“Development of computational methods to investigate the stochastic dynamic behavior of microbial communities”.

Abstract

Methods and software tools for the analysis of cell movies that allow the accurate estimation of bacterial attributes at the single-cell level, the visualization and statistical modeling of the extracted cell attributes, and creating in silico experiments simulating microbial communities growth, remain currently a challenge in computational systems biology. Methods reported in the literature are lacking in many respects. They exhibit low success rates in cell segmentation of dense cell movies with thousands of cells, often require laborious parameterization, and are characterized as not user-friendly by biologists with no bioimage analysis expertise. Moreover, there is no end-to-end computational pipeline to automatically analyze cell movies, model the variability of extracted cell attributes, explore cell diversity visually, and facilitate the realistic simulation of microbial communities' behavior while considering the inherent stochasticity of the underlying biological phenomena.

This doctoral dissertation's main goal was to fill this gap by developing an end-to-end computational strategy with three stages for studying microbial communities' dynamic growth behavior. The first stage provides an automated bioimage analysis platform, which extracts, collects, and organizes the estimated cell attributes data hierarchically, without the need for human-user intervention. The second stage provides an information system for single-cell analytics and visualization. It allows the user to visualize in different ways and infer models for the cell attributes extracted by the image analysis stage at different community organization levels. The third stage supports the development of multiscale "digital twin" model realizations to investigate in silico microbial communities' growth. It can consider the individual cells' genetic "program", the microenvironment's conditions, and the simulated biological phenomena' inherent stochasticity for high fidelity purposes.

First, we developed a new cell tracking algorithm inspired by motion estimation for video compression that successfully associates cell instances in consecutive frames and can accurately construct dense and deep forests of lineage trees with many cell colonies and generations. Our method can map on the extracted trees cell instance attributes (e.g., cell surface, length, width) and cell life attributes (e.g., cell division time) with high fidelity. The very high bacterial matching accuracy it achieves (98.7%) for complex cell movies exceeds that of prior methods.

Our complete, end-to-end cell movies analysis methodology, codenamed BaSCA (Bacterial Single-Cell Analytics), covers cell

segmentation, cell tracking, and lineage trees construction for complex cell movies. BaSCA was thoroughly evaluated using datasets generated by different labs and achieved an F-measure rate of 98%. The F-measure remains very high (over 96.7%) even for overcrowded cell movies with many merging colonies and thousands of bacteria in the field of view of the microscope.

The information extracted from the cell movie analysis is organized in a relational bio-database, allowing data mining at the single-cell level (single-cell analytics). Another innovative feature is creating different "views" in the data for the direct and friendly visualization of information (visual analytics). In this way, the system allows the user to select subpopulations (colonies, generations, relative cells in the trees) and perform statistical analyses and best distributions parameter estimation. Additionally, this work also contributed to an R package creation codenamed ViSCAR (Visualization and Single-Cell Analytics in R). Using Viscar, one can also correct inevitable segmentation and tracking errors introduced by the image analysis of dense cell movies.

Next, we had to enhance the capabilities of CellModeller, a popular open-source systems biology tool, to be able to recreate in silico with high fidelity the physical interaction of cells as they grow and divide to form a dense bacterial community while taking into account the genetic "program" of every individual cell, the underlying stochasticity of cell properties (e.g., division time), the microenvironment conditions, potential cell motion, etc. Thus, it became possible to generate in silico experiments of microbial communities growing in two dimensions based on stochastic atomic evolution models for each cell entering the simulation. We created a "digital twin" microbial community prototype to implement the proposed unified strategy for Salmonella's case (Salmonella enterica serovar Typhimurium). To that end, the regulatory networks of gene expression related to the mechanism of intercellular chemical communication (quorum sensing), the mechanism of virulence development of S. Typhimurium, and the interaction between the two were integrated into the individual-based cell models we developed. The proof of concept community level "digital twin" simulation model we developed also considers the inherent stochasticity that governs single-cells' growth and division, which affects the patterns of switching their phenotypic state from non-virulent to virulent. Using the digital twin, it is possible to investigate different scenarios in silico.

