NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

INTERDEPARTMENTAL POSTGRADUATE PROGRAM
"INFORMATION TECHNOLOGIES IN MEDICINE AND BIOLOGY"

MASTER THESIS

Splice site recognition between different organisms

Despoina I. Kalfakakou

Supervisors: **Stavros Perantonis,** Research Director, NCSR Demokritos
George Paliouras, Research Director, NCSR Demokritos
Anastasia Krithara, Post-Doctoral Researcher, NCSR Demokritos

ATHENS

OCTOBER 2015

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΜΕΑΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
"ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αναγνώριση θέσεων ματίσματος μεταξύ διαφορετικών οργανισμών

Δέσποινα Η. Καλφακάκου

Επιβλέποντες: **Σταύρος Περαντώνης,** Διευθυντής Ερευνών, ΕΚΕΦΕ Δημόκριτος
Γεώργιος Παλιούρας, Διευθυντής Ερευνών, ΕΚΕΦΕ Δημόκριτος
Αναστασία Κριθαρά, Μεταδιδακτορική Ερευνήτρια, ΕΚΕΦΕ Δημόκριτος

ΑΘΗΝΑ

ΟΚΤΩΒΡΙΟΣ 2015

MASTER THESIS

Splice site recognition between different organisms

Despoina I. Kalfakakou
R.N.: PIV0119

SUPERVISORS: **Stavros Perantonis,** Research Director, NCSR Demokritos
George Paliouras, Research Director, NCSR Demokritos
Anastasia Krithara, Post-Doctoral Researcher, NCSR Demokritos

EXAMINATION COMITTEE:
Stavros Perantonis, Research Director, NCSR Demokritos
George Paliouras, Research Director, NCSR Demokritos
Anastasia Krithara, Post-Doctoral Researcher, NCSR Demokritos

October 2015

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αναγνώριση θέσεων ματίσματος μεταξύ διαφορετικών οργανισμών

Δέσποινα Η. Καλφακάκου
Α.Μ.: ΠΙΒ0119

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Σταύρος Περαντώνης,** Διευθυντής Ερευνών, ΕΚΕΦΕ Δημόκριτος
Γεώργιος Παλιούρας, Διευθυντής Ερευνών, ΕΚΕΦΕ Δημόκριτος
Αναστασία Κριθαρά, Μεταδιδακτορική Ερευνήτρια, ΕΚΕΦΕ Δημόκριτος

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Σταύρος Περαντώνης,** Διευθυντής Ερευνών, ΕΚΕΦΕ Δημόκριτος
Γεώργιος Παλιούρας, Διευθυντής Ερευνών, ΕΚΕΦΕ Δημόκριτος
Αναστασία Κριθαρά, Μεταδιδακτορική Ερευνήτρια, ΕΚΕΦΕ Δημόκριτος

Οκτώβριος 2015

ABSTRACT

The protein synthesis for most eukaryotic genes consists of different steps including transcription, post-transcriptional modification and translation. In the post-transcriptional modification step, the pre-mRNA is transformed into mRNA. One necessary step in the process of obtaining mature mRNA is called splicing. The latter involves the removal or "splicing out" of certain sequences referred to as intervening sequences, or introns. The final mRNA consists of the remaining sequences, called exons, which are connected to one another through the splicing process. Splice sites mark the boundaries (donor and acceptor) between exons and introns. Splice site recognition refers to the task of identifying these boundaries.

In this thesis, an investigation of transfer learning approaches for splice site recognition will take place. We apply transfer learning in order to face the problem of poorly annotated genomes and the lack of labeled data for predicting splice sites. The study involves the sequence analysis of the true splice sites of five different organisms, in order to extract the most representative features, and the development of two transfer learning classification models.

SUBJECT AREAS: Computational Biology, Genomics, Machine Learning

KEYWORDS: splicing, splice sites, transfer learning, DNA sequence analysis, DNA sequence representation

ΠΕΡΙΛΗΨΗ

Η πρωτεϊνική σύνθεση στα περισσότερα γονίδια των ευκαρυωτικών οργανισμών αποτελείται από διάφορα βήματα, συμπεριλαμβανομένης της μεταγραφής, της μετα-μεταγραφικής τροποποίησης και της μετάφρασης. Κατά τη διάρκεια του βήματος της μετα-μεταγραφικής τροποποίησης συγκεκριμένα, το pre-mRNA μετατρέπεται σε mRNA. Ένα απαραίτητο βήμα στη διαδικασία της απόκτησης ώριμου mRNA είναι η συναρμογή ή μάτισμα. Το μάτισμα περιλαμβάνει την αφαίρεση συγκεκριμένων υποακολουθιών οι οποίες αναφέρονται ως παρεμβαλλόμενες αλληλουχίες (εσώνια). Το τελικό ώριμο mRNA αποτελείται από τις εναπομείνασες υποακολουθίες (εξώνια), οι οποίες συνδέονται μεταξύ τους κατά τη διάρκεια του ματίσματος. Οι θέσεις ματίσματος αποτελούν τα σύνορα μεταξύ των εσωνίων και των εξωνίων. Η αναγνώριση των θέσεων ματίσματος αφορά στην ταυτοποίηση αυτών των συνόρων.

Στην παρούσα διπλωματική εργασία, λαμβάνει χώρα μια εξερεύνηση των μεθόδων μεταφοράς μάθησης για το πρόβλημα της αναγνώρισης των θέσεων ματίσματος. Οι μέθοδοι μεταφοράς μάθησης εφαρμόζονται ώστε να αντιμετωπιστεί το πρόβλημα των ανεπαρκώς αποκωδικοποιημένων γονιδιωμάτων και της έλλειψης επισημασμένων δεδομένων για την πρόβλεψη θέσεων ματίσματος. Ολόκληρη η μελέτη περιλαμβάνει την ανάλυση των ακολουθιών πραγματικών θέσεων ματίσματος από πέντε διαφορετικούς οργανισμούς, ώστε να εξαχθούν τα πιο αντιπροσωπευτικά χαρακτηριστικά, και την ανάπτυξη δύο μοντέλων ταξινόμησης μεταφοράς μάθησης.

ΘΕΜΑΤΙΚΕΣ ΠΕΡΙΟΧΕΣ: Υπολογιστική Βιολογία, Γενομική, Μηχανική Μάθηση

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: συναρμογή, μάτισμα, θέσεις ματίσματος, μεταφορά μάθησης, ανάλυση ακολουθιών DNA, αναπαράσταση ακολουθιών DNA