Integrating Approximate Bayesian Computation with Complex Agent-Based Models for Cancer Research

In *COMPSTAT 2010 – Proceedings in Computational Statistics*, eds. Saporta G & Lechevallier Y. Springer, Physica Verlag, pp. 57–66, 2010.

Andrea Sottoriva[1] and Simon Tavaré[2]

[1] Department of Oncology, University of Cambridge, CRUK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK, *as949@cam.ac.uk*

[2] Department of Oncology and DAMTP, University of Cambridge, CRUK Cambridge Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK, *st321@cam.ac.uk*

Abstract. Multi-scale agent-based models such as hybrid cellular automata and cellular Potts models are now being used to study mechanisms involved in cancer formation and progression, including cell proliferation, differentiation, migration, invasion and cell signaling. Due to their complexity, statistical inference for such models is a challenge. Here we show how approximate Bayesian computation can be exploited to provide a useful tool for inferring posterior distributions. We illustrate our approach in the context of a cellular Potts model for a human colon crypt, and show how molecular markers can be used to infer aspects of stem cell dynamics in the crypt.

Keywords: ABC, cellular Potts model, colon crypt dynamics, stem cell modeling

1 Introduction

1.1 Agent-based modeling in cancer research

In recent years, cancer research has become a multi-disciplinary field. As well as biological and medical advances, mathematical and computational modeling and advanced statistical techniques have been employed to deal with the ever-increasing amount of data generated by experimental labs.

Recently, the concept of mathematical oncology has taken shape as an emerging field that integrates cancer biology with computational modeling, statistics and data analysis (Anderson and Quaranta (2008)). However, the use of mathematical modeling in cancer research is not completely new; since the 1960s population growth models have been developed to explain the growth kinetics of tumours (cf. Laird (1964), Burton (1966)).

Despite the importance of these models in explaining the basic growth dynamics of solid malignancies, they often fail to represent the intricate underlying mechanisms involved in the disease. Cancer is in fact driven by a large number of complex interactions spanning multiple space and time scales. All these interactions among molecules, such as transcription promoters and repressors, and among cells, such as cell-to-cell signaling, give rise to several emergent behaviors of tumours, most importantly tissue invasion and metastasis (Hanahan and Weinberg (2000)).

In this scenario, multi-scale agent-based models become necessary to study many of the mechanisms involved in cancer formation and progression. In particular, hybrid cellular automata models (Anderson et al. (2006), Sottoriva, Verhoeff et al. (2010)) and cellular Potts models (Jiang et al. (2005), Sottoriva, Vermeulen and Tavaré (2010)) have proved suitable to model cancer cell proliferation, differentiation, migration, invasion and cell signaling. These models represent cancer as an evolutionary process (Merlo et al. (2006)) with emergent behaviour that results from the interplay of several underlying mechanisms at the cellular and extra-cellular level.

1.2 Coupling biological data and models with ABC

Approximate Bayesian Computation (ABC) provides a valuable tool to infer posterior distributions of parameters from biological data when using stochastic models for which likelihoods are infeasible to calculate. Agent-based models are able to incorporate many of the processes occurring in cancer, most of which show non-linear behavior and are therefore impossible to treat analytically. The integration of ABC and agent-based models therefore seems natural, yet there are some important issues to discuss.

In the past ABC has been extensively and successfully employed with population genetics models (Beaumont et al. (2002), Marjoram and Tavaré (2006)). In such applications the models are often relatively simple because they aim to simulate a few crucial underlying processes. In contrast, in some cancer modeling scenarios the models are complex and computationally expensive; even with large computational resources simulating the model millions of times is infeasible. In this paper we discuss how to make use of ABC with complex agent-based models by exploiting parallelization and by reducing the complexity of the ABC algorithms to the minimum. We illustrate our approach using the human colon crypt as an example.

