Novel descriptive and model based statistical approaches in immunology and signal transduction

Submitted for the degree of Doctor of Philosophy

by

Juliane Liepe

Imperial College London

Division of Molecular Biosciences

Center for Bioinformatics

Typeset in Times by LaTeX

This report is the result of my own work.

No part of this dissertation has already been, or is currently being submitted by the author for any other degree or diploma or other qualification.

This dissertation does not exceed 100,000 words, excluding appendices, bibliography, footnotes, tables and equations. It does not contain more than 150 figures.

This work is supported by a Wellcome Trust grant and completed in the Division of Molecular Biosciences at Imperial College, London.

Abstract

Biological systems are usually complex nonlinear systems of which we only have a limited understanding. Here we show three different aspects of investigating such systems. We present a method to extract detailed knowledge from typical biological trajectory data, which have randomness as a main characteristic. The migration of immune cells, such as leukocytes, are a key example of our study. The application of our methodology leads to the discovery of novel random walk behaviour of leukocyte migration.

Furthermore we use the gathered knowledge to construct the underlying mathematical model that captures the behaviour of leukocytes, or more precisely macrophages and neutrophils, under acute injury. Any model of a biological system has little predictive power if it is not compared to collected data. We present a pipeline of how complex spatio-temporal trajectory data can be used to calibrate our model of leukocyte migration. The pipeline employs approximate methods in a Bayesian framework. Using the same approach we are able to learn additional information about the underlying signalling network, which is not directly apparent in the cell migration data.

While these two methods can be seen as data processing and analysis, we show in the last part of this work how to assess the information content of experiments. The choice of an experiment with the highest information content out of a set of possible experiments leads us to the problem of optimal experimental design. We develop and implement an algorithm for simulation based Bayesian experimental design in order to learn parameters of a given model. We validate our algorithm with the help of toy examples and apply it to examples in immunology (Hes1 transcription regulation) and signal transduction (growth factor induced MAPK pathway).

Contents

List of Figures

List of Tables

Abbreviations

ABC	approximate Bayesian computation.
Akt	Protein Kinase B
ATP	adenosine triphosphate
BM	Brownian motion
BRW	biased random walk
BPRW	biased persistent random walk
DMSO	dimethyl sulfoxide
dpf	days post fertilisation
EGF	Epidermal growth factor
EGFP	enhanced green fluorescent protein
EGFR	Epidermal growth factor receptor
GPU	graphical processing units
h	hours
hpw	hours post wounding
IL 8	interleukin 8
JNK	c-Jun N-terminal kinase
LPS	lipopylzsaccharid
MAPK	mitogen-activated protein kinases
MCMC	Markov chain Monte Carlo
min	minutes
p38	P38 mitogen-activated protein kinases
PC	principal component
PCA	principal component analysis

PRW	persistent random walk
RW	random walk
S6	Ribosomal protein S6
sec	seconds
SMC	sequential Monte Carlo

Acknowledgements

I woud like to thank Prof. Michael P.H. Stumpf for guiding and supervising my research. Thank you also to the Theoretical Systems Biology group for a great working environment, in particular Dr. Sarah Filippi and Dr. Michal Komorowski, who worked with me on experimental design methodologies. I also would like to thank Dr. Maxime Huvet for the many discussions about cell migration in zebrafish and Dr. Chris Barnes for the work on ABC implementations. Thank you to Adam MacLean for proofreading this thesis.

I would like to thank Dr. Harriet B. Taylor, Dr. Laurence Bugeon, Prof. Jonathan R. Lamb and Prof Magaret J. Dallman for working out the experimental setup and generating the zebrafish movies analysed in this project.

I acknowledge the financial support by the Wellcome Trust Fundation.

Introduction

1.1 OVERVIEW

The mathematical description of biological systems has over recent years enabled better understanding of many biological processes. Typically we start by studying individual processes, which we can consider as embedded in the complete biological system; we can refer to such individual processes as subnetwork. Each subnetwork exists on a specific scale, for example based on interactions between molecules, or the analysis of a cellular population. However, living organisms are organised on multiple scales that are correlated with each other, i.e. the organism's processes should be modelled on a multiscale basis. Due to the rise of "omics" technologies [1] and high-throughput methods, experimental data are becoming available for all types of cellular processes on a systems-level. These large datasets require us to integrate the current state of knowledge from the molecular scale to the whole organism level [2]. Another level of complexity is introduced by single molecule and single cell data. Analysing the average behaviour of a population of molecules or cells often hides the detailed dynamics of the system at the single cell level. The reason can be found in cell to cell variability, i.e. heterogeneity, but also in spatio-temporal effects. The nature of the experimental data and the diverse levels of complexity require advanced methodologies to extract the full information content in modern life sciences data.

In biological systems where little information is available the first goal is to analyse data using appropriate statistical methods. This process of explorative data analysis usually involves the computation of simple statistics, such as the mean and the standard deviation of a system output. However, the levels of complexity of the biological system as well as of the experimental data may require the development of more specialised approaches. The knowledge gained by such an initial analysis is then often incorporated into the process of constructing mathematical models.

Mathematical models of biomolecular systems are by design and necessity abstractions of a much more complicated reality [3, 4]. In mathematics, and the theoretical sciences more generally, such abstraction is seen primarily as a virtue which allows us to capture the essential features or defining mechanisms underlying the workings of natural systems and processes. But while qualitative agreement between even very simple models and very complex systems is easily achieved, formally assessing whether a given model is indeed good (or even just useful) is notoriously difficult. These difficulties are exacerbated in no small measure for many of the most important and topical research areas in biology [5–7]. The regulatory, metabolic and signalling processes involved in cell-fate and other biological decision-making processes are often only indirectly observable; moreover, when studied in isolation their behaviour can often be markedly altered compared to the experimentally more challenging *in vivo* contexts.

These challenges have prompted the development of novel statistical and inferential tools, required to construct (or improve) mathematical models of such systems. We can loosely group these methods into (i) those aimed at reconstructing network models [8–10] (using correlations or statistical dependencies in observed datasets), (ii) methods to estimate (biochemical reaction) rate parameters of models describing the dynamics of biological systems [11–13], and (iii) approaches that allow us to rank or discern between different candidate models/hypotheses [14, 15]. The first set of challenges is typically faced when dealing with new systems where little information is known, and where network-inference algorithms offer a convenient way of generating novel mechanistic hypotheses from data. The latter two types of problems are frequently (and perhaps should be generally) considered together, and in most instances rely on our ability to formulate suitable candidate models based e.g. on prior knowledge or biophysical/biochemical reasoning.

We can summarise the above described aspects in two main categories: (i) describing and understanding the data to understand the biological system; and (ii) mathematical modelling of the system using the data. A third important topic arises almost naturally: Which experiments should be performed in order to gather the most information about the biological system? This is investigated in the field of experimental design [16]. Once a mathematical description of a system is achieved, it is the aim to calibrate it or to distinguish between competitive models. It is important to understand what type of experiments can be performed that are most informative. While some experiments contain no or little additional information to previously collected data, others will allow to fully answer the biological question of interest. For small networks, this question might seem trivial and the answer is often intuitive. However, the majority of biological systems are complex and nonlinear. To find the optimal experiment turns into an equally complex problem.

1.2 OBJECTIVES

The aim of this PhD thesis is to find new ways to describe and model complex biological systems. In detail we will develop methods for:

- analysis of spatio-temporal leukocyte trajectory data
- model calibration and model choice using spatio-temporal cell trajectory data
- experimental design to calibrate mathematical models of biological systems.

In this thesis we will introduce the main concepts to achieve these goals. We will present novel statistical approaches and illustrate their usefulness by applying them to systems in immunology and signal transduction. A key system will be the migration of leukocytes in response to acute wounding.

1.3 THE IMMUNE SYSTEM

In this section we will outline the basic concepts of the immune system. More specifically we will introduce the innate immune system with its main cell types, i.e. macrophages and neutrophils. This leads us finally to the principals of cell migration via chemotaxis.

1.3.1 *Innate versus adaptive immunity*

The mammalian immune system is comprised of a set of components that defend the organism against foreign substances and organisms such as pathogens. The organs of the immune system include the skin, the bone marrow, the spleen, the thymus and the lymph nodes [17]. Another important component are specialised immune cells and molecular components that affect pathogens. Most mammalian immune systems are composed of the innate and the adaptive immune system, but the innate immune response is always present [17].

The innate immune system is a non-specific response system and it represents the first layer of defence against pathogens. The main parts are tissue barriers (e.g. the skin), innate immune cells (e.g. macrophages and neutrophils) and secreted soluble molecules to attack foreign organisms.

In contrast to the innate immune system, the adaptive immune system works in a specific manner [17]. It is the acquired immunity that is able to generate memory of specific pathogens after a previous response to that pathogen. Characteristic of the adaptive immune system are the production of specific antigens and immune cells, which recognise specific pathogens. The adaptive immune response is slower than the first innate immune response. However, once an immunological memory against a certain pathogen is created the adaptive response is enhanced and more effective in case of a repeated contact. This is the fundamental basis of vaccination.

Our main focus here will be on the innate immune response and we describe some of its components in more detail below.

1.3.2 *Macrophages and neutrophils*

Leukocytes, more commonly known as white blood cells, are important cells during nearly all stages of the immune response. These cells can be divided into several cell types: neutrophils, basophils, eosinophils, lymphocytes, macrophages, monocytes and dendritic cells [17]. The cells which play a crucial role during inflammation and therefore during the innate immune response are macrophages and neutrophils (figure 1.1). The more general term leukocytes is often used in the literature to refer to these two cell types when discussing the innate immune response during inflammation. However, a third cell type, monocytes, which are highly mobile leukocytes

circulating in the blood vessels, initiate the immune response. They leave the blood vessels and enter the tissue after receiving a signal. This process is described as leukocyte recruitment. In the tissue the monocytes develop into the more sedentary macrophages.

Macrophages can survive for long periods of time as sentinels guarding against foreign pathogens [18]. In addition to their phagocytic duties, macrophages posses a large number of receptors for pathogens and pathogen-related substances, which can trigger the secretion of proinflammatory cytokines and chemokines after activation. Macrophages are important cells during wound healing processes. Their detailed functions include the promotion of inflammation. After macrophages are activated by for example pro-inflammatory cytokines, LPS or interferons, macrophages produce various cytokines. These cytokines include interleukin-1, interleukin-6, interleukin-12 and TNFα. [19]. Furthermore macrophages are known to produce chemoattractants. The specific secreted chemokines into the extracellular matrix can activate further cell types of the innate immune response, such as neutrophils [18, 20], but also cells of the adaptive immune response and therefore act as a link between the two systems. Apart from their pro-inflammatory function, macrophages that are related to wound healing also have an anti-inflammatory function. It is suggested by *in vitro* studies that macrophages can switch from the pro-inflammatory state to an reparative state [21, 22]. Latter shows the expression of anti-inflammatory mediators, such as IL-1R antagonist or interleukin-10. Furthermore they produce growth factors (e.g. vascular endothelial growth factor, insulin-like growth factor). These growth factors promote the cell proliferation and protein synthesis [23]. Several studies provide strong evidence that macrophages promote angiogenesis, fibroblast proliferation and synthesis of the extracellular matrix [24–26]. Another important function of macrophages during wound healing is the removal of neutrophils in order to facilitate the repair processes.

Neutrophils have a shorter life span than macrophages (on average 5 days), but appear in much higher numbers during acute inflammation [17]. Neutrophils appear mainly in early wounds. The main function of neutrophils is to generate an antimicrobial environment. This is achieved by phagocytosis, a mechanism by which they engulf microbes [27]. As part of this process, neutrophils negatively influences the process of wound healing, mainly because they destroy surrounding tissue [28].

They contain proteases, which degrade major components of the extracellular matrix and even cytokines. this strongly reduces repair mechanisms during wound healing. Furthermore neutrophils are involved in the production of hydrogen peroxide and other oxygen radicals, which are toxic for microbes. This so-called oxidative burst functions as a defence mechanism [20, 29]. This oxidative burst can cause further tissue damage. All this indicates, that the main function of neutrophils is the decontamination of the wound, rather than the repair of the wound. For this reason they appear in high abundance in poorly healing wounds, while the number of neutrophils in wounds that heal well is low [30]. As mentioned above, it is important that neutrophils are removed from the site of the wound after decontamination, which is supported by macrophages.

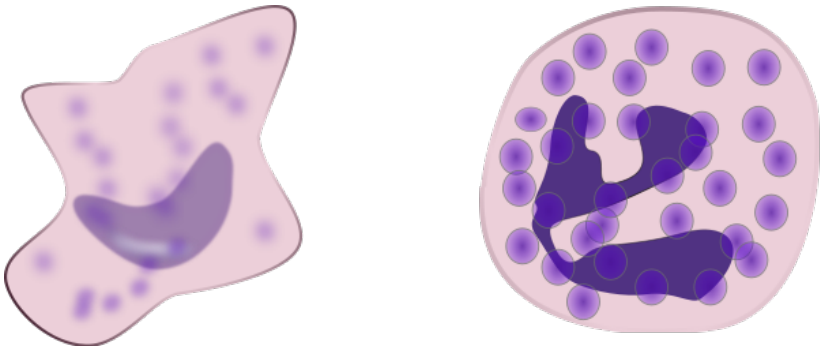


Figure 1.1: Cartoon representations of macrophages and neutrophils. The morphological differences of macrophages and neutrophils are clearly visible in this representation: the macrophage (left) has a large kidney-shaped nucleus, while the neutrophil (right), a polymorphonuclear granulocyte, has a lobed nucleus and toxic granules within its cytoplasm. (These images, which are in the public domain, were taken from http://openclipart.org/user-detail/keikannui).

1.3.3 Cell migration: Chemotaxis

A vital component of the immune system is the ability of cells to migrate in the extracellular matrix. Any immune response which involves immune cells requires the migration of these cells. Examples are the migration of tissue "patrolling" macrophages to protect the organism from pathogens, or the migration of leukocytes during wound healing towards the wound.

The molecular mechanism that leads to the migration of a cell is still not well

understood. The majority of migrating cells produce a protrusion of the cytoplasm and cell membrane, a so called leading edge, towards the direction of movement. Two main theories exist that try to explain the molecular details of this process. The first theory is the cytoskeletal model. It is based on the polymerisation of actin filaments at the front of the cell, which guides the cell towards a specific direction [31–34]. The second theory is the membrane flow model, which is based on the observation that membrane of the migrating cell flows towards the front of the cell [35].

A specific form of cell migration is chemotaxis, which was first described in 1881 by Engelmann [36] and in 1884 by Pfeffer [37]. Chemotaxis describes the migration of cells in response to an external stimulus. This response can be attracting (positive chemotaxis), i.e. the migration occurs towards the stimulus, or repellent (negative chemotaxis), i.e. the migration goes away from the stimulus. Accordingly we distinguish between chemoattractants and chemorepellents, which are in general called chemokines. The behaviour of migrating cells in response to different chemokines was studied in isolation intensively using cell cultures [38–40]. However, immune cells in an organism are subject to a combination of different chemokines. Because of that the cell migration process results from the integration of all surrounding signals.

1.3.4 Chemokines and Cytokines in inflammation and wound healing

During inflammation and the early stages of wound healing chemokines are produced, which result in the recruitment of immune cells to the site of inflammation. The chemokines activate specific receptors and regulate downstream mechanisms. Apart from the release of cytokines, reactive oxygen species (ROS) are produced, which help to decontaminate the tissue [41]. The cytokines, as well as chemokines, which are released during inflammation can be distinguished in pro-inflammatory and anti-inflammatory cytokines / chemokines. Leukocytes release in the first stages of inflammation pro-inflammatory cytokines, interleukin-1α and -β. Other pro-inflammatory molecules are interleukin-6, the chemokine CX3C, TGF-β and TNF-α [42]. Latter is mainly produced by macrophages and regulates other immune cells. TNF-α is directly involved in the apoptotic cell death and therefore links back to the depletion of neutrophils to facilitate the repair processes during wound healing [42]. The main anti-inflammatory cytokine is interleukin-27, which is produced by macrophages.

Interleukin-27 regulates the Il-27 receptor as well as the STAT3 pathway and therefore influence the activity of B-lymphocytes and T-lymphocytes [43].

1.4 INVESTIGATION OF CELL MIGRATION

In this section we will present classical methods to investigate cell migration. We will show the advantages of the zebrafish as a suitable experimental model and introduce the mathematical models of leukocyte migration that have been considered thus far.

1.4.1 Cell migration assays

Leukocyte recruitment is an essential and early component of the inflammatory response to local injury [17]. It was first described in 1881 (T. W. Engelmann) that leukocytes perform chemotaxis, i.e. move upwards a chemical gradient to find the source of an injury [44]. Leukocyte migration has been studied extensively using the so-called chamber system [45, 46, 46, 47]. Isolated cells are positioned in one part of the chamber, while a chemoattractant is placed in the other part. Several types of chambers have been created. The Boyden chamber uses a top-bottom configuration [38]. The Zigmond chamber is based on a sidewards organisation [48], while the Dunn chamber has a centric arrangement [49]. These experimental setups are able to evaluate how many cells migrate from one side of the chamber to the other in a given time. In this way several chemical components can be tested whether or not they can act as chemoattractants. These simple assays do not, however, aim to probe the actual process of migration.

A more advanced technology are microfluidic devices [40, 50]. Jeon *et. al.* constructed a system, which produces a stable gradient of a chemoattractant over time [40]. The gradient shape can be chosen by the experimentalist. This allows researchers to study the migration behaviour of a cell in gradients of chemoattractants as well as chemorepellents and facilitates a dynamical analysis. The major problem of these methods is the unnatural environment that does not reflect the presence of an organism, which is known to impact the migration process. An example would be the structure of the extracellular matrix: a migrating cell is interacting via membrane receptors with components of the matrix. These components are difficult to mimic in a microfluidic device. A novel platform that aims to collect and integrate data about

the molecular details of cell migration is the cell migration gateway (CMG) [51].

We will use these concepts in chapters 3 and 4, where we will investigate the migration of macrophages and neutrophils.

1.4.2 Zebrafish as an experimental model

Many, if not most, open problems in biomedical research involve questions related to whole organism biology. Despite the wealth of insights provided by molecular and cell biology, genetics and genomics, we still do not understand most of the tissue-level, physiological and organism-level processes underlying e.g. development, health and disease. *In-vivo* studies allow us to investigate biological processes at the level of the organism.

The zebrafish is an attractive model in which to study inflammation and the immune system as we can combine *in vivo* imaging of immune processes with molecular analyses that probe or interfere with the molecular signalling processes underlying the immune response.

The innate immune system of a zebrafish, which closely resembles that of mammals, is fully competent at early embryological stages before the emergence of lymphocytes allowing dissection of innate responses in the absence of adaptive immunity. Several studies have demonstrated that tail fin wounding in zebrafish embryos, including tail transections, medial fin incisions and laser induced wounds result in the migration of leukocytes to the site of tissue damage [52–56]. This migration is dependent in part on a hydrogen peroxide gradient produced at the wound margin [57]. How leukocytes respond to this and other signals and reach the decision to migrate towards and remain at the site of injury or infection is only incompletely understood. In particular, it is not known what determines the precise nature and magnitude of the innate immune response, or how individual cells "decide" whether to engage with immune stimuli. Different mechanisms have been proposed: cytokine signalling or sensing of extracellular hydrogen peroxide concentrations could (i) either alter the speed at which leukocytes migrate towards the origin of these signals, (ii) or the number of cells recruited to and retained at the site of injury. Being able to distinguish between these two mechanisms is of fundamental importance for understanding the processes by which immune system cells reach decisions. Ultimately, being able to

manipulate or guide these processes will open up novel therapeutic opportunities.

Previous studies in zebrafish have addressed several aspects of leukocyte migration in isolation: we know, for example, that innate immune cells in zebrafish embryos activated by wounding alone or by bacterial infection express adaptor molecules associated with TLR (Toll-like receptor) mediated signalling [58] and that LPS can induce the expression of inflammatory cytokines [59] and the activation of NF-κB [60]. The p38 MAPKs are stress-activated proteins central to cellular responses induced by external stimuli and due to their role in the response to pro-inflammatory cytokines are considered candidate drug targets for treatment of a broad range of inflammatory diseases [39, 61–63].

The MAPK pathways are important mediators of cellular responses to inflammatory signals including leukocyte migration behaviour. During inflammation the MAPK signal transduction pathway is activated. The activation of p38 MAPK as well as regulating the production of inflammatory mediators regulates the effector function of leukocytes by controlling their migration in response to inflammatory stimuli [64, 65]. p38 MAPK is involved in the regulation of pro-inflammatory cytokine expression, and therefore directly influences the behaviour of macrophages and neutrophils [66]. Furthermore p38 MAPK regulates the inducible NO synthase, which is essential for neutrophils during the oxidative burst phase [66]. A much studied but still little understood anti-inflammatory component is the p38 inhibitor SB203580. Many in vitro studies, as well as studies on cell cultures show that inhibiting p38 results in a decreased straightness index, velocity and recruitment of cell numbers under acute injury and/or when stimulated with LPS [67, 68]. However, some studies report contradictory effects [69, 70]. The use of p38 inhibitors in clinical trials did also not fully confirm the anti-inflammatory behaviour.

Another molecule of the MAPK pathway is JNK (JUN N-terminal kinase). Several JNK substrates are known to impact actin regulation and cytoskeleton remodelling, including MAP1B, MAPA2, DCS and SCG10. These substrates are likely to play an important role during inflammation and cell migration processes [67, 71–73]. JNK has recently been shown to play an important role in insulin-resistance induced by obesity [74]. Studies have been conducted using the anti-inflammatory JNK-inhibitor SP600125. However, the overall role of JNK during inflammation still remains unclear.

The optical translucency of zebrafish, together with the availability of transgenic lines that express fluorescent proteins under myeloid lineage specific promoters, allows real-time imaging of migrating leukocytes in response to inflammatory stimuli [53, 54, 56, 58, 71]. These imaging data can be analysed using automated image analysis and cell tracking [75]. We will make use of these principles in chapter 3.

1.4.3 Models of cell migration

The underlying mechanisms that lead to migration of eukaryotic cells have been studied over the last decades. Several types of mathematical models were formulated and investigated. In this section we will focus on the biological motivation of these modelling approaches, while we will present the mathematical background in chapter 2.

The mathematical modelling approaches can be classified as follows: (i) stochastic models of random walks; (ii) Monte Carlo modelling approaches; (iii) force based dynamical models and (iv) biochemical models of cell migration. The first class, stochastic models, have been investigated by Tranquillo *et. al.* [76, 77] and Stokes *et. al.* [78, 79]. Both describe cellular motion as persistent random walks by numerically solving the Langevin equation. In this way the model generates single cell tracks, but also captures whole population dynamics. However, the details about the biochemical or biophysical mechanisms can not be investigated with such an approach. The initial 2D migration model was later extended to 3D by Parkhurst *et. al.* [80]. In recent years these type of models have been further developed to account for more complicated types of random walks and for random walks in crowded environments. Examples include the work by Painter [81] and Murray [82].

The Monte Carlo modelling approaches have been applied by Zaman *et. al.* [83, 84] to describe cell migration on lattices in 2D as well as in 3D. Short simulation times are the major advantages of these type of models. Furthermore the model description can incorporate simple rules that determine cell migration. Recently, Kim *et. al.* [85] published a dynamic model of cell migration on planar substrates, which combines Monte Carlo simulations with a force based description of the cell migration process.

The force based dynamical models aim to describe the biophysical processes of

cell migration. This includes the interaction of the cell with the surface (matrix) that it is located on, as well as the resulting change in cell shape. The cell tracks are simulated based on internally generated traction forces. The model describes the biophysical process by introducing parameters that describe the matrix stiffness, the matrix density and the cell-matrix adhesivity. This type of cell migration model provides more detailed insights into the cell migration process of a single cell, but it is so far not suited to describe the population behaviour. A well established example of a forced based dynamical model was published by Zaman *et. al.* [86].

The biochemical models of cell migration aim to describe the intracellular and/or extracellular signalling processes that lead to cell migration. The most described processes are the polarisation of intracellular signalling, actin cytoskeleton remodelling and focal adhesion signalling. Although first steps towards a detailed mathematical model have been made [87], the main task here is still to reconstruct the pathways that regulate these processes. The cell migration gateway [88] aims to collect all available information in order to facilitate a thorough mathematical analysis of these signalling cascades.

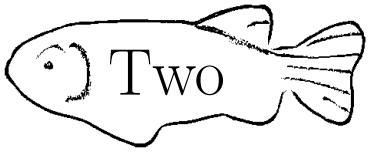
Several approaches have been published that model more specifically the migration of leukocytes in response to chemokines. The above mentioned work by Tranquillo *et. al.* [77] in 1988 presents a stochastic model for leukocyte chemotaxis. The model is based on receptor binding fluctuations and captures two observations from cell culture experiments: (i) in a uniform chemoattractant source the migrating cell shows a persistent random walk; (ii) in a gradient of chemoattractants the cell is migrating towards higher chemoattractant concentration in a biased random walk manner. They assume that a cell polarises during the migration process and that the cell polarity is maintained during every step. The polarity is modelled based on noise in the cellular signal response mechanisms. This model ignores the details of the underlying molecular processes such as kinase signalling cascades.

Since more experimental data about chemotaxis became available, Onsum *et. al* [87] constructed a mathematical model of neutrophil gradient sensing, which is based on the underlying signalling cascades. Using partial differential equations they model actin polymerisation at the front of a cell. This polymerisation results from the local activation of the phosphatidylinositol-3-kinase (PI3K) and Ras signalling pathway, which ion turn leads to the translation of an external stimulus gradient to

an intracellular gradient of signalling molecules. Although the authors dramatically simplify the shape of a migrating cell as a constant two dimensional box, the model shows the importance of spacial models to explain the migration process.

This already complex model ignores cellular morphology (and its changes) and it does not describe the actual migration process of the cells. A detailed analysis that describes changes in the cellular morphology was proposed in the same year [89]. Later different groups focussed on the capability of leukocytes to sense external signals [90]. Recently the movement of leukocytes was described by modelling the extension of pseudopodia [91].

These modelling approaches consider either the sub-cellular processes (e.g. production of internal gradients) or the cellular migration processes in response to chemical gradients. Here we present the first approach to integrate these two scales into one model that also contains intracellular signalling processes (3).



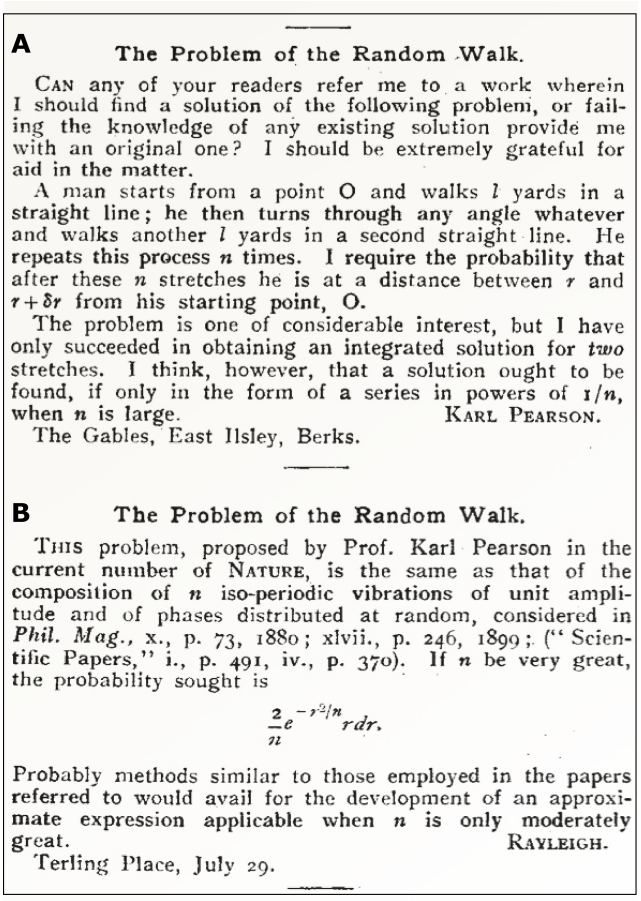
Theoretical background

2.1 RANDOM WALKS

2.1.1 History

In 1827 the scottish botanist Robert Brown (1773 - 1858) observed a jittery motion of particles suspended in water under the microscope. The movement seemed to be continuous and without any regularity and therefore completely random. Even though Brown himself did not propose any theory that describes this motion, the process by which a molecular particle is moving was later described as Brownian molecular motion. This problem of such a random process, although observed much earlier, was not mathematically investigated until 1905, when Karl Pearson sent a letter to the journal Nature formulating the problem of a random walker and asking for help on this matter (figure 2.1A). The main important response to this proposal was from Lord Rayleigh, pointing out that this problem was published and investigated in 1880 and 1899 in the field of soundwaves in heterogeneous materials (see figure 2.1B). Indeed, random walks had been investigated under a different terminology.

Karl Pearson summarises the letter from Lord Rayleigh in the same issue of Nature with: "The lesson of Lord Rayleigh's solution is that in open country the most probable place to find a drunken man who is at all capable of keeping on his feet is somewhere near his starting point!" The communication between Karl Pearson and



A

The Problem of the Random Walk.

CAN any of your readers refer me to a work wherein I should find a solution of the following problem, or failing the knowledge of any existing solution provide me with an original one? I should be extremely grateful for aid in the matter.

A man starts from a point O and walks l yards in a straight line; he then turns through any angle whatever and walks another l yards in a second straight line. He repeats this process n times. I require the probability that after these n stretches he is at a distance between r and $r+\delta r$ from his starting point, O.

The problem is one of considerable interest, but I have only succeeded in obtaining an integrated solution for *two* stretches. I think, however, that a solution ought to be found, if only in the form of a series in powers of $1/n$, when n is large. KARL PEARSON.

The Gables, East Ilsley, Berks.

B

The Problem of the Random Walk.

THIS problem, proposed by Prof. Karl Pearson in the current number of NATURE, is the same as that of the composition of n iso-periodic vibrations of unit amplitude and of phases distributed at random, considered in *Phil. Mag.*, x., p. 73, 1880; xlvii., p. 246, 1899; ("Scientific Papers," i., p. 491, iv., p. 370). If n be very great, the probability sought is

$$\frac{2}{n}e^{-r^2/n}r\,dr.$$

Probably methods similar to those employed in the papers referred to would avail for the development of an approximate expression applicable when n is only moderately great. RAYLEIGH.

Terling Place, July 29.

Figure 2.1: Random walks in early literature. Text passage from the journal Nature (Volume 72, p. 294 and p. 318, 1905). Communication between Karl Pearson (A) and Lord Rayleigh (B).

Lord Rayleigh drew attention to mathematicians and physicists like Einstein (1905, 1906), Smoluchowski (1916) or Ornstein and Uhlenbeck (1930), to name just a few. We will introduce several types of random walks, which are also used in biology to describe animal movement, cell movement, as well as the motion of molecules. A good overview about random walks in biology can be found in [92, 93].

2.1.2 Isotropic random walks

The most basic random walk is the isotropic random walk. Let us imagine a random walker on a straight line, starting at position, $x = 0$. In each time step, τ, our random walker can either move a step to the left or a step to step right with equal probability

q, i.e. $q = 1/2$. We assume for now that the step length is constant, l. We then want to know the probability that our random walker is at position $x^* = nl$ after N time steps, i.e. after time $N\tau$. This problem has been solved in several ways [94–96]. In the discrete case, after one time step τ the walker can be either at position $x = l$ or $x = -l$ with probability $1/2$. In the following tie step τ the walker will be at position $x = 2l$ with probability $1/4$, at position $x = -2l$ with probability $1/4$, or at its origin $x = 0$ with probability $1/2$. Continuing this logic for N time steps, the probability of the walker being at position $x = nl$ is described by the binomial distribution

$$p(n, N) = \left(\frac{1}{2}\right)^N \frac{N!}{(\frac{1}{2}(N+n))!(\frac{1}{2}(N-n))!} = \left(\frac{1}{2}\right)^N \binom{N}{\frac{N-n}{2}}. \tag{2.1}$$

This equation is valid for even n and N. For large N the binomial distribution converges to the Normal distribution [97]:

$$\lim p(n, N) = \left(\frac{2}{N\pi}\right)^{1/2} \exp\left(\frac{-n^2}{2N}\right). \tag{2.2}$$

Using the fact that n is even as well as the above described relations $x = nl$ and $t = N\tau$ we define.

$$P(x, t)dx = p\left(\frac{x}{l}, \frac{t}{\tau}\right)\frac{1}{2l}dx. \tag{2.3}$$

The probability of being between the positions x and $x + dx$ is then given by

$$p(x \in (x, x+dt), t) = \frac{1}{\sqrt{2\pi l^2 \frac{t}{\tau}}} e^{-\frac{x^2\tau}{2l^2 t}} dx. \tag{2.4}$$

Analysing the limits $\tau, l \to 0$ with the constant $l^2/\tau = 2D$, we obtain

$$p(x \in (x, x+dt), t) = \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{x^2}{4Dt}} dx. \tag{2.5}$$

Equation 2.5 is the standard solution of the diffusion equation. Apart from its probability distribution, several characteristics can be measured of the one-dimensional isotropic random walk, such as the mean position

$$\langle x \rangle = \int_{-\infty}^{\infty} xP(x, t)dx = 0 \tag{2.6}$$

and the mean square displacement

$$\langle x^2 \rangle = \int_{-\infty}^{\infty} x^2 P(x, t)dx = 2Dt. \tag{2.7}$$

This description of the one-dimensional isotropic random walk can be extended to higher dimensions. While the diffusion equation in one dimension is given by

$$\frac{\partial P}{\partial t} = D\frac{\partial^2 P}{\partial x^2} \tag{2.8}$$

the same process in r dimensions is described by

$$\frac{\partial P}{\partial t} = D\left(\sum_{i=1}^{r} \frac{\partial^2}{\partial^2 x_i}\right) P. \tag{2.9}$$

The standard solution of equation 2.9 is then given by

$$p(x,t) = \frac{1}{\sqrt{(4\pi D t)^r}} exp\left(-\frac{\sum_{i=1}^{r} x_i^2}{4Dt}\right). \tag{2.10}$$

Figure 2.2 visualises the solution of the diffusion equation for the diffusion coefficients $D = 1$ (A) and $D = 5$ (B) for different time points, t. It is important to note that these solutions are only valid for a large number of time steps, N, because the derivation of equations 2.5 and 2.10 assumes the limit $N \to \infty$ to obtain the normal distribution from the binomial distribution. The same fact was stated by Lord Rayleigh (1880, 1899) as mentioned in the previous section 2.1.1.

