	114	1014	3.9%
9	59370	104	940	3.6%
10	54290	95	870	3.4%
11	49639	87	804	3.1%
13	41536	73	682	2.6%

Πίνακας 25 Έξι διαφορετικά πειράματα του Pictar2 ανάλογα με διάφορες τιμές κατωφλίων

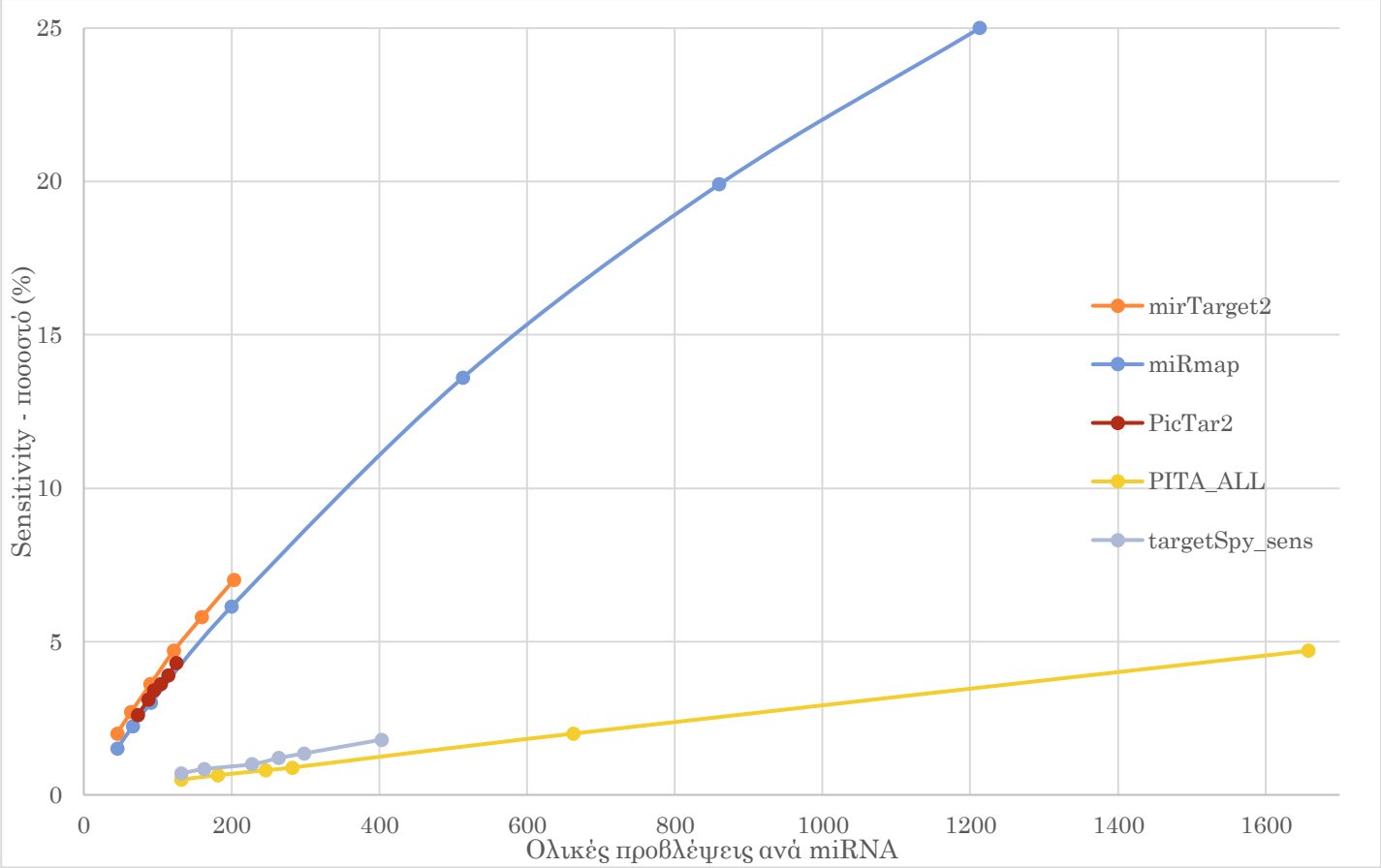
PITA ALL κατώφλια	Total Predictions	Total Predictions για κάθε miRNA	True Positives	Sensitivity
-4.77	499253	1658	1207	4.7%
-9.37	199547	663	539	2%
-12.53	85093	282	229	0.89%
-13	74047	246	206	0.8%
-14	54657	181	164	0.64%
-15	39838	132	127	0.5%

Πίνακας 26 Έξι διαφορετικά πειράματα του PITA_ALL ανάλογα με διάφορες τιμές κατωφλίων

targeStpy_sens κατώφλια	Total Predictions	Total Predictions για κάθε miRNA	True Positives	Sensitivity
0.9957028	123775	403	459	1.8%
0.9984270	91502	298	346	1.35%
0.99881423	81133	264	311	1.2%
0.9991257	70116	228	276	1%
0.9995421	50136	163	216	0.84%
0.99968	40529	132	174	0.7%

Πίνακας 27 Έξι διαφορετικά πειράματα του targetSpy_sens ανάλογα με διάφορες τιμές κατωφλίων

Σύμφωνα με τους Πίνακες 23 – 27, δημιουργήθηκε το παρακάτω Γράφημα 5. Πρόκειται για μια σύγκριση των προγραμμάτων με βάση τα διαφορετικά κατώφλια που δοκιμάστηκαν, όσον αφορά το πλήθος των ολικών προβλέψεων ανά microRNA και το ποσοστό ευαισθησίας – Sensitivity. Η κάθε γραμμή αντιστοιχεί και σ' ένα πρόγραμμα, Πίνακα 2 έως 6. Το κάθε σημείο της γραμμής αντιστοιχεί και σε μια γραμμή των αντίστοιχων πινάκων.



