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

Δημιουργία πακέτου R ανάλυσης γονιδιακής έκφρασης κυττάρων που αναγνωρίζει τις κυτταρικές καταστάσεις και ανακατασκευάζει ρυθμιστικά δίκτυα για τις πιθανές μεταβάσεις καταστάσεων με μη-εποπτευόμενη μηχανική μάθηση

Ευθυμία Μαλέσιου

Επιβλέπων: **Ηλίας Μανωλάκος,** Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

ΑΘΗΝΑ

ΔΕΚΕΜΒΡΙΟΣ 2019

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS & TELECOMMUNICATIONS

POSTGRADUATE PROGRAM
"INFORMATION TECHNOLOGIES IN MEDICINE AND BIOLOGY"

MASTER THESIS

R package for single-cell data analysis that identifies cellular states and reconstructs regulatory networks for potential state transitions with unsupervised machine learning

Efthymia Malesiou

Supervisor: **Elias Manolakos,** Professor, Department of Informatics and Telecommunications, National and Kapodistrian University of Athens

ATHENS

DECEMBER 2019

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Δημιουργία πακέτου R ανάλυσης γονιδιακής έκφρασης κυττάρων που αναγνωρίζει τις κυτταρικές καταστάσεις και ανακατασκευάζει ρυθμιστικά δίκτυα για τις πιθανές μεταβάσεις καταστάσεων με μη-εποπτευόμενη μηχανική μάθηση

Ευθυμία Μαλέσιου

Α.Μ.: ΠΙΒ0171

ΕΠΙΒΛΕΠΩΝ: **Ηλίας Μανωλάκος,** Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Martin Reczko,** Ειδικός Λειτουργικός Επιστήμονας Α', Ερευνητικού Κέντρου Βιοϊατρικών Επιστημών «Αλέξανδρος Φλέμινγκ»
Έμα Αναστασιάδου, Ερευνήτρια Δ', Ίδρυμα Ιατροβιολογικών Ερευνών, Ακαδημίας Αθηνών
Ηλίας Μανωλάκος, Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Δεκέμβριος 2019

MASTER THESIS

R package for single-cell data analysis that identifies cellular states and reconstructs regulatory networks for potential state transitions with unsupervised machine learning

Efthymia Malesiou

S.R.N.: ΠΙΒ0171

SUPERVISOR: **Elias Manolakos,** Professor, Department of Informatics and Telecommunications, National and Kapodistrian University of Athens

EXAMINATION COMMITTEE:

Ema Anastasiadou, Researcher-Lecturer Level, Biomedical Research Foundation, Academy of Athens
Elias Manolakos, Professor, Department of Informatics and Telecommunications, National and Kapodistrian University of Athens
Martin Reczko, Staff Researcher professor level, Biomedical Sciences Research Center "Alexander Fleming"

December 2019

ΠΕΡΙΛΗΨΗ

Η δυνατότητα ποσοτικοποίησης κι ανάλυσης των προφίλ γονιδιακής έκφρασης σε επίπεδο μονήρων κυττάρων (single-cells), έχει επιτρέψει τη μελέτη της ετερογένειας των κυτταρικών πληθυσμών στους ιστούς, την αναγνώριση σπάνιων καταστάσεων και τη διερεύνηση του ρόλου τους και των αποτελεσμάτων της αλληλεπίδρασής τους με το μικρο-περιβάλλον. Ιδιαίτερο ενδιαφέρον, παρουσιάζει η μελέτη των δυναμικών μεταβάσεων ή τροχιών που σχηματίζονται μεταξύ δύο κυτταρικών καταστάσεων. Πρόσφατα, αναπτύχθηκαν αρκετοί αλγόριθμοι για την ανακατασκευή τροχιών, οι κύριες διαφορές μεταξύ των οποίων, είναι η απαίτηση εκ των προτέρων πληροφορίας, ο τρόπος διαμόρφωσης της τοπολογίας, η διάταξη των κυττάρων και το μαθηματικό πλαίσιο στο οποίο βασίζονται.

Στη δημοσίευση των Τσακανίκα Π., Μανατάκη Δ. και Μανωλάκου Η.Σ., «Machine learning methods to reverse engineer dynamic gene regulatory networks governing cell state transitions», bioRxiv, 2018 (DOI: http://dx.doi.org/10.1101/264671), περιγράφεται ένα πιθανοτικό πλαίσιο μη-εποπτευόμενης μηχανικής μάθησης για την ανακατασκευή δυναμικών γονιδιακών ρυθμιστικών δικτύων που καθοδηγούν τη μετάβαση μεταξύ κυτταρικών καταστάσεων, εισαγάγοντας, ταυτόχρονα, την έννοια των μικρο-καταστάσεων σε μία τροχιά. Για τη δημιουργία του προτύπου που περιγράφει το «επιγενετικό τοπίο», χρησιμοποιείται ένα μείγμα κανονικών κατανομών με τις εκ των υστέρων πιθανότητες που προκύπτουν να καθορίζουν τις κυτταρικές καταστάσεις και τις πιθανές μεταβάσεις μεταξύ τους. Περαιτέρω, σε κάθε τροχιά μετάβασης που σχηματίζεται (μετάβαση από την κατάσταση "έναρξης" προς την κατάσταση "προορισμού"), προσδιορίζονται διαδοχικές μικρο-καταστάσεις (φάσεις μετάβασης) κι αναγνωρίζονται τα κύρια γονίδια – ρυθμιστές, καταλήγοντας στη δημιουργία στοχευμένων αιτιατών γονιδιακών ρυθμιστικών δικτύων ανά μικρο-κατάσταση.

