NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS & TELECOMMUNICATIONS

POSTGRADUATE PROGRAM
"INFORMATION TECHNOLOGIES IN MEDICINE AND BIOLOGY"

MASTER THESIS

Automatic extraction of novel lncRNA-miRNA interactions by adopting a text mining approach.

Elissavet P. Zacharopoulou

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

ATHENS

NOVEMBER 2019

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

Αυτόματη εξαγωγή των αλληλεπιδράσεων lncRNA-miRNA υιοθετώντας μια προσέγγιση εξόρυξης δεδομένων κειμένου.

Ελισσάβετ Π. Ζαχαροπούλου

Επιβλέπουσα : **Άρτεμις Χατζηγεωργίου,** Καθηγήτρια Βιοπληροφορικής, Τμήμα Μηχανικών Η/Υ, Πανεπιστήμιο Θεσσαλίας

ΑΘΗΝΑ

ΝΟΕΜΒΡΙΟΣ 2019

MASTER THESIS

Automatic extraction of novel lncRNA-miRNA interactions by adopting a text mining approach.

Elissavet P. Zacharopoulou

SRN.: PIV0189

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

EXAMINATION COMMITTEE:

Prof. Artemis Hatzigeorgiou, Professor of Bioinformatics, Department of Electrical & Computer Engineering, University of Thessaly
Dr. Martin Reczko, Head of the bioinformatics group of the genomics facility, Alexander Fleming
Dr. Dimitra Karagkouni, Post-Doctoral Research Assistant, Hellenic Pasteur Institute, Univeristy of Thessaly

ATHENS
NOVEMBER 2019

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αυτόματη εξαγωγή των αλληλεπιδράσεων lncRNA-miRNA υιοθετώντας μια προσέγγιση εξόρυξης δεδομένων κειμένου.

Ελισσάβετ Π. Ζαχαροπούλου

Α.Μ.: ΠΙΒ0189

Επιβλέπουσα : **Άρτεμις Χατζηγεωργίου,** Καθηγήτρια Βιοπληροφορικής, Τμήμα Μηχανικών Η/Υ, Πανεπιστήμιο Θεσσαλίας

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:
Καθ. Άρτεμις Χατζηγεωργίου, Καθηγήτρια Βιοπληροφορικής, Τμήμα Μηχανικών Η/Υ, Πανεπιστήμιο Θεσσαλίας
Δρ. Martin Reczko, Επικεφαλής Τμήματος Βιοπληροφορικής Ερευνητικού Κέντρου Βιοϊατρικών Επιστημών, Αλέξανδρος Φλέμινγκ
Δρ. Δήμητρα Καραγκούνη, Μεταδιδακτορική Ερευνήτρια, Ελληνικό Ινστιτούτο Pasteur, Πανεπιστήμιο Θεσσαλίας

ΑΘΗΝΑ
ΝΟΕΜΒΡΙΟΣ 2019

ABSTRACT

Many studies have shown that microRNAs (miRNAs) and more recently long non-coding RNAs (lncRNAs) have multiple functions in a wide range of biological processes, such as proliferation, apoptosis, cell cycle arrest, cell migration, and invasion. miRNAs can induce mRNA degradation and/or translational suppression by binding to the 3' untranslated region (3’ UTR), the coding region (CDS) and in some cases, the 5' untranslated region (5' UTR) of the mRNA target. miRNAs also interact with non-coding RNAs, such as long-non coding transcripts. The role of lncRNA-miRNA interacting activity is still unexplored. LncRNAs may act as “sponges” for miRNAs, also known as “competing endogenous RNA” (ceRNA) and indirectly regulate the expression of target-mRNAs. Recently, more and more studies focus on the experimental validation of miRNA-lncRNA interactions and their endogenous activity. This valuable information is hidden in numerous publications. To this end, a text mining pipeline with full-text capacity has been deployed, dedicated to index miRNA coding and non-coding targets. The pipeline processes a prime total of publications from NCBI and stand outs papers and the respective representative sentences that denote miRNA-target interacting pairs.

The text-pipeline was trained on papers from DIANA-LncBase v2 and DIANA-TarBase v8, databases devoted to the cataloguing of miRNA targets. During the preprocessing, the term frequency-inverse document frequency (TFIDF) statistic tool was utilized to estimate a threshold to distinguish the optimal papers. Papers for evaluation are given by a query to the PMC database. After training sentence segmentation and part-of-speech tagging, a lexicon with keywords collects sentences from papers that contain the required features. MongoDB and Python were the main tools applied for the above pipeline. Ensembl v95 was utilized to construct the lexicon for genes and miRBase for miRNAs. This pipeline can provide the primary step before the manual configuration.

The text-mining approach was primary used to collect miRNA-lncRNA pairs. 155 papers of which 132 contained the desired information were exported. The collected sentences form the papers were further preprocessed and manually curated to gather additional information characterizing the interacting pairs, such as cell types and tissues. The final set comprises 570 entries, corresponding to 4 distinct experimental methodologies, 117 cell types and 40 tissues for human species. This compilation of interactions was incorporated into the 3rd version of DIANA-LncBase.

The text mining pipeline can be utilized for the constant update of databases with miRNA-targets such as LncBase and TarBase, providing a valuable asset for ncRNA research. Through changes and extensions, the ability to reduce the required manual curation is promising.

