NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATION

POSTGRADUATE PROGRAM
"DATA SCIENCE AND INFORMATION TECHNOLOGIES"
SPECIALIZATION
“BIOINFORMATICS – BIOMEDICAL DATA SCIENCE”

MASTER THESIS

MLscAN
A flexible tool for single-cell data analysis pipelines and model selection using unsupervised machine learning methods

Arsenios P. Chatzigeorgiou

Supervisor: **Elias S. Manolakos,** Professor, Department of Informatics and Telecommunication, National and Kapodistrian University of Athens

ATHENS

JULE 2021

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΑΣ"
ΕΙΔΙΚΕΥΣΗ
“ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ – ΕΠΙΣΤΗΜΗ ΒΙΟΙΑΤΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ”

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

MLscAN
Εργαλείο ανάλυσης δεδομένων μεμονωμένων κυττάρων και εξερεύνησης εναλλακτικών μοντέλων με χρήση μη-εποπτικών μεθόδων μηχανικής μάθησης

Αρσένιος Π. Χατζηγεωργίου

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

ΑΘΗΝΑ

ΙΟΥΛΙΟΣ 2021

MASTER THESIS

MLscAN
A flexible tool for single-cell data analysis pipelines and model selection using unsupervised machine learning methods

Arsenios P. Chatzigeorgiou

SRN: DS2.18.0019

Supervisor: **Elias S. Manolakos,** Professor, Department of Informatics and Telecommunication, National and Kapodistrian University of Athens

EXAMINATION COMMITEE:
Elias S. Manolakos, Professor , Department of Informatics and Telecommunication, National and Kapodistrian University of Athens
Ema Anastasiadou, Investigator - Assistant Professor Level, Biomedical Research Foundation of the Academy of Athens (BRFAA)
Dimitris Konstantopoulos, Postdoctoral Researcher, Biomedical Sciences Research Center "Alexander Fleming"

July **2021**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

MLscAN
Εργαλείο ανάλυσης δεδομένων μεμονωμένων κυττάρων και εξερεύνησης εναλλακτικών μοντέλων με χρήση μη-εποπτικών μεθόδων μηχανικής μάθησης

Αρσένιος Π. Χατζηγεωργίου

Α.Μ.: DS2.18.0019

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

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:
Ηλίας Σ. Μανωλάκος, Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Έμα Αναστασιάδου, Ερευνητής Γ', Ίδρυμα Ιατροβιολογικών Ερευνών Ακαδημίας Αθηνών
Δημήτρης Κωνσταντόπουλος, Μεταδιδακτορικός Ερευνητής, Ερευνητικό Κέντρο Βιοϊατρικών Επιστημών "Αλέξανδρος Φλέμινγκ"

Ιούλιος 2021

ABSTRACT

Single-cell RNA-seq (scRNA-seq) technologies provide us with the ability to extract snapshots of gene expression at very high resolution, i.e., at the cellular level. One of the fields that scRNA-seq revolutionizes is developmental biology, i.e., studying the developmental stages and their gene regulation. The same is true for cancer biology where the stages represent the progression of cancer. Specifically, the introduction of Trajectory Inference (TI), i.e., the study of the transition of cells from one state to the other, allows us to reconstruct the “epigenetic landscape” of a biological process suggested by a complex scRNA-seq dataset. A lot of TI methods have been proposed, but few of them provide an unsupervised probabilistic approach.

MLscAN (Machine Learning Single-Cell ANalytics) is asset of methods and R-package for unsupervised machine learning single-cell data analysis using Gaussian Mixture Models. MLscAN identifies the different cell-states in a scRNA-seq dataset and infers possible pairwise trajectories between them without requiring any prior knowledge. The pipeline can partition state-to-state transitions into phases (micro-states), identify the key genes governing the state transitions, and reconstruct their Gene Regulatory Networks (GRNs). The analysis is fully automated, and the only required input is the pre-processed expression matrix. Yet, the user may intervene in the pipeline on multiple occasions by adding any prior knowledge, using their own alternative algorithms, or manually fine-tuning pipeline parameters. MLscAN emphasizes effective model space exploration, identifying the optimal model that “best” describes the data and providing powerful insights via effortlessly produced advanced visualizations at every stage of the analysis.

The main objectives of this thesis were the improvement of the package’s flexibility, scalability, and robustness, with a focus on model exploration. New methods and capabilities were introduced, pre-existing options were enriched, and defaults were calibrated. The second goal was the generation of a relevant use-case with a well-known large dataset to demonstrate the utility of implemented MLscAN additions, and improvements. The final objective was the thorough evaluation of MLscAN relative to state-of-the-art trajectory inference (TI) methods, following benchmark analysis pipelines, where MLscAN proved to be a very competitive TI method.

