Approximate Bayesian Computing for Parameter Inference in Hybrid Discrete-Continuum Models

Nick Jagiella[1], Dennis Rickert[1], Fabian J. Theis[1,2], Jan Hasenauer[1,2,*]

1 Helmholtz Zentrum München - German Research Center for Environmental Health, Institute of Computational Biology, 85764 Neuherberg, Germany
2 Technische Universität München, Center for Mathematics, Chair of Mathematical Modeling of Biological Systems, 85748 Garching, Germany

*** jan.hasenauer@helmholtz-muenchen.de**

Contents

Abstract

The accurate description of multi-scale biological process requires sophisticated computational models. A variety of tools for the construction and simulations of such models are available. The inference of the unknown parameters of multi-scale models however remains and open problem. Key challenges are stochasticity and computational complexity of most multi-scale models. In this manuscript we present a parallel Approximate Bayesian Computations (pABC) sequential Monte Carlo (SMC) algorithm for the inference of hybrid discrete-continuum models of biological tissue. The propose pABC SMC algorithm is tailored for large computing clusters with a queuing systems, and allows for the study of stochastic processes. In a simulation example, we verify that the parameters of hybrid discrete-continuum models of tumor spheroids can be inferred reliably. Accordingly, we use the pABC SMC algorithm to study tumor spheroid growth in droplets, a model for in vivo tumor spread. Interestingly, we find that 2D and 3D models provide similar parameter estimates. Furthermore, the inference results can be used for experimental planning. These results illustrate the feasibility of data-driven modeling of complex multi-scale processes and the reliability of ABC methods.

Author Summary

To do.

Introduction

Systems and computational biology aims at a mechanistic understanding of biological systems. To achieve this, biological processes on a wide range of time and length scales have to be captured [1]. This established the need for multi-scale models. World-wide interdisciplinary initiatives have been formed to develop multi-scale models and modeling approaches for basic research, diagnosis and therapy [2]. Among the most

well-known projects are the whole-heart [1,3,4] and the whole-cell modeling initiatives [?,5]. Furthermore, software toolboxes such as MCell [6], VCell [7], FLAME [], Deutsch, ... for the implementation and simulation of multi-scale models have been made available. All this resulted in a tremendous increase of the availability and popularity or multi-scale models. A problem which is however largely unsolved is the parameterization of multi-scale models. To enable truly quantitative predictions, the parameters of multi-scale models have to be inferred from experimental data.

For deterministic multi-scale models obtained by coupling ordinary and partial differential equations (O/PDE) promising successes have be achieved. An integrated physiologically based whole-body model of the glucose-insulin-glucagon regulatory system has been developed and parameterized for individual patients to improve the understanding of type 1 diabetes [8]. Similarly, whole-heart models could be used to infer ischemic regions from body surface potential maps to provide early diagnosis of heart infarction [9]. These and other applications demonstrated that the automated parameterization of multi-scale model from experiment data is feasible. This is however mostly limited to coupled ODE and PDE models, which are deterministic and allow for efficient gradient-based optimization. The parameterization of computationally demanding stochastic model is unsolved, as the recent DREAM challenge revealed [].

Biological processes such as gene expression [10,11], signal transaction [12,13], cell division [14] and cell movement [15,16] are intrinsically stochastic. This stochasticity renders stochastic multi-scale [?,17,18] essential. The analysis and parameterization of these stochastic models is more sophisticated than of their deterministic counterparts. Key reasons are that (i) the simulation of stochastic models is often computationally demanding and that (ii) the likelihood function and its gradients cannot be assessed. The simulation of sophisticated agent-based models of liver regeneration [19] and tumor growth [16,20] takes days to months. To assess the expected behavior of models a large number of such stochastic simulations are necessary. Even worst, the evaluation of the likelihood function of the data given the model – the objective function for parameter optimization – requires the integration over all possible trajectories of the systems. This is already for simple models infeasible and researchers restores to approximations [21]. In practice approximations of the likelihood are mostly based on a few realizations of the processes and therefore corrupted by large statistical noise. This statistical noise renders the reliable calculation of mostly infeasible, limiting the use of scalable gradient-based optimization methods [22]. Instead simple manual line search methods are used in practice (see e.g. [?,5]). These methods are known to be inefficient, possess convergence problems and do not provide reliable information about the parameter uncertainty. ToDo: Add something regarding the DREAM challenge.

