NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATION

INTERDISCIPLINARY POSTGRADUATE PROGRAM
"INFORMATION TECHNOLOGIES IN MEDICINE AND BIOLOGY"

MASTER THESIS

R package development for the analysis and visualization of single-cell data extracted from bacterial time-lapse microscopy cell-movies

Viktoria N. Stefanou

Supervisor: **Dr. Elias Manolakos,** Professor Level, National and Kapodistrian University of Athens (UoA), Department of Informatics and Telecommunications (DIT)

ATHENS

OCTOBER 2018

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

Ανάπτυξη πακέτου R για την ανάλυση και οπτικοποίηση δεδομένων επιπέδου μονήρων κυττάρων που εξάγονται από βακτηριακές ταινίες time-lapse μικροσκοπίας

Βικτώρια Ν. Στεφάνου

Επιβλέπων: **Δρ. Ηλίας Μανωλάκος,** Καθηγητής, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών (ΕΚΠΑ), Τμήμα Πληροφορικής και Τηλεπικοινωνιών

ΑΘΗΝΑ

ΟΚΤΩΒΡΙΟΣ 2018

MASTER THESIS

R package development for the analysis and visualization of single-cell data extracted from bacterial time-lapse microscopy cell-movies

Viktoria N. Stefanou

S.N.: ΠΙΒ0158

ADVISOR: **Dr. Elias Manolakos,** Professor, National and Kapodistrian University of Athens (UoA), Department of Informatics and Telecommunications (DIT)

EXAMINATION COMMITTEE:

Dr. Elias Manolakos, Professor, National and Kapodistrian University of Athens (UoA), Department of Informatics and Telecommunications (DIT)

Dr. Ioannis Emiris, Professor, National and Kapodistrian University of Athens (UoA), Department of Informatics and Telecommunications (DIT)

Dr. Ema Anastasiadou, Researcher - Lecturer Level, Biomedical Research Foundation of the Academy of Athens (BRFAA)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάπτυξη πακέτου R για την ανάλυση και οπτικοποίηση δεδομένων επιπέδου μονήρων κυττάρων που εξάγονται από βακτηριακές ταινίες time-lapse μικροσκοπίας

Βικτώρια Ν. Στεφάνου

Α.Μ.: ΠΙΒ0158

ΕΠΙΒΛΕΠΩΝ: **Δρ. Ηλίας Μανωλάκος,** Καθηγητής, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών (ΕΚΠΑ), Τμήμα Πληροφορικής και Τηλεπικοινωνιών

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΣΤΡΟΠΗ: **Δρ. Ηλίας Μανωλάκος,** Καθηγητής, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών (ΕΚΠΑ), Τμήμα Πληροφορικής και Τηλεπικοινωνιών

Δρ. Ιωάννης Εμίρης, Καθηγητής, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών (ΕΚΠΑ), Τμήμα Πληροφορικής και Τηλεπικοινωνιών

Δρ. Έμα Αναστασιάδου, Ερευνήτρια Δ', Ίδρυμα Ιατροβιολογικών Ερευνών Ακαδημίας Αθηνών (ΙΙΒΕΑΑ)

ABSTRACT

Time-lapse microscopy is an essential imaging method for monitoring the evolution of living bacteria at the single-cell level. It enables researchers to capture detailed information about the morphology and gene expression of each cell at every time point of an experiment. The big data generated by the analysis of such videos (bacterial "cell-movies") can help us study the growth dynamics of bacterial communities and identify the sources and role of intra- and inter-subpopulation heterogeneity. Recent research has highlighted the importance of this technology for investigating the role of biological “noise” in gene regulation, cell growth and proliferation etc. Single-cell data analysis of complex cell-movies, with multiple micro-colonies and thousands of cells in the field of view that may interact during the experiment can shed light on important phenomena for human health, such as the competition between pathogens and benign microbiome cells, dormant cells (persisters) emergence under different stress conditions, biofilms formation that promote pathogens’ resistance etc.

However, highly accurate and fully automated biomage analysis and single-cell data analytics methods remain elusive and will be required before we can really exploit the abundance of cell-movies big data. In this graduate thesis we present the capabilities of ViSCAR, a package we developed in the R programming language which along with our group's ongoing efforts on bacterial image analysis (BaSCA project) allows the visualization and single-cell statistical analysis of data derived from complex bacterial cell movies.

Users of the R package can easily explore visually and correlate the spatiotemporal trends of single-cell attributes at different levels of community organization (i.e. whole population, colony, generation, etc.), construct the forest of lineage and generation trees, discover possible epigenetic information transfer across cell generations, extract mathematical and statistical models describing various stochastic phenomena (e.g. cell growth, cell division), and even identify and correct unavoidable errors produced during the image analysis of a complex movie with thousands of cells. All these unique functionalities empower research towards deciphering the dynamic behavior of heterogeneous microbial communities and uncovering mechanisms that lead to specific phenotypes. To the best of our knowledge there is no other similar R package available in the literature today. We plan to make the R package freely available to the scientific community once it is completed.