Γράφημα 5 Μια γραφική παράσταση Sensitivity σε σχέση με τον αριθμό των συνολικών προβλέψεων ανά miRNA για κάθε διαφορετικό κατώφλι, χρησιμοποιώντας το test dataset 1.

Παρατηρείται ότι ο PITA ALL και targetSpy_sens παρουσιάζουν μικρά ποσοστά sensitivities σε σχέση με υψηλές τιμές ολικών προβλέψεων ανά miRNA κάτι που δεν είναι επιθυμητό.

Για παράδειγμα, το μεγαλύτερο ποσοστό Sensitivity του PITA ALL είναι το 4.7% στις 1658 προβλέψεις ανά microRNA, ενώ παράλληλα το ίδιο ποσοστό sensitivity ο mirTarget2 το παρουσιάζει στις 122 προβλέψεις ανά microRNA και ο PicTar2 (με 4.3%) στις 125 προβλέψεις ανά miRNA. Διαπιστώνουμε, λοιπόν εύκολα ότι ο mirTarget2 και ο PicTar2 είναι αρκετά πιο αποδοτικοί σε σχέση με PITA ALL και targetSpy_sens.

Επιπρόσθετα, οι PicTar2 και mirTarget2 έχουν πιο υψηλά sensitivities με βάση τις ίδιες τιμές περίπου σε προβλέψεις ανά miRNA σε σύγκριση με τον miRmap. Ο miRmap παρουσιάζει αρκετά υψηλά ποσοστά sensitivities στη συνέχεια αλλά με βάση αρκετά μεγάλες τιμές προβλέψεων ανά miRNA, γεγονός που τον καθιστά ανακριβή σύμφωνα πάντα με τα σύνολα επαληθευμένων δεδομένων που έχουν συμπεριληφθεί σε αυτή την εργασία.

Τέλος, με αυτά τα δεδομένα ο mirTarget2 και ο PicTar2 συνεχίζουν να κρατούν τις πρώτες θέσεις όσον αφορά την απόδοσή τους. Η σχέση όμως μεταξύ τους δεν είναι ακριβής και σε κάποιο μελλοντικό σημείο θα χρειαζόταν να πραγματοποιηθούν περεταίρω συγκρίσεις, ανάμεσά τους.

5. ΚΕΦΑΛΑΙΟ – ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

5.1 ΣΥΜΠΕΡΑΣΜΑΤΑ MIRTARGET2

- Χρησιμοποιήθηκε ένα μεγάλο σύνολο δεδομένων μικροσυστοιχιών για την εκπαίδευση της επιλογής χαρακτηριστικών πρόβλεψης στόχων.
- Η γενική συμφωνία στην επιλογή κάποιων χαρακτηριστικών με προηγούμενες μελέτες παρέχει πρόσθετη υποστήριξη για την εγκυρότητα της χρήσης δεδομένων μεταγραφικού προφίλ για την εκπαίδευση μοντέλων.
- Εντοπίστηκαν νέα χαρακτηριστικά και με τον τρόπο αυτό η έλλειψη ορισμένων χαρακτηριστικών θα μπορούσε να αντισταθμιστεί από την παρουσία άλλων σημαντικών χαρακτηριστικών για τον προσδιορισμό του στόχου.
- Ο αλγόριθμος αναγνώρισε σωστά πάνω από το 90% των αρνητικών δειγμάτων εκπαίδευσης. Εντούτοις, μόνο τα μισά από τα θετικά δείγματα εκπαίδευσης εντοπίστηκαν.

