Mathematical Statistics
Stockholm University

Stochastic modelling of cell migration

Jens Malmros

Examensarbete 2010:4

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Examensarbete **2010:4**,
http://www.math.su.se/matstat

Stochastic modelling of cell migration

Jens Malmros*

September 2010

Abstract

Cell migration is a central process in normal human cellular development as well as in numerous disease states. Metastatic spread of cancer tumors occurs as a direct result of changes in cell migration, and further insight into the mechanisms behind cell migration is of great importance in cancer research. CMACs (cell-matrix adhesion complexes) are at the heart of the migratory system of the cell; elucidation of CMAC behaviour is essential in understanding cell migration. In this thesis, results from analysis of quantitative live cell microscopy data are used together with modern biological theory to develop a stochastic model describing the behaviour of the CMAC population of the wild-type cell with respect to CMAC areas and the number of CMACs. Analytical results are derived and simulations are performed to validate model performance. It is shown that the model is able to mimic CMAC behaviour with respect to most aspects of the properties described above, and also can predict the behaviour of new perturbed experimental conditions.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: jensm@math.su.se. Supervisor: Olivia Eriksson, Joanna Tyrcha.

Acknowledgements

I would like to thank my supervisors, Dr. Olivia Eriksson, who has supported my work with great enthusiasm and many hours of rewarding discussion, and Professor Joanna Tyrcha, whose help and feedback has been of much assistance.

I give my thanks to Professor Ola Hössjer for valuable suggestions on model analysis and development, and for reading and feedback. Thanks to Professor Rolf Sundberg for help with model analysis.

Thanks to Dr. Tom Andersson for his initial ideas on modelling and feedback and to Mehrdad Jafari Mamaghani for help with understanding and processing data.

Thanks to Dr. John Lock of Karolinska Institutet for his explanations of CMAC behaviour and modelling ideas, and thanks to the other members of Staffan Strömblad's research group for support and for providing data, without which this thesis could not have been written.

Contents

1 Introduction

Cell migration is the process of cellular motion in multicellular organisms. It is vital to normal human cellular development; furthermore, deregulated cell migration plays an important role in many disease conditions. In cancer, changes in cell migration causes metastatic dissemination [7]; thus, elucidation of the mechanisms behind cell migration is highly desirable.
CMACs (cell-matrix adhesion complexes) are central in cell migration. CMACs connect the cell to its surroundings and control several intracellular processes vital to cell migration [7]. Therefore, further insight into cell migration is dependent on the study of CMAC behaviour.

Recent developments in experimental methodology has substantially advanced the study of cell migration. At Staffan Strömblad's laboratory at Karolinska Institutet (KI), Huddinge, advanced methodology utilizing fluorescence microscopy and quantitative image analysis has been developed, providing large amounts of multivariate cell migration (CMAC) data on multiple resolution levels. Analysis of these data requires advanced statistical methods and substantial work on this has been done within the cell stochastics group of Stockholm University.

In this thesis, we will attempt to construct a stochastic model describing CMAC behaviour in the unperturbed cell. Setting out from a top-down perspective, we will build our model from biological theory, results from exploratory data analysis and assumptions about population mechanics. As we aim for model simplicity, we will select key variables with high descriptive qualities, i.e. CMAC area, CMAC life time and the number of CMACs in the cell. We propose a model in discrete time, where an independent multiplicative development is suggested for individual CMAC area, and the number of CMACs will be regulated by a kind of birth and death process, also generating CMAC life times.

We will use analytical results to compare model output with corresponding results from exploratory data analysis. Where no analytical results have been obtained, simulation studies will be performed to validate the model. Simulations are also used to study the capability of the model in an expanded setting.

We conclude that the model is able to mimic CMAC behaviour with respect to most aspects of the properties described above, and also that it is able to resemble the behaviour of a perturbed population.

2 Background

Cell migration is central in the development and maintenance of multicellular organisms. In humans, it is vital to such varying situations as renewal of skin, embryonic development and immune surveillance among others [15]. Cell migration also plays a role in many disease conditions, e.g. cancer, where deregulated cell migration is a key component in metastatic dissemination, the primary cause for mortality in most cancer types [7]. To understand these pathologies, elucidation of the mechanisms behind cell migration is of great importance.

In multicellular organisms cells reside and migrate in a connective tissue called the extracellular matrix (ECM), see Figure 1. The ECM supports and anchors the cells and also influences cellular life in e.g. intercellular communication [1]. The points of contact between cell and ECM are thus very important in cell migration.

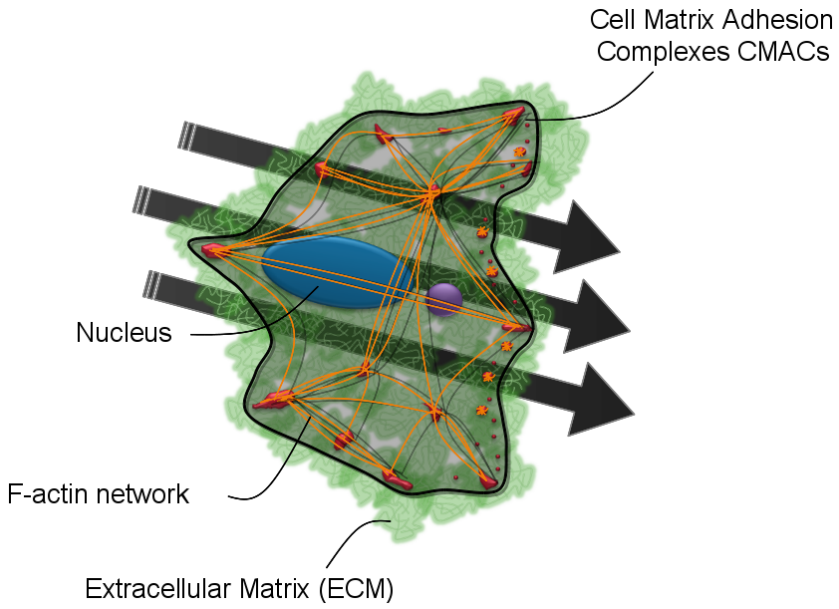

Figure 1: A cell migrating in the ECM. The cell adheres to the ECM through CMACs, which are connected by the microfilament system.

Another vital factor in cell migration is the microfilament (F-actin) system residing within the cell. The microfilament system is a part of the cytoskeleton, a kind of scaffolding, which maintains cell shape, protects the cell and also enables cellular motion.
Located in the cellular membrane, influencing both the ECM and the microfilament system, cell-matrix adhesion complexes (CMACs) are at the heart of the cell migration machinery, see Figure 1; throughout this and the next

paragraph we will use [7] to explain CMAC functionality. CMACs are composed primarily of a type of receptor proteins called integrins and are formed by integrin binding to the ECM and subsequent integrin clustering. Aside from their central role in cell migration, CMACs regulate several vital cell life processes such as cell adhesion, spreading and survival. By physically attaching the cell to the ECM and being linked to the microfilament system in the cell, CMACs simultaneously affect and bind together these key components of cell migration; this gives CMACs a great influence over the migratory process.

Physical and internal properties like size, shape, location and componentry vary significantly between different CMACs. Although CMACs are continuous in their characteristics they can be divided into categories with different properties. For example, small, newly formed CMACs residing in the cellular periphery are called nascent adhesions and these can mature into larger adhesions with properties disparate from those of the nascent adhesions. The various categories of adhesions also link to different parts of the microfilament system, thus providing diverse functionality. Altering the biochemical content and physical appearance of CMACs through environmental cues may enforce selective adaptation, modifying CMAC properties such as size, shape, intensity and composition. Thus, it is possible to create CMAC populations where the proportions of different categories of CMACs may vary, facilitating a systematic approach to elucidating CMAC functionality.

Being a highly complex process in both space and time, cell migration has previously suffered from a lack of thorough investigation due to limitations in experimental methodology. However, recent developments in quantitative live cell microscopy provide the tools needed for a comprehensive analysis of cell migration and of CMAC behaviour. Quantitative live cell microscopy enables the systematic study of cellular and intracellular properties on a large scale, using fluorescence microscopy together with sophisticated imaging software to obtain characteristics of cells and subcellular components.
At the host laboratory of Staffan Strömblad at Karolinska Institutet, Huddinge, quantitative live cell microscopy methodology facilitating ongoing research aiming to elucidate the role of integrins in cellular behaviour, and especially their function in cancer progression, has been developed. From fluorescence microscopy images, unique software extracts quantitative data for 135 key variables incorporating spatiotemporal information from the molecular (CMAC) to the cellular level. This generates huge amounts of data at multiple resolution levels requiring advanced multivariate statistical methods for analysis, which has led to the establishment of a cooperation between Staffan Strömblad's research group and the Division of mathematical statistics, Department of Mathematics at Stockholm University, in 2009. At Stockholm University, the Cell stochastics research group has been formed, including both senior researchers as well as graduate students.

Activity in the Cell stochastics group has mostly been concentrated on exploratory data analysis, characterization of distributions and data mining activities. In this thesis, cell migration is instead approached from a mechanistic point of view, with the aim to provide a stochastic model that attempts to explain CMAC growth in the cell.
Given that our data contain information about the distributions of subcellular properties, stochastic modelling is preferred, since we, as opposed from in deterministic modelling, are able to describe the distributions of the variables of interest. A stochastic model may facilitate the understanding of a cellular system and result in a simple description of it, as it is able to provide the necessary abstraction needed to identify key features influencing the system [10]. Stochastic modelling facilitates the integration of theoretical and experimental results and the reduction of a biological system to its most important parts, and, in the ideal case, gives model predictions that can be used as guidance for setting up new experimental situations [10].
As this is the first try at modelling CMAC behaviour in this way, our ambitions are very moderate, but it is certainly our wish that this thesis will be a starting point for the development of cell migration theory through stochastic modelling.

3 Description of the cell migration data

3.1 Experimental methodology

In the host laboratory, data are recorded through quantitative live cell microscopy. This incorporates automated fluorescence microscopy and quantitative image analysis.
Fluorescence microscopy in biology utilizes fluorophores, molecular components that omit light when absorbing radiation of a different wavelength, to study primarily subcellular processes [5]. The fluorophore is attached to the molecule of interest, often a protein or an enzyme, and the specimen is illuminated with light of a wavelength specific to the fluorophore, which then emits fluorescent light detectable by the microscope [5].
Fluorescent substances are present in nature, but it was the discovery of green fluorescent protein, GFP, in 1962, and subsequent development, together rewarded with the Nobel Prize in Chemistry 2008, that ultimately initiated a whole new era in cell biology. Previously, microscopy was mostly performed on thinned, dead tissue, but, using fluorescence microscopy, researchers are able to investigate cellular processes in living cells, and today, fluorescence microscopy has become an indispensable tool in biological research [19].
Quantitative image analysis is the companion of modern microscopy, using computational techniques enabling automatic identification of the molecular

objects of interest, facilitating effective analysis of vast amounts of digital pictures. The most obvious benefit of quantitative image analysis is the speed compared to human evaluation; another advantage is the ability to generate quantitative measurements, enabling detection of subtle differences between specimens, differences that might be invisible to the human eye [6]. Furthermore, measurements from quantitative image analysis are not biased by human intervention. Quantitative image analysis thus facilitates biological experiments on a much larger scale than before and provides a solid platform for statistical analysis [6].
Together, fluorescence microscopy and quantitative image analysis are the cornerstones of the methodology that enables the production of cell migration data in Staffan Strömblad's laboratory. The quantitative image analysis is focused on CMAC properties and an automated software (Patch Morphology Dynamic (PAD) [4]; patch stands for CMAC) recognizes individual CMACs and tracks them over time while measuring CMAC properties as well as whole-cell properties, see Figure 2. Thus, the progression of individual CMACs can be followed, and by combining individual CMAC properties the population characteristics may be established.

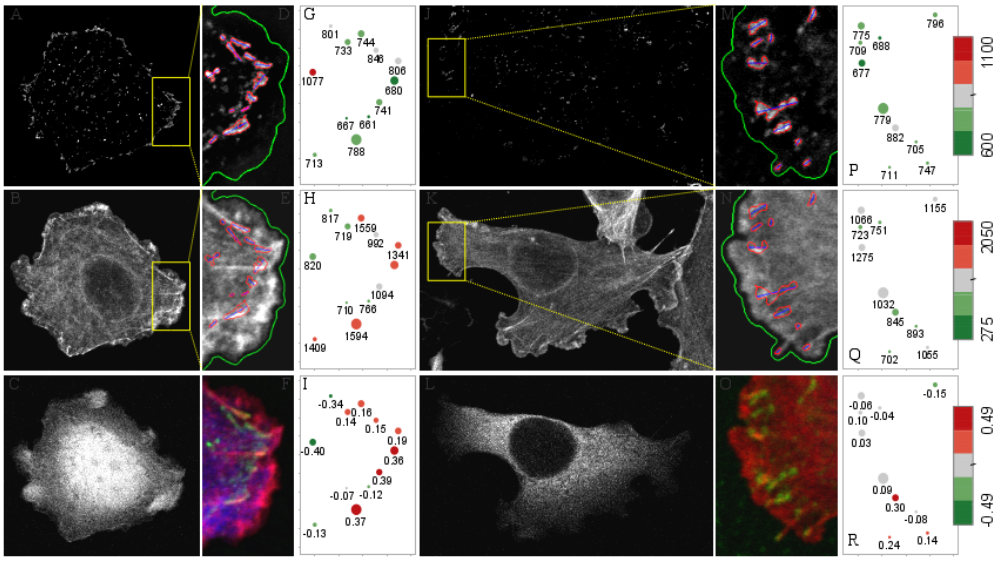

Figure 2: Quantitative image analysis in Staffan Strömblad's laboratory, using pictures from fluorescence microscopy. CMACs are identified by automated software and their characteristics recorded.

The research group at KI uses live human cancer cells in their studies of cellular behaviour. Cancer cells represent the majority of human cell cultures grown in laboratories and cancer progression in humans is the focus of Staffan Strömblad's research group.

