ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εντοπισμός RNA τροποποιήσεων σε δεδομένα μεταγραφώματος και εκτίμηση της επίδρασής τους στους στόχους των miRNA

Μάριος Σ. Μηλιώτης

Επιβλέπουσα: **Καθ. Άρτεμις Χατζηγεωργίου,** Καθηγήτρια Βιοπληροφορικής, Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας

ΑΘΗΝΑ

ΝΟΕΜΒΡΙΟΣ 2019

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS & TELECOMMUNICATIONS

POSTGRADUATE PROGRAM
"INFORMATION TECHNOLOGIES IN MEDICINE AND BIOLOGY"

MASTER'S THESIS

RNA editing identification in transcriptomics data and assessment of impact in miRNA targeting

Marios S. Miliotis

Supervisor: **Prof. Artemis Hatzigeorgiou,** Professor of Bioinformatics, Department of Electrical & Computer Engineering, Telecommunications and Networks, University of Thessaly

ATHENS

NOVEMBER 2019

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εντοπισμός RNA τροποποιήσεων σε δεδομένα μεταγραφώματος και εκτίμηση της επίδρασής τους στους στόχους των miRNA

Μάριος Σ. Μηλιώτης

Α.Μ.: ΠΙΒ0183

ΕΠΙΒΛΕΠΟΥΣΑ: **Καθ. Άρτεμις Χατζηγεωργίου,** Καθηγήτρια Βιοπληροφορικής, Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Καθ. Άρτεμις Χατζηγεωργίου,** Καθηγήτρια Βιοπληροφορικής, Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας
Δρ. Martin Reczko, Ερευνητής Καθηγητής, Ερευνητικό Κέντρο Βιοϊατρικών Επιστημών “Αλέξανδρος Φλέμινγκ”
Δρ. Αλέξανδρος Δημόπουλος, Μεταδιδακτορικός ερευνητής, Ερευνητικό Κέντρο Βιοϊατρικών Επιστημών “Αλέξανδρος Φλέμινγκ”

ΝΟΕΜΒΡΙΟΣ 2019

MASTER'S THESIS

RNA editing identification in transcriptomics data and assessment of impact in miRNA targeting

Marios S. Miliotis

SRN: ΠΙΒ0183

SUPERVISOR: **Prof. Artemis Hatzigeorgiou,** Professor of Bioinformatics, Department of Electrical & Computer Engineering, Telecommunications and Networks, University of Thessaly

EXAMINATION COMMITTEE:

Prof. Artemis Hatzigeorgiou, Professor of Bioinformatics, Department of Electrical & Computer Engineering, Telecommunications and Networks, University of Thessaly
Dr. Martin Reczko, Staff research scientist professor level at Biomedical Sciences Research Center "Alexander Fleming"
Dr. Alexandros Dimopoulos, Postdoctoral Researcher at Biomedical Sciences Research Center "Alexander Fleming"

NOVEMBER 2019

ΠΕΡΙΛΗΨΗ

Η μεταγραφική τροποποίηση είναι μια μεταγραφική/μετα-μεταγραφική διαδικασία, κατά την οποία ένα μόριο RNA υπόκειται στη μεταλλαγή της ακολουθίας του μέσω της εισαγωγής, της απαλοιφής ή της μεταβολής των βάσεων της. Στα μετάζωα, η πλειονότητα των μεταγραφικών τροποποιήσεων που συμβαίνουν αφορά τη μετατροπή του νουκλεοτιδίου αδενοσίνη (A) σε ινοσίνη (I), φαινόμενο που καταλύεται από τα μέλη της οικογένειας των γονιδίων των απαμινάσεων της αδενοσίνης (ADAR) που δρουν σε RNA με διπλή έλικα (dsRNA). Το φαινόμενο της μεταγραφικής τροποποίησης εμφανίζει σχετικά αυξημένη συχνότητα σε μόρια που φέρουν περιοχές ρετροτρανσποζονίων Alu στην ακολουθία τους.

Η τροποποίηση της κωδικής περιοχής των pre-mRNA μπορεί να οδηγήσει στην ενσωμάτωση διαφορετικού αμινοξέος κατά τη μετάφραση και να συμβάλει έτσι στην ποικιλότητα των πρωτεϊνικών προϊόντων και λειτουργιών. Ωστόσο, οι περισσότερες A-σε-I τροποποιήσεις απαντώνται σε μη κωδικές περιοχές των pre-mRNA και των mRNA, καθώς και σε μη κωδικά RNA. Οι μετατροπές στην UTR (μη μεταφραζόμενη περιοχή) ενός mRNA μπορούν να ρυθμίσουν τη μετάφραση, το μάτισμα και την αποικοδόμησή τους. Επίσης, τροποποιήσεις σε ακολουθίες microRNA (miRNA) και long non-coding RNA (lncRNA), καθώς και τροποποιήσεις στις θέσεις πρόσδεσής τους, μπορούν να επηρεάσουν τη βιογένεσή τους, την αναγνώριση των στόχων τους, τη δομή και τη σταθερότητά τους.