5.2 ΣΥΜΠΕΡΑΣΜΑΤΑ MIRMAP

- Εξετάσανε την απόδοση 12 χαρακτηριστικών για να προβλέπουν ανεξάρτητα την ισχύ της καταστολής του miRNA σε στοχευμένα mRNAs και τα συνδυάσανε σε ένα γραμμικό μοντέλο, το λεγόμενο miRmap.
- Αυτή η προσέγγιση επέτρεψε να εκτιμήσουνε την ακρίβεια των χαρακτηριστικών για να ταξινομήσουνε τους στόχους miRNA και να αποφύγουνε την επιλογή ενός κατωφλίου ή τον ορισμό ενός αρνητικού συνόλου δεδομένων.
- Συνολικά, τα συνδυασμένα χαρακτηριστικά προβλέπουν με μεγαλύτερη ακρίβεια τη δύναμη της καταστολής του miRNA στόχου.
- Δοκιμάσανε μια πιο περίπλοκη μέθοδο από την γραμμική παλινδρόμηση, τον κανόνα του συνόλου, αλλά δεν βελτίωσε τις προβλέψεις (τα δεδομένα δεν φαίνονται).
- Οι συσχετισμοί μεταξύ των χαρακτηριστικών και της ατομικής απόδοσής τους σε διαφορετικά σύνολα δεδομένων αποκάλυψαν 5 διαφορετικές ομάδες χαρακτηριστικών πρόβλεψης. Για παράδειγμα, ένας καλός συσχετισμός είναι “ΔG open” και “AU content”. Είναι ενδιαφέρον το γεγονός ότι το «ΔG open» είναι ανώτερο από το «AU content» στην κατάταξη.
- Άλλα εργαλεία πρόβλεψης στόχων miRNA εξετάζουν ένα μόνο ή υποσύνολο των χαρακτηριστικών αυτών. Για παράδειγμα, το PITA θεωρεί μόνο το σύνολο "ΔG total" και το PACMIT ένα συνδυασμό των "ΔG open" και "P.over binomial". Καθώς η απόδοση καθενός από αυτά τα χαρακτηριστικά είναι χαμηλότερη από τη συνδυασμένη προσέγγιση του miRmap, καθιστούν τα εργαλεία αυτά να έχουν μικρότερη προβλεπτική ισχύ.
- Η φυσική επιλογή που μετριέται είτε από το χαρακτηριστικό «BLS» είτε από το «PhyloP» είναι εξαιρετικά καλά συσχετισμένη με την ισχύ της καταστολής: οι επιλεγμένοι τόποι στόχοι είναι επίσης περιοχές ισχυρότερης καταστολής. Είναι επίσης γνωστό ότι τα παλαιότερα miRNAs έχουν υψηλότερα επίπεδα έκφρασης. Η φυσική επιλογή ενεργεί τόσο στο επίπεδο έκφρασης του miRNA όσο και στην δύναμη καταστολής για να μεγιστοποιήσει την αποτελεσματικότητα της καταστολής. Επιπλέον, μια συσχέτιση μεταξύ της προσβασιμότητας του mRNA και της διατήρησης της θέσης στόχου έχει δειχθεί στην Drosophila η οποία μπορεί να εξηγήσει εν μέρει την καλή απόδοση των χαρακτηριστικών

προσβασιμότητας («ΔG open» και «AU content») καθώς αυτή η παράμετρος επιλέγεται φυσικά .

- Η εξάρτηση μεταξύ των χαρακτηριστικών εξηγεί εν μέρει γιατί η ατομική τους απόδοση δε συμβάλλει θετικά στο γενικό μοντέλο. Τα πιθανοτικά χαρακτηριστικά συσχετίζονται επίσης με τα χαρακτηριστικά διατήρησης, αλλά συνήθως υπερέχουν από τα χαρακτηριστικά διατήρησης, ακόμη και αν μερικές φορές έχουν παρόμοια απόδοση (π.χ. τα πιθανοτικά χαρακτηριστικά είναι παρόμοια με την απόδοση «BLS» για το σύνολο δεδομένων Trans.Grimson). Από την άποψη του υπολογισμού και, κυρίως, των δεδομένων εισόδου (πολλαπλές ευθυγραμμίσεις κ.λπ.), τα πιθανοτικά χαρακτηριστικά είναι αναμφισβήτητα λιγότερο δαπανηρά από τα χαρακτηριστικά διατήρησης. Μπορούν επομένως να θεωρηθούν ως μια εναλλακτική λύση σε μια εξελικτική προσέγγιση, ειδικά για τους οργανισμούς με μακρά 3'-UTRs (μεταξύ Drosophila και ανθρώπου, η ακρίβεια πέφτει σημαντικά στη Drosophila).
- Ενώ παρατηρήσανε γενικά σταθερά αποτελέσματα μεταξύ των πειραματικών μεθόδων – transcriptomics**, polysome fractionation και proteomics, διακρίθηκαν τα πειράματα IP.
- Παρ' όλα αυτά, το μοντέλο δείχνει ότι η συλλογή περισσότερων πληροφοριών με συμπληρωματικά χαρακτηριστικά ήδη βελτιώνει σημαντικά την προβλεπτική ισχύ.

5.3 ΣΥΜΠΕΡΑΣΜΑΤΑ TARGETSPY

- Μια νέα υπολογιστική προσέγγιση για την πρόβλεψη θέσεων στόχων microRNA που δεν απαιτεί την ύπαρξη περιοχής seed, ούτε χρησιμοποιούν φυλογενετικά αποτυπώματα. Αποκομίσθηκε ένα σύνολο αντικειμενικών χαρακτηριστικών που θα χρησιμοποιηθούν για μηχανική μάθηση.
- Το TargetSpy **i)** είναι σε θέση να προβλέψει ειδικές (δηλαδή μη διατηρημένες) θέσεις στόχους, **ii)** είναι κατάλληλο για επεξεργασία ανεπαρκώς διατηρημένων / χαμηλής ποιότητας γονιδιωματικών αλληλουχιών για τις οποίες μέθοδοι που βασίζονται σε πληροφορίες σχετικά με τη διατήρηση και την πληροφορία ειδών δεν θα λειτουργούν.
- Συγκεντρώθηκαν προσεγγίσεις υπολογιστικής πρόβλεψης σε τρεις κατηγορίες, ανάλογα με τη χρήση ενός κριτηρίου αντιστοίχισης seed και της διατηρησιμότητας της περιοχής αυτής, έτσι ώστε να παρέχεται μια σύγκριση της απόδοσης μεταξύ διαφόρων προσεγγίσεων.

** Τεχνικές που χρησιμοποιούνται για τη μελέτη του μεταγραφικού κειμένου ενός οργανισμού, το άθροισμα όλων των μεταγραφών RNA του οργανισμού.