To infer parameters of stochastic processes, Approximate Bayesian Computation (ABC) algorithms have been developed [23]. These ABC algorithms circumvent the evaluation of the likelihood function by assessing the distance between summary statistics of measured and simulated data. If the distance measure exceeds a threshold the parameter values used to simulate data is rejected otherwise they are accepted. This concept can be used in rejection sampling [23] but as the acceptance rates are generally low, Markov chain Monte Carlo sampling [24] and in sequential Monte Carlo methods [25]. If the summary statistics are informative enough, samples obtained using ABC algorithms converge to the true posterior as the threshold approaches zero [26]. ABC algorithms have been used in a multitude of systems biology applications including gene expression and signal transtuction [27–31]. The inference of multi-scale models has however not been approached. A potential reason is the large number of necessary simulation, limiting the study of computationally intensive models.

In this study we introduce a parallel Approximate Bayesian Computations sequential Monte Carlo (pABC SMC) algorithm. This simple extension of ABC SMC facilitates

the use of multi-core systems and computing clusters, thereby enabling the analysis of computationally demanding stochastic multi-scale models. Convergences of the samples is ensured by rigorous ordering. Using the pABC SMC algorithm we address parameter inference for the widely used class of hybrid discrete-continuum models [**?**, 16, 19, 20] - MORE WOULD BE BETTER -. These computational multi-scale models describes cells as interacting agents with intracellular information processing. The dynamics of extracellular substances such as nutrition or extracellular matrix are captured by diffusion reaction equations, namely PDEs. The individual subsystems are coupled via boundary conditions. This model call is highly complex and analytical methods and no analytical methods are available.

The performance of the method is demonstrated for models of tumor spheroid growth. In separate studies the parameters of the model influencing tumor growth are inferred from artificial and real experiment data. These studies provide a proof-of-principle that the parameter inference for computationally demanding models with complex internal structure is feasible using ABC methods.

Results

ABC implementation for computationally demanding models

We used a parallelized version of the well established ABC SMC method for parameter inference (see Material and Methods section). Fig. 1 illustrates the pipeline used for parallelization. There are many possibilities how to parallelize (multi-core, GPU, cluster, etc.). As we aim for resource-wise and computationally expensive models, we make use of a queue-mediated cluster architecture. Here a master is running the actual ABC SMC routine and is outsourcing the time and resource consuming

viele Parallelisierungsansätze, point out that m consequtive ones are only accepted if all intermingled simulations finished

can we design a simple example in which we see the bias which is introduced if we do not keep track of the order?

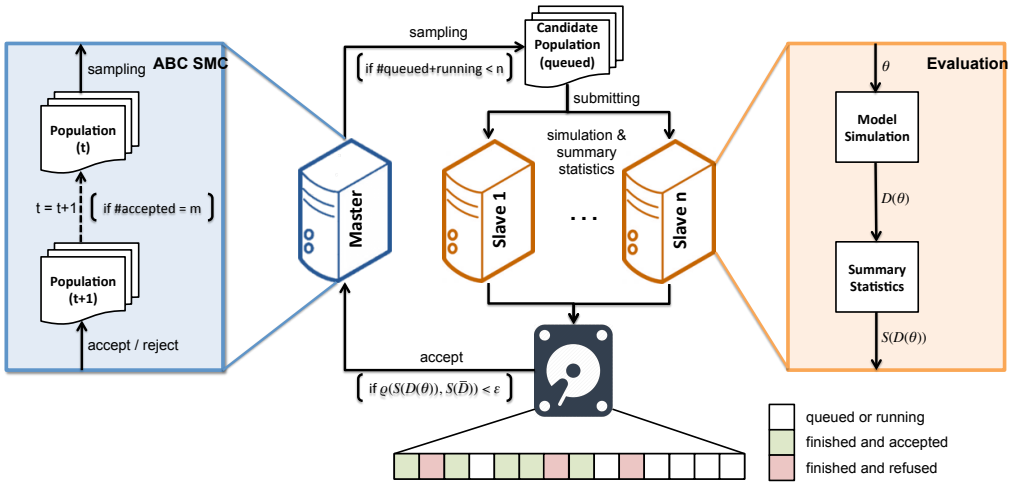


Figure 1. Pipeline using cluster. The ABC method runs on a master machine. Each iteration t new candidate parameters are drawn from a prior distribution and submitted to a queue. Slaves pick up candidates, simulate the model and calculate the summary statistics / distance to data. Then the master accepts those candidate with distances below a threshold ϵ and replaces all finished model evaluations with new samples on the queue.

Artificial data (2D)

Parameter Inference Fig. 2 shows the comparison of data and model prediction and its evolution over the iterations.

