**Mathematical modeling of time course single cell and pooled gene expression data in cancer with CancerInSilico**

Thomas D Sherman1, Raymond Cheng2, Matthew Satriano3, Yong Tang4, Sandra Jablonski5, Michael Considine6, Emily L Flam7, Luciane T Kagohara6, Dylan Kelley7, Genevieve Stein-O'Brien8,9, Daria A Gaykalova7, Cristian Tomasetti6,10, Louis M Weiner5,and Elana J Fertig6

1 Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD USA

2 Science, Math and Computer Science Magnet Program, Poolesville High School, Poolesville, MD USA

3

4

5 Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC USA

6 Department of Oncology, Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD USA

7 Department of Otolaryngology-Head and Neck Surgery, Johns Hopkins University School of Medicine, Baltimore, MD USA

8 Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD USA

9 Lieber Institute for Brain Development, Baltimore, MD USA

10 Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD USA

**Contact:**

Elana J Fertig  
[ejfertig@jhmi.edu](mailto:ejfertig@jhmi.edu)  
@FertigLab

**Summary**

**Highlights**

* An R/Bioconductor package implementing a lattice-free mathematical model of cellular growth in cancer.
* Simulates statistical distributions of cellular growth rates with and without therapeutic agents that inhibit cellular proliferation.
* Simulates time-course gene expression data resulting from cellular growth, cellular density, and growth factor signaling to benchmark the performance of time-course bioinformatics analysis algorithms.
* Generates simulated processed and raw, single cell and pooled RNA-sequencing data with realistic error models.

**eTOC Blurb**

CancerInSilico is an R/Bioconductor package to simulate time-course transcriptional data of cancer cell growth in response to targeted therapeutics. Data is simulated as single cell and pooled RNA-sequencing data with realistic error models to benchmark the performance of bioinformatics analysis algorithms for time-course RNA-sequencing.

**Introduction**

Cancer is a disease with evolving molecular alterations in multiple cell types. As a tumor grows, changes in active cellular signaling pathways alter the tumor and its microenvironment support its progression and acquisition of therapeutic resistance [citations]. High throughput transcriptional and proteomic profiling enable unprecedented characterization cellular signaling processes in the cells comprising tumors. Analysis of such data can facilitate personalized therapeutic selection for precision medicine in clinical practice [citation]. Obtaining the genomics data serially throughput a patient’s disease course are essential to optimize therapeutic selection and mitigate acquired therapeutic resistance [citation]. Simultaneously, time course bioinformatics analysis techniques are emerging to delineate cellular composition [citations] and precise pathway activation for such dynamic therapeutic selection [citations]. However, the emerging time course genomics datasets generated from patient samples or experimentally [citations] lack a known ground truth. Therefore, mathematical models are essential to simulate time course high throughput datasets that can benchmark the performance of time-course bioinformatics algorithms in cancer.

Recently, network models of cellular signaling can accurately simulate proteomic and transcriptional data. Most prominently, the XXX model was used to generate *in silico* phospho-proteomic data to benchmark the performance of new algorithms for network inference and future predictions of proteomics response to cellular signaling activation and targeted therapeutics in the HPN DREAM8 contest [citation]. However, these network-based models are typically optimized for a single cell. As a result, they frequently disregard the pervasive intercellular interactions that alter long-term cellular signaling responses [citation – check Don’s paper!!!]. Even relatively simple covariates, such as cellular density and proliferation rates impact measured therapeutic sensitivity [citation to Hafner paper] in addition to multiple cell types that comprise tumors, which must also be accounted for in simulating time course high throughput data.

At the same time, numerous mathematical models of cellular growth in tumor progression and therapeutic response have emerged [citations]. These models utilizing techniques ranging from differential equations [citations], agent-based techniques [citations], and stochastic simulation [citations] of the distribution of cellular populations within a tumor. DESCRIBE WHERE DRASDO-HOLME FITS WITHIN THIS AND WHY IT IS APPEALING. To further capture the complexity of biological systems, numerous multiscale models linking cellular signaling to the equations of the cellular composition are emerging [citations]. However, these models often require numerous parameters leading to significant overfitting [citation to C elegans paper]. This extensive parameterization becomes even more extreme when simulating high throughput proteomic and transcriptional data. Therefore, these models that simulate high throughput data often have similar complexity to real biological systems. Thus, it is nearly as intractable to benchmark the performance of time course bioinformatics algorithms against these simulated data as it is for high throughput time course data generated from biological experiments.