5.4 ΣΥΜΠΕΡΑΣΜΑΤΑ PITA

- Θεωρήσανε ότι εάν η προσβασιμότητα του χώρου έχει τόσο σημαντική επίδραση στις αλληλεπιδράσεις microRNA, τα γονιδιώματα μπορεί να έχουν εξελιχθεί για να αντιμετωπίσουν αυτόν τον περιορισμό, ίσως με την κατά προτίμηση τοποθέτηση στόχων σε περιοχές που έχουν ανοικτές δομές και είναι επομένως πιο προσιτές.
- Πράγματι, βρήκαν ότι οι περιοχές seed microRNA και στους τέσσερις οργανισμούς έδειξαν αξιοσημείωτη προτίμηση σε περιοχές εξαιρετικά προσιτές, σε σύγκριση με συλλογή seeds ίσου μεγέθους των οποίων οι θέσεις γονιδιώματος επιλέχθηκαν τυχαία.
- Συμπερασματικά, τα αποτελέσματα δείχνουν ότι η προσβασιμότητα στον τόπο είναι εξίσου σημαντική με την αντιστοίχιση αλληλουχίας σε seed για τον προσδιορισμό της αποτελεσματικότητας της μεταφραστικής καταστολής που προκαλείται από microRNA.
- **Εισάγεται ένα θερμοδυναμικό μοντέλο χωρίς παραμέτρους που εξηγεί αυτά τα αποτελέσματα και δείχνει ότι η προτιμησιακή τοποθέτηση θέσεων στόχου microRNA σε**

περιοχές υψηλής προσβασιμότητας είναι ένα διατηρημένο χαρακτηριστικό στα γονιδιώματα. Από τα συμπεράσματα αυτά προκύπτει ότι η θερμοδυναμική του ενδοδιαμοριακού ζευγαρώματος βάσεων αντιπροσωπεύει ένα σημαντικό τμήμα της αλληλεπίδρασης microRNA-στόχου, σύμφωνα με τις παρατηρήσεις για τις αλληλεπιδράσεις siRNA-στόχου [27].

- Ωστόσο, το μοντέλο PITA δεν εξηγεί τη συνολική διακύμανση στα πειράματα. Αυτό μπορεί εν μέρει να οφείλεται σε περιορισμούς των αλγορίθμων πρόβλεψης δομής RNA και στην ανικανότητά τους να λαμβάνουν υπόψη τις επιδράσεις που έχουν οι πρωτεΐνες δέσμευσης RNA στις δευτερογενείς δομές.
- Επομένως, θα χρειαστούν περαιτέρω πειράματα για να κατανοήσουμε πώς οι διαφορετικές πλευρές της αλληλεπίδρασης συμβάλλουν στη δύναμη της καταστολής με τη μεσολάβηση microRNA. Παρόλα αυτά, τα αποτελέσματα παρέχουν έναν σημαντικό ακρογωνιαίο λίθο για την αποκρυπτογράφηση των κανόνων που διέπουν τις αλληλεπιδράσεις του mRNA - microRNA.

5.5 ΣΥΜΠΕΡΑΣΜΑΤΑ EIMMO2

- Συγκεκριμένα, εξακολουθεί να είναι ελάχιστα κατανοητό ποιοι περιορισμοί πέρα από την αντιστοίχιση του seed miRNA καθορίζουν τη λειτουργικότητα των υποθετικών θέσεων στόχων.
- Σε αυτή τη μελέτη, αναπτύξανε μια γενική μέθοδο πρόβλεψης στόχου miRNA που επεκτείνει τις ήδη διαθέσιμες μεθόδους με διάφορους τρόπους. **Πρώτον**, αντιμετωπίζουν τις φυλογενετικές σχέσεις μεταξύ των ειδών με αυστηρό και γενικό τρόπο, χωρίς οποιεσδήποτε μετρήσιμες παραμέτρους. Δηλαδή, η Bayesian διαδικασία προσδιορίζει μοναδικά τις posterrior πιθανότητες για κάθε πρότυπο συντήρησης και τύπο seed από την άποψη των παρατηρούμενων συνθηκών συντήρησης των θέσεων στόχων για κάθε miRNA. Έτσι, σε αντίθεση με πολλές άλλες μεθόδους πρόβλεψης στόχων, οι οποίες είναι ειδικά προσαρμοσμένες ώστε να λειτουργούν σε ένα συγκεκριμένο είδος ειδών, η μέθοδος μας μπορεί να εφαρμοστεί σε οποιαδήποτε ομάδα ειδών και οι φυλογενετικές σχέσεις μεταξύ των ειδών θα ληφθούν αυτομάτως υπόψη κατά την εκτίμηση της σημασίας των προτύπων διατήρησης του τόπου. Αυτό, για παράδειγμα, θα επιτρέψει να ενημερωθούν εύκολα οι προβλέψεις, καθώς θα γίνουν διαθέσιμα περισσότερα γονιδιώματα, χωρίς να χρειάζεται να προσαρμοστεί η μέθοδος.
- Σημειώστε επίσης ότι η Bayesian διαδικασία για την ενσωμάτωση πληροφοριών από τα στατιστικά στοιχεία διατήρησης είναι γενικά ανεξάρτητη από τον ορισμό της "τοποθεσίας" που χρησιμοποιούν και μπορεί εύκολα να εφαρμοστεί σε άλλους ορισμούς των τοποθεσιών στόχων. Έτσι, αν βελτιωθεί ο ορισμός των τοποθεσιών-στόχων στο μέλλον, για παράδειγμα μέσω μιας καλύτερης κατανόησης των απαιτήσεων για τις λειτουργικές περιοχές στόχους miRNA, τότε μπορούμε εύκολα να προσαρμόσουμε τη μέθοδο για να συμπεριλάβουμε στατιστικά στοιχεία διατήρησης ουσιαστικά με τον ίδιο τρόπο. Οι γενικότερες θέσεις, δεδομένης μιας δυαδικής συνάρτησης που διακρίνει "τοποθεσίες" από "μη τοποθεσίες" σε αλληλουχίες RNA, και δίνεται ένα σύνολο "συχνοτήτων background" p (c | bg) με τις θέσεις που ορίζονται, C τυχαία, μπορούν να εφαρμόσουν την ίδια μεθοδολογία για την εκχώρηση posterrior πιθανοτήτων σε όλες τις πιθανές τοποθεσίες, ενσωματώνοντας τις πληροφορίες από τα στατιστικά στοιχεία διατήρησης αυτών των τοποθεσιών.
- **Δεύτερον**, εκτιμούν την εξέλιξη των πιέσεων επιλογής στις θέσεις-στόχους με ειδικό τρόπο για το miRNA. Αυτό επιτρέπει να αντιμετωπίζουν σωστά τα miRNA που εμφανίζονται σε διαφορετικά στάδια της εξέλιξης και των οποίων οι στόχοι μπορεί να έχουν υποστεί διαφορετικές πιέσεις επιλογής σε διαφορετικές γενεές. Συγκεκριμένα, δείχνουν ότι τα διαφορετικά miRNAs εμφανίζουν αξιοσημείωτα διαφορετικές κατανομές των λειτουργικών θέσεων στόχων σε όλο το φυλογενετικό δένδρο και παρέχουν την πρώτη ολοκληρωμένη.