2 Material and methods

2.1 Methylation data

To study the evolutionary dynamics of a human colon crypt we first need to collect data that contain information about the basic processes occurring in

it, such as proliferation, differentiation and migration of cells. Neutral DNA methylation patterns at CpG sites have proved to be suitable candidates as molecular clocks of the cells in the crypt (Yatabe et al. (2001)). By using a population genetics model combined with Markov chain Monte Carlo, Nicolas et al. (2007) showed that it is possible to infer the parameters regulating some of the mechanisms occurring in the human colon crypt.

Nicolas et al. (2007) collected methylation patterns from a total of 57 colon crypts from 7 male patients aged between 40 and 87 years. The dataset is divided into two subgroups: the first contains 8 cells sampled from each of 37 crypts obtained from 5 distinct individuals, the second has 24 cells sampled from each of 20 crypts taken from 3 individuals; one individual is common to both subsets. Each sampled pattern is 9 CpGs long and has been sequenced from a 77 bp locus upstream of the BGN gene on the X chromosome. Because BGN is not expressed in neoplastic or normal colon tissue (Yatabe et al. (2001)) we consider it an epigenetically neutral locus.

2.2 Modeling the colon crypt

Colorectal cancer is one of the most common cancers in humans and it is known to originate from cells in a colon crypt, the units responsible for renewing the colon lining (Barker et al. (2009)). These tubular structures form the colon epithelium and continuously generate new cells that repopulate the fast-renewing colon tissue. At the base of the colon crypt there are stem cells that generate a compartment of transient amplifying cells that in turn give rise to the fully differentiated colon cells. These cells migrate to the top of the crypt and become part the colon epithelial tissue before being shed into the colon lumen (Figure 1). Colorectal cancer is triggered by the disruption of some of the pathways that regulate crypt homeostasis, such as *Wnt* and *APC* (Barker et al. (2009), Reya and Clevers (2005)).

Despite the crucial role played by colon crypts in colorectal carcinogenesis, several mechanisms and parameters of crypt dynamics are unknown, including the number of stem cells present in the crypt, the number of transient amplifying stages and the rate of symmetrical division of stem cells in the crypt (Potten et al. (2009)).

Here we present a newly developed model that simulates cell proliferation, differentiation, migration in the colon crypt. In addition to these processes our model, which we call the *VirtualCrypt*, simulates the occurrence of methylation mutations at each cell division. To model the colon crypt, we unfold the crypt and represent it as a two-dimensional sheet of cells with periodic boundary conditions on the sides and fixed at the bottom (Figure 2A). Cells that exit the top of the lattice are shed into the colon lumen and are therefore deleted from the simulation.

The VirtualCrypt is a Cellular Potts Model that models the colon crypt as a two-dimensional lattice Ω with $N \times M$ sites. Each biological cell in the crypt has a unique identifier or spin σ, and adjacent lattice sites with the

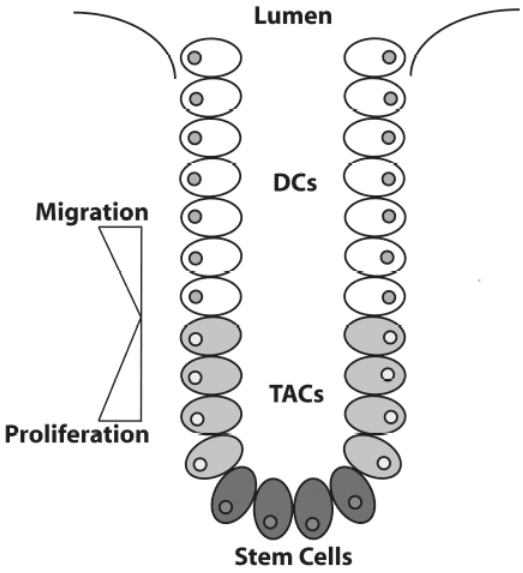

Fig. 1. Cartoon of a colon crypt. DC, differentiated cell; TAC, transient amplifying cell. Stem cells at the bottom of the crypt spawn all the other crypt cells that differentiate and migrate towards the top of the crypt to form the colon lining.

same spin define a single cell volume V_σ and its shape (Glazier and Graner (1993)). Each cell has also a type $\tau(\sigma)$ that identifies a cell as a stem cell, a transient amplifying (TA) cell or a differentiated cell (DC).