SUBJECT AREA: Text Mining

KEYWORDS: lncRNA, miRNA, natural language processing, feature extraction

ΠΕΡΙΛΗΨΗ

Πολλές μελέτες έχουν δείξει ότι τα microRNAs (miRNAs) και τα μακρά μη κωδικά RNAs (lncRNAs) έχουν πολλαπλές λειτουργίες σε ένα ευρύ φάσμα βιολογικών διεργασιών, όπως ο πολλαπλασιασμός, η απόπτωση, η διακοπή κυτταρικού κύκλου, η μετανάστευση κυττάρων και η εισβολή. Τα miRNAs μπορούν να προκαλέσουν αποικοδόμηση και/ή μεταφραστική καταστολή του mRNA με δέσμευση στην 3 'αμετάφραστη περιοχή (3' UTR), την κωδικοποιητική περιοχή (CDS) και σε μερικές περιπτώσεις την 5' αμετάφραστη περιοχή (5' UTR) του mRNA στόχου. Επίσης, τα miRNAs αλληλεπιδρούν με μη κωδικά RNAs, όπως τα μακρά μη κωδικά μετάγραφα. Ο ρόλος της αλληλεπίδρασης των lncRNA-miRNA είναι ακόμα ανεξερεύνητος. Τα LncRNAs μπορούν να λειτουργήσουν ως "σπόγγοι" για τα miRNAs, επίσης γνωστά ως "ανταγωνιστικά ενδογενή RNAs" (ceRNAs) και να ρυθμίζουν έμμεσα την έκφραση των mRNA στόχων. Πρόσφατα, όλο και περισσότερες μελέτες επικεντρώνονται στην πειραματική εξακρίβωση των αλληλεπιδράσεων μεταξύ lncRNA-miRNA και στην ενδογενή δραστηριότητά τους. Αυτή η πολύτιμη πληροφορία κρύβεται σε πολλές δημοσιεύσεις. Ως εκ τούτου, αναπτύσσεται ένας αλγόριθμος, που με τεχνικές text mining και με δεδομένα ολόκληρα τα κείμενα δημοσιεύσεων που εμπεριέχονται στην TarBase v8 και στην LncBase v2, επικεντρώνεται στους κωδικούς και μη κωδικούς στόχους των miRNA. Ο αλγόριθμος επεξεργάζεται ένα πρωταρχικό σύνολο δημοσιεύσεων από την NCBI και ξεχωρίζει αυτές που θεωρούνται σημαντικές και τις αντίστοιχες αντιπροσωπευτικές προτάσεις που υποδηλώνουν ζεύγη αλληλεπιδράσεων miRNA-στόχου.

Ο αλγόριθμος είναι εκπαιδευμένος σε δημοσιεύσεις από τις DIANA-LncBase v2 και DIANA-TarBase v8, βάσεις δεδομένων που αφορούν την καταλογογράφηση στόχων miRNA. Κατά την προ-επεξεργασία, χρησιμοποιήθηκε το στατιστικό εργαλείο συχνότητας-αντίστροφης συχνότητας εγγράφου (TFIDF) για την εκτίμηση ενός κατωφλίου για τη διάκριση των βέλτιστων δημοσιεύσεων. Οι δημοσιεύσεις για αξιολόγηση δίδονται από ένα ερώτημα στη βάση δεδομένων PMC. Μετά τον κατακερματισμό του κάθε κειμένου σε προτάσεις και τον ορισμό του σχήματος λόγου της κάθε λέξης στην πρόταση, με την χρήση λεξικού με λέξεις-κλειδιά γίνεται η συλλογή των προτάσεων που περιέχουν τα απαιτούμενα χαρακτηριστικά. Η MongoDB και η Python ήταν τα κύρια εργαλεία που εφαρμόστηκαν για τον αλγόριθμο. Η Ensembl v95 χρησιμοποιήθηκε για την κατασκευή του λεξικού για τα γονίδια και η miRBase για τα miRNAs. Αυτός ο αλγόριθμος μπορεί να παρέχει το κύριο βήμα πριν από τη εξαγωγή πληροφορίας από τις δημοσιεύσεις μη αυτοματοποιημένα.

Η μέθοδος εξόρυξης κειμένου χρησιμοποιήθηκε πρωταρχικά για τη συλλογή ζευγών miRNA-lncRNA. Εξήχθησαν 155 δημοσιεύσεις από τις οποίες οι 132 επεριέχουν τις επιθυμητές πληροφορίες. Οι συλλεγμένες προτάσεις από τα άρθρα προ-επεξεργάστηκαν περαιτέρω και επιθεωρήθηκαν για την συλλογή πρόσθετων πληροφοριών που χαρακτηρίζουν τα ζεύγη αλληλεπίδρασης, όπως τύποι κυττάρων και ιστοί. Η τελική βάση περιλαμβάνει 570 εγγραφές, που αντιστοιχούν σε 4 διαφορετικές πειραματικές μεθοδολογίες, 117 κυτταρικούς τύπους και 40 ιστούς για το ανθρώπινο είδος. Αυτή η συλλογή αλληλεπιδράσεων ενσωματώθηκε στην τρίτη έκδοση του DIANA-LncBase.

Ο αλγόριθμος εξόρυξης κειμένου μπορεί να χρησιμοποιηθεί για τη συνεχή ενημέρωση βάσεων δεδομένων με στόχους miRNA όπως η LncBase και η TarBase, παρέχοντας ένα πολύτιμο πλεονέκτημα για την έρευνα ncRNA. Μέσω αλλαγών και επεκτάσεων, η δυνατότητα μείωσης της απαιτούμενης μη αυτοματοποιημένης επεξεργασίας είναι υποσχόμενη.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Text Mining

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: lncRNA, miRNA, επεξεργασία φυσικής γλώσσας, εξαγωγή χαρακτηριστικών