ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάπτυξη υπολογιστικών μεθόδων για τον χαρακτηρισμό κωδικών περιοχών και UTR’s

ΑΙΚΑΤΕΡΙΝΗ Γ. ΣΤΥΛΛΟΥ

Επιβλέπουσα: Άρτεμις Χατζηγεωργίου, Ερευνήτρια Β’, ΕΚΕΒΕ - Φλέμινγκ

ΑΘΗΝΑ

ΦΕΒΡΟΥΑΡΙΟΣ 2011

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάπτυξη υπολογιστικών μεθόδων για τον χαρακτηρισμό κωδικών περιοχών και UTR's

Αικατερίνη Γ. Στύλλου

Α.Μ.: ΠΙΒ14

ΕΠΙΒΛΕΠΟΥΣΑ:

Άρτεμις Χατζηγεωργίου, Ερευνήτρια Β', ΕΚΕΒΕ - Φλέμινγκ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Άρτεμις Χατζηγεωργίου, Ερευνήτρια Β', ΕΚΕΒΕ - Φλέμινγκ

Ηλίας Μανωλάκος, Αναπληρωτής Καθηγητής ΕΚΠΑ

Κάτια Καραλή, Ερευνήτρια Β', Ίδρυμα Ιατροβιολογικών Ερευνών Ακαδημίας Αθηνών - ΙΙΒΕΑΑ

ΦΕΒΡΟΥΑΡΙΟΣ 2011

ΠΕΡΙΛΗΨΗ

Πρόσφατες μελέτες έδειξαν ότι το μεγαλύτερο μέρος του γονιδιώματος που εκφράζεται δεν ανήκει στο 2% των περιοχών που κωδικοποιούν πρωτεΐνες (coding RNA's) αλλά κυρίως σε περιοχές που δεν έχουν αυτή την πληροφορία (noncoding RNA's). Τα non-coding RNAs παρουσιάζουν μεγάλο ενδιαφέρον λόγο της λειτουργίας τους σαν ρυθμιστές σημαντικών βιολογικών λειτουργιών όπως η ρύθμιση της μεταγραφής και της μετάφρασης. Το γεγονός αυτό, σε συνδυασμό με τον τεράστιο αριθμό RNAs που προκύπτει με τις νέες τεχνολογίες ανάγνωσης του γονιδιώματος καθιστά απαραίτητη την ανάπτυξη εφαρμογών που αυτοματοποιούν τον διαχωρισμό των RNAs που κωδικοποιούν από αυτά που δεν κωδικοποιούν πρωτεΐνες.

Στην παρούσα διπλωματική εργασία, εξετάζουμε την απόδοση τεσσάρων τέτοιων εφαρμογών, των Coding Potential Calculator (CPC), ESTExplorer, OrfPredictor και OrfFinder. Σκοπός μας είναι να βρούμε την εφαρμογή ή το συνδυασμό εφαρμογών που επιτυγχάνει την καλύτερη πρόβλεψη coding και παράλληλα non-coding RNAs του ανθρώπινου γονιδιώματος. Η πρόβλεψη των coding RNAs περιλαμβάνει και τον εντοπισμό των σωστών ορίων των αντίστοιχων κωδικών περιοχών και κατ' επέκταση των σωστών ορίων των 3' και 5' αμετάφραστων περιοχών.

Στα πλαίσια της προσπάθειας αυτής, αναπτύχθηκε λογισμικό, σε Java, το οποίο υλοποιεί το μεγαλύτερο μέρος της διαδικασίας αξιολόγησης της επίδοσης των εφαρμογών. Τα δεδομένα εισόδου του προγράμματος είναι οι ακολουθίες συμπληρωματικού DNA (RNA / cDNA) του ανθρώπινου γονιδιώματος, οι οποίες ανακτήθηκαν από την αντίστοιχη βάση της ENSEMBL, μέσω του εργαλείου διαχείρισης βιολογικών δεδομένων BioMart. Το πρόγραμμα επεξεργάζεται τα cDNAs και τα προετοιμάζει για την εισαγωγή τους στις τέσσερις εφαρμογές πρόβλεψης. Επιπλέον, συλλέγει τα αποτελέσματα/προβλέψεις των εφαρμογών και τα αναλύει. Τέλος, παρέχει στατιστικά αποτελέσματα ενδεικτικά των ικανοτήτων πρόβλεψης της κάθε εφαρμογής ξεχωριστά καθώς και όλων των πιθανών συνδυασμών εφαρμογών.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Βιοπληροφορική

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Βιοπληροφορική, συμπληρωματικό DNA, ανθρώπινο γονιδίωμα. πρόβλεψη κωδικής περιοχής, πρόβλεψη 3' και 5' αμετάφραστων περιοχών

ABSTRACT

Recent studies have shown that the majority of transcripts do not belong in the 2% of protein encoding transcripts (coding RNAs) but function as noncoding RNAs instead. In vivo experiments have demonstrated important biological roles of noncoding RNAs, such as regulation of transcription and translation. This, coupled with the huge numbers of transcripts generated by cDNA and EST sequencing projects, necessitates the development of methods that automate the distinguishing of protein-coding RNAs from noncoding RNAs.

The main purpose of this thesis is to examine the performance of four such methods. The four methods are: Coding Potential Calculator (CPC), ESTExplorer, OrfPredictor and OrfFinder. Our main goal is to find the method or combination of the four methods that achieves the best prediction of coding and noncoding RNAs of the human genome. The prediction of the coding RNAs involves identifying the correct boundaries of the respective coding regions and thereby the right limits of the 3' and 5' un-translated regions.

As part of this effort, we developed software in Java, which implements the main part of the process of assessing the performance of the four methods. The input of the program is the complementary DNA sequences of the human genome which have been recovered by the respective database of ENSEMBL through BioMart, a management system of biological data. The software processes the cDNAs and prepares them for submission in the four predictor. Furthermore, it collects the outputs/predictions of the methods and analyses them. Finally, it provides statistical results indicative of the prediction abilities of each method individually and of all the possible combinations of the methods.

SUBJECT AREA: Bioinformatics

KEYWORDS: Bioinformatics, complementary DNA, human genome, coding region prediction, 3' and 5'-UTR prediction