Cells in an experimental environment reside in a substance mimicking the physiological extracellular matrix. This substance can have varying rigidity and is known as the cell substrate. In order to investigate the cell migratory system, the cells can be perturbed in different ways, including environmental changes, such as changing substrate density, protein overexpression and the use of siRNA (small interfering RNA) [18]. In this study we mainly focus on the unperturbed system, the wild-type cell, and we will also look at two perturbed conditions, overexpression of the proteins Rac and Rho respectively.
Data come from two different experimental methodologies. The process of creating suitable fluorophores for proteins in live cells is often difficult, with tedious and time-consuming experiments having to be performed for every new type of protein to be investigated. Therefore, alternative techniques using fixed (dead) cells are also used. This means that data produced in the experiments can be either from a time-lapse experiment, with subsequent measurements of live cells, or from a fixed experiment, generating only one group of measurements. The variables recorded in the data can be divided into two categories, static and dynamic, where static variables correspond to instantaneous measurements and dynamic variables are derived from two subsequent instantaneous observations of static variables. For example CMAC area, intensity and distance to border are static variables, whereas CMAC velocity and area growth are dynamic variables. Thus, data from a time-lapse experiment will contain several groups of observations of both static and dynamic variables and data from a fixed experiment will contain one group of observations of static variables. For simplicity we call data originating from time-lapse experiments dynamic and data from fixed experiments static.

3.2 Exploratory data analysis

We will use data from two time-lapse experiments, represented by two dynamic datasets, and from one fixed experiment, represented by one static dataset. The dynamic datasets are produced 090618 and 090527, and the static dataset 090316. The datasets are named e.g. *090618 U2OS stables 20hrs deep red PM combined refch2 vinc*, representing the date and the duration of the experiment, and the fluorophores used.
The dynamic datasets contain observations on unperturbed cancer cells, which are the cells that we are attempting to model, and therefore, our analysis will focus on these datasets. The static dataset contains observations on cells in different perturbation states and is one of the first produced in the laboratory. Therefore, it is subject to more variation due to experimental circumstances which since then have been further developed and standardized, but as we will use this dataset for rough comparison with the

dynamic data this is not critical; it is not itself subject to an analysis. It will also be used for comparison with simulations in Section 5.

The dynamic data are produced in a time interval of, in our case, 20 hours for 090618 and 24 hours for 090527, in which observations are made in blocks with some duration between them, resulting in 120 and 96 indexed times of observation respectively. During one time index, the microscope takes subsequent pictures as it moves around the plate on which the cells reside. Each time index will thus contain observations from one sequence of images recorded as the microscope passes over all the cells and will in fact be a time interval, which has a mean length of 116 and 82 seconds respectively in our dynamic datasets. An excerpt of the 090527 data can be seen in Table 1. Every row in the dynamic data is one observation of one CMAC at one time index, and includes 194 variables. CMACs are given unique ID:s and thus, it is possible to follow changes in properties of individual CMACs over time, since CMACs are always observed in at least two subsequent time indexes. Aside from static and dynamic variables, the observations also contain aggregated variables, derived from the instantaneous variables. The variables have, by the experimentalists, been divided into the following categories:

- **Instantaneous static CMAC properties**
  Static CMAC variables observed in every time point.
- **Instantaneous dynamic CMAC properties**
  Dynamic CMAC variables derived from static variables observed in two subsequent time points.
- **Aggregated properties derived over the whole life time of a CMAC**
- **Instantaneous static cell properties**
  Static cell variables observed in every time point.
- **Aggregated CMAC population and cell properties**

As our modeling will be concerned with only a few selected instantaneous static properties, CMAC area, life time and population size, we will only study a small number of the available variables in the data analysis. In future model development, one might wish to include more variables and ongoing work in the cell stochastics group might be helpful when selecting these variables; this will be further developed in the discussion.

The dynamic datasets contain 91548 (090618) and 20045 (090527) observations, representing a total of 13224 and 3506 unique CMACs (with unique trace ID:s) in 18 and 10 cells respectively. Added during the whole experiment, the cells in the 090618 data contain between 276 and 1393 unique

Idx	Time	ID	Label	Row	Col	Cell no	Area	Maj axis	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
34	3e7	63	1	131.80	273.38	1	0.83	0.72	⋯
34	3e7	20	2	143.04	253.66	1	0.80	1.04	⋯
34	3e7	72	4	147.93	297.54	1	0.55	0.65	⋯
34	3e7	46	5	149.82	287.21	1	0.99	0.68	⋯
34	3e7	102	6	159.38	234.72	1	1.63	1.18	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Table 1: The first 9 variables in a few observations from time index 34 in the 090527 data. The time passed (ms) is in column 2 (3e7=$3 \cdot 10^7$) and CMAC trace ID is in column 3. Label is another, for us unimportant, identification of CMACs. Row and Col are the coordinates of the CMAC in the grid system set up within the image analysis. Cell no represents the number of the cell observed within the present position of the microscope. Individual CMAC properties, here represented by area and Maj axis, the length of the major axis of the observed CMAC, start in column 8 and are followed by other properties as described above. Values have been truncated and times in column 2 rounded.

CMACs each, with a mean of 735 CMACs, and the cells in the 090527 data contain between 259 and 507 CMACs each, with a mean of 351 CMACs. Unfortunately, we are unable to use all observations, due to experimental and physiological issues, such as cell division, which changes the behaviour of CMACs. In plots showing development over time, cell division can often be seen either in the absence of observations for the time indexes concerned or in that CMAC behaviour becomes very unstable compared to non-dividing cells. Plots of variables representing aggregated data will come from undisturbed time indexes only unless otherwise mentioned. All plots can be found in Appendix C.

3.2.1 CMAC population properties

We will start by looking at some aggregated properties describing the number of CMACs in the cells and their area, namely the total CMAC area, the number of CMACs, the average CMAC area multiplied by 10 and the maximum CMAC area multiplied by 10; the multiplication is done to achieve a more readable scale in the plots. In Figure C.1, the development over time for these variables in the 090527 data is shown; note that some of the cells (in the plot denoted by their field names) undergo cell division, additionally, all data after time index 85 should not be taken into account since they are corrupt due to experimental circumstances. Here, and later on, the figures referred to should be thought of as examples showing behaviour that can be

seen throughout data.
Aside from the behaviour of the CMAC population around those time indexes where cell division occurs, we see in most of the plots that the total CMAC area and the number of CMACs remain fairly stable over time with respect to different cells, meaning that they show modest variation around their average values, which are roughly constant over time. This is most evident in fields 3, 16, 17 and 23, the fields corresponding to cells which do not undergo cell division. In all fields, we see that the average CMAC area seems to be constant over time with respect to different cells (remember that in the plots, the value of average CMAC area is multiplied by 10); this means that the total area depends more on the number of CMACs than on the areas of individual CMACs. There is more variation in the maximum size of CMACs over time (also multiplied by 10), both in stable and dividing cells, but this is not surprising since this property is derived from a single extreme value in each time index.
The distribution of the number of CMACs in the cells over time is seen to be fairly normal in our data. Figure C.2 shows normal quantile-quantile plots of observations of the number of CMACs from each time index in fields 3, 16, 17 and 23 in the 090527 data.

When attempting to characterize the development of the CMAC population in one cell over time without considering any individual CMAC properties, it is, in addition to looking at the number of CMACs in the cells, also of interest to study formation and assembly of CMACs.
In Figure C.3, the number of formations (births) and disassemblies (deaths) of CMACs over time (defined as the number of CMAC ID:s that appear and disappear between time indexes) in 8 fields from the 090618 data is plotted together with simple moving averages with a windows size of 16 (the mean of the previous 15 time indexes' values and the present value) for both births and deaths. We see in most of the cells that the number of births and the number of deaths are of almost the same magnitude over time and that they vary around roughly the same average in every time index, since the moving averages of births and deaths tend to follow each other. Aside from a few time indexes in some fields, e.g. the behaviour in field 8 after time index 100 and in field 25, the change in average value is small over time for both births and deaths. Combined with the behaviours seen in the number of CMACs, and with some simple autocorrelation studies on the number of births and deaths in different cells indicating that autocorrelation is present between these, this suggests that the number of CMACs in one cell roughly fluctuates around an average, being regulated by subsequent CMAC formation and disassembly.

3.2.2 CMAC area and life time

Besides the development of the number of CMACs, we would also like to study individual CMAC characteristics in our model, and for reasons further explained in Subsection 4.1, we will focus on the areas of CMACs in the cell. We will also study CMAC life times.
First, we will look at the distribution of CMAC area in the cells over time, since the composition of CMAC area changes in every time index, both from births and deaths in the population, and from individual stochastic variation. To study the change of CMAC area, we will plot its empirical distribution in subsequent equidistant time indexes $0, 10, 20, ...$, looking at fields 5 and 29 from the 090618 data, see Figures C.4 and C.5. Due to limitations in experimental technique, the lowest value of CMAC area that can be read by the microscope is 0.16 μm^2, meaning that we are unable to capture CMAC area development when CMACs are very small. We will also have quite a few values of 0.16 μm^2, since the sampling process actually is discrete, meaning that many small but still measurable values will be rounded down to the lowest value. As our analysis will focus on giving a few general characteristics of CMAC area, this is not critical for us.
In field 5, we have a rather large number of CMACs in all time indexes and it is clear from the plots that the empirical distributions are severely skewed to the right, with much weight being put on the smallest area values. In field 29, there are fewer CMACs, and the skewness is not as prominent, although it is clearly present in those time indexes where there are relatively many CMACs. Although not shown here, strong right-skewness is also present in the empirical distributions of CMAC area in the static unperturbed 090316 data. We can also see in the plots that, while the empirical distributions are right-skewed, there are also, at least in some time indexes, some relatively high CMAC area values present in the empirical distributions, pointing to a heavy-tail character of the empirical distributions.
These characteristics of the empirical distributions will be a starting point when trying to find the mechanisms and distributions that can describe our population. Some common distributions having a right-skewed, heavy-tailed appearance are the power-law and lognormal distributions, which both are used to describe many processes in biological systems [9], and we will look further into this in Subsection 4.1.

The life time of a CMAC is defined as the number of times a CMAC survives from one time index to another. We see in Figure C.6 that the empirical distributions for CMAC life time from 8 fields in the 090618 data are right-skewed, with emphasis on the shortest life times. As it is of interest in our modelling, we have also plotted the estimated geometric distribution for each field in the corresponding histogram, and we see that the geometric distributions does not fully capture the skewness of the empirical distributions,

although it is a fairly good approximation at least to the middle part of the distributions. It is possible that a more heavy-tailed distribution, such as the Pareto, would provide a better fit to CMAC life time, but this will not be covered in this thesis.
As CMACs are measured in at least two subsequent time indexes, data contain no measurements of CMACs younger than 15 minutes. Thus, we can not capture the behaviour of very young CMACs. This is not crucial, since our aim is to make a broad study of population characteristics.

To further study the development of CMAC area, we will look at individual CMAC area progression for some CMACs and the correlation between CMAC life time and average area in some cells. In Figure C.7, CMAC areas over time from field 16 in the 090527 data are shown for CMACs with an average area larger than 1 and a life time longer than 4; aside from reducing the number of CMACs that are considered, these limits are set because we are interested in long-lived CMACs and how they behave when becoming large. It can be seen that the progress of CMAC area tend to form concave trajectories, as CMAC areas often initially increase and then decrease towards the end of the life time of the CMAC in question. There are few CMAC areas becoming significantly larger than 2 for any longer period of time, and many CMACs seem to have an area development that peak around 2, suggesting some kind of threshold for CMAC area.
In Figure C.8, plots of CMAC life time versus average area with linear correlation coefficients are shown for 8 fields from the 090618 data. All the correlation coefficients are significant on the 5% level, but as the data are heteroscedastic this is not so interesting. We see in the plots that short-lived CMACs can have a large average area, and that long-lived CMACs can not have a too small average area; there is an almost linear relationship between the life time of a CMAC and the lower limit of its average area.

3.3 Difficulties with data

There are difficulties with data, both due to experimental limitations and characteristics of the biological system that we work in. We have already touched on some challenges; the multivariate hierarchical structure of data; the large amounts of observations with many variables in every observation; the experimental differences between static and dynamic data, making comparison difficult. Another aggravating circumstance is that we are working in very complex biological systems, the variation of which might be hard to explain or unexpected; this is also important to remember in modelling. In biological systems like these, it is almost impossible to get a full overview of the workings of the system, and one must often be satisfied with only looking at parts of it under certain conditions and then limit conclusions to those conditions.

Considering the experimental limitations in data, the biggest drawback is the time resolution. In e.g. the 090527 data, the time between measurements is around 15 minutes, which allows plenty of things to happen in the cell. A higher resolution would facilitate greater accuracy in our analysis, we would e.g. get a more accurate notion of individual CMAC area variation over time, and CMAC life times could be divided into smaller parts, enabling more detailed study of especially short life times.

4 A model for the CMAC population

4.1 Model background

Starting from a top-down perspective, our intention is to build a simple model that attempts to explain the behaviour of the CMAC population in one wild-type cell. This means that we will set out from a whole-cell context, trying to formulate our model from assumptions about population mechanics rather than from theories concerning basic biological processes in the cell. Thus, we will avoid the increased model complexity often seen when using a low-level mechanistic approach, resulting in a large number of parameters [17].
To further facilitate model simplicity we will prefer basic, well-known components in our modelling and keep the number of variables small. This will give us a solid starting point for future development and model analysis, and also facilitate low model complexity.

Since we initially want to keep the number of variables in the model to a minimum, it is especially important that we select key population variables with high descriptive qualities. A few variables have been put forward by biologists as especially important CMAC characteristics. Among these are area, mean intensity and velocity for individual CMACs and the number of CMACs in the cell. We will choose CMAC area as the main response variable in our model as it has good descriptive qualities, varying between CMACs in different maturation states and different parts of the cell thus facilitating characterization of the underlying population. We have also seen earlier that the empirical distribution of CMAC area has interesting properties, such as right skewness, see Subsection 3.2.2.
As the number of CMACs in the cell is constantly fluctuating due to formation and disassembly of CMACs, affecting the distribution of individual CMAC characteristics such as CMAC area, it is vital to also include the number of CMACs and the lifespans of individual CMACs when attempting to describe the population in one cell.