Overall the computational pipeline we have developed can automatically image-analyze cell movies from live-cell microscopy experiments, display efficiently and extract statistical models of single-cell attributes, and utilize them to create realistic synthetic movies. It can also stochastically simulate a microbial community's

behavior in space and time. This was demonstrated with a digital twin of a community of pathogen S. Typhimurium cells, growing in a microenvironment that promotes stochastic phenotypic switching of the single-cells based on their own personalized genetic network "logic."

«Ανάπτυξη υπολογιστικών στρατηγικών για τη μελέτη της δυναμικής συμπεριφοράς μικροβιακών κοινοτήτων».

Περιλήψη

Η ανάπτυξη μεθόδων και εργαλείων πληροφορικής για την αξιοποίηση των κυτταρικών ταινιών (cell movies) που να επιτρέπουν την ακριβή εξαγωγή των χαρακτηριστικών των βακτηρίων σε επίπεδο του μεμονωμένου κυττάρου (single-cell), την οπτικοποίηση και τη στατιστική μοντελοποίηση αυτών των χαρακτηριστικών, καθώς και την διεξαγωγή πειραμάτων ανάπτυξης μικροβιακών κοινοτήτων in silico, μέσω ρεαλιστικής προσομοίωσης, αποτελεί σήμερα πρόκληση του τομέα της υπολογιστικής συστημικής μικροβιολογίας. Στη διεθνή βιβλιογραφία απαντώνται μέθοδοι που αναλύουν κυτταρικές ταινίες με βακτήρια και ποσοτικοποιούν την πληροφορία που εξάγεται από χρονοσειρές εικόνων που απεικονίζουν αναπτυσσόμενες μικροβιακές αποικίες. Παραταύτα αυτές οι μέθοδοι μειονεκτούν σε αρκετά σημεία, π.χ. παρουσιάζουν χαμηλά ποσοστά επιτυχίας κατάτμησης πολύπλοκων κυτταρικών ταινιών με χιλιάδες κύτταρα, ενώ συχνά απαιτούν λεπτομερή παραμετροποίηση και έτσι χαρακτηρίζονται ως μη φιλικές προς το χρήστη-βιολόγο. Επιπλέον είναι σαφής η έλλειψη μιας ενιαίας υπολογιστικής στρατηγικής (pipeline), ικανής να αναλύει αυτόματα κυτταρικές ταινίες, να χαρακτηρίζει με αξιόπιστο τρόπο τη στοχαστικότητα στις κυτταρικές ιδιότητες, να παρέχει δυνατότητες οπτικοποίησης της κυτταρικής ποικιλομορφίας, αλλά και να επιτρέπει τη ρεαλιστική προσομοίωση της συμπεριφοράς μικροβιακών κοινοτήτων, λαμβάνοντας υπόψιν την εγγενή στοχαστικότητα των βιολογικών φαινομένων.

Βασικό στόχο της εν λόγω διδακτορικής διατριβής λοιπόν αποτέλεσε η ανάπτυξη μιας τέτοιας ενιαίας στρατηγικής με τρία στάδια για τη μελέτη της δυναμικής συμπεριφοράς μικροβιακών κοινοτήτων. Το πρώτο στάδιο αποτελεί η ανάπτυξη αυτοματοποιημένης μεθόδου ανάλυσης κυτταρικών ταινιών, η οποία συλλέγει και οργανώνει ιεραρχικά την εξαγόμενη πληροφορία χωρίς να απαιτείται η παρέμβαση του ανθρώπου-χρήστη. Το δεύτερο στάδιο περιλαμβάνει τη δημιουργία ενός πληροφοριακού συστήματος (single-cell analytics and visualization platform) το οποίο επιτρέπει στο χρήστη να χαρακτηρίζει στατιστικά και να οπτικοποιήσει με διαφορετικούς τρόπους τις εξαγόμενες πληροφορίες από την ανάλυση εικόνας. Το τρίτο στάδιο υποστηρίζει την ανάπτυξη ενός πρότυπου «ψηφιακού διδύμου» που προσομοιώνει την ανάπτυξη μικροβιακών κοινοτήτων λαμβάνοντας υπόψιν την εγγενή στοχαστικότητα του φαινομένου αλλά και τις συνθήκες του μικροπεριβάλλοντος.