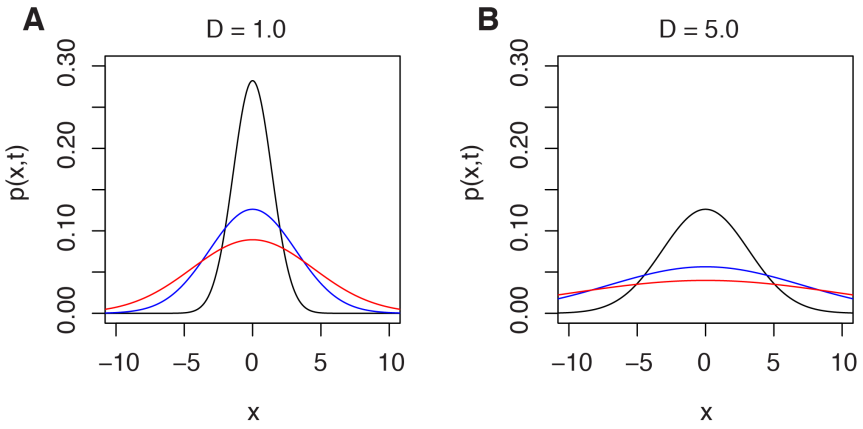


Figure 2.2: Solution of the diffusion equation. Equation 2.10 is plotted for three different time points: $t = 1$ black lines; $t = 5$ blue lines; $t = 10$ red lines. The diffusion coefficient was chosen $D = 1$ (A) and $D = 5$ (B).

2.1.3 Derivation of the diffusion equation

We will now derive the one-dimensional diffusion equation as defined in equation 2.8, as done for example in Strauss (2008) [98]. Imagine we have a liquid in a pipe to which we add a coloured dye. The dye is then diffusing through the liquid from higher to lower concentrations. We describe the concentration of the dye at position x and time t with $P(x,t)$. The total amount of the dye between x_0 and x_1 at time t is then given by

$$A(t) = \int_{x_0}^{x_1} P(x,t)dx. \tag{2.11}$$

We then have

$$\frac{\partial A}{\partial t} = \int_{x_0}^{x_1} \frac{\partial P(x,t)}{\partial t} dx. \tag{2.12}$$

Applying Fick's Law [99], which states

$$\frac{\partial A}{\partial t} = D\frac{\partial P(x_1,t)}{\partial x} - D\frac{\partial P(x_0,t)}{\partial x}, \tag{2.13}$$

with $D > 0$. Therefore combining equation 2.12 and 2.12 results in

$$\frac{\partial A}{\partial t} = \int_{x_0}^{x_1} \frac{\partial P(x,t)}{\partial t} dx = D\frac{\partial P(x_1,t)}{\partial x} - D\frac{\partial P(x_0,t)}{\partial x}. \tag{2.14}$$

The derivative of equation 2.14 with respect to x_1 is then

$$\frac{\partial P(x_1,t)}{\partial t} = D\frac{\partial^2 P(x_1,t)}{\partial x^2} \tag{2.15}$$

which can be written as $\frac{\partial P(x,t)}{\partial t} = D\frac{\partial^2 P(x,t)}{\partial x^2}$ and therefore results in the diffusion equation as presented in equation 2.8.

2.1.4 Biased random walks

Using similar methods as for the derivation of the simple isotropic random walk described in the previous section we can now look at more complicated random walk behaviours. So far we assumed that for a one-dimensional random walk the probability of going left or right is equal to $\frac{1}{2}$. Let us now consider a situation in which the random walker is subject to a constant force, F, which drags him stronger to the right. The movement is then a biased random walk, which is also described as a random walk with drift. In each time step, τ, the random walker is moving (with constant step size, l) to the right with probability q_1, and to the left with probability q_2, where $q_1 + q_2 = 1$. In the case the drift is towards the right we have $q_1 > q_2$. Note that if $q_1 = q_2$ the walk is unbiased and reduces to the isotropic random walk. This process can then be described with the drift-diffusion equation

$$\frac{\partial P}{\partial t} = D\frac{\partial^2 P}{\partial x^2} - v\frac{\partial P}{\partial x}, \tag{2.16}$$

where $v = q_1 - q_2$. The solution of this equation is

$$p(x,t) = \frac{1}{\sqrt{4\pi D t}} e^{-\frac{(x-vt)^2}{4Dt}}. \tag{2.17}$$

For the one dimensional biased random walk mean position is $\langle x \rangle = vt$ and the mean square displacement is $\langle x^2 \rangle = 2Dt + v^2t^2$. A biased random walk in r dimensions can

be described by the partial differential equation

$$\frac{\partial P}{\partial t} = D\left(\sum_{i=1}^{r}\frac{\partial^2}{\partial^2 x_i}\right)P - \nu\left(\sum_{i=1}^{r}\frac{\partial}{\partial x_i}\right)P, \tag{2.18}$$

which has the standard solution

$$p(x,t) = \frac{1}{\sqrt{(4\pi Dt)^r}}exp\left(-\frac{\sum_{i=1}^{r}(x_i-\nu t)^2}{4Dt}\right). \tag{2.19}$$

Equation 2.18 is a special case of the so called Fokker-Planck equation with constant diffusion coefficient D [100].

2.1.5 Persistent random walks

The types of random walk introduced so far can be described as a Markov process, in which the step of the walker at time $t+1$ is independent of the previous step at time t, but only in the position where it is at time t, i.e. the walk is uncorrelated. Other possible types of motion are correlated random walks, where the walker has higher probability of keeping its direction than changing it. This tendency is also described as persistence and the corresponding motion as persistent random walk. Such a process is fully governed by

$$\frac{\partial^2 p}{\partial t^2} + 2\lambda\frac{\partial p}{\partial t} = \nu^2\frac{\partial^2 p}{\partial x^2}, \tag{2.20}$$

where λ is the probability to keep the direction and ν is the velocity. Equation 2.20 is the so called unbiased telegraph equation. In 1951 Goldstein showed that the telegraph equation delivers the full description of the persistent random walk [101], although it was first investigated by Lord Kelvin in the field of signal propagation in the transatlantic telegraph cable. A solution of the telegraph equation in one dimension was provided 1953 by Morse and Feshbach [102]. They showed that given the initial conditions $p(x,0) = \delta(x)$ and $\frac{\partial p}{\partial t}(x,0) = 0$ the solution

$$p(x,t) = \begin{cases} \frac{e^{-\lambda t}}{2}\left\{\delta(x+\nu t) + \delta(x-\nu t) + \frac{\lambda}{\nu}\left(I_0(Z) + \frac{\lambda t}{Z}I_1(Z)\right)\right\} & \text{if } |x| < \nu t \\ 0 & \text{if } |x| \geq \nu t \end{cases} \tag{2.21}$$

satisfies equation 2.20 with $Z = \lambda\sqrt{t^2 - x^2/\nu^2}$ and the modified Bessel function of first kind I_0 and I_1. By applying equation 2.6 to equation 2.20 we obtain the mean position of the persistent random walk, which is $\langle x\rangle = 0$. If we compute the mean square displacement according to equation 2.7 we obtain

$$\frac{\partial^2\langle x^2\rangle}{\partial t^2} + 2\lambda\frac{\partial\langle x^2\rangle}{\partial t^2} = 2\nu^2. \tag{2.22}$$

Under the assumptions $\frac{\partial p(x,0)}{\partial t} = 0$ and $p(x,0) = \delta(x)$ we obtain the initial conditions of equation 2.22 as

$$\langle x^2(0)\rangle = \frac{\partial\langle x^2(0)\rangle}{\partial t} = 0 \tag{2.23}$$

and the solution of the same equation 2.22 results in the mean square displacement

$$\langle x^2(t)\rangle = \frac{v^2}{\lambda}\left(t - \frac{1}{2\lambda}(1 - e^{-2\lambda t})\right). \tag{2.24}$$

One can show that for very small t the mean square displacement is approximately $\langle x^2(t)\rangle \sim v^2t^2$, and for large t it is $\langle x^2(t)\rangle \sim \frac{v^2}{\lambda}t$.

Both the uncorrelated biased random walker and the correlated (persistent) unbiased random walker have at each step a higher probability for a certain direction. However, the persistent random walk is not biased, because the direction does not depend on a global force that drives the walker. It only depends of the previous step and could be seen as a localised bias.

A combination of the latter two described random walks results in a process called biased persistent random walk. This type of motion can be described in one dimension by the biased telegraph equation

$$\frac{\partial^2 p}{\partial t^2} + (\lambda_1 + \lambda_2)\frac{\partial p}{\partial t} + (\lambda_1 - \lambda_2)\frac{\partial p}{\partial t} = v^2\frac{\partial^2 p}{\partial x^2}, \tag{2.25}$$

where λ_1 and λ_2 are the frequencies of keeping the direction and changing the direction, respectively. Note that for $\lambda_1 = \lambda_2$ equation 2.25 simplifies to the equation of the unbiased persistent random walk 2.20.

In chapter 3 and 4 we will apply the random walk theory to investigate leukocyte migration as part of the innate immune response.

2.2 BAYESIAN INFERENCE

2.2.1 History

Bayesian inference is based on Bayes theorem, which was named after Thomas Bayes (1702 - 1761). Thomas Bayes provided the proof of a special case of the theorem. Later Pierre-Simon Laplace (1749 - 1827) proved the general Bayes theorem and applied it in several fields such as statistics and mechanics [103]. There are two major

schools of thought in statistics: the frequentist and the Bayesian approaches [104–106]. In the early 20th Century the frequentist approach was dominating most fields in statistics, and only slowly throughout the century statisticians payed increased attention to Bayesian approaches. In Bayesian inference there are two main interpretations of probability, which divide the community into those favouring objective Bayesian inference and subjective Bayesian inference, respectively [107–109]. In the last 30 years, as new computational resources as well as Markov chain Monte Carlo methods [110] became available, Bayesian inference started to receive increased attention, and the research field of Bayesian inference grew among mathematicians, statisticians as well as physicists [15, 111, 112].

2.2.2 Bayes theorem

Let A and B be two random variables. We want to determine the probability of A given a specific B, i.e. $P(A|B)$. Bayes theorem states:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{2.26}$$

where $P(*|*)$ is the conditional probability, $P(A)$ and $P(B)$ are the probabilities of A and B. In the following section we will show the main concepts of Bayesian inference, which is based on equation 2.26

2.2.3 Model based Bayesian inference

Considering equation 2.26 and replacing A by parameters, Θ, of a given model, B by the evidence in form of collected data D, and the probabilities P by probability density functions p we can write

$$p(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{p(D)}, \tag{2.27}$$

where $p(\Theta)$ is the prior probability distribution, or short prior, of the model parameters before observing the data D, and $p(D|\Theta)$ is the likelihood of observing the data under the model with parameters Θ. We will call the conditional probability density, $p(\Theta|D)$, the posterior distribution. Bayes theorem shows that the posterior distribution is proportional to the likelihood and the prior. $p(D)$ is the probability of the data, which is independent of the model and sometimes referred to as the marginal

likelihood. It can be seen as a constant factor c, where

$$c = \int p(D|\Theta)p(\Theta)d\Theta. \tag{2.28}$$

This constant is important to normalise the posterior distribution so that it integrates to 1. In summary, to obtain the posterior distribution, $p(\Theta|D)$, we have to compute the product of the likelihood and the prior. In the following sections we will explain the main three components of Bayesian inference in more detail and show how we can use approximate methods in cases where the computation of the likelihood is costly or not possible.

2.2.4 The prior distribution

The prior distribution, $p(\Theta)$, (or prior) expresses the amount of uncertainty in the parameter Θ before the data are taken into account. Because in Bayesian inference the prior is multiplied by the likelihood to obtain the posterior distribution, the prior affects the posterior directly (see equation 2.27). Depending on the characteristics of the prior it contains different amounts of information about the parameter Θ. The parameters that describe the actual prior distribution, if an analytic or otherwise convenient form is available, are called hyperparameters. According to the information content of the prior we distinguish between different types of priors: the informative prior and the least informative prior [113].

The informative prior is often used when pre-evidence is taken into account. Previously collected data might guide the settings of the prior for new data. Note that in any case the conventional Bayesian inference scheme requires that the prior is not informed by the data used during the inference [113]. However, a first inference could result in a posterior, which then could be used as an informative prior when new data are collected. Informative priors are often normal distributions with the mean guided by an expert or by previous knowledge [113].

In the case of the least informative prior the aim is to reduce the information content as much as possible to obtain a posterior distribution which mainly depends on the likelihood and therefore on the model and data only. An example of least informative priors are flat priors, where

$$\Theta \sim U(-\infty, +\infty). \tag{2.29}$$

Such a prior is not really informative by itself. A variation would be to define the boundaries of the uniform distribution as finite values. This is often done in biological problems when we know a parameter can not be negative or larger than a certain number. Note that the class of least informative priors used to be considered and referred to as the non-informative prior. The term non-informative prior was however misleading, because the majority of priors contain some information [107].

2.2.5 The likelihood

The likelihood is defined as [114]

$$p(D|\Theta) = \prod_{i=1}^{N} p(D_1|\Theta), \tag{2.30}$$

where N is the amount of collected data D. In order to obtain the posterior distribution, the likelihood must be completely specified. In cases where this might not be possible or computationally very expensive, the likelihood can be approximated as we will explain later in this chapter. The likelihood only depends on the model and the data. Because of that Bayesian inference follows the rule that any two models, which have the same likelihood, will result in the same posterior distribution for the parameter Θ. This is also known as the likelihood principle.

2.2.6 The posterior distribution

The posterior distribution $p(\Theta|D)$ balances the prior information with the information contained in the data via the likelihood. In most real-world applications it is a multivariate distribution. The main advantage of the Bayesian posterior compared to other methods is that the posterior distribution carries more information than a single point estimate. The univariate components, $p(\Theta_i|D)$, called marginal posterior distributions, can be summarised into a single value using e.g. the mean or the median. But the spread, such as the variance, does provide additional information about the confidence of the estimate. The spread of the distribution also contains information about the sensitivity of the system to a given parameter Θ_i: a large spread indicates low sensitivity, while a small spread indicates high sensitivity of the system to the parameter [115].

2.2.7 Numerical approximations

In many systems it is difficult to compute the likelihood [116]. This can have several reasons, for example, because the system is too complex, the data structure is too complex or the system is formulated with stochastic differential equations (SDE) rather than with ordinary differential equations (ODE). In these situations it is still possible to apply Bayesian inference by approximating the likelihood and therefore the posterior distribution. Several so called approximate Bayesian computation (ABC) methods have been developed [117–120]. Here we will introduce the most simple ABC rejection sampler [117] and the more advanced ABC based on a sequential Monte Carlo framework (SMC) [15].

The ABC rejection sampler contains the general principle of all ABC methods. The algorithm is:

```
step 1:  sample θ* from the prior p(θ)
step 2:  simulate the system with θ* to obtain y*
step 3:  if d(D,y*) ≤ ϵ, accept θ otherwise reject
step 4:  return to step 1
```

In this algorithm the calculation of the likelihood is replaced by comparing the collected Data D with simulated data y (step 3). This introduces a tolerance, ϵ, which represents the minimal distance, d, between the collected data, D, and the simulated data, y^*, using parameter Θ^*.

ABC SMC was developed to estimate parameter of dynamical models. Therefore the most common task here is to compare time series data. In this case the Euclidean distance is often used for the distance function d, but in general any metric is possible [111]. So far no detailed studies have been presented to investigate the influence of the chosen distance function on the approximated posterior distribution. Sometimes it might be of use to apply a distance function, which describes the relative deviation from the reference data, instead of the absolute deviation (as it is the case for the euclidean distance). The distance function closely links to the summary statistic of the data. In the case of time series data, we directly use each data point to compute the distance. However, in other fields, for example population genetics, it is necessary to compute a summary statistic. Here, the problem of defining the sufficient summary

statistic appears, which was also discussed in [121].

The resulting distribution from the ABC rejection sampler is a sample from $p(\Theta|d(D, y^*) \leq \epsilon)$ and one can show that if $\epsilon \to 0$ the distribution $p(\Theta|d(D, y^*) \leq \epsilon) \to p(\Theta|D)$. The simple ABC rejection sample can be computationally very inefficient in cases where the posterior differs strongly from the prior, which decreases the acceptance rate. To increase the acceptance rate and therefore the efficiency other ABC based algorithms have been developed: ABC MCMC [119] and ABC SMC [15]. ABC MCMC uses a Markov chain Monte Carlo approach to efficiently sample from the parameter space. We will focus on the ABC SMC scheme, as this algorithm has been used and was further developed throughout this thesis.

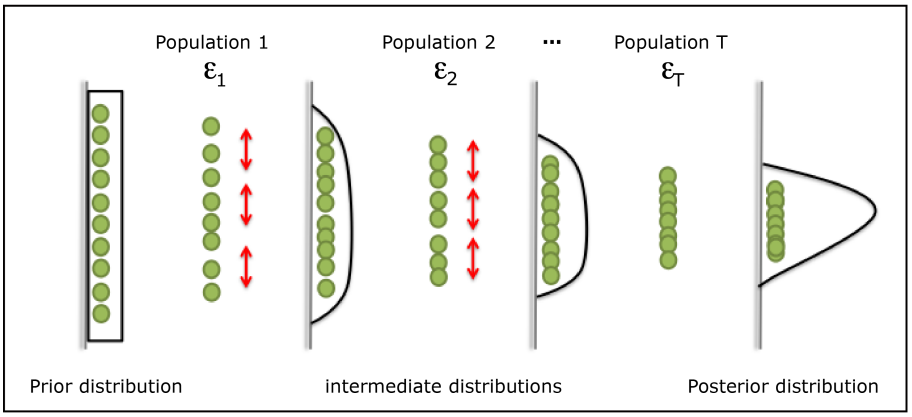


Figure 2.3: Inference scheme ABC SMC. The parameters, also called particles, are initially sampled from a prior distribution and tested wether or not they result into $d(D, y^*) \leq \epsilon_1$. The accepted particles build the first intermediate distribution. Particles are then sampled from the intermediate distribution, perturbed and tested with the next tolerance value. This procedure is repeated until population T, where ϵ_T is small enough to approximate the posterior distribution. (Figure adapted from Secrier *et al.* [122].)

The aim of ABC SMC compared to the simple ABC rejection sampler is to decrease the computational cost. This is achieved by increasing the acceptance rate by introducing a tolerance schedule $\{\epsilon_1, \epsilon_2, ...\epsilon_T\}$, so that at each step the acceptance rate is sufficiently high. This introduces a set of so called populations. In the first population the ABC SMC proceeds like the ABC rejection sampler, i.e. a parameter sample from the prior distribution is drawn; the model is simulated and the output is compared to the collected data using the first tolerance, ϵ_1. In the next population the parameter sample is drawn from the first accepted intermediate distribution, after

perturbing the parameters. This will continue until population T. At each step the accepted distributions get closer to the final approximated posterior distribution (see figure 2.3).

The ABC SMC scheme proceeds as follows [111]:

step 1: define a tolerance schedule $\epsilon_1, ...\epsilon_T$
set $t = 0$
step 2: set $i = 1$
step 3: if $t = 0$: sample θ^* from prior $p(\Theta)$
else: sample θ^* from previous population $\{\Theta_{t-1}^{(i)}\}$
with weights w_{t-1}
perturbation of θ^* results in θ^*_{pert}
if $p(\theta^*_{pert})$=0: return to step 3
else: simulate y^* using $p(\theta^*_{pert})$
if $d(D, y^*) \leq \epsilon_t$: go to step 4
else: return to step 3
step 4: set $\theta_t^{(i)} = \theta^*_{pert}$
calculate the particle weight $w_t^{(i)}$
if $i < N$: set $i = i + 1$ and go to step 3
else: go to step 5
step 5: normalise w_t
if $t < T$: set $t = t + 1$ and go to step 2

The perturbation of the particle Θ^* is done using a perturbation kernel $K_t(\Theta|\Theta^*)$. In Toni *et. al* (2010) [15] the K_t was set to be a random walk, either Gaussian or uniform distributed. The resulting weight of particle $\Theta_t^{(i)}$ is computed as

$$w_t = \begin{cases} 1, & \text{if } t = 0 \\ \frac{p(\Theta_t^{(i)})}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\Theta_{t-1}^{(j)}, \Theta_t^{(i)})} & \text{if } t > 0 \end{cases} \tag{2.31}$$

Further details about how to perturb the sampled parameters and how to compute the weights can be found in [15]. In recent years the perturbation of the particles has been investigated to further increase the efficiency of the ABC SMC algorithm [123].

2.2.8 *Bayesian model comparison*

The Bayesian equivalent of classical frequentist hypothesis testing is Bayesian model selection, which revolves often around the so-called Bayes factors. The Bayes factors where introduced by Harold Jeffreys in 1961 [116]. Using the Bayes theorem in equation 2.26 the posterior model probability $p(M|D)$ of a model M is

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)}, \tag{2.32}$$

where the model M is parameterised with parameters Θ. If we want to compare two models M_1 and M_2 with their parameters Θ_1 and Θ_2 respectively, we can calculate the Bayes factor B as

$$B = \frac{p(D|M_1)}{p(D|M_2)} = \frac{\int p(\Theta_1|M_1)p(D|\Theta_1, M_1)d\Theta_1}{\int p(\Theta_2|M_2)p(D|\Theta_2, M_2)d\Theta_2}. \tag{2.33}$$

A value of $B > 1$ indicates that the data D support M_1 with higher probability than M_2, and vice versa if $B < 1$. It can be difficult to compute Bayes factors, but they can be approximated with the Laplace-Metropolis estimator [124, 125].

The concepts of Bayesian parameter inference and model comparison will be used in chapter 4 and 5 to investigate models of cell migration and other signalling cascades.

2.3 BAYESIAN EXPERIMENTAL DESIGN

2.3.1 *History*

Ever since experiments have been performed the problem of finding the experiment with the highest information arose naturally. The first significant publications on experimental design in a mathematical framework appeared in the early 20th century. In 1918, Smith [126] investigated the problem of optimal design in polynomial

regression. In general, decision theory was very popular during the time of the second world war [127]. In 1943, Wald formulated the problem of experimental design in a mathematical way [128]. Throughout the following years statisticians, like Bose (1948), Karlin and Studden (1966), Stigler (1971) or Kurotschka (1978), developed concepts and methods for optimal experimental design [129–132]. Increasingly these approaches borrowed from information theory, first developed by Shannon in the 1940s. A important theory for optimal design was developed by Kiefer and Wolfowitz [133]. Although their theory had a strong impact on the development of new concepts, only little attention was paid to it in practice. The main criticism was that their concepts assumes the model is absolutely exact. This assumption raised doubts about the robustness of the derived optimal designs. Most optimal design strategies until then aimed to find the best experimental design to estimate the parameters of a system. Only later the concept of assessing the model predictive power was developed.

The first publication about Bayesian experimental design appeared 1956 by Lindley [134]. He measures the information that is provided by an experiment under the consideration of prior knowledge. His measure is based on earlier work by Shannon in 1948 [135]. In the following section we will introduce the concept of Bayesian experimental design and relate it to the work of Shannon and Lindley.

2.3.2 Mathematical formulation

We have a system that we can describe mathematically with parameters Θ. We want to estimate Θ using observed data D resulting from an experiment q. We then want to chose q from a set of possible experiments Q, so that the data under q provide us with the most information about Θ. Using Bayes theorem (equation 2.26) we can write:

$$p(\Theta|D, q) = \frac{p(D|\Theta, q)p(\Theta)}{p(D|q)}, \tag{2.34}$$

where $p(\Theta|D, q)$ is the posterior distribution under experiment q and its data D, $p(D|\Theta, q)$ is the likelihood, $p(\Theta)$ is the prior distribution and $p(D|q)$ is the probability density of the data D under the experiment q. $p(D|q)$ can be evaluated as

$$p(D|q) = \int p(\Theta)p(D|\Theta, q)d\Theta \tag{2.35}$$

and equation 2.34 can be written as

$$p(\Theta|D, q) = \frac{p(D|\Theta, q)p(\Theta)}{\int p(\Theta)p(D|\Theta, q)d\Theta}. \tag{2.36}$$

To determine the experiment q in Q, which provides the most information about Θ we have to define the utility of the experiment $U(q)$:

$$U(q) = \int U(D, q)p(D|q)dD, \tag{2.37}$$

where $U(D, q)$ is a utility function, which contains the cost and gain of performing experiment q with the resulting data D and parameter Θ. This means $U(D, q)$ is a function of the posterior distribution $p(\Theta|D, q)$, which is obtained from collected data D of the experiment q. The aim is now to maximise the expected utility over all experiments q in Q.

2.3.3 The expected utility function

The expected utility function can be defined as the information gain from the prior distribution to the posterior distribution. In the case of parameter estimation, or estimation of functions of these parameter, the Shannon entropy can be used to asses the information gain [135]. The Shannon entropy H is defined for a discrete random variable X as

$$H(X) = E_X(I(X)) = E(-\ln(p(X))), \tag{2.38}$$

where E_X is the expected value of X, I is the information content of X and $p(X)$ is the probability mass function of X. In the continuous case equation 2.38 can be extended to

$$H(X) = -\int_{-\infty}^{+\infty} p(x)\ln(p(x))dx. \tag{2.39}$$

The utility function is then

$$U(D, q) = -H(p(\Theta|D, q))+H(p(\Theta)) = \int \ln(p(\Theta|D, q))p(\Theta|D, q)d\Theta - \int \ln(p(\Theta))p(\Theta)d\Theta \tag{2.40}$$

The equivalent to computing the change in Shannon entropy from prior to posterior distribution is the Kullback-Leibler divergence [136] D_{KL} as

$$U(D, q) = D_{KL}(p(\Theta|D, q)||p(\Theta|q)). \tag{2.41}$$

For continuous distributions A and B the Kullback-Leibler is defined as

$$D_{KL}(A||B) = \int_{-\infty}^{\infty} \ln\left(\frac{a(x)}{b(x)}\right) a(x)dx, \tag{2.42}$$

with a and b being the probability densities of A and B [136]. Lindley showed that by using equation 2.40 in equation 2.37 the expected utility of an experiment $U(q)$ can be written as

$$U(q) = \int\int \ln(p(\Theta|D,q))p(\Theta,D|q)d\Theta dD - \int \ln(p(\Theta))p(\Theta)d\Theta. \tag{2.43}$$

To determine the optimal experiment it is necessary to maximise $U(q)$, i.e. to maximise the right hand-site of equation 2.43. The second term, however, only depends on the prior distribution but not on the experiment q and therefore does not need to be computed to select the optimal experiment.

To calculate such a utility function it is often necessary to solve high dimensional integrals and optimisation problems. For linear models it might be possible to solve the integrals analytically, but nonlinear models require approximations of the integrals. Examples of such approximations are Monte Carlo integration, Laplace integration or numerical quadrature [16, 137].

Other utility functions, such as quadratic loss functions and asymmetric loss functions, have been explored and applied to various design problems [16, 137]. The choice of the utility function always depends on the proposed design problem.

2.3.4 Mutual information

The above described concept can be directly linked to the concept of mutual information. For two continuous random variables x and y the mutual information is defined as

$$I(X,Y) = \int_Y \int_X \ln\left(\frac{p(x,y)}{p(x)p(y)}\right) p(x,y)dxdy. \tag{2.44}$$

It results that the mutual information between the parameters Θ and the collected data D is

$$I(\Theta,D) = \int\int \ln\left(\frac{p(\Theta,D)}{p(\Theta)p(D)}\right) p(\Theta,D)d\Theta dD. \tag{2.45}$$

Using

$$p(\Theta,D) = p(\Theta|D)p(D), \tag{2.46}$$

which can be written as

$$p(\Theta|D) = \frac{p(\Theta,D)}{p(D)}, \tag{2.47}$$

we obtain from equation 2.45

$$I(\Theta,D) = \int\int \ln\left(\frac{p(\Theta|D)}{p(\Theta)}\right) p(\Theta,D)d\Theta dD. \tag{2.48}$$

Separation of variables results in

$$I(\Theta, D) = \int\int \ln\left(p(\Theta|D)\right) p(\Theta, D) d\Theta dD - \int\int \ln\left(p(\Theta)\right) p(\Theta, D) d\Theta dD, \tag{2.49}$$

which is

$$I(\Theta, D) = \int\int \ln\left(p(\Theta|D)\right) p(\Theta, D) d\Theta dD - \int \ln\left(p(\Theta)\right) p(\Theta) d\Theta = U(q). \tag{2.50}$$

This shows that the expected utility of an experiment q is equal to the mutual information between Θ and the collected data D.

In chapter 5 we will meld the concepts of Bayesian inference and information theory to develop a framework of experimental design for complex biological models.

Analysis of Biological Trajectory Data

Parts of this chapter have been published in [138].

3.1 INTRODUCTION

In this chapter we examine the behaviour of leukocyte dynamics in zebrafish embryos in response to injury. The zebrafish, which has long been an important model system in developmental biology, has also become an attractive model in which to study inflammation and the immune system. Analyses in zebrafish allow *in vivo* imaging of immune processes to be combined with molecular studies that target signalling processes regulating leukocyte migration.

The innate immune system of zebrafish closely resembles that of mammals and is fully competent at early embryological stages before the emergence of lymphocytes. For the first few weeks of their life zebrafish embryos rely solely on their innate immune system as the adaptive system becomes functional four weeks after fertilisation. Here we focus on the spatio-temporal response of myeloid cells in zebrafish following surgical injury to the tail fin. Several studies have demonstrated that injury in zebrafish embryos results in the migration of leukocytes to the site of tissue damage [52–56]. Although the migration is dependent in part on a hydrogen peroxide (H_2O_2)

gradient produced at the side of injury [57], clearly other signals also contribute to the decision making that results in cell migration.

What becomes apparent from these studies of leukocyte recruitment in zebrafish embryos is that cells exhibit a panoply of different types of migratory behaviours. These behaviours will be influenced by the time since and distance from the wound site. Here our aim is to capture and rationalize this richness in immune cell chemotaxis. The simple statistics, such as the number of recruited cells, the velocity, the mean square displacement or the straightness index, that are often used to analyse these trajectory data do not capture the whole information content of such rich data [139–142].

Random walks have been used to model animal movement and cell migration [92]. They are often described as uncorrelated random walks with diffusion [93], or as Levy flights [143, 144], which are isotropic random walks with characteristic distributions of the step length (e.g. Brownian motion vs. Levy flights). Another possibility is to model the change in direction rather than considering the step length, which leads to the analysis of isotropic vs. non-isotropic random walks. In this context it was recently reported that living mammary epithelial cells in a tissue display a bimodal persistent random walk [145]. Here we use automatic image analysis to capture and analyse a sufficiently large number of leukocyte trajectories in wounded zebrafish embryos to obtain reliable statistical interpretations of the leukocyte recruitment and migration under different conditions.

We investigate the the influence of the MAPK pathway on leukocyte migration. The MAPK (Mitogen-activated protein kinases) pathway is a sequence of kinases that transports a signal from a cell surface receptor to the nucleus of the cell by phosphorylation and dephosphorylation processes. In the nucleus it activates the transcription of specific genes and the translation of its proteins, which are related to processes such as stress response, cell division, cell growth and apoptosis. Apart from the MAPK's MAP1K, MAP2K and MAP3K, the components of the MAPK pathway include ERK, p38 and JNK.

The discovery of selective ATP-competitive inhibitors made it possible to dissect the individual roles of the JNK and p38 MAPK families. The anthrapyrazolone SP600125 is now widely used as an inhibitor of JNK signalling [145] and SB203580, a pyridinyl imidazole, is commonly used to inhibit p38 MAPK dependent signalling

[146]. These inhibitors are useful tools to study the function of these protein kinases in cell signalling and other physiological processes. For instance, it has recently been shown, that the JNK inhibitor SP600125, but not the p38 inhibitor SB203580 plays an important role in the recruitment of tissue-resident primitive macrophages to the site of acute injury induced by tail transection [71].

To investigate the diverse dynamics of leukocyte migration we apply transition matrices as a novel statistical approach to analyse *in vivo* trajectories of migrating cells. A transition matrix is used to describe the transition, in this case cell movement, from one state to another. Transition matrices have been previously used to define and model different types of random walks [147]. Here we use transition matrices as a data analysis tool to analyse leukocyte migration data produced in zebrafish injured and treated with pharmacological inhibitors of signalling proteins. This allows us to study how different molecular components can modulate the immune response by influencing the migratory behaviour of leukocytes. In addition to the analysis described, we show that migration behaviours are dependent on space and time. Our approach can be applied to analyse any kind of biological trajectories. We finally use the same approach to investigate differences between the migration patterns of macrophages and neutrophils using transgenic zebrafish with the *mpo* marker for heterophil granulocytes (here mainly neutrophills) and the *fms* marker for macrophages.

3.2 METHODS

3.2.1 Zebrafish care and breeding

Tg(-9.0spi1:EGFP)zdf11 (pu.1:EGFP) zebrafish [148] were bred and maintained according to the Animals (Scientific Procedures) Act 1986.

3.2.2 Tail transection and image acquisition

All experiments used in this chapter were performed by the group of Prof. M. Dallman (Imperial College London). pu.1:EGFP zebrafish embryos (5 dpf) were pre-treated in system water only (untreated) or system water containing either 0.002 % (v/v) Dimethyl Sulfoxide (DMSO) (vehicle control), 20 μM SP600125 or 10 μM SB203580, both dissolved in DMSO (Sigma-Aldrich), for two hours at 28.5 C. After two hours, they were anaesthetised in 0.6 mM MS-222 (Tricaine methanesulfonate, from Sigma-Aldrich) and the tail fin was transacted using a sterile scalpel. The fish were then transferred to fresh treatment media for 2 hours 28.5C before transferral to 0.8% low melt agarose (Flowgen, Lichfield, UK) for time lapse imaging experiments. Images were captured using a Zeiss Axiovert 200 inverted microscope controlled by the C-Imaging Simple-PCI acquisition software for up to 14 hpw. The temperature was maintained at 28.5C throughout the experiment using a full incubation chamber with temperature control. The time gap between two consecutive images was 18 seconds.

3.2.3 Image processing and data transformation

Imaging resulted in image stacks with dark background and fluorescent pu.1:EGFP++ cells. The image processing was done in R using the package EBImage [149]. An edge detection method was used to automatically extract the information of the cells from the images. We used a manually set threshold of the light intensity per image stack. Each detected cell was described as an object with the coordinates of its geometrical centre describing the cell location and the occurrence time (figure 3.1A). A surface algorithm was used to track the cells over time, which is based on the shortest distance between cells from two consecutive images. The algorithm calculates the distance between a chosen detected cell and all remaining cells at the next time point. The two cells with the smallest distance are connected, *i. e.* they are part of the same trajectory. Our time-lapse microscopy data were optimised in the experimental setup, e.g. 18 sec time gap between two consecutive images, so that the cell area from one time point to the next one overlapped, in order to reduce the typical tracking errors described in [75]. When two cells overlapped in the same image (due to 2D data) we stopped tracking them to avoid incorrect cell paths. We excluded all cell trajectories that included time points in which the cell was located at the edge of the

image. The images contain the zebrafish tail with the whole injury and only those trajectories with a distance to the injury of less than 650 µm were included. Only trajectories that contain more than 40 time steps have been used for the analysis to improve the reliability of our results. Datasets in which tissue deformation occurred during acquisition were excluded from analysis. However, no minimum path length was required as long as the cell was tracked over minimum 40 time steps. In this way it is also possible to observe possible resting cells. We produced bright field and fluorescent images at each time point. Comparing/overlapping two consecutive bright field images allowed us to detect shift and rotation due to small movements of the zebrafish, which was used to correct the absolute position of tracked cells. An overview about the analysed data is given in figure 3.1. Although the zebrafish tail has a depth of a few cell layers we performed our analysis in 2D for 2 reasons: (i) the majority of leukocytes that respond to injury in this tissue region can be imaged in one focal plane (at 10x magnification) if the fin tissue is mounted flush with the plate (cells not in focus were excluded from analysis), (ii) the acquisition of 3D data leads to a longer time gap between two consecutive images, which results in more tracking errors and less information about the migration process. To analyse the extracted image data more efficiently and to combine or compare data from several movies it was necessary to normalize them, e.g. the reorientation of the object positions in respect to the notochord of the fish. This was achieved by using linear transformation. The transformation describes the rotation and shifting of the new coordinate bases in the way that the blood flow describes the y-axis and orthogonal to it the x-axis, which was located approximately parallel to the injury (figure 3.1A and B). The orientation was based on the bright field images. For the analysis only zebrafish embryos with the wound approximately orthogonal to the notochord were included. Because the embryos were injured manually we accepted small deviations and assumed them to be orthogonal.

