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 an 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

Mathematical modeling aims at a mechanistic understanding of biological processes. As biological processes span a wide range of timescales and length scales [1], multi-scale models are necessary to achieve this goal.

Different types of multi-scale models examples for multi-scale models all computationally demanding deterministic, e.g., whole-heart model, whole-body

physiology-based pharmacokinetic/pharmacodynamic (PK/PD) -¿ gradient-based optimization stochastic, e.g., whole cell model, cancer growth models, liver lobule model sensitivity calculation in general not feasible no application of gradient-based methods global optimization uncertainty analysis no possible

spatial moments not available / approximation

likelihood function cannot be evaluated integration over stochastic realizations

simple distance measures Approximate Bayesian Computation (ABC) circumvents evaluation of likelihood functions simple methods in general low acceptance rate sequential Monte Carlo convergences to true posterior

computationally efficiency individual simulations are time consuming parallelization on multi-core system

computational efficiency demonstrated for more simple stochastic processes -¿ INSIGHT, Toni, time-lapse (Carolin)...

use of parallel architecture design ABC which ensure convergence

hybrid discrete-continuum model cells are models by discrete agents nutrition concentration is model using PDE computationally demanding

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

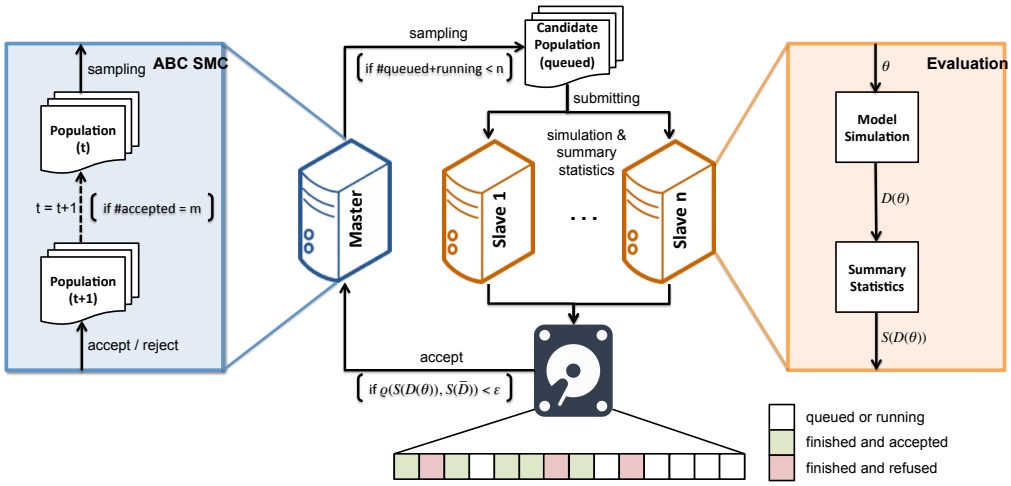


Figure 1. Pipeline using cluster. The ABC method runs on a master machine. Each iteration t new candidate parameters are drawn from a prior distribution and submitted to a queue. Slaves pick up candidates, simulate the model and calculate the summary statistics / distance to data. Then the master accepts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