Multiplicative processes are well-known mechanisms for describing the de-

velopment of organism size in biological systems [3], [9], a few examples in microbiology being protein sequence length and telomere length [13], [12]. By a multiplicative process we mean that if the size of an organism at a discrete timepoint k is X_k, the size at $k+1$ will be $X_{k+1} = X_k \cdot Y_{k+1}$, for some random multiplicative factor Y_{k+1}. Thus, an organism whose size evolves according to a multiplicative process has, in every time step, a random multiplicative change in size which is dependent of its current size.
A multiplicative process results in a lognormal or power law distribution for the quantity whose behaviour is governed by the process. Consider the logarithm of the organism size process described above; the logarithm of size at time k is given by

$$\log X_k = \log X_0 + \sum_{j=1}^{k} \log Y_j,$$

for an organism starting with size X_0 at time 0. Now, if X_0 and the Y_j:s are independent and lognormally distributed, X_k is also lognormally distributed by the definition of the lognormal distribution, see Appendix A.1. By virtue of the central limit theorem, this result also holds for independent and identically distributed random variables Y_j with finite mean and variance. If a lower limit for size had been set, we would instead have had a power law distribution for size, showing how closely related the lognormal and power law distributions are [9].
CMACs are composed of integrins binding to other integrins, giving CMACs with a large area more possibilities to bind to new components and grow even larger. This, and the lognormal/power law behaviour seen in CMAC area data, suggests a multiplicative process as a first choice for modelling CMAC area development. Also, being an ubiquitous phenomenon in e.g. biology, the multiplicative process is a well-known stochastic process, following our intentions for model simplicity. Thus, we suggest a multiplicative process to describe the development of individual CMAC areas over time.

In order for the cell to survive, CMACs must be prevented from growing too large, suggesting a mechanism inducing CMAC area saturation. We have not found experimental evidence for such an area saturation mechanism in the literature, but it has been theoretically predicted [11] that CMAC area should be proportional to the rigidity of the cell substrate. Therefore, it is of interest to include the effect of substrate rigidity in the model and we thus propose an upper threshold for CMAC area in the model over which the areas of CMACs will have zero or negative expected growth.
The introduction of this threshold is also supported by our studies of CMAC area development in data, see Subsection 3.2.2, where we have seen that the majority of the areas of large and long-lived CMACs in a cell seem not to exceed a certain value.

As it is our wish to combine the multiplicative development of individual CMAC areas with the ever-changing number of CMACs in the population due to formation and disassembly, we will introduce a form of birth and death process [16] in the model. The birth and death process is a straightforward approach to modelling CMAC lifespans through the appearing and disappearing of CMACs in the cell, following our intentions for model simplicity.
As we will see later by simulation in Section 5, the introduction of this birth and death process preserves the appearance of the distribution of CMAC area well enough for our purposes.

4.2 Model description

Our aim is to model the behaviour in time of the CMAC population in one wild-type cell, as defined in Subsection 3.1, with respect to the areas of individual CMACs, their lifespans within the cell, and the number of CMACs in the population.
Following the reasoning in Subsection 4.1, we will let CMAC area evolve according to a multiplicative process in time. We assume that the area of a single CMAC existing in the population at any given timepoint will develop following its own multiplicative process, independent of the areas of the other CMACs in the population and of the number of CMACs in the cell at that time.
The growth of the areas of CMACs residing in the cell are restricted by a saturation mechanism that is proportional to the rigidity of the cell substrate. As proposed in Subsection 4.1, this will be illustrated by a threshold, which we from now on will refer to as the substrate threshold, for CMAC area size. When the area of a CMAC grows larger than the substrate threshold, the multiplicative process for the area of this CMAC will have zero or negative expected growth.

CMAC formation and disassembly in the cell will be described by a birth and death process, as suggested in Subsection 4.1. We assume that new CMACs are born into the population independently of the CMACs alive at that time and that the existing CMACs in the population die independently of other CMACs and of their own area.

Thus, we propose a model in discrete time as follows. Let the number of CMACs in the cell at time k, $k = 1, 2, ...$, be $N(k)$, and let the number of CMACs at time 0, $N(0)$, be fixed. Let the $N(0)$ CMACs alive at time 0 be $J(0) = \{1, ..., N(0)\}$, and let the areas of the CMACs alive at time 0 be represented by $\{X_j(0); j \in J(0)\}$, i.e. $\{X_1(0), .., X_{N(0)}(0)\}$, which are sampled from $\text{logN}(\mu_0, \sigma_0^2)$. Let the CMACs alive at subsequent timepoints be $J(1), J(2), ...$, with $|J(k)| = N(k)$, as further described below.
In every time step k the following happens:

- **A new CMAC is born into the population with probability $\mathbf{p_b}$.** Its area at birth will be $X_{j_b}(k)$, where the index $j_b = \max(J(k-1))+1$ is the number of CMACs that has been born into the model so far including the newly born CMAC. For instance, the first CMAC born after time 0 will be indexed $N(0)+1$. The value of $X_{j_b}(k)$ is sampled from $\text{logN}(\mu_b, \sigma_b^2)$. Thus, every CMAC that is born into the population is given a random area and a unique index (the area progression of CMAC j_b is $X_{j_b}(k), X_{j_b}(k+1), X_{j_b}(k+2), \ldots$ for all time points $t \geq k$ when j_b is still alive, i.e. $j_b \in J(t)$).

- **Every CMAC that is alive dies with probability p**. These events are independent of the other CMACS alive and of the corresponding CMAC area. From this we have that CMAC life times will be $Ge(p)$-distributed, counting the number of times that a CMAC survives from one time index to another. Thus, if the CMACs that were killed are given by $D(k)$, we have that $J(k) = (J(k-1) \cup B(k)) \setminus D(k)$, where $B(k) = j_b$, where j_b is as above, if a birth did occur, or $B(k) = \varnothing$ if no birth occurred. This and the above paragraphs constitute our birth and death process for the number of CMACs in the cell.

- **The area of a CMAC that was alive at time $k-1$, and which is not killed in k, will be multiplied by a random factor**. This random factor will be unique for every CMAC and independent of other CMACs and their multiplicative factors, so that, following the description in Subsection 4.1, the random factor multiplied with $X_j(k-1)$, $j \in \{J(k-1) \cap J(k)\}$, the area of CMAC number j at time $k-1$, will be called $Y_j(k)$. The distribution of $Y_j(k)$ will depend on the area of the CMAC,

$$Y_j(k) \sim \begin{cases} \text{logN}(\mu, \sigma^2) & \text{if } X_j(k-1) \leq s, \\ \text{logN}(\nu, \sigma^2) & \text{if } X_j(k-1) > s, \end{cases}$$

  where $\nu \leq 0$ and s is the substrate threshold. Thus we have that

$$X_j(k) = X_j(k-1) \cdot Y_j(k),\ j \in \{J(k-1) \cap J(k)\}.$$

For future purposes we will let the time CMAC j has lived so far at time step k be called $L_j(k)$, $j \in J(k)$. A schematic overview of the model can be seen in Figure 3.

Our model combines the number of CMACs in the population with individual multiplicative processes for CMAC areas and individual births and life times. The number of CMACs in the model is independent of CMAC areas, since CMACs die independently of their area and new CMACs are born into the population with the same probability at every time step.

The model has parameters $p_b, p, \mu_b, \sigma_b^2, \mu, \sigma^2, \nu$ and s. The parameters p_b and p determines the pace of change of our birth and death process, the turnover of the model system. The parameter μ gives the growth of CMAC areas in every time step while ν and s regulates the behaviour at the upper threshold for CMAC area size.
The parameters μ_0 and σ_0^2 of the lognormal distribution from which the areas of CMACs initially are sampled are not further considered, since we are primarily interested in the behaviour of the population at stationarity. That is, all $N(0)$ CMACs alive at time 0 will eventually die out as time tends to infinity.

Since the model is in discrete time, the appearance of the model changes with the scaling of time. Thus, the effects of changes in the turnover or the growth of CMAC areas may be hard to interpret as the time scale is changed, and it might be hard to find a customary way of relating model results to data. As it is not eligible that the time scale should affect the model in this way, a future undertaking will be to take the model into continuous time, which also might reduce the number of parameters in the model. We will present a few results facilitating this in Subsection 4.4.

From the description of the model, we see that it not fully captures the characteristics of some of the CMAC properties that we intend to describe. CMAC lifetimes will, as seen in Subsection 3.2.2, not be entirely comprised by the geometric distribution, and model CMAC area trajectories will not resemble the concave character of real CMAC area trajectories, seen in Figure C.7, to its full extent. Instead, CMAC area development in the model will typically be given by an initial growth, followed by a stage of zero growth, and an instant death; this can be seen in simulations in Section 5.
These discrepancies between model and data are obvious disadvantages, but they are also consequences of our intentions to build a simple model. A more complicated model, which may or may not correct for these discrepancies, will be time-consuming to construct. We believe that it is better to keep these drawbacks in mind and focus on analysis of this model, providing a solid ground for future model development. It is also important to remember that there are many possible extensions and advancements of the model; to build a complex model at this stage might not be fruitful.

4.3 Analytical results

The number of CMACs in our model cell population is independent of the areas of CMACs. This allows us to analyze the birth and death process representing the number of CMACs and the multiplicative processes describing individual CMAC area development separately. We will start by giving some results for the number of CMACs in the model and then continue with results for individual CMAC area.

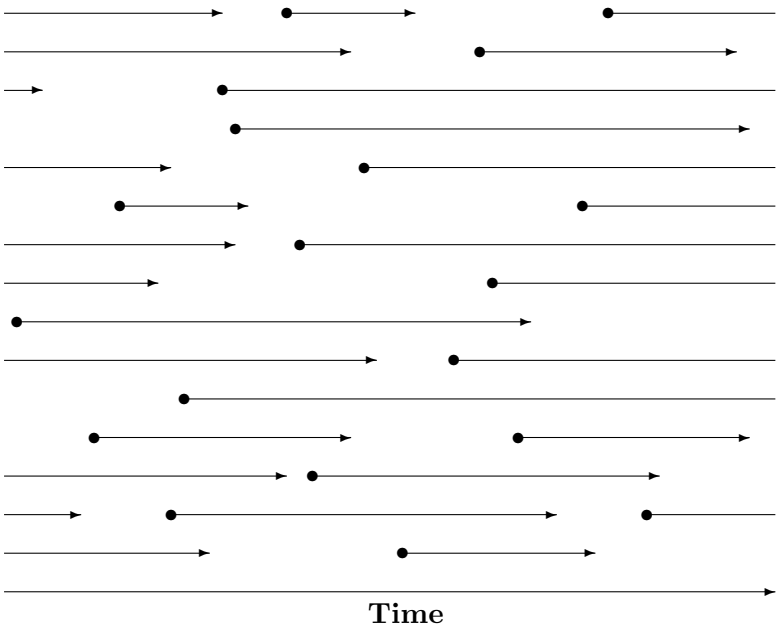

Figure 3: Schematic overview of the model. The circles represent births and the lines the multiplicative processes of CMACs. The arrows represent deaths. The vertical alignment is for aesthetical reasons only.

4.3.1 Number of CMACs in the population

The number of CMACs in the population, $\{N(k), k \geq 0\}$, is a time homogeneous Markov chain. This follows since new CMACs are born into the population with the same probability in every time step and the distribution of the number of dead CMACs in every time step only depends on the current number of CMACs alive. This Markov chain has an infinite state space $0, 1, 2, ...$ and the transition probabilities

$$P_{mn} = P(X_j(k+1) = n | X_j(k) = m)$$

are given by

$$\begin{aligned}
P_{mn} &= 0 \ , n \geq m+2 \\
P_{m,m+1} &= P(N(k) = m+1 | N(k-1) = m) = (1-p)^m p_b \\
P_{mn} &= P(N(k) = n | N(k-1) = m) \\
&= \tbinom{m}{n} p^{m-n} (1-p)^{n-1} \left(1 - p + p_b \left(\tfrac{m+1}{m+1-n} p - 1\right)\right) \quad , n \leq m,
\end{aligned}$$

see Appendix A.2. The transition matrix P is thus a form of triangular matrix and given by

$$P = \begin{bmatrix} P_{00} & P_{01} & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\ P_{10} & P_{11} & P_{12} & 0 & \dots & \dots & \dots & \dots & \dots \\ P_{20} & P_{21} & P_{22} & P_{23} & 0 & \dots & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \dots & \dots & \dots \\ P_{k0} & P_{k1} & P_{k2} & P_{k3} & \dots & P_{k,k+1} & 0 & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \dots \end{bmatrix}.$$

When the Markov chain is in state 0, there are no CMACs in the cell, which in reality is not likely to happen as the cell is unable to exist without CMACs. Therefore, it might seem unreasonable to allow transitions from this state, but it is also in the model unlikely to reach state 0 when the expected value of $N(k)$ is reasonably large (> 10), as can be seen in model simulations, although not performed in this thesis.

There is an apparent asymmetry in the development of $N(k)$, since $P_{m,n} = 0$ if $n \geq m + 2$, and $P_{m,n} > 0$ otherwise, so that, if $N(k) = m$, $m \geq 1$, $N(k)$ may always decrease to states $m - 1, ..., 0$, but never increase further than $m + 1$. This obviously simplifies the model, and keeping in mind that time steps in the model are thought to be very small, with negligible probability for two births or two deaths, it is unproblematic. Another way to consider this is that all CMACs have individual births and deaths. Then we may think of CMAC lifespans as individual, and refrain from considering the whole population.

The expected value of $N(k)$ satisfies

$$E(N(k)) \to \frac{p_b}{p} \text{ as } k \to \infty.$$

If, in addition, $E(N(0)) = p_b/p$, it follows that $E(N(k)) = p_b/p$ holds exactly for all k. See Appendix A.3 for details.
This means that the expected number of CMACs in the model is constant over time, which corresponds to results from data seen in Subsection 3.2.1.