- Έχουν επιδείξει επιπλέον ότι, ειδικά σε μακριές 3 'UTR αλληλουχίες σε σπονδυλωτά, οι περιοχές στόχοι miRNA δείχνουν μια σημαντική προκατάληψη προς την εμφάνιση κοντά στην αρχή και στο τέλος του 3' UTR.
- Όσον αφορά την απόδοση του αλγορίθμου, έχουν δείξει ότι στη μύγα, όπου έχουν γίνει εκτεταμένες συγκρίσεις της απόδοσης των αλγορίθμων πρόβλεψης στόχου, η μέθοδος εκτελεί τουλάχιστον όσο και τις πιο ακριβείς διαθέσιμες σήμερα μεθόδους, με υψηλή εξειδίκευση με μια σχετικά μεγάλη ποικιλία ευαισθησιών.
- Τέλος, αναπτύξανε μια μέθοδο για την αναγνώριση βιοχημικών μονοπατιών που είναι σημαντικά εμπλουτισμένα ή εξαντλημένα σε στόχους ενός συγκεκριμένου miRNA. Δείξαν ότι, για μελέτες miRNAs, αυτή η προσέγγιση ανακάμπτει τις γνωστές λειτουργικές ενώσεις.

5.6 Συμπερασματα PicTar2

- Συνοπτικά, αναπτύξανε μια υπολογιστική προσέγγιση που αναγνωρίζει με επιτυχία όχι μόνο τα γονίδια στόχου microRNA για μεμονωμένα microRNA αλλά και τους στόχους που είναι πιθανόν να ρυθμιστούν από microRNAs που συν-εκφράζονται ή ενεργούν σε μια κοινή οδό.
- Έχουν δείξει ότι οι συγκρίσεις μαζικών αλληλουχιών που χρησιμοποίησαν προηγουμένως μη διαθέσιμες ευθυγραμμίσεις σε όλο το μήκος των οκτώ ειδών σπονδυλωτών μείωσαν σημαντικά τα ψευδώς θετικά ποσοστά των προβλέψεων στόχων microRNA, επιτρέποντας στο PicTar να προβλέψει (πάνω από τον θόρυβο) κατά μέσο όρο 200 στοχευμένες μεταγραφές ανά microRNA.
- Οι συνδυαστικές προβλέψεις στόχων του PicTar για το microRNA οδήγησαν στην πειραματική επικύρωση του Mtpn ως το πρώτο γονίδιο των θηλαστικών που φαίνεται ότι ρυθμίζεται συντονισμένα από τρία microRNAs. Τα αποτελέσματα παρέχουν έτσι ένα υπολογιστικό και πειραματικό μοντέλο για τη μελέτη της μεταφραστικής γονιδιακής ρύθμισης από πολλαπλά microRNAs και μια πρώτη ματιά στην πολυπλοκότητα της μεταφραστικής γονιδιακής ρύθμισης που εκτελείται από τα microRNAs.

5.7 Μελλοντικη Εργασια

Στα πλαίσια της εργασίας και βασιζόμενοι μόνο σε δύο επαληθευμένα σύνολα δεδομένων περιορισμένου μεγέθους παρουσιάζονται κάποια βασικά αποτελέσματα που ήταν εφικτό να εξαχθούν με βάση το πλήθος των αληθή θετικών αποτελεσμάτων και το πλήθος των ολικών προβλέψεων.