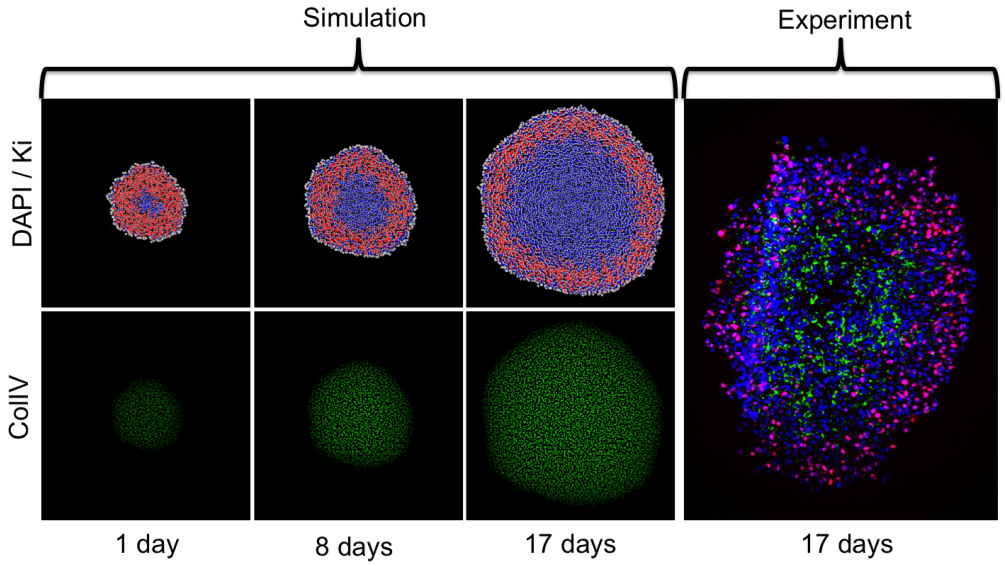


Figure 2. Histological information in in-silicio and in-vitro data.

Fig. 3 shows the comparison of data and model prediction of the parameter populations evolving over iterations.

For the inference of the model parameters used to create the artificial data all parameters can be identified. Nevertheless, the coupling parameter between cellular model and extracellular matrix, e^{crit}, needs a lot of evaluations and remains with a very large uncertainty.

We also can observe that the acceptance rate of candidate samples becomes dramatically low for epsilon $<$ distance at the optimum itself.

Sensitivity to Population/Sample Size Fig. 4 illustrates that the population size has an critical impact on the convergence of the algorithm. If the population size is chosen too small, in our case 20, then convergence can not be assured. On the other hand we observe no significant improvement for an increase from 100 to 200. So for the means of limiting the cluster load per iteration all following parameter inference runs will be done with a population size of 100.

Experimental data (2D)

Fig. 5

no histological info on proliferation (Ki67) leads to wrong predictions cellular kinetic param. kinetic ECM param. not identifiable without hist. info. on ECM (ColV)

Experimental data (3D)