The evolution of the system is modeled using a thermodynamical approach borrowed from statistical mechanics in which all the components of the system seek the point of lowest energy. In other words, at each step we propose a large number of random variations to the system and we are more likely to accept those which are more advantageous, in terms of energy, for the cells. For example a cell will seek to expand to maintain its original volume when it is compressed, or it will tend to migrate along a chemotactic gradient if attracted by it.

In summary, we can describe the total energy of the system with a simple Hamiltonian:

$$H = E_v + E_a + E_c, \tag{1}$$

where E_v is the volume elastic energy, E_a is the cell membrane contact energy and E_c is the chemotactic energy. These values represent the energy cost of

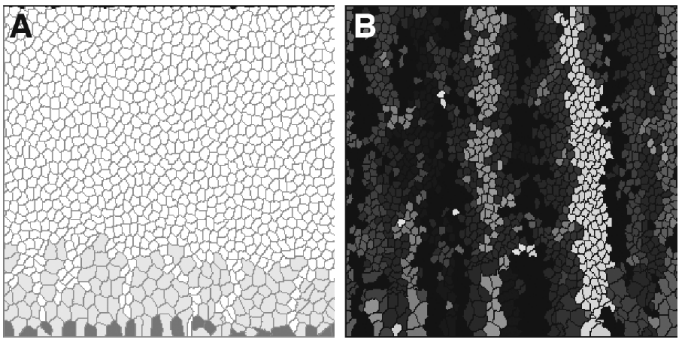

Fig. 2. The cellular Potts model. Panel A: stem cells are dark shade, transient amplifying cells intermediate shade, differentiated cells are white. Panel B: Methylation patterns in the cells in the crypt. Different shades correspond to different alleles in the BGN locus.

a certain cell state. The Volume Elastic Energy E_v is defined by

$$E_v = \sum_{\sigma} \lambda_{\tau(\sigma)} |V_\sigma - V_T|, \tag{2}$$

In the absence of external forces the cell volume V_σ is equal to its target volume V_T and therefore the cell elastic energy $E_v = 0$. When a cell is compressed or stretched its elastic energy increases proportionally to the change in volume and its elastic coefficient $\lambda_{\tau(\sigma)}$, which depends on the cell type. The Cell Adhesion Energy E_a is given by

$$E_a = \sum_{(i,j),(i',j') \text{ neighbours}} J(\tau(\sigma_{i,j}), \tau(\sigma_{i',j'}))(1 - \delta_{\sigma_{i,j}\sigma_{i',j'}}). \tag{3}$$

A certain energy cost or credit $J(\tau_1, \tau_2)$ is associated with each contact point between cells, in a cell-type dependent manner. The δ term in (3) ensures that only contact points between two different cells are considered and not points within the same cell. In this way we can simulate cell adhesion to neighboring cells or to other surfaces in an elegant and straightforward manner. The Chemotactic and Haptotactic Energy is given by

$$E_c = \sum_{(i,j)} \nu_{\tau(\sigma_{i,j})} C(i,j). \tag{4}$$

The chemotactic or haptotactic response of cells to underlying concentration gradients is modeled by assuming that the energy cost of a certain cell state depends on the cell taxis coefficient ν and on the underlying chemical or extracellular matrix concentration $C(i,j)$. The higher the gradient and the migration coefficient, the less it would cost in terms of energy for the cell to migrate rather than stay still.

