NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS & TELECOMMUNICATIONS

POSTGRADUATE PROGRAM
"INFORMATION TECHNOLOGIES IN MEDICINE AND BIOLOGY"

MASTER THESIS

A method for identifying TSS from CAGE data using a Genomic Signal Processing approach

Dimitris N. Grigoriadis

Supervisor: **Prof. Artemis Hatzigeorgiou,** Professor of Bioinformatics, Department of Electrical & Computer Engineering, Telecommunications and Networks, University of Thessaly

ATHENS

FEBRUARY 2019

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

Μέθοδος αναγνώρισης θέσεων έναρξης της μεταγραφής των γονιδίων από CAGE δεδομένα χρησιμοποιώντας τεχνικές επεξεργασίας σήματος

Δημήτρης Ν. Γρηγοριάδης

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

ΑΘΗΝΑ

ΦΕΒΡΟΥΑΡΙΟΣ 2019

MASTER THESIS

A method for identifying TSS from CAGE data using a Genomic Signal Processing approach

Dimitris N. Grigoriadis

Student Registration Number: ΠΙΒ0170

SUPERVISOR: **Prof. Artemis Hatzigeorgiou,** Professor of Bioinformatics, Department of Electrical & Computer Engineering, Telecommunications and Networks, University of Thessaly

EXAMINING COMMITTEE: **Prof. Artemis Hatzigeorgiou,** Professor of Bioinformatics, Department of Electrical & Computer Engineering, Telecommunications and Networks, University of Thessaly
Martin Reczko, Head of the bioinformatics group of the genomics facility, Alexander Fleming
Theodore Dalamagas, Researcher A' and Director at "Athena" Research Center, Information Management Systems Institute

February 2019

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μέθοδος αναγνώρισης θέσεων έναρξης της μεταγραφής των γονιδίων από CAGE δεδομένα χρησιμοποιώντας τεχνικές επεξεργασίας σήματος

Δημήτρης Ν. Γρηγοριάδης

ΑΜ: ΠΙΒ0170

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

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Καθ. Άρτεμις Χατζηγεωργίου,** Καθηγήτρια Βιοπληροφορικής,
Τμήμα Μηχανικών Υ/Η, Τηλεπικοινωνιών και Δικτύων του Πανεπιστημίου Θεσσαλίας
Martin Reczko, Επικεφαλής Τμήματος Βιοπληροφορικής Ερευνητικού Κέντρου Βιοϊατρικών Επιστημών, Αλέξανδρος Φλέμινγκ
Θοδωρής Δαλαμάγκας, Ερευνητής Α' και Αναπληρωτής Διευθυντής του Ινστιτούτου Πληροφοριακών Συστημάτων του Ερευνητικού Κέντρου «ΑΘΗΝΑ»

Φεβρουάριος 2019

ABSTRACT

Genomic signal processing (GSP) can solve various biological problems in low computational cost. There are many mathematical algorithms mostly used for gene identification and comparison between sequences, making use of several DNA representations that did not evolve due to lack of efficiency.

The knowledge of the exact position of the transcription start sites (TSS), which is the location where transcription starts at the 5'-end of a gene sequence in an RNA molecule, is critical for the identification of the regulatory regions that flank it. Many approaches have been mentioned in the literature about locating the TSS positions.

This study presents a novel method for identifying transcription start sites (TSS) from CAGE (Cap Analysis of Gene Expression) data and applying features and techniques borrowed form GSP that also aim to identify TSSs. A fairly new representation method for nucleotides has been introduces able to extract and represent the information in a time series signal vector.

Signals were transferred from time to frequency domain, which allows for filtering artifacts in an efficient robust way, and vice versa. Several filters have been used and their parameters were optimized to maximize the accuracy and performance in results.

In the context of this work a fully modular computational tool has been designed and implemented using GSP techniques and mathematical algorithms able to detect TSSs with high accuracy.

The method was tested in real human cells (H9 line) with data downloaded from FANTOM5 repository and the accuracy has been compared with other algorithms and the ground truth. All the results are presented in this study.

SUBJECT AREA: Computational Biology, Bioinformatics

KEYWORDS: Transcription start sites, genomics, signal processing, CAGE, FFT, GSP

ΠΕΡΙΛΗΨΗ

Η επεξεργασία σήματος σε επίπεδο γονιδιώματος μπορεί να λύσει διάφορα βιολογικά προβλήματα με χαμηλό υπολογιστικό κόστος. Υπάρχουν πολλοί μαθηματικοί αλγόριθμοι οι οποίοι χρησιμοποιούνται κυρίως για την αναγνώριση των γονιδίων και τη σύγκριση μεταξύ ακολουθιών, χρησιμοποιώντας αρκετές αναπαραστάσεις του DNA που δεν εξελίχθηκαν λόγω έλλειψης αποτελεσματικότητας.

Η γνώση της ακριβούς θέσης των θέσεων έναρξης μεταγραφής (TSS), η οποία είναι η θέση όπου η μεταγραφή ξεκινά στο 5'- άκρο μιας αλληλουχίας γονιδίου σε ένα RNA μόριο, είναι απαραίτητη για την αναγνώριση των ρυθμιστικών περιοχών που το επηρεάζουν. Στη βιβλιογραφία έχουν αναφερθεί πολλές προσεγγίσεις σχετικά με τον εντοπισμό των θέσεων έναρξης μεταγραφής.

Αυτή η μελέτη παρουσιάζει μια νέα μέθοδο για την αναγνώριση των θέσεων έναρξης μεταγραφής (TSS) από τα δεδομένα CAGE (Cap Analysis of Gene Expression) και την εφαρμογή χαρακτηριστικών και τεχνικών δανεισμένων από την επεξεργασία σήματος, που αποσκοπούν επίσης στην αναγνώριση των θέσεων έναρξης μεταγραφής. Παρουσιάζεται μία νέα μέθοδος για την αναπαράσταση των νουκλεοτιδίων ικανή να εξάγει και να παρουσιάσει την πληροφορία σε μορφή σήματος στο πεδίο του χρόνου.

Τα σήματα μεταφέρθηκαν από το πεδίο του χρόνου στο πεδίο των συχνοτήτων, οπού μπορεί να γίνει φιλτράρισμα αυτών και να απομακρυνθεί ο θόρυβος με αποτελεσματικό τρόπο. Ύστερα τα σήματα επανήλθαν στο πεδίο του χρόνου. Χρησιμοποιήθηκαν αρκετά φίλτρα και οι παράμετροί τους βελτιστοποιήθηκαν για να μεγιστοποιήσουν την ακρίβεια και την απόδοση των αποτελεσμάτων.

Στα πλαίσια αυτής της εργασίας έχει σχεδιαστεί και υλοποιηθεί ένα πλήρως παραμετροποιήσιμο υπολογιστικό εργαλείο με τη χρήση τεχνικών επεξεργασίας σήματος και μαθηματικών αλγορίθμων ικανών να ανιχνεύουν θέσεις έναρξης μεταγραφής με υψηλή ακρίβεια.

Η μέθοδος αυτή δοκιμάστηκε σε πραγματικά ανθρώπινα κύτταρα (H9) με δεδομένα από το αποθετήριο FANTOM5 και η ακρίβειά τους συγκρίθηκε με άλλους αλγορίθμους όπως επίσης και την πραγματικότητα (Ground Truth). Όλα τα αποτελέσματα παρουσιάζονται στη παρούσα μελέτη.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Υπολογιστική Βιολογία, Βιοπληροφορική

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Θέσεις έναρξης μεταγραφής, γονιδίωμα, επεξεργασία σήματος, CAGE, FFT, GSP