Fig. 6

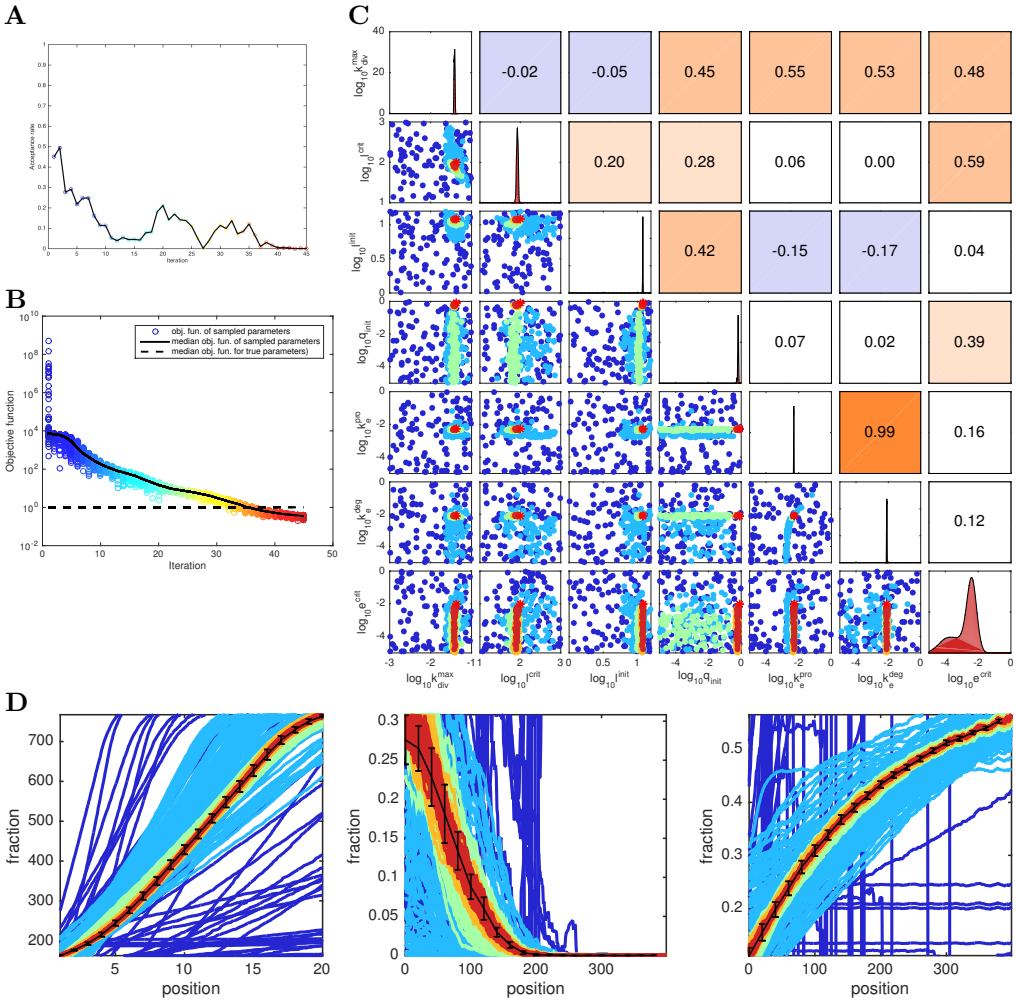


Figure 3. Artificial data and fits for 5 (color-coded) generation (complete dataset). **A** acceptance rate over iteration; **B** ϵ threshold (cyan) and distance of the accepted population (blue) over iteration; for comparison the median distance at the optimum is depicted (red); **C** scatter matrix for 5 (color-coded) generation. todo: color spectrum. **D** Artificial data and fits for 5 (color-coded) generations (complete dataset).