To evolve the system, at each time step τ we propose and eventually accept a certain number of random local changes in a Monte Carlo fashion, proceeding according to the following Metropolis algorithm (Beichl and Sullivan (2000)):

1. Compute system energy H
2. Pick a random lattice site (i, j)
3. Set the content σ of (i, j) to that of its neighbor (i', j'), chosen at random
4. Calculate the new energy $\Delta H = H_{\text{new}} - H$
5. If $\Delta H < 0$ accept the new state because the total energy is lower
6. If $\Delta H \geq 0$ accept the new state with probability $p = \exp(-\Delta H/(\kappa T))$, where κ is the Boltzmann constant and T is the temperature of the system
7. If the cell is growing, increase the target volume to $V_T = V_T + \delta V$
8. If $V_\sigma > 2V_T$ the cell divides (V_σ automatically tends to V_T, for energetic reasons)
9. Go to 1

In addition to the mechanisms handled by the cellular Potts algorithm, at each cell division we simulate the occurrence of methylation mutations with a rate μ (see Table 1). Each of the 9 CpG sites forming the methylation pattern we collected in the data has a probability μ of being methylated or demethylated at each cell division. If no methylation error occurs, the original methylation pattern is passed on to the daughter cell from the mother cell.

2.3 Inferring colon crypt dynamics with ABC

With our cellular Potts model we are able to simulate the evolution of methylation patterns for long periods of time, up to the age of the patient from which the data have been collected. The two main parameters we are interested in inferring are the number of stem cells N present in the bottom of the crypt and their symmetrical division rate ρ. The rest of the parameters are assumed to be fixed and are reported in Table 1.

Parameter	Symbol	Value
TACs and DCs migration speed	ν	1000 (1 cell position per day)
Cell cycle time	t_c	24h (Potten and Loeffler (1990))
Methylation rate	μ	2×10^{-5} (Yatabe et al. (2001))
Methylation pattern length	γ	9 CpGs (Nicolas et al. (2007))

Table 1. Fixed parameters in the VirtualCrypt simulations.

To fit the two parameters to our methylation data we use Approximate Bayesian Computation. The prior distributions are taken to be uniform, with

$$N \sim U(2, 30), \quad \rho \sim U(0, 1).$$

Initially all cells in the crypts are assumed to be unmethylated (Yatabe et al. (2001)). To compare the multi-dimensional data from the simulations and the patients we define a summary measure $S(\cdot)$ by

$$S(d, p, w, u, g) = \sqrt{d^2 + p^2 + w^2 + u^2 + g^2}, \tag{5}$$

where d is the number of distinct patterns, p number of polymorphic sites, w the average pairwise distance between the patterns, u the number of completely unmethylated patterns and g the number of singletons (patterns that appear only once in a crypt). We note that these statistics are normalized to have common range before use. The ABC algorithm that we applied is as follows:

1. Sample a parameter set θ from the prior
2. Sample a random seed r for the simulation
3. Run the model until the correct patient age is reached
4. Repeat from step 2 until the number of simulated crypts is the same as in the data
5. Compute the summary statistics $X = (d, p, w, u, g)$ of the observed data, with d, p, w, u, g averaged over all crypts
6. Compute the summary statistics $X' = (d', p', w', u', g')$ from the simulated crypts, with d, p, w, u, g averaged over all crypts
7. If $|S(X) - S(X')| < \epsilon$ accept θ as a sample from the posterior distribution
8. Go to 1

This simple ABC approach allows for heavy parallelization due to the independence of the simulations and the accept/reject step that can be performed a posteriori, together with other signal extraction techniques.

3 Results

We generated a total of 80,000 single colon crypt simulations, grouped in sets of 16 having the same parameter set but different random seeds (5,000 different θs in all). This allows us to compare the simulations with the data by reproducing the sampling performed on the patients, where up to 14 crypts were analyzed from a single patient. We are assuming crypts from a single patient have similar parameters. In particular, for each single patient dataset, we calculate the mean summary statistics for a group of simulated crypts with the same parameter set θ. Such a group must be the same size as the number of crypts present in the dataset. We then accept those instances according to the ABC algorithm previously described, with a threshold value of $\epsilon = 0.01$. At the end of this step we obtain posterior distributions of the parameters for each patient. Under the assumption that the physiological parameters of the crypt do not vary among different individuals, we finally average all the posterior distributions to obtain a global posterior distribution of the

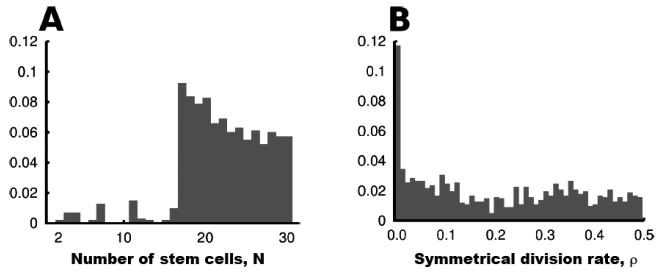

Fig. 3. Posterior histograms of the number of stem cells N and the symmetrical division rate ρ.

parameters given the whole dataset. Such a global posterior distribution is plotted in Figure 3.