Στόχος αυτής της μελέτης είναι να γίνει σύγκριση μιας ομάδας εργαλείων για τον εντοπισμό RNA τροποποιήσεων, να διαχωριστούν τα πραγματικά συμβάντα μεταλλαγής στις 3'UTR των mRNA και να εκτιμηθεί η επίδρασή τους στην ειδικότητα και στην αποτελεσματικότητα της πρόσδεσής των miRNA.

Αρχικά, χρησιμοποιήθηκαν ζευγάρια από σύνολα δεδομένων αλληλούχισης του RNA και του DNA του ίδιου δείγματος ώστε να εντοπιστούν A-σε-I RNA τροποποιήσεις σε 3'UTR περιοχές. Η χρήση ζευγών έγινε με σκοπό να εξεταστούν συμβάντα σε επίπεδο δείγματος, αυξάνοντας έτσι την ειδικότητα. Το σύνολο των δεδομένων που χρησιμοποιήθηκε αποτελούνταν από 2 δείγματα για δοκιμή, 1 ADAR enzyme knockdown δείγμα για έλεγχο και τη RADAR, μία περιεκτική συλλογή A-σε-I δεδομένων στα μεταγραφώματα του ανθρώπου, του ποντικιού και της μύγας, με τους δύο προαναφερθέντες πόρους να αποτελούν τα δεδομένα αντικειμενικής αλήθειας για τη μελέτη. Στη συνέχεια, αναζητήθηκε ο καλύτερος αλγόριθμος στον εντοπισμό RNA τροποποιήσεων. Το ADAR knockdown δείγμα χρησιμοποιήθηκε ώστε να επισημανθούν τα υψηλά ψευδώς θετικά ποσοστά. Η σύγκριση περιλάμβανε το RES-Scanner, που χρησιμοποιεί τον Burrows-Wheeler aligner (BWA), το REDItools που τρέχει με τον aligner GSNAP και το RNAEditor, το οποίο εκτελέστηκε τόσο με τον BWA (προκαθορισμένη επιλογή) όσο και με τον GSNAP. Οι δύο πρώτοι αλγόριθμοι υποστηρίζουν εκ κατασκευής ζευγάρια RNA-DNA συνόλων δεδομένων, ενώ ο τρίτος τροποποιήθηκε ώστε να λαμβάνει υπόψιν και την DNA πληροφορία. Πιο σταθερή συμπεριφορά σε σχέση με την ειδικότητα και την ευαισθησία στα αποτελέσματα αναδείχθηκε να έχει το RNAEditor αξιοποιώντας τον aligner BWA.

Μετέπειτα, 3'UTR που βρέθηκαν να φέρουν τροποποίηση δόθηκαν ως είσοδος στους αλγορίθμους πρόβλεψης στόχων των miRNA ώστε να εκτιμηθούν στατιστικά διαφορές στις υπολογισμένες περιοχές πρόσδεσης που δημιουργήθηκαν εξαιτίας των φαινομένων τροποποίησης. Σε αυτό το στάδιο η σύγκριση επεκτάθηκε προσθέτοντας ένα ακόμα δείγμα με φυσιολογική (wild-type) έκφραση του ADAR από το ίδιο πείραμα με το ADAR

knockdown δείγμα. Συμβάντα τα οποία καταγράφηκαν σε 3'UTR χρησιμοποιήθηκαν ώστε να παραχθούν 2 ισάριθμα σύνολα από ακολουθίες, από τις οποίες 2062 ανήκαν σε περιοχές με υψηλό αριθμό διαδοχικών επαναλήψεων Alu και 144 σε non-Alu. Επίσης, χρησιμοποιήθηκαν τα πρώτα 50 σε έκφραση miRNA για κάθε δείγμα, ώστε να περιοριστεί το εύρος των περιοχών πρόβλεψης στόχων miRNA στην ανάλυση που έγινε με τους αλγορίθμους TargetScan και MIRZA-G. Και οι 2 εκτελέστηκαν χωρίς να λαμβάνονται υπόψιν εξελικτικά χαρακτηριστικά, καθότι αυτά δεν είναι δυνατόν να υπολογιστούν για τις περιοχές που υφίστανται τροποποίηση.

Τα αποτελέσματα υποδηλώνουν ότι οι τροποποιήσεις κυρίως μεταβάλλουν τα χαρακτηριστικά των υφιστάμενων περιοχών πρόσδεσης, ενώ σε πολύ μικρότερο βαθμό τις καθιστούν εντελώς μη λειτουργικές ή δημιουργούν νέες περιοχές. Επιπροσθέτως, παρατηρήθηκε ήπια μεταβολή της κατασταλτικής δράσης των miRNA που στοχεύουν τροποποιημένες UTR. Ξεχωριστή ανάλυση των UTR που εμφανίζουν υψηλό αριθμό τροποποιήσεων δεν υπέδειξε σημαντική συσχέτιση με το βαθμό αυξομείωσης της καταστολής. Το γεγονός ότι η κατασταλτική δράση των miRNA δε φάνηκε να επηρεάζεται καθολικά προς μία κατεύθυνση, υποδηλώνει πως ο ρυθμιστικός ρόλος των RNA τροποποιήσεων δεν ακολουθεί ένα γενικό κανόνα, αντιθέτως, δρα ως μηχανισμός βελτιστοποίησης, κατά περίπτωση ισχυροποιώντας ή αποδυναμώνοντας την πρόσδεση.