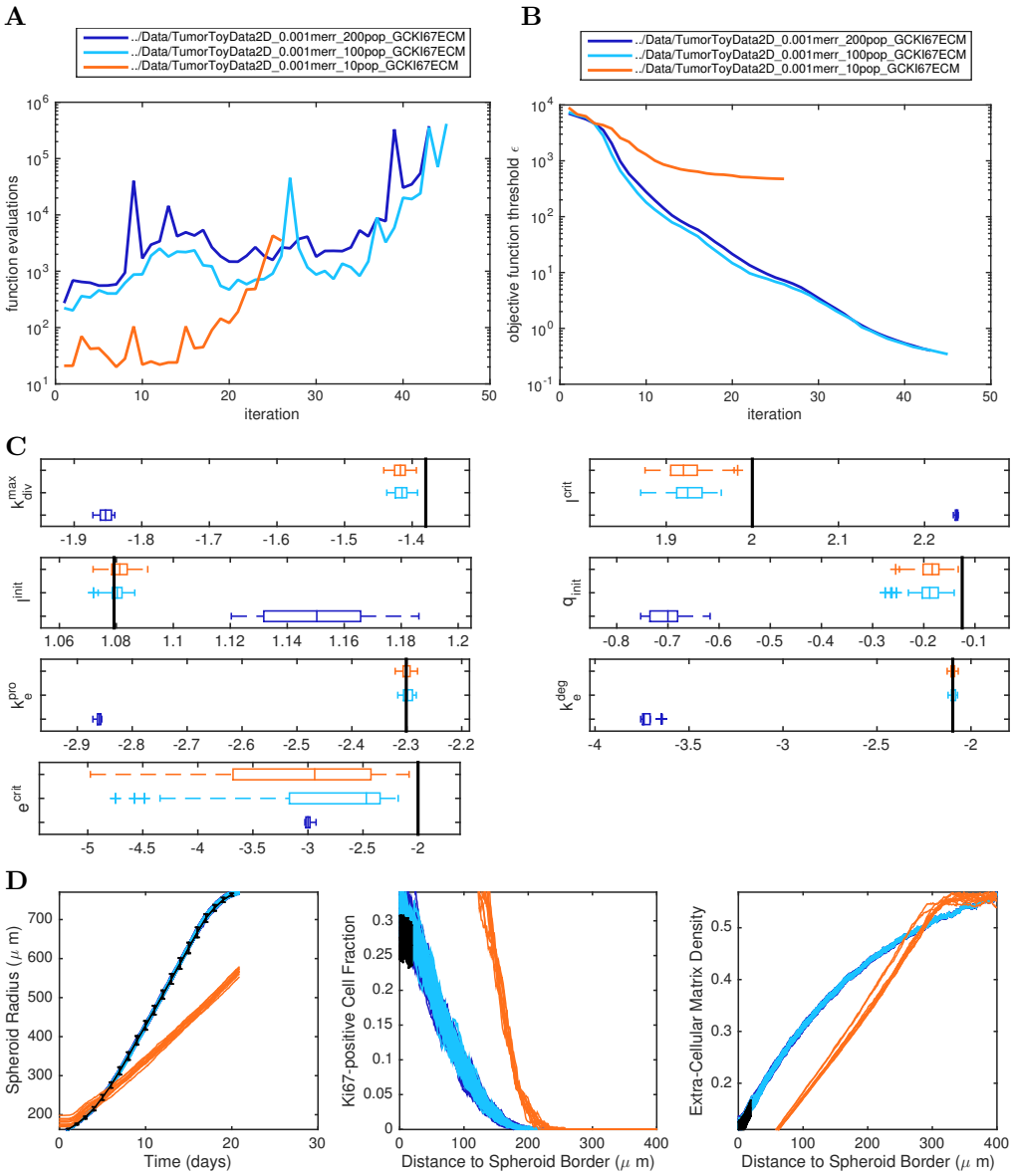
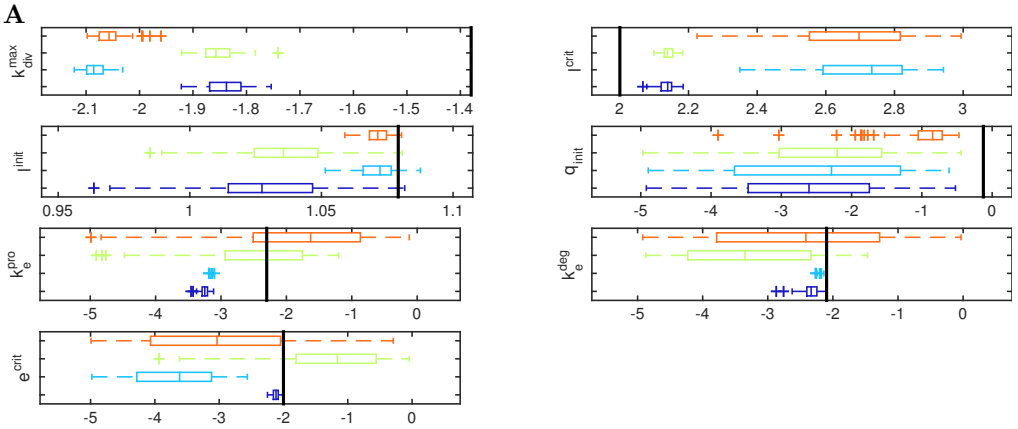


Figure 4. Sensitivity to Population/Sample Size. **A** number of function evaluations over iteration; **B** threshold over iteration; **C** box plot of final sample for different population sizes. **D** final fits.

A



B

		KI67	ECM	KI67ECM			KI67	ECM	KI67ECM
k_{div}^{max}	+	+	++	+	k_{div}^{max}	0.815	0.660	1.000	0.509
l^{crit}	+	++	+	+	l^{crit}	0.102	1.000	0.122	0.779
l^{init}	++	+	+	+	l^{init}	1.000	0.237	0.602	0.176
q_{init}	++	+	+	+	q_{init}	1.000	0.476	0.387	0.468
k_e^{pro}	- -	- -	++	+	k_e^{pro}	0.014	0.019	1.000	0.260
k_e^{deg}	- -	- -	++	+	k_e^{deg}	0.017	0.025	1.000	0.198
e^{crit}	- -	-	-	++	e^{crit}	0.046	0.063	0.081	1.000

C

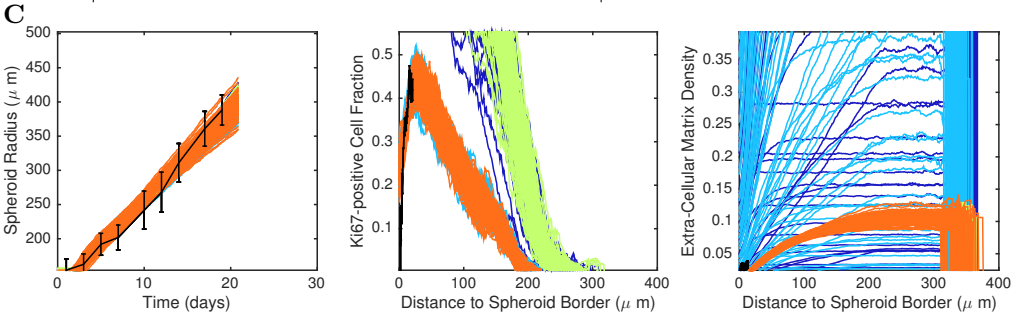


Figure 5. Different combinations of experimental data sets. A box plot for different combinations of data sets; **B** identifiablitity table (+ identifiable, o large uncertainty, - unidentifiable); **C** final fits and scenarios.

