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

Υπολογιστική μέθοδος εύρεση θέσεων έναρξης της μεταγραφής των γονιδίων με χρήση δεδομένων CAGE.

Νικόλαος Παναγιώτου Περδικοπάνης

Επιβλέπουσα **Άρτεμις Χατζηγεωργίου,** Καθηγήτρια Βιοπληροφορικής, Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας.

ΑΘΗΝΑ

ΦΕΒΡΟΥΑΡΙΟΣ 2017

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εύρεση Θέσεων Έναρξης της Μεταγραφής

Νικόλαος Π. Περδικοπάνης
Α. Μ. : ΠΙΒ 0141

ΕΠΙΒΛΕΠΩΝ **Άρτεμις Χατζηγεωργίου ,**Καθηγήτρια

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ **Άρτεμις Χατζηγεωργίου,** Καθηγήτρια Βιοπληροφορικής, Τμήμα Μηχανικών Η/Υ, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας.
Martin Reczko, Επικεφαλής Τμήματος Βιοπληροφορικής Ερευνητικού Κέντρου Αλέξανδρος Φλεμινγκ
Μαρία Παρασκευοπούλου, Ερευνήτρια/Επιστημονικός Συνεργάτης

Φεβρουάριος 2017

ΠΕΡΙΛΗΨΗ

Στόχος της παρούσας εργασίας είναι ο σχεδιασμός και η υλοποίηση ενός αλγόριθμου ικανού να εντοπίσει με ακρίβεια, τις θέσεις έναρξης μεταγραφής των γονιδίων. Ο αλγόριθμος που σχεδιάστηκε χρησιμοποιεί δεδομένα τα οποία έχουν εξαχθεί με την μέθοδο CAGE (Cap Analysis of Gene Expression) καθώς και χαρακτηριστικά της πρωτοταγούς και δευτεροταγούς δομής του DNA των υπό διερεύνηση περιοχών, σε συνδυασμό με μηχανική Μάθηση. Ο αλγόριθμος δοκιμάστηκε σε Human ES cells (H9 line) και η απόδοση του αξιολογήθηκε χρησιμοποιώντας δεδομένα καλά σχολιασμένων περιοχών έναρξης Γονιδίων από τη βιβλιογραφία.

Τα αποτελέσματα του αλγόριθμου συγκρίθηκαν ως προς τα ποιοτικά και ποσοτικά χαρακτηριστικά τους με αντίστοιχους αλγορίθμους εντοπισμού περιοχών έναρξης μεταγραφής (CAGEr Reclu Paraclu) που συναντώνται στην βιβλιογραφία.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Βιοπληροφορική, Μηχανική Μάθηση

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Θέσεις Έναρξη Μεταγραφής Γονιδίων, Μηχανική Μάθηση Μεταγραφή, Υποκινητής

ABSTRACT

Aim of this thesis Is the design and implementation of an algorithm efficient to identify Genes Transcription start Sites (TSS).

The algorithm use data extracted with CAGE method as well as features of the primary and secondary structure of DNA of the screening regions. Algorithm use Machine learning technics (SVM, Regression Models) for the proper TSS identification.

For the training and testing of the algorithm we use Human ES cells (H9 and H1 cell lines) and its efficiency evaluated using well annotated genes TSS, collected from the bibliography.

The algorithm compared with related algorithm (CAGEr Reclu Paraclu) for a set quantitative and qualitative features.