Many of these models have been developed in open source software in Matlab [citations], C++ [citations], and JAVA [citations] widely accessible for the systems biology community. Some of these software packages are advanced, enabling easy drag and drop interfaces for intercellular pathways [cite virtual cell] or inclusion of new user defined classes for custom cellular behavior [cite Chaste]. However, to our knowledge none to data have been programmed in the R language most widely used by the bioinformatics community, creating a language barrier between the research communities to utilize these models to interpret genomics data.

To address these limitations, we present a new mathematical model to simulate time course transcriptional data. The underlying model for cellular growth in this model is an off-lattice, cell-center Monte Carlo mathematical model [ to Drasdo Holme citation]. We extend this model to include a distribution of cellular growth rates and cell types. Pathway activity is simulated in based upon the simulated distribution of growth factor, state in the cell cycle, and cellular type. The model also enables inclusion of a targeted therapeutic, to simulate repression in the growth factor signaling pathway that further alters the growth rate of each cellular type. Transcriptional data is simulated using annotated gene sets for each of these processes [cite MSigDB and TRANSFAC] or cellular types [cite paper with cell types] using established statistical models to simulate microarray [cite context-dependent TF] and RNA-seq [cite limma voom and Polyester] data. We validate the resulting cellular behavior and transcriptional data against time course *in vitro* data resulting from inhibition of proliferation rates with the EGFR inhibitor cetuximab in a single cell line XXX. After benchmarking thus the model for one cellular type, we apply CancerInSilico to predict the impact of multiple cellular types on the resulting transcriptional data. The model is implemented in the R/Bioconductor package CancerInSilico, and includes extensions to simulate single cell RNA-seq data. The package is developed generically, using S4 classes to enable users to easily add custom cellular behavior and signaling pathways in the simulation.

**Results**

***CancerInSilico cell-based model***

The core cellular growth model in *CancerInSilico* is an off-lattice, stochastic cell-based model. The model inputs data from three distinct classes of parameters: (1) cellular population distributions, (2) cellular properties, and (3) cellular mechanics. Each of the parameters for these classes is described in detail in **Supplemental File 1**. Briefly, cellular population distributions represent the relative density of cells and experimental conditions such as cellular synchronization or presence of a boundary. Cellular properties describe the expected cell cycle length and cell size, whereas cellular mechanics encode parameters for cellular movement and cellular response to contact inhibition. The former may be freely set by the user to reflect different cellular types and therapeutic response, while the latter are fixed from parameterization in previous studies (CITE DRASDO HOLME). Each of these parameters is input to CancerInSilico through the core function *runCellSimulation* for a specified run time, and the model outputs the size, shape, cell cycle state, and pathway activity (e.g, contact inhibition) for each simulated cell (**Figure 1a**).

The *cancerInSilico* code is general for all classes of cell-based models, using distinct modeling processes for cellular division and cellular mechanics. By default, the model for cellular mechanics to control location, movement, and division are adapted from CITE DRASDO HOLME. We have added additional features that framework to enable further cancer-specific extensions such as therapeutic effect and multiple cellular types (**Figure 1b, Supplemental File 1**). Nonetheless, users may develop their own code to implement alternative modeling strategies for any of these components of the model (**Figure 1c, CancerInSilico package vignette**).

**Supplemental File 1**: Describe in detail how the model makes decisions, with equations and include comprehensive sensitivity analysis

**Figure 1a.** Overview of model inputs to runCellSimulationFunction and output as pictures of cellular distributions colored according to where cells are in the cycle. **b.** Adaptation from Drasdo Holme of how the model works. **c.** Hierarchy of the model and what can be replaced.