Αρχικά, αναπτύχθηκε πρωτότυπη μέθοδος αντιστοίχισης κυττάρων (cell tracking) μεταξύ συνεχόμενων στιγμιότυπων εικόνας (frames) κυτταρικών ταινιών και δημιουργίας δέντρων κυτταρικής γενεαλογίας (cell lineage tree construction). Η μέθοδος αυτή μπορεί να παρακολουθεί και να ποσοτικοποιεί τις ιδιότητες των κυττάρων (π.χ. κυτταρική επιφάνεια, μήκος, πλάτος κ.α.) ανά γενιά και απεικία. Η πολύ μεγάλη ακρίβεια

αντιστοίχισης βακτηρίων που επιτυγχάνεται (98,7%) είναι υψηλότερη σε σύγκριση με αυτή των υπαρχόντων μεθόδων.

Κατά τη διάρκεια της κατάτμησης και της παρακολούθησης των βακτηρίων (δηλαδή με τη δημιουργία των δέντρων κυτταρικής γενεαλογίας), η προτεινόμενη μεθοδολογία ανάλυσης κυτταρικών ταινιών παράγει πληθώρα «μεγάλων δεδομένων» (“big data”), τα οποία χαρακτηρίζουν κάθε μεμονωμένο κύτταρο σε κάθε χρονική στιγμή (στιγμιότυπο εικόνας) μιας πολύπλοκης κυτταρικής ταινίας με πιθανά εκατοντάδες στιγμιότυπα και χιλιάδες κύτταρα. Δημιουργήθηκε λοιπόν ένα ολοκληρωμένο σύστημα ανάλυσης υψηλής ακρίβειας το οποίο συμβάλλει καθοριστικά στην πρόοδο της συστημικής βιολογίας και της ανάλυσης δεδομένων μεγάλης κλίμακας, όπως οι ταινίες ανάπτυξης βακτηριακών κοινοτήτων πολλαπλών στιγμιότυπων. Με την ολοκλήρωση των παραπάνω σταδίων, είχε αναπτυχθεί πλέον μια ολοκληρωμένη μέθοδος ανάλυσης κυτταρικών ταινιών, η οποία έλαβε την κωδική ονομασία BaSCA (Bacterial Single-Cell Analytics). Το σύστημα BaSCA αξιολογήθηκε ενδελεχώς και επετεύχθη ιδιαίτερα υψηλό ποσοστό F-measure, 98%, γεγονός που αποδεικνύει την ευρωστία (robustness) του αλγορίθμου κατάτμησης. Σε σχέση με υπάρχοντα λογισμικά, το BaSCA επιτυγχάνει σημαντικά υψηλότερο ποσοστό F-measure στα δεδομένα από διαφορετικά εργαστήρια που χρησιμοποιήθηκαν στην αξιολόγηση. Το ποσοστό F-measure παραμένει πολύ υψηλό (πάνω από 96.7%) ακόμα και για πολύπλοκες ταινίες που παρουσιάζουν μεγάλο συνωστισμό βακτηρίων.

Η πληροφορία που μπορεί να εξαχθεί από την ανάλυση ταινιών, οργανώνεται σε σχεσιακή βάση βιοδεδομένων (relational bio-database), που επιτρέπει την εξόρυξη πληροφοριών από τα δεδομένα σε επίπεδο μεμονωμένων κύτταρων (single-cell analytics). Ένα ακόμα καινοτόμο χαρακτηριστικό, είναι η δυνατότητα δημιουργίας διαφορετικών «όψεων» στα δεδομένα, με στόχο την άμεση και φιλική οπτικοποίηση της πληροφορίας (visual analytics). Με τον τρόπο αυτό το σύστημα επιτρέπει στο χρήστη να επιλέγει υποπληθυσμούς (απεικίες, γενιές, συγγενείς στο δέντρο) και να διεξάγει στατιστικές αναλύσεις και εκτιμήσεις κατανομών. Επιπλέον για να επεκτείνουμε τις δυνατότητες οπτικοποίησης του BaSCA, η εργασία αυτή υποστήριξε και τη δημιουργία R πακέτου που ονομάστηκε ViSCAR (Visualization and Single-Cell Analytics in R).

