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

Εκπαίδευση αλγορίθμου εξόρυξης κειμένου για αυτοματοποιημένη αναγνώριση δημοσιεύσεων που περιλαμβάνουν αλληλεπιδράσεις μεταξύ microRNAs και γονιδίων

Ιωάννης-Λαέρτης Ν. Αναστασόπουλος

***Επιβλέπουσα*: Άρτεμις Χατζηγεωργίου,** Καθηγήτρια

ΑΘΗΝΑ

ΜΑΡΤΙΟΣ 2016

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εκπαίδευση αλγορίθμου εξόρυξης κειμένου για αυτοματοποιημένη αναγνώριση επιστημονικών δημοσιεύσεων που περιλαμβάνουν πιθανές αλληλεπιδράσεις μεταξύ microRNAs και γονιδίων

Ιωάννης-Λαέρτης Ν. Αναστασόπουλος

Α.Μ.: ΠΙΒ116

ΕΠΙΒΛΕΠΟΥΣΑ: **Άρτεμις Χατζηγεωργίου,** Καθηγήτρια

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Άρτεμις Χατζηγεωργίου,** Καθηγήτρια
Γεώργιος Σπύρου, Ειδικός Λειτουργικός επιστήμονας
Γιάννης Βλάχος, Ερευνητής/Επιστημονικός Συνεργάτης

Μάρτιος 2016

ΠΕΡΙΛΗΨΗ

Τα microRNA είναι μία μεγάλη ομάδα μικρών μορίων RNA που λειτουργούν ως ρυθμιστικοί παράγοντες της έκφρασης των γονιδίων. Ο ρόλος τους είναι να ρυθμίζουν τη σταθερότητα των mRNA στόχων τους σε μετα-μεταγραφικό στάδιο, καθώς και την μετάφραση αυτών σε πρωτεΐνες.

Η ενημέρωση μίας βάσης δεδομένων με πειραματικά επιβεβαιωμένες αλληλεπιδράσεις μεταξύ microRNA και γονιδίων, όπως το TarBase, είναι μία χρονοβόρα διαδικασία. Ο στόχος μας είναι να εκπαιδεύσουμε έναν αλγόριθμο εξόρυξης κειμένου που ονομάζεται TarMiner, χρησιμοποιώντας διάφορα σετ αλληλεπιδράσεων που στηρίζονται σε μία ποικιλία μεθοδολογιών, έτσι ώστε να αυτοματοποιήσουμε τη διαδικασία αναγνώρισης πιθανών νέων αλληλεπιδράσεων. Αυτά τα σετ δεδομένων προέρχονται από το TarBase, τη μεγαλύτερη διαθέσιμη βάση δεδομένων για αλληλεπιδράσεις μεταξύ microRNA και γονιδίων.

Ωστόσο, πολλές από τις καταγεγραμμένες αλληλεπιδράσεις προέρχονται από την ανάλυση NGS (Next-Generation-Sequencing) δεδομένων, ή κρύβονται σε εικόντες, λεζάντες, πίνακες ή συμπληρωματικό υλικό επιστημονικών δημοσιεύσεων, οπότε δεν μπορούν να αναγνωριστούν από μία τέτοια εφαρμογή. Είναι λοιπόν υψίστης σημασίας να σχεδιαστούν σύνολα θετικών και αρνητικών παραδειγμάτων, έτσι ώστε να εκπαιδευτεί ο αλγόριθμος με αλληλεπιδράσεις που μπορούν να αναγνωρισθούν μέσω του κειμένου, και να επιτευχθεί το καλύτερο δυνατό αποτέλεσμα.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Βιοπληροφορική

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: microRNA, βάση δεδομένων, εξόρυξη κειμένου, εκπαίδευση αλγορίθμου, αυτοματοποίηση

ABSTRACT

MicroRNAs are a large class of short RNA molecules that act as regulators of gene expression. Their role is to post-transcriptionally modulate the stability of mRNA targets and their rate of translation into proteins.

Updating a database with microRNA:gene validated interactions from scientific publications, such as TarBase, can be a very time-consuming process. Our aim is to train a text mining algorithm, called TarMiner, by utilizing different sets of experimentally supported interactions from various methodologies, in order to automate the process of identifying putative interactions. These datasets are derived from TarBase, the largest available repository of microRNA:gene interactions.

However, many of the catalogued interactions are derived from the analysis of NGS (Next Generation Sequencing) datasets, or are hidden in figures, captions, tables, or supplementary material of scientific publications, and cannot be processed by such an implementation. Therefore, it is important to design sets of positive and negative instances, so as to train the algorithm with interactions that can be text-mined, in order to achieve the best results possible.

SUBJECT AREA: Bioinformatics

KEYWORDS: microRNA, database, text mining, algorithm training, automization