***Parameter sensitivity analysis demonstrates that untreated cells from a single cell type are expectedly most sensitive to presence of a boundary, cellular density, and cell cycle length***

We first perform parameter sensitivity analysis of our default model (**Figure 1b**). We first perform parameter sensitivity on only core parameters for a single cellular type that are independent of the modeled therapeutic effect. This simple simulation enables us to benchmark the performance of the model under conditions in which the population dynamics are readily interpretable and predictable. Sensitivity analysis is described in detail for all parameters in **Supplemental File 1**.

We observe that the model is most sensitive to presence of a boundary, density, and cell cycle length (**Figure 2**). Presence of a boundary models the difference in between whole population growth in a dish and local population growth within a region that enables unconstrained growth. Without a boundary, cells initially grow exponentially and then growth rates decrease below exponential due to the presence of contact inhibition. In contract, the presence of a boundary switches cellular growth to logistic (**Figure 2a**). This growth rate increases with reduced cell cycle length. In the case with the boundary, decreasing growth rates cause cells to hit maximum capacity faster. Total cellular density is capped in simulations with and without a boundary (**Figure 2b**). This cap is lower in cells without a boundary than those that have one. In this scenario, cells are more likely to be expanding rather than dividing to reach an optimal configuring. This expansion is possible only in configurations without a boundary and keeps density low.

Mechanical parameters are adapted from previous publications (CITE DRASDO HOLME). They have been well tuned, and are not recommended to be altered for different simulations. We nonetheless perform additional sensitivity analysis for these parameters (**Supplemental File 1**). Altering these parameters may produce artificial effects in the model that mirror sensitivity to interpretable, cellular parameters. For example, altering the expected distribution of cellular configurations by changing the Monte Carlo probability thresholds can decrease growth in a similar manner to increasing cell cycle length. However, this parameter is a technical parameter in the model coding and not directly measured in experiments. Therefore, it is optimal to alter cellular growth parameters directly and retain well-tuned technical parameters for simulation.

**Figure 2** **a.** Multiple growth rates in boundary and no boundary (show every curve) **b.** Shows cap on density in unbounded case.

***Modeling growth inhibition by targeted therapeutics***

We extend the model to incorporate targeted therapeutics that block oncogenic pathways. We simulate their effect by assuming that they decrease cellular growth rates in cancer cells, which has been reported as the mechanism of action of FDA approved inhibitors such as cetuximab. In practice, these therapeutic effects are encoded increasing the length of cell cycle when the drug is present. The code enables users to input a distribution of impacts on cell cycle lengths in the cellular population. This distribution enables therapeutics to impact the growth rate of each cell differently. Drugs may also be introduced to the population at a specified simulation time, to simulate cancer cellular growth prior to treatment. Because the therapeutic effects only cell cycle length, inducing the drug at the first time point is equivalent to modifying the distribution of cell cycle lengths explored in **Figure 2**).

To test our model, we compare simulated growth rates to *in vitro* growth rates observed in the XXX cell line at varying doses of cetuximab treatment. We observe that increasing the length of the cell cycle as a surrogate for therapeutic concentration accurately models the growth rates that are observed in the *in vitro* experiment (**Figure 3a**). We note that in the experiment cells were synchronized 24 hours prior to treatment. Modifying the input to CancerInSilico to encode this synchronization results to better fits between the model and data at early time points (**Figure 3b**). We observe in both cases that although the therapeutic is effective, it only delays cells reaching maximum growth capacity rather than decreasing that overall capacity consistent with its mechanism of action.

**Figure 3a.** Points which are the real data for experiments, lines which are the model fits at different cell lengths. Cell number vs time at different dosages. W/O synchronization. **b.** With synchronization.

