NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

POSTGRADUATE PROGRAM
"INFORMATION TECHNOLOGIES IN MEDICINE AND BIOLOGY"

MASTER THESIS

Splice Site Prediction Using Transfer Learning

Simos I. Kazantzidis

Supervisors: **Stavros Perantonis,** Research Director, NCSR Demokritos
Elias Manolakos, Assoc. Professor, Department of Informatics and Telecommunications, University of Athens
Anastasia Krithara, Associate researcher, NCSR Demokritos

ATHENS

MARCH 2016

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

Πρόβλεψη Θέσεων Ματίσματος με τη χρήση Μεταφοράς Μάθησης

Σίμος Η. Καζαντζίδης

Επιβλέποντες: **Σταύρος Περαντώνης,** Διευθυντής Ερευνών, ΕΚΕΦΕ Δημόκριτος
Ηλίας Μανωλάκος, Αν. Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών, ΕΚΠΑ
Αναστασία Κριθαρά, Συνεργαζόμενη ερευνήτρια, ΕΚΕΦΕ Δημόκριτος

ΑΘΗΝΑ

ΜΑΡΤΙΟΣ 2016

MASTER THESIS

Splice Site Prediction Using Transfer Learning

Simos I. Kazantzidis
R.N.: PIV0127

SUPERVISORS: **Stavros Perantonis,** Research Director, NCSR Demokritos
Elias Manolakos, Assoc. Professor, Department of Informatics and Telecommunications, University of Athens
Anastasia Krithara, Associate researcher, NCSR Demokritos

EXAMINATION COMMITTEE: **Stavros Perantonis,** Research Director, NCSR Demokritos
Elias Manolakos, Assoc. Professor, Department of Informatics and Telecommunications, University of Athens
Anastasia Krithara, Associate researcher, NCSR Demokritos

March 2016

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Πρόβλεψη Θέσεων Ματίσματος με τη χρήση Μεταφοράς Μάθησης

Σίμος Η. Καζαντζίδης
Α.Μ.: ΠΙΒ0127

ΕΠΙΒΛΕΠΩΝΤΕΣ: **Σταύρος Περαντώνης,** Διευθυντής Ερευνών, ΕΚΕΦΕ Δημόκριτος
Ηλίας Μανωλάκος, Αν. Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών, ΕΚΠΑ
Αναστασία Κριθαρά, Συνεργαζόμενη ερευνήτρια, ΕΚΕΦΕ Δημόκριτος

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Σταύρος Περαντώνης,** Διευθυντής Ερευνών, ΕΚΕΦΕ Δημόκριτος
Ηλίας Μανωλάκος, Αν. Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών, ΕΚΠΑ
Αναστασία Κριθαρά, Συνεργαζόμενη ερευνήτρια, ΕΚΕΦΕ Δημόκριτος

Μάρτιος 2016

ABSTRACT

One of the open problems in the field of bioinformatics, is the automatic gene prediction (nucleotide sequence that encodes proteins). More specifically, researchers are trying to predict those positions that correspond to the beginning and the end of genes within a genome. These positions are known as splice sites. Several machine learning techniques have been used for the specific problem. Nevertheless, the acquisition of annotated data, necessary to implement supervised learning techniques, is a significant challenge, as the cost is very large. One of the approaches for addressing this problem is the transferring of knowledge (transfer learning approach). The aim of this work is the study of the representation of genes in order to take into account the sequence of nucleotides within a genome and the role of this representation in transfer learning methods.

SUBJECT AREA: Splice Site Prediction, Computational Biology

KEYWORDS: transfer learning, splice site, machine learning, n-gram graphs

ΠΕΡΙΛΗΨΗ

Ένα από τα ανοιχτά προβλήματα της βιοπληροφορικής, είναι η αυτόματη πρόβλεψη γονιδίων (αλληλουχία νουκλεοτιδίων που κωδικοποιεί πρωτεΐνες). Πιο συγκεκριμένα, οι ερευνητές προσπαθούν να προβλέψουν τις θέσεις που αντιστοιχούν στην αρχή και το τέλος των γονιδίων σε ένα γονιδίωμα. Οι θέσεις αυτές είναι γνωστές ως σήματα ματίσματος (splice sites). Διάφορες τεχνικές της μηχανικής μάθησης έχουν χρησιμοποιηθεί για το συγκεκριμένο πρόβλημα. Παρόλα αυτά, η απόκτηση των επισημειωμένων δεδομένων που είναι αναγκαία για να εφαρμοστούν οι τεχνικές επιβλεπόμενης μάθησης, αποτελεί μια σημαντική πρόκληση, καθώς το κόστος είναι πολύ μεγάλο. Μία από τις προσεγγίσεις για την αντιμετώπιση αυτού του προβλήματος είναι η μεταφορά μάθησης (transfer learning). Στόχος της παρούσας εργασίας είναι η μελέτη της αναπαράστασης των γονιδίων, ώστε να λαμβάνεται υπόψιν η αλληλουχία των νουκλεοτιδίων σε ένα γονιδίωμα, και ο ρόλος της αναπαράστασης αυτής σε μεθόδους μεταφοράς μάθησης.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Πρόβλεψη Θέσεων Ματίσματος, Υπολογιστική Βιολογία

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: μεταφορά μάθησης, θέσεις ματίσματος, μηχανική μάθηση, γράφοι ν-γραμμάτων