4.3.2 CMAC area development

We will start by considering the multiplicative CMAC area development when $s = \infty$, i.e. when there is no upper threshold for CMAC area size. Then, following the description in Subsection 4.2, the logarithm of the area of CMAC j at time k, given $j \in J(k)$ and $L_j(k) \geq 1$, can be expressed as

$$\log X_j(k) = \log X_j(k - L_j(k)) + \sum_{i=k-L_j(k)+1}^{k} \log Y_j(i)$$

and has a normal distribution with expected value $\mu_b + L_j(k)\mu$ and variance $\sigma_b^2 + L_j(k)\sigma^2$ (the distribution at birth is $N(\mu_b, \sigma_b^2)$).

We are interested in the distribution of the area of a randomly chosen CMAC when stationarity has been reached. A randomly chosen CMAC at time k, where k is large, will have lived for a geometrically distributed time until k, see Appendix A.4. Now, let $F_i = \mathcal{L}(\log X_j(k))$, $L_j(k) = i$, $j \in \{J(k-i) \cap J(k)\}$, be the distribution of the logarithm of the area of a CMAC that has lived for i time steps. Since $L_j(k)$ has a geometric distribution, we have for the distribution F of the logarithm of the area of a CMAC that is alive at k that

$$F = \sum_{i=0}^{\infty} (1-p)^i p F_i.$$

When $s = \infty$ we have $F_i \sim N(\mu_b + i\mu, \sigma_b^2 + i\sigma^2)$, so that F is a geometrically weighted sum of normal distributions. F will, since the number of CMACs in the population is independent of CMAC areas, represent the distribution of the logarithm of the area in the whole population when stationarity has been reached. Figure 4 shows a histogram of simulated values from F transformed back to the original scale. Since F is affected by the lack of a threshold for CMAC area size and also depends on the choice of model parameters, further characterized in Section 5, we show it for comparison only and refrain from further analysis of its properties.

If we let $Q(y; x) = P(\log X_j(k) \leq x | \log X_j(k-1) = y)$, for $j \in \{J(k-1) \cap J(k)\}$, then $Q(y; x)$, viewed as a function of y, will be the distribution function of a normally distributed random variable Z with $Var(Z) = \sigma^2$ and, when $s = \infty$, $E(Z) = y + \mu$, which depends only on $y = \log X_j(k-1)$. Therefore, $\{X_j(l); j \in J(l)\}$ is a Markov chain with continuous state space having transition densities as described by $Q(y; x)$, see [8] for more information on these types of Markov chains. Now, the Chapman-Kolmogorov equations can be used to calculate F_i recursively as

$$F_i(x) = \int Q(y; x) dF_{i-1}(y),$$

for $i \geq 1$, with starting distribution $F_0 \sim N(\mu_b, \sigma_b^2)$. These calculations can be carried out numerically by discretizing the system, i.e. limiting and dividing the state space into (small) parts, and calculating the transition probabilities in the discretized state space by numerical integration; this will not be considered in this thesis, instead we will look at the corresponding distribution for CMAC area (when $s < \infty$) by means of simulations.

The expected value of $\log X_j(k)$ given $\log X_j(k-1) = y$, $j \in \{J(k-1) \cap J(k)\}$, is

$$\mu(y) = \int_{-\infty}^{\infty} x dQ(y; x),$$

which for $s = \infty$ will be a linear function $\mu(y) = y + \mu$. When $s < \infty$, we have that

$$\mu(y) = \begin{cases} y + \mu, & y \leq \log(s) \\ y + \nu, & y > \log(s) \end{cases},$$

as shown in Figure 5 for $\nu = 0$.

If $s < \infty$, $Q(y; x)$ will be the distribution function of a $N(\mu(y), \sigma^2)$-distributed random variable and we will have to use the Chapman-Kolmogorov equations to calculate F_i.

Under certain conditions ($s < \infty$, $\nu < 0$), the Markov chain with transition kernel Q has an asymptotic distribution, corresponding to the logarithm of the area of a randomly picked CMAC that has lived for a very long time. However, due to the geometric life length of CMACs, this asymptotic distribution is not too important for us.

4.4 Future model development

There are several ways to expand and develop the model, motivated by biological as well as statistical considerations, which we will elaborate further in the discussion. It is, however, eligible to take the model into continuous

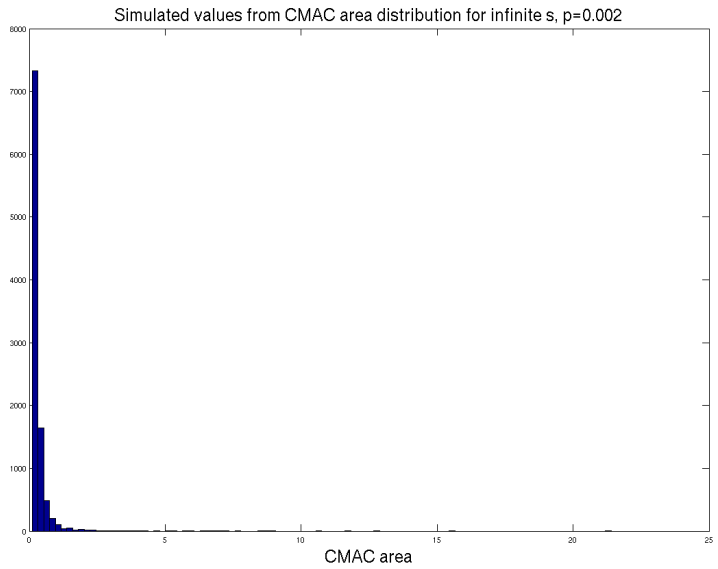

Figure 4: Histogram of 10000 simulated and transformed values from F. Here, $p = 0.002$, $\mu = \log 1.001$, $\sigma = 0.01$, $\mu_b = -1.83$, $\sigma_b = 0.1$ and $s = \infty$; except for s, these parameter values are the same as in simulation i) of Section 5.

time before proceeding with other developments, but the full realization of this is beyond the scope of this thesis. We will instead present a few results on the number of CMACs facilitating this accomplishment; as these are rather technical they are shown in this section.

Given time k and $N(k) = m > 0$, the time to the next event, i.e. birth of a new CMAC or death of a living one, is the minimum of two first success-distributed random variables $W = W_g + 1$ and $V = V_g + 1$, where $W_g \sim Ge(p_b)$ and $V_g \sim Ge(1 - (1 - p)^m)$; V_g represents the minimum of the m geometrically distributed life times of the CMACs alive.
Now, let the time be rescaled with p so that W and V are scaled to pW and pV. Then, if $p_b \to 0$ such that $p_b/p \to \rho$, we have that pW and pV converge to exponentially distributed random variables with intensity ρ and m respectively, see Appendix A.5, and thus that the time to the next event in the process converges to the minimum of these two random variables, which is $Exp(\rho + m)$-distributed.

This means that the system converges to a M/1/M/∞ queue with arrival intensity ρ representing births and departure intensity m representing deaths; the departure intensity depends on the number of CMACs in the cell population. The stationary distribution of the M/1/M/∞ system will be Poisson(ρ) [16]. This convergence is not surprising; choosing memoryless geometric distributions for CMAC life times in discrete time gives us convergence to memoryless exponential distributions in continuous time, and our discrete process for the number of CMACs in the cell population is the analogue of

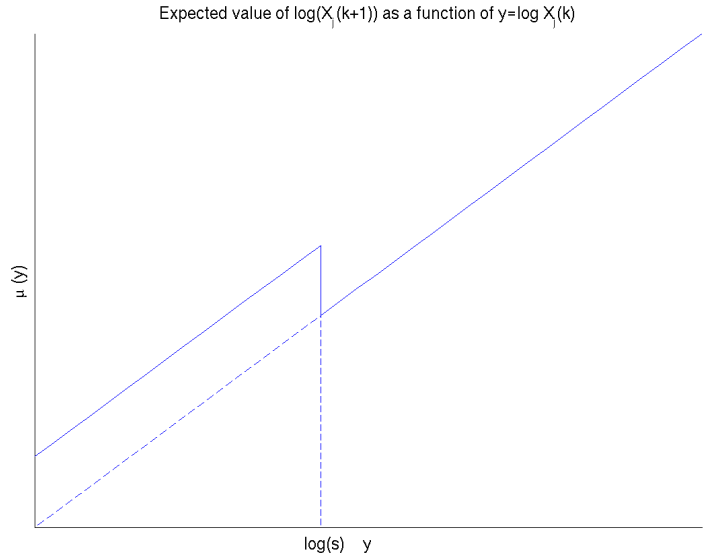

Figure 5: $\mu(y)$, $E(\log X_j(k) | \log X_j(k-1) = y)$, $\nu = 0$.

a M/1/M/∞-process in continuous time.
In fact, the number of CMACs in our discrete-time model also shows a good fit to the Poisson distribution, see Figure C.9, resembling the behaviour seen in data, see Figure C.2; these quantile-quantile plots illustrate the normal approximation to the Poisson distribution. This, together with the results on the expected value of the number of CMACs in the model shown in Subsection 4.3.1, shows that the model depicts the development of the number of CMACs well.

5 Model simulations

Computer based model simulation can be used to validate a model and concretize analytical results, providing visualization of and insight into (long-time) model behaviour. Simulations can also be used to study model performance in situations where no analytical results can be obtained. Here, computer simulations will be used to illustrate model behaviour in three different situations.
i) When the substrate threshold s is set to some finite value, the distribution of CMAC area differs from the corresponding distribution for infinite s, which were derived analytically in Subsection 4.3.2. It is of interest to study this distribution and compare it to the corresponding empirical distributions from data.
ii) It is eligible to estimate the parameters of a model. When a new model is constructed, parameter estimation is often an important step in confirming the ability of the model to mirror empirical results. As meticulous parameter estimation is beyond the scope of this thesis, we will instead derive rough parameter estimates, further described below, to study the relation between model and data.
iii) When the cellular system is perturbed, as described in Subsection 3.1, CMAC behaviour may vary. Two perturbations facilitating different behaviour of the CMAC population are overexpression of the proteins rac and rho respectively. We will examine the ability of the model to qualitatively resemble the behaviour of the CMAC population in these perturbed states.
In i) and iii), the values of the parameters p_b, p and μ will be intuitively chosen or motivated by biological reasons. In all simulations, the parameter ν, regulating expected CMAC area growth as CMAC area becomes larger than s, will be set to 0 for simplicity, and σ_b and σ, the standard deviations of the distribution of CMAC area at birth and of the multiplicative factor respectively, will be set to small values, 0.1 and 0.01, to keep the system robust. The parameter μ_b will be set to -1.83, as this corresponds to the smallest experimentally detectable value ($exp(-1.83) \approx 0.16$) of CMAC area; other values could be used here, e.g. the smallest area at which an integrin cluster can be considered to be a CMAC, but as this is no more than a matter of

scaling we will use -1.83.

i) In this situation, we will study the ability of the model distribution of CMAC area, when $s < \infty$, to resemble the corresponding empirical distribution. We will also explain the setup and the output of the simulations. Figure 6 shows an example of a simulation with $p_b = 0.08$, $p = 0.002$, $s = 5$, $\mu = \log 1.001$ and $\mu_b = -1.83$; the parameter values are arbitrarily chosen to provide a reasonable simulation setup. As we are interested in the behaviour of the process at stationarity, the first 20000 time steps will not be considered; in the plots, time steps 20000 to 40000 are used. The upper plot shows the progress of the simulation, with the coloured lines representing the development of individual CMAC areas, and the lower left plot shows, in line with Figure C.1, the progress of total CMAC area, the number of CMACs, the average CMAC area multiplied by 10 and the maximum CMAC area multiplied by 10. The lower right plot shows the simulated distribution of CMAC area at the end of the simulation.

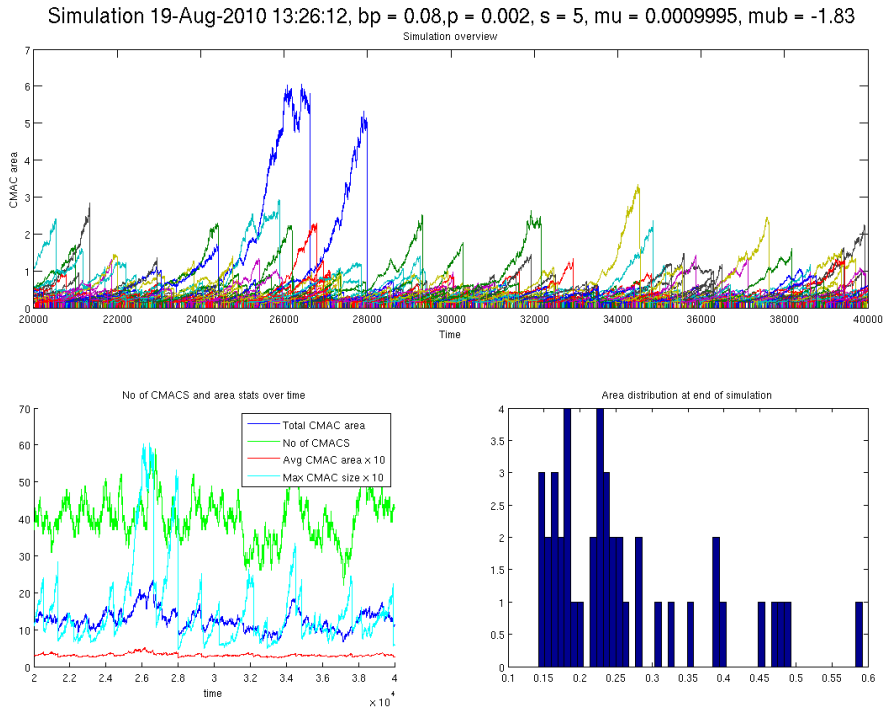

Figure 6: Example of model simulation with $p_b = 0.08$, $p = 0.002$, $s = 5$, $\mu = \log 1.001$ and $\mu_b = -1.83$.

In the lower left plot, we see that the simulation resembles data fairly well. The number of CMACs and the total CMAC area vary around their averages, and the average CMAC area is almost constant. The lower right

plot also shows a good resemblance to the empirical distributions of CMAC area seen in data. In the upper plot, we see model CMAC area progression, which differs from real CMAC area progression, as discussed in Subsection 4.2, although here the trajectories are visually amplified by the scaling. The threshold specified by $s = 5$ is only reached by the areas of two CMACs, but we see that their increase is halted there. Another simulation example which may be used for comparison is shown in Figure C.11.