***Simulating time-course gene expression data***

CancerInSilico uses the cellular states in the mathematical model as the basis to simulate gene expression data (**Figure 4a**). Specifically, the model assumes that there are pathways that are activated at each time point corresponding to: (1) the percentage of cells in G to S, (2) the percentage that have successfully completed division, (3) the local density of cells (“contact inhibition”), and (4) expected length of the cell cycle (“activation growth pathway”). The modeled effect of drug changes only the expected length of the cell cycle to reflect a decreased activation of a growth pathway. Indirectly, this change will also impact the state of the cell cycle due to interdependencies of model parameters (**Figures 2** and **3**). The “pathway activity” is modeled as a value between zero and one. The pathway is zero if it is not active in any cell and one if it is active in all cells, with intermediate values depending on the pathway (**Figure 4b**). Therefore, introduction of therapeutics will have a complex impact on all of the simulated pathways (**Figure 4c**). Pathway values are used to simulate gene expression in target genes with a generalization of previously described methods (CITE FERTIG TF PAPER), described in the methods and summarized in **Figure 4a**. In this case, we note that gene expression data simulated with CancerInSilico mirrors the processes observed in real, microarray gene expression data of cells with and without cetuximab treatment for all simulated pathways (**Figure 4d**). The model in CancerInSilico also enables simulation of gene counts from RNA-sequencing data with a negative binomial error model or raw reads with Polyester (CITE).

**Figure 4**. **a.** Cartoon overview of how we simulate gene expression. **b.** Illustrating how the scaling varies between 0 to 1 for each pathway. **c.** Pick one dosage (10 um/mL) from **Figure 3** and plot each of the effective pathways vs time for that selection. Y-axis is % of pathway activity and all pathways can be plotted together. **d.** Comparison to real gene expression data (ELANA).

***Simulating multiple cell types***

Tumors and cancer cell lines are mixtures of distinct cell types. In the case of tumors, this may include diverse sets of cancer cells and cells in the microenvironment. Cell lines are composed of a mixture of cancer cells with heterogeneous genetic backgrounds. In both systems, each cell type has distinct growth properties, therapeutic response, and effect on gene expression. To model this effect, CancerInSilico enables users to encode a set of cell types as input. Each of these cell types may have their own initial population distribution and cellular properties. We perform parameter sensitivity analysis to assess the impact of these parameters in the simplest scenario with two cellular types. In this simulation, we vary both relative cell cycle lengths and initial distributions of cells of each type (**Figure 5a,b**). In absence of a boundary, the faster growing cell type will always dominate the population independent of the initial density (**Figure 5a**). In the case with a boundary, the cells hit carry capacity causing cells to be locked at a fixed final density regardless of the relative expected cell cycle length depending upon the initial density and initial relative density (**Figure 5b, Supplemental Figure 2**).

In addition, inclusion of multiple cell types enables evaluation of the impact of relative sensitivity of therapeutics in these cell types. To explore this, we extend the parameter sensitivity analysis to therapeutic sensitivity. Specifically, we model cases in which therapeutic effect alters the mean of the expected cell cycle length between the cellular types (**Figure 5c**)and cases in which the therapeutic effect varies the standard deviation of the expected cell cycle length between the cell types (**Figure 5d**). In this simulation, we find that cells that with longer expected cell cycle lengths with therapeutics have similar properties to varying the initial relative cell cycle length between the cell types (**Figure 5c**). Even if the means are identical, cell types with increased variation in expected cell cycle lengths with drug treatment also ultimately dominant the population (**Figure 5d**).

**Figure 5a** As Raymon is generating heatmap for 2 cell types w/o boundary. **b.** W/boundary. **c.** Varying mean of drug effect w/o boundary in a simulation where you would typically observe a balance in the cellular populations. **d.** Varying SD of drug effect w/o boundary where you would typically observe a balance in the cellular populations. Both **c** and **d** drug should have an effect which is a normal distribution on the population.

**Supplemental Figure 2** a. Similar figure to what Raymon is generating for 2 cell types w/ boundary & varying initial overall cellular density. **b.** As for Figure **5c** w/ boundary. **c.** As for Figure 5d w/ a boundary.

***Simulation of single-cell RNA-sequencing with CancerInSilico***

CancerInSilico also encodes an option to simulate RNA-sequencing data from multiple cell types. In this case, gene sets are defined for each cell type and “pathway activity” is the proportion of cells in that cell type to simulate gene expression as described above (**Figure 4**). In addition, the software can return single-cell RNA-sequencing data. This simulation models the “pathway” activity and corresponding gene expression changes with a negative binomial error model or Polyester independently for each cell. The model then randomly samples a pre-specified number of cells. In a comparison of single cell relative to bulk RNA-sequencing data for multiple cell types we observed (**Figure 6**).

