



**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**

**Υπολογιστικές μέθοδοι πρόβλεψης μικρών RNA σε νέα  
γονιδιώματα**

**Δημήτρης Γ Κωνσταντόπουλος**

**Επιβλέποντες:** Άρτεμις Χατζηγεωργίου, Καθηγήτρια

**ΑΘΗΝΑ**

**ΦΕΒΡΟΥΑΡΙΟΣ 2016**

## **ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

Υπολογιστικές μέθοδοι πρόβλεψης μικρών RNA σε νέα γονιδιώματα

**Δημήτρης Γ Κωνσταντόπουλος**

**A.M.: 112**

**ΕΠΙΒΛΕΠΟΝΤΕΣ:** Άρτεμις Χατζηγεωργίου, Καθηγήτρια

**ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:** Άρτεμις Χατζηγεωργίου, Καθηγήτρια  
Γιώργος Βερνίκος, Διδάκτωρ-Ερευνητής  
Γιάννης Βλάχος, Ερευνητής/Επιστημονικός Συνεργάτης

Φεβρουάριος 2016

## ΠΕΡΙΛΗΨΗ

Τα μικρά RNA, microRNA, είναι μικρά σε μέγεθος (21-22 νουκλεοτίδια) μονόκλωνα μη κωδικά μετάγραφα που εμφανίζονται σε πολυκύτταρους ευκαρυωτικούς οργανισμούς και ιούς. Τα microRNA αποτελούν ρυθμιστές της γονιδιακής έκφρασης, καθώς αλληλεπιδρούν με έναν μεγάλο αριθμό ώριμων κωδικών μεταγράφων mRNA, προκαλώντας την καταστολή της μετάφρασης τους σε πρωτεΐνες ή την πλήρη αποδόμησή τους. Κατά συνέπεια τα microRNA επηρεάζουν είτε άμεσα είτε έμμεσα έναν μεγάλο αριθμό βιολογικών λειτουργιών των κυττάρων και γενικότερα του οργανισμού.

Η ανάπτυξη τεχνολογιών γονοτύπησης DNA μεγάλης κλίμακας, έχει αυξήσει τον αριθμό των διαθέσιμων νουκλεοτιδικών ακολουθιών γονιδιωμάτων, για πολλούς οργανισμούς ενδιαφέροντος. Οι απαιτήσεις της έρευνας για τους οργανισμούς αυτούς στις επιστήμες της υγείας, έχει σαν συνέπεια την ανάγκη για χαρτογράφηση των γενομικών περιοχών ενδιαφέροντος, όπως τα γονίδια, οι υποκινητές, τα μεταθετά στοιχεία, οι επαναλαμβανόμενες περιοχές, κ.α. , καθώς επίσης και οι περιοχές που περιέχουν τα μικρά RNA.

Ο σκοπός της συγκεκριμένης μελέτης είναι να διερευνήσει τις διαθέσιμες στην βιβλιογραφία μεθόδους αναγνώρισης μικρών μορίων RNA, επικεντρώνοντας στις υπολογιστικές. Πραγματοποιείται περιγραφή των μεθόδων αυτών, σε επίπεδο ανάλυσης των αλγορίθμων που χρησιμοποιούν, καθώς και τελικά η αξιολόγησή τους όσον αφορά την δυνατότητα σωστών προβλέψεων στον οργανισμό μοντέλο *drosophila melanogaster*. Τελικά επιλέγεται η πιο αξιόπιστη μέθοδος, ώστε να πραγματοποιηθεί η αναγνώριση των πιο πιθανών pre-microRNA δομών στον πρόσφατα γονοτυπημένο οργανισμό *Ceratitis Capitata* (μεσογειακή φρουτόμυγα).

Αρχικά περιγράφονται συνοπτικά κάποια χαρακτηριστικά των οργανισμών *drosophila melanogaster* και *Ceratitis Capitata*, οι οποίοι ανήκουν στην τάξη Δίπτερα και στην οικογένεια της φρουτόμυγας. Στην συνέχεια περιγράφονται τα γενικά χαρακτηριστικά των microRNA γονιδίων, η διαδικασία βιογένεσης των microRNA και οι βιολογικές λειτουργίες τους, κυρίως όσον αφορά τα έντομα. Έπειτα περιγράφονται οι μέθοδοι αναγνώρισης microRNA γονιδίων, αναλύονται 15 διαθέσιμα στην βιβλιογραφία προγράμματα πρόβλεψης και αξιολογείται η επίδοσή τους σε επίπεδο υπολογιστικού χρόνου, καθώς και αξιοπιστίας προβλέψεων, χρησιμοποιώντας διάφορες μετρικές μεθόδους όπως η ευαισθησία, η ειδικότητα, η πιστότητα, ο γεωμετρικός μέσος κ.α., για προβλέψεις στον οργανισμό μοντέλο *drosophila melanogaster*. Τελικά επιλέγεται η μέθοδος πιο αξιόπιστη μέθοδος με βάση τις παραπάνω μετρικές, ώστε να πραγματοποιηθεί σάρωση του γονιδιώματος του οργανισμού *Ceratitis Capitata*, και χαρτογράφηση των πιο πιθανών ακολουθιών που αντιστοιχούν σε pre-microRNA δομές.

**ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ:** Βιοπληροφορική

**ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ:** microRNA, pre-miRNA, *drosophila melanogaster*, πρόβλεψη, αξιολόγηση

## ABSTRACT

MicroRNAs (miRNAs) are small single stranded RNAs found in eukaryotes and viruses. miRNAs are considered as key regulators of gene expression as they interact with mRNA molecules that results to translational repression or mRNA degradation. miRNAs affect a great variety of biological processes of the cell and the organism

As next generation sequencing technologies are on the edge, and the scientific community necessity for more sequenced genomes is rising, new organisms are available for the research fields of Biology. In order to address the research on these newly sequenced genomes, consistent annotations of genomic loci, such as genes, promoters, enhancer, repeated elements etc, as also and microRNA genes, are demanded. Cost-effective and fast predictions of functional genomic regions can be handled by computational methods, and verified in the future by experimental procedures.

The current study is willing to investigate the available microRNA prediction techniques, and emphasize to the implemented computational pipelines. After deep search on the field of de novo microRNA identification, a set of prediction programs is collected and tested on the model organism drosophila melanogaster nucleotide sequences. The most reliable algorithm in terms of several statistical metrics is then used on a genome wide scan of candidate pre-miRNA sequences of the newly sequenced genome of Ceratitis Capitata.

In the beginning of this study, the characteristics of the insects drosophila melanogaster and Ceratitis Capitata are briefly described. They both belong in the order Diptera and the family of fruit flies. Next, evolution characteristics, biogenesis and implication in biological processes of microRNAs are analyzed, focusing on insects. A cohort of 15 prediction programs is collected, set up on a high performance computer cluster, used to identify pre-microRNA on drosophila melanogaster nucleotide sequences, and evaluated in the terms of sensitivity, specificity, accuracy, precision and other statistical measures. Finally, the most reliable program is used for a genome wide scan on the species Ceratitis Capitata and a set of pre-miRNA candidates is identified and collected.

**SUBJECT AREA:** Bioinformatics

**KEYWORDS:** microRNA, pre-miRNA, drosophila melanogaster, prediction, validation