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

"ΠΡΟΒΛΕΨΗ ΜΕΤΑΓΡΑΦΩΝ microRNA ΑΠΟ ΓΕΝΩΜΙΚΑ ΔΕΔΟΜΕΝΑ"

ΓΕΩΡΓΙΟΣ Κ. ΓΕΩΡΓΑΚΙΛΑΣ

Επιβλέπων: Άρτεμις Χατζηγεωργίου, Ερευνήτρια Β' Φλεμινγκ ΕΚΕΒΕ

ΑΘΗΝΑ

ΙΟΥΝΙΟΣ 2011

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΠΡΟΒΛΕΨΗ ΜΕΤΑΓΡΑΦΩΝ microRNA ΑΠΟ ΓΕΝΩΜΙΚΑ ΔΕΔΟΜΕΝΑ

ΓΕΩΡΓΙΟΣ Κ. ΓΕΩΡΓΑΚΙΛΑΣ

Α.Μ.: 046

ΕΠΙΒΛΕΠΩΝ: **Αρτεμις Χατζηγεωργίου,** Ερευνήτρια Β' Φλεμινγκ ΕΚΕΒΕ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Ιωάννης Εμίρης,** Καθηγητής ΕΚΠΑ
Σπυρίδων Γαρμπής, Ερευνητής Δ' ΙΙΒΕΑΑ

Ιούνιος 2011

Abstract

microRNAs are sort length (~22 nucleotides) endogenously produced RNA molecules which regulate gene transcription by binding, in a sequence related way, on 3' UnTranslated Region of messenger RNA. There are lots of microRNAs in eukaryotic cells, regulating a wide variety of gene-targets. During the past few years, microRNAs have been related with the regulation of many biological processes.
This thesis is part of a larger project, whose goal is to identify mechanisms that regulate microRNA transcription and therefore place them in a wider biological pathway of gene transcription and regulation. The transcription start sites of these microRNA genes are largely unknown. In order to explore this currently unknown scientific area, certain models need to be developed, whose training is based on already largely explored scientific areas with common features as microRNA transcription, test these models and finally evaluate them on novel experimental data.
The goal of the thesis is to develop a model, on data based on known protein coding gene transcription start sites. For this, we use information related to the methylated version of a histone (protein), called H3K4me3, which has the ability to bind on DNA regions that are transcription start sites [1]. Out of these data, the appropriate features are extracted, and the model is trained and tested.

The results are very good, since even the simple additive algorithm accomplishes a high prediction performance level (precision up to 70%) while on the same time by using a more advanced machine learning algorithm this performance level is enhanced even more.

SUBJECT AREA: Bioinformatics

KEYWORDS: microRNA, ChIP-seq, transcription, prediction algorithm, machine learning, computational biology

Περίληψη

Τα microRNAs είναι μικρού μήκους (~22 νουκλεοτίδια) ενδογενώς παραγόμενα μόρια RNA τα οποία ρυθμίζουν την γονιδιακή έκφραση δένοντας, με τρόπο σχετικό με την ακολουθία, στην 3' μη μεταγραφόμενη περιοχή του μεταφορικού RNA. Υπάρχει πληθώρα από microRNAs στα ευκαρυωτικά κύταρρα, τα οποία ελέγχουν τη μετάφραση μιας μεγάλης γκάμας γονιδίων στόχων. Τα τελευταία χρόνια, τα microRNAs έχουν συσχετιστεί με τη ρύθμιση πολλών βιολογικών διεργασιών.
Η παρούσα διπλωματική είναι τμήμα ενός ευρύτερου πρότζεκτ, του οποίου ο στόχος είναι να αναγνωριστούν οι μηχανισμοί που ελέγχουν την μεταγραφή των microRNA ώστε να τους δωθεί θέση σε ένα ευρύτερο βιολογικό μονοπάτι γονιδιακής έκφρασης και ελέγχου. Για την εξερεύνηση αυτής της προς το παρόν άγνωστης περιοχής, πρέπει να αναπτυχθούν συγκεκριμένα μοντέλα η εκπαίδευση των οποίων βασίζεται σε ευρέως μελετημένες επιστημονικές περιοχές που μοιράζονται κοινά χαρακτηριστικά με την άγνωστη περιοχή την οποία εξερευνούμε, να ελεγχθούν αυτά τα μοντέλα και τελικώς να αξιολογηθούν μέσω της εφαρμογής τους σε καινούρια πειραματικά δεδομένα.
Ο στόχος της διπλωματικής είναι η ανάπτυξη ενός μοντέλου το οποίο εκπαιδεύεται και ελέγχεται σε δεδομένα από γνωστές θέσεις έναρξης μεταγραφής γονιδίων που παράγουν πρωτείνες. Για να συμβεί αυτό, δοκιμάζεται η ιδιότητα μιας μεθυλιωμένης ιστόνης (πρωτείνη), γνωστή ως H3K4me3, να δένει πάνω στο DNA σε θέσεις όπου ξεκινά η μεταγραφή γονιδίων [1], εξάγονται τα κατάλληλα χαρακτηριστικά, εκπαιδεύεται και ελέγχεται το μοντέλο.

Τα αποτελέσματα είναι εξαιρετικά, μιας και ο σχετικά απλός προσθετικός αλγόριθμος κατάφερε υψηλό ποσοστό ακρίβειας (έως και 70%) στην πρόβλεψη ενώ χρησιμοποιώντας ένα πιο εξελιγμένο αλγόριθμο μηχανικής μάθησης σημειώθηκε σημαντική βελτίωση αυτής της απόδοσης.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Βιοπληροφορική

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: μικράRNA, ανοσοκατακρύμνηση παραγόντων, μεταγραφή, αλγόριθμος πρόβλεψης, μηχανική μάθηση, υπολογιστική βιολογία