3.2.4 Statistical analysis and random walk models

The detailed description of the random walk models and their analysis is present in section 3.3.3. Simulation of the sample paths from the models was done in R. The extracted leukocyte trajectories were split into subtrajectories of 20 time steps.

Table 3.1: Extracted data

exp. condition	number of movies	number of cell tracks (length 20 steps)
untreated	12	3265
+ DMSO	7	1427
+SB203580	7	1244
+SP600125	10	2122

List of extracted data, which were used in this study. The cells were automatically detected and tracked from time-lapes fluorescent microscopy movies. The tracked trajectories were split into subtrajectories of length 20 steps.

All analysis was performed on the subtrajectories, to avoid effects due to different trajectory length. The velocity of the trajectories has been computed using local polynomial regression. The straightness index S_D was calculated as the coefficient of the shortest distance between the start and end point of a trajectory and the actual length of the trajectory. A straightness index close to 1 indicates movement along a straight line. Note that a straightness index close to zero does not necessarily imply that the cell performs a random walk as later described in equations 3.1-3.4. We defined the correlation time τ as the time when the average autocorrelation function (over all trajectories) of β (angle between a motion vector and the negative y-axis) is zero. We compute the autocorrelation function for each cell trajectory in *R* (function acf from package *nlme*) and average over all cells that belong to the same analysis group. We use local linear regression to estimate when the average autocorrelation function reaches zero. Bars are the 5% and 95% bootstrap confidence interval of the mean. All statistics and graphics were generated in *R*. The analysis was here performed in 2D. Simple statistics such as the straightness index and the velocity can be computed analogously in 3D. To investigate the random walk process, some adaptations are necessary. While the mathematical extensions to 3D are straightforward by using spherical coordinates, the visualisation of the results would lack the intuitive appeal compared to 2D. The data were clustered depending on time after wounding (T1-T4) and distance from the wound (S1-S4). The analysis was repeated with shifted intervals to test for independence of the clustering scheme. An initial analysis showed that the cell movement does not vary along the x-direction (parallel to the wound).

3.3 RESULTS

3.3.1 Cell tracking and the acquisition of trajectory information

Developments in the field of live imaging of single cell migration enable us to observe cellular processes and their temporal evolution in unprecedented detail. It is now possible to image the rich diversity of cellular dynamics inside living organisms.

We previously developed an automated cell tracking system in live zebrafish embryos to analyse leukocyte recruitment at the single cell level from trajectory data produced by time-lapse imaging. Details about the tracking algorithm are provided in section 3.2.3. The data acquisition protocol was developed during an MSc project and is summarised in figure 3.1. Trajectories were analysed mathematically to produce detailed information about leukocyte migration. Time lapse imaging of pu.1:EGFP transgenic zebrafish embryos was performed to record the recruitment of pu.1:EGFP+ leukocytes to an injury produced by tail transection (figure 3.1A). The pu.1 promoter is a marker for mainly neutrophils, but also macrophages, *i.e.* we observe a mixed cell population of neutrophils and macrophages. The data acquired were processed and normalised using the automated cell tracking system to produce information on the trajectories of individual cells migrating in response to injury.

Cell trajectory data are usually noisy, discretised and error prone, which needs to be taken into account when analysing them. The errors often result during the data acquisition and data processing stage. Our protocol took into account typical tracking errors, recently reviewed by Beltman and colleagues [75], and controlled for their effects, resulting in reliable trajectory data. The automated cell tracking system acquired cell shape and cell movement information that allowed the generation of image sequences documenting change in cell shape and trajectory over time (figure 3.1B). A cell trajectory is then represented as a sequence of coordinates, here 2 dimensional $\mathtt{x}$ and $\mathtt{y}$, over time.

3.3.2 General analysis of cell migration data

Biological trajectory data have been collected for many years, mainly investigating animal movement. However, surprisingly little work has been done to analyse these data. We next describe the typical analysis parameters and show their advantages as well as limits to capture the whole trajectory information. We apply them to our

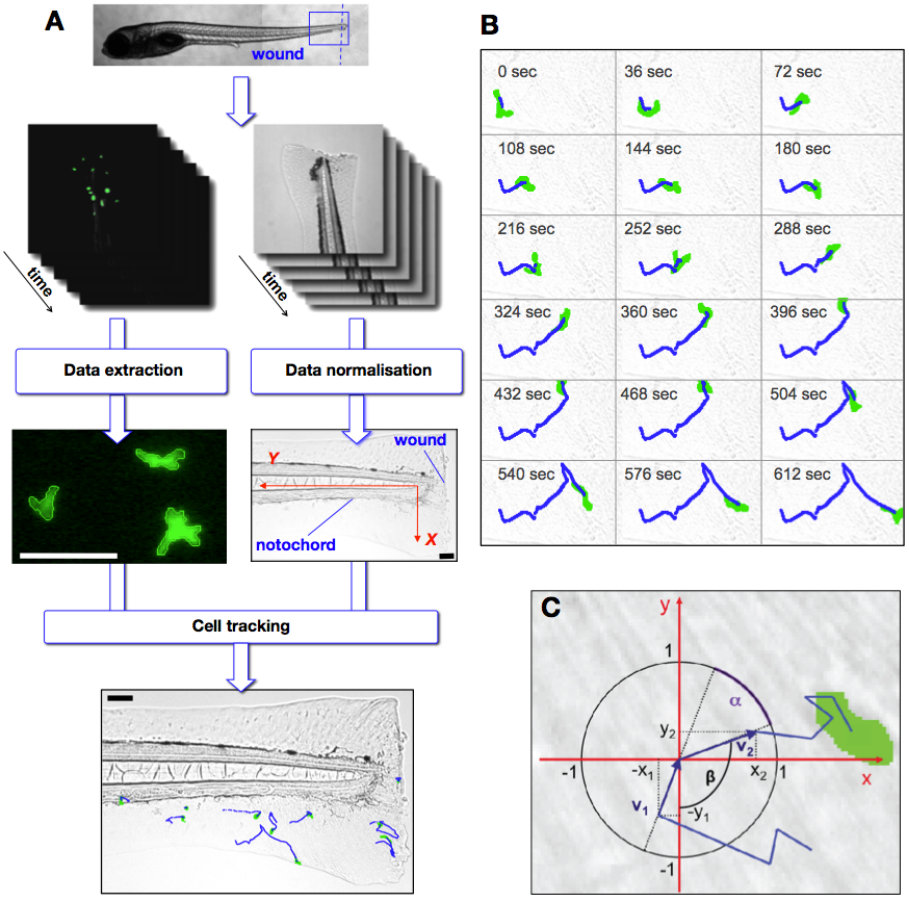


Figure 3.1: The automated leukocyte tracking system. pu.1:EGFP zebrafish embryos were injured via tail transection (blue dashed line). The blue-framed region was captured using time-lapse fluorescent microscopy resulting in image sequences with green fluorescent pu.1:EGFP positive cells. In addition, bright field images were generated to normalise the data by linear transformation of the trajectory data into the new axes shown in red. Resulting cell tracks are shown in blue. Scale bars are 100 µm (A). Time-lapse bright field images overlaid with images of single pu.1:EGFP positive cell automatically detected (green cell) and tracked (blue trajectory line) (B). A trajectory (blue) of a pu.1:EGFP positive cell (green) that was tested for random walk characteristics. Two motion vectors v1 and v2 (dark blue) with their projections onto the x-axis and the y-axis (x1, y1 and x2, y2) were used to test for isotropy which is achieved by calculating the angle α between v1 and v2. If the BM random walk model holds, the one-dimensional projections of the motion vectors onto the axes are Gaussian distributed (C).

extracted leukocyte trajectory data during acute wounding. Simple statistics can be directly computed from the cell coordinates over time such as the velocity and the straightness index of a cell. The distance and direction in which a cell moves between frames, i.e. the motion vector, was determined and the angle α between consecutive motion vectors was calculated (figure 3.1C). We can now compute the velocity and straightness index of a cell as described in the method section 3.2. Both parameters give us an idea about the spread of cells and can provide an expectation how far a cell is moving from a starting point in a certain time. The straightness index provides first insights weather a cell moves randomly or not.

The inhibition of MAPK proteins, known to play a role in leukocyte migration, altered migration behaviour as determined by our new statistical approach. Transgenic pu.1:EGFP zebrafish embryos were used to acquire the *in vivo* experimental data we presented 19. PU.1 is myeloid cell selective allowing investigation of the migratory behaviour of a heterogeneous population of myelomonocytes.

MAPK pathways are known to play important roles in leukocyte migration [69]. JNK but not p38 MAPK has been shown to influence the number of macrophages and neutrophils recruited to an injury in zebrafish [71]. However, the role MAPKs play in modulating leukocyte migration dynamics is poorly understood. We compared different characteristics of cell migration trajectories extracted from zebrafish embryos treated with the p38 MAPK inhibitor SB203580, the JNK inhibitor SP600125 (both soluble in DMSO) with DMSO control treated embryos. We also acquired recruitment data from untreated embryos to determine the effect of DMSO alone on leukocyte migration, as DMSO is known to have modulatory effects on inflammatory processes. We analysed trajectory information produced by two common methods of quantifying cell migration behaviour. Of the treatment groups analysed only the p38 MAPK inhibitor SB203580 had a significant effect, an increase in velocity and straightness index when compared to untreated fish (figure 3.2). Velocity and straightness index contain only little information about the actual dynamics of a cell and are because of that not a sufficient analysis tool to detect differences in treatment groups.

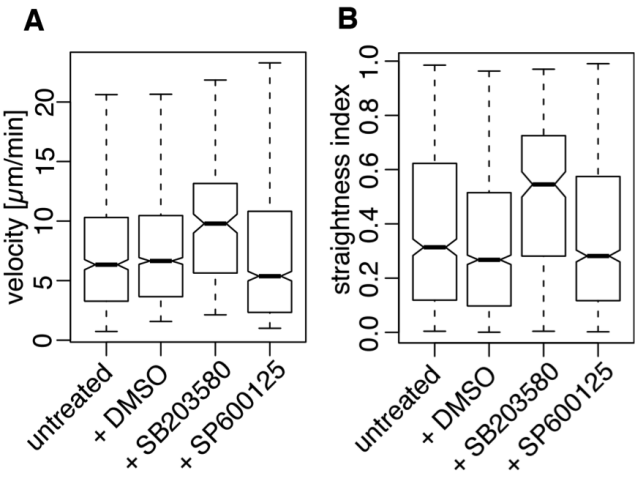


Figure 3.2: P38 MAPK and JNK influence the velocity and the straightness of leukocytes. Boxplots representing the distribution of the velocity (A) and the straightness index (B) of individual leukocyte trajectories for the 4 treatment groups (number of trajectories: 497 untreated, 727 +DMSO, 581 +SB203580, 690 +SP600125) are shown. All cells were detected between 3 and 7 hpw. Dotted lines represent the 5% and 95 percent interval, boxes represent the 25% and 95% interval, bold lines are the medians and notches represent the confidence interval of the median.

3.3.3 Random walks in cell biology

We repeat the key formulae of random walks that were already discussed in section 2.1. Random walks have been used to model animal movement and cell migration [92]. They are often described as uncorrelated random walks with diffusion [93] or Levy flights [143]. Recently it was reported that living mammary epithelial cells in a tissue display a bimodal persistent random walk [144]. To analyse the leukocyte trajectories we considered 4 different random walks [92, 150]. The simple Brownian motion (BM) defined as

$$\frac{\partial P(x,y,t)}{\partial t} = D\nabla^2 P \tag{3.1}$$

with the vector D of the diffusion coefficients, and the probability density function $P(x,y,t)$ of the object location at time t. Considering the possibility of drift, equation 3.1 becomes a biased random walk (BRW)

$$\frac{\partial P(x,y,t)}{\partial t} = -v\nabla P + D\nabla^2 P \tag{3.2}$$

where v describes the expected velocity of the drift. These two types are uncorrelated random walks. However, it is described that animal and cell trajectories often show

correlated random walks. This so-called persistence random walk (PRW) is defined through the Telegraph equation

$$\frac{\partial^2 P(x,y,t)}{\partial t^2} + 2\lambda \frac{\partial P}{\partial t} = D^2 \nabla^2 P \tag{3.3}$$

where λ is the rate of turning events. In case a certain direction is favoured additionally to the persistence, an observed trajectory would be a sample path of the biased Telegraph equation, here referred to as a biased persistence random walk (BPRW)

$$\frac{\partial^2 P(x,y,t)}{\partial t^2} + (\lambda_1 + \lambda_2)\frac{\partial P}{\partial t} + (\lambda_1 - \lambda_2)\frac{\partial P}{\partial y} = D^2 \nabla^2 P. \tag{3.4}$$

The equations 3.1 - 3.4 are extensions of the one dimensional random walk models introduced in chapter 2.1. They describe the probability densities of the cell populations over space and time, i.e. they summarise the walk. They do not explicitly capture the movement of a single cell over time, also called the sample path. The specific sample paths can be obtained by numerical simulations (figure 3.3). We performed all analysis and modelling work based on such sample paths because they closest resemble the characteristics of biological cell trajectories.

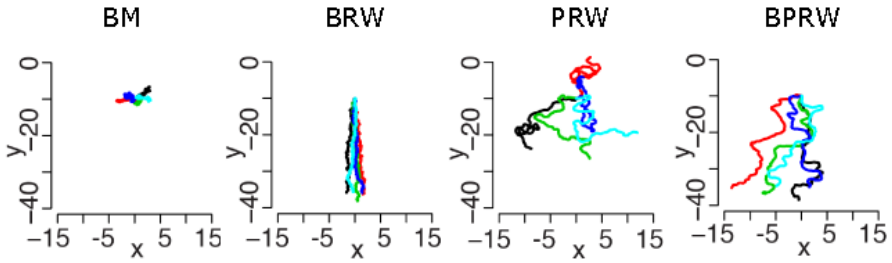


Figure 3.3: Sample paths of random walk models. Simulated trajectories of the 4 described random walk models: BM Brownian motion, BRW biased random walk, PRW persistence random walk, BPRW biased persistence random walk. Initial conditions for numeric simulation: x = 0, y = -10.

When we want to investigate whether or not a cell performs a certain type of random walk, we need to consider that usually cell trajectory data are discretised with a given constant time step Δt. This leads to a scaling problem, which we will further consider in the next section.

3.3.4 *Analysis of cell migration dynamics using transition matrices*

Leukocyte migration dynamics exhibit diverse types of behaviour that are affected by a multitude of factors. These patterns are only poorly described by simple statistics; computing the average straightness index or velocity across a whole population of cells can only give coarse insights into cell migration behaviour and how this differs as conditions are changed. Here we aim to capture more subtle and nuanced changes in the migration behaviour over time and space. We therefore developed the application of transition matrices, which can provide more information and are better suited to elucidating mechanisms and effects of leukocyte behaviour than simple statistics. We furthermore consider how these transition matrices differed at a range of distances from the wound site and at a range of times after injury.

Leukocyte migration in the presence and absence of a signal has been described as a random walk [46, 91, 151–153] and we considered the four different random walk processes described in the previous section. We investigate if and how the migration deviates from a isotropic random walk, where all directions of movement are equally favoured. Our models do not consider the distribution of the step length, because the distribution of the step length is not a characteristic that differs between the investigated types of random walk. However, the step length could be used to distinguish between for example Brownian motion and Levy flights, where the distribution of the step length differs.

We transform all the trajectories so that the zebrafish notochord defines the y-direction and the x-direction is parallel to the injury (see methods 3.2) and therefore the y-axis was used as a reference to determine if a cell moves towards or away from the injury, i.e. the cell movement is biased in a certain direction and is therefore directed.

Next we investigate which dynamical model dominated under the different treatment groups in the real data. The investigated dynamical models are continuous-time stochastic processes $\{\alpha_t\}$, where α_t is the angle between a motion vector and the negative y-axis at time t (figure 3.1). Because we observe data only every 18 seconds, we need to discretize the stochastic process $\{\alpha_t\}$ by sampling intervals of length $l = 18$ seconds and therefore obtain $\{\beta_t\}$, where $\beta_t = \alpha_{kl}$. We can now analyse the process $\gamma_k = i$ if $\beta_k \in [-\pi + 2\pi/15(i-1), -\pi + 2\pi/15i]$ for $i \in [1, ..., 15]$ by computing the

probability matrix of transitions from step t to $t+1$ as $T_{i,j} = P(\gamma_t = i, \gamma_{t+1} = j)$, where $T_{i,j}$ is the (i,j)-th entry in T . A β close to 0 describes a movement towards the injury, $\beta > 0$ and $\beta < 0$ describes an angle to the right and left side, respectively, and β close to $\pm\pi$ (180 degrees) describes movement away from the injury. The schematic in figure 3.4A shows some of the possible transitions using arrows to indicate motion vectors.

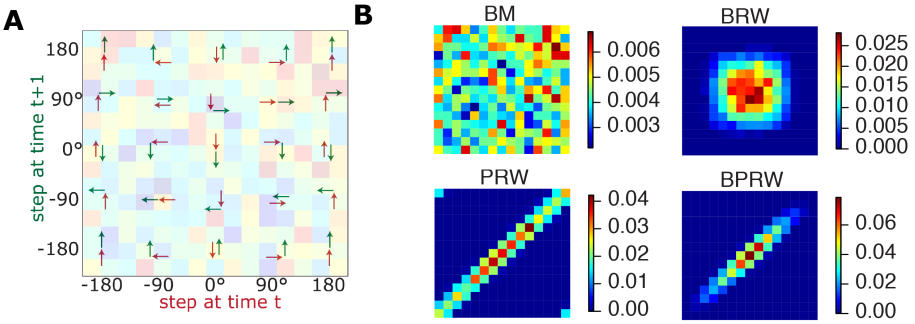


Figure 3.4: Transition matrix as a tool to capture complex dynamics in cell migration behaviours. A key to indicate the cell migration transitions captured by the transition matrix (A). Red arrows indicate the first step, followed by green arrows representing the consecutive step. The angles provide the absolute orientation in the fish, where the negative y-axis (notochord) is used as a reference (A). Sample paths of the 4 described random walk models: BM Brownian motion, BRW biased random walk, PRW persistence random walk, BPRW biased persistence random walk. Initial conditions for numeric simulation: $x = 0$, $y = -10$ (B). Probability matrices for transitions of β for the 4 random walk models plotted as heat maps (blue lowest probability, red highest probability). Matrices are computed from 100 trajectories over 50 time units. The matrices show clearly distinctive patterns and can therefore be used to distinguish between the different random walk types (C).

This approach distinguishes sets of migration patterns that are based on the transition from a given state, in this case the angle between a reference axis (notochord) and the leukocyte step (direction), into another (figure 3.1A). We first used simulations to determine the nature of transition matrices for different types of migration behaviour (see methods 3.2).

Figure 3.4A is a schematic that illustrates the location on the matrix of a representative selection of the step transitions that were captured by the transition matrices. We used Monte Carlo simulations to generate trajectories for each of the four types of random walk described above (figure 3.4B) to generate probability matrices for the four random walk models. We sample a sequence of β_t from a circular normal

distribution with mean μ (π for BRW, β_{t-1} for PRW) and variance σ (0.5 for BRW, PRW and BPRW). We generated 100 sample paths of 20 steps for each random walk model, to capture the characteristics of the extracted cell migration data, which are also limited in number and length. We could clearly distinguish the different patterns produced by the different types of random walks from the resulting transition matrices (figure 3.4C).

Simulated trajectories for Brownian motion produced transition matrices where all transitions have equally high probability, i.e. no overall pattern is discernible. By contrast, simulated trajectories for a biased random walk generated transition matrices with higher probability for the transitions in the centre of the matrix, representing the bias towards the injury, in this case. Trajectories simulated using a persistent random walk produced transition matrices with high probabilities of transitions along the diagonal, where consecutive steps at time t and $t+1$ have very similar directions, indicating persistence. Trajectories from a biased persistent random walk produced transition matrices with high probabilities of transitions along the diagonal with highest probability in the centre of the matrix (figure 3.4C). We used these patterns as dictionaries to compare the patterns produced in transition matrices generated from real experimentally acquired trajectory data to determine the types of walk demonstrated by real leukocytes over time and space and under different treatment conditions.

The investigated models of random walks are stationary processes, i.e. their characteristics do not change in time or space. We next investigate whether or not the observed migration process in zebrafish embryos is stationary as well, in which case the properties of the transition matrices do not change over time or space. It is important to investigate if stationarity is given, not only to gain insights into the biological processes, but also to understand if simple summary statistics are applicable or not.

3.3.5 Temporal dependence of leukocyte migration dynamics

We applied this transition matrix analysis to the dynamics of leukocyte movement. Signals sensed by each leukocyte will change over time depending on the balance of pro-inflammatory and anti-inflammatory/pro-resolution mediators [154, 155]. We

might therefore have expected the dynamics of leukocyte migration behaviour to change over time and we investigated this using experimentally extracted trajectory information from live-imaging data to produce transition matrices. We grouped extracted trajectories taken at a distance of 0 to 155 µm from the wound from 3 hpw (hours post wounding) to 14 hpw into four equally distributed intervals (T1, $3.00 - 5.72$ hpw, T2, $5.72 - 8.44$ hpw, T3, $8.44 - 11.16$ hpw and T4, $11.16 - 14.00$ hpw, figure 3.5A) and computed the corresponding transition matrices for each set of trajectories (the grouping was used to yield roughly equivalent statistical power across all time-windows; the overall picture emerging from this analysis is, however, robust to varying the time-windows). Leukocytes from untreated zebrafish embryos showed a persistent random walk (PRW), as demonstrated by the high probability of transitions along the diagonal (figure 3.5). Over time this high probability along the diagonal is reduced, indicating that the migration type is a non-stationary process, where the level of persistence decreased with time.

Another measure of the level of persistence is the correlation time of a trajectory (for details see chapter 3.2.4). This is a measure of how long it takes until a cell changes its direction. To compute the correlation time, τ, the autocorrelation function of β (the angle between a motion vector and the negative y-axis) was computed. We define the time until this function reaches zero (no correlation) as the correlation time. Figure 3.5B shows the correlation time per time interval (T1-T4) after injury. In an untreated zebrafish the correlation time decreased from 60 sec at T1 to 18 sec at T4. This is in line with the reduction in persistence demonstrated over time by the transition matrices.

The transition matrices for fish that underwent DMSO treatment showed weak persistence, i.e. somewhat lower probability along the diagonal, in comparison to untreated fish (figure 3.5A, 2nd row). A higher probability for transitions in the centre of the matrix was observed, which showed a bias in the leukocyte movement towards the injury site. The pattern of the transition matrices did not change significantly over time, meaning that the temporal dependence of the leukocyte behaviour was ablated in the presence of DMSO. The correlation time for DMSO was also lower than in untreated, and did not decrease over time (figure 3.5B, 2nd row). Treatment with the p38 MAPK inhibitor SB203580 (dissolved in DMSO) restored the persistence and the decrease in correlation time. In fact, it increased the level of persistence

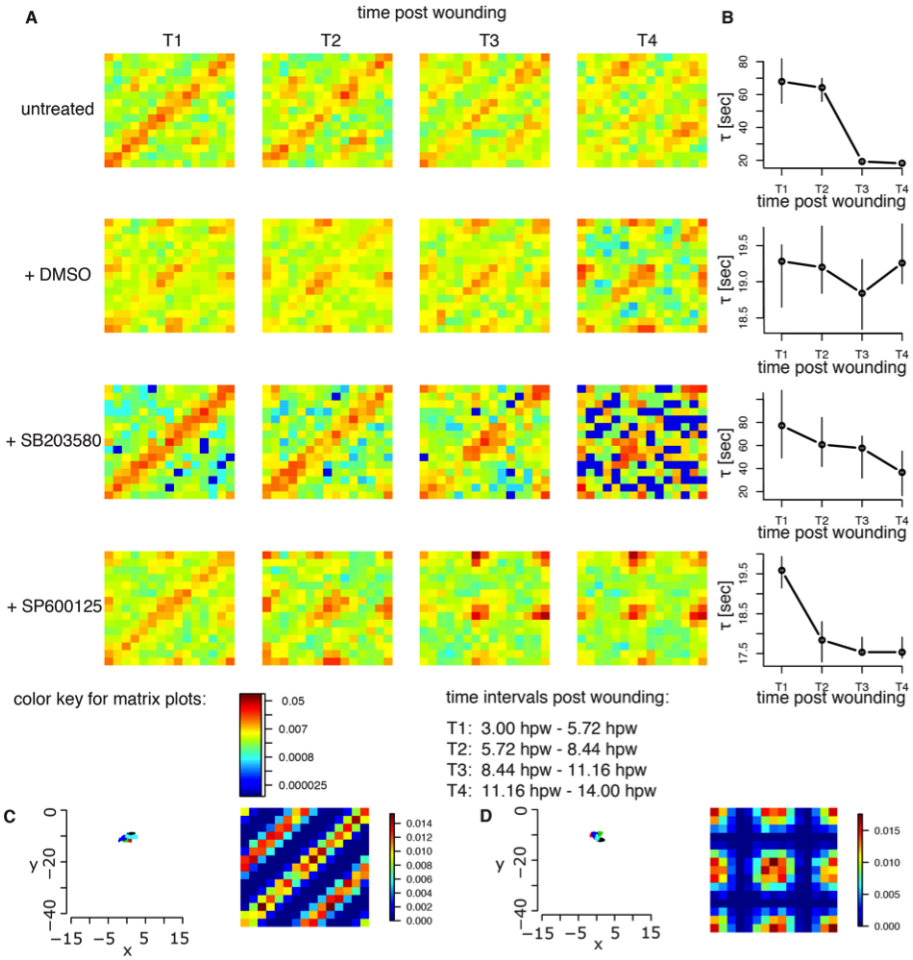


Figure 3.5: Leukocyte dynamics change with time after injury. Transition matrices as heat maps for the 4 treatment groups are presented. Leukocyte trajectories detected at the injury site (distance from injury between 0 and 300 µm) were divided into four time intervals post injury T1 - T4 (see legend) and transition matrices plotted for each (A). We compute the average correlation time at each time interval (circles) with its bootstrap confidence interval of the mean (vertical lines). Note that the scales differ in between the treatment groups (B). To explain the unexpected dynamical patterns that appear in some of the transition matrix, we formulated 2 models, forward-backward random walk (FBRW) (C) and trafficking (D), to numerically simulate trajectories and compute their transition matrices for comparison. Initial conditions for numeric simulation: x = 0, y =-10.

compared to the untreated condition. Inhibiting p38 restored the temporal dependencies. Compared to untreated trajectories we observed a bias towards the injury site at later time points in these trajectories. Since this pattern is also present in the

DMSO control, this is likely a DMSO effect.

Leukocytes exposed to the JNK inhibitor SP600125 (dissolved in DMSO) exhibit correlation times comparable to DMSO only treated zebrafish embryos at T1. At time T1 a moderate level of persistence was observed similar to that seen in untreated fish (figure 3.5A, 4th row). In the presence of SP600125 the persistence decreased rapidly over time following injury: at T1 there was a higher probability on the diagonal of the matrix that was not seen at T2-T4. Instead high probability was observed in the centre of the transition matrix, meaning that there was bias towards the injury. The patterns in the transition matrices observed in the SP600125 treated cells were similar to the DMSO control over time indicating that JNK inhibition does not have a strong effect on migration behaviour when analysed in this way.

In the untreated and DMSO treated groups we also observed an increased probability along two further diagonals (in addition to the high probability along the diagonal of the transition matrix). Such transitions indicated a forward-backward movement along the same axis. This had not been expected or previously been described; but this behaviour can also be described in a Monte Carlo simulation model (figure 3.5C): in this scenario a leukocyte has a high probability of keeping its direction or moving into the opposite direction in consecutive steps. Changing direction along the x-axis was observed with low probability, i.e. there was a low probability for cells to move along the x-axis parallel to the injury. This forward-backward random walk was clearly apparent in untreated zebrafish and also the DMSO alone group. Inhibiting p38 MAPK decreased this behaviour, while inhibiting JNK removed this characteristic completely.

In both the DMSO and the JNK inhibitor treated groups a high probability in a further 4 positions of the matrix (figure 3.5A, 4th row) was observed. These four areas on the matrix represent movement where the first step is towards the injury and the consecutive step away from the injury (and vice versa). This type of behaviour had not been expected but may represent increased leukocyte trafficking at later time points. We were able to generate simulated cell trajectories that display this type behaviour and computed the corresponding transition matrix (figure 3.5D).

3.3.6 *Spatial dependence of leukocyte migration dynamics*

We investigated the dependence of leukocyte migration behaviour on the location of the cell in relation to the wound site. In our experimental setup the tail transection wound is approximately orthogonal to the notochord in the embryo. We investigated how the migration dynamics change depending on the distance of the cell from the wound, i.e. along the y-axis (see figure 3.1A for orientation). To do this we grouped leukocyte trajectories detected between 3 hpw and 5.72 hpw (T1) into four equal distance intervals from the wound and computed the transition matrix and the correlation time for each interval (S1, $0-155\mu m$, S2 $155-310\mu m$, S3, $310-465\mu m$ and S4, $465-620\mu m$). We found that migration behaviour strongly depends on the distance from the wound in all four groups (figure 3.6A). Leukocyte persistence, high probability along the diagonal of the matrices, decreased with distance from the injury (figure 3.6A, S2-S4). The correlation time also reflects this aspect of the leukocyte dynamics and decreases with distance from wound (figure 3.6B). Note that, as seen in the temporal analysis, we observed high levels of persistence in untreated and fish treated with p38 MAPK inhibitor, while treatment with DMSO and JNK inhibitor resulted in lower overall persistence.

Leukocytes from the untreated, DMSO and JNK inhibitor treatment groups had similar spatial dependencies (figure 3.6A, rows 1-2 and 4). When treated with p38 MAPK inhibitor the spatially resolved dynamics showed a different pattern (figure 3.6A, 3rd row). The persistence decreased with distance and at distances S2-S4 (distance between $155\mu m$ and $620\mu m$) we observed a higher probability for movement towards the injury. This bias was increased, while the level of persistence was decreased, for cells further away from the injury site. Inhibiting p38 MAPK leads therefore to biased and persistent migration behaviour (BPRW, figure 3.4B and C), with both bias and persistence depending on the distance to the wound. In general we found that leukocytes observed at greater distances from the injury site (distance $> 465\mu m$) displayed Brownian motion type random walk (BM) across all groups (figure 3.6A, 4th column and figure 3.4B and C).

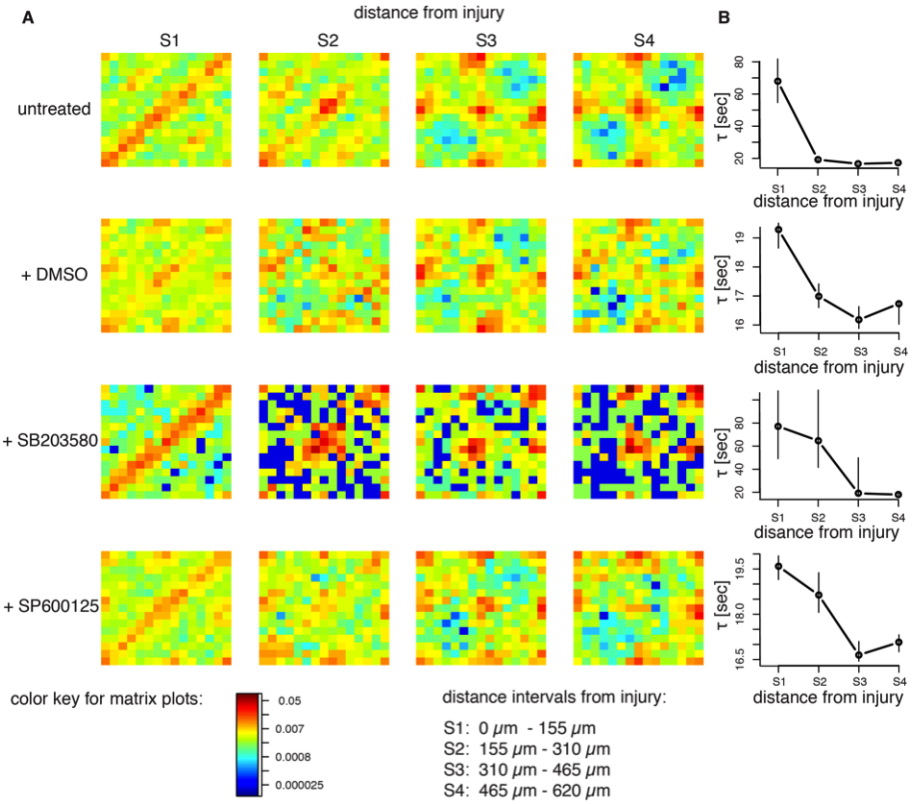


Figure 3.6: The level of persistence in the migration of leukocytes is spatially dependent. Transition matrices as heat maps for the 4 treatment groups. Trajectories detected between 3 and 7 hpw were divided into four equally distributed spatial clusters S1 - S4 according to the distance from the injury, (see legend) and the transition matrices plotted (A). Average correlation time at each cluster (circles) with its bootstrap confidence interval of the mean (vertical lines) (B).

3.3.7 Spatio-temporal leukocyte migration patterns

We studied the combined influence of time since injury (temporal) and distance from injury (spatial) on the level of persistence of the leukocytes. We grouped the leukocyte trajectories into four temporal clusters (T1 - T4), each of these was then split into 4 spatial clusters (S1 - S4), resulting in 16 spatio-temporal clusters. For each cluster we computed the correlation time as a measure of the level of persistence (figure 3.7) for each treatment group.