Σε αυτή τη μελέτη συγκρίναμε εργαλεία για τον εντοπισμό RNA τροποποιήσεων, καταλήξαμε με ένα σύνολο τροποποιημένων 3'UTR και πραγματοποιήσαμε ανάλυση για την πρόβλεψη στόχων των miRNA σε αυτές, η οποία υπέδειξε άλλοτε ενίσχυση και άλλοτε εξασθένηση της κατασταλτικής δράσης των miRNA στους στόχους τους, με ένταση ανεξάρτητη του αριθμού των συμβάντων τροποποίησης στην περιοχή πρόσδεσης. Περαιτέρω αναλύσεις με περισσότερα δείγματα και καταστάσεις θα φανούν χρήσιμες ώστε να επιβεβαιωθούν και να καταστούν στατιστικά σημαντικότερα τα ευρήματα της παρούσας εργασίας.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Βιοπληροφορική, Υπολογιστική Βιολογία

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: RNA τροποποιήσεις, αλληλούχιση επόμενης γενιάς, μεταγράφωμα, A-σε-I μετατροπή, πρόβλεψη στόχων miRNA

ABSTRACT

RNA editing is a co/post-transcriptional process, during which an RNA molecule is undergone an alteration of its sequence by insertion, deletion or modification. The majority of such changes in metazoans is comprised by adenosine (A) to inosine (I) nucleotide transitions, which are catalyzed by members of the adenosine deaminase gene family (ADAR) acting on double-stranded RNA (dsRNA). RNA editing is relatively widespread in Alu-containing mRNA molecules.

Editing of the coding sequence in pre-mRNAs can modify codons and lead to the incorporation of different amino acids during translation, contributing to protein function diversity. However, most A-to-I editing events occur in non-coding regions of pre-mRNAs and mRNAs, as well as in non-coding RNAs. Editing in the UTR (untranslated region) of mRNAs can regulate their translation, splicing and degradation. Also, events in microRNA (miRNA) and long non-coding RNA (lncRNA) sequences, as well as their binding sites, can affect their biogenesis, target recognition, structure and stability.

The goal of this study was to compare a set of RNA editing identification tools, distinguish true substitution events in 3'UTR of mRNAs and assess their impact on miRNA specificity and binding efficacy.

Initially, we used matching RNA and DNA sequencing data to identify A-to-I RNA editing events in 3'UTR regions. This was done to investigate event calls in individual level, increasing specificity. Our dataset consisted of 2 test samples, 1 ADAR enzyme knockdown control sample and RADAR, a comprehensive collection of A-to-I editing events in human, mouse and fly transcripts, with the last two resources being used as ground truth. Then we went on to find the best algorithm to identify events. The ADAR knockdown dataset was useful to pinpoint high false positive rates. The comparison included RES-Scanner employing the Burrows-Wheeler Aligner (BWA), REDItools running with GSNAP aligner and RNAEditor which was run with BWA (default option) and GSNAP. The first two approaches natively support paired/matched RNA-DNA datasets, while the latter was modified to include DNA information. The most robust behaviour in terms of sensitivity and specificity was observed from RNAEditor with BWA aligner.

Edited and non-edited 3'UTR were subsequently used as input in miRNA target prediction algorithms to statistically assess differences in the computed binding sites that arose due to the editing phenomena. The wildtype counterpart of the ADAR knockdown experiment was employed here to further enhance the comparison. Events annotated in 3'UTR were used to generate 2 equally numbered sets of sequences, 2062 of which belonged to highly repetitive Alu regions and 144 in non-Alu. The top 50 expressed miRNA in each sample were used to confine the target prediction analysis that was performed using TargetScan and MIRZA-G algorithms. Both of them were run without incorporating evolutionary features, which cannot be effectively measured in the case of the edited sequences.

The results show a strong preference towards modification of binding site feature distributions, rather than generating new or depleting existing sites. Moreover, we observed a mild alteration of the repressive action of miRNA targeting edited UTR. A separate analysis of highly edited UTR, i.e. UTR subjected to multiple editing events, did not indicate any correlation with the degree of the change. The lack of a global trend in the alteration of miRNA repressive activity implies RNA editing can serve distinct roles in miRNA efficacy, fine-tuning their targeting action on a case-by-case basis.

In this study, we did a benchmark of RNA editing identification tools, we came up with a set of edited UTRs and performed miRNA target prediction on them. This analysis indicated alteration of the targeting efficacy by miRNA, irrespective of the number of editing events in the region. Further analyses of more samples and conditions will be useful to validate and empower our findings.

SUBJECT AREA: Bioinformatics, Computational Biology

KEYWORDS: RNA editing, NGS, transcriptomics, A-to-I substitution, miRNA target prediction