SUBJECT AREA: Computational Biology

KEYWORDS: time-lapse microscopy, bacteria image analysis, lineage tree construction, cytometry, single-cell analytics, visualization

ΠΕΡΙΛΗΨΗ

Η time-lapse μικροσκοπία είναι μια εξέχουσα απεικονιστική μέθοδος για την παρακολούθηση της εξέλιξης ζωντανών βακτηρίων μέσα στο χρόνο σε επίπεδο μονήρων κυττάρων (single-cell). Μας επιτρέπει να αποτυπώσουμε λεπτομερείς πληροφορίες σχετικά με τη μορφολογία και την έκφραση γονιδίων κάθε κυττάρου σε κάθε χρονική στιγμή ενός πειράματος. Τα μεγάλα δεδομένα που παρέχονται από την επεξεργασία τέτοιων βίντεο (bacterial cell movies) μπορούν να μας βοηθήσουν να μελετήσουμε τη δυναμική των κυτταρικών κοινωνιών και να προσδιορίσουμε τις πηγές και τη σημασία της ετερογένειας που παρουσιάζουν οι κυτταρικοί υποπληθυσμοί, τόσο εσωτερικά όσο και μεταξύ τους. Πρόσφατες έρευνες επισημαίνουν τη χρήση και τη σημασία αυτής της τεχνολογίας για τη διερεύνηση του ρόλου του βιολογικού “θορύβου” στη ρύθμιση γονιδίων, την ανάπτυξη και τη διαίρεση των κυττάρων κ.ά. Η ανάλυση περίπλοκων κυτταρικών ταινιών σε επίπεδο μονήρων κυττάρων, με πολλαπλές μικρο-αποικίες στο οπτικό πεδίο και χιλιάδες κύτταρα που αλληλεπιδρούν, μπορεί να διαλευκάνει σημαντικά φαινόμενα για την ανθρώπινη υγεία, όπως π.χ. τον ανταγωνισμό μεταξύ παθογόνων και "αγαθών" κυττάρων του μικροβιώματος, την εμφάνιση μεταβολικά “ανενεργών” βακτηρίων (persisters) σε συνθήκες στρες και το σχηματισμό βιοϋμενίων (biofilms) που συνεισφέρει στη ανθεκτικότητα των παθογόνων στα αντιβιωτικά.

Ωστόσο, προκειμένου να μπορέσουμε πραγματικά να αξιοποιήσουμε την αφθονία αυτών των μεγάλων δεδομένων, η ύπαρξη αυτοματοποιημένων τεχνικών ανάλυσης βιοιατρικής εικόνας μεγάλης ακρίβειας είναι απαραίτητη, όπως επίσης και μεθόδων ανάλυσης δεδομένων σε επίπεδο μονήρων κυττάρων. Σε αυτή την διπλωματική εργασία παρουσιάζουμε τις δυνατότητες του ViSCAR, ενός πακέτου που αναπτύξαμε στη γλώσσα προγραμματισμού R το οποίο επιτρέπει τη στατιστική ανάλυση και οπτικοποίηση δεδομένων single-cell που προέρχονται από την ανάλυση σύνθετων βακτηριακών time-lapse ταινιών. Η προσπάθεια αυτή εντάσσεται στο πλαίσιο της ερευνητικής δραστηριότητα της ομάδας μας στην ανάλυση βακτηριακών εικόνων μεγάλης πολυπλοκότητας (Bacterial Single-Cell Analytics, BaSCA, project).

Οι χρήστες του πακέτου R μπορούν εύκολα να συσχετίσουν και να διερευνήσουν οπτικά τις χωροχρονικές τάσεις των κυτταρικών χαρακτηριστικών σε διάφορα επίπεδα οργάνωσης της βακτηριακής κοινότητας (π.χ. σε επίπεδο συνολικού πληθυσμού, αποικίας, γενιάς κλπ.), να κατασκευάσουν το δάσος των δέντρων γενεαλογίας και κυτταρικών διαιρέσεων, να ανακαλύψουν πιθανή επιγενετική μεταφορά πληροφορίας μεταξύ των γενιών, να εξάγουν αυτόματα μαθηματικά και στατιστικά μοντέλα που περιγράφουν ποικίλα στοχαστικά φαινόμενα (π.χ. την κυτταρική ανάπτυξη και διαίρεση), ακόμη και να εντοπίζουν και να διορθώνουν σφάλματα που αναπόφευκτα συμβαίνουν κατά την ανάλυση μιας

πολύπλοκης ταινίας με χιλιάδες κύτταρα. Όλες αυτές οι μοναδικές λειτουργίες ενισχύουν την έρευνα για την αποκρυπτογράφηση της δυναμικής συμπεριφοράς των μικροβιακών κοινοτήτων και την ανάδειξη και χαρακτηρισμό των μηχανισμών που οδηγούν σε συγκεκριμένους φαινότυπους. Από όσο γνωρίζουμε δεν υπάρχει στη διεθνή βιβλιογραφία παρόμοιο πακέτο R, το οποίο σκοπεύουμε να διαθέσουμε ελεύθερα στην επιστημονική κοινότητα μόλις ολοκληρωθεί.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Υπολογιστική Βιολογία

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: time-lapse μικροσκοπία, ανάλυση βακτηριακής εικόνας, κατασκευή δέντρων γενεαλογίας, κυτταρομετρία, στατιστική ανάλυση δεδομένων επιπέδου μονήρων κυττάρων, οπτικοποίηση