Due to the heterogeneity of the methylation patterns present in the crypt, our study suggests a relatively high number of stem cells (Figure 3A). These findings confirm the results previously reported by our group using a population genetics model on the same dataset (Nicolas et al. (2007)).

In contrast to the common assumption of a colon crypt driven by a small number of stem cells, our model suggests a crypt controlled by quite a large number of stem cells that are responsible for generating the observed heterogeneous methylation patterns. Homeostasis in the crypt appears rather more complex than expected, in the sense that it involves a significant number of stem cells, likely between 18 and 25.

Regarding the symmetrical division rate, our study is in agreement with the common assumption that symmetrical division is a relatively rare event, with a probability per cell division of $\rho << 1$. Our results suggest a value smaller than 0.02 (Figure 3B). Hence, crypt homeostasis appears to be driven by a population of stem cells at the bottom of the crypt that most of the time divide asymmetrically, but occasionally undergo symmetrical division, either for self-renewal or differentiation, about once in every 400 cell divisions.

4 Discussion

The role of the colon crypt as an initiator of colorectal carcinogenesis makes it an important and interesting biological system to study. Nonetheless, characterizing the types of cells in the crypt using reliable biomarkers is often a challenging task. Using an in silico approach incorporating modeling and inference, here using methylation patterns as the marker, has proved a good complementary approach to wet lab techniques.

In this study we have shown that it is possible to infer biological features of a structure such as the colon crypt by using agent-based models that reduce the number of approximations and assumptions we need to make to

simulate a biological system. We found that the high level of methylation pattern heterogeneity observed in human colon crypts can be induced only by a relatively high number of stem cells, in agreement with the classical stochastic model of proliferation in the crypt. Furthermore, we confirm the common assumption that stem cells undergo rare (< 0.025 times per cell division) events of symmetrical division that yields either stem cell self-renewal or the differentiation of both mother and daughter cells.

To our knowledge ABC methods have not previously been used for inference in agent-based models. This new framework needs a different approach to modeling and the Bayesian inference itself. Agent-based models are often complex and time consuming to simulate, and this makes them computationally slow even when extensive computational resources are available. In the past ABC has been employed with relatively simple population genetics models that do not contain complex inter-cellular or spatial communication. Such models are fast and easy to simulate, and have led to the development of adaptive ABC algorithms that allow posterior distributions to be obtained more rapidly (cf. Beaumont et al. (2009)).

In a scenario where a single simulation takes tens of minutes instead of mere seconds, the efficiency of the ABC technique is overwhelmed by the bottleneck induced by the model. In our study we found that adaptive ABC methods are not suitable for computationally expensive models, due to the slow convergence caused by the simulation time of the model.

We found that a more convenient approach was first to run all the simulations in parallel with the parameters sampled from the priors. Once a sufficient number of simulations have been computed, any rejection algorithm can be used to analyze the data, from a simple threshold method to more advanced signal extraction techniques.

Another advantage of the simple approach is that it is embarrassingly parallel, and can be directly implemented in a high-performance computing environment by scheduling the simulations independently. Using an adaptive scheme in a computer cluster may require complex job scheduling scripts that would be able to carry information throughout the process of adaptation.

Here we have reported on the feasibility of extending the ABC framework to inference for complex systems described by multi-scale and agent-based models. These provide a powerful tool for investigating biological systems that are otherwise impossible to study using wet lab techniques, especially in humans. Although we have shown that simple yet elegant ABC techniques can be applied successfully in this context, there still a lot of room for improvement and analyses of these methods. For instance, in systems with variable numbers of cells, such as a growing tumour, the scalability of the model may be a crucial bottleneck that could make ABC unfeasible for realistically large numbers of cells.