To further validate our results, we will consider instantaneous empirical distributions of CMAC area from three different simulation setups, plots of which can be seen in Figure C.10. We see that, for all the choices of parameters, the simulated instantaneous empirical distributions of CMAC area correspond well to those seen in data, cf Figure C.5, and therefore conclude that the model is able to provide reasonable distributions for CMAC area when $s < \infty$.

ii) Parameter estimation in the model is difficult, both due to the quality of data and the nature of the parameter setup in the model. Despite this, we will present estimates of the parameters. These estimates are very rough, meaning that they lack analytically derived properties, and the methods used in obtaining them are also questionable. Furthermore, we have chosen our estimators intuitively as it is beyond the scope of this thesis to conduct a wider study of possible estimators. Nevertheless, we want to include these estimates to show that this model have the ability to resemble data, and therefore might capture the relevant biological mechanisms behind CMAC behaviour. The estimation procedure is presented in Appendix B.
Using data from field 3 of the 090527 dataset we get the parameter estimates $p = 0.00092$, $p_b = 0.026$, $s = 2.25$ and $\mu = 1.0008$. An example of a simulation using these parameters can be seen in Figure C.11, in which we see that model behaviour using our rough parameter estimates resembles data.

iii) It is of interest to study the capability of our model to describe CMAC behaviour in situations outside its usual scope. Therefore, we will describe CMAC behaviour in perturbed cells; these are not wild-type (unperturbed) cells, the behaviour of the CMAC population of which we attempt to model. We will qualitatively study the ability of the model to resemble CMAC behaviour in a system that has been perturbed with either overexpression of the protein rac or overexpression of the protein rho, comparing our results with the corresponding data in the static 090316 dataset.
The cellular proteins rac and rho are important in regulating the formation and disassembly, and growth of CMACs [7], having different functionality in these processes. Rac induces formation and growth, shorter life times and lower maximum area, while rho gives fewer, long-lived CMACs, which has increased probability to grow large. Rac and rho exist naturally in the cell, although in varying concentration in different parts.

Figure C.12 shows the empirical distributions of CMAC area in 4 different cells from each perturbation and from the unperturbed system, all from the 090316 data. The two perturbed systems differ in the number of CMACs, with the rac perturbation having 144 to 446 CMACs and the rho perturbation 18 to 40 CMACs. The distribution of CMAC area also differ with much emphasis on the small values in the rac perturbation and larger values in the rho perturbation.
The two perturbations correspond to a varying turnover and an alternating number of CMACs in the model, both of which are related to the parameters p and p_b. Additionally, the distribution of CMAC area will also be regulated by these if μ is unchanged. As this is a qualitative comparison, p, p_b and the other parameters of the model are set according to our biological knowledge and then fine-tuned for illustrative purposes. Figures C.13 and C.14 shows instantaneous distributions from the simulations corresponding to overexpression of rac and rho respectively, and we see a good resemblance to results in data.

All in all, our simulations show that the model is able to produce output that are similar to that which we have seen in data. Our analysis has certainly been very brief, but as the main purpose of our simulations is to illustrate model behaviour and not to make strong conclusions, we are satisfied with this.

6 Discussion

In this thesis, we have made a first try at modelling the development of the CMAC population in one wild-type cell, with respect to the areas of the CMACs in the population, their life times and the number of CMACs in the cell.

Setting out from exploratory data analysis on CMAC populations in real cells, and biological theory, we have deduced a stochastic model for the CMAC population with respect to the properties described above. Analytical results has been derived to validate and increase the understanding of the model, and simulations has been performed to provide visualization and further insight into model behaviour. We have seen that the model is able to mimic the behaviour of the CMAC population in one cell, with respect to most aspects of the properties described above, in different situations.

As noted in Subsection 4.2, the model does not capture the characteristics of CMAC life time and individual CMAC area progression seen in data to their full extent.
The empirical distributions of CMAC life time are of a more heavy-tailed character than the geometric distribution of CMAC life time in the model, which is also less skewed than the empirical distributions. This means that

the model fails to encompass the large amount of short life times and the longest life times seen in data, which also affects e.g. the model distribution of CMAC area. A possible extension of the model correcting for this problem would be to change the distribution of CMAC life time and possibly also the distribution of the time that passes between births; this would mean that the Markov property is lost, making model analysis more difficult. Another addition that might change the distribution of CMAC life time would be to allow different death probabilities of CMACs, dependent of life time or area, corresponding to a biological mechanism with different CMAC phenotypes; this has to be further experimentally studied though.
The progression of CMAC area over time in data, shown in Figure C.7, shows a predominantly concave character, while CMAC area trajectories in the model typically show an initial increase followed by a stage of zero expected growth and instant death. Thus, our model is unable to capture CMAC area development fully. It is known by biologists that the life of CMACs end from different reasons, affecting area progression in various ways. Therefore, implementing more possibilities for CMAC death in the model, inducing different behaviours of CMAC area, might help in correcting this issue. It should be remembered though, in both of these situations, that changing the model may give rise to new discrepancies.
Aside from these dissimilarities, we have seen that the model captures the behaviour of CMAC areas and the number of CMACs in one cell well, using both analytical results and simulations. The distribution of CMAC area in the model shows a good resemblance to data, see e.g. Figure C.10, and the average CMAC area is almost constant over time, see e.g. Figure 6. The number of CMACs in the model shows a good fit to the Poisson distribution, as described in Subsection 4.4, and the average value of the number of CMACs in the model is constant over time. Considering our modest ambitions for modelling and the simplicity of our model, the performance of the model meets our expectations by far.

The properties of the parameter space and estimation of the parameters has only been briefly considered in this thesis. It would be of interest to further study the connections between the parameters and to identify subsets of the parameter space that yields specific model results. Improved parameter estimation would aid an increased understanding of the parameter space and also facilitate more insight into the relationship between model and reality; furthermore, it is very important in developing the predictive qualities of the model. It is important to remember though, that there might be some model improvements, further discussed below, that will facilitate further understanding of the parameters, that might be considered first.

There are many ways to gain an increased analytical understanding of the model and/or develop the model setup; this may also facilitate insight into the properties of the CMAC population. As stated in Subsection 4.2, it

is eligible to take the model into continuous time to eliminate time scale dependency. Additionally, this may also reduce the number of parameters, which will facilitate parameter estimation and understanding of the parameter space. Thus, taking the model into continuous time will also aid other analytical undertakings, such as inferential procedures and sensitivity analysis, for both of which improved parameter estimation also is vital.

Expansion of the model through the inclusion of new explanatory variables, facilitating an increased understanding of the model and the behaviour of the CMAC population, can be done in several ways. When adding new variables to the model, it is important that they have good descriptive qualities of their own, contributing to a deeper insight into CMAC behaviour. Also, a close collaboration with biologists, facilitating increased theoretical insight and further developments in experimental design, as well as widened exploratory analysis, will be of great importance. Adding new variables to the model will increase model complexity and thorough analysis of both data and model is critical in this process.
There are a few variables, other than CMAC area and the number of CMACs, that has been suggested to have good explanatory features, e.g. the velocity and mean intensity of CMACs. The velocity and the life length of CMACs has been put forward by biologists as together being important in explaining CMAC area, and the mean intensity is a, for Staffan Strömblad's research group unique, measure of CMAC concentration that is believed to be of importance.
Another way to expand the model is to add a spatial dimension. This will make it possible to describe the development of the CMAC population in different parts of the cell more closely instead of restricting the model to some part of the cell.

Modelling biological systems is complicated, both from the complexity of the system and, often, from limitations in data. Although our data are impaired with some difficulties, as discussed in Subsection 3.3, their single-molecule resolution in both space and time provides unique information enabling modelling and analysis of CMAC behaviour. That said, there are also areas of improvement, e.g. the time resolution, which is already experimentally enhanced.
How to handle biological complexity raises several issues in modelling. Classically, cellular systems are often modelled by a reductionist, bottom-up approach, where subsystems are added to an increasing model. Cellular systems are, however, complex by nature, being built up from networks of interacting genes and proteins, including interwined positive and negative feed-back loops [2]. There is also the possibility of emergent phenomena, i.e. that the system as a whole is more than the sum of its parts. It has therefore been suggested that reductionist approaches will not suffice in order to describe biological systems [17]. We have therefore used a top-down

approach in trying to find the simplest possible population model that can account for the observed distributions.

References

[1] Alberts, B. et al. (1994) *Molecular Biology of the Cell, 3rd edition*, Garland Science

[2] Bruggeman, F.J. and Westerhoff, H.V. (2007) The nature of systems biology, *Trends in Microbiology, Vol. 15, No. 1*, 45-50

[3] Crow, E.L. and Shimizu, K. (1988) *Lognormal distributions - Theory and applications*, Marcel Dekker inc.

[4] dcilabs Patch Morphology Dynamic (PAD) Assay, *Guide to the image analysis software used in the experiments, provided on request*

[5] Lichtman, J.W. and Conchello, J-A. (2005) Fluorescence microscopy, *Nature methods, Vol.2, No. 12*, 910-919

[6] Ljosa, V. and Carpenter A.E. (2009) Introduction to the Quantitative Analysis of Two-Dimensional Fluorescence Microscopy Images for Cell-Based Screening, *PLoS Computational biology, Vol. 5, Iss. 12*, e1000603

[7] Lock, J.G. et al. (2008) Cell-matrix adhesion complexes: Master control machinery of cell migration, *Seminars in Cancer Biology 18*, 65-76

[8] Medhi, J. (1994) *Stochastic processes*, New Age International

[9] Mitzenmacher, M. (2003) A Brief History of Generative Models for Power Law and Lognormal Distributions, *Internet mathematics, Vol. 1, No. 2*, 226-251

[10] Mogilner, A. et al. (2006) Quantitative Modeling in Cell Biology: What is it good for?, *Developmental Cell 11*, 279-287

[11] Nicolas, A. et al. (2008) Dynamics of Cellular Focal Adhesions on Deformable Substrates: Consequences for Cell Force Microscopy, *Biophysical Journal, Vol. 95*, 527-539

[12] Oexle, K. (1998) Telomere Length Distribution and Southern Blot Analysis, *Journal of Theoretical Biology, Vol. 190, Iss. 4*, 369-377

[13] Ramakumar, S. and Jain, R. (1999) Stochastic dynamics modeling of the protein sequence length distribution in genomes: implications for microbial evolution, *Physica A, Vol. 273, Iss. 3-4*, 476-485

[14] Resnick, Sidney I. (1992) *Adventures in Stochastic Processes*, Birkhäuser

[15] Ridley, A.J. et al. (2003) Cell Migration: Integrating signals from Front to Back, *Science 302*, 1704-1709

[16] Ross, S. (2007) *Introduction to Probability Models*, Academic Press

[17] Szallasi, Z. et al. (2006) *System Modeling in Cellular Biology - From Concepts to Nuts and Bolts*, MIT Press

[18] Winograd-Katz, S et al. (2009) Multiparametric analysis of focal adhesion formation by RNAi-mediated gene knockdown, *Journal of Cell Biology, Vol. 186, No. 3*, 423-436

[19] Yuste, R. (2005) Fluorescence microscopy today, *Nature methods, Vol.2, No. 12*, 902-904

A Analytical results

A.1 Notes on the lognormal distribution

The lognormal distribution is closely related to the normal distribution; the definition of a lognormal random variable X is that $\log X$ is normally distributed [3]. The lognormal distribution is often parameterized by the parameters for the corresponding normal distribution, μ and σ^2, and correspondingly, the notation used within this thesis is $\text{logN}(\mu, \sigma^2)$. The lognormal distribution is positive-valued, with density

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0. \end{cases}$$

The expected value of the lognormal distribution is

$$E(X) = e^{\mu + \frac{1}{2}\sigma^2}$$

and the variance is

$$Var(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}.$$

A.2 Transition probabilities of $\{N(k), k \geq 0\}$

Let $Y \sim Be(p_b)$ and $Z \sim Bin(m, p)$ be independent, and let $N(k-1) = m$, then $N(k) = m + Y - Z$. This follows since Y will be the number of new CMACs and Z will be the number of dead CMACs in time step k. The transition probabilities can then be derived as follows.

$$\begin{aligned} P_{m,m+1} &= P(N(k) = m+1 | N(k-1) = m) \\ &= P(Y = 1, Z = 0 | N(k-1) = m) \\ &= P(Y = 1 | N(k-1) = m) \cdot P(Z = 0 | N(k-1) = m) \\ &= p_b(1-p)^m \end{aligned}$$

$$\begin{aligned} P_{m,n} &= P(N(k) = n | N(k-1) = m) \\ &= P(\{Y = 0, Z = m-n\} \cup \{Y = 1, Z = m-n+1\} | N(k-1) = m) \\ &= (1-p_b)\tbinom{m}{m-n} p^{m-n}(1-p)^n + p_b \tbinom{m}{m-n+1} p^{m-n+1}(1-p)^{n-1} \\ &= p^{m-n}(1-p)^{n-1}\left(\tbinom{m}{n}(1-p) + p_b\left(\tbinom{m}{m-n+1}p - \tbinom{m}{n}(1-p)\right)\right) \\ &= p^{m-n}(1-p)^{n-1}\left(\tbinom{m}{n} - p\tbinom{m}{n} + p_b\left(\tbinom{m+1}{n}p - \tbinom{m}{n}\right)\right) \\ &= \tbinom{m}{n} p^{m-n}(1-p)^{n-1}\left(1 - p + p_b\left(\tfrac{m+1}{m+1-n}p - 1\right)\right), \; n \leq m. \end{aligned}$$

A.3 Expected value of $N(k)$

Let $Y \sim Be(p_b)$ and $Z \sim Bin(m, p)$ be independent. The expected value of $N(k)$ conditional on $N(k-1)$ is

$$\begin{aligned} E(N(k)|N(k-1)) &= N(k-1) + E(Y - Z) \\ &= N(k-1) + p_b - pN(k-1) \\ &= (1-p)N(k-1) + p_b. \end{aligned}$$

Now put $k = 1$ and take expectation of both sides of the equation above. Then, if $E(N(0)) = p_b/p$, it follows that

$$\begin{aligned} E(N(1)) &= (1-p)E(N(0)) + p_b \\ &= (1-p)N(0) + p_b \\ &= (1-p)\tfrac{p_b}{p} + p_b \\ &= \tfrac{p_b}{p} \end{aligned}$$

and, assuming that $E(N(l)) = p_b/p$, we have for $k = l + 1$ that

$$\begin{aligned} E(N(l+1)) &= E(E(N(l+1)|N(l))) \\ &= E((1-p)E(N(l)) + p_b) \\ &= (1-p)\tfrac{p_b}{p} + p_b \\ &= \tfrac{p_b}{p} \end{aligned}$$

and the result $E(N(k)) = p_b/p$ follows by induction.