**Discussion**

* SUMMARY OF WHAT THE PACKAGE DOES AND WHY IT’S USEFUL
* DISCUSS OBSERVATIONS OF PARAMETER SENSITIVITY.
* Discuss how single-cell RNA-sequencing data includes drop out via the negative binomial error model, but not transcriptional bursting or cell size dependent effects (future work).
* NOTE ROOM FOR GROWTH TO INCLUDE MORE COMPLEX MECHANISMS OF CELLULAR RESPONSE, CROSS-TALK BETWEEN CELL TYPES, AND NETWORK MODELS ALL FACILITATED BY THE GENERAL STRUCTURE OF THE PACKAGE AND EASY TO DO.

**Experimental Procedures**

***CancerInSilico Package***

* BRIEFLY DESCRIBE THE STRUCTURE OF CANCER IN SILICO, VERSION USED HERE, ETC. INCLUDE A RMD THAT CAN GENERATE ALL FIGURES FROM THE MANUSCRIPT AS A SUPPLEMENT AND DESCRIBE HERE.

***DRASDO MODEL***

* BRIEFLY DESCRIBE THE MODEL AS IMPLEMENTED, NOTING ALL PARAMETERS, ETC.

***Pathway simulation***

* NOTE THAT WE DECIDE TO SIMULATE G->M, G->S, CONTACT INHIBITION, AND GROWTH FACTOR SIGNALING. DESCRIBE HOW GENES FOR THESE PATHWAYS ARE SELECTED IN THIS CASE AND AVERAGE VALUES. NOTE THAT PACKAGE IS GENERAL, SO CAN SELECT OTHER INPUTS AND/OR DATASETS FROM WHICH TO ESTIMATE THE VALUES.
* DESCRIBE BOTH SINGLE CELL AND POOLED CELL.

***Simulating gene expression data***

Once pathways are simulated, gene expression data is simulated based upon annotated gene sets for each pathway generalizing methods developed previously (CITE FERTIG TF PAPER). Each gene has a pre-specified expression range (Gmin to Gmax), determined either from a reference dataset or set according to a specified distribution. If a gene G is annotated to only one pathway P, then its expression value is given as G = Gmin + P\*(Gmax – Gmin). If a gene is regulated by multiple pathways, its expression is determined by combining all pathways using a user-defined function. By default, gene expression is set to be the maximum expression of that gene across all pathways. Additional genes unaffected by any pathway are also simulated with a pre-specified gene expression value. Technical error is then simulated in all genes. A normal error model is used to simulate microarray data and negative binomial error model is used to simulate RNA-sequencing data. In addition, the expected values can be input to Polyester (CITE) to simulate raw fastq files of reads for time course RNA-seq data.

***Experiments***

* DESCRIBE EXPERIMENTS.

**Author Contributions**

Conceptualization, TDS, CT, LMW, and EJF; Methodology, TDS, RC, MS, YT, CT, LMW, and EJF; Software, TDS, RC, MC, ELF, GSO, and EJF; Formal Analysis, TDS and EJF; Investigation, TDS, YT, SJ, LMW, and EJF; Data Curation, YT, SJ, ELF, LTK, DK, LMW, and EJF; Writing – Original Draft, TDS and EJF; Writing – Reviewing and Editing, TDS, RC, MS, YT, SJ, MC, ELF, LTK, DK, GSO, DAG, CT, and LMW, EJF; Resources, YT, SJ, LMW, EJF; Funding Acquisition – MS, DAG, CT, LMW, and EJF; Visualization, TDS, RC, and EJF. Supervision – DAG, CT, LMW, and EJF; Project Administration, EJF.

**Acknowledgements**

EJF R01, CLEVELAND FOUNDATION, HNSCC SPORE, HOPKINS FOUNDATION AWARD, CANCER CENTER CORE GRANT. OTHERS?