Με στόχο την υλοποίηση in silico μοντελοποίησης της ανάπτυξης μικροβιακών κοινοτήτων σε δύο διαστάσεις με ρεαλιστικά και ατομοστραφή μοντέλα (Individual-based models), τροποποιήθηκε και επεκτάθηκε σημαντικά το λογισμικό ανοιχτού κώδικα *Cellmodeller*. Έτσι διαμορφώθηκε ένα ευέλικτο υπολογιστικό περιβάλλον στο οποίο μπορούμε να «προγραμματιζουμε» διαφορετικές συμπεριφορές κυττάρων, και να προσομοιώνουμε την ανάπτυξη μιας ποικιλόμορφης κοινωνίας ενόσω αυτά αλληλεπιδρούν στοχαστικά κάτω από διαφορετικές συνθήκες μικροπεριβάλλοντος και πιθανά αλλάζουν φαινότυπο δυναμικά. Έτσι, κατέστη δυνατή η in silico προσομοίωση μικροβιακών κοινοτήτων που αναπτύσσονται με βάση στοχαστικά ατομοστραφή μοντέλα εξέλιξης για κάθε κύτταρο που εισέρχεται στη προσομοίωση. Δημιουργήθηκε ένας «ψηφιακός δίδυμος» (digital twin) μικροβιακής κοινότητας και

εφαρμόστηκε η προτεινόμενη ενιαία στρατηγική για την περίπτωση της Σαλμονέλλας (*Salmonella enterica* serovar Typhimurium). Επιπλέον, ενσωματωθήκαν στα ατομοστραφη μοντέλα τα ρυθμιστικά δίκτυα γονιδιακής έκφρασης (gene regulatory networks) που αφορούν το μηχανισμό διακυτταρικής χημικής επικοινωνίας (quorum sensing), τον μηχανισμό ανάπτυξης της λοιμοτοξικότητας (virulence) της *S.* Typhimurium, αλλά και την μεταξύ τους αλληλεπίδραση. Στη προσομοίωση λαμβάνεται υπόψη η εγγενής στοχαστικότητα που διέπει την ανάπτυξη και διαίρεση των μεμονωμένων βακτηρίων, η οποία επηρεάζει την αλλαγή φαινοτυπικής κατάστασης των κυττάρων ως προς τη λοιμοτοξικότητα.

Επιβεβαιώθηκε έτσι η πρακτική εφαρμογή της ενιαίας προτεινόμενης στρατηγικής (pipeline) που είναι ικανή να αναλύει αυτόματα κυτταρικές ταινίες από πειραματα μικροσκοπίας ζωντανών κυττάρων (live cell imaging), να απεικονίζει με αποτελεσματικό τρόπο και να εξάγει στατιστικά μοντέλα για τα χαρακτηριστικά των κυττάρων σε αυτές. Επιπλέον επιδείχθηκε πως τα μοντέλα αυτά συμβάλουν στη δημιουργία ρεαλιστικών συνθετικών ταινιών που προσομοιώνουν στοχαστικά τη δυναμική συμπεριφορά μικροβιακών κοινοτήτων στο χώρο και στο χρόνο, όπως αυτές του παθογόνου *S.* Typhimurium, σε μικροπεριβάλλον που οδηγεί στην αλλαγή φαινοτυπικής κατάστασης (phenotypic switching) κύττατα που το καθένα ακολουθεί τη δική του εξατομικευμένη «λογική».