References

ANDERSON, A. R. and QUARANTA, V. (2008): Integrative mathematical oncology. *Nature Reviews Cancer 8, 227–234.*

ANDERSON, A. R., WEAVER, A. M., CUMMINGS, P. T. and QUARANTA, V. (2006): tumour morphology and phenotypic evolution driven by selective pressure from the microenvironment. *Cell 127, 905–915.*

BARKER, N., RIDGWAY, R. A., van ES, J. H., van de WETERING, M., BEGTHEL, H., van den BORN, M., DANENBERG, E., CLARKE, A. R., SANSOM, O. J. and CLEVERS, H. (2009): Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature 457, 608–611.*

BEAUMONT, M., CORNUET, J.-M., MARIN, J.-M. and ROBERT, C. P. (2009): Adaptive approximate Bayesian computation. *Biometrika 96, 983–990.*

BEAUMONT, M. A., ZHANG, W. and BALDING, D. J. (2002): Approximate Bayesian computation in population genetics. *Genetics 162, 2025–2035.*

BEICHL, I. and SULLIVAN, F. (2000): The Metropolis algorithm. *Computing in Science and Engineering 2, 65–69.*

BURTON, A. C. (1966): Rate of growth of solid tumours as a problem of diffusion. *Growth 30, 157–176.*

GLAZIER, J. A. and GRANER, F. (1993): Simulation of the differential adhesion driven rearrangement of biological cells. *Physical Review E 47, 2128–2154.*

HANAHAN, D. and WEINBERG, R. A. (2000): The hallmarks of cancer. *Cell 100, 57–70.*

JIANG, Y., PJESIVAC-GRBOVIC, J., CANTRELL, C. and FREYER, J. P. (2005): A multiscale model for avascular tumour growth. *Biophysics Journal 89, 3884–3894.*

LAIRD, A. K. Dynamics of tumour growth. (1964): *British Journal of Cancer 13, 490–502.*

MARJORAM, P. and TAVARÉ, S. (2006): Modern computational approaches for analysing molecular-genetic variation data. *Nature Reviews Genetics 7, 759–770.*

MERLO, L. M. F., PEPPER, J. W., REID, B. J. and MALEY, C. C. (2006): Cancer as an evolutionary and ecological process. *Nature Reviews Cancer 6, 924–935.*

NICOLAS, P., KIM, K. M., SHIBATA, D. and TAVARÉ, S. The stem cell population of the human colon crypt: analysis via methylation patterns. *PLoS Computational Biology 3, e28.*

POTTEN, C. S., GANDARA, R., MAHIDA, Y. R., LOEFFLER, M. and WRIGHT, N. A. (2009): The stem cells of small intestinal crypts: where are they? *Cell Proliferation 42, 731–750.*

POTTEN, C. S. and LOEFFLER, M. (1990): Stem cells: attributes, cycles, spirals, and uncertainties. Lessons for and from the crypt. *Development 110, 1001–1020.*

REYA, T. and CLEVERS, H. (2005): Wnt signalling in stem cells and cancer. *Nature 434, 843–850.*

SOTTORIVA, A., VERHOEFF, J. J. C., BOROVSKI, T., McWEENEY, S. K., NAUMOV, L., MEDEMA, J. P., SLOOT, P. M. A. and VERMEULEN, L. (2010): Cancer stem cell tumor model reveals invasive morphology and increased phenotypical heterogeneity. *Cancer Research 70, 46–56.*

SOTTORIVA, A., VERMEULEN, L. and TAVARÉ, S. (2010): Modeling evolutionary dynamics of epigenetic mutations in hierarchically organized tumours. Submitted.

YATABE, Y., TAVARÉ, S. and SHIBATA, D. (2001): Investigating stem cells in human colon by using methylation patterns. Proceedings of the National Academy of Sciences of the United States of America 98, 10839–10844.