Untreated and p38 MAPK inhibitor treated leukocytes showed a clear spatio-temporal dependency in their correlation time: the strength of persistence decreased

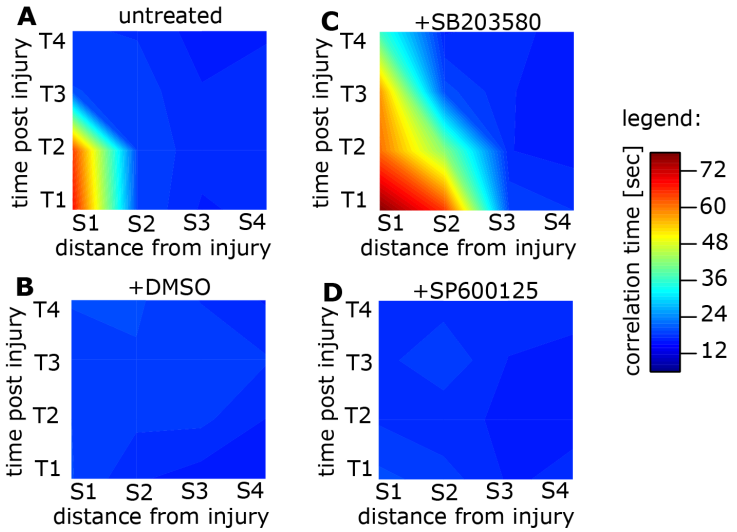


Figure 3.7: The correlation time of migrating leukocyte is spatio-temporal dependent and was modified by DMSO and MAPK inhibitors. The correlation time τ is plotted as a function of the distance from injury and the time post injury for the 4 treatment groups untreated (A), +DMSO (B), +SB203580 (C) and +SP600125 (D). The surfaces represent the interpolation of the measurements, where red is the longest and blue the shortest correlation time.

with increasing distance from the injury (from S1 to S4) and increasing time after injury (from T1 to T4) (figure 3.7A and C). The correlation time in untreated leukocytes rapidly decreased close to the injury (S1-S2) and shortly time after the injury (T2). Inhibiting p38 MAPK slowed this effect considerably and e.g. at T4 ($>$ 11 hpw) leukocytes had a correlation time of 36 seconds at the injury site (53% decrease from T1) (figure 3.7A) compared to 18 sec (73% decrease from T1) in the untreated group (figure 3.7C).

Inhibiting JNK or treating with DMSO alone also reduces correlation times and for early time points after injury (T1) we observe a pronounced spatial dependency with a decrease in the correlation time as we move further away from the wound with increasing distance from the injury (from S1 to S4) (figure 3.7B and D), as mentioned before (figure 4B). This spatial dependence was lost at later time points (T2-T4).

3.3.8 *Heterogeneity in cell populations: characterising macrophages and neutrophils*

The above described analysis was performed using zpu.1:EGFP++ trangenic zebrafish, in which pu.1+ cells are GFP marked. Pu.1+ cells are a mixed population of macrophages and neutrophils, which are known to have different roles and behaviours and even different underlying molecular networks. This fact leads to heterogeneity among the investigated cell population. To further understand the heterogeneity effects based on the mixed population, subsequent *in vivo* imaging was performed with separate markers for neutrophils and macrophages. In this way we wanted to identify differences between macrophage and neutrophil migration patterns with respect to space and time.

Firstly we compute the speed v and the straightness index S_D for both cell populations. Figure 3.8 clearly shows that on average macrophages move slower (0.05 µm/sec) than neutrophils (0.2 µm/sec). Furthermore macrophage have a lower straightness index than neutrophils (figure 3.8B). This simple analysis already shows highly significant differences between the two cell populations.

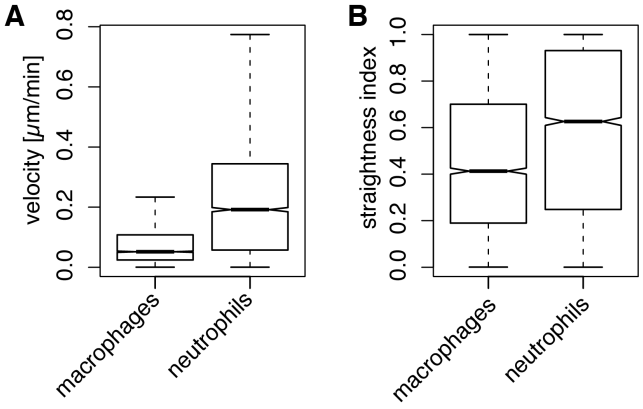


Figure 3.8: Velocity and straightness of macrophages and neutrophils. The velocity (A) and straightness index (B) were computed per detected trajectory and their distribution was plotted as box plots for macrophages and neutrophils, respectively.

Next, the dynamics of the two cell populations have been investigated in more detail, using the above introduced concept of transition matrices. This analysis also takes into account spatio-temporal dependencies. We compare the transition matri-

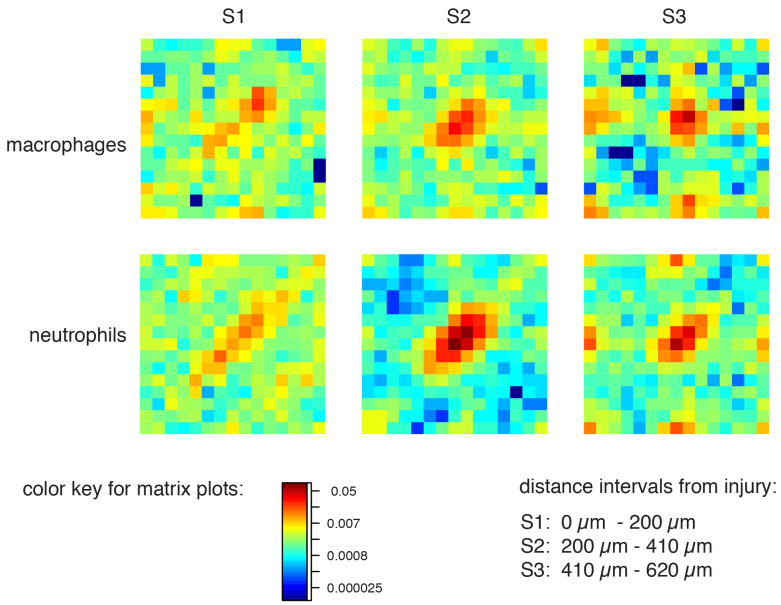


Figure 3.9: Spatial dynamics of macrophages and neutrophils. The transition matrices for the trajectories of macrophages and neutrophils were calculated in dependence of distance from the injury.

ces of macrophages and neutrophils for cell trajectories that were extracted between 1.5h and 4h at different distances from the wound: S1: 0 - 200 µm; S2: 200 - 410 µm; S3: 410 - 620 µm. Macrophages and neutrophils close to the wound (group S1) show a biased-persistent random walk. However, the level of persistence is higher in neutrophils compared to macrophages. In both cell populations the level of persistence decreases with increasing distance from the wound (figure 3.9). To investigate temporal effects we cluster the extracted cells close to the wound (spatial group S1) according to time passed after injury. The time clusters used for this analysis were T1: 1.5-4 hours; T2: 4-6 hours; T3: 6-8 hours and T4: 8+ hours after wounding. The level of persistence increases over time for neutrophils (up to 6h), but decreases again at later time points. On the contrary, the transision matrices computed for macrophages show few temporal dependencies. The level of persistence is similar low over time, with a small bias towards the wound (figure 3.10).

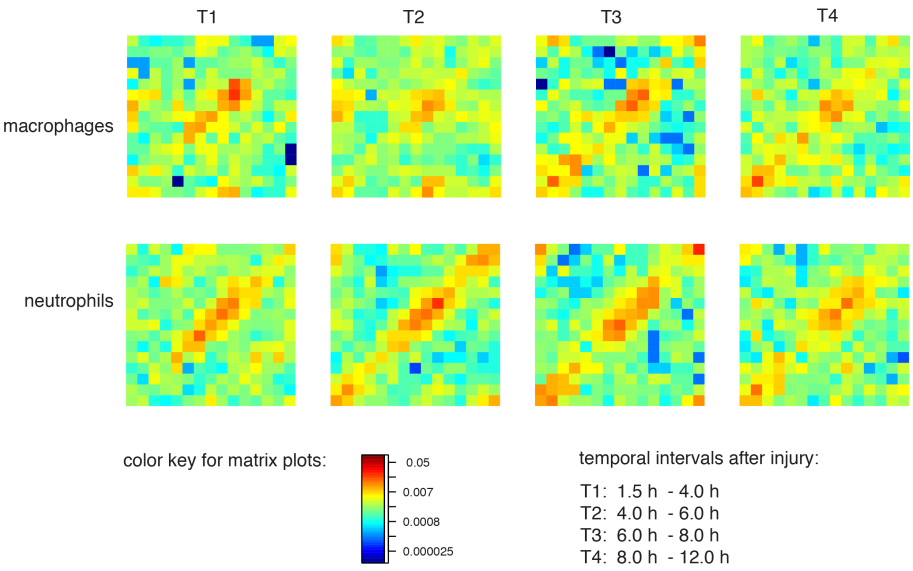


Figure 3.10: Temporal dynamics of macrophages and neutrophils. The transition matrices for the trajectories of macrophages and neutrophils were calculated in dependence of time passed after wounding. Only trajectories close to the wound (0 - 200 µm) were included in the analysis.

3.4 DISCUSSION

The development of sophisticated live imaging facilities enables us to collect high quality cell migration data in live zebrafish embryos. These data are often rich in detail and behaviour, but this also makes them challenging subjects for analysis and statistical investigations. Here we have aimed to capture the spatio-temporal dependence of leukocyte migration in response to wounding. We believe that the new tool of transition matrices which capture the change in directionality of a migrating cell/particle afford a more nuanced description of such processes than previously used statistical measures such as velocity or straightness index. On the one hand this is not surprising as our statistics are multi-dimensional ($n \times n$ if n different intervals of angles are considered); but on the other hand, especially when interpreted in light of the "dictionaries" (presented in figure 3.4) of simulated trajectories, these matrices can be directly linked to certain modes of chemotaxis. Crucially, this methodology enabled us not only to distinguish between the previously described types of random walk, but also revealed new unpredicted migration patterns.

We can therefore interpret cell migration patterns, their change over time and

space, and their dependence on molecular mechanisms in a more straightforward way. The comparison between real data and our dictionaries has enabled us to detect nuances in the migratory behaviour of leukocytes that had hitherto not been observed. The new types of migration behaviour forward and backward is not commensurate with any of the classical random walk behaviour hypotheses [55, 93], but seems to dominate cell migration at later time-points in untreated fish. This forward-backward motion is furthermore biased in the direction perpendicular to the wound in some instances: notably at distances far from the wound in untreated fish, and it is especially pronounced at intermediate and later time points for fish in which JNK is inhibited. Such a dependence is lost if simpler statistical measures, such as the straightness index, are used.

The transition matrices offer a convenient and self-explanatory representation of many aspects of cell migration behaviour and how this is affected by different factors. Here we have used them to test for consistency between hypothetical/theoretical models of random walk behaviour and actual *in vivo* observations of leukocyte migration; and we have been able to propose new models of random walk behaviour that are in better agreement with the observed behaviour under some conditions, when the cell migration clearly deviates from classical, biased, persistent or biased persistent walks.

Computing conventional statistics, such as velocity and straightness index etc., would have failed to detect this nuanced behaviour, which becomes so apparent in the transition matrices. However, these matrices also only become truly useful with the aid of dictionaries (or comparison to numerical simulations). This increased level of detail, however, also comes at a price: we observe pronounced spatio-temporal dependence of the transition matrices, and thus infer concomitant changes in the migratory behaviour of cells, which need to be considered: heterogeneity between cells is to some extent a function of time since and distance to the wound, it appears. This has thus far not received the level of attention it deserves in light of our findings.

This heterogeneity (and the way in which leukocytes respond to wound injury) can be tempered with by selectively inhibiting signalling proteins in the zebrafish embryos. Here we have focussed on DMSO, and inhibitors of the p38 MAPK and JNK MAPK signalling proteins to exemplify how this approach allows us to connect molecular processes and migration phenotypes. However, our approach can more

generally be used to study any kind of cell migration data, collected under diverse experimental conditions.

To understand the impact of the specific molecules on the cell migratory behaviour more detailed and comprehensive (multifactorial) inhibitor studies will be necessary. The platform we present here will help to complete such studies.

Another reason for the heterogeneity is the mixed cell population that is marked in the pu.1:EGFP zebrafish. The separate analysis of macrophages and neutrophils allowed us to distinguish the dynamical patterns of both cell types. The higher motility of neutrophils compared to macrophages is clearly apparent in this analysis. These expected differences can be linked to the differences in the underlying signalling cascades in both cell types.

In conclusion, this work serves to demonstrate the uses and potential insights to be gained from considering transition matrices as descriptions of random walks. While they are staple methods in the simulation of random walks (and a plethora of other stochastic phenomena), this is, to our knowledge, the first time that they have been used in this inverse or reverse engineering capacity. This is, of course, more widely applicable than just to the present context of leukocyte response. While visualising the transition matrices has the additional benefit of serving as a convenient way of exchanging ideas and concepts between experimentalists and modellers, their use in reverse engineering tasks more generally seems equally promising and has here not really been explored at depth. Here recent years have seen advances in connecting simulation studies more directly and immediately to data [15, 156–160] in order to parameterise or infer structures of mechanistic models (here, for example, signalling pathways regulating the cell migratory behaviour).

Bayesian Inference with Biological Trajectory Data

Parts of this chapter have been published in [159].

4.1 INTRODUCTION

Mathematical models of biological systems are abstractions of much more complicated processes [161]. Such models allow us to summarise our state of knowledge about biological systems and processes in a concise manner; to explore likely dynamics of biological systems; and to elucidate experimentally inaccessible aspects of the molecular machinery underlying complex phenotypes and the function of biological organisms more generally. In mathematical studies abstraction is not so much seen as a necessity but as a virtue, which enables us to focus on the principal underlying mechanisms. However, even most experimental analyses are performed under conditions that are very different from those encountered in natural systems. *In vivo* analyses are often performed under as close to realistic conditions as possible, but even here many interactions, e.g. with the environment, are controlled or suppressed. As biological research moves closer to clinical applications it becomes necessary to include more of these details in the analysis of biological systems. In principle it is straightforward to add detail to mathematical models, too. But in practice it

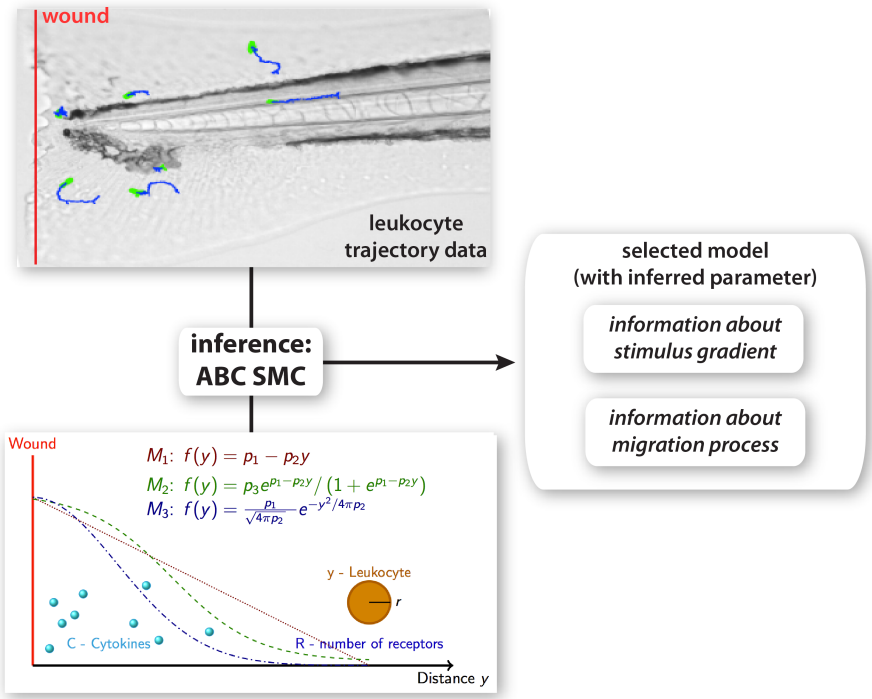


Figure 4.1: Overview of methodology. Leukocyte trajectory data were extracted from time lapes microscopy experiments and used for model selection and parameter inference using the ABC SMC framework. A model for leukocyte migration was constructed, which involves the production of a cytokine gradient after wounding (red line), the sensing of the gradient using receptor binding kinetics and the translation of the signal into movement of the leukocyte. Three different models for the stimulus gradients (M1 - M3) were proposed. From the model that explains the experimental data best information about the migration mechanism can be obtained as well as information about the stimulus gradient.

then becomes more difficult to calibrate (or "fit") these more detailed models against complex and potentially highly resolved data [14].

Here we introduce a statistical methodology that is able to estimate parameters of mathematical models from spatio-temporally resolved *in vivo* data. We employ a Bayesian framework, which also allows us to rank an arbitrary number of alternative models in light of data [14, 15]. Our approach does not require evaluation of the likelihood which is forgone in favour of a simulation-based approximate Bayesian computation procedure [111]. Here simulated data are compared with observed data, and this way parameter (and model) posterior distributions can be constructed even

for cases where conventional statistical approaches are currently unfeasible. We illustrate our approach in the context of leukocyte migration inside zebrafish embryos [53, 54, 56]. We make use of the same imaging facility as introduced in chapter 3 to extract leukocyte trajectory data from injured zebrafish. By coupling the spatio-temporal leukocyte dynamics with models of intra-cellular signal transduction models, true multi-scale analysis becomes within reach [162, 163].

Over the past decade several approaches for analysing and modelling leukocyte migration have been published [75]. As early as 1980 Alt used mathematical descriptions of a biased random walks to model cell chemotaxis [164]. In 1987 Tranquillo *et. al.* [77, 153] proposed a first model for leukocyte chemotaxis. Later other groups constructed leukocyte migration models regarding different aspects of the migration process in response to external stimuli [87, 89, 91].

Here, we combine modelling of leukocyte migration in a living organism in response to wounding with intercellular signalling processes. We construct a model that describes the leukocyte dynamics. Our dynamical model depends on a stimulus gradient shape, which is unknown. We propose 3 different gradient shapes (M1 - M3) and compare the combined model output to the leukocyte trajectories extracted from life imaging data (figure 4.1). We use approximate Bayesian computation (ABC) for model selection to infer the stimulus gradient shape. ABC allows us additionally to gain further details about the leukocyte dynamics during the migration process.

We outline a generic statistical framework that allows us to discriminate between different competing mechanistic models, estimate model parameters, and understand biological processes at a physiological/whole organism level. We illustrate the insights that can be obtained in this framework in the context of leukocyte migration patterns following injury to zebrafish embryos, and conclude with a discussion of how informative such data are about mechanistic models.

4.2 METHODS

4.2.1 *Data acquisition and image processing*

Two datasets were generated and analyzed. The first dataset was used to validate the statistical approach and the simple model of leukocyte migration. This dataset was provided by Daniel Irimia (Harvard Medical School, Boston) and was generated using a microfluidic device, as described in [40]. It contains trajectory data of human neutrophils in a linear interleukin 8 gradient. The second dataset describes the migration of zpu.1:EGFP+ + cells in a living zebrafish embryo after tail transection. The zebrafish were treated exactly as described in chapter 3. The images analysed contained fluorescent zpu.1:EGFP+ + cells. The image processing also corresponds to that in chapter 3.

4.2.2 *Statistical analysis of datasets*

The ABC SMC algorithm is implemented in the `Python` package `ABC-SysBio` [165] and was adapted to allow for comparison between summary statistics and model simulation in *R*. ABC SMC was applied to both datasets. The inference on the second dataset was applied to all 5 temporal groups simultaneously, resulting in 5 different gradient shapes for the 5 time points, and one set of model parameters. The model parameters and their prior distributions are summarised in table 4.1. The gradient specific parameters for the 3 models are: $p_1 = U[0, 100]$ (in rel. gradient concentration), $p_2 = U[-1, 0]$ (unitless) (M1); $p_1 = U[-100, 100]$ (in rel. gradient concentration), $p_2 = U[-1000, 1000]$ (unitless), $p_3 = U[0, 100]$ (in $1/\mu m$) (M2) and $p_1 = U[0, 1000]$ (in μm^2), $p_2 = U[0, 1000]$ (in rel. gradient conc./μm) (M3), where $U[a, b]$ is the uniform prior distribution with minimum a and maximum b.

The first *in vitro* dataset contains 122 trajectories that are spatially distributed in the microfluidic device. The IL 8 gradient is known and constant over the measurement time. The second *in vivo* live-imaging dataset contains 341 trajectories extracted from 18 zebrafish embryos, which are spatially distributed in the zebrafish tail (Figure 4.1). The data were captured from 2 to 10 hours after wounding. For this dataset the gradient is unknown and can not be assumed to be constant over time. Because of that we group the trajectories in 5 equally distributed intervals over time to account for the temporal resolution. To analyse the spatial effects we group

the trajectories according to their distance from the gradient source and wound for the first and second dataset, respectively. We use again equally distributed intervals. Compared to the first dataset we have now 5 temporal groups in dataset 2 (instead of 1) and each of them contains 3 spatial groups.

The following analysis is the same for both datasets. Since we are interested in the characteristics of the cells that describe the dynamics in dependence to the distance from the gradient source / wound, we analyse the motion parallel to the y-axis only. For each spatio-temporal group we compute the distribution of the straightness indices $S_D^{(i)}$ for the extracted trajectories accordingly to equations 4.16 - 4.18. As a result we obtain 3 spatial distributions for the first dataset and 5 times 3 spatio-temporal distributions for the second dataset (Figure 4.4 bottom row).

4.2.3 Robustness analysis

To understand how the dynamical behaviour depends on simultaneous changes to model parameters we perform a robustness analysis on the posterior distributions. The posterior parameter distribution allows us to evaluate the Fisher Information matrix [166], and the eigenvalues and the corresponding eigenvectors correspond to the information content. To determine so called "stiff" and "sloppy" parameter/parameter combinations [115] we performed a principal component analysis (PCA) on the posterior parameter distributions, focusing on those parameters that are relevant for leukocyte migration (table 4.1). The marginalised posterior distributions for parameters $\mathsf{K_d}$ and R are very close to their prior distributions, which means that they are not inferable given the provided datasets. Because of that we exclude these two parameters from the robustness analysis. The PCA was done on the correlation matrix of the remaining five parameter distributions. We can use the correlation matrix here, because all parameters can have values between $\mathsf{0}$ and $\mathsf{1}$ only. The 1st principal component (PC1) shows the "sloppiest" parameter vector, i.e. the parameter combination that carries the least information. We contrast this with the 5th principal component (PC5), which provides the "stiffest" parameter vector and therefore the parameter combination for which the data exhibit the highest information content (figure 4.7 B-D). These vectors can be visualised as in figure 4.7 A, where the pairwise probability density of the parameters is plotted with the vector

for the "sloppy" (red line) and the "stiff" (blue line) direction. The length of the vector represents the inverse of the information content.

4.3 RESULTS

4.3.1 Modelling leukocyte migration

The investigated biological system is illustrated in Figure 4.1. In case of an injury to the zebrafish embryo a stimulus, e.g. cytokines, is released from the injured tissue and/or surrounding tissue, which leads to the establishment of a stimulus gradient. In our experiments we introduce the wound by tail transection so that the wound is orthogonal to the blood vessels of the zebrafish. Because of this simple wound geometry we can assume that the generated distribution of the stimulus is uniform in the direction parallel to the wound (here $\mathtt{x}$ direction), but changes with the distance to the wound, i.e. the direction orthogonal to the wound (here $\mathtt{y}$ direction). Now the concentration of the stimulus is a (unknown) function of the distance to the injury $\mathtt{f}(\mathtt{y},\mathtt{t})$, where $\mathtt{t}$ is the time passed after wounding. A leukocyte is here described by its centre at position $\mathtt{y}$ and its radius $\mathtt{r}$ (more detailed descriptions are possible, of course). Leukocytes move randomly until they are stimulated, e.g. when they sense a cytokine signal. This local external gradient will be translated into an internal cellular signal gradient of signalling molecules that activate F-actin polymerisation in areas with high gradient concentrations, and myosin aggregation in areas with low gradient concentrations, as well as microtubule assembly and disassembly, which leads to a movement of the leukocyte in the direction of the highest stimulus. Thus the direction of leukocyte movement depends on the slope of its internally generated gradient. We describe the behaviour of leukocytes as a sequences of steps, where each step includes: (i) sensing of the gradient by random protrusion of pseudopodia [91], (ii) collapsing pseudopodia in low gradient concentrations while keeping them in high gradient concentrations, and finally (iii) moving towards the high concentration. This discretisation of a, in principle, continuous process is motivated by the type of data we use for the analysis. The data contain leukocyte trajectories with values every

15 seconds. We can compute the average distance between two consecutive steps (average step size) from the data. Now our model needs to describe the direction of the cell, which depends on the directional bias and persistence. We do so by modelling the cell movement as a sequence of angles α_t between consecutive steps t and $t+1$, which is a weighted mean of two processes, persistence and bias:

$$\alpha_t = \underbrace{w_p N_c(\alpha_{t-1}, \mathrm{Var}_p)}_{\text{persistence}} + \underbrace{w_b N_c(0, \mathrm{Var}_b)}_{\text{bias}} \tag{4.1}$$

with the first term describing the directional persistence of the leukocyte (i.e. its tendency to keep moving in the same directions) and the second term describing the directional bias of the leukocyte towards the direction of the highest stimulus, where w_p and w_b are the weights and N_c is the circular normal distribution with its mean and variance. Note, that by using the weights w_p and w_b our model captures not only different levels of a biased persistent random walk, but also the more simple processes of only persistence random walk (if $w_b = 0$) and only biased random walk (if $w_p = 0$). Both directional persistence and bias were previously used to model leukocyte migration [75, 91, 92, 151, 152, 164]. However, in our model the level of the persistence and of the bias is not constant for a given cell, but instead depends on the gradient concentration the cell is sensing. The strength of the persistence and of the bias is here expressed as the variance of the two normal distributions in equation 4.1. Since this model is mechanistic and we do not have much information about the biological parameters that regulate these dependencies, we aim to express our model with normalised parameters in a form as concise as possible. The variance of the two processes can be any positive number. To normalise it we introduce the concentration parameters ρ_p and ρ_b:

$$\mathrm{Var}_p = -2\log(\rho_p) \text{ and } \mathrm{Var}_b = -2\log(\rho_b), \tag{4.2}$$

with $\rho \in (0, 1)$, to compute the variance for directional persistence and bias, respectively. If ρ_p and ρ_b are close to 1 a cell's migratory behaviour will exhibit high persistence and high bias. We assume that both processes, directional persistence and bias depend to some level on the external gradient such that we observe lower variance for high levels of persistence and bias. However, the (effective) gradient concentration is described by an unknown analytical function $f(y)$. We prefer to use the term "effective gradient" as the real gradient will be more complex and presum-

ably depend on a multitude of variables, as well as being more irregular/noisy than the forms considered here. This effective gradient subsumes these complications and describes the input that is sensed by the cell and translated into a internal gradient. The sensing happens via receptors responsive to the external stimuli, *i.e.* cytokine receptors, which are assumed to be uniformly distributed on the surface of the cell, i.e. $R_{\text{front}} \approx R_{\text{rear}} \approx R$. Because of the wound geometry we assume that the leukocyte dynamics are dependent on the distance to the injury (y-direction) but not on the direction parallel to the injury (x-direction), the model of gradient sensing depends only on the y-direction. The leukocyte movement however is described in the x and y direction. Each binding event of a cytokine will lead to the activation of a signalling cascade to generate the internal gradient. Experimental data provide only little (and in part contradictory) information about the involved signalling cascades. Furthermore our experimental leukocyte trajectory data alone are unlikely to provide enough information to infer the details of these signalling cascades. For these two reasons we will simplify the signalling processes and assume that the translation from the external gradient into the internal gradient is linear. The internal gradient then depends on ligand-receptor binding kinetics, i.e. on the amount of ligand-receptor complexes C. The amount of bound ligand in steady-state can be derived using the following scheme:

$$R^* + L^* \underset{k_2}{\overset{k_1}{\rightleftarrows}} C, \tag{4.3}$$

with the number of receptors per cell R^*, the amount of ligand around the cell L^* and the amount of bound ligand per cell C. This reaction can be described by the ordinary differential equations:

$$-\frac{dR}{dt} = -\frac{dL}{dt} = \frac{dC}{dt} = R^*L^*k_1 - Ck_2. \tag{4.4}$$

In steady-state and by defining $K_d = \frac{k_1}{k_2}$ equation 4.4 results in:

$$C = \frac{R^*L^*}{K_d}. \tag{4.5}$$

Using the conservation rules $R^* = R_0 - C$ and $L^* = L_0 - C$, where R_0 and L_0 are the initial amount of receptor and ligand, respectively, we obtain the quadratic equation:

$$0 = C^2 - C(R_0 + L_0 + K_d) + R_0L_0. \tag{4.6}$$

Solving equation 4.6 and using $R = R_0$ and $f(y,t) = L_0$ we obtain

$$C(f(y)) = \frac{1}{2}(R + f(y,t) + K_d) - \sqrt{\frac{1}{4}(R + f(y,t) + K_d)^2 - Rf(y,t)}. \tag{4.7}$$

with the number of receptors R, the receptor binding constant K_d and the relative ligand concentration $f(y,t)$.

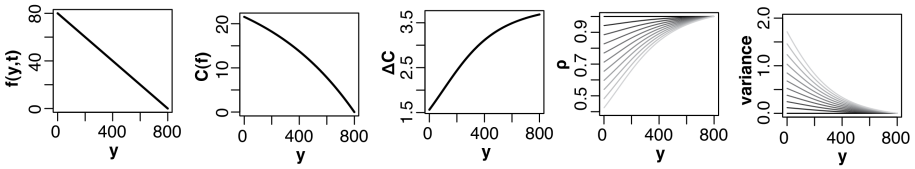


Figure 4.2: Dependencies between strength of persistence on the gradient shape. Shown is an example gradient shape ($f(y,t)$ and the resulting sensed internal gradient $C(f)$, the concentration difference between front and back of a leukocyte ΔC, the concentration parameter ρ and the variance of the circular normal distribution according to equations 4.2 - 4.10. The parameters R, K_d and p_{max} are fixed as an example. The parameter d_p is ranged from 0 to 1 represented from dark to light.

4.3.2 Parameterisation of a leukocyte migration model

We assume that the level of persistence and bias depends to some level on the slope of the gradient regulated by the parameters d_p and d_b. To normalise this dependency we divide the gradient slope $C(f(y-r,t))-C(f(y+r,t))$ by the largest existing slope $\triangle C_{max}$:

$$\triangle C_{max} = \underbrace{\text{argmax}}_{y}(C(f(y,t)) - C(f(y+2r,t))), \tag{4.8}$$

where r is the radius of the cell to describe front and rear. Now the parameters ρ_p and ρ_b can be expressed as:

$$\rho_p = p_{max}\Big(1 + d_p\Big(\frac{C(f(y-r,t)) - C(f(y+r,t))}{\triangle C_{max}} - 1\Big)\Big), \tag{4.9}$$

$$\rho_b = b_{max}\Big(1 + d_b\Big(\frac{C(f(y-r,t)) - C(f(y+r,t))}{\triangle C_{max}} - 1\Big)\Big), \tag{4.10}$$

where p_{max} and b_{max} describe the maximum possible persistence and bias, respectively, and r is the radius of a cell. The term $\frac{C(f(y-r,t))-C(f(y+r,t))}{\triangle C_{max}}$ describes the relative gradient slope at time t and position y, which can have any value between 0 and 1. If the gradient slope is the highest (around 1) then for example equation 4.9 is:

$$\rho_p = p_{max}\Big(1 + d_p\Big(1 - 1\Big)\Big) = p_{max}, \tag{4.11}$$

i.e. the level of persistence is the maximum level of persistence p_{max}. On the contrary, if the relative gradient slope is very small (around 0) then equation 4.9 is:

$$\rho_p = p_{max}\Big(1 + d_p\Big(0 - 1\Big)\Big) = p_{max}(1 - d_p), \tag{4.12}$$

i.e. the level of persistence is the maximum level of persistence reduced by the dependence on the gradient slope $p_{max} - d_p$. The same relationship is described for the bias in equation 4.10. These equations show that the parameters p_{max}, b_{max}, d_p and d_b can all take values between 0 and 1. To understand the relationship between the effective gradient concentration and the resulting variance as a function of y we plot equations 4.7 - 4.10 for an example gradient $f(y,t)$ (figure 4.2).

Here we discuss how the mathematical model can be calibrated against observed trajectories of leukocyte migration from live imaging data. As mentioned above, the leukocyte migration model includes an unknown function, $f(y)$, which describes the spatial (and, implicitly, temporal) distribution of the stimulus with respect to the site of the injury. Several alternative phenomenological models have been proposed [87, 90] and three different distributions are considered here (figure 4.1),

$$M_1: \quad f(y) = p_1 - p_2 y \tag{4.13}$$

$$M_2: \quad f(y) = p_3 \times e^{p_1 - p_2 y} / \left(1 + e^{p_1 - p_2 y}\right) \tag{4.14}$$

$$M_3: \quad f(y) = \frac{p1}{\sqrt{4\pi p_2}} e^{-y^2/4\pi p_2} \tag{4.15}$$

where y is always the distance to the injury and p_1, p_2 and p_3 are unknown parameters that define the effective gradient shape, and which here need to be inferred from the data. The models describe a linear gradient (M_1), a sigmoidal gradient (M_2) and a gradient generated by a standard diffusion process (M_3).

In order to estimate the gradient shape that explains the leukocyte dynamics best, we apply an approximate Bayesian computation (ABC) approach as the likelihood for random-walk processes in unknown gradients is too cumbersome to evaluate exactly. ABC methods have been developed for just this case but where simulation of the (plausible) data generating process is still possible.