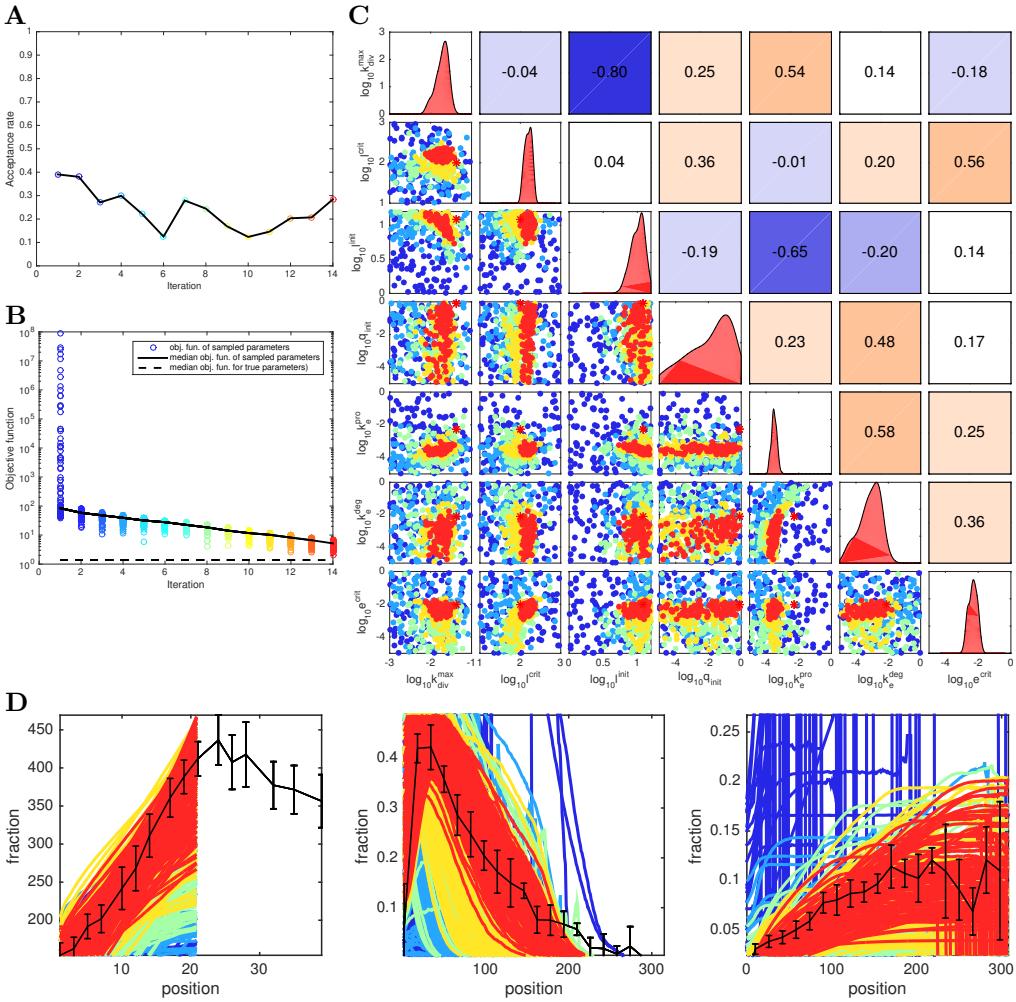


Figure 6. Artificial data and fits for 5 (color-coded) generation (complete dataset). **A** acceptance rate over iteration; **B** ϵ threshold (cyan) and distance of the accepted population (blue) over iteration; for comparison the median distance at the optimum is depicted (red); **C** scatter matrix for 5 (color-coded) generation. todo: color spectrum. **D** Artificial data and fits for 5 (color-coded) generations (complete dataset).

Discussion

spatial moments

Materials and Methods

Model

Tempo-spatial multi-scale model to simulate in-vitro tutor growth. Hybrid-model of individual-based model for tutor cells and a continuum model describing the molecular kinetics of extra-cellular matrix (e) and the key metabolites, glucose (g), oxygen (o) and ATP (a).