Σε μελλοντικό στάδιο, σίγουρα μπορούν να γίνουν περισσότερες συγκρίσεις αυτών των προγραμμάτων, έχοντας και ως δεδομένα τα αρνητικά επαληθευμένα δεδομένα έτσι ώστε να ληφθεί υπόψη και η μετρική Precision.

Αναμφίσβητητα, οι έρευνες για την πρόβλεψη των miRNA θέσεων στόχων σε ανθρώπινα γονίδια δε θα σταματήσουν να γίνονται μιας και η σπουδαιότητά τους και η έντονη ρυθμιστική τους ιδιότητα επηρεάζει άμεσα τον ανθρώπινο οργανισμό.

Μέσα από όλη αυτή την ανάλυση αυτό που συμπεραίνουμε είναι ότι τα αντίστοιχα προγράμματα πρόβλεψης θέσεων στόχων miRNA εξαρτώνται άμεσα από τα χαρακτηριστικά των διαφόρων θέσεων στόχων που πρέπει να συλλεχθούν και να συνδυαστούν με κατάλληλο τρόπο για να μπορέσουμε να έχουμε το καλύτερο επιθυμητό αποτέλεσμα, δηλαδή να προβλέπουμε όσο το δυνατόν περισσότερους και ακριβής στόχους. Με άλλα λόγια, σε τέτοιου είδους προγράμματα είναι αναγκαίο το sensitivity αλλά και το specificity.

Η διαδικασία της καλύτερης και πιο αποτελεσματικής συλλογής χαρακτηριστικών έχει αποδειχθεί μέσα από την πράξη ότι γίνεται με τη βοήθεια της μηχανικής μάθησης. Επομένως, η μεγαλύτερη εξοικείωση με τέτοιου είδους τεχνολογία και κατ' επέκταση η ανάπτυξη προγραμμάτων βασισμένη σε νευρωνικά δίκτυα γενικότερα θα ωφελήσει ιδιαίτερα την κατάσταση που επικρατεί στον κλάδο με τα miRNAs στη βιοπληροφορική.

Επιπρόσθετα, το πρόβλημα της συλλογής χαρακτηριστικών, της ομαδοποίησής τους και κατ' επέκταση της κατάταξης των θέσεων στόχων miRNA είναι στην ουσία ένα πρόβλημα ταξινόμησης και για τέτοιου είδους προβλήματα στις μέρες μας έχουν γίνει «ειδικοί» τα νευρωνικά δίκτυα χάρη στις ιδιότητές τους.

6. ΚΕΦΑΛΑΙΟ – ΠΑΡΑΡΤΗΜΑ

Στο κεφάλαιο αυτό περιλαμβάνεται μια σύντομη περιγραφή του format για το κάθε πρόγραμμα που παράγει προβλεπόμενους στόχους – θέσεις miRNA.

Τα δεδομένα εισόδου που έχουμε στη διάθεσή μας είναι τα εξής:

Σειρά πειραμάτων 1:

Human_experimentally_validated_interactions.txt

Format

1. miRNA: Το όνομα του microRNA. Η έκδοση απ' όπου έχουμε πάρει την ονοματολογία είναι miRBase 18. http://www.mirbase.org/

2. Ensembl_Gene_id: Το Ensembl gene id του γονιδίου. Η έκδοση είναι Ensembl 83. http://dec2015.archive.ensembl.org/index.html

3. Method: Η πειραματική μεθοδολογία με την οποία έχει επιβεβαιωθεί ο στόχος του συγκεκριμένου miRNA.

4. MIMAT: Το όνομα του microRNA.

5. Gene_name: Το όνομα του γονιδίου.

Size

25,671 αλληλεπιδράσεις miRNA:γονιδίων
886 miRNAs

Σειρά πειραμάτων 2

Human_experimentally_validated_interactions_Reporter_Chimeric.txt

Format

Επιπρόσθετα με το προηγούμενο format συν :

1. chr: Το χρωμόσωμα

2. start: Η γονιδιωματική θέση που ξεκινάει η περιοχή πρόσδεσης του miRNA

3. end: Η γονιδιωματική θέση που τελειώνει η περιοχή πρόσδεσης

4. strand: Σε ποια αλυσίδα προσδένεται («+», «-»).

Size
2,175 αλληλεπιδράσεις miRNA:γονιδίων

257 miRNAs

Τα στοιχεία που επακολουθούν αφορούν μόνο στόχους για τον άνθρωπο (hsa -).

mirTarget2

Data Sets

Files

MirTarget2_v4.0_prediction_result.txt

Format

1. miRNA name από miRBase 18
2. Gene association από NCBI
3. Mirtarget2 score

Size
1,196,548 αλληλεπιδράσεις miRNA:γονιδίων

miRmap

Data Sets

Files
mirmap201202e_homsap

_targets	All predicted target sites with their genomic and 3'-UTR coordinates
_targets_1to1	All miRNA-mRNA predicted regulations
_targets_1to1_pt	Same as above, as percentiles over the prediction set
_mirnas	miRNA names and sequences
_transcripts	mRNA transcripts names and sequences