Typically observed, x, and simulated data, x'_θ, where θ is a parameter drawn from its appropriate prior distribution, $\pi(\theta)$, are compared via some distance measure, $d(x, x')$. Only when $d(x, x'_\theta) < \epsilon$, where ϵ is the desired tolerance level, is θ considered as a valid draw from the (approximate) posterior distribution, $Pr(\theta|x)$. When the

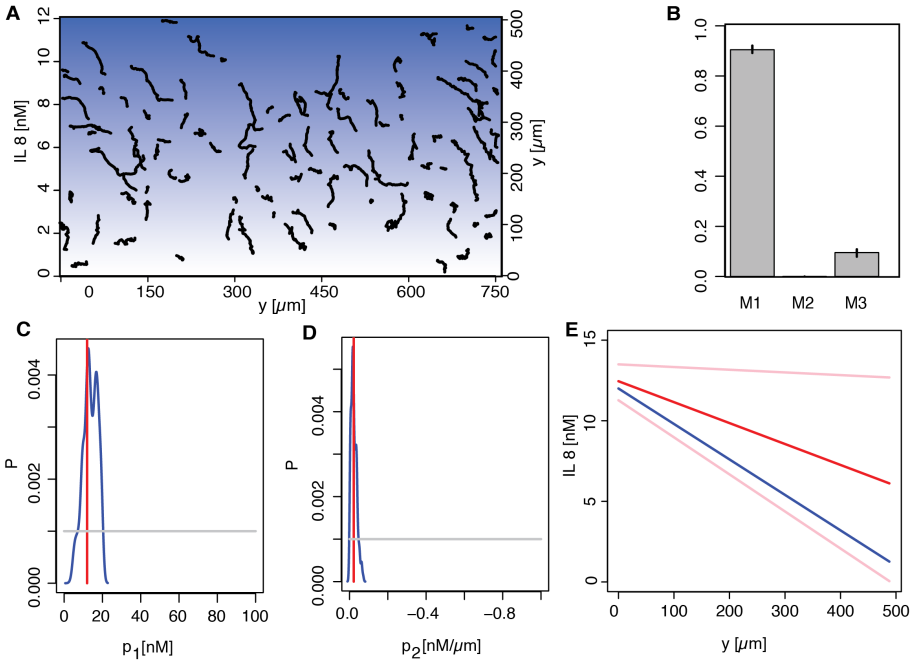


Figure 4.3: Validation of migration model and ABC approach. Trajectories of human neutrophils (black lines) in an interleukin 8 (IL8) gradient were extracted (A). The blue background visualises the IL8 gradient (from white to dark blue for 0 nM to 12 nM). All cells have a tendency to migrate towards high IL8 concentration. We used ABC SMC to obtain the posterior model probabilities (B). The prior distribution was uniform among all 3 gradient models. Shown are the mean probabilities over 5 runs, error bars are the lowest and highest probability of the 5 runs. Model 1 in the last population (population 22) has a probability of 0.89 to fit the experimental data best compared to the remaining two models (B). The estimates of the gradient parameters (blue lines) (shift, p_1, and slope, p_2, of a line) are shown in the prior range (C - D). The prior distribution (grey lines) was for both parameters uniform. The experimental measured parameters (red lines) are in the posterior parameter range. The inferred gradient shape (blue line) with 95 percentiles (pink lines) (interleukin concentration as a function of the distance from the source) and the experimental measured gradient shape (red line) is shown (E).

data are detailed or have a complicated structure, then the probability of generating a simulated dataset that resembles the observed data closely becomes vanishingly small if θ is drawn from the prior. In this case it is, with a number of *caveats*, possible to make some progress by only comparing summary statistics of real and simulated data, $S(x)$ and $S(x'_\theta)$, respectively; especially when $S(.)$ is a sufficient statistic of the data, this compression is loss-less and considerable speed-gains are obtained while still retaining a valid approximation to the posterior (subject to the tolerance level ϵ).

Even if $S(.)$ is not *a priori* sufficient it is possible to perform parameter estimation and a range of methods has been proposed that allow the construction of (approximately) sufficient statistics by pooling information captured by different summary statistics [112, 167, 168].

ABC approaches can also be used for model selection, where we seek to evaluate $Pr(M_i|x)$, *i.e.* the posterior distribution of a model (chosen from a set of candidate models, $M = \{M_1, \ldots, M_q\}$. Here sufficiency across models is also a problem, but here, too, methods to construct sufficient sets of statistics exist. As it turns out, the construction of sufficient statistics is straightforward for random-walk like processes [112]. For each scenario we compute the distribution $\mathcal{S}$ of the straightness indices $S_D^{(i)}$ for the extracted trajectories,

$$S_D^{(i)} = \frac{d_i}{l_i}, \tag{4.16}$$

where l_i is the total length of the trajectory and

$$d_i = |y_0, y_{end}| \tag{4.17}$$

is the Euclidian distance between start and end point of each trajectory i and equation 4.16 can be written as

$$S_D^{(i)} = \frac{|y_0, y_{end}|}{l_i}. \tag{4.18}$$

The straightness index is dependent on the number of observed steps nl of the trajectory thus we split all trajectories so that the resulting trajectories have all the same number of steps of $nl = 30$ steps. The precise value of nl used for the analysis does not seem to matter but for $nl = 30$ steps we are able to use the vast majority of trajectories.

In its simple rejection scheme ABC is too slow to cope with real-world problems and several computational improvements have been suggested, including regression-adaptation, Markov chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC) approaches. We adopt the latter approach, in particular the ABC SMC procedure of Toni et al. [15, 111] as implemented in the *ABC-SysBio* package [165], which was adapted to allow for comparison between summary statistics for random walks and model simulation in *R*. This approach samples parameter combinations (particles) from a non-informative prior distribution, simulates the model and compares the simulation results with the experimental data using a distance function. For classical dynamical systems the distance function is usually the Euclidean distance between

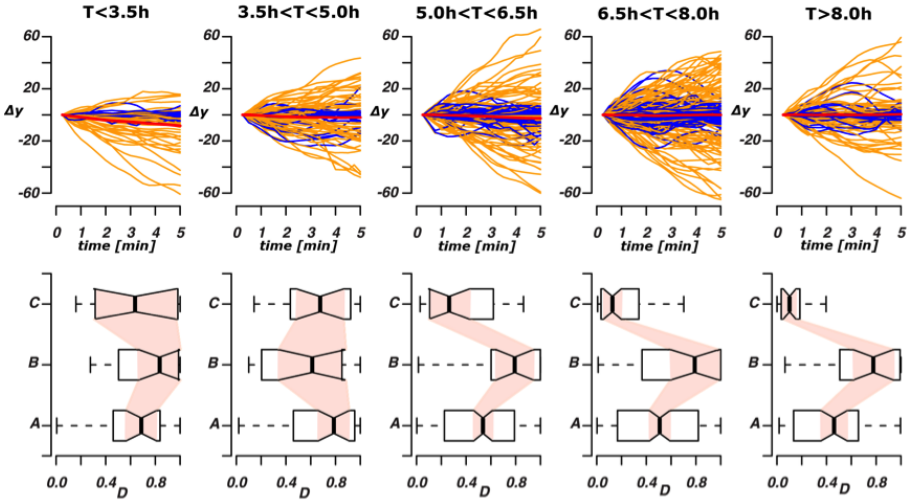


Figure 4.4: Spatio-temporal heterogeneity in chemotactic leukocyte migration behaviour. The top row shows the relative displacement of leukocytes following the start of the tracking (time measured in hpw); each trajectory was considered in non-overlapping 5 min intervals in order to capture any temporal effects acting over longer time scales. Blue trajectories have a straightness index $S_D < 0.5$, orange trajectories have $S_D > 0.5$. The red line indicates the average behaviour across all trajectories shown in the panel. Values $\Delta y > 0$ indicates movement towards the wound, $\Delta y > 0$ indicates movement away from the wound. In the bottom row we show the distributions of the straightness index divided into three different classes according to the distance from the wound (A: $y < 250\mu m$, B: $250\mu m < y < 500\mu m$, C: $y < 500\mu m$) and grouped according to time post-wounding (as top row).

the simulated trajectories and the experimental measurements. However, as we deal with spatio-temporal data that have a considerable random component, the Euclidean distance between single trajectories does not contain sufficient information about whether different trajectories were generated from the same process. To distinguish between different forms of random walk behaviour we compare the distributions, $\mathcal{S}$, of $S_D^{(i)}$. These distributions are generated by simulating 200 trajectories for each sampled particle and we process these in the same way as the experimental dataset. Now we can compare the distribution using the Kolmogorov-Smirnov distance between their respective histograms. The resulting distance function is

$$d = \sum_{t=1}^{N_t} \sum_{s=1}^{N_s} K(\mathcal{S}^{s,t}, \mathcal{S}^{*s,t}), \tag{4.19}$$

where $\mathcal{S}$ and $\mathcal{S}^*$ are the distributions of S for the experimental data and the simulated data, respectively, N_s is the number of spatial groups (here 3) and N_t is the number

of temporal groups, and K is the Kolmogorov-Smirnov distance for pairs of empirical distribution functions. The Kolmogorov-Smirnov distance is defined as:

$$K(\mathcal{S}, \mathcal{S}^*) = \sup_x |\mathcal{S}(x) - \mathcal{S}^*(x)|, \tag{4.20}$$

with $\mathcal{S}$ and $\mathcal{S}^*$ are the two empirical distribution functions that we aim to compare. Details of the implementation are given in section 4.2.

To validate this approach under controlled conditions we first applied it to leukocyte trajectory data extracted from migration patterns in a microfluidic device with a known linear interleukin 8 (IL 8) gradient. Figure 4.3 A shows the extracted trajectories. Our method identifies the experimentally applied linear gradient (model 1, equation 4.13, has 89 % of the posterior probability associated) correctly (figure 4.3 B). The linear gradient has two parameters (p_1 and p_2). The estimates for these parameters are shown in figure 4.3 C-D. The experimental parameter values are covered by the estimated posterior distribution and the resulting predicted gradient shape (figure 4.3 E) is in good agreement with the actual experimental gradient. These results serve as a proof of principle for our model, our statistical approach and demonstrate the ability to extract hidden information from rather simple cell migration data. Purely *in-silico* analyses, where the true model is known by definition, results in similar credible intervals for the gradient parameters (data not shown).

4.3.3 Spatio-temporal analysis of leukocyte migration

We tracked zpu.1:EGFP positive cells in living zebrafish embryos to study the spatio-temporal dynamics of leukocytes in response to wounding. We show that the dynamics of leukocyte movement is dependent upon the position of the cell in relation to the site of inflammation and on the time that has elapsed since the injury. Heterogeneity in spatio-temporal dependencies of leukocyte migratory behaviour is illustrated in Fig. 4.4. Each trajectory is presented as a line with the distance movement towards ($\Delta y < 0$) or away from ($\Delta y > 0$) the wound plotted on the Y axis and time on the X axis (Fig. 4.4 top row). Cells are migrating towards and away from the wound at all time points, which reflects the presence of retrograde chemotactic behaviour [54]. The straightness index (see above) is indicated by the colour of the line: trajectories with a low straightness index ($D < 0.5$) are shown in blue and those that have high directionality ($S_D > 0.5$) in orange. At earlier time points post injury ($T < 3.5$ hpi)

more cells are traveling towards the wound (i.e. $\Delta y < 0$) than at later time points. Displaying each trajectory in this graphical form allows us to appreciate the diversity rather than merely the (not very informative) average behaviour of the immune cell population: while the average population behaviour may always suggest no net movement in the direction perpendicular to the wound (as indicated by the red lines in the top row), the individual migratory behaviour at the single cell level becomes most diverse between 3.5-6.5 hpi. We also plot the distributions of the straightness index of the trajectories at different times, divided into 3 classes according to their distance from the wound (Fig. 4.4 bottom row). At later time points (> 5.0 hpi), the straightness index, S_D, is large for trajectories at intermediate distances and significantly higher than for leukocytes close or far away from the wound.

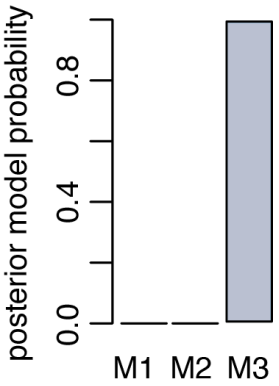


Figure 4.5: Posterior model probablity distribution. Using a uniform prior distribution among the 3 models in the ABC SMC approach model 3 (M3) has a probability of 1 to represent the data best. Model 3 represents a diffusion type gradient that changes over time.

Using different divisions of time results in qualitatively identical behaviour. The chosen partition, however, has the advantage of capturing both the biological phenomenon of a dynamic gradient and resulting in good statistical power (similar numbers of trajectories) for the mechanistic analysis detailed next.

4.3.4 Spatio-temporal characteristics of stimulus gradients

In light of the spatio-temporal behaviour of the leukocytes in response to wounding we next infer the stimulus gradient, for which in general no direct measurements exist. Linking these extracted trajectories with our leukocyte migration model using the

Table 4.1: Leukocyte migration model parameters

parameter	prior	posterior Mean (1st dataset)	posterior Mean (2nd dataset)
d_b (unitless)	$U[0,1]$	$0.42\ [9.6 \cdot 10^{-4}, 0.84]$	$0.60\ [0.05, 0.92]$
d_p (unitless)	$U[0,1]$	$0.41\ [1.3 \cdot 10^{-3}, 0.92]$	$0.67\ [0.26, 0.93]$
b_{max} (unitless)	$U[0,1]$	$0.45\ [1.6 \cdot 10^{-2}, 0.93]$	$0.53\ [0.03, 0.82]$
p_{max} (unitless)	$U[0,1]$	$0.92\ [0.84, 0.97]$	$0.86\ [0.71, 0.95]$
K_d (1/amount per cell)	$U[0,10000]$	$4831.9\ [151.2, 9345.5]$	$5329.03\ [39.2, 8939.4]$
R (amount per cell)	$U[0,10000]$	$5086.3\ [53.1, 9408.4]$	$5052.25\ [226.3, 8612.5]$
w (unitless)	$U[0,1]$	$0.81\ [0.58, 0.98]$	$0.79\ [0.69, 0.88]$

List of parameters specific to leukocyte migration model with priors used in the Bayesian framework (U - uniform prior).

approach described above we gather not only information about the spatio-temporal dynamics of the stimulus gradient but also gain more detailed mechanistic insights into the leukocytes dynamics. The same models for the gradient shape were assumed and tested. The main difference between this dataset and the microfluidic dataset used in the validation of our approach is that the stimulus gradient is now a function of both the distance to the wound and of the time that has elapsed since the injury, $f(y,t)$. Therefore the problem becomes far more complex due to the high dimensional parameter spaces for each of the three models. When we divide time into five intervals each model has now five times as many parameters, which describe the gradient shape (resulting in 17 parameters for model 1 and 3, 22 parameters for model 2). The biophysical reaction parameters that describe the molecular processes inside leukocytes can be assumed to be constant over the time spans considered here [76].

Using ABC SMC with a uniform prior model distribution we find that model 3, which uses a diffusion-process gradient, represents the available datasets best (figure 4.5). The posterior parameter distributions for model 3 reveal more details about the *in vivo* dynamics of leukocyte migration (figure 4.7 A). Parameters d_b and d_p are both higher than 0.5, which indicates that both bias and persistence are dependent on the gradient and therefore spatial characteristics. Parameters b_{max} and p_{max} describe the maximum level of bias and persistence, respectively. The posterior distribution shows that the level of persistence is higher than the level of the bias. This is also seen from parameter w (mean: 0.79), which is the relative weighting between bias and persistence. This parameter is clearly shifted towards 1, i.e. persistence is favoured over the bias. Parameters K_d and R are not inferable. These parameters describe the

binding of the stimulus molecules to the surface receptors, i.e. the stimulus sensing. We can conclude that the trajectory data of the leukocytes do not carry information about these molecular details.

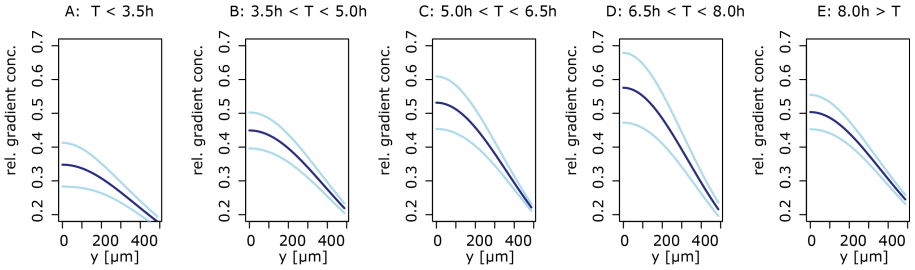


Figure 4.6: Spatio-temporal characteristics of chemokine gradients. Shown are the estimated chemokine gradients for 5 time intervals (A-E) using the mean estimate of the gradient parameters (dark blue lines) and the 5 and 95 percentiles (light blue lines). The gradient concentrations are plotted in relative units.

Finally, we can use the inferred 10 gradient specific parameters to investigate the gradient dynamics over time. Figure 4.6 shows the inferred gradient-specific shapes with their 95 percentiles for 5 time intervals after wounding. For all time points, we observe the classical diffusion shape gradient [93]. Interestingly, the process that generates the stimulus gradient can not result from the analytical form of the heat equation, because the concentration at the source (distance = 0 µm, wound) increases until around 7 hours after wounding. This shows an active production of the stimulus at the site of the injury. After around 7 hours this production decreases, which leads to a decreasing concentration at the wound; this shapes the temporal development of the effective stimulus gradient.

4.3.5 Analysis of leukocyte migration model

The models for leukocyte migration (equations 4.1 - 4.7) allow us to gain information about the stimulus gradient. Next we can use the same model and the corresponding parameter estimates to learn more about the characteristics of the leukocyte movement. Because the parameters, $\mathtt{K_d}$ and $\mathtt{R}$, show flat posterior distributions, they are not inferable from the present data and we can exclude them from the following analysis.

The posterior distribution including only the remaining 5 parameters ($\mathtt{d_b}$, $\mathtt{d_p}$,

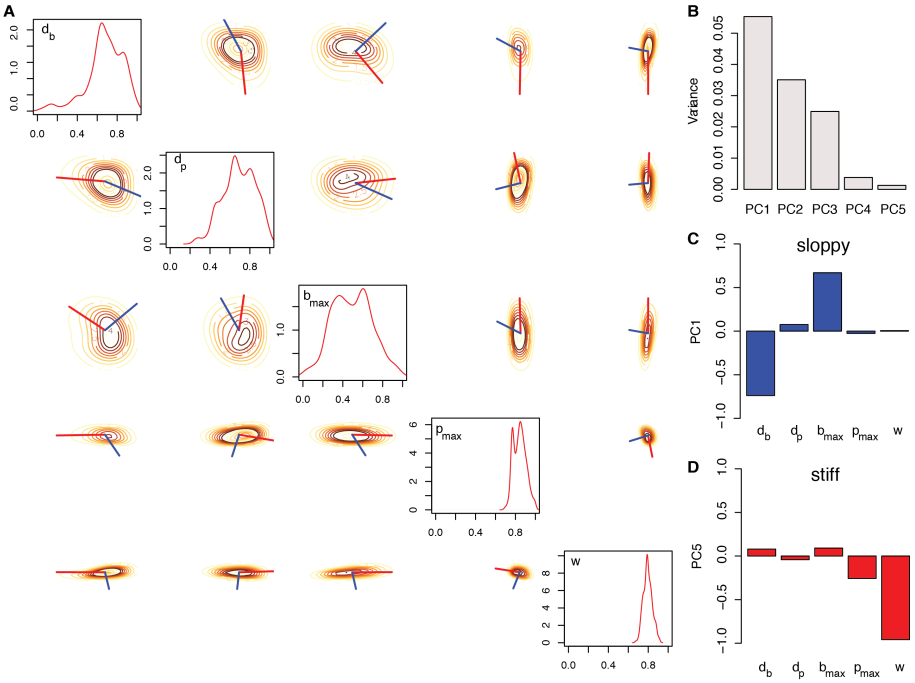


Figure 4.7: Robustness analysis. The marginalized density estimates for posterior parameter distributions related to chemotaxis are plotted on the diagonal in the range of the uniform prior distributions (A). The pairwise posterior density estimates are plotted on the off-diagonal. The red lines show the projection of the first principal component (PC), i.e. the robust direction. The "stiff" direction is displayed by the blue line and represents the vector of the fifth principal component (A). The variance of the principal components are shown in (B). PCA was performed on the correlation matrix of the posterior parameter distribution. The corresponding vectors of the first (C) and fifth (D) PC are visualised by its projections onto the parameters (red bar plots for "sloppy" directions and blue bar plots for "stiff" directions) are shown.

b_{max}, p_{max} and w) was used to determine the relative dependencies of the migratory dynamics on the parameters. This in turn is related to how much information the data carry about the parameters and allows us to determine "stiff" and "sloppy" directions [115]. We calculated the correlation matrix of the posterior distribution and used principal component analysis (PCA) to determine the directions with highest and lowest variance of the overall posterior [122, 169]. Figure 4.7 B shows the 5 principal components (PC) with their corresponding variances (upper row, right). PC1 has the highest variance, therefore the corresponding vector represents the "sloppiest" parameter combination (or the combination of parameters least constrained by the available data). The projections of PC1 onto the "raw" parameter vectors are

shown in figure 4.7 C. The two parameters d_b and b_{max} have the highest projection onto the first principal component, i.e. they are "sloppy" parameters and have thus comparatively minor influence over the leukocyte migration dynamics observed here. These two parameters determine the bias of leukocyte migration towards higher gradient concentrations. This suggests that the movement of leukocytes is not primarily regulated by directional bias. On the contrary, the level of persistence is the dominating characteristic of the leukocyte movement. This results from PC5 (figure 4.7 D), which represent the "stiffest" parameter combination, i.e. the combination of parameters for which the data exhibits the highest information content (and hence the collection of parameters that have the highest impact on system dynamics). The largest projections onto these two principal components are by w and p_{max}. The parameter w shows, as mentioned above, that persistence appears to be more pronounced than bias, while the parameter p_{max} quantifies the level of persistence and also suggests overall highly persistent leukocyte movement. Note that the level of persistence also depends on the slope of the local gradient (mean d_p: 0.67). The importance of bias and persistence can also be seen in figure 4.7 A. Here we show the pairwise posterior probability densities of the parameters. The red and blue lines indicate the "sloppy" and "stiff" directions, respectively. The "stiff" direction is almost parallel to w (representing the persistence), whereas the "sloppy" direction is almost parallel to b_{max} (representing the bias). This indicates that they affect the dynamics independently and that persistence exerts greater influence on the trajectories than bias.

4.4 DISCUSSION

The application of the ABC-centered approach on trajectory data extracted from living zebrafish leukocytes during acute injury provided us with detailed insights into the dynamics of the stimulus gradient. In particular we found evidence that the stimulus changes as a function of space and time. Biologically this is plausible but again current experimental setups cannot routinely overcome the technical difficulties

in measuring such gradients let alone their change (even if all chemokines etc. involved in immune signalling were known). Our results suggest that the stimulus is produced at the site of injury until 7 hours post injury (hpi). Because the stimulus concentration increases at the wound until that time, the diffusion of the stimulus is weaker than its production. This can also be seen from the increasing slope of the gradient until 7 hpi. This increased slope leads to a stronger persistence in the leukocyte movement at intermediate distances and allows a more efficient leukocyte recruitment to the site of injury, i.e. more cells are at the site of the injury.

We then performed robustness analysis on the leukocyte migration model: this reveals the most important characteristics of the leukocyte dynamics, but also safeguards against potential over-interpretation of the dynamics in light of the inferred parameters. From this analysis we learn that the persistence of leukocytes exerts the largest influence on the dynamical behaviour *in vivo*. On the other hand the leukocyte dynamics are robust to changes in the level of bias towards the wound. This means that even with low bias a leukocyte would still manage to migrate towards the wound as long as the level of persistence is high enough and dependent on the local slope of the gradient. Persistent movement seems to be an optimal search strategy for leukocytes during inflammation.

The overarching problem as to how stimulus gradients change with space and time is hard to solve without some further assumptions: here we have partitioned both space and time into discrete intervals. Without this it would be impossible to achieve the statistical power required for the inference of the unknown gradient/model parameters. It may in principle be possible to use an explicitly spatio-temporal parametric model of the gradient but this will require much more detailed knowledge as to what are the actual signalling molecules than is presently known. Inferential techniques, such as ABC, allow us to model immune-response processes by conditioning mathematical models on available data. In the present context, for example, the central finding of a spatially and temporally varying stimulus gradient is probably not surprising or unexpected. Without this type of modelling, however, it would not be possible to determine the relative balance between e.g. persistence and bias in the migratory behaviour. Such mechanistic insights (or hypotheses) cannot be derived from verbal/qualitative models alone.

We present a statistical framework that allows us to calibrate mathematical mod-

els of biological systems against *in vivo* data. Here we study leukocyte dynamics in zebrafish embryos following injury to their tail-fin. We show this response is mediated by a stimulus gradient, which emanates from the wound site and changes with space and time. While this change is not directly observable in experiments, we can infer the spatio-temporal behaviour of the stimulus using an approximate Bayesian computation framework, which is able to reliably infer an experimentally validated *in vitro* stimulus gradient. In chapter 3 we introduced the concept of transition matrices in order to study the types of random walks exhibited my leukocytes in response to wounding. These transition matrices could potentially be used in future instead of the distribution over the straightness index in order to infer more detailed dynamical processes. With a detailed mathematical description of the temporal processes related to the signalling gradient, as well as the internal and external cellular signalling dynamics, we could estimate the temporal evolution of the transition probabilities (for example by applying the forward Kolmogorov equation). However, at the present state this is not possible.

Experimental Design in Systems Biology

Parts of this chapter are published in [170].

5.1 INTRODUCTION

Performing different experiments is costly in terms of both money and time, and not all experiments are equally informative. Ideally we would like to perform only those experiments which yield *substantial* and *relevant* information. Here we have to consider what is meant by both of these terms: we regard any information that decreases our uncertainty about model parameters or model predictions as *relevant*. As we will show below, what is *substantial* information is then easily and naturally resolved. We will show, for example, that experimental interventions differ in the amount of information they provide e.g. about model parameters. Equally some experiments provide insights that are more useful for making predictions about system behaviour than others.

Models in systems biology typically describe how the abundances of a set of molecular entities, $\mathtt{x}$, change with time, $\mathtt{t}$; here the rate of change in $\mathtt{x}(\mathtt{t})$ over time is typically described in terms of (ordinary, partial or stochastic) differential equation

systems, where

$$\frac{dx(t)}{dt} = F_q(x, \theta);$$

here x is a k-dimensional vector describing the system's state and $\theta = (\theta_1 \ldots, \theta_l)$ is an l-dimensional vector containing the model parameters. Finally q denotes the particular experimental setup under which data are being collected. This dependence is generally tacitly ignored but, as we will show below, explicitly incorporating the experimental approach (and the fact that different experimental choices are typically available) into the model and any down-stream statistical analysis allows us to develop strategies that yield more detailed insights into biological systems, and better models thereof.

We therefore require inferential tools that, given some observed biological data and a suitable mathematical candidate model, provide us with parameters that best describe the system's dynamics [111, 118, 119]. Unfortunately obtaining reliable parameter estimates is plagued with difficulties. Usually sparse and notoriously noisy data are fitted using models with large number of parameters [171]. As a result, over-parameterized models tend to fit to the noisy data but may loose predictive power. Conventional fitting approaches to such data routinely fail to capture this complexity by underestimating the uncertainty in the estimated parameters, which substantially decreases their predictive power; the so-called "inverse problem" is often considered (see e.g. Brenner [172]) as one of the major problems facing systems biologists. What we set out here is a rational strategy — together with an associated set of statistical and computational tools — that allow researchers to probe biological systems and develop better, more realistic and predictive mechanistic models. This approach enables experimenters to choose the most appropriate experimental approach to fulfil their respective needs, and ultimately results in better understanding of biological systems at reduced cost and experimental burden.

We use $Q = \{q_1, \ldots, q_{|Q|}\}$ to denote the total set of available experimental assays that could be used to probe a system in a given situation. These might, for example, include knock-out or knock-down mutants, transcriptomic or proteomic assays, different time-courses or different environmental conditions (or both), *etc.* Here we remain very flexible as to which type of experimental setup is included in Q, but merely acknowledge that it is rarely possible to probe all important aspects of a system si-

multaneously. Instead different techniques require different sample preparations *etc*, and therefore separate experiments. Here we account also for the possibility that, for two different experimental set-ups, q and q', the mathematical model may differ; therefore the dependence on q is made explicit in our notation F_q. For example, if species x_k is knocked out in experiment q^*, we can ignore any terms referring to it when modelling F_{q^*}.

Here we show how we can meld concepts from Bayesian inference and information theory to guide experimental investigations into biological systems to arrive at better parameter estimates, better model predictions, and, ultimately better models. We first develop the theoretical concepts before demonstrating the use (and usefulness) of the Bayesian experimental design approach in the context of a number of biological systems that exemplify the set of problems encountered in practice. In order to demonstrate the practical applicability of our approach we investigate two simple models (repressilator and Hes1 systems), as well as a complex signalling pathway (AKT) with experimentally measured dynamics.

5.2 METHODS

5.2.1 Information theoretic design criteria

Our aim is to choose an experiment q from a set of candidate experimental setups, Q, which either reduces uncertainty about model parameters or uncertainty of an outcome of a particular condition q^* for which data are impossible or difficult to obtain. In the information theory framework, these two goals boil down to determining an experiment $q \in Q$ which contains maximal information about the parameter or the desired predictions for condition q^*. In order to present in more details these two goals let us recall some concepts of information theory and experimental design introduced in chapter 2.3. As in equation 2.39 we first define the entropy $H(X)$ of a random variable X, which measures the uncertainty of the random variable,

$$H(X) = -E_X(\log(p(X))) = -\int \log(p(x))p(x)dx \tag{5.1}$$

and the mutual information $I(X, Y)$ between two random variables X, Y, which is the reduction of the uncertainty that knowing Y provides about X,

$$I(X, Y) = H(X) - E_Y(H(X|Y)) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy, \tag{5.2}$$

where $p(x, y)$ is the joint probability density function of X and Y while $p(x)$ and $p(y)$ are the marginal probability density functions. We denote by E_X the expectation with respect to the probability distribution of X. Here we follow the convention where capital letters stand for random variables while lower-case stand for a particular realisation of a random variable.

5.2.2 Reducing uncertainty in model parameters

We first consider the task of choosing an experiment that will on average provide most information about model parameters measured through the reduction in their respective uncertainties. In the information theoretic language, as by Lindley [134] and later by Sebastiani and Wynn [173], the initial (prior) uncertainty is given by the entropy $H(\Theta)$ of the prior distribution $p(\theta)$, which after data x_q have been collected (in experimental setup q) gives rise to the entropy $H(\Theta|x_q)$ of the posterior distribution $p(\theta|x_q)$. The information gained about the parameter by collecting the data x_q is then $H(\Theta) - H(\Theta|x_q)$. On average, however, the decrease of uncertainty about Θ after data are collected in an experiment q is given by $I(X_q, \Theta)$. Therefore, in order to reduce parameters' uncertainties one should choose an experiment that maximises the mutual information between X_q and Θ.

Here we specifically consider models such that the output is of the form

$$x_q = \mu(\theta, q) + \epsilon \tag{5.3}$$

where μ is a deterministic function and ϵ an uncorrelated, zero mean, gaussian random variable with variance σ^2. In such a model maximisation of mutual information $I(\Theta, X_q)$ is equivalent to maximisation of the entropy $H(x_q)$. This observation described first in [173] results directly from the fact that the mutual information $I(\Theta, X_q)$ can be written as the difference between $H(X_q)$ and $E_\Theta(H(X_q|\Theta))$ and that

$$E_\Theta(H(X_q|\Theta)) = -\iint p(\theta)p(x_q|\theta) \log\left(p(x_q|\theta)\right) dx_q d\theta$$

does not depend on the experiment q. Indeed, equation (5.3) implies that $p(x_q|\theta)$ is the probability of the experimental noise ϵ. In this study we use gaussian distributed

noise. But one could use any distribution of noise, for which the probability density function is known in an analytic form, in order to compute $p(x_q|\theta)$. The only assumption here is, that the noise has to be independent for all terms.

Maximisation of $I(X_q, \Theta)$ is equivalent to maximisation of $H(X_q)$. However, this is only the case for the mutual information between the output x_q of an experiment q and the parameter of the system Θ. Whenever we are interested in the increase of information about only one component of the parameter vector, or in reducing uncertainty about an experimental outcome, we need to use the mutual information and not the entropy.