Η παρούσα διπλωματική εργασία, αφορά στη δημιουργία πακέτου R (MLscAN: Machine Learning single-cell ANalytics) που βασίζεται στη μεθοδολογία της παραπάνω δημοσίευσης (Tsakanikas P. et al., 2018), με δυνατότητα ευέλικτης εκτέλεσης όλων των βημάτων, με μόνη απαιτούμενη είσοδο, τα προ-επεξεργασμένα δεδομένα έκφρασης. Εκτός των προκαθορισμένων επιλογών, δίνεται η ευχέρεια στους χρήστες να ενσωματώσουν σε οποιοδήποτε βήμα της διαδικασίας, δικούς τους αλγορίθμους ή ήδη διαθέσιμα αποτελέσματα, αλλά, και να παρέμβουν μετά τη δημιουργία του προτύπου MLscAN, τροποποιώντας στοιχεία στοχευμένα. Το πακέτο, μπορεί να χρησιμοποιηθεί για την παραγωγή κι οπτικοποίηση των αποτελεσμάτων ανάλυσης σε διαφορετικά στάδια της ροής επεξεργασίας· από τη διερεύνηση του προ-επεξεργασμένου πίνακα δεδομένων έως τη μείωση της διαστατικότητας, τον προσδιορισμό των κυτταρικών καταστάσεων και των πιθανών μεταβάσεων, την εξαγωγή των τροχιών και των μικρο-καταστάσεων, την αναγνώριση των κύριων γονιδίων και την κατασκευή των αιτιατών γονιδιακών ρυθιστικών δικτύων στο επίπεδο της μικρο-κατάστασης, με χρήση μη-εποπτευόμενων μεθοδολογιών μηχανικής μάθησης.

Τέλος, το πακέτο R χρησιμοποιήθηκε στην εργασία, για την ανάλυση ενός δημοσιευμένου συνόλου δεδομένων που αφορά στην τροχιά από-διαφοροποίησης β-κυττάρων των νησιδίων του Langerhans ατόμων με σακχαρώδη διαβήτη τύπου 2, με στόχο τη διερεύνηση των αποτελεσμάτων που παραγάγονται σε σχέση με τις επιλεγμένες παραμέτρους.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: μηχανική μάθηση, ανάλυση δεδομένων μονήρων κυττάρων, βιοπληροφορική

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: μονήρη κύτταρα, μετάβαση μεταξύ καταστάσεων, επιγενετικό τοπίο, τροχιά, μικρο-κατάσταση, γονιδιακό ρυθμιστικό δίκτυο

ABSTRACT

Our ability to measure and analyze gene expression profiles at the single-cell level has enabled the study of the heterogeneity of cell populations in tissues, the identification of rare cell states, as well as their role a formed between pairs of cell states. Recently, many trajectory inference algorithms have been proposed; their main differences lie in requiring or not prior information, the methodology applied to determine the topology, the ordering of the cells and the mathematical frameworks they are based upon.

In their recent paper, “Machine learning methods to reverse engineer dynamic gene regulatory networks governing cell state transitions”, bioRxiv, 2018 (DOI: http://dx.doi.org/10.1101/264671), Tsakanikas P., Manatakis D. and Manolakos E.S., have proposed a probabilistic machine learning framework for the reconstruction of dynamic gene regulatory networks (GRNs) governing cell state transitions, without supervision, while introducing the concept of a trajectory’s micro-states. Furthermore, each transition’s trajectory (from a “ground” cell-state to a “landing” cell-state), is partitioned into consecutive micro-states, and after the transition’s key-genes are identified, a causal GRN can be inferred per micro-state.

The main objective of this thesis was the development of an R package (MLscAN: Machin the methodology of the aforementioned article, to execute the full workflow, only requiring the pre-processed expression data as input. Besides the default settings, the users may incorporate, at each stage of the process, their own algorithms or previously generated results. Also, the users may focus on any object and specifically alter it. The package can be used to generate and visualize the results of the top-down analysis at different stages of the workflow, from the pre-processed data matrix exploration to dimensionality reduction, states and possible transitions identification, trajectories and micro-states extraction, key-genes identification and causal GRNs inference down to the micro-state level, based on unsupervised machine learning methods.

Finally, the developed R package was used to analyse a published dataset concerned with the dedifferentiation trajectory of β-cells of the islets of Langerhans of subjects with type 2 diabetes mellitus, aiming at exploring the results generated in conjunction with the selected parameters.

SUBJECT AREA: machine learning, single-cell data analysis, bioinformatics

KEYWORDS: single-cells, state transition, epigenetic landscape, trajectory, micro-state, gene regulatory networks