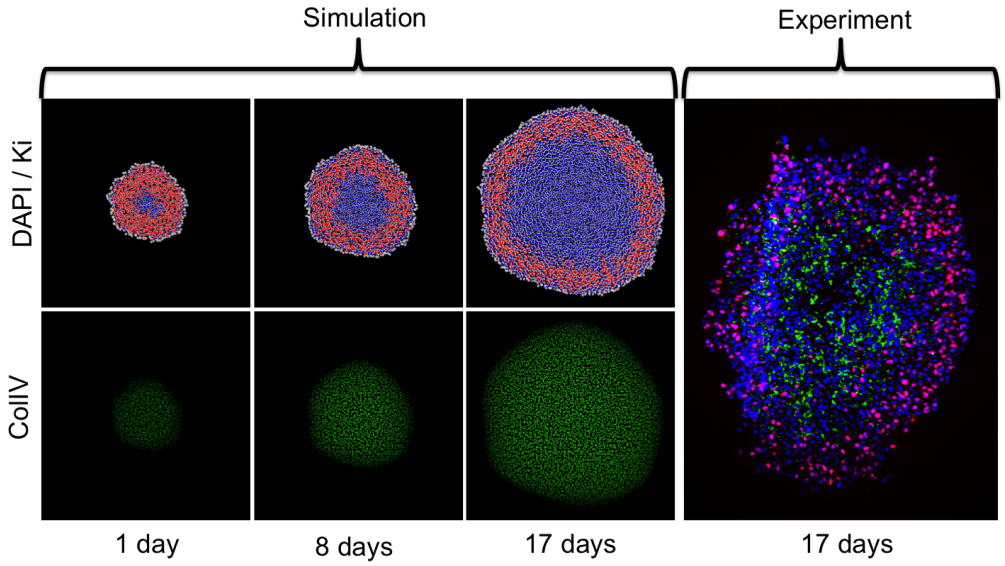


Figure 2. Histological information in in-silicio and in-vitro data.