5.2.3 *Reducing uncertainty in an experimental outcome*

Similar reasoning leads us to a criterion for selecting an experiment q that reduces uncertainty about predictions for the system output under a different set of conditions or experiment q^*. Choosing q that maximises $I(X_q, X_{q^*})$ leads to an experiment that on average reduces the uncertainty of predictions for condition q^* most. This can be seen by rewriting (5.2) as

$$I(X_q, X_{q^*}) = H(X_{q^*}) - E_{X_q}\left(H(X_{q^*} \mid X_q)\right). \tag{5.4}$$

5.2.4 *Estimation of the mutual information.*

The mutual information for models of type (5.3) can be estimated using Monte Carlo simulations. We first focus on the mutual information between parameters Θ and the output X_q of an experiment q which can be written as a function of the prior distribution $p(\theta)$, the probability of the output given the parameter $p(x_q|\theta)$ and the evidence $p(x_q)$ as follows

$$I(\Theta, X_q) = \iint p(\theta, x_q) \log \frac{p(\theta, x_q)}{p(\theta)p(x_q)} d\theta\, dx_q = \iint p(\theta)p(x_q|\theta) \log \frac{p(x_q|\theta)}{p(x_q)} d\theta\, dx_q\,. \tag{5.5}$$

Drawing a sample $\{\theta^{(i)}\}_{1 \le i \le N_1}$ from the prior distribution $p(\theta)$ we obtain a Monte-Carlo estimate,

$$I(\Theta, X_q) \approx \frac{1}{N_1} \sum_{i=1}^{N_1} \log \frac{p(x_q^{(i)}|\theta^{(i)})}{p(x_q^{(i)})}, \tag{5.6}$$

where for all $1 \le i \le N_1$, $x_q^{(i)}$ is an output of the system for the parameter $\theta^{(i)}$. For models of type (5.3) $p(x_q|\theta)$ is the probability density function of a Gaussian

distribution with mean $\mu(\theta, q)$ and covariance $\sigma^2\mathbf{I}$ taken at x_q. To compute the quantity in (5.6) we have to estimate the evidence $p(x_q^{(i)})$, which can be done via Monte Carlo simulation: given a N_2-sample $\{\theta^{(j)}\}_{N_1+1\leq j\leq N_1+N_2}$ drawn independently from the prior distribution $p(\theta)$ with $\{\theta^{(i)}\}_{1\leq i\leq N_1}$ we have

$$p(x_q^{(i)}) = \int p(x_q^{(i)}|\theta)p(\theta)d\theta \approx \frac{1}{N_2}\sum_{j=N_1+1}^{N_1+N_2} p(x_q^{(i)}|\theta^{(j)}) \,. \tag{5.7}$$

Combining equations (5.6) and (5.7), we obtain the following estimate of the mutual information between the parameter θ and the output x_q,

$$I(\Theta, X_q) \approx \frac{1}{N_1}\sum_{i=1}^{N_1}\left[\log\left(p(x_q^{(i)}|\theta^{(i)}),\right) - \log\left(\frac{1}{N_2}\sum_{j=N_1+1}^{N_1+N_2} p(x_q^{(i)}|\theta^{(j)})\right)\right] \,. \tag{5.8}$$

Similarly, we can estimate the mutual information between any single component θ_c of the d-dimensional parameter vector, Θ, and the output X_q. We denote by $\Theta_{\bar{c}}$ the $d-1$-dimensional vector containing all the components of Θ except the c-th one. Integrating over $\theta_{\bar{c}}$, the mutual information between Θ_c and X_q is equal to

$$\begin{aligned} I(\Theta_c, X_q) &= \iint p(\theta_c, x_q)\log\frac{p(\theta_c, x_q)}{p(\theta_c)p(x_q)}d\theta_c\, dx_q \\ &= \iint p(\theta, x_q)\log\frac{p(\theta_c, x_q)}{p(\theta_c)p(x_q)}d\theta\, dx_q \\ &= \iint p(\theta)p(x_q|\theta)\log\frac{p(x_q|\theta_c)}{p(x_q)}d\theta\, dx_q \end{aligned}$$

and can be estimated through Monte Carlo simulation by

$$I(\Theta_c, X_q) \approx \frac{1}{N_1}\sum_{i=1}^{N_1}\log\frac{p(x_q^{(i)}|\theta_c^{(i)})}{p(x_q^{(i)})} \tag{5.9}$$

As previously the evidence $p(x_q^{(i)})$ can be estimated from a sample drawn from the prior without great difficulty. On the other hand, the estimation of the numerator $p(x_q^{(i)}|\theta_c^{(i)})$ requires an additional integration over $\theta_{\bar{c}}$. We then have

$$p(x_q^{(i)}|\theta_c^{(i)}) = \int p(x_q^{(i)}|\theta_c^{(i)}, \theta_{\bar{c}})p(\theta_{\bar{c}}|\theta_c^{(i)})d\theta_{\bar{c}} \approx \frac{1}{N_3}\sum_{j=1}^{N_3} p(x_q^{(i)}|\theta^{(i,j)}) \tag{5.10}$$

where for each $1 \leq i \leq N_1$, $\{\theta^{(i,j)}\}_{1\leq j\leq N_3}$ is a sample drawn from $p(\theta)$ under the constraint that $\theta_c^{(i,j)} = \theta_c^{(i)}$. Putting all the terms together, we obtain the following estimation of the mutual information between Θ_c and X_q: given a sample $\{\theta^{(i)}\}_{1\leq i\leq N_1+N_2}$ drawn from $p(\theta)$ and a $N_1 \times N_3$-sample $\{\theta^{(i,j)}\}_{1\leq i\leq N_1,\, 1\leq j\leq N_3}$ such that for all i, j, $\theta^{(i,j)}$ is drawn from $p(\theta|\theta_c^{(i)})$,

$$I(\Theta_c, X_q) \approx \frac{1}{N_1}\sum_{i=1}^{N_1}\left[\log\left(\frac{1}{N_3}\sum_{j=1}^{N_3} p(x_q^{(i)}|\theta^{(i,j)})\right) - \log\left(\frac{1}{N_2}\sum_{k=N_1+1}^{N_1+N_2} p(x_q^{(i)}|\theta^{(k)})\right)\right] \,. \tag{5.11}$$

To finish we consider the estimation of the mutual information between the output of the system for two different experiments q and q^*. We have

$$\begin{aligned} I(X_q, X_{q^*}) &= \iint p(x_q, x_{q^*}) \log \frac{p(x_q, x_{q^*})}{p(x_q)p(x_{q^*})} \, dx_q \, dx_{q^*} \\ &= \iiint p(x_q|\theta)p(x_{q^*}|\theta)p(\theta) \left(\log p(x_q, x_{q^*}) - \log p(x_q) - \log p(x_{q^*})\right) d\theta \, dx_q \, dx_{q^*} \end{aligned}$$

where we used the fact that X_q and X_{q^*} are independent conditionally to a parameter θ. This equation leads to a Monte Carlo estimation from a N-sample $\{\theta^{(i)}\}_{1 \leq i \leq N}$ drawn from the prior given by

$$\begin{aligned} I(X_q, X_{q^*}) \approx \frac{1}{N_1} \sum_{i=1}^{N_1} \Bigg[&\log \left(\frac{1}{N_2} \sum_{j=N_1+1}^{N_1+N_2} p(x_q^{(i)}|\theta^{(j)}) p(x_{q^*}^{(i)}|\theta^{(j)}) \right) \\ &- \log \left(\frac{1}{N_3} \sum_{k=N_1+N_2+1}^{N_1+N_2+N_3} p(x_q^{(i)}|\theta^{(k)}) \right) \\ &- \log \left(\frac{1}{N_4} \sum_{l=N_1+N_2+N_3+1}^{N} p(x_{q^*}^{(i)}|\theta^{(l)}) \right) \Bigg], \end{aligned}$$

where $N_1 + N_2 + N_3 + N_4 = N$.

For the repressilator, we assume measurement noise ϵ with variance $\sigma^2 = 5$. To estimate the mutual information between the output and the parameter, we use $N_1 = 100000$ and $N_2 = 4500000$. The mutual information between the output of the system and each parameter in the Hes1 example is computed using $\sigma^2 = 0.5$, and $N_1 = N_2 = N_3 = 5000$. For the AKT model, the variance of the measurement noise is equal to 10, $N_1 = 1000$, and $N_2 = N_3 = N_4 = 4500$.

5.2.5 Approximate Bayesian Computation (ABC).

Once data x^* have been collected, we use the Approximate Bayesian Computation (ABC) framework to infer the posterior parameter distribution $p(\theta|x^*)$. We introduced this concept in chapter 2.2.7. To summarise, ABC is a simulation-based method, which mainly consists of sampling the parameter space from a prior distribution $p(\theta)$, simulating the system for each sampled parameter (often called particle) and selecting the particles such that the simulated data are less than some maximal

distance away from the observed data. Those particles define an estimate of the posterior distribution given the observed data:

$$p(\theta|x^*) \approx \frac{p(\theta)\int_x p(x|\theta)1_{\Delta(x,x^*)\leq\delta}\,dx}{p(x^*)}. \tag{5.12}$$

Specifically we use an ABC scheme based on sequential Monte Carlo (SMC) (see chapter 2.2.7), which has been developed for likelihood-free parameter inference in deterministic and stochastic systems [111]. We use the implementation of this method in the package $ABC-SysBio$ [165].

5.2.6 Estimation of the entropy

The estimation of entropy has been performed only to test and confirm our experimental choice, which is based on Monte Carlo estimation of mutual information. For each experiment q we compute the difference between the entropy $H(\Theta)$ of the prior distribution $p(\theta)$ and the entropy $H(\Theta|x_q)$ of the posterior distribution $p(\theta|x_q)$. The entropies $H(\Theta)$ and $H(\Theta|x_q)$ are approximated using a histogram-based estimator. This discretization of the parameter space leads to a change of scale in the entropy measure. This explains why the scales of the differences between estimated entropies and the estimated mutual information differ despite the fact that the mutual information $I(\Theta, X_q)$ is the expectation over all possible data, x_q, of the difference between $H(\Theta)$ and $H(\Theta|x_q)$. It is also well known that such a histogram approach leads to a biased estimate of the entropy [174]. However, since the bias only depends on the number of bins and the sample size, we can compare the estimation results among experiments as long as these algorithm parameters are kept the same.

To compute the entropy $H(\Theta|x_q)$ for each experiment q in the repressilator example we compute a 4-dimensional histogram to discretise the posterior distribution (for all 4 model parameters) using 100 bins for each dimension resulting in a total of 10^8 bins. We use the R package $entropy$ [175] to estimate the entropy. For the $Hes\,1$ model we computed histograms over the marginals posterior distribution, to measure the entropy of each parameter separately. Here we used 1000 bins.

5.2.7 *Experimental data*

The experimental datasets used to investigate the Akt model were collected and published by the lab of S. Kuroda. The data are normalised Western blot measurements as described in [158].

5.3 RESULTS

5.3.1 *Information Content of Experimental Data*

To achieve their full functionality mathematical models require parameter values which generally need to be inferred from experimental data. The extraction of this information is, however, a nontrivial task and is further compounded by the need to assess the statistical confidence of such parameter estimates. In chapter 2.2 we introduced the concept of Bayesian inference. We will now link the ideas of Bayesian inference to the concept of experimental design and refer to the theory introduced in chapter 2.2.

We seek to evaluate the conditional probability distribution, $p(\theta|x)$, which relates to the prior knowledge $p(\theta)$ and the distribution of data, x, given parameters, $p(x|\theta)$. This is achieved via Bayes theorem (see also equations 2.26 and 2.27)

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}. \tag{5.13}$$

The probability density function, $p(\theta|x)$ describes the probability of finding a parameter θ in the volume element $d\theta$ of parameter space, given the data, the model and the prior information. Finding the posterior probability distribution $p(\theta|x)$ is usually achieved by means of powerful (if costly) computational algorithms such as Markov chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC) methods.

Rather than providing a single parameter estimate the posterior distribution allows us to assess how well a parameter is constrained by data (see figure 5.1 A). More formally, we measure the uncertainty about a parameter information-theoretically in terms of the Shannon entropy [135],

$$H(\Theta) = -\int p(\theta)\log(p(\theta))d\theta, \tag{5.14}$$

for the prior and

$$H(\Theta|x) = -\int p(\theta|x)\log(p(\theta|x))d\theta, \tag{5.15}$$

for the posterior. The information gained by collecting data x can then be expressed as the information gain from prior to posterior distribution $H(\Theta) - H(\Theta|x)$. The output of the experiment, however, is in turn "random" with distribution $p(x)$, and therefore the average posterior uncertainty is

$$H_X(\Theta|X) = \int H(\Theta|x)p(x)dx \tag{5.16}$$

which leads to the average information gain called mutual information between X and Θ,

$$I(X,\Theta) = H(\Theta) - H_X(\Theta|X). \tag{5.17}$$

When faced with different experimental setups, q, and hence different datasets, X_q, choosing the set(s) which maximise $I(X_q,\Theta)$ will provide the best insights into the system via improved parameter estimates. Note that in the general concept of optimal design, the aim is to maximise the utility function, which can be directly related to mutual information as we demonstrated in chapter 2.3.4. This observation is the basis of our experimental design methodology, which consists of computing the mutual information $I(X_q,\Theta)$ for every experiment q and selecting the experiment resulting in the highest mutual information (see chapter 5.2 for computational details). Once the chosen experiment has been carried out, the new data are used to update the model and the posterior distribution of the parameters (see figure 5.1 B).

Given the importance of the predictive role of mathematical modelling it is also of interest to reduce the uncertainty of model predictions; intriguingly and counter-intuitively — but demonstrably and provably — better parameter estimates are not necessarily required for better, more secure model predictions. We can thus also seek to identify the experimental condition q^* which maximises the predictive power of the model under (potentially very different) circumstances to predict new data Y. Analogously to the previous case minimising uncertainty in predictions of Y means to maximise mutual information between X and Y (see chapter 5.2):

$$I(X,Y) = H(Y) - H(Y|X) \tag{5.18}$$

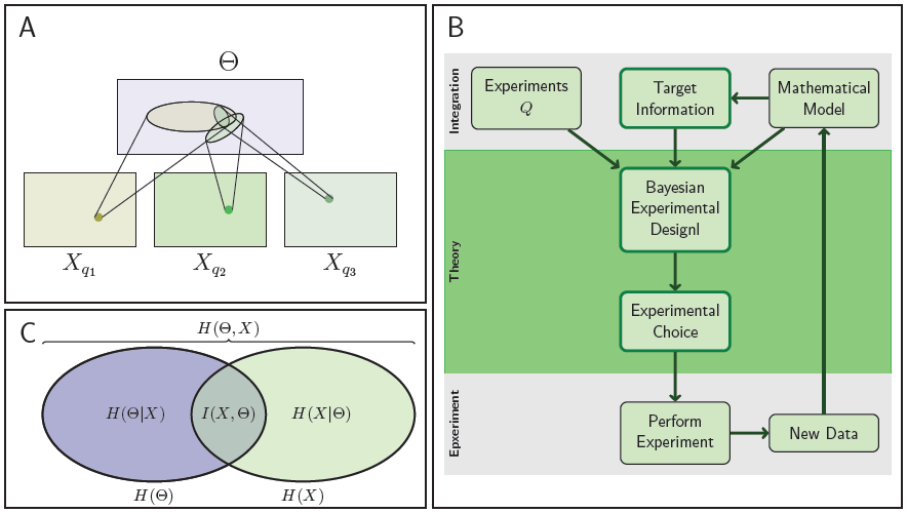


Figure 5.1: Information content of experimental data and flow chart of the experimental design method. (A) The regions of plausible parameters values for three different experiments. Each ellipse defines the set of parameters which are commensurate with the output $\mathsf{X_q}$ of an experiment q. In this example, the data $\mathsf{X_{q_3}}$ leads to the most precise inference of the parameters. The parameters which explain the output of all the three experiment are at the intersection of the three ellipsoids. (B) Flowchart of the experimental design method. Given a mathematical model of the biological system, a set of experiments and the target information —which can be either a set of parameters to infer or a description of the experiment to predict — the Bayesian Experimental Design method determine the experiment to carry out. Once the experiment has been performed, the experimental data are then used to provide target information and to improve the model. Thereafter, the process can be iterated to select other experiments in order to improve the accuracy of the target information. (C) Link between the total and conditional entropies and the mutual information of experimental data X and parameters θ.

5.3.2 Implementation and Validation

The algorithms to estimate the mutual information (computation of the Monte Carlo estimates) were implemented in *Python*. Because a high number (N^2 and N^4) of pairwise probabilities needs to be computed, which are all independent of each other, the algorithm was implemented in a parallel manor using graphical processing units (GPUs). The coupling to *Python* was achieved by using the package *Pycuda*. The numerical solutions of the models were obtained using the solver for ordinary differential equations of the package $\mathtt{cuda-sim}$ [176], which allows for parallel implementation on graphical processing units (GPUs).

We use three representative models to test our implementation. For all three

models we can exactly calculate the expected mutual information analytically, which allows us to determine the accuracy of our implemented algorithm.

Mutual information between parameter space and model system output

We define model 1 as

$$X = \Theta + \epsilon, \tag{5.19}$$

where Θ is the model parameter, and ϵ is a realisation of noise sampled from a normal distribution $N(0, \sigma_\epsilon^2)$. The form of this model corresponds to equation 5.3. We define the prior distribution of Θ as a normal distribution $N(0, \sigma_\Theta^2)$. We now need to determine $I(\Theta, X)$, which is the mutual information between Θ and X. We can write equation 5.17 as:

$$I(\Theta, X) = H(\Theta) + H(X) - H(\Theta, X), \tag{5.20}$$

Since Θ and X are normally distributed we have

$$H(\Theta) = \frac{1}{2}(\ln(2\pi\sigma_\Theta^2) + 1), \tag{5.21}$$

$$H(X) = \frac{1}{2}(\ln(2\pi\sigma_\Theta^2 + \sigma_\epsilon^2) + 1) \tag{5.22}$$

and

$$H(\Theta, X) = \frac{1}{2}\ln((2\pi e\sigma_\Theta + \sigma_\epsilon)^2). \tag{5.23}$$

Using equation 5.20 we obtain

$$I(\Theta, X) = \frac{1}{2}\ln\left(1 + \frac{\sigma_\Theta^2}{\sigma_\epsilon^2}\right) \tag{5.24}$$

We evaluated analytically the mutual information between the parameter Θ and the model output X using equation 5.24 for given σ_Θ and σ_ϵ and compared it with the estimates of our algorithm. Figure 5.2A shows that the estimates correspond to the exact solution for a wide range of σ_ϵ.

Mutual information between a specific parameter and model system output

We tested the algorithm to estimate the mutual information between a specific parameter of the system Θ_s and the model output X with the following model 2:

$$X = \Theta_1 + \Theta_2 + \epsilon. \tag{5.25}$$

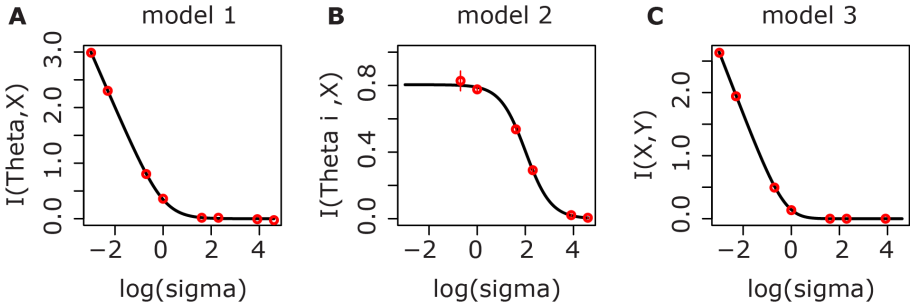


Figure 5.2: Validation of the implemented algorithms. Shown are the theoretical solutions of the mutual information according to equation 5.24 (A), equation 5.26 (B) and equation 5.29 (C) (black lines) and the mean estimates resulting from our implemented algorithms (red circles). Red lines in (B) are the standard deviation of 5 repeated estimations for model 2, in (A) and (C) standard deviation were too small to be shown. The standard deviations of the normal distributions were chosen as $\sigma_\Theta = 1.0$; $\sigma_{\Theta_1} = 10.0$; $\sigma_{\Theta_2} = 5.0$; $\sigma_{m_1} = 0.01$ and $\sigma_{m_2} = 0.001$.

We want to determine $I(\Theta_1, X)$. As in model 1 we assume that ϵ is a realisation of noise sampled from $N(0, \sigma_\epsilon^2)$ and Θ_1 and Θ_2 are as well normally distributed with $N(0, \sigma_{\Theta_1}^2)$ and $N(0, \sigma_{\Theta_2}^2)$, respectively. Following the same principle as in equations 5.20 - 5.23 we obtain:

$$I(\Theta_1, X) = \frac{1}{2} \ln \left(\frac{\sigma_{\Theta_1}^2 + \sigma_{\Theta_2}^2 + \sigma_\epsilon^2}{\sigma_{\Theta_2}^2 + \sigma_\epsilon^2} \right). \tag{5.26}$$

The solution of this equation for a given $\sigma_{\Theta_1}^2$ and $\sigma_{\Theta_2}^2$ is in accordance with the estimated values (figure 5.2B). It is important to notice that for very small values of σ_ϵ^2 it is necessary to increase the number of particles N to obtain stable estimates. This becomes computationally very expensive, because the total number of probabilities to estimate is N^4 (as compared to N^2 for the other two implemented algorithms).

Mutual information between two different model system outputs

In this last case we tested the estimation of mutual information between the outputs X and Y of two different systems, which share some parameter Θ. We define the test model as:

$$X = \Theta + m_1 + \epsilon \text{ and } Y = \Theta + m_2 + \epsilon, \tag{5.27}$$

where m_1 and m_2 are normal distributed random variables with mean 0 and variance $\sigma_{m_1}^2$ and $\sigma_{m_2}^2$, respectively. The noise term ϵ and the parameter Θ follow as well a

normal distribution, $N(0, \sigma_\epsilon^2)$ and $N(0, \sigma_\Theta^2)$, respectively. We can then obtain I(X,Y) by using equation 5.18, which results in:

$$I(X, Y) = H(X) + H(Y) - H(X, Y). \tag{5.28}$$

The resulting mutual information is

$$I(X, Y) = \frac{1}{2} \ln \left(\frac{(\sigma_\Theta^2 + \sigma_{m_1}^2 + \sigma_\epsilon^2)(\sigma_\Theta^2 + \sigma_{m_2}^2 + \sigma_\epsilon^2)}{(\sigma_\Theta^2 + \sigma_{m_1}^2 + \sigma_\epsilon^2)(\sigma_\Theta^2 + \sigma_{m_2}^2 + \sigma_\epsilon^2) - \sigma_\Theta^4} \right). \tag{5.29}$$

Figure 5.2C shows that our estimation results are in agreement with the analytical solution of equation 5.29.

In the following section we use three examples of various complexity to show how this combination of rigorous Bayesian and information theoretical frameworks allows us to design/choose optimal experimental setups for parameter/model inference and prediction, respectively.

5.4 APPLICATIONS

5.4.1 Experiment selection for parameter inference

To investigate the potential of our experimental design method for parameter estimation we first apply it to the repressilator model, a canonical model for gene regulatory systems. The repressilator model was first introduced by Elowitz and Leibler in 2000 [177]. It consists of three genes connected in a feedback loop, where each gene transcribes the repressor protein for the next gene in the loop (see figure 5.3 A and B). The deterministic model describes the change of mRNA and protein per cell as a function of time. In order to reduce the dimensionality of the model, time is rescaled in units is rescaled in units of the mRNA lifetime. As a result, all parameter in this model are unitless, and the abundance of mRNA and protein is measured in arbitrary units (AU).

To infer the parameters of this model, h, α, α_0, and β, we propose 5 sets of possible experiments: the wild type experiment (set 1) which is described in figure 5.3 A and corresponds to the ordinary differential equations in figure 5.3 B, and 4

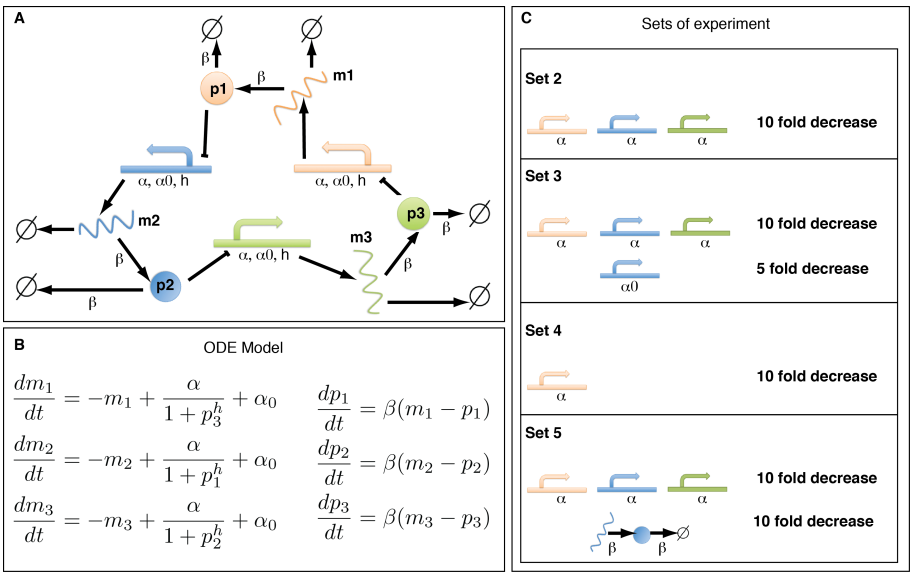


Figure 5.3: The repressilator model and the set of possible experiments. (A) Illustration of the wild type repressilator model. The model consists of 3 mRNA species and their corresponding proteins (shown in the same colour). (B) The ordinary differential equations which describes the evolution of the concentration of the mRNAs and proteins over time. (C) Four potential modifications of the wild-type model. For each experimental intervention the modified parameters are listed (colours are as in A). The modifications of the wild-type model consist of decreasing one or several of the parameters of the system: in sets 2, 3 and 5, the regime of the parameter α is changed; in sets 3, 4 and 5, respectively, parameters α_0, α and β are modified for only one gene which breaks the symmetry of the system.

modifications of the wild type model, see figure 5.3 C. These modifications can lead to different dynamics, and this in turn can lead to a higher mutual information between the parameters and the output of the system: the information content increases as differences in the outputs resulting from different parameter values increase. In figure 5.4, we illustrate the link between the increase in mutual information and the dynamics of the system for three different regimes. The mutual information $I(x, \theta)$ depends on the dynamics of the system given the prior of the system parameters: the more the dynamics for different parameter values differ from each other, the higher is the information content. To visualise this we compute the mutual information between one parameter and the outcome of the system for three different regimes in the repressilator example. Noting that α_0 is a bifurcation parameter and $\alpha_0 = 2.55$ is a Hopf bifurcation point (for the remaining parameters held at $h = 2, \alpha = 1000$

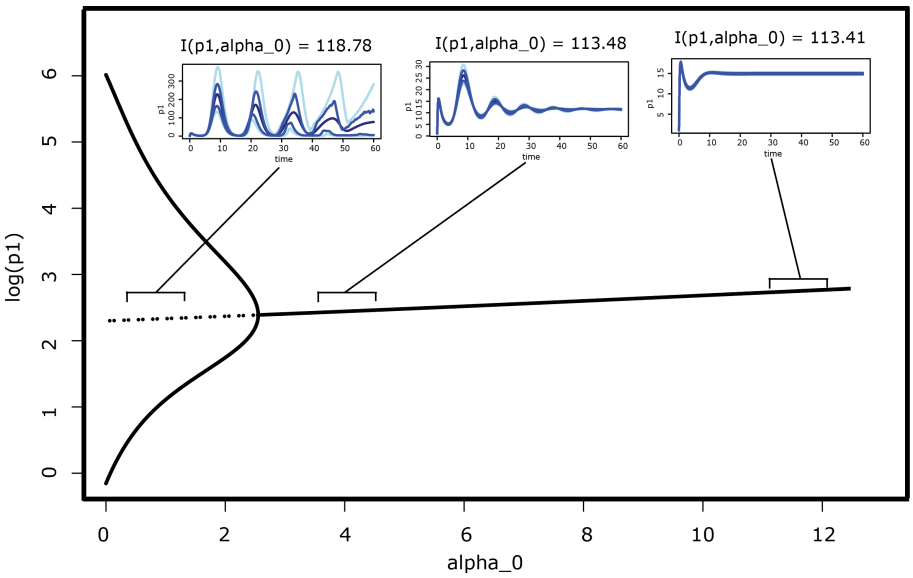


Figure 5.4: Information content of different parameter regimes. Shown is the bifurcation diagram for parameter α_0 with its stable (solid lines) and unstable (dashed lines) states. Estimation of mutual information was performed for 3 different parameter regimes. For illustration we plot the mean (dark blue), 25 and 75 percentiles (blue) and the 5 and 95 percentiles (light blue) of trajectories simulated with 10000 parameter sets, where α_0 is uniformly sampled and the remaining parameters are kept constant ($h = 2$, $\alpha = 1000$ and $\beta = 5$).

and $\beta = 5$), we choose 3 different prior regimes for α_0: $\pi_1(\alpha_0) \sim U(0.5, 1.5)$, $\pi_2(\alpha_0) \sim U(3, 4)$ and $\pi_3(\alpha_0) \sim U(10, 11)$. We keep the remaining 3 parameters constant: $h = 2$, $\alpha = 1000$ and $\beta = 5$. For these three priors we estimate the mutual information $I(x, \alpha_0)$ and represent the dynamics of the output of the system. We observe that the dynamics resulting from the first prior regime are most diverse and therefore $I(x, \alpha_0)$ has the highest value (118.78 ± 2.2710^{-3}) compared to the remaining two parameter regimes (113.48 ± 2.4210^{-3} for $\pi_2(\alpha_0)$ and 113.41 ± 4.9210^{-4} for $\pi_3(\alpha_0)$).

To determine which experiment to carry out, we compute the mutual information between the parameter prior distribution and the system output via Monte-Carlo estimation. We use uniform priors over $[1, 10]$ for h, over $[0, 20]$ for α_0, over $[500, 2000]$ for α and over $[0, 10]$ for β. Figure 5.5 shows that experiment 2 and 5 have highest mutual information, i.e. carrying out those experiments will decrease the uncertainty in the parameter estimates most. To confirm this we simulate data for the 5 experiments using the parameter $(h^*, \alpha^*, \alpha_0^*, \beta^*) = (2, 10, 1000, 5)$; see figure 5.8. Based on these data we perform parameter inference using an approximate Bayesian com-

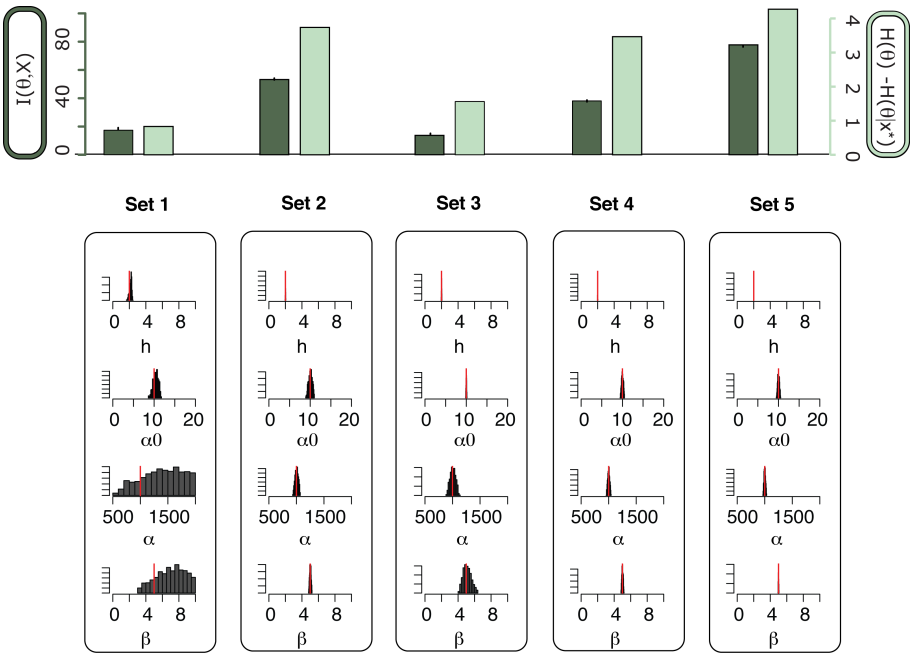


Figure 5.5: Experiment choice for parameter inference in the repressilator model. Top: The mutual information $I(\theta, X)$ between the parameters θ and the output of each set of experiment (in dark green), and the entropy difference between the prior distribution and the posterior distribution for a data obtained from simulation of the system for each experiment (in light green). The error bars on the mutual information barplots show the variance of the mutual information estimations over 3 independent simulations. Bottom: For each set of experiment we show the histogram of the marginal of the posterior distribution of every parameters. The red line indicates the true parameter value.

putation approach [111] for each experiment separately and compare the posterior distributions shown in figure 5.5. We observe that using the data generated from set 1 (wild type) only 2 parameters can be inferred: h and α_0. By contrast, the data generated by set 2 and set 5 allow us to estimate all 4 parameters with high confidence. A more detailed representation of the posterior distribution for each experimental set is given in figure 5.6.

In addition, for each experiment we compute the reduction of uncertainty from the prior to the posterior distribution. The results are consistent with the results using mutual information and confirm that we should choose experiment 2 or 5 for parameter inference. In practice not all molecular species may be experimentally accessible and it is therefore also of interest to decide which species carries most

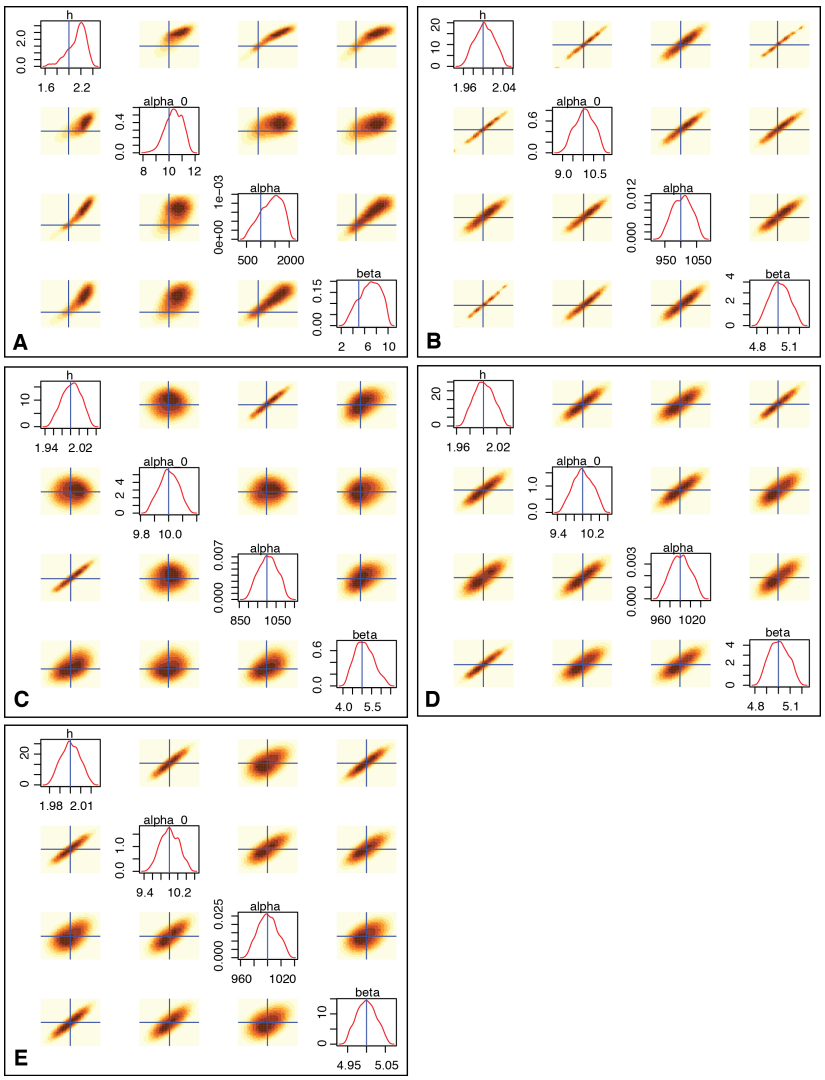


Figure 5.6: Surface plots of the estimated posterior distributions. The posterior distribution is shown given the data represented in Figure 5.8. Each subfigure (A to E) corresponds to an experiment (1 to 5). In each subfigure, the diagonal represents the marginal posterior distribution for each parameter and the off-diagonal elements show the correlations between pairs of parameters.

information about the parameters. We can estimate the mutual information between the parameter and each species independently, and, for example, for experimental set 5 we observe that mRNA m_1 and m_2 as well as protein p_1 carry equally high information; see figure 5.7.

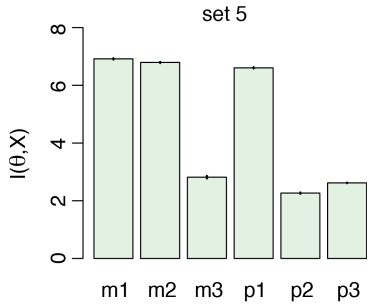


Figure 5.7: Species dependent mutual information measurements. The mutual information between the parameters and each species (3 mRNA and 3 protein measurements) is measured for experiment 5, which had the highest total mutual information.

5.4.2 Experiment selection to infer a specific parameter

Sometimes we are interested in estimating only some of the parameters, e.g. those that have a direct physiological meaning or are under experimental control. To investigate this aspect we consider the Hes1 transcription factor that plays a number of important roles, including in the cell differentiation and segmentation of vertebrate embryos. In 2002 oscillations in the Hes1 system were observed by [178] and such oscillations might be connected with formation of spatial patterns during development. The Hes1 oscillator can be modelled by a simple three-component ODE model [179] as shown in figure 5.9 A. This model contains 4 parameters, k_1, P_0, ν, and h, and 3 species: Hes1 mRNA, m, Hes1 nuclear protein, p_1, and Hes1 cytosolic protein, p_2.Hes1 proteins and Hes1 mRNA are measured as fold-change. Similar to the repressilator model, the time is rescaled in units of the mRNA birth rate, therefore all parameters in this model are dimensionless quantities. It is possible to measure either the mRNA (using real-time PCR) or the total cellular Hes 1 protein concentration $p1 + p2$ (using Western blots). We investigate whether protein or mRNA measurements provide more information about the model parameters. Thus we estimate the mutual information between mRNA and parameters, and between protein and parameters. Figure 5.9 B shows that mRNA measurements carry more information about all of the parameters.