Format
mirmap201202e_homsap_mirnas.csv

1. miRNA name
2. sequence

mirmap201202e_homsap_targets.csv

1. transcript_sΠίνακας_id
2. gene_sΠίνακας_id
3. gene_name
4. transcript_chr

5. transcript_strand
6. mature_name
7. site_id
8. site_end
9. transcript2genome
10. seed_length
11. seed_mismatches_nogu
12. seed_gu
13. tgs_au
14. tgs_position
15. tgs_pairing3p
16. dg_duplex
17. dg_binding
18. dg_open
19. dg_total
20. prob_exact
21. prob_binomial
22. cons_bls
23. selec_phylop
24. mirmap_score

Total Size

1,311,886 αλληλεπιδράσεις miRNA:γονιδίων

1,921 miRNAs

targetSpy

Data Sets

Files

hsa_refseq_seed_sens.txt

hsa_refseq_seed_spec.txt

Format

1. miRNA name without species prefix (hsa-)
2. Gene accession από NCBI (RefSeq)
3. Position in 3'UTR start
4. Position in 3'UTR end
5. Sequence
6. Energy
7. TargetSpy score

Size

hsa_refseq_seed_sens.txt

155,730 αλληλεπιδράσεις miRNA:γονιδίων

hsa_refseq_seed_spec.txt

67,455 αλληλεπιδράσεις miRNA:γονιδίων

PITA

Data Sets

Files

No flank:

PITA_targets_hg18_0_0_TOP.tab

PITA_targets_hg18_0_0_ALL.tab

3/15 flank:

PITA_targets_hg18_3_15_TOP.tab

PITA_targets_hg18_3_15_ALL.tab

Format

1. Gene association από NCBI (RefSeq)
2. Gene name
3. microRNA name
4. Sites, αριθμός από sites
5. PITA score

Size

PITA_targets_hg18_0_0_TOP.tab, PITA_targets_hg18_3_15_TOP.tab

49.142 αλληλεπιδράσεις miRNA:γονιδίων

PITA_targets_hg18_0_0_ALL.tab, PITA_targets_hg18_3_15_ALL.tab

866,191 αλληλεπιδράσεις miRNA:γονιδίων

EIMMO2

Data Sets

Δεν υπάρχει διαθέσιμη ιστοσελίδα για την δημόσια χρήση του αλγορίθμου αυτού.

PicTar2

Data Sets

Files

pictar2_hg19.bed

Format

1. Chromosome
2. Start
3. End
4. miRNA name
5. number of anchor sites

6. Strand

Size

pictar2_hg19.bed

75,409 αλληλεπιδράσεις miRNA:γονιδίων

7. ΚΕΦΑΛΑΙΟ - ΒΙΒΛΙΟΓΡΑΦΙΑ

ΑΝΑΦΟΡΕΣ

mirTarget2

1) Gaidatzis,D. . (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. BMC Bioinformatics, 8, 69.
2) Griffiths-Jones,S. . (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res., 34, D140–D144.
3) Grimson,A. . (2007) MicroRNA Targeting specificity in mammals: determinants beyond seed pairing. Mol. Cell, 27, 91–105.
4) He,L. and Hannon,G.J. (2004) MicroRNAs: small RNAs with a big role in gene regulation. Nat. Rev. Genet., 5, 522–531.
5) Hofacker,I.L. (2003) Vienna RNA secondary structure server. Nucleic Acids Res., 31, 3429–3431.
6) Kertesz,M. . (2007) The role of site accessibility in microRNA target recognition. Nat. Genet., 39, 1278–1284.
7) Ioannis S. Vlachos, Artemis G. Hatzigeorgiou, (2013) Online resources for miRNA analysis
8) Wang X, El Naqa IM (2008) Prediction of both conserved and nonconserved microRNA targets in animals. Bioinformatics, Oxford, England.

miRmap

9) Selbach,M., Schwanha¨ usser,B., Thierfelder,N., Fang,Z., Khanin,R. and Rajewsky,N. (2008) Widespread changes in protein synthesis induced by microRNAs. Nature, 455, 58–63.
10) Chi,S.W., Zang,J.B., Mele,A. and Darnell,R.B. (2009) Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. Nature, 460, 479–486.
11) Friedman,R.C., Farh,K.K., Burge,C.B. and Bartel,D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. Genome Res., 19, 92–105.
12) Wen,J., Parker,B.J., Jacobsen,A. and Krogh,A. (2011) MicroRNA transfection and AGO-bound CLIP-seq data sets reveal distinct determinants of miRNA action. RNA, 17, 820–834.
13) Linsley,P.S., Schelter,J., Burchard,J., Kibukawa,M., Martin,M.M., Bartz,S.R., Johnson,J.M., Cummins,J.M., Raymond,C.K., Dai,H. . (2007) Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. Mol. Cell. Biol., 27, 2240–2252.
14) Hendrickson,D.G., Hogan,D.J., McCullough,H.L., Myers,J.W., Herschlag,D., Ferrell,J.E. and Brown,P.O. (2009) Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. PLoS Biol., 7, e1000238.
15) Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI reference sequences: current status, policy and new initiatives. Nucleic Acids Res., 37, D32–D36.