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

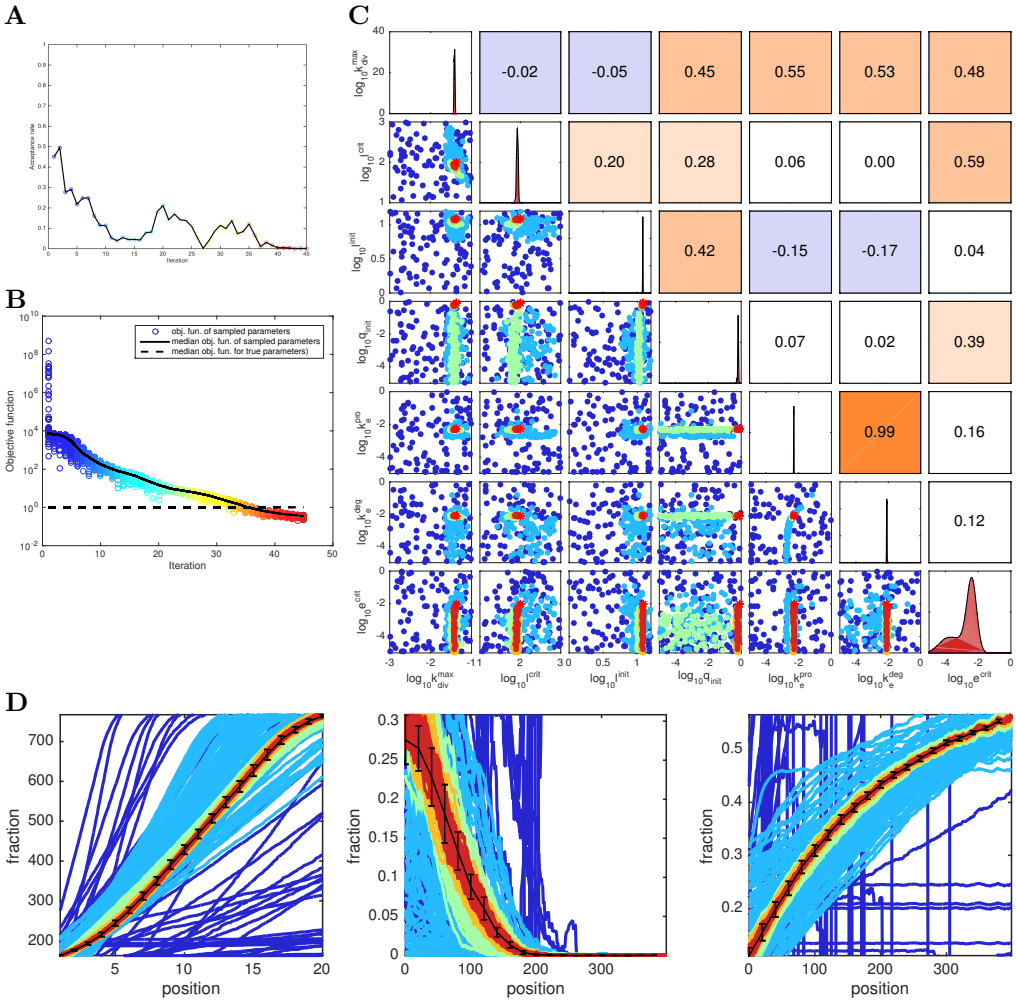


Figure 3. Artificial data and fits for 5 (color-coded) generation (complete dataset). **A** acceptance rate over iteration; **B** ϵ threshold (cyan) and distance of the accepted population (blue) over iteration; for comparison the median distance at the optimum is depicted (red); **C** scatter matrix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

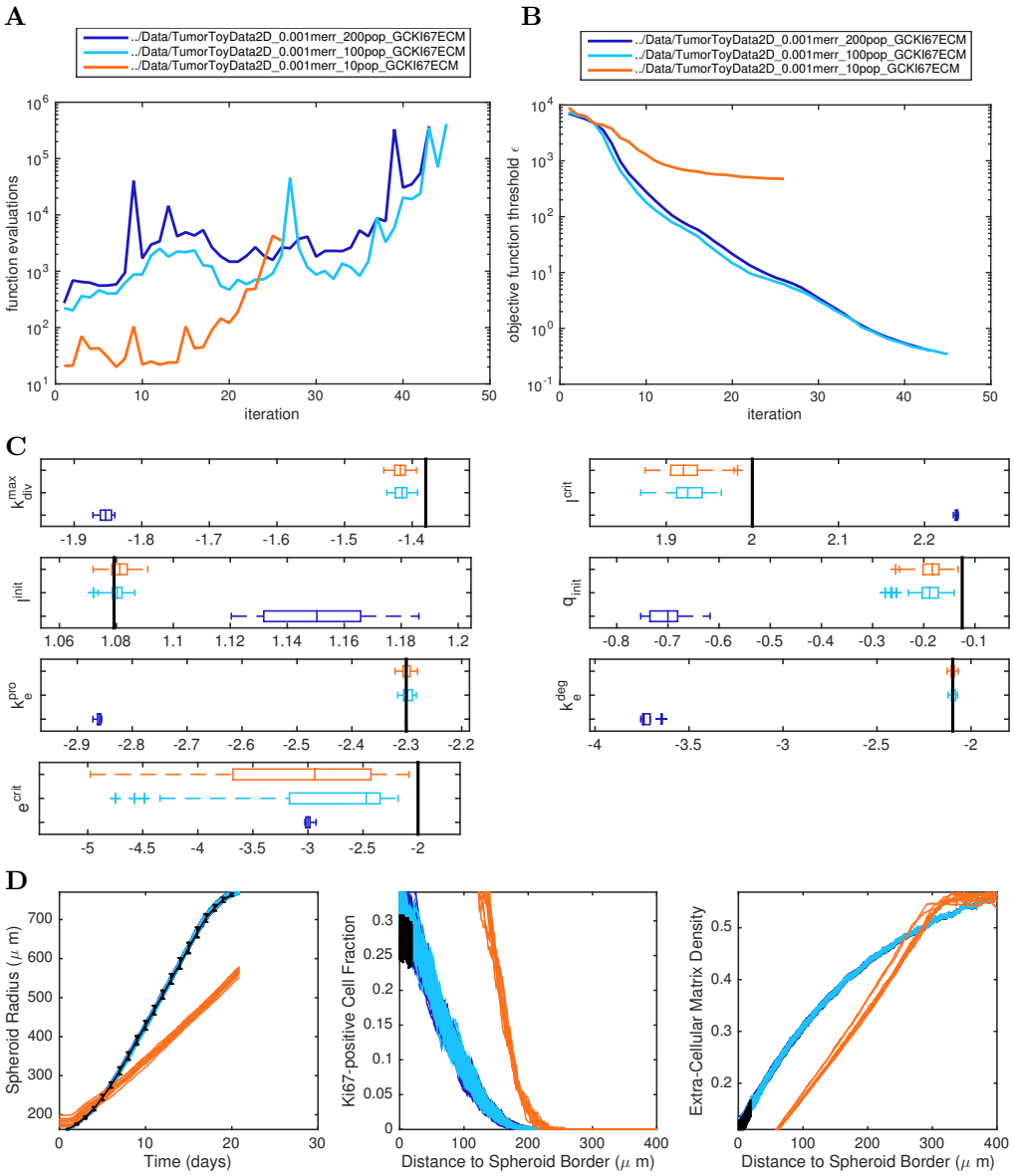
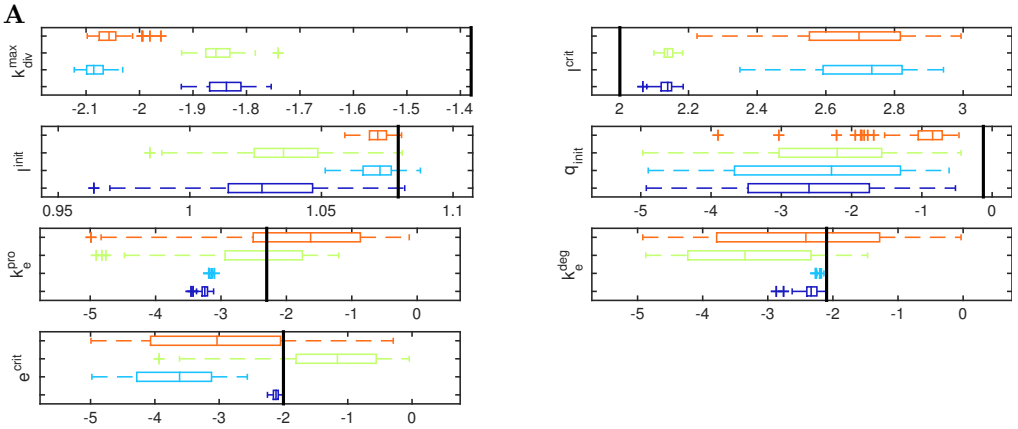