If $N(0) \neq p_b/p$ we will show that

$$E(N(k)) = (1-p)^k N(0) + p_b \frac{1-(1-p)^k}{p}.$$

Then it follows that $E(N(k)) \to p_b/p$ as $k \to \infty$. We will again use induction; for $k = 1$ we have

$$E(N(1)) = (1-p)^1 N(0) + p_b \frac{1-(1-p)^1}{p}.$$

Assuming that

$$E(N(l)) = (1-p)^l N(0) + p_b \frac{1-(1-p)^l}{p},$$

we have

$$\begin{aligned} E(N(l+1)) &= E(E(N(l+1)|N(l))) \\ &= E\left((1-p)\left((1-p)^l N(0) + p_b \tfrac{1-(1-p)^l}{p}\right) + p_b\right) \\ &= (1-p)^{l+1} N(0) + p_b \tfrac{1-(1-p)^l}{p} - p_b(1-(1-p)^l) + p_b \\ &= (1-p)^{l+1} N(0) + p_b \left(\tfrac{1-(1-p)^l - p(1-(1-p)^l) + p}{p}\right) \\ &= (1-p)^{l+1} N(0) + p_b \left(\tfrac{1-(1-p)^l + p(1-p)^l}{p}\right) \\ &= (1-p)^{l+1} N(0) + p_b \tfrac{1-(1-p)^{l+1}}{p} \end{aligned}$$

and then we are done.

A.4 Distribution of the life time of a randomly chosen CMAC

Fix a time point k and let $K >> k$. Let $b_1 < b_2 < \ldots < b_N \leq K$ denote the birth times of CMAC $j = 1, \ldots, N$ born before or at time K and $d_1, \ldots, d_N$ the corresponding death times. Then $I_j = [b_j, d_j]$ denotes the life time of CMAC j. Choose J randomly from $\{1, \ldots, N\}$. Let $L_J = \max(0, k - b_J)$. As we are interested in the distribution of the time that a CMAC has lived so far given that it is alive at k, $P(L_J = i | I_J \text{ covers } k)$ is the quantity of interest. For $i \geq 0$ we have that

$$\begin{aligned} P(L_J = i | I_J \text{ covers } k) &\propto P(L_J = i, I_J \text{ covers } k) \\ &= \textstyle\sum_{b=0}^{K} P(L_J = i, I_J \text{ covers } k | b_J = b) P(b_J = b) \\ &= P(|I_J| \geq i + 1) P(b_J = k - i) \\ &= (1-p)^i P(b_J = k - i). \end{aligned}$$

Since $\{b_j\}$ constitutes a discrete time renewal process, we have as a consequence of the Key Renewal Theorem [14] that, as K grows, $P(b_J = k - i)$ will tend to a constant function of i when k is large enough, in the sense that

$$\max_{0 \leq i \leq i_{max}} \lim_{k \to \infty} \lim_{K \to \infty} K | P(b_J = k - i) - 1/K | = 0$$

for any fixed (and large) i_{max}. Asymptotically, we may therefore drop the term $P(b_J = k - i)$, and so the right-hand side is proportional to the probability function of a geometric distribution.

A.5 Weak convergence of first success random variables

For $X \sim Fs(p)$ we have that, $pX \xrightarrow{d} Z$, where $Z \sim Exp(1)$, when $p \to 0$. For $x \in \mathbb{R}_+$, consider

$$\begin{aligned} P(pX > x) &= P(X > \tfrac{x}{p}) = P(X > \left\lceil \tfrac{x}{p} \right\rceil) = (1-p)^{\left\lceil \frac{x}{p} \right\rceil} \\ &\Leftrightarrow \log P(pX > x) = \left\lceil \tfrac{x}{p} \right\rceil \log(1-p), \end{aligned}$$

which goes to $-x$ as $p \to 0$. Thus, $P(pX > x)$ goes to e^{-x} as $n \to \infty$ and the result is shown.

As for the situation in Subsection 4.4, since $p_b/p \to \rho$ and $(1-(1-p)^m)/p \to m$, we have that

$$\begin{aligned} pY &= \min\left(\left(\tfrac{p}{p_b} \right) p_b W, \left(\tfrac{p}{1-(1-p)^m} \right) (1-(1-p)^m) V \right) \\ &\xrightarrow{d} \min\left(\tfrac{Z_1}{\rho}, \tfrac{Z_2}{m} \right) \end{aligned}$$

as $p \to 0$, where Z_1 and Z_2 are independent $Exp(1)$-distributed random variables, and thus, $pY \sim Exp(\rho + m)$.

B Rough parameter estimates

In this section we assume that we have data from one cell observed in k time steps with a total of m CMACs observed during the whole experiment.

The parameter s is estimated from data with the average of the maxima of the areas of those CMACs whose average area is larger than 0.8, and whose life times are at least 4 (these values are chosen to match field 3 of the 090527 data). These boundaries are set because we want the areas of the CMACs used in the estimation to be reasonably large and that the their life time should be so long that they have been able to reach the substrate threshold. Assuming that CMAC j in data lives in the time index interval $[b_j, d_j]$, and that we have $n \leq m$ observations $(x_{j_1}(b_{j_1}), ..., x_{j_1}(d_{j_1})), ..., (x_{j_n}(b_{j_n}), ..., x_{j_n}(d_{j_n}))$ of the areas over time of CMACs $j_1, ..., j_n$, having the abovementioned properties, we have

$$\hat{s} = \frac{\sum_{i=j_1}^{j_n} \max(x_i(b_i), ..., x_i(d_i))}{n}.$$

The parameter p is estimated by the reciprocal of the average of CMAC life lengths corrected with the difference in resolution between simulation and data; as there are approximately 210 time steps in the model for each time index in data when simulating 20000 time steps we multiply our estimate with 1/210. Given m observations $y_1, ..., y_m$ of CMAC life lengths, we have

$$\hat{p} = \frac{1}{210 \cdot \bar{y}}.$$

To estimate p_b, we use the fact that the expected value of the number of CMACs in the model is $\lambda = p_b/p$ and that it, with observations $z_1, ..., z_k$ of the number of CMACs in time indexes $1, ..., k$, can be estimated by $\hat{\lambda} = \bar{z}$. Together with $\hat{p}$, this estimate is used to derive the estimate of p_b as

$$\hat{p_b} = \hat{\lambda} \cdot \hat{p}.$$

When estimating μ, we will have to consider the discrepancy between CMAC area progression in the model, giving multiplicative growth up to a threshold, and in data, which show a predominantly concave appearance of CMAC area trajectories. Thus, we will restrict our use of data to two observations of the area of every CMAC, its area at birth, and its area when one of the following events occur: the area reaches the estimate of s or, if the estimate of s is never reached, the area reaches its maximum over the life time of the CMAC in question. The value of the area of CMAC j when one of these events occurs is called a_j^{max}. As before, we will only use data on CMACs with an average area over 0.8 and a life length of at least 4 time steps.
With observations of CMACs $j_1, ..., j_n$ as above, our estimate of μ is

$$\hat{\mu} = \frac{\sum_{i=j_1}^{j_n} \left(\frac{a_i^{\max}}{x_i(b_i)} \right)^{\frac{1}{(d_i - b_i) \cdot 210}}}{n},$$

i.e. we use the quotient of CMAC area at that time index when one of the criterions above has been fulfilled and CMAC area at the first time index at which the CMAC in question is measured. These quotients are raised to 1/(CMAC life length*210), giving the average of the multiplicative growth factor over the whole life time of the CMAC in question, corrected for the difference in resolution. Finally, we take the average of these quotients to yield $\hat{\mu}$.

C Figures

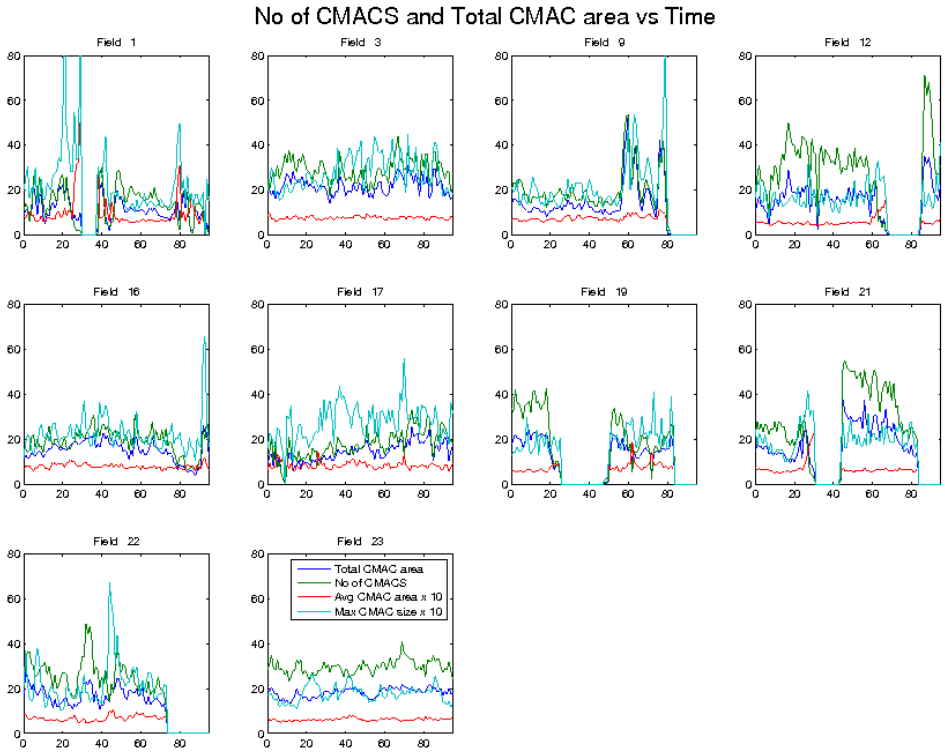

Figure C.1: The total CMAC area, number of CMACs, average CMAC area multiplied by 10 and maximum CMAC area multiplied by 10 in 10 cells (fields; each field corresponds to one cell) from the 090527 data. Fields 1, 9, 12, 19 and 21 undergo cell division at some point.

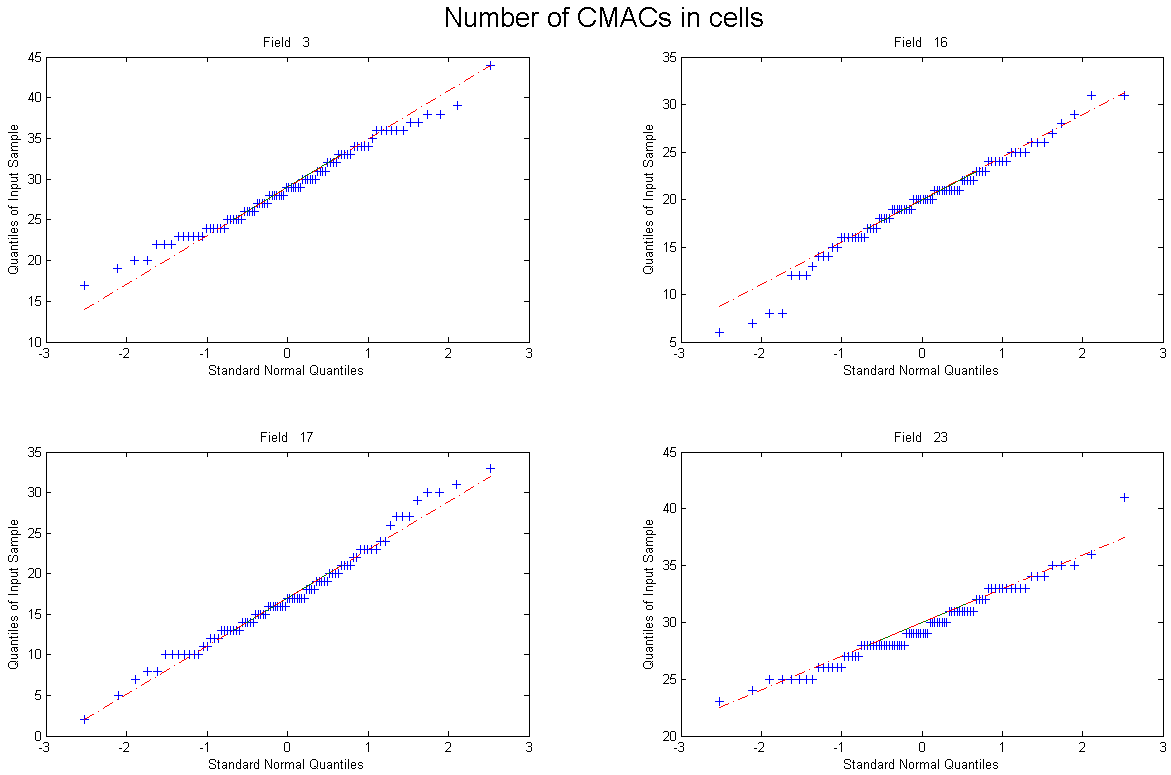

Figure C.2: Normal quantile-quantile plot of the number of CMACs over time in fields 3, 16, 17 and 23 in the 090527 data. Each field corresponds to one cell.