This can again be further substantiated by simulations shown in figure 5.10. We perform parameter inference based on such simulated data and compute the difference between the entropy of the prior and that of the resulting posterior distribution.

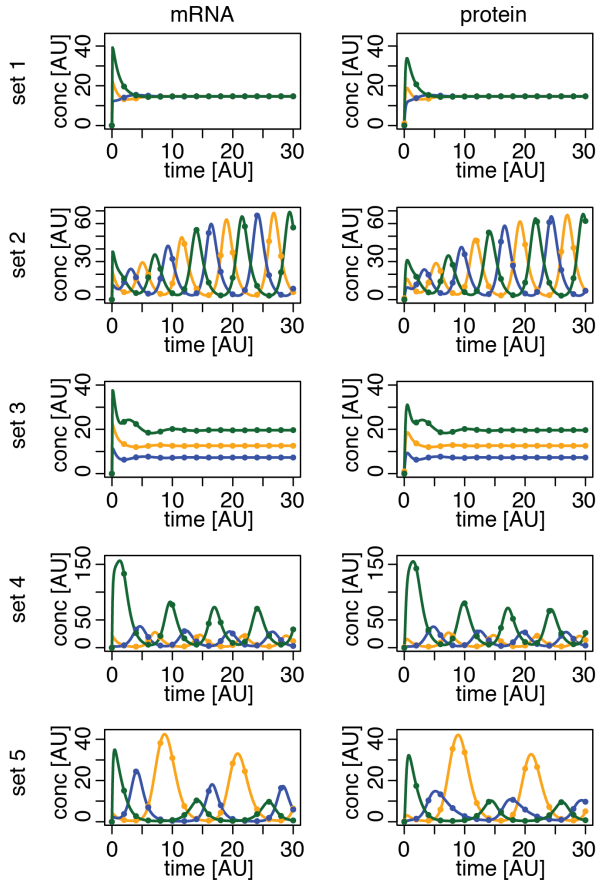


Figure 5.8: Repressilator mRNA and protein concentration time course for each experimental setup. The parameter vector used for simulations is $(h^*, \alpha^*, \alpha_0^*, \beta^*) = (2, 10, 1000, 5)$. The colours correspond to those in figure 5.4. The dots represent the simulated data and the lines corresponds to the mean of the species for 1000 parameters sampled from the posterior distribution computed using ABC SMC.

The results shown in figure 5.9 C are consistent with the predictions based on mutual information: mRNA measurements carry more information for parameter inference. Interestingly, however, although the mutual information computation indicates that the protein measurements should contain more information about parameter k_1 than about the other parameters, this is not confirmed by the difference in entropy result for this simulated dataset. This divergence is due to the fact that the mutual information measures the amount of information contained *on average* over all the possible behaviours of the system whereas figure 5.9 C represents the decrease in entropy from the prior to the posterior distribution given *specific data*. The differences

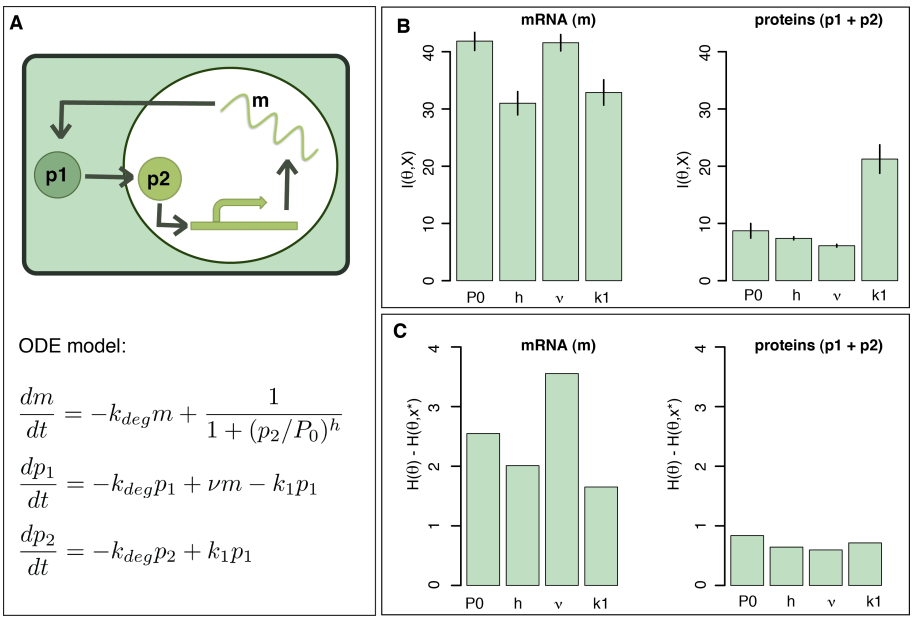


Figure 5.9: Experiment selection for parameter inference in the Hes1 model. (A) Diagram of the Hes 1 model and the associated ordinary differential equations. (B) The mutual information between each parameters and the output of the system for repectively mRNA measurement (left) and total cellular Hes1 measurement (right). The barplot represents the mean and the variance over 3 repetition of the Monte-Carlo estimation. (C) Estimation of the difference between the entropy of the parameter prior and the entropy of the posterior distribution given one dataset. The parameter k_{deg} was set to be 0.03 as experimentally determined by [178].

in entropy for other datasets simulated using different parameter regimes are thus in better agreement with the mutual information results (see figure 5.11).

5.4.3 Experiment selection for prediction

We next focus on a scenario where we aim to predict the behaviour of a biological system [180] under conditions for which it is not possible to obtain direct measurements. We consider as an example the phosphorylation of Akt and ribosomal binding protein S6 in response to a epidermal growth factor (EGF) signal. The Akt pathway regulates many molecular processes, e. g. cell survival, growth and proliferation. These mechanisms are controlled by growth factors. Especially in tumour development, the Akt pathway is a potential target for drug development [181–185]. In this system we analyse data that result from PC12 cells that undergo cell proliferation and differenti-

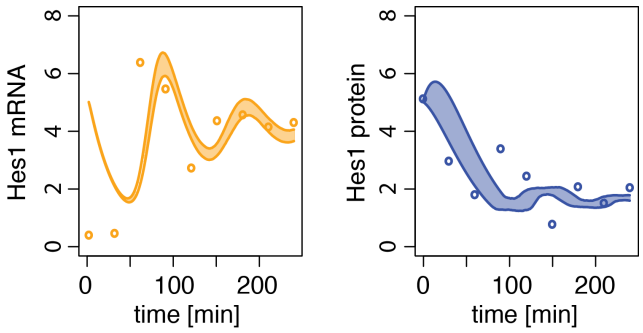


Figure 5.10: Simulation results of Hes1 model. Simulated trajectories of the mRNA and protein concentrations (dots). The parameter used for simulation is $(P_0, h, \nu, k_1) = (1.4, 5.7, 0.02, 0.09)$. The lines represent the 5% and 95% percentiles of the species abundances for 1000 parameters sampled from the posterior distribution computed using ABC SMC.

ation in response to epidermal growth factor (EGF). Depending on the Akt pathway both cell proliferation and differentiation require an increase in cell size. Binding of EGF to its receptor (EGFR) leads to phosphorylation of Akt via phosphatidylinositol 3-kinase (PI3K). Akt is wired to the mTOR complex 1 that phosphorylates and activates the ribosomal protein S6 kinase (S6K). The latter one phosphorylates ribosomal protein S6. Figure 5.12 A shows the pathway of interest: the EGF growth factor binding to the activated receptor EGFR leads to phosphorylation of EGFR and a signal cascade, which results in the phosphorylation of Akt (pAkt) which in turn can activate downstream signalling cascades and leads to the phosphorylation of S6 (pS6); a corresponding mathematical model is shown in Figure 5.13 [158]. The system is modelled using ODE's, which result from first-order mass action kinetics. That means, for the reaction $A + B \rightarrow C$ with rate k, the change of A and B over time is described by

$$\frac{dA}{dt} = \frac{dB}{dt} = -kAB, \tag{5.30}$$

while the change of C is modelled as

$$\frac{dC}{dt} = +kAB. \tag{5.31}$$

We are interested in predicting the dynamics of the Akt system under multiple pulsed stimuli with EGF in the presence of background noise, as shown in figure 5.15 A. We consider 5 pulses of intensity 1 ng/ml and length 60 seconds spaced by 400 seconds with additive background noise. This input is difficult to realise in an

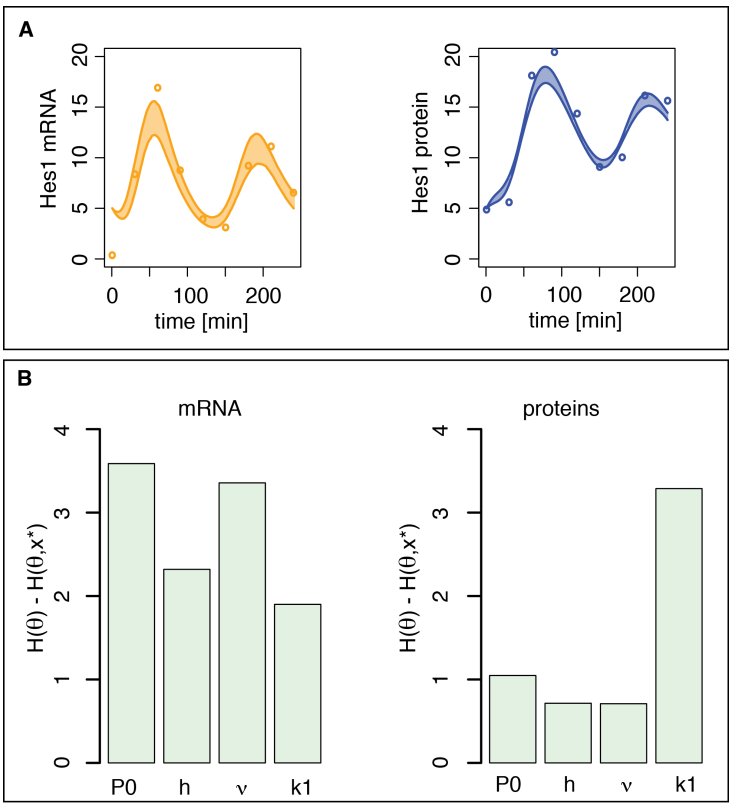


Figure 5.11: (a) Simulated trajectories of the mRNA and protein concentration (dots) for the parameter $(P_0, h, \nu, k_1) = (1.25, 7.54, 5.36 \times 10^{-2}, 4.02 \times 10^{-3})$. The lines represent the 5% and 95% percentiles of the species abundances for 1000 parameters sampled from the posterior distribution computed using ABC SMC. (B) Estimates of the differences between the entropies of the prior and posteriors.

experimental system (let alone in an animal or clinical setting). Using an initial dataset, see figure 5.12 B, we can infer system parameters using ABC SMC (see figure 5.14). From the resulting posterior distribution we then sample 1000 parameter combinations and simulate the model with the 5-pulse-stimulus in order to predict the time courses of phosphorylated EGF receptor (EGFR), phosphorylated Akt and phosphorylated S6; based on just the estimated parameters these predictions are, however, associated with high uncertainty, see figure 5.15 B.

To obtain better predictions we can use data from other experiments measuring the time course of the 3 species of interest for a experimentally more straightforward input signals chosen from among 12 possible stimuli: impulse, step or ramp stimuli with different EGF concentrations (see figure 5.15). To determine which of

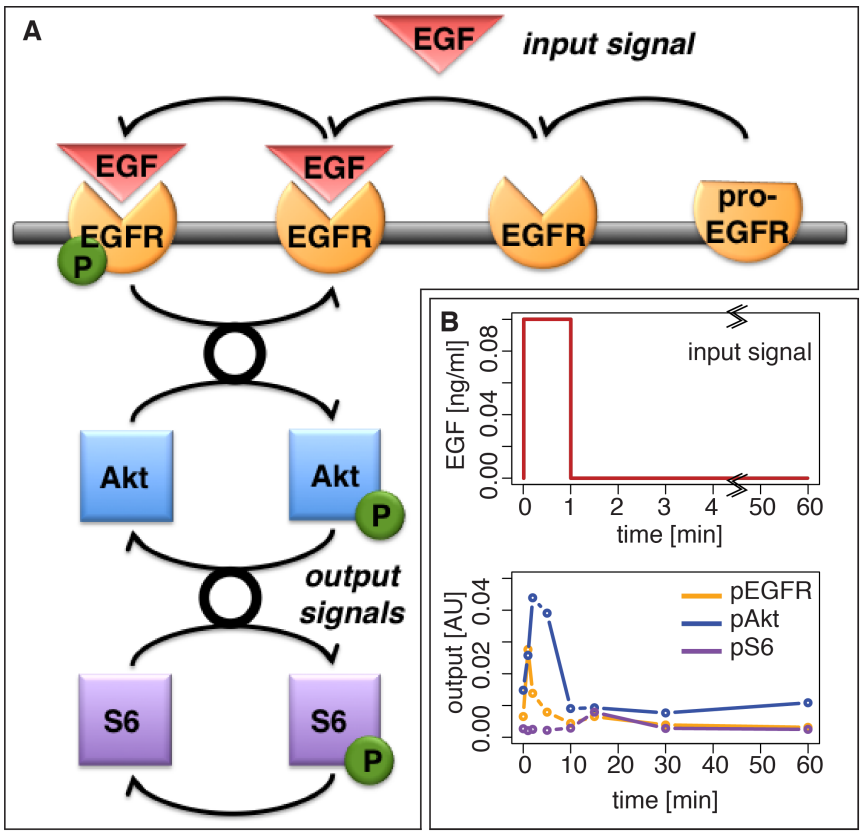


Figure 5.12: The EGF-dependent AKT pathway and an initial dataset. (A) Diagram of the model of the EGF-dependent AKT pathway. Epidermal growth factor (EGF, red triangle) is a stimulus for a signalling cascade, which results in the phosphorylation (green circle) of Akt (blue square) and S6 (purple square). EGF binds to the EGF membrane receptor EGFR (orange), which is generated from a pro-EGFR. The Binding results in the phosphorylation of the receptor, which consequently leads to the activation of downstream cascades (thick black circle). This simplified model was shown to capture the experimentally determined dynamics [158]. (B) A impulse input of EGF over 60 seconds with an intensity of 0.1 ng/ml (top) and the resulting time course of phosphorylated EGF receptor (pEGFR), phosphorylated Akt (pAKT) and phosphorylated S6 (pS6) in response to this stimulus (bottom). Data were provided by the authors of [158].

those inputs would result in the most reliable predictions we compute the mutual information between the time courses for the different potential experimental inputs and the time-course of the target 5-impulse noisy stimulus. We incorporate initial information about model parameters by computing the mutual information based on the posterior distribution inferred above. Figure 5.15 C shows that a step stimulus of

$$\frac{dy_1}{dt} = -p_1 y_0 y_1 + p_2 y_{10} + p_0 y_9 - p_0 y_1 + p_5 y_3$$
$$\frac{dy_2}{dt} = -p_3 y_2 y_4 + p_4 y_3 - p_6 y_2 + p_{12} y_{10}$$
$$\frac{dy_3}{dt} = p_3 y_2 y_4 - p_4 y_3 - p_5 y_3$$
$$\frac{dy_4}{dt} = -p_3 y_2 y_4 + p_4 y_3 + p_{10} y_5 + p_9 y_7$$
$$\frac{dy_5}{dt} = p_5 y_3 - p_7 y_5 y_6 + p_8 y_7 - p_{10} y_5$$
$$\frac{dy_6}{dt} = -p_7 y_5 y_6 + p_8 y_7 + p_{11} y_8$$
$$\frac{dy_7}{dt} = p_7 y_5 y_6 - p_8 y_7 - p_9 y_7$$
$$\frac{dy_8}{dt} = p_9 y_7 - p_{11} y_8$$
$$\frac{dy_{10}}{dt} = p_1 y_0 y_1 - p_2 y_{10} - p_{12} y_{10}$$

Figure 5.13: Ordinary differential equations of the Akt model. The equations describe the dynamics of the 11 species of the AKT model. The model contains 12 parameters denoted p_i, $1 \leq i \leq 12$. The concentration of the following species (in this order) are denoted by y_i, $0 \leq i \leq 10$: EGF, EGFR, pEGFR, pEGFR-AKT, AKT, pAKT, S6, pAKT-S6, pS6, pro-EGFR and EGF-EGFR.

intensity 3 ng/ml has the highest predictive power about the behaviour of our target stimulus pattern.

In figure 5.15 D we show that this does indeed yield vastly improved predictive power compared to the initial prediction. This increase in predictive performance results from the difference in the posterior distributions resulting from different stimuli; by focussing on predictive ability we focus implicitly on data that are informative about those parameters that will affect the system behaviour most under the target (5-pulse) stimulus. The posterior distributions are represented in figure 5.15 E for two parameters, the EGFR turn over and the EGFR initial concentration, which appear to be essential for the prediction of the evolution of Akt/S6 phosphorylation patterns under the 5-pulse stimulus. Those two parameters were not inferred using the initial

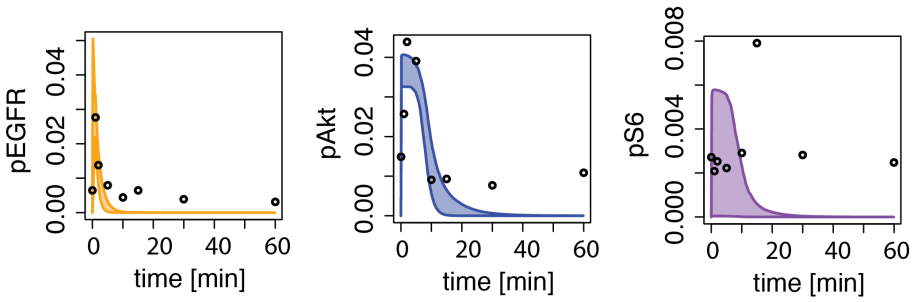


Figure 5.14: Fits of initial dataset. The time course of phosphorylated EGF receptor (pEGFR), phosphorylated Akt (pAKT) and phosphorylated S6 (pS6) in response to an impulse input of EGF over 60 seconds with an intensity of 0.1 ng/ml (dots). Data are Western blots measurements, described in [158]. The lines represent the 5% and 95% percentiles of the evolution of the species for 1000 parameters sampled from the posterior distribution computed using ABC SMC.

dataset alone, whereas the addition of the outcome of the step stimulus experiment suggested by our methods infers these parameters with the required high precision.

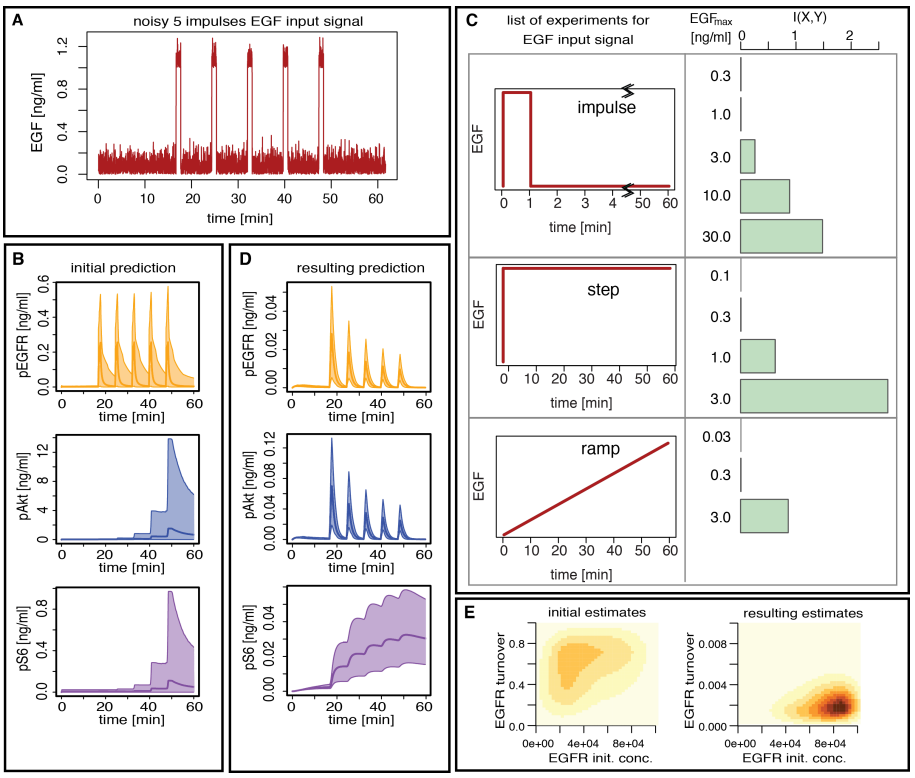


Figure 5.15: Experiment selection for prediction in the EGF-dependent AKT pathway. (A) The noisy 5-impulses EGF input signal: the 5 pulses are of intensity 1 ng/ml and length 60 seconds spaced by 400 seconds with an additive background noise which is the absolute value of a gaussian white noise of variance $\tau^2 = 0.1$. (B) The predicted time course of the proteins pEGFR, pAKT and pS6 under the noisy 5-impulses EGF input signal based on the intial dataset. (C) The mutual information between the time course of the 3 species of interest under the noisy 5-impulses EGF input signal and the time course of the species under each of the following 12 possible experiments: an impulse stimuli of length 60 seconds with 5 possible intensity (0.3, 1,3, 10 and 30 ng/ml), a step stimuli of length 60 minutes with 4 possible intensity (0.1, 0.3, 1 and 3 ng/ml) and a ramp stimuli of length 60 minutes with 3 possible final intensity (0.03, 0.3 and 3 ng/ml). (D) The predicted time course of the proteins pEGFR, pAKT and pS6 under the noisy 5-impulses EGF input signal based on the outcome of the step stimuli with intensity 3 ng/ml, which is the experiment with the highest mutual information. The scale of the y-axis is different for figures (B) and (D). (E) The posterior distribution of two parameters (EGFR turnover and EGFR initial condition) when using the initial dataset alone (left) and when using the initial dataset and the outcome of the step stimuli with intensity 3 ng/ml (right). The scale of the EGFR turnover is the prior range for the figure on the left panel whereas it is 100 times smaller for the figure in the right panel.

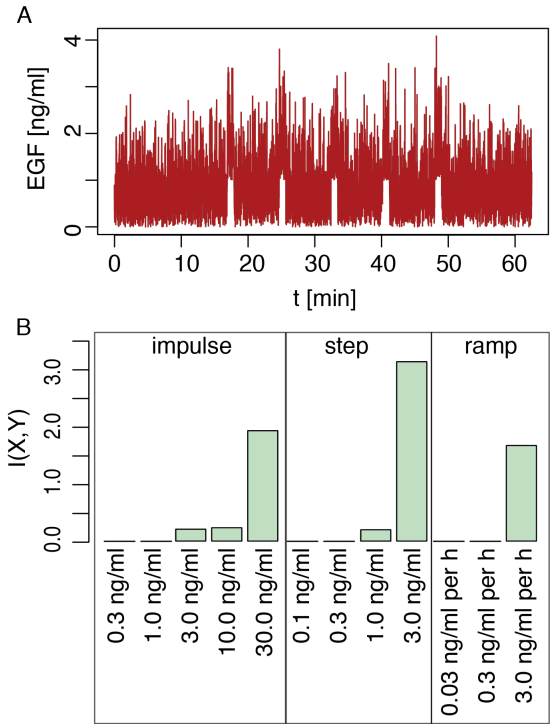


Figure 5.16: Experiment selection for prediction of a signal with high level of noise. (A) A noisy 5-impulses EGF input signal: the 5 pulses are of intensity 1 ng/ml and length 60 seconds spaced by 400 seconds with an additive background noise which is the absolute value of a gaussian white noise of variance $\tau^2 = 1$. (B) The mutual information between the time course of the 3 species of interest under the noisy input signal represented in (A) and the time course of the species under each of the following 12 possible experiments: an impulse stimulus of length 60 seconds with 5 possible intensity (0.3, 1,3, 10 and 30 ng/ml), a step stimulus of length 60 minutes with 4 possible intensity (0.1, 0.3, 1 and 3 ng/ml) and a ramp stimulus of length 60 minutes with 3 possible final intensity (0.03, 0.3 and 3 ng/ml).

This predictive power even extends to much greater signal distortion and even with a noise level of 100 percent of the signal intensity (see figure 5.16 A) we find that our experimental design method yields similar improvements in the predictions (see figure 5.16 and 5.17). We observe that the direct target (EGFR receptor) as well as activated AKT (pAKT) efficiently filter out the noise but capture the 5 pulses; EGFR activation quickly returns to base level in response to the higher frequency background noise. This indicates that there might be a constant low concentration of activated EGFR (pEGFR), but the activation of S6 has very different characteristics and is far less robust to noise. The level of noise is amplified as can be observed in the

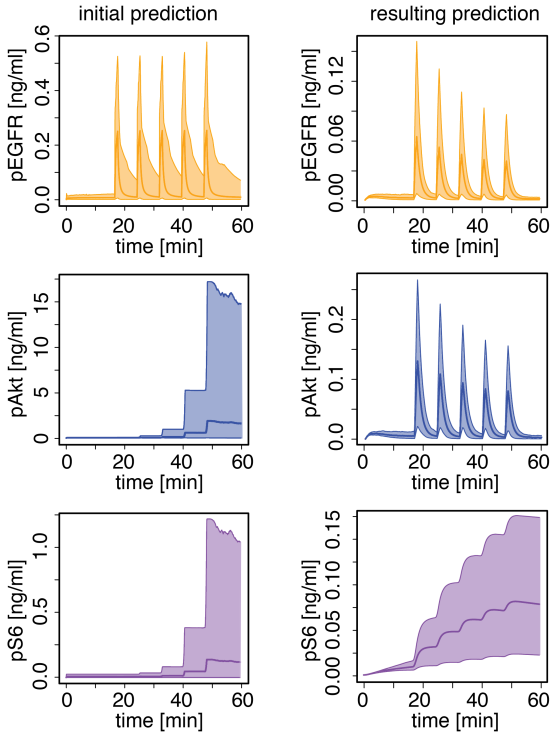


Figure 5.17: Prediction of system response to highly noisy input signal. The predicted time course of the proteins pEGFR, pAKT and pS6 under the noisy 5-impulses EGF input signal with a noise of high intensity represented figure 5.16 A. In the left panel, the prediction is based on the initial dataset whereas in the right panel in addition to the initial data it is also based on the outcome of the step stimulus with intensity 3 ng/ml, which is the experiment with the highest mutual information. The scale of the y-axis is different for each figure.

pS6 time course. This might suggest that the downstream molecule pS6 has a longer time delay to react to a signal. Moreover, pS6 does not have time to relax to its baseline between the 5 pulses, which leads to incremental signal amplification. This behaviour fits with the low-pass filter characteristics previously described [158]. In further support earlier studies [186] found that a downstream molecule can be more sensitive to an upstream activator than the direct target molecule of the activator. This might explains that the activation of EGFR and AKT is more robust to noise than the downstream molecule, S6.

5.5 DISCUSSION

We have found that maximising the mutual information between our target information — here either model parameter values or predictions of system behaviour — and the (simulated) output of potentially available experiments offers a means of arriving at optimally informative experiments. The experiments that are chosen from a set of candidates are always those that add most to existing knowledge: they are, in fact, the experiments that most challenge our current understanding of a system.

This framework has a number of advantages: First, we can simulate cheaply any experimental set-up that can in principle be implemented; second, using simulations allows us to propagate the model dynamics and to quantify rigorously the amount of (relevant) information that is generated by any given experimental design; third, our information measure gives us a means of meaningfully comparing different designs; finally, our approach can be used to design experiments sequentially — our preferred route as this will enable us to update iteratively our knowledge of a system along the way — or in parallel, i.e. selecting more than one experiment. Previous approaches had taken a more local approach [187–190] that relied on initial parameter guesses and often data; our approach also readily incorporates different stimulus patterns [191].

Here we have focussed on designing experiments that increase our ability to estimate model parameters or to predict model behaviour. The latter depends on model parameters in a very subtle way: not all parameters affect system output equally and under all conditions. But with optimal design we can overcome the problem of sloppy parameters [115] (which are, of course, dependent on the experimental intervention chosen [169]). By making predictive power the target of our analysis we directly home in onto those system parameters that are most relevant for making successful predictions under the target conditions. Equally, however, we could make model discrimination or checking the target of our analysis [191, 192], and, for example, choose experimental designs that maximise our ability to distinguish between competing alternative hypotheses or models. All of this is straightforwardly reconciled in the Bayesian framework which also naturally lends itself to such iterative procedures where “today’s posterior” is “tomorrow’s prior ”.

Experimental design in systems biology is different from classical experimental

design studies. The latter theory was first developed at a time when the number of alternative hypotheses was smaller than the amount of available data and replicates [193]. Systems biology, on the other hand is hypotheses rich and data rarely suffice to decide clearly in favour of one model unambiguously. Moreover for dynamical systems, as a host of recent studies have demonstrated, generally less than half of the parameters are tightly confined by experimental data [115, 169]. Together these two challenges have given rise to a number of approaches aimed at improving our ability to develop mechanistic models of such systems.

Our approach improves on previous methods [188–190, 194–198] in a number of ways: first we are able to incorporate but do not require preliminary experimental data; second, it is a global approach that is not limited to some neighbourhood in parameter space unlike approaches solely based on e.g. the Fisher information [188, 199]; third, we obtain comprehensive statistical predictions (including confidence, sensitivity and robustness assessments if desired); and we are very flexible in the type of information that we seek to optimize.

Conclusion

When analysing experimental data resulting from cell populations we often observe effects due to heterogeneity in the cell population. Various sources of heterogeneity exist and their relevance should be carefully explored. Here we have investigated the cell migratory behaviour of leukocytes during wounding. This system serves as a good example of how to analyse such data. The observed heterogeneity in cellular behaviour has several sources.

First of all we investigated a population of cells that share the same marker. The marked cells belong to different cell types, here macrophages and neutrophils. However, even when separating these two cell types, we still observe a high level of heterogeneity. Another source of variation in cellular behaviour is the extra-cellular environment. Spatial as well as temporal differences in the extracellular matrix, neighbouring cells and external signals affect the behaviour of each particular leukocyte and result in large variation across the entire leukocyte population. The presented study here investigates the spatio-temporal effects and elucidates the resulting changes in cell migration.

A further source of heterogeneity is the noise of the external signals, which guide the migratory behaviour of leukocytes. A noisy external signal (due to a non-homogenous extracellular matrix) needs to be integrated and translated into a cellular decision. This noisy component in part leads to the randomness of leukocyte migration. In this study it becomes clearly apparent that investigating the average cellular

response of an entire population provides little information about the biological system.

A detailed analysis of leukocyte migration using transition matrices enabled us to observe several random walk models. We detected previously described Brownian motion, the biased random walk and the persistent random walk. Furthermore we observe two novel types of random walk, which had previously not been described: the forward-backward random walk and the process of trafficking. The main argument throughout our study, however, is that leukocytes do not perform only one specific type of random walk. The walk changes depending on space and time, i.e. with he changing cellular environment and the changing internal signals. We observe mainly mixture models of several random walk types, such as the biased persistent random walk. These fine nuances of leukocyte migration were apparent, because the analysis with transition matrices does not require an initially assumed model of migration. This type of analysis is not specific for leukocyte migration, but can be applied to any kind of biological trajectory data.

The random walk observations describe processes on the cellular scale. However, using molecular interventions it is possible to connect the cellular level with the molecular scale. We demonstrated this for example, in the presented study of how specific inhibitors can modulate the cellular migratory behaviour. This specific example again serves as a case study, which allows the systematic analysis of signalling pathways connected to cell migration.

The information we gain by analysing experimental data is the basis on which we construct a mathematical model of the investigated system. Lack of information often leads to several hypothesis (or models) and it is an important challenge to discriminate between them. We employed a model selection algorithm in an approximate Bayesian framework (ABC SMC) to distinguish between three candidate models of signalling gradients that drive leukocyte migration during wounding. The novelty of our constructed cell migration model is the description of leukocyte recruitment in dependence of the location in the whole organism, i.e. the model takes into account spatial effects. These effects result from the spatio-temporal distribution of signalling molecules. This combines the molecular level with the cellular level under consideration of the whole organism. We apply for the first time the ABC SMC framework for model ranking to spatio-temporal models. A further level of complexity are the

random characteristics of biological trajectory data. Two trajectories generated by the same underlying random walk model can not be compared with the standard euclidian distance, but a sufficient summary statistic needs to be defined, which captures the characteristics of the underlying random walk process based on a set of short sample paths.

To increase the predictive power of a mathematical model it is vital to calibrate it using experimental data. However, different experiments contain different information about the investigated biological system. In this work we meld concepts from information theory and Bayesian inference to select experiments with the highest information content of the resulting data. Our global approach allows us to incorporate already gathered knowledge, i.e. prior information, in a flexible manner. The classical approaches in the field of experimental design are restricted to small and linear systems, which can be treated analytically. These approaches find little application in modern systems biology. The theoretical framework developed here is simulation-based and therefore applicable to a wide range of biological systems. We demonstrate how this approach is used to study complex systems such as the Akt signalling pathway. This systematic method to choose experiments will allow researchers to target experimental efforts optimally.

Bibliography

[1] Joyce, A. R. and Palsson, B. Ø. The model organism as a system: integrating omics data sets. *Nat Rev Mol Cell Biol* 7, 198–210 (2006).

[2] Walker, D. C., Georgopoulos, N. T., and Southgate, J. From pathway to population - a multiscale model of juxtacrine egfr-mapk signalling. *BMC Syst Biol* 2, 102–103 (2008).

[3] Bruggeman, F. and Westerhoff, H. The nature of systems biology. *Trends Microbiol* 15(1), 45–50 (2007).

[4] Nurse, P. and Hayles, J. The cell in an era of systems biology. *Cell* 144(6), 850–854, March (2011).

[5] Silver, P. and Way, J. Molecular systems biology in drug development. *Clin Pharmacol Ther* 82(5), 586–590 (2007).

[6] Spiller, D., Wood, C., Rand, D., and White, M. Measurement of single-cell dynamics. *Nature* 465(7299), 736–745 (2010).

[7] Del Sol, A., Balling, R., Hood, L., and Galas, D. Diseases as network perturbations. *Curr Opin Biotechnol* 21(4), 566–571 (2010).

[8] Beal, M., Falciani, F., Ghahramani, Z., Rangel, C., and Wild, D. A bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics* 21, 349–356 (2005).

[9] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D., and Nolan, G. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308(5721), 523–529 (2005).

[10] Lèbre, S., Becq, J., Devaux, F., Stumpf, M. P. H., and Lelandais, G. Statistical inference of the time-varying structure of gene-regulation networks. *BMC Syst Biol* 4, 130 (2010).

[11] Mendes, P. and Kell, D. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* 14(10), 869–883 (1998).

[12] Reinker, S., Altman, R. M., and Timmer, J. Parameter estimation in stochastic biochemical reactions. *IEE Proc Syst Biol* 153(4), 168–178 (2006).

[13] Kreutz, C. and Timmer, J. Systems biology: experimental design. *FEBS J* 276(4), 923–942 (2009).