Figure 4. Sensitivity to Population/Sample Size. **A** number of function evaluations over iteration; **B** threshold over iteration; **C** box plot of final sample for different population sizes. **D** final fits.

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

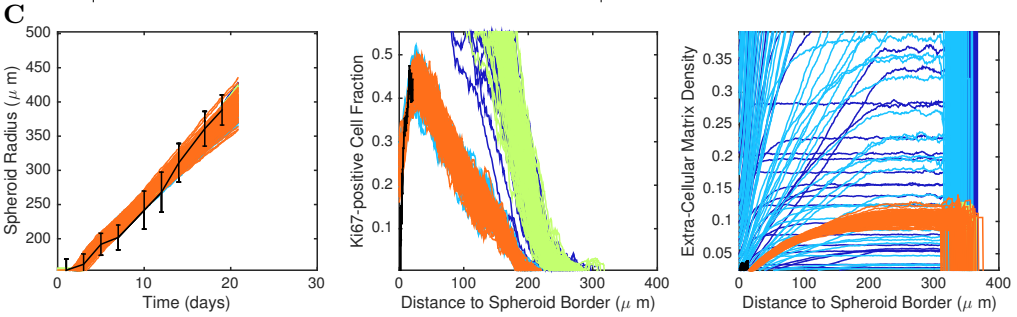


Figure 5. Different combinations of experimental data sets. A box plot for different combinations of data sets; **B** identifiablitity table (+ identifiable, o large uncertainty, - unidentifiable); **C** final fits and scenarios.