Cell model: Cells a) populate a static unstructured lattice (max. 1 cell per lattice site), b) can be either viable or necrotic, and c) can perform a birth and death process if viable or a lysis process if necrotic. The respective transition rates are k_{div}, k_{nec}, k_{lys} and depend on local molecular concentrations

$$k_{div} = k_{div}^{max} H(e - e^{crit}) H(a) H(l - l^{crit}) \quad (1)$$
$$k_{nec} = k_{nec}^{max} H(-a). \quad (2)$$

Molecular model: The molecular dynamics is described by system of partial differential equations as follows

$$\partial_t e = D\nabla e + k_{pro}^e c - k_{deg}^e e \quad (3)$$
$$\partial_t g = D\nabla g - k_{con}^g c \quad (4)$$
$$\partial_t o = D\nabla o - k_{con}^o c \quad (5)$$
$$\partial_t a = k_{pro}^a c - k_{con}^a c, k_{pro}^a = 2k_{con}^g c + 17/3 k_{con}^o. \quad (6)$$

Initial & boundary conditions: The initial cell population occupies all lattice sites within a sphere of radius l^{init}. A fraction of those cells q_{init} is quiescent, while the rest enters the cell cycle.

Parameters Resulting model is stochastic and only numerically solvable. Here we use the Gillespie algorithm and solve the stead state problem for the molecular system after each update. The parameters that are subject to optimisation are indicated in red.

Data

We used two types of data: the growth curves of multi-cellular spheroids over time and histological information on the spatial distribution of proliferating cells and extra-cellular matrix at a certain moment of the experiment.

ABC

The ABC SMC method used in this paper is based on Toni et al. [**?**]. The idea can be briefly summarised as the following:

S1) initialize: $t = 1, p_1(\theta) = \pi(\theta)$, $\epsilon_1 = \infty$

S2.0) set $i = 1$ *#new generation*

S2.1) draw θ_t^i from proposal distribution p_t *#sample*

S2.2) draw $y_t^i = f(\hat{x}, \theta_t^i)$ *#simulate*

S2.3) if $d(y_t^i, \hat{y}) \geq \epsilon_t$ go to S2.1. *#reject*

S2.4) if $i = n$ set $t = t + 1$ and go to S2.0. *#next generation*

Each iteration $t \geq 1$ we sample candidate parameter vectors θ^* from a prior distribution $p^{(t)}(\theta)$ and evaluate the model $y^* = f(x, \theta^*)$. If the corresponding objective function value is $\delta(y^*, y) < \varepsilon^{(t)}$, then θ^* will be accepted and added to $\Theta^{(t)}$. If a minimal number of n accepted parameter vectors is reached, then the algorithm will proceed to the next iteration, $t = t + 1$. The main extension of ABC SMC compared to ABC is to itertively adapt both, the prior distribution $p^{(t)}(\theta)$ and the acceptance threshold $\varepsilon^{(t)}$.

Choice of Prior Distribution $p^{(t)}(\theta)$ Here we chose a Gaussian perturbation kernel:

$$p^{(t)}(\theta) = \sum_i \frac{1}{w_i^{(t)}} (2\pi)^{-k/2} |\Sigma|^{-1/2} e^{-1/2(x-\mu)^T \Sigma^{-1}(x-\mu)}, \quad (7)$$

where μ is the mean and Σ is the standard deviation of the current population t.

Choice of ε-thresholds The threshold distance ϵ for accepting a candidate parameter is chosen to be the median among the actual population

$$\epsilon^{(t)} = \text{median}\{\theta^{(t)}\} \quad (8)$$

(Surrogate Approximation)

Supporting Information

Acknowledgements

This work was supported by the Postdoctoral Fellowship Program (PFP) of the Helmholtz Zentrum München (J.H.).

Author Contributions

Conceived and Designed the Methods: NJ DR FJT JH. Developed Models: NJ. Performed Numerical Experiments: NJ DR. Wrote the Paper: NJ FJT JH.

References

1. Hunter PJ, Borg TK. Integration from proteins to organs: the Physiome Project. Nat Rev Mol Cell Biol. 2003 Mar;4(3):237–243.

2. Physiome Project, http://physiomeproject.org;.

3. Noble D. Modeling the heart – from genes to cells to the whole organ. Science. 2002;295(37):26–33.

4. Trayanova NA. Whole-heart modeling: applications to cardiac electrophysiology and electromechanics. Circulation Research. 2011 Jan;108(1):113–128.

5. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival Jr B, et al. A whole-cell computational model predicts phenotype from genotype. Cell. 2012 July;150(2):389–401.