16) Stark,A., Lin,M.F., Kheradpour,P., Pedersen,J.S., Parts,L., Carlson,J.W., Crosby,M.A., Rasmussen,M.D., Roy,S., Deoras,A.N. . (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature, 450, 219–232.
17) Guo,H., Ingolia,N.T., Weissman,J.S. and Bartel,D.P. (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature, 466, 835–840.
18) Huntzinger,E. and Izaurralde,E. (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat. Rev. Genet., 12, 99 110.
19) Charles E. Vejnar and Evgeny M. Zdobnov,(2012) miRmap: Comprehensive prediction of microRNA target repression strength, Department of Genetic Medicine and Development, University of Geneva, Swiss Institute of Bioinformatics, Imperial College London, South Kensington Campus

targetSpy

20) Witten IH, Frank E: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, 2 2005.
21) Yousef M, Jung S, Kossenkov AV, Showe LC, Showe MK: Naive Bayes for microRNA target predictions–machine learning for microRNA targets. Bioinformatics (Oxford, England) 2007, 23(22):2987-2992.
22) Lewis BP, Burge CB, Bartel DP: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005, 120(1):15-20.
23) Kim SK, Nam JW, Rhee JK, Lee WJ, Zhang BT: miTarget: microRNA target gene prediction using a support vector machine. BMC bioinformatics 2006, 7:411.
24) Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP: The impact of microRNAs on protein output. Nature 2008, 455(7209):64-71.
25) Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N: Widespread changes in protein synthesis induced by microRNAs. Nature 2008, 455(7209):58-63.
26) Webb GI: MultiBoosting: A Technique for Combining Boosting and Wagging. Machine Learning 2000, 40(2):159-196.
27) Daniel W. Thomson, Cameron P. Bracken and Gregory J. Goodal, Experimental strategies for microRNA target identification (2011)
28) Martin Sturm, Michael Hackenberg, David Langenberger, Dmitrij Frishman, TargetSpy: a supervised machine learning approach for microRNA target prediction, (2010)

PITA

29) Stark, A., Brennecke, J., Bushati, N., Russell, R.B. & Cohen, S.M. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3¢UTR evolution. Cell 123, 1133–1146 (2005).
30) Sethupathy, P., Corda, B. & Hatzigeorgiou, A.G. TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA 12, 192–197 (2006).
31) Rehwinkel, J. . Genome-wide analysis of mRNAs regulated by Drosha and Argonaute proteins in Drosophila melanogaster. Mol. Cell. Biol. 26, 2965–2975 (2006).
32) Alexiou P., Maragkakis M., Papadopoulos G., Reczko M., Hatzigeorgiou G., Lost in translation: an assessment and perspective for computational microRNA target identification, Institute of Molecular Oncology, Biomedical Sciences Research Center 'Alexander Fleming' (2009)

33) Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul & Eran Segal, The role of site accessibility in microRNA target recognition (2007)

EIMMO2

34) Lim L, Lau N, Garrett-Engele P, Grimson A, Schelter J, Castle J, Bartel D, Linsley P, Johnson J: Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 2005, 433:769-773.
35) Doench J, Sharp P: Specificity of microRNA target selection in translational repression. *Genes Dev* 2004, 18:504-511.Brennecke J, Stark A, Russell R, Cohen S: Principles of microRNAtarget recognition. *PLoS Biol* 2005, 3:85.
36) Lai E: Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet* 2002, 30:363-364.
37) Sethupathy P., Megraw M., Hatzigeorgiou A., A guide through present computational approaches for the identification of mammalian microRNA targets (2006)
38) Gaidatzis D., Nimwegen E., Hausser J., Zavolan M. Inference of miRNA targets using evolutionary conservation and pathway analysis, (2007)

PicTar2

39) Rajewsky, N. & Socci, N.D. Computational identification of microRNA targets. Dev. Biol. 267, 529–535 (2004).
40) Enright, A.J. . MicroRNA targets in Drosophila. Genome Biol. 5, R1 (2003).
41) Kiriakidou, M. . A combined computational-experimental approach predicts human microRNA targets. Genes Dev. 18, 1165–1178 (2004).
42) John, B. . Human MicroRNA targets. PLoS Biol. 2, e363 (2004).
43) Doench, J.G. & Sharp, P.A. Specificity of microRNA target selection in translational repression. Genes Dev. 18, 504–511 (2004).
44) Banerjee, D. & Slack, F. Control of developmental timing by small temporal RNAs: a paradigm for RNA-mediated regulation of gene expression. Bioessays 24, 119–129 (2002).
45) Reinhart, B.J. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature 24, 901–906 (2000).
46) Griffiths-Jones, S. The microRNA Registry. Nucleic Acids Res. 32, D109–D111 (2004).
47) Azra Krek, Dominic Grun, Matthew N Poy, Rachel Wolf, Lauren Rosenberg, Eric J Epstein, Philip MacMenamin, Isabelle da Piedade, Kristin C Gunsalus, Markus Stoffel & Nikolaus Rajewsky, Combinatorial microRNA target predictions (2005)
48) Pathan, M., Keerthikumar, S., Ang, C.S., Gangoda, L., Quek, C.M.J., Williamson, N.J., Mouradov, D., Sieber, O.M., Simpson, R.J., Salim, A., Bacic, A., Hill, A.F., Stroud, D.A., Ryan, M.T., Agbinya, J.A., Mariadasson, J.M., Burgess, A.W. and Mathivanan, FunRich: a standalone tool for functional enrichment analysis. Proteomics.15, 2597-2601.
49) I. S. Vlachos, M. D. Paraskevopoulou, D. Karagkouni, G. Georgakilas, T. Vergoulis, I. Kanellos, I-L. Anastasopoulos, S. Maniou, K. Karathanou, D. Kalfakakou, A. Fevgas, T. Dalamagas and A. G. Hatzigeorgiou. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucl. Acids Res. (2014)