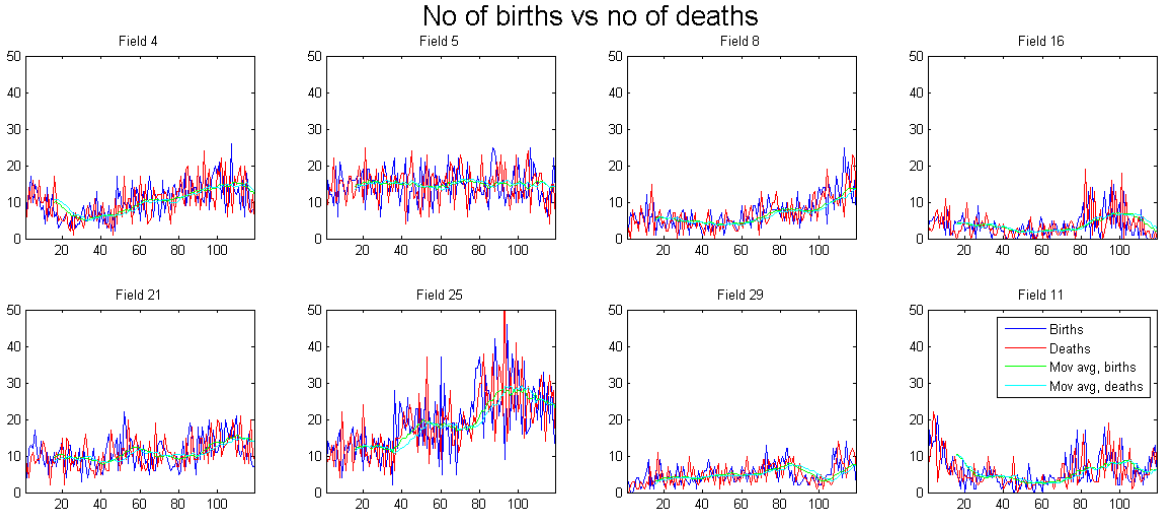

Figure C.3: Formation and disassembly of CMACs over time and their simple moving averages in 8 cells from the 090618 data. Cell division occurs in time index 0-8 in field 11.

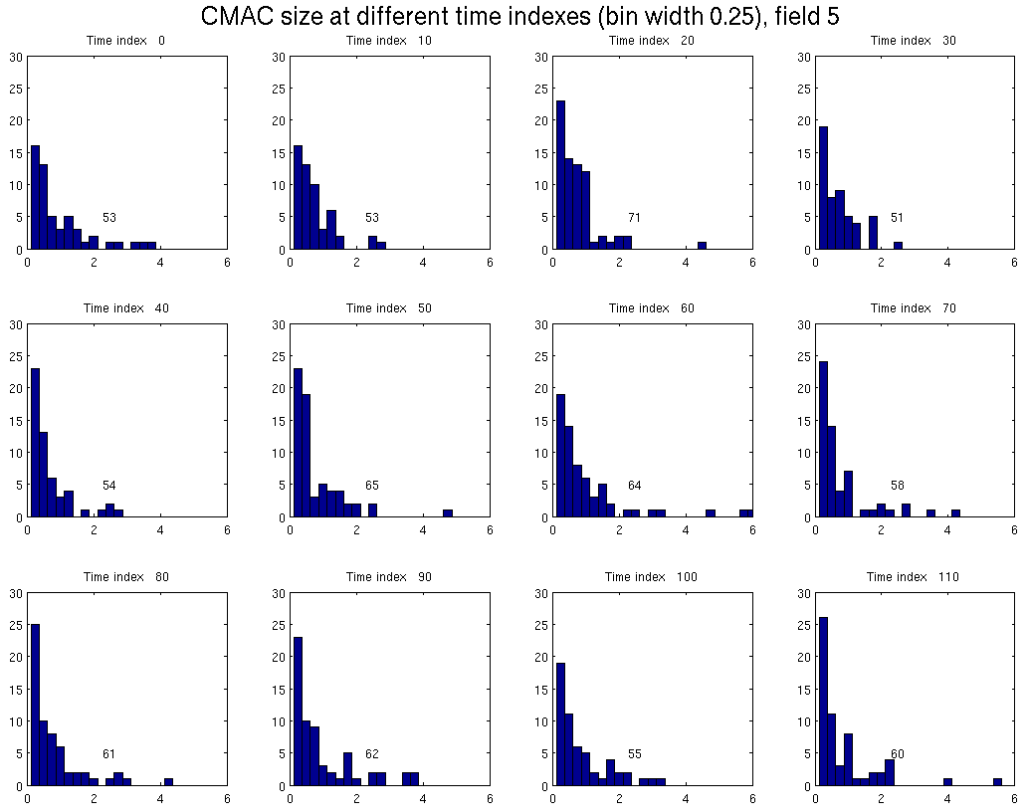

Figure C.4: CMAC area distribution in subsequent equidistant time indexes of the cell corresponding to field 5 from the 090618 data. The number of CMACs is shown inside the histogram.

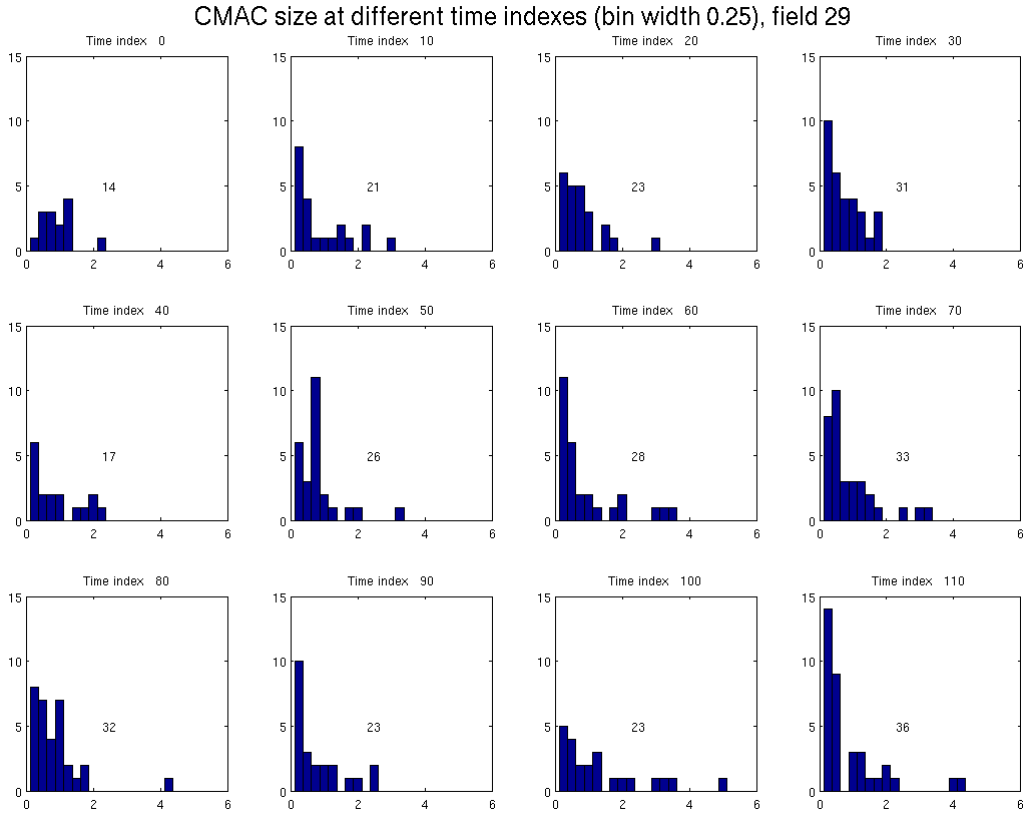

Figure C.5: CMAC area distribution in subsequent equidistant time indexes of the cell corresponding to field 29 from the 090618 data. The number of CMACs is shown inside the histogram.

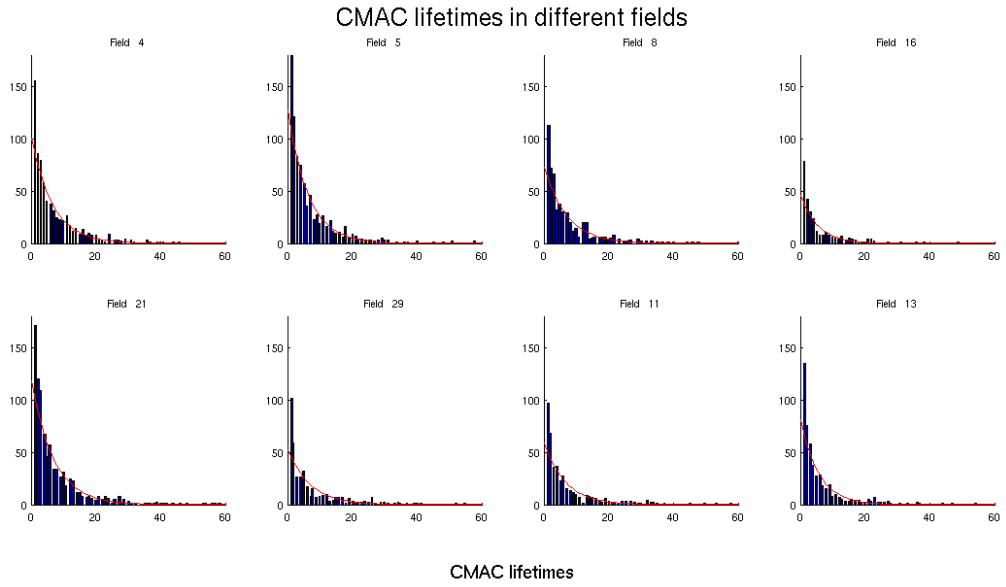

Figure C.6: CMAC life times in 8 cells from the 090618 data. The lines represent the estimated geometric distribution from the data for the corresponding field. Field 11 might be erroneous due to inclusion of corrupt data from time index 0-8. Note that, since there are no CMACs with a life time (defined as the number of time indexes that a CMAC lives minus one) shorter than 1, the geometric distributions are estimated outside the range of data.

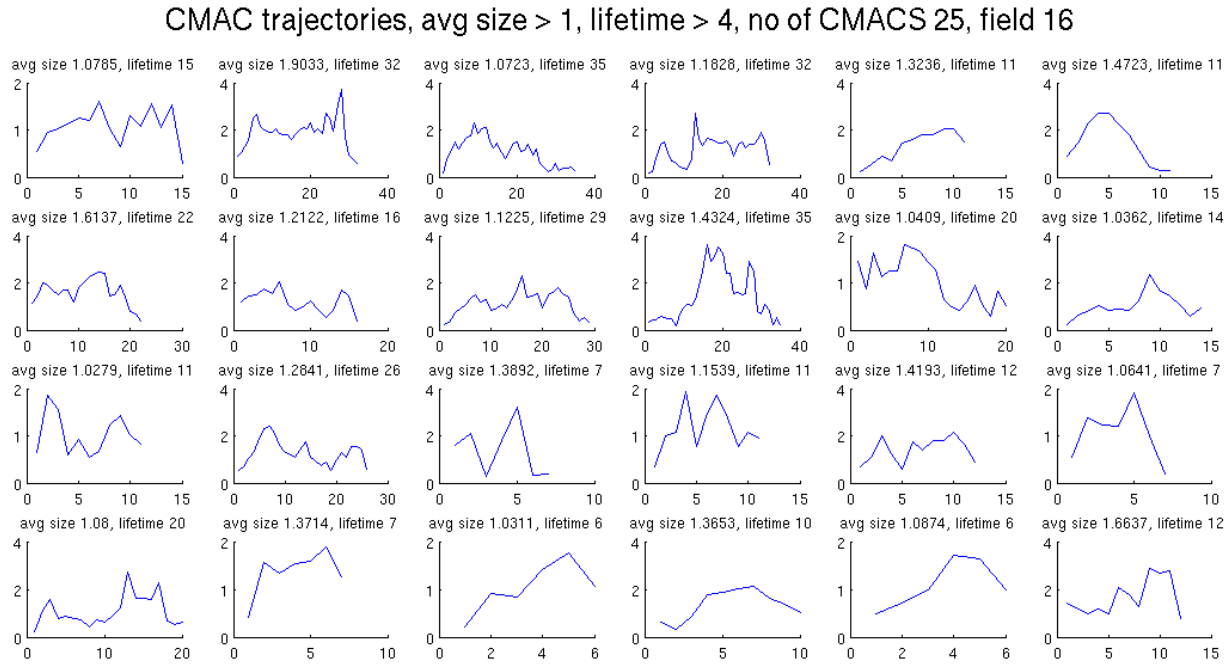

Figure C.7: CMAC area trajectories for CMACs with an average size larger than 1 and a life time of more than 4 (the plots show the number of time indexes that the CMAC in question is alive) in field 16 from the 090527 data.

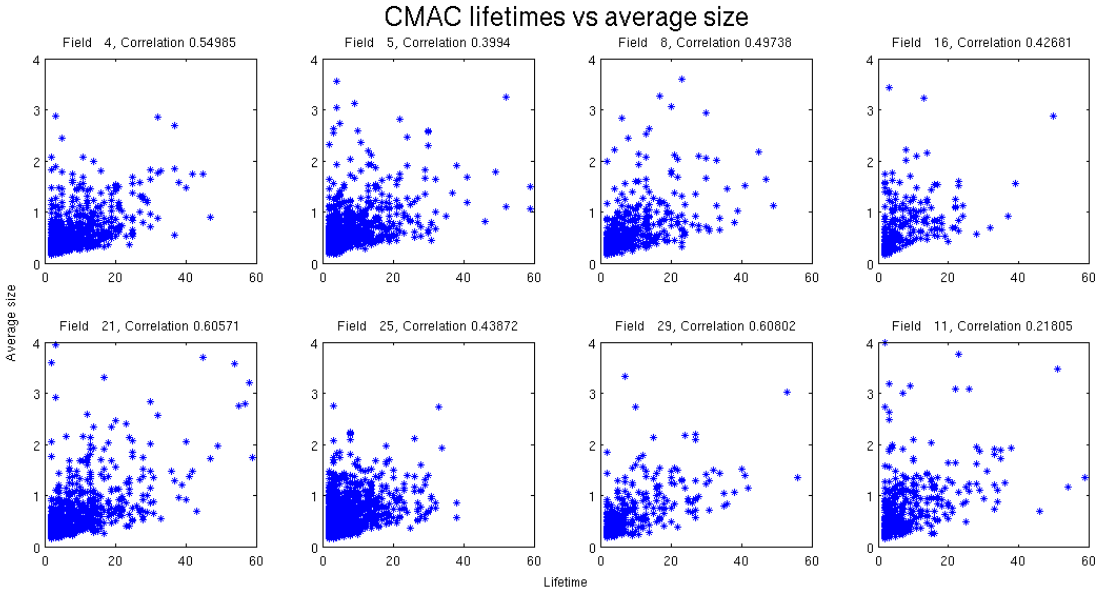

Figure C.8: CMAC life time vs average area in 8 fields from the 090618 data. Field 11 might be erroneous due to inclusion of corrupt data from time index 0-8.