6. Stiles JR, Bartol TM. Monte Carlo methods for simulating realistic synaptic microphysiology using MCell. In: De Schutter E, editor. Computational Neuroscience: Realistic Modeling for Experimentalists. CRC Press, Boca Raton; 2001. p. 87–127.

7. Schaff J, Fink CC, Slepchenko B, Carson JH, Loew LM. A general computational framework for modeling cellular structure and function. Biophys J. 1997 Sept;73(3):1135–1146.

8. Schaller S, Willmann S, Lippert J, Schaupp L, Pieber TR, Schuppert A, et al. A generic integrated physiologically based whole-body model of the glucose-insulin-glucagon regulatory system. CPT Pharmacometrics Syst Pharmacol. 2013 June;2:e65.

9. Nielsen BF, Lysaker M, Grøttum P. Computing ischemic regions in the heart with the bidomain model – first steps towards validation. IEEE Trans Med Imaging. 2013 June;32(6):1085–1096.

10. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science. 2002 Aug;297(5584):1183–1186.

11. Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. Nature. 2010 Sept;467(9):1–7.

12. Niepel M, Spencer SL, Sorger PK. Non-genetic cell-to-cell variability and the consequences for pharmacology. Curr Opin Biotechnol. 2009 Dec;13(5–6):556–561.

13. Klann MT, Lapin A, Reuss M. Stochastic simulation of signal transduction: Impact of the cellular architecture on diffusion. Biophys J. 2009 June;96(12):5122–5129.

14. Huh D, Paulsson J. Non-genetic heterogeneity from stochastic partitioning at cell division. Nat Gen. 2011 Feb;43(2):95–102.

15. Graner F, Glazier J. Simulation of biological cell sorting using a two-dimensional extended Potts model. Phys Rev Lett. 1992 Sept;69(13):2013–2016.

16. Anderson ARA, Quaranta V. Integrative mathematical oncology. Nat Rev Cancer. 2008 Mar;8(3):227–234.

17. Dada JO, Mendes P. Multi-scale modelling and simulation in systems biology. Integr Biol. 2011;3:86–96.

18. Walpole J, Papin JA, Peirce SM. Multiscale computational models of complex biological systems. Annu Rev Biomed Eng. 2013;15:137–154.

19. Hoehme S, Brulport M, Bauer A, Bedawy E, Schormann W, Gebhardt R, et al. Prediction and validation of cell alignment along microvessels as order principle to restore tissue architecture in liver regeneration. Proc Natl Acad Sci U S A. 2010 June;107(23):10371–10376.

20. Jagiella N. Parameterization of lattice-based tumor models from data [Ph.D. thesis]. Université Pierre et Marie Curie. Paris, France; 2012.

21. Dargatz C. Inference for diffusion processes with applications in life sciences. Springer; 2013.

22. Raue A, Schilling M, Bachmann J, Matteson A, Schelke M, Kaschek D, et al. Lessons learned from quantitative dynamical modeling in systems biology. PLoS ONE. 2013 Sept;8(9):e74335.

23. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian Computation in population genetics. Genetics. 2002 Dec;162(4):2025–2035.

24. Marjoram P, Molitor J, Plagnol V, Tavare S. Markov chain Monte Carlo without likelihoods. Proc Nati Acad Sci U S A. 2003 Dec;100(26):15324–15328.

25. Sisson SA, Fan Y, Tanaka MM. Sequential Monte Carlo without likelihoods. Proc Natl Acad Sci U S A. 2007 Jan;104(6):1760–1765.

26. Marin JM, Pillai NS, Robert CP, Rousseau J. Relevant statistics for Bayesian model choice. JR Statist Soc B. 2014 Nov;76(5):833—859.

27. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J R Soc Interface. 2009 July;6:187–202.

28. Toni T, Jovanovic G, Huvet M, Buck M, Stumpf MPH. From qualitative data to quantitative models: analysis of the phage shock protein stress response in Escherichia coli. BMC Syst Biol. 2011 May;5(69).

29. Lillacci G, Khammash M. The signal within the noise: efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations. Bioinf. 2013 July;29(18):2311–2319.

30. Liepe J, Filippi S, Komorowski M, Stumpf MPH. Maximizing the information content of experiments in systems biology. PLoS Comput Biol. 2013 Jan;9(1):e1002888.

31. Loos C, Marr C, Theis FJ, Hasenauer J. Approximate Bayesian Computation for stochastic single-cell time-lapse data using multivariate test statistics. submitted to 13th Conference on Computational Methods in Systems Biology. 2015;in submission.