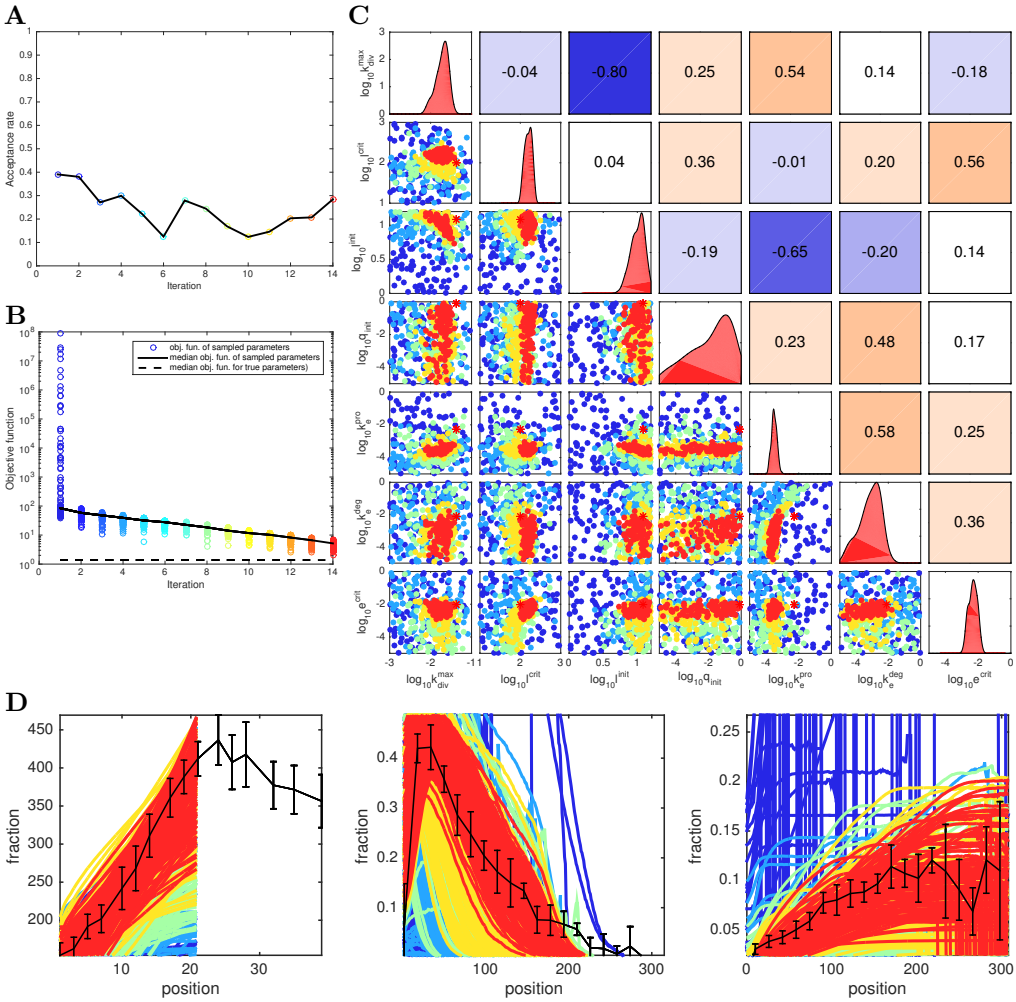


Figure 6. Artificial data and fits for 5 (color-coded) generation (complete dataset). **A** acceptance rate over iteration; **B** ϵ threshold (cyan) and distance of the accepted population (blue) over iteration; for comparison the median distance at the optimum is depicted (red); **C** scatter matrix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

Discussion

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:

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

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

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

(Surrogate Approximation)

Supporting Information

Acknowledgments

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.