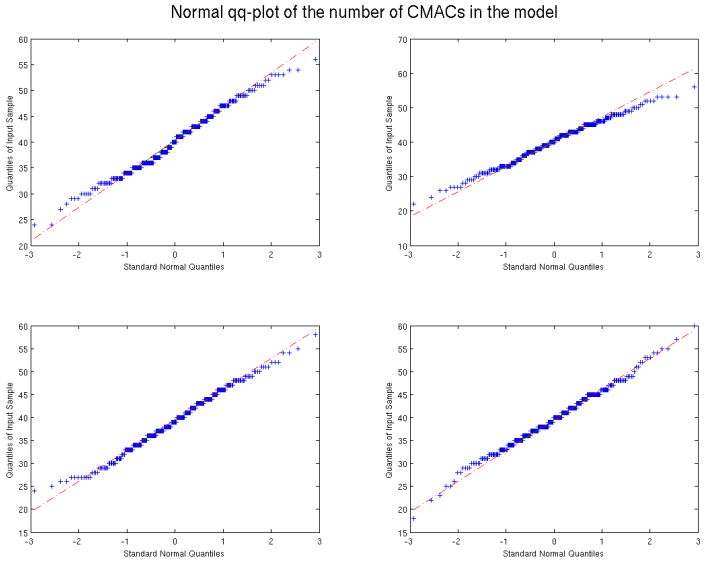

Figure C.9: Normal quantile-quantile plot of the number of CMACs from 4 simulations with $p_b = 0.08$ and $p = 0.002$. The number of CMACs is recorded every 1000 time steps, starting from time step 10000, giving a total of 291 observations in every simulation.

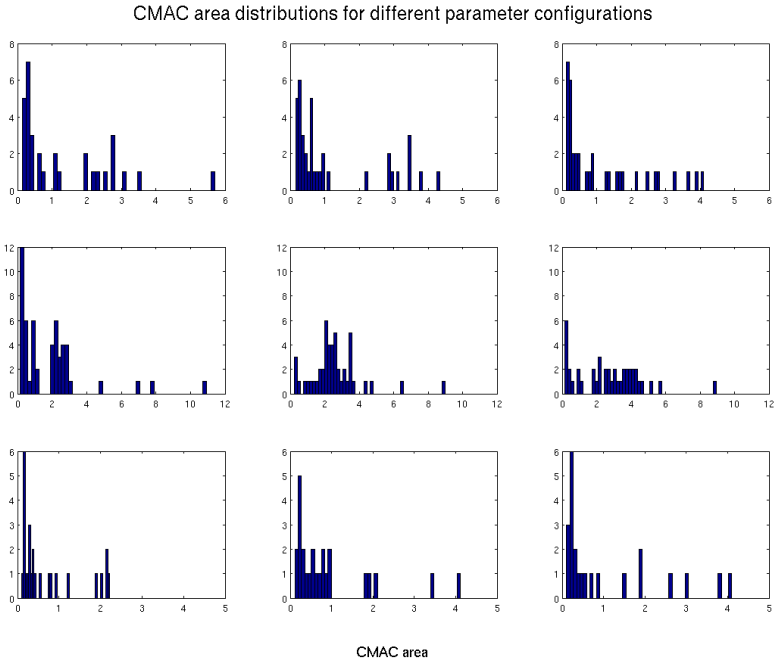

Figure C.10: Instantaneous empirical distribution of CMAC area at time points 10000, 13000, 16000 from three different simulations using the following parameter configurations: row 1, $p = 0.01$, $p_b = 0.0002$, $\mu = \log 1.0005$, $s = 2$; row 2, $p = 0.01$, $p_b = 0.0002$, $\mu = \log 1.001$, $s = 2$; row 3, $p = 0.004$, $p_b = 0.0002$, $\mu = \log 1.0003$, $s = 2$.

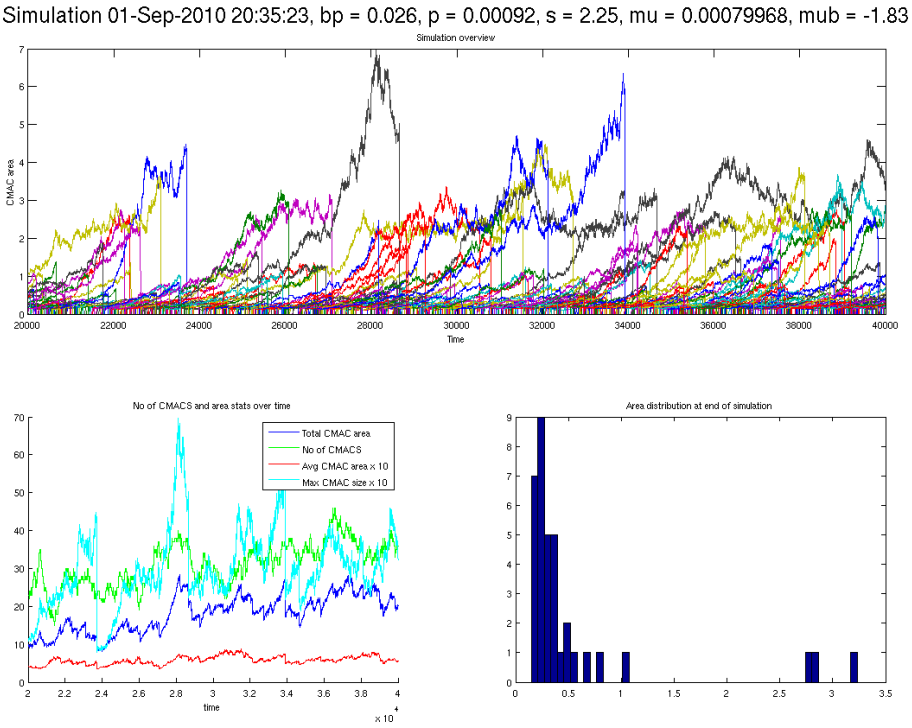

Figure C.11: Example of a model simulation with $p_b = 0.026$, $p = 0.00092$, $s = 2.25$, $\mu = \log 1.0008$ and $\mu_b = -1.83$.

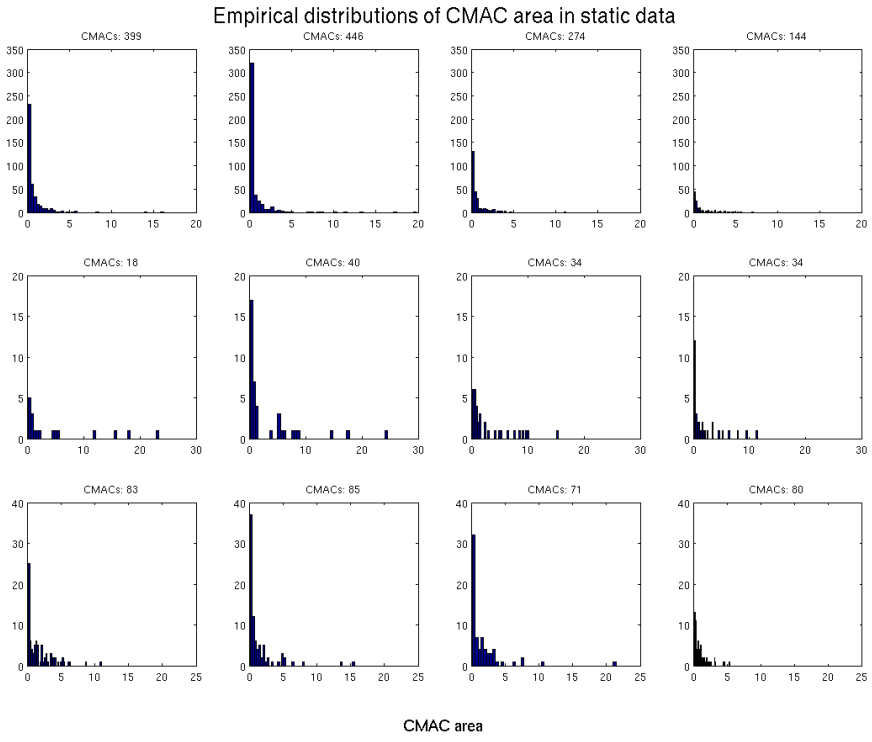

Figure C.12: Empirical distributions of CMAC area in different cells from the 090316 dataset. The upper row shows 4 cells perturbed with overexpression of the protein rac, the middle row shows 4 cells perturbed with overexpression of the protein rho and the lower row shows 4 unperturbed cells for comparison. The scaling of the histograms differ between the rows. The structure of the static data is somewhat involved and we will not present any details of the data used.

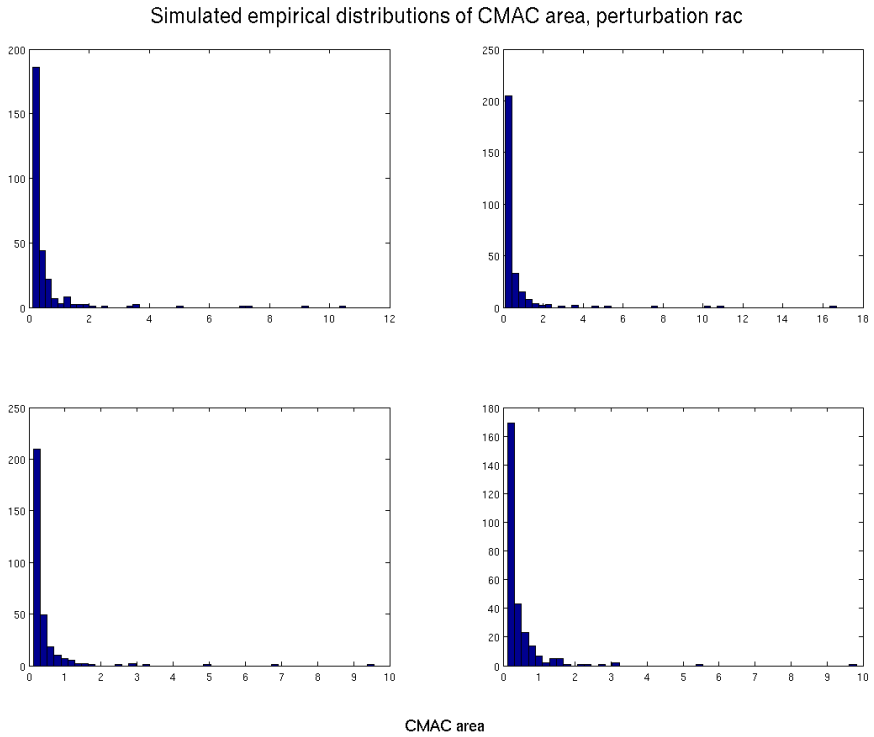

Figure C.13: Empirical distribution of CMAC area at time points 10000, 13000, 16000 and 2000 from a simulation with parameters $p = 0.3$, $p_b = 0.001$, $\mu = \log 1.0007$ and $s = 10$.

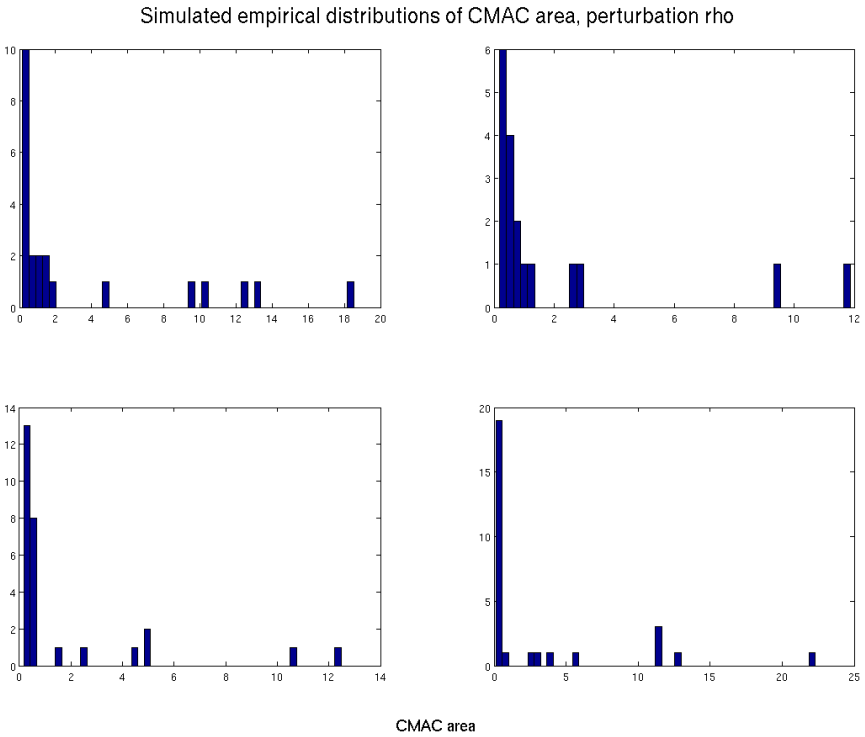

Figure C.14: Empirical distribution of CMAC area at time points 10000, 13000, 16000 and 2000 from a simulation with parameters $p = 0.01$, $p_b = 0.0004$, $\mu = \log 1.0008$ and $s = 10$.