[14] Vyshemirsky, V. and Girolami, M. A. Bayesian ranking of biochemical system models. *Bioinformatics* 24(6), 833–9, Mar (2008).

[15] Toni, T. and Stumpf, M. P. H. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics* 26, 104–110 (2010).

[16] Chaloner, K. and Verdinelli, I. Bayesian experimental design: A review. *Stat Sci* 10(3), 273–304 (1995).

[17] Janeway, C., Travers, P., Walport, M., and Capra, J. *Immunobiology: the immune system in health and disease*, volume 1. Current Biology, (2001).

[18] Parham, P. *The Immune System.* New York, Garland Science., 2 edition, (2005).

[19] Koh, T. J. and DiPietro, L. A. Inflammation and wound healing: the role of the macrophage. *Expert Rev Mol Med* 13, e23 (2011).

[20] Dale, D. C., Boxer, L., and Liles, W. C. The phagocytes: neutrophils and monocytes. *Blood* 112(4), 935–945 (2008).

[21] Brancato, S. K. and Albina, J. E. Wound macrophages as key regulators of repair: origin, phenotype, and function. *The American journal of pathology* 178(1), 19–25 (2011).

[22] Stein, M., Keshav, S., Harris, N., and Gordon, S. Interleukin 4 potently enhances murine macrophage mannose receptor activity: a marker of alternative immunologic macrophage activation. *The Journal of experimental medicine* 176(1), 287–292 (1992).

[23] Rappolee, D. A., Mark, D., Banda, M. J., and Werb, Z. Wound macrophages express tgf-alpha and other growth factors in vivo: analysis by mrna phenotyping. *Science* 241(4866), 708–712 (1988).

[24] Mosser, D. M. The many faces of macrophage activation. *Journal of leukocyte biology* 73(2), 209–212 (2003).

[25] Barrientos, S., Stojadinovic, O., Golinko, M. S., Brem, H., and Tomic-Canic, M. Growth factors and cytokines in wound healing. *Wound Repair and Regeneration* 16(5), 585–601 (2008).

[26] Diegelmann, R. F., Evans, M. C., et al. Wound healing: an overview of acute, fibrotic and delayed healing. *Front Biosci* 9(1), 283–289 (2004).

[27] Lee, W., Harrison, R., and Grinstein, S. Phagocytosis by neutrophils. *Microbes and infection* 5(14), 1299–1306 (2003).

[28] Dovi, J. V., He, L.-K., and DiPietro, L. A. Accelerated wound closure in neutrophil-depleted mice. *Journal of leukocyte biology* 73(4), 448–455 (2003).

[29] Edwards, S. *Biochemistry and Physiology of the Neutrophil*. Cambridge University Press, (2005).

[30] Dovi, J. V., Szpaderska, A. M., and DiPietro, L. A. Neutrophil function in the healing wound: adding insult to injury? *THROMBOSIS AND HAEMOSTASIS-STUTTGART*- 92, 275–280 (2004).

[31] Wang, Y. L. Exchange of actin subunits at the leading edge of living fibroblasts: possible role of treadmilling. *J Cell Biol* 101(2), 597–602 (1985).

[32] Mitchinson, T. and Cramer, L. P. Actin-based cell motility and cell locomotion. *Cell* 84(3), 371–379 (1996).

[33] Pollard, T. D. and Borisy, G. G. Cellular motility driven by assembly and disassembly of actin filaments. *Cell* 112(4), 453–465 (2003).

[34] Doherty, G. J. and McMahon, H. T. Mediation, modulation, and consequences of membrane-cytoskeleton interactions. *Annu Rev Biophys* 37, 65–95 (2008).

[35] Bretscher, M. S. Distribution of receptors for transferrin and low density lipoprotein on the surface of giant hela cells. *Proc Natl Acad Sci U S A* 80(2), 454–458 (1983).

[36] Engelmann, G. Some additions to the north american flora. *Bot Gaz* 6, 223–224 (1881).

[37] Pfeffer, W. Locomotoniche richtungsbewegungen durch chemische reize. *Unter Bot Inst Tubingen* 1, 304–382 (1884).

[38] Boyden, S. The chemotactic effect of mixtures of antibody and antigen on polymorphonuclear leucocytes. *J Exp Med* 1(115), 453–466 (1962).

[39] Lee, J. C. and Young, P. R. Role of csb/p38/rk stress response kinase in lps and cytokine signaling mechanisms. *J Leukoc Biol* 59, 152–157 (1996).

[40] Li Jeon, N., Baskaran, H., Dertinger, S., Whitesides, G., Van de Water, L., and Toner, M. Neutrophil chemotaxis in linear and complex gradients of interleukin-8 formed in a microfabricated device. *Nat Biotechnol* 20, 826–830 (2002).

[41] Schreml, S., Szeimies, R., Prantl, L., Karrer, S., Landthaler, M., and Babilas, P. Oxygen in acute and chronic wound healing. *British Journal of Dermatology* 163(2), 257–268 (2010).

[42] Behm, B., Babilas, P., Landthaler, M., and Schreml, S. Cytokines, chemokines and growth factors in wound healing. *Journal of the European Academy of Dermatology and Venereology* 26(7), 812–820 (2012).

[43] Larousserie, F., Charlot, P., Bardel, E., Froger, J., Kastelein, R. A., and Devergne, O. Differential effects of il-27 on human b cell subsets. *The Journal of Immunology* 176(10), 5890–5897 (2006).

[44] Drews, G. Contributions of theodor wilhelm engelmann on phototaxis, chemotaxis, and photosynthesis. *Photosynth Res* 83(1), 25–34 (2005).

[45] Zigmond, S. Ability of polymorphonuclear leukocytes to orient in gradients of chemotactic factors. *The Journal of Cell Biology* 75(2), 606–616 (1977).

[46] Varani, J., Orr, W., and Ward, P. A comparison of the migration patterns of normal and malignant cells in two assay systems. *The American journal of pathology* 90(1), 159 (1978).

[47] Lauffenburger, D., Rothman, C., and Zigmond, S. Measurement of leukocyte motility and chemotaxis parameters with a linear under-agarose migration assay. *The Journal of Immunology* 131(2), 940–947 (1983).

[48] Zigmond, S. H. Orientation chamber in chemotaxis. *Methods Enzymol* 162, 65–72 (1988).

[49] Zicha, S. H., Dunn, G. A., and Brown, A. F. A new direct-viewing chemotaxis chamber. *J Cell Sci* 99, 769–775 (1991).

[50] Jowhar, D., Wright, G., Samson, P. C., Wikswo, J. P., and Janetopoulos, C. Open access microfluidic device for the study of cell migration during chemotaxis. *Integr Biol (Camb)* 2(11), 648–658 (2010).

[51] The Cell Migration Consortium (CMC) and Nature Publishing Group (NPG). Cell migration gateway., (2008).

[52] Lieschke, G., Oates, A., Crowhurst, M., Ward, A., and Layton, J. Morphologic and functional characterization of granulocytes and macrophages in embryonic and adult zebrafish. *Blood* 98(10), 3087–96, Nov (2001).

[53] Renshaw, S. A. A transgenic zebrafish model of neutrophilic inflammation. *Blood* 108, 3976–3978 (2006).

[54] Mathias, J. R. Resolution of inflammation by retrograde chemotaxis of neutrophils in transgenic zebrafish. *J Leukoc Biol* 80, 1281–1288 (2006).

[55] Brown, S., Tucker, C., Ford, C., Lee, Y., Dunbar, D., and Mullins, J. Class iii antiarrhythmic methanesulfonanilides inhibit leukocyte recruitment in zebrafish. *J Leukoc Biol* 82(1), 79–84, Jul (2007).

[56] Redd, M. J., Kelly, G., Dunn, G., Way, M., and Martin, P. Imaging macrophage chemotaxis in vivo: studies of microtubule function in zebrafish wound inflammation. *Cell Motil Cytoskeleton* 63, 415–422 (2006).

[57] Niethammer, P., Grabher, C., Look, A. T., and Mitchison, T. J. A tissue-scale gradient of hydrogen peroxide mediates rapid wound detection in zebrafish. *Nature* 459, 996–999 (2009).

[58] Hall, C., Flores, M. V., Chien, A., Davidson, A., Crosier, K., and Crosier, P. Transgenic zebrafish reporter lines reveal conserved toll-like receptor signaling potential in embryonic myeloid leukocytes and adult immune cell lineages. *J Leukoc Biol* 85, 751–765 (2009).

[59] Watzke, J., Schirmer, K., and Scholz, S. Bacterial lipopolysaccharides induce genes involved in the innate immune response in embryos of the zebrafish (danio rerio). *Fish Shellfish Immunol* 23, 901–905 (2007).

[60] Sepulcre, M. P., Alcaraz-Perez, F., Lopez-Munoz, A., Roca, F. J., Meseguer, J., Cayuela, M. L., and Mulero, V. Evolution of lipopolysaccharide (lps) recognition and signaling: fish tlr4 does not recognize lps and negatively regulates nf-kappab activation. *J Immunol* 182, 1836–1845 (2009).

[61] Adcock, I. M., Caramori, G., and Chung, K. F. New targets for drug development in asthma. *Lancet* 372, 1073–1087 (2008).

[62] Thalhamer, T., McGrath, M. A., and Harnett, M. M. Mapks and their relevance to arthritis and inflammation. *Rheumatology (Oxford)* 47, 409–414 (2008).

[63] Osman, N., Ballinger, M. L., Dadlani, H. M., Getachew, R., Burch, M. L., and Little, P. J. p38 map kinase mediated proteoglycan synthesis as a target for the prevention of atherosclerosis. *Cardiovasc Hematol Disord Drug Targets* 8, 287–292 (2008).

[64] Cara, D. C., Kaur, J., Forster, M., McCafferty, D.-M., and Kubes, P. Role of p38 mitogen-activated protein kinase in chemokine-induced emigration and chemotaxis in vivo. *The Journal of Immunology* 167(11), 6552–6558 (2001).

[65] Aomatsu, K., Kato, T., Fujita, H., Hato, F., Oshitani, N., Kamata, N., Tamura, T., Arakawa, T., and Kitagawa, S. Toll-like receptor agonists stimulate human neutrophil migration via activation of mitogen-activated protein kinases. *Immunology* 123(2), 171–180 (2008).

[66] Kaminska, B. et al. Mapk signalling pathways as molecular targets for anti-inflammatory therapy–from molecular mechanisms to therapeutic benefits. *Biochimica et biophysica acta* 1754(1-2), 253 (2005).

[67] Huang, C., Jacobson, K., and Schaller, M. D. Map kinases and cell migration. *Journal of cell science* 117(20), 4619–4628 (2004).

[68] Liu, X., Ma, B., Malik, A. B., Tang, H., Yang, T., Sun, B., Wang, G., Minshall, R. D., Li, Y., Zhao, Y., et al. Bidirectional regulation of neutrophil migration by mitogen-activated protein kinases. *Nature immunology* (2012).

[69] Stadtmann, A., Brinkhaus, L., Mueller, H., Rossaint, J., Bolomini-Vittori, M., Bergmeier, W., Van Aken, H., Wagner, D. D., Laudanna, C., Ley, K., et al. Rap1a activation by caldag-gefi and p38 mapk is involved in e-selectin-dependent slow leukocyte rolling. *European journal of immunology* 41(7), 2074–2085 (2011).

[70] Hannigan, M. O., Zhan, L., Ai, Y., Kotlyarov, A., Gaestel, M., and Huang, C.-K. Abnormal migration phenotype of mitogen-activated protein kinase-activated protein kinase 2-/- neutrophils in zigmond chambers containing formyl-methionyl-leucyl-phenylalanine gradients. *The Journal of Immunology* 167(7), 3953–3961 (2001).

[71] Zhang, Y., Bai, X., Zhu, K., Jin, Y., Deng, M., Le, H., Fu, Y., Chen, Y., Zhu, J., Look, A., Kanki, J., Chen, Z., Chen, S., and Liu, T. In vivo interstitial migration of primitive macrophages mediated by jnk-matrix metalloproteinase 13 signaling in response to acute injury. *J Immunol* 181(3), 2155–64, Aug (2008).

[72] Dong, C., Davis, R. J., and Flavell, R. A. Map kinases in the immune response. *Annual review of immunology* 20(1), 55–72 (2002).

[73] Johnson, G. L. and Lapadat, R. Mitogen-activated protein kinase pathways mediated by erk, jnk, and p38 protein kinases. *Science* 298(5600), 1911–1912 (2002).

[74] Han, M. S., Jung, D. Y., Morel, C., Lakhani, S. A., Kim, J. K., Flavell, R. A., and Davis, R. J. Jnk expression by macrophages promotes obesity-induced insulin resistance and inflammation. *Science* 339(6116), 218–222 (2013).

[75] Beltman, J. B., Maree, A. F., and de Boer, R. J. Analysing immune cell migration. *Nat Rev Immunol* 9, 789–798 (2009).

[76] Tranquillo, R., Zigmond, S., and Lauffenburger, D. Measurement of the chemotaxis coefficient for human neutrophils in the under-agarose migration assay. *Cell motility and the cytoskeleton* 11(1), 1–15 (1988).

[77] Tranquillo, R. T., Lauffenburger, D. A., and Zigmond, S. H. A stochastic model for leukocyte random motility and chemotaxis based on receptor binding fluctuations. *J Cell Biol* 106, 303–309 (1988).

[78] Stokes, C. L., Lauffenburger, D. A., and Williams, S. K. Migration of individual microvessel endothelial cells: stochastic model and parameter measurement. *Journal of Cell Science* 99(2), 419–430 (1991).

[79] Stokes, C. L. and Lauffenburger, D. A. Analysis of the roles of microvessel endothelial cell random motility and chemotaxis in angiogenesis. *Journal of theoretical biology* 152(3), 377–403 (1991).

[80] Parkhurst, M. R. and Saltzman, W. M. Quantification of human neutrophil motility in three-dimensional collagen gels. effect of collagen concentration. *Biophysical journal* 61(2), 306–315 (1992).

[81] Painter, K. J. Continuous models for cell migration in tissues and applications to cell sorting via differential chemotaxis. *Bulletin of mathematical biology* 71(5), 1117–1147 (2009).

[82] Murray, J. D. On the mechanochemical theory of biological pattern formation with application to vasculogenesis. *Comptes rendus biologies* 326(2), 239–252 (2003).

[83] Zaman, M. H., Matsudaira, P., and Lauffenburger, D. A. Understanding effects of matrix protease and matrix organization on directional persistence and translational speed in three-dimensional cell migration. *Annals of biomedical engineering* 35(1), 91–100 (2007).

[84] Zaman, M. H. et al. A multisclae probabilisitc framework to model early steps in tumor metastasis. *MOLECULAR AND CELLULAR BIOMECHANICS* 4(3), 133 (2007).

[85] Kim, M.-C., Neal, D. M., Kamm, R. D., and Asada, H. H. Dynamic modeling of cell migration and spreading behaviors on fibronectin coated planar substrates and micropatterned geometries. *PLOS Computational Biology* 9(2), e1002926 (2013).

[86] Zaman, M. H., Kamm, R. D., Matsudaira, P., and Lauffenburger, D. A. Computational model for cell migration in three-dimensional matrices. *Biophysical journal* 89(2), 1389–1397 (2005).

[87] Onsum, M. and Rao, C. V. A mathematical model for neutrophil gradient sensing and polarization. *PLoS Comput Biol* 3, 16;3(3):e36 (2007).

[88]

[89] Andrew, N. and Insall, R. H. Chemotaxis in shallow gradients is mediated independently of ptdins 3-kinase by biased choices between random protrusions. *Nat Cell Biol* 9, 193–200 (2007).

[90] Endres, R. G. and Wingreen, N. S. Accuracy of direct gradient sensing by single cells. *Proc Natl Acad Sci U S A* 105, 15749–15754 (2008).

[91] Van Haastert, P. J. M. A model for a correlated random walk based on the ordered extension of pseudopodia. *PLoS Comput Biol* 6, e1000874 (2010).

[92] Codling, E. A., Plank, M. J., and Benhamou, S. Random walk models in biology. *J R Soc Interface* 5, 813–834 (2008).

[93] Berg, H. C. Random walks in biology. *Princeton University Press* , 164–165 (1983).

[94] Chandrasekhar, S. Stochastic problems in physics and astronomy. *Rev Mod Phys* 15, 1–89, Jan (1943).

[95] Lin, C. C. and Segel, L. A. *Mathematics applied to deterministic problems in the natural sciences*. Macmillan New York, (1974).

[96] Murray, M. P. A drunk and her dog: an illustration of cointegration and error correction. *Am Stat* 48(1), 37–39 (1994).

[97] Clarke, G. M. and Cooke, D. *A basic course in statistics.*, volume 29. Edward Arnold, (1992).

[98] WA, S. *Partial differential equations.* John Wiley Sons Ltd., 2nd edition, (2008).

[99] A, F. Ueber diffusion. *Ann Phys* 170, 59–86 (1855).

[100] Fokker, A. Die mittlere energie rotierender elektrischer dipole im strahlungsfeld. *Ann Phys* 348(5), 810–820 (1914).

[101] Goldstein, S. On diffusion by discontinuous movements, and on the telegraph equation. *Q J Mech Appl Math* 4(2), 129–156 (1951).

[102] Morse, P. M. and Feshbach, H. *Methods of theoretical physics.* New York: McGraw-Hill, (1953).

[103] Stigler, S. M. *The History of Statistics: The Measurement of Uncertainty before 1900*, volume 8. Harvard University Press, (1986).

[104] Bayarri, M. and Berger, J. The interplay of bayesian and frequentist analysis. *Statistical Science* 19(1), 58–80 (2004).

[105] Efron, B. Why isn't everyone a bayesian? *The American Statistician* 40(1), 1–5 (1986).

[106] Bolstad, W. and Wiley, J. *Introduction to Bayesian statistics*, volume 2. Wiley-Interscience New York, (2007).

[107] Berger, O. J., Bernardo, J. M., and Sun, D. The formal definition of reference priors. *Ann Stat* 37(2), 905–938 (2006).

[108] Anscombe, F. J. and Aumann, R. J. A definition of subjective probability. *Ann Math Stat* 34(1), 199–205 (1963).

[109] Goldstein, M. Subjective bayesian analysis: Principles and practice. *Bayesian Anal* 1(3), 406–420 (2006).

[110] Berg, B. A. *Markov Chain Monte Carlo Simulations And Their Statistical Analysis.* World Scientific Publishing, (2004).

[111] Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf., M. P. H. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *J R Soc Interface* 6(31), 187–202 (2009).

[112] Barnes, C., Filippi, S., Stumpf, M. P., and Thorne, T. Considerate Approaches to Achieving Sufficiency for ABC model selection. *arXiv* stat.CO, June (2011).

[113] Berger, J. *Statistical decision theory and Bayesian analysis.* Springer, (1985).

[114] MacKay, D. J. C. *Information theory, inference and learning algorithms.* Cambridge University Press, (2003).

[115] Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., and Sethna, J. P. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol* 3, 1871–1878 (2007).

[116] Jeffreys, H. *Theory of Probability.* Oxford University Press, (1961).

[117] Pritchard, J. K. and Rosenberg, N. A. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 65(1), 220–228 (1999).

[118] Beaumont, M. A., Zhang, W., and Balding, D. J. Approximate bayesian computation in population genetics. *Genetics* 162(4), 2025–2035 (2002).

[119] Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. Markov chain monte carlo without likelihoods. *Proc Natl Acad Sci U S A* 100(26), 15324–15328 (2003).

[120] Sisson, S. A., Fan, Y., and Tanaka, M. M. Sequential monte carlo without likelihoods. *Proc Natl Acad Sci U S A* 106(39), 16889 (2007).

[121] Robert, C. P., Cornuet, J.-M., Marin, J.-M., and Pillai, N. S. Lack of confidence in approximate bayesian computation model choice. *Proceedings of the National Academy of Sciences* 108(37), 15112–15117 (2011).

[122] Secrier, M., Toni, T., and Stumpf, M. The abc of reverse engineering biological signalling systems. *Mol Biosyst* 5(12), 1925–1935, Dec (2009).

[123] Filippi, S., Barnes, C., Cornebise, J., and Stumpf, M. P. H. On optimality of kernels for approximate bayesian computation using sequential monte carlo. *arXiv* (2012).

[124] Kass, R. E. and Raftery, A. E. Bayes factors. *Am Stat* 90(430), 773–795 (1995).

[125] Lewis, S. M. and Raftery, A. E. Estimating bayes factors via posterior simulation with the laplace-metropolice estimator. *Am Stat* 92(438), 648–655 (1997).

[126] Smith, K. On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika* 12(1), 1–85 (1918).

[127] DasGupta, A. 29 review of optimal bayes designs. *Handbook of Statistics* 13, 1099–1147 (1996).

[128] Wald, A. On the efficient design of statistical investigations. *Ann Math Stat* 14(2), 134–140 (1943).

[129] Bose, R. C. The design of experiments. *Proceedings of the Thirty-Fourth Indian Science Congress* (1948).

[130] Karlin, S. and Studden, W. J. *Tchebycheff systems: with applications in analysis and statistics*. Interscience New York, (1966).

[131] Stigler, S. M. Optimal experimental design for polynomial regression. *Am Stat* 66(334), 311–318 (1971).

[132] Kurotschka, V. Optimal design of complex experiments with quantitative factors of influence. *Commun Stat Theory Methods* , 1363–1378 (1978).

[133] Kiefer, J. and Wolfowitz, J. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann Math Stat* 27(4), 887–906 (1956).

[134] Lindley, D. V. On a measure of the information provided by an experiment. *Ann Math Stat* 27(4), 986–1005 (1956).

[135] Shannon, C. E. A mathematical theory of communication. *Bell Sys Tech J* 27, 379–423, 623–656 (1948).

[136] Kullback, S. and Leibler, R. A. On information and sufficiency. *Ann Math Stat* 22(1), 79–86 (1951).

[137] Clyde, M. A. *Experimental Design: A Bayesian Perspective. In nternational Encyclopedia of the Social and Behavioral Sciences.* Elsevier Science New York, (2001).

[138] Taylor, H., Liepe, J., Barthen, C., Bugeon, C., Huvet, M., Kirk, P., Brown, S., Lamb, J., Dallman, M., and Stumpf, M. p38 and jnk have opposing effects on persistence of in vivo leukocyte migration in zebrafish. *Immunol Cell Biol* [In press] (2012).

[139] Chtanova, T., Schaeffer, M., Han, S., van Dooren, G., Nollmann, M., Herzmark, P., Chan, S., Satija, H., Camfield, K., Aaron, H., Striepen, B., and Robey, E. Dynamics of neutrophil migration in lymph nodes during infection. *Immunity* 29(3), 487–496 (2008).

[140] Harms, B., Bassi, G., Horwitz, A., and Lauffenburger, D. Directional persistence of egf-induced cell migration is associated with stabilization of lamellipodial protrusions. *Biophys J* 88(2), 1479–88, Feb (2005).

[141] Lokuta, M., Nuzzi, P., and Huttenlocher, A. Calpain regulates neutrophil chemotaxis. *Proc Natl Acad Sci U S A* 100(7), 4006–11, Apr (2003).

[142] Woodfin, A., Voisin, M., Beyrau, M., Colom, B., Caille, D., Diapouli, F., Nash, G., Chavakis, T., Albelda, S., Rainger, G., Meda, P., Imhof, B., and Nourshargh, S. The junctional adhesion molecule jam-c regulates polarized transendothelial migration of neutrophils in vivo. *Nat Immunol* 12(8), 761–9 (2011).

[143] Edwards, A. M., Phillips, R. A., Watkins, N. W., P., F. M., Murphy, E. J., Afanasyev, V., Buldyrev, S. V., Luz, M. G. E. D., Raposo, E. P., E., S. H.,

and Viswanathan, G. M. Revisiting lévy flight search patterns of wandering albatrosses, bumblebees and deer. *Nature* 449, 1044–1048 (2007).

[144] Potdar, A. A., Jeon, J., Weaver, A. M., Quaranta, V., and Cummings, P. T. Human mammary epithelial cells exhibit a bimodal correlated random walk pattern. *PLoS One* 5, e9636. doi:10.1371/journal.pone.0009636. (2010).

[145] Bennett, C., Kanki, J., Rhodes, J., Liu, T., Paw, B., Kieran, M., Langenau, D., Delahaye-Brown, A., Zon, L., Fleming, M., and Look, A. Myelopoiesis in the zebrafish, danio rerio. *Blood* 98(3), 643–51, Aug (2001).

[146] Young, P., McLaughlin, M., Kumar, S., Kassis, S., Doyle, M., McNulty, D., Gallagher, T., Fisher, S., McDonnell, P., Carr, S., Huddleston, M., Seibel, G., Porter, T., Livi, G., Adams, J., and Lee, J. Pyridinyl imidazole inhibitors of p38 mitogen-activated protein kinase bind in the atp site. *J Biol Chem* 272(18), 12116–21, May (1997). Journal Article United states.

[147] Gardiner, C. *Handbook of Stochastic Methods - for Physics, Chemistry and the Natural Sciences*, volume 3. (2004).

[148] Hsu, K., Traver, D., Kutok, J., Hagen, A., Liu, T., Paw, B., Rhodes, J., Berman, J., Zon, L., Kanki, J., and Look, A. The pu.1 promoter drives myeloid gene expression in zebrafish. *Blood* 104(5), 1291–1297 (2004).

[149] Pau, G., Fuchs, F., Sklyar, O., Boutros, M., and Huber, W. Ebimage - an r package for image processing with applications to cellular phenotypes. *Bioinformatics* 26, 979–81 (2010).

[150] Rudnick, S. and Gaspari, G. *Elements of the Random Walk: An introduction for Advanced Students and Researchers.* Cambridge University Press., (2004).

[151] Campos, D., Méndez, V., and Llopis, I. Persistent random motion: uncovering cell migration dynamics. *J Theor Biol* 267(4), 526–534 (2010).

[152] Li, L., Nørrelykke, S. F., and Cox, E. C. Persistent cell motion in the absence of external signals: A search strategy for eukaryotic cells. *PLoS One* 3(5), e2093. doi:10.1371/journal.pone.0002093 (2008).

[153] Tranquillo, R. T. and Lauffenburger, D. A. Stochastic model of leukocyte chemosensory movement. *J Math Biol* 25(3), 229–262 (1987).

[154] Medzhitov, R. Origin and physiological roles of inflammation. *Nature* 454, 428–435 (2008).

[155] Serhan, C., Chiang, N., and Dyke, T. V. Resolving inflammation: dual anti-inflammatory and pro-resolution lipid mediators. *Nat Rev Immunol* 8, 349–361 (2008).

[156] Bumgarner, S., Neuert, G., Voight, B., Symbor-Nagrabska, A., Grisafi, P., van Oudenaarden, A., and Fink, G. Single-cell analysis reveals that noncoding rnas contribute to clonal heterogeneity by modulating transcription factor recruitment. *Mol Cell* , Jan (2012). Mol Cell. 2012 Jan 18.

[157] Cheong, R., Rhee, A., Wang, C., Nemenman, I., and Levchenko, A. Information transduction capacity of noisy biochemical signaling networks. *Science* 334(6054), 354–8, Oct (2011).

[158] Fujita, K., Toyoshima, Y., Uda, S., Ozaki, Y., Kubota, H., and Kuroda, S. Decoupling of receptor and downstream signals in the akt pathway by its low-pass filter characteristics. *Sci Signal* 3(132), ra56 (2010).

[159] Liepe, J., Taylor, H., Barnes, C., Huvet, M., Bugeon, L., Thorne, T., Lamb, J., Dallman, M., and Stumpf, M. Calibrating spatio-temporal models of leukocyte dynamics against in vivo live-imaging data using approximate bayesian computation. *Integr Biol (Camb)* 10 (2012).

[160] Xu, T., Vyshemirsky, V., Gormand, A., von Kriegsheim, A., Girolami, M., Baillie, G., Ketley, D., Dunlop, A., Milligan, G., Houslay, M., and Kolch, W. Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. *Sci Signal* 3(134), ra20 (2010).

[161] May, R. M. Uses and abuses of mathematics in biology. *Science* 303(5659), 790–793 (2004).

[162] Martins, M. L., Jr, F., and Vilela, M. J. Multiscale models for biological systems. *Curr Opin Colloid Interface Sci* 15(1-2), 18–23 (2010).

[163] Schnell, S., Grima, R., and Maini, P. K. Multiscale modeling in biology - new insights into cancer illustrate how mathematical tools are enhancing the understanding of life from the smallest scale to the grandest. *Am Sci* 95(2), 134–142, March (2007).

[164] Alt, W. Biased random walk models for chemotaxis and related diffusion approximations. *J Math Biol* 9(2), 147–177 (1980).

[165] Liepe, J., Barnes, C., Cule, E., Erguler, K., Kirk, P., Toni, T., and Stumpf, M. P. H. Abc-sysbio - approximate bayesian computation in python with gpu support. *Bioinformatics* 26, 1797–1799 (2010).

[166] Plyasunov, S. and Arkin, A. P. Efficient stochastic sensitivity analysis of discrete event systems. *J Comp Phys* 2, 724–738 (2007).

[167] Joyce, P. and Marjoram, P. Approximately sufficient statistics and bayesian computation. *Stat Appl Genet Mol Biol* 7(1), January (2008). Look up nearly sufficient statistics.

[168] Nunes, M. A. and Balding, D. J. On Optimal Selection of Summary Statistics for Approximate Bayesian Computation. *Stat Appl Genet Mol Biol* 9(1), – (2010).

[169] Erguler, K. and Stumpf, M. P. H. Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models. *Mol Biosyst* 7, 1593–1602, March (2011).

[170] Liepe, J., Filippi, S., Komorowski, M., and Stumpf, M. Maximizing the information content of experiments in systems biology. *PLoS Comput Biol* 9(1), e1002888.

[171] Wilkinson, D. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat Rev Genet* 10(2), 122–133 (2009).

[172] Brenner, S. Sequences and consequences. *Philos Trans R Soc Lond B Biol Sci* 365(1537), 207–212, January (2010).

[173] Sebastiani, P. and Wynn, H. Maximum entropy sampling and optimal bayesian experimental design. *J R Stat Soc Series B Stat Methodol* 62(1), 145–157 (2000).

[174] Paninski, L. Estimation of entropy and mutual information. *Neural Comp* 15(6), 1191–1253 (2003).

[175] Hausser, J. and Strimmer, K. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *J Mach Learn Res* 10, 1469–1484 (2009).

[176] Zhou, Y., Liepe, J., Sheng, X., Stumpf, M., and Barnes, C. Gpu accelerated biochemical network simulation. *Bioinformatics* 27(6), 874–876 (2011).

[177] Elowitz, M., Leibler, S., et al. A synthetic oscillatory network of transcriptional regulators. *Nature* 403(6767), 335–338 (2000).

[178] Hirata, H., Yoshiura, S., Ohtsuka, T., Bessho, Y., Harada, T., Yoshikawa, K., and Kageyama, R. Oscillatory expression of the bhlh factor hes1 regulated by a negative feedback loop. *Sci STKE* 298(5594), 840 (2002).

[179] Silk, D., Kirk, P., Barnes, C., Toni, T., Rose, A., Moon, S., Dallman, M., and Stumpf, M. Designing attractive models via automated identification of chaotic and oscillatory dynamical regimes. *Nat Commun* 2, 489 (2011).

[180] Bazil, J. N., Buzzard, G. T., and Rundell, A. E. A global parallel model based design of experiments method to minimize model output uncertainty. *Bull Math Biol* 74(3), 688–716, March (2012).

[181] Vivanco, I. and Sawyers, C. L. The phosphatidylinositol 3-kinase–akt pathway in human cancer. *Nature Reviews Cancer* 2(7), 489–501 (2002).

[182] Shaw, R. J. and Cantley, L. C. Ras, pi (3) k and mtor signalling controls tumour cell growth. *Nature* 441(7092), 424–430 (2006).

[183] Manning, B. D. and Cantley, L. C. Akt/pkb signaling: navigating downstream. *Cell* 129(7), 1261 (2007).

[184] Huang, J. and Manning, B. D. The tsc1–tsc2 complex: a molecular switchboard controlling cell growth. *The Biochemical journal* 412(2), 179 (2008).

[185] Li, Y., Corradetti, M. N., Inoki, K., and Guan, K.-L. Tsc2: filling the gap in the mtor signaling pathway. *Trends in biochemical sciences* 29(1), 32–38 (2004).

[186] Toyoshima, Y., Kakuda, H., Fujita, K., Uda, S., and Kuroda, S. Sensitivity control through attenuation of signal transfer efficiency by negative regulation of cellular signalling. *Nat Commun* 3, 743 (2012).

[187] Hengl, S., Kreutz, C., Timmer, J., and Maiwald, T. Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics* 23(19), 2612–2618 (2007).

[188] Chu, Y. and Hahn, J. Integrating parameter selection with experimental design under uncertainty for nonlinear dynamic systems. *AIChE J* 54(9), 2310–2320 (2008).

[189] Bandara, S., Schloeder, J., Eils, R., Bock, H., and Meyer, T. Optimal Experimental Design for Parameter Estimation of a Cell Signaling Modell. *PLoS Comput Biol* 5(11), e1000558 (2009).

[190] Vanlier, J., Tiemann, C., Hilbers, P., and van Riel, N. A bayesian approach to targeted experiment design. *Bioinformatics* (2012).

[191] Apgar, J. F., Toettcher, J. E., Endy, D., White, F. M., and Tidor, B. Stimulus design for model selection and validation in cell signaling. *PLoS Comput Biol* 4(2), e30, February (2008).

[192] Mélykúti, B., August, E., Papachristodoulou, A., and El-Samad, H. Discriminating between rival biochemical network models: three approaches to optimal experiment design. *BMC Syst Biol* 4, 38, January (2010).

[193] Cox, D. *Principles of Statistical Inference*. Cambridge University Press, Cambridge, (2006).

[194] Kutalik, Z., Cho, K.-H., and Wolkenhauer, O. Optimal sampling time selection for parameter estimation in dynamic pathway modeling. *Biosystems* 75(1-3), 43–55, July (2004).

[195] Casey, F. P., Baird, D., Feng, Q., Gutenkunst, R. N., Waterfall, J. J., Myers, C. R., Brown, K. S., Cerione, R. A., and Sethna, J. P. Optimal experimental design in an epidermal growth factor receptor signalling and down-regulation model. *IET Syst Biol* 1, 190–202, January (2007).

[196] Apgar, J. F., Witmer, D. K., White, F. M., and Tidor, B. Sloppy models, parameter uncertainty, and the role of experimental design. *Mol Biosyst* 6(10), 1890 (2010).

[197] Huan, X. and Marzouk, Y. Simulation-based optimal bayesian experimental design for nonlinear systems. *Arxiv preprint arXiv:1108.4146* (2011).

[198] Vanlier, J., Tiemann, C., Hilbers, P., and van Riel, N. An integrated strategy for prediction uncertainty analysis. *Bioinformatics* (2012).

[199] Komorowski, M., Costa, M. J., Rand, D. A., and Stumpf, M. P. H. Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proc Natl Acad Sci U S A* 108(21), 8645–8650, May (2011).