SUBJECT AREA: unsupervised machine learning, single-cell RNA-seq analytics, bioinformatics

KEYWORDS: single-cells, state transitions, epigenetic landscape, trajectory inference, gene regulatory network, R package

ΠΕΡΙΛΗΨΗ

Οι τεχνολογίες αλληλούχισης RNA μεμονωμένων κυττάρων (single-cell RNA-seq), έδωσαν την δυνατότητα να ανιχνεύσουμε τη γονιδιακή έκφραση μιας δεδομένης στιγμής με μεγάλη διακριτική ικανότητα, μέχρι και το κυτταρικό επίπεδο. Ένα από τα ερευνητικά πεδία όπου εξώθησε η συγκεκριμένη τεχνολογία είναι η αναπτυξιακή βιολογία και η μελέτη της ρύθμισης των αναπτυξιακών σταδίων. Αντίστοιχα αλματώδης είναι η βελτίωση στην μελέτη των ετερογενών καρκινικών τύπων . Συγκεκριμένα, η εξαγωγή των τροχιών μεταξύ των κυτταρικών τύπων οδηγεί στην αναπαράσταση ενός εποπτικού επιγενετικού τοπίου μιας βιολογικής διεργασίας, που οδηγείται από τα δεδομένα. Πολλές μέθοδοι εξαγωγής τροχιών έχουν δημιουργηθεί, λίγες όμως ακολουθούνε μια μη-εποπτευόμενη και πιθανοτική προσέγγιση.

Το MLscAN (Machine Learning Single-Cell Analytics) είναι ένα πακέτο της γλώσσας R για μη-εποπτευόμενη ανάλυση δεδομένων μεμονωμένων κυττάρων με χρήση Μεικτών Κανονικών Κατανομών. Το MLscAN αναγνωρίζει διαφορετικές κυτταρικές καταστάσεις σε ένα σύνολο δεδομένων single-cell RNA-seq, και εξάγει πιθανά συζυγή ζεύγη τροχιών μεταξύ των καταστάσεων, χωρίς την απαίτηση κάποιας πρότερης γνώσης. Η ροή του αλγορίθμου, αναγνωρίζει τις μικρο-καταστάσεις εντός μιας τροχιάς, ανιχνεύει τα κύρια γονίδια και τέλος στην εξάγει τα Ρυθμιστικά Γονιδιακά Δίκτυά τους. Η ανάλυση είναι πλήρως αυτοματοποιημένη και η μόνη απαιτούμενη είσοδος είναι ο προ-επεξεργασμένος πίνακας έκφρασης. Δίνεται όμως δυνατότητα στον χρήστη να παρέμβει σε κάθε ένα από τα στάδια της μεθόδου με την προσθήκη πρότερης γνώσης ή πληροφορίας, χρήση διαφορετικών αλγορίθμων ή ρύθμιση των υπαρχόντων παραμέτρων. Το MLscAN δίνει έμφαση στην εξερεύνηση των εναλλακτικών μοντέλων και την αναγνώριση του βέλτιστου που ερμηνεύει «καλύτερα» τα δεδομένα.

Στην παρούσα εργασία, οι κύριοι στόχοι ήταν η βελτιστοποίηση της ευελιξίας, κλιμάκωσης και ευρωστίας. Προστέθηκαν νέες μέθοδοι, εμπλουτίστηκαν προϋπάρχουσες επιλογές και επαναρρυθμίστηκαν οι προεπιλογές. Δεύτερος στόχος ήταν η δημιουργία ενός αρχείου υπόθεσης χρήσης που θα καταδεικνύονταν οι δυνατότητες, προσθήκες και βελτιώσεις του αλγορίθμου. Τέλος, το MLscAN αξιολογήθηκε με άλλους αλγορίθμους εξαγωγής τροχιών, με βάση τα πρότυπα αξιολόγησης TI μεθόδων. Από τα αποτελέσματα φάνηκε ότι το MLscAN ανταγωνίζεται τις κορυφαίες μεθόδους.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: μηχανική μάθηση, ανάλυση δεδομένων single-cell RNA-seq

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: μεμονωμένα κύτταρα, επιγενετικό τοπίο, εξαγωγή τροχιών, ρυθμιστικά γονιδιακά δίκτυα